

# QUELLE EST LA RELATION ENTRE LES GAINS HEBDOMADAIRES MOYENS ET LA PAIE MENSUELLE MOYENNE ? UNE APPROCHE PAR VARIABLE LATENTE

Serge Godbout<sup>1</sup>

## RÉSUMÉ

Dans l'Enquête sur l'emploi, la rémunération et les heures de travail, l'estimation des gains hebdomadaires totaux se fait à l'aide d'un estimateur par régression utilisant la paie mensuelle moyenne comme donnée auxiliaire. Le modèle actuellement utilisé inclut une ordonnée à l'origine, ce choix étant basé sur des résultats empiriques à partir des données observées. À partir de leur définition, nous poserons des hypothèses sur les variables basées sur la présence d'une variable latente. Nous développerons la distribution de densité conditionnelle entre les variables observées. Nous comparerons la performance de différents modèles à partir de résultats de simulations.

MOTS CLÉS : Estimateur par la régression généralisée; modèle de régression; modélisation à variables latentes.

## ABSTRACT

In the Survey of Employment, Payrolls and Hours, the estimation of the total weekly earnings is done using a regression estimator with average monthly earnings as an auxiliary variable. The model currently used, a regression line, includes an intercept. The choice to use an intercept was based on empirical results from observed data. We propose a hypothesis on the distribution of the variables based on their definition and the latent variable. We will calculate the conditional density function of the observed variables. Based on a simulation, we will evaluate the performance of different models.

KEY WORDS: Generalized regression estimator, Latent variable modelling, Regression model

## 1. INTRODUCTION

### 1.1 Description du problème

L'Enquête sur l'emploi, la rémunération et les heures de travail (EERH) de Statistique Canada est une enquête mensuelle utilisant deux sources de données : un recensement de dossiers administratifs et une enquête auprès d'un échantillon d'établissements. L'EERH a pour objectif de produire des estimations de niveaux et de tendances pour l'emploi, les gains, les heures et autres variables connexes et ce, par province et par industrie (Grondin *et al.*, 2005).

La source administrative est constituée de formulaires de retenues à la source fournis par l'Agence du revenu du Canada (ARC). Nous avons ainsi, pour chaque établissement  $k$  de l'univers des établissements  $U$ , le nombre d'employés  $E_k$ , la paye mensuelle brute  $P_k$  et la paie mensuelle moyenne par employé  $x_k = P_k / E_k$ . La paie mensuelle moyenne par employé est définie comme étant le « *montant total avant les retenues de toute la rémunération versée aux salariés au cours du mois de référence [...] (salaires réguliers, paiements réguliers pour les heures supplémentaires ainsi que les paiements spéciaux)* » (Statistique Canada, 2005) divisé par le nombre d'employés de l'établissement. La portion enquête, quant à elle, consiste en un échantillon,  $s$ , stratifié d'environ 11 000 établissements choisis à partir d'une base liste tirée du Registre des entreprises (RE) de Statistique Canada. Ces établissements peuvent être reliés à la source administrative. Le poids de sondage des unités  $k$  de la strate  $h$  est  $w_{hk} = 1 / \pi_{hk}$ , où  $\pi_{hk}$  correspond à la probabilité de sélection de l'unité  $k$  appartenant à la strate  $h$ . Parmi les variables recueillies pour chaque unité  $k \in s$ , nous comptons les gains hebdomadaires moyens par employé  $y_k$  définis comme étant « *la partie de la rémunération mensuelle brute qui*

<sup>1</sup> Serge Godbout, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Immeuble R.-H.-Coats 11<sup>e</sup> étage, 100, promenade du Pré Tunney, Ottawa, K1A 0T6, serge.godbout@statcan.ca

