Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Diagnostic de colinéarité dans les modèles linéaires généralisés ajustés aux données d’enquête

par Dan Liao et Richard Valliant

Date de diffusion : le 23 décembre 2025

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, décembre 2025
Vol. 51, nº 2, pp. 613-643
Statistique Canada, nº 12-001-X au catalogue

Diagnostic de colinéarité dans les modèles linéaires généralisés ajustés aux données d'enquête

Dan Liao et Richard Valliant[1]

Résumé

La catégorie des modèles linéaires généralisés (GLM) est une généralisation flexible de la régression par les moindres carrés ordinaires qui permet de relier le modèle linéaire à la variable réponse à l'aide d'une fonction de lien et suppose que l'ampleur de la variance de chaque mesure est fonction de sa valeur prédite. La multicolinéarité dans les GLM peut faire augmenter les variances des coefficients estimés et peut provoquer une mauvaise prédiction dans certaines régions de l'espace de régression. Elle peut également entraîner une statistique de Wald non significative, même lorsque les variables explicatives sont hautement prédictives dans un modèle de la famille des GLM. Peu de recherches ont étudié de près le diagnostic de multicolinéarité dans les GLM, en particulier lorsque des données d'enquête complexes sont utilisées. Dans le présent article, nous développons des facteurs d'inflation de la variance (VIF) qui mesurent l'augmentation de la variance d'un estimateur de paramètre due à la multicolinéarité dans les GLM. De plus, nous étendons les VIF et les indices de conditionnement pour les appliquer à des données d'enquête complexes, en tenant compte des caractéristiques du plan, par exemple les poids, les grappes et les strates. Ces méthodes sont illustrées au moyen de données provenant d'une enquête auprès des ménages sur la santé et la nutrition.

Mots-clés : Diagnostics de données d'enquête; facteurs d'inflation de la variance; indices de conditionnement; multicolinéarité; régression avec données d'enquêtes complexes.

1. Introduction

La catégorie des modèles linéaires généralisés (GLM) (McCullagh et Nelder, 1989; McCulloch et Searle, 2001) est une généralisation flexible de la régression par les moindres carrés ordinaires qui permet de relier le modèle linéaire à la variable réponse à l'aide d'une fonction de lien et suppose que l'ampleur de la variance de chaque mesure est fonction de sa valeur prédite. La colinéarité dans les GLM peut faire augmenter les variances des coefficients estimés et entraîner une mauvaise prédiction dans certaines régions de l'espace de régression. Hauck et Donner (1977) ont aussi souligné que la colinéarité peut entraîner une statistique de Wald non significative même lorsque les variables explicatives sont hautement prédictives dans un modèle logistique, et Væth (1985) a souligné que cet effet de colinéarité peut se produire pour les autres membres de la famille des modèles linéaires généralisés.

Au cours des décennies antérieures, plusieurs auteurs ont examiné les problèmes de colinéarité dans les modèles logistiques (voir Schaefer, Roi et Wolfe, 1984; Schaefer, 1986; Marx et Smith, 1990b). D'autres se sont penchés sur ce problème dans le cadre des GLM (voir Mackinnon et Puterman, 1989; Marx et Smith, 1990a; Weissfeld et Sereika, 1991; Lesaffre et Marx, 1993). Tous ces articles soulignent que la colinéarité dans les GLM n'est pas simplement due à des relations colinéaires entre les variables explicatives (ou, de manière équivalente, entre les colonnes de la matrice de plan $\mathbf{X}$), mais plutôt à un mauvais conditionnement de la matrice d'information observée. Lesaffre et Marx (1993), suivant une proposition de Mackinnon et

1. Dan Liao, RTI International, 701 13th Street, N.W., Suite 750, Washington DC, 20005; Richard Valliant, professeur émérite, University of Michigan et University of Maryland, 4620 N Park Ave, #1406W, Chevy Chase MD 20815. Courriel : valliant@umich.edu.

Puterman (1989) concernant le mauvais conditionnement, ont étudié la dépendance des problèmes de mauvais conditionnement vis-à-vis la valeur particulière du paramètre de régression $\boldsymbol{\beta}$. Ils ont conclu que dans les GLM, la réponse et le choix du modèle contribuent également au degré de colinéarité (mauvais conditionnement) de la matrice d'information. Le terme « ML-collinearity » (multicolinéarité) est employé dans leur article pour décrire le scénario d'une estimation par le maximum de vraisemblance du paramètre du modèle où il existe une colinéarité entre les variables explicatives pondérées; les poids sont propres à la forme de la fonction de lien et à la variance du modèle dans les GLM, et ils sont différents des poids de sondage.

Le facteur d'inflation de la variance (VIF) est l'une des techniques conventionnelles de diagnostic de colinéarité les plus populaires pour les régressions linéaires par les moindres carrés ordinaires ou pondérés. Bien que l'idée soit assez ancienne (voir par exemple Theil, 1971, page 166), les VIF sont encore largement utilisés. Dans les sciences appliquées (par exemple économie, science du sol, sociologie, marketing et science de l'environnement), les articles récents où les VIF sont utilisés dans les analyses de régression linéaire sont ceux d'Abaidoo et Agyapong (2024), d'Alhakim-Naser, Hasan et Zaifoglu (2024), de Dobos et Sasvari (2024), de Gao, Xu et Pan (2024), de Parasyris, Stankovic et Stankovic (2024) et de Zhang, Li, Li et Yin (2024).

Un VIF mesure l'inflation de la variance de l'estimation d'une pente, $\hat{\beta}_k$, due à la non-orthogonalité des variables explicatives en comparant la variance de modèle de $\hat{\beta}_k$ dans la régression multivariée avec celle obtenue lorsque la k^{e} variable explicative est orthogonale à toutes les autres variables explicatives, ou, de manière équivalente, s'il n'y avait que la k^{e} variable explicative dans la régression. Dans une régression par les moindres carrés ordinaires, le VIF est $1/(1-R_k^2)$, où R_k^2 est le carré de la corrélation multiple provenant de la régression de la k^{e} variable explicative par rapport aux autres variables explicatives. Fox et Monette (1992) ont présenté une généralisation des VIF des modèles linéaires qui est disponible dans les paquets R `car` (Fox, Weisberg, Price, Adler, Bates, Baud-Bovy, Bolker, Ellison, Firth, Friendly, Gorjanc, Graves, Heiberger, Krivitsky, Laboissiere, Maechler, Monette, Murdoch, Nilsson, Ogle, Ripley, Short, Venables, Walker, Winsemius et Zeileis, 2023) et `glmtoolbox` (Vanegas, Rondón et Paula, 2024). Les indices de conditionnement d'une matrice des variables explicatives constituent un autre outil qui peut être utilisé pour diagnostiquer les problèmes de colinéarité dans les modèles linéaires (voir par exemple Belsley, Kuh et Welsch, 1980, chapitre 3), et ils peuvent être calculés dans les paquets `klaR` (Roever, Raabe, Luebke, Ligges, Szepannek, Zentgraf et Meyer, 2023) et `olsrr` (Hebbali, 2024). Liao et Valliant (2012b) et Liao et Valliant (2012a) ont étendu aux données d'enquête complexes les VIF et les indices de conditionnement utilisés dans les modèles linéaires.

Lors de leur introduction, les VIF et les indices de conditionnement n'étaient pas seulement utilisés comme outils pour détecter les problèmes de colinéarité dans les modèles, mais aussi souvent utilisés pour sélectionner le modèle en éliminant les variables explicatives qui affichaient des corrélations élevées avec d'autres covariables dans un modèle. Cependant, les progrès réalisés depuis lors sur le plan de la sélection du modèle – en particulier avec l'essor des algorithmes d'apprentissage automatique – font remettre en

question le maintien de l'utilisation des VIF comme outil essentiel pour la sélection des modèles. Les VIF et les indices de conditionnement n'en demeurent pas moins utiles à l'analyse exploratoire, car ils aident les analystes à détecter les problèmes de colinéarité et à comprendre les relations entre les variables explicatives et leurs effets individuels sur un modèle. Ces deux diagnostics détectent la multicolinéarité, qui peut fausser l'interprétation de l'effet d'une variable explicative unique dans les modèles de régression. En revanche, les algorithmes d'apprentissage automatique sont mieux adaptés à la réduction de dimension lorsque le but principal est la prédiction plutôt que l'inférence. Les techniques telles que la régularisation, qui imposent des pénalités en cas de surajustement (par exemple lasso, méthode de régression ridge) ou de sélection des caractéristiques, visent à réduire la redondance des caractéristiques afin d'améliorer le rendement et la généralisation du modèle plutôt qu'à comprendre les effets particuliers des variables explicatives. Ainsi, les VIF et les indices de conditionnement sont mieux adaptés pour comprendre le rôle de variables explicatives précises, tandis que l'apprentissage automatique excelle dans l'amélioration des prédictions.

Dans la pratique, les analystes utilisent également les VIF comme indicateurs du degré d'inflation de la variance et comme diagnostic des problèmes de colinéarité lorsqu'un modèle de la famille des GLM est utilisé. Cependant, comme nous le révélons dans la section 2, le terme $1/(1-R_k^2)$ n'est ni un moyen précis de mesurer l'inflation de la variance dans les GLM, ni un moyen adéquat de refléter la relation de colinéarité sous-jacente dans les GLM. Nous calculons les VIF des GLM à partir d'un estimateur de variance par linéarisation pour les GLM qui est couramment utilisé dans les analyses statistiques. Le VIF nouvellement calculé est déterminé par la relation entre les variables explicatives pondérées dans la matrice d'information observée et reflète le degré d'inflation de la variance causé par la coexistence d'autres variables explicatives dans le modèle.

Pour appliquer les VIF et les indices de conditionnement des GLM aux données d'enquête complexes, un certain nombre de facteurs doivent être pris en considération. En règle générale, les données d'enquête complexes proviennent de plans de sondage complexes ou reflètent des structures de population sous-jacentes complexes, comme définies par Skinner, Holt et Smith (1989). Le plan de sondage, par exemple, peut comporter un échantillonnage à plusieurs degrés et entraîner une corrélation intragrappe positive pour une variable à l'étude donnée parmi les unités de l'échantillon. Autre possibilité : l'adoption d'un échantillonnage stratifié pour lequel différentes structures de variance-covariance moyenne et résiduelle peuvent convenir dans des strates différentes. Des poids inégaux sont souvent inévitables dans les données d'enquête en raison de la variabilité des probabilités d'inclusion et de la repondération qui tient compte des ajustements pour les non-réponses ou les non-contacts, de la poststratification et des ajustements par calage. Pour tenir compte de ces divers aspects complexes, les analystes peuvent utiliser des méthodes spécialisées d'analyse de régression afin d'estimer les paramètres et les erreurs-types, comme l'expliquent Skinner et coll. (1989) dans *Analysis of Complex Surveys* et Chambers et Skinner (2003) dans *Analysis of Survey Data*. Si le modèle de régression pour les données de l'échantillon est le même que pour l'ensemble de la population, le plan d'échantillonnage est souvent qualifié d'*ignorable* et le modèle peut être ajusté légitimement sans utiliser de poids de sondage. En revanche, si le modèle d'échantillonnage s'écarte du modèle de la population, le

plan est *non-ignorable* ou *informatif* (Pfeffermann et Sverchkov, 2003). Dans ce dernier cas, des poids d'échantillonnage sont utilisés au moment d'ajuster le modèle afin de garantir que le modèle de la population est bien estimé. Dans le présent article, nous traitons à la fois des diagnostics non pondérés et pondérés.

Le reste de l'article est organisé comme suit. Dans la section 2, nous établissons des VIF fondés sur un modèle pour des données ne provenant pas d'enquêtes. Dans la section 3, nous examinons le cas d'un analyste qui estime les paramètres de modèle d'un GLM à l'aide de poids de sondage (SWGLM) afin de tenir compte de la complexité des données d'enquête, et par conséquent, nous établissons des VIF et des indices de conditionnement adaptés aux SWGLM. Notre approche prévoit l'élaboration d'estimateurs qui offrent à la fois une interprétation du modèle et une interprétation du plan, comme nous le décrivons dans la section 3. Dans la quatrième section, nous présentons des illustrations numériques des techniques, et la conclusion figure dans la cinquième section.

2. Le diagnostic de colinéarité dans le modèle linéaire généralisé

Dans cette section, nous passons en revue les GLM et présentons des résultats du diagnostic de colinéarité fondé sur un modèle pour des données ne provenant pas d'enquêtes. Dans la section 2.1, nous présentons les définitions et la notation à utiliser, qui sont principalement les mêmes que celles de McCullagh et Nelder (1989). Dans la section 2.2, nous passons en revue les variances d'estimateurs de paramètres dans des modèles à large échantillon, et nous établissons les VIF associés.

2.1 Définitions et notations

Supposons que les données d'une population finie de N éléments soient générées comme la réalisation d'un modèle de superpopulation. Le vecteur de dimension N des variables de réponse $\mathbf{y}_{N\times 1} = (y_1,\ldots,y_N)$ est censé contenir des mesures indépendantes dans un modèle avec une moyenne μ_i et une densité issue de la famille exponentielle. Les calculs effectués par rapport à un modèle se verront assigner l'indice M; les calculs fondés sur le plan se verront assigner l'indice π. La densité de y_i peut s'écrire :

$$f_{Y_i}(y_i;\theta_i,\tau) \;=\; \exp\left\{\left[y_i\theta_i - b(\theta_i)\right]\big/\tau^2 - c(y_i,\tau)\right\}, \quad i=1,\ldots,N; \tag{2.1}$$

où, pour plus de commodité, nous avons écrit la densité dans ce qui est appelé sa *forme canonique*, $b(\cdot)$ et $c(\cdot)$ étant des fonctions connues et τ étant un paramètre dérangeant inconnu qui est constant pour tous les y_i. Comme le montrent McCullagh et Nelder (1989, section 2.2.2), $E_M(y_i) = \partial b(\theta_i)/\partial\theta_i \equiv \mu_i$ et $\mathrm{var}_M(y_i) = \tau^2\partial^2 b(\theta_i)/\partial\theta_i^2 \equiv \tau^2 v(\mu_i)$, où $v(\mu_i) = \partial^2 b(\theta_i)/\partial\theta_i^2$. Ainsi, la moyenne et la variance de y_i ne font intervenir que le paramètre naturel θ_i et le paramètre dérangeant τ.

