Do Dropouts Suffer from Dropping Out? Estimation and Prediction of Outcome Gains in Generalized Selection Models

Mingliang Li
University of California - Irvine
minglial@uci.edu

Dale J. Poirier
University of California-Irvine
dpoirier@uci.edu

Justin L. Tobias[1]
Department of Economics
University of California-Irvine
3151 Social Science Plaza
Irvine, CA 92697-5100
jtobias@uci.edu

April 8, 2002

Abstract

In this paper we describe methods for predicting distributions of outcome gains in the framework of a latent variable selection model. We describe such procedures for the "textbook" Gaussian selection model, Student-t selection models, and a finite mixture of Gaussian selection models. Importantly, our algorithms for fitting these models are simple to implement in practice, and also permit learning to take place about the non-identified cross-regime correlation parameter. Using data from High School and Beyond, we apply our methods to determine the impact of dropping out of high school on a math test score taken at the senior year of high school. Our results show that selection bias is an important feature of this data, that our beliefs about this non-identified correlation are updated from the data, and that generalized models of selectivity offer an improvement over the "textbook" Gaussian model. Further, our results indicate that on average dropping out of high school has a large negative impact on senior-year test scores. However, for those individuals who actually drop out of high school, the act of dropping out of high school does not have a significantly negative impact on test scores. This suggests that policies aimed at keeping students in school may not be as beneficial as first thought, since those individuals who must be induced to stay in school are not the ones who benefit significantly (in terms of test scores) from staying in school.

JEL Classifications: C11, C15, C51

[1]Corresponding Author.

1 Introduction

Since the early 1970's, great strides have been made in the econometrics literature in the estimation of "treatment-response" or "selection" models when the assignment to treatment is not random. The statistical problem in such models is that the treatment decision itself is endogenous, and thus simple OLS estimates will confound the actual treatment impact with the effect of unobserved factors that influence both the treatment decision and the outcome(s) of interest.

In the early stages of this binary treatment - continuous outcome literature, the primary focus was on consistent estimation of structural parameters in the presence of selection bias (*e.g.* Gronau (1974) and Heckman (1976)). In subsequent work, criticisms emerged about the potential limitations of the normality assumption often used to facilitate estimation, (*e.g.* Goldberger (1983) and Paarsch (1984)), while the focus of several studies began to move away from estimation of structural parameters to the estimation of various *treatment parameters* such as the Average Treatment Effect (ATE), the effect of Treatment on the Treated (TT) and the Local Average Treatment Effect (LATE) (*e.g.* Rosenbaum and Rubin (1983), Heckman and Robb (1985), Bjorklund and Moffitt (1987), Imbens and Angrist (1994), Hahn (1998), Dehejia and Wahba (1999), Heckman and Vytlacil (1999,2000a,2000b), Heckman, Tobias and Vytlacil (2000), Hirano, Imbens and Ridder (2000), Vytlacil (2000), Abadie, Angrist and Imbens (2001)). These treatment parameters measure various expected outcome gains from receipt of treatment for different subpopulations, and offer a nice way to summarize average program impacts. Further developments were made in describing procedures for accommodating unobserved heterogeneity in these models (Heckman and Robb (1985) and Bjorklund and Moffitt (1987)), while others developed more flexible estimation methods and provided alternatives for relaxing the normality assumption (*e.g.* Lee (1982, 1983), Heckman (1990), Heckman, Ichimura, Smith and Todd (1998), and Heckman, Tobias and Vytlacil (2001)). Finally, recent research has made use of "natural experiments" to break the selection problem (*e.g.* Angrist (1990), Card (1990), Angrist and Krueger (1994), Angrist (1998), Meghir and Palme (1999)), though the lack of availability of quasi-experimental data and the existence of non-random assignment remains common to data available to social scientists.

Despite these numerous and important advances, relatively little attention has been given to the estimation of quantities other than mean treatment parameters for various subpopulations. In our view, the nearly exclusive focus on mean impacts is attributable to an unidentified parameter

problem. That is, for every individual in the sample, we will only observe his or her "treated" or "untreated" outcome, but never both. For this reason, the correlation between the treated and untreated outcomes is not identified, and does not enter the likelihood function for the observed data. As a result, *distributions of quantities of interest, such as the outcome gain resulting from receipt of treatment will depend on this non-identified parameter, while means of these distributions will not.* For this reason, mean impacts have dominated the literature, while relatively little attention has been given to characterizing *distributions of outcome gains.*

Notable exceptions exist, however, as Heckman and Honore (1990), Heckman and Smith with Clements (1997), Chib and Hamilton (2000) and Poirier and Tobias (2001) have discussed the issue of estimating distributions of outcome gains in the presence of this unidentified parameter. Heckman and Honore (1990) and Heckman and Smith with Clements (1997) discuss identification and estimation in the context of the Roy (1951) model, which is based upon income maximization and thus requires the decision to take treatment to be based upon the sign of difference of the treated and untreated outcomes.[2] Chib and Hamilton (2000) discuss within-sample distributions of outcome gains subject to the restriction that the non-identified correlation parameter is equal to zero. Poirier and Tobias (2001) relax this restriction, focus on predictive distributions of outcome gains, but only obtain results for the "textbook" Gaussian selection model.

In the following sections, we go beyond previous work in this area and offer several contributions to the existing literature. In so doing, we advocate an estimation approach that places a prior over the "full" covariance matrix - despite the fact that the cross-regime correlation parameter is not identified - and show that *learning can take place about the non-identified correlation parameter through information contained in the identified correlation parameters.* That is, the priors and posteriors for this non-identified correlation parameters can differ. In this sense, it is unreasonable and unnecessary to fix this parameter in value *a priori*, as the data update our beliefs about the values of this correlation. Additionally, we point out that when working with the "full" covariance matrix, the resulting Markov Chain Monte Carlo algorithms are relatively easy to implement in practice as the complete conditionals can be easily sampled.

Second, we extend the methods of Poirier and Tobias (2001) to include algorithms for fitting non-Gaussian selection models, particularly Student-t models as well as a finite mixture of Normals. In

[2]Interestingly, this requires the covariates in the selection equation to be exactly the same as those in the outcome equations, while the existence of an "instrument" is usually argued to be the source of identification in empirical work.

all cases, we work with the "full" covariance matrix, describe the resulting algorithms under this approach, and show that they are simple to implement in practice.

For these generalized models we also derive expressions for various *posterior predictive distributions of outcome gains resulting from the receipt of treatment*. We imagine that the goal of our empirical analysis is to use the available data to learn about the parameters of our model, and then use this information to predict the *distribution* of outcome gains for some future populations. We will use the learning that takes place about the non-identified correlation and incorporate that information into our predictive distributions of interest. We link all of these distributions to conventional *mean* treatment effects often used in the program evaluation literature. This includes the predictive distributions associated with the Average Treatment Effect (ATE), the effect of Treatment on the Treated (TT), and the Local Average Treatment Effect (LATE).

Finally, we apply our methods to predict the impact of dropping out of high school on a senior-year math test score using data from High School and Beyond (HSB). What makes this data set very interesting for our purposes is that the HSB administrators were careful to collect test score information for high school dropouts by re-interviewing the dropouts in groups outside the classroom. That is, we are able to obtain a "senior-year" math test score for those individuals who are currently completing their senior year of high school, as well as for those individuals who dropped out of school, but would be in their senior year had they remained in school.

For this application, one might suspect that unobservable factors making an individual more likely to drop out of high school might also make her more likely to earn low test scores had she chosen to remain in high school. Given this potential selectivity problem, we apply our algorithms for estimating Gaussian and non-Gaussian selection models to investigate these issues. We obtain results for the "textbook" Gaussian model, Student-t models, as well as finite Normal mixture models with varying numbers of mixture components. To determine which of those models receive the most support from the data, we calculate marginal likelihoods associated with each of these competing specifications.

As shown in section 7, this application is of significant economic interest, and also illustrates our econometric points amazingly well. We find that selection bias is an important feature of this data, and that the widely-used Gaussian selection model is inferior to the generalized selection models described in this paper. We find preference for a two component Normal mixture model, and that

within the component receiving the majority of the weight, *a substantial amount of learning takes place about the non-identified correlation parameter.* For the other component, which seems to primarily accommodate outlying individuals in the sample, very little learning takes place as the priors and posteriors for this non-identified correlation are virtually identical. Taken together, this application *simultaneously illustrates cases where the learning about the non-identified correlation parameter is both highly informative and uninformative.*

We then use this learning to calculate various predictive distributions of test score gains, and tie these distributions to various "treatment parameters" often reported in the program evaluation literature. We find that dropping out of high school has a large negative impact on senior-year test scores for an "average" individual. However, for those individuals who actually drop out of high school, the act of dropping out does not have a significantly negative impact on test scores. That is, the difference between the (counterfactual) test score dropouts would receive had they stayed in school and the (observed) test score they receive after dropping out of school seems nearly centered at zero. This suggests that policies aimed at keeping students in school may not be as beneficial as first thought, since those individuals who need to be induced to stay in school are not the ones who benefit significantly (in terms of test scores) from remaining in school.

The outline of the paper is as follows. In section 2, we introduce our standard model of potential outcomes. In section 3, we review and extend our theoretical analysis which shows how learning can take place about the non-identified cross-regime correlation parameter. In section 4 we briefly describe our algorithms for fitting Student-t and Normal mixture selection models, and note that since these algorithms work with the "full" covariance matrix, the data can serve to update our beliefs about this correlation parameter. Section 5 derives expressions for various predictive distributions of outcome gains for both the Student-t and mixture models, and discusses methods for calculating all of these predictive distributions. Section 6 describes the High School and Beyond data used in our application, and section 7 presents the empirical results. The paper concludes with a summary in section 8.

2 The Model

In this paper, we focus on a standard model of *potential outcomes* (*e.g.* Rubin (1978), Poirier and Ruud (1981), Rosenbaum and Rubin (1983), Dawid (2000)) with a binary treatment decision (D), and a continuous outcome (Y):

$$D^* = Z\theta + U_D \tag{1}$$

$$Y_1 = X\beta_1 + U_1 \tag{2}$$

$$Y_0 = X\beta_0 + U_0. \tag{3}$$

The last two equations are the outcome equations in the *treated* and *untreated* states, respectively, where the 1 subscript is used to denote variables and parameters associated with the treated state and the 0 subscript with the untreated state. We assume for simplicity that the variables appearing in X are constant across states. This assumption is essentially without loss of generality, and the subsequent analysis is not affected by replacing (2) and (3) with state-specific design matrices X_1 and X_0.

In our potential outcomes framework, D^* is a *latent variable* that generates an observed dichotomous treatment decision $D(Z)$:

$$D(Z) = I(D^* > 0) = I(Z\theta + U_D > 0).$$

Here $I(\cdot)$ is an indicator variable equal to one if the statement within the parentheses is true and is otherwise zero, $D(Z) = 1$ implies receipt of treatment, and $D(Z) = 0$ implies nonreceipt. The latent variable D^* has the interpretation as the net desire for receipt of treatment - individuals take the treatment if $D^* > 0$, but otherwise do not.

We also assume the existence of an exclusion restriction or instrument, and let Z_{k^*} denote an element of Z which is not contained in X. Though this assumption is not strictly required for identification (given some set of distributional assumptions), the practical importance of such an instrument has been widely-documented, and the recent movement toward "natural experiments" is entirely based on the availability of such an exogenous source of variation. Further, the instrument itself will serve to define the Local Average Treatment Effect parameter (*e.g.,* Imbens and Angrist (1994), Heckman and Vytlacil (1999, 2000a, 2000b)), as we will discuss in section 5.

For a given individual, we observe either their treated or untreated outcome, but never both.

Letting Y denote the observed outcome, we can write:

$$Y = DY_1 + (1 - D)Y_0.$$

To characterize the effectiveness of the program or treatment, we would like to learn about the outcome gain resulting from the receipt of treatment, (*i.e.* $\Delta \equiv Y_1 - Y_0$). Immediately, one recognizes that *distributions* associated with Δ depend on the non-identified correlation parameter $\rho_{10} \equiv \text{Corr}(Y_1, Y_0)$, though the *means* of these distributions will not. For this reason, previous work has focused on methods for estimating various average returns to the receipt of treatment (*i.e.* $E(\Delta|\cdot) = E(Y_1 - Y_0|\cdot)$) in order to evaluate the merits of the program. Here our goal is more ambitious as we seek to describe methods for calculating various *predictive distributions* of outcome gains for a variety of non-Gaussian models.

