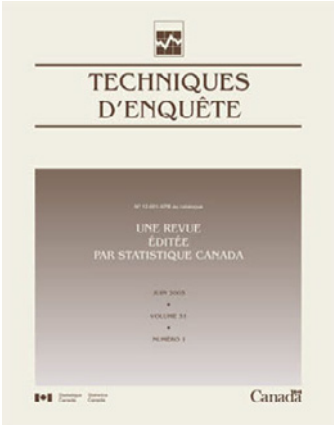
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Nouveau mode d’estimation d’un modèle logistique cumulatif avec des données d’enquêtes à plans complexes



par Phillip S. Kott et Peter Frechtel

Date de diffusion : le 27 juin 2019

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 STATCAN.infostats-infostats.STATCAN@canada.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

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

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, juin 2019
Vol. 45, nº 2, pp. 361-369
Statistique Canada, nº 12-001-X au catalogue

Nouveau mode d'estimation d'un modèle logistique cumulatif avec des données d'enquêtes à plans complexes

Phillip S. Kott et Peter Frechtel[1]

Résumé

Quand on ajuste une variable catégorique ordonnée à $L > 2$ niveaux à un ensemble de covariables sur données d'enquêtes à plans complexes, on suppose communément que les éléments de la population suivent un modèle simple de régression logistique cumulative (modèle de régression logistique à cotes proportionnelles). Cela signifie que la probabilité que la variable catégorique se situe à un certain niveau ou au-dessous est une fonction logistique binaire des covariables du modèle. Ajoutons, sauf pour l'ordonnée à l'origine, les valeurs des paramètres de régression logistique sont les mêmes à chaque niveau. La méthode « fondée sur le plan » classique servant à ajuster le modèle à cotes proportionnelles est fondée sur le pseudo-maximum de vraisemblance. Nous comparons les estimations calculées par cette méthode à celles d'un traitement dans un cadre basé sur un modèle robuste sensible au plan. Nous indiquons par un simple exemple numérique en quoi les estimations tirées de ces deux traitements peuvent différer. La nouvelle méthode peut facilement s'élargir pour ajuster un modèle logistique cumulatif général où l'hypothèse du parallélisme peut ne pas se vérifier. Un test de cette hypothèse peut aisément s'ensuivre.

Mots-clés : Hypothèse du parallélisme; estimation sensible au plan; modèle standard; modèle élargi.

1 Introduction : Ajustement d'un modèle de régression avec des données d'enquêtes à plans complexes

Notre propos est de présenter un nouveau mode d'estimation d'un modèle logistique cumulatif (aussi appelé modèle logistique ordinal ou modèle de régression ordinale), soit un modèle de régression avec une variable dépendante catégorique comptant plus de deux catégories ordonnées, sur des données d'enquêtes à plans complexes. Les méthodes standard d'estimation ne peuvent s'appliquer avec la plupart des logiciels « basés sur le plan » classiques, comme SAS (SAS Institute Inc., 2015), sauf si l'« hypothèse de parallélisme » se vérifie comme nous le verrons.

Ce sont Fuller (1975) pour la régression linéaire et Binder (1983) plus généralement qui ont introduit le cadre standard « fondé sur le plan » pour ajuster un modèle de régression à des données d'enquête. Un tel cadre traite la population finie comme la réalisation d'essais indépendants d'une population conceptuelle. En principe, il serait possible de calculer un estimateur de régression en maximum de vraisemblance à partir des valeurs d'une population finie. Avec le cadre Fuller/Binder, le but est d'estimer à partir de données d'enquête l'estimateur conceptuel du maximum de vraisemblance ou sa limite quand la population croît arbitrairement. C'est ce que Skinner (1989) appelle l'approche par « pseudo-maximum de vraisemblance ».

