UNIVERSITE DE LIEGE
FACULTE DES SCIENCES

Accord entre observateurs et coefficient Kappa de Cohen

Mémoire présenté par
Sophie Vanbelle
en vue de l'obtention du grade de
licenciée en Sciences Mathématiques

Année académique 2001-2002

Je tiens à remercier mon promoteur, Monsieur le Professeur Adelin Albert, pour avoir accepté de diriger mon mémoire, et surtout pour m'avoir permis de découvrir un sujet très intéressant.

Mes remerciements vont aussi à Mademoiselle Gisèle Merch qui, grâce à sa disponibilité et ses conseils judicieux, m'a permis de mener mon travail à bien.

Enfin, je remercie toutes les personnes de mon entourage qui m'ont chacune soutenue à leur manière lors de la réalisation de ce travail.

Introduction

Ce mémoire est consacré au célèbre "coefficient Kappa (κ)" introduit par Cohen en 1960 pour mesurer le degré d'accord entre observateurs (inter-observer agreement). Ce concept est né dans le domaine des sciences humaines. En psychologie ou en psychiatrie, par exemple, l'état d'un individu ne se mesure pas à l'aide d'un appareil mais il s'évalue le plus souvent à l'aide d'une échelle qualitative ordinale. Ainsi, dans la dépression, un sujet peut être considéré comme "non déprimé", "peu déprimé" ou "fortement déprimé" par un psychiatre sur base d'un questionnaire rempli par le patient. Dans une épreuve de sélection de projets de recherche, il est classique de soumettre chaque dossier à l'avis de plusieurs experts chargés de fournir une appréciation qualitative. Les dossiers retenus sont ceux qui auront recueilli l'accord favorable d'une majorité d'experts.

Dans toute évaluation qualitative, on suppose que les observateurs (on dit aussi "juges" ou "experts") ont un jugement objectif et aussi homogène que possible entre eux. Si ce n'est pas le cas, on peut mettre en cause la partialité de certains juges (comme ce fut le cas aux derniers Jeux Olympiques d'hiver à Salt Lake City où la juge française de patinage artistique fut exclue par le Comité Olympique) ou au contraire une certaine ambiguïté sur l'interprétation de l'échelle d'évaluation. Ce problème est fréquent lorsqu'on traduit un questionnaire dans une autre langue (par exemple de français en néerlandais ou d'anglais en français), la traduction de certaines questions pouvant changer le sens des phrases et dès lors modifier l'interprétation par l'évaluateur. En médecine, lorsqu'on sollicite l'avis d'un autre médecin ("second opinion") pour un problème de santé, on met implicitement en cause le degré d'agrément entre le premier et le second médecin.

Il ressort des considérations précédentes que la mesure de l'accord entre observateurs est un élément essentiel dans la validation d'une échelle d'évaluation ou d'une procédure de classement.

Quarante ans après l'introduction du coefficient Kappa de Cohen, des progrès considérables viennent d'être réalisés par Shoukri et Mian (1996) et Lipsitz et al (2001), grâce aux développements des modèles linéaires généralisés. Ces travaux ont permis d'étudier l'effet de cofacteurs sur le coefficient d'accord entre observateurs. La problématique est double. En effet, le coefficient Kappa de Cohen pourrait varier en fonction des caractéristiques des individus à évaluer. Ainsi, les juges pourraient être davantage d'accord entre eux lors-

qu'il s'agit d'évaluer des candidats masculins plutôt que féminins, ou des sujets jeunes plutôt qu'âgés. D'autre part, le coefficient Kappa de Cohen pourrait varier en fonction des caractéristiques des juges. Par exemple, l'expérience des juges est un facteur essentiel, dans la mesure où des juges expérimentés ont un meilleur accord que des juges non expérimentés. En radiologie, le degré d'accord entre radiologues chevronnés pour évaluer la présence d'une tumeur sur un cliché est meilleur que celui obtenu avec des radiologues en formation. Les travaux récents permettent d'aborder cette ancienne problématique.

Ce mémoire est structuré en cinq chapitres. Le premier chapitre rappelle quelques mesures d'association entre échelles qualitatives. Il introduit le coefficient Kappa de Cohen entre deux observateurs dans le cas de variables binaires (situation la plus simple) et pour des variables qualitatives à plusieurs modalités. Le deuxième chapitre aborde la distribution d'échantillonnage du Kappa de Cohen (erreur type) et les différents tests d'hypothèses. Le plus classique consiste à tester l'hypothèse que le coefficient Kappa théorique est nul (κ=0, absence d'accord entre observateurs) ou encore égal à une valeur donnée ($\kappa = \kappa_0$). D'autres tests permettent de comparer plusieurs coefficients Kappa dans différentes populations.

Le troisième chapitre envisage l'extension du coefficient Kappa de Cohen de deux à plusieurs observateurs, en suivant une approche similaire à l'analyse de la variance à un critère. Il est également montré comment le coefficient Kappa dépend dans le cas d'un test binaire (et plus particulièrement d'un test diagnostic) de la prévalence de la maladie dans la population, de la spécificité et de la sensibilité du test.

Le quatrième chapitre rappelle les notions fondamentales des modèles linéaires généralisés (GLM), tels qu'introduits par Nelder et Wedderburn (1972). Il décrit aussi brièvement le modèle logistique classique et le modèle de régression logistique ordinale. Ceci permet d'introduire sans difficulté la méthode, récemment proposée par Shoukri et Mian (1996), d'estimer, par le principe du maximum de vraisemblance, le coefficient κ pour une table 2×2 lorsque les jugements dichotomiques dépendent de covariables relatives aux sujets et/ou aux examinateurs. Plusieurs exemples sont présentés.

Enfin, le cinquième chapitre décrit les travaux récents de Lipsitz et al (2001) qui permettent de modéliser le coefficient κ comme une fonction de covariables relatives aux observateurs et /ou aux sujets. Les auteurs proposent d'utiliser deux régressions logistiques et une régression linéaire pour les données binaires. Nous avons pu reproduire en SAS l'analyse des données fournies par les auteurs relatives à une enquête menée par Smith (1996) sur la similarité du niveau d'enseignement des membres d'un couple (époux et épouse).

Chapitre 1

Echantillonnage et tests d'hypothèses

1.1 Introduction

Dans ce chapitre, les erreurs types des coefficients introduits au Chapitre **??** seront estimées à l'aide de la méthode *Delta* multivariée. Cette méthode sera exposée de manière générale et dans le cas particulier d'une distribution multinomiale. Le lecteur intéressé par la méthode *Delta* peut se référer à Agresti (1990).

Une méthode basée sur la décomposition du Chi-carré et visant à effectuer des tests d'égalité de plusieurs coefficients d'association sera abordée. Cette méthode sera illustrée par son application au coefficient Kappa de Cohen.

1.2 Version multivariée de la méthode *Delta*

1.2.1 Cas général

Soient
$\boldsymbol{\theta}$ un vecteur de paramètres de population de dimension $T \times 1$: $\boldsymbol{\theta}=(\theta_1, \ldots, \theta_T)'$;
$\hat{\boldsymbol{\theta}}_n$ un vecteur de dimension $T \times 1$ d'estimateurs du vecteur de paramètres $\boldsymbol{\theta}$ pour un effectif n : $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{nT})'$.

Supposons que $\hat{\boldsymbol{\theta}}_n$ suive asymptotiquement une loi normale c'est-à-dire

$$\mathcal{L}[\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})] \longrightarrow \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \tag{1.1}$$

où le symbole $\mathcal{L}$ désigne la convergence en loi
et où $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ est la matrice de variances-covariances asymptotique $T \times T$ de $\hat{\boldsymbol{\theta}}_n$. Cette matrice est singulière si $\hat{\boldsymbol{\theta}}_n$ a une distribution incluse dans un sous-espace de l'espace T-dimensionnel.

Supposons que $\mathbf{f}$ soit une fonction définie sur un sous-ensemble ouvert d'un espace T-dimensionnel et à valeurs dans un espace R-dimensionnel

$$\boldsymbol{f} : \mathbb{R}^T \ \rightarrow \ \mathbb{R}^R \ : \ \boldsymbol{\theta} \mapsto \boldsymbol{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), \ldots, f_R(\boldsymbol{\theta}))'.$$

Supposons aussi que $\boldsymbol{f}$ soit au moins une fois différentiable en $\boldsymbol{\theta}$, c'est-à-dire

$$f_i(\boldsymbol{x}) = f_i(\boldsymbol{\theta}) + \sum_{j=1}^{T} (x_j - \theta_j) \frac{\partial f_i}{\partial x_j} \Bigg| \ \boldsymbol{x} = \boldsymbol{\theta} \ + o(||\boldsymbol{x} - \boldsymbol{\theta}||) \text{ si } \boldsymbol{x} \rightarrow \boldsymbol{\theta} \text{ pour } i = 1, \ldots, R.$$

Si $\left(\dfrac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \right)$ désigne la matrice $R \times T$ où l'élément (i, j) est la dérivée partielle de f_i par rapport à la jème coordonnée de $\boldsymbol{x} = (x_1, \ldots, x_T)'$ évaluée en $\boldsymbol{x} = \boldsymbol{\theta}$ c'est-à-dire

$$\left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \right)_{ij} = \frac{\partial f_i}{\partial x_j} \Bigg| \ \boldsymbol{x} = \boldsymbol{\theta},$$

alors,

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{\theta}) + \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \right) (\boldsymbol{x} - \boldsymbol{\theta}) + o(||\boldsymbol{x} - \boldsymbol{\theta}||) \text{ si } \boldsymbol{x} \rightarrow \boldsymbol{\theta}. \tag{1.2}$$

Bishop et Fienberg (1975) ont montré que

Théorème 1.2.1 . *Si $\hat{\boldsymbol{\theta}}_n$, $\boldsymbol{\theta}$ et $\boldsymbol{f}$ ont été définis comme ci-dessus et que (1.1) et (1.2) sont vérifiés, alors la distribution asymptotique de $\boldsymbol{f}(\hat{\boldsymbol{\theta}}_n)$ est donnée par*

$$\mathcal{L}[\sqrt{n}(\boldsymbol{f}(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{f}(\boldsymbol{\theta}))] \ \longrightarrow \ \mathcal{N}\left(0, \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \right) \Sigma(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \right)' \right) \tag{1.3}$$

1.2.2 Cas particulier : distribution multinomiale

Les calculs se simplifient considérablement dans le cas d'une distribution multinomiale.

Supposons que l'on dispose d'une table de contingence $k \times k$ et que le comptage dans les cellules

$$(n_{11}, \ldots, n_{1k}, \ldots, n_{k1}, \ldots, n_{kk})'$$

suive une distribution multinomiale avec des probabilités dans les cellules

$$\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1k}, \ldots, \pi_{k1}, \ldots, \pi_{kk})'.$$

Soient

$$n = n_{11} + \ldots + n_{1k} + \ldots + n_{k1} + \ldots + n_{kk}$$

et

$$\boldsymbol{p} = (p_{11}, \ldots, p_{1k}, \ldots, p_{k1}, \ldots, p_{kk})' \tag{1.4}$$

le vecteur des proportions d'échantillon où

$$p_{ij} = \frac{n_{ij}}{n}.$$

Soit la ième observation

$$\boldsymbol{y_i} = (y_{i11}, \ldots, y_{i1k}, \ldots, y_{ik1}, \ldots, y_{ikk})$$

où $y_{ijl} = 1$ si le ième objet est placé dans la catégorie j par l'observateur 1 et dans la catégorie l par l'observateur 2, $y_{ijl} = 0$ sinon.

Dès lors,

$$\sum_{j=1}^{k}\sum_{l=1}^{k} y_{ijl} = 1$$

et

$$y_{ijl}y_{imr} = 0 \text{ si } j \neq m \text{ ou si } l \neq r.$$

De plus, nous avons

$$p_{jl} = \frac{1}{n}\sum_{i=1}^{n} y_{ijl}\ . \tag{1.5}$$

Aussi,

$$E(y_{ijl}) = \pi_{jl} = E(y_{ijl}^2) \text{ et } E(y_{ijl}y_{imr}) = 0 \text{ si } j \neq m \text{ ou si } l \neq r.$$

Donc,

$$E(\boldsymbol{y_i}) = \boldsymbol{\pi} \text{ et } cov(\boldsymbol{y_i}) = \boldsymbol{\Sigma}, \quad i = 1, \ldots, n$$

$$\text{où } \boldsymbol{\Sigma} = (\sigma_{jl}) \text{ avec}$$

$$\sigma_{jj} = var(y_{ijl}) = \pi_{jl}(1 - \pi_{jl})$$

$$\sigma_{jl} = cov(y_{ijl}, y_{imr}) = -\pi_{jl}\pi_{mr} \text{ si } j \neq m \text{ ou si } l \neq r.$$

La matrice $\boldsymbol{\Sigma}$ a la forme

$$\boldsymbol{\Sigma} = diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' \tag{1.6}$$

où $diag(\boldsymbol{\pi})$ est la matrice diagonale des éléments de $\boldsymbol{\pi}$.

En vertu de (1.5), nous avons

$$cov(\boldsymbol{p}) = \frac{1}{n}(diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}').$$

Puisque

$$\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij} = 1,$$

cette matrice est singulière.

Le théorème central-limite multivarié (Rao, 1973) implique que

$$\mathcal{L}\left[\sqrt{n}\left[\boldsymbol{p} - \boldsymbol{\pi}\right]\right] \longrightarrow \mathcal{N}\left[0, diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\right].$$

Si $g(t_{11}, \cdots, t_{1k}, \cdots, t_{k1}, \cdots, t_{kk})$ est une fonction différentiable et

$$\phi_{ij} = \frac{\partial g}{\partial \pi_{ij}} \quad i, j = 1, \ldots, k$$

est $\dfrac{\partial g}{\partial \boldsymbol{t}}$ évalué en $\boldsymbol{t} = \boldsymbol{\pi}$, par la méthode *Delta*,

$$\mathcal{L}\left[\sqrt{n}\left[g(\boldsymbol{p}) - g(\boldsymbol{\pi})\right]\right] \longrightarrow \mathcal{N}\left(0, \boldsymbol{\phi}'\left[(diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}')\right]\boldsymbol{\phi}\right)$$

où $\phi' = (\phi_{11}, \cdots, \phi_{1k}, \cdots, \phi_{k1}, \cdots, \phi_{kk})$.

La matrice asymptotique de variances et covariances est égale à

$$\boldsymbol{\phi}' diag(\boldsymbol{\pi})\boldsymbol{\phi} - (\boldsymbol{\phi}'\boldsymbol{\pi})^2 = \sum_{i=1}^{k}\sum_{j=1}^{k}\pi_{ij}\phi_{ij}^2 - (\sum_{i=1}^{k}\sum_{j=1}^{k}\pi_{ij}\phi_{ij})^2.$$

1.3 Erreur type des coefficients d'association

1.3.1 Erreur type du coefficient ϕ^2

Après de longs et fastidieux calculs qui ne seront pas reproduits ici, Bishop et Fienberg (1975) ont montré que

$$\begin{aligned}(s.e.(\hat{\phi}^2))^2 &= \frac{1}{n}\left(4\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^3}{p_{i.}^2 p_{.j}^2} - 3\sum_{i=1}^{I}\frac{1}{p_{i.}}\left[\sum_{j=1}^{J}\frac{p_{ij}^2}{p_{i.}p_{.j}}\right]^2 - 3\sum_{j=1}^{J}\frac{1}{p_{.j}}\left[\sum_{i=1}^{I}\frac{p_{ij}^2}{p_{i.}p_{.j}}\right]^2\right. \\ &+ \left. 2\sum_{i=1}^{I}\sum_{j=1}^{J}\left[\frac{p_{ij}}{p_{i.}p_{.j}}\left(\sum_{l=1}^{I}\frac{p_{lj}^2}{p_{l.}p_{.j}}\right)\left(\sum_{m=1}^{J}\frac{p_{im}^2}{p_{i.}p_{.m}}\right)\right]\right) \qquad (1.7)\end{aligned}$$

et

$$(s.e.(\hat{V}))^2 = \frac{1}{2(min[(I-1),(J-1)])^{1/2}V}(s.e.(\hat{\phi}^2))^2. \qquad (1.8)$$

1.3.2 Erreur type du coefficient $\lambda_{C|R}$

Goodman et Kruskal (1963) ont montré que

$$(s.e.(\hat{\lambda}_{C|R}))^2 = \frac{(1-\sum_{i=1}^{I} p_{im})(\sum_{i=1}^{I} p_{im} + p_{.m} - 2\sum^{\star} p_{im})}{(1-p_{.m})^3} \tag{1.9}$$

où
$\sum^{\star} p_{im}$ est la somme des p_{im} qui apparaissent dans la colonne de $p_{.m}$.

Les calculs sont reproduits dans l'annexe A.

1.3.3 Erreur type du coefficient λ

Goodamn et Kruskal (1963) ont montré que

$$\begin{aligned}(s.e.(\hat{\lambda}))^2 &= \frac{1}{n(2-\Psi_{.})^4}\left[(2-\Psi_{.})^2\left[\Psi_{\Sigma}(1-\Psi_{\Sigma}) + 2\sum^{\star} p_{im}\right]\right. \\ &+ (2-\Psi_{\Sigma})^2\left[\Psi_{.}(1-\Psi_{.}) + 2p_{\star\star}\right] \\ &- \left. 2(2-\Psi_{.})(2-\Psi_{\Sigma})\left[\Psi_{\star} - \Psi_{.}\Psi_{\Sigma}\right]\right]\end{aligned} \tag{1.10}$$

où

$\Psi_{.} = p_{.m} + p_{m.}$;

$\Psi_{\Sigma} = \sum_{i=1}^{I} p_{im} + \sum_{j=1}^{J} p_{mj}$;

$\Psi_{\star} = \sum_{\star} p_{mj} + \sum^{\star} p_{im} + p_{m\star} + p_{\star m}$;

où

$p_{\star m}$ désigne les p_{im} qui sont dans la colonne pour laquelle $p_{i.}$ est maximal ;

$p_{m\star}$ désigne les p_{mj} qui sont dans la ligne pour laquelle $p_{.j}$ est maximal ;

$\sum_{\star} p_{mj}$ désigne la somme des p_{mj} qui apparaissent dans la ligne correspondant à $p_{m.}$;

$\sum^{\star} p_{im}$ est la somme des p_{im} qui apparaissent dans la colonne de $p_{.m}$.
Les calculs sont reproduits dans l'annexe A.

1.4 Erreur type des coefficients Kappa

1.4.1 Cas général : Kappa pondéré

Appliquons la méthode *Delta* à

$$g(\boldsymbol{p}) = \hat{\kappa}_w(\boldsymbol{p}) = \frac{\theta_1 - \theta_2}{1 - \theta_2}$$

où

$$\theta_1 = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{ij}$$

$$\theta_2 = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i.}p_{.j}$$

On a

$$\begin{aligned}
\frac{\partial\theta_1}{\partial p_{lm}} &= w_{lm} \\
\frac{\partial\theta_2}{\partial p_{lm}} &= \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{.j}\delta_{li} + \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i.}\delta_{mj} \\
&= \sum_{j=1}^{k} w_{lj}p_{.j} + \sum_{i=1}^{k} w_{im}p_{i.} \\
&= \overline{w}_{l.} + \overline{w}_{.m} \qquad (1.11)
\end{aligned}$$

où $\overline{w}_{.m} = \sum_{i=1}^{k} w_{im}p_{i.}$, $\overline{w}_{l.} = \sum_{i=1}^{k} w_{li}p_{.i}$ et le symbole δ désigne le symbole de Kroneker.

Ainsi,

$$\phi_{lm} = \frac{1}{(1-\theta_2)^2}\left[w_{lm}\left(1-\theta_2\right) - \left(\overline{w}_{l.} + \overline{w}_{.m}\right)\left(1-\theta_1\right)\right].$$

Donc,

$$\sum_{l=1}^{k}\sum_{m=1}^{k} p_{lm}\phi_{lm}^2 = \frac{1}{(1-\theta_2)^4}\left[\sum_{l=1}^{k}\sum_{m=1}^{k} p_{lm}\left[w_{lm}\left(1-\theta_2\right) - \left(\overline{w}_{l.} + \overline{w}_{.m}\right)\left(1-\theta_1\right)\right]^2\right].$$

Par de simples manipulations algébriques, on trouve que

$$\sum_{l=1}^{k}\sum_{m=1}^{k} p_{lm}\phi_{lm} = \theta_1 + \theta_1\theta_2 - 2\theta_2$$

Ainsi,

$$\begin{aligned}(s.e.(\hat{\kappa}_w))^2 &= \frac{1}{n\left(1-\theta_2\right)^4}\{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left[w_{ij}\left(1-\theta_2\right)-\left(\overline{w}_{i.}+\overline{w}_{.j}\right)\left(1-\theta_1\right)\right]^2 \\ &- \left(\theta_1\theta_2-2\theta_2+\theta_1\right)^2\} \end{aligned} \tag{1.12}$$

Supposons que l'on désire tester les hypothèses suivantes :

H_0 : les accords sont dus au hasard c'est-à-dire $p_{ij} = p_{i.}p_{.j}$ $(i,\ j = 1, \cdots, k)$. Il en résulte que $\kappa = 0$ et $\theta_1 = \theta_2$.

vs

H_1 : les accords ne sont pas dus au hasard c'est-à-dire $\exists i,\ j$ tq $p_{ij} \neq p_{i.}p_{.j}$ et $\kappa \neq 0$.

Sous l'hypothèse H_0, (1.12) se réduit à

$$(s.e.^{\star}(\hat{\kappa}_w))^2 = \frac{1}{n\left(1-\theta_2\right)^2}\{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{i.}p_{.j}\left[w_{ij}-\left(\overline{w}_{i.}+\overline{w}_{.j}\right)\right]^2-\theta_2^2\} \tag{1.13}$$

Cette expression s'obtient directement en remplaçant θ_1 par θ_2.