3 Learning About the Non-Identified Correlation Parameter

In this section, we review and extend the arguments of Vijverberg (1993), Koop and Poirier (1997), and Poirier and Tobias (2001) to show how learning takes place about the non-identified correlation parameter through information learned from the identified correlation parameters. When proceeding we will work with the correlation parameters, letting ρ_{10} denote the non-identified correlation, and ρ_{1D} and ρ_{0D} the identified correlations:

$$\rho_{10} \equiv \text{Corr}(U_1, U_0), \quad \rho_{1D} \equiv \text{Corr}(U_1, U_D), \quad \rho_{0D} \equiv \text{Corr}(U_0, U_D).$$

We let Σ denote the covariance matrix associated with the 3×1 disturbance vector from (1)-(3) and write:

$$\Sigma = \begin{bmatrix} 1 & \rho_{1D}\sigma_1 & \rho_{0D}\sigma_0 \\ \rho_{1D}\sigma_1 & \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 \\ \rho_{0D}\sigma_0 & \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 \end{bmatrix}.$$

We will also let ξ denote all remaining parameters of this model. We begin by noting that

$$|\Sigma| = \sigma_1^2\sigma_0^2 \left[(1 - \rho_{1D}^2)(1 - \rho_{0D}^2) - (\rho_{10} - \rho_{1D}\rho_{0D})^2\right]. \tag{4}$$

It follows that the covariance matrix Σ is positive definite *iff* this determinant is positive. This requires us to choose a prior for the non-identified correlation ρ_{10}, denoted $p(\rho_{10}|\rho_{1D}, \rho_{0D})$, which is *not* independent of the other correlation parameters, but instead is defined over the support $\underline{\rho}_{10} \leq \rho_{10} \leq \overline{\rho}_{10}$ where:

$$\underline{\rho}_{10} = \underline{\rho}_{10}(\rho_{1D}, \rho_{0D}) = \rho_{1D}\rho_{0D} - [(1 - \rho_{1D}^2)(1 - \rho_{0D}^2)]^{1/2} \tag{5}$$

$$\overline{\rho}_{10} = \overline{\rho}_{10}(\rho_{1D}, \rho_{0D}) = \rho_{1D}\rho_{0D} + [(1 - \rho_{1D}^2)(1 - \rho_{0D}^2)]^{1/2}. \tag{6}$$

As shown in Koop and Poirier (1997) and further described in Poirier and Tobias (2001), conditioned on the values of the identified correlations and remaining parameters, no learning takes place about the non-identified correlation parameter ρ_{10}. That is,

$$p(\rho_{10} \mid \rho_{1D}, \rho_{0D}, \xi, \text{Data}) \quad = \quad p(\rho_{10}|\rho_{1D}, \rho_{0D}, \xi) = p(\rho_{10}|\rho_{1D}, \rho_{0D}). \tag{7}$$

However, the *marginal* priors and posteriors of ρ_{10} can be quite different. To further describe this point and separate the contributions of the data and the prior in affecting the behavior of the marginal posterior for ρ_{10}, first let $R(\rho_{10})$ denote the conditional support of (ρ_{1D}, ρ_{0D}) given ρ_{10}, defined as the set of all (ρ_{1D}, ρ_{0D}) such that (4) is positive given ρ_{10}. It follows that:

$$\begin{aligned}
p(\rho_{10}|\text{Data}) &= \int_{R(\rho_{10})} \int_{\xi} p(\rho_{10}, \rho_{1D}, \rho_{0D}, \xi|\text{Data}) d\rho_{1D}\, d\rho_{0D}\, d\xi \\
&= \int_{R(\rho_{10})} \int_{\xi} p(\rho_{10}|\rho_{1D}, \rho_{0D}) p(\rho_{1D}, \rho_{0D}, \xi|\text{Data}) d\rho_{1D}\, d\rho_{0D}\, d\xi \\
&= \int_{R(\rho_{10})} \int_{\xi} p(\rho_{10}|\rho_{1D}, \rho_{0D}) p(\xi|\rho_{1D}, \rho_{0D}, \text{Data}) p(\rho_{1D}, \rho_{0D}|\text{Data}) d\rho_{1D}\, d\rho_{0D}\, d\xi \\
&= \int_{R(\rho_{10})} p(\rho_{10}|\rho_{1D}, \rho_{0D}) p(\rho_{1D}, \rho_{0D}|\text{Data}) d\rho_{1D}\, d\rho_{0D}.
\end{aligned}$$

This second line factors the joint posterior into the conditional for ρ_{10} times the marginal, and uses the result above - that the conditional priors and posteriors of the non-identified correlation are identical. The last line of this derivation suggests that as the identified correlation parameters asymptotically "collapse" around some limiting values, the marginal posterior for ρ_{10} would reduce to the conditional prior for ρ_{10} evaluated at those values of the identified correlations.

That is, if the joint posterior for ρ_{1D} and ρ_{0D} is highly informative or "tight" about some point $(\hat{\rho}_{1D}, \hat{\rho}_{0D})$ then

$$p(\rho_{10}|\text{Data}) \approx p(\rho_{10}|\rho_{1D} = \hat{\rho}_{1D}, \rho_{0D} = \hat{\rho}_{0D}). \tag{8}$$

This derivation helps to separate out the influence of the data and the prior in determining the behavior of ρ_{10}. That is, information conveyed from the data regarding the identified correlation parameters ρ_{1D} and ρ_{0D} spills over and revises our beliefs about the conditional support of ρ_{10}. Thus, *in general, the marginal priors and posteriors will differ, suggesting that learning has taken place.* However, within these conditional support bounds, the prior clearly matters. In fact, the above derivation suggests that if the joint posterior for ρ_{1D} and ρ_{0D} was degenerate (which might approximately be the case in sufficiently large samples), then the marginal posterior of ρ_{10} is simply the conditional prior evaluated and those limiting values of ρ_{1D} and ρ_{0D}.

4 Estimation in Non-Gaussian Selection Models

In the preceding section, we showed how learning can take place about the non-identified cross-regime correlation parameter ρ_{10} through the positive definiteness of Σ. In the appendix, we outline in complete detail our algorithms for estimating selection models as in (1) - (3) when the errors are assumed to follow a multivariate Student-t distribution, or when the distribution of outcomes is assumed to follow a finite mixture of Normals. The computational "tricks" we employ for estimating these generalized selection models are standard ones - adding Gamma mixing variables to the disturbance term variance to extend to the class of Student-t distributions (*e.g.* Carlin and Polson (1991), Albert and Chib (1993), Geweke (1993), Chib and Hamilton (2000)), and augmenting with component indicator variables to estimate mixture models (*e.g.* McLachlan and Peel (2000)).

It is important to recognize that the algorithms we describe work with the "full" 3×3 covariance matrix, and as such, permit learning to take place about ρ_{10}. Though Chib and Hamilton (2000) also discussed procedures for extending the textbook Gaussian selection model, their algorithms fix ρ_{10} *a priori*, and thus can not permit any learning to take place about this parameter. Our algorithms embrace the potential for learning about ρ_{10}, and enable researchers to specify a range of prior beliefs about this parameter. Further, the resulting algorithms when working with the "full" covariance matrix are comparatively simple to implement (as described in the appendix), since all of the complete conditionals are easily sampled and no Metropolis-Hastings substeps are required.

5 Predictive Distributions of Outcome Gains

In this section we show how our estimation results can be used to characterize the *predictive distributions* of outcome gains associated with the receipt of treatment. To this end, we imagine the outcomes of a future population and suppose our selection model as described in (1)-(3) applies to analysis of that population. From the past data we will learn about the parameters of our model, and we will use this updated parameter information to characterize predictive distributions of outcome gains resulting from receipt of treatment. These predictive distributions can be used in order to evaluate the benefits associated with making the treatment available to some future population. Our analysis will not only provide us with mean effects, but will enable us to compute

a wealth of other quantities of interest such as variances and quantiles associated with the outcome gain from receipt of treatment. Further, we will characterize the impact of the treatment on a variety of different subpopulations, and relate these distributions to widely-used "treatment parameters."

To these ends, we define the following variables:

$$\Delta_f \equiv Y_{1f} - Y_{0f}, \quad \gamma_1 \equiv \sigma_{1D} - \sigma_{0D}, \quad \gamma_2 \equiv \sigma_1^2 + \sigma_0^2 - 2\sigma_{10},$$

where the f subscript is used to denote future, as yet unobserved outcomes. We also let x_f and z_f denote the future covariates in the outcome and selection equations, respectively.

We focus on describing methods for obtaining three different predictive distributions of outcome gains and tie these into the previous program evaluation literature. Specifically, we wish to characterize

$$p(\Delta_f | x_f, \text{Data}) \tag{9}$$

$$p(\Delta_f | x_f, z_f, D_f(z_f) = 1, \text{Data}) \tag{10}$$

$$p(\Delta_f | x_f, z_f, \tilde{z}_f, D_f(\tilde{z}_f) = 1, D_f(z_f) = 0, \text{Data}) \tag{11}$$

The first density describes the predictive distribution of outcome gains for a random person, the second describes the distribution for those taking the treatment at z_f, and the third describes the distribution for those taking the treatment at $\tilde{z}_f$ but not at z_f.[3] The means of these distributions correspond to the Average Treatment Effect (ATE), the Effect of Treatment on the Treated (TT) (*e.g.* Rubin (1978), Heckman and Robb (1985)), and Local Average Treatment Effect (LATE) (*e.g.* Imbens and Angrist (1994)) predictive parameters, respectively, widely reported in the program evaluation literature.

To further describe these predictives via a concrete example, let us jump ahead to our application of section 7. Suppose that the decision to drop out of high school is regarded as a decision to receive "treatment." Further, suppose the outcome variable of interest is a test score variable that measures student achievement. The densities in (9)-(11) would then characterize the test score loss from dropping out of high school (relative to remaining in school) for someone chosen at random (ATE), for those actually choosing to drop out (TT), or for those that can be induced to drop out as a result of a change in some instrument. In our application, we will use local labor market

[3] We assume that $\tilde{z}_f\theta > z_f\theta$, or that the change in instrument leads to a higher propensity for the individual to take the treatment.

conditions as an instrument that affects the decision to drop out of high school (better conditions would imply that it is easier to obtain attractive employment opportunities), but does not directly affect test scores. Thus, LATE would characterize the test score loss received by individuals "riding the fence" who would not drop out at one value of labor market conditions, but would drop out at some other "improved" value of labor market conditions.

Of course, the densities in (9)-(11) are not conditioned on any parameters. As such, we will first need to obtain expressions for these densities conditioned on all parameters, and then integrate the results over joint posterior distribution. Formally, then, we will obtain these predictives as follows:

$$p(\Delta_f|x_f, \text{Data}) = \int_\eta p(\Delta_f|x_f, \eta, \text{Data}) p(\eta|\text{Data}) d\eta$$

$$p(\Delta_f|x_f, z_f, D_f(z_f) = 1, \text{Data}) = \int_\eta p(\Delta_f|x_f, z_f, D_f(z_f) = 1, \eta, \text{Data}) p(\eta|z_f, D_f(z_f) = 1, \text{Data}) d\eta$$

$$\begin{aligned} p(\Delta_f|x_f, z_f, \tilde{z}_f, D_f(\tilde{z}_f) = 1, D_f(z_f) = 0, \text{Data}) &= \int_\eta [p(\Delta_f|x_f, z_f, \tilde{z}_f, D_f(\tilde{z}_f) = 1, D_f(z_f) = 0, \text{Data}, \eta) * \\ &\quad p(\eta|z_f, \tilde{z_f}, D_f(\tilde{z}_f) = 1, D_f(z_f) = 0, \text{Data})] \, d\eta, \end{aligned}$$

where η denotes all of the parameters in the model. In the above, we are careful to recognize that the events $\tilde{D}_f = 1$, $D_f = 1$ or $D_f = 0$ involve elements of the parameter vector η, and so the distribution we average over in the expressions above must also condition on these restrictions. It is straight-forward to show that

$$p(\eta|z_f, D_f(z_f) = 1, \text{Data}) \propto \Pr(D_f(z_f) = 1|z_f, \eta) p(\eta|\text{Data}) \quad (12)$$

$$p(\eta|\tilde{z}_f, z_f, \tilde{D}_f(\tilde{z}_f) = 1, D_f(z_f) = 0, \text{Data}) \propto \Pr(\tilde{D}_f(\tilde{z}_f) = 1, D_f(z_f) = 0|\tilde{z}_f, z_f, \eta) p(\eta|\text{Data}) \quad (13)$$

These results can be substituted into the expressions above so that in all cases we perform the integration over the joint posterior $p(\eta|\text{Data})$, suggesting that simulation-based methods will enable us to perform this high-dimensional integration. Expressions for the terms $\Pr(D_f(z_f) = 1|z_f, \eta)$ and $\Pr(\tilde{D}_f(\tilde{z}_f) = 1, D_f(z_f) = 0|\tilde{z}_f, z_f, \eta)$ are easily obtainable from each model. For example, if we were assuming a Gaussian model, these would be $\Phi(z_f\theta)$ and $\Phi(\tilde{z}_f\theta) - \Phi(z_f\theta)$, respectively.

