Chapter 5. Nonlinear and Related Panel Data Models

William Greene[1], Qiushi Zhang[2]

Abstract

The panel data linear regression model has been exhaustively studied in a vast literature that originates with Nerlove (1966) and spans the entire range of empirical research in Economics. This chapter describes the application of panel data methods to some nonlinear models such as binary choice and nonlinear regression, where the treatment has been more limited. Some of the methodology of linear panel data modeling can be carried over directly to nonlinear cases, while other aspects must be reconsidered. The ubiquitous fixed effects linear model is the most prominent case of this latter point. Familiar general issues including dealing with unobserved heterogeneity, fixed and random effects, initial conditions and dynamic models are examined here. Practical considerations such as incidental parameters, latent class and random parameters models, robust covariance matrix estimation, attrition, and maximum simulated likelihood estimation are considered. We review several practical specifications that have been developed around a variety of specific nonlinear models including binary and ordered choice, models for counts, nonlinear regressions, stochastic frontier and multinomial choice models.

Keywords: Binary Choice; Correlated Random Effects; Count Data; Dynamic Model; Fixed Effects; Heterogeneity; Incidental Parameters; Latent Class; Nonlinear Model; Panel Data; Random Effects

[1] Department of Economics, Stern School of Business, New York University
[2] Department of Economics, Duke University

5.1 Introduction

This chapter explores the intersection of two topics: nonlinear modeling and the treatment of panel data. Superficially, nonlinearity merely compels parameter estimation to use methods more involved than linear least squares. But, in many ways, *nonlinear models* are qualitatively different from linear ones – it is more than a simple matter of functional form. (I.e., nonlinearity is more than the simple difference between [$y = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$] and [$y = h(\boldsymbol{\beta},\mathbf{x}) + \varepsilon$].) Analysis often involves reinterpreting the objects of estimation Most of the received analysis of *panel data models* focuses on the treatment of unobserved heterogeneity. The full set of issues that appear in the (fixed or random effects) linear panel data regression appear in more complicated forms in nonlinear contexts.

The application of panel data methods to nonlinear models is a subarea of *microeconometrics*. (See Cameron and Trivedi (2005).) The analyst is interested in the behavior of individual units, such as people, households, firms, etc., where the typical model examines the outcome of an individual decision. We are interested in *nonlinear models*, using methods and models defined for *panel data*. To cite a template example, many researchers have analyzed health outcomes data, including *health satisfaction* (a discrete, ordered, categorical outcome), *retirement* (a discrete, binary outcome) and *health system utilization* (usually a discrete count of events), in the context of the *German Socioeconomic Panel* data set or the *European Community Household Panel* data set. These are repeated surveys of a large number of households gathered over a number of years. We are interested in models and methods that extend beyond linear regression.

Many of the *longitudinal data sets* that are used in contemporary microeconometric research provide researchers with rich studies of outcomes such as fertility, health decisions and outcomes, income, wealth and labor market experiences, subjective health and well being and consumption decisions. Most of these variables are discrete or discontinuous and not amenable to conventional linear regression modeling. The literature provides a wide variety of theoretical and empirical frameworks for nonlinear modeling, such as *binary*, *ordered* and *multinomial choice*, *censoring*, *truncation*, *attrition* and *sample selection*. These nonlinear models have adapted econometric methods to more complicated settings than linear regression and simple instrumental variable (IV) techniques. This chapter will provide an overview of these applications. Some theoretical developments are presented to give context to the practical implementations. The particular interest is in the extension of 'panel data' methods to these nonlinear models that have long provided the econometric platforms. This includes development of treatments of *fixed and random effects* models and *random parameter* forms for *unobserved heterogeneity*, models that involve *dynamic effects* and *sample attrition*. We are also interested in the theoretical issues and complications that define this area of analysis and in a number of specific kinds of applications such as *random utility based discrete choice models*, *random parameter and latent class models* and applications of the *stochastic frontier model*.

Overall, we are interested in a general arena of models that have appeared in empirical applications. The treatment leans more toward the parametric treatments than some recent treatments such as Honoré (2002) and Arellano and Hahn (2006). Some essential theory is presented, as well as a variety of applications. The selection of topics in this survey is wider than in some others (e.g., Honore (2002, 2013), Honoré and Kesina (2017)), but not exhaustive. A large literature on deeper theory (see, e.g., Wooldridge (2010)) and results that advance the fundamental methodology, such as set vs. point identification in discrete choice models (e.g., Chesher (2013)) is left for more advanced treatments. Many additional practical results appear in Cameron and Trivedi (2005). One of the important features of the

analysis described here is that familiar results for the linear model cannot be carried over to nonlinear ones. We begin in Section 5.2 by examining the interpretation of parameters and partial effects in nonlinear models. Specific aspects of panel data modeling, notably heterogeneity under different assumptions, the incidental parameters problem and dynamic effects are treated in Section 5.3. Section 5.4 describes features that are common to most nonlinear panel data models. Applications, including the essential layout of longitudinal data sets are treated in Sections 5.5 and 5.6. The last two sections also consider the problem of attrition and issues related to robust estimation and inference.

The following notation is used throughout the survey:

Panel Data Set Dimensions:

i = index for observations (individuals),
t = index for periods, or replications,
n = sample size; $i = 1,\ldots,n$,
T_i = number of observations in group i, not assumed constant,
$N = \sum_{i=1}^{n} T_i$;

Panel Data:

$y_{i,t}$ = variable of interest in the 'model,' may be one or more than one outcome,
$\mathbf{x}_{i,t}$ = exogenous variables = $(1,\mathbf{z}_{i,t}')'$, column vectors,
$\mathbf{y}_i$ = $(y_{i,1},\ldots,y_{i,Ti})'$ = sequence of realizations of $y_{i,t}$,
$\mathbf{X}_i$ = sequence of observations on exogenous variables, $T_i \times K$; $\mathbf{x}_{i,t}'$ = row t of $\mathbf{X}_i$,
$\mathbf{d}(i) = \mathbf{d}(i)_{j,t} = \mathbf{d}_i = \mathbf{1}[j = i, t = 1,\ldots,T_i]$ = sample length dummy variable for i,
$\mathbf{i}$ = constant term = column of ones;

Functions:

$\phi(t)$, $\Phi(t)$ = standard normal pdf, cdf,
$\Lambda(t)$ = logistic cdf,
$N[\mu,\sigma^2]$ = normal distribution,
$N^+[\mu,\sigma^2]$ = truncated at zero normal distribution = $|u|$ where $u \sim N[0,\sigma^2]$,
$f(c|\mathbf{X})$ = conditional density of c given $\mathbf{X}$,
$f(c:\sigma)$ or $f(c|\mathbf{X}:\sigma)$ = density of variable that involves parameter σ,
$f(y_{i,t}|\ldots)$ = density for $y_{i,t}$, used generally for the model for $y_{i,t}$,
$\mathbf{1}$[condition] = 1 if condition is true, 0 if false,
$E[c]$ = expected value,
$E_c[g(x,c)]$ = $h(x)$ = expected value over c;

Model Components:

$\varepsilon_{i,t}$ = general idiosyncratic disturbance in model,
c_i = unobserved heterogeneity, usually univariate,
α_i = fixed effects version of c_i, $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_n)'$,
η_i = $\exp(\alpha_i)$,
u_i = random effects version of c_i,
$\boldsymbol{\beta}$ = slope vector in index function model, appears as $\boldsymbol{\beta}'\mathbf{x}_{i,t} = \pi + \boldsymbol{\gamma}'\mathbf{z}_{i,t}$,
$\boldsymbol{\gamma}$ = subvector of $\boldsymbol{\beta}$ omitting the constant,
π = constant term, $\boldsymbol{\beta} = (\pi,\boldsymbol{\gamma}')'$,
$\phi_{i,t}$ = $\exp(\boldsymbol{\gamma}'\mathbf{z}_{i,t})$,

$\lambda_{i,t}$ = $\exp(\boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i) = \eta_i\phi_{i,t}$, $c_i = \alpha_i$ or u_i,
$\boldsymbol{\theta}$ = one or more ancillary parameters in parametric model,
σ^2 = variance of u_i in random effects model,
σ_ε^2 = variance of $\varepsilon_{i,t}$ in random index function model.

5.2 Nonlinear Models

The *linear panel data regression model* is

$$y_{i,t} = \boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i + \varepsilon_{i,t},\ i = 1,\ldots,n,\ t = 1,\ldots,T_i,$$

where $y_{i,t}$ is the outcome variable of interest, $\mathbf{x}_{i,t}$ is a vector of time varying and possibly time invariant variables, also possibly including $y_{i,t-1}$, c_i is unobserved time invariant heterogeneity that is independent of $\varepsilon_{i,t}$ and $\varepsilon_{i,t}$ is a classical disturbance. Since c_i is unobserved, there is no coefficient or scale attached to it. The 'linearity' of the model relates to (1) the way that the natural estimator of the parameter vector of interest, $\boldsymbol{\beta}$, is computed, that is, by using some variant of linear least squares or instrumental variables (IV) to solve a set of linear equations and (2) the way that the unobserved heterogeneity, c_i enters the function of interest, here the conditional mean function.

We are interested in models in which the function of interest, such as a conditional mean, is intrinsically nonlinear. This would include, for example, the *Poisson regression model*:

$$\begin{aligned}&\text{(Data Generating Process)} \quad \text{Prob}(y_{i,t} = j \mid \mathbf{x}_{i,t}, c_i) = \left[\exp(-\lambda_{i,t})\lambda_{i,t}^{j}\right] / j!;\\ &\text{(Function of Interest)} \quad E[y_{i,t} \mid \mathbf{x}_{i,t}, c_i] = \lambda_{i,t} = \exp(\boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i).\end{aligned}$$

(See Cameron and Trivedi (2005) and Greene (2018).) Most models of interest in this area involve missing data in which $y_{i,t}$, the outcome of some underlying process involving $\boldsymbol{\beta}$ as well as c_i, passes through a filter between the data generating process (DGP) and the observed outcome.[1] The most common example is the familiar (semiparametric) *random effects binary choice model*:

$$\begin{aligned}&\text{(Random Utility DGP)} \quad y_{i,t}^* = \boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i + \varepsilon_{i,t},\ y_{i,t}^* = \text{unobserved random utility};\\ &\text{(Revealed Preference)} \quad y_{i,t} = \mathbf{1}[y_{i,t}^* > 0].\end{aligned}$$

(The model becomes parametric when distributions are specified for c_i and $\varepsilon_{i,t}$.) In this case, the nonlinear function of interest is

$$\text{Prob}[y_{i,t} = 1 \mid \mathbf{x}_{i,t}, c_i] = F\left[(\boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i) / \sigma_\varepsilon\right].$$

where $F[.]$ is the cdf of $\varepsilon_{i,t}$. This example also fits into category (1).[2] It will not be possible to use least squares or IV for parameter estimation; (2) Some alternative to group mean deviations or first differences

[1] Nearly all of the models listed above in Section 4.6 are of this type.
[2] In cases in which the function of interest is a nonlinear conditional mean function, it is sometimes suggested that a 'linear approximation' to quantities of intrinsic interest, such as partial effects, be obtained by simply using linear

is needed to proceed with estimation in the presence of the unobserved, heterogeneity. In the most familiar cases, the issues center on persuasive forms of the model and practicalities of estimation, such as how to handle heterogeneity in the form of fixed or random effects. The linear form of the model involving the unobserved heterogeneity is a considerable advantage that will be absent from all of the extensions we consider here. A panel data version of the *stochastic frontier model* (Aigner, Lovell and Schmidt (1977)) is

$$\begin{aligned} y_{i,t} &= \beta'\mathbf{x}_{i,t} + c_i + v_{i,t} - u_{i,t} \\ &= \beta'\mathbf{x}_{i,t} + c_i + \varepsilon_{i,t}, \end{aligned}$$

where $v_{i,t} \sim N[0,\sigma_v^2]$ and $u_{i,t} \sim N^+(0,\sigma_u^2)$. (See Greene (2004a, 2004c).) Superficially, this is a linear regression model with a disturbance that has a skew normal distribution,

$$f(\varepsilon_{i,t}) = \frac{2}{\sigma}\phi\left(\frac{\varepsilon_{i,t}}{\sigma}\right)\Phi\left(\frac{-\lambda\varepsilon_{i,t}}{\sigma}\right),\ \lambda = \frac{\sigma_u}{\sigma_v},\ \sigma^2 = \sigma_v^2 + \sigma_u^2.$$

In spite of the apparent linearity, the preferred estimator is (nonlinear) maximum likelihood. A second, similar case is Graham et al.'s (2015) quantile regression model, $y_{i,t}(\tau) = \beta(\tau, c_i)'\mathbf{x}_{i,t} + \varepsilon(\tau)_{i,t}$. (See Geraci and Bottai (2007).) The model appears to be intrinsically linear. However, the preferred estimator is, again, not linear least squares – it is usually based on a linear programming approach. For present purposes, in spite of appearances this model is intrinsically nonlinear.

5.2.1 Coefficients and Partial Effects

The feature of interest will usually be a nonlinear function, $g(\mathbf{x}_{i,t},c_i)$ derived from the probability distribution, $f(y_{i,t}|\mathbf{x}_{i,t},c_i)$, such as the conditional mean function, $E[y_{i,t}|\mathbf{x}_{i,t},c_i]$ or some derivative function such as a probability in a discrete choice model, $\text{Prob}(y_{i,t} = j|\mathbf{x}_{i,t},c_i) = F(\mathbf{x}_{i,t},c_i)$. In general, the function will involve structural parameters that are not, themselves, of primary interest; $g(\mathbf{x}_{i,t},c_i) = g(\mathbf{x}_{i,t},c_i : \theta)$ for some vector of parameters, θ. The partial effects will then be $\text{PE}(\mathbf{x},c) = \delta(\mathbf{x},c : \theta) = \partial g(\mathbf{x},c:\theta)/\partial\mathbf{x}$. In the probit model, the function of interest is the probability, and the relevant quantity is a partial effect,

$$\text{PE}(\mathbf{x},c) = \partial\text{Prob}(y_{i,t} = 1|\mathbf{x},c)/\partial\mathbf{x}.$$

Estimation of partial effects is likely to be the main focus of the analysis. Computation of partial effects will be problematic even if θ is estimable in the presence of c, because c is unobserved and the distribution of c remains to be specified. If enough is known about the distribution of c, computation at a specific value, such as the mean, may be feasible. The *partial effect at the average* (of c) would be

$$\text{PEA}(\mathbf{x}) = \delta(\mathbf{x},E[c] : \theta) = \partial\text{Prob}(y_{i,t} = 1|\mathbf{x}_{i,t}, E[c_i])/\partial\mathbf{x},$$

while the average (over c) partial effect would be

---

least squares. See, e.g., Angrist and Pischke (2009) for discussion of the canonical example, the binary probit model.

$$\text{APE}(\mathbf{x}) = E_c[\delta(\mathbf{x},c:\theta)] = E_c[\partial \text{Prob}(y_{i,t}=1|\mathbf{x},c)/\partial \mathbf{x}].$$

One might have sufficient information to characterize $f(c_i|\mathbf{x}_{i,t})$ or $f(c_i|\mathbf{X}_i)$. In this case, the PEA could be based on $E[c_i|\mathbf{X}_i]$ or the APE might be based on the conditional distribution, rather than the marginal. Altonji and Matzkin (2005) identify this as a *local average response* (LAR, i.e., local to the subpopulation associated with the specific realization of $\mathbf{X}_i$). If c_i and $\mathbf{X}_i$ are independent, then the conditional and marginal distributions will be the same and the LAR and APE will also be the same.

In *single index function models*, in which the covariates enter the model in a linear index function, $\boldsymbol{\beta}'\mathbf{x}_{i,t}$, the partial effects usually simplify to a multiple of $\boldsymbol{\beta}$;

$$\text{PEA}(\mathbf{x}) = \boldsymbol{\beta}\, h(\boldsymbol{\beta}'\mathbf{x},E[c])], \text{ where } h(\boldsymbol{\beta}'\mathbf{x},E[c])] = \frac{\partial g(t,E[c])}{\partial t}\Big|_{t=\boldsymbol{\beta}'\mathbf{x}},$$

$$\text{APE}(\mathbf{x}) = \boldsymbol{\beta}\, E_c[h(\boldsymbol{\beta}'\mathbf{x},c)].$$

For the normalized ($\sigma_\varepsilon = 1$) probit model, $\text{Prob}(y_{i,t}=1|\mathbf{x}_{i,t},c_i) = \Phi(\boldsymbol{\beta}'\mathbf{x}_{i,t}+c_i)$. Then, $g(\boldsymbol{\beta}'\mathbf{x},c) = \Phi(\boldsymbol{\beta}'\mathbf{x}+c)$ and $h(\boldsymbol{\beta}'\mathbf{x},c) = \boldsymbol{\beta}\phi(\boldsymbol{\beta}'\mathbf{x}+c)$. The coefficients have the same signs as partial effects, but their magnitude may be uninformative;

$$\text{APE}(\mathbf{x}) = \boldsymbol{\beta}\int_c \phi(\boldsymbol{\beta}'\mathbf{x}+c)dF(c\,|\,\mathbf{x}).$$

To complete this example, if $c \sim N[0,\sigma^2]$ and $\varepsilon \sim N[0,1^2]$. Then, $y^* = \boldsymbol{\beta}'\mathbf{x} + c + \varepsilon = \boldsymbol{\beta}'\mathbf{x} + w$, where $w \sim N[0,1+\sigma^2]$. It follows that

$$\text{Prob}[y=1|\mathbf{x},c] = \text{Prob}(\varepsilon \le \boldsymbol{\beta}'\mathbf{x}+c) = \Phi(\boldsymbol{\beta}'\mathbf{x}+c),$$

$$\text{Prob}(y=1|\mathbf{x}) = \text{Prob}(w \le \boldsymbol{\beta}'\mathbf{x}) = \Phi(\boldsymbol{\beta}'\mathbf{x}/\sigma_w) = \Phi[\boldsymbol{\beta}'\mathbf{x}/(1+\sigma^2)^{1/2}].$$

Then $\text{PEA}(\mathbf{x}) = \boldsymbol{\beta}\,\phi(\boldsymbol{\beta}'\mathbf{x}+0) = \boldsymbol{\beta}\,\phi(\boldsymbol{\beta}'\mathbf{x})$ while

$$\begin{aligned}\text{APE}(\mathbf{x}) &= \boldsymbol{\beta}\int_c \phi(\boldsymbol{\beta}'\mathbf{x}+c)(1/\sigma)\phi\left(c/\sigma\right)dc \\ &= (\boldsymbol{\beta}/(1+\sigma^2)^{1/2}) \times \phi[\boldsymbol{\beta}'\mathbf{x}/(1+\sigma^2)^{1/2}] = \delta\,\phi(\boldsymbol{\delta}'\mathbf{x}).\end{aligned}$$

5.2.2 Interaction Effects

Interaction effects arise from second order terms; $y_{i,t} = \beta x_{i,t} + \gamma z_{i,t} + \delta x_{i,t}z_{i,t} + c_i + \varepsilon_{i,t}$, so that

$$\text{APE}(x|z) = E_c\{\partial E[y|x,z,c]/\partial x\} = E_c[\partial(\beta x_{i,t} + \gamma z_{i,t} + \delta x_{i,t}z_{i,t} + c_i)/\partial x] = \beta + \delta z_{i,t}.$$

The *interaction effect* is $\partial \text{APE}(x|z)/\partial z = \delta$. What appear to be interaction effects will arise unintentionally in nonlinear index function models. Consider the nonlinear model, $E[y_{i,t}|x_{i,t},z_{i,t},c_i] = \exp(\beta x_{i,t} + \gamma z_{i,t} + c_i)$.

