Discrete Choices with Panel Data

Manuel Arellano
CEMFI

Working Paper No. 0101
January 2001

This paper has been written for presentation as the *Investigaciones Económicas* Lecture at the XXV Simposio de Análisis Económico, Universitat Autònoma de Barcelona, Bellaterra, 19-21 December 2000. I wish to thank Pedro Albarrán, Olympia Bover, Raquel Carrasco, Jesús Carro, and Enrique Sentana for helpful discussions. I am also grateful to Pedro Albarrán and Jesús Carro for very able research assistance. (Email address: arellano@cemfi.es).

CEMFI, Casado del Alisal 5, 28014 Madrid, Spain
www.cemfi.es

Abstract

This paper reviews the existing approaches to deal with panel data binary choice models with individual effects. Their relative strengths and weaknesses are discussed. Much theoretical and empirical research is needed in this area, and the paper points to several aspects that deserve further investigation. In particular, I illustrate the usefulness of asymptotic arguments in providing both approximately unbiased moment conditions, and approximations to sampling distributions for panels of different sample sizes.

JEL classification: C23.
Keywords: Binary choice, panel data, fixed effects, modified likelihood, asymptotic corrections.

1 Introduction

This paper reviews the existing approaches to deal with panel data discrete choice models with individual effects. Their relative strengths and weaknesses are discussed. Much theoretical and empirical research is needed in this area, and the paper points to several aspects that deserve further investigation. In particular, I illustrate the usefulness of time series asymptotic arguments in providing both approximately unbiased moment conditions, and approximations to sampling distributions even for fairly short panels. I will focus on the static binary case for simplicity and because many results are only available for this case.

2 Models and Parameters of Interest

I begin by considering the following static binary choice model

$$y_{it} = 1\left\{x_{it}'\beta_0 + \eta_i + v_{it} \geq 0\right\} \ \ (t = 1, ..., T; i = 1, ..., N) \qquad (1)$$

where the errors v_{it} are independently distributed with *cdf* F conditional on η_i and $x_i = (x_{i1}', ..., x_{iT}')'$, so that

$$\Pr(y_{it} = 1 \mid x_i, \eta_i) = F\left(x_{it}'\beta_0 + \eta_i\right). \qquad (2)$$

The Linear Model as a Benchmark In a linear model of the form

$$E(y_{it} \mid x_i, \eta_i) = x_{it}'\beta_0 + \eta_i, \qquad (3)$$

β_0 is identifiable from the regression in first differences or deviations from means in a cross-sectional population for fixed T, regardless of the form of the distribution of $\eta_i \mid x_i$. That is, we have

$$p \lim_{N\to\infty} \frac{1}{TN} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it} - \overline{x}_i\right) \left[\left(y_{it} - \overline{y}_i\right) - \left(x_{it} - \overline{x}_i\right)' \beta_0\right] = 0, \qquad (4)$$

which is uniquely satisfied by the true value β_0 provided

$$p \lim_{N \to \infty} \frac{1}{TN} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \overline{x}_i)(x_{it} - \overline{x}_i)' \qquad (5)$$

is non-singular. So, the value $\widehat{\beta}$ that solves

$$\frac{1}{TN} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \overline{x}_i) \left[(y_{it} - \overline{y}_i) - (x_{it} - \overline{x}_i)' \widehat{\beta} \right] = 0 \qquad (6)$$

(the "within-group" estimator) is a consistent estimator of β_0 for large N, no matter how small is T as long as $T \geq 2$ (see, for example, Hsiao, 1986).

This is of economic interest if one hopes that by conditioning on η_i, β_0 measures a more relevant (causal or structural) effect of x on y. The consistency result matters because one wants to make sure that gets the right answer when calculating $\widehat{\beta}$ from a large cross-sectional panel with a small time series dimension, which is a typical situation in microeconometrics.

The motivation and aim in a binary choice fixed effects model is to get similar results as in the linear case when the form of the model is given by (1). In our context, the term "fixed effects" has nothing to do with the nature of sampling. It just refers to a model for the effect of x on y given x and η, in which we observe y and x but not η, and the distribution of $\eta \mid x$ is left unrestricted. Following the usage in the econometric literature, the term "random effects" will be reserved for models in which some knowledge about the form of the distribution of $\eta \mid x$ is assumed.

Parameters of Interest The micropanel data literature has emphasized the large-N-short-T identification of β_0 with an unspecified distribution of $\eta_i \mid x_i$. However, a natural parameter of interest is the mean effect on the probability of $y_{it} = 1$ of changing x_{1it} from z_a to z_b, say. A consistent estimator of this is:

$$\frac{1}{N} \sum_{i=1}^{N} \int \left[F\left(z_a \beta_{01} + x_{2it}' \beta_{02} + \eta\right) - F\left(z_b \beta_{01} + x_{2it}' \beta_{02} + \eta\right) \right] dG(\eta \mid x_{2it}) \quad (7)$$

where $G(. \mid x_{2it})$ is the *cdf* of η_i conditional on x_{2it}, and x_{1it} denotes the first component of x_{it}. Thus, measuring this effect would require us to specify G, which is not in the nature of the fixed effects approach.[1]

The direct information we can get from the β coefficients only concerns the relative impacts of explanatory variables on the probabilities. If x_{1it} and x_{2it} are continuous variables we have:

$$\frac{\beta_{02}}{\beta_{01}} = \frac{\partial \Pr ob(y_{it} = 1 \mid x_i, \eta_i)}{\partial x_{2it}} / \frac{\partial \Pr ob(y_{it} = 1 \mid x_i, \eta_i)}{\partial x_{1it}}. \tag{8}$$

3 The Problem

The log-likelihood function from (1) assuming that the y_{it} are independent conditional on x_i and η_i is given by

$$\sum_{i=1}^{N} \ell_i (\beta, \eta_i) \tag{9}$$

where

$$\ell_i (\beta, \eta_i) = \sum_{t=1}^{T} \{ y_{it} \log F_{it} + (1 - y_{it}) \log (1 - F_{it}) \} \tag{10}$$

and $F_{it} = F(x'_{it}\beta + \eta_i)$. Moreover, the scores are

$$d_{\eta i} (\beta, \eta_i) \equiv \frac{\partial \ell_i (\beta, \eta_i)}{\partial \eta_i} = \sum_{t=1}^{T} \frac{f_{it}}{F_{it}(1 - F_{it})} (y_{it} - F_{it}) \tag{11}$$

$$d_{\beta i} (\beta, \eta_i) \equiv \frac{\partial \ell_i (\beta, \eta_i)}{\partial \beta} = \sum_{t=1}^{T} \frac{f_{it}}{F_{it}(1 - F_{it})} x_{it} (y_{it} - F_{it}) \tag{12}$$

where f_{it} denotes the *pdf* corresponding to F_{it}.

For the logit model F is the logistic *cdf* $\Lambda(r) = e^r / (1 + e^r)$ and we have

$$\frac{f_{it}}{F_{it}(1 - F_{it})} = 1$$

[1] An alternative is to obtain the difference in probabilities for specific values of η and x_{2t} (e.g. their means), but this may only be relevant for a small part of the population (see Chamberlain, 1984).

so that in this case the scores are simply $d_{\eta i}(\beta, \eta_i) = \sum_{t=1}^{T}(y_{it} - F_{it})$ and $d_{\beta i}(\beta, \eta_i) = \sum_{t=1}^{T} x_{it}(y_{it} - F_{it})$.

Let the MLE of η_i for given β be

$$\widehat{\eta}_i(\beta) = \arg\max_{\eta} \ell_i(\beta, \eta_i) \tag{13}$$

so that $\widehat{\eta}_i(\beta)$ solves

$$d_{\eta i}(\beta, \widehat{\eta}_i(\beta)) = 0. \tag{14}$$

Therefore, the MLE of β is given by the maximizer of the concentrated (or profile) log-likelihood

$$\widehat{\beta} = \arg\max_{\beta} \sum_{i=1}^{N} \ell_i(\beta, \widehat{\eta}_i(\beta)) \tag{15}$$

which solves the first order conditions

$$\begin{aligned} b_{TN}(\beta) &= \frac{1}{TN}\sum_{i=1}^{N}\left\{ d_{\beta i}(\beta, \widehat{\eta}_i(\beta)) + d_{\eta i}(\beta, \widehat{\eta}_i(\beta))\frac{\partial \widehat{\eta}_i(\beta)}{\partial \beta}\right\} \\ &= \frac{1}{TN}\sum_{i=1}^{N} d_{\beta i}(\beta, \widehat{\eta}_i(\beta)) \end{aligned} \tag{16}$$

The problem is that $b_{TN}(\beta)$ evaluated at $\beta = \beta_0$ does not converge to zero in probability when $N \to \infty$ for T fixed (although it does converge to zero when $T \to \infty$). This situation is known as the *incidental parameters problem* since Neyman and Scott (1948). A discussion of this problem for discrete choice models is in Heckman (1981).

An Example As a classic illustration let us consider a logit model in which $T = 2$, β is scalar, and x_{it} is a time dummy such that $x_{i1} = 0$ and $x_{i2} = 1$ (Andersen, 1973). When $T = 2$, $\widehat{\eta}_i(\beta)$ solves

$$\Lambda(\widehat{\eta}_i(\beta) + x_{i1}\beta) + \Lambda(\widehat{\eta}_i(\beta) + x_{i2}\beta) = y_{i1} + y_{i2}. \tag{17}$$

Therefore, for observations with $y_{i1} + y_{i2} = 0$ we have $\widehat{\eta}_i(\beta) \to -\infty$ and $\Lambda(\widehat{\eta}_i(\beta) + x_{i1}\beta) = \Lambda(\widehat{\eta}_i(\beta) + x_{i2}\beta) = 0$. For observations with $y_{i1} + y_{i2} = 2$

we have $\widehat{\eta}_i(\beta) \to \infty$ and $\Lambda(\widehat{\eta}_i(\beta) + x_{i1}\beta) = \Lambda(\widehat{\eta}_i(\beta) + x_{i2}\beta) = 1$. Finally, for observations with $y_{i1} + y_{i2} = 1$ $\widehat{\eta}_i(\beta)$ satisfies

$$\Lambda(\widehat{\eta}_i(\beta) + x_{i1}\beta) = 1 - \Lambda(\widehat{\eta}_i(\beta) + x_{i2}\beta),$$

so that $\widehat{\eta}_i(\beta) + x_{i1}\beta = -\widehat{\eta}_i(\beta) - x_{i2}\beta$, and

$$\widehat{\eta}_i(\beta) = -(x_{i1} + x_{i2})\beta/2. \tag{18}$$

The implication is that the contributions of observations $(0,0)$ and $(1,1)$ to the concentrated log-likelihood are equal to zero, a $(0,1)$ observation contributes a term of the form $2\log\Lambda(\Delta x_{i2}\beta/2)$, and a $(1,0)$ observation contributes with $2\log[1-\Lambda(\Delta x_{i2}\beta/2)]$. So the concentrated log-likelihood is given by

$$2\sum_{i=1}^{N}\{d_{10i}\log[1-\Lambda(\Delta x_{i2}\beta/2)] + d_{01i}\log\Lambda(\Delta x_{i2}\beta/2)\} \tag{19}$$

where $d_{10i} = 1(y_{i1} = 1, y_{i2} = 0)$ and $d_{01i} = 1(y_{i1} = 0, y_{i2} = 1)$.