Voici des exemples du modèle (2.1). Si y_i est normalement distribué avec la moyenne μ et la variance connue σ_o^2, alors $\theta_i = \mu$, $\tau^2 = \sigma_o^2$, $b(\theta_i) = \mu^2$ et $c(y_i,\tau) = -\left(\frac{y^2}{2\sigma_o^2} + \log\left(\sqrt{2\pi}\sigma_o\right)\right)$. Si y_i est binomial avec x succès dans les n essais, chaque essai ayant une probabilité de succès p, alors $y_i = x$,

$\theta_i = \log\big(p/(1-p)\big)$, $\tau^2 = 1$, $b(\theta_i) = n\log(1-p)$ et $c(y_i,\tau) = \log\binom{n}{x}$. D'autres lois de la famille exponentielle ainsi que certaines des notations qui seront utilisées plus loin et qui sont associées à chaque loi sont énumérées dans le tableau 2.1

Pour tenir compte des variables auxiliaires qui sont prédictives de y, une fonction de *lien* est utilisée, reliant les variables auxiliaires à la moyenne μ_i comme nous le décrivons ci-dessous. Supposons que p variables explicatives (des valeurs de x) soient disponibles pour chaque élément de la population. La matrice $N \times p$ des variables explicatives de tous les éléments de la population est $\mathbf{X}$. La rangée i de $\mathbf{X}$ est désignée par $\dot{\mathbf{x}}_i$, qui est un vecteur de dimension p. La k^{e} colonne de $\mathbf{X}$ est un vecteur de dimension N désigné par $\mathbf{x}_k$ et correspond à la variable explicative k. La matrice $\mathbf{X}$ peut être écrite de deux façons avec cette notation : $\mathbf{X} = (\dot{\mathbf{x}}_1,\ldots,\dot{\mathbf{x}}_N)^T$ ou $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_p)$. Définissons la variable explicative linéaire η_i, qui est une transformation de la moyenne μ_i, par :

$$\eta_i = \dot{\mathbf{x}}_i\boldsymbol{\beta} = g(\mu_i) \tag{2.2}$$

où $\boldsymbol{\beta}$ est un vecteur de dimension p de paramètres, $\eta_i = \theta_i$ avec la restriction supplémentaire qu'il s'agit d'une combinaison linéaire de β comme dans (2.2), et où $g(\cdot)$ est une fonction de lien qui est monotone et deux fois dérivable à l'intérieur d'un intervalle de nombres réels. (Bien que nous puissions utiliser uniquement la notation θ_i, la convention dans la littérature sur les GLM, que nous suivons dans le cas présent, consiste à utiliser θ_i dans la définition de la densité de la famille exponentielle et η_i pour préciser la valeur explicative linéaire.) Des exemples de fonctions de lien sont présentés dans le tableau 2.1 : l'identité pour la loi normale, le logarithme pour la loi de Poisson et le logit ou le probit pour la loi binomiale.

La méthode du maximum de vraisemblance est utilisée pour estimer successivement les paramètres de régression $\boldsymbol{\beta}$ et leurs fonctions, les variables explicatives linéaires et les valeurs prédites (McCullagh et Nelder, 1989; McCulloch et Searle, 2001). Comme le montrent McCullagh et Nelder (1989, section 2.5), les estimations du maximum de vraisemblance (EMV) de $\boldsymbol{\beta}$ peuvent être obtenues par un type de moindres carrés repondérés itératifs. Le logarithme de la vraisemblance de la population dans le modèle linéaire généralisé est la somme du logarithme naturel de chacune des composantes définies par (2.1) pour les N observations :

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N}\big[y_i\theta_i - b(\theta_i)\big]\big/\tau^2 - \sum_{i=1}^{N} c(y_i,\tau). \tag{2.3}$$

En faisant correspondre à zéro les données partielles du premier degré pour $\boldsymbol{\beta}$, les équations des estimations du maximum de vraisemblance (EMV) pour $\boldsymbol{\beta}$ peuvent être écrites sous forme de notation matricielle comme

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{\tau^2}\mathbf{X}^T\boldsymbol{\Gamma}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{0}^T, \tag{2.4}$$

avec $\boldsymbol{\Gamma} = \text{diag}(\gamma_i)$, $\gamma_i = \big\{v(\mu_i)[g'(\mu_i)]^2\big\}^{-1}$, $\boldsymbol{\Delta} = \text{diag}(g'(\mu_i))$, $\mathbf{y} = (y_i)_{N\times 1}$ et $\boldsymbol{\mu} = (\mu_i)_{N\times 1}$. La solution à (2.4) est l'EMV de la population de $\boldsymbol{\beta}$ et est désignée par $\hat{\boldsymbol{\beta}}$.

Tableau 2.1
Caractéristiques de certaines lois univariées courantes dans la famille exponentielle

	Normale	Poisson	Binomiale		Gamma	Gaussienne inverse
Notation	$N(\mu_i,\sigma^2)$	$P(\mu_i)$	$B(m,\mu_i)/m$		$G(\mu_i,v)$	$GI(\mu_i,\sigma^2)$
Intervalle de y	$(-\infty,\infty)$	$(0(1),\infty)$	$\frac{(0(1),m)}{m}$		$(0,\infty)$	$(0,\infty)$
τ^2	σ^2	1	$1/m$		v^{-1}	σ^2
Fonction de lien	Identité	Logarithme	Logit	Probit	Réciproque	μ_i^{-2}
Lien canonique $g(\mu_i)$	μ_i	$\log(\mu_i)$	$\log(\frac{\mu_i}{1-\mu_i})$	[a] $\Phi^{-1}(\mu_i)$	μ_i^{-1}	μ_i^{-2}
Fonction de variance $v(\mu_i)$	1	μ_i	$\mu_i(1-\mu_i)$	$\mu_i(1-\mu_i)$	μ_i^2	μ_i^3
[b] $g'(\mu_i)$	1	μ_i^{-1}	$[\mu_i(1-\mu_i)]^{-1}$	[c] $\varphi^{-1}(\dot{\mathbf{x}}_i^T\boldsymbol{\beta})$	$-\mu_i^{-2}$	$-2\mu_i^{-3}$
[d] $\gamma_i=\{v(\mu_i)[g'(\mu_i)]^2\}^{-1}$	1	μ_i	$\mu_i(1-\mu_i)$	$\frac{\varphi^2(\dot{\mathbf{x}}_i^T\boldsymbol{\beta})}{\mu_i(1-\mu_i)}$	μ_i^2	$\frac{1}{4}\mu_i^3$
[e] $\gamma_i g'(\mu_i)$	1	1	1	$\frac{\varphi(\dot{\mathbf{x}}_i^T\boldsymbol{\beta})}{\mu_i(1-\mu_i)}$	-1	$-\frac{1}{2}$

Notes : [a] Φ est la fonction de distribution cumulative de la loi normale centrée réduite.
[b] $\boldsymbol{\Delta}=\text{diag}[g'(\mu_i)]$.
[c] $\varphi=\Phi'$ est la fonction de distribution de probabilités de la loi normale centrée réduite et $\mathbf{x}_i^T\boldsymbol{\beta}=g(\mu_i)$.
[d] $\boldsymbol{\Gamma}=\text{diag}(\gamma_i)$.
[e] $\boldsymbol{\Gamma\Delta}=\text{diag}[\gamma_i g'(\mu_i)]$.

Il existe des cas où l'EMV est introuvable. Si $\boldsymbol{\beta}$ se situe dans les limites de P, il existe au moins un i, où $g^{-1}(\dot{\mathbf{x}}_i\boldsymbol{\beta})=\mu_i=\mu_{\min}$ ou $\mu_{\max}$, comme nous l'avons défini précédemment. Alors, $g'(\mu_i)=0$, car g est une fonction monotome et deux fois dérivable. Par conséquent, γ_i dans (2.4) sera infini et l'EMV n'existera pas. Cette situation peut se produire lorsque y_i est binaire (0, 1) si toutes les unités sont 0 ou 1 pour une configuration particulière de $\mathbf{X}$ – une condition connue sous le nom de séparation complète. Un autre inconvénient des EMV est qu'elles peuvent afficher un piètre rendement avec de petits échantillons.

2.2 La variance asymptotique de l'estimateur de paramètres et son facteur d'inflation de la variance

Dans l'inférence basée sur un modèle, la matrice de variance-covariance asymptotique de $\hat{\boldsymbol{\beta}}$ est (voir McCulloch et Searle, 2001) :

$$\text{avar}_M(\hat{\boldsymbol{\beta}}) = \tau^2(\mathbf{X}^T\boldsymbol{\Gamma}\mathbf{X})^{-1} \tag{2.5}$$

où avar_M représente la variance de modèle limite ou asymptotique. Dans les modèles linéaires généralisés, le paramètre d'intérêt, $\boldsymbol{\beta}$, est estimé et utilisé à son tour pour estimer l'espérance de y_i, μ_i, conditionnellement à $\mathbf{x}_i$. Désignons la valeur estimée μ_i par $\hat{\mu}_i$, $\hat{\gamma}_i=\{v(\hat{\mu}_i)[g'(\hat{\mu}_i)]^2\}^{-1}$ avec $g'(\hat{\mu}_i)=\partial g(\hat{\mu}_i)/\partial\hat{\mu}_i$ et $\hat{\boldsymbol{\Gamma}}=\text{diag}(\hat{\gamma}_i)$.

Avec $\hat{\mathbf{A}}=\mathbf{X}^T\hat{\boldsymbol{\Gamma}}\mathbf{X}$, la matrice de variance-covariance asymptotique de $\hat{\boldsymbol{\beta}}$ peut être estimée par :

$$v_M(\hat{\boldsymbol{\beta}}) = \hat{\tau}^2\hat{\mathbf{A}}^{-1}, \tag{2.6}$$

et son k^{e} élément sur la diagonale principale est la variance de modèle asymptotique de $\widehat{\beta}_k$, qui, de manière équivalente, peut être estimée par :

$$v_M(\widehat{\beta}_k) = \hat{\tau}^2 a^{kk}, \tag{2.7}$$

où a^{kk} est le k^{e} élément sur la diagonale principale de la matrice $\widehat{\mathbf{A}}^{-1}$. Pour les familles Bernoulli et Poisson, τ est identiquement égal à 1. Pour les familles exponentielles gamma, normale et gaussienne inverse, τ est un paramètre inconnu. L'estimation de τ peut dépendre des covariables incluses dans le modèle et peut être nécessaire ou non, selon le modèle. Aux fins de l'analyse dans le cas présent, aucune estimation n'est nécessaire, car le VIF présenté ci-dessous ne comprend pas τ.

La dérivation du VIF dans l'expression (2.10) ci-dessous suit de près les étapes décrites dans Liao et Valliant (2012b, annexe A). L'annexe A.1 du présent article expose certains détails relatifs aux GLM et aux SWGLM. Dans le cas où il n'y a pas de poids de sondage, a^{kk} peut s'exprimer en fonction du coefficient de détermination d'une régression par les moindres carrés pondérés de $\mathbf{x}_k$ sur les autres x :

$$a^{kk} = \mathbf{i}_k^T \widehat{\mathbf{A}}^{-1} \mathbf{i}_k = \mathbf{i}_k^T (\mathbf{X}^T \widehat{\mathbf{\Gamma}} \mathbf{X})^{-1} \mathbf{i}_k = \frac{1}{(1 - R^2_{\Gamma(k)}) \mathbf{x}_k^T \widehat{\mathbf{\Gamma}} \mathbf{x}_k} \tag{2.8}$$

où $\mathbf{i}_k$ est un vecteur de dimension p où 1 figure à la position k et des zéros figurent ailleurs; et $R^2_{\Gamma(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{\Gamma(k)} \mathbf{X}^T_{(k)} \widehat{\mathbf{\Gamma}} \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{\Gamma(k)}}{\mathbf{x}_k^T \widehat{\mathbf{\Gamma}} \mathbf{x}_k}$ où $\widehat{\beta}_{\Gamma(k)} = (\mathbf{X}^T_{(k)} \widehat{\mathbf{\Gamma}} \mathbf{X}_{(k)})^{-1} \mathbf{X}^T_{(k)} \widehat{\mathbf{\Gamma}} \mathbf{x}_k$ et $\mathbf{X}_{(k)}$ sont la matrice $\mathbf{X}$, la k^{e} colonne étant omise. Le terme $R^2_{\Gamma(k)}$ est un coefficient de détermination et varie de 0 à 1.

Lorsqu'il n'y a que $\mathbf{x}_k$ dans la régression, selon (2.5), l'estimateur de variance asymptotique de $\widehat{\beta}_k$ est :

$$v_M(\widehat{\beta}_k) = \hat{\tau}^2 (\mathbf{x}_k^T \widehat{\mathbf{\Gamma}} \mathbf{x}_k)^{-1}. \tag{2.9}$$

En comparant (2.8) et (2.9) et en notant $v_M(\widehat{\beta}_k)$ dans (2.6), nous pouvons constater que l'estimateur de variance asymptotique fondé sur un modèle de $\widehat{\beta}_k$ est gonflé par un multiple de

$$\text{VIF}_{M(k)} = \frac{1}{1 - R^2_{\Gamma(k)}} \tag{2.10}$$

dans (2.8) en prenant toutes les autres $p-1$ variables explicatives dans la régression avec $\mathbf{x}_k$. Il convient de mentionner que la régression de $\mathbf{x}_k$ sur les autres variables explicatives $\mathbf{x}$ utilise des poids $\widehat{\mathbf{\Gamma}}$, mais que ceux-ci ne sont pas des poids de sondage. Ce facteur d'inflation de la variance est toujours supérieur ou égal à 1, en raison de l'intervalle de $R^2_{\Gamma(k)}$. Pour les données ne provenant pas d'enquêtes, Özkale (2021, section 3.1) propose une mesure de VIF pour les GLM, similaire à (2.10), qui repose sur des éléments diagonaux de l'inverse de la matrice de corrélation pour les variables explicatives x, ce qui se rapproche de la proposition de Chatterjee et Price (1991) pour les modèles linéaires. Son VIF est fondé sur la matrice complète centrée et échelonnée $\mathbf{X}$, suivant une proposition de Smith et Marx (1990). De plus, Özkale a étendu aux GLM l'indicateur de l'efficacité relative du plan (RED) – élaboré à l'origine pour la régression linéaire dans Kovács, Petres et Tóth (2005). Elle a également proposé des VIF corrigés fondés sur des

estimations de la méthode de régression ridge, qui contribuent à stabiliser les estimations de la variance en présence de multicolinéarité en introduisant un paramètre de régularisation.

Dans le modèle (2.1), la variance de y_i dans le GLM est proportionnelle à $v(\mu_i)$. Par exemple, la variance de y_i dans un modèle logistique est $p_i(1-p_i)$ selon l'hypothèse de la loi de Bernoulli. Cette hypothèse n'est pas respectée lorsque les données sont regroupées dans la population, ce qui entraîne une *surdispersion*, c'est-à-dire que la variance de y_i dépasse la variance $v(\mu_i)$. Une méthode couramment utilisée pour aborder l'échantillonnage en grappe selon une inférence basée sur un modèle consiste à utiliser des modèles linéaires mixtes généralisés (MLMG) en traitant les grappes comme des effets aléatoires. Agresti (2002, chapitre 12) et Shao (2003, chapitre 4) analysent plus en détail cette approche. Dans le présent article, nous ne traitons que les GLM généraux et nous réservons à de futures recherches les diagnostics appropriés des MLMG. Dans la section suivante, nous examinons toutefois un autre type d'échantillonnage en grappe : celui qui est induit par un plan d'échantillonnage.

