Étude de la confusion résiduelle et erreur de mesure dans des modèles de régression.

par

Mariem FOURATI

mémoire présenté au département de mathématiques en vue de l'obtention du grade de maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, 17 juillet 2015

Membres du jury

Professeur Taoufik Bouezmarni
Directeur de recherche
Département de mathématiques

Professeur Alan.A Cohen
Co-directeur de recherche
Département de médecine de famille et de médecine d'urgence
Faculté de médecine

Professeur Bernard Colin
Évaluateur interne
Département de mathématiques

Professeur Ernest Monga
Président-rapporteur
Département de mathématiques

À mon père

À ma mère

À mes soeurs

À mon époux

À mon fils Youssef

Remerciements

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma reconnaissance.

Je tiens tout d'abord à remercier Monsieur Taoufik Bouezmarni, mon directeur de recherche qui m'a accueilli à l'université de Sherbrooke et m'a offert sa confiance, je lui adresse toute ma gratitude pour sa patience et tous ses conseils judicieux.

Je remercie aussi Monsieur Alan Cohen, mon co-directeur de recherche pour sa disponiblité, son savoir-faire et son soutien.

Je les remercie tous les deux pour l'inspiration, l'aide et le temps qu'il ont bien voulu me consacrer, je leur en suis très reconnaissante, ils m'ont soutenu, encouragé malgré les moments difficiles et je leur exprime aussi toute ma gratitude pour leur compréhension et leur présence tout au long de ma maîtrise.

Je remercie mes très chers parents, qui ont toujours été là pour moi, ' Vous avez tout sacrifié pour vos enfants n'épargnant ni santé ni efforts. Vous m'avez donné un magnifique modèle de labeur et de persévérance. Je suis redevable d'une éducation dont je suis fière '.

Je remercie mon très cher époux Walid, mon compagnon, mon âme sœur, sans qui je n'aurais pas pu continuer, je le remercie pour son soutien, sa présence, tous ses sacrifices et son amour sans égal.

Je remercie également mes deux chères soeurs Fatma et Hejer pour leur soutien et leurs encouragements en tout temps.

Je tiens aussi à remercier, bien qu'il ne pourra pas encore en être conscient, mon fils Youssef, mon rayon de soleil et la prunelle de mes yeux parce qu'il a illuminé ma vie, m'a

donné aussi la force de vouloir avancer et aller de l'avant.
Je remercie tous les professeurs du département de mathématiques de l'université de Sherbrooke desquels j'ai beaucoup appris.
Je remercie également tout le personnel de la faculté des sciences de l'université de Sherbrooke.

Mariem Fourati
Sherbrooke, Mai 2015

Sommaire

Dans nombreuses situations dans les études épidémiologiques, on note la présence de certaines erreurs de mesure qui peuvent biaiser les résultats recherchés.
Un grand nombre de méthodes de correction de l'effet de ces erreurs a été développé mais en pratique elles ont été rarement appliquées, probablement à cause du fait que leur capacité de correction et leur mise en oeuvre sont peu maîtrisées. Ce mémoire s'intéresse aux erreurs de mesures présentes dans les variables de confusion et particulièrement dans les modèles de régression linéaire et logistique. La motivation principale réside dans le fait de trouver les paramètres les plus affectés par une telle confusion dans un modèle de régression et de trouver une méthode simple et efficace pour corriger le biais présent dans ce modèle.

Dans une première partie de ce mémoire, seront traitées les notions de régression linéaire et logistique et les méthodes d'estimation de leurs paramètres. Dans une deuxième partie, on va définir la notion d'erreur de mesure et de variables de confusion et les traiter dans leurs différents cas. On explorera aussi le cas du modèle linéaire dont l'une des variables est affectée par une erreur de mesure ou/et une confusion. On définira une méthode de correction dite l'étalonnage de régression. Dans une troisième partie, seront présentées les simulations faites dans le cadre de cette problématique et les interprétations et résultats

qui en découlent et nous définirons une méthode de correction du biais d'estimation des paramètres de la régression linéaire

Table des matières

LISTE DES FIGURES

Terminologie

Covariable: Une covariable, est une variable qui n'est pas d'un intérêt direct dans une étude de recherche, mais qui doit être prise en compte dans le cadre de la recherche, car elle a le potentiel d'influencer la variable de réponse dite aussi d'intérêt.

Variable de contrôle: Une variable de contrôle, est une variable ajoutée dans une régression dans le but d'éviter un biais dans l'estimation du paramètre d'intérêt. Par exemple, si vous vous intéressez à l'effet du taux de change euro/dollars sur la croissance économique en France, faire une régression avec simplement le taux de change euro/dollars en variable explicative (x) et la croissance économique en variable expliquée (y) a beaucoup de chance de s'avérer fallacieux : le paramètre reliant les deux variables ne sera pas mesuré correctement car d'autres variables expliquant la croissance économique ne sont pas spécifiées dans la régression. Les spécifier (on dit aussi "contrôler par d'autres variables") permettra d'éviter un biais dans l'estimation du paramètre d'intérêt, celui reliant la croissance au taux de change.

Variable endogène : Une variable endogène, est une variable qui apparaît comme une variable dépendante dans au moins une équation du modèle structurel.

Variable exogène : Une variable exogène, est une variable qui n'apparaît jamais en tant que variable dépendante dans les équations d'un modèle de régression.

Cohorte : Une cohorte désigne un ensemble d'individus ayant vécu un même événement au cours d'une même période.

Variable dépendante : Une variable dépendante, dite aussi variable de réponse ou d'intérêt est la variable que l'on cherche à expliquer en fonction d'autres variables. La variable dépendante correspond souvent à la variable endogène.

Variable indépendante : Une variable indépendante sert à expliquer la variable dépendante et correspond le plus souvent à la variable exogène.

Odd : L'odd, également appelé rapport des fréquences, est une mesure statistique, souvent utilisée en épidémiologie, exprimant le degré de dépendance entre des variables aléatoires qualitatives. Il est utilisé en régression logistique, et permet de mesurer l'effet d'un facteur.

Biais : En statistique, un biais est une erreur systématique entre une estimation et la véritable valeur du paramètre estimé. Ainsi, le biais est une erreur qui se reproduit à l'identique (systématique) et qui, contrairement aux erreurs aléatoires, ne se compense pas en moyenne

Erreur de type I : Il se produit une erreur de type I lorsque le test statistique d'hypothèse mène à rejeter une hypothèse nulle alors que celle ci est vraie.

Erreur de type II : Il se produit une erreur de type II lorsqu'on omet de accepter une hypothèse nulle et que celle-ci est en réalité fausse.

Intercept : L'intercept correspond à l'ordonnée à l'origine.

CHAPITRE 1

Introduction

Au 20^e siècle, les avancées médicales se sont succédées sur tous les fronts : biologie, chimie, physiologie, pharmacologie et technologies. Elles ont d'ailleurs souvent profité à plusieurs branches de la médecine. Grâce à une meilleure compréhension des maladies, on a souvent pu élaborer de nouveaux traitements et remèdes. L'espérance de vie s'est allongée dans la plupart des régions du monde, avec pour conséquence l'augmentation des maladies liées au vieillissement, en particulier, les cancers et les maladies cardiovasculaires. La médecine s'est donc attelée au traitement et à la prévention de ces maladies. D'après Rothman (2002), l'épidémiologie, est une discipline scientifique qui étudie la fréquence des maladies, leur répartition dans la société, les facteurs de risque et les décès liés à cette maladie. Ces informations sont indispensables pour la médecine préventive. En épidémiologie, comme dans plusieurs domaines d'étude, on cherche les causes des maladies, et ce à partir des causes qui peuvent modifier le risque de ces maladies.

Dans les recherches épidémiologiques, la maladie ou un évènement lié à la santé, est considéré comme une variable de résultat. D'après les travaux de Axelson (1985), cette variable peut être facilement définie contrairement à la variable d'exposition, qui, considérée aussi comme un facteur de risque, est plus complexe et plus variable. Une variable d'exposition peut être associée soit à une augmentation ou à une diminution de l'apparition de la maladie ou à d'autres effets sur la santé. Elle peut se rapporter

à l'environnement, à un mode de vie ou à des caractéristiques innées ou héréditaires. Généralement, les études sont réalisées sur les gens avec toutes les contraintes pratiques et éthiques qui peuvent en découler et donc sont presque toujours sujettes à des biais. Et il est malheureusement toujours possible de commettre des erreurs. Par exemple, si nous menons une comparaison entre deux médicaments, nous pouvons conclure qu'il existe une différence entre les deux alors qu'en réalité, il n'y en a pas, cette erreur est appelée une erreur de Type I. Certaines erreurs peuvent à tort augmenter ou diminuer l'ampleur d'une association entre deux variables ou même inverser la direction de l'association. On parle dans ce cas des facteurs de confusion.

On peut donc définir la variable de confusion comme une variable qui modifie l'association entre la variable d'exposition et la variable de résultat.

La confusion telle que définie dans Lindsay (2014) se produit lorsqu'un facteur différent de celui que l'on étudie est associé avec la maladie et le facteur à l'étude. Cela peut porter à croire que le facteur étudié est une cause de la maladie, alors qu'en réalité, ce n'est pas le cas.

Rothman (2002), a défini la confusion en disant "Confounding is confusion, or mixing, of effects; the effect of the exposure is mixed together with the effect of another variable, leading to bias".

Par exemple, des épidémiologistes ont trouvé que boire du café, est associé aux maladies cardiovasculaires mais la question qui a été posée par la suite est : Est-ce que boire du café est réellement une cause de la cardiopathie? La réponse est qu'il existe peut-être un autre facteur derrière cette causalité, comme le tabagisme, i.e, et si boire du café est associé au tabagisme (peut-être que les gens qui boivent du café ont également tendance à fumer), le tabagisme constitue une variable confusionnelle parce que nous savons qu'il influe de façon distincte sur la cardiopathie comme l'illustre le cas de Lindsay (2014) ci-dessous.

Exemple et Mise en situation

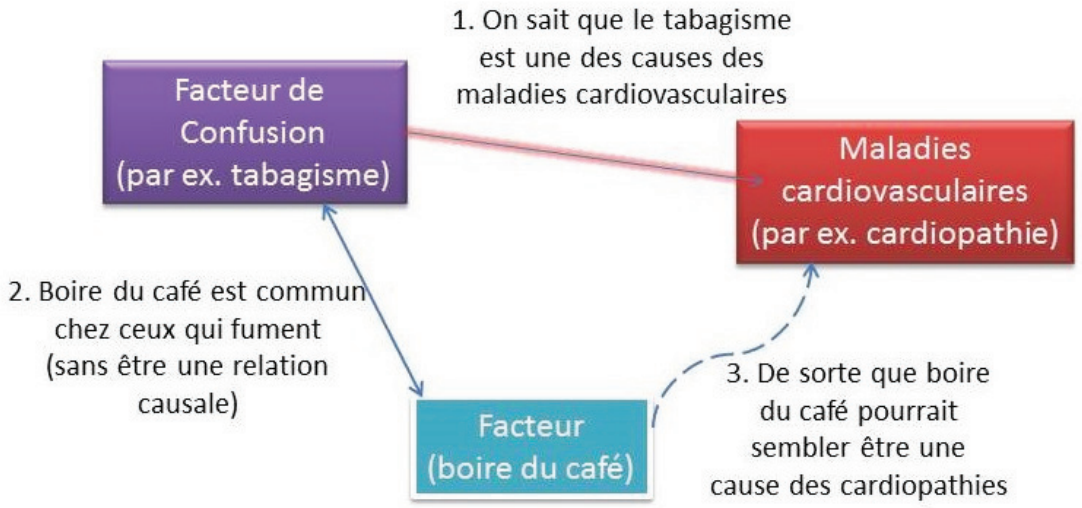

Figure 3.1: Exemple illustrant la présence d'un facteur de confusion

D'après cette figure, il semblerait avoir une association entre la cardiopathie et boire du café, alors qu'en réalité, ce lien peut plutôt être attribuable au fait que les personnes qui boivent du café fument également et que leur cardiopathie découle de leur habitude de tabagisme, non pas de leur consommation de café.

Ce genre d'erreurs de mesure dans les facteurs de confusion peut mener à une confusion résiduelle. La confusion résiduelle peut être causée soit par un manque de contrôle de la confusion, soit par un manque de précision des données sur la variable de confusion ou par la présence d'erreurs dans le classement des sujets par rapport aux variables de confusion.

Afin de traiter une erreur de mesure dans les variables de confusion, plusieurs méthodes sont utilisées. Et certains ont recours à la régression, qui est considérée la plus utilisée de toutes les techniques statistiques.

Barnwell et al. (2014), affirment que les erreurs dues à la catégorisation ou tout autre

mesure imparfaite d'une variable continue sont bien connues, et des études précédentes ont démontré que cela peut conduire à une inflation de l'erreur de type I lorsque cette variable est un facteur de confusion dans une analyse de régression. Cependant, on ne sait pas comment l'erreur de type I peut varier dans des circonstances différentes, y compris en variant plusieurs paramètres dans une régression linéaire ou une régression logistique. Dans le cadre des recherches faites auparavant par Barnwell et al. (2014) au sein de Groupe PRIMUS dans le département de médecine et en collaboration avec le département de mathématique de l'université de Sherbrooke, s'introduit le travail de ce rapport de mémoire de maîtrise. Dans ce travail, nous avons étudié l'analyse de régression multiple comme méthode de traitement des facteurs de confusion, qui peut servir à déterminer les effets d'une erreur de mesure dans une variable de confusion. Dans ce contexte, la régression multiple a permis d'estimer l'association entre une variable indépendante donnée et le résultat tout en laissant constantes toutes les autres variables. Ce qui nous a permis par la suite d'estimer le biais de confusion dans le cas d'une régression linéaire en premier lieu, et dans le cas d'une régression logistique en second lieu. Une correction de ce biais dans un modèle de régression linéaire a été possible par la suite.

CHAPITRE 2

Estimation de la fonction de régression

Au cours de ce chapitre, sera définie en première partie, la régression linéaire suivie de l'estimation de ses paramètres par la méthode des moindres carrés. Une description de la régression logistique et l'estimation de ses paramètres par la méthode de vraisemblance feront l'objet d'une seconde partie

2.1 La régression linéaire

2.1.1 Description

Comme a été défini dans Worster et al. (2007), l'analyse de régression, également appelée modèle de régression, est une méthode statistique de plus en plus commune utilisée pour décrire et quantifier la relation entre un résultat d'intérêt et une ou plusieurs autres variables.

Dans Selme & Masson (2013), on dit que les sciences exactes sont fondées sur la notion de relations répétables, qui peut s'énoncer ainsi : dans les mêmes conditions, les mêmes

causes produisent les mêmes effets.

En statistiques, un modèle de régression linéaire est un modèle qui relie la variable de réponse (dépendante) à une ou plusieurs variables explicatives (indépendantes) dans lequel on fait l'hypothèse que la fonction qui relie les variables explicatives à la variable expliquée est linéaire dans ses paramètres. La régression linéaire multiple est une analyse statistique qui décrit les variations de la variable d'intérêt en fonction de plusieurs variables explicatives. Par exemple, une analyse de régression multiple peut révéler une relation positive entre la demande de lunettes de soleil et différents caractéristiques démographiques (âge, salaire, etc) des acheteurs de ce produit. La demande augmente ou baisse avec les variations de ces caractéristiques.

2.1.2 Spécification du modèle

Le but de l'analyse de régression linéaire multiple consiste à expliquer les fluctuations d'une variable dépendante Y par une combinaison linéaire de certaines variables $X_1, ..., X_p$, appelées variables explicatives ou indépendantes. Etant donné un échantillon $(Y_i, X_{i,1}, ..., X_{i,p})$ pour $i = 1, ..., n$, où $\beta_0, \cdots, \beta_p$ sont les paramètres à estimer et $\boldsymbol{\epsilon}$ est un bruit aléatoire représentant le terme d'erreur du modèle qui résume l'information manquante dans l'explication linéaire des valeurs de Y_i à partir des $X_{i,1}, \cdots, X_{i,p}$. Le modèle de régression multiple, prend la forme suivante :

$$Y_i = \beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p} + \epsilon_i, \quad i = 1, \cdots, n \tag{2.1}$$

Pour étudier ce modèle, on suppose les hypothèses suivantes :

Hypothèses du modèle

$\boldsymbol{H_1}$: Les variables X_i sont aléatoires ou fixes.

$\boldsymbol{H_2}$: $\mathbb{E}(\epsilon_i) = 0$ pour tout i.

$\boldsymbol{H_3}$: $\boldsymbol{var}(\epsilon_i) = \sigma^2$ pour tout i (homoscédasticité des erreurs).

$\boldsymbol{H_4}$: $\boldsymbol{cov}(\epsilon_i, \epsilon_j) = 0$ pour tout $i \neq j$.

$\boldsymbol{H_5}$: X_i et ϵ_i sont indépendantes.

$\boldsymbol{H_6}$: Pour l'inférence, le vecteur aléatoire $\epsilon = (\epsilon_i)_{i=1,..,n}$ suit une loi normale $N(0, \sigma^2 I_n)$.

On suppose que le modèle de régression linéaire peut être réécrit de la forme :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

où :

- $\boldsymbol{Y} = (y_1, \cdots, y_n)$ est un vecteur aléatoire de réponse de dimension n,
- $\boldsymbol{X}$ est une matrice de taille $n \times p$ connue, appelée matrice du plan d'expérience composée des lignes $(1, X_{i,1}, ..., X_{i,p})$ pour $i = 1, .., n$,
- $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ est le vecteur de dimension $p+1$ des paramètres inconnus du modèle,
- $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ est le vecteur de dimension n des erreurs..

