NBER WORKING PAPER SERIES

Analysis of Covariance with Qualitative Data

Gary Chamberlain

Working Paper No. 325

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge MA 02138

March 1979

The research reported here is part of the NBER's research program in Labor Economics. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research. Financial Support was provided by the National Science Foundation (Grant No. SOC77-15624).

NBER Working Paper 325
March 1979

Analysis of Covariance with Qualitative Data

ABSTRACT

In data with a group structure, incidental parameters are included to control for missing variables. Applications include longitudinal data and sibling data. In general, the joint maximum likelihood estimator of the structural parameters is not consistent as the number of groups increases, with a fixed number of observations per group. Instead a conditional likelihood function is maximized, conditional on sufficient statistics for the incidental parameters. In the logit case, a standard conditional logit program can be used. Another solution is a random effects model, in which the distribution of the incidental parameters may depend upon the exogenous variables.

Gary Chamberlain
Department of Economics
Harvard University
Littauer Center
Cambridge, MA 02138
617/495-3203

ANALYSIS OF COVARIANCE WITH QUALITATIVE DATA

by

Gary Chamberlain
Harvard University

1. Introduction

This paper deals with data that has a group structure. A simple example in the context of a linear regression model is

$$E(y_{it}|\underset{\sim}{x}, \underset{\sim}{\beta}, \underset{\sim}{\alpha}) = \underset{\sim}{\beta}'\underset{\sim}{x}_{it} + \alpha_i \quad (i=1, \ldots, N;\ t=1, \ldots, T),$$

where there are T observations within each of N groups. The α_i are group specific parameters. Our primary concern is with the estimation of $\underset{\sim}{\beta}$, a parameter vector common to all groups. The role of the α_i is to control for group specific effects; i.e., for omitted variables that are constant within a group. The regression function that does not condition on the group will not in general identify $\underset{\sim}{\beta}$:

$$E(y_{it}|\underset{\sim}{x}, \underset{\sim}{\beta}) \neq \underset{\sim}{\beta}'\underset{\sim}{x}_{it}.$$

In this case there is an omitted variable bias.

An important application is generated by longitudinal or panel data, in which there are two or more observations on each individual. Then the group is the individual, and the α_i capture individual differences. If these person effects are correlated with $\underset{\sim}{x}$, then a regression function that fails to control for them will not identify $\underset{\sim}{\beta}$. In another important application the group is a family, with observations on two or more siblings within the family. Then the α_i capture omitted variables that are family specific, and they give a concrete representation to family background.

We shall assume that observations from different groups are independent.

Then the α_i are incidental parameters (Neyman and Scott [31]), and $\underset{\sim}{\beta}$, which is common to the independent sampling units, is a vector of structural parameters. In the application to sibling data, T is small, typically T=2, whereas there may be a large number of families. Small T and large N are also characteristic of many of the currently available longitudinal data sets. So a basic statistical issue is to develop an estimator for $\underset{\sim}{\beta}$ that has good properties in this case. In particular, the estimator ought to be consistent as $N \to \infty$ for fixed T.

It is well-known that analysis of covariance in the linear regression model does have this consistency property. The problem of finding consistent estimators in other models is non-trivial, however, since the number of incidental parameters is increasing with sample size. We shall work with the following probability model: y_{it} is a binary variable with

$$\text{Prob}(y_{it} = 1 | \underset{\sim}{x}, \underset{\sim}{\beta}, \underset{\sim}{\alpha}) = F(\underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i),$$

where F( ) is a cumulative distribution function such as a unit normal or a logistic. For example, y may indicate labor force participation, unemployment, job change, marital status, health status, or a college degree. Section 2 considers maximum likelihood (ML) estimation of the fixed effects version of this model. A simple algorithm is available which involves a weighted analysis of covariance at each iteration. The ML estimator of $\underset{\sim}{\beta}$ is not consistent (for fixed T), however, and we present a simple example with T=2 in which the ML estimator of $\underset{\sim}{\beta}$ converges to $2\underset{\sim}{\beta}$.

Section 3 presents one solution to this problem by working with a conditional likelihood function that conditions on sufficient statistics for the incidental parameters. This likelihood function does not depend upon the incidental parameters, and hence standard asymptotic theory for maximum likelihood estimation applies. This approach is applied to a

multinomial logit model for grouped data and to the multivariate log-linear probability model. Section 4 develops an alternative approach, based on a random effects model in which the incidental parameters are assumed to follow a distribution. The important point here is that the distribution of the α_i is not assumed to be independent of $\underset{\sim}{x}$; otherwise the problem of omitted variable bias would be assumed away from the beginning. Throughout the paper we shall use the familiar linear regression case to guide the exposition.

2. Fixed Effects: Maximization of the Joint Likelihood Function

We shall begin with a brief review of the linear regression case. Let

$$y_{it} = \underset{\sim}{\beta}'\underset{\sim}{x}_{it} + \alpha_i + \varepsilon_{it},$$

where ε_{it} is i.i.d. $N(0, \sigma^2)$. So in addition to assuming independence across the groups, we are assuming that observations within a group are independent as well, conditional on the group effects. The dependence of different observations within a group is assumed to be due to their common dependence on the group specific α_i. More general forms of dependence are, of course, possible; for example, there could be serial correlation in addition to the α_i in the longitudinal case.

Maximum likelihood for this model is simply a multiple regression of y on $\underset{\sim}{x}$ and a set of group indicator dummy variables. A useful computational simplification is that the ML estimator of $\underset{\sim}{\beta}$ can be obtained from a regression of $y_{it}-\bar{y}_i$ on $\underset{\sim}{x}_{it}-\bar{\underset{\sim}{x}}_i$, where $\bar{y}_i$ and $\bar{\underset{\sim}{x}}_i$ are group means ($\bar{y}_i = \Sigma_t y_{it}/T$). In the case of T=2, this is equivalent to a regression of $y_{i2}-y_{i1}$ on $\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1}$. Since we have

$$y_{i2} - y_{i1} = \underset{\sim}{\beta}'(\underset{\sim}{x}_{i2} - \underset{\sim}{x}_{i1}) + \varepsilon_{i2} - \varepsilon_{i1},$$

with the ε's independent of $\underset{\sim}{x}$, it is clear this provides a consistent estimator of $\underset{\sim}{\beta}$ as $N \to \infty$ (provided that there is sufficient variation in $\underset{\sim}{x}_2 - \underset{\sim}{x}_1$).

There is a comparable computational simplification for the probability models. We shall discuss ML estimation using either a Newton-Raphson or a scoring algorithm, and shall show that each iteration reduces to a weighted analysis of covariance. The binary y_{it} are assumed to be independent (conditional on $\underset{\sim}{x}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{\alpha}$) both between and within groups, with $\text{Prob}(y_{it} = 1 | \underset{\sim}{x}, \underset{\sim}{\beta}, \underset{\sim}{\alpha}) = F(\underset{\sim}{\beta}'\underset{\sim}{x}_{it} + \alpha_i)$. Let $\underset{\sim}{\theta}'\underset{\sim}{z}_{it} = \underset{\sim}{\beta}'\underset{\sim}{x}_{it} + \alpha_i$. Then the log-likelihood function is

$$L = \sum_{i,t} \{y_{it} \ln F(\underset{\sim}{\theta}'\underset{\sim}{z}_{it}) + (1 - y_{it}) \ln[1 - F(\underset{\sim}{\theta}'\underset{\sim}{z}_{it})]\}.$$

Note that if $y_{it} = 1$ for all t then the ML estimate of α_i is ∞, and if $y_{it} = 0$ for all t then the ML estimate of α_i is $-\infty$. Hence the observations on such groups do not affect the ML estimate of $\underset{\sim}{\beta}$, and we can simplify by only including in L the groups within which y varies.

