Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Facteurs d'inflation de la variance dans l'analyse des données d'enquêtes complexes

par Dan Liao et Richard Valliant

Juin 2012

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 57-67
Statistique Canada, N° 12-001-X au catalogue

Facteurs d'inflation de la variance dans l'analyse des données d'enquêtes complexes

Dan Liao et Richard Valliant [1]

Résumé

Les données d'enquêtes servent souvent à ajuster des modèles de régression linéaire. Les valeurs des covariables utilisées dans la modélisation n'étant toutefois pas contrôlées comme elles pourraient l'être dans une expérience, la colinéarité entre les covariables est un problème inévitable dans l'analyse des données d'enquêtes. Même si de nombreux livres et articles ont décrit le problème de la colinéarité et proposé des stratégies en vue de comprendre, d'évaluer et de traiter sa présence, la littérature sur les méthodes d'enquête n'a livré aucun outil diagnostique approprié pour évaluer son incidence sur l'estimation par la régression quand il est tenu compte de la complexité de l'enquête. Nous avons élaboré des facteurs d'inflation de la variance qui mesurent l'augmentation (« l'inflation ») de la variance des estimateurs des paramètres attribuable au fait que les variables explicatives ne sont pas orthogonales. Les facteurs d'inflation de la variance conviennent pour les estimateurs par la régression pondérée par les poids de sondage et tiennent compte des caractéristiques du plan de sondage complexe, par exemple, les pondérations, les grappes et les strates. Ces méthodes sont illustrées en utilisant un échantillon probabiliste provenant d'une enquête-ménage sur la santé et la nutrition.

Mots clés : Échantillon en grappes ; diagnostics de colinéarité ; estimateur de la variance par linéarisation ; moindres carrés pondérés par les poids de sondage ; échantillon stratifié.

1. Introduction

Dans une régression linéaire, la colinéarité des variables explicatives s'entend d'une situation où ces variables sont corrélées les unes aux autres. Les termes multicolinéarité et mauvais conditionnement sont également utilisés pour désigner ce genre de situation. La colinéarité est préoccupante pour des raisons tant numériques que statistiques. Les estimations des coefficients de pente peuvent être numériquement instables dans certains jeux de données en ce sens que de petites variations dans les $\mathbf{X}$ ou les $\mathbf{Y}$ peuvent produire de grandes variations dans les valeurs de ces estimations. Statistiquement, la corrélation entre les variables explicatives peut mener à des estimations de pente dont la variance est grande. En outre, quand les $\mathbf{X}$ sont fortement corrélés, le R^2 d'une régression peut être grand alors que les estimations de pente individuelles ne sont pas statistiquement significatives. Même si elles le sont, elles peuvent être de signe opposé à celui attendu (Neter, Kutner, Wasserman et Nachtsheim 1996). La colinéarité influe parfois aussi sur les prévisions (Smith 1974 ; Belsley 1984).

Dans les plans expérimentaux, il peut être possible de créer des situations où les variables explicatives sont orthogonales les unes par rapport aux autres. Par contre, dans de nombreuses enquêtes, des données sur des variables fortement corrélées sont recueillies pour l'analyse. Ainsi, le revenu total et ses composantes (par exemple, traitements et salaires, gains en capital, intérêts et dividendes) sont recueillis dans le cadre de la Panel Survey of Income Dynamics (http://psidonline.isr.umich.edu/) afin de suivre le bien-être économique au fil du temps. Lorsqu'une variable explicative est une combinaison linéaire des autres, on parle de colinéarité (ou de multicolinéarité) parfaite et celle-ci est facile à déceler. Les cas présentant un intérêt en pratique sont ceux où la colinéarité est imparfaite, mais a néanmoins une incidence sur la précision des estimations (Kmenta 1986, section 10.3).

Alors que la littérature sur les diagnostics de régression est abondante pour les données ne provenant pas d'enquêtes, elle l'est nettement moins pour les données d'enquêtes. Au cours de la dernière décennie, quelques articles ont présenté des techniques d'évaluation de la qualité de la régression appliquée à des données d'enquêtes complexes, principalement pour repérer les points et les groupes influents présentant des valeurs de données ou de poids de sondage anormales. Ainsi, Elliot (2007) a élaboré des méthodes bayésiennes de troncature des poids des estimateurs par la régression linéaire et par la régression linéaire généralisée pour les plans d'échantillonnage avec probabilités d'inclusion inégales. Li (2007a, b), de même que Li et Valliant (2009, 2011) ont adapté et étendu une série de techniques diagnostiques classiques à la régression sur des données d'enquêtes complexes, principalement pour la détection des observations influentes et des groupes d'observations influents. Les travaux de recherche de Li portent sur les résidus et les leviers, DFBETA, DFBETAS, DFFIT, DFFITS, la distance de Cook et l'approche *forward search* (recherche avant). Alors que de nombreuses publications de statistique appliquée offrent des suggestions et des lignes directrices précieuses pour aider les analystes des données à

1. Dan Liao, RTI International, 701 13th Street, N.W., Suite 750, Washington DC, 20005. Courriel : dliao@rti.org ; Richard Valliant, University of Michigan et University of Maryland, Joint Program in Survey Methodology, 1218 Lefrak Hall, College Park, MD, 20742.

diagnostiquer la présence de colinéarité (par exemple, Farrar et Glauber 1967 ; Theil 1971 ; Belsley, Kuh et Welsch 1980 ; Fox 1984 ; Belsley 1991), aucun de ces travaux de recherche ne traite des diagnostics de colinéarité lorsque des modèles sont ajustés au moyen de données d'enquêtes.

Le facteur d'inflation de la variance (VIF, de l'anglais *variance inflation factors*) décrit à la section 2, qui représente l'une des techniques classiques de diagnostic de la colinéarité les plus répandues, s'applique principalement à des régressions par les moindres carrés ordinaires ou pondérés. Le VIF mesure l'augmentation (« inflation ») de la variance de l'estimation d'une pente causée par la non-orthogonalité des variables explicatives en sus de ce que la variance serait dans des conditions d'orthogonalité. À la section 3, nous examinons le cas d'un analyste qui estime les paramètres d'un modèle en appliquant la méthode des moindres carrés pondérés par les poids de sondage (MCPPS) et calculons les VIF appropriés pour cette méthode. Les composantes du VIF peuvent être estimées en utilisant les éléments d'un estimateur de variance d'usage fréquent dans les progiciels pour l'analyse des données d'enquêtes. Dans le cas de la régression linéaire, un estimateur de variance de type sandwich permettra d'estimer à la fois la variance sous le modèle et la variance sous le plan de sondage de l'estimateur MCPPS de la pente. Comme nous le montrerons à la section 3, la variance sous le modèle ou sous le plan de sondage de $\hat{\beta}_k$, un estimateur de la pente associée à la variable explicative $\mathbf{x}_k$, présente une certaine inflation quand diverses variables explicatives sont corrélées les unes aux autres, comparativement à ce qu'elle serait si $\mathbf{x}_k$ était orthogonale aux autres variables explicatives. La mesure de l'inflation, le VIF, est composée de termes qui doivent être estimés d'après l'échantillon. Notre approche consiste à substituer des estimateurs qui ont une interprétation à la fois sous le modèle et sous le plan, comme il est décrit à la section 3.5.

À la quatrième section, nous présentons une étude empirique portant sur des données de la National Health and Nutrition Examination Survey réalisée aux États-Unis. Nous montrons l'application de notre nouvelle approche et comparons les valeurs nouvellement dérivées du VIF pour l'estimateur MCPPS à celles obtenues pour les estimateurs par les moindres carrés ordinaires et par les moindres carrés pondérés, qui sont produites par les progiciels statistiques standard. Les comparaisons montrent que les valeurs du VIF diffèrent selon la méthode de régression et qu'un VIF spécial pour échantillon complexe devrait être utilisé pour évaluer l'effet nuisible de la colinéarité dans l'analyse des données d'enquêtes.

