NBER WORKING PAPER SERIES

SIMPLE ESTIMATORS FOR TREATMENT PARAMETERS IN A LATENT VARIABLE FRAMEWORK WITH AN APPLICATION TO ESTIMATING THE RETURNS TO SCHOOLING

James Heckman
Justin L. Tobias
Edward Vytlacil

Working Paper 7950
http://www.nber.org/papers/w7950

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2000

This research was supported by NSF 97-09-873 and NIH ROI-HD34958-01. The views expressed in this paper are those of the authors and not necessarily those of the National Bureau of Economic Research.

Simple Estimators for Treatment Parameters in a Latent Variable Framework with an Application to Estimating the Returns to Schooling

James Heckman, Justin L. Tobias, and Edward Vytlacil

NBER Working Paper No. 7950
October 2000
JEL No. C10, C34, I21

ABSTRACT

This paper derives simply computed closed-form expressions for the Average Treatment Effect (ATE), the effect of Treatment on the Treated (TT), Local Average Treatment Effect (LATE) and Marginal Treatment Effect (MTE) in a latent variable framework for both normal and non-normal models. The techniques presented in the paper are applied to estimating a variety of treatment parameters capturing the returns to a college education for various populations using data from the National Longitudinal Survey of Youth (NLSY).

James J. Heckman
Department of Economics
University of Chicago
1126 E. 59th Street
Chicago IL 60637

Justin L. Tobias
Department of Economics
University of California--Irvine
3151 Social Science Plaza
Irvine CA 92697-5100

Edward Vytlacil
Department of Economics
Stanford University
579 Serra Mall
Stanford CA 94305

1 Introduction

The problem of evaluating the effectiveness of a social program or a "treatment" is a central problem in social science and medicine. The problem of selection bias potentially arises in any evaluation. Individuals observed participating in a program or receiving treatment often possess different characteristics than an average person. Evaluating the economic return to a program requires accounting for the non-random assignment of individuals into the treated and untreated states.

One popular approach for dealing with selection bias, introduced in Gronau (1974) and Heckman (1976), is to specify a latent index model which relates the rule for assigning individuals to treatment to the potential treatment outcomes. The latent index has the interpretation of the expected net utility derived from receiving treatment; individuals participate in a program if net utility is positive (or non-negative) and do not participate if net utility is negative. This approach is based on assumptions about error distributions and allows for dependence between the errors in outcome and choice equations. While computationally convenient, this approach has been criticized for its reliance on distributional assumptions and lack of robustness to departures from normality (Goldberger (1983) and Paarsch(1984), and later work by Glynn, Laird and Rubin (1986)).

In response to these criticisms, recent analysts have adopted a more robust approach and have attempted to identify and estimate various treatment parameters without imposing strong distributional assumptions (see, for example, the LATE analysis of Imbens and Angrist (1994)). While these methods are free of parametric distributional assumptions, they typically estimate only one treatment parameter and are quite limited in the range of policy questions they can answer (Heckman and Vytlacil (2000b)). Further, the assumptions imposed in LATE analysis are actually equivalent to those required to specify a nonparametric selection model (Vytlacil (1999)).

This paper uses a latent variable framework to unite the recent treatment effect literature with the classical selection bias literature. We obtain simple closed-form expressions for four treatment

parameters of interest: the Average Treatment Effect (ATE), the effect of Treatment on the Treated (TT), the Local Average Treatment Effect (LATE) (Imbens and Angrist (1994)), and the Marginal Treatment Effect (MTE) (Heckman (1997), Heckman and Vytlacil (1999, 2000a-b)). Our analysis is motivated by the observation that despite the recent advances in flexible estimation of selection models (see, for example, Ahn and Powell (1993)), simple two-step correction procedures continue to dominate applied work on this topic (see, e.g. Tunali (2000)).

Since robustness of estimates to maintained distributional assumptions is an important problem, we present closed-form solutions for the four treatment parameters for non-normal models using flexible specifications for the selection equation, allowing the error terms to follow a trivariate Student-t_v distribution. For these generalized selection models, we derive closed-form expressions for the various treatment parameters and show how they can be consistently estimated by two-step methods. This simple generalization allows for considerable departures from normality, and thus offers an alternative to the standard selection model without increasing the computational burden.

The performance of the techniques developed in this paper are evaluated in Monte Carlo experiments. These simulations reveal the flexibility of our approach and assess the performance of a widely used model selection procedure due to Amemiya (1980). Using the NLSY data, we investigate the role of self-selection into higher education and its impact on estimated returns to schooling.

The plan of this paper is as follows. In the next section, we present a general model of potential outcomes, and define and interpret the various treatment parameters within it. In Section 3, expressions for these parameters are derived assuming fully parametric specifications for the outcome and selection equations. We obtain results for the textbook selection model, and for generalizations of this model which yield simple closed-form solutions. In Section 4, results from some Monte Carlo experiments are presented. We report that when selection bias is an empirically important problem, the Amemiya model selection procedure is effective. When selection is not a feature of the data, it is not effective but all models produce essentially the same estimates of treatment parameters. Section 5 estimates various average gains in post-schooling earnings

through the receipt of some form of college education. Using data from the National Longitudinal Survey of Youth (NLSY) we present point estimates of ATE, TT, LATE and MTE. The paper concludes with a summary in Section 6.

2 Treatment Parameters in a Canonical Model

Consider a model of potential outcomes:

$$\begin{aligned} Y^1 &= X\beta^1 + U^1 \\ Y^0 &= X\beta^0 + U^0 \\ D^* &= Z\theta + U^D. \end{aligned} \tag{1}$$

The first two equations denote outcome equations in two possible "states" or "sectors" (college or non-college in our paper). Without loss of generality, we assume that the first state indexed by the "1" superscript represents the treated state and the "0" superscript denotes the untreated state. Each agent is observed in only one state, so that either Y^1 or Y^0 is observed for each person, but the pair (Y^1, Y^0) is never observed for a given person. What we would like to recover is information about various expected gains from the receipt of treatment, where the gain is denoted by $\Delta \equiv Y^1 - Y^0$.

Let $D(Z)$ denote the observed treatment decision, where $D(Z) = 1$ denotes receipt of treatment and $D(Z) = 0$ denotes nonreceipt. The variable D^* is a latent variable which generates $D(Z)$ according to a threshold crossing rule,

$$D(Z) = 1[D^*(Z) \geq 0] = 1[Z\theta + U^D \geq 0], \tag{2}$$

where $1[A]$ is the indicator function which takes the value 1 if the event A is true and the value 0 otherwise. In an extension of the Roy (1951) model, $D^* = Y^1 - Y^0 - C$, where C represents the cost of participating in the treated state, so that agents choose to receive treatment if the gain from participating in the program minus costs is non-negative. We also define the following counterfactual choice variables. For any z which is a potential realization of Z, we define the

variable $D(z) = 1[z\theta \geq U^D]$. $D(z)$ indicates whether or not the individual would have received treatment had her value of Z been externally set to z, holding her unobserved U^D constant. We require an exclusion restriction and denote by Z_k some element of Z which is not contained in X. By varying Z_k, we can manipulate an individual's probability of receiving treatment without affecting potential outcomes. Finally, we assume $(U^D\ U^1\ U^0)$ is independent of X and Z.

Letting Y denote observed earnings,

$$Y = DY^1 + (1 - D)Y^0. \tag{3}$$

This model has been called the switching regression model of Quandt (1972), Rubin's model (Rubin 1978), or the Roy model of income distribution (Roy (1951), Heckman and Honoré (1990)).[1] To illustrate how a model of this type can be applied to evaluate an interesting policy question, consider the problem of estimating the return to a college education. In this case, Y represents log earnings, Y^1 denotes the log earnings of college graduates and Y^0 denotes the log earnings of those not selecting into higher education. The latent index maps people into either the "college" (or treated) state and the "no-college" (or untreated) state. To estimate the return to college, we might estimate the expected college log wage premium for given characteristics X, $E(Y^1 - Y^0 \mid X)$.[2] In general, given the model described by (1) and (2), we would like to have methods for estimating various average gains to program participation. In this paper, we examine four such *treatment parameters*, which measure possibly different average gains to the receipt of treatment. These four parameters are the Average Treatment Effect (ATE), the effect of Treatment on the Treated (TT), the Local Average Treatment Effect (LATE), and the Marginal Treatment Effect (MTE). [3]

The Average Treatment Effect (ATE) is defined as the expected gain from participating in the program for a randomly chosen individual. As before, we let $\Delta \equiv Y^1 - Y^0$ denote the gain from program participation, and note that the average treatment effect conditional on $X = x$ can be expressed as:

$$\text{ATE}(x) = E(\Delta \mid X = x) = x(\beta^1 - \beta^0). \tag{4}$$

The average treatment effect evaluated at the random variable X is ATE(X). This defines the treatment parameter as a function of the characteristics X. We can obtain unconditional estimates

by integrating (4) over the distribution of X,

$$\text{ATE} = E(\Delta) = \int \text{ATE}(X) dF(X) \approx \frac{1}{n} \sum_{i=1}^{n} \text{ATE}(x_i), \tag{5}$$

where n is sample size. A conceptually different parameter is the effect of Treatment on the Treated (TT). This is the average gain from treatment for those that actually select into the treatment:

$$\begin{aligned} \text{TT}(x, z, D(z) = 1) &= E(\Delta \mid X = x, Z = z, D(z) = 1) \\ &= x(\beta^1 - \beta^0) + E(U^1 - U^0 \mid U^D \geq -z\theta, X = x, Z = z) \\ &= x(\beta^1 - \beta^0) + E(U^1 - U^0 \mid U^D \geq -z\theta), \end{aligned} \tag{6}$$

where the third equality follows from the assumption that $(U^D \; U^1 \; U^0)$ is independent of (X, Z). The value of the Treatment on the Treated parameter evaluated at the random variables (X, Z) is $\text{TT}(X, Z, D(Z) = 1)$. As with ATE, we can obtain an unconditional estimate by integrating over the joint distribution of X and Z for those who actually receive treatment. Letting n_t be the number of observations with $D_i = 1$, TT can be approximated as follows:

$$\begin{aligned} \text{TT} &= E(\Delta | D(Z) = 1) \\ &= \int \text{TT}(X, Z, D(Z) = 1) dF(X, Z | D(Z) = 1) \\ &\approx \frac{1}{n_t} \sum_{i=1}^{n} D_i \; TT(x_i, z_i, D(z_i) = 1). \end{aligned} \tag{7}$$