We have the following score vector and Hessian:

$$\frac{\partial L}{\partial \underset{\sim}{\theta}} = \sum_{i,t} \left(\frac{y_{it}}{F} - \frac{1-y_{it}}{1-F}\right) F' \underset{\sim}{z}_{it} \qquad \frac{\partial^2 L}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} = \sum_{i,t} h_{it} \underset{\sim}{z}_{it} \underset{\sim}{z}_{it}',$$

where F and its derivatives are evaluated at $\underset{\sim}{\theta}'\underset{\sim}{z}_{it}$, and

$$h_{it} = -\left[\frac{y_{it}}{F^2} + \frac{1-y_{it}}{(1-F)^2}\right] (F')^2 + \left(\frac{y_{it}}{F} - \frac{1-y_{it}}{1-F}\right) F''.$$

It is well-known that L is concave for probit $[F(u) = \int_{-\infty}^{u} e^{-r^2/2}\, dr/\sqrt{2\pi}\,]$ or for logit $[F(u) = e^u/(1 + e^u)]$. Hence a Newton-Raphson algorithm is expected to be effective:

$$\Delta\underset{\sim}{\theta} = -\left(\frac{\partial^2 L}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'}\right)^{-1} \frac{\partial L}{\partial \underset{\sim}{\theta}}.$$

Also of interest is a scoring algorithm which replaces $\partial^2 L/\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'$

by its expectation:[1]

$$E\left(\frac{\partial^2 L}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'}\right) = \sum_{i,t} E(h_{it})\, \underset{\sim}{z}_{it}\, \underset{\sim}{z}_{it}',$$

where

$$E(h_{it}) = -\frac{(F')^2}{F(1-F)}.$$

In either case the computational burden at each iteration comes from inverting $\Sigma_{i,t} s_{it}\, \underset{\sim}{z}_{it}\, \underset{\sim}{z}_{it}'$, where s_{it} is either h_{it} or $E(h_{it})$. Simplifying the partitioned inverse gives the following formulas for up-dating $\underset{\sim}{\beta}$ and α_i:

$$\Delta\underset{\sim}{\beta} = \left(\sum_{i,t} s_{it}\, \underset{\sim}{x}_{it}\, \underset{\sim}{x}_{it}' - \sum_i s_i\, \bar{\underset{\sim}{x}}_i^*\, \bar{\underset{\sim}{x}}_i'^*\right)^{-1} \left(\sum_{i,t} s_{it}\, \underset{\sim}{x}_{it}\, \psi_{it} - \sum_i s_i\, \bar{\underset{\sim}{x}}_i^*\, \bar{\psi}_i^*\right)$$

$$\Delta\alpha_i = \bar{\psi}_i^* - (\Delta\underset{\sim}{\beta})'\, \bar{\underset{\sim}{x}}_i^* \quad (i=1,\ldots,N),$$

where

$$\psi_{it} = (y_{it} - F)/F'$$

$$s_i = \sum_t s_{it},\ \bar{\underset{\sim}{x}}_i^* = \frac{1}{s_i} \sum_t s_{it}\, \underset{\sim}{x}_{it},\ \bar{\psi}_i^* = \frac{1}{s_i} \sum_t s_{it}\, \psi_{it}.$$

At each iteration, F and its derivatives are evaluated at the current values for $\underset{\sim}{\beta}'\underset{\sim}{x}_{it} + \alpha_i$.

This iterated, weighted analysis of covariance algorithm is computationally effective.[2] Unfortunately, the consistency property (for fixed T) of the ML estimator of $\underset{\sim}{\beta}$ in the linear regression model does not carry over to this case. That maximum likelihood need not be consistent in the presence of incidental parameters can be illustrated in the linear regression model. The ML estimator of σ^2 does not adjust for degrees of freedom, and hence

$$\operatorname*{plim}_{N\to\infty} \hat{\sigma}^2 = \frac{T-1}{T}\, \sigma^2.$$

For T=2, the ML estimator is inconsistent by a factor of two.[3]

Another example is a first order autoregression:

$$y_{it} = \beta y_{i,t-1} + \alpha_i + \varepsilon_{it}.$$

We shall condition on y_{i0}. In that case the likelihood function with ε_{it} i.i.d. $N(0, \sigma^2)$ is formally identical to the previous case. The log-likelihood function is quadratic in β and $\underset{\sim}{\alpha}$ (given σ^2), and the ML estimator of β is analysis of covariance. With T=2, it can be obtained from a least squares regression of $y_{i2}-y_{i1}$ on $y_{i1}-y_{i0}$. Given that the log-likelihood function is quadratic, it is rather surprising that the ML estimator for β is not consistent. The inconsistency follows immediately since

$$y_{i2} - y_{i1} = \beta(y_{i1} - y_{i0}) + \varepsilon_{i2} - \varepsilon_{i1},$$

and ε_1 is correlated with y_1. If the joint distribution of (y_0, y_1, y_2) is stationary, then the estimator converges to $(\beta-1)/2$ as $N\to\infty$.

As an example of the inconsistency of maximum likelihood in the probability models, consider the following logit model: $F(u) = e^u/(1 + e^u)$, T=2, $x_{i1}=0$, $x_{i2}=1$, i=1, . . ., N. So the "treatment" is administered only to the second observation in the group. Assume that the sequence of α_i's is such that the following limits exist:

$$\lim_{N\to\infty} \frac{1}{N} \sum_i E[y_{i1}(1-y_{i2})|\alpha_i] = m_1$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_i E[(1-y_{i1})y_{i2}|\alpha_i] = m_2 ,$$

where $E[y_{i1}(1-y_{i2})|\alpha_i] = F(\alpha_i)F(-\alpha_i -\beta)$ and $E[(1-y_{i1})y_{i2}|\alpha_i] = F(-\alpha_i)F(\alpha_i + \beta)$. Then Andersen [1973, p. 66] shows that the ML estimator of β almost surely satisfies

$$\hat{\beta} = 2\beta$$

as $N\to\infty$. A simple extension of his argument shows that if F is a distribution function corresponding to a symmetric, continuous, non-zero probability density, then

$$\hat{\beta} = 2F^{-1}\left(\frac{m_2}{m_1+m_2}\right)$$

almost surely as $N\to\infty$. The logit case is special in that $m_2/m_1 = e^{\beta}$, independently of the sequence of α_i's. In general the limiting $\hat{\beta}$ depends on this sequence; but if all of the $\alpha_i = 0$, then once again we obtain $\hat{\beta} = 2\beta$ almost surely as $N\to\infty$.

We conclude that the linear regression model is very special. The consistency of the ML estimator of $\underset{\sim}{\beta}$ does not carry over to other models. The next section interprets this result by introducing a conditional likelihood function that conditions on sufficient statistics for the incidental parameters. In the linear regression case, the conditional ML estimator of $\underset{\sim}{\beta}$ is identical to the ML estimator based on the original joint likelihood function. Then we show that the idea of using such a conditional likelihood function can be applied to other models.

