JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **19**: 203–225 (2004)
Published online 12 March 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jae.731

DO DROPOUTS SUFFER FROM DROPPING OUT? ESTIMATION AND PREDICTION OF OUTCOME GAINS IN GENERALIZED SELECTION MODELS

MINGLIANG LI, DALE J. POIRIER AND JUSTIN L. TOBIAS*
Department of Economics, University of California, Irvine, USA

SUMMARY

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 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. Copyright © 2004 John Wiley & Sons, Ltd.

1. INTRODUCTION

Since the early 1970s, 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. In recent work in this binary treatment/continuous outcome literature, considerable attention has been given 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., Heckman and Robb, 1985; Bjorklund and Moffitt, 1987; Heckman, 1990; Imbens and Angrist, 1994; Dehejia, 1999; Dehejia and Wahba, 1999; Heckman and Vytlacil, 1999, 2000; Heckman *et al.*, 2002). These treatment parameters measure various *expected* outcome gains from receipt of treatment for different subpopulations.

Despite the numerous and important advances made in the estimation of these parameters, 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 treatment impacts is attributable to a non-identified parameter problem. That is, for every individual in the sample, we will only observe his or her 'treated' or 'untreated' outcome, but never both, and thus the correlation between the treated and untreated outcomes is not identified. As a result,

* Correspondence to: Justin L. Tobias, Department of Economics, University of California-Irvine, 3151 Social Science Plaza, Irvine, CA 92697-5100, USA. E-mail: jtobias@uci.edu

Copyright © 2004 John Wiley & Sons, Ltd.
Received 26 April 2002
Revised 14 January 2003

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*.

Several methods have been advanced for dealing with this non-identified parameter problem. Heckman and Honoré (1990), Heckman *et al.* (1997), Heckman and Smith (1998),[1] Chib and Hamilton (2000) and Poirier and Tobias (2001) discuss the issue of estimating *distributions of outcome gains* in the presence of this unidentified parameter. Using a Bayesian approach, Chib and Hamilton (2000) discuss parametric within-sample distributions of outcome gains subject to the restriction that the non-identified correlation parameter is equal to zero, and Chib and Hamilton (2002) discuss semiparametric estimation of longitudinal data treatment effects under this assumption. Poirier and Tobias (2001) discuss how this prior restriction can be relaxed, 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 continue to 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*. 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 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. For these generalized models we derive expressions for various *posterior predictive distributions of outcome gains resulting from the receipt of treatment*. We link all of these predictive distributions to conventional *mean* treatment effects often used in the program evaluation literature. This includes the predictive distributions associated with the ATE, the TT, and the LATE.

Finally, and most importantly, 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). 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*.

Our results indicate 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 centred at zero. This suggests that policies aimed at keeping students in school

[1] In the context of the Roy (1951) model based on outcome maximization, Heckman *et al.* (1997) and Heckman and Smith (1998) move beyond mean effects and discuss non-parametric identification of the joint outcome distribution over a given support of the explanatory variables, thus enabling non-parametric identification of the treatment effects.

Copyright © 2004 John Wiley & Sons, Ltd.

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* with a binary treatment decision (D), and a continuous outcome (Y):[2]

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

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

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

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, without loss of generality, that the variables appearing in X are constant across states.

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 zero otherwise, $D(Z) = 1$ implies receipt of treatment, and $D(Z) = 0$ implies non-receipt. 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^* denote an element of Z which is not contained in X. Though this assumption is not strictly required for identification[3] (given some set of distributional assumptions), the practical importance of such an instrument has been widely documented. Further, the instrument itself will serve to define the

[2] We discuss estimation, learning and prediction in the context of this model with separable errors. For more on identification of treatment effects in models with non-separable errors, see, for example, Vytlacil (2000).

[3] See, for example, Heckman *et al*. (1997) and Heckman and Smith (1998) who discuss non-parametric identification in the context of the Roy (1951) model.

Copyright © 2004 John Wiley & Sons, Ltd.

LATE parameter (e.g., Imbens and Angrist, 1994; Heckman and Vytlacil, 1999, 2000), 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 $\rho_{10} \equiv \text{Corr}(U_1, U_0)$, though the *means* of these distributions will not. In the following section, we show that learning takes place about ρ_{10} and will use this learning to calculate various predictive outcome gain distributions.

3. LEARNING ABOUT THE NON-IDENTIFIED CORRELATION PARAMETER

In this section, we review and extend the arguments of Vijverberg (1993), Koop and Poirier (1997), Poirier (1998) 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:

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

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

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

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[4] $\underline{\rho}_{10} \leq \rho_{10} \leq \overline{\rho}_{10}$ where:

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

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

[4] In a similar spirit, Manski (1990, 1994) uses bounds on the outcomes and properties of the instrument to bound ATE. Here the source of the bounding is different, as it arises from the imposed positive definiteness of the covariance matrix.

