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

Techniques d’enquête

La vraisemblance pénalisée de Firth pour les régressions à risques proportionnels en cas d’enquêtes complexes

par Pushpal K. Mukhopadhyay

Date de diffusion : le 15 décembre 2020

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, décembre 2020
Vol. 46, nº 2, pp. 229-256
Statistique Canada, nº 12-001-X au catalogue

La vraisemblance pénalisée de Firth pour les régressions à risques proportionnels en cas d'enquêtes complexes

Pushpal K. Mukhopadhyay[1]

Résumé

Le présent article propose une méthode de mise à l'échelle des poids pour la vraisemblance pénalisée de Firth pour des modèles de régression à risques proportionnels. La méthode calcule une relation entre la vraisemblance pénalisée utilisant des poids mis à l'échelle et la vraisemblance pénalisée utilisant des poids non mis à l'échelle, et elle montre que la vraisemblance pénalisée utilisant des poids mis à l'échelle possède certaines propriétés souhaitables. Une étude par simulations indique que la vraisemblance pénalisée utilisant des poids mis à l'échelle produit des biais plus petits dans les estimations ponctuelles et les erreurs-types que les biais produits par la vraisemblance pénalisée utilisant des poids non mis à l'échelle. La vraisemblance pénalisée pondérée est appliquée à l'estimation des taux de risque pour les crises cardiaques au moyen d'un ensemble de données à grande diffusion provenant de la *National Health and Epidemiology Follow up Study* (NHEFS, Étude de suivi épidémiologique et de santé nationale). L'annexe contient les instructions SAS[MD] servant à estimer les taux de risque à l'aide de données d'enquêtes complexes.

Mots-clés : Vraisemblance monotone; jackknife avec suppression d'une UPE; mise à l'échelle de poids.

1 Introduction

Le modèle de régression à risques proportionnels de Cox (Cox, 1972) est couramment utilisé dans l'analyse des données de survie. Il s'agit d'un modèle semi-paramétrique qui explique l'effet des variables explicatives sur les taux de risque. Le modèle suppose que l'effet des variables explicatives a une forme linéaire, mais il permet que la fonction de survie sous-jacente ait une forme non spécifiée. On estime les paramètres du modèle en maximisant une vraisemblance partielle (Cox, 1972, 1975).

Pour estimer les paramètres canoniques dans les distributions de la famille exponentielle, Firth (1993) a proposé de multiplier la vraisemblance par la loi a priori de Jeffreys afin d'obtenir une estimation par le maximum de vraisemblance qui soit du premier ordre sans biais. La vraisemblance pénalisée prend la forme

$$L_p(\boldsymbol{\beta}) = L(\boldsymbol{\beta})|I(\boldsymbol{\beta})|^{0,5}$$

où $L(\boldsymbol{\beta})$ est la vraisemblance non pénalisée, I est la matrice d'information et $\boldsymbol{\beta}$ est un vecteur des paramètres de régression. La vraisemblance pénalisée de Firth est une technique très utile en pratique, non seulement aux fins de réduction du biais, mais aussi de correction des vraisemblances monotones.

Souvent, les modèles de régression à risques proportionnels souffrent de vraisemblances monotones, dans lesquelles la vraisemblance converge vers une valeur finie, mais un paramètre au moins diverge (Heinze, 1999). La vraisemblance pénalisée de Firth sert également à corriger les vraisemblances monotones et à obtenir des estimations de paramètres qui convergent (Heinze, 1999; Heinze et Schemper, 2001; Heinzel, Rüdiger et Schilling, 2002).

1. Pushpal K. Mukhopadhyay, Advanced Analytics Division, SAS Institute Inc., Cary (NC), États-Unis. Courriel : pushpal.mukhopadhyay@sas.com.

Bien que la vraisemblance pénalisée de Firth soit utile aux fins de réduction des biais et d'obtention d'estimations à partir de vraisemblances monotones, elle n'a pas été étudiée sur des enquêtes complexes comportant des poids inégaux. Il est raisonnable d'utiliser une vraisemblance pondérée pour les enquêtes complexes afin de compenser la pondération inégale (Fuller, 1975; Binder et Patak, 1994). Les ensembles de données d'enquête comprennent couramment des poids de sondage ou des poids d'analyse pour lesquels la somme des poids est un estimateur de la taille de la population. Toutefois, ces pondérations non mises à l'échelle ne mettent pas adéquatement à l'échelle la matrice d'information utilisée dans le terme de pénalité. Il est souhaitable que les paramètres d'une régression à risques proportionnels pour données d'enquête aient les deux propriétés suivantes.

- *Invariance :* les estimations ponctuelles et les erreurs-types pour les paramètres de régression doivent être invariantes par rapport à l'échelle des poids.
- *Précision :* la variance linéarisée de Taylor pour les paramètres de régression estimés doit être proche de la variance du jackknife avec suppression d'une UPE.

Dans l'article, nous montrons d'abord que si l'on n'utilise pas la correction de Firth, alors l'invariance et la précision sont satisfaites, mais que si la correction de Firth est utilisée avec les poids non mis à l'échelle, alors les estimations ponctuelles et les erreurs-types ne sont pas invariantes par rapport à l'échelle des poids. Autrement dit, si les poids sont multipliés par une constante et que la correction de Firth est utilisée, les estimations ponctuelles et les erreurs-types seront différentes. Nous proposons ensuite une méthode de mise à l'échelle des poids de sens commun pour démontrer que la correction de Firth utilisant des poids mis à l'échelle possède les deux propriétés souhaitables. La seule différence entre les poids mis à l'échelle et ceux non mis à l'échelle est que la somme des poids mis à l'échelle est égale à la taille de l'échantillon, alors que la somme des poids non mis à l'échelle est un estimateur de la taille de la population.

1.1 Exemple d'utilisation de poids non mis à l'échelle

Nous avons utilisé un ensemble de données tiré d'une étude portant sur 65 patients atteints de myélome qui ont été traités au moyen d'agents alkylants (Lee, Wei et Amato, 1992) pour démontrer les propriétés de la vraisemblance pénalisée de Firth utilisant des poids non mis à l'échelle. Les durées de survie en mois ont été enregistrées pour chaque patient. Les patients qui étaient vivants après la période de l'étude étaient considérés comme des données censurées. Nous disposions des variables suivantes pour chaque patient :

- Durée [Time] : durée de la survie en mois,
- Statut vital [Vstatus] : état du patient, zéro ou un, indiquant respectivement si le patient était vivant ou mort,
- LogBUN : log du niveau d'azote uréique sanguin,
- HGB : taux d'hémoglobine dans le sang.

Pour créer une vraisemblance monotone, nous avons ajouté une nouvelle variable explicative, Contrived [Artificielle], telle que sa valeur à tout moment de l'événement est la plus grande de toutes les

valeurs de l'ensemble de risques (voir l'exemple « Correction de Firth pour vraisemblance monotone » dans « The PHREG Procedure » [La procédure PHREG] dans SAS Institute Inc. (2018)). La variable Contrived [Artificielle] a la valeur 1 si la durée de survie observée est inférieure ou égale à 65; sinon, elle a la valeur 0.

Pour démontrer l'effet des poids dans la vraisemblance pénalisée de Firth, nous avons créé trois variables de poids, $w1$, $w3$ et $w5$, avec respectivement des valeurs de 1, 1 000 et 100 000 pour chaque observation. On estime les paramètres de régression à risques proportionnels en maximisant une vraisemblance pondérée comme cela est décrit dans la section 1.2. Parce que $w1$ a la valeur 1 pour toutes les observations, l'utilisation de $w1$ dans l'analyse équivaut à l'exécution de l'analyse non pondérée.

Nous avons ajusté les deux modèles à risques proportionnels suivants à l'aide de la procédure PHREG dans SAS/STAT^{MD} (voir « The PHREG Procedure » dans SAS Institute Inc. (2018) :

$$\lambda(t, \mathbf{Z}) = \lambda_0(t)\exp(\beta_1 \text{LogBUN} + \beta_2 \text{HGB})$$

$$\lambda(t, \mathbf{Z}) = \lambda_0(t)\exp(\beta_1 \text{LogBUN} + \beta_2 \text{HGB} + \beta_3 \text{Contrived})$$

où $\lambda(t)$ et $\lambda_0(t)$ sont respectivement la fonction de risque et la fonction de risque de référence. La vraisemblance pénalisée de Firth n'est pas requise pour l'ajustement du premier modèle sans la variable Contrived (la probabilité converge en trois étapes d'itération), mais le deuxième modèle contenant la variable Contrived ne converge pas sans la pénalité de Firth dans la vraisemblance. Le tableau 1.1 présente la valeur de la vraisemblance et les trois coefficients de régression pour 14 itérations. Bien que la fonction objective et les coefficients de LogBUN et HGB convergent vers une valeur finie après la quatrième itération, ceux de Contrived divergent. Il s'agit d'un exemple de vraisemblance monotone pour la variable Contrived. En raison de cette monotonie, il faut utiliser la vraisemblance pénalisée de Firth pour ajuster le deuxième modèle contenant Contrived.

Tableau 1.1
Historique de l'itération du maximum de vraisemblance montrant une vraisemblance monotone pour la variable Contrived

Nombre d'itérations	Valeur de la vraisemblance	LogBUN	HGB	Contrived
1	-140,693405	1,994882	-0,084319	1,466331
2	-137,784163	1,679468	-0,109068	2,778361
3	-136,971190	1,714061	-0,111564	3,938095
4	-136,707893	1,718174	-0,112273	5,003054
5	-136,616426	1,718755	-0,112370	6,027436
6	-136,583520	1,718829	-0,112382	7,036445
7	-136,571515	1,718839	-0,112384	8,039764
8	-136,567113	1,718841	-0,112384	9,040985
9	-136,565495	1,718841	-0,112384	10,041434
10	-136,564900	1,718841	-0,112384	11,041600
11	-136,564681	1,718841	-0,112384	12,041660
12	-136,564601	1,718841	-0,112384	13,041683
13	-136,564571	1,718841	-0,112384	14,041691
14	-136,564560	1,718841	-0,112384	15,041694

Si Contrived n'est pas utilisée comme variable explicative, les trois ensembles de poids produisent des estimations ponctuelles et des estimations de variance linéarisées de Taylor identiques (tableau 1.2). Les estimations de la variance du jackknife avec suppression d'une UPE sont également identiques pour les trois ensembles de poids. Ainsi, les estimations ponctuelles et les erreurs-types sont invariantes par rapport à l'échelle des poids quand la correction de Firth n'est pas utilisée.

Tableau 1.2
Estimations des paramètres et erreurs-types sans correction de Firth pour les trois ensembles de poids

	Estimation	Erreur type
LogBUN	1,674	0,583
HGB	-0,119	0,060

En revanche, si l'on utilise des poids non mis à l'échelle, les estimations ponctuelles pour Contrived ne sont pas invariantes par rapport à l'échelle des poids. Le tableau 1.3 présente les estimations des paramètres pour trois ensembles de poids quand Contrived est utilisée comme variable explicative (et que la vraisemblance pénalisée de Firth est appliquée). Parce que la vraisemblance n'est pas monotone (tableau 1.1) pour LogBUN et HGB, les estimations ponctuelles pour ces deux coefficients ne sont pas affectées par l'échelle des poids.