Kott (2018) décrit une nouvelle méthode basée sur un modèle pour estimer des modèles de régression avec des données d'enquêtes à plans complexes qu'on peut qualifier d'estimation par modèle robuste « sensible au plan ». À la suite de Kott (2007), nous définissons le *modèle standard* de cette approche de la manière suivante :

1. Phillip S. Kott et Peter Frechtel, RTI International, 6110, Executive Blvd., Rockville, MD 20852, États-Unis. Courriel : pkott@rti.org.

$$y_k = f(\mathbf{x}_k^T\boldsymbol{\beta}) + \varepsilon_k, \text{ où } E(\varepsilon_k | \mathbf{x}_k) = 0. \tag{1.1}$$

Sous un aspect très général, le modèle standard impose une restriction clé dans l'équation (1.1) : $E(\varepsilon_k) = 0,$ quelle que soit la valeur de $\mathbf{x}_k$. L'hypothèse peut ne pas se vérifier et le modèle standard pourrait ne pas convenir à la population analysée.

Dans le *modèle élargi,* $E(\varepsilon_k | \mathbf{x}_k) = 0$ à l'équation (1.1) est remplacé par $E(\mathbf{x}_k\varepsilon_k) = \mathbf{0}$. À la différence du modèle standard, le modèle élargi robuste plus général échoue rarement.

Avec une population indépendante et identiquement distribuée (iid) U de N éléments, on voit d'emblée que

$$p\lim\left\{N^{-1}\sum\nolimits_U [y_k - f(\mathbf{x}_k^T\boldsymbol{\beta})]\,\mathbf{x}_k\right\} = \mathbf{0}$$

sous le modèle élargi. Avec un échantillon complexe S à poids $\{w_k\},$ dont chacun est (presque) égal à l'inverse de la probabilité de sélection de l'élément correspondant,

$$p\lim\left\{N^{-1}\sum\nolimits_S w_k\,[y_k - f(\mathbf{x}_k^T\boldsymbol{\beta})]\,\mathbf{x}_k\right\} = \mathbf{0} \tag{1.2}$$

sous des conditions légères sur le plan de sondage. Le « presque » entre parenthèses doit être ajouté lorsque les poids comprennent des corrections pour non-réponse totale ou erreurs de couverture dans la base de sondage dont l'analyste suppose qu'on a tenu compte d'une manière asymptotiquement non biaisée. Les corrections de poids par calage à des fins d'efficacité statistique sont une autre raison d'ajouter ce « presque ».

Que l'analyste suppose que le modèle standard ou le modèle élargi vaut pour la population, résoudre pour $\mathbf{b}$ dans l'*équation d'estimation pondérée* (Godambe et Thompson, 1986)

$$\sum\nolimits_S w_k\,[y_k - f(\mathbf{x}_k^T\mathbf{b})]\,\mathbf{x}_k = \mathbf{0} \tag{1.3}$$

procure un estimateur cohérent pour $\boldsymbol{\beta}$ sous des conditions légères.

L'équation d'estimation en pseudo-maximum de vraisemblance dans Binder est

$$\sum\nolimits_S w_k \frac{f'(\mathbf{x}_k^T\mathbf{b})}{v_k}[y_k - f(\mathbf{x}_k^T\mathbf{b})]\,\mathbf{x}_k = \mathbf{0},$$

où $v_k = E(\varepsilon_k^2 | \mathbf{x}_k)$. Pour une régression logistique, de Poisson ou linéaire par les moindres carrés ordinaires (MCO), $f'(\mathbf{x}_k^T\boldsymbol{\beta})/v_k = 1.$ Cette égalité pourrait néanmoins ne pas se vérifier dans le cas d'une régression linéaire par les moindres carrés généraux (MCG), et ce, même lorsque les éléments ne sont pas corrélés. Elle peut aussi ne pas tenir pour un modèle de régression logistique cumulative.

Celui-ci est un modèle de régression logistique multinomiale pour L catégories dans un ordre naturel (toujours, souvent, parfois, jamais, etc.). L'appartenance à la première catégorie, à la première ou deuxième et à la première, deuxième ou troisième, etc., suivrait par hypothèse dans chaque cas un modèle logistique.

