"I listen to what I am not told"

A.H. Parola

Truncation Bias *

Nir Billfeld† Moshe Kim‡

July 19, 2016

In the case of truncation, which is the widespread phenomenon plaguing the majority of all fields of empirical research, the observed data distribution function is truncated and related to participants' covariates only, rendering Heckman's seminal and known correction procedure not implementable. Thus, for the correction of endogenous selectivity bias propagated by truncation we introduce a new methodology that recovers the unobserved part of the data distribution function, using only its observed truncated part. The correlation patterns among the non-participants' covariates (which are all functions of the recovered non-participants' density function) are recovered as well. The rationale underlying the ability to recover the unobserved complete density function from the observed truncated density function relies on the fact that the latter is obtained by conditioning the former on the selection rule. Consequently, the parameters set which characterizes the truncated density function contains all the parameters characterizing the unobserved non-truncated density function. Thus, it is possible to characterize the unobserved non-participants' density function in terms of the parameters estimated using the truncated data soley. Once this unobserved part is recovered one can estimate the selection rule equation for the hazard rate calculation as if the full sample consisting of both participants and non-participants is observable. Monte-Carlo simulations attest to the high accuracy of the estimates and above conventional $\sqrt{n}$ consistency.

Keywords: Selectivity bias correction, Truncated Probit.

*We would like to thank Shaul K.Bar-Lev, Bill Green, and Ben Reiser, for their many constructive and valuable comments. This is a revised version of SSRN no. 2786263 with same title.

†University of Haifa nbillfeld@staff.haifa.ac.il

‡University of Haifa kim@econ.haifa.ac.il

1. Introduction

The Heckman (1976, 1979) correction for selectivity bias has become an important tool routinely employed in the analysis of censored data. Problems of selection bias are of great importance in research in general and in applied economic research in particular. The problem arises when observations are not randomly selected into the sample thus potentially propagating endogenous selectivity bias. The aforementioned bias arises from a correlation between two error terms, one of which is generated by the participation equation and the other is generated by the substantive equation of interest.

Heckman (1976) introduced a two-step procedure correcting for the selectivity bias. The first step is based on estimation of a Probit regression for the propensity to participate in the sample in order to generate the inverse Mills ratio (the hazard rate), which in the second step, is inserted in the equation of interest as a covariate. In order to implement the Heckman's correction procedure, the Probit model of participation must be identified, meaning that there is a need to observe *both participants and non-participants*. This, however is rarely the case as the data available to researchers are truncated rather than censored and as such no information regarding non participant observations exists.

This paper develops a methodology to correct for selectivity bias in *truncated* samples stemming from endogenous selection rule. We formulate the relationship among the observed truncated density function, the selection rule (which depends on unknown estimable parameters) and the unobserved non-truncated density function to be recovered using the underlying data. Specifying an indicator variable $\mathcal{S} = 1$ for being selected, the probability to be selected for each observation with a covariates vector $(z_1, .., z_p)$ is $P_{\boldsymbol{\gamma}}(\mathcal{S} = 1|z_1, ..., z_p)$. We assume that the observations in the non-truncated sample are randomly drawn from a multivariate normal distribution function, with the density function $f_{\boldsymbol{\eta}}(z_1, .., z_p)$. However, only the selected observations are observed in the truncated sample. Hence, one can characterize (using the same unknown parameters) the truncated density function $f_{\boldsymbol{\theta}}(z_1, ..., z_p|\mathcal{S} = 1)$, where each parameters vector $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ represents the selection rule's and the non-truncated density function's unknown parameters respectively. The entire set of parameters to be estimated is denoted by $\boldsymbol{\theta} = [\boldsymbol{\eta}, \boldsymbol{\gamma}]^T$. The proportion of participants in the complete sample, $P_{\boldsymbol{\theta}}(\mathcal{S} = 1)$, is a function of $\boldsymbol{\theta}$ and describes the probability of a randomly chosen observation (from the complete sample) to be selected.[1] Thus, for a participant with the covariates vector $(z_1, ..., z_p)$, the relationship between the observed truncated density function (the posterior) and the unobserved non-truncated density function (the prior) given Bayes' rule is:

$$f_{\boldsymbol{\theta}}(z_1, ..., z_p|\mathcal{S} = 1) = \frac{f_{\boldsymbol{\eta}}(z_1, ..., z_p)P_{\boldsymbol{\gamma}}(\mathcal{S} = 1|z_1, ..., z_p)}{P_{\boldsymbol{\theta}}(\mathcal{S} = 1)}. \tag{1.1}$$

[1]The probability of a randomly chosen observation from the non-truncated sample to be selected is: $P_{\boldsymbol{\theta}}(\mathcal{S} = 1) = \int_{v_1=-\infty}^{\infty} ... \int_{v_p=-\infty}^{\infty} P_{\boldsymbol{\gamma}}(\mathcal{S} = 1|v_1, ...v_p) f_{\boldsymbol{\eta}}(v_1, .., v_p) dv_1 ... dv_p$.

The density function product $\prod_{i=1}^{n} f_{\boldsymbol{\theta}}(z_{1i},..,z_{pi}|\mathcal{S}_i=1)$ based on (1.1) thus reflects the conditional (on participation) density function which generated the truncated sample of participants. Using this density product function we search for the parameters vector $\boldsymbol{\theta}$ which maximizes the likelihood to obtain the truncated sample we observe.[2]

In a similar fashion to (1.1), the conditional density function of a non-participant with the covariates vector $(z_1,...,z_p)$ given Bayes' rule is:

$$f_{\boldsymbol{\theta}}(z_1,...,z_p|\mathcal{S}=0)=\frac{f_{\boldsymbol{\eta}}(z_1,...,z_p)P_{\boldsymbol{\gamma}}(\mathcal{S}=0|z_1,...,z_p)}{P_{\boldsymbol{\theta}}(\mathcal{S}=0)}. \tag{1.2}$$

The unobserved non-participants' truncated density function in (1.2) depends on the same unknown parameters as the observed participants' truncated density function in (1.1). Thus, once estimating these common unknown parameters (using the sample of participants) it is possible to characterize the unobserved non-participants' covariates correlation patterns and their vector of expectations which are all functions of the unobserved non-participants' truncated density function.

The rationale underlying the ability to recover the unobserved complete density function from the observed truncated density function relies on the fact that the latter is obtained by conditioning the former on the selection rule.[3]

In order to allow for dichotomous variables, which are routinely used in the social sciences, we introduce a cohorts set $\mathcal{J}\equiv\{1,2,...,J-1,J\}$ consisting of J mutually exclusive cohorts such that each observation i belongs to a specific one cohort $j\in\mathcal{J}$. The number of participants and the number of observations in the j'th cohort are denoted by n_j and N_j respectively. Nevertheless, only the former is observed in the truncated sample, while the aggregate share of each cohort in the population (or in the complete data) is unknown. Additionally, the total number of observations $N=\sum_{j=1}^{J}N_j$ in the complete sample is unknown, but will be recovered post estimation, conditional on the fact the selection rule is endogenous and not random.[4]

Denote the observations indices belonging to j'th cohort by I_j for each $j\in\mathcal{J}$, and a dichotomous random variable for being a memeber in the $j'th$ cohort,

[2] This parameter vector contains both the $\boldsymbol{\gamma}$ parameters vector for the selection rule and the parameters vector $\boldsymbol{\eta}$ characterizing the non-truncated density function.

[3] Cosslett (1981a,b) introduced a non-parametric estimation methodology for an endogenous selectivity bias correction in a multinomial discrete choice model in cases where all the alternatives are included in the sample or alternatively by completing the data (enriching the sample) using an additional sample consisting of alternatives that are underrepresented in the first one (non-participants). However, we are dealing with a truncated sample, in which the non-participants' data do not exist and thus cannot utilize any enriched data.

[4] If in addition to the endogenous selection rule, there is also a random selection process (focusing only on a random sub-sample of the participants), it is still possible to estimate the model, but not to recover the various shares of the sub-populations.

defined as follows:

$$D_{ji} = \begin{cases} 1, & \text{if } i \in I_j \\ 0, & \text{if } i \notin I_j \end{cases}. \tag{1.3}$$

It must hold that $\sum_{j=1}^{J} D_{j_i} = 1 \; \forall i$ since all the cohorts are mutually exclusive by construction. For simplicity of notation every cohort $j \in \mathcal{J}$ can be represented by a unique dichotomous variables vector $\boldsymbol{D_j}$ of size $((J-1) \times 1)$ such that all of its elements are zero, except for the j'th element, and $\boldsymbol{D_J}$ is a vector of zeros.

Our methodology treats each group of observations (cohort) as a separate subsample drawn from a unique truncated multivariate distribution function, such that each cohort $j \in \mathcal{J}$ is characterized by a unique parameters set $\boldsymbol{\theta_j} \equiv [\boldsymbol{\eta_j}, \boldsymbol{\gamma_j}]^T$ and $\boldsymbol{\theta} \equiv (\boldsymbol{\theta_1}, ..., \boldsymbol{\theta_J})$ is the entire parameters set to be estimated.

Additionally, the decomposition of the truncated data into cohorts accommodates for the binary dependent variable, participation, to be completely separated by the cohorts[5]. For instance, if the data consist of two cohorts: employed men and unemployed women solely, and one is interested in estimating the contribution of gender to the probability to be employed, conventional binary response models (such as Probit or Logit) cannot estimate the coefficient of gender since unemployment, the binary outcome variable, is completely separated by gender. However, if the complete separation is applied only to the truncated data, reconstructing the data and completing the information regarding the omitted control groups can solve the problem of complete separation. In this example, there are two omitted control groups to be recovered: unemployed men and employed women. Once these omitted control groups are recovered, one can estimate the effect of gender on the probability of being employed as if one utilizes the complete data consisting of all four cohorts.

Formally, the omitted control groups problem in the case of dichotomous variables is modeled using a participation's state variable (two mutually exclusive states such as employed and unemployed) $s_j \in \{0, 1\}$ defined individually for each cohort $j \in \mathcal{J}$. This variable does not vary within groups (cohorts), although it can be varied among cohorts, leading to a complete separation.

The entire data on the continuous covariates in the presence of dichotomous variables, is distributed as a Gaussian mixture, defined as follows:

$$f_{\boldsymbol{\theta}}(\boldsymbol{z}) = \sum_{J=1}^{J} P(\boldsymbol{D} = \boldsymbol{D_j}) f_{\boldsymbol{\theta_j}}(\boldsymbol{z} | \boldsymbol{D} = \boldsymbol{D_j}), \tag{1.4}$$

In practice, the dichotomous variables probability function $P(\boldsymbol{D} = \boldsymbol{D_j})$ is unknown since it represents the proportion of any sub-population in the unobserved complete sample. Thus, we introduce a procedure that does not assume a specific

[5] A complete separation occurs when the outcome variable separates a combination of predictor variables completely (or to a certain degree) (Albert and Anderson, 1984). As a result, the maximum likelihood estimate for the set of completely separate covariates does not exist.

distribution function for the dichotomous variables. The estimation procedure is based on two steps. First, we obtain the density function product in the case of dichotomous variables which is calculated as follows:[6]

$$L = \prod_{J=1}^{J} \prod_{i=1}^{n} \left[P(\boldsymbol{D} = \boldsymbol{D_j} | \mathcal{S}_i = s_j) f_{\boldsymbol{\theta_j}}(\boldsymbol{z} | \mathcal{S}_i = s_j, \boldsymbol{D} = \boldsymbol{D_j}) \right]^{D_{ji}}, \tag{1.5}$$

Once all of the sub-samples' truncated distribution functions are estimated using the observed data, the entire data unobserved complete distribution function is recovered.

The j'th cohort's density function for a participant with the covariates vector $\boldsymbol{z} \equiv [z_1, ..., z_p]^T$ of size $(p \times 1)$ is:

$$f_{\boldsymbol{\theta_j}}(\boldsymbol{z} | \mathcal{S} = 1, \boldsymbol{D} = \boldsymbol{D_j}) = \frac{f_{\boldsymbol{\eta_j}}(\boldsymbol{z} | \boldsymbol{D} = \boldsymbol{D_j}) P_{\boldsymbol{\gamma_j}}(\mathcal{S} = 1 | \boldsymbol{z}, \boldsymbol{D} = \boldsymbol{D_j})}{P_{\boldsymbol{\theta_j}}(\mathcal{S} = 1 | \boldsymbol{D} = \boldsymbol{D_j})}. \tag{1.6}$$

In a similar fashion to equation (1.2), the j'th cohort's density function for a non-participant with the covariates vector $\boldsymbol{z}$ is:

$$f_{\boldsymbol{\theta_j}}(\boldsymbol{z} | \mathcal{S} = 0, \boldsymbol{D} = \boldsymbol{D_j}) = \frac{f_{\boldsymbol{\eta_j}}(\boldsymbol{z} | \boldsymbol{D} = \boldsymbol{D_j}) P_{\boldsymbol{\gamma_j}}(\mathcal{S} = 0 | \boldsymbol{z}, \boldsymbol{D} = \boldsymbol{D_j})}{P_{\boldsymbol{\theta_j}}(\mathcal{S} = 0 | \boldsymbol{D} = \boldsymbol{D_j})}. \tag{1.7}$$

Using this density product function in (1.5) we search for the parameters vector $\boldsymbol{\theta}$ which maximizes the likelihood to obtain the truncated sample we observe. In order to reduce the parameters' dimensionality in case where every two cohorts $j, j' \in J$ satisfies $s_j = s_{j'}$, one can impose restrictions on subsets of the parameters to be the same across cohorts. For instance, imposing a restriction such as $\boldsymbol{\gamma_j} = \boldsymbol{\gamma}$ $\forall j \in \mathcal{J}$, implies that there is no discrimination in the selection rule among cohorts. Yet, the cohorts are differentiated according to their characteristics' distribution function.

The proportion of participants $P(\boldsymbol{D} = \boldsymbol{D_j} | \mathcal{S}_i = s_j)$ which belong to cohort j conditional on being selected and denoted by π_j, can be estimated non-parameterically, as follows:

$$\hat{\pi}_j = \frac{n_j}{\sum_{j=1}^{J} n_j}, \quad \forall j \in \mathcal{J}. \tag{1.8}$$

Since the $\pi_1, ..., \pi_J$ estimators are not functions of the parameters vector $\boldsymbol{\theta}$ to be estimated, one only needs to maximize the simplified version of the likelihood

[6] In cases where there are at least two different cohorts j and j' satisfying $s_j \neq s_{j'}$, the conditional density product in (1.5) is estimated under the assumption of unique parameters set for each cohort in order to simplify the likelihood function. In other cases, one can impose restrictions on the parameters to be the same across cohorts using the aforementaioned density fuction product.

function in (1.5), which is:

$$L^* = \prod_{J=1}^{J} \prod_{i=1}^{n} \left[f_{\boldsymbol{\theta_j}}(\boldsymbol{z}|\mathcal{S}_i = s_j, \boldsymbol{D} = \boldsymbol{D_j}) \right]^{D_{ji}}, \tag{1.9}$$

In the second step (post-estimation) one can calculate the probability for a random observation to participate for each cohort, separately, as follows:

$$\hat{P}_{\boldsymbol{\theta}_j}(\mathcal{S} = 1|\boldsymbol{D} = \boldsymbol{D_j}) = \frac{n_j}{N_j}, \tag{1.10}$$

implying that we obtain the following estimator: $\hat{N}_j = \frac{n_j}{P_{\boldsymbol{\theta}_j}(\mathcal{S} = 1|\boldsymbol{D} = \boldsymbol{D_j})}$ for the total number of observations in the j'th cohort (participants and non-participants).

Then we introduce the following estimator for the proportion of participants in the complete sample:

$$\hat{P}(\mathcal{S} = 1) = \frac{\sum_{j=1}^{J} n_j}{\sum_{j=1}^{J} \hat{N}_j} \tag{1.11}$$

Additionally, an estimator for the proportion of observations in the complete data which belong to the j'th cohort can be obtained, as follows:

$$\hat{P}(\boldsymbol{D} = \boldsymbol{D_j}) = \frac{\hat{N}_j}{\sum_{j=1}^{J} \hat{N}_j} \tag{1.12}$$

We depart from the existing selectivity bias literature (Heckman, 1974, 1976, 1979) which (unlike in the case of truncation) has access to the entire distribution of the data and assumes a joint normal distribution function for the selection model's disturbances. Instead of the disturbances' joint normality distribution assumption, our approach specifies a multivariate normal distribution function (and its density function) depicting the likelihood for each observation to occur in the complete (albeit unobserved non-truncated) data. Then, based on the chosen multivariate normal density function and the participation equation we construct the truncated distribution function (and its density function), showing the likelihood for each observation to occur in the incomplete (truncated) data. We estimate the parameters of that truncated multivariate density function by the maximum likelihood method, such that some of the parameters to be estimated are the Probit regression coefficients and the others are incidental (nuisance) parameters (characterizing the multivariate density). Once these Probit coefficients are estimated by our new approach, we calculate the probability of being sampled for each individual observation, as if we regressed the

conventional Probit model on the entire data, consisting of both participants and non-participants. Based on these predicted probabilities we calculate the inverse Mills ratio in a similar fashion to Heckman's methodology and estimate the equation of interest using the inverse Mills ratio as an additional covariate. This solves the endogenous selectivity bias stemming from truncation.

Due to the generality of the methodology, this approach can be extended for selectivity bias correction based on any binary response model (Probit, Logit, etc). For sake of compatibility with Heckman's correction, we implement our methodology for the case of Probit regression. Additionally, an important contribution of our methodology arises from the information embodied in the incidental parameters which are estimated for the purpose of the recovery of the complete (non-truncated) multivariate normal density function. We show that the incidental parameters can be utilized to 'reverse engineer' the unrestricted data generation process, and as a byproduct recover the correlation patterns among all covariates, and each individual covariates' vector of variance and expectation. Moreover, we can infer the correlation patterns among the non-participants and estimate the proportion of observations which are not included in the sample due to selectivity bias.

