Techniques d'enquête, juin 1998
Vol. 24, n° 1, pp. 53-58
Statistique Canada



# Estimateurs de régression généralisés logistiques

RISTO LEHTONEN et ARI VEIJANEN[1]

RÉSUMÉ

Dans cet article, les auteurs examinent comment estimer la fréquence des classes d'une variable discrète associée aux réponses par le biais d'un modèle. L'estimation fait appel à une nouvelle méthode d'estimation des données d'enquête, étroitement liée à l'estimation par régression généralisée. Dans cette dernière, les données sur les variables auxiliaires sont intégrées à la méthode d'estimation par ajustement au moyen d'un modèle linéaire. Au lieu de recourir à un modèle linéaire pour les indicateurs de classe, nous décrivons la distribution combinée des indicateurs de classe par un modèle logistique multinominal. Les auteurs présentent des estimateurs de régression généralisés logistiques pour la fréquence des classes au sein d'une population et de divers domaines. Ils ont entrepris des essais de Monte Carlo sur des données simulées et les données réelles issues de l'Enquête sur la population active menée chaque mois par Statistics Finland. L'estimation de régression généralisée logistique donne de meilleurs résultats que l'estimation de régression ordinaire pour les petits domaines, en particulier les classes à faible fréquence.

MOTS CLÉS: Information auxiliaire; fréquence des classes; modèles linéaires généralisés; enquête sur la population active; estimation assistée par modèle; estimateurs de régression.

## 1. INTRODUCTION

Envisageons l'estimation de la fréquence des classes d'une variable discrète associée à une réponse dans le cadre d'une enquête par sondage. Le nombre de sujets dans la classe correspond à la somme des indicateurs de la classe pour la population, bref au total de cet indicateur. On peut donc recourir à des méthodes conçues pour estimer la population totale afin de résoudre le problème. Pour accroître la précision de l'estimation, le statisticien d'enquête fera souvent appel aux données auxiliaires existantes. Si la valeur probable de la variable associée à la réponse présente un lien linéaire avec les variables auxiliaires, comme cela peut se produire avec les variables continues, il vaut la peine de recourir à l'estimateur de régression généralisé (Särndal, Swensson et Wretman 1992; Estevao, Hidiroglou et Särndal 1995). En effet, la régression généralisée peut donner une meilleure estimation et atténuer le biais attribuable à la non-réponse unitaire si les variables secondaires présentent de fortes corrélations avec la variable principale de la réponse.

Du point de vue du concepteur, un modèle linéaire s'avère fort restrictif et pourrait ne pas constituer le choix idéal pour les variables binaires comme la situation d'emploi d'une personne (occupée, au chômage) ni pour les variables discrètes comme sa situation sur le marché du travail (occupée, au chômage, inactive), d'une manière plus générale. Nous proposons pour ces variables un genre d'estimateur de régression généralisé logistique s'inspirant du modèle logistique multinominal qui décrit la distribution combinée des indicateurs de classe. La raison pour laquelle ce modèle a été retenu est donc identique à celle évoquée pour les modèles linéaires généralisés (McCullagh et Nelder 1989).

Nous estimerons les paramètres du modèle logistique en maximisant le logarithme du rapport de vraisemblance pondéré d'un échantillon, soit l'estimateur de Horvitz-Thompson pour la fonction du rapport de vraisemblance de la population (Godambe et Thompson 1986; Nordberg 1989; Skinner, Holt et Smith 1989; Särndal et coll. 1992, p. 517).

Comme application, nous prendrons l'estimation du taux de chômage dans le cadre de l'Enquête sur la population active effectuée chaque mois par Statistics Finland. Les dossiers administratifs indiquant si une personne à la recherche d'un emploi est inscrite au bureau d'emploi local peuvent servir de données auxiliaires d'un registre et ont été combinés aux données d'enquête sur chaque sujet grâce au numéro d'identité personnel, unique dans chaque source de données. La variable auxiliaire correspondante présente une étroite corrélation avec les résultats de l'enquête sur le chômage. L'usage de ces données administratives dans l'estimation devrait donc améliorer cette dernière et atténuer le biais. D'autres données auxiliaires (sexe, âge, données régionales) viennent du Registre de la population. Elles aussi ont été combinées aux données d'enquête au niveau individuel.