3. Adaptations aux GLM pondérés par les poids de sondage

Les résultats obtenus à la section 2 pour des données ne provenant pas d'enquêtes peuvent être adaptés pour être appliqués à des enquêtes complexes, comme nous le décrivons dans la présente section. Dans cette section, nous présentons à la fois les résultats fondés sur un modèle et ceux de type modèle-plan en ce qui concerne les VIF. Par estimateurs « modèle-plan », nous faisons référence aux estimateurs qui maintiennent l'efficacité et l'intelligibilité associées aux méthodes fondées sur un modèle, tout en offrant une certaine robustesse relativement aux caractéristiques d'un plan de sondage complexe. Cette approche s'apparente au concept d'inférence sensible au plan de Kott (2018), qui conserve les techniques de l'analyse fondée sur le plan (telles que les équations d'estimation pondérées et les estimateurs de la variance de type sandwich) dans un cadre fondé sur un modèle. Selon cette approche, les estimations sont traitées comme des paramètres de superpopulation, les poids étant utilisés pour renforcer la robustesse de l'ajustement du modèle, tandis que dans les méthodes fondées sur le plan, les estimations sont considérées comme des quantités de population finie. Que les estimations soient considérées selon la perspective de l'approche fondée sur le plan ou de celle sensible au plan, nos estimateurs de VIF de type modèle-plan restent intelligibles pour évaluer la multicolinéarité dans le cadre conceptuel correspondant.

3.1 Plan d'échantillonnage

Supposons qu'un plan d'échantillonnage stratifié à plusieurs degrés soit utilisé. Il y a $h=1,\ldots,H$ strates dans la population, $i=1,...,M_h$ grappes dans la strate h et $t=1,\ldots,N_{hi}$ éléments dans la grappe hi. Nous sélectionnons $i=1,\ldots,m_h$ grappes dans la strate h et $t=1,\ldots,n_{hi}$ éléments dans la grappe hi. Désignons l'ensemble des grappes échantillonnées dans la strate h par s_h, et désignons l'échantillon des éléments de la grappe hi par s_{hi}. Le nombre total de grappes échantillonnées est $m=\sum_h m_h$, le nombre total d'éléments échantillonnés dans la strate h est $n_h=\sum_{i\in s_h} n_{hi}$, et le total dans l'échantillon est $n=\sum_{h=1}^{H} n_h$. Supposons

que des échantillons probabilistes soient sélectionnés à partir de grappes et d'éléments au sein de grappes; les méthodes de sélection particulières n'ont pas à être précisées pour les analyses ultérieures. On suppose que les grappes sont sélectionnées avec remise à l'intérieur des strates et indépendamment entre les strates. L'hypothèse « avec remise » pour la sélection des grappes est courante dans les analyses théoriques, car elle simplifie les formules de variance.

3.2 Modèles linéaires généralisés pondérés par les poids de sondage

Dans cette section, pour plus de commodité, nous simplifions la notation par rapport à celle de la section 3.1, mais nous appliquerons cette notation plus élaborée dans les sections qui suivront. Dans un plan de sondage complexe, désignons l'ensemble des données échantillonnées par s et supposons qu'il y ait n éléments dans s. Le logarithme de la vraisemblance de la population dans le modèle linéaire généralisé dans (2.3) est estimé en additionnant le logarithme naturel de chacune des composantes définies par (2.1) pour les n observations pondérées par les poids de sondage w_i :

$$\hat{l}(\boldsymbol{\beta}) = \sum_{i \in s} w_i \left[y_i \theta_i - b(\theta_i) \right] / \tau^2 - \sum_{i \in s} w_i c(y_i, \tau). \tag{3.1}$$

Supposons que les poids, $\{w_i\}_{i \in s}$, sont construits de manière à fournir des estimateurs sans biais ou approximativement sans biais des totaux de population.

En faisant correspondre à zéro les données partielles du premier degré pour $\boldsymbol{\beta}$ et en étendant le résultat dans (2.4), les équations d'estimation par le pseudomaximum de vraisemblance pour $\boldsymbol{\beta}$ sont données par :

$$\begin{aligned} \frac{\partial \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{1}{\tau^2} \sum_{i \in s} w_i \frac{(y_i - \mu_i)}{v(\mu_i) g'(\mu_i)} \mathbf{x}_i^T \\ &= \frac{1}{\tau^2} \sum_{i \in s} w_i (y_i - \mu_i) \gamma_i g'(\mu_i) \mathbf{x}_i^T \\ &= \mathbf{0}^T \end{aligned} \tag{3.2}$$

où $\gamma_i = \left\{ v(\mu_i) [g'(\mu_i)]^2 \right\}^{-1}$ comme dans (2.4). Dans la notation matricielle, il s'agit de

$$\frac{\partial \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{\tau^2} \mathbf{X}^T \mathbf{W} \boldsymbol{\Gamma} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}^T, \tag{3.3}$$

où $\mathbf{X}$ est la matrice $n \times p$ des variables explicatives pour les éléments de l'échantillon, $\mathbf{W} = \text{diag}(w_i)_{i \in s}$, $\boldsymbol{\Gamma} = \text{diag}(\gamma_i)_{i \in s}$, $\boldsymbol{\Delta} = \text{diag}\left[g'(\mu_i) \right]$ et $\boldsymbol{\mu} = (\mu_i)_{i \in s}$. La valeur de $\boldsymbol{\beta}$ qui résout les équations d'estimation pondérées par les poids de sondage sera désignée par $\hat{\boldsymbol{\beta}}_{SW}$. Pour distinguer les modèles linéaires généralisés des plans de sondage complexes des modèles linéaires généralisés ordinaires, nous ferons référence aux *modèles linéaires généralisés pondérés par les poids de sondage* (SWGLM), car les poids de sondage sont intégrés à l'estimation par le pseudomaximum de vraisemblance.

Les résultats des VIF fondés sur un modèle pour la régression pondérée par les poids de sondage sont parallèles à ceux de la section 2.2 pour les GLM ne provenant pas d'enquêtes, et ils sont exposés dans le cas présent. La matrice de variance-covariance asymptotique fondée sur un modèle de $\hat{\boldsymbol{\beta}}_{SW}$ est :

$$\text{avar}_M(\hat{\boldsymbol{\beta}}_{SW}) = \tau^2(\mathbf{X}^T\mathbf{W}\boldsymbol{\Gamma}\mathbf{X})^{-1} \equiv \tau^2\mathbf{A}^{-1}. \tag{3.4}$$

Même si cet estimateur fait effectivement intervenir les poids de sondage, un échantillonnage non informatif est nécessaire pour qu'il soit cohérent dans le modèle (2.1). Le k^{e} élément de la diagonale principale, $\tau^2 a^{kk}$, est la variance asymptotique de modèle $\hat{\beta}_{SWk}$. L'expression (3.4) peut être estimée par $\hat{\tau}^2\widehat{\mathbf{A}}^{-1}$ où $\widehat{\mathbf{A}} = \mathbf{X}^T\mathbf{W}\widehat{\boldsymbol{\Gamma}}\mathbf{X}$ est l'estimateur de la matrice d'information avec $\widehat{\boldsymbol{\Gamma}} = \text{diag}\left(\hat{\gamma}_i\right)_{i\in s}$. Un estimateur de variance de modèle de $\hat{\beta}_{SWk}$ est :

$$\text{avar}_M(\hat{\beta}_{SWk}) = \hat{\tau}^2\hat{a}^{kk}, \tag{3.5}$$

où $\hat{a}^{kk}$ est le k^{e} élément sur la diagonale principale de la matrice $\widehat{\mathbf{A}}^{-1}$. Dans certains modèles, il faut estimer τ; toutefois, aux fins de l'analyse dans le cas présent, aucune estimation n'est nécessaire, car le VIF présenté ci-dessous n'inclut pas τ.

En utilisant des étapes semblables à celles nécessaires pour dériver (2.8),

$$\hat{a}^{kk} = \frac{1}{(1-R^2_{W\Gamma(k)})\,\mathbf{x}_k^T\mathbf{W}\widehat{\boldsymbol{\Gamma}}\mathbf{x}_k} \tag{3.6}$$

où $R^2_{W\Gamma(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{W\Gamma(k)}\mathbf{X}^T_{(k)}\mathbf{W}\widehat{\boldsymbol{\Gamma}}\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)}}{\mathbf{x}_k^T\mathbf{W}\widehat{\boldsymbol{\Gamma}}\mathbf{x}_k}$ avec $\hat{\beta}_{W\Gamma(k)} = (\mathbf{X}^T_{(k)}\mathbf{W}\widehat{\boldsymbol{\Gamma}}\mathbf{X}_{(k)})^{-1}\mathbf{X}^T_{(k)}\mathbf{W}\widehat{\boldsymbol{\Gamma}}\mathbf{x}_k$. $R^2_{W\Gamma(k)}$ est le coefficient de détermination dans la régression linéaire de $\mathbf{x}_k$ sur les $p-1$ autres variables explicatives avec la matrice de poids $\mathbf{W}\widehat{\boldsymbol{\Gamma}}$. La variance asymptotique de modèle de $\hat{\beta}_{SWk}$ est gonflé par

$$\text{VIF}_{MWk} = \frac{1}{1-R^2_{W\Gamma_{(k)}}} \tag{3.7}$$

par rapport à une régression pondérée dans laquelle toutes les variables explicatives sont orthogonales.

3.3 Estimateur de variance par linéarisation de type modèle-plan pour $\hat{\boldsymbol{\beta}}_{SW}$

Pour calculer un VIF de type modèle-plan dans un plan complexe, il faut un estimateur asymptotique de la variance de $\hat{\boldsymbol{\beta}}_{SW}$. Un choix fondé sur un modèle est l'estimateur de la matrice d'information qui est obtenu en substituant les estimateurs de τ et $\boldsymbol{\Gamma}$ dans (3.4). Un choix de type modèle-plan, c'est-à-dire un choix qui présente de bonnes propriétés de plan et de modèle, est l'estimateur de variance par linéarisation pour $\hat{\boldsymbol{\beta}}_{SW}$ à l'aide de la méthode « Binder » (Binder, 1983). L'estimateur de type sandwich de Binder est pratique, car il est disponible dans des logiciels largement utilisés tels que Stata (https://www.stata.com/manuals/svyvarianceestimation.pdf) et le paquet R `survey` (Lumley, 2024). L'estimateur de type sandwich est convergent par rapport au modèle si la spécification du GLM est correcte et convergente par rapport au plan selon les conditions imposées par Binder (1983, annexe).

Binder et Roberts (2003, section 3.3.1) observent également que lorsque la fraction de sondage est faible, un estimateur de variance convergent par rapport au plan de sondage d'un estimateur ponctuel $\hat{t}$ convient pour estimer la variance de modèle de type modèle-plan, $V_{M\pi}\left(\hat{t}\right) = V_M E_\pi\left(\hat{t}\right) + E_M V_\pi\left(\hat{t}\right)$. Le résultat de

Binder-Roberts exige que l'estimateur de variance convergent par rapport au plan de sondage soit aussi asymptotiquement sans biais par rapport au modèle pour $E_M V_\pi\left(\hat{t}\right)$ – une exigence à laquelle l'estimateur de type sandwich de Binder satisfait.

Les équations d'estimation pour $\hat{\boldsymbol{\beta}}_{SW}$ dans les modèles linéaires généralisés sont obtenues en faisant correspondre $\partial\hat{l}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ à zéro, comme indiqué dans (3.2) et (3.3). Nous pouvons réécrire (3.3) en définissant $\mathbf{z}_i = (y_i - \mu_i)\gamma_i g'(\mu_i)\mathbf{x}_i^T = e_i\gamma_i g'(\mu_i)\mathbf{x}_i^T$, où $e_i = y_i - \mu_i$ est le résidu de l'élément i, et $\hat{\mathbf{z}}_i^* = w_i\mathbf{z}_i$, $\hat{\mathbf{z}}^* = \sum_{i\in s}\mathbf{z}_i^* = \frac{1}{\hat{\tau}^2}\mathbf{X}^T\mathbf{W}\boldsymbol{\Gamma}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})$ avec $\boldsymbol{\mu} = (\mu_i)_{n\times 1}$ et $\boldsymbol{\Delta} = \text{diag}[g'(\mu_i)]$. Les équations d'estimation dans (3.3) peuvent alors être écrites brièvement comme $\mathbf{z}^* = \mathbf{0}$. Il convient de mentionner que pour évaluer $\hat{\mathbf{z}}_i^*$, il faut utiliser des estimations d'échantillon de μ_i.

Sous réserve des conditions énumérées dans Binder (1983, annexe), la matrice de variance-covariance approximative du plan pour $\hat{\boldsymbol{\beta}}_{SW}$ est la suivante :

$$V_\pi(\hat{\boldsymbol{\beta}}_{SW}) = \mathbf{J}(\boldsymbol{\beta})^{-1}V_\pi(\hat{\mathbf{z}}^*)\mathbf{J}(\boldsymbol{\beta})^{-1} \tag{3.8}$$

où V_π désigne la variance par rapport au plan d'échantillonnage, $\mathbf{J}(\boldsymbol{\beta})$ est la matrice dérivée partielle (jacobienne) des équations d'estimation pour $\hat{\boldsymbol{\beta}}_{SW}$ évaluées au niveau de la population $\boldsymbol{\beta}$. La matrice jacobienne est $E_M\left[\frac{\partial^2\hat{l}(\boldsymbol{\beta}_{SW})}{\partial\boldsymbol{\beta}_{SW}\partial\boldsymbol{\beta}_{SW}^T}\right] = \tau^{-2}\mathbf{X}^T\boldsymbol{\Gamma}\mathbf{X}$. L'estimateur de variance par linéarisation de la matrice de variance-covariance pour $\hat{\boldsymbol{\beta}}_{SW}$ est obtenu en remplaçant les estimateurs de la variance fondés sur le plan pour $\mathbf{J}$ et $V_\pi(\hat{\mathbf{z}}^*)$, ce qui donne

$$\begin{aligned} v_L(\hat{\boldsymbol{\beta}}_{SW}) &= \hat{\tau}^4[\mathbf{X}^T\mathbf{W}\hat{\boldsymbol{\Gamma}}\mathbf{X}]^{-1}v_\pi(\hat{\mathbf{z}}^*)[\mathbf{X}^T\mathbf{W}\hat{\boldsymbol{\Gamma}}\mathbf{X}]^{-1} \\ &= \hat{\tau}^4\hat{\mathbf{A}}^{-1}v_\pi(\hat{\mathbf{z}}^*)\hat{\mathbf{A}}^{-1} \end{aligned} \tag{3.9}$$

par $\hat{\mathbf{A}} = \mathbf{X}^T\mathbf{W}\hat{\boldsymbol{\Gamma}}\mathbf{X}$ comme dans (3.4). Il convient de mentionner que cet estimateur de variance ne nécessite pas de formuler, pour le modèle (2.1), l'hypothèse selon laquelle la variance de modèle de y est une fonction de la moyenne du modèle de y. L'estimateur (3.9) peut également être appliqué à des données ne provenant pas d'enquêtes, pour lesquelles aucun poids de sondage n'est utilisé.

