La régression logistique en épidémiologie

Jean Bouyer

edp sciences

Du même auteur :

Bouyer J, Dreyfus J, Gueguen S, Lazar P, Papiernik E : La prématurité. Enquête périnatale de Haguenau. Inserm et Doin Éditeurs 1987

Papiernik E, Keith L G, Bouyer J, Dreyfus J, Lazar P : Effective prevention of preterm birth: The French experience measured at Haguenau. March of Dimes Foundation. Birth Defects: Original Article Series. Volume 25 Number 1, 1989

Bouyer J, Hémon D, Cordier S, Derriennic F, Stücker I, Stengel B, Clavel J : Épidémiologie – Principes et méthodes quantitatives. Éditions Inserm, 1993 (2e édition : Lavoisier 2009)

Schwartz D, Bouyer J : Statistique en médecine et en biologie – Exercices corrigés et commentés. Flammarion Médecine-Sciences 1994

Bouyer J : Méthodes statistiques. Médecine – Biologie. Exercices corrigés. Estem, 1999

Bouyer J : Épidémiologie. Exercices corrigés. Estem, 2001

Bouyer J : Méthodes statistiques. Médecine – Biologie. Avec exercices corrigés. Vuibert, 2017 (1re édition Estem, Éditions Inserm, 1996)

Biographie : https://www.inserm.fr/portrait/histoire/jean-bouyer/

Publications scientifiques : https://orcid.org/0000-0003-4861-0783

Composition et mise en pages : Flexedo
Conception graphique de la couverture : CB Defretin, Lisieux.

Imprimé en France
ISBN (papier) : 978-2-7598-3817-2
ISBN (ebook) : 978-2-7598-3818-9
DOI : https://doi.org/10.1051/978-2-7598-3817-2

La régression logistique en épidémiologie

Jean Bouyer

edp sciences

À Clémence, Adèle et Suzanne

À Marion

Sommaire

Chapitre 3

Codage des variables et interprétation des coefficients

Chapitre 4

Modélisation des variables quantitatives

Chapitre 5

Choix des variables à inclure dans un modèle logistique

Chapitre 6

Régressions logistiques multinomiale et ordinale

Chapitre 7

Adéquation du modèle

Préface

Un nouveau livre de Jean Bouyer, c'est toujours un événement.

Qu'on ne se trompe pas à la lecture du titre et du sommaire, (trop ?) sobres. C'est un livre sur la régression logistique, mais comme il est rappelé dans le chapitre 1, ce qui est développé par la suite s'applique plus largement à d'autres modèles, tels la régression linéaire multiple ou le modèle de Cox. En réalité, il s'agit d'un livre sur la stratégie d'analyse quantitative en épidémiologie analytique, et à ce titre c'est un livre qu'il est indispensable de lire. Car, ainsi que Jean le rappelle, « les logiciels ne sont pas attentifs, c'est au chercheur/utilisateur de l'être ».

Nous sommes nombreux à avoir suivi les cours de Jean Bouyer, j'en ai fait partie, cours où l'humour pointe rapidement derrière le sérieux de l'enseignant. J'ai bénéficié de ses merveilleux « polys », à une époque où peu d'enseignants se donnaient la peine d'écrire leurs cours, ni de les actualiser chaque année. Car, en effet, les cours de Jean Bouyer ne sont pas inertes, ils vivent et ils évoluent, avec l'évolution des connaissances et des problématiques, avec chaque nouvelle promotion d'étudiants et leurs questions/demandes/suggestions, avec l'évolution des logiciels (il a ainsi accueilli dans sa palette de compétences, à côté de son cher Stata, le désormais incontournable R). En tant que responsable du M2 « Recherches en santé publique » de Paris-Saclay, je l'ai vu chaque année et sans surprise caracoler largement en tête des évaluations des enseignements par les étudiants.

Ce livre est à son image : scientifique, argumenté, intègre, allant loin dans les réponses à des questions « simples » ; ouvrant des portes, apportant des nuances, jamais des diktats. C'est un livre nourri de sa très riche expérience d'enseignement et de recherche en épidémiologie.

Lecteurs débutants comme chercheurs chevronnés y trouveront comment démarrer, poursuivre, s'améliorer, se recycler (j'ai déjà commencé pour ma part !). Je recommande particulièrement les petites notes et les remarques, à savourer. On retrouvera avec un immense plaisir l'humour de Jean dans nombre d'entre elles.

Très bonne(s) lecture(s) !

Pr. Laurence Meyer
Faculté de médecine, université Paris-Saclay

Préambule

Historiquement, l'épidémiologie était consacrée à l'étude des épidémies. L'intérêt était centré sur les maladies transmissibles et sur la comptabilisation du nombre de cas de maladie. Aujourd'hui, l'épidémiologie s'intéresse toujours aux maladies transmissibles, et son nom, jusqu'alors confidentiel, est devenu commun avec la pandémie de Covid-19, mais elle s'intéresse aussi plus largement à l'ensemble des événements de santé, notamment aux maladies chroniques (Morabia A, (ed), 2004), et ses concepts et méthodes ont beaucoup évolué depuis la fin du XIXe siècle (Leplège A et al., 2011).

Elle peut se définir comme l'étude de la distribution des maladies et des facteurs qui influencent cette distribution (Last J, 1983, Leclerc A et al., 1990).

L'étude de la distribution des maladies fait l'objet de l'épidémiologie descriptive. Cette distribution peut être caractérisée par la prévalence ou l'incidence de la pathologie. En épidémiologie analytique, on privilégie l'incidence, plus proche d'une relation causale, en s'intéressant aux cas apparus dans un intervalle de temps fixé ou depuis une date fixée.

Les facteurs qui influencent la distribution d'une maladie sont ce qu'on appelle des facteurs de risque. Ce terme ne se limite pas aux causes de la maladie, bien qu'elles soient le but ultime des recherches, mais englobe les variables associées statistiquement à la maladie. C'est ce qui explique le rôle central des statistiques comme outil de l'épidémiologie. La plupart du temps, il est cependant insuffisant de se limiter à l'existence d'un lien statistique, et de donner une réponse dichotomique pour l'existence d'une association (significatif oui/non).

L'épidémiologie quantitative vise à quantifier la force de l'association entre un facteur de risque et la maladie et à déterminer sa forme dans le cas d'un facteur de risque quantitatif, ce qui peut permettre de mieux comprendre les mécanismes le liant à la maladie.

Dans ce livre, je me limite aux cas où la maladie est caractérisée par une variable en 0/1 (malade non/oui), à part une petite digression (un chapitre quand même !) pour introduire les modèles logistiques multinomiaux et ordinaux. Les facteurs de risque peuvent, quant à eux, être qualitatifs (dichotomiques, nominaux ou ordinaux) ou quantitatifs.

Le contenu du livre est issu d'un cours intitulé « Épidémiologie quantitative », que j'ai donné pendant plusieurs années (je préfère ne pas les compter…) dans le cadre

du master 2 « Recherches en santé publique » de la faculté de médecine de l'université Paris-Sud, devenue ensuite l'université Paris-Saclay. Les données qui servent d'exemples viennent d'enquêtes épidémiologiques auxquelles j'ai participé au sein des unités de recherche de l'Inserm auxquelles j'ai appartenu. Ces unités ont contribué en 2010 à la création du CESP (Centre de recherche en épidémiologie et santé des populations), auquel je suis très attaché.

Le cours du master 2 a sensiblement évolué au cours de temps grâce aux étudiants, auprès de qui j'ai cherché à être le plus précis et complet possible tout en restant clair et compréhensible, sans recourir à un bagage mathématique approfondi. Je les remercie beaucoup pour leurs questions et leurs commentaires. Mon contact avec eux a été mon plaisir et ma récompense au long de ces années.

Le contenu et la forme du cours doivent aussi beaucoup aux enseignants qui ont animé les TP et à qui je suis très reconnaissant de leur fidélité : Guillemette Antoni, Béatrice Ducot, Isabelle Jaussent, Nathalie Lelong, Jean-Paul Teglas. J'adresse des remerciements particuliers à Guillemette Antoni, qui m'a beaucoup aidé à apprendre la programmation en R et à qui les lignes de codes R qui accompagnent ce livre (voir chapitre 1, § VI) doivent beaucoup.

Enfin, tous mes remerciements et mon amitié vont à Freddy Spira et Laurence Meyer, qui ont dirigé successivement le M2R de santé publique et m'y ont donné une place dont j'ai longtemps profité.

Ce livre et ce cours ont aussi une histoire plus ancienne qui a commencé à la fin des années 1980, notamment avec un polycopié rédigé pour un séminaire à Pékin organisé par Joseph Lellouch, Pierre Ducimetière, Denis Hémon et Monique Kaminski (Lellouch J et al., 1988). Un enseignement formidable de nouveauté à cette époque. Ce cours a ensuite été donné dans le cadre de DEA de santé publique de l'université Paris-Sud dirigé par J. Lellouch, où j'ai beaucoup appris, et qui est l'ancêtre (le père en réalité…) du master 2 « Recherches en santé publique » de Paris-Saclay. Je suis un des enfants de cette histoire. J'ai grandi et vieilli avec elle et j'espère que ce livre contribuera à ce qu'elle se poursuive.

Novembre 2024

Chapitre 1

Introduction – Le modèle logistique

• • •

I. Brève présentation des modèles multivariés

I.1. Introduction

En épidémiologie quantitative, on cherche à quantifier la force d'une association entre une (ou des) exposition(s) et la maladie. Les tests et la signification statistiques gardent bien sûr un intérêt. Malgré son côté arbitraire, le seuil « inamovible » de 5 %

permet d'éviter une dérive qui conduirait à présenter et à interpréter des associations de moins en moins significatives sans imaginer qu'elles peuvent ne résulter que du hasard (Bouyer J, 2017). La quantification de la force de l'association est cependant beaucoup plus informative. Un OR (Odds Ratio) et son intervalle de confiance donnent à la fois la force de l'association et le fait qu'elle soit significative si l'intervalle de confiance ne contient pas 1. Cela ne donne pas le degré de signification p, mais p n'est pas un outil privilégié pour interpréter la force d'une association car sa valeur dépend, notamment, de la taille de l'échantillon (van Rijn MH et al., 2017).

La question centrale des études épidémiologiques est causale : est-ce que tel facteur est la cause (ou une cause) de la maladie ou de sa guérison ? Il faut peut-être mettre un peu à part les études descriptives dont l'objectif est de donner une photographie de la situation, tout en notant que dès qu'on décrit des variations de l'incidence d'une maladie, par exemple géographiques ou temporelles, on met un pied dans la causalité. La causalité est hors de portée des enquêtes d'observation, qui sont les plus fréquentes en épidémiologie. Il est cependant possible de s'en approcher, ou du moins d'éliminer des mécanismes d'association alternatifs.
D'une part, au moment du protocole, en limitant les principaux biais. C'est pourquoi ce premier chapitre contient une longue annexe consacrée aux biais en épidémiologie, avec un accent particulier sur les biais de confusion.
D'autre part, et surtout, au moment de l'analyse, en utilisant des modèles multivariés incluant les variables qui permettent de contrôler, au moins en partie, les phénomènes de confusion et d'étudier la forme ou le mode de cheminement de l'association entre les expositions et la maladie.

Ce livre est consacré à un des modèles multivariés les plus utilisés en épidémiologie : le modèle logistique[1].

I.2. Généralités sur les modèles multivariés

Un modèle multivarié permet d'exprimer une variable Y en fonction de plusieurs variables X_i.
En épidémiologie, Y « est » une maladie, ou du moins un événement de santé. Je mets des guillemets car la façon de caractériser la maladie n'est pas unique. Cela peut être une variable dichotomique (malade oui/non), une variable qualitative à plus de deux classes (type ou gravité d'un cancer par exemple) ou une variable quantitative, comme un dosage sanguin. On verra dans le paragraphe suivant que faire varier la nature de Y conduit à des modèles multivariés différents.
Les X_i caractérisent les facteurs dont on veut étudier l'association avec la maladie et peuvent être des variables qualitatives ou quantitatives. Le fait qu'il y ait plusieurs X_i permet une analyse fine des facteurs de risque de la maladie tenant compte de

1. Les termes « modèle logistique » et « régression logistique » sont employés indifféremment par les épidémiologistes. Dans ce livre, le terme « modèle logistique » est le plus fréquent. Sauf dans le titre, allez savoir pourquoi !

l'ensemble des facteurs et de phénomènes de confusion ou d'interaction. Trois chapitres préciseront sous quelle forme les X_i peuvent être inclus dans un modèle multivarié (chapitres 3 et 4) et comment les choisir (chapitre 5).
Les modèles multivariés supposent une certaine *modélisation* de la réalité. C'est une sorte d'évidence, puisque c'est contenu dans leur nom ! Mais il faut bien réaliser que cela veut dire que le recours à ces modèles suppose que certaines hypothèses sur la relation entre Y et les X_i sont satisfaites. Par exemple, l'utilisation de la régression linéaire suppose que la relation entre Y et les X_i est une droite, ou du moins que sa représentation (ou sa modélisation) par une droite est acceptable. De même, la façon dont les variables X_i sont incluses dans un modèle logistique peut impliquer l'absence d'interaction entre ces variables. Ou du moins, là aussi, que la représentation des relations entre Y et les X_i sans interaction est acceptable.

Les conclusions qu'on tire des analyses multivariées sont donc en partie conditionnées par le bien-fondé des hypothèses faites, notamment sur la forme de la relation entre Y et X (par exemple, linéarité). Ces hypothèses sont souvent implicites, ou du moins pas entièrement exprimées. Certains aspects statistiques interviennent aussi. On verra par exemple que les paramètres d'un modèle logistique sont estimés par la méthode du maximum de vraisemblance, dont la validité nécessite que la taille de l'échantillon soit suffisamment grande.
En pratique, il est fréquent qu'on ne puisse pas vérifier les hypothèses faites autrement que par la non-significativité d'un test dont on ne connaît pas la puissance. Je pense en particulier à la linéarité ou à l'absence d'interaction.
Il ne faut pas que ce soit bloquant, car les modèles sont le plus souvent robustes, c'est-à-dire que leurs résultats restent valides même si on s'écarte (raisonnablement…) des hypothèses nécessaires. Mais il faut avoir cela en tête dans l'interprétation des résultats et l'analyse de leur solidité, et conserver un certain recul.

I.3. Principaux modèles multivariés utilisés en épidémiologie

De nombreux modèles multivariés sont utilisés en épidémiologie pour permettre d'analyser les situations les plus variées, que ce soit du point de vue de l'organisation des données (par exemple structure hiérarchique, observations longitudinales) ou de la question posée (par exemple recherche de facteurs de risque, enchaînements de causes et d'effets avec les modèles d'équations structurales) – voir par exemple les livres de Commenges (Commenges D et al., 2015) et de Falissard (Falissard B, 2019).

Je me limiterai ici aux trois principaux modèles utilisés en épidémiologie, en les classant selon la nature de la variable Y : modèle (ou régression) linéaire, modèle logistique et modèle de Cox.
Ces modèles expriment la *moyenne* de Y en fonction de variables X_1, …, X_p. Ils ne sont pas faits pour prédire les valeurs individuelles de Y en fonction des X_i, qui restent soumises à la variabilité biologique autour de la moyenne.

I.3.a. Modèle linéaire (Y est quantitative)

Lorsque la maladie ou la caractéristique de santé est mesurée par une variable quantitative (indice de masse corporelle, dosage biologique), on s'intéresse le plus souvent à la variation de sa moyenne et on utilise le modèle linéaire, qui s'écrit :

$$E(Y|X_1,...,X_p) = \alpha + \beta_1 X_1 + ... + \beta_p X_p = \alpha + \sum_{i=1}^{p} \beta_i X_i$$

$E(Y|X_1, ..., X_p)$, qui se lit « espérance de Y sachant X_1, ..., X_p » est la moyenne de Y connaissant X_1, ..., X_p. Le mot espérance est le terme mathématique pour désigner la moyenne.

> Le modèle peut aussi s'écrire sous une autre forme, qui permet (peut-être ?) de mieux comprendre et qui exprime Y (plutôt que l'espérance de Y) en fonction des X_i : $Y = \alpha + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon = \alpha + \sum_{i=1}^{p} \beta_i X_i + \varepsilon$. ε s'appelle le résidu. C'est l'écart entre la valeur observée de Y_j pour un sujet particulier et la moyenne $E(Y|X_1, ..., X_p)$. Dans le modèle linéaire, on suppose que ε a une distribution normale de moyenne 0 et de même variance quelle que soit la combinaison des valeurs X_1, ..., X_p.

I.3.b. Modèle logistique (Y = 0/1)

Lorsque le paramètre de santé auquel on s'intéresse s'exprime en deux catégories (« malades »/« non-malades »), que l'on notera par la suite M^+ et M^-, et que la fréquence de la maladie est mesurée par un risque, c'est le modèle logistique qui est adapté. Il s'écrit :

$$P(M^+|X_1,...,X_p) = \frac{1}{1+\exp\{-(\alpha + \sum \beta_i x_i)\}}$$

$P(M^+)$ est la probabilité de la maladie, c'est-à-dire la probabilité que Y = 1, qui est aussi la moyenne de Y. On a donc bien, comme pour le modèle linéaire :

$$P(M^+|X_1,...,X_p) = E(Y|X_1,...,X_p) = \frac{1}{1+\exp\{-(\alpha + \sum \beta_i x_i)\}}$$

> Lorsque Y a plus de deux classes (Y = 0, 1, ..., k), le modèle logistique se généralise pour devenir un modèle logistique multinomial (ou polytomique). Si, de plus, les catégories de Y sont ordonnées, on peut utiliser l'un des modèles logistiques ordinaux. Voir chapitre 6.

I.3.c. Modèle de Cox (Y = incidence instantanée λ)

Lorsqu'on veut tenir compte non seulement de la survenue de la maladie, mais aussi du moment où elle survient, on a recours à l'incidence instantanée $\lambda(t)$, qui mesure le risque de maladie à un instant t. On utilise, pour l'analyse de ces données, le modèle de Cox, qui s'écrit : $\lambda(t|X_1,...,X_p) = \lambda_0(t) \exp\{\sum_{i=1}^{p} \beta_i X_i\}$.

Les trois modèles précédents ont en commun le « bloc linéaire » $\sum \beta_i X_i$, qui indique la façon de faire figurer les variables dont on veut étudier l'association avec Y. Ici, elles apparaissent sous forme d'une combinaison linéaire. Cela implique notamment que le

choix d'inclure ou non les variables X_i dans le modèle, et sous quelle forme, ainsi que l'interprétation de leur coefficient, répond à des principes très proches d'un modèle à l'autre. Ce qui sera développé à ce sujet dans les chapitres suivants avec le modèle logistique s'applique ainsi à ces trois modèles.
Les deux premiers modèles (linéaire et logistique) font partie de la famille des modèles linéaires généralisés (voir § III).

II. Le modèle logistique

Le modèle logistique (ou régression logistique) est un modèle qui permet d'exprimer une variable Y qualitative à deux classes en fonction de variables X_i qui peuvent être quantitatives ou qualitatives. Dans la suite, les deux catégories de Y seront toujours codées 0 et 1 (voir chapitre 3, § II). Ce modèle est particulièrement utilisé dans le domaine de la santé (et donc en épidémiologie), où le paramètre de santé se répartit très souvent en deux catégories : « malades » (M^+ ou $Y = 1$) ou « non-malades » (M^- ou $Y = 0$).
Le modèle logistique permet d'exprimer, non pas Y en tant que telle, mais la probabilité que $Y = 1$, c'est-à-dire la probabilité de survenue de la maladie quand la valeur des variables X_i est connue : $P(Y = 1 | X_1, X_2, \ldots, X_p)$ ou $P(M^+ | X_1, X_2, \ldots, X_p)$. Comme cela a été souligné plus haut, cette probabilité est aussi la moyenne (ou espérance de Y), puisque Y est codée 0/1.

Dans ce chapitre, on va voir plus précisément comment s'écrit le modèle logistique et ce qui motive sa place privilégiée dans le domaine des études quantitatives en santé. On se limitera dans un premier temps à une seule variable X, de façon à mieux comprendre comment interpréter les coefficients du modèle. Ce point sera plus longuement développé dans le chapitre 3.
Je montrerai ensuite comment l'inclusion de plusieurs variables dans un modèle logistique permet de prendre en compte des facteurs de confusion. La prise en compte d'interactions sera présentée dans le chapitre 3.
Enfin, je dirai un mot de l'utilisation du modèle logistique dans les différents types d'enquêtes épidémiologiques, les enquêtes cas-témoins notamment.

II.1. Écriture du modèle logistique avec une variable X

La fonction logistique est définie par son équation $f(x) = \frac{1}{1 + e^{-(\alpha + \beta x)}}$. Elle a une forme en S comme l'indique la Figure 1.1, dont le tracé précis dépend des valeurs de α et de β.

Le modèle de logistique avec une seule variable X s'écrit : $P(M^+|X) = f(X)$, où f est la fonction logistique, c'est-à-dire : $P(M^+|x) = \frac{1}{1 + e^{-(\alpha + \beta x)}}$.

L'association entre X et la maladie (M^+ ou Y) est mesurée (ou quantifiée) par β. En effet, si $\beta = 0$, il n'y a pas d'association puisque $P(M^+ | X)$ ne dépend pas de X. Inversement, si $\beta \neq 0$, la variation de Y en fonction de X est d'autant plus rapide que β est grand,

ce qui est bien une façon de quantifier la force de l'association sans se limiter à une réponse dichotomique « association significative oui/non ».

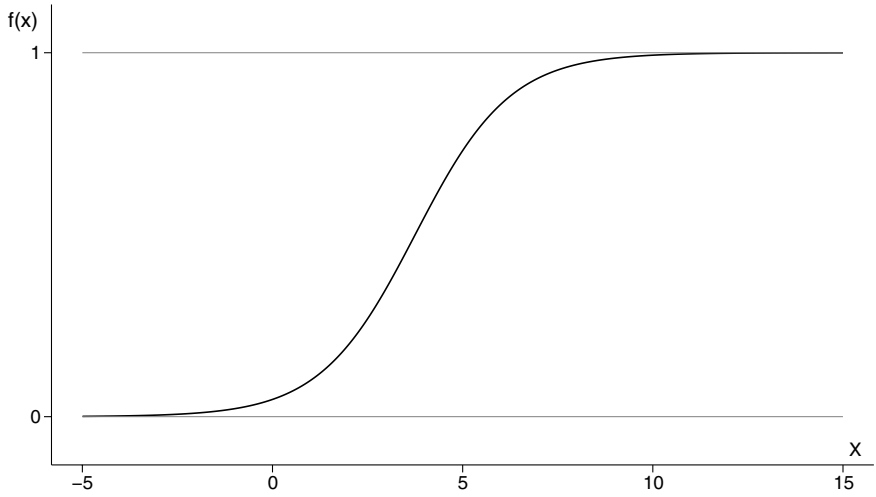


Figure 1.1 : Fonction logistique pour $\alpha = -3$ et $\beta = 0{,}8$.

Deux raisons principales ont conduit au choix de la fonction et du modèle logistique en épidémiologie :

- La fonction logistique a une forme sigmoïde qui correspond à une forme de relation souvent observée entre une dose quantitative X (de produit toxique par exemple) et la fréquence d'une maladie Y ou du décès. Elle revient en pratique à ce qu'il y ait deux seuils pour la dose X : une dose en dessous de laquelle il n'y a pratiquement aucun malade (f(x) est pratiquement égale à 0) et une dose au-dessus de laquelle pratiquement tous les sujets sont malades (f(x) est pratiquement égale à 1). Entre les deux, la relation entre X et la proportion de malades est proche de la linéarité.
- Lorsque X est une variable dichotomique ($X = 1$: exposé et $X = 0$: non exposé), β permet de retrouver l'odds ratio associé à X, qui est une mesure d'association très fréquemment utilisée en épidémiologie (Bouyer J et al., 1993).

En effet, par définition, l'odds ratio est égal à $OR = \dfrac{P_1 / (1-P_1)}{P_0 / (1-P_0)}$, avec $P_0 = P(M^+ \mid X = 0)$ et $P_1 = P(M^+ \mid X = 1)$.

D'après le modèle logistique : $P(M^+ \mid X) = \dfrac{1}{1+e^{-(\alpha+\beta x)}}$. On obtient donc :

- Lorsque $X = 1$ (exposés) : $P_1 = P(M^+ \mid X = 1) = \dfrac{1}{1+e^{-(\alpha+\beta)}}$ et $1-P_1 = \dfrac{e^{-(\alpha+\beta)}}{1+e^{-(\alpha+\beta)}}$
- Lorsque $X = 0$ (non-exposés) : $P_0 = P(M^+ \mid X = 0) = \dfrac{1}{1+e^{-\alpha}}$ et $1-P_0 = \dfrac{e^{-\alpha}}{1+e^{-\alpha}}$

Ce qui donne : $OR = \dfrac{P_1/(1-P_1)}{P_0/(1-P_0)} = \dfrac{1/e^{-(\alpha+\beta)}}{1/e^{-\alpha}} = e^{\beta}$.

Le coefficient β du modèle logistique est ainsi « la même chose » que l'OR habituel utilisé pour mesurer l'association entre une exposition et la maladie (toutes les deux dichotomiques). On passe de l'un à l'autre par l'exponentielle ou le logarithme.

Même si les calculs précédents ne sont pas très compliqués, on voit que l'écriture du modèle logistique conduit à manipuler l'expression $\frac{1}{1+e^{-(\alpha+\beta x)}}$, ce qui n'est pas très agréable. C'est pour cela qu'il est fréquent d'écrire le modèle logistique d'une autre façon, équivalente, mais plus commode en pratique, en simplifiant l'écriture de la partie droite $\frac{1}{1+e^{-(\alpha+\beta x)}}$, quitte à compliquer celle de la partie gauche.

On définit pour cela : $\text{logit } P = \ln \frac{P}{1-P} = \ln P - \ln (1-P))$.

Dans le cas du modèle logistique où $P = \frac{1}{1+e^{-(\alpha+\beta x)}}$ et donc $\frac{P}{1-P} = e^{(\alpha+\beta x)}$, on obtient logit $P = \alpha + \beta X$, ce qui fait apparaître le modèle logistique comme un modèle linéaire, mais avec logit P à la place du Y habituel (on parle de modèle linéaire généralisé, voir paragraphe suivant).

Avec cette écriture du modèle logistique, le calcul précédent pour établir la relation entre β et OR devient nettement plus simple :

$$\ln OR = \ln\left(\frac{P_1/(1-P_1)}{P_0/(1-P_0)}\right) = \ln(P_1/1-P_1) - \ln(P_0/1-P_0) = \text{logit } P_1 - \text{logit } P_0 = (\alpha+\beta) - \alpha = \beta$$

Cette formule ($\ln OR = \text{logit } P_1 - \text{logit } P_0$) doit être retenue, car elle est toujours utile pour exploiter et interpréter les résultats d'un modèle logistique.
Dans le cas où interviennent plusieurs variables, c'est elle qui permet, à partir d'un modèle logistique et même dans des cas compliqués, de calculer l'odds ratio associé à une exposition où les exposés correspondent aux valeurs des X_i qui donnent P_1 et les non-exposés aux valeurs qui donnent P_0. Ici, avec une seule variable, les valeurs de X qui donnent P_1 et P_0 sont 1 et 0.

II.2. Plusieurs variables X_i, prise en compte de facteurs de confusion

Le modèle logistique se généralise à plusieurs variables X_i ($i = 1, \ldots, p$) sous la forme suivante, qui permet d'exprimer la probabilité de survenue de la maladie en fonction des valeurs prises par les variables X_i :

$$P(M^+ \mid X_1,\ldots,X_p) = \frac{1}{1+\exp\{-(\alpha+\sum_{i=1}^{p}\beta_i x_i)\}} \text{ ou encore : } \text{logit } P = \alpha + \sum_{i=1}^{p}\beta_i x_i$$

Comme dans le cas d'une seule variable, les X_i peuvent être dichotomiques, qualitatives à plus de deux classes ou quantitatives. Je reviendrai dans le chapitre 3 sur la forme sous laquelle les variables peuvent (et parfois doivent) être incluses dans un

modèle logistique et sur l'interprétation de leur(s) coefficient(s) ; toutefois, on peut déjà noter que la valeur du coefficient β_i de la variable X_i, et donc l'interprétation de ce dernier, dépend de la présence des autres variables.

Dans le cas vu précédemment où l'on n'est en présence que d'une seule variable, lorsque X augmente d'une unité, logit P augmente de β. Mais, dans un problème à plusieurs variables, lorsque, par exemple, X_1 augmente d'une unité, logit P n'augmente de β_1 *qu'à condition que les autres X_i restent fixes*. Cela signifie que β_1 mesure l'association entre X_1 et la maladie *lorsque les autres variables X_i sont constantes (ou fixées)*. C'est-à-dire précisément que β_1 mesure cette association *ajustée* sur les autres variables.
Lorsqu'il n'y a qu'une variable X et qu'elle est dichotomique, on sait que e^{β} est l'odds ratio associé à X. Avec plusieurs variables, e^{β_1} est toujours l'odds ratio associé à X_1 (si X_1 est dichotomique), mais ajusté sur les autres variables.

Il importe de bien comprendre que, dans la réalité, si deux sujets ont des valeurs de X_1 qui diffèrent d'une unité, il est possible (fréquent même) que les autres X_i soient aussi différents. Ce n'est donc que par calcul (ou modélisation) qu'on « force » les X_i à être identiques lorsque X_1 varie d'une unité.

> Remarque
> On peut d'ores et déjà noter que le raisonnement précédent suppose, sans que cela soit entièrement explicite, que logit P varie d'une même quantité β_1 lorsque X_1 varie d'une unité, quelles que soient les valeurs des autres X_i du moment qu'elles sont fixées.
> En d'autres termes, il n'y a pas d'interaction entre X_1 et les autres variables X_i. Je reviendrai dans le chapitre 3 sur la question de l'interaction, mais cela permet de souligner ici l'importance d'être attentif à la façon dont les variables figurent dans un modèle logistique pour en interpréter correctement les résultats. Les logiciels ne le font pas (« être attentif ») ; c'est à l'utilisateur de s'en charger.

En conclusion, au-delà de sa valeur numérique, c'est l'interprétation du coefficient d'une variable qui dépend de la présence des autres variables. Un modèle logistique (un modèle multivarié de façon générale) doit donc être considéré et interprété dans son ensemble, en tenant compte de l'ensemble des variables qui y sont incluses.

III. Modèles linéaires généralisés

Comme je l'ai indiqué au § I.3, les modèles linéaire et logistique font partie des modèles linéaires généralisés (en anglais GLM, *Generalized Linear Models*), qui comprennent toute une famille de modèles multivariés fréquemment utilisés en statistique et en particulier en biostatistique. Comme leur nom l'indique, ces modèles généralisent le modèle linéaire, ce qui permet d'unifier la théorie statistique et leur utilisation. Ils ont été développés dans le livre de Peter McCullagh et John Nelder (McCullagh P et al., 1989). Je ne donne ici qu'une présentation très résumée, dont l'intérêt

principal est de replacer le modèle logistique dans un contexte plus général et de comprendre les commandes des logiciels statistiques, en particulier R, pour le mettre en œuvre.

Pour présenter les modèles linéaires généralisés, partons du modèle linéaire dont l'écriture est $E(Y|X_1,...,X_p) = \alpha + \sum_{i=1}^{p} \beta_i x_i$. En notant $\mu = E(Y|X_1,...,X_p)$ et $\eta = \alpha + \sum_{i=1}^{p} \beta_i x_i$, ce modèle peut être interprété comme ayant trois composantes :

- La variable réponse Y, dite aussi composante aléatoire ou variable dépendante, qui caractérise la maladie en épidémiologie. Y est elle-même formée des observations y_j ($j = 1,...,n$ pour les n sujets de l'échantillon), qui doivent être indépendantes et avoir une distribution normale de moyenne μ et de variance constante.
- Les prédicteurs X_i, ou variables explicatives, ou variables indépendantes, ou composantes systématiques (on dit aussi déterministes), qui interviennent sous forme de combinaisons linéaires : $\eta = \alpha + \sum_{i=1}^{p} \beta_i x_i$.
- La fonction de lien g qui relie les deux composantes précédentes sous la forme : $g(\mu) = \eta$. Pour le modèle linéaire, g est la fonction identité.

Les modèles linéaires généralisés permettent deux extensions des modèles linéaires :

- ✓ la distribution des y_j peut faire partie de la famille exponentielle sans être limitée à la loi normale ;
- ✓ la fonction de lien peut être une fonction monotone dérivable, autre que l'identité.

Remarques

- Il n'est pas nécessaire ici d'entrer dans le détail de ce qu'est la famille exponentielle. Disons seulement qu'elle inclut la plupart des lois de probabilités usuelles telles que les lois normale, normale inverse, binomiale, Poisson, gamma.
- Le modèle logistique est un GLM (Hilbe JM, 2020), où :
  - Y est la loi binomiale, pour laquelle on a :
    $$\mu = E(Y|X_1,...,X_p) = P(M^+|X_1,...,X_p)|$$
  - la fonction de lien est la fonction logit : $g(x) = \ln\left(\frac{x}{1-x}\right)$

À partir de $g(\mu) = \eta$, on obtient bien le modèle logistique :

$$\ln\left(\frac{\mu}{1-\mu}\right) = \text{logit } P = \alpha + \sum \beta_i x_i.$$

Pour utiliser les GLM avec les logiciels statistiques, il faut parfois (dans R par exemple) indiquer dans la syntaxe de la commande la loi de probabilité de la famille exponentielle (option « family ») et la fonction de lien (option « link »). À titre d'exemple, voici trois combinaisons fréquemment utilisées :

Family	Link	Modèle
normale (gaussienne)	identité	modèle linéaire
binomiale	logit	modèle logistique
Poisson	ln	modèle de Poisson

Remarques

- Dans les trois exemples ci-dessus, la fonction de lien est celle qui est dite « naturellement » associée à la loi de probabilité (on dit que c'est le lien canonique). En général, ce lien est pris par défaut par les logiciels et il suffit d'indiquer la loi de probabilité (family).
- Pour les modèles courants, les logiciels ont parfois une commande « directe » (sans passer par l'écriture d'un GLM). Par exemple, « logit » ou « logistic » pour le modèle logistique dans Stata.

IV. Modèle logistique et type d'enquête

Le modèle logistique modélise la probabilité de survenue de la maladie en fonction de caractéristiques X_i des sujets choisies pour être des facteurs de risque potentiels. Au-delà du fait qu'il est utilisé lorsque la maladie est caractérisée par une variable dichotomique (malade/non malade), il est donc a priori adapté aux cas où les données observées permettent d'estimer cette probabilité, c'est-à-dire ceux où l'enquête est de type transversal ou cohorte.

IV.1. Enquêtes cas-témoins

Dans les enquêtes cas-témoins, on ne peut pas estimer la probabilité de la maladie. Le résultat important est que le modèle logistique peut quand même être utilisé dans les enquêtes cas-témoins.

De façon plus précise, le fait que le modèle logistique peut être utilisé dans les enquêtes cas-témoins signifie que les coefficients β estimés par le modèle logistique à partir de données d'une telle enquête donnent bien les odds ratios ($OR = e^{\beta}$), comme ils le feraient dans d'autres types d'enquête. On peut d'ailleurs noter que cela est cohérent avec le fait que OR peut être estimé dans tous les types d'enquête. La constante α reste cependant non interprétable.

La démonstration est la suivante.

L'échantillonnage d'une enquête cas-témoins peut être considéré comme étant le résultat de deux échantillons tirés au sort séparément dans la population des non-malades ($Y = 0$) et dans celle des malades ($Y = 1$). On note f_0 et f_1 les fractions de sondage correspondantes.

Dans la « vraie vie », il y a rarement tirage au sort, mais on doit se préoccuper de ce que les cas et les témoins soient « représentatifs » de leur population pour qu'il n'y ait pas de biais de sélection (Bouyer J et al., 1993), de sorte que la présentation de l'échantillonnage par un tirage au sort avec des fractions de sondage soit fidèle à la réalité, même si la valeur des fractions de sondage est inconnue.

Bien que, dans une enquête cas-témoins, la probabilité que Y = 1 ne puisse pas être estimée à partir des observations, puisque les effectifs de cas et de témoins ont été fixés arbitrairement par le protocole de l'enquête, on peut cependant écrire, en utilisant la fraction de sondage :

$$P(Y=1|X) = f_1 \frac{1}{1+\exp\{-(\alpha+\sum\beta_i X_i)\}} \text{ et}$$

$$P(Y=0|X) = f_0 \frac{\exp\{-(\alpha+\sum\beta_i X_i)\}}{1+\exp\{-(\alpha+\sum\beta_i X_i)\}} = 1-P(Y=1|X)$$

On en déduit : $$\frac{P(Y=1|X)}{1-P(Y=1|X)} = \frac{f_1}{f_0}\exp\{+(\alpha+\sum\beta_i X_i)\}$$

et donc : $$\text{logit } P = \ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \ln\left(\frac{f_1}{f_0}\right) + (\alpha+\sum\beta_i X_i) = \alpha' + \sum\beta_i X_i.$$

Cela montre que l'échantillonnage de type cas-témoins ne modifie pas les coefficients β_i (et donc les odds ratios OR_i) et leur estimation. Cela montre aussi que la constante α' n'est pas interprétable et donc que la probabilité de survenue de la maladie dans la population ne peut pas être estimée à partir des seules données de l'enquête. La constante résulte en effet directement de la proportion de cas dans l'échantillon d'étude, qui est par exemple de 50 % s'il y a un témoin par cas. Et elle ne peut pas être corrigée, puisqu'elle dépend de fractions de sondage qui sont inconnues.

IV.2. Enquêtes de cohorte ou transversales et estimation du risque relatif

Dans les enquêtes de cohorte ou transversales, il n'y a aucun obstacle de principe à l'utilisation du modèle logistique. Ce dernier quantifie cependant l'association entre les variables X et la maladie par des odds ratios, alors que ce type d'enquête permet de calculer des risques relatifs qui peuvent être préférés en raison de leur interprétation plus facile.

On sait que, si la maladie est peu fréquente, l'OR est peu différent du RR (Bouyer J et al., 1993, Zhang J et al., 1998), ce qui, d'une certaine manière, résout le problème. Dans les cas où la maladie est plus fréquente (disons supérieure à 15 % ou 20 %), des méthodes de calcul permettant de passer de l'OR au RR ont été proposées (Zhang J et al., 1998) et critiquées (McNutt L-A et al., 1999). Avec le développement des moyens de calcul, d'autres méthodes ont été publiées, qui reposent sur des GLM avec des liens autres que le lien logit (Skov T et al., 1998, Barros A et al., 2003, McNutt L-A et al., 2003, Zou G, 2004, Lumley J et al., 2006, Petersen M et al., 2008) ou sur l'utilisation « détournée » d'autres modèles, comme le modèle de Poisson.

IV.3. Enquête avec données non indépendantes

L'estimation des paramètres du modèle logistique par la méthode du maximum de vraisemblance suppose que les observations sont indépendantes. Lorsque ce n'est

pas le cas, en particulier pour des données appariées ou organisées de façon hiérarchique, il faut avoir recours à des méthodes particulières. Le cas particulier des enquêtes multicentriques est présenté dans le chapitre 5, § VI.

V. Annexe 1 : Les biais en épidémiologie

La survenue de biais dans les enquêtes épidémiologiques est une question majeure, que ce soit pour l'analyse ou l'interprétation des résultats. Mon intention ici n'est pas de faire de longs développements à ce sujet. Ce sont des notions connues et on peut se référer à des ouvrages « généralistes » d'épidémiologie (Bouyer J et al., 1993, Rothman KJ et al., 2008). Je ne donne ici que quelques rappels, en mettant l'accent sur les biais de confusion dont la prise en compte est un des objectifs de l'analyse des données. Je serai aussi amené à discuter de l'interaction, qui n'est pas un biais, mais qui est très intriquée avec la confusion.

V.1. Les trois types de biais

Dans une enquête épidémiologique, il y a deux sources d'erreur lorsqu'on on cherche à estimer une fréquence, une moyenne ou une association entre une exposition et une maladie : les erreurs aléatoires et les biais. On prend souvent l'image d'une cible dont on vise le centre avec un pistolet à grenaille, chaque impact représentant un échantillon. Les erreurs aléatoires sont nulles en moyenne (on vise bien le centre de la cible) et leur dispersion diminue quand la taille de l'échantillon augmente ; elles sont prises en compte par les tests statistiques. Les biais (ou erreurs systématiques) ont une moyenne non nulle (on ne vise pas le centre de la cible) et indépendante de la taille de l'échantillon. En général, on préfère un estimateur sans biais, car il est toujours possible d'améliorer sa précision en augmentant la taille de l'échantillon (Bouyer J, 2017).

On distingue trois types de biais en épidémiologie (Bouyer J et al., 1993, Schwartz S et al., 2015) :

- le biais de sélection, qui résulte de la façon dont les sujets sont choisis pour faire partie de l'échantillon analysé (choix des groupes à comparer, non-réponses, *healthy worker effect* pour les études d'exposition professionnelle…) ;
- le biais de classement, qui résulte d'une erreur de mesure sur une ou plusieurs variables ;
- le biais de confusion (ou dû à des facteurs de confusion), qui résulte de ce que le facteur étudié est en partie « mélangé » avec d'autres facteurs.

On trouve d'autres dénominations dans les publications épidémiologiques (biais d'indication, de mémoire, de désirabilité sociale…), mais qui se ramènent toujours à l'un de ces trois types.

Les biais de confusion sont d'une nature différente de celles des biais de sélection et de classement. Ces derniers résultent de l'intervention des investigateurs (épidémiologistes, enquêteurs), par exemple dans le choix du groupe témoin ou des critères

et procédures de diagnostic. Les biais de confusion n'ont pas la même origine. Si, par exemple, les sujets exposés sont en moyenne plus âgés que les sujets non exposés, c'est « la nature des choses », indépendamment de l'intervention de l'épidémiologiste. Le risque relatif brut (sans tenir compte de l'âge) mesure bien le rapport des risques de maladie *dans la population* entre exposés et non-exposés du moment qu'il n'y a pas d'erreur de mesure ou de mauvais choix du ou des échantillons. En ce sens, le risque relatif brut peut avoir un intérêt en tant que tel et peut être pris en considération. Mais il ne peut pas prétendre estimer sans biais le risque relatif propre de l'exposition, puisque cette dernière n'est pas le seul facteur de différence entre les exposés et les non-exposés.
Il arrive cependant que la distinction entre biais de sélection et biais de confusion soit ténue. Par exemple, le *healthy worker effect* (Li CY et al., 1999), qui est habituellement considéré comme un biais de sélection, est le constat du fait que les sujets exerçant une activité professionnelle ont, en moyenne, une meilleure santé que les autres. La seule chose qui le différencie d'un biais de confusion, c'est que ce meilleur état de santé n'est pas caractérisé par une variable bien identifiée (comme l'âge, par exemple). On ne peut donc pas corriger ce biais au moment de l'analyse statistique.

V.2. Confusion, interaction

V.2.a. Association entre maladie et exposition en présence d'autres facteurs de risque

Pour qu'il y ait interaction ou confusion, il faut que plusieurs variables soient prises en compte simultanément. Pour la présentation, je me limiterai ici à des variables qualitatives. La maladie et l'exposition d'intérêt sont des variables dichotomiques qui sont notées M et E. Le tiers facteur (confusion ou interaction) est noté F ; il peut se répartir en plus de deux classes. C'est ce que présente le Tableau 1.1, dans lequel les R_{ij} figurant dans chaque case sont les risques de maladie (prévalence ou incidence selon l'indicateur retenu).

		E		
		–	+	
F	1	R_{10}	R_{11}	$R_{1.}$
	2	R_{20}	R_{21}	$R_{2.}$
	…	…	…	…
	k	R_{k0}	R_{k1}	$R_{k.}$
		$R_{.0}$	$R_{.1}$	$R_{..}$

R_{ij} = risque de maladie
$R_{i.}$ et $R_{.j}$ = risques « marginaux » de maladie

Tableau 1.1 : Risque de maladie en fonction des catégories de E et F

L'association entre un facteur de risque et la maladie est mesurée ici par le risque relatif (RR), mais tout ce qui suit s'applique aussi pour une autre mesure d'association, l'odds ratio OR ou la différence Δ, par exemple.

En présence de F, la mesure de l'association entre l'exposition et la maladie ne peut pas se réduire à un indice unique. On distingue :

- le risque relatif brut (ou risque relatif marginal) $RR = \frac{R_{.1}}{R_{.0}}$: c'est le risque relatif associé à E si on ne tient pas compte de F ;
- les risques relatifs conditionnels $RR_i = \frac{R_{i1}}{R_{i0}}$: ce sont les risques relatifs associés à E selon le niveau i de F ;
- le risque relatif associé à E ajusté sur F (ou risque relatif propre de E). Noté RR_a, c'est la valeur commune des RR_i lorsqu'ils sont égaux.

V.2.b. Définition de l'interaction

On dit qu'il y a interaction entre les facteurs E et F si les RR_i ne sont pas égaux, c'est-à-dire si l'association (ou plus précisément la mesure de l'association) entre E et la maladie n'est pas la même selon le niveau de F.

Il faut noter que la notion d'interaction n'est pas toujours directement liée au mécanisme d'action de l'exposition sur la maladie. En effet, l'existence d'une interaction entre E et F dépend de la mesure d'association choisie et pas seulement du mécanisme biologique (ou social) régissant la relation entre E, F et M. C'est ainsi qu'il ne peut pas y avoir simultanément absence d'interaction avec RR et Δ : s'il n'y a pas d'interaction avec RR, il y en a toujours une avec Δ et réciproquement (sauf si tous les RR_{ij} sont égaux à 1, c'est-à-dire qu'il n'y a pas d'association). C'est la même chose avec RR et OR. De façon plus générale, on peut trouver des situations où l'interaction est nettement significative avec RR, et non significative avec OR.

V.2.c. Définition de la confusion

Par définition, F est un facteur de confusion pour la relation entre E et M si $RR_a \neq RR$ (ou si $OR_a \neq OR$ si on a pris l'odds ratio comme mesure d'association), c'est-à-dire si le risque relatif ajusté est différent du risque relatif brut. On peut noter que cette définition suppose qu'on peut définir RR_a, et donc que les RR_i sont égaux, c'est-à-dire qu'il n'y a pas d'interaction entre E et F. Cette intrication entre confusion et interaction complique (un peu) les choses. J'y reviendrai dans la suite (§ V.2.d).

Le phénomène de confusion associé à F est donc un biais puisque, si on ne tient pas compte de F, la quantité estimée est RR au lieu de RR_a.

Pour décrire le phénomène de confusion, on dit souvent, avec un vocabulaire très (trop) causal que les effets de E et F sur la maladie se « mélangent » selon le schéma triangulaire de la Figure 1.2. La liaison entre E et F (dont le sens importe peu, d'où les doubles flèches) et le fait que F soit facteur de risque de la maladie peut induire une liaison indirecte entre E et M, ou du moins modifier la liaison propre existante.

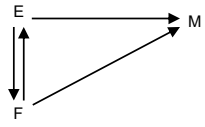


Figure 1.2 : Liaison de F avec E et avec M pouvant induire un phénomène de confusion pour l'association entre E et M.

Il y a ainsi deux façons d'aborder la confusion, voire deux définitions :

- La définition 1 fait explicitement référence à un biais dans la mesure de l'association entre E et M ($RR_a \neq RR$). C'est elle que nous retiendrons en épidémiologie étiologique ou quantitative où l'intérêt principal est la quantification de l'association.
- La définition 2 consiste à dire que F est facteur de confusion parce qu'il est lié à la fois à l'exposition E et à la maladie M (schéma triangulaire). Elle a l'avantage de la commodité pour lister a priori les variables qui sont des facteurs de confusion potentiels ; elle est souvent utilisée dans ce but, comme on le verra dans le chapitre 5. On peut aussi noter que c'est cet abord de la confusion qu'utilise, souvent implicitement, le tirage au sort des essais randomisés.

Ces deux définitions coexistent dans la littérature épidémiologique (Jenicek et Cléroux 1982, Rothman 1986, Kelsey et al. 1986, Kleinbaum et al. 1982, Breslow et Day 1980 – les trois premiers travaux prennent la définition de la Figure 1.1). Tout en n'étant pas strictement équivalentes, elles sont très proches en pratique (Bouyer J et al., 1993).

Pour comprendre la différence entre les deux définitions, il faut calculer le rapport de confusion $RC = \frac{RR}{RR_a}$, qui quantifie le degré de confusion entraîné par F pour l'association entre E et M. Si RC = 1, F n'est pas facteur de confusion ; inversement, si RC ≠ 1, F est facteur de confusion. On se limite ici au cas où F a 2 niveaux.

- Considérons tout d'abord le cas où la mesure d'association est le risque relatif et supposons qu'il n'y a pas d'interaction. On peut montrer (voir un calcul détaillé au §VI du chapitre 5) que le rapport de confusion est égal à $RC = \frac{p_1 RR_{FM/E} + (1-p_1)}{p_2 RR_{FM/E} + (1-p_2)}$, où p_1 est le pourcentage de sujets exposés à F parmi les sujets exposés à E et p_2 est le pourcentage de sujets exposés à F parmi les sujets non exposés à E.

  On voit que RC est égal à 1 si $p_1 RR_{FM/E} + (1-p_1) = p_2 RR_{FM/E} + (1-p_2)$, ce qui équivaut à : $(1-RR_{FM/E})(p_1-p_2) = 0$. On en déduit que F est facteur de confusion (c'est-à-dire que RC ≠ 1) si les deux conditions suivantes sont réunies simultanément :
  - ✓ $RR_{FM/E} \neq 1$, c'est-à-dire que F est associé à M par classe de E. Autrement dit, il y a une association propre entre F et M ;
  - ✓ $p_1 \neq p_2$, ce qui équivaut à $RR_{EF} = \frac{p_1}{p_2} \neq 1$, c'est-à-dire qu'il y a une association brute entre E et F.
- Dans le cas où la mesure d'association est l'odds ratio, on obtient : $RC = \frac{p'_1 OR_{FM/E} + (1-p'_1)}{p'_2 OR_{FM/E} + (1-p'_2)}$, où p'_1 est le pourcentage de sujets non exposés à F parmi les sujets malades et non exposés à E et p'_2 est le pourcentage de sujets non exposés à F parmi les sujets malades et exposés à E.

De même que plus haut, on montre que F est facteur de confusion si les deux conditions suivantes sont réunies simultanément :

- ✓ $OR_{FM/E} \neq 1$, c'est-à-dire que F est associé à M par classe de E – autrement dit, il y a une association propre entre F et M ;
- ✓ $p'_1 \neq p'_2$, qui équivaut à $OR_{EF/M} \neq 1$, c'est-à-dire qu'il y a une association propre entre E et F.

On peut résumer ces résultats en disant que F est facteur de confusion pour l'association entre E et M s'il y a une association entre E et F et une association entre F et M, mais que la nature de cette association dépend de la définition choisie pour la confusion et de la mesure d'association choisie selon le Tableau 1.2.

	Définition 1 ($RR_a \neq RR$)		Définition 2 (schéma triangulaire)
Mesure d'association	RR	OR	RR ou OR
Association entre E et F	brute	propre	brute
Association entre F et M	propre	propre	brute

Tableau 1.2 : Conditions pour que F soit facteur de confusion pour l'association entre E et M selon la définition choisie

On voit donc que les différences entre les deux définitions de la confusion sont ténues. Il est utile de les connaître, mais, en pratique, on utilise les deux définitions indifféremment.

V.2.d. Interaction et confusion

Comme je l'ai souligné plus haut, la première définition de la confusion ($RR_a \neq RR$) suppose qu'on peut définir RR_a, et donc que les RR_i sont égaux, c'est-à-dire qu'il n'y a pas d'interaction. La vérification de l'absence d'interaction paraît donc un point de passage obligé pour pouvoir parler de confusion. En pratique, la situation est beaucoup moins tranchée.

Il y a plusieurs raisons à cela :

- ✓ Sauf dans les enquêtes dont l'objectif est l'étude de l'interaction entre E et F, il n'y a, en général, pas d'hypothèses scientifiques intéressantes sous-jacentes au test de l'interaction. Il ne peut donc pas être satisfaisant de tester l'interaction uniquement dans le but de pouvoir prendre en compte la confusion.
- ✓ L'interaction statistique n'est pas toujours liée à un phénomène biologique ; le choix de sa prise en compte dans les analyses ultérieures doit reposer d'abord sur les mécanismes d'action possibles plutôt que sur des tests statistiques.
- ✓ Même en cas d'interaction, une valeur moyenne des RR_i peut avoir un sens si ces RR_i ont des valeurs proches, ou du même ordre de grandeur. On parle alors parfois d'interaction quantitative. Cela s'oppose à une interaction qualitative (ou irréductible) lorsqu'il n'y a pas d'association dans une catégorie de F alors qu'il y en a une dans une autre catégorie.

Pour ces raisons, il est recommandé de ne pas examiner ou tester systématiquement toutes les interactions entre deux facteurs de risque (voir aussi chapitre 5, § III.4) et de se limiter aux interactions reposant sur des mécanismes d'action possibles et pré-spécifiés. La question de l'interaction ne doit pas constituer un blocage pour la prise en compte d'un biais de confusion.

VI. Annexe 2 : Données d'exemple utilisées et codes informatiques avec Stata et R

Les différentes méthodes d'analyse statistique présentées dans les chapitres de ce livre seront illustrées par des exemples s'appuyant sur les données de plusieurs enquêtes épidémiologiques. Elles sont présentées ici rapidement ; des explications supplémentaires seront données dans les chapitres correspondants.

Les données de ces enquêtes sont accessibles sous différents formats permettant de les analyser avec les logiciels les plus courants (formats Stata, R, et .csv). Il s'agit de données partielles, que ce soit pour le nombre de sujets ou le nombre de variables, et elles sont bien sûr anonymisées. Les fichiers sont disponibles sur https://laboutique.edpsciences.fr/produit/1504/9782759838189/la-regression-logistique-en-epidemiologie.

Je réserve une place particulière au registre des GEU (grossesses extra-utérines) en Auvergne et à l'enquête cas-témoins qui lui est associée. J'y suis « spécialement » attaché parce que j'ai participé au recueil des données et à l'analyse pendant une dizaine d'années aux côtés de Nadine Job-Spira et Joël Coste ; c'est avec ces données que le cours de master, puis le contenu de ce livre, ont été construits.

VI.1. Grossesse extra-utérine

La GEU est une grossesse qui s'implante en dehors de l'utérus, le plus souvent dans les trompes de Falope. Elle ne peut pas s'y poursuivre, mais elle est surtout dangereuse pour la femme, car la trompe de Falope est un tissu très innervé et vascularisé. En grossissant, l'œuf provoque des douleurs intenses, puis, si rien n'est fait, des hémorragies importantes en cas de rupture de la trompe. Sans intervention pour arrêter la grossesse et éliminer l'œuf mal implanté, la vie de la femme est en danger. Le traitement de la GEU peut être chirurgical (incision et conservation ou ablation de la trompe) ou médical (injection intramusculaire de méthotrexate). La fréquence des GEU est de l'ordre de 2 GEU pour 100 naissances.
La GEU représente la complication principale des salpingites (infections des trompes, gonococciques et surtout à *Chlamydia trachomatis*). Le suivi de son incidence peut constituer un indicateur indirect, mais facile à mesurer, d'infection sexuellement transmissible. Or, en France, au début des années 1990, on ne disposait d'aucun outil permettant d'estimer l'incidence de la GEU dans la population générale, ni d'aucun instrument de mesure de ses fluctuations.

C'est ce qui a motivé la mise en place d'un registre des GEU en Auvergne en 1992. Pour améliorer la connaissance des facteurs de risque et des conséquences de la GEU, une étude cas-témoin lui a été associée et la fertilité des femmes a été suivie au cours du temps après le traitement de leur GEU.

VI.1.a. Registre des grossesses extra-utérines en Auvergne

Le registre des GEU en Auvergne couvre les départements du Puy-de-Dôme, du Cantal et de l'Allier. Il a été mis en place en 1992 (en 1993 pour l'Allier) par Nadine Job-Spira, Joël Coste et les obstétriciens et sage-femmes de la région (Coste J et al., 1994). J'y ai contribué à partir de 1996. Le nombre de cas notifiés était d'environ 200 par an.
Toutes les femmes âgées de 15 à 44 ans dont la résidence principale se trouvait dans la zone concernée par le registre et qui étaient traitées chirurgicalement ou médicalement pour une GEU ont été incluses jusqu'en 2005. La notification des cas était effectuée par les maternités publiques et privées et les services de chirurgie générale de la zone couverte.
Les données recueillies comportaient des informations concernant les antécédents médicaux des femmes (en particulier les maladies sexuellement transmissibles et les infections pelviennes), les antécédents gynéco-obstétricaux et chirurgicaux (notamment appendicectomie et chirurgie pelvienne), l'histoire de la contraception, l'activité sexuelle, les conditions de la conception (induction de l'ovulation, contraception au moment de la conception), les habitudes tabagiques, les caractéristiques sociodémographiques du couple. Étaient également relevées des informations concernant la GEU elle-même (signes cliniques, localisation, état de la trompe, prise en charge).
L'exhaustivité du registre a été évaluée chaque année par la méthode capture-recapture. Les données recueillies dans le cadre du Programme de médicalisation du système d'information (PMSI) fournissaient la deuxième source de données, indépendante du registre. Le taux d'exhaustivité est resté stable, autour de 88 % (Coste J et al., 2004).

VI.1.b. Enquête cas-témoins sur les facteurs de risque de GEU

Une enquête cas-témoins associée au registre comprenait tous les cas de GEU enregistrés, à chacun desquels étaient associées deux femmes ayant accouché dans les jours qui suivaient et dans le même centre que celui où la GEU avait été traitée. Dans quelques cas, il n'y avait qu'un seul témoin, et même (rarement) aucun si le cas avait été enregistré très tardivement. Les mêmes informations étaient recueillies chez les témoins et chez les cas. L'enquête a duré jusqu'en 2000 avec 1 065 cas et 1 881 témoins (Bouyer J et al., 2003).
Le choix des témoins dans une enquête sur les facteurs de risque de GEU pose un problème particulier de biais de sélection. Si les femmes qui ont accouché sont bien des « non-malades » (pas de GEU), elles ont aussi choisi ou eu la possibilité de

poursuivre leur grossesse (pas d'IVG ni de fausse couche). Ce n'est pas le cas des GEU, car la GEU est diagnostiquée avant qu'une fausse couche ait pu survenir ou qu'une IVG ait été décidée. La femme apprend souvent en même temps qu'elle est enceinte et que c'est une GEU. Pour éviter le biais de sélection que cela aurait pu induire, l'analyse de l'enquête cas-témoin a été réalisée, selon les recommandations de l'équipe de N. Weiss (Weiss NS et al., 1985), uniquement chez les femmes vivant en couple et ne suivant pas de traitement contraceptif au moment de la conception.

Les données figurant à titre d'exemples dans les chapitres qui suivent portent sur une partie des variables de l'enquête cas-témoins et sur les cas et témoins sélectionnés comme indiqué ci-dessus pour les années 1993 à 1998, soit 574 cas et 1151 témoins.

VI.2. Fichiers de données utilisés

L'ensemble des données utilisées à titre d'exemples dans ce livre est le suivant :

- ✓ Grossesses extra-utérines (fichier geu, voir ci-dessus). Il s'agit d'une enquête cas-témoins réalisée entre 1993 et 1998 dans la région Auvergne dans le but de rechercher les facteurs de risque de la GEU. Elle comprend 574 cas et 1151 témoins. Ces données sont utilisées dans les chapitres 2, 3, 5 et 7.
- ✓ Accouchement après FIV (fécondation in vitro) et âge de la femme (fichier cycles3). Il s'agit des données de 6 400 tentatives de fécondation in vitro réalisées entre 1998 et 2002 dans deux centres français d'assistance médicale à la procréation. Les variables utilisées sont les suivantes : le résultat de la tentative de FIV (accouchement oui/non, variable dichotomique), l'âge de la femme (en années), le nombre d'ovocytes prélevés et le nombre d'embryons de bonne qualité obtenus (c'est-à-dire des embryons que les cliniciens considèrent comme pouvant être réimplantés ou congelés). Ces données sont utilisées dans le chapitre 4.
- ✓ Prématurité après FIV (fichier premafiv). Ce sont les données de 7 601 naissances uniques obtenues après fécondation in vitro en France entre 1986 et 1994 (le fait qu'elles soient survenues après FIV n'a pas d'importance ici). Seules les grossesses singletons sont analysées en raison du risque augmenté de prématurité pour les grossesses multiples, assez fréquentes en FIV. Ces données sont utilisées dans le chapitre 6 pour la régression logistique multinomiale.
- ✓ Rang de succès en FIV (fichier suc_fiv). Ce fichier contient le rang de succès en FIV de 68 566 couples dans 116 centres entre 1986 et 2000, c'est-à-dire le nombre de tentatives successives avant l'obtention d'une grossesse. Le rang de succès est compté jusqu'à 4. Au-delà, ou en cas d'échec, il est codé 9. Ces données sont utilisées dans le chapitre 6 pour la régression logistique ordinale.
- ✓ Santé des adolescents (fichier ado). Il s'agit d'une enquête épidémiologique menée auprès de 15 235 adolescents en milieu scolaire en 2013 pour étudier leur santé et leur vécu par une méthodologie de recherche mixte, quantitative et qualitative.

Seule la partie quantitative est utilisée dans le chapitre 6 pour la régression logistique ordinale.

✓ Vieillissement chez des personnes porteuse du VIH (fichier septavih). Ces données sont un extrait de celles de l'enquête ANRS-Septavih qui porte sur le vieillissement précoce des personnes vivant avec le VIH (virus d'immunodéficience humaine). L'enquête a porté sur 491 sujets. Le vieillissement précoce est quantifié par l'état de fragilité, qui est caractérisé par le score de Fried. Ces données sont utilisées dans le chapitre 6 pour la régression logistique ordinale.

VI.3. Codes informatiques avec Stata et R

Pour la régression logistique comme pour les modélisations statistiques de façon générale, il est impératif de recourir à des logiciels d'analyse spécialisés pour réaliser les calculs. Il est donc important de savoir les utiliser et en lire les résultats. La plupart des logiciels disponibles ont des fonctions semblables, du moins pour ce qui concerne les méthodes présentées dans le présent ouvrage.
Les chapitres de ce livre sont illustrés par les sorties du logiciel Stata. Les codes qui ont servi à les obtenir, ainsi que les codes équivalents en langage R, sont disponibles sur https://laboutique.edpsciences.fr/produit/1504/9782759838189/la-regression-logistique-en-epidemiologie.

Chapitre 2

Estimation et test des paramètres

• • •

L'estimation et le test des paramètres d'un modèle de régression (et du modèle logistique en particulier) utilisent la notion de vraisemblance. En gros, la vraisemblance d'un échantillon est la probabilité de l'observer, certaines hypothèses étant faites sur les vraies valeurs des paramètres dans la population étudiée. L'estimation fait appel à la méthode du maximum de vraisemblance et les tests recourent, entre autres, au rapport entre les vraisemblances maximales.
Je vais commencer par présenter succinctement la notion de vraisemblance, avant d'indiquer comment elle permet d'estimer les paramètres et de les tester.

I. Vraisemblance d'un échantillon

Dans le cas où, comme pour le modèle logistique, on s'intéresse à une variable dichotomique (M^+ / M^-), la vraisemblance d'un échantillon est la probabilité de l'observer étant donné le pourcentage vrai de malades dans la population. Elle est notée V (et en anglais L pour *likelihood*).
Pour montrer explicitement comment se calcule la vraisemblance, considérons une population où le pourcentage vrai de malades, c'est-à-dire la probabilité $P(M^+)$, est égal à P, ainsi qu'un échantillon tiré au sort (représentatif) de cette population.

Commençons par un échantillon ne comprenant qu'un seul sujet, même si en pratique on n'a jamais affaire à un tel échantillon… La vraisemblance de l'échantillon, qui est la probabilité de l'observer, dépend de si le sujet est malade ou pas :

- sujet malade : $V = P$
- sujet non malade : $V = 1 - P$.

Si on prend maintenant un échantillon de deux sujets (tout aussi peu fréquent qu'un échantillon d'un sujet, mais cela permet d'avancer…), la vraisemblance dépend de nouveau de la composition de l'échantillon :

- deux sujets malades (M^+M^+) : $V = P^2$
- deux sujets non malades (M^-M^-) : $V = (1-P)^2$
- un sujet malade et un sujet non malade : il faut tenir compte de ce qu'il y a deux types d'échantillons (M^+M^- et M^-M^+). On a donc : $V = P(1-P) + (1-P)P = 2P(1-P)$.

On peut noter que, dans les deux cas, la somme des vraisemblances associées à chaque type d'échantillon est égale à 1. C'est normal, car tous les échantillons possibles sont décrits ; la somme de leur probabilité doit donc être égale à 1.

Ce résultat se généralise. La vraisemblance d'un échantillon de n sujets avec k malades s'écrit : $V = C_n^k P^k (1-P)^{n-k}$, où k peut varier de 0 à n et où C_n^k est le nombre de compositions possibles d'échantillons de n sujets avec k malades. On montre que $C_n^k = \frac{n!}{k!(n-k)!}$ où $n! = n \times (n-1) \times \ldots \times 1$ (avec $0! = 1$).
On a, là aussi, même si c'est un peu plus difficile à démontrer : $\sum_{k=0}^{n} C_n^k P^k (1-P)^{n-k} = 1$.

On voit que la valeur de la vraisemblance dépend de la composition de l'échantillon (nombre de malades et de non-malades), mais aussi de P. On parle donc plutôt de *fonction vraisemblance*, qui est une fonction de P, c'est-à-dire de la vraie valeur dans la population. Par exemple, pour un échantillon de 20 sujets comprenant cinq malades :

- si $P = 10\,\%$, la vraisemblance (c'est-à-dire la probabilité d'obtenir cet échantillon par tirage au sort parmi tous les échantillons de 20 sujets) est : $V_1 = C_{20}^5 P^5 (1-P)^{15} = \frac{20!}{5!\,15!} \times 0{,}10^5 \times 0{,}90^{15} = 0{,}03$;
- si $P = 25\,\%$, on a $V_2 = C_{20}^5 \times 0{,}25^5 \times 0{,}75^{15} = 0{,}20$.

La vraisemblance V_2 est nettement plus élevée que V_1. Cela signifie qu'il est beaucoup plus probable d'observer cinq malades dans un échantillon de 20 sujets lorsque P = 25 % que lorsque P = 10 %. Réciproquement, si on a effectivement observé un échantillon de 20 sujets comprenant cinq malades, la valeur vraie du pourcentage de malades dans la population est plus vraisemblablement égale à 25 % qu'à 10 %.

On peut noter que, au coefficient C_n^k près, qui reste constant pour un échantillon de composition donnée, la vraisemblance est le produit des probabilités p_i d'observer chacun des sujets i de l'échantillon. En effet, la probabilité d'observer un sujet est P s'il s'agit d'un malade (et il y a k sujets de ce type) et (1 – P) s'il s'agit d'un non-malade (il y a n – k sujets de ce type). On a donc : $V = c\prod_{i=1}^{n} p_i$ où $c = C_n^k$ et $p_i = P$ si le sujet i est malade et $p_i = 1 - P$ si le sujet i est non malade.
Cette écriture sera utile pour calculer la vraisemblance dans le cas d'un modèle logistique. Elle permet aussi de souligner qu'il faut que les observations soient indépendantes pour que la probabilité de l'échantillon soit égale au produit des probabilités de chacun de ses sujets (c'était déjà le cas depuis le début avec des échantillons de deux sujets). C'est pour cette raison que les données appariées (et plus généralement les données non indépendantes) doivent être analysées avec des méthodes particulières.

Pour des raisons de commodité de calcul mathématique, c'est souvent le logarithme de la vraisemblance qui est utilisé. C'est lui qui est donné par les logiciels. C'est un nombre négatif, dont la valeur absolue est d'autant plus grande que la vraisemblance est petite. Ce qui arrive, en particulier, quand la taille de l'échantillon est grande.

On définit aussi la déviance D = – 2 ln(V), qui est donc d'autant plus grande que la vraisemblance est grande.

II. Estimation d'un pourcentage par la méthode du maximum de vraisemblance

II.1. Principe général

Lorsqu'on ne connaît pas P, et qu'on cherche à estimer sa valeur, on prend un échantillon aléatoire de n sujets dans la population. Si on observe k malades dans cet échantillon, on sait que sa vraisemblance (la probabilité de l'observer) est égale à $V = C_k^k P^k (1-P)^{n-k}$.

Le principe de l'estimation du maximum de vraisemblance consiste à choisir comme estimation de P la valeur p_0 qui maximise V, c'est-à-dire qui rend maximum la probabilité d'observer l'échantillon. La « philosophie » est donc de considérer que les observations contiennent l'information et que le « modèle statistique » (ici la valeur de P décrivant la fréquence de la maladie) doit être le plus compatible possible avec cette information.

Avant d'indiquer comment trouver, de façon générale, la valeur p_0 pour laquelle V est maximale, nous allons examiner un exemple.

Exemple

Supposons que n = 15 et k = 3. La vraisemblance s'écrit : $V_1 = C_{15}^3 P^3(1-P)^{12} = 455 P^3(1-P)^{12}$. Comme annoncé, sa valeur dépend de P. On a, par exemple : V = 0,129 lorsque P = 0,10, V = 0,250 lorsque P = 0,20 et V = 9,54 10^{-7} lorsque P = 0,80. Le graphe de la vraisemblance V en fonction de P est représenté sur la Figure 2.1.

On constate que la vraisemblance est maximum pour P = 0,20 (elle vaut alors 0,250). L'estimation du pourcentage de malades dans la population faite à partir de l'échantillon observé est donc $p_0 = 0,20$. On remarque que c'est la valeur observée du pourcentage de malades[1].

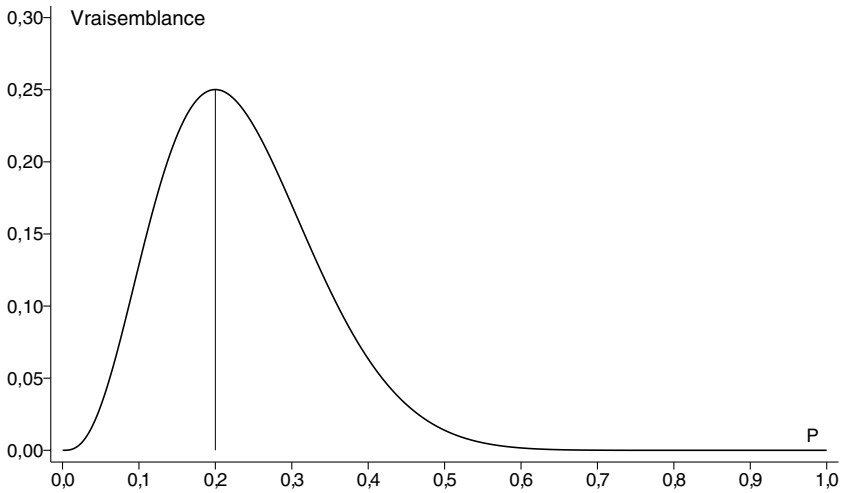


Figure 2.1 : Graphe de la vraisemblance d'un échantillon de 15 sujets comprenant trois malades en fonction de la probabilité P de la maladie dans la population.

Dans le cas général (n et k quelconques), la méthode pour trouver la valeur p_0 de P pour laquelle la vraisemblance est maximale repose sur des calculs mathématiques simples, exposés ci-dessous. On trouve $p_0 = \frac{k}{n}$, c'est-à-dire, comme dans l'exemple précédent, le pourcentage observé de malades.

Démonstration de $p_0 = \frac{k}{n}$

On sait[2] que le maximum d'une fonction est atteint en un point où sa dérivée est nulle.

1. C'est toujours une satisfaction que de constater dans un cas simple qu'une méthode compliquée donne un résultat attendu !

2. Après un petit effort pour se remémorer les cours de maths du lycée…

La dérivée de V est compliquée, mais celle de ln V est beaucoup plus simple, car il s'agit de dériver une somme et non un produit comme avec V. C'est une des raisons pour lesquelles les calculs utilisent ln V plutôt que V. Or, il est équivalent de rechercher le maximum de V et celui de son logarithme, car V et ln V sont maximum pour la même valeur de P.

On a $\ln V = \ln C_n^k + k \ln P + (n-k) \ln (1-P)$.

La dérivée de ln V par rapport à P est : $\frac{d\ln V}{dP} = \frac{k}{P} - \frac{n-k}{1-P}$. Elle est nulle pour $\frac{k}{P} = \frac{n-k}{1-P}$, ce qui donne : $k(1-P) = (n-k)P$ ou encore $P = \frac{k}{n} = p_0$.

II.2. Propriétés des estimateurs du maximum de vraisemblance

Deux des propriétés des estimateurs du maximum de vraisemblance les rendent particulièrement intéressants : ils sont asymptotiquement sans biais et leur distribution est asymptotiquement normale.

Rappelons que le terme « asymptotiquement » signifie que ces propriétés ne sont vraies que lorsque la taille de l'échantillon tend vers l'infini, c'est-à-dire en pratique lorsqu'elle est suffisamment grande. C'est de là que viennent certaines conditions d'application ($n \geq 30$ ou $nP, nQ \geq 5$) habituellement utilisées pour les formules d'intervalle de confiance ou la réalisation des tests (Bouyer J, 2017), qui concrétisent (de façon assez arbitraire il est vrai) ce que signifie « suffisamment grande ».

Enfin, les estimateurs du maximum de vraisemblance ont asymptotiquement la variance la plus faible parmi tous les estimateurs sans biais, et la méthode du maximum de vraisemblance permet de calculer la variance des paramètres estimés (voir § II.3).

Ces propriétés permettent de calculer l'intervalle de confiance des paramètres auxquels on s'intéresse. Elles donnent par ailleurs une justification théorique à l'utilisation des estimateurs intuitifs du pourcentage, de la moyenne et de la variance.

II.3. Intervalle de confiance

On a dit que les estimateurs du maximum de vraisemblance ont, asymptotiquement, une distribution normale. Pour trouver l'intervalle de confiance d'un paramètre estimé par la méthode du maximum de vraisemblance, on peut donc utiliser la formule générale (à condition que l'effectif de l'échantillon sur lequel l'estimation est faite soit assez grand) : l'intervalle de confiance à $1-\alpha$ de μ_A est : $m_A \pm z_{\alpha/2}\sqrt{s_A^2}$ si la distribution de A est normale (Bouyer J, 2017).

Il reste, pour y parvenir, à expliquer comment calculer la variance d'un estimateur du maximum de vraisemblance.

Le principe est le suivant. Soit A le paramètre à estimer. On a vu que la vraisemblance V s'exprime en fonction de la valeur de A, et que, pour en trouver une estimation, on cherche la valeur « a » pour laquelle la dérivée de ln V est nulle. Pour obtenir la variance de l'estimateur de A, on calcule la dérivée seconde de ln V qui, au même

titre que la vraisemblance elle-même, dépend de la valeur de A. Soit $(\ln(V_a))''$ la valeur de cette dérivée seconde au point « a ». On montre qu'une estimation de la variance de A est $\frac{-1}{(\ln(V_a))''}$.

Exemple : intervalle de confiance d'un pourcentage

On a vu (§ II.1) que : $\ln(V) = \ln(C_n^k) + k\ln(P) + (n-k)\ln(1-P)$

et que : $\frac{d(\ln V)}{dP} = \frac{k}{P} - \frac{n-k}{1-P}$.

On en a déduit l'estimateur de P : $p_0 = \frac{k}{n}$.

La dérivée seconde de ln(V) est : $\frac{d^2(\ln V)}{dP^2} = -\frac{k}{P^2} + \frac{n-k}{(1-P)^2}$. En calculant sa valeur au point $a = p_0$ (c'est-à-dire en remplaçant P par p_0 dans la formule) et en remplaçant k par np_0 puisque $p_0 = \frac{k}{n}$, on obtient : $\frac{d^2(\ln V)_{p_0}}{dP^2} = -\frac{n}{p_0(1-p_0)} = -\frac{n}{p_0 q_0}$. Ce qui donne la formule habituelle : $\text{var}(p_0) = \frac{1}{\frac{d^2(\ln V)_{p_0}}{dP^2}} = \frac{p_0 q_0}{n}$.

III. Application au modèle logistique

III.1. Estimation des paramètres

Le modèle logistique s'écrit $P(M^+ \mid X_1,...,X_k) = \frac{1}{1+\exp-\left(\alpha + \sum_{j=1}^{k} \beta_j x_j\right)}$. Il exprime la probabilité d'être malade en fonction des valeurs prises pour les variables X_j et des paramètres α et β_j. Pour estimer ces paramètres, on dispose des observations faites sur un échantillon de n sujets, c'est-à-dire de k-uplets de valeurs pour chaque sujet i (y_i, x_{1i}, ..., x_{ki}) où y_i est égal à 1 ou 0 selon que le sujet i est malade ou non malade et où x_{ji} est la valeur du sujet i pour la variable X_j.

Les estimations des paramètres sont notées $\hat{\alpha}$ et $\hat{\beta}_j$. Ce sont les valeurs qui rendent la vraisemblance de l'échantillon maximum. Pour les obtenir, il faut donc calculer la vraisemblance, ce que je vais faire dans un premier temps dans le cas d'une seule variable X pour des raisons de simplification des écritures.

Le modèle logistique avec une seule variable X s'écrit : $P(M^+ \mid X) = \frac{1}{1+e^{-(\alpha+\beta x)}}$ Les observations faites sur un échantillon de n sujets sont les couples de valeurs (y_i, x_i), avec i = 1, ..., n. On a vu que la vraisemblance de l'échantillon peut s'écrire $V = c\prod_{i=1}^{n} p_i$, où p_i est la probabilité d'observer le sujet i étant donné le modèle logistique, c'est-à-dire sous l'hypothèse que le modèle logistique $P(M^+ \mid X) = \frac{1}{1+e^{-(\alpha+\beta x)}}$ décrit correctement la

réalité. En utilisant le modèle logistique, on peut donc exprimer p_i en fonction de x_i et des paramètres α et β.
On a :

- $p_i = P(M^+ \mid x_i) = \dfrac{1}{1+e^{-(\alpha+\beta x_i)}}$ si le sujet est malade
- $p_i = P(M^- \mid x_i) = 1 - P(M^+ \mid x_i) = \dfrac{e^{-(\alpha+\beta x_i)}}{1+e^{-(\alpha+\beta x_i)}}$ s'il est non malade.

La vraisemblance de l'échantillon est donc : $V = c \dfrac{\prod_{\text{non-malades}} e^{-(\alpha+\beta x_i)}}{\prod_{\text{malades}} (1+e^{-(\alpha+\beta x_i)})}$.

Les estimations du maximum de vraisemblance de α et β sont les valeurs $\hat{\alpha}$ et $\hat{\beta}$ qui rendent V maximum.
Il n'y a de solutions explicites pour $\hat{\alpha}$ et $\hat{\beta}$ que dans le cas d'un seul coefficient β et d'une variable X dichotomique (voir Annexe 1). Sinon (quand il y a plusieurs variables X), il faut procéder par itérations, et les valeurs obtenues sont des approximations numériques.

Remarques :

- On peut trouver une présentation assez simple de la méthode d'itération au début de l'article de Roy et al. (Roy SS et al., 2008).
- Dans le cas du modèle logistique, l'analogue des « conditions d'application » pour utiliser l'estimation par maximum n'est pas aussi universel que « $n \geq 30$ ou nP, $nQ \geq 5$ ». On considère souvent que l'échantillon est assez grand quand il y a au moins 10 événements par variable (voir chapitre 5, § II). C'est-à-dire que si le modèle contient k variables X_i, l'échantillon doit avoir au moins 10 k malades (et 10 k non-malades).
- En réalité, les logiciels donnent ln V et non V, et ne font pas figurer « c » (ce qui ne change rien pour la recherche du maximum). On obtient alors $\ln V = -\sum_{\text{non-malades}} (\alpha+\beta x_i) - \sum_{\text{tous}} \ln(1+e^{-(\alpha+\beta x_i)})$, ce qui peut s'écrire, de façon plus ramassée : $\ln V = \sum_{\text{tous}} [(y_i - 1)(\alpha+\beta x_i) - \ln(1+e^{-(\alpha+\beta x_i)})]$.