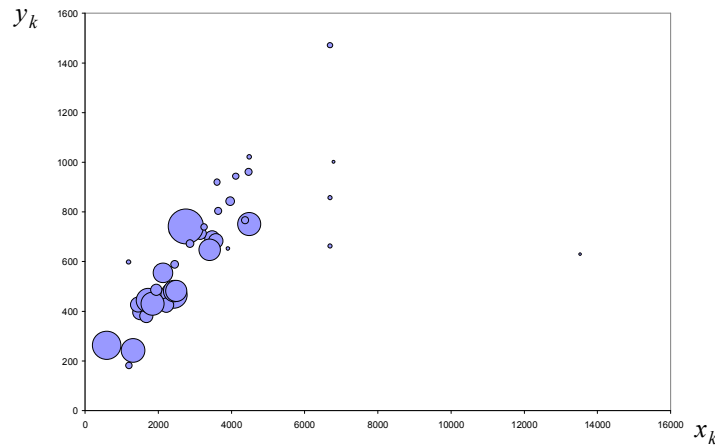
reflète le travail accompli durant la semaine de référence [divisée par le nombre d'employés] [...] Les paiements annuels spéciaux sont exclus tandis que les paiements forfaitaires spéciaux sont rajustés de façon à coïncider avec la semaine de référence. » (Statistique Canada, 2007).

Même s'il existe des différences importantes au niveau de la période de couverture et des différents paiements spéciaux, les  $x_k$  représentent une variable auxiliaire très intéressante pour l'estimation des gains hebdomadaires totaux ( $\hat{t}_y$ ). Ceux-ci sont actuellement estimés à l'aide d'un estimateur par régression du type par projection simple tel que décrit par Särndal *et al.* (1992) dont le modèle  $\xi$  est estimé au niveau de groupes de modélisation  $g$  :

$$\hat{t}_y^{REG} = \sum_U E_k \hat{y}_k \text{ avec } E_\xi(y_k) = \mathbf{x}'_k \mathbf{B}_g = B_{0g} + B_{1g} x_k \text{ et } V_\xi(y_k) = \sigma_g^2 / E_k \quad (1)$$

Un bon modèle aura pour effet de réduire la variance de l'estimateur (Särndal *et al.*, 1992). Notre modèle actuel a été établi à partir de l'étude de données empiriques et de la significativité des paramètres estimés. Cependant, les observations montrent que la relation demeure quelque peu ambiguë (comme l'illustre la figure 1) ayant pour résultat que ce modèle a souvent été mis en doute.

**Figure 1 – Dispersion des gains hebdomadaires moyens par rapport à la paie mensuelle moyenne**



Chaque bulle correspond à un établissement échantillonné et sa taille est déterminée par l'emploi pondéré ( $w_k E_k$ ). Nous voyons qu'un grand nombre de points suivent une relation linéaire entre les  $x_k$  et les  $y_k$ . Cependant, certains points se détachent vers la droite de cette ligne, généralement en raison de la présence de paiements spéciaux ou de périodes de paie supplémentaires dans les  $x_k$ . L'ajout de ces deux concepts à la paie mensuelle moyenne pose un problème majeur : quelle est la relation entre les  $y_k$  et les  $x_k$  ? Et surtout, quel est l'impact du modèle sur la précision de l'estimateur par régression ?

Dans ce document, nous poserons des hypothèses sur les variables  $x$  et  $y$  basées sur la présence d'une variable latente  $z$ . Nous développerons la distribution de densité conditionnelle entre les variables observées et déduirons les fonctions d'espérance et de variance conditionnelles, nous donnant notre modèle théorique. Finalement, nous comparerons la performance de ce modèle avec d'autres sur l'estimation des  $\hat{t}_y$  à l'aide de résultats de simulations.

