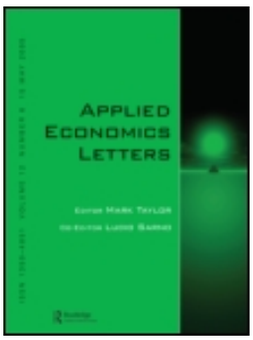

Routledge
Taylor & Francis Group

Applied Economics Letters

ISSN: 1350-4851 (Print) 1466-4291 (Online) Journal homepage: http://www.tandfonline.com/loi/rael20

On the weighted maximum likelihood estimator for endogenous stratified samples when the population strata probabilities are unknown

Esmeralda A. Ramalho & Joaquim J. S. Ramalho

To cite this article: Esmeralda A. Ramalho & Joaquim J. S. Ramalho (2007) On the weighted maximum likelihood estimator for endogenous stratified samples when the population strata probabilities are unknown, Applied Economics Letters, 14:3, 171-174, DOI: 10.1080/13504850500426194

To link to this article: https://doi.org/10.1080/13504850500426194

Applied Economics Letters, 2007, **14**, 171–174

Routledge
Taylor & Francis Group

On the weighted maximum likelihood estimator for endogenous stratified samples when the population strata probabilities are unknown

Esmeralda A. Ramalho and Joaquim J. S. Ramalho*

Departamento de Economia, Universidade de Evora and CEMAPRE, Portugal

The popular weighted maximum likelihood estimator for endogenous stratified samples requires knowledge on the population proportions of each stratum. In this paper we extend their estimator for cases where such information is not available.

I. Introduction

In many research settings, economists are often interested in estimating parametric models based on endogenously stratified samples (ESS), where the probability of being sampled depends on the value of the variable of interest. Although several estimators have been proposed to deal with ESS (see *inter alia* Cosslett, 1981; Manski and McFadden, 1981; Imbens and Lancaster, 1996), certainly due to its simplicity, only Manski and Lerman's (1977) weighted maximum likelihood (WML) estimator has been widely used in empirical work (e.g. Artis *et al*., 1999; Early, 1999; Kitamura *et al*., 2003). However, this estimator may only be employed in cases where the proportions of each stratum in the population are known. In this article, we propose a simple extension of the WML estimator for cases where this information is not available, suggesting a generalized method of moments (GMM) estimator that uses as moment conditions the same weighted score functions that define the WML estimator and, in addition, a weighted version of the equations that define the nonparametric maximum likelihood (ML) estimator of the population strata probabilities under random sampling (RS). The weights used in both cases are the same.

The article is organized as follows. Section II briefly reviews the main characteristics of ESS. Section III describes the new weighted GMM estimator. Section IV investigates its performance in finite samples through a Monte-Carlo simulation study. Finally, Section V presents some concluding remarks.

II. Endogenous Stratified Samples

Consider a sample of $i = 1, \ldots, N$ individuals and let Y be the variable of interest, continuous or discrete and X a vector of k exogenous variables. Both Y and X are random variables defined on $\mathcal{Y} \times \mathcal{X}$ with population joint density function

$$f(y, x; \theta) = f(y|x, \theta) f(x) \tag{1}$$

where the conditional density function $f(y|x, \theta)$ is known up to the parameter vector θ and the marginal density function $f(x)$ is unknown. Our interest is

*Corresponding author. E-mail: ela@uevora.pt

estimation of and inference on the parameter vector θ in $f(y|x,\theta)$.

Assume that the population of interest is divided into J nonempty and possibly overlapping strata. Each stratum is designated as $C_s = \mathcal{Y}_s \times \mathcal{X}$, with $s \in \mathcal{S}$, $\mathcal{S} = \{1, \ldots, J\}$ and $\mathcal{Y}_s$ defined as the subset of $\mathcal{Y}$ for which the observation y lies in $\mathcal{C}_s$. The proportion of the stratum s in the population is given by

$$Q_s = \int_{\mathcal{X}} \int_{\mathcal{Y}_s} f(y, x; \theta) \mathrm{d}y \, \mathrm{d}x \tag{2}$$

where $Q_s(\theta) > 0$.

One of the mechanisms which may be employed for drawing an ESS is the so-called multinomial sampling. In this sampling scheme, considered, for example, by Manski and Lerman (1977) and Manski and McFadden (1981), the stratum indicators S are drawn independently from a multinomial distribution. The sampling agent randomly selects a stratum $\mathcal{C}_s$ with a pre-defined probability H_s, where $H_s > 0$ and $\sum_{s\in\mathcal{S}} H_s = 1$ and, then, randomly samples from that stratum. In this setting, the variable of interest, the covariates and the stratum indicator are observed according to

$$h(y, x, s) = \frac{H_s}{Q_s} f(y|x,\theta) f(x) \tag{3}$$

III. Weighted GMM Estimation

Manski and Lerman's (1977) WML estimator results from solving the just-identified system of k equations

$$g(\theta) = \frac{Q_s}{H_s} g(\theta)_{RS} \tag{4}$$

where $g(\theta)_{RS} \equiv \nabla_\theta \ln f(y|x,\theta)$ are the score functions that define the random sampling maximum liklihood (RSML) estimator for θ. Note that under ESS, as the data are observed according to $h(y,x,s)$ of (3), while $E[g(\theta)_{RS}] \neq 0$ in general, $E[g(\theta)] = 0$ since the weight factor (Q_s/H_s) promotes the reconstruction of the population structure. As it is clear from (4), application of the WML estimator presumes knowledge of Q_s.