2. Diagnostics de colinéarité dans l'estimation par les moindres carrés ordinaires

Supposons que l'échantillon s contient n unités sur chacune desquelles sont observées p variables explicatives $\mathbf{x}$ et une variable d'analyse Y. Le modèle linéaire classique, dans un contexte autre qu'une enquête, est $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, où $\mathbf{Y}$ est un vecteur de dimension $n \times 1$ d'observations sur une variable réponse, ou variable dépendante ; $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ est une matrice de plan de dimensions $n \times p$ de constantes fixes avec $\mathbf{x}_k$, le vecteur de dimension $n \times 1$ des valeurs de la variable explicative k pour les n unités échantillonnées ; $\boldsymbol{\beta}$ est un vecteur de dimension $p \times 1$ des paramètres à estimer ; et $\boldsymbol{\epsilon}$ est un vecteur de dimension $n \times 1$ de termes d'erreur statistiquement indépendants de moyenne nulle et de variance constante σ^2. Nous supposons, pour simplifier, que $\mathbf{X}$ est de plein rang-colonne. L'estimation par les moindres carrés ordinaires (MCO) de $\boldsymbol{\beta}$ est $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, pour lequel la variance sous le modèle est $\mathrm{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Ici, nous utilisons l'indice M pour désigner l'espérance sous le modèle.

Les colinéarités des variables explicatives accroissent la variance des coefficients de régression du modèle comparativement à la situation où les $\mathbf{X}$ sont orthogonales. On peut voir cet effet dans la formule de la variance de l'estimation d'un coefficient de régression particulier autre que l'ordonnée à l'origine $\hat{\beta}_k$ (Theil 1971),

$$\mathrm{Var}_M(\hat{\beta}_k) = \frac{\sigma^2}{\sum_{i \in s} x_{ik}^2} \frac{1}{1 - R_k^2} \qquad (1)$$

où R_k^2 est le carré de la corrélation multiple provenant de la régression de la k^{e} colonne de $\mathbf{X}$ sur les autres colonnes. Ce R-carré, défini comme étant $R_k^2 = \hat{\beta}_{(k)}\mathbf{X}_{(k)}^T\mathbf{X}_{(k)}\hat{\beta}_{(k)} / \mathbf{x}_k^T\mathbf{x}_k$, où $\hat{\beta}_{(k)}$ est l'estimation par les MCO de la pente quand on fait la régression de $\mathbf{x}_k$ sur les autres $\mathbf{x}$ et que $\mathbf{X}_{(k)}$ est la matrice $\mathbf{X}$ dont la k^{e} colonne a été supprimée. Le terme $\sigma^2 / \sum x_{ik}^2$ est la variance de $\hat{\beta}_k$ sous le modèle si la k^{e} variable explicative est orthogonale à toutes les autres variables explicatives. La valeur de R_k^2 peut être non nulle parce que la k^{e} variable explicative est corrélée à une autre variable explicative ou à cause d'une relation plus complexe de dépendance entre $\mathbf{x}_k$ et plusieurs autres variables explicatives. Par conséquent, la colinéarité entre $\mathbf{x}_k$ et certaines autres variables explicatives peut entraîner l'inflation de la variance de $\hat{\beta}_k$ au-delà de la valeur qui serait obtenue si les $\mathbf{X}$ étaient orthogonales. Le deuxième terme de (1), $(1 - R_k^2)^{-1}$, est appelé facteur d'inflation de la variance, symbolisé par VIF, pour *variance inflation factor* (Theil 1971).

Une référence fondamentale sur la colinéarité et d'autres diagnostics de la régression par les MCO est le document de Belsley et coll. (1980). Les diagnostics de colinéarité sont abordés dans de nombreux autres traités, y compris Fox (1984) et Neter et coll. (1996). Dans certains cas, il est souhaitable de pondérer les cas différemment dans une analyse de régression afin d'intégrer une variance résiduelle non constante. Cette forme de pondération, qui est fondée

sur un modèle, est la méthode des moindres carrés pondérés (MCP). La plupart des progiciels statistiques courants (par exemple, SAS, Stata, S-Plus et R) utilisent $(1 - R^2_{k(\mathrm{MCP})})^{-1}$ comme VIF pour les MCP, où $R^2_{k(\mathrm{MCP})}$ est le carré de la corrélation multiple provenant de la régression par les MCP de la k^{e} colonne de $\mathbf{X}$ sur les autres colonnes. Fox et Monette (1992) ont également généralisé ce concept d'inflation de la variance en tant que mesure de colinéarité à un sous-ensemble de paramètres dans $\mathbf{b}$ et dérivé un facteur d'inflation de la variance généralisé (GVIF pour *generalized variance-inflation factor*). De surcroît, certains travaux intéressants ont abouti à l'élaboration de mesures de type VIF, telles que les *indices de colinéarité* de Steward (1987) qui sont simplement les racines carrées des VIF et la *tolérance* définie comme étant l'inverse du VIF dans Simon et Lesage (1988).

3. Le VIF dans la régression par les moindres carrés pondérés par les poids de sondage

3.1 Estimateurs par les moindres carrés pondérés par les poids de sondage

Supposons que le modèle structurel sous-jacent à la superpopulation est $\mathbf{Y} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{e}$, où les termes d'erreur du modèle ont une structure de variance générale $\mathbf{e} \sim (0, \sigma^2\mathbf{V})$ avec $\mathbf{V}$ et σ^2 connues. Soit $\mathbf{W}$ la matrice diagonale des poids de sondage. Nous supposons tout au long de l'exposé que les poids de sondage sont construits de telle façon qu'ils peuvent être utilisés pour estimer les totaux de population finie. L'estimateur par les moindres carrés pondérés par les poids de sondage (MCPPS) est $\hat{\boldsymbol{\beta}}_{\mathrm{PPS}} = (\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{WY}$, en supposant que $\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X}$ est inversible. Fuller (2002) décrit les propriétés de cet estimateur.

L'estimateur $\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$ est un estimateur sans biais par rapport au modèle de $\boldsymbol{\beta}$ sous le modèle $\mathbf{Y} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{e}$, que $\mathrm{Var}_M(\mathbf{e}) = \sigma^2\mathbf{V}$ soit spécifiée correctement ou non, et est approximativement sans biais sous le plan pour le paramètre de population (recensement) $\mathbf{B}_U = (\mathbf{X}_U^T\mathbf{X}_U)^{-1}\mathbf{X}_U^T\mathbf{Y}_U$, dans la population finie de N unités. L'indice U désigne la population finie, $\mathbf{Y}_U = (Y_1, ..., Y_N)^T$, et $\mathbf{X}_U = (\mathbf{x}_1, ..., \mathbf{x}_p)$ avec $\mathbf{x}_k$ le vecteur de dimension $N \times 1$ des valeurs de la covariable *k*.

3.2 Variance des estimations des coefficients sous le modèle

La variance sous le modèle de l'estimateur du paramètre $\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$, en supposant que $\mathrm{Var}_M(\mathbf{e}) = \sigma^2\mathbf{V}$, peut s'exprimer

$$\begin{aligned}\mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}) &= \sigma^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1} \\ &= \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\sigma^2 = \mathbf{G}\sigma^2, \qquad (2)\end{aligned}$$

où $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$, $\tilde{\mathbf{V}} = \mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}$, $\mathbf{A} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$, $\mathbf{B} = \tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}$, et $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$.