The average partial effect of $x|z$ is $\text{APE}(x|z) = E_c\{\partial E[y|x,z,c]/\partial x\} = \beta\exp(\beta x + \gamma z)E[\exp(c)]$. The second order (interaction) effect of z on the partial effect of x is $\beta\gamma\exp(\beta x + \gamma z)E[\exp(c)]$, which will generally be nonzero even in the absence of a second order term. The situation is worsened if an interaction effect is built into the model. Consider $E[y|x,z,c] = \exp(\beta x + \gamma z + \delta xz + c)$. The average partial effect is

$$\begin{aligned}\text{APE}(x|z) &= E_c\{\partial E[y|x,z,c]/\partial x\} \\ &= E[\exp(c)](\beta+\delta z)\exp(\beta x + \gamma z + \delta xz)].\end{aligned}$$

The interaction effect is, now,

$$\partial\text{APE}(x|z)/\partial z = E[\exp(c)]\exp(\beta x + \gamma z + \delta xz)]\{\delta + (\beta+\delta z)(\gamma + \delta x)\}.$$

The effect contains what seems to be the anticipated part plus an effect that clearly results from the nonlinearity of the conditional mean. Once again, the result will generally be nonzero even if δ equals zero. This creates a considerable amount of ambiguity about how to model and interpret interactions in a nonlinear model. (See Mandic, Norton and Dowd (2012), Pinar, Norton and Dowd (2012), Ai and Norton (2003) and Greene (2010a) for discussion.)

5.2.3 Identification through Functional Form

Results in nonlinear models may be identified through the form of the model rather than through covariation of variables. This is usually an unappealing result. Consider the triangular model of health satisfaction and SNAP (food stamp) program participation by Gregory and Deb (2015);

$$\begin{aligned}SNAP &= \boldsymbol{\beta}_S'\mathbf{x} + \boldsymbol{\delta}'\mathbf{z} + \varepsilon \\ HSAT &= \boldsymbol{\beta}_H'\mathbf{x} + \gamma SNAP + w.\end{aligned}$$

Note that $\mathbf{x}$ is the same in both equations. If $\boldsymbol{\delta}$ is nonzero, then this *linear simultaneous equations model* is identified by the usual rank and order conditions. Two stage least squares would likely be the preferred estimator of the parameters in the *HSAT* equation (assuming that *SNAP* is endogenous – that is, if ε and w are correlated). However, if $\boldsymbol{\delta}$ equals $\mathbf{0}$, the *HSAT* equation will fail the order condition for identification and be inestimable. But, the model in the application is not linear – *SNAP* is binary and *HSAT* is ordered and categorical – both outcome variables are discrete. In this case, the parameters *are* fully identified even if $\boldsymbol{\delta}$ equals $\mathbf{0}$. Maximum likelihood estimation of the full set of parameters is routine in spite of the fact that the regressors in the two equations are identical. The parameters are identified by the likelihood equations under mild assumptions (essentially that the Hessian of the full information log likelihood with respect to $(\boldsymbol{\beta}_S,\boldsymbol{\delta},\boldsymbol{\beta}_H,\gamma)$ is nonsingular at $\boldsymbol{\delta} = \mathbf{0}$). This is *identification by functional form*. The causal effect, γ is identified when $\boldsymbol{\delta} = \mathbf{0}$, even though there is no instrument ($\mathbf{z}$) that drives *SNAP* participation independently of the exogenous influences on *HSAT*. The authors note this, and suggest that the nonzero $\boldsymbol{\delta}$ (exclusion of $\mathbf{z}$ from the *HSAT* equations) is a good idea to "improve" identification, in spite of result.[3]

[3] Scott et al. (2009) who make the same observation. Rhine and Greene (2013) is a similar application. See also Filippini et al. (2018), Wilde (2000) and Mourifie and Meango (2014) for discussion of some special cases.

5.2.4 Endogeneity

In the linear regression model, $y_{i,t} = \alpha + \beta x_{i,t} + \delta z_{i,t} + \varepsilon_{i,t}$, there is little ambiguity about the meaning of endogeneity of x. There may be various theories to motivate it, such as omitted variables or heterogeneity, reverse causality, nonrandom sampling, and so on. But, in any of these events, the ultimate issue is tied to some form of covariation between $x_{i,t}$ (the observable) and $\varepsilon_{i,t}$ (the unobservable). Consider, instead, the Poisson regression model described above, where, now, $\lambda_{i,t} = \exp(\alpha + \beta x_{i,t} + \delta z_{i,t})$. For example, suppose $y_{i,t}$ equals hospital or doctor visits (a health outcome) and $x_{i,t}$ equals income. This should be a natural application of reverse causality. But, there is no mechanism within this Poisson regression model that supports the notion of endogeneity suggested above. The model leaves open the question of what (in the context of the model) is correlated with $x_{i,t}$ that induces the endogeneity. (See Cameron and Trivedi (2005, p. 687).) For this particular application, a common approach is to include the otherwise absent unobserved heterogeneity in the conditional mean function, as $\lambda_{i,t}|w_{i,t} = \exp(\beta x_{i,t} + \delta z_{i,t} + w_{i,t})$.

As a regression framework, the Poisson model has a shortcoming – it specifies the model for observed heterogeneity, but lacks a coherent specification for unobserved heterogeneity (a disturbance). The model suggested above is a *mixture model*. For the simpler case of exogenous x, the feasible empirical specification is obtained by analyzing

$$\text{Prob}(y_{i,t} = j \mid x_{i,t}, z_{i,t}) = \int_{w_{i,t}} \text{Prob}(y_{it} = j \mid x_{i,t}, z_{i,t}, w_{i,t}) dF(w_{i,t}).$$

This parametric approach would require a specification for $F(w)$. The traditional approach is a log-gamma that produces a closed form, the negative binomial model, for the unconditional probability. Recent applications use the normal distribution. A semiparametric approach could be taken as well if less is known about the distribution of w. This might seem less ad hoc than the parametric model, but the assumption of the Poisson distribution is not innocent at the outset. To return to the earlier question, a parametric approach to the endogeneity of $x_{i,t}$ would mandate a specification of the joint distribution of w and x, $F(w_{i,t},x_{i,t})$. For example, it might be assumed that $x_{i,t} = \boldsymbol{\theta}'\mathbf{f}_{i,t} + v_{i,t}$ where w and v are bivariate normally distributed with correlation ρ. This completes a mechanism for explaining how $x_{i,t}$ is endogenous in the Poisson model. This is precisely the approach taken in Gregory and Deb's *SNAP*/*HSAT* model shown earlier.

5.3 Panel Data Models

The objective of analysis is some feature(s) of the joint conditional distribution of a sequence of outcomes for individual *i*;

[5.3-1] $$f\left(y_{i,1}, y_{i,2}, ..., y_{i,T_i} \mid \mathbf{x}_{i,1}, \mathbf{x}_{i,2}, ..., \mathbf{x}_{i,T_i}, c_{i,1}, ..., c_{i,M}\right) = f\left(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{c}_i\right).$$

The sequence of random variables, $y_{i,t}$ is the outcome of interest. Each will typically be univariate, but need not be. In Riphahn, Wambach and Million's (2003) study, $\mathbf{y}_{i,t}$ consists of two count variables that jointly record health care system utilization, counts of doctor visits and counts of hospital visits. In order to have a compact notation, in [5.3-1], $\mathbf{y}_i$ denotes a column vector in which the observed outcome $y_{i,t}$, is either univariate or multivariate – the appropriate form will be clear in context. The observed conditioning effects are a set of time varying and time invariant variables, $\mathbf{x}_{i,t}$. (See, e.g., *EDUC* and *FEMALE*, respectively in Table 5.2 below.) The matrix $\mathbf{X}_i$ is $T_i \times K$ containing the K observed variables $\mathbf{x}_{i,t}$ in each row. To accommodate a constant term, $\mathbf{X}_i = [\mathbf{i}, \mathbf{Z}_i]$.

For now, $\mathbf{x}_{i,t}$ is assumed to be strictly exogenous. The scalars, $c_{i,m}$ are unobserved, time invariant heterogeneity. The presence of the time invariant, unobserved heterogeneity is the signature feature of a 'panel data model.' For present purposes, with an occasional exception noted later, it will be sufficient to work with a single unobserved variate, c_i.

Most cases of practical interest depart from an initial assumption of *strict exogeneity*. That is, for the marginal distribution of $y_{i,t}$, we have

[5.3-2] $$f(y_{i,t} \mid \mathbf{x}_{i,1}, \mathbf{x}_{i,2}, ..., \mathbf{x}_{i,T_i}, c_i) = f(y_{i,t} \mid \mathbf{x}_{i,t}, c_i).$$

That is, after conditioning on $(\mathbf{x}_{i,t}, c_i)$, $\mathbf{x}_{i,r}$ for $r \neq t$ contains no additional information for the determination of outcome $y_{i,t}$.[4] Assumption [5.3-2] will suffice for nearly all of the applications to be considered here. The exception that will be of interest below will be dynamic models, in which, perhaps, *sequential exogeneity*,

[5.3-3] $$f(y_{i,t} \mid \mathbf{x}_{i,1}, \mathbf{x}_{i,2}, ..., \mathbf{x}_{i,T_i}, c_i) = f(y_{i,t} \mid \mathbf{x}_{i,1}, \mathbf{x}_{i,2}, ..., \mathbf{x}_{i,t}, c_i),$$

is sufficient.

Given [5.3-2], the natural next step is to characterize $f(\mathbf{y}_i \mid \mathbf{X}_i, c_i)$. The *conditional independence assumption* adds that $y_{i,t} \mid \mathbf{x}_{i,t}, c_i$ are independent within the cross section group, $t = 1, \ldots, T_i$. It follows that

[5.3-4] $$f(y_{i,1}, y_{i,2}, ..., y_{i,T_i} \mid \mathbf{X}_i, c_i) = \prod_{t=1}^{T_i} f(y_{i,t} \mid \mathbf{x}_{i,t}, c_i).$$

The large majority of received applications of nonlinear panel data modeling are based on fully parametric specifications. With respect to the model above, this adds a sufficient description of the DGP for c_i that estimation can proceed.

[4] For some purposes, only the restriction on the derived function of interest, such as the conditional mean, $E[y_{i,t} \mid \mathbf{X}_i, c_i] = E[y_{i,t} \mid \mathbf{x}_{i,t}, c_i]$ is necessary. (See Wooldridge (1995).) Save for the linear model, where this is likely to follow by simple construction, obtaining this result without [5.3-2] is likely to be difficult. That is, asserting the mean independence assumption while retaining the more general [5.3-1] is likely to be difficult.

5.3.1 Objects of Estimation

In most cases, the platform for the analysis is the distribution for the observed outcome variable in [5.3-1]. The desired target of estimation is some derivative of that platform, such as a conditional mean or variance, a probability function defined for an event, a median, or some other conditional quantile, a hazard rate or a prediction of some outcome related to the variable of interest. For convenience, we restrict attention to a univariate case. In many applications, interest will center on some feature of the distribution of y_{it}, $f(y_{it}|\mathbf{x}_{i,t}, c_i)$, such as the conditional mean function, $g(\mathbf{x}, c) = E[y|\mathbf{x}, c]$. The main object of estimation will often be partial effects, $\boldsymbol{\delta}(\mathbf{x}, c) = \partial g(\mathbf{x}, c)/\partial \mathbf{x}$, for some specific value of $\mathbf{x}$ such as $E[\mathbf{x}]$ if $\mathbf{x}$ is continuous, or $\Delta(\mathbf{x},d,c) = g(\mathbf{x},1, c) - g(\mathbf{x},0, c)$ if the margin of interest relates to a binary variable.

A strictly nonparametric approach to $\boldsymbol{\delta}(\mathbf{x},c)$ offers little promise outside the narrow case in which no other variables confound the measurement.[5] Without at least some additional detail about distribution of c, there is no obvious way to isolate the effect of c from the impact of the observable $\mathbf{x}$. Since c is unobserved, as it stands, $\boldsymbol{\delta}$ is inestimable without some further assumptions. For example, if it can be assumed that c has mean μ_c (zero, for example) and is independent of $\mathbf{x}$, than a partial effect at this mean, $\text{PEA}(\mathbf{x}, \mu) = \boldsymbol{\delta}(\mathbf{x}, \mu)$ may be estimable. If the distribution of c can be more completely specified, then it may be feasible to obtain an average partial effect,

$$\text{APE}(\mathbf{x}) = E_c[\boldsymbol{\delta}(\mathbf{x}, c)].$$

Panel data modeling is complicated by the presence of unobserved heterogeneity in estimation of parameters and functions of interest. This situation is made yet worse because of the nonlinearity of the target feature. In most cases, the results gained from the linear model are not transportable. Consider the linear model with strict exogeneity and conditional independence, $E[y_{it}|\mathbf{x}_{it},c_i] = \boldsymbol{\beta}'\mathbf{x}_{it} + c_i + \varepsilon_{it}$. Regardless of the specification of $f(c)$, the desired partial effect is $\boldsymbol{\beta}$. Now consider the (nonlinear) probit model,

(DGP)	$y_{i,t}^*$	$= \boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i + \varepsilon_{i,t}$, $\varepsilon_{i,t}\|\mathbf{x}_{i,t},c_i \sim N[0,1^2]$,
(Observation)	$y_{i,t}$	$= \mathbf{1}[y_{i,t}^* > 0]$,
(Function of Interest)	$\text{Prob}(y_{i,t} = 1\|\mathbf{x}_{i,t},c_i)$	$= \Phi(\boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i)$.

With sufficient assumptions about the generation of c_i, such as $c_i \sim N[0,\sigma^2]$, estimation of $\boldsymbol{\beta}$ will be feasible. The relevant partial effect is now

$$\boldsymbol{\delta}(\mathbf{x},c) = \partial\Phi(\boldsymbol{\beta}'\mathbf{x} + c)/\partial\mathbf{x} = \boldsymbol{\beta}\phi(\boldsymbol{\beta}'\mathbf{x} + c).$$

If $f(c)$ is sufficiently parameterized, then an estimator of $\text{PE}(\mathbf{x}|\hat{c}) = \boldsymbol{\beta}\phi(\boldsymbol{\beta}'\mathbf{x} + \hat{c})$ such as

$$\text{PEA}(\mathbf{x}|\hat{c}) = \boldsymbol{\beta}\phi[\boldsymbol{\beta}'\mathbf{x} + \hat{E}(c)]$$

[5] If there are no $\mathbf{x}$ variables in $E[y|\mathbf{x}, c]$, then with independence of d and c and binary y, there may be scope for nonparametric identification.

may be feasible. If c can be assumed to have a fixed conditional mean, $\mu_c = E[c|\mathbf{x}] = 0$, and if $\mathbf{x}$ contains a constant term, then the estimator might be PEA($\mathbf{x}$,0) = $\boldsymbol{\beta}\phi(\boldsymbol{\beta}'\mathbf{x})$. This is not sufficient to identify the average partial effect. If it is further assumed that c is normally distributed (and independent of $\mathbf{x}$) with variance σ^2, then,

$$\begin{aligned}\text{APE}(\mathbf{x}) &= \boldsymbol{\beta}/(1+\sigma^2)^{1/2}\,\phi[\boldsymbol{\beta}'/(1+\sigma^2)^{1/2}\mathbf{x}] \\ &= \boldsymbol{\beta}\,(1-\rho)^{1/2}\,\phi[\boldsymbol{\beta}'(1-\rho)^{1/2}\mathbf{x}] \\ &= \boldsymbol{\gamma}\,\phi(\boldsymbol{\gamma}'\mathbf{x}),\end{aligned}$$

where ρ is the *intragroup correlation*, $\text{Corr}[(\varepsilon_{i,t} + c_i),(\varepsilon_{i,s} + c_i)] = \sigma^2/(1+\sigma^2)$. In the context of this model, what will be estimated with a random sample (of panel data)? Will APE and PEA be substantively different? In the linear case, PEA($\mathbf{x}|\hat{c}$) and APE($\mathbf{x}$) will be the same $\boldsymbol{\beta}$. It is the nonlinearity of the function that implies that they might be different here..

If c_i were observed data, then fitting a probit model for $y_{i,t}$ on ($\mathbf{x}_{i,t}$,c_i) would estimate ($\boldsymbol{\beta}$,1). We have not scaled c_i, but since we are treating c_i as observed data (and uncorrelated with $\mathbf{x}_{i,t}$), we can use, instead, $c^* = c_i/s_c$ as the variable, and attach the parameter σ_c to c_i^*. So a fully specified parametric model might estimate ($\boldsymbol{\beta}$,σ_c). If c_i were simply ignored, we would fit a 'pooled' probit model. The true underlying structure is $y_{i,t} = \mathbf{1}\{\boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i + \varepsilon_{i,t} > 0|\varepsilon_{i,t} \sim N[0,1^2]\}$. The estimates, shown above, would reveal $\boldsymbol{\gamma} = \boldsymbol{\beta}(1-\rho)^{1/2}$. Each element of $\boldsymbol{\gamma}$ is an attenuated (biased toward zero) version of its counterpart in $\boldsymbol{\beta}$. If the model were linear, then omitting a variable that is uncorrelated with the included $\mathbf{x}$, would not induce this sort of 'omitted variable bias.' Conclude that the pooled estimator estimates $\boldsymbol{\gamma}$ while the MLE estimates ($\boldsymbol{\beta}$,σ_c), and the attenuation occurs even if $\mathbf{x}$ and c are independent.