3. Fixed Effects: The Conditional Likelihood Function

We have seen that maximization of the fixed effects likelihood function can give seriously inconsistent estimators if there are only a small number of observations per group. This section will develop an alternative approach using a conditional likelihood function. The key idea is to base the likelihood function on the conditional distribution of the data, conditioning on a set of sufficient statistics for the incidental parameters.[4]

We shall begin by applying this idea to the familiar linear regression case. Let

$$y_{it} = \underset{\sim}{\beta}'\underset{\sim}{x}_{it} + \alpha_i + \varepsilon_{it},$$

with ε_{it} i.i.d. $N(0, \sigma^2)$. Then a sufficient statistic for α_i is $\Sigma_t y_{it}$. It is straightforward to check that the conditional density for $y_{i1}, \ldots, y_{iT}$, conditional on $\Sigma_t y_{it}$, is

$$f(y_{i1}, \ldots, y_{iT} | \Sigma_t y_{it}) =$$

$$\sqrt{T}(2\pi)^{-(T-1)/2}\sigma^{-(T-1)} \exp\{-\frac{1}{2\sigma^2} \Sigma_t[(y_{it} - \bar{y}_i) - \underset{\sim}{\beta}'(\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i)]^2\}.$$

Note that this conditional density does not depend upon α_i. Hence the conditional log-likelihood function depends only upon $\underset{\sim}{\beta}$ and σ:

$$L = -N(T-1)\ln\sigma - \frac{1}{2\sigma^2} \Sigma_{i,t}[(y_{it} - \bar{y}_i) - \underset{\sim}{\beta}'(x_{it} - \bar{x}_i)]^2;$$

there is no incidental parameter problem, and so maximum likelihood will give consistent estimates provided that the usual regularity conditions are satisfied. The conditional ML estimator of $\underset{\sim}{\beta}$ is the analysis of covariance estimator that results from maximization of the joint likelihood function. Hence the consistency of that estimator, which was surprising given the incidental parameter problem, follows immediately from the coincidence of the joint and the conditional ML estimators.

The advantage of the conditional likelihood approach can be seen in the conditional ML estimator for σ^2:

$$\hat{\sigma}^2 = \frac{1}{N(T-1)} \Sigma_{i,t}[(y_{it} - \bar{y}_i) - \hat{\underset{\sim}{\beta}}'(\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i)]^2.$$

Unlike the joint ML estimator, here there is a correction for degrees of freedom which ensures that $\hat{\sigma}^2$ is a consistent estimator of σ^2.

The conditional likelihood approach can be applied directly to the fixed effects logit probability model, since $\Sigma_t y_{it}$ is again a sufficient statistic for α_i.[5] Consider first the case of T=2. If $y_{i1} + y_{i2} = 0$ or 2, then y_{i1} and y_{i2} are both determined given their sum. So the only case of interest is $y_{i1} + y_{i2} = 1$. Then the two possibilities are $w_i = 1$ if $(y_{i1}, y_{i2}) = (0,1)$ and $w_i = 0$ if $(y_{i1}, y_{i2}) = (1,0)$. The conditional

density is

$$\text{Prob}(w_i = 1 \mid y_{i1} + y_{i2} = 1) = \text{Prob}(w_i = 1)/[\text{Prob}(w_i = 0) + \text{Prob}(w_i = 1)]$$

$$= \frac{e^{\underset{\sim}{\beta}'(\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1})}}{1 + e^{\underset{\sim}{\beta}'(\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1})}} = F[\underset{\sim}{\beta}'(\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1})],$$

which does not depend upon α_i. The conditional log-likelihood function is

$$L = \sum_{i \in I_1} \{w_i \ \ln F[\underset{\sim}{\beta}'(\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1})] + (1-w_i) \ \ln F[-\underset{\sim}{\beta}'(\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1})]\},$$

where $I_1 = \{i | y_{i1} + y_{i2} = 1\}$.

This conditional likelihood function does not depend upon the incidental parameters. In fact, it is in the form of a binary logit likelihood function in which the two outcomes are (0,1) and (1,0) with explanatory variables $\underset{\sim}{x}_2-\underset{\sim}{x}_1$. This is the analog of differencing in the two period regression model. The conditional ML estimate of $\underset{\sim}{\beta}$ can be obtained simply from a standard ML binary logit program.

The conditional ML estimator of $\underset{\sim}{\beta}$ is consistent provided that the conditional likelihood function satisfies regularity conditions, which impose mild restrictions on the α_i. These restrictions, which constrain the rate at which the sequence of α_i's is allowed to become unbounded, are discussed in Andersen [1], [2]. Furthermore, the inverse of the information matrix based on the conditional likelihood function provides an asymptotic (as $N \to \infty$) covariance matrix for the conditional ML estimator of $\underset{\sim}{\beta}$.[6] In deriving this information matrix, one must be careful to note that I_1 is a random set of indices. This can be made more explicit by defining $d_i = 1$ if $y_{i1} + y_{i2} = 1$ and $d_i = 0$ otherwise. Then we have

$$\frac{\partial^2 L}{\partial \underset{\sim}{\beta} \partial \underset{\sim}{\beta}'} = -\sum_i d_i F(1-F)(\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1})(\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1})',$$

where F and its derivatives are evaluated at $\underset{\sim}{\beta}'(\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1})$. The information matrix is

$$\underset{\sim}{J} = \sum_i P_i F(1-F)(\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1})(\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1})',$$

where

$$P_i = E(d_i|\alpha_i) = F(\alpha_i + \underset{\sim}{\beta}'\underset{\sim}{x}_{i1})F(-\alpha_i - \underset{\sim}{\beta}'\underset{\sim}{x}_{i2}) + F(-\alpha_i - \underset{\sim}{\beta}'\underset{\sim}{x}_{i1})F(\alpha_i + \underset{\sim}{\beta}'\underset{\sim}{x}_{i2}).$$

This information matrix is difficult to evaluate since we do not have a consistent estimator for α_i, which appears in P_i. Moreover, a standard ML binary logit program will be evaluating

$$\underset{\sim}{J}_d = -E\left(\frac{\partial^2 L}{\partial\underset{\sim}{\beta}\partial\underset{\sim}{\beta}'}\Big|\underset{\sim}{d}\right) = -\frac{\partial^2 L}{\partial\underset{\sim}{\beta}\partial\underset{\sim}{\beta}'}$$

(since the Hessian of the logit log-likelihood function is non-stochastic), which depends only upon $\underset{\sim}{\beta}$ (given $\underset{\sim}{d}$). In fact, $\underset{\sim}{J}_d^{-1}$ is an appropriate asymptotic covariance matrix for the conditional ML estimator of $\underset{\sim}{\beta}$, since we can apply the strong law of large numbers to establish that

$$\frac{1}{N}\underset{\sim}{J}_d - \frac{1}{N}\underset{\sim}{J} \xrightarrow{a.s.} 0 \quad \text{as } N \to \infty$$

$$\text{if } \sum_i \underset{\sim}{m}_i\underset{\sim}{m}_i'/i^2 < \infty,$$

where $\underset{\sim}{m}_i$ replaces each element of $(\underset{\sim}{x}_{i2}-\underset{\sim}{x}_{i1})$ by its square. This follows since the d_i are independent with $Ed_i = P_i$, and both F and the variance of d_i are uniformly bounded. The condition for convergence clearly holds if the $\underset{\sim}{x}_{it}$ are uniformly bounded.[7]