Si les colonnes de $\mathbf{X}$ sont orthogonales, $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \mathrm{diag}(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)$ et $\mathbf{A}^{-1} = \mathrm{diag}(1/\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)$, où $\tilde{\mathbf{x}}_k = \mathbf{w}_k^{1/2}\mathbf{x}_k$. Le ij^{e} élément de $\mathbf{G}$ devient alors $\tilde{\mathbf{x}}_i^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_j/(\tilde{\mathbf{x}}_i^T\tilde{\mathbf{x}}_i)^2$. Donc, quand les $\mathbf{X}$ sont orthogonales, la variance de $\hat{\beta}_{\mathrm{PPS}_k}$ sous le modèle est

$$\mathrm{Var}_M(\hat{\beta}_{\mathrm{PPS}_k}) = \sigma^2\,\tilde{\mathbf{x}}_k^T\,\tilde{\mathbf{V}}\,\tilde{\mathbf{x}}_k/(\tilde{\mathbf{x}}_k^T\,\tilde{\mathbf{x}}_k)^2, \qquad (3)$$

un fait que nous utiliserons plus tard. De manière plus générale, la variance sous le modèle de $\hat{\beta}_{\mathrm{PPS}_k}$, l'estimation du coefficient de la k^{e} variable explicative, est

$$\mathrm{Var}_M(\hat{\beta}_{\mathrm{PPS}_k}) = \mathbf{i}_k'\,\mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}})\mathbf{i}_k = \sigma^2\,\mathbf{i}_k'\,\mathbf{G}\mathbf{i}_k = \sigma^2 g^{kk} \quad (4)$$

où $\mathbf{i}_k$ est un vecteur de dimension $p \times 1$ avec 1 en position k et des 0 ailleurs, et g^{kk} est le k^{e} élément diagonal de la matrice $\mathbf{G}$.

3.3 VIF fondé sur le modèle

Comme nous le montrons à l'annexe A, la variance sous le modèle de $\hat{\beta}_{\mathrm{PPS}_k}$ dans (4) peut s'écrire :

$$\mathrm{Var}_M(\hat{\beta}_{\mathrm{PPS}_k}) = g^{kk}\sigma^2 = \frac{\zeta_k\rho_k}{1 - R^2_{\mathrm{PPS}(k)}}\frac{\sigma^2\,\tilde{\mathbf{x}}_k^T\,\tilde{\mathbf{V}}\,\tilde{\mathbf{x}}_k}{(\tilde{\mathbf{x}}_k^T\,\tilde{\mathbf{x}}_k)^2}, \quad (5)$$

où

$$\zeta_k = \frac{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{V}}\tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{e}}_{xk}} = \frac{\mathbf{e}_{xk}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}\mathbf{e}_{xk}},$$

avec $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{\mathrm{PPS}(k)}$ représentant le résidu de la régression par les MCPPS de $\mathbf{x}_k$ sur $\mathbf{X}_{(k)}$, et $\tilde{\mathbf{e}}_{xk} = \tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\mathrm{PPS}(k)} = \mathbf{W}^{1/2}\mathbf{e}_{xk}$,

$$\rho_k = \frac{\tilde{\mathbf{x}}_k^T\,\tilde{\mathbf{x}}_k}{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k} = \frac{\mathbf{x}_k^T\mathbf{W}\mathbf{x}_k}{\mathbf{x}_k^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{x}_k}$$

et $R^2_{\mathrm{PPS}(k)}$, défini à l'annexe A, est le carré de la corrélation multiple provenant de la régression pondérée de la k^{e} colonne de $\mathbf{X}$ sur les autres colonnes. Donc, ζ_k et ρ_k dépendent de $\mathbf{W}$ et de $\mathbf{V}$. La variance sous orthogonalité donnée par (3) est augmentée

$$\mathrm{VIF}_k = \frac{\zeta_k\rho_k}{1 - R^2_{\mathrm{PPS}(k)}} \qquad (6)$$

fois lorsque l'on introduit les $p - 1$ autres variables explicatives dans les MCPPS. Dans ces derniers, le VIF fondé sur le modèle inclut non seulement le coefficient de corrélation multiple $R^2_{\mathrm{PPS}(k)}$, mais aussi deux coefficients d'ajustement, ζ_k et ρ_k, qui ne sont pas présents dans le cas des MCO et des MCP.

En procédant à la décomposition de $\tilde{\mathbf{V}}$ en valeurs singulières, nous pouvons borner le facteur $\zeta_k\rho_k$, qui est l'ajustement du VIF dans les MCP. Sur la base des extrema

du ratio des formes quadratiques (Lin 1984), les bornes du terme ζ_k sont celles de l'intervalle $\mu_{\min}(\tilde{\mathbf{V}}) \le \zeta_k \le \mu_{\max}(\tilde{\mathbf{V}})$, et les bornes de ρ_k, celles de l'intervalle

$$\frac{1}{\mu_{\max}(\tilde{\mathbf{V}})} \le \rho_k \le \frac{1}{\mu_{\min}(\tilde{\mathbf{V}})},$$

où $\mu_{\min}(\tilde{\mathbf{V}})$ et $\mu_{\max}(\tilde{\mathbf{V}})$ sont les valeurs singulières minimale et maximale de la matrice $\tilde{\mathbf{V}}$. En combinant ces résultats, le coefficient conjoint $\zeta_k \rho_k$ a pour bornes l'intervalle :

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} \le \zeta_k \rho_k \le \frac{\mu_{\max}(\tilde{\mathbf{V}})}{\mu_{\min}(\tilde{\mathbf{V}})}.$$

Notons que, quand $\tilde{\mathbf{V}} = \mathbf{I}$, $\zeta_k = \rho_k = 1$ et (6) se réduit à

$$\frac{1}{1 - R^2_{\text{PPS}(k)}} \frac{\sigma^2}{\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k},$$

qui est la variance sous le modèle de l'estimation par les MCP quand $\mathbf{V}$ est diagonale et que $\mathbf{W}$ est correctement spécifiée comme étant $\mathbf{W} = \mathbf{V}^{-1}$. Dans ce cas inhabituel, le VIF calculé les progiciels courants sera approprié pour les MCPPS. Cependant, l'hypothèse que $\mathbf{W} = \mathbf{V}^{-1}$ est rarement raisonnable dans l'estimation d'après des données d'enquête. Si $\tilde{\mathbf{V}} \ne \mathbf{I}$, alors ζ_k et ρ_k ne sont pas égaux à 1 et un calcul spécial du VIF demeure nécessaire. Quand $\mathbf{V} = \mathbf{I}$, soit l'application habituellement considérée par les analystes,

$$\tilde{\mathbf{V}} = \mathbf{W}, \zeta_k = \frac{\tilde{\mathbf{e}}_{xk}^T \mathbf{W} \tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{e}}_{xk}}, \rho_k = \frac{\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k}{\tilde{\mathbf{x}}_k^T \mathbf{W} \tilde{\mathbf{x}}_k}$$

et

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} = \frac{w_{\min}}{w_{\max}},$$

où $w_{\min}$ est la valeur minimale des poids de sondage et $w_{\max}$ est leur valeur maximale. Dans ce cas, l'intervalle de $\zeta_k \rho_k$ est borné par

$$\left[\frac{w_{\min}}{w_{\max}}, \frac{w_{\max}}{w_{\min}} \right].$$

Quand tous les poids de sondage sont constants, $\zeta_k \rho_k = 1$ et les VIF produits par les logiciels standard, $(1 - R^2_{\text{PPS}})^{-1}$, n'ont pas besoin d'être ajustés sous les MCPPS ; cependant, si l'intervalle des poids de sondage est grand, $\zeta_k \rho_k$ peut être très petit ou très grand, et peut être supérieur ou inférieur à 1. Dans ce cas, le VIF produit par les logiciels standard ne convient pas et un calcul spécial est nécessaire. Ces faits seront illustrés dans nos études expérimentales.

Le VIF donné par (6) est approprié, que le modèle comprenne ou non une ordonnée à l'origine. Une autre version peut être écrite en supposant que le modèle contient une ordonnée à l'origine quand on effectue la régression de $\tilde{\mathbf{x}}_k$ sur les autres $\mathbf{x}$. L'établissement de cette forme figure dans Liao (2010). Nous résumons le résultat ci-dessous.