1.4.2 Cas particulier : Kappa de Cohen non pondéré

Dans ce cas,

$$g(\boldsymbol{p}) = \hat{\kappa}(\boldsymbol{p}) = \frac{\theta_1-\theta_2}{1-\theta_2}$$

où

$$\theta_1 = \sum_{i=1}^{k} p_{ii}$$

$$\theta_2 = \sum_{i=1}^{k} p_{i.}p_{.i}$$

On a alors

$$\frac{\partial\theta_1}{\partial p_{lm}} = \delta_{lm}$$

où le symbole δ désigne le symbole de Kroneker ;

$$\begin{aligned}
\frac{\partial \theta_2}{\partial p_{lm}} &= \sum_{i=1}^{k} \frac{\partial}{\partial p_{lm}} (\sum_{s=1}^{k} p_{is})(\sum_{t=1}^{k} p_{ti}) \\
&= \sum_{i=1}^{k} (\sum_{t=1}^{k} p_{ti} \sum_{s=1}^{k} \frac{\partial}{\partial p_{lm}} p_{is}) + \sum_{i=1}^{k} (\sum_{s=1}^{k} p_{is} \sum_{t=1}^{k} \frac{\partial}{\partial p_{lm}} p_{ti}) \\
&= \sum_{i=1}^{k} (\sum_{t=1}^{k} p_{ti}) \delta_{li} + \sum_{i=1}^{k} (\sum_{s=1}^{k} p_{is}) \delta_{mi} \\
&= p_{.l} + p_{m.} \qquad (1.14)
\end{aligned}$$

On calcule également

$$\begin{aligned}
\phi_{lm} &= \frac{1}{(1-\theta_2)^2} \left[[1-\theta_2] \frac{\partial(\theta_1 - \theta_2)}{\partial p_{lm}} + [\theta_1 - \theta_2] \frac{\partial \theta_2}{\partial p_{lm}} \right] \\
&= \frac{1}{(1-\theta_2)^2} \left[\delta_{lm} (1-\theta_2) + (\theta_1 - 1)(p_{.l} + p_{m.}) \right].
\end{aligned}$$

On obtient alors directement

$$\sum_{l=1}^{k} \sum_{m=1}^{k} p_{lm} \phi_{lm}^2 = \frac{1}{(1-\theta_2)^4} \left[\theta_1 (1-\theta_2)^2 + 2\theta_3 (1-\theta_2)(\theta_1 - 1) + \theta_4 (1-\theta_1)^2 \right]$$

où

$$\theta_3 = \sum_{l=1}^{k} p_{ll} (p_{l.} + p_{.l})$$

$$\theta_4 = \sum_{l=1}^{k} \sum_{m=1}^{k} p_{lm} (p_{.l} + p_{m.})^2.$$

On a aussi

$$\sum_{l=1}^{k} \sum_{m=1}^{k} p_{lm} \phi_{lm} = \left[\theta_1 (1-\theta_2) + (\theta_1 - 1) \sum_{l=1}^{k} \sum_{m=1}^{k} p_{lm} (p_{.l} + p_{m.}) \right].$$

Or, on obtient facilement par développement

$$\sum_{l=1}^{k} \sum_{m=1}^{k} p_{lm} (p_{.l} + p_{m.}) = 2\theta_2$$

Ainsi,

$$\left(\sum_{l=1}^{k} \sum_{m=1}^{k} p_{lm} \phi_{lm} \right)^2 = \frac{1}{(1-\theta_2)^4} \left[\theta_1^2 (1-\theta_2)^2 + 4\theta_1 (1-\theta_2)(\theta_1 - 1)\theta_2 + 4\theta_2^2 (1-\theta_1)^2 \right].$$

Finalement,

$$(s.e.(\hat{\kappa}))^2 = \frac{\theta_1(1-\theta_1)}{n(1-\theta_2)^2} + \frac{2(\theta_1-1)(\theta_3-2\theta_1\theta_2)}{n(1-\theta_2)^3} + \frac{(\theta_1-1)^2(\theta_4-4\theta_2^2)}{n(1-\theta_2)^4} \tag{1.15}$$

où

$$\theta_1 = \sum_{i=1}^{k} p_{ii}$$

$$\theta_2 = \sum_{i=1}^{k} p_{i.}p_{.i}$$

$$\theta_3 = \sum_{i=1}^{k} p_{ii}(p_{i.} + p_{.i})$$

$$\theta_4 = \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}(p_{.i} + p_{j.})^2.$$

Sous l'hypothèse H_0, $\theta_1 = \theta_2$ et $\theta_3 = \sum_{i=1}^{k} p_{i.}p_{.i}(p_{i.} + p_{.i})$

Le carré de l'erreur type de $\hat{\kappa}$ est alors

$$(s.e.^{\star}(\hat{\kappa}))^2 = \frac{1}{n(1-\theta_2)^2}\left[\theta_2(1-\theta_2) - 2\theta_3 + \theta_4\right]$$

En développant le terme $\theta_4 - 2\theta_3$, on obtient facilement

$$\theta_4 - 2\theta_3 = 2\theta_2^2 - \theta_3$$

Finalement,

$$(s.e.^{\star}(\hat{\kappa}))^2 = \frac{1}{n(1+\theta_2)^2}\left[\theta_2(1-\theta_2) - \theta_3\right] \tag{1.16}$$

où

$$\theta_2 = \sum_{i=1}^{k} p_{i.}p_{.i}$$

$$\theta_3 = \sum_{i=1}^{k} p_{i.}p_{.i}(p_{i.} + p_{.i}).$$

1.4.3 Kappa conditionnel

Bishop et Fienberg (1975) ont montré que

$$(s.e.(\hat{\kappa}_i))^2 = \frac{p_{i.} - p_{ii}}{np_{i.}^3\left(1-p_{.i}\right)^3}\left[\left(p_{i.} - p_{ii}\right)\left(p_{i.}p_{.i} - p_{ii}\right) + p_{ii}\left(1 - p_{i.} - p_{.i} + p_{ii}\right)\right] \tag{1.17}$$

Sous l'hypothèse H_0 de non association, la formule (1.17) se simplifie en

$$(s.e.^{\star}(\hat{\kappa}_i))^2 = \frac{p_{.i}\,(1-p_{i.})}{np_{i.}\,(1-p_{.i})} \tag{1.18}$$

1.5 Test d'hypothèse portant sur Kappa

Testons l'hypothèse suivante où κ_0 est une valeur de Kappa fixée a priori,

$$H_0 :\ \kappa = \kappa_0 \text{ vs } H_1 :\ \kappa \neq \kappa_0$$

La statistique

$$Z = \frac{\hat{\kappa} - \kappa_0}{s.e.(\hat{\kappa})} \tag{1.19}$$

suit asymptotiquement une loi normale $Z \sim N(0,1)$ où $(s.e.(\hat{\kappa}))^2$ est donnée par (1.15).

Remarquons que pour tester l'hypothèse $H_0 :\ \kappa = 0$ vs $H_1 :\ \kappa \neq 0$, l'expression de $(s.e.(\hat{\kappa}))^2$ est donnée par (1.16).

Dès lors, on rejette H_0 si la statistique Z observée (Z_{obs}) est telle que

$$\mid Z_{obs} \mid \geq Q_Z(1-\alpha/2)$$

où $Q_Z(1-\alpha/2)$ est le $(1-\alpha/2)$-quantile de la distribution gaussienne. Sinon on ne rejette pas H_0.

1.6 Comparaison de plusieurs kappas

En pratique, la valeur de κ peut dépendre d'une (ou plusieurs) covariable(s) nominale(s) à g modalités.

Par exemple : le sexe ($g = 2$), la classe d'âge ($g \geq 2$), la tranche de revenus ($g \geq 2$).

Supposons que l'on désire tester l'hypothèse

$H_0 \ :\ \kappa_1 = \cdots = \kappa_g$
vis-à-vis de l'alternative
$H_1 \ :\ \exists i \in (1, \cdots, g)$ et $j \in (1, \cdots, g)$ tels que $\kappa_i \neq \kappa_j$

La méthode exposée dans le paragraphe ci-dessous est inspirée par l'analyse de variance classique à un critère de classification. Cette méthode de décomposition d'un χ^2 est assez empirique.

L'étude de l'effet de covariables sur la valeur de κ sera affinée aux Chapitres 3 et 4.

1.6.1 Décomposition du Chi-carré en deux termes de κ

Fleiss (1981) propose la décomposition suivante.

Désignons par m_i la valeur de la mesure d'association choisie pour le ième groupe (c'est-à-dire la ième modalité d'une covariable).

Désignons par $s.e.(m_i)$ l'erreur type de m_i et définissons

$$w_i = \frac{1}{(s.e.(m_i))^2} \tag{1.20}$$

La quantité w_i représente le poids attaché à m_i.

En outre, supposons que $m_i = 0$ indique qu'il n'y ait pas d'association. Ainsi, sous l'hypothèse de non association (c'est-à-dire $m_i = 0$) dans le ième groupe, la statistique

$$\chi_i = \frac{m_i}{s.e.(m_i)} = m_i\sqrt{w_i} \tag{1.21}$$

suit approximativement une loi normale et la statistique

$$\chi_i^2 = w_i m_i^2 \tag{1.22}$$

suit approximativement une loi du Chi-carré à un degré de liberté pourvu que les effectifs n_i $(i = 1, \cdots, g)$ de chaque groupe d'individus soient assez "grands". Si l'hypothèse d'association prévaut dans le ième groupe, χ_i^2 prendra de "grandes" valeurs de façon à ce que l'hypothèse de non association soit rejetée si on effectue un test du Chi-carré.

Notre intérêt porte ici sur l'ensemble des groupes. Calculons, dès lors, la statistique suivante

$$\chi_{total}^2 = \sum_{i=1}^{g} \chi_i^2 \tag{1.23}$$

Si l'hypothèse nulle est vraie dans chacun des g groupes, χ_{total}^2 suit une loi Chi-carré à g degrés de liberté.

χ_{total}^2 peut être subdivisé en deux composantes

$$\chi_{total}^2 = \chi_{homog}^2 + \chi_{assoc}^2 \tag{1.24}$$

où

χ_{homog}^2 représente le degré d'homogénéité, ou d'égalité, entre les g mesures d'association ;
χ_{assoc}^2 représente un degré moyen d'association.

Le terme χ^2_{assoc} est calculé comme suit : une mesure totale d'association pour tous les groupes est définie. C'est une moyenne pondérée par les poids définis en (1.20)des g mesures individuelles.

$$\overline{m} = \frac{\sum_{i=1}^{g} w_i m_i}{\sum_{i=1}^{g} w_i}. \tag{1.25}$$

Sous l'hypothèse d'une mesure d'association globale égale à 0, $\overline{m}$ a une valeur moyenne de zéro et

$$s.e.(\overline{m}) = \frac{1}{\sqrt{\sum_{i=1}^{g} w_i}}.$$

Ainsi,

$$Z_{assoc} = \frac{\overline{m}}{s.e.(\overline{m})} = \frac{\sum_{i=1}^{g} w_i m_i}{\sqrt{\sum_{i=1}^{g} w_i}}$$

est distribué approximativement normalement sous l'hypothèse nulle. Fleiss (1981) considère que la statistique définie par

$$\chi^2_{assoc} = Z^2_{assoc} = \overline{m}^2 \sum_{i=1}^{g} w_i = \frac{\left(\sum_{i=1}^{g} w_i m_i\right)^2}{\sum_{i=1}^{g} w_i}$$

suit approximativement une loi du Chi-carré à un degré de liberté. Le terme χ^2_{homog} est obtenu par soustraction

$$\chi^2_{homog} = \chi^2_{total} - \chi^2_{assoc} = \sum_{i=1}^{g} w_i m_i^2 - \overline{m}^2 \sum_{i=1}^{g} w_i = \sum_{i=1}^{g} w_i (m_i - \overline{m})^2. \tag{1.26}$$

Ce terme suit approximativement une loi du Chi-carré à $g-1$ degrés de liberté sous l'hypothèse d'une association homogène et permettrait d'éprouver l'hypothèse H_0.

1.6.2 Exemple : application à κ

Considérons g estimations indépendantes de κ : $\hat{\kappa}_1, \cdots, \hat{\kappa}_g$.

L'estimation combinée des kappas (1.25) est

$$\hat{\kappa}_{assoc} = \frac{\sum_{m=1}^{g} w_m \hat{\kappa}_m}{\sum_{m=1}^{g} w_m} \tag{1.27}$$

Afin de tester l'hypothèse $H_0 : \kappa_1 = \cdots = \kappa_g$ vs $H_1 : \exists\, i \neq j : \kappa_i \neq \kappa_j$ $(i, j \in \{1, \cdots, g\})$, il suffit de comparer la valeur de

$$\chi^2_{obs} = \sum_{m=1}^{g} \frac{(\hat{\kappa}_m - \hat{\kappa}_{asso})^2}{(s.e.(\hat{\kappa}_m))^2} \tag{1.28}$$

aux tables du Chi-carré à $g-1$ degrés de liberté, l'hypothèse nulle étant rejetée au niveau d'incertitude α si la valeur de (1.28) est supérieure à $Q_{\chi^2}(1-\alpha; g-1)$, le $(1-\alpha)$-quantile de la distribution du χ^2 à $g-1$ degrés de liberté.

Remarquons que $(s.e.(\hat{\kappa}_m))^2$ est donné par l'expression (1.15).

1.7 Discussion

Le classement des valeurs de κ établi dans la table **??** du Chapitre **??** est arbitraire. Lorsqu'on souhaite apprécier une valeur calculée de κ, il est préférable de disposer de critères plus objectifs.

Dans le présent chapitre, nous avons utilisé la méthode *Delta* pour calculer les erreurs types des coefficients d'association et d'accord afin d'être apte à formuler et à construire des tests d'hypothèses.

Ces tests d'hypothèses sont de deux types. Ils portent, soit sur l'égalité d'un coefficient m à une valeur m_0 donnée a priori, soit sur l'égalité entre eux de plusieurs coefficients de même type.

Il est important de remarquer que les distributions des statistiques qui permettent d'effectuer les tests ne sont connues qu'asymptotiquement c'est-à-dire lorsque les effectifs tendent vers l'infini.

Pour un test du Chi-carré d'indépendance, il est connu depuis longtemps que tous les effectifs doivent être supérieurs à 5.

Pour les coefficients Kappa de Cohen et Kappa pondéré, les règles relatives aux effectifs de la table et aux totaux marginaux, qui permettraient de dire que l'approximation asymptotique est justifiée, sont détaillées dans les articles de Cichetti (1977) et Fleiss (1978). Ces détails ne seront pas repris dans notre travail.

Des études de simulation plus récentes existent comme, par exemple, l'étude de Donner (1998) ou Blackman et Koval (2000).

Chapitre 2

Extension du coefficient Kappa de Cohen

2.1 Introduction

Dans ce chapitre, la notion de coefficient Kappa de Cohen sera étendue au cas de plus de deux observateurs au moyen d'une méthode analogue à l'analyse de la variance (ANOVA) à un critère de classification pour les variables continues.

Un modèle de population sera proposé pour le coefficient Kappa de Cohen. Ce modèle repose sur l'adaptation du modèle de population du coefficient de corrélation intra-classe ρ au cas de variables nominales.

Nous verrons aussi, de manière analytique, que le coefficient Kappa de Cohen dépend de la spécificité et de la sensibilité des observateurs. Cette propriété sera utilisée dans les chapitres 3 et 4 afin d'obtenir une estimation du coefficient Kappa de Cohen en présence de covariables.

2.2 ANOVA à un critère de classification

Pour plus de clarté, les termes aléatoires seront soulignés dans le présent paragraphe.

Considérons un échantillon de n objets répartis en g groupes d'effectifs k_i ($i = 1, \cdots, g$).

Considérons le modèle usuel d'analyse de variance à un critère aléatoire de classification.

$$\underline{x}_{ij} = \mu + \underline{\alpha}_i + \underline{e}_{ij},\ i = 1, \cdots, g \text{ et } j = 1, \cdots, k_i$$

où
μ est l'effet moyen général,
$\underline{\alpha}_i$ est une variable aléatoire relative au ième groupe,

$\underline{e}_{ij}$ est l'erreur relative à l'observation (i, j).

Nous supposons que

1. $E(\underline{\alpha}_i) = 0$, $var(\underline{\alpha}_i) = \sigma_\alpha^2$ $(i = 1, \cdots, g)$ et que les $\underline{\alpha}_i$ sont mutuellement indépendantes ;
2. $E(\underline{e}_{ij}) = 0$, $var(\underline{e}_{ij}) = \sigma_e^2$ $(i = 1, \cdots, g\,;\, j = 1, \cdots, k_i)$ et les $\underline{e}_{ij}$ sont mutuellement indépendantes ;
3. Les variables aléatoires $\underline{\alpha}_i$ sont indépendantes des $\underline{e}_{ij}$.

Si on note $BSS = \sum_{i=1}^{g}\sum_{j=1}^{k_i}(\overline{\underline{x}}_{i.} - \overline{\underline{x}}_{..})^2$ et $WSS = \sum_{i=1}^{g}\sum_{j=1}^{k_i}(\underline{x}_{ij} - \overline{\underline{x}}_{i.})^2$ les sommes de carrés entre populations et des erreurs, respectivement, et $\sum_{i=1}^{g} k_i = n$, la table 2.1 résume l'analyse de la variance à un critère. Le calcul des espérances des sommes de carrés $(E(MS))$ et la définition de k_0 sont repris dans l'annexe B.

TABLE 2.1 – Table d'analyse de la variance à un critère

Source de variabilité	Somme des carrés	Degrés de liberté	Carrés moyens	E(MS)
Entre population	BSS	$g-1$	BMS	$\sigma_e^2 + k_0\sigma_\alpha^2$
Erreur	WSS	$n-g$	WMS	σ_e^2
Total	TSS	$n-1$		

On définit également le *coefficient de corrélation intra-classe* par

$$\rho = \frac{cov(\underline{x}_{ij}, \underline{x}_{il})}{\sqrt{var(\underline{x}_{ij})}\sqrt{var(\underline{x}_{il})}} = \frac{\sigma_\alpha^2}{\sigma_e^2 + \sigma_\alpha^2} \tag{2.1}$$

En termes de carrés moyens, l'estimation de (2.1), r, est donnée par

$$r = \frac{BMS - WMS}{BMS + (k_0 - 1)WMS}$$

Puisque ρ est une fonction de σ_α^2, on teste l'hypothèse suivante

$$H_0 : \sigma_\alpha^2 = 0 \text{ vs } \; H_1 : \sigma_\alpha^2 \neq 0$$

$$\text{ou } \; H_0 : \rho = 0 \text{ vs } \; H_1 : \rho \neq 0$$

Si on ajoute aux hypothèses existantes une hypothèse de normalité, à savoir les variables aléatoires $\underline{\alpha}_i$ et $\underline{e}_{ij}$ sont normales, alors, sous l'hypothèse nulle, $F_{obs} = \dfrac{BMS}{WMS}$ est distribué comme un F de Snedecor à $g-1$ et $n-g$ degrés de liberté. H_0 est rejetée si

$$F_{obs} \geq Q_F(1-\alpha; g-1, n-g)$$

sinon, H_0 n'est pas rejetée.

2.3 Plusieurs observateurs et critère binaire

Soient m observateurs qui, de manière indépendante, sont appelés à classer n sujets dans deux catégories et m_i le nombre d'observations par sujet. Nous supposerons que les différents sujets ne sont pas toujours classés par tous les observateurs. Ainsi, les nombres m_i ne sont pas nécessairement identiques pour chaque sujet. Une des deux catégories sera appelée la catégorie positive et l'autre la négative. Désignons par n_i le nombre d'observations positives sur le sujet i, $(i = 1, \cdots, n)$ et $(m_i - n_i)$ le nombre de classements négatifs sur celui-ci.

Fleiss (1981) propose d'utiliser une analyse de variance à un critère sur ce type de données particulières afin d'obtenir un coefficient κ en codant les observations positives par 1 et les observations négatives par 0.

Nous remarquerons qu'un groupe d'observations est formé par les m_i observations sur le même sujet $(i = 1, \cdots, n)$.

Dans ce cas particulier, on a

$$\overline{x}_{..} = \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} m_i} \quad \text{et} \quad \overline{x}_{i.} = \frac{n_i}{m_i} \quad i = 1, \cdots, n.$$

Les sommes de carrés deviennent

$$\begin{aligned} BSS &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}(\overline{x}_{i.} - \overline{x}_{..})^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}(\frac{n_i}{m_i} - \overline{x}_{..})^2 \\ &= \sum_{i=1}^{n}\frac{(n_i - m_i\overline{x}_{..})^2}{m_i} \end{aligned} \tag{2.2}$$

avec $n-1$ degrés de liberté,

et

$$\begin{aligned} WSS &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}(x_{ij}-\overline{x}_{i.})^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}(x_{ij}^2+\frac{n_i^2}{m_i^2}-2x_{ij}\frac{n_i}{m_i}) \\ &= \sum_{i=1}^{n}(n_i-\frac{n_i^2}{m_i}) \\ &= \sum_{i=1}^{n}\frac{n_i(m_i-n_i)}{m_i} \end{aligned} \tag{2.3}$$

avec $\sum_{i=1}^{n} m_i - n$ degrés de liberté, ou encore si on pose $\overline{m} = \sum_{i=1}^{n}\frac{m_i}{n}$, le nombre de degrés de liberté de WSS est $n(\overline{m}-1)$.

Pour simplifier l'expression analytique du coefficient de corrélation intra-classe, Fleiss propose de définir le carré moyen BMS par

$$BMS = \frac{1}{n}\sum_{i=1}^{n}\frac{(n_i - m_i\overline{x}_{..})^2}{m_i} \tag{2.4}$$

et par analogie avec le cas des variables continues, on définit

$$r = \frac{BMS - WMS}{BMS + (m_0-1)WMS} \tag{2.5}$$

où

$$m_0 = \overline{m} - \frac{\sum_{i=1}^{n}(m_i-\overline{m})^2}{n(n-1)\overline{m}}$$

Si n est "grand", les nombres m_0 et $\overline{m}$ sont peu différents ; ils sont égaux si le nombre d'observateurs est constant pour toutes les observations. Si m_0 est remplacé par $\overline{m}$ dans (2.5), le coefficient Kappa est estimé par

$$\hat{\kappa} = r = \frac{BMS - WMS}{BMS + (\overline{m}-1)WMS} \tag{2.6}$$

Il est important de remarquer qu'il n'est pas incorrect de calculer des sommes de carrés et des coefficients de corrélation intra-classe pour des données indépendantes et homoscédastiques qui ne sont pas normales. Une erreur méthodologique consisterait à calculer des statistiques analogues au F de Snedecor et à effectuer des tests au moyen de cette statistique.

Après quelques manipulations algébriques élémentaires, la formule (2.6) devient

$$\hat{\kappa} = 1 - \frac{\sum_{i=1}^{n} \frac{n_i(m_i - n_i)}{m_i}}{n(\overline{m} - 1)\overline{x}_{..}\overline{y}_{..}}$$

où $\overline{y}_{..} = 1 - \overline{x}_{..}$

ou encore

$$\hat{\kappa} = 1 - \frac{\sum_{i=1}^{n} \overline{x}_{i.}\overline{y}_{i.}m_i}{n(\overline{m} - 1)\overline{x}_{..}\overline{y}_{..}}$$

où
$\overline{x}_{i.} = \dfrac{n_i}{m_i}$ et $\overline{y}_{i.} = \dfrac{(m_i - n_i)}{m_i}$, $i = 1, \cdots, n$.

$\hat{\kappa}$ possède les propriétés suivantes :

1. Si $\forall\, i,\ \overline{x}_{i.} = \overline{x}_{..}$, avec $\overline{x}_{..} \neq 0$ et $\overline{x}_{..} \neq 1$, alors, il existe plus de désaccords intra-sujets qu'entre sujets. Dans ce cas, $\hat{\kappa}$ prend sa valeur minimale, à savoir $-1/(\overline{m} - 1)$.
2. Si chaque proportion $\overline{x}_{i.}$ est égale à 0 ou est égale à 1, alors, l'accord sur les sujets est parfait et $\hat{\kappa} = 1$.

2.4 Coefficient kappa et paramètres de population

En général, le coefficient κ n'est pas défini en termes de paramètres de population mais en termes de procédures de calcul conduisant à son élaboration.

Kraemer (1979) propose de définir κ en termes de paramètres de population, de manière analogue au coefficient de corrélation intra-classe, ρ, mais adaptée au cas de variables nominales.

2.4.1 Modèle de population

Soit $\underline{\boldsymbol{x}}_i = (\underline{x}_{i1}, \cdots, \underline{x}_{ik})$ un vecteur de variables aléatoires binaires attaché au ième sujet. $\underline{x}_{ij} = 1$ si le ième sujet est classé dans la cellule j, $\underline{x}_{ij} = 0$ sinon, $j = 1, 2, \cdots, k$.
On a $\displaystyle\sum_{j=1}^{k} \underline{x}_{ij} = 1$.

