REPUBLIQUE DU CAMEROUN
Paix-Travail-Patrie
****
UNIVERSITE DE YAOUNDE I
****
ECOLE NORMALE SUPERIEURE DE YAOUNDE
****
DEPARTEMENT DE MATHEMATIQUES
****

REPUBLIC OF CAMEROON
Peace-Work-Fatherland
****
UNIVERSITY OF YAOUNDE I
****
HIGHER TEACHER TRAINING COLLEGE
****
DEPARTMENT OF MATHEMATICS
****

N

E S

Sur-dispersion dans les modèles de régression logistique

Mémoire de DI.P.E.S. II de Mathématiques

De

SAMADINE Djallo
Matricule : 02Y640

Sous la direction de :
Dr NGUEFACK –TSAGUE Georges
Chargé de cours
Université de Yaoundé I

ANNEE ACADEMIQUE: 2015-2016

♣ Dédicace ♣

A mon papa :

DJALLO HAMAN

♣ Remerciements ♣

Ce mémoire est le fruit d'efforts et de soutiens convergents et je voudrais ici rendre hommage aux principaux acteurs :

1. Je remercie vivement mon Directeur de mémoire ; Dr. Georges NGUEFACK-TSAGUE, pour sa disponibilité, son soutien sans relâche, ainsi que ses conseils ;
2. Je tiens aussi à remercier le Chef du Département de Mathématiques ; Pr. Lawrence DIFFO LAMBO et tous les enseignants du Département de Mathématiques de l'Ecole Normale Supérieure de Yaoundé pour leurs encadrements durant ma formation ;
3. Ma gratitude va à l'endroit de mes grands frères le **Pr. HAMAN DJALO** ; l'ingénieur **VAGAÏ DJALLO** et **NDODI DJALLO** pour leur constant soutien ;
4. Mes remerciements vont aussi à l'endroit de ma chère épouse MECKONG Tatiana Fabrice pour son soutien et sa compréhension ;
5. Mes remerciements vont aussi à l'endroit de mes deux bambinos **Josepha** et **Francky** de qui j'ai puisé cette force qui m'a permis de tenir ;
6. Je voudrais également ici exprimer ma gratitude à la grande famille **DJALLO HAMAN** de Salak, tous mes frères et soeurs pour leurs soutiens ;
7. Je remercie toutes mes mamans DJIVID ; AMBA ; TSINABI et les autres épouses de Papa pour leur amour et leurs encouragements ;
8. Je trouve ici le lieu d'exprimer ma réconnaissance à tous mes amis tels que ESSOMBA ; TCHOPMO ; NGUETIO ; TSAFACK Eric ; ***MENANA MBIDA Damien*** ; **FOTUE TABUE Alexandre** et tous mes camarades de promotion pour les merveilleux moments que nous avons passé ensemble.

♣ Déclaration sur l'honneur ♣

Le présent travail est une oeuvre originale du candidat et n'a été soumis nulle part ailleurs, en partie ou en totalité, pour une autre évaluation académique. Les contributions externes ont été dûment mentionnées et recensées en bibliographie.

Signature du candidat

SAMADINE DJALLO

♣ Table des matières ♣

♣ Liste des abréviations ♣

1. **V.A** : variable aléatoire
2. **E.M.V** : Estimateur du maximum de vraisemblance

♣ Liste des figures ♣

♣ Liste des tableaux ♣

♣ Résumé ♣

La sur-dispersion est un phénomène régulièrement rencontré dans des analyses statistiques. Très souvent connue dans les modèles de Poisson, la sur-dispersion est généralement non prise en compte dans les analyses des modèles binaires. L'impact de la prise en compte de la présence de la sur-dispersion sur les résultats des tests des analyses binaires n'est cependant pas à négliger. Il est donc question pour nous de présenter tout d'abord les modèles de régression logistique et les différents tests qui s'appliquent sans la prise en compte de la présence de la sur-dispersion ; ensuite nous allons mettre en évidence la présence de cette sur-dispersion en déterminant le paramètre de sur-dispersion et réeffectuer les tests en tenant compte de ce paramètre de sur-dispersion. Enfin nous allons comparer les résultats afin de voir les améliorations apportées par la prise en compte de ce paramètre. Par exemple, l'inflation de la variance des coefficients du modèle peut être parfois de l'ordre de 80% pour des échantillons de grandes tailles. Pour terminer, nous avons illustré les conséquences de la présence de la sur-dispersion à travers une étude pratique.

mots clés : régression logistique ; sur-dispersion ; analyse statistique ; inflation de la variance.

♣ Abstract ♣

Overdispersion is often overlooked in statistical analyses. Sometimes not taken into account, its impact in the tests results is not however to be neglected. The work aims to study the impact of the presence of this overdispersion in the logistic regression. In there, we first present the logistic regression and the various tests applied without taking account the presence of overdispersion. Then to determine the overdispersion parameter and to remake all these tests by taking account the overdispersion parameter, then to compare the results to see the precise details brought by the taking into account of this parameter. Sometimes, the inflation of standard error can be to 80%. Finally, we have illustrated the consequences of the overdispersion presence through a practical study.

Keywords : logistic regression ; statisticals analysis ; overdispersion ; inflation of standard error

♣ Introduction générale ♣

Les modèles de régression logistique font partie des membres des modèles d'analyse de la grande famille des modèles linéaires généralisés qui permettent d'étudier la liaison entre une variable dépendante appelée **variable réponse Y** et un ensemble de variables dépendantes ou indépendantes entre elles appelées **variables explicatives** ou **prédicteurs** $\mathbf{X}_1; ...; \mathbf{X}_p$. Ils englobent :

— le modèle linéaire qui comprend la régression linéaire simple et multiple permet l'analyse de la variance et l'analyse de la co-variance ;

— le modèle log −linéaire ;

— la régression logistique

— le modèle de Poisson.

Les modèles linéaires généralisés sont formés de trois composantes :

— **la variable de réponse Y**, composante aléatoire à laquelle est associée une loi de probabilité ;

— **les variables explicatives** $\mathbf{X}_1; ...; \mathbf{X}_p$ utilisées comme prédicteurs dans le modèle, définissent sous forme d'une combinaison linéaire **la composante déterministe ;**

— **le lien** décrit la relation fonctionnelle entre la combinaison linéaire des variables $X_1; ...; X_p$ et l'espérance mathématique de la variable de **réponse Y**.

La régression logistique est couramment utilisée en épidémiologie (voir [1], [2], [6]) ; en économie, [17] et en biométrie [11]. Elle permet d'étudier la nature de la rélation qui existe entre une variable aléatoire dichotomique et une ou plusieurs autres variables indépendantes encore appelées variables explicatives. Son emploi, rendu aisé par l'utilisation de logiciels statistiques, permet le contrôle des biais de confusion. La mesure d'association calculée dans ce modèle est l'***odds-ratio*** (ou rapport de cotes en français), qui quantifie la force de l'association entre la survenue d'un événement, représentée par la variable dichotomique, et les facteurs susceptibles

de l'influencer, représentés par des variables explicatives. Le choix des variables explicatives intégrées au modèle repose sur une connaissance préalable du phénomène étudié afin de ne pas omettre des facteurs de confusion déjà identifiés. Ainsi, le modèle logistique utilisé pour l'analyse d'un phénomène doit être basé sur des hypothèses et des connaissances du réseau de "causalité" qui se tisse autour du phénomène.

Lorsqu'une distribution suit une **loi de Poisson** ou **une loi Binomiale**, la connaissance de l'espérance mathématique de cette variable aléatoire permet de déterminer sa variance. Des études récentes ont montré que lorsqu'on adopte le modèle de régression logistique pour effectuer l'analyse sur des données groupées voir [3] ou bien si on l'adopte pour effectuer l'analyse sur des données corrélées voir [4], la variance de la variable aléatoire calculée dans ces conditions ne respecte plus automatiquement les propriétés de la variance de la loi binomiale dont elle est issue. La sur-dispersion décrit un phénomène où la variance calculée est supérieure à la variance prévue par le modèle théorique considéré. Le plus souvent non prise en compte dans les analyses des données binaires, la prise en compte de la présence de la sur-dispersion dans les analyses a cependant un important impact sur les différents résultats trouvés. Notre travail consiste à étudier l'influence de la prise en compte de la présence de cette sur-dispersion sur les résultats des tests d'analyses à l'aide des modèles de régression logistique, puis apporter des solutions aux différents problèmes causés par la présence de la sur-dispersion et enfin présenter des méthodes d'analyses qui prennent en compte la présence de cette sur-dispersion. Le présent mémoire comporte trois chapitres : le **Chapitre 1** est consacré à la présentation du modèle de régression logistique ; à la description des algorithmes permettant d'estimer les coefficients d'un modèle de régression logistique ; à présenter les différents tests effectués lors d'analyses à l'aide d'un modèle de régression logistique ; le **Chapitre 2** décrit les méthodes permettant d'indiquer la présence de la sur-dispersion en régression logistique, les méthodes d'estimation du paramètre de sur–dispersion et présente la notion d'estimation d'equations généralisées. Ici, sont aussi présentées les méthodes qui permettent de prendre en compte la présence de la sur-dispersion dans le modèle logistique lors des analyses des données. Le **Chapitre 3** nous avons effectué une analyse pratique pour mettre en évidence l'impact de la prise en compte de la présence de la sur-dispersion sur la variation de l'écart-type des coefficients du modèle logistique. Et enfin, dans le **chapitre** 4 de ce mémoire, nous avons donné les implications didactiques de notre thème.

Chapitre 1

Variables aléatoires conditionnelles et modèles de régression

Définition **1.1.** Soient $\mathbf{Y}$ une variable aléatoire et $\mathbf{X}$ un vecteur de variables qui peuvent être dépendantes ou indépendantes entre elles. La variable aléatoire $\mathbf{Y}$ est dite conditionnelle par rapport aux coordonnées de $\mathbf{X}$ si sa réalisation ou non peut être expliquée à l'aide des coordonnées de $\mathbf{X}$.

On dit alors que $\mathbf{Y}$ est une **variable à expliquer** et que les coordonnées de $\mathbf{X}$ sont les **variables explicatives**.

1.1 Notation

— $X = (1; X_1; ...; X_p)$: le vecteur de dimension $p+1$ où les $X_j \quad j = 1; ...; p$ sont les variables explicatives. On notera par $x = (1; x_1; ...; x_p)$ une réalisation de X;

— $\mathbf{Y}$ est la variable à expliquer,

— $f(\mathbf{Y}|X; \theta)$ représente la fonction de probabilité ou la densité de probabililté de la variable aléatoire Y et θ est le vecteur des paramètres à estimer,

— N assez grand, représente le nombre d'individus observés pour effectuer l'analyse,

— $(X_1; Y_1); ...; (X_N; Y_N)$: un $N-$échantillon aléatoire (indépendant et identiquement distribué (i.i.d) et de même loi que $(X; Y)$) tel que $X_i = (1; X_{i1}; ...; X_{ip})$,

— $(x_1; y_1); ...; (x_N; y_N)$ est une réalisation de $(X_1; Y_1); ...; (X_N; Y_N)$,

— X : la matrice des observations :

$$\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{pmatrix}$$

1.2 La vraisemblance d'un échantillon

Soit $\mathbf{Y}$, une variable aléatoire à expliquer admettant une distribution conditionnelle par rapport aux coordonnées de la variable explicative X et dont la **densité conditionnelle** ou **fonction de probabilité conditionnelle** est notée $f(\mathbf{Y}|X;\theta)$; que nous noterons parfois par $f(X;\theta)$ et où θ est le paramètre que l'on cherche à estimer. On dispose d'un échantillon de N observations indépendantes et identiquement distribuées.

1.2.1 Définitions et théorème

Définition 1.2. La vraisemblance de l'échantillon, représentant la probabilité d'observer l'ensemble de l'échantilon, notée L, est définie par :

$$L(\mathbf{Y}|\mathbf{X};\theta) = \prod_{i=1}^{N} (f(y_i|x_i;\theta)) \tag{1.1}$$

La log-vraisemblance de cet échantillon, notée $\mathcal{L}$, est égale à son logarithme népérien et est :

$$\mathcal{L}(\mathbf{Y}|\mathbf{X};\theta) = \sum_{i=1}^{N} \ln(f(y_i|x_i;\theta)) \tag{1.2}$$

Définition 1.3. On appelle **estimateur du maximum de vraisemblance de l'échantillon** une valeur $\widehat{\theta} \in \Theta$; s'il en existe une ; telle que

$$L\left(Y|X;\widehat{\theta}\right) \succeq L(Y|X;\theta) \qquad \forall\theta \in \Theta \tag{1.3}$$

où Θ est l'ensemble des valeurs admissibles du paramètre θ. Cet estimateur sera parfois noté par **E.M.V**. Notons que l'inégalité (1.3) est encore équivalente à :

$$\mathcal{L}\left(Y|X;\widehat{\theta}\right) \succeq \mathcal{L}(Y|X;\theta) \qquad \forall\theta \in \Theta \tag{1.4}$$

de sorte que l'on puisse maximiser la log-vraisemblance $\mathcal{L}(Y|X;\theta)$ pour déterminer l'**E.M.V.** au lieu de maximiser la vraisemblance $L(Y|X;\theta)$.

Définition 1.4. On appelle biais de l'estimateur $\widehat{\theta}$ pour θ la valeur :

$$b_\theta\left(\widehat{\theta}\right) = E_\theta\left(\widehat{\theta}\right) - \theta.$$

Si $b_\theta\left(\widehat{\theta}\right) = 0$ quel que soit $\theta \in \Theta$, on dit que $\widehat{\theta}$ est sans biais pour θ.

Le **théorème** ci-dessous nous assure de la convergence d'un algorithme itératif vers $\widehat{\theta}$ et nous guide sur le comportement asymptotique de l'estimateur du maximum de vraisemblance (voir [6]). Considérons le jeu d'hypothèses suivant :

— H_1 : rang(X) $= p + 1$

— H_2 : on est en situation de recouvrement

— H_3 : la matrice **E**[XX'] existe et est définie positive

Théorème 1.1. –Sous les hypothèses H_1 et H_2, la log −vraisemblance $\theta \longmapsto \mathcal{L}_N(\theta)$ est strictement concave : $\widehat{\theta}$ existe et est unique.

–Sous H_3, $\widehat{\theta}$ converge vers θ lorsque $N \longrightarrow +\infty$.

1.2.2 Quelques propriétés de l'Estimateur du Maximum de Vraisemblance

Soient $N \in \mathbb{N}$ et $f(x;\theta)$, la fonction de probabilité de la variable aléatoire conditionnelle X vérifiant les conditions de régularité suivantes :

— $\mathbf{I}(\theta)$ existe pour tout $\theta \in \Theta$

— La dérivée par rapport à θ d'une intégrale sur la densité conjointe

$$\int ... \int_{\mathbb{R}} f(x_1, x_2, ..., x_N; \theta)\, dx_1 dx_2 ... dx_N$$

peut s'obtenir en dérivant à l'intérieur de l'intégrale

— La dérivée par rapport à θ de $\mathrm{E}_\theta\left(\widehat{\theta}\right)$ peut s'obtenir en dérivant à l'intérieur de l'intégrale correspondante

— Le support de $f(x;\theta)$ est indépendant de θ.