La variance de $\hat{\beta}_{\text{PPS}_k}$ dans un modèle M2 qui contient une ordonnée à l'origine et dans lequel $\mathbf{x}_k$ est orthogonale aux autres $\mathbf{x}$ est :

$$\text{Var}_{M2}(\hat{\beta}_{\text{PPS}_k}) = \frac{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)^T \tilde{\mathbf{V}} (\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)}{\text{SST}^2_{\text{PPS}_{m(k)}}} \qquad (7)$$

où $\tilde{\mathbf{1}} = (w_1^{1/2}, ..., w_n^{1/2})$, $\bar{\tilde{x}}_k = \sum_{i \in s} w_i x_{ki} / \hat{N}$, $\hat{N} = \sum_{i \in s} w_i$ et $\text{SST}_{\text{PPS}_{m(k)}} = \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k - \hat{N} \bar{\tilde{x}}_k^2$. La variance de $\hat{\beta}_{\text{PPS}_k}$ peut alors s'écrire

$$\text{Var}_M(\hat{\beta}_{\text{PPS}_k}) = \frac{\zeta_k \rho_{mk}}{1 - R^2_{\text{PPS}m(k)}} \text{Var}_{M2}(\hat{\beta}_{\text{PPS}_k}) \qquad (8)$$

où $R^2_{\text{PPS}m(k)}$ est le R-carré de la régression par les MCPPS de $\tilde{\mathbf{x}}_k$ sur les $\mathbf{x}$ dans la partie restante de $\tilde{\mathbf{X}}$ (en excluant une colonne pour l'ordonnée à l'origine). Le terme ζ_k a été défini en suivant (5) et

$$\rho_{mk} = \frac{(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k - \hat{N} \bar{\tilde{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)^T \tilde{\mathbf{V}} (\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)}.$$

La plupart des progiciels donnent systématiquement $(1 - R^2_{\text{PPS}m(k)})^{-1}$ comme VIF dans la sortie de la régression par les MCP. Il convient de souligner que ce paramètre est différent du VIF égal à $(1 - R^2_{\text{PPS}(k)})^{-1}$ présenté à la section 3.3, qui ne suppose pas qu'une ordonnée à l'origine est gardée dans le modèle. Habituellement, les progiciels ne fournissent pas $(1 - R^2_{\text{PPS}(k)})^{-1}$.

Au moyen d'arguments similaires à ceux de la section précédente, nous pouvons borner $\zeta_k \rho_{mk}$ par

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} \le \zeta_k \rho_{mk} \le \frac{\mu_{\max}(\tilde{\mathbf{V}})}{\mu_{\min}(\tilde{\mathbf{V}})}.$$

La variance sous le modèle de $\hat{\beta}_{\text{PPS}_k}$ est augmentée d'une valeur

$$\text{VIF}_{mk} = \frac{\zeta_k \rho_{mk}}{1 - R^2_{\text{PPS}m(k)}}$$

comparativement à sa variance sous le modèle (M2) contenant uniquement la variable explicative $\tilde{\mathbf{x}}_k$ et l'ordonnée à l'origine. Le nouveau VIF_{mk} ajusté pour tenir compte de l'ordonnée à l'origine garde certaines propriétés du VIF_k donné en (6). Quand $\tilde{\mathbf{V}} = \mathbf{I}$, nous avons $\zeta_k = 1$, $\rho_{mk} = 1$ et le VIF corrigé pour l'ordonnée à l'origine dans (8) pour les MCPPS est égal au VIF corrigé pour l'ordonnée à l'origine classique : $(1 - R^2_{m(k)})^{-1}$. Quand $\mathbf{V} = \mathbf{I}$, nous avons $\tilde{\mathbf{V}} = \mathbf{W}$,

$$\zeta_k = \frac{\tilde{\mathbf{e}}_{xk} \mathbf{W} \tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{e}}_{xk}}, \rho_{mk} = \frac{(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k - \hat{N} \bar{\tilde{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)^T \mathbf{W} (\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)}$$

et

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} = \frac{w_{\min}}{w_{\max}}.$$

L'intervalle de $\zeta_k \rho_{mk}$ dépend aussi de l'intervalle des poids de sondage, comme celui de $\zeta_k \rho_k$.

3.4 Estimation du VIF pour un modèle avec échantillonnage en grappes stratifié quand $\mathbf{V}$ est inconnue

Aux sections qui précèdent, nous avons utilisé des arguments fondés sur le modèle pour calculer les VIF. Ces derniers contiennent des termes, $\tilde{\mathbf{V}}$ en particulier, qui sont inconnus et doivent être estimés. À la présente section, nous construisons des estimateurs des composantes du VIF, de nouveau en nous servant d'arguments fondés sur le modèle. Cependant, un estimateur par linéarisation classique de la variance sous le plan permet aussi d'estimer la variance sous le modèle, comme nous le montrons plus bas, et fournit les composantes nécessaires pour estimer le VIF. Dans la suite de la présente section, nous présentons des estimateurs qui conviennent pour un modèle possédant une structure de covariance en grappes stratifiée.

Supposons que, dans un plan d'échantillonnage à plusieurs degrés stratifié, il existe $h = 1, ..., H$ strates dans la population, $i = 1, ..., N_h$ grappes dans la strate correspondante h et $t = 1, ..., M_{hi}$ unités dans la grappe hi. Nous sélectionnons $i = 1, ..., n_h$ grappes dans la strate h et $t = 1, ..., m_{hi}$ unités dans la grappe hi. Désignons l'ensemble de grappes échantillonnées dans la strate h par s_h et l'échantillon d'unités dans la grappe hi par s_{hi}. Le nombre total d'unités échantillonnées dans la strate h est $m_h = \sum_{i \in s_h} m_{hi}$, et le total dans l'échantillon est $m = \sum_{h=1}^{H} m_h$. Nous supposons que les grappes sont sélectionnées avec remise dans les strates et indépendamment entre les strates. Considérons le modèle qui suit :

$$\begin{aligned} &E_M(Y_{hit}) = \mathbf{x}_{hit}^T \boldsymbol{\beta} \\ &\qquad h = 1, \ldots, H, \quad i = 1, \ldots, N_h, \quad t = 1, \ldots, M_{hi} \\ &\mathrm{Cov}_M(Y_{hit}, Y_{h'i't'}) = 0 \\ &\qquad h \neq h', \text{ où, } h = h' \text{ et } i \neq i'. \end{aligned} \tag{9}$$

Nous supposons que les unités sont corrélées dans chaque grappe, mais il n'est pas nécessaire de spécifier la corrélation particulière des covariances pour la présente analyse. L'estimateur du paramètre de régression est :

$$\hat{\boldsymbol{\beta}}_{\mathrm{PPS}} = \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{A}^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{Y}_{hi} \tag{10}$$

où $\mathbf{X}_{hi}$ est la matrice de dimensions $m_{hi} \times p$ des covariables pour les unités échantillonnées dans la grappe hi, $\mathbf{W}_{hi} = \mathrm{diag}(w_t)$, $t \in s_{hi}$ est la matrice diagonale des poids de sondage pour la grappe hi et $\mathbf{Y}_{hi}$ est le vecteur de dimensions $m_{hi} \times 1$ de variables réponse dans la grappe hi. La variance sous le modèle de $\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$ est :

$$\begin{aligned} \mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}) &= \mathbf{A}^{-1} \left[\sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{V}_{hi} \mathbf{W}_{hi} \mathbf{X}_{hi} \right] \mathbf{A}^{-1} \\ &= \mathbf{A}^{-1} \left[\sum_{h=1}^{H} \mathbf{X}_h^T \mathbf{W}_h \mathbf{V}_h \mathbf{W}_h \mathbf{X}_h \right] \mathbf{A}^{-1}, \end{aligned} \tag{11}$$

où $\mathbf{V}_{hi} = \mathrm{Var}_M(\mathbf{Y}_{hi})$ et $\mathbf{V}_h = \mathrm{Blkdiag}(\mathbf{V}_{hi})$, $i \in s_h$. L'expression (11) est un cas particulier de (2) avec $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T, ..., \mathbf{X}_H^T)$, $\mathbf{X}_h$ est la matrice de dimensions $m_h \times p$ des covariables pour les unités échantillonnées dans la strate h, $\mathbf{W} = \mathrm{diag}(\mathbf{W}_{hi})$, pour $h = 1, ..., H$ et $i \in s_h$, et $\mathbf{V} = \mathrm{Blkdiag}(\mathbf{V}_h)$.

