Sharp Bounds on the Distribution of the Treatment Effect in Switching Regimes Models*

Yanqin Fan and Jisong Wu
Department of Economics
Vanderbilt University
VU Station B #351819
2301 Vanderbilt Place
Nashville, TN 37235-1819

This version: July 2007

Abstract

In this paper, we establish sharp bounds on the distribution of the treatment effect in switching regimes models or generalized sample selection models in Heckman (1990). These bounds depend on the identified model parameters only and hence are themselves identified. Their estimation is straightforward once the identified model parameters are estimated. We compare our bounds when the identified bivariate marginal distributions are either both normal or both Student's t with those assuming trivariate normal or trivariate Student's t distribution, where the latter bounds follow from existing sharp bounds on the correlation between the outcome errors. To illustrate the usefulness of the distribution bounds established in this paper, we apply them to a wage earnings model for child laborers in the early 1900s, where regimes are governed according to literacy.

Keywords: Average treatment effect; treatment effect for the treated; copula; Frechet-Hoeffding inequality; correlation bounds

JEL codes: C31, C35, C14

*We thank Murray Smith for providing us the data set used in the empirical application in this paper, and Bo Honore for helpful discussions. Fan gratefully acknowledges financial support from the National Science Foundation.

1 Introduction

The class of switching regimes models (SRMs) or generalized sample selection models extends the Roy model of self-selection by allowing a more general decision rule for selecting into different states. The income maximizing Roy model of self-selection was developed to explain occupational choice and its consequences for the distribution of earnings when individuals differ in their endowments of occupation-specific skills. Heckman and Honore (1990) demonstrated that the identification of the joint distribution of potential outcomes is essential to the empirical content of the Roy model.

By allowing a more general decision/selection rule, SRMs enjoy a much wider scope of application than the Roy model, but in any particular application, they are also limited in their ability to address a wide range of interesting economic questions because of the non-identifiability of the joint distribution of potential outcomes in SRMs. Even in the 'textbook' Gaussian SRM, the correlation coefficient between the potential outcomes or equivalently the joint distribution of the potential outcomes is not identifiable. When used to study treatment effect, important distributional aspects of the treatment effect other than its mean are not identified in SRMs. This partly explains why the current literature has focussed on various measures of average treatment effect including the average treatment effect (ATE), the treatment effect for the treated (TT), the local average treatment effect (LATE), and the marginal treatment effect (MTE). Heckman, Tobias, and Vytlacil (2003) derived expressions for these four average treatment effect parameters for a Gaussian copula SRM and a Student's t copula SRM with normal outcome errors and non-normal selection errors[1]. Heckman and Vytlacil (2005), among other things, showed that in a latent variable framework, ATE, TT, and LATE can be expressed in terms of MTE.

Recently two approaches have been proposed to deal with the non-identifiability problem of the joint distribution of potential outcomes in the 'textbook' Gaussian SRM and some of its extensions. By employing the positive semidefiniteness of the covariance matrix of the outcome errors and the selection error, Vijverberg (1993) showed that in the 'textbook' Gaussian SRM, although unidentified, useful bounds can be placed on the correlation coefficient between the potential outcomes. Koop and Poirier (1997), Poirier (1998), and Poirier and Tobias (2003) demonstrated via Bayesian approach that these bounds allow learning on the unidentified correlation to take place through the identified correlation coefficients. Since the joint distribution of the potential outcomes in the 'textbook' Gaussian SRM depends on the unidentified correlation coefficient only (besides the identified marginal parameters), learning is possible on the joint distribution of the potential outcomes and on the distribution of the difference between the potential outcomes, see Poirier and Tobias (2003) for details. In the second approach, restrictions are imposed on the dependence structure between the potential outcomes such that their joint distribution and the distribution of the treatment effect

[1]They didn't use the concept of copulas, but their models can be interpreted this way.

are identified, see, e.g., Heckman, Smith, and Clements (1997), Biddle, Boden, and Reville (2003), Carneiro, Hansen, and Heckman (2003), Aakvik, Heckman, and Vytlacil (2003), among others.

The work by Vijverberg (1993), Koop and Poirier (1997), Poirier (1998), and Poirier and Tobias (2003) provide a useful alternative approach to addressing the non-identifiability problem of the joint distribution of the potential outcomes in a SRM which may be used to address important questions in economics where self-selection is present. However, the applicability of this approach is limited by its dependence on the trivariate normality assumption or more generally on the assumption that the trivariate distribution of the errors is solely determined by the unidentified correlation such as those in Li, Poirier, and Tobias (2004). The fact that the joint distribution of the potential outcomes in a SRM is not identifiable and thus can never be verified empirically calls for (i) a study of the robustness of this approach to the implied joint distribution of the potential outcomes and (ii) the development of a general approach to bounding the joint distribution of the potential outcomes and the distribution of the treatment effect that is robust to distributional assumptions on the outcome errors and the selection error. The contribution of this paper is to accomplish both tasks.

The new tool we employ to establish our distribution bounds is the Fréchet-Hoeffding inequality on copulas. A straightforward application of this inequality allows us to bound the joint distribution of potential outcomes using the bivariate distributions of each outcome error and the selection error, where the latter distributions are known to be identified, see Joe (1997) and Lee (2002). Bounds on the joint distribution for various populations of interest are also developed. To bound the distribution of the treatment effect, we make use of existing results on sharp bounds on functions of two random variables including the four simple arithmetic operations, see Williamson and Downs (1990). For a sum of two random variables, Makarov (1981), Rüschendorf (1982), and Frank, Nelsen, and Schweizer (1987) establish sharp bounds on its distribution, see also Nelsen (1999). These results have been used in Fan and Park (2006) to bound the distribution of the treatment effect and the quantile function of the treatment effect in the context of ideal social experiments where selection is random. Other applications of the Fréchet-Hoeffding inequality include Heckman, Smith, and Clements (1997) in which they bound the variance of the treatment effect under the assumption of random selection; Manski (1997b) in which he established bounds on the mixture of two potential outcomes when the distribution of each outcome is known; and Fan (2005a) in which she provided a systematic study on the estimation and inference on the correlation bounds. Fan (2005b) studied nonparametric estimation and inference on Fréchet-Hoeffding distribution bounds when random samples are available from each marginal distribution.

In SRMs with trivariate normal or Student's t errors, existing sharp bounds on the correlation coefficient between the potential outcomes imply sharp bounds on their joint distribution and the distribution of the treatment effect. Interestingly, we find that the sharp bounds on the joint

distribution of potential outcomes in SRMs with trivariate normal or Student's t errors are robust to the implied joint distribution of the potential outcomes (normal or Student's t) in the sense that these bounds remain valid for any distribution of the trio of errors as long as the implied bivariate distributions for each outcome error and the selection error are either bivariate normal or Student's t. In contrast, the sharp bounds on the treatment effect distribution in SRMs with trivariate normal or Student's t errors are not robust to the implied joint distribution of the potential outcomes. We establish the sharp bounds without relying on any assumption on the joint distribution of the trio of errors and provide a detailed numerical comparison between sharp bounds on the treatment effect distribution relying on the trivariate normal or Student's t distribution with those that do not specify the non-refutable joint distribution of potential outcomes. Our numerical results show that bounds relying on the trivariate normal or Student's t assumption can be misleading.

The sharp bounds on the treatment effect distributions developed in this paper allow us to go beyond simple average treatment effects. As an example, we demonstrate via both synthetic and real data how they can be used to investigate the minimum or maximum probability that a person in a subpopulation would benefit from the treatment by participating in it.

The rest of this paper is organized as follows. In Section 2, we extend existing work on correlation bounds in SRMs with trivariate normal or Student's t errors by establishing sharp bounds on the joint distribution of potential outcomes and on the distribution of the treatment effect for the whole population and various subpopulations in these models. In Section 3, we first present existing results on sharp bounds on the difference between two random variables and then use these results to establish sharp bounds on the joint distribution of potential outcomes and on the distribution of the treatment effect for the whole population and various subpopulations in SRMs without restricting the joint distribution of the potential outcomes. In Section 4 we provide a systematic comparison of the two sets of bounds when the two identified bivariate marginal distributions in the SRM are respectively normal and Student's t. Section 5 presents an empirical application of our bounds to evaluating the effect of literacy on weekly wages of child laborers. The last section concludes. Some technical proofs are relegated in Appendix A. In the main text, we focus on the treatment effect distribution corresponding to ATE and TT respectively. Results on the treatment effect distribution corresponding to LATE and MTE are provided in Appendix B.

2 Distribution Bounds in SRMs with Trivariate Gaussian and Student's t Errors

Consider the following SRM:

$$\begin{aligned} Y_{1i} &= X_i'\beta_1 + U_{1i}, \\ Y_{0i} &= X_i'\beta_0 + U_{0i}, \\ D_i &= I_{\{W_i'\gamma+\epsilon_i>0\}},\ i = 1, \ldots, n, \end{aligned} \tag{1}$$

where $\{X_i, W_i\}$ denote individual i's observed covariates and $\{U_{1i}, U_{0i}, \epsilon_i\}$ individual i's unobserved covariates. Here, D_i is the binary variable indicating participation of individual i in the program or treatment; it takes the value 1 if individual i participates in the program and takes the value zero if she chooses not to participate in the program, Y_{1i} is the outcome of individual i we observe if she participates in the program, and Y_{0i} is her outcome if she chooses not to participate in the program. For individual i, we always observe the covariates $\{X_i, W_i\}$, but observe Y_{1i} if $D_i = 1$ and Y_{0i} if $D_i = 0$. The errors or unobserved covariates $\{U_{1i}, U_{0i}, \epsilon_i\}$ are assumed to be independent of the observed covariates $\{X_i, W_i\}$. We also assume the existence of an exclusion restriction, i.e., there exists at least one element of W_i which is not contained in X_i. Parametric SRMs supplement model (1) by distributional specifications for the errors $\{U_{1i}, U_{0i}, \epsilon_i\}$. The textbook Gaussian model assumes that $\{U_{1i}, U_{0i}, \epsilon_i\}$ is trivariate normal. Other commonly used distributions include the trivariate Student's t and mixtures of normal distributions, see e.g., Li, Poirier, and Tobias (2004).

Since either Y_{1i} or Y_{0i} is observed for any given individual i but never both, the joint distribution of U_{1i} and U_{0i} is not identified in a SRM even with parametric distributional assumptions on $\{U_{1i}, U_{0i}, \epsilon_i\}$ such as normality. For example, consider the Gaussian SRM in which $\{U_{1i}, U_{0i}, \epsilon_i\}$ follows a trivariate normal distribution:

$$\begin{pmatrix} U_{1i} \\ U_{0i} \\ \epsilon_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_0\rho_{10} & \sigma_1\rho_{1\epsilon} \\ \sigma_1\sigma_0\rho_{10} & \sigma_0^2 & \sigma_0\rho_{0\epsilon} \\ \sigma_1\rho_{1\epsilon} & \sigma_0\rho_{0\epsilon} & 1 \end{pmatrix} \right]. \tag{2}$$

Based on the sample information alone, ρ_{10} is not identified and all the other parameters are identified. As a result, the marginal distributions of U_{1i} and U_{0i} are identified, but their joint distribution is not. One member of the Gaussian SRM in which ρ_{10} is identified is the well-known Roy model, in which the additional information is provided by the relation between the selection error and the outcome errors: $\epsilon_i = U_{1i} - U_{0i}$.

Heckman, Tobias, and Vytlacil (2003) derived expressions for four treatment parameters of interest for a Gaussian copula model and a Student's t copula model with normal outcome errors and non-normal selection errors. They are respectively ATE, TT, LATE, and MTE. Let $\Delta_i = Y_{1i} - Y_{0i}$ denote the gain from program participation for individual i. Then the average treatment effect

conditional[2] on X_i, W_i is given by

$$ATE_i \equiv E\left(\Delta_i | X_i, W_i\right) = X_i'(\beta_1 - \beta_0).$$

The effect of the treatment on the treated is the effect from treatment for those that actually select into the treatment. In the Gaussian SRM,

$$\begin{aligned} TT_i &\equiv E\left(\Delta_i | X_i, W_i, D_i = 1\right) = ATE_i + E\left(U_{1i} - U_{0i} | \epsilon_i > -W_i'\gamma\right) \\ &= ATE_i + \left(\rho_{1\epsilon}\sigma_1 - \rho_{0\epsilon}\sigma_0\right)\lambda\left(W_i'\gamma\right), \end{aligned}$$

where $\lambda(\cdot)$ is the inverse mills ratio. Since both ATE_i and TT_i depend on the identified parameters only, they themselves are identified. Likewise, $LATE_i$ and MTE_i are also identified, see Appendix B for their expressions.

The distribution of Δ_i, however, depends on ρ_{10} and hence is not identified: (1) implies that the individual i's treatment effect is given by

$$\Delta_i = ATE_i + (U_{1i} - U_{0i}).$$

As a result, the individual treatment effect Δ_i may differ across individuals with the same ATE_i or TT_i because of the unobserved heterogeneity $(U_{1i} - U_{0i})$. This motivates the study of the distribution of treatment effect Δ_i.

In the rest of this section, we first characterize the class of joint distributions of $\{Y_{1i}, Y_{0i}\}$ and the distributions of Δ_i in the Gaussian SRM consistent with the empirical evidence. Then we discuss extensions to the SRM with Student's t errors.

2.1 Sharp Bounds on the Joint Distribution of Potential Outcomes

For notational compactness, we omit the subscript i in the rest of Section 2. Let F_{10}^Y denote the joint distribution of potential outcomes Y_1, Y_0 conditional on $X = x$. To simplify the notation, we keep the conditioning on the regressors implicit unless it serves to clarify the exposition otherwise. In a Gaussian SRM,

$$F_{10}^Y(y_1, y_0) = \Phi_{\rho_{10}}\left(\frac{y_1 - x'\beta_1}{\sigma_1}, \frac{y_0 - x'\beta_0}{\sigma_0}\right),$$

where $\Phi_\rho(\cdot,\cdot)$ is the distribution function of a bivariate normal variable with zero means, unit variances, and correlation coefficient ρ. Vijverberg (1993), Koop and Poirier (1997), Poirier (1998), and Poirier and Tobias (2003) point out that learning can take place about ρ_{10} through the restriction that the covariance matrix of the errors is positive semi-definite. This restriction places

[2] We focus on conditional treatment effects, as the unconditional treatment effects can be obtained from averaging the corresponding conditional treatment effects over observations in the sample.

bounds on the unidentified correlation between the potential outcomes given what we learn about the correlation between the outcome and selection errors. Specifically, let

$$\rho_L = \rho_{1\epsilon}\rho_{0\epsilon} - \sqrt{(1-\rho_{1\epsilon}^2)(1-\rho_{0\epsilon}^2)},\ \rho_U = \rho_{1\epsilon}\rho_{0\epsilon} + \sqrt{(1-\rho_{1\epsilon}^2)(1-\rho_{0\epsilon}^2)}.$$

Vijverberg (1993) showed that the positive semi-definiteness of the covariance matrix of the errors $\{U_{1i}, U_{0i}, \epsilon_i\}$ implies

$$\rho_L \leq \rho_{10} \leq \rho_U. \tag{3}$$

Note that ρ_L and ρ_U depend on the identified parameters only and hence are themselves identified. The following lemma follows from simple algebra.

Lemma 2.1 *(i) If $\rho_{1\epsilon} = \rho_{0\epsilon} = 0$, then $\rho_L = -1$ and $\rho_U = 1$; (ii) If $\rho_{1\epsilon}^2 + \rho_{0\epsilon}^2 > 1$ and $\rho_{1\epsilon}, \rho_{0\epsilon}$ have the same sign, then $\rho_L > 0$; (iii) If $\rho_{1\epsilon}^2 + \rho_{0\epsilon}^2 > 1$ and $\rho_{1\epsilon}, \rho_{0\epsilon}$ have the opposite sign, then $\rho_U < 0$.*

Lemma 2.1 (i) implies that the bounds ρ_L, ρ_U are informative as long as at least one of the potential outcomes is correlated with the selection error; otherwise, no learning takes place about ρ_{10}. In addition, Lemma 2.1 (ii) implies that it is possible to identify the sign of ρ_{10}. (3) characterizes the class of Gaussian SRMs consistent with the sample information; any Gaussian SRM with ρ_{10} violating (3) is inconsistent with the sample information. It is interesting to observe that the Roy model with $\epsilon = U_1 - U_0$ is consistent with the sample information, since it is a simple algebra to show that $\rho_{10} = \rho_U$ in the Roy model.

Obviously, bounds on ρ_{10} place bounds on the joint distribution $F_{10}^Y(y_1, y_0)$. In terms of the Gaussian copula, we can rewrite the expression for $F_{10}^Y(y_1, y_0)$ as:

$$F_{10}^Y(y_1, y_0) = C^{Gau}\left(\Phi(\frac{y_1 - x\beta_1}{\sigma_1}), \Phi(\frac{y_0 - x\beta_0}{\sigma_0}), \rho_{10}\right),$$

where $\Phi(\cdot)$ is the distribution function of a standard normal random variable and C^{Gau} denotes the Gaussian copula defined as

$$C^{Gau}(u, v, \rho) = \Phi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right),\ (u, v) \in [0, 1]^2.$$

Since the Gaussian copula is increasing in concordance in ρ_{10} (see Joe (1997)), we obtain the following sharp bounds on the joint distribution of Y_1, Y_0:

$$C^{Gau}\left(\Phi(\frac{y_1 - x\beta_1}{\sigma_1}), \Phi(\frac{y_0 - x\beta_0}{\sigma_0}), \rho_L\right) \leq F_{10}^Y(y_1, y_0) \leq C^{Gau}\left(\Phi(\frac{y_1 - x\beta_1}{\sigma_1}), \Phi(\frac{y_0 - x\beta_0}{\sigma_0}), \rho_U\right). \tag{4}$$

For any fixed x and y_1, y_0, the bounds above are informative as long as $\Phi_{\rho_L}\left(\frac{y_1 - x\beta_1}{\sigma_1}, \frac{y_0 - x\beta_0}{\sigma_0}\right) \neq 0$, or $\Phi_{\rho_U}\left(\frac{y_1 - x\beta_1}{\sigma_1}, \frac{y_0 - x\beta_0}{\sigma_0}\right) \neq 1$. (4) characterizes the class of joint distributions $F_{10}^Y(y_1, y_0)$ in Gaussian SRMs consistent with the empirical evidence. One such distribution is the joint distribution of

the potential outcomes in the Roy model, which provides the sharp upper bound on the joint distribution of potential outcomes in all Gaussian SRMs consistent with the sample information.