## 2. MODÉLISATION DES GAINS HEBDOMADAIRES MOYENS SELON LA PAIE MENSUELLE MOYENNE

### 2.1 Hypothèses de base

Nous allons travailler avec un modèle simplifié mais dont les conclusions demeureront valides pour la définition du modèle réel et de la compréhension de son rôle dans l'estimation. Nous éliminerons l'emploi comme poids économique et nous poserons que les  $x_k$  comprennent un nombre fixe de semaines de paie. Nous sommes intéressés à estimer le total de la variable  $y_k$  à l'aide d'un estimateur par régression :

$$\hat{t}_y^{REG} = \sum_U \hat{y}_k \text{ où } \hat{y}_k = f_\xi(x_k) \quad (2)$$

À partir des définitions données à la section 1.1, posons d'abord une variable latente  $z_k$  permettant de dériver à la fois  $x_k$  et  $y_k$ . Nous définirons cette variable comme étant la paie mensuelle « régulière » moyenne, i.e. excluant les

paiements spéciaux inclus dans les  $x_k$  mais pas dans les  $y_k$ . De cette façon, nous pouvons décrire  $x_k$  et  $y_k$  à partir de  $z_k$  de la manière suivante :

$$x_k = z_k + SP_k + \varepsilon_{1,k} \text{ et } y_k = \beta_k z_k + \varepsilon_{2,k} \quad (3)$$

Nos hypothèses seront construites sans ajout d'erreur aléatoire, c'est-à-dire que les termes d'erreur  $\varepsilon_{1,k}$  et  $\varepsilon_{2,k}$  sont fixés à 0 afin de simplifier la dérivation du modèle sans perdre de généralité. La variable  $SP_k$  correspond aux paiements spéciaux inclus dans les  $x_k$  mais pas dans les  $y_k$ . La variable  $\beta_k$  correspond au nombre de semaines de paie dans un mois. Encore pour des raisons de simplification, nous allons fixer  $\beta_k = \beta$  à une constante. Afin de finaliser les hypothèses de base de notre modèle, nous allons poser que  $z_k$  suit une loi normale et que  $SP_k$  suit une loi exponentielle de paramètre  $\lambda$ . Ces distributions simples ressemblent à celles observées dans les données réelles. Ainsi, le modèle pour lequel nous chercherons la relation entre  $x_k$  et  $y_k$  sera le suivant :

$$x_k = z_k + SP_k \text{ et } y_k = \beta z_k \text{ où } \beta = \text{Constante}, z_k \rightarrow N(\mu_z, \sigma_z^2) \text{ et } SP_k \rightarrow EXP(\lambda) \quad (4)$$

Puisque  $SP_k > 0$ , nous devons noter que  $x_k > z_k$  et que  $\beta x_k > y_k$ .

## 2.2 Fonction de densité conditionnelle

Pour connaître le modèle reliant les variables  $x$  et  $y$ , nous trouverons la fonction de densité conditionnelle  $f_{Y|X}(y|x)$  avec l'aide de la théorie sur les variables aléatoires (Casella et Berger, 2002). Tout d'abord, à partir des hypothèses du modèle (4), nous pouvons obtenir les fonctions de densité suivantes :

$$f_Z(z) = \frac{1}{\sigma_z \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{z - \mu_z}{\sigma_z}\right)^2\right) \text{ et } f_{X|Z}(x|z) = \frac{1}{\lambda} \exp\left(-\frac{1}{\lambda}(x - z)\right) \quad (5)$$

La fonction de densité de  $Y$  est obtenue à partir d'une transformation de la fonction  $f_Z$  :

$$f_Y(y) = f_Z(y/\beta) \frac{d(y/\beta)}{dy} = \frac{1}{\beta \sigma_z \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y - \beta \mu_z}{\beta \sigma_z}\right)^2\right) \quad (6)$$

Puisque les fonctions  $f_Y(y)$  et  $f_{X|Z}(x|z)$  sont indépendantes, la fonction de densité conjointe  $f_{XY}(x, y)$  est donnée par  $f_{XY}(x, y) = f_{X|Z}(x|z) f_Y(y)$  sur la région  $y < \beta x$ . Ainsi, nous trouvons la fonction de densité marginale  $f_X(x)$  en intégrant la fonction de densité conjointe  $f_{XY}(x, y)$  par rapport à la variable  $y$  sur la région  $y < \beta x$  :

$$f_X(x) = \int_{-\infty}^{\beta x} f_{XY}(x, y) dy = \frac{1}{\lambda} \exp\left(\frac{\sigma_z^2 + 2\mu_z \lambda - 2x\lambda}{2\lambda^2}\right) \Phi\left(\frac{\beta x - \frac{\beta(\sigma_z^2 + \mu_z \lambda)}{\lambda}}{\beta \sigma_z}\right) \quad (7)$$