Désignons les résidus au niveau de la grappe comme un vecteur, $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$. Un estimateur par linéarisation sous le plan de sondage est donné par :

$$\begin{aligned} \mathrm{var}_L(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}) &= \mathbf{A}^{-1} \left[\sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i \in s_h} (\mathbf{z}_{hi} - \bar{\mathbf{z}}_h)(\mathbf{z}_{hi} - \bar{\mathbf{z}}_h)^T \right] \mathbf{A}^{-1} \\ &= \mathbf{A}^{-1} \left[\sum_{h=1}^{H} \frac{n_h}{n_h - 1} \left(\sum_{i \in s_h} \mathbf{z}_{hi} \mathbf{z}_{hi}^T - n_h \bar{\mathbf{z}}_h \bar{\mathbf{z}}_h^T \right) \right] \mathbf{A}^{-1}, \end{aligned} \tag{12}$$

où

$$\bar{\mathbf{z}}_h = \frac{1}{n_h} \sum\nolimits_{i \in s_h} \mathbf{z}_{hi}$$

et $\mathbf{z}_{hi} = \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{e}_{hi}$ avec $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$. Cette expression peut être réduite à la formule pour un plan de sondage stratifié à seul degré quand les tailles des échantillons de grappes sont toutes égales à 1, $m_{hi} = 1$. L'expression (12) est utilisée par les progiciels Stata et SUDAAN, entre autres. L'estimateur $\mathrm{var}_L(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}})$ est convergent et approximativement sans biais par rapport au plan sous un plan dans lequel les grappes sont sélectionnées avec remise (Fuller 2002). Li (2007a, b) a montré que (12) est également un estimateur approximativement sans biais par rapport au modèle sous le modèle (11).

Le terme entre crochets dans (12) sert d'estimateur de la matrice $\mathbf{B}$ dans l'expression (2). Les composantes de $\mathrm{var}_L(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}})$ peuvent être utilisées pour construire les estimateurs de ζ_k et ρ_k dans (5) et de ρ_{mk} dans (8). En particulier,

$$\hat{\zeta}_k = \frac{\mathbf{e}_{xk}^T \mathbf{W} \hat{\mathbf{V}} \mathbf{W} \mathbf{e}_{xk}}{\mathbf{e}_{xk}^T \mathbf{W} \mathbf{e}_{xk}}, \tag{13}$$

où

$$\hat{\mathbf{V}} = \mathrm{Blkdiag}\left[\frac{n_h}{n_h - 1} \left(\hat{\mathbf{V}}_h - \frac{1}{n_h} \mathbf{e}_h \mathbf{e}_h^T \right) \right], \; h = 1, 2, ..., H$$

avec $\hat{\mathbf{V}}_h = \mathrm{Blkdiag}(\mathbf{e}_{hi} \mathbf{e}_{hi}^T)$ et

$$\hat{\rho}_k = \frac{\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{W} \hat{\mathbf{V}} \mathbf{W} \mathbf{x}_k},$$

avec $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{\mathrm{PPS}(k)}$. L'estimation de $\hat{\rho}_{mk}$, défini selon (8), est

$$\hat{\rho}_{mk} = \frac{(\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k - \hat{N} \bar{\bar{x}}_k^2)}{(\mathbf{x}_k - \mathbf{1}\bar{\bar{x}}_k)^T \mathbf{W} \hat{\mathbf{V}} \mathbf{W} (\mathbf{x}_k - \mathbf{1}\bar{\bar{x}}_k)}. \qquad (14)$$

Étant donné ces estimateurs des composantes, VIF_k est estimé par

$$\widehat{\mathrm{VIF}}_k = \frac{\hat{\zeta}_k \hat{\rho}_k}{1 - R^2_{\mathrm{PPS}(k)}}$$

et VIF_{mk} est estimé par

$$\widehat{\mathrm{VIF}}_{mk} = \frac{\hat{\zeta}_k \hat{\rho}_{mk}}{1 - R^2_{\mathrm{PPS}m(k)}}.$$

4. Étude expérimentale

Nous allons maintenant illustrer les diagnostics de colinéarité modifiés proposés et étudier leur comportement en utilisant des données sur l'apport alimentaire provenant de la National Health and Nutrition Examination Survey (NHANES) réalisée aux États-Unis en 2007-2008 (http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/datadoc_changes_0708.htm). Les données sur l'apport alimentaire sont utilisées pour estimer les types et les quantités d'aliments et de boissons consommés durant la période de 24 heures qui précède l'entrevue (de minuit à minuit) et pour estimer les apports d'énergie, de nutriments et d'autres composants alimentaires provenant de ces aliments et boissons. La NHANES est réalisée selon un plan d'échantillonnage probabiliste à plusieurs degrés complexe. Certains sous-groupes de population sont suréchantillonnés afin d'accroître la fiabilité et la précision des estimations des indicateurs de l'état de santé pour ces groupes. Parmi les participants qui répondent à l'interview sur place au centre d'examen mobile (CEM), environ 94 % fournissent des renseignements complets sur les apports alimentaires. Les poids de sondage pour ces données ont été construits en prenant les poids de sondage ajustés pour le CEM et en les rajustant en outre pour tenir compte de la non-réponse supplémentaire et de la différence de répartition selon le jour de la semaine pour la collecte des données sur les apports alimentaires. Ces derniers poids sont plus variables que les poids produits pour le CEM. Le jeu de données utilisé dans notre étude est un sous-ensemble des données de 2007-2008 composé de femmes de 26 à 40 ans ayant répondu à l'enquête. Les observations comportant des valeurs manquantes pour les variables choisies ont été exclues de l'échantillon qui, en dernière analyse, contient 672 réponses complètes. Les poids finaux dans notre échantillon varient de 6 028 à 330 067, avec un ratio de 55 pour 1. Le National Center for Health Statistics des États-Unis recommande que le plan de sélection de l'échantillon s'approche de la sélection stratifiée avec remise de 32 UPE dans 16 strates, avec 2 UPE dans chaque strate.

Pour notre étude empirique, nous avons ajusté un modèle de régression linéaire du poids corporel (en kg) par la méthode des moindres carrés pondérés par les poids de sondage. Les variables explicatives considérées comprennent l'âge, la race noire et les variables d'apport nutritionnel total quotidien, qui sont les calories (100 kcal), les protéines (100 g), les glucides (100 g), les sucres (100 g), les lipides totaux (100 g), les acides gras saturés totaux (100 g), les acides gras mono-insaturés totaux (100 g), les acides gras poly-insaturés totaux (100 g) et l'alcool (100 g). Toutes les variables d'apport nutritionnel total quotidien sont corrélées les unes aux autres à divers degrés comme l'illustre la figure 1.

Trois méthodes de régression ont été comparées dans l'étude. La première s'appuie sur la méthode des *moindres carrés ordinaires* (MCO) et ne tient pas compte des aspects complexes de l'échantillonnage, y compris la pondération. La deuxième est la méthode des *moindres carrés pondérés* (MCP), dans laquelle sont intégrés les poids de sondage en supposant que $\mathbf{V} = \mathbf{W}^{-1}$, mais qui ne tient pas compte de tous les aspects complexes de l'échantillonnage. La troisième est la méthode des *moindres carrés pondérés par les poids de sondage* (MCPPS), qui s'appuie sur le plan d'échantillonnage complexe réel comme il est décrit à la section 3.4. Les matrices des poids, les estimateurs des variances des coefficients et les diagnostics de colinéarité de ces trois méthodes sont présentés au tableau 1.