It is interesting to observe that there are two types of learning on the joint distribution of potential outcomes: (i) learning through the correlation between the outcome and selection errors or through learning on ρ_{10}; (ii) learning through the marginal distributions of the potential outcomes. To understand the second type of learning on $F_{10}^{Y}(y_1, y_0)$, recall the Fréchet-Hoeffding inequality:

$$C_L(s,t) \leq C(s,t) \leq C_U(s,t), \text{ for all } (s,t) \in [0,1]^2,$$

where $C(\cdot,\cdot)$ is any copula function, $C_L(s,t) = \max(s+t-1, 0)$ is the Fréchet lower bound copula, and $C_U(s,t) = \min(s,t)$ is the Fréchet upper bound copula. Now, suppose selection is random so that $\rho_{1\epsilon} = 0$ and $\rho_{0\epsilon} = 0$. In this case, $\rho_L = -1$ and $\rho_U = 1$ so no learning takes place on ρ_{10}. However, since the Gaussian copula equals the Fréchet lower bound copula $C_L(u,v)$ when $\rho_{10} = -1$ and equals the Fréchet upper bound copula when $\rho_{10} = 1$, (4) implies

$$C_L\left(\Phi(\frac{y_1 - x\beta_1}{\sigma_1}), \Phi(\frac{y_0 - x\beta_0}{\sigma_0})\right) \leq F_{10}^{Y}(y_1, y_0) \leq C_U\left(\Phi(\frac{y_1 - x\beta_1}{\sigma_1}), \Phi(\frac{y_0 - x\beta_0}{\sigma_0})\right). \tag{5}$$

For any fixed x and y_1, y_0, the above bounds are informative as long as the lower bound is not zero or the upper bound is not 1. Notice that this is a simple application of the Fréchet-Hoeffding inequality to the joint distribution $F_{10}^{Y}(y_1, y_0)$ without taking into account the bounds on the unidentified correlation coefficient. As a result, in general, the bounds in (4) are sharper than those in (5). That is, taking into account self-selection tightens the bounds.

2.2 Sharp Bounds on the Distribution of the Treatment Effect

We now consider the distribution of the treatment effect $\Delta = Y_1 - Y_0$. Define

$$\gamma_1 = \sigma_1\rho_{1\epsilon} - \sigma_0\rho_{0\epsilon} \text{ and } \gamma_2^2 = \sigma_1^2 + \sigma_0^2 - 2\sigma_1\sigma_0\rho_{10}.$$

Then $\gamma_2^2 = Var(\Delta|X = x)$ satisfies $\sigma_L^2 \leq \gamma_2^2 \leq \sigma_U^2$, where

$$\sigma_U^2 = \sigma_1^2 + \sigma_0^2 - 2\rho_L\sigma_1\sigma_0, \ \sigma_L^2 = \sigma_1^2 + \sigma_0^2 - 2\rho_U\sigma_1\sigma_0.$$

It is worth pointing out that the treatment effect in the Roy model has the smallest variance in all Gaussian SRMs consistent with the sample information, since $\rho_{10} = \rho_U$ in the Roy model.

Poirier and Tobias (2003) showed that the distributions of Δ conditional on $X = x$ corresponding to ATE and TT are given by

$$ATE : F_\Delta(\delta) = \Phi\left(\frac{\delta - ATE}{\gamma_2}\right),$$

$$TT : F_\Delta(\delta|D=1) = \frac{\int_{-\infty}^{\delta} \frac{1}{\gamma_2}\phi\left(\frac{\zeta - ATE}{\gamma_2}\right)\Phi\left(\frac{\gamma_1(\zeta - ATE)/\gamma_2^2 + w'\gamma}{\sqrt{1-\gamma_1^2/\gamma_2^2}}\right)d\zeta}{\Phi(w'\gamma)},$$

where $\phi(\cdot)$ denotes the pdf of the standard normal distribution. Note that while the distribution of Δ corresponding to ATE is normal, the distribution of Δ corresponding to TT is skew-normal unless selection is random in which case $\gamma_1 = 0$. When γ_1 is positive, the distribution of Δ corresponding to TT is right skewed, otherwise it is left skewed. The degree of skewness depends on the value of $w'\gamma$: the smaller $w'\gamma$ is, the larger the skewness of the distribution of Δ corresponding to TT.

Since the only non-identified parameter in both distributions is γ_2, their lower and upper bounds are given respectively by their pointwise minimum and maximum over $\gamma_2 \in [\sigma_L, \sigma_U]$. For $F_\Delta(\delta)$, it can be shown that $F^L_\Delta(\delta) \leq F_\Delta(\delta) \leq F^U_\Delta(\delta)$, where

$$
\begin{aligned}
F^L_\Delta(\delta) &= \begin{cases} \Phi\left(\frac{\delta - ATE}{\sigma_U}\right) & \text{if } \delta \geq ATE \\ \Phi\left(\frac{\delta - ATE}{\sigma_L}\right) & \text{if } \delta < ATE \end{cases}; \\
F^U_\Delta(\delta) &= \begin{cases} \Phi\left(\frac{\delta - ATE}{\sigma_L}\right) & \text{if } \delta \geq ATE \\ \Phi\left(\frac{\delta - ATE}{\sigma_U}\right) & \text{if } \delta < ATE \end{cases}. \qquad (6)
\end{aligned}
$$

For the distribution of Δ corresponding to TT, there is no closed-form expression for its bounds. However, they can be easily computed by numerical optimization algorithms.

Let $F^R_\Delta(\delta)$ denote the distribution of Δ in the Roy model. Then for all δ,

$$F^R_\Delta(\delta) = \Phi\left(\frac{\delta - ATE}{\sigma_U}\right).$$

Since $ATE = x'(\beta_1 - \beta_0)$ in all Gaussian SRMs, (6) implies that to the left of ATE, $F_\Delta(\delta)$ is bounded from above by $F^R_\Delta(\delta)$ and to the right of ATE, $F_\Delta(\delta)$ is bounded from below by $F^R_\Delta(\delta)$. Consequently, the distribution of Δ in the Roy model is a mean preserving spread of the distribution of Δ in all Gaussian SRMs consistent with the sample information and hence second order stochastically dominates the latter distributions.

THEOREM 2.2 *In Gaussian SRMs, F^R_Δ second order stochastically dominates any F_Δ consistent with the sample information.*

For any fixed x and fixed δ, the bounds on $F_\Delta(\delta)$ are informative as long as $F^L_\Delta(\delta) \neq 0$ or $F^U_\Delta(\delta) \neq 1$. In particular, when $\delta = ATE$, $F^L_\Delta(\delta) = F^U_\Delta(\delta) = 0.5$. Hence $F_\Delta(ATE) = 0.5$, implying that the value of the distribution of Δ at the ATE is identified and that the median of the distribution of the outcome gain is the same as ATE.

We note that the two types of learning for the joint distribution occur here as well. When selection is random, learning takes place through the marginals and the bounds are given by

$$
\begin{aligned}
F^L_\Delta(\delta) &= \begin{cases} \Phi\left(\frac{\delta - ATE}{(\sigma_1 + \sigma_0)}\right) & \text{if } \delta \geq ATE \\ \Phi\left(\frac{\delta - ATE}{|\sigma_1 - \sigma_0|}\right) & \text{if } \delta < ATE \end{cases}; \\
F^U_\Delta(\delta) &= \begin{cases} \Phi\left(\frac{\delta - ATE}{|\sigma_1 - \sigma_0|}\right) & \text{if } \delta \geq ATE \\ \Phi\left(\frac{\delta - ATE}{(\sigma_1 + \sigma_0)}\right) & \text{if } \delta < ATE \end{cases}.
\end{aligned}
$$

In general, $(\sigma_1 - \sigma_0)^2 \leq \sigma_L^2 \leq \sigma_U^2 \leq (\sigma_1 + \sigma_0)^2$. Taking into account self-selection tightens the bounds. Moreover, the following simple algebra demonstrates that the stronger the self-selection is, the tighter the bounds. For any δ, the width of the distribution bounds depend on σ_U and σ_L. Noting that

$$\sigma_U^2 - \sigma_L^2 = 4\sigma_1\sigma_0\sqrt{(1-\rho_{1\epsilon}^2)(1-\rho_{0\epsilon}^2)},$$

we conclude that the width of the distribution bounds becomes narrower as the correlation between the selection error and the outcome errors become stronger. In the extreme case where $\rho_{1\epsilon}^2 = 1$ or $\rho_{0\epsilon}^2 = 1$, the lower and upper bounds coincide and the distribution of Δ is identified.

Given the bounds on the distribution of Δ, we get immediately quantile bounds:

$$(F_\Delta^U)^{-1}(q) \leq F_\Delta^{-1}(q) \leq (F_\Delta^L)^{-1}(q),$$

where

$$(F_\Delta^L)^{-1}(q) = \begin{cases} ATE + \sigma_U\Phi^{-1}(q) & \text{if } q \geq 1/2 \\ ATE + \sigma_L\Phi^{-1}(q) & \text{if } q < 1/2 \end{cases};$$

$$(F_\Delta^U)^{-1}(q) = \begin{cases} ATE + \sigma_L\Phi^{-1}(q) & \text{if } q \geq 1/2 \\ ATE + \sigma_U\Phi^{-1}(q) & \text{if } q < 1/2 \end{cases}.$$

2.3 Extensions to Student's t Errors

The results we obtained for Gaussian SRMs can be easily extended to models with Student's t errors. For example, Li, Poirier and Tobias (2004) provides the following expression for the distribution of the treatment effect under the assumption of trivariate Student's t errors:

$$F_\Delta(\delta) = T_{[v]}\left(\frac{\delta - ATE}{\gamma_2}\sqrt{\frac{v}{v-2}}\right),$$

where $T_{[v]}(\cdot)$ denotes the distribution function of the Student's t distribution with v degrees of freedom. Similar to Gaussian models, one can show that $F_\Delta^L(\delta) \leq F_\Delta(\delta) \leq F_\Delta^U(\delta)$, where

$$F_\Delta^L(\delta) = \begin{cases} T_{[v]}\left(\frac{\delta-ATE}{\sigma_U}\sqrt{\frac{v}{v-2}}\right) & \text{if } \delta \geq ATE \\ T_{[v]}\left(\frac{\delta-ATE}{\sigma_L}\sqrt{\frac{v}{v-2}}\right) & \text{if } \delta < ATE \end{cases};$$

$$F_\Delta^U(\delta) = \begin{cases} T_{[v]}\left(\frac{\delta-ATE}{\sigma_L}\sqrt{\frac{v}{v-2}}\right) & \text{if } \delta \geq ATE \\ T_{[v]}\left(\frac{\delta-ATE}{\sigma_U}\sqrt{\frac{v}{v-2}}\right) & \text{if } \delta < ATE \end{cases}.$$

It is obvious that the qualitative conclusions in Gaussian models carry over to the Student's t case. The sharp bounds on the distribution corresponding to TT are given by the minimum and maximum of $F_\Delta(\delta|D=1)$ over $\gamma_2 \in [\sigma_L, \sigma_U]$:

$$F_\Delta(\delta|D=1) = \frac{\int_{-\infty}^{\delta} t_{[v]}\left(\frac{\zeta-ATE}{\gamma_2}\sqrt{\frac{v}{v-2}}\right) T_{[v+1]}\left(\frac{\frac{\gamma_1}{\gamma_2^2}(\zeta-ATE)+w'\gamma}{\Omega}\sqrt{\frac{v+1}{v-1}}\right) d\zeta}{T_{[v]}\left(w'\gamma\sqrt{\frac{v}{v-2}}\right)},$$

where $t_{[v]}(\cdot)$ denotes the pdf of the Student's t distribution with v degrees of freedom and

$$\Omega = \sqrt{\left[v - 2 + \frac{(\delta - ATE)^2}{\gamma_2}\right]\left(\frac{1}{v-1}\right)\left(1 - \frac{\gamma_1^2}{\gamma_2^2}\right)}.$$

3 Bounds in Semiparametric SRMs

The bounds for Gaussian and Student's t SRMs established in Section 2 depend crucially on the parametric distribution assumption, especially the implied joint normal or Student's t distribution of the potential outcomes. Given that the assumption of joint normality or Student's t distribution of the potential outcomes can never be verified empirically, it is important to investigate the robustness of these bounds to the corresponding distributional assumptions and to establish bounds that do not rely on them. This will be accomplished in the current section.

For generality, we adopt Heckman (1990)'s notation and consider the following semiparametric SRM:

$$\begin{aligned} Y_{1i} &= g_1(X_{1i}, X_{ci}) + U_{1i}, \\ Y_{0i} &= g_0(X_{0i}, X_{ci}) + U_{0i}, \\ D_i &= I_{\{(W_i, X_{ci})'\gamma + \epsilon_i > 0\}},\ i = 1, \ldots, n, \end{aligned} \tag{7}$$

where both $g_1(x_1, x_c)$, $g_0(x_0, x_c)$ and the distribution of $\{U_{1i}, U_{0i}, \epsilon_i\}$ are completely unknown. Heckman (1990) provided conditions under which the joint distributions of $\{U_{1i}, \epsilon_i\}$ and $\{U_{0i}, \epsilon_i\}$, $g_1(x_1, x_c)$, $g_0(x_0, x_c)$, and γ are identified from the sample information alone. However, the joint distribution of $\{U_{1i}, U_{0i}\}$ is not identified.

In this section, we provide sharp bounds on the joint distribution of $\{U_{1i}, U_{0i}\}$ or $\{Y_{1i}, Y_{0i}\}$ and the distribution of Δ_i. We assume independence of the errors $\{U_{1i}, U_{0i}, \epsilon_i\}$ and the regressors $\{X_{1i}, X_{0i}, X_{ci}, W_i\}$. The covariance approach used in Section 2 is not applicable here, as the distribution of $\{U_{1i}, U_{0i}, \epsilon_i\}$ is completely unknown. Instead we make use of existing results bounding the distribution of a difference of two random variables each having a given distribution function.

3.1 Sharp Bounds on the Distribution of a Difference of Two Random Variables

Fréchet-Hoeffding Inequality has been used to establish sharp bounds on functions of random variables Y_1 and Y_0 including the four simple arithmetic operations, see Williamson and Downs (1990). For a sum of two random variables, Makarov (1981), Rüschendorf (1982), and Frank, Nelsen, and Schweizer (1987) establish sharp bounds on its distribution, see also Nelsen (1999). Frank, Nelsen, and Schweizer (1987) demonstrate that their proof based on copulas can be extended to more general functions than the sum. These references except Nelsen (1999) make use of the left-continuous version of the distribution function. To avoid any confusion, we will assume that

the random variables Y_1 and Y_0 are continuous with distribution functions F_1 and F_0 respectively. Given that we are interested in the treatment effect, we will present the relevant results for the difference of two random variables. More specifically, let $\Delta = Y_1 - Y_0$ and $F_\Delta(\cdot)$ denote the distribution function of Δ. The following lemma presents sharp bounds on $F_\Delta(\cdot)$ when only F_1 and F_0 are known.

Lemma 3.1 *Let* $F_{\min}(\delta) = \sup_{y_1} \max(F_1(y_1) - F_0(y_1 - \delta), 0)$ *and* $F_{\max}(\delta) = 1 + \inf_{y_1} \min(F_1(y_1) - F_0(y_1 - \delta), 0)$. *Then* $F_{\min}(\delta) \leq F_\Delta(\delta) \leq F_{\max}(\delta)$.

Viewed as an inequality among all possible distribution functions, the sharp bounds $F_{\min}(\delta)$ and $F_{\max}(\delta)$ cannot be improved, because it is easy to show that if either F_1 or F_0 is the degenerate distribution at a finite value, then for all δ, we have $F_{\min}(\delta) = F_\Delta(\delta) = F_{\max}(\delta)$. In fact, given any pair of distribution functions F_1 and F_0, the inequality: $F_{\min}(\delta) \leq F_\Delta(\delta) \leq F_{\max}(\delta)$ cannot be improved, that is, the bounds $F_{\min}(\delta)$ and $F_{\max}(\delta)$ for $F_\Delta(\delta)$ are point-wise best-possible, see Frank, Nelsen, and Schweizer (1987) for a proof of this for a sum of random variables and Williamson and Downs (1990) for a general operation on two random variables. Unlike the sharp bounds on the correlation coefficient between Y_1, Y_0 or the joint distribution of Y_1, Y_0 which are reached at the Fréchet-Hoeffding lower and upper bounds for the distribution of Y_1, Y_0 when Y_1 and Y_0 are perfectly negatively dependent or perfectly positive dependent (see Fan (2005a)), the sharp bounds on the distribution of Δ are not reached at the Fréchet-Hoeffding lower and upper bounds for the distribution of Y_1, Y_0. Frank, Nelsen, and Schweizer (1987) provided explicit expressions for copulas that reach the bounds on the distribution of Δ.

Explicit expressions for bounds on the distribution of a sum of two random variables are available for the case where both random variables have the same distribution which includes the uniform, the normal, the Cauchy, and the exponential families, see Alsina (1981), Frank, Nelsen, and Schweizer (1987), and Denuit, Genest, and Marceau (1999). Below we provide expressions on $F_{\min}(\delta)$ and $F_{\max}(\delta)$ when both Y_1 and Y_0 are normal or Student's t.