The Local Average Treatment Effect (LATE) of Imbens and Angrist (1994) estimates an average gain to program participation without explicitly specifying a latent variable framework or imposing a distributional assumption.[4] LATE is defined as the expected outcome gain for those induced to receive treatment through a change in the instrument from $Z_k = z_k$ to $Z_k = z'_k$. The variable Z_k is assumed to affect the treatment decision (is contained in Z in (1)), but not to affect the outcomes Y^1 and Y^0. Below and throughout this paper, we define the LATE parameter as a change in the index from $Z\theta = z\theta$ to $Z\theta = z'\theta$, where $z'\theta > z\theta$ and z and z' are identical except for their k^{th} coordinate. Because of the latent index structure in (1) and (2), we can equivalently define the treatment parameters in terms of the propensity score, $P(Z) = 1 - F_{U^D}(-Z\theta)$, where F_S denotes

the cdf of the random variable S. The LATE parameter is defined as follows:

$$\begin{aligned} \text{LATE}(D(z)=0, D(z')=1, X=x) &= E(\Delta \mid D(z)=0, D(z')=1, X=x) \\ &= x(\beta^1-\beta^0)+E(U^1-U^0 \mid -z'\theta \leq U^D \leq -z\theta, X=x) \\ &= x(\beta^1-\beta^0)+E(U^1-U^0 \mid -z'\theta \leq U^D \leq -z\theta), \end{aligned} \tag{8}$$

where the third equality follows from the assumption that $(U^D\ U^1\ U^0)$ is independent of (X, Z). There are two ways to define the unconditional version of LATE. First, consider

$$\begin{aligned} E(\Delta|D(z)=0, D(z')=1) &= \int \text{LATE}(D(z)=0, D(z')=1, X)dF(X) \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\text{LATE}(D(z)=0, D(z')=1, x_i). \end{aligned} \tag{9}$$

The parameter $E(\Delta|D(z)=0, D(z')=1)$ corresponds to the treatment effect for individuals who would not select into treatment if their vector Z was set to z but would select into treatment if Z was set to z'. An alternative definition of the unconditional version of LATE is as follows. Let $Z^0(Z)$ equal Z but with the kth element replaced by z_k. Let $Z^1(Z)$ equal Z but with the kth element replaced by z'_k. In this notation the second definition of the unconditional version of LATE,

$$\begin{aligned} E(\Delta|D(Z^0(Z))=0, D(Z^1(Z))=1) &= \int \text{LATE}(D(Z^0(Z))=0, D(Z^1(Z))=1, X)dF(X,Z) \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\text{LATE}(D(Z^0(z_i))=0, D(Z^1(z_i))=1, x_i). \end{aligned} \tag{10}$$

This parameter corresponds to the treatment effect for individuals who would not select into treatment if the kth component of the Z vector is set to z_k (all other components of Z unchanged) but would select into treatment if the kth component of the Z vector is set to z'_k (all other components of Z unchanged).

The Marginal Treatment Effect (MTE) (Heckman (1997), Heckman and Smith (1998), Heckman and Vytlacil (1999, 2000a-b)) is the treatment effect for individuals with a given value of U^D,

$$\begin{aligned} \text{MTE}(x, u^D) &= E(\Delta|X=x, U^D=u^D) \\ &= x(\beta^1-\beta^0)+E(U^1-U^0 \mid U^D=u^D, X=x) \\ &= x(\beta^1-\beta^0)+E(U^1-U^0 \mid U^D=u^D) \end{aligned} \tag{11}$$

where the third equality follows from the assumption that $(U^D \ U^1 \ U^0)$ is independent of X. Evaluation of the MTE parameter at low values of u^D averages the outcome gain for those with unobservables making them least likely to participate, while evaluation of the MTE parameter at high values of u^D is the gain for those individuals with unobservables which make them most likely to participate. Since X is independent of U^D, the MTE parameter unconditional on observed covariates can be written as

$$\text{MTE}(u^D) = \int \text{MTE}(X, u^D) dF(X) \approx \frac{1}{n} \sum_{i=1}^{n} \text{MTE}(x_i, u^D).$$

The MTE parameter can also be expressed as the limit form of the LATE parameter,

$$\begin{aligned} \lim_{z\theta \to z'\theta} \text{LATE}(x, D(z) = 0, D(z') = 1) &= x(\beta^1 - \beta^0) + \lim_{z\theta \to z'\theta} E(U^1 - U^0 \mid -z'\theta \leq U^D \leq -z\theta, X = x) \\ &= x(\beta^1 - \beta^0) + E(U^1 - U^0 \mid U^D = -z'\theta) \\ &= \text{MTE}(x, -z'\theta). \end{aligned}$$

Thus the MTE parameter measures the average gain in outcomes for those individuals who are just indifferent to the receipt of treatment when the $z\theta$ index is fixed at the value $-u^D$.

The four parameters define different average gains to program participation if U^D is not (mean) independent of $U^1 - U^0$ but the four parameters are identical if U^D is mean independent of $U^1 - U^0$ conditional on $X = x$. In this paper, we derive closed-form solutions and simple estimators for these four parameters given certain distributional assumptions for the error terms. These expressions enable researchers to obtain estimates of the various treatment effects using simple methods.

3 Simple Expressions for the Different Treatment Parameters

This section derives expressions for ATE, TT, LATE, and MTE as given in (4) - (11) using two different assumptions regarding the distribution of the unobservables. Estimates of the treatment parameters can be obtained by using the output from a two-step procedure. We begin with the textbook selection model[5] and then present flexible non-normal models that possess the computational simplicity of the normal model.

3.1 Results for the "Textbook" Model

We first present expressions for the textbook normal model:

$$\begin{bmatrix} U^D \\ U^1 \\ U^0 \end{bmatrix} \sim N\left(0, \begin{bmatrix} 1 & \sigma_{1D} & \sigma_{0D} \\ \sigma_{1D} & \sigma_1^2 & \sigma_{10} \\ \sigma_{0D} & \sigma_{10} & \sigma_0^2 \end{bmatrix}\right).$$

The variance parameter in the selection equation is normalized to unity without loss of generality. For all of the values of the parameters, ATE reduces to the form given in (4). Under the normality assumption, the expression for Treatment on the Treated (TT) is:

$$\text{TT}(x, z, D(z) = 1) = x(\beta^1 - \beta^0) + (\rho_1\sigma_1 - \rho_0\sigma_0)\frac{\phi(z\theta)}{\Phi(z\theta)},$$

where $\rho_i \equiv \text{Corr}(U^i, U^D),\ i = 0, 1$. Under the normalization that the variance of the disturbance term in the selection equation is unity, $\rho_i\sigma_i = \sigma_{iD}$. As previously noted, under independence between U^D and $(U^1 - U^0)$ all treatment parameters are the same. Thus, if $\text{Cov}(U^1 - U^0, U^D) = 0$, or $\rho_1\sigma_1 = \rho_0\sigma_0$, Treatment on the Treated reduces to ATE in (4). In this case, people are not selecting into program on the basis of their unobserved (by the econometrician) gain, and all the treatment parameters reduce to ATE. If $\text{Cov}(U^1 - U^0, U^D) > 0$, then TT > ATE. If this condition is true, people are selecting into treatment on the basis of their idiosyncratic gain to treatment, and thus the gain from program participation for those observed in the treated state will exceed the gain for the average person. Also note that as $z\theta \to \infty$, TT $\to$ ATE. In this case, the probability of receiving treatment is one given the observable characteristics $Z = z$ and thus there is no selection problem. In this case, the conditioning information $D = 1$ is redundant given the characteristics $Z = z$ and thus the two parameters in (4) and (6) are equal.

Using standard results (see e.g. Cramer (1946) or Johnson, Kotz and Balakrishnan (1992)), the LATE parameter can easily be derived using the fact that if $(y, z) \sim N(\mu_y, \mu_z, \sigma_y, \sigma_z, \rho)$ and $b > a$

$$E(y \mid a \leq z \leq b) = \mu_y + \rho\sigma_y\left(\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\right),$$

where $\alpha = (a - \mu_z)/\sigma_z$, $\beta = (b - \mu_z)/\sigma_z$, so

$$\begin{aligned} \text{LATE}(x, D(z) = 0, D(z') = 1) &= E(Y_1 - Y_0 \mid x, z\theta < U^D < z'\theta)) \\ &= x(\beta^1 - \beta^0) + (\rho_1\sigma_1 - \rho_0\sigma_0)\frac{\phi(z'\theta) - \phi(z\theta)}{\Phi(z'\theta) - \Phi(z\theta)}. \end{aligned} \tag{12}$$

The Marginal Treatment Effect

$$\begin{aligned} MTE(x, u^D) &= x(\beta^1 - \beta^0) + E(U^1 - U^0 | U^D = u^D) \\ &= x(\beta^1 - \beta^0) + E(U^1 | U^D = u^D) - E(U^0 | U^D = u^D) \\ &= x(\beta^1 - \beta^0) + (\rho_1 \sigma_1 - \rho_0 \sigma_0) u^D. \end{aligned}$$

It is the limit form of LATE, [6]

$$\begin{aligned} MTE(x, u^D) &= x(\beta^1 - \beta^0) + (\rho_1 \sigma_1 - \rho_0 \sigma_0) \lim_{t \to -u^D} \left[\frac{\phi(-u^D) - \phi(t)}{\Phi(-u^D) - \Phi(t)} \right] \\ &= x(\beta^1 - \beta^0) + (\rho_1 \sigma_1 - \rho_0 \sigma_0) \lim_{t \to -u^D} \left[\frac{\left(\phi(-u^D) - \phi(t) \right) / (-u^D - t)}{\left(\Phi(-u^D) - \Phi(t) \right) / (-u^D - t)} \right] \\ &= x(\beta^1 - \beta^0) + (\rho_1 \sigma_1 - \rho_0 \sigma_0) u^D. \end{aligned} \tag{13}$$

Evaluating MTE when u^D is large corresponds to the case where the average outcome gain is evaluated for those individuals with unobservables making them most likely to participate, (and conversely when u^D is small). When $u^D = 0$, MTE = ATE as a consequence of the symmetry of the normal distribution. We next consider non-normal models.