The paper is organized in the following way. In the next section we discuss issues concerning selectivity bias in the various fields of research and existing methods available in the literature for its correction. The third section prepares the ground for our methodology to deal with *truncated* samples. At first, we present the main problem of endogenous selectivity and the limitation of common approaches as a motivation for the new approach. We then discuss our approach for dealing with truncated data. The fourth section presents *truncated Probit* regression (in comparison to *non-truncated Probit*), including model's parameters estimation by the maximum likelihood method. The fifth section introduces Monte-Carlo simulations generating multiple random data sets to be used for the estimation of the *truncated Probit*. We compare results which emanate from the non-truncated and truncated data corrected for selectivity bias. The sixth section presents implications for applied empirical work and the seventh section concludes.

2. Discussion

Truncation and its derivative bias is present in virtually all areas of empirical research. In social sciences selectivity bias had been dealt with in variety of studies. In labor economics, e.g., there exists selection bias stemming from the exclusion of the unemployed from the data (Killingsworth et al., 1986). Another instance of bias may exist in the case of '*publication selection bias*' which may arise because of the tendency in virtually all scientific fields to report statistical results that tend to reject the hypothesis of no effect (Ashenfelter et al., 1999), implying that we may observe biased published stock of knowledge.

Based on a study of gender participation in a competitive environment, there is evidence that despite the absence of gender differences in performance, men are more than twice as likely to enter the tournament (Niederle and Vesterlund, 2007), implying that women self select out of competition, leading to bias in the estimated contribution of gender to performance.

Economics duration models are also faced with incomplete data of different kinds. For example, in the case of a non-participation duration analysis (for e.g. measuring the unemployment duration), individuals who are non-participant for a short period of time (between two successive surveys) are underrepresented in data (Kiefer, 1988), leading to an over estimation of the non-participation length (conditional on being a non-participant). In the case of truncated data as it is dealt with in our model, we observe a sample of participants only, and as a result we do not observe the non-participants' covariates at all. However, in duration models both participants' and non-participants' covariates are observable.

In quantum physics and theoretical chemistry, due to computational limitations it is not feasible to utilize all the required observations (electrons) which can amount to data size of sextillion (10^{21}) terabytes. Kohn and Sham (1965) utilized the density function theorem (DFT) to reduce the data requirement to just a few hundred kilobytes, well within the computation capacity. This however, can introduce selectivity bias.

In the field of computer science there is a use of 'machine learning' for selectivity bias correction. However these methodologies are applicable only if two types of data sets are accessible to the researcher (Huang et al., 2006; Sugiyama et al., 2012): the unbiased data set (*training data*) which enables learning about the population true distribution and another data set which is biased (*test data*) and is intended for making predictions. The relative frequencies in the later data set are corrected using the relative frequencies in the former data set. Otherwise, these methodologies are not feasible in cases where one has no access to the full sample, i.e., truncation.

Another set of approaches dealing with truncation belongs to the field of survival analysis and are commonly employed in actuarial science (Cox, 1972). There are two key differences between survival analysis such as Kaplan Meier[7] (Bland and Altman, 1998; Kaplan and Meier, 1958) and our truncated sample approach. First, in survival analysis, the data consist of both survivors and non-survivors, such that we observe each individual (participant or not) at least one period of time. However, in truncated data dealt with in our model, each observation is sampled only once and non-participants are excluded from the very beginning from the truncated sample. Thus, we solely observe participants. In terms of survival analysis, we have a sample consisting of survivors only.

[7]The survival probability $S(t_j)$ in time t_j can be estimated using: $S(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right)$. Where $S(t_{j-1})$ stands for the survival probability in time t_{j-1}, n_j represents the number of individuals alive (participants) just before t_j, and d_j stands for the number of events (becoming non-participants) at t_j.

Second, if the parameters vector of interest ,$\boldsymbol{\gamma}$, is known or we observe the participation propensity variable Y_{2i}[8] for each observation i preserving $Y_{2i} = \boldsymbol{Z_i'\gamma} + \xi_{2i}$ (see equation (2.2) that follows) we can treat each observation i's covariates mixture as an equivalent to the observable truncated time variable in survival analysis. However, since these parameters are unknown and the variable Y_{2i} is latent (it is not observed in the data), one cannot employ survival analysis in order to deal with the truncation for sake of selectivity bias correction.

Additional widely employed methodology in economics dealing with truncation, is a truncated regression analysis (Amemiya, 1973). This approach is applicable if the truncation depends entirely on the dependent variable, *in the absence of an endogenous selection* rule dictating which observations are included or excluded. In many cases, the dependent variable has a lower or upper limit (truncation point),[9] implying that conditional on the covariates, the dependent variable's support is bounded in a deterministic fashion, such that Y_1 can take only certain values conditional on X. Hausman and Wise (1977) introduced a methodology to estimate a truncated regression in the case that the dependent variable in each observation individually is required to satisfy a constraint. For instance, we observe a sample consisting of observations with current family income (the dependent variable) below their last year family income (an individual cut-off for each observation).

Incidental truncation models deal with an endogenous selection rule that determines which observations are included in the truncated sample, instead of a direct truncation on the dependent variable. In these models the selection rule is determined endogenously (Heckman, 1979) as a function of the observation's covariates vector Z (the systemic part) as well as a random disturbance (the non-systemic part), the truncation is only in probability space (it is not a deterministic function of the covariates). Thus, due to the random disturbance term in the selection rule, there is no implicit restriction on the Y_1 variable (taking only values above or below some constant y_0) that can be imposed to correct for the selectivity bias. As a result, the substantive equation's dependent variable Y_1 can take (in a large sample) any value conditional on the covariates vector X.

In the case of truncation, the sample is restricted to participants only. Thus, all the approaches above dealing with a direct truncation on the dependent variable or utilizing the information embodied in the non-participants' covariates are not applicable for the endogenous selectivity bias correction.

A methodology for endogenous truncation bias correction in the absence of non-participants data, was introduced by Bloom and Killingsworth (1985). That

[8] The participation propensity is a continuous random variable measuring, for each observation, its likelihood to be included in the truncated sample.

[9] For instance, in the case of a left truncation one can only observe a sub-population of the dependent variable satisfying $Y_{1i} \geq y_0$, implying that y_0 is served as a truncation point. The estimation method involves the maximization of truncated density function product $\prod_{i=1}^{n} f(Y_i|Y_i > y_0)$ with respect to the $\boldsymbol{\beta}$ parameter values, where $Y_{1i} = \boldsymbol{X_i'\beta} + \xi_{1i}$ and f is a density function, truncated at the point y_0.

approach was intended for selectivity bias correction in the special case of joint normally distributed disturbances.[10] However, in the general case, the joint distribution of the two equations' disturbances is unknown and a violation of the disturbances' joint normality distribution assumption leads to biased estimators, especially in the case of truncation as was shown by Arabmazar and Schmidt (1982). Lee (1982) showed that when the distributions of the disturbances are misspecified to be normal, one may not be able to even detect the presence of selectivity bias in the regression equations. Additional disadvantage of that approach is that one cannot estimate the proportion of participants and measure the magnitude of truncation. This however requires the non-participants data to be recovered which we do in the present paper.

Thus, we suggest a new methodology that is applicable in cases where the disturbances' joint-distribution function is unknown and it is not necessarily bivariate normal.[11] Our methodology enables estimating the selection equation using data reconstruction, implying that one can recover the unobserved part of the data distribution function and calculate the percentage of missing data (the proportion of non-participants). Moreover, the selection equation is estimated without specifying any distribution function for the substantive equation's disturbance. The aforementioned disturbance is not restricted to be symmetric about zero given the covariates or identically and independently distributed (heteroskedasticity is allowed). Additionally, as will be shown in section (4.3) the estimation procedure allows for a complete separation of the outcome variable when one employed dichotomous (dummy) variables which are frequently used in social research.

Several semi-parametric methods can be employed for selectivity bias correction in censored data using a two steps procedure. These methods are appropriate for dealing with a general dependence between the substantive equation's disturbance and the selection equation's disturbance. One method approximates in the second step the selection term in the substantive equation using a polynomial of the probability to participate estimated in the first step (Newey, 2009). Another method is referred to as a *pairwise difference* and is based on matching any two observations in the second step with a similar probability to participate to eliminate the selection term by first order difference (Ahn and Powell, 1993; Powell, 1987).

For the truncated selection model in semi-parametric methods, identification and estimation of the unknown parameters is much more difficult. The methods involve one-step procedure which imposes some restrictions on the disturbances. The model is referred to as "Type 3 Tobit" (Amemiya, 1985) and imposes the restrictions on the conditional symmetry of the disturbances about zero given

[10] That methodology involves both equations' unknown parameters estimation by a maximization of the likelihood function based on the conditional (on participation) joint density function of the two equations' dependents variables. This density function is obtained under the disturbances' joint normality distribution.

[11] As is assumed in Bloom and Killingsworth (1985).

the covariates (Powell, 1994). Another method is based on a procedure suggested by Honoré and Powell (1994), where the model is estimated without intercepts and the disturbances are symmetric or independently and identically distributed (Honore et al., 1997).

In the present paper we introduce a two step procedure: in the first step one estimates only the selection rule's unknown parameters (regardless of the substantive equation) and in the second step estimating the substantive equation's unknown parameters, using the probability to participate (estimated in the previous step) as an additional covariate.

In next section, we will distinguish between two types of missing data: censored and truncated. In both cases, we have missing data on the dependent variable in the equation of interest for all the observations related to the non-participants. However, in a truncated sample we do not observe the non-participants covariates (which are related to the selection rule equation), implying that the observed data on these covariates are drawn from a truncated distribution function, while in a censored sample the observed data on these covariates are drawn from the complete distribution function. Due to the difficulty of not being able to employ the information embodied in the entire distribution function, stemming from its absence, one has to deliver methodology capable of recovering the non-truncated density function. This is the main innovation of the present paper in the correction for endogenous selectivity bias under truncation.

2.1. Two Equations Selection Model

We first present the general structure of an endogenous selection model, by characterizing the set of two regression equations to be estimated, each depicting a dependent variable denoted by Y_{1i} and Y_{2i} and disturbances ξ_{1i} and ξ_{2i} respectively. The first equation is a linear function of a $(1 \times (k+1))$ covariates vector $\boldsymbol{X_i}$ and is referred to as the substantive equation (of interest), for k number of covariates excluding the intercept. The second equation is a linear function of a $(1 \times (p+1))$ covariates vector $\boldsymbol{Z_i}$ and is referred to as the selection equation, for p number of covariates excluding the intercept.

For a random sample of N observations, the regression equations (substantive and selection) for individual i may be written as (Heckman, 1974, 1976, 1979) :

$$Y_{1i} = \boldsymbol{X_{1i}}\boldsymbol{\beta} + \xi_{1i} \tag{2.1}$$

the substantive equation, where $\boldsymbol{\beta}$ is a $(1 \times (p+1))$ parameters vector consisting of $k+1$ coefficients including an intercept, and,

$$Y_{2i} = \boldsymbol{Z_i}\boldsymbol{\gamma} + \xi_{2i} \tag{2.2}$$

the selection equation, where $\boldsymbol{\gamma}$ is a $(1 \times (p+1))$ parameters vector consisting of $p+1$ coefficients including an intercept.

In order to estimate the substantive equation (2.1), one must estimate the selection equation (2.2) while assuring zero mean and homoscedasticity.

The population regression (2.1) is:

$$\mathbb{E}\left[Y_{1i}|\boldsymbol{X_{1i}}\right] = \boldsymbol{X_{1i}}\boldsymbol{\beta}, i = 1, .., N \tag{2.3}$$

However, the dependent variable in the substantive equation Y_{1i} is observed if and only if $Y_{2i} > 0$, implying that our sample is incomplete.

The regression equation for the incomplete sample for $i = 1, .., n$ is:

$$\mathbb{E}\left[Y_{1i}|\boldsymbol{X_{1i}}, Y_{2i} > 0\right] = \boldsymbol{X_{1i}}\boldsymbol{\beta} + \mathbb{E}\left[\xi_{1i}|Y_{2i} > 0\right] \tag{2.4}$$

Heckman (1979) obtained the result described in equation (2.4) that the selection bias in the substantive equation is caused by a mis-specification when neglecting the conditional expectation of its random disturbance on the selection rule. In what follows we present the Heckman's correction for selectivity bias dealing with the special case of a joint-normality distribution function of the disturbances.

Based on equation (2.4), the conditional expectation of the dependent variable is a function of two terms as in the following equation:

$$\mathbb{E}\left(Y_{1i}|\boldsymbol{X_{1i}}, Y_{2i} \geq 0\right) = \boldsymbol{X_{1i}}\boldsymbol{\beta} + \mathbb{E}\left(\xi_{1i}|\xi_{2i} \geq -\boldsymbol{Z_i'}\boldsymbol{\gamma}\right). \tag{2.5}$$

Using the joint normal distribution (Heckman, 1976):

$$\mathbb{E}\left(\xi_{1i}|Y_{2i} > 0\right) = \mathbb{E}\left(\xi_{1i}|\xi_{2i} > -\boldsymbol{Z_i'}\boldsymbol{\gamma}\right) = \frac{\sigma_{\xi_{12}}}{\sqrt{\sigma_{\xi_{22}}}}\lambda_i \tag{2.6}$$

Equation (2.6) implies that the random disturbance's conditional (on participation) expectation in equation (2.1) is a function of the inverse Mills ratio denoted by the parameter λ_i[12], defined as follows:

$$\lambda_i = \frac{\phi\left(-\frac{\boldsymbol{Z_i'}\boldsymbol{\gamma}}{\sqrt{\sigma_{\xi_{22}}}}\right)}{1 - \Phi\left(-\frac{\boldsymbol{Z_i'}\boldsymbol{\gamma}}{\sqrt{\sigma_{\xi_{22}}}}\right)}. \tag{2.7}$$

The regression equation for the incomplete sample is simplified to:

$$\mathbb{E}\left(Y_{1i}|\boldsymbol{X_{1i}}, Y_{2i} \geq 0\right) = \boldsymbol{X_{1i}}\boldsymbol{\beta} + \frac{\sigma_{\xi_{12}}}{\sqrt{\sigma_{\xi_{22}}}}\lambda_i \tag{2.8}$$

Thus, if we know λ_i (or ϕ_i), or an estimate of such (based on the complete sample consisting of both participants and non-participants), least squares estimators could be applied to equation (2.8) for the selectivity bias correction.[13]

[12] The hazard rate.

[13] In case our sample is restricted to participants only and a joint-normal distribution is assumed for the disturbances, one can employ the procedure introduced by Bloom and Killingsworth (1985) to correct for the selectivity bias, in a similar fashion to the conventional

However, the disturbances' joint normality distribution assumption has some limitations. A modest hetroscedasticity in the disturbance ξ_{1i} causes the parameters to be mis-estimated by a substantial amount (Hurd, 1979). Moreover, if the expectation of ξ_{1i} conditional on ξ_{2i} is not linear or the disturbances' joint normality distribution is violated, λ_i misspecifies the relationship between $\boldsymbol{Z_i'\gamma}$ and Y_{2i}, and might leading to biased estimates, especially for a truncated sample (Arabmazar and Schmidt, 1982).

Thus, one can apply a similar analysis to Heckman's selectivity bias correction without assuming a specific function for the aforementioned conditional expectation using some unknown general function $\mathcal{M}$ instead of λ_i, as follows:

$$\mathbb{E}\left[\xi_{1i}|Y_{2i} > 0\right] \equiv \mathcal{M}(\boldsymbol{Z_i'\gamma}) \tag{2.9}$$

Then, the substantive equation to be estimated is a partially linear regression (Robinson, 1988), defined as follows:

$$Y_{1i} = \boldsymbol{X_{1i}\beta} + \mathcal{M}(\boldsymbol{Z_i'\gamma}) + \upsilon_{1i}, \tag{2.10}$$

where υ_{1i} is a random disturbance which satisfies $\mathbb{E}[\upsilon_{1i}|\boldsymbol{X_{1i}, Z_i}] = 0$, and $\mathcal{M}$ is some unknown function.

This equation is a special case of a semi-parametric model because it is a partially linear function of two components: a linear function of unknown parameters and some non-linear unknown function which captures the conditional expectation of the disturbance given the covariates. The general non-linear function component enables the estimation of the model without assuming a specific conditional distribution function for the disturbance ξ_1 given the selection rule.

Next we discuss the difference between truncate and censored data.

3. Methodology

3.1. Truncated vs. censored data

In the case of a censored sample, it is possible to compute the probability that an observation has data missing from Y_1 and hence it is possible to use Probit analysis to estimate ϕ_i and λ_i. Both observations for which $Y_{2i} > 0$ (participants) and observations for which $Y_{2i} < 0$ (non-participants) are accessible in the data. In such a case Heckman's correction is feasible.

In the case of a truncated sample, however, Heckman's correction is no longer feasible, because there is no information regarding the non-participants. Only observations for which $Y_{2i} > 0$ (participants) are observable. This rather more complicated situation requires a different methodology which can compute the probability that an observation has data missing on Y_1, using a truncated Probit estimation. Thus, it is possible to to estimate ϕ_i and λ_i in equation (2.8).

In both aforementioned cases, we have an endogenous variable Y_2 which determines the selection process (the selection equation) and the equation for Y_1,

Heckman's correction procedure using only the participants data.

the equation of interest (or the substantive equation). We indicate participants by $\mathcal{S} = 1$, and all others are non-participants (as will be described in equation (4.3) to follow).

If the selection equation's and the substantive equation's error terms are entirely not correlated, then by definition a selection bias is not a problem. Under most circumstances, however, the assumption of independent error terms will not be met because of specification problems. If any factor that affects both the selection and substantive equations is omitted from the model, this factor will enter both error terms and induce correlation between them. Ignoring this correlation, is leading to biased estimates.