Soient

$E(\underline{x}_{ij}) = P_j,\ (P'_j = 1 - P_j),$

$var(\underline{x}_{ij}) = \sigma_j^2,\ j = 1, \cdots, k.$

Le coefficient Kappa de Cohen pour la jème catégorie est défini par

$$\kappa_j = \frac{\sigma_j^2}{P_j P'_j} \tag{2.7}$$

et le coefficient κ global pour les k catégories est donné par

$$\kappa = \frac{\sum_{j=1}^k \sigma_j^2}{\sum_{j=1}^k P_j P'_j}. \tag{2.8}$$

Ces définitions sont en accord avec la définition usuelle de κ. En effet, la probabilité que deux observations indépendantes soient en accord pour le sujet i est

$p_o = \sum_{j=1}^k E(\underline{x}_{ij}^2) = \sum_{j=1}^k (\sigma_j^2 + P_j^2)$

et l'accord dû au hasard est

$p_e = \sum_{j=1}^k (E(\underline{x}_{ij})E(\underline{x}_{ij})) = \sum_{j=1}^k P_j^2$

Par conséquent,

$$\kappa = \frac{\sum_{j=1}^k \sigma_j^2}{\sum_{j=1}^k P_j P'_j} = \frac{\sum_{j=1}^k P_j P'_j \kappa_j}{\sum_{j=1}^k P_j P'_j} \tag{2.9}$$

2.4.2 Effet de la prévalence, de la sensibilité et de la spécificité

Soit M une maladie dont la prévalence est notée π.

$$\pi = \mathbb{P}[M]$$

Dans le cadre de nombreuses applications médicales, on dispose d'un test T tel qu'un sujet donné sera déclaré malade si le test est positif (T_+) et "non malade" si le test est négatif (T_-). On se trouve donc dans le cas d'une dichotomie.

On définit la sensibilité du test par la probabilité qu'un sujet donné soit déclaré positif s'il est malade. On a

$$S_e = \mathbb{P}[T_+|M] = E_{|M}\underline{x}_{i1}.$$

De même la spécificité est définie par la probabilité qu'un individu donné soit déclaré négatif s'il n'est pas malade.

$$S_p = \mathbb{P}[T_-|\overline{M}] = E_{|\overline{M}}\underline{x}_{i2}$$

avec $\underline{x}_{i1} + \underline{x}_{i2} = 1 \quad \forall i$.

On a

$$\mathbb{P}[T_+] = \mathbb{P}[T_+|M]\mathbb{P}[M] + \mathbb{P}[T_+|\overline{M}]\mathbb{P}[\overline{M}]. \tag{2.10}$$

Si on note $\pi' = 1 - \pi, \quad S'_e = 1 - S_e, \quad S'_p = 1 - S_p$, il vient

$$P = \mathbb{P}[T_+] = \pi S_e + \pi' S'_p.$$

P est l'espérance de $\underline{x}_{i1}$. En effet,

$$\begin{aligned} P &= E(\underline{x}_{i1}) = E_{|M}\underline{x}_{i1}\mathbb{P}[M] + E_{|\overline{M}}\underline{x}_{i1}\mathbb{P}[\overline{M}] \\ &= S_e\pi + E_{|\overline{M}}(1 - \underline{x}_{i2})\mathbb{P}[\overline{M}] = S_e\pi + S'_p\pi'. \end{aligned} \tag{2.11}$$

P est la probabilité d'observer un test positif et est donc un estimateur biaisé de la prévalence π.

Calculons la variance de $\underline{x}_{i1}$.

On a

$$var(\underline{x}_{i1}) = E(\underline{x}_{i1}^2) - (E(\underline{x}_{i1}))^2.$$

Or,

$$\begin{aligned} E(\underline{x}_{i1}^2) &= (E_{|M}\underline{x}_{i1})^2\mathbb{P}[M] + (E_{|\overline{M}}\underline{x}_{i1})^2\mathbb{P}[\overline{M}] \\ &= S_e^2\pi + S_p'^2(1 - \pi). \end{aligned} \tag{2.12}$$

Dès lors, il vient, après quelques manipulations algébriques élémentaires,

$$var(\underline{x}_{i1}) = \pi(1 - \pi)(S_e + S_p - 1)^2. \tag{2.13}$$

En vertu de (2.9), on a

$$\kappa = \frac{\pi(1 - \pi)(S_e + S_p - 1)^2}{P(1 - P)}. \tag{2.14}$$

Il est évident que $\kappa = 1$ si et seulement si $S_e = 1$ et $S_p = 1$, c'est-à-dire que le test T ne produit aucune erreur de malclassification. Si $\pi = 0$ ou $\pi = 1$, $\kappa = 0$.

En dehors de ces valeurs extrêmes, κ présente un maximum si $\pi = \dfrac{\sigma_{S_p}}{\sigma_{S_e} + \sigma_{S_p}}$ où $\sigma^2_{S_p} = S_p(1 - S_p)$ et $\sigma^2_{S_e} = S_e(1 - S_e)$.

La démonstration de cette propriété est longue et fastidieuse mais elle ne fait appel qu'à des manipulations algébriques élémentaires. Elles ne seront reproduites que partiellement ici.

Nous devons résoudre l'équation

$$\frac{\partial \kappa}{\partial \pi} = \frac{1}{P^2(1-P)^2}\left\{(P-P^2)[1-2\pi] - \pi(1-\pi)(1-2P)(S_e+S_p-1)\right\} = 0 \qquad (2.15)$$

par rapport à π afin de trouver l'extremum de κ.

Ceci revient à devoir résoudre l'équation du second degré

$$(S_e+S_p-1)(S_e-S_p)\pi^2 - 2S_p(1-S_p)\pi + S_p(1-S_p) = 0. \qquad (2.16)$$

Les solutions sont

$$\pi = \frac{\sigma_{S_p}(\sigma_{S_p} \pm \sigma_{S_e})}{(\sigma_{S_p}+\sigma_{S_e})(\sigma_{S_p}-\sigma_{S_e})}. \qquad (2.17)$$

Donc,

$$\pi_1 = \frac{\sigma_{S_p}}{\sigma_{S_p}+\sigma_{S_e}} \quad \text{et} \quad \pi_2 = \frac{\sigma_{S_p}}{\sigma_{S_p}-\sigma_{S_e}}.$$

Une étude de signe classique conduit à la conclusion suivante : le maximum de (2.16) est atteint dans les cas $\sigma_{S_p} < \sigma_{S_e}$ et $\sigma_{S_p} > \sigma_{S_e}$ en π_1.

En effectuant les remplacements adéquats, on obtient après des calculs longs et fastidieux mais élémentaires

$$\kappa_{max} = [(S_eS_p)^{1/2} - (S'_eS'_p)^{1/2}]^2.$$

2.5 Plusieurs observateurs et critère qualitatif non binaire

Supposons maintenant que le nombre k de catégories soit supérieur ou égal à deux. Landis et Koch (1977b) proposent de prendre la moyenne pondérée

$$\hat{\overline{\kappa}} = \frac{\sum_{j=1}^{k} \overline{x}_{.j}\overline{y}_{.j}\hat{\kappa}_j}{\sum_{j=1}^{k} \overline{x}_{.j}\overline{y}_{.j}} \qquad (2.18)$$

comme mesure globale d'accord entre observateurs où $\overline{x}_{.j} = \dfrac{\sum_{i=1}^{n} x_{ij}}{nm}$ représente la proportion globale d'observations dans la catégorie j et $\overline{y}_{.j} = 1 - \overline{x}_{.j}$.

Il est intéressant de remarquer l'analogie entre les formules (2.9) et (2.18).

2.6 Nombre d'observations constant par sujet

Dans le cas où le nombre d'observations par sujet est constant, les expressions de $\hat{\kappa}_j$ et $\hat{\overline{\kappa}}$ se simplifient.

En effet, notons n_{ij} le nombre d'observations sur le sujet i pour la catégorie j.

$$\forall\ i,\ \sum_{j=1}^{k} n_{ij} = m$$

Dès lors,

$$\hat{\kappa}_j = 1 - \frac{\sum_{i=1}^{n} n_{ij}(m - n_{ij})}{nm(m-1)\overline{x}_{.j}\overline{y}_{.j}} \tag{2.19}$$

et

$$\hat{\overline{\kappa}} = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} n_{ij}^2}{nm(m-1)\sum_{j=1}^{k}\overline{x}_{.j}\overline{y}_{.j}} \tag{2.20}$$

2.7 Erreur type de κ

Fleiss et Cuzick (1979) ont montré que sous l'hypothèse nulle
$H_0 :\ \kappa = 0$,

$$s.e._0(\hat{\kappa}) = \frac{1}{(\overline{m}-1)\sqrt{n\overline{m}_H}}\sqrt{2(\overline{m}_H - 1) + \frac{(\overline{m} - \overline{m}_H)(1 - 4\overline{x}_{..}\overline{y}_{..})}{\overline{m}\overline{x}_{..}\overline{y}_{..}}} \tag{2.21}$$

où

$$\overline{m}_H = \frac{n}{\sum_{i=1}^{n}\frac{1}{m_i}} \tag{2.22}$$

est la moyenne harmonique du nombre d'observations par sujet.

Dans le cas où le nombre d'observations par sujet est constant, Fleiss, Nee et Landis (1979) ont montré que

$$s.e._0(\hat{\overline{\kappa}}) = \frac{\sqrt{2}}{\sum_{j=1}^{k}\overline{x}_{.j}\overline{y}_{.j}\sqrt{nm(m-1)}}\sqrt{\left(\sum_{j=1}^{k}\overline{x}_{.j}\overline{y}_{.j}\right)^2 - \sum_{j=1}^{k}\overline{x}_{.j}\overline{y}_{.j}(\overline{y}_{.j} - \overline{x}_{.j})} \tag{2.23}$$

et

$$s.e._0(\hat{\kappa}_j) = \sqrt{\frac{2}{nm(m-1)}} \tag{2.24}$$

Remarquons que $s.e._0(\hat{\kappa}_j)$ est indépendant de $\overline{x}_{.j}$ et $\overline{y}_{.j}$.

2.8 Exemple

Fleiss (1981) propose l'exemple artificiel suivant portant sur $n = 10$ sujets. Cinq observations sont effectuées sur chaque sujet ($m = 5$). Le nombre de catégories possibles de classement est $k = 3$. Les données sont reprises dans la table 2.2.

TABLE 2.2 – Exemple : 5 observations sur chacun des 10 sujets dans une des 3 catégories

	Nombre d'observations dans la catégorie			
Sujet	1	2	3	$\sum_{i=1}^{3} x_{ij}^2$
1	1	4	0	17
2	2	0	3	13
3	0	0	5	25
4	4	0	1	17
5	3	0	2	13
6	1	4	0	17
7	5	0	0	25
8	0	4	1	17
9	1	0	4	17
10	3	0	2	13
total	20	12	18	174

Les 3 proportions globales d'observations dans chaque catégorie sont

$$\overline{x}_{.1} = 20/50 = 0.40 \quad , \quad \overline{x}_{.2} = 12/50 = 0.24 \quad \text{et} \quad \overline{x}_{.3} = 18/50 = 0.26$$

Ainsi,

$$\hat{\kappa}_1 = 1 - \frac{\sum_{i=1}^{10} x_{i1}(5 - x_{i1})}{10 \times 5 \times 4 \times 0.40 \times 0.60} = 1 - \frac{34}{48} = 0.29$$

$$\hat{\kappa}_2 = 1 - \frac{\sum_{i=1}^{10} x_{i2}(5 - x_{i2})}{10 \times 5 \times 4 \times 0.24 \times 0.76} = 1 - \frac{12}{36.48} = 0.67$$

$$\hat{\kappa}_3 = 1 - \frac{\sum_{i=1}^{10} x_{i3}(5 - x_{i3})}{10 \times 5 \times 4 \times 0.36 \times 0.64} = 1 - \frac{30}{46.08} = 0.35$$

En vertu de (2.20),

$$\hat{\overline{\kappa}} = 1 - \frac{10 \times 25 - 174}{10 \times 5 \times 4 \times (0.40 \times 0.60 + 0.24 \times 0.76 + 0.36 \times 0.64)} = 0.42$$

Estimons l'erreur type de $\hat{\overline{\kappa}}$ et de $\hat{\kappa}_i$ $(i = 1, \cdots, 3)$ à l'aide des formules (2.23) et (2.24) respectivement.

$$\sum_{i=1}^{3} \overline{x}_{.j}\overline{y}_{.j} = 0.40 \times 0.60 + 0.24 \times 0.76 + 0.36 \times 0.64 = 0.6528$$

et

$$\sum_{i=1}^{3} \overline{x}_{.j}\overline{y}_{.j}(\overline{y}_{.j} - \overline{x}_{.j}) = 0.40 \times 0.60 \times 0.20 + 0.24 \times 0.76 \times 0.52 + 0.36 \times 0.64 \times 0.28 = 0.2074$$

Dès lors, en vertu de (2.23), on a

$$s.e._0(\hat{\overline{\kappa}}) = \frac{\sqrt{2}}{0.6528\sqrt{10 \times 5 \times 4}}\sqrt{0.6528^2 - 0.2078} = 0.072$$

Puisque

$$z = \frac{\hat{\overline{\kappa}}}{s.e._0(\hat{\overline{\kappa}})} = \frac{0.42}{0.072} = 5.83$$

la valeur globale de κ est significativement différente de 0 ($p < 0.001$), mais, selon le tableau **??**, l'accord global est modéré.

En vertu de (2.24), on a

$$s.e._0(\hat{\kappa}_i) = \sqrt{\frac{2}{10 \times 5 \times 4}} = 0.10$$

En conclusion, chaque κ_i $(i = 1, \cdots, 3)$ est significativement différent de 0 ($p < 0.001$). Mais, selon le tableau **??**, $\hat{\kappa}_1$ et $\hat{\kappa}_3$ représentent un accord médiocre et $\hat{\kappa}_2$, un accord modéré.

2.9 Discussion

Dans ce chapitre, nous avons étudié deux aspects intéressants des propriétés des coefficients de Cohen.

Le premier aspect est la définition de κ en fonction de paramètres de population. Nous avons montré au paragraphe 2.4.2 que le coefficient κ de Cohen dépend de la prévalence, de la spécificité et de la sensibilité. De nombreuses méthodes pour estimer ces variables existent dont la méthode de Hui et Walter (1980).

Nous approfondirons ce problème dans les chapitres suivants et nous montrerons comment on peut estimer le coefficient Kappa de Cohen à l'aide de paramètres relatifs à des covariables liées aux observateurs et/ou aux sujets.

Le deuxième aspect que nous avons étudié est relatif au nombre d'observateurs. Lorsque ce dernier est supérieur à deux, les méthodes que nous avons trouvées dans la littérature sont peu nombreuses, peu originales et souvent directement inspirées par des modèles applicables aux données continues, comme par exemple au paragraphe 2.2.

Bishop et Fienberg (1975) ont cité dans leur bibliographie des modèles qui semblaient prometteurs et qui faisaient appel aux modèles log-linéaires. Malheureusement, tous les articles cités ont été publiés dans des rapports techniques ou des revues locales pratiquement inaccessibles. Nous n'avons pas trouvé trace de ces développements dans la littérature.

Chapitre 3

Modèle linéaire généralisé

3.1 Introduction

Nous définirons dans ce chapitre la notion de famille exponentielle généralisée. Les deux premiers moments d'une variable aléatoire appartenant à une telle famille seront déterminés. Nous introduirons aussi les notions de modèle linéaire généralisé pour données indépendantes et de régression logistique pour variables dichotomiques et ordinales. Nous appliquerons ces modèles pour estimer le coefficient κ.

Le modèle linéaire généralisé a été introduit par Nelder et Wedderburn (1972). Ce modèle est une généralisation à la famille exponentielle généralisée du modèle linéaire défini pour les populations normales.

Nous déterminerons d'abord les équations du maximum de vraisemblance pour le modèle linéaire généralisé et ensuite nous exposerons la méthode itérative des scores de Fisher qui permet d'estimer les paramètres de ce modèle.

Nous exposerons brièvement l'extension du modèle linéaire généralisé à des données appariées.

3.2 Famille exponentielle généralisée

3.2.1 Définition

Soit une variable aléatoire Y. Supposons que $(y_1, \cdots, y_n)$ désigne les valeurs de n observations indépendantes de la variable aléatoire Y.

Y est une variable aléatoire exponentielle généralisée si la densité de chacun des y_i, $(i = 1, \cdots, n)$ par rapport à une mesure λ de Lebegue ou de comptage peut s'écrire sous la forme

$$f(y_i; \theta_i, \phi) = exp\left[(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)\right] \tag{3.1}$$

Le paramètre θ_i est appelé *paramètre canonique*, le paramètre ϕ est appelé *paramètre de dispersion*. On suppose que les fonctions $a(\phi)$ et $b(\theta_i)$ sont au moins deux fois continûment différentiables dans l'espace des paramètres.

3.2.2 Expression des deux premiers moments de Y

Soit la contribution de la ième observation au logarithme de la vraisemblance

$$l(\theta_i, \phi; y_i) = \ln(f(y_i; \theta_i, \phi)) = [(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)]. \tag{3.2}$$

Nous avons

$$\partial l/\partial\theta_i = [y_i - b'(\theta_i)]\,/a(\phi) \tag{3.3}$$

$$\partial^2 l/\partial\theta_i^2 = -b''(\theta_i)/a(\phi) \tag{3.4}$$

où $b'(\theta_i)$ et $b''(\theta_i)$ désignent respectivement les deux premières dérivées de b évaluées en θ_i.

On montre aisément, en toute généralité, que si les conditions de permutation des opérateurs d'intégration et de dérivation sont satisfaites,

$$E\left(\frac{\partial l}{\partial\theta_i}\right) = 0 \tag{3.5}$$

et

$$-E\left(\frac{\partial^2 l}{\partial\theta_i^2}\right) = E\left(\frac{\partial l}{\partial\theta_i}\right)^2. \tag{3.6}$$

Ainsi, en vertu de (3.3) et (3.5), si μ_i désigne $E(Y_i)$,

$$\mu_i = b'(\theta_i) \tag{3.7}$$

et en vertu de (3.4) et (3.6),

$$E\left[(y_i - b'(\theta_i))^2\,/a^2(\phi)\right] = var(Y_i)/a^2(\phi) = b''(\theta_i)/a(\phi).$$

Donc,

$$var(Y_i) = b''(\theta_i)a(\phi). \tag{3.8}$$

3.3 Composante systématique et fonction de lien

Soient $x_{i1}, \cdots, x_{it}$ les valeurs de t variables explicatives (appelées également covariables) relatives à la ième observation. La *composante systématique* relie les paramètres inconnus aux variables explicatives en utilisant un prédicteur linéaire

$$\eta_i = \sum_{j=1}^{t} \beta_j x_{ij},\ i = 1, \cdots, n$$

ou encore, sous forme matricielle,

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} \tag{3.9}$$

où $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)'$ est un vecteur de prédicteurs linéaires, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_t)'$ est un vecteur de paramètres inconnus et $\boldsymbol{X}$ est une matrice $n \times t$ de variables explicatives.

On définit également

$$\eta_i = g(\mu_i) \tag{3.10}$$

où g est une fonction strictement monotone et au moins deux fois continûment différentiable par rapport à $\mu_1, \cdots, \mu_n$. g est appelée *fonction de lien*.

La fonction g pour laquelle $g(\mu_i) = \theta_i$ est appelée *lien canonique*.

3.4 Estimation des paramètres du modèle linéaire généralisé

Pour n observations indépendantes, le logarithme de la vraisemblance est

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln(f(y_i; \theta_i, \phi)) = \sum_{i=1}^{n} l_i \tag{3.11}$$

où $l_i = l(\theta_i, \phi, y_i),\ i = 1, \cdots, n$.

Les équations du maximum de vraisemblance s'obtiennent en calculant

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{3.12}$$

On a $\dfrac{\partial l_i}{\partial \theta_i} = [y_i - b'(\theta_i)]/a(\phi)$, $\mu_i = b'(\theta_i)$ et $var(Y_i) = b''(\theta_i)a(\phi)$,
donc,

$$\frac{\partial l_i}{\partial \theta_i} = (y_i - \mu_i)/a(\phi)$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{var(Y_i)}{a(\phi)}$$

De plus,

$$\eta_i = \sum_{j=1}^{t} \beta_j x_{ij} \quad \text{donc} \quad \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

Dès lors,

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)}\frac{a(\phi)}{var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} \tag{3.13}$$

Finalement, les équations du maximum de vraisemblance s'écrivent

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{var(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} = 0, \ j = 1, \cdots, t \tag{3.14}$$

ou, sous forme matricielle,

$$\boldsymbol{X}'\boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0} \tag{3.15}$$

où $\boldsymbol{\Delta}$ est la matrice diagonale composée des éléments $(\frac{\partial \mu_i}{\partial \eta_i} \frac{1}{var(Y_i)})$.

De telles équations n'ont pas de solution analytique. La méthode itérative utilisée pour ajuster le modèle linéaire généralisé est la méthode des scores de Fisher décrite au paragraphe suivant.

Le taux de convergence de $\hat{\boldsymbol{\beta}}$ vers $\boldsymbol{\beta}$ dépend de la matrice d'information.

On a

$$\begin{aligned} -E\left(\frac{\partial^2 l_i}{\partial \beta_h \partial \beta_j}\right) &= E\left[\left(\frac{\partial l_i}{\partial \beta_h}\right)\left(\frac{\partial l_i}{\partial \beta_j}\right)\right] \\ &= E\left[\frac{(Y_i - \mu_i)x_{ih}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \frac{(Y_i - \mu_i)x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}\right] \\ &= \frac{x_{ih}x_{ij}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \end{aligned} \tag{3.16}$$

Donc,

$$-E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_h \partial \beta_j}\right) = \sum_{i=1}^{n} \frac{x_{ih}x_{ij}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{3.17}$$

Sous forme matricielle,

$$\boldsymbol{H} = \boldsymbol{X'WX} \tag{3.18}$$

où

$\boldsymbol{H} = E\left(-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_h \partial \beta_j}\right)$ est appelée *matrice d'information* ;

$\boldsymbol{W}$ est la matrice diagonale avec les éléments $w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 / var(Y_i)$ sur la diagonale principale.

3.5 Méthode des scores de Fisher

Désignons par $\boldsymbol{\beta}^{(k)}$ la kème approximation de l'estimation $\hat{\boldsymbol{\beta}}$ obtenue par la méthode de Newton-Raphson où la matrice des dérivées secondes est remplacée par son espérance $-\boldsymbol{H}$.

On a

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + (\boldsymbol{H}^{(k)})^{-1}\boldsymbol{q}^{(k)} \tag{3.19}$$

où
$\boldsymbol{H}$ est la matrice définie en (3.18) et supposée non singulière ;
$\boldsymbol{q}$ est le vecteur composé des éléments $\dfrac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}$;
$\boldsymbol{H}^{(k)}$ et $\boldsymbol{q}^{(k)}$ sont $\boldsymbol{H}$ et $\boldsymbol{q}$ évalués en $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}$.

En vertu des formules (3.14) et (3.17) il vient

$$\boldsymbol{H}^{(k)}\boldsymbol{\beta}^{(k)} + \boldsymbol{q}^{(k)} = \sum_{j=1}^{t}\left(\sum_{i=1}^{n}\frac{x_{ih}x_{ij}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \beta_j^{(k)}\right) + \sum_{i=1}^{n}\frac{(y_i - \mu_i^{(k)})}{var(Y_i)}x_{ik}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

où μ_i et $\left(\dfrac{\partial \mu_i}{\partial \eta_i}\right)$ sont évalués en $\boldsymbol{\beta}^{(k)}$.