Où  $\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$  est la fonction de distribution cumulative de la loi normale. Maintenant, la fonction de densité conditionnelle  $f_{Y|X}(y|x)$  devient :

$$f_{Y|X}(y|x) = f_{XY}(x, y) / f_X(x) = \frac{1}{\beta \sigma_z \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y - \frac{\beta(\sigma_z^2 + \mu_z \lambda)}{\lambda}}{\beta \sigma_z}\right)^2\right) \Bigg/ \Phi\left(\frac{\beta x - \frac{\beta(\sigma_z^2 + \mu_z \lambda)}{\lambda}}{\beta \sigma_z}\right) \text{ sur } -\infty < y < \beta x \quad (8)$$

Cette fonction de densité conditionnelle est une distribution normale, de moyenne et de variance indépendantes de  $x$ , mais tronquée sur le domaine  $y < \beta x$ . Cette distribution est fortement asymétrique lorsque  $x$  prend des petites valeurs mais tend vers une distribution normale lorsque  $x$  tend vers l'infini.

## 2.2 Espérance et variance de $y$ conditionnelles à $x$

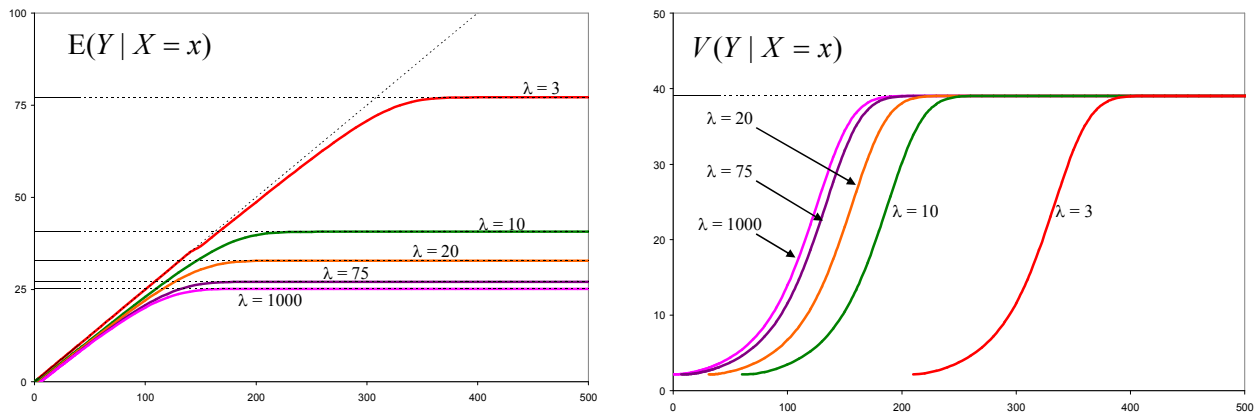
Pour connaître la forme du modèle de  $y$  conditionnel à  $x$  (que nous nommerons A), nous allons d'abord calculer l'espérance et la variance conditionnelles à partir de la fonction de densité conditionnelle (8).

$$E_A(Y | X = x) = \int_{-\infty}^{\beta x} y f_{Y|X}(y | x) dy = \frac{\beta(\sigma_z^2 + \mu_z \lambda)}{\lambda} - \beta \sigma_z P\left(\frac{\beta x - \frac{\beta(\sigma_z^2 + \mu_z \lambda)}{\lambda}}{\beta \sigma_z}\right) \quad (9)$$

$$V_A(Y | X = x) = \int_{-\infty}^{\beta x} y^2 f_{Y|X}(y | x) dy - E^2(Y | X = x) = \beta^2 \sigma_z^2 \left(1 + \frac{\sigma_z^2 + \lambda(\mu_z - x)}{\sigma_z \lambda} P(x) - P^2(x)\right) \quad (10)$$