Le problème à traiter par la suite consiste à résoudre :

$$\min_{\beta} \mathbb{E}[(Y - X\beta)^2] = \min_{\beta} \mathbb{E}[(Y - X\beta)^T (Y - X\beta)]$$

2.1.3 Estimation des paramètres par la méthode des moindres carrés

Comme dans Palm & Iemma (1995), on procède dans cette partie à l'estimation des paramètres $\beta_0, \cdots, \beta_p$ par la méthode des moindres carrés ordinaires. Les estimateurs par maximum de vraisemblance sont les mêmes que les moindres carrés sous l'hypothèse $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$. Les estimateurs de moindres carrés $\hat{\beta}_0, \cdots, \hat{\beta}_p$ associés aux paramètres $\beta_0, \cdots, \beta_p$ sont obtenus en minimisant l'erreur quadratique moyenne relative aux termes résiduels du modèle. Ce qui nous ramène à résoudre le problème d'optimisation suivant :

$$\min_{\boldsymbol{\beta}} \sum_{i=0}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_p x_{i,p})\right)^2 .$$

Pour ce faire, posons :

$$L(\boldsymbol{\beta}) = \sum_{i=0}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_p x_{i,p}))^2.$$

Les estimateurs $\hat{\beta}_0, \cdots, \hat{\beta}_p$ sont ainsi obtenus en résolvant le système :

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0.$$

$$\vdots$$

$$\left.\frac{\partial L}{\partial \beta_p}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0$$

Ce qui est équivalent au système linéaire suivant :

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1} + ... + \hat{\beta}_p \sum_{i=1}^{n} x_{i,p} &= \sum_{i=1}^{n} y_i, \\ \hat{\beta}_0 \sum_{i=1}^{n} x_{i,1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1}^2 + ... + \hat{\beta}_p \sum_{i=1}^{n} x_{i,1}x_{i,p} &= \sum_{i=1}^{n} x_{i,1}y_i, \end{aligned}$$

$$\vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i,p} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1}x_{i,p} + ... + \hat{\beta}_p \sum_{i=1}^{n} x_{i,p}^2 = \sum_{i=1}^{n} x_{i,p}y_i.$$

Ce qui peut s'écrire sous la forme :

$$\begin{pmatrix} n & \sum_{i=1}^n x_{i,1} & \cdots & \sum_{i=1}^n x_{i,p} \\ \sum_{i=1}^n x_{i,1} & \sum_{i=1}^n x_{i,1}^2 & \cdots & \sum_{i=1}^n x_{i,1}x_{i,p} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{i,p} & \sum_{i=1}^n x_{i,1}x_{i,p} & \cdots & \sum_{i=1}^n x_{i,p}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i,1}y_i \\ \vdots \\ \sum_{i=1}^n x_{i,p}y_i \end{pmatrix}$$

En remarquant que :

$$\begin{pmatrix} n & \sum_{i=1}^n x_{i,1} & \cdots & \sum_{i=1}^n x_{i,p} \\ \sum_{i=1}^n x_{i,1} & \sum_{i=1}^n x_{i,1}^2 & \cdots & \sum_{i=1}^n x_{i,1}x_{i,p} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{i,p} & \sum_{i=1}^n x_{i,1}x_{i,p} & \cdots & \sum_{i=1}^n x_{i,p}^2 \end{pmatrix} = \boldsymbol{X^T X}$$

et que

$$\begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i,1} y_i \\ \vdots \\ \sum_{i=1}^{n} x_{i,p} y_i \end{pmatrix} = \boldsymbol{X^T Y},$$

l'estimateur $\hat{\beta} = (\hat{\beta}_0, \cdots, \hat{\beta}_p)^T$ est obtenu en résolvant le système suivant :

$$\boldsymbol{X^T X \hat{\beta}} = \boldsymbol{X^T Y}.$$

Et par suite, si $\boldsymbol{X}^T\boldsymbol{X}$ est inversible, on obtient :

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}.$$

2.1.4 Propriétés des estimateurs des moindres carrés ordinaires

Les propriétés des estimateurs par la méthode des moindres carrés ont été largement étudiées comme dans Cornillon & Matzner-Lober (2007). D'après les hypothèses considérées au départ de H_1 à H_6, on peut établir certaines propriétés statistiques des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$.

Proposition : L'estimateur $\hat{\beta}$ est sans biais de matrice de variance covariance que l'on écrit au général sous la forme :

$$\Sigma_{\hat{\boldsymbol{\beta}}} = \sigma^2 (X^T X)^{-1}.$$

Preuve : En effet,

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}((X^TX)^{-1}X^TY) \\
&= (X^TX)^{-1}X^T\mathbb{E}(Y) \\
&= (X^TX)^{-1}X^T\mathbb{E}(X\beta+\epsilon) \\
&= (X^TX)^{-1}X^TX\beta \\
&= \beta.
\end{aligned}$$

Donc, l'estimateur $\hat{\beta}$ de β est sans biais.

Pour la matrice de variance covariance, on a

$$\begin{aligned}
\boldsymbol{var}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{var}((X^TX)^{-1}X^TY) \\
&= ((X^TX)^{-1}X^T\boldsymbol{var}(Y)[(X^TX)^{-1}X^TY]^T) \\
&= (X^TX)^{-1}X^TX(X^TX)^{-1}\sigma^2I \\
&= \sigma^2(X^TX)^{-1}.
\end{aligned}$$

■

Sous les hypothèses de H_1 à H_6, on a :

$$\hat{\beta} \sim N_k(\beta, \sigma^2(X^TX)^{-1}).$$

Finalement,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (X^TX)^{-1}X^T(X\beta+\epsilon) \\
&= \beta + (X'X)^{-1}X^T\epsilon \sim N(\beta; \sigma^2(X^TX)^{-1})
\end{aligned}$$

Exemple : Soit le modèle de régression linéaire simple suivant

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

On a :

$$\boldsymbol{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\boldsymbol{var}(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

et

$$\boldsymbol{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^{n}}(x_i - \bar{x})^2.$$

Remarques

- Plus σ^2 est petit, plus $\boldsymbol{var}(\hat{\boldsymbol{\beta}}_{\mathbf{0}})$ et $\boldsymbol{var}(\hat{\boldsymbol{\beta}}_{\mathbf{1}})$ diminuent.
- Plus x_i est dispersé, plus les variances deviennent petites.
- Si $\bar{x} \geq 0$ alors $\hat{\beta}_0$ et $\hat{\beta}_1$ sont corrélés négativement.

2.2 La régression logistique

2.2.1 Description

D'après Cramer (2002), la régression logistique a été proposée au 19^e siècle pour la description de la croissance des populations et le cours des réactions chimiques autocatalytiques. On définit dans Wiki.stat (2013), d'un point de vue historique, la régression logistique ou régression binomiale comme la première méthode utilisée, notamment en

marketing pour le "scoring" et en épidémiologie, pour aborder la modélisation d'une variable binaire binomiale (nombre de succès pour n_i essais) ou de Bernoulli (avec $n_i = 1$). Par exemple, possession ou non d'un produit, bon ou mauvais client, décès ou survie d'un patient, absence ou présence d'une pathologie, etc.

Bien connue dans ces types d'applications et largement répandue, la régression logistique conduit à des interprétations pouvant être complexes mais rentrées dans les usages pour quantifier, par exemple, des facteurs de risque liés à une pathologie, une faillite, etc. Cette méthode reste donc la plus utilisée même si, en terme de qualité prévisionnelle, d'autres approches sont susceptibles, en fonction des données étudiées, d'apporter de bien meilleurs résultats. Il est donc important de bien maîtriser les différents aspects de la régression logistique comme l'interprétation des paramètres, la sélection de modèle par sélection de variables ou par régularisation (pour plus de détails voir Tibshirani (1996)).

D'un point de vue statistique, la régression logistique est une technique explicative et prédictive en même temps. D'un côté, elle vise à construire un modèle permettant d'expliquer les valeurs prises par une variable cible qualitative (le plus souvent binaire, on parle alors de régression logistique binaire et si elle possède plus de 2 modalités, on parle de régression logistique polytomique) à partir d'un ensemble de variables explicatives quantitatives ou qualitatives (un codage est nécessaire dans ce cas). D'un autre côté, selon Desjardins (2005), cette technique permet de connaître les facteurs associés à un phénomène en élaborant un modèle de prédiction. La popularité de cette méthode est bien connue dans les sciences de la santé et en sciences humaines, où la variable à prédire est la présence ou l'absence d'une maladie. Par exemple, il peut s'agir d'une étude sur la dépression majeure où l'on désire connaître les facteurs la prédisant le mieux possible, en étudiant des variables telles que l'âge, le sexe, l'estime de soi, les relations interpersonnelles et, comme cité dans Tabachnick & Fidell (2000), la régression logistique a pour objectifs de chercher à expliquer la survenue d'un évènement et de calculer la probabilité de succès.

2.2.2 Spécification du modèle

Les recherches dans Preux et al. (2014) montrent que l'intérêt majeur du modèle de la régression logistique est de quantifier la force de l'association entre chaque variable indépendante et la variable dépendante, en tenant compte de l'effet des autres variables intégrées dans le modèle.
Dans cette section, on supposera que la variable Z à laquelle on s'intéresse est une variable qualitative à 2 modalités : 1 ou 0, succès ou échec, présence ou absence de maladie, ect. On désigne par π la probabilité tel que :

$$\pi = \mathbb{P}(Z = 1) \quad ou \quad 1 - \pi = \mathbb{P}(Z = 0).$$

On considère par la suite pour $i = 1, .., N$, différentes valeurs fixées $x_{i,1}, .., x_{i,K}$ des variables explicatives $X_1, .., X_K$ éléments de la matrice X composé de N lignes et de $K + 1$ colonnes et l'on désigne par $\boldsymbol{\beta}$ un vecteur paramètre de longueur K .
Pour chaque groupe, c'est-à-dire pour chacune des combinaisons de valeurs ou facteurs, on réalise n_i observations ($n = \sum_{i=1}^{N} n_i$) de la variable Z où les y_i sont les éléments du vecteur y et qui désignent le nombre de "succès" observés lors des n_i essais.

La régression logistique assimilée à la transformation *logit* se définit comme suit :

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \sum_{k=0}^{K} x_{i,k}\beta_k, \quad i = 1, ..., N. \tag{2.2}$$

Odd-Ratio ou Rapport des cotes

Dans le but de pouvoir discuter et interpréter les résultats d'une régression logistique, Christensen (1997) a mentionné qu'il est important de définir le terme "Odd".
Supposons qu'un évènement A ait une probabilité π de se produire. Le rapport des cotes

est défini comme le rapport de la probabilité que A se produise par la probabilité que A ne se produise pas, i.e.

$$Odds(A) = \frac{\mathbb{P}(A)}{\mathbb{P}(A^c)} = \frac{\pi}{1-\pi},$$

où A^c représente le complémentaire de A.
L'Odds ratio est défini comme le rapport des cotes tel que :

$$OR = \frac{Odds(A)}{Odds(B)}$$

2.2.3 Estimation par la méthode du maximum de vraisemblance

La mise en oeuvre de la régression logistique est l'estimation des $(K+1)$ paramètres inconnus $\boldsymbol{\beta}$ dans (2.2). Ceci se fait à travers la maximisation de la vraisemblance. Czepiel (2010) a établi l'estimateur par la méthode du maximum de vraisemblance du modèle logistique. Cette méthode consiste en premier lieu, à définir la fonction de vraisemblance, soit la fonction de probabilité conjointe, celle ci est obtenue à partir de la fonction de probabilité de chaque observation individuellement en considérant l'hypothèse que les observations sont indépendantes. La fonction de densité de la probabilité jointe de Y s'écrit comme suit :

$$f(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i-y_i)!} \pi_i^{y_i} (1-\pi_i)^{n_i-y_i}. \tag{2.3}$$

Sachant que π_i représente la probabilité de succès. La fonction de maximisation de vraisemblance s'écrit alors :

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i-y_i)!} \pi_i^{y_i} (1-\pi_i)^{n_i-y_i}. \tag{2.4}$$

Les estimateurs du maximum de vraisemblance sont les valeurs de $\boldsymbol{\beta}$ qui maximisent la fonction de vraisemblance dans (2.4). Donc, on va chercher cette équation qui est proportionnelle à :

$$\prod_{i=1}^{N} \left(\frac{\pi_i}{1-\pi_i} \right)^{y_i} (1-\pi_i)^{n_i}. \tag{2.5}$$

En tenant compte de 2.2, on obtient

$$\frac{\pi_i}{1-\pi_i} = \exp\left(\sum_{k=0}^{K} x_{ik}\beta_k \right) \tag{2.6}$$

Alors,

$$\pi_i = \frac{\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k\right)}{1+\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k\right)}$$

On peut donc en conclure que (2.5) devient :

$$\prod_{i=1}^{N} \exp\left(\sum_{k=0}^{K} x_{ik}\beta_k \right)^{y_i} \left(1 - \frac{\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k\right)}{1+\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k\right)} \right)^{n_i}.$$

En remplacant 1 par $\frac{1+\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k\right)}{1+\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k\right)}$,

$$\prod_{i=1}^{N} \exp\left(\sum_{k=0}^{K} x_{ik}\beta_k \right)^{y_i} \left(1+\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k \right) \right)^{-n_i}.$$

Par la suite, en prenant le logarithme, on obtient :

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[y_i \left(\sum_{k=0}^{K} x_{ik}\beta_k \right) - n_i . \log\left(1+\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k \right) \right) \right].$$

L'étape suivante consiste à calculer la dérivée de la log-vraisemblance pour chaque point,

$$\begin{aligned} \frac{\partial(\beta)}{\partial\beta_k} &= \sum_{i=1}^{N} \left[y_i x_{ik} - n_i \left(\frac{1}{1+\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k\right)} \right) \frac{\partial}{\partial\beta_k} \left(1+\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k \right) \right) \right] \\ &= \sum_{i=1}^{N} \left(y_i x_{ik} - n_i \pi_i x_{ik} \right). \end{aligned} \tag{2.7}$$

Les estimateurs du maximum de vraisemblance pour $\boldsymbol{\beta}$ peuvent être trouvés en fixant chacune des $(K+1)$ équations égale à 0 dans (2.7) et le point critique sera maximum si la matrice hessienne est définie négative. Cette dernière est de la forme

$$\begin{aligned}\frac{\partial^2 l(\beta)}{\partial\beta_k\partial\beta_{k'}} &= \frac{\partial}{\partial\beta_{k'}}\sum_{i=1}^{N}(y_i x_{ik} - n_i x_{ik}\pi_i) \\ &= -\sum_{i=1}^{N} n_i x_{ik}\frac{\partial}{\partial\beta_{k'}}\left(\frac{\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k\right)}{1+\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k\right)}\right) \\ &= -\sum_{i=1}^{N} n_i x_{ik}\pi_i(1-\pi_i)x_{ik'}.\end{aligned}$$

2.2.4 Propriétés des estimateurs du maximum de vraisemblance

Un estimateur du maximum de vraisemblance est un estimateur possédant des propriétés distinctes telles que citées dans Russel & Mackinnon (1993).

Propriétés asymptotiques de l'estimateur

Si $\hat{\beta}$ est l'estimateur de β alors on dit que $\hat{\beta}$ est consistant s'il converge vers sa vraie valeur β_0 quand le nombre d'observations $n \to \infty$,
En définissant D comme la matrice diagonale de format telle que

$$d_{ij} = \begin{cases} \pi_i \quad si \quad i = j \\ 0 \quad sinon. \end{cases}$$

Et en supposant que les hypothèses suivantes définies dans Gourieroux & Monfort (1981) et Amemiya (1985) soient vérifiées :

- H1 : Les variables explicatives sont uniformément bornées, i.e, il existe $C \leq \infty$ tel que $|| x ||\leq C$.

- H2 : Soient λ_{1n} et λ_{2n} les valeurs propres respectivement minimale et maximale de la matrice $X^T D(\beta_0)X$. Alors, il existe une constante $K \leq \infty$ telle que $\frac{\lambda_{pn}}{\lambda_{1n}} \leq K$, pour tout n.

Théorème (Normalité assymptotique)
Sous les hypothèses H1 et H2, et si $\hat{\beta}$ est consistant, alors

$$\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, \tau(\beta_0)^{-1}), \quad quand \quad n \to \infty,$$

où

$$\tau(\beta) = -\mathbb{E}\left[\frac{\partial^2 \log(\beta)}{\partial\beta\partial\beta^T}\right]$$

est la matrice d'information de Fisher.

CHAPITRE 3

Confusion résiduelle et erreur de mesure

Dans ce chapitre, nous allons définir la confusion et quelques types de facteurs de confusion. Par la suite, on traitera le cas de modèles de régression avec une confusion. Dans la dernière partie, seront définies la confusion résiduelle et l'erreur de mesure et quelques modèles de régression avec une variable affectée par une erreur de mesure ou un facteur de confusion.

3.1 Problème de Confusion en épidémiologie

3.1.1 Définition de la confusion

Selon Huff (1993), la confusion est définie comme étant une situation dans laquelle les effets de deux processus ne sont pas séparés. Le mot est d'origine latine 'confundere' et qui signifie "mélanger ensemble".
En statistique, une variable de confusion est définie dans Lindsay (2014) comme une variable étrangère dans un modèle statistique qui est corrèlée (directement ou inversement) à

la fois avec la variable dépendante et la variable indépendante qui sont respectivement la variable d'exposition et la variable de résultat. Cela peut entraîner le chercheur à analyser les résultats d'une étude d'une manière incorrecte. Ces résultats peuvent montrer une fausse corrélation entre les variables dépendantes et indépendantes.

3.1.2 Quelques types de facteurs de confusion

Facteurs de confusion connus

Supposons maintenant que nous soyons principalement intéressés à la relation entre une variable particulière X et un résultat Y. Un facteur de confusion mesuré est une variable Z qui peut être mesurée et inclue dans un modèle de régression avec X. Un facteur de confusion mesuré ne pose généralement pas un problème pour estimer l'effet de X, sauf s'il est fortement colinéaire avec X. Par exemple, supposons que nous étudions les effets de la santé de l'exposition au tabagisme passif. Nous mesurons la variable réponse de santé Y directement. Les sujets qui fument sont bien sûr concernés par un bon nombre des mêmes risques qui peuvent être associés à l'exposition de la fumée secondaire. Ainsi, il serait très important de déterminer qui est soumis à la fumée, et inclure cette information comme une covariable (un facteur de confusion mesuré) dans un modèle de régression utilisé pour évaluer les effets de l'exposition de la fumée secondaire.

Facteurs de confusion inconnus

Un facteur de confusion inconnu est une variable qui affecte le résultat d'intérêt, mais qui nous est inconnu.

Par exemple, il peut y avoir des facteurs génétiques qui produisent des effets similaires à ceux des effets du tabagisme passif mais dans ce cas, nous n'avons pas connaissance des gènes spécifiques qui sont impliqués.

Facteurs de confusion non mesurés

Un facteur de confusion non mesuré est une variable dont on connaît l'existence et dont on a accès à certaines de ses propriétés statistiques, mais c'est une variable qui n'est pas mesurée. Par exemple, nous pouvons savoir que certaines professions (comme travailler dans certains types d'usines) peuvent produire des risques similaires aux risques d'exposition au tabagisme passif. Si toutes les données ne sont pas collectées dans une étude particulière, ceci est un facteur de confusion non mesuré. Puisque nous ne disposons pas de données pour les facteurs de confusion non mesurés, leur omission peut produire des biais dans les effets estimés pour les variables d'intérêt. Si nous avons une certaine compréhension de la façon dont un certain facteur de confusion non mesuré fonctionne, nous pourrions être en mesure d'utiliser une analyse de sensibilité pour obtenir une idée approximative de combien le biais est présent.