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,

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

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 behaviour of the marginal posterior for ρ_{10}, first let $R(\rho_{10}) \subset [0, 1] \times [0, 1]$ 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

$$p(\rho_{10}|\text{Data}) \approx p(\rho_{10}|\rho_{1D} = \hat{\rho}_{1D}, \rho_{0D} = \hat{\rho}_{0D}), \qquad \Pr(\underline{\hat{\rho}}_{10} \leq \rho_{10} \leq \overline{\hat{\rho}}_{10}|\text{Data}) \approx 1 \tag{8}$$

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, then the marginal posterior of ρ_{10} would simply be the conditional prior evaluated at those limiting values of ρ_{1D} and ρ_{0D}. Though ρ_{10} is *not identified*, and its posterior will not collapse asymptotically, it is important to recognize that we still update our beliefs regarding this parameter through the p.d. restriction. The upper and lower conditional support bounds implied by this restriction are identified, and will serve to limit the conditional support of ρ_{10}.[5]

[5] Interestingly, note that if treatment were randomly assigned in the sense that $\rho_{1D} = \rho_{0D} = 0$, the conditional support bounds would be completely uninformative. Intuitively, this makes sense as we learn about ρ_{10} 'indirectly' through the outcomes correlations with the treatment decision. That is, if the outcome unobservables move together sufficiently with the selection unobservables, then to some degree they must also move together with each other. In randomized experiments, the design itself breaks the outcomes association with the treatment decision, and thus provides no vehicle for learning about ρ_{10}.

4. ESTIMATION IN NON-GAUSSIAN SELECTION MODELS

The model described in (1)–(3) is estimated using the Gibbs sampler, and we describe the details of our algorithm in the appendix. To generalize our previous algorithm (Poirier and Tobias, 2002) which assumed Normality, we employ standard computational 'tricks' to extend our analysis to multivariate Student-*t* errors (e.g., Carlin and Polson, 1991; Albert and Chib, 1993; Geweke, 1993) or to a finite mixture of Normals (e.g., Chib and Hamilton, 2000; McLachlan and Peel, 2000).

What is important to recognize is that the algorithms we describe work with the 'full' 3×3 covariance matrix, and as such, permit learning to take place about ρ_{10}. The resulting algorithms when working with the 'full' covariance matrix also turn out to be quite 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

Since our goal is to use the given data to *predict* distributions of outcome gains for future populations, it remains to discuss how the predictive distributions are calculated in our generalized selection models.

To these ends, suppose the model in (1)–(3) applies to our future population, and define

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

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.[6] The means of these distributions correspond to the ATE, the TT (e.g., Rubin, 1978; Heckman and Robb, 1985), and the LATE (e.g., Imbens and Angrist, 1994) predictive parameters, respectively, widely reported in the program evaluation literature.

We will calculate these predictives by marginalizing the conditional predictive densities over the joint posterior:

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

[6] The change from z_f to $\tilde{z}_f$ results from a change in the instrument from z^* to $\tilde{z}^*$. 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.

Copyright © 2004 John Wiley & Sons, Ltd.

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

$$p(\Delta_f|x_f, z_f, \tilde{z}_f, D_f(\tilde{z}_f) = 1, D_f(z_f) = 0, \text{Data}) \\ = \int_\eta [p(\Delta_f|x_f, z_f, \tilde{z}_f, D_f(\tilde{z}_f) = 1, D_f(z_f) = 0, \text{Data}, \eta) \\ \times p(\eta|z_f, \tilde{z}_f, D_f(\tilde{z}_f) = 1, D_f(z_f) = 0, \text{Data})]d\eta$$

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 straightforward to show that

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

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

These results can be substituted into the expressions above so that in all cases we perform the integration over the joint posterior $p(\eta|\text{Data})$. 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.