An experiment based on 'real' data will be suggestive. The data in Table 5.1 below are a small subsample from the data used in Riphahn et al. (2003).[6] The sample contains 27,326 household/year observations in 7,293 groups ranging in size from one to seven. We have fit simple pooled and panel probit models based on

$$Doctor_{i,t}^* = \beta_1 + \beta_2 Age_{i,t} + c_i + \varepsilon_{i,t};\ Doctor = \mathbf{1}[Doctor_{i,t}^* > 0]$$

where *Doctor* = $\mathbf{1}$[*Doctor Visits* > 0]. The results are

$$\begin{aligned}&\text{(Pooled)} && Doctor_{i.t}^* = -0.37176 + 0.01625 Age_{i,t} \\ &\text{(Panel)} && Doctor_{i,t}^* = -0.53689 + 0.02338 Age_{i,t} + 0.90999 c_i^*,\end{aligned}$$

[6] The original data set may be found at the Journal of Applied Econometrics data archive, **http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/**. The raw data set contains variables INCOME and HSAT (self reported health satisfaction) that contain a few anomalous values. In the 27,326 observations, three values of income were reported as zero. The minimum of the remainder was 0.005. These three values were recoded to 0.0015. The health satisfaction variable is an integer, 0,..,10. In the raw data, 40 observations were recorded between 6.5 and 7.0. These 40 values were rounded up to 7.0. The data set used here, with these substitutions is at **http://people.stern.nyu.edu/wgreene/text/healthcare.csv**. Differences between estimators computed with the uncorrected and corrected values are trivial.

where c_i*. is normalized to have variance 1.[7] The estimated value of $\rho = \sigma^2/(1+\sigma^2)$ is 0.45298, so the estimated value of σ is 0.90999. The estimator of the attenuation factor, $(1 - \rho)^{1/2}$, is 0.73961. Based on the results above, then, we obtain the estimate of γ based on the panel model, 0.02338×0.73961 = 0.01729. The finite sample discrepancy is about 6%. The average value of *Age* is 43.5 years. The average partial effects based on the pooled model and the panel model, respectively, would be

$$\begin{aligned}
&\text{(Pooled)} \quad \text{APE}(Age{:}\gamma) = 0.01625 \times \phi(-0.37176 + 0.01625\times 43.5) = 0.00613\\
&\text{(Panel)} \quad \text{APE}(Age{:}\beta,\sigma) = 0.02338(1 - .45298)^{1/2} \times\\
&\qquad\qquad \phi[(1 - 0.45298)^{1/2}(-0.53689 + 0.02338\times 43.5)] = 0.00648.
\end{aligned}$$

The estimate of APE(*Age*:γ) should not be viewed as PEA(*Age*,*E*[*c*]) = PEA(*Age*,0). That estimator would be PEA(*Age*,0:β,σ) = $0.02338 \times \phi(-0.53689 + 0.02338\times 43.5) = 0.008312$.[8] This estimator seems to be misleading. Finally, simple least squares estimation produces

$$\text{(Linear PM)} \quad Doctor_{i,t} = 0.36758 + 0.00601 Age_{i,t} + e_{i,t}.$$

This appears to be a reasonable approximation.[9]

Most situations to be considered in the subject of this chapter focus on nonlinear models such as the probit or Poisson regression, and pursue estimates of appropriate partial effects (or 'causal' effects) in many cases. As we will see in Section 5.6, there are a variety of situations in which something other than partial effects is interest. In the stochastic frontier model,

$$\begin{aligned}
y_{i,t} &= \alpha + \boldsymbol{\gamma}'\mathbf{z}_{i,t} + c_i + v_{i,t} - u_{i,t},\\
&= \alpha + \boldsymbol{\gamma}'\mathbf{z}_{i,t} + c_i + \varepsilon_{i,t},
\end{aligned}$$

the object of interest is an estimator of the inefficiency term, $u_{i,t}$. The estimator used is $\hat{u}_{i,t} = E_c[E[u_{i,t} \mid \varepsilon_{i,t}]]$. The various panel data formulations focus on the role of heterogeneity in the specification and estimation of the inefficiency term. In the analysis of individual data on multinomial choice, the counterpart to 'panel data modeling' in many studies is the *stated choice experiment*. The random utility based multinomial logit model with heterogeneity takes the form

$$\text{Prob}[Choice_{i,t} = j] = \frac{\exp(\alpha_{i,j} + \boldsymbol{\gamma}'\mathbf{z}_{i,t,j})}{1 + \sum_{j=1}^{J} \exp(\alpha_{i,j} + \boldsymbol{\gamma}'\mathbf{z}_{i,t,j})}, \; j = 1,\ldots,J.$$

[7] The model was estimated as a standard 'random effects probit model' using the Butler and Moffitt (1982) method. The estimate of σ was 0.90999. With this in hand, the implied model is as shown above. When the model is estimated in precisely that form ($\boldsymbol{\beta}'\mathbf{x} + \sigma c$*) using maximum simulated likelihood, the estimates are 0.90949 for σ and
(-0.53688,0.02338) for $\boldsymbol{\beta}$. Quadrature and simulation give nearly identical results, as expected.

[8] The slope in the OLS regression of *Doctor* on (1,*Age*) is 0.00601. This suggests, as observed elsewhere, that to the extent OLS estimates any defined quantity in this model, it will likely resemble APE($\mathbf{x}$).

[9] There is no econometric framework available within which it can be suggested that the OLS slope is a *consistent* estimator of an average partial effect (at the means, for example). It just 'works' much of the time.

Some applications involve 'mixed logit' modeling, in which not only the alternative specific constants, $\alpha_{i,j}$ but also the marginal utility values, $\boldsymbol{\gamma}_i = \boldsymbol{\gamma} + \mathbf{u}_i$ are heterogeneous. Quantities of interest include willingness to pay for specific attributes (such as trip time), WTP= $E_c[E[\gamma_{i,k}/\gamma_{i,income}]]$ and elasticities of substitution, $\eta_{j,l|k} = E_c[-\gamma_i P_{i,j} P_{i,l}]$, and entire conditional distributions of random coefficients.

5.3.2 General Frameworks

Three general frameworks are employed in empirical applications of panel data methods. Save for the cases we will note below, they depart from strict exogeneity and conditional independence.

Fixed Effects

If no restriction is imposed on the relationship between c and $\mathbf{X}$, then the conditional density $f(c|\mathbf{x}_1,\ldots,\mathbf{x}_T)$ depends on $\mathbf{X}$ in some unspecified fashion. The assumption that $E[c|\mathbf{X}]$ is not independent of $\mathbf{X}$ is sufficient to invoke the 'fixed effects' setting. With strict exogeneity and conditional independence, the application takes the form

$$f(y_{it}|\mathbf{x}_{i,t},c_i) = f_y(y_{i,t},\boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i),$$

such as in the linear panel data regression.[10] In most cases, the models are estimated by treating the effects as parameters to be estimated, using a set of dummy variables, $\mathbf{d}(j)$. The model is thus

$$f(y_{it}|\mathbf{x}_{i,t},c_i) = f_y(y_{i,t},\boldsymbol{\beta}'\mathbf{x}_{i,t} + \Sigma_j \alpha_j \mathbf{d}(j)_{i,t}).$$

The dummy variable approach presents two obstacles. First, in practical terms, estimation involves at least $K+n$ parameters Many modern panels involve tens or hundreds of thousands of units, which might make the physical estimation of $(\boldsymbol{\beta},\boldsymbol{\alpha})$ impractical. Some considerations are suggested below. The more important problem arises in models estimated by M estimators – that is, by optimizing a criterion function such as a log likelihood function. The *incidental parameters problem* (IP) arises when the number of parameters in the model (α_i) increases with the number of observation units. In particular, in almost all cases, it appears that the maximum likelihood estimator of $\boldsymbol{\beta}$ in the fixed effects model is inconsistent when T is 'small' or fixed, even if the sample is large (in n), and the model is correctly specified.

Random Effects

The random effects model specifies that $\mathbf{X}$ and c are independent so $f(c|\mathbf{X}) = f(c)$. With strict independence between $\mathbf{X}$ and c, the model takes the form $f(y_{it}|\mathbf{x}_{i,t},c_i) = f(y_{i,t},\boldsymbol{\beta}'\mathbf{x}_{i,t} + u_i)$. Estimation of

[10] Greene (2004c) labels index function models in this form 'true fixed effects' and 'true random effects' models. There has been some speculation as to what the author meant by effects models that were not 'true.' The use of the term was specifically meant only to indicate linear index function models in contrast to models that introduced the effects by some other means. The distinction was used to highlight certain other models, such as the 'fixed effects negative binomial regression model' in Hausman, Hall and Griliches (1984). In that specification, there were fixed effects defined as above in terms of $f(c|\mathbf{x})$, but the effects were not built into a linear index function.

parameters can still be problematic. But, pooled estimation (ignoring u_i) may reveal useful quantities such as average partial effects. More detailed assumptions, such as a full specification of $u_i \sim N[0,\sigma^2]$ will allow full estimation of $(\beta',\sigma)'$. It will still be necessary to contend with the fact that u_i remains unobserved. The Butler and Moffitt (1982) and maximum simulated likelihood approaches are based on the assumption that

$$E_{c_i}[f(y_{i,1},...,y_{i,t} \mid \mathbf{X}_i, c_i)] = \int_{c_i} \prod_{t=1}^{T_i} f(y_{i,t} \mid \beta'\mathbf{x}_{i,t} + c_i : \theta) d F(c_i : \sigma)$$

depends on $(\beta',\theta',\sigma)'$ in a way that the expected likelihood can be the framework for the parameters of interest.

Correlated Random Effects

The fixed effects model is appealing for its weak restrictions on $f(c_i|\mathbf{X}_i)$. But, as noted, there are practical and theoretical shortcomings that follow. The random effects approach remedies these shortcomings, but rests on an assumption that might be unreasonable, that the heterogeneity is uncorrelated with the included variables. The *correlated random effects model* places some structure on $f(c_i|\mathbf{X}_i)$. Chamberlain (1980) suggested that the unstructured $f(c_i|\mathbf{X}_i)$ be replaced with

$$c_i|\mathbf{Z}_i = \pi + \theta_1'\mathbf{z}_{i,1} + \theta_2'\mathbf{z}_{i,2} + \ldots + \theta_{Ti}'\mathbf{z}_{i,Ti} + u_i.$$

with $f(u_i)$ to be specified – u_i would be independent of $\mathbf{z}_{i,t}$. A practical problem with the Chamberlain approach is the ambiguity of unbalanced panels. Substituting $\mathbf{z}_i = 0$ for missing observations or deleting incomplete groups from the data set, are likely to be unproductive. The amount of detail in this specification might be excessive – in a modern application with moderate T and large K (say 30 or more) this implies a potentially enormous number of parameters. Mundlak (1978) and Wooldridge (2005,2010) suggest a useful simplification,

$$c|\mathbf{X}_i = \pi + \theta' \overline{\mathbf{z}}_i + u_i.$$

Among other features, it provides a convenient device to distinguish fixed effects ($\theta \neq \mathbf{0}$) from random effects ($\theta = \mathbf{0}$).

5.3.3 Dynamic Models

Dynamic models are useful for their ability (at least in principle) to distinguish between state dependence such as the dominance of initial outcomes and dependence induced by the stickiness of unobserved heterogeneity. In some cases such as in stated choice experiments, the dynamic effects might themselves be an object of estimation. (See, as well, Contoyannis et al. (CRJ, 2004).)

A general form of dynamic model would specify $f(y_{i,t}|\mathbf{X}_i,c_i,y_{i,t-1},y_{i,t-2},\ldots y_{i,0})$. Since the time series is short, the dependence on the initial condition, $y_{i,0}$, is likely to be substantive. Strict exogeneity is not feasible, since $y_{i,t}$ depends on $y_{i,t-1}$ in addition to $\mathbf{x}_{i,t}$, it must also depend on $\mathbf{x}_{i,t-1}$. A minor simplification in

terms of the lagged values produces the density $f(y_{i,t}|\mathbf{X}_i,c_i,y_{i,t-1},y_{i,0})$. The joint density of the sequence of outcomes is then

$$f(y_{i,1}, y_{i,2},\ldots,y_{i,Ti} \,|\mathbf{X}_i, y_{i,t-1},c_i, y_{i,0}) \;=\; \prod_{t=1}^{T_i} \; f(y_{i,t} \mid \mathbf{X}_i, y_{i,t-1}, c_i, y_{i,0}).$$

It remains to complete the specification for c_i and y_{i0}. A pure fixed effects approach that treats $y_{i,0}$ as 'predetermined' (or exogenous) would specify

$$f(y_{i,t}|\mathbf{X}_i,y_{i,t-1},c_i,y_{i,0}) \;=\; f(y_{i,t}|\boldsymbol{\gamma}'\mathbf{z}_{i,t} + \theta y_{i,t-1} + \gamma y_{i,0} + \alpha_i),$$

with $\mathbf{Z}_i$ implicitly embedded in α_i. This model cannot distinguish between the time invariant heterogeneity and the persistent initial conditions effect. Moreover, as several authors (e.g., Carro (2007)) have examined, the incidental parameters problem is made worse than otherwise in dynamic fixed effects models. Wooldridge (2005) suggests an extension of the correlated random effects model,

$$c_i|\mathbf{X}_i,y_{i,0} \;=\; \pi \;+\; \boldsymbol{\pi}'\,\overline{\mathbf{z}}_i \;+\; \theta y_{i,0} \;+\; u_i.$$

This approach overcomes the two shortcomings noted earlier. At the cost of the restrictions on $f(c|\mathbf{X},y_0)$, this model can distinguish the effect of the initial conditions from the effect of state persistence due to the heterogeneity. Cameron and Trivedi (2005) raise a final practical question – how should a lagged dependent variable appear in a nonlinear model? They propose, for example, a Poisson regression that would appear

$$\text{Prob}[y_{i,t} = j \mid \mathbf{X}_i, y_{i,0}, c_i] = \frac{\exp(-\lambda_{i,t})\lambda_{i,t}^{j}}{j!}, \lambda_{i,t} = \exp(\boldsymbol{\eta}'\mathbf{z}_{i,t} + \rho y_{i,t-1} + \theta_0 y_{i,0} + \pi + \boldsymbol{\theta}'\overline{\mathbf{z}}_i + u_i)$$

CRJ (2004) proposed a similar form for their ordered probit model.

5.4 Nonlinear Panel Data Modeling

Some of the methodological issues in nonlinear panel data modeling have been considered in Sections 5.2 and 5.3. We examine some of the practical aspects of common effects models.

5.4.1 Fixed Effects

The fixed effects model is semiparametric. The model framework, such as the probit or Tobit model is fully parameterized. [See Ai et al. (2015).] But, the conditional distribution of the fixed effect, $f(c|\mathbf{X})$ is unrestricted. We can treat the common effects as parameters to be estimated with the rest of the model. Assuming strict exogeneity and conditional independence, the model is

$$f(y_{i,1}, y_{i,2}, ..., y_{i,T_i} \mid \mathbf{X}_i, c_i) = \prod_{t=1}^{T_i} \; f(y_{i,t} \mid \mathbf{x}_{i,t}, c_i) = \prod_{t=1}^{T_i} \; f(y_{i,t} \mid \boldsymbol{\gamma}'\mathbf{z}_{i,t} + \alpha_i : \boldsymbol{\theta}),$$

where θ is any ancillary parameters in the model such as σ_ε in a Tobit model. Denote the number of parameters in (γ,θ) as $K^*=K+M$. A full maximum likelihood estimator would optimize the criterion function,

$$[5.4\text{-}1]\quad \ln L(\gamma,\alpha,\theta)=\sum_{i=1}^{n}\ \sum_{t=1}^{T_i}\ \ln f(y_{i,t}\mid \mathbf{z}_{i,t}:\gamma,\alpha_i,\theta)=\sum_{i=1}^{n}\ \sum_{t=1}^{T_i}\ \ln f(y_{i,t},\gamma'\mathbf{z}_{i,t}+\alpha_i:\theta),$$

where α is the $n\times 1$ vector of fixed effects. The *unconditional estimator* produces all K^*+n parameters of the model directly using conventional means.[11] The conditional approach operates on a criterion function constructed from the joint density of $(y_{i,t},\ t=1,\ldots,T_i)$ conditioned on a sufficient statistic, such that the resulting criterion function is free of the fixed effects.

Unconditional Estimation

The general log likelihood in [5.4-1] is not separable in γ and α. (For present purposes, θ can be treated the same as γ, so it is omitted for convenience.) Unconditional maximum likelihood estimation requires the dummy variable coefficients to be estimated along with the other structural parameters. For example, for the Poisson regression,

$$\text{Prob}(y_{i,t}=j\mid \mathbf{z}_{i,t}:\gamma,\alpha_i)=\frac{\exp(-\lambda_{i,t})\lambda_{i,t}^{j}}{j!},\lambda_{i,t}=\exp(\alpha_i+\gamma'\mathbf{z}_{i,t})\cdot$$

The within transformation or first differences of the data does not eliminate the fixed effects. The same problem will arise in any other nonlinear model in which the index function is transformed or the criterion function is not based on deviations from means to begin with.[12]

For most cases, full estimation of the fixed effects model requires *simultaneous* estimation of β and α_i. The likelihood equations are

$$\frac{\partial \ln L}{\partial \gamma}=\sum_{i=1}^{n}\ \sum_{t=}^{T_i}\ \frac{\partial \ln f(y_{i,t}\mid \mathbf{z}_{i,t}:\gamma,\alpha_i)}{\partial \gamma}=\mathbf{0},$$

$$[5.4-2]\qquad \frac{\partial \ln L}{\partial \alpha}=\sum_{t=1}^{T_i}\ \frac{\partial \ln f(y_{i,t}\mid \mathbf{z}_{i,t}:\gamma,\alpha_i)}{\partial \alpha_i}=0,\quad i=1,\ldots,n.$$

[11] If the model is linear, the full unconditional estimator is the within groups least squares estimator. If $\mathbf{z}_{i,t}$ contains any time invariant variables (TIVs), it will not be possible to compute the within estimator – the regressors will be collinear; the TIV will lie within the column space of the individual effects, $\mathbf{D}=(\mathbf{d}_1,\ldots,\mathbf{d}_n)$. The same problem arises for other true fixed effects nonlinear models. The collinearity problem arises in the column space of the first derivatives of the log likelihood. The Hessian for the log likelihood will be singular. The OPG matrix will be also. A widely observed exception is the negative binomial model proposed in Hausman et al. (1984) which is not a 'true' fixed effects model.

[12] If the model is a nonlinear regression of the form $y_{i,t}=\eta_i h(\gamma'\mathbf{z}_{i,t})+\varepsilon_{i,t}$, then, $\mathrm{E}[y_{i,t}/\overline{y}_i]\approx h_{i,t}/\overline{h}_i$, does eliminate the fixed effect. See Cameron and Trivedi (2005, p. 782).

Maximum likelihood estimation can involve matrix computations involving vastly more memory than would be available on a computer. Greene (2005) noted that this assessment overlooks a compelling advantage of the fixed effects model. The large submatrix of the Hessian, $\partial^2 \ln L/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'$ is diagonal, which allows a great simplification of the computations. The resulting algorithm reduces the order of the computations from $(K+n)\times(K+n)$ to $K\times K + n$. Fernandez-Val (2009) used the method to fit a fixed effects probit model with 500,000 fixed effects coefficients.[13] The method can be easily used for most of the models considered here.

Unconditional fixed effects estimation is, in fact, straightforward in principle. However, it is still often an unattractive way to proceed. The disadvantage is not the practical difficulty of the computation. In most cases – the linear regression and Poisson regression are exceptions – the unconditional estimator encounters the *incidental parameters problem*. Even with a large sample (n) and a correctly specified likelihood function, the estimator is inconsistent when T is small, as assumed here.