In order to correct for selectivity bias, Heckman (1979) introduced a methodology treating the sample selection bias problem as an example of omitting variable bias that can be solved in two step procedure. First a selection equation is estimated by Probit regression, and its predicted values are the observations' probabilities to participate in the sample. The inverse Mills ratio (which is the hazard rate[14]) is then estimated based on these estimated probabilities. The second step is estimating an ordinary least square (OLS) regression with the covariates X as well as the inverse Mills ratio. As is evident, the aforementioned Heckman's procedure is applicable only for cases of censored data when observations on both the included and secluded (non-participants) are available. However, in the case of truncated data no information as to non participants exists, implying that Heckman's Inverse Mills ratio cannot be estimated and hence no correction for the selectivity bias is feasible since the Probit model is no longer identified (Heckman and Singer, 1984).

Thus, we offer the following procedure, instead of estimation of the participation's probability of each individual, we treat our observations in the complete (non-truncated) sample as if they were drawn randomly from a specific distribution function, such that only observations which meet the selection rule (for which $Y_2 > 0$) are observed in the incomplete (truncated) sample. Two sets of parameters are unknown and are needed to be estimated: the first set involves the parameters that characterize the complete sample density function. The second set involves the selection equation coefficients (in the censored case these are estimated by Probit regression). By conditioning the complete sample distribution function on the selection rule, we obtain the *truncated distribution function* and as a by-product the *truncated density function*. Then, we provide a solution for the following problem: what is the set of parameters which maximize the likelihood to obtain the truncated sample we observe.

Each observation is characterized by a different propensity to be included in the sample (as is depicted in equation (2.2) above). The aforementioned propensity is a continuous latent variable which is a weighted sum of participation covariates with unknown estimable coefficients and a random disturbance, such that whenever this propensity is above zero the particular observation is observed

[14]The instantaneous probability of being excluded.

in the data.

Due to the random disturbance in the selection process, some observations with low weighted sum of covariates (denoted by $\boldsymbol{Z_i'\gamma}$) are selected into the sample and some other observations with high weighed sum of covariates are not selected into the sample, however, the higher the weighted sum of covariates, the more likely an observation is selected into the sample. As a matter of fact, the random disturbance term in the selection equation plays an important role. Given that an observation is or is not included in the truncated sample it may not necessarily imply that every observation with the same covariates will or will not be included. This is a result from the existence of the random disturbance. For this reason, we can address the main issue: how likely it is for an observation i to appear in the data, given that this observation preserves the participation condition.

4. Selectivity bias correction for truncated data

This section introduces our proposed methodology for the correction of truncation bias. It is assumed that the observations in the complete sample consists of $(p \times 1)$ variables (describing the selection process into the incomplete sample) were drawn from a known multivariate distribution denoted by $\mathcal{F}_p(\boldsymbol{\eta})$, but since we observe only observations preserving the participation constraint $\mathcal{S} = 1$, our incomplete data observations are distributed as if they were drawn from a truncated distribution function, denoted by $\mathcal{F}_p(\boldsymbol{\eta}|\mathcal{S} = 1)$. The vector of unknown parameters $\boldsymbol{\eta}$ characterizes the non-truncated multivariate distribution function, and is referred to as incidental (nuisance) parameters vector, because it is used only for the sake of identification of unknown parameters of interest denoted by $\boldsymbol{\gamma}$ vector (as described in equation (2.2)). Then, we use maximum likelihood method in order to estimate the truncated multivariate density function denoted by $f_{\boldsymbol{\theta}}(z_1, ..., z_p|\mathcal{S} = 1)$, such that we are looking for the vector of parameters $\boldsymbol{\theta} = [\boldsymbol{\eta}, \boldsymbol{\gamma}]^T$ which maximizes the likelihood for generating the truncated sample we observe. Following theorem 2.1 in (Mäkeläinen et al., 1981) the maximum likelihood estimate is shown to exist and to be unique if it is constant on the boundary of the parameter space and if the Hessian matrix is negative definite whenever the gradient vector vanishes.

Once those parameters are identified, we calculate the probability for every observation in the sample for being selected into the sample in the first place. This probability is equivalent to estimating a Probit regression on the non-truncated data consisting of both participants and non-participants. This estimated probability makes it possible to calculate the Inverse Mill's ratio, and then to correct the selectivity bias as in Heckman (1979).

In the next section, we discuss the difference between conventional Probit regression estimation and the truncated Probit regression estimation.

4.1. Probit regression estimation

4.1.1. The conventional Probit - Both participants and non-participants are observable

As mentioned by Heckman (1976) the regression of the incomplete sample observations yields an unbiased estimates of $\boldsymbol{\beta}$ if the conditional expectation of ξ_{1i} on selection rule is zero. Of course, this is not always the case. More generally, Y_{1i} is observed only if $Y_{2i} \geq C$, for some arbitrary constant C. In our present case it is normalized to zero due to the intercept's coefficient in the selection equation. Y_{2i} may be interpreted as the propensity to participate or entering into the sample. Otherwise, we cannot observe Y_{1i} and its observed value Y_{1i}^* is zero.

Denote a dichotomous random variable for being selected $\mathcal{S}_i = \begin{cases} 1, & \text{if } Y_{2i} \geq 0 \\ 0, & \text{if } Y_{2i} < 0 \end{cases}$.

Observation i's participation probability in the case of a Probit model is specified as function of its covariates $\boldsymbol{Z_i}$:

$$P_{\boldsymbol{\gamma}}(\mathcal{S}_i = 1|z_i) = P\left(\xi_{2i} + \boldsymbol{Z_i'\gamma} > 0|\boldsymbol{Z_i'}\right) = P\left(\xi_{2i} > -\boldsymbol{Z_i'\gamma} > 0|\boldsymbol{Z_i'}\right) = 1 - \Phi\left(-\boldsymbol{Z_i'\gamma}\right). \tag{4.1}$$

The likelihood function based on equation (4.1) is a product of two terms. The first one is the probability to not-participate which is calculated for non-participants only. The second one, is the complimentary probability to participate which is calculated for participants only. More formally, the likelihood function is:

$$L\left(\boldsymbol{\gamma}; Y_2|Z\right) = \prod_{i:\mathcal{S}_i=0} \Phi\left(-\boldsymbol{Z_i'\gamma}\right) \prod_{i:\mathcal{S}_i=1} \left(1 - \Phi\left(-\boldsymbol{Z_i'\gamma}\right)\right) \tag{4.2}$$

For presentation convenience we denote the participant observations by:

$$\mathcal{S}_i = I\left(\xi_{2i} + \boldsymbol{Z_i'\gamma} > 0\right), \tag{4.3}$$

where $I(\cdot)$ is the indicator function.

The main disadvantage of estimating the Probit model presented above is the requirement that the sample consists of non-participants which is not always accessible.

4.1.2. Truncated Probit - Only participants are observable

In a more general case we assume that only observations for which $\mathcal{S}_i = 1$ are observable for both the participation equation covariates $\boldsymbol{Z_i}$ and dependent variable Y_{1i} of the main equation, indicating that our sample is no longer a censored sample but it is a truncated one.

Correction for truncation requires a new methodology, since the conventional Probit regression requires both observations for participants and non-participants.

We postulate two essential assumptions: (i) The data generation function is well known up to unknown estimable parameters, such that it can be represented as a multivariate density function. (ii) The mechanism of entering into the sample can be fully characterized as a random variable which is a weighted sum of all participation covariates together with a random disturbance. The weights are estimable unknown parameters. We refer to this random variable as the propensity of an individual observation to be included in the sample.

If both assumptions hold, all it takes is to characterize the truncated (constrained) data generation function, or equivalently the conditional density of data generation function on entering into the sample. In other words, the first requirement represents the full data generation function, while the second requirement represents the restriction.

Our approach is to estimate the parameters of the conditional density function (on the event of participation) of a given observation i of being sampled. The event of participation is a function of covariates and a random disturbance (see equation (2.2)).

Before we present the application of our approach to the multivariate Probit regression, we begin with the univariate Probit regression.

4.1.3. The univariate latent Probit regression case

Let z_i, $i = 1, ..., N$, be an i.i.d random variable normally distributed which determines the participation propensity for each observation i with expectation μ_z and variance σ_z^2, such that the incidental parameters vector is denoted by $\boldsymbol{\eta} = [\mu_z, \sigma_z^2]$. We denote the vector $\boldsymbol{Z_i} = [1, z_i]^T$ and likewise we denote $\boldsymbol{\delta_z} = [1, \mu_z]^T$. The weighted sum of participation determinant and a random disturbance denoted by ξ_{2i}[15] is represented by $\boldsymbol{Z_i'\gamma} + \xi_{2i}$. This weighted sum is referred to as the convolution of two independent random variables.[16] The convolution coefficients vector is denoted by $\boldsymbol{\gamma} = [\gamma_0, \gamma_1]^T$.

Without loss of generality (for a univariate participation equation), the density function for the maximum likelihood is:

$$L(\boldsymbol{\theta}|Z) = \prod_{i:\mathcal{S}_i=1} f_{\boldsymbol{\theta}}(z_i|\mathcal{S}_i = 1) = \prod_{i:\xi_{2i}+\boldsymbol{Z_i'\gamma}>0} f_{\boldsymbol{\theta}}\left(z_i|\boldsymbol{Z_i'\gamma} + \xi_{2i} > 0\right) \tag{4.4}$$

Although Y_{2i} is observed as a dichotomous variable $\mathcal{S}_i$ which is equal to 1 for every observation in the incomplete sample, we utilize the fact that its unobserved value can be represented as convolution of participation explanatory variables with a random disturbance (see equation (2.2)). Based on equation (4.4), we treat this convolution as a constraint making it possible to characterize the density function of the truncated (constrained) data generation process. Applying theorem 2.1 in (Mäkeläinen et al., 1981), the maximum likelihood function should

[15] The disturbance term is randomly and normally distributed with zero mean and unitary variance, as in the conventional Probit.

[16] The vector Z is consisting of exogenous covariates with respect to ξ_{2i}.

preserve constant boundaries and negative definiteness of the Hessian when the gradient vanishes to establish uniqueness. Alternatively uniqueness can be established under regularity conditions given in Chanda (1954), by applying theorem 2 in (Rai and Van Ryzin, 1982). These characteristics have been verified numerically.[17]

4.1.4. The Truncated (constrained) data generation function

Our goal is to recover the unobservable complete sample density function. However, we have access only to a sub-sample of observations (truncated sample) which meet the selection rule condition. The complete sample as well as the truncated sample are characterized each by their own specific density functions. Nevertheless, the two densities are interrelated through common parameters. This is the case because the truncated sample density function is obtained by conditioning the complete sample density function on the selection rule. The aforementioned density function depends both on the parameters characterizing the complete sample density function as well as on the selection rule parameters. Thus, the parameter set characterizing the truncated density function contains the parameter set characterizing the complete sample density function. Consequently, we can infer the complete sample density function through the linkage between the truncated and non-truncated density functions.

Therefore, we need to find the parameters which maximize the likelihood to obtain the truncated data we observe using the truncated density function which generated the observations in the incomplete sample. Once identifying these parameters, we use the estimated selection rule parameters to evaluate each observation's participation probability. By estimating these parameters, one can reconstruct the entire data on the covariates as if one observes the (non-truncated) data distribution function.

The first step is specifying the distribution function which generates the complete (non-truncated) data. The cumulative distribution function of the observation in the complete sample is denoted by $\mathcal{F}(z_i)$.

The probability of an observation to be below z_i and to participate is:

$$P\left(Z < z_i, \boldsymbol{Z_i'}\boldsymbol{\gamma} + \xi_{2i} > 0\right) = P\left(Z < z_i\right) - P\left(Z < z_i, \boldsymbol{Z_i'}\boldsymbol{\gamma} + \xi_{2i} < 0\right). \tag{4.5}$$

[17] In order to verify both the uniqueness and existence of maximum likelihood function's estimators, we have used Monte-Carlo simulations in which random data sets have been generated (under different parameter settings we have chosen arbitrarily). For each data set we have generated, we searched over the entire parameter space the vector of parameters which maximizes the MLE (maximum likelihood function) in equation (4.19) to follow. In each simulation we found that: the Hessian matrix of the MLE is always negative definite in every point in which the gradient vector vanishes (indicating a local maximum) and that the gradient vector vanishes only in one interior point (uniqueness). The estimated vector was found to be consistent and approaches the true parameter vector as number of observations increases (as will be depicted in section (5)).

For notification convenience we define the following:

$$\mathcal{F}(z_i) \equiv P(Z < z_i) \tag{4.6}$$

and,

$$\mathcal{F}_2(z_i, 0) \equiv P\left(Z < z_i, \boldsymbol{Z_i'\gamma} + \xi_{2i} < 0\right). \tag{4.7}$$

Thus, we can express equation (4.5) in the following way:

$$P\left(Z < z_i, \boldsymbol{Z_i'\gamma} + \xi_{2i} > 0\right) = \mathcal{F}(z_i) - \mathcal{F}_2(z_i, 0). \tag{4.8}$$

The generalized truncated univariate distribution function is therefore:

$$\mathcal{F}(z_i|\mathcal{S}_i = 1) = \frac{\mathcal{F}(z_i) - \mathcal{F}_2(z_i, 0)}{P_{\boldsymbol{\theta}}(\mathcal{S}_i = 1)}. \tag{4.9}$$

It is more convenient to express $\mathcal{F}_2(z_i, 0)$ in the following way:

$$\mathcal{F}_2(z_i, 0) = \int_{v_1=-\infty}^{z_i} f_{\boldsymbol{\eta}}(v_1) P\left(\boldsymbol{Z_i'\gamma} + \xi_{2i} < 0|z_i = v_1\right) dv_1. \tag{4.10}$$

Since our goal is to obtain the truncated density function for the sake of using the maximum likelihood method, we need to obtain the derivative of the cumulative distribution function in equation (4.9) with respect to z_i. For that purpose, we first take the derivative of equation (4.10) with respect to z_i:

$$\frac{\partial \mathcal{F}_2(z_i, 0)}{\partial z_i} = f_{\boldsymbol{\eta}}(z_i) P\left(\boldsymbol{Z_i'\gamma} + \xi_{2i} < 0|z_i\right) = f_{\boldsymbol{\eta}}(z_i) \Phi\left(-\boldsymbol{Z_i'\gamma}\right). \tag{4.11}$$

Next, we would like to find the participation probability (in the denominator of equation (4.9)):

$$\begin{aligned} &P_{\boldsymbol{\theta}}(\mathcal{S}_i = 1) = P(\boldsymbol{Z_i'\gamma} + \xi_{2i} > 0) = \\ &\int_{v_1=-\infty}^{z_i} f_{\boldsymbol{\eta}}(v_1) P\left(\boldsymbol{Z_i'\gamma} + \xi_{2i} > 0|z_i = v_1\right) dv_1. \end{aligned} \tag{4.12}$$

We denote the vector $\boldsymbol{V_1} = [1, v_1]^T$, and thus the generalized participation probability is:

$$P_{\boldsymbol{\theta}}(\mathcal{S}_i = 1) = P(\boldsymbol{Z_i'\gamma} + \xi_{2i} > 0) = \int_{v_1} f_{\boldsymbol{\eta}}(v_1) \left[1 - \Phi\left(-\boldsymbol{V_1'\gamma}\right)\right] dv_1. \tag{4.13}$$

After taking the derivative of the cumulative distribution function (equation (4.9)) with respect to z_i as depicted in equation (4.11) we get the constrained data generation density function:

$$f_{\boldsymbol{\theta}}\left(z_i|\boldsymbol{Z_i'\gamma} + \xi_{2i} > 0\right) = \frac{f_{\boldsymbol{\eta}}(z_i)\left[1 - \Phi\left(-\boldsymbol{Z_i'\gamma}\right)\right]}{\int_{v_1} f_{\boldsymbol{\eta}}(v_1)\left[1 - \Phi\left(-\boldsymbol{V_1}'\gamma\right)\right] dv_1}. \tag{4.14}$$

Note that the term $f_{\boldsymbol{\eta}}(z_i)$ in equation (4.14) represents the non-truncated data generation density function; the term $[1 - \Phi(-Z_i'\boldsymbol{\gamma})]$ represents the participation probability of individual-specific observation i (based on a Probit binary response model). The term in the denominator represents the probability of a randomly

selected individual to be included in truncated data.

4.1.5. Simple example: The case of bivariate normal distribution

We demonstrate an application of our model to the case of a bivariate normal distribution.