Les résultats de l'ajustement du modèle en utilisant les trois méthodes de régression différentes sont présentés au tableau 2. Le modèle contenant toutes les variables explicatives figure dans la partie supérieure du tableau. Dans le tiers inférieur du tableau, un modèle réduit, dans lequel le problème de quasi-dépendance est moindre, est ajusté au moyen de trois variables explicatives seulement : l'âge, la race noire et les calories. Dans le modèle réduit, la valeur du coefficient pour les calories est positive et significative quand les MCP ou les MCPPS sont utilisés, ce qui paraît logique et reflète la relation positive attendue entre le poids corporel d'une répondante et son apport calorique total quotidien. Cependant, lorsque l'on inclut les autres variables d'apport nutritionnel total dans le modèle, la valeur du coefficient des calories est négative et non significative à cause de l'inflation de sa variance. Il s'agit d'un exemple type dans lequel il y a inflation de la variance d'un coefficient et où son signe est illogique à cause de la colinéarité.

Le tableau 3 donne les valeurs du VIF pour les trois méthodes de régression utilisées. Les formules du VIF pour ces méthodes sont présentées au tableau 1. Quand toutes les

variables explicatives sont incluses dans le modèle, la variable des calories obtient la valeur du VIF la plus élevée dans toutes les régressions en raison de sa forte quasi-dépendance avec toutes les autres variables d'apport nutritionnel total. Comme le montre le tableau 1, le VIF pour les MCPPS peut être obtenu en multipliant le VIF pour les MCP par le coefficient d'ajustement $\zeta_k \rho_k$. Au tableau 3, les coefficients d'ajustement $\zeta_k \rho_k$ pour toutes les variables d'apport nutritionnel sauf les lipides sont inférieurs à 1, surtout celui des glucides, qui est de 0,46. Cela indique que les valeurs du VIF pour ces variables dans les MCPPS sont nettement plus faibles que celles dans les MCP, et que la colinéarité des variables explicatives dans le modèle a moins d'effet sur l'estimation des coefficients lorsque l'on utilise les MCPPS que quand on applique les MCP. Par contre, pour les variables d'apport nutritionnel reliées aux lipides, les $\zeta_k \rho_k$ sont tous plus grand que 1. Donc, la colinéarité entre ces variables est plus nuisible à l'estimation des coefficients sous les MCPPS que sous les MCP. Afin d'examiner ce problème de plus près, nous avons également ajusté un modèle ne contenant que deux variables d'apport nutritionnel : les lipides totaux et les acides gras mono-insaturés totaux. Les valeurs du VIF sous les MCPPS sont trois fois plus grandes que sous les MCO ou les MCP pour ces deux variables nutritionnelles. Si un analyste applique la méthode des MCPPS aux données d'enquête examinées ici et qu'il utilise les valeurs non ajustées du VIF fournies par les progiciels statistiques standard pour les MCO (présentées dans la première colonne) ou pour les MCP (présentées dans la deuxième colonne), son appréciation de la gravité de la colinéarité dans le modèle sera quelque peu erronée. Brièvement, même si les pentes estimées et les prédictions de la régression par les MCP et par les MCPPS sont les mêmes, les VIF risquent d'être sous-estimés ou surestimés si l'on ne tient pas compte des aspects complexes de l'enquête.

5. Conclusion

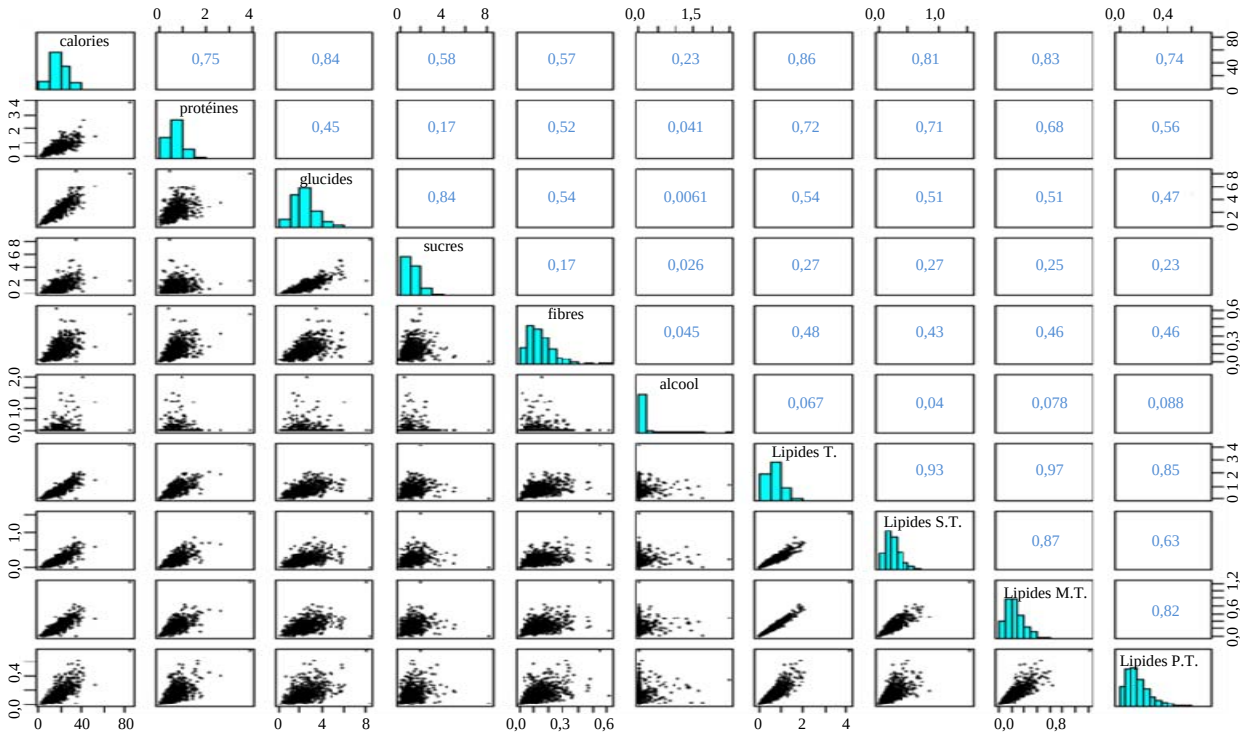
Afin qu'ils soient applicables à des modèles estimés d'après des données d'enquête, les diagnostics de régression doivent être adaptés de manière à tenir compte de la pondération et des caractéristiques du plan d'échantillonnage, telles que la stratification et la mise en grappes. Dans le présent article, nous avons élaboré une nouvelle formule pour calculer un facteur d'inflation de la variance (VIF) approprié pour les modèles linéaires. Un VIF est une mesure de l'augmentation, ou inflation, de la variance de l'estimateur d'un paramètre attribuable au fait que les variables explicatives sont corrélées au lieu d'être orthogonales. Bien que les progiciels standard permettent de produire des estimations de la pente d'une droite de régression pondérée par les poids de sondage en suivant des procédures qui reposent sur les moindres carrés pondérés, les VIF calculés par les routines développées pour des données ne provenant pas d'enquêtes sont incorrects. Le VIF pour un échantillon complexe est égal au VIF issu des moindres carrés pondérés multiplié par un facteur d'ajustement. Ce dernier est positif, mais peut être plus grand ou plus petit que 1, selon la nature des données analysées.

Dans une étude empirique, nous avons illustré l'application de notre nouvelle approche en utilisant des données provenant de la National Health and Nutrition Examination Survey de 2007-2008. Nous avons donné un exemple simple de la façon dont la colinéarité entre les variables explicatives influence l'estimation des coefficients dans la régression linéaire et démontré que, même si les coefficients estimés (et les valeurs prédites) sont les mêmes quand on utilise les moindres carrés pondérés ou les moindres carrés pondérés par les poids de sondage, leurs variances estimées et leurs valeurs du VIF (qui reflètent l'effet de la colinéarité sur l'estimation du coefficient) peuvent être différentes.

