econometrics

MDPI

Article

Response-Based Sampling for Binary Choice Models With Sample Selection

Maria Felice Arezzo *,† **and Giuseppina Guagnano** †

Department of Methods and Models for Economics, Territory and Finance, Faculty of Economics, Sapienza University of Rome, Rome, 00161, Italy; giuseppina.guagnano@uniroma1.it

* Correspondence: mariafelice.arezzo@uniroma1.it; Tel.: +39-06-49766424

† These authors contributed equally to this work.

Received: 29 December 2017; Accepted: 1 March 2018; Published: 7 March 2018

Abstract: Sample selection models attempt to correct for non-randomly selected data in a two-model hierarchy where, on the first level, a binary selection equation determines whether a particular observation will be available for the second level (outcome equation). If the non-random selection mechanism induced by the selection equation is ignored, the coefficient estimates in the outcome equation may be severely biased. When the selection mechanism leads to many censored observations, few data are available for the estimation of the outcome equation parameters, giving rise to computational difficulties. In this context, the main reference is Greene (2008) who extends the results obtained by Manski and Lerman (1977), and develops an estimator which requires the knowledge of the true proportion of occurrences in the outcome equation. We develop a method that exploits the advantages of response-based sampling schemes in the context of binary response models with a sample selection, relaxing this assumption. Estimation is based on a weighted version of Heckman's likelihood, where the weights take into account the sampling design. In a simulation study, we found that, for the outcome equation, the results obtained with our estimator are comparable to Greene's in terms of mean square error. Moreover, in a real data application, it is preferable in terms of the percentage of correct predictions.

Keywords: selection bias; response-based sampling; observational study; credit scoring

JEL Classification: C13; C34; C25

1. Introduction

Most empirical work in the social sciences is based on observational data that are incomplete. There are many types of selection mechanisms that result in a non random sample. Some of them are due to sample design, while others depend on the behavior of the units being sampled, other than non-response or attrition.

In the first case, data are usually missing for all the variables of interest; for example, in estimating a saving function for all the families of a given country, a bias would arise if only families whose household head shows certain characteristics were sampled. However, when causes of missingness are appropriately exogenous, using a sub-sample has no serious consequences. In the second case, instead, there is a self-selection of the sample units and data availability on a key variable depends on the behavior of the units for another variable. The classical example is the estimation of the wage offer equation for people of working age, where we want to estimate the expected wage of an individual using a set of exogenous characteristics (gender, age, education, etc). This equation, by definition, should be valid for people of working age, independently of their working conditions at the time of the survey. On the contrary, we can only observe the wage offer for *employed* individuals; in regressing wages on their characteristics, we are not making inferences for the population as a whole. In other

words, people in employment are a selected sample of the population and their wages are higher than the unemployed would have had. Hence, the results will tend to be biased (sample selection bias).To avoid this bias, we should take into account the selection mechanism due to the individual's decision to take a job and then receive a wage.

As is well known, Heckman (1979) proposed a useful framework for handling estimation when the sample is subject to a selection mechanism. In the original framework, the dependent variable in the outcome equation (the wage equation in the above example) is continuous and can be explained by a linear regression model with a normal random component. In addition to the output equation, a selection equation describes the selection rule by means of a binary choice model (probit).

The original Heckman model was extended in many directions and a survey would be beyond the scope of this paper, but the interested reader can refer to Vella (1998) and Lee (2003). For our purposes, the relevant framework is the one where both the output and the selection equations are defined as a binary choice model (Dubin and Rivers 1989). The applications are countless and span every field of applied research: from political science (Dubin and Rivers 1989; Grier et al. 1994; Jacobs and O'Brien 1998) to health economics (Van de Ven and Van Praag 1981), transport (Ingram 1999; Kayser 2000), and credit scoring (Boyes et al. 1989; Greene 2008), just to make a very brief and non-exhaustive list.

In our work, we focus on the problem that arises when the selection mechanism is particularly severe and gives rise to a large amount of censored observations. This situation might occur for example when the event of interest is infrequent or fragmented or no frames are available for standard sampling procedures. Relevant examples are the surveys on elusive populations (such as working children, homeless persons, illegal immigrants, tax evaders, and drug users); in these populations, by virtue of their characteristics or difficulties in obtaining the required information, adequate samples cannot be defined, drawn or implemented using the standard procedures of random sampling. In this case, no or very partial frames are available for sampling, or many units are not available or willing to participate in the survey. Consequently, random sampling is either not feasible or inefficient, because it is very costly (due to the high number of total units to be sampled to obtain a sufficient number of uncensored observations).

One solution is given by the response-based sampling scheme, also known as case-control setting. In this design (see Hosmer and Lemeshow 2013), samples of a fixed size are randomly chosen from the two strata identified by the dependent variable.

In the more specific context of binary choice models with response-based sampling and sample selection bias, a solution is given by Greene (2008), who extended the work of Manski and Lerman (1977). In both proposals, however, the estimator requires the knowledge of the true proportion of occurrences in the outcome equation. In some applications, this requirement can be a serious limitation.

Our work fits into this last framework and aims to overcome these shortfalls.

The remainder of the paper is organized as follows. In Section 2, we start by describing the methodological background and then we illustrate the proposed estimation procedure. In Section 3, a simulation study compares our approach with Greene's. In Section 4, the proposed method is applied to real data on credit scoring. In Section 5, a discussion and some concluding remarks are provided.

2. Methods

2.1. General Background

2.1.1. Sample Selection

Let us first introduce some notations and briefly illustrate the sample selection framework with a binary choice model for both the selection and the output equations (Dubin and Rivers 1989).

Let Y^* and A^* be two latent (unobservable) variables characterizing the output and the selection equations, respectively. The model, in its general form, is:

$$Y_i^* = \boldsymbol{X_{1i}}\beta + \epsilon_{1i} \tag{1a}$$
$$A_i^* = \boldsymbol{X_{2i}}\theta + \epsilon_{2i} \tag{1b}$$

where $\boldsymbol{X_i} = (\boldsymbol{X_{1i}}, \boldsymbol{X_{2i}})$ is a vector of exogenous variables (namely, $\boldsymbol{X_{1i}}$ for Y_i and $\boldsymbol{X_{2i}}$ for A_i), containing all the relevant covariates, and β and θ are the vectors of regression coefficients. Let us define Y_i and A_i as two observable variables such that:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$A_i = \begin{cases} 1 & \text{if } A_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The p.d.f. of Y_i and A_i is Bernoulli, with probability of success depending on the parameters β and θ respectively.

Model (3) defines the mechanism which governs the censoring process: we can observe Y_i if and only if $A_i = 1$. On the contrary, if $A_i = 0$, Y_i will be missing.

In the general case, if we were to estimate the parameters of Equation (1a) without considering the selection process in (1b), a bias would arise. This is because the processes represented by the two equations are related, i.e., $corr(\epsilon_1, \epsilon_2) = \rho$ is not null (see for example Cameron and Trivedi (2005) for further details).

The likelihood function for model (1a-1b) is:

$$\begin{aligned} L(\eta) =& \prod_{i=1}^{n} \Big[Pr(A_i^* < 0) \Big]^{1-A_i} \cdot \Big[P(Y_i = y_i | A_i^* > 0) \cdot Pr(A_i^* > 0) \Big]^{A_i} = \\ =& \prod_{i=1}^{n} \Big[1 - {}_A\pi(\boldsymbol{X_i}) \Big]^{1-A_i} \cdot \Big[P(Y_i = y_i | A_i = 1) \cdot {}_A\pi(\boldsymbol{X_i}) \Big]^{A_i} \end{aligned} \tag{4}$$

where $\eta = (\beta, \theta, \rho)$ is the vector of parameters to be estimated, $y_i = 0, 1$ and the function ${}_A\pi(\cdot)$ gives the probability that an observation is uncensored.

2.1.2. Response-Based Sampling

Let us now briefly review the main characteristics of the response-based sampling scheme relevant to our method. In this context, samples of fixed size are randomly chosen from the two strata identified by the dependent variable A (note that, for our purposes, in the context of sample selection, it corresponds to the dependent variable of the selection equation). In particular, n_A units are drawn at random from the N_A cases and $n_{\bar{A}}$ units from the $N_{\bar{A}}$ controls.

The likelihood function is the product of the stratum-specific likelihoods, and depends on the probability that the individual is in the sample and on the joint density of the covariates (Hosmer and Lemeshow 2013):

$$\begin{aligned} \prod_{i=1}^{n} f(\boldsymbol{X_i} | A_i, S_i = 1) =& \prod_{i=1}^{n_A} Pr(\boldsymbol{X_i} | A_i = 1, S_i = 1) \cdot \\ & \cdot \prod_{i=1}^{n_{\bar{A}}} Pr(\boldsymbol{X_i} | A_i = 0, S_i = 1), \end{aligned} \tag{5}$$

where S_i is a binary variable which takes value 1 if the i-th individual is in the sample and 0 otherwise.

2.2. A Binary Choice Model with Sample Selection under a Response-Based Sampling

2.2.1. The Weighted Endogenous Sampling Likelihood

The estimator, first proposed by Manski and Lerman (1977), is designed for a binary response model under a response based sampling framework and it is called the Weighted Endogenous Sampling Maximum Likelihood (WESML) estimator, because it assigns weights to the likelihood function. The weights are given by the ratio between the proportion of individuals in the population for which $A = 1$ and the corresponding proportion in the sample.