Example 3.1. Let $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_0 \sim N(\mu_0, \sigma_0^2)$. Fan and Park (2006) provide the following expressions for the bounds $F_{\min}(\delta)$ and $F_{\max}(\delta)$:

(i) If $\sigma_1 = \sigma_0 = \sigma$, then

$$F_{\min}(\delta) = \begin{cases} 0 & \text{if } \delta < \mu_1 - \mu_0, \\ 2\Phi\left(\frac{\delta - (\mu_1 - \mu_0)}{2\sigma}\right) - 1 & \text{if } \delta \geq \mu_1 - \mu_0, \end{cases} \tag{8}$$

$$F_{\max}(\delta) = \begin{cases} 2\Phi\left(\frac{\delta - (\mu_1 - \mu_0)}{2\sigma}\right) & \text{if } \delta < \mu_1 - \mu_0, \\ 1 & \text{if } \delta \geq \mu_1 - \mu_0. \end{cases} \tag{9}$$

(ii) If $\sigma_1 \neq \sigma_0$, then

$$\begin{aligned} F_{\min}(\delta) &= \Phi\left(\frac{\sigma_1 s - \sigma_0 t}{\sigma_1^2 - \sigma_0^2}\right) + \Phi\left(\frac{\sigma_1 t - \sigma_0 s}{\sigma_1^2 - \sigma_0^2}\right) - 1, \\ F_{\max}(\delta) &= \Phi\left(\frac{\sigma_1 s + \sigma_0 t}{\sigma_1^2 - \sigma_0^2}\right) - \Phi\left(\frac{\sigma_1 t + \sigma_0 s}{\sigma_1^2 - \sigma_0^2}\right) + 1, \end{aligned}$$

where $s = \delta - (\mu_1 - \mu_0)$ and $t = \left(s^2 + 2\left(\sigma_1^2 - \sigma_0^2\right)\ln\left(\frac{\sigma_1}{\sigma_0}\right)\right)^{\frac{1}{2}}$.

Example 3.2. For $j = 0, 1$, we assume $\frac{Y_j - \mu_j}{\sigma_j}\sqrt{\frac{v_j}{v_j - 2}} \sim t_{[v_j]}$, where $v_j > 2$, so that $E(Y_j) = \mu_j$, $Var(Y_j) = \sigma_j^2$ and $F_j(\delta) = T_{[v_j]}\left(\left(\frac{\delta - \mu_j}{\sigma_j}\right)\sqrt{\frac{v_j}{v_j - 2}}\right)$.

By lemma 3.1, $F_{\min}(\delta) = \max(F_1(x_1^*) - F_0(x_1^* - \delta), 0)$ and $F_{\max}(\delta) = 1 + \min(F_1(x_2^*) - F_0(x_2^* - \delta), 0)$, where x_1^* and x_2^* are the maximizer and minimizer of the function $[F_1(x) - F_0(x - \delta)]$ respectively, i.e., x_1^*, x_2^* satisfy the equation:

$$\frac{1}{\sigma_1}\sqrt{\frac{v_1}{v_1 - 2}}t_{[v_1]}\left(\left(\frac{x - \mu_1}{\sigma_1}\right)\sqrt{\frac{v_1}{v_1 - 2}}\right) = \frac{1}{\sigma_0}\sqrt{\frac{v_0}{v_0 - 2}}t_{[v_0]}\left(\left(\frac{x - \mu_0 - \delta}{\sigma_0}\right)\sqrt{\frac{v_0}{v_0 - 2}}\right).$$

In general, one must solve the above equation and hence evaluate $F_{\min}(\delta)$ and $F_{\max}(\delta)$ numerically. But when $v_1 = v_0 \equiv v$ (say), we are able to get closed-form expressions for $F_{\min}(\delta)$ and $F_{\max}(\delta)$ as follows:

(i) If $\sigma_1 = \sigma_0 = \sigma$, then

$$F_{\min}(\delta) = \begin{cases} 0 & \text{if } \delta < \mu_1 - \mu_0, \\ 2T_{[v]}\left(\left(\frac{\delta - (\mu_1 - \mu_0)}{2\sigma}\right)\sqrt{\frac{v}{v-2}}\right) - 1 & \text{if } \delta \geq \mu_1 - \mu_0, \end{cases} \tag{10}$$

$$F_{\max}(\delta) = \begin{cases} 2T_{[v]}\left(\left(\frac{\delta - (\mu_1 - \mu_0)}{2\sigma}\right)\sqrt{\frac{v}{v-2}}\right) & \text{if } \delta < \mu_1 - \mu_0, \\ 1 & \text{if } \delta \geq \mu_1 - \mu_0. \end{cases} \tag{11}$$

(ii) If $\sigma_1 \neq \sigma_0$, then

$$\begin{aligned} & F_{\min}(\delta) \\ = \; & T_{[v]}\left(\left(\frac{\sigma_1^{2\kappa-1}s - \sigma_0^{\kappa}\sigma_1^{\kappa-1}t}{\sigma_1^{2\kappa} - \sigma_0^{2\kappa}}\right)\sqrt{\frac{v}{v-2}}\right) + T_{[v]}\left(\left(\frac{\sigma_1^{\kappa}\sigma_0^{\kappa-1}t - \sigma_0^{(2\kappa-1)}s}{\sigma_1^{2\kappa} - \sigma_0^{2\kappa}}\right)\sqrt{\frac{v}{v-2}}\right) - 1, \\ & F_{\max}(\delta) \\ = \; & T_{[v]}\left(\left(\frac{\sigma_1^{2\kappa-1}s + \sigma_0^{\kappa}\sigma_1^{\kappa-1}t}{\sigma_1^{2\kappa} - \sigma_0^{2\kappa}}\right)\sqrt{\frac{v}{v-2}}\right) - T_{[v]}\left(\left(\frac{\sigma_1^{\kappa}\sigma_0^{\kappa-1}t + \sigma_0^{(2\kappa-1)}s}{\sigma_1^{2\kappa} - \sigma_0^{2\kappa}}\right)\sqrt{\frac{v}{v-2}}\right) + 1, \end{aligned}$$

where $s = \delta - (\mu_1 - \mu_0)$, $\kappa = \frac{v}{v+1}$, and

$$t = \left(s^2 + \left(\sigma_1^{2\kappa} - \sigma_0^{2\kappa}\right)\left(\sigma_1^{2(1-\kappa)} - \sigma_0^{2(1-\kappa)}\right)(v-2)\right)^{\frac{1}{2}}.$$

It is easy to see that in both cases, the expressions for $F_{\min}(\delta)$ and $F_{\max}(\delta)$ reduce to those in Example 3.1 as $v \to +\infty$. For instance, consider the case where $\sigma_1 \neq \sigma_0$. As $v \to +\infty$, we have $\kappa \to 1$, $\sqrt{\frac{v}{v-2}} \to 1$, and $\left(\sigma_1^{2(1-\kappa)} - \sigma_0^{2(1-\kappa)}\right)(v-2) \to 2\log\left(\frac{\sigma_1}{\sigma_0}\right)$.

3.2 Semiparametric SRMs

Let $F_{1\epsilon}(u_1,\epsilon)$ and $F_{0\epsilon}(u_0,\epsilon)$ denote respectively the distribution functions of $\{U_{1i},\epsilon_i\}$ and $\{U_{0i},\epsilon_i\}$ in model (7). Since $F_{1\epsilon}(u_1,\epsilon)$ and $F_{0\epsilon}(u_0,\epsilon)$ are identified from the sample information, the joint distribution of $\{U_{1i},U_{0i},\epsilon_i\}$ belongs to the Frechet class of trivariate distributions for which the (1,3) and (2,3) bivariate margins are given or fixed, denoted as $\mathcal{F}(F_{1\epsilon},F_{0\epsilon})$. Joe (1997) showed that for any $F_{10\epsilon}\in\mathcal{F}(F_{1\epsilon},F_{0\epsilon})$, it must satisfy

$$\int_{-\infty}^{\epsilon} C_L\left[F_{1|\epsilon}(u_1),F_{0|\epsilon}(u_0)\right]dF_\epsilon(\epsilon)\le F_{10\epsilon}(u_1,u_0,\epsilon)\le\int_{-\infty}^{\epsilon} C_U\left[F_{1|\epsilon}(u_1),F_{0|\epsilon}(u_0)\right]dF_\epsilon(\epsilon), \tag{12}$$

where $F_{j|\epsilon}(u_j)$ denote the conditional distribution of U_{ji} given $\epsilon_i=\epsilon$, $j=1,0$ and $F_\epsilon(\epsilon)$ the marginal distribution function of ϵ_i. Inequality (12) follows from the Frechet-Hoeffding inequality and the expression: $F_{10\epsilon}(u_1,u_0,\epsilon)=\int_{-\infty}^{\epsilon}F_{10|\epsilon}(u_1,u_0)dF_\epsilon(\epsilon)$, where $F_{10|\epsilon}(u_1,u_0)$ is the conditional joint distribution of U_{1i},U_{0i} given $\epsilon_i=\epsilon$.

THEOREM 3.2 *In a semiparametric SRM, the following inequalities hold.*

(i) *ATE: The joint distribution of potential outcomes satisfies*

$$F_{10}^L(y_1,y_0)\le F_{10}^Y(y_1,y_0)\le F_{10}^U(y_1,y_0), \tag{13}$$

where

$$\begin{aligned}F_{10}^L(y_1,y_0) &= \int_{-\infty}^{\infty} C_L\left[F_{1|\epsilon}(y_1-g_1(x_1,x_c)),F_{0|\epsilon}(y_0-g_0(x_0,x_c))\right]dF_\epsilon(\epsilon),\\ F_{10}^U(y_1,y_0) &= \int_{-\infty}^{\infty} C_U\left[F_{1|\epsilon}(y_1-g_1(x_1,x_c)),F_{0|\epsilon}(y_0-g_0(x_0,x_c))\right]dF_\epsilon(\epsilon).\end{aligned} \tag{14}$$

(ii) *TT: The joint distribution of potential outcomes for the treated satisfies*

$$F_{10}^L(y_1,y_0|D=1)\le F_{10}^Y(y_1,y_0|D=1)\le F_{10}^U(y_1,y_0|D=1),$$

where

$$F_{10}^L(y_1,y_0|D=1)=\frac{\int_{-(w,x_c)'\gamma}^{\infty}C_L\left(F_{1|\epsilon}(y_1-g_1(x_1,x_c)),F_{0|\epsilon}(y_0-g_0(x_0,x_c))\right)dF_\epsilon(\epsilon)}{1-F_\epsilon(-(w,x_c)'\gamma)},$$

$$F_{10}^U(y_1,y_0|D=1)=\frac{\int_{-(w,x_c)'\gamma}^{\infty}C_U\left(F_{1|\epsilon}(y_1-g_1(x_1,x_c)),F_{0|\epsilon}(y_0-g_0(x_0,x_c))\right)dF_\epsilon(\epsilon)}{1-F_\epsilon(-(w,x_c)'\gamma)}.$$

The result in (i) is presented in Lee (2002). It is an immediate consequence of (12) when $\epsilon=\infty$. To prove (ii), we note that

$$\begin{aligned}&F_{10}^Y(y_1,y_0|D=1)=P(U_{1i}\le y_1-g_1(x_1,x_c),U_{0i}\le y_0-g_0(x_0,x_c)|\epsilon_i>-(w,x_c)'\gamma)\\ =\;&\frac{P(U_{1i}\le y_1-g_1(x_1,x_c),U_{0i}\le y_0-g_0(x_0,x_c),\epsilon_i>-(w,x_c)'\gamma)}{P(\epsilon_i>-(w,x_c)'\gamma)}\\ =\;&\frac{\int_{-(w,x_c)'\gamma}^{\infty}F_{10|\epsilon}(y_1-g_1(x_1,x_c),y_0-g_0(x_0,x_c))dF_\epsilon(\epsilon)}{1-F_\epsilon(-(w,x_c)'\gamma)}.\end{aligned} \tag{15}$$

Now since $F_{10|\epsilon}(y_1 - g_1(x_1, x_c), g_0(x_0, x_c))$ satisfies the Frechet-Hoeffding inequality, we obtain the inequality in (ii).

The bounds in Theorem 3.2 are reached when the two potential outcomes are conditionally (on ϵ) perfectly dependent on each other. One example is $\epsilon = Y_1 - Y_0$ in which Y_1, Y_0 are perfectly positively dependent conditional on ϵ. These bounds take into account the self-selection process and are tighter than the bounds obtained under random selection. For instance, if selection is random, i.e., both U_{1i} and U_{0i} are independent of ϵ_i, then the bounds in Theorem 3.2 (i) become

$$F_{10}^{LI}(y_1, y_0) = C_L\left[F_1(y_1 - g_1(x_1, x_c)), F_0(y_0 - g_0(x_0, x_c))\right], \tag{16}$$

$$F_{10}^{UI}(y_1, y_0) = C_U\left[F_1(y_1 - g_1(x_1, x_c)), F_0(y_0 - g_0(x_0, x_c))\right]. \tag{17}$$

In general, $F_{10}^{LI}(y_1, y_0) \leq F_{10}^{L}(y_1, y_0)$ and $F_{10}^{UI}(y_1, y_0) \geq F_{10}^{U}(y_1, y_0)$ implying that the dependence between the outcome errors and the selection error improves on the bounds on $F_{10}^{Y}(y_1, y_0)$. But even when selection is random, learning can take place about $F_{10}^{Y}(y_1, y_0)$ through its marginals provided that $F_{10}^{LI}(y_1, y_0)$ is not zero or $F_{10}^{UI}(y_1, y_0)$ is not 1.

We now consider bounds on the distribution of $\Delta = Y_1 - Y_0$. Note that

$$ATE \equiv E(\Delta|X = x) = g_1(x_1, x_c) - g_0(x_0, x_c)$$

and $F_\Delta(\delta) = E\left[P(U_1 - U_0 \leq \{\delta - ATE\}|\epsilon)\right]$. Applying Lemma 3.1 to $P(U_1 - U_0 \leq \{\delta - ATE\}|\epsilon)$, we obtain the sharp bounds on the distribution of the treatment effect in Theorem 3.3 (i) below. Other bounds presented in Theorem 3.3 can be obtained in the same way.

THEOREM 3.3 *In a semiparametric SRM, the following inequalities hold.*

(i) *ATE:* $F_\Delta^L(\delta) \leq F_\Delta(\delta) \leq F_\Delta^U(\delta)$, *where*

$$F_\Delta^L(\delta) = \int_{-\infty}^{+\infty}\left[\sup_u \max\left\{F_{1|\epsilon}(u) - F_{0|\epsilon}(u - \{\delta - ATE\}), 0\right\}\right] dF_\epsilon(\epsilon),$$

$$F_\Delta^U(\delta) = \int_{-\infty}^{+\infty}\left[\inf_u \min\left\{1 - F_{0|\epsilon}(u - \{\delta - ATE\}) + F_{1|\epsilon}(u), 1\right\}\right] dF_\epsilon(\epsilon).$$

(ii) *TT: The distribution of* Δ *for the treated satisfies*

$$F_\Delta^L(\delta|D = 1) \leq F_\Delta(\delta|D = 1) \leq F_\Delta^U(\delta|D = 1),$$

where

$$F_\Delta^L(\delta|D = 1) = \frac{\int_{-(w,x_c)'\gamma}^{\infty}\left[\sup_u \max\left\{F_{1|\epsilon}(u) - F_{0|\epsilon}(u - \{\delta - ATE\}), 0\right\}\right] dF_\epsilon(\epsilon)}{1 - F_\epsilon(-(w, x_c)'\gamma)},$$

$$F_\Delta^U(\delta|D = 1) = \frac{\int_{-(w,x_c)'\gamma}^{\infty}\left[\inf_u \min\left\{1 - F_{0|\epsilon}(u - \{\delta - ATE\}) + F_{1|\epsilon}(u), 1\right\}\right] dF_\epsilon(\epsilon)}{1 - F_\epsilon(-(w, x_c)'\gamma)}.$$

In contrast to sharp bounds on the joint distribution of potential outcomes, the sharp bounds on the distribution of the treatment effect are not reached at conditional perfect positive or negative dependence. Again, two types of learning take place here through self-selection and the identified marginals of $F_{10}^{Y}(y_1, y_0)$.

When $\epsilon = Y_1 - Y_0$, the potential outcomes are perfectly positively dependent conditional on ϵ. Let $F_{\Delta|\epsilon}^{R}$ and F_{Δ}^{R} denote respectively the conditional distribution of Δ on ϵ and the unconditional distribution of Δ in this case. Fan and Park (2006) shows that $F_{\Delta|\epsilon}^{R}$ second order stochastically dominates any outcome gain distribution conditional on ϵ, $F_{\Delta|\epsilon}$. Taking expectation with respect to ϵ, we obtain the following theorem.

THEOREM 3.4 *In a semiparametric SRM, F_{Δ}^{R} second order stochastically dominates any F_{Δ} consistent with the sample information.*

Unlike the average treatment parameters such as ATE and TT, the quantile of Δ is in general not identified. By inverting the distribution bounds in Theorem 3.3, we obtain sharp bounds on the quantile of the treatment effect[3] for the whole population and the subpopulation receiving treatment.

3.3 Some Applications of the Distribution Bounds

By using the distribution bounds established in the previous subsection, we can provide informative bounds on many interesting effects other than the average treatment effect. Some illustrative examples are discussed below, see Heckman, Smith, and Clements (1997) for more examples.

