Evaluation and Identification of Semiparametric Maximum Likelihood Models of Dynamic Discrete Choice

Stephen V. Cameron [1]
Department of Economics,
Columbia University

Christopher R. Taber
Department of Economics and
Center of Urban Affairs and Policy Research,
Northwestern University

Revised: November, 1998 (Original: November 1994)

[1] This project arose from joint work on this estimator with Jim Heckman. We thank him for suggesting it and for many useful comments. We thank Joesph Altonji and Hide Ichimura and seminar participants at Northwestern University and the University of Chicago for useful comments. This research has been supported by grants from the National Science Foundation and the Bradley Foundation to James J. Heckman.

Abstract

This paper investigates a nonparametric maximum likelihood estimator of dynamic discrete-choice models, which is easily extended to a wide variety of economic and social science contexts. Specifically, we explore longitudinal data models of binary choice in the presence of unobserved heterogeneity of unknown distribution. Several estimation schemes have been developed which account for the presence of unobserved effects. However, previous estimators are either based on strict parametric assumptions about the distributional form of the heterogeneity or do not generalize beyond narrow specifications and cannot, for example, be used to distinguish between the effects of heterogeneity versus state-dependence. Furthermore, as opposed to many previous estimators, because the distribution of the heterogeneity is estimated within the model, policy counterfactual simulations may also be investigated. The estimator investigated in this paper makes no parametric assumptions about the form of the heterogeneity. The distribution of the heterogeneity is estimated nonparametrically, using an approach similar to the one proposed for continuous-time models by Heckman and Singer (1984). This estimator has the additional advantage that it can be easily applied in practice with a minimal computational burden.

Two chief drawbacks to applying these estimators in practice have been the lack of asymptotic distribution theory and the scarcity of knowledge about their small-sample properties. This paper address both these topics. First, we present identification theorems that are needed to not only establish asymptotic consistency but also illuminate contexts in which the estimator is likely to perform well. Second, we investigate the small-sample performance of the estimator under a variety of assumptions using Monte-Carlo methods. Along the way, we also investigate several rules for choosing the proper nonparametric approximation to the unknown heterogeneity distribution.

Our paper makes the following contributions. First, our theorems establish identification results for a broader set of distributional assumptions than those already in the literature. Second, while there is not yet a general asymptotic distribution theory, our Monte-Carlo evidence suggests that assuming the estimated slope parameters converge to normality at rate root-n provides a close approximation in practice to the true behavior of the estimator. Third, our Monte-Carlo results suggest that not only are the parameters of interest recovered accurately but so are their standard errors, and hence, inferences regarding those parameters are likely to be valid. This finding contradicts previous research, which is pessimistic about the ability of these estimators to produce accurate standard errors of the estimated parameters. Fourth, we compare how well this estimator performs against alternative parametric maximum-likelihood estimators that assume the true heterogeneity distribution to be known. We find little difference between the performance of the semiparametric estimator and the parametric estimator. However, when the true heterogeneity is not assumed known, the semiparametric estimator outperforms the parametric alternatives hands down.

1 Introduction

Dynamic discrete choice problems in economics and other social sciences appear in many contexts. In this paper, we investigate the properties of a semiparametric maximum likelihood estimator of dynamic choice models in a discrete time-period framework. In particular, we assume the error term consists of an unobserved random effect of unknown distribution and a period-specific error term drawn from a known distribution such as normal or logistic. For each individual the random effect is common across periods, but may have an intertemporal correlation pattern that depends on time. Omitted variables or, more generally, unobserved heterogeneity gives rise to the intertemporal correlation of the error terms influencing individual behavior. A similar approach to the one we use to correct for "heterogeneity bias" has been analyzed for continuous-time models by Heckman and Singer (1984). In the discrete-time context Follmann (1985) and Follmann and Lambert (1989) have used Monte Carlo methods to examine a subset of the models we investigate. Despite Follmann and Lambert's finding that these models estimate the unknown parameter well but estimate its variance poorly, we find a variety of contexts in which both the parameter and its variance are estimated with surprising accuracy. Furthermore we demonstrate that the parameters are approximately normally distributed and that we lose very little precision using our approach as compared to full maximum likelihood.

These types of models are of great value in analyzing discrete outcomes in longitudinal data. For example, an analyst with access to monthly or quarterly spell data on employment behavior may be interested in the dynamics of the labor force participation or job-changing decision. Controlling for unobserved heterogeneity when producing estimates of the behavioral parameters is a key element in the construction and interpretation of unbiased counterfactual policy simulations. If an omitted variable, such as motivation or past work history, affects the participation decision, then failure to control for that effect will bias estimated parameters and invalidate policy forecasts. Econometricians refer to this problem as unobserved heterogeneity bias.

We address two separate issues in this paper: the first is the identification and asymptotic behavior of the estimator, and the second is the small sample bias and precision with which unknown parameters are recovered. Given that identification is satisfied, consistency for this class of estimators follows from the work of Heckman and Singer (1984) and Follmann (1985), who applied the Heckman and Singer approach to models of dynamic discrete choice in which the period-specific component of the error term is logistic. Using Monte Carlo methods, we will examine how well the estimator and its standard error recover the true parameters. Furthermore, since the asymptotic distribution and rate of convergence of the estimator is unknown, we will present Monte Carlo evidence that suggests the slope parameter is distributed asymptotically normal and converges at rate $\sqrt{n}$ or a rate close to $\sqrt{n}$. Even more, we present evidence that normality of the estimates is not a bad assumption in small samples.

We examine the behavior of this type of estimator in three different data environments that commonly arise in social science and other disciplines. First, we consider a binary time series on longitudinal data–for example, an analyst may have data available on monthly or annual labor force participation rates from a panel of individuals. The bulk of the Monte Carlo results in this paper is devoted to this type of data. The second model we will consider is a binary time series with lagged dependent variables. The third is a birth-death model that has been used to analyze birth transitions, time to strike settlement (Melino 1989), or schooling transitions (Cameron and Heckman 1992 and 1993, Mare 1981, Bartholemew, 1973). We will compare outcomes for each model when the vector of covariates is constant across time periods and when the vector of covariates is allowed to vary across time (exclusion restrictions).

Our paper proceeds as follows. Section 1 discusses the basic model and its optimization. Section 2 presents a discussion of identification and asymptotic consistency. Section 3 presents detailed Monte Carlo results for the dynamic discrete choice models with time-varying explanatory variables and time-constant exogenous variables. We focus on how well the nonparametric estimator recovers the parameters of the underlying model and their standard errors in each of the three data environments mentioned above. We examine how well our model does with both discrete and continuous omitted variables.

2 The Model

We assume that for each individual in a sample we observe a binary response vector d_i of length T, where T may be 1. (This model is extensively described in Heckman, 1981). Let the elements of d_i be denoted by d_{it} and take on values 0 or 1. In general, we may have more than two states in each period, but we focus on the binary case in the following. We also observe a matrix of covariates X_i of dimension T*K where K is the number of observable independent variables. Let the vector of exogenous variables observed in period t (the t^{th} column of X_i) be denoted by x_{it}. Let f_i represent an individual-specific omitted variable not observable to the econometrician but observed by the individual. If T>1, we allow a factor loading term α_t on the f_i, and normalize α_1 to 1 for identification. (A standard random effects model is a special case in which α_t=1 for all t.) Let α signify the vector of α_t and β the vector of β_t. The econometrician observes d_{it} equal to one if the individual's underlying utility in time-period t is positive:

$$d_{it} = 1(x'_{it}\beta_t + \varepsilon_{it} + \alpha_t f_i \geq 0) \tag{2.1}$$

where $1(\cdot)$ is the indicator function which takes the value 1 if its argument is true and the value 0 otherwise. (We will drop the individual subscript i for the remainder when it is clear enough to do so.) Assuming ε_t is independent across time with a logistic distribution gives us a simple logit with a random effect of unknown distribution, denoted by $H(f)$. We make no assumption about the distribution of f except $\mathrm{E}(f^2) < \infty$ and f is distributed independently of x_t and ε_t for all t. We will use nonparametric maximum likelihood estimation (NPMLE) to estimate β_t. An individual's contribution to the likelihood function is thus

$$\int \left\{ \prod_{t=1}^{T} F(x'_t\beta_t + \alpha_t f)^{d_t}(1 - F(x'_t\beta_t + \alpha_t f))^{1-d_t} \right\} dH(f), \tag{2.2}$$

where F is the logistic distribution function. We obtain consistent estimates of $(\beta_1, \beta_2, \ldots, \beta_T, \alpha_2, \ldots, \alpha_t, H$ by maximizing the likelihood with respect to these arguments.

In practice we approximate $H(f)$ nonparametrically with a mixing distribution defined on a finite but unknown number of support points. Let $G(f_c)$ denote the mixing distribution and C the number of support points, where c subscripts the points of support (atoms) of the mixing distribution from 1 to C. Thus (2.2) is approximated by

$$(2.3) \qquad \sum_{c=1}^{C} \left\{ \prod_{t=1}^{T} F(x_t'\beta_t + \alpha_t f_c)^{d_t} (1 - F(x_t'\beta_t + \alpha_t f_c))^{1-d_t} \right\} g_c$$

where f_c are the support points of $G(f_c)$, and g_c are the probability masses associated with each point (the g_c are restricted to be nonnegative and sum to one in order to guarantee that G is a proper distribution function). Along with β_t, we estimate parameters of the mixing distribution: α_t, f_c, g_c, and C.

Discreteness imposes no restriction on the maximum likelihood procedure for producing accurate estimates of β and its standard error. Laird (1978) provides conditions under which the NPMLE estimator of H takes the form (2.3). To see the intuition behind this result, suppose that for each individual we observe f_i and we want to estimate H. For any finite amount of data of dimension n, regardless of the true H, the empirical distribution of f_i is discrete with at most n support points, so the empirical distribution of the heterogeneity could be fit perfectly with $C \leq n$. The NPMLE estimator of H in this case is simply the empirical distribution function of f_i.

The model presented above can be easily extended in a number of ways. We will discuss the relaxation of the logistic assumption later in the paper. Furthermore, endogenous continuous variables can be integrated into these models, but we will not discuss them here (see Cameron and Heckman, 1991 for a framework more general than that used here).

Mixing distributions have a long history in statistics, going at least to the work of Pearson (1894). Recent contributions have been made by Lindsay (1983a, 1983b, and 1989) and others. Heckman and Singer (1984) analyze the behavior of the NPMLE estimator for continuous time duration models. Follman (1985) and Follman and Lambert (1989) examine the NPMLE estimator of β for logistic regression models. Their work is closest to that presented here. A modern introduction can be found in Everitt and Hand (1981) or Titterington, Smith, and Makov (1985). For our purposes, these models are relatively simple to estimate and nest standard logit, probit, and linear probability models as special cases.

