Une interprétation de la pseudo-vraisemblance

Monique Graf

Institut de statistique, Université de Neuchâtel, 2000 Neuchâtel, Suisse
Elpacos statistique, 2520 La Neuveville, Suisse
monique.p.n.graf@bluewin.ch

Résumé. Considérons un modèle statistique de super-population dans lequel une variable d'intérêt connue sur une population de taille N est considérée comme un ensemble de N réalisations aléatoires indépendantes du modèle. La log-vraisemblance au niveau de la population s'écrit alors comme une somme. Si on ne dispose que d'un échantillon, tiré selon un plan de sondage à probabilités inégales, la log-pseudo-vraisemblance est l'estimateur de Horvitz-Thompson de la log-vraisemblance de la population. En général, les poids sont multipliés par un facteur de normalisation, de telle sorte qu'ils somment à la taille de l'échantillon. Dans le cas d'un seul niveau, cela ne change pas la valeur des paramètres estimés. Le problème du choix des facteur de normalisation dans les plans en grappes a été abondamment traité dans la littérature, sans aboutir à des directives claires. On propose de calculer ces facteurs de telle sorte que la pseudo-vraisemblance soit une vraisemblance au sens propre en tenant compte du caractère aléatoire des poids.

Mots-clés. Analyse de données d'enquêtes, modèle multi-niveaux, modèle mixte généralisé, pondération.

Abstract. Consider a super-population statistical model, where a variable of interest, known on a finite poppulation of size N is considered a set of N independent realizations of the model. The log-likelihood at the population level is then written as a sum. If only a sample is observed, drawn according to a design with unequal inclusion probabilities, the log-pseudo-likelihood is the Horvitz-Thompson estimate of the population log-likelihood. In general, the extrapolation weights are multiplied by a normalization factor, in such a way that normalized weights sum to the sample size. If the design just has one level, the value of estimated model parameters is unchanged. The problem of the choice of the normalization factors in cluster sampling has been largely addressed in the literature, without clear recommendations having been issued. It is proposed here to compute these factors in such a way that the pseudo-likelihood becomes a proper likelihood taking the random nature of the weights into account.

Keywords. Analysis of survey data, multi-level model, generalized mixed model, weighting.

1 Introduction

Les modèles multi-niveaux sont un cas particulier du modèle mixte généralisé. Ils sont utilisés pour l'analyse de données d'enquêtes où plusieurs niveaux (strates, grappes,unités) sont présents. Ils trouvent leur origine dans les travaux de Binder (1983), Gourieroux et al. (1984), Skinner et al. (1989), Pfeffermann et al. (1998). Ces auteurs introduisent le concept de pseudo-vraisemblance pour l'estimation des paramètres, applicable au cas d'une enquête à probabilités inégales. Alors que pour une enquête à un niveau unique, les poids d'extrapolation peuvent être multipliés de manière arbitraire par une constante (normalisés) sans changer les estimateurs de pseudo-vraisemblance, dans le cas multi-niveau, tout facteur de normalisation aux niveaux inférieurs a un impact. La question du choix du facteur de normalisation prend donc toute son importance. Pour une discussion sur ces facteurs, voir Rabe-Hesketh et Skrondal (2006) et Asparouhov (2006). Ces auteurs considèrent des normalisations qui tiennent compte de la taille ou de la taille effective de l'échantillon ou des grappes, alors que Kovačević et Rai (2003) considèrent l'estimation de la vraisemblance de la population finie. Guadarrama et al. (2018) par exemple utilisent la taille effective dans le cadre d'estimation par petits domaines. La méthode (avec une normalisation à choisir par l'utilisateur) est applicable avec la procédure SAS GLIMMIX.