Nous avons étudié les propriétés des estimateurs de régression généralisés avec les techniques de simulation de Monte Carlo en vertu desquelles des échantillons EASSR ont été prélevés de façon répétitives d'une population construite à partir des données de l'Enquête sur la population active. Nous avons recouru à une stratification a posteriori incomplète ou à une procédure itérative reposant sur un modèle d'analyse de la variance des effets principaux. Les résultats des essais révèlent que la formulation logistique donne de meilleurs résultats que la formulation

[1] Risto Lehtonen et Ari Veijanen, Statistics Finland, P.O. Box 5A, FIN-00022 Statistics Finland, Finlande.

Monte Carlo de l'estimateur HT (Lehtonen et Pahkinen 1996). Nous avons déterminé la précision globale des estimations relatives aux domaines grâce à l'erreur moyenne absolue relative au domaine (EMARD), pour $D$ domaines et $K$ échantillons $s_j$:

$$\text{EMARD}(i) = \frac{1}{D}\sum_{p=1}^{D}\frac{1}{K}\sum_{j=1}^{K}\frac{100\left|\hat{t}_{(d_p)i}(s_j) - t_{(d_p)i}\right|}{t_{(d_p)i}}.$$

Dans les estimations REGG (2), la variance correspondait à une constante $\sigma_{ki}^2 = \sigma^2$, dont les valeurs se sont annulées. Avec REGGL, on a estimé la fréquence des domaines au moyen de l'équation (5) et la variance avec l'équation (7). Pour la méthode REGG et l'estimateur HT, lire Särndal et coll. (1992, p. 401). On a calculé les intervalles de confiance des fréquences comme si les indicateurs de classe étaient indépendants. Au seuil de signification nominal de 95 %, le taux de couverture acceptable se situe à l'intérieur de [93,65 %, 96,35 %] pour $K$ = 1 000 échantillons simulés.

### 4.2 Essai avec les données simulées

Pour comparer REGGL et REGG, nous avons simulé un ensemble de données dans lequel la variable auxiliaire $X$ correspondait à une variable aléatoire continue, distribuée uniformément dans (-3.3). La variable $Y$, à laquelle on s'intéresse, représentait trois classes suivant la distribution (1) spécifiée par $\boldsymbol{x}_k'\boldsymbol{\beta}_1 = 0$, $\boldsymbol{x}_k'\boldsymbol{\beta}_2 = 3X_k - 1$ et $\boldsymbol{x}_k'\boldsymbol{\beta}_3 = -2X_k$ pour $N$ = 10 000 éléments ($k$ = 1, 2, ..., $N$). On a prélevé un millier d'échantillons de taille $n$ = 1 000 indépendamment par EASSR. $X_k$ et $X_k^2$ servaient de variables auxiliaires. Aucun estimateur ne semble biaisé (tableau 1). Par ailleurs les estimations de la variance présentaient un biais empirique inférieur à 3 % et un écart-type de moins de 5 %.

**Tableau 1**
Effets de plan d'échantillonnage (Deff) sur les estimateurs de la fréquence des classes et taux de couverture empiriques (TC) (%) à l'intervalle de confiance de 95 % pour les classes $i$ = 1, 2, 3

| Méthode | Deff | | | TC | | |
|---|---|---|---|---|---|---|
| | $\hat{t}_1$ | $\hat{t}_2$ | $\hat{t}_3$ | $\hat{t}_1$ | $\hat{t}_2$ | $\hat{t}_3$ |
| HT | 1 | 1 | 1 | 95,2 | 95,3 | 94,7 |
| REGG | 0,93 | 0,55 | 0,57 | 95,0 | 94,3 | 95,6 |
| REGGL | 0,89 | 0,45 | 0,50 | 94,9 | 93,7 | 95,3 |

Les meilleurs résultats sont ceux obtenus avec la méthode REGGL, vraisemblablement parce que les fréquences proportionnelles des classes varient considérablement pour la gamme des variables auxiliaires. La probabilité de chaque classe dépendait tellement de la variable auxiliaire continue que le modèle de régression linéaire s'ajustait mal aux données.

### 4.3 Application aux données de l'Enquête sur la population active finlandaise

#### 4.3.1 Population artificielle

Nous avons examiné l'estimation du taux de chômage au moyen des données de l'Enquête sur la population active (EPA) finlandaise couvrant trois mois successifs de 1994. Cette population comprenait 33 329 sujets. Le Registre de la population nous a permis d'établir le groupe d'âge (15-24, 25-34, 35-44, 45-54 et 55-64 ans), le sexe et la région (trois régions) de chaque sujet. Nous avons aussi tiré un indicateur recherche d'un emploi du registre que garde le ministère du Travail indiquant, quelles personnes au chômage recherchent un emploi. Cette source de données administratives se caractérise par un décalage d'environ deux semaines. La proportion de personnes dont la situation réelle sur le marché du travail change au cours d'un si bref laps de temps devrait être relativement faible. Il convient de souligner que le statut de personne à la recherche d'un emploi dans le registre n'a pas la même définition que dans l'Enquête sur la population active. Dans cette dernière, la mesure repose sur la définition normalisée par le Bureau international du travail (BIT). Toutes les données auxiliaires ont été amalgamées aux données d'enquête, pour chaque sujet.