2.2.2. The WESML Estimator Corrected for Selection Bias

Greene (2008) extended the work of Manski and Lerman to the context of a binary response model with selection bias and response-based sampling; in that work, the goal was to estimate the probability of loan defaults $P(Y = 1|\boldsymbol{X})$ from a sample of individuals whose credit card application was accepted ($A = 1$). The corresponding likelihood is:

$$\log L(\eta) = \sum_{i=1}^{n} \frac{N_{\bar{A}}/N}{n_{\bar{A}}/n} \cdot \log\Big[1 - {}_A\pi(\boldsymbol{X}_i)\Big]^{1-A_i} + \frac{(N_{yA}/N)}{n_{yA}/n} \cdot \log\Big[P(Y_i = y_i|A_i = 1) \cdot {}_A\pi(\boldsymbol{X}_i)\Big]^{A_i}$$

where again $y_i = 0, 1$ and, according to Manski and Lerman, the weights are given by the ratio of two proportions: we have population-level quantities at the numerator and the corresponding sampling quantities at the denominator. More precisely, in Greene's application $N_{\bar{A}}/N$ represents the fraction of non-cardholders in the population and $n_{\bar{A}}/n$ is the homologous in the sample; coherently, N_{yA}/N is the prevalence of defaults ($y = 1$) and non defaults ($y = 0$) in the population, while n_{yA}/n is the sample counterpart.

It is important to note that Greene's estimator (GE) requires knowledge of the proportion in the population not only for the controls (i.e., $N_{\bar{A}}/N$), but also for the response variable (namely N_{1A}/N and N_{0A}/N). This requirement can be an insurmountable obstacle in some applications.

2.2.3. The Sample-Selection Response-Based-Sampling Likelihood

In the following, we provide our main result which is the likelihood function in the framework of interest, i.e., a sample selection mechanism with a severe censoring process assuming that the population prevalences $\frac{N_{0A}}{N_A}$ and $\frac{N_{1A}}{N_A}$ are unknown. The full proof is given in the Appendix.

We make the following very general and non restrictive assumptions:

1. We have a set of fully informative and exogenous covariates $\boldsymbol{X_i} = (\boldsymbol{X_{1i}}, \boldsymbol{X_{2i}})$.
2. Conditional on the covariates, the probability that an observation is uncensored does not depend on its value, i.e., $P(A_i = 1|S_i = 1, \boldsymbol{X_i}, Y_i) = P(A_i = 1|S_i = 1, \boldsymbol{X_i})$.
3. The set of covariates $\boldsymbol{X_{1i}}$, specific for Y_i, and the set $\boldsymbol{X_{2i}}$, specific for A_i, may have common elements but they cannot fully overlap. In particular, we assume that there is at least one covariate in the selection equation which is not in the outcome.

Assumption 1 means that a correlation between the covariates and the residual terms in Equations (1a) and (1b) does not exist . Assumption 2 is justified because, as the covariates are informative, all the information brought by Y_i is contained in $\boldsymbol{X_i}$. Assumption 3 is necessary for parameter identification (exclusion condition).

Under the conditions stated, the likelihood function for a binary-choice model with sample-selection response-based sampling is:

$$\begin{aligned} L(\eta) &= \prod_{i=1}^{n} f\left(\boldsymbol{X_i}|S_i=1\right)\left\{\left(1-{}_A\pi(\boldsymbol{X_{2i}})\right)\cdot\frac{N}{N_{\bar{A}}}\right\}^{1-A_i} \\ &\cdot\left\{\frac{N}{N_A}\cdot\frac{n_A}{n_{y_iA}}{}_A\pi(\boldsymbol{X_{2i}})\cdot P\left(Y_i=y_i|A_i=1,\boldsymbol{X_i}\right)\right\}^{A_i} \\ &= \prod_{i=1}^{n} f\left(\boldsymbol{X_i}|S_i=1\right)\left\{\left(1-{}_A\pi(\boldsymbol{X_{2i}})\right)\cdot\frac{N}{N_{\bar{A}}}\right\}^{1-A_i} \\ &\cdot\left\{\left[{}_Y\pi(\boldsymbol{X_{1i}})\cdot\frac{N}{N_A}\frac{n_A}{n_{1A}}\right]^{y_i}\cdot\left[\left(1-{}_Y\pi(\boldsymbol{X_{1i}})\right)\cdot\frac{N}{N_A}\frac{n_A}{n_{0A}}\right]^{1-y_i}\cdot{}_A\pi(\boldsymbol{X_{2i}})\right\}^{A_i} \end{aligned} \tag{6}$$

where ${}_Y\pi(\boldsymbol{X_{1i}})$ is the probability of observing $Y=1$ given that the observation is uncensored and n_{yA} is the amount of uncensored units in the sample having $Y=y$, with $y=0,1$. Moreover, as previously said, ${}_A\pi(\boldsymbol{X_{2i}})$ is the probability that an observation is uncensored, n_A is the number of units sampled from the N_A uncensored observations and $n_{\bar{A}}$ is the number of units sampled from the $N_{\bar{A}}$ censored observations.

It is easy to see that the likelihood (6) is a weighted version of (4), and the weights simply take into account the sampling scheme. In addition, note that, in the maximization process, the term $f\left(\boldsymbol{X_i}|S_i=1\right)$ is non influential, as it does not contain any information on the vector of parameters η, and that the only known quantities at the population level are N_A and N.

The estimator for η obtained by maximizing this likelihood will be referred to as Sample Selection Response-based Sampling (SSRS hereafter).

3. Simulation Results

In this section, we will compare the results obtained using the SSRS and GE estimators through a Monte Carlo experiment.

The generating model is:

$$Y_i^* = 0.5 + 1.5X_{11i} - 1.8X_{12i} + \epsilon_{1i} \tag{7a}$$

$$A_i^* = \theta_0 + 0.8X_{21i} - 0.5X_{22i} + \epsilon_{2i} \tag{7b}$$

where all Xs are independently generated from a univariate standard normal. Note that, in Equation (7b), θ_0 governs the proportion of uncensored observations in the population (i.e., N_A/N). In particular, in the following, we consider three (approximate) proportions: 4% (ensured by $\theta_0=-2$), 15% ($\theta_0=-1.43$) and 30% ($\theta_0=-0.72$).

Even though in the derivation of Equation (6) we have not assumed any probability model for ${}_Y\pi(\boldsymbol{X_{1i}})$ and for ${}_A\pi(\boldsymbol{X_{2i}})$, we had to do it for the Monte Carlo experiment. More precisely, we set:

$$\begin{pmatrix}\epsilon_{1i}\\ \epsilon_{2i}\end{pmatrix} \sim NID\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&\rho\\ \rho&1\end{pmatrix}\right] \tag{8}$$

From Equation (8) and referring to the probabilities ${}_A\pi(\boldsymbol{X_2})$ and ${}_Y\pi(\boldsymbol{X_1})$ in (6), it follows:

$$P(A=0) = 1 - {}_A\pi(\boldsymbol{X_2}) = \Phi\left(-\theta'\boldsymbol{X_2}\right) \tag{9}$$

$$\begin{aligned} P(Y=1, A=1) &= P(Y=1|A=1)\cdot P(A=1) = {}_Y\pi(\boldsymbol{X_1})\cdot {}_A\pi(\boldsymbol{X_2}) \\ &= \Phi_2\left(\beta'\boldsymbol{X_1}, \theta'\boldsymbol{X_2}, \rho\right) \end{aligned} \tag{10}$$

$$\begin{aligned} P(Y=0, A=1) &= P(Y=0|A=1)\cdot P(A=1) = (1 - {}_Y\pi(\boldsymbol{X_1}))\cdot {}_A\pi(\boldsymbol{X_2}) \\ &= \Phi_2\left(\beta'\boldsymbol{X_1}, -\theta'\boldsymbol{X_2}, -\rho\right) \end{aligned} \tag{11}$$

where Φ and Φ_2 are c.d.f. of the univariate and the bivariate normal respectively.

For each censoring scheme (N_A/N), we performed 500 replications for any $\rho \in [-0.8, 0.8]$ with a step of 0.1 and for three different sampling proportions of cases and controls (i.e., $pc = \frac{n_a}{n} \in (0.25, 0.5, 0.75)$). From a population of size $N = 1,000,000$, we drew samples of different size to evaluate the performances of the two estimators. More precisely, we drew sample with dimension spanning from $n = 2000$ to $n = 10,000$ by 500.

In commenting on the results, we begin by comparing the empirical densities of SSRS and GE estimators for $n = 2000$ and $n = 10,000$, for each censoring scheme[1]. Rather than overwhelm the reader with statistics, we prefer to present information in a summary graphical form.

As regards β_1, for $N_A/N = 0.04$ (see Figure 1), we note that the patterns are similar for $pc = 0.25$, whereas the behavior differs more as pc increases. Furthermore, the SSRS estimator is almost always more concentrated except for few combinations of ρ and pc.

As expected, when N_A/N increases, the discrepancies between SSRS and GE estimators fade away (see Figure 2).

Analogous considerations apply for β_2 (see Figures 3 and 4). In summary, Figures 1–4 show that the SSRS and GE estimators have comparable behavior, even though the former uses less information at population level. It should be underlined that a direct comparison between the two sets of coefficient estimates may not always be appropriate. In fact, as noted by Mroz and Zayats (2008), a better comparison would be based on the relative effect, that is on the coefficients ratio. For these reasons we computed the average ratios $\hat{\theta}_2/\hat{\theta}_1$ and $\hat{\beta}_2/\hat{\beta}_1$ in the simulations, for each $k, pc, N_A/N$. The results, available upon request, are very close to the true ratios, $\theta_2/\theta_1 = -0.63$ and $\beta_2/\beta_1 = -1.2$, for both GE and SSRS estimators and they show once again that GE slightly outperforms SSRS in the selection equation, while the reverse seems to happen for the outcome equation.