Donc,

$$\boldsymbol{H}^{(k)}\boldsymbol{\beta}^{(k)} + \boldsymbol{q}^{(k)} = \boldsymbol{X'W}^{(k)}\boldsymbol{Z}^{(k)} \tag{3.20}$$

où
$\boldsymbol{W}^{(k)}$ est $\boldsymbol{W}$ évalué en $\boldsymbol{\beta}^{(k)}$ et $\boldsymbol{Z}^{(k)}$ est composé des éléments

$$\begin{aligned} z_i^{(k)} &= \sum_{j=1}^{t} x_{ij}\beta_j^{(k)} + (y_i - \mu_i^{(k)})\left(\frac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}}\right) \\ &= \eta_i^{(k)} + (y_i - \mu_i^{(k)})\left(\frac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}}\right). \end{aligned} \tag{3.21}$$

Finalement, les équations de Fisher (3.19) ont la forme suivante :

$$\begin{aligned} \boldsymbol{H}^{(k)}\boldsymbol{\beta}^{(k+1)} &= \boldsymbol{H}^{(k)}\boldsymbol{\beta}^{(k)} + \boldsymbol{q}^{(k)} \\ \boldsymbol{X'W}^{(k)}\boldsymbol{X}\boldsymbol{\beta}^{(k+1)} &= \boldsymbol{X'W}^{(k)}\boldsymbol{Z}^{(k)}. \end{aligned} \tag{3.22}$$

Si elle existe et est unique, la solution de ces équations est

$$\boldsymbol{\beta}^{(k+1)} = (\boldsymbol{X'W}^{(k)}\boldsymbol{X})^{-1}\boldsymbol{X'W}^{(k)}\boldsymbol{Z}^{(k)}. \tag{3.23}$$

Le vecteur $\boldsymbol{Z}^{(k)}$ représente, dans cette expression, un forme linéarisée de la fonction de lien en $\boldsymbol{\mu}$ évaluée en $\boldsymbol{y}$.

$$\begin{aligned} g(y_i) &\sim g(\mu_i) + (y_i - \mu_i) g'(\mu_i) \\ &\sim \eta_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} = z_i \end{aligned} \tag{3.24}$$

Il est important de remarquer que la méthode des scores de Fisher ne converge pas toujours. En pratique, on se donne une valeur $\epsilon > 0$ et un nombre maximum d'itérations K. Le processus itératif s'arrête dès que

$$\forall j \in \{1, \cdots, t\} \quad \frac{\left\| \beta_j^{(k+1)} - \beta_j^{(k)} \right\|}{\|\beta^{(k)}\|} < \epsilon$$

ou si $k = K$.

3.6 Equations d'estimation généralisées

Le but de ce paragraphe est d'étendre les modèles linéaires généralisés au cas de données appariées. Nous développerons l'approche préconisée par Liang et Zeger (1986), précurseurs dans ce domaine. Ceux-ci définissent les équations d'estimation généralisées (GEE) en se reposant sur un modèle dit *moyen de population*.

Considérons I blocs de données appariées mutuellement indépendants.

Nous noterons

$$\begin{aligned} \boldsymbol{y} &= (y_{11}, \cdots, y_{1n_1}, \cdots, y_{I1}, \cdots, y_{In_I}) \\ &= (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_I) \end{aligned}$$

le vecteur des observations,

$$\begin{aligned} \boldsymbol{\mu} &= (\mu_{11}, \cdots, \mu_{1n_1}, \cdots, \mu_{I1}, \cdots, \mu_{In_I}) \\ &= (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_I) \end{aligned} \tag{3.25}$$

le vecteur des moyennes et

$$\boldsymbol{X}_i = \begin{pmatrix} x_{i11} & \cdots & x_{i1p} \\ \vdots & & \vdots \\ x_{in_i1} & \cdots & x_{in_ip} \end{pmatrix}$$

la matrice $n_i \times p$ des covariables relatives au ième individu ($i = 1, \cdots, I$) où $\boldsymbol{x}_{ij} = (x_{ij1}, \cdots, x_{ijp})$ est un vecteur de covariables correspondant à la jème observation du ième individu.

Supposons que le vecteur $\boldsymbol{y}$ soit extrait d'une population exponentielle généralisée et qu'il existe le modèle linéaire suivant :

$$\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}, \quad i = 1, \cdots, I$$

où $\boldsymbol{\eta}_i = (\eta_{i1}, \cdots, \eta_{in_i})'$ est un vecteur de prédicteurs linéaires et $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ est un vecteur de paramètres inconnus. De plus,

$$\eta_{im} = g(\mu_{im}) \;\; i = 1, \cdots, I \quad m = 1, \cdots, n_i$$

est la fonction lien.

L'approche de Liang et Zeger (1986) nécessite des hypothèses sur la nature de la corrélation entre les données appariées.

Une des matrices de corrélation $n_i \times n_i$, proposée par Liang et Zeger (1986) est notée $\boldsymbol{R}_i(\alpha)$ et possède la forme suivante

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

où la corrélation inconnue entre les données appariées est supposée constante et notée α.

Beaucoup d'autres matrices de corrélation peuvent convenir. Elles ne seront pas commentées dans notre travail.

Notons

$\boldsymbol{A}_i = diag(b''(\theta_{im})a(\phi)), \quad i = 1, \cdots, I$ la matrice $n_i \times n_i$ correspondant aux variances sous le modèle d'exponentiel généralisé.

$\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\alpha)\boldsymbol{A}_i^{1/2}$.

Par analogie avec l'équation (3.15), Liang et Zeger (1986) proposent les équations d'estimation généralisées

$$\sum_{i=1}^{I} \boldsymbol{X}_i'\boldsymbol{\Delta}_i\boldsymbol{A}_i\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0} \tag{3.26}$$

Ces équations ne possèdent pas de solution analytique. Si la solution existe et est unique, alors nous pouvons appliquer la méthode itérative de Newton-Raphson pour résoudre ces équations.

3.7 Régression logistique

3.7.1 Régression logistique pour les données dichotomiques

Beaucoup de variables catégorielles possèdent seulement deux catégories. Chaque observation sur chacun des sujets peut être classée comme un succès (représenté par la valeur 1) ou un échec (représenté par la valeur 0). Pour les variables aléatoires binaires, la distribution de Bernoulli spécifie les probabilités $P(Y = 1) = \pi$, $P(Y = 0) = 1 - \pi$. Il en résulte que $E(Y) = \pi$ et $var(Y) = \pi(1 - \pi)$.

Lorsque Y_i possède une distribution de Bernoulli de paramètre π_i, la densité de distribution est

$$\begin{aligned} f(y_i, \pi_i) &= \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\ &= (1-\pi_i)\left[\frac{\pi_i}{1-\pi_i}\right]^{y_i} \\ &= \exp\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + ln(1-\pi_i)\right] \end{aligned} \tag{3.27}$$

pour $y_i = 0$ et $y_i = 1$.

Cette distribution appartient à la famille exponentielle généralisée. Le paramètre canonique, θ_i, est $\ln(\frac{\pi_i}{1-\pi_i})$. Ce terme est appelé $logit(\pi_i)$.

a) Régression linéaire

Pour des réponses dichotomiques, le modèle de régression linéaire est

$$E(Y) = \pi(x) = \beta_0 + \beta_1 x. \tag{3.28}$$

Lorsque les observations y sont indépendantes, ce modèle correspond au modèle linéaire généralisé avec la fonction identité comme fonction de lien.

Ce modèle a un défaut majeur. Alors que les proportions $\pi(x)$ doivent être comprises entre 0 et 1, les fonctions linéaires prennent des valeurs sur toute la droite réelle. La relation (3.28) peut prédire des valeurs de π inférieures à zéro ou supérieures à un. On propose donc de prendre une relation non linéaire entre $\pi(x)$ et x. Le modèle approprié est introduit dans le paragraphe suivant.

b) Régression logistique pour une seule covariable

Pour une variable explicative (covariable), le modèle proposé est le suivant :

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{3.29}$$

appelé la fonction de *régression logistique*.

Lorsque

$$\begin{aligned} x \to \infty \quad &, \quad \pi(x) \downarrow 0 \text{ si } \beta < 0; \\ &, \quad \pi(x) \uparrow 1 \text{ si } \beta > 0. \\ \pi(x) \quad &= \quad \frac{1}{2} \text{ si } x = -\frac{\beta_0}{\beta_1}. \end{aligned}$$

Ce modèle est représenté par une courbe sigmoïde et possède les propriétés d'une fonction de répartition continue.

La fonction de lien pour laquelle le modèle de régression logistique est un modèle linéaire généralisé est

$$\ln(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \beta_1 x. \tag{3.30}$$

Ainsi la fonction de lien correspond à la fonction $logit(\pi)$. On est donc en présence d'un lien canonique.

Supposons qu'il y ait g groupes de n_i observations ($i = 1, \cdots, g$). Soit y_i la ième observation de la variable aléatoire binomiale étudiée.

$$\ln\left(\frac{y_i}{n_i - y_i}\right) \tag{3.31}$$

n'est pas défini lorsque $y_i = 0$ ou $y_i = n_i$. C'est pourquoi on utilise souvent le *logit empirique*, qui est un estimateur légèrement biaisé du logit réel :

$$\ln\left(\frac{y_i + 1/2}{n_i - y_i + 1/2}\right). \tag{3.32}$$

La généralisation de la fonction logit à plusieurs variables explicatives est très simple. Soit $\boldsymbol{x} = (x_1, \cdots, x_k)'$ les valeurs de k variables explicatives. Le modèle de régression logistique est alors :

$$\ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{3.33}$$

3.7.2 Estimateurs du maximum de vraisemblance

Soient $(y_1, \cdots, y_n)$ les valeurs réalisées de n variables binaires. Nous supposerons que ces variables aléatoires sont indépendantes et possèdent une distribution de Bernoulli.

Soit $\boldsymbol{x}_i = (x_{i0}, \cdots, x_{ik})$ le ième ensemble de k variables explicatives, $i = 1, \cdots, I$ et $x_{i0} = 1$. Lorsque les covariables sont continues, il peut exister un ensemble de covariables

différent pour chaque sujet auquel cas $I = n$.

Le modèle de régression logistique est :

$$\pi(\boldsymbol{x}_i) = \frac{\exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)} \tag{3.34}$$

Soit n_i le nombre d'observations pour une valeur fixée de $\boldsymbol{x}_i = (x_{i0}, \cdots, x_{ik})$. Y_i est une variable aléatoire qui compte le nombre de succès. Les variables aléatoires Y_i, $(i = 1 \cdots, I)$ sont des variables aléatoires binomiales indépendantes où $E(Y_i) = n_i \pi(\boldsymbol{x_i})$
et $n_1 + \cdots + n_I = n$.

La densité de probabilité est

$$f(y_i, \pi) = \exp\left[y_i \ln\left(\frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)}\right) + n_i \ln(1 - \pi(\boldsymbol{x}_i))\right]. \tag{3.35}$$

Donc, on a

$$\eta_i = \theta_i; \quad \mu_i = n_i \pi(\boldsymbol{x}_i); \quad b(\theta_i) = -n_i \ln(1 - \pi(\boldsymbol{x}_i));$$
$$var(Y_i) = n_i \pi(\boldsymbol{x}_i)(1 - \pi(\boldsymbol{x}_i)) \quad \text{et} \quad a(\phi) = 1.$$

Ainsi,

$$\theta_i = \ln\left(\frac{n_i \pi(\boldsymbol{x}_i)}{n_i - n_i \pi(\boldsymbol{x}_i)}\right) \quad \text{et} \quad \pi(\boldsymbol{x}_i) = \frac{e^{\theta_i}}{1 + e^{\theta_i}}$$

Ce qui donne

$$b(\theta_i) = n_i \ln(1 + e^{\theta_i})$$

De plus,

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \left(\frac{\partial \theta_i}{\partial \mu_i}\right)^{-1}$$

Or,

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{n_i \pi(\boldsymbol{x}_i)} + \frac{1}{n_i(1 - \pi(\boldsymbol{x}_i))} = \frac{1}{n_i \pi(\boldsymbol{x}_i)(1 - \pi(\boldsymbol{x}_i))}$$

Donc

$$\frac{\partial \mu_i}{\partial \theta_i} = n_i \pi(\boldsymbol{x}_i)(1 - \pi(\boldsymbol{x}_i)) = var(Y_i)$$

Les équations du maximum de vraisemblance sont donc

$$\sum_{i=1}^{I} (y_i - n_i \pi_i) x_{ij} = 0, \quad j = 0, \cdots, k \tag{3.36}$$

ou, sous forme matricielle, si $\boldsymbol{X}$ est la matrice $I \times (k+1)$ composée des valeurs $\{x_{ij}\}$, (3.36) a la forme

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{\mu} \tag{3.37}$$

La matrice d'information s'écrit alors

$$H_{kj} = \sum_{i=1}^{I} \frac{x_{ik}x_{ij}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \tag{3.38}$$

ou

$$\boldsymbol{H} = \boldsymbol{X}'\boldsymbol{diag}(var(Y_i))\boldsymbol{X} \tag{3.39}$$

Pour la régression logistique, les estimateurs du maximum de vraisemblance existent et sont uniques sauf dans certains cas limites (Wedderburn (1976), Albert et Anderson (1984)). Les équations du maximum de vraisemblance sont des fonctions non linéaires des estimations du maximum de vraisemblance $\hat{\boldsymbol{\beta}}$. La résolution de ces équations se fait par la méthode de Newton-Raphson modifiée.

3.8 Régression logistique ordinale

Il existe différents modèles de régressions logistiques pour données ordinales. Seul sera mentionné celui dont on se servira dans le Chapitre 4. Le lecteur intéressé par les autres modèles peut se référer à l'ouvrage de Hosmer D. et Lemeshow S. (2000).

Soient une variable aléatoire ordinale Y, pouvant prendre $K+1$ valeurs, notées $0, 1, \cdots, K$ et $\boldsymbol{x} = (x_1, \cdots, x_p)'$ un vecteur de p covariables.

Notons

$$\mathbb{P}[Y = k|\boldsymbol{x}] = \phi_k(\boldsymbol{x}). \tag{3.40}$$

Supposons que l'on désire comparer les probabilités $\mathbb{P}[Y \leq k|\boldsymbol{x}]$ et $\mathbb{P}[Y > k|\boldsymbol{x}]$.

On définit alors

$$\begin{aligned} c_k(\boldsymbol{x}) &= \ln\left[\frac{\mathbb{P}[Y \leq k|\boldsymbol{x}]}{\mathbb{P}[Y > k|\boldsymbol{x}]}\right] \\ &= \ln\left[\frac{\phi_0(\boldsymbol{x}) + \cdots + \phi_k(\boldsymbol{x})}{\phi_{k+1}(\boldsymbol{x}) + \cdots + \phi_K(\boldsymbol{x})}\right] \\ &= \ln\left[\frac{\gamma_k(\boldsymbol{x})}{1 - \gamma_k(\boldsymbol{x})}\right] \\ &= \tau_k - \boldsymbol{x}'\boldsymbol{\beta} \end{aligned} \tag{3.41}$$

pour $k = 0, \cdots, K-1$, où $\gamma_k = \phi_0(\boldsymbol{x}) + \cdots + \phi_k(\boldsymbol{x})$ et τ_k est l'ordonnée à l'origine.

Ce modèle est le modèle logistique cumulatif linéaire, il possède la propriété suivante

$$\ln\left(\frac{\gamma_k(\boldsymbol{x}_1)}{1-\gamma_k(\boldsymbol{x}_1)}\right) - \ln\left(\frac{\gamma_k(\boldsymbol{x}_2)}{1-\gamma_k(\boldsymbol{x}_2)}\right) = \boldsymbol{\beta}'(\boldsymbol{x}_1 - \boldsymbol{x}_2) \tag{3.42}$$

c'est-à-dire que la différence entre les deux logistiques ne dépend pas de la catégorie k.

Mac Cullagh (1980) a défini les équations du maximum de vraisemblance et le procédé itératif de Newton-Raphson pour les régressions logistiques ordinales. Ces équations ont été détaillées par Noiroux (2000). Elles ne seront pas reproduites ici.

De nombreux logiciels sont programmés pour calculer les régressions logistiques binaires et ordinales. Parmi ceux-ci, citons SAS, S-PLUS, STATISTICA et GENSTAT.

3.9 Qualité de l'ajustement

3.9.1 La déviance

Un critère de la qualité d'ajustement d'une régression logistique ou ordinale est donné par

$$D = -2\sum_{i=1}^{n}[l_i(\hat{\mu}_i) - l_i(y_i)] \tag{3.43}$$

où les l_i sont définis par (3.11).

La statistique D est appelée *la déviance*. On peut démontrer que D est asymptotiquement distribué comme une variable Chi-carré à $n-g$ degrés de liberté où g est le nombre de paramètres estimés.

Considérons deux modèles emboîtés

$$M_0 : \quad \ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 \quad i = 1, \cdots, n$$

et

$$M_p : \quad \ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \quad i = 1, \cdots, n.$$

Soient $D(M_0)$ et $D(M_p)$ les déviances relatives aux modèles M_0 et M_p respectivement.

On peut démontrer que

$$D(M_p|M_0) = D(M_0) - D(M_p) \tag{3.44}$$

est asymptotiquement distribuée comme un Chi-carré à p degrés de liberté.

$D(M_p|M_0)$ est un indice de la qualité d'ajustement de la régression qui comporte p covariables.

3.9.2 Erreurs types des estimateurs des paramètres

Il est bien connu que l'erreur type de l'estimateur $\hat{\underline{\beta}}_i$, $i = 1, \cdots, p$ est donnée par

$$s.e.(\hat{\underline{\beta}}_i) = (H^{-1})_{ii}, \quad i = 1, \cdots, p \tag{3.45}$$

où $\boldsymbol{H}^{-1}$ est l'inverse de la matrice d'information
et que

$$Z(\hat{\underline{\beta}}_i) = \frac{\hat{\underline{\beta}}_i - \beta_i}{s.e.(\beta_i)} \tag{3.46}$$

est asymptotiquement normale de moyenne nulle et d'écart-type égal à 1.

On peut donc éprouver l'hypothèse

$$H_0 : \quad \beta_i = 0 \text{ vs } \quad H_1 : \quad \beta_i \neq 0 \quad i = 0, \cdots, p$$

au niveau d'incertitude α en comparant $|Z(\beta_i)|$ à $Q_Z(1 - \alpha/2)$.

De nombreux auteurs testent H_0 au moyen de la statistique dite de "Wald" définie par

$$\chi^2_{\beta_i} = Z^2(\beta_i) \tag{3.47}$$

On rejette H_0 au niveau d'incertitude α si

$$\chi^2_{\beta_i} > Q_{\chi^2}(1 - \alpha; 1)$$

sinon on ne rejette pas H_0.

3.10 Coefficient Kappa et régression logistique

Shoukri et Mian (1996) ont proposé un estimateur du maximum de vraisemblance du coefficient κ pour une table 2×2 lorsque les jugements dichotomiques dépendent de covariables relatives aux sujets et /ou aux examinateurs.

3.10.1 Modèle pour données bivariées binaires

Soit $\underline{Y}_{ij}$ une variable aléatoire binaire telle que $\underline{Y}_{ij} = 1$ représente un jugement positif émis par le jème observateur ($j = 1, 2$) sur le ième sujet ($i = 1, \cdots, n$), et $\underline{Y}_{ij} = 0$ sinon.

Les probabilités assignées aux cellules de la table 2×2 sont représentées dans la table 3.1.

TABLE 3.1 – Réponses des observateurs relatives au ième sujet

		Observateur 1		
		1	2	Total
Observateur 2	1	P_{i11}	P_{i21}	π_{i2}
	2	P_{i12}	P_{i22}	π'_{i2}
Total		π_{i1}	π'_{i1}	1

La probabilité conjointe que la réponse k soit observée par le premier examinateur et que la réponse l soit observée par le second examinateur sur le sujet i est $P_{ikl} = \mathbb{P}[\underline{Y}_{i1} = k, \underline{Y}_{i2} = l]$ où $i = 1, \cdots, n$; $k = 0, 1$ et $l = 0, 1$.

Nous avons

$$\begin{aligned} P_{i11} &= \mathbb{P}\{[\underline{Y}_{i1} = 1] \cap [\underline{Y}_{i2} = 1]\} = E[\underline{Y}_{i1}\underline{Y}_{i2}] \\ P_{i12} &= E[\underline{Y}_{i1}(1 - \underline{Y}_{i2})] \\ P_{i21} &= E[\underline{Y}_{i2}(1 - \underline{Y}_{i1})] \\ P_{i22} &= E[(1 - \underline{Y}_{i2})(1 - \underline{Y}_{i1})] = 1 - P_{i11} - P_{i12} - P_{i21}. \end{aligned}$$

Le vecteur aléatoire $\underline{\mathbf{Y}}_i = (\underline{Y}_{i1}\underline{Y}_{i2}, \underline{Y}_{i1}(1 - \underline{Y}_{i2}), \underline{Y}_{i2}(1 - \underline{Y}_{i1}))'$ possède une distribution multinomiale dont le vecteur des paramètres de population est $\mathbf{P}_i = (P_{i11}, P_{i21}, P_{i12})'$.

Remarquons que ce vecteur ne possède que 3 composantes en vertu de la relation linéaire $P_{i22} = E[(1 - \underline{Y}_{i2})(1 - \underline{Y}_{i1})] = 1 - P_{i11} - P_{i12} - P_{i21}$.

On a

$$E(\underline{Y}_{i1}) = (1, 1, 0)\mathbf{P}_i = \pi_{i1},$$
$$E(\underline{Y}_{i2}) = (1, 0, 1)\mathbf{P}_i = \pi_{i2}.$$

L'association linéaire entre $\underline{Y}_{i1}$ et $\underline{Y}_{i2}$ peut être mesurée, entre autre, par le coefficient de corrélation :

$$\rho_i = \frac{cov(\underline{Y}_{i1}, \underline{Y}_{i2})}{(var(\underline{Y}_{i1}))^{1/2}(var(\underline{Y}_{i2}))^{1/2}} \tag{3.48}$$

On a

$$\begin{aligned} cov(\underline{Y}_{i1}, \underline{Y}_{i2}) &= E[(\underline{Y}_{i1} - \pi_{i1})(\underline{Y}_{i2} - \pi_{i2})] \\ &= E[\underline{Y}_{i1}\underline{Y}_{i2} - \pi_{i2}\underline{Y}_{i1} - \pi_{i1}\underline{Y}_{i2} + \pi_{i1}\pi_{i2}] \\ &= P_{i11} - \pi_{i1}\pi_{i2} \end{aligned} \tag{3.49}$$

et

$$\begin{aligned} var(\underline{Y}_{ij}) &= (1-\pi_{ij})^2\pi_{ij} + (0-\pi_{ij})^2(1-\pi_{ij}) \\ &= \pi_{ij}\pi'_{ij} \qquad (j=1,2) \end{aligned} \tag{3.50}$$

Donc,

$$\rho_i = \frac{P_{i11} - \pi_{i1}\pi_{i2}}{(\pi_{i1}\pi'_{i1}\pi_{i2}\pi'_{i2})^{1/2}}. \tag{3.51}$$

D'où

$$P_{i11} = \pi_{i1}\pi_{i2} + \tau_i, \quad P_{i12} = \pi_{i1}\pi'_{i2} - \tau_i \tag{3.52}$$

$$P_{i21} = \pi_{i2}\pi'_{i1} - \tau_i, \quad P_{i22} = \pi'_{i1}\pi'_{i2} + \tau_i \tag{3.53}$$

où
$\tau_i = \rho_i(\pi_{i1}\pi'_{i1}\pi_{i2}\pi'_{i2})^{1/2}$.