Où la fonction  $P(x)$ , le rapport entre la fonction de densité de probabilité et sa fonction cumulative, est décroissante et tend vers 0 lorsque  $x$  tend vers l'infini. Ainsi, l'espérance conditionnelle  $E(Y | X = x)$  a un comportement asymptotique : elle suit  $y = \beta x$  lorsque  $x$  est près de zéro et elle tend vers  $\beta(\sigma_z^2 + \mu_z \lambda) / \lambda$  lorsque  $x \rightarrow \infty$ . La variance conditionnelle  $V(Y | X = x)$  a également un comportement asymptotique : elle débute par une valeur près de zéro puis croît rapidement en convergeant vers  $y = \beta^2 \sigma_z^2$  lorsque  $x \rightarrow \infty$ . Pour  $\beta = 1/4$ ,  $\mu_z = 100$  et  $\sigma_z^2 = 625$ , nous avons tracé les fonctions  $E(Y | X = x)$  et  $V(Y | X = x)$  lorsque  $\lambda = \{3, 10, 20, 75, 1000\}$  (voir figure 2).

**Figure 2 – Espérance et variance conditionnelles pour différentes valeurs de lambda**

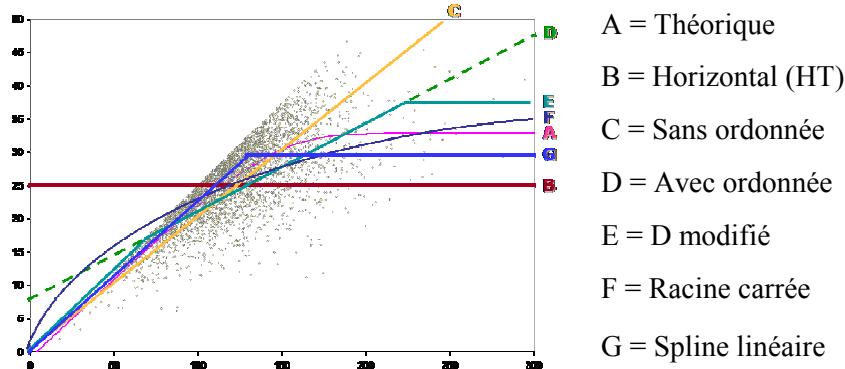


### 2.3 Modèles approximatifs

Le modèle théorique dérivé en (9) et (10) n'a pas une forme standard et il n'est pas possible de l'estimer directement par régression. Bien que nous sommes intéressés à connaître le modèle reliant  $x_k$  et  $y_k$ , notre objectif n'est pas de l'estimer mais bien d'estimer le total de la variable  $y$  en utilisant le modèle via l'estimateur par régression. Ainsi, toute erreur dans le modèle estimé sera acceptée en autant que l'estimateur par régression du total demeure performant.

Nous allons donc adopter 2 stratégies. Premièrement, nous allons estimer le modèle théorique (noté A) en estimant séparément ses paramètres  $\hat{\mu}_z$ ,  $\hat{\sigma}_z^2$  et  $\hat{\lambda}$ . Ce modèle ne fonctionnera que si la population suit exactement les hypothèses initialement posées (4) ayant servi à le dériver. Deuxièmement, nous allons utiliser des modèles approximatifs plus simples qui pourraient remplacer le modèle théorique. Nous étudierons 6 modèles approximatifs choisis pour leur forme semblable au modèle théorique et pour leur simplicité d'utilisation : horizontal (B, équivalent à un estimateur Horvitz-Thompson), linéaire sans ordonnée à l'origine (C, équivalent à un estimateur par ratio), linéaire avec ordonnée à l'origine (D), linéaire par morceaux dérivé à partir du modèle D (E), par racine carrée (F) et par spline linéaire (G). Le tableau A-1 (en annexe) décrit en détails ces modèles et leur méthode d'estimation et la figure 3 les illustre.