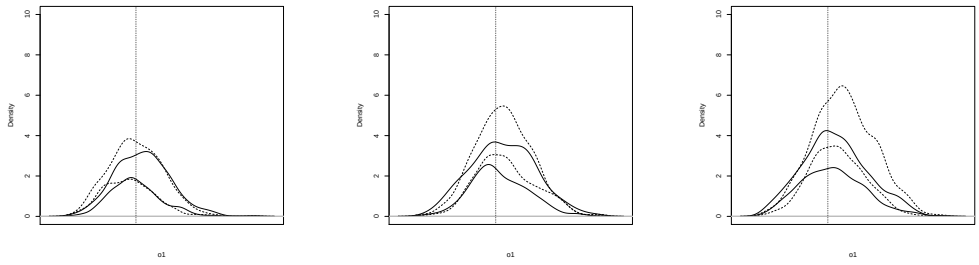
Coming to the comparison of the estimated MSEs, as shown in Figures 5 and 6, the overall results obtained by the SSRS estimator are quite similar to Greene's. Specifically, it seems that SSRS gives better results in estimating the parameters of the outcome equation (which are, generally speaking, those of greater interest), especially in the case of a severe censoring mechanism ($N_A/N = 0.04$). The situation is reversed for the selection equation.

When $N_A/N = 0.30$, the two estimators are substantially equivalent for the selection equation, while SSRS continues to slightly outperform GE for the outcome (except for $pc = 0.25$).

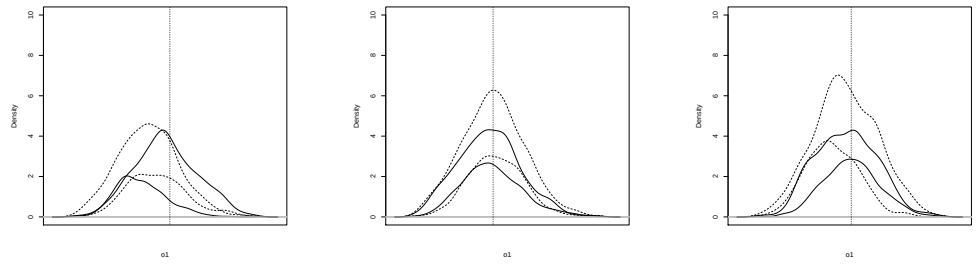
As expected, when N_A/N increases, the differences between the MSEs of the two estimators decrease in absolute value. Everything we said holds for both sample sizes.

---

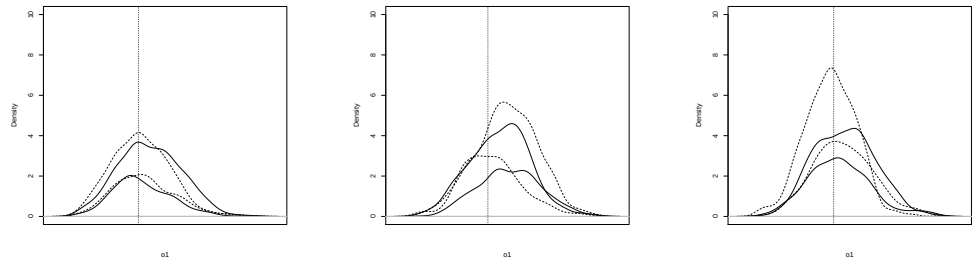
[1] For the sake of brevity, we do not report the results for $N_A/N = 0.15$ and for all values of ρ, but they are available upon request.



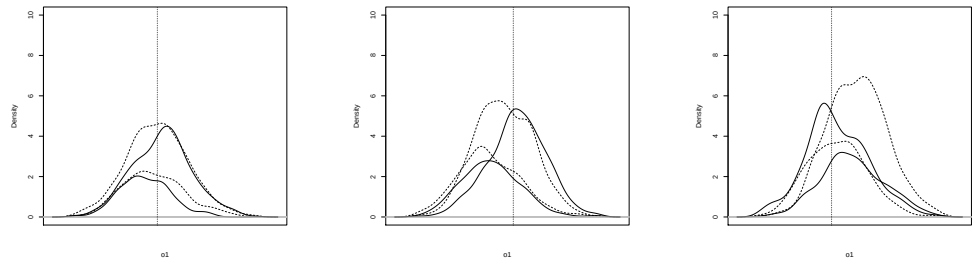
(**a**) $\rho = 0.8$, $pc = 0.25$, $\frac{N_A}{N} = 0.04$ (**b**) $\rho = 0.8$, $pc = 0.50$, $\frac{N_A}{N} = 0.04$ (**c**) $\rho = 0.8$, $pc = 0.75$, $\frac{N_A}{N} = 0.04$



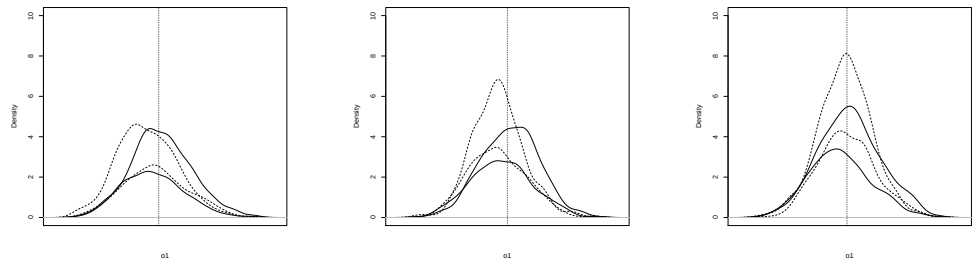
(**d**) $\rho = -0.8$, $pc = 0.25$, $\frac{N_A}{N} = 0.04$ (**e**) $\rho = -0.8$, $pc = 0.50$, $\frac{N_A}{N} = 0.04$ (**f**) $\rho = -0.8$, $pc = 0.75$, $\frac{N_A}{N} = 0.04$



(**g**) $\rho = 0.4$, $pc = 0.25$, $\frac{N_A}{N} = 0.04$ (**h**) $\rho = 0.4$, $pc = 0.50$, $\frac{N_A}{N} = 0.04$ (**i**) $\rho = 0.4$, $pc = 0.75$, $\frac{N_A}{N} = 0.04$

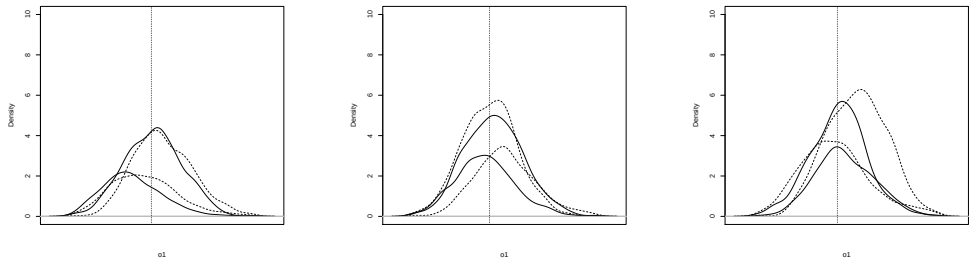


(**j**) $\rho = -0.4$, $pc = 0.25$, $\frac{N_A}{N} = 0.04$ (**k**) $\rho = -0.4$, $pc = 0.50$, $\frac{N_A}{N} = 0.04$ (**l**) $\rho = -0.4$, $pc = 0.75$, $\frac{N_A}{N} = 0.04$

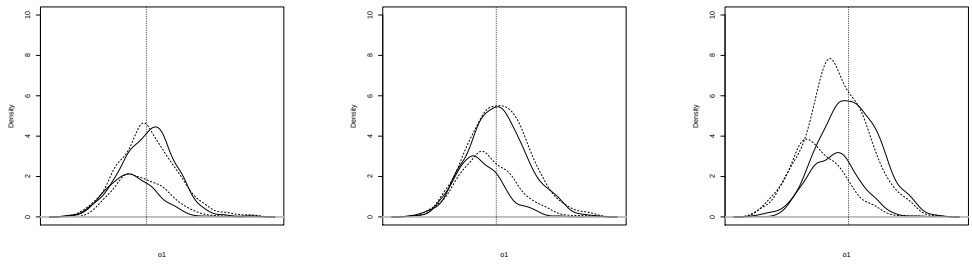


(**m**) $\rho = 0$, $pc = 0.25$, $\frac{N_A}{N} = 0.04$ (**n**) $\rho = 0$, $pc = 0.50$, $\frac{N_A}{N} = 0.04$ (**o**) $\rho = 0$, $pc = 0.75$, $\frac{N_A}{N} = 0.04$

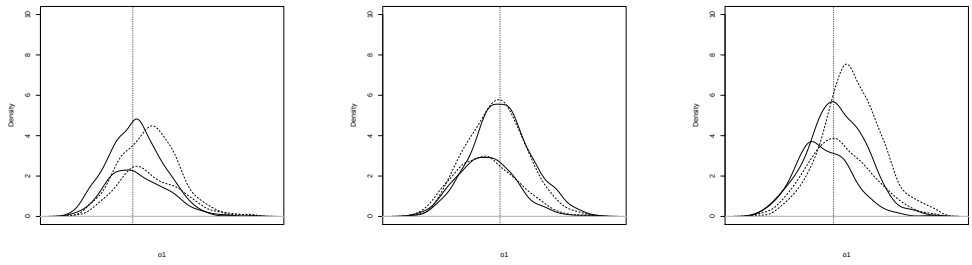
Figure 1. Densities of $\widehat{\beta}_1$ for some values of ρ, $N_A/N = 0.04$ and three sampling proportions of cases pc (SSRS: dashed lines; GE: continuous line; $n = 2000$ for the two flatter distributions; $n = 10,000$ for the two more sharpened distributions).