Moreover, since $\Delta x_{i2} = 1$ for all observations, the MLE of $p = \Lambda(\beta/2)$ is

$$\widehat{p} = \frac{\sum_{i=1}^{N} d_{01i}}{\sum_{i=1}^{N} 1(y_{i1} + y_{i2} = 1)}, \tag{20}$$

so that

$$\widehat{\beta} = 2\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right). \tag{21}$$

Note that $\widehat{p}$ is the sample counterpart of $p_0 = \Pr(y_{i1} = 0, y_{i2} = 1 \mid y_{i1} + y_{i2} = 1)$. Thus the MLE $\widehat{\beta}$ satisfies

$$p\lim_{N\to\infty}\widehat{\beta} = 2\log\left(\frac{p_0}{1-p_0}\right) = 2\beta_0. \tag{22}$$

The last equality follows from the fact that $p_0 = \Lambda(\beta_0)$ where β_0 is the true value. Therefore, ML would be estimating a relative log odds ratio that is twice as large as its true value. This form of inconsistency for $\widehat{\beta}$ also holds for more general two-period logit models with multiple regressors.

4 Fixed T Solutions

4.1 Conditional MLE

A sufficient statistic for η_i, S_i say, is a function of the data such that the distribution of the data given S_i does not depend on η_i. The idea is to use the likelihood conditioned on S_i to make inference about β_0 (Andersen, 1970). This works as long as β_0 is identified from the conditional likelihood of the data, which obviously requires that the conditional likelihood depends on β_0. Unfortunately, this is not the case except for the logit model.

In the logit model $\sum_{t=1}^{T} y_{it}$ is a sufficient statistic for η_i. Indeed, we have

$$\Pr\left(y_{i1},...,y_{iT} \mid \sum_{t=1}^{T} y_{it}, x_i\right) = \frac{\exp\left(\sum_{t=1}^{T} y_{it}x_{it}'\beta_0\right)}{\sum_{(d_1,...,d_T)\in B_i}\exp\left(\sum_{t=1}^{T} d_t x_{it}'\beta_0\right)} \tag{23}$$

where B_i is the set of all $0-1$ sequences such that $\sum_{t=1}^{T} d_t = \sum_{t=1}^{T} y_{it}$. This result was first obtained by Rasch (1960, 1961) (for surveys see Chamberlain, 1984, or Arellano and Honoré, 2000). For example, with $T=2$ we have

$$\Pr\left(y_{i1}, y_{i2} \mid y_{i1}+y_{i2}, x_i\right) = \begin{cases} 1 & \text{if } (y_{i1}, y_{i2}) = (0,0) \text{ or } (1,1) \\ 1-\Lambda\left(\Delta x_{i2}'\beta_0\right) & \text{if } (y_{i1}, y_{i2}) = (1,0) \\ \Lambda\left(\Delta x_{i2}'\beta_0\right) & \text{if } (y_{i1}, y_{i2}) = (0,1). \end{cases} \tag{24}$$

Therefore, the log-likelihood conditioned on $y_{i1}+y_{i2}$ is given by[2]

$$L_c(\beta) = \sum_{i=1}^{N}\left\{d_{10i}\log\left[1-\Lambda\left(\Delta x_{i2}'\beta\right)\right] + d_{01i}\log\Lambda\left(\Delta x_{i2}'\beta\right)\right\} \tag{25}$$

and the score takes the form

$$\frac{\partial L_c(\beta)}{\partial\beta} = \sum_{i=1}^{N}\Delta x_{i2}\left\{d_{01i} - \Lambda\left(\Delta x_{i2}'\beta\right)1(y_{i1}+y_{i2}=1)\right\}. \tag{26}$$

[2] The contributions of $(0,0)$ or $(1,1)$ observations is zero.

4.2 Maximum Score Estimation

The previous technique crucially relied on the logit assumption. Manski (1987) considered a more general model of the form (1) in which the *cdf* of $-v_{it} \mid x_i, \eta_i$ was non-parametric and could depend on x_i and η_i in a time-invariant way. Namely, for all t and s

$$\Pr(-v_{it} \leq r \mid x_i, \eta_i) = \Pr(-v_{is} \leq r \mid x_i, \eta_i) = F(r \mid x_i, \eta_i), \tag{27}$$

so that $F(r \mid x_i, \eta_i)$ does not change with t but is otherwise unrestricted.

This assumption imposes stationarity and strict exogeneity, but allows for serial dependence in the errors v_{it}. It also allows for a certain kind of conditional heteroskedasticity, though not a very plausible one, since $Var(v_{it} \mid x_i, \eta_i)$ may depend on x_i but v_{it} is not allowed to be more sensitive to x_{it} than to other x's. Similarly if the expectations $E(v_{it} \mid x_i, \eta_i)$ exist, they may depend on x_i but not their first-differences $E(\Delta v_{it} \mid x_i, \eta_i) = 0$.

The time-invariance of F implies that for $T = 2$:[3]

$$med(y_{i2} - y_{i1} \mid x_i, y_{i1} + y_{i2} = 1) = \operatorname{sgn}(\Delta x'_{i2}\beta_0). \tag{28}$$

To see this note that, given $y_{i1} + y_{i2} = 1$, the difference $y_{i2} - y_{i1}$ can only equal 1 or -1. So the median will be one or the other depending on whether $\Pr(y_{i2} = 1, y_{i1} = 0 \mid x_i) \lesseqgtr \Pr(y_{i2} = 0, y_{i1} = 1 \mid x_i)$. Thus[4]

$$\begin{aligned} med(y_{i2} - y_{i1} \mid x_i, y_{i1} + y_{i2} = 1) &= \operatorname{sgn}[\Pr(y_{i2} = 1, y_{i1} = 0 \mid x_i) - \\ - \Pr(y_{i2} = 0, y_{i1} = 1 \mid x_i)] &= \operatorname{sgn}[\Pr(y_{i2} = 1 \mid x_i) - \Pr(y_{i1} = 1 \mid x_i)]. \end{aligned}$$

[3] The sign function is defined as

$$\operatorname{sgn}(u) = 1(u > 0) - 1(u < 0),$$

i.e. $\operatorname{sgn}(u) = -1$ if $u < 0$, $\operatorname{sgn}(u) = 0$ if $u = 0$ and $\operatorname{sgn}(u) = 1$ if $u > 0$.

[4] The second equality follows from

$$\begin{aligned} \Pr(y_{i2} = 1 \mid x_i) &= \Pr(y_{i2} = 1, y_{i1} = 0 \mid x_i) + \Pr(y_{i2} = 1, y_{i1} = 1 \mid x_i) \\ \Pr(y_{i1} = 1 \mid x_i) &= \Pr(y_{i2} = 0, y_{i1} = 1 \mid x_i) + \Pr(y_{i2} = 1, y_{i1} = 1 \mid x_i). \end{aligned}$$

Moreover, from the model's specification, i.e.

$$\begin{aligned}\Pr(y_{i1}=1\mid x_i,\eta_i) &= F(x_{i1}'\beta_0+\eta_i\mid x_i,\eta_i)\\ \Pr(y_{i2}=1\mid x_i,\eta_i) &= F(x_{i2}'\beta_0+\eta_i\mid x_i,\eta_i),\end{aligned}$$

and the monotonicity of F, we have that for any η_i (the constancy of F over time becomes crucial at this point):

$$\Pr(y_{i2}=1\mid x_i,\eta_i) \lesseqgtr \Pr(y_{i1}=1\mid x_i,\eta_i) \Leftrightarrow x_{i2}'\beta_0 \lesseqgtr x_{i1}'\beta_0.$$

Therefore, the implication also holds unconditionally relative to η_i:

$$\Pr(y_{i2}=1\mid x_i) \lesseqgtr \Pr(y_{i1}=1\mid x_i) \Leftrightarrow x_{i2}'\beta_0 \lesseqgtr x_{i1}'\beta_0.$$

or

$$\operatorname{sgn}[\Pr(y_{i2}=1\mid x_i)-\Pr(y_{i1}=1\mid x_i)] = \operatorname{sgn}(\Delta x_{i2}'\beta_0).$$

Manski showed that the true value of β_0 uniquely maximizes (up to scale) the expected agreement between the sign of $\Delta x_{i2}'\beta$ and that of Δy_{i2} conditioned on $y_{i1}+y_{i2}=1$. This identification result required an unbounded support for at least one of the explanatory variables with a non-zero coefficient. That is, letting $x_{it}'=(z_{it},w_{it}')$ and $\beta_0'=(\gamma_0,\alpha_0')$, the minimal requirement for identification is that z_{it} has unbounded support and $\gamma_0\neq 0$. Identification fails at $\gamma_0=0$, so that $\gamma_0=0$ is not a testable hypothesis. Manski's identification result implies that we can learn about the relative effects of the variables w_{it} under the maintained assumption that $\gamma_0\neq 0$.

Manski then proposed to estimate β_0 by selecting the value that matches the sign of $\Delta x_{i2}'\beta$ with that of Δy_{i2} for as many observations as possible in the subsample with $y_{i1}+y_{i2}=1$. The suggested estimator is

$$\widehat{\beta}=\arg\max_{\beta}\sum_{i=1}^{N}\operatorname{sgn}(\Delta x_{i2}'\beta)(y_{i2}-y_{i1}) \tag{29}$$

subject to the normalization $\| \beta \| = 1$.[5] This is the maximum score estimator applied to the observations with $y_{i1} + y_{i2} = 1$ (notice that the estimation criterion is unaffected by removing observations having $y_{i1} = y_{i2}$). It is consistent under the assumption that there is at least one unbounded continuous regressor, but it is not root-N consistent, and not asymptotically normal.