De nombreux auteurs ont démontré depuis très longtemps que

$$-\min\left\{\left(\frac{\pi_{i1}\pi_{i2}}{\pi'_{i1}\pi'_{i2}}\right)^{1/2}, \left(\frac{\pi'_{i1}\pi'_{i2}}{\pi_{i1}\pi_{i2}}\right)^{1/2}\right\} \leq \rho_i \leq \min\left\{\left(\frac{\pi_{i1}\pi'_{i2}}{\pi'_{i1}\pi_{i2}}\right)^{1/2}, \left(\frac{\pi'_{i1}\pi_{i2}}{\pi_{i1}\pi'_{i2}}\right)^{1/2}\right\} \tag{3.54}$$

Bahadur (1961) a montré que la connaissance des deux probabilités marginales π_{i1} et π_{i2} et du coefficient de corrélation ρ_i (3.51) permet de caractériser la distribution bivariée d'observations binaires.

En effet, la connaissance de π_{i1} et π_{i2} permet de caractériser les totaux marginaux en vertu des relations linéaires

$$\pi_{i1} + \pi'_{i1} = 1 \quad \text{et} \quad \pi_{i2} + \pi'_{i2} = 1$$

et la connaissance du coefficient de corrélation ρ_i permet de déterminer P_{i11}. Les autres probabilités sont déterminées par les relations linéaires suivantes

$$P_{i21} = \pi_{i2} - P_{i11}, \quad P_{i12} = \pi_{i1} - P_{i11} \quad \text{et} \quad P_{i22} = P_{i11} - P_{i12} - P_{i21}$$

Soient y_{i1} et y_{i2} les valeurs observées de $\underline{Y}_{i1}$ et $\underline{Y}_{i2}$ respectivement. La densité de probabilité de $\underline{Y}_{i1}$ et $\underline{Y}_{i2}$ par rapport à une mesure de comptage est proportionnelle à

$$f(\underline{Y}_{i1}, \underline{Y}_{i2}) = P_{i11}^{y_{i1}y_{i2}} P_{i10}^{y_{i1}(1-y_{i2})} P_{i01}^{y_{i2}(1-y_{i1})} P_{i00}^{(1-y_{i1})(1-y_{i2})} \tag{3.55}$$

avec $0 < \pi_{i1}, \pi_{i2} < 1$.

Le coefficient κ est défini de manière habituelle par l'équation (**??**).

En utilisant les relations (3.52) et (3.53), on obtient aisément l'expression suivante du coefficient κ :

$$\kappa_i = \frac{2\rho_i(\pi_{i1}\pi'_{i1}\pi_{i2}\pi'_{i2})^{1/2}}{\pi_{i1}\pi'_{i2} + \pi_{i2}\pi'_{i1}} \tag{3.56}$$

Le coefficient κ donne donc des informations sur l'association par l'intermédiaire du coefficient ρ_i et le biais existant entre les observateurs en fonction de la différence entre π_{i1} et π_{i2}.

Une association parfaite ($\rho_i = 1$) n'implique pas un accord parfait ($\kappa_i = 1$), à moins que $\pi_{i1} = \pi_{i2}$.

Vu la relation entre κ_i et ρ_i, nous pouvons écrire la distribution de probabilité conjointe (3.55) comme une fonction de κ_i. La somme des probabilités devant être égale à 1, κ doit satisfaire aux contraintes suivantes :

$$a_i \leq \kappa_i \leq b_i \tag{3.57}$$

où

$$a_i = -2\min\left\{\frac{\pi_{i1}\pi_{i2}}{\pi_{i1}\pi'_{i2} + \pi_{i2}\pi'_{i1}}, \frac{\pi'_{i1}\pi'_{i2}}{\pi_{i1}\pi'_{i2} + \pi_{i2}\pi'_{i1}}\right\}$$

$$b_i = 2\min\left\{\frac{\pi_{i1}\pi'_{i2}}{\pi_{i1}\pi'_{i2} + \pi_{i2}\pi'_{i1}}, \frac{\pi'_{i1}\pi_{i2}}{\pi_{i1}\pi'_{i2} + \pi_{i2}\pi'_{i1}}\right\}$$

Puisque le coefficient κ dépend de l'indice i de l'individu, le nombre de paramètres nécessaires $(\rho_i, \pi_{i1}, \pi_{i2})$ pour estimer κ augmente avec la taille de l'échantillon. Pour réduire ce nombre de paramètres nous supposerons qu'il existe une valeur de κ commune à tous les sujets et que les valeurs permises pour ce κ commun sont comprises entre la valeur maximale des a_i et la valeur minimale des b_i.

3.10.2 Principe de la régression logistique

Lorsque les deux jugements dichotomiques sur le sujet i dépendent des covariables associées aux observateurs et/ou aux sujets, chaque sujet possède un vecteur de covariables $\boldsymbol{x}_i$ spécifique aux sujets et deux vecteurs de covariables $(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$ relatifs aux observateurs, à savoir un vecteur par observateur. Nous pouvons regrouper ces vecteurs de covariables comme suit $\boldsymbol{z}'_{ij} = (\boldsymbol{x}'_i, \boldsymbol{x}'_{ij})$, $i = 1, \cdots, n$ et $j = 1, 2$. Nous supposerons que la logistique de π_{ij} est une fonction linéaire à coefficients connus $\boldsymbol{z}_{ij}$ de paramètres $\boldsymbol{\beta}$ inconnus :

$$logit(\pi_{ij}) = \boldsymbol{z}'_{ij}\boldsymbol{\beta} \tag{3.58}$$

3.10.3 Estimateurs du maximum de vraisemblance

Soient (y_{i1}, y_{i2}), $i = 1, \cdots, n$ un échantillon simplement fortuit de n paires de réponses binaires corrélées de densité conjointe $f(y_{i1}, y_{i2})$ donnée par la fonction (3.55) où ρ_i est remplacé par $\kappa(\pi_{i1}\pi'_{i2} + \pi_{i2}\pi'_{i2})/2\sqrt{(\pi_{i1}\pi'_{i1}\pi_{i2}\pi'_{i2})}$. Le logarithme de la vraisemblance est donné par

$$L(\kappa, \boldsymbol{\beta}) = \sum_{i=1}^{n} L_i(\kappa, \boldsymbol{\beta}) \tag{3.59}$$

où

$$\begin{aligned} L_i(\kappa, \boldsymbol{\beta}) &= y_{i1}y_{i2} \ln \left[\pi_{i1}\pi_{i2} + \frac{\kappa}{2}((\pi_{i1}\pi'_{i2} + \pi_{i2}\pi'_{i1}))\right] \\ &+ y_{i1}(1 - y_{i2}) \ln \left[\pi_{i1}\pi'_{i2} - \frac{\kappa}{2}((\pi_{i1}\pi'_{i2} + \pi_{i2}\pi'_{i1}))\right] \\ &+ y_{i2}(1 - y_{i1}) \ln \left[\pi'_{i1}\pi_{i2} - \frac{\kappa}{2}((\pi_{i1}\pi'_{i2} + \pi_{i2}\pi'_{i1}))\right] \\ &+ (1 - y_{i1})(1 - y_{i2}) \ln \left[\pi'_{i1}\pi'_{i2} + \frac{\kappa}{2}((\pi'_{i1}\pi_{i2} + \pi_{i2}\pi'_{i1}))\right] \end{aligned}$$

et où en vertu du modèle (3.58), les π_{ij} sont des fonctions connues des paramètres $\boldsymbol{\beta}$.

Dès lors, les estimateurs du maximum de vraisemblance $\hat{\kappa}$ et $\hat{\boldsymbol{\beta}}$ de κ et $\boldsymbol{\beta}$ respectivement, s'obtiennent en résolvant les équations

$$\frac{\partial L(\kappa, \boldsymbol{\beta})}{\partial \kappa} = 0 \tag{3.60}$$

et

$$\frac{\partial L(\kappa, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0} \tag{3.61}$$

où

$$\begin{aligned} \frac{\partial L(\kappa, \boldsymbol{\beta})}{\partial \kappa} &= \frac{1}{2} \sum_{i=1}^{n} \left\{ y_{i1}y_{i2} \frac{\nu_i}{P_{i11}} - y_{i1}(1 - y_{i2}) \frac{\nu_i}{P_{i12}} \right. \\ &- \left. y_{i2}(1 - y_{i1}) \frac{\nu_i}{P_{i21}} + (1 - y_{i1})(1 - y_{i2}) \frac{\nu_i}{P_{i22}} \right\} \end{aligned} \tag{3.62}$$

$$\begin{aligned} \frac{\partial L(\kappa, \boldsymbol{\beta})}{\partial \beta_r} &= \sum_{i=1}^{n} \left\{ y_{i1}y_{i2} \frac{e_{ir}}{P_{i11}} + y_{i1}(1 - y_{i2}) \frac{(d_{i2r} - e_{ir})}{P_{i12}} \right. \\ &+ \left. y_{i2}(1 - y_{i1}) \frac{(d_{i1r} - e_{ir})}{P_{i21}} + (1 - y_{i1})(1 - y_{i2}) \frac{(e_{ir} - d_{i1r} - d_{i2r})}{P_{i22}} \right\} \end{aligned} \tag{3.63}$$

et

$$\begin{aligned}
\nu_i &= \pi_{i1}(1-\pi_{i2}) + \pi_{i2}(1-\pi_{i1}) \\
e_{ir} &= \frac{\partial}{\partial \beta_r}\left[\pi_{i1}\pi_{i2} + \frac{\kappa}{2}\nu_i\right] \\
&= z_{i1r}\pi_{i1}\pi'_{i1}[\pi_{i2} - \kappa(\pi_{i2} - 1/2)] + z_{i2r}\pi_{i2}\pi'_{i2}[\pi_{i1} - \kappa(\pi_{i1} - 1/2)] \\
d_{irj} &= \frac{\partial \pi_{ij}}{\partial \beta_r} = z_{ijr}\pi_{ij}\pi'_{ij}
\end{aligned}$$

$i = 1, \cdots, n; \quad j = 1, 2 \;\text{ et }\; r = 0, \cdots, k.$

Les équations (3.60) et (3.61) ne possèdent pas de solution analytique. Les estimateurs $\hat{\kappa}$ et $\hat{\boldsymbol{\beta}}$ doivent être calculés par une méthode itérative d'analyse numérique comme par exemple la méthode de Newton-Raphson généralisée. Les conditions d'existence et d'unicité des solutions n'ont jamais été étudiées. L'application de la méthode de Newton-Raphson implique le calcul des dérivées secondes de $l(\kappa, \boldsymbol{\beta})$ par rapport à $\boldsymbol{\beta}$ et κ.
La matrice d'information $\boldsymbol{H}$ est composée des éléments

$$\begin{aligned}
H_{\kappa\kappa} &= -E\left[\frac{\partial^2 L(\kappa, \boldsymbol{\beta})}{\partial \kappa^2}\right] \\
&= \frac{1}{4}\sum_{i=1}^{n} \nu_i^2 \left\{ \frac{P_{i21}P_{i12}P_{i22} + P_{i11}P_{i21}P_{i22} + P_{i11}P_{i12}P_{i22} + P_{i11}P_{i12}P_{i21}}{P_{i11}P_{i12}P_{i21}P_{i22}} \right\} \qquad (3.64) \\
H_{rs} &= -E\left[\frac{\partial^2 L(\kappa, \boldsymbol{\beta})}{\partial \beta_r \beta_s}\right] \\
&= \sum_{i=1}^{n} \left\{ \frac{e_{ir}e_{is}}{P_{i11}} + \frac{(d_{i1r} - e_{ir})(d_{i1s} - e_{is})}{P_{i21}} + \frac{(d_{i2r} - e_{ir})(d_{i2s} - e_{is})}{P_{i12}} \right. \\
&+ \left. \frac{(d_{i1r} + d_{i2r} - e_{ir})(d_{i1s} + d_{i2s} - e_{is})}{P_{i22}} \right\} \qquad (3.65) \\
H_{r\kappa} &= -E\left[\frac{\partial^2 L(\kappa, \boldsymbol{\beta})}{\partial \beta_r \kappa}\right] \\
&= \frac{1}{2}\sum_{i=1}^{n} \nu_i \left\{ \frac{e_{ir}}{P_{i11}} - \frac{(d_{i1r} - e_{ir})}{P_{i21}} - \frac{(d_{i2r} - e_{ir})}{P_{i21}} + \frac{(e_{ir} - d_{i1r} - d_{i2r})}{P_{i22}} \right\} \qquad (3.66)
\end{aligned}$$

Si l'algorithme de Newton-Raphson modifié ne converge pas, Shoukri et Mian (1996) proposent une procédure d'estimation en deux étapes. La première consiste à estimer $\boldsymbol{\beta}$ et donc les paramètres π_{ij} en utilisant une régression logistique. La seconde à estimer κ avec un estimateur intuitif :

$$\widetilde{\kappa} = 2\sum_{i=1}^{n} \frac{(y_{i1} - \hat{\pi}_{i1})(y_{i2} - \hat{\pi}_{i2})}{\hat{\pi}_{i1}\hat{\pi}'_{i2} + \hat{\pi}_{i2}\hat{\pi}'_{i1}} \qquad (3.67)$$

Cet estimateur $\widetilde{\kappa}$ existe pour autant que la régression logistique converge et que les estimations $\hat{\pi}_{i1}$ et $\hat{\pi}_{i2}$ ne soient pas proches de leurs bornes inférieure et supérieure.

3.11 Exemples

3.11.1 Exemple 1

Hui et Walter (1980) ont fourni des données relatives à la détection de la tuberculose au moyen de tests cutanés. L'observateur 1 utilise le test Standard de Mantoux. L'observateur 2 utilise un nouveau test dénommé "Tine" test. Une réaction cutanée se traduit par un gonflement de la peau plus grand qu'une taille fixée après 48 heures et constitue un résultat positif noté "+". Les données de la population 1 proviennent d'une étude menée dans une école du sud des Etats-Unis par Greenberg et Jekel (1969). Selon le même protocole médical, la seconde étude a été menée dans le sanatorium de l'Etat du Missouri par Capobres et al (1962). Les données sont représentées dans la table 3.2.

TABLE 3.2 – Résultats des deux tests de détection de la tuberculose dans deux populations indépendantes

		Population 1			Population 2		
		Tine test (Observateur 2)			Tine test (Observateur 2)		
		+	−	Totaux	+	−	Totaux
Test de Mantoux (observateur 1)	+	14	4	18	887	31	918
	−	9	528	537	37	367	404
	Totaux	23	532	555	924	398	1322

Puisqu'ils sont en présence de deux populations et deux observateurs indépendants, Shoukri et Mian (1996) définissent deux covariables au moyen des variables factices suivantes.

$$z_1 = \begin{cases} 1 \text{ si le Tine test est appliqué} \\ 0 \text{ si le test de Mantoux est appliqué} \end{cases}$$

$$z_2 = \begin{cases} 1 \text{ si la population 1 est étudiée} \\ 0 \text{ si la population 2 est étudiée} \end{cases}$$

Le modèle proposé est le suivant

$$logit(\pi_{ij}) = \beta_0 + \beta_1 z_{ij1} + \beta_2 z_{ij2} \tag{3.68}$$

$i = 1, 2$, $j = 1, 2$.

Remarquons que les π_{ij} ne sont pas relatifs à chaque sujet. L'indice i désigne la population ($i = 1, 2$) et l'indice j désigne l'observateur ($j = 1, 2$).

ou, sous forme matricielle,

$$\ln \overbrace{\begin{pmatrix} \frac{\pi_{11}}{1-\pi_{11}} \\ \frac{\pi_{12}}{1-\pi_{12}} \\ \frac{\pi_{21}}{1-\pi_{21}} \\ \frac{\pi_{22}}{1-\pi_{22}} \end{pmatrix}}^{} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Le vecteur des observations est

$$\ln \overbrace{\begin{pmatrix} \frac{\pi_{11}}{1-\pi_{11}} \\ \frac{\pi_{12}}{1-\pi_{12}} \\ \frac{\pi_{21}}{1-\pi_{21}} \\ \frac{\pi_{22}}{1-\pi_{22}} \end{pmatrix}}^{} = \ln \begin{pmatrix} \frac{18}{537} \\ \frac{23}{532} \\ \frac{918}{404} \\ \frac{924}{398} \end{pmatrix}$$

TABLE 3.3 – Estimations du maximum de vraisemblance obtenues à partir des données de la table 3.2

Paramètre	Estimation	Erreur type	Z-score	p-value
β_0	0.8547	0.0596	14.3406	< 0.0001
β_1	-0.0366	0.0302	-1.2119	0.2256
β_2	-3.9501	0.2137	-18.4843	< 0.0001
κ	0.8651	0.0148	58.4527	< 0.0001

Aux vues du tableau 3.3, on peut remarquer que $\hat{\beta}_1$ est non significativement différent de 0 (niveau d'incertitude : $\alpha = 5\%$) puisque la p-value correspondant à β_1 est supérieure à 0,05. Cela signifie qu'il n'existe pas de différence significative entre le Tine test et le test de Mantoux. Par contre $\hat{\beta}_2$ est significativement différent de 0. Ainsi, le fait que l'on effectue le test de détection de la tuberculose dans le sanatorium de l'état du Missouri ou dans une école du sud des Etats-Unis contribue à la prédiction de la valeur de κ. Ajoutons que selon l'échelle **??** définie par Landis et Koch (1977a), l'accord entre les observateurs peut être qualifié de "très bon".

Il est aussi intéressant de remarquer que le terme d'interaction $z_3 = z_1 \times z_2$ n'apparaît pas dans le modèle. Celui-ci est considéré comme non-significatif ce qui implique l'hypothèse d'un κ constant à travers les populations. Nous pouvons renforcer notre opinion

sur cette hypothèse en comparant les estimations obtenues à partir du modèle (3.68) à l'estimation combinée des κ définie au paragraphe 1.6.2 par Fleiss. On a

$$\hat{\kappa}_{asso} = 0.8730 \quad \text{et} \quad s.e.(\hat{\kappa}_{asso}) = 0.0145$$

3.11.2 Exemple 2

Oden (1991) a fourni des données binaires sur l'absence et la présence d'une atrophie géométrique dans les yeux de 840 patients, où les deux mêmes examinateurs calibraient chaque oeil.

TABLE 3.4 – Données binoculaires d'Oden

		Oeil gauche			Oeil droit		
		Observateur 2			Observateur 2		
		+	−	Totaux	+	−	Totaux
Observateur 1	+	6	5	11	9	4	13
	−	12	817	829	11	816	827
	Totaux	18	822	840	20	820	840

Oden propose deux estimateurs d'un κ commun qui tiennent compte de la corrélation existante entre les yeux d'une même personne. Le premier, κ_{pooled}, a été obtenu sous l'hypothèse d'homogénéité, c'est-à-dire que toute différence entre le κ estimé pour l'oeil gauche et l'oeil droit est supposée être due simplement au hasard. Le second estimateur, κ_{ave}, a été obtenu en faisant une moyenne pondérée de κ_g obtenu sur les yeux gauches et de κ_d obtenu sur les yeux droits.

Shoukri et Mian (1996) utilisent les données d'Oden pour trouver l'estimateur du maximum de vraisemblance du coefficient κ ajusté aux effets des yeux et des examinateurs. De la même manière que dans l'exemple 1, deux covariables sont introduites :

$$z_1 = \begin{cases} 1 \text{ pour l'observateur 2} \\ 0 \text{ pour l'observateur 1} \end{cases}$$

$$z_2 = \begin{cases} 1 \text{ pour l'oeil gauche} \\ 0 \text{ pour l'oeil droit} \end{cases}$$

Le même modèle (3.68) est proposé. Les résultats sont présentés dans la table 3.5.

TABLE 3.5 – Estimations du maximum de vraisemblance obtenues à partir des données d'Oden

Paramètre	Estimation	Erreur type	Z-score	p-value
β_0	-4.2104	0.2466	-17.0378	< 0.0001
β_1	0.4680	0.1905	2.4567	0.0140
β_2	-0.0479	0.2975	-0.1610	0.8721
κ	0.4747	0.0794	5.9786	< 0.0001

Dans cet exemple, on peut remarquer que le fait d'examiner l'oeil gauche ou l'oeil droit des différents patients n'aide pas à prédire la valeur de κ car $\hat{\beta}_2$ n'est pas significativement différent de 0. Par contre la covariable z_1 relative aux observateurs est significativement prédictive de la valeur de κ. Ajoutons que l'accord entre les deux observateurs est qualifié par Landis et Koch (1977a) de "modéré".

Remarquons que l'approche de Shoukri et Mian (1996) est uniquement valable lorsque la corrélation entre les yeux est nulle. Le fait d'ignorer une corrélation non nulle, dans une situation réelle, peut être une erreur méthodologique.

3.12 Discussion

Dans ce chapitre, nous avons résumé les notions de famille exponentielle, de modèles linéaires généralisés et d'estimateurs du maximum de vraisemblance afin de pouvoir les adapter à un modèle logistique d'estimation de κ en présence de covariables.

Le modèle de Shoukri et Mian a été établi pour les tables de contingence 2×2 uniquement.

Pour les tables générales $k \times k$ ($k > 2$), la complexité de la fonction de vraisemblance est telle que le problème n'a jamais été envisagé. Nous verrons au Chapitre 4 une méthode qui permet de contourner ces difficultés.

Il est important de remarquer que puisque la vraisemblance comporte κ parmi ses paramètres, les équations du maximum de vraisemblance (3.60) et (3.61) ne sont pas identiques à celles d'une régression logistique classique.

Contrairement aux modèles linéaires généralisés, il n'existe actuellement pas de logiciel pour résoudre ces équations.

Enfin, les conditions d'existence et d'unicité des solutions n'ont jamais été envisagées.

Chapitre 4

Modèles de régression pour κ en présence de covariables

4.1 Introduction

Lipsitz et al (2001) ont développé une méthode simple qui permet de modéliser le coefficient κ comme une fonction de covariables relatives aux observateurs et/ou aux sujets. Afin d'estimer le modèle de régression pour le coefficient κ, Lipsitz et al proposent d'utiliser deux régressions logistiques et une régression linéaire pour les données binaires. Nous nous proposons de décrire cette méthode dans le présent chapitre.

4.2 Observations dichotomiques : notations et modèle

Soient $i = 1, \cdots, n$ sujets indépendants. Deux observateurs observent une caractéristique discrète dichotomique. Soit la variable aléatoire $\underline{Y}_{ir}$ telle que $\underline{Y}_{ir} = 1$ si le sujet i est jugé positif par l'observateur r, $\underline{Y}_{ir} = 0$ sinon $(r = 1, 2)$.

On suppose que chaque sujet possède un vecteur de covariables spécifiques aux sujets noté $\mathbf{x}_i$. On suppose également qu'il existe deux vecteurs de covariables spécifiques aux observateurs $\mathbf{x}_{ir},\ r = 1, 2$ comme dans l'approche de Shoukri et Mian (1996). On pose $\mathbf{z}_i' = (\mathbf{x}_i', \mathbf{x}_{i1}', \mathbf{x}_{i2}')$.