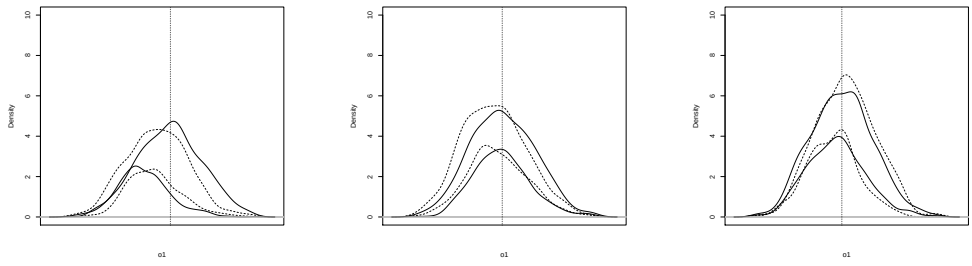
(**a**) $\rho = 0.8$, $pc = 0.25$, $\frac{N_A}{N} = 0.30$ (**b**) $\rho = 0.8$, $pc = 0.50$, $\frac{N_A}{N} = 0.30$ (**c**) $\rho = 0.8$, $pc = 0.75$, $\frac{N_A}{N} = 0.30$



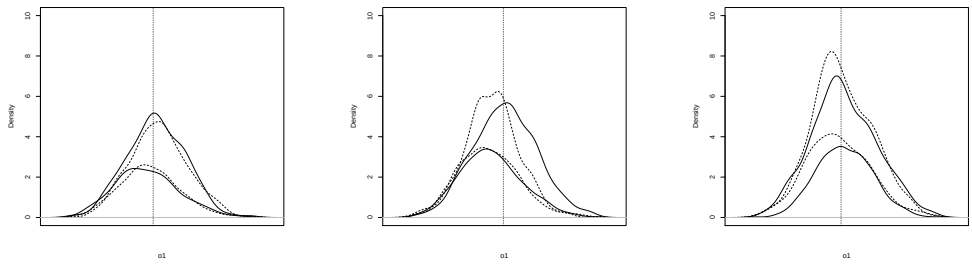
(**d**) $\rho = -0.8$, $pc = 0.25$, $\frac{N_A}{N} = 0.30$ (**e**) $\rho = -0.8$, $pc = 0.50$, $\frac{N_A}{N} = 0.30$ (**f**) $\rho = -0.8$, $pc = 0.75$, $\frac{N_A}{N} = 0.30$



(**g**) $\rho = 0.4$, $pc = 0.25$, $\frac{N_A}{N} = 0.30$ (**h**) $\rho = 0.4$, $pc = 0.50$, $\frac{N_A}{N} = 0.30$ (**i**) $\rho = 0.4$, $pc = 0.75$, $\frac{N_A}{N} = 0.30$

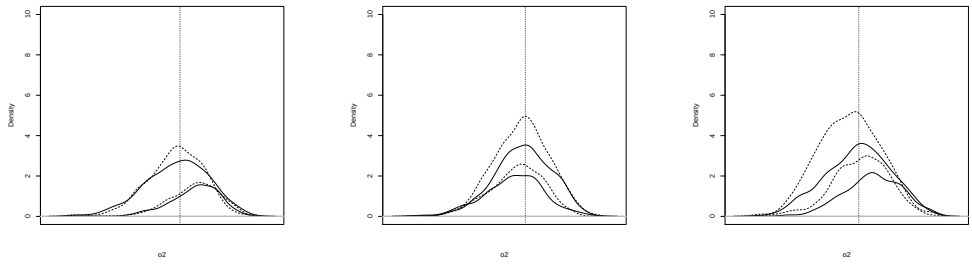


(**j**) $\rho = -0.4$, $pc = 0.25$, $\frac{N_A}{N} = 0.30$ (**k**) $\rho = -0.4$, $pc = 0.50$, $\frac{N_A}{N} = 0.30$ (**l**) $\rho = -0.4$, $pc = 0.75$, $\frac{N_A}{N} = 0.30$

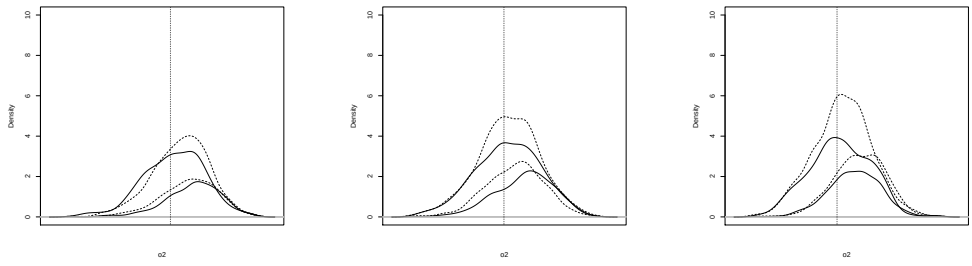


(**m**) $\rho = 0$, $pc = 0.25$, $\frac{N_A}{N} = 0.30$ (**n**) $\rho = 0$, $pc = 0.50$, $\frac{N_A}{N} = 0.30$ (**o**) $\rho = 0$, $pc = 0.75$, $\frac{N_A}{N} = 0.30$

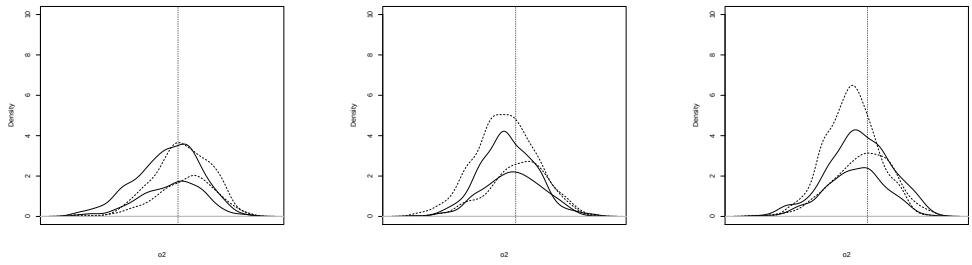
Figure 2. Densities of $\widehat{\beta}_1$ for some values of ρ, $N_A/N = 0.30$ and three sampling proportions of cases pc (SSRS: dashed lines; GE: continuous line; $n = 2000$ for the two flatter distributions; $n = 10,000$ for the two more sharpened distributions).



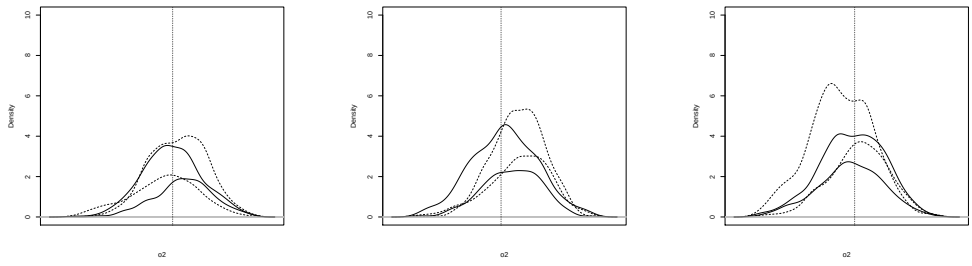
(**a**) $\rho = 0.8$, $pc = 0.25$, $\frac{N_A}{N} = 0.04$ (**b**) $\rho = 0.8$, $pc = 0.50$, $\frac{N_A}{N} = 0.04$ (**c**) $\rho = 0.8$, $pc = 0.75$, $\frac{N_A}{N} = 0.04$



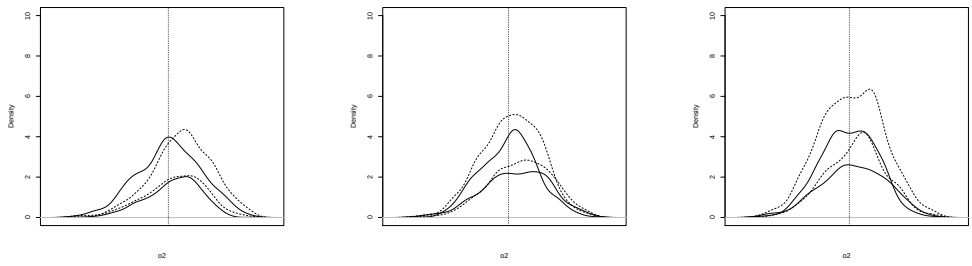
(**d**) $\rho = -0.8$, $pc = 0.25$, $\frac{N_A}{N} = 0.04$ (**e**) $\rho = -0.8$, $pc = 0.50$, $\frac{N_A}{N} = 0.04$ (**f**) $\rho = -0.8$, $pc = 0.75$, $\frac{N_A}{N} = 0.04$



(**g**) $\rho = 0.4$, $pc = 0.25$, $\frac{N_A}{N} = 0.04$ (**h**) $\rho = 0.4$, $pc = 0.50$, $\frac{N_A}{N} = 0.04$ (**i**) $\rho = 0.4$, $pc = 0.75$, $\frac{N_A}{N} = 0.04$

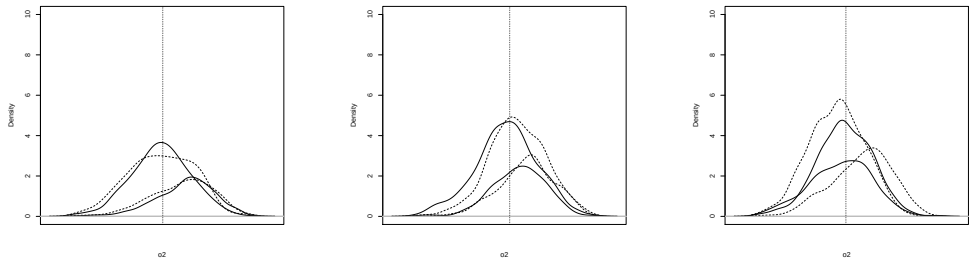


(**j**) $\rho = -0.4$, $pc = 0.25$, $\frac{N_A}{N} = 0.04$ (**k**) $\rho = -0.4$, $pc = 0.50$, $\frac{N_A}{N} = 0.04$ (**l**) $\rho = -0.4$, $pc = 0.75$, $\frac{N_A}{N} = 0.04$