Tableau 1.3
Estimations des paramètres avec correction de Firth et poids non mis à l'échelle

	Poids $w1$		Poids $w3$		Poids $w5$	
	Estimation	**Erreur type**	**Estimation**	**Erreur type**	**Estimation**	**Erreur type**
LogBUN	1,722	0,584	1,719	1,85E-2	1,719	1,85E-3
HGB	-0,112	0,061	-0,112	1,93E-3	-0,112	1,93E-4
Contrived	3,815	1,558	10,629	1,38	14,633	1,02

Si Contrived n'est pas utilisée comme variable explicative, le rapport entre les erreurs-types du jackknife et les erreurs-types de la linéarisation de Taylor est de 1,13 et 1,10 pour les trois ensembles de poids des variables LogBUN et HGB, respectivement. Ainsi, le rapport entre la variance par la méthode du jackknife et la variance linéarisée de Taylor pour la vraisemblance non pénalisée est invariant par rapport à l'échelle des poids, et il est raisonnable de penser que le rapport est invariant quand on utilise la vraisemblance pénalisée.

1.2 Bref examen des estimations ponctuelles et des estimations de variance pour les paramètres de régression de populations finies

Avant de discuter de la méthode de mise à l'échelle des poids, nous examinerons brièvement les estimations ponctuelles et les estimations de variance pour les paramètres de régression dans une régression à risques proportionnels sur des enquêtes complexes comportant des poids inégaux. Lin et Wei (1989), Binder (1990, 1992), Lin (2000) et Boudreau et Lawless (2006) ont traité de l'estimation par la méthode du pseudo-maximum de vraisemblance des paramètres de régression à risques proportionnels

pour les données d'enquête. On trouve une description plus générale de l'estimation des paramètres de régression pour les enquêtes complexes dans Kish et Frankel (1974); Godambe et Thompson (1986); Pfeffermann (1993), Korn et Graubard (1999, chapitre 3), Chambers et Skinner (2003, chapitre 2) et Fuller (2009, section 6.5). Wolter (2007) a décrit plusieurs techniques d'estimation de la variance pour les données d'enquête.

Soit $\mathcal{U}_N = \{1, 2, \ldots, N\}$ l'ensemble des indices et $\mathcal{F}_N$ l'ensemble des valeurs pour une population finie de taille N. On suppose que la durée de survie de chaque membre de la population finie suit sa propre fonction de risque, $\lambda_i(t)$, exprimée comme suit :

$$\lambda_i(t) = \lambda(t; \mathbf{Z}_i(t)) = \lambda_0(t) \exp\left(\mathbf{Z}_i'(t)\, \boldsymbol{\beta}\right)$$

où $\lambda_0(t)$ est une fonction de risque de référence arbitraire et non spécifiée, $\mathbf{Z}_i(t)$ est un vecteur de taille P des variables explicatives de l'unité i^{e} au temps t, et $\boldsymbol{\beta}$ est un vecteur de paramètres de régression inconnus.

La fonction de vraisemblance partielle introduite par Cox (1972, 1975) élimine le risque de référence inconnu $\lambda_0(t)$ et tient compte des durées de survie censurées. Si toute la population est observée, cette fonction de vraisemblance partielle peut servir à estimer $\boldsymbol{\beta}$. Soit $\boldsymbol{\beta}_N$ l'estimateur souhaité.

En supposant un modèle de travail avec des réponses non corrélées, on obtient $\boldsymbol{\beta}_N$ en maximisant la log-vraisemblance partielle,

$$l_N(\boldsymbol{\beta}) = \sum_{i \in \mathcal{U}_N} \log\{\mathrm{L}(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)\}$$

par rapport à $\boldsymbol{\beta}$, où $\mathrm{L}(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)$ est la fonction de vraisemblance partielle de Cox.

Supposons qu'un échantillon probabiliste A_N est sélectionné dans la population finie $\mathcal{U}_N$. Soit π_i la probabilité de sélection et $w_i\,(= \pi_i^{-1})$ le poids d'échantillonnage pour l'unité i. Supposons ensuite que les variables explicatives $\mathbf{Z}_i(t)$ et la durée de survie t_i sont disponibles pour chaque unité de l'échantillon A_N. Un estimateur sans biais par rapport au plan pour la log-vraisemblance de la population finie est

$$l(\boldsymbol{\beta}) = \sum_{i \in A_N} \pi_i^{-1} \log\{\mathrm{L}(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)\} = \sum_{i \in A_N} w_i \log\{\mathrm{L}(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)\}.$$

On peut obtenir un estimateur basé sur un échantillon $\hat{\boldsymbol{\beta}}_N$ pour la quantité finie de population $\boldsymbol{\beta}_N$ en maximisant la pseudo-log-vraisemblance partielle $l(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)$ par rapport à $\boldsymbol{\beta}$. On obtient la variance fondée sur le plan pour $\hat{\boldsymbol{\beta}}_N$ en supposant que l'ensemble de valeurs de la population finie $\mathcal{F}_N$ est fixe.

La vraisemblance de Breslow pondérée peut être exprimée comme suit :

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\exp\left(\boldsymbol{\beta}' \sum_{\mathcal{D}_k} w_i \mathbf{Z}_i(t)\right)}{\left\{\sum_{\mathcal{R}_k} w_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))\right\}^{\sum_{\mathcal{D}_k} w_i}}$$

où $\mathcal{R}_k$ est l'ensemble de risques juste avant le k^{e} temps d'événement ordonné $t_{(k)}$, $\mathcal{D}_k$ est l'ensemble des personnes qui fait défaillance au temps $t_{(k)}$, et K est le nombre de temps d'événements distincts.

On obtient les estimations ponctuelles pour $\boldsymbol{\beta}$ en maximisant $l(\boldsymbol{\beta}) = \log[L(\boldsymbol{\beta})]$.

Bien que les poids suffisent aux fins d'estimation des coefficients de régression pour la population finie, on doit aussi utiliser les données de stratification et de corrélation intra-grappe pour estimer la variabilité d'échantillonnage. Afin d'estimer la variabilité d'échantillonnage, on peut utiliser la méthode de linéarisation en séries de Taylor ou une méthode de rééchantillonnage.

1.2.1 Estimateur de la variance analytique par la méthode de linéarisation en séries de Taylor

La méthode de linéarisation en séries de Taylor utilise la somme des carrés des scores résiduels pondérés pour estimer la variabilité d'échantillonnage.

Définissons $\bar{\mathbf{Z}}(\boldsymbol{\beta}, t) = \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}$, où

$$S^{(0)}(\boldsymbol{\beta}, t) = \sum_{A_N} w_i I(t_i \geq t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))$$

et

$$\mathbf{S}^{(1)}(\boldsymbol{\beta}, t) = \sum_{A_N} w_i I(t_i \geq t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t)) \mathbf{Z}_i(t).$$

Le score résiduel pour le i^{e} sujet est

$$\mathbf{u}_i(\boldsymbol{\beta}) = \Delta_i \{\mathbf{Z}_i(t_i) - \bar{\mathbf{Z}}(\boldsymbol{\beta}, t_i)\} - \sum_{j \in A_N} \left[\Delta_j \frac{w_j I(t_i \geq t_j) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t_j))}{S^{(0)}(\boldsymbol{\beta}, t_j)} \{\mathbf{Z}_i(t_j) - \bar{\mathbf{Z}}(\boldsymbol{\beta}, t_j)\} \right]$$

où Δ_i est l'indicateur d'événement.

Alors, l'estimateur de variance linéarisé de Taylor est

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}}) \, \mathrm{G} \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$$

où $\mathcal{I}(\hat{\boldsymbol{\beta}})$ est la matrice d'information observée et la $p \times p$ matrice $\mathbf{G}$ est définie comme étant

$$\mathbf{G} = \sum_{i,\, j \in A_N : i < j} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\hat{\mathbf{u}}_i}{\pi_i} - \frac{\hat{\mathbf{u}}_j}{\pi_j} \right)' \left(\frac{\hat{\mathbf{u}}_i}{\pi_i} - \frac{\hat{\mathbf{u}}_j}{\pi_j} \right)$$

où π_{ij} sont les probabilités d'inclusion conjointe pour les unités i et j.

En particulier, pour les plans d'échantillonnage en grappes stratifiés dans lesquels les UPE sont sélectionnées au moyen d'un échantillon aléatoire simple sans remise, la $p \times p$ matrice $\mathbf{G}$ se réduit à

$$\mathbf{G} = \sum_{h=1}^{H} \frac{n_h (1 - f_h)}{n_h - 1} \sum_{i=1}^{n_h} (\mathbf{e}_{hi+} - \bar{\mathbf{e}}_{h\cdot\cdot})' (\mathbf{e}_{hi+} - \bar{\mathbf{e}}_{h\cdot\cdot})$$

où $\mathbf{e}_{hi+}$ est la somme pondérée des scores résiduels, $\hat{\mathbf{u}}_{hij}$, dans la strate h et l'UPE i; $\bar{\mathbf{e}}_{h\cdot\cdot}$ est la moyenne de $\mathbf{e}_{hi+}$; n_h est le nombre d'UPE; et f_h est la fraction d'échantillonnage dans la strate h.

Ces estimateurs sont largement étudiés par la littérature sur les enquêtes-échantillons. Par exemple, Binder (1992) et Lin (2000) fournissent des conditions en vertu desquelles $\hat{\boldsymbol{\beta}}$ et $\hat{\mathbf{V}}(\hat{\beta})$ sont convergents. Chambless et Boyle (1985) ont calculé la variance fondée sur le plan et la normalité asymptotique pour des modèles à risques proportionnels discrets.

1.2.2 Estimateur de la variance par répliques au moyen de la méthode du jackknife avec suppression d'une UPE

La méthode du jackknife est une méthode d'estimation de la variance par répliques couramment utilisée en cas d'enquêtes complexes. Pour créer des rééchantillonnages, elle supprime (en lui attribuant un poids nul) une UPE à la fois de l'échantillon complet. Dans chaque rééchantillonnage, les poids d'échantillonnage des UPE restantes sont modifiés par le coefficient du jackknife α_r. Les poids modifiés sont appelés *poids de rééchantillonnage*.

Supposons que l'UPE i_r dans la strate h_r soit omise du r^{e} rééchantillonnage, alors les poids de rééchantillonnage et les coefficients du jackknife sont donnés par

$$w_{hij}^{(r)} = \begin{cases} 0 & i = i_r \text{ et } h = h_r \\ w_{hij} / \alpha_r & i \neq i_r \text{ et } h = h_r \\ w_{hij} & h \neq h_r \end{cases}$$

et $\alpha_r = \frac{n_{h_r} - 1}{n_{h_r}}$, respectivement, pour toutes les unités d'observation j dans la strate h et l'UPE i. Le nombre d'UPE dans la strate h_r est de n_{h_r}.

On peut appliquer la méthode du jackknife pour estimer les variances des paramètres de régression estimés pour le modèle de Cox parce que les paramètres du modèle sont les solutions d'un ensemble d'équations d'estimation qui sont des fonctions lisses de totaux (les fonctions de score correspondantes sont données dans la section 2). Les propriétés des estimateurs de la variance par la méthode du jackknife pour les modèles de régression à risques proportionnels sont étudiées dans Shao et Tu (1995, section 8.3).

Pour appliquer la méthode du jackknife, on estime les paramètres du modèle au moyen de l'échantillon complet et en utilisant chaque échantillon répété. Soit $\hat{\boldsymbol{\beta}}$ les coefficients de régression à risques proportionnels estimés à partir de l'échantillon complet et soit $\hat{\boldsymbol{\beta}}_r$ les coefficients de régression estimés à partir du r^{e} rééchantillonnage. Alors, la matrice de covariance de $\hat{\boldsymbol{\beta}}$ est estimée par

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \sum_{r=1}^{R} \alpha_r (\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})'.$$

Si les fractions de sondage ne sont pas ignorables, la matrice de covariance de $\hat{\boldsymbol{\beta}}$ est estimée par

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \sum_{r=1}^{R} \alpha_r (1 - f_r)(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})'$$

où $f_r = \frac{n_{h_r}}{N_{h_r}}$ est la fraction de sondage dans la strate h_r.