An alternative form of the score objective function is

$$S_N(\beta) = \sum_{i=1}^{N} \left\{ d_{10i} 1\left(\Delta x_{i2}' \beta < 0\right) + d_{01i} 1\left(\Delta x_{i2}' \beta \geq 0\right) \right\}. \tag{30}$$

The score $S_N(\beta)$ gives the number of correct predictions we would make if we predicted (y_{i1}, y_{i2}) to be $(0, 1)$ whenever $\Delta x_{i2}' \beta \geq 0$. In contrast, $\sum_{i=1}^{N} \operatorname{sgn}\left(\Delta x_{i2}' \beta\right) \Delta y_{i2}$ gives the number of successes minus the number of failures. Yet another form of the estimator suggested by the median regression interpretation is as the minimizer of the number of failures, which is given by

$$\frac{1}{2} \sum_{i=1}^{N} 1\left(y_{i1} \neq y_{i2}\right) \left| \Delta y_{i2} - \operatorname{sgn}\left(\Delta x_{i2}' \beta\right) \right|. \tag{31}$$

Smoothed Maximum Score It is possible to consider a smoothed version of the maximum score estimator along the lines of Horowitz (1992), which does have an asymptotic normal distribution, although the rate of convergence remains slower than root-N (Charlier, Melenberg and van Soest, 1995, and Kyriazidou, 1997).[6] The idea is to replace $S_N(\beta)$ with a smooth function $S_N^*(\beta)$ whose limit *a.s.* as $N \to \infty$ is the same as $S_N(\beta)$. This is of the form

$$S_N^*(\beta) = \sum_{i=1}^{N} \left\{ d_{10i} \left[1 - K\left(\Delta x_{i2}' \beta / \gamma_N\right)\right] + d_{01i} K\left(\Delta x_{i2}' \beta / \gamma_N\right) \right\} \tag{32}$$

[5] In the logit case the scale normalization is imposed through the variance of the logistic distribution. More generally, if F is a known distribution a priori, the scale normalization is determined by the form of F. Comparisons can be made by considering ratios of coefficients.

[6] Chamberlain (1986) showed that there is no root-N consistent estimator of β under the assumptions of Manski for his maximum score method.

where $K(.)$ is analogous to a *cdf* and γ_N is a sequence of positive numbers such that $\lim_{N\to\infty}\gamma_N = 0$. Notice the similarity between $S_N^*(\beta)$ and the conditional log likelihood for logit $L_c(\beta)$.

4.3 Random Effects

In general

$$\Pr(y_{i1},...,y_{iT} \mid x_i) = \int \Pr(y_{i1},...,y_{iT} \mid x_i,\eta_i)\, dG(\eta_i \mid x_i) \qquad (33)$$

where $G(\eta_i \mid x_i)$ is the *cdf* of $\eta_i \mid x_i$. The substantive model specifies $\Pr(y_{i1},...,y_{iT} \mid x_i,\eta_i)$, but only $\Pr(y_{i1},...,y_{iT} \mid x_i)$ has an empirical counterpart. For example, we may have specified

$$\Pr(y_{i1},...,y_{iT} \mid x_i,\eta_i) = \prod_{t=1}^{T} \Pr(y_{it} \mid x_i,\eta_i) = \prod_{t=1}^{T} F_{it}^{y_{it}} (1-F_{it})^{(1-y_{it})}.$$

In a fixed effects model we seek to make inferences about parameters in $\Pr(y_{i1},...,y_{iT} \mid x_i,\eta_i)$ without restricting the form of G. In a random effects model G is typically parametric or semiparametric, and the parameters of interest may or may not be identified with G unrestricted. Thus a fixed effects model can be regarded as a random effects model that leaves the distribution of the effects unrestricted.

The choice between fixed and random effects models often involves a trade-off between robustness in the specification of $\Pr(y_{i1},...,y_{iT} \mid x_i,\eta_i)$ and robustness in G, in the sense that achieving fixed-T identification with unrestricted G usually requires a more restrictive specification of $\Pr(y_{i1},...,y_{iT} \mid x_i,\eta_i)$.

Chamberlain (1980, 1984) considered a random effects model in which the effects are of the form

$$\eta_i = \mu(x_i) + \varepsilon_i \qquad (34)$$

and ε_i is independent of x_i. He also made the normality assumptions

$$v_{it} \mid x_i,\eta_i \sim \mathcal{N}(0,\omega_{tt}) \qquad (35)$$

$$\varepsilon_i \mid x_i \sim \mathcal{N}\left(0, \sigma_\eta^2\right), \tag{36}$$

which imply that

$$\Pr\left(y_{it} = 1 \mid x_i\right) = \Phi\left[\sigma_t^{-1}\left(x_{it}'\beta_0 + \mu\left(x_i\right)\right)\right]. \tag{37}$$

where $\sigma_t^2 = \sigma_\eta^2 + \omega_{tt}$ and $\Phi(.)$ is the standard normal *cdf.* In this model the v_{it} may be serially dependent and heteroskedastic over time.

Chamberlain assumed a linear specification $\mu(x_i) = \lambda_0 + x_i'\lambda$, and Newey (1994) generalized the model to a non-parametric $\mu(x_i)$. In the linear case, β_0, λ_0, λ, and the σ_t^2 can be estimated subject to the normalization $\sigma_1^2 = 1$ by combining the period-by-period probit likelihood functions (see Bover and Arellano, 1997, for a discussion of alternative estimators). In the semi-parametric case, Newey used the fact that

$$\sigma_t\Phi^{-1}\left[\Pr\left(y_{it} = 1 \mid x_i\right)\right] - \sigma_{t-1}\Phi^{-1}\left[\Pr\left(y_{i(t-1)} = 1 \mid x_i\right)\right] = \Delta x_{it}'\beta_0 \tag{38}$$

together with non-parametric estimates of the probabilities $\Pr(y_{it} = 1 \mid x_i)$ to obtain an estimator of β_0 and the relative scales. A further generalization of the model is to drop the normality assumptions and allow the distribution of the errors $\varepsilon_i + v_{it} \mid x_i$ to be unknown. This case has been considered by Chen (1998).

Another semi-parametric approach has been followed by Lee (1999). Under certain assumptions on the joint distribution of x_i and η_i, Lee proposed a maximum rank correlation-type estimator which is $\sqrt{N}$-consistent and asymptotically normal.

5 Identification Problems with Fixed T

It would be useful to know which models for $\Pr(y_{i1}, ..., y_{iT} \mid x_i, \eta_i)$ are identified without placing restrictions in the form of $G(\eta_i \mid x_i)$ (i.e. *fixed-effects identification with fixed T*) and which are not.

A model is given by a $2^T \times 1$ vector $p(x_i, \eta_i, \beta_0)$ with elements that specify the probabilities

$$\Pr\left((y_{i1}, ..., y_{iT}) = d_j \mid x_i, \eta_i\right) \;\; (j = 1, ..., 2^T) \tag{39}$$

where d_j is a $0-1$ sequence of order T. Let the true *cdf* of $\eta_i \mid x_i$ be $G_0(\eta \mid x)$. Identification will fail at β_0 if for all x in the support of x_i there is a *cdf* $G^*(\eta \mid x)$ and $\beta^* \neq \beta_0$ in the parameter space, such that

$$\int p(x, \eta, \beta_0)\, dG_0(\eta \mid x) = \int p(x, \eta, \beta^*)\, dG^*(\eta \mid x)\,. \tag{40}$$

If this is so, (β_0, G_0) and (β^*, G^*) give the same conditional distribution for $(y_{i1}, ..., y_{iT})$ given x_i. Therefore, they are observationally equivalent relative to such distribution.

Chamberlain (1992) studied the identification of a fixed effects binary choice model with $T = 2$. He considered the model

$$y_{it} = 1\left(x_{it}'\beta_0 + \eta_i + v_{it} \geq 0\right) \;\; (t = 1, 2)$$

together with the assumption that the $-v_{it}$ are independent of x_i, η_i and are i.i.d. over time with a known *cdf* F. The distribution F is strictly increasing on the whole line, with a bounded, continuous derivative. Moreover, we have the partitions $x_{it}' = (d_t, z_{it}')$ and $\beta_0' = (\alpha_0, \gamma_0')$, where d_t is a time dummy such that $d_1 = 0$ and $d_2 = 1$, and z_i is a continuous random vector with bounded support.

With these assumptions Chamberlain showed that if F is not logistic, then there is a value of α such that identification fails for all β_0 in a neighborhood of $(\alpha, 0)$. This seems puzzling since Manski (1987) proved identification under less restrictive assumptions. He required, however, the presence of an explanatory variable with unbounded support. Indeed, the difference between the identification result of Manski and the underidentification result of Chamberlain is due to the bounded support for the explanatory variables.

The line between identification and underidentification in this context is very subtle. Under Manski's assumptions identification will fail at $\beta_0' = (\alpha_0, 0)$ even if z_{it} has unbounded support, but there will be identification as long as a component of γ_0 is different from zero. Chamberlain shows that if z_{it} is bounded β_0 is underidentified not only when $\beta_0' = (\alpha_0, 0)$, but also for all β_0 in a neighborhood of $(\alpha_0, 0)$ for a certain value of α_0. So it seems to be a case of local underidentification at zero versus local underidentification in a neighborhood around zero.

The lesson from these findings is the fragility of fixed-T identification results and the special role of the logistic assumption. Chamberlain (1992) also showed that when the support of z_{it} is unbounded (so that identification holds to the exclusion of $\gamma_0 = 0$ from the parameter space) the information bound for β_0 is zero unless F is logistic. Thus, root-N consistent estimation is possible only for the logit model.

Chamberlain's proof can be sketched as follows. In his case $p(x, \eta, \beta_0)$ is

$$p(x, \eta, \beta_0) = \begin{pmatrix} (1 - F_1)(1 - F_2) \\ (1 - F_1) F_{i2} \\ F_1 (1 - F_2) \\ F_1 F_2 \end{pmatrix}$$

where $F_1 = F(z_1'\gamma_0 + \eta)$ and $F_2 = F(\alpha_0 + z_2'\gamma_0 + \eta)$.

Let $\beta^* = (\alpha, 0)$ and define the 4×4 matrix

$$H(x, \eta_1, ..., \eta_4, \beta^*) = [p(x, \eta_1, \beta^*), ..., p(x, \eta_4, \beta^*)] .$$

which does not vary with x when evaluated at β^*.

The proof proceeds by showing that unless $H(x, \eta_1, ..., \eta_4, \beta^*)$ is singular for every α and $\eta_1, ..., \eta_4$, there will be lack of identification for all β_0 in a neighborhood of some β^*. Next it is shown that $H(x, \eta_1, ..., \eta_4, \beta^*)$ can only be singular if F is logistic.