We assume that the random variables z_i and its convolution with ξ_{2i} are bivariate normally distributed[18], or more formally:

$$\begin{bmatrix} z_i \\ \boldsymbol{Z_i'}\boldsymbol{\gamma} + \xi_{2i} \end{bmatrix} \sim \mathcal{N}_2 \left(\begin{bmatrix} \mu_z \\ \boldsymbol{\delta_z'}\boldsymbol{\gamma} \end{bmatrix}, \begin{bmatrix} \sigma_z^2 & \rho\sigma_Z\sqrt{1+{\gamma_1}^2\sigma_z^2} \\ \rho\sigma_z\sqrt{1+{\gamma_1}^2\sigma_z^2} & 1+{\gamma_1}^2\sigma_z^2 \end{bmatrix} \right).$$

We denote the bivariate cumulative distribution function of two random variables which are distributed bivariate normal as:

$$P\left(z_1 < z_i, z_2 < 0\right) = P\left(z_1 < z_i, \boldsymbol{Z_i'}\boldsymbol{\gamma} + \xi_{2i} < 0\right) = \Phi_2\left(\frac{z_i - \mu_z}{\sigma_z}, -\frac{\boldsymbol{\delta_z'}\boldsymbol{\gamma}}{\sqrt{1+\gamma_1^2\sigma_z^2}}, \rho\right), \tag{4.15}$$

where the ρ parameter is the correlation coefficient between z_i and the convolution:[19]

$$\rho = \frac{\gamma_1\sigma_z}{\sqrt{1+\gamma_1^2\sigma_z^2}}. \tag{4.16}$$

The cumulative distribution function of the constrained data generation process can equivalently be defined as the conditional cumulative distribution function of the unrestricted data generation process which is:

$$P\left(Z_i < z_i | \boldsymbol{Z_i'}\boldsymbol{\gamma} + \xi_{2i} > 0\right) = \frac{\Phi\left(\frac{z_i - \mu_z}{\sigma_z}\right) - \Phi_2\left(\frac{z_i - \mu_z}{\sigma_z}, -\frac{\boldsymbol{\delta_z'}\boldsymbol{\gamma}}{\sqrt{1+\gamma_1^2\sigma_z^2}}, \rho\right)}{1 - \Phi\left(-\frac{\boldsymbol{\delta_z'}\boldsymbol{\gamma}}{\sqrt{1+\gamma_1^2\sigma_z^2}}\right)}. \tag{4.17}$$

The constrained data generation density function is:

$$f_{\boldsymbol{\theta}}\left(z_i | \boldsymbol{Z_i'}\boldsymbol{\gamma} + \xi_{2i} > 0\right) = \frac{\frac{1}{\sigma_z}\phi\left(\frac{z_i - \mu_z}{\sigma_z}\right)\left[1 - \Phi\left(-\boldsymbol{Z_i'}\boldsymbol{\gamma}\right)\right]}{1 - \Phi\left(-\frac{\boldsymbol{\delta_z'}\boldsymbol{\gamma}}{\sqrt{1+\gamma_1^2\sigma_z^2}}\right)}. \tag{4.18}$$

[18] The convolution of multivariate normal distributed random variables is distributed normally (see theorem (8.1) in the Appendix). This property makes it easier to evaluate the probability to participate which is based on a convolution of normally distributed variables.

[19] $\mathbb{COV}(z_i, \boldsymbol{Z_i'}\boldsymbol{\gamma} + \xi_{2i}) = \mathbb{COV}(z_i, \gamma_1 z_i) = \gamma_1\sigma_z^2$

Note that equation (4.18) is composed of three parts: the first part is the data generation function, the second part is the probability of participation (based on the Probit regression probability to participate $\left(1-\Phi\left(-\boldsymbol{Z_i'\gamma}\right)\right)$) and the third one represents the probability of a randomly selected individual observation to be included in truncated data (in the denominator). Both scalars μ_z and σ_z are incidental parameters (employed to recover the unobserved data distribution function regarding the non-participants). The parameters of interest are capitalized in vector $\boldsymbol{\gamma}$, these parameters are the conventional Probit regression coefficients.

4.1.6. The likelihood function

In order to estimate the data generation density function we apply maximum likelihood method. The log-likelihood function applied to the density function depicted in equation (4.18) is:

$$\ln L = -n\ln\sigma_z + \sum_{i=1}^{n}\ln\phi\left(\frac{z_i-\mu_z}{\sigma_z}\right) + \sum_{i=1}^{n}\ln\left[1-\Phi\left(-\left(\gamma_0+\gamma_1 z_i\right)\right)\right] - n\ln\left[1-\Phi\left(-\frac{\boldsymbol{\delta_z'\gamma}}{\sqrt{1+\gamma_1^2\sigma_z^2}}\right)\right] \tag{4.19}$$

Denoting $\lambda_\delta \equiv \frac{\phi\left(-\frac{\boldsymbol{\delta_z'\gamma}}{\sqrt{1+\gamma_1^2\sigma_z^2}}\right)}{1-\Phi\left(-\frac{\boldsymbol{\delta_z'\gamma}}{\sqrt{1+\gamma_1^2\sigma_z^2}}\right)}$, the first order conditions obtained from the log-likelihood maximization are:

$$\frac{\partial\ln L}{\partial\sigma_z} = -\frac{n}{\sigma_z} + \sum_{i=1}^{n}\frac{(z_i-\mu_z)^2}{\sigma_z^3} + n\lambda_\delta\frac{\gamma_1^2\sigma_z(\gamma_0+\gamma_1\mu_z)}{\left[1+\gamma_1^2\sigma_z^2\right]^{3/2}} = 0 \tag{4.20}$$

$$\frac{\partial\ln L}{\partial\mu_z} = \sum_{i=1}^{n}\frac{z_i-\mu_z}{\sigma_z^2} - n\lambda_\delta\frac{\gamma_1}{\sqrt{1+\gamma_1^2\sigma_z^2}} = 0 \tag{4.21}$$

$$\frac{\partial\ln L}{\partial\gamma_0} = \sum_{i=1}^{n}\frac{\phi\left(-\left(\gamma_0+\gamma_1 z_i\right)\right)}{1-\Phi\left(-\left(\gamma_0+\gamma_1 z_i\right)\right)} - n\lambda_\delta\frac{1}{\sqrt{1+\gamma_1^2\sigma_z^2}} = 0 \tag{4.22}$$

$$\frac{\partial\ln L}{\partial\gamma_1} = \sum_{i=1}^{n}\frac{\phi\left(-\left(\gamma_0+\gamma_1 z_i\right)\right)}{1-\Phi\left(-\left(\gamma_0+\gamma_1 z_i\right)\right)}z_i + n\lambda_\delta\frac{\gamma_0\gamma_1\sigma_z^2-\mu_z}{\left[1+\gamma_1^2\sigma_z^2\right]^{3/2}} = 0. \tag{4.23}$$

We search for the parameter values $\gamma_0, \gamma_1, \mu_z, \sigma_z$ which solve equations (4.20) to (4.23). These parameters are needed for identification of the truncated Probit, to evaluate each observation's probability to participate in truncated sample.

4.1.7. Explicit expectation and variance

Our motivation is to show how the inverse Mills-ratio, denoted by λ_i, and μ_z are interrelated, especially for large value of the $\hat{\gamma}_1$ estimator. It is important to examine the case where the inverse Mills ratio approaches zero for every observation in the sample, indicating the absence of truncation in the first place. In order to study the impact of large γ_1 on the μ_z estimator, we denote by $\bar{\lambda}$ the average of the inverse Mills-ratio (described in equation (2.7)) over all the observations in the truncated sample :

$$\bar{\lambda} \equiv \frac{1}{n}\sum_{i=1}^{n}\lambda_i \quad \text{and the ratio: } C \equiv -\frac{\frac{1}{n}\sum_{i=1}^{n}\lambda_i z_i}{\bar{\lambda}} \tag{4.24}$$

We divide equation (4.21) by equation (4.22) to extract μ_z, and divide equation (4.22) by equation (4.23) to extract σ_z^2. We obtain the following two close-form solutions:

$$\mu_z = -\sigma_z^2\gamma_1\bar{\lambda} + \bar{z} \tag{4.25}$$

and

$$\sigma_z^2 = \frac{C+\mu_z}{-C\gamma_1^2+\gamma_0\gamma_1}. \tag{4.26}$$

By using both equations (4.25) and (4.26) it is possible to express μ_z and σ_z in terms of γ_0 and γ_1. We obtain the following two close-form solutions:

$$\mu_z = -\frac{\gamma_1\bar{\lambda}\left(C+\bar{z}\right)}{-C\gamma_1^2+\gamma_0\gamma_1+\gamma_1\bar{\lambda}} + \bar{z} \tag{4.27}$$

and

$$\sigma_z^2 = \frac{C+\bar{z}}{-C\gamma_1^2+\gamma_0\gamma_1+\gamma_1\bar{\lambda}}. \tag{4.28}$$

In order to prove that $\lim_{\gamma_1\to\infty}\hat{\mu}_z = \bar{z}$ (in theorem (4.1) to follow), we examine separately each one of the expressions $\lim_{\gamma_1\to\infty}\lambda_i$, $\lim_{\gamma_1\to\infty}\lambda_i\gamma_1$ and $\lim_{\gamma_1\to\infty}C$ (described in equation (4.27)) using propositions (4.1) to (4.3).

Taking the derivative of λ_i with respect to γ_1; we get:

$$\lambda_i' = -z_i\left[\frac{\phi'}{\phi}\lambda_i + \lambda_i^2\right]. \tag{4.29}$$

Using the following result:

$$\frac{\phi'}{\phi} = \gamma_0 + \gamma_1 z_i, \tag{4.30}$$

we obtain the derivative of λ_i with respect to γ_i:

$$\lambda_i' = -z_i\left[(\gamma_0+\gamma_1 z_i)\lambda_i + \lambda_i^2\right]. \tag{4.31}$$

We now present and prove the following results:

PROPOSITION 4.1 $\lim_{\gamma_1\to\infty}\lambda_i = 0.$

PROOF: Since $\lim_{\gamma_1\to\infty}\phi(-\gamma_0-\gamma_1 z_i) = 0$ and $\lim_{\gamma_1\to\infty} 1-\Phi(-\gamma_0-\gamma_1 z_i) = 1$ it implies that $\lim_{\gamma_1\to\infty}\lambda_i = 0.$ ■

PROPOSITION 4.2 $\lim_{\gamma_1\to\infty}\lambda_i = 0$ *implies that* $\lim_{\gamma_1\to\infty}\lambda_i\gamma_1 = 0.$

PROOF: Expressing $\lim_{\gamma_1\to\infty}\lambda_i\gamma_1$ as $\lim_{\gamma_1\to\infty}\frac{\gamma_1}{\frac{1}{\lambda_i}}$ and applying L'Hopital's rule on this limit using equation (4.31) we obtain: $\lim_{\gamma_1\to\infty}-\frac{1}{\frac{\lambda_i'}{\lambda_i^2}} = \lim_{\gamma_1\to\infty}-\frac{\lambda_i}{-z_i\left[(\gamma_0+\gamma_1 z_i)\lambda_i+\lambda_i^2\right]}$ and $\lim_{\gamma_1\to\infty}\frac{1}{z_i[\gamma_0+\gamma_1 z_i+\lambda_i]} = 0$, implying that: $\lim_{\gamma_1\to\infty}\lambda_i\gamma_1 = 0.$ ■

PROPOSITION 4.3 $\lim_{\gamma_1\to\infty} C = (n\times\min(z_1,...,z_n)).$

PROOF: Let $a_i \equiv \frac{\lambda_i}{\sum_{i=1}^n\lambda_i}$, $i=1,...,n$, then clearly $\sum_{i=1}^n a_i = 1$ and $\lim_{\gamma_1\to\infty} a_i = 0\ \ \forall i\in\{i|\lambda_i<\max(\lambda_1,...,\lambda_n)\}$. Additionally, $\lim_{\gamma_1\to\infty} a_i = 1,\ \ \forall i=\{i|\lambda_i=\max(\lambda_1,...,\lambda_n)\}$. Then, we can express C as a weighted average over $z_1,...,z_n$: $\lim_{\gamma_1\to\infty} C = \lim_{\gamma_1\to\infty} n\sum_{i=1}^n a_i z_i$. Since λ is a negative strictly monotonic function of z, it follows that λ is maximized when $z=\min(z_1,...,z_n)$ and consequently $a_i=1$ for the minimal z_i only, and thus $\lim_{\gamma_1\to\infty} C = n\min(z_1,...,z_n)$. ■

Now, we show the linkage between the observable sample average and the $\hat{\mu}_z$ estimator in the case of non-truncated sample using all the propositions above to get:

THEOREM 4.1 $\lim_{\gamma_1\to\infty}\hat{\mu}_z = \bar{z}.$

PROOF: $\lim_{\gamma_1\to\infty}\hat{\mu}_z = \lim_{\gamma_1\to\infty}-\frac{\bar{\lambda}(C+\bar{z})}{-C\gamma_1+\gamma_0+\bar{\lambda}}+\bar{z},$

where the denominator satisfies: $\lim_{\gamma_1\to\infty}-C\gamma_1+\gamma_0+\bar{\lambda} = -\infty$ and the numerator satisfies: $\lim_{\gamma_1\to\infty}\bar{\lambda}\,(C+\bar{z}) = 0$. Hence, $\lim_{\gamma_1\to\infty}\hat{\mu}_z = \bar{z}.$ ■

The interesting outcome is that when $\hat{\gamma}_1$ approaches infinity, the inverse Mills ratio approaches zero and $\hat{\mu}_z$ approaches the sample average. In this case there is no truncation bias in the first place.

Moreover, for every parameters vector of interest $\boldsymbol{\gamma} = [\gamma_0,\gamma_1]^T$ there is only one incidental parameters vector $\boldsymbol{\eta} = \left[\mu_z,\sigma_z^2\right]^T$.

Next, we extend our methodology to deal with a multivariate truncated Probit case.

4.1.8. The multivariate likelihood function

Let $\boldsymbol{z_i} = [z_{1i}, .., z_{pi}]^T$ be an i.i.d $(p \times 1)$ multivariate normal random variables vector which determines the participation propensity for each observation i with expectation vector $(p \times 1)$ denoted by $\boldsymbol{\mu_z} = [\mu_{z_1}, ..., \mu_{z_p}]$ and a covariance matrix denoted by $\boldsymbol{\Sigma}_{p\times p}$, such that $\boldsymbol{\eta} = [\boldsymbol{\mu_z}, \boldsymbol{\Sigma}_{p\times p}]$. We denote the vector $\boldsymbol{Z_i} = [1, \boldsymbol{z_i}]^T$ and $\boldsymbol{\delta_z} = [1, \boldsymbol{\mu_z}]$. The weighted sum of participation determinant and a random disturbance denoted by ξ_{2i} is represented by $\boldsymbol{Z_i'\gamma} + \xi_{2i}$. This weighted sum is referred to as the convolution of p random variables (multivariate distributed) and a disturbance term.[20] The convolution coefficients vector is denoted by $\boldsymbol{\gamma} = [\gamma_0, \gamma_1, .., \gamma_p]^T$.

Without loss of generality (for a multivariate regression in the participation equation), the density function for the maximum likelihood is:

$$L(\boldsymbol{\theta}|z_1, ..., z_p) = \prod_{i:\mathcal{S}_i=1} f_{\boldsymbol{\theta}}(z_{1i}, ..., z_{pi}|\mathcal{S}_i = 1) = \prod_{i:\xi_{2i}+\boldsymbol{Z_i'\gamma}>0} f_{\boldsymbol{\theta}}\left(z_{1i}, ..., z_{pi}|\boldsymbol{Z_i'\gamma} + \xi_{2i} > 0\right). \tag{4.32}$$

Next we present the generalized truncated density function in the case of several covariates in the participation equation.

4.1.9. The multivariate truncated (constrained) data generation function

In the present section we demonstrate the truncated multivariate distribution function which generates the data.

The observations in the complete sample, are randomly drawn from a specific multivariate distribution function, and only some of them meet the participation rule (with participation propensity above zero, see subsection (4.1.1)), implying that each observation has its own propensity to be included. Thus, it is possible to calculate the probability to randomly sample an observation such that all of its covariates are below the corresponding covariates of an observation i and to be included (participate) in the truncated data is:

$$P\left(Z_1 < z_{1i}, ..., Z_p < z_{pi}, \boldsymbol{Z_i'\gamma} + \xi_{2i} > 0\right) = P(Z_1 < z_{1i}, ..., Z_p < z_{pi}) - P\left(Z_1 < z_{1i}, ..., Z_p < z_{pi}, \boldsymbol{Z_i'\gamma} + \xi_{2i} < 0\right). \tag{4.33}$$

For notification convenience we define the following:

$$\mathcal{F}_p(z_{1i}, ..., z_{pi}) \equiv P(Z_1 < z_{1i}, ..., Z_p < z_{pi}) \tag{4.34}$$

and,

$$\mathcal{F}_{p+1}(z_{1i}, ..., z_{pi}, 0) \equiv P\left(Z_1 < z_{1i}, ..., Z_p < z_{pi}, \boldsymbol{Z_i'\gamma} + \xi_{2i} < 0\right). \tag{4.35}$$

[20] The vector Z is orthogonal to vector ξ, as assumed in Probit.

where the lower script p stands for the number of covariates in the participation equation, and $p+1$ represents the inclusion of observation i's participation propensity as additional random variable.

Thus, we can express equation (4.33) in the following way:

$$P\left(Z_1 < z_{1i}, ..., Z_p < z_{pi}, \boldsymbol{Z_i'\gamma} + \xi_{2i} > 0\right) = \mathcal{F}_p\left(z_{1i}, ..., z_{pi}\right) - \mathcal{F}_{p+1}\left(z_{1i}, ..., z_{pi}, 0\right). \tag{4.36}$$

The generalized truncated multivariate distribution function is:

$$\mathcal{F}_p\left(z_{1i}, ..., z_{pi}, |\mathcal{S}_i = 1\right) = \frac{\mathcal{F}_p\left(z_{1i}, ..., z_{pi}\right) - \mathcal{F}_{p+1}\left(z_{1i}, ..., z_{pi}, 0\right)}{P_{\boldsymbol{\theta}}(\mathcal{S}_i = 1)}. \tag{4.37}$$

Next, we would like to find the participation probability:

$$P_{\boldsymbol{\theta}}(\mathcal{S}_i = 1) = P(\boldsymbol{Z_i'\gamma} + \xi_{2i} > 0) = \int_{v_p} ... \int_{v_1} f_{\boldsymbol{\eta}}\left(v_1, ...v_p\right) P\left(\boldsymbol{Z_i'\gamma} + \xi_{2i} > 0 | z_{1i} = v_1, ..., z_{pi} = v_p\right) dv_1...dv_p. \tag{4.38}$$

We obtain the probability of a random observation to be included as a weighted sum of probabilities to participate over all possible observation's realizations, using the multivariate density function f as a weight. We denote the vector $\boldsymbol{V} = [1, v_1, ..., v_p]^T$ as a specific realization for a randomly drawn observation from the multivariate density f.