En pratique, on utilise les estimations de la variance par linéarisation de Taylor et les estimations de la variance par la méthode du jackknife pour construire les intervalles de confiance t de Wald avec $R - H$

degrés de liberté, où R est le nombre d'UPE (ou le nombre de rééchantillonnages) et H est le nombre de strates.

On montre facilement que l'estimateur de la variance par la méthode du jackknife équivaut algébriquement à l'estimateur linéarisé de Taylor pour les estimateurs linéaires du plan de sondage. En revanche, pour ce qui est des estimateurs non linéaires du plan de sondage, comme les coefficients de régression pour les modèles de régression à risques proportionnels, la méthode du jackknife tend à produire des estimations de la variance légèrement plus élevées que la méthode linéarisée de Taylor (Fuller, 2009).

Notons que si l'estimation de l'échantillon complet présente une vraisemblance monotone, il est très probable que la plupart des échantillons répétés présentent également des vraisemblances monotones. Il en résultera de nombreuses estimations par répliques « inutilisables ».

Les procédures d'analyse des données d'enquête dans SAS/STAT prennent en charge à la fois les méthodes d'estimation de la variance linéarisée de Taylor et d'estimation de la variance par répliques (Mukhopadhyay, An, Tobias et Watts, 2008).

2 Mise à l'échelle des poids

Soit w_i le poids de l'unité i. Nous proposons d'utiliser $\tilde{w}_i = \left(\sum_{A_N} 1 \Big/ \sum_{A_N} w_i\right) w_i = \left(n \Big/ \sum_{A_N} w_i\right) w_i$ comme poids mis à l'échelle. Par construction, les poids mis à l'échelle sont invariants par rapport à l'échelle du poids. C'est-à-dire que $\tilde{w}_i^* = \left(n \Big/ \sum_{A_N} \gamma w_i\right) \gamma w_i = \left(n \Big/ \sum_{A_N} w_i\right) w_i = \tilde{w}_i$ pour tous les $\gamma \neq 0$.

La vraisemblance pénalisée de Firth est donnée par $L_p(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) |\mathcal{I}(\boldsymbol{\beta})|^{0,5}$, où $L(\boldsymbol{\beta})$ et $\mathcal{I}(\boldsymbol{\beta})$ sont respectivement la probabilité non pénalisée et la matrice d'information. La log-vraisemblance pénalisée est

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + 0{,}5 \log(|\mathcal{I}(\boldsymbol{\beta})|).$$

En particulier, quand les poids mis à l'échelle sont utilisés, la log-vraisemblance partielle non pénalisée de Breslow (Breslow, 1974) est

$$l(\boldsymbol{\beta}) = \sum_{k=1}^{K} \left\{ \boldsymbol{\beta}' \sum_{i \in \mathcal{D}_k} \tilde{w}_i \mathbf{Z}_i(t_k) - \left(\sum_{i \in \mathcal{D}_k} \tilde{w}_i \right) \log \sum_{i \in \mathcal{R}_k} \tilde{w}_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t_k)) \right\}$$

où w_i est le poids non mis à l'échelle de l'unité i.

Notons

$$\mathbf{S}_k^{(a)}(\boldsymbol{\beta}) = \sum_{i \in \mathcal{R}_k} \tilde{w}_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t_k)) [\mathbf{Z}_i(t_k)]^{\otimes a}$$

où k est le k^{e} temps d'événement ordonné, $a = 0, 1, 2$, $[\mathbf{Z}_i(t_k)]^{\otimes 0}$ est 1, $[\mathbf{Z}_i(t_k)]^{\otimes 1}$ est le vecteur $\mathbf{Z}_i(t_k)$, et $[\mathbf{Z}_i(t_k)]^{\otimes 2}$ est la matrice $[\mathbf{Z}_i(t_k)][\mathbf{Z}_i(t_k)]'$.

La fonction de score est alors donnée par

$$\begin{aligned}\mathbf{U}(\boldsymbol{\beta}) &\equiv \left(U(\beta_1), \ldots, U(\beta_p)\right)' \\ &= \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \sum_{k=1}^{K} \left\{ \sum_{i \in \mathcal{D}_k} \tilde{w}_i \mathbf{Z}_i(t_k) - \sum_{i \in \mathcal{D}_k} \tilde{w}_i \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^0(\boldsymbol{\beta})} \right\}\end{aligned}$$

et la matrice d'information de Fisher est donnée par

$$\begin{aligned}\mathcal{I}(\boldsymbol{\beta}) &= -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \\ &= \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \tilde{w}_i \left\{ \frac{\mathbf{S}_k^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right]\left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right]' \right\}.\end{aligned}$$

Notons

$$\mathbf{Q}_{kp}^{(a)}(\boldsymbol{\beta}) = \sum_{i \in \mathcal{R}_k} \tilde{w}_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t_k)) Z_{i,p}(t_k) [\mathbf{Z}_i(t_k)]^{\otimes a}$$

où $a = 0, 1, 2$; $p = 1, \ldots, P$; et $\mathbf{Z}_i(t) = \left(Z_{i,1}(t), \ldots, Z_{i,p}(t)\right)$. Alors,

$$\begin{aligned}\frac{\partial \mathcal{I}(\boldsymbol{\beta})}{\partial \beta_p} = \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \tilde{w}_i \Bigg\{ &\left[\frac{\mathbf{Q}_{kp}^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kp}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \frac{\mathbf{S}_k^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right] \\ &- \left[\frac{\mathbf{Q}_{kp}^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kp}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right]\left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right]' \\ &- \left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right]\left[\frac{\mathbf{Q}_{kp}^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kp}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right]' \Bigg\}\end{aligned}$$

où $p = 1, \ldots, P$.

Les estimations ponctuelles et les erreurs-types linéarisées de Taylor pour la vraisemblance pénalisée sont obtenues à partir des fonctions de score et de la matrice hessienne comme le décrit la section 1.2. Les erreurs-types « jackknife » sont obtenues par la maximisation de la vraisemblance pénalisée dans chaque échantillon répété.

L'annexe 1 montre que dans certaines conditions de régularité, les estimateurs ponctuels obtenus par la maximisation de la vraisemblance pénalisée de Firth convergent par rapport au plan de sondage.

2.1 Vraisemblances pénalisées et échelle des poids

Dans la présente section, nous calculons une relation entre la log-vraisemblance pénalisée utilisant des poids mis à l'échelle et la log-vraisemblance pénalisée utilisant des poids non mis à l'échelle, et nous démontrons que la vraisemblance pénalisée de Firth utilisant des poids non mis à l'échelle ne possède pas la propriété d'invariance.

Soit $l(\boldsymbol{\beta}_{\tilde{w}}; \tilde{w})$ la log-vraisemblance utilisant des poids $\tilde{w}$, et soit $l(\boldsymbol{\beta}_w; w)$ la log-vraisemblance utilisant des poids w, où $\tilde{w}_i = \alpha w_i$ pour tous les i et $\alpha \neq 0$. La log-vraisemblance de Breslow peut s'écrire comme suit :

$$
\begin{aligned}
l(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w}) &= \sum_{k=1}^{K}\left\{\boldsymbol{\beta}_{\tilde{w}}' \sum_{i\in\mathcal{D}_k} \tilde{w}_i \mathbf{Z}_i(t_k) - \left(\sum_{i\in\mathcal{D}_k} \tilde{w}_i\right)\log \sum_{i\in\mathcal{R}_k} \tilde{w}_i \exp\left(\boldsymbol{\beta}_{\tilde{w}}' \mathbf{Z}_i(t_k)\right)\right\} \\
&= \sum_{k=1}^{K}\left\{\boldsymbol{\beta}_{\tilde{w}}' \alpha \sum_{i\in\mathcal{D}_k} w_i \mathbf{Z}_i(t_k) - \left(\alpha \sum_{i\in\mathcal{D}_k} w_i\right)\log \sum_{i\in\mathcal{R}_k} \alpha w_i \exp\left(\boldsymbol{\beta}_{\tilde{w}}' \mathbf{Z}_i(t_k)\right)\right\} \\
&= \alpha \sum_{k=1}^{K}\left\{\boldsymbol{\beta}_{\tilde{w}}' \sum_{i\in\mathcal{D}_k} w_i \mathbf{Z}_i(t_k) - \left(\sum_{i\in\mathcal{D}_k} w_i\right)\log \sum_{i\in\mathcal{R}_k} w_i \exp\left(\boldsymbol{\beta}_{\tilde{w}}' \mathbf{Z}_i(t_k)\right)\right\} \\
&\quad - \sum_{k=1}^{K}\left(\alpha \sum_{i\in\mathcal{D}_k} w_i\right)\log \alpha \\
&= \alpha l(\boldsymbol{\beta}_{\tilde{w}};\, w) - \sum_{k=1}^{K}\left(\alpha \sum_{i\in\mathcal{D}_k} w_i\right)\log \alpha.
\end{aligned}
$$

Parce que le deuxième terme du deuxième membre ne contient pas $\boldsymbol{\beta}$, la dérivée et la matrice hessienne de la log-vraisemblance ne sont qu'un multiplicateur de α et les estimations des paramètres et les erreurs-types sont invariantes par rapport à l'échelle des poids.

Cependant, la relation suivante montre que les estimations ponctuelles obtenues par la maximisation de la log-vraisemblance pénalisée ne sont pas invariantes par rapport à l'échelle des poids :

$$
\begin{aligned}
l_p(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w}) &= l(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w}) + 0{,}5 \log\left|I(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w})\right| \\
&= \alpha l(\boldsymbol{\beta}_{\tilde{w}};\, w) + 0{,}5 \log\left|\alpha I(\boldsymbol{\beta}_{\tilde{w}};\, w)\right| - \sum_{k=1}^{K}\left(\alpha \sum_{i\in\mathcal{D}_k} w_i\right)\log \alpha \\
&= \alpha l(\boldsymbol{\beta}_{\tilde{w}};\, w) + 0{,}5\left\{\log\left|I(\boldsymbol{\beta}_{\tilde{w}};\, w)\right| + p \log \alpha\right\} - \sum_{k=1}^{K}\left(\alpha \sum_{i\in\mathcal{D}_k} w_i\right)\log \alpha \\
&= \alpha\left\{l(\boldsymbol{\beta}_{\tilde{w}};\, w) + 0{,}5 \log\left|I(\boldsymbol{\beta}_{\tilde{w}};\, w)\right|\right\} - \sum_{k=1}^{K}\left(\alpha \sum_{i\in\mathcal{D}_k} w_i\right)\log \alpha \\
&\quad + 0{,}5\left\{p \log \alpha + (1-\alpha)\log\left|I(\boldsymbol{\beta}_{\tilde{w}};\, w)\right|\right\} \\
&= \alpha l_p(\boldsymbol{\beta}_{\tilde{w}};\, w) - \sum_{k=1}^{K}\left(\alpha \sum_{i\in\mathcal{D}_k} w_i\right)\log \alpha \\
&\quad + 0{,}5\left\{p \log \alpha + (1-\alpha)\log\left|I(\boldsymbol{\beta}_{\tilde{w}};\, w)\right|\right\}.
\end{aligned}
$$

Le terme supplémentaire du deuxième membre de l'équation précédente comporte les paramètres de régression. Par conséquent, les estimations ponctuelles et les erreurs-types ne sont pas invariantes par rapport à l'échelle des poids.

Par construction, les estimations ponctuelles qui utilisent la log-vraisemblance et les poids mis à l'échelle sont invariantes par rapport à l'échelle des poids.