For general T, conditioning on $\Sigma_t y_{it}$, i=1, . . ., N, gives the following conditional log-likelihood function:

$$L = \sum_i \ln\left[\exp(\underset{\sim}{\beta}'\Sigma\underset{\sim}{x}_{it}y_{it}) \Big/ \sum_{\underset{\sim}{d}\in B_i} \exp(\underset{\sim}{\beta}'\Sigma\underset{\sim}{x}_{it}d_t)\right],$$

where $B_i = \{d = (d_1, \ldots, d_T) | d_t = 0 \text{ or } 1 \text{ and } \Sigma_t d_t = \Sigma_t y_{it}\}$. L is in conditional logit form with the alternative set (B_i) varying across the observations.[8] There are T+1 distinct alternative sets corresponding to $\Sigma_t y_{it} = 0, 1, \ldots, T$. Groups for which $\Sigma_t y_{it} = 0$ or T contribute zero to L, however, and so only T-1 alternative sets are relevant. The alternative set for groups with $\Sigma_t y_{it} = s$ has $\binom{T}{s}$ elements, corresponding to the distinct sequences of T trials with s successes. For example, with T=3 and s=1 there are three alternatives with the following conditional probabilities:

$$\text{Prob}(1,0,0 | \textstyle\sum_t y_{it} = 1) = \frac{e^{\underset{\sim}{\beta}'(\underset{\sim}{x}_{i1} - \underset{\sim}{x}_{i3})}}{D}$$

$$\text{Prob}(0,1,0 | \textstyle\sum_t y_{it} = 1) = \frac{e^{\underset{\sim}{\beta}'(\underset{\sim}{x}_{i2} - \underset{\sim}{x}_{i3})}}{D}$$

$$\text{Prob}(0,0,1 | \textstyle\sum_t y_{it} = 1) = \frac{1}{D}, \quad D = e^{\underset{\sim}{\beta}'(\underset{\sim}{x}_{i1} - \underset{\sim}{x}_{i3})} + e^{\underset{\sim}{\beta}'(\underset{\sim}{x}_{i2} - \underset{\sim}{x}_{i3})} + 1.$$

Since L is in the form of a conditional logit log-likelihood function, it can be maximized by standard programs. The information matrix evaluated by such a program will implicitly condition on the alternative sets, which are random in our problem. So the program will evaluate $\underset{\sim}{J}_B = -E(\partial^2 L / \partial \underset{\sim}{\beta} \partial \underset{\sim}{\beta}' | B)$. Since the Hessian of the log-likelihood function in conditional logit is non-stochastic, we have $\underset{\sim}{J}_B = -\partial^2 L / \partial \underset{\sim}{\beta} \partial \underset{\sim}{\beta}'$. Hence $\underset{\sim}{J}_B^{-1}$ is an appropriate asymptotic covariance matrix for the conditional ML estimator of $\underset{\sim}{\beta}$ provided that J_B/N converges to its expectation. This will follow from the strong law of large numbers if, for example, the $\underset{\sim}{x}_{it}$ are uniformly bounded.

In the remainder of this section we shall first extend our conditional likelihood approach from the binary to the multinomial case; then we shall apply our approach to the multivariate log-linear probability model, thereby relaxing the assumption that the observations within a group are independent.

Multinomial Logit for Grouped Data. Say that y_{it} can take on three values: a, b, c. Then we have

$$\text{Prob}(y_{it} = j) = \frac{e^{\alpha_{ij} + \beta' x_{itj}}}{\sum_j e^{\alpha_{ij} + \beta' x_{itj}}} \quad (j = a, b, c).$$

We assume that the y's are independent both within and between groups. We shall condition on the number of occurrences within the i^{th} group of each of the three events.

If T=2, then the only cases of interest are those in which two of the three events each occurs once, for otherwise there is no stochastic variation. Conditioning on a and b each occurring once gives (suppressing the i subscript):

$$\text{Prob}[(a,b) \mid (a,b) \text{ or } (b,a)] = \frac{e^{\beta' z}}{1 + e^{\beta' z}},$$

where $z = (x_{2b} - x_{2a}) - (x_{1b} - x_{1a})$. Hence we have a binary logit problem with (a,b) and (b,a) as the two alternatives and with z as the explanatory variables. The incidental parameters do not appear in this conditional probability. There is a similar result when we condition on a and c each occurring once, and also when b and c each occur once.

In the general case of T independent observations on each group with y_{it} taking on J values, we define $w_{itj} = 1$ if $y_{it} = j$ and $w_{itj} = 0$ otherwise. We condition on $s_{ij} = \Sigma_t w_{itj}$, j=1, . . ., J. This gives the following conditional log-likelihood function:

$$L = \sum_i \ln[\exp(\beta' \sum_{t,j} x_{itj} w_{itj}) / \sum_{d \in B_i} \exp(\beta' \sum_{t,j} x_{itj} d_{tj})],$$

where

$$B_i = \{d = (d_{11}, \ldots, d_{TJ}) \mid d_{tj} = 0 \text{ or } 1, \ \Sigma_j d_{tj} = 1, \ \Sigma_t d_{tj} = s_{ij}, \ j=1, \ldots, J\}.$$

This is in the form of a conditional logit log-likelihood function and can be maximized by standard programs.

The Log-Linear Probability Model. We shall relax the assumption that the y_{it} are independent within a group by extending the conditional likelihood approach to the general log-linear model.[9] We begin by illustrating the log-linear model for the binary case (y_{it} = 0 or 1) with T=3 (the i subscripts are suppressed):

$$\ln \text{Prob}(y_1, y_2, y_3) = \mu + \gamma_1 y_1^* + \gamma_2 y_2^* + \gamma_3 y_3^* + \gamma_{12} y_1^* y_2^* + \gamma_{13} y_1^* y_3^* + \gamma_{23} y_2^* y_3^* + \gamma_{123} y_1^* y_2^* y_3^*,$$

where $y^* = 1$ if $y = 1$ and $y^* = -1$ if $y = 0$. This is a saturated model since there are $2^3-1 = 7$ independent probabilities, and there are seven free parameters with μ determined by the constraint that the probabilities sum to one.

A common way to impose structure on this model is to specify the main effects in terms of a set of explanatory variables: $\gamma_{it} = \beta' x_{it}$, and to assume that the interaction terms are constant: $\gamma_{ist} = \gamma_{st}$, for s, t=1, 2, 3, and $\gamma_{i123} = \gamma_{123}$. Additional structure can be imposed by specifying that the

interaction terms beyond some order are zero; for example, that $\gamma_{123} = 0$.

We shall introduce group specific effects by letting $\gamma_{it} = \alpha_i + \underset{\sim}{\beta}'\underset{\sim}{x}_{it}$. It is straightforward to check that

$$\ln \frac{\text{Prob}(y_{i1} = 1 | y_{i2}, y_{i3})}{1-\text{Prob}(y_{i1}=1 | y_{i2}, y_{i3})} = 2\alpha_i + 2\underset{\sim}{\beta}'\underset{\sim}{x}_{i1} + 2\gamma_{12}y^*_{i2} + 2\gamma_{13}y^*_{i3} + 2\gamma_{123}y^*_{i2}y^*_{i3}.$$

So if the interaction terms $\gamma_{12} = \gamma_{13} = \gamma_{123} = 0$, then y_1 is independent of y_2 and y_3, and the probability of $y_1=1$ takes the logistic form that we have been using (except for a scale factor of 2).