When it is identified, proving consistency for this model is a straight forward extension of work by Heckman and Singer (1994) and Coslett (1983) who verify the assumptions of Kiefer and Wolfowitz (1956). Consistency for a special case of this model is shown by Follman (1986) and for a more general case by Cameron, Heckman, and Taber (1994). Some identification issues (e.g identification of the factor loading terms) are nontrivial and are considered in detail in the next section. While results of the consistency of this estimator have existed for quite some time, very little is known about its rate of convergence and asymptotic distribution. This remains an active topic in the statistics literature. However, we provide Monte Carlo evidence that $\sqrt{n}$ asymptotic normality is not a bad assumption in finite samples.

3 Identification

In the previous section we presented a general model for longitudinal binary choice data. In its most general form the model is not identified. In this section we provide conditions under which it is identified and supply counter examples of models that are not identified.

We take Cameron and Heckman (1993) as our point of departure. They explore identification of a longitudinal binary choice model with only partial observability. Define[1]

$$d_t = 1(X_t'\tilde{\beta}_t - u_t \geq 0). \tag{3.1}$$

In their discrete duration model, for $t > 1$ the Econometrician observes (d_t, X_t) if and only if $d_{t-1} = 1$. They establish conditions which deliver identification of $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_T)$ and the joint distribution of $u = (u_1, \ldots, u_T)$ up to scale and location. Since they show identification using a subset of the information available to us, we take their results as given and assume that β and the distribution of u are identified up to scale. We refer interested readers to their results. The question still remains whether identification of the joint distribution u is sufficient for identification of our factor structure (2.1). This is the question we address here.

The problem turns out to be the non-identification of the scale in a way that we make precise below. In the binary choice model we can just normalize the scale, however with this specification no normalization delivers identification without imposing restrictions on the underlying model. We fix the scale and location of (3.1) by assuming that

[1] Comparing this specification with (2.1), β and $\tilde{\beta}$ will be proportional, but not necessarily equal.

for $t = \{1, 2, \dots, T\}$, $\|\tilde{\beta}_t\| = 1$ and that X_t does not contain an intercept so no location normalization is required on the unobservables. Under these normalizations we appeal to Cameron and Heckman (1993) and make the following assumption.

Assumption 1 *The joint distribution of the vector $(u_1, u_2, \dots, u_T)$ is identified.*

We define our factor structure with the following assumption,

Assumption 2

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix} = \begin{bmatrix} \gamma_1(\varepsilon_1 + f) \\ \gamma_2(\varepsilon_2 + \alpha_2 f) \\ \vdots \\ \gamma_T(\varepsilon_T + \alpha_T f) \end{bmatrix} \tag{3.2}$$

where for each $t = \{1, 2, \dots, T\}, \varepsilon_t$ is independent of f and independent of ε_τ for $\tau \neq t$. The non-degenerate random variable ε_t has a known distribution function F_t.

The scalars γ_t enter equation (3.2) to represent the non-identification of the scale. Since we used the normalization $\|\tilde{\beta}_t\| = 1$, it is easy to show that

$$\gamma_t = \frac{1}{\|\beta_t\|}$$

where β_t is defined by (3.2). We will discuss the cases under which these assumptions are sufficient for identification of the scalar terms γ_t and α_t, and the distribution of f.

Define $\gamma \equiv (\gamma_1, \gamma_2, \dots, \gamma_T)$ and $\alpha \equiv (\alpha_2, \dots, \alpha_T)$. Let $\Gamma \subset (\Re^+)^T$, $\mathcal{A} \subset \Re^{\mathcal{T}-\infty}$, and define $\mathcal{H}$ to be the class of distribution functions with median zero. We want to show identification of (γ, α, H) in the class $\Omega \equiv (\Gamma, \mathcal{A}, \mathcal{H})$.

Under Assumptions 1 and 2, we say (γ, α, H) is identified in Ω when for any $(\gamma^1, \alpha^1, H^2) \in \Omega$ and $(\gamma^2, \alpha^2, H^2) \in \Omega$,

$$\int \prod_{t=1}^{T} F_t\left(\frac{y_t}{\gamma_t^1} - \alpha_t^1 f\right) dH^1(f) = \int \prod_{t=1}^{T} F_t\left(\frac{y_t}{\gamma_t^2} - \alpha_t^2 f\right) dH^2(f) \qquad \forall\, y \in \Re^T, \tag{3.3}$$

if and only if $(\gamma^1, \alpha^1, H^1) = (\gamma^2, \alpha^2, H^2)$.

We begin by considering the one period version of the model

$$d_1 = 1(X_1'\beta_1 - \varepsilon_1 - f \geq 0). \tag{3.4}$$

This is a simple binary choice model where the distribution of ε_1 is as assumed known, but the distribution of f is unrestricted.Under Assumption 2,

$$\Pr(d_1 = 1) = \int F_1(X_1'\beta_1 - f)dH(f).$$

We know that β_1 and the distribution of $(\varepsilon_1 + f)$ are identified up to scale and location (Coslett 1983). Complete knowledge of the distribution of ε_1 presumes a knowledge of its scale, this factor structure imposes the scale normalization on the binary choice model through the specification of the distribution of ε_1. However, this scale normalization is not sufficient for identification of the model. We illustrate with the following two simple counter examples .

Example 1: For any $\sigma > 0$ suppose the random variable (σf) has distribution F_1. Then

$$\begin{aligned}\Pr(X_1'\beta_1 - f - \varepsilon_1 \geq 0) &= \Pr(X_1'(\sigma\beta_1) - \sigma f - \sigma\varepsilon_1 \geq 0)\\ &= \int F_1(X_1'(\sigma\beta_1) - \sigma\varepsilon_1)dF_1(\varepsilon_1).\end{aligned}$$

So this model is indistinguishable from model (3.4). Normalizing the scale of ε_1 is not sufficient for identification of the scale of β_1.

■

The point of this example is simply that if there is some scalar $\sigma \neq 1$ such that (σf) has the same distribution as ε_1, then we can never tell which error term corresponds with ε_1. The other examples uses properties of Gaussian distributions.

Example 2: Let both ε_1 and f have a standard normal distribution. For any $\sigma > \frac{1}{2}$ we can not distinguish model (3.4) from

$$d^* = 1(X_1'(\sigma\beta) + \varepsilon^* + f^* \geq 0)$$

where ε^* has standard normal distribution and f^* is distributed normally with mean zero and variance $2\sigma - 1$.

■

These counter examples are both very special, but the basic principle behind the lack of identification is actually quite general. Before proceeding we first establish some necessary notation. Let ϕ_t and ϕ_H denote the characteristic functions associated with the distribution of ε_t and the distribution function H. We make one more additional assumption,

Assumption 3 *The characteristic functions $\phi_1, \phi_2, \ldots, \phi_t$ do not vanish.*

Since the distribution of these functions is pre-specified, this condition is easy to check and will be satisfied for both normal and logistic characteristic functions.

We claimed above that the non-identification of the one period model results from the non-identification of the scale. The next proposition makes this precise by showing that when the scale can be identified, the model is identified.

Proposition 1 *When $T = 1$ and γ_1 is identified, under Assumptions 1-3 we can identify the distribution of f.*
(Proof in Appendix)

We can summarize the results of this proposition and the two counter examples by providing a necessary and sufficient condition for identification in the one period model. While the condition may not be very intuitive, it illustrates the generality of the non-identification result .

Proposition 2 *When T=1, for any $(\gamma_1, H) \in \Omega$ and any $\gamma_1^* \in \Gamma$, under Assumptions 1-3 there exists a distribution function H^* such that (γ_1, H) is not identified relative to (γ_1^*, H^*) if and only if*

$$\phi^*(t) \equiv \frac{\phi_1(\gamma_1 t)\phi_H(\gamma_1 t)}{\phi_1(\gamma_1^* t)}$$

is a characteristic function.
(Proof in Appendix)

Another avenue to achieve identification is to consider restrictions on the class of permissible distribution functions. One possibility is to bound the support of the heterogeneity . Define $\tilde{\mathcal{H}}$ to be the class of distribution functions with median 0 and bounded support. So for any $H \in \tilde{\mathcal{H}}$, there exists $-\infty < \underline{f} < \bar{f} < \infty$ such that $H(\underline{f}) = 0$ and $H(\bar{f}) = 1$.

Proposition 3 *If ε_1 has support $\Re$, $X_1'\beta_1$ has support $\Re$, and $H \in \tilde{\mathcal{H}}$. Then the scale of β is identified from the first period.*
(Proof in Appendix)

Identification of the scale is essentially obtained in Proposition 3 by restricting the tails of the distribution of f to be thinner than the tails for the distribution of ε_1. A similar strategy is used by Heckman and Singer (1984) for continuous time duration models and is extended to binary choice models by McCall (1992) .Using our notation, they restrict the tails of the distribution of f by assuming that $E(e^f) < \infty$. Ishwaran (1994) generalizes this strategy by restricting the tails of the heterogeneity to be thinner than the tails of the known distribution. As an indicator of the size of the tails defines the radius of continuity

$$r_0(H) \equiv \sup_{r \geq 0} \left(r : \int e^{r|f|} dH(f) < \infty \right).$$

He shows that the scale is identified in model (3.4) when $r_0(H)$ is restricted to be strictly greater than $r_0(F_1)$[2].

We have shown that in general the model is not identified with only one period. We now consider the two period model in which we obtain much more identifying information. Not only do we observe the marginal distributions of $\varepsilon_1 + f$ and $\varepsilon_2 + \alpha_2 f$, but also their joint distributions. Since ε_1 is independent of ε_2, the dependence of $\varepsilon_1 + f$ and $\varepsilon_2 + \alpha_2 f$ will only operate through the distribution of f. However, this additional information is still not sufficient to identify the full model. As an example we present the case where the random variables are all normally distributed and show that even though we know that all the distributions are Gaussian, we can not identify all of the parameters.