What remains is for us to derive expressions for the predictive distributions of outcome gains *conditioned* on the values of these parameters. In the following two sections we derive expressions for these conditionals in our Student-t and Normal mixture models.

5.1 Student-t predictives

With a bit of work, we derive expressions for these conditional predictives within our Student-t model:

$$[ATE] \equiv p_t(\Delta_f|x_f,\eta) \sim t_v(x_f(\beta_1-\beta_0),\gamma_2),$$

where $t_v(a,b)$ denotes a Student-t density with v degrees of freedom, mean a and variance $vb/(v-2)$. To obtain the densities for TT and LATE, we first need to define the following variables:

$$\mu_{D^*|\Delta_f}(x_f,z_f,\Delta_f,\eta) \equiv z_f\theta + (\gamma_1/\gamma_2)[\Delta_f - x_f(\beta_1-\beta_0)].$$

$$\Omega_{D^*|\Delta_f}(x_f,\Delta_f,v,\eta) \equiv \left[v + \frac{(\Delta_f - x_f(\beta_1-\beta_0))^2}{\gamma_2}\right]\left(\frac{1}{v+1}\right)\left(1-\frac{\gamma_1^2}{\gamma_2}\right).$$

Given this notation, we obtain the following conditional predictives:

$$[TT] = p(\Delta_f|x_f,z_f,D_f(z_f)=1,\eta) = \left(\frac{p_t(\Delta_f|x_f,\eta)}{T_v(z_f\theta)}\right) T_{v+1}\left(\frac{\mu_{D^*|\Delta_f}(x_f,z_f,\Delta_f,\eta)}{\sqrt{\Omega_{D^*|\Delta_f}(x_f,\Delta_f,v,\eta)}}\right).$$

$$\begin{aligned}[LATE] &= p(\Delta_f|x_f,z_f,\tilde{z}_f,D_f(z_f)=0,D_f(\tilde{z}_f)=1,\eta) \\ &= \left(\frac{p_t(\Delta_f|x_f,\eta)}{T_v(\tilde{z}_f\theta)-T_v(z_f\theta)}\right)\left[T_{v+1}\left(\frac{\mu_{D^*|\Delta_f}(x_f,\tilde{z}_f,\Delta_f,\eta)}{\sqrt{\Omega_{D^*|\Delta_f}(x_f,\Delta_f,v,\eta)}}\right) - T_{v+1}\left(\frac{\mu_{D^*|\Delta_f}(x_f,z_f,\Delta_f,\eta)}{\sqrt{\Omega_{D^*|\Delta_f}(x_f,\Delta_f,v,\eta)}}\right)\right].\end{aligned}$$

In the expressions for TT and LATE, $p_t(\Delta_f|x_f,\eta)$ refers to the ATE density for the Student-t model as derived above.

Since the *conditional* predictives have the above closed-form solutions, we can obtain the *unconditional* predictives via "Rao-Blackwellization." That is, taking the ATE density as an example, we can use

$$\hat{p}(\Delta_f^0|x_f,\text{Data}) = \frac{1}{m}\sum_{i=1}^{m} p(\Delta_f^0|x_f,\eta=\eta^i,\text{Data}),$$

where η^i is the i^{th} post-convergence draw from the sampler, and m denotes the total number of draws. This is repeated for a variety of different Δ_f^0, thus providing density ordinates over a fine grid of values. A similar process is used to obtain the unconditional TT and LATE predictives.

5.2 Predictives for the Mixture Model

To derive expressions for the predictive distributions of outcome gains using Normal mixtures, we first note that for estimation purposes we have introduced a set of component indicator variables,

say $\{c_{ig}\},\ i = 1, 2, \cdots n,\ g = 1, 2, \cdots G$ into our model. The variable $c_{ig} = 1$ denotes that the i^{th} individual is drawn from the g^{th} component of the mixture, and is otherwise zero.

In terms of prediction, *conditioned on the future component indicator values* $c_{gf} = 1$ we are in the framework of the "textbook" Gaussian selection model, and thus the expressions for the conditional predictives follow identically to the Gaussian case. These expressions for the textbook Gaussian model were previously derived in Poirier and Tobias (2001). Thus, we obtain:

$$[ATE]:\ p(\Delta_f|c_{gf}=1, x_f, \eta, \text{Data}) = \phi(\Delta_f; x_f(\beta_1^g - \beta_0^g), \gamma_2^g) = [\gamma_2^g]^{-1/2}\phi\left(\frac{\Delta_f - x_f(\beta_1^g - \beta_0^g)}{\sqrt{\gamma_2^g}}\right). \quad (14)$$

$$[TT]:\ p(\Delta_f|c_{gf}=1, x_f, z_f, D(z_f)=1, \eta, \text{Data}) = \frac{\phi(\Delta_f; x_f(\beta_1^g - \beta_0^g), \gamma_2^g)}{\Phi(z_f\theta^g)}\Phi\left[\frac{z_f\theta^g + (\gamma_1^g/\gamma_2^g)[\Delta_f - x_f(\beta_1^g - \beta_0^g)]}{\sqrt{1-([\gamma_1^g]^2/\gamma_2^g)}}\right]. \quad (15)$$

$$[LATE]: p(\Delta_f|c_{gf}=1, x_f, z_f, \tilde{z}_f, D(z_f)=0, D(\tilde{z}_f)=1, \eta, \text{Data}) = \frac{\phi(\Delta_f; x_f(\beta_1^g - \beta_0^g), \gamma_2^g)}{\Phi(\tilde{z}_f\theta^g) - \Phi(z_f\theta^g)} \cdot A(\Delta_f, x_f, z_f, \tilde{z}_f, \eta) \quad (16)$$

where

$$A(\Delta_f, x_f, z_f, \tilde{z}_f, \eta) \equiv \left[\Phi\left(\frac{\tilde{z}_f\theta^g + (\gamma_1^g/\gamma_2^g)[\Delta_f - x_f(\beta_1^g - \beta_0^g)]}{\sqrt{1-([\gamma_1^g]^2/\gamma_2^g)}}\right) - \Phi\left(\frac{z_f\theta + (\gamma_1^g/\gamma_2^g)[\Delta_f - x_f(\beta_1^g - \beta_0^g)]}{\sqrt{1-([\gamma_1^g]^2/\gamma_2^g)}}\right)\right],$$

and we have used the notation $\phi(x; \mu, \sigma^2)$ to denote that x has a normal distribution with mean μ and variance σ^2. Each component of the mixture is permitted to contain its own regression parameters and covariance matrix, so the "g" superscript is used to denote parameters associated with the g^{th} component of the mixture.

The desired predictives *given the parameters but marginalized over the component indicators* follows as a weighted average of the conditional predictives above, where the component probabilities serve as the weights. We then integrate this result over the posterior distribution of the parameters to obtain the predictive density of interest. Focusing on ATE as an example we would proceed to obtain:

$$\begin{aligned} p(\Delta_f|x_f, \text{Data}) &= \int_\eta p(\Delta_f|x_f, \eta, \text{Data})p(\eta|\text{Data})d\eta & (17)\\ &= \int_\eta \left[\sum_{g=1}^{G} p(\Delta_f|x_f, \eta, c_{gf}=1, \text{Data})\Pr(c_{gf}=1|\eta, \text{Data})\right] p(\eta|\text{Data})d\eta & (18)\\ &= \int_\eta \sum_{g=1}^{G} \left[\pi_g p(\Delta_f|x_f, \eta, c_{gf}=1, \text{Data})\right] p(\eta|\text{Data})d\eta. & (19) \end{aligned}$$

Note that the component probability π_g is an element of the complete parameter vector η, and thus we use our updated posterior beliefs regarding the weights associated with each component when calculating these predictives. Rao-Blackwellization can again be used to obtain ordinates of this predictive, since the *conditional* (on the parameters and component indicators) predictives are known, as given in (39)-(41). Calculation of the TT and LATE predictives in the mixture model follows similarly.

6 The Data

We apply our procedures described in the previous sections to assess the impact of dropping out of high school on a mathematics exam administered in the senior year of high school. We acquire data to investigate these issues from the High School and Beyond (HSB) data set.

HSB is a survey conducted on behalf of the National Center for Education Statistics, and was constructed with the intent of yielding a sample of students that are representative of the population of American high school students. HSB is a biennial survey that begins in 1980, and in this base-year, two large cohorts of sophomore and senior high school students are interviewed.[4] To focus on the impact of dropping out of high school on student achievement, we confine our attention to the sophomore cohort, as some (approximately 8 percent) of this original cohort will be observed to dropout of high school prior to their 1982 (senior-year) interview. It is also very important to recognize a somewhat unusual feature of this data - that "senior year" test scores are available for *both those that drop out of high school as well as those that do not.* Given the geographical "clustering" of the sample, HSB was able to conduct group interviews for the dropouts outside of school at approximately the same time as the in-school students were interviewed, and test score data was collected for each group. This enables us to determine if the act of dropping out of high school between the sophomore and senior years has important consequences on senior year student achievement.

In both the base year and first-follow up survey, the sophomore cohort is given a variety of tests in several different areas. In this paper, we focus only on two sections of those tests which involve mathematical and quantitative reasoning. These tests specifically involve quantitative comparisons

[4]The sophomore cohort, for example, consists of approximately 30,000 individuals.

in which the student indicates which of two quantities is greater, asserts their equality, or indicates lack of sufficient data to determine which quantity is greater. We calculate both a sophomore and senior year test score as an average of the two mathematics test scores taken in each year. Each of the test scores is then standardized to have mean zero and unit variance. The senior-year mathematics test score is used as the outcome variable in both the "treated" (dropout) and "untreated" (non-dropout) states. We will include the base-year (sophomore) math test score from the 1980 interviews as an explanatory variable in the outcome and selection equations to pick up initial differences in "ability" across individuals. In our outcome (test score) equations, we also add dummy variables for being female or white, highest grade completed by the individual's mother and father, family income, and number of siblings as explanatory variables.

We also add this same set of variables to explain the decision to drop out of high school. This is motivated by our belief that those individuals with high sophomore-year test scores coming from well-educated and wealthy families are probably less likely to drop out of high school between their sophomore and senior years. We additionally include the race and gender dummies as well as a control for number of siblings in this dropout equation, as these variables might also be empirically important explanatory variables describing dropout behavior.

Our exclusion restriction which enters this dropout equation but does not appear in the outcome (senior test score) equations is the percentage of employment growth in the local labor market over the period 1980-1982. Our expectation is that a large amount of local employment growth over this period suggests prosperous local labor market conditions, making it more attractive for someone to drop out of high school and begin full-time employment. Specifically we imagine that individuals who are just indifferent to dropping out or staying in school might be induced to drop out if the local labor market conditions were to improve. We do not expect, however, that employment growth itself will have a direct effect on senior-level test scores, and thus omit it from the senior-year test score equations.[5]

We restrict the sample to students in the sophomore cohort attending public high schools who participated in both the base-year and follow-up mathematics tests. Further excluding observations where other key covariates are missing produced a final sample of 12,459 observations. Among this final sample, approximately 8.1 percent of the individuals (1,006) dropped out of high school

[5]To test this supposition, we added this employment growth rate to the test score equations, and found that it played virtually no role in those equations, and its inclusion had no effect on the estimates obtained for the remaining parameters.

between their sophomore and senior years, so that a substantial amount of observations exist for the estimation of parameters in both the "treated" and "untreated" states.

7 Empirical Results

Our goal is to take the HSB data and use it to address two questions which seem to be of primary interest: (1) How does dropping out of high school impact the test scores of a randomly chosen individual? (2) How are the test scores of those that actually choose to drop out of high school affected by dropping out? To address these questions, we adopt our potential outcome framework outlined in equations (1) - (3). Thus, we imagine that for each of these questions the primary object of interest is the difference between the test scores that one would receive had they completed high school and the test score they would get if they dropped out. For every individual, only one of those two outcomes is observed, though we can imagine a counterfactual state where the missing outcome would be observed. To address question (1) we will calculate the posterior predictive distribution of outcome gains for a randomly chosen person (ATE), and for question (2) we will calculate this distribution for those who actually drop out (TT).