5.1. Student-*t* Predictives

With a bit of work, we can derive expressions for the conditional predictives above within our model with Student-*t* errors:

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

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:

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

$$[LATE] = p(\Delta_f|x_f, z_f, \tilde{z}_f, D_f(z_f) = 0, D_f(\tilde{z}_f) = 1, \eta)$$

Copyright © 2004 John Wiley & Sons, Ltd.

$$= \left(\frac{p_t(\Delta_f|x_f,\eta)}{T_v(\tilde{z}_f\theta) - T_v(z_f\theta)}\right)\left[T_{v+1}\left(\frac{\mu_{D^*|\Delta_f}(x_f,\tilde{z}_f,\Delta_f,\eta)}{\sqrt{\Omega_{D^*|\Delta_f}(x_f,\Delta_f,v,\eta)}}\right) - T_{v+1}\left(\frac{\mu_{D^*|\Delta_f}(x_f,z_f,\Delta_f,\eta)}{\sqrt{\Omega_{D^*|\Delta_f}(x_f,\Delta_f,v,\eta)}}\right)\right]$$

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 ith 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 can be 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 introduce 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 ith individual is drawn from the gth component of the mixture, and is otherwise zero.

In terms of prediction, *conditioned on the future component indicator value* $c_{gf} = 1$ 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. Thus, we obtain (Poirier and Tobias, 2002):

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

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

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

Copyright © 2004 John Wiley & Sons, Ltd.

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 gth component of the mixture.

The desired predictives *given the parameters but marginalized over the component indicators* follow as a weighted average of the conditional predictives above, where the component probabilities serve as the weights. Focusing on ATE as an example we note:

$$p(\Delta_f|x_f, \text{Data}) = \int_\eta p(\Delta_f|x_f, \eta, \text{Data})p(\eta|\text{Data})d\eta \tag{17}$$

$$= \int_\eta \left[\sum_{g=1}^{G} p(\Delta_f|x_f, \eta, c_{gf} = 1, \text{Data})\Pr(c_{gf} = 1|\eta, \text{Data})\right] p(\eta|\text{Data})d\eta \tag{18}$$

$$= \int_\eta \sum_{g=1}^{G}[\pi_g p(\Delta_f|x_f, \eta, c_{gf} = 1, \text{Data})]p(\eta|\text{Data})d\eta \tag{19}$$

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 (14)–(16). 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 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.[7] 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%) of this original cohort will be observed to drop out 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

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

Copyright © 2004 John Wiley & Sons, Ltd.

for *both individuals who drop out of high school as well as those that do not*. 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 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) and selection 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.

Our excluded variable which enters the dropout equation but does not appear in the outcome (senior test score) equations is the percentage of employment growth in the local labour market over the period 1980–1982. Our expectation is that a large amount of local employment growth over this period suggests prosperous local labour 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 labour 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.[8]

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.[9] Among this final sample, approximately 8.1% of the individuals (1006) dropped out of high school 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 question (1) we will calculate the posterior predictive

[8] 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.

[9] Over 12% of the original observations were from private schools and were discarded. When students moved between their sophomore and senior years, no follow-up information was available, and 40 of the schools originally sampled dropped out of the survey. Approximately 4% of the remaining students were not re-surveyed in the follow-up. Parental education and parental income had a high non-response rate, though the distributions of these variables in our final selected sample were found to be similar to those obtained in the entire sample. We do not take up the issue of modelling the missing observations, as this would additionally require us to model the process generating these observations.

Copyright © 2004 John Wiley & Sons, Ltd.

distribution of test score gains for a randomly chosen person (ATE), and for question (2) we will calculate this distribution for those who actually drop out (TT).[10]

Throughout this analysis we specify independent priors of the form

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

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

where $\beta = [\theta\prime\, \beta_1'\, \beta_0']'$ (as defined in (1)–(3)). The first line proposes a Normal prior for the elements of β, while the second specifies a Wishart prior[11] for the inverse covariance matrix Σ^{-1} subject to the normalization that the scale parameter in the selection equation is unity. We set all elements of the prior mean vector $\underline{\beta}$ to zero, but set the coefficient associated with the employment growth rate to 0.01 to reflect our prior expectation that more favourable local labour market conditions will induce some marginal individuals to drop out of school. The employed prior is quite diffuse, as we set the prior standard deviation of each element equal to 2 (setting $\underline{V}_\beta = 4I_k$, with I_k denoting the $k \times k$ identity matrix), so that the data information is predominant. As for the prior for the inverse covariance matrix, we set $\rho = 12$, and centre the variance parameters in both outcome equations over 0.25. All correlation parameters in R are set equal to zero so that our prior 'centres' our model over one where selection bias is not important, though our prior is diffuse enough to let the data revise our beliefs and reveal to us the importance of unobservable selectivity.

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 favour 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 favoured by the data. The results of these marginal likelihood calculations are presented in Table I.

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. 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

Table I. Log marginal likelihoods, and posterior model probabilities for alternate models

All models	Log marginal likelihoods	Model Pr (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

[10] It would also be possible to calculate the predictive joint distribution of (y_1, y_0), and thereby calculate a variety of other quantities of interest. Here we restrict our attention to the outcome gain distributions described previously.