The generalized participation probability is:

$$P_{\boldsymbol{\theta}}(\mathcal{S}_i = 1) = \int_{v_p} ... \int_{v_1} f_{\boldsymbol{\eta}}\left(v_1, ...v_p\right) \left[1 - \Phi\left(-\boldsymbol{V'\gamma}\right)\right] dv_1...dv_p. \tag{4.39}$$

For sake of brevity we express $\mathcal{F}_{p+1}\left(z_{1i}, ..., z_{pi}, 0\right)$ in the following way:

$$\mathcal{F}_{p+1}\left(z_{1i}, ..., z_{pi}, 0\right) = \int_{v_p=-\infty}^{z_{pi}} ... \int_{v_1=-\infty}^{z_{1i}} f_{\boldsymbol{\eta}}\left(v_1, .., v_p\right) P\left(Z_i'\boldsymbol{\gamma} + \xi_{2i} < 0 | z_{1i} = v_1, ..., z_{pi} = v_p\right) dv_1...dv_p. \tag{4.40}$$

Then, for expressing the density function, we calculate the derivative of the cumulative distribution function (4.40) with respect to observation i's covariates vector z_i:

$$\frac{\partial^p \mathcal{F}_{p+1}\left(z_{1i}, ..., z_{pi}, 0\right)}{\partial z_{1i}...\partial z_{pi}} = f_{\boldsymbol{\eta}}\left(z_{1i}, ..., z_{pi}\right) P\left(\boldsymbol{Z_i'\gamma} + \xi_{2i} < 0 | z_i\right) = f_{\boldsymbol{\eta}}\left(z_{1i}, ..., z_{pi}\right) \Phi\left(-\boldsymbol{Z_i'\gamma}\right). \tag{4.41}$$

After taking the derivative of the truncated cumulative distribution function (equation (4.37)) with respect to z_i using the results obtained in equation (4.41) we get the truncated (constrained) data generation density function:

$$f_{\boldsymbol{\theta}}\left(z_{1i},...,z_{pi}|\boldsymbol{Z_i'\gamma}+\xi_{2i}>0\right)= \tag{4.42}$$
$$\frac{f_{\boldsymbol{\eta}}\left(z_{1i},...,z_{pi}\right)\left[1-\Phi\left(-\boldsymbol{Z_i'\gamma}\right)\right]}{\int_{v_p}...\int_{v_1}f_{\boldsymbol{\eta}}\left(v_1,...v_p\right)\left[1-\Phi\left(-\boldsymbol{V'\gamma}\right)\right]dv_1...dv_p}.$$

Note that in equation (4.42) the term $f_{\boldsymbol{\eta}}\left(z_{1i},...,z_{pi}\right)$ represents the non-truncated data generation density function;[21] the term $\left[1-\Phi\left(-\boldsymbol{Z_i'\gamma}\right)\right]$ represents the participation probability of individual-specific observation i (based on a Probit binary response model). The term in the denominator represents the probability of a randomly selected individual observation to be included in the truncated data.[22]

In the next section we demonstrate an implementation of the truncated multivariate model for the case of observations which are multivariate normally distributed.

4.1.10. Implementation for a multivariate normal distribution

The methodology used so far can be generalized to the multivariate case. Let's denote our non-truncated sample as a vector of random variables representing the participation covariates, such that each one is distributed normally. Without loss of generality, we allow for correlation among them (each covariate can be correlated with any other covariate) as characterized by the non-truncated data multivariate density function. As in the univariate truncated Probit model, we assume that this vector is multivariate normally distributed. Conditional on the participation constraint we can characterize the constrained data generation process.

For sake of brevity, we denote $\boldsymbol{\omega_v}=[v_1,...,v_p]$ and $\boldsymbol{\omega_{z_i}}=[z_{1i},...,z_{pi}]$ each describing the covariates vectors v and z respectively. The cumulative distribution function of the covariates vector z_i $(p\times 1)$ and Y_{2i} is:

$$D_i=P\left(Z_{1i}<z_{1i},...,Z_{pi}<z_{pi},Y_{2i}<0\right)= \tag{4.43}$$
$$\int_{v_p=-\infty}^{z_{pi}}...\int_{v_1=-\infty}^{z_{1i}}\phi_p\left(\boldsymbol{\omega_v},\boldsymbol{\mu_z},\boldsymbol{\Sigma}_{p\times p}\right)\left[1-\Phi\left(-\boldsymbol{V'\gamma}\right)\right]dv_1...dv_p.$$

Inserting (4.43) into the cumulative constrained distribution of our covariates vector $\boldsymbol{Z_i}$ to get the following conditional (on participation) cumulative distribution of the truncated data generation process:

[21] One possible way to obtain such a multivariate density is by using a Copula function.

[22] Despite high-dimensionality of this multiple integration, there is a simple method to evaluate it by random sampling of a vector from the multivariate distribution and averaging the resulting vector.

$$P\left(Z_{1i} < z_{1i}, ..., Z_{pi} < z_{pi} | \boldsymbol{Z_i'\gamma} + \xi_{2i} > 0\right) = \tag{4.44}$$

$$\frac{\Phi_p\left(\boldsymbol{\omega_{z_i}}, \boldsymbol{\mu_z}, \boldsymbol{\Sigma}_{p\times p}\right) - D_i}{\int_{v_p} ... \int_{v_1} \phi_p\left(\boldsymbol{\omega_v}, \boldsymbol{\mu_z}, \boldsymbol{\Sigma}_{p\times p}\right)\left[1 - \Phi\left(-\boldsymbol{V'\gamma}\right)\right] dv_1 ... dv_p}.$$

This enables us to construct the truncated (data generation process) density function:[23]

$$f_{\boldsymbol{\theta}}\left(z_{1i}, ..., z_{pi} | \boldsymbol{Z_i'\gamma} + \xi_{2i} > 0\right) = \tag{4.45}$$

$$\frac{\phi_p\left(\boldsymbol{\omega_{z_i}}, \boldsymbol{\mu_z}, \boldsymbol{\Sigma}_{p\times p}\right)\left[1 - \Phi\left(-\boldsymbol{Z_i'\gamma}\right)\right]}{\int_{v_p} ... \int_{v_1} \phi_p\left(\boldsymbol{\omega_v}, \boldsymbol{\mu_z}, \boldsymbol{\Sigma}_{p\times p}\right)\left[1 - \Phi\left(-\boldsymbol{V'\gamma}\right)\right] dv_1 ... dv_p}.$$

This truncated density function is required for the implementation of the maximum likelihood estimation method (which will be used in the ensuing section).

4.1.11. A generalization of the Likelihood function for multivariate normal regression

Similar to the log-likelihood function described in the univariate truncated Probit, we can express the log-likelihood function for the multivariate truncated Probit with p covariates:

$$\ln L = \sum_{i=1}^{n} \ln \phi_p\left(z_{1i}, ..., z_{pi}, \boldsymbol{\mu_z}, \boldsymbol{\Sigma}_{p\times p}\right) + \sum_{i=1}^{n} \ln\left[1 - \Phi\left(-\boldsymbol{Z_i'\gamma}\right)\right] \tag{4.46}$$

$$-n \ln\left(\int_{v_p} ... \int_{v_1} \phi_p\left(\boldsymbol{\omega_v}, \boldsymbol{\mu_z}, \boldsymbol{\Sigma}_{p\times p}\right)\left[1 - \Phi\left(-\boldsymbol{V'\gamma}\right)\right] dv_1 ... dv_p\right).$$

The arguments of the function $\phi_p(.)$ in (4.46) are the covariates $\boldsymbol{(z_1, ..., z_p)}$ expressed without any coefficients in order to avoid the problem of no identification of the $\boldsymbol{\gamma}$ vector. If however these covariates are expressed with coefficients it might leads to different estimates for the parameters of the vector $\boldsymbol{\gamma}$, due to the existence of an orthogonal transformation of the $\boldsymbol{(z_1, ..., z_p)}$ and the random disturbance ξ_2 that generated the transformed data, which is also a multivariate normally distributed. Another important restriction we impose by construction of the covariance matrix is that these covariates are orthogonal to the random disturbance. Additionally, in a similar fashion to the conventional Probit analysis, the random disturbance is assumed to be standardized normally distributed, implying that one does not need to estimate its variance.

The log-likelihood function in (4.46) involves a cumbersome high dimensional integration calculation and can be simplified as follows (see theorem (8.1) in the

[23] For computational convenience the multivariate normal density function is normalized by the factor $\left(\prod_{i\leq p} \frac{1}{\sigma_{z_i}}\right)$, such that we use the standardized normal vector.

Appendix):

$$\ln L = \sum_{i=1}^{n} \ln \phi_p \left(\boldsymbol{z_i}, \boldsymbol{\mu_z}, \boldsymbol{\Sigma}_{p\times p} \right) + \sum_{i=1}^{n} \ln \left[1 - \Phi \left(-\gamma_0 - \boldsymbol{z_i'} \boldsymbol{\gamma^*} \right) \right] - n \ln \left(1 - \Phi \left(- \frac{\gamma_0 + \boldsymbol{\mu_z}' \boldsymbol{\gamma^*}}{\sigma_w} \right) \right), \tag{4.47}$$

where $\sigma_w \equiv \sqrt{1 + \boldsymbol{\gamma'^*} \boldsymbol{\Sigma} \boldsymbol{\gamma^*}}$, $\boldsymbol{\gamma} \equiv [\boldsymbol{\gamma_0}, \boldsymbol{\gamma^*}]^T$ and $\boldsymbol{\gamma^*} \equiv [\gamma_1, ..., \gamma_p]^T$.

Denoting: $\lambda_\delta \equiv \dfrac{\phi \left(- \frac{\gamma_0 + \boldsymbol{\mu_z}' \boldsymbol{\gamma^*}}{\sigma_w} \right)}{1 - \Phi \left(- \frac{\gamma_0 + \boldsymbol{\mu_z}' \boldsymbol{\gamma^*}}{\sigma_w} \right)}$ and $\lambda_i \equiv \frac{\phi \left(-\gamma_0 - \boldsymbol{z_i'} \boldsymbol{\gamma^*} \right)}{1 - \Phi \left(-\gamma_0 - \boldsymbol{z_i'} \boldsymbol{\gamma^*} \right)}$.

The first order conditions obtained from the maximization of (4.47) are:

$$\frac{\partial \log L}{\partial \boldsymbol{\mu_z}} = \boldsymbol{\Sigma^{-1}} \sum_{i=1}^{n} \left(\boldsymbol{z_i} - \boldsymbol{\mu_z} \right) - n \lambda_\delta \frac{1}{\sigma_w} \boldsymbol{\gamma^*} = 0 \tag{4.48}$$

$$\frac{\partial \log L}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2} \left(n \boldsymbol{\Sigma^{-1}} - \boldsymbol{\Sigma^{-1}} \sum_{i=1}^{n} \left(\boldsymbol{z_i} - \boldsymbol{\mu_z} \right) \left(\boldsymbol{z_i} - \boldsymbol{\mu_z} \right)' \boldsymbol{\Sigma^{-1}} \right) + \frac{1}{2} n \lambda_\delta \frac{\gamma_0 + \boldsymbol{\mu_z}' \boldsymbol{\gamma^*}}{\sigma_w^3} \boldsymbol{\gamma^*} \boldsymbol{\gamma'^*} = 0 \tag{4.49}$$

$$\frac{\partial \log L}{\partial \gamma_0} = \sum_{i=1}^{n} \lambda_i - n \lambda_\delta \frac{1}{\sigma_w} = 0 \tag{4.50}$$

$$\frac{\partial \log L}{\partial \boldsymbol{\gamma^*}} = \sum_{i=1}^{n} \lambda_i \boldsymbol{z_i} + n \lambda_\delta \frac{(\gamma_0 + \boldsymbol{\mu_z}' \boldsymbol{\gamma^*}) \boldsymbol{\Sigma} \boldsymbol{\gamma^*} - \boldsymbol{\mu_z} \sigma_w^2}{\sigma_w^3} = 0. \tag{4.51}$$

We search for the parameter values $\boldsymbol{\mu_z}, \boldsymbol{\Sigma_{p\times p}}, \boldsymbol{\gamma}$ which solve equations (4.48) to (4.51). There are $p + \frac{1}{2}p(p-1)$ parameters to be estimated in $\boldsymbol{\Sigma}_{p\times p}$ and p parameters to be estimated in vector $\boldsymbol{\mu_z}$. Overall, we have $2p + \frac{1}{2}p(p-1)$ incidental parameters and $p+1$ parameters of interest (including the intercept).[24]

In the next section we will extend our theory by an inclusion of dichotomous variables.

4.2. Implementation for dichotomous variables

Until now we have shown our model's estimation procedure relying on the assumption that the entire covariates data in the complete (non-truncated) sample are multivariate normally distributed, which implies that each one of the covariates is treated as a continuous variable. However, in many applications in

[24] The covariance between any two non-participants' covariates (z_i, z_j) is: $\mathbb{COV}(z_j, z_j) = \int ... \int v_i v_j f(v_1, ..., v_p | \mathcal{S} = 0) dv_1 ... dv_p - \mathbb{E}(z_i)\mathbb{E}(z_j)$.

social sciences, dichotomous (dummy) variables are employed in order to classify the samples into distinguishable cohorts of interest. Dichotomous variables are useful to predict the causal effect of a new policy or a treatment on an outcome variable. In many cases, counterfactual analysis (DiNardo et al., 1996) is employed in order to measure the contribution of each factor separately on the variation in the outcome variable. Lets assume that in order to measure the discrimination between A and B in the selection process, one needs to estimate the counterfactual probability of random observation from cohort A to be selected given another state of nature, for instance, being a member in cohort B. In the censored sample case one can estimate directly the selection rule's equation using a binary response model utilizing dichotomous variables to distinguish among the cohorts. Since both participants and non-participants are observed, the complete data distribution function on the selection equation's covariates are observed. Post estimation of the selection equation both cohort A's actual and counterfactual expected probabilities to participate can be evaluated using cohort A's data. The procedure involves averaging the predicted probability to participate using cohort A's coefficients and doing the same calculation using cohort B's coefficients.

However, in the case of a truncated sample, the control groups, the non-participants of any cohort, are omitted from the sample in the very beginning. Therefore, data reconstruction can be helpful in recovering the characteristics' distribution function for each one of the cohorts A and B. For instance, one can estimate what would be the percentage of employed women if women were selected into the labor force according to men's selection rule. Using a simulation we can sample women observations according to women characteristics distribution functions, and evaluate the percentage of employed women, by using men's selection rule equation.

Amending our methodology to cases where dichotomous variables are present we introduce a cohorts set $\mathcal{J} \equiv \{1, 2, ..., J-1, J\}$ consisting of J mutually exclusive cohorts such that each observation i belongs to a specific one cohort $j \in \mathcal{J}$. The number of participants and the number of observations in the j'th cohort' are denoted by n_j and N_j respectively.

Denoting the observations indices which belong to j'th cohort by I_j for each $j \in \mathcal{J}$, and a dichotomous random variable for being a memeber in the $j'th$ cohort, defined as:

$$D_{ji} = \begin{cases} 1, & \text{if } i \in I_j \\ 0, & \text{if } i \notin I_j \end{cases}. \tag{4.52}$$

It must hold that $\sum_{j=1}^{J} D_{ji} = 1 \ \forall i$ since all the cohorts are mutually exclusive by construction.

Without loss of generality, each cohort $j \in \mathcal{J}$ is characterized by its own coefficients vector $\boldsymbol{\gamma}_{\boldsymbol{j}}^{*} \equiv [\gamma_{j1}, ..., \gamma_{jp}]^T$ of size $(p \times 1)$, and an intercept γ_{0j} for the j'th cohort.

Thus, the selection rule equations in the presence of dichotomous variables can be characterized in the following way:[25]

$$Y_{2i} = \sum_{j=1}^{J} D_{ji}(\gamma_{0j} + \boldsymbol{z_i}\boldsymbol{\gamma_j^*}) + \xi_{2i}, \tag{4.53}$$

By construction, the specification of equation (4.53) allows each cohort to have a different parameters set, such that it is possible that the same covariates impact differently on the probability to be selected for different cohorts.

Conditional on belonging to the j'th cohort, the continuous variables are multivariate normally distributed, such that $\boldsymbol{z}|\boldsymbol{D} = \boldsymbol{D_j} \sim \mathcal{N}_p(\boldsymbol{\mu_z^j}, \boldsymbol{\Sigma_{p\times p}^j})$, where $\boldsymbol{\eta_j}$ is composed of the nuisance parameters $\boldsymbol{\mu_z^j}$ and $\boldsymbol{\Sigma_{p\times p}^j}$.

Using a weighted sum of all the conditional densities it follows that the unconditional (on cohort) complete data distribution of the continuous variables is a Gaussian mixture, defined as:

$$f_{\boldsymbol{\eta}}(\boldsymbol{z}) = \sum_{j=1}^{J} P(\boldsymbol{D} = \boldsymbol{D_j}) f_{\boldsymbol{\eta}_j}(\boldsymbol{z}|\boldsymbol{D} = \boldsymbol{D_j}), \tag{4.54}$$

where $\boldsymbol{\eta} \equiv [\boldsymbol{\eta_1}, ..., \boldsymbol{\eta_J}]^T$ is a vector of size $(J \times 1)$ consisting of the parameters which characterize each cohort's characteristics distribution function. Consequently, the dichotomous variables have impact on the multivariate density function of the continuous variables.