Suppose that $H(x, \eta_1, ..., \eta_4, \beta^*)$ is nonsingular for some α and $\eta_1, ..., \eta_4$. Since x is bounded, for $\beta_0 \neq \beta^*$ in a neighborhood of β^*, $H(x, \eta_1, ..., \eta_4, \beta_0)$

will also be nonsingular for *all* admissible values of x. We can now choose a *pmf* $\pi^* = (\pi_1^*, ..., \pi_4^*)'$, $\pi_j^* > 0$, $\sum_{j=1}^4 \pi_j^* = 1$ and define

$$\pi_0(x) = H(x, \eta_1, ..., \eta_4, \beta_0)^{-1} H(x, \eta_1, ..., \eta_4, \beta^*) \pi^*,$$

such that $\pi_{0j}(x) > 0$ for all admissible x. Moreover, since $\iota' H = \iota'$ where ι is a 4×1 vector of ones, we also have $\iota' H^{-1} = \iota'$ and $\iota' \pi_0(x) = 1$. Therefore,

$$\sum_{j=1}^4 p(x, \eta_j, \beta_0) \pi_{0j}(x) = \sum_{j=1}^4 p(x, \eta_j, \beta^*) \pi_j^*$$

which implies that β_0 cannot be distinguished from β^*.

The singularity of $H(x, \eta_1, ..., \eta_4, \beta^*)$ requires that

$$\psi_1 [1 - F(\eta)][1 - F(\alpha + \eta)] + \psi_2 [1 - F(\eta)] F(\alpha + \eta)$$

$$+ \psi_3 F(\eta)[1 - F(\alpha + \eta)] + \psi_4 F(\eta) F(\alpha + \eta) = 0$$

for all η and some scalars $\psi_1, ..., \psi_4$ that are not all zero. Taking limits as η tends to $\pm\infty$ gives $\psi_1 = \psi_4 = 0$. Thus we are left with

$$\psi_2 Q(\alpha + \eta) + \psi_3 Q(\eta) = 0$$

where $Q \equiv F/(1 - F)$. For $\eta = 0$ we obtain $\psi_3/\psi_2 = -Q(\alpha)/Q(0)$. Therefore the singularity of H requires that for all α and η we have

$$q(\alpha + \eta) = q(\alpha) + q(\eta) - q(0).$$

This can only happen if the log odd ratios $q \equiv \log Q$ are linear or equivalently if F is logistic.

6 Adjusting the Concentrated Likelihood

Cox and Reid (1987) considered the general problem of doing inference for a parameter of interest in the absence of knowledge about nuisance parameters. They proposed a first-order adjustment to the concentrated likelihood to take account of the estimation of the nuisance parameters (the *modified profile likelihood*). Their formulation required information orthogonality between the two types of parameters. That is, that the expected information matrix be block diagonal between the parameters of interest and the nuisance parameters; something that may be achieved by transformation of the latter (Cox and Reid explained how to construct orthogonal parameters). A discussion of orthogonality in the context of panel data models and a Bayesian perspective have been given by Lancaster (1997, 2000). The nature of the adjustment in a fixed effects model and some examples are also discussed in Cox and Reid (1992).

6.1 Orthogonalization

Let $\ell_i(\beta, \eta_i)$ be the log-likelihood for unit i (conditional on x_i and η_i). A strong form of orthogonality arises when for some parameterization of η_i we have

$$\ell_i(\beta, \eta_i) = \ell_{1i}(\beta) + \ell_{2i}(\eta_i), \tag{41}$$

for in this case the MLE of η_i for given β does not depend on $\beta, \widehat{\eta}_i(\beta) = \widehat{\eta}_i$. The implication is that the MLE of β is unaffected by lack of knowledge of η_i. In this case $\partial^2 \ell_i(\beta, \eta_i) / \partial\beta\partial\eta_i = 0$ for all i. Unfortunately, such factorization does not hold for binary choice models. In contrast, information orthogonality just requires the cross derivatives to be zero on average.

Suppose that a reparameterization is made from (β, η_i) to (β, λ_i) chosen so that β and λ_i are information orthogonal. Thus $\eta_i = \eta(\beta, \lambda_i)$ is chosen such that the reparameterized log likelihood

$$\ell_i^*(\beta, \lambda_i) = \ell_i(\beta, \eta(\beta, \lambda_i)) \tag{42}$$

satisfies (at true values):

$$E\left(\frac{\partial^2 \ell_i^*(\beta_0, \lambda_i)}{\partial\beta\partial\lambda_i} \mid x_i, \eta_i\right) = 0. \tag{43}$$

Since we have

$$\frac{\partial \ell_i^*}{\partial \beta} = \frac{\partial \ell_i}{\partial \beta} + \frac{\partial \eta_i}{\partial \beta}\frac{\partial \ell_i}{\partial \eta_i} \tag{44}$$

and[7]

$$E\left(\frac{\partial^2 \ell_i^*}{\partial \beta \partial \lambda_i} \mid x_i, \eta_i\right) = \frac{\partial \eta_i}{\partial \lambda_i} E\left(\frac{\partial^2 \ell_i}{\partial \beta \partial \eta_i} \mid x_i, \eta_i\right) + \frac{\partial \eta_i}{\partial \lambda_i}\frac{\partial \eta_i}{\partial \beta} E\left(\frac{\partial^2 \ell_i}{\partial \eta_i^2} \mid x_i, \eta_i\right), \tag{45}$$

following Cox and Reid (1987) and Lancaster (1997), the function $\eta(\beta, \lambda_i)$ must satisfy the partial differential equations

$$\frac{\partial \eta_i}{\partial \beta} = -E\left(\frac{\partial^2 \ell_i}{\partial \beta \partial \eta_i} \mid x_i, \eta_i\right) / E\left(\frac{\partial^2 \ell_i}{\partial \eta_i^2} \mid x_i, \eta_i\right). \tag{46}$$

Orthogonal Effects in Binary Choice Let us now consider the form of information orthogonal fixed effects for model (1)-(2). These have been obtained by Lancaster (1998, 2000). For binary choice we have

$$E\left(\frac{\partial^2 \ell_i(\beta_0, \eta_i)}{\partial \beta \partial \eta_i} \mid x_i, \eta_i\right) = -\sum_{t=1}^{T} h(x_{it}'\beta_0 + \eta_i) x_{it} \tag{47}$$

$$E\left(\frac{\partial^2 \ell_i(\beta_0, \eta_i)}{\partial \eta_i^2} \mid x_i, \eta_i\right) = -\sum_{t=1}^{T} h(x_{it}'\beta_0 + \eta_i) \tag{48}$$

where

$$h(r) = \frac{f(r)^2}{F(r)[1 - F(r)]}. \tag{49}$$

Since in general (47) is different from zero, β and η_i are not information orthogonal. In view of (46), an orthogonal transformation of the effects will satisfy

$$\frac{\partial \eta_i}{\partial \beta} = -\frac{1}{\sum_{t=1}^{T} h_{it}} \sum_{t=1}^{T} h_{it} x_{it} \tag{50}$$

where $h_{it} = h(x_{it}'\beta + \eta_i)$.

[7]Note that there is a term that vanishes: $(\partial^2 \eta_i / \partial \beta \partial \lambda_i) E(\partial \ell_i / \partial \eta_i \mid x_i, \eta_i) = 0$.

Moreover, letting $\phi(r) = h'(r)$ and $\phi_{it} = \phi(x'_{it}\beta + \eta_i)$, since

$$\frac{\partial^2 \eta_i}{\partial \beta \partial \lambda_i} = -\frac{\partial \eta_i}{\partial \lambda_i} \left[\frac{1}{\sum_{t=1}^T h_{it}} \sum_{t=1}^T \phi_{it} \left(x_{it} + \frac{\partial \eta_i}{\partial \beta} \right) \right]$$

and

$$\frac{\partial^2 \eta_i}{\partial \beta \partial \lambda_i} / \frac{\partial \eta_i}{\partial \lambda_i} = \frac{\partial}{\partial \beta} \log \left| \frac{\partial \eta_i}{\partial \lambda_i} \right|, \tag{51}$$

it turns out that

$$\frac{\partial \eta_i}{\partial \lambda_i} = \frac{1}{\sum_{t=1}^T h_{it}}. \tag{52}$$

Hence, Lancaster's orthogonal reparameterization is

$$\lambda_i = \sum_{t=1}^T \int_{-\infty}^{x'_{it}\beta + \eta_i} h(r)\, dr. \tag{53}$$

When $F(r)$ is the logistic distribution $h(r)$ coincides with the logistic density, so that an orthogonal effect for the logit model is

$$\lambda_i = \sum_{t=1}^T \Lambda(x'_{it}\beta + \eta_i). \tag{54}$$

6.2 Modified Profile Likelihood

The modified profile log likelihood function of Cox and Reid (1987) can be written as

$$L_M(\beta) = \sum_i \ell_{Mi}(\beta)$$

and

$$\ell_{Mi}(\beta) = \ell_i^* \left(\beta, \widehat{\lambda}_i(\beta) \right) - \frac{1}{2} \log \left[-d^*_{\lambda\lambda i} \left(\beta, \widehat{\lambda}_i(\beta) \right) \right], \tag{55}$$

where $\widehat{\lambda}_i(\beta)$ is the MLE of λ_i for given β, and $d^*_{\lambda\lambda i}(\beta, \lambda_i) = \partial^2 \ell_i^* / \partial \lambda_i^2$. Intuitively, the role of the second term is to penalize values of β for which the information about the effects is relatively large.

An individual's modified score is of the form

$$d_{Mi}(\beta) = d_{\mathrm{Ci}}(\beta) - \frac{1}{2d^{*}_{\lambda\lambda i}\left(\beta, \widehat{\lambda}_i(\beta)\right)}\left(d^{*}_{\lambda\lambda\beta i}\left(\beta, \widehat{\lambda}_i(\beta)\right) + d^{*}_{\lambda\lambda\lambda i}\left(\beta, \widehat{\lambda}_i(\beta)\right)\frac{\partial \widehat{\lambda}_i(\beta)}{\partial \beta}\right) \tag{56}$$

where $d_{\mathrm{Ci}}(\beta)$ is the standard score from the concentrated likelihood, $d^{*}_{\lambda\lambda\beta i}(\beta, \lambda_i) = \partial^3 \ell^{*}_i / \partial \lambda_i^2 \partial \beta$ and $d^{*}_{\lambda\lambda\lambda i}(\beta, \lambda_i) = \partial^3 \ell^{*}_i / \partial \lambda_i^3$.

The function (55) was derived by Cox and Reid as an approximation to the conditional likelihood given $\widehat{\lambda}_i(\beta)$. Their approach was motivated by the fact that in an exponential family model, it is optimal to condition on sufficient statistics for the nuisance parameters, and these can be regarded as the MLE of nuisance parameters chosen in a form to be orthogonal to the parameters of interest. For more general problems the idea was to derive a concentrated likelihood for β conditioned on the MLE $\widehat{\lambda}_i(\beta)$, having ensured via orthogonality that $\widehat{\lambda}_i(\beta)$ changes slowly with β.