Lorsqu'une estimation sans biais ou convergente par rapport au plan de $V_\pi(\hat{\mathbf{z}}^*)$ est substituée dans (3.9), l'estimateur de variance par linéarisation pour $\hat{\boldsymbol{\beta}}_{SW}$ sera lui aussi approximativement sans biais et convergent par rapport au plan. Pour obtenir un VIF de type modèle-plan, nous devons intégrer notre méthode de diagnostic à un plan de sondage particulier, car l'estimation de l'estimateur de variance fondé sur le plan $v_\pi(\hat{\mathbf{z}}^*)$ est reliée au plan de sondage. L'annexe A.2 présente l'estimateur, $v_\pi(\hat{\mathbf{z}}^*)$, dans un échantillonnage stratifié et à plusieurs degrés.

3.4 Facteur d'inflation de la variance dans les SWGLM

Supposons que $\widehat{\mathbf{V}}$ est l'estimateur, $v_\pi(\hat{\mathbf{z}}^*)$, conformément au plan d'échantillonnage, quel qu'il soit. L'estimateur de variance par linéarisation $v_L(\hat{\boldsymbol{\beta}}_{SW})$ peut alors être réécrit comme suit :

$$v_L(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\tau}^4\widehat{\mathbf{A}}^{-1}\mathbf{X}^T\mathbf{W}\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Delta}}\widehat{\mathbf{V}}\hat{\boldsymbol{\Delta}}\hat{\boldsymbol{\Gamma}}\mathbf{W}\mathbf{X}\widehat{\mathbf{A}}^{-1}. \tag{3.10}$$

Désignons $\mathbf{V}_\Delta = \hat{\mathbf{\Delta}}\hat{\mathbf{V}}\hat{\mathbf{\Delta}}$, $\mathbf{W}_\Gamma = \mathbf{W}\hat{\mathbf{\Gamma}} = \text{diag}(w_i\hat{\gamma}_i)$, $\mathbf{B} = \mathbf{X}^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{X}$, $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ et nous avons $v_L(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\tau}^4\mathbf{G}$. L'estimateur de variance par linéarisation pour $\hat{\beta}_{SWk}$ est le k^{e} élément de la diagonale principale de $v_L(\hat{\boldsymbol{\beta}}_{SW})$; il est donc directement lié au k^{e} élément de la diagonale principale de la matrice $\mathbf{G}$, comme suit :

$$v_L(\hat{\beta}_{SWk}) = \hat{\tau}^4 g_{kk}, \tag{3.11}$$

en définissant $\mathbf{G} = (g_{ij})$. Comme le montre l'annexe A.1, g_{kk} est

$$g_{kk} = \frac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{W\Gamma(k)}}\frac{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{x}_k}{(\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{x}_k)^2} \tag{3.12}$$

où $R^2_{\mathbf{W}\Gamma(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{W\Gamma(k)}\mathbf{X}^T_{(k)}\mathbf{W}_\Gamma\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)}}{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{x}_k}$ avec $\hat{\beta}_{\mathbf{W}\Gamma(k)} = (\mathbf{X}^T_{(k)}\mathbf{W}_\Gamma\mathbf{X}_{(k)})^{-1}\mathbf{X}^T_{(k)}\mathbf{W}_\Gamma\mathbf{x}_k$, $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)}$ est le résidu de la régression pondérée par les poids de sondage de $\mathbf{x}_k$ sur $\mathbf{X}_{(k)}$ à l'aide de la matrice de poids $\mathbf{W}_\Gamma$,

$$\hat{\zeta}_k = \frac{(\mathbf{x}_k - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma(\mathbf{x}_k - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})}{(\mathbf{x}_k - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T\mathbf{W}_\Gamma(\mathbf{x}_k - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})} = \frac{\mathbf{e}_{xk}^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}_\Gamma\mathbf{e}_{xk}},$$

et

$$\hat{\varrho}_k = \frac{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{x}_k}{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{x}_k}.$$

Il convient de mentionner que si $\mathbf{V}_\Delta = \mathbf{W}_\Gamma^{-1}$, le VIF pour l'approche fondée sur le plan sera $\left(1-R^2_{W\Gamma(k)}\right)^{-1}$, ce qui est similaire au VIF ne provenant pas d'enquêtes dans un modèle linéaire. Cependant, cette hypothèse est peu probable dans la pratique.

Considérons un modèle avec uniquement $\mathbf{x}_k$. En utilisant (3.10), l'estimateur de variance par linéarisation pour $\hat{\beta}_{SWk}$ est :

$$v_{0L}(\hat{\beta}_{SWk}) = \hat{\tau}^4\frac{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{x}_k}{(\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{x}_k)^2}. \tag{3.13}$$

Ainsi, $v_{0L}(\hat{\beta}_{SWk})$ est gonflé par un multiple de

$$\text{VIF}_{SWk} = \frac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{W\Gamma(k)}} \tag{3.14}$$

lorsque d'autres variables explicatives sont incluses dans le modèle, en comparant $v_{0L}(\hat{\beta}_{SWk})$ dans (3.13) à $v_L(\hat{\beta}_{SWk})$ dans (3.12). Ce faisant, on suppose que τ est identique dans le modèle qui comprend uniquement $\mathbf{x}_k$ et dans celui qui comprend tous les $\mathbf{x}$.

Le terme $\hat{\zeta}_k\hat{\varrho}_k\big/1-R^2_{W\Gamma(k)}$ est appelé le VIF pour la k^{e} variable explicative du modèle. Le VIF dans (3.14) est analogue à ceux produits à partir de données ne provenant pas d'enquêtes pour des modèles linéaires par des progiciels tels que SAS[MD] et Stata[MD], et il sera appelé dans le cas présent VIF *sans l'ordonnée à l'origine*. Dans le cas particulier d'un modèle linéaire avec variance homogène, c'est-à-dire $\text{var}_M(\mathbf{y}) = \sigma^2\mathbf{I}$, nous avons $\mathbf{W}_\Gamma = \mathbf{W}$ et VIF_{SWk} se réduira au VIF pour un modèle linéaire dans Liao et Valliant (2012b, équation (6)).

Le VIF dans (3.14) repose sur un modèle avec $\mathbf{x}_k$ uniquement comme comparaison pour déterminer dans quelle mesure la variance de $\widehat{\beta}_{SWk}$ est gonflé par la non-orthogonalité des valeurs x. Une méthode de rechange consiste à réaliser une comparaison avec un modèle qui comporte à la fois $\mathbf{x}_k$ et l'ordonnée à l'origine. L'argument en faveur de l'utilisation de ce modèle comme modèle de comparaison est qu'un analyste typique inclura toujours une ordonnée à l'origine dans un modèle. Ainsi, la comparaison appropriée se fait entre un modèle comprenant une ordonnée à l'origine et un x et un modèle comprenant une ordonnée à l'origine et tous les x. Dans ce cas, en utilisant (3.12), l'estimateur de variance par linéarisation pour $\widehat{\beta}_{SWk}$ peut s'écrire :

$$v_{1L}(\widehat{\beta}_{SWk}) = \hat{\tau}^4 \frac{(\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)}{(\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2)^2} \tag{3.15}$$

où $\widehat{\widehat{N}} = \sum_{i\in s} w_i \hat{\gamma}_i$ et $\bar{\tilde{x}}_k = \sum_{i\in s} w_i \hat{\gamma}_i x_{ki} / \widehat{\widehat{N}}$.

La variance $v_L(\widehat{\beta}_{SWk})$ dans (3.15) peut être décomposée en plusieurs facteurs en conséquence :

$$v_L(\widehat{\beta}_{SWk}) = \tau^4 g_{kk}$$

avec (3.16)

$$\begin{aligned} g_{kk} &= \frac{\hat{\zeta}_k}{1 - R^2_{W\Gamma m(k)}} \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2}{(\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)} \times \frac{(\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)}{(\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2)^2} \\ &= \frac{\hat{\zeta}_k \hat{\varrho}_{km}}{1 - R^2_{W\Gamma m(k)}} v_{1L}(\widehat{\beta}_{SWk}) \end{aligned}$$

où $R^2_{W\Gamma m(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{W\Gamma(k)} \mathbf{X}^T_{(k)} \mathbf{W}_\Gamma \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2}{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2}$ est le coefficient de détermination corrigé pour la moyenne dans la régression par les moindres carrés pondérés lorsque la matrice de pondération est $\mathbf{W}_\Gamma$, et $\hat{\varrho}_{km} = \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k - \widehat{\widehat{N}}\bar{\tilde{x}}_k^2}{(\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)}$. Le terme

$$\text{VIF}_{SWm(k)} = \frac{\hat{\zeta}_k \hat{\varrho}_{km}}{1 - R^2_{W\Gamma m(k)}} \tag{3.17}$$

est appelé le VIF *ajusté avec l'ordonnée à l'origine* pour la k^{e} variable explicative dans le modèle.

Diverses règles empiriques ont été proposées pour détecter les signes d'une colinéarité nuisible dans la régression linéaire. Marquardt (1970) et Kutner, Nachtsheim et Neter (2004) estiment qu'un VIF de 10 ou plus est le seuil à partir duquel il y a colinéarité importante. Chatterjee et Price (1991) proposent une règle plus élaborée en deux parties : 1) le VIF le plus élevé est égal ou supérieur à 10; 2) la moyenne de tous les VIF pour toutes les variables explicatives est considérablement supérieure à 1. Bien sûr, ces règles sont quelque peu arbitraires et peuvent être adaptées, en fonction de l'analyse. Si, par exemple, l'objectif est la prédiction, la colinéarité peut représenter un problème que dans certains domaines de prédiction (Snee et Marquardt, 1984). Dans l'inférence causale, la colinéarité est un problème général, car elle peut masquer les variables explicatives qui expliquent le phénomène modélisé. Ainsi, dans l'inférence causale, un seuil plus strict peut être justifié que lors de la prédiction ou de la prévision.

L'approche fondée sur le plan pour l'ajustement des GLM dont il est question ci-dessus est l'une des méthodes typiques où les poids de sondage sont utilisés directement dans les équations d'estimation. D'autres approches fondées sur le plan ont également été élaborées pour l'ajustement des GLM dans un échantillonnage informatif. Pfeffermann et Sverchkov (2003) ont examiné deux autres approches qui nécessitent la modélisation et l'estimation de l'espérance des poids de sondage, soit en tant que fonction du résultat et des variables explicatives, soit en tant que fonction des variables explicatives seulement. Les espérances des poids de sondage sont utilisées comme poids dans les équations d'estimation. Ainsi, lorsque ces deux approches sont utilisées, nous pouvons simplement traiter les espérances des poids comme la matrice $\mathbf{W}$ dans (3.14) ou (3.17) pour obtenir les VIF.

Certaines des lois univariées courantes de la famille exponentielle et leurs fonctions de lien spéciales sont énumérées dans le tableau 2.1. En remplaçant leurs éléments correspondants des matrices $\mathbf{\Delta}$ et $\mathbf{\Gamma}$ dans (3.14) et (3.17), nous pouvons obtenir leurs formules de VIF respectives. Pour démontrer ces techniques dans un modèle GLM spécial, nous prendrons l'exemple d'un modèle logistique dans la section 4.

3.5 Indices de conditionnement

Une autre méthode de diagnostic de colinéarité consiste à utiliser des indices de conditionnement comme proposé par Belsley (1984) pour les modèles linéaires et appliqués aux données d'enquêtes complexes par Liao et Valliant (2012a). Pour détecter les colinéarités problématiques dans d'autres membres de la catégorie des GLM, Mackinnon et Puterman (1989), Weissfeld et Sereika (1991) et Lesaffre et Marx (1993) ont modifié cette approche en utilisant une version échelonnée de la matrice d'information dans un GLM afin de détecter l'instabilité des estimations de $\boldsymbol{\beta}$.

Dans le modèle linéaire généralisé pondéré par les poids de sondage, nous utiliserons la matrice d'information évaluée à $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{SW}$. Comme le montre (3.4), la matrice d'information estimée est $\widehat{\mathbf{A}} = \mathbf{X}^T \mathbf{W} \widehat{\mathbf{\Gamma}} \mathbf{X}$. Désignons la matrice pondérée $\mathbf{W}^{1/2} \widehat{\mathbf{\Gamma}}^{1/2} \mathbf{X} \equiv \mathbf{W}_{\Gamma}^{1/2} \mathbf{X}$ par $\widetilde{\mathbf{X}}$. La k^{e} colonne de $\widetilde{\mathbf{X}}$ est $\tilde{\mathbf{x}}_k$. La décomposition en valeurs singulières (DVS) de $\widetilde{\mathbf{X}}$ est $\widetilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, où $\mathbf{U}$ est une matrice orthogonale $n \times p$, $\mathbf{V}$ est une matrice orthogonale $p \times p$ et $\mathbf{D} = \text{diag}(\mu_k)$ est la matrice diagonale $p \times p$ de valeurs singulières.

Le nombre de conditionnements de $\widetilde{\mathbf{X}}$ est défini comme $\kappa(\widetilde{\mathbf{X}}) = \mu_{\max} / \mu_{\min}$, où $\mu_{\max}$ et $\mu_{\min}$ sont les valeurs singulières maximales et minimales de $\widetilde{\mathbf{X}}$. Le nombre de conditionnements de $\widetilde{\mathbf{X}}$ est habituellement différent du nombre de conditionnements de la matrice de données $\mathbf{X}$ en raison des poids de sondage et des $\hat{\gamma}_i$ inégaux. Les indices de conditionnement sont définis comme

$$\psi_k = \mu_{\max} / \mu_k, \quad k = 1, \ldots, p \tag{3.18}$$

où μ_k est l'une des valeurs singulières de $\widetilde{\mathbf{X}}$. Comme le mentionnent Belsley et coll. (1980), la taille d'un indice de conditionnement dépend des échelles des colonnes de $\mathbf{X}$. Pour les régressions linéaires, ils recommandent d'échelonner les colonnes de $\mathbf{X}$ afin qu'elles soient toutes de la même longueur – une procédure que nous adoptons dans le cas présent pour $\widetilde{\mathbf{X}}$. La k^{e} colonne de $\widetilde{\mathbf{X}}$ est divisée par $\left(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k\right)^{1/2}$ afin que la longueur de chaque colonne soit égale à 1. En conséquence, tous les indices de conditionnement

seraient égaux à 1 lorsque les colonnes sont orthogonales entre elles. Dans les résultats ci-dessous, les termes « indices de conditionnement échelonnés » et « nombres de conditionnements échelonnés » font référence aux indices de conditionnement et aux nombres de conditionnements des $\widetilde{\mathbf{X}}$ échelonnés.