Example 3: Let

$$\varepsilon_1 \sim \mathrm{N}(0,1)$$
$$\varepsilon_2 \sim \mathrm{N}(0,1)$$
$$f \sim \mathrm{N}(0,\sigma_f^2)$$

where we take $\sim \mathrm{N}(0,\sigma^2)$ to mean distributed normally with mean 0 and variance σ^2. The question is whether knowledge of the joint distribution of (u_1, u_2) will suffice for identification of the scale terms (γ_1, γ_2), the factor loading term α_2, and the variance of the

[2] This result does not strictly generalize Proposition 3. For instance if F_1 is Normal then $r_0(F_1) = \infty$.

heterogeneity σ_f^2. Since all of the error terms are Gaussian, from the joint distribution of (u_1, u_2) all that we can hope to identify is

$$\begin{aligned} \mathrm{Var}(u_1) &= \gamma_1^2(\sigma_f^2 + 1) \\ \mathrm{Var}(u_2) &= \gamma_2^2(\alpha_2^2\sigma_f^2 + 1) \\ \mathrm{Cov}(u_1, u_2) &= \gamma_1\gamma_2\alpha_2\sigma_f^2 \end{aligned}$$

This yields three equations in four unknowns, so we can not identify all four parameters. ∎

Though this example crucially depends on the Gaussian structure, it demonstrates that the general model is not identified with two periods. Even though normality may be very special, we do not want to rule it out since most of the previous work in this area has relied on the normality assumption.

Notice that in the example if we normalize $\alpha_2 = 1$ we are left with three equations in three unknowns and we can identify the full model. Invoking this assumption restricts the specification to the standard random effect model. We will show below that this one normalization will provide identification in the general two period model. There is a sense in which showing identification is most difficult when all of the error terms are normal. This can be seen in the following lemmas which we will use to show identification. The first two we take directly from Kagan, Linnik, and Rao (1973) , and the third is simple fact about polynomials.

Lemma 1 (Kagan, Linnik, Rao (1973) pp. 29-31) *Consider the equation, assumed valid for $|u| < \delta_0, |v| < \delta_0$,*

$$\Psi_1(u + b_1 v) + \ldots + \Psi_r(u + b_r v) = A(u) + B(v) + P_k(u, v)$$

where P_k is a polynomial of degree k; Ψ_j,A, and B are complex valued functions of two real variables u and v. We assume that the numbers b_j are all distinct without loss of generality and that the functions A, B, and the Ψ_j are continuous. Then, in some neighborhood of the origin, the functions A, B, and the Ψ_j are all polynomials of degree $\leq$ max (r, k).

This lemma was proved originally by both Linnik (1964) and by Rao (1966) and is also proved in Kagan, Linnik, Rao (1973). We will use this it to extend identification of the

Gaussian model to more general models. The importance of normality becomes clear from the following lemma.

Lemma 2 (Kagan, Linnik, Rao (1973) pp. 82-83) *Let the characteristic function ϕ of some random variable have the form*

$$\phi(t) = \exp(Q(t))$$

in some neighborhood $|t| < \delta$ of the origin where Q is a polynomial. Then

$$Q(t) = At^2 + iCt \tag{3.5}$$

where $A \leq 0$ and C are real constants, and relation (3.5) holds for all real t.

Recall that the characteristic function of a normal random variable with mean μ and variance σ^2 is $\exp(it\mu - \frac{(\sigma t)^2}{2})$. Thus if we can show that the log of a characteristic function of a nondegenerate random variable is a polynomial, then that random variable must be normal. We will also use the following fact.

Lemma 3 *If for some continuous complex function Ψ, for some $\gamma_1 \neq \gamma_2$, and for some $\delta > 0$, $A(t) = \Psi(\gamma_1 t) - \Psi(\gamma_2 t)$ is a polynomial of degree k when $|t| < \delta$, then Ψ must be a polynomial of degree k in some neighborhood of zero.*
(Proof in Appendix)

In the proof of the following proposition we will show that if equation (3.3) holds with $H \neq H^*$ then H and H^* must be normal. However, as can be seen from Example 3, if the error terms are all normal than the model is identified.

Proposition 4 *In the model above with $T = 2$, under Assumptions 1-3 and $\alpha_2 = 1$, (γ, H) is identified in Ω.*
(Proof in Appendix)

With one more period we can show identification of the vector α as well. Consider the normal case.

Example 4: Let

$$\begin{aligned}\varepsilon_1 &\sim \mathrm{N}(0,1)\\ \varepsilon_2 &\sim \mathrm{N}(0,1)\\ \varepsilon_3 &\sim \mathrm{N}(0,1)\\ f &\sim \mathrm{N}(0,\sigma_f^2)\end{aligned}$$

From the joint distribution of (u_1, u_2, u_3) we can identify

$$\begin{aligned}\mathrm{Var}(u_1) &= \gamma_1^2(\sigma_f^2+1)\\ \mathrm{Var}(u_2) &= \gamma_2^2(\alpha_2^2\sigma_f^2+1)\\ \mathrm{Var}(u_3) &= \gamma_3^2(\alpha_3^2\sigma_f^2+1)\\ \mathrm{Cov}(u_1,u_2) &= \gamma_1\gamma_2\alpha_2\sigma_f^2\\ \mathrm{Cov}(u_1,u_3) &= \gamma_1\gamma_3\alpha_3\sigma_f^2\\ \mathrm{Cov}(u_2,u_3) &= \gamma_1\gamma_2\alpha_2\alpha_3\sigma_f^2\end{aligned}$$

This yields six equations in six unknowns and it is easy to show all six parameters are identified.

∎

Using the three Lemmas we can show that since the Gaussian version of the model is identified, the general model is identified. The form of the proof is almost identical to the previous one.

Proposition 5 *In the model above with $T \geq 3$, under Assumptions 1-3 , (γ, α, H) is identified in Ω.*
(Proof in Appendix)

4 Optimization

Let $(\hat{\beta}, \hat{G}, \hat{C})$ denote a maximum likelihood solution to the product over the sample of individual likelihoods (2.3), where $\hat{G}$ represents the parameters of the mixing distribution in (2.3): α_t, g_c, and f_c. Even conditioned on C, in general the solution will not be unique and will depend on the initial values of the parameters; moreover, poor starting values may cause numerical problems for algorithms like quasi-Newton. One widespread technique, used by Follmann and Lambert (1989) for example, starts searching the parameter space with the EM algorithm (Dempster, Laird, and Rubin 1977). But since the EM algorithm converges slowly (linear convergence), once successive iterates become "close" the algorithm switches to a quasi-Newton routine that converges more quickly (quadratic convergence). However, since we had little problem in finding reasonable starting values, and since the computation needed by the EM algorithm was excessive, we used a quasi-Newton routine for all the results we present in Section 3. We found virtually no difference between results using EM or quasi-Newton.

Maximization precedes by first maximizing the log-likelihood when $C = 1$. In this case the likelihood for each individual simplifies to a product of independent logits in each time period. The MLE estimator $\hat{\beta}$ is unique in this case. Next, C is incremented to two and $\hat{\beta}$ is used as the starting value for β, and the log-likelihood is maximized over β again and the parameters of G_2–a 2-point mixing distribution. (Our method for finding starting values for G_2 is explained below.) Iteration continues in this way until the log-likelihood no longer increases.

We use a standard method of finding starting values for the heterogeneity distribution whenever C is incremented to $C+1$ (see Simar, 1976). Let $L[\hat{G}_C, \hat{\beta}_C]$ denote the log-likelihood with C support points, and let $\hat{G}_C$ and $\hat{\beta}_C$ be a set of maximum likelihood estimates. The effect on the log-likelihood of taking ε mass from the distribution $\hat{G}_C$ and placing it on a new point δ_θ is described by the directional derivative

$$D((1-\varepsilon)\hat{G}_C + \varepsilon\delta_\theta, \hat{G}_C) = \lim_{\varepsilon \downarrow 0} \frac{L[(1-\varepsilon)\hat{G}_C + \varepsilon\delta_\theta; \hat{\beta}_C] - L[\hat{G}_C; \hat{\beta}_C]}{\varepsilon}. \tag{4.1}$$

The δ_θ that maximizes $D(\hat{G}_C(1-\varepsilon)\hat{G}_C + \varepsilon\delta_\theta)$ is the best initial guess for a $(C+1)$st point. For mixture problems with no covariates or with β fixed, Lindsey (1983a, Theorem 4.1) showed that finding a $\hat{G}$ such that

$$\max_{\delta_\theta} D(\hat{G}_C, (1-\varepsilon)\hat{G}_C + \varepsilon\delta_\theta) = 0$$

is equivalent to finding a $\hat{G}$ that maximizes the log-likelihood.

Lesperance and Kalblfleish (1992) recently proposed a twist on the Lindsey-Simar algorithm that promises to greatly reduce the computational burden of maximum likelihood estimation of mixture models in a variety of situations. Based on a result of Lindsey (1983), they propose adding a new point of support at every local maximum of the directional derivative, rather than only one point at the global maximum as is done in the Lindsey-Simar algorithm. Lesperance and Kalbfleisch present promising Monte Carlo results, but do not investigate the model in which we are interested, nor do we investigate their technique in this paper. Future investigation in the context of our model may be fruitful.

5 Monte Carlo Results

5.1 Introduction

We present a series of Monte Carlo results for the three types of data generating environments previously discussed. In Section B, we present a set of baseline results. We assume the time-independent component of the error term is drawn from a logistic distribution and generate heterogeneity from a continuous distribution (normal) and alternatively from a discrete distribution (binomial). We also show Monte Carlo results for each environment when both the observed exogenous variables vary with time (exclusion restrictions) and when the observed exogenous variables are fixed across time periods. We also compare how the estimated standard errors of the slope parameter matches up with its true coefficients.

Since the asymptotic distribution of the estimated slope parameter is unknown, we are interested in examining not only how well the estimated slope parameter recovers the truth, but also how the empirical distribution of the estimated slope parameter compares to a normal distribution. We do this in Section C. In Section D, we contrast our baseline results with results obtained when the number of support points in the heterogeneity distribution are estimated using three different stopping rules: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and full Maximum Likelihood estimation (MLE), where convergence is determined when we can no longer numerically increase the log of the likelihood by adding more points to the estimated heterogeneity distribution. We discuss the advantages and disadvantages of each rule. We also examine performance when we alter the assumptions of our baseline model by varying the number of observations in each Monte Carlo draw, decreasing the variance of the exogenous covariates varies, and altering the algorithm that generates starting values each time the a new point is added to the support distribution. The findings of the above section support the $\sqrt{n}$ asymptotic normality of $\hat{\beta}$. We strongly reject the hypothesis that any of the estimators of the parameters of the mixing distribution, $\hat{G}$, possess the same property. Furthermore, we find that the NPMLE estimator not only accurately recovers β but that its standard error is recovered as well.