III.2. Exemple

Dans l'enquête sur les facteurs de risque de grossesse extra-utérine, on s'intéresse à l'antécédent de salpingite (infection des trompes de Fallope). Cet antécédent est connu à partir de l'interrogatoire de la femme au cours duquel plusieurs questions permettent de déterminer si celle-ci a eu une « salpingite clinique prouvée ou suspectée ». Le diagnostic médical précis n'est pas disponible.

- La variable maladie est notée « ct » (0 = accouchement ; 1 = GEU) (« ct » pour cas-témoins).
- La variable « exposition » est notée « salp » (0 = non ; 1 = oui).

Le modèle s'écrit donc : $\text{logit } P = \alpha + \beta\, \text{salp}$ ou $P(M^+ \mid \text{salp}) = \dfrac{1}{1+e^{-(\alpha+\beta\, \text{salp})}}$.

La répartition des femmes de l'enquête selon la maladie et l'exposition est donnée dans le Tableau 2.1, issu du logiciel Stata.

Lorsqu'on ne considère que les deux variables ct et salp, il y a donc quatre catégories de sujets selon que ces variables valent 0 ou 1. Les sujets d'une même catégorie ne sont pas distinguables les uns des autres ; en particulier, ils ont la même probabilité d'être observés et donc la même contribution à la vraisemblance de l'ensemble de l'échantillon.

```
. tab ct salp

           |  salp clinique prouv
    0:acc  |       ou susp
    1:GEU  |         0          1 |     Total
-----------+----------------------+----------
         0 |     1,122         26 |     1,148
         1 |       466         91 |       557
-----------+----------------------+----------
     Total |     1,588        117 |     1,705
```

Tableau 2.1 : Répartition des sujets de l'enquête GEU selon les deux variables ct (GEU oui/non) et salp (« salpingite clinique prouvée ou suspectée » oui/non)

Remarque : on peut noter qu'il manque 20 sujets (1705 au lieu de 1725). Cela provient de ce qu'il y a des données manquantes pour la variable salp, qui ne figurent pas sur ce tableau.

Le bilan des contributions p_i à la vraisemblance est ainsi le suivant :

$p_1 = \dfrac{1}{1+e^{-(\alpha+\beta)}}$ pour les 91 sujets $E^+ M^+$

$p_2 = \dfrac{1}{1+e^{-\alpha}}$ pour les 466 sujets $E^- M^+$

$p_3 = \dfrac{e^{-(\alpha+\beta)}}{1+e^{-(\alpha+\beta)}}$ pour les 26 sujets $E^+ M^-$

$p_4 = \dfrac{e^{-\alpha}}{1+e^{-\alpha}}$ pour les 1122 sujets $E^- M^-$

La vraisemblance totale est donc : $V = C p_1^{91} p_2^{466} p_3^{26} p_4^{1122} = C \dfrac{\left[e^{-(\alpha+\beta)}\right]^{26}\left[e^{-\alpha}\right]^{1122}}{\left[1+e^{-(\alpha+\beta)}\right]^{117}\left[1+e^{-\alpha}\right]^{1588}}$

Dans l'expression de V, on peut remarquer que les inconnues sont α et β, dont les valeurs seront estimées en cherchant le maximum de V. Les chiffres indiqués en puissance correspondent aux observations faites sur l'échantillon (les x_i, y_i). L'ensemble de ces quatre chiffres permet de reconstituer le Tableau 2.1 de répartition des sujets, qui décrit complètement l'échantillon.

Pour estimer α et β, le programme procède par itération successives. La commande correspondante de Stata (logit) commence par fournir les logs de vraisemblance successifs de ces itérations (qui ne servent en pratique à rien…) avant de donner les résultats des estimations (Tableau 2.2).

Remarques :

- Dans les résultats, les coefficients apparaissent toujours sous le nom des variables qui leur sont associées dans l'écriture du modèle. On trouve donc « salp » et non « β ».
- Il est donc important de connaître l'écriture explicite du modèle pour bien interpréter les résultats.
- Pour α qui n'est associé à aucune variable, l'intitulé de la ligne est _cons (pour constante).
- Il y a ici 557 malades et 1148 non-malades et une variable (salp). Les conditions d'utilisation de l'estimation par maximum de vraisemblance sont amplement satisfaites.

```
. logit ct salp

Iteration 0:   log likelihood = -1077.2309
Iteration 1:   log likelihood = -1023.1123
Iteration 2:   log likelihood = -1023.0534
Iteration 3:   log likelihood = -1023.0534

Logistic regression                             Number of obs   =       1705
                                                LR chi2(1)      =     108.36
                                                Prob > chi2     =     0.0000
Log likelihood = -1023.0534                     Pseudo R2       =     0.0503

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   2.131445    .229102     9.30   0.000     1.682413    2.580476
       _cons |  -.8786825   .0551107   -15.94   0.000    -.9866974    -.7706675
------------------------------------------------------------------------------
```

Tableau 2.2 : Modèle logistique pour analyser la relation entre ct et salp

On obtient donc ici : $\hat{\alpha} = -0{,}879$ et $\hat{\beta} = 2{,}131$
On peut en déduire l'odds ratio, qui est souvent en pratique plus parlant que $\hat{\beta}$: $\widehat{OR} = e^{\hat{\beta}} = 8{,}42$

On note que les résultats contiennent aussi les écarts-types des coefficients estimés : $s_{\hat{\alpha}} = 0{,}0551$ et $s_{\hat{\beta}} = 0{,}229$.

III.3. Intervalle de confiance

Comme les estimateurs de maximum de vraisemblance ont asymptotiquement une distribution normale, on peut obtenir leurs intervalles de confiance avec les formules « habituelles » associées à la loi normale (si l'échantillon est assez grand, ce qui est le cas ici). Par exemple, l'intervalle de confiance de β au risque α est $\hat{\beta} \pm z_{\alpha/2} s_{\hat{\beta}}$, où $z_{\alpha/2}$ est le seuil de la loi normale au risque α/2.
Ici, l'intervalle de confiance à 95 % (ou au risque 5 %) de β est donc : $2{,}131 \pm 1{,}96 \times 0{,}229 = [1{,}682\ ;\ 2{,}580]$. C'est ce qu'on trouve dans les deux dernières colonnes des résultats du Tableau 2.2.

On en déduit l'intervalle de confiance de l'OR : $\left[e^{1{,}682}\ ;\ e^{2{,}580}\right] = [5{,}38\ ;\ 13{,}20]$.

Remarques :

- On peut obtenir directement l'odds ratio et son intervalle de confiance avec la commande « logistic » de Stata. Attention cependant : dans ce cas, la colonne « Std. Err. » ne contient pas l'écart-type de l'odds ratio, mais « autre chose », qui n'est là que par souci d'homogénéité de présentation du tableau de résultats et qui n'a pas d'interprétation. Les autres colonnes correspondent en revanche à leur intitulé (Tableau 2.3).
- On retrouve les mêmes résultats avec le tableau initial (Tableau 2.1) et la formule $OR = \frac{ad}{bc} = \frac{91 \times 1122}{26 \times 466} = 8,43$. Le fait qu'on trouve les mêmes résultats avec les données « brutes » du tableau de répartition des sujets (sans modélisation donc) vient de ce que, dans le cas d'une seule variable dichotomique (salp), le modèle logistique ne modélise en réalité rien. Il ne fait que décrire les données.

```
. logistic ct salp

Logistic regression                               Number of obs   =       1705
                                                  LR chi2(1)      =     108.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -1023.0534                       Pseudo R2       =     0.0503

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   8.427032   1.93065     9.30   0.000     5.378519    13.20343
       _cons |   .4153298   .0228891   -15.94   0.000     .3728059    .4627041
------------------------------------------------------------------------------
```

Tableau 2.3 : Modèle logistique pour analyser la relation entre ct et salp avec la commande logistic

IV. Tests des paramètres du modèle logistique

Le modèle logistique s'écrit logit $P = \alpha + \beta X$, où P est la fréquence de la maladie et X la variable qui caractérise l'exposition. Comme on l'a vu au chapitre 1, l'association de l'exposition avec la maladie est quantifiée par le coefficient (ou paramètre) β. Pour tester l'existence cette association, il faut donc tester l'hypothèse $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.

Je vais présenter différentes méthodes pour réaliser ce test (en réalité toutes équivalentes lorsque la taille de l'échantillon est assez grande).

Dans certains cas, on peut être amené à tester simultanément plusieurs paramètres β. Il peut s'agir de tester simultanément l'association de plusieurs expositions avec la maladie ou, plus fréquemment, des situations dans lesquelles l'exposition est représentée par deux variables X_1 et X_2, comme on le verra dans le chapitre 3 (§ III). J'indiquerai donc comment la démarche peut s'étendre au test de plusieurs paramètres simultanément.

IV.1. Test d'un seul paramètre

Lorsque β est estimé par la méthode du maximum de vraisemblance, il y a trois tests pour le comparer à une valeur donnée (0 dans le cas du modèle logistique). Ils sont présentés en Annexe 2. Je vais en montrer deux dans ce paragraphe : le test de Wald et le test du rapport des vraisemblances. Dans le cas particulier où la variable X est dichotomique, il est aussi possible de recourir au test « classique » du χ^2 à un degré de liberté.
Mais je vais tout d'abord dire quelques mots sur le lien entre test et intervalle de confiance.

IV.1.a. Test et intervalle de confiance

On peut tout d'abord noter que le fait que l'intervalle de confiance à 95 % de β ne contienne pas 0 est équivalent à rejeter H_0 au risque $\alpha = 5\,\%$. Une première façon de tester un paramètre du modèle logistique est donc de calculer son intervalle de confiance. Cette méthode présente l'inconvénient de ne pas donner le degré de signification, mais il ne faut pas oublier que la valeur du degré de signification résulte non seulement de la force de l'association mais aussi (et peut-être surtout) de la taille de l'échantillon (van Rijn MH et al., 2017). L'intervalle de confiance possède l'avantage, de plus en plus privilégié en épidémiologie, de quantifier la réponse en donnant β et un intervalle autour de β qui donnent à la fois la force de l'association et sa précision plutôt que d'apporter une réponse en oui/non (significatif ou pas), qui revient à réduire un résultat quantitatif en une réponse dichotomique après avoir fixé un seuil.

Ajoutons cependant que le recours à l'intervalle de confiance présente surtout un intérêt si β a une interprétation concrète. C'est le cas par exemple quand X est en 0/1, car e^{β} est alors égal à l'odds ratio entre l'exposition et la maladie.
Précisons aussi que le recours à l'intervalle de confiance ne se généralise pas au test simultané de plusieurs paramètres.

> Remarque
> Il arrive que les résultats du test et de l'intervalle de confiance ne soient pas tout à fait cohérents. Il peut arriver, par exemple, que le test soit significatif ($p < 5\,\%$) alors que l'intervalle de confiance contient la valeur 0. Cela vient de ce que les méthodes de calcul sont parfois différentes et que l'équivalence entre test et intervalle de confiance est vraie « asymptotiquement » (quand l'échantillon « tend » vers l'infini). Le plus souvent, les résultats sont en réalité très proches : p est « tout juste » supérieur à 5 % et 0 est « tout proche » de la limite de l'intervalle de confiance. Cela peut avoir une importance réelle pour publier ces résultats (malheureusement…), mais ne change en réalité pas grand-chose pour la conclusion qu'on peut en tirer.

IV.1.b. Test de Wald

Cette méthode repose sur les propriétés des estimateurs du maximum de vraisemblance, qui ont une distribution normale lorsque la taille de l'échantillon est assez grande, et sur le fait que la méthode du maximum de vraisemblance fournit aussi la variance des estimateurs.

Soit donc $\hat{\beta}$ l'estimation du maximum de vraisemblance de β et $s_{\hat{\beta}}$ son écart-type. On sait que si H_0 est vraie, la quantité $z = \frac{\hat{\beta} - 0}{s_{\hat{\beta}}}$ suit une loi normale centrée réduite. Le test de Wald consiste donc à calculer $z_0 = \frac{\hat{\beta}}{s_{\hat{\beta}}}$ à partir des observations et à voir si elle dépasse la valeur seuil de la loi N(0,1) pour le risque d'erreur fixé (1,96 pour $\alpha = 5\,\%$), et à en déduire le degré de signification.

Souvent, ce test est exprimé sous forme de χ^2 en élevant z_0 au carré. Il s'écrit alors, de façon équivalente : si H_0 est vrai, $\chi_0^2 = \frac{\hat{\beta}^2}{s_{\hat{\beta}}^2} = \frac{\hat{\beta}^2}{\text{var}(\hat{\beta})}$ suit une loi de χ^2 à un degré de liberté.

Cette méthode se généralise au test de plusieurs paramètres, comme on le verra plus loin (§ IV.2).

Les conditions d'application sont les mêmes que celles de l'utilisation de la méthode du maximum de vraisemblance pour estimer β. En général, elles ne sont pas données explicitement et reposent sur le nombre de variables incluses dans le modèle (voir § III.1). Ici, avec une seule variable, il faut que l'échantillon contienne au moins 10 malades et 10 non-malades.

IV.1.c. Test du rapport des vraisemblances

Comme la précédente, cette méthode repose sur les propriétés des estimateurs du maximum de vraisemblance.

Le principe est de calculer la vraisemblance V_0 lorsque H_0 est vraie (c'est-à-dire $\beta = 0$) et de la comparer à la vraisemblance V_1 observée sur l'échantillon, c'est-à-dire lorsque $\beta = \hat{\beta}$ où $\hat{\beta}$ est l'estimation du maximum de vraisemblance. On rejette l'hypothèse H_0 si V_0 est significativement différente de V_1.

Pour comprendre ce qui fonde ce test, considérons les modèles M_0 et M_1 obtenus respectivement lorsque $\beta = 0$ et lorsque $\beta \neq 0$. Ils s'écrivent :

M_0 : logit $P = \alpha$ (vraisemblance : V_0)

M_1 : logit $P = \alpha + \beta X$ (vraisemblance : V_1)

Si H_0 est vrai, les deux modèles sont identiques et les vraisemblances sont égales. Sinon, les vraisemblances sont différentes. La plus grande est alors V_1. On a en effet $V_1 \geq V_0$ car, pour obtenir la vraisemblance maximum V_1 du modèle M_1, on peut choisir entre toutes les valeurs possibles de α et de β. On obtient donc une vraisemblance supérieure ou égale à V_0, pour laquelle le choix est limité en imposant $\beta = 0$.

Le test de comparaison entre V_0 et V_1 passe par leur rapport. On montre que, si H_0 est vrai, $2\ln\left(\frac{V_1}{V_0}\right) = 2\ln(V_1) - 2\ln(V_0)$ suit une loi de χ^2 à un degré de liberté.

Comme on le verra plus loin (§ IV.2.b), le test du rapport des vraisemblances se généralise au test simultané de plusieurs paramètres.

IV.1.d. Cas particulier d'une variable X dichotomique, test du χ^2

Lorsque la variable X est dichotomique (l'exposition est en oui/non), les observations peuvent se présenter comme dans le tableau à quatre cases suivant :

	E^+	E^-	
M+	a	b	m_1
M-	c	d	m_0
	n_1	n_0	n

Les méthodes précédentes sont bien sûr toujours applicables, mais on peut aussi utiliser le test « classique » du χ^2 à un degré de liberté.

Dans ce cas particulier, l'hypothèse nulle peut s'écrire de trois façons équivalentes :

- H_0 : $P_0 = P_1$, où P_0 et P_1 sont les pourcentages d'exposés chez les non-malades et chez les malades ;
- H_0 : OR = 1, où OR est l'odds ratio quantifiant l'association entre l'exposition et la maladie ;
- H_0 : $\beta = 0$, où β est le coefficient de X dans le modèle logistique.

Le test lui-même s'écrit : $\chi^2_0 = \frac{(ad - bc)^2 n}{n_1 n_2 m_1 m_2}$.

Cette méthode ne se généralise pas au test simultané de plusieurs paramètres.

IV.1.e. Exemple

Considérons l'association entre la survenue d'une GEU et l'antécédent de salpingite dont j'ai parlé plus haut (Tableau 2.1).

L'antécédent de salpingite étant une variable dichotomique, les trois méthodes de test présentées ci-dessus sont donc applicables. Voici les résultats obtenus avec le logiciel Stata, accompagnés de quelques commentaires qui ne sont pas tous propres à ce dernier.

Test de Wald

Le test de Wald est donné en même temps que les résultats de l'estimation des paramètres et leur intervalle de confiance. La valeur de z_0 (appelé z dans la sortie du logiciel) est affichée, ainsi que le degré de signification (noté P > |z|) (voir Tableau 2.3).

On remarque que la sortie du logiciel donne aussi l'intervalle de confiance, ce qui souligne de nouveau l'équivalence entre les deux et montre leurs apports respectifs (voir § IV.1.a).
Il faut noter que la valeur de z_0 qui est donnée (9,30) est bien celle du test de Wald, c'est-à-dire $\frac{\hat{\beta}}{s_{\hat{\beta}}} = \frac{2,13}{0,229}$. C'est tout à fait clair quand on utilise la commande logit qui donne la valeur de β (voir Tableau 2.2). Cela l'est moins avec la commande logistic qui donne l'OR. Dans le Tableau 2.3, le z affiché n'est pas le rapport entre OR et Std. Err. Les logiciels sont parfois déroutants…

Comme cela a été indiqué précédemment, le test de Wald est parfois donné sous forme de χ^2. On obtient ici : $\chi_0^2 = z_0^2 = 9,30^2 = 86,49$.

> Remarque
> Le degré de signification affiché est 0,000. Il faut comprendre que c'est un arrondi et que cela signifie que p < 0,0001. Le degré de signification ne peut pas être nul.

Test du rapport des vraisemblances maximum

Il s'agit de comparer les vraisemblances maximum des modèles :
M_0 : logit P = α et M_1 : logit P = α + β salp
Le test s'exécute en trois temps : calcul des vraisemblances maximum des modèles M_1 et M_0, puis calcul de leur rapport avec la commande lrtest, comme il est indiqué dans le Tableau 2.4.

```
. logistic ct salp

Logistic regression                               Number of obs   =       1705
                                                  LR chi2(1)      =     108.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -1023.0534                       Pseudo R2       =     0.0503

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       salp |   8.427032    1.93065     9.30   0.000     5.378519    13.20343
      _cons |   .4153298   .0228891   -15.94   0.000     .3728059    .4627041
------------------------------------------------------------------------------

. est store a

. logistic ct if salp!=.

Logistic regression                               Number of obs   =       1705
                                                  LR chi2(0)      =      -0.00
                                                  Prob > chi2     =          .
Log likelihood = -1077.2309                       Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .4851916    .025054   -14.01   0.000     .4384899    .5368674
------------------------------------------------------------------------------

. est store b


. lrtest a b

Likelihood-ratio test                             LR chi2(1)  =     108.36
(Assumption: b nested in a)                       Prob > chi2 =     0.0000
```

Tableau 2.4 : Test du rapport des vraisemblances maximum

Remarques

- Les vraisemblances maximum qu'on compare doivent être calculées sur le même échantillon. C'est tout à fait indispensable, car la vraisemblance est le produit des probabilités que chaque sujet de l'échantillon soit observé. Sa valeur dépend donc en partie de la taille de l'échantillon.
- Dans l'exemple ci-dessus, la variable « salp » est inconnue pour 20 sujets. La vraisemblance du modèle M_1 est donc calculée sur 1705 sujets, alors que celle de M_0 pourrait être calculée sur 1725 sujets. Il faut donc contraindre la vraisemblance de M_0 à être calculée sur les 1705 sujets de M_1. C'est le sens de l'option « if salp! = ». qui veut dire « si "salp" n'est pas égale à "donnée manquante" ».
- La commande « est store » (*estimates store*) signifie qu'on met en mémoire les résultats du modèle qui vient d'être estimé. C'est une commande très générale ; les résultats mis en mémoire contiennent notamment la vraisemblance maximum, qui est ensuite utilisée par lrtest. Ici, les résultats de M_1 ont été mis dans la mémoire « a » et ceux de M_0 dans la mémoire « b ».
- Les modèles M_0 et M_1 qu'on compare doivent être emboîtés (c'est une notion sur laquelle je reviendrai dans l'annexe au chapitre 3). C'est ce que signale la parenthèse « Assumption: b nested in a », qui indique de plus que c'est à l'utilisateur de s'assurer de cet emboîtement.
- Dans les résultats du modèle 1, on trouve, en haut à droite, « LR chi2(1) = 108.36 ». Il s'agit en réalité du test des rapports de vraisemblance maximum du modèle estimé (ici M_1) et du modèle sans aucune variable X, qui est en fait ici M_0. On retrouve donc logiquement exactement le résultat du test comparant M_0 et M_1.

Test du χ^2 « habituel »

On l'obtient en ajoutant l'option chi à la commande donnant la répartition de sujets (Tableau 2.5).

```
. tab ct salp, chi

           |  salp clinique prouv
    0:acc  |        ou susp
    1:GEU  |         0          1 |     Total
-----------+----------------------+----------
         0 |     1,122         26 |     1,148
         1 |       466         91 |       557
-----------+----------------------+----------
     Total |     1,588        117 |     1,705

          Pearson chi2(1) = 116.2093   Pr = 0.000
```

Tableau 2.5 : Répartition des sujets et test du χ^2 « habituel »

IV.1.f. Comparaison des tests

Nous avons donc vu trois tests de la même hypothèse $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$. Leurs résultats s'expriment tous sous la forme d'un χ^2 à un degré de liberté parce qu'un seul paramètre est testé. Ils sont résumés dans le tableau ci-dessous.

χ^2 « habituel »	Wald	Rapport des vraisemblances
116,2	86,49 (9,30^2)	108,36

On constate que les valeurs numériques obtenues avec les trois tests sont différentes, même si elles conduisent toutes à la même conclusion : rejet de H_0 avec $p < 1‰$.

Le résultat théorique général est que ces trois tests sont asymptotiquement équivalents (Buse A, 1982), c'est-à-dire que leurs résultats sont les mêmes lorsque la taille de l'échantillon tend vers l'infini. Les différences sont dues ici à ce que les échantillons sont de taille finie.

Remarques

- Il peut donc se produire qu'un des tests soit significatif et un autre pas. Comme cela a été souligné dans le § IV.2.a à propos de l'intervalle de confiance, lorsque cela arrive, les résultats sont très proches de la limite de signification, mais de part et d'autre (un peu au-dessus ou un peu en dessous). La conclusion en termes d'association entre l'exposition et la maladie qu'on peut en tirer est en réalité la même.
- D'un point de vue pratique, il n'y a pas de raison de choisir l'un plutôt que l'autre. Le mieux est de garder une cohérence dans l'ensemble d'une analyse en utilisant le même test partout. Et bien sûr de ne pas surinterpréter des résultats proches de la signification statistique dans un sens ou dans l'autre.

IV.2. Test simultané de plusieurs paramètres

Il arrive qu'on soit intéressé à tester plusieurs paramètres (ou coefficients) du modèle logistique en même temps. Je donnerai un exemple plus bas (§ IV.2.c), mais ce sera surtout dans le chapitre 3 (§ III), avec les variables indicatrices, que ce type de test prendra tout son intérêt.

Considérons par exemple le modèle logistique logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, dans lequel on souhaite tester simultanément si les coefficients β_2 et β_3 sont égaux à 0.

Les hypothèses testées s'écrivent : $H_0 : \beta_2 = 0$ et $\beta_3 = 0$, et $H_1 : \beta_2 \neq 0$ ou $\beta_3 \neq 0$.

Cet exemple peut bien sûr se généraliser au test simultané de plus de deux coefficients.

Comme on va le voir, les tests de Wald et du rapport des vraisemblances s'étendent au test de plusieurs coefficients à la fois. Mais ce n'est pas le cas du test habituel de χ^2.

IV.2.a. Test de Wald

Partons de l'écriture du test de Wald sous la forme d'un χ^2. On a vu que, si H_0 est vraie, $\chi_0^2 = \dfrac{\hat{\beta}^2}{\text{var}(\hat{\beta})}$ suit une loi de χ^2 à un degré de liberté. Cette expression peut s'écrire

sous forme matricielle en considérant une matrice colonne contenant les coefficients qu'on veut tester : $\theta = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix}$ dans le cas de deux coefficients.

L'extension de $\frac{\hat{\beta}^2}{\text{var}(\hat{\beta})}$ est alors $(\hat{\theta})' \left[\text{var}(\hat{\theta})\right]^{-1} (\hat{\theta})$ où $(\hat{\theta})' = (\beta_2 \ \ \beta_3)$ est la transposée de $\hat{\theta}$ et var(θ) est la matrice de variance-covariance de θ.

Dans le cas de deux coefficients : $\text{var}(\theta) = \begin{pmatrix} \text{var}(\beta_2) & \text{cov}(\beta_2, \beta_3) \\ \text{cov}(\beta_2, \beta_3) & \text{var}(\beta_3) \end{pmatrix}$

On montre que, si $H_0 : \beta_2 = 0$ et $\beta_3 = 0$ et $H_1 : \beta_2 \neq 0$ ou $\beta_3 \neq 0$ est vraie, $\chi_0^2 = (\hat{\theta})' \left[\text{var}(\hat{\theta})\right]^{-1} (\hat{\theta})$ suit une loi de χ^2 à deux degrés de liberté.
De façon plus générale, si le test porte sur k coefficients, l'expression est la même et x_0^2 suit une loi de χ^2 à k degrés de liberté.

IV.2.b. Test du rapport des vraisemblances

Si on poursuit l'exemple précédent, il s'agit de comparer deux modèles :

- M_0 : logit $P = \alpha + \beta_1 X_1$, qui correspond à H_0 ;
- M_1 : logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, qui correspond à H_1.

Comme dans le cas du test d'un seul paramètre, la vraisemblance V_0 du modèle M_0 est inférieure ou égale à celle, V_1, du modèle M_1. Le test pour déterminer si ces vraisemblances sont différentes est le même que dans le cas d'un seul paramètre. On montre que, si H_0 est vraie, $2 \ln\left(\frac{V_1}{V_0}\right) = 2 \ln V_1 - 2 \ln V_0$ suit une loi de χ^2 à deux degrés le liberté.
De façon plus générale, si le test porte sur k coefficients, l'expression du test du rapport des vraisemblances est la même et suit une loi de χ^2 à k degrés de liberté.

> Remarque
> Le test du rapport des vraisemblances a été présenté ici pour tester simultanément plusieurs coefficients, c'est-à-dire pour comparer deux modèles dont l'un est obtenu en enlevant des variables à l'autre. C'est en réalité un cas particulier de la comparaison de deux modèles emboîtés, qui a une portée plus générale qui sera exposée dans le chapitre 3.

IV.2.c. Exemple

Pour illustrer les méthodes de test précédentes, considérons les trois variables dichotomiques suivantes (toujours dans l'enquête sur les grossesses extra-utérines) :

- ✓ salp : antécédent de salpingite (oui/non),
- ✓ univf : niveau d'études universitaires de la femme (oui/non),
- ✓ fprof : activité professionnelle de la femme (oui/non).

Le modèle logistique incluant ces trois variables est :
logit $P = \alpha + \beta_1$ salp $+ \beta_2$ univf $+ \beta_3$ fprof.

On s'intéresse ici au test simultané des paramètres β_2 et β_3. Comme les deux variables univf et fprof sont des indicateurs du niveau socio-professionnel de la femme, l'idée générale de ce test est de déterminer s'il y a un lien global entre ce niveau socio-professionnel et le risque de GEU. Si H_0 est vraie ($\beta_2 = 0$ et $\beta_3 = 0$), il n'y a pas de lien. Si H_1 est vraie ($\beta_2 \neq 0$ ou $\beta_3 \neq 0$), il y a un lien, sans qu'on puisse dire précisément s'il est « dû » à univf ou fprof.

> Remarque
> Ajoutons que, dans le cas présent, le test simultané de β_2 et β_3 n'a qu'un intérêt limité. Il est peu probable qu'on procèderait de cette façon si on s'intéressait au niveau socio-professionnel des femmes. Il n'est donné qu'à titre d'exemple. On verra plus loin des situations où il est plus utile, et même indispensable, de tester plusieurs paramètres simultanément (chapitre 3, § III, variables indicatrices).

L'estimation des paramètres du modèle logistique complet (avec les trois variables) est donnée dans le Tableau 2.6.

```
. logistic ct salp univf fprof

Logistic regression                             Number of obs     =      1,230
                                                LR chi2(3)        =      71.66
                                                Prob > chi2       =     0.0000
Log likelihood = -730.93713                     Pseudo R2         =     0.0467

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   7.392138    1.94466     7.60   0.000     4.414111    12.37932
       univf |   .8476072   .1323978    -1.06   0.290      .624073    1.151208
       fprof |   1.369012    .198271     2.17   0.030     1.030694    1.818381
       _cons |   .3314003   .0405161    -9.03   0.000     .2607878    .4211322
------------------------------------------------------------------------------
Note:  _cons estimates baseline odds.
```

Tableau 2.6 : Modèle logistique avec les variables salp unifv et fprof

> Remarque
> Même si ce n'est pas directement lié au principe des tests qu'on illustre ici, on peut remarquer que l'analyse avec les trois variables porte sur 1230 sujets (il est toujours utile de penser à regarder ce point). Cela veut dire qu'on a perdu environ un tiers des sujets par rapport à l'échantillon total, ce qui est considérable. En regardant plus en détail, on constaterait que c'est la variable fprof qui est inconnue pour 27 % des sujets, ce qui doit mettre en cause son utilisation pour les analyses ultérieures (ou du moins faire réfléchir à ce qu'on peut en faire).

Les résultats des tests de Wald et du rapport des vraisemblances sont donnés dans le Tableau 2.7.
On note que la syntaxe du test de Wald (avec la commande testparm) fait intervenir le nom des variables et non celui des coefficients. Cela est dû au fait que le nom des coefficients (et leur numérotation) est arbitraire, alors qu'il n'y a pas d'ambiguïté sur le nom des variables.
Cela souligne aussi l'importance qu'il y a à écrire explicitement le modèle pour éviter toute erreur sur l'expression du test. Avec le test de deux coefficients, comme ici,

on peut éventuellement s'en passer, mais cela peut devenir crucial dans des cas plus complexes.

```
. qui logistic ct salp univf fprof
. testparm univf fprof

 ( 1)  [ct]univf = 0
 ( 2)  [ct]fprof = 0
           chi2(  2) =    5.06
         Prob > chi2 =    0.0797

. est store a  // cette commande met en mémoire le modèle précédent avec les 3 variables salp, univf,fprof

. logistic ct salp if univf!=. & fprof!=.

Logistic regression                             Number of obs     =      1,230
                                                LR chi2(1)        =      66.53
                                                Prob > chi2       =     0.0000
Log likelihood = -733.50452                     Pseudo R2         =     0.0434

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   7.151568   1.872462     7.51   0.000       4.2809    11.94724
       _cons |   .3995128   .0260965   -14.05   0.000     .3515034    .4540795
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

. est store b  // cette commande met en mémoire le modèle précédent avec la seule variable salp

. lrtest a b

Likelihood-ratio test                                 LR chi2(2)  =      5.13
(Assumption: b nested in a)                           Prob > chi2 =    0.0767
```

Tableau 2.7 : Tests de Wald et du rapport des vraisemblances.

Les deux tests donnent ici des résultats très proches. La remarque faite plus haut dans le cas du test d'un seul paramètre reste cependant d'actualité : il peut y avoir des écarts, car les échantillons n'ont pas une taille infinie.
La conclusion est donc la même : on ne rejette pas H_0 ($p = 0{,}08$). On est cependant proches de la limite de significativité et il serait imprudent d'en tirer une conclusion péremptoire sur l'absence de lien entre le niveau socio-professionnel et le risque de GEU.
Notons enfin que l'estimation des paramètres du modèle complet indique que le coefficient de univf n'est pas significativement différent de 0 ($p = 0{,}29$) et que celui de fprof est significativement différent de 0 ($p = 0{,}03$). La règle générale, pour des raisons de contrôle des risques d'erreur de première espèce, est de n'interpréter les tests séparés des coefficients que lorsque le test global est significatif. On ne doit donc pas s'occuper ici des tests séparés. Si on voulait le faire, il faudrait le faire d'emblée et ne pas faire de test global.

V. Annexe 1 : Estimation des paramètres du modèle logistique avec une seule variable X, dichotomique

Les calculs qui suivent viennent du polycopié de Joseph Lellouch et al. que j'ai déjà mentionné (Lellouch J et al., 1988).

Dans le cas général (c'est-à-dire avec plusieurs variables X_i ou une seule variable non dichotomique), il n'y a pas de solutions explicites aux équations auxquelles on aboutit avec la vraisemblance pour estimer les paramètres d'un modèle logistique (voir § III.1).

Dans le cas d'une seule variable X dichotomique, le modèle logistique s'écrit $P(M^+ \mid X) = \frac{1}{1+e^{-(\alpha+\beta x)}}$ et sa vraisemblance est $V = \frac{\prod_{\text{non malades}} e^{-(\alpha+\beta x_i)}}{\prod_{\text{tous}} (1+e^{-(\alpha+\beta x_i)})}$. Il y a deux paramètres, α et β. Pour trouver leurs valeurs qui rendent V maximum, on passe par le logarithme de la vraisemblance L et on écrit que ses deux dérivées par rapport à α et β sont nulles.

Lorsque la variable X est dichotomique (l'exposition est en 0/1), les observations peuvent se présenter comme dans le tableau à quatre cases suivant :

	X = 1	X = 0	
M+	a	b	m_1
M-	c	d	m_0
	n_1	n_0	n

On obtient :

$$L = \sum_{\substack{\text{non-}\\ \text{malades}}} \ln(e^{-(\alpha+\beta x_i)}) - \sum_{\text{tous}} \ln(1+e^{-(\alpha+\beta x_i)}) = -m_0\alpha - \beta \sum_{\substack{\text{non-}\\ \text{malades}}} x_i - \sum_{\text{tous}} \ln(1+e^{-(\alpha+\beta x_i)})$$

En utilisant les effectifs du tableau ci-dessus, cela donne :

$$L = -m_0\alpha - \beta\, c - n_1 \ln(1+e^{-(\alpha+\beta)}) - n_0 \ln(1+e^{-\alpha})$$

Les deux équations correspondant au fait que les dérivées de L par rapport à α et β sont nulles sont les suivantes :

$$\frac{dL}{d\alpha} = -m_0 + n_1 \frac{e^{-(\alpha+\beta)}}{(1+e^{-(\alpha+\beta)})} + n_0 \frac{e^{-\alpha}}{(1+e^{-\alpha})} = 0$$

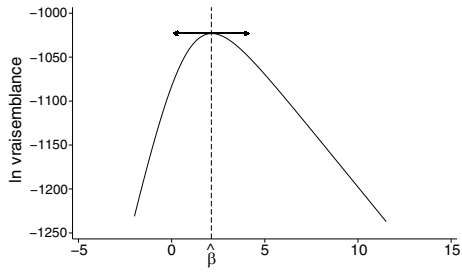
$$\frac{dL}{d\beta} = -c + n_1 \frac{e^{-(\alpha+\beta)}}{(1+e^{-(\alpha+\beta)})} = 0$$

La deuxième équation permet de trouver $e^{\alpha+\beta} = \frac{a}{c}$. En remplaçant cette valeur dans la première équation, on trouve $e^{\alpha} = \frac{b}{d}$. On en déduit les estimations de α et β : $\alpha = \ln\left(\frac{b}{d}\right)$ et $\beta = \ln\left(\frac{ad}{bc}\right) = \ln OR$.

VI. Annexe 2 : Les trois tests issus de la méthode du maximum de vraisemblance

Lorsqu'un paramètre est estimé par la méthode du maximum de vraisemblance, comme le coefficient β du modèle logistique, il existe trois tests, dérivés de cette méthode d'estimation, pour le comparer à une valeur fixée : le test de Wald, le test du rapport des vraisemblances et le *score test*. Ces trois tests sont asymptotiquement équivalents (Rayner JCW, 1997), c'est-à-dire qu'ils donnent le même résultat si la taille de l'échantillon tend vers l'infini.

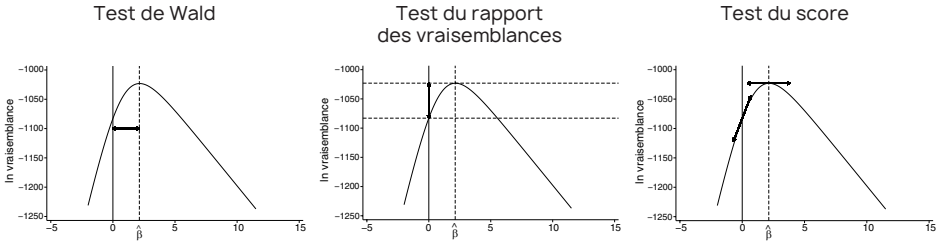
Sans entrer dans le détail de leur construction mathématique, nous allons décrire leur principe en partant de l'exemple de l'association entre le risque de GEU et l'antécédent de salpingite vu précédemment (voir le Tableau 2.1).



Le modèle correspondant s'écrit donc : logit P = $\alpha + \beta$ salp et la fonction vraisemblance a la forme de cloche ci-contre. L'estimation de β par la méthode du maximum de vraisemblance est $\hat{\beta} = 2{,}131$, valeur obtenue lorsque la vraisemblance est maximum ou, ce qui est équivalent, lorsque la dérivée est nulle, c'est-à-dire que la tangente à la courbe est horizontale.

Si on veut comparer β à une valeur fixe, 0 par exemple – mais le principe serait le même pour une autre valeur –, les trois tests cités plus haut pourraient être représentés graphiquement comme sur la figure suivante (Buse A, 1982).

- Le test de Wald consiste à calculer la « distance horizontale » entre 0 et $\hat{\beta}$ et à tester si elle est égale à 0.
- Le test du rapport des vraisemblances consiste à calculer la distance verticale entre la valeur maximum de la vraisemblance (obtenue pour $\hat{\beta}$) et la valeur pour $\beta = 0$ et à tester si elle est égale à 0.
- Le test du score consiste à comparer la pente de la tangente de la fonction de vraisemblance pour $\beta = 0$ à 0, qui est la valeur de la pente au maximum de la fonction (c'est-à-dire en $\hat{\beta}$).



Chapitre 3

Codage des variables et interprétation des coefficients

• • •

Comme je l'ai dit dans les chapitres précédents, le modèle logistique permet de prendre en compte tous les types de variables numériques, qualitatives ou quantitatives. Il y a cependant des règles à respecter sur la façon d'inclure les variables dans un modèle. De façon plus précise, la forme sous laquelle une variable doit être incluse dans un modèle ainsi que l'interprétation que l'on doit faire du ou des coefficients correspondant(s) dépendent, notamment, de la nature qualitative ou quantitative de la variable.

Cette question présente plusieurs volets. Le premier est que les résultats portant sur une variable ne doivent pas être interprétés isolément. J'en ai déjà parlé en indiquant que l'OR associé à une variable est un OR ajusté si d'autres variables sont présentes dans le modèle (chapitre 1). J'y reviendrai dans le § VI à propos de l'interaction. Le second, qui fait l'objet principal de ce chapitre, est l'interprétation du coefficient lui-même (ou des coefficients). Nous verrons qu'il faut distinguer les variables dichotomiques, qualitatives nominales ou ordinales, et quantitatives (auxquelles le chapitre 4 est consacré). Enfin, le choix de la liste des variables à inclure dans un modèle, question cruciale en épidémiologie, est discuté dans le chapitre 5.

I. Règle générale d'interprétation du coefficient d'une variable

Considérons d'abord un modèle ne comprenant qu'une seule variable. Même si, comme nous le verrons dans les paragraphes suivants, il y a d'autres façons de procéder, je m'intéresserai dans ce paragraphe au cas où X est incluse sous sa forme initiale : logit $P = \alpha + \beta X$.

De façon générale, avec ce modèle, l'odds ratio entre les catégories $X = x_1$ et $X = x_0$ (qui peuvent par exemple représenter les exposés et les non-exposés) est donné par :

$$\ln OR = \text{logit } P_1 - \text{logit } P_0 = (\alpha + \beta\, x_1) - (\alpha + \beta\, x_0) = \beta(x_1 - x_0).$$

On obtient donc : $OR = e^{\beta(x_1 - x_0)}$.

Ce résultat général pour calculer l'odds ratio permet d'interpréter les résultats de tous les modèles logistiques, même les plus complexes. Son utilisation demande d'écrire explicitement le modèle auquel on s'intéresse et me permet d'insister sur l'importance qu'il y a à le faire. Avec l'expérience, cela ne devient réellement nécessaire que pour les modèles les plus compliqués, mais j'invite le lecteur à se prêter à l'exercice au début, même pour les modèles simples, justement pour acquérir cette expérience.

Illustrons le calcul de l'odds ratio dans le cas où X est l'âge de la femme, ce qui permettra d'insister sur des précautions qu'il faut quand même prendre.

Le modèle logistique correspondant s'écrit : logit $P = \alpha + \beta$ age. L'estimation de ses coefficients est donnée dans le Tableau 3.1.

```
. logit ct age

.....

Logistic regression                             Number of obs     =      1,721
                                                LR chi2(1)        =      55.07
                                                Prob > chi2       =     0.0000
Log likelihood = -1068.1375                     Pseudo R2         =     0.0251

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0772177   .0105778     7.30   0.000     .0564855    .0979499
       _cons |  -3.003536   .3236634    -9.28   0.000    -3.637904   -2.369167
------------------------------------------------------------------------------
```

Tableau 3.1 : Modèle logistique pour analyser l'association entre la GEU (ct) et l'âge de la femme, variable quantitative

Le modèle s'écrit donc finalement : logit P = – 3,00 + 0,0772 age.

Si on veut estimer l'OR de GEU entre $x_0 = 29$ ans et $x_1 = 30$ ans, on obtient $OR = e^{0,0772 \times (30-29)} = 1,08$, ce qui peut s'interpréter comme un risque relatif en raison de la faible fréquence de la GEU, c'est-à-dire comme une multiplication du risque de GEU par 1,08 entre 29 et 30 ans.
Ce résultat mérite les deux commentaires suivants (au moins…) :

- On peut remarquer que l'augmentation du risque de GEU est la même entre 35 et 36 ans, ou d'ailleurs pour n'importe quelle variation d'une année. Mais il ne faut pas considérer qu'on a démontré quoi que ce soit quant à la linéarité de la relation entre l'âge et logit P. Ce résultat est contenu dans l'écriture du modèle sous la forme linéaire logit P = α + β age. Tout ce que l'on peut dire, c'est que, si la linéarité de la relation est vraie (ce que les calculs précédents ne montrent absolument pas), alors le facteur d'augmentation annuel du risque de GEU est $e^{0,077} = 1,08$.
  Cela souligne qu'il faut distinguer l'interprétation des coefficients d'un modèle et la véracité du modèle lui-même (au lieu de véracité, on parle aussi d'adéquation du modèle aux observations).
- Une augmentation de risque de 1,08 peut paraître modeste. Si on refait le même calcul entre 25 et 40 ans, on obtient : $OR = e^{0,0772 \times (40-25)} = 3,18$ ($= 1,08^{15}$), ce qui représente une augmentation du risque beaucoup plus importante. Et pourtant, ce sont les mêmes résultats et on dit en réalité la même chose, même si l'information peut être reçue de façon très différente par l'auditoire ou le lecteur. Sans qu'il s'agisse de manipulation, il est de la responsabilité de l'épidémiologiste de communiquer les résultats sous la forme qui lui paraît la plus adaptée selon ce qui lui semble le plus pertinent de transmettre (1,08 par an ou 3,18 au bout de 15 ans).

II. Variable dichotomique

Une variable dichotomique X a 2 catégories, généralement codées 0/1 (mais pas toujours). On a donc $x_0 = 0$ (non-exposés) et $x_1 = 1$ (exposés).
Avec le modèle logistique logit P = α + βX, on trouve donc pour l'odds ratio associé à l'exposition : $OR = e^{\beta}$. Le coefficient β est donc « identique » à l'odds ratio en ce sens que la connaissance de l'un est équivalente à la connaissance de l'autre.
Mais X peut aussi être codée 1/2, comme c'est souvent le cas pour le sexe ou parfois pour des variables recodées automatiquement par le logiciel. Dans ce cas, on a toujours $x_1 - x_0 = 1$ et donc, avec le modèle logit P = α + βX, l'odds ratio entre les deux catégories de X est encore $OR = e^{\beta}$. Il faut cependant noter que la valeur de la constante α est changée par rapport au codage 0/1.

La situation est différente si l'écart entre x_0 et x_1 n'est pas égal à 1. De façon générale, si la variable X est codée a/b, l'odds ratio entre les deux catégories de X n'est pas e^{β} mais $e^{\beta(b-a)}$. Le résultat est fondamentalement le même, il y a équivalence entre l'odds

ratio et le coefficient β du modèle logistique logit $P = \alpha + \beta X$, mais il fait faire attention en faisant les calculs et en lisant ce qu'affiche le logiciel, qui peut continuer à appeler « odds ratio » ce qui n'est que e^{β}.

Illustrons cela avec l'exemple de l'odds ratio associé à l'antécédent de salpingite (variable salp) déjà vu au chapitre précédent. Avec le modèle logistique logit $P = \alpha + \beta$ salp et la variable salp codée 0/1, le logiciel Stata donne : $\beta = 2{,}13$, d'où on déduit : $OR = e^{\beta} = 8{,}4$ (voir chapitre 2, § III.2).
Si l'antécédent de salpingite est codé 0/2 (variable salp0_2), le logiciel donne le résultat présenté dans le Tableau 3.2.

```
. logit ct salp0_2
...

Logistic regression                             Number of obs     =      1,705
                                                LR chi2(1)        =     108.36
                                                Prob > chi2       =     0.0000
Log likelihood = -1023.0534                     Pseudo R2         =     0.0503

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     salp0_2 |    1.065722    .114551     9.30   0.000     .8412065    1.290238
       _cons |   -.8786825   .0551107   -15.94   0.000    -.9866974   -.7706675
------------------------------------------------------------------------------

. logistic ct salp0_2

Logistic regression                             Number of obs     =      1,705
                                                LR chi2(1)        =     108.36
                                                Prob > chi2       =     0.0000
Log likelihood = -1023.0534                     Pseudo R2         =     0.0503

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     salp0_2 |   2.902935   .3325341     9.30   0.000     2.319163    3.633652
       _cons |   .4153298   .0228891   -15.94   0.000     .3728059    .4627041
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Tableau 3.2 : Modèle logistique avec la variable salp codée 0/2 (et notée salp0_2)

Le coefficient de la variable salp0_2 est bien la moitié du précédent (1,065 = 2,13/2) et, lorsqu'on utilise la commande logistic, la valeur qui figure dans la colonne « Odds Ratio » est bien la racine carrée du vrai odds ratio ($2{,}90 = \sqrt{8{,}4}$).

Normalement, on ne devrait pas être « pris au piège », car on est censé connaître le codage des variables lorsqu'on analyse un fichier, mais cela permet d'insister sur le fait que la façon dont une variable est incluse dans le modèle (y compris son codage) a de l'importance pour l'interprétation des résultats.

En pratique, mon conseil est de systématiquement coder 0/1 une variable dichotomique pour éviter les difficultés mentionnées ci-dessus, ainsi que d'autres qu'on verra plus loin (notamment lorsqu'il y a des termes d'interaction).

III. Variable qualitative nominale à plus de deux classes

Les catégories d'une variable qualitative nominale n'ont pas d'ordre. Cela s'oppose à une variable qualitative ordinale dont les catégories ont un ordre « naturel » qui fera l'objet du paragraphe suivant. Considérons par exemple une variable X ayant 4 catégories non ordonnées. Nous allons voir pourquoi il n'est pas correct de l'inclure dans un modèle logistique sous la forme logit $P = \alpha + \beta X$.
À titre d'illustration, je prendrai l'exemple de l'induction de la grossesse. Dans l'enquête sur les GEU, cette variable est notée gind avec 4 catégories, selon le traitement utilisé : « non induite », « induite par hCG », « induite par Clomid », « induite par une autre méthode ». Elles sont codées respectivement de 0 à 3 sans pour autant que cela indique un ordre[1].

Si on utilise le modèle logit $P = \alpha + \beta X$ où X a 4 catégories, il faut interpréter le coefficient β en suivant la règle générale indiquée au § I :

- L'odds ratio $OR_{0,1}$ comparant les catégories $X = 1$ et $X = 0$ est égal à e^{β}. En effet, on a : $\ln OR_{0,1} = \text{logit } P_1 - \text{logit } P_0$, où P_1 est la catégorie « exposés », c'est-à-dire $X = 1$, et P_0 la catégorie « non-exposés », c'est-à-dire $X = 0$. On obtient donc : $\ln OR_{0,1} = (\alpha + \beta) - \alpha = \beta$
- De même, l'odds ratio $OR_{0,2}$ comparant les catégories $X = 2$ et $X = 0$ est égal à $e^{2\beta}$, puisque $\ln OR_{0,2} = \text{logit } P_2 - \text{logit } P_0 = \ln OR_{0,1} = (\alpha + 2\beta) - \alpha = 2\beta$.
- Enfin, l'odds ratio $OR_{0,3}$ comparant les catégories $X = 3$ et $X = 0$ est égal à $e^{3\beta}$.

Avec l'exemple de l'induction de la grossesse, et en prenant la catégorie « non induite » pour référence, on obtient donc : $OR_{\text{« hCG »}} = e^{\beta}$; $OR_{\text{« Clomid »}} = e^{2\beta} = (OR_{\text{« hCG »}})^2$ et $OR_{\text{« Autre »}} = (OR_{\text{« hCG »}})^3$. C'est-à-dire que l'odds ratio associé à Clomid est le carré de celui associé à Hcg. Cela peut paraître un résultat remarquable, mais il faut noter que ces relations entre les odds ratios sont contenues dans l'écriture du modèle sous la forme logit $P = \alpha + \beta$ gind et restent les mêmes indépendamment des données recueillies. Si l'induction de la grossesse avait été enregistrée dans une autre variable gind2, contenant exactement la même information, mais codée en permutant les catégories 1 et 2 : 0 pour « non induite », 1 pour « induite par Clomid », 2 pour « induite par hCG » et 3 pour « induite par une autre méthode », le modèle logit $P = \alpha' + \beta'$ gind2 aurait donné $OR_{\text{« Clomid »}} = e^{\beta'}$; $OR_{\text{« hCG »}} = (OR_{\text{« Clomid »}})^2$ et $OR_{\text{« Autre »}} = (OR_{\text{« Clomid »}})^3$. Résultat tout aussi remarquable, mais très différent. Et tout aussi indépendant des données.

Cette modélisation de X ne doit donc pas être acceptée, car ses résultats sont liés à l'arbitraire du codage choisi pour numéroter les catégories de la variable X, indépendant des observations.

1. Le codage numérique des variables qualitatives est quasiment la règle dans les logiciels d'analyse, dont certaines commandes peuvent l'exiger.

La règle intangible est ainsi qu'une variable X qualitative nominale à plus de deux classes ne doit jamais être incluse sous sa forme initiale dans un modèle logistique, ni d'ailleurs dans aucun autre type de modèle de régression.

III.1. Décomposition d'une variable qualitative en variables indicatrices

Pour faire figurer X dans un modèle logistique sans se heurter au problème précédent, il faut tenir compte du fait que les catégories de X n'ont pas de lien (en particulier pas d'ordre) et qu'il est vain de vouloir les représenter avec une seule variable. La méthode consiste à remplacer la variable initiale X à k+1 classes en k variables X_i, chacune en 0/1 et définie comme indiqué dans le Tableau 3.3. L'inclusion de X dans le modèle est alors faite au travers des k variables X_i.

		X_1	X_2	…	X_k	
	0	0	0	…	0	
	1	1	0	…	0	
X =	2	0	1	…	0	logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$
	…	…	…	…	…	
	k	0	0	…	1	

Tableau 3.3 : Décomposition d'une variable qualitative X en variables indicatrices X_1, …, X_k

Les variables X_i sont appelées variables indicatrices, car chacune d'elles permet de repérer une des catégories de X. Le terme anglais *dummy variables* indique bien que ces variables n'ont pas de sens propre et qu'elles ne doivent être considérées dans leur ensemble que comme un moyen de faire figurer X dans le modèle (j'y reviendrai dans la suite).

On note qu'il suffit de k variables pour repérer les k+1 catégories initiales. C'est-à-dire qu'il est équivalent de connaître la valeur de X pour un sujet ou l'ensemble de ses k valeurs pour les variables indicatrices.
On peut même ajouter que, non seulement il est suffisant d'avoir k variables indicatrices, mais qu'il ne faut pas en ajouter une (k+1)-ième, X_0, qui serait la variable indicatrice de X = 0. En effet, on aurait alors le schéma et le modèle logistique suivants.

		X_0	X_1	X_2	…	X_k	
	0	1	0	0	…	0	
	1	0	1	0	…	0	
X =	2	0	0	1	…	0	logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$
	…	…	…	…	…	…	
	k	0	0	0	…	1	

On peut constater que le modèle logit $P = (\alpha + a) + (\beta_0 - a)X_0 + (\beta_1 - a)X_1 + (\beta_2 - a)X_2 + \ldots + (\beta_k - a)X_k$ donne les mêmes valeurs de logit P que le précédent, quelle que soit la valeur de X. Il y a donc une indétermination pour estimer les coefficients de ce modèle avec k variables. Il y a en réalité une variable en trop. Cela peut se comprendre en remarquant que, si on connaît les valeurs de X_1 à X_k pour un sujet, on connaît automatiquement la valeur de X_0, qui est donc une variable inutile. On dit parfois que les variables X_0 à X_k sont colinéaires. On a en effet : $X_0 = 1 - (X_1 + \ldots + X_k)$.

En pratique, pour utiliser des variables indicatrices dans les analyses, on peut les construire par programmes comme de nouvelles variables. Il est cependant plus facile d'utiliser les fonctions mises à disposition par les logiciels, et c'est surtout plus pertinent, car ces fonctions permettent aussi de « garder une trace » de ce que ce sont des variables indicatrices de X et de les traiter correctement dans les calculs et les tests ultérieurs (voir § III.4).
Avec Stata, les variables indicatrices sont générées en ajoutant i. devant le nom de la variable. Avec R, on a recours à la fonction « factor ».

III.2. Interprétation des coefficients d'une variable indicatrice

Dans le modèle logistique logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$, où les X_i sont les variables indicatrices de X, les coefficients β_i sont les odds ratios comparant les catégories $X = i$ et $X = 0$.

En effet, si on note $OR_{0,i}$ cet odds ratio, on a $\ln OR_{0,i} = \text{logit } P_1 - \text{logit } P_0$, où P_i correspond à la catégorie « exposés », qui est définie par $X_i = 1$ et $X_k = 0$ si $k \neq i$, et P_0 à la catégorie « non-exposés », définie par $X_k = 0$ quel que soit k. Ce qui donne bien : $\ln OR_{0,i} = (\alpha + \beta_i) - \alpha = \beta_i$.

Dans l'exemple de l'induction de la grossesse, c'est ce que l'on obtient avec la commande logistic, qui donne directement les OR (Tableau 3.4).

```
. logistic ct i.gind

Logistic regression                             Number of obs     =      1,719
                                                LR chi2(3)        =       9.20
                                                Prob > chi2       =     0.0267
Log likelihood = -1088.8724                     Pseudo R2         =     0.0042

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gind |
           1 |   1.601107   .6787768     1.11   0.267     .6975257    3.675197
           2 |   1.947153     .51349     2.53   0.012     1.161253    3.264925
           3 |   2.601799   1.750769     1.42   0.155     .6958158    9.728665
             |
       _cons |   .4804368   .0254398   -13.84   0.000     .4330757    .5329772
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Tableau 3.4 : Modèle logistique avec la variable gind décomposée en variables indicatrices

On peut noter que les OR obtenus avec le modèle logistique sont les mêmes que ceux qu'on obtient avec le tableau de données de base en prenant la catégorie « non induite » comme référence et en calculant les OR correspondant aux autres catégories (Tableau 3.5).
On a bien, par exemple, $OR_{hCG} = \frac{10 \times 1099}{13 \times 528} = 1{,}60$. Dans ce cas, le modèle logistique « ne modélise rien ». Il se contente de refléter les observations. La décomposition de X en variables indicatrices permet donc d'inclure X dans le modèle sans aucune hypothèse sur les liens éventuels entre les catégories de X.

```
. tab ct gind

    0:acc |       0:non ind 1:hcg 2:clomid 3:autr
    1:GEU |         0         1         2         3 |     Total
----------+--------------------------------------------+---------
        0 |     1,099        13        31         4 |     1,147
        1 |       528        10        29         5 |       572
----------+--------------------------------------------+---------
    Total |     1,627        23        60         9 |     1,719
```

Tableau 3.5 : Répartition des sujets selon les variables ct et gind

III.3. OR entre les catégories d'une variable indicatrice

On pourrait croire que la décomposition en variables indicatrices limite l'analyse parce qu'elle impose de fixer une catégorie de référence X = 0 et ne permet de calculer que les odds ratios des autres catégories par rapport à cette référence. En réalité, il n'en est rien, le choix de la catégorie de référence n'a aucune importance théorique. C'est-à-dire qu'en changeant de catégorie de référence (on verra comment faire au paragraphe suivant), on obtient un modèle équivalent : la vraisemblance est la même, on peut obtenir les mêmes OR (quitte à ce que ce soit un peu laborieux).

Pour s'en convaincre, voyons comment calculer l'OR entre deux catégories quelconques de X. Par exemple, si on veut l'odds ratio $OR_{2,3}$ entre les catégories 2 et 3 de X, il suffit d'utiliser la méthode générale : $\ln(OR_{2,3}) = \text{logit } P_1 - \text{logit } P_0$, où P_1 correspond à X = 3, c'est-à-dire $X_1 = 0$, $X_2 = 0$, $X_3 = 1$ et P_0 correspond à X = 2, c'est-à-dire $X_1 = 0$, $X_2 = 1$, $X_3 = 0$. On obtient ainsi : $\ln(OR_{2,3}) = (\alpha + \beta_3) - (\alpha + \beta_2) = \beta_3 - \beta_2$.
Pour la variable gind, l'OR entre les catégories 2 (hCG) et 3 (Autre) est ainsi égal à : $\ln(OR_{2,3}) = 0{,}96 - 0{,}67 = 0{,}29$, soit $OR_{2,3} = 1{,}34$.

Si on veut éviter ce calcul « manuel » (et, en plus, pouvoir déterminer l'intervalle de confiance de l'OR), il y a deux possibilités, qui seront présentées dans le paragraphe suivant :

- ✓ changer la catégorie de référence en l'indiquant dans la commande créant les variables indicatrices – c'est ce qu'il faut faire quand ce changement est à faire sur de nombreux calculs,
- ✓ calculer ponctuellement l'OR correspondant à $\beta_3 - \beta_2$ et son IC avec la commande lincom de Stata (qui est aussi présente dans R) (voir § III.3.b).

III.3.a. Changement et choix de la catégorie de référence

Le principe général est de considérer autant de variables indicatrices X_i qu'il y a de catégories de X. On a vu plus haut qu'il ne fallait pas les mettre toutes dans le modèle (dont les coefficients deviendraient indéfinis). L'exclusion de la variable X_0 qui a été faite précédemment revient en réalité à choisir X = 0 comme catégorie de référence. Si on veut que la catégorie de référence soit X = i, il suffit d'exclure X_i et de garder les autres variables indicatrices.

Prenons l'exemple de la variable gind et du logiciel Stata. La syntaxe i.gind créée en réalité quatre variables indicatrices pour les quatre classes 0, 1, 2 et 3 de gind. Elles n'apparaissent pas dans la liste des variables, mais sont accessibles avec les noms 0.gind, 1.gind, 2.gind et 3.gind. Par défaut, Stata exclut la variable 0.gind, car elle correspond à la catégorie avec la plus petite valeur. On peut imposer une autre catégorie de référence (ou de base) en remplaçant i. par ib(#), où # est la valeur de la catégorie de référence voulue.
Par exemple, pour prendre comme référence la catégorie codée 2, il faut écrire ib(2). gind (Tableau 3.6). La catégorie de référence est celle dont la valeur n'apparaît pas dans la première colonne de résultats (sous l'intitulé gind). Les OR de chaque ligne sont ceux qui les comparent à cette catégorie de référence. On a donc $OR_{2,3} = 1{,}34$.

Il s'agit bien du même modèle que celui du Tableau 3, avec 0 comme catégorie de référence. La vraisemblance est la même, ce qui n'est qu'une preuve indirecte. Mais surtout, on obtient les mêmes OR. On a retrouvé $OR_{2,3} = 1{,}34$, exactement comme à partir du modèle du Tableau 3.4.

```
. logistic ct ib(2).gind

Logistic regression                             Number of obs     =      1,719
                                                LR chi2(3)        =       9.20
                                                Prob > chi2       =     0.0267
Log likelihood = -1088.8724                     Pseudo R2         =     0.0042

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gind |
          0  |   .5135703   .1354353    -2.53   0.012     .3062858    .8611386
          1  |   .8222812   .4058971    -0.40   0.692     .3124982    2.163681
          3  |   1.336207   .9605281     0.40   0.687     .3265804    5.467103
             |
       _cons |   .9354839   .2416752    -0.26   0.796     .5638126    1.552165
------------------------------------------------------------------------------
Note:  _cons estimates baseline odds.
```

Tableau 3.6 : Modèle logistique avec la catégorie 2 de gind comme catégorie de référence

Finalement, le choix de la catégorie de référence n'est utile que pour la présentation des résultats et pour obtenir, avec plus ou moins de facilité, les OR qui paraissent pertinents ainsi que leur intervalle de confiance, et les interpréter. Mais c'est important !
Il serait ici par exemple peu pertinent de choisir la catégorie X = 3 (« Autre ») comme référence, comme cela est fait dans le Tableau 3.7.

```
. logistic ct ib(3).gind

Logistic regression                             Number of obs     =      1,719
                                                LR chi2(3)        =       9.20
                                                Prob > chi2       =     0.0267
Log likelihood = -1088.8724                     Pseudo R2         =     0.0042

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gind |
           0 |   .3843494   .2586314    -1.42   0.155      .102789    1.437162
           1 |   .6153846   .4872522    -0.61   0.540     .1303702    2.904792
           2 |   .7483871   .5379757    -0.40   0.687     .1829122    3.062033
             |
       _cons |        1.25   .8385255     0.33   0.739     .3356655    4.654932
------------------------------------------------------------------------------
Note:  _cons estimates baseline odds.
```

Tableau 3.7 : Modèle logistique avec la catégorie 3 de gind comme catégorie de référence

D'une part, on n'en voit pas la justification (et c'est une raison majeure pour ne pas le faire). D'autre part, l'effectif de cette catégorie est très petit (n = 9), ce qui conduirait à ce que tous les OR calculés par rapport à cette catégorie soient très imprécis (grand intervalle de confiance). Cela ne changerait rien au test global de l'association entre X et Y (voir § III.4), mais chaque OR serait « touché ». Enfin, tous les OR seraient inférieurs à 1. Ce n'est pas en soi un problème, mais l'interprétation d'un OR inférieur est toujours plus difficile. Il faut toujours faire un effort mental pour réaliser qu'un OR égal à 0,20 représente un lien aussi fort qu'un OR égal à 5.

III.3.b. Commande lincom

Considérons le modèle logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$, où la variable X à (k+1) classes a été décomposée en k variables indicatrices X_i, en prenant la catégorie X = 0 comme référence. Pour calculer l'OR entre les catégories X = 2 et X = 3, il faut calculer $\beta_3 - \beta_2$, c'est-à-dire une combinaison linéaire particulière entre les coefficients du modèle. Cette opération est en fait assez courante et Stata (comme R) dispose de la commande lincom (*linear combination*) pour le faire. « Comme d'habitude », dans la syntaxe de cette commande, ce ne sont pas les noms des coefficients inconnus du logiciel qui doivent apparaître, mais ceux des variables associées aux coefficients dans le modèle. On obtient ainsi l'OR cherché et son intervalle de confiance (Tableau 3.8).

```
. quiet logistic ct i.gind

. lincom 3.gind-2.gind, or

 ( 1)  - [ct]2.gind + [ct]3.gind = 0

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    1.336207   .9605281     0.40   0.687     .3265804    5.467103
------------------------------------------------------------------------------
```

Tableau 3.8 : Résultats de la commande lincom

III.4. Test de l'association entre Y et X décomposée en variables indicatrices

III.4.a. Les variables indicatrices sont indissociables

Lorsqu'on décompose X en variables indicatrices, celles-ci doivent être toutes présentes ou absentes *en même temps* dans le modèle.
Considérons, par exemple, le modèle avec la variable X à 4 classes décomposée en variables indicatrices, donc avec le modèle logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$. Si on constate, par exemple, que le coefficient β_3 est non significativement différent de 0, cela ne doit pas conduire à retirer du modèle la variable indicatrice correspondante.

La raison en est la suivante. Le modèle complet est logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, où les variables sont définies selon le Tableau 3.3. Si on retire la variable X_3, la partie grisée disparaît comme dans le tableau ci-dessous et le modèle devient $\text{logit } P = \alpha' + \beta'_1 X_1 + \beta'_2 X_2$.

		X_1	X_2	X_3
	0	0	0	0
	1	1	0	0
X =	2	0	1	0
	3	0	0	1

Or, en l'absence de X_3, les catégories 0 et 3 de X ne sont plus distinguables avec les variables restantes X_1 et X_2 : les sujets de ces deux catégories ont les mêmes valeurs pour X_1 et X_2. Le coefficient β'_1 de la variable X_1 n'a plus le sens du coefficient β_1 du modèle complet. e^{β_1} était l'odds ratio entre les catégories 0 et 1 de X, alors que $e^{\beta'_1}$ devient l'odds ratio entre la catégorie X = 1 et les catégories X = 0 et X = 3 réunies.
Retirer X_3 revient donc à regrouper les catégories X = 0 et X = 3. Ce n'est bien sûr pas interdit, mais cela doit se faire en connaissance de cause, sans se fonder uniquement sur des résultats statistiques et lorsque les résultats restent interprétables avec la catégorie réunie. Dans le cas de la variable gind, est-ce que cela aurait, a priori, un sens d'avoir une catégorie de référence composée des femmes dont la grossesse n'a pas été induite et des femmes dont elle a été induite par un autre traitement que Clomid ou hCG ?
La non-significativité du coefficient d'une variable indicatrice n'est pas une raison suffisante pour retirer cette variable du modèle. Une raison clinique ou épidémiologique est nécessaire. Il faut notamment être vigilant sur ce point lors de l'utilisation de méthodes automatiques de sélection des variables (méthodes pas-à-pas par exemple).

III.4.b. Tests de Wald et du rapport des vraisemblances

De façon cohérente avec le paragraphe précédent, le test de l'association entre Y et la variable X décomposée en variables indicatrices doit être un test global. Les variables indicatrices X_i sont là pour représenter X sous une forme adéquate et ne doivent pas être testées individuellement.

Les hypothèses à tester s'écrivent :

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

$H_1 : \beta_1 \neq 0$ ou $\beta_2 \neq 0$ ou… ou $\beta_k \neq 0$.

D'un point de vue pratique, on utilise le test de Wald ou le test du rapport des vraisemblances, qui sont tous les deux des χ^2 à k degrés de liberté (il y a k coefficients à tester simultanément), comme le montre l'exemple ci-dessous avec la variable gind (Tableau 3.9). Il n'y a pas lieu de réaliser des tests séparés des coefficients β_i (ou des OR correspondants). Cela ne correspond pas à la question initiale, qui est celle de l'association entre X et la maladie Y. De plus, cela conduirait à des tests multiples et à un mauvais contrôle du risque de première espèce.

Dans le cas de la variable gind, le modèle logistique est celui du Tableau 3.4 ; les résultats des tests de Wald et du rapport des vraisemblances sont donnés respectivement dans les Tableaux 3.9 et 3.10. Comme cela est fréquent (voir chapitre 2, § IV.1.f), ils aboutissent à la même conclusion : un rejet de l'hypothèse d'une absence de lien entre l'induction de la grossesse et le risque de GEU.

```
. quiet logistic ct i.gind
. testparm i.gind

 ( 1)  [ct]1.gind = 0
 ( 2)  [ct]2.gind = 0
 ( 3)  [ct]3.gind = 0

           chi2(  3) =    9.36
         Prob > chi2 =    0.0248
```

Tableau 3.9 : Test de Wald pour la variable gind du modèle logistique du Tableau 3.4

```
. quiet logistic ct i.gind
. est store a
. quiet logistic ct if gind!=.
. est store b
. lrtest a b

Likelihood-ratio test                           LR chi2(3)  =      9.20
(Assumption: b nested in a)                     Prob > chi2 =    0.0267
```

Tableau 3.10 : Test du rapport des vraisemblances pour la variable gind du modèle logistique du Tableau 3.4

III.4.c. Présentation des résultats

La présentation des résultats concernant la variable X (gind ici) doit tenir compte des considérations précédentes et avoir la forme suivante, avec un seul test (Tableau 3.11). On ne devrait pas voir de test sur chacune des lignes hCG, Clomid et Autre.

Cette règle générale ne doit cependant pas empêcher de commenter le résultat global, comme on le fait souvent. Ici, on montre qu'il y a un lien global entre le risque de

GEU et l'induction de la grossesse (avec p = 0,03). On peut noter que ce lien s'explique principalement par l'augmentation du risque chez les femmes dont la grossesse a été induite par du Clomid, mais cela reste une observation, pas un test au sens propre. De plus, on s'abstiendrait de ce commentaire si le test global était non significatif, quel que soit le résultat ligne par ligne.

Induction de la grossesse	n	OR et IC	p
non	1627	1	0,03
hCG	23	1,6 [0,7 ; 3,7]	
Clomid	60	1,9 [1,2 ; 3,3]	
Autre	9	2,6 [0,7 ; 9,7]	

Tableau 3.11 : Présentation dans un tableau des résultats de l'association entre gind et le risque de GEU (résultats du modèle logistique du Tableau 3.4 et test du Tableau 3.9)

IV. Variable qualitative ordinale

Les catégories d'une variable qualitative ordinale suivent un ordre « naturel », c'est-à-dire un ordre que personne ne conteste raisonnablement. Par exemple, une exposition peut être notée comme nulle, modérée ou forte. Tout le monde s'accorde pour dire que la catégorie modérée se situe entre les deux autres catégories, même si les distances entre les catégories ne sont pas toujours faciles, voire sont impossibles, à fixer et peuvent en tout cas être discutées.
Souvent, il s'agit d'une variable sous-jacente quantitative mise en classe, soit dans le questionnaire lui-même (on n'a pas alors la valeur quantitative de base), soit au moment de l'analyse. Dans ce second cas, des alternatives aux classes existent, comme on le verra dans le chapitre 4.

Je prendrai l'exemple dans ce paragraphe de la consommation de tabac de la femme, comptée en cigarettes par jour et mise en quatre classes : variable tabfc, dont la distribution est indiquée dans le Tableau 3.12.

Comme dans le paragraphe précédent, si on veut rendre compte des observations sans faire d'hypothèse sur les relations entre les catégories de fumeuses, il faut trois odds ratios. En prenant les non-fumeuses comme catégorie de référence, on obtient : $OR_{1-9} = \frac{78 \times 809}{158 \times 264} = 1,51$, $OR_{10-19} = 3,27$ et $OR_{\geq 20} = 4,06$. Ces OR peuvent être donnés par un modèle logistique où la variable tabfc est décomposée en variables indicatrices, comme si c'était une variable qualitative nominale.

Elle présente l'inconvénient de ne pas tenir compte de l'ordre des catégories de X et donc de ne pas exploiter toute l'information disponible, ce qui peut se traduire par une perte de puissance s'il y a une relation de type dose-effet.

```
. tab ct tabfc

    0:acc |          0:nf;1:1-9;2:10-19;3:>=20
    1:GEU |         0          1          2          3 |     Total
----------+--------------------------------------------+----------
        0 |       809        158        106         71 |     1,144
        1 |       264         78        113         94 |       549
----------+--------------------------------------------+----------
    Total |     1,073        236        219        165 |     1,693
```

Tableau 3.12 : Répartition de la consommation de cigarettes par jour selon que la femme a une GEU ou pas

On peut donc envisager de « modéliser » l'association entre Y (survenue d'une GEU) et X (consommation de tabac) pour représenter la relation sous forme dose-effet (de façon similaire à une régression linéaire). Cela nécessite alors de faire (et si possible de vérifier) l'hypothèse de la linéarité de la relation entre X et Y.

IV.1. Modélisation de l'association GEU-tabac

Avant de modéliser l'association GEU–tabac, on peut commencer par la représenter graphiquement. Cette représentation est possible parce que X est ordonné et peut donc constituer l'abscisse du graphique. Ce ne serait pas le cas avec une variable nominale. L'échelle des ordonnées doit être le logit, puisque la modélisation se fera avec le modèle logistique. On calcule donc logit P pour les quatre catégories de consommation de tabac.

Pour la catégorie X = 0 (non-fumeuses), on a $\text{logit } P_0 = \ln\left(\frac{P_o}{1-P_o}\right) = \ln\left(\frac{264/1073}{809/1073}\right) = -1{,}12$

et de même logit $P_1 = -0{,}71$; logit $P_2 = 0{,}064$ et logit $P_3 = 0{,}28$. Ce qui donne les points de la Figure 3.1.

On peut raisonnablement envisager de représenter la variation de logit P en fonction de la quantité de cigarettes par une droite. Le modèle correspondant est : logit $P = \alpha + \beta X$, et il faudra bien sûr tester s'il est adéquat (voir § IV.3). On parle de modèle linéaire bien que ce soit la relation entre X et logit P qui est linéaire et pas celle entre X et le risque P de GEU. Après estimation des coefficients sur les données, le modèle s'écrit : logit $P = -1{,}12 + 0{,}51\ X$, ce qui permet de compléter la Figure 3.1 par la droite qui représente le mieux les points observés (parmi toutes les droites possibles).

Le coefficient β est la pente de la droite (0,506) ; il est significativement différent de 0 (avec p < 1‰). Son exponentielle ($e^{0{,}506} = 1{,}66$) est l'odds ratio comparant deux catégories successives de X. Il est le même pour toutes les catégories, ce qui correspond à l'hypothèse de linéarité, et c'est en cela qu'il y a modélisation de l'association entre X et Y. Pour obtenir les odds ratios donnés par le modèle entre les différentes catégories de fumeuses (X = 1, 2 ou 3) et les non-fumeuses (X = 0), il suffit de calculer e^{β}, $e^{2\beta}$ et $e^{3\beta}$, par exemple avec la commande lincom, qui donne aussi directement leurs intervalles de confiance (Tableau 3.13).

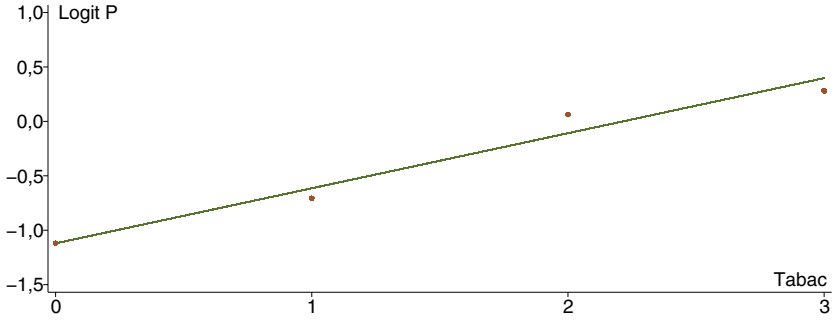


Figure 3.1 : Représentation graphique de la relation entre la consommation de tabac en 4 classes, codées 0, 1, 2, 3 et 4, et le logit du risque de GEU : points observés et modélisation linéaire.

Les odds ratios donnés par le modèle sont différents des odds ratios observés. Les différences sont dues aux éventuels écarts à l'hypothèse de linéarité, et aux (inévitables) fluctuations d'échantillonnage des OR observés. Il faudra tester si ces différences sont significatives, mais elles paraissent ici assez peu importantes.

Nombre de cigarettes par jour	0	1–9	10–19	≥ 20
X	0	1	2	3
OR observé et intervalle de confiance	1	1,51 [1,12–2,05]	3,27 [2,42–4,41]	4,07 [2,89–5,69]
OR donné par le modèle logit $P = \alpha + \beta X$ et intervalle de confiance	1	1,66 [1,50–1,83]	2,75 [2,27–3,35]	4,56 [3,39–6,12]

Tableau 3.13 : Odds ratios de GEU pour les trois catégories de fumeuses codées 1, 2, 3

L'avantage de la modélisation (si l'hypothèse de linéarité est acceptable) est qu'on résume l'association GEU–tabac par un seul paramètre : $\beta = 0{,}506$ (ou OR = 1,66), au lieu des trois paramètres du modèle avec des variables indicatrices. Cette forme de résumé est plus informative pour l'interprétation des résultats qu'une description brute des valeurs observées et de leurs fluctuations d'échantillonnage. La modélisation conserve la possibilité de calculer les OR pour chacune des catégories de X, ce qui est très important pour la présentation des résultats, et ceux-ci sont plus précis (les intervalles de confiance sont plus petits). Cette amélioration de la précision, qui est une règle générale avec la modélisation, résulte de l'ajout d'une hypothèse supplémentaire (ici la linéarité) et du fait que l'ensemble des observations sont utilisées pour l'estimation de chaque OR et de son intervalle de confiance.

Enfin, le test de l'association entre X et Y avec la modélisation est celui de l'unique paramètre β (c'est un test de tendance) ; il est plus puissant que le test simultané des k paramètres β_i avec des variables indicatrices.

IV.2. Choix des valeurs de X

Dans le paragraphe précédent, les catégories de X étaient codées 0, 1, 2 et 3. Cela respecte l'ordre des catégories de X, ce qui est essentiel, et cela pouvait sembler être une numérotation « neutre » de ces catégories. En réalité, cela n'est pas neutre. Ce codage contient l'hypothèse que l'écart entre deux catégories adjacentes est le même. Ce n'est pas forcément le cas et cela a une influence sur la modélisation de l'association entre X et Y.

Pour s'en rendre compte, on peut par exemple, de façon alternative, coder X en prenant le centre des classes : 0, 5, 15 et 25. Cela a le mérite d'être plus proche de la variable quantitative sous-jacente. Les odds ratios observés sont bien sûr inchangés, mais la nouvelle variable X' est différente et les odds ratios estimés par le modèle logit $P = \alpha' + \beta' X'$ sont différents (Tableau 3.14). Attention, ils doivent être calculés en utilisant $e^{5\beta'}$, $e^{15\beta'}$ et $e^{25\beta'}$.

La représentation graphique de la relation entre X et Y est aussi modifiée. Assez peu, semble-t-il, sur la Figure 3.2, mais il faut faire attention à ce que l'impression visuelle est très dépendante de la façon de représenter l'échelle des abscisses. On remarque, par exemple, que les points observés paraissent atteindre un « plafond » au-delà de 15 cigarettes par jour et seraient peut-être mieux représentés par une courbe de type logarithme que par une droite.

Nombre de cigarettes par jour	0	1–9	10–19	≥ 20
X'	0	5	15	25
OR observé et intervalle de confiance	1	1,51 [1,12–2,05]	3,27 [2,42–4,41]	4,07 [2,89–5,69]
OR donné par le modèle logit $P = \alpha' + \beta' X'$ et intervalle de confiance	1	1,37 [1,28–1,45]	2,54 [2,12–3,06]	4,74 [3,50–6,43]

Tableau 3.14 : Odds ratios de GEU pour les trois catégories de fumeuses codées 5, 15 et 25

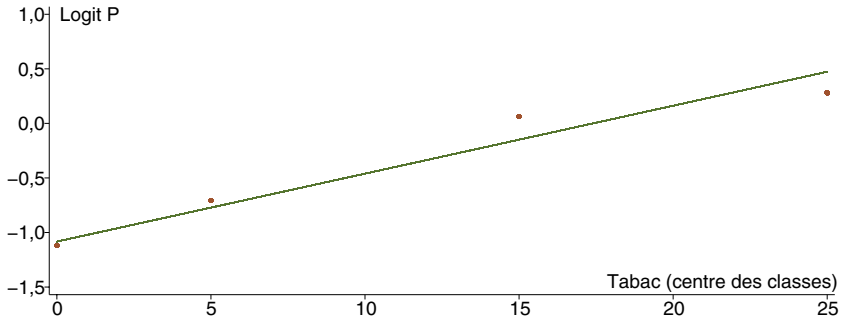


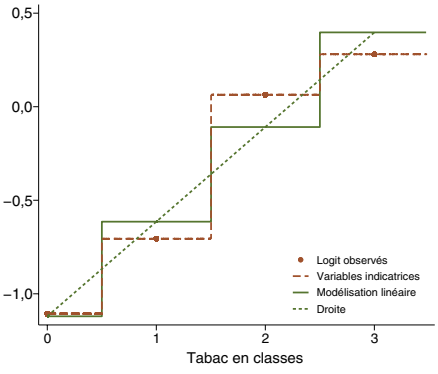
Figure 3.2 : Représentation graphique de la relation entre la consommation de tabac, codée en cigarettes par jour, et le logit du risque de GEU : points observés et modélisation linéaire.