Concentrated Log Likelihood and Uninformative Observations

For some models, it is possible to form a concentrated log likelihood for $(\boldsymbol{\gamma},\alpha_1,\ldots,\alpha_n)$. The strategy is to solve each element of [5.4-2] for $\alpha_i(\boldsymbol{\gamma}|\mathbf{y}_i,\mathbf{X}_i)$, then insert the solution into [5.4-1] and maximize the resulting log likelihood for $\boldsymbol{\gamma}$. The implied estimator of α_i can then be computed. For the Poisson model, define

$$\lambda_{i,t} = \exp(\alpha_i + \boldsymbol{\gamma}'\mathbf{z}_{i,t}) = \eta_i \exp(\boldsymbol{\gamma}'\mathbf{z}_{i,t}) = \eta_i\phi_{i,t}.$$

The log likelihood function is

$$\ln L(\boldsymbol{\gamma},\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left[-\eta_i\phi_{i,t} + y_{i,t}\ln\eta_i + y_{i,t}\ln\phi_{i,t} - \ln y_{i,t}!\right].^{14}$$

The likelihood equation for η_i is $\partial \ln L/\partial\eta_i = -\Sigma_t \phi_{i,t} + \Sigma_t y_{i,t}/\eta_i$. Equating this to zero produces

$$[5.4\text{-}3] \quad \hat{\eta}_i = \frac{\Sigma_{t=1}^{T_i} y_{i,t}}{\Sigma_{t=1}^{T_i}\phi_{i,t}} = \frac{\bar{y}_i}{\bar{\phi}_i}.$$

Inserting this solution into the full log likelihood produces the *concentrated log likelihood*,

$$\ln L_{conc} = \sum_{i=1}^{n}\left[-\frac{\bar{y}_i}{\bar{\phi}_i}\sum_{t=1}^{T_i}\phi_{i,t} + \ln\left(\frac{\bar{y}_i}{\bar{\phi}_i}\right)\sum_{t=1}^{T_i} y_{i,t} + \sum_{t=1}^{T_i}\left(y_{i,t}\ln\phi_{i,t} - \ln\left(y_{i,t}!\right)\right)\right]$$

[13] The Hessian for a model with n = 500,000 will, by itself, occupy about 950gb of memory if the symmetry of the matrix is used to store only the lower triangle. Exploiting the special form of the Hessian reduces this to less than 4mb.

[14] The log likelihood in terms of $\eta_i = \exp(\alpha_i)$ relies on the invariance of the MLE to 1:1 transformations. See Greene (2018)

The concentrated log likelihood can now be maximized to estimate $\boldsymbol{\gamma}$. The solution for $\boldsymbol{\gamma}$ can then be used in [5.4-3] to obtain each estimate of η_i and $\alpha_i = \ln(\eta_i)$.

Groups of observations in which $\Sigma_t y_{i,t} = 0$ contribute zero to the concentrated log likelihood. In the full log likelihood, if $y_{i,t} = 0$ for all t, then $\partial \ln L/\partial \eta_i = \Sigma_t \phi_{i,t}$ which cannot equal zero. The implication is that there is no estimate of α_i if $\Sigma_t y_{i,t} = 0$. Surprisingly, for the Poisson model, estimation of a nonzero constant does not require within group variation of $y_{i,t}$ but it does require that there be at least one nonzero value. Notwithstanding the preceding issue, this strategy will not be available for most models, including the one of most interest, the fixed effects probit model.

Conditional Estimation

For a few cases, the joint density of the T_i outcomes conditioned on a *sufficient statistic*, A_i, is free of the fixed effects;

$$f(y_{i,1},...,y_{i,T_i} \mid \mathbf{X}_i, c_i, A_i) = g(y_{i,1},...,y_{i,T_i} \mid \mathbf{X}_i, A_i).$$

The most familiar example is the linear regression with normally distributed disturbances, in which, after the transformation,

$$f(y_{i,1},\ldots,y_{i,Ti}|\mathbf{X}_i,c_i,\overline{y}_i) \;=\; N[\boldsymbol{\gamma}'(\mathbf{z}_{i,t}-\overline{\mathbf{z}}_i),\sigma_\varepsilon^2).$$

The within groups estimator is the conditional maximum likelihood estimator, then the estimator of c_i is $\overline{y}_i - \hat{\boldsymbol{\gamma}}'\overline{\mathbf{z}}_i$. The Poisson regression model is another.[15] For the sequence of outcomes, with $\lambda_{i,t} = \exp(\alpha_i)\exp(\boldsymbol{\gamma}'\mathbf{z}_{i,t}) = \eta_i\phi_{i,t}$,

$$f(y_{i,1},...,y_{i,T_i} \mid \mathbf{X}_i, \Sigma_{t=1}^{T_i} y_{i,t}) = \frac{(\Sigma_{t=1}^{T_i} y_{i,t})!}{\Pi_{t=1}^{T_i}(y_{i,t}!)} \times \Pi_{t=1}^{T_i}\left(\frac{\phi_{i,t}}{\Sigma_s \phi_{i,s}}\right)^{y_{i,t}}.$$

(See Cameron and Trivedi (2005, p.807).)

Maximization of the conditional log likelihood produces a consistent estimator of $\boldsymbol{\gamma}$, but none of the fixed effects. Computation of a partial effect, or some other feature of the distribution of $y_{i,t}$, will require an estimate of α_i or $E[\alpha_i]$ or a particular value. The conditional estimator provides no information about the distribution of α_i. For index function models, it may be possible to compute ratios of partial effects, but these are generally of limited usefulness. With a consistent estimator of $\boldsymbol{\gamma}$ in hand, one might reverse the concentrated log likelihood approach. Taking $\boldsymbol{\gamma}$ as known, the term of the log likelihood relevant to estimating α_i is

$$\frac{\partial \ln L}{\partial \boldsymbol{\alpha}}\Big|\hat{\boldsymbol{\gamma}} = \sum_{t=1}^{T_i} \frac{\partial \ln f(y_{i,t} \mid \mathbf{x}_{i,t}, \hat{\boldsymbol{\gamma}} : \alpha_i)}{\partial \alpha_i} = 0, \quad i = 1,...,n.$$

[15] The exponential regression model, $f(y_{i,t}|\mathbf{x}_{i,t}) = \lambda_{i,t}\exp(-y_{i,t}\lambda_{i,t})$, $y_{i,t} \geq 0$, is a third. This model appears in studies of duration, as a base case specification, unique for its feature that its constant hazard function, $h(y_{i,t}|\mathbf{x}_{i,t}) = f(y_{i,t}|\mathbf{x}_{i,t})/[1 - F(y_{i,t}|\mathbf{x}_{i,t})] = \lambda_{i,t}$, independent of $y_{i,t}$.

In principle, one could solve each of these in turn to provide an estimator of α_i that would be consistent in *T*. Since *T* is small (and fixed), estimation of the individual elements is still dubious. However, by this solution, $\hat{\alpha}_i = \alpha_i + w_i$ where Var(w_i) = O(1/*T*). Then $\bar{\hat{\alpha}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\alpha}_i$ could be considered the mean of a sample of observations from the population generating α_i. (Each term could be considered an estimator of $E[\alpha_i|\mathbf{y}_i]$. Based on the law of iterated expectations, $\bar{\hat{\alpha}}$ should estimate $E_{\mathbf{y}}[E[\alpha|\mathbf{y}_i]] = E[\alpha]$. The terms in the mean are all based on common $\hat{\boldsymbol{\gamma}}$. But by assumption Plim$_n$ $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}$. Then, plim $\bar{\hat{\alpha}}(\hat{\boldsymbol{\gamma}}) = \text{plim}\ \bar{\hat{\alpha}}(\boldsymbol{\gamma}) = E[\alpha]$, which is what will be needed to estimated partial effects for fixed effects model.[16]

The Incidental Parameters Problem and Bias Reduction

The disadvantage of the unconditional fixed effects estimator is the incidental parameters (IP) problem. [See Lancaster (2000).] The unconditional maximum likelihood estimator is generally inconsistent in the presence of a set of incidental (secondary) parameters whose number grows with the dimension of the sample (*n*) while the number of cross sections, *T* is fixed. The phenomenon was first identified by Neyman and Scott (1948), who noticed that the unconditional maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 in the linear fixed effects model are the within groups estimator for $\boldsymbol{\gamma}$ and $\hat{\sigma}^2 = \mathbf{e}'\mathbf{e}/(nT)$, with no degrees of freedom correction. The latter estimator is inconsistent; plim $\hat{\sigma}^2 = [(T\text{-}1)/T]\ \sigma^2 < \sigma^2$. The downward bias does not diminish as *n* increases, though it does decrease to zero as *T* increases. In this particular case, plim $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}$**.** No bias is imparted to $\hat{\boldsymbol{\gamma}}$. Moreover, the estimators of the fixed effects, $\hat{\alpha}_i = \Sigma_t(y_{i,t} - \hat{\boldsymbol{\gamma}}'\mathbf{x}_{i,t})$, are unbiased, albeit inconsistent because Asy.Var[$\hat{\alpha}_i$] is O(1/*T*)

There is some misconception about the IP problem. The bias is usually assumed to be transmitted to the entire parameter vector and away from zero. The inconsistency of the estimators of α_i taints the estimation of the common parameters, $\boldsymbol{\gamma}$. But, this does not follow automatically. The nature of the inconsistencies of $\hat{\alpha}_i$ and $\hat{\boldsymbol{\gamma}}(\hat{\boldsymbol{\alpha}})$ are different. The FE estimator, $\hat{\alpha}_i$, is inconsistent because its *asymptotic variance* does not converge to zero as the sample (*n*) grows. There is no obvious sense in which the fixed effects estimators are systematically *biased* away from the true values. (In the linear model, the fixed effects estimators are actually unbiased.) But, in many nonlinear settings, the common parameters, $\boldsymbol{\gamma}$, are estimated with a systematic bias that does not diminish as *n* increases. No internally consistent theory implies this result. It varies by model. In the linear regression case, there is no systematic bias. In the binary logit case, the bias in the common parameter vector is proportional for the entire vector, away from zero. The result appears to be the same for the probit model, though this remains to be proven

[16] Wooldridge (2010, p. 309) makes this argument for the linear model. There is a remaining complication about this strategy for nonlinear models that will be pursued again in Section 4.6. Broadly, $\bar{\hat{\alpha}}$ estimates α_i for the subsample for which there is a solution for $\hat{\alpha}_i$. For example, for the Poisson model, the likelihood equation for α_i has no solution if $\Sigma_t y_{it} = 0$. These observations have been dropped for purposes of estimation. The average of the feasible estimators would estimate $E[\alpha_i|\Sigma_t y_{i,t} \neq 0]$. This may represent a nontrivial truncation of the distribution. Whether this differs from $E[\alpha_i]$ remains to be explored.

analytically. Monte Carlo evidence (Greene (2005) for the Tobit model suggests, again, that the scale parameter, σ_ε is biased, but the common slope estimators are not. In the true fixed effects stochastic frontier model, which has common parameters $\boldsymbol{\gamma}$ and two variance parameters, σ_u and σ_v, the IP problem appears to reside only in σ_v, which resembles the Neyman and Scott case.

As suggested by the Neyman and Scott application, it does seem that the force of the result is actually exerted on some explicit or embedded scaling parameters in index models. (E.g., the linear regression, Tobit, stochastic frontier, and even in binary choice, where the bias appears equally in the entire vector.) The only theoretically verified case is the binary logit model, for which it has been shown that plim $\hat{\boldsymbol{\gamma}} = 2\boldsymbol{\gamma}$ when $T = 2$. [See Abreveya (1997).] It can also be shown that plim $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}$ as $(n,T) \to \infty$. What applies between 2 and ∞, and what occurs in other models has been suggested experimentally. (See e.g., Greene (2004a).) A general result that does seem widespread is suggested by Abrevaya's result, that the IP bias is away from zero. But, in fact, this seems not to be the case either. In the Tobit case, for example, and in the stochastic frontier, the effect seems to reside in the variance term estimators. In the truncated regression, it appears that both slopes and standard deviation parameters are biased *downward*. Table 5.1 below shows some suggestive Monte Carlo simulations from Greene (2004a, 2005). All simulations are based on a latent single index model $y_{i,t}^* = \alpha_i + \beta x_{i,t} + \delta d_{i,t} + \sigma\varepsilon_{i,t}$ where $\varepsilon_{i,t}$ is either a standardized logistic variate or standard normal, $\beta = \delta = 1$, $x_{i,t}$ is continuous, $d_{i,t}$ is a dummy variable and α_i is a correlated random effect – i.e., the DGP is actually a true fixed effects model. Table entries in each case are percentage 'biases' of the unconditional estimators, computed as $100\%[(b - \beta)/\beta]$ where β is the quantity being estimated (1.0) and b is the unconditional FE estimator. The simulation also estimates the scale factor for the partial effects. The broad patterns that emerge are, first, when there is discrete variation in $y_{i,t}$, the slopes are biased away from zero. When there is continuous variation, the bias, if there is any, in the slopes, is toward zero. The bias in $\hat{\sigma}_\varepsilon$ in the censored and truncated regression models is toward zero. Estimates of partial effects seem to be more accurate than estimates of coefficients. Finally, the IP problem obviously diminishes with increases in T. Figure 5.1 shows the results of a small experimental study for a stochastic frontier model, $y_{i,t} = \alpha_i + \beta x_{i,t} + \sigma_v v_{i,t} - \sigma_u |u_{i,t}|$ where, again, this is a true fixed effects model, and $v_{i,t}$ and $u_{i,t}$ are both standard normally distributed. The true values of the parameters β, σ_u and σ_v are 0.2, 0.18 and 0.10, respectively. For β and σ_u, the deviation of the estimator from the true value is persistently only 2-3%. Figure 5.1 compares the behavior of a consistent method of moments estimator of σ_v to the maximum likelihood estimator. The results strongly suggest that the bias of the true fixed effects estimator is relatively small compared to the models in Table 5.1, and it resides in the estimator of σ_v.

Proposals to 'correct' the unconditional fixed effects estimator have focused on the probit model. Several approaches have been suggested that involve operating directly on the estimates, maximizing a 'penalized log likelihood,' or modifying the likelihood equations. Hahn and Newey's (2004) jackknife procedure provides a starting point. The central result for an unconditional estimator based on n observations and T periods is

$$\text{plim}_{n\to\infty}\ \hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma} + \tfrac{1}{T}\mathbf{b}_1 + \tfrac{1}{T^2}\mathbf{b}_2 + O\left(\tfrac{1}{T^3}\right),$$

where $\hat{\boldsymbol{\gamma}}$ is the unconditional MLE, $\mathbf{b}_1$ and $\mathbf{b}_2$ are vectors and the final term is a vector of order $(1/T^3)$.[17] For any t, a 'leave one period out' estimator without that t, has

$$\text{plim}_{n\to\infty}\hat{\boldsymbol{\gamma}}_{(t)} = \boldsymbol{\gamma} + \tfrac{1}{T-1}\mathbf{b}_1 + \tfrac{1}{(T-1)^2}\mathbf{b}_2 + O\left(\tfrac{1}{T^3}\right).$$

It follows that

$$\text{plim}_{n\to\infty}\ T\hat{\boldsymbol{\gamma}}_T - (T-1)\hat{\boldsymbol{\gamma}}_{(t)} = \boldsymbol{\gamma} - \tfrac{1}{T(T-1)}\mathbf{b}_2 + O\left(\tfrac{1}{T^3}\right) = \boldsymbol{\gamma} + O\left(\tfrac{1}{T^2}\right).$$

This reduces the bias to $O(1/T^2)$. In order to take advantage of the full sample, the jackknife estimator would be

$$\hat{\hat{\boldsymbol{\gamma}}} = T\hat{\boldsymbol{\gamma}}_T - (T-1)\bar{\hat{\boldsymbol{\gamma}}} \text{ where } \bar{\hat{\boldsymbol{\gamma}}} = \tfrac{1}{T}\Sigma_{t=1}^{T}\hat{\boldsymbol{\gamma}}_{(t)}.$$

Based on the simulation results above, one might expect the bias in this estimator to be trivial if T is in the range of many contemporary panels (say 15 or so). Imbens and Wooldridge (2012) raise a number of theoretical objections that together might limit this estimator, including a problem with $\hat{\boldsymbol{\gamma}}_{(t)}$ in dynamic models and the assumption that $\mathbf{b}_1$ and $\mathbf{b}_2$ will be the same in all periods. Several other authors, including Fernandez-Val (2009) and Carro (2007, 2014), have provided refinements on this estimator.

Table 4.1. Bias of Unconditional Fixed Effects Estimators in Limited Dependent Models

		T=2		*T*=8		*T*=20	
		Parameter	APE	Parameter	APE	Parameter	APE
Logit	β	+102.00	+67.60	+21.70	+19.10	+06.90	+3.40
	δ	+103.00	+66.00	+19.10	+12.80	+06.20	+5.20
Probit	β	+108.30	+47.40	+32.80	+24.10	+10.80	+8.80
	δ	+93.80	+38.80	+24.30	+15.20	+6.80	+4.70
Ordered Probit	β	+132.80	–	+16.60	–	+5.80	–
	δ	+160.50	–	+12.20	–	+6.80	–
Tobit	β	+0.67	+15.33	+ 0.29	+1.30	+0.05	+0.08
	δ	+0.33	+19.67	+ 0.54	+2.16	+0.14	+0.27
	σ	–36.14	–	–8.40	–	–3.30	–
Truncated	β	–17.13	–7.52	–4.92	–1.72	–2.11	–0.67
Regression	δ	–22.81	–11.64	–7.51	–3.64	–3.27	–1.53
	σ	–35.36	–	–9.12	–	–3.75	–

[17] For the probit and logit models, it appears that the relationship could be plim $\hat{\boldsymbol{\gamma}}$ = $\boldsymbol{\gamma}\, g(T)$ where $g(2) = 2$, $g'(T) < 0$ and $\lim_{T\to\infty} g(T) = 1$. This simpler alternative approach remains to be explored.