Le taux de non-réponse varie avec le statut recherche d'un emploi. Ainsi, il s'établissait à 1,4 % pour les personnes inscrites à la recherche d'un emploi et à 7,6 % pour les autres. On a modélisé la probabilité de non-réponse au moyen d'un modèle d'analyse de la variance logistique et les estimations du taux de non-réponse le plus vraisemblable (variant de 2,9 % à 22,8 %) ont servi de modèle de non-réponse dans les simulations.

Pour les simulations, nous avons bâti une population artificielle de $N$ = 30 835 personnes. La situation d'emploi pouvait être de trois sortes: "occupé", "au chômage" et "pas dans la population active" dont la fréquence, au sein de la population, s'établissait respectivement à $t_1$ = 17 373, $t_2$ = 4 433, et $t_3$ = 9 029, Le taux de chômage était défini par $R = t_2/(t_1 + t_2) = 20{,}33$ %. Comme domaine, nous avons utilisé les cellules des totalisations croisées selon le groupe d'âge, le sexe et la situation d'emploi notée au registre.

De la population artificielle, on a prélevé $K$ = 1 000 échantillons indépendants de $n$ = 1 000 (ÉASSR). Le modèle de non-réponse a été ajusté à la population originale pour simuler la non-réponse dans chaque échantillon. Ensuite, on a estimé la probabilité de réponse de chaque échantillon par régression logistique en recourant au même modèle d'analyse de la variance que pour le modèle de non-réponse. Enfin, nous avons multiplié chaque probabilité $\pi_k$ par la probabilité de réponse estimée.

La comparaison de REGGL et REGG repose sur trois modèles. Les composantes de $\boldsymbol{x}_k$ étaient des variables nominales correspondant à l'âge (5 groupes), au sexe, à la région (3 régions) et au statut "recherche d'un emploi".

Pour la stratification a posteriori incomplète, ou la procédure itérative, le modèle d'analyse de la variance des effets principaux reposait sur des variables auxiliaires classées. Nous avons comparé les modèles avec et sans l'indicateur de recherche d'un emploi. Le troisième modèle comprenait un polynôme du quatrième degré pour l'âge.

### 4.3.2 Résultats

Sans les variables auxiliaires, les estimateurs HT aboutissent habituellement à une variance plus élevée que les estimateurs de régression généralisés (tableau 2). Les résultats sont légèrement meilleurs pour ces derniers que pour les estimateurs HT quand on recourt à une procédure itérative combinant l'âge, le sexe et la région. Les résultats étaient nettement supérieurs avec les modèles intégrant l'indicateur de recherche d'un emploi, qui présente une corrélation plus étroite ($r = 0{,}83$) avec l'indicateur du chômage du BIT que les autres variables auxiliaires. Les variables auxiliaires améliorent donc la précision de l'estimation (voir Djerf 1997).

**Tableau 2**
Propriétés des estimations du taux de chômage ($\hat{T}$(%)) selon la procédure itérative du quotient (R) et le modèle incluant le polynôme pour l'âge (P), avec (E) ou sans (N) indicateur de recherche d'un emploi. É.-T. désigne l'écart-type et (TC) (%), le taux de couverture à l'intervalle de confiance de 95 %

| Modèle | Méthode | $\hat{T}$ | Biais | É.-T. | Deff | TC | EMARD |
|---|---|---|---|---|---|---|---|
| | HT | 20,32 | -0,0081 | 1,461 | 1 | 95,7 | 35,28 |
| RN | REGG | 20,30 | -0,0262 | 1,454 | 0,995 | 95,3 | 46,03 |
| RN | REGGL | 20,31 | -0,0229 | 1,454 | 0,995 | 95,3 | 45,93 |
| RE | REGG | 20,30 | -0,0244 | 0,895 | 0,612 | 96,0 | 35,74 |
| RE | REGGL | 20,29 | -0,0419 | 0,901 | 0,617 | 94,8 | 34,80 |
| PE | REGG | 20,30 | -0,0259 | 0,887 | 0,607 | 95,6 | 35,41 |
| PE | REGGL | 20,29 | -0,0421 | 0,896 | 0,613 | 95,1 | 34,76 |