Il n'y a pas de test pour comparer les deux codages, mais on peut, pour chacun d'eux, tester s'il s'écarte de la linéarité (ici le test est non significatif dans les deux cas). Le choix du codage de X repose en partie sur l'adéquation du modèle aux observations, mais aussi (et je dirais, surtout) sur des considérations épidémiologiques et/ou de mécanisme d'action de l'exposition X sur la maladie.
Ajoutons que ce choix a de l'importance pour quantifier la relation entre Y et X et estimer les odds ratios, bien que souvent de façon relativement modeste. Cependant, en pratique, le test de tendance de l'association entre X et Y n'est que peu affecté.

IV.3. Choix entre variables indicatrices et modélisation linéaire (test de linéarité)

Avec une variable qualitative ordinale X, on a donc le choix de la décomposer en variables indicatrices comme on le fait pour une variable nominale ou de l'inclure dans le modèle logistique sous sa forme originelle (modélisation linéaire) au prix d'une hypothèse sur la linéarité de la relation entre X et logit P. Si cette hypothèse est satisfaite, le modèle linéaire est le meilleur choix, ainsi qu'on vient de le voir : meilleure précision, meilleure puissance.
C'est le test de cette hypothèse, qui est donc un test de linéarité[2], qu'on va exposer ici. Je vais poursuivre avec l'exemple de la variable « tabac » présenté au début de ce § IV.

Il faut tout d'abord préciser qu'il est abusif de représenter la relation tabac–GEU par une courbe continue, car la variable « tabac » est en classes et ne peut donc prendre que quatre valeurs. Il est préférable de représenter la relation par une fonction en escalier, contrairement à ce qui a été fait dans les figures précédentes, qui peuvent donc, à ce titre, être trompeuses[3].



Modèle avec variables indicatrices :
$\text{logit } P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
Modèle « linéaire » : $\text{logit } P = \alpha + \beta X$ (avec X = 0, 1, 2 ou 3)
Droite : pas de modèle correspondant, car X ne prend que 4 valeurs

Figure 3.3 : Trois modélisations d'une variable qualitative ordinale.

2. Ce test correspond au fait que X est une variable qualitative ordonnée. D'autres tests de linéarité sont utilisés lorsque X est une variable quantitative (voir chapitre 4).

3. Une fonction en escalier est souvent présentée sous forme continue, mais les parties verticales ne font en réalité pas partie de la courbe et ne sont là que par habitude.

C'est ce qu'indique la Figure 3.3. Le terme « modélisation linéaire », s'il correspond bien à l'écriture formelle du modèle logit $P = \alpha + \beta X$, ne doit pas être pris au pied de la lettre. Il correspond en réalité à la fonction en escalier en trait continu, escalier avec des marches de hauteur constante de façon à ce qu'elle corresponde à la droite en petits pointillés.

Lorsqu'on veut tester l'hypothèse de linéarité, il faut comparer le modèle linéaire, celui avec fonction en escalier régulier (trait continu), et le modèle avec des variables indicatrices (trait en pointillés larges), qui est aussi avec une fonction en escalier, mais avec des marches de hauteurs différentes, déterminées pour que l'escalier passe par tous les points observés. Ces deux modèles ont les caractéristiques suivantes :

- Modèle linéaire
  - Il utilise l'information concernant l'ordre des catégories de X.
  - Il représente la relation de façon linéaire, ce qui paraît un qualificatif justifié bien qu'il s'agisse en réalité d'un escalier régulier, comme cela a été expliqué plus haut.
  - Il résume et synthétise la relation par un seul coefficient, la pente de la droite (qui est aussi la hauteur de chaque marche), dont le test (comparaison à 0) constitue un test de tendance pour la relation entre X et la maladie.
- Modèle avec variables indicatrices
  - L'information concernant l'ordre des catégories de X n'intervient pas dans l'estimation de ses coefficients.
  - il est le plus proche possible des observations (le plus adéquat aux données), puisqu'il passe par tous les points observés.
  - il représente la relation par un escalier (ce qui a peu de chances d'être conforme à la réalité, mais le fait de répartir la consommation de tabac en classes n'est pas non plus très réaliste…).
  - il est sensible aux fluctuations d'échantillonnage : les hauteurs des « marches » peuvent être très variables selon l'échantillon, surtout si une classe a un effectif petit.

Les deux modèles s'écrivent :

- Modèle « linéaire » : $\text{logit } P = \alpha + \beta X$ (avec $X = 0, 1, 2$ ou 3),
- Modèle avec variables indicatrices : $\text{logit } P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

Le modèle avec des variables indicatrices est plus général. Il ne se confond avec le modèle linéaire que si ses marches sont de hauteurs égales, c'est-à-dire si la hauteur totale où arrive la deuxième marche est le double de la hauteur de la première marche ($\beta_2 = 2\,\beta_1$), et ainsi de suite pour les marches suivantes. On montre (voir annexe) que les deux modèles sont emboîtés. Ils peuvent donc être comparés au moyen d'un test statistique, qui est donc un test de linéarité, et qui peut être réalisé par un test du rapport des vraisemblances ou par un test de Wald portant sur plusieurs paramètres, dont les hypothèses sont : H_0 : $\beta_2 = 2\beta_1$ et $\beta_3 = 3\beta_1$ et H_1 : $\beta_2 \neq 2\beta_1$ ou $\beta_3 \neq 3\beta_1$.

En poursuivant avec l'exemple de la variable tabfc, les deux tests donnent les résultats présentés dans le Tableau 3.15.

Comme c'est le cas de façon générale (voir chapitre 2, § IV.1.f), on constate que leurs résultats sont équivalents, même s'ils ne sont pas toujours identiques. Ici, la différence entre les modèles est non significative. Il n'y a donc pas d'écart significatif à la linéarité, c'est-à-dire qu'il n'y a pas de raison statistique de préférer l'un ou l'autre. En pratique, on retient la modélisation linéaire qui est plus simple et plus synthétique, comme expliqué plus haut.

```
Test du rapport des vraisemblances

. qui logit ct tabfc
. est store lin

. qui logit ct i.tabfc
. est store indic

. lrtest indic lin

Likelihood-ratio test                                 LR chi2(2)  =      2.62
(Assumption: lin nested in indic)                     Prob > chi2 =    0.2702


Test de Wald

. qui logit ct i.tabfc
. test (2.tabfc=2*1.tabfc) (3.tabfc=3*1.tabfc)

 ( 1)  - 2*[ct]1.tabfc + [ct]2.tabfc = 0
 ( 2)  - 3*[ct]1.tabfc + [ct]3.tabfc = 0

           chi2(  2) =    2.62
         Prob > chi2 =    0.2705
```

Tableau 3.15 : Test de linéarité avec le rapport des vraisemblances et avec le test de Wald

Remarque

Il faut faire attention à l'interprétation de ce test de linéarité. Tel que je l'ai exprimé, on pourrait croire que « non significatif » veut dire pas de différence (entre les deux modèles) et donc qu'on a prouvé la linéarité. Cela serait contradictoire avec la conclusion habituelle d'un test où « non significatif » ne veut pas dire qu'il n'y a pas de différence, mais qu'on n'en a pas mis en évidence, peut-être à cause d'un manque de puissance.
Ici, il s'agit d'une question de choix entre deux modèles. Un test non significatif veut dire que les statistiques ne permettent pas de les distinguer. Il faut donc un autre critère de choix. On a choisi ici le modèle le plus simple.

V. Variable quantitative

Lorsque la variable X est quantitative, l'écriture « naturelle » du modèle logistique, logit $P = \alpha + \beta X$, suppose que la relation entre X et logit P est linéaire. Comme ce n'est pas toujours le cas, il faut donc envisager d'autres formes d'association et choisir celle qui est la plus pertinente. Les méthodes correspondantes sont discutées en détail dans le chapitre 4.

VI. Prise en compte d'une interaction

Pour que la question de l'interaction se pose, il faut qu'il y ait trois variables : la maladie M et deux facteurs, l'exposition E et un autre facteur F. Rappelons (voir chapitre 1, § V.2) qu'il y a interaction entre deux facteurs E et F lorsque l'association entre E et M n'est pas la même selon les niveaux de F. Rappelons aussi que l'interaction, définie de cette façon, n'a pas de support biologique (ou seulement partiellement) et que son existence peut être liée à la mesure d'association choisie, à savoir l'odds ratio pour le modèle logistique.
Nous allons voir ici comment prendre en compte une interaction dans l'écriture du modèle logistique, comment interpréter les résultats et estimer les odds ratios correspondants.

La première chose importante à noter est que, dans un modèle logistique où les variables E et F sont incluses sous la forme logit $P = \alpha + \beta E + \gamma F$, il n'y a pas, par construction, d'interaction entre ces deux variables. En effet, en prenant à titre d'exemple E et F dichotomiques (codées en 0/1), les odds ratios associés à E selon que $F = 0$ ou $F = 1$ sont les suivants :

- pour $F = 0$: $\ln OR_0 = \text{logit } P_1 - \text{logit } P_0 = \alpha + \beta - \alpha = \beta$
- pour $F = 1$: $\ln OR_1 = \text{logit } P_1 - \text{logit } P_0 = \alpha + \beta + \gamma - (\alpha + \gamma) = \beta$

On a donc $OR_0 = OR_1$ c'est-à-dire qu'il n'y a pas d'interaction entre E et F par construction du modèle, indépendamment des données observées.

Si l'on veut rendre possible l'existence d'une interaction, il faut que OR_1 puisse être différent de OR_0. Le modèle logistique correspondant est : $\text{logit } P = \alpha + \beta E + \gamma F + \delta EF$, avec $EF = 0$ si $E = 0$ ou $F = 0$ et $EF = 1$ si $E = F = 1$.
Avec ce modèle, on a :

- pour $F = 0$: $\ln OR_0 = \beta$
- pour $F = 1$: $\ln OR_1 = \beta + \delta$.

EF est appelé terme d'interaction. C'est une variable qui n'a pas d'interprétation concrète. Son coefficient δ est estimé à partir des données observées. Si on rejette l'hypothèse $H_0 : \delta = 0$, c'est qu'il y a une interaction significative entre E et F.

> Remarque
> L'interaction est représentée de façon additive avec un coefficient δ ($\ln OR_1 = \beta + \delta$) plutôt que sous une autre forme, par exemple $\beta \times \delta$. Cela est cohérent avec le fait que le modèle logistique est un modèle *linéaire* généralisé.

Pour poursuivre cette introduction, je vais commencer par présenter en détail la façon de prendre en compte une interaction entre variables qualitatives. Je dirai ensuite un mot sur l'interaction lorsqu'une des variables est quantitative.

VI.1. Interaction entre variables qualitatives

Dans un modèle logistique avec une interaction ($\text{logit } P = \alpha + \beta E + \gamma F + \delta EF$), le terme d'interaction se « comporte » comme une variable supplémentaire, mais tient une place particulière. D'une part, la présence de EF et l'interprétation de son coefficient

supposent que les variables E et F soient aussi présentes dans le modèle. Un modèle avec seulement EX n'est pas impossible théoriquement, mais il impliquerait des hypothèses spécifiques et très fortes sur les associations entre E, F et M. Il n'est jamais utilisé en pratique courante. D'autre part, la présence de EF modifie l'interprétation des coefficients β et γ, ainsi qu'on le verra plus loin.
D'un point de vue pratique, la variable d'interaction EF peut être construite par un programme (« à la main »), mais les logiciels ont des commandes pour le faire qui présentent l'avantage de conserver le lien entre les variables initiales E et F et le terme d'interaction, et donc d'inclure dans le modèle à la fois l'interaction et les variables initiales[4].

VI.1.a. Exemple

Dans cet exemple, la variable E est clomid (induction de la grossesse par Clomid, non/oui) et la variable F age30 (âge ≥ 30 ans, non/oui). La maladie M est ct (GEU, non/oui) comme dans les exemples précédents. Les femmes dont la grossesse a été induite par un autre traitement que le Clomid sont exclues de l'analyse.
Les résultats de l'estimation des coefficients du modèle logit $P = \alpha + \beta E + \gamma F + \delta EF$ sont donnés dans le Tableau 3.16.

```
. logit ct i.clomid##i.age30

Logistic regression                             Number of obs   =       1683
                                                LR chi2(3)      =      44.67
                                                Prob > chi2     =     0.0000
Log likelihood =  -1046.124                     Pseudo R2       =     0.0209

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.clomid |   1.234339   .3807449     3.24   0.001     .4880927    1.980585
    1.age30 |   .6511233   .1072584     6.07   0.000     .4409006    .8613459
             |
clomid#age30 |
       1 1  |  -1.127027   .5354195    -2.10   0.035    -2.17643    -.0776236
             |
       cons |    -1.0267   .0744555   -13.79   0.000    -1.17263    -.8807696
------------------------------------------------------------------------------
```

Tableau 3.16 : Modèle logistique avec une interaction entre l'âge de la femme et l'induction de la grossesse avec du Clomid

Outre le fait qu'il faut penser à mettre deux signes # et le préfixe « i. » dans la commande, il est utile de comprendre ce que fait le logiciel Stata pour interpréter correctement les résultats. La commande i.clomid##i.age30 crée :

- ✓ des variables indicatrices pour clomid et age30 (notées 0.clomid, 1.clomid, 0.age30 et 1.age30), en excluant du calcul celles correspondant aux catégories de référence ;
- ✓ des termes d'interaction combinant les variables indicatrices notés : 0.clomid#0.age30, 0.clomid#1.age30, 1.clomid#0.age30 et 1.clomid#1.age30 (attention, il n'y

4. • Avec Stata, la variable d'interaction est construite avec le système de *factor variables* par la syntaxe i.var1##i.var2.
• Avec R, l'interaction est construite avec le signe * : var1*var2.

a qu'un seul # pour le nom des variables). Seul le dernier, qui ne porte pas sur les catégories de référence, est conservé.

Remarques

- Dans le cas présent de variables dichotomiques, la création (puis l'élimination partielle) de toutes ces variables peut paraître inutile. Par exemple, 1.clomid et 1.age30 sont identiques aux variables initiales clomid et age30. Ce principe de décomposition en variables indicatrices servira pour des variables à plus de deux classes.
- Les variables créées n'apparaissent pas dans la liste des variables de Stata. Dans les résultats, 1.clomid et 1.age30 sont indiqués à la place de clomid et age30 ; pour l'interaction, ce n'est pas 1.clomid#1.age30 qui apparaît, mais clomid#age30 avec en dessous 1 1. La raison sera plus facile à comprendre dans le cas de variables à plus de deux classes.
  Le logiciel R est, dans ce cas, plus lisible en faisant figurer dans les résultats clomid, age30 et, pour le terme d'interaction, clomid:age30.

VI.1.b. Interprétation des coefficients du modèle

Le modèle s'écrit : logit P = −1,027 + 1,234 clomid + 0,651 age30 − 1,127 clomid*age30. Pour interpréter correctement ses coefficients, il ne faut pas oublier qu'il y a **deux** odds ratios associés au Clomid, puisque le modèle contient un terme d'interaction :

- chez les femmes de moins de 30 ans (age30 = 0) : $OR_0 = e^{\beta} = e^{1,234} = 3,43$
- chez les femmes de plus de 30 ans (age30 = 1) : $OR_0 = e^{\beta + \delta} = e^{1,234 - 1,127} = 1,11$.

On voit donc que la ligne 1.clomid (ou clomid dans R) n'indique pas l'OR associé au Clomid, mais l'OR associé au Clomid dans la catégorie de référence de l'autre variable intervenant dans l'interaction, c'est-à-dire ici chez les moins de 30 ans.

Le test du coefficient de la variable d'interaction est un test de comparaison de deux odds ratios OR_0 et OR_1, c'est-à-dire un test d'interaction. Il est ici tout juste significatif (p = 0,035).

Remarque

Le « tout juste significatif » de la ligne précédente fait référence au § VI.1.d ci-dessous, mais aussi au fait qu'il faut toujours garder une certaine réserve en tirant des conclusions d'un test dont le degré de signification est proche du seuil de 5 %. Il s'agit de ne pas transformer dans son raisonnement un continuum (p) en une dichotomie (significative ou pas).

L'interaction est une notion symétrique entre les deux variables E et F. Il y a deux odds ratios associés au Clomid, ainsi qu'on vient de le voir, mais il y a aussi deux odds ratios associés à l'âge supérieur à 30 ans :

- chez les femmes sans Clomid : $OR = e^{0,651} = 1,92$
- chez les femmes avec Clomid : $OR = e^{0,651 - 1,127} = 0,62$.

Le test d'interaction est le même. Il indique que ces deux OR sont (tout juste…) significativement différents.

VI.1.c. Intervalle de confiance des odds ratios en cas d'interaction

Les deux odds ratios associés au Clomid sont obtenus par : ln $OR_0 = \beta$ et ln $OR_1 = \beta + \delta$. Pour avoir leur intervalle de confiance, il faut passer par ceux de β et $\beta + \delta$:

IC de β : $\hat{\beta} \pm 1{,}96\sqrt{\text{var}(\hat{\beta})}$

IC de $\beta + \delta$: $\hat{\beta} + \hat{\delta} \pm 1{,}96\sqrt{\text{var}(\hat{\beta} + \hat{\delta})} = \hat{\beta} + \hat{\delta} \pm 1{,}96\sqrt{\text{var}(\hat{\beta}) + \text{var}(\hat{\delta}) + 2\,\text{cov}(\hat{\beta}, \hat{\delta})}$

L'intervalle de confiance de β (et, par suite, celui de $OR_0 = e^{\beta}$) se lit directement dans les résultats de la commande logit (ou logistic). Ce n'est pas le cas de l'intervalle de confiance de $\beta + \delta$, qui fait intervenir la covariance entre β et δ. Le plus simple pour l'obtenir est d'utiliser la commande qui permet d'estimer une combinaison linéaire de plusieurs coefficients (lincom avec Stata comme avec R) et donc l'OR correspondant et son intervalle de confiance (Tableau 3.17).

On obtient finalement : $OR_0 = 3{,}44$ [1,63 ; 7,25] et $OR_1 = 1{,}11$ [0,53 ; 2,33].

```
. qui logistic ct i.clamid##i.age30
. lincom 1.clamid + 1.clamid#1.age30

 ( 1)  [ct]1.clamid + [ct]1.clamid#1.age30 = 0

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .1073125   .3764404     0.29   0.776    -.6304971    .8451221
------------------------------------------------------------------------------
```

Tableau 3.17 : Estimation et intervalle de confiance de l'OR en cas d'interaction

VI.1.d. Interaction et facteur à plus de deux classes

Lorsque la variable F est qualitative avec plus de deux classes, l'expression de son interaction avec E suit le même principe, après avoir décomposé F en variables indicatrices.

Si F a 3 classes (on généralise facilement à un plus grand nombre), les variables indicatrices sont F_1 et F_2. Deux termes d'interaction sont construits, EF_1 et EF_2, définis par $EF_i = 0$ si $E = 0$ ou $F_i = 0$ et $EF_i = 1$ si $E_i = F_i = 1$, pour $i = 1, 2$.

Le modèle logistique s'écrit : $\text{logit}\, P = \alpha + \beta E + \gamma_1 F_1 + \gamma_2 F_2 + \delta_1 EF_1 + \delta_2 EF_2$.

Il y a donc 3 odds ratios associés à E, un par niveau de F :

- ✓ $OR_0 = e^{\beta}$ pour le niveau 0 de F,
- ✓ $OR_1 = e^{\beta + \delta_1}$ pour le niveau 1 de F,
- ✓ $OR_2 = e^{\beta + \delta_2}$ pour le niveau 2 de F.

En poursuivant avec l'exemple précédent, mais cette fois avec, pour l'âge, la variable agec3 en trois classes codées 0 : < 30 ans, 1 : 30-34 ans, 2 : ≥ 35 ans, on obtient (Tableau 3.18) :

- ✓ chez les femmes de moins de 30 ans (agec3 = 0) : $OR_0 = e^{\beta} = e^{1{,}234} = 3{,}44$
- ✓ chez les femmes de 30 à 34 ans (agec3 = 1) : $OR_0 = e^{\beta + \delta_1} = e^{1{,}234 - 1{,}292} = 0{,}94$
- ✓ chez les femmes de plus de 35 ans (agec3 = 2) : $OR_2 = e^{\beta + \delta_2} = e^{1{,}234 - 0{,}701} = 1{,}70$.

Le même procédé que précédemment, avec la commande lincom, permet d'obtenir leur intervalle de confiance : $OR_0 = 3,44$ [1,63 ; 7,25], $OR_1 = 0,94$ [0,39 ; 2,29] et $OR_2 = 1,70$ [0,40 ; 7,35].

L'OR chez les femmes de moins de 30 ans est logiquement le même que dans les calculs précédents avec age30, puisque ce sont les mêmes classes.

Le test de l'interaction est un test global portant sur δ_1 et δ_1, comme c'était le cas pour une variable indicatrice (voir § III.4). Les hypothèses testées sont :

- ✓ $H_0 : \delta_1 = 0$ et $\delta_1 = 0$
- ✓ $H_1 : \delta_1 \neq 0$ ou $\delta_1 \neq 0$.

Les résultats, donnés à la fin du Tableau 3.18 avec un test de Wald, auraient aussi pu être obtenus par le test du rapport des vraisemblances. L'interaction entre l'induction de la grossesse par Clomid et l'âge en trois classes n'est pas significative, alors qu'elle l'était avec l'âge en deux classes. Dans les deux cas, on est à la limite du seuil de signification (une fois en dessous, l'autre au-dessus). Cela souligne que l'écart entre les deux situations n'est pas considérable, même si la limite de signification passe entre les deux.

```
. logit ct i.clomid##i.agec3

(...)

Logistic regression                                     Number of obs =  1,683
                                                        LR chi2(5)    =  53.76
                                                        Prob > chi2   = 0.0000
Log likelihood = -1041.5811                             Pseudo R2     = 0.0252

------------------------------------------------------------------------------
          ct | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    1.clomid |   1.234339   .3807449     3.24   0.001     .4880927    1.980585
             |
       agec3 |
          2  |   .5251477   .1173772     4.47   0.000     .2950926    .7552029
          3  |   1.004227   .1673869     6.00   0.000     .6761546    1.332299
             |
clomid#agec3 |
        1 2  |  -1.292403   .5912944    -2.19   0.029    -2.451319   -.1334871
        1 3  |   -.701041   .8371229    -0.84   0.402    -2.341772    .9396897
             |
       _cons |    -1.0267   .0744555   -13.79   0.000     -1.17263   -.8807696
------------------------------------------------------------------------------


. testparm (1.clomid#2.agec3) (1.clomid#3.agec3)    //  ou testparm 1.c*#1.a* 1.c*#2.a*

 ( 1)  [ct]1.clomid#2.agec3 = 0
 ( 2)  [ct]1.clomid#3.agec3 = 0

           chi2(  2) =    4.82
         Prob > chi2 =    0.0898
```

Tableau 3.18 : Modèle logistique avec une interaction entre l'âge de la femme en trois classes agec3 (< 30 ans ; 30-34 ans ; ≥ 35 ans) et l'induction de la grossesse avec du Clomid

VI.2. Interaction avec F et analyses séparées selon les niveaux de F

L'inclusion de termes d'interaction dans un modèle logistique, ainsi que cela a été fait dans les paragraphes précédents, est équivalente à réaliser des analyses séparées

dans les catégories de la variable d'interaction. En prenant deux variables dichotomiques comme dans le § VI.1.a, cela signifie précisément qu'on obtient les mêmes résultats avec le modèle logit $P = \alpha + \beta E + \gamma F + \delta EF$ et avec le modèle logit $P = \alpha + \beta E$ appliqué successivement sur les deux sous-échantillons définis par $F = 0$ et $F = 1$ (Tableau 3.19).

```
1. Analyse avec un terme d'interaction

. logit ct i.clamid##i.age30

Logistic regression                             Number of obs   =       1683
                                                LR chi2(3)      =      44.67
                                                Prob > chi2     =     0.0000
Log likelihood =  -1046.124                     Pseudo R2       =     0.0209

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.clamid |   1.234339   .3807449     3.24   0.001     .4880927    1.980585
     1.age30 |   .6511233   .1072584     6.07   0.000     .4409006    .8613459
             |
clamid#age30 |
       1 1   |  -1.127027   .5354195    -2.10   0.035    -2.17643    -.0776236
             |
       _cons |    -1.0267   .0744555   -13.79   0.000    -1.17263    -.8807696
------------------------------------------------------------------------------

2. Analyses séparées

X=0 (moins de 30 ans)

. logit ct clamid if age30==0

Logistic regression                             Number of obs   =        958
                                                LR chi2(1)      =      10.35
                                                Prob > chi2     =     0.0013
Log likelihood = -555.90136                     Pseudo R2       =     0.0092
------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      clamid |   1.234339   .3807449     3.24   0.001     .4880928    1.980585
       _cons |    -1.0267   .0744555   -13.79   0.000    -1.17263    -.8807696
------------------------------------------------------------------------------
X=1 (plus de 30 ans)

. logit ct clamid if age30==1

Logistic regression                             Number of obs   =        725
                                                LR chi2(1)      =       0.08
                                                Prob > chi2     =     0.7761
Log likelihood = -490.22259                     Pseudo R2       =     0.0001
------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      clamid |   .1073125   .3764404     0.29   0.776    -.6304971     .845122
       _cons |  -.3755765   .0772059    -4.86   0.000    -.5268973   -.2242556
------------------------------------------------------------------------------
```

Tableau 3.19 : Interaction et analyses séparées

Il est important de noter qu'on obtient ici strictement les mêmes résultats (mêmes tests et mêmes intervalles de confiance par exemple). En particulier, on ne gagne pas de puissance avec une analyse avec un terme d'interaction par rapport à des analyses séparées. La seule chose qu'on gagne, c'est la possibilité de tester l'existence de l'interaction. Concernant l'interprétation des résultats, une analyse avec un (ou des) terme(s) d'interaction revient donc à faire des analyses séparées.
Lorsqu'il y a d'autres variables que E et F dans le modèle, le résultat général est le même : dans son principe, une analyse avec interaction équivaut à plusieurs analyses

séparées, mais pour que l'équivalence soit complète, il faut inclure des termes d'interaction pour toutes les variables. C'est ainsi qu'avec une variable X_1 supplémentaire, l'utilisation du modèle logit $P = \alpha + \beta E + \beta_1 X_1$ successivement dans les sous-échantillons définis par $F = 0$ et $F = 1$ n'est pas équivalente à celle du modèle logit $P = \alpha + \beta E + \beta_1 X_1 + \gamma F + \delta EF$, mais à celle du modèle logit $P = \alpha + \beta E + \beta_1 X_1 + \gamma F + \delta EF + \delta' X_1 F$
En pratique, il est très fréquent de n'inclure que l'interaction d'intérêt, EF par exemple, et d'interpréter quand même les résultats comme des analyses séparées pour $F = 0$ et $F = 1$ (c'est-à-dire comme si le terme d'interaction X_1F était aussi inclus).

VI.3. Interaction avec une variable quantitative

Si F est une variable quantitative, la notion d'interaction entre E et F est plus compliquée à formuler et devient moins symétrique. Cela n'a pas de sens de dire qu'il y a interaction entre E et F si la relation entre E et M est différente selon les niveaux de F puisque F comporte, a priori, autant de niveaux que de sujets dans l'échantillon. En prenant la question dans « l'autre sens », on peut dire qu'il y a interaction si la relation entre F et M n'est pas la même selon les niveaux de E (deux niveaux si E est dichotomique), mais on ne peut pas se limiter à étendre la modélisation précédente sous la forme logit $P = \alpha + \beta E + \gamma F + \delta EF$, où EF est le produit des variables E et F. En effet, ce modèle suppose que la relation entre F et logit P est linéaire à la fois lorsque $E = 0$ et lorsque $E = 1$, et que les pentes de ces droites sont différentes s'il y a une interaction. Ces hypothèses sont trop restrictives.
Des méthodes plus complètes ont été développées pour prendre en compte, ou tester, des interactions lorsqu'au moins une des variables est quantitative (Royston P et al., 2004, Royston P et al., 2008, Royston P et al., 2014). Elles passent par la modélisation de la relation entre F et logit P par des polynômes fractionnaires. Leur compréhension nécessite d'avoir lu le § VIII du chapitre 4.

VII. Annexe : Comment déterminer si deux modèles sont emboîtés ?

Par définition, deux modèles sont emboîtés si l'un est un cas particulier (ou un sous-modèle) de l'autre.
Le cas le plus simple est celui où M' est déduit de M par retrait d'une des variables. Considérons par exemple M : logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2$ et M' : logit $P = \alpha' + \beta' X_1$. En imposant $\beta_2 = 0$ dans M, on retrouve M'. En effet, bien que les paramètres ne soient pas notés de la même façon, logit $P = \alpha + \beta_1 X_1$ est identique au modèle M' car, appliqué à un même échantillon, il donnera les mêmes estimations du coefficient de la variable X_1 et de la constante.

Dans les autres cas, il faut un moyen général pour déterminer si un modèle est un sous-modèle d'un autre. Pour cela, je vais prendre l'exemple du niveau d'études « etu », variable qualitative ordinale à trois classes, codée 0 : niveau d'études primaire ou moins, 1 : secondaire, 2 : supérieur.

On va considérer deux autres moyens de représenter le niveau d'études :

- ✓ une variable en deux classes prim (niveau d'études primaire ou moins : oui/non), qui revient donc à regrouper les deux dernières classes de etu ;
- ✓ une décomposition en deux variables indicatrices : etu_1 et etu_2.

Ces trois représentation (ou codages) du niveau d'études sont résumées dans le tableau ci-dessous.

		prim	et_1	et_2
etu	0 : niveau d'études ≤ primaire	0	0	0
	1 : niveau d'études secondaire	1	1	0
	2 : niveau d'études supérieur	1	0	1

Les trois modèles logistiques correspondants s'écrivent :

- ✓ M_1 : logit $P = \alpha_0 + \beta_1 et_1 + \beta_2 et_2$
- ✓ M_2 : logit $P = \alpha' + \beta'$ prim
- ✓ M_3 : logit $P = \alpha_3 + \beta_3$ etu.

Ils permettent, chacun de façon différente, d'analyser la relation entre le niveau d'études et la maladie. Pour les comparer deux à deux sur le plan statistique, on peut utiliser le test du rapport des vraisemblances à condition de vérifier auparavant qu'ils sont emboîtés.

M' est un sous-modèle de M (ou est emboîté dans M) si, en partant de M et en imposant certaines relations entre ses coefficients, on retrouve le modèle M'. Pour déterminer si c'est bien le cas, il faut écrire que les valeurs de logit P sont égales pour toutes les catégories de niveau d'études. Si les équations obtenues ont une solution générale, c'est-à-dire une solution qui ne porte que sur des relations entre les coefficients de M, les deux modèles sont emboîtés et on peut déduire de la solution trouvée les relations nécessaires entre les coefficients pour passer d'un modèle à l'autre.

Les valeurs de logit P pour les trois catégories de niveau d'études et les trois modèles considérés sont indiquées dans le tableau ci-dessous.

	logit P		
	M_1	M_2	M_3
niveau d'études ≤ primaire	α_0	α'	α_3
niveau d'études secondaire	$\alpha_0 + \beta_1$	$\alpha' + \beta'$	$\alpha_3 + \beta_3$
niveau d'études supérieur	$\alpha_0 + \beta_2$	$\alpha' + \beta'$	$\alpha_3 + 2\,\beta_3$

On en déduit les équations à résoudre pour comparer deux modèles l'un avec l'autre.

- Comparaison de M_1 et M_3
  Les équations sont les suivantes :

$$\begin{array}{lcl} \alpha_0 = \alpha_3 & & \alpha_0 = \alpha_3 \\ \alpha_0 + \beta_1 = \alpha_3 + \beta_3 & \Leftrightarrow & \beta_1 = \beta_3 \\ \alpha_0 + \beta_2 = \alpha_3 + 2\,\beta_3 & & \beta_2 = 2\,\beta_3 \end{array}$$

  On voit donc que ces égalités sont satisfaites si on a la relation $\beta_2 = 2\,\beta_1$ entre les coefficients de M_1. Les modèles M_1 et M_3 sont donc emboîtés. M_3 est un cas particulier de M_1. Le test du rapport des vraisemblances est un test de χ^2 à 1 degré de liberté (différence entre le nombre de coefficients de M_1 et M_3). Il correspond au test des hypothèses : $H_0 : \beta_2 = 2\,\beta_1$ et $H_1 : \beta_2 \neq 2\,\beta_1$.

  Si les vraisemblances sont significativement différentes ($\chi_0^2 \geq 3{,}84$), le modèle M_1 est le meilleur, au sens où il est plus adéquat aux données.
  Si les vraisemblances ne dont pas significativement différentes, il n'y a pas de raison statistique de choisir l'un plutôt que l'autre. On prend donc le modèle le plus simple, c'est-à-dire M_3.

  Remarques
  - De façon générale, si on considère une variable qualitative X à k classes décomposée en variables indicatrices, le modèle avec la variable X sous sa forme initiale (non décomposée) est toujours emboîté dans le modèle avec les (k – 1) variables indicatrices. Le test de comparaison est un χ^2 à (k – 2) degrés de liberté.
  - La comparaison entre M_1 et M_3 est le test de linéarité du § IV.3.
  - Le résultat obtenu indique aussi qu'on peut utiliser le test de Wald pour comparer les modèles M_1 et M_3. Il ne s'écrit pas sous la forme d'un coefficient égal à 0, mais d'une égalité entre coefficients : $\beta_2 = 2\,\beta_1$.

- Comparaison de M_1 et M_2
  Les équations sont les suivantes :

$$\begin{array}{lcl} \alpha_0 = \alpha' & & \alpha_0 = \alpha' \\ \alpha_0 + \beta_1 = \alpha' + \beta' & \Leftrightarrow & \beta_1 = \beta' \\ \alpha_0 + \beta_2 = \alpha' + \beta' & & \beta_2 = \beta' \end{array}$$

  On voit donc que ces égalités sont satisfaites si on a la relation $\beta_2 = \beta_1$ entre les coefficients de M_1. Les modèles M_1 et M_2 sont donc emboîtés. M_2 est un cas particulier de M_1. On peut comparer ces deux modèles sur le plan statistique par la méthode du rapport des vraisemblances. Le test du rapport des vraisemblances est un χ^2 à 1 degré de liberté (différence entre le nombre de coefficients de M_1 et M_2). Il correspond au test des hypothèses : $H_0 : \beta_2 = \beta_1$ et $H_1 : \beta_2 \neq \beta_1$.
  Si les vraisemblances sont significativement différentes ($\chi_0^2 \geq 3{,}84$), le modèle M_1 est le meilleur, au sens où il est plus adéquat aux données.

Si les vraisemblances ne sont pas significativement différentes, il n'y a pas de raison statistique de choisir l'un plutôt que l'autre. On prend donc le modèle le plus simple, c'est-à-dire M_2.

Remarques

- De façon générale, si on considère une variable qualitative X à k classes, le modèle avec la variable X sous sa forme initiale est toujours emboîté dans le modèle, avec une variable X' à k' classes obtenues en regroupant des catégories de X. Le test de comparaison est un χ^2 à (k – k') degrés de liberté.
- Là aussi, un test de Wald est possible.

- Comparaison de M_2 et M_3
  Les équations sont les suivantes :
  $\alpha' = \alpha_3$
  $\alpha' + \beta' = \alpha_3 + \beta_3$
  $\alpha' + \beta' = \alpha_3 + 2\,\beta_3$

Ces équations n'ont pas de solution générale car les deux dernières conduisent à $\beta_3 = 2\,\beta_3$, c'est-à-dire à $\beta_3 = 0$ (et aussi à $\beta' = 0$). Cela ne correspond ni au modèle général M_2, ni au modèle général M_3, mais à une absence d'association entre la maladie et le niveau d'études, cas où le niveau d'études n'a pas à être pris en compte.
Les modèles M_2 et M_3 ne sont donc pas emboîtés, et par conséquent, pas comparables sur le plan statistique par la méthode du rapport des vraisemblances.

Remarque : cela tombe bien si j'ose dire, car le test χ^2 qui comparerait ces deux modèles aurait zéro degré de liberté (ils auraient le même nombre de paramètres), ce qui n'aurait pas de sens.

Chapitre 4

Modélisation des variables quantitatives

• • •

I. Introduction

L'étude de l'association entre la maladie Y et une variable quantitative X se présente de façon différente et, disons-le, plus compliquée que lorsque X est une variable dichotomique (et même qualitative, de façon plus générale). Au-delà du test de l'existence de l'association et de sa quantification, la forme de la relation est importante pour tirer parti au mieux du caractère quantitatif de X, et surtout pour éviter des erreurs d'interprétation, en particulier si la relation entre X et Y n'est pas monotone.

Une des méthodes les plus fréquemment utilisées est de faire des classes de X pour transformer cette variable en une variable qualitative ordonnée. Cela ramène aux méthodes présentées dans le chapitre 3. Je montrerai l'importance des limites et des inconvénients de cette méthode dans le § III, mais je reviendrai sur ses caractéristiques dans le § VI, car elle reste très utilisée.
Si on ne scinde pas X en classes, il est possible d'inclure X sous sa forme initiale dans le modèle logistique sous la forme : logit $P = \alpha + \beta X$. On lit ou on entend souvent que X est alors pris sous forme quantitative, ce qui est abusif. Il vaudrait mieux dire que X est pris en compte sous forme linéaire, et surtout il faudrait vérifier qu'il n'y a pas d'écart à la linéarité, ce qui n'est pas toujours réalisé. Il n'empêche que le modèle linéaire occupe une place à part, sur laquelle je reviendrai dans le § IV.3.
On comprend donc qu'il y a besoin de méthodes permettant de conserver X sous sa forme quantitative initiale et de modéliser la forme de l'association entre X et Y sans la supposer linéaire a priori. C'est le rôle des méthodes présentées dans les § VII à IX, et en particulier des polynômes fractionnaires et des fonctions splines.

Dans ce chapitre, comme dans les autres, la variable Y est dichotomique et analysée par régression logistique (sauf pour quelques illustrations dans les § II et III). Les méthodes présentées sont cependant applicables pour tous les modèles linéaires généralisés (notamment pour le modèle linéaire lorsque Y est quantitative) ainsi que pour le modèle de Cox.

II. Représentation graphique de la relation entre X et Y

Pratiquement toutes les méthodes que nous verrons modélisent la relation entre X et Y par une fonction du type $Y = f(X)$ où f a une expression plus ou moins compliquée. Si par exemple $f(X) = \alpha + \beta X$, c'est-à-dire que la modélisation est de type linéaire, on voit facilement que la relation est représentée par une droite, et qu'elle est croissante

ou décroissante selon le signe de β. Mais si $f(X) = \alpha + \beta_1 X^3 + \beta_2 X^3 \ln X$, modélisation par polynôme fractionnaire, on ne peut pas « deviner » la forme de la relation à la lecture de la fonction, ni même si cette relation est croissante ou pas. Il est donc nécessaire de représenter la relation graphiquement pour interpréter correctement les résultats. Bien que les outils des logiciels d'analyse statistique soient parfois « casse-tête » pour obtenir un résultat graphique agréable, il est important d'associer des représentations graphiques à l'analyse de la relation entre X et Y, comme nous le verrons tout au long de ce chapitre.

Prenons l'exemple de la relation entre le terme de naissance X (en semaines) et le poids de naissance Y (en kg). Ici Y est quantitative pour la commodité des illustrations qui vont suivre, mais je reviendrai ensuite sur la situation où Y est dichotomique. Les données sont celles des naissances de l'année 1985 à la maternité de Haguenau en Alsace (Bouyer J et al., 1987). Lorsqu'on modélise la relation entre X et Y par une droite, la fonction est $f(X) = -2{,}22 + 0{,}139\ X$. Lorsqu'on la modélise par un polynôme fractionnaire, on obtient $f(X) = -4{,}54 + 0{,}540\ X^3 - 0{,}300\ X^3 \ln X$. Il n'y a qu'avec la représentation graphique ci-contre (Figure 4.1) qu'on peut visualiser la position des courbes, et en particulier se rendre compte de ce qu'elles sont relativement proches.

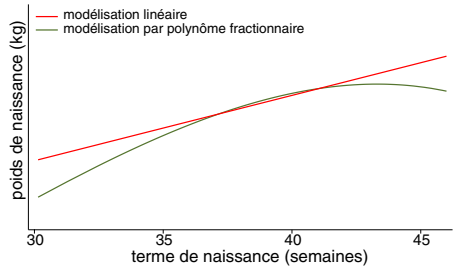


Figure 4.1 : Deux modélisations de la relation entre terme et poids de naissance.

On souhaite aussi souvent représenter les observations sur le graphique. On espère, même si c'est en partie illusoire, que cela va donner des idées sur la forme de la relation entre X et Y ou permettre de vérifier que la modélisation retenue est adéquate aux valeurs observées.

Poursuivons tout d'abord avec l'exemple des termes et poids de naissance. Les observations sont représentées par le nuage de points de la Figure 4.2a, où chaque point représente un sujet.

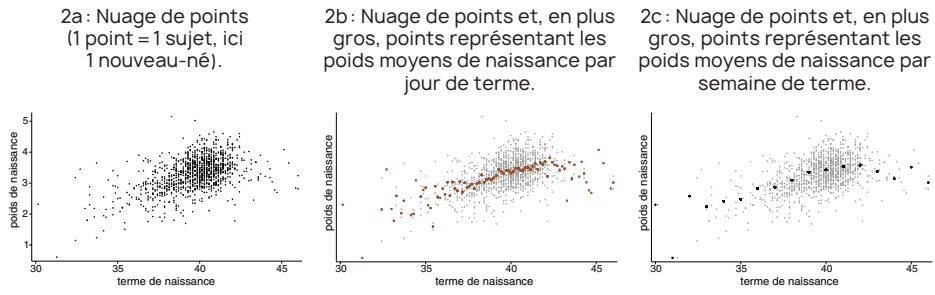


Figure 4.2 : Terme et poids de naissance des 1235 nouveau-nés de la maternité de Haguenau (Alsace) de l'année 1985.

Le nuage de points indique une grande tendance générale (le poids de naissance augmente avec le terme), mais ne permet pas d'en dire beaucoup plus sur la forme de la relation, en particulier si elle est linéaire. Cela est en partie dû au fait que la régression de Y en fonction de X ne vise pas à représenter l'ensemble des données individuelles, mais modélise la moyenne (ou espérance) de Y à X fixé, E(Y|X). Il est donc utile d'ajouter au nuage de points les points représentant la moyenne de Y à X fixé. C'est ce qui est fait sur la Figure 4.2b, avec des regroupements des valeurs de X par jour de terme, qui est la précision maximum du terme. On voit que la pertinence de cette représentation des moyennes de Y à X fixé est limitée par le peu d'observations pour les termes extrêmes, ce qui rend les points moyens correspondants très sensibles aux fluctuations d'échantillonnage. On préfèrera un regroupement en catégories plus larges, par exemple par semaine de terme (Figure 4.2c).

> Remarque
> Il faut bien noter que ce regroupement en classes de terme ne sert que pour représenter les points « observés » ; la relation entre X et Y doit être modélisée avec les valeurs individuelles du terme (en jours ici), sans établir de classes.

On peut aussi noter que le nuage de points ne permet pas non plus à lui seul de choisir visuellement entre modélisation linéaire ou polynôme fractionnaire (Figure 4.3). C'est déjà plus facile avec le regroupement par semaine de terme, mais il est préférable de recourir aux méthodes statistiques que nous verrons plus loin, et surtout de ne pas oublier les connaissances cliniques indiquant une diminution du poids de naissance moyen au-delà de 42 semaines de gestation, qui induisent à adopter la courbe par polynôme fractionnaire.

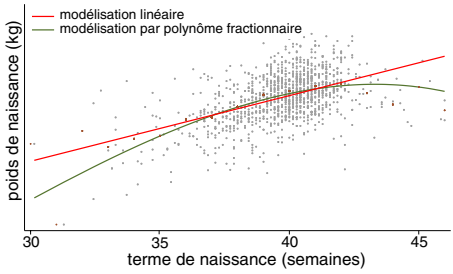


Figure 4.3 : Deux modélisations de la relation entre terme et poids de naissance

Lorsque Y est 0/1, comme c'est souvent le cas en épidémiologie où Y est la variable « malade oui/non » ou « succès/échec », le nuage de points est non informatif. On obtient en effet deux séries de points alignés, l'un à l'ordonnée 0 et l'autre à l'ordonnée 1 (Figure 4.4a). On ne peut pas non plus utiliser logit P, qui est une quantité non observable au niveau individuel (elle serait égale à $-\infty$ ou $+\infty$). On peut alors procéder comme précédemment, en regroupant les valeurs de X en classes (ici par exemple par semaine de terme) et en représentant les valeurs moyennes de logit P (Figure 4.4b).

Cette représentation ne résout cependant pas tous les problèmes, car logit P n'est pas calculable quand $P = 0$ ou $P = 1$ (tous les sujets d'une même classe ont la même valeur de Y). Pour les graphiques de ce chapitre, j'ai choisi de représenter ces points par des losanges situés en bas du graphique si $P = 0$ (et donc logit $P = -\infty$) ou en

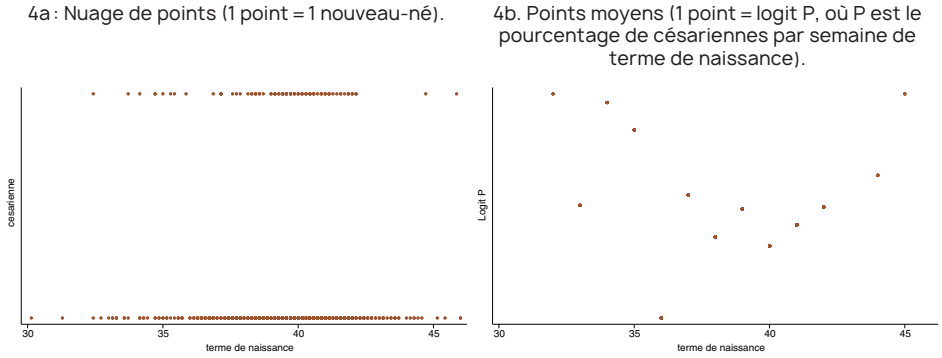


Figure 4.4 : Accouchement par césarienne en fonction du terme de naissance des nouveau-nés de la maternité de Haguenau (Alsace) de l'année 1985.

haut si P = 1 (et donc logit P = +∞) et surmontés du nombre de sujets concernés (voir Figure 4.5).

Notons de nouveau que ce n'est qu'un problème de représentation graphique. Dans les analyses où X est prise comme une variable quantitative, l'ensemble des données sont considérées (celles représentées par des losanges ne sont pas exclues).

Ce choix fait pour représenter les données observées a le mérite de la simplicité et de permettre une représentation facilement interprétable des données. Il présente cependant l'inconvénient de ne pas complètement conserver la nature quantitative de X, puisqu'on fait des classes pour représenter les données observées. Le logiciel Stata utilise une autre méthode (qui n'est pas sans inconvénients) pour les représentations graphiques des courbes obtenues par modélisation avec des polynômes fractionnaires. Elle est explicitée en annexe (§ XII.2).

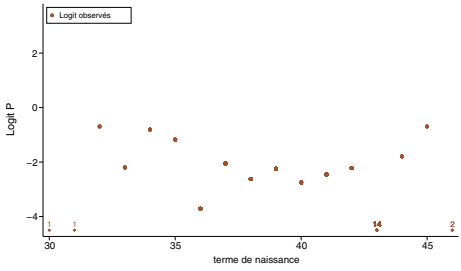


Figure 4.5 : Représentation des logit du nombre de césariennes observées par classe de terme d'une semaine.

III. Transformer (ou pas) une variable quantitative en classes

La pratique qui reste la plus courante en épidémiologie pour inclure une variable quantitative X dans un modèle de régression est la catégorisation, c'est-à-dire la transformer en variable qualitative en faisant des classes. Cette nouvelle variable qualitative est, de plus, souvent analysée comme une variable nominale en la remplaçant par des variables indicatrices (voir chapitre 3, § III), donc sans tenir compte de l'ordre des classes, de sorte que la notion même de variable quantitative est quasiment perdue.

Même si une évolution de cette pratique dans le temps a pu être notée, avec une légère décroissance (Del Priore G et al., 1997), l'impression générale qu'il s'agit de la méthode de référence ressort de la lecture des articles publiés (Mabikwa OV et al., 2017), mais aussi des mémoires d'étudiants ou des thèses (c'est donc ce que retiennent de leur formation les plus jeunes épidémiologistes, ou ce que de plus anciens épidémiologistes leur recommandent de faire…). La fréquence de la catégorisation dans les publications scientifiques en épidémiologie a été quantifiée par Turner et al. (Turner E et al., 2010), qui ont revu les articles d'épidémiologie observationnelle de cinq grands journaux d'épidémiologie et de cinq grands journaux généralistes de médecine en décembre 2007 et janvier 2008. Sur les 58 articles pertinents, 86 % transformaient la variable quantitative en classes, dont moins de la moitié l'étudiaient aussi quantitativement. On peut noter que 6 % seulement se limitaient à deux classes (dichotomisation exposés/non-exposés), et que la majorité (60 %) constituaient 4 ou 5 classes. À la fin de leur article, les auteurs proposent des recommandations en 13 points pour guider le choix de la catégorisation, le dernier point étant d'ailleurs de considérer la possibilité de ne pas catégoriser. Ces résultats sont confirmés par Mabikwa et al. (Mabikwa OV et al., 2017), qui ajoutent que plus des deux tiers des articles ne donnent pas de raison pour avoir catégorisé Y.

Les avantages le plus souvent mis en avant en faveur de la catégorisation d'une variable quantitative relèvent de deux grands types d'arguments :

- La présentation et l'interprétation des résultats sont plus simples et/ou mieux adaptées aux besoins.
  Il est en effet exact que la présentation des résultats en donnant des odds ratios par catégorie est plus facilement compréhensible, notamment (mais pas seulement) pour des non-statisticiens ou des non-épidémiologistes. Cela permet de plus d'avoir une vision assez claire de la progression du risque avec les valeurs de X, même si la relation dose-effet n'est pas toujours testée. En outre, les classes constituées sont souvent les catégories utilisées de façon habituelle dans les autres publications (par exemple, l'âge en classes de cinq ans) et facilitent donc les comparaisons et la discussion des résultats. Enfin, il existe souvent une catégorie « non-exposés », qui sert naturellement de référence.
- Il s'agit d'une méthode non paramétrique, en ce sens qu'elle ne nécessite pas d'hypothèses sur la forme de la relation entre les variables X et Y. En particulier, aucune hypothèse de linéarité n'est requise.

En réalité, ces deux types d'arguments sont assez « datés », et nous verrons plus loin que les méthodes actuelles de modélisation des variables quantitatives sont suffisamment souples pour ne pas être handicapées par des hypothèses paramétriques et sont en mesure de concilier la qualité de la modélisation et la simplicité de la présentation des résultats et de leur l'interprétation (§ X.3).

La tendance actuelle est nettement de souligner **les inconvénients** de la transformation d'une variable continue en catégories (Greenland S, 1995b, Greenland S, 1995c, Weinberg CR, 1995, Dinero TE, 1996, American College of Obstetricians and

Gynecologists, 1997, Altman DG et al., 2006, Royston P et al., 2006, Froslie K et al., 2010, Bennette C et al., 2012). Les arguments sont multiples et, lorsqu'ils sont pertinents, ont été validés sur le plan statistique par des simulations.
Lorsque la variable transformée en catégories est l'exposition d'intérêt principal, les deux grands types d'arguments contre sa catégorisation sont la mauvaise modélisation de la relation avec Y et le mauvais contrôle des risques d'erreurs statistiques (en particulier une perte de puissance). Lorsqu'il s'agit d'une variable d'ajustement, la prise en compte du biais de confusion n'est que partielle après catégorisation.

- Mauvaise modélisation de la relation entre X et Y
  La catégorisation de X et l'utilisation de variables indicatrices revient à représenter l'association entre X et Y par un escalier. Si on reprend l'exemple donné plus haut du terme (X) et du poids de naissance (Y), la constitution de classes de X (ici ≤ 35 ; [36-37] ; [38-40] ; [41-43] ; ≥ 44) donne la modélisation de la Figure 4.6.
  On est habitué à cette représentation en escalier, dont il faut reconnaître qu'elle résume assez bien la situation. Il faut cependant avouer également qu'elle n'a pratiquement aucune de chance de correspondre à la réalité. On ne peut pas penser que le poids moyen de naissance fait de brusques sauts aux bornes des classes.
  L'argument selon lequel, dans le cas d'une exposition X mesurée avec erreur, on peut réduire les effets de ces erreurs de mesure en faisant des classes est intuitivement séduisant, mais n'a pas fait la preuve de sa réalité (Weinberg CR, 1995). On peut par exemple se dire qu'une consommation de tabac rapportée en nombre de cigarettes par jour est sujette à des erreurs d'arrondi, avec des réponses trop fréquentes pour cinq ou dix cigarettes par jour, et qu'il est préférable de faire des classes de consommation. En réalité, au contraire, en faisant des classes, on rajoute encore de l'erreur de mesure et cela n'arrange rien.

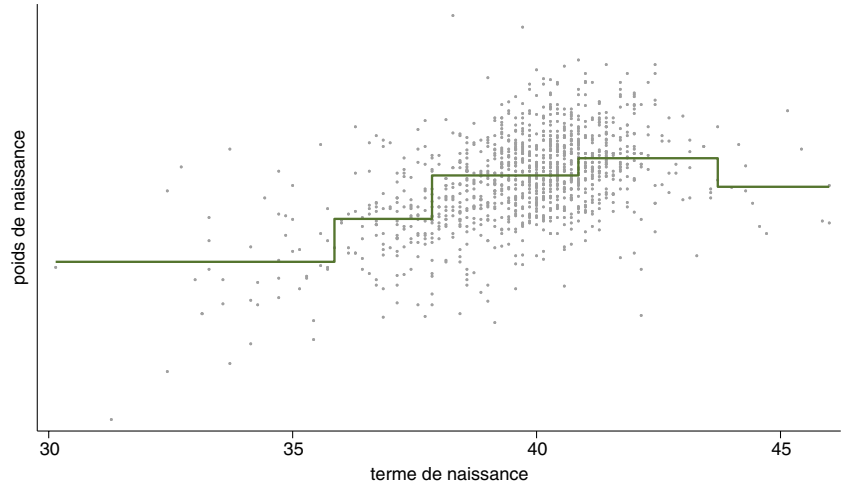


Figure 4.6 : Relation entre terme et poids de naissance après catégorisation du terme en cinq classes.

Une des formes « extrêmes », mais courante en épidémiologie, de la catégorisation est la dichotomisation : X est transformée en une variable à deux classes, de type exposés/non-exposés (obèses, hypertendus, fumeurs…). Le choix de la valeur seuil déterminant ces classes est bien sûr crucial et on a montré que les méthodes dites de « choix optimal » de ce seuil peuvent conduire à des biais (MacCallum RC et al., 2002, Royston P et al., 2006).

Le choix des bornes des classes est souvent arbitraire et les résultats peuvent en dépendre. Non seulement la forme de la courbe en escalier, mais aussi l'interprétation qu'on en fait.
Si, par exemple, on fait deux classes de terme, on obtient des résultats différents selon que la limite entre les classes est 36 ou 37 semaines. C'est ce que l'on voit sur la Figure 4.7 ci-contre. Il est clair que les courbes sont différentes, ce qui n'est peut-être pas trop grave car elles indiquent toutes les deux que le poids de naissance moyen est plus faible chez les nouveau-nés de petit terme. En revanche, l'écart de poids entre les petits termes et les autres n'est pas le même : 941 g pour une limite à 36 semaines et 739 g pour une limite à 37 semaines, soit des résultats nettement différents.

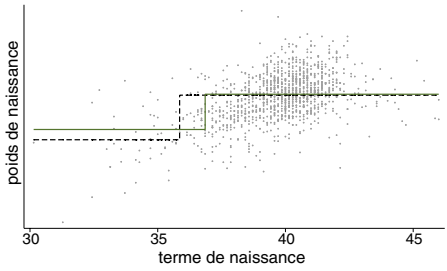


Figure 4.7 : Rôle du choix du seuil lorsque le terme de naissance est catégorisé en deux classes.

De façon plus générale, il n'y a pas de consensus sur le choix du nombre de catégories. Le plus souvent, les auteurs prennent quatre ou cinq classes (Mabikwa OV et al., 2017) pour limiter la perte d'information (Connor RJ, 1972). Il n'y a pas consensus non plus sur les bornes des classes, bien qu'on utilise souvent des percentiles (terciles, quartiles, quintiles) (Altman DG, 1998). Des considérations cliniques peuvent bien sûr intervenir, mais elles résultent souvent d'habitudes plus ou moins anciennes, qui n'ont éventuellement pas été reconsidérées, et/ou des connaissances sur le lien entre la variable X et la maladie Y. Par exemple, la définition habituelle de la prématurité (< 37 semaines de gestation) repose sur l'observation, maintenant en partie dépassée, de l'augmentation des problèmes pédiatriques chez les nouveau-nés à partir de 37 semaines.

Le choix de seuils « optimaux » (Barrio I et al., 2017), au sens où ils rendent l'adéquation du modèle ou sa capacité de prédiction les meilleurs, n'est pas recommandé, car ils peuvent varier considérablement d'une étude à l'autre (Buettner P et al., 1997) et rendre difficile toute comparaison ou combinaison entre leurs résultats. Ce type de choix accroît aussi de façon importante le risque d'erreur de première espèce (Altman DG et al., 1994, Harrell FE, 2001, Royston P et al., 2006).

Au-delà du nombre de catégories, le choix de la catégorie de référence ne fait pas l'unanimité. Dans un certain nombre de cas, elle s'impose d'elle-même lorsqu'il y

a une catégorie « non-exposés », par exemple les non-fumeurs ou les sujets non exposés à une pollution environnementale. Mais, pour des variables comme l'âge, le poids ou le BMI, il n'y a pas de telles évidences. Le choix de la catégorie de référence ne change rien aux tests d'association entre X et Y, qui doivent être des tests globaux, mais cela peut modifier la façon dont les résultats sont présentés et donc dont ils sont compris par le lecteur, voire interprétés par leurs auteurs. Froslie et al. en donnent un exemple à propos de la relation entre l'hyperglycémie du nouveau-né et le BMI de la mère (Froslie K et al., 2010).

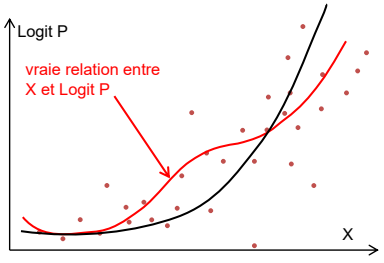
- Perte de puissance
  La catégorisation d'une variable continue peut être assimilée à une erreur de mesure, puisqu'on attribue la même valeur à toutes les observations d'une même catégorie (Altman DG, 1998). C'est donc une valeur erronée dans la plupart des cas. Comme pour toutes les erreurs de mesure, cela induit une perte de puissance pour le test de l'existence d'une association entre X et Y (Greenland S, 1995a, Moser BK et al., 2004). Elle a été évaluée comme équivalente à la perte d'environ un tiers des sujets, ce qui est considérable (Lagakos SW, 1988, Zhao LP et al., 1992). La perte de puissance semble cependant limitée avec cinq catégories ou plus (Brenner H et al., 1997).

- Mauvais ajustement
  La catégorisation d'une variable continue a aussi des conséquences lorsque cette variable intervient comme variable d'ajustement dans l'analyse. Le résultat général est que la qualité de l'ajustement est dégradée, c'est-à-dire que les coefficients des autres variables restent biaisés après ajustement sur une variable catégorisée. Il y a donc une mauvaise prise en compte des phénomènes de confusion (Taylor JMG et al., 2002, Sullivan TR et al., 2024), et il reste de la confusion résiduelle (Becher H, 1992, Brenner H et al., 1997, Brenner H, 1998, Royston P et al., 2006).
  De façon plus précise, on montre (Cochran WG, 1968) que, dans la comparaison de deux moyennes ajustée sur une variable continue Z, le biais de confusion est réduit de 64 %, 79 %, 86 %, 90 % et 92 % lorsque que Z est catégorisée respectivement en 2, 3, 4, 5 et 6 classes.
  De même, Austin et al. et Barnwell-Menard et al. (Austin PC et al., 2004a, Barnwell-Menard JL et al., 2015) ont montré qu'il y a une importante augmentation du risque d'erreur de première espèce lorsqu'on teste, avec une régression logistique, l'association entre Y et une variable continue X après ajustement sur une variable continue Z qui a été catégorisée. Dans le cas où la corrélation entre X et Z est forte, le risque d'erreur peut valoir jusqu'à 40 % (au lieu des 5 % attendus), même avec cinq catégories de Z. On a donc un risque très élevé de conclure à tort à l'existence d'une association entre X et Y.

En conclusion, malgré des habitudes parfois solidement ancrées, et malgré sa plus grande facilité d'utilisation, les inconvénients de la catégorisation des variables quantitatives pour modéliser la relation entre X et Y l'emportent. Il est préférable d'utiliser

des méthodes qui conservent la nature quantitative de X comme les polynômes fractionnaires (§ VIII) (Greenland S, 1995b, Greenland S, 1995a, Royston P et al., 2008) ou les fonctions splines (§ IX) (Harrell FE, 2001). Il faut cependant éviter de perdre la facilité de présentation des résultats, qui est un atout essentiel de la catégorisation. Je montrerai comment y parvenir dans le § X.3.

IV. Les différentes méthodes de modélisation d'une variable quantitative



Lorsqu'on modélise la relation entre X et Y, on cherche à approcher le mieux possible la vraie relation (qui est inconnue) à partir d'observations faites sur un échantillon. Sur le schéma ci-contre, la vraie relation a une forme globale « en J », avec un palier pour les valeurs intermédiaires de X. La relation modélisée est représentée par la courbe continue noire. Les observations sont les points. En réalité, ce sont des valeurs moyennes par classes de X dont on a vu au paragraphe précédent qu'elles étaient souvent plus pertinentes pour la représentation graphique que les valeurs individuelles et qui peuvent être des guides pour décider si la modélisation retenue est satisfaisante.

Il n'y a pas une façon unique de définir ce qu'est une « bonne » modélisation, de même qu'il n'y a pas de méthode universelle pour choisir le meilleur modèle (quelles variables, codées comment ?) (Royston P et al., 2008). Les règles qu'on retient habituellement pour choisir une modélisation reposent sur sa vraisemblance, c'est-à-dire un critère qu'on peut résumer comme étant la probabilité d'observer les données de l'échantillon si la modélisation était la bonne (voir chapitre 1) et qui doit être maximum.

Il ne faut cependant pas chercher à trop « coller » aux données observées. Cela donne certes une bonne vraisemblance et une bonne adéquation de la modélisation aux observations, mais il ne faut pas oublier que les observations sont sujettes à des fluctuations d'échantillonnage. Particulièrement pour les valeurs de X où elles sont peu nombreuses (souvent les valeurs extrêmes). À trop coller aux données, on risque alors de modéliser le hasard, ou du moins d'essayer de le faire car c'est peine perdue… Ainsi que j'aurai l'occasion de le redire, cela vaut pour pratiquement toutes les méthodes présentées dans ce chapitre.

Au-delà de ces considérations générales, on distingue habituellement, pour la modélisation d'une variable quantitative, les méthodes d'ajustement local et d'ajustement global, le modèle jouant un rôle à part.

IV.1. Ajustement local

Il s'agit de « découper » X en intervalles et de modéliser la relation entre X et Y séparément dans chaque intervalle. On parle d'ajustement local parce que la valeur de Y prédite par la modélisation en un point x_0 dépend des valeurs de Y dans l'intervalle où se trouve x_0, mais pas des autres. En réalité, certaines de ces méthodes tiennent aussi compte de valeurs de Y en dehors de l'intervalle, mais avec un poids faible ou de façon indirecte.
Ces méthodes d'ajustement local n'aboutissent pas à une représentation globale de l'ensemble de la relation entre X et Y par une fonction. Parfois, aucune équation de fonction n'est disponible, parfois plusieurs fonctions sont nécessaires. Les principales méthodes sont les suivantes :

- ✓ Modélisation par une fonction en escalier (§ VI). C'est un autre nom de la catégorisation présentée au § III. Elle généralise l'utilisation de variables indicatrices pour la modélisation des variables qualitatives ordinales (chapitre 3, § IV). C'est une méthode « purement » locale : les valeurs de Y extérieures à un intervalle n'interviennent pas dans la modélisation au sein de cet intervalle. Malgré ses inconvénients soulignés au § III, c'est une méthode utilisée très fréquemment.
- ✓ Régressions locales pondérées (lowess). Il s'agit d'une méthode descriptive qui généralise celle des moyennes mobiles (Wikipédia, 2023) qui sera présentée plus en détail plus bas.
- ✓ Modélisation par fonctions splines (§ IX). Sa place dans les méthodes d'ajustement local est discutable. Si elle est caractérisée par la modélisation de la relation entre X et Y au sein de chaque intervalle par un polynôme, les contraintes sur ces polynômes sont telles que l'ensemble des valeurs de Y interviennent dans le résultat final.

Les **avantages** des méthodes d'ajustement local sont la flexibilité (capacité à représenter de façon satisfaisante toutes les sortes de courbes, notamment celles ayant plusieurs variations de pente) et une bonne adaptation aux données observées.
Les **inconvénients** sont liés au fait qu'il n'y a pas d'approche reconnue pour sélectionner le meilleur modèle, notamment en ce qui concerne le choix du nombre et de la limite des intervalles et, pour les fonctions splines, pour le degré du polynôme au sein de chaque intervalle. Il y a, de plus, des risques de surajustement si, pour coller aux données, on prend des intervalles trop petits ; c'est le pendant de la flexibilité.

Régressions locales pondérées
Quelques lignes pour présenter la méthode de régression locale pondérée. Pour la culture générale et pour exposer une méthode descriptive qui peut être utile.
Il s'agit d'une méthode d'ajustement local des valeurs observées proposée initialement par Cleveland (Cleveland W, 1979, Cleveland W et al., 1988) et appelée lowess (*locally weighted scatterplot smoothing*).

Le principe général est le suivant. Pour représenter graphiquement la relation entre deux variables quantitatives X et Y à partir d'observations (x_i, y_i) $i = 1, \ldots, N$, chaque valeur y_i est remplacée par une valeur « lissée » y_i^s calculée de la façon suivante :

- ✓ on retient les observations (x_k, y_k) pour lesquelles x_k est « proche » de x_i
- ✓ on calcule la régression linéaire de Y en X pour ces observations en les pondérant selon la distance entre x_k et x_i (les plus éloignées ayant une pondération plus petite)
- ✓ y_i^s est la valeur prédite par cette régression en x_i.

Lorsque Y est une variable dichotomique (0/1), le principe est le même, de sorte qu'on obtient une valeur y_i^s comprise en 0 et 1, qui est une estimation du pourcentage de succès à X fixé et qu'on peut donc transformer en logit si besoin.

Cette méthode nécessite de préciser ce qu'on considère comme les observations proches de x_i (celles qui doivent être retenues pour estimer y_i^s). Les logiciels ont pour cela une option « bandwith » qui doit être comprise entre 0 et 1 et qui représente le pourcentage de l'ensemble des observations qui sont retenues. Par exemple, bandwith = 0.8 veut dire qu'on prend les x_k les plus proches de x_i jusqu'à ce que cela représente 80 % de l'ensemble de l'échantillon. Plus la valeur de bandwith est petite, plus la courbe obtenue est proche des observations (et moins le lissage est « efficace »). La Figure 4.8 montre, par exemple, le résultat obtenu avec les valeurs 0,8 et 0,1. Même si le deuxième cas est clairement plus proche des données observées, on comprend bien qu'il modélise à la fois la relation entre X et Y et les fluctuations d'échantillonnage. Le premier choix semble donc préférable.

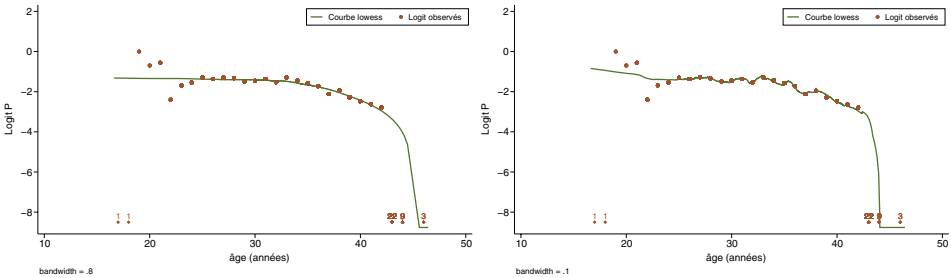


Figure 4.8 : Représentation de la relation entre l'âge et le succès en FIV (voir § V) par lowess avec deux choix de la valeur de bandwith.

> Remarque : Dans les deux cas, la forme de la courbe au-delà de 45 ans doit être prise avec réserves. Elle est en grande partie due au fait que la fonction « lowess » de Stata remplace les valeurs de logit P inférieures à 0,0001 par 1/N. C'est notamment le cas pour les 34 valeurs égales à $-\infty$. Le module lowess de R ne fait pas ce choix.

On peut sophistiquer la méthode en remplaçant la régression linéaire par une régression avec un polynôme de degré p (Gutierrez R et al., 2003). La commande correspondante est locpoly dans Stata et dans R.

L'avantage majeur de la méthode par régressions locales pondérées est d'être une méthode non paramétrique. Il n'y a besoin de spécifier ni formule ni forme de fonction pour représenter la relation entre X et Y.
L'inconvénient principal de lowess est de ne pas être réellement une modélisation et de ne pas fournir une fonction de régression de Y en fonction de X qui puisse être utilisée sur d'autres données ou même de tenir compte de covariables sur les données d'origine. Lowess demande par ailleurs des échantillons de taille suffisamment grande (ce qui est le cas de notre exemple).
Au total, en raison notamment de sa simplicité, cela fait de cette méthode un outil surtout utile d'un point de vue descriptif.

IV.2. Ajustement global

Dans les méthodes d'ajustement global, l'ensemble des valeurs de X et Y sont prises en compte pour prédire la valeur de Y en x_0 et aboutir à une modélisation par une fonction unique de X. Il s'agit de représenter globalement la courbe par une seule fonction de X.

La modélisation par une droite en est un exemple (voir § IV.3). Les autres méthodes, que nous détaillerons dans les paragraphes suivants, sont principalement :

- ✓ la modélisation par des polynômes de degré supérieur à 1, qui est une extension « naturelle » du modèle linéaire.
- ✓ la modélisation par polynômes fractionnaires (§ VIII). C'est une extension de la méthode précédente qui l'assouplit en autorisant que les degrés du polynôme ne soient pas des nombres entiers positifs.
- ✓ la modélisation par fonctions splines (§ IX). J'en ai déjà parlé dans les méthodes d'ajustement local, car on peut hésiter sur son classement entre ajustement local et global.

Les **avantages** principaux des méthodes d'ajustement global sont leur meilleure précision (bande de confiance plus étroite autour de la courbe) et leur capacité à donner une courbe « lisse » qui tient compte de l'ensemble des observations et qui, pour ces raisons, a plus de chances d'être proche de la vraie relation entre X et Y. Ce lissage de la relation correspond à « l'esprit » de la méthode statistique, qui ne vise pas à supprimer la variabilité individuelle ni les fluctuations d'échantillonnage qui sont indissociablement liées aux sciences de la vie, mais « à dépasser le désordre apparent qui en résulte au niveau individuel, d'une part en donnant des résultats moyens qui sont autant de points de repère pour la pratique du clinicien ou du biologiste, ou pour le chercheur, et d'autre part en mettant en évidence des phénomènes réguliers et stables étayant l'existence de lois biologiques générales » (Bouyer J, 2017).

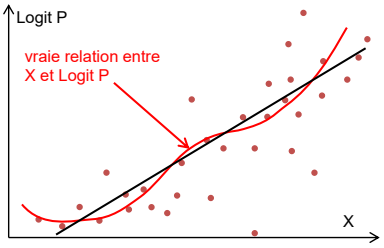
Leurs **inconvénients** sont leur manque de flexibilité (surtout vrai pour l'utilisation de polynômes, nettement moins pour les polynômes fractionnaires et les fonctions splines) et la possibilité d'artefacts, c'est-à-dire de formes de courbes s'écartant de

la « vraie » courbe pour s'adapter de façon « abusive » aux données observées pour les valeurs extrêmes de X.

IV.3. Modèle linéaire

Le modèle linéaire, qui s'écrit logit $P = \alpha + \beta X$, a une place à part dans la recherche d'une modélisation de la relation entre X et Y en raison de sa simplicité et du fait qu'il permet de résumer l'association entre X et Y avec un seul paramètre, facile à interpréter, la pente β de la régression.

C'est souvent, de plus, une représentation tout à fait satisfaisante de la relation entre X et Y, même si la linéarité n'est pas strictement respectée (Brenner H et al., 1997, Royston P et al., 2008).



La figure ci-contre, qui reprend l'exemple fictif du début de ce § IV, montre que la vraie relation, dont j'ai dit qu'elle avait une forme de J avec un palier pour les valeurs intermédiaires de X, peut raisonnablement être représentée par une droite qui donne la tendance générale.

Par ailleurs, Brenner et al. (Brenner H et al., 1997) ont montré que la modélisation linéaire d'un facteur de confusion quantitatif X_2 était préférable à la catégorisation de X_2 dans pratiquement tous les cas, y compris si la relation entre X_2 et Y avait une forme de U. Et aussi, de façon plus précise, que le contrôle de la confusion résiduelle est mauvais lorsque X_2 est catégorisé, alors qu'il est très satisfaisant avec une modélisation linéaire, et, bien sûr, encore meilleur avec une modélisation par fonctions splines ou avec des polynômes fractionnaires.

Pour l'ensemble de ces raisons, le modèle linéaire garde une place privilégiée dans le choix de la modélisation. C'est le modèle « de référence », à retenir, même si la linéarité n'est pas strictement respectée.

Les méthodes de modélisation doivent ainsi toujours être comparées au modèle linéaire et n'être retenues que s'il y a suffisamment d'évidence de ce qu'elles sont meilleures…

V. Données d'exemple

Pour présenter les méthodes de ce chapitre, je m'appuierai sur quelques variables issues de données de tentatives de fécondation in vitro (FIV) réalisées entre 1998 et 2002 dans deux centres français d'assistance médicale à la procréation (voir chapitre 1, § VI.2).

Les données portent sur 6 400 tentatives parmi lesquelles il y a eu 1 070 succès (Y = accouchement), soit 16,7 %. Les variables X sont les suivantes :

- ✓ X : âge de la femme (en années)
- ✓ X_1 : nombre d'ovocytes prélevés

✓ X_2 : nombre d'embryons de bonne qualité obtenus (c'est-à-dire des embryons dont les cliniciens considèrent qu'ils peuvent être réimplantés ou congelés).

Je vais tout d'abord m'intéresser à la modélisation du taux de succès P en fonction de l'âge X, avant de prendre en considération les autres variables comme facteurs de risque concomitants ou comme facteurs de confusion dans les paragraphes consacrés à la stratégie de choix du modèle (§ VIII.5 et IX.7).

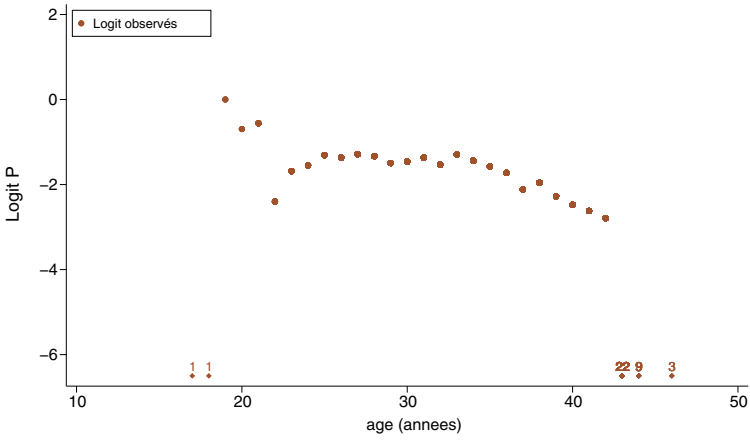


Figure 4.9 : Représentation des logit P observés selon l'âge de la femme en années.

La Figure 4.9 donne la représentation graphique des observations (logit P en fonction de X) en utilisant la méthode indiquée à la fin du § II qui fait apparaître les situations où P = 0 (pas de succès) comme des losanges en bas de la figure, surmontés de la valeur du nombre de sujets concernés. Ce sont ces observations qu'il s'agit de modéliser.

VI. Modélisation avec une fonction en escalier

Le principe général est de représenter la relation entre logit P et X par une courbe en escalier (en noir sur la Figure 4.10 ci-contre, qui reprend l'exemple fictif du début du § IV), qui suit au mieux les points observés. Mais qui peut s'écarter de la vraie courbe en raison des fluctuations d'échantillonnage.

La modélisation avec une fonction en escalier revient à considérer que logit P est constant au sein de chacun des intervalles de X qui ont été définis a priori. C'est donc la même chose que ce que j'ai appelé la catégorisation de X dans le § III. C'est une méthode très couramment utilisée en pratique, en particulier au début de l'analyse de la relation entre X et Y. Elle permet une première description quantitative de l'association entre X et Y, qui va guider les étapes

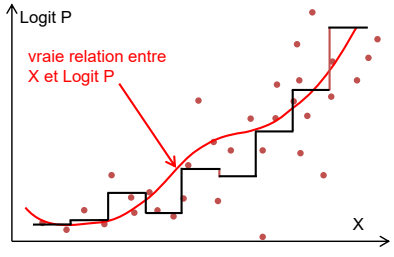


Figure 4.10 : Représentation de la relation entre X et Y par une fonction en escalier.

suivantes, souvent réalisées, il est vrai, avec moins de classes que dans l'exemple de la Figure 4.10.

Plus les classes sont étroites, plus l'escalier est proche des valeurs observées. La méthode est non paramétrique au sens où aucune hypothèse a priori n'est faite sur la forme de la relation entre X et Y, l'escalier se contentant de suivre l'évolution des observations faites sur l'échantillon. Le revers de la médaille est une moins bonne précision sur l'estimation de la hauteur de chaque marche de l'escalier (et des odds ratios correspondants) liée au petit nombre de sujets dans chaque classe et aux hasards d'échantillonnage. Comme je l'ai déjà indiqué plus haut, une trop grande adaptation aux données observées revient alors à « modéliser » les fluctuations d'échantillonnage, et surtout à chercher à leur donner un sens en essayant, par exemple, d'interpréter les quelques marches descendantes de la Figure 4.10, alors que la tendance générale va dans le sens d'une augmentation de Y avec X.

Pour illustrer et développer ces propos généraux, je vais prendre l'exemple de l'âge et du succès en FIV avec les données décrites dans le § V. La variable âge initiale, notée X, est continue. C'est le nombre de jours entre la naissance et la tentative de FIV qu'on a l'habitude de diviser par 365,25 pour l'exprimer en années. L'analyse de son association avec le succès de la FIV par un modèle linéaire aboutit au résultat suivant :

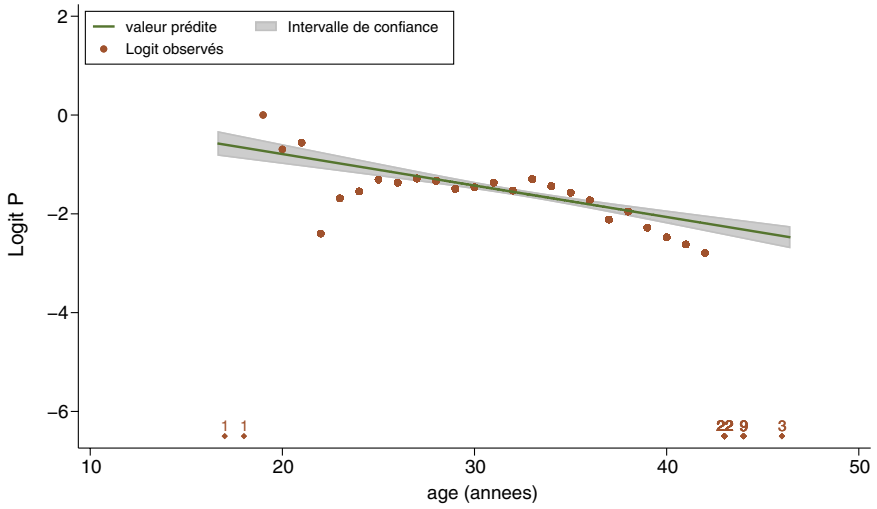


Figure 4.11 : Représentation de la relation entre l'âge et le pourcentage P de succès en FIV par un modèle linéaire.