En général les estimateurs de variance du maximum de vraisemblance sont biaisés. Les estimateurs basés sur la pseudo-vraisemblance n'échappent pas à ce travers. Pfeffermann et al. (1998) discutent de la convergence des estimateurs dans le cas à deux niveaux. Alors que la convergence des estimateurs des paramètres de régression ne dépend que du nombre de grappes, la convergence des estimateurs de variance intra et inter grappes demandent encore que la taille des grappes elle aussi tende vers l'infini. Rao et al. (2013) proposent une méthode d'estimation par des fonctions estimantes et montrent sur un exemple sa supériorité sur la pseudo-vraisemblance avec deux choix de facteurs étudiés par Asparouhov (2006). Ils obtiennent également une vraisemblance composite qui est plus simple que la vraisembalnce d'origine, mais donne les mêmes estimateurs des paramètres, voir aussi Varin et al. (2011). D'autres approches existent. Pfeffermann (2011) résume, dans le contexte d'un plan de sondage non-ignorable, l'approche par la modélisation des poids d'extrapolation à l'aide des variables disponibles. La loi conditionnelle des poids ainsi obtenue modifie la densité postulée sur la population pour obtenir une densité de l'échantillon tenant compte du plan. Bonnéry et al. (2018) établissent les propriétés asymptotiques de la vraisemblance obtenue avec cette densité.

Remarquons que dans les travaux sur la densité d'échantillon, le fait que les poids soient approchés par un modèle n'est pas pris en compte. Or il est probable que les analystes n'aient pas à disposition toute l'information pour reconstituer les poids. De plus, il peut être lourd pour l'analyste de devoir modéliser les poids avant d'entamer ses recherches.

On reprend donc ici la méthode de la pseudo-vraisemblance. Il n'y a pas de justification théorique péremptoire pour le choix d'une normalisation des poids. Les simulations publiées ne donnent pas d'indice clair sur la meilleure façon de normaliser, voir Pfeffer-

mann et al. (1998), Asparouhov (2006) ; Rabe-Hesketh et Skrondal (2006), Lucas et al. (2014). Le but est ici d'obtenir à partir de la pseudo-vraisemblance, une vraisemblance au sens propre appartenant à la même famille de distributions que la loi postulée sur la population. Cet article présente une méthode pour choisir rationnellement la normalisation, de telle sorte que la pseudo-vraisemblance soit une vraisemblance à part entière. Cette normalisation tient compte du fait que, selon le plan, les poids d'échantillonnage varient d'un tirage à l'autre, même si la taille de l'échantillon est fixe.

2 Motivation

Considérons un seul niveau. Dans ce cas, la normalisation n'a pas d'effet sur les estimateurs, mais le principe des calculs peut s'expliquer simplement. On suppose que la variable d'intérêt sur la population est un ensemble de N réalisations $y_i, i = 1, ..., N$ de $Y_i, i = 1, ..., N$ qu'on suppose i.i.d. $N(\mu, \sigma^2)$. La log-vraisemblance est donnée par

$$\ell(\mu, \sigma) = \sum_{i=1}^{N} \log \left(\frac{1}{\sigma\sqrt{2\pi}} \exp \left(-\frac{1}{2} \frac{(y_i - \mu)^2}{\sigma^2} \right) \right)$$

On n'observe qu'un échantillon de taille n. Soit $w_i, i = 1, ..., n$ les poids d'enquête. La log-pseudo-vraisemblance,

$$\ell_w(\mu, \sigma) = \sum_{i=1}^{n} w_i \log \left(\frac{1}{\sigma\sqrt{2\pi}} \exp \left(-\frac{1}{2} \frac{(y_i - \mu)^2}{\sigma^2} \right) \right) \quad (1)$$

est l'estimateur de Horvitz-Thompson de $\ell(\mu, \sigma)$. On peut interpréter la log-pseudo-vraisemblance en exprimant les termes de la somme dans l'expression (1) comme une somme de log-densités ayant des paramètres dépendant de i,

$$\begin{aligned} \ell_w(\mu, \sigma) &= \sum_{i=1}^{n} \log \left(\frac{1}{(\sigma/\sqrt{w_i})\sqrt{2\pi}} \exp \left(-\frac{1}{2} \frac{(y_i - \mu)^2}{(\sigma/\sqrt{w_i})^2} \right) \right) + \\ &+ \sum_{i=1}^{n} (w_i - 1) \log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2} \log \left(\prod_{i=1}^{n} w_i \right). \end{aligned} \quad (2)$$