In this article, we assume that there is no information available on Q_s and propose an extension of the WML estimator for such case. Namely, we propose a weighted GMM (WGMM) that treats Q_s as unknowns to be estimated simultaneously with the parameters of interest θ. Let $Q' = (Q_1, Q_2, \ldots, Q_J)'$ be a J-dimensional vector containing the population strata probabilities and $\beta' = (\theta', Q')$ be the $(k+J)$-dimensional vector of parameters to be estimated. A set of at least $(k+J)$ moment indicators is required to estimate β. In addition to the k functions (4) that define the WML estimator, where Q_s is no longer replaced by its true value, we suggest the J moment indicators

$$g(Q) = \frac{Q_s}{H_s} g(Q)_{RS} \tag{5}$$

where $g(Q_t)_{RS} \equiv Q_t - \int_{\mathcal{Y}_t} f(y|x,\theta)\mathrm{d}y$ are the estimating functions that define the RSML estimators $\hat{Q}_t = N^{-1} \sum_{i=1}^{N} \int_{\mathcal{Y}_t} f(y_i|x_i,\hat{\theta})\mathrm{d}y$, $t = 1, \ldots, J$, which use the empirical distribution function as the nonparametric ML estimator of $f(x)$, see (2). Again, it is straightforward to show that, under ESS, while $E[g(Q)_{RS}] \neq 0$ in general, $E[g(Q)] = 0$:

$$\begin{aligned}
E[g(Q_t)] &= \sum_{s\in\mathcal{S}} \int_{\mathcal{X}} \int_{Y_s} \frac{Q_s}{H_s} g(Q_t)_{RS} \frac{H_s}{Q_s} f(y|x,\theta) f(x) \mathrm{d}y \, \mathrm{d}x \\
&= \sum_{s\in\mathcal{S}} \int_{Y_s} \int_{Y_s} \frac{Q_s}{H_s} Q_t \frac{H_s}{Q_s} f(y|x,\theta) f(x) \mathrm{d}y \, \mathrm{d}x \\
&\quad - \sum_{s\in\mathcal{S}} \int_{\mathcal{Y}_s} \int_{Y_s} \int_{Y_t} \frac{Q_s}{H_s} f(y|x,\theta)\mathrm{d}y \frac{H_s}{Q_s} f(y|x,\theta) f(x) \mathrm{d}y \, \mathrm{d}x \\
&= Q_t \sum_{s\in\mathcal{S}} \int_{Y_t} \int_{\mathcal{Y}_t} f(y|x,\theta) f(x) \mathrm{d}y \, \mathrm{d}x \\
&\quad - \int_{\mathcal{Y}_t} \int_{\mathcal{Y}_t} f(y|x,\theta) f(x) \mathrm{d}x \, \mathrm{d}y \sum_{s\in\mathcal{S}} \int_{\mathcal{Y}_s} f(y|x,\theta)\mathrm{d}y \\
&= Q_t - Q_t = 0, \quad t = 1, \ldots, J
\end{aligned}$$

Let $g(y,x,s,\beta)' \equiv [g(\theta)', \; g(Q)']$, $\hat{g}(\beta) \equiv \sum_{i=1}^{N} g(y_i, x_i, s_i, \beta)/N$ and $\hat{W}$ denote a symmetric, positive definite matrix that converges almost surely to a nonrandom, positive definite matrix W. As the number of moment conditions and unknown parameters is identical, an efficient GMM estimator is defined by

$$\hat{\beta} = \arg\min_{\beta\in\mathcal{B}} \hat{g}(\beta)' \hat{W}^{-1} \hat{g}(\beta) \tag{6}$$

where $\mathcal{B}$ denotes the parameter space. Under suitable regularity conditions, see Newey and McFadden (1994), we have

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}\left[0, (G'\Omega^{-1}G)^{-1}\right] \tag{7}$$

where $G \equiv E[\nabla_\beta g(y,x,s,\beta)]$, $\Omega \equiv E[g(y,x,s,\beta)g(y,x,s,\beta)']$ and $\xrightarrow{d}$ denotes convergence in distribution.