3.2 Extensions to Non-Normal Models

We first note that the trivariate normal case presented in the previous section can be generalized by exploiting the natural flexibility of the selection equation. In the latent variable framework, the selection rule assigns people to the treated state ($D_i = 1$) provided $U_i^D \geq -Z_i\theta$. This is equivalent to setting $D_i = 1$ when $J(U_i^D) \geq J(-Z_i\theta)$ for some strictly increasing function J.[7]

Suppose that $U^D \sim F$, where F is an absolutely continuous distribution function which can be non-normal. For simplicity assume symmetry of U^D about zero so that $F(-a) = 1 - F(a)$. This model trivially maps into an equivalent model where the normal results apply. Define $\tilde{U}^D \equiv J_\Phi(U^D)$, and let $J_\Phi(u) \equiv \Phi^{-1}F(u)$. Clearly, J_Φ is left-continuous and strictly increasing and $J_\Phi(-u) = -J_\Phi(u)$ given the assumed symmetry of F. The transformed variable, $\tilde{U}^D$, is easily seen to be a standard normal random variable. Thus, the original model in (1) is *equivalent* to

the transformed model:

$$\begin{aligned} Y^1 &= X\beta^1 + U^1 \\ Y^0 &= X\beta^0 + U^0 \\ D_i^{**} &= J_\Phi(Z\theta) + \tilde{U}^D \end{aligned} \tag{14}$$

where we now assume that the transformed error vector $[\tilde{U}_D,\ U^1,\ U^0]'$ is trivariate normal so we can again use the normal framework. We thus obtain the following selection-corrected conditional mean functions:

$$E(Y^1 \mid D(Z) = 1, X = z, Z = z) = x\beta^1 + \rho_1\sigma_1\frac{\phi\left(J_\Phi(z\theta)\right)}{F(z\theta)}, \tag{15a}$$

and

$$E(Y^0 \mid D(Z) = 0, X = x, Z = z) = x\beta^0 - \rho_0\sigma_0\frac{\phi\left(J_\Phi(z\theta)\right)}{1 - F(z\theta)}, \tag{15b}$$

and obtain the treatment parameters:[8]

$$TT(x, z, D(z) = 1) = x(\beta^1 - \beta^0) + (\rho_1\sigma_1 - \rho_0\sigma_0)\frac{\phi\left(J_\Phi(z\theta)\right)}{F(z\theta)}, \tag{16}$$

$$LATE(x, D(z) = 0, D(z') = 1) = x(\beta^1 - \beta^0) + (\rho_1\sigma_1 - \rho_0\sigma_0)\frac{\phi\left(J_\Phi(z'\theta)\right) - \phi\left(J_\Phi(z\theta)\right)}{F(z'\theta) - F(z\theta)}, \tag{17}$$

and

$$MTE(x, u^D) = x(\beta^1 - \beta^0) + (\rho_1\sigma_1 - \rho_0\sigma_0)J_\Phi(u^D). \tag{18}$$

If $F = \Phi$, we obtain the normal model presented in Section 3.1. This model trivially generalizes the trivariate normal model and is applicable if there is concern that the errors in the selection equation are non-normal. All the parameters necessary for estimation of the treatment parameters for given x and z can be consistently estimated using a standard two-step procedure.

A less straightforward generalization can be achieved if we follow Lee (1982, 1983) and allow the error terms in (1) or (14) to be jointly distributed according to the Student-t_v distribution. By varying the degrees of freedom parameter, v, he produces a flexible class of models which can depart quite significantly from the textbook normal case. Since many of these parameters

are defined in terms of the tail behavior of the error terms, the family of t_v distributions offers a very attractive and potentially more appropriate class of models for the treatment parameters than those implied by the benchmark normal model, especially since wage data tend to be fat tailed (see e.g. Lydall (1968)).[9] We are able to obtain closed-form expressions for the various treatment parameters in the Student-t_v case, and can also estimate these expressions using output from simple two-step procedures.

Let $t_v(\mu, \text{-})$ denote the multivariate Student-t_v density function with mean μ, scale matrix -(variance equal to $[v/(v-2)]\text{-}^{-1}$) and v degrees of freedom.[10] We retain the notation used to define the covariance matrix for the normal model, and parameterize the scale matrix - in the same fashion. Finally, let t_v denote the standardized univariate Student t_v density with mean 0 and scale parameter equal to 1, and let T_v denote the associated cdf. To obtain expressions for the treatment parameters and derive the appropriate two-step estimators, we need to evaluate the truncated mean $E(U^D \mid U^D > -u)$ when U^D has a univariate t_v distribution. As shown in Raiffa and Schlaifer (1961) if $U^D \sim t_v$,

$$E(U^D \mid U^D > -u) = \left(\frac{v+u^2}{v-1}\right)\frac{t_v(u)}{T_v(u)}, \quad v > 1. \tag{19}$$

Using this result, now derive the treatment parameters for the more general model. To ensure that $\tilde{U}^D$ has a t_v density in the general case when $U^D \sim F$, we define $J_{T_v}(u) \equiv T_v^{-1}(F(u))$, again noting that $J_{T_v}(-u) = -J_{T_v}(u)$. We then assume that for this transformed model, $[\tilde{U}^D,\ U^1,\ U^0]'$ has a trivariate $t_v(0, \text{-})$ density. Given (19), we obtain the selection-corrected conditional mean functions:

$$E(Y^1 \mid D(Z)=1, X=x, Z=z) = x\beta^1 + \rho_1\sigma_1\left[\left(\frac{v+[J_{T_v}(z\theta)]^2}{v-1}\right)\left(\frac{t_v(J_{T_v}(z\theta))}{F(z\theta)}\right)\right], \tag{20a}$$

and

$$E(Y^0 \mid D(Z)=0, X=x, Z=z) = x\beta^0 - \rho_0\sigma_0\left[\left(\frac{v+[J_{T_v}(z\theta)]^2}{v-1}\right)\left(\frac{t_v(J_{T_v}(z\theta))}{1-F(z\theta)}\right)\right]. \tag{20b}$$

For convenience in notation, define the following function:

$$g(u,v) \equiv \left(\frac{v+[J_{T_v}(u)]^2}{v-1}\right)t_v(J_{T_v}(u)).$$

In this notation, the following expressions for the three treatment effects are easily derived:[11]

$$TT(x,z,D(z)=1) = x(\beta^1-\beta^0) + (\rho_1\sigma_1 - \rho_0\sigma_0)\frac{g(z\theta,v)}{F(z\theta)}. \tag{21}$$

$$LATE(x, D(z) = 0, D(z') = 1) = x(\beta^1 - \beta^0) + (\rho_1\sigma_1 - \rho_0\sigma_0)\frac{g(z'\theta, v) - g(z\theta, v)}{F(z'\theta) - F(z\theta)}. \tag{22}$$

$$MTE(x, u^D) = x(\beta^1 - \beta^0) + (\rho_1\sigma_1 - \rho_0\sigma_0)J_{T_v}(u^D). \tag{23}$$

For a given function F, the Student-t_v case converges to the case previously analyzed as $v \to \infty$, since the t_v density and cdf approach the normal for large v and thus the term in parentheses in the definition of $g(u, v)$ approaches one as $v \to \infty$. The difference in the expressions for the various treatment parameters across the normal and Student-t_v cases is determined by the difference in the selection correction terms. In the textbook normal model, these are the well-known Mills ratio terms, while in the Student-t_v case, these terms take the form $g(z\theta, v)/F(z\theta)$ and $g(z\theta, v)/(1 - F(z\theta))$. Figure 1 plots the truncated means $E(\tilde{U}^D \mid \tilde{U}^D > -J(u))$. The plots are labeled in the figure according to how J and the outcome variables are constructed. The first argument refers to the distribution assigned to the outcome errors and $\tilde{U}^D$, while the second argument following the "/" refers to the choice of link function (*i.e.*, the J function). In the "normal / normal" case, $J(u) = J_\Phi(u) = \Phi^{-1}\Phi(u) = u$, and the truncated mean reduces to the standard Mills ratio term: $\phi(u)/\Phi(u)$. In the $t_{v=2}$/ normal case, $J(u) = J_{T_{v=2}}(u) = T_{v=2}^{-1}(\Phi(u))$, and the truncated mean reduces to the expression used in (15a) and (15b).

The general models are quite flexible; the $t_{v=2}$ (or other low values v) results can depart quite significantly from the benchmark normal case and could produce treatment parameter estimates which are quite different from those obtained using normal errors. For large v, the results obtained are quite similar to those obtained for the normal case, as expected. In the following subsection, we present consistent estimators of the parameters of the models.