Le *modèle logistique cumulatif général* est (si on sépare l'ordonnée à l'origine du reste des covariables)

$$E(y_{\ell k} \mid \mathbf{x}_k) = \frac{\exp(\alpha_\ell + \mathbf{x}_k^T \boldsymbol{\beta}_\ell)}{1 + \exp(\alpha_\ell + \mathbf{x}_k^T \boldsymbol{\beta}_\ell)} \text{ pour } \ell = 1, \ldots, L-1,$$

où $y_{\ell k} = 1$ quand k appartient à une des premières catégories ℓ (la valeur est 0 dans les autres cas). L'*hypothèse du parallélisme* est que $\boldsymbol{\beta}_\ell = \boldsymbol{\beta}$ pour toutes les valeurs entières de ℓ inférieures à L, chacune des valeurs en question ayant sa propre ordonnées à l'origine (α_ℓ). Avec l'hypothèse du parallélisme, le modèle logistique cumulatif est souvent ce qu'on appelle un *modèle à cotes proportionnelles*. Nous emploierons le terme « modèle logistique cumulatif simple », bien qu'on parle couramment du modèle logistique cumulatif (ou du modèle logistique ordinal).

On peut trouver les a_ℓ et $\mathbf{b}_\ell$ qui satisfont l'équation d'estimation :

$$\sum_{k \in S} w_k \left[y_{\ell k} - \frac{\exp(a_\ell + \mathbf{x}_k^T \mathbf{b}_\ell)}{1 + \exp(a_\ell + \mathbf{x}_k^T \mathbf{b}_\ell)} \right] \begin{bmatrix} 1 \\ \mathbf{x}_k \end{bmatrix} = \mathbf{0} \text{ pour } \ell = 1, \ldots, L-1 \tag{1.4}$$

pour estimer le modèle logistique cumulatif général. Ce n'est *pas* là l'équation d'estimation en pseudo-maximum de vraisemblance dans la routine *surveylogistic* de SAS/STAT 14.1 (An (2002, page 7) examine l'équation d'estimation en pseudo-maximum de vraisemblance multivariée ajustée par cette procédure) dans la routine *logistic* de SUDAAN 11 (Research Triangle Institute, 2012) ou dans la routine *gologit2* de STATA (Williams, 2005) pour le modèle logistique cumulatif simple. Seule la routine de STATA permet la variation de $\mathbf{b}_\ell$.

Avec L catégories *nominales* et des données d'enquêtes à plans complexes, SAS et SUDAAN *peuvent* ajuster le *modèle logistique multinomial général*,

$$E(y_{\ell k} \mid \mathbf{x}_k) = \frac{\exp(\alpha_\ell + \mathbf{x}_k^T \boldsymbol{\beta}_\ell)}{1 + \sum_{j=1}^{L-1} \exp(\alpha_j + \mathbf{x}_k^T \boldsymbol{\beta}_j)} \text{ pour } \ell = 1, \ldots, L-1,$$

avec $y_{\ell k} = 1$, où k est dans la ℓ^{e} catégorie, 0 sinon. Ce n'est pas là la même chose que le modèle logistique *cumulatif* général que ces mêmes programmes ne peuvent estimer avec des données d'enquêtes à plans complexes.

Dans la suite du texte, nous présentons un modeste exemple de modèle logistique cumulatif simple. Avec des données d'enquêtes à plans complexes, nous ajustons le modèle tant par la technique de pseudo-maximum de vraisemblance que par l'équation (1.4). Dans ce dernier cas, nous créons un jeu de données avec $L-1$ observations pour chaque répondant k (à noter que $y_{1k}, \ldots, y_{L-1k}$ sont dans la même unité primaire d'échantillonnage). Nous suivons Kott (2018) et appelons ce mode d'ajustement la technique « sensible au plan », bien que, strictement parlant, il soit basé sur un modèle. L'approche par pseudo-maximum de vraisemblance est également sensible aux poids de sondage et aux autres aspects du plan de sondage.

L'article teste ensuite l'hypothèse du parallélisme. Un exemple simple sera présenté à la section 2 et une discussion conclura l'article à la section 3.

2 Un exemple simple

La *National Survey on Drug Use and Health* (NSDUH) est une enquête annuelle auprès de la population civile hors établissement âgée de 12 ans et plus aux États-Unis. À l'aide des données de 2006 à 2010 de la NSDUH, nous nous attachons à une question d'enquête posée à des adolescents (de 12 à 17 ans) ayant reçu un traitement contre la dépression dans la dernière année :

Dans les 12 derniers mois, à quel point le traitement ou le counselling vous a-t-il aidé?