La première somme dans l'expression (2) est la log-vraisemblance d'observations indépendantes hétéroscedastiques de loi $N(\mu, \sigma^2/w_i)$. Notons cette log-vraisemblance $\sum \ell_i$. Si les poids w_i ont été normalisés de telle sorte qu'ils somment à n,

$$\sum_{i=1}^{n} (w_i - 1) = 0, \quad (3)$$

et la deuxième somme est nulle. Conditionnellement aux poids, la solution des équations de vraisemblance basées sur $\sum \ell_i$ est la même que celle de de la log-pseudo-vraisemblance à l'expression (1). Avec $\sum w_i = n$, on peut donc considérer la loi $N(\mu, \sigma^2/w_i)$ comme la loi de Y_i, tenant compte du plan d'échantillonnage, conditionnellement aux poids. Cependant, une autre solution existe qui tient compte de la variabilité des poids d'un tirage de l'échantillon à l'autre. Soit x un facteur de normalisation multipliant les poids (déjà normalisés de telle sorte qu'ils somment à n). Nous cherchons x tel que

$$\begin{aligned}
\ell_{xw}(\mu, \sigma) &= \sum_{i=1}^{n} \log\left(\frac{1}{(\sigma/\sqrt{xw_i})\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y_i-\mu)^2}{(\sigma/\sqrt{xw_i})^2}\right)\right) + C(x), \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{(\sigma_x/\sqrt{w_i})\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y_i-\mu)^2}{(\sigma_x/\sqrt{w_i})^2}\right)\right) + C(x) \qquad (4)
\end{aligned}$$

où $\sigma_x^2 = \sigma^2/x$, et

$$\begin{aligned}
C(x) &= \sum_{i=1}^{n}(xw_i - 1)\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\log\left(\prod_{i=1}^{n}(xw_i)\right) \\
&= -\frac{n}{2}\left[(x-1)\log\left(2\pi\sigma^2\right) + \log(x) + \log(G)\right] \quad \text{par l'équation (3)} \\
&= -\frac{n}{2}\left[x\log(x) + (x-1)\log\left(2\pi\sigma_x^2\right) + \log(G)\right], \qquad (5)
\end{aligned}$$

avec $G = (\prod_{i=1}^{n} w_i)^{1/n,}$, la moyenne géométrique des poids.

Ainsi nous pouvons chercher le maximum de la log-vraisemblance d'observations hétéroscédastiques d'espérance μ et de variance (σ_x^2/w_i) sous la contrainte $C(x) = 0$. Dans ce cas, l'estimation des paramètres μ et σ_x^2 est indépendante de celle de x et peut s'obtenir en optimisant la log-vraisemblance (1er terme de l'expression 4), puis x peut être trouvé comme solution de $C(x) = 0$ avec σ_x donné. On démontre que

1. Le maximum de $C(x)$ se situe en $x_{max} = \exp[-(1 + \log(2\pi\sigma_x^2)/2] = \exp[-h/2]$ où h est l'entropie de la loi $N(\mu, \sigma_x^2)$.
2. $C(x) = 0$ a au plus deux solutions et une seule si $C(0)$ est du même signe que $\max(C(x))$.

Dans les applications, nous choisissons toujours la solution la plus proche de 1. Soit x_0 cette solution. Finalement la variance estimée est donnée par

$$\hat{\sigma}^2 = x_0\, \hat{\sigma}_x^2,$$

où $\hat{\sigma}_x^2$ peut être calculé par un estimateur sandwich, c'est-à-dire en utilisant la variance sous le plan des scores (dérivées partielles de la log-vraisemblance par rapport aux paramètres) et la linéarisation de l'estimateur de la variance.