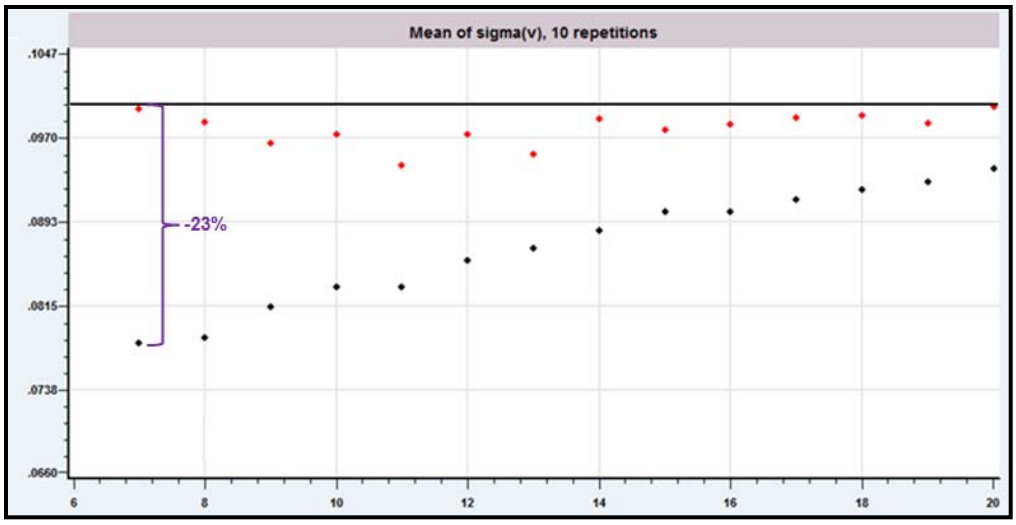

Figure 5.1. Unconditional Fixed Effects Stochastic Frontier Estimator

5.4.2 Random Effects Estimation and Correlated Random Effects

The random effects model specifies that c_i is independent of the entire sequence $\mathbf{x}_{i,t}$. Then, $f(c_i|\mathbf{X}_i) = f(c)$. Some progress can be made analyzing functions of interest, such as $E[y|\mathbf{x},c]$ with reasonably minimal assumptions. For example, if only the conditional mean, $E[c]$ is assumed known (typically zero), then estimation can sometimes proceed semiparametrically, by relying on the law of iterated expectations and averaging out the effects of heterogeneity. Thus, if sufficient detail is known about $E[y|\mathbf{x},c]$, then partial effects such as APE = $E_c\,[\partial E[y|\mathbf{x},c]/\partial \mathbf{x}]$ can be studied by averaging away the heterogeneity. However, most applications are based on parametric specifications of c_i.

Parametric Models

With strict exogeneity and conditional independence,

$$f(y_{i,1},...,y_{i,T_i} \mid \mathbf{X}_i, c_i) = \prod_{t=1}^{T_i} f(y_{i,t} \mid \mathbf{x}_{i,t}, c_i).$$

The conditional log likelihood for a random effects model is, then,

$$\ln L(\boldsymbol{\beta}, \theta, \sigma) = \sum_{i=1}^{n} \ln\left(\prod_{t=1}^{T_i} f(y_{i,t} \mid \boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i : \theta, \sigma)\right).$$

It is not possible to maximize the log likelihood with the unobserved c_i present. The unconditional density will be

$$\int_{c_i} \left(\prod_{t=1}^{T_i} f(y_{i,t} \mid \boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i : \theta)\right) f(c_i : \sigma) dc_i.$$

The unconditional log likelihood is

$$\ln L_{unconditional}(\boldsymbol{\beta},\theta,\sigma) = \sum_{i=1}^{n} \ln \int_{c_i} \left(\prod_{t=1}^{T_i} f(y_{i,t} \mid \boldsymbol{\beta}'\mathbf{x}_{i,t} + c_i : \theta) \right) f(c_i : \sigma) dc_i .$$

The maximum likelihood estimator is now computed by maximizing the unconditional log likelihood. The remaining obstacle is computing the integral. Save for the two now familiar cases, the linear regression with normally distributed disturbances and normal heterogeneity and the Poisson regression with log-gamma distributed heterogeneity, integrals of this type do not have known closed forms, and must be approximated.[18] Two approaches are typically used, Gauss-Hermite quadrature and Monte Carlo simulation.

If c_i is normally distributed with mean zero and variance σ^2, the unconditional log likelihood may be written

$$\ln_{unconditional} L(\boldsymbol{\beta},\theta,\sigma) = \sum_{i=1}^{n} \ln \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} f(y_{i,t} \mid \mathbf{x}_{i,t}, c_i : \boldsymbol{\beta},\theta) \right] \frac{1}{\sigma} \phi\left(\frac{c_i}{\sigma} \right) dc_i$$

With a change of variable and some manipulation, this can be transformed to

$$\ln L_{unconditional}(\boldsymbol{\beta},\theta,\sigma) = \sum_{i=1}^{n} \ln \int_{-\infty}^{\infty} g(h_i) e^{-h_i^2} dh_i ,$$

which is in the form needed to use Gauss-Hermite quadrature. The approximation to the unconditional log likelihood is

$$\ln L_{quadrature}(\boldsymbol{\beta},\theta,\sigma) = \sum_{i=1}^{n} \ln \sum_{h=1}^{H} \left[\prod_{t=1}^{T_i} f(y_{i,t} \mid \mathbf{x}_{i,t}, a_h : \boldsymbol{\beta},\theta) \right] w_h ,$$

where a_h and w_h are the nodes and weights for the quadrature. The method is fast and remarkably accurate, even with small numbers (H) of quadrature points. Butler and Moffitt (1982) proposed the approach for the random effects probit model. It has since been used in many different applications.[19]

Monte Carlo simulation is an alternative method. The unconditional log likelihood is,

$$\begin{aligned} \ln L(\boldsymbol{\beta},\theta,\sigma) &= \sum_{i=1}^{n} \ln \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} f(y_{i,t} \mid \mathbf{x}_{i,t}, c_i : \boldsymbol{\beta},\theta) \right] \frac{1}{\sigma} \phi\left(\frac{c_i}{\sigma} \right) dc_i \\ &= \sum_{i=1}^{n} \ln E_c \left[\prod_{t=1}^{T_i} f(y_{i,t} \mid \mathbf{x}_{i,t}, c_i : \boldsymbol{\beta},\theta) \right]. \end{aligned}$$

By relying on a law of large numbers, it is possible to approximate this expectation with an average over a random sample of observations on c_i. The sample can be created with a pseudo-random number generator. The simulated log likelihood is

[18] See Greene (2018)
[19] See, e.g., Stata (2018) and Econometric Software (2017).

$$\ln L_{simulation}(\boldsymbol{\beta},\boldsymbol{\theta},\sigma) = \sum_{i=1}^{n} \ln \frac{1}{R}\sum_{r=1}^{R} \left[\prod_{t=1}^{T_i} f(y_{i,t} \mid \mathbf{x}_{i,t}, \tilde{c}_{i,r} : \boldsymbol{\beta},\boldsymbol{\theta},\sigma)\right]$$

where $\tilde{c}_{i,r}$ is the rth pseudo random draw.[20] Maximum simulated likelihood has been used in a large and growing number of applications. Two advantages of the simulation method are, first, if integration must be done over more than one dimension, the speed advantage of simulation over quadrature becomes overwhelming and, second, the simulation method is not tied to the normal distribution – it can be applied with any type of population that can be simulated.

In most applications, the parameters of interest are partial effect of some sort, or some other derivative function of the model parameters. In random effects models, these functions will likely involve c_i. For example, for the random effects probit model, the central feature is Prob($y_{i,t}$ = 1|$\mathbf{x}_{i,t}c_i$) = $\Phi(\boldsymbol{\beta}'\mathbf{x}_{i,t} + \sigma v_i)$ where $c_i = \sigma v_i$ with $v_i \sim N[0,1]$. As we have seen earlier, the average partial effect is

APE = $E_v[\boldsymbol{\beta}\phi(\boldsymbol{\beta}'\mathbf{x} + \sigma v)]$ = $\boldsymbol{\beta}(1 - \rho)^{1/2}\,\phi(\boldsymbol{\beta}'\mathbf{x}(1 - \rho)^{1/2})$.

The function could also be approximated using either of the methods noted above. In more involved cases that do not have closed forms, that would be a natural way to proceed.

Correlated Random Effects

The fixed effects approach, with its completely unrestricted specification of $f(c|\mathbf{X})$ is appealing, but difficult to implement empirically. The random effects approach, in contrast imposes a possibly unpalatable restriction. The payoff is the detail it affords as seen in the previous section. The *correlated random effects* approach suggested by Mundlak (1978), Chamberlain (1980)) and Wooldridge (2010) is a useful middle ground. The specification is $c_i = \pi + \boldsymbol{\theta}'\overline{\mathbf{z}}_i + u_i$. This augments the random effects model shown above.

$$\ln L(\boldsymbol{\gamma},\pi,\boldsymbol{\theta},\sigma) = \sum_{i=1}^{n} \ln\left(\prod_{t=1}^{T_i} f(y_{i,t} \mid \pi + \boldsymbol{\gamma}'\mathbf{z}_{i,t} + \boldsymbol{\theta}'\overline{\mathbf{z}}_i + u_i)\right)$$

For example, if $u_i \sim N[0,\sigma^2]$, as is common, the log likelihood for the correlated random effects probit model would be

$$\ln L(\boldsymbol{\gamma},\pi,\boldsymbol{\theta},\sigma) = \sum_{i=1}^{n} \ln\int_{-\infty}^{\infty} \left(\prod_{t=1}^{T_i} \Phi[(2y_{i,t} - 1)(\pi + \boldsymbol{\gamma}'\mathbf{z}_{i,t} + \boldsymbol{\theta}'\overline{\mathbf{z}}_i + \sigma v_i)]\right)\phi(v_i)dv_i$$

Post estimation, the partial effects for this model would be based on

$$PE = \frac{\partial\Phi(\pi + \boldsymbol{\gamma}'\mathbf{z} + \boldsymbol{\theta}'\overline{\mathbf{z}} + \sigma v)}{\partial \mathbf{z}} = \boldsymbol{\gamma}\phi(\pi + \boldsymbol{\gamma}'\mathbf{z} + \boldsymbol{\theta}'\overline{\mathbf{z}} + \sigma v) = \boldsymbol{\delta}(\mathbf{z},\overline{\mathbf{z}},v).$$ [21]

[20] See Cameron and Trivedi (2005, p. 394) for some useful results on properties of this estimator.

[21] We note, in application, $\partial\Phi(\pi + \boldsymbol{\gamma}'\mathbf{z} + \boldsymbol{\theta}'\overline{\mathbf{z}} + \sigma v)/\partial\mathbf{z}$ should include a term $\frac{1}{T_i}\boldsymbol{\theta}$. For purpose of the partial effect, the variation of z is not taken to be variation if a component of $\overline{\mathbf{z}}$.

Empirically, this can be estimated by simulation or, as before, with

$$\widehat{PE} = \gamma\left(1-\rho\right)^{1/2}\,\phi[(1-\rho)^{1/2}(\pi+\gamma'\mathbf{z}+\theta'\overline{\mathbf{z}})]$$

The CRE model relaxes the restrictive independence assumption of the random effects specification, while overcoming the complications of the unrestricted fixed effects approach.

Random Parameters Models

The random effects model may be written $f(y_{i,t}|\mathbf{x}_{i,t},c_i) = f[y_{i,t}|\gamma'\mathbf{z}_{i,t} + (\pi + u_i):\theta]$. That is, as a nonlinear model with a randomly distributed constant term. We could extend the idea of heterogeneous parameters to the other parameters. A random utility based multinomial choice model might naturally accommodate heterogeneity in marginal utilities over the attributes of the choices with a random specification $\boldsymbol{\gamma}_i = \boldsymbol{\gamma} + \mathbf{u}_i$ where $E[\mathbf{u}_i] = \mathbf{0}$, $\text{Var}[\mathbf{u}_i] = \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$ and $\boldsymbol{\Gamma}$ is a lower triangular Cholesky factor for $\boldsymbol{\Sigma}$. The log likelihood function for this random parameters model is

$$\ln L(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\Sigma}) = \sum_{i=1}^{n} \ln \int_{\mathbf{v}_i} \left[\prod_{t=1}^{T_i} f(y_{i,t} \mid (\boldsymbol{\beta}+\Gamma\mathbf{v}_i)'\mathbf{x}_{i,t}:\boldsymbol{\theta})\right] f(\mathbf{v}_i)d\mathbf{v}_i$$

The integral is over K (or fewer) dimensions, which makes quadrature unappealing – the amount of computation is $O(H^K)$ while the amount of computation needed to use simulation is roughly linear in K.

A Semiparametric Random Effects Model

The preceding approach is based on a fully parametric specification for the random effect. Heckman and Singer (1984) argued (in the context of a duration model), that the specification was unnecessarily detailed. They proposed a semiparametric approach using a finite discrete support over c_i, c_q, $q = 1,\ldots,Q$, with associated probabilities, τ_q. The approach is equivalent to a *latent class*, or *finite mixture model*. The log likelihood, would be

$$\ln L(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{c},\boldsymbol{\tau}) = \sum_{i=1}^{n} \ln \frac{1}{Q}\sum_{q=1}^{Q} \tau_q\left[\prod_{t=1}^{T_i} f(y_{i,t} \mid \mathbf{x}_{i,t}:c_q,\boldsymbol{\beta},\boldsymbol{\theta})\right]., \; 0 < \tau_q < 1, \; \Sigma_q\tau_q = 1.$$

Willis (2006) applied this approach to the fixed effects binary logit model proposed by Cecchetti (1986). The logic of the discrete random effects variation could be applied to more than one, or all of the elements of $\boldsymbol{\beta}$. The resulting latent class model has been used in many recent applications.

5.4.3 Robust Estimation and Inference

In nonlinear (or linear) panel data modeling, 'robust' estimation arises in two forms. First, the difference between fixed or correlated random effects and pure random effects arises from the assumption

about restrictions on $f(c_i|\mathbf{X}_i)$. In the correlated random effects case, $f(c_i|\mathbf{X}_i) = f(c_i| \pi + \boldsymbol{\theta}'\overline{\mathbf{z}}_i)$ and in the pure random effects, case, $f(c_i|\mathbf{X}_i) = f(c_i)$. A consistent fixed effects estimator should be robust to the other two specifications. This proposition underlies much of the treatment of the linear model. The issue is much less clear for most nonlinear models because, at least in the small T case, there is no sharply consistent fixed effects estimator– because of the incidental parameters problem. This forces the analyst to choose between the inconsistent fixed effects estimator and a possibly nonrobust random effects estimator. In principle, at the cost of a set of probably mild, reasonable assumptions, the correlated random effects approach offers an appealing approach.

The second appearance of the idea of robustness in nonlinear panel data modeling will be the appropriate covariance matrix for the ML estimator. The panel data setting is the most natural place to think about clustering and robust covariance matrix estimation. [See Abadie et al. (2017), Cameron and Miller (2015) and Wooldridge (2003).] In the linear case, where the preferred estimator is OLS,

$$\mathbf{b} - \boldsymbol{\beta} = \left[\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i}\mathbf{x}_{i,t}\mathbf{x}_{i,t}'\right)\right]^{-1}\left[\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i}\mathbf{x}_{i,t}\varepsilon_{i,t}\right)\right].$$

The variance estimator would be

$$\text{Est.Var}[\mathbf{b}|\mathbf{X}] = \left[\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i}\mathbf{x}_{i,t}\mathbf{x}_{i,t}'\right)\right]^{-1}\left[\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i}\mathbf{x}_{i,t}e_{i,t}\right)\left(\Sigma_{t=1}^{T_i}\mathbf{x}_{i,t}'e_{i,t}\right)\right]\left[\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i}\mathbf{x}_{i,t}\mathbf{x}_{i,t}'\right)\right]^{-1}.$$

The correlation accommodated by the *cluster correction* in the linear model arises through the within group correlation of $(\mathbf{x}_{i,t}e_{i,t})$. Abadie et al. (2017) discuss the issue of when clustering "matters." For the linear model with normally distributed disturbances, the first and second derivatives of the log likelihood function are $\mathbf{g}_{i,t} = \mathbf{x}_{i,t}\varepsilon_{i,t}/\sigma^2$ and $\mathbf{H}_{i,t} = -\mathbf{x}_{i,t}\mathbf{x}_{i,t}'/\sigma^2$. In this case, whether clustering matters would turn on whether $(-\Sigma_{t=1}^{T_i}\hat{\mathbf{H}}_{i,t}) = \mathbf{X}_i'\mathbf{X}_i/\hat{\sigma}^2$ differs substantially from

$$\left(\Sigma_{t=1}^{T_i}\hat{\mathbf{g}}_{i,t}\right)\left(\Sigma_{t=1}^{T_i}\hat{\mathbf{g}}_{i,t}'\right) = \Sigma_{t=1}^{T_i}\Sigma_{s=1}^{T_i}e_{i,t}e_{i,s}\mathbf{x}_{i,t}\mathbf{x}_{i,s}'/\hat{\sigma}^4 = \Sigma_{t=1}^{T_i}\Sigma_{s=1}^{T_i}\hat{\mathbf{g}}_{i,t}\hat{\mathbf{g}}_{i,t}'$$

(apart from the scaling $\hat{\sigma}^2$). This, in turn depends on the within group correlation of $(\mathbf{x}_{i,t}e_{i,t})$, not necessarily on that between $e_{i,t}$ or $\mathbf{x}_{i,t}$ separately.

For a maximum likelihood estimator, the appropriate estimator is built up from the Hessian and first derivatives of the log likelihood. By expanding the likelihood equations for the MLE $\hat{\boldsymbol{\gamma}}$ around $\boldsymbol{\gamma}$,

$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \approx \left[\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i}\mathbf{H}_{i,t}\right)\right]^{-1}\left[\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i}\mathbf{g}_{i,t}\right)\right]$$

The estimator for the variance of $\hat{\boldsymbol{\gamma}}$ is then

$$\text{Est.Var}[\hat{\boldsymbol{\gamma}}] = \left[\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i}\hat{\mathbf{H}}_{i,t}\right)\right]^{-1}\left[\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i}\hat{\mathbf{g}}_{i,t}\right)\left(\Sigma_{t=1}^{T_i}\hat{\mathbf{g}}_{i,t}'\right)\right]\left[\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i}\hat{\mathbf{H}}_{i,t}\right)\right]^{-1}$$

where the terms are evaluated at $\hat{\gamma}$. The result for the nonlinear model mimics that for the linear model. In general, clustering matters with respect to the within group correlation of the scores of the log likelihood. It may be difficult to interpret this in natural terms such as membership in a group. Abadie et al. also take issue with the idea that clustering is harmless, arguing it should be "substantive." We wholeheartedly agree with this, especially given the almost reflexive (even in cross section studies) desire to secure credibility by finding something to 'cluster on.' The necessary and sufficient condition is that some form of unobservable be autocorrelated within the model. (I.e., the mere existence of some base similarity within defined groups in a population is not alone sufficient to motivate this correction.)

Clustering appears universally to be viewed as 'conservative.' The desire is to protect against being too optimistic in reporting standard errors that are too small. It seems less than universally appreciated that the algebra of the 'cluster correction' (and robust covariance matrix correction more generally) does not guarantee that the resulting estimated standard errors will be larger than the uncorrected version.