[11] For the Normal mixture models, we specify identical priors of this form for each mixture component. We also parameterize the Wishart so that (in the absence of the normalization) $E(\Sigma^{-1}) = \underline{R}^{-1}$.

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.

Estimation results from the two-component mixture are presented in Table II. To interpret these and subsequent results, we regard the decision to drop out as the decision to receive 'treatment' so that, in the notation of 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. Thus, negative values associated with the treatment effect $\Delta \equiv Y^1 - Y^0$ indicate a reduction in test scores as a result of dropping out.

As shown in the first row of Table II, the second component receives the vast majority of the weight, as the posterior mean of the probability associated with this second component was 0.91.

Table II. Posterior means, standard deviations and probabilities of being positive: two-component Normal mixture model

Variable/posterior	First component			Second component		
	Mean	Std.	$\Pr(\cdot > 0 \mid D)$	Mean	Std.	$\Pr(\cdot > 0 \mid D)$
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
% Employment 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

The remaining 9% 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 directions of the effects suggested by Table II 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-dropout equation 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 equation describing the dropout decision, 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.

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 -0.25, while the upper bound is approximately 0.75. 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. For the second component (right-most graphs), the priors and posteriors clearly

Copyright © 2004 John Wiley & Sons, Ltd.

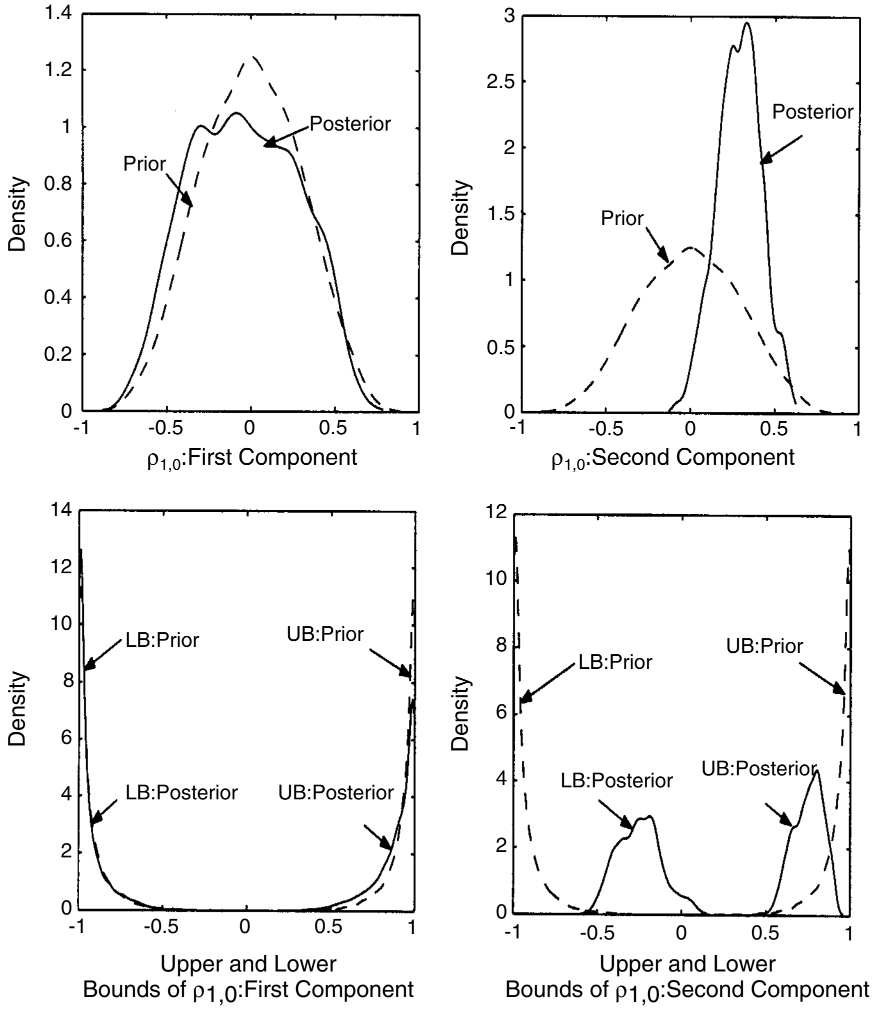

Figure 1. Posterior (solid) and prior (dashed) distributions of ρ_{10} and its upper ($\overline{\rho}_{10}$) and lower ($\underline{\rho}_{10}$) bounds: two-component model

differ, and the marginal posterior of ρ_{10} places virtually no mass to the left of -0.25 and to the right of 0.75, 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 the 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}. For the first component of the mixture (left-most graphs), the priors and posteriors are quite similar, indicating that no information has been conveyed regarding the identified correlations which serve to limit or restrict these support bounds. However, for the second component of the mixture (right-most graphs), the priors and posteriors of the bounds are quite different, suggesting that learning has taken place. Further, the lower bound is approximately centred at -0.25, and the upper bound at 0.75, which is again consistent with our quick calculation at posterior mean values.