Tableau 1
Méthodes de régression et leurs statistiques de diagnostic de la colinéarité utilisées dans l'étude expérimentale

Type de régression	Matrice de pondération $\mathbf{W}^a$	Estimation de la variance de $\hat{\boldsymbol{\beta}}$	Formule du VIF
MCO	$\mathbf{I}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$	$\text{VIF} = \dfrac{1}{1-R^2_{m(k)}}$ [b]
MCP	$\mathbf{W}^c$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\text{VIF} = \dfrac{1}{1-R^2_{\text{PPS}m(k)}}$
MCPPS	$\mathbf{W}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\text{VIF} = \dfrac{\hat{\zeta}_k\hat{\rho}_{mk}}{1-R^2_{\text{PPS}m(k)}}$
		avec	avec $\hat{\zeta}_k = \dfrac{\mathbf{e}_{xk}^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}\mathbf{e}_{xk}}$,
		$\hat{\mathbf{V}} = \sum_{h=1}^{H}\dfrac{n_h}{n_h-1}\left[\text{Blkdiag}(\mathbf{e}_{hi}\,\mathbf{e}_{hi}^T) - \dfrac{1}{n_h}\mathbf{e}_h\,\mathbf{e}_h^T\right]$	$\hat{\rho}_{mk} = \dfrac{(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\hat{\mathbf{V}}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}$

[a] Dans tous les modèles de régression, les paramètres sont estimés par : $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$.
[b] $R^2_{m(k)}$ est le R-carré de la régression par les MCO de $\mathbf{x}_k$ sur les $\mathbf{x}$ dans la partie restante de $\mathbf{X}$ (en excluant une colonne pour l'ordonnée à l'origine).
[c] $\mathbf{W}$ est la matrice diagonale qui contient les poids de sondage w_i sur la diagonale principale.



[a] Lipides T. : lipides totaux
Lipides S.T. : acides gras saturés totaux
Lipides M.T. : acides gras mono-insaturés totaux
Lipides P.T. : acides gras poly-insaturés totaux

Figure 1 Diagrammes de dispersion par paire et coefficients de corrélation des variables nutritionnelles[a]

Tableau 2
Estimations des paramètres et leurs erreurs-types connexes pour trois méthodes de régression différentes

	Modèle complet					
	MCO		MCP		MCPPS	
Variable	**Bêta**	**E.-T.**	**Bêta**	**E.-T.**	**Bêta**	**E.-T.**
Ordonnée à l'origine	63,90***[a]	6,95	67,47***	6,36	67,47***	8,76
Âge	0,26	0,19	0,08	0,18	0,08	0,25
Race noire	10,39***	2,07	10,59***	2,38	10,59***	2,20
Calories	-6,41	5,76	-8,19	5,56	-8,19	5,75
Protéines	25,72	24,76	40,98	23,60	40,98	25,38
Glucides	26,67	23,93	32,31	22,96	32,31	22,65
Sucres	-1,90	3,06	-0,30	2,82	-0,30	4,06
Fibres	-41,17	20,23	-34,20	17,98	-34,20	19,05
Alcool	38,84	39,45	49,37	38,28	49,37	40,10
Lipides totaux	150,25*	69,53	161,78*	72,12	161,78	94,76
Acides gras saturés totaux	-113,20*	49,81	-101,40	56,26	-101,40	82,71
Acides gras mono-insaturés totaux	-72,05	48,03	-92,44	51,52	-92,44	83,51
Acides gras poly-insaturés totaux	-92,60*	46,13	-75,55	51,16	-75,55	78,76
	Modèle réduit					
	MCO		MCP		MCPPS	
Variable	**Bêta**	**E.-T.**	**Bêta**	**E.-T.**	**Bêta**	**E.-T.**
Ordonnée à l'origine	62,26***	6,88	67,52***	6,29	67,52***	8,48
Âge	0,27	0,19	0,07	0,18	0,07	0,25
Race noire	12,54***	1,98	11,74***	2,32	11,74***	2,05
Calories	0,15	0,10	0,23*	0,09	0,23*	0,10

[a] valeurs p de signification : * $p = 0{,}05$; ** $p = 0{,}01$; *** $p = 0{,}005$.

Tableau 3
Valeurs du VIF pour trois méthodes de régression différentes

	Modèle complet			
	MCO	MCP	MCPPS	
Variable	VIF	VIF	VIF	$\zeta_k \rho_k$
Âge	1,02	1,03	0,96	0,94
Race noire	1,10	1,07	1,12	1,05
Calories	3 411,61	3 562,70	2 740,83	0,77
Protéines	123,12	127,35	103,50	0,81
Glucides	1 074,87	1 007,40	462,08	0,46
Sucres	8,37	7,03	4,87	0,69
Fibres	4,59	3,94	2,37	0,60
Alcool	120,56	115,67	89,92	0,78
Lipides totaux	1 190,24	1 475,27	2 513,69	1,70
Acides gras saturés totaux	76,80	112,61	202,91	1,80
Acides gras mono-insaturés totaux	82,37	107,34	286,24	2,67
Acides gras poly-insaturés totaux	34,73	49,45	118,21	2,39
	Modèle réduit			
	MCO	MCP	MCPPS	
Variable	VIF	VIF	VIF	$\zeta_k \rho_k$
Âge	1,00	1,00	0,98	0,98
Race noire	1,02	1,01	0,97	0,96
Lipides totaux	20,10	20,22	63,15	3,12
Acides gras mono-insaturés totaux	20,16	20,26	61,57	3,04
	Modèle réduit			
	MCO	MCP	MCPPS	
Variable	VIF	VIF	VIF	$\zeta_k \rho_k$
Âge	1,00	1,00	0,98	0,97
Race noire	1,00	1,03	1,00	1,00
Calories	1,00	1,01	0,96	0,95

Les objectifs de l'analyse doivent être pris en considération pour décider de la façon d'utiliser les VIF. Si l'objectif principal est la prédiction, le fait d'inclure des variables colinéaires ou de choisir des variables incorrectes n'est pas très préoccupant. Si des conclusions plus fondamentales sont souhaitées, l'analyste devrait examiner quelles variables il serait logique d'inclure comme variables explicatives au lieu de se fier à un algorithme automatique pour sélectionner les variables. Les VIF sont un outil utile pour repérer les variables explicatives dont les coefficients estimés ont une variance inutilement grande. Quoiqu'ils puissent être considérés comme un outil pour la sélection automatique des variables, les simulations présentées dans Liao (2010), non reproduites ici, montrent que leur utilisation n'est pas un moyen fiable d'identifier un vrai modèle sous-jacent.

Remerciements

Les auteurs remercient le rédacteur associé et les examinateurs dont les commentaires ont apporté des améliorations importantes à l'article.

Annexe A

Détermination de g^{kk}

De façon analogue à la détermination du VIF pour les MCO classiques dans Theil (1971), la somme des carrés des écarts et la matrice des produits croisés $\mathbf{A} = \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$, qui peut être partitionné comme

$$\mathbf{A}_{p\times p} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T \tilde{\mathbf{X}}_{(k)} \\ \tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{x}}_k & \tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{X}}_{(k)} \end{pmatrix} \qquad (15)$$

où les colonnes de $\tilde{\mathbf{X}}$ sont réordonnées de manière que $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_k \; \tilde{\mathbf{X}}_{(k)})$ avec $\tilde{\mathbf{X}}_{(k)}$ désignant la matrice de dimensions $n \times (p-1)$ contenant toutes les colonnes sauf la k^{e} colonne de $\tilde{\mathbf{X}}$.