3.2 Modèle de régression avec une variable de confusion

3.2.1 Modèle de régression linéaire

Comme dans Shedden (2014), on considère un modèle généré comme suit:

$$Y = \alpha + \beta X + \gamma Z + \epsilon$$

où X et Z sont des variables aléatoires corrélées tels que: $\boldsymbol{var}(X) = 1$ et $\mathbb{E}(X) = 0$.
L'erreur ϵ du modèle est telles que $\mathbb{E}(\epsilon) = 0$ et $\boldsymbol{var}(\epsilon) = \sigma^2$ et Z et ϵ sont indépendantes. On suppose ici que seuls X et Y sont observées. La variable Z de moyenne 0 et de variance égale à 1 qui est associée avec les deux variables, dépendante et indépendante dans un modèle de régression est appelée "variable de confusion". Supposons que X et Z soient normalisées, et que $\boldsymbol{cor}(X, Z) = r$. Supposons en outre que $\mathbb{E}[Z|X] = rX$. En raison de la linéarité de $\mathbb{E}[Y|X, Z]$:
($\boldsymbol{i}$) Si X augmente d'une unité et Z reste fixe, la réponse attendue Y augmente de β unités.
($\boldsymbol{j}$) Si Z augmente d'une unité et X reste fixe, la réponse attendue Y augmente de γ unités.
Toutefois, si nous choisissons le cas présentant des valeurs de X différentes par une unité au hasard (sans contrôler Z), les valeurs de Z vont différer en moyenne de r unités. Par conséquent, Y pour ces cas diffère de $\beta + r\gamma$ unités.

L'estimateur $\hat{\beta}$ des moindres carrés ordinaires s'écrit sous la forme

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_i Y_i(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \\
&= \frac{\sum_i (\alpha + \beta X_i + \gamma Z_i + \epsilon_i)(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \\
&= \frac{\alpha \sum_i (X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} + \beta \frac{\sum_i X_i(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} + \gamma \frac{\sum_i Z_i(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} + \frac{\sum_i \epsilon_i(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \\
&= \beta \frac{\sum_i (X_i - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} + \gamma \frac{\sum_i (Z_i - \bar{Z})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} + \frac{\sum_i (\epsilon_i - \bar{\epsilon})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}
\end{aligned}$$

Or Or d'après le théorème de Slutsky:

$$\frac{\sum_i (Z_i - \bar{Z})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \xrightarrow{P} \frac{\boldsymbol{cov}(X, Z)}{var(X)} = r$$

et vu l'indépendance entre Z et ϵ

$$\frac{\sum_i (\epsilon_i - \bar{\epsilon})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \xrightarrow{P} \frac{\boldsymbol{cov}(\epsilon, Z)}{\boldsymbol{var}(X)} = 0$$

D'où

$$\hat{\beta} \xrightarrow{P} \beta + \gamma r.$$

Notons que si $\gamma = 0$ ou bien si $r = 0$ alors β est correctement estimé.

Par la suite, on va s'intéresser à la limite en probabilité de α :

$$\begin{aligned}
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\
&= \alpha + \beta\bar{x} + \gamma\bar{z} - \hat{\beta}\bar{x} \\
&= \alpha - (\hat{\beta} - \beta)\bar{x} + \gamma\bar{z} \\
&= \alpha - (\hat{\beta} - \beta - \gamma r)\bar{x} - \gamma r\bar{x} + \gamma\bar{z} \\
&= \alpha - (\hat{\beta} - \beta - \gamma r)\bar{x} + \gamma(\bar{z} - r\bar{x})
\end{aligned}$$

Or $\hat{\beta} - \beta - \gamma r \xrightarrow{P} 0$ et $\bar{z} - r\bar{x} \xrightarrow{P} \mathbb{E}(Z) - r\mathbb{E}(X) = 0$ puisque

$$\begin{aligned}
\mathbb{E}(Z) &= \mathbb{E}(\mathbb{E}(Z|X)) \\
&= \mathbb{E}(rX) \\
&= r\mathbb{E}(X) = 0.
\end{aligned}$$

Et par la suite

$$\hat{\alpha} \xrightarrow{P} \alpha.$$

Variance du modèle

La variance du modèle est donnée par:

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2}\sum_i (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 \\
&= \frac{1}{n-2}\sum_i (\alpha + \beta X_i + \gamma Z_i + \epsilon_i - \hat{\alpha} - \hat{\beta}X_i)^2 \\
&= \frac{1}{n-2}\sum_i ((\alpha - \hat{\alpha}) + (\beta - \hat{\beta})X_i + \gamma Z_i + \epsilon_i)^2 \\
&= \frac{1}{n-2}\sum_i (-\gamma r X_i + \gamma Z_i + \epsilon_i)^2 \\
&\approx \frac{1}{n}\sum_i (\gamma(Z_i - rX_i) + \epsilon_i)^2 \\
&= \gamma^2 \frac{1}{n}\sum_i (Z_i - rX_i)^2 + \frac{1}{n}\sum_i {\epsilon_i}^2 + 2\gamma\frac{1}{n}\sum_i \epsilon_i(Z_i - rX_i)
\end{aligned}$$

Or :

$$\frac{1}{n}\sum_i {\epsilon_i}^2 \xrightarrow{P} \sigma^2, \quad \frac{1}{n}\sum_i \epsilon_i(Z_i - rX_i) \xrightarrow{P} \boldsymbol{cov}(\epsilon, Z - rX) = 0$$

et

$$\begin{aligned}
\frac{1}{n}\sum_i (Z_i - rX_i)^2 \xrightarrow{P} \mathbb{E}(Z_i - rX_i)^2 &= \mathbb{E}(Z^2) - 2r\mathbb{E}(XZ) + r^2\mathbb{E}(X^2) \\
&= 1 - 2r^2\mathbb{E}(X^2) + r^2\mathbb{E}(X^2) \\
&= 1 - r^2\mathbb{E}(X^2) \\
&= 1 - r^2.
\end{aligned}$$

On en conclut alors que :

$$\hat{\sigma^2} \longrightarrow \sigma^2 + \gamma^2(1 - r^2).$$

3.2.2 Modèle de régression logistique

On considère deux modèles logistiques tels que:

$$\textit{logit}[\mathbb{P}(Y = 1|X_1)] = \beta_0 + \beta_1 X_1 \tag{3.1}$$

$$\textit{logit}[\mathbb{P}(Y = 1|X_1, X_2)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{3.2}$$

Si on compare l'effet de la variable X_1 dans chacun des modèles en calculant le rapport des cotes à chaque fois OR, on obtient:

Effet brut de X_1 dans (3.1): $OR_1 = \exp(\beta_1)$.

Effet de X_2 ajusté de X_2 dans (3.2): $OR_2 = \exp(\beta_1)$.

Il y a confusion si $OR_1 \neq OR_2$.

Par la suite, on calcule la variation relative:

$$VR = \frac{\hat{OR}_2 - \hat{OR}_1}{\hat{OR}_2}. \tag{3.3}$$

En supposant τ compris entre 0.1 et 0.2.

Si $VR \geq \tau$ alors X_2 est un facteur de confusion.

Si $VR \leq \tau$ on vérifie $\beta_2 = 0$. Si oui, on retire X_2 de l'étude.

3.3 Confusion résiduelle et Erreur de mesure

3.3.1 Confusion résiduelle

D'un point de vue statistique, le but de l'analyse épidémiologique est souvent d'estimer l'effet causal d'une variable d'exposition sur un résultat d'intérêt. La confusion peut

être causée par des variables qui sont associées à la fois à la variable de résultat et à la variable d'exposition. Une erreur de mesure dans les variables explicatives et les facteurs de confusion non mesurés peuvent causer des problèmes considérables dans les études épidémiologiques. L'erreur de mesure dans les facteurs de confusion peut conduire à une confusion résiduelle, mais ce genre de cas n'est pas traitable d'une manière simple et évidente parce qu'il n'est pas souvent clair dans quelle direction le problème pointera. Dans Last (1995), le terme "confusion résiduelle" est utilisé pour décrire une confusion non mesurée. Plus précisément, cette dernière est définie comme : la confusion potentielle causée par des facteurs ou des variables pas encore pris en compte dans l'analyse mais aussi par l'erreur de mesure des variables prises en compte, ceux-ci peuvent être directement observables ou non; Dans ce dernier cas, ils sont considérés comme des facteurs de confusion résiduels latents.

Étant en désaccord avec cette définition, en jugeant une non-clarté de la différence entre la confusion et la confusion résiduelle, certains comme Olsen & Basso (1989) proposent qu'il soit plus approprié de définir la confusion résiduelle comme la confusion qui persiste suite à un échec de la contrôler. Les sources de confusion résiduelle seraient alors à la fois des informations insuffisamment détaillées (également issues d'une catégorisation incorrecte), une erreur de classification de la variable, des variables non-incluses ou des erreurs de mesure des variables de confusion incluses.

D'une manière plus claire, comme décrit dans LaMorte & Sullivan (2014), la confusion résiduelle peut être définie comme la distorsion qui subsiste après la prise en compte de confusion dans la conception ou l'analyse d'une étude. Elle peut être causée par exemple par des facteurs de confusion supplémentaires qui ne sont pas pris en compte, car les données sur ces facteurs n'ont pas été recueillies ou parce que le contrôle de confusion n'était pas assez robuste.

Un exemple apparaît d'après Chen et al. (1999) explique plus ce phénomène dans lequel, il s'est avéré que les femmes qui fument pendant la grossesse présentent un risque diminué

d'avoir des bébés naissant avec une trisomie 21. Ceci semble contradictoire en soit, puisque tout le monde sait que le tabagisme n'est pas souvent considéré comme une bonne chose à faire. Devrions-nous demander aux femmes de commencer à fumer pendant la grossesse?

Il s'est avéré donc qu'il existe une relation entre l'âge et le tabagisme pendant la grossesse, et que les plus jeunes femmes sont plus susceptibles de commencer cette mauvaise habitude. Les jeunes femmes sont également moins susceptibles de donner naissance à un enfant atteint de trisomie 21. Par la suite, lorsqu'on traite le modèle concernant le tabagisme et la trisomie 21 avec la covariable âge, alors l'effet du tabagisme disparaît. Mais quand on fait l'ajustement du modèle en utilisant une variable binaire (âge $\leq$ 35 ans, l'âge $\geq$ 35 ans), l'effet protecteur du tabagisme semble rester. Ceci est un exemple de confusion résiduelle si on est capable de démontrer que l'effet disparait avec La variable continue âge.

Etant donné l'importance de la confusion en épidémiologie, des méthodes statistiques ciblent spécifiquement ce problème. La Confusion résiduelle est un problème unique à la régression, donc évidemment toute approche pour le cibler porte sur la régression .

3.3.2 Erreur de mesure

Shawky (2000) affirme que l'erreur aléatoire est définie comme étant la différence due au hasard, d'une observation sur un échantillon de la vraie valeur de la population, conduisant à un manque de précision dans la mesure d'association.

Les erreurs de mesure se produisent lorsque la valeur mesurée diffère de la valeur réelle. De telles erreurs peuvent être attribuables à plusieurs facteurs. Ces erreurs peuvent être aléatoires ou elles peuvent entraîner un biais systématique si elles ne sont pas aléatoires.

En particulier, on suppose que toute observation est composée de la valeur réelle plus une certaine valeur d'erreur aléatoire. Mais est-ce raisonnable? Que faire si toutes les erreurs ne sont pas aléatoires? Une façon de traiter cette notion est de réviser la nature de l'erreur et la décomposer en deux sous-catégories, erreur aléatoire et erreur systématique. L'erreur aléatoire est causée par des facteurs qui affectent de façon aléatoire la mesure de la variable dans l'échantillon. Par exemple, l'humeur de chaque personne peut affecter sa performance en toute occasion. Dans un test particulier, et dans le cas où l'humeur est indépendante de chaque biais éventuel, certains enfants peuvent se sentir de bonne humeur et d'autres peuvent se sentir déprimés. Si l'humeur affecte leurs performances dans un examen par exemple, elle peut augmenter artificiellement les scores observés pour certains enfants et les diminuer artificiellement pour les autres.

Pour cette raison, l'erreur aléatoire est parfois considérée comme le bruit.

Tandis que, l'erreur systématique est causée par des facteurs qui affectent systématiquement la mesure de la variable dans l'échantillon. Par exemple, s'il y a du trafic fort qui passe juste à l'extérieur d'une salle de classe où les étudiants passent un test, ce bruit est susceptible d'affecter tous les scores des étudiants, dans ce cas, systématiquement il va les abaisser d'où une erreur dans la mesure de la moyenne de chaque étudiant. Contrairement à l'erreur aléatoire, les erreurs systématiques ont tendance à être systématiquement positives ou négatives, de ce fait, une erreur systématique est parfois considérée comme un biais dans la mesure.

Il se peut aussi dans certains cas, qu'une erreur de mesure soit aléatoire dans un contexte et systématique dans un autre contexte. Prenons l'exemple d'une coupure de courant qui se produit dans une usine, si cette coupure se produit pendant la journée, cette coupure va affecter systématiquement le fonctionnement de toutes les machines et donc le rendement au sein de l'entreprise, et donc elle est considérée comme une erreur systématique. Par contre, si cette erreur se produit pendant la nuit, elle n'affectera que les machines fonctionnant la nuit comme le système d'alarme par exemple, dans ce cas, c'est une erreur

aléatoire.

Modèle de régression linéaire avec erreur de mesure

Dans la littérature, on trouve différents modèles de régression reliant Y à $(X_1, .., X_p)$ comme par exemple un modèle de régression logistique, modèle de Cox ou modèle de régression linéaire et ceci selon la distribution de Y.
Notre intérêt se porte sur la régression linéaire:

$$\boldsymbol{Y} = \boldsymbol{Z}\beta + \boldsymbol{\epsilon}, \tag{3.4}$$

et $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ et $\beta = (\beta_0, \cdots, \beta_q)$
qui satisfait les hypothèses classiques du modèle linéaire. Souvent, la variable $\boldsymbol{Z}$ est contaminée par une perturbation e, c'est à dire qu'on va observer une nouvelle variable X au lieu de Z telle que

$$\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{e}$$

où $\boldsymbol{e}$ est un vecteur d'erreurs de mesure avec $\mathbb{E}(\boldsymbol{e}) = 0$. Ce modèle s'appelle un modèle avec erreurs de mesure. On note par β^* le nouveau paramètre associé à X après la contamination de Z. L'estimation de β^* par la méthode des moindres carrés est :

$$\begin{aligned}
\hat{\beta}^* &= \frac{\boldsymbol{cov}(Y, X)}{\boldsymbol{var}(X)} \\
&= \frac{\boldsymbol{cov}(Y, Z) + \boldsymbol{cov}(Y, e)}{\boldsymbol{var}(Z) + \boldsymbol{var}(e)} \\
&= \frac{\boldsymbol{cov}(Y, Z)}{\boldsymbol{var}(Z)} \\
&\equiv \hat{\beta} \xrightarrow{P} \beta.
\end{aligned}$$

Remarque: Plus $\boldsymbol{var}(e)$ augmente, plus $\hat{\beta}^*$ sous estime β.

En tenant compte des conditions listées ci-dessus, on obtient que

$$\hat{\beta}^* \xrightarrow{P} \frac{\boldsymbol{cov}(Y,Z)}{\boldsymbol{var}(Z)}$$

Remarque

Le modèle d'erreur de mesure "classique" est

$$\boldsymbol{X} = \boldsymbol{Z} + e \tag{3.5}$$

où $\boldsymbol{Z}$ est la valeur réelle et $\boldsymbol{X}$ est la valeur observée. Ce modèle est le plus couramment considéré. Alternativement, dans le cas d'une expérience, il peut être plus judicieux d'utiliser le modèle d'erreur de Berkson (1950).

$$\boldsymbol{Z} = \boldsymbol{X} + e \tag{3.6}$$

Par exemple, supposons que nous désirions à étudier une réaction chimique quand une concentration $\boldsymbol{X}$ donnée de substrat est présente. Toutefois, en raison de notre incapacité à contrôler complètement le processus, la concentration réelle du substrat $\boldsymbol{Z}$ diffère aléatoirement de $\boldsymbol{X}$ d'une quantité inconnue e. Dans ce cas, on ne peut pas simplement réorganiser $\boldsymbol{Z} = \boldsymbol{X} + e$ à $\boldsymbol{X} = \boldsymbol{Z} + e$ et affirmer que les deux situations sont équivalentes. Dans le modèle (3.5), e est indépendante de $\boldsymbol{Z}$ mais dépend de $\boldsymbol{X}$. Tandis que dans le modèle (3.6), e est indépendante de $\boldsymbol{X}$ mais dépend de $\boldsymbol{Z}$.

3.4 Modèle de régression linéaire avec une erreur de mesure dans une variable de confusion

Dans les études épidémiologiques, plusieurs variables sont susceptibles d'être contaminées par des erreurs de mesure. De ce fait, les résultats des modèles étudiées peuvent y être affectés.

Dans cette section, on s'intéresse à un problème de confusion dans un modèle de régression linéaire en prenant en considération la présence d'une erreur de mesure dans la variable de confusion. Le modèle schématisant ce scénario est le suivant:

$$Y_i = \alpha + \beta X_i + \gamma Z_i + \epsilon_i \tag{3.7}$$

avec X est la variable d'exposition et Z est la variable de confusion affectée par une erreur de mesure b tel que $Z_i = W_i + b_i$.

Les hypothèses du modèle sont définies par:

- $\boldsymbol{cor}(X_i, Z_i) = r$
- $\mathbb{E}(X_i\epsilon_i) = 0$
- $\mathbb{E}(Z_i\epsilon_i) = 0$
- $b \sim N_n(0, \tau^2 I)$
- $\epsilon_i \sim N_n(0, \sigma^2 I)$
- $\boldsymbol{var}(X) = 1$

Or, on observe le modèle tel que Y_i' est la nouvelle variable de réponse après la contamination :

$$Y_i' = \alpha + \beta \boldsymbol{X_i} + \gamma(\boldsymbol{Z_i} - b_i) + \boldsymbol{\epsilon_i}$$

où $\boldsymbol{Z_i} = \boldsymbol{W_i} + b_i$ tels que: b est indépendant de W et $\boldsymbol{cov}(b, Z) \neq 0$.
L'estimateur $\hat{\beta}$ des moindres carrés ordinaires s'écrit sous la forme

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_i Y_i'(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \\
&= \frac{\sum_i (\alpha + \beta X_i + \gamma Z_i - \gamma b_i + \epsilon_i)(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \\
&= \frac{\alpha \sum_i (x_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} + \beta \frac{\sum_i X_i (X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} + \gamma \frac{\sum_i Z_i (X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} - \gamma \frac{\sum_i b_i (X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} + \frac{\sum_i \epsilon_i (X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \\
&= \beta \frac{\sum_i (X_i - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} + \gamma \frac{\sum_i (Z_i - \bar{Z})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} - \gamma \frac{\sum_i (b_i - \bar{b})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} + \frac{\sum_i (\epsilon_i - \bar{\epsilon})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}
\end{aligned}$$

Or d'après le théorème de Slutsky : Premièrement,

$$\frac{\sum_i (Z_i - \bar{Z})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \xrightarrow{P} \frac{\boldsymbol{cov}(X, Z)}{var(X)} = r$$

Deuxièmement,

$$\frac{\sum_i (\epsilon_i - \bar{\epsilon})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \xrightarrow{P} \frac{\boldsymbol{cov}(\epsilon, X)}{\boldsymbol{var}(X)} = 0$$

Troisièmement,

$$\frac{\sum_i (b_i - \bar{b})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \xrightarrow{P} \frac{\boldsymbol{cov}(X, b)}{var(X)} = \boldsymbol{cov}(X, b).$$

Donc:

$$\beta \xrightarrow{P} \beta + \gamma r - \gamma \boldsymbol{cov}(X, b)$$

CHAPITRE 4

Effets d'une erreur de mesure dans une variable de confusion dans un modèle de régression

Dans les études épidémiologiques, les erreurs de mesure peuvent biaiser les résultats. La présence de ces erreurs peut être due à plusieurs facteurs, entre autres, la nature de la variable étudiée (qualitative ou quantitative..). Par exemple, pour la variable catégorie socio-professionnelle, il se présente difficile d'avoir l'information précise pour cette variable et les erreurs de mesure qui y sont associées sont aussi parfois difficiles à calculer. Dans d'autres cas, comme par exemple, pour la variable âge, l'erreur de mesure associée peut être plus facilement détectée.