Modèle 1 : logit $P = \alpha + \beta X = 0{,}485 - 0{,}0637\ X$ avec $\ln V = -2855{,}03$
La vraisemblance V (ou du moins son logarithme) est donnée en vue de comparaisons entre différents modèles. Sa valeur numérique n'a pas de sens en elle-même.
La courbe de la relation entre X et logit P est donnée par la Figure 4.11.

Pour modéliser la relation entre X et Y par une fonction en escalier, il faut catégoriser X. Pour cela, X est remplacé par une variable X' en classes d'un an, l'âge en années révolues (c'est-à-dire tronqué à l'année inférieure, comme on le fait dans la vie courante). Les effectifs par classe sont indiqués dans le Tableau 4.1.a.
On voit que, pour les deux plus petites catégories d'âge et les trois plus grandes, il n'y a pas de succès. On a donc P = 0, de sorte que logit P ne peut pas être calculé. Si on garde les classes sous cette forme, l'analyse avec des variables indicatrices exclut les classes correspondantes (et les sujets qui les composent). L'analyse serait donc faite sur un plus petit nombre total de sujets (6 364 au lieu de 6 400), ce qui n'est pas très grave, mais surtout en excluant des catégories avec aucun succès, ce qui est problématique pour l'interprétation des résultats.

Remarque : Stata signale ce fait par un message d'une clarté moyenne, mais somme toute compréhensible :

```
. logit acc i.agea
note: 17.agea != 0 predicts failure perfectly
      17.agea dropped and 1 obs not used
(...)
note: 46.agea != 0 predicts failure perfectly
      46.agea dropped and 3 obs not used
```

« predicts failure perfectly » se comprend en effet puisque tous les sujets de ces classes ont un échec. Savoir qu'un sujet est dans une de ces classes permet donc de prédire « parfaitement » qu'il a un échec…

R ne signale rien… mais donne des estimations et des intervalles de confiance aberrants pour les coefficients correspondant à ces classes, mais exacts pour les autres.

Si on veut garder tous les sujets, il faut effectuer des regroupements. En en faisant le moins possible, on obtient la variable X", dont la distribution est indiquée dans le Tableau 4.1.b. C'est ce choix qui est fait dans la suite pour la modélisation avec une fonction en escalier, de façon à pouvoir comparer les différents modèles entre eux sur le même échantillon.

Il y a donc deux modèles pour modéliser la relation entre âge et succès en FIV avec une fonction en escalier :

- ✓ Modèle A : logit $P = \alpha_A + \beta_A X"$. C'est en réalité un modèle linéaire semblable au modèle 1 vu précédemment, mais cette fois avec une variable âge en classes d'un an (arrondies à l'année entière) au lieu d'une variable continue où l'âge n'est pas un entier.
- ✓ Modèle B : Logit $P = \alpha'' + \sum_{i=1}^{23} \beta_i X_i''$ où l'âge, en classes d'un an avec les regroupements indiqués plus haut (variable X"), est décomposé en variables indicatrices X_i'', avec i = 1, …, 23.

a. Variable X'

Classe d'âge	Échec	Accouchement
17	1	0
18	1	0
19	1	1
20	2	1
21	7	4
22	11	1
23	27	5
24	47	10
25	74	20
26	118	30
27	177	49
28	265	70
29	325	73
30	405	94
31	404	103
32	448	97
33	446	122
34	408	97
35	400	83
36	354	63
37	323	39
38	324	46
39	254	26
40	225	19
41	151	11
42	98	6
43	22	0
44	9	0
46	3	0
Total	5330	1070

b. Variable X"

Classe d'âge	Échec	Accouchement
18	3	1
20	2	1
21	7	4
22	11	1
23	27	5
24	47	10
25	74	20
26	118	30
27	177	49
28	265	70
29	325	73
30	405	94
31	404	103
32	448	97
33	446	122
34	408	97
35	400	83
36	354	63
37	323	39
38	324	46
39	254	26
40	225	19
41	151	11
44	132	6
Total	5330	1070

Tableau 4.1 : Histogramme de la variable âge en classes d'un an (variable X') et après regroupement de classes (variable X")

En ajoutant le modèle linéaire 1, les paramètres des trois modèles que nous avons vus, estimés par la méthode du maximum de vraisemblance, sont les suivants :

Modèle 1 : $\text{logit } P = \alpha + \beta X = 0{,}485 - 0{,}0637\ X$ $\ln V = -2855{,}03$

Modèle A : $\text{logit } P = \alpha_A + \beta_A X'' = 0{,}502 + 0{,}0642\ X''$ $\ln V = -2854{,}72$

Modèle B : $\text{logit } P = \alpha'' + \sum_{i=1}^{23} \beta_i X''_i$ $\ln V = -2832{,}54$

En prenant la classe 30 ans comme référence, les coefficients estimés sont les suivants :

$\alpha'' = -1{,}461$; $\beta_1 = 0{,}362$; $\beta_2 = 0{,}767$; $\beta_3 = 0{,}901$; $\beta_4 = -0{,}937$; $\beta_5 = -0{,}226$; $\beta_6 = -0{,}087$; $\beta_7 = 0{,}152$; $\beta_8 = 0{,}091$; $\beta_9 = 0{,}176$; $\beta_{10} = 0{,}129$; $\beta_{11} = -0{,}033$; $\beta_{12} = 0{,}094$; $\beta_{13} = -0{,}069$; $\beta_{14} = 0{,}164$; $\beta_{15} = 0{,}024$; $\beta_{16} = -0{,}112$; $\beta_{17} = -0{,}266$; $\beta_{18} = -0{,}653$; $\beta_{19} = -0{,}492$; $\beta_{20} = -0{,}819$; $\beta_{21} = -1{,}011$; $\beta_{22} = -1{,}159$; $\beta_{23} = -1{,}630$.

Les trois modèles sont représentés sur la Figure 4.12 (sans les intervalles de confiance qui la rendraient illisible).

On remarque que les modèles 1 et A sont très proches l'un de l'autre, aussi bien pour l'estimation de leurs coefficients que pour leur vraisemblance, ce qui est normal car X et X" sont quasiment les mêmes variables, à l'arrondi de l'âge et au regroupement près. Ces deux modèles ne sont pas cependant pas comparables par un test statistique comme celui du rapport des vraisemblances, car ils ne sont pas emboîtés. Pour la même raison, les modèles 1 et B ne sont pas non plus comparables par un test.

> Remarque
> Il faut prêter attention au fait que les logiciels ne signalent pas que les modèles ne sont pas emboîtés. Stata, par exemple, se contente de mentionner « Assumption: A nested within B ». C'est à l'utilisateur de vérifier que cette hypothèse est vraie !

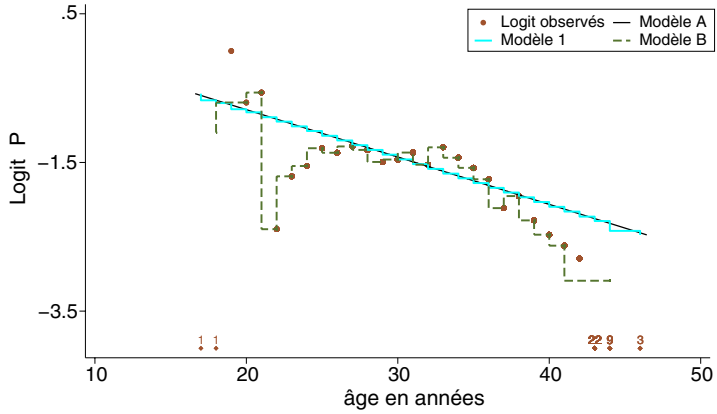


Figure 4.12 : Représentation graphique des modèles 1, A et B. L'âge est la variable X pour le modèle 1 et la variable X' pour les modèles A et B.

On peut en revanche comparer les modèles A et B (du moins parce qu'on a fait les regroupements conduisant à la variable X" et qu'ils portent donc sur le même nombre de sujets). Comme le modèle A est un proxy du modèle linéaire (c'est pour cela qu'il

a été introduit ici), sa comparaison avec le modèle B revient en réalité à un test de linéarité.

Le test du rapport des vraisemblances montre que les modèles A et B sont significativement différents :

$$\chi_0^2 = 2\ln\frac{V_B}{V_A} = 2\ln(V_B) - 2\ln(V_A) = 2(-2832{,}54 - (-2854{,}72)) = 44{,}36 \text{ à } 22 \text{ ddl, } p < 0{,}01$$

En conclusion, le modèle B est meilleur sur le plan statistique que le modèle A. On rejette la linéarité de la relation entre X et Y et on retient le modèle B.

En pratique, les résultats du modèle B permettent d'écarter la linéarité de la relation entre X et Y, mais ils sont peu lisibles, difficiles à « communiquer » et aussi (et peut-être surtout) difficiles à interpréter. Ils demandent en effet de donner un tableau de résultats avec les 23 coefficients ou les 23 odds ratios et leur intervalle de confiance, comme indiqué dans le Tableau 4.2, où la classe 30 ans a été prise comme référence, ce qui est très peu synthétique. On n'imagine pas ces résultats dans le tableau d'un article. On a, de plus, bien du mal à justifier la brusque variation à 22 ans, même si l'intervalle de confiance est très large.

On retrouvera ce problème dans la communication des résultats d'une modélisation de façon quasiment constante, en particulier avec les polynômes fractionnaires et les fonctions splines. Dans le cas présent, on préfèrera souvent considérer des classes plus larges, ce qui en diminue le nombre, et donc l'adéquation aux données de la modélisation, mais permet de limiter les « à-coups » (fluctuations d'échantillonnage) dus aux petits effectifs dans chaque classe. Le modèle B n'aura alors servi qu'à tester la linéarité.

Age (X")	OR	Age (X")	OR
18	1,44 [0,15 ; 13,96]	31	1,10 [0,80 ; 1,50]
20	2,15 [0,19 ; 24,01]	32	0,93 [0,68 ; 1,28]
21	2,46 [0,71 ; 8,58]	33	1,18 [0,87 ; 1,59]
22	0,39 [0,05 ; 3,07]	34	1,02 [0,75 ; 1,40]
23	0,80 [0,30 ; 2,13]	35	0,89 [0,65 ; 1,24]
24	0,92 [0,45 ; 1,88]	36	0,77 [0,54 ; 1,09]
25	1,16 [0,68 ; 2,00]	37	0,52 [0,35 ; 0,78]
26	1,10 [0,69 ; 1,73]	38	0,61 [0,42 ; 0,90]
27	1,19 [0,81 ; 1,76]	39	0,44 [0,28 ; 0,70]
28	1,14 [0,81 ; 1,61]	40	0,36 [0,22 ; 0,61]
29	0,97 [0,69 ; 1,36]	41	0,31 [0,16 ; 0,60]
30	1	44	0,20 [0,08 ; 0,46]

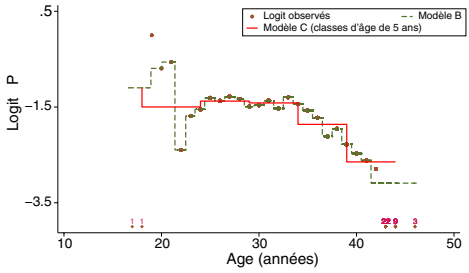
Tableau 4.2 : Odds ratios (et intervalles de confiance) de succès associés aux classes d'âge d'un an (avec 30 ans pris comme catégorie de référence)

Par exemple, avec des classes de cinq ans, et les classes [25–29] prises comme références (modèle C), on obtient les odds ratios du Tableau 4.3. Le graphique correspondant est donné sur la Figure 4.13.

Tableau 4.3 : Odds ratios (et intervalles de confiance) de succès associé à l'âge en classes de 5 ans.

Age	OR
≤ 19	1,32 [0,14-12,8]
20-24	0,89 [0,54-1,45]
25-29	1
30-34	0,96 [0,81-1,14]
35-39	0,61 [0,51-0,75]
≥ 40	0,28 [0,19-0,41]

Figure 4.13 : Graphique permettant de comparer les modélisations avec l'âge en classes de 1 et 5 ans.



Remarque
Le modèle C est emboîté dans le modèle B (et on peut vérifier qu'ils ne sont pas significativement différents), mais les modèles A et C ne sont pas emboîtés. On ne peut donc pas tester la linéarité en comparant les modèles A et C.

VII. Modélisation avec des polynômes

L'idée « naturelle » pour une modélisation s'écartant de la linéarité tout en respectant la nature quantitative de X est d'étendre la fonction linéaire en ajoutant des puissances successives de X. On obtient ainsi une modélisation avec un polynôme de degré m qui s'écrit : $\text{logit } P = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_m X^m$.

Cette modélisation présente l'avantage de la simplicité et aussi celui de permettre un test simple d'écart à la linéarité en testant l'hypothèse $H_0 : \beta_2 = \ldots = \beta_m = 0$ par la méthode du rapport des vraisemblances ou par un test de Wald global de l'ensemble de ces coefficients. C'est par ailleurs une modélisation rencontrée assez fréquemment en pratique avec $m = 2$ ou $m = 3$.

En modélisant l'âge avec un polynôme de degré 3, on obtient ainsi la courbe représentée sur la Figure 4.14a, plus satisfaisante qu'une droite ou qu'une fonction en escalier. L'équation de cette courbe est $\text{logit}P = 8{,}978 - 1{,}177\ X + 0{,}044\ X^2 - 0{,}0005\ X^3$. Le test global des deux derniers coefficients est significatif, ce qui confirme que la relation entre X et logit P s'écarte significativement de la linéarité.

L'inconvénient principal de la modélisation polynômiale est son manque de souplesse. D'une part, les polynômes ne peuvent pas prétendre représenter certaines formes de relation entre Y et X, en particulier celles qui tendent vers un plateau ou une droite oblique (asymptote) lorsque x tend vers $+\infty$ ou $-\infty$. D'autre part, l'augmentation du nombre de degrés du polynôme pour obtenir une meilleure adéquation

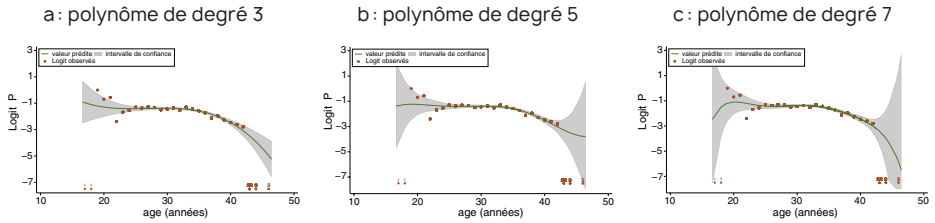


Figure 4.14 : Modélisation de la relation entre l'âge et le succès avec un polynôme

avec les observations peut conduire à des artefacts, c'est-à-dire à des formes de relation entre X et Y qui s'adaptent aux fluctuations aléatoires des observations et ne peuvent que s'éloigner ainsi de la réalité. On constate ce phénomène avec un polynôme de degré 5 (Figure 4.14b) et encore plus avec un polynôme de degré 7 (Figure 4.14c). Les « points d'inflexion » (ondulations) dans la partie centrale de la courbe de la Figure 4.14c ont peu de chance de correspondre à la réalité. On note surtout l'augmentation considérable de l'imprécision de la courbe (augmentation de la largeur de l'intervalle de confiance) pour les valeurs extrêmes de X lorsque le degré du polynôme augmente. Cela incite à la prudence dans l'interprétation des résultats, bien sûr pour ces valeurs de X, mais aussi plus globalement puisque cela peut aller jusqu'à une incapacité à reconnaître l'écart à la linéarité (le test d'écart à la linéarité avec le polynôme de degré 7 est non significatif).

Ici, le polynôme de degré 3 apparaît comme le meilleur compromis.

VIII. Modélisation avec des polynômes fractionnaires

Les polynômes fractionnaires sont une extension des polynômes ordinaires du § VII obtenue en autorisant les exposants de X à être négatifs et/ou non entiers (ce ne sont donc plus des polynômes à proprement parler). Ils ont été introduits dans les années 1990 et largement développés et popularisés par P. Royston, W. Sauerbrei et D. Altman (Sauerbrei W et al., 2010), qui en ont fait un ensemble susceptible non seulement de modéliser la relation entre Y et une ou plusieurs variables quantitatives X simultanément, mais aussi de sélectionner les variables à inclure dans un modèle et leur meilleure modélisation. On en trouve une présentation synthétique et pédagogique dans l'article de Royston et al. de 1999 (Royston P et al., 1999), plus théorique dans Royston et al. 1994 (Royston P et al., 1994) ou dans Sauerbrei et al. 1999 (Sauerbrei W et al., 1999). Pour un développement plus complet, qui inclut l'ensemble des questions que pose la modélisation en épidémiologie quantitative, voir l'excellent livre de Royston et Sauerbrei (Royston P et al., 2008).

VIII.1. Définition et écriture d'un polynôme fractionnaire

Les exposants d'un polynôme fractionnaire sont choisis dans un ensemble prédéfini de huit valeurs : $S = [-2; -1; -0{,}5; 0; 0{,}5; 1; 2; 3]$ où, par convention, la valeur 0

correspond à ln X. Il est possible de « répéter » une puissance, comme on le verra plus loin. L'ensemble S peut paraître restreint, mais l'expérience montre qu'il est suffisant pour la quasi-totalité des situations rencontrées en pratique (Royston P et al., 2008).
Un polynôme fractionnaire de degré (ou d'ordre) d, noté PF_d (ou FP_d pour *fractional polynomial* en anglais), comprend m termes et s'écrit donc $FP_d = \beta_0 + \sum_{j=1}^{d} \beta_j x^{(p_j)}$ où $p_j \in S$, avec les conventions suivantes :

- ✓ $x^{(p_j)} = x^{p_j}$ si $p_j \neq 0$
- ✓ $x^{(0)} = \ln(x)$
- ✓ si (p_j) est répété m fois, le premier terme est x^{p_j} et les suivants sont $x^{p_j} (\ln(x))^{i-1}$ pour i = 2, …, m.

Par exemple, le polynôme fractionnaire de degré 3 qui s'écrit $PF_3 = \beta_0 + \beta_1 x^{(0)} + \beta_2 x^{(0,5)} + \beta_3 x^{(0,5)}$ est égal à $PF_3 = \beta_0 + \beta_1 \ln x^{(0)} + \beta_2 \sqrt{x} + \beta_3 \sqrt{x} \ln x$.
De façon conventionnelle, un polynôme fractionnaire est noté avec la suite des d puissances p_j. Le PF_3 précédent est donc noté : (0 0,5 0,5).

Il faut préciser que les PF supposent que X > 0 pour être définis, ce qui est un de leurs point faibles. Si ce n'est pas le cas, les auteurs recommandent de remplacer X par X' = X + a, où a est choisi pour que X' soit strictement positif (voir § XI).

Il y a au total huit PF_1 et 36 PF_2. En pratique, on constate qu'il n'est pas nécessaire de considérer des degrés supérieurs à 2 pour représenter de façon satisfaisante la relation entre Y et une variable continue X. On peut en effet obtenir la plupart des formes de fonctions avec des polynômes PF_2, ainsi que le montre la Figure 4.15.
On peut aussi remarquer, comme l'illustre la Figure 4.16, qu'il est possible d'obtenir des courbes différentes avec les mêmes puissances (ici –0,5 –0,5) et des coefficients différents. Cela rassure sur le fait que les puissances sélectionnées pour modéliser la relation brute entre X et Y pourront être conservées dans un modèle où d'autres variables X_i seront incluses, même si cette inclusion modifie la forme de la relation brute.

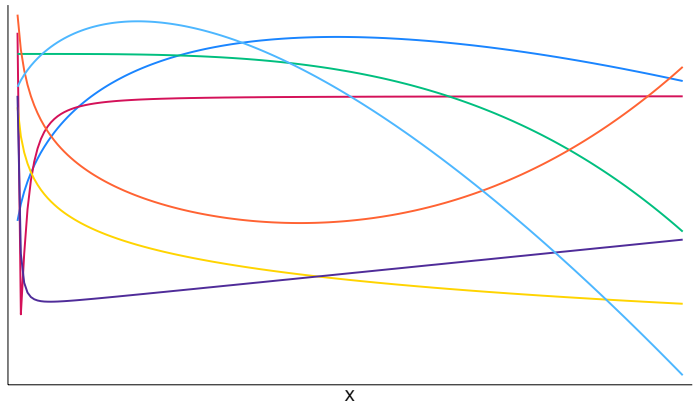


Figure 4.15 : Différentes formes de courbes obtenues avec des polynômes fractionnaires de degré 2.

Courbe en trait plein
Coefficients : 2 ; −1
Équation de la courbe : $2\sqrt{x} - \sqrt{x}\ln(x)$
Courbe pointillée
Coefficients : −2 ; 2
Équation de la courbe : $-2\sqrt{x} + 2\sqrt{x}\ln(x)$

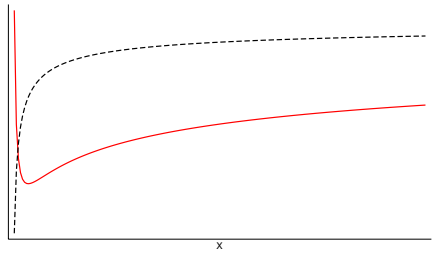


Figure 4.16 : Deux formes de courbes obtenues avec un polynôme fractionnaire de degré 2 et les puissances (−0,5 −0,5), mais des coefficients différents.

Sur le plan logiciel, les polynômes fractionnaires ont été développés initialement dans Stata par Patrick Royston, qui a écrit les premières commandes. Elles ont ensuite été réécrites et intégrées dans le cœur officiel du logiciel. Dans Stata il y a deux commandes principales : fp, qui permet de déterminer et de construire le meilleur polynôme fractionnaire de degré fixé, et mfp, dont je parlerai plus loin (§ VIII.3), qui permet de sélectionner le meilleur polynôme fractionnaire (degré et puissances) et surtout s'étend à plusieurs variables X. La commande mfp existe aussi en package R (et en macro SAS). Davantage de détails sur les aspects logiciels sont donnés en annexe.

VIII.2. Choix des puissances d'un polynôme fractionnaire de degré donné

Le choix du meilleur polynôme fractionnaire est bien sûr un des problèmes importants qui se posent pour la modélisation de la relation entre X avec Y (on retrouvera la même question avec les fonctions splines dans le § IX). Pour un PF de degré donné, la (ou les) puissance(s) retenue(s) est (sont) celle(s) qui condui(sen)t à la vraisemblance la plus grande (ou à la déviance la plus petite). Il n'y a cependant pas de test statistique formel pour le comparer aux autres PF de même degré, car ils ne sont pas emboîtés. Concernant le degré, on ne gagne pas toujours à l'augmenter, ainsi que je vais l'illustrer ci-dessous en montrant les avantages et les inconvénients de polynômes fractionnaires de degrés 1, 2 et 4 pour modéliser la relation entre l'âge et le succès dans l'exemple des données de FIV. Je finirai en présentant la stratégie générale proposée par P. Royston pour choisir degré et puissances.
Rappelons qu'il est très utile, voire indispensable, de ne pas s'en tenir à l'équation du PF, mais de tracer aussi la courbe de la relation, car il est quasi impossible de se rendre compte de son allure à partir de son équation.

VIII.2.a. Polynôme fractionnaire de degré 1

En se limitant à des PF de degré 1 pour modéliser la relation entre l'âge de la femme et le succès de la FIV, la puissance retenue est 3 (c'est celle pour laquelle la vraisemblance est maximum). Le modèle est le suivant : logit $P = -0{,}814 - 0{,}0000214 \times \text{âge}^3$. Le graphique correspondant est donné sur la Figure 4.17.

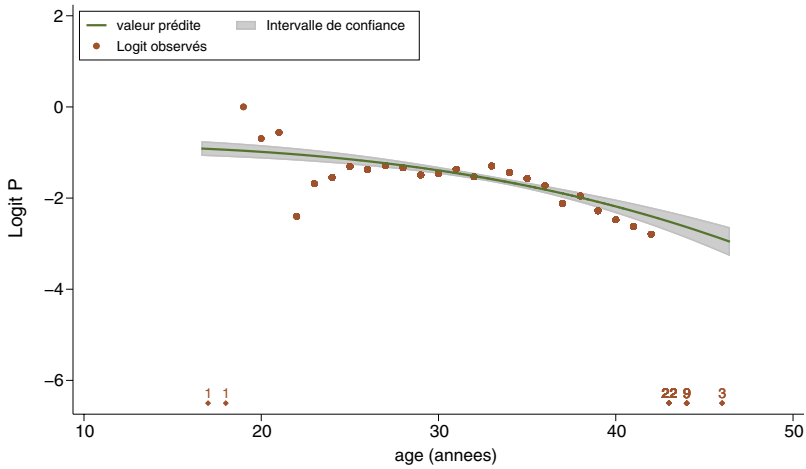


Figure 4.17 : Relation entre l'âge et le succès en FIV modélisée par un polynôme fractionnaire de degré 1.

Remarque
Il arrive que age soit remplacé par x = age/10, comme dans les premières versions de Stata et dans R. Cela se justifiait par limitations numériques qui sont dépassées pour les ordinateurs actuels et cela n'a rien d'obligatoire. Cela ne change ni la courbe ni la constante du modèle et permet juste d'avoir un coefficient multiplié par 1000 pour x^3, plus facile à écrire qu'avec age. Il faut cependant faire attention à ne pas se tromper dans les utilisations ultérieures du modèle en remplaçant x par x/10.

VIII.2.b. Polynôme fractionnaire de degré 2

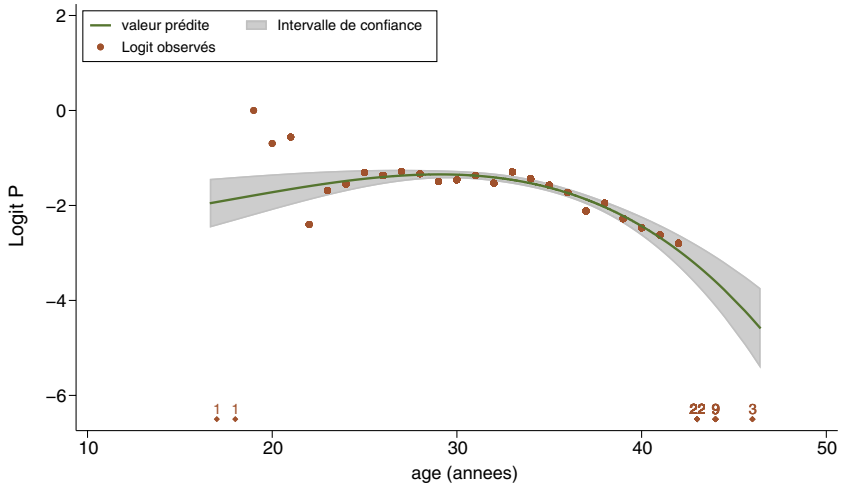


Figure 4.18 : Relation entre l'âge et le succès en FIV modélisée par un polynôme fractionnaire de degré 2.

En prenant des PF de degré 2, les puissances retenues sont 3 et 3. Cela signifie que le modèle est le suivant : logit P = – 2,56 + 0,000552 x^3 – 0,000149 x^3 ln(x). La figure correspondante est la Figure 4.18.

Le polynôme fractionnaire (3 3) est le meilleur des PF_2 dans le sens où il a la plus grande vraisemblance. Bien qu'il n'y ait pas de test statistique formel pour les comparer, on peut constater que d'autres PF_2, avec d'autres combinaisons de deux puissances, donnent des vraisemblances (ou des déviances) très proches (Tableau 4.4), et finalement des modélisations quasiment identiques en pratique.

Le modèle retenu est le modèle n° 44 (attention, les modèles ne sont pas tout à fait classés dans l'ordre des déviances décroissantes). On voit que les modèles 41 et 43 sont pratiquement aussi bons que le modèle 44 et ont des puissances « plus simples » (pas de log dans l'écriture du modèle). On peut donc choisir de les prendre à la place du modèle retenu. On constate d'ailleurs sur la Figure 4.19 qu'ils donnent quasiment la même modélisation.

Model#	Deviance	Power1	Power2
1	5732.988	-2.0	.
2	5724.741	-1.0	.
3	5720.735	-0.5	.
4	5716.837	0.0	.
5	5713.068	0.5	.
6	5709.445	1.0	.
7	5702.703	2.0	.
8	5696.717	3.0	.
9	5699.433	-2.0	-2.0
10	5695.801	-1.0	-2.0
11	5694.056	-0.5	-2.0
12	5692.373	0.0	-2.0
13	5690.760	0.5	-2.0
14	5689.226	1.0	-2.0
15	5686.416	2.0	-2.0
16	5683.983	3.0	-2.0
17	5692.616	-1.0	-1.0
18	5691.098	-0.5	-1.0
19	5689.641	0.0	-1.0
20	5688.252	0.5	-1.0
21	5686.938	1.0	-1.0
22	5684.555	2.0	-1.0
23	5682.517	3.0	-1.0
24	5689.693	-0.5	-0.5
25	5688.348	0.0	-0.5
26	5687.071	0.5	-0.5
27	5685.866	1.0	-0.5
28	5683.689	2.0	-0.5
29	5681.842	3.0	-0.5
30	5687.115	0.0	0.0
31	5685.946	0.5	0.0
32	5684.848	1.0	0.0
33	5682.874	2.0	0.0
34	5681.210	3.0	0.0
35	5684.885	0.5	0.5
36	5683.890	1.0	0.5
37	5682.111	2.0	0.5
38	5680.623	3.0	0.5
39	5682.995	1.0	1.0
40	5681.404	2.0	1.0
41	5680.084	3.0	1.0
42	5680.161	2.0	2.0
43	5679.150	3.0	2.0
44	5678.408	3.0	3.0

Tableau 4.4 : Déviance des 44 PF de degré 1 ou 2 pour modéliser la relation entre le succès et l'âge avec les différentes combinaisons de puissances (la déviance est égale à – 2 fois la vraisemblance)

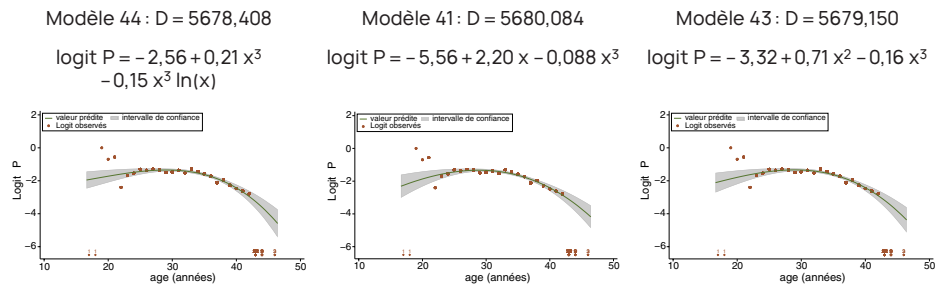


Figure 4.19 : Relation entre l'âge et le succès en FIV modélisée par des polynômes fractionnaires de degré 2 avec des puissances différentes et des déviances d proches (ici x = age/10).

VIII.2.c. Polynôme fractionnaire de degré 4

Les puissances retenues sont (– 2 – 2 – 2 – 1). Le modèle est donc le suivant :

$$\text{logit } P = -203{,}9 - 1079279\, x^{-2} + 887001\, x^{-2} \ln(x) - 258732\, x^{-2} \ln(x)^2 + 41259/x$$

Il est représenté sur la Figure 4.20. On peut noter que ce modèle colle effectivement mieux aux observations. Mais au prix de variations difficiles à interpréter. On constate une légère diminution de la probabilité de succès entre 20 et 25 ans, qui est peu explicable et qui reflète le faible taux de succès observé chez les femmes de 22 ans (voir Tableau 4.1), qui peut n'être qu'une fluctuation d'échantillonnage. Chez les plus jeunes, la forte diminution du taux de succès s'explique principalement par la présence de deux femmes de 17 et 18 ans ayant toutes connu un échec. C'est, là aussi, proche d'un phénomène aléatoire (ou d'un mécanisme de sélection des femmes qui ont besoin d'une FIV à cet âge). Il y a d'ailleurs une très forte imprécision de la courbe pour ces catégories les plus jeunes.
Lorsqu'on confronte les courbes (Figure 4.21), on voit que le gain en passant de FP_2 à FP_4 est modeste sur la majeure partie de la courbe. Il consiste essentiellement à mieux représenter les observations (ou les fluctuations d'échantillonnage…) pour les âges les plus jeunes, pour lesquels les effectifs sont faibles et la précision de la courbe FP_4 mauvaise.

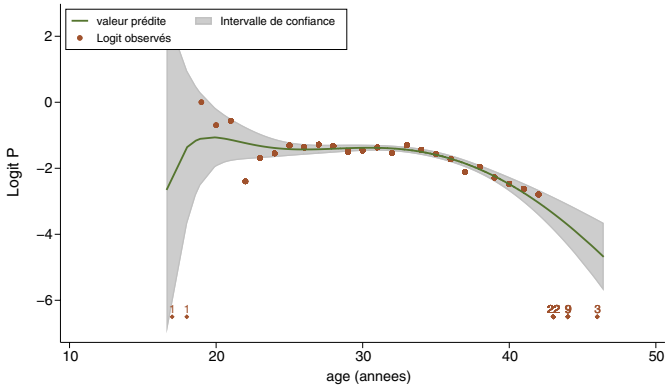


Figure 4.20 : Relation entre l'âge et le succès en FIV modélisée par un polynôme fractionnaire de degré 4.

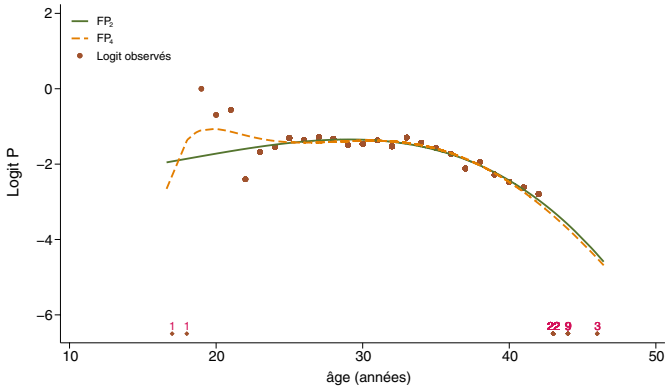


Figure 4.21 : Comparaison graphique des modélisations par des polynômes fractionnaires de degrés 2 et 4.

VIII.3. Choix du meilleur polynôme fractionnaire pour une variable

Le choix doit reposer sur une stratégie qui soit la plus « objective » possible pour déterminer le degré du polynôme, puis les puissances qu'on va retenir, et qui permette de contrôler le risque d'erreur de première espèce. C'est ce qu'ont proposé Royston et ses collègues, d'abord pour la modélisation d'une seule variable quantitative X, puis dans le cas de plusieurs variables X_1, X_2, …, X_p à modéliser simultanément (Ambler G et al., 2001, Royston P et al., 2005).

Les principes généraux de la stratégie de choix du polynôme fractionnaire avec une seule variable quantitative X sont, d'une part, de considérer qu'il est inutile de prendre des polynômes fractionnaires de degré supérieur à 2 et, d'autre part, de privilégier le modèle linéaire s'il n'est pas significativement moins bon qu'un autre. Le premier point, dont nous avons vu un exemple plus haut avec l'inutilité d'aller jusqu'au degré 4, a été conforté par l'expérience (Royston P et al., 2008).

Le processus se déroule en quatre étapes :

- ✓ Étape 1 : choisir le meilleur modèle FP_2.
- ✓ Étape 2 : comparer le modèle obtenu avec celui n'incluant pas la variable X.
  Si la différence est non significative, X n'est pas significativement lié à Y. Le processus de sélection s'arrête là et X n'est pas inclus dans l'analyse. Au-delà du choix du meilleur FP, ce processus est donc aussi un moyen de sélection des variables à inclure dans le modèle.
- ✓ Étape 3 : sinon, comparer le meilleur FP_2 avec le modèle linéaire. Si la différence est non significative, la linéarité n'est pas rejetée et le modèle linéaire retenu.
- ✓ Étape 4 : sinon, comparer le meilleur FP_2 et le meilleur FP_1. Si la différence est non significative, on retient le modèle FP_1, sinon le modèle FP_2. Cette étape est un moyen de simplifier le modèle final en retenant un FP_1 plutôt qu'un FP_2 si c'est possible.

Pour les étapes ci-dessus :

- ○ le meilleur polynôme fractionnaire FP_m est celui des FP_m qui a la plus grande vraisemblance ;
- ○ pour comparer deux polynômes fractionnaires de degrés différents, on ne peut pas utiliser le test du rapport des vraisemblances, car les modèles ne sont pas emboîtés (du moins pas toujours). On utilise le cependant le rapport des vraisemblances en considérant (cela a été vérifié à l'aide de simulations) qu'un modèle FP_m a 2m degrés de liberté (1 pour chaque coefficient, et 1 pour le choix de chaque puissance) et que le modèle linéaire a 1 degré de liberté. Le test de comparaison du meilleur FP_2 au modèle linéaire (étape 3 précédente) est ainsi un test de χ^2 à 3 degrés de liberté.

Il a été vérifié par simulation que cette stratégie permet de contrôler le risque d'erreur α global dans le choix du modèle (Ambler G et al., 2001).

Exemple

La stratégie précédente de choix du meilleur FP a été programmée dans Stata et dans R. Le nom de la commande est mfp. Elle est exécutée ci-dessous avec Stata sur l'exemple de la relation entre l'âge de la femme et le succès (accouchement) en FIV avec les données décrites précédemment (Tableau 4.5).

Ici, on retient le modèle FP_2 avec les puissances (3 3). C'est le même résultat que celui obtenu au § VIII.2.b en cherchant le meilleur FP_2, mais ici la procédure mfp permet en outre de rejeter la linéarité. Il faut cependant être attentif au fait que la variable age est ici remplacée par age/10, ce qui change les coefficients par rapport à ceux du § VIII.2.b (sans changer la modélisation).

```
. mfp, select(.05) center(no) : logit acc age

Deviance for model with all terms untransformed =  5710.067, 6400 observations

Variable     Model (vs.)   Deviance  Dev diff.   P      Powers   (vs.)
-------------------------------------------------------------------------
age          null   FP2    5777.901    99.924  0.000*   .         3 3
             lin.          5710.067    32.090  0.000+   1
             FP1           5697.090    19.113  0.000+   3
             Final         5677.977                     3 3

Transformations of covariates:
-> gen double Iage__1 = X^3 if e(sample)
-> gen double Iage__2 = X^3*ln(X) if e(sample)
   (where: X = age/10)

Final multivariable fractional polynomial model for acc
-------------------------------------------------------------------------
    Variable |     -----Initial-----          -----Final-----
             |   df     Select   Alpha    Status    df     Powers
-------------+-----------------------------------------------------------
         age |    4     0.0500   0.0500     in       4     3 3
-------------------------------------------------------------------------

Logistic regression                             Number of obs     =      6,400
                                                LR chi2(2)        =      99.92
                                                Prob > chi2       =     0.0000
Log likelihood = -2838.9885                     Pseudo R2         =     0.0173

------------------------------------------------------------------------------
         acc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Iage__1 |   .2085767   .0546516     3.82   0.000     .1014616    .3156918
     Iage__2 |  -.1489945   .0354207    -4.21   0.000    -.2184178   -.0795712
       _cons |  -2.563884   .4265015    -6.01   0.000    -3.399811   -1.727956
------------------------------------------------------------------------------
Deviance: 5677.977.
```

Tableau 4.5 : Procédure mfp pour le choix de la modélisation de la relation entre l'âge de la femme et le succès en FIV

VIII.4. Modélisation simultanée de plusieurs variables quantitatives, procédure mfp

La procédure mfp se généralise à la stratégie de choix du meilleur modèle lorsqu'on veut utiliser plusieurs variables quantitatives simultanément. C'est même là son objectif premier : mfp sont les initiales de *Multivariable Fractional Polynomials*.

En présence de plusieurs variables, les étapes du processus de sélection sont les suivantes :

- Si, parmi les variables X, certaines sont qualitatives, elles doivent être transformées en variables indicatrices, considérées comme un bloc et donc non modélisées.
- Détermination de l'ordre dans lequel les variables sont traitées : on considère pour cela les modèles logistiques linéaires successifs où les variables sont considérées une par une. Les variables sont prises dans l'ordre croissant du degré de signification p du test de leur coefficient (le plus petit p d'abord…).
- Choix du meilleur FP pour la première variable ou élimination de cette variable : c'est le même processus qu'avec une seule variable, sachant que les autres variables sont incluses dans le modèle sous forme linéaire.
- Même chose pour les variables suivantes en les ajustant sur celles qui ont déjà été traitées avec la modélisation qui leur a été attribuée, et en continuant à inclure celles qui n'ont pas été traitées sous forme linéaire.
- Fin du cycle 1 : lorsque toutes les variables ont été traitées, c'est la fin du cycle 1. On recommence « au début » en prenant les variables dans le même ordre, mais cette fois les variables sont incluses dans le modèle avec la modélisation obtenue au cours du cycle précédent (et non sous forme linéaire). On arrive ainsi à la fin du cycle 2.
- Modèle final : le processus s'arrête lorsque la modélisation des variables ne change plus entre deux cycles successifs (très souvent un petit nombre de cycles suffit, parfois deux).

Exemple

Poursuivons l'exemple du succès en FIV avec, cette fois-ci, les variables suivantes :

✓ Y : résultat d'une tentative de FIV (accouchement ou échec)

✓ X_1 : âge de la femme ; X_2 : nombre d'ovocytes prélevés ; X_3 : nombre d'embryons de bonne qualité obtenus.

La modélisation par mfp de la relation entre Y et la seule variable âge a déjà été donnée au paragraphe précédent. Pour chacun des deux autres variables X_i prises séparément, on obtient le résultat de la Figure 4.22.

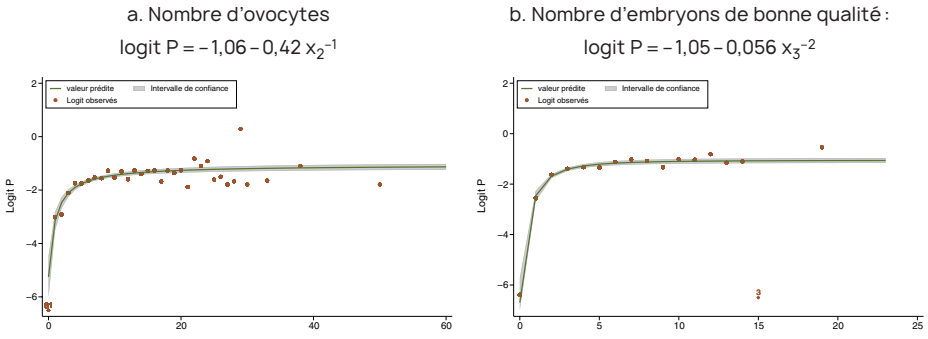


Figure 4.22 : Modélisation des relations entre Y et le nombre d'ovocytes et le nombre d'embryons de bonne qualité (pris comme des variables quantitatives).

```
. mfp, center(no) select(0.05) : logit acc age ovo emb
Deviance for model with all terms untransformed = 5576.866, 6366 observations

Variable     Model (vs.)   Deviance  Dev diff.   P      Powers   (vs.)
-------------------------------------------------------------------------
emb          null   FP2    5665.430   278.830  0.000*   .         -2 3
             Lin.          5576.866   190.266  0.000+   1
             FP1           5387.000     0.400  0.819    -2
             Final         5387.000                     -2

age          null   FP2    5432.668    74.335  0.000*   .         3 3
             Lin.          5387.000    28.667  0.000+   1
             FP1           5376.395    18.062  0.000+   3
             Final         5358.333                     3 3

ovo          null   FP2    5359.763     3.812  0.432    .         -.5 0
             Final         5359.763                     .
-------------------------------------------------------------------------
End of Cycle 1: Deviance =  5359.763

emb          null   FP2    5659.665   300.048  0.000*   .         -2 -.5
             Lin.          5547.598   187.981  0.000+   1
             FP1           5359.763     0.147  0.929    -2
             Final         5359.763                     -2

age          null   FP2    5432.741    72.977  0.000*   .         3 3
             Lin.          5388.550    28.787  0.000+   1
             FP1           5378.302    18.539  0.000+   3
             Final         5359.763                     3 3

ovo          null   FP2    5359.763     3.812  0.432    .         -.5 0
             Final         5359.763                     .

Fractional polynomial fitting algorithm converged after 2 cycles.

Transformations of covariates:

-> gen double Iage__1 = X^3 if e(sample)
-> gen double Iage__2 = X^3*ln(X) if e(sample)
   (where: X = age/10)
-> gen double Iemb__1 = X^-2 if e(sample)
   (where: X = (emb+1)/10)

Final multivariable fractional polynomial model for acc
-------------------------------------------------------------------------
    Variable |     -----Initial-----          -----Final-----
             |   df     Select   Alpha    Status    df     Powers
-------------+-----------------------------------------------------------
         age |    4     0.0500   0.0500     in      4     3 3
         ovo |    4     0.0500   0.0500     out     0
         emb |    4     0.0500   0.0500     in      2     -2
-------------------------------------------------------------------------

Logistic regression                                     Number of obs =  6,366
                                                        LR chi2(3)    = 399.26
                                                        Prob > chi2   = 0.0000
Log likelihood = -2679.8817                             Pseudo R2     = 0.0693
------------------------------------------------------------------------------
         acc | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     Iage__1 |    .211946   .0554319     3.82   0.000     .1033016    .3205905
     Iage__2 |  -.1489283   .0359064    -4.15   0.000    -.2193035   -.078553
     Iemb__1 |  -.0539564   .0053172   -10.15   0.000     -.064378   -.0435349
       _cons |   -2.16091   .4346931    -4.97   0.000    -3.012892   -1.308927
------------------------------------------------------------------------------
Deviance = 5359.763.
```

Tableau 4.6 : Détails de la procédure mfp pour le choix de la modélisation de la relation entre le succès en FIV et l'âge de la femme, le nombre d'ovocytes et le nombre d'embryons de bonne qualité

La commande pour sélectionner le meilleur modèle avec les variables X_1, X_2, X_3 est mfp, comme dans le cas avec une variable. Les résultats figurent dans le Tableau 4.6.

La convergence est obtenue au bout de deux cycles (en réalité, le résultat est acquis dès le premier cycle, mais on ne s'en rend compte qu'à la fin du deuxième).
La variable nombre d'ovocytes est exclue au cours du premier cycle, puisque sa modélisation par le meilleur FP_2 qui est (–0,5 0) n'est pas différente du modèle où elle est absente (p = 0,432).
La modélisation finale est la suivante :

$$\text{logit } P = -2{,}16 + 0{,}212\,X_1'^{3} - 0{,}148\,X_1'^{3}\ln\left(X_1'\right) - 0{,}0530\,X_2'^{-2},$$

$$\text{avec } X_1' = \text{age}/10 \text{ et } X_2' = (\text{emb}+1)/10.$$

La division par 10 est liée à de possibles problèmes numériques avec des grandes valeurs des variables. L'ajout de 1 à emb est dû au fait que certains sujets ont emb = 0, ce qui n'est pas admissible pour des polynômes factionnaires dont les puissances peuvent être négatives.

On voit donc que la procédure mfp est un moyen à la fois de choisir les puissances pour modéliser les variables et de sélectionner les variables qui doivent être incluses dans le modèle (ou en être exclues) – j'y reviendrai dans le chapitre 5, § V.2.

IX. Modélisation avec des fonctions splines

Le mot anglais *spline* désigne une latte flexible utilisée par les dessinateurs pour matérialiser des lignes à courbure variable et passant par des points fixés a priori ou « à proximité » de ceux-ci. Le tracé ainsi réalisé minimise l'énergie de déformation de la latte. Par analogie, ce mot désigne également des familles de fonctions permettant de représenter des courbes observées avec des propriétés « optimales » de régularité. L'idée originale date du début du XX^{e} siècle (Whittaker 1923), et elle connaît ses premières applications en statistique dans les années 1970 (tiré et adapté de Besse P et al. (1989)).

Le principe général des fonctions splines en analyse statistique des données est de remplacer la variable X par une combinaison linéaire de fonctions de X, dites fonctions splines de base, dont l'écriture, assez compliquée, est expliquée dans la suite. Les caractéristiques et contraintes portent sur la fonction spline globale. Ce sont les suivantes :

- On commence par définir (k + 1) intervalles pour la variable X en choisissant k valeurs numériques (appelées « nœuds »), qui définissent les bornes des intervalles. Il est souvent préférable de choisir ces nœuds en fonction de la distribution de X, ainsi qu'on le verra plus bas, mais cela n'a rien d'obligatoire.
- Au sein de chaque intervalle, la fonction spline globale doit être un polynôme de degré d. Le plus souvent, d = 3 (on parle alors de splines cubiques), mais on verra aussi le cas d = 1, splines linéaires ou fonctions linéaires par morceaux.

- La courbe de la fonction spline globale doit être la plus régulière possible, comme la latte flexible qui lui a donné son nom, c'est-à-dire continue (sans rupture) et « lisse », ou encore « sans angle »[1]. Sur le plan mathématique, cela se traduit par le fait que la fonction spline doit être continue, et dérivable (d – 1) fois.

Une des difficultés avec les fonctions splines est qu'il existe plusieurs façons de construire les fonctions de base tout en gardant la même fonction spline globale. Leur nombre lui-même peut varier selon les « modules splines » des logiciels, dont il n'est pas toujours facile de savoir ce qu'ils font.

D'un point de vue pratique, la fonction spline globale elle-même est construite comme une combinaison linéaire de fonctions de base déterminées à partir des nœuds choisis. Sa forme générale ne dépend donc pas de Y. Ce n'est que dans un second temps que les coefficients de la combinaison linéaire sont estimés en fonctions des observations (x_i, y_i).

Un exemple de ces fonctions de base est donné par la Figure 4.23 ci-contre pour des splines cubiques restreintes (voir § IX.6) d'une variable X avec 5 nœuds placés aux valeurs 26, 30, 33, 36 et 40. Les valeurs de X vont de 16 à 46 (cela peut être l'âge, par exemple) mais cela n'intervient pas dans la construction des fonctions splines de base, ainsi qu'on le verra plus loin (seuls les nœuds comptent).

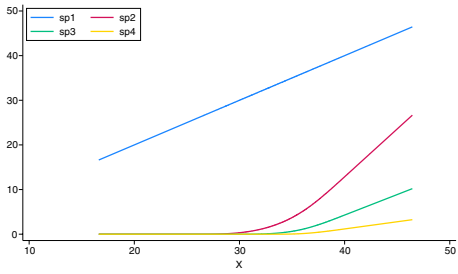


Figure 4.23 : Fonctions splines de base pour des splines cubiques restreintes avec cinq nœuds.

Il y a 4 fonctions de base, sp1 à sp4. Attention, sp1 est ici linéaire, ce qui est propre aux splines cubiques restreintes et n'est pas le cas pour d'autres façons de construire les fonctions splines (voir § IX.6).

Comme pour les polynômes fractionnaires, on constate empiriquement que les fonctions splines globales obtenues comme combinaisons linéaires des fonctions de base permettent de représenter à peu près toutes les formes de courbes. C'est ce qui fait la force de ces méthodes. La Figure 4.24 ci-contre montre des courbes qu'on peut obtenir avec, comme précédemment, des splines cubiques restreintes avec des nœuds placés aux valeurs 26, 30, 33, 36 et 40 et différents coefficients pour les fonctions de base.

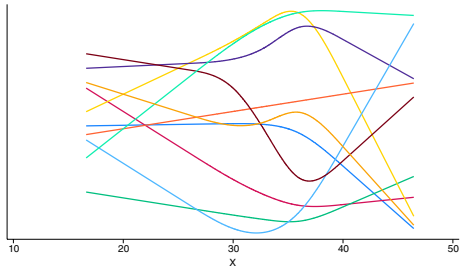


Figure 4.24 : Différentes formes de courbes obtenues avec des fonctions splines cubiques avec cinq nœuds.

1. Sauf pour les splines linéaires (voir § IX.4).

On voit que ces courbes sont très variées. Cela permet d'être assuré que, quelle que soit la forme de la relation entre X et Y, on trouvera une courbe qui la représente en choisissant correctement les coefficients des polynômes.

Remarques :

- Il faut insister sur le fait que la fonction spline ne dépend que des nœuds choisis, pas de la distribution de X ni de l'étendue de ses valeurs (j'y reviendrai dans le § X.3). Il peut être cependant utile que le choix des nœuds tienne compte des valeurs de X (voir § IX.3).
- S'il n'est pas nécessaire de savoir dans le détail comment la fonction spline est construite, il ne faut pas oublier que c'est un processus en deux temps : création des fonctions de base qui donnent des variables qui sont ajoutées à la liste des variables du fichier de données, puis inclusion (à la place de X) de ces variables dans le modèle logistique pour estimer leur coefficient, puis obtenir les valeurs prédites et tracer la courbe.
- Concernant le vocabulaire, les fonctions de base sont parfois aussi appelées fonctions splines, y compris par moi, d'où une certaine confusion possible. Comme on dit d'habitude, le contexte permet de s'y retrouver…
- La modélisation par fonctions splines est à la fois une méthode locale et globale. Locale à cause des intervalles de X et parce qu'on impose que la fonction spline soit un polynôme d'ordre d dans chaque intervalle. Globale, et c'est l'aspect qui domine, parce que finalement, c'est bien une fonction globale dont les coefficients sont estimés pour modéliser la relation entre X et Y.

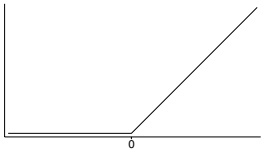
IX.1. Écriture d'une fonction spline

Il y a plusieurs façons d'écrire une fonction spline qui satisfasse aux propriétés énoncées ci-dessus. Nous en verrons des exemples plus loin, et il n'est pas toujours facile de s'y retrouver dans les différents articles qui présentent les splines, ni dans les différentes commandes des logiciels qui construisent ces fonctions (Smith PL, 1979, Wegman EJ et al., 1983). Je vais commencer par présenter une des écritures des fonctions splines qui est assez simple, même si ce n'est pas celle qui est utilisée par défaut par le logiciel R, comme on le verra avec les splines cubiques restreintes (voir § IX.6).

On a besoin des notations suivantes :

– les k nœuds sont notés s_1, …, s_k

– on définit la fonction « plus » par : $(u)_+ = \begin{cases} u & \text{si } u \geq 0 \\ 0 & \text{si } u < 0 \end{cases}$



– pour chaque nœud s_j, on définit $P_d(x;s) = (x - s_j)_+^d$, qui est une fonction de base pour construire la fonction spline globale. Il y a donc autant de fonctions de base que de nœuds. On voit cependant que $P_d(x;s_j)$ n'est pas seulement définie sur le j-ième intervalle, mais sur l'ensemble des valeurs de X. Ce n'est un polynôme de degré d que pour $x \geq s_j$, sinon elle est égale à 0.

La fonction spline globale S(x) de degré d avec les nœuds s_1, …, s_k est alors définie par (Harrell FE, 2001) : $S(x) = \alpha_0 + \sum_{i=1}^{d} \alpha_i x^i + \sum_{j=1}^{k} \beta_j P_d(x; s_j)$.

> Remarque
> J'ai retenu cette forme de la fonction spline, car elle contient un terme $\alpha_1 x$, ce qui permet un test simple de linéarité, en testant l'hypothèse que les coefficients autres que α_1 (et α_0) sont nuls simultanément.

Ce choix des fonctions $P_d(x;s_j)$ permet que soient satisfaites les contraintes de régularité de S(x) (continue et (d – 1) fois dérivable). On voit aussi qu'au sein de chaque intervalle $[s_j ; s_{j+1}]$, S(x) est bien un polynôme de degré d, car il est égal à la somme de $\sum \alpha_i x^i$ et des $P_d(x;s_j)$ pour $i \geq j$.

Les (k + d + 1) coefficients α_i et β_j sont estimés à partir des observations pour que l'adéquation avec la relation entre X et Y soit la meilleure (vraisemblance maximum).

On montre que les fonctions $P_d(x;s_j)$ présentent des inconvénients, car elles peuvent prendre des valeurs numériques très grandes lorsque d et/ou x sont grands. Cela conduit à une instabilité des estimations des coefficients. On préfère alors remplacer les $P_d(x;s_j)$ par d'autres fonctions appelées B-splines $S(x) = \alpha_0 + \sum_{i=1}^{d} \alpha_i x^i + \sum_{j=1}^{k} \beta_j B_d(x; s_j)$, dont l'écriture est plus compliquée (Newson R, 2012), mais qui sont proposées par la plupart des logiciels.

IX.2. Utilisation pratique avec des logiciels

En pratique, il n'est pas nécessaire de programmer soi-même l'écriture des fonctions splines de base. Les logiciels d'analyse statistique disposent d'une ou de plusieurs commandes qui le font (voir Annexe XII.1). Selon la commande, ce sont les P_d qui sont créées ou des B-splines, ou encore d'autres fonctions comme pour les « *restricted cubic splines* ».

Il n'est pas nécessaire de connaître dans le détail l'équation de la fonction spline qui est construite, mais il est utile de savoir quels types de fonctions splines de base sont créées par le logiciel, ne serait-ce que pour connaître leur nombre attendu (cela permet de corriger des erreurs) et pour savoir comment tester l'écart à la linéarité. Cela nécessite une lecture attentive des descriptions des modules des logiciels, qui ne sont malheureusement pas toujours explicites.

IX.3. Choix des nœuds et du degré des polynômes

Une fonction spline S est définie par deux paramètres : le nombre et la position des k nœuds qui définissent les (k + 1) intervalles de X, et le degré d des polynômes au sein de chaque intervalle.

En dehors des splines linéaires (voir § IX.4), qui représentent la relation entre X et Y par des segments de droite jointifs, la valeur de d est le plus souvent fixée à 3, qui paraît être le meilleur compromis entre la flexibilité de la courbe pour représenter la

relation entre X et Y et sa complexité (on parle alors de splines cubiques). En pratique, les logiciels ne proposent que ces deux types de splines (linéaires et cubiques), ce qui limite le choix de toute façon.
En ce qui concerne les nœuds, leur position importe moins que leur nombre (Stone C, 1986), notamment dans le cas de splines cubiques restreintes (voir § IX.6). On considère généralement qu'il n'est pas utile qu'il y en ait plus de 10 et que le bon compromis se trouve entre 3 et 5, le choix dépendant en partie de la taille de l'échantillon, afin que les intervalles contiennent suffisamment d'observations.
Les nœuds sont souvent placés aux percentiles de X, ce qui permet d'équilibrer les effectifs dans les intervalles. Ainsi, k nœuds sont placés aux percentiles $100/k+1$, $2\times100/(k+1)$, …, $k\times100/(k+1)$. Pour $k=3$, cela donne les percentiles 25, 50 et 75.
Dans le cas de splines cubiques restreintes, F.E. Harrell (Harrell FE, 2001) recommande d'écarter de façon plus importante les nœuds extrêmes de façon à mieux modéliser la courbe pour les petites et grandes valeurs de X (Tableau 4.7). Cette recommandation, qui paraît tout à fait pertinente pour une meilleure modélisation, ne repose cependant pas sur des arguments statistiques s'appuyant sur des comparaisons avec d'autres répartitions des nœuds. En réalité, il n'y a pas stratégie de choix de la « meilleure » fonction spline (quels nœuds, quel degré pour les polynômes ?). On se contente en général des règles indiquées plus haut : splines cubiques restreintes avec trois à cinq nœuds. Seuls Royston et al. ont proposé une procédure de choix, calquée sur celle des polynômes fractionnaires (Royston P et al., 2007), dont je reparlerai dans la suite (§ IX.7).

Nombre de nœuds	Percentiles où les nœuds sont placés						
3			10	50	90		
4			5	35	65	95	
5		5	27,5	50	72,5	95	
6	5	23	41	59	77	95	
7	2,5	18,33	34,17	50	65,83	81,67	97,5

Tableau 4.7 : Emplacements des nœuds recommandés par F.E. Harrell selon le nombre de nœuds

IX.4. Splines linéaires

Lorsque le degré des polynômes est 1 ($d=1$), leur représentation est une droite au sein de chaque intervalle. La relation entre Y (logit P) et X est alors modélisée par une fonction linéaire par morceaux. Contrairement aux fonctions en escalier, les morceaux sont jointifs, cela fait partie des contraintes des splines, et (donc…) non horizontaux. Les splines linéaires ont l'avantage de la simplicité (et correspondent à la primauté donnée à la droite par rapport aux autres formes de courbes) et peuvent permettre de repérer des ruptures de pentes. La courbe reste continue, mais n'est plus vraiment « lisse », elle a des angles à chaque nœud.

Si on prend l'exemple de trois nœuds positionnés en a, b et c, la fonction spline s'écrit :

$$S(X) = \beta_0 + \beta_1 X + \beta_2 (X-a)_+ + \beta_3 (X-b)_+ + \beta_4 (X-c)_+.$$

Avec cette écriture, les coefficients β_j sont égaux aux changements de pente à chaque nœud : β_1 est la pente initiale et β_{j+1} est la variation de pente après le nœud j. En effet, on peut écrire la fonction spline sous la forme équivalente :

$$S(X) \begin{cases} = \beta_0 + \beta_1 X & \text{si } X < a \\ = (\beta_0 - a\beta_2) + (\beta_1 + \beta_2) X & \text{si } a \le X < b \\ = (\beta_0 - a\beta_2 - b\beta_3) + (\beta_1 + \beta_2 + \beta_3) X & \text{si } b \le X < c \\ = (\beta_0 - a\beta_2 - b\beta_3 - c\beta_4) + (\beta_1 + \beta_2 + \beta_3 + \beta_4) X & \text{si } c \le X \end{cases}$$

D'autres écritures sont possibles pour des splines linéaires, notamment celle où les coefficients des fonctions splines sont égaux aux pentes des droites au sein de chaque intervalle (option « marginal » des logiciels qui la proposent).

Les exemples qui suivent ont été établis avec les données de FIV, en choisissant la règle des percentiles pour l'emplacement des nœuds.

IX.4.a. Un nœud (k = 1, donc 2 droites)

Le nœud est placé au 50e percentile, c'est-à-dire à la médiane de l'âge, qui est égale ici à 33,1. Deux fonctions splines sont donc construites : X et $(X-33{,}1)_+$ (Figure 4.25a). La relation entre X (âge) et le succès en FIV est représentée par deux segments de droite (Figure 4.25b).

Attention cependant au fait que l'angle sur la courbe de la Figure 4.25b correspond à l'emplacement du nœud (ici la médiane de l'âge), qui a été choisi indépendamment de la relation entre X et Y. On observe donc, par construction, une rupture de pente à cet endroit (qui est ici significative). Vu la position des logits observés, on peut penser qu'il y a effectivement une rupture de pente, mais rien ne dit qu'elle est située à cette valeur particulière de X.

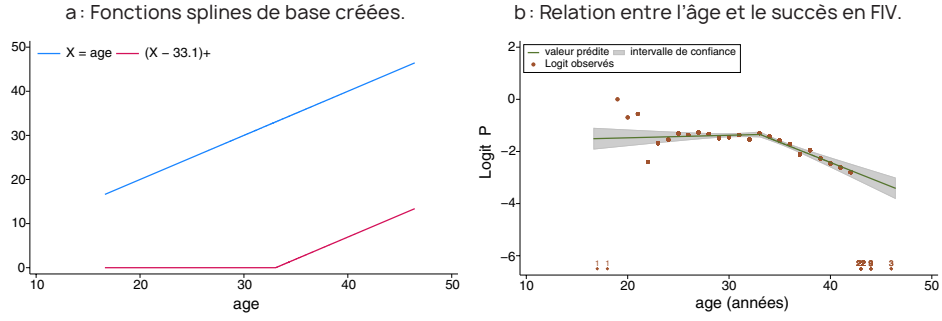
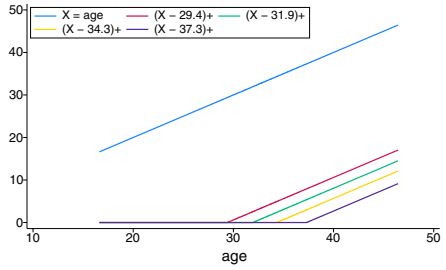


Figure 4.25 : Modélisation par une fonction spline linéaire avec 1 nœud (situé à 33,1).

IX.4.b. Quatre nœuds (k = 4, donc 5 droites)

On peut essayer de mieux représenter les données en prenant davantage d'intervalles. Avec 4 nœuds, toujours situés aux percentiles de l'âge de façon à répartir l'échantillon en 5 groupes de même effectif, 5 fonctions splines sont créées, représentées sur la Figure 4.26a où l'emplacement des nœuds est indiqué dans la légende. La relation entre l'âge et le succès en FIV est modélisée sur la Figure 4.26b. On constate que la courbe passe effectivement (un peu) plus près des points observés, mais c'est au prix de changements de pente qui forment des « zigzags », modérés, mais peu crédibles. On constate par ailleurs une augmentation de la largeur de l'intervalle de confiance aux extrémités de la courbe. De nouveau, en augmentant le nombre de nœuds, on modélise à la fois la « vraie » relation et le « bruit » des fluctuations d'échantillonnage. Dans le cas présent, on peut raisonnablement conclure qu'on va trop loin en prenant 4 nœuds.

a : Fonctions splines de base créées.



b : Relation entre l'âge et le succès en FIV.

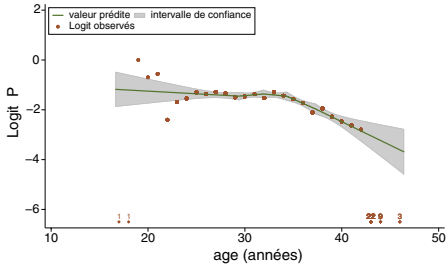


Figure 4.26 : Modélisation par une fonction spline linéaire avec 4 nœuds placés aux percentiles.

IX.4.c. Réduction du nombre de nœuds avec une procédure pas-à-pas

L'augmentation du nombre de nœuds, si elle n'est pas utile à la modélisation de la relation entre X et Y, peut permettre de préciser l'emplacement de la (ou des) rupture(s) de pente, en évitant l'arbitraire qui a été souligné dans le cas d'un seul nœud choisi a priori.

Une méthode possible est de commencer avec un grand nombre (k) de nœuds. La fonction spline avec k nœuds s_1, …, s_k s'écrit : $S(X) = \beta_0 + \beta_1 X + \sum_{j=2}^{k+1} \beta_j (X - s_{j-1})_+$. (k+1) variables sont créées : X et les $(X - s_j)_+$, dont les coefficients dans une régression logistique avec logit P sont les modifications de la pente à chaque nœud. Si un coefficient d'une variable $(X - s_j)_+$ est non significatif, cela signifie que le changement de pente après le nœud s_j n'est pas significativement différent de 0. On peut alors considérer qu'il n'est pas nécessaire de prendre en compte cette variable, dont la suppression dans le modèle revient à supprimer le nœud s_j et à réunir les deux intervalles de part et d'autre du nœud s_j de façon à n'avoir qu'un seul intervalle et qu'une seule pente de droite.

Sur le plan pratique, on peut y parvenir avec une procédure pas-à-pas (*stepwise*, voir chapitre 5, § V.4)), qui conduira à retirer la variable $(X - s_j)_+$, dont le coefficient est non significatif.
Il faut cependant que la première variable (X) reste dans le modèle final pour qu'on conserve une modélisation par une fonction spline. Cela évite aussi que la fonction retenue finalement commence par une droite horizontale jusqu'au premier nœud, ce qui est une hypothèse forte et non souhaitable.
La procédure *stepwise* doit donc être écrite de sorte que la variable X ne puisse pas être sortie.

Exemple
Cet exemple porte toujours sur les données de FIV.
On commence avec 11 nœuds placés aux percentiles. La courbe initiale, qui est une fonction linéaire par morceaux avec 12 morceaux, n'a pas d'intérêt en elle-même ; elle permet juste de choisir au départ un ensemble de nœuds suffisamment varié pour limiter l'arbitraire de celui (ou de ceux) qui sera (seront) finalement retenu(s). Les résultats apparaissent dans le Tableau 4.7.
L'emplacement des 11 nœuds est donné juste après la construction des fonctions splines de base. La procédure *stepwise* comprend l'option lockterm1, qui permet de conserver la première variable, qui est l'âge lui-même. Avec un seuil « classique » d'élimination des variables fixé à 10 %, toutes les variables sauf une sont retirées successivement, ce qui conduit à ne retenir qu'un seul nœud. C'est le sixième nœud qui correspond à la variable vsp_1_6, soit 33,1. Le modèle final est le même que celui du § IX.4.a, mais cette fois-ci, c'est avec une plus grande assurance qu'on peut dire qu'il y a une rupture de pente à cette valeur particulière de l'âge.

```
. makespline linear age, order(1) basis(vsp) knots(11) replace
. matrix l r(knots), format(%6.1f)
r(knots)[1,11]
     c1    c2    c3    c4    c5    c6    c7    c8    c9   c10   c11
r1  27.2  28.8  30.1  31.1  32.1  33.1  34.1  35.2  36.4  37.8  39.6

. sw, pr(.10) lockterm1 : logit acc `r(regressors)'

Wald test, begin with full model:
p = 0.8712 >= 0.1000, removing vsp_1_11
p = 0.6175 >= 0.1000, removing vsp_1_3
p = 0.2191 >= 0.1000, removing vsp_1_4
p = 0.3660 >= 0.1000, removing vsp_1_5
p = 0.2523 >= 0.1000, removing vsp_1_7
p = 0.2222 >= 0.1000, removing vsp_1_8
p = 0.4609 >= 0.1000, removing vsp_1_9
p = 0.4389 >= 0.1000, removing vsp_1_10
p = 0.1814 >= 0.1000, removing vsp_1_1
p = 0.2366 >= 0.1000, removing vsp_1_2

Logistic regression                                     Number of obs =  6,400
                                                        LR chi2(2)    = 102.16
                                                        Prob > chi2   = 0.0000
Log likelihood = -2837.8689                             Pseudo R2     = 0.0177

------------------------------------------------------------------------------
         acc | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .0098092   .0152023     0.65   0.519    -.0199867    .0396051

      vsp_1_6 |  -.1642499   .0286578    -5.73   0.000     -.220418   -.1080817
       _cons |  -1.672471   .4633235    -3.61   0.000    -2.580569   -.7643741
------------------------------------------------------------------------------
```

Tableau 4.7 : Élimination successives des nœuds par *stepwise* d'une fonction spline avec 11 nœuds initialement.

IX.5. Splines cubiques

Les splines cubiques correspondent à d = 3. Elles modélisent la relation entre X et Y au sein de chaque intervalle par des polynômes de degré 3 qui permettent de mieux « coller » aux observations qu'une droite, tout en limitant le nombre de coefficients nécessaires par rapport à des polynômes de degré supérieur.

L'écriture d'une fonction spline cubique avec k nœuds est la suivante : $S(x) = \alpha_0 + \sum_{i=1}^{3} \alpha_i x^i + \sum_{j=1}^{k} \beta_j (x - s_j)_+^3$. k variables $(x - s_j)_+^3$ sont créées en plus des puissances de X et (k + 4) coefficients doivent être estimés.

Avec l'exemple du succès en FIV, et en prenant 3 nœuds placés aux percentiles, on obtient la courbe de la Figure 4.27.

On constate habituellement que les fonctions splines cubiques ont tendance à être sensibles aux valeurs observées dans le premier et le dernier intervalle. Cela signifie que la forme de la courbe aux extrémités de la zone de variation de X dépend de façon très forte (et peut-être trop forte) des valeurs de Y dans ces intervalles, sans « correction » possible par des valeurs au-delà, car il n'y en a pas ou elles sont non observées[2]. C'est particulièrement vrai lorsqu'il y a peu d'observations dans les intervalles extrêmes. Cela se traduit, comme on le voit sur la Figure 4.27, par des variations importantes de la courbe et par un fort élargissement de l'intervalle de confiance.

C'est ce qui a motivé le recours à des fonctions splines cubiques restreintes.

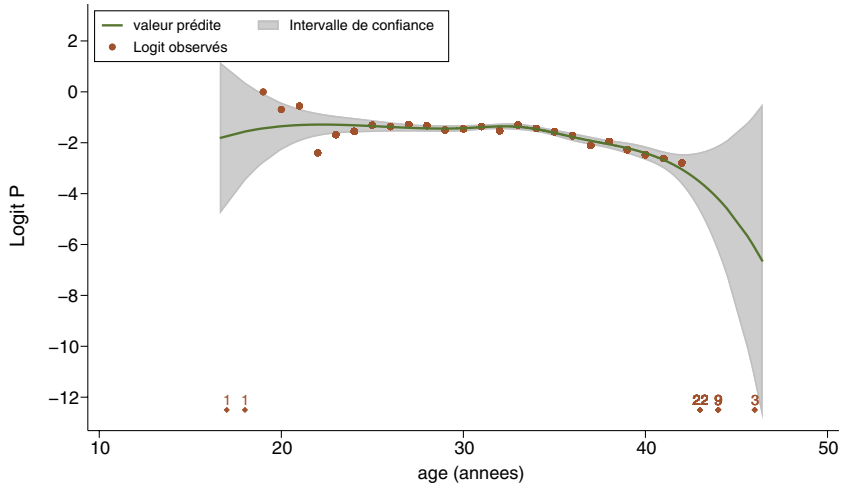


Figure 4.27 : Relation entre l'âge et le succès en FIV par des splines cubiques avec 3 nœuds.

2. Ce phénomène ne se produit pas pour les intervalles au centre de la courbe, car il y a toujours des observations à droite et à gauche qui interviennent sur sa forme.

IX.6. Splines cubiques restreintes

Pour tenir compte de la sensibilité des fonctions splines cubiques aux valeurs extrêmes de X, on utilise plutôt les splines cubiques restreintes (*restricted cubic splines* en anglais). Elles sont aussi appelées « *natural splines* ». Ce sont les mêmes fonctions que précédemment (d = 3), mais auxquelles on ajoute la contrainte d'être linéaires pour les deux intervalles extrêmes. C'est bien sûr une hypothèse supplémentaire difficile, voire impossible, à vérifier, mais qui évite l'amplitude des « dérives » des débuts ou fins de courbes observées avec des splines cubiques.

Pour obtenir les fonctions splines cubiques restreintes avec k nœuds, on part de la forme habituelle des fonctions splines cubiques que j'ai rappelée plus haut :

$$S(x) = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \beta_{03}x^3 + \sum_{j=1}^{k}\beta_{1j}(x - s_j)_+^3$$

En ajoutant les contraintes de linéarité pour les deux intervalles extrêmes[3], on montre que cela introduit des contraintes sur les coefficients qui permettent de simplifier la formule de S(x) (Durrleman S et al., 1989, Harrell FE, 2001, Royston P et al., 2002, Royston P et al., 2007). Il n'y a plus alors besoin que de k coefficients (4 de moins) et la fonction s'écrit : $S(x) = \beta'_{00} + \beta'_{01}x + \sum_{j=2}^{k-1}\beta'_{1j}\upsilon_j(x)$, avec :

$$\upsilon_j(x) = \frac{1}{(s_k - s_1)^2}\left[(x - s_{j-1})_+^3 - (x - s_{k-1})_+^3\frac{(s_k - s_{j-1})}{(s_k - s_{k-1})} + (x - s_k)_+^3\frac{(s_{k-1} - s_{j-1})}{(s_k - s_{k-1})}\right] \text{ pour } j = 2, \ldots, k-1$$

ou, plus fréquemment $S(x) = \beta'_{00} + \sum_{j=1}^{k-1}\beta'_{1j}\upsilon_j(x)$ avec $\upsilon_1(x) = x$.

Il y a donc (k–1) variables υ_j créées lorsqu'il y a k nœuds, la première étant X elle-même. Avec ces splines cubiques restreintes, il est très facile de tester l'écart à la linéarité : il suffit de tester simultanément si tous les paramètres autres que β_{01} (ou β_{11} avec la seconde écriture) sont nuls.

Exemple de splines cubiques restreintes avec trois nœuds et les données de FIV

Deux fonctions sont créées : X et υ_2 (Figure 4.28a). La relation entre l'âge et le succès en FIV est représentée sur la Figure 4.28b. On voit que le gain est surtout dans la forme de la courbe, linéaire aux deux extrémités, et dans la réduction de son intervalle de confiance.

3. En pratique, cela revient à écrire que la dérivée seconde de S(x) est nulle sur les intervalles extrêmes.

a. Fonctions splines cubiques restreintes créées.

b. Représentation graphique de la relation.

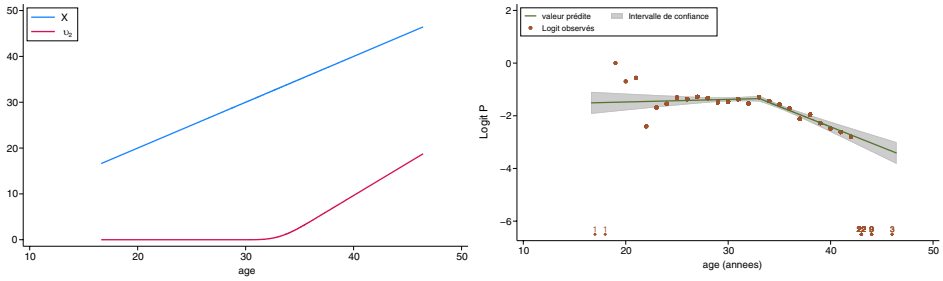


Figure 4.28 : Fonctions de base splines cubiques restreintes créées avec trois nœuds.

IX.7. Stratégie de choix du modèle avec des fonctions splines

Lorsque la modélisation est faite avec des fonctions splines, le choix du modèle est celui du nombre et de la position des nœuds et du degré des polynômes dans chaque intervalle.

Pour le degré des polynômes, le choix se limite en pratique à degré 1 (qui correspond au choix de privilégier une droite par morceaux) ou degré 3 (dont on a constaté qu'il permettait de représenter correctement pratiquement toutes les formes de courbes).

Pour les nœuds, la situation est plus compliquée car, contrairement aux polynômes fractionnaires où le nombre de puissances possibles est limité à huit, il n'y a pas de limite au nombre de nœuds, même si on admet qu'il est rarement utile qu'il dépasse 5, et surtout pas de limite à leurs emplacements, dont le choix est infini.

Royston et al. (Royston P et al., 2007) ont proposé une stratégie de choix, inspirée de celle qu'ils ont construite pour les polynômes fractionnaires, que j'ai décrite dans les § VIII.3 et VIII.4. Elle consiste à choisir les nœuds qu'on retient parmi un ensemble déterminé à l'avance. Si cet ensemble est assez grand, l'arbitraire de l'emplacement de chacun n'a pratiquement pas d'importance. Je vais présenter cette stratégie en commençant par le cas d'une seule variable X, qui sera étendu à la modélisation simultanée de plusieurs variables X_i.

Cette stratégie est, à ma connaissance, programmée uniquement dans Stata.

IX.7.a. Choix d'une fonction spline pour modéliser une variable

Le processus proposé par Royston et al. consiste à commencer par fixer le degré des polynômes. Le plus fréquent est de prendre 3 et de considérer les splines cubiques restreintes. Mais on pourrait aussi prendre 1 pour que la modélisation soit une fonction linéaire par morceaux ou même 0, qui correspond au fait que la fonction spline est constante au sein d'un intervalle, c'est-à-dire qu'on aurait une modélisation par

des variables indicatrices (et donc une fonction en escalier). La méthode n'inclut pas de comparaison (et donc pas de choix) entre ces différentes puissances.
On fixe ensuite le nombre maximum de nœuds, m, et leur emplacement. Ils peuvent être placés tous les 100/(m+1)-ièmes percentiles ou à des emplacements choisis, par exemple, en suivant la règle de Harrell (§ IX.3).
La suite du processus consiste à réduire le nombre de nœuds en ayant pour référence la modélisation linéaire, comme pour les polynômes fractionnaires.