Proposition 1.1. Soit $f(Y|X;\theta)$ la densité de probabilité conditionnelle de la variable expliquée Y. Elle vérifie :

$$E_\theta\left[\frac{\partial \ln f(y|x;\theta)}{\partial \theta}\right] = 0$$

Preuve.

$$\begin{aligned}
E_\theta\left[\frac{\partial \ln f(y|x;\theta)}{\partial \theta}\right] &= E_\theta\left[\frac{1}{f(y|x;\theta)}\frac{\partial f(y|x;\theta)}{\partial \theta}\right] \\
&= \int \frac{1}{f(y|x;\theta)}\frac{\partial f(y|x;\theta)}{\partial \theta} f(y|x;\theta)\, dy \\
&= \int \frac{\partial f(y|x;\theta)}{\partial \theta} dy \\
&= \frac{\partial}{\partial \theta}\underbrace{\int f(y|x;\theta)\, dy}_{1} \\
&= 0
\end{aligned}$$

■

Proposition 1.2. Soit $f(Y|X;\theta)$ la densité conditionnelle de la variable expliquée ; elle vérifie :

$$E_\theta\left[\frac{\partial \ln f(y|x;\theta)}{\partial\theta}\cdot\frac{\partial \ln f(y|x;\theta)}{\partial\theta}\right]=E_\theta\left[-\frac{\partial^2 \ln f(y|x;\theta)}{\partial\theta^2}\right]$$

Preuve. Posons

$$U=\frac{\partial \ln f(y|x;\theta)}{\partial\theta}=\frac{\frac{\partial}{\partial\theta}f(y|x;\theta)}{f(y|x;\theta)}.$$

On a

$$\begin{aligned}\mathbf{E}_\theta(U) &= \int_{\mathbb{R}}\frac{\frac{\partial}{\partial\theta}f(y|x;\theta)}{f(y|x;\theta)}f(y|x;\theta)\,dy=\int_{\mathbb{R}}\frac{\partial}{\partial\theta}f(y|x;\theta)\,dy\\ &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}}f(y|x;\theta)\,dy=0\end{aligned}$$

car $\int_{\mathbb{R}}f(y|X;\theta)\,dy=1$

De plus,

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\ln f(y|x;\theta) &= \frac{\frac{\partial^2}{\partial\theta^2}f(y|x;\theta).f(y|x;\theta)-\left[f(y|x;\theta)\right]^2}{\left[f(y|x;\theta)\right]^2}\\ &= \frac{\frac{\partial^2}{\partial\theta^2}f(y|x;\theta)}{f(y|x;\theta)}-\left[\frac{\frac{\partial}{\partial\theta}f(y|x;\theta)}{f(y|x;\theta)}\right]^2\end{aligned}$$

D'où

$$\begin{aligned}\mathbf{E}_\theta\left[\frac{\partial \ln f(y|x;\theta)}{\partial\theta}\cdot\frac{\partial \ln f(y|x;\theta)}{\partial\theta}\right] &= \mathbf{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f(y|x;\theta)\right]\\ &= \mathbf{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(y|x;\theta)}{f(y|x;\theta)}\right]-\mathbf{E}_\theta\left[U^2\right].\end{aligned}$$

Or

$$\mathbf{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(y|x;\theta)}{f(y|x;\theta)}\right]=\int_{\mathbb{R}}\frac{\partial^2}{\partial\theta^2}f(y|x;\theta)\,dy=\frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}}f(y|x;\theta)\,dy=0$$

D'où

$$\mathbf{E}_\theta\left[\frac{\partial \ln f(y|x;\theta)}{\partial\theta}\cdot\frac{\partial \ln f(y|x;\theta)}{\partial\theta}\right]=-\mathbf{E}_\theta\left[\frac{\partial^2 \ln f(y|x;\theta)}{\partial\theta^2}\right].$$

■

Proposition 1.3. Sous les hypothèses de régularité ci-dessus que nous supposerons vérifiées par la suite, les estimateurs du maximum de vraisemblance sont convergents, asymptotiquement normaux et asymptotiquement efficaces (c'est-à-dire à variance minimale parmi les estimateurs sans biais) :

$$\sqrt{N}\left(\widehat{\theta}_N-\theta\right)\underset{N\to+\infty}{\xrightarrow{L}}\mathcal{N}\left(0;I_1^{-1}(\theta)\right)$$

où $I_1(\theta)$ est la matrice d'information de Fisher définie par :

$$I_1(\theta) = \mathbf{E}_\theta\left[\left(\frac{\partial \ln f(y|x;\theta)}{\partial\theta}\right)^2\right] = \mathbf{E}_\theta\left[\frac{\partial \ln f(y|x;\theta)}{\partial\theta}\cdot\frac{\partial \ln f(y|x;\theta)}{\partial\theta}\right]$$

De plus, en vertu de l'égalité de la matrice d'information, on peut aussi utiliser la matrice $\mathrm{J}_1(\theta)$:

$$J_1(\theta) = \mathbf{E}_\theta\left[-\frac{\partial^2 \ln f(y|x;\theta)}{\partial\theta^2}\right]$$

car $J_1(\theta) = I_1(\theta)$. La distribution de $\widehat{\theta}_N$ peut être approximée par :

$$\widehat{\theta}_N \overset{\mathbf{A}}{\rightsquigarrow} \mathcal{N}\left(\theta; \frac{1}{N}\times I_1^{-1}\left(\widehat{\theta}_N\right)\right) = \mathcal{N}\left(\theta; I_N^{-1}\left(\widehat{\theta}_N\right)\right)$$

où $\overset{\mathbf{A}}{\rightsquigarrow}$ désigne la distribution asymptotique (signifie utilisable pour les grands échantillons).

Preuve. Remarquons tout d'abord que la matrice d'information de l'ensemble de l'échantillon est définie par $I_N(\theta) = N \times I_1(\theta)$; on a donc

$$\frac{1}{N} I_1^{-1}(\theta) = (N \times I_1(\theta))^{-1} = I_N^{-1}(\theta)$$

Soit $\widehat{\theta}$ un estimateur du maximum de vraisemblance de $L(y|X;\theta)$; on a alors

$$\frac{\partial \ln L\left(y|X;\hat{\theta}\right)}{\partial\theta} = \sum_{i=1}^{N}\frac{\partial \ln f\left(y_i|X_i;\widehat{\theta}\right)}{\partial\theta} = 0$$

En effectuant le dévéloppement limité de $\partial \ln L/\partial\theta$ au voisinage de $\widehat{\theta}$ on obtient :

$$\frac{\partial \ln L\left(y|x;\hat{\theta}\right)}{\partial\theta} \overset{\mathbf{A}}{=} \frac{\partial \ln L(y|x;\theta)}{\partial\theta} + \frac{\partial^2 \ln L(y|x;\theta)}{\partial\theta^2}\left(\widehat{\theta}-\theta\right)$$

On remarque que ce dévéloppement limité devient exact quand $N \longrightarrow +\infty$. L'égalité précédente devient

$$0 \overset{\mathbf{A}}{=} \frac{\partial \ln L(y|x;\theta)}{\partial\theta} + \frac{\partial^2 \ln L(y|x;\theta)}{\partial\theta^2}\left(\widehat{\theta}-\theta\right)$$

de sorte qu'on écrive

$$\left(\widehat{\theta}-\theta\right) \overset{\mathbf{A}}{=} -\frac{\partial \ln L(y|x;\theta)}{\partial\theta}\cdot\left(\frac{\partial^2 \ln L(y|x;\theta)}{\partial\theta^2}\right)^{-1}$$

$$\Longleftrightarrow \quad \sqrt{N}\left(\widehat{\theta}-\theta\right) \overset{\mathbf{A}}{=} \left[-\frac{1}{N}\frac{\partial^2 \ln L(y|x;\theta)}{\partial\theta^2}\right]^{-1}\cdot\frac{1}{\sqrt{N}}\frac{\partial \ln L(y|x;\theta)}{\partial\theta}$$

La première quantité du membre de droite de l'équation est une moyenne qui converge en probabilité vers l'espérance mathématique correspondante. En appliquant la loi des grands nombres

$$-\frac{1}{N}\frac{\partial^2 \ln L(y|x;\theta)}{\partial\theta^2} = \frac{1}{N}\sum_{i=1}^{N}\left[-\frac{\partial^2 \ln f(y_i|x_i;\theta)}{\partial\theta^2}\right] \xrightarrow{p} \mathbf{E}_\theta\left[-\frac{\partial^2 \ln f(y|x;\theta)}{\partial\theta^2}\right] = \mathbf{J}_1(\theta)$$

Le second terme du membre de droite de l'équation suit, asymptotiquement, une loi normale. On peut écrire :

$$\begin{aligned}\frac{1}{\sqrt{N}}\frac{\partial \ln L\left(y|x;\theta\right)}{\partial\theta} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\frac{\partial \ln f\left(y_i|x_i;\theta\right)}{\partial\theta}\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\sqrt{N}\frac{\partial \ln f\left(y_i|x_i;\theta\right)}{\partial\theta} \\ &= \frac{1}{N}\sum_{i=1}^{N}z_i\end{aligned}$$

où z_i est la variable dont on cherche la distribution. On a alors sous les hypothèses usuelles :

$$\sqrt{N}\left(\overline{z}-E\left(\overline{z}\right)\right)\underset{N\to+\infty}{\overset{L}{\to}}N(0;\mathbf{V}\left(\overline{z}\right)).$$

■

1.3 Le modèle linéaire (simple ou multiple)

Contexte d'adoption

Le but d'un modèle de régression linéaire est de chercher à expliquer la réalisation d'une variable aléatoire $\mathbf{Y}$ par p variables $X=(1;X_1;...;X_p)$. Pour cela, on dispose de n réalisations $(x_1;y_1);...;(x_n;y_n)$ du couple $(X;Y)$. Le but est de modéliser la dépendance de la variable réponse Y sur les variables explicatives $X_1;...;X_p$. Plusieurs raisons peuvent motiver cette volonté de modélisation :

— **la description** : on veut un modèle qui permette de décrire la relation entre X et Y;

— l'**évaluation** des contributions relatives de chaque prédicteur pour expliquer Y;

— la **prédiction** : Prévoir la valeur de Y pour des nouvelles valeurs des variables explicatives.

On rapelle que le modèle linéaire s'écrit :

$$Y=X'\beta+\varepsilon=\beta_0+\beta_1X_1+...+\beta_pX_p+\varepsilon,$$

avec $\beta=(\beta_0;\beta_1;...;\beta_p)'\in\mathbb{R}^{p+1}$ et $\varepsilon\sim\mathcal{N}(0;\sigma^2)$. On distingue alors deux cas :

— les variables X_i sont déterministes (non aléatoires) :

$$Y\sim\mathcal{N}(X\beta';\sigma^2I_n),\qquad E\left[Y\right]=X\beta'$$

— les variables X_i sont aléatoires :

$$(Y|X)\sim\mathcal{N}(X\beta';\sigma^2I_n),\qquad E\left[Y|X\right]=X^{\beta\prime}$$

Limites du modèle linéaire simple ou multiple

Plaçons nous maintenant dans le cas où la variable à expliquer Y est qualitative (sexe, couleur, présence ou absence d'une maladie...) et possède un nombre fini de modalités $g_1, ..., g_m$. Le problème consiste alors à expliquer l'appartenance d'un individu à un groupe à partir des p variables explicatives.

Il est bien entendu impossible de modéliser directement la variable Y par une relation linéaire (imaginons que Y soit le sexe d'une personne ou la couleur de ses cheveux). Dans le but de trouver une solution à cette difficulté, on va s'intéresser aux probabilités $P(Y = g_k|X = x)$. Supposons pour simplifier que la variable Y ne prenne uniquement que deux valeurs : 0 ("groupe 0") ou 1 ("groupe 1"). La connaissance de $P(Y = 1|X = x)$ implique celle de $P(Y = 0|X = x)$: il suffit par conséquent de modéliser la probabilité $\pi(x) = P(Y = 1|X = x)$. On peut par exemple envisager une relation de la forme :

$$\pi_\beta(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon = x'\beta + \varepsilon.$$

Mais cette approche possède plusieurs inconvénients : en effet,

— Remarquons tout d'abord que la variance de $Y|X = x$ vaut $\pi_\beta(x)(1 - \pi_\beta(x))$. Contrairement au modèle linéaire traditionnel, cette variance n'est pas constante et par conséquent l'hypothèse d'homoscédasticité des résidus ne sera pas vérifiée.

— Le fait qu'aucune restriction ne soit effectuée sur les β implique que $x'\beta$ peut prendre n'importe quelle valeur dans $\mathbb{R}$. Ce fait est gênant pour l'estimation d'une probabilité (imaginer une estimation du genre $P_{\widehat{\beta}}(Y = 1|X = x) = -127,332$!!!).

Pour ces raisons, nous devons étendre le modèle linéaire classique aux cas où :

— Y peut être une variable qualitative (présence ou absence d'une maladie, appartenance à une catégorie...) ;

— les erreurs peuvent ne pas avoir la même variance (s'affranchir de l'hypothèse d'homoscédasticité).

1.4 Le modèle de régression logistique

Définition **1.5.** *Un modèle de* ***régression logistique*** *est un modèle d'analyse en statistique qui permet d'étudier la relation entre une variable aléatoire Y à réponse binaire appelée* ***variable à expliquer*** *et une ou plusieurs variables dites* ***explicatives*** *pouvant être quantitatives*

ou qualitatives en utilisant une ***fonction de distribution logistique.*** *Lorsqu'on est en présence d'une seule variable explicative, on dit que le modèle de régression logistique st* ***simple*** *et lorsqu'on a plusieurs variables explicatives, le modèle de régression logistique est* ***multiple****.*

Dans toute la suite, la variable aléatoire Y est dichotomique.
Soit $\mathbf{Y}$ une variable aléatoire à valeurs dans $\{0,1\}$ qu'on cherche à expliquer par p variables explicatives $X = (1, X_1, ..., X_p)'$. Le modèle logistique propose une modélisation de la loi de $Y|X = x$ par une loi de Bernoulli de paramètre $\pi_\beta(x) = P_\beta(Y = 1|X = x)$ telle que :

$$\log \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = x'\beta, \tag{1.5}$$

soit encore

$$\log it(\pi(x)) = x'\beta \tag{1.6}$$

Dans un modèle de régression logistique, nous effectuons deux choix pour définir le modèle :

1. le choix d'une loi pour $Y|X = x$, ici la loi de Bernoulli ;
2. le choix de la modélisation de $P(Y = 1|X = x)$ par

$$\log it(P_\beta(Y = 1|X = x)) = x'\beta$$

Remarque 1.4.1. *On peut remarquer que* $\left\{ \begin{array}{c} \mathbf{E}_\beta[Y|X = x] = P_\beta(Y = 1|X = x) = \pi_\beta(x) \\ \mathbf{V}_\beta(Y|X = x) = \pi_\beta(x)(1 - \pi_\beta(x)) \end{array} \right\}$.
ce qui implique que la variance n'est pas constante et varie selon x.

1.4.1 Estimation des coefficients du modèle de régression logistique

Soit $(x_1; y_1); ...; (x_n; y_n)$ un n-échantillon où $\forall i \in \{1; ...; n\}$, $y_i \in \{0; 1\}$. La vraisemblance du modèle logistique $\mathcal{M}$ de cet échantillon est définie par :

$$\begin{aligned} L_n \quad : \quad & \{0,1\}^n \times \mathbb{R}^{p+1} \longrightarrow \mathbb{R}^+ \\ & (y_1, ..., y_n; \beta) \longmapsto \prod_{i=1}^{n} \pi_\beta(x_i)^{y_i} (1 - \pi_\beta(x_i))^{1-y_i} \end{aligned}$$

Remarque 1.4.2. *Si on désigne par $\mathbf{Y}_i$ une variable aléatoire de loi $\mathcal{B}(\pi_\beta(x_i))$, alors les variables $\mathbf{Y}_1, ..., \mathbf{Y}_n$ sont indépendantes, mais pas de même loi.*

Lorsqu'il n'y a pas de confusion possible, on notera $L_n(y_1, ..., y_n; \beta) = L_n(\beta)$ et on notera également $\ln(L_n(\beta)) = \mathcal{L}_n(\beta)$. Calculons la vraisemblance de l'échantillon ; on a

$$L_n(\beta) = \prod_{i=1}^{n} P_\beta(Y = y_i|X = x_i) = \prod_{i=1}^{n} \pi_\beta(x_i)^{y_i} (1 - \pi_\beta(x_i))^{1-y_i} \tag{1.7}$$

En passant au log, nous obtenons

$$\begin{aligned}\mathcal{L}_n(\beta) &= \sum_{i=1}^{n}\{y_i\log(\pi_\beta(x_i))+(1-y_i)\log(1-\pi_\beta(x_i))\}\\ &= \sum_{i=1}^{n}\left\{y_i\log\left(\frac{\pi_\beta(x_i)}{1-1-\pi_\beta(x_i)}\right)+\log(1-\pi_\beta(x_i))\right\}\\ &= \sum_{i=1}^{n}\{y_i x_i'\beta-\log(1-\exp(x_i'\beta))\}\end{aligned}$$

d'où

$$\mathcal{L}_n(\beta)=\sum_{i=1}^{n}\{y_i x_i'\beta-\log(1-\exp(x_i'\beta))\} \tag{1.8}$$

Le vecteur gradient au point β défini par

$$\nabla\mathcal{L}_n(\beta)=\left[\frac{\partial\mathcal{L}_n}{\partial\beta_0}(\beta),...,\frac{\partial\mathcal{L}_n}{\partial\beta_p}(\beta)\right]' \tag{1.9}$$

s'obtient par dérivation

$$\begin{aligned}\frac{\partial\mathcal{L}_n}{\partial\beta_p}(\beta) &= \sum_{i=1}^{n}\left[y_i x_{ij}-\frac{x_{ij}\exp(x_i'\beta)}{1+x_{ij}\exp(x_i'\beta)}\right]\\ &= \sum_{i=1}^{n}[x_{ij}(y_i-\pi_\beta(x_i'))]\end{aligned}$$

ce qui donne en écriture matricielle

$$\nabla\mathcal{L}_n(\beta)=\sum_{i=1}^{n}[x_{ij}(y_i-\pi_\beta(x_i'))]=\mathbf{X}'(\mathbf{Y}-\mathbf{P}_\beta) \tag{1.10}$$

où $\mathbf{Y}=(y_1,...,y_n)'$ et $\mathbf{P}_\beta=(\pi_\beta(x_i),...,\pi_\beta(x_n))$.