Figure 3 – Modèles de régression étudiés

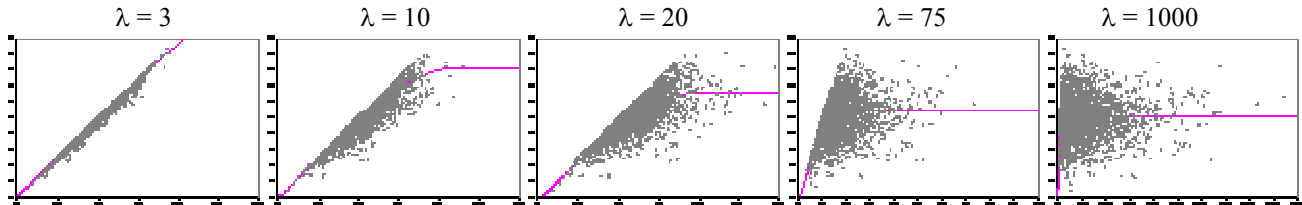


### 3. ÉTUDE PAR SIMULATION

#### 3.1 Simulation Monte Carlo

Nous avons créé aléatoirement 5 populations différentes de 4000 unités à partir des hypothèses (4) avec  $\sigma_z = 25$ ,  $\mu_z = 100$  et  $\beta = 0,25$ . Nous avons d'abord généré un ensemble de 4000 valeurs de  $z_k$  et de  $y_k = 0,25z_k$  communes aux 5 populations. Seuls les  $x_k = z_k + SP_k$  diffèrent entre les 5 populations, étant générés par des  $\lambda$  respectivement égaux à 3, 10, 20, 75 et 1000. Ces 5 valeurs de  $\lambda$  visent à simuler des populations contenant peu, passablement ou une part majeure de paiements spéciaux. Les 5 populations de points  $(x_k, y_k)$  ainsi créées sont illustrées dans la figure 4, tout comme la courbe correspondant au modèle théorique A.

Figure 4 – Dispersion des  $x$  et  $y$  et modèle théorique selon les 5 populations créées



Par la suite, nous avons sélectionné de façon synchronisée pour les 5 populations un total de 500 échantillons aléatoires simples indépendants de taille 25. Nous avons estimé à chaque fois le total  $\hat{t}_y^{REG}$  pour les 7 modèles. La performance des modèles a été mesurée par un coefficient d'adéquation moyenne (CAM), par le biais relatif moyen Monte Carlo ( $BRM^{MC}$ ) et par la racine de l'erreur quadratique moyenne relative Monte Carlo ( $REQMR^{MC}$ ) calculés comme suit :

$$CAM = \frac{1}{500} \left( \sum_{i=1}^{500} \left( 1 - \frac{\sum_{k \in S_i} w_k (y_k - \mathbf{x}'_k \hat{\mathbf{B}}_i)^2}{\sum_{k \in S_i} w_k (y_k - \hat{y}_i)^2} \right) \right), \quad BRM^{MC} = \frac{1}{\hat{t}_{yi}} \left| \frac{1}{500} \sum_{i=1}^{500} \hat{t}_{yi} - t_y \right| \quad \text{et} \quad REQMR^{MC} = \frac{1}{\hat{t}_{yi}} \left( \frac{1}{500} \sum_{i=1}^{500} (\hat{t}_{yi} - t_y)^2 \right)^{1/2} \quad (11)$$

Le CAM est équivalent à la moyenne des coefficients de détermination  $\hat{R}^2$  dans le cas de modèles linéaires avec ordonnées à l'origine (comme le modèle D). Une adéquation parfaite aura un CAM égal à 1. Une valeur égale à 0 signifie une performance équivalente à un modèle par la moyenne des  $y$  (horizontal) alors qu'une valeur négative signifie une performance pire que le modèle horizontal. Cependant, nous considérons que le meilleur estimateur est celui qui a un  $REQMR^{MC}$  le plus près de 0 tout en conservant un  $BRM^{MC}$  négligeable par rapport à la  $REQMR^{MC}$ . Le tableau 1 présente les résultats de l'étude par simulation sur la performance des modèles.