En désignant par M_m le modèle le plus complet (avec l'ensemble des m nœuds) et M_0 le modèle linéaire (sans nœud), les étapes sont les suivantes.

- ✓ Étape 1 : Comparer M_m au modèle ne contenant pas la variable X. Si la différence est non significative, X n'est pas significativement lié à Y. Le processus de sélection s'arrête là et X n'est pas inclus dans l'analyse.
- ✓ Étape 2 : Sinon, comparer M_m à M_0. Si la différence est non significative, la linéarité n'est pas rejetée et le modèle linéaire M_0 est retenu.
- ✓ Étape 3 : Sinon, on considère les m fonctions splines possibles en utilisant un seul des m nœuds et on retient celui qui est le plus adéquat (c'est-à-dire qui a la plus grande vraisemblance), noté $\widetilde{M}_1$. Si $\widetilde{M}_1$ n'est pas significativement différent de M_m, il n'y a pas d'évidence que le modèle le plus complexe soit meilleur et on conserve $\widetilde{M}_1$.
- ✓ Étape 4 : Sinon, on « augmente » $\widetilde{M}_1$ en ajoutant successivement chacun des m–1 nœuds restants. Parmi les m–1 modèles obtenus, on retient celui qui est le plus adéquat (qui a la plus grande vraisemblance), noté $\widetilde{M}_2$, et on le compare à M_m. Si la différence est non significative, on retient $\widetilde{M}_2$; sinon, on réitère l'étape 4 avec les m–2 nœuds restants.

```
. uvrs logit acc age, alpha(0.05) degree(3) df(4) trace
  Knot    Deviance   Loss   P
  All     5674.755   0.000   .
Linear    5710.067  35.313  0.000
-
3 candidate knots considered
  30.08   5681.932   7.178  0.028
  33.06   5679.962   5.208  0.074
  36.40   5678.833   4.078  0.130
-
Adding knot at 36.4 and stopping search.

Iteration 0:   log likelihood = -2888.9503
Iteration 1:   log likelihood = -2841.6906
Iteration 2:   log likelihood = -2839.4302
Iteration 3:   log likelihood = -2839.4164
Iteration 4:   log likelihood = -2839.4164

Logistic regression                             Number of obs   =       6400
                                                LR chi2(2)      =      99.07
                                                Prob > chi2     =     0.0000
Log likelihood = -2839.4164                     Pseudo R2       =     0.0171

------------------------------------------------------------------------------
         acc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_0 |  -.3304828   .0386539    -8.55   0.000    -.4062431   -.2547224
       age_1 |    .207445   .0389672     5.32   0.000     .1310708    .2838193
       _cons |  -1.652334   .0351207   -47.05   0.000    -1.721169   -1.583499
------------------------------------------------------------------------------
Deviance:  5678.833. Best knots:  36.4
```

Tableau 4.8 : Procédure uvrs de sélection de la meilleure fonction splines cubiques restreintes avec une seule variable

La procédure se poursuit et s'arrête lorsqu'un test non significatif survient ou, si tous les tests sont significatifs, qu'on est amené à choisir finalement M_m.

Tous les tests précédents sont des χ^2 du rapport des vraisemblances.

Cette méthode est programmée dans Stata avec la commande uvrs. Dans l'exemple qui suit, réalisé sur les données de FIV, degree() indique le degré de la fonction spline et df() le nombre d'intervalles. Le nombre de nœuds est égal à df() –1; ils sont placés aux percentiles correspondants.
On part de quatre nœuds, placés aux quartiles. À l'issue de la procédure, le nœud du troisième quartile (placé à 36,4) est le seul retenu (Tableau 4.8).

On peut noter que si on avait pris des splines linéaires (degree(1)), un seul nœud aurait été retenu parmi les trois, et cela aurait été 30,06. Un résultat conforme à ce qui avait été obtenu dans le § IX.4.

IX.7.b. Modélisation simultanée de plusieurs variables avec des fonctions splines

Lorsqu'on veut modéliser plusieurs variables simultanément, le processus est généralisé d'une manière similaire à celle présentée pour les polynômes fractionnaires (§ VIII.4).

- ✓ On choisit pour chaque variable le nombre maximum de nœuds et leurs emplacements. Ce choix peut être différent d'une variable à l'autre, bien qu'il y ait rarement une raison de le faire.
- ✓ On ordonne les variables selon la valeur croissante du degré de signification p de la pente de leur relation avec logit P dans des modèles linéaires successifs où les variables sont considérées une par une.
- ✓ On choisit alors la meilleure modélisation par fonction spline pour la première variable (les autres étant incluses dans le modèle sous forme linéaire) en utilisant la méthode indiquée pour une seule variable.
- ✓ On recommence pour les variables suivantes en ajustant sur celles qui ont déjà été traitées avec la modélisation qui leur a été attribuée, et en continuant à inclure celles qui n'ont pas été traitées sous forme linéaire.
- ✓ Lorsque toutes les variables ont été traitées, on recommence un nouveau cycle en prenant les variables dans le même ordre, mais cette fois les variables sont incluses dans le modèle avec la modélisation obtenue dans le cycle précédent (et non sous forme linéaire).
- ✓ Le processus s'arrête lorsque la modélisation des variables ne change plus entre deux cycles successifs.

Cette méthode est programmée dans Stata avec la commande mvrs.

```
. mvrs logit acc age ovo emb, alpha(0.05) degree(3) df(4) select(0.05)

Deviance for model with all terms untransformed =  5576.866, 6366 observations

------------------------------------------------------------
Variable  Final  Deviance   Dev.diff    P     Final knot
             df            cf. null          positions
------------------------------------------------------------
emb           3  5394.698   270.731   0.000   2 3
age           2  5367.248    72.767   0.000   36.4
ovo           0  5369.204     1.957   0.372
------------------------------------------------------------
End of Cycle 1: deviance =    5369.204
------------------------------------------------------------

emb           3  5369.204   291.343   0.000   2 3
age           2  5369.204    70.811   0.000   36.4
ovo           0  5369.204     1.957   0.372

Regression spline fitting algorithm converged after 2 cycles.

Transformations of covariates:

Final multivariable spline model for acc
------------------------------------------------------------------------------
    Variable |     -----Initial-----          -----Final-----
             |   df     Select   Alpha    Status   df    Knot positions
-------------+----------------------------------------------------------------
         age |    4     0.0500   0.0500     in      2     [lin] 36.4
         ovo |    4     0.0500   0.0500     out     0
         emb |    4     0.0500   0.0500     in      3     [lin] 2 3
------------------------------------------------------------------------------

Logistic regression                             Number of obs   =       6366
                                                LR chi2(5)      =     389.82
                                                Prob > chi2     =     0.0000
Log likelihood = -2684.6022                     Pseudo R2       =     0.0677

------------------------------------------------------------------------------
         acc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_0 |  -.2786657   .0394025    -7.07   0.000    -.3558932   -.2014382
       age_1 |   .1984333   .0394864     5.03   0.000     .1210415    .2758252
       emb_0 |   .6243414   .0524218    11.91   0.000     .5215965    .7270863
       emb_1 |   .5552767   .0560337     9.91   0.000     .4454526    .6651007
       emb_2 |  -.5310683   .0635433    -8.36   0.000    -.6556109   -.4065257
       _cons |  -1.883959   .0490448   -38.41   0.000    -1.980085   -1.787833
------------------------------------------------------------------------------
Deviance: 5369.204.
```

Tableau 4.9 : Procédure mvrs de sélection de la meilleure fonction splines cubiques restreintes avec plusieurs variables.

X. Présentation des résultats issus de la modélisation

Ainsi qu'on a pu s'en rendre compte dans les présentations des polynômes fractionnaires et des fonctions splines, les variables créées ne sont pratiquement pas interprétables. Dans le cas des fonctions splines, il est même souvent impossible en pratique d'écrire leur équation. Les coefficients associés aux variables créées qui apparaissent dans les sorties des logiciels sont eux aussi ininterprétables ainsi que les tests séparés qui les accompagnent. Il est même difficile de se rendre compte de l'allure de la courbe au seul vu des variables et de leurs coefficients (ne serait-ce que de savoir si elle croissante ou pas).

Il n'y a guère que pour les splines linéaires, pour lesquelles on peut choisir que les coefficients des splines de base soient égaux aux variations de pente d'un intervalle à l'autre, que leur valeur et leur test peuvent avoir une interprétation utile.

On ne peut donc pas se contenter de donner les sorties des logiciels en guise de résultats. Cette règle a en réalité une portée générale, mais est encore plus vraie ici, si je puis dire !
Je vais revenir ci-dessous sur trois aspects particuliers de la présentation des résultats après modélisation :

- test de l'existence d'une association entre X et Y et de sa linéarité,
- forme de cette association,
- expression quantitative des résultats (pour les faire apparaître dans les tableaux d'un article, par exemple).

X.1. Test de l'association entre X et Y, test de linéarité

Pour tester l'existence d'une association entre X et Y, le test doit porter globalement sur l'ensemble des coefficients. L'hypothèse nulle (absence d'association entre X et Y) est que tous les coefficients des variables créées pour les fonctions splines ou pour les polynômes fractionnaires sont nuls. C'est ce test unique qui doit apparaître dans les résultats. Les tests séparés des coefficients ne servent à rien pour le test de l'association entre X et Y, voire sont trompeurs.

La modélisation linéaire de l'association entre X et Y occupe une place privilégiée, comme je l'ai expliqué dans le § IV.3. Avec les polynômes fractionnaires, le test de linéarité est intégré dans la procédure mfp (§ VIII.5) : son résultat indique directement si une relation linéaire peut être retenue ou doit être rejetée. Avec les fonctions splines cubiques restreintes, il faut choisir la paramétrisation où la première variable créée est X. Les logiciels Stata et R le permettent (voir § XII.1). Il suffit alors de tester si les coefficients des autres variables sont nuls pour avoir un test de linéarité.

X.2. Représentation graphique de la relation entre X et Y

Une représentation graphique est essentielle pour connaître la forme de la relation entre X et Y, que les équations des variables créées ou leurs coefficients ne permettent pas de deviner.
Dans le cas des polynômes fractionnaires et du logiciel Stata, la commande « fracplot » permet d'obtenir directement le graphe de la courbe. Sinon, ou si on veut des graphes différents (ajout d'une légende, modification de l'intitulé de l'axe des Y et de la représentation des observations…), il faut utiliser les commandes qui donnent la valeur de logit P prédite et son écart-type. Les commandes de graphiques des logiciels permettent ensuite de tracer la courbe elle-même et son intervalle de confiance, avec les difficultés dont j'ai déjà parlé si on veut faire figurer les valeurs observées. Des programmes Stata (logit_crb.ado) et R (logit_crb.R) permettant d'obtenir les courbes avec les points observés telles qu'elles ont été présentées dans le § II sont disponibles sur le web, au même endroit que les codes Stata et R correspondant aux tableaux et aux exemples du livre[4].

4. https://laboutique.edpsciences.fr/produit/1504/9782759838189/la-regression-logistique-en-epidemiologie.

Le plus souvent les courbes sont données avec logit P en ordonnée. C'est ce qui a été fait pour toutes celles de ce chapitre. La raison principale en est que c'est cette quantité qui est modélisée. Cette habitude est entrée dans le vocabulaire : lorsqu'on parle d'un modèle linéaire ou d'écart à la linéarité, cela concerne logit P et pas la probabilité P. Il est bien sûr possible de faire une courbe avec P comme ordonnée ; cela peut paraître plus parlant, mais c'est plutôt une source de complications. D'une part, il faut bien expliquer ce qu'on fait, ce qui n'est pas toujours facile en face d'un lecteur ou d'un auditoire non averti, en particulier lorsqu'il s'agit de le convaincre qu'il est normal qu'un modèle « linéaire » ne soit pas représenté par une droite. D'autre part, il faut éviter de le faire pour une enquête cas-témoin où la valeur de P estimée par le modèle n'est alors pas interprétable, ce qui rend trompeuse l'interprétation de l'échelle des ordonnées.

> Remarque
> Dans une enquête cas-témoin, la valeur de logit P n'est pas plus interprétable que celle de P. On est « sauvé » par le fait que logit P n'est pas observable au niveau individuel et que l'échelle des ordonnées en logit P ne signifiant pas grand-chose, on ne regarde que la courbe…

Dans la plupart des courbes de ce chapitre, j'ai fait apparaître les valeurs observées en considérant que cela peut constituer un garde-fou pour le choix de la modélisation. C'est souvent le cas lorsque Y est une variable quantitative, mais c'est plus discutable lorsque Y est une variable dichotomique et que la modélisation porte sur logit P, car logit P n'est pas observable au niveau individuel. Il est possible de « contourner » la difficulté, ainsi que cela est détaillé dans les § II et XII.2 et comme cela a été fait dans l'ensemble de ce chapitre. Cela demande un effort de programmation et surtout d'expliquer ce qu'on fait, ce qui n'est pas toujours facile à faire car ce n'est pas l'objet principal des résultats qu'on veut montrer (et aussi pour des raisons de temps ou de place dans une communication à un congrès ou dans un article). Il est alors possible (et légitime) de représenter la courbe sans les points observés, en gardant quand même, autant que possible, l'intervalle (ou bande) de confiance.

X.3. Présentation quantitative des résultats dans un tableau

Les représentations graphiques ne résolvent pas la question de la publication des résultats dans des tableaux avec un contenu quantitatif, comme on le fait habituellement avec des odds ratios lorsque X est en classes.
Il n'est pas réellement envisageable de donner les coefficients des variables créées pour les polynômes fractionnaires ou les fonctions splines. Ces variables ne sont pas interprétables individuellement et leurs coefficients ne donnent aucune indication quantitative sur la relation entre X et Y. C'est d'ailleurs la raison principale pour laquelle une variable quantitative est souvent transformée en variables indicatrices après avoir été découpée en classes (voir §VI). Malgré les inconvénients que cette

transformation représente, les odds ratios par catégorie de la variable X sont, eux, faciles à interpréter.
Une bonne solution de compromis est de modéliser la relation entre X et Y avec des splines ou des polynômes fractionnaires en gardant X quantitative, puis de faire des classes X et de calculer les OR correspondant aux centres des classes, en se servant de la modélisation (Royston P et al., 1999).

De façon pratique, notons la modélisation de la relation entre X et Y, $\text{logit } P = \alpha + \sum_{i=1}^{m} \beta_i v_i(x)$, où les $v_i(x)$ sont les variables créées par les polynômes fractionnaires ou les splines, selon la méthode utilisée. Considérons deux classes de X dont les centres sont respectivement x_0 et x_1. L'odds ratio entre ces classes est estimé par : $\ln OR = \text{logit } P_1 - \text{logit } P_0 = \sum_{i=1}^{m} \beta_i (v_i(x_1) - v_i(x_0)) = \sum_{i=1}^{m} D_i \beta_i$. Les D_i sont des valeurs fixes ne dépendant que des classes de X et des fonctions v_i utilisées pour la modélisation. ln OR est donc une combinaison linéaire des coefficients β_i qu'on peut estimer ainsi que son intervalle de confiance avec les fonctions classiques des logiciels (lincom pour Stata et R).

Pour illustrer ce propos un peu théorique, prenons l'exemple de la relation entre l'âge et le succès en FIV et considérons des classes d'âge de 5 ans : [15–19[; [20–24[; … ; [40–46[. Les centres des classes sont 17 ; 22 ; … ; 42 (on peut discuter de prendre 17,5 ; 22,5…, mais ce n'est pas essentiel).
L'odds ratio entre la classe [25–29[prise comme référence et la classe [15–19[est donné par : $\ln OR = \sum_{i=1}^{m} \beta_i (v_i(17) - v_i(27))$. Le calcul des $D_i = v_i(17) - v_i(27)$ dépend de l'expression des v_i, qui est différente selon le type de modélisation.

- Modélisation par un polynôme fractionnaire
  Le modèle retenu par la procédure mfp (voir § VIII.4.a) a les puissances (3 3) et s'écrit donc : $\text{logit } P = \alpha + \beta_1 x^3 + \beta_2 x^3 \ln x$ (avec $x = \text{age}/10$). On a donc $v_1(x) = x^3$ et $v_2(x) = x^3 \ln x$. Ce qui donne :
  $D_1 = 1{,}7^3 - 2{,}7^3 = -14{,}770$
  $D_2 = 1{,}7^3 \ln(1{,}7) - 2{,}7^3 \ln(2{,}7) = -16{,}943$.
  On en déduit : $\ln OR = -14{,}770\ \beta_1 - 16{,}943\ \beta_2$.

Cela permet de calculer OR après avoir remplacé β_1 et β_2 par leur estimation ($\beta_1 = 0{,}21$; $\beta_2 = -0{,}15$). En pratique, il est plus simple d'utiliser la commande lincom, qui donne OR et son intervalle de confiance : 0,57 [0,38-0,86].
Les commandes sont les suivantes :
avec Stata :

```
mfp, center(no) : logistic acc age
lincom -14.770*Iage_1-16.943*Iage_2
```

avec R :

```
cycles3$age10 <- cycles3$age/10
fp2 <-mfp(acc~fp(age10,df=4,scale=FALSE),family=binomial(),data=cycles3,select=0.05)
OR.17.22 <- lincom(fp2,"- 14.770*age10.1-16.943*age10.2",eform=T)
```

- Modélisation par une fonction spline
  Le modèle avec des fonctions splines cubiques restreintes et trois nœuds a été donné au § IX.6. Il contient deux variables et s'écrit : $\text{logit } P = \alpha + \beta_1 v_1(x) + \beta_2 v_2(x)$, avec $v_1(x) = x$ et $v_2(x) = \upsilon(x)$, où $\upsilon(x)$ a l'expression donnée au § IX.6, compliquée et peu utilisable en pratique pour calculer D_1 et D_2 comme on l'a fait avec les polynômes fractionnaires. Il faut donc procéder autrement. Si l'échantillon étudié comprend des sujets dont la valeur de X est exactement x_0 ou x_1, on a directement les valeurs de v_1 et v_2 en ces points, mais cette situation est peu fréquente. Il faut alors se rappeler que la construction des fonctions splines ne dépend que de la position des nœuds, pas de la distribution de X. Il suffit donc de reconstruire les mêmes fonctions splines (c'est-à-dire avec les mêmes nœuds) sur un échantillon qui contient x_0 et x_1 pour avoir les valeurs souhaitées de v_1 et v_2. La mise en œuvre demande plus de manipulations de fichiers et de lignes de commandes, mais est, somme toute, assez simple dans son principe. Elle se termine par l'utilisation de la commande lincom comme avec les polynômes fractionnaires.

Il n'empêche que, tout en étant d'un principe (relativement) simple, ces méthodes pour obtenir des OR par classes sont fastidieuses et peuvent décourager de modéliser une variable quantitative par splines ou polynômes fractionnaires.
C'est pourquoi j'ai écrit des fonctions, une pour les polynômes fractionnaires (ORcl_pf), l'autre pour les fonctions splines (ORcl_sp), chacune en Stata et en R, qui réalisent ce travail et qu'on peut trouver à la même adresse que les codes Stata et R correspondant aux exemples de ce livre[5]. Ces fonctions permettent en outre d'inclure d'autres variables que X pour avoir des OR ajustés.

En faisant les calculs « à la main » comme indiqué plus haut, ou en utilisant ces fonctions avec des classes d'âge de 5 ans, on obtient les résultats du Tableau 4.11 (la première colonne correspondant aux calculs classiques avec des variables indicatrices).

Âge	OR « non modélisés »*	PF_2**	Splines***
15–19	1,32 [0,14–12,75]	0,57 [0,38–0,86]	0,80 [0,57–1,13]
20–24	0,89 [0,54–1,45]	0,80 [0,66–0,98]	0,90 [0,75–1,06]
25–29	1	1	1
30–34	0,96 [0,81–1,14]	0,96 [0,85–1,10]	1,02 [0,89–1,18]
35–39	0,62 [0,51–0,75]	0,61 [0,51–0,72]	0,60 [0,50–0,71]
40–44	0,28 [0,19–0,41]	0,21 [0,15–0,29]	0,25 [0,19–0,34]

* C'est-à-dire obtenus par transformation de X en classes avec variables indicatrices.

** Polynômes fractionnaires de degré 2 (voir aussi § VIII.2.b et Figure 4.18).

*** Splines cubiques restreintes avec 3 nœuds placés selon Harrell (voir § IX.6 et Figure 4.28b).

Tableau 4.11 : Odds ratios par classes selon la modélisation de X (variables indicatrices ou polynômes fractionnaires) avec la classe d'âge [25 ; 29[comme classe de référence.

5. https://laboutique.edpsciences.fr/produit/1504/9782759838189/la-regression-logistique-en-epidemiologie.

Le Tableau 4.11 permet de comparer les résultats selon qu'on se contente de faire des classes de X et des variables indicatrices (première colonne), ainsi qu'on le fait très souvent, ou qu'on modélise la relation entre X et Y par un polynôme fractionnaire (deuxième colonne) ou des fonctions splines (troisième colonne).
On note que les écarts entre les OR « non modélisés » et les OR obtenus grâce aux polynômes fractionnaires ou aux fonctions splines sont assez réduits, ce qui, d'une certaine manière, est rassurant, et d'ailleurs assez attendu lorsqu'il y a un nombre assez important de classes (6 ici).
Mais surtout, il faut mettre la comparaison dans l'ensemble de son contexte et ne pas oublier les points suivants :

- L'écart est important pour les deux premières classes d'âge, en particulier pour la largeur de l'intervalle de confiance et la signification statistique des OR.
  Ce résultat est intéressant, car ce sont des femmes jeunes, qui ont probablement des problèmes particuliers de reproduction pour recourir à la FIV à cet âge. La modélisation permet, malgré leurs effectifs réduits, d'attirer l'attention sur le fait que leur risque d'échec est plus grand, alors que l'utilisation de variables indicatrices n'indique rien d'autre que l'imprécision qui découle de ces effectifs réduits.
- Pour les autres catégories d'âge, il y a peu de différence, si ce n'est des intervalles de confiance un peu plus larges avec les variables indicatrices, conformément à la perte de puissance quand on scinde X en classes. Mais il ne faut pas oublier que les centres des classes avec lesquels les calculs d'OR du Tableau 4.11 ont été faits sont précisément les points où la fonction en escalier croise les autres courbes. Ce sont donc les points où les différences entre l'utilisation de variables indicatrices et la modélisation sont minimum. Pour les autres points des courbes, les différences sont plus grandes, assez nettement parfois, en raison de la forme en escalier de la courbe avec des variables indicatrices (Figure 4.30). Cela a de l'importance, car ce sont les valeurs de ces courbes qui seront utilisées dans tout le reste de l'analyse et seront sources d'erreurs ou de biais avec les variables indicatrices (voir § III).
- L'utilisation des variables indicatrices est plus « consommatrice » de variables. Dans l'exemple que j'ai pris avec 6 classes, il faut 5 variables, alors que les polynômes fractionnaires ou les fonctions splines n'en demandent que 2. Cela peut avoir une réelle importance avec des échantillons de taille limitée et s'il y a plusieurs variables X à modéliser.

En conclusion, il y a un certain prix à payer pour présenter les résultats par classes en se servant de la modélisation, car le calcul des odds ratios n'est pas intégré dans les logiciels. Avec les fonctions Stata et R ORcl_pf et ORcl_sp, le prix devient modeste. La présentation des résultats par classes est alors la présentation de choix, car l'alternative est soit de présenter les résultats de façon incompréhensible, soit de s'exposer à une perte de puissance importante, ou à une augmentation du risque de première espèce lorsque X est un facteur de confusion.

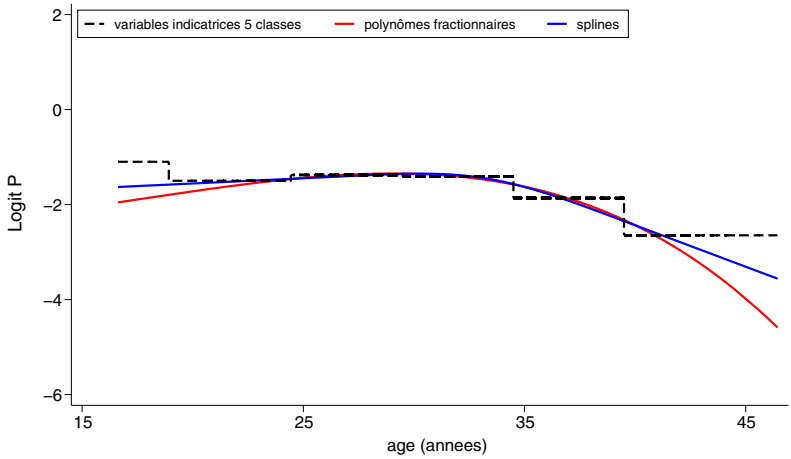


Figure 4.30 : Relation entre l'âge et le succès en FIV avec des variables indicatrices, des polynômes fractionnaires ou des fonctions splines.

XI. Fonctions splines ou polynômes fractionnaires ?

Une comparaison par simulation assez exhaustive des modélisations par polynômes fractionnaires et par fonctions splines a été faite par Binder et al. (Binder H et al., 2013). Elle concerne une variable Y quantitative, mais les auteurs soulignent que ses résultats s'appliquent aussi à des modèles logistiques, de Poisson ou au modèle de Cox. Plusieurs variables X, quantitatives ou qualitatives, ont été incluses dans les simulations.

Les résultats montrent que les deux méthodes ont des performances très proches pour la prédiction de Y :

- Lorsque l'information apportée par les variables X pour expliquer Y (mesurée par le coefficient R^2) est faible (R^2 de l'ordre de 20 %)[6], aucun modèle ne peut espérer prédire Y correctement, mais les PF et les splines donnent des résultats similaires (assez mauvais donc…).
- Lorsque l'information est modérée (R^2 de l'ordre de 50 %), la sélection des variables et la prédiction de Y est meilleure avec des PF et la procédure mfp qu'avec les fonctions splines. Il faut cependant noter que les fonctions splines peuvent modéliser certaines fonctions qui ne peuvent pas l'être raisonnablement par les PF.
- Lorsque l'information apportée par les variables X est bonne (R^2 de l'ordre de 80 %), FP et splines donnent en général des modélisations semblables.

6. L'utilisation de R^2 dans le cas de la régression logistique est discutée, en particulier l'interprétation de sa valeur comme une mesure de la force de l'association (Hosmer DW et al., 2013). Voir aussi chapitre 7, § 1.

Les auteurs ajoutent que, pour la sélection des variables à inclure dans un modèle, les deux méthodes sont aussi assez proches. La comparaison doit cependant être nuancée par le fait que la procédure mfp pour les PF comme les procédures utilisables pour les splines en Stata peuvent être « réglées » pour fixer l'équilibre entre l'adéquation du modèle et le nombre de variables incluses. Mais les deux méthodes de sélection classent les variables à inclure dans des ordres semblables.

Au-delà de ces considérations statistiques, on peut ajouter quelques autres considérations plus « pratiques » qui différencient les polynômes fractionnaires des fonctions splines :

- Les polynômes fractionnaires imposent que X ait des valeurs > 0 pour que les puissances non entières ou négatives et les logarithmes puissent être calculés.

  C'est très souvent le cas en sciences de la vie. Sinon, il faut remplacer X par $X' = X + a$ de façon à ce que X' soit positif. Royston et Altman (Royston P et al., 1994) ont ainsi proposé de prendre $X' = X - x_{(1)} + \delta$, où $\delta = \min\{x_{(i)} - x_{(i-1)}\}$ et $x_{(i)}$ est la i-ème valeur de la variable X par ordre croissant ($x_{(1)}$ est donc négatif ou nul). Ce choix de δ donne de bons résultats, en particulier pour le contrôle du risque de première espèce (Ambler G et al., 2001). Il est intégré dans les modules de commandes des polynômes fractionnaires de Stata.
  Cette contrainte que X soit strictement positif reste cependant un argument fort pour les détracteurs des polynômes fractionnaires, pour qui ce n'est plus vraiment X qu'on modélise s'il y a un changement d'origine de X.

- Il y a plus de flexibilité avec les fonctions splines pour représenter toutes les formes d'associations entre X et Y.
  C'est ce que montrent les Figures 4.15 et 4.24, et que reconnaissent les auteurs des polynômes fractionnaires (Royston P et al., 2008). Cette flexibilité n'est a priori pas nécessaire pour les situations rencontrées habituellement en épidémiologie, mais il faut garder ce point en tête.

- La stratégie de sélection et de modélisation de plusieurs variables est plus convaincante avec les polynômes fractionnaires.
  Bien que débordant un peu du cadre de ce chapitre consacré à la modélisation d'une variable X, ce point me paraît très important, car l'inclusion de plusieurs variables dans un modèle est le cœur de l'analyse statistique en épidémiologie. La stratégie de sélection et de modélisation simultanée de plusieurs variables est « constitutive » des polynômes fractionnaires, alors qu'elle n'existe pas (ou mal) avec les fonctions splines. Il existe cependant d'autres procédures pour choisir les variables à inclure dans un modèle ; elles sont présentées dans le chapitre 5.

- Aspects « logiciels »
  Il n'y a qu'un module dans les logiciels Stata, R et SAS pour utiliser les polynômes fractionnaires : mfp, accompagné dans Stata de quelques autres modules

pratiques (tels que fp ou fracplot). À l'inverse, il y a de nombreux modules (ou package d'utilisateurs) dans ces logiciels pour les fonctions splines, notamment dans R, où ils sont très peu hiérarchisés. Leurs différences, pas toujours faciles à cerner, portent le plus souvent sur la construction des fonctions splines de base. Il n'est pas sûr qu'une telle abondance soit un avantage pour l'utilisateur…

L'idée générale de cette comparaison rapide peut être résumée en disant qu'on obtient (en général) des résultats similaires avec les deux types de modélisation, mais que les polynômes fractionnaires sont plus faciles à utiliser, bien que moins couramment présents dans les logiciels. C'est un peu une question de goût.
Et c'est de toute façon mieux que de transformer une variable quantitative en classe !

XII. Annexes

XII.1. Logiciels

XII.1.a. Polynômes fractionnaires

Les modules de PF ont surtout été développés dans Stata, mais ils existent aussi dans SAS et dans R, avec parfois des options ou des valeurs par défaut différentes (Sauerbrei W et al., 2006).

Pour SAS, la macro est téléchargeable depuis https://mfp.imbi.uni-freiburg.de/software.

Pour R, le package est mfp (https://cran.r-project.org/web/packages/mfp/mfp.pdf). Deux articles de Benner donnent de précieux exemples et explications de son utilisation (Benner A, 2005, Benner A, 2022).

Pour Stata, les commandes officielles sont fp et mfp. Dans les versions antérieures, fp était remplacé par fracpoly, qui est toujours disponible. Ces commandes sont accompagnées de plusieurs autres, dont on trouvera la liste sur https://mfp.imbi.uni-freiburg.de/software.

XII.1.b. Fonctions splines

Les commandes pour créer des fonctions splines sont assez nombreuses avec des fonctions splines de base qui peuvent être différentes, ce qui ne facilite pas le choix, car les documentations sont souvent sibyllines. Certaines commandes intègrent la construction des fonctions splines et leur inclusion dans le modèle logistique (ou linéaire) qu'on veut utiliser.

Pour les splines cubiques restreintes, je privilégie ici les commandes qui créent des fonctions splines dont la première est la variable X elle-même, ainsi que je l'ai décrit dans le § IX.6.

Pour SAS, ma connaissance est trop parcellaire, je renvoie donc à la documentation de ce logiciel.

Pour R, l'article de Perperoglou et al. (Perperoglou A et al., 2019) fait une synthèse des packages et fonctions existants. Il ne contient cependant pas la commande que j'ai utilisée dans le § IX.6 pour les splines cubiques restreintes, qui est rcspline.eval(), issue du package Hmisc.
Les splines linéaires peuvent être obtenues avec la commande lspline() du package lspline.

Pour Stata, la commande makespline permet de construire des splines linéaires, des B-splines, des splines cubiques et des splines cubiques restreintes dont la première fonction de base est la variable X. Dans les versions antérieures de Stata, la commande mkspline, toujours disponible, avait des fonctions similaires. D'autres commandes d'utilisateurs, plus anciennes, existent aussi.

XII.2. Représentation graphique des données observées avec la courbe modélisée

Comme je l'ai souligné dans le § X.2, il n'est pas toujours nécessaire, ni même souhaitable, de faire figurer les données observées à côté de la représentation graphique de la relation entre X, variable quantitative, et Y, variable dichotomique. Si on veut cependant le faire, par exemple pour conforter ou interpréter la modélisation obtenue ou pour satisfaire la demande de reviewers, il n'y a pas de solution qui s'impose, car la probabilité P de Y à X fixé ou logit P, qui est la quantité modélisée, ne peuvent pas être observés au niveau individuel.

Dans ce chapitre, j'ai choisi de grouper X en catégories de petite amplitude (un an pour l'âge de la femme dans les données de FIV ou une semaine pour la durée de grossesse dans le § II), de façon à avoir un nombre assez grand de catégories pour une représentation suffisamment précise des observations. On peut alors calculer la valeur de logit P pour chacune d'entre elles, sauf celles pour lesquelles P = 0 ou P = 1. Pour ces catégories, par convention, les points sont représentés par des losanges en bas ou en haut du graphique, accompagnés du nombre de sujets correspondant (§II, Figure 4.5). Avec cette méthode, on ne peut pas éviter la perte de la nature quantitative de X. Cette perte ne concerne que la représentation graphique ; l'analyse, elle, est faite sur les données individuelles. La perte n'est cependant que partielle, car la variable en classes est qualitative ordonnée avec un grand nombre de classes. Si on voulait conserver la nature quantitative de X en ne faisant pas de classes (et donc en gardant l'âge en jours), il n'y aurait quasiment que des observations avec P = 0 ou P = 1, car presque tous les sujets ont des âges différents au jour près, et donc aucune représentation graphique ne serait possible avec cette méthode (voir Figure 4.4a).

Le logiciel Stata procède autrement, en s'inspirant de ce qu'on pourrait faire si Y était quantitative. Le principe de la méthode est intéressant à comprendre, car il intervient aussi pour l'étude de l'adéquation du modèle (chapitre 7). La commande correspondante est fracplot, qu'on peut utiliser à la suite de mfp.

Le principe est que l'ordonnée du point « observé » qui doit figurer sur le graphique pour la valeur x_0 de X a pour ordonnée $\eta(x_0) + d(x_0)$, où $\eta(x_0)$ est la valeur prédite de Y par le polynôme fractionnaire et $d(x_0)$ est le résidu de la déviance. L'origine de la méthode se comprend si on l'applique à la régression linéaire (Y quantitative), dans laquelle le résidu de la déviance est, par définition, égal à $y_0 - \eta(x_0)$, de sorte que $\eta(x_0) + d(x_0) = y_0$ et qu'on retrouve ainsi exactement le nuage de points des (x_i, y_i).
Dans le cas de la régression logistique, les choses sont plus compliquées, en raison de l'expression du résidu de la déviance. Pour le définir, on groupe ensemble par « profils »[7] les sujets qui ont la même valeur de X. Tous les sujets d'un même profil ont la même valeur prédite de logit P par le modèle logistique. Le résidu de la déviance pour le profil j est égal à : $d_j = \pm\left\{2\left[n_{1j}\ln\left(\frac{n_{1j}}{m_j\hat{P}_j}\right) + n_{0j}\ln\left(\frac{n_{0j}}{m_j\left(1-\hat{P}_j\right)}\right)\right]\right\}^{1/2}$, où $\hat{P}_j$ est le pourcentage prédit de succès pour le profil j avec le polynôme fractionnaire et où ± est le même signe que celui de $n_{1j} - m_j\hat{P}_j$, avec m_j le nombre de sujets dans le profil j et n_{1j} le nombre de succès.
Pour les profils tels que $n_{1j} = 0$ (uniquement des succès), on prend $d_j = -\sqrt{2m_j\left|\ln\left(1-\hat{P}_j\right)\right|}$
Pour les profils tels que $n_{0j} = 0$ (uniquement des échecs), on prend $d_j = \sqrt{2m_j\left|\ln\left(\hat{P}_j\right)\right|}$
Pour un même profil j, tous les sujets sont représentés par un même point observé de coordonnées $(x_j, \eta(x_j) + d(x_j))$, c'est-à-dire ici $(x_j, \text{logit } P_j + d(x_j))$. Il est normal qu'il n'y ait qu'un seul point, car tous ces sujets ont le même x_j et la même valeur prédite logit P_j. Ils sont donc indifférenciables pour la représentation graphique qui nous occupe.
Ce procédé a l'avantage de fournir un nuage de points en gardant les valeurs individuelles de X sans faire de classes. Sa construction explique l'intitulé de l'axe des ordonnées (« Prediction + residual of x »), qu'il est souvent utile de changer si on veut être compris (et éviter des questions qui demanderaient tout le développement précédent pour y répondre…). Cependant, pour une variable comme l'âge non arrondi en années des données de FIV, la plupart des profils ont un seul sujet. On observe alors que le nuage de points ressemble à deux lignes quasiment parallèles au polynôme fractionnaire (Figure 4.31). Non seulement cela ne donne pas d'indication visuelle claire sur l'adéquation de la courbe aux données, mais c'est incommunicable sans une longue explication très technique et (donc ?) peu convaincante.
On remarque aussi sur la Figure 4.31 que le nuage de points n'est pas tout à fait le même selon le polynôme fractionnaire retenu pour la modélisation. C'est général ; ce serait la même chose avec des profils comprenant un plus grand nombre de sujets et avec d'autres polynômes fractionnaires. Cela vient de ce que la méthode calcule une pseudo-ordonnée des points observés qui dépend de la modélisation (voir § XII.2). Et, bien sûr, cela rajoute une difficulté pour présenter ces résultats avec des « observations » qui ne sont pas vraiment observées.

7. La notion de « profils » et le résidu de la déviance sont discutés plus en détail dans le chapitre 7.

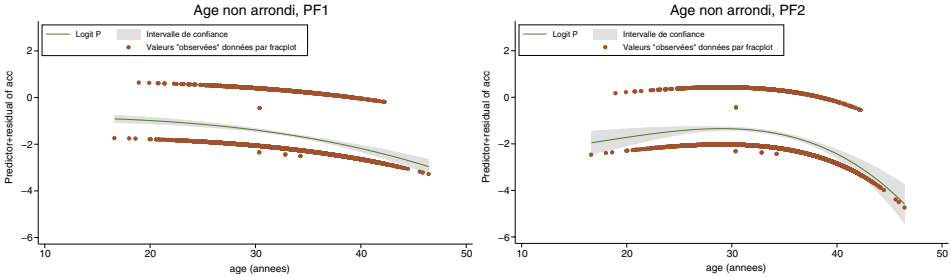


Figure 4.31 : Courbe et nuage de points de la relation entre l'âge et le succès en FIV modélisée par un FP1 et un FP2.

Les inconvénients de la procédure fracplot que je viens de décrire ne doivent pas empêcher de l'utiliser (si on dispose de Stata), car elle est réellement commode pour avoir une première idée de la forme de la courbe, mais il me semble préférable de ne pas faire apparaître les valeurs « observées » ainsi obtenues. Si on veut des points observés, il faut le faire soi-même, par exemple avec les programmes logit_crb.

Chapitre 5

Choix des variables à inclure dans un modèle logistique

• • •

Le choix des variables à inclure dans un modèle logistique (ou plus généralement dans un modèle multivarié) est le cœur de l'analyse des enquêtes épidémiologiques. Il est très lié à la notion de biais de confusion qui a été présentée dans le § V du chapitre 1. Le choix des variables inclut aussi celui de leur codage (voir chapitre 3) ou de la façon de les modéliser lorsqu'il s'agit de variables qualitatives ordonnées ou quantitatives (voir chapitre 4). J'ai choisi ici de présenter ces différents aspects dans des chapitres différents pour mieux les détailler, mais l'ensemble peut être inclus dans une stratégie générale de choix des variables et de leur modélisation (Slama R et al., 2005).

Le choix des variables est une question difficile, toujours discutée aujourd'hui et sans solution uniformément meilleure que les autres (Greenland S et al., 1999, Harrell FE, 2001, Sauerbrei W et al., 2007, Greenland S, 2008, Royston P et al., 2008, Walter S et al., 2009, Greenland S et al., 2015). Le plus important, finalement, est d'avoir un fil directeur et quelques principes généraux pour guider ce choix et surtout d'expliciter clairement ce qu'on fait, de façon à ce que les lecteurs puissent comprendre et se faire leur propre idée de la pertinence du choix (Collins GS et al., 2015).

Un dernier paragraphe de ce chapitre est consacré à la question, tout à fait à part, des variables qui doivent être incluses du fait de la structure de l'échantillon.

I. Principes généraux

I.1. Pas de modèle universellement meilleur

Choisir les variables à inclure dans un modèle logistique suppose de définir des critères précisant ce qu'est le « meilleur » modèle.

Or, il n'y a pas de modèle qui soit universellement meilleur. Les objectifs d'un modèle ne sont pas les mêmes selon qu'il s'agit de faire une prédiction ou un pronostic ou selon qu'il s'agit d'une recherche à visée étiologique (Moons KG et al., 2009). Même si pronostic et étiologie ont des liens forts, ce ne sont pas les mêmes caractéristiques du modèle qui sont privilégiées dans un cas ou dans l'autre (voir (Royston P et al., 2008) §1.7.2).

Dans une première approche, on peut dire que, lorsqu'il s'agit de prédire ou de pronostiquer (une maladie, une guérison, une grossesse…), on privilégie l'adéquation du modèle avec des indicateurs de qualité tels que la sensibilité et la spécificité à prédire l'événement d'intérêt. La limitation du nombre de variables n'est pas toujours un objectif prioritaire.

Lorsqu'il s'agit d'étiologie, situation plus fréquente en épidémiologie analytique, discipline sur laquelle ce chapitre est centré, l'objectif est d'identifier les causes de la maladie ou, de façon plus réaliste, ses facteurs de risque importants, et de comprendre et de quantifier leur effet. Il faut donc que les variables scientifiquement importantes soient prises en compte, que les phénomènes de confusion soient contrôlés, et que les résultats soient suffisamment stables pour être extrapolables. Ajoutons qu'il ne faut pas que le modèle soit une « boîte noire », même performante. Il faut que ses paramètres aient une interprétation concrète dans le mécanisme connu

ou supposé menant de la cause potentielle à la maladie. S'il s'agit de variables quantitatives, il faut que la forme de la relation avec la maladie soit correctement modélisée et interprétable. Par exemple, est-elle croissante ou présente-t-elle un plateau ?

Dans ce cadre général où on cherche à répondre aux objectifs de l'étude en prenant en compte au mieux les phénomènes de confusion et en gardant la possibilité d'interpréter ses coefficients, le choix des variables repose sur un « mélange » :

- ✓ de connaissances scientifiques indépendantes des données de l'enquête analysée. Ce qu'on appelle en anglais *subject-matter knowledge* et qui repose sur une bibliographie la plus exhaustive possible et sur des discussions scientifiques avec des spécialistes des pathologies et des expositions étudiées. Ces connaissances sont parfois (souvent ?) limitées, d'autant plus qu'elles portent sur des domaines où, justement, on en manque et où des enquêtes et des études sont réalisées pour les améliorer.
- ✓ d'utilisation de méthodes statistiques appliquées aux données (*data-driven procedures*). On en verra des exemples plus loin. Ces méthodes doivent être utilisées avec un certain recul, de façon à ce que leurs résultats ne soient pas trop dépendants de la particularité des données analysées et soient donc suffisamment généralisables.
- ✓ d'expérience et de bon sens. Ce sont des aspects à ne pas négliger, bien qu'ils soient souvent passés sous silence. Ils expliquent qu'une bonne analyse statistique ne peut heureusement pas être automatisée (du moins, pas entièrement). L'expérience est un des privilèges de l'âge, mais le bon sens n'attend pas le nombre des années !

I.2. Équilibre entre trop ou trop peu de variables prises en compte

Finalement, même s'il y a là une forme d'évidence, il faut trouver un équilibre entre trop ou trop peu de variables prises en compte dans le modèle final. Le fait qu'on puisse prendre en compte trop peu de variables est une notion intuitive que j'essaierai de préciser plus loin. Mais il ne faut pas non plus inclure trop de variables dans un modèle multivarié.

I.2.a. Trop de variables

On pourrait envisager d'inclure toutes les variables dans un modèle en se disant qu'on ne risque pas ainsi d'exclure un facteur de confusion important pour l'estimation de l'association entre un facteur de risque et la maladie. Du moins, parmi les variables figurant dans l'enquête. C'est souvent infaisable à cause de la taille de l'échantillon, qui limite, pour des raisons méthodologiques, le nombre de variables qu'il est possible d'inclure (voir § II). De toute façon, pratiquement tous les auteurs affirment que cette stratégie entraînerait des estimations instables et biaisées. Bien que ces affirmations soient rarement étayées, on peut souligner quelques

inconvénients liés à l'inclusion d'un trop grand nombre de variables dans un modèle de régression :

- ✓ Il y a une perte de puissance. Sur le plan théorique, la précision de l'estimation de chaque paramètre diminue avec le nombre total de paramètres à estimer. Cela reste cependant souvent marginal en pratique sur le plan quantitatif. Le problème majeur est très pratique. Il est lié aux données manquantes qui augmentent le nombre de sujets exclus quand le nombre de variables augmente (du moins si on ne recourt pas à des méthodes d'imputation).
  Par exemple, dans l'étude du lien entre GEU et antécédent d'IVG, l'OR brut est estimé sur 1725 sujets. Après prise en compte de 9 variables qui sont des facteurs de confusion potentiels (âge, tabagisme, niveau d'études, activité professionnelle, antécédents de fausses couches, de GEU, d'infection, de pathologie tubaire, induction de la grossesse), l'OR ajusté n'est estimé que sur les 1187 sujets qui ne présentent pas de données manquantes (soit une perte de 31 % de l'effectif). Et on est loin d'avoir inclus toutes les variables !
- ✓ Il y a un risque de surajustement, qui conduit à ce que l'OR ajusté associé à l'exposition d'intérêt se rapproche artificiellement de 1 et/ou devienne très imprécis. Cela peut se comprendre en se rappelant que l'ajustement vise à ce que « tout se passe comme si » les exposés et les non-exposés étaient semblables pour tous les facteurs autres que l'exposition. Si certains de ces facteurs sont très liés à l'exposition (par exemple, en caractérisant certains comportements des sujets), il se peut qu'à force d'être semblables pour tous ces facteurs, les sujets le soient aussi vis-à-vis de l'exposition. L'estimation de l'OR ajusté a alors une grande variance. Elle n'est pas forcément biaisée, mais est tellement imprécise que la valeur obtenue sur un échantillon est souvent éloignée de la vraie valeur par fluctuations aléatoires.
- ✓ Les résultats sont instables et donc moins interprétables. Le terme « résultats instables » est un peu vague. Il doit être compris ici comme faisant référence à des résultats qui sont susceptibles de changer si les calculs sont refaits sur une partie de l'échantillon (par exemple en enlevant aléatoirement une partie des sujets), ou en retirant ou en ajoutant un petit nombre de variables qui peuvent paraître d'importance secondaire. Cette instabilité est souvent liée à une forme de redondance des variables qui rend instable l'estimation de chaque coefficient et même la sélection de telle ou telle variable. Dans les cas les plus marqués, on parle aussi de colinéarité (voir § IV.3). Les résultats peuvent alors devenir difficiles à interpréter.

I.2.b. Trop peu de variables, oubli de certaines variables

Ne pas prendre en compte assez de facteurs dans un modèle multivarié est ce qu'on craint a priori le plus. On risque en effet de ne pas suffisamment éliminer les biais de confusion et donc d'avoir une estimation biaisée de l'association entre l'exposition et la maladie. On parle de confusion résiduelle (voir § IV.6).

I.3. Deux étapes

Partant d'un ensemble de données recueillies dans une enquête qui peut compter plusieurs dizaines, voire quelques centaines de variables, on procède en général en deux étapes (Royston P et al., 2008), qui seront détaillées dans les paragraphes suivants.

1. Choix des variables « candidates »
   L'objectif est de faire un premier tri parmi l'ensemble des variables disponibles, mais aussi de décider comment elles seront codées (en catégories par exemple) ou combinées (Blettner M et al., 1993). On ne trouve souvent que des traces de cette étape, pourtant essentielle, dans les publications des résultats.
2. Sélection des variables à inclure dans le modèle final
   Lorsqu'il reste encore trop de variables à l'étape précédente, on poursuit la sélection parmi les variables candidates pour aboutir à un modèle qui soit le plus concis possible. Cet objectif de parcimonie du modèle final est justifié par des raisons statistiques, mentionnées plus haut (§ I.2.a), mais aussi par la préoccupation qu'il reste interprétable. On recourt souvent à des procédures de sélection basées sur des procédures statistiques (Royston P et al., 2008), bien qu'elles aient leurs détracteurs (Sribney B, 1998, Harrell FE, 2001, Rothman KJ et al., 2008).

II. Nombre maximum de variables

Comme cela a été indiqué plus haut, un trop grand nombre de variables dans un modèle multivarié pose des problèmes de surajustement, de stabilité de l'estimation des coefficients, de redondance entre les variables et de difficulté pour les sélectionner. Cela reste vrai même avec un grand nombre de sujets. Quand le nombre de sujets ou le nombre de malades est plus petit, ces problèmes sont accentués, mais s'ajoutent aussi des contraintes de méthodologie statistique. Le nombre de variables est limité par le respect des propriétés de la méthode du maximum de vraisemblance pour estimer les paramètres du modèle. Le nombre de sujets doit en effet être assez grand pour que les conditions asymptotiques de cette méthode soient satisfaites et qu'elle donne des résultats non biaisés (estimations, intervalles de confiance et tests). Voir aussi chapitre 2, § II.

La limite habituellement retenue porte sur le nombre d'« événements par variable » (EPV). Le nombre d'événements est le nombre de malades (ou de non-malades s'il est plus petit).

La règle générale pour la régression logistique est EPV ≥ 10 (Harrell FE, Jr. et al., 1985, Peduzzi P et al., 1996, Steyerberg EW et al., 2001). Elle reste discutée (van Smeden M et al., 2016). Certains auteurs affirment que, dans la plupart des cas, EPV ≥ 5 est suffisant (Concato J et al., 1995, Vittinghoff E et al., 2007). D'autres expliquent que EPV ne résume pas tout. Il faut aussi prendre en considération la corrélation entre les variables, ou les coefficients des variables dans le modèle : s'ils sont élevés, il peut

être nécessaire que EPV soit au moins égal à 20 (Courvoisier DS et al., 2011, Steyerberg EW et al., 2011).

> Remarque
> Dans le cas du modèle de Cox, la règle EPV ≥ 10 est aussi proposée (Peduzzi P et al., 1995). Pour la régression linéaire, la question se pose différemment, puisqu'il n'y a pas d'événement en oui/non à proprement parler. On s'intéresse alors au nombre de sujets par variable (SPV), et il semble qu'il suffise qu'il soit supérieur à 2 (Austin PC et al., 2015).

Dans le cas de l'enquête sur les facteurs de risque de la GEU qui me sert d'exemple, il y a 574 cas. Si on applique la règle EPV ≥ 10, il ne faut donc pas dépasser 57 variables dans le modèle (ce qui laisse de la marge…).

Lorsque EPV est trop petit, il y a deux possibilités principales :

- ✓ Utiliser des méthodes d'estimation dites « exactes » adaptées aux petits échantillons (cf. par exemple Mehta CR et al. (1995) pour le modèle logistique). Elles sont non paramétriques, c'est-à-dire construites de manière à ne pas avoir besoin des hypothèses de distribution statistique nécessaires à la méthode d'estimation du maximum de vraisemblance. Elles résolvent (en partie du moins) les problèmes statistiques d'estimation, mais elles ne résolvent pas l'instabilité des modèles et celle de la sélection des variables.
- ✓ Avoir recours au score de propension (Rosenbaum PR et al., 1983, D'Agostino RB, Jr., 1998, Joffe MM et al., 1999, Cepeda MS et al., 2003, Sturmer T et al., 2006, Austin PC, 2007, Kwiatkowski F et al., 2007, Garrido MM et al., 2014, Lu B et al., 2018). On part d'un modèle $\text{logit } P = \alpha + \beta E + \sum \beta_i X_i$, où E est l'exposition d'intérêt (en 0/1) et où il y a un grand nombre de variables X_i. Le principe est de construire un modèle logistique de E en fonction des X_i : $\text{logit } E = \alpha' + \sum \beta'_i X_i$, d'en déduire une nouvelle variable (le score) $S = \sum \beta'_i X_i$, et d'inclure cette variable dans le modèle logistique de la maladie M en fonction de E ($\text{logit } P = \alpha' + \beta' E + \gamma S$), ou bien d'apparier les sujets sur S pour analyser l'association entre E et M, ou bien de faire une analyse pondérée (D'Agostino RB, Jr., 1998, Lunceford JK et al., 2004). Pour dire les choses rapidement, cette méthode permet que la distribution des facteurs de confusion potentiels X_i soit semblable chez les exposés et les non-exposés. Un peu comme si l'exposition avait été tirée au sort. C'est pour cette raison que le score de propension est parfois présenté comme une méthode d'analyse « causale », malgré le côté « trompeur » ou « formule marketing » que contient cette expression.

  Si le nombre d'individus exposés est plus grand que le nombre de malades, ce qui est fréquent, la régression logistique $\text{logit } E = \alpha' + \sum \beta'_i X_i$ n'est pas (ou est moins) limitée en nombre de variables et la régression initiale n'a plus qu'une variable, S, au lieu des variables X_i.

III. Choix des variables « candidates »

III.1. Au moment du protocole

La réflexion sur les variables qui figureront dans les analyses multivariées doit être faite dès le protocole de l'étude. Il faut en effet – c'est une évidence – que ces variables apparaissent dans le ou les questionnaires.

Cette réflexion repose sur la connaissance scientifique de la question : les aspects cliniques de la maladie étudiée, ses facteurs de risque connus, ce qu'on sait (ou ce qu'on suppose) des mécanismes d'action des expositions auxquelles on s'intéresse.

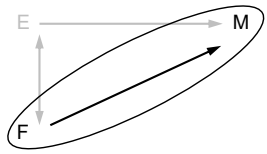
Cela passe par une revue bibliographique complète et détaillée, par le choix de résultats déjà connus qu'on veut confirmer ou qui serviront de « contrôle qualité » des données recueillies.

Il faut bien sûr inclure aussi les variables qui serviront à tester les nouvelles hypothèses qui sont à l'origine de l'enquête.

III.2. Après le recueil des données, examen systématique des variables

C'est après le recueil des données, au moment de commencer leur analyse, que le choix des variables candidates susceptibles de figurer dans un modèle multivarié est principalement fait.

De façon générale, les variables que l'on retient sont les facteurs de risque connus de la maladie et les facteurs qui lui sont liés dans l'échantillon étudié. Cela revient donc à s'appuyer sur le schéma triangulaire « classique » ci-contre en considérant que F est un facteur de confusion potentiel dès qu'il s'agit d'un facteur lié à la maladie.



L'objectif est de faire une première sélection, suffisamment large pour ne pas laisser de côté un facteur qui pourrait apporter de la confusion, la sélection finale des variables candidates se poursuivant à l'étape suivante (§ IV).

Comme à l'étape précédente, les connaissances scientifiques, la revue bibliographique et les mécanismes d'action supposés sont toujours présents. En pratique, on procède à une analyse univariée détaillée et complète qui permet d'étudier de façon systématique les liens entre les différentes variables de l'enquête et la maladie (Heinze G et al., 2024).

On commence par les facteurs de risque reconnus de la maladie. Ces facteurs seront de toute façon pris en compte lors le choix des variables candidates, car leur présence dans l'analyse des facteurs de risque ou dans la prédiction d'une maladie est requise. On ne voit pas, par exemple, comment on pourrait étudier une maladie pulmonaire sans tenir compte du tabagisme (à moins, peut-être, d'une argumentation solide et convaincante). Cette vérification du lien entre les facteurs de risque reconnus et la maladie permet aussi de s'assurer qu'on les retrouve bien dans les données de

l'enquête. C'est une sorte de contrôle de qualité et un gage de crédibilité pour les résultats ultérieurs.
La sélection des variables issue de ces analyses univariées systématiques doit être large. On retient les variables dont le test d'association avec la maladie a un degré de signification inférieur à 20 % ou 25 %. Des simulations ont en effet montré, d'une part, que les variables ainsi sélectionnées ne sont pas trop nombreuses lorsqu'il n'y a pas de confusion et, d'autre part, que se limiter au degré de signification habituel de 5 % laisse de côté trop de facteurs de confusion potentiels (Dales LG et al., 1978, Mickey RM et al., 1989, Maldonado G et al., 1993, Sun GW et al., 1996, Royston P et al., 2008). Précisons qu'il ne s'agit pas ici de changer le seuil de signification habituel des tests, mais de choisir un seuil de sélection des variables dont le statut de facteur de risque ou de confusion sera discuté pour l'établissement du modèle final.
Cette procédure est critiquée par certains auteurs (Sun GW et al., 1996, Harrell FE, 2001), Une des raisons invoquées est la multiplication des tests qu'elle implique, ce qui conduirait à des intervalles de confiance des paramètres trop optimistes. L'autre argument de ces auteurs est que cette méthode de sélection est biaisée : elle tend à sur-sélectionner les variables dont le coefficient *estimé* d'association avec la maladie est grand et à sous-sélectionner les variables dont le coefficient *estimé* d'association avec la maladie est petit, alors qu'il faudrait se baser sur les coefficients *vrais*. Ces critiques ont un réel fondement théorique, mais restent justement trop théoriques. Elles ne proposent pas d'alternative, si ce n'est de ne s'appuyer que sur la littérature et les connaissances des mécanismes d'action, ce qui n'est pas toujours possible. De plus, il ne s'agit pas à proprement parler de faire des tests successifs dont on voudrait contrôler le risque d'erreur individuel, mais d'une règle pratique de sélection des variables. Le risque d'erreur qu'il faut contrôler est celui d'oublier des variables qui devraient être incluses. Les simulations réalisées indiquent que c'est bien le cas avec la règle de 20 % à 25 % (Mickey RM et al., 1989, Maldonado G et al., 1993).

J'ajoute que cet examen systématique des variables a aussi le mérite d'obliger à consacrer du temps à l'examen des données elles-mêmes. Ce temps est indispensable pour bien les connaître, avoir vu (au moins une fois !) leur distribution, identifier les variables ayant des données manquantes, repérer les erreurs de saisie ou les incohérences. Ce temps, qui n'apparaît pas et n'est jamais mentionné dans les rapports, mémoires ou thèses et encore moins dans les publications, est un temps long. Il peut s'étendre sur plusieurs semaines, selon le nombre de variables et la connaissance qu'on a du sujet, mais l'expérience montre que ce n'est pas du temps perdu.

III.3. Choix du codage des variables

L'examen systématique des variables conduit « naturellement » à s'interroger sur leur codage. Le chapitre 3 est consacré à cette question et à l'interprétation des coefficients qui en résultent. Quelques phrases pour résumer. Dans certains cas, la question ne se pose pas : par exemple, pour les variables dichotomiques, si ce n'est qu'il est vivement conseillé de les coder 0/1. Par rapport à un codage avec d'autres valeurs

(par exemple 1/2), cela facilite les analyses ultérieures et évite des erreurs. Pour les variables ayant plus de deux classes, le regroupement de catégories peut se discuter, notamment si les classes sont nombreuses ou si certaines d'entre elles ont des effectifs petits. Si les classes sont ordonnées, le choix des valeurs numériques à attribuer à chaque classe doit être réfléchi. Quant aux variables quantitatives, le mieux est de les conserver sous leur forme initiale et de les modéliser avec des méthodes appropriées (voir le chapitre 4 qui leur est consacré). Si elles sont malgré tout découpées en classes, il est souvent pertinent de faire de l'ordre de cinq classes ; les remarques qui précèdent sur les variables qualitatives ordinales s'appliquent encore.
Il est fréquent qu'au-delà du codage de chaque variable il soit nécessaire de construire des variables synthétiques à partir de plusieurs variables initiales. Cela peut permettre de limiter les phénomènes de colinéarité dont on reparlera plus loin (§ IV.3), mais aussi de construire des variables pertinentes à partir du questionnaire de l'enquête. C'est par exemple ce qui a été fait dans l'enquête sur les facteurs de risque de GEU pour la variable « antécédent de salpingite ». Le questionnaire issu du dossier médical contenait trois variables dans la rubrique « antécédents d'infections gynécologiques » : nombre d'infections basses ayant fait l'objet d'un traitement médical, nombre de salpingites sans certitude, nombre de salpingites certaines (prouvées à la cœlioscopie). La variable « antécédent de salpingite » a été construite en considérant les réponses « oui » aux deux dernières questions, mais sans prendre en compte la première question.
Les choix concernant le codage pourront être rediscutés dans la suite de l'analyse, lorsque d'autres variables auront été prises en compte.

III.4. Recherche des interactions à prendre en compte

La sélection des interactions entre les variables étudiées est beaucoup moins formalisée et discutée que celle des variables elles-mêmes.
Je mets à part les études dont le but est précisément d'analyser une interaction, par exemple celles qui portent sur les interactions entre gènes et environnement. Dans les autres cas, la stratégie est différente de celle que j'ai présentée pour les variables à prendre en compte au titre d'un risque potentiel de confusion. On ne procède pas à une recherche systématique, mais on ne teste l'existence d'une interaction qu'au cas par cas, lorsqu'on a des hypothèses a priori sur un mécanisme d'action qui peuvent l'expliquer. Il y a plusieurs raisons à cela :

- ✓ L'interaction entre deux facteurs concerne leurs mécanismes d'action sur la maladie, pas la correction d'un biais comme la confusion. La multiplication des tests se met donc à avoir une importance beaucoup plus grande. Il y aurait beaucoup de tests à faire, puisque pour p variables, il y a $p(p-1)/2$ interactions possibles (par exemple, 190 interactions pour 20 variables). Il y a donc un risque important (pratiquement 100 %) de trouver à tort une ou plusieurs interactions significatives qu'il sera en pratique très difficile d'ignorer, même si on ne parvient pas à l'interpréter (ou à les interpréter).

- ✓ La notion statistique d'interaction ne correspond pas toujours à l'existence d'un phénomène biologique. Elle peut notamment varier selon la mesure d'association, quantitativement ou même sur le plan de la signification statistique (voir, par exemple, chapitre 1, § V.2.b).
- ✓ L'interprétation des résultats devient compliquée. Ce n'est pas une raison en soi pour ne pas se préoccuper des interactions éventuelles, mais on doit aussi se dire qu'une analyse dont on ne sait pas interpréter les résultats n'a pas une grande utilité.
- ✓ La prise en compte d'une interaction revient à faire des analyses séparées par catégories de la variable d'interaction (voir chapitre 3, § VI.2). Cela conduit à une perte de puissance pour l'étude de l'exposition d'intérêt principal et ne se justifie donc que pour les variables pour lesquelles l'interaction peut être intéressante et informative.

Finalement, la décision de prendre en compte une interaction est un mélange de considérations statistiques et de réflexions sur la nature de la question. Avec ce que j'ai dit ci-dessus, on comprend qu'on peut quand même ajuster s'il y a une interaction, à condition qu'on puisse donner un sens à l'effet moyen entre les strates de la variable d'interaction. Même si on a testé l'interaction (parfois par erreur, faiblesse ou esprit de système), le fait qu'elle soit significative ne doit donc pas être un blocage pour poursuivre l'analyse sans l'inclure dans les modèles.
La règle qu'on peut retenir est de ne rechercher et de ne conserver que le minimum d'interactions possible. Parfois aucune…

IV. Problèmes à considérer lors de la sélection des variables candidates

Au-delà de la procédure et des considérations générales présentées dans le paragraphe précédent, un certain nombre de questions doivent être prises en compte et une réponse doit leur être apportée lors de la sélection des variables candidates. Comme on va le voir, une partie des réponses font appel aux connaissances bibliographiques, mais aussi, et parfois surtout, à un examen minutieux des données analysées.

IV.1. Trop de données manquantes, distribution très déséquilibrée

Il faut se poser la question de conserver ou pas des variables ayant un grand nombre de données manquantes. De façon générale, les valeurs manquantes peuvent être sources de biais si elles ne le sont pas complètement au hasard (Rubin BB, 1987).
Il n'y a pas de règle précise, mais la question se pose de conserver une variable particulière dans l'analyse au-delà d'un certain pourcentage de données manquantes, que je ne me hasarderais pas à fixer, mais qui peut être de l'ordre de 15 % à 25 %.

Lorsque plusieurs variables sont concernées, ce qui est le cas général, même si la proportion de données manquantes est faible pour chaque variable, leur cumul peut réduire le nombre de sujets sur lesquels porte l'analyse des sujets complets et donc la puissance statistique. On peut ainsi perdre de l'ordre de 30 à 40 % des sujets pour l'analyse multivariée, ce qui pose alors une question de validité des résultats et de la possibilité de les généraliser.

On peut envisager deux façons de pallier cet inconvénient ; la première est une bonne méthode, mais l'autre pas :

- ✓ Il est possible d'imputer les valeurs manquantes. Il y a plusieurs méthodes, les meilleures prenant en compte la variabilité supplémentaire introduite par l'imputation (Little RJA, 1992, Little RJA et al., 2002, Molenberghs G et al., 2007, White IR et al., 2011). Les hypothèses faites avec l'imputation sont moins fortes qu'avec l'analyse restreinte aux sujets sans données manquantes (analyse en cas complets), qui suppose que les données manquent complètement au hasard (*MCAR missing completely at random*) (Rubin BB, 1987). Pour l'imputation, il « suffit » que les données soit MAR (*missing at random*), c'est-à-dire MCAR conditionnellement aux variables utilisées pour imputer.
- ✓ Lorsque la variable qui a des données manquantes est en classes, on peut envisager d'ajouter une catégorie supplémentaire « inconnu ». Cela n'est pas recommandé. Certes, cela évite de perdre des sujets, mais les résultats obtenus sont biaisés (Greenland S et al., 1995), parfois de façon importante, même si les données manquent au hasard.

La question de conserver ou pas une variable se pose aussi lorsqu'il s'agit d'une variable en classes dont la distribution est très déséquilibrée. Par exemple, dans des études sur les expositions professionnelles, il peut y avoir très peu de femmes dans certains secteurs de la métallurgie ou très peu d'hommes dans certains secteurs de la santé. L'ajustement sur le genre reviendrait pratiquement à éliminer de l'analyse la catégorie la plus petite. Cela reviendrait donc à restreindre l'analyse à l'autre catégorie en croyant ajuster.
De façon analogue, l'inclusion d'une variable quantitative ayant une variabilité très réduite nécessite des hypothèses fortes sur la forme de la relation de cette variable avec la maladie, même si la modélisation de cette relation a été choisie avec soin.

IV.2. Erreurs de mesure

Rappelons que lorsqu'il y a des erreurs de mesure sur un facteur de risque potentiel X, cela induit un biais sur l'estimation du coefficient associé à cette variable. Si l'erreur de mesure est non différentielle (c'est-à-dire identique chez les malades et les non-malades), le biais va dans le sens de rapprocher le coefficient de 0 (ou d'augmenter sa variance s'il s'agit d'une variable quantitative) (Bouyer J et al., 1993). Si l'erreur de mesure est différentielle, le sens du biais n'est pas toujours le même, ce qui rend plus difficile l'interprétation de l'association estimée entre X et la maladie.

Lorsque X est pris en compte comme un facteur de confusion potentiel, une erreur de mesure sur X induit un biais dans l'estimation du coefficient de l'exposition d'intérêt principal E ajusté sur X. Même si l'erreur de mesure sur X est non différentielle, le sens du biais sur le coefficient de E peut être dans n'importe quel sens (Lellouch J et al., 1988). Pour les situations les plus fréquentes cependant, le biais est dans le sens d'une moins bonne prise en compte de la confusion due à X (Brenner H et al., 1993).
Par exemple, si X est l'âge codé en classes, l'erreur de mesure faite sur l'âge (par rapport à l'âge en continu) est non différentielle. Le plus souvent, elle a pour conséquence que l'effet de confusion dû à l'âge est mal corrigé, d'autant plus que les classes sont peu nombreuses (et donc larges), voir chapitre 4, § III.
Ce phénomène doit donc être pris en considération lorsqu'on choisit les variables candidates pour figurer dans un modèle multivarié ainsi que leur codage.

IV.3. Colinéarité

Il faut éviter d'inclure dans un modèle multivarié des variables « trop » liées entre elles. On parle alors de colinéarité et une des difficultés est de savoir ce que veut dire « trop ». Ce phénomène peut concerner deux variables ou un plus grand nombre. On parle alors de multi-colinéarité, qui est plus difficile à identifier et à corriger. La colinéarité peut être à l'origine de problèmes numériques et d'instabilité dans les analyses ultérieures au sens où l'ajout ou le retrait d'une variable impliquée dans un phénomène de colinéarité peut avoir une forte influence sur l'estimation des coefficients ou des odds ratios associés aux autres variables.

La situation « simple » est celle où la colinéarité est « mathématique », c'est-à-dire qu'il y a une équation linéaire exacte qui relie les variables concernées. La liaison entre les variables est alors plus que « trop » forte, elle est parfaite. C'est le cas par exemple (mais pas seulement) de la décomposition d'une variable à k classes en variables indicatrices. On sait que k – 1 variables indicatrices suffisent et que si on en prend k, elles sont colinéaires. En effet, la valeur d'une des k variables peut se déduire exactement des valeurs des k – 1 autres. D'un point de vue mathématique, cela correspond au fait que la matrice (X' X), qui sert à calculer les variances et covariances des coefficients du modèle, n'est pas inversible (Rakotomalala R, 2017). C'est un peu comme si on voulait diviser par 0. Dans ce cas, le logiciel élimine une des variables (ou plusieurs si nécessaire).

Sans être non inversible, la matrice (X' X) peut être « presque » non inversible, ce qui conduit à des variances très grandes des coefficients du modèle ainsi qu'à de grandes covariances entre eux. C'est pour cela que leurs estimations sont instables. Cela correspond à des situations où les variables, sans être reliées par une équation mathématique, sont fortement liées entre elles et donc « trop » liées.

La détection de la colinéarité est difficile, car il n'y a pas de seuil clair au-delà duquel les variables sont « trop » liées, ni même d'indicateur pratique de la liaison entre les variables.

Dans le cas du modèle de régression linéaire et de variables X quantitatives, un indicateur a été développé, le VIF (*Variance Inflation Factor*) (Rakotomalala R, 2017). Le VIF donne une mesure de la force du lien entre chaque variable et l'ensemble des autres. Il y a donc un VIF associé à chaque variable et on calcule aussi la moyenne des VIF. Cela permet de donner quelques guides pour décider si une ou plusieurs variables doi(ven)t être retirée(s) de l'ensemble des variables candidates (Chatterjee S et al., 1986).

Cependant, l'utilisation de VIF pour le modèle logistique ou des variables X qualitatives est, pour le moins, très discutée, et souvent non recommandée. Une des raisons est que la variance d'une variable dichotomique (pq/n) n'est pas indépendante de sa moyenne (p) et que sa variation ne peut pas se discuter de la même façon que celle d'une variable quantitative.
On ne doit donc compter que sur des études de liaisons entre les variables et son « bon sens », et aussi sur des essais réalisés en enlevant des variables ou en en ajoutant.
L'exemple suivant, tout en étant assez particulier, illustre ce qui peut se passer lorsque deux variables sont très liées et montre comment c'est parfois au cas par cas que le problème peut se résoudre. Il s'agit du l'étude de l'association entre le risque de GEU (Y = ct), d'une part, et l'intervention chirurgicale sur les trompes (ctub : chirurgie tubaire) et l'antécédent de GEU (ageu), d'autre part. À l'époque où l'enquête a été faite (entre les années 1990 et 2000), les GEU antérieures avaient presque toutes été traitées chirurgicalement. Bien qu'elles ne correspondent pas à la même chose, les deux variables ctub et ageu sont donc très liées, ainsi que le montre le tableau ci-dessous.

```
. tab ctub ageu

     chir |       atcd geu
  tubaire |         0          1 |     Total
----------+----------------------+----------
        0 |     1,540          1 |     1,541
        1 |        70        114 |       184
----------+----------------------+----------
    Total |     1,610        115 |     1,725
```

Lorsqu'on étudie leur association avec le risque de GEU, les deux variables ctub et ageu sont de très forts facteurs de risque de GEU, avec des OR bruts très élevés.

	brut	ajusté sur l'autre variable
ORctub	9,30 [6,44–13,4]	4,82 [2,93–7,95]
ORageu	14,8 [8,63–25,3]	3,39 [1,66–6,93]

Lorsque les deux variables sont incluses ensemble dans un modèle logistique, les OR ajustés varient fortement par rapport aux OR bruts. Cela doit alerter sur l'existence possible d'un problème qui est ici la colinéarité entre ctub et ageu[1]. Quand on regarde

1. Notons qu'on devrait s'en rendre compte plus tôt lors de l'examen détaillé des données de l'étude, ce qui permet de souligner à nouveau l'importance de cet examen.

les données plus en détail, on constate que l'OR de ctub ajusté sur ageu est quasiment égal à l'OR de ctub calculé chez les femmes avec ageu = 0, et réciproquement que l'OR de ageu ajusté sur ctub est quasiment égal à l'OR de ageu calculé chez les femmes avec ctub = 1. Ce qui montre l'impossibilité pratique qu'il y a à ajuster une variable sur l'autre et donc de les maintenir dans un même modèle. En pratique, il est impossible de séparer les rôles de ces deux variables.

Deux solutions sont possibles :

- ✓ ne garder qu'une des deux variables, par exemple ctub,
- ✓ construire une variable combinée qu'on peut appeler « trompe lésée » (ltub), qui vaut 1 quand ctub ou ageu est égale à 1, et 0 sinon. Cette nouvelle variable remplacera les deux précédentes dans les analyses ultérieures. C'est ce qui a été fait pour cette étude.

IV.4. Facteurs intermédiaires

Un facteur intermédiaire est un facteur qui intervient comme une étape dans la chaîne causale qui relie un facteur de risque et une maladie. Pour mieux comprendre la notion de facteur intermédiaire, considérons l'exemple de l'étude de l'association entre maladies coronariennes et exposition au sulfure de carbone (CS_2), tiré d'une enquête de Hernberg et al. (Hernberg S et al., 1973). L'hypertension artérielle (HTA) est un facteur de risque connu de maladies coronariennes. Si l'exposition au CS_2 est cause de ce type de maladies, deux schémas peuvent être envisagés. Dans le schéma 1, le CS_2 et l'hypertension sont deux facteurs de risque indépendants. Prendre en compte l'hypertension comme facteur de confusion permet « classiquement » de se rapprocher de la mesure de l'association propre entre l'exposition au CS_2 et la survenue d'une maladie coronarienne.

Schéma 1 : HTA et CS_2 sont deux facteurs de risque indépendants de maladie coronarienne.

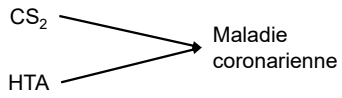
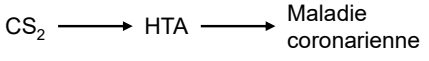


Schéma 2 : HTA est un facteur de risque intermédiaire dans la chaîne causale reliant l'exposition au CS_2 aux maladies coronariennes.



Dans le schéma 2, l'hypertension est un intermédiaire dans la chaîne causale entre l'exposition au CS_2 et l'apparition de maladies coronariennes. Ajuster sur ce facteur intermédiaire comme on le ferait pour un facteur de confusion revient alors à considérer les associations entre CS_2 et maladie, d'un côté chez les hypertendus, de l'autre chez les non-hypertendus. Dans le cas extrême où tous les exposés au CS_2 seraient hypertendus et aucun non-exposé ne le serait, cet ajustement ferait disparaître l'association entre CS_2 et maladie. Dans les cas plus courants, l'association serait fortement diminuée par l'ajustement sur l'hypertension.

La disparition par ajustement de l'association entre CS_2 et maladie coronarienne après ajustement sur l'hypertension ne conduit pas à la même interprétation dans les schémas 1 et 2. Dans le schéma 1, elle remet en cause la causalité de l'association

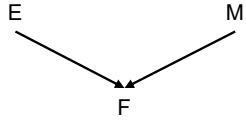
CS_2–maladie coronarienne. Dans le second, elle ne remet pas en cause cette causalité, mais apporte une information complémentaire : si la relation CS_2–maladie coronarienne est causale, elle passe – au moins en partie – par l'intermédiaire d'une élévation de la tension artérielle.

La distinction entre un facteur de risque indépendant et un facteur intermédiaire ne peut reposer que sur des arguments cliniques, biologiques ou temporels. Sur le plan statistique, on ne peut pas les distinguer. En effet, dans le schéma 2, l'hypertension répond à la définition d'un facteur de confusion : c'est un facteur lié, d'une part, à l'exposition au CS_2 et, d'autre part, à la maladie, indépendamment de l'exposition au CS_2. Lorsqu'on est en présence d'un facteur qui peut être intermédiaire, on doit adopter des stratégies différentes selon la question à laquelle on veut répondre. S'il s'agit de recherche étiologique (recherche des causes de la maladie), il est recommandé de ne pas ajuster sur un facteur intermédiaire, car cela masquerait de façon artificielle l'association entre le facteur de risque étudié et la maladie. S'il s'agit d'étudier le mécanisme d'action d'un facteur de risque, il faut prendre en compte un facteur intermédiaire. L'interprétation correcte d'un test ajusté non significatif dans le cadre du schéma 2 n'est pas l'absence de lien entre CS_2 et maladie coronarienne, mais le fait que HTA soit une étape intermédiaire. Des analyses plus sophistiquées, et en général très informatives, ont été développées pour étudier les facteurs intermédiaires : les analyses de médiation ou de cheminement (Buis ML, 2010, Lange T et al., 2012, Richiardi L et al., 2013).

Pour conclure avec l'étude de Hernberg et al. qui a servi d'exemple, précisons que les auteurs ont considéré que l'HTA au début de l'étude, au moment où les sujets commençaient à être exposés au CS_2, devait être traitée comme un facteur de confusion, mais que l'HTA diagnostiquée après 5 ans de suivi était un facteur intermédiaire.

IV.5. *Colliders*

Un troisième schéma de relation entre l'exposition E, la maladie M et un autre facteur F est possible : celui où F est une conséquence commune de E et de M.



On dit alors, dans la théorie des graphes, que F est un *collider*. J'en reparlerai plus loin dans le § V.1. Le résultat important à noter dès maintenant est qu'il ne faut pas ajuster sur un *collider*.

IV.6. Confusion résiduelle

L'existence d'une confusion résiduelle, c'est-à-dire qui reste à l'issue de l'analyse multivariée des données, a déjà été évoquée au § I.2.b. Une de ses premières causes est bien sûr l'oubli de certains facteurs de confusion. Ce risque d'oubli est bien réel, mais, en pratique, le risque que cela ait des conséquences fortes sur l'estimation

de l'OR d'intérêt reste limité. En effet, il faudrait avoir oublié un facteur majeur ou plusieurs facteurs d'importance moindre (voir § VII annexe), ce qui est peu probable si le protocole, et en particulier l'analyse bibliographique, ont été correctement faits.

Il existe cependant d'autres causes de confusion résiduelle :

- Des erreurs de mesure ou de classement sur le facteur de confusion (voir § IV.2). Une des formes fréquentes de ce type d'erreur est le choix de catégories trop larges lorsqu'on transforme une variable quantitative en classes ;
- Une mauvaise modélisation de la relation entre le facteur de confusion potentiel et la maladie (Becher H, 1992, Brenner H et al., 1997). Le plus fréquent est d'inclure sous forme linéaire une variable quantitative ou une variable qualitative ordonnée dont le lien avec la maladie s'écarte de la linéarité. La modélisation des variables quantitatives (voir chapitre 4) est donc un aspect important du choix des variables candidates à l'inclusion dans un modèle multivarié.

V. Sélection des variables à inclure dans le modèle final

Partant de la liste des variables candidates établie précédemment, il y a deux grandes catégories de méthodes pour sélectionner les variables à inclure dans le modèle final : celles qui sont basées sur l'utilisation des connaissances scientifiques et des mécanismes de la question étudiée, biologiques, médicaux, sociaux, par exemple (*subject-matter knowledge*), et celles qui font appel à des procédures statistiques (*data-driven procedures*).