3.3 Estimation

A general recipe for obtaining two-step estimators of the various treatment parameters is as follows:

1. Obtain $\hat{\theta}$ from a binary choice model using F as the distribution of U^D.

2. Compute the appropriate selection correction terms evaluated at $\hat{\theta}$. In the classical normal selection model, these terms are $\phi(Z_i\hat{\theta})/\Phi(Z_i\hat{\theta})$ when $D_i = 1$, and $\phi(Z_i\hat{\theta})/(1-\Phi(Z_i\hat{\theta}))$ when $D_i = 0$. For the generalized models, the corresponding terms to be used are in (15a-b) or (20a-b).

3. Run treatment-outcome-specific regressions (for the groups $\{i : D_i = 1\}$ and $\{i : D_i = 0\}$) with the inclusion of the appropriate selection-correction terms obtained from the previous step.

4. Given $\hat{\beta}^0, \hat{\beta}^1, \rho_1\hat{\sigma}_1$ and $\rho_0\hat{\sigma}_0$ obtained from step 3 and $\hat{\theta}$ from step (1), use these parameter estimates to obtain point estimates of the treatment parameters for given X, Z, and Z'.[12] Standard errors can be obtained using the Delta method or the parametric bootstrap, as discussed below in Section 5.

4 Monte Carlo Simulations

In this section we assess the performance of a simple model selection procedure and also assess the performance of our treatment parameter estimators under correct and incorrect model specification. We obtain sampling distributions of the estimators of different treatment parameters using both generated normal and Student-t_v data. We show that the different models discussed in Section 3 can give different estimates of the various treatment effects. Further, we demonstrate the intuitively plausible result that our ability to correctly differentiate among competing models is increasing in the sample size and the degree of selectivity in the model.

The model that we employ in the experiments below is a basic selection model with few covariates, given as follows:

$$Y^1 = \alpha^1 + \alpha^2 + U^1 \tag{24}$$

$$Y^0 = \alpha^1 + U^0$$

$$D^* = \theta^0 + \theta^1 Z + U^D. \tag{25}$$

We generate the data by setting $\alpha^1 = 2$, $\alpha^2 = 1$, $\theta^0 = 0$, $\theta^1 = 1$, and $Z \sim N(0,1)$. With this structure, the average treatment effect is $\alpha_2 = 1$. For the first experiment we obtain a data set with 1,500 observations by drawing the error term vector from a trivariate normal distribution. Given these draws, we determine the individuals' treatment choice, and given this choice, calculate the observed value of y. For each simulated data set, we estimate the Marginal Treatment Effect and Treatment on the Treated for various values of Z and u^D. To introduce selection bias, the data are drawn such that $\rho_{1D} = .95$, $\rho_{0D} = .1$, We choose $\text{Var}(U^1) = \text{Var}(U^0) = .4$, normalize $\text{Var}(U^D) = 1$, and set the unidentified correlation coefficient between Y^1 and Y^0 equal to 0. New data sets are drawn 1,000 times given the specification and parameter values above, and for each iteration, values for the above treatment parameters are obtained and stored. Sampling distributions of these treatment parameters are then estimated by kernel smoothing the resulting 1,000 parameter estimates. Results obtained from the true model (the normal model) are compared with those obtained using the misspecified, heavy-tailed trivariate t_2 model in Figures 2 and 3.

In Figures 2 and 3, we see that the sampling distributions are centered around the correct values when the normal model is appropriate while the heavy-tailed t_2 misses the mark, and often places extremely small weight near the true values. Although not shown in the two figures, the degree of discrepancy between the normal and t_2 models increases as the parameter of interest moves farther into the tail of the distribution. For example, if the parameter of interest is MTE with $u^D = 2$, then the distribution of MTE associated with the t_2 model places virtually no mass on the correct values. In Figures 4 and 5, the same experiment is run, except the data are generated from a trivariate t_4 distribution. We then compare results from the true t_4 model to those obtained from the misspecified normal model. Again we see that the true model outperforms the misspecified model, and the normal results generally place small mass around the true value. Thus for parameters of interest such as TT and MTE, which are defined in the tails, the normal and Student-t_v results can give quite different predictions. Given this result, it is of some interest to present a way for choosing among competing models.

A simple model selection procedure (given equal numbers of parameters across the various models) is to obtain estimates of the selection-corrected conditional mean functions for a variety

of competing models and then select the one which minimizes the sum of squared residuals (SSR). This approach to model selection chooses the model whose conditional mean function provides the best fit to the observed data (Amemiya, (1980)). Formally, we choose the model m which minimizes the criterion:

$$\sum_{i=1}^{n} \left[(y_i - D_i \hat{m}(X_i, Z_i \mid D_i = 1) - (1 - D_i) \hat{m}(X_i, Z_i \mid D_i = 0) \right]^2 ,$$

where $\hat{m}(X_i, Z_i \mid D_i = 1)$ corresponds to the estimated selection-corrected conditional mean function in the treated state, and $\hat{m}(X_i, Z_i \mid D_i = 0)$ corresponds to this conditional mean function in the untreated state.

Several Monte Carlo experiments were conducted to examine the performance of this model selection procedure. We generated 1,000 data sets of sizes 50, 250, 500 and 1,000 and determined the probability of choosing the correct model for each sample size. These results are presented in Figure 6. The data are generated from a normal distribution, and we carry along the t_2 model as a competitor to the normal model. The experiments are repeated for three different correlation structures, each depicting varying degrees of the importance of selection bias.

The performance of the proposed model selection procedure improves with the sample size n, and also with the degree of selectivity in the model. With little role for selection bias, it is difficult to differentiate among the models, even with a fairly large sample size. However, distinguishing among the models may not be important in this case, since our treatment parameter estimates will be similar in the absence of selection bias, and controlling for self-selection may not be important to the evaluation of the given program. The results displayed in Figure 6 also suggest that one can assess the degree of confidence about the ability to differentiate among the competing models by investigating the empirical importance of selectivity. When selectivity is most important, the models discussed here will give different estimates of the treatment parameters. It is reassuring that our Monte Carlo analysis suggests that we can differentiate among these models using our MSE criterion given a reasonable sample size. When selectivity is not an important feature of the data, treatment parameter estimates across these models will be similar, and thus the problem of model selection is not important. For intermediate cases, where one is not confident about the ability to choose among competing models, yet estimates of the treatment parameters differ across

the models, one could place bounds on the treatment parameter estimates within the flexible class of models described in this paper.

5 The Returns to College

We next present estimates of the return to some form of college education using our flexible scheme. The problem of selection bias has long been recognized as important in assessing the returns to education (see, for example Willis and Rosen (1979)). We seek to provide robust yet simple estimates for various returns to schooling while controlling for self-selection into higher education. Data are taken from the National Longitudinal Survey of Youth (NLSY). In our analysis, Y^1 denotes the log of 1991 hourly earnings for those individuals completing at least 13 years of schooling by 1991, and Y^0 is the log of hourly wages for those with 12 or fewer years of schooling. The sample is restricted to white males who are not enrolled in school in the current year and report hourly earnings between \$1 and \$100. Observations are also deleted when other explanatory variables used in the analysis are missing, resulting in a final sample of 1,230 observations.

The variables in X include an intercept, two indicators for residence in the Northeast and South,[13] potential labor market experience and its square,[14] an indicator for residence in an urban area, the local unemployment rate in 1990, and a measure of "ability" denoted as g. This ability measure is constructed from the 10 component tests of the ASVAB (Armed Services Vocational Aptitude Battery) provided in the NLSY. Since people vary in age at the time of the test, each component test is first regressed on age. The residuals from this regression are then standardized, and g is defined as the first principal component of the standardized residuals.[15] We choose a parsimonious specification for the variables in the selection equation (Z), which includes an intercept, g, indicator variables denoting if the respondent's mother and father attended college, an indicator for residence in an urban area at age 18 and number of siblings. The last variable serves as our exclusion restriction and is assumed to affect the college entry decision without affecting post-schooling earnings.[16]

We obtain estimates of the four treatment parameters discussed in this paper using a variety of models. These include the "textbook" normal model, Student-t_v models with a logit link function, and Student-t_v models with a T_v link function. For the Student-t_v cases, results are obtained for $v \in \{2, 3, 4, 5, 6, 8, 12, 24\}$. For small values of v, results could potentially be quite different from those obtained from the normal model.

Point estimates of the Average Treatment Effect (ATE) are obtained by averaging the conditional treatment effects (given X) over the sample distribution of characteristics, as in equation (5). For Treatment on the Treated, point estimates are obtained as in (7) by averaging over the joint distribution of characteristics (given X and Z) for the subsample that actually selects into college. To estimate LATE, we average over the joint distribution of characteristics after setting the number of siblings variable in $Z = z$ equal to four, and equal to 0 in $Z = z'$ (this is the second form of the unconditional LATE parameter previously discussed). This estimates the average college log wage premium for persons induced to attend college when the number of siblings has been lowered from four to zero. Finally, for each value of U^D, we construct the Marginal Treatment Effect parameter not conditioning on observable characteristics by averaging $\text{MTE}(X, u^D)$ over the sample distribution of X characteristics. We plot the resulting Marginal Treatment Effect (MTE) parameter over values of U^D from -3 to 3 in Figure 7. Point estimates of the treatment parameters are scaled by the difference in average years of schooling across the college and no-college groups (≈ 3.8) to estimate the return to schooling. Large sample standard errors of the estimated treatment parameters are computed using the parametric bootstrap.[17]

Point Estimates of ATE, TT and LATE across the alternative models are presented in Table 1 of the appendix. We first see that the receipt of some form of college education tends to raise the hourly wage of a randomly selected person by 6-9 percent. For those who actually select into college, the results are lower, ranging from 2.8-4 percent. Point estimates of LATE are similar to ATE, and range from 5.3-7.9 percent. The similarity between LATE and ATE results from the fact that the change from 4 to 0 siblings does not significantly alter the propensity score, and thus this treatment parameter is very similar to the Average Treatment Effect. Figure 7 presents a plot of the MTE over the interval [-3,3] across a variety of models. As the degrees of freedom parameter

increases, the estimated MTE tends to approach what is obtained from the benchmark normal case. Further, the average treatment effects are obtained for the special cases where $u^D = 0$. The upward slope of the plots indicates that those individuals with unobservables making them least likely to attend college receive the highest percent increase in hourly wages, due to a negative selection effect. For the best-fitting normal model, we test and reject (with a t-statistic equal to -2.1) the hypothesis of a constant MTE $\text{Cov}(U^D, U^1 - U^0) = 0$, and conclude that selection is an important feature of this data. Marginal entrants get lower returns than those who precede them in attending college. Similar results are reported in Carneiro, Hansen, Heckman and Vytlacil (2000). The methods used here are easily implemented and can be applied to robustly estimate or bound a variety of policy-relevant average gains to program participation in the presence of selectivity bias.