Plus généralement, considérons un modèle univarié représenté par la densité $f(y, \theta)$, où θ est un vecteur de paramètres. Les observations sont supposées i.i.d. de loi f. Soit D le domaine de définition de f. Soit w_i le poids d'extrapolation de l'observation i. La log-pseudo-vraisemblance est donnée par

$$\ell(\theta|y_1, ..., y_n) = \sum_{i=1}^{n} w_i \log(f(y_i, \theta) = \sum_{i=1}^{n} \log(f(y_i, \theta)^{w_i}). \tag{6}$$

Pour transformer la fonction $f(y, \theta)^{w_i}$ en une densité, il faut corriger la constante d'intégration. Posons $C_i = \int_D f(y, \theta)^{w_i} dy$.

$$\ell(\theta|y_1, ..., y_n) = \sum_{i=1}^{n} \log\left(\frac{1}{C_i} f(y_i, \theta)^{w_i}\right) + \sum_{i=1}^{n} \log(C_i), \tag{7}$$

où $f_i(., \theta) = \frac{1}{C_i} f(., \theta)^{w_i}$ est une densité. Si les poids d'extrapolation sont tels que $\sum_i \log(C_i) = 0$, la solution des équations de pseudo-vraisemblance est la même que celle des équations de vraisemblance basées sur les densités f_i. Celles-ci peuvent donc être vues comme les densités tenant compte de l'échantillonnage. Par exemple,

1. Soit des observations i.i.d. suivant une loi exponentielle de densité $\frac{1}{\lambda} \exp(-\frac{y}{\lambda})$. On démontre que $C(x)$ vaut

   $$C(x) = -n \left[x \log(x) + (x - 1) \log(\lambda_x) + \log(G)\right],$$

   où $\lambda_x = \lambda/x$ et G est la moyenne géométrique des poids normalisés de telle sorte qu'ils somment à n, la taille de l'échantillon. On remarque la très grande analogie avec le cas normal (équation 5).
   Le maximum de C se situe en $x = \exp[-(1 + \log(\lambda_x))] = \exp(-h)$, où h est l'entropie de la loi exponentielle de paramètre λ_x.
2. Il n'est pas toujours possible de séparer l'estimation des paramètres de celle du facteur x. C'est le cas par exemple pour une loi gamma. On pourrait alors traiter le problème avec des multiplicateurs de Lagrange ou par la méthode des fonctions implicites.
3. On généralise facilement le cas normal à une enquête multi-niveaux dans laquelle on suppose que les niveaux correspondent aux types d'unités (strates, unités primaires, unités secondaires).

On montrera quelques exemples.

Bibliographie

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods*, 35(3) :439–460.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51 :279–292.

Bonnéry, D., Breidt, F. J., et Coquet, F. (2018). Asymptotics for the maximum sample likelihood estimator under informative selection from a finite population. *Bernoulli*, 24(2) :929–955.

Gourieroux, C., Monfort, A., et Trognon, A. (1984). Pseudo maximum likelihood methods : theory. *Econometrica*, 52 :681–700.

Guadarrama, M., Molina, I., et Rao, J. (2018). Small area estimation of general parameters under complex sampling designs. *Computational Statistics and Data Analysis*, 121 :20–40.

Kovačević, M. S. et Rai, S. N. (2003). A pseudo maximum likelihood approach to multilevel modelling of survey data. *Communications in Statistics - Theory and Methods*, 32(1) :103–121.

Lucas, J.-P., Sébille, V., Tertre, A. L., Strat, Y. L., et Bellanger, L. (2014). Multilevel modelling of survey data : impact of the two-level weights used in the pseudolikelihood. *Journal of Applied Statistics*, 41(4) :716–732.

Pfeffermann, D. (2011). Modélisation des données d'enquêtes complexes : Pourquoi les modéliser ? pourquoi est-ce un problème ? comment le résoudre ? *Techniques d'enquête*, 37(2) :123–146.

Pfeffermann, D., Skinner, C., Holmes, D. J., Goldstein, H., et Rasbash, J. (1998). Weighting for unequal selection probabilities in multlevel models. *J. R. Statist. Soc. B*, 60(1) :23–40.

Rabe-Hesketh, S. et Skrondal, A. (2006). Multilevel modelling of complex survey data. *J. R. Statist. Soc. A*, 169(4) :805–827.

Rao, J., Verret, F., et Hidiroglou, M. A. (2013). Une approche d'inférence fondée sur la vraisemblance composite pondérée pour des modèles à deux niveaux issus de données d'enquête. *Techniques d'enquête*, 39(2) :291–311.

Skinner, C. J., Holt, D., et Smith, T. M. F. (1989). *Analysis of Complex Surveys*. Wiley, New York.

Varin, C., Reid, N., et Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21 :5–42.