1. The proportion of people participating in the program who benefit from it,

$$P(Y_1 > Y_0|D = 1) = P(\Delta > 0|D = 1) = 1 - F_{\Delta}(0|D = 1).$$

2. The proportion of the total population that benefits from the program,

$$P(Y_1 > Y_0|D = 1)P(D = 1) = \{1 - F_{\Delta}(0|D = 1)\} P(D = 1).$$

3. The share of 'productive' workers employed in sector 1,

$$P(D = 1|Y_1 > Y_0) = \frac{\{1 - F_{\Delta}(0|D = 1)\} P(D = 1)}{1 - F_{\Delta}(0)}.$$

[3]Recently, $\left[F_1^{-1}(q) - F_0^{-1}(q)\right]$ has been used to study treatment effect heterogeneity and is referred to as the quantile treatment effect (QTE), see e.g., Heckman, Smith, and Clements (1997), Abadie, Angrist, and Imbens (2002), Chen, Hong, and Tarozzi (2004), Chernozhukov and Hansen (2005), Firpo (2005), Imbens and Newey (2005), among others, for more discussion and references on the estimation of QTE. Manski (1997a) referred to QTE as ΔD-parameters and the quantile of the treatment effect distribution as $D\Delta$-parameters. Assuming monotone treatment response, Manski (1997a) provided sharp bounds on the quantile of the treatment effect distribution.

4. The distribution of the potential outcome Y_1 of an individual with an above average Y_0,

$$P(Y_1 \leq y_1 | U_0 > 0) = \frac{F_1\left(y_1 - g_1(x_1, x_c)\right) - F_{10}\left(y_1 - g_1(x_1, x_c), 0\right)}{1 - F_0(0)}.$$

5. The variance of the treatment effect,

$$\sigma_1^2 + \sigma_0^2 - 2\sigma_1\sigma_0\rho_{10}^U \leq Var(\Delta) \leq \sigma_1^2 + \sigma_0^2 - 2\sigma_1\sigma_0\rho_{10}^L,$$

where ρ_{10}^U (ρ_{10}^L) is the correlation coefficient of the distribution F_{10}^U (F_{10}^L).

6. The variance of the treatment effect for participants (Lee, 2002),

$$\sigma_{L,D=1}^2 \leq Var(\Delta | D = 1) \leq \sigma_{U,D=1}^2,$$

where

$$\begin{aligned}
\sigma_{U,D=1}^2 &= Var\left(Y_1 | D = 1\right) + Var\left(Y_0 | D = 1\right) \\
&\quad -2\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(F_{10}^L(y_1, y_0 | D = 1) - F_1\left(y_1 | D = 1\right) F_0\left(y_0 | D = 1\right)\right) dy_1 dy_0, \\
\sigma_{L,D=1}^2 &= Var\left(Y_1 | D = 1\right) + Var\left(Y_0 | D = 1\right) \\
&\quad -2\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(F_{10}^U(y_1, y_0 | D = 1) - F_1\left(y_1 | D = 1\right) F_0\left(y_0 | D = 1\right)\right) dy_1 dy_0.
\end{aligned}$$

Similar techniques used in the previous subsection may help to establish bounds on other parameters of interest. For example, the distribution of the potential outcome Y_1 of an individual with an above average Y_0 who selects into the program is given by

$$P(Y_1 \leq y_1 | D = 1, U_0 > 0) = \frac{P\left(Y_1 \leq y_1, \epsilon \geq -(w, x_c)'\gamma\right) - \int_{-(w,x_c)'\gamma}^{\infty} F_{10|\epsilon}(y_1 - g_1(x_1, x_c), 0) dF_\epsilon(\epsilon)}{P\left(\epsilon \geq -(w, x_c)'\gamma, U_0 > 0\right)},$$

where the probability in the denominator and the first probability in the numerator are identified from the sample information and the second term in the numerator can be bounded by applying the Frechet-Hoeffding inequality to $F_{10|\epsilon}(y_1 - g_1(x_1, x_c), 0)$.

4 A Comparison of the two sets of Bounds

The distribution bounds developed in Section 3 depend on the bivariate distributions of $\{U_{1i}, \epsilon_i\}$ and $\{U_{0i}, \epsilon_i\}$ which can be parametric or nonparametric. In this section, we first study these bounds when $\{U_{ji}, \epsilon_i\}, j = 1, 0$, follows either the bivariate normal or bivariate Student's t distribution and then compare them with those established in Section 2 for Gaussian or Student's t models. The difference between these two sets of bounds is that our bounds are valid for any joint distribution of the errors $\{U_{1i}, U_{0i}, \epsilon_i\}$ provided the bivariate marginal distributions corresponding to $\{U_{1i}, \epsilon_i\}$ and

$\{U_{0i}, \epsilon_i\}$ are bivariate normal or bivariate Student's t, while the bounds in Section 2 depend crucially on the joint normality or Student's t distribution for the trio of errors $\{U_{1i}, U_{0i}, \epsilon_i\}$. Robustness of our results to the joint distribution of U_{1i} and U_{0i} is a desirable property, as this joint distribution is not identifiable from the sample information alone and any distributional assumption imposed on it can never be verified empirically.

4.1 Bounds on F_Δ in Semiparametric SRMs with Bivariate Normal Distributions

Assume (ϵ_i, U_{ji}) follows a bivariate normal distribution:

$$\begin{pmatrix} U_{ji} \\ \epsilon_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_j^2 & \sigma_j\rho_{j\epsilon} \\ \sigma_j\rho_{j\epsilon} & 1 \end{pmatrix}\right].$$

By making use of the fact that the conditional distribution of U_{ji} given ϵ_i is normal for $j = 0, 1$, we show in Appendix A that the following theorem holds.

THEOREM 4.1 *In a SRM with bivariate normal distributions for $\{U_{ji}, \epsilon_i\}$ for $j = 0, 1$, we have:*

$$\begin{aligned} F_{10}^L(y_1, y_0) &= \int_{-\infty}^{\infty} \max\{F_{1|\epsilon}(y_1 - g_1(x_1, x_c)) + F_{0|\epsilon}(y_0 - g_0(x_0, x_c)) - 1, 0\} dF_\epsilon(\epsilon) \\ &= \Phi_{\rho_L}\left(\frac{y_1 - g_1(x_1, x_c)}{\sigma_1}, \frac{y_0 - g_0(x_0, x_c)}{\sigma_0}\right), \end{aligned} \tag{18}$$

$$\begin{aligned} F_{10}^U(y_1, y_0) &= \int_{-\infty}^{\infty} \min\{F_{1|\epsilon}(y_1 - g_1(x_1, x_c)), F_{0|\epsilon}(y_0 - g_0(x_0, x_c))\} dF_\epsilon(\epsilon) \\ &= \Phi_{\rho_U}\left(\frac{y_1 - g_1(x_1, x_c)}{\sigma_1}, \frac{y_0 - g_0(x_0, x_c)}{\sigma_0}\right). \end{aligned} \tag{19}$$

We observe immediately that these bounds are the same as the bounds on the joint distribution of potential outcomes in Gaussian models presented in Section 2. This is interesting, because it implies that the non-refutable Gaussian assumption on the joint distribution of the potential outcomes in Gaussian models does not improve on the bounds of this joint distribution. Heuristically, this is because the conditional copula for $\{U_{1i}, U_{0i}\}$ given ϵ_i implied by the trivariate normality assumption in Gaussian models is Gaussian with parameter given by the partial correlation between U_{1i} and U_{0i}. Since the partial correlation between U_{1i} and U_{0i} ranges from -1 to 1, the conditional copula for $\{U_{1i}, U_{0i}\}$ given ϵ_i interpolates between the lower bound copula to the upper bound copula, resulting in the same bounds as if the conditional copula for $\{U_{1i}, U_{0i}\}$ is unrestricted at all.

The distribution of U_{ji} given $\epsilon_i = \epsilon$ follows a univariate normal distribution with mean $\sigma_j\rho_{j\epsilon}\epsilon$ and variance $\sigma_j^2(1 - \rho_{j\epsilon}^2)$, $j = 1, 0$. Example 3.1 provides bounds on the distribution of Δ given ϵ, i.e., expressions for

$$\sup_u \max\Big\{F_{1|\epsilon}(u) - F_{0|\epsilon}(u - \{\delta - ATE\}), 0\Big\}$$

and

$$\inf_u \min\left\{1 - F_{0|\epsilon}(u - \{\delta - ATE\}) + F_{1|\epsilon}(u), 1\right\}$$

in Theorem 3.3. Taking their expectations with respect to ϵ leads to the following bounds on $F_\Delta(\delta)$.

THEOREM 4.2 *In a SRM with bivariate normal distributions for $\{U_{ji}, \epsilon_i\}$ for $j = 0, 1$, we have:*

(i) *If $\sigma_1\sqrt{1-\rho_{1\epsilon}^2} = \sigma_0\sqrt{1-\rho_{0\epsilon}^2}$ and $\rho_{j\epsilon}^2 \neq 1$, then*

$$F_\Delta^L(\delta) = 2\int_A \Phi\left(\frac{\{\delta - ATE\} - (\sigma_1\rho_{1\epsilon} - \sigma_0\rho_{0\epsilon})\,\epsilon}{2\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}\right)\phi(\epsilon)d\epsilon - P(A),$$

$$F_\Delta^U(\delta) = 2\int_{A^C} \Phi\left(\frac{\{\delta - ATE\} - (\sigma_1\rho_{1\epsilon} - \sigma_0\rho_{0\epsilon})\,\epsilon}{2\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}\right)\phi(\epsilon)d\epsilon + P(A),$$

where $A = \{\epsilon : \{\delta - ATE\} \geq (\sigma_1\rho_{1\epsilon} - \sigma_0\rho_{0\epsilon})\,\epsilon\}$ and A^C is the complement of A. When $(\sigma_1\rho_{1\epsilon} - \sigma_0\rho_{0\epsilon}) = 0$, A is the whole real line if $\delta \geq ATE$, else A is an empty set;

(ii) If $\sigma_1\sqrt{1-\rho_{1\epsilon}^2} \neq \sigma_0\sqrt{1-\rho_{0\epsilon}^2}$ and $\rho_{j\epsilon}^2 \neq 1$, then

$$\begin{aligned} F_\Delta^L(\delta) &= \int_{-\infty}^{+\infty} \Phi\left(\frac{\sigma_1\sqrt{(1-\rho_{1\epsilon}^2)}s - \sigma_0\sqrt{(1-\rho_{0\epsilon}^2)}t}{\sigma_1^2(1-\rho_{1\epsilon}^2) - \sigma_0^2(1-\rho_{0\epsilon}^2)}\right)\phi(\epsilon)d\epsilon \\ &\quad + \int_{-\infty}^{+\infty} \Phi\left(\frac{\sigma_1\sqrt{(1-\rho_{1\epsilon}^2)}t - \sigma_0\sqrt{(1-\rho_{0\epsilon}^2)}s}{\sigma_1^2(1-\rho_{1\epsilon}^2) - \sigma_0^2(1-\rho_{0\epsilon}^2)}\right)\phi(\epsilon)d\epsilon - 1, \end{aligned}$$

$$\begin{aligned} F_\Delta^U(\delta) &= \int_{-\infty}^{+\infty} \Phi\left(\frac{\sigma_1\sqrt{(1-\rho_{1\epsilon}^2)}s + \sigma_0\sqrt{(1-\rho_{0\epsilon}^2)}t}{\sigma_1^2(1-\rho_{1\epsilon}^2) - \sigma_0^2(1-\rho_{0\epsilon}^2)}\right)\phi(\epsilon)d\epsilon \\ &\quad - \int_{-\infty}^{+\infty} \Phi\left(\frac{\sigma_1\sqrt{(1-\rho_{1\epsilon}^2)}t + \sigma_0\sqrt{(1-\rho_{0\epsilon}^2)}s}{\sigma_1^2(1-\rho_{1\epsilon}^2) - \sigma_0^2(1-\rho_{0\epsilon}^2)}\right)\phi(\epsilon)d\epsilon + 1, \end{aligned}$$

where $s = \{\delta - ATE\} - (\sigma_1\rho_{1\epsilon} - \sigma_0\rho_{0\epsilon})\,\epsilon$ and

$$t = \left(s^2 + 2\left[\sigma_1^2(1-\rho_{1\epsilon}^2) - \sigma_0^2(1-\rho_{0\epsilon}^2)\right]\ln\left(\frac{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}\right)\right)^{\frac{1}{2}}.$$

In contrast to the sharp bounds on the joint distribution of the potential outcomes in Theorem 4.1, the bounds given above on the distribution of the outcome gain differ from the corresponding bounds in Gaussian models and are in general wider, because they are valid for any trivariate distribution with bivariate normal marginals for (U_{1i}, ϵ_i) and (U_{0i}, ϵ_i), not necessarily the trivariate Normal distribution in Gaussian models. On the one hand, imposing the trivariate normality assumption narrows the width of the bounds, but on the other hand, it may lead to misleading conclusions if the implied normality assumption for the joint distribution of potential outcomes is violated. To see the seriousness of this problem, remember in Gaussian models, the value of the treatment effect distribution at its mean is always identified: $F_\Delta(ATE) = 0.5$. However, if the joint distribution of the potential outcomes is unknown, then $F_\Delta(ATE)$ is not identified and the bounds on $F_\Delta(ATE)$ depend on the parameters of the identified bivariate distributions.

In Figure 1, we plotted the two sets of bounds on $F_\Delta(\cdot)$ in Gaussian models and semiparametric models with bivariate normal marginals. We fixed $ATE = 0$, $\sigma_1^2 = 1$ and $\sigma_0^2 = 1$. For $\rho_{1\epsilon} = 0.5$, we chose a range of values for $\rho_{0\epsilon}$. We also plotted the bounds when $\rho_{1\epsilon} = \rho_{0\epsilon} = 0$. Solid curves are bounds in Theorem 4.2 assuming bivariate normality (BN) for (U_{ji}, ϵ_i) only, while dashed curves are bounds in (6) assuming trivariate normality (TN) for $(U_{ji}, U_{0i}, \epsilon_i)$. Several general conclusions emerge from Figure 1. First, for any given set of parameter values, the bounds under bivariate normal marginals are always wider than the bounds under the trivariate normal assumption; Second, for given δ, the bounds in general become narrower as the dependence between U_{0i} and ϵ_i as measured by the magnitude of $\rho_{0\epsilon}$ increases except when $\delta = 0$ in Gaussian models in which case the lower and upper bounds coincide and become 0.5. In the extreme cases where either $\rho_{1\epsilon}^2 = 1$ or $\rho_{0\epsilon}^2 = 1$, the two sets of bounds coincide and both identify the distribution of Δ. More specifically, we have

$$F_\Delta(\delta) = F_\Delta^L(\delta) = F_\Delta^U(\delta) = \Phi\left(\frac{\delta - ATE}{\sqrt{\sigma_1^2 + \sigma_0^2 - 2\rho_{0\epsilon}\sigma_1\sigma_0}}\right), \text{ if } \rho_{1\epsilon} = 1, \tag{20}$$

$$F_\Delta(\delta) = F_\Delta^L(\delta) = F_\Delta^U(\delta) = \Phi\left(\frac{\delta - ATE}{\sqrt{\sigma_1^2 + \sigma_0^2 + 2\rho_{0\epsilon}\sigma_1\sigma_0}}\right), \text{ if } \rho_{1\epsilon} = -1. \tag{21}$$

To see why, consider $\rho_{1\epsilon} = 1$. Then the conditional distribution of U_{1i} given $\epsilon_i = \epsilon$ is degenerate at $\sigma_1\epsilon$. Let $G_z(\cdot)$ denote this degenerate distribution function at z. Then

$$\begin{aligned} F_\Delta^L(\delta) &= \int_{-\infty}^{+\infty}\left[\sup_u \max\left\{G_{\sigma_1\epsilon}(u) - \Phi\left(\frac{u - (\delta - ATE) - \rho_{0\epsilon}\sigma_0\epsilon}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}\right), 0\right\}\right]\phi(\epsilon)d\epsilon \\ &= \int_{-\infty}^{+\infty}\left[\Phi\left(\frac{\delta - ATE - (\sigma_1 - \rho_{0\epsilon}\sigma_0)\epsilon}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}\right)\right]\phi(\epsilon)d\epsilon, \end{aligned}$$

$$
\begin{aligned}
F_{\Delta}^{U}(\delta) &= \int_{-\infty}^{+\infty}\left[\inf_{u}\min\left\{1-\Phi\left(\frac{u-(\delta-ATE)-\rho_{0\epsilon}\sigma_0\epsilon}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}\right)+G_{\sigma_1\epsilon}(u),1\right\}\right]\phi(\epsilon)d\epsilon \\
&= \int_{-\infty}^{+\infty}\left[\Phi\left(\frac{\delta-x(\beta_1-\beta_0)-(\sigma_1-\rho_{0\epsilon}\sigma_0)\epsilon}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}\right)\right]\phi(\epsilon)d\epsilon.
\end{aligned}
$$

Hence $F_{\Delta}(\delta)=F_{\Delta}^{L}(\delta)=F_{\Delta}^{U}(\delta)$. Since

$$
\begin{aligned}
\frac{\partial F_{\Delta}(\delta)}{\partial\delta} &= \int_{-\infty}^{+\infty}\frac{1}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}\phi\left(\frac{\delta-ATE-(\sigma_1-\rho_{0\epsilon}\sigma_0)\epsilon}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}\right)\phi(\epsilon)d\epsilon \\
&= \phi\left(\frac{\delta-ATE}{\sqrt{\sigma_1^2+\sigma_0^2-2\rho_{0\epsilon}\sigma_1\sigma_0}}\right),
\end{aligned}
$$

we get the result for $\rho_{1\epsilon}=1$. The result for $\rho_{1\epsilon}=-1$ follows suit. Third, the bounds corresponding to $(\rho_{1\epsilon},\rho_{0\epsilon})=(0,0)$ are wider than the bounds when $(\rho_{1\epsilon},\rho_{0\epsilon})\neq(0,0)$, because the former does not account for the information through self-selection.

To see how these bounds change with the variance parameters. In Figure 2, we plotted the bounds on $F_{\Delta}(\delta)$ against σ_0 at $\delta=0,1,4$ when $\sigma_1^2=1$, $\rho_{1\epsilon}=0.5$ and $\rho_{0\epsilon}=0.5$. One interesting fact we observe is that the distribution bounds under both trivariate normality and bivariate normality become wider to some point and then narrower as σ_0 goes to ∞.