To illustrate how our algorithms presented in section 3 can be applied, we first fit our selection model under a variety of different distributional assumptions. In all cases, we choose our prior hyperparameters to center both the identified and non-identified correlations over zero, yet remain reasonably diffuse. Thus, our prior "centers" our model over one where selection bias is not present, yet is diffuse enough so that the data can revise our beliefs and reveal to us that selectivity is present. Priors for the regression parameters are also chosen to be quite diffuse (yet proper) so that the data information is predominant.[6]

We obtain results for the "textbook" Gaussian model, a Student-t model with 2, 5, and 16 degrees of freedom, and also two and three component Normal mixture models. Our prior view is that the three component mixture model should be general enough to capture the key features of this data, and as shown below, the data do tend to favor specifications that are more parsimonious than this most general specification. For each of these models we calculate log marginal likelihoods to determine those specifications most favored by the data. The results of these marginal likelihood

[6]Details are available upon request.

calculations are presented in Table 1 below:

Table 1:
Log Marginal Likelihoods, and Posterior Model Probabilities
For Alternate Models

All Models	Log Marginal Likelihoods	Model Prob. (L-M)
Gaussian	-14,106.538	0.000000
t(v=2)	-14,511.080	0.000000
t(v=5)	-14,035.599	0.000000
t(v=16)	-14,025.568	0.000000
Two-Component	-13,996.708	0.999960
Three-Component	-14,006.835	0.000040

As shown in the table, the widely-used Gaussian model ranks second-to-last relative to its competitors, and the two component mixture model produces the highest log marginal likelihood. These results suggest that for this particular application the use of generalized models of selectivity are clearly supported by the data. Further, values of the marginal likelihoods imply that the associated posterior model probabilities virtually place probability one on the two-component Normal mixture, and thus model averaged quantities would simply reduce to model-specific ones. For this reason, we focus our remaining attention on specific results obtained from the two-component mixture model, though details of results from these alternate models are available upon request.

Estimation results from the two component mixture are presented in Table 2. To interpret these and subsequent results, we regard the decision to drop out as the decision to receive "treatment" so that, in the notation in equations (1) - (3), Y_1 represents the test score for the dropouts, and Y_0 represents the test score for those remaining in high school. The first set of rows of Table 2 provide posterior means, standard deviations and probabilities of being positive for parameters in the test score equation for the dropouts (the "treated" outcome). The second set provides posterior quantities for the test scores of the non-dropouts (the "untreated" outcome), and the third set provides posterior quantities for the equation governing the decision to drop out (the "selection" equation). The final portion of the table provides posterior means of the elements of the 3×3 covariance matrix Σ.

As shown in the first row of Table 2, the second component receives the vast majority of the weight, as the posterior mean of the probability associated with this second component was .91. The remaining 9 percent is allocated to the first component of the mixture, and as seen from

inspection of the elements of Σ, the group of people who "comprise" this first component appear to be characterized by relatively high-variance outcomes.

As the second component receives the vast majority of the weight, we confine most of our discussion to estimation results obtained within that component. As a general rule, the direction of the effects suggested by table 2 are highly consistent with our prior expectations. Individuals scoring higher on the sophomore exam, raised in families with higher family income and fewer siblings are more likely to score higher on the senior mathematics test. *Note that the empirical importance of the family characteristics remains even after controlling for sophomore-year test scores, suggesting that family environment between the sophomore and senior years matters in terms of senior-year student achievement.* Also note that for the non-drop out equation, which contains the majority of our observations, the probabilities of being positive are virtually one or zero for all of the coefficients. In this sense, the posterior suggests overwhelming evidence that family education, income and size, and initial test scores are important predictors of senior year student achievement.

For the dropout equation, the coefficient estimates are again very similar to what we would expect. Those achieving higher sophomore test scores with more educated and wealthier parents from smaller families appear to be significantly less likely to drop out of high school between their sophomore and senior years. On another important issue, our exclusion restriction appears to be an empirically important factor in the decision to drop out. Higher employment growth over the period from 1980 to 1982 is associated with an increased propensity for students to drop out of high school from their sophomore to senior years. This is consistent with our prior view that favorable local labor market conditions may induce some individuals to drop out of high school who otherwise would not, so that local employment growth rates should be positively correlated with the decision to drop out of high school.

Inspection of the elements of the covariance matrix reveals some very important results. For the second component of the mixture, posterior means of the identified correlations ρ_{1D} and ρ_{0D} are negative, and the marginal posterior distributions show that virtually all of their mass is placed over negative values. The negative coefficient estimates indicate that *unobservable factors making it more likely for an individual to drop out of high school also make it less likely for him or her to receive high senior year test scores.* Thus, in order to accurately characterize the impact of dropping out of high school on senior-year test scores, one needs to estimate a model like this one which accounts for the endogeneity of dropout choice and features the role of the unobservables.

For the relatively few individuals belonging to the first component, however, selection does not seem to be empirically important, as the identified correlations are not clearly bounded away from zero. For the vast majority of individuals, however, *selection bias is a key feature of this data.*

The fact that the selection effect differs across the components of the mixture illustrates our theoretical points made in section 4 extremely well. For the second component, the identified correlations are bounded away from zero, and ρ_{0D} is quite large in value. This provides a vehicle for learning about ρ_{10}, as the lower and upper conditional support bounds will be informative. Recall that the positive definiteness of the 3×3 covariance matrix Σ implies $\underline{\rho}_{10} \leq \rho_{10} \leq \overline{\rho}_{10}$, where these upper and lower limits depend only on the identified correlations ρ_{1D} and ρ_{0D}, as described in (5) and (6).

Evaluated at posterior means, the lower bound is found to be approximately -.25, while the upper bound is approximately .75. This clearly restricts the (conditional) support of the non-identified correlation, and as these identified correlations are rather precisely estimated, the resulting marginal posterior distribution of ρ_{10} should "live" mostly within these upper and lower limits.

In Figure 1, we provide graphical evidence to support this point. In the first row of this figure, we plot the priors and posteriors of the non-identified correlation parameter ρ_{10} for both the first (left) and second (right) components of the mixture. The priors were chosen to be identical across the components, so the prior plots are identical across the figures. For the second component (rightmost graphs), the priors and posteriors clearly differ, and the marginal posterior of ρ_{10} places virtually no mass to the left of -.25 and to the right of .75, which is consistent with our quick calculations for the values of the upper and lower bounds. *This clearly shows that learning about the identified correlations leads us to learn about this non-identified correlation through information conveyed in the p.d. restriction on* Σ. The first component, however, shows that the priors and posteriors for ρ_{10} are nearly identical, as the posterior distributions of the identified correlations do not yield informative support bounds. Thus, this application seems to be an ideal one for our purposes, as it *simultaneously illustrates cases where learning does and does not take place about the non-identified correlation parameter* ρ_{10}.

In the bottom portion of Figure 1 we present plots of priors and posteriors associated with the conditional support bounds $\underline{\rho}_{10}$ and $\overline{\rho}_{10}$. For both cases, the priors place a large mass over 1 or -1, reflecting relative non-informativeness *a priori* regarding values of the identified correlation

parameters ρ_{1D} and ρ_{0D}. That is, the choice of relatively non-informative priors for the identified correlations will not restrict the conditional support of the non-identified correlation, so that we see prior masses for the lower and upper bound clustered around -1 and 1, respectively. For the first component of the mixture (leftmost graphs), the priors and posteriors are quite similar, indicating that no information has been conveyed regarding the identified correlations which serves to limit or restrict these support bounds. However, for the second component of the mixture (rightmost graphs), the priors and posteriors of the bounds are quite different, suggesting that learning has taken place. Further, the lower bound is approximately centered at -.25, and the upper bound at .75, which is again consistent with our quick calculation at mean values.

Predictive Distributions of Outcome Gains: ATE and TT

In Figure 2, we plot the ATE and TT posterior predictive distributions of test score gains fixing the covariates at mean values. These predictives average over the predictions obtained from each component of the mixture, and as shown previously, the second component receives the vast majority of the weight. We have already shown that a substantial amount of learning takes place about ρ_{10} within this second component, and thus we will incorporate this learning into our unconditional predictives as in (19). For this application, ATE represents the loss in test scores from dropping out of high school for someone chosen at random, while TT represents the test score loss that exists for those actually dropping out of high school.

From the figure, we see that the ATE predictive is centered near -1 (specifically, its posterior mean is -.80), indicating that dropping out of high school clearly hurts student achievement on average. The size of the effect is quite large, as the sample standard deviation of the test scores is 1, since the senior-year mathematics test score was standardized to have unit variance. *Interestingly, when using our methods, we are also able to calculate quantities such as the posterior probability that (on average), dropping out of high school leads to a reduction in test scores*: $\Pr(\Delta_f < 0 | x_f = \overline{x}_f, \text{Data}) =$ .90. When looking at just mean effects, as is widely done in the program evaluation literature, such parameters can not be uncovered - their calculation requires the predictive distribution of outcome gains, which is the focus of our analysis.

The TT predictive in Figure 2 is shifted to the right relative to ATE, *indicating that the test score loss that occurs as a result of dropping out of high school is much smaller for those who actually decide to drop out of high school then for an average person.* Specifically, the posterior

mean of TT is approximately .24, suggesting that dropping out actually increases the test scores of the dropouts! We can not make this claim with any large degree of confidence, however, as the posterior probability that test sores *increase* from dropping out for those actually dropping out of high school is $\Pr(\Delta_f > 0 | x_f = \overline{x}_f, z_f = \overline{z}_f, D_f(\overline{z}_f) = 1, \text{Data}) = .69$.[7]

Note that what creates the right-shift of TT relative to ATE is the fact that the covariance between U^0 and U^D is very large, and large relative to the covariance between U^1 and U^D. This implies that *unobservable factors* making it more likely for an individual to drop out of school is strongly negatively correlated with her test scores if she remains in school, while these unobservables have a much smaller negative correlation with test score outcomes if she drops out of school. In other words, knowing that an individual would drop out conveys strong information that such an individual would likely get low test scores had she remained in school. As a result, the TT predictive distribution is shifted to the right relative to ATE. Said differently, ***our results suggest that staying in school matters in terms of test score outcomes, but it matters primarily for those who are inclined to stay in school and graduate.*** Thus, any intervention implemented with the intent of keeping individuals in school to raise their test scores is perhaps questionable, since those individuals who drop out are less likely to do well on tests if they were to remain in school. We are able to further support this claim by computing $\Pr(TT > ATE | \text{Data}) = .89$. Thus, even though the standard deviations associated with each predictive is quite large, we see strong evidence that dropouts benefit less from remaining in school in terms of test scores than an average student. Again, it is important to note that quantities like this one, which seem to have the significant policy-relevance, can not be obtained when looking only at mean effects.

Finally, we note that all of our previous discussion reported predictive results for the two-component mixture model, where we averaged over the predictions within each component as in (19). Since the second component of the model receives the vast majority of the weight, the results presented in Figure 2 will closely resemble the results obtained within this second component. Thus, our predictives largely characterize the impact of dropping out of high school for the "majority," while an inspection of Figure 2 does not by itself reveal any information about the impact of dropping out received by the "minority" group in the first component.

To this end we obtain the ATE and TT predictives associated with the first component of the mixture, and thus characterize the impact of dropping out of high school for the subgroup of

[7] Again, note that it is not possible to calculate this quantity with knowledge of only mean effects.

individuals ascribed to this component. That is, we plot in Figure 3 the *conditional* predictives given $c_{1f} = 1$ as in (14) and (15) rather than ones which marginalize over the component indicators.

What we see from Figure 3 is that for this "minority" group of individuals, selection is not empirically important - TT is not shifted relative to ATE (the posterior probability that ATE>TT was .45), and the two predictives look very similar. Further, both predictives are nearly centered at zero, suggesting that for this subpopulation, dropping out of high school does not obviously lower test scores for an average individual, or for those individuals that actually choose to drop out of high school. Both predictives are also quite "flat" in general, partly owing to the high variances associated with outcomes in this component of the mixture.

We can interpret the preference for the two component mixture model as the data revealing its preference for two different "groups" or "subpopulations" rather than just one. By looking at these groups *individually* instead of averaging over the group-specific predictions, we learn that the impact of dropping out of school is heterogeneous across the two groups. In the "majority" group, students clearly suffer on average as a result of dropping out of high school, while an average person chosen from the "minority" group does not appear to suffer substantially from dropping out. However, for those individuals in either group who actually decide to drop out of school, we find no convincing evidence that they would have benefited in terms of test scores had they remained in school.