Another motivation for using (55) is that the corresponding expected score has a bias of a smaller order of magnitude than the standard ML score (cf. Liang, 1987, McCullagh and Tibshirani, 1990, and Ferguson, Reid, and Cox, 1991). Seen in this way, the objective of the adjustment is to center the concentrated score function to achieve consistency up to a certain order of magnitude in T. Specifically, while the difference between the score with known λ_i and the concentrated score is in general of order $O_p(1)$, the corresponding difference with the modified concentrated score is of order $O_p\left(T^{-1/2}\right)$ (see Appendix). This leads to a bias of order $O\left(T^{-1}\right)$ in the expected modified score, as opposed to $O(1)$ in the concentrated score without modification.

The Adjustment in Terms of the Original Parameterization Cox and Reid's motivation for modifying the concentrated likelihood relied on the orthogonality between common and nuisance parameters. Nevertheless, the *mpl* function (55) can be expressed in terms of the original parameterization.

Firstly, note that because of the invariance of MLE $\widehat{\eta}_i(\beta) = \eta(\beta, \widehat{\lambda}_i(\beta))$ and

$$\ell_i^*\left(\beta, \widehat{\lambda}_i(\beta)\right) = \ell_i(\beta, \widehat{\eta}_i(\beta)). \tag{57}$$

Next, the term $d_{\lambda\lambda i}^*\left(\beta, \widehat{\lambda}_i(\beta)\right)$ can be calculated as the product of the Fisher information in the (β, η_i) parameterization and the square of the Jacobian of the transformation from (β, η_i) to (β, λ_i) (Cox and Reid, 1987, p. 10). That is, since the second derivatives of ℓ_i^* and ℓ_i are related by the expression

$$\frac{\partial^2 \ell_i^*}{\partial \lambda_i^2} = \frac{\partial^2 \ell_i}{\partial \eta_i^2}\left(\frac{\partial \eta_i}{\partial \lambda_i}\right)^2 + \frac{\partial \ell_i}{\partial \eta_i}\left(\frac{\partial^2 \eta_i}{\partial \lambda_i^2}\right),$$

and $\partial \ell_i / \partial \eta_i$ vanishes at $\widehat{\eta}_i(\beta)$, letting $d_{\eta\eta i}(\beta, \eta_i) = \partial^2 \ell_i / \partial \eta_i^2$ we have

$$d_{\lambda\lambda i}^*\left(\beta, \widehat{\lambda}_i(\beta)\right) = d_{\eta\eta i}(\beta, \widehat{\eta}_i(\beta))\left(\frac{\partial \eta_i}{\partial \lambda_i}|_{\lambda_i = \widehat{\lambda}_i(\beta)}\right)^2. \tag{58}$$

Thus, the *mpl* can be written as

$$\ell_{Mi}(\beta) = \ell_i(\beta, \widehat{\eta}_i(\beta)) - \frac{1}{2}\log\left[-d_{\eta\eta i}(\beta, \widehat{\eta}_i(\beta))\right] + \log\left(\frac{\partial \lambda_i}{\partial \eta_i}|_{\eta_i = \widehat{\eta}_i(\beta)}\right). \tag{59}$$

Finally, in view of (46) and (51), the derivative with respect to β of the Jacobian term (the required term for the modified score) can be expressed as

$$\frac{\partial}{\partial \beta}\log\left|\frac{\partial \lambda_i}{\partial \eta_i}\right| = \frac{\partial}{\partial \eta_i} q_i(\beta, \eta_i), \tag{60}$$

where $q_i(\beta, \eta_i) = -\kappa_{\beta\eta i}(\beta, \eta_i) / \kappa_{\eta\eta i}(\beta, \eta_i)$ and

$$\kappa_{\beta\eta i}(\beta_0, \eta_i) = E\left[\frac{1}{T} d_{\beta\eta i}(\beta_0, \eta_i) \mid x_i, \eta_i\right] \tag{61}$$

$$\kappa_{\eta\eta i}(\beta_0, \eta_i) = E\left[\frac{1}{T} d_{\eta\eta i}(\beta_0, \eta_i) \mid x_i, \eta_i\right]. \tag{62}$$

Modified Profile Likelihood for Binary Choice Replacing (52) in (59) we have

$$\ell_{Mi}(\beta) = \ell_i(\beta, \widehat{\eta}_i(\beta)) - \frac{1}{2}\log\left[-d_{\eta\eta i}(\beta, \widehat{\eta}_i(\beta))\right] + \log\left(\sum_{t=1}^{T} \widehat{h}_{it}(\beta)\right) \tag{63}$$

where $\widehat{h}_{it}(\beta) = h(x_{it}'\beta + \widehat{\eta}_i(\beta))$,

$$\ell_i(\beta, \eta_i) = \sum_{t=1}^{T} \{y_{it} \log F_{it} + (1 - y_{it}) \log(1 - F_{it})\}$$

and

$$d_{\eta\eta i}(\beta, \eta_i) = -\sum_{t=1}^{T} [h_{it} - \rho_{it}(y_{it} - F_{it})] \tag{64}$$

where $\rho_{it} = \rho(x_{it}'\beta + \eta_i)$ and

$$\rho(r) = \frac{f'(r) - h(r)[1 - 2F(r)]}{F(r)[1 - F(r)]}. \tag{65}$$

For logit, the MLE $\widehat{\lambda}_i(\beta)$ for given β solves

$$\widehat{\lambda}_i(\beta) = \sum_{t=1}^{T} \Lambda(x_{it}'\beta + \widehat{\eta}_i(\beta)) = \sum_{t=1}^{T} y_{it} \tag{66}$$

so that it does not vary with β. Therefore, the likelihood conditioned on $\widehat{\lambda}_i(\beta)$ coincides with the conditional logit likelihood given a sufficient statistic for the fixed effect discussed in Section 4.1.

For the logistic distribution $\rho(r) = 0$. The modified profile likelihood (mpl) for logit is therefore

$$\ell_{Mi}(\beta) = \ell_i(\beta, \widehat{\eta}_i(\beta)) + \frac{1}{2} \log\left(\sum_{t=1}^{T} f_\Lambda(x_{it}'\beta + \widehat{\eta}_i(\beta))\right) \tag{67}$$

where $f_\Lambda(r) = \Lambda(r)[1 - \Lambda(r)]$ is the logistic density and $\ell_{Mi}(\beta)$ is defined for observations such that $\sum_{t=1}^{T} y_{it}$ is not zero or T.[8]

Comparisons for the Two-Period Logit Model The *mpl* for logit (67) differs from Andersen's conditional likelihood, and the estimator $\widehat{\beta}_{MML}$ that maximizes the *mpl* is inconsistent for fixed T. Pursuing the example in

[8]If $\widehat{\eta}_i(\beta) \to \pm\infty$, then $\log\left(\sum_{t=1}^{T} f_\Lambda(x_{it}'\beta + \widehat{\eta}_i(\beta))\right)$ tends to $-\infty$ for any β. So observations for individuals that never change state are uninformative about β.

Section 3, we compare the large-N biases of ML and MML for $T = 2$ and $\Delta x_{i2} = 1$. Thus we are assessing the value of the large-T adjustment in (67) when $T = 2$.

When $T = 2$, for individuals who change state $\widehat{\eta}_i(\beta) = -\beta/2$ so that the second term in (67) becomes

$$\frac{1}{2}\log\left[f_\Lambda(-\beta/2) + f_\Lambda(\beta/2)\right]. \tag{68}$$

Collecting terms and ignoring constants, the modified profile log-likelihood takes the form

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\ell_{Mi}(\beta) &= \frac{1}{N}\sum_{i=1}^{N}\{2d_{10i}\log\left[1-\Lambda(\beta/2)\right] + 2d_{01i}\log\Lambda(\beta/2) \\
&\quad + (d_{10i}+d_{01i})\frac{1}{2}\left(\log\Lambda(\beta/2) + \log\left[1-\Lambda(\beta/2)\right]\right)\} \\
&\propto \frac{1}{N}\sum_{i=1}^{N}\{(5d_{10i}+d_{01i})\log\left[1-\Lambda(\beta/2)\right] + (5d_{01i}+d_{10i})\log\Lambda(\beta/2)\} \\
&\propto (5-4\widehat{p})\log\left[1-\Lambda(\beta/2)\right] + (4\widehat{p}+1)\log\Lambda(\beta/2)
\end{aligned} \tag{69}$$

where $d_{10i} = 1(y_{i1} = 1, y_{i2} = 0)$, $d_{01i} = 1(y_{i1} = 0, y_{i2} = 1)$ and $\widehat{p}$ is as defined in (20). This is maximized at

$$\widehat{\beta}_{MML} = 2\Lambda^{-1}\left(\frac{4\widehat{p}+1}{6}\right) = 2\log\left(\frac{4\widehat{p}+1}{5-4\widehat{p}}\right). \tag{70}$$

Therefore,

$$p\lim_{N\to\infty}\widehat{\beta}_{MML} = 2\log\left(\frac{4p_0+1}{5-4p_0}\right) = 2\log\left(\frac{4\Lambda(\beta_0)+1}{5-4\Lambda(\beta_0)}\right). \tag{71}$$

Figure 1 shows the probability limits of MML for positive values of β_0, together with those of ML (the $2\beta_0$ line) and conditional ML (the 45° line) for comparisons.[9] In this example the adjustment produces a surprisingly

[9] See McCullagh and Tibshirani (1990, pp. 337-8) for a similar exercise using different adjusted likelihood functions.

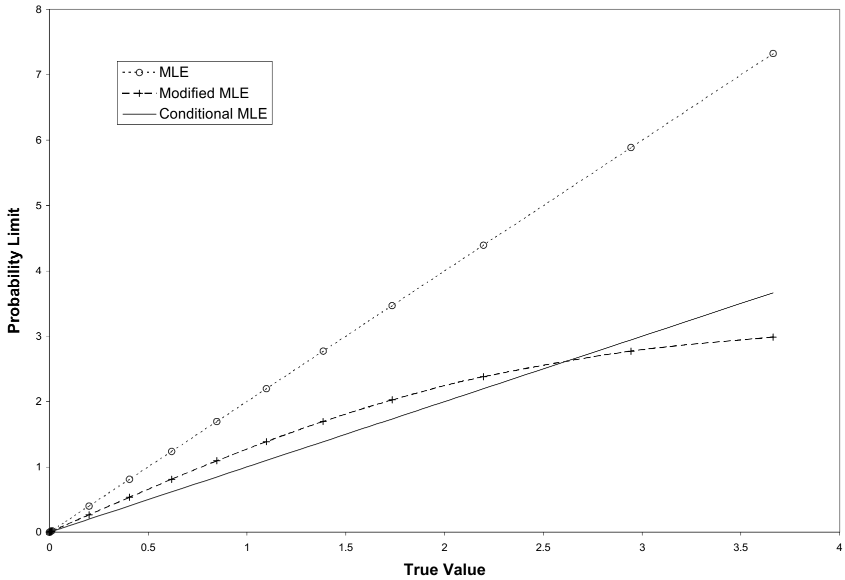

Figure 1: Probability limits for a logit model with $T = 2$

good improvement given that we are relying on a large T argument with $T = 2$. For example, for $p_0 = 0.65$, we have $\beta_0 = 0.62$, $\beta_{ML} = 1.24$ and $\beta_{MML} = 0.81$. Since the MML biases are of order $O(1/T^2)$, the result suggests that, although the biases are not negligible for $T = 2$, they may be so for values of T as small as 5 or 6.