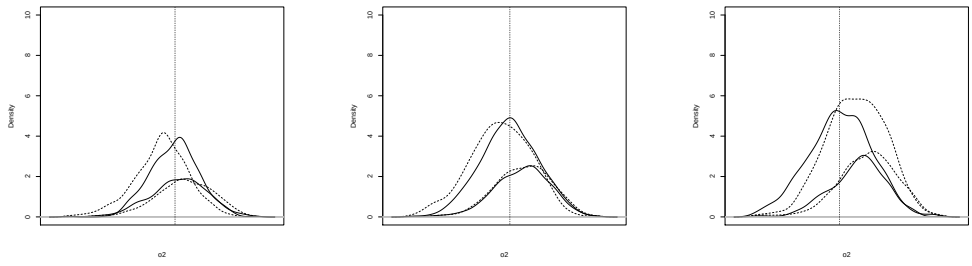


(**m**) $\rho = 0$, $pc = 0.25$, $\frac{N_A}{N} = 0.04$ (**n**) $\rho = 0$, $pc = 0.50$, $\frac{N_A}{N} = 0.04$ (**o**) $\rho = 0$, $pc = 0.75$, $\frac{N_A}{N} = 0.04$

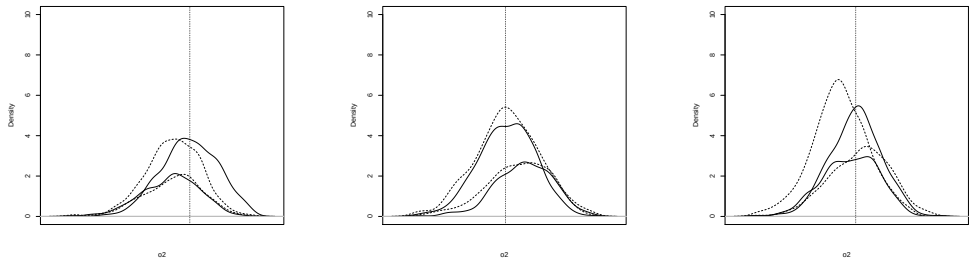
Figure 3. Densities of $\widehat{\beta}_2$ for some values of ρ, $N_A/N = 0.04$ and three sampling proportions of cases pc (SSRS: dashed lines; GE: continuous line; $n = 2000$ for the two flatter distributions; $n = 10,000$ for the two more sharpened distributions).



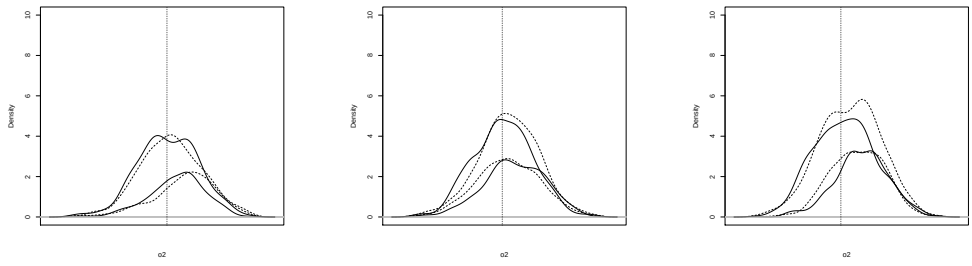
(**a**) $\rho = 0.8$, $pc = 0.25$, $\frac{N_A}{N} = 0.30$ (**b**) $\rho = 0.8$, $pc = 0.50$, $\frac{N_A}{N} = 0.30$ (**c**) $\rho = 0.8$, $pc = 0.75$, $\frac{N_A}{N} = 0.30$



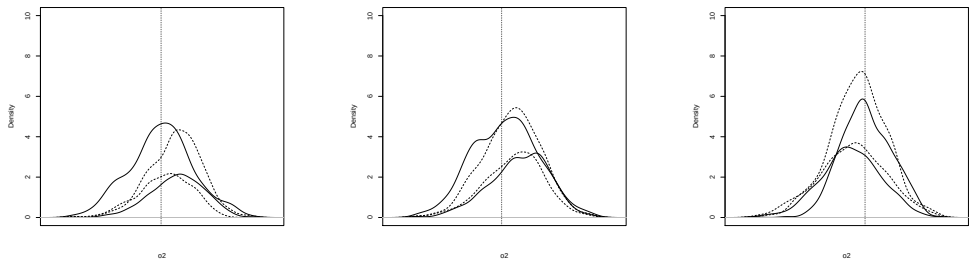
(**d**) $\rho = -0.8$, $pc = 0.25$, $\frac{N_A}{N} = 0.30$ (**e**) $\rho = -0.8$, $pc = 0.50$, $\frac{N_A}{N} = 0.30$ (**f**) $\rho = -0.8$, $pc = 0.75$, $\frac{N_A}{N} = 0.30$



(**g**) $\rho = 0.4$, $pc = 0.25$, $\frac{N_A}{N} = 0.30$ (**h**) $\rho = 0.4$, $pc = 0.50$, $\frac{N_A}{N} = 0.30$ (**i**) $\rho = 0.4$, $pc = 0.75$, $\frac{N_A}{N} = 0.30$

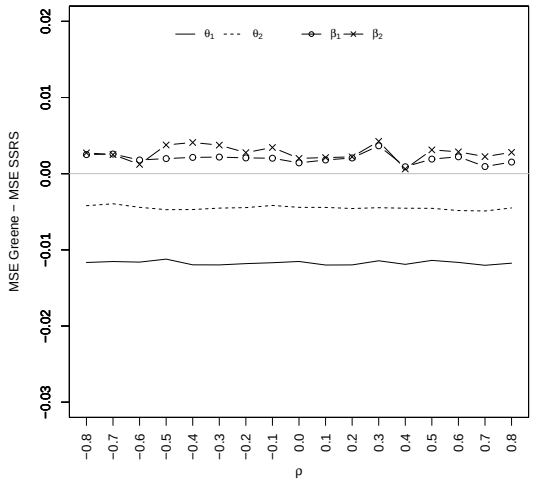


(**j**) $\rho = -0.4$, $pc = 0.25$, $\frac{N_A}{N} = 0.30$ (**k**) $\rho = -0.4$, $pc = 0.50$, $\frac{N_A}{N} = 0.30$ (**l**) $\rho = -0.4$, $pc = 0.75$, $\frac{N_A}{N} = 0.30$

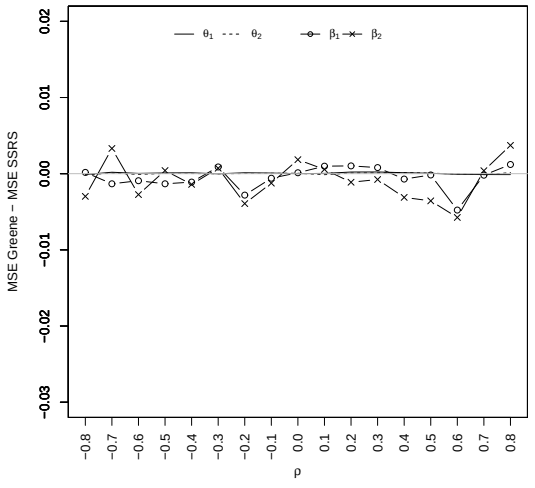


(**m**) $\rho = 0$, $pc = 0.25$, $\frac{N_A}{N} = 0.30$ (**n**) $\rho = 0$, $pc = 0.50$, $N_A/N = 0.30$ (**o**) $\rho = 0$, $pc = 0.75$, $\frac{N_A}{N} = 0.30$

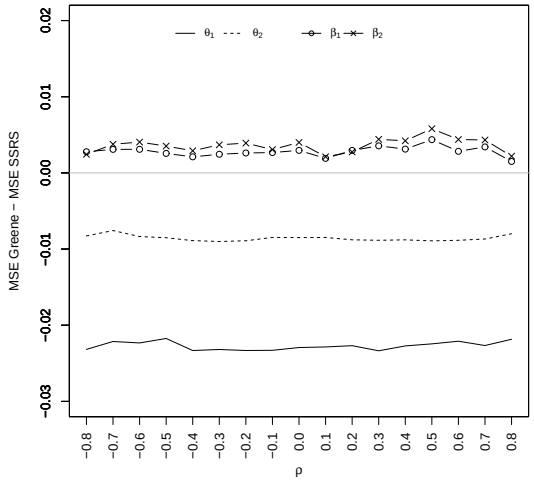
Figure 4. Densities of $\widehat{\beta}_2$ for some values of ρ, $N_A/N = 0.30$ and three sampling proportions on cases pc (SSRS: dashed lines; GE: continuous line; $n = 2000$ for the two flatter distributions; $n = 10,000$ for the two more sharpened distributions).