For the general case of T binary variables we have (suppressing the i subscripts):

$$\ln \text{Prob}(y_1, \ldots, y_T) = \mu + \sum_{k=1}^{T} \sum_{t \in M_k} \gamma_{\underset{\sim}{t}} y^*_{t_1} \cdots y^*_{t_k},$$

where $M_k = \{(t_1, \ldots, t_k)\}$ is the set consisting of the $\binom{T}{k}$ groups of k integers that can be formed from the integers 1, . . ., T. We shall specify the first order terms as $\gamma_{it} = \alpha_i + \underset{\sim}{\beta}'\underset{\sim}{x}_{it}$. The interaction terms may depend upon $\underset{\sim}{x}$ but with coefficients that do not vary in i, so that the incidental parameters are confined to the first order terms.

Since $\Sigma_t y_{it}$ is a sufficient statistic for α_i, we form the following conditional density:

$$\text{Prob}(y_{i1}, \ldots, y_{iT} | \textstyle\sum_t y_{it}) = \frac{\exp[\Sigma(\alpha_i + \underset{\sim}{\beta}'\underset{\sim}{x}_{it})y^*_{it} + g(\underset{\sim}{y}_i)]}{\sum_{\underset{\sim}{d} \in B_i} \exp[\Sigma(\alpha_i + \underset{\sim}{\beta}'\underset{\sim}{x}_{it})d^*_t + g(\underset{\sim}{d})]}$$

$$= \frac{\exp[\underset{\sim}{\beta}'\Sigma_t \underset{\sim}{x}_{it} y^*_{it} + g(\underset{\sim}{y}_i)]}{\sum_{\underset{\sim}{d} \in B_i} \exp[\underset{\sim}{\beta}'\Sigma_t \underset{\sim}{x}_{it} d^*_t + g(\underset{\sim}{d})]},$$

where $B_i = \{\underset{\sim}{d} = (d_1, \ldots, d_T) | d_t = 0 \text{ or } 1 \text{ and } \Sigma_t d_t = \Sigma_t y_{it}\}$, $\underset{\sim}{y}_i = (y_{i1}, \ldots, y_{iT})$, and g( ) does not depent upon α_i. We see that the conditional density does not depend upon α_i. The corresponding log-likelihood function differs from the one for independent y's

only in the g( ) terms. For example, with T=3 and $\Sigma_t y_{it} = 1$ we have $g(1, 0, 0) = -\gamma_{12} - \gamma_{13} + \gamma_{23} + \gamma_{123}$; $g(0, 1, 0) = -\gamma_{12} + \gamma_{13} - \gamma_{23} + \gamma_{123}$; $g(0, 0, 1) = \gamma_{12} - \gamma_{13} - \gamma_{23} + \gamma_{123}$. Rescaling all the coefficients by one-half, we can write the conditional probabilities as

$$\text{Prob}(1, 0, 0 \mid \Sigma_t y_{it} = 1) = \frac{1}{D} \exp[\underset{\sim}{\beta}'(\underset{\sim}{x}_{i1} - \underset{\sim}{x}_{i3}) + \gamma_{23} - \gamma_{12}]$$

$$\text{Prob}(0, 1, 0 \mid \Sigma_t y_{it} = 1) = \frac{1}{D} \exp[\underset{\sim}{\beta}'(\underset{\sim}{x}_{i2} - \underset{\sim}{x}_{i3}) + \gamma_{13} - \gamma_{12}]$$

$$\text{Prob}(0, 0, 1 \mid \Sigma_t y_{it} = 1) = \frac{1}{D} ,$$

with D determined so that the probabilities sum to one. So this differs from the independence case by introducing alternative specific constants into the conditional probabilities.

We have seen that it is fruitful to base the likelihood function on a conditional distribution that conditions on sufficient statistics for the incidental parameters. It is not always possible, however, to find a sufficient statistic for α_i such that the conditional distribution is sufficiently informative about $\underset{\sim}{\beta}$.[10] The next section examines a random effects model in which a consistent estimator for $\underset{\sim}{\beta}$ can be obtained without relying upon sufficient statistics for the α_i.

4. Random Effects: the Marginal Likelihood Function

An alternative approach is to assume that the incidental parameters follow a distribution. Then the likelihood function can be based on the density for $\underset{\sim}{y}$, given $\underset{\sim}{x}$, $\underset{\sim}{\beta}$, and G, the distribution function for $\underset{\sim}{\alpha}$. If we specify a parametric family for G, indexed by a finite parameter vector $\underset{\sim}{\tau}$, then we have the following log-likelihood function for $\underset{\sim}{\beta}$, $\underset{\sim}{\tau}$:

$$L = \ln\int f(\underset{\sim}{y}|\underset{\sim}{x}, \underset{\sim}{\beta}, \underset{\sim}{\alpha})dG(\underset{\sim}{\alpha}|\underset{\sim}{x}, \underset{\sim}{\tau}).$$

So the density function for $\underset{\sim}{y}$ conditional on $\underset{\sim}{\alpha}$ has been replaced by a density function that is marginal on $\underset{\sim}{\alpha}$.[11] The maximization of this likelihood function will, under weak regularity conditions, give consistent (as $N \to \infty$) estimators for $\underset{\sim}{\beta}$ and $\underset{\sim}{\tau}$.[12]

This approach introduces additional information and is most naturally formulated in Bayesian terms. A potentially appealing prior distribution specifies that the α's are independent and identically distributed. This can often be justified by de Finetti's [16] exchangeability criterion. If (for arbitrary N) the distribution of the α_i's is not affected by permuting them, so that the subscript is purely a labeling device with no substantive content, then the joint distribution of the α's must be expressable as random sampling from a univariate distribution. This criterion will often be satisfied when i indexes individuals (longitudinal data) or families (sibling data).

The main point I want to make here is that the random sampling specification is appropriate only as a marginal distribution for $\underset{\sim}{\alpha}$. We must, however, specify a distribution for $\underset{\sim}{\alpha}$ conditional on $\underset{\sim}{x}$. The conventional random effects model assumes that $\underset{\sim}{\alpha}$ is independent of $\underset{\sim}{x}$. But our interest in introducing the incidental parameters was motivated by missing variables that are correlated with $\underset{\sim}{x}$. If one mistakenly models $\underset{\sim}{\alpha}$ as independent of $\underset{\sim}{x}$, then the omitted variable bias is not eliminated. So we want to specify a conditional distribution for $\underset{\sim}{\alpha}$ given $\underset{\sim}{x}$ that allows for dependence.[13] A convenient possibility is to assume that the dependence is only via a linear regression function: $\alpha_i = \underset{\sim}{\pi}'\underset{\sim}{x}_i + v_i$, with $\underset{\sim}{x}_i' = (\underset{\sim}{x}_{i1}', \ldots, \underset{\sim}{x}_{iT}')$, and where v_i is independent of $\underset{\sim}{x}$. We appeal to exchangeability to argue that the v_i are independent and identically distributed. A restriction on the regression function that may be appropriate is $\underset{\sim}{\pi}'\underset{\sim}{x}_i = \underset{\sim}{\delta}'\bar{\underset{\sim}{x}}_i$.

We shall illustrate this approach with a production function example that leads to a linear regression model.[14] Say that a farmer is producing a product under the following Cobb-Douglas technology: $Y = L^{\beta}Q^{\gamma}e^{\varepsilon}$, where Y is output, L is a variable factor (labor), Q is a fixed factor (soil quality), ε is stochastic (rainfall), and $0 < \beta < 1$. Assume that ε is distributed independently of Q; persistent differences in average rainfall can be incorporated into Q. We assume that the farmer knows the product price (P) and the factor price (W), which do not depend on his decisions, and that he knows Q. The factor input decision, however, is made before knowing ε, and we assume that L is chosen to maximize expected profit: $E(PY - WL|P, W, Q)$.