L'idée de ce projet fait suite à l'étude faite dans Barnwell et al. (2014), et a comme but d'évaluer l'impact de la présence d'une erreur de mesure d'une variable de confusion et de son effet sur les résultats d'un modèle de régression linéaire et logistique.

Dans certaines études épidémiologiques, les simulations peuvent être utilisées comme outils pour aider à la compréhension de nombreux concepts statistiques.

Le résultat des simulations de ce chapitre auront pour but d'évaluer l'estimation de l'erreur de type I du modèle selon la variance de l'erreur de mesure associée à la variable de confusion et ceci dans plusieurs scénarios en faisant varier les différents paramètres du modèle de régression. Par exemple, si on prend le cas du modèle considéré au début de ce mémoire où le café est considéré comme la variable d'exposition, le tabagisme comme

la variable de confusion et les maladies cardio-vasculaires comme notre variable d'intérêt, nous cherchons dans ce cas à savoir l'effet qu'a l'erreur associée au tabagisme sur les résultats du modèle de régression et si le contrôle de cette variable affecte le résultat final et c'est pour cette raison, que nous allons nous intéresser à l'estimation de β_1 associée à X_1 (Café). D'après les simulations proposées, l'intérêt est de définir l'effet de cette perturbation sur les résultats du modèle. Plus spécifiquement, cette approche de simulation a pour but de permettre :

- La génération de deux variables aléatoires.
- L'ajout d'une erreur de mesure dans l'une des deux variables.
- L'étude de l'impact de l'ajout des erreurs de mesures sur les paramètres du modèle de régression et leurs estimateurs.

Pour ce faire, dans le cas de la régression linéaire et logistique, on considère le cas de deux variables normales X_1 et X_2 de moyennes et de variance connues tout en ajoutant une erreur de mesure à X_2.
Au cours de ces simulations, les valeurs des paramètres β_0 et β_2 sont limitées à des valeurs positives.
On fait par la suite varier les différents paramètres du modèle de régression tels que:

- β_0: L'ordonnée à l'origine.
- β_1: Le paramètre associé à X_1.
- β_2: Le paramètre associé à X_2.
- $\sigma_{1,2}$: la corrélation entre X_1 et X_2 .

Pour les simulations, le logiciel R a été utilisé. ce logiciel permet de faire des analyses statistiques et de produire des graphiques. Il a été choisi parce qu'il est également un language de programmation complet, un des aspects qui le diffère des autres logiciels statistiques.

4.1 Cas de régression linéaire

4.1.1 Génération des données

Les simulations ont été modélisées pour un échantillon de taille $n = 1000$, dans le scénario général du modèle de régression linéaire suivant:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon. \tag{4.1}$$

La variable d'exposition X_1 et la variable de confusion X_2 ont été générées avec un processus normal. Nous avons généré selon une normale multivariée la variable

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

avec une moyenne égale à

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

et de matrice de variance covariance :

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \sigma_{1,2} \\ \sigma_{1,2} & 1 \end{pmatrix}.$$

$\boldsymbol{\epsilon}$ est définie comme l'erreur du modèle, associée à $\boldsymbol{Y}$, qui signifie aussi la portion de $\boldsymbol{Y}$ qui a été mal expliquée par X_1 et X_2. On suppose que cette erreur suit une loi normale de moyenne 0 et de variance fixée à l'avance égale à 0,5.
Une autre erreur e de moyenne nulle et de variance égale à sig est ensuite ajoutée à X_2.
On fera varier la valeur de sig de 0 à 0,9 afin de voir son effet sur les estimateurs par la méthode des moindres carrés des paramètres β_0, β_1 et β_2.
On notera $X_{2,2}$ la variable observée avec l'erreur e telle que $X_{2,2} = X_2 + e$.
Nous ferons ensuite une régression linéaire de X sur Y.

4.1.2 Résultats et interprétations

Nous observons à chaque fois la variation de $(\hat{\beta}_1 - \beta_1)$ selon la variation d'autres paramètres du modèle cités auparavant. Les figures ont été faites à l'aide de la commande *boxplot* sur le logiciel R.
Le *boxplot*, appelé aussi boîte à moustaches, est une interprétation graphique des données statistiques basées sur le minimum, premier quartile, médiane, troisième quartile, et maximale.
Les simulations on été faites sur le logiciel R en utilisant le package "MASS" et les commandes "rmvnorm", "boxplot" et "lm".

Résultats basés sur un échantillon avec une seule répétition

Dans cette section, nous visons à observer le comportement de l'estimateur du paramètre β_1 selon tout en faisant varier les paramètres du modèle linéaire.
On définit le vecteur $\tau = (n, \beta_0, \beta_1, \beta_2, \sigma_{1,2})$ qui sera modifié selon le cas.

Dans ce qui suit, nous exposons les résultats des simulations avec différents scénarios avec une seule répétition:

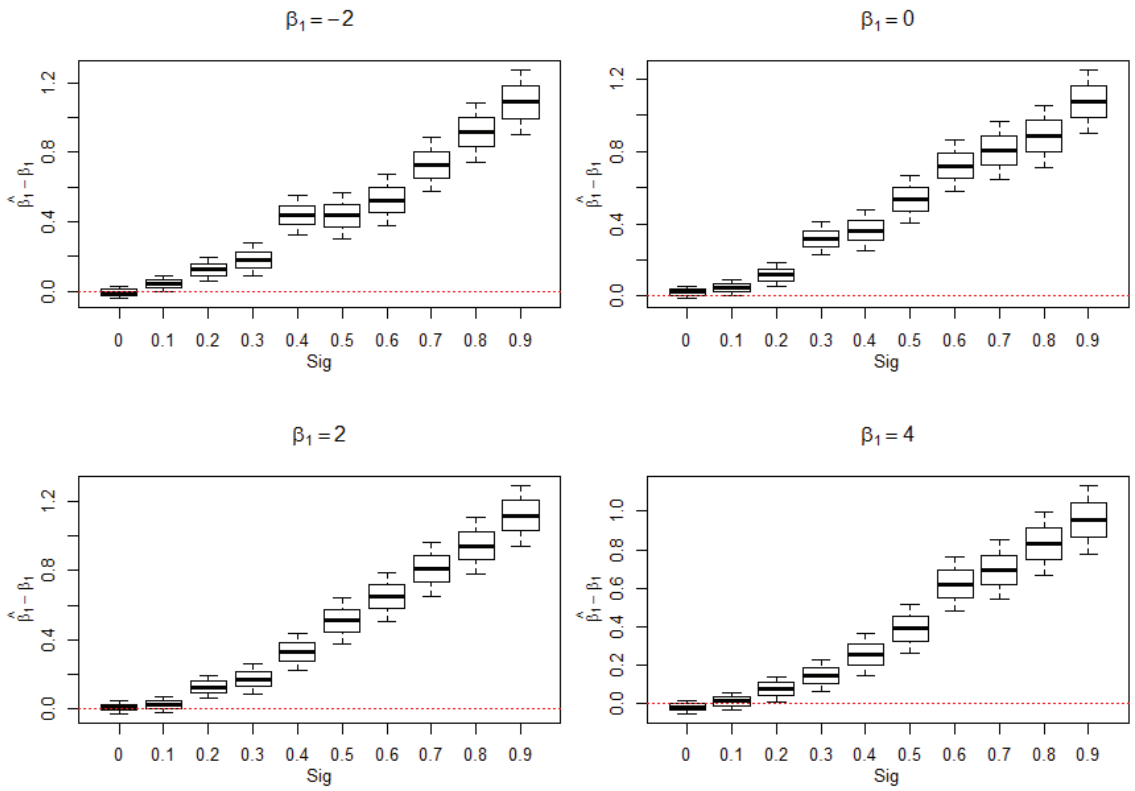

Figure 4.1: **Effet de la variation de β_1 sur les résultats du modèle**: Variation de $(\hat{\beta}_1 - \beta_1)$ selon *sig* pour le modèle linéaire (4.1) tel que $\tau = (1000, 4, ., 4, 0.5)$.

La figure (4.1) illustre le *boxplot* représentant la variation de $(\hat{\beta}_1 - \beta_1)$ selon la variance de l'erreur de mesure, *sig*, pour différentes valeurs de β_1. On peut remarquer que la valeur de β_1 dans ce cas n'a pas d'effet significatif sur la variation de $(\hat{\beta}_1 - \beta_1)$ après avoir ajouté la contamination à la variable X_2. Le coefficient β_1 est toujours sur-estimé et on remarque aussi que la variance de $\hat{\beta}_1$ augmente avec *sig*.

La figure (4.2) illustre le *boxplot* représentant la variation de $(\hat{\beta}_1 - \beta_1)$ selon la variance de l'erreur de mesure, *sig*, pour différentes valeurs de β_2. D'après cette figure, on constate que quand β_2 augmente, $\hat{\beta}_1$ augmente globalement d'où la croissance de $(\hat{\beta}_1 - \beta_1)$ et ceci selon *sig*. On remarque aussi que pour le cas où $\beta_2 = 0$, β_1 est légèrement sous-estimé.

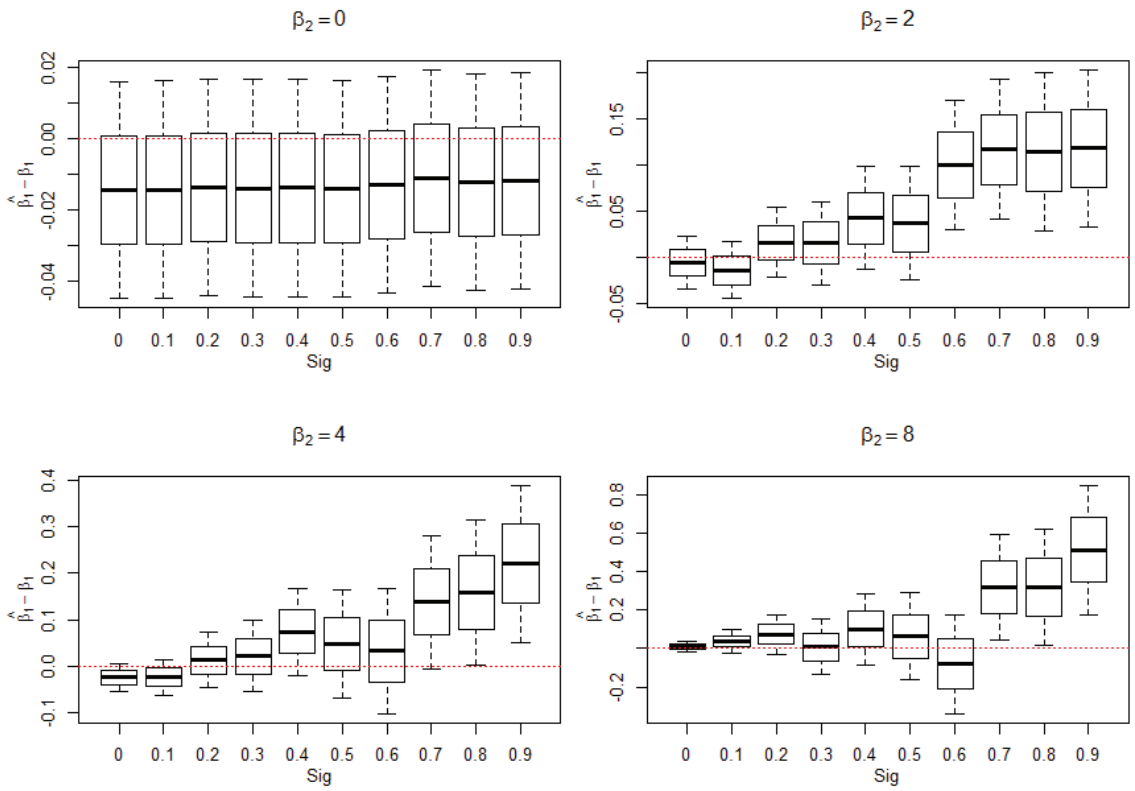

Figure 4.2: **Effet de la variation de β_2 sur les résultats du modèle**: Variation de $(\hat{\beta}_1 - \beta_1)$ selon *sig* pour le modèle linéaire (4.1) avec $\tau = (1000, 2, 4, , 0.1)$.

La figure (4.3) illustre le *boxplot* représentant la variation de $(\hat{\beta}_1 - \beta_1)$ selon la variance de l'erreur de mesure, *sig*, avec différentes valeurs de β_2. D'après cette figure, sachant que la corrélation entre X_1 et X_2 est dans ce cas assez forte et égale à 0.9, on constate que pour une valeur de β_2 non nulle, $(\hat{\beta}_1 - \beta_1)$ est positive, c'est à dire que β_1 est sur-estimé et augmente plus on augmente la valeur de β_2. Pour $\beta_2 = 0$, $(\hat{\beta}_1 - \beta_1)$ est tantôt positive tantôt négative et ne nous permet pas d'en tirer une claire conclusion.

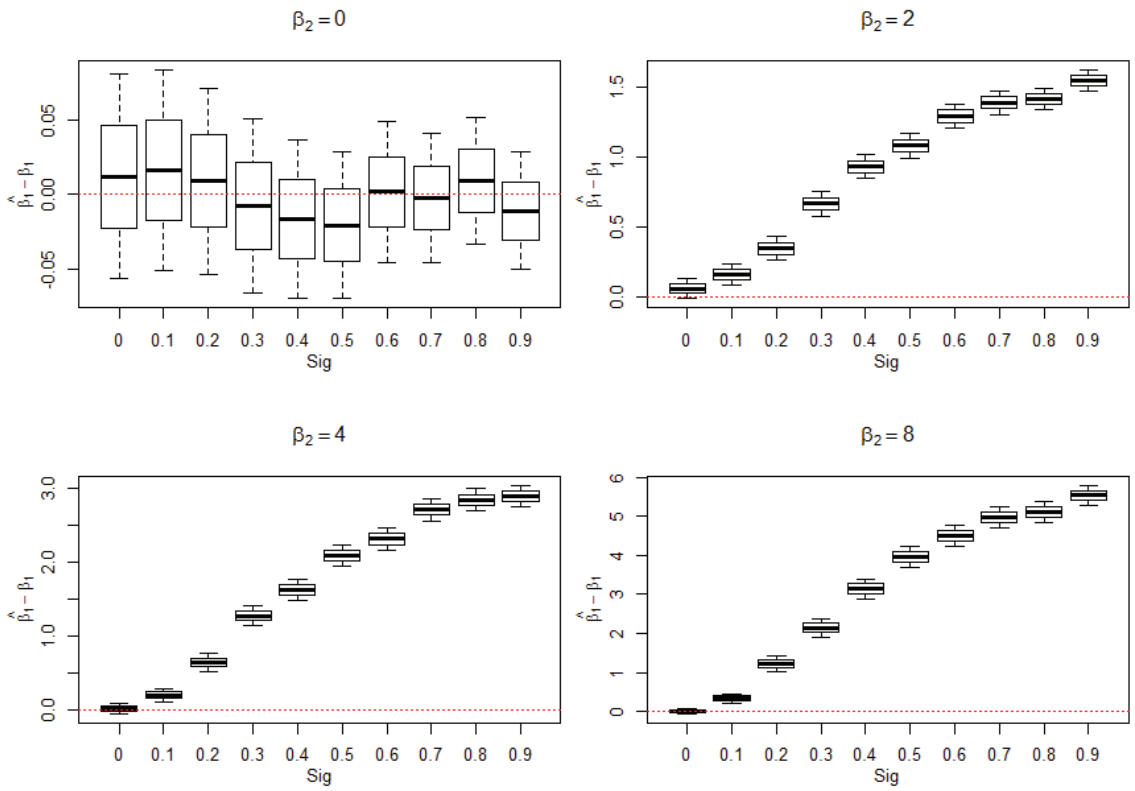

Figure 4.3: **Effet de la variation de β_2 sur les résultats du modèle**: Variation de $(\hat{\beta}_1 - \beta_1)$ selon *sig* pour le modèle linéaire (4.1) avec $\tau = (1000, 2, 4, ., 0.9)$.

Les deux figures ci-dessus (4.4) et (4.5) représentent dans des *boxplots*, la variation de $(\hat{\beta}_1 - \beta_1)$ selon *sig* pour différentes valeurs de la corrélation entre X_1 et X_2 notée $\sigma_{1,2}$ tout en laissant les autres paramètres fixes. On remarque clairement ici que, quand $\sigma_{1,2}$ augmente, la valeur de $\hat{\beta}_1$ augmente, c'est à dire que la corrélation entre X_1 et X_2 a une influence sur le comportement de l'estimateur de β_1.