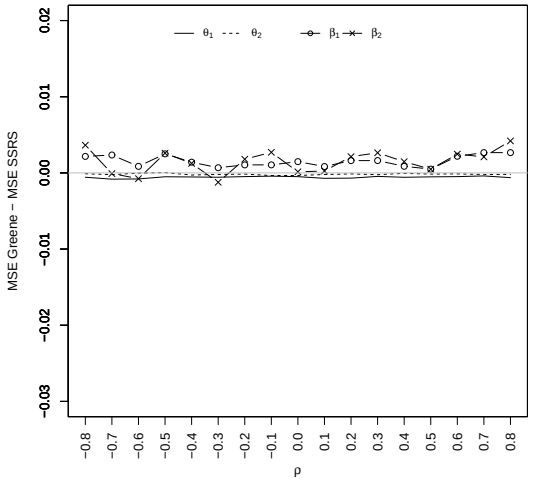
(**a**) $n = 2000$; $pc = 0.25$; $N_A/N = 0.04$



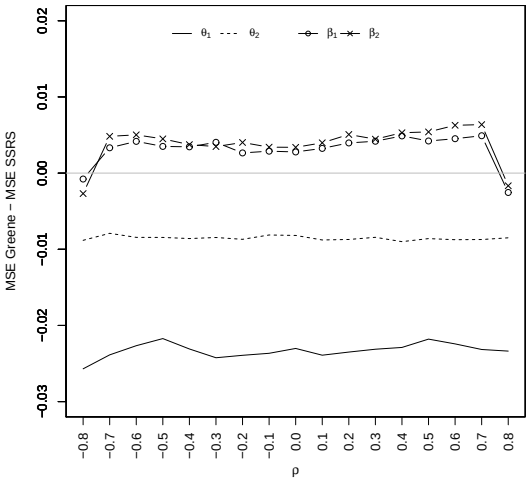
(**b**) $n = 2000$; $pc = 0.25$; $N_A/N = 0.3$



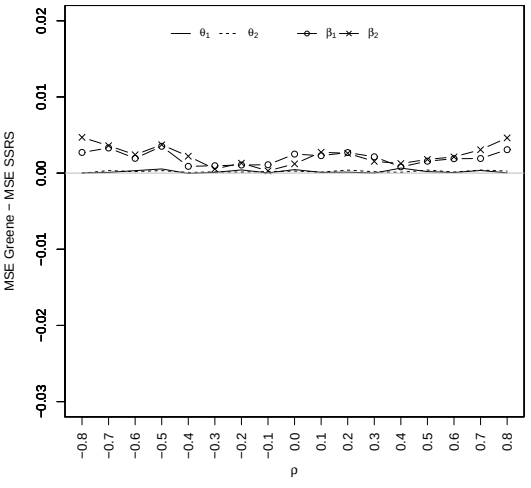
(**c**) $n = 2000$; $pc = 0.50$; $N_A/N = 0.04$



(**d**) $n = 2000$; $pc = 0.50$; $N_A/N = 0.3$

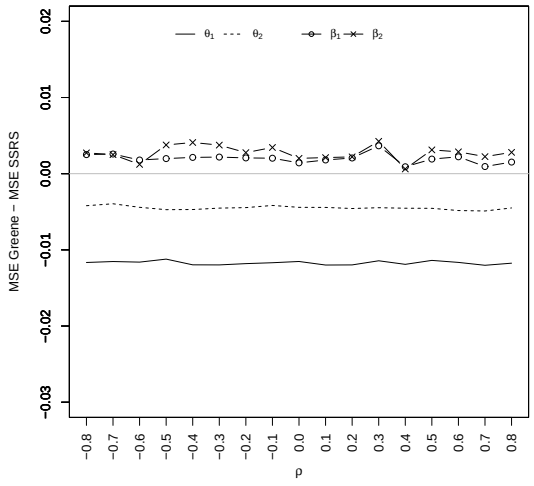


(**e**) $n = 2000$; $pc = 0.75$; $N_A/N = 0.04$

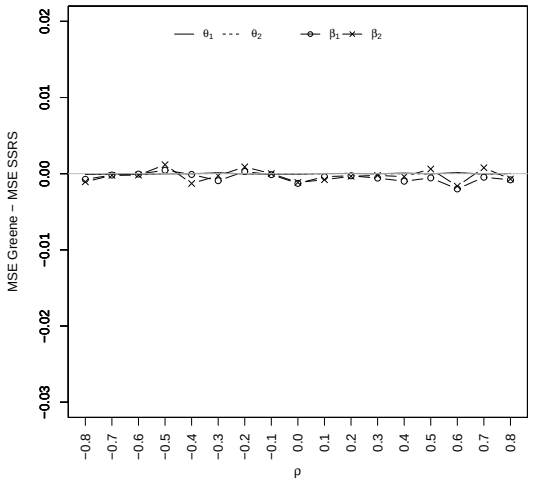


(**f**) $n = 2000$; $pc = 0.75$; $N_A/N = 0.3$

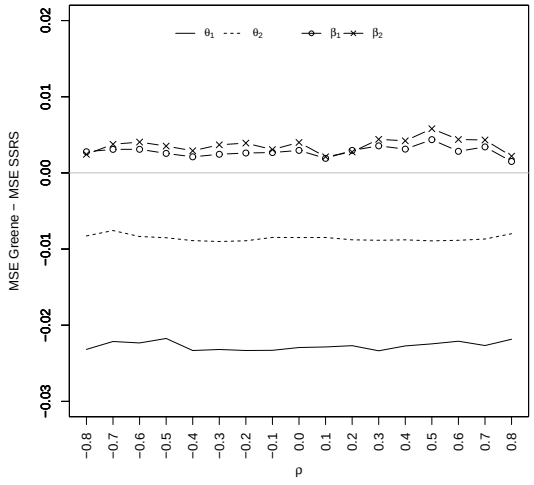
Figure 5. Performance comparison: $MSE_{GE} - MSE_{SSRS}$ for $n = 2000$, for three sampling proportions of cases (pc) and for two proportions (N_A/N) of uncensored observations.



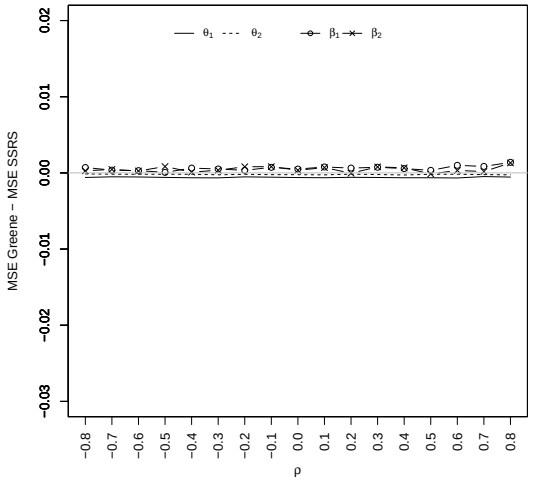
(**a**) $n = 10,000$; $pc = 0.25$; $N_A/N = 0.04$



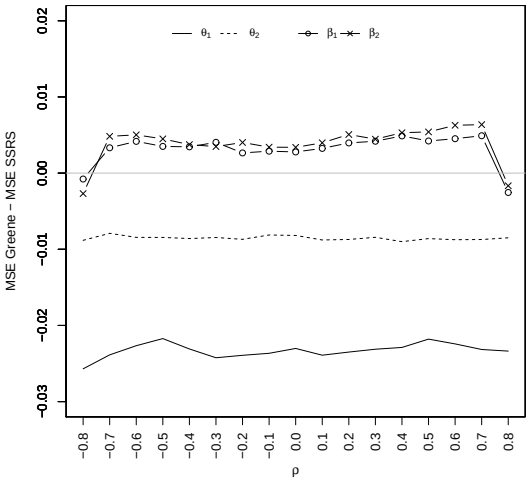
(**b**) $n = 10,000$; $pc = 0.25$; $N_A/N = 0.3$



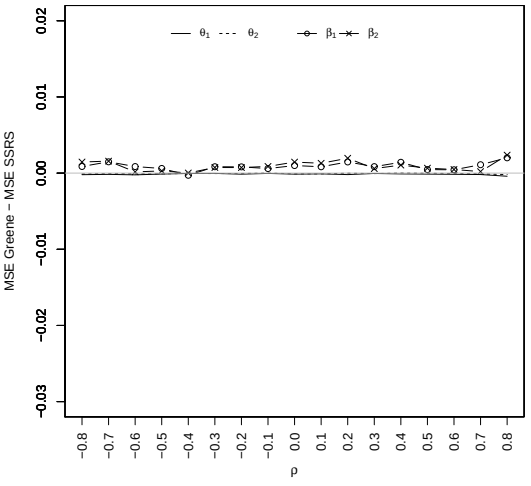
(**c**) $n = 10,000$; $pc = 0.50$; $N_A/N = 0.04$



(**d**) $n = 10,000$; $pc = 0.50$; $N_A/N = 0.3$



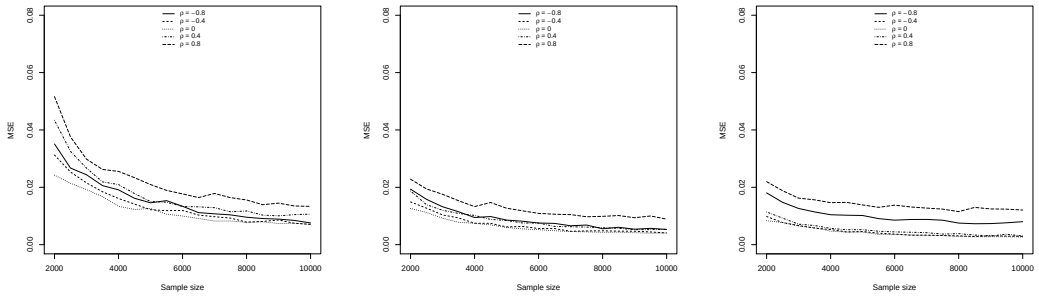
(**e**) $n = 10,000$; $pc = 0.75$; $N_A/N = 0.04$



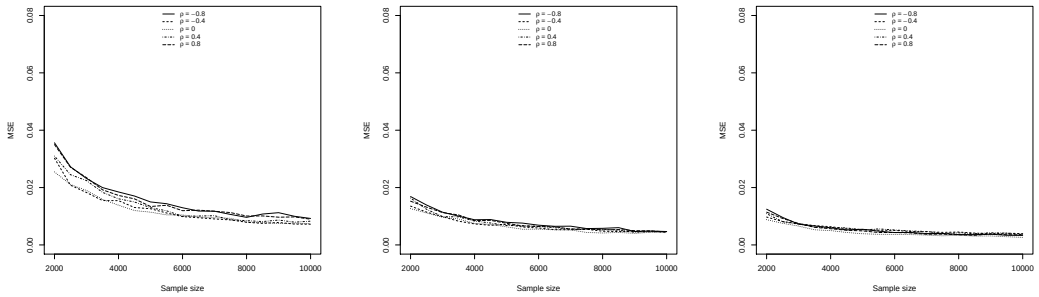
(**f**) $n = 10,000$; $pc = 0.75$; $N_A/N = 0.3$

Figure 6. Performance comparison: $MSE_{GE} - MSE_{SSRS}$ for $n = 10,000$, for three sampling proportions of cases (pc) and for two proportions (N_A/N) of uncensored observations.