On définit une variable aléatoire indicatrice $\underline{Y}_i$ telle que $\underline{Y}_i = 1$ si les deux observateurs sont d'accord sur le sujet i et $\underline{Y}_i = 0$ sinon. En terme de $\underline{Y}_{i1}$ et $\underline{Y}_{i2}$, on a

$$\underline{Y}_i = \underline{Y}_{i1}\underline{Y}_{i2} + (1 - \underline{Y}_{i1})(1 - \underline{Y}_{i2}). \tag{4.1}$$

Le coefficient Kappa de Cohen pour mesurer l'accord entre $\underline{Y}_{i1}$ et $\underline{Y}_{i2}$ est défini de manière habituelle

$$\kappa_i = \frac{p_{oi} - p_{ei}}{1 - p_{ei}} \tag{4.2}$$

où

$$p_{oi} = \mathbb{P}[\underline{Y}_i = 1|\mathbf{z}_i] = \mathbb{P}[\underline{Y}_{i1} = 1, \underline{Y}_{i2} = 1|\mathbf{z}_i] + \mathbb{P}[\underline{Y}_{i1} = 0, \underline{Y}_{i2} = 0|\mathbf{z}_i]$$

et

$$p_{ei} = p_{i1}p_{i2} + (1 - p_{i1})(1 - p_{i2})$$

où $p_{ir} = \mathbb{P}[\underline{Y}_{ir} = 1|\mathbf{x}_i, \mathbf{x}_{ir}], \quad r = 1, 2.$

Pour une valeur donnée de p_{ei}, nous avons vu au paragraphe **??** que

$$-1 \leq \kappa_i \leq 1. \tag{4.3}$$

En effet, ici, nous sommes dans le cas d'observations binaires effectuées sur un sujet i.

Un modèle linéaire en les paramètres pourrait être défini au moyen d'une fonction de lien $g(.)$ telle que

$$g(\kappa_i) = \mathbf{z}_i'\boldsymbol{\gamma}$$

où $\boldsymbol{\gamma}$ est un vecteur de paramètres inconnus. La fonction de lien évite la nécessité d'exprimer les contraintes sur le paramètre $\boldsymbol{\gamma}$ afin que les inégalités (4.3) soient satisfaites. Puisque κ présente certaines analogies avec un coefficient de corrélation, on pourrait définir une fonction de lien proportionnelle à la transformation bien connue de Fisher, à savoir

$$g(\kappa_i) = arcth\,(\kappa_i) = \ln(\frac{1 + \kappa_i}{1 - \kappa_i}) \qquad i = 1, \cdots, n.$$

Malheureusement cette fonction de lien et les paramètres $\boldsymbol{\gamma}$ qui y sont associés ne sont pas aisés à interpréter en termes concrets.

C'est pour cette raison que Lipsitz et al (2001) ont préféré utiliser le lien identité

$$\kappa_i = \mathbf{z}_i'\boldsymbol{\gamma} \quad i = 1, \cdots, n \tag{4.4}$$

sans aucune contrainte sur les valeurs de κ dans le procédé d'estimation.

Pour tous les jeux de données que les auteurs précités ont expérimentés jusqu'à présent, les valeurs estimées de κ ont satisfait aux contraintes (4.3). Ceci ne signifie pas qu'il n'existe pas d'échantillon concret ou artificiel pour lequel κ_i pourrait être supérieur à un. A l'heure actuelle, le modèle (4.4) est toujours au stade expérimental et aucune démonstration de ses propriétés n'a été envisagée.

Montrons que nous pouvons estimer le vecteur de paramètres $\boldsymbol{\gamma}$ au moyen du modèle

$$p_{oi} = E[\underline{Y}_i|\mathbf{z}_i] = \mathbb{P}[\underline{Y}_i = 1|\mathbf{z}_i].$$

En utilisant les équations (4.2) et (4.4), le modèle de p_{oi} en terme de κ_i peut s'écrire

$$\begin{aligned} p_{oi} &= p_{ei} + \kappa_i(1 - p_{ei}) \\ &= p_{ei} + \mathbf{z}_i'\boldsymbol{\gamma}(1 - p_{ei}) \\ &= p_{ei} + \mathbf{z}_i^{*'}\boldsymbol{\gamma} \end{aligned} \tag{4.5}$$

où

$$\mathbf{z}_i^* = (1 - p_{ei})\mathbf{z}_i \ .$$

Si p_{ei} était connu, alors le modèle de p_{oi} serait un modèle linéaire en les paramètres avec
- une fonction de lien identité ;
- une ordonnée à l'origine connue p_{ei} ;
- un vecteur de covariables connu $\mathbf{z}_i^*$;
- un vecteur de paramètres inconnus $\boldsymbol{\gamma}$.

Par conséquent, afin d'estimer $\boldsymbol{\gamma}$, si p_{ei} est connu, on peut utiliser la méthode du maximum de vraisemblance basée sur la distribution de Bernoulli de la variable aléatoire $\underline{Y}_i$. La vraisemblance est, dans ce cas,

$$L(\kappa, \boldsymbol{\gamma}) = \prod_{i=1}^{n} p_{oi}^{y_i}(1 - p_{oi})^{1-y_i}.$$

Malheureusement, p_{ei} est rarement connu. Une des possibilités est alors d'utiliser l'estimation du maximum de vraisemblance basée sur la distribution conjointe de $(\underline{Y}_{i1}, \underline{Y}_{i2})$ pour estimer conjointement (p_{i1}, p_{i2}) et $\boldsymbol{\gamma}$, comme l'ont fait Shoukri et Mian (cfr 3.10.3).

L'approche de Lipsitz est différente, il remplace p_{ei} dans l'équation (4.5) par son estimation $\hat{p}_{ei}$ et estime alors $\boldsymbol{\gamma}$ en utilisant un modèle linéaire. En particulier, on peut estimer p_{i1} et p_{i2}, en utilisant une régression logistique avec le modèle

$$logit(p_{ir}) = \mathbf{x}_{ir}'\boldsymbol{\beta}_{1r} + \mathbf{x}_i'\boldsymbol{\beta}_{2r} \quad r = 1, 2 \tag{4.6}$$

et estimer p_{ei} comme suit

$$\hat{p}_{ei} = \hat{p}_{i1}\hat{p}_{i2} + (1 - \hat{p}_{i1})(1 - \hat{p}_{i2}). \tag{4.7}$$

Dès lors, le modèle linéaire pour p_{oi} est

$$p_{oi} \approx \hat{p}_{ei} + (1 - \hat{p}_{ei})\mathbf{z}_i'\boldsymbol{\gamma}. \tag{4.8}$$

(Vu que $\hat{p}_{ei}$ est une estimation, le signe "$\approx$" remplace le signe "=" dans l'expression (4.8.))

En considérant la valeur estimée $\hat{p}_{ei}$ dans l'expression (4.8) comme connue, $\boldsymbol{\gamma}$ peut être estimé en utilisant un modèle linéaire pour données binaires.

On procède donc comme suit

- On utilise une régression logistique de $\underline{Y}_{i1}$ sur les covariables $(\mathbf{x}_{i1}, \mathbf{x}_i)$ afin d'obtenir $\widehat{p}_{i1}$;
- On utilise une régression logistique de $\underline{Y}_{i2}$ sur les covariables $(\mathbf{x}_{i2}, \mathbf{x}_i)$ afin d'obtenir $\widehat{p}_{i2}$;
- Pour chaque sujet, on détermine alors $\hat{p}_{ei} = \hat{p}_{i1}\hat{p}_{i2} + (1 - \hat{p}_{i1})(1 - \hat{p}_{i2})$;
- On utilise une régression linéaire pour données binaires de $\underline{Y}_i$ sur les covariables $\mathbf{z}_i^*$ avec une ordonnée à l'origine connue $\hat{p}_{ei}$ (en anglais : an offset). Le processus itératif utilisé fait appel aux équations d'estimation généralisées que nous avons introduites au Chapitre 3.

Remarquons que puisque le modèle pour p_{ir} (4.6) n'est pas une fonction de $\boldsymbol{\gamma}$, pour un modèle donné de p_{ir}, l'estimation $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ sera la même, en dépit du modèle de κ_i.

Par contre, l'estimation de κ_i dépend du modèle utilisé pour estimer les paramètres p_{ir} et peut être biaisé si le modèle n'est pas adéquat.

En traitant de nombreux échantillons, Lipsitz et al (2001) ont constaté qu'il était préférable d'introduire trop de covariables dans le modèle (4.6) que trop peu, même si certaines covariables ne sont pas significatives au niveau d'incertitude α donné a priori.

Cette manière de procéder entraîne une légère augmentation de l'erreur type estimée de $\hat{\boldsymbol{\gamma}}$.

Il n'existe aucune théorie formelle ni aucune démonstration des propriétés énoncées ci-dessus.

Posons $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_{11}, \boldsymbol{\beta}'_{21}, \boldsymbol{\beta}'_{12}, \boldsymbol{\beta}'_{22})$.

En utilisant le développement en séries de Taylor comme Prentice (1988) et en supposant que les modèles de p_{i1}, p_{i2} et κ_i sont correctement spécifiés, on peut démontrer que $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')$ est un estimateur convergent en probabilité pour $(\boldsymbol{\beta}', \boldsymbol{\gamma}')$ et que $n^{1/2}((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})', (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})')$ suit une distribution asymptotiquement multinormale avec un vecteur moyen $\mathbf{0}$ et une matrice de variances-covariances pouvant être estimée par un estimateur robuste de la variance comme l'estimateur du Jackknife (Wu, 1986).

Une forme de l'estimateur du Jackknife est

$$var(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')' = \sum_{i=1}^{n}((\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i} - (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}'))'((\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i} - (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')) \tag{4.9}$$

où $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i}$ est l'estimation de $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')$ obtenue en supprimant les deux observations relatives au ième individu et recalculant $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$.

4.3 Extension à un nombre de catégories supérieur à deux.

Supposons que $\underline{Y}_{ij}$ puisse prendre les valeurs $1, \cdots, J$ à la place de 0 ou 1. Dans ce cas le coefficient κ est défini par l'expression (4.2) avec

$$p_{oi} = \sum_{j=1}^{J} \mathbb{P}[\underline{Y}_{i1} = j, \underline{Y}_{i2} = j | \mathbf{z}_i] \tag{4.10}$$

et

$$p_{ei} = \sum_{j=1}^{J} \mathbb{P}[\underline{Y}_{i1} = j | \mathbf{x}_i, \mathbf{x}_{i1}] \mathbb{P}[\underline{Y}_{i2} = j | \mathbf{x}_i, \mathbf{x}_{i1}]. \tag{4.11}$$

Si on introduit aussi une variable aléatoire $\underline{Y}_i$ qui vaut 1 si les deux observateurs sont d'accord sur le sujet i et $\underline{Y}_i = 0$ sinon, alors,

$$\begin{aligned} \mathbb{P}[\underline{Y}_i = 1 | \mathbf{z}_i] &= p_{oi} \\ &= p_{ei} + \kappa_i (1 - p_{ei}) \\ &= p_{ei} + \mathbf{z}_i^{*'} \boldsymbol{\gamma} \end{aligned} \tag{4.12}$$

pour le modèle linéaire $\kappa_i = \mathbf{z}_i' \boldsymbol{\gamma}$ où $\mathbf{z}_i^{*'} = \mathbf{z}_i (1 - p_{ei})$.

Afin d'estimer $\boldsymbol{\gamma}$, si p_{ei} est connu, on peut aussi utiliser la méthode du maximum de vraisemblance basée sur la distribution de $\underline{Y}_i$. La vraisemblance vaut

$$L(\kappa, \boldsymbol{\gamma}) = \prod_{i=1}^{n} p_{oi}^{y_i} (1 - p_{oi})^{(1-y_i)}.$$

Les estimateurs du maximum de vraisemblance des paramètres s'obtiennent en utilisant un logiciel capable de traiter les modèles linéaires généralisés et les GEE. Parmi ceux-ci, citons SAS. Cependant, p_{ei} est rarement connu. En étendant la méthode du paragraphe précédent, on peut

- Utiliser une régression logistique ordinale ou polynomiale de $\underline{Y}_{i1}$ sur les covariables $(\mathbf{x}_i, \mathbf{x}_{i1})$ afin d'obtenir $\hat{\mathbb{P}}[\underline{Y}_{i1} = j | \mathbf{x}_i, \mathbf{x}_{i1}]$;
- Utiliser une régression logistique ordinale ou polynomiale de $\underline{Y}_{i2}$ sur les covariables $(\mathbf{x}_i, \mathbf{x}_{i2})$ afin d'obtenir $\hat{\mathbb{P}}[\underline{Y}_{i2} = j | \mathbf{x}_i, \mathbf{x}_{i2}]$;
- Pour chaque sujet, déterminer $\hat{p}_{ei} = \sum_{j=1}^{J} \hat{\mathbb{P}}[\underline{Y}_{i1} = j | \mathbf{x}_i, \mathbf{x}_{i1}] \hat{\mathbb{P}}[\underline{Y}_{i2} = j | \mathbf{x}_i, \mathbf{x}_{i1}]$
- Utiliser des équations d'estimation généralisées pour les données binomiales et avec un lien identé et une ordonnée à l'origine connue pour estimer les p_{oi}, $i = 1, \cdots, n$.

4.4 Modèle de régression logistique modifiée

Lipsitz et al (article soumis à *Psychometrika* en 2001) proposent une méthode qui permet d'estimer la probabilité d'accord comme fonction de covariables. Coughlin et al (1992) avaient proposé d'utiliser une régression logistique ordinaire. Lipsitz et al proposent de modifier cette régression logistique.

La modification est motivée par le fait suivant. Supposons que deux observations binaires soient effectuées indépendamment sur un même sujet. Supposons aussi que la prévalence du signe "+" (qui est une covariable) soit grande dans certains sous-groupes et petites dans d'autres sous-groupes. Les covariables peuvent paraître significativement prédictives d'accord dans la régression logistique alors que celui-ci est uniquement dû au hasard.

Pour pallier ce problème, Lipsitz et al proposent une régression logistique modifiée qui soustrait la probabilité d'accord dû au hasard dans le modèle au moyen d'un "offset", c'est-à-dire un coefficient de régression connu. Le modèle de régression logistique modifiée de l'accord est estimé par 3 régressions logistiques ordinaires.

4.4.1 Notations et modèle

Supposons comme au paragraphe 4.2 l'existence d'un vecteur $\mathbf{z}_i$ de covariables et reprenons la variable aléatoire indicatrice d'accord $\underline{Y}_i$ décrite en (4.1). $\underline{Y}_i$ possède une distribution de Bernoulli

$$f(y_i|\mathbf{x}_i, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \boldsymbol{\beta}) = p_i^{y_i}(1-p_i)^{1-y_i}$$

où $p_i = \mathbb{P}[\underline{Y}_i = 1|\mathbf{x}_i, \mathbf{x}_{i1}, \mathbf{x}_{i2}]$.

Coughlin et al (1992) modélisent la probabilité d'accord à partir du modèle de régression logistique suivant

$$logit(p_i) = \mathbf{x}'_{i1}\boldsymbol{\beta}_1 + \mathbf{x}'_{i2}\boldsymbol{\beta}_2 + \mathbf{x}'_i\boldsymbol{\beta}_3 \tag{4.13}$$

où $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \boldsymbol{\beta}'_3)$ est un vecteur de paramètres inconnus.

Lipsitz et al ont développé un modèle de régression logistique pour lequel $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ et $\boldsymbol{\beta}_3$ sont égaux à $\mathbf{0}$ lorsque l'accord est uniquement dû au hasard, ce qui n'est pas nécessairement le cas dans le modèle (4.13).

Nous savons que lorsque les observateurs opèrent indépendamment,

$$p_i = p_{i1}p_{i2} + (1-p_{i1})(1-p_{i2}). \tag{4.14}$$

Le modèle logistique modifié proposé par Lipsitz et al est le suivant

$$logit(p_i) = \eta_i + \mathbf{x}'_{i1}\boldsymbol{\beta}_1 + \mathbf{x}'_{i2}\boldsymbol{\beta}_2 + \mathbf{x}'_i\boldsymbol{\beta}_3 \tag{4.15}$$

où

$$\eta_i = logit[p_{i1}p_{i2} + (1 - p_{i1})(1 - p_{i2})].$$

Dans le modèle (4.15), si les probabilités marginales p_{ir} $(i = 1, \cdots, n,\ r = 1, 2)$ sont connues, le terme η_i est appelé un "offset". Les probabilités marginales sont estimées par le modèle (4.6).

Remarquons que le modèle (4.15) est un modèle corrigé de l'accord dû uniquement au hasard. En effet, si l'accord est uniquement dû au hasard, alors

$$logit(p_i) = \eta_i. \tag{4.16}$$

La procédure proposée afin d'estimer le modèle (4.15) est

- Utiliser une régression logistique de $\underline{Y}_{i1}$ sur les covariables $(\mathbf{x}_i, \mathbf{x}_{i1})$ afin d'estimer p_{i1} ;
- Utiliser une régression logistique de $\underline{Y}_{i2}$ sur les covariables $(\mathbf{x}_i, \mathbf{x}_{i2})$ afin d'estimer p_{i2} ;
- Estimer "l'offset" η_i par $\hat{\eta}_i = logit[\hat{p}_{i1}\hat{p}_{i2} + (1 - \hat{p}_{i1})(1 - \hat{p}_{i1})]$;
- Utiliser une régression logistique de $\underline{Y}_i$ sur les covariables $(\mathbf{x}_i, \mathbf{x}_{i1}, \mathbf{x}_{i2})$ et une ordonnée à l'origine connue $\hat{\eta}_i$ afin de déterminer $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3)$ au moyen du modèle (4.15).

Remarquons que la procédure ne fait pas intervenir de régression linéaire contrairement à la méthode décrite au paragraphe 4.2. Par conséquent, le problème posé par les contraintes sur $\hat{p}_{ei}$ $(\hat{p}_{ei} \in [0, 1])$ est résolu.

La variance de $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3)$ peut être estimée par l'estimateur du Jackknife.

4.4.2 Comparaison entre accord obtenu et accord dû uniquement au hasard

Pour chaque individu, Lipsitz et al proposent l'expression suivante, afin de mesurer la différence entre l'accord obtenu et l'accord dû au hasard :

$$\hat{\xi} = logit(\hat{p}_i) - \hat{\eta}_i = \mathbf{x}'_{i1}\hat{\boldsymbol{\beta}}_1 + \mathbf{x}'_{i2}\hat{\boldsymbol{\beta}}_2 + \mathbf{x}'_i\hat{\boldsymbol{\beta}}_3. \tag{4.17}$$

Dès lors, on peut éprouver l'hypothèse H_0 : l'accord est uniquement dû au hasard c'est-à-dire $logit(p_i) - \eta_i = 0$ vis-à-vis de l'alternative $H_1 : logit(p_i) - \eta_i \neq 0$.

Pour une fonction monotone bornée $g(.)$, Lipsitz et al proposent d'utiliser la mesure

$$\xi_i* = \frac{g[logit(p_i)] - g[\eta_i]}{\{\max g[logit(p_i)]\} - g[\eta_i]} \tag{4.18}$$

afin d'éprouver l'hypothèse H_0. Cette mesure (4.18) vaut 0 si l'accord est dû au hasard et vaut 1 si l'accord est parfait.

En particulier, pour une valeur donnée a, en choisissant

$$g[a] = \frac{e^a}{1+e^a}$$

si on évalue l'expression (4.18) en $(\hat{p}_i, \hat{p}_{i1}, \hat{p}_{i2})$, on obtient un estimateur du coefficient Kappa de Cohen pour chaque individu

$$\begin{aligned} \hat{\kappa}_i &= \frac{\hat{p}_i - \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}}}{1 - \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}}} \\ &= \frac{\hat{p}_i - [\hat{p}_{i1}\hat{p}_{i2} + (1-\hat{p}_{i1})(1-\hat{p}_{i2})]}{1 - [\hat{p}_{i1}\hat{p}_{i2} + (1-\hat{p}_{i1})(1-\hat{p}_{i2})]}. \end{aligned} \tag{4.19}$$

Ainsi, si $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ et $\boldsymbol{\beta}_3$ sont égaux à $\mathbf{0}$, $p_i = p_{i1}p_{i2} + (1-p_{i1})(1-p_{i2})$ et $\kappa_i = 0$.

4.5 Exemple

On s'intéresse ici à la similarité du plus haut niveau d'enseignement reçu par un homme et son épouse. Le niveau d'enseignement est une variable ordinale à 3 catégories : pas d'études supérieures (below HS), études supérieures (HS) ou collège (COL). Le "sujet" dans cette étude est le couple mari-femme. L'épouse est considérée comme un observateur et le mari comme un autre. Les données proviennent de l'étude sociale générale de 1991 (Smith, 1996) constituée d'interviews personnels menés sur 332 couples par le centre de recherche de l'opinion nationale de Chicago. Le but de cette analyse était de déterminer si l'accord entre le plus haut niveau d'enseignement reçu par le mari et son épouse pouvait être prédit par deux covariables relatives aux sujets (c'est-à-dire à l'environnement familial) et trois covariables relatives aux observateurs. Les covariables relatives aux sujets sont les revenus de la famille (moins de 25000$, entre 25000 et 50000$ et plus de 50000$) et le nombre d'enfants dans le ménage (0, 1, 2, 3 ou 4). Les covariables relatives aux observateurs sont l'âge auquel se sont mariés les époux, le plus haut niveau d'enseignement du père de l'épouse et du père du mari, et si le mari et l'épouse vivaient encore à la maison avec leurs parents à l'âge de 16 ans (oui ou non).

Dans le cadre de cet exemple, il est raisonnable d'espérer un accord plus grand pour des époux dont les pères ont le même niveau d'enseignement, pour un couple qui s'est marié alors que l'épouse et le mari avaient le même âge et pour un couple dont les deux membres vivaient ou ne vivaient plus à la maison à l'âge de 16 ans.

La table 4.1 donne des estimations de κ pour différents niveaux des covariables, non ajustées pour les autres covariables.

On peut remarquer que c'est pour les différents niveaux d'instruction du père que les valeurs de κ varient le plus entre elles.

Une disquette où sont reprises les données détaillées relatives à l'exemple et une macro en langage SAS qui permet d'effectuer les calculs est disponible pour celui qui le souhaite. Le nombre de données multivariées sans valeur manquante est égal à 332.

Puisque les variables indépendantes sont ordinales, il convient de modéliser les probabilités marginales $p_{ir,j} = \mathbb{P}[\underline{Y}_{ir} = j|\mathbf{x}_i, \mathbf{x}_{ir}]$ à partir de la fonction lien logistique cumulative (3.41).

TABLE 4.1 – Estimation du coefficient Kappa de Cohen à partir des données fournies par l'étude de Smith (1996)

Covariables	Nombre	% d'accord	κ	Erreur type	p-value
Revenus de la famille					
$< 25000\$$	50	64	0.420	0.110	0.325
$25000\$ - 50000\$$	127	61	0.312	0.071	
$> 50000\$$	157	72	0.458	0.069	
Education du père (épouse, mari)*					
(below HS, below HS)	80	70	0.518	0.081	0.002
(below HS, HS)	31	48	0.192	0.134	
(below HS, COL)	8	44	-0.231	0.192	
(HS, below HS)	36	72	0.354	0.144	
(HS, HS)	80	64	0.270	0.107	
(HS, COL)	27	52	-0.073	0.171	
(COL, below HS)	15	73	0.492	0.199	
(COL, HS)	28	79	0.569	0.143	
(COL, COL)	29	83	0.356	0.220	
*Age au mariage** (épouse, mari)*					
$(\leq 22, \leq 22)$	129	70	0.476	0.070	0.503
$(\leq 22, > 22)$	80	64	0.400	0.090	
$(> 22, \leq 22)$	9	56	0.053	0.296	
$(> 22, > 22)$	116	66	0.387	0.074	
Vivre chez les parents à 16 ans (épouse, mari)					
(non, non)	6	50	-0.200	0.150	0.299
(non, oui)	29	59	0.374	0.132	
(oui, non)	27	67	0.449	0.148	
(oui, oui)	270	68	0.441	0.049	
Nombre d'enfants					
0-1	244	66	0.433	0.051	0.872
2-4	90	69	0.417	0.085	

*HS : école supérieure, COL : collège ;

**22 ans est l'âge médian ;

la p-value est relative au test de Wald

Le modèle est le suivant :

$$\begin{aligned}\ln(\frac{\gamma_{ir,j}}{1-\gamma_{ir,j}}) &= \beta_{0rj} + \beta_{1r}KIDS_i + \beta_{2r}INCOME2_i + \beta_{3r}INCOME3_i \\ &+ \beta_{4r}FATHED2_{ir} + \beta_{5r}FATHED3_{ir} + \beta_{6r}AGE_{ir} + \beta_{7r}HOME_{ir} \\ &+ \beta_{8r}INCOME2_i \times FATHED2_{ir} + \beta_{9r}INCOME2_i \times FATHED3_{ir} \\ &+ \beta_{10r}INCOME3_i \times FATHED2_{ir} + \beta_{11r}INCOME3_i \times FATHED3_{ir} \\ &+ \beta_{12r}INCOME2_i \times AGE_{ir} + \beta_{13r}INCOME3_i \times AGE_{ir} \qquad (4.20)\end{aligned}$$

pour $j = 1, 2$ et $r = 1, 2$, où

$KIDS_i$ représente le nombre d'enfants dans le ménage ;

$INCOME2_i = 1$ si le revenu est compris entre 25000\$ et 50000\$, $INCOME2_i = 0$ sinon ;

$INCOME3_i = 1$ si le revenu est supérieur à 50000\$, $INCOME3_i = 0$ sinon ;

$FATHED2_{ir} = 1$ si le plus haut niveau d'étude du père est l'école supérieure, $FATHED2_{ir} = 0$ sinon ;

$FATHED3_{ir} = 1$ si le plus haut niveau d'étude du père est le collège, $FATHED2_{ir} = 0$ sinon ;

AGE_{ir} représente l'âge lors du mariage ;

$HOME_{ir}$ traduit le fait que le mari ou l'épouse vivaient encore ou non chez leurs parents à l'âge de 16 ans.