7 N and T Asymptotics

The panel data literature has probably overemphasized the quest for fixed-T large-N consistent estimation of non-linear models with fixed effects. We have already seen the difficulties that arise in trying to obtain a root-N consistent estimator for a simple static fixed effects probit model. Not surprisingly, the difficulties become even more serious for dynamic binary choice models. In a sense, insisting on fixed T consistency has similarities with (and may be as restrictive as) requiring exactly unbiased estimation in non-linear models. Panels with $T = 2$ are more common in theoretical discussions than in econometric practice. For a micro panel with 7 or 8 time series observations, whether estimation biases are of order $O(1/T)$ or $O(1/T^2)$ may make all the difference. So it seems useful to consider a wider class of estimation methods than those providing fixed-T consistency, and assess their merits with regard to alternative N and T asymptotic plans. There are multiple possible asymptotic formulations, and it is a matter of judgement to decide which one provides the best approximation for the sample sizes involved in a given application.

Here we consider the asymptotic properties of the estimators that maximize the concentrated likelihood (ML) and the modified concentrated likelihood (MML) when T/N tends to a constant (related results for autoregressive models are in Alvarez and Arellano, 1998, and Hahn, 1998).

Consistency The ML estimator of β can be shown to be consistent as $T \to \infty$ regardless of N using the arguments and the consistency theorem in Amemiya (1985, pp. 270-72). The consistency of MML follows from noting that the concentrated likelihood and the *mpl* converge to the same objective function uniformly in probability as $T \to \infty$.

Letting $\widehat{\rho}_{it}(\beta) = \rho(x'_{it}\beta + \widehat{\eta}_i(\beta))$ and $\widehat{F}_{it}(\beta) = F(x'_{it}\beta + \widehat{\eta}_i(\beta))$, from (63) we have

$$\begin{aligned} p\lim_{T\to\infty}\frac{1}{T}\ell_{Mi}(\beta) &= p\lim_{T\to\infty}\frac{1}{T}\ell_i(\beta,\widehat{\eta}_i(\beta)) + p\lim_{T\to\infty}\frac{1}{T}\log\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{h}_{it}(\beta)\right) \qquad (72)\\ &\quad - p\lim_{T\to\infty}\frac{1}{T}\log\left(\frac{1}{T}\sum_{t=1}^{T}\left[\widehat{h}_{it}(\beta) - \widehat{\rho}_{it}(\beta)\left(y_{it} - \widehat{F}_{it}(\beta)\right)\right]\right)^{1/2}, \end{aligned}$$

where the convergence is uniform in β in a neighborhood of β_0, and the last two terms vanish.

Asymptotic Normality When $T/N \to c$, $0 < c < \infty$, both ML and MML are asymptotically normal but, unlike MML, the ML estimator has a bias in the asymptotic distribution. An informal calculation of the terms arising in the asymptotic distributions is given in the Appendix. The results are as follows:

$$\left(H_{NT}' V_{NT}^{-1} H_{NT}\right)^{1/2} \sqrt{NT} \left(\widehat{\beta}_{ML} - \beta_0 + \frac{1}{T} H_{NT}^{-1} b_N\right) \xrightarrow{d} \mathcal{N}(0, I) \tag{73}$$

$$\left(H_{NT}^{\dagger\prime} V_{NT}^{-1} H_{NT}^{\dagger}\right)^{1/2} \sqrt{NT} \left(\widehat{\beta}_{MML} - \beta_0\right) \xrightarrow{d} \mathcal{N}(0, I). \tag{74}$$

where $\kappa^*_{\beta\lambda\lambda i} = E\left[T^{-1} d^*_{\beta\lambda\lambda i}(\beta_0, \lambda_{i0}) \mid x_i, \lambda_i\right]$, $\kappa^*_{\lambda\lambda i} = E\left[T^{-1} d^*_{\lambda\lambda i}(\beta_0, \lambda_{i0}) \mid x_i, \lambda_i\right]$,

$$b_N = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\kappa^*_{\beta\lambda\lambda i}}{2\kappa^*_{\lambda\lambda i}}\right), \tag{75}$$

$$V_{NT} = \frac{1}{NT} \sum_{i=1}^{N} d^*_{\beta i}(\beta_0, \lambda_{i0}) \, d^*_{\beta i}(\beta_0, \lambda_{i0})', \tag{76}$$

$$H_{NT} = \frac{1}{NT} \sum_{i=1}^{N} \frac{\partial}{\partial \beta'} d^*_{\beta i}\left(\beta_0, \widehat{\lambda}_i(\beta_0)\right), \tag{77}$$

and

$$H_{NT}^{\dagger} = \frac{1}{NT} \sum_{i=1}^{N} \frac{\partial}{\partial \beta'} d_{Mi}(\beta_0). \tag{78}$$

Thus, the asymptotic distribution of the ML estimator will contain a bias term unless $\kappa^*_{\beta\lambda\lambda i} = 0$.

8 Concluding Remarks

In this paper we have considered ML and modified ML estimators, but the estimation problem can be put more generally in terms of moment conditions in a GMM framework. Fixed-T consistent estimators rely on exactly unbiased moment conditions. When T/N tends to a constant, a GMM estimator from moment conditions with a $O(1/T)$ bias will typically exhibit a bias in the asymptotic distribution, but not if the estimator is based on moment conditions with a $O(1/T^2)$ bias. Thus, in the context of binary choice and other non-linear microeconometric models, a search for optimal orthogonality conditions that are unbiased to order $O(1/T^2)$ or greater seems a useful research agenda.

But do these biases really matter? Heckman (1981) reported a Monte Carlo experiment for ML estimation of a probit model with strictly exogenous variables and fixed effects, $T = 8$ and $N = 100$. Using a random effects estimator as a benchmark, he concluded that the MLE of the common parameters (jointly estimated with the effects) performed well. According to this, it would seem that even for fairly small panels there is not much to be gained from the use of fixed-T unbiased or approximately unbiased orthogonality conditions. For models with only strictly exogenous explanatory variables this may well be the case. But these are models that are found to be too restrictive in many applications.

When modelling panel data, state dependence, predetermined regressors, and serial correlation often matter. Heckman (1981) found that when a lagged dependent variable was included the ML probit estimator performed badly. This is not surprising since similar problems occur with linear autoregressive models. The difference is that while standard tools are available in the literature that ensure fixed T consistency for linear dynamic models, very little is known for dynamic binary choice.[10] This is therefore a promising area of application of asymptotic arguments to both the construction of estimating equations and useful approximations to sampling distributions.

[10] See Keane (1994), Arellano and Carrasco (1996), Magnac (1997), Hyslop (1999), Honoré and Kyriazidou (2000), Honoré and Lewbel (2000), and Arellano and Honoré (2000) for a survey and more references.

References

[1] Alvarez, J. and M. Arellano (1998): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators", Working Paper 9808, CEMFI, Madrid.

[2] Amemiya, T. (1985): *Advanced Econometrics*, Basil Blackwell.

[3] Andersen, E. B. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society*, Series B, 32, 283-301.

[4] Andersen, E. B. (1973): *Conditional Inference and Models for Measuring*, Mentalhygiejnisk Forsknings Institut, Copenhagen.

[5] Arellano, M. and R. Carrasco (1996): "Binary Choice Panel Data Models with Predetermined Variables," Working Paper 9618, CEMFI, Madrid.

[6] Arellano, M. and B. Honoré (2000): "Panel Data Models: Some Recent Developments", in J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, vol. 5, North Holland, forthcoming.

[7] Bover, O. and M. Arellano (1997): "Estimating Dynamic Limited Dependent Variable Models from Panel Data", *Investigaciones Económicas*, 21, 141-165.

[8] Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data", *Review of Economic Studies*, 47, 225-238.

[9] Chamberlain, G. (1984): "Panel Data", in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics*, vol. 2, Elsevier Science, Amsterdam.

[10] Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring", *Journal of Econometrics*, 32, 189-218.

[11] Chamberlain (1992): "Binary Response Models for Panel Data: Identification and Information", unpublished manuscript, Department of Economics, Harvard University.

[12] Charlier, E., B. Melenberg, and A. van Soest (1995): "A Smoothed Maximum Score Estimator for the Binary Choice Panel Data Model and an Application to Labour Force Participation", *Statistica Neerlandica*, 49, 324–342.

[13] Chen, S. (1998): "Root-N Consistent Estimation of a Panel Data Sample Selection Model", unpublished manuscript, The Hong Kong University of Science and Technology.

[14] Cox, D. R. and N. Reid (1987): "Parameter Orthogonality and Approximate Conditional Inference" (with discussion), *Journal of the Royal Statistical Society*, Series B, 49, 1-39.

[15] Cox, D. R. and N. Reid (1992): "A Note on the Difference between Profile and Modified Profile Likelihood", *Biometrika*, 79, 408-411.

[16] Ferguson, H., N. Reid, and D. R. Cox (1991): "Estimating Equations from Modified Profile Likelihood", in Godambe, V. P. (ed.), *Estimating Functions*, Oxford University Press.

[17] Hahn, J. (1998): "Asymptotically Unbiased Inference of Dynamic Panel Data Model with Fixed Effects When Both n and T are Large", unpublished manuscript, University of Pennsylvania.

[18] Heckman, J. J. (1981): "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time–Discrete Data Stochastic Process" in C. F. Manski and D. McFadden (eds.),

Structural Analysis of Discrete Data with Econometric Applications, MIT Press.

[19] Honoré, B. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables", *Econometrica*, 68, 839-874.

[20] Honoré, B. and A. Lewbel (2000): "Semiparametric Binary Choice Panel Data Models without Strictly Exogeneous Regressors", unpublished manuscript, Department of Economics, Boston College.

[21] Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model", *Econometrica*, 60, 505-531.

[22] Hsiao, C. (1986): *Analysis of Panel Data*, Cambridge University Press.

[23] Hyslop, D. R. (1999): "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women", *Econometrica*, 67, 1255-1294.

[24] Keane, M. (1994): "A Computationally Practical Simulation Estimator for Panel Data", *Econometrica*, 62, 95-116.

[25] Kyriazidou, E. (1997): "Estimation of a Panel Data Sample Selection Model", *Econometrica*, 65, 1335–1364.

