Available online at www.sciencedirect.com
ScienceDirect
Economics Letters 96 (2007) 70–76
economics letters
www.elsevier.com/locate/econbase

Binary models with misclassification in the variable of interest and nonignorable nonresponse[☆]

Esmeralda A. Ramalho[*]

Universidade de Évora, Portugal
CEMAPRE, Portugal

Received 26 January 2005; received in revised form 8 November 2006; accepted 12 December 2006
Available online 5 April 2007

Abstract

In this paper we propose a general framework to deal with datasets where a binary outcome is subject to misclassification and, for some sampling units, neither the error-prone variable of interest nor the covariates are recorded. A model to describe the observed data is formalized and efficient likelihood-based generalized method of moments estimators are suggested.

Keywords: Nonignorable nonresponse; Misclassification; Generalized method of moments estimation

JEL classification: C25; C51

1. Introduction

In this paper we propose a general framework to deal with datasets where a binary outcome is subject to misclassification and, for some sampling units, neither the error-prone variable of interest nor the covariates are recorded. We assume that misclassification is due to the nature of the variable of interest and, thus, may be

[☆] The author gratefully acknowledges partial financial support from Fundação para a Ciência e Tecnologia, program POCTI, partially funded by FEDER. Helpful comments and suggestions were given by Joaquim Ramalho and participants at the 59th Econometric Society European Meeting, Madrid.

[*] Departamento de Economia, Universidade de Évora, Largo dos Colegiais 2, 7000-803 Évora, Portugal. Tel.: +351 266740894; fax: +351 266740807.

E-mail address: ela@uevora.pt.

0165-1765/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2006.12.012

described by the conditional probability of the observable outcome given its true value. On the other hand, we consider that nonresponse depends on the error-prone alternative revealed and define a missing data mechanism in terms of the conditional probability of a response indicator given the error-prone outcome.

Similarly to Ramalho and Smith (2003), we reinterpret the missing data problem in discrete choice models by analogy with the choice-based (CB) sampling framework. Then, we extend this methodology to handle also misclassification in a similar way to that employed by Ramalho (2002) to adapt the estimators proposed by Imbens (1992) for CB samples. A model to describe the observed data is formalized and efficient likelihood-based generalized method of moments (GMM) estimators are suggested.

2. The model

Let $Y^*\in\mathcal{Y}^* = \{0,1\}$ be a binary response variable and X a vector of k exogenous variables defined on $\mathcal{X}$. Employing also the superscript "*" to denote the latent version of all probabilities and densities, the population joint density function of Y^* and X may be written as $f^*(y^*,x)=\Pr^*(y^*|x,\theta)f(x)$, where the marginal density function $f(x)$ for X is unknown and $\Pr^*(y^*|x,\theta)$ is known up to the parameter vector θ. Our interest is consistent estimation of and inference on the parameter vector θ. The marginal probability of observing an individual for which $Y^*=y^*$ in the population is $Q^*_{y^*} = \int_{\mathcal{X}} \Pr^*(y^*|x,\theta)f(x)\mathrm{d}x$, with $\sum_{y^*=0}^{1} Q^*_{y^*} = 1$.

In presence of misclassification, let $Y\in\mathcal{Y} = \{0,1\}$ represent the binary observable outcome. The error model is described by the conditional probability

$$\Pr(Y=y|Y^*=y^*,x) = \Pr(Y=y|Y^*=y^*) = \alpha_{yy^*}, \tag{1}$$

$0\leq\alpha_{yy^*}\leq 1$ and $\sum_{y=0}^{1}\alpha_{yy^*} = 1$. Hence, the conditional probability of the observable variable Y given X and the marginal probability of Y are, respectively, $\Pr(y|x,\theta,\alpha) = \sum_{y^*=0}^{1}\alpha_{yy^*}\Pr^*(y^*|x,\theta)$ and $Q_y = \sum_{y^*=0}^{1}\alpha_{yy^*}Q^*_{y^*}$, where the vector $\alpha=(\alpha_{10},\ \alpha_{01})$ contains the two misclassification probabilities. Similarly to Hausman et al. (1998), we adopt the identification condition $\alpha_{10}+\alpha_{01}<1$.

Assume also that a RS of size N on Y and X is to be collected, but only n individuals accept to participate in the survey. The n sampling units for which (Y, X) is recorded form the so-called complete sample. Moreover, in order to cope with the case where a given error-prone outcome is never observed, we assume that an independent SRS of all covariates of size m is drawn from the population of interest and define $N_m=N+m$ and $n_m=n+m$. While n_m, n and m are observable in all cases, the total number of individuals involved in the main survey, N, may or may not be known. Throughout this paper we assume that N is known, since all the results may be straightforwardly simplified for the case where that information is not available; see Section 4.