5.2 Monte-Carlo results for a 10 Period Panel of Binary Indicators

Table 1 displays simple Logit estimates of the slope parameter when no heterogeneity controls are applied. One-hundred Monte Carlo draws were estimated; each draw contains 500 observations, each observation observed for 10 periods. The heterogeneity distribution is assumed normal with mean -1.0 and variance 1.0 (panel A), and binomial with equal mass at -1.0 and 1.0 (panel B). Full details of the model can be found at the base of the Table. Each panel displays the results when no heterogeneity correction is applied, results obtained by applying the nonparametric heterogeneity correction, and disaggregated results for each estimated value of C when applying the heterogeneity correction. Reading across panel A, "Support Points" is the value of C; "Number of Runs" is the number of draws (out of 100) that converged at the given number of support points; "Mean" is the mean of the Monte Carlo distribution of the estimated slope parameter and "Std Dev" is the standard deviation that distribution; "Median" is the median of the estimated slope parameter; "Mean of Std Error" is the mean of the Monte Carlo distribution of the standard error calculated from the estimated information matrix in each run; and finally the associated "Std Dev" is the standard deviation of Monte Carlo distribution of the standard error. The bias in the estimated parameter is substantial: the mean of the estimates is 0.6248 for normal heterogeneity (see the row "No Heterogeneity Correction") and 0.8138 for the binomial case, compared with the true at 1.0. The standard deviation is less than .04 in both cases. The next row ("All Runs") in each panel presents comparable estimates using the same generated data and applying the nonparametric heterogeneity correction. These results evidence the accuracy with which we can recover the true parameter value and the small deviation of the estimated distribution (about .05 in each case). This accuracy is surprising given the small number of observations on which we estimate β. It is also important to notice the small number of heterogeneity support points needed for these estimates even when the underlying heterogeneity is drawn from a continuous distribution (normal heterogeneity). One draw converged with 3 points of support, 92 with 4 points, and 7 with 5 points. For the binomial case, 91 runs converged at the true number of points, and 9 runs finished with 3 points of support. β is accurately recovered at each estimated point of support. Even when the binomial heterogeneity distribution is estimated with three points of support, the mean estimate is within 0.26% of the true value (bottom row of panel B). Even though we may not do well in recovering the true distribution, we will see that in both cases this estimator does a good job in approximating the true mean and variance of the underlying heterogeneity distribution.

Table 2 shows the means of the estimated points of support and their associated masses. For the normal case he support points are fairly symmetric around the mean -1.0. This is true for the binomial case when C stops at 2, but not for the other cases. We will further discuss these results later in the paper.

Table 3 compares several standard error computations to the mean estimated standard error for both normal and binomial heterogeneity. We find that the nonparametric estimator does extremely well in estimating the true value. The first row displays the heterogeneity-corrected results from Table 1 for comparison; the second row shows results for the estimator when the true distribution is known up to its mean and variance. In other words, for the model with normal heterogeneity, we assumed the heterogeneity in the model to be normally distributed with unknown mean and variance, and for binomial heterogeneity we assumed the true number of points of support to be 2. The mean estimate and the mean standard error of the estimate are virtually identical in both cases indicating that little if anything is lost in this case from not knowing the true form of the heterogeneity. The third and forth rows compare the mean standard error to the true standard error. The third row is the mean over the 100 draws when the information matrix is constructed with the same data (both x and d) used in the estimation and evaluated at the true parameter values of the true model. For the fourth row computations, we use only the draws on the exogenous variables from above and then integrate the information matrix over the heterogeneity and the logistic error. This computation represents the value to which the mean of the information matrices would converge conditional on the values of the exogenous covariate and shows how close the row 3 value is to its asymptotic value. The results in rows 3 and 4 are almost identical. By comparing the mean standard error from row 1 with the corresponding value in row 3, we find that the difference between the standard error of the nonparametric estimator and the standard error from the true model is negligible. By contrast, Table 1 indicates that the estimated standard error from the uncorrected logit is biased downward about 20% for the normal and 10% for the binomial case. The nonparametric estimator produces standard errors that are exceptionally close to their true values and can be used confidently in hypothesis testing on the estimated parameters.

Table 4 displays results using three alternative distributions of the heterogeneity factor: exponential, mixture of normals, and normal. In addition, the model is estimated when no heterogeneity correction is applied, when the heterogeneity term is assumed normal, and when the nonparametric correction is applied. In all cases, the nonparametric correction recovers the slope parameter accurately. The normal correction works well when the true heterogeneity is drawn from a normal distribution and, surprisingly, when it is drawn from an exponential distribution. It does not work well when the true heterogeneity distribution is a mixture of normals.

Since these models are nonlinear, focusing solely on how estimated parameter values compare to the true value may obscure the true test of the model, which is how well it predicts. Table 5 fills this gap. Using the estimates summarized in Table 4, Table 5 shows results from a simple predictive test. Panel A displays our findings when the true heterogeneity distribution is a mixture of normal evaluated at three different points in the covariate distribution (x_{it} equals 0.0 , 0.675, and -0.675) [3] Column two shows how the predicted probability of $d_t = 1$ compares to the true, column three shows the probability of the period one and two indicators both being one, and column four compares the probability of the period one to five indicators all being one. A '*' indicates that the model fails the prediction test at the 95% level. Panels B and C show the same when the heterogeneity term is drawn from an exponential and from a normal.

[3] Recall that the true distribution of the x_{it} are standard normal.

A glance at the table reveals that the nonparametric model outperforms the other models hands down. A model that does not fit the data will not necessarily yield correct behavioral inferences or correct policy simulations. This finding is given added credence because we have conditioned the tests on different values of the x distribution. An analyst interested in doing policy simulations on how people from poor family backgrounds perform in the labor market, for example, would be interested in exactly this kind of conditioning.

We continue changing the assumptions of Table 1 and examining the findings. Table 6 replicates the results for Table 1 on the same model except the covariate x is fixed for each period (no exclusion restrictions). Panels A and B show the results when no heterogeneity correction is applied for both normal and binomial heterogeneity, and Panels C and D shows the results with the nonparametric correction. The results reported here are also encouraging. Analogous to Table 2, Table 6 aligns the findings of Table 3 with the the true standard errors and the outcomes when we estimate β knowing the distribution from which the heterogeneity is drawn. Comparing rows 1 and 2 of Panel A shows, as Table 2 did, that little is lost by estimating β with a nonparametric correction–the mean of each estimated β distribution is virtually identical. The Monte Carlo mean is slightly low for both the model with the nonparametric correction, however, but still within 4% of its true value. This compares with the x-varying case in which we found the mean estimate to be almost identical to the true value. Panel B also compares favorably, but the mean estimate is again dead on target. Not only do the estimates perform well, but the mean standard errors are also very close to the true values: all means are within 2% of the true value.

5.3 Asymptotic Normality

Table 7 shows the results for 1000 runs when the sample sizes are 500 and 250. In Table 8 we report the p-values of a Shapiro, Wilk Normality test on these empirical distributions (row 1): for the 500 observation runs we get .87 and .61 for the 250 observation runs, supporting the hypothesis that $\hat{\beta}$ is asymptotically normal. However, we reject normality of all of the estimated parameters of the mixing distribution at the .005 level. Figure 1 presents a standard quantile-quantile plot of the distribution of the $\hat{\beta}$ from the 500 observation samples: any deviation from the normality line is evidence against the hypothesis of normality, and as the figure shows, there is virtually none. Furthermore, if the NPMLE estimator of β

possesses $\sqrt{n}$ normality, we would expect that the ratio of the mean standard errors from the 500 and 250 observation samples to be approximately equal to $\sqrt{2}$. From Table 7, $\sqrt{2}$ times the standard error from panel B is .0684, which is very close to the panel A mean standard error of .0685. A final check of normality is how often we reject the hypothesis that $\hat{\beta}$ over $\hat{\sigma}$ rejects the true hypothesis of β equal to 1. These results are presented in Table 8 and also support the notion that $\hat{\beta}$ follows something close to a normal distribution.

5.4 Further Perturbations

In Table 9 we change the stopping rule for estimating the number of points of support, c. We try four different stopping rules: BIC, AIC (these numbers correspond to those found in Table 1), MLE assuming convergence when the change in the likelihood value was less than 0.01 (since we maximize log-likelihoods, this implies an approximately 1% change), and MLE with convergence set at a more stringent 0.0001 (an approximately 0.01% change).

Deciding when to stop adding points is somewhat of an inexact science. as c increases, the Hessian can become numerically ill-behaved or singular, particularly for models where c becomes relatively large. Since we can never decrease the likelihood by adding a point of support, forcing more points into the likelihood to increase its value will eventually result in a singular model. One advantage of conservative stopping rules such as BIC or AIC is that they are more successful at avoiding ill-behaved Hessians. Another obvious advantage is that they simply converge more quickly. For the MLE with 0.0001 convergence , 1 in 100 runs became singular for the normal case, and 3 in 100 for the binomial model. All other stopping rules avoided singularity though the strictest rule, BIC, was best at avoiding any kind of poorly behaved Hessian.

Panel A reports results for normal heterogeneity, and Panel B for binomial heterogeneity. The mean number of points of support for each rule is given next to the heading ($\bar{C}$), and below are the results for all runs and each point of support. Each stopping rule produces a mean estimate of the slope parameter within 0.5% of the true, and a mean estimate of the standard error within 1% of the true value. In general, as the mean number of points of support increases so does the estimate and its standard error. The mean number of support points in the normal case ranges from 3.7 for BIC to 4.76 for the last case. For the binomial case, BIC recovers exactly 2 points in every draw and MLE with 0.0001 convergence has a mean of 2.76. Nonetheless, the estimates are all excellent–the mean is within 0.5% of the true, and the mean standard error is within about 1% of the true value reported in Table 2.

Further perturbations on the model discussed thus far are laid out in Tables 10 and 11. Table 10 shows results for sample sizes ranging from 125 to 5000. In Table 11 we first try two different types of starting values; second we alter the variance of x; and finally we alter the probability of d=1. None of these perturbations alters our conclusions: we find exceptional performance in every case. As expected, increasing N only helps. Next, as discussed in Section 1, final estimates may be sensitive to starting values in these models so we reran the experiments changing the starting values in two ways. First, we dropped the Lindsay-Simar algorithm described in Section 1 for locating starting values when adding a point of support to the heterogeneity distribution. Rather, we made each point in the new distribution equiprobable and made each point equidistant without altering the variance of the estimated heterogeneity distribution (see "Diffuse Starting Values" in the table). [4] Next, when adding a point of support, we not only assumed equidistant points with equiprobable mass at each point but also set the values of β to zero (see "Zero Starting Values"). Full details are located at the base of the table. As no change is recorded for the normal heterogeneity case and only the smallest difference for the binomial case, these findings reinforce our confidence in the algorithm we use to locate the global maximum of the log-likelihood. Third, increasing the variance of x shrinks the standard error of the estimated β and increases c for the case with continuous heterogeneity; shrinking the variance does the opposite. Finally, as the probability of d=1 changes toward 1 or 0, the total amount of information in the data will fall and we expect and find larger standard errors and slightly less accurate estimates. Changing the probability from 0.43 in the basic model to 0.85 makes little difference, and the the Monte Carlo mean is still within about 1% of the true value with a standard deviation of about 0.05 in each case. In other results not reported, we altered the values of β and the variance of the heterogeneity. We found no appreciable degradation in the results. As expected, increasing the variance of the heterogeneity, decreased slightly the precision with which we estimate β.