6 Conclusion

This paper presents simple expressions for the parameters often used to evaluate the effectiveness of a given program or treatment: the Average Treatment Effect (ATE), the effect of Treatment on the Treated (TT), the Local Average Treatment Effect (LATE), and the Marginal Treatment Effect (MTE). These expressions were obtained for the "textbook" selection model, and also for generalizations of this model which enable departures from normality. The appeal of our approach is that practitioners can obtain consistent estimates of these parameters using a two-step estimator or a simple generalization of that estimator.

The modern approach to program evaluation focuses on the estimation of narrowly defined parameters without having to impose strong distributional assumptions. The approach adopted in this paper permits estimation of a variety of policy-relevant parameters as well as estimation of the four treatment effects listed above, rather than one or the other parameters featured in the recent treatment effect literature. We provide generalized yet computationally simple alternatives to the often-used and often criticized normal model. The approach presented in this paper maintains the flexibility of the structural model in terms of the number of parameters which we can estimate,

while relaxing the dependence on normality assumptions.

The methods presented in this paper are applied to estimate the returns to a college education. Using data from the NLSY, we obtain point estimates of ATE, TT, LATE, and MTE using both the two-step procedure and generalized two-step methods. The results suggest that for the flexible class of models analyzed, a college education raises hourly wages from 6-9 percent for a randomly selected person, and between 2.8-4 percent for those actually selecting into higher education.

7 Derivation of MTE: Student-t_v Case

Using the notation of Section 3.2, let $J_{T_v}(x) = T_v^{-1}(F(x))$. Note the following results:

$$t_v'(x) = -(v+1)x(v+x^2)^{-1}t_v(x) \quad \text{and} \quad J_{T_v}'(x) = \frac{f(x)}{t_v(J_{T_v}(x))}. \tag{A-1}$$

The last statement follows by noting

$$T_v(J_{T_v}(x)) = F(x).$$

By the chain rule,

$$\frac{\partial T_v(J_{T_v}(x))}{\partial x} = \frac{\partial T_v(J_{T_v}(x))}{\partial J_{T_v}(x)}\frac{\partial J_{T_v}(x)}{\partial x} = f(x),$$

so

$$\frac{\partial J_{T_v}(x)}{\partial x} = \frac{f(x)}{t_v(J_{T_v}(x))}.$$

Consider $\lim_{z\theta \to z'\theta} \text{LATE}(D(z)=0, D(z')=1, X=x)$,

$$\begin{aligned} & x(\beta_1-\beta_0) + (\rho_1\sigma_1 - \rho_0\sigma_0)\lim_{z\theta\to z'\theta}\frac{g(z'\theta,v)-g(z\theta,v)}{F(z'\theta)-F(z\theta)} \\ = \; & x(\beta_1-\beta_0) + (\rho_1\sigma_1 - \rho_0\sigma_0)\lim_{z\theta\to z'\theta}\frac{(g(z'\theta,v)-g(z\theta,v))/(z'\theta-z\theta)}{(F(z'\theta)-F(z\theta))/(z'\theta-z\theta)} \\ = \; & x(\beta_1-\beta_0) + (\rho_1\sigma_1 - \rho_0\sigma_0)\frac{\partial g(z\theta,v)/\partial z\theta}{f(z\theta)}, \end{aligned}$$

since the limits exist and equal the derivatives of g (with respect to its first argument) and F. With $g(z\theta, v)$ defined as below(20), it follows that

$$\frac{\partial g(z\theta,v)}{\partial z\theta} = \frac{v+J_{T_v}^2(z\theta)}{v-1}\frac{\partial t_v(J_{T_v}(z\theta))}{\partial z\theta} + \frac{2J_{T_v}(z\theta)}{v-1}\frac{\partial J_{T_v}(z\theta)}{\partial z\theta}t_v(J_{T_v}(z\theta)).$$

Substituting the two results in (A-1) above and canceling terms, and using the relationship between MTE and the limit of the LATE parameter, we obtain

$$MTE(x,u^D) = x(\beta_1-\beta_0) + (\rho_1\sigma_1-\rho_0\sigma_0)J_{T_v}(u^D),$$

as claimed.

References

[1] Ahn, Hyungtaik and James Powell, "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics* 58 (1993), 3-29.

[2] Amemiya, Takeshi, "Selection of Regressors," *International Economic Review* 21 (1980), 331-354.

[3] Amemiya, Takeshi, *Advanced Econometrics* (Cambridge: Harvard University Press, 1985).

[4] Carneiro, Pedro, Karsten Hansen, James Heckman and Edward Vytlacil, "Marginal Treatment Effects and the Returns to Schooling," forthcoming, 2000.

[5] Cawley, John, Karen Coneely, James Heckman and Edward Vytlacil, "Cognitive Ability, Wages, and Meritocracy," in Devlin, Bernie, Stephen E. Fienberg, Daniel P. Resnick and Kathryn Roeder, (eds.), *Intelligence, Genes and Success: Scientists Respond to the Bell Curve* (New York: Springer, 1997), 178-192.

[6] Cramer, Harold, *Mathematical Methods of Statistics* (Princeton: Princeton University Press, 1946).

[7] Glynn, Robert, Nan Laird and Donald Rubin, "Selection Models Versus Mixture Modeling with Nonignorable Nonresponse," in Wainer, Howard, (ed.), *Drawing Inference from Self-Selected Samples* (New York: Springer, 1986).

[8] Goldberger, Arthur, "Abnormal Selection Bias," in Karlin, Samuel, Takeshi Amemiya and Leo Goodman, (eds.), *Studies in Econometrics, Time Series, and Multivariate Statistics* (New York: Academic Press, 1983).

[9] Gronau, Reuben, "Wage Comparisons - A Selectivity Bias," *Journal of Political Economy* 82:6 (1974), 1119-1143.

[10] Heckman, James, "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5 (1976), 475-492.

[11] Heckman, James, "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources* 32 (1997), 441-462.

[12] Heckman, James and Bo Honoré, "The Empirical Content of the Roy Model," *Econometrica* 50 (1990), 1121-1149.

[13] Heckman, James and Jeffrey Smith, " Evaluating the Welfare State," in Strom, S. (ed.), *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*, Econometric Society Monograph Series (Cambridge: Cambridge University Press, 1998).

[14] Heckman, James and Edward Vytlacil, "Instrumental Variable Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling with the Return is Correlated with Schooling," *Journal of Human Resources* 33 (1998), 974-987.

[15] Heckman, James and Edward Vytlacil, "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences* 96 (1999), 4730-4734.

[16] Heckman, James and Edward Vytlacil, "The Relationship Between Treatment Parameters within a Latent Variable Framework," *Economics Letters* 66 (2000a), 33-39.

[17] Heckman, James and Edward Vytlacil, "Local Instrumental Variables," in Hsiao, C., K. Morimune, and J. Powell, (eds.), *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya* (Cambridge: Cambridge University Press, 2000b).

[18] Imbens, Guido and Joshua Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62 (1994), 467-476.

[19] Johnson, Norman, Samuel Kotz and N. Balakrishnan, *Continuous Univariate Distributions* (New York: John Wiley and Sons, 1992).

[20] Johnston, John and John DiNardo, *Econometric Methods* (New York: Magraw-Hill, 1997).

[21] Lee, Lung-Fei, "Unionism and Wage Rates: A Simultaneous Model with Qualitative and Limited Dependent Variables," *International Economic Review* 19:2 (1979), 415–433.

[22] Lee, Lung-Fei, "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies* 49:3 (1982), 355-372.

[23] Lee, Lung-Fei, "Generalized Econometric Models With Selectivity," *Econometrica* 51:2 (1983), 507-512.

[24] Lydall, Harold, *The Structure of Earnings*, (Oxford: Clarendon Press, 1968).

[25] Paarsch, Harry J., "A Monte Carlo Comparison of Estimators for Censored Regression Models," *Journal of Econometrics* 24 (1984), 197-213.

[26] Quandt, Richard, "Methods for Estimating Switching Regressions," *Journal of the American Statistical Association* 67:338 (1972), 306-310.

[27] Raiffa, Howard and Robert Schlaifer, *Applied Statistical Decision Theory* (Boston: Graduate School of Business, Harvard University, 1961).

[28] Roy, A. D., "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers* 3 (1951), 135-146.

[29] Rubin, Donald, "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics* 6 (1978), 34-58.

[30] Tunali, Insan, "Rationality of Migration," forthcoming in *International Economic Review* (2000).

[31] Tobias, Justin L., "Three Essays on Bayesian Inference in Econometrics with an Application to Estimating the Returns to Schooling Quality," Ph.D. Dissertation, Department of Economics, University of Chicago (1999).