2.2 Exemple d'utilisation de poids mis à l'échelle

Examinons l'étude du myélome décrite à la section 1.1. Nous avons ajusté de nouveau le même modèle de régression à risques proportionnels en utilisant les variables explicatives LogBUN, HGB et Contrived, mais en prenant maintenant des poids mis à l'échelle pour construire la vraisemblance pénalisée de Firth.

Le tableau 2.1 présente les estimations ponctuelles et les erreurs-types avec une vraisemblance pénalisée de Firth utilisant des poids mis à l'échelle et l'estimateur de variance linéarisé de Taylor. Ces statistiques sont invariantes par rapport à l'échelle des poids.

Tableau 2.1
Estimations des paramètres et de leurs erreurs-types au moyen de la méthode linéarisée de Taylor avec correction de Firth et poids mis à l'échelle

	Poids $w1$		Poids $w3$		Poids $w5$	
	Estimation	**Erreur type**	**Estimation**	**Erreur type**	**Estimation**	**Erreur type**
LogBUN	1,722	0,564	1,722	0,564	1,722	0,564
HGB	-0,112	0,064	-0,112	0,064	-0,112	0,064
Contrived	3,815	0,458	3,815	0,458	3,815	0,458

Les erreurs-types utilisant des répliques jackknife sont également invariantes par rapport à l'échelle des poids. Pour les méthodes d'estimation de la variance par répliques, chaque ensemble de poids de rééchantillonnage doit être mis à l'échelle au moyen du facteur d'échelle qui a servi à mettre à l'échelle les poids de l'échantillon complet. Le tableau 2.2 présente les estimations ponctuelles et les erreurs-types de la vraisemblance pénalisée de Firth utilisant des poids mis à l'échelle et de l'estimateur de la variance par répliques jackknife.

Tableau 2.2
Estimations des paramètres et de leurs erreurs-types au moyen de répliques jackknife avec correction Firth et poids mis à l'échelle

	Poids $w1$		Poids $w3$		Poids $w5$	
	Estimation	**Erreur type**	**Estimation**	**Erreur type**	**Estimation**	**Erreur type**
LogBUN	1,722	0,653	1,722	0,653	1,722	0,653
HGB	-0,112	0,074	-0,112	0,074	-0,112	0,074
Contrived	3,815	0,642	3,815	0,642	3,815	0,642

Les estimations de la log-vraisemblance pénalisée utilisant des poids mis à l'échelle possèdent également la propriété de précision. Les rapports entre les erreurs-types jackknife et les erreurs-types linéarisées de Taylor sont de 1,16, 1,17 et 1,40 pour les trois ensembles de poids pour les variables LogBUN, HGB et Contrived, respectivement (tableaux 2.1 et 2.2).

3 Applications sur des enquêtes complexes

Souvent, les données d'enquêtes complexes contiennent des poids, des strates et des grappes inégaux. Il est recommandé d'utiliser les poids et d'autres caractéristiques du plan à l'étape de l'analyse. Les données pondérées fournissent une meilleure représentation de la population étudiée que les données non pondérées. Dans la présente section, nous comparons les poids mis à l'échelle et non mis à l'échelle pour estimer les coefficients de régression à risques proportionnels au moyen d'une étude par simulations, et nous appliquons la vraisemblance pénalisée de Firth utilisant des poids mis à l'échelle pour estimer les durées de survie à partir d'un ensemble de données de la NHEFS.

3.1 Étude par simulations

Nous avons réalisé une petite étude par simulations pour comparer les biais dans les estimations des paramètres et les erreurs-types avec et sans mise à l'échelle des poids au moyen de la probabilité pénalisée de Firth. Nous avons utilisé deux méthodes d'échantillonnage pour sélectionner des échantillons à partir d'une population finie fixe : un échantillonnage aléatoire simple (EAS) sans remise dans lequel un poids égal est attribué à chaque unité d'observation, d'une part, et un échantillon avec probabilité proportionnelle à la taille (PPT) sans remise dans lequel le poids de sondage pour une unité d'observation dépend de la valeur de la mesure de taille associée à la fonction de risque pour cette unité, d'autre part. Aux fins de l'inférence de la population finie, nous traitons les paramètres de régression à risques proportionnels estimés dans la population finie comme les « vraies » valeurs des paramètres. Les biais sont mesurés à partir de ces valeurs vraies.

Des populations finies de taille 10 000 sont générées comme suit :

- $\mathrm{Z}_1, \mathrm{Z}_2, \ldots, \mathrm{Z}_{10} \sim \text{Bernoulli}(0{,}75)$,
- $h = \exp(-0{,}69Z_1 - 0{,}69Z_2 - \cdots - 0{,}69Z_{10})$,
- $u \sim \text{uniform}(0, 1)$,
- $t = \log(u)/h$,
- $c \sim \text{Bernoulli}(v)$,
- $m \sim \text{uniform}(10h, 10h + 0{,}1)$

où h est la fonction de risque, t est la durée de survie, c est un indicateur de censure et m est une mesure de taille pour chaque unité. Six populations finies sont générées au moyen de différentes valeurs de censure ($v = 0{,}1;\ 0{,}3;\ 0{,}5;\ 0{,}7;\ 0{,}8;\ 0{,}9$). Voir Bender, Augustin et Blettner (2005) à propos des méthodes de génération de durées de survie. On génère dix variables explicatives $(\mathrm{Z}_1, \mathrm{Z}_2, \ldots, \mathrm{Z}_{10})$ au moyen de distributions de Bernoulli pour créer des vraisemblances monotones, surtout quand la taille de l'échantillon est petite et que le taux de censure est élevé.

On sélectionne les échantillons dans chaque population finie au moyen de deux méthodes d'échantillonnage : un échantillonnage aléatoire simple sans remise et des échantillons avec probabilité proportionnelle à la taille sans remise, où la variable m est utilisée comme mesure de la taille. Quatre tailles d'échantillon sont utilisées pour chaque méthode d'échantillonnage : 50, 100, 500 et 1 000. Les poids de sondage de toutes les unités pour l'EAS dépendent uniquement de la taille de l'échantillon, alors que le poids de sondage d'une unité pour l'échantillonnage avec PPT dépend à la fois de la taille de l'échantillon et de la valeur observée de la variable m pour cette unité correspondante. Pour que la distribution des observations censurées soit identique dans les données échantillonnées et dans la population, on sélectionne les échantillons indépendamment des unités censurées et non censurées de la population.

Enfin, les paramètres de régression du modèle de régression à risques proportionnels

$$\lambda(t, \mathbf{Z}) = \lambda_0(\mathbf{t}) \exp(\beta_1 \mathbf{Z_1} + \beta_2 \mathbf{Z_2} + \cdots + \beta_{10} \mathbf{Z_{10}})$$

sont estimés à partir de chaque ensemble de données échantillonnées, où $\lambda_0(t)$ est le risque de référence, t est la durée de survie et c est l'indicateur censuré. Les paramètres de régression sont estimés par maximalisation de la vraisemblance pénalisée de Firth pondérée. Notons que la vraisemblance non pénalisée ne converge pas dans la plupart des cas en raison de la monotonie de la vraisemblance dans les données simulées. Quand la vraisemblance n'est pas monotone, nous avons constaté que les estimations ponctuelles obtenues au moyen de la vraisemblance pénalisée sont très proches des estimations ponctuelles obtenues au moyen de la vraisemblance non pénalisée. Heinze et Schemper (2001) ont fait état de résultats semblables en cas de données non pondérées.

Nous comparons les biais relatifs des estimations ponctuelles et des erreurs-types à l'aide de la méthode du jackknife pour les poids mis à l'échelle et non mis à l'échelle. Les biais relatifs (BR) sont définis ci-dessous (Sitter, 1992).

Soit $\hat{\beta}_s$ l'estimation ponctuelle et $\hat{v}_s$ l'estimation de la variance pour une composante de $\boldsymbol{\beta}$ tirée de l'ensemble de données s. Définissons ce qui suit :

Biais relatif pour les estimations ponctuelles, $\hat{\beta}$,

$$\mathrm{BR}\left(\hat{\beta}\right) = S^{-1}\sum_{s=1}^{S}\frac{\left|\left(\hat{\beta}_s - \beta_T\right)\right|}{\left|\beta_T\right|}.$$

Biais relatif pour les estimations de la variance, $\hat{v}$

$$\mathrm{BR}\left(\hat{v}\right) = S^{-1}\sum_{s=1}^{S}\frac{\left|\left(\hat{v}_s - \mathrm{EQM}_T\right)\right|}{\mathrm{EQM}_T}$$

où la vraie EQM est

$$\mathrm{EQM}_T\left(\hat{\beta}\right) = S^{-1}\sum_{s}\left(\hat{\beta}_s - \beta_T\right)^2$$

et β_T est la valeur « vraie » du paramètre obtenue par l'ajustement du modèle de régression à risques proportionnels utilisant toutes les unités de la population finie. Le rapport des BR est défini comme étant le rapport entre le BR utilisant des poids non mis à l'échelle et le BR utilisant des poids mis à l'échelle.

La médiane des rapports des BR sur plus de 5 000 répétitions est présentée dans la section. Nous indiquons la médiane en raison de certains « mauvais » échantillons dans lesquels les convergences sont douteuses, y compris en cas de correction de Firth. Ces « mauvais » échantillons produisent un petit nombre d'estimations comportant de très grands biais. En raison de ces biais importants, la moyenne du rapport des BR est une statistique plus instable que la médiane. Sans « mauvaises » répliques, la moyenne et les médianes sont très proches. Nous avons aussi constaté que la log-vraisemblance pénalisée qui utilise des poids non mis à l'échelle produit davantage de ces « fausses » convergences.

Les résultats de toutes les variables explicatives $\mathrm{Z}_1, \mathrm{Z}_2, \ldots, \mathrm{Z}_{10}$ sont semblables. Pour simplifier la lecture, nous présentons les résultats de seulement deux variables explicatives, Z_3 et Z_8.

Les rapports des BR dans les estimations des paramètres avec poids non mis à l'échelle et mis à l'échelle pour les variables Z_3 et Z_8 sont présentés dans les figures 3.1, 3.2, 3.3 et 3.4. En cas

d'échantillons de petite taille et de grand nombre d'observations censurées, les BR utilisant des poids mis à l'échelle sont nettement plus petits que les BR utilisant des poids non mis à l'échelle. En cas de grandes tailles d'échantillon, les BR des deux sortes de poids sont semblables principalement parce que l'option Firth n'est pas nécessaire, puisque la convergence ne pose pas problème en cas de grands ensembles de données.

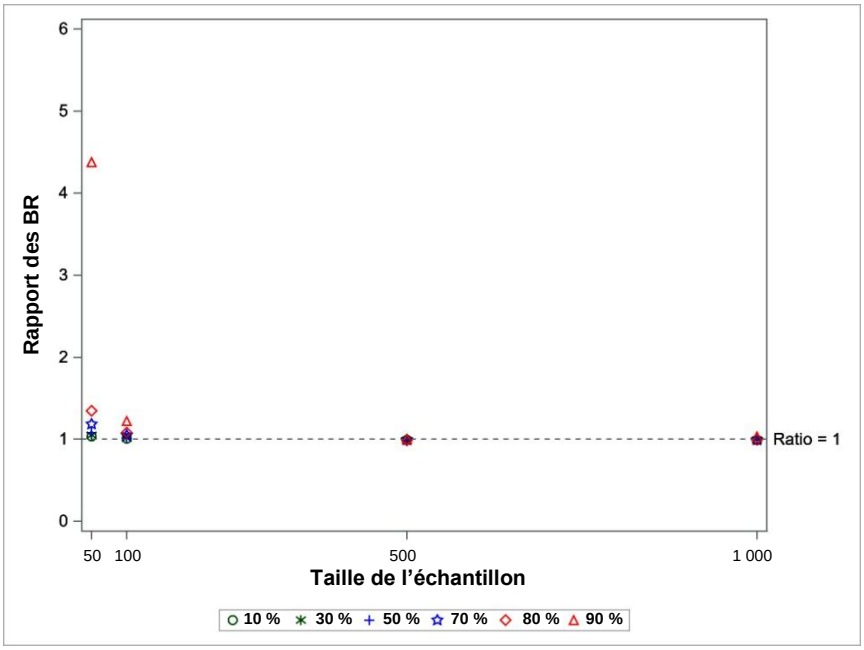

Figure 3.1 Rapport des biais relatifs dans les estimations des paramètres avec échantillons EAS pour Z3.