En fonction des extrema du rapport des formes quadratiques (Lin, 1984), le nombre de conditionnements $\kappa(\widetilde{\mathbf{X}})$ est limité dans l'intervalle suivant :

$$\frac{(w\cdot\hat{\gamma})_{\min}^{1/2}}{(w\cdot\hat{\gamma})_{\max}^{1/2}}\kappa(\mathbf{X}) \le \kappa(\widetilde{\mathbf{X}}) \le \frac{(w\cdot\hat{\gamma})_{\max}^{1/2}}{(w\cdot\hat{\gamma})_{\min}^{1/2}}\kappa(\mathbf{X}), \tag{3.19}$$

où $(w\cdot\hat{\gamma})_{\min}$ et $(w\cdot\hat{\gamma})_{\max}$ sont les valeurs minimale et maximale de $w_i\hat{\gamma}_i$ pour les différentes unités observées.

L'expression (3.19) indique que si les poids de travail utilisés dans le SWGLM (c'est-à-dire $w_i\hat{\gamma}_i$) ne varient pas trop, le nombre de conditionnements que contient le SWGLM sera similaire à celui des moindres carrés ordinaires (MCO). Si les procédures relatives au plan et à l'estimation conduisent à un large éventail de poids de sondage ou si les $\hat{\gamma}_i$ ont un large intervalle, le nombre de conditionnements peut être très différent. Lorsque le SWGLM a un nombre de conditionnements élevé, ce n'est pas nécessairement le cas des MCO. Il n'existe pas de règles absolues pour déterminer quand un indice de conditionnement est « élevé ». Si la matrice d'information estimée est singulière, au moins une valeur singulière sera égale à 0 et l'indice de conditionnement sera infini. En pratique, les indices de conditionnement compris entre 5 et 10 sont considérés comme associés à des dépendances faibles, tandis que toute valeur supérieure à 30 indique des relations modérées à fortes.

En utilisant la DVS de $\widetilde{\mathbf{X}}$, la variance approximative de modèle de (3.4) est

$$\text{avar}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right) = \tau^2\left[(\mathbf{UDV})^T(\mathbf{UDV})\right]^{-1} = \tau^2\mathbf{VD}^{-2}\mathbf{V}^T. \tag{3.20}$$

La variance approximative de modèle de $\widehat{\beta}_{SWk}$ est le k^{e} élément diagonal de (3.20) :

$$\text{avar}_M\left(\widehat{\beta}_{SWk}\right) = \tau^2\sum_{j=1}^{p}\frac{v_{kj}^2}{\mu_j^2} \tag{3.21}$$

où $\mathbf{V} = (v_{kj})_{p\times p}$. Supposons que $\phi_{kj} = v_{kj}^2/\mu_j^2$ et $\phi_k = \sum_{j=1}^{p}\phi_{kj}$. Les proportions de décomposition de la variance sont $\pi_{jk} = \phi_{jk}/\phi_k$. Mentionnons qu'en établissant $\mathbf{W} = \mathbf{I}$, (3.20) devient la décomposition de la variance approximative du GLM sans poids de sondage dans (2.5).

Comme au moins un v_{kj} doit être non nul dans (3.21), on peut en déduire qu'une proportion élevée d'une variance, quelle qu'elle soit, peut être associée à une valeur singulière élevée, même en l'absence de colinéarité. L'approche standard consiste à vérifier si un indice de conditionnement élevé est associé à une proportion élevée de la variance d'au moins deux coefficients pendant le diagnostic de colinéarité. En effet, au moins deux colonnes de $\mathbf{X}$ doivent intervenir pour créer une quasi-dépendance. Liao et Valliant (2012a), suivant Belsley et coll. (1980), ont proposé de présenter une matrice des proportions de décomposition et

des indices de conditionnement de $\widetilde{\mathbf{X}}$ dans un tableau de décomposition de la variance – une démarche que nous suivons dans l'exemple empirique de la section 4. Il convient de mentionner que la variance approximative de type modèle-plan dans (3.8) pourrait également être décomposée et que des proportions de variance pourraient être calculées de façon analogue aux étapes ci-dessus, mais nous ne poursuivrons pas cette solution de rechange dans la présente étude.

4. Exemple empirique

Nous illustrons les techniques susmentionnées et étudions leur comportement en utilisant des données sur l'apport alimentaire provenant de la National Health and Nutrition Examination Survey (NHANES) réalisée aux États-Unis. Deux cycles de la NHANES, 2015-2016 et 2017-2018, ont été combinés pour augmenter la taille de l'échantillon aux fins de l'analyse, comme l'ont proposé Johnson, Paulose-Ram et Ogden (2012). 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'interview (de minuit à minuit) et pour estimer les apports d'énergie, de nutriments et d'autres constituants alimentaires provenant de ces aliments et boissons. La NHANES est réalisée selon un plan d'échantillonnage probabiliste complexe à plusieurs degrés (Curtin, Mohadjer et Dohrmann, 2012). Certains sous-groupes de population sont suréchantillonnés afin d'accroître la fiabilité et la précision des estimations de l'état de santé pour ces groupes. Les taux de réponse obtenus pour différentes années de la NHANES peuvent être consultés sur le site https://wwwn.cdc.gov/nchs/nhanes/ResponseRates.aspx. Les poids de sondage ont été construits en prenant les poids de sondage initiaux et en les rajustant pour tenir compte de la non-réponse au niveau du ménage, de la non-réponse aux questions sur le régime alimentaire et de la différence de répartition selon le jour de la semaine pour la collecte des données sur les apports alimentaires. Les poids ont aussi été stratifiés a posteriori par rapport aux totaux de population pour un ensemble de groupes démographiques afin de tenir compte de la sous-représentation et de la surreprésentation. Les champs disponibles dans les fichiers à usage public ne permettent pas de tenir pleinement compte de ces complexités d'estimation lors de l'estimation des variances. L'ensemble de données utilisé dans notre étude est un sous-ensemble de données de 2015 à 2018 composé de répondants de 18 à 65 ans. Les observations comportant des valeurs manquantes pour les variables choisies ont été exclues de l'échantillon qui, en dernière analyse, contient 6 682 réponses complètes. Les poids finaux dans notre échantillon varient de 1 358 à 345 570, avec un ratio de 254 pour 1. Conformément à la documentation de la NHANES, nous traitons l'échantillon comme une sélection stratifiée de 60 unités primaires d'échantillonnage (UPE) provenant de 30 strates, avec 2 UPE échantillonnées avec remise dans chaque strate.

Pour la présente étude empirique, nous considérons un modèle logistique binaire. Les variables explicatives comprennent une variable catégorique pour la race-ethnicité (Noir, Blanc et autres, « Blanc » étant le groupe de référence), le genre (homme = 1, femme = 0), l'âge, l'interaction entre le genre et l'âge,

ainsi que neuf variables de l'apport alimentaire quotidien : les calories (kcal), les protéines (grammes), les glucides (grammes), les sucres (grammes), les lipides totaux (grammes), les acides gras saturés totaux (grammes), les acides gras mono-insaturés totaux (grammes), les acides gras polyinsaturés totaux (grammes) et l'alcool (grammes). Les modèles de régression logistique sont ajustés selon l'obésité/la non-obésité comme variable de réponse binaire. L'obésité est définie comme l'indice de masse corporelle (IMC) de 30 ou plus chez le répondant. Les variables relatives à l'apport alimentaire étant corrélées, des problèmes de colinéarité sont à prévoir lorsqu'elles sont toutes utilisées ensemble comme valeurs explicatives de l'obésité.

Trois méthodes de régression sont appliquées et comparées dans la présente étude. Chaque méthode utilise l'estimation du maximum de vraisemblance mise en œuvre à l'aide des moindres carrés repondérés itératifs, mais les caractéristiques de l'enquête y sont traitées différemment. La première est appelée les *moindres carrés généralisés* (MCG), mais elle ne tient pas compte de tous les aspects complexes de l'échantillonnage, notamment la stratification, le regroupement en grappes et la pondération. La deuxième est appelée les *moindres carrés pondérés généralisés* (MCPG). Elle intègre les poids de sondage, mais part du principe que les personnes sont indépendantes et ne tient pas compte des strates et de l'échantillonnage en grappes. Les MCG et les MCPG sont tous deux victimes d'une erreur de spécification en ce sens où l'on ne tient pas compte de la surdispersion causée par l'échantillonnage en grappes. L'omission de variables explicatives importantes est un autre type d'erreur de spécification, mais elle n'est pas prise en compte dans cet exemple. La troisième méthode, décrite à la section 3.1, est la nouvelle méthode qui utilise le plan d'échantillonnage complexe en tant que tel et est représentée par SWGLM dans les tableaux. Les formules applicables aux matrices de pondération, les estimateurs de la variance des coefficients et les diagnostics de colinéarité de ces trois méthodes sont présentés dans le tableau 4.1.

Les estimations des coefficients et des erreurs-types du paquet R `survey` (Lumley, 2024) pour les modèles ajustés à l'aide des trois méthodes de régression distinctes sont présentées dans le tableau 4.2. Il convient de mentionner que les erreurs-types des MCG et des MCPG ont une interprétation fondée sur un modèle (en partant du principe que le modèle est correctement précisé). Les erreurs-types des SWGLM ont à la fois une interprétation fondée sur le plan et une interprétation fondée sur un modèle. Les modèles *complets* accompagnés de toutes les variables explicatives se trouvent dans la partie supérieure du tableau. Ensuite, un modèle *réduit* présentant moins de problèmes de quasi-dépendance est ajusté à l'aide d'une variable indicatrice pour la race (Noir, toutes les autres races constituant le groupe de référence), l'âge, les acides gras saturés totaux, les acides gras mono-insaturés totaux et l'alcool; il est présenté dans la partie inférieure. Le modèle réduit ne vise pas à trouver le meilleur modèle pour prédire l'obésité, mais sert plutôt à illustrer l'effet que la colinéarité peut avoir sur la taille des coefficients estimés et des erreurs-types.

Tableau 4.1
Modèles de régression et leurs statistiques de diagnostic de la colinéarité utilisée dans l'étude expérimentale

Modèle	Y	Type de régression	$\mathbf{W}_\Gamma$	Estimation de la variance de $\hat{\boldsymbol{\beta}}$	Formule du VIF
Logistique	Obèse	MCG	$\hat{\boldsymbol{\Gamma}}$ [a]	$(\mathbf{X}^T\hat{\boldsymbol{\Gamma}}\mathbf{X})^{-1}$ [b]	$\text{VIF} = \dfrac{1}{1-R^2_{\Gamma(k)}}$
(Binaire)	/Non-obèse	MCPG	$\mathbf{W}\hat{\boldsymbol{\Gamma}}$ [c]	$(\mathbf{X}^T\mathbf{W}_\Gamma\mathbf{X})^{-1}$	$\text{VIF} = \dfrac{1}{1-R^2_{W\Gamma(k)}}$
		(fondé sur un modèle) SWGLM	$\mathbf{W}\hat{\boldsymbol{\Gamma}}$	$(\mathbf{X}^T\mathbf{W}_\Gamma\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_\Gamma\mathbf{V}\Delta\mathbf{W}_\Gamma\mathbf{X}(\mathbf{X}^T\mathbf{W}_\Gamma\mathbf{X})^{-1}$	$\text{VIF} = \dfrac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{W\Gamma(k)}}$ avec $\hat{\zeta}_k = \dfrac{\mathbf{e}_{xk}^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}_\Gamma\mathbf{e}_{xk}}$, $\hat{\varrho}_k = \dfrac{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{x}_k}{\mathbf{x}_k^T\mathbf{W}_\Gamma\mathbf{V}_\Delta\mathbf{W}_\Gamma\mathbf{x}_k}$
		(fondé sur le plan)		$\mathbf{V}_\Delta = \hat{\boldsymbol{\Delta}}\hat{\mathbf{V}}\hat{\boldsymbol{\Delta}}$ [d] $= \hat{\boldsymbol{\Delta}}\left\{\sum_{h=1}^{H}\dfrac{n_h}{n_h-1}\left[\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T) - \dfrac{1}{n_h}\mathbf{e}_h\mathbf{e}_h^T\right]\right\}\hat{\boldsymbol{\Delta}}$	

Notes : [a] $\hat{\boldsymbol{\Gamma}} = \text{diag}[\hat{p}_i(1-\hat{p}_i)]$.
[b] $\tau^2 = 1$ dans un modèle logistique.
[c] Matrice diagonale qui contient des poids de sondage $w_i\gamma_i$ sur la diagonale principale.
[d] $\hat{\boldsymbol{\Delta}} = \text{diag}\{[\hat{p}_i(1-\hat{p}_i)]^{-1}\}$.
- MCG = moindres carrés généralisés; MCPG = moindres carrés pondérés généralisés; SWGLM = modèles linéaires généralisés avec poids de sondage; VIF = facteurs d'inflation de la variance.

Tableau 4.2
Estimations des paramètres et leurs erreurs-types connexes (entre parenthèses) pour les modèles complets et réduits à l'aide de trois méthodes de régression différentes

Variable	MCG		MCPG		SWGLM	
Modèle complet						
(Ordonnée à l'origine)	-0,887	**	-	*	-1,043	*
	(0,297)		(0,456)		(0,470)	
Autre race	-0,125	*	0,111		0,111	
	(0,060)		(0,082)		(0,108)	*
Noir	0,276	***	0,396	***	0,396	**
	(0,069)		(0,088)		(0,108)	
Genre	0,080		0,049		0,049	
	(0,173)		(0,256)		(0,250)	
Âge (années)	0,002		0,012		0,012	
	(0,006)		(0,010)		(0,009)	
Gender*Age			0,001		0,001	
	(0,004)		(0,006)		(0,005)	
Calories	0,006	***	0,007	***	0,007	***
	(0,001)		(0,002)		(0,002)	
Protéines	-0,027	***	-0,031	***	-0,031	***
	(0,005)		(0,008)		(0,007)	
Glucides	-0,026	***	-0,029	***	-0,029	***
	(0,005)		(0,007)		(0,006)	
Sucres	0,001	*	0,001		0,001	
	(0,001)		(0,001)		(0,001)	
Lipides totaux	0,007		0,023		0,023	
	(0,016)		(0,024)		(0,024)	
Acides gras saturés totaux	-0,057	***	-0,075	***	-0,075	**
	(0,014)		(0,021)		(0,023)	
Acides gras mono-insaturés totaux	-0,068	***	-0,101	***	-0,101	***
	(0,013)		(0,020)		(0,021)	
Acides gras polyinsaturés totaux	-0,060	***	-0,086	***	-0,086	**
	(0,013)		(0,019)		(0,023)	
Alcool	-0,045	***	-0,052	***	-0,052	***
	(0,008)		(0,013)		(0,011)	

Notes : - Valeurs p : *** $0 < p \leq 0{,}001$; ** $0{,}001 < p \leq 0{,}01$; * $0{,}01 < p \leq 0{,}05$.
- MCG = moindres carrés généralisés; MCPG = moindres carrés pondérés généralisés; SWGLM = modèles linéaires généralisés avec poids de sondage.