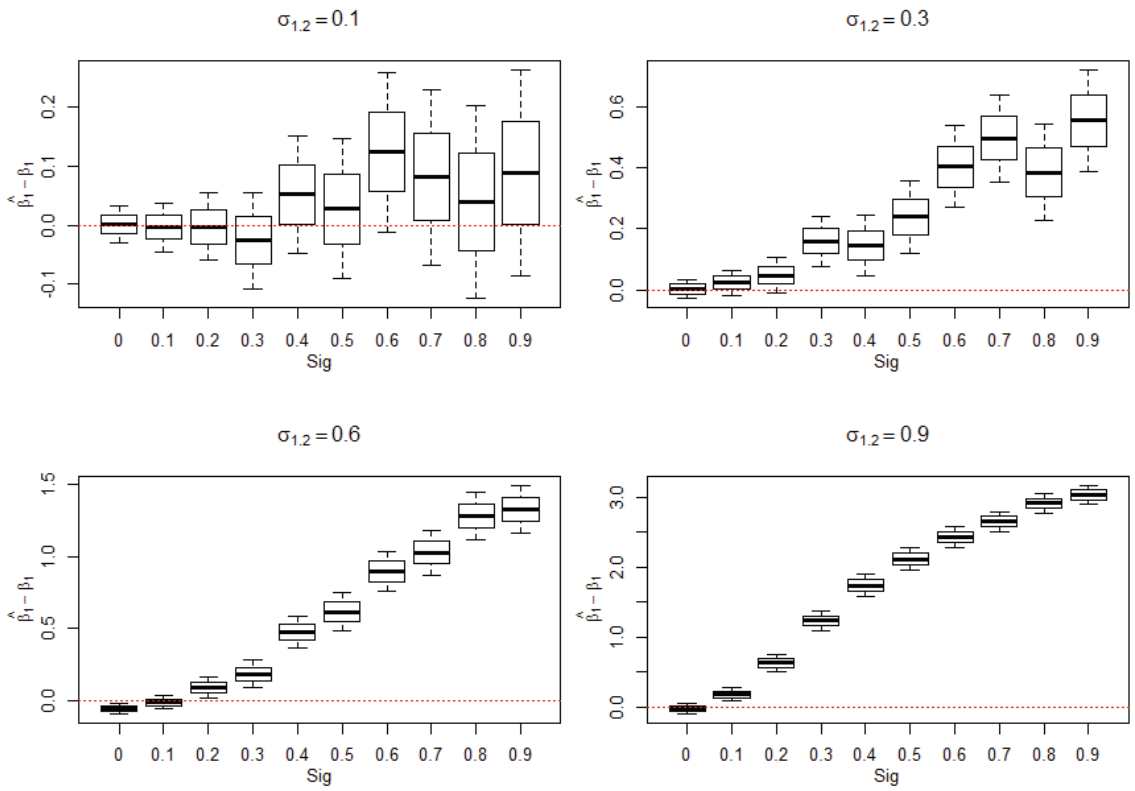

Figure 4.4: **Effet de la variation de $\sigma_{1,2}$ sur les résultats du modèle**: Variation de $(\hat{\beta}_1 - \beta_1)$ selon *sig* pour le modèle linéaire (4.1) avec $\tau = (1000, 4, 4, 4)$.

Résultats basés sur un échantillon avec plusieurs répétitions

Afin de confirmer les résultats obtenus précédemment, les simulations ont été refaites avec 200 répétitions.

Pour différentes valeurs des paramètres β_0, β_1, β_2 et $\sigma_{1,2}$, on procède à une régression linéaire, avec une différente valeur de la variance de l'erreur *sig* à chaque fois, qui varie de 0,1 à 0,9 et on s'intéresse cette fois aux comportements des estimateurs des trois paramètres β_0, β_1 et β_2.

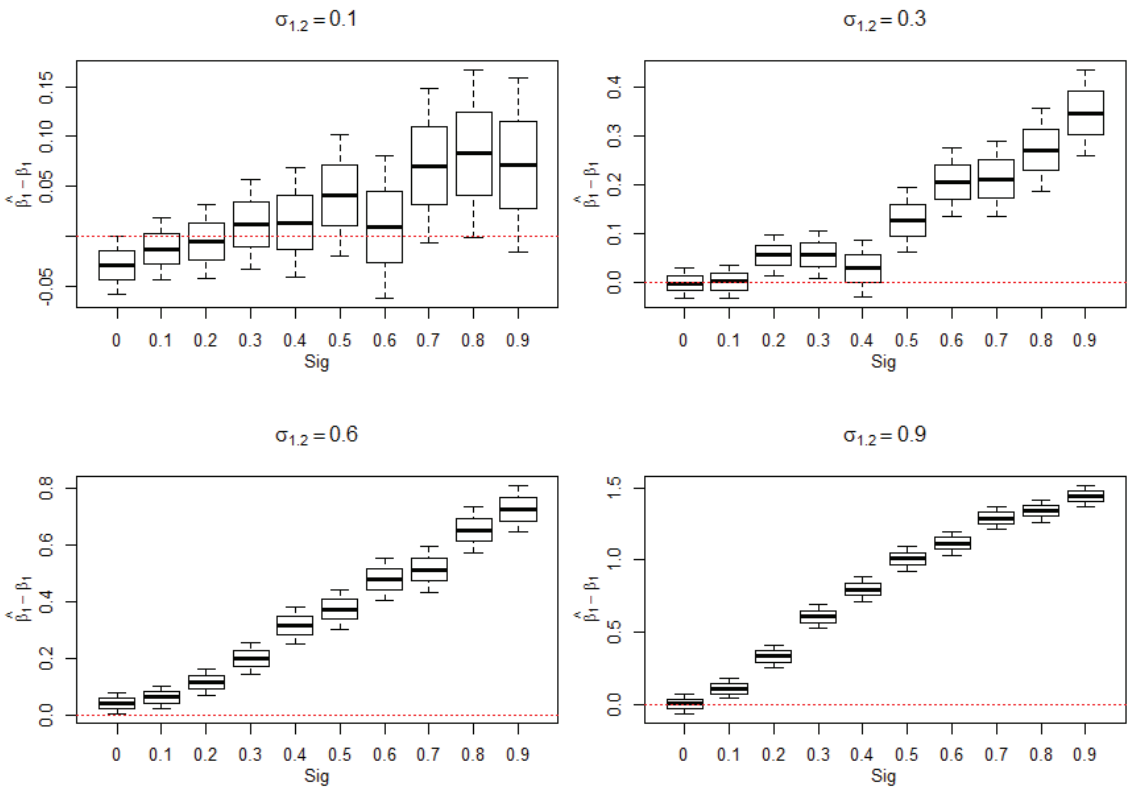

Figure 4.5: **Effet de la variation de $\sigma_{1,2}$ sur les résultats du modèle**: Variation de $(\hat{\beta}_1 - \beta_1)$ selon sig pour le modèle linéaire (4.1) avec $\tau = (1000, 4, 4, 2)$.

La figure (4.6) représente les *boxplots* (a), (b) et (c) illustrant respectivement les variations de $(\hat{\beta}_0 - \beta_0)$, $(\hat{\beta}_1 - \beta_1)$ et $(\hat{\beta}_2 - \beta_2)$ selon la variance de l'erreur de mesure, sig, variant de 0.1 à 0.9. La figure (a) montre que l'estimateur $\hat{\beta}_0$ n'est pas affecté par la contamination de X_2. La figure (b) reflète une légère augmentation de $(\hat{\beta}_1 - \beta_1)$ d'où une sur-estimation de β_1 avec l'augmentation de l'erreur de mesure affectée à X_2. Et d'après la figure (c), on constate que $\hat{\beta}_2$ décroît avec la valeur de sig.

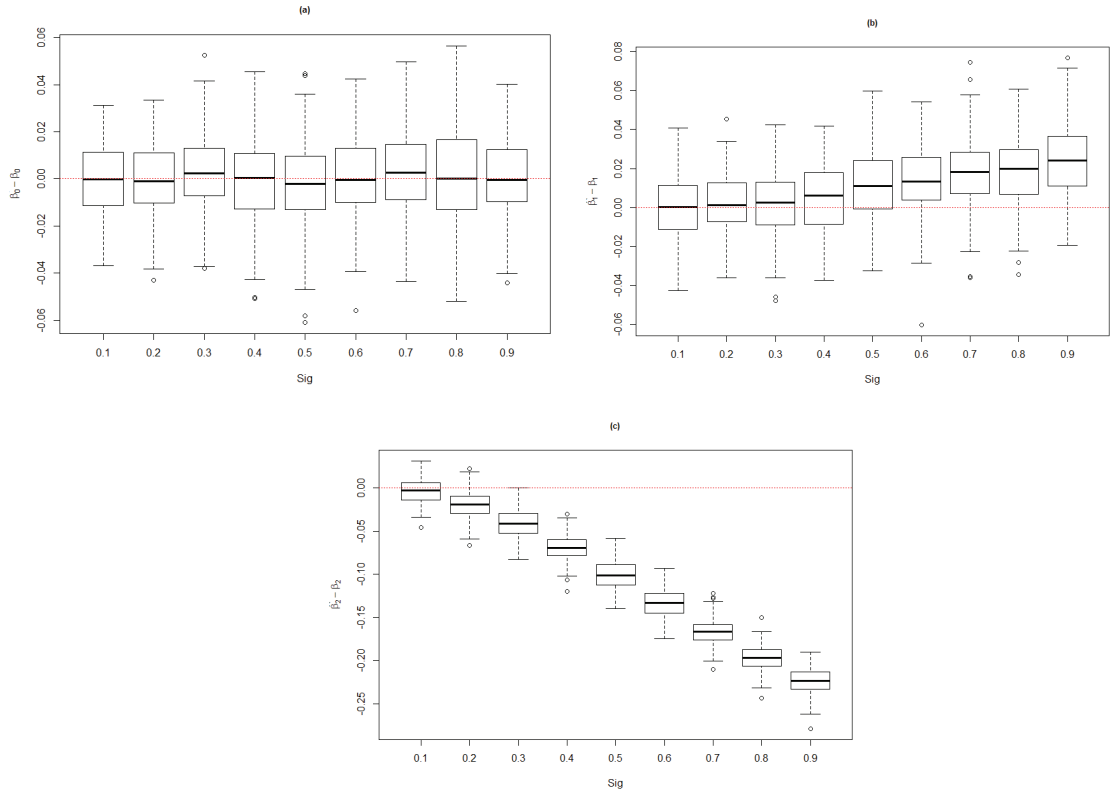

Figure 4.6: **Variation de $(\hat{\beta}_0-\beta_0), (\hat{\beta}_1-\beta_1)$ et $(\hat{\beta}_2-\beta_2)$**: Variation de $(\hat{\beta}_0-\beta_0), (\hat{\beta}_1-\beta_1)$ *et* $(\hat{\beta}_2-\beta_2)$ respectivement dans $(a), (b)$ et (c) selon sig pour le modèle linéaire (4.1) avec $\tau = (1000, 0, 2, 0.5, 0.1)$.

Les deux figures (4.7) et (4.8) représentent les *boxplots* $(a), (b)$ et (c) illustrant respectivement les variations de $(\hat{\beta}_0 - \beta_0), (\hat{\beta}_1 - \beta_1)$ et $(\hat{\beta}_2 - \beta_2)$ selon la variance de l'erreur de mesure, sig, variant de 0.1 à 0.9. Dans le cas de ces figures, le *boxplot* (a) montre que le paramètre β_0 est bien estimé pour toutes les valeurs de sig. D'après le *boxplot* représenté dans (b), on voit bien que $(\hat{\beta}_1 - \beta_1)$ augmente avec l'augmentation de sig et le *boxplot* représenté dans la figure (c) montre que $(\hat{\beta}_2 - \beta_2)$ diminue avec la variation croissante de sig, c'est à dire que le paramètre β_2 est sous-estimé suite à l'ajout de l'erreur de mesure à X_2.

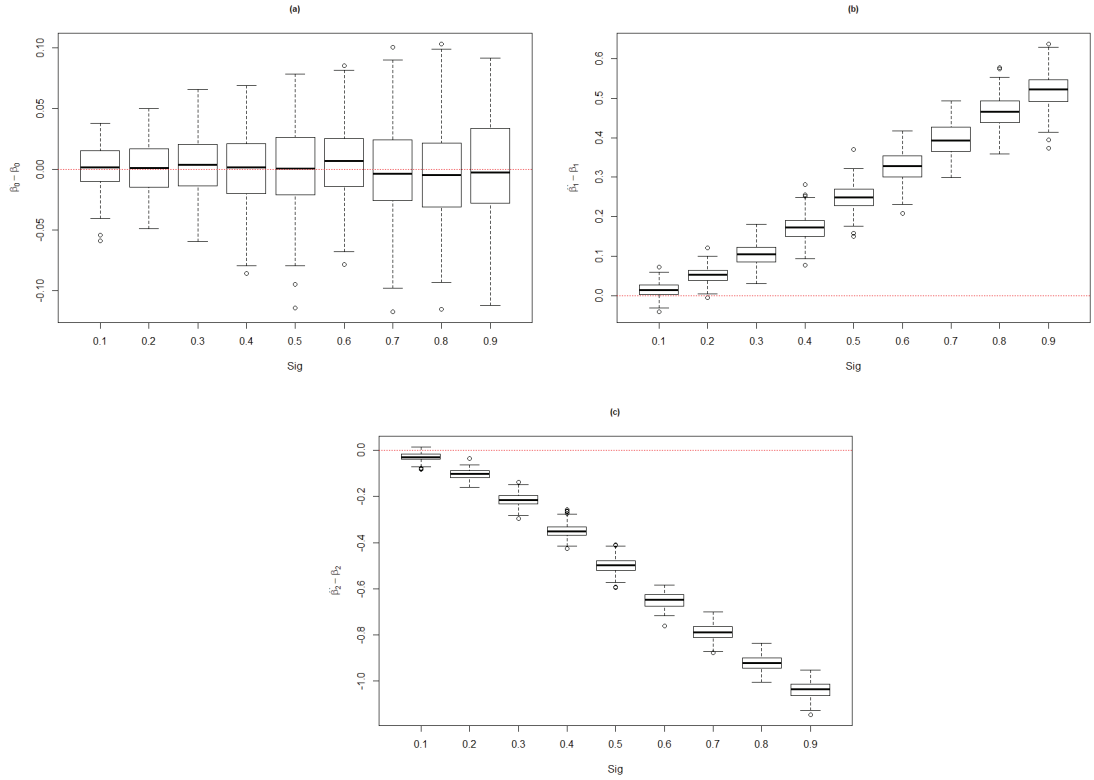

Figure 4.7: **Variation de** $(\hat{\beta}_0-\beta_0), (\hat{\beta}_1-\beta_1)$ **et** $(\hat{\beta}_2-\beta_2)$: Variation de $(\hat{\beta}_0-\beta_0), (\hat{\beta}_1-\beta_1)$ *et* $(\hat{\beta}_2-\beta_2)$ respectivement dans $(a), (b)$ et (c) selon la variance de l'erreur de mesure, sig, pour le modèle linéaire (4.1) avec $\tau = (1000, 4, -2, 2, 0.5)$.

Conclusion

Les simulations faites avec 200 répétitions illustrant les variations des estimateurs des trois paramètres β_0, β_1 et β_2 montrent que seuls $\hat{\beta}_1$ et $\hat{\beta}_2$ sont affectés par l'erreur de mesure ajoutée à la variable de confusion X_2 dans le modèle de régression linéaire.

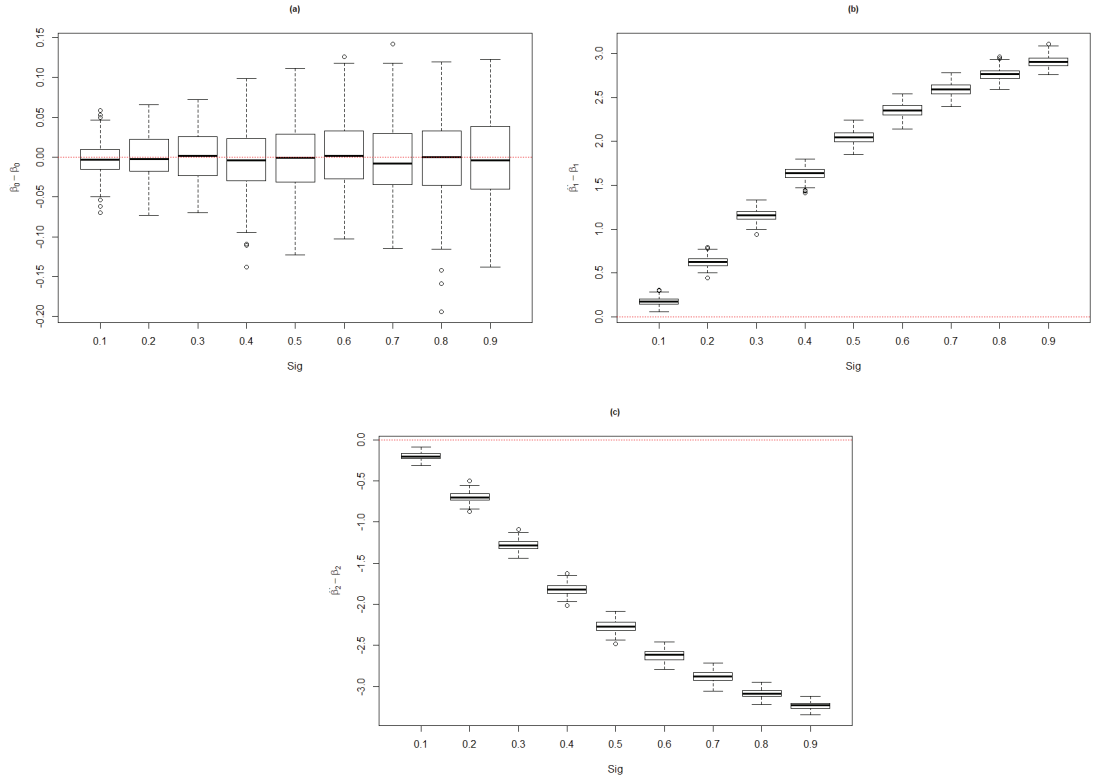

Figure 4.8: **Variation de $(\hat{\beta}_0-\beta_0), (\hat{\beta}_1-\beta_1)$ et $(\hat{\beta}_2-\beta_2)$**: Variation de $(\hat{\beta}_0-\beta_0), (\hat{\beta}_1-\beta_1)$ *et* $(\hat{\beta}_2-\beta_2)$ respectivement dans $(a), (b)$ et (c) selon sig pour le modèle linéaire (4.1) avec $\tau = (1000, 4, -2, 2, 0.5)$.

4.2 Cas de régression logistique

Nous avons déjà vu dans le paragraphe précédent le cas où une variable de confusion est contaminée par une erreur de mesure dans un modèle de régression linéaire. Toutes les définitions et les questions demeurent essentiellement inchangées en régression logistique comme c'est confirmé dans (Joseph, 2010).

En régression linéaire, le but était d'évaluer l'effet de l'ajout d'une erreur de mesure dans une variable de confusion sur les différents paramètres du modèle dans plusieurs scénarios.

Dans cette section, on s'intéresse à une étude similaire pour le cas de la régression logistique, autrement dit, on cherche à connaître l'effet de la contamination d'une variable de confusion par une erreur de mesure et les effets qu'elle peut avoir sur les résultats du modèle de régression logistique.

Le sujet en question a beaucoup été traité dans Austin & Brunner (2004) et Austin & Brunner (2009).