L'estimateur du maximum de vraisemblance (s'il en existe) s'obtient en résolvant l'équation (1.11) ci-dessous (appelée équation de score) :

$$s(\beta)=\nabla\mathcal{L}_n(\beta)=\mathbf{X}'(\mathbf{Y}-\mathbf{P}_\beta)=0 \tag{1.11}$$

On rappelle que si cette équation admet une solution en β notée $g(y_1,...,y_n)$ (et que cette solution est un maximum de $\mathcal{L}_n(\beta)$) alors l'**E.M.V.** sans biais est $\widehat{\beta}=g(Y_1,...Y_n)$.

Toutefois, trouver explicitement $\widehat{\beta}$ n'est pas assez aisé. En effet, l'équation (1.11) ci-dessus se réécrit :

$$\begin{cases} x_{11}y_1+...+x_{n1}y_n=x_{11}\frac{\exp(\beta_1x_{11}+...+\beta_px_{1p})}{1+\exp(\beta_1x_{11}+...+\beta_px_{1p})}+...+x_{n1}\frac{\exp(\beta_1x_{n1}+...+\beta_px_{np})}{1+\exp(\beta_1x_{n1}+...+\beta_px_{np})}\\ \vdots\\ \vdots\\ x_{1p}y_1+...+x_{np}y_n=x_{1p}\frac{\exp(\beta_1x_{11}+...+\beta_px_{1p})}{1+\exp(\beta_1x_{11}+...+\beta_px_{1p})}+...+x_{np}\frac{\exp(\beta_1x_{n1}+...+\beta_px_{np})}{1+\exp(\beta_1x_{n1}+...+\beta_px_{np})}\end{cases}$$

Ce système (qui n'est pas linéaire en β) n'admet pas de solution analytique. On a donc recours à des methodes numériques qui nécessitent de connaître la connaissance des différentes propriétés sur la régularité de la fonction à optimiser.

Il existe plusieurs algorithmes de calcul utilisés par les ordinateurs et qui conduisent à la résolution de cette équation ; notamment l'algorithme de Newton-Raphson et le scoring de Fisher éxécuté par le logiciel **R ;** le logiciel **SAS** ; le logiciel **LOGISTIC**,... (voir les pages 114 à 116 de [8] pour les détails). Toutefois, nous présentons ici l'algorithme de Newton-Raphson qui conduit à déterminer l'**E.M.V.**

L'algorithme de Newton-Raphson qui conduit à la résolution numérique des équations de score.

Pour simplifier les notations, nous supposons que β est univarié. On part tout d'abord d'une valeur initiale arbitraire de β notée β^0 et on désigne par $\beta^1 = \beta^0 + h$ une valeur candidate pour être solution de $s(\beta) = 0$, c'est-à-dire $s(\beta^0 + h) = 0$. Par un dévéloppement limité d'ordre un, de la fonction s, on obtient l'approximation suivante :

$$s\left(\beta^0 + h\right) \approx s\left(\beta^0\right) + hs'\left(\beta^0\right).$$

Comme $s(\beta^0 + h) = 0$, on obtient pour h la valeur suivante :

$$h = -\left[s'\left(\beta^0\right)\right]^{-1} s\left(\beta^0\right)$$

et donc

$$\beta^1 = \beta^0 - \left[s'\left(\beta^0\right)\right]^{-1} s\left(\beta^0\right)$$

Dans le cas qui nous concerne, $\beta \in \mathbb{R}^{p+1}$ et $s(\beta) = \nabla\mathcal{L}_n(\beta)$. La formule de récurrence se traduit par

$$\beta^1 = \beta^0 - \left[\nabla^2\mathcal{L}_n\left(\beta^0\right)\right]^{-1} \nabla\mathcal{L}_n\left(\beta^0\right)$$

où $\nabla^2\mathcal{L}_n(\beta^0)$ désigne la matrice hessienne de la log $-$vraisemblance au point β^0 :

$$\nabla^2\mathcal{L}_n\left(\beta^0\right)_{kl} = \left[\frac{\partial^2\mathcal{L}}{\partial\beta_k\partial\beta_l}\left(\beta^0\right)\right] \quad 0 \preceq k, l \preceq p$$

où nous commettons ici l'abus de désigner par $\nabla^2\mathcal{L}_n(\beta^0)_{kl}$ le terme de la $(k+1)^{ème}$ ligne et $(l+1)^{ème}$ colonne de $\nabla^2\mathcal{L}_n(\beta^0)$.

Le processus est ensuite itéré jusqu'à la convergence et [8] propose qu'on s'arrête lorsque $\left|\beta_l^{(k+1)} - \beta_l^{(k)}\right| < 10^{-6}$. L'algorithme d'estimation de β se résume comme suit :

1. Le choix d'un point de départ β^0;

2. On construit β^{k+1} à partir de β^k

$$\beta^{k+1} = \beta^k + \mathbf{A}^k \nabla \mathcal{L}_n\left(\beta^k\right)$$

où $\nabla \mathcal{L}_n\left(\beta^k\right)$ est le gradient au point β^k et $\mathbf{A}^k = -\left[\nabla^2 \mathcal{L}_n\left(\beta^k\right)\right]^{-1}$ est la matrice de "pas" de l'algorithme. (l'inverse du hessien de $\mathcal{L}_n$ au point β^k).

Algorithme 1 : Maximisation de la vraisemblance

require : β^0

k$\longleftarrow$ 1

repeat

$\beta^{k+1} \longleftarrow \beta^k + \mathbf{A}^k \nabla \mathcal{L}_n\left(\beta^k\right)$

until $\beta^{k+1} \approx \beta^k$ et/ou $\mathcal{L}_n\left(\beta^{k+1}\right) \approx \mathcal{L}_n\left(\beta^k\right)$

Calculons la matrice hessienne $\nabla^2 \mathcal{L}_n(\beta) = \left\{\frac{\partial^2 \mathcal{L}_n}{\partial \beta_k \partial \beta_l}(\beta)\right\}_{0 \preceq k,l \preceq p}$

$\frac{\partial^2 \mathcal{L}_n}{\partial \beta_k \partial \beta_l}(\beta) = -\sum_{i=1}^{n} x_i^k x_i^l \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)} = -\sum_{i=1}^{n} x_i^k x_i^l \pi_\beta(x_i)\left(1-\pi(x_i)\right)$.

En écriture matricielle, nous obtenons

$$\nabla^2 \mathcal{L}_n(\beta) = -\sum_{i=1}^{n} x_i x_i' \pi_\beta(x_i)\left(1-\pi_\beta(x_i)\right) = \mathbf{X}' W_\beta \mathbf{X}$$

où W_β est la matrice diagonale $diag\left(\pi_\beta(x_i)\left(1-\pi_\beta(x_i)\right)\right)$; $i = 1, ..., n$. Nous pouvons maintenant exprimer β^{k+1} en fonction de β^k;

$$\begin{aligned}
\beta^{k+1} &= \beta^k + \left(\mathbf{X}' W_{\beta^k} \mathbf{X}\right)^{-1} \mathbf{X}'\left(\mathbf{Y} - \mathbf{P}_{\beta^k}\right) \\
&= \left(\mathbf{X}' W_{\beta^k} \mathbf{X}\right)^{-1} \mathbf{X}' W_{\beta^k}\left(\mathbf{X}\beta^k + W_{\beta^k}^{-1}\left(\mathbf{Y} - \mathbf{P}_{\beta^k}\right)\right) \\
&= \left(\mathbf{X}' W_{\beta^k} \mathbf{X}\right)^{-1} \mathbf{X}' W_{\beta^k} Z^k
\end{aligned}$$

ou $Z^k = \mathbf{X}\beta^k + W_{\beta^k}^{-1}\left(\mathbf{Y} - \mathbf{P}_{\beta^k}\right)$.

L'algorithme ci dessus est encore appelé **IRLS**ou encore ; **Algorithme des moindres carrés podérés avec itération**.

Cette équation est simplement une régression pondérée du vecteur Z^k où les poids W_{β^k} dépendent de X et de β^k.

Proposition 1.4. l'estimateur du maximum de vraisemblance permet d'estimer la valeur de $\pi(x)$ du modèle de régression logistique.

Preuve. Il suffit de résoudre (en exécutant l'algorithme de Newton-Raphson tel qu'indiqué ci-dessus) l'équation (1.11) et déterminer les coordonnées de $\widehat{\beta}$ solutions de cette équation. Ces

solutions sont ensuite remplacées dans l'expression initiale de $\pi_i(x)$. On obtient alors

$$\widehat{\pi}_i(x) = \frac{\exp(\widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}^j x_i^j)}{1 + \exp(\widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}^j x_i^j)}$$

■

Proposition 1.5. Le maximum de vraisemblance du modèle logistique permet de déterminer la matrice des variances-covariances des coefficients du modèle.

Preuve. De la **proposition 1.1**, la matrice des variances-covariances de β; est égale à

$$\widehat{\mathbf{V}}\left(\widehat{\beta}\right) \simeq \left[\mathbf{I}\left(\widehat{\beta}\right)\right]^{-1}$$

où

$$[\mathbf{I}(\beta)] = -\mathbf{E}_\beta\left(\frac{\partial^2}{\partial\beta^2} l(\beta; y, x)\right)$$

est la matrice d'information de Fisher.

Pour besoin de simplification, considérons le cas où nous sommes en présence d'une seule variable explicative. Toutefois, les calculs peuvent aisement s'étendre pour les cas des modèles logistiques à plusieurs variables explicatives.

Nous allons utiliser les propriétés asymptotiques de l'estimateur du maximum de vraisemblance. Soit $\mathbf{I}(\beta)$ la 2×2-matrice d'information de Fisher de β, on a alors $\widehat{\mathbf{V}}\left(\widehat{\beta}\right) \simeq \left[\mathbf{I}\left(\widehat{\beta}\right)\right]^{-1}$.

Posons $f(y,\beta)$ la fonction de probabilité conjointe du vecteur aléatoire $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ au point $y = (y_1, y_2, ..., y_n)$.

$$\mathbf{I}(\beta) = \begin{pmatrix} -E\left(\frac{\partial^2}{\partial^2\beta_0}\ln f(y,\beta)\right) & -E\left(\frac{\partial^2}{\partial\beta_0\partial\beta_1}\ln f(y,\beta)\right) \\ -E\left(\frac{\partial^2}{\partial\beta_0\partial\beta_1}\ln f(y,\beta)\right) & -E\left(\frac{\partial^2}{\partial^2\beta_1}\ln f(y,\beta)\right) \end{pmatrix}$$

Or $\ln f(y,\beta)$ n'est autre que la log $-$vraisemblance vue un peu plus haut. Il nous faut donc calculer ses dérivées partielles secondes, soit :

$$\frac{\partial^2}{\partial^2\beta_0}\ln f(y,\beta) = \frac{\partial}{\partial\beta_0}\sum_{i=1}^{n}[y_i - \pi(x_i)] = -\sum_{i=1}^{n}\frac{\partial}{\partial\beta_0}\pi(x_i) = -\sum_{i=1}^{n}\pi(x_i)[1-\pi(x_i)]$$

De même :

$$\frac{\partial^2}{\partial^2\beta_1}\ln f(y,\beta) = \frac{\partial}{\partial\beta_1}\sum_{i=1}^{n}x_i[y_i - \pi(x_i)] = -\sum_{i=1}^{n}x_i^2\pi(x_i)[1-\pi(x_i)]$$

et :

$$\frac{\partial^2}{\partial\beta_0\partial\beta_1}\ln f(y,\beta) = \frac{\partial}{\partial\beta_1}\sum_{i=1}^{n}[y_i - \pi(x_i)] = -\sum_{i=1}^{n}x_i\pi(x_i)[1-\pi(x_i)].$$

Ces dérivées ne dépendent plus des y_i et elles restent inchangées quand on passe aux espérances mathématiques ; d'où :

$$\mathbf{I}(\beta) = \begin{pmatrix} \sum_{i=1}^{n} \pi(x_i)[1-\pi(x_i)] & \sum_{i=1}^{n} x_i \pi(x_i)[1-\pi(x_i)] \\ \sum_{i=1}^{n} x_i \pi(x_i)[1-\pi(x_i)] & \sum_{i=1}^{n} x_i^2 \pi(x_i)[1-\pi(x_i)] \end{pmatrix}.$$

Comme β est inconnu, il nous faut estimer $\mathbf{I}(\beta)$ par $\mathbf{I}\left(\widehat{\beta}\right)$, c'est-à-dire substituer dans l'écriture de $\mathbf{I}(\beta)$ ci-dessus $\widehat{\pi}(x_i)$ à $\pi(x_i)$. En inversant $\mathbf{I}\left(\widehat{\beta}\right)$, on obtient une estimation de $\mathbf{V}\left(\widehat{\beta}\right)$ que nous notons :

$$\mathbf{V}\left(\widehat{\beta}\right) = \begin{pmatrix} s^2\left(\widehat{\beta}_0\right) & s^2\left(\widehat{\beta}_0, \widehat{\beta}_1\right) \\ s^2\left(\widehat{\beta}_0, \widehat{\beta}_1\right) & s^2\left(\widehat{\beta}_1\right) \end{pmatrix}$$

Où $s^2\left(\widehat{\beta}_0\right)$ est une estimation de la variance de $\widehat{\beta}_0$, $s^2\left(\widehat{\beta}_1\right)$ est une estimation de la variance de $\widehat{\beta}_1$ et $s^2\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$ est une estimation de la covariance entre $\widehat{\beta}_0$ et $\widehat{\beta}_1$.

Pour le cas d'une régression logistique multiple, on procède ainsi qu'il suit :

L'élément à la position $(j;j)$ de cette matrice de Fisher est obtenu par :,

$$\begin{aligned} \frac{\partial^2 l(\beta;y,x)}{\partial \beta_j^2} &= \frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} x_{ij}(y_i - \pi_i) = \sum_{i=1}^{n} x_{ij} \frac{\partial}{\partial \beta_j}(y_i - \pi_i) = -\sum_{i=1}^{n} x_{ij} \frac{\partial}{\partial \beta_j} \pi_i \\ &= -\sum_{i=1}^{n} x_{ij} \frac{\partial}{\partial \beta_j} \cdot \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \\ &= -\sum_{i=1}^{n} x_{ij} \left\{ \frac{x_{ij}\exp(x_i'\beta)}{(1+\exp(x_i'\beta))^2} \right\} \\ &= -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i) \end{aligned}$$

$$\frac{\partial^2 l(\beta;y,x)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i) \tag{1.12}$$

et, par un raisonnement similaire à celui ci-dessus, on détermine l'élément en position $(j;l)$ de cette matrice qui est obtenu par

$$\frac{\partial^2 l(\beta;y,x)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1-\pi_i)$$

On a alors la matrice d'information de Fisher :

$$\mathbf{I}(\beta) = \begin{pmatrix} -\mathbf{E}\left(\frac{\partial^2}{\partial^2 \beta_0} l(\beta;y;x)\right) & -\mathbf{E}\left(\frac{\partial^2}{\partial \beta_0 \partial \beta_1} l(\beta;y;x)\right) & \cdots & -\mathbf{E}\left(\frac{\partial^2}{\partial \beta_0 \partial \beta_{k-1}} l(\beta;y;x)\right) \\ -\mathbf{E}\left(\frac{\partial^2}{\partial \beta_0 \partial \beta_1} l(\beta;y;x)\right) & -\mathbf{E}\left(\frac{\partial^2}{\partial^2 \beta_1} l(\beta;y;x)\right) & \cdots & -\mathbf{E}\left(\frac{\partial^2}{\partial \beta_1 \partial \beta_{k-1}} l(\beta;y;x)\right) \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{E}\left(\frac{\partial^2}{\partial \beta_0 \partial \beta_{k-1}} l(\beta;y;x)\right) & -\mathbf{E}\left(\frac{\partial^2}{\partial \beta_1 \partial \beta_{k-1}} l(\beta;y;x)\right) & \cdots & -\mathbf{E}\left(\frac{\partial^2}{\partial^2 \beta_{k-1}} l(\beta;y;x)\right) \end{pmatrix}$$

On remplace ensuite les valeurs estimées des β_i dans l'expression de la matrice d'information de Fisher pour obtenir $\mathbf{I}\left(\widehat{\beta}\right)$. L'inverse de cette matrice égale à $\widehat{\mathbf{Var}}\left(\widehat{\beta}\right)$, est un estimateur convergent de $\mathbf{Var}\left(\widehat{\beta}\right)$. ■

En général, ces calculs sont effectués à l'aide des programmes de calcul numérique intégrés dans certains logiciels d'analyse statistique tels **R**, **SAS, STATISTICA**, **LOGISTIC**....

