Techniques d'enquête, décembre 2000
Vol. 26, N° 2, pp. 153-155
Statistique Canada, N° 12-001 au catalogue

Un effet de plan de sondage approximatif pour une pondération inégale en cas de corrélation possible entre les mesures et les probabilités de sélection

Bruce D. Spencer [1]

Résumé

On estime couramment l'effet de plan de sondage attribuable à la pondération par 1 plus la variance relative des poids de l'échantillon. Cette formule a été justifiée en l'absence de corrélation entre les probabilités de sélection et la variable d'intérêt. Une approximation de l'effet de plan de sondage est présentée pour tenir compte de la présence d'une corrélation.

Mots clés : Pondération; effet de plan de sondage; variance d'échantillonnage; échantillons complexes.

1. Introduction

Une pratique courante consiste à pondérer les observations dans un échantillon à probabilités inégales par les valeurs inverses des probabilités de sélection. On le fait parce que l'absence des poids entraîne une erreur systématique s'il y a une corrélation entre les poids d'échantillonnage et la variable d'intérêt. Un inconvénient du processus de pondération est l'augmentation de la variance d'échantillonnage si les poids varient de façon excessive dans l'échantillon. Cette augmentation peut être quantifiée à l'aide de l'effet de plan de sondage. Celui-ci est le rapport entre la variance de la statistique d'intérêt pour le plan de sondage en question et la variance de la statistique pour un échantillonnage aléatoire simple et la même taille d'échantillon (Kish 1965). Les effets de plan de sondage sont importants tant pour l'approximation des erreurs types une fois que l'échantillon est disponible que pour la prédiction préalable des erreurs types, étape critique d'un plan de sondage efficace.

Kish (1965, 1992) a décrit une approximation de l'effet de plan de sondage pour des estimations pondérées tirées d'échantillons à probabilités inégales : $1 + rvw$, où rvw est défini comme la variance relative des poids de l'échantillon. Si donc w_i est le poids de l'unité i de l'échantillon, et si $\overline{w}$ est la moyenne de l'échantillon, $rvw = n^{-1}\sum_{i=1}^{n} (w_i - \overline{w})^2/\overline{w}^2$. Gabler, Haeder et Lahiri (1999) ont utilisé un modèle de superpopulation pour obtenir un effet de plan de sondage en présence d'une mise en grappes également. Leur formule, qui correspond à des résultats fondés sur un plan de sondage figurant dans Kish (1965), se réduit à $1 + rvw$ en présence d'une corrélation intraclasse nulle. L'approximation $1+rvw$ pour l'effet de plan de sondage se fonde sur un modèle ou un plan dans lequel il n'y a pas de corrélation entre les poids et la variable d'intérêt (de sorte qu'une estimation non pondérée serait tout aussi bonne ou meilleure que l'estimation pondérée). Nous présentons ici une approximation de l'effet de plan de sondage pour un modèle dans lequel il peut y avoir une corrélation. Ce faisant, nous ne supposons pas que la population est tirée d'une superpopulation. L'exactitude de l'approximation dépend uniquement des caractéristiques du plan de sondage et de la population d'intérêt.

Par souci de simplicité, nous avons recours à l'échantillonnage à un degré à probabilités inégales avec remise. L'élargissement heuristique des résultats à l'échantillonnage sans remise est abordé à la section 4.

2. Représentation de régression du plan de sondage et de la population

Soit y_i, la mesure d'intérêt, P_i, la probabilité de sélection (d'un tirage à l'autre) pour un échantillon de taille n, et $w_i = 1/(nP_i)$, le poids d'échantillonnage pour l'unité i d'une population de taille N, $1 \le i \le N$. À noter que $\overline{P} = \sum_{i=1}^{N} P_i / N = N^{-1}$. Considérons la droite de régression des moindres carrés