Tableau 4.2 (suite)
Estimations des paramètres et leurs erreurs-types connexes (entre parenthèses) pour les modèles complets et réduits à l'aide de trois méthodes de régression différentes

Variable	MCG		MCPG		SWGLM	
Modèle réduit						
(Ordonnée à l'origine)	-0,855	***	-0,966	***	-0,966	***
	(0,099)		(0,149)		(0,187)	
Noir	0,407	***	0,408	***	0,408	***
	(0,059)		(0,077)		(0,091)	
Âge en années	0,008	***	0,012	***	0,012	**
	(0,002)		(0,003)		(0,004)	
Acides gras saturés totaux	0,013	***	0,022	***	0,022	***
	(0,003)		(0,004)		(0,005)	
Acides gras mono-insaturés totaux	-0,008	**	-0,017	***	-0,017	**
	(0,003)		(0,004)		(0,005)	
Alcool	-0,004	***	-0,004	*	-0,004	*
	(0,001)		(0,002)		(0,001)	

Notes : - Valeurs p : *** $0 < p \leq 0{,}001$; ** $0{,}001 < p \leq 0{,}01$; * $0{,}01 < p \leq 0{,}05$.
- MCG = moindres carrés généralisés; MCPG = moindres carrés pondérés généralisés; SWGLM = modèles linéaires généralisés avec poids de sondage.

Le tableau 4.3 présente les VIF calculés à partir des équations (3.7) et (3.14). Les VIF étiquetés « FM » sont ceux fondés sur le modèle de l'équation (3.7) dans le cas des régressions qui reposent sur les poids de sondage, mais ne tiennent pas compte de la stratification et de l'échantillonnage en grappes. Les VIF étiquetés MCG, MCPG et SWGLM proviennent de (3.14). Le tableau 4.4 énumère les VIF ajustés avec l'ordonnée à l'origine qui ont été calculés à partir de (3.17) pour les modèles complet et réduit. Tous les VIF ont été calculés à l'aide du paquet R `svydiags` (Valliant, 2024). Les VIF des MCG sont calculés comme si un échantillon aléatoire simple avait été utilisé, c'est-à-dire sans utiliser de poids de sondage, de strates ni de grappes. Les MCPG intègrent uniquement les poids de sondage, tandis que le SWGLM tient compte de toutes les caractéristiques du plan.

Tableau 4.3
Valeurs des VIF sans l'ordonnée à l'origine pour le modèle de régression logistique utilisant trois méthodes d'estimation distinctes; les valeurs des VIF FM proviennent de (3.7), et les valeurs des MCG, des MCPG et du SWGLM proviennent de (3.14)

Variable	VIF FM (3.7)	VIF MCG (3.14)	VIF MCPG (3.14)	VIF SWGLM (3.14)	VIF $\zeta_k \hat{\varrho}_k$ (3.14)
Modèle complet					
Autre race	1,5	2,6	2,3	5,3	3,42
Noir	1,2	1,8	1,8	3,3	2,67
Genre	118,0	123,8	114,8	106,3	0,90
Âge en années	115,9	119,9	113,4	113,4	0,98
Genre*Âge	117,1	120,0	110,5	107,7	0,92
Calories	11,469,7	11,776,7	12,405,7	19,545,3	1,70
Protéines	320,6	327,3	337,0	511,9	1,60
Glucides	2,494,7	2,630,2	2,728,0	4,174,7	1,67
Sucres	12,0	11,8	11,5	16,2	1,35
Lipides totaux	4,187,4	3,833,8	3,934,1	6,005,3	1,43
Acides gras saturés totaux	348,6	311,2	323,4	380,3	1,09
Acides gras mono-insaturés totaux	346,0	322,3	333,2	427,1	1,23
Acides gras polyinsaturés totaux	162,8	148,3	144,1	162,5	1,00
Alcool	76,9	78,8	78,0	109,8	1,43

Note : FM = fondés sur un modèle; MCG = moindres carrés généralisés; MCPG = moindres carrés pondérés généralisés; SWGLM = modèles linéaires généralisés avec poids de sondage; VIF = facteurs d'inflation de la variance.

Tableau 4.3 (suite)
Valeurs des VIF sans l'ordonnée à l'origine pour le modèle de régression logistique utilisant trois méthodes d'estimation distinctes; les valeurs des VIF FM proviennent de (3.7), et les valeurs des MCG, des MCPG et du SWGLM proviennent de (3.14)

Variable	VIF FM (3.7)	VIF MCG (3.14)	VIF MCPG (3.14)	VIF SWGLM (3.14)	VIF $\zeta_k \hat{\varrho}_k$ (3.14)
Modèle réduit					
Noir	1,1	1,3	1,5	2,1	1,86
Âge en années	11,4	11,6	10,7	13,9	1,22
Acides gras saturés totaux	12,2	11,8	12,3	20,3	1,66
Acides gras mono-insaturés totaux	12,9	12,6	12,9	21,5	1,67
Alcool	1,2	1,2	1,2	1,1	0,93

Note : FM = fondés sur un modèle; MCG = moindres carrés généralisés; MCPG = moindres carrés pondérés généralisés; SWGLM = modèles linéaires généralisés avec poids de sondage; VIF = facteurs d'inflation de la variance.

Tableau 4.4
Valeurs des VIF ajustés avec l'ordonnée à l'origine à partir de (3.17) pour le modèle de régression logistique utilisant trois méthodes d'estimation distinctes

Variable	VIF MCG	VIF MCPG	VIF SWGLM	VIF $\zeta_k \hat{\varrho}_k$
Modèle complet				
Autre race	1,4	1,3	1,4	1,27
Noir	1,4	1,3	1,5	1,41
Genre	11,7	10,8	8,7	0,76
Âge en années	10,4	10,7	8,3	0,80
Genre*Âge	20,3	19,3	16,1	0,77
Calories	2,140,7	2,225,3	3,390,9	1,77
Protéines	73,2	76,2	117,3	1,75
Glucides	529,9	539,7	754,0	1,52
Sucres	4,0	3,8	5,1	1,26
Lipides totaux	956,3	980,8	1,334,2	1,39
Acides gras saturés totaux	90,9	92,9	98,1	1,06
Acides gras mono-insaturés totaux	87,4	89,7	99,3	1,14
Acides gras polyinsaturés totaux	46,1	44,2	41,8	0,86
Alcool	69,0	67,7	100,1	1,51
Modèle réduit				
Noir	1,0	1,0	1,0	0,99
Âge en années	1,0	1,0	1,0	0,99
Acides gras saturés totaux	3,4	3,5	5,1	1,57
Acides gras mono-insaturés totaux	3,4	3,5	4,8	1,48
Alcool	1,0	1,0	1,0	1,01

Note : MCG = moindres carrés généralisés; MCPG = moindres carrés pondérés généralisés; SWGLM = modèles linéaires généralisés avec poids de sondage; VIF = facteurs d'inflation de la variance.

Une règle empirique citée précédemment prévoit qu'un VIF sans l'ordonnée à l'origine de 10 ou plus indique une colinéarité dommageable (Kutner et coll., 2004; Marquardt, 1970). Certains VIF sont extrêmement élevés dans le modèle complet. Les VIF du SWGLM ajustés avec l'ordonnée à l'origine pour les calories, les glucides et les lipides totaux sont respectivement de 3 391, 754 et 1 334 dans le tableau 4.4. Les VIF sans l'ordonnée à l'origine figurant dans le tableau 4.3 pour les calories, les glucides et les lipides totaux sont encore plus élevés, à savoir 19 545, 4 175 et 6 005. Les VIF sans ordonnée à l'origine figurant dans le tableau 4.3 sont toujours plus élevés que les VIF ajustés avec l'ordonnée à l'origine qui sont présentés dans le tableau 4.4.

Comme le montre le tableau 4.1, le VIF du SWGLM peut être obtenu en multipliant le VIF des MCPG par le coefficient de correction $\zeta_k \varrho_k$. Pour certains coefficients, par exemple les calories dans le modèle complet avec $\zeta_k \varrho_k = 1{,}77$ dans le tableau 4.4, cette correction peut être importante, ce qui signifie que le VIF des MCPG serait trop faible. Les VIF du SWGLM sont souvent plus élevés que ceux des VIF FM, des MCG et des MCPG; toutefois, si l'une de ces valeurs est élevée, toutes les valeurs seront élevées. Les VIF FM, les MCG et les MCPG présentent tous à peu près les mêmes renseignements diagnostiques dans cet exemple, car ils sont tous trois de taille similaire.

La colinéarité peut entraîner des anomalies dans certains coefficients. Par exemple, dans le modèle complet, le signe des acides gras saturés totaux est négatif et l'estimation du coefficient est très importante avec les MCG, les MCPG et les SWGLM, même si l'on sait qu'une consommation importante de gras saturés a un lien positif avec l'obésité (voir par exemple Phillips, Kesse-Guyot, McManus, Hercberg, Lairon, Planells et Roche, 2012).

Le tableau 4.5 présente les indices de conditionnement et les proportions de décomposition de la variance pour le SWGLM complet. Si toutes les valeurs explicatives du modèle étaient orthogonales, les indices de conditionnement seraient tous égaux à 1; or c'est loin d'être le cas pour le modèle complet, où l'indice de conditionnement le plus élevé est de 407,2. Les proportions inférieures ou égales à 0,01 sont illustrées par des points afin d'améliorer la lisibilité. Les colonnes de la deuxième à la dernière totalisent (approximativement) 1. Les proportions sont utilisées pour trouver les colinéarités en localisant d'abord les rangées présentant des indices de conditionnement « élevés », puis en parcourant vers le bas les colonnes des variables explicatives en y recherchant deux proportions ou plus de la même rangée qui sont importantes. Par exemple, pour les glucides, les acides gras saturés totaux et les acides gras mono-insaturés totaux, les proportions de 0,21, 0,18 et 0,35 sont associées à un indice de conditionnement de 18,7. Ainsi, ces trois variables explicatives sont relativement colinéaires. D'autre part, pour la variable explicative de l'alcool, 0,82 de sa variance est associée à un indice de conditionnement de 3,4, mais aucune autre variable explicative ne présente une proportion de variance élevée qui soit associée à cet indice de conditionnement. Ainsi, l'alcool ne semble pas présenter une colinéarité nuisible avec d'autres covariables, même si ses VIF sont de 109,8 et 100,1 dans les tableaux 4.3 et 4.4.

La manière de former un modèle réduit en supprimant les variables explicatives colinéaires n'est pas un processus unique, et des analystes différents peuvent prendre des décisions différentes. Pour cet exemple, nous avons supprimé la plupart des variables présentant les VIF les plus élevés et codé de nouveau la variable de la race-ethnicité en deux catégories : Noir et autres races-ethnicités, cette dernière étant utilisée comme point de référence. Dans le modèle réduit, tous les VIF du SWGLM ajustés avec l'ordonnée à l'origine figurant dans le tableau 4.4 sont inférieurs à 10, et le signe des acides gras saturés totaux est positif. La partie inférieure du tableau 4.5 présente les indices de conditionnement et les proportions de décomposition de la variance pour le modèle réduit. L'indice de conditionnement le plus élevé est de 8,4, ce qui est un autre indicateur de la colinéarité limitée. L'absence des variables les plus corrélées dans le modèle

réduit rend les erreurs-types des coefficients estimés sensiblement plus faibles pour les variables explicatives retenues. Par exemple, dans le modèle complet, l'erreur-type de l'âge dans le SWGLM est de 0,009 dans le tableau 4.2, contre 0,004 dans le modèle réduit; son coefficient estimé est de 0,012 tant dans le modèle complet que le modèle réduit. Cependant, la colinéarité influe sur la taille des autres coefficients. L'estimation du SWGLM pour la consommation d'alcool est de -0,052 dans le tableau 4.2 du modèle complet, mais de -0,004 dans le modèle réduit. Puisque l'erreur-type passe de 0,011 dans le modèle complet à 0,001 dans le modèle réduit, le coefficient de l'alcool reste significativement différent de zéro dans le modèle réduit.

Le fait que, dans le modèle réduit, les lipides saturés totaux et les acides gras mono-insaturés totaux ont des VIF du SWGLM ajustés avec l'ordonnée à l'origine de 5,1 et 4,8 et des VIF sans l'ordonnée à l'origine de 20,3 et 21,5 nous porte à croire que le modèle pourrait être affiné davantage afin de réduire la colinéarité. La corrélation non pondérée de ces deux variables est de 0,84, ce qui confirme qu'elles sont quelque peu redondantes. Nous n'explorons pas la possibilité d'affiner davantage le modèle dans le cas présent.

Tableau 4.5
Indices de conditionnement échelonnés provenant de (3.18) et proportions de décomposition de la variance provenant du SWGLM

Indice de condition-nement échelonné	Autre race	Noir	Genre	Âge en années	Genre* Âge	Calories	Protéines	Glucides	Sucres	Lipides totaux	Acides gras saturés totaux	Acides gras mono-insaturés totaux	Acides gras poly-insaturés totaux	Alcool
Modèle complet														
1,0	0,03	.	0,08	0,08	0,07	0,10	0,09	0,09	0,08	0,09	0,09	0,09	0,09	0,02
3,2	0,30	0,69	.	.	.	.	.	.	.	.	.	.	.	.
3,4	0,05	0,04	0,03	.	0,03	.	.	.	.	.	.	.	.	0,82
3,7	0,23	0,07	0,10	0,06	0,13	0,02	.	.	0,02	0,05	0,06	0,05	0,04	0,15
4,3	0,38	0,18	0,10	0,09	0,21	.	.	.	.	.	.	.	.	.
5,5	.	.	0,00	.	.	.	0,02	0,17	0,58	0,05	0,03	0,06	0,07	.
8,7	.	.	.	.	.	.	0,29	.	.	.	0,12	.	0,55	.
10,2	.	.	0,09	0,11	.	.	0,35	.	.	0,03	0,28	0,03	0,08	.
11,0	.	.	0,40	0,38	.	.	0,13	.	.	.	0,03	0,02	.	.
17,3	.	.	.	.	.	0,05	0,04	0,31	0,19	.	0,09	0,29	.	.
18,7	.	.	.	.	.	0,03	0,04	0,21	0,10	.	0,18	0,35	0,09	.
24,8	.	.	0,19	0,25	0,51	.	.	0,03	.	.	.	.	.	.
170,1	.	.	.	.	.	0,10	.	0,02	.	0,61	0,11	0,11	0,05	.
407,2	.	.	.	.	.	0,68	0,02	0,15	.	0,15	.	.	.	.

Modèle réduit					
	Noir	**Âge en années**	**Lipides totaux**	**Acides gras mono-insaturés totaux**	**Alcool**
1,0	0,06	0,27	0,29	0,30	0,08
1,8	0,77	.	0,01	0,00	0,23
2,0	0,17	0,03	0,06	0,05	0,69
3,7	0,00	0,71	0,16	0,13	0,00
8,4	.	0,00	0,48	0,52	.