Les réponses viables étaient pas du tout (l), un peu (2), dans une certaine mesure (3), beaucoup (4) et extrêmement (5).

Nous avons écarté les réponses manquantes ou non valides tant à cette question qu'à la question de savoir si le répondant avait reçu ou non un traitement contre la dépression la dernière année. Nous reviendrons sur cette pratique à la section « Discussion ».

En nous servant de SAS, nous avons estimé le modèle logistique cumulatif simple qui suit :

$$E\left(y_{\ell k} \mid \mathbf{x}_k\right) = \frac{\exp\left(\alpha_\ell + meds_k \beta\right)}{1 + \exp\left(\alpha_\ell + meds_k \beta\right)} \text{ pour } \ell = 1, \ldots, L-1, \tag{2.1}$$

où $meds = 1$ quand le répondant k avait des médicaments contre la dépression (la valeur est 0 dans les autres cas) avec l'approche en pseudo-maximum de vraisemblance et avec la technique sensible au plan. Dans le cas de l'estimation en pseudo-maximum de vraisemblance, nous avons inversé l'ordre des réponses avec $y_{1k} = 1$ où k a répondu que le traitement (ou le counselling) l'avait extrêmement aidé, avec $y_{2k} = 1$ où k a répondu que le traitement l'avait extrêmement ou beaucoup aidé, avec $y_{3k} = 1$ où k a répondu que le traitement l'avait aidé plus qu'un peu, et avec $y_{4k} = 1$ où k a répondu que le traitement l'avait aidé au moins un peu. Il y avait enfin $y_{5k} = 1 - y_{4k} = 1$ où k a répondu que le traitement ne l'avait pas aidé du tout. Dans SAS, cela veut dire que la variable dépendante Y était égale à 1 en cas d'aide extrême, à 2 en cas d'aide importante, etc., jusqu'à 5 en cas d'aide tout à fait inefficace.

Pour la technique srensible au plan, nous avons créé quatre observations à partir de k dans un nouveau jeu de données. Avec la i^{e} observation notée C = i dans SAS, variable (catégorique) de classe ajoutée à l'énoncé du modèle, nous avons créé une variable dépendante (D) égale à y_{ik} à l'équation (2.1). Nous devions ajouter EVENT = « 1 » après D dans l'énoncé du modèle, parce que nous modélisons quand D = 1.

Le code SAS des deux techniques d'estimation figure en annexe. Le jeu de données NSDUH que nous avons utilisé comptait 60 strates de variance avec deux unités primaires d'échantillonnage (UPE) de variance dans chacune et des poids d'analyse fondés sur les probabilités de sélection et la réponse des unités.

Nous présentons respectivement aux tableaux 2.1 et 2.2 les estimations des paramètres obtenues avec SAS avec les approches en pseudo-maximum de vraisemblance et sensible au plan. Au tableau 2.1, l'*Ordonnée à l'origine = i* est l'estimation en pseudo-maximum de vraisemblance de α_{ik} à l'équation (2.1). La somme de l'*Ordonnée à l'origine* et C = i au tableau 2.2 est l'estimation sensible au plan pour α_{ik} quand $i = 1, 2$ ou 3. L'estimation sensible au plan pour α_{4k} est l'ordonnée à l'origine au tableau 2.2,

moins la somme $[C = 1] + [C = 2] + [C = 3]$. Enfin (et plus simplement), *meds* dans les deux tableaux estime β.

Dans tous les cas, les estimations d'un même paramètre sont proches dans les deux tableaux. La hausse en pourcentage de chaque niveau de satisfaction à l'égard du traitement par administration de médicaments contre la dépression (estimation pour β) est en gros de 45 % (dans notre examen des résultats des régressions logistiques, nous traitons les différences de cotes logarithmiques comme égales aux différences en pourcentage des cotes, bien que ce ne soit qu'approximativement vrai). La quasi-égalité fait voir que l'hypothèse de parallélisme n'est pas violée par nos données NSDUH.