Define the binary indicators R, which takes the value 1 if (Y, X) is observed or 0 otherwise, and S, which takes the value 1 or 0 when the sampling unit belongs to, respectively, the main or the supplementary dataset. We assume that the nonignorable missing data mechanism is given by

$$\Pr(R=1|Y=y,Y^*=y^*,x) = \Pr(R=1|Y=y) = \delta_y, \tag{2}$$

where $0\leq\delta_y\leq 1$. Thus, the data would be missing completely at random only when $\delta_1=\delta_0=\Pr(R=1)$. Note also that due to the independence of the main and the supplementary samples, $\Pr(R=1|Y=y, Y^*=y^*, x, S=1)=\delta_y$.

In order to handle the problem of interest by analogy with the CB sampling framework, for each of the two observable outcomes Y, we reinterpret as strata the set of respondents and the set of nonrespondents.

The error-prone proportion of each stratum of respondents and nonrespondents in the population is the same, Q_y, and in the sample is, respectively, H_y=Pr(Y=y, R=1, S=1) and H_y^{nr}=Pr(Y=y, R=0, S=1). Additionally, the SRS form another stratum with proportion 1 in the population, because this sample is random, and H_S=Pr(S=0) in the sample. Due to the independence of the supplementary and the main sample we may reexpress the missing data mechanism in Eq. (2) as $\delta_y = \frac{H_y}{Q_y(1-H_S)}$.

In this setup, the likelihood function for an individual in the available dataset,

$$\begin{aligned} l(y,x,r,s) &= \left[h(y,x,r=1,s=1)^r \mathrm{Pr}(r=0,s=1)^{1-r}\right]^s h(x,s=0)^{1-s} \\ &= \left\langle [H_y h(x|y)]^r \left\{ \sum_{y=0}^{1} \int_{\mathcal{X}} [Q_y(1-H_S)-H_y] h(x|y)\mathrm{d}x \right\}^{1-r} \right\rangle^s [H_S f(x)]^{1-s} \\ &= \left\{ \left[\frac{H_y}{Q_y} \mathrm{Pr}(y|x,\theta,\alpha) f(x)\right]^r (1-H_S-H_1-H_0)^{1-r} \right\}^s [H_S f(x)]^{1-s}. \end{aligned} \quad (3)$$

is similar to that in Ramalho and Smith (2003), with the crucial difference that some of the densities and probabilities are now error-prone. From the density functions of (R, S) and X derived from Eq. (3), respectively,

$$\mathrm{Pr}(R=r,S=s) = \left[(H_1+H_0)^r (1-H_S-H_1-H_0)^{1-r}\right]^s H_S^{1-s}$$

and

$$h(x) = f(x)\left[H_S + \sum_{y=0}^{1} \frac{H_y}{Q_y} \mathrm{Pr}(y|x,\theta,\alpha)\right] + 1-H_S-H_1-H_0,$$

we may conclude that although the indicators R and S are ancillary for θ and α, the covariates do not share this property. Thus, the efficient GMM estimators proposed in the next section are based on the likelihood (3), which is not conditional on X. Moreover, the analysis is conditional on R and S, since H=(H_0, H_1, H_S) is estimated together with the remaining parameters of interest instead of being estimated separately from $\hat{H}_y = \frac{n_y}{N_m}$ and $\hat{H}_S = \frac{m}{N_m}$, where n_y is the number of fully observed subjects reporting Y=y; for a discussion on this procedure of conditioning the analysis on ancillary statistics, see Imbens and Lancaster (1996).

3. Generalized method of moments estimation

In order to avoid the specification of $f(x)$, assume that the covariates follow a discrete distribution with L points of support x^l, l=1, 2…, L, and associated probability mass parameters Pr(X=x^l)=π_l, π_l>0, l=1, 2…, L. The resultant log-likelihood function based on Eq. (3),

$$\begin{aligned} L(H,\theta,\pi) = &\sum_{i=1}^{N_m} s_i r_i \left[\ln H_{y_i} + \ln \mathrm{Pr}\left(y_i|x^{l_i},\theta,\alpha\right) - \ln \sum_{l=1}^{L} \pi_l \, \mathrm{Pr}\left(y_i|x^l,\theta,\alpha\right) + \ln \pi_{l_i}\right] \\ &+ \sum_{i=1}^{N_m} s_i(1-r_i)\ln(1-H_{S_i}-H_{1_i}-H_{0_i}) + \sum_{i=1}^{N_m}(1-s_i)(\ln H_{S_i} + \ln \pi_{l_i}), \end{aligned} \quad (4)$$