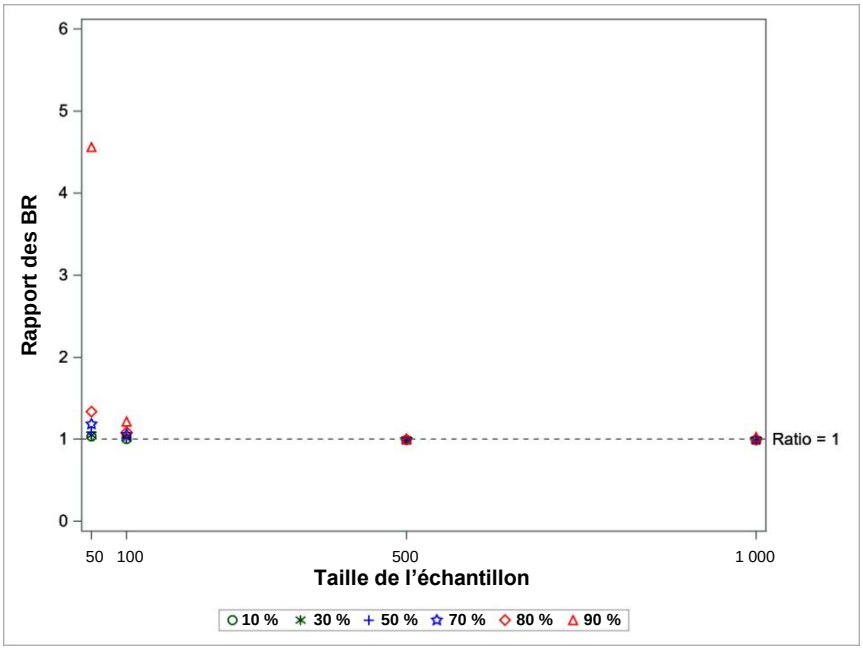

Figure 3.2 Rapport des biais relatifs dans les estimations des paramètres avec échantillons EAS pour Z8.

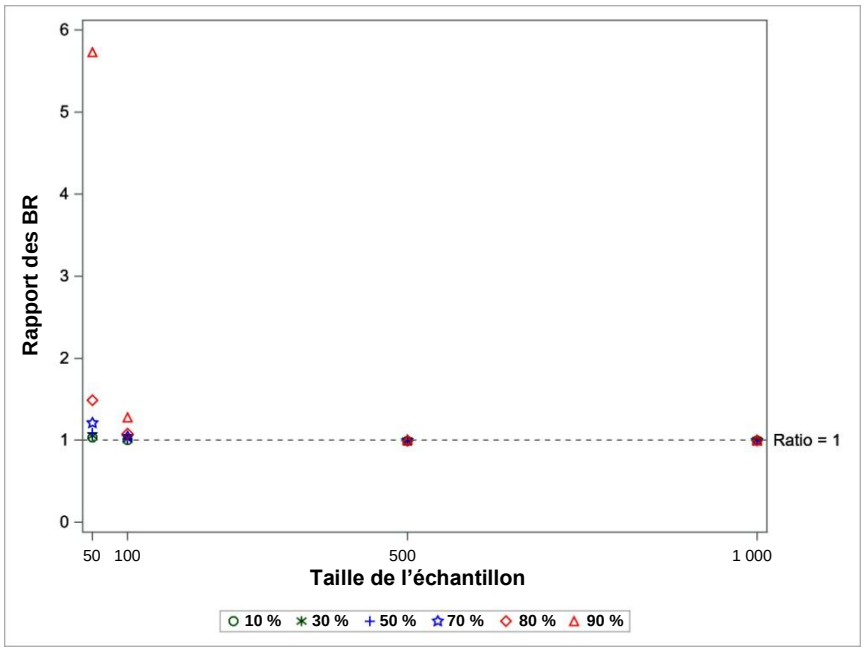

Figure 3.3 Rapport des biais relatifs dans les estimations des paramètres en cas d'échantillons avec PPT pour Z3.

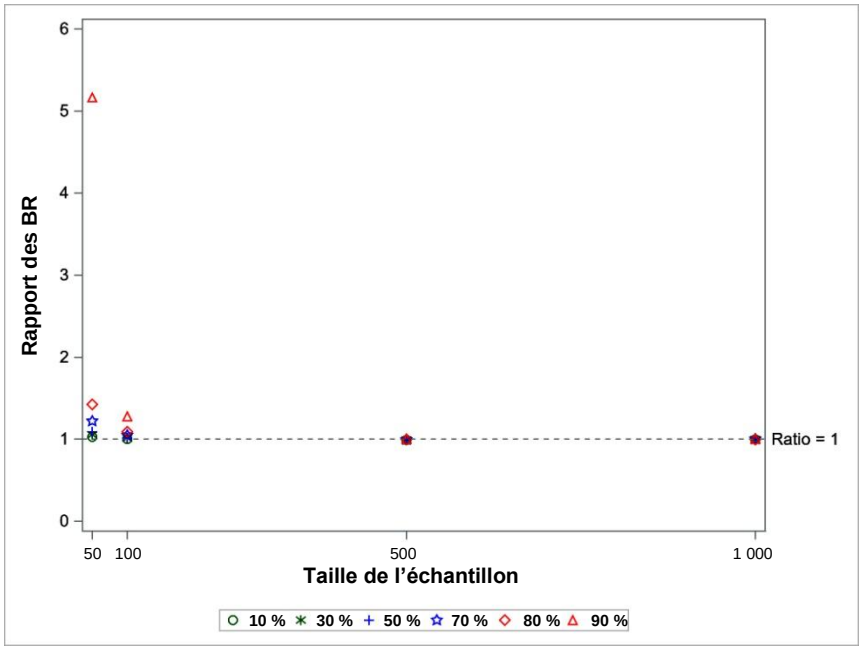

Figure 3.4 Rapport des biais relatifs dans les estimations des paramètres en cas d'échantillons avec PPT pour Z8.

Les rapports des biais relatifs dans les erreurs-types avec poids non mis à l'échelle et mis à l'échelle pour les variables Z_3 et Z_8 sont présentés dans les figures 3.5, 3.6, 3.7 et 3.8. Les BR des erreurs-types suivent la même tendance que les BR des estimations ponctuelles. Les premiers sont toutefois plus élevés

que les BR des estimations ponctuelles. En cas d'échantillons de petite taille et de grand nombre d'observations censurées, les BR utilisant des poids mis à l'échelle sont nettement plus petits que les BR utilisant des poids non mis à l'échelle. En cas de grandes tailles d'échantillon, les BR avec poids mis à l'échelle et non mis à l'échelle sont semblables principalement parce que l'option Firth n'est pas nécessaire, puisque la convergence ne pose pas problème en cas de grands ensembles de données.

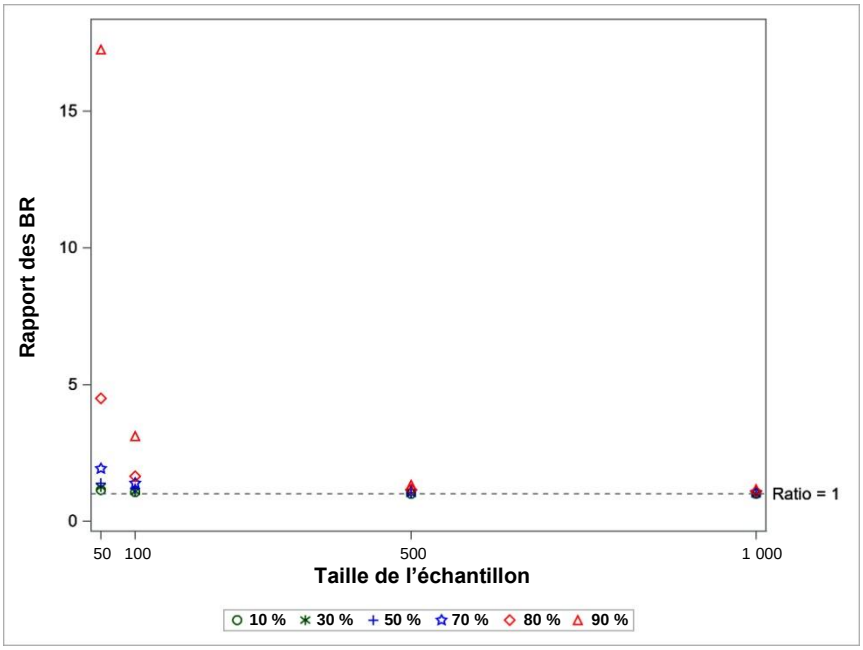

Figure 3.5 Rapport des biais relatifs dans les erreurs-types avec échantillons EAS pour Z3.

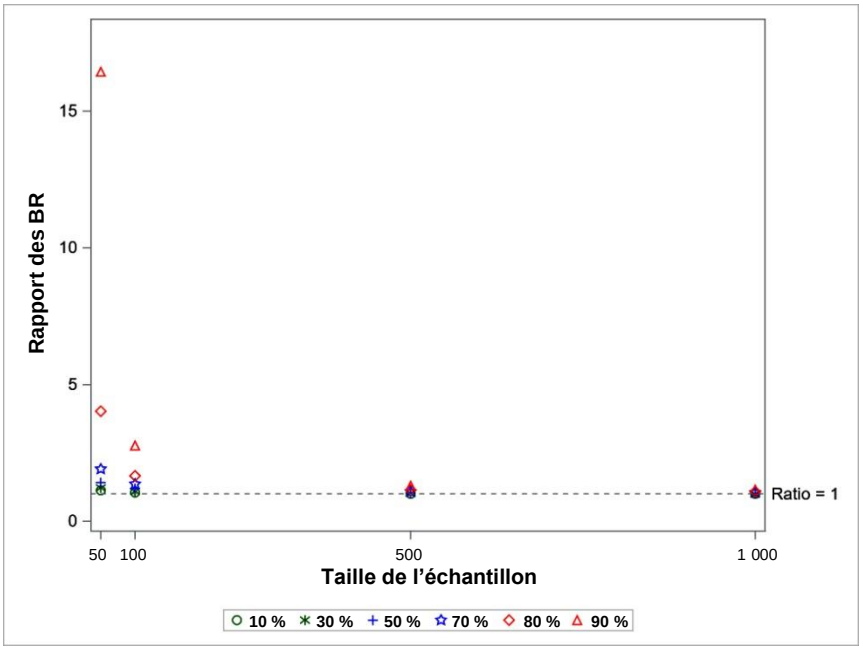

Figure 3.6 Rapport des biais relatifs dans les erreurs-types avec échantillons EAS pour Z8.