1.4.2 Interprétation des coefficients du modèle de régression logistique

Les odds ratio servent à mesurer l'effet d'une variable quantitative ou le contraste entre les effets d'une variable qualitative. L'idée générale est de raisonner en termes de probabilités ou de rapport de cotes (odds). Si on a par exemple, une probabilité $p = 1/4$ de gagner à un jeu, cela signifie que sur 4 personnes participant à ce jeu, une gagne et les trois perdent, soit un rapport de 1 gagnant sur 3 perdants, c'est-à-dire $\frac{p}{1-p} = \frac{1}{3}$. Ce rapport varie entre 0 (0 gagnant) et l'infini (que des gagnants) en passant par 1 (autant de gagnants que de perdants).

Définition **1.6.** – L'odds (chance) pour un individu x d'obtenir la réponse Y=1 est défini par :

$$odds(x) = \frac{\pi(x)}{1-\pi(x)}, \quad \text{où} \quad \pi(x) = P(Y = 1|X = x).$$

– L'odds ratio (rapport des chances) entre deux individus x et $\overline{x}$ est :

$$OR(x, \overline{x}) = \frac{odds(x)}{odds(\overline{x})} = \frac{\frac{\pi(x)}{1-\pi(x)}}{\frac{\pi(\overline{x})}{1-\pi(\overline{x})}}$$

Les odds ratio peuvent être utilisés de plusieurs manières :

1. **Pour comparer les probabilités de succès entre deux individus :**
   - $OR(x, \overline{x}) \succ 1 \Longleftrightarrow \pi(x) \succ \pi(\overline{x})$
   - $OR(x, \overline{x}) = 1 \Longleftrightarrow \pi(x) = \pi(\overline{x})$
   - $OR(x, \overline{x}) \prec 1 \Longleftrightarrow \pi(x) \prec \pi(\overline{x})$
2. **Pour l'interprétation du risque relatif :** dans le cas où $\pi(x)$ et $\pi(\overline{x})$ sont très petits par rapport à 1, comme dans le cas d'une maladie très rare, on peut faire l'approximation $OR(x, \overline{x}) \sim \frac{\pi(x)}{\pi(\overline{x})}$ et interpréter simplement. Par exemple, si $OR(x, \overline{x}) = 4$, alors la réponse (maladie) est 4 fois plus probable dans le cas où $X = x$ que dans le cas où $X = \overline{x}$.
3. **Pour la mesure de l'impact d'une variable :** Sachant que

$$\log it(\pi_\beta(x)) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p,$$

on vérifie aisement que

$$OR\left(x,\overline{x}\right)=\exp(\beta_1\left(x_1-\overline{x}_1\right))...\exp\left(\beta_p\left(x_p-\overline{x}_p\right)\right)$$

Pour mesurer l'influence d'une variable sur l'odds ratio, il suffit de considérer deux observations x et $\overline{x}$ qui diffèrent uniquement de la $j^{ème}$ variable. On obtient alors

$$OR\left(x,\overline{x}\right)=\exp\left(\beta_j\left(x_j-\overline{x}_j\right)\right)$$

Ainsi, une variation d'une unité (sur l'échelle de cette variable) correspond à un odds ratio $\exp(\beta_j)$ qui est uniquement fonction de β_j. Le coefficient β_j permet ainsi de mesurer l'influence de la $j^{ème}$ variable sur le rapport $\frac{\pi(x)}{1-\pi(x)}$ lorsque x_j varie d'une unité, et ceux indépendamment de la valeur de x_j. Une telle analyse peut se révéler intéressante pour étudier l'influence d'un changement d'état d'une variable qualitative.

On peut donc donner les interprétations suivantes des coefficients du modèle logistique :

– Une première interprétation de β_0 est que β_0 correspond au logarithme de la cote pour un individu ayant les coordonnées de X toutes nulles. Car en effet, si $\forall i=1;...;k\,;\ x_i=0$, alors

$$\log\left(\frac{\pi\left(0\right)}{1-\pi\left(0\right)}\right)=\beta_0\Longleftrightarrow\frac{\pi\left(0\right)}{1-\pi\left(0\right)}=e^{\beta_0}.$$

Cependant, cette interprétation connaît une limite d'interprétation lorsqu'on est dans le cas où la variable X ne peut s'annuler (comme le cas de l'âge).

1. La seconde interprétation est que β_0 correspond au logarithme de la cote pour un individu dont toutes les variables X_i sont ignorées.
2. Généralement, si nous nous concentrons sur n'importe quel coefficient β_L pour $i=L$, nous pouvons fournir l'interprétation suivante :

β_L correspond au changement dans le $logit\pi\left(X\right)$ pour chaque unité de changement dans X_L sachant que toutes les autres valeurs de X_i pour $i\neq L$ sont fixées.

1.4.3 Tests et intervalle de confiance

Définition 1.7. *Un modèle $\mathcal{M}_1$ est emboîté dans un autre modèle $\mathcal{M}$ plus général (ou plus grand), lorsque le modèle $\mathcal{M}_1$ est un cas particulier qui impose des restrictions aux paramètres de ce modèle $\mathcal{M}$ plus général.*

Exemple 1.4.1. $\mathcal{M}_1:\log it\left(\pi_\beta\left(x\right)\right)=\beta_0+\beta_1x_1+\beta_2x_2$ et
$\mathcal{M}_2:\log it\left(\pi_\gamma\left(x\right)\right)=\gamma_0+\gamma_1x_1+\gamma_2x_2+\gamma_3x_3$ sont deux modèles emboîtés l'un de l'autre.

Définition **1.8.** *Le modèle saturé $\mathcal{M}_s$ est celui qui illustre de façon parfaite l'échantillon observé. Les probabilités du modèle saturé sont les fréquences observées : donc* $p_i = \frac{\sum_{i=1}^{k} y_i}{n}$.

Définition **1.9.** *Le modèle complet est le modèle qui prend en compte l'ensemble des variables explicatives condérées pour effectuer l'analyse qui a permis d'obtenir l'échantillon considéré.*

Soient $\mathcal{L}_1$ et $\mathcal{L}_2$; les fonctions log-vraisemblances des modèles $\mathcal{M}_1$ et $\mathcal{M}_2$. La déviance est

$$D = -2\left(\widehat{\mathcal{L}}_1 - \widehat{\mathcal{L}}_2\right)$$

où $\widehat{\mathcal{L}}_1$ et $\widehat{\mathcal{L}}_2$ sont les log $-$vraisemblance maximisées de $\mathcal{M}_1$ et $\mathcal{M}_2$.

Tests de nullité de q coefficients du modèle

La théorie du maximum de vraisemblance nous donnant la loi (asymptotique) des estimateurs, il est possible de tester la significabilité des variables explicatives. Pour cela, trois tests sont généralement utilisés :

1. le test de **Wald ;**
2. le test du **rapport de vraisemblance ;**
3. le **score test.**

Les hypothèses s'écrivent :

$$\mathbf{H}_0 : \beta_{j1} = \beta_{j2} = ... = \beta_{jq} = 0 \ \text{ contre } \quad \mathbf{H}_1 : \exists k \in \{1; ...; q\} : \beta_{jk} \neq 0.$$

Pour alléger les notations, nous supposons sans perte de généralité que nous testons la nullité des q coeffficients du modèle

$$\mathbf{H}_0 : \beta_1 = \beta_2 = ... = \beta_q = 0 \ \text{ contre } \quad \mathbf{H}_1 : \exists k \in \{1; ...; q\} : \beta_k \neq 0.$$

Test de Wald Il est basé sur la **proposition 1.1**. On note par $\beta_{0,...,q-1}$ le vecteur composé des q premières composantes de β et $\widehat{\sum}^{-1}_{0,...,q-1}$ la matrice bloc composée des q premières lignes et colonnes de $\widehat{\sum}^{-1}$.

$$\beta'_{0,...,q-1}\widehat{\sum}_{0,...,q-1}\beta_{0,...,q-1} \xrightarrow{\mathcal{L}} \mathcal{X}^2_q$$

Test du rapport de vraisemblance ou test de la déviance

La statistique de test est basée sur la différence des rapports de vraisemblance entre le modèle complet et le modèle sous $\mathbf{H}_0$. On note $\widehat{\beta}_{\mathbf{H}_0}$ l'estimateur du maximum de vraisemblance contraint par $\mathbf{H}_0$ (il s'obtient en supprimant les q premières variables du modèle). On a alors sous $\mathbf{H}_0$

$$2\left(\mathcal{L}_n\left(\widehat{\beta}\right) - \mathcal{L}_n\left(\widehat{\beta}_{\mathbf{H}_0}\right)\right) \longrightarrow \mathcal{X}^2_q$$

(q ddl. car il y a q coefficients $(\beta_0, \beta_1, ..., \beta_{q-1})$ testés en même temps).

Notons tout de même que ce test ne s'applique que pour deux modèles "emboîtés" c'est-à-dire que l'un est un cas particulier de l'autre.

Test du score On cherche ici à vérifier si la fonction de score (gradient de la log-vraisemblance) est proche de 0 sous H_0. Sous H_0 on a

$$S\left(\widehat{\beta}_{\mathbf{H}_0}\right)' \widehat{\sum}_{\mathbf{H}_0}^{-1} S\left(\widehat{\beta}_{\mathbf{H}_0}\right) \xrightarrow{\mathcal{L}} \mathcal{X}_q^2$$

où $\widehat{\sum}_{\mathbf{H}_0}^{-1} = \mathbf{X}' W_{\widehat{\beta}_{\mathbf{H}_0}} \mathbf{X}$.

Pour ces 3 tests, on rejette l'hypothèse nulle si la valeur observée de la statistique de test dépasse le quantile d'ordre $1-\alpha$ de la loi $\mathcal{X}_q^2$. La **figure 1.1** (voir [7]) ci-dessous permet de visualiser les trois tests. Le test du score revient à tester la nullité de la pente en $\widehat{\beta}_{H_0}$ ($\widehat{\beta}$ H_0), le test de Wald la nullité de la distance entre $\widehat{\beta}$ et $\widehat{\beta}_{H_0}$ et le test du rapport de vraisemblance la nullité de la différence entre les vraisemblances en ces deux points.

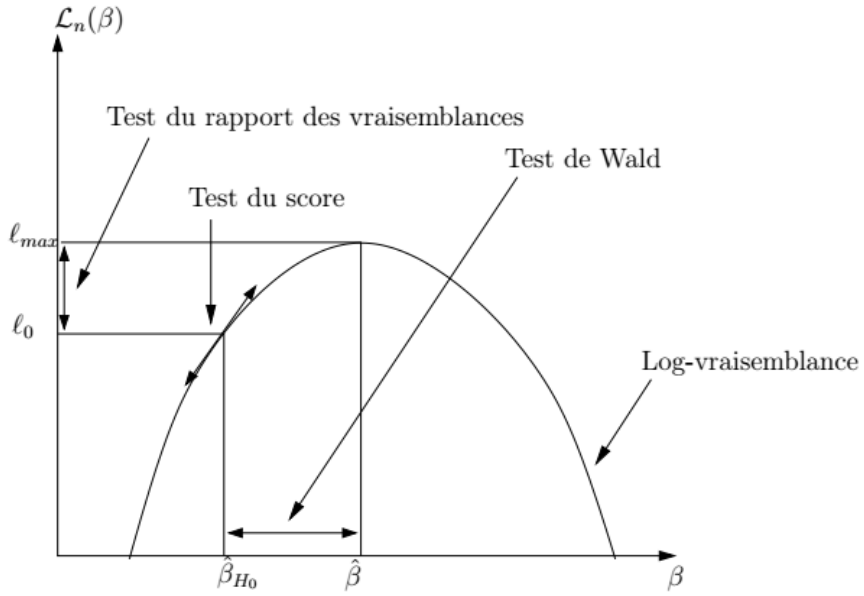

FIG 1.1 - Test du Rapport de vraisemblance, Score test et test de Wald

Le test du khi-deux de Pearson

La statistique de Pearson appelée encore Khi-deux de Pearson est une autre mesure définie par

$$X^2 = \sum_{i=1}^{k} \frac{(y_i - n_i\widehat{\pi})^2}{\widehat{\pi}_i (n_i - n_i\widehat{\pi}_i)} = \sum_{i=1}^{k} \frac{n_i (p_i - \widehat{\pi})^2}{\widehat{\pi}_i (1 - \widehat{\pi}_i)} \tag{1.13}$$

Proposition 1.6. La statistique de Pearson $X^2 \sim \chi_\upsilon^2$, où υ est le nombre de restrictions imposées par le modèle logistique sur les k paramètres du modèle saturé.

Preuve. $X^2 = \sum_{i=1}^{k} \frac{(y_i - n_i\widehat{\pi})^2}{n_i\widehat{\pi}_i} + \sum_{i=1}^{k} \frac{(y_i - n_i\widehat{\pi})^2}{n_i(1-\widehat{\pi}_i)} = \sum_{i=1}^{k} \frac{(y_i - n_i\widehat{\pi})^2}{n_i\widehat{\pi}_i} + \sum_{i=1}^{k} \frac{[(n_i - y_i)(n_i - n_i\widehat{\pi})]^2}{n_i(1-\widehat{\pi}_i)}$ qui représente la statistique χ^2. Lorsque les n_i sont assez grands, $X^2 \sim \chi^2_v$. ■

Test de la déviance résiduelle

Définition **1.10.** On appelle **déviance résiduelle** le nombre noté D_r tel que

$$D_r = -2\log\widehat{L}_m - \left(-2\log\widehat{L}_s\right)$$

où $\widehat{L}_m$ et $\widehat{L}_s$ sont les fonctions de vraisemblance maximisées sous $\mathcal{M}$ et $\mathcal{M}_s$. D_r exprime l'écart entre la log-vraisemblance du modèle complet de la régression logistique avec la log-vraisemblance du modèle saturé. Elle est égale à 2 fois la différence entre les log-vraisemblances évaluées de l'**E.M.V.** et les Y_i observées

$$\mathbf{D}\left(p, \widehat{\pi}\right) = -2\ln\left(\frac{\text{vraisemblance du modèle logistique}}{\text{vraisemblance du modèle saturé}}\right)$$

soit

$$\mathbf{D}\left(p, \widehat{\pi}\right) = 2\sum_{i=1}^{k}\left[y_i \ln\left(\frac{p_i}{\widehat{\pi}_i}\right) + (n_i - y_i)\ln\left(\frac{1-p_i}{1-\widehat{\pi}_i}\right)\right] \tag{1.14}$$

Proposition **1.7.** Lorsque les n_i sont assez grands $\mathbf{D}_r \sim \chi^2_{k-(p+1)}$ où p est le nombre de variables à expliquer du modèle logistique complet.