4.2 Bounds on F_{Δ} in Semiparametric SRMs with Bivariate Student's t Distributions

Suppose $\{U_{ji},\epsilon_i\}$ follows a bivariate Student's t distribution:

$$
\left\{\sqrt{\frac{v}{v-2}}\frac{U_{ji}}{\sigma_j},\sqrt{\frac{v}{v-2}}\epsilon_i\right\}\tilde{}\ t_{[v]}\left(\bullet,\bullet,\rho_{j\epsilon}\right),\ j=1,0.
$$

Then using Example 3.2, we can show:

THEOREM 4.3 *In the generalized sample selection model with bivariate Student's t distributions for $\{U_{ji},\epsilon_i\}$ for $j=0,1$, we have:*

$$
\begin{aligned}
F_{10}^{L}(y_1,y_0) &= \int_{-\infty}^{\infty}\max\{F_{1|\epsilon}(y_1-g_1(x_1,x_c))+F_{0|\epsilon}(y_0-g_0(x_0,x_c))-1,0\}dF_{\epsilon}(\epsilon) \\
&= T_{[v]}\left(\sqrt{\frac{v}{v-2}}\frac{(y_1-g_1(x_1,x_c))}{\sigma_1},\sqrt{\frac{v}{v-2}}\frac{(y_0-g_0(x_0,x_c))}{\sigma_0},\rho_L\right), \qquad (22)
\end{aligned}
$$

$$
\begin{aligned}
F_{10}^{U}(y_1,y_0) &= \int_{-\infty}^{\infty}\min\{F_{1|\epsilon}(y_1-g_1(x_1,x_c)),F_{0|\epsilon}(y_0-g_0(x_0,x_c))\}dF_{\epsilon}(\epsilon) \\
&= T_{[v]}\left(\sqrt{\frac{v}{v-2}}\frac{(y_1-g_1(x_1,x_c))}{\sigma_1},\sqrt{\frac{v}{v-2}}\frac{(y_0-g_0(x_0,x_c))}{\sigma_0},\rho_U\right). \qquad (23)
\end{aligned}
$$

To derive bounds on the distribution of Δ in this case, we make use of the fact that $U_{ji}|\epsilon_i = \epsilon$ follows the univariate Student's t distribution with degrees of freedom $v+1$, mean $\sigma_j \rho_{j\epsilon} \epsilon$, and variance $\sigma_j^2(1-\rho_{j\epsilon}^2)\left(\frac{(v-2)+\epsilon^2}{v-1}\right)$, $j = 1, 0$, see Mardia (1970). Example 3.2 provides bounds on the distribution of Δ given ϵ, i.e., expressions for

$$\sup_u \max\left\{F_{1|\epsilon}(u) - F_{0|\epsilon}(u - \{\delta - ATE\}), 0\right\}$$

and

$$\inf_u \min\left\{1 - F_{0|\epsilon}(u - \{\delta - ATE\}) + F_{1|\epsilon}(u), 1\right\}$$

in Theorem 3.3. Taking their expectations with respect to ϵ leads to the bounds on $F_\Delta(\delta)$.

THEOREM 4.4 *In a SRM with bivariate Student's t distributions for $\{U_{ji}, \epsilon_i\}$ for $j = 0, 1$, we have:*

(i) *Suppose $\sigma_1\sqrt{1-\rho_{1\epsilon}^2} = \sigma_0\sqrt{1-\rho_{0\epsilon}^2} \equiv \sigma$ and $\rho_{j\epsilon}^2 \neq 1$. Let $\overline{\sigma} = \sigma\sqrt{\left(\frac{(v-2)+\epsilon^2}{v-1}\right)}$. Then*

$$F_\Delta^L(\delta) = 2\int_A T_{[v+1]}\left(\left(\frac{\delta - ATE - (\sigma_1\rho_{1\epsilon} - \sigma_0\rho_{0\epsilon})\,\epsilon}{2\overline{\sigma}}\right)\sqrt{\frac{v+1}{v-1}}\right) t_{[v]}(\epsilon)d\epsilon - P(A),$$

$$F_\Delta^U(\delta) = 2\int_{A^C} T_{[v+1]}\left(\left(\frac{\delta - ATE - (\sigma_1\rho_{1\epsilon} - \sigma_0\rho_{0\epsilon})\,\epsilon}{2\overline{\sigma}}\right)\sqrt{\frac{v+1}{v-1}}\right) t_{[v]}(\epsilon)d\epsilon + P(A),$$

where $A = \{\epsilon : \{\delta - ATE\} \geq (\sigma_1\rho_{1\epsilon} - \sigma_0\rho_{0\epsilon})\,\epsilon\}$ and A^C is the complement of A.

(ii) *Suppose $\sigma_1\sqrt{1-\rho_{1\epsilon}^2} \neq \sigma_0\sqrt{1-\rho_{0\epsilon}^2}$ and $\rho_{j\epsilon}^2 \neq 1$. Let $\overline{\sigma}_1 = \sigma_1\sqrt{(1-\rho_{1\epsilon}^2)\left(\frac{(v-2)+\epsilon^2}{v-1}\right)}$ and $\overline{\sigma}_0 = \sigma_0\sqrt{(1-\rho_{0\epsilon}^2)\left(\frac{(v-2)+\epsilon^2}{v-1}\right)}$. Then*

$$\begin{aligned} F_\Delta^L(\delta) &= \int_{-\infty}^{+\infty} T_{[v+1]}\left(\left(\frac{\overline{\sigma}_1^{2\kappa-1}s - \overline{\sigma}_0^{\kappa-1}\overline{\sigma}_1^{\kappa-1}t}{\overline{\sigma}_1^{2\kappa} - \overline{\sigma}_0^{2\kappa}}\right)\sqrt{\frac{v+1}{v-1}}\right) t_{[v]}(\epsilon)d\epsilon \\ &\quad + \int_{-\infty}^{+\infty} T_{[v+1]}\left(\left(\frac{\overline{\sigma}_1^{\kappa}\overline{\sigma}_0^{\kappa-1}t - \overline{\sigma}_0^{2\kappa-1}s}{\overline{\sigma}_1^{2\kappa} - \overline{\sigma}_0^{2\kappa}}\right)\sqrt{\frac{v+1}{v-1}}\right) t_{[v]}(\epsilon)d\epsilon - 1, \end{aligned}$$

$$\begin{aligned} F_\Delta^U(\delta) &= \int_{-\infty}^{+\infty} T_{[v+1]}\left(\left(\frac{\overline{\sigma}_1^{2\kappa-1}s + \overline{\sigma}_0^{\kappa}\overline{\sigma}_1^{\kappa-1}t}{\overline{\sigma}_1^{2\kappa} - \overline{\sigma}_0^{2\kappa}}\right)\sqrt{\frac{v+1}{v-1}}\right) t_{[v]}(\epsilon)d\epsilon \\ &\quad - \int_{-\infty}^{+\infty} T_{[v+1]}\left(\left(\frac{\overline{\sigma}_1^{\kappa}\overline{\sigma}_0^{\kappa-1}t + \overline{\sigma}_0^{2\kappa-1}s}{\overline{\sigma}_1^{2\kappa} - \overline{\sigma}_0^{2\kappa}}\right)\sqrt{\frac{v+1}{v-1}}\right) t_{[v]}(\epsilon)d\epsilon + 1, \end{aligned}$$

where

$$s = \{(\delta - ATE) - (\sigma_1\rho_{1\epsilon} - \sigma_0\rho_{0\epsilon})\,\epsilon\},\ \kappa = \frac{v+1}{v+2},$$

and

$$t = \left(s^2 + \left(\overline{\sigma}_1^{2\kappa} - \overline{\sigma}_0^{2\kappa}\right)\left(\left(\overline{\sigma}_1^{2(1-\kappa)} - \overline{\sigma}_0^{2(1-\kappa)}\right)(v-1)\right)\right)^{\frac{1}{2}}.$$

We evaluated these bounds for the same set of parameters used in the normal case for $v = 4$, see Figure 3. The same general qualitative conclusions hold as in the normal case. Comparing Figures 1 and 3, we observe that the degree of freedom parameter has little effect on the bounds at the ATE, but it has large effect on the bounds away from ATE. This is due to the fact that Student's t distribution has fatter tails than the normal distribution.

4.3 Bounds on $F_\Delta(\cdot|D=1)$ and the Propensity Score

In a SRM, the propensity score is given by

$$P(D=1|W,X_c) = P(\epsilon > -(W,X_c)'\gamma) = 1 - F_\epsilon(-(W,X_c)'\gamma).$$

Hence the smaller the value of $(W,X_c)'\gamma$, the less likely the individual with the value of $(W,X_c)'\gamma$ will participate in the program or the smaller the propensity score. Since there is a one-to-one relation between the propensity score and $(W,X_c)'\gamma$, we can group individuals in the population via their propensity score. For a given value of the propensity score, Theorem 3.3 (ii) provides sharp bounds on the distribution of Δ for participants with the given propensity score in semiparametric SRMs. Figure 4 depicts the distribution bounds for Δ for participants with $(W,X_c)'\gamma = -1.28$ or propensity score 0.1. Figures 4(a) and 4(b) are based on the normal assumption, while Figures 4(c) and 4(d) are based on the Student's t assumption with degree of freedom 4. We observe that the distribution bounds in Student's t case are generally wider than those in normal case. Moreover, plots with different values of the propensity score and/or the degree of freedom in the Student's t case reveal that the degree of skewness of each bound increases as the propensity score decreases and the bounds get tighter as the degree of freedom increases.

One important and potentially useful application of the distribution bounds established in Theorem 3.3 (ii) is to predict or bound the probability that an individual with a given propensity score will benefit from participating in the program. Note that

$$F_\Delta^L(0|D=1) \le P(\Delta \le 0|D=1) \le F_\Delta^U(0|D=1).$$

Hence $1-F_\Delta^L(0|D=1)$ is the maximum probability that an individual with a given propensity score will benefit from participating in the program and $1-F_\Delta^U(0|D=1)$ is the minimum probability that an individual with a given propensity score will benefit from participating in the program. To

see how these probabilities change with respect to the propensity score, we plotted them against the propensity score in SRMs with bivariate normal distributions in Figures 5(b)-10(b). The expressions[4] for $F_{\Delta}^{L}(\delta|D=1)$ and $F_{\Delta}^{U}(\delta|D=1)$ are derived by using Theorem 3.3 (ii) and a similar argument to Theorem 4.2. Using these expressions, one can show[5] that the bounds $F_{\Delta}^{L}(\delta|D=1)$ and $F_{\Delta}^{U}(\delta|D=1)$ approach either 0 or 1 as the propensity score approaches zero. As a result, the bounds are informative for individuals with low propensity score and once they participate, with high probability, they either get hurt or benefit from the treatment.

In a SRM with bivariate normal distributions, TT is given by

$$TT = ATE + (\rho_{1\epsilon}\sigma_1 - \rho_{0\epsilon}\sigma_0)\,\lambda\left((W, X_c)'\,\gamma\right),$$

where $\lambda(\cdot)$ is the inverse mills ratio. For a given value of $(W, X_c)'\,\gamma$ or a given value of the propensity score, TT measures the average treatment effect for the subpopulation of participants with the given propensity score. It is composed of two terms: the first term is the average treatment effect for the population with covariates X_1, X_0, X_c, W and the second term is the effect due to selection on unobservables. Figures 5(a)-10(a) plotted TT and the second term in TT due to unobservables against the propensity score. Also plotted in each graph are the bounds on the median of the distribution $F_{\Delta}(\cdot|D=1)$.

In Figures 5 and 6, ATE is zero. In Figure 5, $(\rho_{1\epsilon}, \rho_{0\epsilon}) = (0.5, -0.5)$ and TT is non-negative for all values of the propensity score. However, when the propensity score is greater than 0.54, there is a positive probability that an individual with the given propensity score will get hurt by participating in the program. This probability increases as the value of the propensity score increases. And for all values of the propensity score, there is always a positive probability that an individual with the given propensity score will benefit from participating in the program and this probability decreases as the value of the propensity score increases. Consequently, people with low propensity score would benefit from the program with high probability once they participate. In Figure 6, $(\rho_{1\epsilon}, \rho_{0\epsilon}) = (-0.5, 0.5)$ and TT is non-positive for all values of the propensity score. However, when the propensity score is less than 0.54, there is a positive probability that an individual with the given propensity score will benefit from participating in the program and this probability increases as the value of the propensity score increases. In addition, for all values of the propensity score, there is always a positive probability that an individual with the given propensity score will get hurt from participating in the program and this probability decreases as the value of the propensity score increases. The seemingly reversal roles of the two probabilities in Figures 5 and 6 are due to the reversal of the correlation values. Consider Figure 5 with $(\rho_{1\epsilon}, \rho_{0\epsilon}) = (0.5, -0.5)$. Heuristically, for small values of the propensity score, individuals participating in the program tend to have large selection errors ϵ. Given the positive correlation between Y_1 and ϵ, Y_1 would tend to be large

[4] They are tedious and hence not provided here, but they are available upon request.

[5] The proofs are elementary, but tedious. They are available upon request.

for those participants. By the same token, the negative correlation between Y_0 and ϵ imply small Y_0. As a result, Δ tend to be large for participants with small propensity score. Figures 5 and 6 demonstrate clearly that average treatment effect parameters such as ATE and TT do not provide a complete picture of the effects of treatment when there is selection on unobserved variables, and the distribution bounds we established in this paper provide useful information that are missed by ATE and TT.

Figures 7 and 8 further support the conclusions we drew from Figures 5 and 6. They are similar to Figures 5 and 6 except that $ATE = -0.5$ in Figure 7 and $ATE = 0.5$ in Figure 8. In both figures, TT is positive for some values of the propensity score and negative for other values of the propensity score. The patterns of $\left[1 - F^L_\Delta(0|D=1)\right]$ and $\left[1 - F^U_\Delta(0|D=1)\right]$ as functions of the propensity score remain the same as in Figures 5 and 6. It is interesting to observe from Figures 7 and 8 that even when the ATE for the whole population is negative (-0.5) or positive (0.5), some subpopulations (those with the propensity score greater or less than 0.73) will in general benefit or get hurt from the program if they join the program. The proportion of people in each subgroup who will benefit or get hurt from being in the program will also change with the level of ATE.

In Figures 9 and 10, we increased $\rho_{1\epsilon}$ to 0.95. Comparing these figures with Figures 5-8, we see clearly that the distribution bounds get tighter as $\rho_{1\epsilon}$ ($\rho_{0\epsilon}$) gets larger. When the magnitudes of $\rho_{1\epsilon}, \rho_{0\epsilon}$ are the same, the bounds are more informative when $\rho_{1\epsilon}$ and $\rho_{0\epsilon}$ have different signs than when they have the same sign.

Summarizing Figures 5-10, we conclude that the unobserved selection error has a large effect on those with low propensity score. That is, those who are less likely to participate in the program will most likely be affected by the program once they participate in the program. Whether they gain or lose from participating in the program once they participate depends on the sign of $(\rho_{1\epsilon}\sigma_1 - \rho_{0\epsilon}\sigma_0)$.

5 An Empirical Application

The effect of literacy on the wages of child labourers has been studied in Poirier and Tobias (2003) and Smith (2005) using SRMs. In terms of our notation, $D = 1$ if the child is literate; $D = 0$ if the child is illiterate. The logarithm of current weekly wage earnings was then modelled for the literate group (Y_1) and for the illiterate group (Y_0). Poirier and Tobias (2003) analyzed a portion of the survey data from the database collected by New Jersey Bureau of Statistics of Labor and Industrials in 1903. Using the Normal SRM, they find a positive but insignificant ATE. Smith (2005) uses the same data source with more observations by including female child labourers, resulting in 873 observations in total. The covariates used in Smith (2005) are described as follows: Labour market experience (experience) is current age less age at which work was begun; Literacy S is the self-reported ability to read, write and do math; Years of education (Dedu) and its square (Dedu2) are measured as deviations about the sample mean (5.61 years) to get rid of the collinearity. The data

sample statistics are presented in the following table.

Table 1: Sample Data Statistics

Name	Mean	Std dev	Min	Max
Weekly Wage($)	4.68	1.63	1.57	17.50
Literacy (Yes= 1, No= 0)	0.795	0.405	0	1
Sex (Female= 1, Male= 0)	0.48	0.500	0	1
Age (years)	15.50	1.16	12	19
Education (years)	5.61	1.98	0.25	8
Age Began work (years)	13.04	0.88	10	16
Experience (years)	2.45	1.01	0	7

In addition to the Normal SRM, Smith (2005) also estimated several other models constructed using the copula approach, see Smith (2003, 2005) for a detailed introduction to copula-based sample selection models or SRMs. To summarize it briefly, each pair of errors $\{U_{ji}, \epsilon_i\}$ in a copula-based SRM is assumed to follow a bivariate distribution of the form: $C_j(F_j(u,\alpha_j), F(\epsilon,\alpha), \theta_j)$, $j = 1, 0$, where $C_j(u, v, \theta_j)$: $0 \leq u, v \leq 1$, is a copula function with parameter θ_j and $F_j(u,\alpha_j), F(\epsilon,\alpha)$ are univariate distribution functions with $\{F(\epsilon,\alpha) : \alpha \in \mathcal{A}\}$ being any family of parametric distribution functions with zero mean and variance 1 and $\{F_j(u,\alpha_j) : \alpha_j \in \mathcal{A}_j\}$ any family of parametric distribution functions with zero mean and variance σ_j^2, $j = 1, 0$. By Sklar's theorem, $F(\epsilon,\alpha)$ and $F_j(u,\alpha_j)$ are respectively the distribution functions of ϵ_i and U_{ji}.