[4] Keeping the variance the same simply locates the distance between each point.

6 Conclusions

Our findings demonstrate the exceptional small sample properties of the nonparametric maximum likelihood estimator for dynamic discrete choice problems. We find that both the slope parameters and their standard errors are recovered accurately and, hence, inferences regarding those parameters will be valid. When the true distribution of the heterogeneity is know, the semiparametric estimator performs as well as the parametric estimator that is based on the true distribution. When the true heterogeneity distribution is unknown, our model recovers the parameters of interest with exceptional accuracy, and outperforms misspecified parametric estimators hands down. Though there is not yet a general theory of asymptotic convergence, our evidence strongly supports the notion that the estimator converges to normality at a rate in the neighborhood of $\sqrt{n}$.

Appendix

Proof of Proposition 1: Let f have distribution H and f^* have distribution H^* where both H and H^* are elements of $\mathcal{H}$. Suppose that the distribution of $\varepsilon_1 + f$ is identical to the distribution of $\varepsilon_1 + f^*$. Then the characteristic function associated with these convolutions must be identical,

$$\phi_1(t)\phi_H(t) = \phi_1(t)\phi_{H^*}(t).$$

Dividing both sides by $\phi_1(t)$ we see that,

$$\phi_H(t) = \phi_{H^*}(t).$$

From the uniqueness theorem for characteristic functions, H is identified.

■

Proof of Proposition 2: Suppose there exists a distribution function H^* such that (γ_1, H) is not identified relative to (γ_1^*, H^*). This means that

$$\int F_1\left(\frac{X_1'\beta_1 + f}{\gamma_1}\right) dH(f) = \int F_1\left(\frac{X_1'\beta_1 + f}{\gamma_1^*}\right) dH^*(f)$$

which implies that

$$\phi_1(\gamma_1 t)\phi_H(\gamma_1 t) = \phi_1(\gamma_1^* t)\phi_H^*(\gamma_1^* t)$$

or

$$\phi_H^*(\gamma_1^* t) = \frac{\phi_1(\gamma_1 t)\phi_H(\gamma_1 t)}{\phi_1(\gamma_1^* t)}$$

Since H^* is distribution function $\phi^*(t) = \phi_H^*(\gamma_1 t)$ is a characteristic function.

Now suppose $\phi^*(t)$ is a characteristic function and let $\phi_H^*(\gamma_1^* t) = \phi^*(t)$. Then

$$\phi_1(\gamma_1 t)\phi_H(\gamma_1 t) = \phi_1(\gamma_1^* t)\phi_H^*(\gamma_1^* t)$$

and from the uniqueness theorem of characteristic functions, we can invert $\phi_H^*(\gamma_1 t)$ to obtain the distribution function H^*, so (γ_1, H) is not identified relative to (γ_1^*, H^*).

■

Proof of Proposition 3: Suppose the scale of β were not identified. Then there exists a $\gamma^* \neq 1$ and a distribution function $H^* \in \tilde{\mathcal{H}}$ such that

$$\int_{\underline{f}}^{\bar{f}} F_1(X_1'\beta_1 + f)dH(f) = \int_{\underline{f}^*}^{\bar{f}^*} F_1(X_1'(\gamma^*\beta_1) + f)dH^*(f)$$

for all $X_1'\beta_1 \in \Re$, where the support of f is contained in $(\underline{f}, \bar{f})$ under H and is contained in $(\underline{f}^*, \bar{f}^*)$ under H^* . But since

$$F_1(X_1'\beta_1 + \underline{f}) \leq \int_{\underline{f}}^{\bar{f}} F_1(X_1'\beta_1 + f)dH(f) = \int_{\underline{f}^*}^{\bar{f}^*} F_1(\gamma^* X_1'\beta_1 + f)dH^*(f) \leq F_1(\gamma^* X_1'\beta_1 + \bar{f}^*)$$

and since F_1 is strictly increasing over all of $\Re$, this can only be true if

$$(y + \underline{f}) \leq (\gamma_1^* y + \bar{f}^*)$$

for all $y \in \Re$. But this can only be possible if $\gamma^* = 1$.

■

Proof of Lemma 3: We know A(0)=0 so define $(a_1, a_2, \ldots, a_k)$ so that

$$A(t) = \sum_{n=1}^{k} a_n t^n.$$

Consider a possible candidate for Ψ,

$$\tilde{\Psi}(t) = \sum_{n=1}^{k} \frac{a_n}{\gamma_1^n - \gamma_2^n} t^n.$$

Then

$$A(t) = \tilde{\Psi}(\gamma_1 t) - \tilde{\Psi}(\gamma_2 t).$$

Thus if for some constant c, $\Psi = \tilde{\Psi} + c$ in some neighborhood of zero, then it must be a polynomial of degree k in a neighborhood of zero. Therefore we only need to show that

$$V(t) = \Psi(t) - \tilde{\Psi}(t)$$

is constant in some neighborhood of zero. Without loss of generality assume $|\gamma_2| > |\gamma_1|$.

Suppose there is no neighborhood around zero for which V is constant , in particular suppose that there exists $\epsilon > 0$ and $t^* < \delta|\gamma_2|$ such that $|V(t^*) - V(0)| > \epsilon$. Note that ,

$$A(t) = \Psi(\gamma_1 t) - \Psi(\gamma_2 t) = \tilde{\Psi}(\gamma_1 t) - \tilde{\Psi}(\gamma_2 t)$$

for $|t| < \delta$. This implies that $V(\gamma_1 t) = V(\gamma_2 t)$ for $|t| < \delta$, so

$$\begin{aligned} V(t^*) &= V\left(\gamma_2\left(\frac{1}{\gamma_2}\right)t^*\right) = V\left(\gamma_1\left(\frac{1}{\gamma_2}\right)t^*\right) = V\left(\left(\frac{\gamma_1}{\gamma_2}\right)t^*\right) \\ &= V\left(\gamma_2\left(\frac{\gamma_1}{\gamma_2^2}\right)t^*\right) = \ldots = V\left(\left(\frac{\gamma_1}{\gamma_2}\right)^n t^*\right) \end{aligned}$$

But since V is continuous at zero and

$$\lim_{n\to\infty}\left(\frac{\gamma_1}{\gamma_2}\right)^n = 0$$

then there must be some n for which

$$|V(t^*) - V(0)| = \left|V\left(\left(\frac{\gamma_1}{\gamma_2}\right)^n t^*\right) - V(0)\right| < \epsilon$$

a contradiction. In some neighborhood of zero $V(t)$ must be constant, so $\Psi(t)$ must be a polynomial of degree k.

■

Proof of Proposion 4: Let $\phi(t_1, t_2)$ be the characteristic function of (u_1, u_2), then under Assumption 1

$$\phi(t_1, t_2) = \phi_1(\gamma_1 t_1)\phi_H(\gamma_1 t_1 + \gamma_2 t_2)\phi_2(\gamma_2 t_2)$$

where H is the distribution of f Suppose there exists $(\gamma^*, H^*) \in \Omega$ such that

$$\begin{aligned} \phi(t_1, t_2) &= \phi_1(\gamma_1^* t_1)\phi_{H^*}(\gamma_1^* t_1 + \gamma_2^* t_2)\phi_2(\gamma_2^* t_2) \\ &= \phi_1(\gamma_1 t_1)\phi_H(\gamma_1 t_1 + \gamma_2 t_2)\phi_2(\gamma_2 t_2) \end{aligned} \tag{6.1}$$

We will first show, using Lemma 1, that (6.1) implies that

$$\frac{\gamma_2}{\gamma_1} = \frac{\gamma_2^*}{\gamma_1^*}.$$

Suppose not, suppose that

$$\frac{\gamma_2}{\gamma_1} \neq \frac{\gamma_2^*}{\gamma_1^*} \tag{6.2}$$

then in a neighborhood of the origin let

$$\begin{aligned}\Psi_1(u) &\equiv \log(\phi_H(\gamma_1 u))\\ \Psi_2(u) &\equiv -\log(\phi_{H^*}(\gamma_1^* u))\\ A(u) &\equiv \log(\phi_1(\gamma_1^* u)) - \log(\phi_1(\gamma_1 u))\\ B(u) &\equiv \log(\phi_2(\gamma_2^* u)) - \log(\phi_2(\gamma_2 u))\end{aligned}$$

So (6.1) implies that

$$\Psi_1(t_1 + \frac{\gamma_2}{\gamma_1} t_2) + \Psi_2(t_1 + \frac{\gamma_2^*}{\gamma_1^*} t_2) = A(t_1) + B(t_2)$$

Therefore, from Lemma 1 Ψ_1, Ψ_2, A and B must all be polynomials of degree ≤ 2 in a neighborhood of the origin. But then Lemma 1 and Lemma 2 imply that $\varepsilon_1, \varepsilon_2$, and f must all be normally distributed [5].

Since they are all normal with median zero, they can be completely characterized by their variance. Let σ_f^2 be the variance associated with the distribution function H, and let σ_f^{*2} be the variance associated with H^*, σ_1^2 be the variance of ε_1, and σ_2^2 be the variance of ε_2. Thus if (γ_1, γ_2, H) is not identified relative to $(\gamma_1^*, \gamma_2^*, H^*)$ then

$$\begin{aligned}\gamma_1^2(\sigma_f^2 + \sigma_1^2) &= \gamma_1^{*2}(\sigma_f^{*2} + \sigma_1^2)\\ \gamma_2^2(\sigma_f^2 + \sigma_2^2) &= \gamma_2^{*2}(\sigma_f^{*2} + \sigma_2^2)\\ \gamma_1\gamma_2\sigma_f^2 &= \gamma_1^*\gamma_2^*\sigma_f^{*2}\end{aligned}$$

Solutions of this system of equations yields $\sigma_f^2 = \sigma_f^{*2}$. So

$$\frac{\gamma_2}{\gamma_1} = \frac{\gamma_2^*}{\gamma_1^*}$$

which contradicts (6.2).