Conditional on being selected, a truncated multivariate density mixture is obtained:

$$f_{\boldsymbol{\theta}}(\boldsymbol{z}|\mathcal{S}_i = 1) = \sum_{j=1}^{J} P(\boldsymbol{D} = \boldsymbol{D_j}|\mathcal{S}_i = 1) f_{\boldsymbol{\theta_j}}(\boldsymbol{z}|\mathcal{S}_i = 1, \boldsymbol{D} = \boldsymbol{D_j}), \tag{4.55}$$

Additionally, the conditional (on j'th cohort) truncated multivariate density function is:

$$\begin{aligned} &f_{\boldsymbol{\theta_j}}(\boldsymbol{z_i}|\mathcal{S}_i = 1, D_{ji} = 1) \\ &\qquad = \frac{f_{\boldsymbol{\eta_j}}(\boldsymbol{z_i}|D_{ji} = 1) P_{\boldsymbol{\gamma_j}}(\mathcal{S}_i = 1|D_{ji} = 1, \boldsymbol{z_i}) P(D_{ji} = 1)}{P_{\boldsymbol{\theta_j}}(\mathcal{S}_i = 1|D_{ji} = 1) P(D_{ji} = 1)}, \end{aligned} \tag{4.56}$$

which can be simplified to:

$$f_{\boldsymbol{\theta_j}}(\boldsymbol{z_i}|\mathcal{S}_i = 1, D_{ji} = 1) = \frac{f_{\boldsymbol{\eta_j}}(\boldsymbol{z_i}|D_{ji} = 1) P_{\boldsymbol{\gamma_j}}(\mathcal{S}_i = 1|D_{ji} = 1, \boldsymbol{z_i})}{P_{\boldsymbol{\theta_j}}(\mathcal{S}_i = 1|D_{ji} = 1)} \tag{4.57}$$

In practice, the probability function $P(\boldsymbol{D} = \boldsymbol{D_j})$ in (4.54) is unknown since it represents the proportion of any sub-population $j \in \mathcal{J}$ in the unobserved com-

[25] In case that all the cohorts belonging to the cohorts subset $J' \subset J$ are only differentiated by the intercept, one can impose a restriction on the slope parameters to be same for all of these cohorts, by setting $\boldsymbol{\gamma_j^*} = \boldsymbol{\gamma_{j'}^*}$ for all $j, j' \in J'$.

plete data. Thus, we suggest an estimation method to estimate this unknown proportions without assuming a specific distribution function for the dichotomous variables.

At first, we define the the proportion of participants belonging to each cohort conditional on being selected:

$$\pi_j \equiv P(\boldsymbol{D} = \boldsymbol{D_j}|S = 1). \tag{4.58}$$

This proportion can be estimated non-parameterically (since it is observed in the truncated sample) as:

$$\hat{\pi_j} = \frac{n_j}{\sum_{j=1}^{n} n_j}, \quad j = 1, ..., J. \tag{4.59}$$

Then, we formulate the likelihood function as a product of the truncated density in (4.55), using the estimators calculated in (4.59):

$$L = \prod_{j=1}^{J}\prod_{i=1}^{n}\left[\hat{\pi}_j f_{\boldsymbol{\theta_j}}(\boldsymbol{z_i}|\mathcal{S}_i = 1, D_{ji} = 1)\right]^{D_{ji}} \tag{4.60}$$

We obtain the following likelihood function:

$$L = \prod_{j=1}^{J}\prod_{i=1}^{n}\left[\hat{\pi}_j \frac{f_{\boldsymbol{\eta_j}}(\boldsymbol{z_i}|D_{ji} = 1)P_{\boldsymbol{\gamma_j}}(\mathcal{S}_i = 1|D_{ji} = 1, \boldsymbol{z_i})}{P_{\boldsymbol{\theta_j}}(\mathcal{S}_i = 1|D_{ji} = 1)}\right]^{D_{ji}} \tag{4.61}$$

This formulation is based on being able to classify the observations into cohorts and since the $\pi_1, ..., \pi_j$ estimators are not functions of the unknown parameters to be estimated $\boldsymbol{\theta}$, one can simplify the likelihood function and formulate it as a product of the *conditional* (on cohort) truncated densities solely:

$$L^* = \prod_{j=1}^{J}\prod_{i=1}^{n}\left[\frac{f_{\boldsymbol{\eta_j}}(\boldsymbol{z_i}|D_{ji} = 1)P_{\boldsymbol{\gamma_j}}(\mathcal{S}_i = 1|D_{ji} = 1, \boldsymbol{z_i})}{P_{\boldsymbol{\theta_j}}(\mathcal{S}_i = 1|D_{ji} = 1)}\right]^{D_{ji}} \tag{4.62}$$

Thus, one does not need to assume a prior probability function for the dichotomous variables.

The probability to participate for an observation from the j'th cohorts with the covariates vector $\boldsymbol{z_i} = [z_{1_1}, ..., z_{p_1}]^T$ is defined as follows:

$$P_{\boldsymbol{\gamma_j}}(\mathcal{S}_i = 1|\boldsymbol{z_i}, \boldsymbol{D} = \boldsymbol{D_j}) = 1 - \Phi\left(-\gamma_{j_0} - \boldsymbol{z_i'}\boldsymbol{\gamma_j^*}\right), \tag{4.63}$$

where $\boldsymbol{\gamma_j} = [\gamma_{j_0}, \boldsymbol{\gamma_j^*}]^T$ is the j'th cohort selection equation's coefficients vector which contains a scalar γ_{j_0} (for the intercept) and a vector $\boldsymbol{\gamma_j^*}$ of size $(p \times 1)$.

The probability to participate for a random observation from the j'th cohort

as a function of the parameters vector $\boldsymbol{\theta_j} \equiv [\boldsymbol{\eta_j}, \boldsymbol{\gamma_j}]^T$ is:

$$P_{\boldsymbol{\theta}_j}(\mathcal{S}=1|\boldsymbol{D}=\boldsymbol{D_j}) = 1 - \Phi\left(-\frac{\gamma_{j_0} + \boldsymbol{\mu_z'^j}\boldsymbol{\gamma_j^*}}{\sqrt{1+\boldsymbol{\gamma_j'^*}\boldsymbol{\Sigma^j}\boldsymbol{\gamma_j^*}}}\right). \tag{4.64}$$

The log-likelihood function to be maximized with respect to the parameters $\boldsymbol{\theta}$ is defined as follows:

$$\begin{aligned}\ln L^* = &\sum_{j=1}^{J}\sum_{i=1}^{n} D_{ji} \ln \phi_p\left(\boldsymbol{z_i}, \boldsymbol{\mu_z^j}, \boldsymbol{\Sigma}^j_{p\times p}\right) + \sum_{j=1}^{J}\sum_{i=1}^{n} D_{ji} \ln\left[1-\Phi\left(-\boldsymbol{Z_i'}\boldsymbol{\gamma_j}\right)\right] \\ &- \sum_{j=1}^{J} n_j \ln\left(1-\Phi\left(-\frac{\gamma_{j_0}+\boldsymbol{\mu_z'^j}\boldsymbol{\gamma_j^*}}{\sigma_{jw}}\right)\right),\end{aligned} \tag{4.65}$$

where $\sigma_{jw} \equiv \sqrt{1+\boldsymbol{\gamma_j'^*}\boldsymbol{\Sigma^j}\boldsymbol{\gamma_j^*}}$.

Then, post-estimation one can calculate the probability for a random observation to participate for each cohort separately using (4.64) and compare it to the participants proportion in the j'th cohort as follows:

$$\hat{P}_{\boldsymbol{\theta}_j}(\mathcal{S}=1|\boldsymbol{D}=\boldsymbol{D_j}) = \frac{n_j}{N_j}, \tag{4.66}$$

implying that by extracting N_j from (4.66) we obtain the following estimator:

$$\hat{N}_j = \frac{n_j}{\hat{P}_{\boldsymbol{\theta}_j}(\mathcal{S}=1|\boldsymbol{D}=\boldsymbol{D_j})}, \tag{4.67}$$

for the total number of observations in the j'th cohort (participants and non-participants).

Then we suggest the following estimator for the proportion of participants in the complete sample as:

$$\hat{P}(\mathcal{S}=1) = \frac{\sum_{j=1}^{J} n_j}{\sum_{j=1}^{J} \hat{N}_j} \tag{4.68}$$

Additionally, an estimator for the proportion of observations in the complete data which belong to the j'th cohort can be obtained, as follows:

$$\hat{P}(\boldsymbol{D}=\boldsymbol{D_j}) = \frac{\hat{N}_j}{\sum_{j=1}^{J} \hat{N}_j} \tag{4.69}$$

Next, we discuss the estimation procedure for the case where the binary response model is completely separated by some dichotomous covariates.

4.3. Completely separating dichotmous variables

Let examine the case where the sample consists of J mutually exclusive cohorts and as before, can be split into two complement subsets $\mathcal{J}_1$ and $\mathcal{J}_2$, which are

differentiated also by the binary outcome variable:

$$s_j = \begin{cases} 1, & \text{if } j \in \mathcal{J}_1 \\ 0, & \text{if } j \notin \mathcal{J}_1 \end{cases}, \quad j = 1, ..., J. \tag{4.70}$$

For instance, the outcome variable is a binary employment status: employed or unemployed, such that the cohorts belonging to $\mathcal{J}_1$ subset are employed men ($s_j = 1$) and all other cohorts belonging to $\mathcal{J}_2$ are unemployed women ($s_j = 0$). If one is interested in estimating the effect of gender on the probability of being employed, a conventional binary response model cannot be applied to estimate this coefficient using the observed data, because the outcome variable is completely separated by gender. However, a complete separation in the truncated sample, does not imply a complete separation in the complete sample. In the truncated sample two control groups are unobserved: the unemployed men and the employed women. In order to bypass the problem of complete separation one needs to recover the unobserved control groups. The methodology to recover the complete data distribution function relies on the principle of characterizing the complete data distribution function as a function of the parameters estimated using the observed truncated data. The unobserved complete data density function is the same as in equation (4.54).

Conditional on the selection variable, a truncated multivariate density mixture is obtained:

$$f_{\boldsymbol{\theta}}(\boldsymbol{z}|\mathcal{S}_i = s_j) = \sum_{j=1}^{J} P(\boldsymbol{D} = \boldsymbol{D_j}|\mathcal{S}_i = s_j) f_{\boldsymbol{\theta_j}}(\boldsymbol{z}|\mathcal{S}_i = s_j, \boldsymbol{D} = \boldsymbol{D_j}). \tag{4.71}$$

The probabilities $P(\boldsymbol{D} = \boldsymbol{D_j}|\mathcal{S}_i = s_j)$ are estimated non-parameterically in a similar fashion to (4.59) and are not functions of the unknown parameters to be estimated in vector $\boldsymbol{\theta}$. Thus, the likelihood function to be maximized with respect to the unknown parameters is:[26]

$$L^* = \prod_{j=1}^{J} \prod_{i=1}^{n} \left[\frac{f_{\boldsymbol{\eta_j}}(\boldsymbol{z_i}|D_{ji} = 1) P_{\boldsymbol{\gamma_j}}(\mathcal{S}_i = s_j|D_{ji} = 1, \boldsymbol{z_i})}{P_{\boldsymbol{\theta_j}}(\mathcal{S}_i = s_j|D_{ji} = 1)} \right]^{D_{ji}}. \tag{4.72}$$

The probability function of the selection variable conditional on the covariates vector $\boldsymbol{z_i}$ and belonging to the j'th cohort, is described as follows:

$$P_{\boldsymbol{\gamma_j}}(\mathcal{S}_i = s_j|D_{ji} = 1, \boldsymbol{z_i}) = \begin{cases} 1 - \Phi(-\boldsymbol{Z_i'\gamma_j}), & \text{if } s_j = 1 \\ \Phi(-\boldsymbol{Z_i'\gamma_j}), & \text{if } s_j = 0 \end{cases}, \quad j = 1, ..., J. \tag{4.73}$$

The probability function of the selection variable conditional on belonging to

[26] We assume a unique parameters set for each cohort in order to simplify the likelihood function.

the j'th cohort for a random observation is described as follows:

$$P_{\boldsymbol{\theta_j}}(\mathcal{S}_i = s_j | D_{ji} = 1) = \begin{cases} 1 - \Phi(-\frac{\gamma_{j0} + \boldsymbol{\mu_z'^j \gamma_j^*}}{\sigma_{jw}}), & \text{if } s_j = 1 \\ \Phi(-\frac{\gamma_{j0} + \boldsymbol{\mu_z'^j \gamma_j^*}}{\sigma_{jw}}), & \text{if } s_j = 0 \end{cases}, \quad j = 1, ..., J. \tag{4.74}$$

In both equations (4.73) and (4.74) the probabilities for the mutually exclusive events $s_j = 1$ and $s_j = 0$ to be selected and not selected respectively are complimentary probabilities.

It follows that the log-likelihood function obtained from (4.72) using (4.73) and (4.74) is:

$$\begin{aligned} \ln L^* = &\sum_{j=1}^{J}\sum_{i=1}^{n} D_{ji} \ln \phi_p^j(\boldsymbol{z_i}) + \sum_{j=1}^{J}\sum_{i=1}^{n} D_{ji} \ln \left[P_{\boldsymbol{\gamma_j}}(\mathcal{S}_i = s_j | D_{ji} = 1, \boldsymbol{z_i})\right] \\ &- \sum_{j=1}^{J} n_j \ln \left(P_{\boldsymbol{\theta_j}}(\mathcal{S}_i = s_j | D_{ji} = 1)\right), \end{aligned} \tag{4.75}$$

where $\phi_p^j(\boldsymbol{z_i}) \equiv \phi_p(\boldsymbol{z_i}, \boldsymbol{\mu^j}, \boldsymbol{\Sigma^j_{p\times p}})$.

Once the parameters vector $\boldsymbol{\theta}$ is estimated, each dichotomous variable effect on the probability to participate is obtained in a similar fashion to the conventional Probit analysis by calculation of the probability to be selected (post estimation) for a given observation, as if the complete data with all the control groups are accessible. In the next section we will validate our theory and test for robustness by using Monte-Carlo simulations.

5. Validation and robustness

In this section we examine how well our methodology performs in the case of truncated data relative to conventional Probit analysis. In order to validate our methodology we use Monte-carlo simulations. In each iteration the simulation is based on generation of random data consisting of covariates vectors which are drawn independently from the specified multivariate density. For simplicity we use the multivariate normal density function and construct an endogenous selection rule as a monotonic function of the generated sample covariates and a random disturbance as in equation (2.2). By construction each one of the covariates is an exogenous variable which is uncorrelated with the selection equation's random disturbance. This procedure is intended to generate two nested samples (for each iteration). The first sample is the non-truncated data, in the sense that we have both participants and non-participants. The second sample is a subsample of the first one, consisting solely on observations preserving the selection rule (participants) referred to as the truncated data. The selection rule depends on a weighted sum of the observation covariates (the endogenous component) and a random disturbance. As before, the weights are the $((p+1)\times 1)$ coefficients vector, $\boldsymbol{\gamma}$, in the participation equation. However, this time we deliberately

set these coefficients' values (the true parameters in the population), in order to make a comparison between their estimated values and their theoretical values. For that purpose, we estimate the standard deviation for each estimated parameter as well. The procedure involves estimation of the following two different models: (i) The conventional Probit estimation using the non-truncated data. All the participants' and non-participants' covariates are used in the estimation. (ii) The Truncated Probit estimation using the truncated data. Only participants' observations are used in the estimation.

In table (II) we present simulations results. We sample 200,000 observations from two independent distributions: the covariate z_i is drawn from a normal distribution with the two parameters: μ_z (stands for its expectation), and σ_z (stands for its standard deviation). The random disturbance is drawn from a standard normal distribution ξ_{2i}. Then based on these two random variables we construct the participation latent variable $Y_{2i} = \boldsymbol{Z_i'\gamma} + \xi_{2i}$,[27] such that $\boldsymbol{Z_i} = [1, z_i]^T$. Several parameters settings are examined. In the first parameters setting, we set $[\mu_z, \sigma_z]^T = [5, 2.1]^T$ and $\boldsymbol{\gamma} = [-2.55, 1.23]^T$. In the second parameters setting, we set $[\mu_z, \sigma_z]^T = [-8, 4]^T$ and $\boldsymbol{\gamma} = [-1.55, -0.75]^T$.[28] Then we estimate the two models as described above, and repeat this procedure 20,000 times.

Based on 20,000 samples randomly generated, in the first parameters setting, the conventional Probit estimators produce values of -2.5501 and 1.2301 for the intercept and the slope respectively (on average) and the corresponding standard deviations of 0.0215 and 0.0081 respectively. The Truncated Probit estimators produce values of -2.5508 and 1.2304 on average for the intercept and the slope respectively, with corresponding standard deviations of 0.0353 and 0.0223 respectively.

In the second parameters setting, the conventional Probit estimators produce values of -1.5501 and -0.75 for the intercept and the slope respectively (on average) and the corresponding standard deviations of 0.0172 and 0.0055 respectively. The Truncated Probit estimators produce values of -1.5505 and -0.7504 on average for the intercept and the slope respectively, with corresponding standard deviations of 0.0246 and 0.0143 respectively. We have also calculated the bootstraps standard deviations in order to measure how accurately that likelihood standard deviations were calculated.

We conduct sensitivity test to measure the influence of an increase in number of observations on the accuracy of the truncated sample estimators. Using the

[27] We denote participant observations by $\mathcal{S}_i = 1$ when $Y_{2i} > 0$.
[28] These figures are set arbitrarily for the sake of the Monte Carlo simulations.

following root mean square gap $R_j(n)$ measure:[29]

$$R_j(n) = \left(\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{\gamma}_{i,j}^{tp} - \hat{\gamma}_{i,j}^{p}}{\hat{\gamma}_{i,j}^{p}} \right)^2 \right)^{1/2}, \tag{5.1}$$

where $\gamma_{i,j}^{p}$ stands for the Probit's j'th coefficient estimated using the full sample, and where $\gamma_{i,j}^{tp}$ stands for the truncated Probit's j'th coefficient estimated using the truncated sample. The sample size is denoted by n_k.