Tableau 2.1
Estimation en pseudo-maximum de vraisemblance du modèle logistique cumulatif simple

Paramètre	Estimation	Erreur-type	Valeur t	Pr > | t |
Ordonnée à l'origine 1	-2,2917	0,0913	-25,10	< 0,0001
Ordonnée à l'origine 2	-0,7617	0,0685	-11,11	< 0,0001
Ordonnée à l'origine 3	0,2511	0,0624	4,02	0,0002
Ordonnée à l'origine 4	1,3695	0,0739	18,53	< 0,0001
meds	0,4516	0,0965	4,68	< 0,0001

NOTA : Le nombre de degrés de liberté des tests *t* est de 60.

Tableau 2.2
Estimation sensible au plan du modèle logistique cumulatif simple

Paramètre	Estimation	Erreur-type	Valeur t	Pr > | t |
Ordonnée à l'origine	-0,3591	0,0583	-6,16	< 0,0001
C 1	-1,9329	0,0592	-32,63	< 0,0001
C 2	-0,4039	0,0356	-11,33	< 0,0001
C 3	0,6087	0,0392	15,52	< 0,0001
meds	0,4498	0,0955	4,71	< 0,0001

NOTA : Le nombre de degrés de liberté des tests *t* est de 60.

On peut directement vérifier l'hypothèse du parallélisme en ajoutant une variable de classe M au jeu de données sensible au plan avec :

$$M = 1 \text{ si } C = 1 \text{ et } meds = 1,$$
$$M = 2 \text{ si } C = 2 \text{ et } meds = 1,$$
$$M = 3 \text{ si } C = 3 \text{ et } meds = 1, \text{ et}$$
$$M = 4 \text{ dans les autres cas.}$$

Si elle est ajoutée à l'énoncé du modèle dans SAS, la variable de classe M appréhende les différences d'incidence de l'administration de médicaments contre la dépression dans la dernière année sur les niveaux de satisfaction à l'égard du traitement. Ainsi, la hausse des cotes estimée en pourcentage d'une satisfaction extrême est, selon le tableau 2.3 de 0,3816 (à partir de *meds*), plus 0,0717 (à partir de M = 1) ou de 45,33 %. Les autres hausses en pourcentage sont moindres, mais aucune n'est statistiquement différente des autres.

Nous le constatons par la valeur F extrêmement faible de M au tableau 2.4. Ajoutons qu'aucune des valeurs t pour un M au tableau 2.3 n'est significative même au niveau 0,5 (10 fois la valeur standard de 0,05).

Tableau 2.3
Estimation du modèle logistique cumulatif général

Paramètre	Estimation	Erreur-type	Valeur t	Pr > | t |
Ordonnée à l'origine	-0,2919	0,1270	-2,30	0,0251
C 1	-1,9636	0,0806	-24,37	< 0,0001
C 2	-0,4104	0,0440	-9,33	< 0,0001
C 3	0,6202	0,0490	12,66	< 0,0001
Meds	0,3816	0,1452	2,63	0,0109
M 1	0,0717	0,1273	0,56	0,5754
M 2	0,0234	0,0652	0,36	0,7215
M 3	-0,0236	0,0719	-0,33	0,7439

NOTA : Le nombre de degrés de liberté des tests *t* est de 60.

Tableau 2.4
Tests F du modèle logistique cumulatif général

Effet	Valeur F	DL numérateur	DL dénominateur	Pr > F
C	280,39	3	58	< 0,0001
Meds	6,91	1	60	0,0109
M	0,16	3	58	0,9239

3 Discussion

S'il y a plus d'une variable explicative dans le modèle logistique cumulatif, chacune doit être testée comme *meds* l'a été à la section précédente. Il s'agit d'ajouter une variable de classe analogue dans chaque cas. On peut effectuer un test F général pour vérifier si chaque variable de classe n'est pas significative (au niveau 0,05, disons). Une meilleure approche avec des données d'enquêtes à plans complexes consisterait à suivre Korn et Graubard (1990) et à employer le test t simple ajusté de Bonferroni. Pour une signification au niveau 0,05, nous calculerons les valeurs t pour chaque composante testée de chaque variable de classe ajoutée (il y en a trois au tableau 2.3), puis nous comparerons la valeur p de la plus petite à 0,05/le nombre de composantes testées.