7.1. Predictive Distributions of Outcome Gains: ATE and TT

In Figure 2 we plot the ATE and TT posterior predictive distributions of test score gains for white males, fixing the continuous covariates at mean values and rounding the means of the integer-valued variables to the nearest integer.[12] 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.

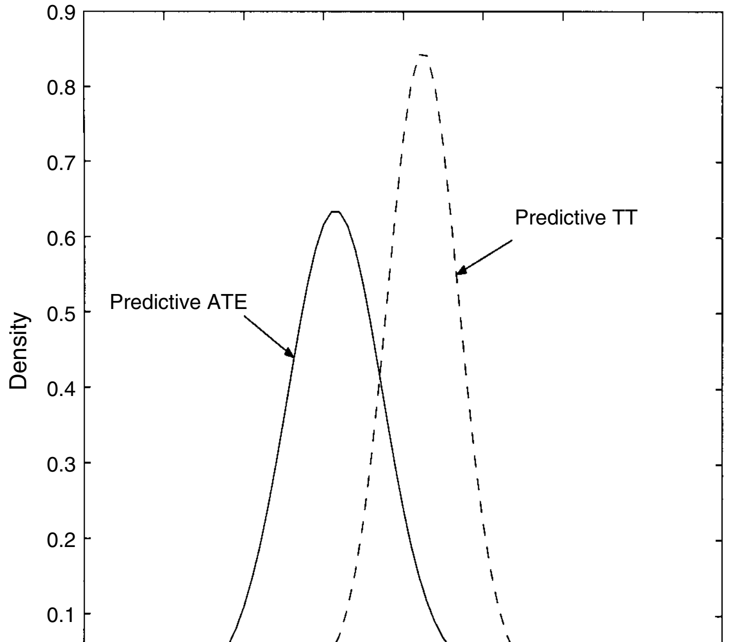

Figure 2. Predictive distributions of test score gain resulting from dropping out of high school: ATE (solid) and TT (dashed) [negative values indicate a LOSS in test scores as a result of dropping out]

[12] Alternatively, one could also marginalize over the covariates by specifying some distribution for their future values.

Copyright © 2004 John Wiley & Sons, Ltd.

From the figure, we see that the ATE predictive is centred near -1 (specifically, its posterior mean is -0.84), indicating that dropping out of high school clearly hurts student achievement on average. The size of the effect is quite large, as the test scores have been 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*: $\Pr(\Delta_f < 0|x_f = \overline{x}_f, \text{ Data}) = 0.90$. When looking at just mean effects, as is widely done in the program evaluation literature, such parameters cannot 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 than for an average person*. Specifically, the posterior mean of TT is approximately 0.19, suggesting that dropping out actually increases the test scores of the dropouts! We cannot make this claim with any large degree of confidence, however, as the posterior probability that test scores *increase* from dropping out for those actually dropping out of high school is $\Pr(\Delta_f > 0|x_f = \overline{x}_f, z_f = \overline{z}_f, D_f(\overline{z}_f) = 1, \text{ Data}) = 0.65$.[13]

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 are 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. 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 $\Pr(TT > ATE|\text{Data}) = 0.88$. Thus, even though the standard deviations associated with each predictive are 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, cannot be obtained when looking only at mean effects.

7.2. Prior Sensitivity Analysis

To provide evidence that our key substantive conclusions were not affected by choice of prior for the non-identified parameter, we re-estimated our model using a different prior. In this prior, we choose the hyperparameters to centre the non-identified correlation at 0.5, but leave the remaining prior hyperparameters unchanged. Centring the prior for ρ_{10} over 0.5 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 *et al.* (1997, p. 510) state, this prior reflects a seemingly reasonable and widespread belief that '... good persons are good at whatever they do'.

[13] Again, note that these probabilities are calculated for white males and integer-valued variables are rounded to the nearest integer.

Copyright © 2004 John Wiley & Sons, Ltd.