Intervalle de confiance

1. **Estimation par approximation normale, dite methode de Wald** D'après la **proposition 1.5**, un estimateur de la variance de $\widehat{\beta}_j$ est donné par le $j^{ème}$ terme de la diagonale de la matrice des variances-covariances que nous avons noté $\sigma^2 = s^2\left(\widehat{\beta}_j\right)$. Si on pose $Z_j = \frac{\left(\widehat{\beta}_j - \beta_j\right)^2}{s^2\left(\widehat{\beta}_j\right)}$. Il vient que

$$Z_j = \frac{\left(\widehat{\beta}_j - \beta_j\right)^2}{s^2\left(\widehat{\beta}_j\right)} \xrightarrow{\mathcal{L}} \mathcal{X}^2_1 \text{ ou encore } \frac{\widehat{\beta}_j - \beta_j}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Un intervalle de confiance (asymptotique) de niveau $1-\alpha$ pour β_j est donné par $IC_{1-\alpha}\left(\beta_j\right) = \left[\beta_j - u_{1-\alpha/2}\widehat{\sigma}_j; \beta_j + u_{1-\alpha/2}\widehat{\sigma}_j\right]$, où $u_{1-\alpha/2}$ représente le quantile du niveau $(1-\alpha/2)$ de la loi normale $\mathcal{N}(0,1)$.

$\frac{\widehat{\beta}_j - \beta_j}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ nous permet de déduire de la nullité d'un coefficient du modèle . En effet, si on note $\mathrm{H}_0 : \beta_j = 0$ et $\mathrm{H}_1 : \beta_j \neq 0$, alors sous l'hypothèse H_0, $\widehat{\beta}_j/\widehat{\sigma}_j \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. On rejettera H_0 si la valeur observée de $\widehat{\beta}_j/\widehat{\sigma}_j$ dépasse en valeur absolue le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

2. **Estimation par la methode du rapport de vraisemblance.**

Elle est obtenue de la manière suivante. Les limites de l'intervalle de confiance du RC, pour un niveau $100(1-\alpha)$ sont obtenues en calculant d'abord β_j, puis par exponentiation celles de **RC** correspondantes sont obtenues.

Les limites de confiance du coefficient β_j sont les deux valeurs qui satisfont l'équation suivante :

$$2\left[L_{\max} - L\left(\beta_j\right)\right] - \chi^2_{(1-\alpha,1)} = 0$$

où L(β) et $L_{\max}$ désignent le log de la fonction de vraisemblance évalué respectivement en β_j et à l'estimateur du maximum de vraisemblance de $\widehat{\beta}_j$.

Les logiciels d'analyse statistique comme [20] l'utilisent pour déterminer l'intervalle de validité de chaque coefficient du modèle logistique.

Chapitre 2

La sur-dispersion dans les modèles de régression logistique

Considérons un $n-$échantillon $(X_1; Y_1); ...; (X_n; Y_n)$ tel que $\forall i \in \{1; ...; n\}, Y_i \in \{0; 1\}$ et X_i est le vecteur des variables explicatives de l'individu i de l'échantillon. Le modèle de régression logistique stipule que

$$P(Y_i = 1|X_i = x_i) = \pi(x_i) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)} \text{ et } P(Y_i = 0|X_i = x_i) = 1 - \pi(x_i)$$

Ainsi,

$$\forall i \in \{1; ...; n\}, Y_i \sim B(\pi(x_i))$$

On a donc $\forall i \in \{1; ...; n\}$;

$$\mathbf{E}(Y_i) = \pi(x_i) \text{ et que } Var[Y_i] = \pi(x_i)(1 - \pi(x_i))$$

Lorsque nous travaillons sur des données individuelles, cette condition est naturellement satisfaite. Cependant, quand nous travaillons sur des données groupées ou bien corrélées, cette caractéristique peut ne plus être respectée et ceci pour plusieurs raisons.

2.1 Données corrélées

Nous supposerons dans toute la suite que les n individus de l'échantillon sont reparties en k groupes se présentant ainsi qu'il suit :

- $n_i =$ à l'effectif du groupe i; $i = 1; ...; k$,
- $Y_i =$ au nombre de succès observés dans le groupe i;
- $\pi_i =$ à la probabilité de succès dans le groupe i préconisée par le modèle logistique; on a alors $\pi_i = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$;

- $\widehat{Y}_i = n_i\widehat{\pi}_i$ représente le nombre de succès attendu dans le groupe i suivant le modèle, et $\widehat{\pi}_i$ est la probabilité de succès estimée dans le groupe i par le modèle.

Pour tout $i \in \{1; ...; k\}$, la variable à expliquer $Y_i \sim \mathcal{B}(n_i; \pi_i)$. On doit donc avoir alors $\mathbf{E}(Y_i) = n_i\pi_i$ et de variance $\mathbf{Var}(Y_i) = n_i\pi_i(1-\pi_i)$.
Cependant, pour des raisons diverses, on se retrouve à avoir souvent $\mathbf{Var}(Y_i) \prec n_i\pi_i(1-\pi_i)$ ou $\mathbf{Var}(Y_i) \succ n_i\pi_i(1-\pi_i)$.

Définition 2.1. La **sur-dispersion** est un phénomène rencontré dans les analyses statistiques et qui décrit une situation où la variance calculée est supérieure à la variance attendue imposée par le modèle théorique que la variable aléatoire suit. Puisque la variable aléatoire $\mathbf{Y}$ suit une loi binomiale, alors il y a **sur-dispersion** lorsque $Var[\mathbf{Y}] \succ n\widehat{\pi}(1-\widehat{\pi})$, où $\widehat{\pi}$ est l'estimée de la probabilité de réalisation de la variable aléatoire $\mathbf{Y}$

2.2 Causes de la sur-dispersion

Plusieurs raisons peuvent conduire à la présence de la sur-dispersion dans une analyse utilisant un modèle de régression logistique. On peut citer entre autres :

- la non prise en compte d'une variable explicative très importante ;
- l'échantillon n'est pas assez représentatif ;
- une forte corrélation entre les différentes réponses binaires de la variable à expliquer ;
- la non constance des probabilités dans les différentes classes...etc.

A cause des conditions imposées dans l'adoption d'un modèle de régression logistique pour effectuer l'analyse [3], nous allons considérer les cas où la sur-dispersion est causée par la **non constance des probabilités entre les classes** et celle qui est causée par une **forte corrélation entre les réponses binaires** car les autres peuvent être éviter.

2.2.1 La non constance des probabilités dans les classes

Soit $i = 1; ...; k$, alors $Y_i \sim \mathcal{B}(n_i; \pi_i)$ de moyenne $\mathrm{E}(Y_i|i) = n_i\pi_i$ et de variance $\mathrm{Var}(Y_i|i) = n_i\pi_i(1-\pi_i)$. La non constance des probabilités dans les différentes classes fait que π_i se comporte aussi comme une variable aléatoire qui suit aussi une loi binomiale avec une moyenne $E(\pi_i) = \pi_{i0}$ et une variance $Var(\pi_i) = \tau^2\pi_{i0}(1-\pi_{i0})$

La moyenne non conditionnelle à la classe i est $\mathrm{E}(Y_i) = E(E(Y_i|i)) = E(n_i\pi_i) = n_i\pi_{i0}$

La variance non conditionnelle à la classe i est

$$\begin{aligned}
\mathbf{Var}\,(Y_i) &= E\,(Var\,(Y_i|i)) + Var\,(E\,(Y_i|i)) \\
&= E\,(n_i\pi_i\,(1-\pi_i)) + \mathbf{Var}\,(n_i\pi_i) \\
&= n_i\left[E\,(\pi_i) - E\left(\pi_i^2\right)\right] + n_i^2\tau_i^2\,(1-\pi_{i0}) \\
&= n_i\left(\pi_{i0} - Var\,(\pi_i) - (E\,(\pi_i))^2\right) + n_i^2\tau_i^2\pi_{i0}\,(1-\pi_{i0}) \\
&= n_i\left(\pi_{i0} - \tau_i^2\pi_{i0}\,(1-\pi_{i0}) - \pi_{i0}^2\right) + n_i^2\tau_i^2\pi_{i0}\,(1-\pi_{i0}) \\
&= n_i\pi_{i0} - n_i\tau_i^2\pi_{i0}\,(1-\pi_{i0}) - n_i\pi_{i0}^2 + n_i^2\tau_i^2\pi_{i0}\,(1-\pi_{i0}) \\
&= n_i\pi_{i0}\,(1-\pi_{i0}) + n_i\tau_i^2\pi_{i0}\,(1-\pi_{i0})\,(n_i-1) \\
&= n_i\pi_{i0}\,(1-\pi_{i0})\left[1 + (n_i-1)\,\tau_i^2\right]
\end{aligned}$$

si on pose $\phi = 1 + (n_i - 1)\,\tau_i^2$, on a bien $\phi \succ 1$ dès que $n_i \succ 1$ et

$$\mathbf{Var}\,(Y_i) = n_i\pi_{i0}\,(1-\pi_{i0})\,\phi \tag{2.1}$$

2.2.2 La forte corrélation entre les réponses binaires

Pour une classe i, posons R_{ij} la réponse de l'observation j. On a alors

$$R_{ij} = \begin{cases} 1 \text{ si succès} \\ 0 \text{ si échec} \end{cases}$$

Alors R_{ij} suit une loi binomiale d'espérance $\mathbf{E}(R_{ij}) = \pi_i$ et de variance

$$\mathbf{Var}\,(R_{ij}) = \pi_i\,(1-\pi_i)\,.$$

Le nombre total de succès dans la classe i suit une loi binomiale d'espérance $\mathbf{E}(Y_i) = n_i\pi_i$. Supposons que la constante de corrélation est la même entre deux observations différentes dans la classe i, et notons par ρ cette constante de corrélation.

On a alors

$$\begin{aligned}
cor\,(R_{ij}; R_{ik}) &= \rho\ ,\ \forall\ j \neq k \\
&\Longrightarrow\ cov\,(R_{ij}; R_{ik}) = \rho\sqrt{\mathbf{Var}\,(R_{ij})\,\mathbf{Var}\,(R_{ik})} = \rho\pi_i\,(1-\pi_i)
\end{aligned}$$

$$\begin{aligned}
\mathbf{Var}\,(Y_i) &= \sum_{j=1}^{n_i}\mathbf{Var}(R_{ij}) + \sum_{j=1}^{n_i}\sum_{j\neq k} cov\,(R_{ij}; R_{ik}) \\
&= n_i\pi_i\,(1-\pi_i) + n_i\,(n_i-1)\,[\rho\pi_i\,(1-\pi_i)] \\
&= n_i\pi_i\,(1-\pi_i)\,[1 + (n_i-1)\,\rho]
\end{aligned}$$

$$\mathbf{Var}\,(Y_i) = n_i\pi_i\,(1-\pi_i)\,[1 + (n_i-1)\,\rho] \tag{2.2}$$

En posant $\phi = [1 + (n_i - 1)\,\rho]$, on obtient alors $\mathbf{Var}\,(Y_i) = n_i\pi_i\,(1-\pi_i)\,\phi$ avec $\phi \succ 1$.

2.3 Conséquences de la présence de la sur-dispersion

2.3.1 sur les coefficients du modèle logistique

La vraisemblance de cet échantillon est

$$L\left(\mathbf{y};\pi_{\beta}(x)\right)=\prod_{i=1}^{k}f\left(y_i;\pi_{\beta}(x_i)\right)=\prod_{i=1}^{k}C_{n_i}^{y_i}\left(\pi_{\beta}(x_i)\right)^{y_i}\left(1-\pi_{\beta}(x_i)\right)^{y_i-n_i}$$

L'estimateur du maximum de vraisemblance $\widehat{\beta}$ est la valeur de β qui maximise L (ou son logarithme). Le logarithme de L est

$$\mathcal{L}\left(\mathbf{y};\pi_{\beta}(x)\right)=\sum_{i=1}^{k}\ln C_{n_i}^{y_i}+\sum_{i=1}^{k}\left[y_i x_i'\beta+n_i\ln\left(1+x_i'\beta\right)\right]$$

La valeur de β qui maximise la vraisemblance doit satisfaire l'équation $\frac{\partial\mathcal{L}\left(\mathbf{y};\pi_{\beta}(x)\right)}{\partial\beta}=0$. Sa résolution ; identique à l'équation (1.11) ; se fait numériquement à l'aide des programmes informatiques comme présenté plus haut.

Remarque 2.3.1. *On peut remarquer que la présence de la sur-dispersion ; lorsqu'on ne l'a pas prise en compte ; n'a aucun impact sur l'estimation des coefficients*

2.3.2 Sur la variance et covariance des coefficients du modèle

Notons par $\mathbf{E}(Y_i)=\pi_0\,;\ \ \forall i\in\{1;...;k\}$ alors d'après la **proposition 1.1,** la matrice des variances-covariances de β; est égale à

$$\widehat{\mathbf{V}}'\left(\widehat{\beta}\right)\simeq\left[\mathbf{I}'\left(\widehat{\beta}\right)\right]^{-1}$$

où

$$\left[\mathbf{I}'(\beta)\right]=-\mathbf{E}_{\beta}\left(\frac{\partial^2}{\partial\beta^2}l\left(\beta;y,x\right)\right)$$

est la matrice d'information de Fisher.

Pour besoin de simplification, considérons le cas où nous sommes en présence d'une seule variable explicative. Toutefois, les calculs peuvent aisement s'étendre pour les cas des modèles logistiques à plusieurs variables explicatives.

Nous allons utiliser les propriétés asymptotiques de l'estimateur du maximum de vraisemblance. Soit $\mathbf{I}(\beta)$ la 2×2-matrice d'information de Fisher de β, on a alors $\widehat{\mathbf{V}}'\left(\widehat{\beta}\right)\simeq\left[\mathbf{I}'\left(\widehat{\beta}\right)\right]^{-1}$.

Posons $f(y,\beta)$ la fonction de probabilité conjointe du vecteur aléatoire $\mathbf{Y}=(Y_1,Y_2,...,Y_k)$ au point $y=(y_1,y_2,...,y_k)$.

$$\mathbf{I}'(\beta)=\begin{pmatrix}-E\left(\frac{\partial^2}{\partial^2\beta_0}\ln f(y,\beta)\right) & -E\left(\frac{\partial^2}{\partial\beta_0\partial\beta_1}\ln f(y,\beta)\right)\\ -E\left(\frac{\partial^2}{\partial\beta_0\partial\beta_1}\ln f(y,\beta)\right) & -E\left(\frac{\partial^2}{\partial^2\beta_1}\ln f(y,\beta)\right)\end{pmatrix}$$

Or $\ln f(y,\beta)$ n'est autre que la log $-$vraisemblance vue un peu plus haut. Il nous faut donc calculer ses dérivées partielles secondes, soit :

$$\frac{\partial^2}{\partial^2\beta_0}\ln f(y,\beta)=\frac{\partial}{\partial\beta_0}\sum_{i=1}^{k}\left[y_i-\pi\left(x_i\right)\right]=-\sum_{i=1}^{k}\frac{\partial}{\partial\beta_0}\pi\left(x_i\right)=-\sum_{i=1}^{k}\pi\left(x_i\right)\left[1-\pi\left(x_i\right)\right]=-\phi\sum_{i=1}^{k}\pi_0\left[1-\pi_0\right]$$

De même :

$$\frac{\partial^2}{\partial^2\beta_1}\ln f(y,\beta)=\frac{\partial}{\partial\beta_1}\sum_{i=1}^{n}x_i\left[y_i-\pi\left(x_i\right)\right]=\sum_{i=1}^{k}x_i^2\pi\left(x_i\right)\left[1-\pi\left(x_i\right)\right]=\phi\sum_{i=1}^{k}x_i^2\pi_0\left[1-\pi_0\right]$$

et :

$$\frac{\partial^2}{\partial\beta_0\partial\beta_1}\ln f(y,\beta)=\frac{\partial}{\partial\beta_1}\sum_{i=1}^{k}\left[y_i-\pi\left(x_i\right)\right]=-\sum_{i=1}^{k}x_i\pi\left(x_i\right)\left[1-\pi\left(x_i\right)\right]=-\phi\sum_{i=1}^{k}x_i\pi_0\left[1-\pi_0\right].$$