Plusieurs interactions sont également considérées dans le modèle, à savoir,
$INCOME2_i \times FATHED2_{ir}$;
$INCOME2_i \times FATHED3_{ir}$;
$INCOME3_i \times FATHED2_{ir}$;
$INCOME3_i \times FATHED3_{ir}$;
$INCOME2_i \times AGE_{ir}$;
$INCOME3_i \times AGE_{ir}$.

Les estimations des paramètres du modèle logistique cumulatif (4.20) sont données dans la table 4.2.

TABLE 4.2 – Estimations du modèle logistique cumulatif des marginales sur base des données de l'étude conjugale

Paramètres	Estimation	Erreur type	Z	p-value
Niveau d'étude de l'épouse				
INTERCEPT1	-0.347	1.202	-0.29	0.773
INTERCEPT2	3.489	1.235	2.83	0.005
KIDS	0.100	0.096	1.04	0.299
INCOME2	-1.136	1.402	-0.81	0.418
INCOME3	0.845	1.559	0.54	0.588
FATHED2	-2.079	0.700	-2.97	0.003
FATHED3	-3.990	0.903	-4.42	< 0.0001
AGE AT MAR	0.038	0.053	0.71	0.480
HOME AT 16	-0.767	0.392	-1.96	0.050
INCOME2*FATHED2	1.218	0.810	1.50	0.133
INCOME2*FATHED3	1.760	1.023	1.72	0.085
INCOME3*FATHED2	1.741	0.814	2.14	0.032
INCOME3*FATHED3	2.191	1.043	2.10	0.036
INCOME2*AGE	-0.021	0.062	-0.34	0.737
INCOME3*AGE	-0.193	0.069	-2.81	0.005
Niveau d'étude du mari				
INTERCEPT1	0.076	1.156	0.07	0.948
INTERCEPT2	2.981	1.175	2.54	0.011
KIDS	0.112	0.090	1.24	0.214
INCOME2	-0.978	1.493	-0.66	0.512
INCOME3	0.378	1.578	0.24	0.811
FATHED2	-2.006	0.622	-3.22	0.001
FATHED3	-3.621	0.955	-3.79	< 0.0001
AGE AT MAR	0.001	0.044	0.02	0.988
HOME AT 16	-0.355	0.373	-0.95	0.342
INCOME2*FATHED2	1.245	0.718	1.73	0.083
INCOME2*FATHED3	0.586	1.241	0.47	0.637
INCOME3*FATHED2	1.941	0.727	2.67	0.008
INCOME3*FATHED3	1.854	1.095	1.69	0.090
INCOME2*AGE	-0.028	0.060	-0.46	0.646
INCOME3*AGE	-0.126	0.065	-1.95	0.051

Pour la régression logistique portant sur les épouses, on a

$D(M_p) = 478.510$;
$D(M_p|M_0) = 98.592$ à 13 degrés de liberté.
La p-value est 0.0001.

Pour la régression logistique relative aux maris, on a

$D(M_p) = 530.497$;
$D(M_p|M_0) = 98.896$ à 13 degrés de liberté.
La p-value est 0.0001.

Parmi les coefficients de régression repris dans la table 4.2, certains ne sont pas significatifs au niveau d'incertitude $\alpha = 5\%$. Le modèle a été surestimé. La variable la plus importante est FATHED3.

Voici le modèle pour le coefficient κ :

$$\begin{aligned} \kappa_i &= \gamma_0 + \gamma_1 KIDS_i + \gamma_2 INCOME2_i + \gamma_3 INCOME3_i + \gamma_4 FATHED_i \\ &+ \gamma_5 AGE_i + \gamma_6 HOME_i \end{aligned} \tag{4.21}$$

où

$FATHED_i = 0$ si le niveau d'étude du père du mari et du père de l'épouse est le même et $FATHED_i = 1$ sinon ;

$AGE_i = 0$ si l'âge au mariage du mari et de l'épouse est le même, $AGE_i = 1$ sinon ;

$HOME_i = 0$ si le mari et l'épouse ont la même valeur pour $HOME_{ir}$, $HOME_i = 1$ sinon ;

La table 4.3 donne les estimations de $\boldsymbol{\gamma}$ pour le modèle (4.21) en estimant les erreurs types des paramètres par la méthode du Jackknife.

Les résultats de la table 4.3 sont en accord avec les résultats univariés présentés dans la table 4.1, seul le paramètre correspondant au niveau d'enseignement du père est significativement différent de 0 au niveau d'incertitude $\alpha = 5\%$. Le modèle a été surestimé.

TABLE 4.3 – Estimations des paramètres de la régression

Paramètres	Modèle	Estimation	Erreur type	Z	p-value
INTERCEPT	(4.20)	0.364	0.129	2.83	0.005
KIDS	(4.20)	0.025	0.038	0.65	0.516
INCOME2	(4.20)	-0.112	0.147	-0.76	0.447
INCOME3	(4.20)	0.016	0.145	0.11	0.912
FATHED	(4.20)	-0.140	0.057	-2.47	0.014
AGE	(4.20)	0.005	0.010	0.55	0.582
HOME	(4.20)	-0.037	0.116	-0.32	0.749

4.6 Discussion

Pour des observations dichotomiques, les deux régressions logistiques et la régression linéaire pour données binaires déterminent totalement la distribution conjointe multinomiale des deux observations pour le même sujet. Cette méthode présente un avantage par rapport à l'approche de Soukri et Mian décrite au Chapitre 3 : l'écriture du programme informatique nécessaire à l'obtention des estimations est plus simple.

De plus, l'estimation du coefficient κ pour des observations polynomiales par la méthode du maximum de vraisemblance est plus difficile à étendre car la distribution conjointe des deux observations doit être spécifiée et contient des paramètres de nuisance non désirés. Dans l'approche de Lipsitz et al, le modèle de régression de κ peut être estimé par deux régressions logistiques ordinales ou polynomiales et une régression linéaire pour données binaires et ne nécessite pas la spécification de paramètres de nuisance indispensables dans la méthode d'estimation du maximum de vraisemblance.

Il faudrait spécifier la distribution conjointe multinomiale de $(\underline{Y}_{i1}, \underline{Y}_{i2})$ pour pouvoir utiliser la méthode du maximum de vraisemblance. Cette distribution possède $J^2 - 1$ probabilités non redondantes.

La méthode de Lipsitz ne nécessite qu'un modèle pour les $2(J-1)$ probabilités non redondantes $\mathbb{P}[\underline{Y}_{il} = j|\mathbf{x}_i, \mathbf{x}_{il}]$ $(j = 1, \cdots, J-1,\ \ l = 1, 2)$ et un modèle pour κ_i. Ce modèle ne spécifie donc que $2(J-1)+1$ dimensions des distributions conjointes de $(\underline{Y}_{i1}, \underline{Y}_{i2})$ $(J^2-1)-$dimensionnelles. Donc, nécessite moins de paramètres de nuisance.

De plus, la distribution conjointe nécessaire pour la méthode du maximum de vraisemblance n'a pas une forme simple.

Remarquons que l'article soumis à *Psychometrika* n'a pas encore été publié. Le dernier *Psychometrika* paru à ce jour est l'édition de mars 2002.

Chapitre 5

Résultats bruts de l'exemple sur les époux

```
The SAS System                    17:27 Tuesday, December 18, 2001 1

The LOGISTIC Procedure

Data Set: WORK.ALLZ
Response Variable: HUS_ED
Response Levels: 3
Number of Observations: 332
Link Function: Logit


     Response Profile

Ordered
  Value  HUS_ED     Count

      1       1        33
      2       2       147
      3       3       152



Score Test for the Proportional Odds Assumption

Chi-Square = 15.1004 with 13 DF (p=0.3011)
```

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	633.393	560.497	.
SC	641.004	617.574	.
-2 LOG L	629.393	530.497	98.896 with 13DF(p=0.0001)
Score	.	.	84.472 with 13DF(p=0.0001)

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCP1	1	0.0755	1.1557	0.0043	0.9479	.
INTERCP2	1	2.9806	1.1749	6.4354	0.0112	.
KIDS	1	0.1124	0.0904	1.5452	0.2138	0.080104
_25_50	1	-0.9779	1.4926	0.4293	0.5123	-0.261627
_50	1	0.3779	1.5776	0.0574	0.8107	0.104165
H_PAED2	1	-2.0059	0.6221	10.3954	0.0013	-0.545854
H_PAED3	1	-3.6211	0.9545	14.3911	0.0001	-0.788735
HUS_AGE	1	0.000683	0.0439	0.0002	0.9876	0.001808
HUS_MF16	1	-0.3545	0.3734	0.9013	0.3424	-0.058561
_2550HP2	1	1.2445	0.7175	3.0082	0.0828	0.241661
_2550HP3	1	0.5856	1.2409	0.2227	0.6370	0.067155
_50HP2	1	1.9411	0.7272	7.1251	0.0076	0.437186
_50HP3	1	1.8536	1.0945	2.8684	0.0903	0.340232
_25_50HA	1	-0.0275	0.0598	0.2108	0.6462	-0.182749
_50HA	1	-0.1258	0.0646	3.7937	0.0514	-0.877075

Analysis of Maximum Likelihood Estimates

Variable	Odds Ratio
INTERCP1	.
INTERCP2	.
KIDS	1.119
_25_50	0.376
_50	1.459
H_PAED2	0.135
H_PAED3	0.027
HUS_AGE	1.001
HUS_MF16	0.702
_2550HP2	3.471
_2550HP3	1.796
_50HP2	6.966
_50HP3	6.383
_25_50HA	0.973
_50HA	0.882

Association of Predicted Probabilities and Observed Responses

Concordant = 75.6%	Somers' D = 0.525
Discordant = 23.1%	Gamma = 0.532
Tied = 1.3%	Tau-a = 0.308
(32211 pairs)	c = 0.762

The SAS System 17:27 Tuesday, December 18, 2001 3

The LOGISTIC Procedure

Data Set: WORK.PREDZ
Response Variable: WIF_ED
Response Levels: 3
Number of Observations: 332
Link Function: Logit

Response Profile

Ordered Value	WIF_ED	Count
1	1	30
2	2	177
3	3	125

Score Test for the Proportional Odds Assumption

Chi-Square = 21.5751 with 13 DF (p=0.0623)

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	615.102	508.510	.
SC	622.713	565.587	.
-2 LOG L	611.102	478.510	132.592 with 13DF(p=0.0001)
Score	.	.	108.128 with 13DF(p=0.0001)

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCP1	1	-0.3471	1.2019	0.0834	0.7728	.
INTERCP2	1	3.4885	1.2346	7.9841	0.0047	.
KIDS	1	0.0996	0.0959	1.0771	0.2993	0.070976
_25_50	1	-1.1358	1.4017	0.6566	0.4178	-0.303859
_50	1	0.8454	1.5585	0.2943	0.5875	0.233064

W_PAED2	1	-2.0790	0.6997	8.8293	0.0030	-0.567938
W_PAED3	1	-3.9900	0.9030	19.5232	0.0001	-0.907929
WIF_AGE	1	0.0376	0.0532	0.5000	0.4795	0.110434
WIF_MF16	1	-0.7674	0.3922	3.8276	0.0504	-0.130125
_2550WP2	1	1.2175	0.8102	2.2581	0.1329	0.238443
_2550WP3	1	1.7596	1.0228	2.9598	0.0854	0.241681
_50WP2	1	1.7410	0.8135	4.5801	0.0323	0.405742
_50WP3	1	2.1907	1.0431	4.4108	0.0357	0.406161
_25_50WA	1	-0.0208	0.0620	0.1125	0.7374	-0.132137
_50WA	1	-0.1925	0.0685	7.9043	0.0049	-1.216711

Analysis of Maximum Likelihood Estimates

Variable	Odds Ratio
INTERCP1	.
INTERCP2	.
KIDS	1.105
_25_50	0.321
_50	2.329
W_PAED2	0.125
W_PAED3	0.018
WIF_AGE	1.038
WIF_MF16	0.464
_2550WP2	3.379
_2550WP3	5.810
_50WP2	5.703
_50WP3	8.942
_25_50WA	0.979
_50WA	0.825

Association of Predicted Probabilities and Observed Responses

Concordant = 79.7%	Somers' D = 0.609
Discordant = 18.8%	Gamma = 0.618
Tied = 1.4%	Tau-a = 0.345
(31185 pairs)	c = 0.804

Univariate Procedure

Variable=KAPPA

Moments

N	332	Sum Wgts	332
Mean	0.37016	Sum	122.893
Std Dev	0.135349	Variance	0.018319
Skewness	0.001183	Kurtosis	0.359507
USS	51.55368	CSS	6.063672
CV	36.56494	Std Mean	0.007428
T:Mean=0	49.83152	Pr>|T|	0.0001
Num ^= 0	332	Num > 0	331
M(Sign)	165	Pr>=|M|	0.0001
Sgn Rank	27634	Pr>=|S|	0.0001

Quantiles(Def=5)

100% Max	0.746398	99%	0.723436
75% Q3	0.452188	95%	0.589827
50% Med	0.380228	90%	0.541321
25% Q1	0.27303	10%	0.224522
0% Min	-0.04881	5%	0.127509
		1%	0.036484
Range	0.795212		
Q3-Q1	0.179158		
Mode	0.262592		

Extremes

Lowest	Obs	Highest	Obs
-0.04881(	190)	0.7234(	80)
0.008496(	88)	0.723436(	268)
0.010471(	143)	0.723972(	60)
0.036484(	239)	0.744994(	98)
0.036824(	148)	0.746398(	182)

Conclusion générale

Ce mémoire constitue un essai de synthèse des travaux relatifs au coefficient Kappa de Cohen qui permet de mesurer l'accord entre observateurs. Plusieurs aspects ont été abordés :

- accord entre deux observateurs dans le cas d'un critère binaire (cas le plus simple) ;
- accord entre plusieurs observateurs dans le cas d'un critère à plusieurs modalités (cas le plus complexe) ;
- variabilité d'échantillonnage du Kappa estimé et tests d'hypothèses ;
- modélisation du coefficient Kappa en fonction de cofacteurs liés aux observateurs et aux sujets.

Durant près de 40 ans, on a eu recours au coefficient Kappa de Cohen chaque fois qu'il s'avérait nécessaire de mesurer le degré d'accord entre juges, experts, évaluateurs ou observateurs. La méthode fut aussi largement utilisée pour comparer deux échelles d'évaluation (par exemple, classer n individus dans k catégories sur base de deux échelles différentes).

Le calcul du coefficient Kappa de Cohen et de son erreur type est aujourd'hui réalisé par la plupart des logiciels statistiques (SAS, S-Plus, etc). Ces logiciels cependant ne permettent pas encore d'utiliser les méthodes de régression proposées par Shoukri et Mian (1996) d'une part, par Lipsitz et al (2001) d'autre part.

Au terme de ce travail, il apparaît que des progrès importants ont été accomplis récemment dans le domaine du Kappa de Cohen. Ceux-ci ont été rendus possibles par la théorie des modèles linéaires généralisés. Des problèmes théoriques restent cependant non résolus. Enfin, les nouvelles méthodes ne seront utilisables à grande échelle que lorsque des logiciels validés auront été développés et diffusés.

Annexe A

Erreur type des coefficients λ

A.1 Erreur type des coefficients lambda asymétriques

Afin de déterminer l'erreur type de $\hat{\lambda}_{C|R}$, exprimons $\sqrt{n}(\hat{\lambda}_{C|R} - \lambda_{C|R})$ comme suit :

$$\begin{aligned}\sqrt{n}(\hat{\lambda}_{C|R} - \lambda_{C|R}) &= \sqrt{n}\left(\frac{\sum_{i=1}^{I}\hat{p}_{im} - \hat{p}_{.m}}{1-\hat{p}_{.m}} - \frac{\sum_{i=1}^{I}p_{im} - p_{.m}}{1-p_{.m}}\right) \\ &= \sqrt{n}\frac{\left[\sum_{i=1}^{I}(\hat{p}_{im} - p_{im})\right](1-p_{.m}) - (\hat{p}_{.m} - p_{.m})(1-\sum_{i=1}^{I}p_{im})}{(1-p_{.m})^2} \\ &\quad \cdot \quad \frac{1-p_{.m}}{1-\hat{p}_{.m}} \qquad (A.1)\end{aligned}$$

Le facteur $\alpha = \dfrac{1-p_{.m}}{1-\hat{p}_{.m}}$ converge en probabilité vers 1.

La distribution asymptotique de (A.1) est la même avec ou sans le facteur α.

En effet, Mann et Wald (1943) ont démontré que

théorème de convergence 1 *Si* $\{\underline{X}_n\}$ *et* $\{\underline{Y}_n\}$ *sont deux séries de variables aléatoires, et si* $\underline{X}$ *est une variable aléatoire et y une constante telles que* $\underline{X}_n$ *converge en distribution vers la distribution de* $\underline{X}$ *et* $\underline{Y}_n$ *converge en probabilité vers y,*
alors,
$\underline{X}_n + \underline{Y}_n$ *converge en distribution vers la distribution de* $\underline{X} + y$,
$\underline{X}_n\underline{Y}_n$ *converge en distribution vers la distribution de* $\underline{X}y$,
si, en plus, $y \neq 0$, $\underline{X}_n/\underline{Y}_n$ *converge en distribution vers la distribution de* $\underline{X}/y$.

Puisque les p_{im} et les $p_{.m}$ sont constants, l'expression (A.1) est une combinaison linéaire des $\hat{p}_{im}$ et des $\hat{p}_{.m}$.

Supposons que $\hat{p}_{im} = \hat{p}_{ij}$ pour la valeur de j tq $p_{im} = p_{ij}$, de même pour $\hat{p}_{.m}$.

Notons Δ le numérateur de $((A.1)/\alpha\sqrt{n})$.

Montrons que $\sqrt{n}\Delta$ a une distribution asymptotiquement normale de moyenne nulle et de variance

$$\begin{aligned} var(\sqrt{n}\Delta) &= (1-p_{.m})^2(\sum_{i=1}^{I} p_{im})(1-\sum_{i=1}^{I} p_{im}) \\ &+ (1-\sum_{i=1}^{I} p_{im})^2 p_{.m}(1-p_{.m}) \\ &- 2(1-p_{.m})\left(1-\sum_{i=1}^{I} p_{im}\right)\left[\sum^{\star} p_{im} - p_{.m}\sum_{i=1}^{I} p_{im}\right] \end{aligned} \tag{A.2}$$

où $\sum^{\star} p_{im}$ désigne la somme des p_{im} qui apparaissent dans la colonne correspondant à $p_{.m}$.

Le numérateur $\sqrt{n}\Delta$ est composé de deux termes :

$A = \sqrt{n}\sum_{i=1}^{I}(\hat{p}_{im} - p_{im})(1-p_{.m})$;

$B = \sqrt{n}(\hat{p}_{.m} - p_{.m})(1-\sum_{i=1}^{I} p_{im})$.

De plus, $var(\sqrt{n}\Delta) = var(A) + var(B) - 2cov(A,B)$.
Ainsi,

$$var(A) = var\left[\sqrt{n}\left(\sum_{i=1}^{I}(\hat{p}_{im} - p_{im})\right)(1-p_{.m})\right] = (1-p_{.m})^2 var\left(\sqrt{n}\sum_{i=1}^{I}(\hat{p}_{im} - p_{im})\right)$$

Comme $\sum_{i=1}^{I}\hat{p}_{im}$ et $(1-\sum_{i=1}^{I}\hat{p}_{im})$ peuvent être considérés, respectivement, comme étant les proportions de succès et d'échec de n expériences indépendantes de Bernouilli avec des probabilités constantes $\sum_{i=1}^{I} p_{im}$ et $(1-\sum_{i=1}^{I} p_{im})$,

$$var(\sqrt{n}\sum_{i=1}^{I}(\hat{p}_{im} - p_{im})) = (\sum_{i=1}^{I} p_{im})(1-\sum_{i=1}^{I} p_{im}) \tag{A.3}$$

et

$$var(\sqrt{n}\sum_{i=1}^{I}\hat{p}_{im}) = (\sum_{i=1}^{I} p_{im})(1-\sum_{i=1}^{I} p_{im}) \tag{A.4}$$

Le terme var(A) constitue la première ligne de (A.2). La deuxième ligne de cette expression (var(B)) s'obtient de manière analogue.

$$cov(A,B) = n(1-p_{.m})(1-\sum_{i=1}^{I} p_{im})cov((\sum_{i=1}^{I}(\hat{p}_{im}-p_{im}),(\hat{p}_{.m}-p_{.m}))$$

Le calcul de cette covariance n'est pas aisé mais, en général, si $\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_k$ sont des proportions multinomiales correspondant aux probabilités $p_1, p_2, \cdots, p_k$, alors

$$\begin{aligned} cov(\hat{p}_1+\hat{p}_2, \hat{p}_1+\hat{p}_3) &= var(\hat{p}_1) + cov(\hat{p}_1,\hat{p}_2) + cov(\hat{p}_1,\hat{p}_3) + cov(\hat{p}_2,\hat{p}_3) \\ &= n^{-1}\left[p_1(1-p_1) - p_1p_2 - p_1p_3 - p_2p_3\right] \\ &= n^{-1}\left[p_1 - (p_1+p_2)(p_1+p_3)\right] \end{aligned} \tag{A.5}$$

Afin d'appliquer le résultat (A.5), prenons

$\hat{p}_1$ la somme des $\hat{p}_{im}$ qui apparaissent dans la colonne correspondant à $\hat{p}_{.m}$, notée $\sum^{\star} p_{im}$;
$\hat{p}_2 = \sum_{i=1}^{I} \hat{p}_{im} - \hat{p}_1$;
$\hat{p}_3 = p_{.m} - \hat{p}_{1.}$.