7.1 Prior Sensitivity Analysis

To provide evidence that our key substantive conclusions are not affected by choice of prior, we present in Table 3 and Figures 4 and 5 estimation results using a different prior. In this prior, we choose the hyperparameters to center the non-identified correlation at .5, but leave the remaining priors essentially unchanged. Since the prior chosen for the non-identified correlation should have the most effect on our posterior results, we examine how our results are affected by changes in this prior. Centering the prior for ρ_{10} over .5 accords with a belief that individuals who perform well (or poorly) on tests in either the dropout or non-dropout state would also perform well (or poorly) in the other state. In short, as Heckman, Smith and Clements (1997, page 510) state, this prior reflects a seemingly reasonable and widespread belief that " ... good persons are good at whatever they do."

As shown in the table and figures, key conclusions are not affected by this choice of prior - selection remains empirically important, learning takes place about ρ_{10} in the second but not the first component of the mixture, test scores fall on average as a result of dropping out high school, and drop outs themselves do not seem to face a significant loss in test scores as a result of dropping out. This new prior does reduce the variance of our predictives slightly which is to be expected since, for example, $\text{Var}(ATE|x_f, \eta) = \gamma_2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{10}$. Expressing prior partiality for positive ρ_{10} (and thus σ_{10}) tends to reduce this posterior variability. Further, "conventional" treatment parameters which only look at mean effects are not affected by this change in prior since their expressions do not involve ρ_{10}. As such, this approach not only enables us to recover mean parameters which are commonly reported in studies on program evaluation, but also enables us to estimate a rich set of other quantities of policy-relevance.

8 Conclusion

In this paper we presented algorithms for fitting generalized non-Gaussian selection models, and discussed methods for calculating predictive distributions of outcome gains within the framework of each model.

The ability to recover *distributions* of outcome gains rather than simply *means* of those distributions enables researchers to obtain a new and rich set of quantities useful for policy evaluation. Extending our focus to distributions of outcome gains, however, is a non-trivial effort, since the distributions of interest depend on a non-identified correlation parameter. In this paper, we argued theoretically and illustrated in practice that learning can take place about this non-identified correlation, and this learning is incorporated into the calculation of predictive distributions of outcome gains resulting from the receipt of treatment. We illustrated these points for our generalized selection models, which include finite Normal mixture models, as well as models with Student$-t$ errors.

We applied our methods to estimate the impact of dropping out of high school on a senior-year mathematics test. This application is of significant economic interest, and also illustrated our econometric points extremely well. For this application, selection bias was an empirically important feature of our data, and non-Gaussian models were strongly preferred over the widely-used Gaussian model. Use of a two-component Gaussian mixture model illustrated situations in which learning

did and did not take place about the non-identified correlation parameter. For the component of the mixture receiving the vast majority of the weight, the priors and posteriors of this non-identified correlation differed considerably, as learning about the identified correlations lead us to update our beliefs about this non-identified correlation. Finally, we used our estimation results to calculate various predictive distributions of test score gains resulting from remaining in high school. We found that dropping out of high school has a significant negative impact on test scores on average, while the test score loss for the subgroup of individuals who actually drop out of high school is modest and nearly centered at zero.

References

[1] Abadie, A., Angrist, J. D. and G.W. Imbens (1998). "Instrumental Variables Estimation of Quantile Treatment Effects" *Econometrica*, forthcoming, and *NBER Technical Working Paper # 229.*

[2] Albert, J.H. and S. Chib (1993). "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669-679.

[3] Angrist, J.D. (1990). "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records" *American Economic Review* 80, 313-335.

[4] Angrist, J.D. (1998). "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants" *Econometrica* 66(2), 1998.

[5] Angrist, J. (2001), "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics* 19(1), 2-16.

[6] Angrist, J.D. and A.B. Krueger (1994). "Why do World War Two Veterans Earn More than Nonvetrans?" *Journal of Labor Economics* 12, 74-97.

[7] Bjorklund, A. and Moffitt, R. (1987), "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics* 69(1), 42-49.

[8] Card, D. (1990). "The Impact of the Mariel Boatlift on the Miami Labor Market" *Industrial and Labor Relations Review* 43(2), 245-257.

[9] Carlin, B and Polson, N. (1991). "Inference for nonconjugate Bayesian models using the Gibbs sampler," *Canadian Journal of Statistics,* 19, 399-405.

[10] Carter S.B., R.L. Ransom, R. Sutch, and H. Zhao (1993). *Codebook and User's Manual: A Survey of 943 Child Laborers in New Jersey, 1903*; Reported in the Twenty-Sixth Annual Report of the New Jersey Bureau of Statistics of Labor and Industries. Berkeley: Institute of Business and Economic Research.

[11] Chib, S. and B. Hamilton (2000), "Bayesian Analysis of Cross-Section and Clustered Data Treatment Models," *Journal of Econometrics* 97, 25-50.

[12] Dawid, A.P. (2000), "Casual Inference Without Counterfactuals," *Journal of the American Statistical Association* 95(450), 407-424.

[13] Dehejia, R. and S. Wahba (1999), "Casual Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94, 1053-1062.

[14] Dehejia, R. (2001), "When is ATE Enough? Rules of Thumb and Decision Analysis in Evaluating Training Programs," mimeo, Columbia University Department of Economics.

[15] Geweke J. (1993), "Bayesian Treatment of the Independent Student t Linear Model," *Journal of Applied Econometrics* 8, 19-40.

[16] Goldberger, A.S. (1983), "Abnormal Selection Bias" in *Studies in Econometrics, Time Series and Multivariate Statistics* S. Karlin et al, eds. New York: Academic Press.

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

[18] Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66(2), 315-331.

[19] Heckman, J. (1976), "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, 475-492.

[20] Heckman, J. (1990), "Varieties of Selection Bias," *American Economic Review Papers and Proceedings* 90(2), 313-318.

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

[22] Heckman, J., H. Ichimura, J. Smith and P. Todd (1998). "Characterizing Selection Bias Using Experimental Data" *Econometrica*, 66(5): 1017-1098.

[23] Heckman, J. and R. Robb. (1985), "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data* New York: Ca.

[24] Heckman, J. and J. Smith with 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.

[25] Heckman, J. and J. Smith (1998), " 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).

[26] Heckman, J., Tobias, J. and Vytlacil, E. (2000), "Simple Estimators for Treatment Parameters in a Latent Variable Framework with an Application to Estimating the Returns to Schooling," *NBER Working Paper #7950.*

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

[28] Heckman, J. and E. Vytlacil (2000a), "The Relationship Between Treatment Parameters within a Latent Variable Framework," *Economics Letters*, 33-39.

[29] Heckman, J. and E. Vytlacil (2000b), "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).

[30] Hirano, K., Imbens, G.W. and G. Ridder (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," mimeo, UCLA Department of Economics.

[31] Hirano, K., Imbens, G.W. and D. Rubin and X. Zhou (2000), "Assessing the Effect of an Influenza Vaccine in an Encouragement Design with Covariates " *Biostatistics* 1, 69-88.

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

[33] Imbens, G. and D.B. Rubin (1997), "Bayesian Inference for Casual Effects in Randomized Experiments with Noncompliance." *Annals of Statistics* 25 (1), 305-327.

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

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

[36] Lee, L-F (1983), "Generalized Econometric Models With Selectivity," *Econometrica* 51:2, 507-512.

[37] Manski, C. (1990), "Nonparametric Bounds on Treatment Effects" *American Economic Review Papers and Proceedings* 80, 319-323.

[38] Meghir, C. and M. Palme (1999). "Assessing the Effect of Schooling on Earnings Using a Social Experiment" Working Paper, University College London.

[39] McCulloch, R.E., N.G. Polson and P.E. Rossi (2000), "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," *Journal of Econometrics* 99(1), 173-193.

[40] McLachlan and Peel (2000). *Finite Mixture Models.* New York: Wiley.

[41] Nobile, A. (2000), "Comment: Bayesian Multinomial Probit Models with a Normalization Constraint" *Journal of Econometrics* 99(2), 335-345.

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

[43] Poirier, D.J. (1995), *Intermediate Statistics and Econometrics A Comparative Approach*, Cambridge: MIT Press.

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

[45] Poirier, D.J. and P.A. Ruud (1981), "On the Appropriateness of Endogenous Switching," *Journal of Econometrics* 16:249-256.

[46] Poirier, D.J. and P.A. Ruud (1983), "Diagnostic Testing in Missing Data Models," *International Economic Review* 24:537-546.

[47] Poirier, D.J. and J. L. Tobias (2001), "On the Predictive Distribution of Outcome Gains in the Presence of an Unidentified Parameter," Working Paper, UC-Irvine Department of Economics.

[48] Rosenbaum, P. and D. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Casual Effects," *Biometrica* 70, 41-55.

[49] Rousseeuw, P.J. and G. Molenberghs (1994), "The Shape of Correlation Matricies" *The American Statistician* 48, 276-279.

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

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

[52] Rubin, D. and Thayer (1978), "Relating Tests Given to Different Samples" *Psychometrika* 43, 3-10.

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

[54] Vytlacil, E. (2000), "Independence, Monotonicity, and Latent Variable Models: An Equivalence Result," *Econometrica*, forthcoming.

[55] Willis, R.J. and S. Rosen (1979). "Education and Self-Selection" *Journal of Political Economy* 87(2), S7-S36.

Appendix: Estimation

Student-t Models

As in Koop and Poirier (1997) and Chib and Hamilton (2000), we work with the *complete* or *augmented* outcome data. To this end, we let

$$r_i^* = \begin{bmatrix} D_i^* \\ D_i y_i + (1 - D_i) y_i^{Miss} \\ D_i y_i^{Miss} + (1 - D_i) y_i \end{bmatrix}$$

denote the "complete" set of outcomes for each individual. This consists of the latent desire for receipt of treatment (D^*), and both the observed and potential outcome (y_1 and y_0).

Recall that y_i denoted the *observed* outcome, and we will use y_i^{Miss} to denote the missing *unobserved* or *potential* outcome. Given our ordering of outcomes in (1)-(3), the above expression for r_i^* fills in the value y_i^{Miss} for $[y_1]_i$ when $D_i = 0$ (y_1 is unobserved), and fills in the missing value for $[y_0]_i$ when $D_i = 1$ (y_0 is unobserved). We also note that this particular representation is computationally convenient, as we will not have to determine if y_i^{Miss} should be plugged into the treated or untreated outcome - this is automatically handled through the construction of r_i^*.

We let k_x denote the length of the vector x, and define $k \equiv k_\theta + k_{\beta_1} + k_{\beta_2}$. We also let W_i be the $3 \times k$ matrix with z_i, x_i and x_i on the diagonal and let β denote the $k \times 1$ vector of associated parameters:

$$W_i = \begin{bmatrix} z_i & 0 & 0 \\ 0 & x_i & 0 \\ 0 & 0 & x_i \end{bmatrix}, \quad \beta = \begin{bmatrix} \theta \\ \beta_1 \\ \beta_0 \end{bmatrix}.$$

To specify a model with Student-t errors, it is convenient to work with a conditional Normal model for r_i^*:[8]

$$r_i^* | W_i, \beta, \lambda_i, \Sigma \overset{ind}{\sim} N(W_i\beta, \lambda_i \Sigma) \tag{20}$$

and add the following priors for the λ_i:

$$\lambda_i | v \overset{iid}{\sim} IG(v/2, 2/v), \tag{21}$$

[8]This addition of Gamma or inverted-Gamma mixing variables to the error variance to extend analyses to Student-t distributions, yet maintain computational tractability has been used in previous work by Carlin and Polson (1991), Albert and Chib (1993) Geweke (1993), and Chib and Hamilton (2001), among others.

where $IG(a,b)$ denotes an inverted gamma density with parameters a and b.[9] It follows, then, that marginalized over the mixing variables λ, the complete data follows a Student-t distribution:

$$r_i^*|W_i,\beta.\Sigma \overset{ind}{\sim} t_v(W_i\beta,\Sigma), \tag{22}$$

a multivariate Student-t distribution with v degrees of freedom, mean $W_i\beta$, and covariance matrix $[v/(v-2)]\Sigma$. We parameterize the elements of Σ as follows:

$$\Sigma = \begin{bmatrix} \sigma^2_{D^*} & \sigma_{1D} & \sigma_{0D} \\ \sigma_{1D} & \sigma^2_1 & \sigma_{10} \\ \sigma_{0D} & \sigma_{10} & \sigma^2_0 \end{bmatrix}.$$

To effect a Bayesian analysis, we add priors for the remaining parameters of the model. To this end, we specify the following forms for the priors:

$$\beta \sim N(\underline{\beta},\underline{V}_\beta) \tag{23}$$

$$\Sigma^{-1} \sim W(\underline{\rho},\underline{\rho R})I(\sigma^2_{D^*}=1) \tag{24}$$

The first line proposes a normal prior for the elements of β, while the second specifies a Wishart prior for the inverse covariance matrix Σ^{-1} subject to the normalization that the scale parameter in the "selection equation" is unity. We incorporate this restriction into our prior through the indicator variable $I(\sigma^2_{D^*}=1)$. We recognize at this point the contribution of Nobile (2000) who describes an algorithm for simulating from a Wishart density given such a restriction on a diagonal element of Σ. We thus use his method for simulating from this prior distribution for Σ^{-1} as well as for simulating draws from complete conditional of Σ^{-1} described below. We have also parameterized the Wishart so that (in the absence of the normalization) $E(\Sigma^{-1}) = R^{-1}$.