There are observations on i=1, . . ., N farms in each of t=1, . . ., T periods. Assume that Q is constant over the period of the sample and that the distribution of ε conditional on Q, W, and P is ε_{it} i.i.d. $N(0, \sigma^2)$. Then we have the following production and factor demand functions:

$$y_{it} = \beta x_{it} + \alpha_i + \varepsilon_{it}$$

$$x_{it} = \mu + \frac{1}{1-\beta}(z_{it} + \alpha_i) + u_{it},$$

where $y = \ln Y$, $x = \ln L$, $\alpha = \gamma \ln Q$, $\mu = (\ln\beta + \frac{1}{2}\sigma^2)/(1-\beta)$, $z = \ln(P/W)$, and u is a random term, reflecting optimization and other errors, which is independent of α and ε. Although Q is known to the farmer and affects his factor demand decisions, we assume that it is not observed by the econometrician; α_i is included in order to capture this omitted variable. The example is useful in showing explicitly how a correlation between x and α might arise.

We shall focus on estimating the production function without using whatever price data is available. A pooled least squares regression of y on x, which does not allow for farm effects, is inconsistent. If α is independent of z, then as $N \to \infty$ this estimator converges to

$$\beta + \frac{\sigma_\alpha^2}{(1-\beta)(V_W + V_B)},$$

where

$$V_W = \operatorname*{plim}_{N\to\infty} \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)^2, \quad V_B = \operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_i (\bar{x}_i - \bar{\bar{x}})^2,$$

and σ_α^2 is the marginal variance of α. Now consider a random effects approach, α_i i.i.d. $N(\psi, \sigma_\alpha^2)$, that incorrectly assumes that α is independent of x. Then the ML estimator of β, conditional on $\lambda = \sigma_\alpha^2/\sigma^2$, is generalized least squares. This is equivalent to ordinary least squares using deviations from fractional means: regress $y_{it} - \gamma\bar{y}_i$ on $x_{it} - \gamma\bar{x}_i$, where $\gamma = 1 - (1 + \lambda T)^{-1/2}$. This estimator converges, as $N\to\infty$, to

$$\beta + \frac{(1-\gamma)^2\sigma_\alpha^2}{(1-\beta)[V_W + (1-\gamma)^2 V_B]}.$$

Hence it is consistent only as $T\to\infty$.

So it is essential to allow for a dependence between α and x. Let $w_{it} = z_{it}/(1-\beta) + u_{it}$ and assume that $\underset{\sim}{w}_i$ is i.i.d. $N(\underset{\sim}{m}, \underset{\sim}{\Sigma})$. Then the distribution of α conditional on x is given by $\alpha_i = \kappa + \underset{\sim}{\pi}'\underset{\sim}{x}_i + v_i$, where

$$\underset{\sim}{\pi} = \frac{\sigma_\alpha^2}{1-\beta}\left[\frac{\sigma_\alpha^2}{(1-\beta)^2}\underset{\sim\sim}{\ell\ell}' + \underset{\sim}{\Sigma}\right]^{-1}\underset{\sim}{\ell},$$

$\underset{\sim}{\ell}$ is a Tx1 vector of ones, and v is independent of x with v_i i.i.d. $N(0, \sigma_v^2)$. Note that assuming a stationary $\underset{\sim}{\Sigma}$ does not imply that $\underset{\sim}{\pi}'\underset{\sim}{x}_i = \delta\bar{x}_i$ if $T > 2$. A sufficient condition is that $\underset{\sim}{\Sigma}$ is equicorrelated: $\underset{\sim}{\Sigma} = \rho_1\underset{\sim}{I} + \rho_2\underset{\sim\sim}{\ell\ell}'$.

The ML estimator of $(\underset{\sim}{\beta}, \underset{\sim}{\pi})$, allowing for several variables in $\underset{\sim}{x}_{it}$ and given $\lambda = \sigma_v^2/\sigma^2$, can be obtained from the regression of $y_{it} - \gamma\bar{y}_i$ on $\underset{\sim}{x}_{it} - \gamma\underset{\sim}{\bar{x}}_i$ and $(1-\gamma)\underset{\sim}{x}_i$. The resulting estimator for $\underset{\sim}{\beta}$ can be obtained from the regression of $y_{it} - \gamma\bar{y}_i$ on the residual from the regression of $\underset{\sim}{x}_{it} - \gamma\underset{\sim}{\bar{x}}_i$ on $\underset{\sim}{x}_i$. This residual is $\underset{\sim}{x}_{it} - \underset{\sim}{\bar{x}}_i$; but the regression of $y_{it} - \gamma\bar{y}_i$ on $\underset{\sim}{x}_{it} - \underset{\sim}{\bar{x}}_i$ is equivalent to the regression of $y_{it} - \bar{y}_i$ on $\underset{\sim}{x}_{it} - \underset{\sim}{\bar{x}}_i$.

We have obtained the interesting result that a random effects

specification can give a ML estimator of $\underset{\sim}{\beta}$ that is identical to the fixed effects estimator, if we allow the distribution of the incidental parameters to depend upon $\underset{\sim}{x}$.[15] Of course the linear regression case is special, since the fixed effects estimator is consistent. This is not true for the (joint) ML estimator of $\underset{\sim}{\beta}$ in the linear autoregressive model or in the probability models. So the random effects specification leads to new estimators in those cases.

In the autoregressive case, let

$$y_{i1} = \beta y_{i0} + \alpha_i + \varepsilon_{i1}$$

$$y_{i2} = \beta y_{i1} + \alpha_i + \varepsilon_{i2},$$

where, conditional on y_{i0} and α_i, we have $(\varepsilon_{i1}, \varepsilon_{i2})$ i.i.d. from a normal distribution with mean $\underset{\sim}{0}$ and diagonal covariance matrix: $\text{diag}\{\sigma_1^2, \sigma_2^2\}$. Let $\alpha_i = \pi y_{i0} + v_i$, where, conditional on y_{i0}, we have v_i i.i.d. $N(0, \sigma_v^2)$. Then

$$(y_{i1}, y_{i2}) = (\delta_1, \delta_2)y_{i0} + (u_{i1}, u_{i2}),$$

where $\delta_1 = \beta + \pi$, $\delta_2 = \beta\delta_1 + \pi$, and $\underset{\sim}{u}_i$ is i.i.d. $N(\underset{\sim}{0}, \underset{\sim}{\Sigma})$. This is a multivariate regression model in which the ML estimator of $\underset{\sim}{\delta}$ is obtained from the least squares regressions of y_1 and y_2 on y_0. Then we can solve for the ML estimator of β from $\hat{\beta} = (\hat{\delta}_2 - \hat{\delta}_1)/(\hat{\delta}_1 - 1)$. This estimator is consistent if the y_{i0} have sufficient variation and if $\beta + \pi \neq 1$. It is equivalent to taking first differences, $y_{i2} - y_{i1} = \beta(y_{i1} - y_{i0}) + \varepsilon_{i2} - \varepsilon_{i1}$, and using y_{i0} as an instrumental variable for $y_{i1} - y_{i0}$. If we add the assumption that $\sigma_1 = \sigma_2$, then an additional consistent estimator of β can be obtained from a consistent estimator of $\underset{\sim}{\Sigma}$. Now the ML estimator of β will combine the estimator obtained from the regression coefficients with the estimator obtained from the residual covariance matrix.