is maximized with respect to the vector of parameters (H, θ, α, π) subject to the restriction $\sum_{l=1}^{L}\pi_l = 1$. The first order conditions of Eq. (4) are very similar to those in Ramalho and Smith (2003). Thus, by analogous calculations, π_l is concentrated out from those functions. Hence, the dependence on the discrete distribution assumed for $f(x)$ is removed, since π is replaced by Q_1^* in the vector of parameters of interest, and the following estimating functions are obtained:

$$g(v,\varphi)_{H_t} = srI_{(y=t)} - H_t \tag{5}$$

$$g(v,\varphi)_{H_S} = 1 - s - H_S \tag{6}$$

$$g(v,\varphi)_{\theta} = p\left\{sr\frac{y-P}{P(1-P)} - [1-s(1-r)]\frac{A}{B}\right\} \tag{7}$$

$$g(v,\varphi)_{\alpha_{yy^*}} = [y - \Pr{}^*(y^*|x,\theta)]\left\{sr\frac{y-P}{P(1-P)} - [1-s(1-r)]\frac{A}{B}\right\} \tag{8}$$

$$g(v,\varphi)_{Q_1^*} = Q_1 - [1-s(1-r)]\frac{P}{B}, \tag{9}$$

where $V=(Y,X,R,S)$, $t=\{0,1\}$, $P=\Pr(Y=1|x,\theta,\alpha)$, $p=\nabla_\theta P$, $Q_1=\alpha_{10}+(1-\alpha_{10}-\alpha_{01})Q_1^*$, $A=\frac{H_1}{Q_1}-\frac{H_0}{1-Q_1}$, $B=H_S+\frac{H_0}{1-Q_1}+AP$ and φ is the vector of parameters of interest. φ is defined as $\varphi=(H, \theta, \alpha, Q_1^*)$ when both Q_1 and Q_1^* are unknown, or simply as $\varphi=(H, \theta, \alpha)$, when one of those probabilities is known. In this case, the known probability is replaced in Eqs. (5)–(9).

The estimating functions (5)–(9) are used as moment indicators in the GMM framework. The objective function to be minimized is $\Upsilon_{N_m}(\varphi) = g_{N_m}(v,\varphi)' W_{N_m} g_{N_m}(v,\varphi)$, where $g_{N_m}(v,\varphi)=\frac{1}{N_m}\sum_{i=1}^{N_m} g(v_i,\varphi)$ is the sample counterpart of the moment conditions $E[g(v, \varphi)]=0$, with $E[.]$ denoting expectation taken over $l(y, x, r, s)$ of Eq. (3) and $g(v, \varphi)$ defined in Eqs. (5)–(9), and W_{N_m} is a positive semi-definite weighting matrix. Assume that the usual regularity conditions required for GMM estimation are met; see Newey and McFadden (1994, Theorems 2.6, 3.4). The resulting optimal estimator, $\hat{\varphi}$, obtained from choosing $W_{N_m}=\Psi_{N_m}^{-1}$, where Ψ_{N_m} is a consistent estimator of $\Psi=E[g(v, \varphi)g(v, \varphi)']$, is consistent for the true value φ^0 and satisfies $\sqrt{N_m}(\hat{\varphi}-\varphi^0)\xrightarrow{d}N[0,(G'\Psi^{-1}G)^{-1}]$, where $\xrightarrow{d}$ denotes convergence in distribution and $G=E[\nabla_\varphi g(v, \varphi)']$. Asymptotic efficiency, in the semiparametric sense, can also be proved by an analogous demonstration to that of Imbens (1992, Theorem 3.3).

4. Some particular cases

First, for cases where N is unknown, we need to define $H_y=\Pr(Y=y, S=1|R=1)$, set $R=1$ and $H_y^{nr}=0$, replace N_m by n_m and, since $H_S+H_0+H_1=1$, suppress $g(v, \varphi)_{H_S}$. Moreover, if none of the sampling units for which $Y=0$ responds and all subjects for which $Y=1$ reveal (Y, X), we obtain a generalization of Lancaster and Imbens' (1996) estimators for nonresponse to handle misclassification by setting $Y=1$, $n_m=n_1+m$, and, as $H_0=0$, suppressing $g(v, \varphi)_{H_0}$.