We do not report the results of this sensitivity analysis here,[14] though we note that key conclusions were not affected by this choice of prior. In particular, selection remained empirically important, learning took place about ρ_{10} in the second but not the first component of the mixture, test scores fell on average as a result of dropping out of high school, and dropouts themselves did not seem to face a significant loss in test scores as a result of dropping out. The posterior results were also found to be robust to moderate changes in the remaining prior hyperparameters, including changes in ρ and $\underline{V}_{\beta}$.[15]

It is also important to recognize that 'conventional' treatment parameters which only look at mean effects are not significantly affected by this change in prior since their expressions do not involve ρ_{10}. As such, the approach described in this paper 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

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 empirically that learning can take place about this non-identified correlation parameter.

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. Using a two-component Normal mixture, we showed that the priors and posteriors of this non-identified correlation differed considerably, as learning about the identified correlations led us to update our beliefs about this non-identified correlation. 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 centred at zero.

APPENDIX: ESTIMATION

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

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

[14] Parameter posterior means and standard deviations, marginal posteriors of ρ_{10}, $\overline{\rho}_{10}$ and $\underline{\rho}_{10}$, and ATE and TT predictives using this prior can be obtained at http://orion.uci.edu/~jtobias/Dropouts.

[15] Specifically, multiplying $\underline{V}_{\beta}$ by 10 or 0.1 had a minimal effect on our results, as did setting $\rho = 5$ or $\rho = 20$.

Copyright © 2004 John Wiley & Sons, Ltd.

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. This particular formulation is computationally convenient, as it automatically determines if y_i^{Miss} should be plugged into the treated or untreated outcome.

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:

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

Student-*t* Models

To specify a model with Student-*t* errors, it proves to be convenient to work with a conditional Normal model for r_i^*:[16]

$$r_i^* | W_i, \beta, \lambda_i, \Sigma \overset{ind}{\sim} N(W_i\beta, \lambda_i\Sigma) \tag{A.1}$$

and add the following hierarchical priors for the λ_i:

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

where $IG(a, b)$ denotes an inverted gamma density with parameters a and b.[17] It follows, then, that marginalized over the mixing variables λ, the complete data follows a Student-*t* distribution:

$$r_i^* | W_i, \beta, \Sigma \overset{ind}{\sim} t_v(W_i\beta, \Sigma) \tag{A.3}$$

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:

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

Priors for β and Σ^{-1} are specified as in Section 7. Given the 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

$$p(\Gamma|\text{Data}) \propto \left[\prod_{i=1}^{n} \phi(r_i^*; W_i\beta, \lambda_i\Sigma)p_{IG}(\lambda_i)\right] \phi(\beta; \underline{\beta}, \underline{V}_\beta)p_W(\Sigma^{-1})I(\sigma_{D*} = 1) \tag{A.4}$$

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.

[16] 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 (2000), among others.

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

A Note on Computation

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) Posterior conditionals for augmented data y_i^{Miss} and D_i^*:

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

where

$$\mu_{1i} = \tilde{x}_i\beta_1 + (\tilde{D}_i^* - \tilde{z}_i\theta)\left[\frac{\sigma_0^2\sigma_{1D} - \sigma_{10}\sigma_{0D}}{\sigma_0^2 - \sigma_{0D}^2}\right] + (\tilde{y}_i - \tilde{x}_i\beta_0)\left[\frac{\sigma_{10} - \sigma_{0D}\sigma_{1D}}{\sigma_0^2 - \sigma_{0D}^2}\right] \quad \text{(A.5)}$$

$$\mu_{0i} = \tilde{x}_i\beta_0 + (\tilde{D}_i^* - \tilde{z}_i\theta)\left[\frac{\sigma_1^2\sigma_{0D} - \sigma_{10}\sigma_{1D}}{\sigma_1^2 - \sigma_{1D}^2}\right] + (\tilde{y}_i - \tilde{x}_i\beta_1)\left[\frac{\sigma_{10} - \sigma_{0D}\sigma_{1D}}{\sigma_1^2 - \sigma_{1D}^2}\right] \quad \text{(A.6)}$$

$$w_{1i} = \sigma_1^2 - \frac{\sigma_{1D}^2\sigma_0^2 - 2\sigma_{10}\sigma_{0D}\sigma_{1D} + \sigma_{10}^2}{\sigma_0^2 - \sigma_{0D}^2} \quad \text{(A.7)}$$

$$w_{0i} = \sigma_0^2 - \frac{\sigma_{0D}^2\sigma_1^2 - 2\sigma_{10}\sigma_{0D}\sigma_{1D} + \sigma_{10}^2}{\sigma_1^2 - \sigma_{1D}^2} \quad \text{(A.8)}$$