Ces dérivées ne dépendent plus des y_i et elles restent inchangées quand on passe aux espérances mathématiques ; d'où :

$$\mathbf{I}'\left(\beta\right)=\phi\begin{pmatrix}\sum\limits_{i=1}^{k}\pi_0\left[1-\pi_0\right] & \sum\limits_{i=1}^{k}x_i\pi_0\left[1-\pi_0\right]\\ \sum\limits_{i=1}^{k}x_i\pi_0\left[1-\pi_0\right] & \sum\limits_{i=1}^{k}x_i^2\pi_0\left[1-\pi_0\right]\end{pmatrix}.$$

$$\Longrightarrow\mathbf{I}'\left(\beta\right)=\phi I(\beta)$$

où $I(\beta)$ est la matrice d'information de Fisher de l'échantillon en l'absence de sur-dispersion. Comme β est inconnu, il nous faut estimer $\mathbf{I}'\left(\beta\right)$ par $\mathbf{I}'\left(\widehat{\beta}\right)$, c'est-à-dire substituer dans l'écriture de $\mathbf{I}'\left(\beta\right)$ ci-dessus $\widehat{\pi}_0$ à π_0. En inversant $\mathbf{I}'\left(\widehat{\beta}\right)$, on obtient une estimation de $\mathbf{V}'\left(\widehat{\beta}\right)$ que nous notons :

$$\mathbf{V}'\left(\widehat{\beta}\right)=\phi\begin{pmatrix}\sum\limits_{i=1}^{k}\widehat{\pi}_0\left[1-\widehat{\pi}_0\right] & \sum\limits_{i=1}^{k}x_i\widehat{\pi}_0\left[1-\widehat{\pi}_0\right]\\ \sum\limits_{i=1}^{k}x_i\widehat{\pi}_0\left[1-\widehat{\pi}_0\right] & \sum\limits_{i=1}^{k}x_i^2\widehat{\pi}_0\left[1-\widehat{\pi}_0\right]\end{pmatrix}$$

$$\Longrightarrow\mathbf{V}'\left(\widehat{\beta}\right)=\phi\begin{pmatrix}s^2\left(\widehat{\beta}_0\right) & s^2\left(\widehat{\beta}_0,\widehat{\beta}_1\right)\\ s^2\left(\widehat{\beta}_0,\widehat{\beta}_1\right) & s^2\left(\widehat{\beta}_1\right)\end{pmatrix}=\phi\mathbf{V}\left(\widehat{\beta}\right)$$

Où $s^2\left(\widehat{\beta}_0\right)$ est une estimation de la variance de $\widehat{\beta}_0$, $s^2\left(\widehat{\beta}_1\right)$ est une estimation de la variance de $\widehat{\beta}_1$ et $s^2\left(\widehat{\beta}_0,\widehat{\beta}_1\right)$ est une estimation de la covariance entre $\widehat{\beta}_0$ et $\widehat{\beta}_1$ lorsqu'on n'a pas prise en compte la présence de la sur-dispersion.

On peut étendre ce procédé de calcul pour le cas d'une régression logistique multiple en procèdant ainsi qu'il suit :

L'élément à la position $(j;j)$ de cette marice de Fisher est obtenu par :,

$$\begin{aligned}\frac{\partial^2 l(\beta;y,x)}{\partial\beta_j^2} &= \frac{\partial}{\partial\beta_j}\sum_{i=1}^{k}x_{ij}(y_i-\pi_i)=\sum_{i=1}^{k}x_{ij}\frac{\partial}{\partial\beta_j}(y_i-\pi_i)=-\sum_{i=1}^{k}x_{ij}\frac{\partial}{\partial\beta_j}\pi_i\\ &= -\sum_{i=1}^{k}x_{ij}\frac{\partial}{\partial\beta_j}\cdot\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}\\ &= -\sum_{i=1}^{k}x_{ij}\left\{\frac{x_{ij}\exp(x_i'\beta)}{(1+\exp(x_i'\beta))^2}\right\}\\ &= -\phi\sum_{i=1}^{k}x_{ij}^2\pi_0(1-\pi_0)\end{aligned}$$

$$\frac{\partial^2 l(\beta;y,x)}{\partial\beta_j^2}=-\phi\sum_{i=1}^{n}x_{ij}^2\pi_0(1-\pi_0) \tag{2.3}$$

et, par un raisonnement similaire à celui ci-dessus, on détermine l'élément en position $(j;l)$ de cette matrice qui est obtenu par

$$\frac{\partial^2 l(\beta;y,x)}{\partial\beta_j\partial\beta_l}=-\phi\sum_{i=1}^{m_i}x_{ij}x_{il}\pi_0(1-\pi_0)$$

On a alors la matrice d'information de Fisher :

$$\mathbf{I}(\beta)=\phi\begin{pmatrix}-\mathbf{E}\left(\frac{\partial^2}{\partial^2\beta_0}l(\beta;y;x)\right) & -\mathbf{E}\left(\frac{\partial^2}{\partial\beta_0\partial\beta_1}l(\beta;y;x)\right) & \cdots & -\mathbf{E}\left(\frac{\partial^2}{\partial\beta_0\partial\beta_{k-1}}l(\beta;y;x)\right)\\ -\mathbf{E}\left(\frac{\partial^2}{\partial\beta_0\partial\beta_1}l(\beta;y;x)\right) & -\mathbf{E}\left(\frac{\partial^2}{\partial^2\beta_1}l(\beta;y;x)\right) & \cdots & -\mathbf{E}\left(\frac{\partial^2}{\partial\beta_1\partial\beta_{k-1}}l(\beta;y;x)\right)\\ \vdots & \vdots & \ddots & \vdots\\ -\mathbf{E}\left(\frac{\partial^2}{\partial\beta_0\partial\beta_{k-1}}l(\beta;y;x)\right) & -\mathbf{E}\left(\frac{\partial^2}{\partial\beta_1\partial\beta_{k-1}}l(\beta;y;x)\right) & \cdots & -\mathbf{E}\left(\frac{\partial^2}{\partial^2\beta_{k-1}}l(\beta;y;x)\right)\end{pmatrix}$$

On remplace ensuite les valeurs estimées des β_i dans l'expression de la matrice d'information de Fisher pour obtenir $\mathbf{I}\left(\widehat{\beta}\right)$. L'inverse de cette matrice égale à $\widehat{\mathbf{Var}}\left(\widehat{\beta}\right)$, est un estimateur convergent de $\mathbf{Var}\left(\widehat{\beta}\right)$.

En général, ces calculs sont effectués à l'aide des programmes de calcul numérique intégrés dans certains logiciels d'analyse statistique tels **R**, **SAS, STATISTICA**, **LOGISTIC**...etc.

2.3.3 Sur la statistique du Khi-deux de Pearson

Notons $\mathcal{X}'^2$, la statistique du Khi-deux de Pearson en présence de la sur-dispersion. On sait que $\mathcal{X}'^2=\sum_{i=1}^{k}\frac{(y_i-\widehat{\pi}(x_i))^2}{\widehat{\pi}(x_i)(n_i-n_i\widehat{\pi}(x_i))}$ où $\widehat{\pi}(x_i)$ est l'estimée de $\pi(x_i)$ dans chaque classe i sans la prise en compte de la présence de la sur-dispersion.

On a donc

$$\begin{aligned}
\mathcal{X}'^2 &= \sum_{i=1}^{k} \frac{(y_i - n_i\widehat{\pi}(x_i))^2}{\widehat{\pi}(x_i)(n_i - n_i\widehat{\pi}(x_i))} \\
&= \sum_{i=1}^{k} \frac{n_i\left(\frac{y_i}{m_i} - \widehat{\pi}(x_i)\right)^2}{n_i\widehat{\pi}(x_i)(1-\widehat{\pi}(x_i))} = \sum_{i=1}^{k} \frac{\left(\frac{y_i}{n_i} - \widehat{\pi}(x_i)\right)^2}{\widehat{\pi}(x_i)(1-\widehat{\pi}(x_i))} \\
&= \frac{1}{\phi}\sum_{i=1}^{k} \frac{\left(\frac{y_i}{n_i} - \widehat{\pi}_0\right)^2}{\widehat{\pi}_0(1-\widehat{\pi}_0)} \\
&= \frac{1}{\phi}\mathcal{X}^2
\end{aligned}$$

Remarque 2.3.2. *Ainsi, la statistique du khi-deux de Pearson sera sous évaluée en présence de la sur-dispersion lorsqu'on ne l'a pas prise en compte car* $\phi \succ 1$.

2.3.4 Sur la déviance résiduelle

Notons par $\mathcal{D}'_r$, la déviance résiduelle de cet échantillon. On a

$$\begin{aligned}
\mathcal{D}'_r &= -2\left(\widehat{\mathcal{L}}'_{\mathcal{M}} + \widehat{\mathcal{L}}'_s\right) = -\frac{2}{\phi}\left(\widehat{\mathcal{L}}_{\mathcal{M}} + \widehat{\mathcal{L}}_s\right) \\
&= \frac{1}{\phi}\left(-2\left(\widehat{\mathcal{L}}_{\mathcal{M}} + \widehat{\mathcal{L}}_s\right)\right) = \frac{1}{\phi}\mathcal{D}_r
\end{aligned}$$

où $\mathcal{D}_r$ est la déviance résiduelle de l'échantillon en l'absence de la sur-dispersion.

Remarque 2.3.3. *Ainsi, la statistique de la déviance résiduelle sera sous évaluée en présence de la sur-dispersion lorsqu'on ne l'a pas prise en compte car* $\phi \succ 1$.

2.4 Estimation du paramètre de sur-dispersion

L'estimation du paramètre de sur-dispersion a été démontrée par [8] et [10] qui proposent l'une ou l'autre des deux méthodes suivantes :

- Le paramètre de dispersion est estimé par le rapport du khi-deux de Pearson à son nombre de degrés de liberté ;
- Le paramètre de dispersion est estimé par le rapport de la déviance résiduelle à son nombre de degrés de liberté.

A) *Lorsque les effectifs n_i sont différents*

Supposons que les observations sont indépendantes dans chaque groupe i, alors $Y_i \sim \mathcal{B}(n_i; \pi_i)$ avec une espérance $\mathbf{E}(Y_i) = \mu_i = n_i\pi_i$ et une variance

$$\mathbf{Var}(Y_i) = \phi n_i\pi_i(1-\pi_i) = \phi\mu_i(n_i - \mu_i)/n_i$$

Le calcul de la statistique de Pearson dans ces conditions nous donne :

$$X'^2 = \sum_{i=1}^{k} \frac{(y_i - \widehat{\pi}_i)^2}{\phi n_i \widehat{\pi}_i (1 - \widehat{\pi}_i)} = \frac{X^2}{\phi}$$

D'après la **proposition 2.4**, $X'^2 \sim \chi^2_{k-r-1}$ et donc on a $\frac{X^2}{\phi} \approx k - r - 1$. D'où

$$\widehat{\phi} = \frac{1}{k - r - 1} \sum_{i=1}^{k} \frac{(y_i - \widehat{\pi}_i)^2}{n_i \widehat{\pi}_i (1 - \widehat{\pi}_i)} \tag{2.4}$$

B) *Lorsque les effectifs n_i sont tous égaux*

Dans cette situation, [10] nous montre que le paramètre ϕ est alors obtenu directement à l'aide de l'égalité (1.13). Ici, cette estimation peut aussi être obtenue en utilisant la statistique de la déviance résiduelle définie par

$$\widehat{\phi} = \frac{\chi^2_{\upsilon_0} \left(\textit{Déviance} \text{ résiduelle}\right)}{\upsilon_0} \tag{2.5}$$

où υ_0 est le nombre de degré de libertés.

Sur-dispersion causée par une corrélation entre les réponses binaires

Elle se fait sous plusieurs étapes :

Notation d'un modèle de régression logistique avec des données corrélées

1. n est le nombre d'individus de l'étude ;
2. Y_i représente le vecteur des variables dépendantes du i^e individu ; soit $Y_i = (Y_{i1}; Y_{i2}; ...; Y_{im_i})'$, avec $j = 1; ...; m_i$; où m_i représente le nombre d'observations réalisées sur l'individu i;
3. Le nombre total d'observations est défini par $\sum_{i=1}^{n} m_i = N$; avec N représentant le nombre total d'observations effectuées sur les n individus ;
4. x_i correspond à la matrice des variables indépendantes de l'individu i; où $x_i = (x_{i1}; x_{i2}; ...; x_{im_i})'$;
5. $\mu_{ij}(\beta)$ est l'espérance Y_{ij} sachant x_{ij} pour l'individu i.

On utilise une matrice symétrique communément appelée "matrice de corrélation de travail", notée $\mathrm{R}_i(\alpha)$ (voir [4]) et définie telle que ci-dessous, où α est le vecteur de corrélation à estimer.

$$R_i(\alpha) = corr(Y_i|X_i) = \begin{pmatrix} 1 & & \\ corr(Y_1; Y_2|X_1; X_2) & 1 & \\ \vdots & & \ddots \\ corr(Y_1; Y_n|X_1; X_n) & \cdots & corr(Y_{n-1}; Y_n|X_{n-1}; X_n) \quad 1 \end{pmatrix}$$

L'idée est d'essayer de spécifier la vraie structure de corrélation des Y_i. Si la structure de corrélation est bonne, alors les inférences sur β seront plus précises.

La structure de la matrice de corrélation de travail $R_i(\alpha)$ est en lien avec le plan d'expérience et le type d'association possible entre les observations d'un individu. Voici quelques structures communes pour $R_i(\alpha)$.

Les différents types de matrices de corrélation

[4] nous propose les différents types de matrice en fonction de la nature de cette corrélation

1. La matrice de corrélation ayant la structure de type "**indépendante**" correspond à l'absence de corrélation entre Y_j et Y_k pour $j \neq k$.

$$R_i = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

2. La structure de type "**échangeable**" indique que la corrélation entre deux observations d'un même individu est la même et égale à une valeur constante α pour toute paire d'observations,

$$R_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \cdots \\ \alpha & 1 & \alpha & \cdots \\ \alpha & \alpha & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

3. La structure de type "**non structuré**" qui est le type de corrélation qui diffère pour toutes les paires d'observations,

$$R_i(\alpha) = \begin{pmatrix} 1 & \alpha_{1,2} & \cdots & \alpha_{1,n_i} \\ \alpha_{1,2} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha_{n_{i-1},n_i} \\ \alpha_{1,n_i} & \vdots & \alpha_{n_i-1,n_i} & 1 \end{pmatrix}$$

4. La structure "**auto-régressive d'ordre 1**" considère que la corrélation entre deux observations d'un même individu diminue de manière géométrique lorsque les observations

se distancent ($|j-k|$ augmente) dans le temps (ou l'espace),

$$R_i(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha^{n_i-2} & \alpha^{n_i-1} \\ \alpha & 1 & \ddots & \vdots & \alpha^{n_i-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \alpha^{n_i-2} & \vdots & \ddots & 1 & \alpha \\ \alpha^{n_i-1} & \alpha^{n_i-2} & \cdots & \alpha & 1 \end{pmatrix}$$

C'est ce paramètre estimé que les logiciels d'analyse statistique tels que [20] et d'autres comme (LOGISTIC ; SAS ; ...etc) utilisent pour réeffectuer les analyses.