Les premières sont les plus séduisantes sur le plan intellectuel, car elles font appel à un raisonnement scientifique « complet ». Depuis le début des années 2000, elles ont été systématisées par l'utilisation de DAG (*Directed Acyclic Graph*), dont je dirai quelques mots plus loin.

Les méthodes statistiques ont l'avantage (qui est aussi un inconvénient) de limiter la « subjectivité » et d'avoir un côté automatique et facile à utiliser. De nombreuses méthodes statistiques ont été développées, qui ne diffèrent parfois que par des options. J'en citerai un certain nombre, mais je ne présenterai en détail dans ce chapitre que les méthodes dites « *stepwise* », qui sont très utilisées.

Il n'y a pas de frontière étanche entre ces deux catégories. Les connaissances scientifiques sont présentes lorsqu'on utilise des méthodes statistiques, du moins on peut le souhaiter. Inversement, les raisonnements scientifiques fondés sur les connaissances et les mécanismes biologiques et sociaux s'appuient aussi sur des résultats statistiques. La distinction de ces deux catégories est, pour certains auteurs, une question de principe ou de philosophie (Mickey RM et al., 1989, Rothman KJ et al., 1998). Sans être aussi catégorique, il me semble que la distinction entre deux catégories permet de clarifier la présentation des méthodes et d'en souligner les avantages et inconvénients.

V.1. Utilisation des connaissances scientifiques, *Directed Acyclic Graph* (DAG)

Pour sélectionner les variables à inclure dans le modèle multivarié, on se fonde (même si ce n'est pas le seul outil disponible) sur les connaissances scientifiques ou empiriques pour examiner l'ensemble des variables candidates et statuer sur leur sort, en tenant compte de la littérature, des hypothèses sur les mécanismes d'action entre exposition et maladie ainsi que de la présence des autres variables.
On synthétise souvent l'ensemble des connaissances et des hypothèses par un graphe dont la théorie a été développée dans les années 2000.
Les DAG (*Directed Acyclic Graphs*) relient entre elles les différentes variables considérées (Greenland S et al., 1999, Evans D et al., 2012). Dans le terme DAG, *directed* signifie que chaque lien entre deux variables doit avoir une flèche et *acyclic* signifie qu'il ne doit pas y avoir de boucle, c'est-à-dire qu'une suite de flèches ne doit pas revenir à la variable initiale.
En dehors de la liste elle-même des variables, l'information contenue dans un DAG est soit l'absence de flèche, qui indique qu'on *sait* qu'il n'y a pas de lien entre deux variables, soit, inversement, la présence d'une flèche, dont la direction indique le sens connu ou supposé de la relation.
Les DAG sont un moyen pratique et pertinent de présenter l'ensemble des connaissances scientifiques sur une question sans faire intervenir d'analyse statistique sur les données.

Un exemple de DAG est donné sur la Figure 5.1.

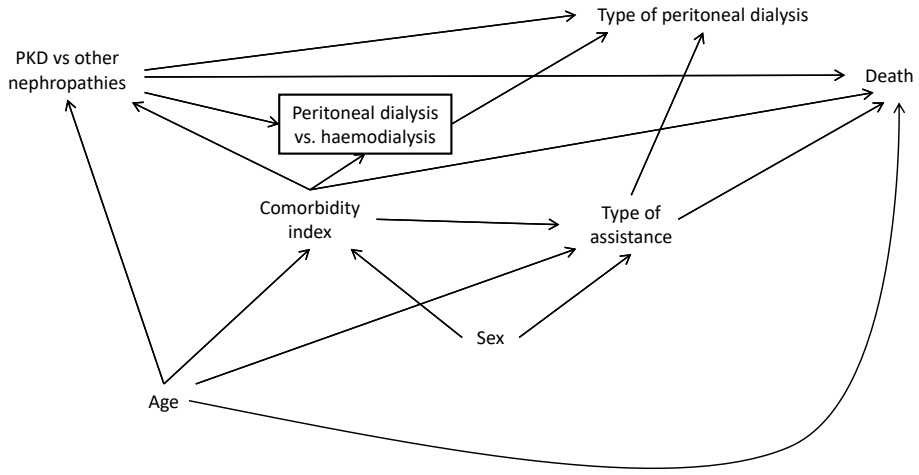


Figure 5.1 : Étude de la mortalité cinq ans après le début de la dialyse péritonéale chez des patients atteints de polykystose rénale (PKD) par rapport aux autres néphropathies. Tiré de Evans D et al. (2012).

Il y a une théorie très développée autour des DAG. Elle vient initialement de la théorie mathématique des graphes, qui est compliquée (Pearl J, 1995) et a été adaptée

par les épidémiologistes pour leurs besoins spécifiques. Elle a donné lieu à de très nombreux articles sur son utilisation en épidémiologie, notamment pour aborder la question de la causalité (Pearl J, 1995, Roese NJ, 1997, Greenland S et al., 1999, Robins JM, 2001, Dawid AP, 2002, Greenland S et al., 2002, Hernan MA et al., 2002, Maldonado G et al., 2002, Hernan MA, 2004, Hernan MA et al., 2004, Hernandez-Diaz S et al., 2006a, Hernandez-Diaz S et al., 2006b, White IR, 2006, Hernan MA et al., 2009, Shahar E, 2009a, Shahar E, 2009b).

Dans le cadre du présent chapitre sur le choix des variables à inclure dans un modèle, un des points importants mis en évidence par les DAG est le fait qu'il ne faut pas ajuster sur un *collider*.
Un *collider* est une variable qui est une conséquence à la fois de l'exposition et de la maladie. C'est ce que représente la partie de la Figure 5.2 à gauche du pointillé, tirée de Haria K (2020). On voit que la différence avec un facteur de confusion réside dans le sens des flèches.

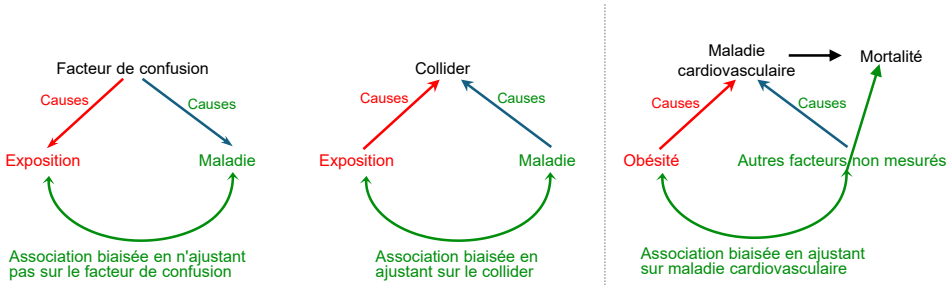


Figure 5.2 : Principe et exemple du phénomène de *collider* (Haria K, 2020).

Lorsqu'il s'agit d'un facteur de confusion, il faut ajuster et donc garder la variable dans le modèle. Lorsqu'il s'agit d'un *collider*, il ne faut pas ajuster et donc retirer la variable du modèle. Haria illustre ce phénomène par plusieurs exemples, dont celui de la partie droite de la Figure 5.2, qui illustre le « paradoxe de l'obésité », plusieurs fois discuté dans la littérature (Sperrin M et al., 2016, Viallon V et al., 2016) : dans la population générale (avec et sans maladies cardiovasculaires), l'obésité augmente le taux de mortalité (risque de décès précoce). L'ajustement sur les maladies cardiovasculaires conduit à une association biaisée entre l'obésité et la mortalité par l'intermédiaire d'autres facteurs de risque de maladies cardiovasculaires non mesurés. Les maladies cardiovasculaires jouent le rôle de *collider*. Cela aboutit, dans certaines études restreintes à des sujets atteints de maladies cardiovasculaires, à ce que l'obésité apparaisse faussement comme faisant diminuer le risque de mortalité.

Au-delà de la description du phénomène que je viens de présenter à l'aide de la Figure 5.2, plusieurs auteurs en ont décortiqué le principe et montré comment l'ajustement sur un *collider* biaise les associations entre les autres facteurs de risque et la maladie (Cole SR et al., 2009, Luque-Fernandez MA et al., 2018, Banack HR et al.,

2023). Le biais peut se produire même s'il n'y a pas d'ajustement à proprement sur un *collider*, mais une restriction de l'étude à une des catégories de ce dernier (comme avec la restriction aux maladies cardiovasculaires ci-dessus). Le bias s'apparente alors à un biais de sélection (Hernan MA et al., 2004, Lu H et al., 2023). On voit donc que l'abord de la question du « biais de *collider* » doit intégrer des variables au-delà de celles qui sont candidates à l'inclusion dans l'analyse et admettre les variables qui interviennent dans la constitution de l'échantillon.

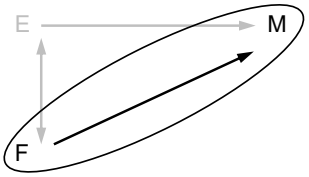
V.2. Méthodes basées sur des procédures statistiques

L'utilisation de procédures statistiques pour sélectionner les variables retenues dans un modèle logistique est très fréquente, parfois combinée aux DAG (Evans D et al., 2012) vus précédemment, ce qui est certainement une très bonne idée.
De nombreuses procédures statistiques ont été proposées. Sans chercher à être exhaustif, certaines de ces procédures sont listées ci-dessous.

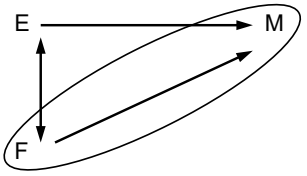
- Examen de l'écart entre OR brut et OR ajusté (voir aussi le § V.3)
  Cette méthode est basée sur la définition première du phénomène de confusion comme un biais. Une variable est facteur de confusion si l'OR ajusté sur cette variable est différent de l'OR brut. On la garde alors dans le modèle ; sinon, on l'enlève.
  Le plus courant est d'examiner les facteurs de confusion potentiels un par un (c'est ce qu'on détaillera plus bas), mais il est aussi possible d'examiner tous les modèles qui contiennent l'exposition d'intérêt et une ou plusieurs des autres variables (Wang Z, 2007). Cette dernière méthode est peu fréquemment présente dans les logiciels classiques.

- Examen de tous les modèles possibles
  Cette procédure, basée sur la comparaison de modèles, est assez peu utilisée de façon générale, et particulièrement dans le cadre du modèle logistique. On la trouve un peu plus fréquemment pour le modèle linéaire. Les modèles possibles sont très nombreux : avec p variables, il y a 2^p modèles (par exemple, p = 20 donne 1 048 576 modèles…). Comme il est en pratique impossible de les examiner un par un, on les ordonne avec des « critères d'information » (IC, *information criterion*) fondés sur la vraisemblance V et qui sont de la forme $IC = -2\ln(V) + a\ \dim(M)$, où dim(M) est le nombre de paramètres du modèle et « a » est une constante de « pénalité ». Deux critères sont utilisés principalement : le critère d'Aikaïke (AIC) avec a = 2 et le critère de Bayes (BIC) avec a = ln(n), où n est le nombre de sujets.
  Notons que cette procédure n'est pas spécifique du phénomène de confusion, puisqu'il s'agit de choisir le meilleur sous-ensemble de variables au sens de sa vraisemblance, c'est-à-dire en réalité de son adéquation aux données. Certains auteurs l'ont complétée avec une approche bayésienne qui prend en compte la distribution a priori des probabilités que les différents modèles possibles soient les bons (Wang D et al., 2004, Swartz MD et al., 2008, Genell A et al., 2010).

- Procédure pas-à-pas (*stepwise*) (voir aussi le § V.4)
  Cette procédure n'est pas non plus spécifique du phénomène de confusion. Elle vise à retenir les variables qui sont facteurs de risque de la maladie. L'idée de base, comme cela a été dit dans le § III.2, est que, pour qu'un facteur F soit facteur de confusion et soit inclus dans le modèle, il faut qu'il soit facteur de risque de la maladie.



- Double test
  Le principe est fondé sur le schéma triangulaire et sa traduction en liaisons conditionnelles entre les variables E, F et M (voir chapitre 1, § V.2) : F est incluse dans le modèle en tant que facteur de confusion s'il y a à la fois une association conditionnelle entre F et M ($OF_{FM/E^-} = OF_{FM/E^+} \neq 1$) et entre E et F ($OF_{EF/M^-} = OF_{EF/M^+} \neq 1$) (Greenland S, 1989, Sauerbrei W, 1999). Cela donne un schéma un peu différent du précédent, plus dans son principe qu'en pratique.



- Utilisation de la procédure MFP (*multivariable fractional polynomial*) (voir aussi chapitre 4, § VIII.4-5)
  Cette procédure a été développée dans le cadre de la modélisation des variables quantitatives par des polynômes fractionnaires (Sauerbrei W et al., 1999, Royston P et al., 2005, Royston P et al., 2008). Son principe est le même que celui des procédures *stepwise*, mais la procédure indique en outre comment les variables quantitatives doivent être modélisées.

- Méthodes spécifiques aux cas où il y a beaucoup de variables
  Sans développer davantage, disons qu'il y a deux catégories de méthodes de ce type :
  - Le score de propension pour les cas où il y a « un peu trop » de variables de confusion et que le nombre de sujets par variable est trop petit. La méthode consiste à calculer un score à partir de ces variables trop nombreuses et de l'inclure dans le modèle à la place des variables initiales (voir § II).
  - Les méthodes du type « lasso » pour les cas où il y a « beaucoup trop » de variables, éventuellement plus, voire beaucoup plus, de variables que de sujets (Ismaili A et al., 2009, Viallon V, 2015, Gaillard P, 2020, Courtois E et al., 2021, Pluntz M et al., 2025). Ces méthodes sont utiles par exemple en génétique statistique ou en pharmaco-surveillance des effets secondaires des médicaments.

V.3. Sélection fondée sur le changement de l'estimation de l'odds ratio

En partant du modèle complet, qui s'écrit : $\text{logit } P = \alpha + \beta E + \sum \beta_i X_i$, le principe général de cette méthode de sélection est de retirer une (ou des) variables X_i du modèle à condition que les effets de confusion soient correctement pris en compte, c'est-à-dire que le retrait de cette ou de ces variables ne modifie pas (du moins pas trop) la valeur de l'odds ratio de l'exposition d'intérêt principal E (Mickey RM et al., 1989, Talbot D et al., 2021).

Cette méthode correspond bien au phénomène de confusion défini comme un biais, puisque si une variable modifie (biaise) l'estimation de l'OR, on la retire du modèle ; sinon, on la garde.

Cette méthode est particulièrement utile s'il y a une exposition d'intérêt E, les autres variables X_i étant des facteurs de confusion potentiels, ce qui n'est pas la situation la plus fréquente. Il peut arriver qu'à l'issue de la mise en œuvre de cette méthode, des variables d'intérêt pour la question étudiée soient retirées parce qu'elles ne jouent pas un rôle de confusion. Si on veut éviter cette situation, qui peut être un véritable inconvénient pour la présentation des résultats, il faut paramétrer la méthode pour que ces variables ne puissent pas être retirées, et donc faire intervenir une part de connaissance scientifique.

Deux paramètres doivent être fixés :

- L'ordre dans lequel on examine les variables X_i.
  Le choix le plus « naturel » est de prendre les variables dans l'ordre croissant du pourcentage de modification de β. On peut aussi, au contraire, adopter une procédure ascendante, dont le principe est le même en partant d'un modèle qui ne contient que la variable E et en ajoutant les variables dans l'ordre décroissant du pourcentage de modification de β. Il y a donc une forme de *stepwise* dans cette méthode.
- Le seuil de variation de β qui est considéré comme « pas trop grand » pour conserver une variable X_i.
  Il n'y a bien sûr pas de règle absolue, mais ce seuil est souvent fixé à une variation de β de 10 % sur la base de simulations ou de comparaisons avec d'autres méthodes (Maldonado G et al., 1993)).

On peut aussi envisager de tester si la variation de β est égale à 0, ce qui revient à tester si X_i est facteur de confusion. Mais la pertinence de ce test est discutée (Greenland S, 2008, Hoffmann K et al., 2008, Pischon T et al., 2008), car ce qui importe, c'est l'ampleur de la différence entre l'OR et l'OR ajusté, qui quantifie le phénomène de confusion, plutôt que le fait qu'elle soit significativement différente de 0. Ce test, d'ailleurs pas si facile à construire, n'est donc pas recommandé et est quasiment absent des logiciels.

La méthode de sélection fondée sur le changement de l'estimation de l'odds ratio est présente dans Stata et dans R avec le module chest.

Considérons par exemple l'association entre le risque de GEU et age30 (âge ≥ 30 ans) ajustée sur les facteurs de confusion potentiels et supposons qu'on souhaite conserver le tabac dans le modèle, même s'il n'est pas facteur de confusion. Le modèle complet est présenté dans le Tableau 5.1.

```
. logit ct age30 tabf univf afcs aivg  ainf clomid ptub
....
Logistic regression                               Number of obs     =      1,619
                                                  LR chi2(8)        =     312.38
                                                  Prob > chi2       =     0.0000
Log likelihood = -858.65002                       Pseudo R2         =     0.1539
------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age30 |   .3315072   .1250582     2.65   0.008     .0863976    .5766167
        tabf |   .9657826   .1229289     7.86   0.000     .7248464    1.206719
       univf |   .1240642   .1426984     0.87   0.385    -.1556195    .4037479
        afcs |    .376094   .1406621     2.67   0.008     .1004013    .6517868
        aivg |   .1160949   .1631794     0.71   0.477    -.2037307    .4359206
        ainf |   .9001355   .1591501     5.66   0.000      .588207    1.212064
      clomid |   .2562968   .3220657     0.80   0.426    -.3749404    .8875339
        ptub |   1.464154   .1388939    10.54   0.000     1.191927    1.736381
       _cons |  -1.972198   .1176644   -16.76   0.000    -2.202816   -1.74158
------------------------------------------------------------------------------
```

Tableau 5.1 : Modèle logistique complet pour étudier l'association entre le risque de GEU et age30

La liste des variables dans l'ordre croissant de la modification du coefficient age30 avec le module chest ainsi que les valeurs modifiées du coefficient sont données dans le Tableau 5.2.

```
. chest age30,lockterms(tabf) backward  format(%5.3g)

Change-in-estimate
logit regression.                       Outcome:  ct
number of obs = 1619                    Exposure: age30

------------------------------------------------------------
Variables |
removed   |    Coef.   [95% Conf.  Interval]   Change,%
----------+-------------------------------------------------
  Adj.All |     .332        .086        .577
  -clomid |     .328        .083        .573       -1.07
   -aivg  |      .34        .097        .582        3.61
  -univf  |     .353        .111        .594        3.78
   -afcs  |     .411        .175        .648        16.6
   -ainf  |     .498        .267         .73        21.2
   -ptub  |      .65        .432        .869        30.5
------------------------------------------------------------
 tabf in all models
```

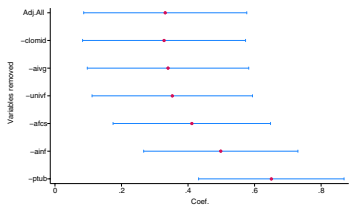


Tableau 5.2 : Résultat du module chest et graphique correspondant

Avec cette méthode, on élimine donc les 3 premières variables (clomid, aivg et unif), qui modifient β de moins de 10 %, et on conserve les 3 dernières ainsi que tabf, qui est forcée dans le modèle.
À noter que la méthode porte sur la variation de β et pas sur celle de l'OR (e^{β}).

V.4. Méthodes de sélection pas-à-pas (*stepwise*)

V.4.a. Procédures descendante et ascendante

En partant du modèle complet $\text{logit}\, P = \alpha + \sum \beta_i X_i$, c'est-à-dire sans distinguer une variable d'exposition particulière E comme dans le paragraphe précédent, le principe

général de la procédure de sélection pas-à-pas est de retirer une variable X_i si le degré de signification du test de comparaison de son coefficient β_i à 0 (ou de l'odds ratio correspondant à 1) est supérieur à une valeur fixée α, qui est en général plus grande que 5 %.
Comme je l'ai souligné plus haut, cette procédure revient à retirer une variable X_i si elle n'est pas (ou « presque pas » si on choisit α supérieur à 5 %) facteur de risque de la maladie après ajustement sur les autres variables présentes dans le modèle. La règle de décision ne porte donc pas sur le fait que la variable X_i est facteur de confusion. Elle est cependant raisonnable parce que, s'il n'y a pas d'association entre X_i et M (X_i n'est pas facteur de risque), il est très peu probable que X_i soit facteur de confusion pour l'association entre M et une autre variable X_j.
Cette procédure revient donc à garder dans le modèle les variables importantes sur le plan statistique pour la question étudiée, sans passer à côté de facteurs de confusion potentiels en choisissant un seuil de sélection au-delà des 5 % habituels (un peu comme dans le choix des variables candidates au § III.2).
Les variables sont examinées successivement une à une dans l'ordre des degrés de signification des tests de leur coefficient. Selon l'ordre choisi, on obtient ainsi des méthodes pas-à-pas descendantes, ou ascendantes.

Procédure descendante

1. On part du modèle « complet » contenant toutes les variables X_i candidates retenues.
2. On retire la variable X_k dont le degré de signification p du test du coefficient dans le modèle complet est le plus grand, tant que $p \geq \alpha$.
3. Dans le modèle sans la variable X_k, on procède de même pour les variables restantes.
4. Et ainsi de suite tant qu'il reste des variables avec $p \geq \alpha$.

Procédure ascendante

Le principe est le même « à l'envers », en partant du modèle vide et avec une liste de variables X_i susceptibles d'être incluses dans le modèle.

1. On part du modèle « vide » sans aucune variable X_i.
2. On ajoute la variable X_k, dont le degré de signification p du test d'association avec la maladie est le plus petit, tant que $p < \alpha'$.
3. Dans le modèle avec la variable ajoutée X_k, on procède de même pour les variables non encore dans le modèle.
4. Et ainsi de suite tant qu'il reste des variables et que $p < \alpha'$.

Le choix des seuils α et α' est bien sûr important. Plus ils sont grands, plus le modèle final comprendra de variables, ce qui permettra à l'utilisateur « d'ajuster » la procédure pas-à-pas à ses besoins. On les prend souvent égaux et de l'ordre de 0,10 à 0,15. (Sun GW et al., 1996, Harrell FE, 2001). Un seuil à 0,157 a été proposé par Royston et al. comme le meilleur compris entre trop et trop peu de variables (Royston P et al., 2008). En réalité, il n'y a pas de valeur du seuil aussi impérative que le 0,05 du seuil de signification.

Les modèles obtenus ne sont pas obligatoirement les mêmes avec les procédures descendantes et ascendantes, même si c'est souvent le cas en pratique. A priori, la procédure ascendante sélectionne moins de variables, ce qui peut être un avantage pour avoir un modèle final économe en variables. La procédure descendante est cependant souvent préférée, car elle présente l'avantage de partir du modèle complet, qui est celui qui prend le mieux en compte l'ensemble des phénomènes de confusion et qui est donc une meilleure référence que le modèle vide.

V.4.b. Adaptations aux données et aux connaissances scientifiques

Les méthodes pas-à-pas sont souvent critiquées à cause de leur côté automatique. Cette critique est fondée pour les méthodes ascendantes et descendantes décrites ci-dessus. Elle s'applique d'ailleurs aux autres méthodes qui reposent sur des procédures statistiques. Elle doit être nuancée par le fait que ces méthodes comprennent toutes l'établissement de la liste des variables candidates, qui nécessite, quant à lui, une connaissance de la question et des recherches publiées sur le sujet.
Il existe en outre des options qui permettent de faire entrer des connaissances scientifiques ou des choix d'analyse dans cet automatisme. J'ai parlé plus haut du choix des seuils α et α' qui permet de piloter le nombre total de variables dans le modèle final ainsi que l'écart souhaité au seuil de signification statistique habituel. Les autres possibilités d'adaptation des méthodes *stepwise* visent à limiter le fait qu'elles mettent les variables sur le même plan sans tenir compte de leur rôle particulier dans les phénomènes étudiés ou de la façon dont elles ont été construites :

- Lorsqu'une variable qualitative a été décomposée en variables indicatrices, on sait que les variables obtenues doivent être considérées comme un bloc et non comme des variables séparées (voir chapitre 3, § III.4). Il faut donc qu'une méthode pas-à-pas respecte cette règle en excluant du modèle l'ensemble des variables indicatrices, ou en les gardant (ou les entrant) toutes. Ce n'est donc pas à proprement parler une adaptation ni une option, mais une obligation pour la procédure *stepwise*.
- Si une ou plusieurs variables sont considérées comme suffisamment importantes pour devoir apparaître de toute façon parmi les variables retenues, on peut les « forcer » dans tous les modèles qui seront examinés par la procédure *stepwise*. On se rapproche alors de la sélection des variables selon les connaissances scientifiques, d'autant plus fortement que le nombre de variables forcées est grand.
- Si les variables candidates sont nombreuses, on peut les organiser en groupes homogènes qu'on commence par analyser séparément. Les procédures pas-à-pas sont appliquées à chaque groupe et les variables retenues sont ensuite incluses dans le modèle final. Certaines variables peuvent intervenir dans plusieurs groupes : par exemple, on peut ajuster systématiquement sur l'âge l'analyse des antécédents médicaux.

À titre d'exemple, les facteurs de risque potentiels de GEU ont été répartis en quatre groupes comme dans le Tableau 5.3 ci-dessous (Bouyer J et al., 2003). L'âge a été inclus dans tous les groupes pour la raison indiquée plus haut et sera forcé dans tous les modèles.
La place de certaines variables dans un groupe plutôt qu'un autre peut se discuter. L'antécédent d'accouchement fait ainsi partie des « antécédents chirurgicaux et obstétricaux », mais peut aussi être vu comme un indicateur de fertilité et être mis dans le quatrième groupe. En cas de doute, il est possible de faire des essais avec des compositions de groupes différentes (on parle souvent d'analyses de sensibilité). Si les résultats sont les mêmes, c'est que la place des variables concernées dans les groupes n'est pas importante pour le résultat final. Sinon, il faut s'interroger sur la raison des différences et sur le choix qui paraît le meilleur.

J'ajoute pour conclure qu'il n'y a pas d'évidence qu'on fasse mieux en regroupant ainsi les variables qu'en les prenant toutes ensemble. Mais cela permet d'introduire de la connaissance scientifique dans une procédure automatique. Et aussi, et peut-être surtout, d'organiser sa réflexion et la présentation des résultats.

Caractéristiques socio-professionnelles	**Exposition potentielle aux infections sexuellement transmissibles**
Âge	
Consommation de tabac	Âge
Niveau d'études	Infection gynécologique
	Sérologie positive pour *C. trachomatis*
	Âge aux premiers rapports sexuels
Antécédents chirurgicaux et obstétricaux	Nombre de partenaires
Âge	**Fertilité**
Appendicectomie	Âge
Chirurgie tubaire	Antécédent de contraception
Antécédent de GEU	Dernier mode de contraception
Antécédent de FCS	Antécédent d'infécondité
Antécédent d'accouchement	Grossesse induite
Antécédent d'IVG	Délai depuis la dernière grossesse

Tableau 5.3 : Répartition en quatre groupes des variables analysées comme facteurs de risque de grossesse extra-utérine (GEU)

Enfin, les procédures ascendantes et descendantes peuvent être mixées ou alternées. C'est possible à partir de l'étape où 2 variables ont été retirées de la procédure descendante ou où 2 variables ont été ajoutées dans la procédure ascendante :

- ✓ si la variable exclue la plus significative vérifie $p < \alpha'$, on la réintègre ;
- ✓ si la variable incluse la moins significative vérifie $p \geq \alpha$, on la retire.

Cette procédure mixte peut avoir l'intérêt d'attirer l'attention sur des variables qui sont sorties dans un premier temps avant d'être réincluses lorsque d'autres variables ont été exclues à leur tour. Il s'agit alors de comprendre pourquoi ce phénomène assez inattendu s'est produit. Il peut s'agir par exemple de colinéarités entre variables non identifiées au préalable.

V.4.c. Avantages et inconvénients des procédures pas-à-pas

Les avantages et inconvénients des procédures pas-à-pas ont déjà été évoqués dans le début de ce chapitre. Ils sont synthétisés et complétés ici :

Avantages des procédures pas-à-pas (Sauerbrei W et al., 2007, Royston P et al., 2008)

- Ce sont des procédures faciles à utiliser et programmées dans tous les logiciels d'analyse statistique.
- Elles examinent toutes les variables de façon systématique et standardisée. Cet aspect automatique est souvent présenté comme un inconvénient. Il a cependant le mérite d'éviter des dérives dans le cas où on ne donnerait pas assez de poids aux données observées par rapport aux connaissances scientifiques, ces dernières étant filtrées par nos propres convictions.
- Les adaptations que j'ai mentionnées plus haut permettent d'assouplir la standardisation et d'introduire dans la procédure ce qui paraît nécessaire de connaissances scientifiques.
- Il n'y pas de réelle alternative en l'absence de connaissances scientifiques fortes des mécanismes gouvernant la question étudiée, ce qui est fréquent, ou d'expérience d'analyse des données, ce qui est aussi fréquent.

Inconvénients des procédures pas-à-pas (Sun GW et al., 1996, Sribney B, 1998, Steyerberg EW et al., 1999, Harrell FE, 2001, Austin PC et al., 2004b, Rothman KJ et al., 2008)

Dans la littérature, les inconvénients des procédures pas-à-pas ont été beaucoup mis en avant avec des arguments de principe, mais aussi des arguments statistiques, parfois assez « pointus ». C'est par exemple clairement le point de vue de l'équipe qui pilote le logiciel Stata, pour qui les méthodes *stepwise* ne devraient pas être choisies (Sribney B, 1998). Ces arguments doivent bien sûr être entendus et pris en compte. Mais il ne faut pas oublier que l'alternative qui les accompagne souvent est de s'appuyer sur les connaissances scientifiques à propos de la question étudiée. Or, il est impossible de comparer quantitativement les méthodes de sélection pas-à-pas avec celles qui s'appuient sur des connaissances scientifiques. Il n'y a donc pas d'évidence qu'elles fassent mieux (voir §V.5.b). Il est en revanche assez évident qu'elles sont peu ou mal reproductibles.

- L'argument principal est que les méthodes pas-à-pas sont fondées sur des critères uniquement statistiques, sans considérations épidémiologiques. Il est

indiscutable et doit conduire à utiliser des méthodes pas-à-pas adaptées aux besoins, comme indiqué plus haut.

- Des arguments plus purement statistiques sont aussi avancés. Harrell explique ainsi que les méthodes pas-à-pas violent certaines conditions statistiques (risque d'erreur, distribution, biais). Elles comportent de nombreux tests qui conduisent à une sous-estimation des intervalles de confiance des coefficients, qui devraient être plus grands pour tenir compte de l'incertitude ajoutée par le choix du modèle (Harrell FE, 2001). Harrell recommande de faire un test global préalable et de s'arrêter sans inclure de variable s'il est non significatif. En réalité, il est très peu probable que le test préalable soit non significatif, ou alors c'est qu'on a été particulièrement maladroit dans le choix des variables candidates (ou dans la question étudiée…).

- La sélection des variables est potentiellement biaisée.
  En effet, elle est basée sur les coefficients estimés des variables et non pas sur leur vraie valeur. Une variable est donc plus vraisemblablement incluse dans le modèle final si son coefficient est surestimé. Cela conduit à ce que les « prédicteurs forts » (facteurs fortement liés à la maladie) soient presque toujours inclus, que leur coefficient soit sous-estimé ou pas (Swartz MD et al., 2008), alors que les « prédicteurs faibles » ne sont inclus que si leur coefficient est surestimé.
  Une possibilité est alors d'utiliser une méthode de *shrinkage* (contraction) (Greenland S, 2008, Royston P et al., 2008) qui vise à ramener les coefficients vers 0 pour éviter une surestimation liée à la sélection des variables ou au maintien de toutes les variables.

V.4.d. Exemples de procédures pas-à-pas

Voici deux exemples de procédures pas-à-pas.
Le premier exemple que je vous propose reprend celui qui a été donné pour la méthode fondée sur le changement de l'estimation de l'odds ratio avec le module chest (voir § V.3). Cette méthode est d'ailleurs une méthode pas-à-pas particulière, puisque là aussi les variables sont retirées ou ajoutées une à une sur des critères statistiques. Pour pouvoir comparer aux résultats des Tableaux 5.1 et 5.2, on part du même modèle complet, avec les variables age30 tabf univf afcs aivg ainf clomid ptub et on force les variables age30 et tabf. Le résultat obtenu figure dans le Tableau 5.4.
Pour la méthode pas-à-pas, le seuil fixé pour exclure une variable est $\alpha = 0{,}10$ (pr(.1) dans la commande). Le seuil de sélection était aussi de 10 % pour la modification du coefficient d'une variable dans la procédure chest (mais voir la remarque ci-dessous).
Finalement, les deux méthodes, si elles suivent des procédures différentes, aboutissent au même modèle final : les 3 mêmes variables sont éliminées, bien que pas tout à fait dans le même ordre.

```
. stepwise, lockterm1 pr(.1) : logit ct (age30 tabf) univf afcs aivg  ainf clamid ptub
                  begin with full model
p = 0.4768 >= 0.1000  removing aivg
p = 0.4303 >= 0.1000  removing clamid
p = 0.3477 >= 0.1000  removing univf

Logistic regression                             Number of obs     =      1,619
                                                LR chi2(5)        =     310.39
                                                Prob > chi2       =     0.0000
Log likelihood = -859.64865                     Pseudo R2         =     0.1529

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age30 |   .3526231   .1230346     2.87   0.004     .1114796    .5937665
        tabf |   .9588128   .1210092     7.92   0.000     .7216391    1.195987
        ainf |   .9231743   .1514091     6.10   0.000     .6264179    1.219931
        afcs |   .3724291   .1405269     2.65   0.008     .0970015    .6478567
        ptub |    1.45788   .1384557    10.53   0.000     1.186512    1.729249
        cons |  -1.923389   .1103932   -17.42   0.000    -2.139756   -1.707023
------------------------------------------------------------------------------
```

Tableau 5.4 : Sélection pas-à-pas (*stepwise*) descendante avec des paramètres semblables à ceux du Tableau 5.2

Remarques

- Il faut néanmoins être attentif à ce que les seuils des deux méthodes, bien qu'ils s'expriment tous les deux avec 0,10, n'ont rien à voir l'un avec l'autre ! L'un est un degré de signification, l'autre un pourcentage de variation de β.
- Les résultats montrent par ailleurs que la sélection des variables resterait la même pour la méthode *stepwise* si on avait pris pr(.30) au lieu de pr(.1), puisque toutes les variables retirées ont un p supérieur à 0,30. Des calculs complémentaires montreraient de plus qu'il faudrait aller jusqu'à pr(.001) pour qu'une variable supplémentaire soit retirée (ce serait alors afcs). Le résultat resterait donc inchangé pour toutes les valeurs de α entre 0,001 et 0,30, ce qui donne une certaine confiance dans la solidité de la méthode pas-à-pas.

```
. stepwise,pr(.1) : logit ct (i.agec) tabf univf afcs aivg ainf clamid ptub
note: 0b.agec omitted because of estimability.

Wald test, begin with full model:
p = 0.5437 >= 0.1000, removing aivg
p = 0.4348 >= 0.1000, removing clamid
p = 0.3691 >= 0.1000, removing univf

Logistic regression                                   Number of obs =  1,619
                                                      LR chi2(7)    = 312.85
                                                      Prob > chi2   = 0.0000
Log likelihood = -858.41549                           Pseudo R2     = 0.1541

------------------------------------------------------------------------------
          ct | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
        agec |
          1  |   .0306916   .1737823     0.18   0.860    -.3099154    .3712985
          2  |   .2739225   .1841362     1.49   0.137    -.0869778    .6348229
          3  |   .5587238   .2091504     2.67   0.008     .1487965    .9686511
             |
        tabf |    .969784   .1214215     7.99   0.000     .7318023    1.207766
        ainf |   .9201408   .1521056     6.05   0.000     .6220193    1.218262
        afcs |   .3560214   .1411075     2.52   0.012     .0794559     .632587
        ptub |   1.440776   .1388064    10.38   0.000      1.16872    1.712831
       _cons |  -1.942585   .1608772   -12.07   0.000    -2.257898   -1.627271
------------------------------------------------------------------------------
```

Tableau 5.5 : Sélection pas-à-pas (*stepwise*) descendante avec une variable qualitative (agec) décomposée en variables indicatrices

Le deuxième exemple porte sur les mêmes variables, mais cette fois l'âge n'est pas pris comme une variable dichotomique (age30), mais comme une variable à quatre classes (agec), qui est décomposée en variables indicatrices. Cet exemple est là pour insister sur le fait que les variables indicatrices doivent être considérées comme un bloc et non examinées une par une.
Le modèle obtenu par sélection *stepwise* descendante est le suivant (Tableau 5.5).
Les parenthèses autour de i.agec indiquent que les trois variables indicatrices doivent être considérées comme un bloc. Si ces parenthèses étaient absentes, ces trois variables ne seraient pas considérées comme liées, et ici les deux premières variables indicatrices (1.agec et 2.agec) seraient retirées et la dernière conservée, ce qui reviendrait à regrouper les trois premières catégories d'âge (voir chapitre 3, § III.4) et n'aurait donc pas lieu d'être (en tout cas pas de cette façon).

V.5. Fréquence d'utilisation et comparaison des méthodes de sélection des variables

V.5.a. Fréquence d'utilisation

L'utilisation des différentes méthodes de sélection des variables dans les articles d'épidémiologie a été étudiée par Walter et Tiermeier (Walter S et al., 2009), qui ont analysé 4 revues majeures d'épidémiologie en 2008 (*The American Journal of Epidemiology*, *Epidemiology*, *The European Journal of Epidemiology* et *The International Journal of Epidemiology*), soit 300 articles. Cette étude a été refaite pour l'année 2015 par Talbot et Massamba (Talbot D et al., 2019), sur les mêmes journaux et 292 articles. Les résultats sont résumés dans le Tableau 5.6.

Méthode de sélection	Articles de 2008[1] (n = 300)	Articles de 2015[1] (n = 292)
Utilisation des connaissances scientifiques	28 %	50 % (40 %[2])
Changement de l'estimation de l'odds ratio	15 %	12 %
Stepwise ou analyse univariée	20 %	14 %
Autre (score de propension, régression pénalisée, shrinkage)	3 %	2 %
Non décrite	35 %	37 %

Notes

(1) Les répartitions des catégories de méthodes sont assez stables d'un journal à l'autre. C'est pourquoi je ne donne ici que la répartition totale.

(2) Les auteurs n'ont pas utilisé tout à fait les mêmes catégories ; j'en ai tenu compte avec un ou deux regroupements selon les commentaires faits par les auteurs. Une des différences importantes concerne la catégorie « Utilisation des connaissances scientifiques ». Les articles de 2015 ont été rangés dans cette catégories dès qu'une « vague » notion de connaissance a priori était mentionnée et même si une autre méthode de sélection était aussi utilisée. D'où les deux chiffres de 50 % et 40 %, le second correspondant aux cas où seules les connaissances a priori étaient utilisées (même formulées vaguement).

Tableau 5.6 : Méthodes de sélection des variables utilisées dans 4 journaux majeurs d'épidémiologie en 2008 et 2015 (Walter S et al., 2009, Talbot D et al., 2019)

On peut tirer de ces deux études les enseignements suivants :

- Un premier tiers (35 % à 37 % de l'ensemble des articles) ne dit rien de la méthode utilisée, ou en tout cas la section « Matériel et méthodes » est trop imprécise pour qu'on sache précisément ce qui a été fait. C'est en fait ce qui me paraît poser le plus de problème.
- Un deuxième « gros » tiers (28 % à 40 % des articles) utilise les connaissances a priori pour sélectionner les variables, mais la moitié de ces publications sont des articles de 2008 qui ne donnent ni justifications ni références qui permettraient de comprendre sur quoi ces connaissances s'appuient. On n'est alors pas très loin de la catégorie précédente.
- Le dernier tiers (28 % à 35 %) utilise des méthodes automatiques de sélection : méthode pas-à-pas ou variation de l'estimation de l'OR après ajustement.
- Les méthodes plus « modernes », regroupées dans la catégorie « autres », sont quasiment absentes (moins de 3 %).
- Les auteurs de l'étude sur l'année 2015 indiquent que l'évolution apparente entre 2008 et 2015 peut s'expliquer par l'écart de définition de la catégorie « Utilisation des connaissances scientifiques » (voir note de bas de tableau). Cela paraît probable en raison du délai assez court.

V.5.b. Comparaison des méthodes

La méthode de sélection des variables basée sur les connaissances scientifiques, bien qu'elle soit recommandée par beaucoup d'épidémiologistes et par des livres de référence, n'est, par construction, pas comparable de façon quantitative aux autres méthodes par simulation, taux de variables bien sélectionnées, etc. Il n'y a en réalité aucune évidence qu'elle fasse mieux que les méthodes basées sur des procédures statistiques (Sauerbrei W et al., 2007). Son choix reste donc une question de philosophie ou de principe, et il y a des arguments intellectuellement séduisants pour le faire. Les détracteurs des procédures statistiques (notamment les méthodes pas-à-pas) leur reprochent un côté automatique et systématique, indépendamment des connaissances épidémiologiques (Rothman KJ et al., 1998, Harrell FE, 2001). Il n'y a pas à réfléchir, juste à appuyer sur un bouton. Pour le logiciel, il n'y a pas de différence autre que statistique entre X_1 et X_2. Il ne faut cependant pas mythifier les méthodes basées sur les connaissances. Il peut arriver que la façon dont elles sont utilisées en pratique revienne finalement à faire péniblement « à la main » ce que la machine fait bien et vite automatiquement ! Cela peut se produire lorsqu'on manque de connaissances, ou plus prosaïquement lorsqu'on manque d'expérience et/ou d'assurance. Il peut alors être utile de recourir à des méthodes plus « automatiques » pour les acquérir.

Les méthodes basées sur des procédures statistiques, dont certaines intègrent aussi des connaissances scientifiques, comme on l'a vu avec les méthodes *stepwise*, ont été assez souvent comparées entre elles (Mickey RM et al., 1989, Maldonado G et al., 1993, Austin PC et al., 2004b, Wang D et al., 2004, Sauerbrei W et al., 2007, Swartz MD et al., 2008, Genell A et al., 2010). On peut cependant noter qu'il s'agissait surtout

de comparaisons entre une méthode classique (telle que la méthode *stepwise*) et des méthodes plus récentes (celles de la catégorie « autre » du Tableau 5.6). En effet, ces nouvelles méthodes avaient besoin de ces comparaisons pour justifier leur apport (et leur publication).
Il est difficile de passer toutes ces comparaisons en revue, mais disons pour synthétiser que, dans l'ensemble, la procédure *stepwise* fait moins bien que des méthodes plus modernes, même si l'écart est limité sur le plan quantitatif. C'est ainsi que le pourcentage de variables incluses à tort ou non incluses à tort est de l'ordre de 10 à 15 % avec les méthodes *stepwise* et 5 % avec des méthodes plus sophistiquées.

VI. Variables à inclure en raison de la structure de l'échantillon, enquêtes multicentriques

Les paragraphes précédents portent sur les variables à inclure (ou pas) dans le modèle logistique pour prendre en compte au mieux les facteurs de risque de la maladie et les facteurs de confusion. Une autre catégorie de variables doit aussi être considérée : celles qui doivent être incluses dans l'analyse pour tenir compte de la façon dont l'échantillon a été constitué. Je vais détailler ici le cas des enquêtes multicentriques, qui est la situation qu'on rencontre le plus fréquemment. Les enquêtes avec appariement pourraient aussi entrer dans ce cadre, mais elles sont maintenant presque toujours analysées avec des modèles mixtes (McCulloch CE et al., 2001) plutôt qu'en incluant les variables d'appariement (Breslow NE et al., 1980).

Une enquête multicentrique est une enquête dont l'échantillon a été constitué séparément à partir de plusieurs sources (ou centres). Cela peut être différentes régions pour des études nationales, de façon à couvrir la diversité des situations, différents hôpitaux ou services hospitaliers pour un essai thérapeutique ou une enquête cas-témoins. Dans certaines de ces situations, comme pour un essai thérapeutique multicentrique, on peut penser que les sujets d'un même service se ressemblent davantage entre eux que les sujets de services différents, même si on n'est pas complètement en mesure de préciser par quel mécanisme. Les sujets de l'ensemble de l'échantillon ne sont alors pas indépendants et les méthodes d'analyse et d'estimation qui font appel à au maximum de vraisemblance ne peuvent pas être utilisées.

La solution consiste à inclure dans le modèle logistique la variable (qualitative nominale) qui caractérise les centres où l'étude a été réalisée, en la décomposant, bien sûr, en variables indicatrices. Les estimations des coefficients des autres variables X_i sont alors faites à valeur constante de la variable centre (on dit aussi conditionnellement à cette variable). Comme, au sein d'un même centre, les sujets sont indépendants, les estimations du maximum de vraisemblance sont utilisables.
Cette solution fait l'hypothèse que les odds ratios entre chaque variable X_i et la maladie (ou l'effet du traitement s'il s'agit d'un essai thérapeutique) sont les mêmes d'un centre à l'autre. C'est-à-dire qu'il n'y a pas d'interaction entre les X_i et la variable centre. A minima, il faut supposer que, s'il y a une interaction, elle est de

type quantitatif (voir chapitre 1, § V.2.d) et que l'estimation d'un odds ratio moyen a un sens.
J'ajoute que cette solution qui consiste à inclure la variable centre est tout à fait admise lorsqu'il y a peu de centres, mais est parfois discutée s'ils sont plus nombreux (Kahan BC, 2014). Surtout depuis que les modèles mixtes sont facilement accessibles dans les logiciels pour analyser des données non indépendantes (McCulloch CE et al., 2001).

VII. Annexe : Conditions pour qu'une association soit expliquée par un facteur de confusion

Le résultat général est que plus une association entre une exposition et une maladie est forte sur le plan quantitatif, moins il est vraisemblable qu'elle soit le fruit de phénomènes de confusion.
On peut montrer que, pour qu'un risque relatif brut RR soit explicable par un facteur de confusion, il faut que ce facteur présente lui-même un risque relatif pour la maladie supérieur ou égal à RR et soit au moins RR fois plus fréquemment présent chez les sujets exposés au facteur de risque étudié que chez les autres.
Supposons, par exemple, qu'on observe une association entre tabac et cancer du poumon mesurée par un risque relatif égal à 9. Pour que cette association soit entièrement explicable par un facteur de confusion, ce dernier doit être associé à la maladie avec un risque relatif au moins égal à 9 et doit être au moins neuf fois plus fréquent chez les fumeurs que chez les non-fumeurs. Il est donc peu vraisemblable, dans ce cas, qu'on « passe à côté » d'un tel facteur de confusion[2].
Un risque relatif fort peut cependant être le fruit, non pas d'un seul, mais de plusieurs facteurs de confusion dont le rôle individuel est modeste et qui sont donc moins facilement identifiables. On est alors dans un cas où les groupes comparés sont dissemblables pour un nombre important de facteurs, ce dont on se rend souvent compte en pratique. Il peut s'agir par exemple d'un mauvais choix des témoins.
En résumé, si l'association entre l'exposition et la maladie étudiées est forte, elle ne peut s'expliquer par un phénomène de confusion que si l'on est passé à côté, soit d'un facteur de risque concurrent massif, soit d'un grand nombre de facteurs de risque moins importants.

Pour démontrer ce résultat tiré de Lellouch J et al. (1988), il faut revenir sur le calcul du rapport de confusion, qui a déjà été abordé au chapitre 1 (§ V.2.c). Je vais me limiter au cas où les variables sont dichotomiques et les associations mesurées par des risques relatifs, et où il n'y a pas d'interactions.

Les observations peuvent être résumées par les deux tableaux suivants, où E et M sont l'exposition et la maladie étudiées et F est le facteur de confusion :

2. Des résultats tout à fait analogues sont obtenus, dans le cas de variables quantitatives, pour le coefficient de corrélation.

	F^+		F^-	
	E^+	E^-	E^+	E^-
M^+	a_1	b_1	a_2	b_2
M^-	c_1	d_1	c_2	d_2
	n_1	n'_1	n_2	n'_2

À partir de ces tableaux, on peut calculer :

- ✓ le risque relatif brut entre E et M : $RR = \frac{a_1 + a_2}{n_1 + n_2} \times \frac{n'_1 + n'_2}{b_1 + b_2}$
- ✓ le risque relatif entre E et M conditionnel à F qui vaut, puisqu'il n'y a pas d'interaction : $RR_a = \frac{a_1}{n_1} / \frac{b_1}{n'_1} = \frac{a_2}{n_2} / \frac{b_2}{n'_2}$
- ✓ le risque relatif entre F et M conditionnellement à E : $RR_{FM/E} = \frac{a_1}{n_1} / \frac{a_2}{n_2} = \frac{b_1}{n'_1} / \frac{b_2}{n'_2}$.

Des deux dernières équations, on déduit : $a_1 + a_2 = RR_a \left(\frac{a_1 n_1}{n'_1} + \frac{n_2 b_2}{n'_2} \right)$ et $b_2 = \frac{b_1 n'_2}{n'_1 RR_{FM}}$

d'où : $RR = RR_a \frac{n'_1 + n'_2}{n_1 + n_2} = \frac{\frac{n_1 b_1}{n'_1} + \frac{n_2 b_1}{n'_1 RR_{FM/E}}}{b_1 + \frac{n'_2 b_1}{n'_1 RR_{FM/E}}} = RR_a \frac{n'_1 + n'_2}{n_1 + n_2} \frac{n_1 RR_{FM/E} + n_2}{n'_1 RR_{FM/E} + n'_2}$

soit : $RR = RR_a \frac{\frac{n_1}{n_1 + n_2} RR_{FM/E} + \frac{n_2}{n_1 + n_2}}{\frac{n'_1}{n'_1 + n'_2} RR_{FM/E} + \frac{n'_2}{n'_1 + n'_2}}$, ce qui peut aussi s'écrire, en notant

$p_1 = \frac{n_1}{n_1 + n_2} = P(F^+/E^+)$ et $p_2 = \frac{n'_1}{n'_1 + n'_2} = P(F^+/E^-)$: $RR = RR_a \frac{p_1 RR_{FM/E} + (1 - p_1)}{p_2 RR_{FM/E} + (1 - p_2)}$

Le rapport $RR_c = \frac{RR}{RR_a} = \frac{p_1 RR_{FM/E} + (1 - p_1)}{p_2 RR_{FM/E} + (1 - p_2)}$ est appelé rapport de confusion ; il quantifie le degré de confusion entraîné par F pour l'association entre E et M.

Pour que la relation entre E et M soit entièrement expliquée par le facteur de confusion F, il faut que RR_a soit égal à 1. On a alors $RR = \frac{p_1 RR_{FM/E} + (1 - p_1)}{p_2 RR_{FM/E} + (1 - p_2)}$ ou encore

$RR_{FM/E} - 1 = \frac{RR - 1}{p_1 - p_2 RR}$ (1)

On se place dans la situation où RR et $RR_{FM/E}$ sont supérieurs à 1. On déduit de (1) que $p_1 - p_2 RR > 0$, ce qui donne $\frac{p_1}{p_2} > RR$. Comme on a $p_1 - p_2\, RR < p_1 < 1$, on déduit aussi de (1) que $RR_{FM/E} > RR$.

Au total, il est donc nécessaire que le facteur de confusion ait dans chaque classe de E un risque relatif de maladie supérieur à celui de E dans l'ensemble de la population et soit au moins RR fois plus fréquent en présence de E qu'en son absence.

Chapitre 6

Régressions logistiques multinomiale et ordinale

• • •

I. Introduction

Dans les chapitres précédents, j'ai montré comment quantifier et modéliser la relation entre X et Y avec le modèle logistique lorsque Y est une variable dichotomique. Si Y est une variable qualitative à plus de deux classes, il faut envisager d'autres modèles, qui font l'objet de ce chapitre. Si Y a p classes et que X est dichotomique, l'association entre X et Y ne peut plus être caractérisée par un seul odds ratio, il faut considérer les (p – 1) OR obtenus en comparant une des classes de Y à chacune des autres. Et on sait que c'est toujours plus compliqué quand plusieurs paramètres sont nécessaires…

Lorsque Y est nominale, l'ordre des catégories de Y est arbitraire et il n'y a pas de relations particulières entre les (p – 1) OR. En choisissant une classe de Y comme référence (souvent les non-malades s'il y a une classe qui leur correspond), le problème se ramène à lui comparer les autres classes (différents types de maladie, par exemple). La méthode correspondante est la régression logistique multinomiale qui est présentée au § II.

Lorsque Y est ordinale, l'ordre des classes peut (et doit) être pris en compte. Il faut cependant noter qu'une variable ordinale n'est pas toujours une variable quantitative discrète. C'est-à-dire que la distance entre deux catégories adjacentes peut ne pas être constante, voire ne pas être définie. Il y a donc besoin de méthodes d'analyse spécifiques, qui ne peuvent pas se limiter à des adaptations des modèles linéaires où Y est quantitative. Ces méthodes sont les modèles dits ordinaux, qui sont présentés au § III et détaillés dans les paragraphes suivants. Elles reposent sur des choix pour comparer les classes de Y, par exemple en considérant la probabilité cumulée $P(Y \leq j)$, qui est la probabilité d'être en deçà de la catégorie j. Le logit qui est modélisé est alors $\text{logit}\, P(Y \leq j) = \ln \frac{P(Y \leq j)}{P(Y > j)} = \alpha_j - \beta_j X$. On cherche souvent à résumer l'information donnée par les (p – 1) coefficients β_j (ou par les $OR_j = e^{\beta_j}$) par un seul coefficient β, ce qui revient à supposer que les β_j sont égaux et à estimer leur valeur commune, comme on le verra avec les différents modèles présentés.

Dans le modèle multinomial, comme dans les modèles ordinaux, les variables indépendantes (« variables X ») peuvent, comme dans le modèle logistique classique, être qualitatives ou quantitatives et doivent être prises en compte, comme cela a été expliqué dans les chapitres 3 et 4. Il est cependant important de réaliser que l'interprétation des résultats est souvent hasardeuse lorsque X n'est pas

dichotomique. Cela est tout à fait similaire à l'analyse « de base » de l'association entre une exposition X à k classes et une maladie à k' classes. Sur le plan technique, on sait tester l'existence d'une association (ici avec un test de χ^2 à $(k-1)(k'-1)$ degrés de liberté), mais on est souvent bien en peine de commenter et de donner un sens au résultat.

II. Régression logistique multinomiale

II.1. Définition et écriture du modèle

Le rôle de la régression logistique multinomiale (on dit aussi parfois polytomique) est de modéliser la relation entre une variable maladie Y qualitative nominale à plus de deux classes et une ou des variables X qui, comme pour la régression classique, peuvent être qualitatives ou quantitatives.
Les k classes de Y sont notées ici 1, 2, …, k et n'ont donc pas d'ordre. La situation qu'on rencontre le plus fréquemment est celle où une classe (la classe 1, par exemple) est composée de non-malades et sert de référence, les autres classes correspondant à des maladies différentes ou à des catégories d'une même maladie, sans qu'il y ait d'ordre entre elles.

La régression logistique multinomiale consiste essentiellement à comparer les classes 2 à k à la classe de référence 1, et donc à « empiler » $(k-1)$ régressions logistiques classiques dans une analyse unique[1]. J'y reviendrai plus loin (§ II.3) pour préciser cette phrase, qui est ici un peu résumée, et pour indiquer quel peut être l'avantage d'une analyse unique par rapport à $(k-1)$ analyses logistiques « dichotomiques » classiques.
Pour écrire le modèle multinomial, il faut un double système d'indices : un pour les classes de Y qui seront repérées par j et un pour les variables incluses dans le modèle, qui seront repérées par i. Avec ces notations, le modèle s'écrit comme une série de modèles logistiques classiques :

$$\ln\left(\frac{P(Y=j\,|\,X)}{P(Y=1\,|\,X)}\right)=\alpha_j+\sum_{i=1}^{p}\beta_{ji}X_i\,,\ \text{où}\ j=2,\ \ldots,\ k\ \text{sont les classes de Y autres que 1 et}$$

$X=(X_1,\ldots,X_p)$ sont les p variables indépendantes incluses dans le modèle.

Le modèle logistique multinomial s'écrit parfois sous une autre forme, qui donne l'expression de $P(Y=j\,|\,X)$ (voir le détail des calculs en Annexe VIII.1) :

$$P(Y=j\,|\,X)=\frac{\exp(\alpha_j+g_j(x))}{\sum_{j=1}^{k}\exp\left\{\alpha_j+g_j(x)\right\}}\ \text{avec}\ g_j(x)=\sum_{i=1}^{p}\beta_{ji}x_i\ \text{et}\ \alpha_1=\beta_{1i}=0$$

1. Notons au passage que cela implique que le modèle multinomial peut être utilisé dans les enquêtes cas-témoins (voir Annexe VIII.2).

C'est cette forme qu'on retrouve dans des livres ou les manuels des logiciels. Elle présente l'avantage de permettre de calculer la vraisemblance (et donc d'estimer les paramètres), mais l'inconvénient de ne pas faire apparaître le modèle multinomial comme une série de régressions logistiques classiques.

II.2. Exemple et interprétation des résultats

L'exemple qui suit utilise des données de grossesses après FIV (fécondation in vitro) en France entre 1986 et 1994 (le fait qu'elles soient après FIV n'a pas d'importance ici). Il s'agit d'étudier le lien entre l'âge de la femme, l'hypertension artérielle (HTA) pendant la grossesse et la naissance prématurée. Seules les grossesses singletons sont analysées, en raison du risque augmenté de prématurité pour les grossesses multiples, assez fréquentes en FIV. Les naissances à terme sont prises comme référence (codée Y = 1). Le type de prématurité est pris en compte : prématurité par rupture des membranes (codée Y = 2), prématurée spontanée (codée Y = 3), prématurée provoquée (codée Y = 4).
L'âge de la femme est enregistré dans une variable dichotomique notée age35 et codée 1 : ≥ 35 ans et 0 sinon. L'hypertension artérielle est aussi une variable dichotomique notée hta et codée (0 : non et 1 : oui). La variable Y (noté cprema) est en 4 classes, avec le codage indiqué ci-dessus.

Les résultats du modèle logistique multinomial donnés par le logiciel Stata sont donnés dans le Tableau 6.1. Y = 1 (naissance à terme) a été pris comme référence, c'est l'option par défaut (classe de Y la plus fréquente). Ici, ce choix « s'impose » mais, si cela n'était pas le cas, il serait possible de prendre une autre classe de référence, il y a une option pour cela. Les résultats seraient équivalents, mais présentés différemment.

La présentation des résultats illustre bien le fait qu'il s'agit de plusieurs régressions logistiques binomiales « empilées » avec la catégorie cprema = 1 prise comme référence (*base outcome*). La présentation des résultats avec R est similaire.
On peut d'ailleurs constater que les résultats du premier bloc du Tableau 6.1 sont strictement identiques à ceux du Tableau 6.2, qui sont les résultats d'une régression logistique dichotomique classique, où Y (rupmemb) est la variable 0/1 correspondant aux deux premières catégories de cprema avec 0 : à terme et 1 : prématuré par rupture des membranes[2].

Comme pour la régression logistique dichotomique, il est possible d'obtenir les odds ratios en prenant les exponentielles des coefficients de la régression. Une option le permet avec Stata (voir Tableau 6.3) ; avec R, il faut calculer explicitement l'exponentielle du coefficient.

2. Je reviendrai sur ce point dans le paragraphe suivant, car c'est un peu plus compliqué que cela lorsqu'il y a plusieurs variables X…

Cela permet de souligner que, dans les résultats de Stata, l'intitulé de la colonne n'est pas OR, mais RRR (*relative risk ratio*). Stata considère en effet que l'odds ratio correspondant à une classe de Y est celui qui oppose cette classe à toutes les autres réunies[3]. Or, ce que donne la régression multinomiale est l'OR qui oppose cette classe à la classe de référence. Cet argument me paraît tout à fait recevable, bien que quelque peu (inutilement ?) puriste, surtout avec l'interprétation de la régression multinomiale comme une série de régressions logistiques binomiales. Je continuerai donc à parler d'odds ratios lorsqu'il s'agit des exponentielles des coefficients.

```
. mlogit cprema age35

....

Multinomial logistic regression                 Number of obs =  7,601
                                                LR chi2(3)    =   8.70
                                                Prob > chi2   = 0.0336
Log likelihood = -3025.1826                     Pseudo R2     = 0.0014

------------------------------------------------------------------------------
      cprema | Coefficient  Std. err.      z    P>|z|     [95 conf. interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
       age35 |   .1761545   .1993624     0.88   0.377    -.2145885    .5668976
       _cons |  -4.211474   .1221634   -34.47   0.000     -4.45091   -3.972038
-------------+----------------------------------------------------------------
3            |
       age35 |    .084688   .1188679     0.71   0.476    -.1482887    .3176648
       _cons |  -3.088647   .0707297   -43.67   0.000    -3.227275   -2.950019
-------------+----------------------------------------------------------------
4            |
       age35 |     .35869   .1277462     2.81   0.005     .1083121    .6090679
       _cons |  -3.407101   .0824436   -41.33   0.000    -3.568688   -3.245514
------------------------------------------------------------------------------
```

Tableau 6.1 : Résultats de la régression multinomiale avec Y = cprema en 4 classes (1 à 4) et X = age35, variable dichotomique (0/1)

```
. logit rupmemb age35

....

Logistic regression                             Number of obs =  7,015
                                                LR chi2(1)    =   0.77
                                                Prob > chi2   = 0.3805
Log likelihood=-561.69019                       Pseudo R2     = 0.0007

------------------------------------------------------------------------------
    rupmemb | Coefficient  Std. err.      z    P>|z|     [95 conf. interval]
-------------+----------------------------------------------------------------
       age35 |   .1761545   .1993624     0.88   0.377    -.2145885    .5668976
       _cons |  -4.211474   .1221634   -34.47   0.000     -4.45091   -3.972038
------------------------------------------------------------------------------
```

Tableau 6.2 : Résultats de la régression logistique binomiale classique modélisant la relation de X = age35, avec rupmemb qui correspond aux 2 premières classes de cprema

3. Voir https://www.stata.com/statalist/archive/2005-04/msg00678.html.

```
. mlogit cprema age35,rrr

....

Multinomial logistic regression                 Number of obs =  7,601
                                                LR chi2(3)    =   8.70
                                                Prob > chi2   = 0.0336
Log likelihood = -3025.1826                     Pseudo R2     = 0.0014

------------------------------------------------------------------------------
      cprema |        RRR   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
       age35 |   1.192622    .237764     0.88   0.377     .8068734     1.76279
       _cons |   .0148245    .001811   -34.47   0.000     .0116679     .018835
-------------+----------------------------------------------------------------
3            |
       age35 |   1.088377   .1293731     0.71   0.476     .8621822    1.373916
       _cons |   .0455635   .0032227   -43.67   0.000     .0396654    .0523387
-------------+----------------------------------------------------------------
4            |
       age35 |   1.431453   .1828626     2.81   0.005     1.114396    1.838717
       _cons |   .0331371   .0027319   -41.33   0.000     .0281928    .0389485
------------------------------------------------------------------------------
Note:  cons estimates baseline relative risk for each outcome.
```

Tableau 6.3 : Résultats de la régression multinomiale identique à celle du Tableau 6.1, mais avec l'exponentielle des coefficients plutôt que les coefficients eux-mêmes

II.3. Régression logistique multinomiale versus plusieurs régressions binomiales

J'ai indiqué que le modèle logistique multinomial est en réalité un empilement de modèles logistiques binomiaux classiques et on a vu dans l'exemple précédent, avec une seule variable X, que ses résultats sont effectivement les mêmes que ceux de plusieurs régressions logistiques successives (Tableaux 6.1 et 6.2).

En fait, lorsqu'on inclut plusieurs variables X_i dans le modèle multinomial, les résultats s'écartent légèrement de ceux de plusieurs régressions logistiques classiques. C'est ce que montre le Tableau 6.4, dans lequel deux variables sont prises de compte : age35 et hta. Cette différence a déjà été signalée dans la littérature (Begg CB et al., 1984, Agresti A, 1990), mais sans qu'il ait été donné d'explication. On voit d'ailleurs sur le Tableau 6.4 que la différence est numériquement minime, ce qui est un constat général dans mon expérience, et peut expliquer qu'elle ait été négligée.

En réalité, cet écart entre la régression multinomiale et plusieurs régressions logistiques binomiales lorsqu'il y a plusieurs variables X_i est dû à des hypothèses différentes faites par ces modèles. De façon générale, un modèle logistique avec les seules variables indépendantes age35 et hta suppose qu'il n'y a pas d'interaction entre ces deux variables (voir chapitre 3, § VI).

Dans le cas du modèle multinomial, cette hypothèse d'absence d'interaction est unique et porte sur l'ensemble des comparaisons entre Y = i et Y = 1, alors que, pour les modèles logistiques binomiaux successifs, il y a plusieurs hypothèses d'absence d'interaction qui portent *séparément* sur chaque comparaison entre Y = 0 et Y = i.

Lorsqu'on estime les coefficients des modèles binomiaux successifs, c'est chaque hypothèse spécifique qui s'applique tout à tour, et non l'hypothèse globale du modèle multinomial.

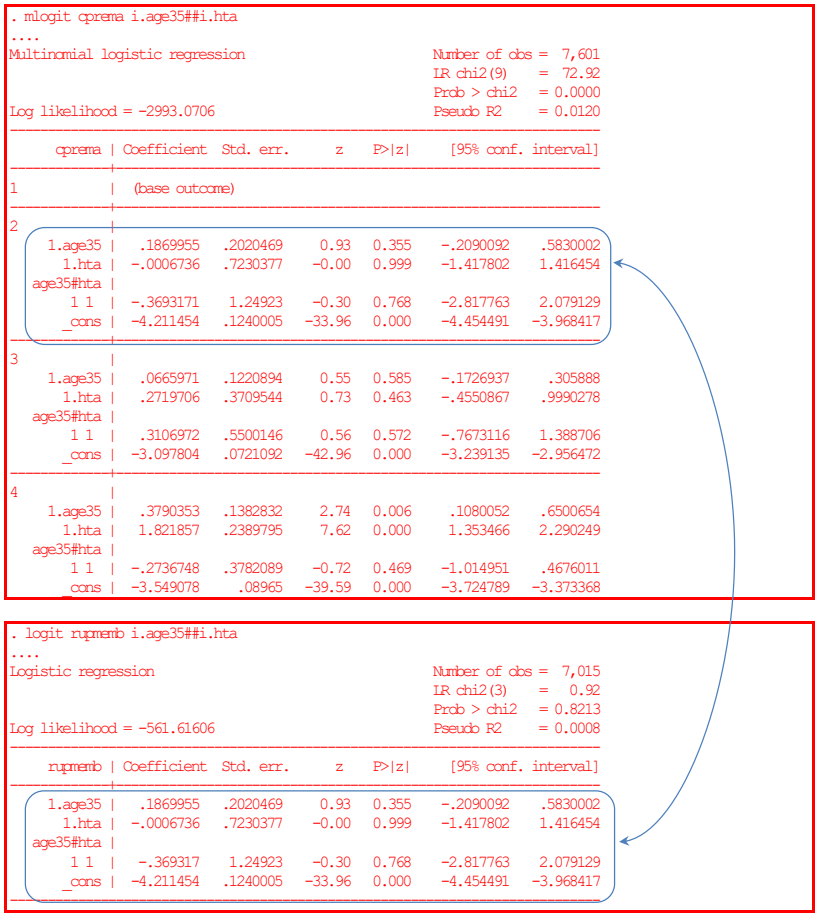
```
. mlogit cprema age35 hta

....

Multinomial logistic regression                 Number of obs =  7,601
                                                LR chi2(6)    =  71.91
                                                Prob > chi2   = 0.0000
Log likelihood = -2993.5779                     Pseudo R2     = 0.0119

------------------------------------------------------------------------------
      cprema | Coefficient  Std. err.      z    P>|z|     [95 conf. interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
       age35 |   .1767704   .1993784     0.89   0.375     -.214004    .5675449
         hta |   -.138074   .5896005    -0.23   0.815    -1.29367     1.017522
       cons  |  -4.207639   .1231218   -34.17   0.000    -4.448953   -3.966324
-------------+----------------------------------------------------------------
3            |
       age35 |   .0823515   .1188986     0.69   0.489    -.1506855    .3153886
         hta |   .4048827   .2733238     1.48   0.139    -.1308221    .9405875
       _cons |  -3.103353   .0717072   -43.28   0.000    -3.243896   -2.962809
-------------+----------------------------------------------------------------
4            |
       age35 |   .3399427    .128752     2.64   0.008     .0875933    .5922921
         hta |   1.706887   .1854845     9.20   0.000     1.343344    2.07043
       _cons |  -3.532773   .0863849   -40.90   0.000    -3.702084   -3.363461
------------------------------------------------------------------------------
```

```
. logit rupmemb age35 hta

Logistic regression                             Number of obs =  7,015
                                                LR chi2(2)    =   0.83
                                                Prob > chi2   = 0.6614
Log likelihood=-561.66125                       Pseudo R2     = 0.0007

------------------------------------------------------------------------------
     rupmemb | Coefficient  Std. err.      z    P>|z|     [95 conf. interval]
-------------+----------------------------------------------------------------
       age35 |   .1768493   .1993825     0.89   0.375    -.2139333    .5676318
         hta |  -.1388484   .5896181    -0.24   0.814    -1.294479    1.016782
       cons  |  -4.207644   .1231079   -34.18   0.000    -4.448931   -3.966357
------------------------------------------------------------------------------
```

Tableau 6.4 : Comparaison des résultats de la régression multinomiale et plusieurs régressions logistiques classiques lorsqu'il y a plusieurs variables X_i

Lorsqu'on ajoute un terme d'interaction au modèle multinomial, il y en a en réalité un par classe de Y, c'est-à-dire autant que pour l'ensemble des modèles binomiaux. Le modèle multinomial redevient identique à une série de modèles binomiaux. C'est ce que montrent les résultats du Tableau 6.5.

Ces résultats ont le mérite d'expliquer le (petit) écart entre la régression multinomiale et des régressions binomiales successives. Ils permettent aussi de se rappeler, de façon générale, que la modélisation peut inclure des hypothèses implicites qu'il est utile d'identifier, même si ce n'est pas toujours facile.
Je pense cependant que cela ne doit pas conduire à abandonner l'interprétation de la régression multinomiale comme un empilement de régressions logistiques binomiales classiques (comme dans les Tableaux 6.1 et 6.2), même si les variables X_i figurent seules (c'est-à-dire sans interaction).



```
. mlogit cprema i.age35##i.hta
....
Multinomial logistic regression                 Number of obs =  7,601
                                                LR chi2(9)    =  72.92
                                                Prob > chi2   = 0.0000
Log likelihood = -2993.0706                     Pseudo R2     = 0.0120
------------------------------------------------------------------------------
      cprema | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
    1.age35 |   .1869955   .2020469     0.93   0.355    -.2090092    .5830002
      1.hta |  -.0006736   .7230377    -0.00   0.999    -1.417802    1.416454
  age35#hta |
        1 1 |  -.3693171    1.24923    -0.30   0.768    -2.817763    2.079129
      _cons |  -4.211454   .1240005   -33.96   0.000    -4.454491   -3.968417
-------------+----------------------------------------------------------------
3            |
    1.age35 |   .0665971   .1220894     0.55   0.585    -.1726937     .305888
      1.hta |   .2719706   .3709544     0.73   0.463    -.4550867    .9990278
  age35#hta |
        1 1 |   .3106972   .5500146     0.56   0.572    -.7673116    1.388706
      _cons |  -3.097804   .0721092   -42.96   0.000    -3.239135   -2.956472
-------------+----------------------------------------------------------------
4            |
    1.age35 |   .3790353   .1382832     2.74   0.006     .1080052    .6500654
      1.hta |   1.821857   .2389795     7.62   0.000     1.353466    2.290249
  age35#hta |
        1 1 |  -.2736748   .3782089    -0.72   0.469    -1.014951    .4676011
      _cons |  -3.549078    .08965   -39.59   0.000    -3.724789   -3.373368
------------------------------------------------------------------------------

. logit rupmemb i.age35##i.hta
....
Logistic regression                             Number of obs =  7,015
                                                LR chi2(3)    =   0.92
                                                Prob > chi2   = 0.8213
Log likelihood = -561.61606                     Pseudo R2     = 0.0008
------------------------------------------------------------------------------
    rupmemb | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    1.age35 |   .1869955   .2020469     0.93   0.355    -.2090092    .5830002
      1.hta |  -.0006736   .7230377    -0.00   0.999    -1.417802    1.416454
  age35#hta |
        1 1 |   -.369317    1.24923    -0.30   0.768    -2.817763    2.079129
      _cons |  -4.211454   .1240005   -33.96   0.000    -4.454491   -3.968417
------------------------------------------------------------------------------
```

Tableau 6.5 : Comparaison entre des résultats de régression multinomiale et ceux de plusieurs régressions logistiques classiques lorsqu'il y a plusieurs variables X_i et une interaction entre les X_i

Il est alors légitime de se demander à quoi sert la régression logistique multinomiale, puisqu'on peut la remplacer par des régressions logistiques binomiales. Au risque d'être provocateur, je dirais : à pas grand-chose ! On peut cependant lui trouver un double et modeste intérêt :

- Elle permet une analyse et une présentation unifiée des comparaisons entre Y = i et Y = 1, parce que les variables indépendantes incluses sont les mêmes. C'est cependant à double tranchant, car il faut bien commenter toutes les comparaisons entre Y = j et Y = 1 et justifier qu'on garde toutes les variables indépendantes, même si elles ne sont pas significatives pour une comparaison particulière.
- La régression multinomiale permet de comparer les OR_j associés à une variable X pour les différentes comparaisons entre Y = j et Y = 1. Cela est rendu possible parce que les OR_j (en réalité les coefficients des variables dont ils sont les exponentielles) sont estimés dans un même modèle. Avec des régressions logistiques

binomiales séparées, on a bien les estimations et les variances, mais pas les covariances. Si les comparaisons entre OR_j sont intéressantes, le paragraphe suivant montre comment procéder.

II.4. Comparaison des OR associés à X selon les catégories de Y

Reprenons les résultats de la partie supérieure du Tableau 6.4. En prenant les exponentielles des coefficients β_j pour avoir des odds ratios, on peut les résumer de la façon suivante.

	Y = 1 (À terme)	Y = 2 (Rupt préma)	Y = 3 (Préma spont)	Y = 4 (Préma prov)
Age ≥ 35	1	1,2 [0,81 ; 1,8]	1,1 [0,86 ; 1,4]	1,4 [1,1 ; 1,8]
HTA	1	0,87 [0,27 ; 2,8]	1,5 [0,88 ; 2,6]	5,5 [3,8 ; 7,9]

Tableau 6.6 : OR associés à age35 et à hta selon la classe de Y (avec comme référence Y = 1, naissance à terme)

Pour comparer les OR associés à hta pour rupture prématurée des membranes (Y = 2) et pour prématurité spontanée (Y = 3), c'est-à-dire comparer $OR_{1,2} = 0,87$ et $OR_{1,3} = 1,5$, il faut, après avoir estimé la régression logistique multinomiale, utiliser la commande de test ci-dessous, qui montre que la différence entre ces deux OR est non significative.

```
. test [2]hta = [3]hta

 ( 1)  [1]hta - [2]hta = 0
           chi2(  1) =    0.71
         Prob > chi2 =    0.3980
```

En revanche, la différence entre les OR associés à hta pour la prématurité provoquée (Y = 4) et pour la prématurité spontanée (Y = 3) est significative.

```
. test [4]hta = [3]hta
 ( 1)  - [3]hta + [4]hta = 0
           chi2(  1) =   17.04
         Prob > chi2 =    0.0000
```

Si on veut tester globalement l'homogénéité entre les trois OR associés à hta, il faut les comparer simultanément. Là aussi, la différence est significative :

```
. test ([2]hta = [3]hta) ([4]hta = [3]hta)

 ( 1)  [2]hta - [3]hta = 0
 ( 2)  - [3]hta + [4]hta = 0

           chi2(  2) =   22.88
         Prob > chi2 =    0.0000
```

Les mêmes calculs pour age35 montrent que les OR ne sont pas significativement différents, ni globalement, ni deux par deux.

II.5. Changement de classe de référence pour Y

Les résultats de la régression logistique multinomiale sont équivalents quel que soit le choix de la classe de référence pour Y. Les OR estimés ne sont pas identiques, puisqu'ils comparent des classes de Y différentes quand on change la référence, mais on peut passer des uns aux autres sans problème.
Prenons par exemple la classe 3 (prématurité spontanée) comme référence. On obtient les résultats du Tableau 6.7.

```
. mlogit cprema age35, b(3)

....

Multinomial logistic regression                 Number of obs =  7,601
                                                LR chi2(3)    =   8.70
                                                Prob > chi2   = 0.0336
Log likelihood = -3025.1826                     Pseudo R2     = 0.0014

------------------------------------------------------------------------------
      cprema | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
1            |
       age35 |   -.084688   .1188679    -0.71   0.476    -.3176648    .1482887
       _cons |   3.088647   .0707297    43.67   0.000     2.950019    3.227275
-------------+----------------------------------------------------------------
2            |
       age35 |   .0914665   .2292956     0.40   0.690    -.3579446    .5408776
       _cons |  -1.122827   .1396086    -8.04   0.000    -1.396454   -.8491986
-------------+----------------------------------------------------------------
3            |  (base outcome)
-------------+----------------------------------------------------------------
4            |
       age35 |    .274002   .1707343     1.60   0.109    -.0606311    .6086351
       _cons |  -.3184537   .1066004    -2.99   0.003    -.5273866   -.1095209
------------------------------------------------------------------------------
```

Tableau 6.7 : Résultats de la même régression multinomiale que celle du Tableau 6.1, avec Y = 3 comme classe de référence

On note que la vraisemblance est la même que celle du Tableau 6.1. On observe aussi que les coefficients de la ligne 1, qui correspondent à la comparaison entre la classe 1 et la classe 3 prise comme référence, sont opposés à ceux de la ligne 3 du Tableau 6.1, ce qui est normal, car ces derniers correspondent à la comparaison inverse entre la classe 3 et la classe 1.
Si on veut le coefficient de la comparaison entre les classes 2 et 4 (dont l'exponentielle est l'odds ratio entre ces deux classes), on trouve le même résultat avec le Tableau 6.7 ci-dessus (0,274 – 0,091 = 0,183) qu'avec le Tableau 6.1 (0,359 – 0,176 = 0,183).

Le choix de la classe de référence de Y « ne change donc rien ». Dans cet exemple, il est logique de prendre les naissances à terme, qui sont, en quelque sorte, les non-malades. Les logiciels ont toujours une option par défaut si on ne fait pas de choix explicite. Stata prend la classe la plus fréquente (qui est celle des naissances à terme dans cet exemple).

III. Les différents modèles de régression logistique ordinale

Cela peut paraître une remarque banale, mais lorsque Y est une variable qualitative ordonnée, une des difficultés pratique avec un modèle ordinal est de parvenir à s'extraire du raisonnement que l'on fait lorsque c'est X qui est une variable ordonnée : il n'y a pas ici de relation dose-effet, mais une recherche de l'« effet »[4] de X sur la distribution de Y, c'est-à-dire la probabilité que Y appartienne à telle ou telle catégorie, sachant que ces catégories sont ordonnées et qu'on veut en tenir compte.

Plusieurs modèles logistiques ordinaux sont possibles selon la façon dont on interprète le caractère ordonné des classes de Y et selon les classes ou groupe de classes de Y qu'on souhaite comparer. Dans tous les cas, les probabilités de Y sont modélisées par la fonction logistique. Par exemple, la probabilité que Y soit inférieure ou égale à j est modélisée par $\text{logit}(P(Y \leq j|X)) = \alpha_j - \sum_{i=1}^{p} \beta_{ji} X_i$. Le signe « – » devant $\sum_{i=1}^{p} \beta_{ji} X_i$ est là pour faciliter l'interprétation de β (voir Annexe § VIII.2). Dans les modèles les plus fréquents, on suppose que les β ne dépendent pas de j (hypothèse dite des *odds* proportionnels entre les classes de Y). Le modèle s'écrit alors :

$$\text{logit}(P(Y \leq j|X)) = \alpha_j - \sum_{i=1}^{p} \beta_i X_i.$$

A.S Fullerton, dans un article très clair, classe les modèles logistiques ordinaux selon deux dimensions (Fullerton AS, 2009) :

(1) Le type d'odds ratios qu'on veut calculer : cumulatif, par étapes ou adjacent ;

(2) L'« étendue » de l'hypothèse des *odds* proportionnels : elle peut s'appliquer à toutes les variables indépendantes X_i, à certaines, ou à aucune d'entre elles.

