SEMIPARAMETRIC MAXIMUM LIKELIHOOD SIEVE ESTIMATOR FOR CORRECTION OF ENDOGENOUS TRUNCATION BIAS

November 18, 2018

Nir Billfeld[*] Moshe Kim[†]

Semiparametric correction for a sample selection bias in the presence of endogenous truncation is known to be much more difficult in the case of a binary selection variable than in the case of a continuous selection variable. This paper proposes a simple bandwidth-free semiparametric methodology to correct for a self-selection bias in a truncated sample, without any prior knowledge of the marginal density functions of the selection model's random disturbances. Each of the unknown marginal density functions is estimated using Sieve estimator, utilizing Hermite polynomials as basis functions. The aforementioned procedure is appropriate for both binary and continuous selection variables cases under the covariate shift assumption. We consider a *double hurdle* model, which is a combination of two selection rules. The first is propagated by a truncation in the dependent variable of the substantive equation. The second is propagated by endogenous self-selection. The suggested correction procedure produces estimates that are of high accuracy and consistent based on Monte Carlo simulations. The random disturbances are not restricted to being symmetric and their marginal distribution functions are unknown. Thus, in the data generation process we verify the applicability of our procedure to cases in which the disturbances are neither jointly nor marginally normally distributed. These disturbances are constructed as realizations of non-symmetric distribution functions.

Keywords: Selectivity bias correction, Hermite Polynomials, Covariate shift.

[*] University of Haifa nbillfeld@staff.haifa.ac.il

[†] University of Haifa kim@econ.haifa.ac.il

1. Introduction

We introduce a semiparametric correction procedure for self-selection truncation bias in binary selection-variable instances. The estimation and identification of semiparametric truncated sample selection models with a binary selection variable are known to be difficult, due to the absence of observed variation in the aforementioned binary variable (Powell, 1994). Thus, various estimation procedures utilize a continuous selection variable to alleviate this difficulty, by imposing different restrictions on the disturbances (Honoré et al., 1997; Powell, 1994), using different kernel estimators (Lee, 1993), as well as including augmented data (Khan and Lewbel, 2007). Alternatively, in the binary selection variable models, a kernel estimator has been employed to estimate the bias term in the substantive equation (Ichimura, 1993; Ichimura and Lee, 1991). However, the resulting estimates can still be biased since the kernel estimator's accuracy depends on selecting the optimal bandwidth, which is hard to find in the semiparametric case (Lewbel and Schennach, 2007). A recent contribution (Ai et al., 2018) allows for the correction of endogenous selectivity bias in the context of 'missing data not at random' using GMM. A similar approach has been introduced by Breunig et al. (2018). These approaches are applicable in censored sample selection models, as they utilize a binary selection variable, which is required to be observed unlike the situation existing under truncation, where the binary selection variable is not observed.

To overcome the aforementioned difficulties in the binary selection variable case, we suggest an alternative approach to the kernel estimator, which can be employed without utilizing augmented data and is bandwidth-free. The generality of our model enables the correction of endogenous selectivity bias in truncated samples, and can be implemented in both binary as well as continuous selection instances.

The underlying model consists of substantive and selection equations, in which the latent (population) dependent variables y^*_{1i} and y^*_{2i} are determined as follows:

(1.1) $$y^*_{1i} = \boldsymbol{x}^{*\prime}_{\boldsymbol{i}}\boldsymbol{\beta} + \xi_{1i} \qquad \text{the substantive equation,}$$

and

(1.2) $$y^*_{2i} = \boldsymbol{z}^{*\prime}_{\boldsymbol{i}}\boldsymbol{\gamma} + \xi_{2i} \qquad \text{the selection equation,}$$

for $i = 1, ..., N$ observations, $(')$ is the transpose operator and ξ_{1i}, ξ_{2i} are jointly distributed random disturbances, which are independent of the latent (population) covariates vectors $\boldsymbol{x}^*_{\boldsymbol{i}} \in \mathbb{R}^{q_1}$ and $\boldsymbol{z}^*_{\boldsymbol{i}} \in \mathbb{R}^{q_2}$, and where $\boldsymbol{\beta} \in \mathbb{R}^{q_1}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{q_2}$ are the coefficient vectors to be estimated. The aforementioned joint distribution function is denoted by F_{ξ_1,ξ_2} and is unknown.

Since the selection variable can be either binary or continuous, two different models are considered. In the continuous selection variable model, one observes the sequence $\left\{y_{1i}, y_{2i}, \boldsymbol{x_i'}, \boldsymbol{z_i'}\right\}$, in which any vector w_i is constructed from a latent variable w_i^*, and defined as:

$$w_i = \begin{cases} w_i^* & \text{if} \quad y_{1i}^* \geq c_1, y_{2i}^* \geq c_2 \\ \text{Unobserved} & \quad \text{Otherwise} \end{cases} \tag{1.3}$$

with c_1 and c_2 specifying the fixed truncation points, $w_i \in \{y_{1i}, y_{2i}, \boldsymbol{x_i}, \boldsymbol{z_i}\}$, and where y_{2i} is the continuous selection variable.

In the binary selection variable model, only the binary outcome of the latent variable, y_{2i}^* equals unity, is observed. Additionally, the sequence $\left\{y_{1i}, S_i, \boldsymbol{x_i'}, \boldsymbol{z_i'}\right\}$ is observed, where the selection variable S_i is defined as:

$$S_i = \begin{cases} 1 & \text{if} \quad y_{1i}^* \geq c_1, y_{2i}^* \geq c_2 \\ \text{Unobserved} & \quad \text{Otherwise} \end{cases} \tag{1.4}$$

The cross equations' covariance is defined as follows:

$$\mathbb{COV}\left[\xi_{1i}, \xi_{2m}\right] = \begin{cases} \sigma_{\xi_1,\xi_2} & \text{if} \quad i = m \\ 0 & \text{if} \quad i \neq m, \qquad\qquad i, m = 1, ..., N, \end{cases} \tag{1.5}$$

with $\sigma_{\xi_1,\xi_2} \neq 0$. The j'th random disturbance ξ_{ji} is of zero mean and finite variance $\sigma_{\xi_j}^2 < \infty$ for all $i = 1, ..., N$ and $j \in \{1, 2\}$. The random disturbances are not restricted to being symmetric and their marginal distribution functions are unknown. Thus, the proposed model is semiparametric.

We introduce a semiparametric Sieve (a mixture of basis functions[1]) estimator to correct for self-selection truncation bias. The estimation procedure is an extension of the methodology introduced by Schwiebert (2013, 2016) for censored sample selection models.[2] The estimation procedure is simplified by decomposing the unknown disturbances' joint density function into two parts using a Copula function:[3] a parametric part and a nonparametric part. The parametric part determines the dependence structure, up to an unknown parameter, which captures the degree of dependence.[4] The nonparametric part consists of the disturbances' marginal density and distribution functions.

[1] Instead of using a mixture of distribution functions, we utilize Hermite polynomials as basis functions.

[2] Based on *Sieves estimators* (Chen, 2007).

[3] A Copula function is a device to decompose any continuous joint density function into two components: the dependence structure and the marginal density functions.

[4] The depedence structure is specified by the Copula function, which is a bivariate uniform density function. Estimating this part parametrically implies one does not need to estimate a bivariate density function.

This use of a Copula function reduces the dimensionality of the parameters set characterizing the joint density function. Instead of approximating a bivariate density function, which involves the estimation of a large set of parameters, one approximates a sequence of univariate densities, utilizing a single parameter to characterize the degree of dependence. Each marginal density is approximated by a different Sieve estimator. Hermite polynomials are chosen as basis functions on the support, $(-\infty, \infty)$ as the disturbances are commonly distributed on this unbounded support. We run Monte Carlo simulations in order to examine our estimator's performance in the presence of a truncated *double hurdle* model, which is a combination of two selection rules. The first is propagated by a truncation in the dependent variable of the substantive equation. The second is propagated by endogenous self-selection. Further, for sake of generality of the offered estimator, we subject it to various distributions in which the disturbances are neither jointly nor marginally normally distributed. These disturbances are constructed as realizations of non-symmetric distribution functions.[5]

The organization of the rest of the paper is as follows: Section 2 presents the literature review and discusses the various estimation methods of semiparametric truncated sample selection models. Section 3 prepares the ground for our methodology by describing the interrelationship between the truncated data and the complete data and is essential for understanding the similarity and dissimilarity between two different model setups: the binary selection variable and the continuous selection variable. Section 4 introduces the methodology which involves a specification of a bivariate Copula density function with unknown marginals to be estimated by a semiparametric maximum likelihood (SPMLE) method. The unknown marginals are approximated by Hermite Polynomials. It also presents the model assumptions and discusses the asymptotic properties. Section 5 presents model extensions in the presence of multiple equations. Section 6 provides Monte Carlo simulations for assessing the accuracy of the suggested procedure. Section 7 provides a practical guide for our procedure implementation and section 8 concludes.

Next we present the literature review, and discuss two different classes of endogenous truncation models, in order to understand the background for our approach.

2. Literature review

A wide variety of methodologies exists in the literature to correct for the selectivity bias propagated by truncated data, and they can be classified into two main classes. The first class is truncated regression models, in which the truncation is determined only as a function of the outcome variable of interest (without self-selection). This outcome variable is the dependent variable of the substantive equation of interest, and is a function of a covariates' vector and a random disturbance. Any

[5] Unlike the practice in some other studies applying only normally distributed disturbances.

restriction imposed on this outcome variable implies a restriction on the aforementioned function, leading to a dependence between its covariates and disturbances, by construction.[6] One can correct for the selection bias by employing a semiparametric procedure (Buckley and James, 1979; Lee and Kim, 1998; Powell, 1986), or a nonparametric one (Lewbel and Linton, 2002; Tsui et al., 1988).

The second class of models is *truncated sample selection models* dealing specifically with self-selection[7] of observations into the sample. The model consists of a substantive equation of interest and a selection (participation) equation. The truncation of the data depends entirely on the selection variable,[8] instead of the outcome variable of interest. Employing an estimation procedure by using the observed data, without taking the self-selection phenomenon into consideration, might lead to biased estimates.[9]

A wide collection of correction procedures can be employed if the selection variable is continuous: for instance, a *symmetric-trimming* approach (Powell, 1994) and the *pairwise difference estimator* approach (Honoré et al., 1997; Honoré and Powell, 1994). The former assumes the disturbances' distribution functions are symmetric about zero given the covariates, and the latter relies on the assumption that the pairwise difference between any two disturbances is symmetrically distributed. Additionally, different kernel estimator procedures[10] can be employed (Lee, 1993), including the procedure introduced by Khan and Lewbel (2007), which utilizes augmented data,[11] allowing for any form of heteroskedasticity in the error distribution function. However, in many cases, the availability of augmented data is limited.

In the absence of augmented data and in the presence of a binary selection model, one can utilize a kernel function estimator for the biased term[12] in the equation of interest and estimate the aforementioned equation semiparametrically, using a numeric search procedure suggested by Ichimura and Lee (1991). This procedure is extended by Ichimura (1993) to allow for multiple selection rules, as well as relying on a kernel estimator.

One of the drawbacks of the above-described methodology is that the kernel estimates' accuracy depends on the bandwidth selection. There is an open question whether there is a way to choose a

[6] Consider a linear regression $y_i = \boldsymbol{\beta'x_i} + \xi_i$, where ξ_i is a random disturbance, $\boldsymbol{x_i}$ and $\boldsymbol{\beta}$ are a covariate and a coefficient vectors, respectively. A specific observation i is selected into the sample if its dependent variable y_i is positive or equivalently, if $\xi_i > -\boldsymbol{\beta'x_i}$, leading to a dependence between the disturbance and the covariates in the truncated data, by construction.

[7] It occurs whenever the substantive equation's covariates are correlated with unobserved factors that affect both the probability to be selected and the outcome variable in the equation of interest.

[8] The selection equation's dependent variable.

[9] The potential bias is due to a bias term which is included in the disturbance, leading to a potential correlation of the disturbance with the regressors.

[10] A kernel estimator is a local estimator, as it must be computed for each data point individually.

[11] For the censored sample selection models a methdology which replace the bandwidth estimator is introduced by Lewbel and Schennach (2007).

[12] The conditional expectation of the disturbance given participation of a specific observation.

bandwidth sequence that is optimal for the estimation of the parameters (Ichimura and Lee, 1991). Semiparametric models that involve a kernel function estimation might lead to biased estimates, due to the difficulty of the optimal bandwidth to be found: "The well-known bandwidth selection rules used in nonparametric estimation, such as cross-validation, are not generally applicable to semiparametric settings"(Lewbel and Schennach, 2007, p. 191). Thus, in practice one needs to use a bandwidth that is "slightly" smaller than the optimal bandwidth obtained using the cross-validation procedure.[13] However, this informal method for bandwidth choice may lead to a non-ignorable bias in the estimates (Lewbel and Schennach, 2007).

Another problem with the kernel approach is its computational burden that might contribute to the bias in the estimates: "The computation burden implies more causal specification of the systemic part, which might lead to a larger bias than would the causal specification of the parametric error term. Therefore it would be desirable to have a method with less computational time" (Ichimura, 1993, p. 104).

The third problem is identification, as the dimension of the nuisance parameters is infinite, and it is not clear at all, a priori, which parameter is identified. As a result, in general one needs to impose different restrictions on the substantive and selection equations' parameters (Ichimura, 1993).

A combination of the classes[14] is referred to as *double hurdle models* (Ichimura, 1993),[15] in which both types of truncations mentioned above are applied: the substantive equation's dependent variable and covariates are observed, if the aforementioned dependent variable is above (or below) a fixed-truncation point and there is a self-selection. The methodology introduced by Ichimura (1993) is appropriate also for this class of models.

The motivation for the present paper is to suggest a simpler alternative to the available semiparametric selectivity bias correction procedures, one that is purged of the aforementioned shortcomings presented in the literature. Our methodology can be employed without augmented data utilization and replaces the kernel local estimator, which is routinely employed in the truncated sample selection models, with a global estimator. The estimation procedure in the latter is *the method of Sieves*.[16] This estimation procedure is an extension of the methodology introduced by Schwiebert

[13] The cross-validation algorithm is a criterion that numerically searches for the bandwidth that minimizes the average square distance function between the kernel estimator computed, by leaving out the i'th observation (calculated over $i = 1, ..., n$) and the empirical function estimated using all the data. For a given candidate bandwidth, the entire model is estimated n times repeatedly. Each time another observation is excluded.

[14] As discussed earlier, the two classes are truncated regression models and truncated sample selection models.

[15] The methodology is introduced by Cragg (1971) for parametric censored sample selection models. For semiparametric application in censored sample selection models see Lewbel (2008).

[16] The method of Sieves is a general approach to solve infinite-dimensional optimization problems. One of its advantages is the possibility to reduce the dimensionality of the problem by substitution of the original problem with an approximated finite-dimensional problem. A Sieve is sequence of finite parmeters spaces constructed so that in the limit the function of interest (e.g., the density function) lies within it (Coppejans, 2001).

(2013, 2016) for censored sample selection models, based on the *Sieves estimators* suggested by Chen (2007). We utilize Sieve space, which is a dimension-reduction methodology. The approximated parameter space is finite instead of an infinite parameter space. All the unknown parameters are estimated simultaneously. The degree of approximation depends on the number of components in the series.

As an alternative to kernel estimators, linear Sieves (or series) are widely used in empirical economics due to its computational simplicity (Chen and Qiu, 2016). Unlike the kernel estimator, which is a local estimator, the Sieve estimator is a global estimator, in the sense that it estimates the function of interest over its domain at once. Sieve methods are easily identified by their use of approximating functions, such as splines, power or trigonometric series. Approximation is of a global nature, so one can estimate the whole function of interest in a single pass, which saves computational time. Additionally, it is much easier to impose certain structure or restrictions in Sieve estimation than in kernel estimation (for e.g., restricting the density functions space to only smooth density functions).

In the described truncated selection models, a population regression equation is assumed for each one of the substantive and the selection equations.[17] We refer to this assumption as *the covariate shift assumption* which will be discussed in detail in section 3.5 to follow.

We rely on the assumption of covariate shift to connect between the observed truncated data set and the unobserved complete data set.

In the former data set, the covariates data are realizations from the truncated distribution function (the participants' characteristics), while in the latter data set, the covariates data are realizations from the complete distribution function (the entire population's characteristics).

Two different cases are examined. In the first case, we consider a continuous dependent variable in the selection equation,[18] while in the second case we consider a binary response dependent variable. Additionally, we also allow for a fixed truncation in the substantive equation's dependent variable.

Next, before getting into technical details, we explain the relationship between the non-truncated (complete) and truncated data distribution functions.

3. Preliminaries

Formalizing the interrelationship between the truncated data and the complete data is of great importance for the understanding of the flexibility of our suggested selectivity bias correction pro-

[17] As the population consists of both excluded and included observations, the same function of the regressors is assumed to determine the dependent variable, whether or not the observation is included or excluded.

[18] In the case of a continuous selection variable, our proposed methodology is an alternative to the *symmetrically trimmed* least square estimation methodology introduced by Powell (1986) which is implemented on a single regression equation and requires a symmetrical disturbances distribution.

cedure, with respect to the assumptions regarding the complete data joint distribution function.

Next we show how the non-truncated data distribution function is related to the random disturbances distribution function.

3.1. The non-truncated data distribution function

In the non-truncated data, the dependent random variables $(\mathrm{y}_1, \mathrm{y}_2)$[19] are constructed as follows:

$$(3.1) \qquad \mathrm{y}_1 = \mathbf{x}'\boldsymbol{\beta} + \xi_1, \qquad \mathrm{y}_2 = \mathbf{z}'\boldsymbol{\gamma} + \xi_2$$

where ξ_1 and ξ_2 are the random disturbances, while $\mathbf{x} \in \mathbb{R}^{q_1}$ and $\mathbf{z} \in \mathbb{R}^{q_2}$ are random variables vectors.