In Smith (2005), the marginal distributions $F(\epsilon,\alpha)$ and $F_j(u,\alpha_j)$ are chosen to be normal, the same as those in Poirier and Tobias (2003), but the copula function is selected from several copula families including the Gaussian copulas for both $\{U_{1i}, \epsilon_i\}$ and $\{U_{0i}, \epsilon_i\}$ and various combinations of copulas in the Archimedean family. Using the maximum likelihood estimation and AIC, Smith (2005) finds that among the models considered, the model that fits the data best is the Gumbel-Clayton model in which the copula of $\{U_{1i}, \epsilon_i\}$ is the Gumbel copula and the copula of $\{U_{0i}, \epsilon_i\}$ is the Clayton copula:

$$C_1(u, v, \theta_1) = \exp\left(-\left[\{-\log u\}^{\theta_1} + \{-\log v\}^{\theta_1}\right]^{1/\theta_1}\right), 1 \leq \theta_1 < \infty,$$

$$C_0(u, v, \theta_0) = \left(u^{-\theta_0} + v^{-\theta_0} - 1\right)^{-1/\theta_0}, 0 \leq \theta_0 < \infty.$$

The maximum likelihood estimation results with t-statistics in the parentheses appear in the following table.

As reported in Smith (2005), the Gumbel-Clayton model leads to a significant negative estimate for ATE. We now extend Smith's study by going beyond ATE to provide bounds on the distribution of the treatment effect for different subpopulations. The ATE conditional on the propensity score varies with the covariates, thus we use the ATE for the whole population, the maximum and minimum ATE among the sample covariates to construct the distribution bounds for each subpopulation. We plot TT for different subgroups corresponding to different values of the propensity

Table 2: ML Estimation Results

Selection	1	Sex	DEdu	DEdu2
	0.887 (12.23)	0.063 (0.71)	−0.080 (−4.13)	−0.032 (−3.81)
Y_1	1	Sex	Experience	σ_1
	1.086 (38.08)	−0.010 (−0.46)	0.138 (14.60)	0.314 (31.24)
Y_0	1	Sex	Experience	σ_0
	1.373 (25.66)	0.071 (1.93)	0.154 (9.75)	0.322 (11.02)
Copula Parameters	Gumbel: θ_1	Clayton: θ_0		
	5.239 (7.68)	2.736 (5.50)		

score and the bounds on the probability that a person in the subgroup benefits from being literate, see Figures 11-13. Focusing on the sample propensity score range ($[0.65, 0.84]$), the three graphs do not show much difference. The choice of the conditional ATE only causes around 5 percent difference in the upper bound and 1 percent difference in the lower bound on the proportion who benefit from being literate. All three graphs show that in the relevant range for the propensity score, the average effect of being literate on weekly wages is negative. Figure 11 reveals that for subpopulations with the propensity score greater than 0.6, a large proportion (over 70%) in the subpopulation get hurt from being literate. However, there is a positive proportion who benefit from being literate. Subpopulations with low propensity score (e.g., less than 0.05) on average benefit from being literate and at least 45% of the subgroups surely benefit from being literate.

Based on estimates of the model parameters, we computed estimates of $\rho_{1\epsilon}$ and $\rho_{0\epsilon}$. They are respectively 0.9503 and 0.7561, resulting in an estimate 0.5149 for ρ_L and an estimate 0.9222 for ρ_U. We know that these bounds are not sharp. The sharp bounds are given by the correlation coefficients of F_{10}^L and F_{10}^U respectively, see Theorem 3.1 and are estimated to be 0.5404 and 0.8795. Both sets of bounds strongly suggest that the weekly wage earnings of an individual in two states are positively related.

6 Conclusion

In this paper we have established sharp bounds on distributions of the treatment effect in SRMs where the identified distributions can take any parametric form or even be nonparametric. The means of these distributions correspond to various average treatment effects that have received most attention in the literature. While our results can be seen as extensions of similar results already established for the Gaussian SRM, they are established using a completely different approach; the existing approach based on the positive semidefiniteness of the covariance matrix does not work in general. Moreover, our sharp bounds do not depend on any parametric specification of either the outcome or the selection equations or any parametric specification of the bivariate distributions, as long as they are identified.

As a first step, this paper has focused on identification. Estimation of the distribution bounds developed in Section 3 is straightforward in view of the identification results in Heckman (1990) and existing work on estimation of parametric/semiparametric sample selection models. Heckman (1990) provides a review of various nonparametric/semiparametric methods for estimating $g_1(x_1, x_c)$ and $g_0(x_0, x_c)$ without specifying the bivariate margins for (U_{1i}, ϵ_i) and (U_{0i}, ϵ_i), see also Ai (1997), Andrews and Schafgans (1998), Schafgans and Zinde-Walsh (2002), Das, Newey, and Vella (2003), Chen (2006), and Chen and Zhou (2006). Gallant and Nychka (1987) provide estimators of the unknown marginal distributions $F_{1\epsilon}$ and $F_{0\epsilon}$.

Given the bounds established in Section 3, statistical inference on the distribution of the treatment effect falls in a currently active research area: inference on partially identified parameters, see e.g., Imbens and Manski (2004), Chernozhukov, V., H. Hong and E. Tamer (2004), and Romano and Shaikh (2006). A complete treatment of this important issue for SRMs is beyond the scope of this paper and left for future research.

Appendix A: Technical Proofs

Proof of Corollary 4.1 (18): Let $u_j = y_j - x'\beta_j$ for $j = 1, 0$. Note that

$$\begin{aligned} F_{10}^L(u_1, u_0) &= \int_{-\infty}^{\infty} \max\{F_{1|\epsilon}(u_1|\epsilon) + F_{0|\epsilon}(u_0|\epsilon) - 1, 0\} dF_\epsilon(\epsilon) \\ &= \int_{-\infty}^{\infty} \max\left\{ \Phi(\frac{u_1 - \rho_{1\epsilon}\sigma_1\epsilon}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}), \Phi(-\frac{u_0 - \rho_{0\epsilon}\sigma_0\epsilon}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}) \right\} d\Phi(\epsilon) + \Phi(\frac{u_0}{\sigma_0}) - 1. \end{aligned}$$

Let $a^* = a(u_1, u_0)$ satisfy

$$\frac{u_1 - \rho_{1\epsilon}\sigma_1 a^*}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}} = -\frac{u_0 - \rho_{0\epsilon}\sigma_0 a^*}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}.$$

Then

$$a^* = \frac{\sqrt{1-\rho_{1\epsilon}^2}\bar{u}_0 + \sqrt{1-\rho_{0\epsilon}^2}\bar{u}_1}{\sqrt{1-\rho_L^2}}, \tag{A.1}$$

where $\bar{u}_1 = \frac{u_1}{\sigma_1}$ and $\bar{u}_0 = \frac{u_0}{\sigma_0}$, and

$$\begin{aligned} & F_{10}^L(u_1, u_0) \\ = & \int_{-\infty}^{a^*} \Phi(\frac{u_1 - \rho_{1\epsilon}\sigma_1\epsilon}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}) d\Phi(\epsilon) + \int_{a^*}^{\infty} \Phi(-\frac{u_0 - \rho_{0\epsilon}\sigma_0\epsilon}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}) d\Phi(\epsilon) + \Phi(\frac{u_0}{\sigma_0}) - 1. \end{aligned}$$

Hence,

$$\begin{aligned} \frac{\partial F_{10}^L(u_1, u_0)}{\partial u_1} &= \frac{\partial a^*}{\partial u_1}\Phi(\frac{u_1 - \rho_{1\epsilon}\sigma_1 a^*}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}})\phi(a^*) + \int_{-\infty}^{a^*} \frac{1}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}\phi(\frac{u_1 - \rho_{1\epsilon}\sigma_1\epsilon}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}})\phi(\epsilon)d\epsilon \\ &\quad -\frac{\partial a^*}{\partial u_1}\Phi(-\frac{u_0 - \rho_{0\epsilon}\sigma_0 a^*}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}})\phi(a^*) \\ &= \frac{1}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}\int_{-\infty}^{a^*} \phi(\frac{u_1 - \rho_{1\epsilon}\sigma_1\epsilon}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}})\phi(\epsilon)d\epsilon \end{aligned}$$

and

$$\frac{\partial^2 F_{10}^L(u_1, u_0)}{\partial u_1 \partial u_0} = \frac{1}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}\frac{\partial a^*}{\partial u_0}\phi(\frac{u_1 - \rho_1\sigma_1 a^*}{\sigma_1\sqrt{1-\rho_1^2}})\phi(a^*). \tag{A.2}$$

It follows from (A.1) that

$$\frac{\partial a^*}{\partial u_0} = \frac{\sqrt{1-\rho_{1\epsilon}^2}}{\rho_{0\epsilon}\sigma_0\sqrt{1-\rho_{1\epsilon}^2} + \rho_{1\epsilon}\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}.$$

Tedious, but straightforward algebra shows:

$$\left(\frac{u_1 - \rho_{1\epsilon}\sigma_1 a^*}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}\right)^2 + a^{*2} = \frac{\bar{u}_1^2}{(1-\rho_L^2)} - \frac{2\rho_L\bar{u}_1\bar{u}_0}{(1-\rho_L^2)} + \frac{\bar{u}_0^2}{(1-\rho_L^2)}.$$

Consequently,

$$\begin{aligned}
&\frac{\partial^2 F_{10}^L(u_1,u_0)}{\partial u_1 \partial u_0} \\
= \quad & \frac{1}{2\pi\sigma_1\sigma_0\left[\rho_{0\epsilon}\sqrt{1-\rho_{1\epsilon}^2}+\rho_{1\epsilon}\sqrt{1-\rho_{0\epsilon}^2}\right]}\exp\left(-\frac{\bar{u}_1^2-2\rho_L\bar{u}_1\bar{u}_0+\bar{u}_0^2}{2\left(1-\rho_L^2\right)}\right) \\
= \quad & \phi_{\rho_L}(\frac{u_1}{\sigma_1},\frac{u_0}{\sigma_0}).
\end{aligned}$$

Proof of Corollary 4.1 (19): It is similar to the proof of (18). We only provide a sketch. Note that

$$F_{10}^U(u_1,u_0)=\int_{-\infty}^{\infty}\min\{\Phi(\frac{u_1-\rho_{1\epsilon}\sigma_1\epsilon}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}),\Phi(\frac{u_0-\rho_{0\epsilon}\sigma_0\epsilon}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}})\}d\Phi(\epsilon).$$

Let $a^*=a(u_1,u_0)$ satisfy

$$\frac{u_1-\rho_{1\epsilon}\sigma_1 a^*}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}=\frac{u_0-\rho_{0\epsilon}\sigma_0 a^*}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}}.$$

Then

$$a^*=\frac{\sqrt{1-\rho_{1\epsilon}^2}\bar{u}_0-\sqrt{1-\rho_{0\epsilon}^2}\bar{u}_1}{\sqrt{1-\rho_U^2}}.$$

Without loss of generality, we assume $\frac{\rho_{1\epsilon}}{\sqrt{1-\rho_{1\epsilon}^2}}\leq\frac{\rho_{0\epsilon}}{\sqrt{1-\rho_{0\epsilon}^2}}$. Then

$$F_{10}^U(u_1,u_0)=\int_{-\infty}^{a^*}\Phi(\frac{u_1-\rho_{1\epsilon}\sigma_1\epsilon}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}})d\Phi(\epsilon)+\int_{a^*}^{\infty}\Phi(\frac{u_0-\rho_{0\epsilon}\sigma_0\epsilon}{\sigma_0\sqrt{1-\rho_{0\epsilon}^2}})d\Phi(\epsilon).$$

It follows that

$$\frac{\partial^2 F_{10}^U(u_1,u_0)}{\partial u_1\partial u_0}=\frac{1}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}\frac{\partial a^*}{\partial u_0}\phi(\frac{u_1-\rho_{1\epsilon}\sigma_1 a^*}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}})\phi(a^*).$$

The result follows from

$$\frac{\partial a^*}{\partial u_0}=\frac{\sqrt{1-\rho_{1\epsilon}^2}}{\rho_{0\epsilon}\sigma_0\sqrt{1-\rho_{1\epsilon}^2}-\rho_{1\epsilon}\sigma_0\sqrt{1-\rho_{0\epsilon}^2}},$$

and

$$\left(\frac{u_1-\rho_{1\epsilon}\sigma_1 a^*}{\sigma_1\sqrt{1-\rho_{1\epsilon}^2}}\right)^2+a^{*2}=\frac{\bar{u}_1^2-2\rho_U\bar{u}_1\bar{u}_0+\bar{u}_0^2}{1-\rho_U^2}.$$

□

Appendix B: Distribution Bounds Corresponding to LATE and MTE

Sharp bounds on the distribution of the treatment effect corresponding to LATE and MTE in SRMs can be established using the methods in Sections 2 and 3 for Gaussian SRM and semiparametric SRMs respectively. For example, for the Gaussian SRM, Poirier and Tobias (2003) provided the distribution of Δ conditional on $X = x$ corresponding to LATE and MTE:

$$\begin{aligned}\text{LATE} \quad : \quad & F_{\Delta|x,-\widetilde{W}'\gamma\leq\epsilon\leq-W'\gamma}(\delta) \\ = \quad & \frac{\int_{-\infty}^{\delta}\frac{1}{\gamma_2}\phi\left(\frac{\zeta-x'(\beta_1-\beta_0)}{\gamma_2}\right)\left[\Phi\left(\frac{\frac{\gamma_1}{\gamma_2}(\zeta-x'(\beta_1-\beta_0))+\widetilde{W}'\gamma}{\sqrt{1-\gamma_1^2/\gamma_2^2}}\right)-\Phi\left(\frac{\frac{\gamma_1}{\gamma_2}(\zeta-x'(\beta_1-\beta_0))+W'\gamma}{\sqrt{1-\gamma_1^2/\gamma_2^2}}\right)\right]d\zeta}{\Phi\left(\widetilde{W}'\gamma\right)-\Phi\left(W'\gamma\right)},\end{aligned}$$

$$\text{MTE: } F_{\Delta|x,\epsilon=-W'\gamma}(\delta)=\int_{-\infty}^{\delta}\frac{1}{\gamma_2\sqrt{1-\gamma_1^2/\gamma_2^2}}\phi\left(\frac{\zeta-x'(\beta_1-\beta_0)+\gamma_1W'\gamma}{\gamma_2\sqrt{1-\gamma_1^2/\gamma_2^2}}\right)d\zeta.$$

Taking the minimum and maximum of the above expressions results in sharp bounds on $F_{\Delta|x,-\widetilde{W}'\gamma\leq\epsilon\leq-W'\gamma}(\delta)$ and $F_{\Delta|x,\epsilon=-W'\gamma}(\delta)$ respectively.

In general SRMs, we have

THEOREM B.1

(i) *LATE: The joint distribution of potential outcomes corresponding to LATE satisfies*

$$\begin{aligned}F_{10}^{L}(y_1,y_0|-(\widetilde{W}_i,\widetilde{X}_{ci})'\gamma \quad &\leq \quad \epsilon_i\leq-(W_i,X_{ci})'\gamma)\leq \\ F_{10}^{Y}(y_1,y_0|-(\widetilde{W}_i,\widetilde{X}_{ci})'\gamma \quad &\leq \quad \epsilon_i\leq-(W_i,X_{ci})'\gamma)\leq \\ F_{10}^{U}(y_1,y_0|-(\widetilde{W}_i,\widetilde{X}_{ci})'\gamma \quad &\leq \quad \epsilon_i\leq-(W_i,X_{ci})'\gamma),\end{aligned}$$

where

$$\begin{aligned}& F_{10}^{L}(y_1,y_0|-(\widetilde{W}_i,\widetilde{X}_{ci})'\gamma\leq\epsilon_i\leq-(W_i,X_{ci})'\gamma) \\ = \quad & \frac{\int_{-(\widetilde{W}_i,\widetilde{X}_{ci})'\gamma}^{-(W_i,X_{ci})'\gamma}C_L\left(F_{1|\epsilon}(y_1-g_1(x_1,x_c)),F_{0|\epsilon}(y_0-g_0(x_0,x_c))\right)dF_\epsilon(\epsilon)}{F_\epsilon(-(W_i,X_{ci})'\gamma)-F_\epsilon(-(\widetilde{W}_i,\widetilde{X}_{ci})'\gamma)}, \\ & F_{10}^{U}(y_1,y_0|-(\widetilde{W}_i,\widetilde{X}_{ci})'\gamma\leq\epsilon_i\leq-(W_i,X_{ci})'\gamma) \\ = \quad & \frac{\int_{-(\widetilde{W}_i,\widetilde{X}_{ci})'\gamma}^{-(W_i,X_{ci})'\gamma}C_U\left(F_{1|\epsilon}(y_1-g_1(x_1,x_c)),F_{0|\epsilon}(y_0-g_0(x_0,x_c))\right)dF_\epsilon(\epsilon)}{F_\epsilon(-(W_i,X_{ci})'\gamma)-F_\epsilon(-(\widetilde{W}_i,\widetilde{X}_{ci})'\gamma)}.\end{aligned}$$

(ii) *MTE: The joint distribution of potential outcomes corresponding to MTE satisfies*

$$\begin{aligned}& F_{10}^{L}(y_1,y_0|\epsilon_i=-(W_i,X_{ci})'\gamma)\leq F_{10}^{Y}(y_1,y_0|\epsilon_i=-(W_i,X_{ci})'\gamma) \\ \leq \quad & F_{10}^{U}(y_1,y_0|\epsilon_i=-(W_i,X_{ci})'\gamma),\end{aligned}$$

where

$$\begin{aligned} F_{10}^{L}(y_1, y_0|\epsilon_i &= -(W_i, X_{ci})'\gamma) \\ &= C_L\left(F_{1|\epsilon}(y_1 - g_1(x_1, x_c)), F_{0|\epsilon}(y_0 - g_0(x_0, x_c))\right), \\ F_{10}^{U}(y_1, y_0|\epsilon_i &= -(W_i, X_{ci})'\gamma) \\ &= C_U\left(F_{1|\epsilon}(y_1 - g_1(x_1, x_c)), F_{0|\epsilon}(y_0 - g_0(x_0, x_c))\right). \end{aligned}$$