2.5 Correction des effets de la présence de la sur-dispersion

Lorsque la présence de la sur-dispersion est constatée, sa correction par la méthode du khi-deux résiduel consiste à faire appel à certaines options de l'énoncé MODEL : AGGREGATE qui permet de structurer les données suivant les modalités des variables concernées, suivi de l'énoncé SCALE=P ou D, suivant que l'on veuille faire appel au khi-deux résiduel de Pearson ou à la déviance résiduelle pour redéfinir l'échelle des variances-covariances en fonction du paramètre de sur-dispersion. [15] propose d'utiliser l'une des méthodes suivantes :

Equations d'estimations généralisées

Considérons le cas où les variables $Y_{ij}|x_{ij} \backsim \mathcal{B}(\mu_{ij})$ où $i = 1;...;n$ et $j = 1;...;n_i$ avec l'hypothèse possiblement erronée d'indépendance entre les observations. On a alors

$$P(Y_{ij} = y_{ij}|x_{ij}) = \mu_{ij}^{y_{ij}} (1-\mu_{ij})^{(1-y_{ij})} \quad , \qquad y_{ij} = 0;1$$

La fonction de lien canonique est le lien logit ; c'est-à-dire

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta' x_{ij} = \eta_{ij} \iff \mu_{ij} = \frac{\exp(\beta' x_{ij})}{1+\exp(\beta' x_{ij})} = \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}$$

et donc

$$\begin{aligned} P(Y_{ij} = y_{ij}|x_{ij}) &= \left(\frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}\right)^{y_{ij}} \left(\frac{1}{1+\exp(\eta_{ij})}\right)^{(1-y_{ij})} \\ &= \exp\left\{y_{ij}\log\left(\frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}\right)\right\} \times \exp\left\{(1-y_{ij})\log\left(\frac{1}{1+\exp(\eta_{ij})}\right)\right\} \\ &= \exp\left\{y_{ij}\log\left(\frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}\right) + (1-y_{ij})\log\left(\frac{1}{1+\exp(\eta_{ij})}\right)\right\} \\ &= \exp\left[y_{ij}\eta_{ij} - \log(1+\exp(\eta_{ij}))\right] \end{aligned}$$

Puisque les observations dans chaque groupe sont supposées indépendantes, nous écrivons la probabilité d'un vecteur de résultat d'un groupe i comme

$$P\left(Y_i = y_i | x_i\right) = \prod_{j=1}^{n_i} \exp\left[y_{ij}\eta_{ij} - \log\left(1 + \exp\left(\eta_{ij}\right)\right)\right]$$

La vraisemblance est

$$\begin{aligned} L\left(\beta\right) &= \prod_{i=1}^{n} P\left(Y_i = y_i | x_i\right) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{n_i} \exp\left[y_{ij}\eta_{ij} - \log\left(1 + \exp\left(\eta_{ij}\right)\right)\right] \\ &= \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[y_{ij}\eta_{ij} - \log\left(1 + \exp\left(\eta_{ij}\right)\right)\right]\right) \end{aligned}$$

La log-vraisemblance est

$$\mathcal{L}\left(\beta\right) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[y_{ij}\eta_{ij} - \log\left(1 + \exp\left(\eta_{ij}\right)\right)\right]$$

Pour $t = 1; ...; p$, on a

$$\frac{\mathcal{L}\left(\beta\right)}{\partial\beta_t} = \frac{\mathcal{L}\left(\beta\right)}{\partial\eta} \cdot \frac{\partial\eta}{\partial\beta_t} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_{ij} - \frac{\exp\left(\eta_{ij}\right)}{1 + \exp\left(\eta_{ij}\right)}\right) x_{ijt}$$

où, sous forme matricielle,

$$\mathbf{U}_{indep}\left(\beta\right) = \frac{\mathcal{L}\left(\beta\right)}{\partial\beta_t} = \sum_{i=1}^{n} X_i'\Delta_i\left\{Y_i - \mu_i\left(\beta\right)\right\} \tag{2.6}$$

où $X_i = \left(x_{i1}; ...; x_{in_i}\right)'$ avec $x_{ij} = \left(x_{ij1}; ...; x_{ijp}\right)'$, $\mu_i\left(\beta\right) = \left(\mu_{i1}\left(\beta\right); ...; \mu_{in_i}\left(\beta\right)\right)'$ et où Δ_i est une matrice diagonale de taille $n_i \times n_i$ donc l'élément en position (j, j) est $\frac{\exp(\eta_{ij})}{(1+\exp(\eta_{ij}))^2} = \partial\mu_{ij}/\partial\eta_{ij}$.

Puisqu'on a supposé qu'il y a indépendance entre les observations, on a $\mu_{ij} = \eta_{ij}$ et donc $\partial\mu_{ij}/\partial\eta_{ij} = 1$, donc $\Delta_i = \mathbf{I}_{n_i \times n_i}$.

Si on définit la matrice $\mathbf{A}_i$ comme étant la matrice diagonale dont l'élément en position (j, j) est $\mathbf{Var}\left(\mu_{ij}\right)$. On a donc

$$\widehat{\mathbf{V}} = \left(\sum_{i=1}^{n} x_i'\Delta_i\mathbf{A}_i\boldsymbol{\Delta}_i x_i'\right)^{-1}$$

La forme matricielle (2.6) ainsi obtenue est appelée **matrice de travail**.

Proposition 2.1.

$$\sqrt{n}\left(\beta_{ind} - \beta\right) \backsim \mathcal{N}\left(0; n\left(\sum_{i=1}^{n} x_i'\Delta_i\mathbf{A}_i\boldsymbol{\Delta}_i x_i'\right)\right)$$

où

$$\mathbf{A}_i = \begin{pmatrix} \pi_{i1}(1-\pi_{i1}) & 0 & \cdots & 0 \\ 0 & \pi_{i2}(1-\pi_{i2}) & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & \cdots & \pi_{in_i}(1-\pi_{in_i}) \end{pmatrix}$$

Preuve. Elle est immédiate en utilisant la preuve de la **proposition 1.1** ■

On pose l'équation (2.6) égale à zéro et sa résolution identique à l'équation (1.10) permet de déterminer l'estimateur du maximum de vraisemblance $\widehat{\beta}$ de β. Cependant, l'indépendance entre les observations n'étant plus vérifiée à cause de la corrélation, on applique la méthode d'**équations d'estimations généralisées** utilisant l'algorithme de la quasi-vraisemblance et dévéloppée par [13] pour l'estimation des paramètres. C'est en fait une généralisation du système d'équations (2.6) et prend en compte l'existence d'une corrélation autre que l'indépendance entre les observations dans les différents groupes . [14] ont monré ensuite que la matrice de variances pour les obervations sur l'individu Y_i est donnée par :

$$\mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i(\alpha) \mathbf{A}_i^{1/2} \tag{2.7}$$

De l'équation (2.7) les paramètres à estimer sont ϕ et α et sont appelés paramètres de sur-dispersion, et [14] montrent qu'on peut les estimer à partir des résidus de Pearson qui sont définis comme suit :

$$r_{ij} = \frac{Y_{ij} - \mu_{ij}}{\sqrt{\mathbf{Var}(\mathbf{Y}_{ij})}} \tag{2.8}$$

et obtenir ainsi l'estimé du paramètre ϕ par :

$$\widehat{\phi} = \frac{1}{N-(k+1)} \sum_{i=1}^{n} \sum_{j=1}^{n_i} r_{ij}^2 \tag{2.9}$$

L'estimation de α dépend de la nature de la forme de $\mathbf{R}_i(\alpha)$ choisie. En effet,

- Si $\mathbf{R}_i(\alpha)$ est de type "**échangeable**", [15] proposent que α est estimé par :

$$\widehat{\alpha} = \frac{1}{\widehat{\phi}} \sum_{i=1}^{n} \left(\frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \widehat{r}_{iu} \widehat{r}_{iv} - \sum_{i=1}^{n} \widehat{r}_{iu}^2}{n_i(1-n_i)} \right)$$

- Si $\mathbf{R}_i(\alpha)$ est de type "**auto régressive**", [16] proposent que α est estimé par :

$$\widehat{\alpha} = \frac{1}{(K_1 - P)\,\widehat{\phi}} \sum_{i=1}^{n} \sum_{j=1}^{n_i - 1} \widehat{r}_{ij} \widehat{r}_{ij+1}$$

où $K_1 = \sum_{i=1}^{n} (n_i - 1)$ et $P = \dim(\beta)$ est le nombre de paramètres imposés par le modèle logistique.

– Si $\mathbf{R}_i(\alpha)$ est de type "**non structuré**", il n'existe pas une méthode d'estimation de α car $R_i(\alpha)$ peut ne pas être inversible.

Réestimation des coefficients du modèle

Pour estimer β, on résout le système (2.11) ci-dessous appelé **système d'équations d'estimation généralisées** en utilisant la **méthode de la quasi-vraisemblace**.

$$\mathbf{U}_{corr}(\beta) = \sum_{i=1}^{N} (\mathbf{A}_i \Delta_i x_i)' \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \mu_i(\beta)\} = 0 \tag{2.10}$$

Pour cela, on a recours à l'algorithme itératif de Newton-Raphson qui s'éxécute comme suit appelé **Algorithme de maximisation par la méthode de quasi-vraisemblance :**

Algorithme 2 : Maximisation par la quasi-vraisemblance

Soit $\widehat{\mathbf{D}}_i = \widehat{\phi}\mathbf{A}_i\Delta_i x_i$ et $\widehat{\mathbf{V}}_i = \widehat{\phi}\mathbf{A}_i^{1/2}\mathbf{R}_i(\widehat{\alpha})\mathbf{A}_i^{1/2}$

1-) Estimer β sous l'hypothèse d'indépendance et dénoter l'estimateur obtenu $\widehat{\beta}^{(0)}$;

2-) Estimer α et ϕ à partir de $\widehat{\beta}$ et des r_{ij}

3-) Poser $\mathbf{V}_i = \widehat{\phi}\mathbf{A}_i^{1/2}\mathbf{R}_i(\widehat{\alpha})\mathbf{A}_i^{1/2}$

4-) Mettre la valeur $\widehat{\beta}$ à jour à partir de l'équation

$$\widehat{\beta}^{(m+1)} = \widehat{\beta}^{(m)} + \left(\sum_{i=1}^{N} \widehat{\mathbf{D}}_i \widehat{\mathbf{V}}_i \widehat{\mathbf{D}}_i'\right)^{-1} \left[\sum_{i=1}^{N} \widehat{\mathbf{D}}_i' \widehat{\mathbf{V}}_i^{-1} \left\{Y_i - \mu_i\left(\beta^{(m-1)}\right)\right\}\right] \tag{2.11}$$

où $\mathbf{D}_i$ **et** $\mathbf{V}_i$ sont évaluées en $\beta = \widehat{\beta}^m$

5-) Itérer les étapes 2 à 4 jusqu'à ce que la différence entre $\widehat{\beta}^{m+1}$ et $\widehat{\beta}^m$ soit négligeable. [8] propose qu'on arrête le processus lorsque $|\widehat{\beta}_s^{(m+1)} - \widehat{\beta}_s^{(m)}| \prec 10^{-6}$.

Estimation de la matrice de variances-covariances

Si la matrice de corrélation $\mathbf{R}_i(\alpha)$ est correctement spécifiée, la matrice de variances-covariances de $\widehat{\beta}$ s'estime de façon convergente par

$$\mathbf{V}_t = \left(\sum_{i=1}^{N} \widehat{\mathbf{D}}_i \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{D}}_i'\right)^{-1} \tag{2.12}$$

Cependant, $\mathbf{R}_i(\alpha)$ ne reflète pas toujours la véritable corrélation qui existe entre les observations, pour rémédier à ce problème, [15] proposent une méthode pour corriger, de façon

empirique, la matrice $\mathbf{V}_t$ en considérant plutôt l'estimateur "sandwich" robuste défini par :

$$\mathbf{V}_s = \mathbf{V}_t \left(\sum_{i=1}^{N} \widehat{\mathbf{D}}_i' \widehat{\mathbf{V}}_i^{-1} \left\{ Y_i - \mu_i \left(\beta \right) \right\} \left\{ Y_i - \mu_i \left(\beta \right) \right\}' \mathbf{V}_i^{-1} \widehat{\mathbf{D}}_i \right) \mathbf{V}_t \tag{2.13}$$

[15] nous disent que le terme "sandwich" vient du fait que dans l'expression (2.13) ; une correction empirique est prise en "sandwich" entre deux estimateurs de la variance basés sur le modèle de la matrice de travail.

⋆ ⋆ Chapitre 3 ⋆ ⋆

Applications

Les données de notre analyse sont tirées de l'article paru en 1976 dans [1] concernant une étude faite entre le 1^{er} juin 1973 et le 31 mai 1974. Pendant cette période, environ 150 femmes se plaignant d'une infertilité secondaire ont été admises au département de Gynécologie et Obstétries de l'Université médicale d'Athènes. Est considérée souffrant d'une infertilité secondaire la patiente qui a déjà accouché ou tout au moins celle qui a été enceinte ; si elle a été mariée ; le sperme du conjoint est jugé de bonne qualité ; de plus si la patiente a essayé de chercher à tomber enceinte depuis une periode d'au moins 18 mois. Cent dix de ces patientes ont rempli ces critères et parmi celles-ci, on choisi cent patientes sur lesquelles on a effectué des études supplémentaires en tenant compte :

1. du niveau d'étude : moins de 6 années d'études, entre 6 et 11 années d'études et celles qui ont plus de 12 années d'études
2. la conduite à terme de la derniere grossesse à travers un bon suivi médical.

Quatre vingt trois de ces patientes ont pu remplir ces conditions et les résultats de cette étude sont dans notre base de données.

Nous allons effectuer une analyse comparative dans le cas d'une régression logistique simple dans un premier temps, puis nous effectuerons une analyse dans le cas de régression multiple et ceci, en l'absence de la sur-dispersion et ensuite en tenant compte de la présence de la sur-dispersion. Pour cela, nous allons utiliser le logiciel **R** qui parait plus adapté pour notre étude car présente une analyse complète de tous les tests pratiqués.

3.1 Comparaison des résultats dans le cas d'un modèle de régression logistique simple

On a effectué une analyse des données de notre base à l'aide d'une fonction de distribution logistique simple en étudiant la relation entre une stérilité secondaire et les avortements

spontanés qu'on a nommé "spontaneous". Dans un premier temps, on a effectué l'analyse avec l'option "naif" de **R** qui ne prend pas en compte la présence de la sur-dispersion, puis on a effectué l'analyse de la relation entre "secondary infertility" et "spontaneous" en faisant appel à l'option "exact" de **R** qui prend en compte la présence de la sur-dispersion. Les résultats comparatifs se trouvent dans le **tableau 4.1** ci-dessous.

On aurait pu aussi utiliser l'option "efron" ; "approximate"ou"breslow" de **R** qui prennent aussi en compte la présence de la sur-dispersion dans l'analyse des données.

Tableau 4.1 : Relation entre "secondary infertility" et "spontaneous"

	"naïf"	"exact"	Variation
β_1	1,0639	1,1768	0,1129
$\exp(\beta_1)$	2,8976	3,2441	0,3465
p	0	0	0
S.E* en %	**0,1964**	**0,2315**	**17,88**
IC(95%)	[5, 002 ;54, 89]	[7, 85 ;165, 17]	
R^{-2}	0,111(max = 0, 449)	0,127(max = 0, 519)	0,016
LRT	26,54(p=$4{,}87.10^{-9}$)	33,76(p=$6{,}23.10^{-9}$)	7,22
Wald	20,18(p=$2{,}89.10^{-7}$)	25,84(p=$3{,}71.10^{-7}$)	5,66
Score test	26,66(p=$4{,}02.10^{-9}$)	34,13(p=$5{,}1510^{-9}$)	7,47

Une analyse de l'étude de la cause de l'infertilité secondaire en fonction d'un avortement spontané sans la prise en compte de la présence de la sur-dispersion donne $\exp(\beta_1) = \mathbf{OR} = \mathbf{2, 89}$. Ceci signifie qu'on a 2,89 fois plus de chance d'avoir une infertilité secondaire dont la cause est due à un avortement spontané ; avec une probabilité $p = 6, 05.10^{-8}$ un ecart-type $S.E = 0, 1964$. De plus, la probabilité d'avoir une infertilité secondaire provoquée par un avortement spontané est de

$$\pi = \frac{\exp(-1, 3739 + 1, 0639x)}{1 + \exp(-1, 3739 + 1, 0639x)}.$$

Elle serait donc

$$\frac{\exp(-1, 3739 + 1, 0639)}{1 + \exp(-1, 3739 + 1, 0639)} = 42, 31\%$$

pour une patiente qui a déjà été victime au moins d'un avortement spontané et elle serait de

$$\frac{\exp(-1, 3739)}{1 + \exp(-1, 3739)} = 20, 19\%$$

pour une patiente qui n'a jamais été victime d'avortement spontané.

Cependant, une analyse de la relation entre l'infertilité secondaire et un avortement spontané en prenant en compte la présence de la sur-dispersion en utilisant l'option "exact" de **R** donne

$\exp(\beta_1) = \mathbf{OR} = \mathbf{3,24}$. Ceci signifie qu'on a **3,24** fois plus de chance d'avoir une infertilité secondaire dont la cause est due à un avortement spontané ; ceci avec une probabilité $p = 3,71.10^{-7}$ et un écart-type $S.E = 0,2315$ qui est largement supérieure aux prédictions faites lorsqu'on n'a pas pris en compte la présence de cette sur-dispersion. De plus, la probabilité d'avoir une infertilité secondaire provoquée par un avortement spontané est de

$$\pi = \frac{\exp(-1,3739 + 1,1768x)}{1 + \exp(-1,3739 + 1,1768x)}.$$

Elle serait donc

$$\frac{\exp(-1,3739 + 1,1768)}{1 + \exp(-1,3739 + 1,1768)} = 45,08\%$$

pour une patiente qui a déjà eu un avortement spontané et elle serait $\frac{\exp(-1,3739)}{1+\exp(-1,3739)} = 20,19\%$ pour une patiente qui n'a jamais été victime d'un avortement spontané.