The joint distribution of y_1 and y_2 given a specific realization $\boldsymbol{x} \in \mathbb{R}^{q_1}$ and $\boldsymbol{z} \in \mathbb{R}^{q_2}$ of the random variables vectors $\mathbf{x}$ and $\mathbf{z}$, respectively, is obtained as follows:

$$(3.2) \qquad F_{\mathrm{y}_1,\mathrm{y}_2}(y_1, y_2 | \mathbf{x} = \boldsymbol{x}, \mathbf{z} = \boldsymbol{z}) = P(\mathrm{y}_1 < y_1, \mathrm{y}_2 < y_2 | \mathbf{x} = \boldsymbol{x}, \mathbf{z} = \boldsymbol{z})$$

$$= P(\mathbf{x}'\boldsymbol{\beta} + \xi_1 < y_1,\ \mathbf{z}'\boldsymbol{\gamma} + \xi_2 < y_2 | \mathbf{x} = \boldsymbol{x}, \mathbf{z} = \boldsymbol{z})$$

$$= P(\xi_1 < y_1 - \mathbf{x}'\boldsymbol{\beta},\ \xi_2 < y_2 - \mathbf{z}'\boldsymbol{\gamma} | \mathbf{x} = \boldsymbol{x}, \mathbf{z} = \boldsymbol{z}) = F_{\xi_1,\xi_2}(y_1 - \boldsymbol{x}'\boldsymbol{\beta}, y_2 - \boldsymbol{z}'\boldsymbol{\gamma}),$$

where y_1 and y_2 are realizations of the random variables y_1 and y_2, respectively.

The random variables y_1 and y_2 in the second equality are replaced with the expressions in (3.1). The last equality is due to the independence assumption of the disturbances and the covariates.

As by construction, the joint distribution function of $\mathrm{y}_1, \mathrm{y}_2 | \mathbf{x} = \boldsymbol{x}, \mathbf{z} = \boldsymbol{z}$ is not affected by the joint distribution function of the random variables, $\mathbf{x}, \mathbf{z}$, which generated the covariates data, one only needs to consider the joint distribution function of the dependent variables, given the covariates.[20]

Next we show the equivalence (given the covariates) between the truncated joint density functions of $\mathrm{y}_1, \mathrm{y}_2$ and ξ_1, ξ_2, using a continuous selection variable.

3.2. The continuous selection variable model

In the continuous selection variable, the truncated distribution function of the random variable y_1 given the covariates vectors is defined as:

[19] The notations y_j, $j = 1, 2$, imply a random variable, while the notation y_j implies a specific realization of this random variable.

[20] By conditioning the distribution function of $\mathrm{y}_1, \mathrm{y}_2$ on the covariates, the covariates are treated as given constants. See a similar treatment for the covariates vector in a parametric truncated sample selection model suggested by Bloom and Killingsworth (1985) and in semiparametric censored selection models by Schwiebert (2013, 2016).

$$
\begin{aligned}
(3.3)\qquad & F_{\mathrm{y}_1,\,\mathrm{y}_2}(y_1, y_2|\mathrm{y}_1 \geq c_1, \mathrm{y}_2 \geq c_2, \mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z}) \\
& = \frac{P(c_1 \leq \mathrm{y}_1 \leq y_1, c_2 \leq \mathrm{y}_2 \leq y_2|\mathbf{x} = \boldsymbol{x}, \mathbf{z} = \boldsymbol{z}) f_{\mathbf{x},\mathbf{z}}(\boldsymbol{x}, \boldsymbol{z})}{P(\mathrm{y}_1 \geq c_1, \mathrm{y}_2 \geq c_2|\mathbf{x} = \boldsymbol{x}, \mathbf{z} = \boldsymbol{z}) f_{\mathbf{x},\mathbf{z}}(\boldsymbol{x}, \boldsymbol{z})} \\
& = \frac{P(-\mathbf{x}'\boldsymbol{\beta} \leq \xi_1 \leq y_1 - \mathbf{x}'\boldsymbol{\beta},\ \ -\mathbf{z}'\gamma \leq \xi_2 \leq y_2 - \mathbf{z}'\gamma|\mathbf{x} = \boldsymbol{x}, \mathbf{z} = \boldsymbol{z})}{P(\xi_1 \geq -\mathbf{x}'\boldsymbol{\beta}, \xi_2 \geq -\mathbf{z}'\gamma|\mathbf{x} = \boldsymbol{x}, \mathbf{z} = \boldsymbol{z})}, \\
& = \frac{P(c_1 - \boldsymbol{x}'\boldsymbol{\beta} < \xi_1 \leq y_1 - \boldsymbol{x}'\boldsymbol{\beta},\ \ c_2 - \boldsymbol{z}'\boldsymbol{\gamma} \leq \xi_2 \leq y_2 - \boldsymbol{z}'\boldsymbol{\gamma})}{P(\xi_1 \geq c_1 - \boldsymbol{x}'\boldsymbol{\beta}, \xi_2 \geq c_2 - \boldsymbol{z}'\gamma)},
\end{aligned}
$$

where $f_{\mathbf{x},\mathbf{z}}(\boldsymbol{x}, \boldsymbol{z})$ is the joint density function of the covariates data, which cancels out and thus has no effect on the conditional joint distribution (3.3).

Simplifying the numerator of the last equality in (3.3) we get:

$$
\begin{aligned}
(3.4)\qquad & P(c_1 - \boldsymbol{x}'\boldsymbol{\beta} \leq \xi_1 \leq y_1 - \boldsymbol{x}'\boldsymbol{\beta}, c_2 - \boldsymbol{z}'\boldsymbol{\gamma} \leq \xi_2 \leq y_2 - \boldsymbol{z}'\boldsymbol{\gamma}) = F_{\xi_1,\xi_2}(y_1 - \boldsymbol{x}'\boldsymbol{\beta}, y_2 - \boldsymbol{z}'\boldsymbol{\gamma}) \\
& + F_{\xi_1,\xi_2}(c_1 - \boldsymbol{x}'\boldsymbol{\beta},\ \ c_2 - \boldsymbol{z}'\boldsymbol{\gamma}) - F_{\xi_1,\xi_2}(y_1 - \boldsymbol{x}'\boldsymbol{\beta},\ \ c_2 - \boldsymbol{z}'\boldsymbol{\gamma}) - F_{\xi_1,\xi_2}(c_1 - \boldsymbol{x}'\boldsymbol{\beta},\ \ y_2 - \boldsymbol{z}'\boldsymbol{\gamma}),
\end{aligned}
$$

where F_{ξ_1,ξ_2} denotes the disturbances bivariate distribution function.

Based on (3.3) using (3.4), the truncated bivariate density function of the random variables vector $\{\mathrm{y}_1, \mathrm{y}_2\}$, given the covariates vectors is defined as:

$$
\begin{aligned}
(3.5)\qquad & f_{\mathrm{y}_1,\,\mathrm{y}_2}(y_1, y_2|\mathrm{y}_1 \geq c_1, \mathrm{y}_2 \geq c_2, \mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z}) \\
& = \frac{\partial^2}{\partial y_1 \partial y_2} F_{\mathrm{y}_1,\,\mathrm{y}_2}(y_1, y_2|\mathrm{y}_1 \geq c_1, \mathrm{y}_2 \geq c_2, \mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z}) \\
& = \frac{f_{\xi_1,\xi_2}(y_1 - \boldsymbol{x}'\boldsymbol{\beta},\ \ y_2 - \boldsymbol{z}'\boldsymbol{\gamma})}{P(\xi_1 \geq c_1 - \boldsymbol{x}'\boldsymbol{\beta}, \xi_2 \geq c_2 - \boldsymbol{z}'\boldsymbol{\gamma})},\ \ \mathrm{y}_1 \geq c_1,\ \ \mathrm{y}_2 \geq c_2,
\end{aligned}
$$

where the f_{ξ_1,ξ_2} is the joint density function of ξ_1 and ξ_2, respectively.

Next we show the equivalence, given the covariates vectors $\mathbf{x} = \boldsymbol{x}$ and $\mathbf{z} = \boldsymbol{z}$ between the truncated joint distribution function of $\{\mathrm{y}_1, \mathrm{y}_2\}$ and the truncated distribution function of $\{\xi_1, \xi_2\}$ in the binary selection variable case.

3.3. The binary selection variable model

In the binary selection variable instances, the truncated distribution function of the random variable y_1, given the covariates vectors, is defined as:

$$
(3.6)\qquad F_{\mathrm{y}_1}(y_1|\mathrm{y}_1 \geq c_1, \mathrm{y}_2 \geq c_2, \mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z})
$$

$$= \frac{P(c_1 \leq \mathrm{y}_1 < y_1, \mathrm{y}_2 \geq c_2 | \mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z}) f_{\mathbf{X},\mathbf{Z}}(\boldsymbol{x}, \boldsymbol{z})}{P(\mathrm{y}_1 \geq c_1, \mathrm{y}_2 \geq c_2 | \mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z}) f_{\mathbf{X},\mathbf{Z}}(\boldsymbol{x}, \boldsymbol{z})}$$

$$= \frac{P(-\mathbf{X}'\boldsymbol{\beta} \leq \xi_1 \leq y_1 - \mathbf{X}'\boldsymbol{\beta},\ \ \xi_2 \geq c_2 - \mathbf{Z}'\boldsymbol{\gamma} | \mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z})}{P(\xi_1 \geq -\mathbf{X}'\boldsymbol{\beta}, \xi_2 \geq -\mathbf{Z}'\boldsymbol{\gamma} | \mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z})},$$

$$= \frac{P(c_1 - \boldsymbol{x}'\boldsymbol{\beta} \leq \xi_1 \leq y_1 - \boldsymbol{x}'\boldsymbol{\beta},\ \ \xi_2 \geq c_2 - \boldsymbol{z}'\boldsymbol{\gamma})}{P(\xi_1 \geq c_1 - \boldsymbol{x}'\boldsymbol{\beta}, \xi_2 \geq c_2 - \boldsymbol{z}'\boldsymbol{\gamma})}$$

where the random variables y_1 and y_2 in the third equality are replaced with the expressions $\mathbf{X}'\boldsymbol{\beta}+\xi_1$ and $\mathbf{Z}'\boldsymbol{\gamma} + \xi_2$, respectively.

Next, we simplify the last equality's numerator in (3.6):

$$\text{(3.7)} \qquad P(c_1 - \boldsymbol{x}'\boldsymbol{\beta} \leq \xi_1 \leq y_1 - \boldsymbol{x}'\boldsymbol{\beta}, \xi_2 \geq c_2 - \boldsymbol{z}'\boldsymbol{\gamma}) = F_{\xi_1}(y_1 - \boldsymbol{x}'\boldsymbol{\beta})$$

$$+ F_{\xi_1,\xi_2}(c_1 - \boldsymbol{x}'\boldsymbol{\beta},\ \ c_2 - \boldsymbol{z}'\boldsymbol{\gamma}) - F_{\xi_1,\xi_2}(y_1 - \boldsymbol{x}'\boldsymbol{\beta},\ \ c_2 - \boldsymbol{z}'\boldsymbol{\gamma}) - F_{\xi_1}(c_1 - \boldsymbol{x}'\boldsymbol{\beta}),$$

Based on (3.6) using (3.7), the truncated density function of the random variable y_1, given the covariates vectors is defined as:

$$\text{(3.8)} \qquad f_{\mathrm{y}_1}(y_1 | \mathrm{y}_1 \geq c_1, \mathrm{y}_2 \geq c_2, \mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z})$$

$$= \frac{\partial}{\partial y_1} F_{\mathrm{y}_1}(y_1 | \mathrm{y}_1 \geq c_1, \mathrm{y}_2 \geq c_2, \mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z})$$

$$= \frac{f_{\xi_1}(y_1 - \boldsymbol{x}'\boldsymbol{\beta}) - \frac{\partial}{\partial y_1} F_{\xi_1,\xi_2}(y_1 - \boldsymbol{x}'\boldsymbol{\beta},\ \ c_2 - \boldsymbol{z}'\boldsymbol{\gamma})}{P(\xi_1 \geq c_1 - \boldsymbol{x}'\boldsymbol{\beta}, \xi_2 \geq c_2 - \boldsymbol{z}'\boldsymbol{\gamma})},\ \ \mathrm{y}_1 \geq c_1,$$

where the f_{ξ_1} and f_{ξ_2} denote the density function of ξ_1 and ξ_2, respectively.

Next, we simplify the last equality's denominator in (3.8):

$$\text{(3.9)} \qquad P(\xi_1 \geq c_1 - \boldsymbol{x}'\boldsymbol{\beta}, \xi_2 \geq c_2 - \boldsymbol{z}'\boldsymbol{\gamma}) = 1 + F_{\xi_1,\xi_2}(c_1 - \boldsymbol{x}'\boldsymbol{\beta},\ \ c_2 - \boldsymbol{z}'\boldsymbol{\gamma})$$

$$- F_{\xi_1}(c_1 - \boldsymbol{x}'\boldsymbol{\beta}) - F_{\xi_2}(c_2 - \boldsymbol{z}'\boldsymbol{\gamma}),$$

The aforementioned disturbances' joint distribution function, F_{ξ_1,ξ_2}, is decomposed into two univariate distribution functions, each approximated using a mixture of basis functions (Sieve estimator).

The decomposition of a joint distribution function into a function that depends on the marginal distribution functions is intended to overcome the curse of dimensionality. The rationale is that one only needs to approximate the marginal distribution functions to recover the disturbances' joint distribution.

Next, we explain the decomposition method using a Copula function, in order to express the disturbances' joint distribution function, in terms of marginal distribution functions and, by doing so, reducing the dimensionality of this joint distribution function.

3.4. A Copula function

Let $\mathrm{V}_1, ..., \mathrm{V}_\mathrm{p}$ be a random variables vector of size $p \times 1$, distributed according to a p dimensional continuous distribution function $F_{\mathrm{V}_1, \mathrm{V}_2, ..., \mathrm{V}_\mathrm{p}}$. According to Sklar's theorem (Sklar, 1959), if F is a p-dimensional distribution function with continuous margins $F_{\mathrm{v}_1}, ..., F_{\mathrm{v}_\mathrm{p}}$, then there exists a copula $\mathcal{C}$ with uniform marginals such that:

$$F_{\mathrm{V}_1, ..., \mathrm{V}_\mathrm{p}}(v_1, ..., v_p) = \mathcal{C}(F_{\mathrm{V}_1}(v_1), ..., F_{\mathrm{V}_\mathrm{p}}(v_p); \tau), \quad 0 \leq F_j(v_j) \leq 1, \quad j = 1, ..., p, \tag{3.10}$$

where $\mathcal{C}(u_1, ..., u_p; \tau)$ is the Copula function, u_j is a specific value that the j'th uniform random variable can take, and τ is a parameter (or alternatively, a set of parameters) that captures the degree of dependence.

Our analysis is focused on two-dimensional Copula functions, because we are interested in modeling two random disturbances. We denote the joint distribution function F_{ξ_1, ξ_2} in terms of a Copula function $\mathcal{C}(u_1, u_2; \tau)$ on the support $[0, 1] \times [0, 1]$ and obtain:

$$F_{\xi_1, \xi_2}(e_1, e_2) = \mathcal{C}\left(F_{\xi_1}(e_1), F_{\xi_2}(e_2); \tau\right), \tag{3.11}$$

where $u_1 = F_{\xi_1}(e_1)$ and $u_2 = F_{\xi_2}(e_2)$.

By taking the derivative of (3.11) with respect to e_1 and e_2, one obtains:

$$f_{\xi_1, \xi_2}(e_1, e_2) = f_{\xi_1}(e_1) f_{\xi_2}(e_2) \boldsymbol{c}\left(F_{\xi_1}(e_1), F_{\xi_2}(e_2); \tau\right) \tag{3.12}$$

where $\boldsymbol{c}(u_1, u_2; \tau) \equiv \frac{\partial^2 \mathcal{C}}{\partial u_1 \partial u_2}\Big|_{\left(u_1 = F_{\xi_1}(e_1), u_2 = F_{\xi_2}(e_2)\right)}$.

These results will be used to construct the bivariate truncated distribution function, to follow, in sections 4.1.1 and 4.2.1.

Next, we introduce the key assumption referred to as *the covariate shift assumption*, which is about the interrelationship between the truncated and the complete data.

3.5. Covariate shift

The key assumption regarding the relationship between the full (unobserved) distribution and the truncated (observed) sample is referred to as the *covariate shift*[21] and requires the truncated and non-truncated data to share the same ***conditional*** distribution functions of the random variables $(\mathrm{y}_1, \mathrm{y}_2)$ given the covariates.[22] In other words, a participant and a non-participant share the same ***decision function*** (whether to participate), but are differentiated by their covariates values (characteristics), due to differences in the realizations of the random variables vectors $\mathbf{x}$ and $\mathbf{z}$. This assumption is common to the various sample selection models discussed in the introduction, by postulating a population regression that must be valid for every observation, unrelated to its inclusion (or exclusion) in the observed truncated data. It implies that if one has access to the entire population and can randomly choose one participant and one non-participant, on average, these two individuals are differentiated by the values of their characteristics only.[23]

Next, we present our proposed methodology to correct for selectivity bias in a truncated selection model, consisting of an endogenous self-selection equation and a substantive equation.

4. Methodology

In this section we propose an estimation procedure for a truncated selection model, consisting of a substantive equation and a selection equation.

The estimation is based on a semiparametric likelihood (SPMLE) method using the truncated joint density function of the aforementioned equations' disturbances. The model is semiparametric, as the joint density is decomposed into a product of two components. The first component is a Copula function that describes, parametrically the dependence structure between the disturbances. The second component consists of the unknown marginal density functions (of the random disturbances), which are approximated nonparametrically, using Hermite-polynomials (see section 4.4 to follow), by the method of Sieves to be described shortly (section 4.5).

Next, we discuss density estimation using a Copula function, to be utilized for a semiparametric estimation of the disturbances' bivariate joint density function presented in sections (3.2) and (3.3) with unknown marginal density functions.

[21] The concept of *covariate shift* is borrowed from the field of computer sciences (Gretton et al., 2009): if the conditional distribution function of the dependent variables, given the covariates, is the same for the biased (observed) and the unbiased (unobserved) samples, one can utilize some data randomly drawn from the complete distribution (referred to as *training data*) and correct for the selection bias in the biased data set (the test data).

[22] The random variables vectors which generated the covariats data are distributed differently for participants and non-participants, it might lead to difference in the ***unconditional*** distribution functions of $\mathrm{y_j}$, $j = 1, 2$.

[23] Unlike the augmented data being utilized in the field of computer sciences (*trianing data*), we do not rely on any auxiliary data.