En utilisant la formule de l'inverse d'une matrice partitionnée, l'élément supérieur gauche de $\mathbf{A}^{-1}$ peut être exprimé par :

$$a^{kk} = \mathbf{i}_k^T \mathbf{A}^{-1} \mathbf{i}_k = \mathbf{i}_k^T (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{i}_k = \frac{1}{(1 - R^2_{\text{PPS}_{(k)}})\, \text{SST}_{\text{PPS}_{(k)}}} = \frac{1}{(1 - R^2_{\text{PPS}_{(k)}})\, \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \quad (16)$$

où

$$R^2_{\text{PPS}_{(k)}} = \frac{\hat{\beta}^T_{\text{PPS}_{(k)}} \tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{PPS}_{(k)}}}{\text{SST}_{\text{PPS}_{(k)}}},$$

avec $\hat{\beta}_{\text{PPS}_{(k)}} = (\tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{X}}_{(k)})^{-1} \tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{x}}_k$ est le coefficient de détermination correspondant à la régression de $\mathbf{x}_k$ sur les $p - 1$ autres variables explicatives. Le terme $\text{SST}_{\text{PPS}_{(k)}} = \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k$, est la somme totale des carrés des écarts dans cette régression.

Le terme $(1 - R^2_{\text{PPS}_{(k)}})^{-1}$ dans (16) est le VIF qui sera produit par les progiciels standard quand est exécutée une régression par les moindres carrés pondérés. Sous le modèle $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ avec $\boldsymbol{\epsilon} \sim (0, \sigma^2 \mathbf{W}^{-1})$, l'expression (16) est égale à $\text{Var}_M(\hat{\beta}_{\text{PPS}_k})/\sigma^2$. Cependant, elle n'est pas appropriée pour les régressions par les moindres carrés pondérés par les poids de sondage, parce que la variance de $\hat{\boldsymbol{\beta}}_{\text{PPS}}$ possède la forme plus complexe donnée par (2).

La matrice $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ peut être exprimée sous la forme :

$$\mathbf{G} = \begin{pmatrix} a^{kk} & \mathbf{a}^{k(k)} \\ \mathbf{a}^{(k)k} & \mathbf{A}^{(k)(k)} \end{pmatrix} \begin{pmatrix} b_{kk} & \mathbf{b}_{k(k)} \\ \mathbf{b}_{(k)k} & \mathbf{B}_{(k)(k)} \end{pmatrix} \begin{pmatrix} a^{kk} & \mathbf{a}^{k(k)} \\ \mathbf{a}^{(k)k} & \mathbf{A}^{(k)(k)} \end{pmatrix} \quad (17)$$

où la matrice inverse $\mathbf{A}^{-1} = [a^{hk}]$, $h, k = 1, ..., p$, $\mathbf{a}^{k(k)}$ est définie comme étant la k^{e} ligne de $\mathbf{A}^{-1}$ en excluant a^{kk}, $(a^{k1}, ..., a^{k(k-1)}, a^{k(k+1)}, ..., a^{kp})$, $\mathbf{a}^{(k)k} = [\mathbf{a}^{k(k)}]^T$ et $\mathbf{A}^{(k)(k)}$ est définie comme étant la partie $(k-1) \times (k-1)$ de la matrice $\mathbf{A}^{-1}$ en excluant les k^{e} ligne et colonne. La version partitionnée de $\mathbf{B}$ est

$$\mathbf{B} = \begin{pmatrix} b_{kk} & \mathbf{b}_{k(k)} \\ \mathbf{b}_{(k)k} & \mathbf{B}_{(k)(k)} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \\ \tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k & \tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \end{pmatrix}. \quad (18)$$

En vertu de la symétrie de $\mathbf{A}$ et de $\mathbf{B}$, le k^{e} élément diagonal de $\mathbf{G}$ est

$$g^{kk} = a^{kk}(a^{kk} b_{kk} + 2\mathbf{b}_{k(k)} \mathbf{a}^{(k)k}) + \mathbf{a}^{(k)kT} \mathbf{B}_{(k)(k)} \mathbf{a}^{(k)k}. \quad (19)$$

En utilisant l'inverse partitionnée de la matrice $\mathbf{A}$, qui représente $(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}$, on peut montrer que

$$\mathbf{a}^{(k)k} = -a^{kk} (\tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{X}}_{(k)})^{-1} \tilde{\mathbf{X}}_{(k)} \tilde{\mathbf{x}}_k = -a^{kk} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)}. \quad (20)$$

En substituant $a^{(k)k}$ dans (19), g^{kk} peut être exprimé de manière compacte en fonction de a^{kk}, $\hat{\boldsymbol{\beta}}_{\text{PPS}(k)}$ et la composante inférieure droite de la matrice $\mathbf{B}$:

$$\begin{aligned} g^{kk} &= (a^{kk})^2 \left(b_{kk} - 2\mathbf{b}_{k(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} + \hat{\boldsymbol{\beta}}^T_{\text{PPS}(k)} \mathbf{B}_{(k)(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right) \\ &= a^{kk} \times \frac{1}{1 - R^2_{\text{PPS}(k)}} \frac{1}{\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \\ &\quad \times \left(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k - 2\tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} + \hat{\boldsymbol{\beta}}^T_{\text{PPS}(k)} \tilde{\mathbf{X}}_{(k)} \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right) \\ &= a^{kk} \times \frac{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)^T \tilde{\mathbf{V}} \left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)}{\left(1 - R^2_{\text{PPS}(k)} \right) \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \\ &= a^{kk} \times \frac{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)^T \tilde{\mathbf{V}} \left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)}{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)^T \left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)} \\ &= \frac{1}{1 - R^2_{\text{PPS}(k)}} \frac{1}{\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \frac{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{V}} \tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{e}}_{xk}}, \end{aligned} \quad (21)$$

où $\tilde{\mathbf{e}}_{xk} = \tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)}$ est le résidu de la régression de $\tilde{\mathbf{x}}_k$ sur $\tilde{\mathbf{X}}_{(k)}$.

Bibliographie

Belsley, D.A. (1984). Collinearity and forecasting. *Journal of Forecasting*, 38, 73-93.

Belsley, D.A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York : John Wiley & Sons, Inc.

Belsley, D.A., Kuh, E. et Welsch, R.E. (1980). *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York : Wiley Interscience.

Elliot, M. (2007). Réduction bayésienne des poids pour les modèles de régression linéaire généralisée. *Techniques d'enquête*, 33, 27-40.

Farrar, D.E., et Glauber, R.R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.

Fox, J. (1984). *Linear Statistical Models and Related Methods, With Applications to Social Research*. New York : John Wiley & Sons, Inc.

Fox, J., et Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417), 178-183.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28(1), 5-25.

Kmenta, J. (1986). *Elements of Econometrics*. New York : Macmillan, 2e Éd.

Li, J. (2007a). Linear regression diagnostics in cluster samples. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3341-3348.

Li, J. (2007b). Regression diagnostics for complex survey data. Unpublished doctoral dissertation, University of Maryland. Available at http://drum.lib.umd.edu/bitstream/1903/7598/1/umi-umd-4863.pdf.

Li, J., et Valliant, R. (2009). Matrice chapeau et effets de levier pondérés par les poids de sondage. *Techniques d'enquête*, 35(1), 17-27.

Li, J., et Valliant, R. (2011). Linear regression influence diagnostics for unclustered survey data. *Journal of Official Statistics*, 20, 99-119.

Liao, D. (2010). *Collinearity Diagnostics for Complex Survey Data*. Thèse de doctorat non-publiée, University of Maryland. Disponible au http://drum.lib.umd.edu/bitstream/1903/10881/1/Liao_umd_0117E_11537.pdf.

Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, 13, 1517-1520.

Neter, J., Kutner, M., Wasserman, W. et Nachtsheim, C. (1996). *Applied Linear Statistical Models*. New York : McGraw-Hill/Irwin, 4e Éd.

Simon, S.D., et Lesage, J.P. (1988). The impact of collinearity involving the intercept term on the numerical accuracy of regression. *Computer Science in Economics and Management*, 1, 137-152.

Smith, G. (1974). Multicollinearity and forecasting. Yale Univeristy Cowles Foundation Disscussion Papier No.383. Disponible au http://cowles.econ.yale.edu/P/cd/d03b/d0383.pdf.

Steward, G.W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(1), 68-84.

Theil, H. (1971). *Principles of Econometrics*. New York : John Wiley & Sons, Inc.