[26] Lancaster, T. (1997): "Orthogonal Parameters and Panel Data", Working Paper No. 97-32, Department of Economics, Brown University.

[27] Lancaster, T. (1998): "Panel Binary Choice with Fixed Effects", unpublished manuscript, Department of Economics, Brown University.

[28] Lancaster, T. (2000): "The Incidental Parameter Problem Since 1948", *Journal of Econometrics*, 95, 391-413.

[29] Lee, M.–J. (1999): "A Root–N Consistent Semiparametric Estimator for Related-Effect Binary Response Panel Data", *Econometrica*, 67, 427–434.

[30] Liang, K.Y. (1987): "Estimating Functions and Approximate Conditional Likelihood", *Biometrika*, 74, 695-702.

[31] Magnac, T. (1997): "State Dependence and Heterogeneity in Youth Employment Histories", Working Paper, INRA and CREST, Paris.

[32] McCullagh, P. and R. Tibshirani (1990): "A Simple Method for the Adjustment of Profile Likelihoods", *Journal of the Royal Statistical Society*, Series B, 52, 325-344.

[33] Manski, C. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data", *Econometrica*, 55, 357-362.

[34] Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators", *Econometrica*, 62, 1349-1382.

[35] Neyman, J. and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations", *Econometrica*, 16, 1–32.

[36] Rasch, G. (1960): *Probabilistic Models for Some Intelligence and Attainment Tests*, Denmarks Pædagogiske Institut, Copenhagen.

[37] Rasch, G. (1961): "On the General Laws and the Meaning of Measurement in Psychology," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, University of California Press, Berkeley and Los Angeles.

Appendix

Expansion for the Score of the Concentrated Likelihood Let us consider a second order expansion of the score of the concentrated likelihood around the true value of the orthogonal effect.

The log likelihood is $\ell_i^*(\beta,\lambda_i)$; its vector of partial derivatives with respect to β is $d_{\beta i}^*(\beta,\lambda_i) = \partial\ell_i^*(\beta,\lambda_i)/\partial\beta$; the concentrated likelihood is $\ell_i^*\left(\beta,\widehat{\lambda}_i(\beta)\right)$ and its score is given by $d_{\beta i}^*\left(\beta,\widehat{\lambda}_i(\beta)\right)$. An approximation at β_0 around the true value λ_{i0} is

$$\begin{aligned} d_{\beta i}^*\left(\beta_0,\widehat{\lambda}_i(\beta_0)\right) &= d_{\beta i}^*(\beta_0,\lambda_{i0}) + d_{\beta\lambda i}^*(\beta_0,\lambda_{i0})\left(\widehat{\lambda}_i(\beta_0)-\lambda_{i0}\right) \\ &\quad +\frac{1}{2}d_{\beta\lambda\lambda i}^*(\beta_0,\lambda_{i0})\left(\widehat{\lambda}_i(\beta_0)-\lambda_{i0}\right)^2 + O_p\left(T^{-1/2}\right) \end{aligned} \quad \text{(A1)}$$

where $d_{\beta\lambda i}^*(\beta,\lambda_i) = \partial^2\ell_i^*(\beta,\lambda_i)/\partial\beta\partial\lambda_i$ and $d_{\beta\lambda\lambda i}^*(\beta_0,\lambda_{i0}) = \partial^3\ell_i^*(\beta,\lambda_i)/\partial\beta\partial\lambda_i^2$. In general, the first three terms are $O_p\left(T^{1/2}\right)$, $O_p\left(T^{1/2}\right)$, and $O_p(1)$, but because of orthogonality $d_{\beta\lambda i}^*(\beta_0,\lambda_{i0})$ is $O_p\left(\sqrt{T}\right)$ as opposed to $O_p(T)$.[11]

Expansion for $\widehat{\lambda}_i(\beta_0)-\lambda_{i0}$ Letting $d_{\lambda i}^*(\beta,\lambda_i) = \partial\ell_i^*(\beta,\lambda_i)/\partial\lambda_i$, the estimator $\widehat{\lambda}_i(\beta_0)$ solves $d_{\lambda i}^*\left(\beta_0,\widehat{\lambda}_i(\beta_0)\right)=0$. Let us also introduce notation for the terms:

$$\begin{aligned} \kappa_{\lambda\lambda i}^* &\equiv \kappa_{\lambda\lambda}^*(\beta_0,\lambda_{i0}) = E\left[\frac{1}{T}d_{\lambda\lambda i}^*(\beta_0,\lambda_{i0}) \mid x_i,\lambda_i\right] \\ \kappa_{\beta\lambda\lambda i}^* &\equiv \kappa_{\beta\lambda\lambda}^*(\beta_0,\lambda_{i0}) = E\left[\frac{1}{T}d_{\beta\lambda\lambda i}^*(\beta_0,\lambda_{i0}) \mid x_i,\lambda_i\right] \end{aligned}$$

Note that $\kappa_{\lambda\lambda i}^*$ and $\kappa_{\beta\lambda\lambda i}^*$ are individual specific because they depend on λ_{i0}, but they do not depend on the y's.[12] Moreover, from the information matrix identity

$$E\left[\frac{1}{T}d_{\lambda i}^*(\beta_0,\lambda_{i0})\,d_{\lambda i}^*(\beta_0,\lambda_{i0}) \mid x_i,\lambda_i\right] = -\kappa_{\lambda\lambda i}^*.$$

[11] Since $\sqrt{T}\left[\frac{1}{T}d_{\beta\lambda i}^*(\beta_0,\lambda_{i0})-0\right] = O_p(1)$, we have $d_{\beta\lambda i}^*(\beta_0,\lambda_{i0}) = O_p\left(\sqrt{T}\right)$.

[12] Also $\frac{1}{T}d_{\lambda\lambda i}^*(\beta_0,\lambda_{i0}) = \kappa_{\lambda\lambda}^*(\beta_0,\lambda_{i0}) + O_p\left(\frac{1}{\sqrt{T}}\right)$, which holds as $\sqrt{T}\left(\frac{1}{T}d_{\lambda\lambda i}^*(\beta_0,\lambda_{i0})-\kappa_{\lambda\lambda}^*(\beta_0,\lambda_{i0})\right) = O_p(1)$.

Expanding $T^{-1/2}d^*_{\lambda i}\left(\beta_0, \widehat{\lambda}_i(\beta_0)\right)$ in the usual way we obtain

$$0 = \frac{1}{\sqrt{T}}d^*_{\lambda i}\left(\beta_0, \widehat{\lambda}_i(\beta_0)\right) = \frac{1}{\sqrt{T}}d^*_{\lambda i}(\beta_0, \lambda_{i0}) + \frac{1}{T}d^*_{\lambda\lambda i}(\beta_0, \lambda_{i0})\sqrt{T}\left(\widehat{\lambda}_i(\beta_0) - \lambda_{i0}\right) + O_p\left(\frac{1}{\sqrt{T}}\right)$$

or

$$0 = \frac{1}{\sqrt{T}}d^*_{\lambda i}(\beta_0, \lambda_{i0}) + \kappa^*_{\lambda\lambda i}\sqrt{T}\left(\widehat{\lambda}_i(\beta_0) - \lambda_{i0}\right) + O_p\left(\frac{1}{\sqrt{T}}\right),$$

Hence, also

$$\sqrt{T}\left(\widehat{\lambda}_i(\beta_0) - \lambda_{i0}\right) = -\frac{1}{\kappa^*_{\lambda\lambda i}}\frac{1}{\sqrt{T}}d^*_{\lambda i}(\beta_0, \lambda_{i0}) + O_p\left(\frac{1}{\sqrt{T}}\right), \quad \text{(A2)}$$

and

$$T\left(\widehat{\lambda}_i(\beta_0) - \lambda_{i0}\right)^2 = \frac{1}{\left(\kappa^*_{\lambda\lambda i}\right)^2}\frac{1}{T}\left[d^*_{\lambda i}(\beta_0, \lambda_{i0})\right]^2 + O_p\left(\frac{1}{\sqrt{T}}\right) = -\frac{1}{\kappa^*_{\lambda\lambda i}} + O_p\left(\frac{1}{\sqrt{T}}\right). \quad \text{(A3)}$$

Combining (A1), (A2) and (A3):

$$\begin{aligned} d^*_{\beta i}\left(\beta_0, \widehat{\lambda}_i(\beta_0)\right) &= d^*_{\beta i}(\beta_0, \lambda_{i0}) - \frac{1}{\kappa^*_{\lambda\lambda i}}d^*_{\beta\lambda i}(\beta_0, \lambda_{i0})\left[\frac{1}{T}d^*_{\lambda i}(\beta_0, \lambda_{i0}) + O_p\left(\frac{1}{T}\right)\right] \\ &\quad + \frac{1}{2}d^*_{\beta\lambda\lambda i}(\beta_0, \lambda_{i0})\frac{1}{T}\left[-\frac{1}{\kappa^*_{\lambda\lambda i}} + O_p\left(\frac{1}{\sqrt{T}}\right)\right] + O_p\left(\frac{1}{\sqrt{T}}\right) \\ &= d^*_{\beta i}(\beta_0, \lambda_{i0}) - \frac{1}{\kappa^*_{\lambda\lambda i}}\frac{1}{T}d^*_{\beta\lambda i}(\beta_0, \lambda_{i0})\, d^*_{\lambda i}(\beta_0, \lambda_{i0}) - \frac{\kappa^*_{\beta\lambda\lambda i}}{2\kappa^*_{\lambda\lambda i}} + O_p\left(\frac{1}{\sqrt{T}}\right) \\ &= d^*_{\beta i}(\beta_0, \lambda_{i0}) + \frac{\kappa^*_{\beta\lambda\lambda i}}{2\kappa^*_{\lambda\lambda i}} + O_p\left(\frac{1}{\sqrt{T}}\right) \qquad \text{(A4)} \end{aligned}$$

where we have made use of the facts that due to the orthogonality between λ_i and β we have $d^*_{\beta\lambda i}(\beta_0, \lambda_{i0}) = O_p\left(\sqrt{T}\right)$ and[13]

$$E\left[\frac{1}{T}d^*_{\beta\lambda i}(\beta_0, \lambda_{i0})\, d^*_{\lambda i}(\beta_0, \lambda_{i0})\right] = -\kappa^*_{\beta\lambda\lambda i}.$$

[13]Let $f = f(x; \beta, \lambda)$ and write information orthogonality as

$$\int \frac{\partial^2 \log f}{\partial\beta\partial\lambda} f dx = 0.$$

Taking derivatives with respect to λ we obtain:

$$\int \frac{\partial^3 \log f}{\partial\beta\partial\lambda^2} f dx + \int \frac{\partial^2 \log f}{\partial\beta\partial\lambda}\frac{\partial \log f}{\partial\lambda} f dx = 0.$$

Finally, given the zero-mean property of the score

$$E\left[d_{\beta i}^{*}\left(\beta_{0},\lambda_{i0}\right)\mid x_{i},\lambda_{i}\right]=0$$

the bias of the concentrated score is $O\left(1\right)$ and can be written as

$$E\left[d_{\beta i}^{*}\left(\beta_{0},\widehat{\lambda}_{i}\left(\beta_{0}\right)\right)\mid x_{i},\lambda_{i}\right]=\frac{\kappa_{\beta\lambda\lambda i}^{*}}{2\kappa_{\lambda\lambda i}^{*}}+O\left(\frac{1}{T}\right).$$

The remainder is $O\left(T^{-1}\right)$ since the $O_{p}\left(T^{-1/2}\right)$ terms in the concentrated score have zero mean (cf. Ferguson et al., 1991, p. 290).