We therefore know that

$$\frac{\gamma_2}{\gamma_1} = \frac{\gamma_2^*}{\gamma_1^*}.$$

[5] Actually f may also be degenerate, but in what follows we can think of this as a normal with mean zero and variance zero.

In equation (6.1), let

$$t_2 = \frac{-\gamma_1 t_1}{\gamma_2}.$$

which implies

$$\phi_1(\gamma_1 t_1)\phi_2(\gamma_1 t_1) = \phi_1(\gamma_1^* t_1)\phi_2(\gamma_1^* t_1)$$

Since the left hand side is the characteristic function for the random variable $\gamma_1(\varepsilon_1 - \varepsilon_2)$ and the right hand side is the characteristic function for the random variable $\gamma_1^*(\varepsilon_1 - \varepsilon_2)$, from the uniqueness theorem for characteristic functions $\gamma_1 = \gamma_1^*$. By similar reasoning we can show $\gamma_2 = \gamma_2^*$. But then from (6.1) and Assumption 3, we can see that $\phi_H = \phi_{H^*}$, and thus $H = H^*$.

■

Proof of Proposition 5: We proceed in a manner very similar to the proof of Proposition 4. Clearly if the model is identified when $T = 3$ it is identified for $T > 3$. So we will assume $T = 3$. Suppose there exists $(\gamma, \alpha, H) \in \Omega$ and $(\gamma^*, \alpha^*, H^*) \in \Omega$ for which

$$\begin{aligned} &\phi_1(\gamma_1 t_1)\phi_2(\gamma_2 t_2)\phi_3(\gamma_3 t_3)\phi_H(\gamma_1 t_1 + \alpha_2\gamma_2 t_2 + \alpha_3\gamma_3 t_3) \\ &\quad = \phi_1(\gamma_1^* t_1)\phi_2(\gamma_2^* t_2)\phi_3(\gamma_3^* t_3)\phi_H(\gamma_1^* t_1 + \alpha_2^*\gamma_2^* t_2 + \alpha_3^*\gamma_3^* t_3). \end{aligned} \tag{6.3}$$

Consider the case,

$$\frac{\alpha_2^*\gamma_2^*}{\gamma_1^*} \neq \frac{\alpha_2\gamma_2}{\gamma_1} \qquad \frac{\alpha_3^*\gamma_3^*}{\gamma_1^*} \neq \frac{\alpha_3\gamma_3}{\gamma_1}$$

First set $t_3 = 0$ and apply the logic above to show that f, ε_1, and ε_2 must all be normal. Setting $t_2 = 0$ we can show ε_3 is normal.

Next consider the case

$$\frac{\alpha_2^*\gamma_2^*}{\gamma_1^*} \neq \frac{\alpha_2\gamma_2}{\gamma_1} \qquad \frac{\alpha_3^*\gamma_3^*}{\gamma_1^*} = \frac{\alpha_3\gamma_3}{\gamma_1}$$

First set $t_3 = 0$ and apply the logic above to show that f, ε_1, and ε_2 must all be normal. Then setting $t_3 = t_1$ we can show ε_3 is normal.

Similarly if

$$\frac{\alpha_2^*\gamma_2^*}{\gamma_1^*} = \frac{\alpha_2\gamma_2}{\gamma_1} \qquad \frac{\alpha_3^*\gamma_3^*}{\gamma_1^*} \neq \frac{\alpha_3\gamma_3}{\gamma_1}$$

we can show that $f, \varepsilon_1, \varepsilon_2$, and ε_3 must all be normal.

If we are in any of these three cases we are left with the model in Example 4. It is easy to show that the parameters are all identified in that case.

This leaves only the following possibility,

$$\frac{\alpha_2^*\gamma_2^*}{\gamma_1^*} = \frac{\alpha_2\gamma_2}{\gamma_1} \qquad \frac{\alpha_3^*\gamma_3^*}{\gamma_1^*} = \frac{\alpha_3\gamma_3}{\gamma_1} \tag{6.4}$$

Let

$$\begin{aligned}\Psi(t) &\equiv \log(\phi_H(\gamma_1 t)) - \log(\phi_{H^*}(\gamma_1^* t))\\ A(t) &\equiv \log(\phi_1(\gamma_1^* t)) - \log(\phi_1(\gamma_1 t))\\ B(t) &\equiv \log(\phi_2(\gamma_2^* t)) - \log(\phi_2(\gamma_2 t))\end{aligned}$$

then by setting $t_3 = 0$ and taking logs of equation (6.3) we know,

$$\Psi(t_1 + \frac{\alpha_2\gamma_2}{\gamma_1} t_2) = A(t_1) + B(t_2)$$

From Lemma 1, A and B must be polynomials of degree zero or one. From Lemmas 2 and 3, if $\gamma_1 \neq \gamma_1^*$ then ϕ_1 must be a polynomial of degree zero or one which is a contradiction to ϕ_1 nondegenerate, so $\gamma_1 = \gamma_1^*$. Similarly, we can show $\gamma_2 = \gamma_2^*$. But then $A(t_1) = 0$ and $B(t_2) = 0$ for all t_1 and t_2, so $\Psi(t) = 0$ for all t, which implies that $\phi_H = \phi_{H^*}$. From the marginal distributions we know

$$\phi_2(\gamma_2 t_2)\phi_H(\alpha_2\gamma_2 t_2) = \phi_2(\gamma_2^* t_2)\phi_H(\alpha_2^*\gamma_2^* t_2)$$

So using Assumption 3 we can identify α_2. We can use exactly the same argument as above to show identification of α_3 and γ_3, so (γ, α, H) is identified in Ω.

■

References

Bartholemew, D. J. (1973), *Stochastic Models for Social Processes,* 2nd Edition.

Cameron, S. V. and Heckman J. J. (1998), "Life-Cycle Schooling and Educational Selectivity: Models and Evidence," *Journal of Political Economy,* **49**, 333-377.

________ and ________ (1992), "The Dynamics of Education among Blacks, Hispanics, and Whites," unpublished manuscript, Columbia University and University of Chicago.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society,* Ser. B, **49**, 1-38.

Everitt, B. S., and Hand, D. J. (1981), *Finite Mixture Distributions,* New York: Chapman & Hall.

Follmann, D. A. (1985), "Nonparametric Mixtures of Logistic Regression Models," unpublished Ph.D. dissertation, Carnegie-Mellon University, Dept. of Statistics.

Follmann, D. A., and Lambert, D. (1989), "Generalizing Logistic Regression by Nonparametric Mixing," *Journal of the American Statistical Association,* **84**, 295-300.

Heckman, J. J. and Singer B. (1984), "A Method for Minimizing the Impact of Distributional Assumptions," *Econometrica,* **52**, 222

Kiefer, J., and Wolfowitz, J. (1956), "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *Annals of Mathematical Statistics,* **27**, 363-366.

Laird, N. (1978), "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association,* **73**, 805-811.

Lesperance M. L., and Kalbfleisch, J. D. (1992), "An Algorithm for Comupting the Nonparametric MLE of a Mixing Distribution," *Journal of the American Statistical Association,* **87**, 120-126.

Lindsey, J.K. (1983a),"The Geometry of Mixture Likelihoods: A General Theory", *Annals of Statistics*, **11**, 86-94.

Lindsey, J.K. (1983b),"The Geometry of Mixture Likelihoods, Part II", *Annals of Statistics.*

Mare, R. D. (1981), "Social Background and School Continuation Decisions," *Journal of the American Statistical Association,* **75**, 295-305.

Pearson, K. (1894), "Contribution to the Mathematical Theory of Evolution," *Phil. Transactions A,* **185**, 71-110.

Simar, L. (1976), "Maximum Likelihood Estimation of a Compound Poisson Process," *The Annals of Statistics*, **4**, 1200-1209.

Taber, Chris (1999), "Identification in Dynamic Models of Schooling," Northwestern University Working Paper.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: John Wiley.

Yakowitz, S. J. and Spragins, J. D. (1968), "On the Identifiability of Finite Mixtures," *Annals of Mathematical Statistics*, **39**, 209-214.

Table 1

Monte Carlo Results from 100 Runs (500 Observations in Each Run)†
Means of the Slope Parameter β and its Standard Error‡
(True Value of β is 1.0)

A. True Heterogeneity Distribution is Normal§

	# of	# of			Mean of Std
	Points	*Runs*	*Mean (Std Dev)*	*Median*	*Error (Std Dev)*
Applying the Nonparametric Heterogeneity Correction	All Runs	100	1.0021 (.053)	0.9935	0.0484 (.002)
	3 Points	1	0.9386 (na)	0.9386	0.0484 (na)
	4 Points	92	1.0028 (.053)	0.9949	0.0484 (.002)
	5 Points	7	1.0016 (.052)	0.9845	0.0475 (.002)
No Heterogeneity Correction (Standard Logit):	0 Points	100	0.6248 (.040)	0.6264	0.0311 (.002)

B. True Heterogeneity Distribution is Binomial§

Applying the Nonparametric Heterogeneity Correction:	All Runs	100	0.9976 (.047)	0.9947	0.0408 (.002)
	2 Points	91	0.9971 (.047)	0.9943	0.0407 (.002)
	3 Points	9	1.0026 (.048)	0.9950	0.0421 (.001)
No Heterogeneity Correction (Standard Logit)	0 Points	100	0.8138 (.038)	0.8068	0.0348 (.002)

na = not applicable: only one run in this category.

† The data was generated as follows. The slope parameter β was estimated 100 times on samples of 500 observations. Each observation consisted of a 10 period long vector (d_{tir}, x_{tir}), where r denotes the run (1 to 100), i denotes the observation in run r (1 to 500), and t denotes the period (1 to 10). Let $U_{tir} = x_{tir}\beta + f_{ir} + \varepsilon_{tir}$. The binary indicator d_{tir}=1 iff $U_{tir} \geq 0$. (Note that x_{tir} varies across periods.) The ε_{tir} is drawn from a logistic distribution with mean 0 and variance 1. The β is equal to 1.0, and x_{tir} is drawn from a N(0,1) for all t,i, and r. The unobserved heterogeneity component f_{ir} is drawn from a normal or binomial distribution (see below).

‡ The mean reported above is simply the mean of the empirical distribution of the 100 estimated β parameters, and the standard deviation of that distribution is reported in parentheses following the mean. The mean of the standard error reported in the right column is the mean of the empirical distribution of the standard error of the estimated β taken from the estimated information matrix for each run. The standard deviation of that distribution is reported in parentheses following the mean. The information matrix was calculated using the outer-product approximation to the hessian matrix (formed from analytic first derivatives) described in Anderson (1969) or Berndt, et al (1977).