[32] Tobias, Justin L., "Are Returns to Schooling Really Concentrated Among the Most Able? A New Look at the Ability-Earnings and Ability-Schooling Relationships," UC-Irvine Department of Economics Working Paper (2000).

[33] Vytlacil, Edward, "Independence, Monotonicity, and Latent Variable Models: An Equivalence Result," working paper, University of Chicago (1999).

[34] Willis, Robert and Sherwin Rosen, "Education and Self-Selection," *Journal of Political Economy* 87:5 (1979), S7-36.

Notes

[1]Amemiya (1985) has classified models of this type as generalized tobit models, and refers to the model in (1) as the Type 5 tobit model.

[2]Other applications which fit directly into this model include Lee (1979) and Willis and Rosen (1979).

[3]For a more general discussion of the parameters and the relationship among them, see Heckman and Vytlacil (1999,2000a-b).

[4]The implications of the assumptions imposed in Imbens and Angrist (1994) which permit estimation of the LATE parameter have been examined by Vytlacil (1999). Vytlacil shows that the independence and monotonicity assumptions used by Angrist and Imbens imply a latent variable specification without parametric restrictions.

[5]Results for this case were first reported in Heckman and Vytlacil (2000b), although they present a more general analysis and do not discuss how estimates of these parameters can be obtained using simple two-step procedures.

[6]The last line follows from L'Hopital's rule.

[7]Lee (1982, 1983) uses this device.

[8]Henceforth, we do not discuss the ATE expression. In all cases ATE(x) is $x(\beta_1 - \beta_0)$.

[9]The fat tail for wages arises, in part, from measurement errors in earnings and hours and because wages are often defined by dividing earnings by hours.

[10]Of course, the mean exists when $v > 1$ and the variance exists when $v > 2$.

[11]The TT expression follows immediately from the result in (20). The LATE expression uses this result and an argument similar to the one used to derive the LATE parameter in the textbook normal case (see appendix). The expression for the MTE is derived in the appendix.

[12]Alternatively, one could integrate over the distribution of the characteristics to obtain unconditional estimates.

[13]The NLSY provides four regional variables - Northcentral, Northeast, South, and West.

[14]Potential experience is defined as Age - Years of Schooling - 6.

[15]For more on the construction and use of this ability measure, see Cawley *et al.* (1997) and Tobias (1999).

[16]The number of siblings variable was found to be a significant determinant of the college entry decision, but was not significant at the 5 percent level when included as a regressor in the outcome equations for the college and no-college states. Other variables, such as distance to college, the local unemployment rate at age 18 and a state-level tuition variable were also constructed and investigated as potential instruments. These variables were found to have surprisingly little power in explaining the college entry decision for this data and thus we selected number of siblings as

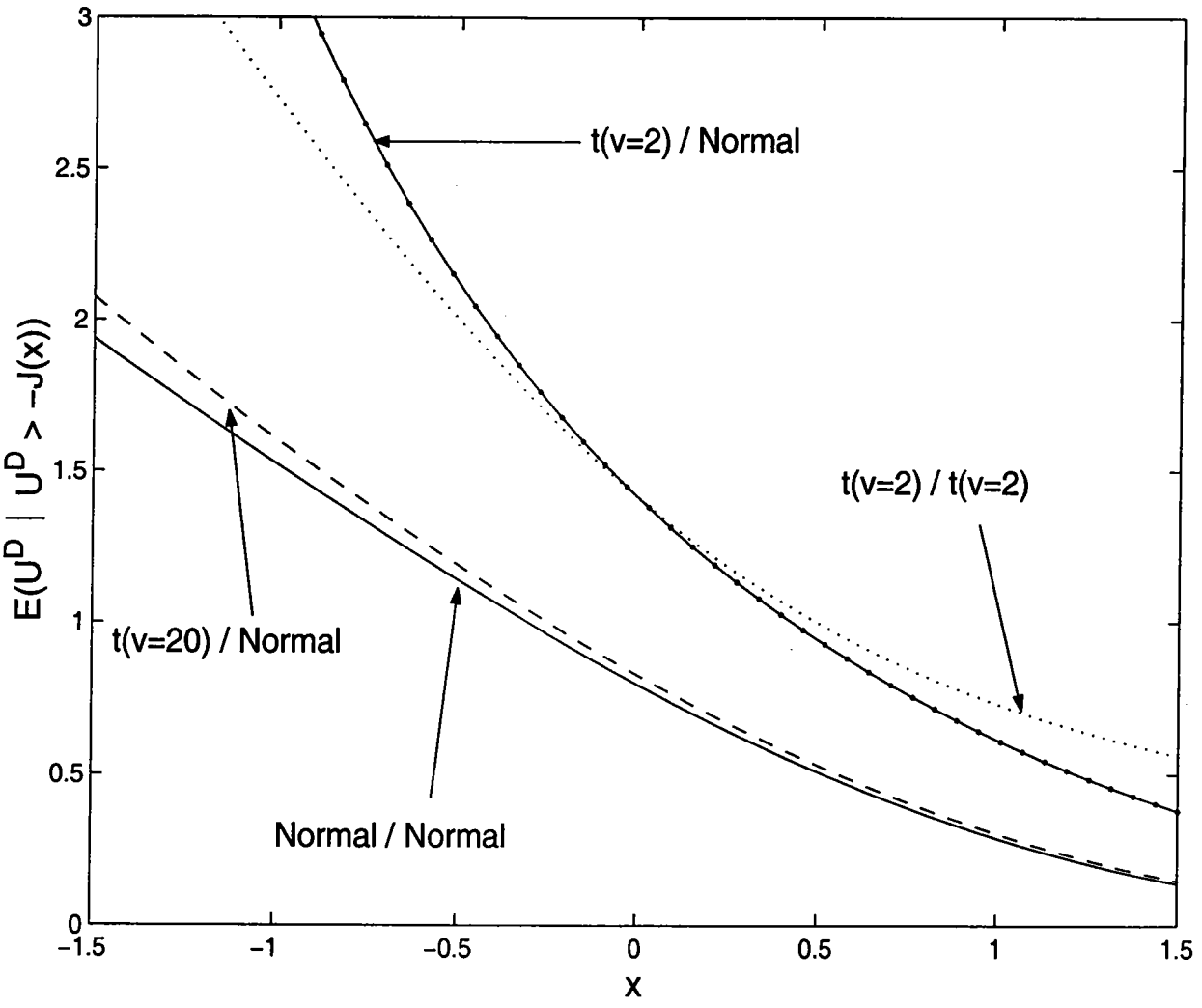
our instrument.

[17]See, for example, Johnston and DiNardo (1997). We obtain 500 draws from the asymptotic distribution of the regression parameters, and evaluate the treatment effects for each draw. Standard errors are computed as the standard deviation of the simulated values of the treatment effects.

Table 1
Point Estimates and Standard Errors of Alternate Treatment Parameters

Outcome Errors / Link Function	ATE	TT	LATE
Normal/Normal	.092	.039	.079
(SSR=345.25)	(.03)	(.04)	(.03)
$t_{v=2}$ / Logit	.061	.036	.053
(SSR = 346.09)	(.02)	(.03)	(.02)
$t_{v=3}$ / Logit	.073	.035	.062
(SSR = 345.79)	(.02)	(.03)	(.02)
$t_{v=4}$ / Logit	.079	.035	.067
(SSR = 345.61)	(.02)	(.04)	(.03)
$t_{v=5}$ / Logit	.082	.034	.069
(SSR = 345.51)	(.03)	(.04)	(.03)
$t_{v=6}$ / Logit	.084	.034	.071
(SSR = 345.44)	(.03)	(.04)	(.03)
$t_{v=8}$ / Logit	.085	.034	.073
(SSR = 345.36)	(.03)	(.04)	(.03)
$t_{v=12}$ / Logit	.087	.034	.073
(SSR = 345.29)	(.03)	(.04)	(.04)
$t_{v=24}$ / Logit	.088	.033	.075
(SSR = 345.23)	(.04)	(.04)	(.03)
$t_{v=2}$ / $t_{v=2}$	.067	.028	.058
(SSR = 345.68)	(.03)	(.04)	(.03)
$t_{v=3}$ / $t_{v=3}$	.075	.030	.063
(SSR = 345.56)	(.03)	(.04)	(.03)
$t_{v=4}$ / $t_{v=4}$	.079	.031	.066
(SSR = 345.48)	(.03)	(.04)	(.03)
$t_{v=5}$ / $t_{v=5}$	.082	.032	.069
(SSR = 345.43)	(.03)	(.04)	(.03)
$t_{v=6}$ / $t_{v=6}$	.084	.033	.070
(SSR = 345.40)	(.03)	(.04)	(.03)
$t_{v=8}$ / $t_{v=8}$	.086	.034	.072
(SSR = 345.36)	(.03)	(.04)	(.03)
$t_{v=12}$ / $t_{v=12}$	.088	.036	.075
(SSR = 345.32)	(.03)	(.04)	(.03)
$t_{v=24}$ / $t_{v=24}$	.090	.037	.077
(SSR = 345.29)	(.03)	(.04)	(.03)

Figure 1: $E(\tilde{U}^D \mid \tilde{U}^D > -J(u))$ for various Specifications of the Outcome Disturbances / and Link Function



Distributions of Treatment on the Treated and Marginal Treatment Effects Using Normal and t_2 Models. Generated NORMAL Data. 1,000 Replications with N = 1,500.