Tableau 1 – Performance du modèle estimé et de l'estimateur par régression

Modèles	CAM (%)					BRM <sup>MC</sup> (%)					REQMR <sup>MC</sup> (%)				
	$\lambda = 3$	10	20	75	1000	3	10	20	75	1000	3	10	20	75	1000
A	98	85	64	25	2	0.0	0.0	0.0	0.0	0.0	0.6	2.0	3.1	4.0	4.6
B	0	0	0	0	0	0.0	0.1	0.1	0.1	0.1	4.6	4.6	4.6	4.6	4.6
C	98	84	46	-246	-1428	0.0	0.0	0.0	0.0	0.0	0.6	2.0	3.6	9.4	21.9

D	98	85	60	13	4	0.0	0.0	0.1	0.8	3.9	0.6	2.0	3.3	5.0	4.7
E	98	85	61	17	-17	0.0	0.1	0.3	0.7	0.1	0.9	2.0	3.1	4.4	4.6
F	74	67	54	-21	-349	0.4	0.4	0.3	0.0	0.1	2.4	2.7	3.2	5.3	10.7
G	98	84	55	16	3	0.1	0.1	0.0	0.1	0.9	0.6	2.0	3.5	4.5	4.8

Nous voyons que le modèle A est toujours le meilleur pour les 5 valeurs de  $\lambda$  et selon les 3 critères. Certains modèles sont aussi efficaces que le A pour des valeurs précises de  $\lambda$  : le modèle B si  $\lambda$  est très grand, C si  $\lambda$  est petit et F si  $\lambda$  est moyen. Les modèles D, E et G sont aussi efficaces que le A pour toutes les valeurs de  $\lambda$ . Cependant, comme le modèle A est fortement dépendant des hypothèses posées, son efficacité relative est complètement annulée par l'incapacité de l'utiliser dans un contexte réel. Par exemple, nous avons généré une sixième population simulant une petite erreur dans les hypothèses de base en posant que les  $SP_k$  suivaient une loi gamma (et non une loi exponentielle). L'impact a été catastrophique sur la  $REQMR^{MC}$  du modèle A tandis que tous les autres modèles ont donné des résultats similaires à ceux du tableau 1. À noter que d'autres modèles utilisant des algorithmes d'estimations plus complexes ont été étudiés, comme des modèles exponentiels ou par morceaux. Cependant, nous avons choisi de ne pas les présenter en raison de l'espace limité disponible pour cet article. Leurs résultats ne modifient en rien nos conclusions.

#### 4. CONCLUSION

Dans le cas de l'estimation des gains hebdomadaires moyens, le modèle reliant les variables, même dérivé d'hypothèses simplifiées, est une courbe complexe et sensible aux hypothèses de base. Cependant, l'utilisation de modèles de régression plus simples et flexibles dans l'estimation par régression permet d'obtenir d'aussi bons résultats. De plus, ces modèles plus robustes s'adaptent aux variations dans les hypothèses de base. Nous croyons que dans la réalité de l'EERH, l'utilisation d'un modèle linéaire avec ordonnée à l'origine (D) répond efficacement aux besoins de l'estimateur par régression. Aussi, les modèles modifiés (E) et par spline linéaire (G) peuvent être considérés comme des améliorations, particulièrement efficaces dans le cas de distributions des paiements spéciaux comportant une queue plus lourde, grâce à son extrapolation horizontale vers la droite.

#### 4.1 Remerciements

L'auteur tient à remercier grandement Yves Morin, Jean-François Beaumont et Carlos Leon de même que les réviseurs.