**Tableau 3**
Erreur moyenne absolue relative au domaine (EMARD) et taux de couverture moyen (TC) (%) des intervalles de confiance à 95 % pour la fréquence estimée des classes des domaines dont la fréquence réelle $t_{(d)i}$ ($i$ = 1, 2, 3) a) est inférieure à 100, et b) est au moins égale à 100. Le modèle inclut le polynôme pour l'âge

| | Méthode | EMARD | | | TC | | |
|---|---|---|---|---|---|---|---|
| | | $\hat{t}_{(d)1}$ | $\hat{t}_{(d)2}$ | $\hat{t}_{(d)3}$ | $\hat{t}_{(d)1}$ | $\hat{t}_{(d)2}$ | $\hat{t}_{(d)3}$ |
| a) | REGG | 96,92 | 67,36 | 121,95 | 88,2 | 77,8 | 84,6 |
| | REGGL | 80,28 | 67,20 | 104,05 | 83,9 | 76,5 | 51,7 |
| b) | REGG | 6,95 | 12,31 | 14,35 | 94,1 | 85,9 | 93,7 |
| | REGGL | 6,88 | 12,34 | 14,29 | 93,9 | 85,4 | 93,3 |

Les estimations REGG et REGGL présentent peu de variations au niveau de la population (tableau 2). La méthode REGGL ne s'est jamais avérée inférieure à la méthode REGG. Elle a donné des estimations plus précises des totaux des domaines que la seconde (tableau 3). Quand le modèle inclut l'âge sous forme de variable auxiliaire continue, l'écart-type du taux de chômage estimé est plus faible avec la méthode REGGL qu'avec la méthode REGG dans 19 domaines sur 20. Malheureusement, les intervalles de confiance calculés avec REGGL sont souvent trop étroits à cause des faibles estimations de la variance (tableau 3).

## 5. RÉCAPITULATION

Nous proposons une nouvelle approche à l'estimation de la fréquence des classes au sein d'une population au moyen d'un modèle, pour une variable discrète associée aux réponses dans le cadre d'une enquête par sondage. Notre méthode d'estimation généralisée par régression logistique (REGGL) repose sur un modèle logistique multinominal qui pourrait s'avérer plus réaliste à l'égard des indicateurs de classe que le modèle linéaire dont on se sert normalement pour l'estimation de régression généralisée (REGG). Les estimateurs REGGL et REGG aboutissent à des résultats identiques avec la stratification a posteriori complète, mais les résultats diffèrent avec d'autres modèles comme la procédure itérative. Comparativement à REGG, la méthode REGGL requiert habituellement plus d'information auxiliaire, et pas seulement les totaux auxiliaires. Quoi qu'il en soit, elle semble préférable à la méthode REGG quand la probabilité des classes varie considérablement avec la fourchette des variables auxiliaires continues, et lorsqu'on a besoin d'estimations pour de petits domaines, en particulier si la fréquence des classes est peu élevée.

## REMERCIEMENTS

Les auteurs remercient le professeur Carl-Erik Särndal, de l'Université de Montréal, pour ses commentaires sur la version antérieure de l'article. Les remarques détaillées d'un rédacteur associé et de deux examinateurs nous ont été d'une grande utilité. Nous nous devons aussi de remercier M. Timo Koskimäki, de Statistics Finland, pour nous avoir fourni les données de l'Enquête sur la population active, ainsi que M. Kari Djerf, pour ses commentaires judicieux.

## BIBLIOGRAPHIE

CHAMBERS, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

DJERF, K. (1997). Effects of post-stratification on the estimates of the Finnish Labour Force Survey. *Journal of Official Statistics*, 13, 29-39.

ESTEVAO, V., HIDIROGLOU, M.A., et SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

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.

LEHTONEN, R., et PAHKINEN, E.J. (1996). *Practical Methods for Design and Analysis of Complex Surveys*. Édition revisée. Chichester: John Wiley & Sons.

LEHTONEN, R., et VEIJANEN, A. (1998). On Multinomial Logistic Generalized Regression Estimators. Jyreäskylä: Preprints from the Department of Statistics, University of Jyreäskylä, 22.

McCULLAGH, P., et NELDER, J.A. (1989). *Generalized Linear Models*. Deuxième édition. London: Chapman and Hall.

NORDBERG, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics*, 5, 223-239.

SKINNER, C.J., HOLT, D., et SMITH, T.M.F. (Éds) (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.