Dans un même principe, on a procédé à des simulations de deux variables normales X_1 et X_2 de moyennes et de variances connues tout en ajoutant un erreur de mesure à X_2 avec le modèle ci-contre

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{k=0}^{K} x_{i,k}\beta_k, \quad i = 1, ..., N \tag{4.2}$$

4.2.1 Génération des données

Comme a été traité dans Barnwell et al. (2014), nous avons généré les données comme suit:

Une variable d'exposition X_1 et un facteur de confusion continu X_2 ont été générés telles que:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

est suit une distribution normale bidimensionnelle (de taille $n = 1000($ de moyenne $\mu = (0, 0)^T$ et de matrice de variance covariance

$$\Sigma = \begin{pmatrix} 1 & \sigma_{1,2} \\ \sigma_{1,2} & 1 \end{pmatrix}$$

On suppose que la variable X à laquelle on s'intéresse est une variable qualitative à 2 modalités: 1 ou 0, succès ou échec... y désigne le nombre de 'succès' observé.

On désigne par π la probabilité telle que :

$$\pi = \mathbb{P}(Z = 1) \quad ou \quad 1 - \pi = \mathbb{P}(Z = 0)$$

telle que

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \tag{4.3}$$

Et d'après (Berkson, 1944)

$$\pi = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2) + 1} \quad et, \quad y \sim Binomiale(p)$$

Par la suite, on ajoute une erreur *sig* qui varie de 0.1 à 0.9 en incréments de 0.1 à la variable d'exposition X_2 telle que $X_{2,2}$ est la variable observée.
On considère un échantillon de taille $n = 1000$ généré à partir du modèle logistique (4.3) et on fait la régression logistique de Y sur X après contamination de X_2. Les simulations on été faites sur le logiciel R.

4.2.2 Résultats et interprétation

Dans cette section, seront montrés les résultats des simulations pour le cas de la régression logistique dans différents scénarios d'abord avec une seule répétition et plusieurs répétitions par la suite. On s'intéresse à la variation de $(\hat{\beta}_1 - \beta_1)$ selon la variation des autres paramètres du modèle. Les figures ont été faites avec la commande *boxplot* sur le logiciel R.

Résultats basés sur un échantillon avec une seule répétition

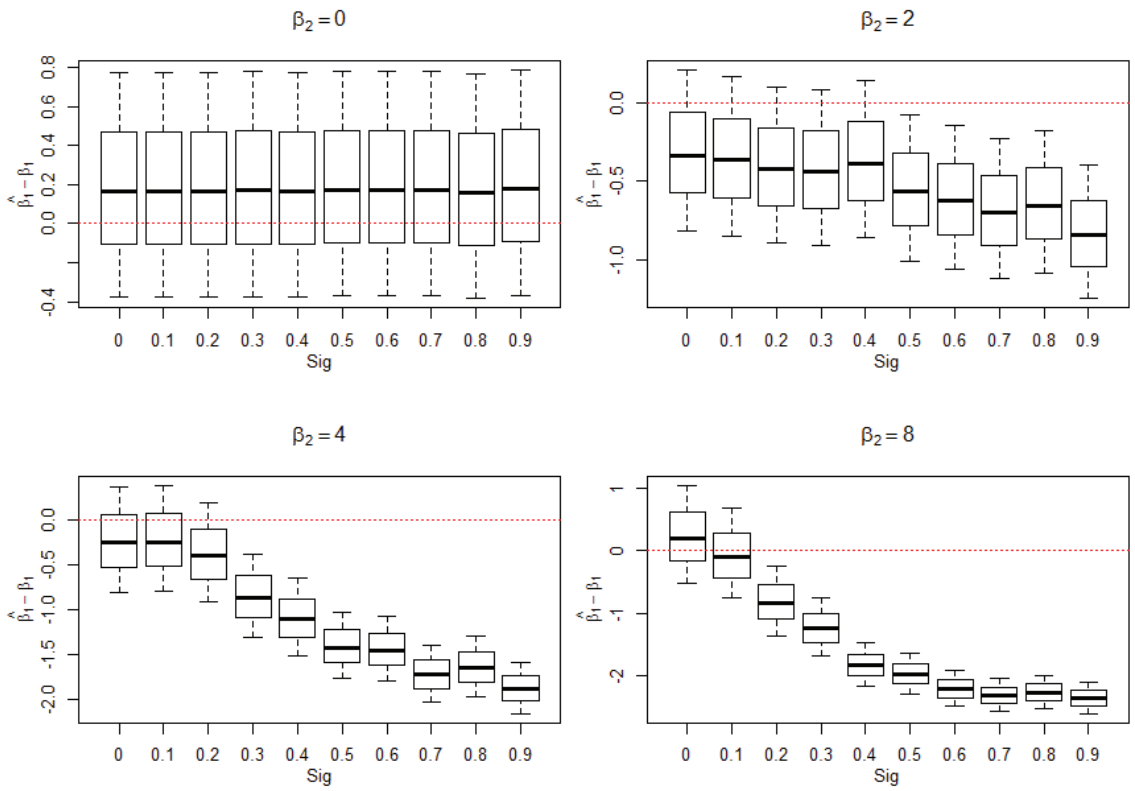

Figure 4.9: **Effet de la variation de β_2 sur les résultats du modèle**: Variation de $(\hat{\beta}_1 - \beta_1)$ selon *sig* pour le modèle logistique (4.3) avec $\tau = (1000, 2, 4, ., 0.1)$.

La figure (4.9) illustre le *boxplot* représentant la variation de $(\hat{\beta}_1 - \beta_1)$ selon la variance de l'erreur de mesure, *sig*. Cette figure montre dans le cas d'une faible corrélation entre X_1 et X_2, on remarque pour une valeur nulle de β_2, que le biais de β_1 est positif ($\hat{\beta}_1$ sur-estime en général β_1) et constant et que sa variance reste constante en fonction de *sig*. Tandis qu'on voit bien que dès que β_2 prend une valeur non nulle, le paramètre β_1 est sous-estimé quelque soit la valeur de *sig* et que le biais est plus sévère quand *sig* ou β_2 augmente. De plus, quand la valeur de β_2 augmente, la variance de l'erreur de mesure de $\hat{\beta}_1$ diminue avec l'augmentation de la valeur de *sig*.

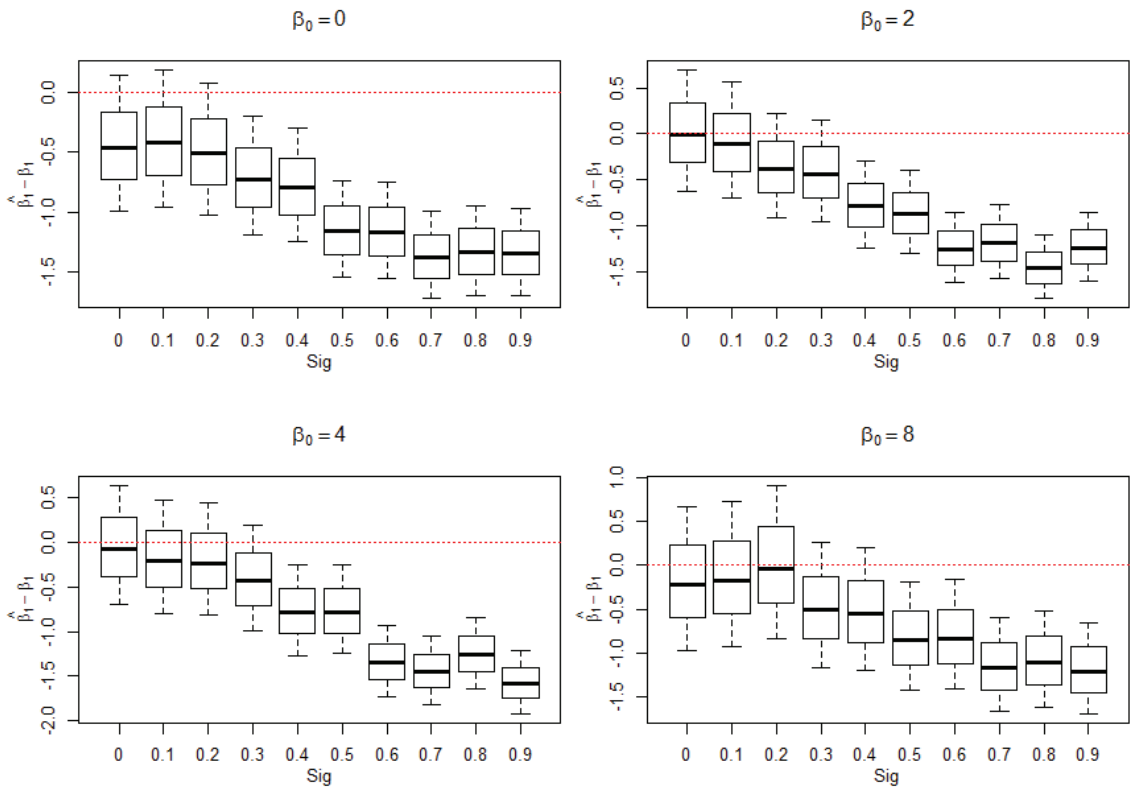

Figure 4.10: **Effet de la variation de β_0 sur les résultats du modèle**: Variation de $(\hat{\beta}_1 - \beta_1)$ selon *sig* pour le modèle logistique (4.3) avec $\tau = (1000, ., 4, 4, 0.4)$.

Dans la figure (4.10) qui illustre le *boxplot* représentant la variation de $(\hat{\beta}_1 - \beta_1)$ selon la variance de l'erreur de mesure, *sig*, on remarque que la variation de β_0 n'a pas un grand effet sur la variation de $\hat{\beta}_1$ et pour toute valeur de β_0, $(\hat{\beta}_1 - \beta_1)$ varie d'une manière décroissante. Aussi, on peut voir que β_1 est largement sous-estimé pour presque toute valeur de *sig*.

La figure (4.11) illustre le *boxplot* de $(\hat{\beta}_1 - \beta_1)$ selon la variance de l'erreur de mesure, *sig*. On remarque qu'avec une forte corrélation entre X_1 et X_2 égale à 0.9, quand β_1 est négatif, le biais devient plus sévère quand *sig* augmente et que la variance diminue avec l'augmentation de *sig*.

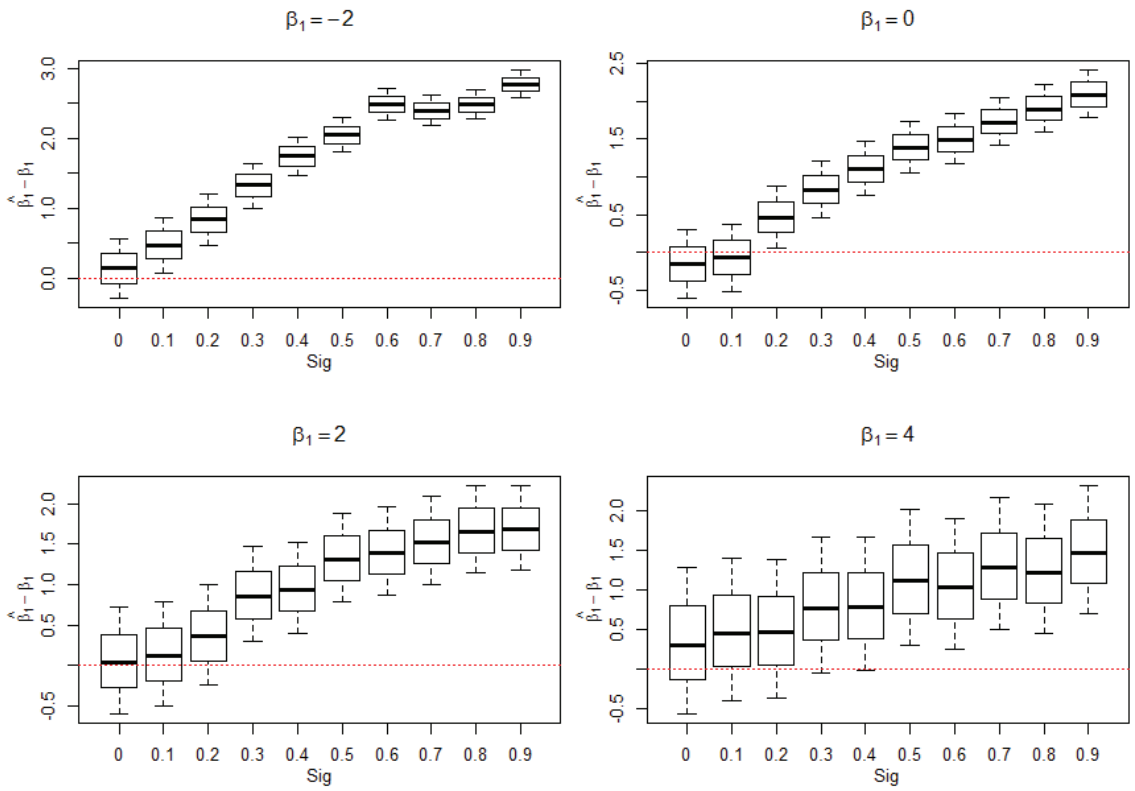

Figure 4.11: **Effet de la variation de β_1 sur les résultats du modèle**: Variation de $(\hat{\beta}_1 - \beta_1)$ selon sig pour le modèle logistique (4.3) avec $\tau = (1000, 0, ., 4, 0.9)$.

Tandis que, lorsque β_1 est positif, on remarque que β_1 est sur-estimé à partir de $sig = 0.2$, pour $\beta_1 = 0$ et pour toute valeur de sig.

La figure (4.12) illustre le *boxplot* de $(\hat{\beta}_1 - \beta_1)$ selon la variance de l'erreur de mesure, sig, avec une corrélation entre X_1 et X_2 est égale à 0.5. Quand β_1 est négatif, le biais devient plus sévère quand sig augmente mais cette fois ci atteint 1.5 pour $sig = 0.9$ comparé à 3 pour le cas de la forte corrélation entre X_1 et X_2 (voir figure 4.11). On remarque aussi, que la variance de $(\hat{\beta}_1)$ augmente avec la variance de l'erreur de mesure sig. Tandis que, le coefficient β_1 est toujours sur-estimé.

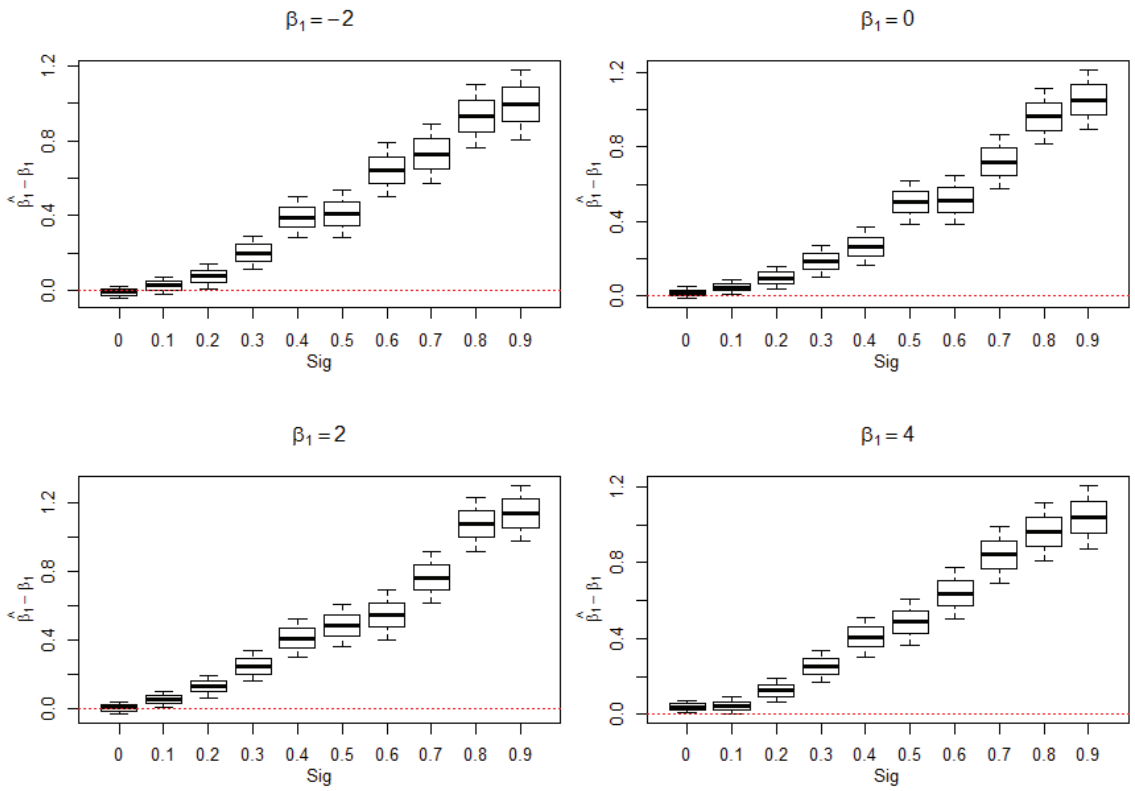

Figure 4.12: **Effet de variation de β_1 sur les résultats du modèle**: Variation de $(\hat{\beta}_1 - \beta_1)$ selon sig pour le modèle logistique (4.3) avec $\tau = (1000, 0, ., 4, 0.5)$.

Résultats basés sur un échantillon avec plusieurs répétitions

Comme dans le cas de la régression linéaire, on refait les simulations avec 200 répétitions. Pour différentes valeurs des paramètres β_0, β_1, β_2 et $\sigma_{1,2}$, on procède à une régression logistique, avec différentes valeurs de la variance de l'erreur sig, qui varie de 0,1 à 0,9 et on s'intéresse cette fois aux comportements des estimateurs des trois paramètres β_0, β_1 et β_2.