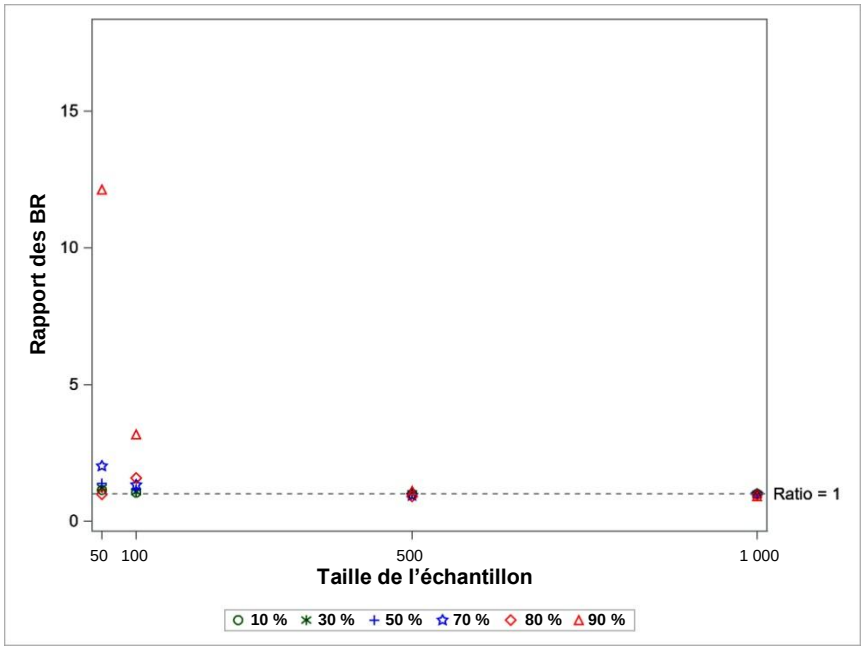

Figure 3.7 Rapport des biais relatifs dans les erreurs types en cas d'échantillons avec PPT pour Z3.

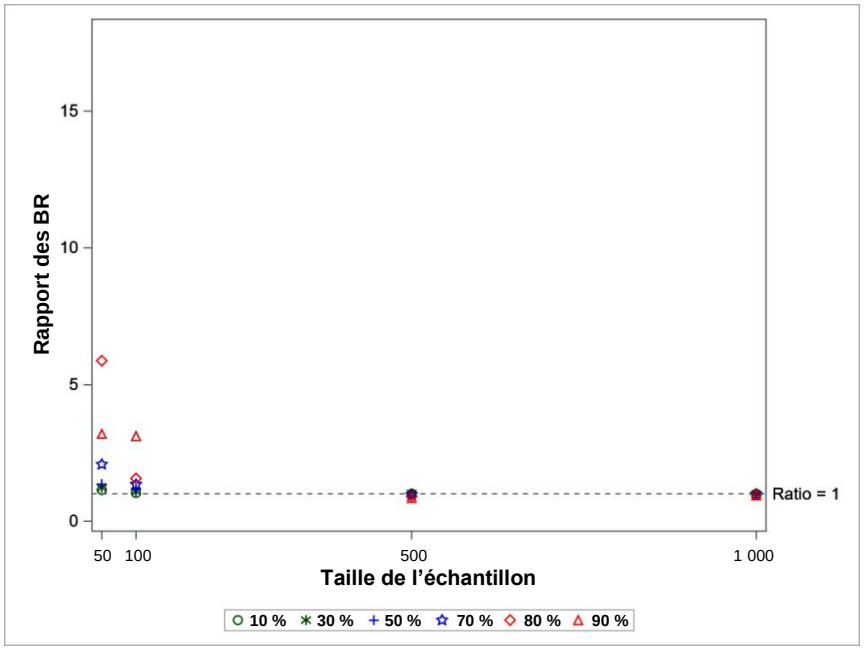

Figure 3.8 Rapport des biais relatifs dans les erreurs types en cas d'échantillons avec PPT pour Z8.

Le tableau 3.1 présente le premier quartile, la médiane et le troisième quartile pour le rapport des BR dans les estimations ponctuelles et les erreurs-types en cas de taille d'échantillon de 50. Le tableau donne les résultats pour la variable Z_3 en cas de censure de 10 % et de 90 %. Nous avons observé que pour

toutes les variables, les premier et troisième quartiles pour le rapport des BR ne contiennent pas 1 quand la taille de l'échantillon est petite et que le pourcentage de censure est élevé. Toutefois, comme prévu, en cas de grands échantillons et de petit nombre d'observations censurées, la différence entre les BR avec poids mis à l'échelle et non mis à l'échelle est faible.

Tableau 3.1
Rapport des BR dans les estimations ponctuelles et les erreurs-types pour la taille d'échantillon de 50 (variable Z_3)

	Rapport des BR dans les estimations ponctuelles					
Conception	**Censure de 90 %**			**Censure de 10 %**		
	Premier quartile	**Médiane**	**Troisième quartile**	**Premier quartile**	**Médiane**	**Troisième quartile**
EAS	1,81	4,38	7,37	1,00	1,03	1,06
PPT	3,36	5,73	11,54	0,99	1,03	1,08
	Rapport des BR dans les erreurs-types					
Conception	**Censure de 90 %**			**Censure de 10 %**		
	Premier quartile	**Médiane**	**Troisième quartile**	**Premier quartile**	**Médiane**	**Troisième quartile**
EAS	9,03	17,26	40,87	1,03	1,15	1,33
PPT	5,57	12,13	29,92	1,00	1,15	1,33

3.2 Application sur les données de la NHEFS

Nous avons étudié la durée avant la survenue d'une crise cardiaque et son lien avec le cholestérol sanguin et le tabagisme au moyen d'un ensemble de données de la NHEFS.

La NHEFS est une enquête longitudinale nationale aux États-Unis qui sert à délimiter les relations entre les facteurs cliniques, nutritionnels et comportementaux, à déterminer les utilisations de l'hôpital, et à surveiller l'évolution des facteurs de risque pour une cohorte initiale qui représente la population NHANES I (National Health and Nutrition Examination Survey I). Une cohorte de 14 407 personnes a été sélectionnée pour la NHEFS. Les données sur le statut vital et le traçage, les données d'entrevue, les données sur les séjours dans les établissements de soins de santé et les données sur la mortalité de 1987 sont accessibles au public. Pour en savoir plus sur l'enquête et les ensembles de données utilisés dans la présente section, consultez le site Web des *Centers for Disease Control and Prevention* (https://www.cdc.gov/).

Nous avons utilisé 4 673 observations tirées des données des entrevues publiques de 1987 de la NHEFS pour étudier l'occurrence de la première crise cardiaque dans la population de l'enquête en 1987 et son lien avec le cholestérol sanguin et le tabagisme. Les variables suivantes ont été employées :

- *Stratum*, identification de la strate;
- *ObservationWeight*, poids d'échantillonnage associé à chaque unité d'observation;
- *PSU*, identification de l'unité primaire d'échantillonnage;
- *Age*, variable temps-événement, définie comme suit :
    - l'âge du sujet quand la première crise cardiaque a été déclarée pour les sujets ayant déclaré une crise cardiaque;

  - l'âge du sujet indiqué lors de l'entrevue pour les sujets qui n'ont jamais déclaré de crise cardiaque;
- *HeartAttack*, indicateur de crise cardiaque (1 = crise cardiaque déclarée);
- *Income*, revenu du ménage normalisé à une moyenne nulle;
- *HighBloodChol*, indique si un sujet a un taux de cholestérol sanguin élevé ou bas;
- *Smoker*, habitude de fumer du sujet (1 = fumeur actuel, 2 = ancien fumeur, –1 = non-fumeur);
- *Race*, race du sujet (1 = noire, 2 = blanche, 3 = autre);
- *Gender*, genre du sujet.

On utilise la procédure SURVEYPHREG dans SAS/STAT (Mukhopadhyay, 2010) pour ajuster un modèle de régression à risques proportionnels pour l'âge selon le revenu, le cholestérol sanguin, l'habitude de fumer, la race, le genre et l'interaction entre race et genre. La crise cardiaque sert d'indicateur censuré. Les poids d'observation varient de 1 164 à 121 040 avec une moyenne de 16 036,51, une médiane de 12 321 et un coefficient de variation de 74,35. Les sujets sont divisés en 644 grappes et 35 strates.

Dans la présente section, on utilise PROC SURVEYPHREG au lieu de PROC PHREG parce que la NHEFS utilise un plan de sondage complexe comportant une stratification, une corrélation intra-grappe et des poids inégaux. PROC SURVEYPHREG prend en charge les instructions STRATA, CLUSTER et WEIGHT pour tenir compte respectivement de la stratification, de la corrélation intra-grappe et des poids inégaux. De plus, PROC SURVEYPHREG prend en charge la linéarisation en séries de Taylor et les estimations de la variance par la méthode du jackknife pour les données d'enquête (Mukhopadhyay, 2010). Aux fins de l'étude, nous avons choisi l'estimation de la variance par la méthode du jackknife. Les instructions SAS ajustées au modèle sont présentées à l'annexe 2.

Les 4 673 sujets de l'échantillon représentent près de 74,9 millions de personnes dans la population étudiée de 1987. Parmi les sujets, 213 ont déclaré au moins une crise cardiaque, et les 4 460 autres sujets sont considérés comme censurés. Les 213 observations d'événements dans l'échantillon représentent environ 3,2 millions d'unités de population, et les 4 460 observations censurées dans l'échantillon représentent environ 71,7 millions d'unités de population. 95,44 % des observations dans l'échantillon n'ont pas déclaré de crise cardiaque, ce qui correspond à 95,68 % d'individus dans la population (tableau 3.2) qui n'ont pas déclaré de crise cardiaque.

Tableau 3.2
Nombre d'observations censurées et non censurées et leur somme de poids

	Total	Événement	Censuré	Pourcentage de censure
Nombre d'observations	4 673	213	4 460	95,44
Somme des poids	74 938 614	3 239 653	71 698 961	95,68

Sans la pénalité de Firth, l'optimisation de Newton-Raphson converge en satisfaisant le critère de convergence du gradient relative (GCONV = 1E–8), mais les coefficients des variables Smoker et Race ne convergent pas. Les coefficients pour Smoker = 2 sont 7,47, 10,87 et 11,83 et ceux pour Race = 1 sont 7,55, 10,95 et 11,17 dans les trois dernières itérations, respectivement. Ce phénomène est très courant en cas de vraisemblance monotone (voir le tableau 1.1). Sur 644 échantillons répétés (= 644 UPE), on observe une vraisemblance monotone dans 542 répliques. La vraisemblance pénalisée de Firth est une bonne solution de rechange en cas de vraisemblances monotones.

Nous utilisons l'option FIRTH dans PROC SURVEYPHREG (voir « The SURVEYPHREG Procedure » dans SAS Institute Inc. (2018)) pour maximiser la vraisemblance pénalisée de Firth. L'option FIRTH dans PROC SURVEYPHREG utilise des poids mis à l'échelle. L'optimisation de la vraisemblance pénalisée converge avec GCONV = 1E–8, ainsi qu'avec une convergence raisonnable dans tous les coefficients. La convergence est également atteinte dans les 644 échantillons répétés avec la pénalité de Firth.

Le tableau 3.3 présente les rapports des risques estimés ainsi que leurs intervalles de confiance de Wald de 95 % pour les taux de cholestérol sanguin et le tabagisme. Dans la population étudiée de 1987, le risque estimé de crise cardiaque chez un sujet ayant un taux de cholestérol sanguin bas est de 0,6 fois le risque estimé de crise cardiaque chez un sujet ayant un taux élevé de cholestérol sanguin. Étant donné que l'intervalle de confiance de 95 % ne contient pas 1, on peut raisonnablement conclure que le risque de crise cardiaque pour un sujet ayant un taux de cholestérol sanguin bas est significativement moins élevé que le risque de crise cardiaque pour un sujet au taux de cholestérol sanguin élevé après ajustement du tabagisme, de la race et d'autres variables explicatives dans la population étudiée de 1987.