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

$$\mu_{Di} = \tilde{z}_i\theta + (D_i\tilde{y}_i + (1-D_i)\tilde{y}_i^{Miss} - \tilde{x}_i\beta_1)\left[\frac{\sigma_0^2\sigma_{1D} - \sigma_{10}\sigma_{0D}}{\sigma_1^2\sigma_0^2 - \sigma_{10}^2}\right] \quad \text{(A.9)}$$

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

$$\omega_{Di} = 1 - \frac{\sigma_{1D}^2\sigma_0^2 - 2\sigma_{10}\sigma_{0D}\sigma_{1D} + \sigma_1^2\sigma_{0D}^2}{\sigma_1^2\sigma_0^2 - \sigma_{10}^2} \quad \text{(A.11)}$$

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) Complete conditional for $\beta \equiv [\theta' \beta_1' \beta_0']$,:

$$\beta | \Gamma_{-\beta}, \text{Data} \sim N(\mu_\beta, \omega_\beta) \tag{A.12}$$

where

$$\mu_\beta = [\tilde{W}'(\Sigma^{-1} \otimes I_n)\tilde{W} + \underline{V}_\beta^{-1}]^{-1}[\tilde{W}'(\Sigma^{-1} \otimes I_n)\tilde{y} + \underline{V}_\beta^{-1}\underline{\beta}] \tag{A.13}$$

$$\omega_\beta = [\tilde{W}'(\Sigma^{-1} \otimes I_n)\tilde{W} + \underline{V}_\beta^{-1}]^{-1} \tag{A.14}$$

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) Complete conditional for 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 a convenient algorithm for drawing a Wishart conditional on a diagonal element. 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) Complete conditional for $\{\lambda_i\}$:

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

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) Complete conditional for v (e.g., Albert and Chib, 1993):

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

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

In the finite mixture framework (see, e.g., McLachlan and Peel, 2000), the contribution of one individual to the likelihood is given as

$$p(r_i^* | \Gamma) = \sum_{g=1}^{G} \pi_g \phi(r_i^*; W_i\beta^g, \Sigma^g) \tag{A.15}$$

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

Copyright © 2004 John Wiley & Sons, Ltd.

(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 one to indicate that the ith individual is drawn from the gth component of the Normal mixture, and are zero otherwise. 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

$$p(r^*|c, \Gamma) = \prod_{i=1}^{n} [\phi(r_i^*; W_i\phi^1, \Sigma^1)]^{c_{1i}} [\phi(r_i^*; W_i\phi^2, \Sigma^2)]^{c_{2i}} \cdots [\phi(r_i^*; W_i\phi^G, \Sigma^G)]^{c_{Gi}} \tag{A.16}$$

We also specify the following priors:

$$p(c|\pi) = \prod_{i=1}^{n} p(c_i|\pi) = \prod_{i=1}^{n}\prod_{g=1}^{G} \pi_g^{c_{gi}} \tag{A.17}$$

$$\pi \sim \text{Dir}(\underline{\alpha}_1, \underline{\alpha}_2, \cdots, \underline{\alpha}_G) \tag{A.18}$$

$$\beta^g \overset{ind}{\sim} N(\underline{\beta}^g, \underline{V}^g), \quad g = 1, 2, \cdots, G \tag{A.19}$$

$$[\Sigma^g]^{-1} \overset{ind}{\sim} W(\underline{\rho}^g, \underline{\rho}^g \underline{R}^g) I(\sigma_{D^*}^2 = 1), \quad g = 1, 2, \cdots, G \tag{A.20}$$

where 'Dir' denotes the Dirichlet distribution (e.g., Poirier, 1995, p. 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 (A.15), 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 (A.16)–(A.20).

Conditioned on the values of the component indicators, 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.[18]

In addition, we obtain the following complete posterior conditionals:

(1) Complete conditional for $\{c_i\}$:

$$c_i|\Gamma_{-c_i}, \text{Data} \overset{ind}{\sim} \text{Mult}\left(1, \frac{\pi_1 |\Sigma^1|^{-1/2} \exp[-0.5(r_i - W_i\beta^1)'(\Sigma^1)^{-1}(r_i - W_i\beta^1)]}{\sum_{g=1}^{G} \pi_g |\Sigma^g|^{-1/2} \exp[-0.5(r_i - W_i\beta^g)'(\Sigma^g)^{-1}(r_i - W_i\beta^g)]}, \ldots\right. \tag{A.21}$$

[18] Of course, we no longer have to scale the data by the inverted gamma mixing variables λ.