Second, when a SRS is not available, we set $S=1$ and $H_S=0$, replace N_m by N, and eliminate $g(v, \varphi)_{H_S}$. In this framework, Ramalho's (2002) estimators for CB samples subject to misclassification are obtained by considering N unknown, which requires setting $R=1$, replacing N by n, and eliminating either $g(v, \varphi)_{H_0}$ or $g(v, \varphi)_{H_1}$.

Finally, in cases where the structural model is a logit, such that $\Pr{}^*(y^*=1|x,\ \theta)=(1+e^{-\ x'\theta})^{-\ 1}$, where $\theta=(\theta_0,\ \theta_1)$, with θ_0 defined as an intercept term and $H_y>0$, by an analogous demonstration to that of Caudill and Cosslett (2004) for CB sampling, it can be shown that the shape of $\Pr(y=j|x,\theta,\alpha)=\frac{\alpha_{j0}e^{-x'\theta}+\alpha_{j1}}{1+e^{-x'\theta}}$ is preserved by the probability of Y given X in the complete error-prone data,

$$\Pr{}_S(y=j|x,R=1,\theta,\alpha,\delta_0,\delta_1)=\frac{\sum\limits_{y^*=0}^{1}\delta_j\alpha_{jy^*}\Pr{}^*(y^*|x,\theta)f(x)}{\sum\limits_{y=0}^{1}\sum\limits_{y^*=0}^{1}\delta_y\alpha_{yy^*}\Pr{}^*(y^*|x,\theta)f(x)}$$
$$=\frac{\delta_j\left(\alpha_{j0}e^{-x'\theta}+\alpha_{j1}\right)}{\sum\limits_{y=0}^{1}\delta_y\alpha_{y0}e^{-x'\theta}+\sum\limits_{y=0}^{1}\delta_y\alpha_{y1}}=\frac{\varpi_{j0}\frac{\varpi_0}{\varpi_1}e^{-x'\theta}+\varpi_{j1}}{1+\frac{\varpi_0}{\varpi_1}e^{-x'\theta}},\tag{10}$$

where $\varpi_0=\sum_{y=0}^{1}\delta_y\alpha_{y0},\varpi_1=\sum_{y=0}^{1}\delta_y\alpha_{y1},$ and $\varpi_{yy^*}=\Pr{}_S(Y=y|Y^*=y,R=1)=\frac{\alpha_{yy^*}\delta_y}{\sum_{y=0}^{1}\alpha_{yy^*}\delta_y}$. The only difference is that now θ_0 and α_{yy^*} are replaced by, respectively, $\gamma=\theta_0-\ln\frac{\varpi_0}{\varpi_1}$ and ϖ_{yy^*}. Thus, for consistent estimation of θ_1, one may utilize the simple likelihood $\Pr(y=j|x,\ \theta,\ \alpha)$, where only the problem of misclassification is accounted for, with the complete dataset.

5. A Monte Carlo simulation study

This section analyzes the performance of the estimation method proposed in this paper in cases where Y^* given X is described by a logit model, the main sample only contains individuals who reported 1, a SRS is available, and the number of individuals choosing 0, n_0, and, consequently, N, are unknown. We replicated two of Lancaster and Imbens' (1996) Monte Carlo experimental designs but we admitted the possibility that some of the observed subjects have chosen alternative zero instead of the reported outcome "one".

The covariates X were generated from a bivariate normal distribution with zero means, unit variances and zero correlation. In the two experimental designs, designated as A and B, the vector of parameters of interest θ contained in $\Pr{}^*(y^*=1|x,\ \theta)=(1+e^{-x'\theta})^{-1}$, where $\theta=(\theta_0,\ \theta_1,\ \theta_2)$ with θ_0 defined as an intercept term, was set equal to, respectively, (0.0, 2.0, 0.5) and (−1.89, 1.0, 1.0), producing $Q_1^*=0.50$ and $Q_1^*=0.20$. In both designs $H_1=H_S=0.5$ (and $H_0=0$) such that $n=n_1=m=2500$, and we performed experiments for three misclassification probabilities: $\alpha_{10}=\alpha_{01}=\bar{\alpha}=\{0.02,\ 0.05,\ 0.20\}$. In all experiments we assumed that the marginal probabilities Q_1^* and Q_1 are unknown and compared Lancaster and Imbens' (1996) estimator (LIE) and its modified version for misclassification developed in this paper (MLIE). The vector of parameters estimated in each case is, respectively, $(H_1,\ \theta,\ Q_1^*)$ and $(H_1,\ \theta,\ \bar{\alpha},\ Q_1^*)$. Similarly to previous studies where the probabilities of misclassification are estimated, e.g. Hausman et al. (1998), Ramalho (2002), we considered a sample size of $n_m=5000$. However, in our experiments, the estimation problem is much more complex: those papers deal with datasets of 5000 observations for which all the information is measured, while we only have complete information for 2500 observations, all of them reporting "one". Obviously, the MLIE would not perform well with