§ The heterogeneity component f_{ir} is drawn from a N(-1,4) distribution for panel A results, and a binomial distribution with equal mass at -1.0 and 1.0 (i.e., mean zero and variance 1) for panel B results.

Table 2

Support and Mass Points of the Estimated Heterogeneity Distribution†
(Mass Points are Ordered Lowest to Highest)

A. True Heterogeneity is Normal‡ (Compare Table 1, Panel A)

	Support Point (θ_i)		*Mass at θ_i (μ_i)*	
Estimated Number of Points	*Mean (Std Dev)*	*Median*	*Mean (Std Dev)*	*Median*
3	-2.67 (na)	-2.67	.45 (na)	.45
	-0.44 (na)	-0.44	.37 (na)	.37
	1.63 (na)	1.63	.19 (na)	.19
4	-4.25 (2.92)	-3.46	.25 (.08)	.27
	-1.32 (.39)	-1.27	.36 (.04)	.37
	0.44 (.36)	0.45	.27 (.05)	.26
	2.54 (.59)	2.46	.12 (.04)	.12
5	-9.99 (7.51)	-3.99	.16 (.05)	.16
	-1.84 (.25)	-1.70	.31 (.04)	.31
	-0.21 (.22)	-0.11	.32 (.03)	.31
	1.44 (.26)	1.37	.17 (.02)	.18
	8.80 (7.60)	4.37	.04 (.02)	.03

B. True Heterogeneity is Binomial§ (Compare Table 1, Panel B)

2	-1.00 (.06)	-1.00	.50 (.03)	.51
	1.01 (.07)	1.01	.50 (.03)	.49
3	-3.96 (4.33)	-1.66	.21 (.20)	.18
	-0.27 (.73)	-0.48	.44 (.08)	.46
	1.40 (.62)	1.15	.35 (.19)	.43

† For a full description of the model and its parameters see the base of Table 1.

‡ The mean of the true normal heterogeneity distribution is -1.0 and its variance is 4.0.

§ The mean of the true binomial heterogeneity distribution is 0 and its variance is 1.0.

Table 3

Comparisons of Standard Error Estimates
Mean Standard Error (Standard Deviation of the Standard Error)

Model	*True distribution is Normal (Table 1.A)*	*True distribution is Binomial (Table 1.B)*
Estimated with Nonparametric Heterogeneity Correction (Table 1)	0.484 (.002)	.0408 (.002)
Estimated When Functional Form of Heterogeneity Distribution Known*	0.482 (.002)	.0407 (.002)
Standard Errors Calculated Using Generated Data and True Parameters†	0.482 (.002)	.0407 (.002)
Standard Errors Computed Using Generated X's and Integrating Out Heterogeneity and Logistic Error‡	0.479 (na)	.0407 (na)

na= Not applicable: only one calculation was made for the entire realized distribution of X.

* We estimate the model assuming the functional form of the heterogeneity distribution is normal with unknown mean and variance (which we estimate) for column one and binomial for column two (i.e., has exactly 2 points of support with unknown distribution).

† Using the same generated X_t and d_t vectors on which the models were estimated, we take the standard error from the inverse of the information matrix for each run evaluated at the true parameter values and then average over all runs.

‡ For these calculations, we retain only the values of the X_t vectors generated for the Table 1 estimates (and not the d_t vectors), and then for each draw in each run we integrated out the heterogeneity term and the logistic error given the true values of the parameters. We then calculate an information matrix by combining all draws from all runs. An alternative procudure would be to to forming inverse information matrices for each run and then average over all runs; in fact, we found no difference between the two procedures.

Table 4

Monte-Carlo Results Using Alternative Heterogeneity Distributions and Applying Different Heterogeneity Corrections (100 Runs)†
(True β = 1.0)

	Mean Estimates					
	No Heterogeneity Correction		*Normal Correction*		*Nonparametric Correction*	
True Heterog. Distribution	*Mean (std dev)*	*Std Err (std dev)*	*Mean (std dev)*	*Std Err (std dev)*	*Mean (std dev)*	*Std Err (std dev)*
Normal (Table 1)	0.6248 (.040)	.0311 (.002)	0.9998 (.053)	.0482 (.002)	1.0021 (.053)	.0484 (.002)
Mixture of Normals‡	0.1632 (.032)	.0275 (.001)	0.8487 (.099)	.0792 (.006)	1.0008 (.093)	.0891 (.007)
Exponential‡	0.5701 (.053)	.0431 (.002)	1.0004 (.070)	.0673 (.004)	1.0006 (.068)	.0681 (.004)

† Only the mean and variance of the normal were estimated along with β. The Monte-Carlo design is identical except for the underlying heterogeneity distribution.

‡ The mean and variance of mixture of normals used to generate the data for the row 2 estimates were 6.0 and 90.0 (a mixture of a N(-3,9) and a N(15,9) normal with equal mass on each). For row 3, the underlying heterogeneity was drawn from an exponential with mean 8 and variance 64.

Table 5

Predictions for the Models in Table 4
Conditional on Values of Exogenous Variables†
Mean Prediction and Standard Deviation of the Prediction in Parentheses
('*' Indicates the Average Prediction Fails a Two-Tailed T-Test at a 95% Level)

A. Basic Model with Mixed-Normal Heterogeneity (Table 4)

Estimation Strategy	*Pr(*$d_1 = 1 \vert x$*)*	*Pr(*$d_1 = 1, d_2 = 1 \vert x$*)*	*Pr(*$d_1 = 1, \ldots, d_5 = 1 \vert x$*)*
	i. x_i *= 0.0 for all i*		
True Probability	.598	.558	.530
No Heterogeneity Control	.600 (.023)	.360*(.027)	.079*(.015)
Normal Hetero. Control	.559 (.126)	.486*(.142)	.428*(.153)
Nonparametric Control	.595 (.023)	.555 (.025)	.527 (.026)
	ii. x_i *= .675 for all i*		
True Probability	.627	.581	.545
No Heterogeneity Control	.626 (.022)	.392*(.027)	.097*(.017)
Normal Hetero. Control	.601*(.118)	.525*(.134)	.454*(.152)
Nonparametric Control	.624 (.022)	.578 (.024)	.542 (.025)
	iii. x_i *= -.675 for all i*		
True Probability	.573	.540	.519
No Heterogeneity Control	.573 (.025)	.329*(.028)	.079*(.015)
Normal Hetero. Control	.519*(.133)	.480*(.148)	.454*(.152)
Nonparametric Control	.571 (.024)	.538 (.025)	.516 (.026)

B. Basic Model with Exponential Heterogeneity (Table 4)

	i. x_i *= 0.0 for all i*		
True Probability	.860	.785	.690
No Heterogeneity Control	.867*(.011)	.751*(.020)	.489*(.032)
Normal Hetero. Control	.865 (.031)	.799*(.039)	.692 (.041)
Nonparametric Control	.860 (.012)	.787 (.017)	.690 (.021)
	ii. x_i *= .675 for all i*		
True Probability	.905	.844	.750
No Heterogeneity Control	.905 (.009)	.819*(.016)	.608*(.030)
Normal Hetero. Control	.904 (.027)	.854*(.035)	.765*(.044)
Nonparametric Control	.906 (.009)	.845 (.019)	.750 (.020)
	iii. x_i *= -.675 for all i*		
True Probability	.807	.726	.633
No Heterogeneity Control	.816*(.015)	.665*(.024)	.362*(.033)
Normal Hetero. Control	.815 (.035)	.732 (.039)	.613*(.034)
Nonparametric Control	.808 (.015)	.727 (.019)	.636 (.022)

C. Basic Model with Normal Heterogeneity (Table 4)

Estimation Strategy	*Pr(*$d_1 = 1 \mid x$*)*	*Pr(*$d_1 = 1, d_2 = 1 \mid x$*)*	*Pr(*$d_1 = 1, \ldots, d_5 = 1 \mid x$*)*
i. x_i *= 0.0 for all i*			
True Probability	.352	.212	.098
No Heterogeneity Control	.352 (.015)	.124*(.010)	.005*(.001)
Normal Hetero. Control	.352 (.015)	.210 (.014)	.094*(.011)
Nonparametric Control	.353 (.015)	.212 (.014)	.097 (.011)
ii. x_i *= .675 for all i*			
True Probability	.451	.299	.157
No Heterogeneity Control	.452 (.016)	.205*(.015)	.005*(.001)
Normal Hetero. Control	.452 (.015)	.299 (.015)	.156 (.014)
Nonparametric Control	.452 (.016)	.301 (.016)	.158 (.004)
iii. x_i *= -.675 for all i*			
True Probability	.263	.139	.055
No Heterogeneity Control	.262 (.014)	.069*(.007)	.001*(.000)
Normal Hetero. Control	.262 (.015)	.138 (.012)	.056 (.008)
Nonparametric Control	.264 (.014)	.140 (.012)	.052*(.009)

† The likelihood and parameters of the above model are specified in Table 4.

Table 6

Monte Carlo Results from 100 Runs with Time-Invariant x
500 Observations in Each Run (True Value of β is 1.0)
Means of the Slope Parameter β and its Standard Error
(Standard Deviations in Parentheses)

A. True Heterogeneity Distribution is Normal†

	# of Points	*# of Runs*	*Mean (Std Dev)*	*Median*	*Mean of Std Error (Std Dev)*
Applying the Nonparametric Heterogeneity Correction	All Runs	100	0.9993 (.257)	1.0052	.1604 (.031)
	3 Points	55	0.9185 (.247)	0.9597	.1453 (.020)
	4 Points	41	1.0919 (.234)	0.9901	.1792 (.030)
	5 Points	4	1.1600 (.280)	1.0112	.1945 (.061)
No Heterogeneity Correction (Standard Logit):	0 Points	100	0.6113 (.062)	0.6174	.0162 (.001)

B. True Heterogeneity Distribution is Binomial†

	# of Points	*# of Runs*	*Mean (Std Dev)*	*Median*	*Mean of Std Error (Std Dev)*
Applying the Nonparametric Heterogeneity Correction	All Runs	100	0.9957 (.059)	0.9939	.0559 (.004)
	2 Points	90	0.9934 (.058)	0.9939	.0554 (.003)
	3 Points	10	1.0164 (.070)	1.0151	.0606 (.005)
No Heterogeneity Correction (Standard Logit)	0 Points	100	0.8133 (.053)	0.8062	.0234 (.001)

C. Comparisons of Standard Error Estimates
Mean of the Standard Error (Standard Deviation in Parentheses)

Model	*Normal (panel A)*	*Binomial (panel B)*
Estimated with Functional Form of Heterogeneity Known (The Mean $\hat{\beta}$=1.0209 for normal and 0.9923 for binomial)	.1870 (.018)	.0550 (.004)
Standard Errors Calculated Using Generated Data and True Parameters	.1884 (.020)	.0550 (.004)

na = not applicable: only one run in this category.