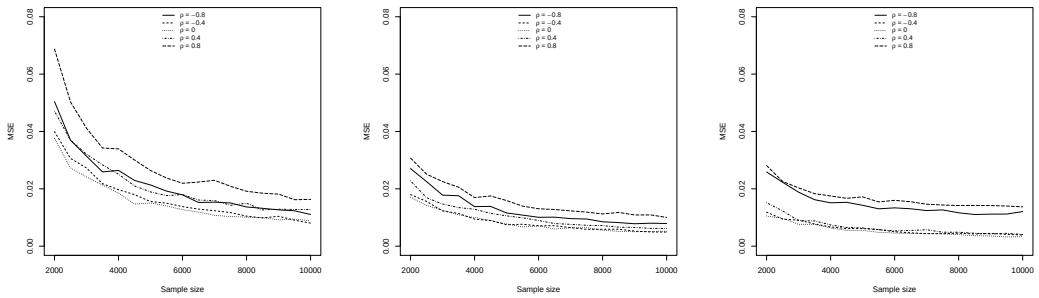
Finally, in Figure 7, we present the trend of the MSEs as the sample size increases, to have a first idea of the possible consistency of our estimator. Clearly, the property of consistency can only be proven analytically. The evaluation has been made only for the SSRS estimator, as it is proven that GE is consistent (Greene 2008; Manski and Lerman 1977). As usual, we consider the three sampling percentages and we compute the MSE for $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\theta}_1$ and $\hat{\theta}_2$. Once again, to save space, we present only the former, but all are available upon request.



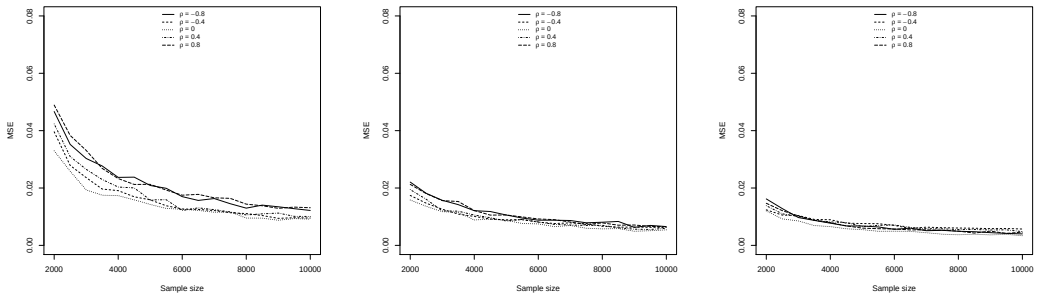
(a) $\hat{\beta}_1$, $pc = 0.25$, $N_A/N = 0.04$ **(b)** $\hat{\beta}_1$, $pc = 0.50$, $N_A/N = 0.04$ **(c)** $\hat{\beta}_1$, $pc = 0.75$, $N_A/N = 0.04$



(d) $\hat{\beta}_1$, $pc = 0.25$, $N_A/N = 0.30$ **(e)** $\hat{\beta}_1$, $pc = 0.50$, $N_A/N = 0.30$ **(f)** $\hat{\beta}_1$, $pc = 0.75$, $N_A/N = 0.30$



(g) $\hat{\beta}_2$, $pc = 0.25$, $N_A/N = 0.04$ **(h)** $\hat{\beta}_2$, $pc = 0.50$, $N_A/N = 0.04$ **(i)** $\hat{\beta}_2$, $pc = 0.75$, $N_A/N = 0.04$



(j) $\hat{\beta}_2$, $pc = 0.25$, $N_A/N = 0.30$ **(k)** $\hat{\beta}_2$, $pc = 0.50$, $N_A/N = 0.30$ **(l)** $\hat{\beta}_2$, $pc = 0.75$, $N_A/N = 0.30$

Figure 7. MSE behavior for SSRS estimator of β_1 and β_2, as the sample size increases.

As suggested by a referee, we made further simulations to study the behavior of the two estimators when the true generating model of the disturbances is not normal, while the likelihood function is based on the probabilities specified in Equations (9)–(11) (normal distribution). In particular, the generating model considered here is a bivariate skew-T; we then simulated 500 samples of size $n = 2000$ and $n = 10,000$, $\rho \in 0, \pm 0.4, \pm 0.8$, $pc \in 0.25, 0.50, 0.75$ and $N_A/N \in 0.04, 0.15, 0.30$. The behavior of the two estimators remains comparable in terms of bias and mean square error. In particular, comparing the mean square errors, GE outperforms SSRS in the estimation of the selection equation; the situation is often reversed for the outcome equation. Furthermore, to compare the robustness to distribution mispecification of the two estimators, we computed the ratio between the bias obtained when the true generating model is a skew-T and the bias obtained when the model is correctly specified: the closer to one the ratio (in absolute value), the more robust the estimator. It emerges that the SSRS estimator is more robust than GE, especially for the selection equation parameters. All results are available upon request.

4. Application on Real Data: Estimation of Credit Scoring

In the following, we consider the problem of estimating the risk of a loan default for credit-card holders. The population of interest is that of loan applicants, for whom some requests are approved and some are rejected according to their loan default risk.

The idea is to use the model to assign a default probability to a random individual who applies for a loan, but the only information that exists about default probabilities comes from previous loan recipients. The problem is that the probability of default for the overall population of applicants is not necessarily the same as for those who have already received a credit card (in fact, we expect that the probability for the whole population is greater than that of cardholders). To avoid this kind of selection bias, we need to consider the selection mechanism explicitly, that is a selection equation which explains the cardholder status.

The dataset considered here[2] contains a subset of the covariates used in Greene (2008). On $n = 13,444$ cardholders, the following characteristics are measured: cardholder status (CH), taking 1 if the application for a credit card was accepted and 0 if not; default status (D), taking 1 if defaulted and 0 if not (observed only when CH = 1, that is for 10,499 observations); age in years plus twelfths of a year (Age); number of dependents (Adepcnt); months living at current address (Acadmos); number of major derogatory reports (Mjrg); number of minor derogatory reports (Mndrg); owner or tenant (Ownrent), taking 1 if the applicant owns his/her home, 0 if renting; monthly income in US dollars divided by 100 (Income); self employment status (Selfempl), taking 1 if self employed, 0 if not; and ratio of monthly credit card expenditure to yearly income (ExpInc).

We used the following specification:

$$D_i = \beta_0 + \beta_1 Adepcnt_i + \beta_2 Income_i + \beta_3 ExpInc_i + \epsilon_{1i}$$
$$CH_i = \theta_0 + \theta_1 Income_i + \theta_2 Ownrent_i + \theta_3 Acadmos_i + \theta_4 Selfempl_i + \theta_5 Mjdrg_i + \theta_6 Mndrg_i + \epsilon_{2i}$$

and we estimated the parameters by SSRS and GE estimators. The results, shown in Table 1, are quite similar for the outcome equation (the one of greater interest because it gives the credit scoring), where the signs are coherent with expectations.

To evaluate the predictive performance of the two models, we computed the confusion matrices (see Table 2), which allowed us to compute the percentages of correct predictions.

---

[2] The data are available in Table 7.3 at http://people.stern.nyu.edu/wgreene/Text/econometricanalysis.htm.

Table 1. Parameter estimates.

	SSRS				**GE**		
	beta	se	pval		beta	se	pval
Age	0.047	0.002	0.000	Age	−0.006	0.001	0.000
Income	−0.002	0.001	0.067	Income	0.018	0.001	0.000
Ownrent	0.039	0.031	0.208	Ownrent	−0.372	0.026	0.000
Acadmos	−0.002	0.000	0.000	Acadmos	0.002	0.000	0.000
Selfempl	0.005	0.060	0.928	Selfempl	0.216	0.045	0.000
Mjdrg	−0.440	0.013	0.000	Mjdrg	−0.385	0.013	0.000
Mndrg	0.393	0.019	0.000	Mndrg	−0.164	0.016	0.000
Cons	−0.144	0.052	0.005	Cons	−0.166	0.042	0.000
Adepcnt	0.430	0.014	0.000	Adepcnt	0.059	0.017	0.001
Income	−0.070	0.002	0.000	Income	−0.012	0.002	0.000
ExpInc	−0.153	0.157	0.329	ExpInc	0.170	0.181	0.349
Cons	−0.048	0.050	0.342	Cons	−0.562	0.050	0.000

Table 2. Confusion matrices: observed vs predicted for non cardholders ($\overline{CH}$), non defaulted ($CH \cap \overline{D}$) and defaulted ($CH \cap D$).