Figure 2: Treatment on the Treated with $Z = -2$. True Value ≈ 2.28

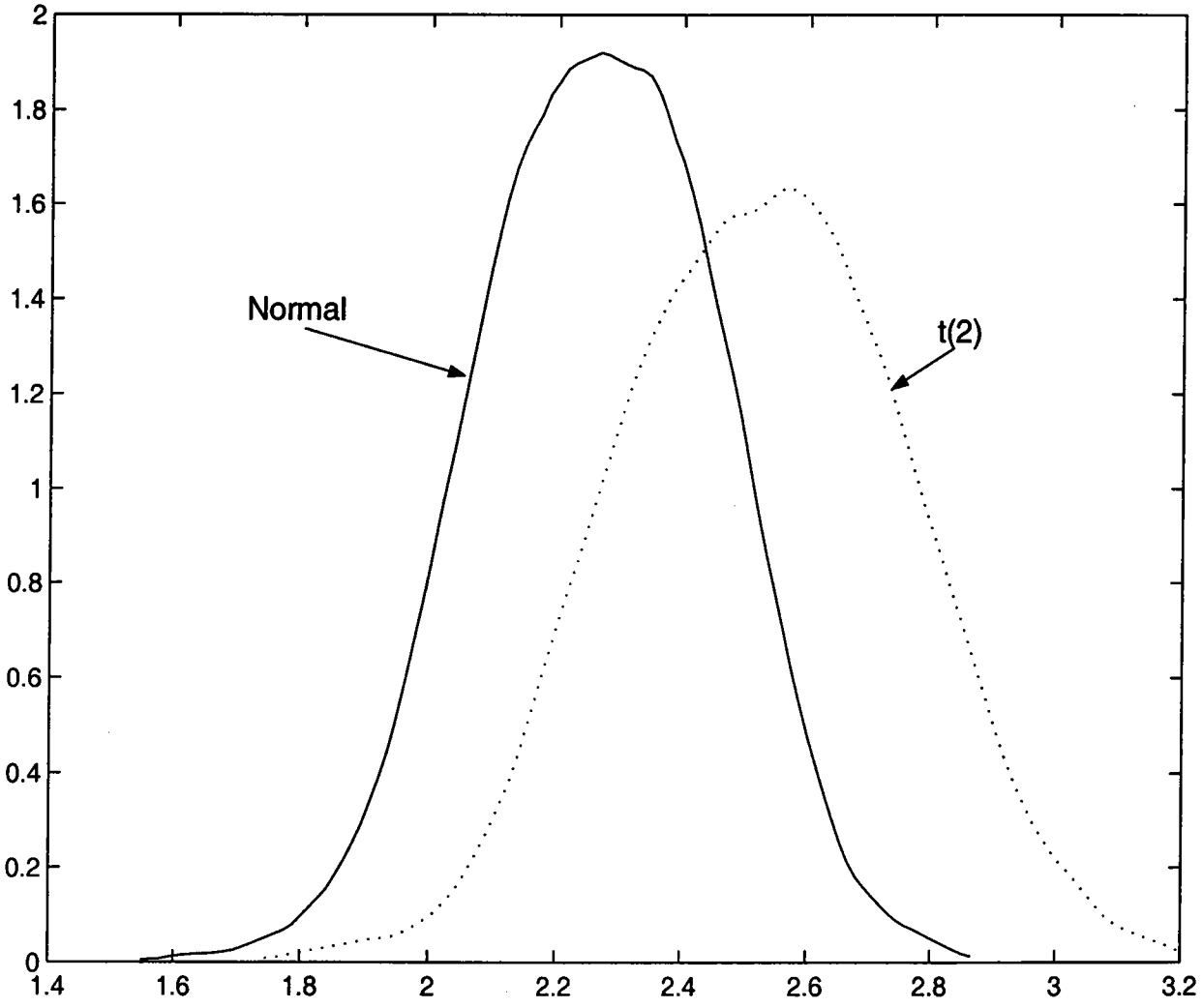
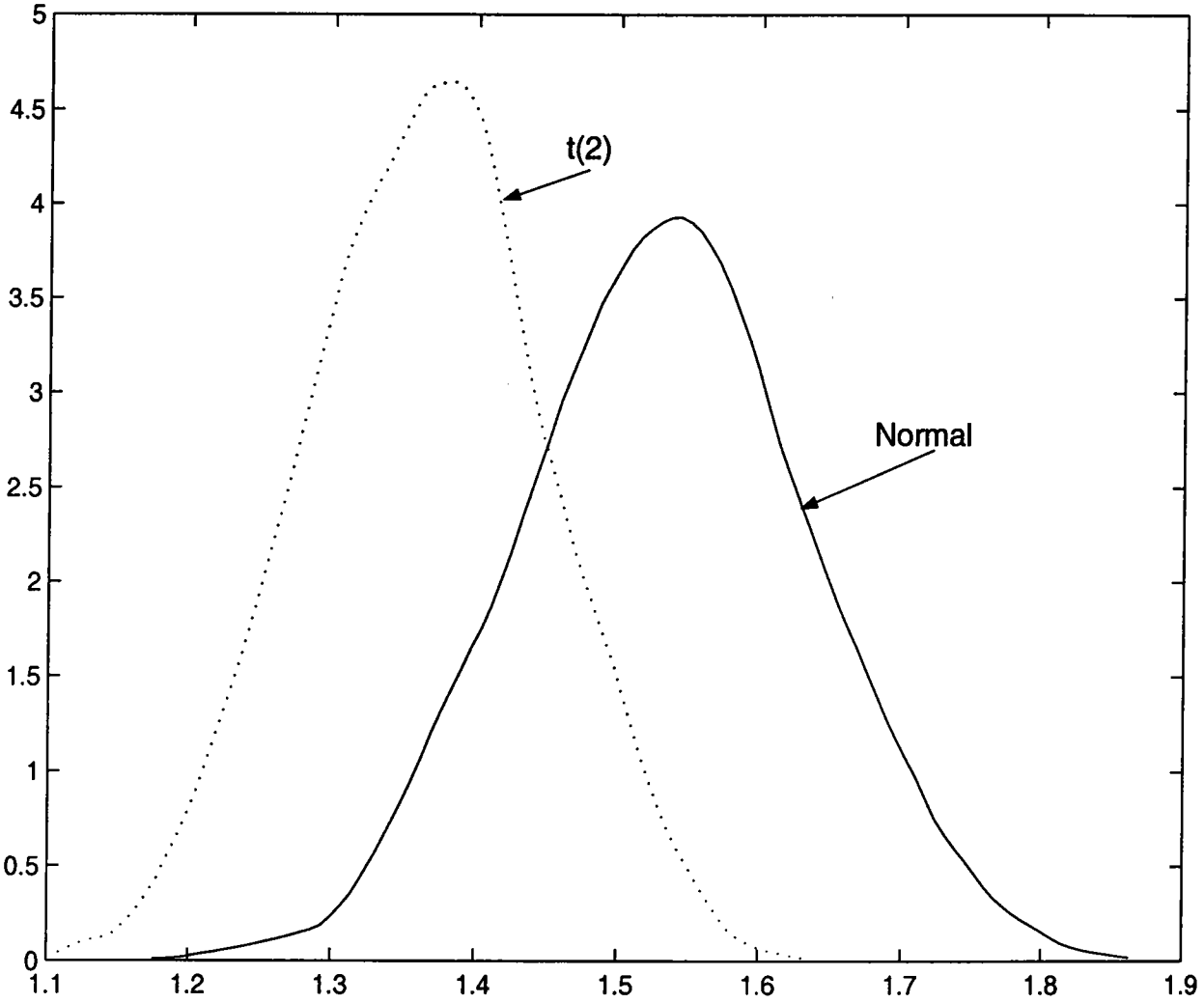


Figure 3: Marginal Treatment Effect with $u^D = 1$. True Value ≈ 1.54



Distributions of Treatment on the Treated and Marginal Treatment Effects Using Normal and t_2 Models. Generated t_4 Data. 1,000 Replications with N = 2,500.