(iii) *LATE: The distribution of* Δ *associated with the LATE satisfies*

$$F_{\Delta}^{L}(\delta|-\widetilde{W}_i'\gamma \leq \epsilon_i \leq -W_i'\gamma) \leq F_{\Delta}(\delta|-\widetilde{W}_i'\gamma \leq \epsilon_i \leq -W_i'\gamma) \leq F_{\Delta}^{U}(\delta|-\widetilde{W}_i'\gamma \leq \epsilon_i \leq -W_i'\gamma),$$

where

$$\begin{aligned} & F_{\Delta}^{L}(\delta|-\widetilde{W}_i'\gamma \leq \epsilon_i \leq -W_i'\gamma) \\ = & \frac{\int_{-\widetilde{W}_i'\gamma}^{-W_i'\gamma}\left[\sup_u \max\left\{F_{1|\epsilon}(u) - F_{0|\epsilon}(u - \{\delta - [g_1(x_1, x_c) - g_0(x_0, x_c)]\}), 0\right\}\right] dF_{\epsilon}(\epsilon)}{F_{\epsilon}(-W_i'\gamma) - F_{\epsilon}(-\widetilde{W}_i'\gamma)}, \\ & F_{\Delta}^{U}(\delta|-\widetilde{W}_i'\gamma \leq \epsilon_i \leq -W_i'\gamma) \\ = & \frac{\int_{-\widetilde{W}_i'\gamma}^{-W_i'\gamma}\left[\inf_u \min\left\{1 - F_{0|\epsilon}(u - \{\delta - [g_1(x_1, x_c) - g_0(x_0, x_c)]\}) + F_{1|\epsilon}(u), 1\right\}\right] dF_{\epsilon}(\epsilon)}{F_{\epsilon}(-W_i'\gamma) - F_{\epsilon}(-\widetilde{W}_i'\gamma)}. \end{aligned}$$

(iv) *MTE: The distribution of* Δ *associated with the MTE satisfies*

$$F_{\Delta}^{L}(\delta|\epsilon_i = -W_i'\gamma) \leq F_{\Delta}(\delta|\epsilon_i = -W_i'\gamma) \leq F_{\Delta}^{U}(\delta|\epsilon_i = -W_i'\gamma),$$

where

$$\begin{aligned} F_{\Delta}^{L}(\delta|\epsilon_i &= -W_i'\gamma) \\ &= \sup_u \max\left\{F_{1|\epsilon}(u) - F_{0|\epsilon}(u - \{\delta - [g_1(x_1, x_c) - g_0(x_0, x_c)]\}), 0\right\}, \\ F_{\Delta}^{U}(\delta|\epsilon_i &= -W_i'\gamma) \\ &= \inf_u \min\left\{1 - F_{0|\epsilon}(u - \{\delta - [g_1(x_1, x_c) - g_0(x_0, x_c)]\}) + F_{1|\epsilon}(u), 1\right\}. \end{aligned}$$

References

[1] Aakvik, A. , J. Heckman, and E. Vytlacil (2003), "Treatment Effects for Discrete Outcomes When Responses to Treatment Vary Among Observationally Identical Persons: An Application to Norwegian Vocational Rehabilitation Programs," Forthcoming in *Journal of Econometrics*.

[2] Abadie, A., J. Angrist, and G. Imbens (2002), "Instrumental Variables Estimation of Quantile Treatment Effects," *Econometrica* 70, 91-117.

[3] Ai, C. (1997), "A Semiparametric Maximum Likelihood Estimator," *Econometrica* 65, 933-963.

[4] Alsina, C. (1981), "Some Functional Equations in the Space of Uniform Distribution Functions," *Equationes Mathematicae* 22, 153-164.

[5] Andrews, D. W. K. and M. M. A. Schafgans (1998), "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies* 65, 497-517.

[6] Biddle, J., L. Boden and R. Reville (2003), "A Method for Estimating the Full Distribution of a Treatment Effect, With Application to the Impact of Workfare Injury on Subsequent Earnings." Mimeo.

[7] Carneiro, P. , K. T. Hansen, and J. Heckman (2003), "Estimating Distributions of Treatment Effects With an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review* 44(2), 361-422.

[8] Chen, S. (2006), "Nonparametric Identification and Estimation of Truncated Regression Models," Working Paper, The Hong Kong University of Science and Technology.

[9] Chen, X., H. Hong, and A. Tarozzi (2004), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects," Working paper.

[10] Chen, S. and Y. Zhou (2006), "Semiparametric and Nonparametric Estimation of Sample Selection Models Under Symmetry," Working Paper, The Hong Kong University of Science and Technology.

[11] Chernozhukov, V. and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," *Econometrica* 73, 245-261.

[12] Chernozhukov, V., H. Hong and E. Tamer (2004), "Inference on Parameter Sets in Econometric Models," Working paper.

[13] Das, M., W. Newey, and F. Vella (2003), "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies* 70, 33-58.

[14] Denuit, M. , C. Genest, and E. Marceau (1999), "Stochastic Bounds on Sums of Dependent Risks," *Insurance: Mathematics and Economics* 25, 85-104.

[15] Fan, Y. (2005a), "Sharp Correlation Bounds and Their Applications," Mimeo.

[16] — (2005b), "Statistical Inference on the Frechet-Hoeffding Distribution Bounds," Mimeo.

[17] Fan, Y. and S. Park (2006), "Sharp Bounds on the Distribution of the Treatment Effect and Their Statistical Inference," Manuscript, Vanderbilt University.

[18] Firpo, S. (2005), "Efficient Semiparametric Estimation of Quantile Treatment Effects," Forthcoming in *Econometrica*.

[19] Frank, M. J. , R. B. Nelsen, and B. Schweizer (1987). "Best-Possible Bounds on the Distribution of a Sum—a Problem of Kolmogorov," *Probability Theory and Related Fields* 74, 199-211.

[20] Gallant, R. and D. Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica* 55, 363-390.

[21] Heckman, J. J. (1990), "Varieties of Selection Bias," *American Economic Review, Papers and Proceedings* 80, 313-318.

[22] Heckman, J. J. and Bo E. Honore (1990), "The Empirical Content of the Roy Model," *Econometrica* 58, 1121-1149.

[23] Heckman, J. J. , J. Smith, and N. Clements (1997), "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *Review of Economic Studies* 64, 487-535.

[24] Heckman, J., J. L. Tobias, and E. Vytlacil (2003), "Simple Estimators for Treatment Parameters in a Latent Variable Framework," *Review of Economics and Statistics* 85, 748-755.

[25] Heckman, J.J. and E. Vytlacil (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica* 73, 669-738.

[26] Imbens, G. W. and C. F. Manski (2004), "Confidence Intervals For Partially Identified Parameters." *Econometrica* 72, 1845–1857.

[27] Imbens, G. W. and W. Newey (2005), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," Working Paper.

[28] Joe, H. (1997). *Multivariate Models and dependence Concepts*. Chapman & Hall/CRC, London.

[29] Koop, G. and D. J. Poirier (1997), "Learning About the Across-Regime Correlation in Switching Regression Models," *Journal of Econometrics* 78, 217-227.

[30] Lee, L. F. (2002), "Correlation Bounds for Sample Selection Models with Mixed Continuous, Discrete and Count Data Variables," Manuscript, The Ohio State University.

[31] Li, M., D. J. Poirier, and J. L. Tobias (2004), "Do Dropouts Suffer From Dropping Out? Estimation and Prediction of Outcome Gains in Generalized Selection Models," *Journal of Applied Econometrics* 19, 203-225.

[32] Makarov, G. D. (1981), "Estimates for the Distribution Function of a Sum of two Random Variables When the Marginal Distributions are Fixed," *Theory of Probability and its Applications* 26, 803-806.

[33] Manski, C. F. (1997a), "Monotone Treatment Effect," *Econometrica* 65, 1311-1334.

[34] Manski, C. F. (1997b), "The Mixing Problem in Programme Evaluation," *Review of Economic Studies* 64, 537-553.

[35] Nelsen, R. B. (1999). *An Introduction to Copulas*. Springer, New York.

[36] Poirier, D. J. (1998), "Revising Beliefs in Non-Identified Models," *Econometric Theory* 14, 483-509.

[37] Poirier, D. J. and J. L. Tobias (2003), "On the Predictive Distributions of Outcome Gains in the Presence of an Unidentified Parameter," *Journal of Business & Economic Statistics* 21, 258-268.

[38] Romano, J. and A. M. Shaikh (2006), "Inference for Identifiable Parameters in Partially Identified Econometric Models," Working Paper.

[39] Rüschendorf, L. (1982), "Random Variables With Maximum Sums," *Advances in Applied Probability* 14, 623-632.

[40] Schafgans, M. M. A. and V. Zinde-Walsh (2002), "On Intercept Estimation in the Sample Selection Model," *Econometric Theory* 18, 40-50.

[41] Smith, M. D. (2003), "Modelling Sample Selection Using Archimedean Copulas," *Econometrics Journal* 6, 99-123.

[42] — (2005), "Using Copulas to Model Switching Regimes with an Application to Child Labour," *The Economic Record* 81, S47-S57.

[43] Vijverberg, W. P. M. (1993), "Measuring the Unidentified Parameter of the Extended Roy Model of Selectivity," *Journal of Econometrics* 57, 69-89.

[44] Williamson, R. C. and T. Downs (1990), "Probabilistic Arithmetic I: Numerical Methods for Calculating Convolutions and Dependency Bounds," *International Journal of Approximate Reasoning* 4, 89-158.

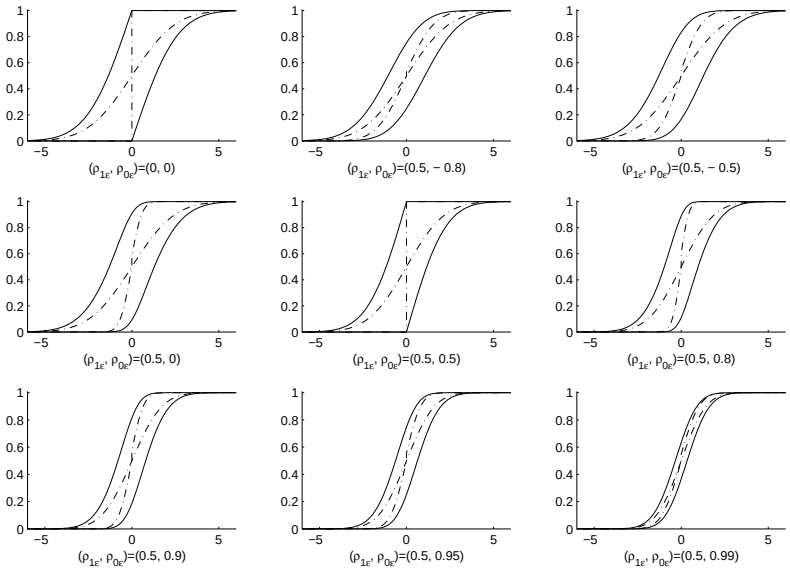

Figure 1: Sharp bounds on the distribution of the treatment effect, $\sigma_1 = \sigma_0 = 1$. Dashed curves are bounds under the trivariate normality assumption for $(U_{1i}, U_{0i}, \epsilon_i)$ and solid curves are bounds assuming bivariate normality for (U_{ji}, ϵ_i), $j = 1, 0$.

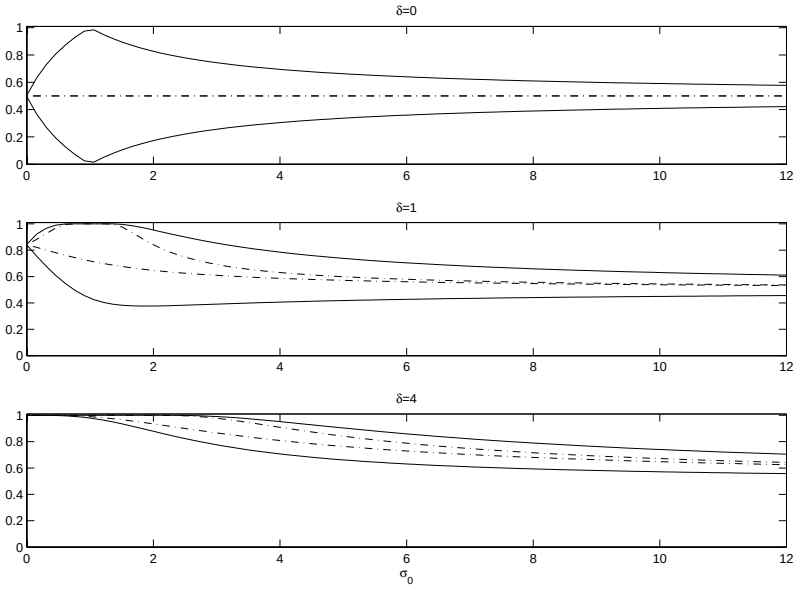

Figure 2: Sharp bounds on the distribution of the treatment effect at a given δ — $\sigma_1 = 1$, $\rho_{1\varepsilon} = 0.5$, and $\rho_{0\varepsilon} = 0.5$. Dashed curves are bounds under the trivariate normality assumption for $(U_{1i}, U_{0i}, \epsilon_i)$ and solid curves are bounds assuming bivariate normality for (U_{ji}, ϵ_i), $j = 1, 0$.

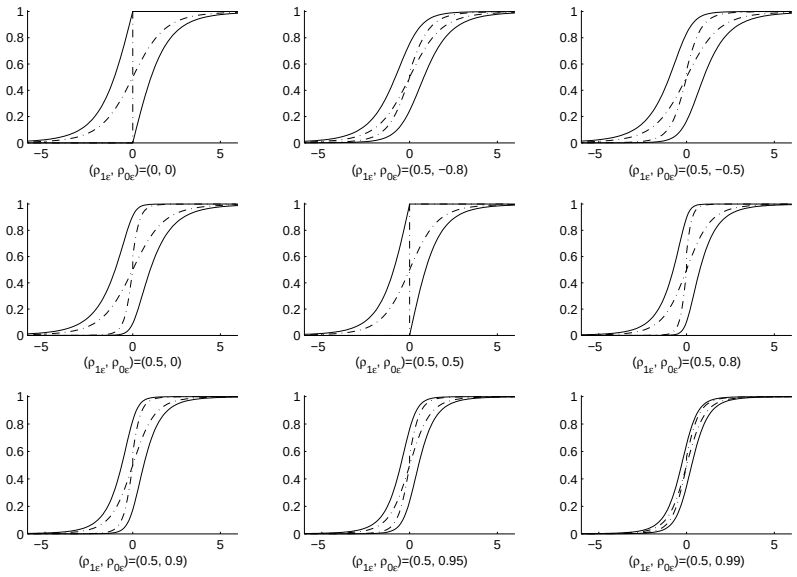

Figure 3: Sharp bounds on the distribution of the treatment effect, $\sigma_1 = \sigma_0 = 1$. Dashed curves are bounds assuming $(U_{1i}, U_{0i}, \epsilon_i)$ follows trivariate Student's t distribution with 4 degrees of freedom and solid curves are bounds assuming (U_{ji}, ϵ_i) follows bivariate Student's t distribution with 4 degrees of freedom.

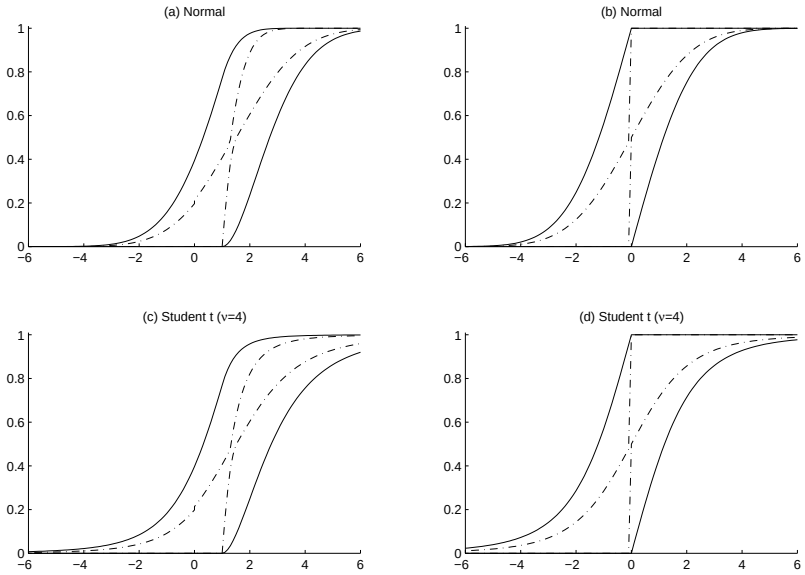

Figure 4: Sharp bounds on the distribution of the treatment effect for the treated — $ATE = 0$, $\sigma_1 = \sigma_0 = 1$, and the Propensity Score = 0.1. In (a) and (c), $\rho_{1\varepsilon} = 0.5$ and $\rho_{0\varepsilon} = -0.5$, while in (b) and (d), $\rho_{1\varepsilon} = \rho_{0\varepsilon} = 0.5$.