Un avantage avec l'approche sensible au plan basée sur un modèle dans l'ajustement d'un modèle logistique cumulatif simple par rapport à une approche en pseudo-maximum de vraisemblance ne ressort pas avec nos données de la NSDUH. Quand l'hypothèse du parallélisme ne se vérifie pas et qu'un modèle élargi est ajusté, la satisfaction de la première « équation » en (1.4) nous assure que

$$\sum_{k\in S} w_k y_{\ell k} = \sum_{k\in S} w_k \frac{\exp\left(a_\ell + \mathbf{x}_k\mathbf{b}\right)}{1+\sum_{j=1}^{L-1}\exp\left(a_j + \mathbf{x}_k\mathbf{b}\right)} \text{ pour } \ell = 1, \ldots, L-1. \qquad (3.1)$$

Quand $\mathbf{x}_k$ même est une variable catégorique à plusieurs niveaux (une seule composante de $\mathbf{x}_k = (x_{k1}, \ldots, x_{kQ})$ est 1 et les autres composantes sont 0), l'équation (3.1) garantit que la moyenne pondérée de $y_{\ell k}$ pour chaque catégorie $\mathbf{x}$ (composante de $\mathbf{x}_k$) et le niveau cumulatif ℓ est égale à la valeur prédite que décrit l'équation

$$\hat{y}_{\ell k} = \exp(a_\ell + \mathbf{x}_k^T \mathbf{b}) \Big/ \left[1 + \sum_{j=1}^{L-1} \exp\left(a_j + \mathbf{x}_k^T \mathbf{b}\right)\right],$$

ce qui a tout d'une propriété raisonnable. L'équation (1.4) est simplement une extension de la propriété à un $\mathbf{x}_k$ plus général.

Dans l'exemple de la NSDUH, on pouvait voir, sans que la portée soit générale, qu'il était légèrement plus efficace d'utiliser l'approche sensible au plan que l'approche en pseudo-maximum de vraisemblance. C'est ce qu'on peut constater en comparant les valeurs t de *meds* (inverses des coefficients de variation estimés respectifs) aux tableaux 2.1 et 2.2. Si nous faisons abstraction des poids d'analyse, des strates et des grappes (en fixant les poids et les strates à 1 et en traitant chaque répondant comme une unité primaire d'échantillonnage), le résultat s'inverse, comme on pouvait s'y attendre. Le point à faire valoir ici est qu'une approche en pseudo-maximum de vraisemblance sur données d'enquêtes à plans complexes est véritablement « pseudo » (dans le cas qui nous occupe, c'est probablement vrai en raison de l'incidence des poids sur les estimations).

Enfin, le jeu de données que nous avons créé retranchait les observations répondantes avec des valeurs manquantes pour les variables dépendante et *meds*. Lorsqu'on ajuste le modèle *élargi*, cela n'est valide (les estimations résultantes sont asymptotiquement non biaisées dans ce cas) que si un répondant du champ d'enquête – un adolescent traité pour dépression l'année précédente – est retranché entièrement au hasard. Dans l'ajustement du modèle *standard*, la probabilité d'être retranché est fonction seulement du fait que l'adolescent dans l'enquête ait pris des médicaments contre la dépression l'année précédente, et de rien d'autre. On userait alors de prudence en ajoutant au modèle des variables qui ne sont jamais manquantes, même quand elles ne sont pas significatives. Si nous ajoutons des variables de classe pour l'âge, le sexe, la race-ethnicité, l'appartenance urbaine et le revenu familial (toutes font l'objet d'une imputation si elles manquent dans le cadre de la NSDUH) à notre modèle logistique cumulatif simple, aucune n'est significative au niveau 0,05. Les principaux résultats ne changent pas outre mesure (l'estimation de β monte d'environ 0,45 à 0,50) bien que la valeur t de meds soit moindre dans l'approche sensible au plan ($b_{meds} = 0{,}4948$; $t_{meds} = 5{,}49$) que dans l'approche en pseudo-maximum de vraisemblance ($b_{meds} = 0{,}4987$; $t_{meds} = 5{,}52$).