Given our assumed independence across observations, the joint posterior distribution of the latent desires for receipt of treatment (D^*), missing outcome data (y^{Miss}), regression parameters β and inverse covariance matrix Σ^{-1} is

$$p(\Gamma|\text{Data}) \propto \left[\prod_{i=1}^{n} \phi(r_i^*; W_i\beta, \lambda_i\Sigma)p_{IG}(\lambda_i)\right] \phi(\beta;\underline{\beta},\underline{V}_\beta)p_W(\Sigma^{-1})I(\sigma_{D^*}=1), \tag{25}$$

with Γ denoting all parameters and augmented data in the joint posterior, and $\phi(x;\mu,\Sigma)$ denoting the multivariate normal density for x with mean μ and covariance matrix Σ.

A Note on Computation

[9] In this paper, we parameterize the inverted gamma density as follows: If $x \sim IG(a,b)$ then $p(x) \propto x^{-(a+1)}\exp[-1/(bx)]$.

Our approach for fitting this model involves transforming the Student-t model back to the Gaussian case by dividing the y, D^*, x and z variables by $\sqrt{\lambda_i}$ in all of the non-λ conditionals. To this end, we let $\tilde{\cdot}$ denote quantities scaled by $\sqrt{\lambda_i}$, *e.g.* $\tilde{x_i} \equiv x_i/\sqrt{\lambda_i}$. We continue to let Γ denote all the parameters and augmented data in our model, and also let Γ_{-x} denote all parameters other than x.

1. To sample the augmented data, we need to obtain the missing outcome y_i^{Miss} as well as the latent desire for receipt of treatment D_i^* for all individuals i. Note that the missing outcome data is sampled by drawing from the following conditional posterior:

$$\tilde{y}_i^{Miss}|\Gamma_{-\tilde{y}_i^{Miss}}, \text{Data} \overset{ind}{\sim} N((1-D_i)\mu_{1i} + (D_i)\mu_{0i}, (1-D_i)\omega_{1i} + (D_i)\omega_{0i})$$

where

$$\mu_{1i} = \tilde{x_i}\beta_1 + (\tilde{D}_i^* - \tilde{z}_i\theta)\left[\frac{\sigma_0^2\sigma_{1D} - \sigma_{10}\sigma_{0D}}{\sigma_0^2 - \sigma_{0D}^2}\right] + (\tilde{y}_i - \tilde{x}_i\beta_0)\left[\frac{\sigma_{10} - \sigma_{0D}\sigma_{1D}}{\sigma_0^2 - \sigma_{0D}^2}\right] \quad (26)$$

$$\mu_{0i} = \tilde{x_i}\beta_0 + (\tilde{D}_i^* - \tilde{z}_i\theta)\left[\frac{\sigma_1^2\sigma_{0D} - \sigma_{10}\sigma_{1D}}{\sigma_1^2 - \sigma_{1D}^2}\right] + (\tilde{y}_i - \tilde{x}_i\beta_1)\left[\frac{\sigma_{10} - \sigma_{0D}\sigma_{1D}}{\sigma_1^2 - \sigma_{1D}^2}\right] \quad (27)$$

$$\omega_{1i} = \sigma_1^2 - \frac{\sigma_{1D}^2\sigma_0^2 - 2\sigma_{10}\sigma_{0D}\sigma_{1D} + \sigma_{10}^2}{\sigma_0^2 - \sigma_{0D}^2} \quad (28)$$

$$\omega_{0i} = \sigma_0^2 - \frac{\sigma_{0D}^2\sigma_1^2 - 2\sigma_{10}\sigma_{0D}\sigma_{1D} + \sigma_{10}^2}{\sigma_1^2 - \sigma_{1D}^2}. \quad (29)$$

As for the latent data $\tilde{D}_i^*$, it is also drawn from its conditional normal, though it is truncated by the observed value of D_i:

$$\tilde{D}_i^*|\Gamma_{-\tilde{D}_i^*}, \text{Data} \overset{ind}{\sim} \begin{cases} TN_{(0,\infty)}(\mu_{Di}, \omega_{Di}) & \text{if } D_i = 1 \\ TN_{(-\infty,0)}(\mu_{Di}, \omega_{Di}) & \text{if } D_i = 0 \end{cases},$$

where

$$\mu_{Di} = \tilde{z}_i\theta + (D_i\tilde{y}_i + (1-D_i)\tilde{y}_i^{Miss} - \tilde{x}_i\beta_1)\left[\frac{\sigma_0^2\sigma_{1D} - \sigma_{10}\sigma_{0D}}{\sigma_1^2\sigma_0^2 - \sigma_{10}^2}\right] + \quad (30)$$

$$((D_i)\tilde{y}_i^{Miss} + (1-D_i)\tilde{y}_i - \tilde{x}_i\beta_0)\left[\frac{\sigma_1^2\sigma_{0D} - \sigma_{10}\sigma_{1D}}{\sigma_1^2\sigma_0^2 - \sigma_{10}^2}\right], \quad (31)$$

$$\omega_{Di} = 1 - \frac{\sigma_{1D}^2\sigma_0^2 - 2\sigma_{10}\sigma_{0D}\sigma_{1D} + \sigma_1^2\sigma_{0D}^2}{\sigma_1^2\sigma_0^2 - \sigma_{10}^2} \quad (32)$$

and $TN_{(a,b)}(\mu, \sigma^2)$ denotes a univariate Normal density with mean μ and variance σ^2, truncated to the interval (a, b).

Given these drawn quantities, we then compute the complete data vector

$$\tilde{r}_i^* = \begin{bmatrix} \tilde{D}_i^* \\ D_i\tilde{y}_i + (1-D_i)\tilde{y}_i^{Miss} \\ D_i\tilde{y}_i^{Miss} + (1-D_i)\tilde{y}_i \end{bmatrix}.$$

2. The conditional posterior distribution of the stacked parameter vector $\beta \equiv [\beta_1' \ \beta_0' \ \theta']$, is straightforward:

$$\beta|\Gamma_{-\beta}, \text{Data} \quad \sim \quad N(\mu_\beta, \omega_\beta), \tag{33}$$

where

$$\mu_\beta = [\tilde{W}'(\Sigma^{-1} \otimes I_n)\tilde{W} + \underline{V}_\beta^{-1}]^{-1}[\tilde{W}'(\Sigma^{-1} \otimes I_n)\tilde{y} + \underline{V}_\beta^{-1}\underline{\beta}] \tag{34}$$

$$\omega_\beta = [\tilde{W}'(\Sigma^{-1} \otimes I_n)\tilde{W} + \underline{V}_\beta^{-1}]^{-1}, \tag{35}$$

where $\tilde{W}$ is the $3n \times k$ block diagonal matrix with $\tilde{Z}$, $\tilde{X}$, and $\tilde{X}$ stacked on the main diagonal, and $\tilde{y}$ is a $3n \times 1$ vector of the stacked $\tilde{D}^*$, $\tilde{y}_1^*$, and $\tilde{y}_0^*$ outcomes.

3. As for the inverse covariance matrix, Σ^{-1}, a slight complication is introduced as the complete conditional is no longer Wishart, given that the (1,1) element must be normalized to unity. We thus use the results of Nobile (2000) who provides an algorithm for drawing from such a Wishart, conditional on the (1,1) element being fixed to one. We express this conditional as

$$\Sigma^{-1}|\Gamma_{-\Sigma}, \text{Data} \sim W\left(n + \underline{\rho}, \left[\sum_{i=1}^{n}(\tilde{r}_i^* - \tilde{W}_i\beta)(\tilde{r}_i^* - \tilde{W}_i\beta)' + \underline{\rho}\underline{R}\right]\right) I(\sigma_{D^*}^2 = 1).$$

4. The λ mixing variables are updated by sampling them from the following inverted gamma density:

$$\lambda_i \mid \Gamma_{-\lambda_i}, \text{Data} \sim IG\left(\frac{v+3}{2}, \left[\frac{v + (r_i - W_i\beta)'\Sigma^{-1}(r_i - W_i\beta)}{2}\right]^{-1}\right), \quad i = 1, 2, \cdots n,$$

where we are using the *untransformed data* r_i and W_i rather than the scaled data $\tilde{r}_i$ and $\tilde{W}_i$.

5. The complete conditional for v takes the somewhat awkward form (*e.g.,* Albert and Chib (1993)):

$$v \mid \Gamma_{-v}, \text{Data} \propto p(v)\prod_{i=1}^{n}\left[\Gamma\left(v/2\right)\left(2/v\right)^{(v/2)}\right]^{-1}\lambda_i^{-(v/2+1)}\exp(-v/(2\lambda_i)),$$

with $p(\cdot)$ denoting the prior for the degrees of freedom parameter. Since this conditional is not easily sampled from, one could discretize the support of v, or use an additional Metropolis step. Alternatively, one can cycle through all but the last conditional after fixing a value of v *a priori*.

A Finite Mixture of Normals

Our second extension of the benchmark Normal model is to the case of a finite mixture of normals. Normal mixtures offer a very flexible modeling alternative, and can approximate multimodal, skewed and asymmetric densities quite accurately (see, *e.g.*, McLachlan and Peel (2000)).

In the finite mixture framework, the contribution of one individual to the likelihood is given as

$$p(r_i^*|\Gamma) = \sum_{g=1}^{G} \pi_g \phi(r_i^*; W_i\beta^g, \Sigma^g), \tag{36}$$

where we have allowed each component of the mixture to possess its own parameter vector β^g and covariance matrix Σ^g, and the π_g are the probabilities of being drawn from each component (*i.e.* $\sum_g \pi_g = 1$). We also define W_i as before to be the $3 \times k$ matrix with z_i along the first row, and the x_i vectors along the last two rows. Finally, we define $\beta^g \equiv [\theta^{g'} \ \beta_1^{g'} \ \beta_0^{g'}]'$.

In terms of estimation, it is desirable to first augment the parameter space with a set of component indicators, denoted $\{c_{gi}\}_{i=1}^n$. These indicator variables take the value of 1 to indicate that the i^{th} individual is drawn from the g^{th} component of the normal mixture, and are otherwise zero. In this case, the likelihood function for the *augmented data* r^* given the set of component label vectors $c = \{c_i\}_{i=1}^n$, $c_i = [c_{1i}\ c_{2i} \cdots c_{Gi}]$ is given as

$$p(r^*|c, \Gamma) = \prod_{i=1}^{n} [\phi(r_i^*; W_i\phi^1, \Sigma^1)]^{c_{1i}} [\phi(r_i^*; W_i\phi^2, \Sigma^2)]^{c_{2i}} \cdots [\phi(r_i^*; W_i\phi^G, \Sigma^G)]^{c_{Gi}}. \tag{37}$$

We also specify the following priors:

$$\begin{aligned}
p(c|\pi) &= \prod_{i=1}^{n} p(c_i|\pi) = \prod_{i=1}^{n}\prod_{g=1}^{G} \pi_g^{c_{gi}} && (38)\\
\pi &\sim \text{Dir}(\underline{\alpha}_1, \underline{\alpha}_2, \cdots, \underline{\alpha}_G) && (39)\\
\beta^g &\overset{ind}{\sim} N(\underline{\beta}^g, \underline{V}^g), \quad g = 1, 2, \cdots G && (40)\\
[\Sigma^g]^{-1} &\overset{ind}{\sim} W(\underline{\rho}^g, \underline{\rho}^g \underline{R}^g) I(\sigma_{D^*}^2 = 1), \quad g = 1, 2, \cdots G, && (41)
\end{aligned}$$

where "Dir" denotes the Dirichlet distribution (*e.g.* Poirier (1995), page 132), and $\pi = [\pi_1\ \pi_2\ \cdots\ (1 - \sum_{g=1}^{G-1} \pi_g)]$. As seen from the above, after integrating over the multinomial prior for the component indicators, we are left with the same density for each r_i^* in (37), so that the component indicators serve the practical purpose of facilitating computation. The joint posterior of the latent and missing data, component indicators, regression parameters and covariance matrices is given as the product of (38) - (42).