III.1. Type d'odds ratios à calculer

On distingue habituellement trois types d'approches pour les odds ratios comparant les classes de Y :

- L'approche cumulative s'intéresse à la distribution de Y par l'intermédiaire de la probabilité cumulée de Y, c'est-à-dire la probabilité que Y soit inférieur au seuil j : $P(Y \leq j|X)$, qui est comparée à la probabilté que Y soit supérieur à j. Le modèle correspondant est le *cumulative-odds model*.

  Il s'écrit : $\text{logit}\left(P(Y \leq j|X)\right) = \ln\left(\frac{P(Y \leq j \mid X)}{P(Y > j \mid X)}\right) = \alpha_j - \sum_{i=1}^{p} \beta_{ji} X_i$, où $j = 1, \ldots, k-1$ désigne une classe de Y.

  C'est le modèle le plus courant ; il est parfois appelé tout simplement « modèle logistique ordinal ». Il est détaillé dans le § IV.

4. Je mets des guillemets car ce mot n'a pas ici de sens causal.

- L'approche par étapes modélise la probabilité que $Y = j$ parmi les sujets tels que $Y \geq j$, c'est-à-dire $\delta_j = \frac{P(Y = j \mid X)}{P(Y \geq j \mid X)}$, qui est en réalité une probabilité conditionnelle. Le modèle correspondant est le *continuation-ratio model*. Il s'écrit : $\text{logit}\left(\frac{P(Y = j \mid X)}{P(Y \geq j \mid X)}\right) = \ln\left(\frac{\delta_j}{1-\delta_j}\right) = \ln\left(\frac{P(Y = j \mid X)}{P(Y > j \mid X)}\right) = \alpha_j - \sum_{i=1}^{p} \beta_{ji} X_i$, où $j = 1, \ldots, k-1$ désigne une classe de Y. Ce modèle est détaillé dans le § V.
- L'approche adjacente modélise aussi une probabilité conditionnelle : la probabilité que $Y = j$ parmi les sujets tels que $Y = j$ ou $Y = j+1$, soit : $\frac{P(Y = j \mid X)}{P(Y = j \mid X) + P(Y = j+1 \mid X)}$. Le modèle correspondant est le *adjacent-category model*. Il s'écrit : $\text{logit}\left(\frac{P(Y = j \mid X)}{P(Y = j \mid X) + P(Y = j+1 \mid X)}\right) = \ln\left(\frac{P(Y = j \mid X)}{P(Y = j+1 \mid X)}\right) = \alpha_j + \sum_{i=1}^{p} \beta_{ji} X_i$, où $j = 1, \ldots, k-1$ désigne une classe de Y. Il sera détaillé dans le § VI.

Ces trois approches correspondent à des points de coupure différents et à différents types d'odds ratios entre les classes de Y. Si, par exemple, Y a quatre classes, les points de coupure dans le modèle cumulatif sont 1/2–4, 1–2/3–4 et 1–3/4, tandis que, dans le modèle *continuation-ratio*, les points de coupure sont 1/2–4, 2/3–4 et 3/4, et que, dans le modèle *adjacent-category*, il s'agit de 1/2, 2/3 et 3/4.

III.2. Hypothèse des *odds* proportionnels

Les *odds* sont dits proportionnels si les β_{ji} ne dépendent pas de j. Si c'est le cas, l'indice j des β doit être supprimé dans l'écriture des modèles vus précédemment et l'odds ratio associé à X_i est constant.
Dans chacun des trois types de modèles ordinaux présentés dans le paragraphe précédent, l'hypothèse des *odds* proportionnels peut s'appliquer à chaque variable X_i, à un sous-ensemble de ces variables, ou à aucune d'entre elles.
Dans le cas où un sous-ensemble de variables ne satisfait pas l'hypothèse des *odds* proportionnels, il est possible qu'une forme de lien existe quand même entre leurs coefficients, ce qui constitue une quatrième catégorie d'hypothèses (Ananth CV et al., 1997).

Le croisement de ces deux dimensions (type d'odds ratios et proportionnalité des *odds*) donne ainsi lieu à une typologie des modèles de régression logistique ordinales en 12 catégories (Fullerton AS, 2009). D'un point de vue pratique, les logiciels proposent des commandes pour les trois principaux types de modèles (*cumulative-odds*, *continuation-ratio* et *adjacent-category*), avec des options correspondant aux hypothèses faites sur la proportionnalité des *odds* (voir (Bauldry S et al., 2018) pour Stata).

Notons enfin que lorsque Y n'a que deux classes, tous les modèles ordinaux sont identiques au modèle logistique binomial classique.

IV. Modèle *cumulative-odds*

Le modèle *cumulative-odds* s'écrit $\text{logit}(P(Y \leq j|X)) = \alpha_j - \sum_{i=1}^{p} \beta_{ji} X_i$. Il correspond à une situation où existe une variable continue sous-jacente T découpée en classes qui définissent les classes de Y (voir Annexe VIII.2).
Je vais, dans un premier temps, m'intéresser au modèle avec *odds* proportionnels, qui s'écrit $\text{logit}(P(Y \leq j|X)) = \alpha_j - \sum_{i=1}^{p} \beta_i X_i$, où $j = 1, \ldots, k-1$ désigne une classe de Y. Le coefficient d'une variable X_i (et l'odds ratio correspondant) est alors indépendant du seuil j. C'est le modèle le plus simple et le plus courant. Je reviendrai plus loin sur le modèle sans cette hypothèse de proportionnalité et sur la façon de le tester.

Le modèle proportionnel *cumulative-odds* possède les propriétés suivantes :

- $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{k-1}$. C'est une conséquence du fait que $P(Y \leq j|X) \leq P(Y \leq j'|X)$ si $j < j'$. Mais il est toujours bon de vérifier ces inégalités dans les résultats, car le contraire indiquerait une erreur.
- Si on inverse l'ordre des classes de Y, les β_i changent de signe, mais restent égaux en valeur absolue.
- Si l'hypothèse des *odds* proportionnels est vraie, les valeurs vraies des β_i ne changent pas si on regroupe des catégories adjacentes de Y. Leurs estimations peuvent bien sûr varier après regroupements.
- Le modèle *cumulative-odds* ne peut pas être utilisé dans des enquêtes de type cas-témoins (voir Annexe VIII.3), c'est-à-dire dans des enquêtes où la distribution de Y dans l'échantillon est choisie par l'investigateur.

IV.1. Interprétation des coefficients du modèle *cumulative-odds*

Comme cela est fréquemment le cas pour les modèles logistisques, il n'y a pas d'interprétation concrète des constantes α_j. Elles sont liées aux points de coupure de la variable T sous-jacente (voir l'Annexe VIII.2), mais comme, en pratique, la variable T est inconnue, on ne sait pas donner de sens à la valeur des α_j.
Les β_i quantifient le lien entre X_i et Y. Le signe de β_i est facile interpréter, mais il faut quand même faire attention, notamment en vérifiant si la modélisation utilisée par le logiciel est avec un – ou avec un + devant les β_i, ce qu'il n'est pas toujours si simple de savoir… En revanche, la valeur numérique de β_i (ou celle de l'OR e^{β_i}) est très difficile à verbaliser et à interpréter. On se contente souvent du signe de β_i et de sa significativité statistique.

IV.2. Exemple : fragilité chez les personnes vivant avec le VIH

Cet exemple utilise les données de l'enquête ANRS-Septavih, qui porte sur le vieillissement précoce des personnes vivant avec le VIH (virus d'immunodéficience humaine) (Achour J et al., 2024). L'enquête porte sur 491 sujets[5]. Le vieillissement précoce est quantifié par l'état de fragilité, qui est caractérisé par une diminution des capacités de l'organisme à s'adapter à des situations de stress.
La fragilité était mesurée par le score clinique de Fried, qui repose sur cinq critères : perte de poids sur les 12 derniers mois, fatigue, niveau d'activité physique, vitesse de marche, force de préhension. Le score est noté sur une échelle numérique discrète de 0 (pas de critère de fragilité) à 5. Trois phénotypes sont définis : « robuste » si le patient a zéro critère de fragilité, « pré-fragile » si le patient présente 1 ou 2 critère(s) de fragilité, « fragile » si le patient a 3 critères de fragilité ou plus. Deux variables Y ont été définies : sf5 pour le score complet à 5 classes et sf3 pour le score en 3 phénotypes. Pour des raisons de commodité dans l'utilisation des logiciels, ces deux variables ont été codées de 1 à 6 et de 1 à 3 (plutôt que de 0 à 5 et de 0 à 3).

La répartition des sujets selon les scores de fragilité est donnée dans le Tableau 6.8 (aucun sujet dans cet échantillon n'a un score de fragilité maximum (égal à 6) avec cinq critères de fragilité).

sf5 (score de Fried en 5 classes)			sf3 (score en 3 classes)		
1	121	24,6 %	1 : robuste	121	24,6 %
2	210	42,8 %	2 : pré-fragile	327	66,6 %
3	117	23,8 %			
4	29	5,9 %	3 : fragile	43	8,8 %
5	14	2,9 %			
Total	491			491	

Tableau 6.8 : Répartition des sujets selon leur score de fragilité

Les facteurs de fragilité étudiés ici sont :

- Précarité socio-économique. Elle est quantifiée par le score EPICES (évaluation de la précarité et des inégalités de santé dans les centres d'examens de santé), variable quantitative continue cotée entre 0 (absence de précarité) et 100 (maximum de précarité) à partir d'un questionnaire standardisé de 11 questions. La variable correspondante est epices.

  Le seuil habituellement utilisé pour définir la précarité est 30. Une variable dichotomique, notée precaire, est donc aussi utilisée avec le codage : 1 si epice $\geq$ 30 ; 0 sinon.

5. Les données manquantes ont été imputées pour cet exemple.

- Taux de CD4 inférieur à 350 cellules/mm^3. La variable (en 0/1) est notée cd4_350.
- Maladie rénale chronique. La variable (en 0/1) est notée mrc.
- Vie en couple ou relation amoureuse suivie. La variable (en 0/1) est notée parten.

IV.3. Résultats avec l'hypothèse des *odds* proportionnels

Les résultats du modèle *cumulative-odds* avec l'hypothèse de proportionnalité des *odds* sont donnés dans le Tableau 6.9.

Le modèle estimé est $\text{logit}(P(Y \leq j|X)) = \alpha_j - \beta\ \text{precaire}$). Les /cut$_j$ sont les α_j du modèle. On voit que le coefficient β est positif et significativement différent de 0, avec p < 1‰.
On en déduit que la précarité est significativement associée à la fragilité en 5 classes avec p < 1‰. Le fait que le coefficient β soit positif indique qu'il y a un lien positif entre precaire et Y (et négatif entre precaire et $P(Y \leq j|X)$), c'est-à-dire qu'un niveau socio-économique plus précaire (precaire = 1) augmente la probabilité d'un score de fragilité Y plus élevé. C'est cette « facilité » d'interprétation qui justifie l'écriture du modèle avec un signe « – » (voir détails dans l'Annexe VIII.2). La valeur quantitative du coefficient, dont l'OR correspondant est $e^{0,949} = 2,58$, n'est pas facile à interpréter. D'une part, il faudrait tenir compte du fait (et l'exprimer) qu'il s'applique à toutes comparaisons entre P(Y ≤ j) et P(Y > j) pour j = 1 à 4. D'autre part, les effectifs des différentes classes de Y ne permettent pas d'exprimer l'OR comme un RR.

```
. ologit sf5 precaire, nolog

Ordered logistic regression                       Number of obs =    491
                                                  LR chi2(1)    =  29.74
                                                  Prob > chi2   = 0.0000
Log likelihood = -632.6304                        Pseudo R2     = 0.0230

------------------------------------------------------------------------------
         sf5 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    precaire |   .9491247   .1762013     5.39   0.000     .6037765    1.294473
-------------+----------------------------------------------------------------
       /cut1 |  -.8165032   .1180844                     -1.047944   -.5850619
       /cut2 |   1.111291   .1234992                      .8692369    1.353345
       /cut3 |   2.783431   .1839658                      2.422864    3.143997
       /cut4 |   3.986472   .2874977                      3.422987    4.549957
------------------------------------------------------------------------------
```

Tableau 6.9 : Enquête Septavih. Modèle *cumulative-odds* avec proportionnalité des *odds*

IV.4. Regroupement de classes de Y et test de l'hypothèse des *odds* proportionnels

J'ai indiqué plus haut que le « vrai » coefficient β ne change pas si on regroupe des classes (et si l'hypothèse de proportionnalité des *odds* est satisfaite).
Si on prend le score de Fried en 3 classes (sf3, voir Tableau 6.8), on obtient :

```
. ologit sf3 precaire, nolog

Ordered logistic regression                       Number of obs =    491
                                                  LR chi2(1)    =  26.14
                                                  Prob > chi2   = 0.0000
Log likelihood = -394.04634                       Pseudo R2     = 0.0321

------------------------------------------------------------------------------
         sf3 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    precaire |   1.044307   .2115657     4.94   0.000     .6296457    1.458968
-------------+----------------------------------------------------------------
       /cut1 |  -.7961434   .1199537                     -1.031248   -.5610385
       /cut2 |   2.839028   .1984283                      2.450116    3.22794
------------------------------------------------------------------------------
```

Tableau 6.10 : Enquête Septavih. Modèle *cumulative-odds* avec Y en 3 classes et proportionnalité des *odds*

On voit que le coefficient β estimé (1,04) est très proche du précédent (0,95), d'autant que les intervalles de confiance se chevauchent largement. C'est attendu si l'hypothèse de proportionnalité des *odds* est vérifiée. Réciproquement, ce résultat conforte cette hypothèse !

Si on veut tester formellement l'hypothèse de proportionnalité des *odds*, il faut utiliser un modèle qui ne la suppose pas – il s'écrit : $\text{logit}(P(Y \leq j|X)) = \alpha_j - \beta_j \text{ precaire}$ – et tester l'égalité des β_j pour j = 1 à 4. On obtient les résultats du Tableau 6.11.

```
. gologit2 sf5 precaire

Generalized Ordered Logit Estimates               Number of obs =    491
                                                  LR chi2(4)    =  31.13
                                                  Prob > chi2   = 0.0000
Log likelihood = -631.93319                       Pseudo R2     = 0.0240

------------------------------------------------------------------------------
         sf5 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
1            |
    precaire |   .9424115   .2423625     3.89   0.000     .4673897    1.417433
       _cons |   .8179974    .123827     6.61   0.000     .5753008    1.060694
-------------+----------------------------------------------------------------
2            |
    precaire |    .897994    .198202     4.53   0.000     .5095252    1.286463
       _cons |  -1.094274   .1316618    -8.31   0.000    -1.352326   -.8362215
-------------+----------------------------------------------------------------
3            |
    precaire |   1.251052   .3349865     3.73   0.000     .5944907    1.907614
       _cons |  -2.968704   .2647478   -11.21   0.000      -3.4876   -2.449808
-------------+----------------------------------------------------------------
4            |
    precaire |   1.133428   .5658064     2.00   0.045     .0244678    2.242388
       _cons |   -4.10099   .4509006    -9.10   0.000    -4.984739   -3.217241
------------------------------------------------------------------------------

. test ([0=1] : precaire) ([0=2] : precaire) ([0=3] : precaire)

 ( 1)  [0]precaire - [1]precaire = 0
 ( 2)  [0]precaire - [2]precaire = 0
 ( 3)  [0]precaire - [3]precaire = 0

           chi2(  3) =    1.35
         Prob > chi2 =    0.7162
```

Tableau 6.11 : Enquête Septavih. Test de l'hypothèse de proportionnalité des *odds* dans le modèle *cumulative-odds*

Ces résultats permettent, d'une part, de se rendre compte de ce que donne un modèle *cumulative-odds* sans l'hypothèse de proportionalité des *odds*. Le Tableau 6.11 contient 4 blocs, correspondant à chacune des 4 premières classes de Y.
Les quatre β_j sont différents ; ils correspondent aux différents seuils de Y. Ils sont assez proches, particulièrement quand on prend en compte leur intervalle de confiance. Le test de comparaison global est non significatif. L'hypothèse de proportionnalité des *odds* est donc acceptable ; on peut conserver le premier modèle avec un seul β.

Remarques

- Les /cut_j n'apparaissent pas dans le Tableau 6.9 ; ils sont indiqués sur les différentes lignes _cons des quatre blocs du Tableau 6.11. Ils ne sont pas identiques à ceux du Tableau 6.9, ce qui est normal, car l'hypothèse des *odds* proportionnels n'est pas faite dans le Tableau 6.11. On peut cependant noter qu'ils sont très proches numériquement[6], mais avec un signe opposé. Cela est dû au fait que la commande gologit2 de Stata modélise $\text{logit}(P(Y > j|X)) = \ln\left(\frac{P(Y > j|X)}{P(Y \leq j|X)}\right) = \alpha_j + \beta_j\ \text{precaire}$ (Williams R, 2006), ce qui est équivalent à $\text{logit}(P(Y \leq j|X)) = \ln\left(\frac{P(Y \leq j|X)}{P(Y > j|X)}\right) = -\alpha_j - \beta_j\ \text{precaire}$. Cela ne change rien pour l'interprétation des β_j, qui sont conservés, mais les signes des α_j sont inversés. Je vous avais dit qu'il fallait faire attention !
- Les résultats du Tableau 6.11 sont les mêmes (au signe des α_j près) que ceux qu'on obtiendrait avec 4 régressions logistiques binomiales $\text{logit}(P(Y \leq j|X)) = \ln\left(\frac{P(Y \leq j|X)}{P(Y > j|X)}\right) = \alpha_j - \beta_j\ \text{precaire}$.

IV.5. Présentation des résultats

La présentation des résultats du Tableau 6.9 faite précédemment respecte parfaitement le modèle utilisé et ses coefficients. Elle présente cependant l'inconvénient de ne pas permettre de quantifier de façon compréhensible la force de la relation entre X et Y puisque, comme je l'ai indiqué, l'interprétation quantitative de β est difficile.
Une solution possible est de donner la distribution de Y selon les valeurs de X, ce qui permet de montrer comment elle évolue. Il faut pour cela calculer la distribution de Y prédite par le modèle selon les classes de X, ce qui a un sens, puisque l'enquête porte sur un échantillon « non sélectionné »[7].
Les résultats, qui figurent dans le Tableau 6.12 pour la variable sf3, indiquent clairement qu'il y a une plus forte proportion d'individus fragiles (14 % versus 6 %) et

6. Cela est lié au fait que l'hypothèse des *odds* proportionnels est acceptable.

7. Je n'ose pas dire représentatif car ce n'est pas le cas, mais en tout cas, ce n'est pas un échantillon de type cas-témoins où la distribution de X est fixée dans le protocole.

une moins forte proportion d'individus robustes (14 % versus 31 %) chez les sujets précaires que chez les sujets non précaires. Au-delà des différences, ces pourcentages témoignent de l'ampleur du décalage de la distribution du score de fragilité quand on passe de sujets non précaires à des sujets précaires.

```
. qui ologit sf3 precaire, nolog

. predict robuste pre_fragile fragile
(option pr assumed; predicted probabilities)

. tabstat robuste pre_fragile fragile, stat(mean) by(precaire) format(%4.2f) nototal

Summary statistics: Mean
Group variable: precaire (score EPICES >=30)

precaire |   robuste  pre_fr~e   fragile
---------+------------------------------
       0 |      0.31      0.63      0.06
       1 |      0.14      0.72      0.14
------------------------------------------
```

Tableau 6.12 : Enquête Septavih. Distribution du score de fragilité selon le statut de précarité

On peut aller un peu plus loin encore, ou en tout cas exprimer les résultats différemment, en tenant compte du fait que la précarité est initialement mesurée par une variable quantitative « epices ».
Après avoir vérifié, avec la procédure mfp (voir chapitre 4, § VIII.4), que la variable epices pouvait être modélisée de façon linéaire, on peut inclure la variable epices dans le modèle sous sa forme initiale avec des valeurs de 0 à 100 et représenter la variation de la distribution de la fragilité avec le degré de précarité, comme cela est fait sur la Figure 6.1. Celle-ci montre que le pourcentage de sujets robustes décroît de 40 % à presque 0 et que le pourcentage de sujets fragiles croît de quelques pourcents à plus de 50 % lorsque le score de précarité varie de 0 à 100.

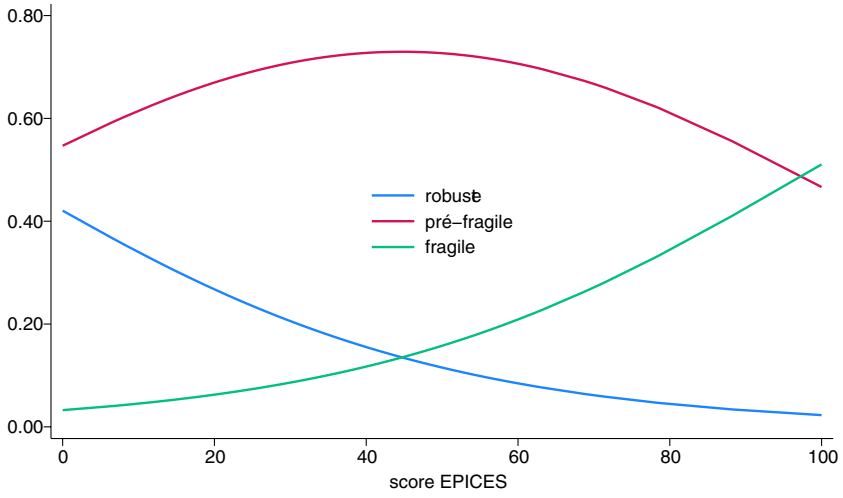


Figure 6.1 : Enquête Septavih : évolution du score de fragilité en 3 classes selon le degré de précarité (score continu EPICES).

IV.6. Plusieurs variables indépendantes

C'est lorsqu'il y a plusieurs variables indépendantes que le modèle ordinal prend tout son intérêt. Le choix des variables et la forme sous laquelle elles sont incluses dans le modèle obéissent aux mêmes règles que celles que j'ai indiquées pour le modèle logistique binomial (voir chapitre 5). S'ajoute ici le choix de l'hypothèse des *odds* proportionnels à faire pour chaque variable.
En poursuivant l'exemple précédent et en considérant les quatre variables indépendantes présentées au § IV.2, on peut estimer leur coefficient β lorsque l'hypothèse des *odds* proportionnels n'est pas imposée et tester cette hypothèse comme on l'a fait au § IV.4. On obtient les résultats du Tableau 6.13, dans lequel les coefficients du modèle avec *odds* proportionnels pour toutes les variables sont aussi indiqués dans la partie droite.

Variable indépendante	Coefficients sans l'hypothèse des *odds* proportionnels					Coefficients avec l'hypothèse des *odds* proportionnels**
	P(Y ≤ 1)/ P(Y > 1)	P(Y ≤ 2)/ P(Y > 2)	P(Y ≤ 3)/ P(Y > 3)	P(Y ≤ 4)/ P(Y > 4)	p*	
precaire	0,93	0,83	1,15	0,98	0,77	0,89
cd4_350	–0,14	0,36	0,20	0,041	0,43	0,15
mrc	0,11	0,13	0,87	0,63	0,11	0,19
parten	–0,17	–0,21	–0,12	–0,16	0,99	–0,19

* Degré de signification de la comparaison des quatre coefficients du modèle sans l'hypothèse des *odds* proportionnels.

** Hypothèse pour l'ensemble des quatre variables.

Tableau 6.13 : Enquête Septavih. Coefficients des 4 variables incluses dans le modèle *cumulative-odds* selon que l'hypothèse des *odds* proportionnels est imposée ou pas

On constate que l'hypothèse des *odds* proportionnels n'est rejetée pour aucune des variables, avec peut-être un doute pour la variable mrc, puisque le degré de signification n'est pas très loin du seuil de 5 %. On peut donc décider (en partie aussi parce que cela permet de poursuivre l'exemple…) d'utiliser un modèle où l'hypothèse des *odds* proportionnels est faite pour les 4 variables, sauf pour la variable mrc. Les résultats figurent dans le Tableau 6.14.

Ceux-ci commencent par la liste des contraintes correspondant à l'hypothèse des *odds* proportionnels, qui s'appliquent donc à toutes les variables, sauf à mrc. Les coefficients des variables autres que mrc sont donc les mêmes pour toutes les classes j de Y, alors que les coefficients de mrc varient. Leur variation peut s'interpréter comme un effet plus fort de mrc au-delà de la classe 3, c'est-à-dire pour le passage à la classe « fragile » (voir Tableau 6.8).

```
. gologit2 sf5 precaire cd4_350 mrc parten, npl(mrc)

Generalized Ordered Logit Estimates                   Number of obs =    491
                                                      LR chi2(7)    =  39.18
                                                      Prob > chi2   = 0.0000
Log likelihood = -627.90852                           Pseudo R2     = 0.0303

 ( 1)  [1]precaire - [2]precaire = 0
 ( 2)  [1]cd4_350 - [2]cd4_350 = 0
 ( 3)  [1]parten - [2]parten = 0
 ( 4)  [2]precaire - [3]precaire = 0
 ( 5)  [2]cd4_350 - [3]cd4_350 = 0
 ( 6)  [2]parten - [3]parten = 0
 ( 7)  [3]precaire - [4]precaire = 0
 ( 8)  [3]cd4_350 - [4]cd4_350 = 0
 ( 9)  [3]parten - [4]parten = 0
------------------------------------------------------------------------------
         sf5 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
1            |
    precaire |    .885921   .1828089     4.85   0.000     .5276222     1.24422
    cd4_350 |    .1567635   .2485991     0.63   0.528    -.3304818    .6440089
         mrc |    .1051412   .2186145     0.48   0.631    -.3233353    .5336178
      parten |   -.1827947   .1764485    -1.04   0.300    -.5286273     .163038
       _cons |    .8851017   .1931284     4.58   0.000      .506577    1.263626
-------------+----------------------------------------------------------------
2            |
    precaire |    .885921   .1828089     4.85   0.000     .5276222     1.24422
    cd4_350 |    .1567635   .2485991     0.63   0.528    -.3304818    .6440089
         mrc |    .1255863   .2007878     0.63   0.532    -.2679505    .5191231
      parten |   -.1827947   .1764485    -1.04   0.300    -.5286273     .163038
       _cons |   -1.056104   .1938097    -5.45   0.000    -1.435964   -.6762441
-------------+----------------------------------------------------------------
3            |
    precaire |    .885921   .1828089     4.85   0.000     .5276222     1.24422
    cd4_350 |    .1567635   .2485991     0.63   0.528    -.3304818    .6440089
         mrc |    .8762085   .3298433     2.66   0.008     .2297275    1.522689
      parten |   -.1827947   .1764485    -1.04   0.300    -.5286273     .163038
       _cons |   -3.113561   .2933373   -10.61   0.000    -3.688491   -2.53863
-------------+----------------------------------------------------------------
4            |
    precaire |    .885921   .1828089     4.85   0.000     .5276222     1.24422
    cd4_350 |    .1567635   .2485991     0.63   0.528    -.3304818    .6440089
         mrc |    .6614359   .5500412     1.20   0.229     -.416625    1.739497
      parten |   -.1827947   .1764485    -1.04   0.300    -.5286273     .163038
       _cons |   -4.205658   .4406252    -9.54   0.000    -5.069268   -3.342049
------------------------------------------------------------------------------
```

Tableau 6.14 : Enquête Septavih. Modèle *cumulative-odds* avec l'hypothèse des *odds* proportionnels, sauf pour la variable mrc

V. Modèle *continuation-ratio*

Comme je l'ai indiqué plus haut, le modèle *continuation-ratio* s'écrit : $\ln\left(\frac{P(Y=j\mid X)}{P(Y>j\mid X)}\right)=\alpha_j-\sum_{i=1}^{p}\beta_{ji}X_i$, où $j=1,\ \ldots,\ k-1$ désigne une classe de Y. Il modélise la probabilité conditionnelle $\delta_j=P(Y=j\mid Y\geq j)$ et convient particulièrement aux situations où l'ordre des k classes de Y correspond à des étapes qui doivent être franchies successivement avant d'atteindre le seuil j. En fait, le modèle *continuation-ratio* est essentiellement le modèle de Cox pour des données de survie avec des temps discrets auquel il est (« quasiment ») équivalent (Armstrong BG et al., 1989).

Contrairement au modèle *cumulative-odds*, il n'est pas invariant lorsqu'on change l'ordre de Y ni lorsqu'on regroupe des catégories de Y. Il n'est pas non plus utilisable dans des enquêtes de type cas-témoins.

V.1. Exemple : rang de succès en FIV

Les données de cet exemple sont celles de tentatives de FIV entre 1986 et 1995 chez 5 000 couples qui ont été suivis lors de leur parcours ; on s'intéresse ici au rang de succès Y, c'est-à-dire le rang de la tentative où une grossesse a été obtenue. Il est compté jusqu'à 4. Au-delà, ou en cas d'échec, Y est codé 9. La variable correspondante est notée rgsuc.

Les facteurs de succès étudiés ici à titre d'exemple sont :

- l'âge de la femme lors de la première tentative, transformé en variable dichotomique notée age35 (0 si l'âge de la femme est inférieur à 35 ans ; 1 sinon) ;
- le nombre d'années d'infécondité avant la première tentative, variable notée durinfec. Elle est transformée, d'une part, en une variable dichotomique dinf5 (0 si durinfec < 5 ; 1 si durinfec ≥ 5) et, d'autre part, en une variable qualitative à 4 classes notée dinf (1 si durinfec = 1–2 ans ; 3 si durinfec = 3–4 ans ; 5 si durinfec = 5–8 ans ; 9 si durinfec ≥ 9 ans).

La répartition des rangs de succès est indiquée dans le Tableau 6.15.

Rang de succès	1	2	3	4	> 4 ou échec	Total
Effectif	797	402	172	85	3544	5000
Fréquence	15,9 %	8,0 %	3,4 %	1,7 %	70,9 %	100 %

Tableau 6.15 : Succès en FIV. Répartition des rangs de succès

Les résultats du modèle *continuation-ratio* avec la contrainte d'*odds* proportionnels, c'est-à-dire que les β_{ij} ne dépendent pas de j, $\ln\left(\frac{P(Y = j \mid X)}{P(Y > j \mid X)}\right) = \alpha_j - \beta\ \text{age35}$, figurent dans le Tableau 6.16.

```
. gencrm rgsuc age35,or nolog

Ordered Logit Estimates                           Number of obs =  5,000
                                                  Wald chi2(1)  =  48.39
Log likelihood = -4597.1565                       Prob > chi2   = 0.0000

------------------------------------------------------------------------------
       rgsuc | Odds ratio   Std. err.      z    P>|z|     [95 conf. interval]
-------------+----------------------------------------------------------------
       age35 |   1.548502   .0973465     6.96   0.000     1.368992     1.75155
-------------+----------------------------------------------------------------
        /tau1 |   -1.52998   .0423257   -36.15   0.000    -1.612937    -1.447024
        /tau2 |  -2.10981   .0553036   -38.15   0.000    -2.218203    -2.001417
        /tau3 |  -2.911437   .0799632   -36.41   0.000    -3.068162    -2.754712
        /tau4 |  -3.591756   .1111391   -32.32   0.000    -3.809584    -3.373927
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to odds ratios.
```

Tableau 6.16 : Succès en FIV. Modèle *continuation-ratio* avec *odds* proportionnels

L'odds ratio associé à age35 est $e^{\beta} = 1{,}55$ et les /tauj sont les constantes α_j du modèle (on peut noter que, malgré l'option « or », c'est α_j qui est donné et non pas e^{α_j}). L'interprétation est qu'à chaque tentative, les femmes de plus de 35 ans ont un rang de succès Y significativement plus élevé que les femmes de moins de 35 ans,

c'est-à-dire un risque d'échec plus élevé à chaque tentative[8]. Cette augmentation du risque est quantifiée par un odds ratio égal à 1,55 [1,37 ; 1,75], qu'on ne peut pas interpréter facilement, en tout cas pas comme un risque relatif, car le risque d'échec à chaque tentative est important (de l'ordre de 80 à 90 %).

V.2. Test de l'hypothèse des *odds* proportionnels

Pour vérifier si l'hypothèse d'*odds* proportionnels est satisfaite, il faut passer par un modèle où elle n'est pas requise. Les résultats figurent dans le Tableau 6.17.

```
. gencrm rgsuc age35,or nolog free(age35)

Ordered Logit Estimates                           Number of obs =  5,000
                                                  Wald chi2(4)  =  49.42
Log likelihood = -4596.4951                       Prob > chi2   = 0.0000

------------------------------------------------------------------------------
       rgsuc | Odds ratio   Std. err.      z    P>|z|     [95 conf. interval]
-------------+----------------------------------------------------------------
eq1          |
       age35 |   1.599288   .1401035     5.36   0.000     1.346971    1.898869
-------------+----------------------------------------------------------------
eq2          |
       age35 |   1.415608   .1637906     3.00   0.003     1.128382    1.775946
-------------+----------------------------------------------------------------
eq3          |
       age35 |   1.516798   .2647585     2.39   0.017     1.077332     2.13553
-------------+----------------------------------------------------------------
eq4          |
       age35 |   1.878228   .4847149     2.44   0.015     1.132605    3.114714
-------------+----------------------------------------------------------------
       /tau1 |  -1.521295   .0452524   -33.62   0.000    -1.609988   -1.432602
       /tau2 |  -2.135152   .0623116   -34.27   0.000    -2.257281   -2.013024
       /tau3 |  -2.917122   .0918298   -31.77   0.000    -3.097105   -2.737139
       /tau4 |  -3.542519   .1258167   -28.16   0.000    -3.789115   -3.295923
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first 4 equations to odds ratios.
```

Tableau 6.17 : Succès en FIV. Modèle *continuation-ratio* sans l'hypothèse des *odds* proportionnels (option free())

On voit bien ici que les 4 odds ratios correspondant aux 4 valeurs des β_j sont très proches. C'est ce que confirme le test de leur comparaison (Tableau 6.18), qui peut être fait en comparant les deux modèles des Tableaux 6.16 et 6.17 par un test de rapport des vraisemblances, car ils sont emboîtés, ou par un test de Wald correspondant aux trois égalités des OR (ou des coefficients) requises par l'hypothèse des *odds* proportionnels sur les coefficients du modèle. On note au passage que ces deux tests donnent des résultats très proches, comme cela est attendu, car ils sont asymptotiquement équivalents (voir chapitre 2, § IV.1.f).
On conclut donc que l'hypothèse des *odds* proportionnels peut être conservée, ce qui revient à dire ici que l'augmentation du risque d'échec associée à un âge supérieur à 35 ans est la même à chaque tentative.

8. Vous pouvez faire une pause à cet endroit pour méditer cette phrase apparemment simple ! Et comprendre que c'est la présence du signe « – » devant β dans l'écriture du modèle qui la rend exacte. Moi-même, je bute à chaque lecture…

```
. qui gencrm rgsuc age35,or nolog
. est store mprop
. qui gencrm rgsuc age35,or nolog free(age35)
. est store mfree
. lrtest mprop mfree

Likelihood-ratio test
Assumption: mprop nested within mfree

 LR chi2(3) =   1.32
Prob > chi2 = 0.7237

. qui gencrm rgsuc age35,or nolog free(age35)
. test _b[eq1:age35] = _b[eq2:age35] = _b[eq3:age35] = _b[eq4:age35]

 ( 1)  [eq1]age35 - [eq2]age35 = 0
 ( 2)  [eq1]age35 - [eq3]age35 = 0
 ( 3)  [eq1]age35 - [eq4]age35 = 0

           chi2(  3) =    1.31
         Prob > chi2 =    0.7267
```

Tableau 6.18 : Succès en FIV. Test de l'hypothèse des *odds* proportionnels

Si on ajoute dinf5, la durée de l'infécondité en deux classes (plus ou moins de 5 ans), comme variable indépendante supplémentaire, on peut vérifier de même qu'un modèle à *odds* proportionnels convient. On compare pour cela un modèle où l'hypothèse des *odds* proportionnels est faite pour les deux variables age35 et dinf5 à un modèle où elle n'est faite que pour age35. C'est l'occasion de souligner que l'hypothèse des *odds* proportionnels peut n'être faite que pour une partie des variables indépendantes, comme on l'a vu pour le modèle *cumulative-odds* (voir § IV.6). Cela donne une certaine souplesse pour l'utilisation des modèles ordinaux, bien qu'il faille admettre qu'on n'a pas toujours (voire rarement) des arguments solides pour faire l'hypothèse des *odds* proportionnels pour telle variable et pas pour telle autre.

V.3. Modélisation de la durée d'infécondité

Au-delà de la question de la proportionnalité des *odds*, on peut constater dans cet exemple que la durée d'infécondité prise sous forme dichotomique (plus ou moins de cinq ans) n'est pas associée au rang de succès (voir Tableau 6.19).

```
. gencrm rgsuc dinf5,or nolog
Ordered Logit Estimates                           Number of obs =  5,000
                                                  Wald chi2(1)  =   1.90
Log likelihood = -4621.6535                       Prob > chi2   = 0.1681
------------------------------------------------------------------------------
       rgsuc | Odds ratio   Std. err.      z    P>|z|     [95 conf. interval]
-------------+----------------------------------------------------------------
       dinf5 |   1.080232    .060483     1.38   0.168     .9679609    1.205526
-------------+----------------------------------------------------------------
       /tau1 |  -1.625883   .0467637   -34.77   0.000    -1.717539   -1.534228
       /tau2 |  -2.209783   .0586525   -37.68   0.000     -2.32474   -2.094827
       /tau3 |   -3.01223   .0823646   -36.57   0.000    -3.173662   -2.850799
       /tau4 |  -3.693342    .112877   -32.72   0.000    -3.914577   -3.472107
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to odds ratios.
```

Tableau 6.19 : Succès en FIV. Relation avec la durée d'infécondité en variable dichotomique

Pour des raisons de puissance, il peut être préférable d'utiliser la variable qualitative à 4 classes dinf définie dans le § V.1. Les résultats obtenus figurent dans le

Tableau 6.20, avec la variable dinf décomposée en variables indicatrices. Ils montrent que l'augmentation du risque d'échec par rapport à la classe 1–2 ans d'infécondité est de plus en plus forte lorsque la durée d'infécondité augmente, puisque l'OR passe de 0,98 à 1,21.

```
. gencrm rgsuc i.dinf,or nolog

Ordered Logit Estimates                           Number of obs =  5,000
                                                  Wald chi2(3)  =   4.71
Log likelihood = -4620.1713                       Prob > chi2   = 0.1945

------------------------------------------------------------------------------
       rgsuc | Odds ratio   Std. err.      z    P>|z|     [95 conf. interval]
-------------+----------------------------------------------------------------
        dinf |
    3-4 ans  |   .9840066    .080206    -0.20   0.843     .8387198    1.154461
    5-8 ans  |   1.026458   .0834904     0.32   0.748     .8751967    1.203861
    >=9 ans  |   1.207243    .129228     1.76   0.079     .9787654    1.489054
-------------+----------------------------------------------------------------
        /tau1 |  -1.636745   .0717682   -22.81   0.000    -1.777408   -1.496082
        /tau2 |    -2.2205   .0799867   -27.76   0.000    -2.377272   -2.063729
        /tau3 |  -3.022766   .0988136   -30.59   0.000    -3.216437   -2.829095
        /tau4 |  -3.703724   .1253554   -29.55   0.000    -3.949416   -3.458032
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to odds ratios.
```

Tableau 6.20 : Succès en FIV. Prise en compte de la durée d'infécondité en cinq classes (dinf) sous forme de quatre variables indicatrices

Cela incite, pour tester cette tendance à l'augmentation des OR, à inclure la variable durée d'infécondité sous sa forme initiale dinf semi-quantitative (codée 1 3 5 9). C'est ce qui est fait dans le Tableau 6.21, qui vérifie d'abord que le modèle avec la variable dinf n'est pas significativement différent du modèle avec une décomposition en variables indicatrices (celui du Tableau 6.20), ce qui autorise à utiliser dinf sans transformation.

Le Tableau 6.21 montre ensuite qu'il existe un lien à la limite de la signification statistique (p = 0,055) entre la durée d'infécondité et le risque d'échec en FIV, avec une relation de type dose-effet.

Remarques

- Les résultats précédents soulignent à nouveau l'importance de prendre en compte le mieux possible le caractère quantitatif des variables indépendantes dans l'analyse. On a ici une meilleure puissance avec dinf qu'avec dinf5 (et même qu'avec dinf décomposée en variables indicatrices, puisqu'on s'est assuré de la linéarité).
- La notion de relation dose-effet dont j'ai parlé s'applique à la « dose » durée d'infécondité. Le fait que le rang de succès Y, qui est ici l'effet, soit aussi une variable ordonnée rend « seulement » l'expression du résultat plus compliquée. Par exemple, la représentation graphique d'une relation dose-effet, que j'ai beaucoup utilisée dans le chapitre 4, devient ici quasi impossible.
- Cet exemple montre aussi les limites de ce que je vous ai dit dans le chapitre 4 à propos de la supériorité de la modélisation d'une variable quantitative par rapport à une transformation en classes. On est ici peu tenté de rajouter des splines ou des polynômes fractionnaires ! Le monde est plein d'injonctions contradictoires…

```
. qui gencrm rgsuc i.dinf,or nolog
. est store ind
. qui gencrm rgsuc dinf,or nolog
. est store lin
. lrtest lin ind

Likelihood-ratio test
Assumption: lin nested within ind

 LR chi2(2) =   1.14
Prob > chi2 = 0.5647

. gencrm rgsuc dinf,or nolog

Ordered Logit Estimates                           Number of obs =  5,000
                                                  Wald chi2(1)  =   3.68
Log likelihood = -4620.7428                       Prob > chi2   = 0.0549

------------------------------------------------------------------------------
       rgsuc | Odds ratio   Std. err.      z    P>|z|     [95 conf. interval]
-------------+----------------------------------------------------------------
        dinf |   1.023738   .0125133     1.92   0.055     .9995035     1.04856
-------------+----------------------------------------------------------------
       /tau1 |  -1.567043   .0626548   -25.01   0.000    -1.689844   -1.444241
       /tau2 |  -2.150911    .071931   -29.90   0.000    -2.291893   -2.009929
       /tau3 |   -2.95309   .0923555   -31.98   0.000    -3.134104   -2.772077
       /tau4 |  -3.634159   .1203623   -30.19   0.000    -3.870065   -3.398253
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to odds ratios.
```

Tableau 6.21 : Succès en FIV. Prise en compte de la durée d'infécondité en 5 classes (dinf) sous forme de la variable d'origine dinf

VI. Modèle *adjacent-category*

Le modèle *adjacent-category* s'écrit : $\ln\left(\frac{P(Y=j\mid X)}{P(Y=j+1\mid X)}\right)=\alpha_j+\sum_{i=1}^{p}\beta_{ji}X_i$, où $j=1, \ldots, k-1$ désigne une classe de Y. Il s'agit en réalité de $(k-1)$ modèles logistiques binomiaux comparant successivement les catégories $Y=j$ et $Y=j+1$.

C'est finalement la même chose que le modèle logistique multinomial. En effet, en partant du modèle multinomial, qui s'écrit $\ln\left(\frac{P(Y=j\mid X)}{P(Y=1\mid X)}\right)=\alpha_j+\sum_{i=1}^{p}\beta_{ji}X_i$, on peut facilement calculer :

$$\ln\left(\frac{P(Y=j+1\mid X)}{P(Y=j\mid X)}\right)=\ln\left(\frac{P(Y=j+1\mid X)}{P(Y=1\mid X)}\right)-\ln\left(\frac{P(Y=j+1\mid X)}{P(Y=1\mid X)}\right)=(\alpha_{j+1}-\alpha_j)+\sum_{i=1}^{p}(\beta_{(j+1)i}-\beta_{ji})X_i=\alpha'_j+\sum_{i=1}^{p}\beta'_{ji}X_i$$

Cela revient à changer de catégorie de référence comme on l'a fait au § II.5, dans un autre contexte. On retrouve alors exactement le modèle *adjacent-category*, qui n'apporte finalement rien de nouveau.

Le modèle *adjacent-category* est souvent présenté avec l'hypothèse des *odds* proportionnels. Il s'écrit alors : $\ln\left(\frac{P(Y=j\mid X)}{P(Y=j+1\mid X)}\right)=\alpha_j+\sum_{i=1}^{p}\beta_iX_i$. Avec ce qui vient d'être dit, on voit que c'est la même chose qu'un modèle multinomial où on imposerait les contraintes $\beta_{ji}=j\times\beta_{1j}$ (Liu I et al., 2005, Fagerland MW, 2014). On retrouve alors le caractère ordonné de Y, puisque les odds ratios comparant la catégorie $Y=j$ à la référence $Y=1$ sont eux-mêmes ordonnés pour chaque variable X_i : $OR_{ij}=OR_{i1}^{j}$.

Le modèle *adjacent-category* n'a donc de réel intérêt qu'avec l'hypothèse des *odds* proportionnels, qu'il faut bien sûr vérifier. Il semble qu'il ait été introduit dans le domaine des sciences sociales pour analyser des réponses telles que « tout à fait en accord », « en accord », « ni en accord, ni en désaccord », « en désaccord », « tout à fait en désaccord », en particulier lorsque l'intérêt se porte sur la comparaison entre les catégories « extrêmes » (les deux premières ou les deux dernières) (Sobel ME, 1997), ce que les modèles ordinaux précédents ne permettent pas de faire.

Je vais illustrer cela avec des résultats obtenus dans le cadre d'une enquête sur la santé des adolescents réalisée en 2013 (Jousselme C et al., 2013, Ibrahim N et al., 2023).
Parmi les questions posées aux adolescents, j'ai retenu ici comme variable Y la satisfaction de la relation avec leurs amis évaluée au moyen d'une variable, notée satis, à 4 classes : 1 « Très satisfait(e) » ; 2 « Satisfait(e) » ; 3 « Ni satisfait(e), ni insatisfait(e) » ; 4 « Pas très satisfait(e) » ; 5 « Pas satisfait(e) du tout ». La variable X est tout d'abord le genre, puis le fait d'avoir une maladie chronique (MC). Elles sont toutes deux dichotomiques, le genre est codé 1 pour les garçons et 2 pour les filles et MC est codée 0/1.

Les résultats du Tableau 6.22 montrent que le niveau de satisfaction de la relation avec leurs amis est plus élevé chez les femmes que chez les hommes. L'odds ratio comparant les catégories adjacentes de satis, égal à $e^{0,104} = 1,11$, est cependant peu élevé, bien que nettement significativement différent de 1 (en partie « à cause » de la grande taille de l'échantillon).

```
. adjcatlogit satis genre

Adjacent-category logistic regression                 Number of obs =   14159
                                                      LR chi2( 1)   =   20.62
                                                      Prob < chi2   =  0.0000
Log likelihood = -1.342e+04                           Pseudo R2     =  0.0008

------------------------------------------------------------------------------
       satis | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
satis        |
       genre |   .1035269   .0228787     4.53   0.000     .0586854    .1483684
-------------+----------------------------------------------------------------
_anc         |
       cons1 |  -.6617739   .0392354   -16.87   0.000    -.7386738    -.584874
       cons2 |   -2.04876   .0703232   -29.13   0.000     -2.18659   -1.910929
       cons3 |  -1.237691   .0961164   -12.88   0.000    -1.426076   -1.049306
       cons4 |  -1.058525   .1241471    -8.53   0.000    -1.301849   -.8152012
------------------------------------------------------------------------------
```

Tableau 6.22 : Analyse de la satisfaction de la relation avec les amis selon le genre (avec l'hypothèse des *odds* proportionnels)

On peut vérifier qu'on retrouverait exactement les mêmes résultats avec un modèle multinomial auquel on ajouterait les contraintes que j'ai indiquées plus haut : $\beta_{ji} = j \times \beta_{1j}$, qui correspondent à l'hypothèse des *odds* proportionnels.
Le modèle multinomial sans contraintes donne les résultats du Tableau 6.23.
On voit que les coefficients des différentes catégories de Y s'écartent un peu des contraintes $\beta_{ji} = j \times \beta_{1j}$ (surtout pour $j = 5$). La comparaison des modèles multinomiaux

avec et sans contraintes est non significative (p = 0,20, non montré dans les tableaux). On peut donc admettre l'hypothèse des *odds* proportionnels et résumer le lien entre satis et genre par l'OR = 1,11 du Tableau 6.22, qui compare deux catégories adjacentes.

```
. mlogit satis genre, b(1)

...

Multinomial logistic regression                 Number of obs = 14,159
                                                LR chi2(4)    =  25.22
                                                Prob > chi2   = 0.0000
Log likelihood = -13418.019                     Pseudo R2     = 0.0009

------------------------------------------------------------------------------
       satis | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
       genre |   .1005915   .0361803     2.78   0.005     .0296794    .1715037
       _cons |  -.6572058   .0580664   -11.32   0.000    -.7710139   -.5433977
-------------+----------------------------------------------------------------
3            |
       genre |   .3202999   .0777345     4.12   0.000      .167943    .4726568
       _cons |  -2.888762   .1288336   -22.42   0.000    -3.141272   -2.636253
-------------+----------------------------------------------------------------
4            |
       genre |   .2954392    .12916     2.29   0.022     .0422903    .5485882
       _cons |  -3.924048   .213859   -18.35   0.000    -4.343204   -3.504892
-------------+----------------------------------------------------------------
5            |
       genre |   .1158167   .1979446     0.59   0.558    -.2721475     .503781
       _cons |  -4.536663   .3200524   -14.17   0.000    -5.163955   -3.909372
------------------------------------------------------------------------------
```

Tableau 6.23 : Analyse de la satisfaction de la relation avec les amis selon le genre par régression logistique multinomiale sans contraintes

Les résultats sont différents avec la variable MC (présence d'une maladie chronique).

```
. adjcatlogit satis MC

Adjacent-category logistic regression           Number of obs =   14159
                                                LR chi2( 1)   =    7.54
                                                Prob < chi2   =  0.0060
Log likelihood = -1.343e+04                     Pseudo R2     =  0.0003

------------------------------------------------------------------------------
       satis | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
satis        |
          MC |   .0921793   .0331155     2.78   0.005     .0272741    .1570845
-------------+----------------------------------------------------------------
_anc         |
       cons1 |  -.5153654   .0185071   -27.85   0.000    -.5516386   -.4790922
       cons2 |  -1.900631   .0346576   -54.84   0.000    -1.968559   -1.832703
       cons3 |  -1.087935   .0521448   -20.86   0.000    -1.190137   -.9857333
       cons4 |  -.9072508   .0768008   -11.81   0.000    -1.057778   -.7567239
------------------------------------------------------------------------------
```

Tableau 6.24 : Analyse de la satisfaction de la relation avec les amis selon la présence ou non d'une maladie chronique (avec l'hypothèse des *odds* proportionnels)

Le Tableau 6.24 montre qu'avec l'hypothèse des *odds* proportionnels, il y a un lien entre la présence d'une maladie chronique et la satisfaction de la relation avec les amis, avec un coefficient positif et significatif (p = 0,05) pour la variable MC. Si on ne fait pas l'hypothèse des *odds* proportionnels, le Tableau 6.25 montre qu'il y a encore un lien entre la présence d'une maladie chronique et la satisfaction de la relation avec

les amis, puisque le χ^2, qui teste globalement les coefficients de la variable MC, est significatif (LR chi2(4) = 15,04, p = 0,005). Mais la variation de ces coefficients est non monotone avec la valeur de Y, et très éloignée de la contrainte $\beta_{ji} = j \times \beta_{1j}$.

```
. mlogit satis MC, b(1)
...
Multinomial logistic regression                     Number of obs = 14,159
                                                    LR chi2(4)    =  15.04
                                                    Prob > chi2   = 0.0046
Log likelihood = -13423.111                         Pseudo R2     = 0.0006
------------------------------------------------------------------------------
       satis | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
          MC |   .0201014   .0554606     0.36   0.717    -.0885994    .1288022
       _cons |  -.5065108   .0192606   -26.30   0.000    -.5442609   -.4687607
-------------+----------------------------------------------------------------
3            |
          MC |   .3405669   .1057011     3.22   0.001     .1333966    .5477372
       _cons |   -2.43977   .0417072   -58.50   0.000    -2.521514   -2.358025
-------------+----------------------------------------------------------------
4            |
          MC |   .0260104   .1947094     0.13   0.894     -.355613    .4076338
       _cons |   -3.47035   .0679985   -51.04   0.000    -3.603624   -3.337075
-------------+----------------------------------------------------------------
5            |
          MC |   .6249215   .2466465     2.53   0.011     .1415032     1.10834
       _cons |  -4.458681   .1103978   -40.39   0.000    -4.675057   -4.242305
------------------------------------------------------------------------------
```

Tableau 6.25 : Analyse de la satisfaction de la relation avec les amis selon la présence ou non d'une maladie chronique par régression logistique multinomiale sans contraintes

Le test de l'hypothèse des risques proportionnels, réalisé comme on l'a fait précédemment avec la variable genre, donne ici p = 0,06 (il ne figure pas dans les tableaux). Sur le plan formel, on ne rejette pas cette hypothèse, mais le degré de signification est très proche du seuil de 5 %. Il serait donc hasardeux de conclure qu'on peut accepter l'hypothèse des risques proportionnels. Cela est confirmé par l'examen des coefficients du Tableau 6.25. Il ne paraît donc pas pertinent de résumer le lien entre MC et la satisfaction de la relation avec les amis par le seul coefficient 0,09 du Tableau 6.24.

VII. Choix du modèle

VII.1. Guides pour choisir un modèle ordinal

Le choix d'un modèle ordinal repose sur la nature de l'ordre entre les classes de Y, ou sur l'interprétation qu'on fait de cet ordre, et sur les types de comparaisons entre les classes qui paraissent intéressantes pour répondre à la question qu'on se pose (Ananth CV et al., 1997, Fullerton AS, 2009, Bauldry S et al., 2018, Bürkner P-C et al., 2019).

- Lorsque Y est une échelle ordinale qui repose sur une mesure continue T sous-jacente, non observable et mise en classes, le modèle correspondant est le modèle *cumulative-odds* (voir détails dans l'Annexe VIII.2). C'est le cas par exemple des échelles d'incapacité, de satisfaction ou de douleur.

- Lorsqu'il n'y a pas de variable continue sous-jacente, mais que l'ordre des classes de Y représente une progression par étapes avec des points de départ et d'arrivée, le modèle correspondant est le modèle *continuation-ratio*.
  Dans ce cas, le fait de parvenir à une étape (c'est-à-dire à un certain niveau de Y) suppose que les étapes précédentes ont été franchies. Pour arriver au niveau Y = 3, il faut déjà être passé par les niveaux 1 et 2, comme, par exemple, pour le niveau d'études ou le nombre de fausses couches. D'autres situations sont envisageables, par exemple, Y = mort-né, vivant malformé, vivant non malformé. On s'intéresse alors à la probabilité d'une naissance vivante (sachant qu'il y a accouchement) et à la probabilité d'une naissance vivante non malformée (sachant qu'il y a une naissance vivante), c'est-à-dire une probabilité conditionnelle. Ce modèle est aussi adapté au cas où Y est une variable de durée ou de comptage en classes et censurée qu'on peut aussi analyser avec un modèle de Cox.

- Lorsqu'on s'intéresse aux comparaisons séquentielles entre les catégories de Y, plutôt qu'aux comparaisons entre catégories cumulées, le modèle de choix est le modèle *adjacent-category*.
  Contrairement aux deux modèles précédents, le modèle *adjacent-category* ne découle pas d'un modèle théorique sur l'ordre des valeurs de Y (variable quantitative sous-jacente ou progression par étapes). Ses propriétés mathématiques en font cependant une alternative intéressante (Bürkner P-C et al., 2019) et il peut s'appliquer au domaine de la santé (Edlinger M et al., 2022). On a vu au § VI que le modèle *adjacent-category* est un cas particulier du modèle multinomial avec des contraintes spécifiques.

Je mentionne enfin le modèle *stereotype*, moins utilisé et qui est aussi un cas particulier du modèle multinomial, qui correspond aux situations où l'ordre des classes de Y est construit à partir de plusieurs facteurs et résulte donc d'une construction multidimensionnelle (Greenland S, 1994, Ananth CV et al., 1997) : par exemple, le score d'Apgar utilisé pour caractériser la vitalité d'un nouveau-né et qui est construit en amalgamant 5 mesures individuelles cotées de 0 à 2 (rythme cardiaque, rythme respiratoire, tonus, coloration cutanée, réactivité aux stimuli).

Au-delà du choix du modèle, il faut choisir à quelles variables indépendantes s'applique l'hypothèse des *odds* proportionnels : à toutes, à aucune, ou à une partie d'entre elles. Il n'y a pas souvent de connaissances fortes pour faire ce choix, mais il est important de se poser la question et de tester si ces hypothèses sont satisfaites. J'en ai montré des exemples dans les § IV à VI.
Enfin, il faut se préoccuper du choix des classes de Y. Ces classes sont souvent prédéfinies, mais la question de leur regroupement éventuel se pose. Il n'y a que pour le modèle *cumulative-odds*, si l'hypothèse des *odds* proportionnels est satisfaite, qu'un regroupement ne change pas les résultats et leur interprétation.

VII.2. Un peu d'humilité sur l'importance du choix…

Après ces pages sur les différents modèles et leurs caractéristiques, et les quelques lignes qui précèdent sur les façons de les différencier et de les choisir, il est salutaire de reconnaître qu'on ne sait pas toujours lequel choisir. Et que ce n'est peut-être pas trop grave…

Reprenons l'exemple du rang de succès en FIV présenté au § V.1 et analysé alors avec le modèle *continuation-ratio*. Les résultats de cette analyse avec les variables age35 et dinf5 et l'hypothèse des *odds* proportionnels est la suivante (Tableau 6.26, identique au Tableau 6.19, mais avec des coefficients au lieu des odds ratios).

```
. gencrm rgsuc age35 dinf5, nolog
Ordered Logit Estimates                           Number of obs =  5,000
                                                  Wald chi2(2)  =  49.05
Log likelihood = -4596.8215                       Prob > chi2   = 0.0000
------------------------------------------------------------------------------
       rgsuc | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       age35 |   .4334064   .0630431     6.87   0.000     .3098442    .5569685
       dinf5 |   .0460338   .0562552     0.82   0.413    -.0642243    .1562919
-------------+----------------------------------------------------------------
        /tau1 |  -1.509206   .0492497   -30.64   0.000    -1.605734   -1.412678
        /tau2 |  -2.089068   .0607337   -34.40   0.000    -2.208104   -1.970032
        /tau3 |  -2.890621   .0838268   -34.48   0.000    -3.054918   -2.726323
        /tau4 |  -3.570946   .1139454   -31.34   0.000    -3.794275   -3.347617
------------------------------------------------------------------------------
```

Tableau 6.26 : Succès en FIV. Modèle *continuation-ratio* avec *odds* proportionnels

Le choix de cette analyse est fondé sur le fait que l'ordre de la variable Y (rang de succès) se comprend comme des étapes qui doivent être franchies successivement. Mais l'ordre de Y peut aussi être compris comme le reflet d'une variable quantitative sous-jacente (le nombre de tentatives de FIV), ce qui conduirait à utiliser le modèle *cumulative-odds*. Si ce choix est fait, on obtient les résultats du Tableau 6.27.

```
. ologit rgsuc age35 dinf5

....

Ordered logistic regression                       Number of obs =  5,000
                                                  LR chi2(2)    =  51.23
                                                  Prob > chi2   = 0.0000
Log likelihood = -4596.9906                       Pseudo R2     = 0.0055

------------------------------------------------------------------------------
       rgsuc | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       age35 |   .4727989   .0683407     6.92   0.000     .3388535    .6067442
       dinf5 |   .0455097   .0617504     0.74   0.461    -.0755189    .1665383
-------------+----------------------------------------------------------------
        /cut1 |  -1.499031   .0511801                     -1.599342   -1.39872
        /cut2 |  -.9866308   .0474003                     -1.079534   -.8937279
        /cut3 |   -.804755   .0465327                     -.8959574   -.7135526
        /cut4 |  -.7200903   .0462067                     -.8106538   -.6295268
------------------------------------------------------------------------------
```

Tableau 6.27 : Succès en FIV. Modèle *cumulative odds* avec *odds* proportionnels

Certes, les modélisations ne sont pas les mêmes et ce n'est pas parce que les coefficients β_i sont très proches que les modèles deviennent les mêmes. Mais, quand on se souvient que l'interprétation qu'on est capable de faire des résultats de ces

modèles porte surtout sur le signe et la signification statistique de ces coefficients, il faut reconnaître que le choix d'un modèle plutôt que de l'autre a une portée limitée sur les conclusions qu'on peut tirer touchant l'association entre les X_i et Y.

Cette constatation, qui pousse à une certaine humilité, n'est pas propre aux modèles logistiques ordinaux. Il arrive que les biostatistiques proposent des modèles faits pour analyser tel ou tel type de données (ici une variable Y qualitative ordonnée), mais dont les paramètres sont difficiles à interpréter sur le plan quantitatif et laissent l'utilisateur (moi compris) avec plus d'interrogations que de solutions.

VIII. Annexes

VIII.1. Deux écritures du modèle multinomial

L'écriture initiale du modèle logistique multinomial, celle qui le fait apparaître comme un cumul de modèles logistiques binomiaux, est la suivante :

$\ln\left(\frac{P(Y=j\mid X)}{P(Y=1\mid X)}\right) = \alpha_j + g_j(X)$, avec $g_j(X) = \sum_{i=1}^{p}\beta_{ji}X_i$ et où j = 2, ..., k sont les classes de Y autres que 1 et $X = (X_1,\ldots,X_p)$ sont les p variables indépendantes incluses dans le modèle.

Cette expression du modèle multinomial est équivalente à :

$$P(Y=j\mid X) = \frac{\exp\{\alpha_j + g_j(x)\}}{1+\sum_{j=1}^{k}\exp\{\alpha_j + g_j(x)\}}, \text{ avec } j=1, \ldots, k \text{ et } \alpha_1 = \beta_{1j} = 0.$$

Pour le montrer, il faut partir de la première expression, d'où on déduit :

$P(Y=j\mid X) = P(Y=1\mid X)\times\exp\left(\alpha_j + g_j(x)\right)$ pour j = 2, ..., k.

Puis, en additionnant ces équations pour j = 1, ..., k, et en tenant compte qu'on a, quel que soit j : $\sum_{j=1}^{k}P(Y=j\mid X) = 1$, on obtient : $1 = P(Y=1\mid X)\left(1+\sum_{j=2}^{k}\exp\left(\alpha_j + g_j(x)\right)\right)$.

Ce qui donne : $P(Y=1\mid X) = \frac{1}{1+\sum_{j=2}^{k}\exp\left(\alpha_j + g_j(x)\right)}$

et

$$P(Y=j\mid X) = P(Y=1\mid X)\times\exp\left(\alpha_j + g_j(x)\right) = \frac{\exp\{\alpha_j + g_j(x)\}}{1+\sum_{j=2}^{k}\exp\{\alpha_j + g_j(x)\}} \text{ pour } j=2, \ldots, k.$$

En prenant $\alpha_1 = \beta_{1j} = 0$, on peut réunir ces deux expressions en une seule formule :

$$P(Y=j\mid X) = \frac{\exp\{\alpha_j + g_j(x)\}}{1+\sum_{j=1}^{k}\exp\{\alpha_j + g_j(x)\}}, \; j=1, \ldots, k.$$

VIII.2. Variable continue sous-jacente et modèle *cumulative-odds*

On se place dans une situation où l'ordre de la variable Y repose sur l'existence d'une variable continue T sous-jacente (cela peut être, par exemple, une échelle de douleur, ou les stades d'une maladie…). Les k classes de Y sont définies par des seuils θ_j de T : $Y = j \Leftrightarrow \theta_{j-1} < T < \theta_j$ avec $j = 1, \ldots, j$ et $-\infty = \theta_0 \leq \theta_1 \leq \ldots \leq \theta_k = +\infty$

Si, comme on le fait souvent, on modélise la relation entre T et des variables X_i par une fonction linéaire, on obtient : $E(T|X) = \alpha + \sum \beta_i X_i$, ce qui peut aussi s'écrire $T = \alpha + \sum \beta_i X_i + \varepsilon$, où ε est le résidu.

On en déduit : $P(Y \leq j|X) = P(T \leq \theta_j|X) = P(\varepsilon \leq \theta_j - \alpha - \sum \beta_i x_i) = F(\theta_j - \alpha - \sum \beta_i x_i)$; où F est la fonction de répartition du résidu ε. En prenant pour F la fonction logistique $F(x) = \frac{1}{1+e^{-x}}$, on retrouve le modèle *cumulative-odds* : $P(Y \leq j|X) = \frac{1}{1+\exp-\{\theta_j - \alpha - \sum \beta_i x_i\}}$, soit $\text{logit}\,(Y \leq j|X) = \alpha_j - \sum \beta_i X_i$, avec $\alpha_j = \theta_j - \alpha$.

Comme on le voit, on aboutit à une écriture du modèle avec un signe « – » devant les β_i. C'est moins habituel que l'écriture avec un signe « + », mais cela a l'avantage de rendre plus intuitive l'interprétation des β_i du modèle *cumulative-odds*. En effet, si β_i est positif, cela signifie qu'un sujet exposé ($X_i = 1$) a une probabilité moins grande qu'un sujet non exposé ($X_i = 0$) que Y soit inférieur à j. C'est-à-dire qu'il y a une relation positive entre X_i et Y avec son codage originel ordonné de 1 à k (et aussi une relation positive entre X_i et T). Tout cela est dû au fait qu'on étudie la probabilité cumulée $P(Y \leq j|X)$ et non $P(Y \geq j|X)$. C'est probablement cette plus grande facilité d'interprétation qui a poussé certains logiciels (mais pas tous, attention !) à paramétrer le modèle *cumulative-ratio* avec un signe « – » devant les β_i. C'est le cas de Stata et de certains modules de R (Parry S, 2020). C'est aussi le choix que j'ai fait.

Le développement précédent repose sur l'hypothèse que la relation entre les X_i et T est linéaire et aboutit au modèle *cumulative-odds* avec *odds* proportionnels. Si la relation entre les X_i et T n'est linéaire que par morceaux, avec des pentes β_{ji} différentes selon les intervalles de T définis par les θ_j, on aboutirait au modèle *cumulative-odds* général (sans hypothèse sur les *odds* proportionnels).

VIII.3. Modèles logistiques multinomial et ordinaux et enquêtes de type cas-témoins

VIII.3.a. Le modèle logistique multinomial est utilisable dans des enquêtes de type cas-témoins

La démonstration est semblable à celle que j'ai faite pour le modèle logistique binomial classique (chapitre 1, § IV.1). On se place dans la situation où l'échantillon est issu par tirage au sort séparément de la population des non-malades (classe 1 de Y) et des différentes populations de malades (classes 2 à k de Y). On note f_j la fraction de sondage pour chaque catégorie $Y = j$.

Dans la « vraie vie », il y a rarement tirage au sort, mais on doit se préoccuper de ce que les cas et les témoins soient « représentatifs » de leur population pour qu'il n'y ait pas de biais de sélection (Bouyer J et al., 1993), de sorte que la représentation de l'échantillonnage par un tirage au sort avec des fractions de sondage soit fidèle à la réalité, même si les valeurs des fractions de sondage sont inconnues.

On sait que, dans un échantillonnage de type cas-témoins, la fréquence de chaque classe de Y, c'est-à-dire la probabilité que $Y = j$, ne peut pas être estimée à partir des observations, puisque les effectifs de chaque classe ont été fixés arbitrairement par le protocole de l'enquête. En utilisant la fraction de sondage, on peut cependant écrire : $P(Y = j|X) = f_j \dfrac{\exp(\alpha_j + g_j(x))}{\sum_{j=1}^{k} \exp\{\alpha_j + g_j(x)\}}$ avec $g_j(x) = \sum_{i=1}^{p} \beta_{ji} X_i$ et $\alpha_1 = \beta_{1i} = 0$.

On en déduit : $\dfrac{P(Y = j|X)}{P(Y = 1|X)} = \dfrac{f_j \exp\{\alpha_j + g_j(x)\}}{f_1 \exp\{\alpha_1 + g_1(x)\}} = \dfrac{f_j}{f_1} \exp\{\alpha_j + g_j(x)\}$

et donc : $\ln\left(\dfrac{P(Y = j|X)}{P(Y = 1|X)}\right) = \ln\left(\dfrac{f_j}{f_1}\right) + \{\alpha_j + g_j(x)\} = \alpha_j' + \sum_{j=1}^{p} \beta_{ji} x_i.$

Cela montre que l'échantillonnage de type cas-témoins ne modifie pas les coefficients β_i (et donc les odds ratios OR_i) et leur estimation. Cela montre aussi que la constante α_j n'est pas interprétable puisqu'elle dépend de fractions de sondage qui sont inconnues.

VIII.3.b. Les modèles cumulative-odds *et* continuation-ratio *ne sont pas utilisables dans des enquêtes de type cas-témoins*

La démonstration précédente n'aboutit pas avec un modèle *cumulative-odds*. En effet, avec un échantillon représentatif, le modèle s'écrit : $\text{logit}\left(P(Y \le j|X)\right) = \ln\left(\dfrac{P(Y \le j \mid X)}{P(Y > j \mid X)}\right) = \alpha_j - \sum_{i=1}^{p} \beta_{ji} X_i$, où $j = 1, \ldots, k-1$ désigne une classe de Y.

Comme $P(Y \le j \mid X) = \sum_{t=1}^{j} P(Y = t \mid X)$, l'estimation de la probabilité que $Y \le j$ sur un échantillon de type cas-témoins avec les fractions de sondage f_t (voir VIII.2.a) s'écrit : $P(Y \le j \mid X) = \sum_{t=1}^{j} f_t \, P(Y = t \mid X)$.

On en déduit : $\text{logit}\, P(Y \le j \mid X) = \ln\left(\dfrac{\sum_{t=1}^{j} f_t \, P(Y = t \mid X)}{\sum_{t=j+1}^{k} f_t \, P(Y = t \mid X)}\right)$, qui ne se simplifie pas. Les coefficients β_j estimés sur un échantillon de type cas-témoins sont donc différents des

coefficients estimés sur un échantillon représentatif ou non sélectionné et ne sont pas interprétables.

Le même phénomène se produit pour le modèle *continuation-ratio*, qui ne peut donc pas être utilisé non plus dans des enquêtes de type cas-témoins (Greenland S, 1994).

Chapitre 7

Adéquation du modèle

• • •

I. Introduction

L'adéquation d'un modèle est sa capacité à bien représenter les observations faites sur l'échantillon d'enquête. J'ai déjà abordé cette question de la qualité d'un modèle (ou de l'adéquation, les termes sont parfois difficiles à distinguer) dans le chapitre 5, notamment dans le § I. Il faut garder en tête, surtout si on est dans une démarche étiologique (recherche des facteurs de risque), qu'un modèle peut être « trop » adéquat parce qu'il « colle » trop aux fluctuations d'échantillonnage. Le risque existe surtout pour les petits échantillons.

Comme cela a été détaillé dans le chapitre 5, la première étape pour assurer la bonne adéquation d'un modèle est de choisir les variables à inclure ainsi que leur modélisation (en particulier pour les variables quantitatives). Je supposerai dans ce chapitre que cette étape a été réalisée soigneusement. La deuxième étape consiste à estimer les coefficients du modèle. Les méthodes d'estimation sont construites pour représenter au mieux les observations. C'est le cas de la méthode des moindres carrés (Armitage P et al., 1987) utilisée pour la régression linéaire et, de façon moins directe, la méthode du maximum de vraisemblance (voir chapitre 2, § II).
La question de l'adéquation du modèle présentée dans ce chapitre se pose à l'issue de ces deux étapes, c'est-à-dire à un moment où beaucoup a déjà été fait pour que les observations soient bien représentées par le modèle. C'est peut-être pour cette raison que les mesures ou les tests d'adéquation sont peu présents dans les articles d'épidémiologie donnant des résultats d'analyses étiologiques.
C'est aussi pour cette raison que l'appréciation de l'adéquation sur le même échantillon que celui qui a servi à estimer les coefficients du modèle est trop optimiste (on parle d'adéquation apparente ou de validité interne). Il n'y a pas d'études détaillées portant sur l'ampleur de cet optimisme, mais il semble que, quelle que soit la mesure, l'adéquation est surestimée d'environ 10 %, voire plus pour les petits échantillons (Bleeker SE et al., 2003, Steyerberg EW et al., 2003). Pour corriger la mesure de l'adéquation, on peut la calculer sur un autre échantillon issu d'une population semblable (validité externe), généralement difficile à trouver. On peut aussi, c'est plus fréquent, modifier la validation interne par des méthodes qui reviennent à peu près à singer le recours à un échantillon externe (Steyerberg EW et al., 2001) par des tirages au sort dans l'échantillon d'origine :

- *Split-sampling* : on sépare l'échantillon aléatoirement en deux parties : une pour estimer les coefficients du modèle (avec le plus souvent 33 % ou 50 % de l'échantillon total), l'autre pour évaluer sa performance. L'inconvénient est qu'on se retrouve avec deux plus petits échantillons pour construire et valider le modèle, ce qui peut faire perdre de la puissance pour ces deux aspects.
- *Cross-validation* : le *split* est répété plusieurs fois avec des tirages au sort différents et la performance du modèle est la moyenne sur les répétitions.
- *Bootstrap* : l'échantillon de validation est construit par tirage au sort avec remise et peut donc contenir plusieurs fois le même sujet. Il peut être de la même taille que l'échantillon initial pour lequel les paramètres sont estimés.

Steyerberg et al. (Steyerberg EW et al., 2001) ont comparé ces méthodes à l'aide de simulations. Ils ont montré qu'elles sont équivalentes et pas meilleures que le calcul de l'adéquation apparente pour les grands échantillons. Pour des échantillons plus petits, le *split-sampling* est inefficace, la meilleure méthode étant le *bootstrap*. Ici, « grands échantillons » veut dire que EPV $\geq$ 40 et « petits échantillons » que EPV $\leq$ 10, où EPV est le nombre d'événements par variable (voir chapitre 5, § II).

L'étude de l'adéquation d'un modèle a été beaucoup développée dans le cadre de la régression linéaire, où elle bénéficie des conditions de distribution de Y normale et de variance constante. Une des mesures très utilisées est le coefficient R^2, qui s'interprète comme la proportion de variance de Y expliquée par le modèle. Pour le

modèle logistique, il n'y a pas de conditions de distribution « contraignantes » pour Y et pour les variables X. Cependant, R^2 n'est pas adapté au modèle logistique. Sa valeur dépend de l'étendue et de la distribution des variables X_i et tend à être faible, même pour un modèle parfaitement adéquat (Cox DR et al., 1992, Mittlböck M et al., 1996). Le R^2 donné par la plupart des logiciels ne s'interprète pas comme une proportion de variance expliquée ; en particulier, sa valeur ne peut pas atteindre 1 (Nagelkerke NJD, 1991). Les R^2 corrigés ou pseudo-R^2 qui ont été proposés, R^2 de Cox et Snell (Cox DR et al., 1989) et de Nagelkerke (Nagelkerke NJD, 1991), ou de McFadden (Allison PD, 2013), ne résolvent pas le problème (Hosmer DW et al., 2013, Allison PD, 2014).

Pour le modèle logistique, il faut définir une distance adaptée entre observations et prédictions (voir § II), qu'on peut utiliser avec une approche globale de l'adéquation (§ III et IV) ou une approche plus « individuelle » avec les diagnostics de régression (§ V). Je me limiterai au modèle logistique binomial « classique ». Des mesures d'adéquation ont été proposées pour d'autres cas : régression logistique multinomiale (Fagerland MW et al., 2008), ordinale (Fagerland MW et al., 2017), mixte (Evans S et al., 2005), ou sur des données appariées (Chen L-C et al., 2013).

Pour conclure cette introduction, je voudrais dire quelques mots sur une question peu ou pas abordée dans les articles et les livres et pour laquelle je n'ai pas non plus de bonne réponse : que peut-on faire si le test d'adéquation est significatif ?
Un modèle peut être incorrect ou inadéquat en raison du non-respect d'hypothèses (par exemple la fonction de lien logit n'est pas la bonne ou les résidus n'ont pas une distribution de Bernouilli), d'oublis de variables importantes ou de surajustement (*overfitting*) (Harrel Jr. FE et al., 1996, Hosmer DW et al., 1997). Dans le modèle logistique, les hypothèses qui pourraient ne pas être respectées portent principalement sur la linéarité de la relation avec X (ou plus généralement de la modélisation de X). Ce point, comme l'oubli de variables ou le surajustement a, en principe, été résolu au moment du choix des variables (voir chapitre 5), c'est-à-dire avant qu'on ne se soit penché sur l'étude de l'adéquation. Si le modèle apparaît « quand même » non adéquat, ce qui est finalement assez peu fréquent, cela peut être que le lien logistique lui-même ne convient pas ; il y a alors assez peu d'alternatives en pratique, car les autres modèles sont assez peu différents dans l'intervalle de 0,2 à 0,8 pour P(Y = 1) (Hosmer DW et al., 2013).

II. Mesure de l'écart entre les observations et les prédictions du modèle logistique

On note les observations $y_1, \ldots, y_n$. Dans le cas du modèle logistique, Y est la variable « maladie » et y_k prend la valeur 1 ou 0 selon que le sujet k est malade ou pas.
Les valeurs prédites par le modèle sont notées $\hat{y}_1, \ldots, \hat{y}_n$. Dans le cas du modèle logistique, $\hat{y}_k$ est la probabilité P_k d'être malade donnée par $\text{logit}(P_k) = \alpha + \sum_{i=1}^{p} \beta_i x_{ik}$ ou encore $P_k = \dfrac{\exp\left(\alpha + \sum_{i=1}^{p} \beta_i x_{ik}\right)}{1 + \exp\left(\alpha + \sum_{i=1}^{p} \beta_i x_{ik}\right)}$, où x_{ik} est la valeur prise par le sujet k pour la variable X_i.

Il s'agit donc d'étudier les écarts entre les y_k et les $\hat{y}_k$. Cela nécessite, comme on va le voir, de définir une distance (c'est-à-dire une mesure de l'écart) entre y_k et les $\hat{y}_k$. On va, de plus, considérer deux types de problèmes :

- ✓ l'étude de l'adéquation globale du modèle, grâce à un indice qui résume l'ensemble des écarts – c'est dans cette catégorie que se rangent les tests d'adéquation à proprement parler, ainsi que les courbes ROC ;
- ✓ l'étude de la contribution de chaque paire $(y_k, \hat{y}_k)$ à un indice résumé : c'est ce qu'on appelle les « diagnostics de régression » (*regression diagnostics* en anglais).

II.1. Définition d'un profil

Notons $X = (x_1, \ldots, x_p)$ le vecteur de l'ensemble des variables explicatives X.
Soit J le nombre de valeurs différentes de X observées, c'est-à-dire le nombre de combinaisons des valeurs des x_i observées. Une valeur particulière de X s'appelle un profil (ou *covariate pattern* en anglais). On peut noter que tous les sujets ayant le même profil ont la même valeur prédite de Y, ce qui les rend, de ce point de vue, indistinguables, et c'est ce qui fait l'intérêt de parler de profil. Pour un profil donné X_j, on note P_j (ou $\hat{Y}_j$) cette valeur commune (j étant l'indice repérant les profils[1]).