Expansion for the Score of the Modified Concentrated Likelihood The *mpf* is given by

$$\ell_{Mi}\left(\beta\right)=\ell_{i}^{*}\left(\beta,\widehat{\lambda}_{i}\left(\beta\right)\right)-\frac{1}{2}\log\left[-d_{\lambda\lambda i}^{*}\left(\beta,\widehat{\lambda}_{i}\left(\beta\right)\right)\right]$$

and the *mpf* score

$$d_{Mi}\left(\beta\right)=d_{\beta i}^{*}\left(\beta,\widehat{\lambda}_{i}\left(\beta\right)\right)-\frac{1}{2}\frac{d}{d\beta}\log\left[-d_{\lambda\lambda i}^{*}\left(\beta,\widehat{\lambda}_{i}\left(\beta\right)\right)\right].$$

Let us consider the form of the difference between the modified and ordinary concentrated scores at β_{0}:

$$\begin{aligned}&d_{Mi}\left(\beta_{0}\right)-d_{\beta i}^{*}\left(\beta_{0},\widehat{\lambda}_{i}\left(\beta_{0}\right)\right)\\=&-\frac{1}{2\frac{1}{T}d_{\lambda\lambda i}^{*}\left(\beta_{0},\widehat{\lambda}_{i}\left(\beta_{0}\right)\right)}\left(\frac{1}{T}d_{\lambda\lambda\beta i}^{*}\left(\beta_{0},\widehat{\lambda}_{i}\left(\beta_{0}\right)\right)+\frac{1}{T}d_{\lambda\lambda\lambda i}^{*}\left(\beta_{0},\widehat{\lambda}_{i}\left(\beta_{0}\right)\right)\frac{\partial\widehat{\lambda}_{i}\left(\beta_{0}\right)}{\partial\beta}\right).\end{aligned}$$

Since $\widehat{\lambda}_{i}\left(\beta_{0}\right)=\lambda_{i0}+O_{p}\left(T^{-1/2}\right)$ we have

$$d_{Mi}\left(\beta_{0}\right)-d_{\beta i}^{*}\left(\beta_{0},\widehat{\lambda}_{i}\left(\beta_{0}\right)\right)=-\frac{1}{2\kappa_{\lambda\lambda i}^{*}}\left(\kappa_{\beta\lambda\lambda i}^{*}+\kappa_{\lambda\lambda\lambda i}^{*}\frac{\partial\widehat{\lambda}_{i}\left(\beta_{0}\right)}{\partial\beta}\right)+O_{p}\left(\frac{1}{\sqrt{T}}\right)$$

Thus,

$$E\left(\frac{\partial^{2}\log f}{\partial\beta\partial\lambda}\frac{\partial\log f}{\partial\lambda}\right)=-E\left(\frac{\partial^{3}\log f}{\partial\beta\partial\lambda^{2}}\right).$$

where $\kappa^*_{\lambda\lambda\lambda i} = E\left[T^{-1}d^*_{\lambda\lambda\lambda i}(\beta_0, \lambda_{i0}) \mid x_i, \lambda_i\right]$.

Now, differentiating $d^*_{\lambda i}\left(\beta, \widehat{\lambda}_i(\beta)\right) = 0$ we obtain

$$d^*_{\beta\lambda i}\left(\beta, \widehat{\lambda}_i(\beta)\right) + d^*_{\lambda\lambda i}\left(\beta, \widehat{\lambda}_i(\beta)\right)\frac{\partial\widehat{\lambda}_i(\beta)}{\partial\beta} = 0$$

or

$$\frac{\partial\widehat{\lambda}_i(\beta)}{\partial\beta} = -\frac{d^*_{\beta\lambda i}\left(\beta, \widehat{\lambda}_i(\beta)\right)}{d^*_{\lambda\lambda i}\left(\beta, \widehat{\lambda}_i(\beta)\right)}.$$

Therefore,

$$\frac{\partial\widehat{\lambda}_i(\beta_0)}{\partial\beta} = -\frac{\kappa^*_{\beta\lambda i}}{\kappa^*_{\lambda\lambda i}} + O_p\left(\frac{1}{\sqrt{T}}\right),$$

but because of orthogonality $\kappa^*_{\beta\lambda i} = E\left[T^{-1}d^*_{\beta\lambda i}(\beta_0, \lambda_{i0}) \mid x_i, \lambda_i\right] = 0$, so that $\partial\widehat{\lambda}_i(\beta_0)/\partial\beta$ is $O_p\left(T^{-1/2}\right)$ and

$$d_{Mi}(\beta_0) - d^*_{\beta i}\left(\beta_0, \widehat{\lambda}_i(\beta_0)\right) = -\frac{\kappa^*_{\beta\lambda\lambda i}}{2\kappa^*_{\lambda\lambda i}} + O_p\left(\frac{1}{\sqrt{T}}\right).$$

Finally, combining this result with (A4) we obtain

$$d_{Mi}(\beta_0) = d^*_{\beta i}(\beta_0, \lambda_{i0}) + O_p\left(\frac{1}{\sqrt{T}}\right). \tag{A5}$$

Thus, the difference between the concentrated likelihood and the modified concentrated likelihood depends primarily on the value of $\kappa^*_{\beta\lambda\lambda i}$. If $\kappa^*_{\beta\lambda\lambda i} = 0$ the scores from both functions will have biases of the same order of magnitude (Cox and Reid, 1992).

Asymptotic Normality of the ML Estimator Let us begin by assuming that, as $T/N \to c$, $0 < c < \infty$, a standard central limit theorem applies to the true score $d^*_{\beta i}(\beta, \lambda_i) = \partial\ell^*_i(\beta, \lambda_i)/\partial\beta$, so that we have

$$V_{NT}^{-1/2}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N} d^*_{\beta i}(\beta_0, \lambda_{i0}) \xrightarrow{d} \mathcal{N}(0, I) \tag{A6}$$

where $V_{NT} = (NT)^{-1} \sum_{i=1}^{N} d_{\beta i}^{*} (\beta_0, \lambda_{i0}) \, d_{\beta i}^{*} (\beta_0, \lambda_{i0})'$.

Using (A4) we can write

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} d_{\beta i}^{*} \left(\beta_0, \widehat{\lambda}_i (\beta_0) \right) = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} d_{\beta i}^{*} (\beta_0, \lambda_{i0}) + \sqrt{\frac{N}{T}} b_N + \sqrt{\frac{N}{T^2}} a_N$$

where $b_N = N^{-1} \sum_{i=1}^{N} \left[\kappa_{\beta\lambda\lambda i}^{*} / \left(2\kappa_{\lambda\lambda i}^{*} \right) \right]$, $a_N = N^{-1} \sum_{i=1}^{N} a_i$, and a_i is an $O_p(1)$ term. Therefore,

$$V_{NT}^{-1/2} \{ \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} d_{\beta i}^{*} \left(\beta_0, \widehat{\lambda}_i (\beta_0) \right) - \sqrt{\frac{N}{T}} b_N \} \xrightarrow{d} \mathcal{N}(0, I) . \qquad \text{(A7)}$$

Next, from a first order expansion of the concentrated score around the true value, we obtain

$$H_{NT} \sqrt{NT} \left(\widehat{\beta} - \beta_0 \right) = - \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} d_{\beta i}^{*} \left(\beta_0, \widehat{\lambda}_i (\beta_0) \right) + O_p \left(\frac{1}{\sqrt{NT}} \right) \qquad \text{(A8)}$$

where

$$H_{NT} = \frac{1}{NT} \sum_{i=1}^{N} \frac{\partial}{\partial \beta'} d_{\beta i}^{*} \left(\beta_0, \widehat{\lambda}_i (\beta_0) \right) .$$

Combining (A7) and (A8) we can write

$$V_{NT}^{-1/2} H_{NT} \sqrt{NT} \left(\widehat{\beta} - \beta_0 + \frac{1}{T} H_{NT}^{-1} b_N \right) =$$

$$- V_{NT}^{-1/2} \{ \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} d_{\beta i}^{*} \left(\beta_0, \widehat{\lambda}_i (\beta_0) \right) - \sqrt{\frac{N}{T}} b_N \} + O_p \left(\frac{1}{\sqrt{NT}} \right) .$$

and finally,

$$\left(H_{NT}' V_{NT}^{-1} H_{NT} \right)^{1/2} \sqrt{NT} \left(\widehat{\beta} - \beta_0 + \frac{1}{T} H_{NT}^{-1} b_N \right) \xrightarrow{d} \mathcal{N}(0, I) .$$

Asymptotic Normality of the MML Estimator We now turn to consider the asymptotic distribution of the modified ML estimator as $T/N \to c$, $0 < c < \infty$. In view of (A5), given (A6) we have

$$V_{NT}^{-1/2} \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} d_{Mi}(\beta_0) \xrightarrow{d} \mathcal{N}(0, I). \tag{A9}$$

Next, from a first order expansion of the modified score around the true value, we obtain

$$H_{NT}^{\dagger} \sqrt{NT} \left(\widehat{\beta}_{MML} - \beta_0 \right) = -\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} d_{Mi}(\beta_0) + O_p\left(\frac{1}{\sqrt{NT}} \right) \tag{A10}$$

where

$$H_{NT}^{\dagger} = \frac{1}{NT} \sum_{i=1}^{N} \frac{\partial d_{Mi}(\beta_0)}{\partial \beta'}.$$

Finally, combining (A9) and (A10) we can write

$$V_{NT}^{-1/2} H_{NT}^{\dagger} \sqrt{NT} \left(\widehat{\beta}_{MML} - \beta_0 \right) = -V_{NT}^{-1/2} \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} d_{Mi}(\beta_0) + O_p\left(\frac{1}{\sqrt{NT}} \right)$$

and

$$\left(H_{NT}^{\dagger\prime} V_{NT}^{-1} H_{NT}^{\dagger} \right)^{1/2} \sqrt{NT} \left(\widehat{\beta}_{MML} - \beta_0 \right) \xrightarrow{d} \mathcal{N}(0, I).$$