Annexe

/* PML est un jeu de données pour des adolescents ayant répondu à la NSDUH dans les années d'enquête 2006 à 2010 et déclaré avoir été traités contre la dépression et avoir pris des médicaments contre cette maladie. Les variables sont les suivantes :

Y = 1, traitement extrêmement utile; Y = 2, traitement largement utile; Y = 3, traitement utile dans une certaine mesure; Y = 4, traitement un peu utile; Y = 5, traitement inutile;
meds = 1 en cas d'administration de médicaments contre la dépression, 0 dans les autres cas;
VESTR, strate de variance;
VEPSU, unité primaire d'échantillonnage de variance;
IDNUM, numéro d'identification du répondant;
ANALWT, poids d'analyse.

Ce jeu de données sert à l'estimation en pseudo-maximum de vraisemblance du modèle logistique cumulatif simple et à la création du jeu de données DS_SIMPLE, qui permet d'estimer ce même modèle logistique cumulatif simple avec une approche sensible au plan. On l'emploie pour créer le jeu de données DS_GENERAL pour une estimation sensible au plan du modèle logistique cumulatif général. */

```
DATA DS_SIMPLE; SET PML; BY VESTR VEPSU IDNUM;
D = 0;
C = 1; IF Y < 2 THEN D = 1; OUTPUT;
C = 2; IF Y < 3 THEN D = 1; OUTPUT;
C = 3; IF Y < 4 THEN D = 1; OUTPUT;
C = 4; IF Y < 5 THEN D = 1; OUTPUT;

DATA DS_GENERAL; SET DS_SIMPLE;
M = 4;
IF C = 1 AND MEDS = 1 THEN M = 1;
IF C = 2 AND MEDS = 1 THEN M = 2;
IF C = 3 AND MEDS = 1 THEN M = 3;

/*PROC ci-après sert à produire le tableau 2.1*/

PROC SURVEYLOGISTIC DATA = PML; CLUSTER VEPSU;
MODEL Y = MEDS;
STRATA VESTR; WEIGHT ANALWT; RUN;

/*PROC ci-après sert à produire le tableau 2.2*/

PROC SURVEYLOGISTIC DATA = DS_SIMPLE; CLASS C;
CLUSTER VEPSU;
MODEL D(EVENT = '1') = C MEDS;
STRATA VESTR; WEIGHT ANALWT; RUN;

/*PROC ci-après sert à produire les tableaux 2.3 et 2.4*/

PROC SURVEYLOGISTIC DATA =DS_GENERAL; CLASS M C;
CLUSTER VEPSU;
MODEL D(EVENT = '1') = C MEDS M;
STRATA VESTR; WEIGHT ANALWT; RUN.
```

Bibliographie

An, A. (2002). Performing logistic regression on survey data with the new SURVEYLOGISTIC procedure. Dans *Proceedings of the Twenty-Seventh Annual SAS® Users Group International Conference*, Cary, NC: SAS Institute Inc. (http://www2.sas.com/proceedings/sugi27/p258-27.pdf).

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

Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhyā-The Indian Journal of Statistics, Series C*, 37, 117-132.

Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *Revue Internationale de Statistique*, 54(2). 127-138.

Korn, E L., et Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *American Statistician*, 44, 270-276.

Kott, P.S. (2007). Clarifying some issues in the regression analysis of survey data. *Survey Research Methods*, 1, 11-18.

Kott, P.S. (2018). A design-sensitive approach to fitting regression models with complex survey data. *Statistics Surveys*, 12, 1-17.

Research Triangle Institute (2012). *SUDAAN Language Manual*, Volumes 1 and 2, Release 11. Research Triangle Park, NC: Research Triangle Institute.

SAS Institute Inc. (2015). *SAS/STAT® 14.1 User's Guide*. Cary, NC: SAS Institute Inc.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. Dans *Analysis of Complex Surveys*, (Éds., C.J. Skinner, D. Holt et T.M.F. Smith). Chichester: John Wiley & Sons, Inc., 59-87.

Williams, R. (2005). Gologit2: A Program for Generalized Logistic Regression/Partial Proportional Odds Models for Ordinal Variables. Récupéré le 3 janvier 2016 (http://www.nd.edu/~rwilliam/strata/gologit2.pdf).