A Note on Estimation

Conditioned on the values of the component indicators $\{c_{gi}\}$, the data sorts itself into G different groups or blocks, and inference on the parameters within the blocks proceeds identically as in the textbook Gaussian model. Hence, the complete posterior conditionals for the regression parameters β^g and inverse covariance matrices $[\Sigma^g]^{-1}$ proceed identically to those described in the previous Student-t section, where the data "belonging to" each component are used to estimate the regression parameters and inverse covariance matrix associated with that component.[10]

However, it remains to derive the complete posterior conditionals for the component indicators $\{c_{gi}\}$ as well as the component probabilities π. The component indicator variables are drawn independently from their conditional multinomial distribution (*e.g.* Poirier (1995, page 118-119)):

$$c_i|\Gamma_{-c_i}, \text{Data} \overset{ind}{\sim} \text{Mult}\left(1, \frac{\pi_1|\Sigma^1|^{-1/2}\exp[-.5(r_i - W_i\beta^1)'(\Sigma^1)^{-1}(r_i - W_i\beta^1)]}{\sum_{g=1}^{G}\pi_g|\Sigma^g|^{-1/2}\exp[-.5(r_i - W_i\beta^g)'(\Sigma^g)^{-1}(r_i - W_i\beta^g)]}, \cdots, \right. \quad (42)$$

$$\left. \frac{\pi_G|\Sigma^G|^{-1/2}\exp[-.5(r_i - W_i\beta^G)'(\Sigma^G)^{-1}(r_i - W_i\beta^G)]}{\sum_{g=1}^{G}\pi_g|\Sigma^g|^{-1/2}\exp[-.5(r_i - W_i\beta^g)'(\Sigma^g)^{-1}(r_i - W_i\beta^g)]}\right). \quad (43)$$

The component probability vector π is drawn from the following conditional posterior:

$$\pi|\Gamma_{-\pi}, \text{Data} \sim \text{Dir}(n_1 + \underline{\alpha}_1, n_2 + \underline{\alpha}_2, \cdots, n_G + \underline{\alpha}_G), \quad (44)$$

where $n_g \equiv \sum_{i=1}^{n} c_{gi}$ denotes the number of people "in" the g^{th} component of the mixture.

[10] Of course, we no longer have to scale the data by the inverted-Gamma mixing variables λ.

Table 2: Posterior Means, Standard Deviations and
Probabilities of Being Positive: Two-Component Normal Mixture Model

	First Component			**Second Component**		
Variable/Posterior	Mean	Std.	$Pr(\cdot > 0\|D)$	Mean	Std.	$Pr(\cdot > 0\|D)$
Component Probability	0.0897	0.0129	1.000	0.910	0.0129	1.000
			Senior Test (Dropouts)			
Intercept	-0.181	0.189	0.165	-0.716	0.147	0
Base Math Score	0.555	0.0505	1.000	0.138	0.0549	0.987
Female	-0.0770	0.0612	0.106	0.000167	0.0379	0.508
White	0.135	0.0661	0.978	0.0250	0.0408	0.729
Father Education	0.0124	0.0112	0.867	-0.0116	0.00934	0.103
Mother Education	0.00109	0.0126	0.526	0.00220	0.00859	0.605
Family Income ($1000)	-0.00564	0.00293	0.0291	0.00331	0.00241	0.925
Number of Siblings	-0.0202	0.0177	0.124	-0.00570	0.0115	0.297
			Senior Test (Non-Dropouts)			
Intercept	-1.689	0.523	0.000208	-0.457	0.0396	0
Base Math Score	0.221	0.117	0.963	0.795	0.00753	1.000
Female	-0.0173	0.139	0.437	-0.0790	0.0126	0
White	0.545	0.181	0.999	0.110	0.0157	1.000
Father Education	0.0385	0.0242	0.947	0.0205	0.00230	1.000
Mother Education	0.0490	0.0296	0.956	0.0108	0.00271	1.000
Family Income ($1000)	0.00633	0.00760	0.805	0.00301	0.000655	1.000
Number of Siblings	-0.0176	0.0455	0.334	-0.00620	0.00393	0.0607
			Dropout Decision			
Intercept	0.815	0.395	0.983	-1.079	0.192	0
Base Math Score	-0.421	0.0887	0	-0.972	0.0587	0
Female	-0.0263	0.131	0.408	-0.0196	0.0560	0.365
White	-0.0143	0.157	0.471	0.0658	0.0605	0.859
Father Education	-0.0374	0.0208	0.0350	-0.0601	0.0120	0
Mother Education	-0.0535	0.0249	0.0137	-0.0558	0.0136	0.000042
Family Income ($1000)	-0.00215	0.00550	0.340	-0.00840	0.00295	0.00204
Number of Siblings	0.0839	0.0384	0.989	0.0711	0.0157	1.000
% Employ. Growth 80-82	0.0130	0.0172	0.773	0.00958	0.00608	0.940
			Correlations, Variances and Bounds			
$\Sigma_{1,1}$	0.268	0.0297	1.000	0.118	0.0143	1.000
$\Sigma_{0,0}$	0.834	0.0958	1.000	0.328	0.00777	1.000
$\rho_{1,0}$	-0.0399	0.318	0.451	0.287	0.131	0.986
$\rho_{1,D}$	0.108	0.233	0.670	-0.296	0.127	0.00454
$\rho_{0,D}$	-0.201	0.253	0.221	-0.849	0.0173	0
$\underline{\rho}_{1,0}$	-0.933	0.0929	0	-0.248	0.133	0.0468
$\overline{\rho}_{1,0}$	0.892	0.123	1.000	0.751	0.0900	1.000

Table 3: Posterior Means, Standard Deviations and
Probabilities of Being Positive: Alternate Prior with ρ_{10} Centered at .5

	First Component			**Second Component**		
Variable/Posterior	Mean	Std.	$Pr(\cdot > 0\|D)$	Mean	Std.	$Pr(\cdot > 0\|D)$
Component Probability	0.0881	0.00867	1.000	0.912	0.00867	1.000
			Senior Test (Dropouts)			
Intercept	-0.151	0.181	0.203	-0.569	0.160	0.000625
Base Math Score	0.581	0.0458	1.000	0.217	0.0456	1.000
Female	-0.0728	0.0626	0.124	-0.000518	0.0373	0.506
White	0.136	0.0699	0.972	0.0145	0.0424	0.634
Father Education	0.0152	0.0115	0.909	-0.00582	0.00903	0.268
Mother Education	0.00402	0.0126	0.623	0.00542	0.00909	0.735
Family Income ($1000)	-0.00579	0.00298	0.0278	0.00417	0.00235	0.963
Number of Siblings	-0.0252	0.0177	0.0731	-0.00980	0.0112	0.205
			Senior Test (Non-Dropouts)			
Intercept	-1.796	0.600	0.000063	-0.450	0.0394	0
Base Math Score	0.227	0.0858	0.998	0.794	0.00717	1.000
Female	0.0650	0.144	0.665	-0.0834	0.0132	0
White	0.629	0.179	1.000	0.107	0.0156	1.000
Father Education	0.0378	0.0265	0.929	0.0203	0.00231	1.000
Mother Education	0.0472	0.0314	0.933	0.0110	0.00268	1.000
Family Income ($1000)	0.00787	0.00703	0.864	0.00297	0.000641	1.000
Number of Siblings	-0.0120	0.0485	0.397	-0.00659	0.00397	0.0493
			Dropout Decision			
Intercept	0.875	0.412	0.988	-1.076	0.204	0
Base Math Score	-0.428	0.0874	0	-0.969	0.0543	0
Female	-0.0422	0.124	0.363	-0.0139	0.0568	0.388
White	-0.0370	0.166	0.418	0.0625	0.0602	0.856
Father Education	-0.0403	0.0208	0.0276	-0.0591	0.0125	0
Mother Education	-0.0524	0.0250	0.0171	-0.0563	0.0133	0
Family Income ($1000)	-0.00193	0.00577	0.366	-0.00861	0.00309	0.00162
Number of Siblings	0.0831	0.0378	0.988	0.0699	0.0165	1.000
% Employ. Growth 80-82	0.0132	0.0176	0.774	0.00897	0.00678	0.907
			Correlations, Variances and Bounds			
$\Sigma_{1,1}$	0.268	0.0293	1.000	0.139	0.0166	1.000
$\Sigma_{0,0}$	0.801	0.0899	1.000	0.331	0.00678	1.000
$\rho_{1,0}$	0.480	0.238	0.952	0.591	0.102	1.000
$\rho_{1,D}$	-0.0738	0.216	0.353	-0.523	0.0968	0
$\rho_{0,D}$	-0.153	0.266	0.297	-0.853	0.0170	0
$\underline{\rho}_{1,0}$	-0.892	0.136	0	0.00485	0.118	0.578
$\overline{\rho}_{1,0}$	0.958	0.0530	1.000	0.887	0.0565	1.000

Figure 1: Posterior (Solid) and Prior (Dashed) Distributions of ρ_{10} and its Upper ($\overline{\rho}_{10}$) and Lower ($\underline{\rho}_{10}$) Bounds: Two-Component Model

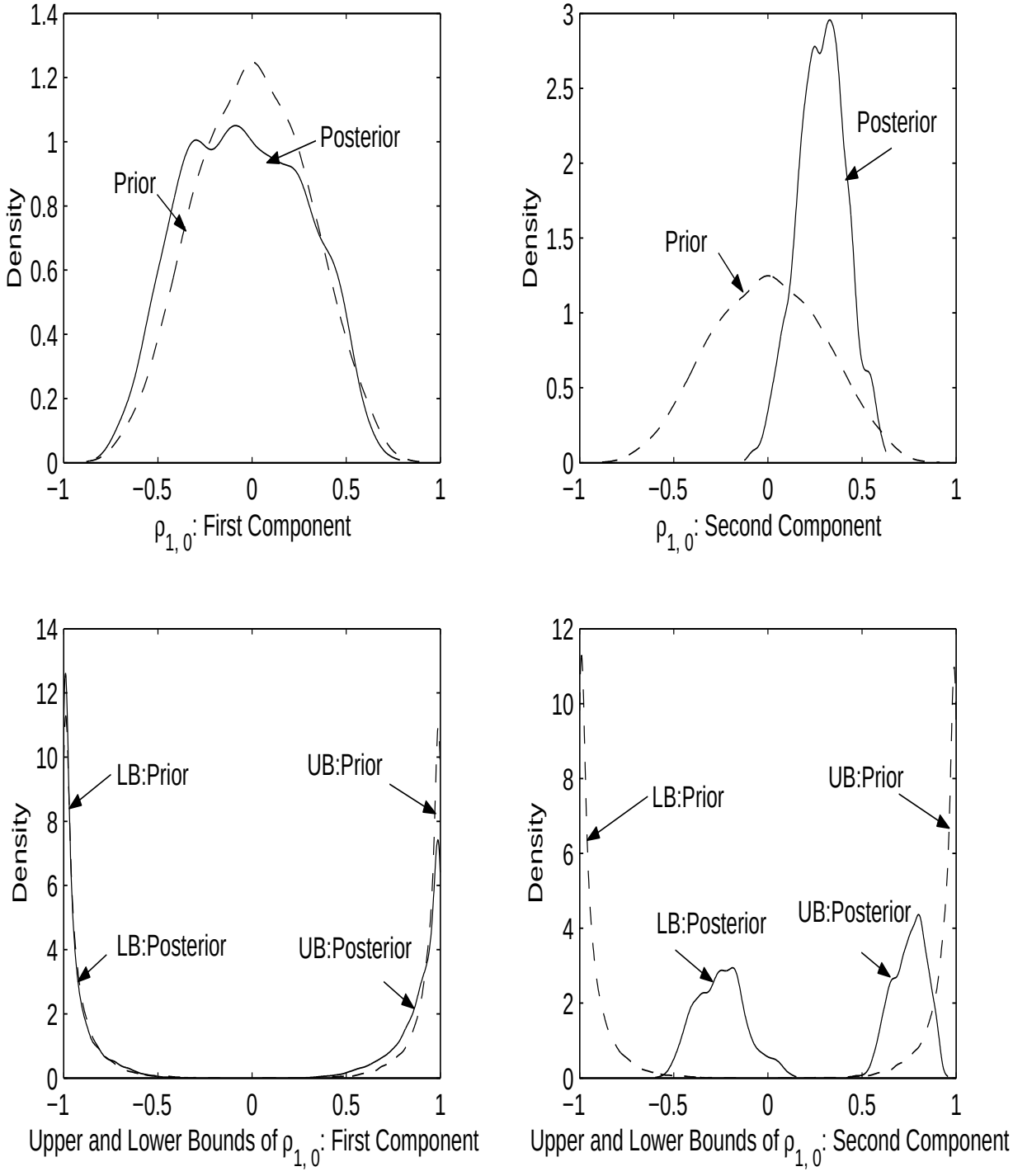

Figure 2: Predictive Distributions of Test Score Gain Resulting from Dropping Out of High School: ATE (Solid) and TT (Dashed). [Negative Values indicate a LOSS in Test Scores as a Result of Dropping Out]