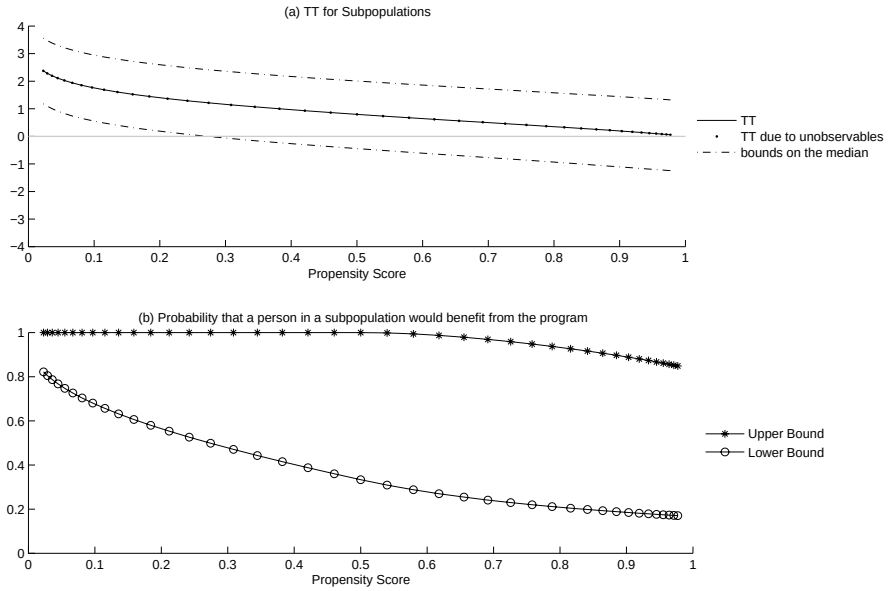

Figure 5: Treatment effect for the treated for subpopulations and the probability that a person in a subpopulation benefits from the treatment, where $ATE = 0$, $\rho_{1\varepsilon} = 0.5$, $\rho_{0\varepsilon} = -0.5$, and $\sigma_1 = \sigma_0 = 1$.

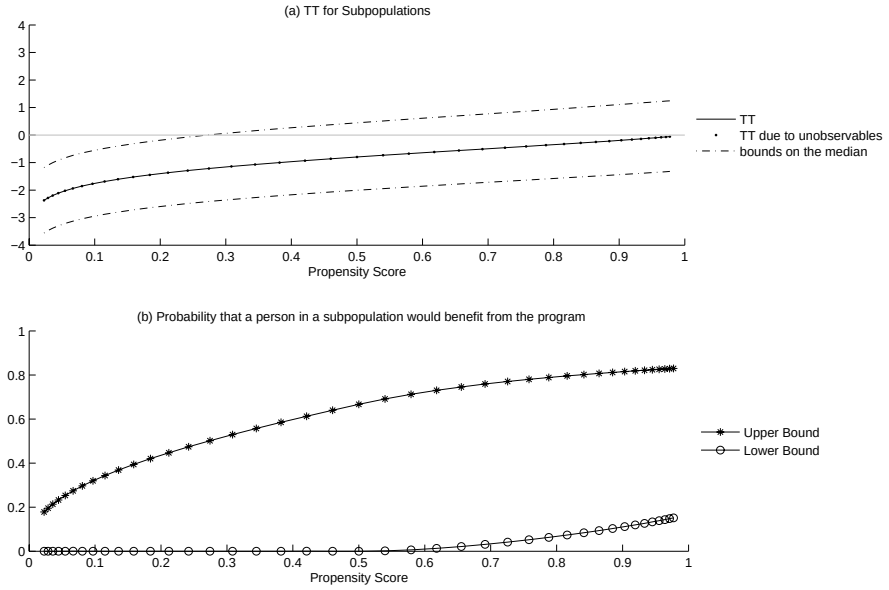

Figure 6: Treatment effect for the treated for subpopulations and the probability that a person in a subpopulation benefits from the treatment, where $ATE = 0$, $\rho_{1\varepsilon} = -0.5$, $\rho_{0\varepsilon} = 0.5$, and $\sigma_1 = \sigma_0 = 1$.

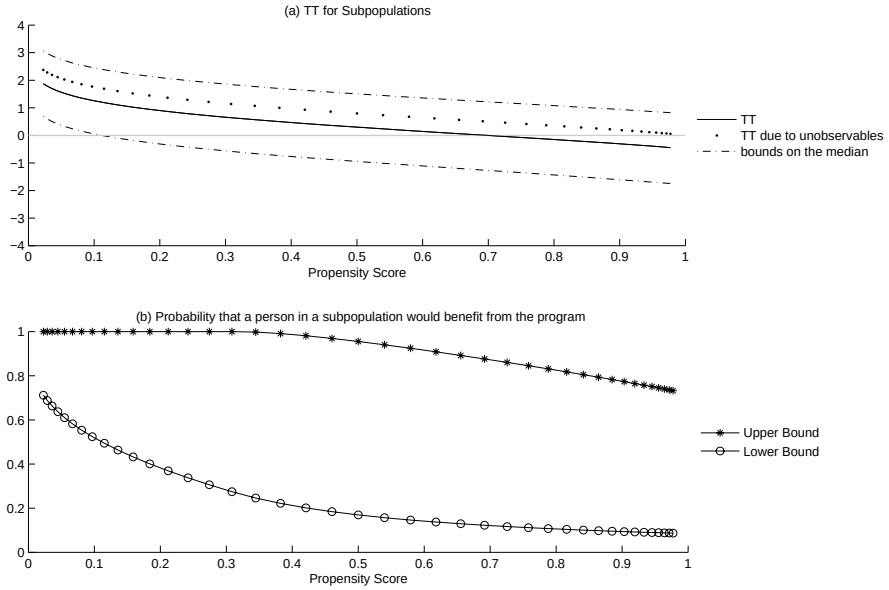

Figure 7: Treatment effect for the treated for subpopulations and the probability that a person in a subpopulation benefits from the treatment, where $ATE = -0.5$, $\rho_{1\varepsilon} = 0.5$, $\rho_{0\varepsilon} = -0.5$, and $\sigma_1 = \sigma_0 = 1$.

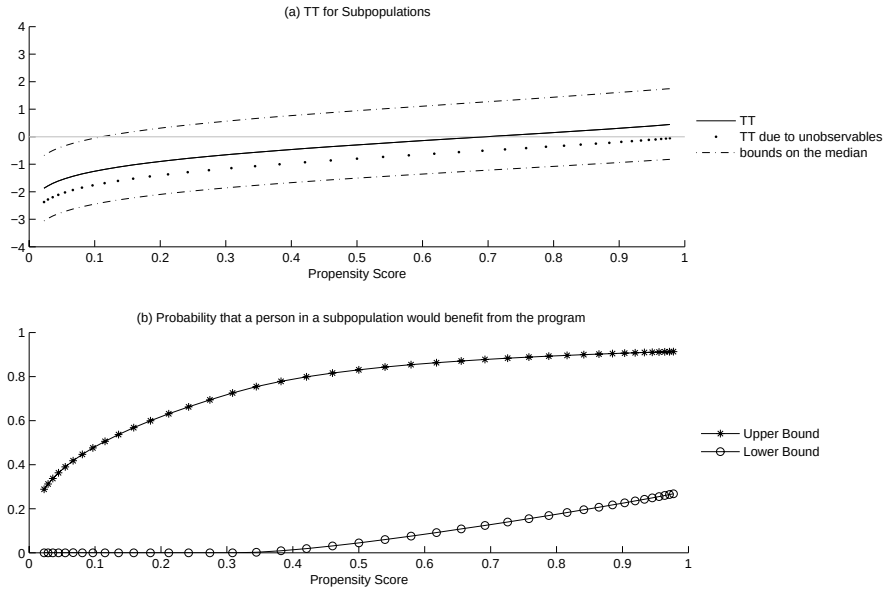

Figure 8: Treatment effect for the treated for subpopulations and the probability that a person in a subpopulation benefits from the treatment, where $ATE = 0.5$, $\rho_{1\varepsilon} = -0.5$, $\rho_{0\varepsilon} = 0.5$, and $\sigma_1 = \sigma_0 = 1$.

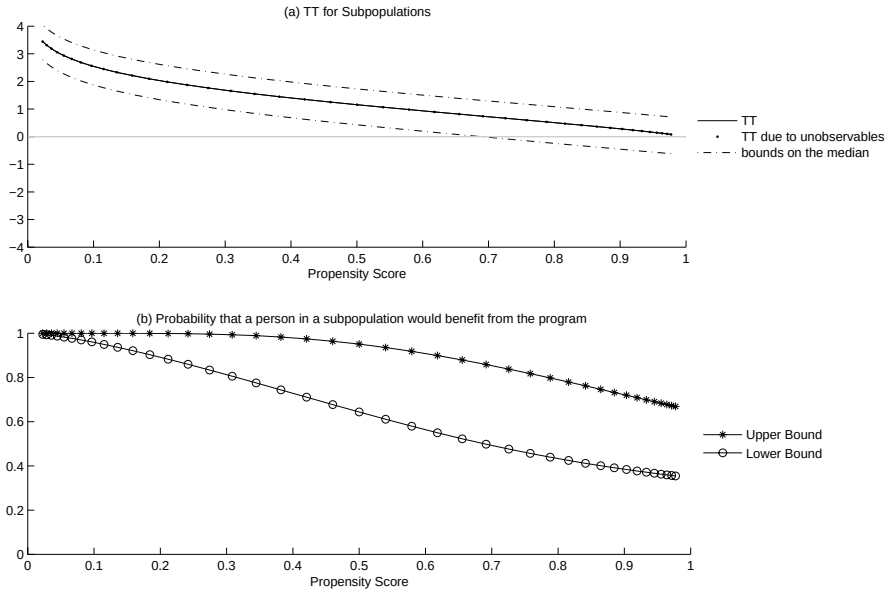

Figure 9: Treatment effect for the treated for subpopulations and the probability that a person in a subpopulation benefits from the treatment, where $ATE = 0$, $\rho_{1\varepsilon} = 0.95$, $\rho_{0\varepsilon} = -0.5$, and $\sigma_1 = \sigma_0 = 1$.

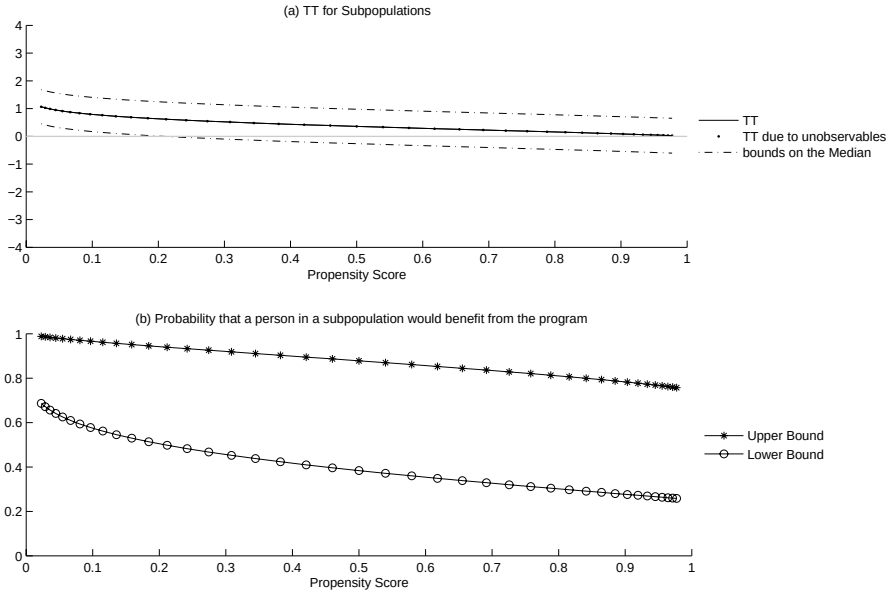

Figure 10: Treatment effect for the treated for subpopulations and the probability that a person in a subpopulation benefits from the treatment, where $ATE = 0$, $\rho_{1\varepsilon} = 0.95$, $\rho_{0\varepsilon} = 0.5$, and $\sigma_1 = \sigma_0 = 1$.

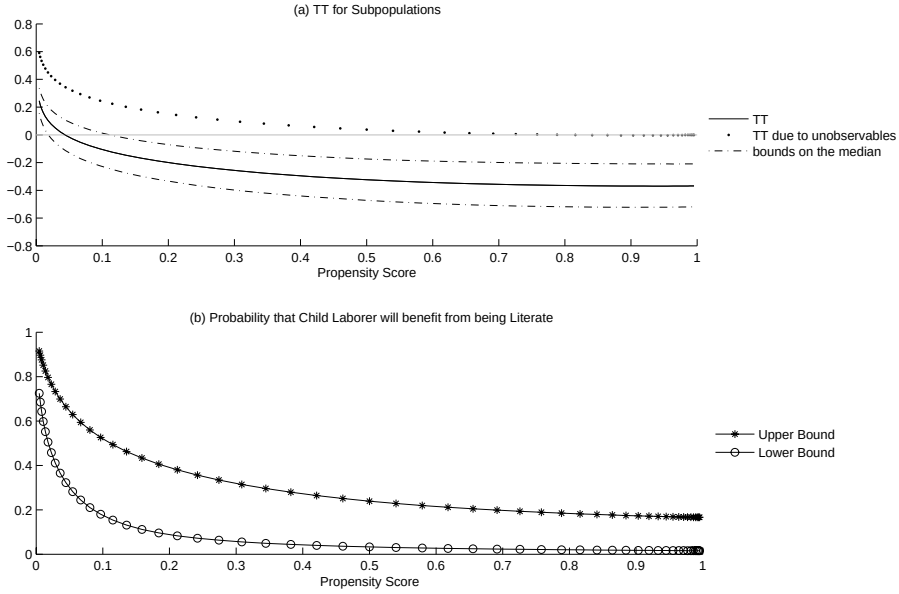

Figure 11: Treatment effect for the treated for subpopulations and the probability that a child laborer in a subpopulation benefits from being literate, where ATE =Conditional Mean, $\rho_{1\varepsilon} = 0.9503\,(\theta_1 = 5.239)$, $\rho_{0\varepsilon} = 0.7561\,(\theta_0 = 2.736)$, $\sigma_1 = 0.314$, $\sigma_0 = 0.322$.

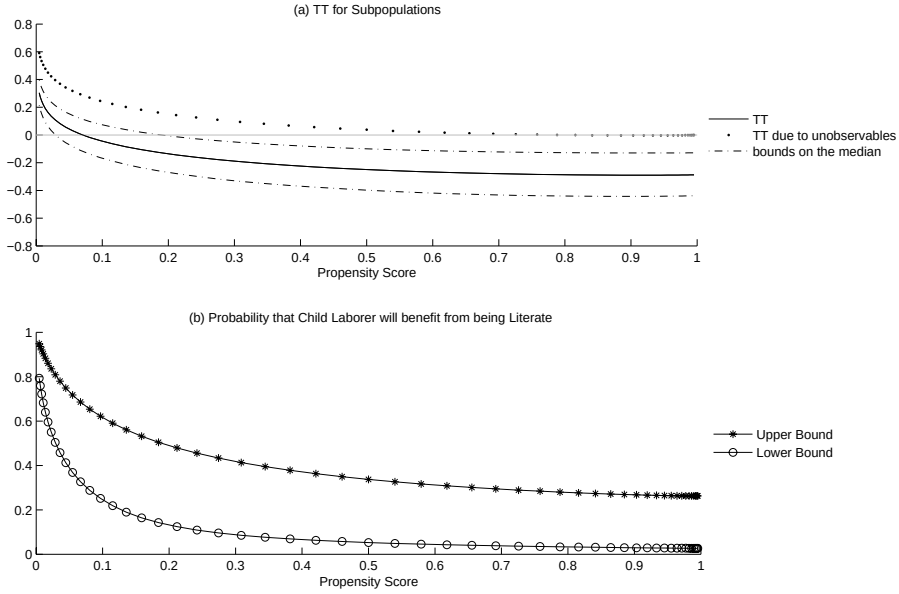

Figure 12: Treatment effect for the treated for subpopulations and the probability that a child laborer in a subpopulation benefits from being literate, where ATE =Maximum possible ATE (-0.2870), $\rho_{1\varepsilon} = 0.9503\,(\theta_1 = 5.239)$, $\rho_{0\varepsilon} = 0.7561\,(\theta_0 = 2.736)$, $\sigma_1 = 0.314$, $\sigma_0 = 0.322$.

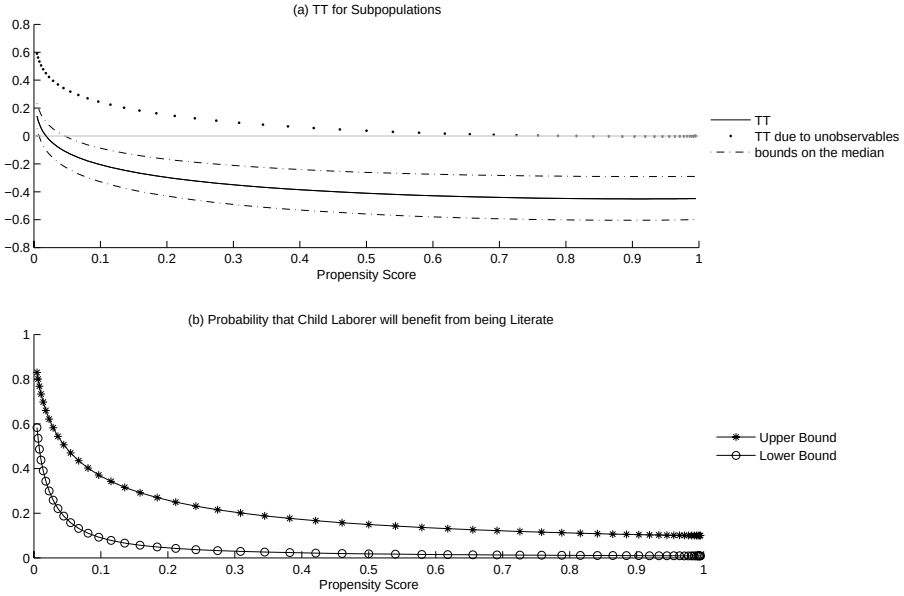

Figure 13: Treatment effect for the treated and the probability that a child laborer in a subpopulation benefits from being literate, where ATE =Minimum possible ATE (-0.4480), $\rho_{1\varepsilon} = 0.9503\,(\theta_1 = 5.239)$, $\rho_{0\varepsilon} = 0.7561\,(\theta_0 = 2.736)$, $\sigma_1 = 0.314$, $\sigma_0 = 0.322$.