4.1. The continuous selection variable model

In this section we deal with a continuous selection variable, where the j'th ($j = 1, 2$) observed dependent variable is defined as:

$$(4.1) \qquad y_{ji} = \begin{cases} y^*_{ji} & \text{if} \quad I(y^*_{1i} \geq c_1)I(y^*_{2i} \geq c_2) = 1 \\ \text{Unobserved} & \text{Otherwise} \end{cases}$$

We treat c_1 and c_2 as known fixed truncation points,[24] such that the truncated data $\{y_{1i}, y_{2i}, \boldsymbol{x_i}, \boldsymbol{z_i}\}_{i=1}^n$ consist of observations satisfying both constraints $y^*_{1i} \geq c_1$ and $y^*_{2i} \geq c_2$. The first constraint, $y^*_{1i} \geq c_1$, is a fixed truncation on the substantive equation's dependent variable, while the second constraint, $y^*_{2i} \geq c_2$ or equivalently $\boldsymbol{z_i'\gamma} + \xi_{2i} \geq c_2$, determines the self-selected observations. Of course, if $c_1 \to -\infty$, the truncation is due to the selection equation only.[25]

4.1.1. A Copula utilization in the continuous selection variable model

The bivariate truncated distribution function of the disturbances can be obtained as follows:

$$(4.2) \qquad P(\xi_1 < u_1, \xi_2 < u_2 | \xi_1 > a, \xi_2 > b)$$

$$= \frac{C(F_{\xi_1}(u_1), F_{\xi_2}(u_2); \tau) + C(F_{\xi_1}(a), F_{\xi_1}(b); \tau) - C(F_{\xi_1}(a), F_{\xi_2}(u_2); \tau) - C(F_{\xi_1}(u_1), F_{\xi_2}(b); \tau)}{1 - C(F_{\xi_1}(a), 1; \tau) - C(1, F_{\xi_2}(b); \tau) + C(F_{\xi_1}(a), F_{\xi_2}(b); \tau)}$$

Let $\boldsymbol{c}(F_{\xi_1}(u_1), F_{\xi_2}(u_2); \tau) \equiv \frac{\partial^2}{\partial F_{\xi_1} \partial F_{\xi_2}} \mathcal{C}(F_{\xi_1}(u_1), F_{\xi_2}(u_2); \tau)$. After taking the derivative of (4.2) with respect to (U_1, U_2), one obtains the truncated copula density function:

$$(4.3) \qquad f_{\xi_1, \xi_2 | \xi_1 > a, \xi_2 > b}(u_1, u_2 | \xi_1 > a, \xi_2 > b) = \frac{f_{\xi_1}(u_1) f_{\xi_2}(u_2) \boldsymbol{c}(F_{\xi_1}(u_1), F_{\xi_2}(u_2); \tau)}{1 + \mathcal{C}(F_{\xi_1}(a), F_{\xi_2}(b); \tau) - F_{\xi_1}(a) - F_{\xi_2}(b)}.$$

By assumptions (2.A) and (2.B) to follow in section 4.3, the true conditional probability density function of (ξ_1, ξ_2) given participation is:

$$(4.4) \qquad f(y_{1i} - \boldsymbol{x_i'\beta}, y_{2i} - \boldsymbol{z_i'}\gamma | \xi_1 > c_1 - \boldsymbol{x_i'\beta}, \xi_2 > c_2 - \boldsymbol{z_i'}\gamma)$$

$$= \frac{f_{\xi_1}(y_{1i} - \boldsymbol{x_i'\beta}) f_{\xi_2}(y_{1i} - \boldsymbol{z_i'\gamma}) \boldsymbol{c}(F_{\xi_1}(y_{1i} - \boldsymbol{x_i'\beta}), F_{\xi_2}(y_{2i} - \boldsymbol{z_i'\gamma}); \tau)}{1 + \mathcal{C}(F_{\xi_1}(c_1 - \boldsymbol{x_i'\beta}), F_{\xi_2}(c_2 - \boldsymbol{z_i'\gamma}); \tau) - F_{\xi_1}(c_1 - \boldsymbol{x_i'\beta}) - F_{\xi_2}(c_2 - \boldsymbol{z_i'\gamma})}$$

Next, we present a model to be employed in the case of a binary selection variable.

[24] Without loss of generality, if c_2 is unknown, one can treat this constant as zero by replacing the original selection equation's intercept γ_0 with a new intercept $\tilde{\gamma}_0$, which satisfies $\tilde{\gamma}_0 = \gamma_0 - c_2$.

[25] The cutoff c_2 is intended to sort the individuals into two distinct groups: participants and non-participants. By taking expectation on both sides of the inequality $\mathrm{y}^*_{2i} \geq c_2$, it must hold that participants are expected to satisfy the constraint: $\mathbb{E}[\mathbf{z_i'}\gamma] \geq c_2$.

4.2. The binary selection variable model

In the binary selection variable model, y_{2i} is unobserved and only an indicator random variable, denoted by S_i is observed, and is defined as follows:

$$S_i = \begin{cases} 1 & \text{if} \quad y_{2i}^* \geq c_2, \ \ y_{1i}^* \geq c_1 \\ \text{Unobserved} & \quad \text{Otherwise} \end{cases} \tag{4.5}$$

where the fixed truncation points c_1 and c_2 are described in section 4.1. It follows that there is no observed variation in the selection equation's dependent variable in the observed data. Whenever S_i is observed, it equals unity. The observed dependent variable in the substantive equation is defined as:

$$y_{1i} = \begin{cases} y_{1i}^* & \text{if} \quad S_i = 1 \\ \text{Unobserved} & \quad \text{Otherwise} \end{cases} \tag{4.6}$$

Unlike the continuous selection variable case (section 4.1), one observes the data $\{y_{1i}, S_i, \boldsymbol{x_i}, \boldsymbol{z_i}\}_{i=1}^n$. In this type of model, the observed data is the set of observations satisfying $S_i = 1$.

4.2.1. A Copula utilization in the binary selection variable model

The truncated distribution function can be obtained as:

$$\begin{aligned} &P(\xi_1 < u_1 | \xi_1 > a, \xi_2 > b) \\ &\qquad = \frac{F_{\xi_1}(u_1) + \mathcal{C}(F_{\xi_1}(a), F_{\xi_2}(b); \tau) - F_{\xi_1}(a) - \mathcal{C}(F_{\xi_1}(u_1), F_{\xi_2}(b); \tau)}{1 + \mathcal{C}(F_{\xi_1}(a), F_{\xi_2}(b); \tau) - F_{\xi_1}(a) - F_{\xi_2}(b)} \end{aligned} \tag{4.7}$$

The truncated density function can be obtained by taking the first order derivative of (4.7) with respect to u_1:

$$f_{\xi_1|\xi_1>a,\xi_2>b}(u_1|\xi_1 > a, \xi_2 > b) = \frac{f_{\xi_1}(u_1)\left(1 - \frac{\partial}{\partial F_{\xi_1}(u_1)}\mathcal{C}(F_{\xi_1}(u_1), F_{\xi_2}(b); \tau)\right)}{1 + \mathcal{C}(F_{\xi_1}(a), F_{\xi_2}(b); \tau) - F_{\xi_1}(a) - F_{\xi_2}(b)}. \tag{4.8}$$

The truncated density function in (4.8) can be approximated without specifying the marginal distribution functions as follows:

$$\begin{aligned} &f_{\xi_1|\xi_1>a,\xi_2>b}(y_{1i} - \boldsymbol{x}_i'\boldsymbol{\beta} | \xi_1 > c_1 - \boldsymbol{x}_i'\boldsymbol{\beta}, \xi_2 > c_2 - \boldsymbol{z}_i'\boldsymbol{\gamma}) \\ &\qquad = \frac{f_{\xi_1}(y_{1i} - \boldsymbol{x}_i'\boldsymbol{\beta})\left(1 - \frac{\partial}{\partial F_{\xi_1}}\mathcal{C}(F_{\xi_1}(y_{1i} - \boldsymbol{x}_i'\boldsymbol{\beta}), F_{\xi_2}(c_2 - \boldsymbol{z}_i'\boldsymbol{\gamma}); \tau)\right)}{1 + \mathcal{C}(F_{\xi_1}(c_1 - \boldsymbol{x}_i'\boldsymbol{\beta}), F_{\xi_2}(c_2 - \boldsymbol{z}_i'\boldsymbol{\gamma}); \tau) - F_{\xi_1}(c_1 - \boldsymbol{x}_i'\boldsymbol{\beta}) - F_{\xi_2}(c_2 - \boldsymbol{z}_i'\boldsymbol{\gamma})} \end{aligned} \tag{4.9}$$

Before presenting the bivariate density function of the disturbances, we present the model assumptions.

4.3. Model assumptions

In order to estimate the model by maximum likelihood method and for identification of the unknown parameters, the following assumptions must be imposed:[26]

Assumption 1 *The set* $\{\boldsymbol{x_i}, \boldsymbol{z_i}, \boldsymbol{\xi_1}, \boldsymbol{\xi_2}\}_{i=1}^{n}$ *is identically and independently distributed from some underlying distribution.*

Assumption 2.A *The joint distribution function of* ξ_1 *and* ξ_2 *are given by* $F_{\xi_1,\xi_2}(e_1, e_2) = \mathcal{C}(F_{0,\xi_1}(e_1)F_{0,\xi_2}(e_2); \tau_0)$*, implying that* $\mathcal{C}(.,.;\tau_0)$ *is the true parametric copula for* ξ_{1i}, ξ_{2i} *up to unknown dependence parameter* τ_0 *and is absolutely continuous with respect to Lebesgue measure on* $[0,1]^2$*.* F_{0,ξ_1} *and* F_{0,ξ_2} *are the true distribution functions of* ξ_1 *and* ξ_2*, respectively. Each one is absolutely continuous with respect to Lebesgue measure on the real line.*

Assumption 2.B *The true marginal density function* f_{0,ξ_j} *of* F_{0,ξ_j} *for* $j = 1, 2$ *is positive on its support; and the true copula density* $\boldsymbol{c}(.,.;\tau_0)$ *of* $\mathcal{C}(.,.;\tau_0)$ *is positive on* $(0,1)^2$*.*[27]

Assumption 3 $(\boldsymbol{x}, \boldsymbol{z})$ *and* (ξ_1, ξ_2) *are independent.*

Assumption 4.A $\boldsymbol{z}$ *contains at least one variable (with a non-zero coefficient) that is not included in x. This is the conventional exclusion restriction in sample selection models.*

Assumption 4.B *The first element of* $\boldsymbol{\gamma}$ *is equal to one in absolute value for the first model setup (a binary response dependent variable).*

Assumption 5 *Covariate Shift. Let the outcome random variable in the j'th equation for* $j \in \{1, 2\}$ *be denoted by* y_j^* *and* y_j *in the complete data and truncated data, respectively. The covariates vector is a realization of the random variables vector* $\mathbf{w}$ *with a distribution function in the complete sample denoted by* $F_{\mathbf{w}}^{C}$*. In the truncated sample, its distribution function is* $F_{\mathbf{w}}^{T}$*, satisfying* $F_{\mathbf{w}}^{\mathrm{T}} \neq F_{\mathbf{w}}^{\mathrm{C}}$*. The following must hold:* $F_{\mathrm{y}_\mathrm{j}^*|\mathbf{w}=\boldsymbol{w}}(y_j|\mathbf{w} = \boldsymbol{w}) = F_{\mathrm{y}_\mathrm{j}|\mathbf{w}=\boldsymbol{w}}(y_j|\mathbf{w} = \boldsymbol{w})$ *for every* y_j *and* $\boldsymbol{w}$*.*

Similar to the censored selection model discussed in Schwiebert (2013), assumptions (1), (2.A) and (2.B) imply that the model can be estimated by maximum likelihood. Assumptions (3) and (4.A) and (4.B) are basic conditions required for identification. Assumption (5) is explained in detail in section 3.5.

In the next section, we present the Hermite polynomials, which will be employed in the estimation of the unknown marginal densities of the disturbances. Conventionally, disturbances are distributed on unbounded support. By utilizing Hermite polynomials as basis functions, we are able to express a density function as a mixture of basis functions on the unbounded support. Consequently, the resulting disturbance function's estimator is also on the unbounded support.

[26] Assumptions (1)-(4.B) are similar to assumptions 1-4 in Schwiebert (2013), which deals with a censored sample selection model. We require an additional asumption (5) as we deal with a truncated sample selection model, rather than a censored sample selection model.

[27] See section 3.4 for a definition of a Copula function.

4.4. Hermite Polynomials

The semiparametric selectivity bias correction procedure is intended to be implemented in the absence of any prior knowledge regrading the disturbances' marginal distribution functions. Each one of the unknown marginal density functions f_{ξ_1} and f_{ξ_2} in (4.4) and (4.9) can be estimated by utilizing Hermite polynomial approximation.

The Hermite polynomial is a polynomial in x with a degree k, defined as:

$$\mathrm{H}_{\mathrm{k}}(x) = (-1)^{\mathrm{k}} \frac{\phi^{\mathrm{k}}(x)}{\phi(x)}, \tag{4.10}$$

where $\phi(x)$ and $\phi^{\mathrm{k}}(x)$ are the standardized normal density function and the k'th derivative of the standardized normal density function, respectively.

The Hermite polynomials are orthogonal polynomials[28] associated with the interval $(-\infty, \infty)$ and an exponential weight function, $\mathrm{w}(x) = e^{-\frac{1}{2}x^2}$(Érdelyi et al., 1953).

Among the classical orthogonal polynomials,[29] Hermite polynomials are the most appropriate for our purposes, to approximate the unknown density functions of the disturbances, as they are commonly assumed to be distributed on the support $(-\infty, \infty)$.

A density estimator based on a Hermite polynomial is known as the inverse Fourier transform and is seen to be (for more details see (Rothenberg, 1984)):

$$f(x) \approx \phi(x) \left[1 + \frac{g_3 \mathrm{H}_3(\mathrm{x})}{6\sqrt{n}} + \frac{3g_4\mathrm{H}_4(\mathrm{x}) + g_3^2 \mathrm{H}_6(\mathrm{x})}{72n} \right] \tag{4.11}$$

where n denotes the sample size and g_j is the j'th cumulant of the x's distribution function.[30]

Next we present Sieve methods to be used for approximating the disturbances' unknown marginal density functions, utilizing Hermite polynomials.

4.5. Unknown marginal density function

In this section, we explain how the disturbances' unknown marginal density functions are approximated by utilizing a finite mixture of basis functions (Sieves). Additionally, we discuss the reason for the choice of Hermite polynomials as the basis functions. The estimation methodology is based

[28] Definition: a set of polynomials in x $\{\mathrm{P_k}(x)\}_{\mathrm{k}=0}^{\mathrm{K_n}}$, $\mathrm{K_n} \leq \infty$ with degree $[\mathrm{P_k}(x)] = \mathrm{k}$ for each k, is called orthogonal with respect to the positive weight function $\mathrm{w}(x)$ on the interval (a, b) with $a < b$ if for each $\mathrm{k}_1, \mathrm{k}_2 \leq \mathrm{K_n}$ (Andrews and Askey, 1985): $\int_a^b \mathrm{P}_{\mathrm{k}_1}(x)\mathrm{P}_{\mathrm{k}_2}(x)\mathrm{w}(x) = \mathrm{h}_{\mathrm{k}_1}$if $\mathrm{k}_1 = \mathrm{k}_2$ and zero otherwise, where the interval (a, b) needs not to be finite.

[29] The classical orthogonal polynomials are named after Hermite, Laguerre and Jacobi. One of their many possible uses is approximating to a density function. The Laguerre polynomial is appropriate for approximating a density function of a non-negative random variable on the support $(0, \infty)$, while Jacobi is appropriate for the bounded support $[-1, 1]$.

[30] The problem with the definition above is that the density can be negative for some x values. However, there is a simple solution to that problem that will be discussed in next section.

on the aforementioned mixtures and is referred to as *the method of Sieves*, where Sieve is a sequence of finite parameter spaces, constructed to approximate a function of interest, using a finite mixture of basis functions (Coppejans, 2001).[31]

Let $f_{n,\eta}$ be an approximation to f_η, $\eta \in \{\xi_1, \xi_2\}$. The approximation is based on a combination of known basis functions denoted by $\{A_{k,\eta}(.) : k \geq 0\}$, and $K_{n,\boldsymbol{\eta}} + 1$ unknown coefficients $\{a_{k,\eta}(.) : k \geq 0\}$, which must be estimated. Chen et al. (2006) proposed the following Sieve space:

$$\mathcal{F}_{n,\eta} = \left\{ f_{n,\eta}(e) = \left[\sum_{k=0}^{K_{n,\eta}} a_{k,\boldsymbol{\eta}} A_{k,\eta}(e) \right]^2, \int f_{n,\eta}(e) dx = 1 \right\}, K_n \to \infty, \frac{K_{n,\eta}}{n} \to 0. \tag{4.12}$$

By construction, the density function in (4.12) is always a non-negative function of e.

For the basis function, Chen et al. (2006) suggest to use Hermite polynomials or splines. However, in our present case, Hermite polynomials are used as the basis functions (mixtures) for computational convenience and due to the unbounded support of the approximated density function.[32]

To insure that the approximation of the density function integrates, one can set:

$$f_{n,\eta}(e) = \frac{\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\boldsymbol{\eta}} A_{k,\eta}(e) \right]^2}{\int \left[\sum_{k=0}^{K_{n,\eta}} a_{k,\boldsymbol{\eta}} A_{k,\eta}(v) \right]^2 dv} \tag{4.13}$$

The density function is approximated using Hermite polynomials as basis functions, as follows:[33]

$$f_{n,\eta}(e) = \frac{\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\boldsymbol{\eta}} (e/\sigma_\eta)^k \right]^2 \phi(e/\sigma_\eta)/\sigma_\eta}{\int_{-\infty}^{\infty} \left[\sum_{k=0}^{K_{n,\eta}} a_{k,\boldsymbol{\eta}} (v/\sigma_\eta)^k \right]^2 \phi(v/\sigma_\eta)/\sigma_\eta dv}, \ \ \eta \in \{\xi_1, \xi_2\} \tag{4.14}$$

where σ_η is a scale parameter to be estimated and $\phi(.)$ is a standard normal density function.

Based on (4.14), the distribution function can be approximated, as well, in the following way:

$$F_{n,\eta}(e) = \frac{\int_{-\infty}^{e} \left[\sum_{k=0}^{K_{n,\eta}} a_{k,\boldsymbol{\eta}} (v/\sigma_\eta)^k \right]^2 \phi(v/\sigma_\eta)/\sigma_\eta dv}{\int_{-\infty}^{\infty} \left[\sum_{k=0}^{K_{n,\eta}} a_{k,\boldsymbol{\eta}} (v/\sigma_\eta)^k \right]^2 \phi(v/\sigma_\eta)/\sigma_\eta dv}, \ \ \eta \in \{\xi_1, \xi_2\}. \tag{4.15}$$

In order to obtain close form expressions for f_{ξ_1} and f_{ξ_1} (similar to Schwiebert (2013)), let:

$$h_{n,\xi_1,\xi_2}(e_1, e_2) = c(F_{n,\xi_1}(e_1), F_{n,\xi_2}(e_2); \tau) f_{n,\xi_1}(e_1) f_{n,\xi_2}(e_2), \tag{4.16}$$

[31] For instance, one can approximate any continuous density function using a finite mixture of normal densities. In such case, the Sieve is the sequence of mixtures (Coppejans, 2001).