Notes: - Les proportions inférieures ou égales à 0,01 sont illustrées par des points.
- SWGLM = modèles linéaires généralisés avec poids de sondage.

L'exemple présenté dans l'étude vise simplement à illustrer les VIF et les indices de conditionnement à l'aide d'un ensemble de données provenant d'enquêtes réelles. Chaque analyste est susceptible d'arriver

différemment à son modèle final. Dans l'exemple ci-dessus, un analyste pourrait s'intéresser particulièrement à la relation entre l'obésité et les calories, les protéines et les glucides. Ainsi, ces covariables pourraient être ajoutées une par une à un modèle réduit plutôt que de les éliminer et de n'inclure que les acides gras saturés et mono-insaturés totaux. L'analyse des données d'enquête comporte également des subtilités que nous ne tenterons pas d'aborder dans la présente étude. Dans la pratique, il peut être important de tenir compte de l'interaction des régresseurs avec les variables du plan de sondage lors de la formulation d'un modèle (voir par exemple Korn et Graubard, 1999; Little, 2004). Par exemple, la densité de population utilisée comme régresseur peut afficher une corrélation élevée avec l'appartenance à une strate dans un plan de sondage où les strates sont fondées selon un milieu urbain ou un milieu rural. Un régresseur peut présenter une corrélation élevée avec des poids (par exemple les revenus des années précédentes lors de la conception d'un plan d'enquête selon une probabilité proportionnelle à la taille). Dans d'autres plans d'enquête, la relation entre les poids de sondage et les variables explicatives potentielles peut être négligeable. Les applications particulières devraient avoir leur propre analyse détaillée afin de déterminer la meilleure façon de réaliser la modélisation.

5. Conclusion

À l'instar des modèles de régression linéaire ordinaires, les modèles linéaires généralisés (GLM) peuvent être influencés par la colinéarité des variables explicatives. La littérature sur la colinéarité dans les modèles linéaires aborde le problème de deux manières : le calcul des facteurs d'inflation de la variance (VIF) ainsi que des indices de conditionnement de la matrice des variables explicatives. La littérature antérieure sur les GLM ne traite principalement que des indices de conditionnement. Dans le présent article, nous avons mis au point de nouvelles formulations des VIF qui conviennent aux GLM ajustés à partir de données provenant d'enquêtes ou non. Un VIF mesure le degré d'ajustement de la variance d'un estimateur de paramètre qui est causé par la corrélation entre les variables prédictives, plutôt que par leur orthogonalité. Le VIF des GLM pour des données ne provenant pas d'enquêtes dépend d'un type de R^2 provenant d'une régression pondérée d'une valeur explicative x sur les autres x, le poids dépendant du paramètre de variance du modèle et de la fonction de lien dérivée. Pour les données d'enquête, le poids de la régression dépend également du poids de sondage.

La méthode VIF conventionnelle repose habituellement sur un modèle avec un seul $\mathbf{x}_k$ comme comparaison pour déterminer dans quelle mesure la variance d'un estimateur de la pente, $\boldsymbol{\beta}_k$, dans le modèle de comparaison peut être gonflé sous l'effet de la non-orthogonalité des $\mathbf{x}$. Mais, dans la pratique, les chercheurs peuvent vouloir utiliser un modèle avec à la fois $\mathbf{x}_k$ et une ordonnée à l'origine comme modèle de comparaison pour examiner les problèmes potentiels de colinéarité lors de l'ajout des variables explicatives au modèle. Nous avons également obtenu des formules de VIF ajustés avec l'ordonnée à l'origine afin de répondre à cet objectif.

De plus, nous avons élaboré une théorie pour les indices de conditionnement liés à la matrice d'information pondérée associée à un GLM estimé à partir de données d'enquêtes complexes. La matrice

d'information est le principal élément sous le modèle de variance approximative de la pente de régression estimée dans un GLM pondéré par les poids de sondage. Les indices de conditionnement complètent les VIF en facilitant le recensement des paires de variables explicatives qui conduisent à la colinéarité.

La théorie des VIF et des indices de conditionnement a été illustrée à l'aide d'une régression logistique prédisant l'obésité à partir d'un ensemble de données contenant un certain nombre de variables corrélées liées à l'apport calorique. Comme prévu, la colinéarité peut influencer à la fois la taille des coefficients estimés et leurs erreurs-types. Les fonctions R qui permettent de calculer les VIF et les indices de conditionnement pour les modèles linéaires ou les GLM se trouvent dans le paquet R `svydiags` (Valliant, 2024).

Les méthodes présentées dans la présente étude ont des limites et ne seront pas uniformément utiles dans toutes les applications. Si la prédiction est le but principal, l'inclusion de variables explicatives colinéaires peut ne pas être préoccupante tant que l'erreur de prédiction est réduite au minimum. Il en va autrement de l'inférence causale, où la taille relative des coefficients estimés peut être d'une importance cruciale. Un modèle avec interactions est un cas intéressant, car il présentera naturellement un certain degré de colinéarité en raison de la corrélation entre une interaction et ses effets principaux. Pour décider s'il convient de conserver un paramètre d'interaction avec un VIF élevé, un analyste peut tenir compte des éléments suivants :

(1) Conserver l'interaction si elle est statistiquement significative, si elle a une grande ampleur d'effet ou si elle améliore l'ajustement du modèle et le rendement de la prédiction.

(2) Envisager de la supprimer si elle n'est pas significative, si elle a une incidence minime sur le résultat ou si la multicolinéarité touche gravement l'interprétation du modèle.

(3) La régularisation peut également être utilisée si l'analyste estime que les interactions peuvent être importantes et que la qualité des prédictions l'emporte sur les préoccupations liées à la multicolinéarité.

Il convient de mentionner que certaines méthodes d'ajustement, telles que les forêts aléatoires qui reposent sur la moyenne des modèles, ne se prêtent pas à l'analyse du VIF ou de l'indice de conditionnement, car les modèles de la moyenne n'incluent pas nécessairement tous les mêmes covariables ou leurs interactions.

Les erreurs de spécification sont toujours une source de préoccupation pendant l'ajustement des modèles, surtout avec des données d'enquête. Dans de tels cas, le modèle estimé sera biaisé dans une certaine mesure, mais l'analyste utilise malgré tout le modèle qui correspond le mieux aux données, compte tenu de la forme du modèle utilisé (par exemple linéaire, logistique) et de l'ensemble de covariables qui sont saisies. Lorsque des covariables importantes sont omises parce qu'elles ont été négligées ou ne sont pas disponibles, les VIF et les indices de conditionnement permettent tout de même de déterminer les $\mathbf{x}$ qui sont colinéaires avec d'autres dans un modèle. L'examen des VIF et d'autres diagnostics comme les résidus peut permettre

d'affiner le modèle utilisé, même en cas d'erreur de spécification. Bien que les diagnostics de colinéarité aient leurs limites, ils jouent un rôle utile dans l'affinement des modèles.

Des travaux futurs dans ce domaine pourraient consister à étendre la théorie des indices de conditionnement dans les GLM aux estimateurs de variance de type modèle-plan; à étendre la théorie des VIF aux modèles de superpopulation à effets mixtes; à étudier de quelle façon les VIF et d'autres diagnostics de modèle peuvent être appliqués à l'étude des propriétés des estimateurs de statistiques descriptives telles que les totaux et les moyennes; et, au besoin, à mettre au point d'autres logiciels destinés à rendre ces outils facilement accessibles aux analystes.

Remerciements

Nous remercions le rédacteur en chef, le rédacteur associé et les examinateurs, dont les commentaires ont permis d'améliorer la présentation.

Annexe A

A.1 Partition de la matrice $\mathbf{A}$ dans le GLM et le SWGLM

Définissons $\widetilde{\mathbf{X}} = \left(\mathbf{W}\boldsymbol{\Gamma}\right)^{1/2}\mathbf{X} = \mathbf{W}_{\Gamma}^{1/2}\mathbf{X}$ et $\tilde{\mathbf{x}}_k = \mathbf{W}_{\Gamma}^{1/2}\mathbf{x}_k$. Le matrice $\mathbf{A} = \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$ dans (3.9) peut être partitionnée comme

$$\mathbf{A} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T\widetilde{\mathbf{X}}_{(k)} \\ \widetilde{\mathbf{X}}_{(k)}^T\tilde{\mathbf{x}}_k & \widetilde{\mathbf{X}}_{(k)}^T\widetilde{\mathbf{X}}_{(k)} \end{pmatrix} \tag{A.1}$$

après avoir réordonné les colonnes de $\widetilde{\mathbf{X}}$ de sorte que $\widetilde{\mathbf{X}} = \left(\tilde{\mathbf{x}}_k \;\; \widetilde{\mathbf{X}}_{(k)}\right)$ avec $\widetilde{\mathbf{X}}_{(k)}$ soit la matrice $n \times (p-1)$ contenant toutes les colonnes, à l'exception de la k^{e} colonne de $\widetilde{\mathbf{X}}$. En utilisant la formule pour l'inverse d'une matrice partitionnée (voir par exemple Theil, 1971, section 1.2) et en fixant $\mathbf{W} = \mathbf{I}_N$, nous obtenons directement (2.8). Pour la version pondérée par les poids de sondage, nous utilisons la forme plus générale qui figure dans (3.6).

Dérivation de g_{kk} dans le SWGLM, section 3.4

Pour dériver g_{kk}, nous partitionnons les composantes de $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ comme

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

où $\mathbf{A}^{-1} = [a^{lk}] \;\; (l,k = 1,\ldots,p)$, $\mathbf{a}^{k(k)}$ est la k^{e} rangée de $\mathbf{A}^{-1}$, en excluant a^{kk}, $\mathbf{a}^{(k)k} = [\mathbf{a}^{k(k)}]^T$, et où $\mathbf{A}^{(k)(k)}$ est la partie $(p-1)\times(p-1)$ de $\mathbf{A}^{-1}$, en excluant la k^{e} rangée et colonne. En utilisant la formule pour

l'inverse d'une matrice partitionnée, les composantes dans le coin supérieur gauche et le coin inférieur droit de $\mathbf{A}^{-1}$ sont les suivantes :

$$a^{kk} = \mathbf{i}_k^T \mathbf{A}^{-1} \mathbf{i}_k = \mathbf{i}_k^T \left(\mathbf{X}^T \mathbf{W}_\Gamma \mathbf{X} \right)^{-1} \mathbf{i}_k = \frac{1}{\left(1 - R^2_{\Gamma W(k)}\right) \mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k} \tag{A.2}$$

$$\mathbf{a}^{(k)k} = -a^{kk} \left(\mathbf{X}_{(k)}^T \mathbf{W}_\Gamma \mathbf{X}_{(k)} \right)^{-1} \mathbf{X}_{(k)} \mathbf{W}_\Gamma \mathbf{x}_k \equiv -a^{kk} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} \tag{A.3}$$

où $R^2_{\Gamma W(k)}$ a été défini dans la section 3.4. La matrice partitionnée $\mathbf{B}$ est :

$$\mathbf{B} = \begin{pmatrix} b_{kk} & \mathbf{b}_{k(k)} \\ \mathbf{b}_{(k)k} & \mathbf{B}_{(k)(k)} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k & \mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{X}_{(k)} \\ \mathbf{X}_{(k)}^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k & \mathbf{X}_{(k)}^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{X}_{(k)} \end{pmatrix}. \tag{A.4}$$

En multipliant les matrices dans $\mathbf{G}$ et en utilisant la symétrie de $\mathbf{A}$ et $\mathbf{B}$, nous obtenons

$$\begin{aligned} g_{kk} &= a^{kk}(a^{kk} b_{kk} + 2\mathbf{b}_{k(k)} \mathbf{a}^{(k)k}) + \mathbf{a}^{(k)kT} \mathbf{B}_{(k)(k)} \mathbf{a}^{(k)k} \\ &= (a^{kk})^2 (b_{kk} - 2\mathbf{b}_{k(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} + \hat{\boldsymbol{\beta}}_{W\Gamma(k)}^T \mathbf{B}_{(k)(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)}). \end{aligned}$$

En utilisant (A.2)-(A.4) et en suivant des étapes analogues à la dérivation du VIF pour la régression linéaire présentée à l'annexe de Liao et Valliant (2012b), nous pouvons décomposer g_{kk} en un paramètre $R^2_{W\Gamma(k)}$ avec deux coefficients de correction :

$$\begin{aligned} g_{kk} &= \left(\frac{1}{1 - R^2_{W\Gamma(k)}} \frac{1}{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k} \right)^2 \times \\ &\quad \left(\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k - 2\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} + \hat{\boldsymbol{\beta}}_{W\Gamma(k)}^T \mathbf{X}_{(k)} \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} \right) \\ &= \frac{1}{1 - R^2_{W\Gamma(k)}} \frac{\left(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} \right)^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \left(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} \right)}{\left(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} \right)^T \mathbf{W}_\Gamma \left(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} \right)} \frac{1}{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k} \\ &= \frac{1}{1 - R^2_{W\Gamma(k)}} \frac{\mathbf{e}_{xk}^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{e}_{xk}}{\mathbf{e}_{xk}^T \mathbf{W}_\Gamma \mathbf{e}_{xk}} \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k} \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k}{\left(\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k \right)^2} \\ &= \frac{\hat{\zeta}_k \hat{\varrho}_k}{1 - R^2_{W\Gamma(k)}} \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k}{\left(\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k \right)^2} \end{aligned} \tag{A.5}$$

où $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)}$ est le résidu de la régression pondérée par les poids de sondage de $\mathbf{x}_k$ sur $\mathbf{X}_{(k)}$ à l'aide de la matrice de pondération $\mathbf{W}_\Gamma$, $\hat{\zeta}_k = \frac{(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})}{(\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T \mathbf{W}_\Gamma (\mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})} = \frac{\mathbf{e}_{xk}^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{e}_{xk}}{\mathbf{e}_{xk}^T \mathbf{W}_\Gamma \mathbf{e}_{xk}}$ et $\hat{\varrho}_k = \frac{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{W}_\Gamma \mathbf{V}_\Delta \mathbf{W}_\Gamma \mathbf{x}_k}$.