Figure 4: Treatment on the Treated with $Z = -2$. True Value ≈ 2.64

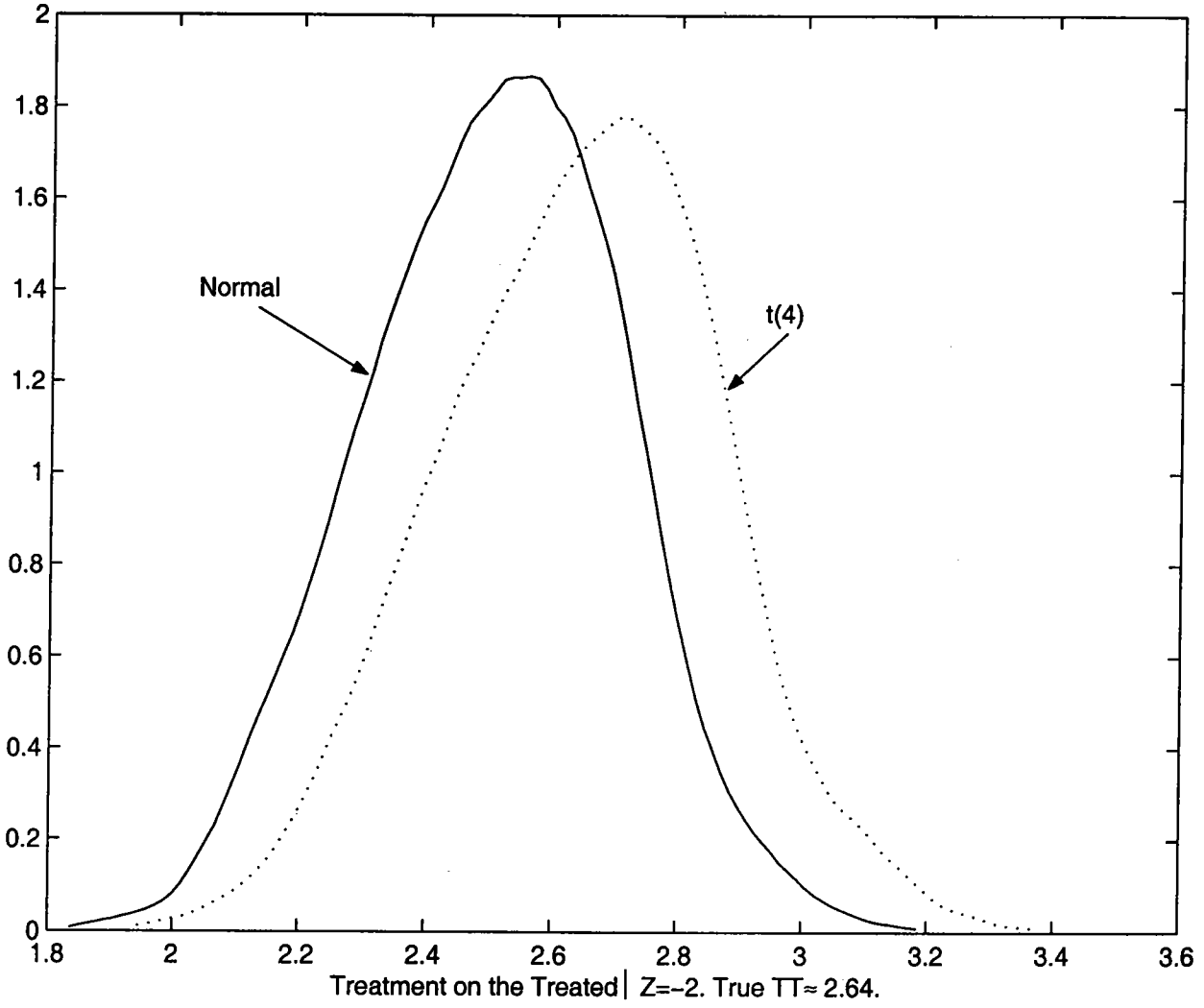


Figure 5: Marginal Treatment Effect with $u^D = 2$. True Value ≈ 2.08

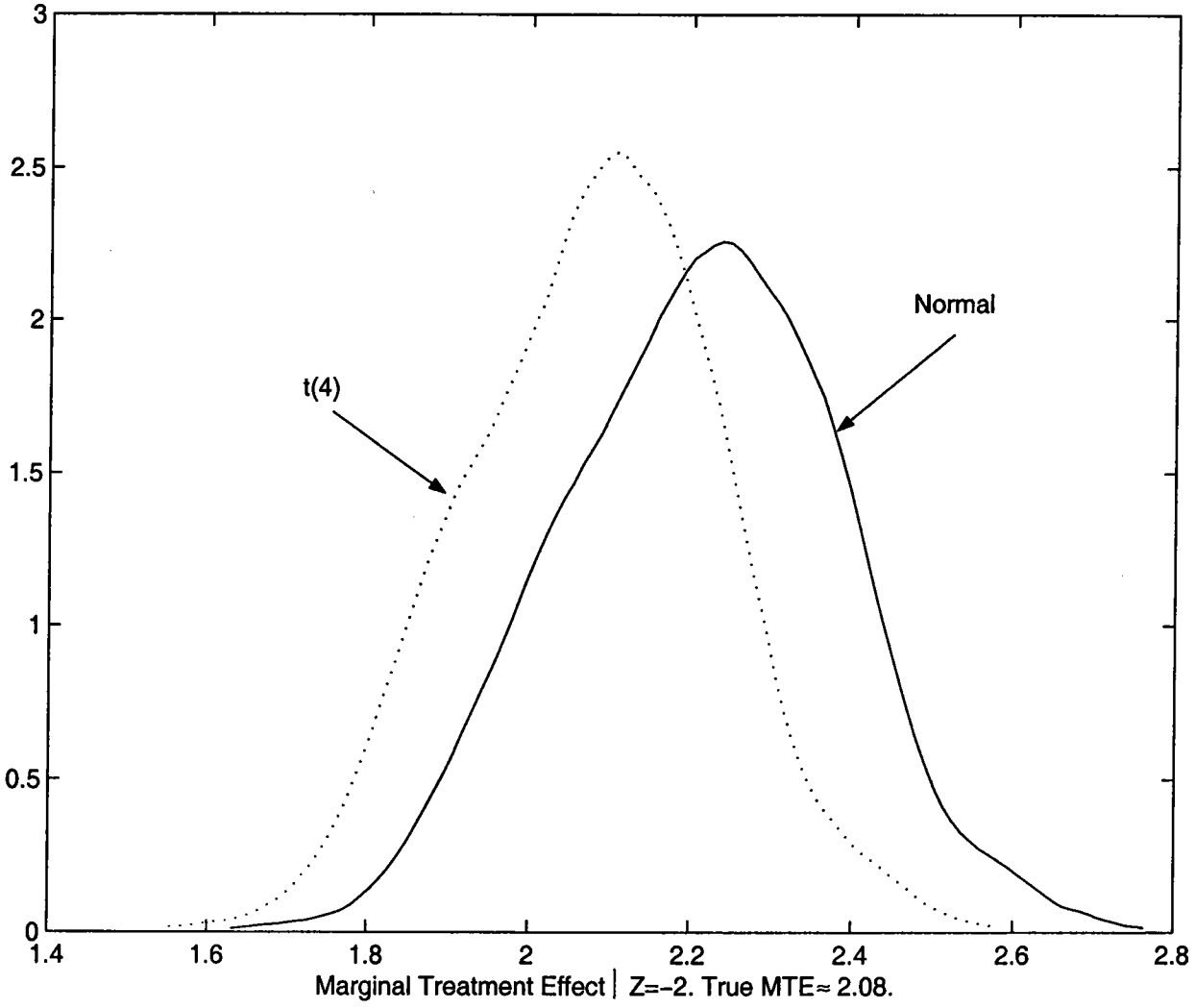


Figure 6: Probability of Correctly Choosing Normal Model Over t_2 Model Using MSE Criterion. 1,000 Iterations

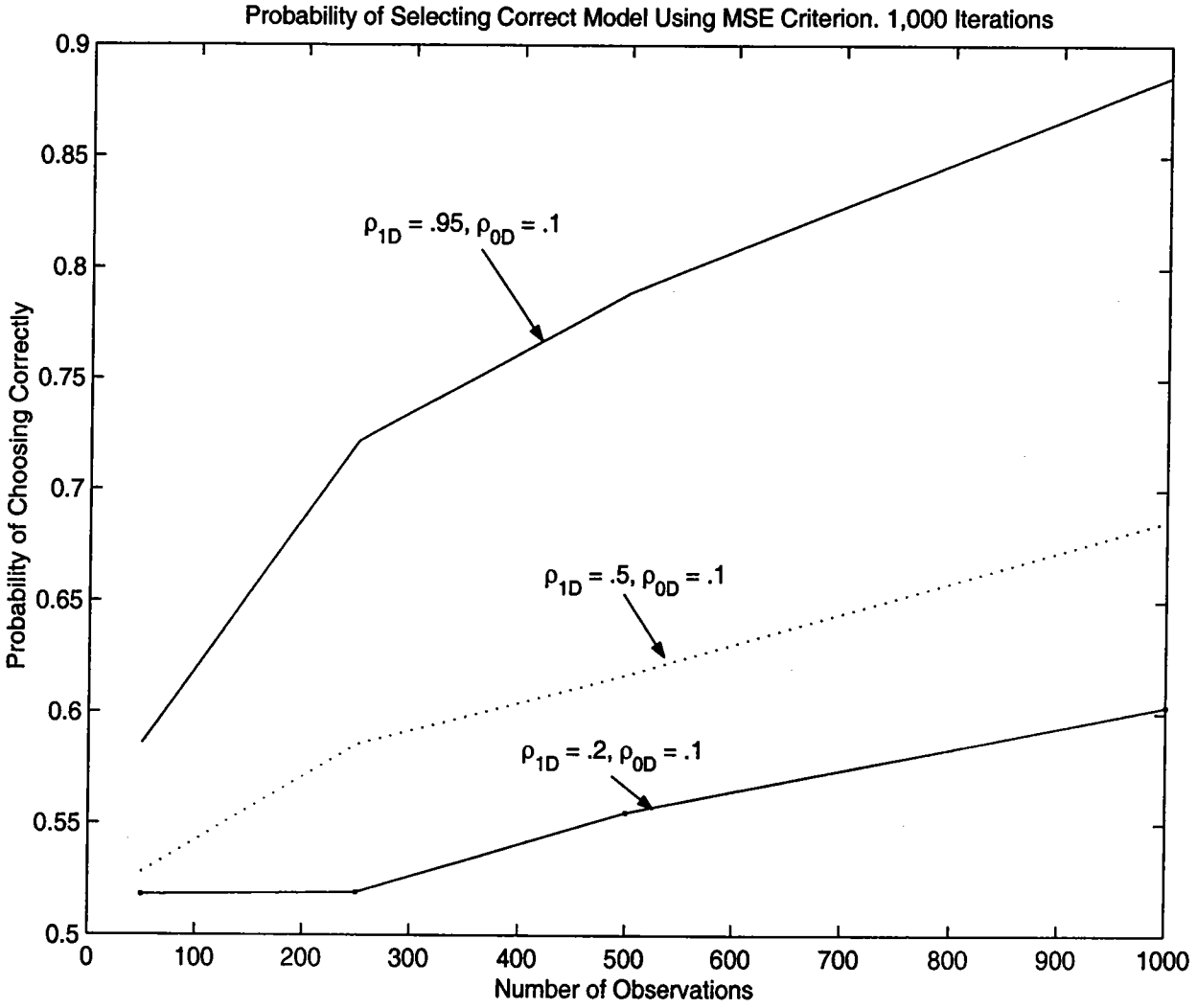


Figure 7: Plots of Marginal Treatment Effects Across Alternate Models (Unscaled)