Dans le cas du modèle logistique, on fait fréquemment appel à des variables explicatives X_i qualitatives. Ces variables ont le plus souvent un petit nombre de catégories, de sorte que, même lorsqu'on considère plusieurs variables X_i dans le même modèle, le nombre total de profils (c'est-à-dire de catégories combinées des variables) reste limité. Si plusieurs sujets ont le même profil, alors le nombre total de profils différents J est strictement inférieur au nombre total de sujets n. Si une (et a fortiori plusieurs) variable(s) X_i est (sont) quantitative(s), elle(s) peu(ven)t avoir autant de catégories que de sujets. Dans ce cas, chaque profil n'a qu'un seul sujet, et la notion de profil se confond avec celle de sujet (et perd de son intérêt).
On note m_j le nombre de sujets ayant le profil j (c'est-à-dire tels que $X = X_j$). Parmi ces m_j sujets, le nombre de malades observés est noté n_{1j} (n_{1j} est aussi la somme des y_i pour les sujets ayant le profil j). Le nombre « théorique » de malades (c'est-à-dire le nombre attendu si l'hypothèse que l'on teste, celle selon laquelle le modèle est adéquat, est vraie) est $m_j P_j$.
De même, pour le profil j, le nombre de non-malades observé est n_{0j} (qui vaut $m_j - n_{1j}$) et le nombre de non-malades attendu est $m_j (1 - P_j)$.

Puisque les sujets ayant un même profil ont les mêmes valeurs pour les variables X, et donc pour la valeur prédite P_j, le problème de l'étude des écarts entre les valeurs observées y_k et les valeurs données par le modèle $\hat{y}_k$ se ramène à l'étude des écarts entre les pourcentages de malades observés et théoriques dans les profils j. Ou, ce qui revient au même, à l'étude des écarts entre les nombres de malades observés n_{1j}

1. Dans la suite, les profils seront repérés par l'indice j et les sujets par l'indice i. Ce n'est pas un système de notation très orthodoxe, aussi j'espère qu'il n'introduira pas de confusion.

et théoriques m_j P_j. On s'intéresse donc à un tableau très similaire à un tableau de χ^2 (Tableau 7.1).

Profil	1	2	…	J
Valeurs de X	X_1	X_2	…	X_J
Y = 1 (malade)	$n_{11}\left(m_1P_1\right)$	$\hat{y}_1,\ldots,\hat{y}_n$		$n_{1J}\left(m_JP_J\right)$
Y = 0 (non malade)	$n_{01}\left(m_1\left(1-P_1\right)\right)$	$n_{02}\left(m_2\left(1-P_2\right)\right)$		$n_{0J}\left(m_J\left(1-P_J\right)\right)$
Total	m_1	m_2		m_J

Tableau 7.1 : Tableau de type χ^2 donnant les écarts par profil entre les effectifs observés et les effectifs attendus (ou prédits) entre parenthèses

Il s'agit d'un tableau de χ^2 habituel, mais l'hypothèse testée n'est pas que le pourcentage de malades est le même dans tous les profils (colonnes), mais que le pourcentage de malades de chaque profil est celui donné par le modèle (P_j). C'est ce qui explique les valeurs des effectifs théoriques (ou attendus) figurant entre parenthèses dans chacune des cases du tableau.

Dans la terminologie des modèles de régression (linéaire ou logistique), l'écart entre valeur observée et valeur prédite (ou attendue) par le modèle est appelé « résidu ». Dans le cas du modèle logistique, on utilise habituellement deux types de résidus : le résidu de Pearson et le résidu de la déviance (Hosmer DW et al., 1997, Mair P et al., 2008, Hilbe JM, 2009).

II.2. Résidus de Pearson et de la déviance

Le résidu de Pearson pour le profil j est égal à : $r_j = \dfrac{\left(n_{1j} - m_jP_j\right)}{\sqrt{m_jP_j\left(1-P_j\right)}}$. Il est toujours défini, puisque P_j ne peut être égal ni à 0 ni à 1. On voit, comme pour le test de χ^2 classique, que l'écart entre l'effectif observé et l'effectif attendu est pondéré par son écart-type.

Le résidu de la déviance pour le profil j est défini par : $d_j = \pm\left\{2\left[n_{1j}\ln\left(\dfrac{n_{1j}}{m_jP_j}\right) + n_{0j}\ln\left(\dfrac{n_{0j}}{m_j\left(1-P_j\right)}\right)\right]\right\}^{1/2}$ où ± est le même signe que celui de $\left(n_{1j} - m_jP_j\right)$.

Pour les profils tels que $n_{1j} = 0$ (uniquement des non-malades), on prend $d_j = -\sqrt{2m_j\left|\ln\left(1-P_j\right)\right|}$.

Pour les profils tels que $n_{0j} = 0$ (uniquement des malades), on prend $d_j = \sqrt{2m_j\left|\ln\left(P_j\right)\right|}$.

Comme le précédent, ce résidu est égal à 0 si les effectifs observés et attendus sont égaux.

III. Tests d'adéquation

Je me limite ici aux tests basés sur le tableau de χ^2 vu plus haut (Tableau 7.1), qui sont les plus courants. Il s'agit de tests globaux de l'adéquation du modèle, ils ne permettent pas, ou mal, de préciser les raisons d'un éventuel défaut d'adéquation.

III.1. Tests de χ^2 de Pearson et de la déviance

Les deux statistiques globales d'adéquation présentées ci-dessus (X^2 et D^2) donnent respectivement le χ^2 de Pearson et le χ^2 de la déviance.

Le χ^2 de Pearson est égal à : $X^2 = \sum_{j=1}^{J} r_j^2$. C'est le χ^2 « classique » qu'on trouve dans tous les logiciels et qui s'écrit habituellement $X^2 = \sum_{i,j} \frac{\left(n_{ij} - \hat{m}_{ij}\right)^2}{\hat{m}_{ij}}$, où n_{ij} et $\hat{m}_{ij}$ sont les effectifs observés et théoriques.

Pour retrouver qu'il s'agit bien de la même chose, il suffit de remarquer qu'on a $n_{1j} - \left(m_j P_j\right) = -\left(n_{oj} - m_j\left(1 - P_j\right)\right)$ puisque $n_{1j} + n_{0j} = m_j$. On obtient donc, en remplaçant les $\hat{m}_{ij}$ par $m_i P_i$:

$$X^2 = \sum_{i,j} \frac{\left(n_{ij} - \hat{m}_{ij}\right)^2}{\hat{m}_{ij}} = \sum_{j=1}^{J} \sum_{i=0}^{1} \frac{\left(n_{ij} - \hat{m}_{ij}\right)^2}{\hat{m}_{ij}} = \sum_{j=1}^{J} \left(\frac{\left(n_{1J} - m_j P_j\right)^2}{m_j P_j} + \frac{\left(n_{0J} - m_j(1 - P_j)\right)^2}{m_j(1 - P_j)} \right)$$

$$= \sum_{j=1}^{J} \left(n_{1j} - m_j P_j\right)^2 \left(\frac{1}{m_j P_j} + \frac{1}{m_j(1 - P_j)} \right) = \sum_{j=1}^{J} r_j^2$$

Le χ^2 de la déviance est égal à : $D^2 = \sum_{j=1}^{J} d_j^2$. Il est aussi appelé χ^2 du rapport des vraisemblances (*likelihood-ratio chi-squared*) (Agresti A, 1990). Son écriture habituelle (équivalente) est : $\chi^2 = 2 \sum_{i,j} n_{ij} \ln\left(\frac{n_{ij}}{\hat{m}_{ij}}\right)$, avec les mêmes règles que pour le résidu de la déviance si n_{1j} ou n_{0j} sont nuls.

Ces deux statistiques devraient a priori suivre une loi de χ^2 à (J – 1) degrés de liberté lorsque l'hypothèse nulle (adéquation du modèle) est vraie. Cet a priori doit cependant être doublement rectifié :

- ✓ d'une part, le nombre de degrés de liberté est J – (p + 1) et non pas J – 1, car les données sont déjà utilisées pour estimer (p + 1) paramètres (α et les β_i), ce qui « fait perdre » (p + 1) degrés de liberté ;
- ✓ d'autre part, X^2 et D^2 ne suivent des lois de χ^2 que si les effectifs théoriques sont assez grands.

Les statistiques d'adéquation définies sur la base des χ^2 précédents ne peuvent donc être étudiées en pratique que quand les m_j sont assez grands (on dit qu'elles sont

asymptotiques)[2]. Il n'y a pas, à ma connaissance, de règle pratique sur les valeurs de m_i, mais on peut prendre celles du test de χ^2 habituel : il faut que les effectifs attendus soient supérieurs à cinq, ou du moins qu'il y en ait moins d'un sur cinq qui soit inférieur à 5 (Cochran WG, 1954, Armitage P et al., 1987). Si ce n'est pas le cas, il faut regrouper des profils entre eux, jusqu'à ce que ces conditions soient satisfaites, comme on le fait habituellement avec un tableau de χ^2. Le principe est de faire des regroupements selon les probabilités estimées de survenue de la maladie, P_j. Pour cela, on range les profils dans l'ordre des P_j croissants, et on regroupe des profils adjacents.

III.2. Test de Hosmer et Lemeshow

Pour tenir compte de la nécessité fréquente de grouper les profils, Hosmer et Lemeshow ont proposé un test qui consiste à regrouper les profils en 10 groupes (Hosmer DW et al., 1980, Hosmer DW et al., 2013). Ils ont envisagé deux façons de construire les groupes, toujours en ordonnant les profils selon les P_j croissants (Hosmer DW et al., 1997) :

- ✓ fixer des valeurs seuils de P_j pour définir les groupes, par exemple 10 %, 20 %, …, 90 %. La statistique de test obtenue est notée $\hat{H}$;
- ✓ faire des groupes de tailles égales en prenant pour valeurs seuils des groupes les déciles de la distribution des P_j : le premier contient les profils correspondant aux 10 % de sujets ayant les valeurs de P_j les plus petites, et ainsi de suite. La statistique obtenue est notée $\hat{C}$. On peut remarquer que, puisqu'on regroupe des profils pouvant contenir plusieurs sujets, il peut arriver qu'un ou plusieurs groupes ne contien(nen)t pas exactement 10 % des sujets. En pratique, on procède aux regroupements qui correspondent le mieux à cette situation.

Korn et al. (Korn L et al., 1986) ont montré que la statistique $\hat{C}$ était préférable, en particulier quand les variables X_i sont qualitatives. C'est celle qui porte le nom de test de Hosmer et Lemeshow, et qui figure dans tous les logiciels.

---

2. En fait, il faudrait considérer deux « façons d'être asymptotique » :
Si le nombre J de profils augmente en même temps que n, alors chaque valeur m_j tend à être ou rester petite. Les résultats obtenus sont dits n-asymptotiques. C'est ce qui se passe quand il y a une variable continue parmi les variables explicatives X_i et que le nombre de profils augmente avec le nombre de sujets.
Si J reste fixe, alors chaque valeur de m_j devient grande. Les résultats obtenus sont dits m-asymptotiques C'est ce qui se passe quand il n'y a que des variables catégorielles parmi les variables explicatives X_i.
En pratique, les résultats ne sont satisfaisants que si les effectifs attendus de malades et de non-malades dans chaque profil sont assez grands (cas m-asymptotique), c'est-à-dire s'il n'y a pas trop de profils différents et s'il y a suffisamment de sujets.

III.3. Exemple

Cet exemple s'appuie sur l'enquête sur les facteurs de risque de la GEU (grossesse extra-utérine (voir chapitre 1). La variable maladie est « ct », codée 1 pour les GEU et 0 pour les témoins (accouchements). Cinq variables explicatives sont considérées ici :

- ✓ ctub : antécédent de chirurgie tubaire (0 : non ; 1 : oui) ;
- ✓ agea : âge en années. Cette variable a une trentaine de valeurs différentes de 15 à 46 ans ;
- ✓ tabfc : consommation de tabac en classes (0 : non-fumeuse ; 1 : 1 à 9 cig/j ; 2 : 10 à 19 cig/j ; 3 : ≥ 20 cig/j) ;
- ✓ afcs : antécédent de fausse couche spontanée (0 : non ; 1 : oui) ;
- ✓ ainf : antécédent d'infécondité (0 : moins d'un an de recherche infructueuse de grossesse ; 1 : plus d'un an).

La variable tabfc se décompose en trois variables indicatrices et la variable agea est modélisée linéairement après avoir vérifié l'absence d'écart à la linéarité par la procédure mfp (chapitre 4, § VII.4). Les coefficients du modèle logistique correspondant sont donnés dans le Tableau 7.2.

Les statistiques d'adéquation sont données par la commande « estat gof ». Un calcul supplémentaire programmé spécifiquement permet d'obtenir le χ^2 de la déviance : $D^2 = 380{,}67$, avec 350 degrés de liberté et $p = 0{,}12$.

```
. logit ct ctub agea i.tabfc afcs ainf

(...)

Logistic regression                                     Number of obs =  1,682
                                                        IR chi2(7)    = 344.62
                                                        Prob > chi2   = 0.0000
Log likelihood = -887.84967                             Pseudo R2     = 0.1625

------------------------------------------------------------------------------
          ct | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
        ctub |   1.831296   .2047665     8.94   0.000     1.429961    2.232631
        agea |    .048838   .0124025     3.94   0.000     .0245294    .0731465
             |
      tabfc |
          1  |   .5313484   .1702519     3.12   0.002     .1976609    .8650359
          2  |   1.322487   .1671586     7.91   0.000      .994862    1.650111
          3  |   1.516687   .1862517     8.14   0.000      1.15164    1.881734
             |
        afcs |    .337114   .1394595     2.42   0.016     .0637785    .6104496
        ainf |   .8166237    .149279     5.47   0.000     .5240422    1.109205
       _cons |   -3.06147   .3723993    -8.22   0.000    -3.791359   -2.331581
------------------------------------------------------------------------------
. estat gof

Goodness-of-fit test after logistic model
Variable: ct

      Number of observations =  1,682
Number of covariate patterns =    358
          Pearson chi2(350) = 331.07
                Prob > chi2 = 0.7592

khi2 de la déviance (350) = 380.67   p = 0.12
```

Tableau 7.2 : Coefficients du modèle logistique et statistiques d'adéquation

On voit qu'il y a 358 profils pour un total de 1682 sujets. On peut donc penser que nombre d'entre eux ont un effectif réduit, ce que confirme le Tableau 7.3 qui indique, par exemple, que 151 profils ne comprennent qu'un seul sujet.
En raison de l'existence de nombreux profils avec peu de sujets, les résultats des tests d'adéquation portant sur l'ensemble des profils ne sont pas interprétables, et il faut utiliser le test de Hosmer–Lemeshow.

Nombre de sujets par profil	Nombre de profils	Nombre de sujets par profil	Nombre de profils
1	151	19	1
2	64	20	1
3	34	21	1
4	24	26	1
5	16	30	1
6	7	32	1
7	10	33	1
8	14	36	1
9	2	51	1
10	1	53	1
11	6	56	1
12	7	57	1
13	2	58	1
14	1	59	1
16	1	62	1
18	3	Total	358

Tableau 7.3 : Distribution du nombre de sujets par profils pour les 358 profils

Dans le Tableau 7.4 :

- ✓ La colonne Prob donne la valeur maximum de la probabilité de maladie P_j qui détermine le seuil de chaque groupe.
- ✓ Les colonnes Obs 1 et Exp 1 donnent les nombres observé et théorique de sujets malades dans chaque groupe. Les deux colonnes suivantes correspondent aux non-malades.
- ✓ La dernière colonne donne le nombre total de sujets par groupe. On voit qu'il est toujours assez proche du dixième de la taille de l'échantillon (168), comme il se doit pour un décile. Pas exactement cependant, car les groupes sont des réunions de profils dont certains comprennent plus d'un sujet. Pour le groupe 1, par exemple, les profils sont regroupés jusqu'à ce que l'effectif total dépasse 168. On obtient

203 sujets. Le dernier profil regroupé contient 53 sujets. Si on l'enlève, il n'y en aurait que 150 (nombre inférieur à 168). D'autres logiciels ont un autre algorithme de construction des groupes, et cela peut modifier le résultat du test (Hosmer DW et al., 1997).

```
. estat gof,group(10) table
note: obs collapsed on 10 quantiles of estimated probabilities.

Goodness-of-fit test after logistic model
Variable: ct

  Table collapsed on quantiles of estimated probabilities
  +-------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.1370 |    34 |  25.7 |   169 | 177.3 |   203 |
  |     2 | 0.1552 |    20 |  27.1 |   162 | 154.9 |   182 |
  |     3 | 0.1754 |    24 |  27.5 |   139 | 135.5 |   163 |
  |     4 | 0.1977 |    29 |  26.3 |   110 | 112.7 |   139 |
  |     5 | 0.2336 |    33 |  33.3 |   121 | 120.7 |   154 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.3056 |    46 |  45.2 |   123 | 123.8 |   169 |
  |     7 | 0.3911 |    63 |  60.1 |   108 | 110.9 |   171 |
  |     8 | 0.4679 |    70 |  71.7 |    98 |  96.3 |   168 |
  |     9 | 0.6830 |    89 |  94.0 |    78 |  73.0 |   167 |
  |    10 | 0.9608 |   138 | 135.1 |    28 |  30.9 |   166 |
  +-------------------------------------------------------+

Number of observations =  1,682
      Number of groups =     10
Hosmer-Lemeshow chi2(8) =   7.41
           Prob > chi2 = 0.4931
```

Tableau 7.4 : Test de Hosmer et Lemeshow

Globalement, il n'y a pas de défaut d'adéquation significatif entre le modèle et les observations, puisque le χ^2 de Hosmer–Lemeshow est non significatif. On peut cependant noter que l'écart entre effectifs observé et théorique a tendance à être plus grand (au moins en valeur relative) pour les valeurs les plus petites de P_j, ce qui signifierait que le modèle est moins adéquat pour les sujets dont la probabilité de GEU est la plus faible.

III.4. Comparaison avec d'autres tests

Le test de Hosmer et Lemeshow est le plus utilisé. Il a été comparé à d'autres et s'est avéré le meilleur quand on tenait compte des performances et de la facilité de calcul (Lemeshow S et al., 1982).

Dans un article plus récent (Hosmer DW et al., 1997), une comparaison complète est faite ainsi que des indications sur les risques d'erreur. Tous les tests comparés ont une taille (c'est-à-dire un risque d'erreur de première espèce) de 5 %. La plupart des tests, dont celui de Hosmer et Lemeshow, ont une bonne puissance pour détecter un défaut d'adéquation associé à une forme quadratique pour une variable X_i alors qu'elle est incluse dans le modèle de façon linéaire (puissance égale à 50 % si $n = 100$ et supérieure à 90 % si $n = 500$), mais une puissance médiocre pour détecter que la fonction de lien logit n'est pas la bonne. Tous les tests ont une très mauvaise puissance pour détecter l'oubli d'une interaction.

IV. Courbe ROC

IV.1. Régression logistique et classement

La régression logistique peut être utilisée pour classer les sujets en malades ou non-malades, dans un but diagnostique ou prédictif. Il suffit pour cela de décider qu'un sujet est classé malade si la probabilité de maladie que lui attribue le modèle logistique dépasse un seuil donné.
Comme pour tout système de classement, on peut alors calculer une sensibilité et une spécificité qui mesurent sa capacité à classer correctement un malade ou un non-malade. Plus la sensibilité et la spécificité sont élevées, meilleure est la capacité du modèle logistique à discriminer entre malades et non-malades. En ce sens, la sensibilité et la spécificité sont des mesures globales d'adéquation du modèle.

Bien entendu, la sensibilité et la spécificité dépendent du seuil choisi. Si on prend 0,5 (valeur par défaut dans Stata), on obtient les résultats du Tableau 7.5. Si on prend un seuil de 0,3 (Tableau 7.6), la sensibilité est améliorée au détriment de la spécificité.

```
. estat classification

Logistic model for ct

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       213            89  |        302
     -     |       333          1047  |       1380
-----------+--------------------------+-----------
   Total   |       546          1136  |       1682

Classified + if predicted Pr(D) >= .5
True D defined as ct != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   39.01%
Specificity                     Pr( -|~D)   92.17%
Positive predictive value       Pr( D| +)   70.53%
Negative predictive value       Pr(~D| -)   75.87%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    7.83%
False - rate for true D         Pr( -| D)   60.99%
False + rate for classified +   Pr(~D| +)   29.47%
False - rate for classified -   Pr( D| -)   24.13%
--------------------------------------------------
Correctly classified                        74.91%
--------------------------------------------------
```

Tableau 7.5 : Classement des sujets par le modèle logistique du Tableau 7.2, avec un seuil à 0,5 (par défaut)

On peut résumer l'ensemble des résultats pour les différentes valeurs possibles du seuil sur un graphique où la sensibilité et la spécificité sont représentées en fonction du seuil choisi, ce qui permet de choisir le compromis jugé le meilleur (Figure 7.1).

> Remarque
> Les valeurs de sensibilité et de spécificité gardent un sens dans une enquête cas-témoins. C'est aussi le cas de l'utilisation de ces caractéristiques pour évaluer l'adéquation du modèle ainsi que celui de la Figure 7.1 (et plus loin de la Figure 7.2). En revanche, une partie des résultats des Tableaux 7.5 et 7.6 n'ont pas de sens, par exemple les valeurs prédictives positive et négative.

```
. estat classification, cutoff(.3)

Logistic model for ct

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       363           322  |        685
     -     |       183           814  |        997
-----------+--------------------------+-----------
   Total   |       546          1136  |       1682

Classified + if predicted Pr(D) >= .3
True D defined as ct != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   66.48%
Specificity                     Pr( -|~D)   71.65%
Positive predictive value       Pr( D| +)   52.99%
Negative predictive value       Pr(~D| -)   81.64%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   28.35%
False - rate for true D         Pr( -| D)   33.52%
False + rate for classified +   Pr(~D| +)   47.01%
False - rate for classified -   Pr( D| -)   18.36%
--------------------------------------------------
Correctly classified                        69.98%
--------------------------------------------------
```

Tableau 7.6 : Classement des sujets par le modèle logistique du Tableau 7.2, avec un seuil à 0,3

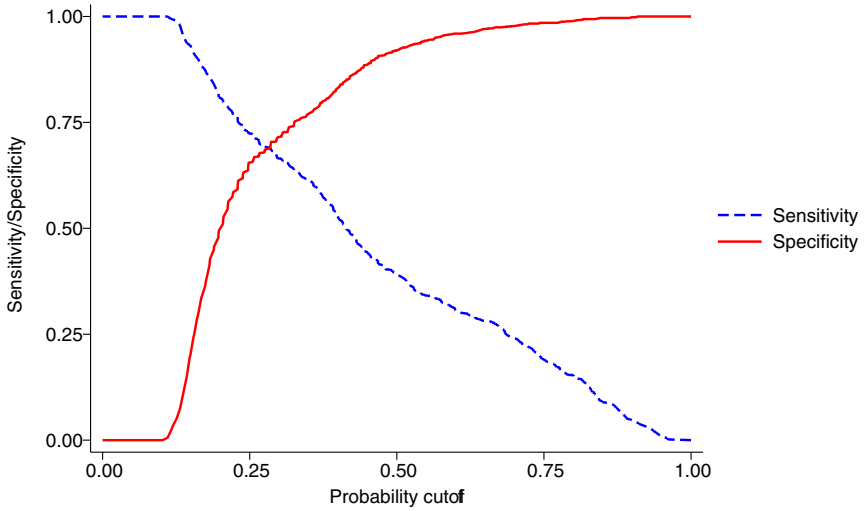


Figure 7.1 : Sensibilité et spécificité du classement par le modèle logistique selon le seuil choisi.

IV.2. Aire sous la courbe ROC

Pour synthétiser avec une seule valeur la capacité globale du modèle à discriminer entre malades et non-malades, on construit la courbe ROC[3] qui donne le graphe de la sensibilité en fonction de (1-spécificité), et on considère la surface (l'aire) sous la courbe (AUC, *area under the curve*), qu'on appelle aussi index C (Harrell FE, Jr et al., 1982).
La courbe ROC correspondant aux données précédentes est présentée sur la Figure 7.2.

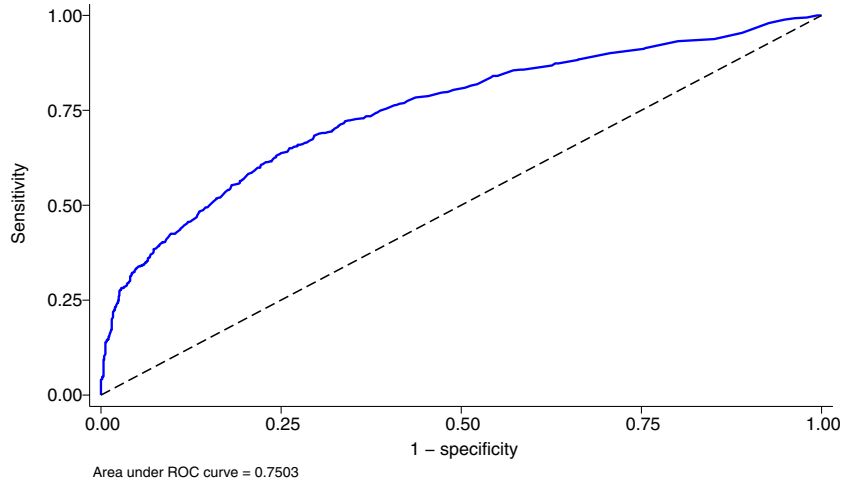


Figure 7.2 : Courbe ROC pour le modèle logistique du Tableau 7.2.

La diagonale hachurée sert de point de repère. Elle représente la situation où, pour chaque valeur seuil, le classement d'un sujet en malade ou non-malade est fait au hasard avec une probabilité de 50 %. On s'attend (on espère !) à ce que le modèle logistique fasse mieux que le hasard, et donc que la courbe ROC se situe au-dessus de la diagonale.

La surface sous la courbe peut valoir de 0 à 1. On montre qu'elle s'interprète de la façon suivante : si on prend au hasard deux sujets, un malade et un non-malade, dont les risques de maladie prédits par le modèle sont respectivement P_1 et P_2, la surface

3. ROC signifie « *Receiver Operating Characteristic* ». Sur le plan historique, l'analyse ROC fait partie de la théorie de la détection du signal développée pendant la Seconde Guerre mondiale pour analyser les images radar. Les opérateurs radar devaient décider si un écho sur leur écran était une cible ennemie, un bateau allié ou seulement du « bruit ». La théorie de la détection du signal mesure la capacité des opérateurs radar (*radar receiver operators*) à faire cette importante discrimination. Cette capacité a été dénommée « *Receiver Operating Characteristic* ». Ce n'est que dans les années 1970 que cette théorie a commencé à être utilisée pour interpréter les résultats des tests médicaux.

sous la courbe est la probabilité que P_1 soit supérieur à P_2, c'est-à-dire la probabilité que les deux sujets soient bien classés.
Si la surface sous la courbe est égale à 50 %, cela signifie que le modèle n'a pas globalement plus de capacité discriminante que le hasard. La règle empirique habituellement retenue pour classer les valeurs de la surface sous la courbe ROC est la suivante :

0,5–0,6 :	très mauvaise	0,8–0,9 :	bonne
0,6–0,7 :	mauvaise	0,9–1,0 :	excellente
0,7–0,8 :	acceptable		

IV.3. Courbe ROC et tests d'adéquation

On dispose finalement de deux catégories d'indices pour l'adéquation du modèle : les tests d'adéquation vus au § 3 et l'aire sous la courbe ROC. Leurs résultats ne sont pas toujours convergents (Steyerberg EW et al., 2005). Cela peut se comprendre de la façon suivante : si on ajoute, par exemple, 0,25 à chaque P_i pour un modèle ayant une bon « fit » (c'est-à-dire une bonne adéquation au sens des tests de χ^2 de Pearson ou de la déviance) et une bonne capacité discriminante (aire sous la courbe ROC suffisamment proche de 1), le nouveau modèle aura un mauvais « fit » alors que sa capacité discriminante globale ne sera pas affectée.
Il est donc conseillé de considérer les deux types d'indices pour apprécier l'adéquation d'un modèle.

V. Diagnostics de régression

Les diagnostics de régression portent sur les observations individuelles (ou plus exactement les profils).

Il ne s'agit pas à proprement parler de savoir si le modèle est « adéquat pour telle ou telle observation ». En effet, un modèle statistique cherche à représenter au mieux l'ensemble des données et conduit donc obligatoirement à des « compromis ». En ce sens, les tests d'adéquation sont par essence des tests globaux.

Il s'agit plutôt de rechercher quelles sont les observations qui ont un poids important dans les statistiques globales d'adéquation du § III (Roy SS et al., 2008). On parle d'observations influentes. Il se trouve que ce sont souvent aussi les observations les moins bien représentées par le modèle, ce qui peut entretenir la confusion. En cas de défaut global d'adéquation, les observations concernées peuvent correspondre à des erreurs de mesure ou de codage. Si ce n'est pas le cas, c'est que les variables prises en compte dans le modèle, ou leur modélisation, ou la forme du lien entre ces variables et Y, ne suffisent pas à décrire correctement les observations faites sur l'échantillon. On peut alors espérer que les observations repérées par les diagnostics de régression donneront des idées sur les modifications à faire. Quoi qu'il en soit, on comprend que ces questions ne relèvent pas de problèmes de fluctuation d'échantillonnage. Il n'est donc pas très surprenant que les diagnostics de

régression ne soient pas assortis de tests. Ce sont plutôt des moyens graphiques qui sont utilisés.

La plupart des notions utilisées viennent de la régression linéaire. Elles ont été adaptées à la régression logistique, notamment dans un article de Pregibon (Pregibon D, 1981, Hosmer DW et al., 2013). On raisonne toujours en profil, puisque les sujets d'un même profil sont « non distinguables », et on s'intéresse à deux types de variations lorsque le profil j est retiré :

- ✓ La variation de la statistique d'adéquation due au profil j ($\Delta\chi^2_j$ ou ΔD^2_j selon la statistique utilisée). Elle mesure le poids de ce profil dans la valeur totale de la statistique d'adéquation.
- ✓ La variation moyenne sur l'estimation des coefficients du modèle $\Delta\hat{\beta}_j$. Elle mesure le poids moyen standardisé du profil j dans les valeurs estimées de l'ensemble des coefficients du modèle.

V.1. Levier associé à un profil

Avant de poursuivre, il est nécessaire d'introduire la notion de levier (*leverage* en anglais), qui sert à calculer les indices de diagnostic que nous allons utiliser. En régression linéaire, le levier « mesure » la distance entre le sujet (ou le profil) observé et la moyenne des observations. On comprend bien intuitivement que plus cette distance est grande, plus l'influence du profil est grande dans le résultat final. C'est en particulier assez clair dans le cas de la régression linéaire simple (une seule variable X), pour laquelle on sait que les points éloignés du centre de gravité du « nuage de points » ont une grande influence sur la valeur de la pente de la droite de régression estimée. Cette droite « pivotant » autour du centre de gravité, ces points ont un fort « bras de levier » pour « influencer la pente de la droite ». Il peut même suffire d'un seul point, ainsi que l'illustre la Figure 7.3.

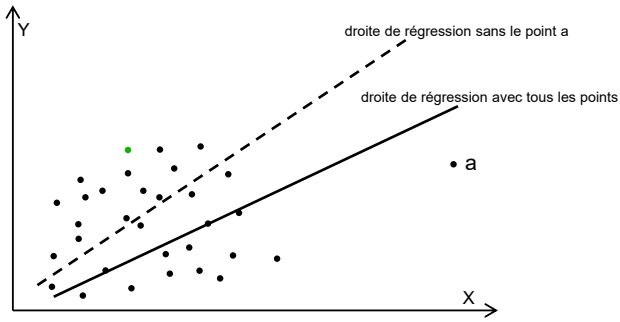


Figure 7.3 : Influence du point a sur la pente de la droite de régression.

Dans le cas de la régression linéaire, la notion de levier est la suivante (Glantz SA et al., 1990). Le modèle linéaire s'écrit $\hat{y} = \alpha + \sum \beta_i x_i$ ou, sous une forme matricielle qui va être plus commode ici, $\hat{y} = Xb$, où b est le vecteur des paramètres ($b' = \left[\alpha \ \beta_1 \ldots \beta_p\right]$) et X la matrice des observations :

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & & x_{p2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{bmatrix}$$

L'estimation des coefficients est donnée par : $b = (X'X)^{-1}X'y$, de sorte qu'on peut écrire : $\hat{y} = Xb = X(X'X)^{-1}X'y = Hy$. La matrice H est appelée en anglais « *hat matrix* », car elle transforme y en $\hat{y}$. Ses coefficients sont notés h_{ij}. Le levier associé à x_j est le terme diagonal h_{jj}.
On peut montrer que h_{jj} est donc compris entre 0 et 1, et que : $\sum_{i=1}^{n} h_{ji} = 1$ et $h_{jj} = \sum_{i=1}^{n} h_{ji}^2$. Dans le cas d'une seule variable, $h_{ij} = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum (x_j - \bar{x})^2}$, ce qui montre bien que h_{jj} est d'autant plus grand que x_i est éloigné de la moyenne $\bar{x}$. Dans le cas de p variables, on montre que la moyenne des h_{jj} est $(p+1)/n$, et l'habitude est de considérer que h_{jj} est grand s'il dépasse $2(p+1)/n$.

Dans le cas de la régression logistique, la matrice H est définie « par extension » par $H = V^{1/2}X(X'VX)^{-1}X'V^{1/2}$, où V est une matrice diagonale dont le j-ième élément est $v_j = m_j P_j(1-P_j)$. On montre (après quelques calculs…) que le levier h_j associé au profil j est égal à $h_j = m_j P_j(1-P_j)x_j'(X'VX)^{-1}x_j'$.
On montre aussi, cependant (Hosmer DW et al., 2013), que h_j possède de moins bonnes propriétés que dans le cas linéaire. Notamment, la valeur de h_j n'augmente pas toujours lorsque le profil s'éloigne de la moyenne (c'est-à-dire pour les grandes et les petites valeurs de P_j). En particulier, on constate que h_j diminue fortement lorsque la valeur P_j associée s'approche de 0 ou de 1. La notion de levier est donc peu utilisable (et peu utilisée) en régression logistique, h_j n'intervenant que dans les expressions pour calculer d'autres indices de diagnostic.

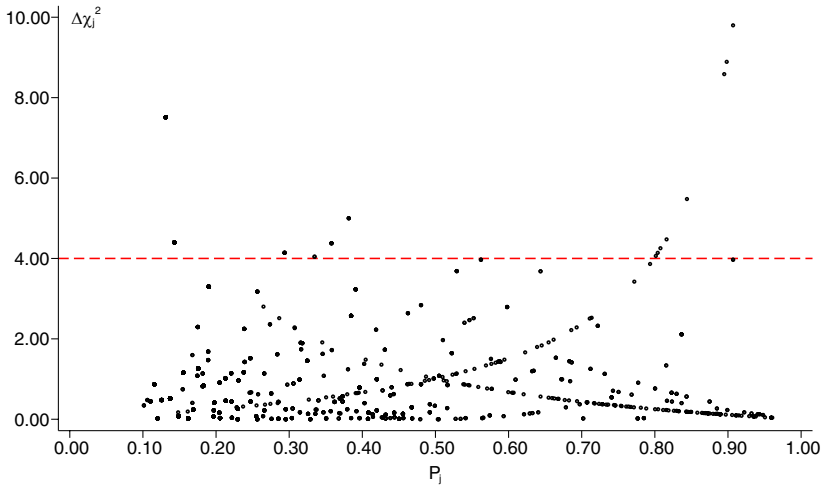
V.2. Influence d'un profil sur les statistiques d'adéquation

Pour calculer la variation du χ^2 d'adéquation associé au profil j, il faut calculer la différence entre les χ^2 d'adéquation avec et sans le profil j. Cette différence est notée $\Delta\chi_j^2$ si on utilise la statistique de Pearson ou ΔD_j^2 si on utilise celle du rapport des vraisemblances. Ces calculs sont cependant assez lourds, car il faut estimer les coefficients du modèle logistique et calculer les statistiques d'adéquation après avoir retiré successivement chacun des profils. Pour les simplifier, D. Pregibon a montré que les modifications des statistiques d'adéquation lorsque le profil j est retiré sont approximativement égales à : $\Delta\chi_j^2 = \frac{r_j^2}{(1-h_j)}$ et $\Delta D_j^2 = \frac{d_j^2}{(1-h_j)}$ (Pregibon D, 1981).

On peut aussi montrer que $\Delta\chi_j^2$ et ΔD_j^2 suivent approximativement une loi de χ^2 à un degré de liberté. Cela signifie qu'on s'attend à ce qu'environ 5 % des valeurs

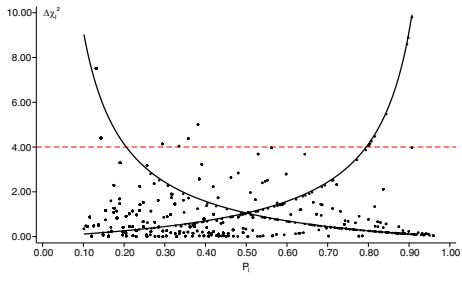
dépassent 4, et que pratiquement aucune ne dépasse 10. Cela donne donc des points de repère, sachant qu'on s'intéresse aux valeurs les plus grandes de $\Delta\chi_j^2$ et ΔD_j^2, qui correspondent aux profils les plus influents pour les statistiques d'adéquation.

Pour représenter les profils influents, on trace le graphe de $\Delta\chi_j^2$ ou ΔD_j^2 en fonction de P_j (Hilbe JM, 2009, Hosmer DW et al., 2013). Dans le cas de $\Delta\chi_j^2$, et avec les données précédentes, le graphique est celui de la Figure 7.4.
Applica



Le trait pointillé à la hauteur 4 n'est qu'un point de repère pour les profils les plus « excentriques ».
Figure 7.4 : Graphique donnant $\Delta\chi_j^2$ en fonction de P_j.

Remarque
On note sur la Figure 7.4 que certains des points font apparaître deux courbes en forme de parabole, qui sont dessinées sur le graphique ci-contre. Elles sont d'autant plus marquées qu'il y a beaucoup de profils ne contenant qu'un seul sujet (c'est-à-dire notamment quand il y a des variables avec beaucoup de classes, comme l'âge en années ici, ou des variables quantitatives).



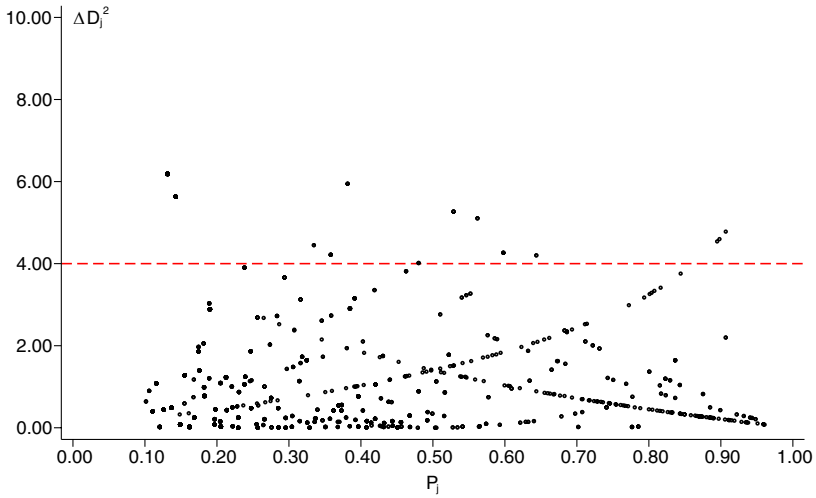
L'explication de la présence de ces courbes tient aux valeurs particulières prises par $\Delta\chi_j^2$ lorsqu'il n'y a qu'un seul sujet dans le profil j ($m_j = 1$) :

- si le sujet est un non-malade ($n_{1j} = 0$), on a $\Delta\chi_j^2 \approx \dfrac{P_j}{(1-P_j)\left(1-CP_j(1-P_j)\right)}$, où C est une constante, ce qui correspond à la courbe de droite ;

- si le sujet est un malade ($n_{1j} = 1$), on a $\Delta\chi_j^2 \approx \dfrac{(1-P_j)}{P_j\left(1-CP_j(1-P_j)\right)}$, ce qui correspond à la courbe de gauche.

Dans le cas de ΔD_j^2, on obtient un graphique analogue, avec toujours les deux mêmes types de courbes (Figure 7.5). Cette fois-ci, lorsque $m_j = 1$, on a :

$\Delta D_j^2 \approx \dfrac{2\left|\ln(1-P_j)\right|}{\left(1-CP_j(1-P_j)\right)}$ si $n_{1j} = 0$ et $\Delta D_j^2 \approx \dfrac{2\left|\ln P_j\right|}{\left(1-CP_j(1-P_j)\right)}$ si $n_{1j} = 1$.



Le trait pointillé à la hauteur 4 n'est qu'un point de repère pour les profils les plus « excentriques ».

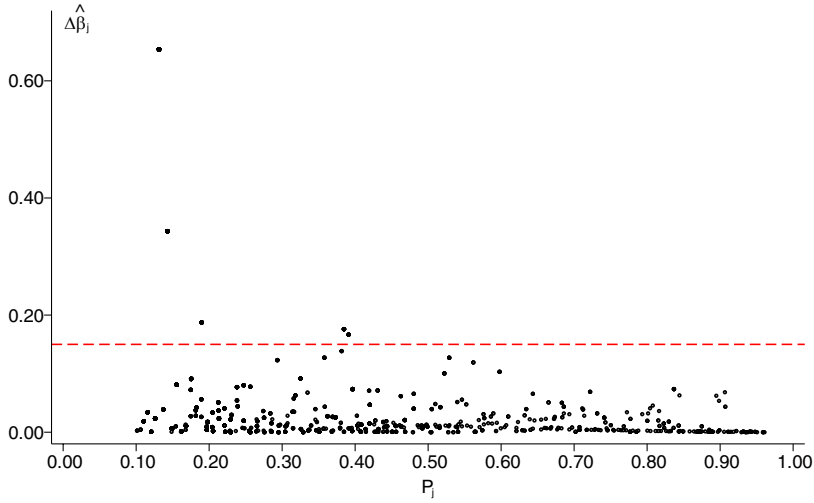
Figure 7.5 : Graphique donnant ΔD_j^2 en fonction de P_j.

V.3. Influence sur l'estimation des coefficients

La modification moyenne des coefficients estimés du modèle lorsque le profil j est retiré est égale à la moyenne, pondérée et standardisée sur la matrice de covariance, des différences entre les valeurs β et β_{-j} des coefficients estimés avec l'ensemble des profils et après avoir retiré le profil j.

On montre (Pregibon D, 1981) que cette modification moyenne peut être obtenue, par une approximation linéaire, en utilisant l'expression $\Delta\hat{\beta}_j = \left(\hat{\beta} - \hat{\beta}_{(-j)}\right)'X'VX\left(\hat{\beta} - \hat{\beta}_{(-j)}\right)$, qui se simplifie en : $\Delta\hat{\beta}_j = \dfrac{r_j^2 h_j}{(1-h_j)^2}$. L'interprétation quantitative de la valeur de $\Delta\hat{\beta}_j$ n'est pas facile en raison de la standardisation. Elle ne donne pas d'idée précise sur la variation des $\hat{\beta}_j$ eux-mêmes. On considère qu'il faut s'intéresser aux profils dont la valeur de $\Delta\hat{\beta}_j$ s'écarte de celle des autres (Hosmer DW et al., 2013).

Dans l'exemple de l'enquête sur les facteurs de risque de GEU, la Figure 7.6 représente les valeurs de $\Delta\hat{\beta}_j$ en fonction de P_j. Le trait pointillé sur le graphique, situé à 0,15, est assez arbitraire. Il isole 5 profils correspondant à une modification moyenne de l'estimation des coefficients supérieure aux autres.



Le trait pointillé à la hauteur 0,15 n'est qu'un point de repère pour les profils les plus « excentriques ».
Figure 7.6 : Graphique donnant $\Delta\hat{\beta}_j$ en fonction de P_j.

V.4. Examen des points influents

Pour poursuivre l'analyse, on peut détailler les caractéristiques des profils ayant les plus grandes valeurs de $\Delta\chi^2_j$, ΔD^2_j et/ou de $\Delta\hat{\beta}_j$ pour les variables incluses dans le modèle.
Le Tableau 7.7 donne les caractéristiques des profils tels que $\Delta\chi^2_j$ ou ΔD^2_j est supérieur à 4 (il y en a 19), classés dans l'ordre décroissant de ΔD^2_j.
On constate qu'il y a principalement deux grands types de profils : ceux pour lesquels l'écart entre les probabilités de maladie observée et prédite n'est pas très grand, mais qui comprennent beaucoup de sujets (par exemple le profil n° 61), et ceux pour lesquels les probabilités de maladie observée et prédite sont très différentes, mais qui ne réunissent que quelques sujets (par exemple le profil n° 251). Chacun joue un rôle important dans l'adéquation globale, mais pour des raisons différentes (à peu près « symétriques »).

Ce tableau donne aussi la modification moyenne des valeurs des coefficients du modèle logistique ($\Delta\hat{\beta}_j$) lorsque le profil j est retiré, dont la première (pour le profil n° 37) paraît importante. On peut préciser ce point en calculant la modification induite par le retrait de chacun des profils sur l'estimation de chaque paramètre (Tableau 7.8).

N° du	Nb de sujets	Variables figurant dans le modèle					$\Delta\hat{\beta}_j$	$\Delta\chi^2_j$	ΔD^2_j	Probabilité de maladie	
profil	m_j	ctub	agea	tabfc	afcs	ainf				observée	prédite (P_j)
37	56	0	24	0	0	0	0,65	7,52	6,18	0,25	0,13
189	3	0	35	1	1	0	0,14	5,00	5,94	1,00	0,38
61	59	0	26	0	0	0	0,34	4,40	5,63	0,05	0,14
170	4	0	34	3	0	0	0,13	3,69	5,27	1,00	0,53
44	3	0	24	2	0	1	0,12	3,97	5,10	0,00	0,56
251	1	1	24	3	0	1	0,07	9,80	4,78	0,00	0,91
261	1	1	26	2	0	1	0,05	8,89	4,60	0,00	0,90
339	1	1	38	3	0	0	0,06	8,58	4,54	0,00	0,89
52	2	0	25	0	1	1	0,07	4,04	4,45	1,00	0,33
83	4	0	27	2	0	1	0,10	2,79	4,27	1,00	0,60
167	5	0	34	0	0	1	0,13	4,38	4,22	0,80	0,36
133	2	0	31	2	0	1	0,07	3,68	4,20	0,00	0,64
172	3	0	34	2	0	0	0,07	2,84	4,02	0,00	0,48
356	1	1	42	1	1	0	0,06	5,48	3,76	0,00	0,84
89	6	0	28	0	0	1	0,12	4,14	3,66	0,67	0,29
344	1	1	39	0	0	1	0,04	4,48	3,42	0,00	0,82
269	1	1	27	1	0	1	0,05	4,25	3,33	0,00	0,81
268	1	1	27	2	0	0	0,04	4,14	3,29	0,00	0,80
333	1	1	37	0	0	1	0,03	4,06	3,26	0,00	0,80

Tableau 7.7 : Caractéristiques des profils tels que $\Delta\chi^2_j$ ou ΔD^2_j est supérieur à 4

Pour ne pas trop alourdir la présentation, le Tableau 7.8 se limite aux sept profils ayant les plus grandes valeurs de ΔD^2_j. Il donne les pourcentages de variations des estimations des coefficients lorsque les sujets d'un profil sont retirés, puis lorsque les sept profils sont tous retirés. Il donne également les variations correspondantes observées des statistiques d'adéquations D^2 et de χ^2, ce qui permet de les confronter aux valeurs approchées calculées ($\Delta\chi^2_j$ et ΔD^2_j). Enfin, le tableau donne les valeurs successives du test de Hosmer–Lemeshow.

		N° du profil retiré de l'analyse							
	Tous les sujets	37	189	61	170	44	251	339	Tous retirés
nbre de sujets dans l'analyse	1682	1626	1679	1623	1678	1679	1681	1681	1551
	Coefficient	Variation (%) de l'estimation des coefficients							
ctub	1,83	0,5 %	0,5 %	– 0,5 %	0,5 %	– 0,5 %	1,6 %	2,2 %	5,5 %
agea	0,0488	11,9 %	– 2,9 %	– 5,5 %	– 3,3 %	– 2,0 %	– 2,0 %	2,3 %	7,6 %
tab1	0,531	10,2 %	– 8,3 %	– 7,5 %	– 0,2 %	0,2 %	0,2 %	0,4 %	7,9 %
tab2	1,32	4,5 %	0,0 %	– 3,0 %	0,0 %	3,0 %	0,0 %	0,8 %	3,8 %
tab3	1,52	3,3 %	0,0 %	– 3,3 %	– 3,9 %	0,0 %	1,3 %	1,3 %	5,9 %
afcs	0,337	6,2 %	– 7,4 %	– 5,9 %	3,0 %	– 2,1 %	– 1,2 %	– 2,1 %	– 6,8 %
ainf	0,817	2,8 %	1,2 %	– 2,6 %	0,5 %	4,0 %	1,7 %	– 1,1 %	– 3,1 %
Cste	– 3,06	7,5 %	– 1,3 %	– 4,2 %	– 1,3 %	– 1,0 %	– 0,7 %	1,0 %	4,9 %
Variation de D^2		Variation absolue des statistiques d'adéquation							
Observée		6,30	5,92	5,53	5,22	5,07	4,81	4,57	
Approchée par ΔD_j^2		6,18	5,94	5,63	5,27	5,1	4,78	4,54	
Variation de χ^2									
Observée		3,59	4,39	7,14	4,8	3,12	7,45	6,20	
Approchée par $\Delta\chi_j^2$		7,52	5,00	4,40	3,69	3,97	9,8	8,58	
Test de Hosmer–Lemeshow	7,41	2,72	7,18	5,17	10,1	6,80	7,52	7,85	3,88

Tableau 7.8 : Variations des coefficients du modèle et diagnostics de régressions quand les profils les plus influents sont retirés de l'échantillon

On constate tout d'abord que les variations observées de D^2 et de χ^2 sont assez proches des valeurs approchées par $\Delta\chi_j^2$ et ΔD_j^2, ce qui conforte, s'il en était besoin, l'article de D. Pregibon (Pregibon D, 1981). Lorsqu'il y a une différence, les valeurs approchées surestiment la réalité, ce qui semble être souvent constaté (Hosmer DW et al., 2013).

On voit aussi que les modifications des valeurs des coefficients liées à chaque profil restent le plus souvent inférieures à 5 % et ne dépassent jamais 12 %. Lorsque les sept profils sont tous retirés de l'analyse, la modification des coefficients reste inférieure à 10 %.

Il ne faut cependant pas oublier qu'on se situe dans un contexte où le test global d'adéquation est non significatif. On ne doit donc pas s'attendre à ce qu'un profil ait une importance particulièrement grande.

Références

Achour J, Abulizi D, Makinson A, Arvieux C, Bonnet F, Goujard C, Lambert O, Slama L, Blain H, Meyer L, Allavena C, Group SS. One-Year Frailty Transitions Among Persons With HIV Aged 70 Years or Older on Antiretroviral Treatment. Open Forum Infect Dis. 2024;11(7):ofae229.

Agresti A. Categorical data analysis. New York; 1990.

Allison PD. What's the Best R-Squared for Logistic Regression Statistical Horizons. 2013 [cited 2024–03–12]; Available from: https://statisticalhorizons.com/r2logistic/

Allison PD. Measures of fit for fogistic regression. 2014 [cited 2024–03–13]; Available from: https://support.sas.com/resources/papers/proceedings14/1485-2014.pdf.

Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994;86:829–35.

Altman DG. Categorizing continuous variables. In: Armitage P, Colton T, editors. Encyclopedia of Biostatistics. Chichester: John Wiley and Sons; 1998. p. 563–7.

Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332(7549):1080.

Ambler G, Royston P. Fractional polynomial model selection procedures: investigation of type I error rate. J Statist Comput Simul. 2001;69(1):89-108.

American College of Obstetricians and Gynecologists. Shoulder Dystocia. Washington, DC; 1997 October. Report No. 7.

Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. Int J Epidemiol. 1997;26(6):1323–33.

Armitage P, Berry G. Stat Methods Med Res. 1987.

Armstrong BG, Sloan M. Ordinal regression models for epidemiologic data. Am J Epidemiol. 1989;129(1):191–204.

Austin PC, Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. Stat Med. 2004a;23(7):1159–78.

Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol. 2004b;57(11):1138–46.

Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med. 2007;26(16):3078–94.

Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. J Clin Epidemiol. 2015;68(6):627–36.

Banack HR, Mayeda ER, Naimi AI, Fox MP, Whitcomb BW. Collider Stratification Bias I: Principles and Structure. Am J Epidemiol. 2023;193(2):238–40.

Barnwell-Menard JL, Li Q, Cohen AA. Effects of categorization method, regression type, and variable distribution on the inflation of Type-I error rate when categorizing a confounding variable. Stat Med. 2015;34(6):936–49.

Barrio I, Arostegui I, Rodríguez-Álvarez M-X, Quintana J-M. A new approach to categorising continuous variables in prediction models: Proposal and validation. Stat Methods Med Res. 2017;26(6):2586–602.

Barros A, Hirakata V. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol. 2003;3(1):21.

Bauldry S, Xu J, Fullerton AS. gencrm: A new command for generalized continuation-ratio models. Stata J. 2018;18(4):924–36.

Becher H. The concept of residual confounding in regression models and some applications. Stat Med. 1992;11:1747–58.

Begg CB, Gray R. Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika. 1984;71(1):11–8.

Benner A. mfp: Multivariable fractional polynomials. R News. 2005;5:20–3.

Benner A. Multivariable Fractional Polynomials. 2022 [cited June 2023]; Available from: https://cran.r-project.org/web/packages/mfp/vignettes/mfp_vignette.pdf.

Bennette C, Vickers A. Against Quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol. 2012;12(1):21.

Besse P, Thomas-Agnan C. Le lissage par fonctions splines en statistiques, revue bibliographique. Statistique et analyse des données. 1989;14(1):55–84.

Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. Stat Med. 2013;32(13):2262–77.

Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, Moons KGM. External validation is necessary in prediction research: A clinical example. J Clin Epidemiol. 2003;56(9):826–32.

Blettner M, Sauerbrei W. Influence of model-building strategies on the results of a case-control study. Stat Med. 1993;12(14):1325–38.

Bouyer J, Dreyfus J, Gueguen S, Lazar P, Papiernik E. La prématurité. Enquête périnatale de Haguenau. Paris: INSERM et DOIN Éditeurs 1987.

Bouyer J, Hémon D, Cordier S, Derriennic F, Stücker I, Stengel B, Clavel J. Épidémiologie – Principes et méthodes quantitatives: INSERM; 1993.

Bouyer J, Coste J, Shojaei T, Pouly JL, Fernandez H, Gerbaud L, Job-Spira N. Risk factors for ectopic pregnancy: a comprehensive analysis based on a large case-control population-based study in France. Am J Epidemiol. 2003;157(3):185–94.

Bouyer J. Méthodes statistiques. Médecine – Biologie. Paris: Vuibert; 2017.

Brenner H, Blettner M. Misclassification bias arising from random error in exposure measurement: implications for dual measurement strategies. Am J Epidemiol. 1993;138:453–61.

Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. Epidemiology. 1997;8:429–34.

Brenner H. A potential pitfall in control of covariates in epidemiologic studies. Epidemiology. 1998;9:68–71.

Breslow NE, Day NE. Statistical methods in cancer research. Volume I. The design and analysis of case control studies. London: Oxford University Press; 1980.

Buettner P, Garbe C, Guggenmoos-Holzmann I. Problems in defining cutoff points of continuous prognostic factors: Example of tumor thickness in primary cutaneous melanoma. J Clin Epidemiol. 1997;50(11):1201–10.

Buis ML. Direct and indirect effects in a logit model. Stata J. 2010;10(1):11–29.

Bürkner P-C, Vuorre M. Ordinal Regression Models in Psychology: A Tutorial. Adv Methods Pract Psychol Sci. 2019;2(1):77–101.

Buse A. The likelihood-ratio, Wald, and Lagrange multiplier tests: an expository note. Amer Statist. 1982;36(3):153–7.

Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol. 2003;158(3):280–7.

Chatterjee S, Hadi AS. Regression analysis by example. 5th ed. Hoboken, NJ: Wiley; 1986.

Chen L-C, Wang J-Y. Testing the fit of the logistic model for matched case-control studies. Comput Stat Data Anal. 2013;57(1):309–19.

Cleveland W. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc. 1979;74(368):829–36.

Cleveland W, Devlin S. Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. J Am Stat Assoc. 1988;83(403):596–610.

Cochran WG. Some Methods for Strengthening the Common c^2 Tests. Biometrics. 1954;10(4):417–51.

Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics. 1968;24(2):295–313.

Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, Poole C. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2009;39(2):417–20.

Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. J Clin Epidemiol. 2015;68(2):112–21.

Commenges D, Jacqmin-Gadda H. Modèles biotatistiques pour l'épidémiologie. Louvain-la-Neuve, Belgique: De Boeck Supérieur; 2015.

Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis I. Background, goals, and general strategy. J Clin Epidemiol. 1995;48(12):1495–501.

Connor RJ. Grouping for Testing Trends in Categorical Data. J Am Stat Assoc. 1972;67(339):601–4.

Coste J, Job-Spira N, Aublet-Cuvelier B, Germain E, Glowaczover E, Fernandez H, Pouly JL. Incidence of ectopic pregnancy. First results of a population-based register in France. Hum Reprod. 1994;9:742–5.

Coste J, Bouyer J, Ughetto S, Gerbaud L, Fernandez H, Pouly J-L, Job-Spira N. Ectopic pregnancy is again on the increase. Recent trends in the incidence of ectopic pregnancies in France (1992–2002). Hum Reprod. 2004;19(9):2014–8.

Courtois E, Tubert-Bitter P, Ahmed I. New adaptive lasso approaches for variable selection in automated pharmacovigilance signal detection. BMC Med Res Methodol. 2021; 21(1):271.

Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron Al, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64(9):993–1000.

Cox DR, Snell EJ. Analysis of Binary Data. Second ed: Chapman & Hall; 1989.

Cox DR, Wermuth N. A Comment on the Coefficient of Determination for Binary Responses. Amer Statist. 1992;46(1):1–4.

D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17(19):2265–81.

Dales LG, Ury HK. An Improper Use of Statistical Significance Testing in Studying Covariables. Int J Epidemiol. 1978;7(4):373–6.

Dawid AP. Commentary: Counterfactuals: help or hindrance? Int J Epidemiol. 2002;31(2):429–31.

Del Priore G, Zandieh P, Lee MJ. Treatment of continuous data as categoric variables in Obstetrics and Gynecology. Obstetrics & Gynecology. 1997;89(3):351–4.

Dinero TE. Seven Reasons Why You Should NOT Categorize Continuous Data. J Health Soc Policy. 1996;8(1):63–72.

Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med. 1989;8(5):551–61.

Edlinger M, van Smeden M, Alber HF, Wanitschek M, Van Calster B. Risk prediction models for discrete ordinal outcomes: Calibration and the impact of the proportional odds assumption. Stat Med. 2022;41(8):1334–60. doi: 10.002/sim.9281. Epub 2021 Dec 12.

Evans D, Chaix B, Lobbedez T, Verger C, Flahault A. Combining directed acyclic graphs and the change-in-estimate procedure as a novel approach to adjustment-variable selection in epidemiology. BMC Med Res Methodol. 2012;12(1):156.

Evans S, Li L. A comparison of goodness of fit tests for the logistic GEE model. Stat Med. 2005;24(8):1245–61.

Fagerland MW, Hosmer DW, Bofin AM. Multinomial goodness-of-fit tests for logistic regression models. Stat Med. 2008;27(21):4238–53.

Fagerland MW. adjcatlogit, ccrlogit, and ucrlogit: Fitting ordinal logistic regression models. Stata J. 2014;14(4):947–64.

Fagerland MW, Hosmer D. How to test for goodness of fit in ordinal logistic regression models. Stata J. 2017;17(3):668–86.

Falissard B. Comprendre et utiliser les statistiques dans les sciences de la vie. 3e ed. Paris: Masson; 2005.

Froslie K, Roislien J, Laake P, Henriksen T, Qvigstad E, Veierod M. Categorisation of continuous exposure variables revisited. A response to the Hyperglycaemia and Adverse Pregnancy Outcome (HAPO) Study. BMC Med Res Methodol. 2010;10(1):103.

Fullerton AS. A Conceptual Framework for Ordered Logistic Regression Models. Sociol Methods Res. 2009;38(2):306–47.

Gaillard P. High-dimensional statistics (The LASSO). 2020 [cited]; Available from: http://pierre.gaillard.me/doc/2020_cours_lasso_telecom.pdf.

Garrido MM, Kelley AS, Paris J, Roza K, Meier DE, Morrison RS, Aldridge MD. Methods for constructing and assessing propensity scores. Health Serv Res. 2014;49(5):1701–20.

Genell A, Nemes S, Steineck G, Dickman P. Model selection in Medical Research: A simulation study comparing Bayesian Model Averaging and Stepwise Regression. BMC Med Res Methodol. 2010;10(1):108.

Glantz SA, Slinker BK. Primer of applied regression and analysis of variance. New York: McGraw-Hill; 1990.

Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health. 1989;79(3):340–9.

Greenland S. Alternative models for ordinal logistic regression. Stat Med. 1994;13:1665–77.

Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology. 1995a;6(4):450–4.

Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology. 1995b;6(4):356–65.

Greenland S. Problems in the average-risk interpretation of categorical dose-response analyses. Epidemiology. 1995c;6(5):563–5.

Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol. 1995;142(12):1255–64.

Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37–48.

Greenland S, Brumback B. An overview of relations among causal modelling methods. Int J Epidemiol. 2002;31(5):1030–7.

Greenland S. Invited Commentary: Variable Selection versus Shrinkage in the Control of Multiple Confounders. Am J Epidemiol. 2008;167(5):523–9.

Greenland S, Pearce N. Statistical Foundations for Model-Based Adjustments. Annu Rev Public Health. 2015;36(1):89–108.

Gutierrez R, Linhart J, Pitblado JS. From the help desk: Local polynomial regression and Stata plugins. Stata J. 2003;3(4):412–9.

Haria K. Collider Bias. 2020 [cited 2021 2021/10/17]; Available from: https://rstudio-pubs-static.s3.amazonaws.com/623981_6589ceca95e64f78a590acf9eeb2f600.html.

Harrel Jr. FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361–87.

Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Spinger; 2001.

Harrell FE, Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. JAMA. 1982;247(18):2543–6.

Harrell FE, Jr., Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep. 1985;69(10):1071–77.

Heinze G, Baillie M, Lusa L, Sauerbrei W, Schmidt CO, Harrell FE, Huebner M, on behalf of TG, initiative TGotS. Regression without regrets – initial data analysis is a prerequisite for multivariable regression. BMC Med Res Methodol. 2024;24(1):178.

Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology. Am J Epidemiol. 2002;155(2):176–84.

Hernan MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health. 2004;58(4):265–71.

Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615–25.

Hernan MA, Cole SR. Invited Commentary: Causal Diagrams and Measurement Bias. Am J Epidemiol. 2009;170(8):959–62.

Hernandez-Diaz S, Schisterman EF, Hernan MA. Hernandez-Diaz et al. Respond to "The Perils of Birth Weight". Am J Epidemiol. 2006a;164(11):1124–5.

Hernandez-Diaz S, Schisterman EF, Hernan MA. The Birth Weight "Paradox" Uncovered? Am J Epidemiol. 2006b;164(11):1115–20.

Hernberg S, Nurminen M, Tolonen M. Excess mortality from coronary heart disease in viscose rayon workers exposed to carbon disulfude. Scand J Work Environ Health. 1973;10:93–9.

Hilbe JM. Logistic regression models. Boca Raton, FL: Chapman & Hall/CRC; 2009.

Hilbe JM. Logistic regression. 2020 [cited 2023; Available from: https://encyclopediaofmath.org/images/6/69/Logistic_regression.pdf.

Hoffmann K, Pischon T, Schulz M, Schulze MB, Ray J, Boeing H. A Statistical Test for the Equality of Differently Adjusted Incidence Rate Ratios. Am J Epidemiol. 2008;167(5):517–22.

Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics – Theory and Methods. 1980;9(10):1043–69.

Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med. 1997;16(9):965–80.

Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Third Edition. New York: John Wiley & Sons; 2013.

Ibrahim N, Hassler C, Jousselme C, Barry C, Lefevre H, Falissard B, Bouyer J, Rouquette A. Chronic conditions, subjective wellbeing and risky sexual behaviour among adolescents and young adults. Eur J Pediatr. 2023;182(3):1163–71.

Ismaili A, Gaillard P. Le Lasso, ou comment choisir parmi un grand nombre de variables à l'aide de peu d'observations. 2009. http://pierre.gaillard.me/doc/Ga09-report.pdf.

Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol. 1999;150(4):327–33.

Jousselme C, Cosquer M, Hassler C. Portraits d'adolescents. Enquête épidémiologique multicentrique en milieu scolaire en 2013. Paris: Inserm; 2013.

Kahan BC. Accounting for centre-effects in multicentre trials with a binary outcome – when, why, and how? BMC Med Res Methodol. 2014;14(1):20.

Korn L, Hosmer DW, Lemeshow S. The performance of goodness of fit tests for logistic tegression with discrete covariates. Biom J. 1986;28:697–708.

Kwiatkowski F, Karem S, Verrelle P, Chamorey E, Kramar A. Le score de propension : intérêt et limites. Bull Cancer. 2007;94(7–8):680–6.

Lagakos SW. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. Stat Med. 1988;7(1–2):257–74.

Lange T, Vansteelandt S, Bekaert M. A simple unified approach for estimating natural direct and indirect effects. Am J Epidemiol. 2012;176(3):190–5.

Last J. A dictionary of epidemiology. Oxford: Oxford Medical Publications; 1983.

Leclerc A, Papoz L, Bréart G, Lellouch J. Dictionnaire d'épidémiologie: Eds Frison-Roche; 1990.

Lellouch J, Ducimetiere P, Hémon D, Kaminski M. Méthodes quantitatives en épidémiologie. Séminaire Inserm et cours pour le DEA de Santé publique. Villejuif: Inserm; 1988. https://hal.science/hal-04464373.

Lemeshow S, Hosmer DW, Jr. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol. 1982;115(1):92–106.

Leplège A, Bizouarn P, Coste J, (sous la direction de). De Galton à Rothman. Les grands textes de l'épidémiologie au XX^{e} siècle. Paris: Hermann; 2011.

Li CY, Sung FC. A review of the healthy worker effect in occupational epidemiology. Occup Med (Lond). 1999;49(4):225–9.

Little RJA. Regression With Missing X's: A Review. J Am Stat Assoc. 1992;87(420):1227–37.

Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken, New Jersey: John Wiley & Sons; 2002.

Liu I, Agresti A. The Analysis of Ordered Categorical Data: A.n Overview and a Survey of Recent Developments. Sociedad de Estadistica e Investigacion Operativa. 2005;14(1):1–73.

Lu B, Cai D, Tong X. Testing causal effects in observational survival data using propensity score matching design. Stat Med. 2018;37(11):1846–58.

Lu H, Gonsalves GS, Westreich D. Selection Bias Requires Selection: The Case of Collider Stratification Bias. Am J Epidemiol. 2023.

Lumley J, Kronmal R. Relative risk regression in medical research: models, contrasts, estimators, and algorithms. UW Biostatistics Working Paper Series. 2006.

Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med. 2004;23(19):2937–60.

Luque-Fernandez MA, Schomaker M, Redondo-Sanchez D, Jose Sanchez Perez M, Vaidya A, Schnitzer ME. Educational Note: Paradoxical collider effect in the analysis of non-communicable disease epidemiological data: a reproducible illustration and web application. Int J Epidemiol. 2018;48(2):640–53.

Mabikwa OV, Greenwood DC, Baxter PD, Fleming SJ. Assessing the reporting of categorised quantitative variables in observational epidemiological studies. BMC Health Serv Res. 2017;17(1):201.

MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychol Methods. 2002;7(1):19–40.

Mair P, Reise SP, Bentler PM. IRT Goodness-of-Fit Using Approaches from Logistic Regression. UCLA: Department of Statistics 2008-01-09 [cited 2024-03-19]; Available from: https://escholarship.org/uc/item/1m46j62q.

Maldonado G, Greenland S. Simulation study of confounder-selection strategies. Am J Epidemiol. 1993;138(11):923–36.

Maldonado G, Greenland S. Estimating causal effects. Int J Epidemiol. 2002;31(2):422–9.

McCullagh P, Nelder JA. Generalized linear models. second ed. London: Chapman & Hall; 1989.

McCulloch CE, Searle SR. Generalized, linear and mixed models. New York: John Wiley & Sons; 2001.

McNutt L-A, Hafner J-P, Xue X. Correcting the Odds Ratio in Cohort Studies of Common Outcomes. [Letter]. JAMA. 1999;282(6):529.

McNutt L-A, Wu C, Xue X, Hafner JP. Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. Am J Epidemiol. 2003;157(10):940–3.

Mehta CR, Patel NR. Exact logistic regression: Theory and examples. Stat Med. 1995;14(19):2143–60.

Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. Am J Epidemiol. 1989;129(1):125–37.

Mittlböck M, Schemper M. Explained variation for logistic regression. Stat Med. 1996;15(19):1987–97.

Molenberghs G, Kenward MG. Missing data in clinical Studies. Chichester: John Wiley & Sons; 2007.

Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375.

Morabia A, (ed). A history of epidemiologic methods and concepts: Springer Basel AG; 2004.

Moser BK, Coombs LP. Odds ratios for continuous outcome variables without dichotomizing. Stat Med. 2004;23(12):1843–60.

Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;78(3):691–2.

Newson R. Sensible parameters for univariate and multivariate splines. Stata J. 2012;12(3):479–504.

Parry S. Ordinal Logistic Regression models and Statistical Software: What You Need to Know. 2020 [cited sdptember 2023]; Available from: https://cscu.cornell.edu/wp-content/uploads/91_ordlogistic.pdf.

Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669–710.

Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48(12):1503–10.

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373–9.

Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. BMC Med Res Methodol. 2019;19(1):46-.

Petersen M, Deddens J. A comparison of two methods for estimating prevalence ratios. BMC Med Res Methodol. 2008;8(1):9.

Pischon T, Schulze MB, Drogan D, Boeing H. Pischon et al. Respond to "Variable Selection versus Shrinkage in Control of Confounders". Am J Epidemiol. 2008;167(5):530–1.

Pluntz M, Dalmasso C, Tubert-Bitter P, Ahmed I. A Simple Information Criterion for Variable Selection in High-Dimensional Regression. Stat Med. 2025;44(1–2):e10275.

Pregibon D. Logistic regression diagnostics. Ann Stat. 1981;9(4):705–24.

Rakotomalala R. Colinéarité et sélection des variables. Régression Linéaire Multiple. 2017 [cited 2022 April, 4]; Available from: http://eric.univ-lyon2.fr/~ricco/cours/slides/Reg_Multiple_Colinearite_Selection_Variables.pdf.

Rayner JCW. The asymptotically optimal tests. The Statistician. 1997;46(3):337–46.

Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. Int J Epidemiol. 2013;42(5):1511–9.

Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology. 2001;12(3):313–20.

Roese NJ. Conterfactual thinking. Psychol Bull. 1997;121:133–48.

Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika. 1983;70(1):41–55.

Rothman KJ, Greenland S. Modern Epidemiology. 1998.

Rothman KJ, Lash T, Greenland S. Modern epidemiology: Lippincott Williams & Wilkins; 2008.

Roy SS, Guria S. Diagnostics in logistic regression models. J Korean Stat Soc. 2008;37(2):89–94.

Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat. 1994;43(3):429–67.

Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol. 1999;28:964–74.

Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21(15):2175–97.

Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Stat Med. 2004;23(16):2509–25.

Royston P, Sauerbrei W. Building multivariable regression models with continuous covariates in clinical epidemiology--with an emphasis on fractional polynomials. Methods Inf Med. 2005;44(4):561–71.

Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127–41.

Royston P, Sauerbrei W. Multivariable modeling with cubic regression splines: A principled approach. Stata J. 2007;7(1):45–70.

Royston P, Sauerbrei W. Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: John Wiley & Sons; 2008.

Royston P, Sauerbrei W. Interaction of treatment with a continuous variable: simulation study of power for several methods of analysis. Stat Med. 2014;33(27):4695–708.

Rubin BB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.

Sauerbrei W, Royston P. Multivariable Fractional Polynomials (MFP). 2010 [cited May 23, 2023]; Available from: https://mfp.imbi.uni-freiburg.fr/fp.

Sauerbrei W. The Use of Resampling Methods to Simplify Regression Models in Medical Statistics. Appl Stat. 1999;48(3):313–29.

Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J R Stat Soc A Stat. 1999;162(1):71–94.

Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. Comput Stat Data Anal. 2006;50(12):3464–85.

Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med. 2007;26(30):5512–28.

Schwartz S, Campbell UB, Gatto NM, Gordon K. Toward a clarification of the taxonomy of "bias" in epidemiology textbooks. Epidemiology. 2015;26(2):216–22.

Shahar E. Shahar Responds to "Causal Diagrams and Measurement Bias". Am J Epidemiol. 2009a;170(8):963–4.

Shahar E. The Association of Body Mass Index With Health Outcomes: Causal, Inconsistent, or Confounded? Am J Epidemiol. 2009b;170(8):957–8.

Skov T, Deddens J, Petersen MR, Endahl L. Prevalence proportion ratios: estimation and hypothesis testing. Int J Epidemiol. 1998;27(1):91–5.

Slama R, Werwatz A. Controlling for continuous confounding factors: non- and semiparametric approaches. Rev Epidemiol Sante Publique. 2005;53(2S):2S65–2S80.

Smith PL. Splines as a Useful and Convenient Statistical Tool. Amer Statist. 1979;33(2):57–62.

Sobel ME. Modeling Symmetry, Asymmetry, and Change in Ordered Scales with Midpoints Using Adjacent Category logit Models for Discrete Data. Sociol Methods Res. 1997;26(2):213–32.

Sperrin M, Candlish J, Badrick E, Renehan A, Buchan I. Collider Bias Is Only a Partial Explanation for the Obesity Paradox. Epidemiology. 2016;27(4):525–30.

Sribney B. What are some of the problems with stepwise regression? 1998 [cited 24/04/2022]; Available from: https://www.stata.com/support/faqs/stat/stepwise.html.

Steyerberg EW, Eijkemans MJC, Habbema JDF. Stepwise Selection in Small Data Sets: A Simulation Study of Bias in Logistic Regression Analysis. J Clin Epidemiol. 1999;52(10):935–42.

Steyerberg EW, Harrell J, Frank E., Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774–81.

Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM. Internal and external validation of predictive models: A simulation study of bias and precision in small samples. J Clin Epidemiol. 2003;56(5):441–7.

Steyerberg EW, Eijkemans MJC, Boersma E, Habbema JDF. Equally valid models gave divergent predictions for mortality in acute myocardial infarction patients in a comparison of logical regression models. J Clin Epidemiol. 2005;58(4):383–90.

Steyerberg EW, Schemper M, Harrell FE. Logistic regression modeling and the number of events per variable: selection bias dominates. J Clin Epidemiol. 2011;64(12):1464–5.

Stone C. Comment : Generalized additive models. Stat Sci. 1986;1(3):312–4.

Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol. 2006;59(5):437.e1–e24.

Sullivan TR, Morris TP, Kahan BC, Cuthbert AR, Yelland LN. Categorisation of continuous covariates for stratified randomisation: How should we adjust? Stat Med. 2024;43(11):2083–95.

Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49(8):907–16.

Swartz MD, Yu RK, Shete S. Finding factors influencing risk: Comparing Bayesian stochastic search and standard variable selection methods applied to logistic regression models of cases and controls. Stat Med. 2008;27(29):6158–74.

Talbot D, Massamba VK. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. Eur J Epidemiol. 2019;34(8):725–30.

Talbot D, Diop A, Lavigne-Robichaud M, Brisson C. The change in estimate method for selecting confounders: A simulation study. Stat Methods Med Res. 2021;30(9):2032–44.

Taylor JMG, Yu M. Bias and Efficiency Loss Due to Categorizing an Explanatory Variable. J Multiv Anal. 2002;83(1):248–63.

Turner E, Dobson J, Pocock S. Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. Epidemiol Perspect Innov. 2010;7(1):9.

van Rijn MH, Bech A, Bouyer J, van den Brand JA. Statistical significance versus clinical relevance. Nephrol Dial Transplant. 2017;32(suppl_2):ii6–ii12.

van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, Reitsma JB. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16(1):163.

Viallon V. Régression pénalisée : le Lasso. 2015 [cited; Available from: http://pbil.univ-lyon1.fr/members/fpicard/franckpicard_fichiers/master/coursLasso.pdf.

Viallon V, Dufournet M. Can collider bias fully explain the obesity paradox? arXiv:161206547 [statME]. 2016.

Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 2007;165(6):710–8.

Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol. 2009;24(12):733–6.

Wang D, Zhang W, Bakhai A. Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. Stat Med. 2004;23(22):3451–67.

Wang Z. Two postestimation commands for assessing confounding effects in epidemiological studies. Stata J. 2007;7(2):183–96.

Wegman EJ, Wright IW. Splines in Statistics. J Am Stat Assoc. 1983;78(382):351–65.

Weinberg CR. How bad is categorization? Epidemiology. 1995;6(4):345–7.

Weiss NS, Daling JR, Chow WH. Control definition in case-control studies of ectopic pregnancy. Am J Public Health. 1985;75:67–8.

White IR. Commentary: Dealing with measurement error: multiple imputation or regression calibration? Int J Epidemiol. 2006;35(4):1081–2.

White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377–99.

Wikipédia. Moyenne mobile, Wikipédia, l'encyclopédie libre. 2023 [cited Page consultée le 18-avril-2023]; Available from: http://fr.wikipedia.org/w/index.php?title = Moyenne_mobile&oldid = 203212355.

Williams R. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. Stata J. 2006;6(1):58–82.

Zhang J, Yu KF. What's the Relative Risk?: A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. JAMA. 1998;280(19):1690–1.

Zhao LP, Kolonel LN. Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. Am J Epidemiol. 1992;136(4):464–74.

Zou G. A modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702–6.

Index

Les termes « régression logistique » et « modèle logistique » sont employés comme des synonymes dans cet Index et dans le livre de façon générale.

D

E

F

G

H

I

K

L

M

N

O

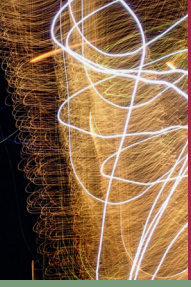
P

R

S

T

V



La régression logistique en épidémiologie

Jean Bouyer

La régression logistique est la méthode la plus utilisée en épidémiologie lorsque la maladie (Y) est caractérisée par une variable en 2 classes ou plus. Elle permet de rechercher et d'analyser des facteurs de risque de la maladie ou ses facteurs pronostics (X), quelle que soit leur nature, qualitative ou quantitative. Les logiciels d'analyse statistique rendent sa mise en œuvre très facile en pratique, mais n'évitent pas les risques de mésusages ni d'erreurs d'interprétation.

Ce livre détaille le modèle logistique pour caractériser et quantifier la relation entre Y et les variables X. Grâce à ses qualités pédagogiques, la présentation n'est pas réservée aux biostatisticiens, tout en donnant aux épidémiologistes les moyens nécessaires à une bonne compréhension.

L'intérêt et les moyens d'intégrer dans un modèle logistique des variables X quantitatives sans les transformer en classes sont largement développés. La modélisation des variables quantitatives par des polynômes fractionnaires ou des fonctions splines est détaillée et accompagnée de fonctions en Stata et en R destinées à présenter les résultats obtenus de façon compréhensible pour le lecteur.

Le choix des variables à inclure dans un modèle logistique, une des questions cruciales de l'analyse des enquêtes épidémiologiques, est longuement présenté et discuté. Cela permet d'aborder la plupart des notions et discussions rencontrées lors de l'analyse des enquêtes épidémiologiques.

Le contenu de ce livre est issu d'un cours intitulé «Épidémiologie quantitative» donné dans le cadre du Master 2 Recherche de Santé publique de la faculté de médecine de l'université Paris-Sud, devenue ensuite université Paris-Saclay. Il s'adresse aux personnes ayant une formation de base en statistique et en épidémiologie (Master 1 de Santé Publique ou CESAM par exemple).

***Jean Bouyer** est ancien élève de l'École normale supérieure de Saint-Cloud, épidémiologiste et biostatisticien, directeur de recherche émérite à l'Inserm. Il a enseigné pendant de nombreuses années au Master de Santé publique des universités Paris-Sud puis Paris-Saclay.*

978-2-7598-3817-2

9 782759 838172

Publié avec le soutien de la Graduate school Santé publique de l'Université Paris-Saclay

université PARIS-SACLAY | GRADUATE SCHOOL Santé publique

edp sciences
www.edpsciences.org