#### ANNEXE

**Tableau A-1 – Modèles de régression étudiés et méthodes d'estimation**

Nom	Modèle	Estimation
A = Théorique	Fonctions dérivées en (9) et (10)	$\hat{\mu}_z = 4\hat{y}$ , $\hat{\sigma}_z^2 = 16\hat{\sigma}_y^2$ , $\hat{\lambda} = \hat{x} - \hat{\mu}_z$ et $\hat{y}_k = E_A(x_k)$
B = Horizontal (HT)	$E_B(y_k) = B_{0B}$ , $V_B(y_k) = \sigma^2$	$\hat{B}_{0B} = \sum_s w_k y_k / \sum_s w_k$ et $\hat{y}_k = \hat{B}_{0B}$
C = Sans ordonnée	$E_C(y_k) = B_{1C} x_k$ , $V_C(y_k) = x_k \sigma^2$	$\hat{B}_{1C} = \sum_s w_k y_k / \sum_s w_k x_k$ et $\hat{y}_k = \hat{B}_{1C} x_k$
D = Avec ordonnée	$E_D(y_k) = B_{0D} + B_{1D} x_k = \mathbf{x}'_k \mathbf{B}_D$ , $V_D(y_k) = \sigma^2$	$\hat{\mathbf{B}}_D = \left( \sum_s w_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \sum_s w_k \mathbf{x}_k y_k$ et $\hat{y}_k = \mathbf{x}'_k \hat{\mathbf{B}}_D$
E = D modifié	$E_E(y_k) = \text{MIN}(0,25x_k, \mathbf{x}'_k \mathbf{B}_D)$ , $V_E(y_k) = V_D(y_k)$	$\hat{y}_k = \begin{cases} \text{MIN}(0,25x_k, \hat{B}_{0D} + \hat{B}_{1D} x_k) & \text{si } x_k < \text{MAX}_s(x_k) \\ \hat{B}_{0D} + \hat{B}_{1D} \text{MAX}_s(x_k) & \text{si } x_k \geq \text{MAX}_s(x_k) \end{cases}$
F = Racine carrée	$E_F(y_k) = B_{1F} \sqrt{x_k}$ , $V_F(y_k) = \sqrt{x_k} \sigma^2$	$\hat{B}_{1F} = \sum_s w_k y_k / \sum_s w_k \sqrt{x_k}$ et $\hat{y}_k = \hat{B}_{1F} \sqrt{x_k}$
G = Spline linéaire	$\begin{cases} E_G(y_k) = B_{1G} x_k \text{ et } V_G(y_k) = x_k \sigma^2 & \text{si } x_k < \psi \\ E_G(y_k) = B_{1G} \psi \text{ et } V_G(y_k) = \psi \sigma^2 & \text{si } x_k \geq \psi \end{cases}$	Pour $\psi \in [0, \text{MAX}_s(x_k)]$ minimisant $\hat{R}^2 = \sum_s w_k (y_k - \hat{y}_k)^2 / \sum_s w_k y_k^2$ , choisir $\hat{B}_{1G} = \sum_s w_k y_k / \sum_s w_k \text{MIN}(x_k, \psi)$ et $\hat{y}_k = \hat{B}_{1G} \text{MIN}(x_k, \psi)$

## RÉFÉRENCES

Casella G. et Berger R.L. (2002). *Statistical Inference* (2<sup>e</sup> édition). Pacific Grove: Duxbury.

Grondin C., Lavallée P. et Godbout S. (2005). « Current Methodology of the Survey of Employment, Payrolls and Hours ». Ottawa, Canada: Statistics Canada.

Särndal C.-E., Swensson B. et Wretman J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag

Statistique Canada (2007). « Emploi, gain et durée du travail (Juin 2007) ». Division de la statistique du travail. No 72-002-XIB au catalogue

Statistique Canada (2005). « Votre guide de l'Enquête sur la rémunération auprès des entreprises ». Division de la statistique du travail.