Les rapports des risques estimés pour les non-fumeurs, les fumeurs actuels et les anciens fumeurs sont respectivement de 0,59, 0,64 et 1,1. Le risque estimé de crise cardiaque pour les non-fumeurs est inférieur au risque estimé pour les fumeurs actuels ou les anciens fumeurs. Nous ne possédons toutefois pas de données probantes suffisantes pour conclure que les rapports des risques associés au tabagisme sont considérablement différents à 95 % après correction pour tenir compte du cholestérol sanguin, de la race et d'autres agents régulateurs dans la population étudiée en 1987.

Tableau 3.3
Rapport des risques pour le cholestérol sanguin et le tabagisme ainsi que leurs intervalles de confiance de Wald de 95 %

		Intervalle de confiance	
	Estimation ponctuelle	Inférieur	Supérieur
HighBloodChol 0 vs 1	0,643	0,469	0,882
Smoker -1 vs 1	0,590	0,259	1,345
Smoker -1 vs 2	0,641	0,361	1,140
Smoker 1 vs 2	1,087	0,359	3,290

4 Résumé

La vraisemblance pénalisée de Firth est utile pour l'obtention d'estimations par le maximum de vraisemblance à partir d'une vraisemblance monotone dans des modèles de régression à risques proportionnels. Nous avons proposé une méthode de mise à l'échelle des poids et montré que la vraisemblance pénalisée de Firth utilisant des poids mis à l'échelle présente certaines propriétés souhaitables quand on étudie des enquêtes complexes. Au moyen d'une étude par simulations, on a montré que les biais estimés dans les estimations ponctuelles et les erreurs-types utilisant des poids mis à l'échelle sont inférieurs aux biais estimés quand les poids ne sont pas mis à l'échelle. Bien que la vraisemblance pénalisée de Firth produise de « bonnes » estimations sur la plupart des ensembles de données simulés, elle n'a pas produit de « bonnes » convergences pour certains ensembles de données. La vraisemblance pénalisée de Firth utilisant des poids mis à l'échelle est parvenue à corriger pour une vraisemblance monotone quand nous avons estimé les taux de risque de crise cardiaque à partir d'un ensemble de données de la NHEFS. Bien que les résultats numériques soient très encourageants, il faut approfondir les recherches pour calculer les distributions asymptotiques des estimateurs obtenus au moyen de la vraisemblance pénalisée de Firth.

Nous recommandons d'utiliser une vraisemblance non pénalisée quand la convergence ne pose pas problème. En revanche, la vraisemblance pénalisée de Firth utilisant des poids mis à l'échelle est préférable en cas de vraisemblance monotone dans l'ajustement des modèles de régression à risques proportionnels pour les enquêtes complexes.

Remerciements

Je remercie Ying So, Randy Tobias et Ed Huddleston du SAS Institute Inc. de l'aide précieuse qu'ils m'ont apportée dans la préparation de l'article. J'aimerais également remercier les deux examinateurs anonymes et le rédacteur adjoint pour leurs suggestions constructives.

Annexe 1

Convergence de l'estimateur par la vraisemblance pénalisée de Firth

Les estimateurs de la section 2 sont définis comme la solution à un système d'équations construites au moyen des fonctions de score des modèles de régression à risques proportionnels. Dans la présente annexe, nous montrons que, dans certaines conditions de régularité, ces estimateurs sont convergents par rapport au plan de sondage. Les propriétés des estimateurs qui sont des solutions à un ensemble d'équations d'estimation sont largement étudiées dans la littérature sur les enquêtes. Voir par exemple Binder (1983), Godambe et Thompson (1986) et Fuller (2009, section 1.3.4).

Toutefois, les équations d'estimation pour les modèles de régression à risques proportionnels sont plus complexes que les équations d'estimation pour les modèles linéaires généralisés, car les fonctions de score impliquent des sommes pondérées sur les unités échantillonnées. Binder (1992) et Lin (2000) ont montré que les estimateurs obtenus par la résolution d'équations d'estimation pour les modèles de régression à risques proportionnels sont convergents. Dans la présente annexe, nous adoptons des arguments semblables à ceux de Lin (2000) et d'Andersen et Gill (1982).

Il faut plusieurs hypothèses techniques pour démontrer que les estimations ponctuelles sont cohérentes. Nous avons besoin d'hypothèses sur les équations d'estimation, la population finie et le plan de sondage selon lesquelles :

- les fonctions définissant les équations d'estimation doivent être lisses et convexes;
- la population finie doit être telle que les moments des quantités de population qu'on utilise pour définir les équations d'estimation existent;
- le plan de sondage doit être tel que les estimateurs de Narain-Horvitz-Thompson (NHT) (Rao, 2005) pour les totaux de population se comportent raisonnablement.

Toutes ces hypothèses sont courantes dans la littérature sur les enquêtes-échantillons comme dans Fuller (2009). Les fonctions de score pour les modèles de régression à risques proportionnels comportent des rapports de moyennes de fonctions exponentielles qui sont différenciables à l'infini.

Soit $\mathcal{U}_N$ et $\mathcal{F}_N$ qui désignent respectivement l'ensemble d'indices et les valeurs pour la N^{e} population finie dans une séquence de populations indexées par N, et soit A_N un échantillon de taille n tiré de $\mathcal{U}_N$. Afin d'étudier les propriétés des grands échantillons pour les estimateurs basés sur un échantillon, nous supposons des séquences de population et des échantillons tels que $N \to \infty$ et $(N - n) \to \infty$, en gardant la fraction de sondage, $\frac{n}{N}$, fixe.

Supposons que $\mathcal{F}_N = \{(t_i, \Delta_i, \mathrm{Z}_i(\cdot))\}_{i=1}^{N}$ est un échantillon aléatoire simple de taille N de la distribution conjointe de $(T, \Delta, \mathrm{Z}(\cdot))$, où t est le temps de défaillance ou le temps de censure, s'il est inférieur; $\Delta = 1$ si le temps de défaillance est inférieur au temps de censure et 0 sinon; et $Z(\cdot)$ est un vecteur de variables explicatives susceptibles de varier avec le temps.

Soit $\boldsymbol{\beta}$ un ensemble de paramètres de régression pour la superpopulation qui est définie par la distribution conjointe de $(T, \Delta, Z(\cdot))$. Soit $\boldsymbol{\beta}_N$ un ensemble de paramètres de population finie obtenus par la résolution des équations d'estimation quand toutes les unités N de la population sont observées, et soit $\hat{\boldsymbol{\beta}}_N$ un estimateur de $\boldsymbol{\beta}_N$ qui est obtenu par la résolution des équations d'estimation pondérées seulement au moyen des unités échantillonnées. Notre objectif est de montrer que $\hat{\boldsymbol{\beta}}_N$ se rapproche de $\boldsymbol{\beta}_N$ et que les deux se rapprochent de $\boldsymbol{\beta}$ à mesure que la taille de l'échantillon et la taille de la population augmentent.

Examinons les équations d'estimation qui correspondent à la vraisemblance pénalisée de Firth décrite à la section 2. Par souci de simplicité, nous écrivons ces équations pour un cas sans événements liés. Afin de

simplifier encore la notation, nous écrivons séparément chaque composante des équations d'estimation. Les paramètres de population finie, $\boldsymbol{\beta}_N$, sont une solution à la fonction de score de la vraisemblance partielle pénalisée, $U_N(\boldsymbol{\beta}) = \left(U_{N,1}(\boldsymbol{\beta}), \ldots, U_{N,P}(\boldsymbol{\beta})\right)'$, où

$$
\begin{aligned}
U_{N,p}(\boldsymbol{\beta}) = N^{-1} \sum_{i \in \mathcal{U}_N} \Delta_i \Bigg[& \mathbf{Z}_i(t_i) - \frac{S_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \\
& + 0{,}5 \operatorname{tr}\Bigg(\Bigg[N^{-1} \sum_{i \in \mathcal{U}_N} \Delta_i \left\{ \frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \right\} \Bigg]^{-1} \\
& \left\{ \left(\frac{Q_p^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \right. \\
& - \left(\frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \\
& \left. - \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left(\frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \right\} \Bigg) \Bigg]
\end{aligned}
$$

où

$$
S^{(a)}(\boldsymbol{\beta}, t) = N^{-1} \sum_{i \in \mathcal{U}_N} I(t_i \geq t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t)) [\mathbf{Z}_i(t)]^{\otimes a}
$$

$$
Q_p^{(a)}(\boldsymbol{\beta}, t) = N^{-1} \sum_{i \in \mathcal{U}_N} I(t_i \geq t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t)) Z_{i,p}(t) [\mathbf{Z}_i(t)]^{\otimes a}
$$

et où $a = 0, 1, 2$; $Z_i(t) = \left(Z_{i,1}(t), \ldots, Z_{i,p}(t)\right)'$; $\mathbf{S}^{(1)}(\boldsymbol{\beta}, t) = (S_1^{(1)}(\boldsymbol{\beta}, t), \ldots, S_P^{(1)}(\boldsymbol{\beta}, t))'$; $\operatorname{tr}(\cdot)$ désigne la trace d'une matrice; $I(\cdot)$ désigne la fonction indicatrice; $p = 1, 2, \ldots, P$; et P est le nombre de paramètres de régression. Notons que $S^{(a)}(\boldsymbol{\beta}, t)$ et $Q_p^{(a)}(\boldsymbol{\beta}, t)$ dépendent de N, bien que la notation ne l'indique pas par souci de simplicité.

En définissant la fonction de score pour la vraisemblance pénalisée, nous supposons que la matrice d'information pour la population finie, $I_N(\boldsymbol{\beta}, t)$, est toujours définie positive.

Il faut toutefois rappeler que dans une situation réaliste, toutes les unités de la population finie ne sont pas disponibles. Soit un échantillon A_N sélectionné au moyen d'un plan de sondage probabiliste qui attribue une probabilité de sélection non nulle, π_i, à chaque unité de la population. Soit $w_i = \pi_i^{-1}$ le poids de sondage. On obtient un estimateur basé sur un échantillon, $\hat{\boldsymbol{\beta}}_N$, en résolvant les équations de score estimées par la vraisemblance partielle pénalisée. En supposant que N est connu, on obtient comme estimateur basé sur un échantillon pour $U_{N,p}(\boldsymbol{\beta})$

$$\hat{U}_{N,p}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in A_N} w_i\Delta_i\Bigg[\mathbf{Z}_i(t_i) - \frac{\hat{S}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}$$
$$+ \ 0{,}5\mathrm{tr}\Bigg(\Bigg[N^{-1}\sum_{i\in A_N} w_i\Delta_i\left\{\frac{\hat{S}^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\right\}\Bigg]^{-1}$$
$$\left\{\left(\frac{\hat{Q}_p^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{S}^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\right.$$
$$-\left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)'$$
$$\left.-\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\right\}\Bigg)\Bigg]$$

où

$$\hat{S}^{(a)}(\boldsymbol{\beta}, t) = N^{-1}\sum_{i\in A_N} w_i I(t_i \geq t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))[\mathbf{Z}_i(t)]^{\otimes a}$$
$$\hat{Q}_p^{(a)}(\boldsymbol{\beta}, t) = N^{-1}\sum_{i\in A_N} w_i I(t_i \geq t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))Z_{i,p}(t)[\mathbf{Z}_i(t)]^{\otimes a}$$

sont les estimateurs de NHT pour $S^{(a)}(\boldsymbol{\beta}, t)$ et $Q_p^{(a)}(\boldsymbol{\beta}, t)$, respectivement.