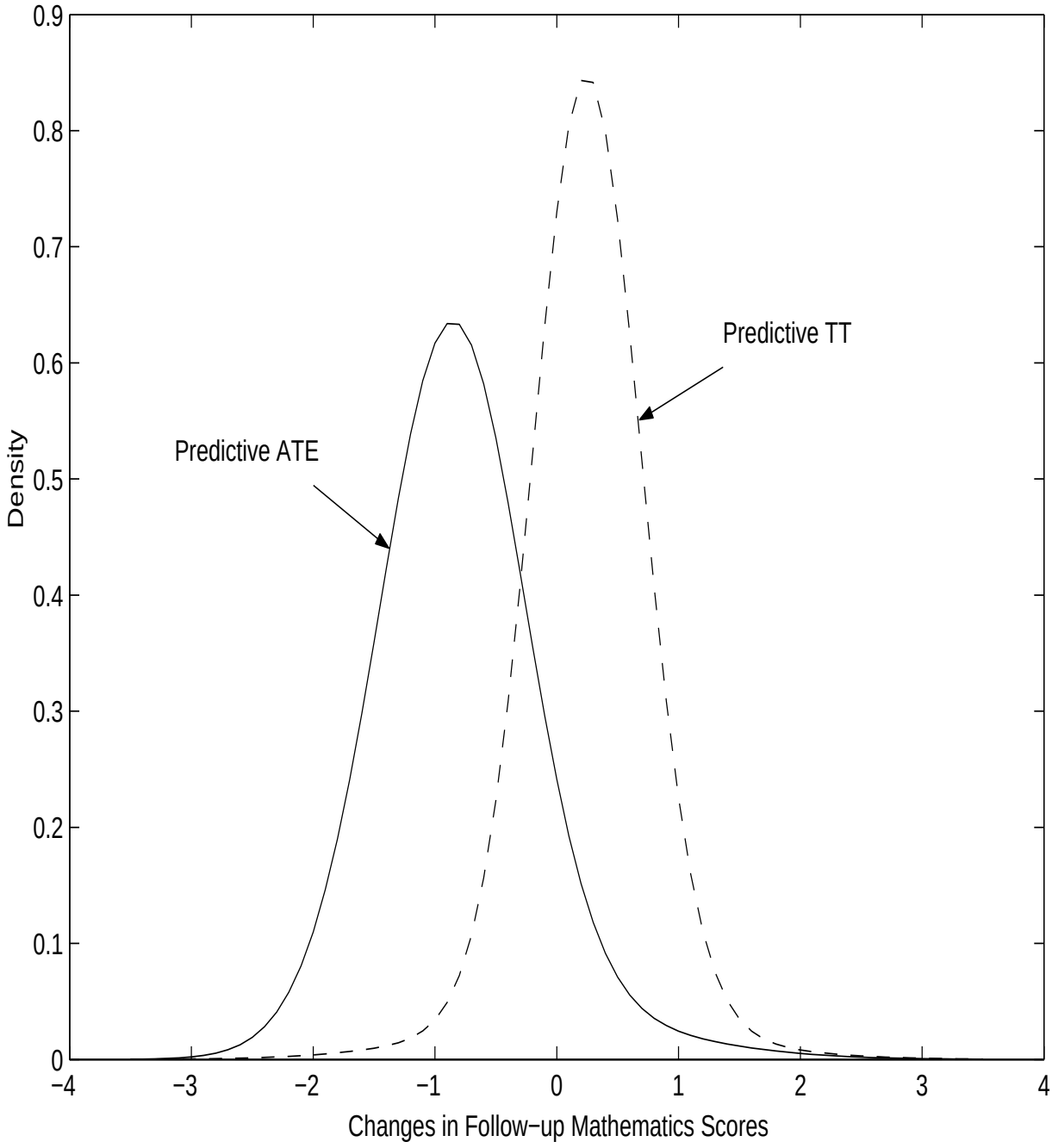

Figure 3: Predictive Distributions of Test Score Gain Resulting from Dropping Out of High School: ATE (Solid) and TT (Dashed) for FIRST COMPONENT Only. [Negative values indicate a LOSS in Test Scores as a Result of Dropping Out]

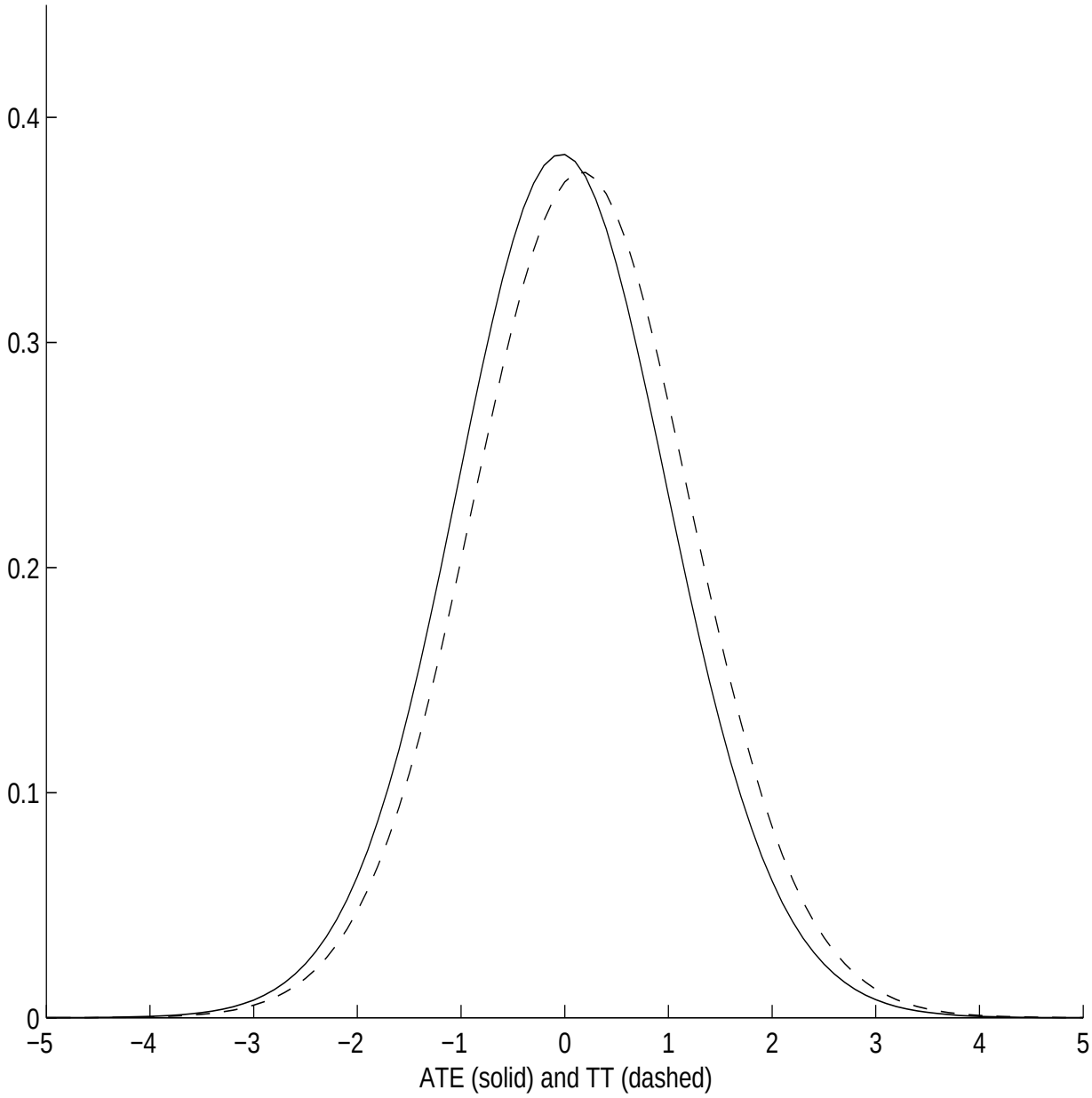

Figure 4: Posterior (Solid) and Prior (Dashed) Distributions of ρ_{10} and its Upper ($\overline{\rho}_{10}$) and Lower ($\underline{\rho}_{10}$) Bounds: Alternate Prior with ρ_{10} Centered at .5.

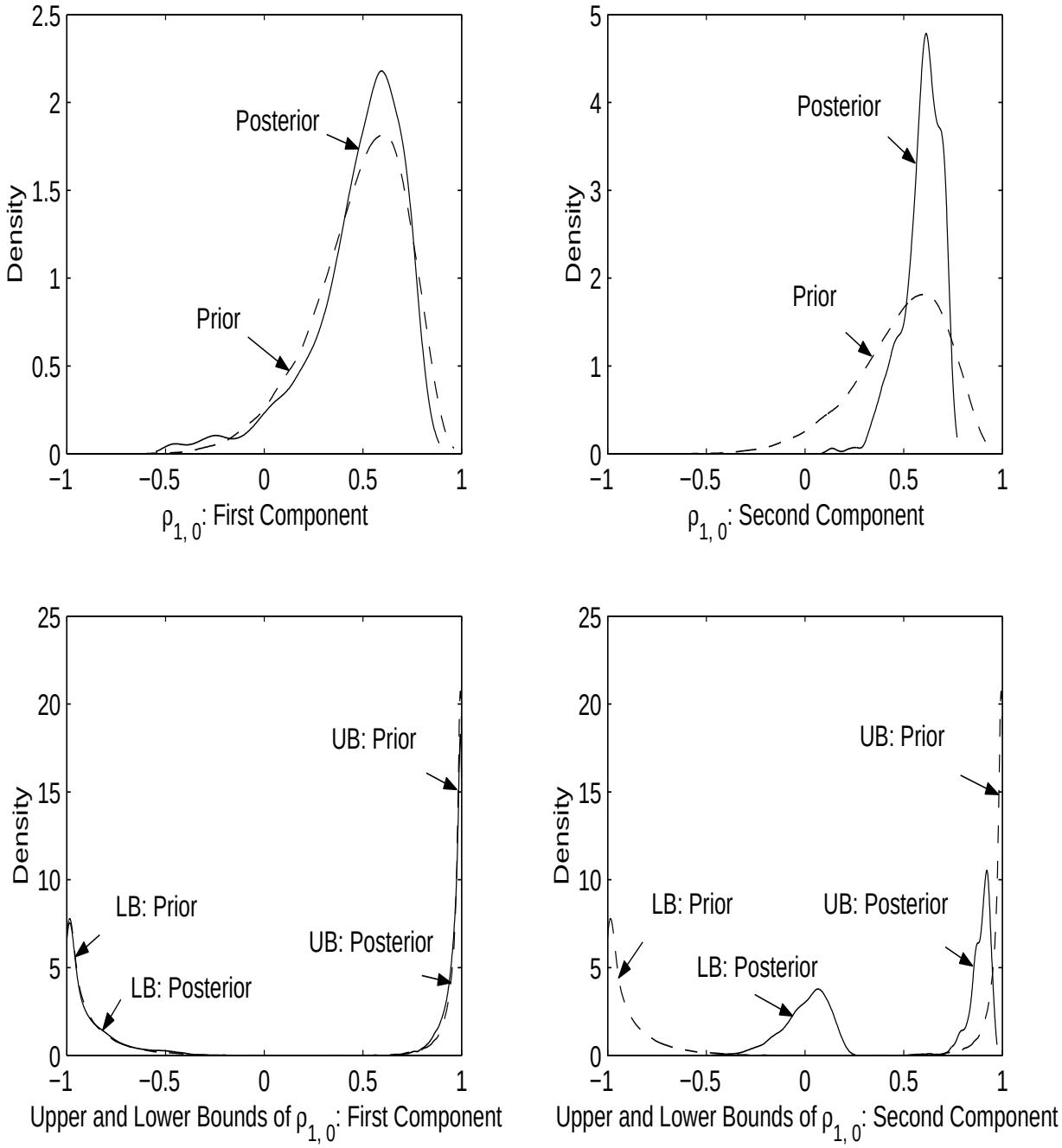

Figure 5: Predictive Distributions of Test Score Gain Resulting from Dropping Out of High School: ATE (Solid) and TT (Dashed): Alternate Prior with ρ_{10} Centered at .5. [Negative Values indicate a LOSS in Test Scores as a Result of Dropping Out.]