[32] Sieve estimator obtained using Hermite polynomials as basis functions is appropriate for approximating a density function with unbounded support. Both the approximated density and distribution function can be obtained analytically, for computation convenience. Thus, it is unnecessary to use numerical integration's methods to obtain the distribution function.

[33] In our procedure, Hermite polynomials are used to approximate the square root of a density. By taking the second power of this estimator, one obtains the non-negative density estimator in (4.14), as suggseted by Gallant and Nychka (1987).

where $F_{n,\eta}(e) = \int_{-\infty}^{e} f_{n,\eta}(v)dv$, $\eta \in \{\xi_1, \xi_2\}$.

The joint cumulative distribution function of ξ_1 and ξ_2 is given by:

$$H_{\xi_1,\xi_2}(e_1, e_2) = C(F_{n,\xi_1}(e_1), F_{n,\xi_2}(e_2); \tau). \tag{4.17}$$

where τ is the parameter that captures the dependence structure between the random variables ξ_1 and ξ_2.

Next we present the likelihood function to be estimated, consisting of the selection model's unknown parameters, including the parameters that are used to approximate the unknown disturbances' marginal density functions, and the unknown disturbances' marginal distribution functions. All these unknown density and distribution functions are approximated using (4.14) and (4.15), respectively.

4.5.1. The likelihood function

For the continuous selection variable model, the likelihood function is:

$$\begin{aligned} &L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}, a_{0,\xi_1}, ..., a_{K_{n,\xi_1},\xi_1}, a_{0,\xi_2}, ..., a_{K_{n,\xi_2},\xi_2}, \sigma_{\xi_1}, \sigma_{\xi_2} | \mathbf{Y_1}, \mathbf{Y_2}, \boldsymbol{X}, \boldsymbol{Z}) \\ &= \prod_{i=1}^{n} \frac{h_{n,\xi_1,\xi_2}(y_{1i} - \boldsymbol{x_i'}\boldsymbol{\beta}, y_{2i} - \boldsymbol{z_i'}\gamma)}{1 + H_{\xi_1,\xi_2}(c_1 - \boldsymbol{x_i'}\boldsymbol{\beta}, c_2 - \boldsymbol{z_i'}\gamma) - f_{n,\xi_1}(c_1 - \boldsymbol{x_i'}\boldsymbol{\beta}) - f_{n,\xi_2}(c_2 - \boldsymbol{z_i'}\gamma)} \end{aligned} \tag{4.18}$$

denoting the parameters vectors by $\boldsymbol{\alpha} = [\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau, a_{0,\xi_1}, ..., a_{K_{n,\xi_1},\xi_1}, a_{0,\xi_2}, ..., a_{K_{n,\xi_2},\xi_2}, \sigma_{\xi_1}, \sigma_{\xi_2}]$, we search for $\boldsymbol{\alpha}$, which maximizes the likelihood function in (4.18).

In the case of the binary selection variable model, y$_2$ is not observed; only S_i is observed, given participation.

In this (binary selection model) case, the likelihood function is:

$$\begin{aligned} &L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}, a_{0,\xi_1}, ..., a_{K_{n,\xi_1},\xi_1}, a_{0,\xi_2}, ..., a_{K_{n,\xi_2},\xi_2}, \sigma_{\xi_1}, \sigma_{\xi_2} | \mathbf{Y_1}, \boldsymbol{X}, \boldsymbol{Z}) \\ &= \prod_{i=1}^{n} \frac{f_{\xi_1}(y_1 - \boldsymbol{x'}\boldsymbol{\beta}) - \frac{\partial}{\partial y_1} H_{\xi_1,\xi_2}(y_1 - \boldsymbol{x'}\boldsymbol{\beta},\ \ c_2 - \boldsymbol{z'}\gamma)}{1 + H_{\xi_1,\xi_2}(c_1 - \boldsymbol{x_i'}\boldsymbol{\beta}, c_2 - \boldsymbol{z_i'}\gamma) - F_{n,\xi_1}(c_1 - \boldsymbol{x_i'}\boldsymbol{\beta}) - F_{n,\xi_2}(c_2 - \boldsymbol{z_i'}\gamma)} \end{aligned} \tag{4.19}$$

denoting the parameters vectors by $\boldsymbol{\alpha} = [\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau, a_{0,\xi_1}, ..., a_{K_{n,\xi_1},\xi_1}, a_{0,\xi_2}, ..., a_{K_{n,\xi_2},\xi_2}, \sigma_{\xi_1}, \sigma_{\xi_2}]$, we search for $\boldsymbol{\alpha}$ which maximizes the likelihood function in (4.19).

Next, we present the asymptotic properties of the estimator $\hat{\boldsymbol{\alpha}}$.

4.6. Asymptotic Properties

In this section, we show the consistency of our estimators using the results in Chen et al. (2006) and Chen (2007). Let $\mathcal{A} = \Theta \times \mathcal{F}_{\xi_1} \times \mathcal{F}_{\xi_2}$ denote the parameter space. The Sieve maximum likelihood function is defined as (see, Schwiebert (2013)):

$$(4.20) \quad \hat{\boldsymbol{\alpha}}_n = \arg \max_{\boldsymbol{\alpha} \in \mathcal{A}_n} \ln L(\boldsymbol{\alpha}; \boldsymbol{Y_1, Y_2, X, Z}) = \sum_{i=1}^{n} \ln l(\boldsymbol{\alpha}; y_{1i}, y_{2i}, x_i, z_i),$$

where $\boldsymbol{\alpha} = (\boldsymbol{\theta}', f_{\xi_1}, f_{\xi_2})'$ and $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\theta}}', \hat{f}_{n,\xi_1}, \hat{f}_{n,\xi_2})' \in \Theta \times \mathcal{F}_{n,\xi_1} \times \mathcal{F}_{n,\xi_2} = \mathcal{A}_n$, such that $\boldsymbol{\alpha_0} = (\boldsymbol{\theta_0}', f_{0,\xi_1}, f_{0,\xi_2})' \in \mathcal{A}$ is the true parameter vector.

Let $\boldsymbol{W} = \{\boldsymbol{Y_1, Y_2, X, Z}\}$, and suppose that the distance function $d(.,.)$ is a (pseudo) metric on $\mathcal{A}$.[34] The following assumptions are taken from conditions 3.1'-3.3' and 3.4-3.5 in Chen (2007) (which appears also in Schwiebert (2013)) and must be imposed to guarantee the consistency of the estimators:

Assumption 6.A *(identification)* $\mathbb{E}[\ln L(\boldsymbol{\alpha; W})]$ *is continuous at* $\boldsymbol{\alpha_0} \in \mathcal{A}$*, and* $\mathbb{E}[\ln L(\boldsymbol{\alpha_0; W})] > -\infty$.

Assumption 6.B *There exists a non-increasing positive function* $\delta(.)$ *and a positive function* $g(.)$ *such that for all* $\epsilon > 0$ *and for all* $k \geq 1$*:*

$$\mathbb{E}[\ln L(\boldsymbol{\alpha_0; W})] - \sup_{\{\boldsymbol{\alpha} \in \mathcal{A}: d(\boldsymbol{\alpha}, \boldsymbol{\alpha_0}) \geq \epsilon\}} \mathbb{E}[\ln L(\boldsymbol{\alpha; W})] \geq \delta(k) g(\epsilon) > 0.$$

The method of Sieves (Chen, 2007) provides one general approach to resolve the difficulties associated with maximizing $\hat{\boldsymbol{\alpha}}$ over an infinite-dimensional space $\mathcal{A}$ by maximizing $\hat{\boldsymbol{\alpha}}$ over a sequence of approximating spaces $\mathcal{A}_k$. We postulate the following assumption:

Assumption 7 *The sequence of approximating spaces are non-decreasing, implying that* $\mathcal{A}_k \subseteq \mathcal{A}_{k+1} \subseteq \mathcal{A}$ *for all* $k \geq 1$*; and for any* $\boldsymbol{\alpha} \in \mathcal{A}$ *there exists a sequence* $\boldsymbol{\pi_k \alpha_0} \in \mathcal{A}_k$ *such that* $d(\boldsymbol{\alpha_0, \pi_k \alpha_0}) \to 0$ *as* $k \to \infty$*. In other words, we assume that asymptotically the difference between an (unknown) function and its Sieve approximation tends to zero, where* π_k *can be regarded as a projection mapping from* $\mathcal{A}_k$ *to* $\mathcal{A}$ *(see Chen (2007)).*

Assumption 8.A $\ln L(\boldsymbol{\alpha, W})$ *is a measurable function of the data* $\boldsymbol{W}$ *for all* $\boldsymbol{\alpha} \in \mathcal{A}_k$.

The continuity assumption:

Assumption 8.B *for any data* $\boldsymbol{W}$, $\ln L(\boldsymbol{\alpha, W})$ *is upper semicontinuous on* $\mathcal{A}_k$ *under the metric* $d(.,.)$.

The Sieve spaces $\mathcal{A}_k$ are required to be closed (Sieve space contains all its limits points) and bounded, leading to the following assumption:

Assumption 9 *The Sieve spaces,* $\mathcal{A}_k$ *are compact under* $d(.,.)$.

[34] A pseudo-metric space is a set $\mathcal{A}$ together with a non-negative real-valued function $d : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$, such that, for every $a_1, a_2, a_3 \in \mathcal{A}$: $d(a_1, a_1) = 0$, $d(a_1, a_2) = d(a_2, a_1)$ (symmetric) and $d(a_1, a_3) \leq d(a_1, a_2) + d(a_2, a_3)$. Unlike a metric space, points in a pseduo-metric space need not be distinguishable: one may have $d(a_1, a_2) = 0$ for distinct values $a_1 \neq a_2$.

The convergence assumption of the sample log-likelihood to its population counterpart over the Sieves:

Assumption 10 *For all* $k \geq 1$, $\text{plim}_{n\to\infty} \sup_{\boldsymbol{\alpha}\in\mathcal{A}_k} |\ln L(\boldsymbol{\alpha}) - \mathbb{E}[\ln L(\boldsymbol{\alpha})]| = 0$ *and* $\eta_{k(n)} = o(\delta(k(n)))$.

Assumptions (6.A) and (6.B) are identification conditions, to ensure that the true parameters vector $\boldsymbol{\alpha_0}$ uniquely maximizes the expected value of the log-likelihood function. Based on Theorem 3.1 in Chen (2007),[35] if assumptions 6.A-10 hold, then $d(\boldsymbol{\hat{\alpha}_n}, \boldsymbol{\alpha_0}) = o_p(1)$, which implies that $\boldsymbol{\hat{\alpha}_n}$ is a consistent estimator for $\boldsymbol{\alpha_0}$.[36] The proof relies on the fact that for all $\epsilon > 0$:

$$(4.21) \qquad Pr(d(\hat{\alpha}_n, \alpha_0) > \epsilon) \leq Pr(\sup_{\{\boldsymbol{\alpha}\in\mathcal{A}_k : d(\boldsymbol{\alpha},\boldsymbol{\alpha_0})\geq\epsilon\}} L(\boldsymbol{\alpha}; \boldsymbol{W}) \geq \ln L(\boldsymbol{\pi_{k(n)}}\boldsymbol{\alpha_0}; \boldsymbol{W}) - O(\eta_{k(n)}))$$

but as the right-hand side approaches zero, as $n \to \infty$, it implies that $Pr(d(\hat{\alpha}_n, \alpha_0) > \epsilon) \to 0$, which proves that $\boldsymbol{\hat{\alpha}_n}$ is a consistent estimator for $\boldsymbol{\alpha}$.

5. Model extensions

In practice, there could be multiple substantive equations (that are correlated in the disturbances across equations) and multiple selection equations. We present the two different model extensions to deal with such cases. The first one is appropriate for continuous selection variables, and the second one is appropriate for binary selection variables.

5.1. The case of multiple equations

In the presence of a system of correlated J equations, consisting of J^* substantive equations and $J - J^*$ selection equations, one can employ an extension of our procedure to allow for multiple equations and correct for the selectivity bias propagated by truncation. For simplicity, we denote the substantive equations' and selection equations' indices by $1, ..., J^*$ and $J^*+1, ..., J$, respectively.

Let the following system of equations be the selection rule and substantive equations:

$$(5.1) \qquad y^*_{ji} = \boldsymbol{w'_{ji}}\boldsymbol{\theta_j} + \xi_{ji}, \;\; j = 1, .., J.$$

where $\boldsymbol{\theta_j}$ is the j'th equation's parameters vector of size $q_j \times 1$, $\boldsymbol{w_{ji}}$ is a covariates vector of size $q_j \times 1$, ξ_j is a random disturbance, and y^*_{ji} is a latent dependent variable.

Both the equations of interest and the selection rules' equations are truncated in a similar fashion to the case of a model consisting of single selection rule discussed in previous sections. The j'th observed dependent variable is defined as $(j = 1, 2, ..., J)$:

[35] See also Theorem 1 in Schwiebert (2013).

[36] See the Proof of Theorem 3.1 in Chen (2007), pp. 5589-5591.

$$(5.2) \qquad y_{ji} = \begin{cases} y^*_{ji} & \text{if} \quad \prod_{q=1}^{J} I(y_{qi} \geq c_q) = 1 \\ \text{Unobserved} & \text{Otherwise} \end{cases}$$

where $c_1, ..., c_J$ are cutoffs, such that c_j denotes a fixed truncation point on y_{ji}.

The multivariate survival function is the probability of each individual observation to be observed (non-truncated), defined as:

$$\begin{aligned} (5.3) \qquad & Pr(\xi_1 > c_1 - \boldsymbol{w'_{1i}\theta_1},\, \dots,\, \xi_J > c_J - \boldsymbol{w'_{Ji}\theta_J}) \\ & = Pr(F_{\xi_1}(c_1 - \boldsymbol{w'_{1i}\theta_1}) < U_1 < 1,\, \dots,\, F_{\xi_J}(c_J - \boldsymbol{w'_{Ji}\theta_J}) < U_J < 1) \\ & = Pr(m_{1i}(1) < U_1 < m_{1i}(0),\, \dots,\, m_{Ji}(1) < U_J < m_{Ji}(0)) \\ & = \sum_{j_1=0}^{1} \dots \sum_{j_J=0}^{1} \left[(-1)^{\sum_{i=1}^{J} j_i}\right] \mathcal{C}\left(m_{1i}(j_1),\, \dots,\, m_{Ji}(j_J); \tau\right) \end{aligned}$$

where $U_t = F_{\xi_t}(\xi_t)$, $m_{ti}(j) = [F_{\xi_t}(c_t - \boldsymbol{w'_{ti}\theta_t})]^j$, $t = 1, ..., J$, $j \in \{0, 1\}$.[37]

The multivariate truncated density function is defined as:

$$(5.4) \qquad f_{\xi_1, \dots, \xi_J}(y_1 - \boldsymbol{w'_{1i}\theta_1},\, \dots,\, y_J - \boldsymbol{w'_{Ji}\theta_J} | \xi_1 > c_1 - \boldsymbol{w'_{1i}\theta_1},\, \dots,\, \xi_J > c_J - \boldsymbol{w'_{Ji}\theta_J})$$

$$= \frac{\left(\prod_{j=1}^{J} f_{\xi_j}(y_j - \boldsymbol{w'_{ji}\theta_j})\right) \boldsymbol{c}(F_{\xi_1}(y_1 - \boldsymbol{w'_{1i}\theta_1}),\, \dots,\, F_{\xi_J}(y_J - \boldsymbol{w'_{Ji}\theta_J}); \tau)}{\sum_{j_1=0}^{1} \dots \sum_{j_J=0}^{1} \left[(-1)^{\sum_{i=1}^{J} j_i}\right] \mathcal{C}\left(m_{1i}(j_1),\, \dots,\, m_{Ji}(j_J); \tau\right)}$$

where $\boldsymbol{c}(u_1,\, \dots,\, u_J; \tau) \equiv \frac{\partial^{(J)}}{\partial u_1,\, \dots,\, \partial u_J} \mathcal{C}(u_1,\, \dots,\, u_J; \tau)$.

The likelihood function is:

$$(5.5) \qquad L(\boldsymbol{\theta_1,\, \dots,\, \theta_j}, \tau, a_{0,\xi_1}, ..., a_{K_{n,\xi_1},\xi_1}, , a_{0,\xi_2}, ..., a_{K_{n,\xi_2},\xi_2} | \boldsymbol{Y_1,\, \dots,\, Y_J, W})$$

$$= \prod_{i=1}^{n} \frac{\left(\prod_{j=1}^{J} f_{\xi_j}(y_j - \boldsymbol{w'_{ji}\theta_j})\right) \boldsymbol{c}(F_{\xi_1}(y_1 - \boldsymbol{w'_{1i}\theta_1}),\, \dots,\, F_{\xi_J}(y_J - \boldsymbol{w'_{Ji}\theta_J}))}{\sum_{j_1=0}^{1} \dots \sum_{j_J=0}^{1} \left[(-1)^{\sum_{i=1}^{J} j_i}\right] \mathcal{C}\left(m_{1i}(j_1),\, \dots,\, m_{Ji}(j_J); \tau\right)}$$

Let $\boldsymbol{\alpha} = [\boldsymbol{\theta_1,\, \dots,\, \theta_J, \tau}, a_{0,\xi_1}, ..., a_{K_{n,\xi_1},\xi_1}, , a_{0,\xi_2}, ..., a_{K_{n,\xi_2},\xi_2}]$ denote the parameters vectors. We search for the parameters vector $\boldsymbol{\alpha}$, which maximizes the likelihood function in (5.5).

In the next section, we discuss the case of multiple substantive and selection equations, in which all selection variables are binary variables.

[37] Last equality in (5.3) is based on proposition (9.1) in the Appendix.

5.2. The case of multiple binary selection equations

In the presence of a system of correlated J equations, consisting of J^* substantive equations and $J - J^*$ binary selection equations, we show that one can apply our theoretic model to correct for the selectivity bias.

Let the following system of equations be the selection rule and substantive equations:

$$y^*_{ji} = \boldsymbol{w'_{ji}\theta_j} + \xi_{ji}, \ \ j = 1, .., J. \tag{5.6}$$

where $\boldsymbol{\theta_j}$ is the j'th equation's parameters vector of size $q_j \times 1$, $\boldsymbol{w_{ji}}$ is a covariates vector of size $q_j \times 1$, ξ_j is a random disturbance, and y^*_{ji} is a latent dependent variable.