Nous pouvons alors directement écrire la troisième ligne de (A.5). Dès lors,

$$2n\ cov((\sum_{i=1}^{I}(\hat{p}_{im}-p_{im}),(\hat{p}_{.m}-p_{.m})) = 2(\sum^{\star} p_{im} - p_{.m}\sum_{i=1}^{I} p_{im})$$

qui est la troisième ligne de l'expression (A.2).
On obtient facilement, après simplifications de l'expression (A.2),

$$var(\sqrt{n}\Delta) = (1-p_{.m})(1-\sum_{i=1}^{I} p_{.m})(\sum_{i=1}^{I} p_{im} + p_{.m} - 2\sum^{\star} p_{im}) \tag{A.6}$$

Ainsi,

$$(s.e.(\hat{\lambda}_{C|R}))^2 = \frac{1}{n(1-p_{.m})^3}(1-\sum_{i=1}^{I} p_{.m})(\sum_{i=1}^{I} p_{im} + p_{.m} - 2\sum^{\star} p_{im}) \tag{A.7}$$

On obtient l'erreur type de $\hat{\lambda_{R|C}}$ de manière analogue.

Remarquons que la méthode *Delta*, qui ne sera pas appliquée ici, donne les mêmes résultats.

A.2 Variance du coefficient λ

De la même manière que dans le paragraphe précédent, nous pouvons écrire

$$\begin{aligned}\sqrt{n}(\hat{\lambda}-\lambda) &= \frac{\sqrt{n}}{(2-p_{.m}-p_{m.})^2}\left[(2-p_{.m}-p_{m.})(\sum_{i=1}^{I}\hat{p}_{im}+\sum_{j=1}^{J}\hat{p}_{mj})\right.\\ &+ (\sum_{i=1}^{I}p_{im}+\sum_{j=1}^{J}p_{mj}-2)(\hat{p}_{.m}+\hat{p}_{m.})\\ &+ \left.2(p_{.m}+p_{m.}-\sum_{i=1}^{I}p_{im}-\sum_{j=1}^{J}p_{mj})\right].\frac{2-p_{.m}-p_{m.}}{2-\hat{p}_{.m}-\hat{p}_{m.}} \qquad (A.8)\end{aligned}$$

Posons

$$\begin{aligned}\sqrt{n}\Delta &= \sqrt{n}\left[(2-p_{.m}-p_{m.})(\sum_{i=1}^{I}\hat{p}_{im}+\sum_{j=1}^{J}\hat{p}_{mj})\right.\\ &+ (\sum_{i=1}^{I}p_{im}+\sum_{j=1}^{J}p_{mj}-2)(\hat{p}_{.m}+\hat{p}_{m.})\\ &+ \left.2(p_{.m}+p_{m.}-\sum_{i=1}^{I}p_{im}-\sum_{j=1}^{J}p_{mj})\right] \qquad (A.9)\end{aligned}$$

et

$A=\sum_{i=1}^{I}\hat{p}_{im}+\sum_{j=1}^{J}\hat{p}_{mj}$;
$B=\hat{p}_{.m}+\hat{p}_{m.}$.

Dès lors,

$$\begin{aligned}var(\sqrt{n}\Delta) &= (2-p_{.m}-p_{m.})^2 var(A)+(\sum_{i=1}^{I}p_{im}+\sum_{j=1}^{J}p_{mj}-2)^2 var(B)\\ &+ 2(2-p_{.m}-p_{m.})(\sum_{i=1}^{I}p_{im}+\sum_{j=1}^{J}p_{mj}-2)cov(A,B) \qquad (A.10)\end{aligned}$$

On a, vu (A.4),

$$var(A)=(\sum_{i=1}^{I}p_{im})(1-\sum_{i=1}^{I}p_{im})+(\sum_{j=1}^{J}p_{mj})(1-\sum_{j=1}^{J}p_{mj})+2cov(\sum_{i=1}^{I}\hat{p}_{im},\sum_{j=1}^{J}\hat{p}_{mj})$$

Or, en utilisant (A.5) avec

$\hat{p}_1 = \sum^{\star} \hat{p}_{im}$;
$\hat{p}_2 = \sum_{i=1}^{I} \hat{p}_{im} - \hat{p}_1$;
$\hat{p}_3 = \sum_{j=1}^{J} \hat{p}_{mj} - \hat{p}_1$;

on obtient

$$2cov(\sum_{i=1}^{I} \hat{p}_{im}, \sum_{j=1}^{J} \hat{p}_{mj}) = 2(\sum^{\star} p_{im} - (\sum_{i=1}^{I} p_{im})(\sum_{j=1}^{J} p_{mj}))$$

Ainsi,

$$var(A) = (\sum_{i=1}^{I} p_{im})(1 - \sum_{i=1}^{I} p_{im}) + (\sum_{j=1}^{J} p_{mj})(1 - \sum_{j=1}^{J} p_{mj}) + 2(\sum^{\star} p_{im} - (\sum_{i=1}^{I} p_{im})(\sum_{j=1}^{J} p_{mj}))$$

Analoguement,

$$var(B) = (p_{.m})(1 - p_{.m}) + (p_{m.})(1 - p_{m.}) + 2(p_{\star\star} - p_{.m}p_{m.})$$

où $p_{\star\star}$ désigne le p_{ij} qui apparaît dans la ligne pour laquelle $p_{i.}$ est maximal et dans la colonne pour laquelle $p_{.j}$ est maximal.

$$\begin{aligned} cov(A, B) &= cov(\sum_{i=1}^{I} \hat{p}_{im}, \hat{p}_{.m}) + cov(\sum_{i=1}^{I} \hat{p}_{im}, \hat{p}_{m.}) + cov(\sum_{j=1}^{J} \hat{p}_{mj}, \hat{p}_{m.}) + cov(\sum_{j=1}^{J} \hat{p}_{mj}, \hat{p}_{.m}) \\ &= D + F + G + H \end{aligned} \tag{A.11}$$

En utilisant (A.5), on trouve

$$D = \sum^{\star} p_{im} - p_{.m}(\sum_{i=1}^{I} p_{im})$$

De même,

$$E = p_{\star m} - p_{m.} \sum_{i=1}^{I} p_{im}$$

où $p_{\star m}$ désigne les p_{im} qui sont dans la colonne pour laquelle $p_{i.}$ est maximal.

$$F = p_{m\star} - p_{.m} \sum_{j=1}^{J} p_{mj}$$

où $p_{m\star}$ désigne les p_{mj} qui sont dans la ligne pour laquelle $p_{.j}$ est maximal.

$$G = \sum_{\star} p_{mj} - p_{m.}(\sum_{j=1}^{J} p_{mj})$$

où $\sum_{\star} p_{mj}$ désigne la somme des $\hat{p}_{mj}$ qui apparaissent dans la ligne correspondant à $p_{m.}$.

En posant

$$\Psi_{.} = p_{.m} + p_{m.}$$
$$\Psi_{\Sigma} = \sum_{i=1}^{I} p_{im} + \sum_{j=1}^{J} p_{mj}$$
$$\Psi_{\star} = \sum_{\star} p_{mj} + \sum^{\star} p_{im} + p_{m\star} + p_{\star m}$$

On obtient finalement,

$$\begin{aligned}(s.e.(\hat{\lambda}))^2 &= \frac{1}{n(2-\Psi_{.})^4}\left[(2-\Psi_{.})^2\left[\Psi_{\Sigma}(1-\Psi_{\Sigma}) + 2\sum^{\star} p_{im}\right]\right. \\ &+ (2-\Psi_{\Sigma})^2\left[\Psi_{.}(1-\Psi_{.}) + 2p_{\star\star}\right] \\ &- \left. 2(2-\Psi_{.})(2-\Psi_{\Sigma})\left[\Psi_{\star} - \Psi_{.}\Psi_{\Sigma}\right]\right] \qquad \text{(A.12)}\end{aligned}$$

Annexe B

ANOVA à un critère de classification : modèle non équilibré

Reprenons le modèle décrit au paragraphe 2.2. Graybill (1961) propose de procéder comme suit pour calculer les espérances des carrés moyens.

Lemme B.1 *Soient $x_1, \cdots, x_n$ des variables aléatoires normales et indépendantes de variance σ^2 et de moyenne μ,*
alors,

$$E[\sum_{i=1}^{n}(x_i - \overline{x})^2] = (n-1)var(x_i) = (n-1)\sigma^2 \tag{B.1}$$

démonstration : Il est bien connu que $\overline{x} \sim N(\mu, \frac{\sigma^2}{n})$. Dès lors,

$$\begin{aligned}
E[\sum_{i=1}^{n}(x_i - \overline{x})^2] &= E[\sum_{i=1}^{n}((x_i - \mu + \mu - \overline{x})^2)] \\
&= E[\sum_{i=1}^{n}(x_i - \mu)^2 + \sum_{i=1}^{n}(\mu - \overline{x})^2 - 2(\overline{x} - \mu)\sum_{i=1}^{n}(x_i - \mu)] \\
&= E[\sum_{i=1}^{n}(x_i - \mu)^2 + n(\mu - \overline{x})^2 - 2(\overline{x} - \mu)(n\overline{x} - n\mu)] \\
&= E[\sum_{i=1}^{n}(x_i - \mu)^2 + n(\mu - \overline{x})^2 - 2n(\overline{x} - \mu)^2] \\
&= E[\sum_{i=1}^{n}(x_i - \mu)^2 - n(\overline{x} - \mu)^2] \\
&= \sum_{i=1}^{n}E[(x_i - \mu)^2] - nE[(\overline{x} - \mu)^2] \\
&= n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2
\end{aligned}$$

Proposition B.1 *Supposons que les variables aléatoires $\underline{x}_{ij}$ soient réparties en g groupes d'effectifs k_i $(i = 1, \cdots, g)$ avec $\sum_{i=1}^{g} k_i = n$.*

Si l'on note

$$TSS = \sum_{i=1}^{g} \sum_{j=1}^{k_i} (\underline{x}_{ij} - \overline{\underline{x}}_{..})^2$$

la somme des carrés totale,

$$BSS = \sum_{i=1}^{g} \sum_{j=1}^{k_i} (\overline{\underline{x}}_{i.} - \overline{\underline{x}}_{..})^2$$

la somme des carrés entre populations et

$$WSS = \sum_{i=1}^{g} \sum_{j=1}^{k_i} (\underline{x}_{ij} - \overline{\underline{x}}_{i.})^2$$

les sommes de carrés des erreurs, alors

$$E[WSS] = \sigma_e^2 \tag{B.2}$$

et

$$E[BSS] = \sigma_e^2 + k_0 \sigma_\alpha^2 \tag{B.3}$$

où $k_0 = \dfrac{n^2 - \sum_{i=1}^{g} k_i^2}{n(g-1)}$.

démonstration :
En vertu du lemme 1, on a

$$\begin{aligned} E[WSS] &= \frac{1}{n-g} \sum_{i=1}^{g} \left\{ E\left(\sum_{j=1}^{k_i} (\underline{x}_{ij} - \overline{\underline{x}}_{i.})^2 \right) \right\} \\ &= \frac{1}{n-g} \sum_{i=1}^{g} (k_i - 1)\sigma_e^2 = \sigma_e^2 \end{aligned}$$

et

$$E[BSS] = \frac{1}{g-1} E\left[\sum_{i=1}^{g} \sum_{j=1}^{k_i} (\overline{\underline{x}}_{i.} - \overline{\underline{x}}_{..})^2 \right]$$

où

$$\begin{aligned} \overline{\underline{x}}_{..} &= \frac{1}{n} \sum_{i=1}^{g} k_i \overline{\underline{x}}_{i.} \\ &= \frac{1}{n} \sum_{i=1}^{g} k_i (\mu + \underline{\alpha}_i + \overline{\underline{e}}_{i.}) \end{aligned}$$

Ainsi,

$$\begin{aligned}
E[BSS] &= \frac{1}{g-1}E\left\{\sum_{i=1}^{g}\sum_{j=1}^{k_i}\left(\mu+\underline{\alpha}_i+\overline{\underline{e}}_{i.}-\mu-\frac{1}{n}\sum_{r=1}^{g}k_r\underline{\alpha}_r-\frac{1}{n}\sum_{q=1}^{g}\sum_{s=1}^{k_q}\overline{\underline{e}}_{qs}\right)\right\} \\
&= \frac{1}{g-1}\left\{E\left[\sum_{i=1}^{g}\sum_{j=1}^{k_i}\left(\underline{\alpha}_i-\frac{1}{n}\sum_{r=1}^{g}k_r\underline{\alpha}_r\right)^2\right]\right. \\
&+ E\left[\sum_{i=1}^{g}\sum_{j=1}^{k_i}\left(\overline{\underline{e}}_{i.}-\frac{1}{n}\sum_{q=1}^{g}\sum_{s=1}^{k_q}\overline{\underline{e}}_{qs}\right)^2\right] \\
&+ \left.2E\left[\sum_{i=1}^{g}\sum_{j=1}^{k_i}\left(\underline{\alpha}_i-\frac{1}{n}\sum_{r=1}^{g}k_r\underline{\alpha}_r\right)\left(\overline{\underline{e}}_{i.}-\frac{1}{n}\sum_{q=1}^{g}\sum_{s=1}^{k_q}\overline{\underline{e}}_{qs}\right)\right]\right\}
\end{aligned}$$

Posons $D = 2E\left[\sum_{i=1}^{g}\sum_{j=1}^{k_i}\left(\underline{\alpha}_i-\frac{1}{n}\sum_{r=1}^{g}k_r\underline{\alpha}_r\right)\left(\overline{\underline{e}}_{i.}-\frac{1}{n}\sum_{q=1}^{g}\sum_{s=1}^{k_q}\overline{\underline{e}}_{qs}\right)\right]$

On a

$$\begin{aligned}
D &= 2E\left[\sum_{i=1}^{g}k_i\underline{\alpha}_i\overline{\underline{e}}_{i.}-\frac{1}{n}\sum_{q=1}^{g}\sum_{s=1}^{k_q}\overline{\underline{e}}_{qs}\sum_{r=1}^{g}k_r\underline{\alpha}_r\right] \\
&= 2E\left[\sum_{i=1}^{g}k_i\underline{\alpha}_i\sum_{j=1}^{k_i}\overline{\underline{e}}_{ij}-\frac{1}{n}\sum_{q=1}^{g}\sum_{s=1}^{k_q}\overline{\underline{e}}_{qs}\sum_{r=1}^{g}k_r\underline{\alpha}_r\right] \\
&= 0
\end{aligned}$$

puisque les variables aléatoires $\underline{\alpha}_i$ et $\overline{\underline{e}}_{ij}$ sont indépendantes.

Donc,

$$
\begin{aligned}
E[BSS] &= \frac{1}{g-1}\left\{\sum_{i=1}^{g}\sum_{j=1}^{k_i} E\left[\underline{\alpha}_i^2 - \frac{2}{n}\underline{\alpha}_i\sum_{r=1}^{g} k_r\underline{\alpha}_r + \frac{1}{n^2}\left(\sum_{r=1}^{g} k_r\underline{\alpha}_r\right)^2\right]\right. \\
&+ \left.\sum_{i=1}^{g}\sum_{j=1}^{k_i} E\left[\overline{\underline{e}}_{i.}^2 - \frac{2}{n}\overline{\underline{e}}_{i.}\sum_{q=1}^{g}\sum_{s=1}^{k_q}\overline{\underline{e}}_{qs} + \frac{1}{n^2}\left(\sum_{q=1}^{g}\sum_{s=1}^{k_q}\overline{\underline{e}}_{qs}\right)^2\right]\right\} \\
&= \frac{1}{g-1}\left\{\sum_{i=1}^{g}\sum_{j=1}^{k_i}\left(\sigma_\alpha^2 - \frac{2}{n}\sum_{r=1}^{g} k_r E(\underline{\alpha}_i\underline{\alpha}_r) + \frac{1}{n^2}\sum_{r=1}^{g} k_r^2\sigma_\alpha^2 + \sum_{r=1}^{g}\sum_{l\neq r} k_r k_l E(\underline{\alpha}_r\underline{\alpha}_l)\right)\right. \\
&+ \left.\sum_{i=1}^{g}\sum_{j=1}^{k_i}\left(\frac{\sigma_e^2}{k_i} + \frac{1}{n^2}\sum_{q=1}^{g}\sum_{s=1}^{k_q}\sigma_e^2 + \frac{1}{n^2}\sum_{q,s}\sum_{m,l\neq q,s} E(\overline{\underline{e}}_{qs}\overline{\underline{e}}_{ml}) - \frac{2}{nk_i}\sum_{q=1}^{g}\sum_{s=1}^{k_q}\sum_{j=1}^{k_i} E(\overline{\underline{e}}_{ij}\overline{\underline{e}}_{qs})\right)\right\} \\
&= \frac{1}{g-1}\left\{\sum_{i=1}^{g}\sum_{j=1}^{k_i}(\sigma_\alpha^2 - \frac{2}{n}k_i\sigma_\alpha^2 + \frac{1}{n^2}\sum_{r=1}^{g} k_r^2\sigma_\alpha^2)\right. \\
&+ \left.\sum_{i=1}^{g}\sum_{j=1}^{k_i}(\frac{\sigma_e^2}{k_i} + \frac{1}{n}\sigma_e^2 - \frac{2}{nk_i}k_i\sigma_e^2)\right\}
\end{aligned}
$$

car les variables aléatoires $\underline{\alpha}_i$ comme les variables aléatoires $\overline{\underline{e}}_{ij}$ sont mutuellement indépendantes.

$$
\begin{aligned}
E[BSS] &= \frac{1}{g-1}\left\{\sum_{i=1}^{g}(k_i\sigma_\alpha^2 - 2\frac{k_i^2}{n}\sigma_\alpha^2 + \frac{k_i}{n^2}\sum_{r=1}^{g} k_r^2\sigma_\alpha^2)\right. \\
&+ \left.\sum_{i=1}^{g}(\sigma_e^2 - \frac{k_i}{n}\sigma_e^2)\right\} \\
&= \frac{1}{g-1}\left\{\sigma_\alpha^2\left(n - \frac{1}{n}\sum_{i=1}^{g} k_i^2\right) + (g-1)\sigma_e^2\right\} \\
&= \sigma_e^2 + \frac{n^2 - \sum_{i=1}^{g} k_i^2}{n(g-1)}\sigma_\alpha^2 \\
&= \sigma_e^2 + k_0\sigma_\alpha^2 \qquad \text{(B.4)}
\end{aligned}
$$

où
$k_0 = \dfrac{n^2 - \sum_{i=1}^{g} k_i^2}{n(g-1)}$.

[1] [**?**][2] [3][4][5] [6][7][8] [9][10][**?**][11][12] [13][14][15][16] [17][18][19][**?**][**?**][20][21] [22][23][24][25][26] [27][**?**][**?**] [28][29][30][31][32][**?**][33][34][35][**?**]

Table des matières

Bibliographie

[1] Agresti A. *Categorical data analysis*. John Wiley and Sons, 1990.

[2] Bahadur R.R. in Solomon H. *Studies in item Analysis and Prediction*. Stanford University Press, 1961.

[3] Bishop M.M.Y., Fienberg E.S., and Holland W.P. *Discrete Multivariate Analysis. Theory and practice*. the MIT Press, 1975.

[4] Blackman N. J-M. and Koval J.J. Interval estimation for cohen's kappa as a measure of agreement. *Statist. Med.*, 19 :723–741, 2000.

[5] Capobres D.B., Tosh F.E., Yates J.L., and H.V. Langeluttig. Experience with the tuberculin tine test in a sanatorium. *J.A.S.A*, 70 :561–567, 1962.

[6] Cicchetti D. and Fleiss J.L. Comparaison of the null distributions of weighted kappa and the c ordinal statistic. *Appl. Psychol. Meas.*, 1 :195–201, 1977.

[7] Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 :37–46, 1960.

[8] Cohen J. Weighted kappa : nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bul.*, 70 :213–220, 1968.

[9] Coughlin S.S., Pickle L.W., Goodman M.T., and Wilkens L.R. The logistic modeling of interobserver agreement. *J Clin Epidemiol*, 45 :1237–1241, 1992.

[10] Cramér H. *Mathematical Methods of Statistics*. Princeton, N.J., Princeton Univ. Press, 1946.

[11] Donner A. Sample size requirements for the comparaison of two or more coefficients of inter-observer agreement. *Statist. Med.*, 17 :1157–1168, 1998.

[12] Fermanian J. Mesure de l'accord entre deux juges. cas qualitatif. *Rev. Epidém. et Santé Publ.*, 32 :140–147, 1984.

[13] Fleiss J.L. Inference about weighted kappa in the non-null case. *Appl. Psychol. Meas.*, 1 :113–117, 1978.

[14] Fleiss J.L. and Cuzick J. The reliability of dichotomous judgements : unequal numbers of judges per subject. *Appl. Psychol. Meas.*, 3 :537–542, 1979.

[15] Fleiss J.L., Nee J.C.M., and Landis J.R. The large sample variance of kappa in the case of different sets of raters. *Psychol. Bull.*, 86 :974–977, 1979.

[16] Fleiss J.L. *Statistical methods for rates and proportions*. John Wiley, New York, 2nd edition, 1981.

[17] Goodman L.A. and Kruskal W.H. Measures of association for cross classifications i, ii, iii, iv. *Journal of the American Statistical Association*, 49, 54, 58, 67 :732–764, 123–163, 310–364, 415–421, 1954, 1959, 1963, 1972.

[18] Graybill F.A. *An introduction to linear statistical models, volume 1*. Mc Graw-Hill Series in probability and statistics, 1961.

[19] Greenberg R.A. and Jekel J.F. Some problems in the determination of false positive ans false negative rates of tuberculin tests. *American Review of Respiratory Disease*, 100 :645–650, 1969.

[20] Hui S.L. and Walter S.D. Estimating the error rates of diagnostic tests. *Biometrics*, 36 :167–171, 1980.

[21] Kraemer H.C. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika*, 44 :461–472, 1979.

[22] Landis J.R. and Koch G.G. The measurement of observer agreement for categorical data. *Biometrics*, 33 :159–174, 1977a.

[23] Landis J.R. and Koch G.G. A one-way components of variance model for categorical data. *Biometrics*, 33 :671–679, 1977b.

[24] Liang K.Y. and Zeger S.L. Longitudinal data analysis using generalized linear models. *Biometrika*, 73 :13–22, 1986.

[25] Lipsitz S.R., Williamson J., Klar N., Ibrahim J., and Parzen M. A simple method for estimating a regression model for κ between a pair of raters. *J.R. Statist. Soc. A*, 164 :449–465, 2001.

[26] Lipsitz S.R., Parzen M., Garrett F., and Klar N. A modified logistic regression model for analysing inter-rater agreement. *soumis à Psychometrika*.

[27] Mann H.B. and Wald A. On stochastic limit and order relationship. *Annals of Mathematical Statistics*, 14 :217–226, 1943.

[28] Nelder J.A. and Wedderburn R.W.M. Generalized linear models. *J.R. Statist. Soc.*, 135 :370–384, 1972.

[29] Noiroux Laurence. *Analyse statistique des courbes ROC*. Université de Liège, 2000.

[30] Oden N.L. Estimating kappa from binocular data. *statist. Med.*, 10 :1303–1313, 1991.

[31] Prentice R.L. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44 :1033–1048, 1988.

[32] Rao C.R. *Linear Statistical Inference and its Applications*. John Wiley, New York, 2nd edition, 1973.

[33] SAS. *User's Guide, Version 6*. Stanford University Press, 1990.

[34] Shoukri M.M. and Mian I.U.H. Maximum likelihood estimation of the kappa coefficient from bivariate logistic regression. *Statistics in medicine*, 15 :1409–1419, 1996.

[35] Smith T.W. Who, what, where, and why : an analysis of usage of the general social survey. Technical report, Chicago : National Opinion Research Center, 1996.