The likelihood function for the joint distribution of (y_0, y_1, y_2)

is obtained by multiplying the likelihood conditional on y_0 by the marginal density of y_0. If the parameters of this marginal density are left unconstrained, then the ML estimator of β is unaffected. Imposing stationarity on the joint distribution will, however, imply constraints. If y_{i0} is i.i.d. normal with variance ρ, then stationarity implies that $\rho = \sigma_v^2/[(1-\delta_1)\pi]$.

In the binary data case, let $\text{Prob}(y_{it} = 1|\underset{\sim}{x}, \underset{\sim}{\beta}, \underset{\sim}{\alpha}) = F(\underset{\sim}{\beta}'\underset{\sim}{x}_{it} + \alpha_i)$. Then the log-likelihood function under our random effects specification is

$$L = \sum_i \ln \int \prod_t F(\underset{\sim}{\beta}'\underset{\sim}{x}_{it} + \underset{\sim}{\pi}'\underset{\sim}{x}_i + v)^{y_{it}}[1 - F(\underset{\sim}{\beta}'\underset{\sim}{x}_{it} + \underset{\sim}{\pi}'\underset{\sim}{x}_i + v)]^{1-y_{it}} \, dH(v|\underset{\sim}{\psi}),$$

where $H(\,|\underset{\sim}{\psi})$ is a family of univariate distribution functions indexed by the parameter vector $\underset{\sim}{\psi}$. For example, if F is a unit normal distribution function and we choose H to be the distribution function of a $N(0, \sigma_v^2)$ random variable, then our specification gives a multivariate probit model:

$$y_{it} = 1 \text{ if } \underset{\sim}{\beta}'\underset{\sim}{x}_{it} + \underset{\sim}{\pi}'\underset{\sim}{x}_i + u_{it} > 0$$

$$\underset{\sim}{u}_i \text{ i.i.d. } N(\underset{\sim}{0}, \sigma_v^2 \underset{\sim}{\ell}_T \underset{\sim}{\ell}_T' + \underset{\sim}{I}_T),$$

where $\underset{\sim}{\ell}_T$ is a T x 1 vector of ones. The novel feature of this model is the inclusion of the term $\underset{\sim}{\pi}'\underset{\sim}{x}_i$ to capture the dependence between the incidental parameters and $\underset{\sim}{x}$.

For example, consider estimating the effect of ability on the probability of attending college, controlling for family background. There is a sample of N families with test scores (x) for each of T=2 brothers per family. The family effect α_i is intended to capture omitted variables such as family wealth and parents' schooling. Under this interpretation, α is likely to be correlated with x. Our procedure in the probit case is to fit a (constrained) bivariate probit model for y_{i1} and y_{i2} on x_{i1} and

x_{i2}. This provides estimates of

$$\begin{bmatrix} \beta + \pi_1 & \pi_2 \\ \pi_1 & \beta + \pi_2 \end{bmatrix},$$

from which we obtain an estimate of β by taking the coefficient of sib 1's test score in sib 1's equation minus the coefficient of sib 1's test score in sib 2's equation. We can do the same with sib 2's test score and hence the constraint on the matrix of probit coefficients.

From the symmetry of this example (ignoring birth order effects), it is appropriate to set $\pi_1 = \pi_2$. Then β can be consistently estimated by taking the coefficient of sib 1's test score in sib 1's equation minus the coefficient of sib 2's test score in sib 1's equation. Hence we only require y for one of the sibs provided that we have x for both. For example, the Michigan Panel Study of Income Dynamics [26] has extensive information on the respondent and much less complete information on his siblings. There is schooling data for the respondent and his oldest brother, but earnings and occupation data only for the respondent. Nevertheless, we can control for family background in assessing the relationship between schooling and earnings by including the schooling of sib 2 in a regression of sib 1's earnings on his schooling. Then β is estimated by the excess of sib 1's schooling coefficient over that of his brother. A probit example could arise in studying the relationship between schooling and occupation, where occupations are classified into two groups corresponding to production and non-production workers.

5. Conclusion

The paper has discussed three approaches to the analysis of grouped data: the joint likelihood function, the conditional likelihood function, and the marginal likelihood function. Throughout the paper, our concern has been with

the parameters ($\underset{\sim}{\beta}$) that are common to all of the groups; the incidental parameters (α_i) are intended to capture group effects whose omission would result in biased estimates of $\underset{\sim}{\beta}$. The objective has been to obtain estimators that converge to $\underset{\sim}{\beta}$ as the number of groups (N) increases, even if the number of observations per group (T) is small. Important applications include longitudinal data, in which there are two or more observations on each individual, and the α_i capture person effects; and sibling data, in which the α_i capture family effects, such as omitted family background variables.

We have illustrated the inconsistency of the joint ML estimator in the fixed effects probability models. One solution, within the fixed effects model, is to maximize a conditional likelihood function that conditions on sufficient statistics for the incidental parameters. This conditional likelihood function does not depend upon the incidental parameters, and so standard asymptotic theory can be applied. In the (normal-theory) linear regression model, the consistency of the joint ML estimator of $\underset{\sim}{\beta}$ corresponds to the coincidence of the joint and the conditional ML estimators. In the logit case, however, the conditional ML estimator of $\underset{\sim}{\beta}$ is consistent whereas the joint ML estimator is not (for fixed T). The conditional ML estimator for the logit case can be implemented with a standard conditional logit program, which allows the alternative set to vary across the observations.

Finally, we discussed random effects models which impose a (prior) distribution on the incidental parameters. Then the likelihood function is based on the distribution for y that is marginal on the incidental parameters. The important point here is that the specification of the conditional distribution for α_i given $\underset{\sim}{x}$ should allow for dependence; the common assumption that α_i is independent of $\underset{\sim}{x}$ assumes away omitted variable bias. In the linear regression model, the ML estimator for $\underset{\sim}{\beta}$ under our random effects specification is once again analysis of covariance. So in this special case, all three of our approaches give identical estimators

for $\underset{\sim}{\beta}$. In the probability models, however, the marginal likelihood specification leads to new estimators.

The marginal likelihood approach has the advantage of not requiring simple sufficient statistics for the incidental parameters. Furthermore, it imposes (stochastic) restrictions on the fixed effects model, which will lead to more precise estimators if the restrictions are valid. The disadvantage is that in order to specify that the α_i are independent of each other (conditional on $\underset{\sim}{x}$), our approach requires a particular parametric class of conditional distributions for α_i given $\underset{\sim}{x}$. Hence some sensitivity analysis is called for. The fixed effects model allows for a very general relationship between the incidental parameters and the explanatory variables.

Footnotes

[1] In the logit case the Hessian does not depend upon y, and so scoring is identical to the Newton-Raphson algorithm.

[2] A program to implement this algorithm is described in Hall [21], along with an example of the computational efficiency of the program. A labor force participation application of a fixed effects probit model is presented in Heckman [22].

[3] This example is discussed in Neyman and Scott [31].

[4] The use of conditional likelihood functions for incidental parameter problems is discussed in Bartlett [8], [9], [10], Andersen [1], [5], Kalbfleisch and Sprott [23], and Barndorff-Nielsen [7].

[5] The conditional likelihood approach in the logit case is closely related to R. A. Fisher's [17] exact test for independence in a 2x2 table. This exact significance test has been extended by Cox [15] and others to the case of several contingency tables. Additional references are in Cox [15] and in Bishop et al. [11]. A conditional likelihood approach was used by Rasch [32], [33] in his model for intelligence tests. The probability that person i gives a correct answer to item number t is $\exp(\alpha_i + \beta_t)/[1 + \exp(\alpha_i + \beta_t)]$; this is a special case in which $\underset{\sim}{x}_{it}$ is a set of dummy indicator variables. An algorithm for conditional maximum likelihood estimation in this model is described in Andersen [4].