Both substantive and selection rules equations are truncated in a similar fashion to the case of a model consisting of single selection rule discussed in previous sections. The observed dependent variable in the j'th substantive equation is defined as $(j = 1, 2, ..., J^*)$:

$$y_{ji} = \begin{cases} y^*_{ji} & \text{if} \quad \prod_{q=1}^{J} I(y_{qi} \geq c_q) = 1 \\ \text{Unobserved} & \text{Otherwise} \end{cases} \tag{5.7}$$

where $c_1, ..., c_J$ are cutoffs, such that c_j denotes a fixed truncation point on y_{ji}.

The observed dependent variable in the j'th selection equation is defined as $(j = J^* + 1, 2, ..., J)$:

$$y_{ji} = \begin{cases} 1 & \text{if} \quad \prod_{q=1}^{J} I(y_{qi} \geq c_q) = 1 \\ \text{Unobserved} & \text{Otherwise} \end{cases} . \tag{5.8}$$

The multivariate survival function is the probability of each individual observation to be observed (non-truncated), defined as:

$$r_{ti}(j) = \begin{cases} F_{\xi_t}(y_t - \boldsymbol{w'_{ti}\theta_t}) & \text{if} \quad j = 0 & \text{and} \quad 1 \leq t \leq J^* \\ F_{\xi_t}(c_t - \boldsymbol{w'_{ti}\theta_t}) & \text{if} \quad j = 1 & \text{and} \quad 1 \leq t \leq J^* \\ \left[F_{\xi_t}(c_t - \boldsymbol{w'_{ti}\theta_t})\right]^j & \text{if} \quad j \in \{0, 1\} & \text{and} \quad J^* + 1 \leq t \leq J \end{cases} \tag{5.9}$$

Based on proposition (9.2) (see the Appendix):

$$\begin{aligned} &Pr(r_{1i}(1) \leq U_1 \leq r_{1i}(0), \ ... \ , r_{Ji}(1) \leq U_J \leq r_{Ji}(0)) \\ &\quad = \sum_{j_1=0}^{1} ... \sum_{j_J=0}^{1} \left[(-1)^{\sum_{i=1}^{J} j_i}\right] \mathcal{C}\left(r_{1i}(j_1), \ ... \ , r_{Ji}(j_J); \tau\right) \end{aligned} \tag{5.10}$$

where $U_t = F_{\xi_t}(\xi_t)$, $t = 1, ..., J$.

The multivariate truncated density function is obtained by setting $j_1 = ... = j_{J^*} = 0$, as follows:

$$f_{\xi_1,\,...\,,\,\xi_{J^*}}(y_1 - \boldsymbol{w'_{1i}\theta_1},\,...\,,\,y_{J^*} - \boldsymbol{w'_{J^*i}\theta_{J^*}}|\xi_1 > c_1 - \boldsymbol{w'_{1i}\theta_1},\,...\,,\,\xi_J > c_J - \boldsymbol{w'_{Ji}\theta_J})$$
$$= \frac{\sum_{j_q=0}^{1} \cdots \sum_{j_J=0}^{1} \left[(-1)^{\sum_{i=q}^{J} j_i}\right] \left(\prod_{j=1}^{J^*} f_{\xi_j}(y_j - \boldsymbol{w'_{ji}\theta_j})\right) \boldsymbol{c^{(J^*)}}\left(r_{1i}(j_1),\,...\,,\,r_{Ji}(j_J);\tau\right)}{\sum_{j_1=0}^{1} \cdots \sum_{j_J=0}^{1} \left[(-1)^{\sum_{i=1}^{J} j_i}\right] \mathcal{C}\left(m_{1i}(j_1),\,...\,,\,m_{Ji}(j_J);\tau\right)} \tag{5.11}$$

where $q \equiv J^* + 1$ and $\boldsymbol{c^{(J^*)}}(u_1,\,...\,,\,u_J;\tau) \equiv \frac{\partial^{(J^*)}}{\partial u_1,\,...\,,\,\partial u_{J^*}}\mathcal{C}(u_1,\,...\,,\,u_J;\tau)$.

The likelihood function is:

$$L(\boldsymbol{\theta_1,\,...\,,\,\theta_j}, \tau, a_{0,\xi_1}, ..., a_{K_{n,\xi_1},\xi_1}, , a_{0,\xi_2}, ..., a_{K_{n,\xi_2},\xi_2}|\boldsymbol{Y_1,\,...\,,\,Y_{J^*}, W})$$
$$= \prod_{i=1}^{n} \frac{\sum_{j_q=0}^{1} \cdots \sum_{j_J=0}^{1} \left[(-1)^{\sum_{i=q}^{J} j_i}\right] \left(\prod_{j=1}^{J^*} f_{\xi_j}(y_j - \boldsymbol{w'_{ji}\theta_j})\right) \boldsymbol{c^{(J^*)}}\left(r_{1i}(j_1),\,...\,,\,r_{Ji}(j_J);\tau\right)}{\sum_{j_1=0}^{1} \cdots \sum_{j_J=0}^{1} \left[(-1)^{\sum_{i=1}^{J} j_i}\right] \mathcal{C}\left(m_{1i}(j_1),\,...\,,\,m_{Ji}(j_J);\tau\right)} \tag{5.12}$$

denoting the parameters vectors by $\boldsymbol{\alpha} = [\boldsymbol{\theta_1,\,...\,,\,\theta_J, \tau}, a_{0,\xi_1}, ..., a_{K_{n,\xi_1},\xi_1}, , a_{0,\xi_2}, ..., a_{K_{n,\xi_2},\xi_2}]$, we search for the parameters vector $\boldsymbol{\alpha}$, which maximizes the likelihood function in (5.12).

The next section presents a simulation conducted in order to validate our theoretic model under different sample sizes and using both model setups: a binary selection variable and a continuous selection variable. The simulation is intended to examine the theoretic model's performance, by making a comparison between the estimates obtained using our suggested methodology for selectivity bias correction and the estimates that would be obtained under no truncation. For the sake of robustness, we measure the proximity of our estimates to the true parameters, using different measures. This will be dealt with in the next section.

6. Simulation

For each of the model's setups (I) and (II), we generate random samples of different sizes $N \in \{1000, 3000, 5000, 10000\}$ and estimate the model's parameters to verify our theoretic model. We repeat this procedure 10,000 times for each sample size. Both model setups do not assume a specific marginal distribution function for each one of the disturbances. However, for the purpose of generating the data, one needs to specify the disturbances' marginal distribution functions. The disturbances' data are generated in two steps. We arbitrarily choose both disturbances to be a demeaned-Gamma distributed,[38] where the marginal distribution functions are denoted by F_{ξ_1} and F_{ξ_2}. The dependence structure is arbitrarily chosen to follow the Clayton Copula, defined as follows:

[38] If one subtracts the expectation from a non zero mean random variable v, the generated random variable $\tilde{\text{v}} = \text{v} - \mathbb{E}[\text{v}]$ is distributed according to the orignal distribution's shape.

$$\mathcal{C}(u_1, u_2) = (u_1^{-\tau} + u_2^{-\tau} - 1)^{-1/\tau}, \qquad 0 \leq u_1, u_2 \leq 1, \tag{6.1}$$

where $\mathcal{C}(.,.)$ denotes a Copula function, the arguments u_1 and u_2 are defined on the unit square ($u_1 \in [0,1]$ and $u_2 \in [0,1]$), the degree of dependence is captured by τ, which is chosen to be 8 (in order to generate highly-correlated disturbances, with a correlation coefficient of 0.85).

We generate a random sample that is a sequence $\{u_{1i}, u_{2i}\}_{i=1}^{N}$, consisting of N independent pairs, obtained as realizations from the bivariate Clayton Copula.

Let $G(e; \lambda_1, \lambda_2)$ denote the demeaned-Gamma distribution function with the shape and scale parameters λ_1 and λ_2, respectively. The argument $e \in (-\lambda_1/\lambda_2, \infty)$ is a specific value that the demeaned Gamma distributed random variable can take.

Then, the disturbances data are constructed using the following transformation: each one of the disturbances e_{1i} and e_{2i} for $i = 1, ..., N$ is generated using the inverse distribution function, denoted by $e_{ji} = G^{-1}(u_{ji}; \lambda_1, \lambda_2)$ $\forall j \in \{1, 2\}$.[39]

We arbitrarily choose the parameters set $[\beta_0, \gamma_0, \beta_1, \beta_2, \gamma_1] = [-12.5, -14, 6, 2, 3]$ to be used in equations (6.4) and (6.5) to follow. In order to generate correlated disturbances, one needs to choose the degree of dependence parameter τ. To demonstrate our suggested procedure's performance, we are interested in highly-correlated disturbances, to guarantee a biased estimates in the absence of correction for the selectivity biased. For this reason, we use $\tau = 8$, leading to a correlation of 0.85 between the disturbances.

We are about to generate a vector consisting of three continuous variables x_i, z_{1i}, z_{2i} and one dichotomous variable D_i, to be used in equations (6.4) and (6.5) to follow. For simplicity, each one of the former continuous variables is randomly drawn from a normal distribution, $\mathrm{x} \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $\mathrm{z}_1 \sim \mathcal{N}(\mu_{z_1}, \sigma_{z_1}^2)$ and $\mathrm{z}_2 \sim \mathcal{N}(\mu_{z_2}, \sigma_{z_2}^2)$, where μ_T and σ_T for $T \in \{X, Z_1, Z_2\}$ denote the expectation's and standard deviation's parameters, respectively. The binary variable is a random drawn from Bernoulli distribution with probability of success 0.2, such that D $\sim$ Bernoulli(0.2). We arbitrarily choose the parameters' values $\mu_x = \mu_{z_1} = 5$ and $\sigma_x = \sigma_{z_1} = 2$, $\mu_{z_2} = 4$, $\sigma_{z_2} = 3$.[40]

As these variables are correlated, as is commonly assumed in many sample selection models, we simplify the process of data generation by splitting it into two steps. Initially, auxiliary data $[v_i^D, v_i^X, v_i^{Z_1}, v_i^{Z_2}]$ are generated as random realizations from a multivariate Clayton Copula with four uniform correlated variables, using a degree of dependence parameter 4 (a correlation coefficient of

[39] The j'th disturbance's realizations are obtained using a transformation on the j'th uniform variable previously generated, as if it is randomly drawn from the demeaned-Gamma distribution function.

[40] The covariates are treated as constants, as we estimate the conditional distribution function of y_1, given the covariates (which is the disturbance's distribution). Thus, the data's distribution function of these covariates is significant only for the purpose of data generation.

0.8).

In the final step, using the auxiliary data, we generate the variable $x_i = 2\phi^{-1}(v_i^X) + 5$ and in a similar fashion, $z_{1i} = 2\phi^{-1}(v_i^{Z_1}) + 5$, where $\phi^{-1}(.)$ is the inverse of the standardized normal distribution function. Additionally, a dichotomous variable D is constructed:

$$(6.2) \qquad D_i = \begin{cases} 1 & \text{if } v_i^D \leq 0.2 \\ 0 & \text{if } v_i^D > 0.2 \end{cases}$$

This dichotomous variable is negatively correlated with all the regressors in the model, with a correlation of -0.6.

We generate an additional variable only for identification of the selection equation's parameters, in the setup for the second model (to be used with a normalized coefficient):

$$(6.3) \qquad z_{2i} = 3\left(0.3\phi^{-1}(v_i^{Z_2}) + 0.7\phi^{-1}(v_i^R)\right) + 4,$$

where v_i^R is randomly drawn from a standardized normal distribution.

We arbitrarily set $\lambda_1 = 20$ and $\lambda_2 = 1/6$, and re-scale each one of the disturbances' standard deviations to 5. The main goal is to demonstrate our procedure's performance when the truncated sample's estimates, which are obtained in the absence of selectivity bias correction, are not close to the true parameter values.

Finally, the latent variables y_1^* and y_2^* are constructed as follows:

$$(6.4) \qquad y_{1i}^* = -12.5 + 6x_i + 2D_i + e_{1i}, \qquad i = 1, ..., N,$$

and,

$$(6.5) \qquad y_{2i}^* = -14 + 3z_{1i} + 1z_{2i} + e_{2i}, \qquad i = 1, ..., N,$$

where the variable z_{2i} is chosen to be the variable with the normalized coefficient.[41]

Without loss of generality, each one of the cutoffs (fixed truncation points) c_1 and c_2 is assumed to be zero. In the truncated sample, the dependent variable in the j'th equation is constructed as follows:

$$(6.6) \qquad y_{ij} = \begin{cases} y_{ij}^* & \text{if} \quad y_{1j}^* \geq c_1, y_{2j}^* \geq c_2 \qquad \forall i \in \{1, ..., N\} \quad \text{and} \quad \forall j \in \{1, 2\} \\ \text{Unobserved} & \text{otherwise} \end{cases}$$

In the binary selection variable case, the truncated sample is: $\{y_{1i}, x_i, D_i, z_{1i}, z_{2i}\}_{i=1}^n$. Similarly,

[41] For the model setup with the binary response selection variable, additional regressor data is generated with a normalized coefficient $\gamma_2 = 1$ (see assumption (4.B)).

in the continuous selection variable case, the truncated sample is: $\{y_{1i}, y_{2i}, x_i, D_i, z_{1i}\}_{i=1}^{n}$.

Our motivation is to demonstrate, for each of the binary and continuous model setups, its performance under non-normally distributed disturbances, and for that purpose the Gamma distribution is arbitrarily chosen as the marginal distribution function.

We conduct a sensitivity test to measure the influence of an increase in number of observations on the accuracy of the truncated sample's estimates.

The first accuracy measure we used is the standardized root mean square error, $RMSE_j$, to measure the bias in the truncated regression estimate $\hat{\theta}_j^{tp}$ relative to the true parameter value θ_j, defined as:

$$(6.7) \qquad RMSE_j\left(n\right) = \left(\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{\hat{\theta}_{i,j}^{tp} - \theta_j}{\theta_j}\right)^2\right)^{1/2},$$

where $\hat{\theta}_{i,j}^{tp}$ stands for the substantive equation's j'th coefficient estimated in the i'th truncated sample. n is the sample size in each one of the n_k Monte Carlo simulations. This measures is calculated using the truncated sample to evaluate separately the model's performance with and without correction for the selection bias.

Another measure is based on a formula similar to the one described in (6.7), and is intended to find the relative accuracy of the truncated sample estimates, in comparison to full sample estimates, defined as:

$$(6.8) \qquad R_j\left(n\right) = \left(\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{\hat{\theta}_{i,j}^{tp} - \hat{\theta}_{i,j}^{p}}{\hat{\theta}_{i,j}^{p}}\right)^2\right)^{1/2},$$

where $\hat{\theta}_{i,j}^{p}$ stands for the substantive equation's j'th coefficient, which is estimated using the full sample. This measure is calculated using the truncated sample, to evaluate separately the model's performance with and without correction for the selection bias.

Another accuracy measure is the change in standardized root mean square due to an increase in the sample size by α additional observations, defined as:

$$(6.9) \qquad \Delta RMSE_j\left(n_1, n_2\right) = \left(\alpha \times \frac{RMSE_j\left(n_2\right) - RMSE_j\left(n_1\right)}{n_2 - n_1}\right),$$

where n_1 and n_2 are the sample size of data set 1 and data set 2, respectively. This is the measure of the contribution of α additional observations to the accuracy of the truncated sample estimates (in terms of proximity to the true parameter). For $\alpha = 1000$, we obtain the results described in table (II).

A similar accuracy measure to the accuracy measure in (6.9) is the change in R_j, due to an increase in the sample size by α additional observations, defined as:

$$\Delta R_j\left(n_1, n_2\right) = \left(\alpha \times \frac{R_j\left(n_2\right) - R_j\left(n_1\right)}{n_2 - n_1}\right), \tag{6.10}$$

This is the measure of the contribution of α additional observations to the truncated sample estimates accuracy (in terms of proximity to full sample estimates). For $\alpha = 1000$ we obtain the results described in table (II).

The last estimates' 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. It depicts the speed of the standard deviation's shrinkage, which is due to increasing the sample size. This coefficient is calculated based on the following ratio:

$$\delta = \frac{\ln\left(\sigma_1/\sigma_2\right)}{\ln\left(n_2/n_1\right)}, \tag{6.11}$$

where σ_1 and σ_2 are the estimator's standard deviations calculated in data set 1 and data set 2, respectively (calculated for a given estimator).

Based on table (III), the distance between the truncated sample's estimates and the full sample's estimates is getting smaller with the number of observations. This is captured by R_j which implies that there is a greater similarity between the truncated sample's estimates and the full sample's estimates as number of observations increases. Additionally, the contribution of 1,000 additional observations to the truncated sample estimates accuracy (in terms of proximity to the full sample's estimates) is higher, the smaller the initial sample. This is embodied in the value of ΔR_j which becomes smaller and eventually approaches zero, as the number of observations increases.

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 point to the validity and robustness of our correction for the selectivity bias generated by truncated samples.