Note: the data and estimates above were generated in the same manner described at the base of Table 1 except that the observed exogenous variable x_{tir} is constant across periods for the above results (i.e., x_{tir} = x_{1ir} for all i,r). The calculations for the standard error comparisons are detailed at the base of Table 3.

† In panel A, 3 runs converged with a singular Hessian matrix and in panel B results, 1 run was had a singular Hessian. These runs were used in forming the $\hat{\beta}$ distribution but not the $\hat{\sigma}$ distribution.

Table 7

Monte Carlo Results from 1000 Runs† when the Heterogeneity is Normal Means of the Slope Parameter β and its Standard Error (True Value of β is 1.0)

A. Estimates with 500 Observations in Each Run

# of Points	*# of Runs*	*Mean (Std Dev)*	*Median*	*Mean of Std Error (Std Dev)*
All Runs	1000	1.0000 (.050)	0.9996	.0484 (0.002)
3 Points	25	0.9694 (.041)	0.9735	.0472 (0.002)
4 Points	899	0.9999 (.049)	0.9995	.0480 (0.002)
5 Points	76	1.0118 (.051)	1.0092	.0491 (0.002)

B. Estimates with 250 Observations in Each Run

# of Points	*# of Runs*	*Mean (Std Dev)*	*Median*	*Mean of Std Error (Std Dev)*
All Runs	1000	1.0021 (.067)	0.9904	.0689 (.004)
3 Points	440	0.9801 (.066)	0.9898	.0674 (.004)
4 Points	550	1.0045 (.068)	1.0001	.0695 (.004)
5 Points	10	1.0483 (.093)	1.0115	.0705 (.004)

† The likelihood and parameters of the above model are specified exactly as those reported in Table 1.A (except that the panel B estimates were obtained on samples with 250 observations in each run).

Table 8

Tests of Normality for the Monte Carlo Distribution of Estimated Slope Parameters for Experiments Reported in Table 7

A. Shapiro, Wilk Normality Tests

	500 Observations *(Compare Table 7.A)*	*250 Observations* *(Compare Table 7.B)*
P-Value	.87	.61

B. Probablity of Rejecting $\hat{\beta}_i/\hat{\sigma}_i$ from a Normal Distribution (H_o: β=1)

Size of Test	*Portion Rejected*	
One Percent	.016	.010
Five Percent	.053	.046
Ten Percent	.100	.089

Table 9

Effect on Estimates of β of Alternative Stopping Rules for the Number of Support Points C (True β = 1.0)

	Normal			*Binomial*		
	Mean (std dev)	*Std Err (std dev)*	$\bar{c}$	*Mean (std dev)*	*Std Err (std dev)*	$\bar{c}$
A. X Varies over Time						
Bayesian Information Criterion	0.9974 (.053)	.0480 (.002)	3.70	0.9964 (.047)	.0407 (.002)	2.00
Akaike Information Criterion (Table 1)	1.0021 (.053)	.0484 (.002)	4.06	0.9976 (.047)	.0408 (.002)	2.09
Log-Likelihood Less Than .01†	1.0045 (.054)	.0486 (.002)	4.72	1.0007 (.047)	.0412 (.002)	2.71
B. X Fixed over Time						
Bayesian Information Criterion	0.9361 (.254)	.1402 (.022)	3.01	0.9923 (.058)	.0550 (.004)	2.00
Akaike Information Criterion (Table 4)	0.9993 (.257)	.1605 (.031)	3.49	0.9957 (.059)	.0559 (.004)	2.10
Log-Likelihood Less Than .01†	1.0897 (.259)	.1906 (.028)	4.34	1.0046 (.060)	.0595 (.004)	2.74

Note: Estimates and standard deviations produced by using a stopping rule based on a log-likelihood change of .0001 were almost identical to those based on a .01 change in log-likelihood as a stopping rule except that the mean number of support points increased by 5 to 10 percent.

† Since we are using log-likelihood as a criterion, a change in its value of less than .01 (.0001) is the same as a 1% (.01%) change in the likelihood.

Table 10

Effect of Changing Sample Size on the Basic Model Presented in Table 1†
True Heterogeneity Distribution is Normal
(True β = 1.0)

Number of Cases	*Mean (std dev)*	*Std Err (std dev)*	$\bar{c}$
125 Cases	0.9957 (.0823)	.0973 (.008)	3.41
250 Cases	1.0043 (.0688)	.0695 (.004)	3.66
500 Cases (Table 1)	1.0021 (.0534)	.0484 (.002)	4.06
2000 Cases	1.0013 (.0281)	.0241 (.002)	4.58
5000 Cases	1.0005 (.0156)	.0152 (.000)	4.95

† Details of the Monte-Carlo design are provided at the base of Table 1.

Table 11

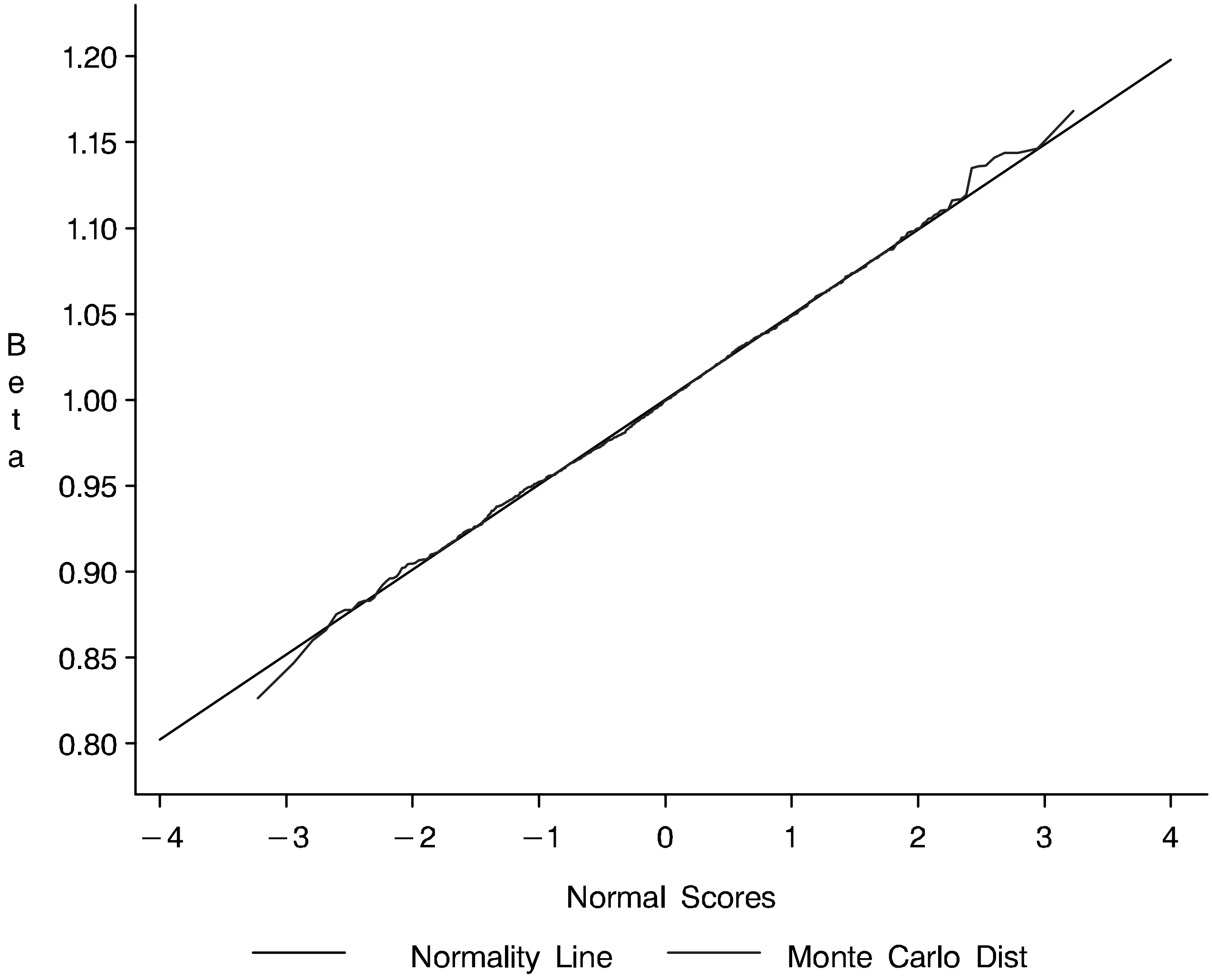
Alternative Versions of the Basic Model Presented in Table 1 (True β = 1.0)

	A. Normal			*B. Binomial*		
	Mean (std dev)	*Std Err (std dev)*	$\bar{c}$	*Mean (std dev)*	*Std Err (std dev)*	$\bar{c}$
Basic Model from Table 1	1.0021 (.053)	.0484 (.002)	4.06	0.9976 (.047)	.0408 (.002)	2.09
Diffuse Starting Values for Heter. Distribution†	1.0021 (.053)	.0484(.002)	4.06	0.9975 (.047)	.0408 (.002)	2.08
Zero Starting Values for β‡	1.0021 (.053)	.0484 (.002)	4.06	0.9976 (.047)	.0408 (.002)	2.09
Decreasing Variance of X from 1 to .25	1.0093 (.102)	.0843 (.002)	3.96	0.9961 (.097)	.0803 (.002)	2.04
Increasing Pr(d_t) from .47 to .85	1.0118 (.053)	.0624 (.002)	3.73	0.9986 (.049)	.0495 (.002)	2.11

† Rather than using the Lindsay-Simar algorithm (see Section 1 of the text) to choose the value of the new support point when increasing the number of support points from C to C+ 1, we let the starting values of the support take values equidistant while holding the variance and mean of the estimated heterogeneity distribution constant (and letting the starting values for β be the optimum values that maximize the likelihood when the number of support points is fixed at C.

‡ We employ the same algorithm used to choose the heterogeneity distribution described for panel C with the exception that starting values for β are set to zero whenever C is increased.

Figure 1. Quantile–Quantile Plot of the Monte Carlo Distribution of Beta (from Table 1, Panel A).



Note: Deviations of normal scores (inverse normal of empirical distribution) represent departures from normality. Plotted against values of a standard normal. The mean (std dev) of the empirical distribution of beta is 1.000 (.0501).