$$y_i = \alpha + \beta P_i + \varepsilon_i, \qquad (1)$$

où $\alpha = \overline{Y} - \beta/N$, $\beta = \sum_{i=1}^{N} (y_i - \overline{Y})(P_i - \overline{P}) / \sum_{i=1}^{N} (P_i - \overline{P})^2$, et $\overline{Y} = \sum_{i=1}^{N} y_i / N$. Nous notons les variances de population des y, des ε, des ε^2 et des w par $\sigma_y^2, \sigma_\varepsilon^2, \sigma_{\varepsilon^2}^2$, et σ_w^2, avec, par exemple, $\sigma_y^2 = \sum_{i=1}^{N} (y_i - \overline{Y})^2 / N$. Nous notons la corrélation de population entre y et P par $\rho_{y,P}$, entre ε et w par $\rho_{\varepsilon,w}$, et entre ε^2 et w par $\rho_{\varepsilon^2,w}$. Il s'ensuit des proprétés des moindres carrés, ou de façon équivalente des définitions de α et de β, que $\sum_{i=1}^{N} \varepsilon_i P_i = \sum_{i=1}^{N} \varepsilon_i / N = 0$ et que $\sigma_\varepsilon^2 = (1 - \rho_{y,P}^2)\sigma_y^2$. Si des données sont disponibles, nous pouvons ajuster la représentation de régression (1) et estimer α, β, σ, et ρ, par exemple, par $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$ et $\hat{\rho}$.

Soit $\hat{Y} = \sum_{i=1}^{n} w_i y_i$, l'estimateur pondéré habituel du total de la population, Y. La variance de $\hat{Y}$ est bien connue (Cochran 1977, 253) :

1. Bruce D. Spencer, Department of Statistics and Institute for Policy Research, 2006 Sheridan Road, Northwestern University, Evanston, IL 60208, États-Unis.

$$V(\hat{Y}) = n^{-1}\sum_{i=1}^{N} P_i(y_i/P_i - Y)^2. \qquad (2)$$

À l'aide de la formulation de régression (1), nous pouvons réexprimer la variance comme suit

$$V(\hat{Y}) = \alpha^2 N(\bar{W} - N/n) + (1-\rho_{y,P}^2)\sigma_y^2 N\bar{W} + N\rho_{\varepsilon^2,w}\sigma_{\varepsilon^2}\sigma_w + 2\alpha N\rho_{\varepsilon,w}\sigma_\varepsilon\sigma_w, \qquad (3)$$

où $\bar{W} = \sum_{i=1}^{N} w_i / N$.

Cette expression ne se fonde sur aucune hypothèse au sujet de l'ajustement du modèle de régression (voir le calcul à la section 5).

Si le modèle de régression est assez bien ajusté pour que $\rho_{\varepsilon^2,w}$ et $\rho_{\varepsilon,w}$ soient nuls, la variance en (3) est simplifiée de façon à donner $V(\hat{Y}) = \alpha^2 N(\bar{W} - N/n) + (1-\rho_{y,P}^2)\sigma_y^2 N\bar{W}$. Si l'on avait eu recours à un échantillonnage aléatoire simple avec remise, la variance aurait été $n^{-1}N^2\sigma_y^2$. Par conséquent, si $\rho_{\varepsilon^2,w}$ et $\rho_{\varepsilon,w}$ sont négligeables, l'effet de plan de sondage est approximativement :

$$\text{deff} = (1-\rho_{y,P}^2)n\bar{W}/N + (\alpha/\sigma_y)^2(n\bar{W}/N - 1). \qquad (4)$$

Cette approximation n'exige pas que les résidus de la régression soient négligeables, et elle reste valable lorsque σ_ε est grand. Un examinateur a fait remarquer que la condition voulant que $\rho_{\varepsilon^2,w}$ et $\rho_{\varepsilon,w}$ soient négligeables peut sembler anormale dans un modèle qui régresse y sur P plutôt que sur $w \propto 1/P$. À noter, toutefois, qu'en présence non seulement d'une corrélation nulle entre ε et P mais également d'indépendance, nous aurions une corrélation nulle entre les fonctions de ε et les fonctions de P, de sorte que $\rho_{\varepsilon^2,w}$ et $\rho_{\varepsilon,w}$ seraient nuls eux aussi.