TABLE I

Monte Carlo Simulation

Parameter	Model setup							
	Binary				Continuous			
	1000	3000	5000	10000	1000	3000	5000	10000
Full sample model's estimates								
$\beta_0 = -12.5$	-12.5052	-12.4986	-12.4987	-12.5006	-12.4924	-12.4884	-12.5021	-12.4983
median	-12.5010	-12.5011	-12.4999	-12.5006	-12.4944	-12.4935	-12.5029	-12.4983
std	(0.6002)	(0.3479)	(0.2624)	(0.1953)	(0.6014)	(0.3455)	(0.2145)	(0.1926)
$\beta_1 = 6$	6.0009	5.9999	5.9998	6.0001	5.9986	5.9979	6.0005	5.9996
median	6.0012	6.0000	5.9999	6.0001	5.9992	5.9995	6.0007	5.9996
std	(0.1051)	(0.0610)	(0.0459)	(0.0342)	(0.1055)	(0.0603)	(0.0377)	(0.0336)
$\beta_2 = 2$	2.0044	1.9960	1.9989	2.0004	1.9968	1.9952	2.0024	2.0017
median	2.0019	2.0001	1.9973	2.0004	1.9945	1.9937	2.0020	2.0017
std	(0.5219)	(0.3036)	(0.2319)	(0.1665)	(0.5218)	(0.3062)	(0.1847)	(0.1702)
Truncated sample model's estimates - without correction								
$\beta_0 = -12.5$	-8.8179	-8.8159	-8.8148	-8.8118	-6.3490	-6.3640	-6.3703	-6.3731
median	-8.8182	-8.8147	-8.8210	-8.8118	-6.3405	-6.3789	-6.3716	-6.3731
std	(0.6860)	(0.3976)	(0.2996)	(0.2170)	(0.7639)	(0.4460)	(0.2633)	(0.2440)
$\beta_1 = 6$	5.5334	5.5333	5.5331	5.5327	5.3076	5.3103	5.3116	5.3119
median	5.5342	5.5329	5.5343	5.5327	5.3069	5.3121	5.3116	5.3119
std	(0.1170)	(0.0678)	(0.0509)	(0.0372)	(0.1272)	(0.0740)	(0.0439)	(0.0405)
$\beta_2 = 2$	4.2677	4.2739	4.2770	4.2747	4.1761	4.1675	4.1642	4.1736
median	4.2685	4.2679	4.2793	4.2747	4.1610	4.1701	4.1637	4.1736
std	(0.6797)	(0.3955)	(0.3012)	(0.2151)	(0.8829)	(0.5098)	(0.3060)	(0.2771)
Truncated sample model's estimates - with correction								
$\beta_0 = -12.5$	-12.4220	-12.6943	-12.7650	-12.8293	-12.4042	-12.5989	-12.6276	-12.6622
median	-12.4447	-12.6925	-12.7464	-12.8051	-12.4356	-12.5701	-12.6187	-12.6112
std	(1.3767)	(0.8188)	(0.6141)	(0.4595)	(0.9356)	(0.5432)	(0.3307)	(0.3294)
$\beta_1 = 6$	6.0098	6.0044	6.0024	6.0010	5.9994	6.0005	5.9999	6.0000
median	6.0095	6.0048	6.0027	6.0009	5.9989	6.0012	6.0000	5.9999
std	(0.1193)	(0.0626)	(0.0461)	(0.0327)	(0.0389)	(0.0218)	(0.0134)	(0.0118)
$\beta_2 = 2$	2.0485	2.0184	2.0085	2.0011	2.0558	2.0204	2.0096	2.0082
median	2.0397	2.0154	2.0051	2.0001	2.0469	2.0223	2.0079	2.0074
std	(0.5809)	(0.3162)	(0.2362)	(0.1664)	(0.6046)	(0.3407)	(0.2066)	(0.1819)
$\gamma_0 = 14$	-13.9869	-13.6990	-13.7852	-13.8548	-14.0108	-14.1607	-14.1784	-14.2024
median	-14.0133	-13.8683	-13.8808	-13.9157	-14.0070	-14.1177	-14.1644	-14.1454
std	(2.8198)	(1.8419)	(1.2504)	(0.7879)	(1.0419)	(0.5977)	(0.3781)	(0.3599)
$\gamma_1 = 3$	3.0437	3.0111	3.0048	3.0018	3.0003	3.0003	3.0006	3.0005
median	3.0127	3.0024	3.0032	3.0008	2.9996	3.0002	3.0006	3.0004
std	(0.4098)	(0.2040)	(0.1481)	(0.1043)	(0.0501)	(0.0280)	(0.0173)	(0.0154)

Note: We estimate by the semiparametric maximum likelihood method the parameters for the truncated sample and the full sample, and compute the standard deviation in every random sample consisting of N observations. For simplicty, each one of the density functions is approximated using four components. Then, we calculate, for these estimates, the mean, median and standard deviation (Std.) over all data sets. The standard deviations are obtained using the estimates from the Monte Carlo simulations.

TABLE II

Convergence Measures (I)

Parameter	Model setup							
	Binary				Continuous			
	1000	3000	5000	10000	1000	3000	5000	10000
Full sample $RMSE_j$								
β_0	0.1103	0.0673	0.0535	0.0452	0.0752	0.0442	0.0284	0.0294
β_1	0.0200	0.0104	0.0077	0.0054	0.0065	0.0036	0.0022	0.0020
β_2	0.2914	0.1584	0.1182	0.0832	0.3036	0.1706	0.1034	0.0910
Truncated sample without correction $RMSE_j$								
β_0	0.2996	0.2964	0.2958	0.2956	0.4959	0.4922	0.4908	0.4905
β_1	0.0802	0.0786	0.0783	0.0781	0.1173	0.1156	0.1150	0.1149
β_2	1.1837	1.1540	1.1484	1.1424	1.1742	1.1133	1.0929	1.0956
Truncated sample with correction $RMSE_j$								
β_0	0.1103	0.0673	0.0535	0.0452	0.0752	0.0442	0.0284	0.0294
β_1	0.0200	0.0104	0.0077	0.0054	0.0065	0.0036	0.0022	0.0020
β_2	0.2914	0.1584	0.1182	0.0832	0.3036	0.1706	0.1034	0.0910
γ_0	0.2014	0.1333	0.0906	0.0572	0.0744	0.0442	0.0299	0.0295
γ_1	0.1374	0.0681	0.0494	0.0348	0.0167	0.0093	0.0058	0.0051
Full sample $\Delta RMSE_j$								
β_0	-0.0215	-0.0069	-0.0017	0.0000	-0.0155	-0.0079	0.0002	0.0000
β_1	-0.0048	-0.0014	-0.0004	0.0000	-0.0014	-0.0007	-0.0001	0.0000
β_2	-0.0665	-0.0201	-0.0070	0.0000	-0.0665	-0.0336	-0.0025	0.0000
Truncated sample without correction $\Delta RMSE_j$								
β_0	-0.0016	-0.0003	-0.0000	0.0000	-0.0018	-0.0007	-0.0001	0.0000
β_1	-0.0008	-0.0002	-0.0000	0.0000	-0.0009	-0.0003	-0.0000	0.0000
β_2	-0.0148	-0.0028	-0.0012	0.0000	-0.0304	-0.0102	0.0005	0.0000
Truncated sample with correction $\Delta RMSE_j$								
β_0	-0.0215	-0.0069	-0.0017	0.0000	-0.0155	-0.0079	0.0002	0.0000
β_1	-0.0048	-0.0014	-0.0004	0.0000	-0.0014	-0.0007	-0.0001	0.0000
β_2	-0.0665	-0.0201	-0.0070	0.0000	-0.0665	-0.0336	-0.0025	0.0000
γ_0	-0.0341	-0.0213	-0.0067	0.0000	-0.0151	-0.0072	-0.0001	0.0000
γ_1	-0.0346	-0.0093	-0.0029	0.0000	-0.0037	-0.0018	-0.0001	0.0000
δ coefficient (n^δ is the convergence rate)								
β_0	-	0.4729	0.5016	0.4766	-	0.4950	0.6461	0.4534
β_1	-	0.5880	0.5908	0.5627	-	0.5296	0.6630	0.5180
β_2	-	0.5535	0.5591	0.5430	-	0.5221	0.6671	0.5216
γ_0	-	0.3876	0.5053	0.5538	-	0.5058	0.6298	0.4617
γ_1	-	0.6351	0.6322	0.5945	-	0.5308	0.6596	0.5112

Note: We examine three different measures. First, the standardized root mean square error of substantive equation's estimates. Calculated separately for the full sample model, truncated sample without correction, and the truncated sample with correction for the selection bias. Second, the marginal effect of increasing the sample by 1000 observations on the $RMSE_j$ measure. The third measure is the δ coefficient which is calculated for the truncated regression 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.

TABLE III

Convergence Measures (II)

Parameter	Model setup							
	Binary				Continuous			
	1000	3000	5000	10000	1000	3000	5000	10000
Truncated sample without correction R_j								
β_0	0.2974	0.2955	0.2952	0.2953	0.4949	0.4915	0.4908	0.4904
β_1	0.0787	0.0780	0.0779	0.0780	0.1162	0.1150	0.1149	0.1147
β_2	1.5546	1.2118	1.1787	1.1570	1.7439	1.1739	1.1102	1.1102
Truncated sample with correction R_j								
β_0	0.1108	0.0686	0.0543	0.0460	0.0924	0.0545	0.0340	0.0340
β_1	0.0191	0.0101	0.0076	0.0055	0.0177	0.0101	0.0064	0.0056
β_2	0.5805	0.1980	0.1458	0.1024	0.7541	0.2260	0.1306	0.1151
Truncated sample without correction ΔR_j								
β_0	-0.0010	-0.0001	0.0000	0.0000	-0.0017	-0.0003	-0.0001	0.0000
β_1	-0.0003	-0.0000	0.0000	0.0000	-0.0006	-0.0000	-0.0000	0.0000
β_2	-0.1714	-0.0166	-0.0043	0.0000	-0.2850	-0.0318	0.0000	0.0000
Truncated sample with correction ΔR_j								
β_0	-0.0211	-0.0072	-0.0017	0.0000	-0.0189	-0.0102	-0.0000	0.0000
β_1	-0.0045	-0.0012	-0.0004	0.0000	-0.0038	-0.0019	-0.0001	0.0000
β_2	-0.1912	-0.0261	-0.0087	0.0000	-0.2640	-0.0477	-0.0031	0.0000

Notes: We examine two additional measures. The relative difference between the truncated sample estimates and the full sample estimates is calculated based on equation (6.8) for the convergence rate's evaluation (as a function of observations' number). The fourth measure is the marginal effect of increasing the sample by 1000 observations on the R_j measure.

The next section presents a practical guide for the sample selection model's estimation.

7. A practical guide

In this section, we discuss the estimation procedure of our theoretic model. The first step is to obtain initial values for the selection model's regressions. One can obtain these initial estimates by OLS estimation of each one of the equations separately, using the observed (truncated) sample, as follows:

$$\hat{\boldsymbol{\beta}}_{\mathbf{0}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{Y_1} \tag{7.1}$$

For the continuous model setup, one can obtain the initial parameters by OLS estimation:

$$\hat{\boldsymbol{\gamma}}_{\mathbf{0}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\mathbf{Y_2} \tag{7.2}$$

Then, one needs to employ the following procedure to obtain initial estimates for Sieve coefficients

$[a_{0,\xi_1},...,a_{K_{n,\xi_1},\xi_1}]$ and $[a_{0,\xi_2},...,a_{K_{n,\xi_2},\xi_2}]$ for the substantive and selection equations, respectively:

$$L(a_{0,\xi_1},...,a_{K_{n,\xi_1},\xi_1}|\boldsymbol{X},\mathbf{Y_1},\hat{\boldsymbol{\beta}}_{\mathbf{0}}) = \sum_{i=1}^{n} \ln(f_{n,\xi_1}(y_{1i} - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}_{\mathbf{0}})) \tag{7.3}$$

and,

$$L(a_{0,\xi_2},...,a_{K_{n,\xi_2},\xi_2}|\boldsymbol{Z},\mathbf{Y_2},\hat{\boldsymbol{\gamma}}_{\mathbf{0}}) = \sum_{i=1}^{n} \ln(f_{n,\xi_2}(y_{2i} - \boldsymbol{z_i'}\hat{\boldsymbol{\gamma}}_{\mathbf{0}})) \tag{7.4}$$

In the case of a binary response variable, one can only estimate (7.3) and use the initial values obtained from (7.3) for the parameters of the second equation in (7.4).

The second step involves the maximization of the semi parametric likelihood function with respect to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and Sieve coefficients, where the initial estimates are obtained from (7.1)-(7.4).

In the case of a continuous selection variable, the log-likelihood function to be maximized is[42]:

$$\begin{aligned} &\ln L(\boldsymbol{\beta},\boldsymbol{\gamma},a_{0,\xi_1},...,a_{K_{n,\xi_1},\xi_1},a_{0,\xi_2},...,a_{K_{n,\xi_2},\xi_2},\tau,\sigma_{\xi_1},\sigma_{\xi_2}|\boldsymbol{Z},\mathbf{Y_2}) \\ &= \sum_{i=1}^{n} \ln f_{0,\xi_1}(y_{1i}-\boldsymbol{x_i'\beta}) + \sum_{i=1}^{n} \ln f_{0,\xi_2}(y_{1i}-\boldsymbol{z_i'}\gamma) + \sum_{i=1}^{n} \ln \boldsymbol{c}(F_{0,\xi_1}(y_{1i}-\boldsymbol{x_i'\beta}),F_{0,\xi_2}(y_{2i}-\boldsymbol{z_i'}\gamma);\tau_0) \\ &\quad - \sum_{i=1}^{n} \ln\left(1+\mathcal{C}(F_{0,\xi_1}(c_1-\boldsymbol{x_i'\beta}),F_{0,\xi_2}(c_2-\boldsymbol{z_i'}\gamma);\tau) - F_{0,\xi_1}(c_1-\boldsymbol{x_i'\beta}) - F_{0,\xi_2}(c_2-\boldsymbol{z_i'}\gamma)\right) \end{aligned} \tag{7.5}$$

Similarly, for a binary selection variable, the log-likelihood function to be maximized is:

$$\begin{aligned} &\ln L(\boldsymbol{\beta},\boldsymbol{\gamma},a_{0,\xi_1},...,a_{K_{n,\xi_1},\xi_1},a_{0,\xi_2},...,a_{K_{n,\xi_2},\xi_2},\tau,\sigma_{\xi_1},\sigma_{\xi_2}|\boldsymbol{Z},\mathbf{Y_2}) \\ &= \sum_{i=1}^{n} \ln f_{0,\xi_1}(y_{1i}-\boldsymbol{x_i'\beta}) + \sum_{i=1}^{n} \ln \boldsymbol{\mathcal{C}}_{\mathbf{1}}(F_{0,\xi_1}(y_{1i}-\boldsymbol{x_i'\beta}),F_{0,\xi_2}(c_2-\boldsymbol{z_i'}\gamma);\tau_0) \\ &\quad - \sum_{i=1}^{n} \ln\left(1+\mathcal{C}(F_{0,\xi_1}(c_1-\boldsymbol{x_i'\beta}),F_{0,\xi_2}(c_2-\boldsymbol{z_i'}\gamma);\tau) - F_{0,\xi_1}(c_1-\boldsymbol{x_i'\beta}) - F_{0,\xi_2}(c_2-\boldsymbol{z_i'}\gamma)\right) \end{aligned} \tag{7.6}$$

The algorithm that we suggest for finding the optimal parameter values is the Trust-Region numeric search method available in R and MATLAB softwares. This method substitutes each one of the functions in (7.6) and (7.5) with a quadratic approximation (second order Taylor expansion) by an evaluation of its gradient and Hessian. For speed of calculation, this function requires the analytic gradient of the minus of log-likelihood to be minimized (see appendix for the analytic gradient).

[42] See sections (9.1) and (9.2) in the Appendix, for obtaining analytic expressions of the log-likelihood function and its Gradient.

8. Conclusions

Selectivity bias in truncated data stemming from self-selection can be corrected using our proposed correction procedure, which involves the decomposition of the disturbances' joint density function into a nonparametric and a parametric component using a Copula function. Both of these components are estimated together using the semiparametric maximum likelihood method to obtain consistent estimates of the truncated sample selection model's unknown coefficients. The nonparametric component is the disturbances' marginal density functions (approximated by the method of Sieves with Hermite polynomials as basis functions). The parametric component includes the truncated sample selection model's unknown coefficients and the dependence structure which is specified by the Copula function and depends on a parameter, which determines the degree of dependence.

Our correction procedure can also be employed in cases where there is a fixed truncation point on the dependent variable of interest, in addition to the self-selection. Additionally, the selection variable can be either binary or continuous. In order to verify our proposed procedure's estimates accuracy, two different models have been examined by Monte Carlo simulations, using different sample sizes. In the first, the selection variable is continuous, and in the second the selection variable is binary. In both models, the observed data is truncated due to the self-selection of observations and the existence of a fixed truncation point on the dependent variable of interest.

The estimates obtained are accurate and consistent estimates. The binary response model's estimates are accurate when the sample size is above 5,000 observations. Similarly, the continuous selection variable model's estimates are accurate for sample sizes above 5,000 observations. This model assumes a Copula function that must be specified correctly, in order to achieve consistent estimates. Next, we are about to extend the theory to a nonparametric approach, and allow for a more flexible model setup that does not require a specification of the Copula function.

9. Appendix

Proposition 9.1 *Let U_1, ... , U_2 be a sequence of q dependent uniform random variables. The multivariate survival function of the random variables U_1, ... , U_q must satisfy:*

$$(9.1) \qquad Pr(U_1 > u_1, ..., U_q > u_q) = \sum_{i_1=0}^{1} ... \sum_{i_q=0}^{1} (-1)^{\sum_{j=1}^{q} i_j} Pr(U_1 < u_1^{i_1}, ..., U_q < u_q^{i_q}).$$

Proof: Let $E = \{1, .., q\}$ be the index set and we define the event $\mathcal{B}_i = \{U_i \leq u_i\}$, $i \in E$. Using c to denote complementarity, one obtains by the inclusion-exclusion principle the probability of the non-survival event:

$$(9.2) \qquad 1 - Pr(\underset{i}{\cap}\mathcal{B}_i^c) = Pr(\underset{i}{\cup}\mathcal{B}_i) = \sum_i Pr(\mathcal{B}_i) - \sum_{i,j} Pr(\mathcal{B}_i \cap \mathcal{B}_j) + \sum_{i,j,k} Pr(\mathcal{B}_i \cap \mathcal{B}_j \cap \mathcal{B}_k) - ...$$

where the summations are over all singletons, doublets, triplets, and any elements subset of E, such that subsets consisting of an even number of events are subtracted.

The multivariate survival probability is the complementary probability of (9.2):

$$(9.3) \qquad Pr(U_1 > u_1, ..., U_q > u_q) = Pr(\underset{i}{\cap}\mathcal{B}_i^c)$$

$$= 1 - \sum_i Pr(\mathcal{B}_i) + \sum_{i,j} Pr(\mathcal{B}_i \cap \mathcal{B}_j) - \sum_{i,j,k} Pr(\mathcal{B}_i \cap \mathcal{B}_j \cap \mathcal{B}_k) + ...$$

$$= \sum_{i_1=0}^{1} ... \sum_{i_q=0}^{1} (-1)^{\sum_{j=1}^{q} i_j} Pr(U_1 < u_1^{i_1}, ..., U_q < u_q^{i_q})$$

Since by definition $Pr(\mathcal{B}_i) = Pr(U_i < u_i)$ implies that $Pr(\mathcal{B}_i \cup \mathcal{B}_i^c) = Pr(U_i < 1)$. The last equality in (9.3) stems from the following property of multivariate probability functions:

$$(9.4) \qquad Pr(\mathcal{B}_i) = Pr(\mathcal{B}_i \cap (\underset{j \neq i}{\cap}(\mathcal{B}_j \cup \mathcal{B}_j^c))) = Pr(U_i < u_i \underset{j \neq i}{\cap} U_j < 1)$$

and similarly, for any indices subset $A \subseteq E$:

$$(9.5) \qquad Pr(\underset{i \in A}{\cap}\mathcal{B}_i) = Pr(\underset{i \in A}{\cap}\mathcal{B}_i \cap (\underset{j \notin A}{\cap}(\mathcal{B}_j \cup \mathcal{B}_j^c))) = Pr(\underset{i \in A}{\cap}(U_i < u_i) \underset{j \notin A}{\cap}(U_j < 1)).$$

Q.E.D.