5.4.6 Attrition

When the panel data set is unbalanced, the question of ignorability is considered. The methodological framework for thinking about attrition is similar to sample selection. If attrition from the panel is related systematically to the unobserved effects in the model, then the observed sample may be 'nonrandom.' (In CRJ's (2004) study of self assessed health, the attrition appeared to be most pronounced among those whose initial health was rated poor or fair.) It is unclear what the implications are for data sets impacted by nonrandom attrition. Verbeek and Nijman (VN, 1992) suggested some variable addition tests for the presence of 'attrition bias.' The authors examined the issue in a linear regression setting. The application of CRJ (2004) to an ordered probit model is more relevant here. The Verbeek and Nijman tests add (one at a time) three variables to the main model: (1) NEXT WAVE is a dummy variable added at observed wave *t* that indicates if the individual is observed in the next wave; (2) ALL WAVES is a dummy variable that indicates whether the individual is present for all waves; (3) NUMWAVES is the total number of waves for which individual *i* is present in the sample. (Note that all of these variables are time invariant, so they cannot appear in a fixed effects model.) The authors note, these 'tests' may have low power against some alternatives and are nonconstructive – they do not indicate what response should follow a finding of attrition bias. A Hausman style of test might work. The comparison would be between the estimator based only on the full balanced panel and the full, larger, unbalanced panel. Contoyannis et al. (CRJ) note that this approach would likely not work because of the internal structure of the ordered probit model. The problem is worse than that, however. The more 'efficient' estimator of the pair is only more efficient because it uses more observations, not because of the some aspect of the model specification, as is generally required for the Hausman (1978) test. It is not clear, therefore, how the right asymptotic covariance matrix for the test should be constructed. This would apply in any modeling framework. The outcome of the VN test suggests whether the analyst should restrict the sample to the balanced panel that is present for all waves, or they can gain the additional efficiency afforded by the full, larger, unbalanced sample.

Wooldridge (2002) proposed an inverse probability weighting scheme to account for nonrandom attrition. For each individual in the sample, $d_{i,t}$ = 1[individual *i* is present in wave *t*, *t*=1,…,*T*]. A probit model is estimated for each wave based on characteristics $\mathbf{z}_{i,1}$ that are observed for everyone at wave 1. For CRJ (2004), these included variables such as initial health status and initial values of several

characteristics of health. At each period, the fitted probability $\hat{p}_{i,t}$ is computed for each individual. The weighted *pooled* log likelihood is

$$\ln L = \sum_{i=1}^{n} \sum_{t=1}^{T_i} (d_{i,t} / \hat{p}_{i,t}) \log L_{i,t} .$$

CRJ suggested some refinements to allow **z** to evolve. Their application of the set of procedures suggested the presence of attrition 'bias' for men in the sample, but not for women. Surprisingly, the difference between the estimates based on the full sample and the balanced panel were negligible.

5.4.7 Specification Tests

The random effects and fixed effects models each encompass the pooled model (linear or not) via some restriction on $f(c_i|\mathbf{X}_i)$. The tests are uncomplicated for the linear case. For the fixed effects model, the linear restriction, H_0:$\alpha_i = \alpha_1$, $i = 2,\ldots,n$ can be tested with an F statistic with (n-1) and N-n-K degrees of freedom. Under the normality assumption, a likelihood ratio statistic, -2ln($\mathbf{e}_{LSDV}'\mathbf{e}_{LSDV}/\mathbf{e}_{POOLED}'\mathbf{e}_{POOLED}$) would have a limiting chi squared distribution with n-1 degrees of freedom under H_0. There is no counterpart to the F statistic for nonlinear models. The likelihood ratio test might seem to be a candidate, but this strategy requires the unconditional fixed effects estimator to be consistent under H_0. The Poisson model is the only clear candidate for this. Cecchetti (1986) proposed a Hausman (1978) test for the binary logit model based on a comparison of the efficient pooled estimator to the inefficient conditional ML estimator.[22] This option will not be available for many other models. It requires the conditional estimator, or some other consistent (but inefficient under H_0) estimator. The logit and Poisson are the only available candidates. The strategy is certainly not available for the probit model. A generic likelihood ratio test will not be available because of the incidental parameters problem and, for some cases, the fixed effects estimator must be based on a smaller sample.

A useful middle ground is provided by the correlated random effects (CRE) strategy. The CRE model restricts the generic fixed effects model by assuming $c_i = \pi_0 + \theta' \overline{\mathbf{z}} + u_i$. If we embed this in the generic fixed effects model, so

$$f(y_{i,1},\ldots,y_{i,Ti}|\mathbf{X}_i,c_i) = \Pi_t f(\pi + \boldsymbol{\gamma}'\mathbf{z}_{i,t} + \boldsymbol{\theta}'\overline{\mathbf{z}}_i + u_i).$$

This model can be estimated as a random effects model if a distribution (such as normal) is assumed for w_i. The Wald statistic for testing H_0:$\boldsymbol{\theta} = \mathbf{0}$ would have a limiting chi squared distribution with K degrees of freedom. (The test should be carried out using a robust covariance matrix owing to the loose definition of c_i.[23])

---

[22] The validity of Cecchetti's test depends on using the same sample for both estimators. The observations with Σ_t $y_{i,t} = 0$ or T_i should be omitted from the pooled sample even though they are useable.

[23] The same test in the linear presents a direct approach. Linear regression of $y_{i,t}$ on ($\mathbf{z}_{i,t}$, $\overline{\mathbf{z}}_i$) is algebraically identical to the within estimator. A Wald test of the hypothesis that the coefficients on $\overline{\mathbf{z}}_i$ equal zero (using a robust covariance matrix) is loosely equivalent to the test described here for nonlinear models. This is the Wu (1973) test, but the underlying logic parallels the Hausman test.

The test for random effects likewise has some subtle complications. For the linear model, with normally distributed random effects, the standard approach is Breusch and Pagan's LM test based on the pooled OLS residuals:

$$LM = \frac{\left(\Sigma_{i=1}^{n} T_i\right)^2}{2\Sigma_{i=1}^{n} T_i(T_i-1)}\left[\frac{\Sigma_{i=1}^{n}(T_i\overline{e}_i)^2}{\Sigma_{i=1}^{n}\Sigma_{t=1}^{T_i} e_{i,t}^2}-1\right]^2 \longrightarrow \chi^2[1].$$

Wooldridge (2010) proposes a method of moments based test statistic that uses $\text{Cov}(\varepsilon_{i,t},\varepsilon_{i,s}) = \text{Var}(\varepsilon_{i,t}) = \sigma^2$,

$$Z = \frac{\frac{1}{n}\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i-1}\Sigma_{s=T_i+1}^{T_i} e_{i,t}e_{i,s}\right)}{\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\Sigma_{t=1}^{T_i-1}\Sigma_{s=T_i+1}^{T_i} e_{i,t}e_{i,s}\right)^2}} \longrightarrow N[0,1]$$

Some manipulation of this reveals that $Z = \sqrt{n}\ \overline{r}\ / s_r$ where $r_i = [(T_i\overline{e}_i)^2 - \mathbf{e}_i'\mathbf{e}_i]$. The difference between the two is that the *LM* statistic relies on variances (and underlying normality) while Wooldridge's relies on the covariance between $e_{i,t}$ and $e_{i,s}$ and the central limit theorem.

There is no direct counterpart to either of these statistics for nonlinear models, generally because nonlinear models do not produce 'residuals' to provide a basis for the test.[24] There is a subtle problem with tests of H_0:$\sigma_c^2 = 0$ based on the likelihood function. The regularity conditions required to derive the limiting chi squared distribution of the statistic require the parameter to be in the interior of the parameter space, not on its boundary, as it would be here. (Greene and McKenzie (2015) examine this issue for the random effects probit model.)

Under the fairly strong assumptions that underlie the Butler and Moffitt or random constants model, a simpler Wald test is available. For example, for the random effects probit model, maximization of the simulated log likelihood,

$$\ln L(\boldsymbol{\beta},\sigma) = \sum_{i=1}^{n} \ln\frac{1}{R}\sum_{r=1}^{R}\left[\prod_{t=1}^{T_i} \Phi[(2y_{i,t}-1)(\boldsymbol{\beta}'\mathbf{x}_{i,t}+\sigma v_{i,r})\right]$$

produces estimates of $\boldsymbol{\beta}$ and σ. The latter can form the basis of a Wald or likelihood ratio test. The Butler and Moffitt estimator produces an estimate of $\rho = \sigma^2/(1+\sigma^2)$ that can be treated similarly.

The random and fixed effects models are not nested without some restrictions – H_0: $f(c|\mathbf{X}) = f(c)$ requires some formal structure to provide a basis for statistical inference. Once again, the correlated random effects model provides a convenient approach. The log likelihood function under a suitable definition of $f(c|\mathbf{X}_i)$ would be

$$\ln L(\boldsymbol{\beta},\theta,\sigma) = \sum_{i=1}^{n} \ln\int_{-\infty}^{\infty}\left[\prod_{t=1}^{T_i} f(y_{i,t}\,|\,(\pi+\boldsymbol{\gamma}'\mathbf{z}_{i,t}+\ \boldsymbol{\theta}'\overline{\mathbf{z}}_i+\sigma u_i)\right] f(u_i)du_i$$

[24] Greene and McKenzie (2015) develop an *LM* test for H_0 for the random effects probit model using *generalized residuals*. [See Chesher and Irish (1986).] For a single index nonlinear (or linear) model, the generalized residual is $u_{i,t} = \partial\ln f(y_{i,t}|\bullet)/\partial(\boldsymbol{\beta}'\mathbf{x})$, i.e., the derivative with respect to the constant term. For the linear model, this is $\varepsilon_{i,t}/\sigma_\varepsilon^2$.

A Wald test of H_0:$\boldsymbol{\theta} = \mathbf{0}$ tests the difference between fixed and random effects under this specification.

5.5 Panel Data

Panel data are found in several forms. Broadly, *n* observational units are each observed *T* times in sequence. One useful distinction can be made by delineating the sampling frame that generates *n* and *T*. In the *longitudinal* data settings of interest here, we treat *T* as 'fixed,' though not necessarily very small. *The Panel Survey of Income Dynamics* (*PSID*) contains over 50 years of data while the *German Socioeconomic Panel* (*GSOEP*) is near 20 years. The *European Community Household Panel* (*ECHP*) data set was ended after eight waves. Econometric considerations in such data are generally based on *n* multivariate (*T*-variate) observations. The statistical theory for longitudinal analysis is labeled 'fixed *T*.' In particular, although some of these data sets might be long enough to be considered otherwise, the time series properties of the data (e.g., stationarity) are not of interest. The Penn World Tables (**`http://www.rug.nl/ggdc/productivity/pwt/)`** consist of $T = 65$ years of data on $n = 182$ countries (as of version 9.0 in 2017). In analyzing these *aggregate time series data*, the time series properties are of paramount importance. These could be regarded as 'fixed *n*,' though the number of countries in any particular analysis is typically not an important feature of the analysis. Asymptotic properties of estimators in this context, for example, hinge on *T*, not *n*. A style of analysis rather different from longitudinal modeling is called for in this setting. In contrast, the *Center for Research in Security Prices* (*CRSP*) data (**`http://www.crsp.com`**) provide financial analysts with extremely wide (large *n*) data on some very long time series (large *T*), such as stock and bond data for corporations. Each of these settings calls for its own classes of models and methods. In this (now, admittedly parochial) survey, we are interested in longitudinal analysis (small or fixed *T* and large *n*). Some examples of these national (or international) data sets are as follows:

- European Community: *SHARE* (*Survey of Health, Ageing and Retirement in Europe*);
- European Community: *ECHP* (*European Community Household Panel*);
- Australia: *HILDA* (*Household Income and Labor Dynamics in Australia*);
- UK: *BHPS* (now, *Understanding Society*, previously the *British Household Panel Survey*);
- Germany: *GSOEP* (*German Socioeconomic Panel*);
- Mexico: *ENEU* (*Encuesta Nacional de Empleo Urbano, Urban Employment Survey*)
- China: *CFPS* (*China Family Panel Study*);
- Italy: *WHIP* (*Work Histories Italian Panel*);
- USA: *PSID* (*Panel Survey of Income Dynamics*);
- USA: *MEPS* (*Medical Expenditure Panel Survey*);
- USA: *NLS* (*National Longitudinal Survey*);
- USA: *SIPP* (*Survey of Income and Program Participation*).

We note an immediate complication in the description above. In practice, most longitudinal data sets do not actually involve a fixed *T* observations on *n* units. Rather, units come and go from the sample for various reasons. This may be by design. In a *rotating panel*, such as the *SIPP* and *ENEU* data, units enter the panel for a fixed number of waves, and the entry of specific units is staggered. In a particular wave of the panel, the number of appearances of any unit may be any of 1,…,*T*. (*T* varies from two to

four years for the *SIPP* data and is five for the *ENEU* data) But, the reasons for exit and possible reentry by any unit might be unexplainable in the context of the study. Full generality would require us to specify that the $i = 1,\ldots,n$ observations are each observed T_i times. In nearly all received cases, this sort of variation merely presents a notational inconvenience for the econometrician and a practical, accounting complication for the model builder. It is necessary, however, to distinguish *randomly missing* observations from *attrition*. For purpose of the analysis, attrition will have two features: (1) It is an *absorbing state* – the unit that attrites from the sample does not return later. (There is 'churn' in some of the data sets listed above.); (2) In the context of whatever model is under consideration, the unobservable features that explain attrition will be correlated with the unobservables that enter the model for the interesting variable(s) under analysis. These two results produce a complication due to *nonrandom sampling*. For an example, it is not simply association of attrition with the 'dependent variable' that creates an attrition 'problem.' The association is with the unobservable effects in the model. In a model for *Income*, if attrition is explainable completely in terms of *Income* - individuals whose income reaches a certain level are asked to exit the panel - then the phenomenon can be modeled straightforwardly in terms of *truncation*. But, if the attrition is associated with the disturbance in the *Income* equation, matters become much more complicated. To continue the example, in an *Income* model, attrition that is related to *Health* might well be nonrandom with respect to *Income*. We will examine an application below.

A panel data set that consists precisely of T observations on N units is said to be a balanced panel. In contrast, if the number of observations T_i varies with i, then the panel is unbalanced. Attrition is a potential problem in unbalanced panels. Table 5.2 below displays an extract from an unbalanced panel data set. The analysis in the remainder of this survey is concerned with data such as these. (The data are extracted from the GSOEP sample that was used in Riphahn, Wambach and Million (2003).) For our purposes, the interesting variables in this data set are *HSAT*, health satisfaction, and *DOCVIS*, number of doctor visits.

ID	FEMALE	YEAR	AGE	EDUC	MARRIED	DOCVIS	HSAT	INCOME	CHILDREN
1	0	1984	54	15	1	1	8	0.305	0
1	0	1985	55	15	1	0	8	0.451005	0
1	0	1986	56	15	1	0	7	0.35	0
2	1	1984	44	9	1	0	7	0.305	0
2	1	1985	45	9	1	1	8	0.318278	0
2	1	1986	46	9	1	2	7	0.35	0
2	1	1988	48	9	1	1	8	0.35305	0
3	1	1984	58	11	0	0	10	0.1434	0
3	1	1986	60	11	0	0	9	0.3	0
3	1	1987	61	11	0	10	10	0.11	0
3	1	1988	62	11	0	3	10	0.1	0
4	1	1985	29	18	0	4	10	0.13	0
5	0	1987	27	11.8182	0	1	9	0.065	0
5	0	1988	28	11.8182	0	2	10	0.06	0
5	0	1991	31	11.8182	0	0	10	0.155	0
6	0	1985	25	9	0	2	10	0.16	1
6	0	1986	26	9	1	3	9	0.3	1
6	0	1987	27	9	1	0	8	0.3	1
6	0	1988	28	9	1	1	10	0.2	1
6	0	1991	31	9	1	18	2	0.18	1
7	1	1987	26	10	1	0	9	0.3	1
7	1	1988	27	10	1	0	7	0.2	1
7	1	1991	30	10	1	2	9	0.18	1
8	0	1984	64	10.5	0	7	0	0.15	0
9	0	1984	30	13	0	6	9	0.24	0
9	0	1987	33	13	0	7	8	0.265	0
9	0	1988	34	13	1	0	8	0.6	1
9	0	1991	37	18	1	4	7	0.7	1
9	0	1994	40	18	1	0	9	0.75	1
10	1	1988	30	18	0	0	6	0.36	0
10	1	1994	36	18	1	0	6	0.92	1

Table 5.2. Unbalanced Panel Data

5.6 Modeling Frameworks and Applications

We illustrate the applications of the panel data methods in several different nonlinear settings. We begin with the binary choice model that dominates the received literature then examine several others. A few relatively uncommon applications such as duration models (Lee (2008)) are left for more extensive treatments.

5.6.1. Binary Choice

The probit and logit models for binary choice are the standard settings for examining 'nonlinear' modeling in general, and panel data modeling in particular. The canonical origin of the topic would be Chamberlain's (1980) development of the fixed effects model and Butler and Moffitt's (1982) treatment of the random effects model.[25] The unconditional fixed effects estimators for the panel probit and logit models (see Greene (2004a,b, 2018)) exemplify the incidental parameters problem and therefore are unappealing approaches. The literature on extensions and less parameterized alternatives to the two models includes Hahn and Kuersteiner(2011), Han and Newey (2004), Carro (2007), Fernandez-Val (2009), Honoré and Lewbel (2002), Honoré and Kesina (2017), Manski (1975), Aguirrebira and Mira (2007) and Lewbel and Dong (2015).

Random and Unconditional Fixed Effects Probit Models

The log likelihood function for a panel probit model[26] is

$$\ln L(\boldsymbol{\beta},\sigma) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \ln \Phi[q_{i,t}(\pi + \boldsymbol{\gamma}'\mathbf{z}_{i,t} + c_i)],\ q_{i,t} = (2y_{i,t} - 1).$$

The pooled estimator was examined earlier. The random effects estimator would be based either on simulation or Hermite quadrature. There is no conditional likelihood estimator for the fixed effects form of this model. To illustrate the model, we will compare the various estimators using the GSOEP health data described earlier. The data are an unbalanced panel with 7,293 groups, 27,326 household/year observations. We have used the 877 households who were observed in all 7 waves (so, there are no issues of attrition embedded in the data). For purposes of computing the dynamic models, the last 6 years of data were used in all cases. The outcome variable is $Doctor_{i,t} = \mathbf{1}[DocVis_{i,t} > 0]$. Groups for which $\Sigma_t\, Doctor_{i,t}$ equals 0 or 6 were then omitted from the sample. This leaves $n^* = 597$ observations.

Estimates for random and unconditional fixed effects for a small specification are shown in Table 5.2. (Standard errors are not shown, as the discussion of the various models is not concerned with

[25] Rasch (1960) is a precursor to the fixed effects logit model.

[26] (We distinguish this from the *panel probit model* described in Bertschuk and Lechner (1998), which was essentially a constrained seemingly unrelated regressions model for a set of T binary choices; $y_{i,t} = \mathbf{1}[\boldsymbol{\beta}'\mathbf{x}_{i,t} + \varepsilon_{i,t} > 0]$ with $\text{Cov}(\varepsilon_{i,t},\varepsilon_{j,s}) = \mathbf{1}[i = j]\rho_{t,s}$ with $\rho_{t,t} = 1$ Their formulation describes cross period correlation, not individual heterogeneity.

efficiency of different estimators.) Overall, the pooled and fixed effects (FE) estimators seem distinctly removed from the random effects (RE) counterparts. The correlated random effects model seems likewise to have substantial effect on the estimated partial effects. Based on the *LM* test, the Pooled approach is rejected for any static or dynamic form. The simple RE form is rejected in favor of the CRE form for both cases as well. This would argue in favor of the FE model. A direct test for the FEM soundly rejects all other forms of the model, static or dynamic. It is not clear whether this is a valid test, however, as the FE log likelihood is not based on a consistent estimator of the parameters estimated by any other form. Still using the LR test, the dynamic CRE rejects the static one, so the preferred model is the dynamic CRE. Comparing to the static pooled model, the extensions substantially change the partial effects.