IV. Monte-Carlo Simulation Study

To investigate the performance of the WGMM estimator in finite samples, we carried out a small

Monte-Carlo analysis of some particular examples based on Cosslett's (1981) design. In all experiments, we deal with binary data with two strata: stratum 0, containing individuals who respond $Y=0$ and stratum 1, containing individuals who respond $Y=1$. We consider logit models, where $\Pr(Y=1|x,\theta)=1/(1+\exp(-x\theta))$, with x generated according to the normal distribution $\mathcal{N}(2,0.5)$ and θ fixed as -1.81044, -1.24301, -0.73171, -0.43284, -0.20376, or 0, giving rise to six different values of Q_1: 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5. In all cases, the sampling proportion of both strata is the same: $H_1=H_0=0.5$. We computed three estimators: the RSML estimator, which, since in our design there is no intercept, is consistent only for $Q_1=0.5$; the Manski and Lerman's (1977) WML estimator, which assumes Q_1 known; and the WGMM estimator derived in this article, which estimates Q_1 simultaneously with θ. All experiments were based on 5000 replications, which were computed using the package S-Plus.

Table 1 reports the median bias and the standard deviation (SD) across replications for each estimator. Apart from the case where the proportions of the strata coincide in the population and in the sample, where it is approximately unbiased, the RSML estimator is clearly upwardly biased in all the other experiments, displaying biases ranging from 85.8 to 88.7%.

In contrast, both the weighted estimators are either unbiased or, in the case of the WGMM estimator, exhibit only small distortions for the smallest values of Q_1. The WGMM estimator for Q_1 is also approximately unbiased. However, the gain in precision that results from knowledge on Q_1 is enormous. Thus, the usefulness of the WGMM estimator is confined to cases where the population strata probabilities are unknown.

Table 1. Monte-Carlo results

	θ		Q_1	
	Bias	SD	Bias	SD
$Q_1=0.05$				
RSML	1.554	0.031	–	–
WML	0.000	0.056	–	–
WGMM	−0.043	0.471	−0.003	0.027
$Q_1=0.1$				
RSML	1.084	0.027	–	–
WML	0.000	0.035	–	–
WGMM	−0.020	0.309	−0.003	0.043
$Q_1=0.2$				
RSML	0.645	0.024	–	–
WML	0.000	0.025	–	–
WGMM	−0.013	0.238	−0.004	0.067
$Q_1=0.3$				
RSML	0.384	0.023	–	–
WML	0.000	0.023	–	–
WGMM	−0.005	0.220	−0.002	0.087
$Q_1=0.4$				
RSML	0.180	0.023	–	–
WML	−0.001	0.022	–	–
WGMM	−0.007	0.208	−0.003	0.096
$Q_1=0.5$				
RSML	0.000	0.022	–	–
WML	0.000	0.022	–	–
WGMM	0.001	0.204	0.000	0.098

Notes: RSML random sampling maximum likelihood estimator; WML weighted maximum likelihood estimator; WGMM weighted GMM estimator.

V. Concluding Remarks

In this article, we have proposed a WGMM estimator to deal with ESS when the marginal strata probabilities in the population are unknown. The use of this new estimator in empirical work is straightforward, since it merely requires that a new set of estimating equations, which employs the same weights as those used by the popular Manski and Lerman's (1977) WML estimator, is considered for the estimation of the proportion of the strata in the population of interest. The WGMM estimator may also be used in cases where the marginal strata probabilities are known. In such cases, the same moment indicators given in (4) and (5) may be used but with the vector Q replaced by its known value. As more information is included in the estimation procedure, the resultant estimator is more efficient than the WML estimator.

Acknowledgements

The authors gratefully acknowledge partial financial support from Fundação para a Ciência e Tecnologia, program POCTI, partially funded by FEDER.

References

Artis, M., Ayuso, M. and Guillén, M. (1999) Modelling different types of automobile insurance fraud behaviour in the spanish market, *Insurance: Mathematics and Economics*, **24**, 67–81.

Cosslett, S. (1981) Efficient estimation of discrete-choice models, in *Structural Analysis of Discrete Data with Econometric Applications* (Eds) C. Manski and D. McFadden, The MIT Press, Cambridge, pp. 51–111.

Early, D. W. (1999) A microeconomic analysis of homelessness: an empirical investigation using choice-based sampling, *Journal of Housing Economics*, **8**, 312–27.

Imbens, G. W. and Lancaster, T. (1996) Efficient estimation and stratified sampling, *Journal of Econometrics*, **74**, 289–318.

Kitamura, R., Yamamoto, T. and Sakai, H. (2003) A methodology for weighting observations from complex endogenous sampling, *Transportation Research Part B*, **37**, 387–401.

Manski, C. and Lerman, S. (1977) The estimation of choice probabilities from choice based samples, *Econometrica*, **45**, 1977–88.

Manski, C. and McFadden, D. (1981) Alternative estimators and sample designs for discrete choice analysis, in *Structural Analysis of Discrete Data with Econometric Applications* (Eds) C. Manski and D. McFadden, The MIT Press, Cambridge, pp. 2–50.

Newey, W. K. and McFadden, D. (1994) Large sample estimation and hypothesis testing, in *Handbook of Econometrics*, Vol. 4 (Eds) R. F. Engle and D. McFadden, Elsevier Science, Amsterdam, pp. 2111–245.