Parce que $\hat{S}^{(a)}(\boldsymbol{\beta}, t)$ et $\hat{Q}_p^{(a)}(\boldsymbol{\beta}, t)$ utilisent des sommes pondérées sur des unités échantillonnées, nous avons besoin de techniques définies dans Lin (2000) pour étudier les propriétés des grands échantillons de ces estimateurs. Définissons $G_i(t) = \Delta_i I(t_i \leq t)$, $G(t) = N^{-1}\sum_{i\in\mathcal{U}_N} G_i(t)$ et $\hat{G}(t) = \sum_{i\in A_N} w_i G_i(t)$. Nous pouvons alors écrire les fonctions de score de population finie au moyen de l'intégration stochastique,

$$U_{N,p}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in\mathcal{U}_N}\int_0^\infty\Bigg[\mathbf{Z}_i(t_i) - \frac{S_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}$$
$$+ \ 0{,}5\mathrm{tr}\Bigg(\Bigg[N^{-1}\sum_{i\in A_N}\int_0^\infty\left\{\frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\right\}dG(t_i)\Bigg]^{-1}$$
$$\left\{\left(\frac{Q_p^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)\right.$$
$$-\left(\frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)'$$
$$\left.-\left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\right\}\Bigg)\Bigg]dG_i(t)$$

et les fonctions de score basées sur l'échantillon sont

$$\hat{U}_{N,p}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in\mathcal{U}_N}\int_0^\infty I(i\in A_n)\,w_i\Bigg[\mathbf{Z}_i(t_i) - \frac{\hat{S}_p^{(1)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)}$$
$$+ 0{,}5\mathrm{tr}\Bigg(\Bigg[N^{-1}\sum_{i\in\mathcal{U}_N}\int_0^\infty I(i\in A_N)\,w_i$$
$$\left\{\frac{\hat{S}^{(2)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)} - \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)}\right)\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)}\right)'\right\} dG_i(t)\Bigg]^{-1}$$
$$\Bigg\{\left(\frac{\hat{Q}_p^{(2)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)}\frac{\hat{S}^{(2)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)}\right)$$
$$-\left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)}\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)}\right)\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)}\right)'$$
$$-\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)}\right)\left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)}\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta},t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta},t_i)}\right)'\Bigg\}\Bigg)\Bigg] dG_i(t).$$

Notons que les quantités $S^{(a)}$ et $Q_p^{(a)}$ sont simplement des moyennes sur des quantités de population finie. Définissons les limites de ces moyennes comme suit :

$$\begin{aligned}
\mathbf{s}^{(a)}(\boldsymbol{\beta},t) &:= \lim_{N\to\infty}\mathbf{S}^{(a)}(\boldsymbol{\beta},t)\\
&= \lim_{N\to\infty}N^{-1}\sum_{i\in\mathcal{U}_N}I(t_i\geq t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))[\mathbf{Z}_i(t)]^{\otimes a}\\
q_p^{(a)}(\boldsymbol{\beta},t) &:= Q_p^{(a)}(\boldsymbol{\beta},t)\\
&= N^{-1}\sum_{i\in\mathcal{U}_N}I(t_i\geq t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))Z_{i,p}(t)[\mathbf{Z}_i(t)]^{\otimes a}\\
g(t) &:= \lim_{N\to\infty}N^{-1}\sum_{i\in\mathcal{U}_N}G_i(t)\\
\alpha &:= \lim_{N\to\infty}N^{-1}\sum_{i\in\mathcal{U}_N}\int_0^\infty\mathbf{Z}_i(t)\,dG_i(t).
\end{aligned}$$

Ainsi, la fonction de score de la population finie, $U_{N,p}(\boldsymbol{\beta})$, converge vers la fonction de score de la superpopulation $u_{N,p}(\boldsymbol{\beta})$, où

$$u_{N,p}(\boldsymbol{\beta}) = \alpha - \int_0^\infty\frac{s_p^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}dg(t)$$
$$+ 0{,}5\mathrm{tr}\Bigg(\int_0^\infty\Bigg[\int_0^\infty\left\{\frac{s^{(2)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)} - \left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}\right)\left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}\right)'\right\}dg(t)\Bigg]^{-1}$$
$$\left\{\left(\frac{q_p^{(2)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)} - \frac{q_p^{(0)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}\frac{s^{(2)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}\right)\right.$$
$$-\left(\frac{\mathbf{q}_p^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)} - \frac{q_p^{(0)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}\right)\left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}\right)'$$
$$\left.-\left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}\right)\left(\frac{\mathbf{q}_p^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)} - \frac{q_p^{(0)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}\right)'\right\}dg(t)\Bigg).$$

Supposons maintenant que les quantités de population, $\mathbf{Z}_i$, qui servent à définir les fonctions des score ont des moments finis et que la séquence des plans de sondage est telle que toutes les fonctions lisses des estimateurs de NHT convergent. Parce que $U_N(\boldsymbol{\beta})$ est une fonction lisse des totaux de population, et que chaque total est estimé au moyen d'un estimateur de NHT, $\hat{U}_N(\boldsymbol{\beta})$, est convergent par rapport au plan de sondage pour $U_N(\boldsymbol{\beta})$. Par conséquent, $\left(U_N(\boldsymbol{\beta}) - \hat{U}_N(\boldsymbol{\beta})\right) \big| \mathcal{F}_N = o(1)$. Ainsi, en utilisant des arguments semblables à ceux de Lin (2000) et d'Andersen et de Gill (1982), on peut montrer que $\boldsymbol{\beta}_N$ et $\hat{\boldsymbol{\beta}}_N$ convergent vers la même limite.

Parce que n/N est fixe, $\sum_{A_N} w_i$ est l'estimateur de NHT pour N, et que $\hat{U}(\boldsymbol{\beta})$ est un estimateur convergent (pas nécessairement sans biais) de 0, $\hat{U}(\boldsymbol{\beta})$ et $(n/N)\hat{U}(\boldsymbol{\beta})$ convergent vers la même limite avec un ordre de convergence identique. On peut facilement montrer que $(n/N)\hat{U}(\boldsymbol{\beta})$ et $\left(n \Big/ \sum_{A_N} w_i\right)\hat{U}(\boldsymbol{\beta})$, les équations d'estimation qui utilisent les poids mis à l'échelle, ont une espérance identique.

Annexe 2

Programme SAS pour l'obtention d'estimations par la vraisemblance pénalisée de Firth

Les instructions SAS à la fin de la section sont ajustées à un modèle de régression à risques proportionnels utilisant des poids mis à l'échelle dans une vraisemblance pénalisée de Firth. L'instruction PROC invoque la procédure et l'option VARMETHOD = JK lance une requête d'estimation de la variance par la méthode du jackknife. On peut aussi spécifier VARMETHOD = TAYLOR, VARMETHOD = BRR ou VARMETHOD = BOOT pour lancer une requête de méthode d'estimation de la variance par linéarisation en séries de Taylor, par répliques répétées équilibrées ou par rééchantillonnage bootstrap, respectivement. La sous-option DETAILS de l'option VARMETHOD = JK imprime les estimations de chaque échantillon répété ainsi que l'état de convergence. L'instruction WEIGHT spécifie les poids d'échantillonnage, l'instruction STRATA spécifie les strates et l'instruction CLUSTER spécifie les UPE. L'instruction MODEL spécifie le modèle d'analyse. L'option FIRTH dans l'instruction MODEL lance une requête de vraisemblance pénalisée de Firth. Les deux instructions HAZARDRATIO lancent des requêtes de rapports des risques pour le cholestérol sanguin et le tabagisme, respectivement. L'instruction ODS OUTPUT stocke les estimations par répliques et l'état de convergence de chaque réplique dans l'ensemble de données SAS RepEstimatesFirth. Cet ensemble de données est utile aux fins de vérification de l'état de convergence de chaque échantillon répété.

```
proc surveyphreg data = NHEFS varmethod=jk (details);
      class        GenderHighBloodChol Race Smoker;
      weight       ObservationWeight;
      strata       Stratum;
      cluster      PSU;
      model        EventTime*HeartAttack(2) = Income HighBloodChol
                   Smoker Race Gender Race*Gender / firth;
      Hazardratio  HighBloodChol;
      Hazardratio  Smoker;
      ods output   repestimates=RepEtimatesFirth;
run;
```

Bibliographie

Andersen, P.K., et Gill, R.D. (1982). Cox's regression model counting process: A large sample study. *Annals of Statistics*, 10, 1100-1120.

Bender, R., Augustin, T. et Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24, 1713-1723.

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

Binder, D.A. (1990). Fitting Cox's proportional hazards models from survey data. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 342-347.

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika*, 79, 139-147.

Binder, D.A., et Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89(427), 1035-1043.

Boudreau, C., et Lawless, J.F. (2006). Survival analysis based on the proportional hazards model and survey data. *Canadian Journal of Statistics*, 34, 203-216.

Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30, 89-99.

Chambers, R.L., et Skinner, C.J. (2003). *Analysis of Survey Data*. Chichester, Royaume-Uni: John Wiley & Sons, Inc.

Chambless, L.E., et Boyle, K.E. (1985). Maximum likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics - Theory and Methods*, 14, 1377-1392.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, (avec discussion), 34, 187-220.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, 62, 269-276.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27-38.

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

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

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

Heinze, G. (1999). *The Application of Firth's Procedure to Cox and Logistic Regression*. Rapport technique 10, mis à jour en janvier 2001, Department of Medical Computer Sciences, Section of Clinical Biometrics, University of Vienna.

Heinze, G., et Schemper, M. (2001). A solution to the problem of monotone likelihood in Cox regression. *Biometrics*, 51, 114-119.

Heinzel, G., Rüdiger, A. et Schilling, R. (2002). *Spectrum and Spectral Density Estimation by the Discrete Fourier Transform (DFT), Including a Comprehensive List of Window Functions and Some New Flat-Top Windows*. Rapport technique, Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), Teilinstitut Hannover.

Kish, L., et Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

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

Lee, E.W., Wei, L.J. et Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. *Survival Analysis: State of the Art*, (Éds., J.P. Klein et P.K. Goel), Dordrecht, Pays-Bas: Kluwer Academic, 237-247.

Lin, D.Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika*, 87, 37-47.

Lin, D.Y., et Wei, L.J. (1989). The robust inference for the proportional hazards model. *Journal of the American Statistical Association*, 84, 1074-1078.

Mukhopadhyay, P.K. (2010). Not hazardous to your health: Proportional hazards modeling for survey data with the SURVEYPHREG procedure. *Proceedings of the SAS Global Forum 2010 Conference*. Cary, Caroline du Nord: SAS Institute Inc. http://support.sas.com/resources/papers/proceedings10/254-2010.pdf.

Mukhopadhyay, P.K., An, A.B., Tobias, R.D. et Watts, D.L. (2008). Try, try again: Replication-based variance estimation methods for survey data analysis in SAS 9.2. *Proceedings of the SAS Global Forum 2008 Conference*. Cary, Caroline du Nord: SAS Institute Inc. http://www2.sas.com/proceedings/forum2008/367-2008.pdf.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

Rao, J.N.K. (2005). Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage. *Techniques d'enquête*, 31, 2, 127-151. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9040-fra.pdf.

SAS Institute Inc. (2018). *SAS/STAT 15.1 User's Guide*. Cary, NC: SAS Institute Inc. http://go.documentation.sas.com/?docsetId=statug&docsetTarget=titlepage.htm&docsetVersion=15.1&locale=en.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135-154.

Wolter, K.M. (2007). *Introduction to Variance Estimation*. 2nd Ed. New York: Springer.