Another accuracy measure is the change in root mean square gap $\Delta R_j(n)$:

$$\Delta R_j(n_1, n_2) = \left(\alpha \times \frac{R_j(n_2) - R_j(n_1)}{n_2 - n_1} \right). \tag{5.2}$$

This is the measure of the contribution of α additional observations to the truncated Probit estimators' accuracy (in terms of proximity to Probit estimators). For $\alpha = 1000$ we obtain the results described in table (I).

The last estimators' accuracy measure is the δ coefficient used for the calculation of the estimators' standard deviations convergence rate n^{δ} with respect to the sample size. This coefficient is calculated based on the following ratio:

$$\delta = \left(\ln \tfrac{\sigma_1}{\sigma_2}\right) / \left(\ln \tfrac{n_2}{n_1}\right). \tag{5.3}$$

Based on table (I), the distance between truncated Probit and full sample Probit estimators is getting smaller with number of observations. This is captured by R_j which implies that there is a much more similarity between truncated Probit and full sample Probit as number of observation increases. Additionally, the contribution of 1,000 additional observations to the truncated Probit estimators' accuracy (in terms of proximity to Probit estimators) is higher the smaller the initial sample. This is embodied in the value of ΔR_j which becomes smaller and eventually approaches zero as number of observations increases. Additionally, the convergence rate in truncated Probit is slightly above $\sqrt{n}$. For a sample of 3,000 observations the estimators $\hat{\gamma}_0$ and $\hat{\gamma}_1$ have both $O(1/n^{0.57})$ standard deviation,[30] implying that for each of these estimators the convergence rate is $n^{0.57}$. When the number of observations increases to 20,000 the convergence rate decreases to the conventional $\sqrt{n}$ convergence rate.[31] Similarly, for comparison we present

[29] The standardized root mean square error (rmse) based on n Monte-Carlo simulations is based on a similar formula, which is: $\mathrm{R}\left(\hat{\theta}\right) = \left(\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{\theta}_i - \theta}{\theta} \right)^2 \right)^{1/2}$. However, our motivation is to find the relative accuracy of the truncated Probit in comparison to full sample Probit.

[30] The O notation's formal definition is as follows: For two sequences of non-negative numbers $\{a_n\}$ and $\{b_n\}$, there exist two positive constants A and B such that $A \leq a_n/b_n \leq B$. Then, if $a_n = O(b_n)$ and b_n converges to zero as n approaches infinity, then a_n also convergences to zero with the same convergence rate. Thus, $O(n^{-0.57})$ implies that the estimator's standard deviation converges to zero at the same rate as $n^{-0.57}$ approaches zero, when n approaches infinity. Consequently, doubling the sample size, shrinks the standard deviation of $\hat{\gamma}_1$ by $2^{0.574} = 1.488$ which is 5% faster than the conventional Probit convergence rate ($2^{0.508} = 1.422$).

[31] Ordinary Least Square errors convergence rate is $\sqrt{n}$ (Murray, 2005). Many semi para-

the full sample Probit convergence rate for the aforementioned estimators, which is almost constant and close to $\sqrt{n}$ convergence rate. It is apparent from the results presented, that both models (the full sample model and the truncated one) produce nearly the same results. This is especially true for samples with over 5,000 observations. These results points to the validity and robustness of our correction for the selectivity bias generated by truncated samples. As can be seen, from entries in table (II) through table (V) the bootstraps standard deviations are almost identical to the conventional standard deviations, pointing to the high accuracy of our proposed method.

TABLE I

Convergence Mesaures

Parameter	Sample size							
	3,000	5,000	8,000	10,000	12,000	15,000	20,000	50,000
R_j								
γ_0	0.1127	0.0768	0.0591	0.0515	0.0469	0.0412	0.0355	0.0219
γ_1	0.1737	0.1186	0.0912	0.0791	0.0717	0.0635	0.0551	0.0341
ΔR_j								
γ_0	-0.0179	-0.0059	-0.0038	-0.0023	-0.0019	-0.0011	-0.0005	-0.0001
γ_1	-0.0276	-0.0091	-0.0061	-0.0037	-0.0028	-0.0017	-0.0007	-0.0001
Truncated Probit: δ coefficient (n^δ is the convergence rate)								
γ_0	0.5711	0.5387	0.5315	0.5165	0.5210	0.5125	0.5064	0.5022
γ_1	0.5738	0.5405	0.5322	0.5173	0.5257	0.5119	0.5066	0.5023
Probit: δ coefficient (n^δ is the convergence rate)								
γ_0	0.5065	0.5037	0.5032	0.5022	0.5036	0.5011	0.5008	0.5002
γ_1	0.5079	0.5041	0.5036	0.5033	0.5044	0.5010	0.5010	0.5002

Note: We examine three different measures. First, the relative difference between Probit and truncated Probit estimators is calculated based on equation (5.1) for covergence rate's evaluation (as a function of observations' number). Second, the marginal effect of increasing the sample by 1000 observations is calculated. Third the convergence rate is calculated for both the truncated Probit and the full sample Probit. The estimators' convergence rate is measured by n^δ, implying that multiplying the sample size by 2, shrinks the estimators' standard deviations by 2^δ.

metric regression models are characterized by $\sqrt{n}$ consistent estimators (Robinson, 1988).

TABLE II

Monte Carlo Simulation.

Parameter	Probit				Truncated Probit			
	Mean	Median	Std.	Std.(b)	Mean	Median	Std.	Std. (b)
N = 200,000								
$\gamma_0 = -2.55$	-2.5501	-2.5499	0.0215	0.0213	-2.5508	-2.5506	0.0353	0.0343
$\gamma_1 = 1.23$	1.2301	1.230	0.0081	0.0080	1.2304	1.2300	0.0223	0.0217
$\sigma_z = 2.1$	0	0	0	0	2.1001	2.1001	0.0081	0.0080
$\mu_z = 5$	0	0	0	0	4.9997	4.9999	0.0139	0.0138
$\gamma_0 = -1.55$	-1.5501	-1.5500	0.0172	0.0172	-1.5505	-1.5502	0.0246	0.0243
$\gamma_1 = -0.75$	-0.7500	-0.7500	0.0055	0.0055	-0.7504	-0.7501	0.0143	0.0141
$\sigma_z = 4$	0	0	0	0	4	3.9998	0.0134	0.0133
$\mu_z = -8$	0	0	0	0	-8	-8.0004	0.0211	0.0210
N = 50,000								
$\gamma_0 = -2.55$	-2.5508	-2.5506	0.043	0.0428	-2.5544	-2.5527	0.0707	0.0689
$\gamma_1 = 1.23$	1.2303	1.2301	0.0162	0.0161	1.2323	1.231	0.0447	0.0439
$\sigma_z = 2.1$	0	0	0	0	2.1004	2.1002	0.0162	0.0161
$\mu_z = 5$	0	0	0	0	4.9991	4.9999	0.0278	0.0277
$\gamma_0 = -1.55$	-1.5513	-1.5511	0.0344	0.034	-1.5532	-1.5533	0.0493	0.0478
$\gamma_1 = -0.75$	-0.7505	-0.7504	0.0111	0.0110	-0.7517	-0.7506	0.0288	0.0283
$\sigma_z = 4$	0	0	0	0	4.0004	4.0000	0.0269	0.0267
$\mu_z = -8$	0	0	0	0	-7.9987	-7.9996	0.0422	0.042
N = 20,000								
$\gamma_0 = -2.55$	-2.5527	-2.5526	0.0680	0.0672	-2.5602	-2.5570	0.1124	0.1097
$\gamma_1 = 1.23$	1.2311	1.2308	0.0256	0.0253	1.2356	1.2331	0.0712	0.0703
$\sigma_z = 2.1$	0	0	0	0	2.1006	2.0996	0.0256	0.0257
$\mu_z = 5$	0	0	0	0	4.9977	5.0001	0.0443	0.0446
$\gamma_0 = -1.55$	-1.5544	-1.5521	0.0772	0.0775	-1.5647	-1.5611	0.1117	0.1134
$\gamma_1 = -0.75$	-0.7517	-0.7512	0.0249	0.0251	-0.7588	-0.7540	0.0657	0.0671
$\sigma_z = 4$	-	-	-	-	4.0021	4.0005	0.0605	0.0607
$\mu_z = -8$	-	-	-	-	-7.9944	-7.9996	0.0955	0.0960

Note: We estimate by maximum likelihood method the parameters for the truncated Probit and the conventional Probit, and compute the standard deviation in every random sample consisting of N observations. Then, we calculate for these estimates the mean, median and standard deviation over all data sets. The Monte Carlo standard deviations of the simulation are the average standard deviations over all data sets (computed as a function of the maximum likelihood Hessian). Std(b) stands for bootstrapping standard deviation.

TABLE III

Monte Carlo Simulation.

Parameter	Probit				Truncated Probit			
	Mean	Median	Std.	Std.(b)	Mean	Median	Std.	Std. (b)
N = 15,000								
$\gamma_0 = -2.55$	-2.5533	-2.5514	0.0786	0.0782	-2.5649	-2.559	0.1303	0.1276
$\gamma_1 = 1.23$	1.2313	1.2305	0.0296	0.0294	1.2382	1.2337	0.0824	0.0813
$\sigma_z = 2.1$	0	0	0	0	2.1009	2.1	0.0296	0.0295
$\mu_z = 5$	0	0	0	0	4.9972	5.0003	0.0512	0.0511
$\gamma_0 = -1.55$	-1.5526	-1.5513	0.0629	0.0621	-1.5589	-1.5567	0.0906	0.0887
$\gamma_1 = -0.75$	-0.7511	-0.7506	0.0203	0.0201	-0.7558	-0.753	0.0531	0.0524
$\sigma_z = 4$	0	0	0	0	4.0011	3.9996	0.0493	0.0491
$\mu_z = -8$	0	0	0	0	-7.9965	-7.9996	0.0776	0.0769
N = 12,000								
$\gamma_0 = -2.55$	-2.555	-2.5534	0.0879	0.0871	-2.569	-2.5632	0.1463	0.1432
$\gamma_1 = 1.23$	1.232	1.231	0.0331	0.0327	1.2407	1.2349	0.0927	0.0909
$\sigma_z = 2.1$	0	0	0	0	2.1008	2.0995	0.0332	0.0331
$\mu_z = 5$	0	0	0	0	4.9969	5.0007	0.0574	0.0574
$\gamma_0 = -1.55$	-1.5532	-1.5523	0.0704	0.0699	-1.5617	-1.5589	0.1016	0.1001
$\gamma_1 = -0.75$	-0.7514	-0.7508	0.0227	0.0227	-0.7571	-0.7534	0.0596	0.0594
$\sigma_z = 4$	0	0	0	0	4.0019	4.0008	0.0552	0.0551
$\mu_z = -8$	0	0	0	0	-7.9949	-7.999	0.087	0.0873
N = 10,000								
$\gamma_0 = -2.55$	-2.5552	-2.5530	0.0963	0.0969	-2.5726	-2.5641	0.1609	0.1628
$\gamma_1 = 1.23$	1.2323	1.2312	0.0363	0.0365	1.2427	1.2353	0.1019	0.1035
$\sigma_z = 2.1$	-	-	-	-	2.1012	2.0998	0.0364	0.0365
$\mu_z = 5$	-	-	-	-	4.9956	5.0003	0.0632	0.0637
$\gamma_0 = -1.55$	-1.5544	-1.5521	0.0772	0.0775	-1.5647	-1.5611	0.1117	0.1134
$\gamma_1 = -0.75$	-0.7517	-0.7512	0.0249	0.0251	-0.7588	-0.7540	0.0657	0.0671
$\sigma_z = 4$	-	-	-	-	4.0021	4.0005	0.0605	0.0607
$\mu_z = -8$	-	-	-	-	-7.9944	-7.9996	0.0955	0.096

Note: We estimate by maximum likelihood method the parameters for the truncated Probit and the conventional Probit, and compute the standard deviation in every random sample consisting of N observations. Then, we calculate for these estimates the mean, median and standard deviation over all data sets. The Monte Carlo standard deviations of the simulation are the average standard deviations over all data sets (computed as a function of the maximum likelihood Hessian). Std(b) stands for bootstrapping standard deviation.

TABLE IV
MONTE CARLO SIMULATION.

Parameter	Probit				Truncated Probit			
	Mean	Median	Std.	Std.(b)	Mean	Median	Std.	Std. (b)
N = 8000								
$\gamma_0 = -2.55$	-2.5567	-2.5534	0.1078	0.1072	-2.5811	-2.5698	0.1809	0.1797
$\gamma_1 = 1.23$	1.2330	1.2317	0.0406	0.0404	1.2474	1.2386	0.1146	0.1157
$\sigma_z = 2.1$	0	0	0	0	2.1017	2.1003	0.0408	0.0411
$\mu_z = 5$	0	0	0	0	4.9945	4.9996	0.0709	0.0716
$\gamma_0 = -1.55$	-1.555	-1.5528	0.0863	0.0857	-1.5666	-1.5632	0.1252	0.1224
$\gamma_1 = -0.75$	-0.7521	-0.7512	0.0279	0.0276	-0.7601	-0.7546	0.0738	0.0738
$\sigma_z = 4$	0	0	0	0	4.0026	4.0008	0.0678	0.0672
$\mu_z = -8$	0	0	0	0	-7.993	-7.9988	0.1073	0.1074
N = 5000								
$\gamma_0 = -2.55$	-2.5605	-2.5564	0.1366	0.137	-2.5999	-2.5824	0.2328	0.2307
$\gamma_1 = 1.23$	1.2344	1.2323	0.0514	0.0516	1.2578	1.2423	0.1477	0.1487
$\sigma_z = 2.1$	0	0	0	0	2.1028	2.1002	0.0519	0.0522
$\mu_z = 5$	0	0	0	0	4.9910	5.0000	0.0909	0.0924
$\gamma_0 = -1.55$	-1.5586	-1.5545	0.1094	0.1095	-1.5762	-1.5708	0.1602	0.1583
$\gamma_1 = -0.75$	-0.7534	-0.7516	0.0354	0.0355	-0.766	-0.7563	0.0949	0.0952
$\sigma_z = 4$	0	0	0	0	4.0039	3.9999	0.0862	0.0854
$\mu_z = -8$	0	0	0	0	-7.9888	-7.999	0.1371	0.1367
N = 3000								
$\gamma_0 = -2.55$	-2.5680	-2.5610	0.1770	0.1774	-2.6330	-2.6041	0.3094	0.3028
$\gamma_1 = 1.23$	1.2376	1.2342	0.0667	0.0672	1.2750	1.2490	0.1964	0.1972
$\sigma_z = 2.1$	0	0	0	0	2.1054	2.1006	0.068	0.0683
$\mu_z = 5$	0	0	0	0	4.9826	4.9972	0.1208	0.1232
$\gamma_0 = -1.55$	-1.5646	-1.5593	0.1420	0.1416	-1.6016	-1.5887	0.2127	0.2107
$\gamma_1 = -0.75$	-0.756	-0.7532	0.0459	0.046	-0.7824	-0.7629	0.1279	0.1315
$\sigma_z = 4$	0	0	0	0	4.0068	4.0005	0.1119	0.1110
$\mu_z = -8$	0	0	0	0	-7.9817	-7.9984	0.1794	0.1802

Note: We estimate by maximum likelihood method the parameters for the truncated Probit and the conventional Probit, and compute the standard deviation in every random sample consisting of N observations. Then, we calculate for these estimates the mean, median and standard deviation over all data sets. The Monte Carlo standard deviations of the simulation are the average standard deviations over all data sets (computed as a function of the maximum likelihood Hessian). Std(b) stands for bootstrapping standard deviation.

TABLE V
Monte Carlo Simulation.

Parameter	Probit				Truncated Probit			
	Mean	Median	Std.	Std.(b)	Mean	Median	Std.	Std. (b)
N = 2000								
$\gamma_0 = -2.55$	-2.5814	-2.5699	0.2180	0.2178	-2.6818	-2.6325	0.3959	0.3843
$\gamma_1 = 1.23$	1.2430	1.2383	0.0822	0.0821	1.3032	1.2604	0.2517	0.2531
$\sigma_z = 2.1$	0	0	0	0	2.1076	2.1004	0.0841	0.084
$\mu_z = 5$	0	0	0	0	4.9772	5.0008	0.1513	0.1543
$\gamma_0 = -1.55$	-1.5753	-1.5643	0.1750	0.1733	-1.6320	-1.6147	0.2679	0.2566
$\gamma_1 = -0.75$	-0.7600	-0.7554	0.0567	0.0567	-0.7991	-0.7723	0.1631	0.1667
$\sigma_z = 4$	0	0	0	0	4.0116	4.0033	0.1386	0.1381
$\mu_z = -8$	0	0	0	0	-7.9686	-7.9929	0.2248	0.2253

Note: We estimate by maximum likelihood method the parameters for the truncated Probit and the conventional Probit, and compute the standard deviation in every random sample consisting of N observations. Then, we calculate for these estimates the mean, median and standard deviation over all data sets. The Monte Carlo standard deviations of the simulation are the average standard deviations over all data sets (computed as a function of the maximum likelihood Hessian). Std(b) stands for bootstrapping standard deviation.

Based on table (V) for a small sample size (below 3000 observations) both the intercept estimator $\hat{\gamma}_0$ and the slope estimator $\hat{\gamma}_1$ are biased upward in absolute values. However, these estimators are less biased in median than in expectation. While the incidental parameters' estimators $[\hat{\mu}_z, \hat{\sigma}_z]$ are less biased than the parameters of interests' estimators. Additionally they are not biased in median. Thus, even in the presence of small samples the parameters characterizing the non-truncated density function can be recovered from the truncated data.

6. Guaidance for empirical work

In this section we sketch the needed procedure which will enable empirical researchers to apply the aforementioned procedure for the correction of parameters estimates which are affected by endogenous truncation bias.