Table. 5.3 Estimated Probit Models. (Estimated partial effects in parentheses)

	Static				Dynamic		
Pooled	Pooled	RE	CRE	FE	Pooled	RE	CRE
Constant	1.603	1.612	2.668		0.648	0.880	1.449
Age	0.007	0.015	0.033	0.040	0.005	0.010	0.030
	(0.002)	(0.004)	(0.009)	(0.008)	(0.002)	(0.003)	(0.008)
Education	-0.042	-0.052	0.178	0.109	-0.026	-0.035	0.165
	(-0.014)	(-0.014)	(0.046)	(0.019)	(-0.008)	(-0.009)	(0.044)
Income	0.051	0.046	-0.119	-0.177	0.005	0.054	-0.116
	(0.018)	(0.012)	(-0.031)	(-0.315)	(0.001)	(-0.014)	(-0.031)
Health	-0.180	-0.197	-0.144	-0.180	-0.141	-0.171	-0.143
	(-0.062)	(-0.052)	(-0.037)	(-0.032)	(-0.044)	(-0.046)	(-0.038)
Married	0.119	0.105	-0.007	0.016	0.099	0.099	-0.146
	(0.041)	(0.028)	(-0.019)	(0.003)	(0.031)	(0.027)	(-0.004)
$\overline{Age}$			-0.029				-0.027
$\overline{Educ}$			-0.221				-0.198
$\overline{Income}$			0.220				0.105
$\overline{Health}$			-0.175				-0.079
$\overline{Married}$			0.250				0.220
$Doctor_{t-1}$					0.667	0.230	0.207
$Doctor_0$					0.475	0.799	0.774
ρ	0.436		0.430			0.300	0.305
Ln*L*	-3212.59	-2923.37	-2898.88	-1965.63	-2898.18	-2826.68	-2815.87
LM		215.754	212.28			112.64	121.03

Logit Model and Conditional Fixed Effects Estimation

The binary logit model is the most familiar of a small handful of models that provide a conditional estimator. [See Lancaster (2000).] The probability with fixed effects is

$$\text{Prob}(y_{i,t}=1\,|\,\mathbf{x}_{i,t},\alpha_i)=\Lambda(\alpha_i+\boldsymbol{\gamma}'\mathbf{z}_{i,t})=e^{\alpha_i+\boldsymbol{\gamma}'\mathbf{z}_{i,t}}/[1+e^{\alpha_i+\boldsymbol{\gamma}'\mathbf{z}_{i,t}}].$$

The unconditional logit log likelihood is

$$\ln L(\boldsymbol{\gamma},\boldsymbol{\alpha})=\sum_{i=1}^{n^*}\sum_{t=1}^{T_i}\ln\Lambda[q_{i,t}(\boldsymbol{\gamma}'\mathbf{z}_{i,t}+\alpha_i)],\ q_{i,t}=(2y_{i,t}-1).$$

Groups for which $\Sigma_t\, y_{i,t}$ equals 0 or T_i do not contribute to this log likelihood, so the sum is over the n^* observations for which $0 < \Sigma_t y_{i,t} < T_i$. The unconditional log likelihood is straightforward to maximize over($\boldsymbol{\gamma}$,$\boldsymbol{\alpha}$) using the remaining observations. The conditional log likelihood is the sum of the logs of the probabilities conditioned on $S_i=\Sigma_{t=1}^{T_i}y_{i,t}$,

$$\text{Prob}\left(y_{i,1}, y_{i,2}, \ldots, y_{i,T_i} \,|S_i\right) = \frac{\exp\left(\sum_{t=1}^{T_i} y_{i,t}\boldsymbol{\gamma}'\mathbf{z}_{i,t}\right)}{\sum_{\Sigma_t d_{i,t}=S_i} \exp\left(\sum_{t=1}^{T_i} d_{i,t}\boldsymbol{\gamma}'\mathbf{z}_{i,t}\right)} = \frac{\exp\left(\sum_{t=1}^{T_i} y_{i,t}\boldsymbol{\gamma}'\mathbf{z}_{i,t}\right)}{\sum \binom{T_i}{S_i} \text{ different ways that } \Sigma_t d_{i,t} \text{ can equal } S_i \quad \exp\left(\sum_{t=1}^{T_i} d_{i,t}\boldsymbol{\gamma}'\mathbf{z}_{i,t}\right)}.$$

The denominator is summed over all the different combinations of T_i values of $y_{i,t}$ that sum to the same total as the observed data. There are $\binom{T_i}{S_i}$ terms. This may be large. With $T = 6$ (as in our example), it reaches 30 at $S = 3$. With $T = 50$, it reaches 10^{14} at $S = 25$.[27] The algorithm by Krailo and Pike (1984) makes the computation extremely fast and simple. The estimators of α_i are not individually consistent, but one might expect $(1/n^*)\Sigma_i \hat{\alpha}_i$ to be a consistent estimator of $E[\alpha_i]$. A remaining question to be considered is whether $E[\alpha_i | 0 < S_i < T_i]$ differs from $E[\alpha_i]$. Assuming not, partial effects for the fixed effects logit model can be estimated with

$$\widehat{APE} = \hat{\boldsymbol{\gamma}}\left\{\frac{1}{n^*}\Sigma_{i=1}^{n^*}\Sigma_{t=1}^{T_i}\left[\Lambda\left(\bar{\hat{\alpha}} + \hat{\boldsymbol{\gamma}}'\mathbf{z}_{i,t}\right)\right]\left[1 - \Lambda\left(\bar{\hat{\alpha}} + \hat{\boldsymbol{\gamma}}'\mathbf{z}_{i,t}\right)\right]\right\}.$$

(The average could be over n^* alone using $\bar{\mathbf{z}}_i$.) Table 5.4 shows the estimates. They are quite close even though n^* is moderate and $T_i = 6$ for all i, which is small by modern standards. The unconditional

Table 5.4 Estimated Fixed Effects Logit Models (Percentage excess in parentheses)

	Unconditional		Conditional	
	Estimate	PEA	Estimate	PEA
Age	0.065 (14)	0.017 (21)	0.057	0.014
Educ	0.168 (17)	0.041 (14)	0.144	0.036
Income	-0.284 (21)	-0.070 (21)	-0.234	-0.058
Health	-0.304 (21)	-0.074 (19)	-0.251	-0.062
Married	0.041 (24)	0.010 (25)	0.033	0.008

estimates are uniformly slightly larger. The percentage differences between the two estimates are shown in parentheses in the table. The results are consistent with the results for $T = 8$ in Table 5.1. This does suggest that the effect diminishes from the benchmark of 100% at $T = 2$ rather rapidly. We also examined the estimated fixed effects. The unconditional estimates are estimated with $\boldsymbol{\gamma}$. The conditional estimates are computed by solving the unconditional likelihood equation for α_i using the consistent conditional estimator of $\boldsymbol{\gamma}$. The means of the conditional and unconditional estimators are –2.4 for the unconditional and -2.1 for the conditional. Figure 5.2 compares the two sets of estimates.

[27] Estimation of a model with $n = 1{,}000$ and $T = 50$ required about 0.5 seconds. Of course, if $T = 50$, the incidental parameters problem would be a moot point.

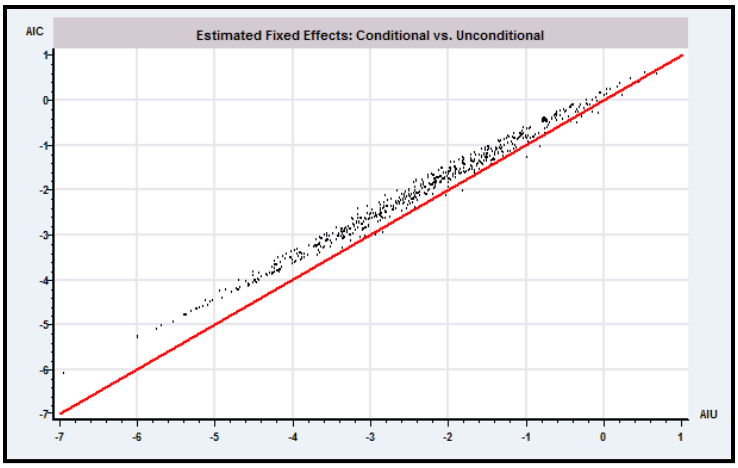

Figure 5.2 Plot of Estimates of α_i Conditional vs. Unconditional

Chamberlain (1980) also proposed a conditional estimator for a multinomial logit model with fixed effects. The model is defined for a sequence of choices from J+1 alternatives by individual i in repetition t, J choices and an 'opt out' or 'none' choice that is taken a substantive number of times. The choice probabilities are then

$$\text{Prob}(y_{i,t,j}=1\,|\,\mathbf{z}_{i,t,j})=\frac{e^{\alpha_{i,j}+\boldsymbol{\gamma}'\mathbf{z}_{i,t,j}}}{1+\Sigma_{m=1}^{J}e^{\alpha_{i,m}+\boldsymbol{\gamma}'\mathbf{z}_{i,t,m}}};\ \text{Prob}(y_{i,t,0}=1\,|\,\mathbf{z}_{i,t,0})=\frac{1}{1+\Sigma_{m=1}^{J}e^{\alpha_{i,m}+\boldsymbol{\gamma}'\mathbf{z}_{i,t,m}}},\ j=1,\ldots,J,$$

where the outcome is $d_{i,t,j}$ = **1**[individual i makes choice j in choice task t] and $\mathbf{z}_{i,t,j}$ = a set of alternative specific attributes of choice j. Individual specific, choice invariant characteristics such as age or income could be introduced into the model by interacting them with J alternative specific constants.) The probability attached to the sequence of choices is constructed similarly but the summing in the denominator of the conditional probability is for the sum of $d_{i,t,j}$ over (J+1)T terms for individual i. The summing for the conditional probability itemizes terms for which the denominator $\Sigma_{j,t}d_{i,j,t}$ equals S_i, subject to the constraint that the terms in each block of (J+1) sum to 1 (only one choice is made) and the sum in the T blocks equals the sum for the observed blocks. The counterpart to the uninformative observations in the binomial case are individuals that make the same choice, j, in every period, t. There is an enormous amount of computation. (See Pforr (2011, 2014).) But, there is a much simpler way to proceed. For each of the J alternatives, there is a set of T blocks of 2 alternatives, each consisting of alternative j and the opt out choice. In each n(2T) set, there is a binary logit model to be constructed, where the individual chooses either alternative j or the opt out choice. Each of these binary choice models produces a consistent estimator of $\boldsymbol{\gamma}$, say $\hat{\boldsymbol{\gamma}}(j)$, j=1,…,J. Since there are J such estimators, they can be reconciled with a minimum distance estimator,

$$\hat{\boldsymbol{\gamma}}_{MD}=\left[\Sigma_{j=1}^{J}\{\hat{\boldsymbol{\Omega}}(j)\}^{-1}\right]^{-1}\left[\Sigma_{j=1}^{J}\{\hat{\boldsymbol{\Omega}}(j)\}^{-1}\hat{\boldsymbol{\gamma}}(j)\right]=\Sigma_{j=1}^{J}\mathbf{W}(j)\hat{\boldsymbol{\gamma}}(j),$$

$$\mathbf{W}(j)=\left[\Sigma_{j=1}^{J}\{\hat{\boldsymbol{\Omega}}(j)\}^{-1}\right]^{-1}\{\hat{\boldsymbol{\Omega}}(j)\}^{-1}\ \text{su}\ \text{h}\ \text{t}\ \text{a}\Sigma_{j=1}^{J}\mathbf{W}(j)=\mathbf{I},$$

where $\hat{\boldsymbol{\Omega}}(j)$ is the estimated asymptotic covariance matrix for the jth estimator. The amount of computation involved is a very small fraction of that developed in Pforr(2011,2014). The reduction in the amount of computation is enormous at the possible cost of some efficiency. For Pforr's example, which is involves 26,200 individual/period choices and J+1 = 2 alternatives, the author reports the full Chamberlain computation requires 101.58 seconds. Partitioning the problem and using the minimum distance estimator produces the numerically identical result in 0.297 seconds.[28]

5.6.2. Bivariate and Recursive Binary Choice

The bivariate probit model (there is no logit counterpart), and recursive bivariate probit (probit model with an endogenous binary variable) has attracted some recent attention.[29] The two equation model with common effects would be

$$y_{1,i,t} = \mathbf{1}[\boldsymbol{\beta}_1'\mathbf{x}_{1,i,t} + \boldsymbol{\gamma}'\mathbf{z}_{i,t} + c_{1,i} + \varepsilon_{1.i,t} > 0]$$
$$y_{2,i,t} = \mathbf{1}[\boldsymbol{\beta}_2'\mathbf{x}_{2,i,t} + \delta y_{1,i,t} + c_{2,i} + \varepsilon_{2,i,t} > 0].$$

A full fixed effects treatment would require two sets of fixed effects and would be affected by the IP problem. There is no conditional estimator available. The random effects model, or the correlated random effects model would be a natural choice. A dynamic model would proceed along the lines developed earlier for the single equation case. (Rhine and Greene (2013) treated y_1 as the initial value and y_2 as the second period value in a two period RBP.)

5.6.3. Ordered choice

Contoyannis et al. (2004) used the dynamic CRE model in their analysis of health satisfaction in the BHPS. One of the complications in their case is the treatment of lagged effects for an ordered choice outcome that takes J+1 values, 0,…,J. The solution is a set of J endogenous lagged dummy variables, one for each category. A fixed effects treatment of the ordered probit (logit) model presents the same complications as the binary probit or logit model. Ferrer-i-Carbonell and Frijters (2004) note that the ordered choice model can be broken up into a set of binary choice models. If

$$\text{Prob}(y_{i,t} = j) = \Lambda(\mu_j - \alpha_i - \boldsymbol{\gamma}'\mathbf{z}_{i,t}) - \Lambda(\mu_{j-1} - \alpha_i - \boldsymbol{\gamma}'\mathbf{z}_{i,t})$$

then

$$\text{Prob}(y_{i,t} > j) = \Lambda(\alpha_i + \gamma' z_{i,t} - \mu_j).$$

The transformed model can be treated with Chamberlain's conditional fixed effects approach. The time invariant threshold becomes an outcome specific constant, and will be lost in the fixed effects. Like the

[28] Pforr's data for this application are obtained from Stata at http://www.stata-press.com/data/r11/r.html under the CLOGIT heading. The data are reconfigured for NLOGIT (Econometric Software (2017)). The data may be downloaded from the author's website at http://people.stern.nyu.edu/wgreene/felogit.csv. A second example involving J=3, T=8 and n=400 required 0.229 seconds using the MDE.

[29] (Wilde (2000), Han and Vytlacil (2017), Mourifie and Meango (2014), Filippini, Greene, Kumar and Martinez-Cruz (2018), Rhine and Greene (2013), Scott, Schurer, Jensen and Sivey (2009), Gregory and Deb (2015).

multinomial logit model considered earlier, this produces multiple estimates of γ, which can be reconciled with a minimum distance estimator. Bias corrections for the fixed effects ordered probit and logit models are developed by Bester and Hansen (2009), Carro (2007), Carro and Trafferri (2014), Muris (2017) and others.

5.6.4. Censored or Truncated Regression

Much less is known (or studied) about the censored (Tobit) and truncated regression models. Greene's (2005) results (in Table .1) suggest that the incidental parameters problem appears, but in a different fashion than in discrete choice models – and the censored and truncated models behave differently from each other. Honoré and Kesina (2017) examine a number of issues in this setting and a semiparametric specification. A serious complication will arise in a dynamic Tobit models – it is unclear how a lagged effect that is either zero or continuous should be built into the model.

5.6.5. Stochastic Frontier: Panel Models,

Panel data considerations in the stochastic frontier model focus on both inefficiency and heterogeneity. The model framework is built from the canonical model

$$y_{i,t} = \beta'\mathbf{x}_{i,t} + v_{i,t} - u_{i,t}$$

where $u_{i,t} < 0$ and typically $v_{i,t}$ is $N[0,\sigma_v^2]$. Aigner, Lovell and Schmidt's (1977) base case specifies $u_{i,t}$ as $N^+(0,\sigma_u^2)$. The early developments for panel data treatments focused on $u_{i,t}$, not on heterogeneity. Pitt and Lee (1981) specified u_i as a time invariant, random one sided term that represented inefficiency. Schmidt and Sickles (1984) and Cornwell, Schmidt and Sickles (1990) developed a fixed effects approach that respecified $u_{i,t}$ as a fixed value, a_i or time varying, $a_i(t)$. Subsequent developments (e.g., Kumbhakar et al. (2014) and Battese and Coelli (1995) and Cuesta (2000) extended the time variation of $u_{i,t}$ by various specifications of $\sigma_u(t)$. These developments oriented the focus on inefficiency measurement while leaving unobserved heterogeneity ambiguous or assumed to be time varying and embedded in $v_{i,t}$. Greene (2005) proposed the 'true random effects' and 'true fixed effects' models

$$y_{i,t} = (\alpha + w_i) + \gamma'\mathbf{z}_{i,t} + v_{i,t} - u_{i,t}$$

where $u_{i,t}$ is as originally specified in Aigner et al. and w_i is treated as either a 'true' fixed or random effect. The latter model, with its combination of normal w_i and skew normal $(v_{i,t} - u_{i,t})$ is estimated by maximum simulated likelihood. Kumbhakar et al. (2014) completed the development with the 'generalized true random effects model,'

$$y_{i,t} = (\alpha + w_i - f_i) + \gamma'\mathbf{z}_{i,t} + v_{i,t} - u_{i,t}$$

where f_i now has a truncated normal distribution like $u_{i,t}$, and the full model is based on the sum of two skew normal variables, which has a closed skew normal distribution. The authors developed a full maximum likelihood estimator. Greene and Filippini (2015) showed how the estimation could be simplified by simulation.

5.6.6 Count Data

With the binary probit and logit models, the Poisson regression model for count data has been the proving ground for methods of nonlinear panel data modeling. A comprehensive early reference is Hausman, Hall and Griliches (1984).[30] The fixed effects conditional estimator is identical to the unconditional estimator, so the latter is consistent. The random effects model (or correlated random effects) is a straightforward application of Butler and Moffitt's method. As a nonlinear regression, the specification provides a convenient framework for modeling multiple equations. Riphahn et al. (2003) specified a two equation random effects Poisson model,

$$y_{i,t,j} \sim \text{Poisson with } \lambda_{i,t,j} = \exp(\pi_j + \boldsymbol{\gamma}_j'\mathbf{z}_{i,t,j} + \varepsilon_{i,t,j} + u_{i,j}),\ j = 1,2,\ i=1,\ldots,n,\ t = 1,\ldots,T_i.$$

The two equations are correlated through the means, $\rho = \text{Cov}(\varepsilon_{i,t,1},\varepsilon_{i,t,2})$. (A natural extension would be to allow correlation between the random effects as well, or instead.) In the univariate, cross section case, the heterogeneous Poisson regression is specified with conditional mean $\lambda_{i,t} = \exp(\pi + \boldsymbol{\gamma}'\mathbf{z}_{i,t} + u_i)$. If $u_i \sim$ log-gamma with mean 1, the unconditional distribution after integrating out u_i is the negative binomial (NB). This convenience has motivated use of the NB form. The log-gamma, while convenient, in that form, is extremely inconvenient (intractable) in a model such as RWM's. Recent applications of mixed models have used the normal distribution, and computed the necessary integrals by Monte Carlo simulation.