Proposition 9.2 *Let U_1, ... , U_2 be a sequence of q dependent uniform random variables. The multivariate probability function of the random variables U_1, ... , U_q must satisfy:*

$$(9.6) \qquad Pr(\underline{u}_1 < U_1 < \overline{u}_1, ..., \underline{u}_q < U_q < \overline{u}_q)$$

$$= \sum_{i_1=0}^{1} ... \sum_{i_q=0}^{1} (-1)^{\sum_{j=1}^{q} i_j} Pr(U_1 < Q_1(i_1), ..., U_q < Q_q(i_q)),$$

where $0 \leq \underline{u}_{q^*} < \overline{u}_{q^*} \leq 1 \; \forall q^* \in \{1, ..., q\}$, such that $Q_{q^*}(i) = \overline{u}_{q^*}$ if $i = 0$ and $Q_{q^*}(i) = \underline{u}_{q^*}$ if $i = 1$.

Proof: Let $E = \{1, .., q\}$ be the index set and we define the events $\mathcal{B}_i = \{U_i \leq \underline{u}_i\}$ and $\mathcal{D}_i = \{U_i \leq \overline{u}_i\}$, $i \in E$. Using c to denote complementarity, one obtains:

$$(9.7) \qquad Pr\left(\underset{m}{\cap}\mathcal{D}_m \cap \left(\underset{i}{\cap}\mathcal{B}_i^c\right)^c\right) = Pr(\underset{m}{\cap}\mathcal{D}_m) - Pr((\underset{i}{\cap}\mathcal{B}_i^c) \cap (\underset{m}{\cap}\mathcal{D}_m))$$

$$= \left[Pr(\underset{m}{\cap}\mathcal{D}_m | \underset{m}{\cap}\mathcal{D}_m) - Pr(\underset{i}{\cap}\mathcal{B}_i^c | \underset{m}{\cap}\mathcal{D}_m)\right] Pr(\underset{m}{\cap}\mathcal{D}_m)$$

$$= \left[1 - Pr(\underset{i}{\cap}\mathcal{B}_i^c | \underset{m}{\cap}\mathcal{D}_m)\right] Pr(\underset{m}{\cap}\mathcal{D}_m)$$

The first term in the last equality can be expressed by using the inclusion-exclusion principle, as follows:

$$(9.8) \qquad 1 - Pr(\underset{i}{\cap}\mathcal{B}_i^c | \underset{m}{\cap}\mathcal{D}_m) = Pr(\underset{i}{\cup}\mathcal{B}_i | \underset{m}{\cap}\mathcal{D}_m) = \sum_i Pr(\mathcal{B}_i | \underset{m}{\cap}\mathcal{D}_m)$$
$$- \sum_{i,j} Pr(\mathcal{B}_i \cap \mathcal{B}_j | \underset{m}{\cap}\mathcal{D}_m) + \sum_{i,j,k} Pr(\mathcal{B}_i \cap \mathcal{B}_j \cap \mathcal{B}_k | \underset{m}{\cap}\mathcal{D}_m) - ...$$

where the summations are over all singletons, doublets, triplets, and any elements subset of E, such that subsets consisting of an even number of events are subtracted.

The following multivariate probability is obtained using (9.7) and (9.8):

$$(9.9) \qquad Pr(\underline{u}_1 < U_1 < \overline{u}_1, ..., \underline{u}_q < U_q < \overline{u}_q) = Pr((\underset{i}{\cap}\mathcal{B}_i^c) \cap (\underset{m}{\cap}\mathcal{D}_m))$$

$$= Pr(\underset{m}{\cap}\mathcal{D}_m) - \sum_i Pr(\mathcal{B}_i \cap (\underset{m}{\cap}\mathcal{D}_m)) + \sum_{i,j} Pr(\mathcal{B}_i \cap \mathcal{B}_j \cap (\underset{m}{\cap}\mathcal{D}_m)) - \sum_{i,j,k} Pr(\mathcal{B}_i \cap \mathcal{B}_j \cap \mathcal{B}_k \cap (\underset{m}{\cap}\mathcal{D}_m)) + ...$$

$$= \sum_{i_1=0}^{1} ... \sum_{i_q=0}^{1} (-1)^{\sum_{j=1}^{q} i_j} Pr(U_1 < Q_1(i_1), ..., U_q < Q_q(i_q))$$

where $0 \leq \underline{u}_{q^*} < \overline{u}_{q^*} \leq 1 \ \forall q^* \in \{1, ..., q\}$, such that $Q_{q^*}(i) = \overline{u}_{q^*}$ if $i = 0$ and $Q_{q^*}(i) = \underline{u}_{q^*}$ if $i = 1$.

The last equality in (9.9) stems from the following property of multivariate probability functions:

$$(9.10) \qquad Pr(\mathcal{B}_i \cap (\underset{m}{\cap}\mathcal{D}_m)) = Pr((\mathcal{B}_i \cap \mathcal{D}_i) \cap (\underset{m \neq i}{\cap}\mathcal{D}_m)) = Pr(U_i < \underline{u}_i \cap (\underset{m \neq i}{\cap} U_m < \overline{u}_m))$$

and similarly, for any indices subset $A \subseteq E$:

$$(9.11) \qquad Pr((\underset{i \in A}{\cap}\mathcal{B}_i) \cap (\underset{m}{\cap}\mathcal{D}_m)) = Pr(\underset{i \in A}{\cap}(\mathcal{B}_i \cap \mathcal{D}_i) \cap (\underset{m \neq A}{\cap}\mathcal{D}_m))$$
$$= Pr(\underset{i \in A}{\cap}(U_i < \underline{u}_i) \cap (\underset{m \notin A}{\cap}(U_m < \overline{u}_m))).$$

Q.E.D.

9.1. Analytic log-likelihood function

The log-likelihood function in (7.5) and (7.5) depends on the disturbances' marginals (density and distribution functions),[43] and depends also on the Copula function. Formalizing the log-likelihood function analytically leads to an analytic gradient expression as described in sections (9.2.1) and (9.2.2) to follow. Obtaining the analytic Gradient and log-likelihood function is required in order to save computation time.

Next we present the analytic expressions for Gaussian Copula to be used in equations (7.5) and (7.5).

[43] The closed-form expressions for these functions are presented in equations (9.45) and (9.46) to follow.

9.1.1. Gaussian Copula

In this section, we show the analytic derivatives of a bivariate Gaussian Copula with respect to all its arguments. The standardized bivariate normal distribution and density functions are denoted by $P(\mathrm{Z}_1 < z_1, \mathrm{Z}_2 < z_2) = \Phi_2(z_1, z_2, \rho)$ and $f_{\mathrm{Z}_1,\mathrm{Z}_2}(z_1, z_2) = \phi_2(z_1, z_2, \rho)$, respectively. ρ is the correlation parameter and Z_1 and Z_2 are the correlated random variables.

Let $u_1 \in [0, 1]$, $u_2 \in [0, 1]$ and τ be the arguments of the Copula function, denoted by C. The Copula function is then defined as:

$$\mathcal{C}(u_1, u_2, \tau) = \Phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \tanh(\tau)\right), \tag{9.12}$$

where $\rho = \tanh(.)$ is the correlation parameter that is bounded by the support $[-1, 1]$, using the hyperbolic tangent function, which is a mapping from $(-\infty, \infty) \to [-1, 1]$.

$$\varrho(u_1, u_2, \tau) = \frac{\Phi^{-1}(u_2) - \tanh(h)\Phi^{-1}(u_1)}{\sqrt{1 - \tanh(h)^2}} \tag{9.13}$$

$$\varsigma_1(\tau) = \tanh^2(\tau)/(1 - \tanh^2(\tau))^{3/2} + 1/(1 - \tanh^2(\tau))^{1/2} \tag{9.14}$$

$$\varsigma_2(\tau) = \tanh(\tau)/(1 - \tanh^2(\tau))^{3/2} \tag{9.15}$$

$$\mathcal{C}_1(u_1, u_2, \tau) = \frac{\partial \mathcal{C}(u_1, u_2, \tau)}{\partial u_1} = \Phi\left(\varrho(u_1, u_2, \tau)\right) \tag{9.16}$$

$$\mathcal{C}_2(u_1, u_2, \tau) = \frac{\partial \mathcal{C}(u_1, u_2, \tau)}{\partial u_2} = \Phi\left(\varrho(u_2, u_1, \tau)\right) \tag{9.17}$$

$$\mathcal{C}_\tau(u_1, u_2, \tau) = \frac{\partial \mathcal{C}_1(u_1, u_2, \tau)}{\partial \tau} = (1 - \tanh^2(\tau))\phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \tanh(\tau)\right) \tag{9.18}$$

$$\mathcal{C}_{1\tau}(u_1, u_2, \tau) = \frac{\partial \mathcal{C}_1(u_1, u_2, \tau)}{\partial \tau} = \frac{(\Phi^{-1}(u_2)\varsigma_2(\tau) - \Phi^{-1}(u_1)\varsigma_1(\tau))\phi\left(\varrho(u_1, u_2, \tau)\right)}{(1 - \tanh^2(\tau))^{-1}} \tag{9.19}$$

$$\mathcal{C}_{11}(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_1(u_1,u_2,\tau)}{\partial u_1}=-\frac{\tanh(\tau)}{\phi(\Phi^{-1}(u_1))\sqrt{1-\tanh^2(\tau)}}\phi\left(\varrho(u_1,u_2,\tau)\right) \tag{9.20}$$

$$\mathcal{C}_{22}(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_2(u_1,u_2,\tau)}{\partial u_2}=-\frac{\tanh(\tau)}{\phi(\Phi^{-1}(u_2))\sqrt{1-\tanh^2(\tau)}}\phi\left(\varrho(u_2,u_1,\tau)\right) \tag{9.21}$$

$$\mathcal{C}_{12}(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_1(u_1,u_2,\tau)}{\partial u_2}=\frac{1}{\phi(\Phi^{-1}(u_2))\sqrt{1-\tanh^2(\tau)}}\phi\left(\varrho(u_1,u_2,\tau)\right) \tag{9.22}$$

$$\begin{aligned}\mathcal{C}_{12\tau}(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_{12}(u_1,u_2,\tau)}{\partial\tau}&=\frac{(1-\tanh^2(\tau))\tan(\tau)}{\phi(\Phi^{-1}(u_2))(1-\tanh^2(\tau))^{3/2}}\phi\left(\varrho(u_1,u_2,\tau)\right)\\&-\frac{(\Phi^{-1}(u_2)\varsigma_2(\tau)-\Phi^{-1}(u_1)\varsigma_1(\tau))}{\phi(\Phi^{-1}(u_2))\sqrt{1-\tanh^2(\tau)}}(1-\tanh^2(\tau))\phi\left(\varrho(u_1,u_2,\tau)\right)\varrho(u_1,u_2,\tau)\end{aligned} \tag{9.23}$$

$$\mathcal{C}_{121}(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_{12}(u_1,u_2,\tau)}{\partial u_1}=\frac{\tanh(\tau)\phi\,(\varrho(u_1,u_2,\tau)\,\varrho(u_1,u_2,\tau)}{\phi(\Phi^{-1}(u_2))\left(1-\tanh^2(\tau)\right)} \tag{9.24}$$

$$\mathcal{C}_{122}(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_{22}(u_1,u_2,\tau)}{\partial u_1}=\frac{\tanh(\tau)\phi\,(\varrho(u_2,u_1,\tau)\,\varrho(u_2,u_1,\tau)}{\phi(\Phi^{-1}(u_1))\left(1-\tanh^2(\tau)\right)} \tag{9.25}$$

9.1.2. Clayton Copula

In this section, we show the analytic derivatives of a bivariate Clayton Copula that belong to the Archimedes Copula family, with respect to all its arguments.

Let $u_1\in[0,1]$, $u_2\in[0,1]$ and τ be the arguments of the Copula function, denoted by C. The Copula function is defined as:

$$\mathcal{C}(u_1,u_2,\tau)=\left(u_1^{-\tau}+u_2^{-\tau}-1\right)^{-1/\tau} \tag{9.26}$$

where τ is the dependence parameter which belongs to the support $(0,\infty)$.

$$\varrho(u_1,u_2,\tau)=1/u_1^{\tau}+1/u_2^{\tau}-1 \tag{9.27}$$

$$\mathcal{C}_1(u_1,u_2,\tau)=\frac{\partial\mathcal{C}(u_1,u_2,\tau)}{\partial u_1}=\left(u_1^{\tau+1}\varrho(u_1,u_2,\tau)^{1/\tau+1}\right)^{-1} \tag{9.28}$$

$$\mathcal{C}_2(u_1,u_2,\tau)=\frac{\partial\mathcal{C}(u_1,u_2,\tau)}{\partial u_2}=\left(u_2^{\tau+1}\varrho(u_1,u_2,\tau)^{1/\tau+1}\right)^{-1} \tag{9.29}$$

$$\mathcal{C}_\tau(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_1(u_1,u_2,\tau)}{\partial\tau}=\frac{\log\left(\varrho(u_1,u_2,\tau)\right)}{\tau^2\varrho(u_1,u_2,\tau)^{1/\tau}}+\frac{\log(u_1)/u_1^\tau+\log(u_2)/u_2^\tau}{\tau\varrho(u_1,u_2,\tau)^{1/\tau+1}} \tag{9.30}$$

$$\mathcal{C}_{1\tau}(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_1(u_1,u_2,\tau)}{\partial\tau}=\frac{\log(\varrho(u_1,u_2,\tau))}{\tau^2(\varrho(u_1,u_2,\tau))^{1/\tau+1}} \tag{9.31}$$

$$+\frac{(\log(u_1)/u_1^\tau+\log(u_2)/u_2^\tau)(1/\tau+1)}{(\varrho(u_1,u_2,\tau))^{1/\tau+2})u_1^{\tau+1}}-\frac{\log(u_1)}{u_1^{\tau+1}(\varrho(u_1,u_2,\tau))^{1/\tau+1}}$$

$$\mathcal{C}_{11}(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_1(u_1,u_2,\tau)}{\partial u_1}=\frac{\tau(1/\tau+1)u_1^{-2\tau-2}}{(\varrho(u_1,u_2,\tau))^{1/\tau+2}}-\frac{(\tau+1)u_1^{-\tau-2}}{(\varrho(u_1,u_2,\tau))^{1/\tau+1}} \tag{9.32}$$

$$\mathcal{C}_{22}(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_2(u_1,u_2,\tau)}{\partial u_2}=\frac{\tau(1/\tau+1)u_2^{-2\tau-2}}{(\varrho(u_1,u_2,\tau))^{1/\tau+2}}-\frac{(\tau+1)u_2^{-\tau-2}}{(\varrho(u_1,u_2,\tau))^{1/\tau+1}} \tag{9.33}$$

$$\mathcal{C}_{12}(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_1(u_1,u_2,\tau)}{\partial u_2}=\frac{\tau(1/\tau+1)}{u_1^{\tau+1}u_2^{\tau+1}(\varrho(u_1,u_2,\tau))^{1/\tau+2}} \tag{9.34}$$

$$\mathcal{C}_{12\tau}(u_1,u_2,\tau)=\frac{\partial\mathcal{C}_{12}(u_1,u_2,\tau)}{\partial\tau}=\frac{1}{u_1^{\tau+1}u_2^{\tau+1}(\varrho(u_1,u_2,\tau))^{1/\tau+2}} \tag{9.35}$$

$$+\frac{\tau(1/\tau+1)\log(\varrho(u_1,u_2,\tau))}{\tau^2(\varrho(u_1,u_2,\tau))^{1/\tau+2}}+\frac{(\log(u_1)/u_1^\tau+\log(u_2)/u_2^\tau)(1/\tau+2)}{(\varrho(u_1,u_2,\tau))^{1/\tau+3})u_1^{\tau+1}u_2^{\tau+1}}$$

$$-\frac{\tau\log(u_1)(1/\tau+1)}{u_1^{\tau+1}u_2^{\tau+1}(\varrho(u_1,u_2,\tau))^{1/\tau+2}}-\frac{\tau\log(u_2)(1/\tau+1)}{u_1^{\tau+1}u_2^{\tau+1}(\varrho(u_1,u_2,\tau))^{1/\tau+2}}$$

$$(9.36)\quad \mathcal{C}_{121}(u_1,u_2,\tau) = \frac{\partial \mathcal{C}_{12}(u_1,u_2,\tau)}{\partial u_1} = \frac{u_1^{-\tau-2}u_2^{-\tau-1}(1+\tau)}{(\varrho(u_1,u_2,\tau))^{1/\tau+2}}\left[\frac{u_1^{-\tau}(1+2\tau)}{\varrho(u_1,u_2,\tau)} - (1+\tau)\right]$$

$$(9.37)\quad \mathcal{C}_{122}(u_1,u_2,\tau) = \frac{\partial \mathcal{C}_{22}(u_1,u_2,\tau)}{\partial u_2} = \mathcal{C}_{121}(u_2,u_1,\tau)$$

Next, we obtain all of the analytic expressions to be used for formalizing, analytically, the gradient of the log-likelihood function in (7.5) and (7.5).

9.2. First order derivatives

In this section, we obtain analytic expressions to construct the log-likelihood function derivatives in both the continuous selection variable model (presented in section 9.2.1, to follow) as well as in the binary selection variable model (presented in section 9.2.2, to follow). These derivatives are intended to be used when employing one of the various numerical search methods for the parameters values (such as the Trust Region method), in order to reduce the computation burden.