The correction for selectivity bias procedure in the case of truncation requires the following four steps: First, specifying both regression equations as a function of unknown parameters to be estimated:

$$Y_{1i} = \beta_0 + \boldsymbol{x_i}\boldsymbol{\beta^*} + \xi_{1i} \tag{6.1}$$

the substantive equation, where the vector $\boldsymbol{x_i}$ of size $(1 \times p)$ contains all the

substantive equation covariates, $\boldsymbol{\beta}^*$ is a $(1 \times p)$ parameters vector consisting of k coefficients and β_0 is an intercept, and,

$$Y_{2i} = \sum_{j=1}^{J} D_{ji}(\gamma_{j_0} + \boldsymbol{z_i}\boldsymbol{\gamma_j^*}) + \xi_{2i} \tag{6.2}$$

the selection equation, where each cohort $j \in \mathcal{J}$ is characterized by its own coefficients vector $\boldsymbol{\gamma_j^*} \equiv [\gamma_{j1}, ..., \gamma_{jp}]^T$ of size $(p \times 1)$, and γ_{j_0} is an intercept for the j'th cohort. The vector $\boldsymbol{z_i}$ of size $(1 \times p)$ contains all the participation equation covariates. The dichotomous variable D_{ji} specifies a membership in the j'th cohort for the i'th observation and is defined as follows:

$$D_{ji} = \begin{cases} 1, & \text{if the observation } i \text{ belongs to the j'th cohort} \\ 0, & \text{otherwise} \end{cases}. \tag{6.3}$$

Second step involves the maximization of the maximum-likelihood function (6.4) with respect to participants covariates' vector of expectation $\boldsymbol{\mu_z^j}$ and covariance matrix $\boldsymbol{\Sigma_{p\times p}^j}$, and the parameters of the selection equation γ_{j_0} and $\boldsymbol{\gamma_j^*}$ for $j = 1, ..., J$:

$$\begin{aligned} \ln L = &\sum_{j=1}^{J}\sum_{i=1}^{n} D_{ji} \ln \phi_p^j(\boldsymbol{z_i}) + \sum_{j=1}^{J}\sum_{i=1}^{n} D_{ji} \ln\left[1 - \Phi\left(-\gamma_{j_0} - \boldsymbol{z_i'}\boldsymbol{\gamma_j^*}\right)\right] \\ &- \sum_{j=1}^{J} n_j D_{ji} \ln\left(1 - \Phi\left(-\frac{\gamma_{j_0} + \boldsymbol{\mu_z'^j}\boldsymbol{\gamma_j^*}}{\sigma_{jw}}\right)\right), \end{aligned} \tag{6.4}$$

where $\phi_p^j(\boldsymbol{z_i}) \equiv \phi_p(\boldsymbol{z_i}, \boldsymbol{\mu^j}, \boldsymbol{\Sigma_{p\times p}^j})$.

Using a numeric global maximum search method (such as the Trust Region algorithm[32]) one can obtain the parameters vector which maximizes the log-likelihood function. For faster optimization of the maximum likelihood, it is recommended to maximize (6.4) using the following first order conditions:

$$\frac{\partial \log L}{\partial \boldsymbol{\mu_z^j}} = \boldsymbol{(\Sigma^j)^{-1}} \sum_{i=1}^{n} D_{ji}\left(\boldsymbol{z_i} - \boldsymbol{\mu_z^j}\right) - n_j \lambda_{j\delta} \frac{1}{\sigma_{jw}} \boldsymbol{\gamma_j^*} = 0, \;\; j = 1, ..., J \tag{6.5}$$

$$\frac{\partial \log L}{\partial \gamma_{j_0}} = \sum_{i=1}^{n} D_{ji}\lambda_{ji} - n_j \lambda_{j\delta} \frac{1}{\sigma_{jw}} = 0, \;\; j = 1, ..., J \tag{6.6}$$

[32]For faster computations, the Trust Region Optimization algorithm (Geyer, 2014) is preferable for global optimization since it is based on a simplification (a quadratic approximation) of the objective function to be maximized. This algorithm is available in MATLAB and in R software (using the 'trust' package).

$$\frac{\partial \log L}{\partial \boldsymbol{\Sigma^j}} = -\frac{1}{2}\left(n_j - \boldsymbol{(\Sigma^j)^{-1}}\sum_{i=1}^{n} D_{ji}\left(\boldsymbol{z_i} - \boldsymbol{\mu_z^j}\right)\left(\boldsymbol{z_i} - \boldsymbol{\mu_z^j}\right)'\right)\boldsymbol{(\Sigma^j)^{-1}} + \frac{1}{2} n_j \lambda_{j\delta} \frac{\gamma_{0j} + \boldsymbol{\mu_z'^j \gamma_j^*}}{\sigma_{jw}^3} \boldsymbol{\gamma_j^* \gamma_j'^*} = 0, \quad j = 1, ..., J \tag{6.7}$$

$$\frac{\partial \log L}{\partial \boldsymbol{\gamma_j^*}} = \sum_{i=1}^{n} D_{ji}\lambda_{ji}\boldsymbol{z_i} + n_j \lambda_{j\delta} \frac{(\gamma_{j_0} + \boldsymbol{\mu_z'^j \gamma_j^*})\boldsymbol{\Sigma^j \gamma_j^*} - \boldsymbol{\mu_z^j}\sigma_{jw}^2}{\sigma_{jw}^3} = 0, \quad j = 1, ..., J \tag{6.8}$$

where $\sigma_{jw} \equiv \sqrt{1 + \boldsymbol{\gamma_j'^* \Sigma^j \gamma_j^*}}$ and $\lambda_{ji} \equiv \frac{\phi\left(-\gamma_{j_0} - \boldsymbol{z_i'\gamma_j^*}\right)}{1-\Phi\left(-\gamma_{j_0} - \boldsymbol{z_i'\gamma_j^*}\right)}$

Once the log-likelihood function (6.4) is maximized, the estimators γ_0, $\hat{\boldsymbol{\gamma}}^*$, $\hat{\boldsymbol{\mu}}_{\boldsymbol{z}}$ and $\hat{\boldsymbol{\Sigma}}_{p\times p}$ are obtained.[33] In the third step, based on the vector of parameters' estimates $\hat{\boldsymbol{\gamma}}$[34] from previous step, the inverse Mills ratio is calculated for each observation, using the following equation: $\hat{\lambda}_i = \frac{\phi\left(-\sum_{j=1}^{J} \boldsymbol{D_{ji} Z_i'\hat{\gamma}_j}\right)}{1-\Phi\left(-\sum_{j=1}^{J} \boldsymbol{D_{ji} Z_i'\hat{\gamma}_j}\right)}$, where ϕ and Φ are the standardized univariate normal density and cumulative density functions respectively.

The final step is estimating the equation of interest $Y_{1i} = \beta_0 + \boldsymbol{x_i\beta^*} + \sigma\hat{\lambda}_i + \tilde{\xi}_{1i}$ using $\hat{\lambda}_i$ calculated in previous step as an additional covariate, where the vector $\boldsymbol{\beta^*}$ and the scalars σ and β_0 are the parameters to be estimated. The random disturbance is $\tilde{\xi}_{1i}$.

Alternatively, one can run a partially linear regression as introduced by Robinson (1988): $Y_{1i} = \beta_0 + \boldsymbol{x_i\beta^*} + \mathcal{M}\ (m_i)$, where $m_i = -\sum_{\boldsymbol{j=1}}^{\boldsymbol{J}} \boldsymbol{D_{ji} Z_i'\hat{\gamma}_j}$.[35]

By employing this procedure, one can correct for the selectivity bias propagated by truncated data.

7. Summary

Selectivity bias propagated by truncated data, can be corrected. The identification of the participation equation parameters, involves the maximization of the multivariate truncated density function with respect to unknown parameters which generated the incomplete data we observe. Some of those parameters are incidental parameters used for sake of identification. The ability to identify the non-participants' characteristics can be of value for decision makers to affect the selection rule. This can be done by utilizing the counterfactual participants'

[33] Initial values for the expectations vector $\boldsymbol{\mu_z}$ and the covariance matrix $\boldsymbol{\Sigma}_{p\times p}$ are the sample averages of all the covariates and the sample covariance matrix respectively. Initial values for the vector $\boldsymbol{\gamma}$ are chosen arbitrarily.

[34] The log-likelihood function in equation (4.46) is constant when $\boldsymbol{V'\gamma} \to \infty$ for every vector $\boldsymbol{V}$, since it implies that $\Phi(-\boldsymbol{V'\gamma}) \to 0$, and consequently $\ln L \to \sum_{i=1}^{n} \ln \phi_p\left(\boldsymbol{z_i}, \boldsymbol{\mu_z}, \boldsymbol{\Sigma}_{p\times p}\right)$.

[35] In case that the disturbances are not i.i.d estimate this equation using the option robust standard errors in Stata.

distribution function, representing the participants' distribution function in an alternative state of nature. Consequently, one can estimate what would be the the percentage of participants and the participants' composition (their characteristics' distribution) in the general population as a result of a change in one (or more) of the selection rule's coefficients.

The methodology introduced in this paper can recover the unobserved non-truncated density function's unknown parameters based only on the observable truncated density function. Robustness verification results indicate that the truncated Probit estimators are very accurate, consistent, their convergence rate is above the conventional $\sqrt{n}$ consistency and are identical to the conventional full sample (non-truncated) Probit estimators.

A rather more general specification of the model to accommodate for dichotomous variables enables the differentiation among various cohorts which may exist in the data, by their characteristics' distribution function and their respective selection rule. This can contribute for policy effects on both the probability to be selected and the outcome variable in the substantive equation. The cases in which participation can be perfectly determined by cohorts imply that the selection rule is completely separated by the cohorts. In conventional binary response models such as Probit and Logit, the maximum likelihood estimate for the set of completely separate covariates does not exist. However, data reconstruction can overcome this problem, by recovering the omitted control groups. For each cohort the omitted control group is its complement, such that the particpants are the complement of non-participants and vice versa.

8. Appendix

THEOREM 8.1 *Let* $\boldsymbol{Y} = [y_1, ..., y_p]^T$ *denote a multivariate normal distributed random vector* $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *where* $\boldsymbol{\mu}$ *is the expectations vector of size* $(p \times 1)$ *and* $\boldsymbol{\Sigma}$ *is the covariance matrix of size* $(p \times p)$*. Then, the weighted sum* $T = \boldsymbol{\gamma'Y}$ *is normally distributed, such that* $T \sim \mathcal{N}(\boldsymbol{\mu'\gamma}, \boldsymbol{\gamma'\Sigma\gamma})$ *where* $\boldsymbol{\gamma}$ *is a coefficients vector of size* $(p \times 1)$*.*

PROOF: For a p dimensional multivariate normal distribution $Y \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = [\mu_1, ..., \mu_p]^T$ and $\sigma_{jk} = \text{cov}(Y_j, Y_k)$ $j, k = 1, .., p$, the characteristic function is given by: $\varphi_Y(\boldsymbol{t}) = E[\exp(i\boldsymbol{t}^T Y)] = \exp(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma t})$ and consequently, $\varphi_Y(\boldsymbol{t}) = \exp(i\sum_{j=1}^{p} t_j\mu_j - \frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{p} t_j t_k \sigma_{jk})$.

Now, suppose we define a new random variable $T = \boldsymbol{\gamma}'Y = \sum_{j=1}^{p}\gamma_j Y_j$. The characteristic function for T is the same as for Y: $\varphi_T(\boldsymbol{t}) = \exp(itT) = \exp(it\boldsymbol{\gamma}'Y)$, implying that $\varphi_T(\boldsymbol{t}) = \varphi_Y(\boldsymbol{t\gamma}) = \exp(it\sum_{j=1}^{p}\gamma_j\mu_j - \frac{1}{2}t^2\sum_{j=1}^{p}\sum_{k=1}^{p}\gamma_j\gamma_k\sigma_{jk})$.

We can compare between the characteristic function of T and the characteristic function of a univariate normal variable $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ which is: $\varphi_{X\boldsymbol{t}} = \exp(it\mu_X - \frac{1}{2}t^2\sigma_X^2)$, by defining $\mu_T = \sum_{j=1}^{p}\gamma_j\mu_j$ and $\sigma_T^2 = \sum_{j=1}^{p}\sum_{k=1}^{p}\gamma_j\gamma_k\sigma_{jk}$.

Hence, because the characteristic function of T is equivalent to the characteristic function of X, the distributions must be also equal. Thus, T is normally distributed. ■

Let $[\boldsymbol{z_{p\times 1}}, \xi_2]^T \sim \mathcal{N}_{p+1}\left(\begin{bmatrix}\boldsymbol{\mu_{z_{p\times 1}}}\\ 0_{1\times 1}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma_{p\times p}} & \boldsymbol{0_{p\times 1}}\\ \boldsymbol{0_{1\times p}} & 1_{1\times 1}\end{bmatrix}\right)$ and $\boldsymbol{\gamma} \equiv [\gamma_0, \boldsymbol{\gamma^*}]$ is a vector composed of a scalar γ_0 and a vector $\boldsymbol{\gamma^*}$ of size $(p \times 1)$, such that $\boldsymbol{\gamma^*} \equiv [\gamma_1, ..., \gamma_p]^T$.

Based on theorem (8.1), using the weighted sum of a random variables vector distributed multivariate normal, it is possible to simplify the high-dimensional integral in the following equation:

$$P_{\boldsymbol{\theta}}(\mathcal{S}_i = 1) = P(\boldsymbol{Z_i'\gamma} + \xi_{2i} > 0) = \int_{v_p} ... \int_{v_1} f_{\boldsymbol{\eta}}(v_1, ...v_p)\, P\left(\boldsymbol{Z_i'\gamma} + \xi_{2i} > 0 | z_{1i} = v_1, ..., z_{pi} = v_p\right) dv_1 ... dv_p. \tag{8.1}$$

to obtain the following close-form formula:

$$P_{\boldsymbol{\theta}}(\mathcal{S}_i = 1) = P(\gamma_0 + \boldsymbol{z_i'\gamma^*} + \xi_{2i} > 0) = 1 - \Phi\left(-\frac{\gamma_0 + \boldsymbol{\mu_z}'\boldsymbol{\gamma^*}}{\sqrt{1 + \boldsymbol{\gamma'^*\Sigma\gamma^*}}}\right). \tag{8.2}$$

References

Ahn, H. and Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *journal of Econometrics*, 58(1-2):3–29.

Albert, A. and Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica: Journal of the Econometric Society*, pages 997–1016.

Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.

Arabmazar, A. and Schmidt, P. (1982). An investigation of the robustness of the tobit estimator to non-normality. *Econometrica: Journal of the Econometric Society*, pages 1055–1063.

Ashenfelter, O., Harmon, C., and Oosterbeek, H. (1999). A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour economics*, 6(4):453–470.

Bland, J. M. and Altman, D. G. (1998). Survival probabilities (the kaplan-meier method). *Bmj*, 317(7172):1572–1580.

Bloom, D. E. and Killingsworth, M. R. (1985). Correcting for truncation bias caused by a latent truncation variable. *Journal of Econometrics*, 27(1):131–135.

Chanda, K. (1954). A note on the consistency and maxima of the roots of likelihood equations. *Biometrika*, 41(1/2):56–61.

Cosslett, S. R. (1981a). Efficient estimation of discrete-choice models. *Structural analysis of discrete data with econometric applications*, pages 51–111.

Cosslett, S. R. (1981b). Maximum likelihood estimator for choice-based samples. *Econometrica: Journal of the Econometric Society*, pages 1289–1316.

Cox, P. R. (1972). *Life Tables*. Wiley Online Library.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, pages 1001–1044.

Geyer, C. J. (2014). Trust: Trust region optimization. *R package version 0.1-6, URL http://CRAN. R-project. org/package= trust*.

Hausman, J. A. and Wise, D. A. (1977). Social experimentation, truncated distributions, and efficient estimation. *Econometrica: Journal of the Econometric Society*, pages 919–938.

Heckman, J. J. (1974). Effects of child-care programs on women's work effort. In *Marriage, Family, Human Capital, and Fertility*, pages 136–169. Journal of Political Economy 82 (2), Part II.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 475–492. NBER.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.

Heckman, J. J. and Singer, B. (1984). Econometric duration analysis. *Journal of Econometrics*, 24(1):63–132.

Honore, B. E., Kyriazidou, E., and Udry, C. (1997). Estimation of type 3 tobit models using symmetric trimming and pairwise comparisons. *Journal of econometrics*, 76(1):107–128.

Honoré, B. E. and Powell, J. L. (1994). Pairwise difference estimators of censored and truncated regression models. *Journal of Econometrics*, 64(1-2):241–278.

Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., and Smola, A. J. (2006). Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608.

Hurd, M. (1979). Estimation in truncated samples when there is heteroscedasticity. *Journal of Econometrics*, 11(2):247–258.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kiefer, N. M. (1988). Economic duration data and hazard functions. *Journal of economic literature*, 26(2):646–679.

Killingsworth, M. R., Heckman, J. J., et al. (1986). Female labor supply: A survey. *Handbook of labor economics*, 1(1):103–204.

Kohn, W. and Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133.

Lee, L.-F. (1982). Some approaches to the correction of selectivity bias. *The Review of Economic Studies*, 49(3):355–372.

Mäkeläinen, T., Schmidt, K., and Styan, G. P. (1981). On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples. *The Annals of Statistics*, pages 758–767.

Murray, M. P. (2005). *Econometrics: A modern introduction*. Pearson Higher Education.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal*, 12(s1):S217–S229.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, pages 1067–1101.

Powell, J. L. (1987). *Semiparametric estimation of bivariate latent variable models*. University of Wisconsin–Madison, Social Systems Research Institute.

Powell, J. L. (1994). Estimation of semiparametric models. *Handbook of econometrics*, 4:2443–2521.

Rai, K. and Van Ryzin, J. (1982). A note on a multivariate version of rolle's theorem and uniqueness of maximum likelihood roots. *Communications in Statistics-Theory and Methods*, 11(13):1505–1510.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.