A.2 Estimateur de variance dans un échantillon stratifié à plusieurs degrés

Supposons que le plan soit celui précisé à la section 3.1. Considérons un GLM avec une ordonnée à l'origine et un vecteur de pente communs à toutes les strates. Dans un échantillon, les équations d'estimation fondées sur le plan sont les suivantes :

$$\sum_{h=1}^{H}\sum_{i\in s_h}\sum_{t\in s_{hi}} w_{hit}(y_{hit}-\mu_{hit})\,\hat{\gamma}_{hit}g'(\mu_{hit})\,\mathbf{x}_{hit} \;=\; \sum_{h=1}^{H}\sum_{i\in s_h}\mathbf{X}_{hi}^{T}\mathbf{W}_{hi}\hat{\mathbf{\Gamma}}_{hi}\hat{\mathbf{\Delta}}_{hi}(\mathbf{y}_{hi}-\boldsymbol{\mu}_{hi}) \;=\; \mathbf{0}; \tag{A.6}$$

où $\mathbf{y}_{hi}=(y_{hi1},\ldots,y_{hin_{hi}})^T$, $\mu_{hi}=(\mu_{hi1},\ldots,\mu_{hin_{hi}})^T$, $\mathbf{W}_{hi}=\text{diag}(w_{hi1},\ldots,w_{hin_{hi}})$, $\hat{\mathbf{\Gamma}}_{hi}=\text{diag}(\gamma_{hi1},\ldots,\gamma_{hin_i})$, $\hat{\mathbf{\Delta}}_{hi}=\text{diag}(g'(\mu_{hi1}),\ldots,g'(\mu_{hin_i}))$, et $\mathbf{X}_{hi}(n_{hi}\times p)=(\mathbf{x}_{1hi},\ldots,\mathbf{x}_{phi})$ avec $\mathbf{x}_{khi}=(x_{khi1},\ldots,x_{khin_{hi}})^T$.

Supposons que $\mathbf{z}_{hi}=\mathbf{X}_{hi}^{T}\hat{\mathbf{\Gamma}}_{hi}\hat{\mathbf{\Delta}}_{hi}(\mathbf{y}_{hi}-\boldsymbol{\mu}_{hi})$. Pour la grappe i, définissons un vecteur pondéré de résidus $\mathbf{z}_{hi}^{*}=\mathbf{X}_{hi}^{T}\mathbf{W}_{hi}\hat{\mathbf{\Gamma}}_{hi}\hat{\mathbf{\Delta}}_{hi}(\mathbf{y}_{hi}-\boldsymbol{\mu}_{hi})$. L'estimateur de variance convergent par rapport au plan pour $\mathbf{z}^{*}$ dans ce plan d'échantillonnage stratifié à plusieurs degrés est :

$$\begin{aligned} v_\pi(\hat{\mathbf{z}}^{*}) \;&=\; \sum_{h=1}^{H}\frac{m_h}{m_h-1}\sum_{i\in s_h}\left(\mathbf{z}_{hi}^{*}-\overline{\mathbf{z}}_{h}^{*}\right)\left(\mathbf{z}_{hi}^{*}-\overline{\mathbf{z}}_{h}^{*}\right)^T \\ &=\; \sum_{h=1}^{H}\frac{m_h}{m_h-1}\left(\sum_{i\in s_h}\mathbf{z}_{hi}^{*}\mathbf{z}_{hi}^{*T}-m_h\overline{\mathbf{z}}_{h}^{*}\overline{\mathbf{z}}_{h}^{*T}\right) \\ &=\; \sum_{h=1}^{H}\frac{m_h}{m_h-1}\mathbf{X}_h^{T}\mathbf{W}_h\hat{\mathbf{\Gamma}}_h\hat{\mathbf{\Delta}}_h\left[\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^{T})-\frac{1}{m_h}\mathbf{e}_h\mathbf{e}_h^{T}\right]\hat{\mathbf{\Delta}}_h\hat{\mathbf{\Gamma}}_h\mathbf{W}_h\mathbf{X}_h, \end{aligned} \tag{A.7}$$

où $\mathbf{X}_h^{T}=(\mathbf{X}_{h1},\ldots,\mathbf{X}_{hm_h})^T$, $\mathbf{W}_h=\text{diag}(\mathbf{W}_{hi})_{i\in s_h}$, $\hat{\mathbf{\Gamma}}_h=\text{diag}(\hat{\mathbf{\Gamma}}_{hi})_{i\in s_h}$, $\hat{\mathbf{\Delta}}_h=\text{diag}(\hat{\mathbf{\Delta}}_{hi})_{i\in s_h}$, $\mathbf{e}_{hi}=\mathbf{y}_{hi}-\boldsymbol{\mu}_{hi}$ et $\mathbf{e}_h=(\mathbf{e}_{h1},\mathbf{e}_{h2},\ldots,\mathbf{e}_{hm_h})^T$ est un vecteur de résidus d'éléments dans une strate h, et $\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^{T})$ est la matrice $m\times m$ avec $\mathbf{e}_{hi}\mathbf{e}_{hi}^{T}$ sur la diagonale principale et 0 ailleurs.

De ce fait, l'estimateur de variance par linéarisation $v_L(\hat{\boldsymbol{\beta}}_{SW})$ est :

$$\begin{aligned} v_L(\hat{\boldsymbol{\beta}}_{SW}) \;&=\; \hat{\tau}^4\sum_{h=1}^{H}\mathbf{A}^{-1}\frac{m_h}{m_h-1}\mathbf{X}_h^{T}\mathbf{W}_h\hat{\mathbf{\Gamma}}_h\hat{\mathbf{\Delta}}_h\left[\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^{T})-\frac{1}{m_h}\mathbf{e}_h\mathbf{e}_h^{T}\right]\times \\ &\quad\; \hat{\mathbf{\Delta}}_h\hat{\mathbf{\Gamma}}_h\mathbf{W}_h\mathbf{X}_h\mathbf{A}^{-1} \\ &=\; \hat{\tau}^4\mathbf{A}^{-1}\mathbf{X}^{T}\mathbf{W}\hat{\mathbf{\Gamma}}\hat{\mathbf{\Delta}}\;\text{Blkdiag}\left\{\frac{m_h}{m_h-1}\left[\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^{T})-\frac{1}{m_h}\mathbf{e}_h\mathbf{e}_h^{T}\right]_{i\in s_h}\right\}_{h=1,\ldots,H}\times \\ &\quad\; \hat{\mathbf{\Delta}}\hat{\mathbf{\Gamma}}\mathbf{W}\mathbf{X}\mathbf{A}^{-1}. \end{aligned} \tag{A.8}$$

Lorsque la valeur de m_h est élevée, $\mathbf{e}_h\mathbf{e}_h^{T}/m_h\xrightarrow{p}\mathbf{0}$ dans des conditions raisonnables et

$$v_\pi(\hat{\mathbf{z}}^{*}) \;=\; \sum_{h=1}^{H}\frac{m_h}{m_h-1}\mathbf{X}_h^{T}\mathbf{W}_h\hat{\mathbf{\Gamma}}_h\hat{\mathbf{\Delta}}_h\;\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^{T})\,\hat{\mathbf{\Delta}}_h\hat{\mathbf{\Gamma}}_h\mathbf{W}_h\mathbf{X}_h. \tag{A.9}$$

Par conséquent,

$$v_L(\hat{\boldsymbol{\beta}}_{SW}) \;\doteq\; \hat{\tau}^4\sum_{h=1}^{H}\mathbf{A}^{-1}\frac{m_h}{m_h-1}\mathbf{X}_h^{T}\mathbf{W}_h\hat{\mathbf{\Gamma}}_h\hat{\mathbf{\Delta}}_h\;\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^{T})\,\hat{\mathbf{\Delta}}_h\hat{\mathbf{\Gamma}}_h\mathbf{W}_h\mathbf{X}_h\mathbf{A}^{-1}. \tag{A.10}$$

Les formules applicables aux plans plus simples, comme l'échantillonnage à un degré utilisant des poids inégaux, s'obtiennent en spécialisant (A.10).

Bibliographie

Abaidoo, R. et Agyapong, E. (2024). Financial market uncertainty and the macro economy: The role of governance and institutional quality. *Economia*, 23. DOI: 10.1108/ECON-02-2023-0034.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc, 2nd edition.

Alhakim-Naser, M., Hasan, G. et Zaifoglu, H. (2024). A geospatial analysis of flood risk zones in Cyprus: Insights from statistical and multi-criteria decision analysis methods. *Environmental Science and Pollution Research*, 31(2), 1-26. DOI: 10.1007/s11356-024-33391-x.

Belsley, D.A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician*, 38(2), 73-77.

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: John Wiley & Sons, Inc.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Binder, D.A. et Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. Dans *Analysis of Survey Data*, (Éds., R.L. Chambers et C.J. Skinner), Chapitre 3, 29-48. Wiley Series in Survey Methodology.

Chambers, R.L. et Skinner, C.J. (2003). *Analysis of Survey Data*. Wiley Series in Survey Methodology. New York: John Wiley & Sons, Inc.

Chatterjee, S. et Price, B. (1991). *Regression Analysis by Example*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc., 2nd edition.

Curtin, L.R., Mohadjer, L. et Dohrmann, S. (2012). The National Health and Nutrition Examination Survey: Sample design. *Vital and Health Statistics, National Center for Health Statistics*, 2(155).

Dobos, I. et Sasvari, P. (2024). Statistical analysis of QS World University Rankings 2021 university rankings using Scopus/SciVal databases. *Regional Statistics*, 14(4), 768-792. DOI: 10.15196/RS140407.

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

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Short, T., Venables, W., Walker, S., Winsemius, D. et Zeileis, A. (2023). *car: Companion to Applied Regression, R package version 3.1-2*. https://CRAN.R-project.org/package=car.

Gao, J., Xu, D. et Pan, Y. (2024). Configurational paths to arouse tourists' positive experiences in smart tourism destinations: A fsQCA approach. *SAGE Open*, 14(2). DOI: 10.1177/21582440241248226.

Hauck, W.W. et Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72(360), 851-853.

Hebbali, A. (2024). *olsrr: Tools for Building OLS Regression Models, R package version 0.6.0*. https://CRAN.R-project.org/package=olsrr.

Johnson, C.L., Paulose-Ram, R. et Ogden, C.L. (2012). National Health and Nutrition Examination Survey: Analytic guidelines, 1999-2010. *Vital and Health Statistics, National Center for Health Statistics*, 2(161).

Korn, E.L. et Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

Kott, P.S. (2018). A design-sensitive approach to fitting regression models with complex survey data. *Statistics Surveys*, 12, 1-17. https://doi.org/10.1214/17-SS118.

Kovács, P., Petres, T. et Tóth, L. (2005). A new measure of multicollinearity in linear regression models. *Revue Internationale de Statistique*, 73(3), 405-412.

Kutner, M., Nachtsheim, C. et Neter, J. (2004). *Applied Linear Statistical Models*. New York: McGraw-Hill/Irwin, 4th edition.

Lesaffre, E. et Marx, B.D. (1993). Collinearity in generalized linear regression. *Communications in Statistics – Theory and Methods*, 22(7), 1933-1952.

Liao, D. et Valliant, R. (2012a). Indices de conditionnement et décompositions des variances pour le diagnostic de la colinéarité dans l'analyse de données d'enquête au moyen de modèles linéaires. *Techniques d'enquête*, 38(2), 205-220. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012002/article/11757-fra.pdf.

Liao, D. et Valliant, R. (2012b). Facteurs d'inflation de la variance dans l'analyse des données d'enquêtes complexes. *Techniques d'enquête*, 38(1), 57-67. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012001/article/11685-fra.pdf.

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

Little, R. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of American Statistical Association*, 99(466), 546-556. DOI: 10.1198/016214504000000467.

Lumley, T. (2024). `survey`: *Analysis of Complex Survey Samples, R package version 4.4-2*. https://CRAN.R-project.org/package=survey.

Mackinnon, M. J. et Puterman, M. L. (1989). Collinearity in generalized linear models. *Communications in Statistics – Theory and Methods*, 18(9), 3463-3472.

Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12, 591-612.

Marx, B.D. et Smith, E.P. (1990a). Principal component estimation for generalized linear regression. *Biometrika*, 77(1), 23-32.

Marx, B.D. et Smith, E.P. (1990b). Weighted multicollinearity in logistic regression: Diagnostics and biased estimation techniques with an example from lake acidification. *Canadian Journal of Fisheries and Aquatic Sciences*, 47(6), 1128-1135.

McCullagh, P. et Nelder, J.A. (1989). *Generalized Linear Models*. Londres: Chapman and Hall, 2nd edition.

McCulloch, C.E. et Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc.

Özkale, M.R. (2021). The red indicator and corrected VIFs in generalized linear models. *Communications in Statistics – Simulation and Computation*, 50(12), 4144-4170. DOI: 10.1080/03610918.2019.1639740.

Parasyris, A., Stankovic, L. et Stankovic, V. (2024). A machine learning-driven approach to uncover the influencing factors resulting in soil mass displacement. *Geosciences*, 14(8), 220-241. DOI: 10.3390/geosciences14080220.

Pfeffermann, D. et Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. Dans *Analysis of Survey Data*, (Éds., R.L. Chambers et C.J. Skinner), Chapitre 12, 175-195. Wiley Series in Survey Methodology.

Phillips, C.M., Kesse-Guyot, E., McManus, R., Hercberg, S., Lairon, D., Planells, R. et Roche, H.M. (2012). High dietary saturated fat intake accentuates obesity risk associated with the fat mass and obesity-associated gene in adults. *Journal of Nutrition*, 142(5). DOI: 10.3945/jn.111.153460.

Roever, C., Raabe, N., Luebke, K., Ligges, U., Szepannek, G., Zentgraf, M. et Meyer, D. (2023). *klaR: Classification and Visualization, R package version 1.7-3*. https://CRAN.R-project.org/package=klaR.

Schaefer, R.L. (1986). Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computations and Simulations*, 25, 75-91.

Schaefer, R.L., Roi, L.D. et Wolfe, R.A. (1984). A ridge logistic estimator. *Communications in Statistics – Theory and Methods*, 13(1), 99-113.

Shao, J. (2003). *Mathematical Statistics*. New York: Springer, 2nd edition.

Skinner, C.J., Holt, D. et Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Wiley Series in Survey Methodology. Wiley Interscience.

Smith, E.P. et Marx, B.D. (1990). III-conditioned information matrices, generalized linear models and estimation of the effects of acid rain. *Environmetrics*, 1(1), 57-71. DOI: 10.1002/env.3170010107.

Snee, R.D. et Marquardt, D.W. (1984). Collinearity diagnostics depend on the domain of prediction, and model, and the data. *The American Statistician*, 2, 88-90.

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

Væth, M. (1985). On the use of Wald's test in exponential families. *Revue Internationale de Statistique*, 53(2), 199-214.

Valliant, R. (2024). *svydiags: Regression Model Diagnostics for Survey Data, R package version 0.7*. https://CRAN.R-project.org/package=svydiags.

Vanegas, L.H., Rondón, L.M. et Paula, G.A. (2024). *glmtoolbox: Set of Tools to Data Analysis using Generalized Linear Models*. https://CRAN.R-project.org/package=glmtoolbox.

Weissfeld, L.A. et Sereika, S.M. (1991). A multicollinearity diagnostic for generalized linear models. *Communications in Statistics – Theory and Methods*, 20(4), 1183-1198.

Zhang, K., Li, F., Li, H. et Yin, C. (2024). Revealing urban residents' intention to pay for the greening of farmland in the urban fringe by extending the theory of planned behavior: Insights from payment for ecosystem services. *Land Use Policy*, 141(2). DOI: 10.1016/j.landusepol.2024.107127.