[6] The efficiency of the conditional ML estimator is maximized by conditioning on minimal sufficient statistics for the incidental parameters. $\Sigma_t y_{it}$ is a minimal sufficient statistic for α_i both in the linear regression model and in the logit model. Even so the conditional ML estimator need not attain the asymptotic Cramer-Rao bound as $N \to \infty$ for fixed T. It does in the linear regression case but not in the logit model. However, I

doubt whether there is another consistent estimator that has smaller asymptotic variance in the fixed effects logit model. The random effects model of section 4, which introduces additional (stochastic) restrictions, can lead to a more efficient estimator of $\underset{\sim}{\beta}$.

[7]An alternative justification for the use of $-E(\partial^2 L/\partial\underset{\sim}{\beta}\partial\underset{\sim}{\beta}' | \underset{\sim}{d})$ can be based on stating the limiting distribution properties in terms of the conditional distribution, in which the observed values of the sufficient statistics are treated as parameters. This approach is pursued in Andersen [3].

[8]The conditional logit model is developed in McFadden [25].

[9]The log-linear model is developed in Goodman [18], [19], Haberman [20], and Nerlove and Press [30]. Additional references are in Bishop et al. [11].

[10]In the probit model, for example, there does not appear to be such a sufficient statistic.

[11]Kalbfleisch and Sprott [23] call this an integrated likelihood function. A marginal likelihood function can also be useful in a fixed effects approach, in which we consider the distribution of some function of $\underset{\sim}{y}_i$, conditional on α_i. For example, in the linear regression case with T=2, the distribution of $y_{i2}-y_{i1}$ does not depend upon α_i. Hence maximizing the associated likelihood function gives consistent (as $N\to\infty$) estimators of $\underset{\sim}{\beta}$ and σ. Once again the ML estimator of $\underset{\sim}{\beta}$ is the standard analysis of covariance estimator.

[12]Note that the original Kiefer and Wolfowitz [24] results were not limited to the parametric case.

[13]Note that the empirical work by Chamberlain and Griliches [13], [14] and Chamberlain [12] does allow the random effects to be correlated with the explanatory variables. Also in the original Balestra and Nerlove [6] model, the autoregressive component is correlated with the random effects.

[14]This example is discussed in Mundlak [27], [28].

[15]This result is discussed in Mundlak [29] for the case $\underset{\sim}{\pi}'\underset{\sim}{x}_i = \underset{\sim}{\delta}'\bar{\underset{\sim}{x}}_i$.

References

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

[2] Andersen, E. B. "Asymptotic Properties of Conditional Likelihood Ratio Tests", Journal of the American Statistical Association, 66 (1971), 630-633.

[3] Andersen, E. B. "A Strictly Conditional Approach in Estimation Theory", Skandinavisk Aktuarietidskrift, (1971), 39-49.

[4] Andersen, E. B. "The Numerical Solution of a Set of Conditional Estimation Equations", Journal of the Royal Statistical Society, Series B, 34 (1972), 42-54.

[5] Andersen, E. B. Conditional Inference and Models for Measuring (Copenhagen: Mentalhygiejnisk Forlag, 1973).

[6] Balestra, P, and Nerlove, M. "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas", Econometrica, 34 (1966), 585-612.

[7] Barndorff-Nielsen, O. Information and Exponential Families in Statistical Theory (New York: Wiley, 1978).

[8] Bartlett, M. S. "The Information Available in Small Samples", Proceedings of the Cambridge Philosophical Society, 32 (1936), 560-566.

[9] Bartlett, M. S. "Statistical Information and Properties of Sufficiency", Proceedings of the Royal Society, Series A, 154 (1936), 124-137.

[10] Bartlett, M. S. "Properties of Sufficiency and Statistical Tests", Proceedings of the Royal Society, Series A, 160 (1937), 268-282.

[11] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. Discrete Multivariate Analysis: Theory and Practice (Cambrige, Mass.: M.I.T. Press, 1975).

[12] Chamberlain, G. "Omitted Variable Bias in Panel Data: Estimating the Returns to Schooling", Annales de l'INSEE, 30-31 (1978), 49-82.

[13] Chamberlain, G. and Griliches, Z. "Unobservables with a Variance-Components Structure: Ability, Schooling, and the Economic Success of Brothers", International Economic Review, 16 (1975), 422-449.

[14] "Chamberlain, G. and Griliches, Z. "More on Brothers" in Taubman, P. (ed.), Kinometrics: The Determinants of Socio-economic Success Within and Between Families (Amsterdam: North Holland Publishing Company, 1977).

[15] Cox, D. R. Analysis of Binary Data (London, Methuen, 1970).

[16] deFinetti, B. "La Prévision: Les Lois Logiques, ses Sources Subjectives", Annales de l'Institut Henri Poincaré, 7 (1937). English translation in Kyburg, H. E. and Smokler, H. E. (eds.), Studies in Subjective Probability (New York: Wiley, 1964).

[17] Fisher, R. A. "The Logic of Inductive Inference", Journal of the Royal Statistical Society, Series B, 98 (1935), 39-54.

[18] Goodman, L. "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications", Journal of the American Statistical Association, 65 (1970), 226-256.

[19] Goodman, L "A Modified Multiple Regression Approach to the Analysis of Dichotomous Variables", American Sociological Review, 37 (1972), 28-46.

[20] Haberman, S. J. The Analysis of Frequency Data (Chicago: University of Chicago Press, 1974).

[21] Hall, B. H. "A General Framework for Time Series-Cross Section Estimation", Annales de l'INSEE, 30-31 (1978), 177-202.

[22] Heckman, J. J. "Statistical Models for Discrete Longitudinal Data", (University of Chicago, 1978).

[23] Kalbfleisch, J. D. and Sprott, D. A. "Application of Likelihood Methods to Models Involving Large Numbers of Parameters", Journal of the Royal Statistical Society, Series B, 32 (1970), 175-208.

[24] Kiefer, J. and Wolfowitz, J. "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters", Annals of Mathematical Statistics, 27 (1956), 887-906.

[25] McFadden, D. "Conditional Logit Analysis of Qualitative Choice Behavior" in Zarembka, P (ed.), Frontiers in Econometrics (New York: Academic Press, 1974).

[26] Morgan, J. N. et al. A Panel Study of Income Dynamics (Ann Arbor: Institute for Social Research, 1972).

[27] Mundlak, Y. "Empirical Production Function Free of Management Bias", Journal of Farm Economics, 43 (1961), 44-56.

[28] Mundlak, Y. "Estimation of Production and Behavioral Functions from a Combination of Cross-Section and Time-Series Data" in Christ, C. et al., Measurement in Economics (Stanford University Press, 1963).

[29] Mundlak, Y. "On the Pooling of Time Series and Cross Section Data", Econometrica, 46 (1978), 69-85.

[30] Nerlove, M. and Press, S. J. "Multivariate Log-Linear Probability Models for the Analysis of Qualitative Data", Center for Statistics and Probability Discussion Paper, No. 1 (Northwestern University, 1976).

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

[32] Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests (Copenhagen: Danmarks Paedagogiske Institut, 1960).

[33] Rasch, G. "On General Laws and the Meaning of Measurement in Psychology", Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 4 (1961), 321-333.