3.2 Comparaisons des résultats de l'analyse dans le cas d'un modèle de régression logistique multiple

Ici, l'analyse des données de la base est effectuée pour établir la nature de la relation qui existe entre "secondary infertility" et deux variables explicatives qui sont "spontaneous" et "induced" simultanement. Comme dans le cas de la régression logistique simple, nous étudierons dans un premier temps cette relation sans tenir compte de la présence de la sur-dispersion à l'aide de l'option "naïf " de **R** ; puis, nous étudierons cette relation en considérant la présence de la sur-dispersion à l'aide de l'option "exact" de **R.**

Les résultats comparatifs de ces deux tests se trouvent dans le tableau 4.2 ci-dessous

Tableau 4.2 : relation entre "secondary infertility" et "spontaneous et induced"

3.2. Comparaisons des résultats de l'analyse dans le cas d'un modèle de régression logistique multiple

		"naïf"	"exact"	Variation
β_1	spontaneous	1,1972	1,9859	0,7887
β_2	induced	0,4181	1,4090	0,9909
$\exp(\beta_1)$	spontaneous	3,331	7,285	3,954
$\exp(\beta_2)$	induced	1,52	4,092	2,572
p_1	spontaneous	$1{,}54.10^{-8}$	$1{,}75.10^{-8}$	
p_2	induced	0,042	$9{,}38.10^{-5}$	
S.E$_1$*en %	**spontaneous**	**0,171**	**0,212**	**20,47**
S.E$_2$*en %	**induced**	**0,182**	**0,206**	**13,18**
I.C$_1$ (95%)	spontaneous	$[24,53; 3,44.10^5]$	$[38,45 ; 2,06.10^6]$	
I.C$_2$ (95%)	induced	$[5,87 ; 1,45.10^3]$	$[7,53 ; 4,02.10^3]$	
R^2		0,169(max = 0,455)	0,193(max = 0,519)	0,24
LRT		47,03	53,15(p=286.10^{-12})	6,12
Wald		28,18	31,84(p=$1{,}22.10^{-7}$)	3,66
Score test		42,86	48,44(p=$3{,}03.10^{-11}$)	5,58

L'option "naif" de **R** nous montre que le coefficient de "spontaneous" $\beta_1 = 1,1972$ d'exponentielle $\exp(\beta_1) = \mathbf{OR} = 3,331$ et que le coefficient de "induced" $\beta_2 = 0,4181$ d'exponentielle $\exp(\beta_2) = \mathbf{OR} = 1,52$. Ceci signifie que lorsqu'on prend compte à la fois les avortements spontanés et les avortements provoqués, on a 3,331 fois plus de chance d'avoir une infertilité secondaire provoquée par avortement spontané avec une probabilité $p = 1,54.10^{-8}$ et un écart-type $S.E = 0,171$. On a aussi 1,52 fois de chance d'avoir une infertilité secondaire provoquée par un avortement induit avec une probabilité $p = 0,042$ et un écart-type $S.E = 0,182$. Dans ces conditions, la probabilité d'avoir une infertilité secondaire est de

$$\pi = \frac{\exp(-1,7079 + 1,1972x_1 + 0,4181x_2)}{1 + \exp(-1,7079 + 1,1972x_1 + 0,4181x_2)}.$$

Elle serait donc

$$\frac{\exp(-1,7079 + 1,1972 + 0,4181)}{1 + \exp(-1,7079 + 1,1972 + 0,4181)} = 47,68\%$$

pour une patiente qui a déjà eue un avortement spontané et un avortement induit ; elle serait

$$\frac{\exp(-1,7079 + 0,4181)}{1 + \exp(-1,7079 + 0,4181)} = 22,63\%$$

pour une patiente qui n'a jamais été victime d'un avortement spontané mais qui a déjà eue des avortements provoqués. Elle serait aussi de

$$\frac{\exp(-1,7079 + 1,1972)}{1 + \exp(-1,7079 + 1,1972)} = 37,50\%$$

pour une patiente qui n'a jamais eue des avortements provoqués, mais qui a déjà été victime des avortements spontanés.

Dans ce tableau, l'analyse des données avec l'option "exact" de **R** nous montre que le coefficient de "spontaneous" $\beta_1 = 1,9859$ d'exponentielle $\exp(\beta_1) = \mathbf{OR} = 7,285$ et le coefficient de "induced" $\beta_2 = 1,4090$ et son exponentielle est $\exp(\beta_2) = \mathbf{OR} = 4,092$. Ceci signifie que lorsqu'on prend compte la présence de la sur-dispersion, on a 7,285 fois plus de chance d'avoir une infertilité secondaire provoquée par avortement spontané avec une probabilité $p = 1,54.10^{-8}$ et un écart-type $S.E = 0,171$. On a aussi 1,52 fois de chance d'avoir une infertilité secondaire provoquée par un avortement induit avec une probabilité $p = 0,042$ et un écart-type $S.E = 0,182$. Dans ces conditions, on estime que la probabilité d'une femme de souffrir d'une infertilité secondaire est de

$$\pi = \frac{\exp(-1,7079 + 1,1972x_1 + 0,4181x_2)}{1 + \exp(-1,7079 + 1,1972x_1 + 0,4181x_2)}$$

. Elle serait donc

$$\frac{\exp(-1,7079 + 1,1972 + 0,4181)}{1 + \exp(-1,7079 + 1,1972 + 0,4181)} = 47,68\%$$

pour une patiente qui a déjà eue un avortement spontané et un avortement induit ; elle serait

$$\frac{\exp(-1,7079 + 0,4181)}{1 + \exp(-1,7079 + 0,4181)} = 22,63\%$$

pour une patiente qui n'a jamais été victime d'un avortement spontané mais qui a déjà eue des avortements provoqués. Elle serait aussi de

$$\frac{\exp(-1,7079 + 1,1972)}{1 + \exp(-1,7079 + 1,1972)} = 37,50\%$$

qui n'a jamais eue d'avortements provoqués, mais qui a déjà été victime d'avortements spontanés.

Il est clair que dans le **Tableau 4.1** comme dans le **Tableau 4.2**, la prise en compte de la présence de la sur-dispersion apporte une réelle amélioration dans l'interprétation des différents résulats des tests. En effet, au **Tableau 4.1** par exemple, **l'inflation de l'écart-type est de**

$$\mathbf{\Delta}_{S.E} = \mathbf{17,88}\%$$

sur les prédictions faites par rapport à la relation entre "**secondary infertility**" et "**spontaneous**" avec une différence de probabilités $\Delta_p = 3,105.10^{-7}$. Le rapport ecart-type en présence de la sur-dispersion sur écart-type en l'absence de la sur-dispersion donne

$$\frac{0,2315}{0,1964} = \frac{\text{Déviance résiduelle}}{ddf} = 1,18$$

Dans le cas du **Tableau 4.2, l'inflation de l'écart-type est de**

$$\mathbf{\Delta}_{S.E_1} = 20,47\%$$

sur les prédictions faites sur la relation entre "secondary infertility" et "sponaneous" lorsqu'on a pas pris en compte la présence de cette sur-dispersion. Aussi, **l'inflation de l'écart-type est de**

$$\Delta_{S.E_2} = 13,18\%$$

sur la relation entre "secondary" et "induced".

Dans le **Tableau 4.1** comme dans le **Tableau 4.2**, la prise en compte de la présence de la sur-dispersion apporte une réelle amélioration sur les différents résultats.

Remarque 3.2.1. Le lien entre la variable à expliquer et la variable explicative est non significatif lorsque l'odds-ratio se réalise avec une probabilité $p \succ 0,05$.

Chapitre 4

Implications Didactiques

Les notions de statistiques et de probabilités, enseignées dans les classes du secondaires des lycées et colléges, ne sont pas toujours bien maîtrisées par les élèves. Ceci est dû en partie ou en totalité à la façon dont ces enseignements sont dispensés dans les salles de classe. En effet, Ces savoirs sont mal appréhendés par les apprenants pour deux raisons : Soit par mauvaise foi de l'enseignant qui refusent de bien dispenser la leçon (ce qui serait vraiment dommage car il a choisi ce metier), ou alors parce que l'enseignant lui-meme ne maitrise pas ces notions. Notre thème : "**Sur-dispersion dans les modèles de régression logistique**" nous a permis de bien comprendre les notions de statistiques, de probabilité et leur applications. Ce qui nous permettra de mieux les enseigner à nos jeunes fréres du secondaire.

Aussi, ce travail nous a permis de nous accommoder à effectuer des recherches en profondeur sur un thème donné tant sur la forme que sur le fond.

La rigueur avec laquelle nous avons écrit ce mémoire nous y conduira à bien préparer nos leçons lorsque nous serons sur le terrain dans l'exercice de notre fonction d'enseignant.

♣ Conclusion ♣

Le thème, soumis à notre étude, nous a amené à nous poser la question : quel impact la prise en compte de la présence de la sur-dispersion dans les analyses des modèles de régression logistique a-t-elle sur les différents résultats des tests trouvés lorsqu'on n'a pas pris en compte cette sur-dispersion ?.

Pour apporter la réponse à notre problématique, nous avons tout d'abord commencé par présenter les modèles de régression logistique et les méthodes qui conduisent à déterminer les coefficients d'un modèle de régression logistique, nous avons ensuite présenter les différents tests pratiqués dans les analyses des modèles de régression logistique : notamment le test du rapport de vraisemblance ; le test basé sur la variance des coefficients du modèle appelé test de Wald et enfin le Score test. Aussi, nous avons présenté les causes de la sur-dispersion.

La présence de la sur-dispersion, en fonction de la nature de sa cause, entraîne un biais important dans l'estimation des paramètres et vient remettre en cause les résultats des tests pratiqués. Nous avons donc essayé de déterminer quelques methodes conduisant à trouver le paramètre de dispersion et avons aussi essayé de présenter une méthode conduisant à déterminer les coefficients du modèle lorsque la sur-dispersion est causée par une forte corrélation entre les réponses binaires. Aussi, nous avons essayé de présenter les methodes qui permettent de corriger les effets causés par la non prise en compte de la sur-dispersion. Les données utilisées pour l'application sont tirées de [2] concernant une étude faite sur la rélation entre "souffrir d'une infertilité secondaire" et les facteurs qui peuvent être des causes de cette infertilité secondaire.

Cependant, nous pouvons déplorer le fait que notre étude rencontre des limites dès que le paramètre de dispersion lui-même n'est pas constant dans les différents groupes formant notre échantillon, car il est difficile de déterminer la variance du paramètre ϕ de dispersion et ensuite l'intégrer dans l'analyse des données.

♣ Bibliographie ♣

[1] Hosmer, D. & Lemeshow, S. (2002) *Applied logistic regression*, Wiley, London.

[2] Trichopoulos, D., et al, (1976) *Induced Abortion And Secondary Infertility,* British Journal of Obstetrics and gyneacology, Vol. 83, 645-650.

[3] Collett, D. (1999) *Modelling Binary data*, Chapman & Hall, London.

[4] Beauregard, B. (2013) *Comparaison des modèles de régression logistique utilisés pour l'analyse des données recueillies dans le cadre d'études du type cas-témoin appariés sur le déplacement animal*, Sherbrooke, Quebec ; Canada.

[5] Bernard, P.-M. (1982) *La régression logistique*. Département de médécine sociale et préventive. Université de Laval. 1982.

[6] Gouriéroux, C. (1972)*Econométrie des variables qualitatives*, Dunod, Paris.

[7] Choubai, R. (2006) *Les fondements théorique de la régression logistique et son utilisation en épidémiologie à l'aide du système SAS*. Sherbrooke, Quebec, Canada. 2006.

[8] McCullagh, P. & Nelder, J. A. (1989) *Generalized linear Models*. Chapman & Hill New York/.

[9] Bhat, C. R. (2001) *Quasi-random maximum simulatted likelihood estimation of the mixed multinomial logit model*. Transportation Research Part B : Methodological, 35 :677-693 cooperative games", Working paper, Faculty of Commerce and Business Administration, University of British Columbia.

[10] Williams, D. (1982a) *Extra binomial Variation in logistic linear models. Applied Statistics* 31, 144-148.

[11] Lopez, C. (2000) *Régression Logistique avec GENMOD*, Service de Biométrie, Vol. 46, pp 42-44.

[12] Agresti, A. (1990) *Categorical data analysis*. Wiley, New-York.

[13] Wedderburn, R. (1974) *Quasi-likelihood functions, generalized linear models and the gauss-newton method*. Biometrika, 61, 439-47.

[14] Zeger, S. L. & Liang, K.Y. (1986) *Longitudinal Data Analysis for Discrete and continuous outcomes*, Biometrics 42(1), 121-130.

[15] Cox, D. R. & Snell, E. J. (1989) *Generalised Linear Models*, seconde edn,Chapman & Hall, London.

[16] Hardin, J. W. & Hilbe, J.M. (2002) *Generalized Estimating Equations*. Chapman & Hall/CRC, Boca Raton, Florida 33431.

[17] Jean Bouyer. (2012) *La régression logistique en épidémiologie*. Master. Epidémiologie Quantitative, M2 recherche en Santé Publique, Universités Paris V, XI, XII, Versailles Saint Quentin.

[18] Dobson, A. J. (1990) *An Introduction to generalized Linear Models*. Chapman & Hall, London.

[19] Marque, S. (2005) *Prise en compte de la surdispersion par des modèles à mélange de poisson*. pp 31-32.

[20] Dean, C. B. (1992) *Testing for overdispersion in Poisson and Binomial regression models*. JASA, 87(418) : 451-7. 1992.

[21] R Core Team R : (2015) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http ://www.R-project.org/.

[22] Jorgensen, B. (1997) *The theory of Dispersion Models*. Chapman & Hall, London.

[23] Turnbull, B. W. & Weiss, L. (1978) *A likelihood ratio statistic for testing goodness of fit with randomly sensored data*. Biometrics, Vol 34(3), 367-75.

[24] Wald, A. (1943) *Tests of statistsical hypothesis concerning several parameters when the number of observations is large*. Trans. Amer. Math. Soc.54 : 426-482.

♣ Annexe ♣

Options de R ayant produit les résultats du tableau 4.1 et du tableau 4.2

```
# Importation du tableau de données sur l'infertilité secondaire
# mdat1<-read.table("C:/Overdispersioncode/infert.txt")library(survival)
# Ajustement des facteurs de confusion
summary(model2 <- glm(case ~age+parity+education+spontaneous+induced, data =
infert, family = binomial()))
```

Cas du tableau 4.1

```
# Analyse des données sur la relation entre "secondary infertility" et
"spontaneous" sans tenir compte de la présence de la sur-dispersion à l'aide
de l'option "naïf" de R ayant produit les résultats de la colonne "naïf" du tableau
4.1
model1 <- glm(case ~spontaneous, data = infert, family = binomial())
summary(model1)
# Analyse des données sur la relation entre "secondary infertility" et
"spontaneous" en prenant en compte la présence de la sur-dispersion à l'aide
de l'option "exact" de R ayant produit les résultats de la colonne "exact" du tableau
4.1
model2 <- clogit(case ~spontaneous+strata(stratum),method="exact",data = infert)
 summary(model2)
```

Cas du tableau 4.2

```
# Analyse des données sur la relation entre "secondary infertility" et
```

```
   "spontaneous et induced" sans tenir compte de la présence de la sur-dispersion
à l'aide de l'option "naïf" de R ayant produit les résultats de la colonne "naïf"
du tableau 4.2
   model1.2 <- glm(case ~spontaneous+induced, data = infert, family = binomial())
   summary(model1.2)
   # Analyse des données sur la relation entre "secondary infertility" et
   "spontaneous et induced" en prenant en compte la présence de la sur-dispersion
à l'aide de l'option "exact" de R ayant produit les résultats de la colonne "exact"
du tableau 4.2
   model2.2 <- clogit(case ~spontaneous+induced+strata(stratum),method="exact",data
= infert)  summary(model2.2)
```