3. Estimation de l'effet de plan de sondage

Afin d'estimer l'effet de plan de sondage lorsque l'échantillon est disponible, nous pouvons utiliser $1 + rvw$ pour estimer $n\bar{W}/N$. Afin d'en comprendre le bien-fondé, notons d'abord que

$$1 + rvw = \frac{n^{-1}\sum_{i=1}^{n} w_i^2}{\bar{w}^2}. \qquad (5)$$

L'espérance du numérateur est $N\bar{W}/n$. L'espérance de $\bar{w}$ est N/n, et donc le dénominateur de (5) peut être considéré comme un estimateur de $(N/n)^2$. Le fait de diviser l'espérance du numérateur par $(N/n)^2$ nous permet d'obtenir $n\bar{W}/N$. Ainsi, l'effet de plan de sondage peut être estimé à partir de l'échantillon comme suit :

$$(1-\hat{\rho}_{y,P}^2)(1 + rvw) + (\hat{\alpha}/\hat{\sigma}_y)^2(rvw). \qquad (6)$$

Notons, à titre de cas spécial, que si nous fixons $\hat{\rho}_{y,P} = 0$, qui est le cas de la « pondération au hasard » (Kish 1992), l'estimation de l'effet de plan de sondage se réduit à

$$1 + rvw + rvw(\hat{\alpha}^2/\hat{\sigma}_y^2). \qquad (7)$$

Cette estimation se rapproche de celle de Kish pour $\hat{\alpha}/\hat{\sigma}_y$ presque nul.

4. Échantillonnage sans remise

Le calcul de l'effet de plan de sondage exact pour un échantillonnage sans remise serait plus complexe, car il supposerait que l'on tienne compte des probabilités de sélection composées pour des paires d'unités. Un élargissement heuristique des résultats, par contre, est facile. Rappelons que le rapport entre la variance d'une moyenne de l'échantillon pour un échantillonnage aléatoire simple sans remise et la variance pour un échantillonnage avec remise est approximativement $(1 - n/N)$.

Les résultats que nous avons calculés pour l'effet de plan de sondage s'appliquent à des échantillons à un degré à probabilités inégales de n unités sans remise si la variance de l'estimateur de Horvitz-Thompson du total est approximativement $(1 - n/N)$ fois la variance en (2), P_i étant considéré comme n^{-1} fois la probabilité de sélection globale pour l'unité i (Särndal, Swensson et Wretman 1992, 154).

5. Calcul de la formule de la variance (3)

À partir de (2) nous avons $V(\hat{Y}) = n^{-1}(\sum_{i=1}^{N} y_i^2/P_i - Y^2)$. Ensuite, nous notons que (1) suppose que

$$Y^2 = (N\alpha + \beta)^2 = N^2\alpha^2 + 2N\alpha\beta + \beta^2 \qquad (8)$$

et que

$$\begin{aligned}\sum_{i=1}^{N} y_i^2/P_i &= \sum_{i=1}^{N}[\alpha^2/P_i + \beta^2 P_i + \varepsilon_i^2/P_i + 2\alpha\beta + 2\alpha\varepsilon_i/P_i + 2\beta\varepsilon_i] \\ &= \alpha^2\sum_{i=1}^{N} P_i^{-1} + \beta^2 + \sum_{i=1}^{N}\varepsilon_i^2/P_i + 2N\alpha\beta \\ &\quad + 2\alpha\sum_{i=1}^{N}\varepsilon_i/P_i \\ &= \alpha^2 n\sum_{i=1}^{N} w_i + \beta^2 + n\sum_{i=1}^{N}\varepsilon_i^2 w_i + 2N\alpha\beta \\ &\quad + 2\alpha n\sum_{i=1}^{N}\varepsilon_i w_i. \qquad (9)\end{aligned}$$

Le fait de soustraire (8) de (9) et de diviser par n donne

$$V(\hat{Y}) = \alpha^2\left(\sum_{i=1}^{N} w_i - N^2/n\right) + \sum_{i=1}^{N}\varepsilon_i^2 w_i + 2\alpha\sum_{i=1}^{N}\varepsilon_i w_i.$$

Afin d'obtenir (3), nous notons que

$$\sum_{i=1}^{N} \varepsilon_i^2 w_i = N\rho_{\varepsilon^2, w}\sigma_{\varepsilon^2}\sigma_w + N\overline{W}\sigma_\varepsilon^2$$

$$= N\rho_{\varepsilon^2, w}\sigma_{\varepsilon^2}\sigma_w + (1 - \rho_{y, P}^2)\sigma_y^2 N\overline{W}$$

et

$$\sum_{i=1}^{N} \varepsilon_i w_i = N\rho_{\varepsilon, w}\sigma_\varepsilon\sigma_w.$$

Bibliographie

Cochran, W.G. (1977). *Sampling Techniques*. 3ième Éd. New York : John Wiley & Sons, Inc.

Gabler, S., Haeder, S. et Lahiri, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 1, 119-120.

Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8, 2, 183-200.

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