For brevity of notation, let $\mathcal{M}_j \equiv \sum_{k=0}^{K_{n,\xi_j}} a_{k,\boldsymbol{\xi_j}}(e/\sigma_{\xi_j})^k$ and $\mathbf{a}_{\boldsymbol{\xi}_\mathbf{j}} \equiv \left[a_{0,\boldsymbol{\xi_j}}, ..., a_{k,\boldsymbol{\xi_j}}\right]^T$ be a vector of size $(K_{n,\xi_j}+1)\times 1$.

$$\mathcal{M}'_{j\mathbf{a}} \equiv \left[\frac{\partial \mathcal{M}_j}{\partial \mathbf{a}_{\boldsymbol{\xi}_\mathbf{j}}}\right]_{(K_{n,\xi_j}+1)\times 1} = \left[1, (e/\sigma_{\xi_j})^1, \ldots, (e/\sigma_{\xi_j})^{K_{n,\xi_j}}\right]^T$$

$$\mathcal{M}'_{j\sigma} \equiv \left[\frac{\partial \mathcal{M}_j}{\partial \sigma_{\xi_j}}\right]_{1\times 1} = -\sum_{k=0}^{K_{n,\xi_j}} ke/\sigma_{\xi_j}^2 a_{k,\boldsymbol{\xi_j}}(e/\sigma_{\xi_j})^{k-1}$$

$$\mathcal{M}'_{je} \equiv \left[\frac{\partial \mathcal{M}_j}{\partial e}\right]_{1\times 1} = \sum_{k=0}^{K_{n,\xi_j}} k/\sigma_{\xi_j} a_{k,\boldsymbol{\xi_j}}(e/\sigma_{\xi_j})^{k-1}$$

We define the following expressions:

$$(9.38)\quad v_{ji} \equiv c_j - \boldsymbol{w'_{ji}\theta_j},$$

and

$$(9.39)\quad e_{ji} \equiv y_{ji} - \boldsymbol{w'_{ji}\theta_j}.$$

The following expression is essential for the gradient derivation:

$$(9.40)\quad \Psi_{\mathrm{R}}(v_{1i},v_{2i}) \equiv \frac{\mathcal{C}'_{\mathrm{R}}(F_{\xi_j}(v_{1i}), F_{\xi_2}(v_{2i}),\tau)}{1+\mathcal{C}(F_{\xi_j}(v_{1i}), F_{\xi_2}(v_{2i}),\tau) - F_{\xi_1}(v_{1i}) - F_{\xi_2}(v_{2i})},$$

where R $\in \{1,2,\tau\}$, $\mathcal{C}'_1 \equiv \frac{\partial \mathcal{C}}{\partial u_1}$, $\mathcal{C}'_2 \equiv \frac{\partial \mathcal{C}}{\partial u_2}$ and $\mathcal{C}'_\tau \equiv \frac{\partial \mathcal{C}}{\partial \tau}$.

$$\psi_{\mathrm{R}}(v_{1i}, v_{2i}) \equiv \frac{c'_{\mathrm{R}}(F_{\xi_1}(e_{1i}), F_{\xi_2}(e_{2i}), \tau)}{c(F_{\xi_j}(e_{ji}), F_{\xi_2}(e_{2i}), \tau)}, \tag{9.41}$$

where R $\in \{1,2,\tau\}$, $c'_1 \equiv \frac{\partial c}{\partial u_1}$, $c'_2 \equiv \frac{\partial c}{\partial u_2}$ and $c'_\tau \equiv \frac{\partial c}{\partial \tau}$.

$$\varphi_{\mathrm{R}}(e_{1i}, v_{2i}) = \frac{\mathcal{C}'_{1\mathrm{R}}(F_{\xi_1}(e_{1i}), F_{\xi_2}(e_{2i}), \tau)}{1 - \mathcal{C}'_1(F_{\xi_j}(e_{ji}), F_{\xi_2}(e_{2i}), \tau)} \tag{9.42}$$

where R $\in \{1,2,\tau\}$, $c'_{11} \equiv \frac{\partial^2 c}{\partial u_1 \partial u_1}$, $c'_{12} \equiv \frac{\partial^2 c}{\partial u_1 \partial u_2}$ and $c'_{1\tau} \equiv \frac{\partial^2 c}{\partial u_1 \partial \tau}$.

Denote $\mathcal{H}_{K_{n,\xi_j}}$ a symmetric matrix of size $(K_{n,\xi_j}+1) \times (K_{n,\xi_j}+1)$, with elements defined as follows:

$$\hbar_{i,j} = 1\left\{i+j-2 \text{ is } \textit{even}\right\}(i+j-2)!!, \;\; \forall 1 \leq i,j \leq K_{n,\xi_j}+1 \tag{9.43}$$

where for an even integer n $n!! = \prod_{k=0}^{n/2}(2k)$ is referred to as the double factorial of n.

Denote $\mathcal{T}^e_{K_{n,\xi_j}}$ a matrix of size $(K_{n,\xi_j}+1) \times (K_{n,\xi_j}+1)$, with elements which are functions of e, defined as follows:

$$\mathcal{T}_{i,j} = b_{i+j-1}, \;\; 1 \leq i,j \leq K_{n,\xi_j}+1 \tag{9.44}$$

where $b_1 = \Phi(e/\sigma_{\xi_j})$, $b_2 = -\phi(e/\sigma_{\xi_j})$, such that for all $3 \leq k \leq 2(K_{n,\xi_j}+1)-1$ it must hold that $b_k = -\phi(e/\sigma_{\xi_j})(e/\sigma_{\xi_j})^{k-2} + (k-2)b_{k-2}$.

A close form representation of the approximated density function (using Hermite polynomial) is:

$$f_{\xi_j}(e) = \frac{M_j^2 1/\sigma_{\xi_j}\phi(e/\sigma_{\xi_j})}{\mathrm{tr}\left(\mathbf{a_{\xi_j}}\mathbf{a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\right)} = \frac{M_j^2 1/\sigma_{\xi_j}\phi(e/\sigma_{\xi_j})}{\mathbf{a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}} \tag{9.45}$$

where the operator tr(.) is the trace (the sum of the main diagonal elements) of the given matrix.

Similar to (9.45), a close form representation of the approximated distribution function is:

$$F_{\xi_j}(e) = \frac{\mathrm{tr}\left(\mathbf{a_{\xi_j}}\mathbf{a_{\xi_j}}'\mathcal{T}^e_{K_{n,\xi_j}}\right)}{\mathrm{tr}\left(\mathbf{a_{\xi_j}}\mathbf{a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\right)} = \frac{\mathbf{a'_{\xi_j}}\mathcal{T}^e_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}}{\mathbf{a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}} \tag{9.46}$$

The denominator in (9.45) is a close form solution of the following integral (see Schwiebert (2013)):

$$\mathrm{tr}\left(\mathbf{a_{\xi_j}}\mathbf{a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\right) = \int_{-\infty}^{\infty}\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\boldsymbol{\eta}}(v/\sigma_\eta)^k\right]^2 \phi(v/\sigma_\eta)/\sigma_\eta dv \tag{9.47}$$

Additionally, the numerator in (9.46) is a close form solution of the following integral (see Schwiebert (2013)):

$$\text{(9.48)}\qquad \mathrm{tr}\left(\mathbf{a_{\xi_j}a'_{\xi_j}}\mathcal{T}^{e}_{K_{n,\xi_j}}\right) = \int_{-\infty}^{e}\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\boldsymbol{\eta}}(v/\sigma_\eta)^k\right]^2 \phi(v/\sigma_\eta)/\sigma_\eta dv$$

Based on the analytic closed from representation in (9.45),the analytic first order derivatives in (9.49) and (9.50) are:

$$\text{(9.49)}\qquad f'^{e}_{\xi_j}(e) \equiv \left[\frac{\partial f_{\xi_j}(e)}{\partial \xi_j}\right]_{1\times 1} = \frac{1/\sigma_{\xi_j}\left(2\mathcal{M}_j\mathcal{M}'_{je} - \mathcal{M}_j^2 e/\sigma^2_{\xi_j}\right)\phi(e/\sigma_{\xi_j})}{\mathrm{tr}\left(\mathbf{a_{\xi_j}a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\right)}$$

$$\text{(9.50)}\qquad f'^{\sigma}_{\xi_j}(e) \equiv \left[\frac{\partial f_{\xi_j}(e)}{\partial \sigma_{\xi_j}}\right]_{1\times 1} = \frac{1/\sigma_{\xi_j}\left(2\mathcal{M}_j\mathcal{M}'_{j\sigma} + \mathcal{M}_j^2 e^2/\sigma^3_{\xi_j} - \mathcal{M}_j^2/\sigma_{\xi_j}\right)\phi(e/\sigma_{\xi_j})}{\mathrm{tr}\left(\mathbf{a_{\xi_j}a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\right)}$$

$$\text{(9.51)}\qquad f'^{\mathbf{a}}_{\xi_j} \equiv \left[\frac{\partial f_{\xi_j}(e)}{\partial \mathbf{a_{\xi_j}}}\right]_{(K_{n,\xi_j}+1)\times 1} = \frac{\phi(e/\sigma_{\xi_j})}{\sigma_{\xi_j}}\left[\frac{\mathcal{M}'_j\mathbf{a}}{\mathbf{a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}} - \frac{2\mathcal{M}_j\mathcal{H}_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}}{\left(\mathbf{a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}\right)^2}\right]$$

$$\text{(9.52)}\qquad F'^{\mathbf{a}}_{\xi_j} \equiv \left[\frac{\partial F_{\xi_j}(e)}{\partial \mathbf{a_{\xi_j}}}\right]_{(K_{n,\xi_j}+1)\times 1} = \frac{2\mathcal{T}^{e}_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}}{\mathbf{a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}} - \frac{2\left(\mathbf{a'_{\xi_j}}\mathcal{T}^{e}_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}\right)\mathcal{H}_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}}{\left(\mathbf{a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}\right)^2}$$

$$\text{(9.53)}\qquad F'^{\sigma}_{\xi_j} \equiv \left[\frac{\partial F_{\xi_j}(e)}{\partial \sigma_{\xi_j}}\right]_{1\times 1} = \frac{\mathbf{a'_{\xi_j}}\dfrac{\partial \mathcal{T}^{e}_{K_{n,\xi_j}}}{\partial \sigma_{\xi_j}}\mathbf{a_{\xi_j}}}{\mathbf{a'_{\xi_j}}\mathcal{H}_{K_{n,\xi_j}}\mathbf{a_{\xi_j}}}$$

For a symmetric matrix, $\mathbf{B}$ of size $(K_{n,\xi_j}+1)\times(K_{n,\xi_j}+1)$, the following result is obtained:

$$\text{(9.54)}\qquad \mathrm{tr}\left(\mathbf{a_{\xi_j}a'_{\xi_j}B}\right) = \mathbf{a'_{\xi_j}Ba_{\xi_j}}$$

$$\text{(9.55)}\qquad \frac{\partial \mathrm{tr}\left(\mathbf{a_{\xi_j}a'_{\xi_j}B}\right)}{\partial \mathbf{a_{\xi_j}}} = 2\mathbf{Ba_{\xi_j}}$$

Next, each of the semiparametric log-likelihood function's derivatives in the continuous model is presented,[44] in order to construct the gradient vector, which consists of these derivatives.

9.2.1. The continuous model specification

The truncated selection model's first order conditions are specified in (9.56)-(9.59) to follow, where $j \in \{1, 2\}$ indicates the equation's number.

$$\frac{\partial \log L}{\partial \boldsymbol{\theta_j}} = \sum_{i=1}^{n} \left[\frac{f'^{e}_{\xi_j}(e_{ji})}{f_{\xi_j}(e_{ji})} + f_{\xi_j}(e_{ji})\psi_j(e_{1i}, e_{2i}) - \Psi_j(v_{1i}, v_{2i})(1 - f_{\xi_j}(v_{ji})) \right] (-\boldsymbol{w_{ji}}) \tag{9.56}$$

$$\frac{\partial \log L}{\partial \mathbf{a}_{\boldsymbol{\xi}_\mathbf{j}}} = \sum_{i=1}^{n} \frac{f'^{\mathbf{a}}_{\xi_j}(e_{ji})}{f_{\xi_j}(e_{ji})} + \sum_{i=1}^{n} F'^{\mathbf{a}}_{\xi_j}(e_{ji})\psi_j(e_{1i}, e_{2i}) - \sum_{i=1}^{n} \Psi_j(v_{1i}, v_{2i})(1 - F'^{\mathbf{a}}_{\xi_j}(v_{ji})) \tag{9.57}$$

$$\frac{\partial \log L}{\partial \sigma_{\xi_j}} = \sum_{i=1}^{n} \frac{f'^{\sigma}_{\xi_j}(e_{ji})}{f_{\xi_j}(e_{ji})} + \sum_{i=1}^{n} F'^{\sigma}_{\xi_j}(e_{ji})\psi_j(e_{1i}, e_{2i}) - \sum_{i=1}^{n} \Psi_j(v_{1i}, v_{2i})(1 - F'^{\sigma}_{\xi_j}(v_{ji})) \tag{9.58}$$

$$\frac{\partial \log L}{\partial \tau} = \sum_{i=1}^{n} \psi_\tau(e_{1i}, e_{2i}) - \sum_{i=1}^{n} \Psi_\tau(v_{1i}, v_{2i}, \tau) \tag{9.59}$$

In the continuous truncated selection model's specification, the general structure of the first order conditions is the same for the selection equation and for the substantive equation.

Next, each of the semiparametric log-likelihood function's derivatives in the binary model is presented,[45] in order to construct the gradient vector, which consists of these derivatives. We show that in the binary truncated selection model's specification, there is a different structure for the substantive and selection equations' first order conditions, as presented in equations (9.60)-(9.66) to follow.

9.2.2. The binary model specification

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{f'^{e}_{\xi_1}(e_{1i})}{f_{\xi_1}(e_{1i})} - f_{\xi_1}(e_{1i})\varphi_1(e_{1i}, v_{2i}) - \Psi_1(v_{1i}, v_{2i})(1 - f_{\xi_1}(v_{1i})) \right] (-\boldsymbol{x_i}) \tag{9.60}$$

[44] The analytic gradient is computed for a general Copula function in order to guarantee that model estimates can be obtained under different Copula specifications.

[45] The analytic gradient is computed for a general Copula function in order to guarantee that model estimates can be obtained under different Copula specifications.

$$\frac{\partial \log L}{\partial \gamma} = \sum_{i=1}^{n} \left[-f_{\xi_2}(e_{2i})\varphi_2(e_{1i}, v_{2i}) - \Psi_2(v_{1i}, v_{2i})(1 - f_{\xi_2}(v_{2i}))\right](-\boldsymbol{z_i}) \tag{9.61}$$

$$\frac{\partial \log L}{\partial \mathbf{a}_{\boldsymbol{\xi_1}}} = \sum_{i=1}^{n} \frac{f'^{\mathbf{a}}_{\xi_1}(e_{1i})}{f_{\xi_1}(e_{1i})} - \sum_{i=1}^{n} F'^{\mathbf{a}}_{\xi_1}(e_{1i})\varphi_1(e_{1i}, v_{2i}) - \sum_{i=1}^{n} \Psi_1(v_{1i}, v_{2i})(1 - F'^{\mathbf{a}}_{\xi_1}(v_{1i})) \tag{9.62}$$

$$\frac{\partial \log L}{\partial \mathbf{a}_{\boldsymbol{\xi_2}}} = -\sum_{i=1}^{n} F'^{\mathbf{a}}_{\xi_2}(v_{2i})\varphi_2(e_{1i}, v_{2i}) - \sum_{i=1}^{n} \Psi_2(v_{1i}, v_{2i})(1 - F'^{\mathbf{a}}_{\xi_2}(v_{2i})) \tag{9.63}$$

$$\frac{\partial \log L}{\partial \sigma_{\xi_1}} = \sum_{i=1}^{n} \frac{f'^{\sigma}_{\xi_1}(e_{1i})}{f_{\xi_1}(e_{1i})} - \sum_{i=1}^{n} F'^{\sigma}_{\xi_1}(e_{1i})\varphi_1(e_{1i}, v_{2i}) - \sum_{i=1}^{n} \Psi_1(v_{1i}, v_{2i})(1 - F'^{\sigma}_{\xi_1}(v_{1i})) \tag{9.64}$$

$$\frac{\partial \log L}{\partial \sigma_{\xi_2}} = -\sum_{i=1}^{n} F'^{\sigma}_{\xi_2}(v_{2i})\varphi_2(e_{1i}, v_{2i}) - \sum_{i=1}^{n} \Psi_2(v_{1i}, v_{2i})(1 - F'^{\sigma}_{\xi_2}(v_{2i})) \tag{9.65}$$

$$\frac{\partial \log L}{\partial \tau} = -\sum_{i=1}^{n} \varphi_\tau(e_{1i}, v_{2i}) - \sum_{i=1}^{n} \Psi_\tau(v_{1i}, v_{2i}, \tau) \tag{9.66}$$

References

Ai, C., Linton, O., and Zhang, Z. (2018). A simple and efficient estimation method for models with nonignorable missing data. *arXiv preprint arXiv:1801.04202*.

Andrews, G. E. and Askey, R. (1985). Classical orthogonal polynomials. In *Polynômes orthogonaux et applications*, pages 36–62. Springer.

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.

Breunig, C., Mammen, E., and Simoni, A. (2018). Nonparametric estimation in case of endogenous selection. *Journal of Econometrics*. Forthcoming.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, 66(3):429–436.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632.

Chen, X., Fan, Y., and Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association*, 101(475):1228–1240.

Chen, X. and Qiu, Y. J. J. (2016). Methods for nonparametric and semiparametric regressions with endogeneity: A gentle guide. *Annual review of economics*, 8:259–290.

Coppejans, M. (2001). Estimation of the binary response model using a mixture of distributions estimator (mod). *Journal of Econometrics*, 102(2):231–269.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, 39:829–844.

Érdelyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. (1953). *Higher Transcendental Functions, Vol. 2.* McGraw Hill, New York.

Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica: Journal of the Econometric Society*, pages 363–390.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009). Covariate shift and local learning by distribution matching. In *Dataset Shift in Machine Learning*, pages 131–160. MIT Press, Cambridge, MA, USA.

Honoré, 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.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71–120.

Ichimura, H. and Lee, L. F. (1991). Semiparametric least squares estimation of multiple index models: single equation estimation. In *Nonparametric and semiparametric methods in econometrics and statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics. Cambridge*, pages 3–49.

Khan, S. and Lewbel, A. (2007). Weighted and two-stage least squares estimation of semiparametric truncated regression models. *Econometric Theory*, 23(02):309–347.

Lee, M. J. (1993). Quadratic mode regression. *Journal of Econometrics*, 57(1):1–19.

Lee, M. J. and Kim, H. (1998). Semiparametric econometric estimators for a truncated regression model: A review with an extension. *Statistica Neerlandica*, 52(2):200–225.

Lewbel, A. (2008). Semiparametric double index model identification and estimation. Technical report, Citeseer.

Lewbel, A. and Linton, O. (2002). Nonparametric censored and truncated regression. *Econometrica*, 70(2):765–779.

Lewbel, A. and Schennach, S. M. (2007). A simple ordered data estimator for inverse density weighted expectations. *Journal of Econometrics*, 136(1):189–211.

Powell, J. L. (1986). Symmetrically trimmed least squares estimation for tobit models. *Econometrica: journal of the Econometric Society*, pages 1435–1460.

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

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of econometrics*, 2:881–935.

Schwiebert, J. (2013). Sieve maximum likelihood estimation of a copula-based sample selection model. Mimeo, available at: http://www.iza.org/conference files/SUMS 2013/schwiebert j8731.pdf.

Schwiebert, J. (2016). Evidence on copula-based double-hurdle models with flexible margins. *Empirical Economics*, 51(1):245–289.

Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

Tsui, K. L., Jewell, N. P., and Wu, C. (1988). A nonparametric approach to the truncated regression problem. *Journal of the American Statistical Association*, 83(403):785–792.