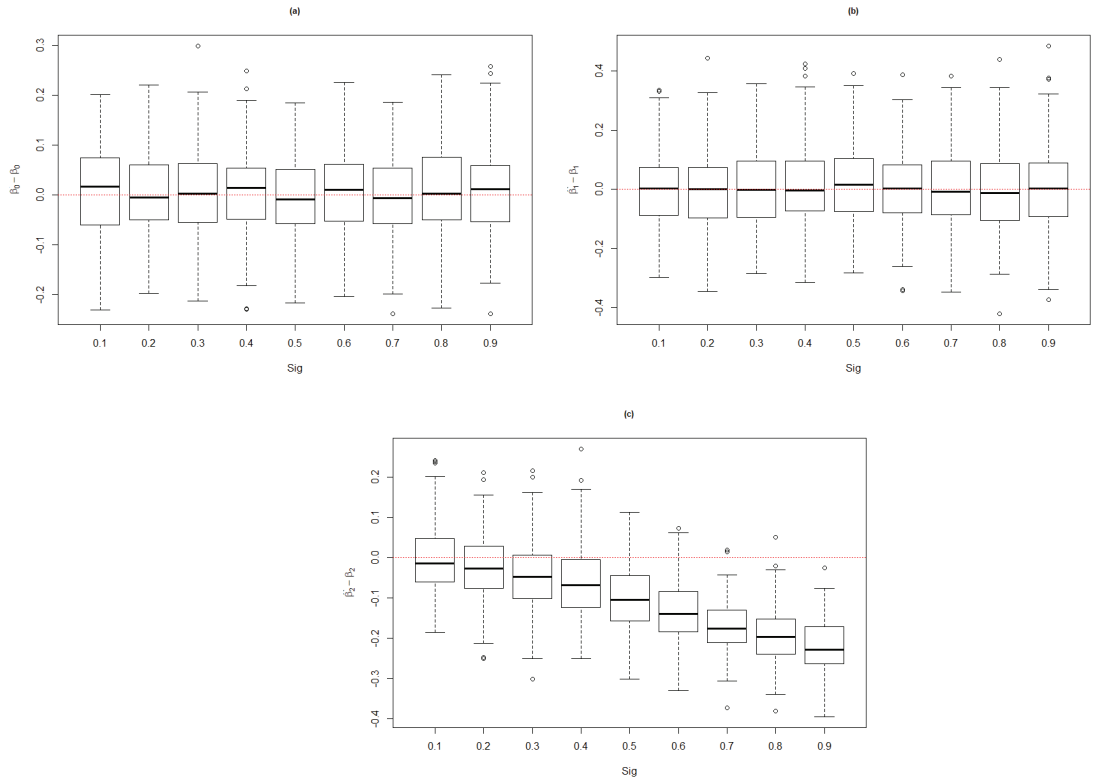

Figure 4.13: **Évaluation des estimateurs du modèle β_0, β_1 et β_2**: Variation de $(\hat{\beta}_0 - \beta_0)$, $(\hat{\beta}_1 - \beta_1)$ et $(\hat{\beta}_2 - \beta_2)$ selon sig pour le modèle logistique (4.3) avec $\tau = (1000, 0, 2, 0.5, 0.1)$.

La figure (4.13) représente les *boxplots* $(a), (b)$ et (c) illustrant respectivement les variations de $(\hat{\beta}_0 - \beta_0), (\hat{\beta}_1 - \beta_1)$ et $(\hat{\beta}_2 - \beta_2)$ selon la variance de l'erreur de mesure, sig, variant de 0.1 à 0.9. On remarque d'après (a) et (b), que les paramètres β_0 et β_1 ne sont pas affectés par la contamination de X_2 pour toute valeur de sig. Tandis que (c) montre que l'estimateur de β_2 est affecté par la présence de l'erreur de mesure dans la variable de confusion et que β_2 subit une très légère sous-estimation quand sig augmente.

La figure (4.14) représente les *boxplots* $(a), (b)$ et (c) illustrant respectivement les variations de $(\hat{\beta}_0 - \beta_0), (\hat{\beta}_1 - \beta_1)$ et $(\hat{\beta}_2 - \beta_2)$ selon la variance de l'erreur de mesure, sig,

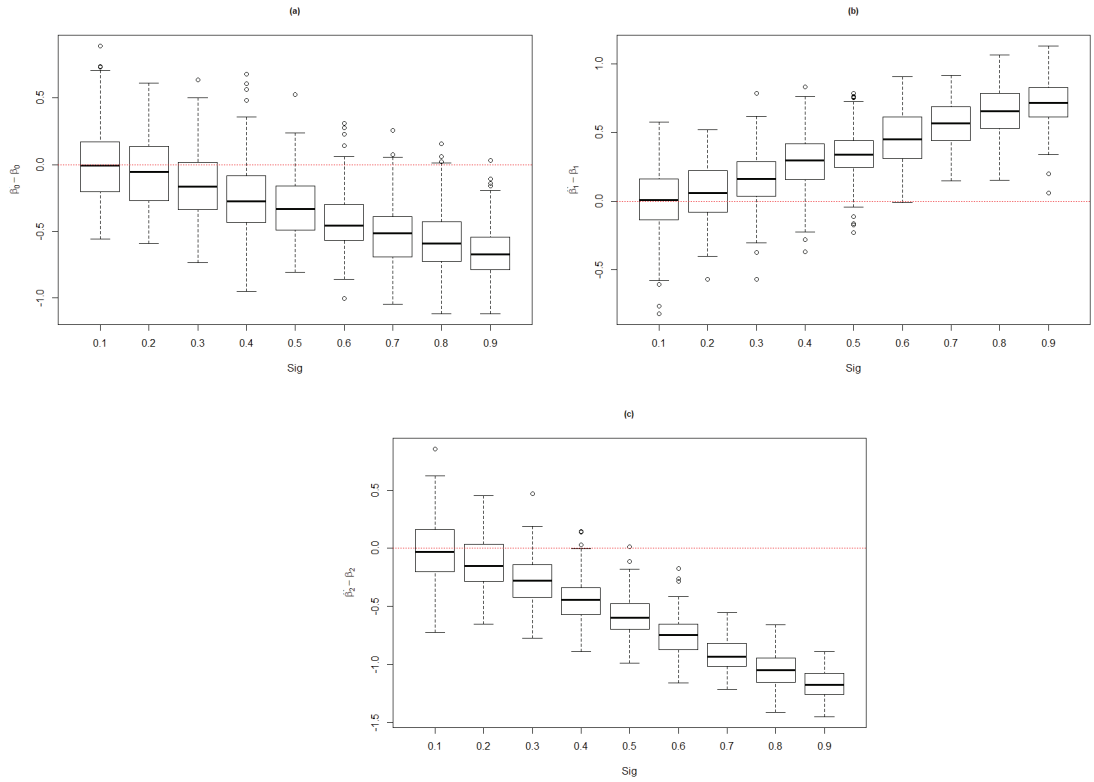

Figure 4.14: **Évaluation des estimateurs du modèle** β_0, β_1 **et** β_2: Variation de $(\hat{\beta}_0 - \beta_0)$, $(\hat{\beta}_1 - \beta_1)$ et $(\hat{\beta}_2 - \beta_2)$ selon *sig* pour le modèle logistique (4.3) avec $\tau = (1000, 4, -2, 2, 0.5)$.

variant de 0.1 à 0.9. D'après (a) et (c), on constate que les paramètres β_0 et β_2 subissent une sous-estimation dans le cas où $(\beta_0 = 4, \beta_1 = -2, \beta_2 = 2, \sigma_{1,2} = 0.5)$. Tandis que, (b) montre que $\hat{\beta}_1$ sur-estime β_1.

La figure (4.15) représente les *boxplots* $(a), (b)$ et (c) illustrant respectivement les variations de $(\hat{\beta}_0 - \beta_0), (\hat{\beta}_1 - \beta_1)$ et $(\hat{\beta}_2 - \beta_2)$ représentés sur l'axe des ordonnés selon la variation de *sig* variant de 0.1 à 0.9 sur l'axe des abscisses. On constate d'après (a) que β_0 n'est pas affecté par la contamination de X_2. D'après (b), le paramètre β_1 subit une sur-estimation quand *sig* augmente. Contrairement à $\hat{\beta}_2$ qui sous-estime β_2. Dans ce cas,

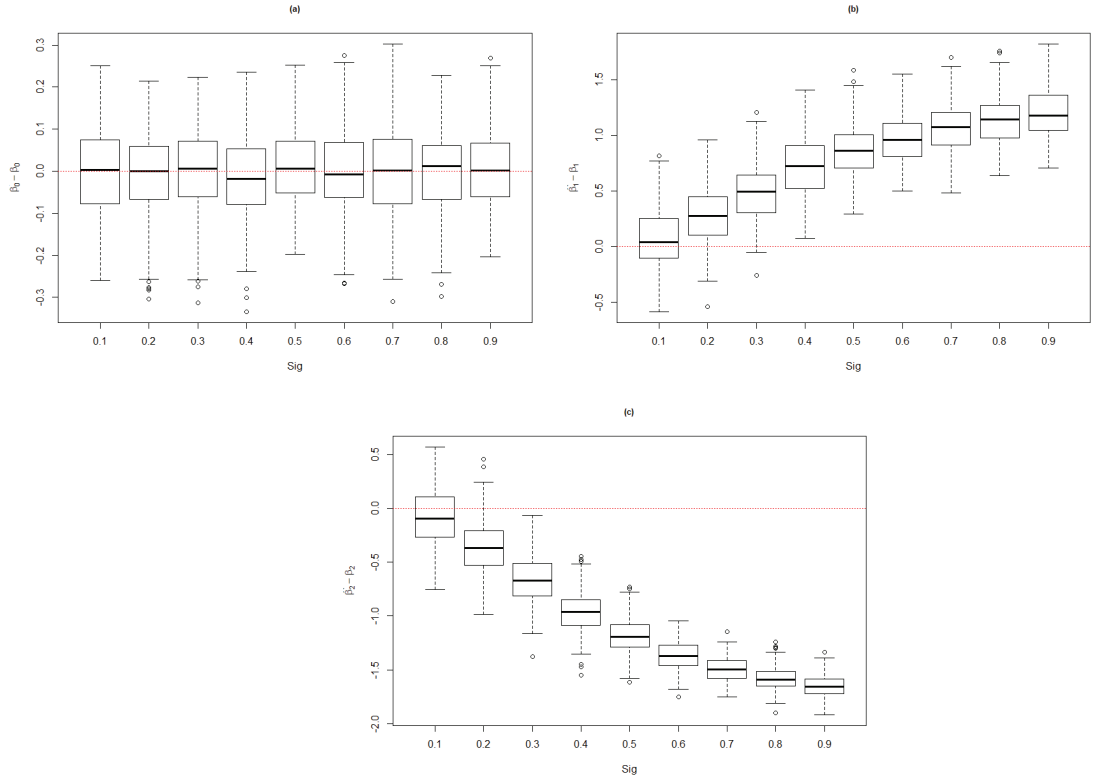

Figure 4.15: **Évaluation des estimateurs du modèle β_0, β_1 et β_2**: Variation de $(\hat{\beta}_0 - \beta_0)$, $(\hat{\beta}_1 - \beta_1)$ et $(\hat{\beta}_2 - \beta_2)$ sur l'axe des ordonnées selon *sig* sur l'axe des abscisses pour le modèle logistique (4.3) avec $\tau = (1000, 0, 2, 2, 0.9)$.

on peut conclure que seuls β_1 et β_2 sont affectés par l'erreur de mesure ajoutée à X_2.

Conclusion

Les résultats précédents correspondants à la régression logistique montrent que dépendamment du cas, varie le comportement des estimateurs des paramètres de la régression et on ne peut désormais pas en tirer une conclusion commune. Pour cette raison, entre autres, qu'on peut dire que le modèle de la régression logistique est plus difficile à traiter dans tous aspects que la régression linéaire.

4.3 Implications et Conclusions

D'après les simulations faites dans le cas du modèle de régression linéaire et logistique et les résultats obtenus, nous pouvons dire que dans le cas de la régression linéaire et logistique, les résultats sont assez différents. Pour la régression linéaire, le paramètre β_0 n'est en général pas affecté par la contamination de la variable de confusion X_2 par une erreur de mesure. Contrairement aux paramètres β_1 et β_2, qui, selon le cas, subissent une mauvaise estimation. Pour le cas de la régression logistique, l'estimateur β_0 est parfois affecté par l'erreur de mesure ajoutée à la variable de confusion et parfois non. Mais les paramètres β_1 et β_2 sont clairement influencés par la contamination de X_2.

Les résultats trouvés dans ce chapitre ont permis d'illustrer l'effet de l'erreur de mesure présente dans la variable de confusion (tabagisme) sur les résultats des modèles de régression linéaire et logistique. Ceci a conduit dans la plupart des cas à une augmentation de l'erreur de type I associée à la variable d'exposition (café) et donc à l'effet du café sur la variable d'intérêt (maladies cardio-vasculaires) et de confirmer ce que nous connaissons à priori, c'est-à-dire, le fait que le café est lié au tabagisme fait en sorte qu'il affecte sur les maladies cardio-vasculaires, ceci est aussi prouvé dans le cas de la régression linéaire par le fait que plus la corrélation entre X_1 et X_2 est forte, plus β_1 est sur-estimé. Pour la régression logistique, les résultats se sont avérés plus diversifiés, et contrairement à la régression linéaire, β_1 est sous-estimé quand β_2 est positif et $\hat{\beta}_1$ sur-estime β_1 seulement quand ce paramètre est nul. Et lorsque, la valeur de β_2 est positive et fixe, selon la variation de β_1, l'estimateur de ce dernier croît. On peut conclure aussi que ces résultats nous amènent à dire que si on ajoute une erreur artificielle dans une étude, on peut observer à quel point ceci a pour résultat d'augmenter la taille d'effet, ce qui nous permettra d'estimer le risque d'effet dans le modèle optimal.

CHAPITRE 5

Correction d'une erreur de mesure dans un modèle linéaire

5.1 Méthode de correction d'une erreur de mesure

Divers sont les approches statistiques utilisées pour contourner le problème de contamination d'une variable dans un modèle de régression, comme dans (Spiegelman et al., 1997) ou (Carroll & Stefanski, 1994).
Dans ce chapitre, sera proposée une méthode de correction du biais présent dans un modèle linéaire suite à la contamination de la variable de confusion par une erreur de mesure.

5.1.1 Démonstration mathématique

Après avoir évalué, dans un modèle de régression linéaire, l'effet d'une erreur de mesure dans une variable de confusion dans le chapitre précédent, on propose dans cette section, une méthode de correction du biais.

On considère un modèle linéaire avec deux variables indépendantes X_1 et X_2 tel que $\boldsymbol{cor}(X_1, X_2) = \rho$. Basé sur un échantillon aléatoire $(y_i, x_{1,i}, x_{2,i})$ pour $i = 1, ..., n$, où n est la taille de l'échantillon, le modèle linéaire qui est le plus répandu dans la littérature est donné par

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + e_i, \quad i = 1, ..., n \tag{5.1}$$

où e_i est une variable normale de moyenne zéro et de variance σ^2. En utilisant la méthode de l'estimation des moindres carrés des paramètres $(\beta_0, \beta_1, \beta_2)$, on obtient que :

$$\hat{\beta} = (\mathbf{X^{*T}X^*})^{-1}\mathbf{X^{*T}Y} \tag{5.2}$$

avec $\mathbf{Y} = (y_1, ..., y_n)'$ et

$$\mathbf{X}^* = \begin{pmatrix} 1 & X_{1,1} & X_{2,1} \\ \vdots & \vdots & \vdots \\ 1 & X_{1,n} & X_{2,n} \end{pmatrix}.$$

Dans beaucoup de situations, l'une des variables indépendantes n'est pas directement observée. Dans cette partie, on suppose qu'on va observer l'échantillon aléatoire $(y_i, x_{1,i}, x_{2,2,i})$ pour $i = 1, ..., n$ où $x_{2,i} = x_{2,2,i} + b$ et où b est une variable normale de moyenne zéro de variance égale à sig indépendante de X_1 et donc $\boldsymbol{cor}(X_{1,i}, X_{2,2,i}) = \rho$. On a aussi $\boldsymbol{E}(X_{1,i}) = \boldsymbol{E}(X_{2,2,i}) = 0$. On étudie l'effet de cette perturbation sur l'estimation des paramètres dans le modèle linéaire dans l'équation (5.1) qui s'écrit comme suit

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

où $\boldsymbol{Y} = (y_1, ..., y_n)'$, $= (\beta_0, \beta_1, \beta_2)'$, avec

$$\mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & X_{2,2,1} \\ \vdots & \vdots & \vdots \\ 1 & X_{1,n} & X_{2,2,n} \end{pmatrix} \quad et \quad \boldsymbol{\epsilon} = \begin{pmatrix} e_1 - \beta_2 b_1 \\ e_2 - \beta_2 b_2 \\ \vdots \\ e_n - \beta_2 b_n \end{pmatrix}.$$

Alors, l'estimateur du vecteur des paramètres $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ par la méthode des moindres carrés ordinaires est:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} \\ &= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\epsilon}.\end{aligned}$$

Par conséquent,

$$\lim_{n\to+\infty}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \lim_{n\to+\infty}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}].$$

Or,

$$\lim_{n\to+\infty}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}] = \lim_{n\to+\infty}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T e - \beta_2 \lim_{n\to+\infty}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T b].$$

D'un côté, d'après les hypothèses du modèle, on a:

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{e} \xrightarrow{P} 0.$$

D'un autre côté,

$$\begin{aligned}\lim_{n\to+\infty}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} &= -\beta_2 \lim_{n\to+\infty}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T b\right) \\ &= -\beta_2 \lim_{n\to+\infty}\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}^T b\right).\end{aligned}$$

D'une part,

$$
\begin{aligned}
\frac{1}{n}\boldsymbol{X}^T b &= \frac{1}{n}\begin{pmatrix} 1 & \cdots & 1 \\ X_{1,1} & \cdots & X_{1,n} \\ X_{2,2,1} & \cdots & X_{2,2,n} \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \\
&= \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n b_i \\ \sum_{i=1}^n b_i X_{1,i} \\ \sum_{i=1}^n b_i X_{2,2,i} \end{pmatrix} = \begin{pmatrix} \mathbb{E}(b) \\ \mathbb{E}(bX_1) \\ \mathbb{E}(bX_{2,2}) \end{pmatrix} \\
&\xrightarrow{P} \begin{pmatrix} 0 \\ 0 \\ \boldsymbol{cov}(x_{2,2}, b) \end{pmatrix}.
\end{aligned}
$$

Or d'autre part,

$$
\begin{aligned}
A_n &\equiv \left(\frac{1}{n}(X^T X)^{-1}\right) \\
&= \left[\frac{1}{n}\begin{pmatrix} 1 & \cdots & 1 \\ X_{1,1} & \cdots & X_{1,n} \\ X_{2,2,1} & \cdots & X_{2,2,n} \end{pmatrix}\begin{pmatrix} 1 & X_{1,1} & X_{2,2,1} \\ \vdots & \vdots & \vdots \\ 1 & X_{1,n} & X_{2,2,n} \end{pmatrix}\right]^{-1} \\
&= \left[\frac{1}{n}\begin{pmatrix} n & \sum X_{1,i} & \sum X_{2,2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,2,i} \\ \sum X_{2,2,i} & \sum X_1 X_{2,2,i} & \sum X_{2,2,i}^2 \end{pmatrix}\right]^{-1} \\
&\xrightarrow{P} \begin{pmatrix} 1 & E(X_1) & E(X_{2,2}) \\ E(X_1) & E(X_1^2) & E(X_1 X_{2,2}) \\ E(X_{2,2}) & E(X_1 X_{2,2}) & E(X_{2,2}^2). \end{pmatrix}^{-1}.
\end{aligned}
$$

Mais puisque, par l'hypothèse, $E(X_1) = E(X_{2,2}) = 0$

$$A_n \xrightarrow{P} A \equiv \begin{pmatrix} 1 & 0 & 0 \\ 0 & E(X_1^2) & E(X_1 X_{2,2}) \\ 0 & E(X_1 X_{2,2}) & E(X_{2,2}^2) \end{pmatrix}^{-1} . \tag{5.3}$$

Un simple calcul donne

$$A = \frac{1}{det} \begin{pmatrix} E(X_1^2)E(X_{2,2}^2) - E((X_1 X_{2,2})^2) & 0 & 0 \\ 0 & E(X_{2,2}^2) & -E(X_1 X_{2,2}) \\ 0 & -E(X_1 X_{2,2}) & E(X_{2,2}^2) \end{pmatrix}$$

où $det = E(X_1^2)E(X_{2,2}^2) - E((X_1 X_{2,2})^2)$.