Table 1
Summary statistics for GMM estimators from 1000 replications

$\theta^{\text{design A}}=(0.0,\ 2.0,\ 0.5)$, $\theta^{\text{design B}}=(-1.89,\ 1.0,\ 1.0)$													
				$\hat{\theta}_1$					$\hat{\theta}_2$				
Design	$\bar{\alpha}$	Estimator	FC	Bias		SD	MAE	RMSE	Bias		SD	MAE	RMSE
				Mean	Med.				Mean	Med.			
A	.02	LIE	0	−.129	−.132	.150	.099	.198	−.128	−.132	.071	.048	.146
		MLIE	4	−.006	−.001	.321	.190	.321	.002	−.008	.119	.067	.119
	.05	LIE	1	−.275	−.277	.133	.087	.305	−.272	−.240	.068	.043	.280
		MLIE	7	.010	.005	.338	.215	.338	.012	.004	.116	.075	117
	.20	LIE	63	−.647	−.651	.105	.069	.655	−.646	−.650	.055	.038	.648
		MLIE	13	.166	.041	2.850	.398	2.855	2	.188	.020	.799	.132
B	.02	LIE	5	−.069	−.073	.069	.048	.098	−.067	−.066	.065	.042	.093
		MLIE	9	.009	−.005	.423	.086	.423	.013	−.003	.449	.081	.449
	.05	LIE	5	−.166	−.169	.065	.043	.178	−.163	−.163	.062	.042	.174
		MLIE	9	.005	.006	.255	.086	.255	.011	.009	.272	.084	.272
	.20	LIE	363	−.467	−.472	.043	.027	.469	−.465	−.468	.041	.028	.467
		MLIE	55	.026	.023	.586	.127	.587	.018	.026	.414	.127	.414

$n_m=400$, the sample size considered by Lancaster and Imbens (1996) in the absence of misclassification.

Table 1 reports for each estimator the mean and the median bias in percentage terms, the standard deviation across 1000 replications, the mean absolute error and the root mean squared error for the slope estimates. The number of replications that failed to converge (FC) is also reported, since it was very large for $\bar{\alpha}=0.2$, mainly for LIE, which ignore the presence of misclassification.[1] The behaviour of the MLIE is very promising, namely for the two smallest misclassification probabilities, where the worst distortion of the MLIE is 1.3%. Naturally, the performance decays with the highest level of misclassification, but even in these cases the median bias of the MLIE is smaller than 4.1%.[2] On the other hand, the LIE exhibits large biases, but presents smaller standard deviations than those of the MLIE, which captures the additional variability induced by misclassification. Therefore, the biased LIE often presents smaller (understated) MAE and RMSE than the MLIE.

References

Caudill, S.B., Cosslett, S.R. 2004, A note on estimation from choice-based samples with misclassification in the response variable. Unpublished Working Paper.

Hausman, J.A., Abrevaya, F., Scott-Morton, F.M., 1998. Misclassification of the dependent variable in a discrete-response setting. Journal of Econometrics 87, 239–269.

Imbens, G., 1992. An efficient method of moments estimator for discrete choice models with choice-based sampling. Econometrica 60, 1187–1214.

[1] This problem is more serious in design B, because the available dataset is near a pure CB sampling design, a situation where the problem of identification of θ_0 and Q_1^* is well known; see Lancaster and Imbens (1996).

[2] In design A, the results for $\alpha=0.2$ were negatively affected by the presence of 4 replications where the estimate for θ_1 was larger than 30. Eliminating these replications, the mean bias for θ_1 and θ_2 is reduced to, respectively, 8.0% and 9.3% and their standard deviations across the replications are 0.713 and 0.240.

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

Lancaster, T., Imbens, G., 1996. Case–control studies with contaminated controls. Journal of Econometrics 71, 145–160.

Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. IV. Elsevier Science, pp. 2113–2245.

Ramalho, E.A., 2002. Regression models for choice-based samples with misclassification in the response variable. Journal of Econometrics 106, 171–201.

Ramalho, E.A., Smith, R.J., 2003. Discrete Choice Nonresponse, CeMMAP Working Paper CWP07/03.