	SSRS	Predicted		
		$\overline{CH}$	$CH \cap \overline{D}$	$CH \cap D$
Observed	$\overline{CH}$	484	2302	159
	$CH \cap \overline{D}$	58	9210	235
	$CH \cap D$	6	947	43
	GE	Predicted		
		$\overline{CH}$	$CH \cap \overline{D}$	$CH \cap D$
Observed	$\overline{CH}$	2747	198	0
	$CH \cap \overline{D}$	7984	1519	0
	$CH \cap D$	897	99	0

From these matrices, it is easy to obtain the percentages of correct predictions for both estimation procedures. In particular, the overall percentages are 0.724 and 0.317 for SSRS and GE, respectively, showing a strong dominance of the former estimation method over the latter. This dominance is even more evident in the percentages conditioned on the default status: for non defaulters, the percentage is 0.969 for SSRS and 0.160 for GE, while, for defaulters, it is 0.043 for SSRS and 0 for GE. On the other hand, GE outperforms SSRS in the prediction of non cardholder status: the values are 0.164 for SSRS and 0.933 for GE.

5. Conclusions

In this paper, we propose a method for estimating the regression coefficients in binary response models with sample selection, when a censoring mechanism intervenes to make a vast majority of units unobservable. We derive the likelihood function analytically, taking advantage of the response-based sampling framework and find that it is a weighted version of Heckman's likelihood.

A simulation study highlights that the finite sample performance of the point estimators are very satisfactory even when compared to a similar estimator (Greene 2008), which in turn assumes knowledge of the population proportion of occurrences in the outcome equation. Specifically, our estimator slightly outperforms Greene's in estimating the parameters of the outcome equation (which are, generally speaking, those of greater interest), especially in the case of a severe censoring mechanism ($NA/N = 0.04$). Under this censoring scenario, the situation is reversed for the selection equation. When the censoring is less severe, the two estimators are substantially equivalent for the selection equation, while in general ours still outperforms Greene's for the outcome.

An analogous result is obtained in an empirical analysis aimed to estimate the risk of loan default: our estimator performs better in the prediction of the dependent variable of the outcome equation (i.e., the default status), while Greene's does better in the selection equation (i.e., the prediction of the cardholder status).

The fact that the best results refer to the outcome equation is an advantage, since it is generally the one of greater interest. The main result is that we do not require knowledge of the true prevalence of occurrences, that is, our estimator can be successfully used when N_{1A} and N_{0A} are unknown.

Future research will be devoted to the problems arising from measurement errors, either in the dependent variable or in the covariates of the outcome equation. In the first case, the problem is obviously simpler and a modified version of the likelihood can take into account the bias arising from this kind of error. In the second case, an endogeneity issue arises and it is necessary to consider further information, for example, instrumental variables.

Acknowledgments: This work has been funded by Sapienza University of Rome, grant 000041-17-RDB-ATENEO2016-AREZZO. The authors wish to thank Judith Turnbull for her precious help with English corrections and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GE	Greene's estimator
SSRS	sample-selection response-based sampling estimator
MSE	Mean Square Error

Appendix A. Derivation of the Likelihood Function

Making use of the Bayes rule, we can rewrite the likelihood function (5) for our sample as:

$$\prod_{i=1}^{n} f(\boldsymbol{X_i}|A_i, Y_i, S_i = 1) = \prod_{i=1}^{n} f\left(\boldsymbol{X_i}|S_i = 1\right) \cdot \frac{P(Y_i, A_i|S_i = 1, \boldsymbol{X_i})}{P(Y_i, A_i|S_i = 1)} \tag{A1}$$

The ratio on the right hand side represents the contribution to the likelihood jointly given by Y_i and A_i. However, since Y_i is observable only when $A_i = 1$, the contribution to the likelihood when $A_i = 0$ is only given by:

$$\begin{aligned} \frac{P(A_i = 0|S_i = 1, \boldsymbol{X_i})}{P(A_i = 0|S_i = 1)} &= \frac{P(A_i = 0|\boldsymbol{X_i})P(S_i = 1|A_i = 0, \boldsymbol{X_i})/P(S_i = 1|\boldsymbol{X_i})}{P(A_i = 0|S_i = 1)} \\ &= \frac{P(S_i = 1|A_i = 0)P(A_i = 0|\boldsymbol{X_i})}{P(A_i = 0|S_i = 1)P(S_i = 1)} = \frac{\frac{n_{\bar{A}}}{N_{\bar{A}}} \cdot (1 - {}_A\pi(\boldsymbol{X_i}))}{\frac{n_{\bar{A}}}{n} \cdot \frac{n}{N}} \\ &= \frac{N}{N_{\bar{A}}} \cdot (1 - {}_A\pi(\boldsymbol{X_i})) \end{aligned} \tag{A2}$$

where, in the second ratio of Equation (A2) ,we used the Bayes rule; furthermore, the probability of being in the sample does not depend on the covariates. More precisely, $P(S_i = 1|\boldsymbol{X_i}, Y_i, A_i = a_i) = P(S_i = 1|A_i = a_i)$, with $a_i = 0, 1$.

For the probabilities in the third ratio, we put:

- $P(S_i = 1|A_i = 0) = n_{\bar{A}}/N_{\bar{A}}$;
- $P(A_i = 0|\boldsymbol{X_i}) = 1 - P(A_i = 1|\boldsymbol{X_i}) = 1 - {}_A\pi(\boldsymbol{X_{2i}})$;
- $P(A_i = 0|S_i = 1) = n_{\bar{A}}/n$;
- $P(S_i = 1) = n/N$.

On the other hand, when $A_i = 1$, we have:

$$\frac{P(Y_i = y_i, A_i = 1|S_i = 1, \boldsymbol{X_i})}{P(Y_i = y_i, A_i = 1|S_i = 1)} = \frac{P\left(S_i = 1|A_i = 1\right) \cdot P\left(Y_i = y_i, A_i = 1|\boldsymbol{X_i}\right)}{P\left(S_i = 1\right) \cdot P\left(Y_i = y_i, A_i = 1|S_i = 1\right)} =$$
$$\frac{\frac{n_A}{N_A} \cdot P\left(Y_i = y_i, A_i = 1|\boldsymbol{X_i}\right)}{\frac{n}{N} \cdot \frac{n_{y_i A}}{n}} = \frac{N}{N_A} \cdot \frac{n_A}{n_{y_i A}} \cdot P\left(A_i = 1|\boldsymbol{X_i}\right) P\left(Y_i = y_i|A_i = 1, \boldsymbol{X_i}\right) =$$
$$\frac{N}{N_A} \cdot \frac{n_A}{n_{y_i A}} \cdot {}_A\pi(\boldsymbol{X_{2i}}) \cdot [{}_Y\pi(\boldsymbol{X}_{1i})]^{y_i} \cdot [1 - {}_Y\pi(\boldsymbol{X}_{1i})]^{1-y_i} \tag{A3}$$

where $y_i = 0, 1$; hence, $n_{y_i A}$ becomes n_{1A} or n_{0A}, according to the Y_i value, and indicates the number of individuals in the sample for which it is observed, respectively, $Y_i = 1$ or $Y_i = 0$.

For the probabilities in Equation (A3), we put:

- $P(A_i = 1|\boldsymbol{X_i}) = {}_A\pi(\boldsymbol{X_{2i}})$, as before;
- $P(Y_i = y_i|A_i = 1, \boldsymbol{X_i})$ can be specified as a binary response model for Y_i in the covariates $\mathbf{X_{1i}}$, so that $P(Y_i = 1|A_i = 1, \boldsymbol{X_i}) = {}_Y\pi(\boldsymbol{X_{1i}})$, while $P(Y_i = 0|A_i = 1, \boldsymbol{X_i}) = 1 - {}_Y\pi(\boldsymbol{X_{1i}})$; and
- $P(S_i = 1|A_i = 1) = n_A/N_A$.

References

Boyes, William J., Dennis L. Hoffman, and Stuart A. Low. 1989. An econometric analysis of the bank credit scoring problem. *Journal of Econometrics* 40: 3–14.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Dubin, Jeffrey A., and Douglas Rivers. 1989. Selection bias in linear regression, logit and probit models. *Sociological Methods & Research* 18: 360–90.

Greene, William. 2008. A statistical model for credit scoring. In *Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction. Quantitative Methods for Applied Economics and Business Research*. Edited by Stewart Jones and David A. Hensher. Cambridge: Cambridge University Press, Chapter 1, pp. 14–43.

Grier, Kevin B., Michael C. Munger, and Brian E. Roberts. 1994. The determinants of industry political activity, 1978–1986. *American Political Science Review* 88: 911–26.

Heckman, James J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–62.

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*, 3rd ed. Hoboken: John Wiley & Sons.

Ingram, Gregory K. 1999. Determinants of Motorization and Road Provision. Policy Research Working Paper No. WPS2042, The World Bank, Washington, DC, USA.

Jacobs, David, and Robert M. O'Brien. 1998. The determinants of deadly force: A structural analysis of police violence. *American Journal of Sociology* 103: 837–62.

Kayser, Hilke A. 2000. Gasoline demand and car choice. estimating gasoline demand using household information. *Energy Economics* 22: 331–48.

Lee, Lung-Fei. 2003. Self-Selection. In *A Companion to Theoretical Econometrics*. Edited by Badi H. Baltagi. Malden: Blackwell Publishing, pp. 383–409.

Manski, Charles, and Steven R. Lerman. 1977. The estimation of choice probabilities from choice based samples. *Econometrica* 45: 1977–88.

Mroz, Thomas M., and Yaraslau V. Zayats. 2008. Arbitrarily normalized coefficients, information sets and false reports of biases in binary outcome models. *Review of Economics and Statistics* 90: 406–13.

Van de Ven, Wynand, and Bernard Van Praag. 1981. The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics* 17: 229–52.

Vella, Francis. 1998. Estimating models with sample selection bias: A survey. *The Journal of Human Resources* 33: 127–69.