Copyright © 2004 John Wiley & Sons, Ltd.

$$\left.\frac{\pi_G|\Sigma^G|^{-1/2}\exp[-0.5(r_i - W_i\beta^G)'(\Sigma^G)^{-1}(r_i - W_i\beta^G)]}{\sum_{g=1}^{G}\pi_g|\Sigma^g|^{-1/2}\exp[-0.5(r_i - W_i\beta^g)'(\Sigma^g)^{-1}(r_i - W_i\beta^g)]}\right) \tag{A.22}$$

with ‘Mult’ denoting the Multinomial distribution (e.g., Poirier, 1995, pp. 118–119).
(2) Complete conditional for π:

$$\pi|\Gamma_{-\pi}, \text{Data} \sim \text{Dir}(n_1 + \underline{\alpha}_1, n_2 + \underline{\alpha}_2, \cdots, n_G + \underline{\alpha}_G) \tag{A.23}$$

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

ACKNOWLEDGEMENTS

We are grateful to seminar participants at the University of Laval, the University of Michigan, the University of Missouri and the University of Wisconsin-Milwaukee. We are particularly grateful to the co-editor and two anonymous referees for valuable comments and suggestions on previous drafts of this paper. All errors are, of course, our own.

REFERENCES

Albert JH, Chib S. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**: 669–679.

Bjorklund A, Moffitt R. 1987. The estimation of wage gains and welfare gains in self-selection models. *Review of Economics and Statistics* **69**(1): 42–49.

Carlin B, Polson N. 1991. Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian Journal of Statistics* **19**: 399–405.

Chib S, Hamilton B. 2000. Bayesian analysis of cross-section and clustered data treatment models. *Journal of Econometrics* **97**: 25–50.

Chib S, Hamilton B. 2002. Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics* **110**: 67–89.

Dehejia R. 1999. Program evaluation as a decision problems. NBER working paper # 6954. Also *Journal of Econometrics*, Forthcoming.

Dehejia R, Wahba S. 1999. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**: 1053–1062.

Geweke J. 1993. Bayesian treatment of the independent Student *t* linear model. *Journal of Applied Econometrics* **8**: 19–40.

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.

Heckman J. 1990. Varieties of selection bias. *American Economic Review Papers and Proceedings* **90**(2): 313–318.

Heckman J, Honoré B. 1990. The empirical content of the Roy model. *Econometrica* **50**: 1121–1149.

Heckman J, Smith J. 1998. Evaluating the welfare state. In: *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*, Strom S (ed.). Econometric Society Monograph Series. Cambridge University Press: Cambridge; 241–318.

Heckman J, Vytlacil E. 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences* **96**: 4730–4734.

Heckman J, Vytlacil E. 2000. The relationship between treatment parameters within a latent variable framework. *Economics Letters* 33–39.

Copyright © 2004 John Wiley & Sons, Ltd.

Heckman J, Smith J, Clements N. 1997. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Review of Economic Studies* **64**: 487–535.

Heckman J, Ichimura H, Smith J, Todd P. 1998. Characterizing selection bias using experimental data. *Econometrica* **66**(5): 1017–1098.

Heckman J, Tobias J, Vytlacil E. 2002. Simple estimators for treatment parameters in a latent variable framework. *Review of Economics and Statistics* **85**(3): 748–754.

Imbens G, Angrist J. 1994. Identification and estimation of local average treatment effects. *Econometrica* **62**: 467–476.

Koop G, Poirier DJ. 1997. Learning about the across-regime correlation in switching regression models. *Journal of Econometrics* **78**: 217–227.

Manski C. 1990. Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings* **80**: 319–323.

Manski C. 1994. The selection problem. In: *Advances in Econometrics: Sixth World Congress* Sims C (ed.). Cambridge University Press: Cambridge; 143–170.

McLachlan G, Peel D. 2000. *Finite Mixture Models*. Wiley: New York.

Nobile A. 2000. Comment: Bayesian multinomial Probit models with a normalization constraint. *Journal of Econometrics* **99**(2): 335–345.

Poirier DJ. 1995. *Intermediate Statistics and Econometrics: A Comparative Approach*. MIT Press: Cambridge, MA.

Poirier DJ. 1998. Revising beliefs in non-identified models. *Econometric Theory* **14**: 483–509.

Poirier DJ, Tobias JL. 2003. On the predictive distribution of outcome gains in the presence of an unidentified parameter. *Journal of Business and Economic Statistics* **21**: 258–268.

Roy AD. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* **3**: 135–146.

Vijverberg WPM. 1993. Measuring the unidentified parameter of the extended Roy model of selectivity. *Journal of Econometrics* **57**: 69–89.

Vytlacil E. 2000. Semiparametric identification of the average treatment effect in nonseparable models. Unpublished manuscript, Stanford University.

Copyright © 2004 John Wiley & Sons, Ltd.