On peut donc conclure que,

$$\begin{aligned} (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\epsilon \quad &\xrightarrow{P} \quad \frac{-\beta_2}{det} \begin{pmatrix} 1 & 0 & 0 \\ 0 & E(X_1^2) & E(X_1 X_{2,2}) \\ 0 & E(X_1 X_{2,2}) & E(X_{2,2}^2) \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ \boldsymbol{cov}(X_{2,2}, b) \end{pmatrix} \\ &= \quad \frac{-\beta_2}{det} \begin{pmatrix} 0 \\ -E(X_1 X_{2,2})\boldsymbol{cov}(X_{2,2}, b) \\ E(X_1^2)\boldsymbol{cov}(X_{2,2}, b) \end{pmatrix}. \end{aligned}$$

On note par la suite B et C les quantités définies par : $B = E(X_1 X_{2,2})\boldsymbol{cov}(X_{2,2}, b)$ et

$C = E(X_1^2)\boldsymbol{cov}(X_{2,2}, b)$. Donc,

$$\begin{aligned}\hat{\beta} &\xrightarrow{P} \beta - \frac{\beta_2}{det}\begin{pmatrix} 0 \\ B \\ C \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -\frac{B}{det} \\ 0 & 0 & 1-\frac{C}{det} \end{pmatrix}}_{M}\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.\end{aligned}$$

Par conséquent, on voit clairement que $\hat{\beta}$ est un estimateur biaisé donc n'est pas consistant sauf si B et C sont nuls. Cependant, $\hat{\beta}_0$ reste un estimateur consistant de β_0 même dans cette situation. En fait, l'erreur de mesure a affecté l'estimation de β_1 et β_2 uniquement. Dans ce travail, nous proposons une simple correction de β afin d'avoir un estimateur convergent de β dans une telle situation. Cette méthode consiste à multiplier l'estimateur $\hat{\beta}$ par l'inverse de la matrice M. On note par $\hat{\beta}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_2^*)$ le nouvel estimateur corrigé de β. En effet, pour estimer β, on propose:

$$\hat{\beta}^* = M^{-1}\hat{\beta}$$

avec,

$$M^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{B}{det-C} \\ 0 & 0 & \frac{det}{det-C} \end{pmatrix}.$$

Remarques

$\hat{\beta}^* = \hat{\beta} \iff M^{-1} = I \iff B = 0$ et $C = 0$.

$$\begin{aligned} B = 0 &\iff \boldsymbol{E}(X_1 X_{2,2})\boldsymbol{cov}(X_{2,2}, b) = 0 \\ &\iff \boldsymbol{E}(X_1 X_{2,2}) = 0 \quad ou \quad \boldsymbol{cov}(X_{2,2}, b) = 0 \\ &\iff b = 0 \end{aligned}$$

$$\begin{aligned} C = 0 &\iff \boldsymbol{E}(X_1^2)\boldsymbol{cov}(X_{2,2}, b) = 0 \\ &\iff \boldsymbol{E}(X_1^2) = 0 \quad ou \quad \boldsymbol{cov}(X_{2,2}, b) = 0 \\ &\iff b = 0 \end{aligned}$$

5.1.2 Simulations et résultats

Dans cette section, nous appliquons la correction du biais calculé précédemment dans la démonstration mathématique et nous obtenons les estimations ajustées suite à cette modification.
Les simulations ont été faites à l'aide du logiciel R. Nous avons généré le modèle linéaire cité dans le chapitre 4 (4.1) où X_1 et X_2 sont deux variables normales. On calcule par la suite la valeur du nouvel estimateur $\hat{\beta}$ pour chaque valeur de sig en variant les paramètres du modèle. Les figures ci-dessous représentent les variations de ces estimateurs selon la variation de sig. La taille de l'échantillon est égale à 1000.

La figure (5.1) illustre les *boxplots* représentant la variation de $(\hat{\beta}_0^* - \beta_0^*)$, $(\hat{\beta}_1^* - \beta_1^*)$ et $(\hat{\beta}_2^* - \beta_2^*)$ respectivement sur (a), (b) et (c). D'après la figure (5.1), on voit clairement

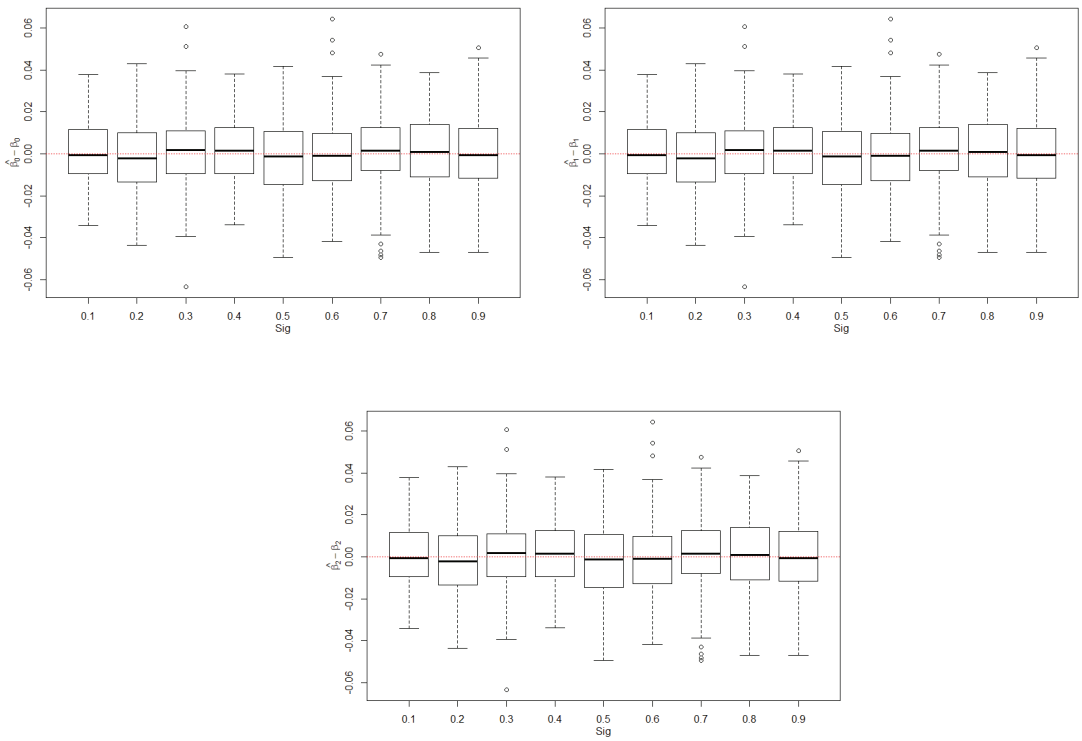

Figure 5.1: **Évaluation des estimateurs après correction**:Variation de $(\hat{\beta}_0^* - \beta_0^*)$ dans (a), $(\hat{\beta}_1^* - \beta_1^*)$ dans (b) et $(\hat{\beta}_2^* - \beta_2^*)$ dans (c) sur l'axe des ordonnées selon *sig* sur l'axe des abscisses pour 200 répétitions pour le modèle linéaire (5.1) avec avec $\tau = (1000, 0, 2, 0.5, 0.1)$.

que l'estimateur proposé est sans biais et n'est pas affecté par l'erreur de mesure.

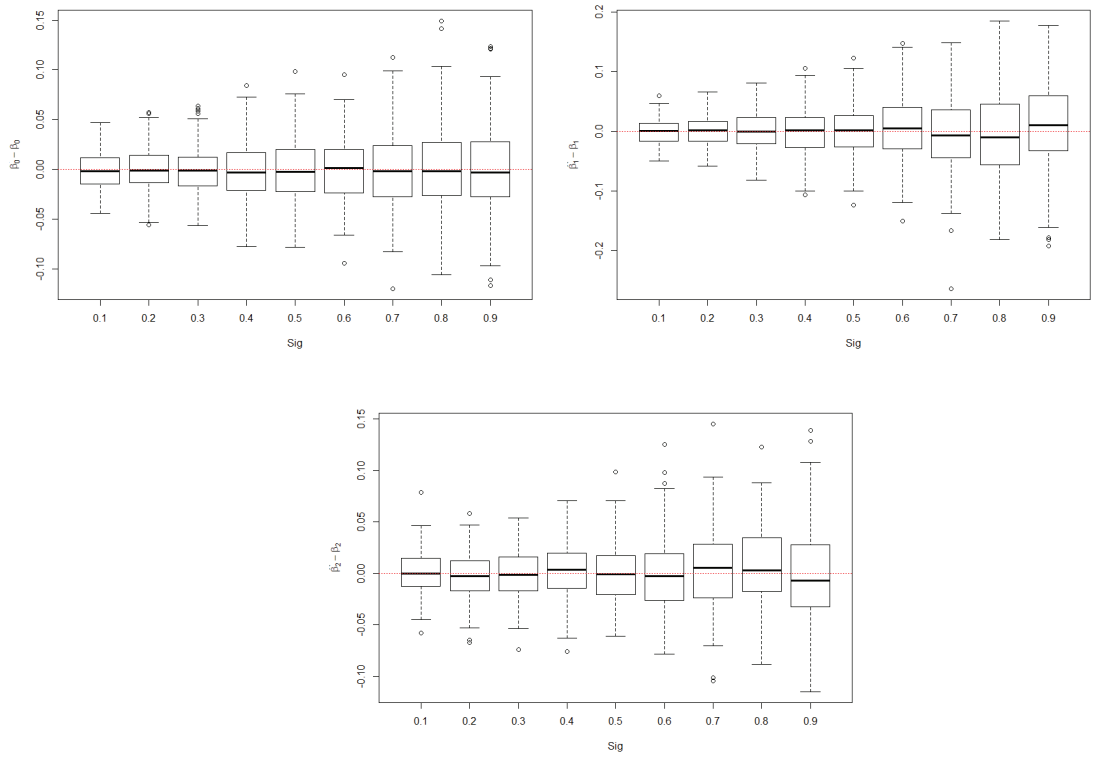

Figure 5.2: **Évaluation des estimateurs après correction**:Variation de $(\hat{\beta}_0^* - \beta_0^*)$ dans (a), $(\hat{\beta}_1^* - \beta_1^*)$ dans (b) et $(\hat{\beta}_2^* - \beta_2^*)$ sur (c) sur l'axe des ordonnées selon *sig* sur l'axe des abscisses pour 200 répétitions pour le modèle linéaire (5.1) avec avec $\tau = (1000, 4, -2, 2, 0.5)$.

Dans la figure (5.2), on fait la même remarque que dans (5.1) concernant le biais tandis qu'on constate que la variance augmente quand *sig* augmente. D'autres configurations ont été étudiées et ont donné les mêmes résultats, ce qui peut nous mener à une généralisation du résultat obtenu.

5.1.3 Cas de variance de l'erreur de mesure inconnue

Bien que dans la correction du biais proposé auparavant, nous nous sommes penchés sur le cas d'une erreur de mesure qui suit une distribution normale de paramètres connus, nous nous intéressons dans cette section au cas échéant, c'est à dire, si l'accès aux distributions sous-jacentes est impossible, l'estimation des corrections est-elle possible?
(Delaigle et al., 2008) a traité la problématique des erreurs de mesures dans les modèles de régression et a montré que dans le cas d'erreurs de mesure avec distribution inconnue, la solution n'est possible que dans le cas de plusieurs répétitions, c'est à dire si on observe

$$y_{jk} = x_j + e_{jk} \quad pour 1 \leq k \leq N_j \quad et \quad 1 \leq j \leq n \tag{5.4}$$

pour un y_{jk} observé, on doit observer x_j N_j fois.
Dans ce cas, non seulement la variance de l'erreur peut être estimée mais aussi la distribution entière de l'erreur.

Chapitre 6

Conclusion

L'un des biais pouvant affecter les études en épidémiologie est le biais de confusion. Dans certaines études, la présence de ce biais peut passer inaperçue et n'affecte en rien les résultats, dans d'autres, une confusion , aussi minime soit-elle peut fausser complètement les résultats si bien que certains épidémiologistes préfèrent contrôler ce facteur de confusion au niveau de la planification de l'étude et au niveau de l'analyse pour assurer des résultats fiables.

Dans ce travail, nous avons pu étudier cette problématique en utilisant deux méthodes. D'un coté, la régression linéaire, qui, à travers les simulations faites dans le chapitre 4, nous avons pu conclure que selon la variation des paramètres du modèle proposé, le biais est souvent sur-estimé ou sous-estimé tout en remarquant que plus les variables sont corrélées, plus le biais est robuste. Toutes ces constatations nous ont mené à chercher une correction adéquate qui vise à corriger ce biais, chose qui a été faite au cours du chapitre 5.

D'un autre côté, on a étudié le cas de la régression logistique, qui s'est avéré plus délicate à traiter vu sa complexité par rapport à la régression linéaire. Les résultats obtenus d'après les simulations faites ont confirmé ceci et le comportement des estimateurs s'est montré différent dans le cas de la régression linéaire malgré quelques similitudes.

Dans cette étude, notre intérêt était concentré sur l'erreur de mesure associée à la variable de confusion. Néanmoins, il est important de mettre l'accent sur la possibilité de la présence d'une erreur de mesure qui peut affecter la variable d'exposition et qui n'a pas été prise en compte dans ce travail. Il est donc possible que cette erreur, qu'elle soit faible ou forte, affecte les résultats du modèle aussi. Une perspective future pourra traiter ce cas.

À travers le chapitre 5, nous avons pu présenter une méthode simple et efficace pour corriger le biais de la régression linéaire et ce dans le cas de la connaissance des distributions sous-jacentes.

Dans le futur, les résultats trouvés peuvent être appliqués dans une étude de cas réel en présence de variables de confusion. La correction du biais de confusion proposé peut être testé sur un vrai échantillon de population ayant des distributions d'erreurs de mesures connues, le cas contraire se présente plus difficile à traiter.

Références

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Austin, P, C. & Brunner, L, J. (2004). Inflation of the type 1 error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics In Medecine* **23**, 1159–1178.

Austin, P, C. & Brunner, L, J. (2009). Inflation of the type 1 error rate in multiple regression when independent variables are measured with error. *The Canadian Journal Of Statistics* **37(1)**, 33–46.

Axelson, O. (1985). Dealing with the exposure variable in occupational and environmental epidemiology. *Scand J Soc Med* **13(4)**, 147–52.

Barnwell, J, L., Li, Q. & Cohen, A, A. (2014). Effects of categorization method regression type, and variable distribution on the inflation of type i error rate when categorizing a confounding variable. *Statistics In Medecine* **34(6)**, 936–949.

Berkson (1944). Application of the logistic function to bio-essay. *Journal Of The American Statistical Association* **39**, 357–365.

Berkson (1950). Are there two regressions? *Journal of the American Statistical Association* **45(250)**, 164–180.

Carroll, R, J. & Stefanski, L, A. (1994). Measurement error instrumental variables and corrections for attenuation with applications to metaanalysess. *Statistics In Medecine* **13**, 1265–1282.

Chen, C, L., Gilbert, T, J. & R, Daling, J. (1999). Maternal smoking and down syndrome: The confounding effect of maternal age. *Am. J. Epidemiol* **149(5)**, 442–446.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*. Springer Texts in Statistics, 2nd ed.

Cornillon, P. & Matzner-Lober, E. (2007). Regression avec r. *Springer* .

Cramer, J, S. (2002). The origins of logistic regression. *Tinbergen Institute Discussion Paper* .

Czepiel, S, A. (2010). Maximum likelihood estimation of logistic regression models: Theory and implementation. *Available at http://czep.net/stat/mlelr.pdf* .

Delaigle, A., Hall, P. & Meister, A. (2008). On deconvolution with repeated measurements. *The Annals Of Statistics* **36(2)**, 665–685.

Desjardins, J. (2005). L'analyse de régression logistique. *Tutorial in Quantitative Methods for Psychology* **1(1)**, 35–41.

Gourieroux, C. & Monfort, A. (1981). Assymptotics properties of the maximimum likelihood estimator in dichotomous logit models. *Journal of Econometrics* **17**, 83–97.

Huff, D. (1993). *How to Lie with Statistics*. Norton and Company.

Joseph, L. (2010). Confounding and collinearity in multivariate logistic regressiong. Available at : http://www.medicine.mcgill.ca/epidemiology/joseph/courses/epib-621/logconfound.pdf.

LaMorte, W, W. & Sullivan, L. (2014). Residual confounding, confounding by indication and reverse causality. Boston University School of Public Health.

Last, J, M. (1995). *A dictionary of Epidemiology*. Oxford University Press, 3rd ed.

Lindsay, L. (2014). Facteurs de confusion. Available online.

Olsen, J. & Basso, O. (1989). Re:residual confounding. vol. 149(3). American Journal Of Epidemiology.

Palm, R. & Iemma, A, F. (1995). Quelques alternatives à la régression classique dans le cadre de la colinéarité. *Revue de Statistique Appliquée* **43(2)**, 5–33.

Preux, P, M., Odermatt, P., Perna, A., Marin, B. & Vergnenègre, A. (2014). Qu'est-ce qu'une régression logistique ? *Revue des Maladies Respiratoires, Elsevier* **22(1)**, 159–162.

Rothman, K, J. (2002). *Epidemiology: An Introduction*. USA: Oxford University Press.

Russel, D. & Mackinnon, J, G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press.

Selme, M, O. & Masson, G. (2013). Décision et prévision statistique. Available at : http://tice.inpl-nancy.fr/modules/unit-stat/.

Shawky, S. (2000). Potential errors in epidemiologic studies random error .

Shedden, K. (2014). Specification errors measurement errors and confounding. Available online.

Spiegelman, D., McDermott, A. & Rosner, B. (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *American Journal Clinical Nutrition* **65(4)**, 1179–1186.

Tabachnick, B. & Fidell, L. (2000). *Using Multivariate Statistics*. Allyn and Bacon., 4th ed.

Tibshirani, R. (1996). Regression shrinkage and selection. *Journal of the Royal. Statistical society* **58(1)**, 267–288.

Wiki.stat (2013). In *Regression Logistique ou Modèle Binomial*. Institut de mathématiques de Toulouse.

Worster, A., Fan, J. & Ismaila, A. (2007). Understanding linear and logistic regression analyses. *Cjiem* **9(2)**, 111–113.