The Poisson and negative binomial models have also been frequently the setting for latent class models. Jones and Schurer (2011) examined the frequency of doctor visits in a two class negative binomial latent class model. Their methodology provides a useful example for using latent class modeling. Two questions that attend this type of modeling are (1) is it possible to characterize the latent classes (other than by number) and (2) is it possible to assign individuals to their respective classes? Strictly, the answer to both classes is no. Otherwise, the classes would not be latent. But, it is possible to do both probabilistically. The latent class Poisson model is

$$\text{Prob}[y_{i,t} = j \mid class = q] = \frac{\exp(-\lambda_{i,t} \mid class = q)(\lambda_{i,t} \mid class = q)^j}{j!},\ (\lambda_{i,t} \mid class = q) = \exp(\boldsymbol{\beta}_q'\mathbf{x}_{i,t})$$

$$\ln L = \sum_{i=1}^{n} \log \sum_{q=1}^{Q} \tau_q \prod_{t=1}^{T_i} \frac{\exp(-\lambda_{i,t} \mid q)(\lambda_{i,t} \mid q)^j}{j!} = \sum_{i=1}^{n} \ln \sum_{q=1}^{Q} \tau_q (H_i \mid q)$$

Maximization of the log likelihood produces estimates of $(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_Q)$ and $(\tau_1,\ldots,\tau_q)$. (A more elaborate specification that bears some similarity to the correlated random effects model would make τ_q a function of exogenous factors, $\mathbf{z}_i$ and/or the group means of $\mathbf{x}_{i,t}$. See Greene (2018, Section 18.4). With the estimates of $(\boldsymbol{\beta}_q,\tau_q)$ in hand, the *posterior class probabilities* for each individual can be computed;

[30] Hausman et al.'s (1984) formulation of the fixed effects NB model embedded the fixed effects in a variance parameter, not as an offset in the conditional mean as is familiar in other models. As a consequence, their FE model permits time invariant variables in the mean function, a result that continues to surprise researchers who are not warned of this. See Greene (2018, p. 901).

$$\hat{\tau}_{i,q} = \frac{\hat{\tau}_q(\hat{H}_i \mid q)}{\Sigma_{s=1}^{Q}\hat{\tau}_s(\hat{H}_i \mid s)}$$

Individuals can then be assigned to the class with the highest posterior probability. Jones and Schurer (2011) then characterized the two classes as 'light users' and 'heavy users' by the average frequency of doctor visits within the classes. They also computed characteristics such as average partial effects by the two groups to characterize the system. Table 5.5 repeats this exercise with the GSOEP data used earlier. The three classes do appear to be separating individuals by the intensity of usage. The pattern of the partial effects suggests

Table 5.5 Latent Class Model for Doctor Visits

	Class 1		Class 2		Class 3	
	Parameter	APE	Parameter	APE	Parameter	APE
Constant	3.253	-	1.524	-	0.116	-
Age	0.015	0.132	0.024	0.102	0.038	0.048
Educ	-0.061	-0.535	-0.035	-0.137	-0.040	-0.050
Income	-0.178	-0.156*	-0.274	-0.107*	0.301	0.038*
HSAT	-0.220	-1.929	-0.178	-0.696	-0.275	-0.347
Married	0.134	1.175	0.080	0.313	0.005	0.006
$\overline{DocVis} \mid \hat{q}_i$	10.423		4.174		1.642	
$Mean\ \hat{E}[\bullet] \mid \hat{q}_i$	8.771		3.914		1.262	
$\hat{\tau}_q$	0.158		0.474		0.368	

5.6.7. A General Nonlinear Regression

Papke and Wooldridge (1996, 2008) proposed a model for aggregates of binary responses. The resulting outcome is a fractional variable. Minimum chi squared methods for fractional variables have long provided a useful consistent approach. The model developed here builds out from a common effects binary choice model. The resulting treatment is a heteroscedastic nonlinear regression that lends itself well to the correlated random effects treatment. (See, also Wooldridge (2010) pp. 748-755 and 759-764.) No obvious likelihood based approach emerges, so the preferred estimator is nonlinear (possibly weighted) least squares.

5.6.9. Sample Selection Models:

Most treatments of sample selection have layered the fixed and/or random effects treatments over Heckman's (1979) sample selection model. Verbeek (1990) and Verbeek and Nijman (1992) proposed a hybrid fixed and random effects specification,

$$d_{i,t} = 1[\gamma' \mathbf{z}_{i,t} + \eta_{i,t} > 0] + \text{(Random effects probit)}$$

$$y_{i,t} \mid (d_{i,t} = 1) = \boldsymbol{\beta}' \mathbf{x}_{i,t} + \text{(Fixed effects regression)}$$

Zabel (1992) argued that the FE model should have appeared in both equations. He then proposed the CRE form for the usual reasons. The system that results is two CRE models with correlation of the idiosyncratic disturbances. A natural extension would be correlation of u_i and v_i.

$$d_{i,t} = 1[\gamma z + \eta' \bar{\mathbf{z}}_i + \theta d_{i,t-1} + u_i + u_{i,t} > 0] \quad \text{(Correlated random effects probit)}$$

$$y_{i,t} \mid (d_{i,t} = 1) = \psi\varepsilon + \boldsymbol{\kappa}'\mathbf{z}_{i,t} + (\boldsymbol{\lambda}'\bar{\mathbf{z}}_i + v_i) + \varepsilon_{i,t} \quad \text{(Correlated random effects regression)}$$

Vella (1998) provides some details on this strand of development. Fernandez-Val and Vella (2009) continue the analysis with bias corrections based on the fixed effects specification. Kyriazidou (1997) suggested a semiparametric approach based on a fixed effects logit selection and weighted least squares with kernel estimators for the weights. Refinements are considered by Vella and Verbeek (1999), Barrachine (1999) and Dustman Rochina Barrachina (2007) and Semykina and Wooldridge (2010)

In all of these treatments, the selection process is run at the beginning of each period – the selection equation is repeated, without autocorrelation, for every *t*. Bravo-Ureta et al. (2012) applied the selection model in a setting in which the selection occurs at the baseline, and is unchanged for all *T* periods. The selection effect becomes a correlated random effect. In their application, the main outcome equation is a stochastic frontier model. Greene (2010) shows how the model can be estimated either by full information maximum likelihood or by Monte Carlo simulation.

5.6.8. Individual choice and stated choice experiments

The choice probability in the multinomial choice model we examined in Section 5.6.1 is

$$\text{Prob}(\textit{choice} = j) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{i,j})}{\Sigma_{s=1}^{J}\exp(\boldsymbol{\beta}'\mathbf{x}_{i,s})}.$$

More than any other model examined in this survey, the coefficients in this model are not of direct use. Once the parameters have been estimated, the model will be used to compute probabilities, simulate market shares under policy scenarios, estimate willingness to pay and distributions of willingness to pay, and compute elasticities of probabilities. Since all of these require a full set of components for the probabilities, the fixed effects model that bypasses computation of the fixed effects does not seem helpful. A random effects approach is considered in Hensher et al. (2007)

The counterpart of a 'panel' in recent applications of choice modeling is the *stated choice experiment*. (See Hensher et al. (2015).) The individual being interviewed is offered a choice task involving *J* alternatives with a variety of attributes, $\mathbf{x}_{i,t,j}$. In the typical experiment, this scenario will be repeated *T* times with widely varying attribute sets in order to elicit the characteristics of the respondent's preferences. The common fixed or random effect that is persistent across choice settings serves to accommodate the feature that this is the same individual with the same latent attributes making the choices with short intervals between tasks. It is unlikely that the random utility formulation of the model could be so complete that the choice tasks would be independent conditioned on the information that appears in the utility functions. The mixed logit is the current standard in the modeling of choice experiments. The model is

$$\text{Prob}(Choice_{i,t} = j \mid \mathbf{X}_i) = \frac{\exp(V_{i,t,j})}{\sum_{s=1}^{J} \exp(V_{i,t,s})},\ V_{i,t,s} = \alpha_j + \boldsymbol{\beta}_i'\mathbf{x}_{i,t,j} + \varepsilon_{i,t,j}$$

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Gamma}\mathbf{u}_i$$

Revelt and Train (1998) modeled results of a survey of California electric utility customers. Train (2009) summarizes the theory and relevant practical aspects of discrete choice modeling with random parameters.

5.6.10 Multilevel Models Hierarchical (Nonlinear) Models

The general methodology of 'multilevel modeling' (often linear modeling) builds a random parameters specification that bears some resemblance to the correlated random effects model. (See Raudebush and Bryk (2002).) A generic form would be

$$f(y_{i,t}|\mathbf{x}_{i,t},\mathbf{u}_i:\boldsymbol{\beta},\boldsymbol{\Sigma}) = f(y_{i,t}, (\boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u}_i)'\mathbf{x}_{i,t}:\boldsymbol{\theta}) = f(y_{i,t}, \boldsymbol{\beta}_i'\mathbf{x}_{i,t}:\boldsymbol{\theta}).$$

A useful extension is $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Gamma}\mathbf{u}_i$, where $\mathbf{z}_i$ indicates exogenous factors; $\mathbf{z}_i$ could also include the correlated random effects treatment with the group means of $\mathbf{x}_{i,t}$. For a linear model, estimation is often based on manipulation of feasible generalized least squares. For a nonlinear model, this will require multivariate integration to deal with the unobserved random effects. This can be done with Monte Carlo simulation.

References

Abadie, A., Athey, S., Imbens, G., Wooldridge, J., 2017. When should you adjust standard errors for clustering? MIT Department of Economics Working Paper 13927. < https://economics.mit.edu/files/13927> (accessed 18.01.18.).

Abrevaya, J., 1997. The equivalence of two estimators of the fixed effects logit model. Economics Letters 55, 41-43.

Aguirrebiria, V., Mira, P., 2007. Sequential estimation of dynamic discrete games. Econometrics 75 (1), 1-53.

Ai, C., Li, H., Lin, Z., Ment, M., 2015. Estimation of panel data partly specified tobit regression with fixed effects, Journal of Econometrics 188 (2), 316-326.

Ai, C., Norton, E., 2003. Interaction terms in logit and probit models, Economics Letters 80, 123-129.

Aigner, D., Lovell, K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production models. Journal of Econometrics, 6, 21-37.

Altonji, J., Matzkin, R., 2005. Cross section and panel data for nonseparable models with endogenous regressors. Econometrica, 73(4)1053-1102.

Angrist, J.,Pischke, J., 2009. Mostly Harmless Econometrics. Princeton University Press, Princeton, NJ.

Arellano, M., Hahn, J., 2006. Understanding bias in nonlinear panel models: some recent developments. Advances in Economics and Econometrics, Ninth World Congress. Cambridge University Press, New York.

Barrachina, M., 1999. A new estimator for panel data sample selection models. Annales d'Economie et de Statistique 55/56, 153-181.

Battese, G., Coelli, T., 1995. A model for technical inefficiency effects in a stochastic frontier production for panel data. Empirical Economics 20, 325-332.

Bertschuk, I.,Lechner, M., 1998. Convenient Estimators for the Panel Probit Model, Journal of Econometrics, 87 (2), 329-372.

Bester, C., Hansen, A., 2009. A penalty function approach to bias reduction in nonlinear panel models with fixed effects. Journal of Business and Economic Statistics 27 (2), 131-148.

Bravo-Ureta, B., Greene, W., Solis, D., 2012. Technical efficiency analysis correcting for biases from observed and unobserved variables: an application to a natural resource management project. Empirical Economics 43 (1), 55-72.

Butler, J.,Moffitt, R., 1982. A computationally efficient quadrature procedure for the one factor multinomial probit model. Econometrica 50, 761-764.

Cameron, C., Miller, D., 2015. A practitioner's guide to cluster robust inference. Journal of Human Resources 50 (2), 317-373.

Cameron, C, Trivedi, P., 2005. Microeconometrics: Methods and Applications. Cambridge University Press, New York, NY.

Carro, J., 2007. estimating dynamic panel data discrete choice models with fixed effects. Journal of Econometrics 140, 503-528.

Carro, J., Browning, M., 2014. Dynamic binary outcome models with maximal heterogeneity. Journal of Econometrics 178 (2), 805-823.

Carro, J., Trafferri, A., 2014. State dependence and heterogeneity in health using a bias corrected fixed effects estimator. Journal of Econometrics 29, 181-207.

Cecchetti, S., 1986. The frequency of price adjustment: a study of newsstand prices of magazines, Journal of Econometrics 31 (3), 255-274.

Chamberlain, G., 1980. Analysis of covariance with qualitative data. Review of Economic Studies 47, 225-238.

Chesher, A., Irish, M., 1987. Residual Analysis in the Grouped Data and Censored Normal Linear Model, Journal of Econometrics, 34, 33-62. (1986)

Chesher, A., 2013 Semiparametric structural models of binary response: shape restrictions and partial identification, Econometric Theory 29, 231-266.

Contoyannis, C., Jones, A., Rice, N., 2004. The dynamics of health in the British household panel survey. Journal of Applied Econometrics 19 (4), 473-503.

Cornwell, C., Schmidt, P., Sickles, R., 1990. Production frontiers with cross-section and time-series variation in efficiency levels, Journal of Econometrics 46, 185-200.

Cuesta, R., 2000. A production model with firm-specific temporal variation in technical inefficiency: with application to Spanish dairy farms. Journal of Productivity Analysis 13 (2), 139-158.

Dustman, C., Marrachine, M., 2007. Selection correction in panel data models: an application to the estimation of females' wage equations. Econometrics Journal 10, 263-293.

Econometric Software Inc., 2017. NLOGIT. ESI Inc., Plainview, NY.

Fernandez-Val, I., 2009. Fixed effects estimation of structural parameters and marginal effects in panel probit models. Journal of Econometrics 150 (1), 71-75.

Fernandez-Val, I., Vella, F., 2009. Bias corrections for two-step fixed effects panel data estimators. Journal of Econometrics 163 (2), 144-162.

Ferrer-i-Carbonell, A., Frijters, P., 2004. The effect of methodology on the determinants of happiness. Economic Journal 1114, 641-659.

Filippini, M., Greene, W., Kumar, N., Martinez-Cruz, A., 2018. A note on the different interpretation of the correlation parameters in the bivariate probit and the recursive bivariate probit. Economics Letters forthcoming 2018.

Graham, B., Hahn, J., Poirier, A.,Powell, J., 2015. Quantile regression with panel data. NBER Working Paper 21034, NBER, Cambridge, MA.

Geraci, M., Bottai, M., 2007. Quantile regression for longitudinal data using the asymmetric Laplace distribution. Biostatistics 8(1) 140-151.

Greene, W., 2004a. The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. Econometrics Journal 7, 98-19.

Greene, W., 2004b. Convenient estimators for the panel probit model. Empirical Economics 29 (1), 21-47.

Greene, W., 2004c. Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the world health organization's panel data on national health care systems. Health Economics 13, 959-980.

Greene, W., 2005. Fixed effects and bias due to the incidental parameters problem in the tobit model. Econometric Reviews 23 (2), 125-147.

Greene 2010a "Testing Hypotheses About Interaction Terms in Nonlinear Models," *Economics Letters*, 107, 2010, pp. 291-296.

Greene, W., 2010b. A sample selection corrected stochastic frontier model. Journal of Productivity Analysis, 41, 15-24.

Greene, W., 2018. Econometric Analysis. Pearson, New York, NY.

Greene, W., Filippini, M., 2015. Persistent and transient productive inefficiency: a maximum simulated likelihood approach. Journal of Productivity Analysis 45 (2)187-196.

Greene, W., McKenzie, C., 2015. An LM test for random effects based on generalized residuals. Economics Letters 127 (1), 47-50.

Gregory, C., Deb, P., 2015. Does SNAP improve your health? Food Policy 50, 11-19.

Han, S., Vytlacil, E., 2017. Identification in a generalization of bivariate probit models with dummy endogenous regressors. Journal of Econometrics 199, 63-73.

Hahn, J., Kuersteiner, G., 2011. Bias reduction for dynamic nonlinear panel models with fixed effects. Econometric Theory 27 (6), 1152-1191.

Hahn, J. and Newey, W., 2004. Jackknife and analytical bias reduction for nonlinear panel models. Econometrica 77, 1295-1313.

Hausman, J., 1978. Specification tests in econometrics. Econometrica 46, 1251-1271.

Hausman, J., Hall, B., Griliches, Z., 1984. Economic models for count data with an application to the patents-r&d relationship. Econometrica 52, 909-938.

Heckman, J., 1979. Sample selection as a specification error. Econometrica 47, 153-161.

Heckman, J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 52, 748-755.

Hensher, D., Jones S., Greene, W., 2007. An error component logit analysis of corporate bankruptcy and insolvency risk in Australia. The Economic Recort 63 (260), 86-103.

Hensher, D., Rose, J., Greene, W., 2015. Applied Choice Analysis, 2nd ed. Cambridge University Press, New York.

Honoré, B., 2002. Nonlinear models with panel data. Portuguese Economic Journal, 2002 (1), 163-179.

Honoré, B., 2013. Non-linear models with panel data. CEMMAP working paper 13/02, IFS, Department of Economics, UCL, London.

Honoré, B., Kesina, M., 2017. Estimation of some nonlinear panel data models with both time-varying and time-invariant explanatory variables. Journal of Business and Economic Statistics, 35 (4), 543-558.

Honoré, B., Lewbel, A., 2002. Semiparametric binary choice panel data models without strictly exogenous regressors. Econometrica 70 (5), 2053-2063.

Imbens, G., Wooldridge, J., 2012. Nonlinear Panel Data Models, Notes 4. National Bureau of Economic Research, Cambridge, MA. < http://www.nber.org/WNE/lect_4_nlpanel.pdf> (accessed 18.01.18).

Jones, A., Schurer, S., 2011. How does heterogeneity shape the socioeconomic gradient in health satisfaction. Journal of Applied Econometrics 26 (4), 549-579.

Krailo, M., Pike, M., 1984. Conditional multivariate logistic analysis of stratified case control studies. Applied Statistics 44 (1), 95-103.

Kumbhakar, S., Colombi, M, Martini, A., Vittadini, S., 2014. Closed skew normality in stochastic frontiers with individual effects and long/short-run efficiency. Journal of Productivity Analysis 42, 123-136.

Kyriazidou, E., 1997. Estimation of a panel data sample selection model. Econometrica, 65 (6), 1335-1364.

Lancaster, T., 2000. The incidental parameters problem since 1948. Journal of Econometrics 95 (2), 391-414.

Lee, S., 2008. Estimating panel data duration models with censored data. Econometric Theory, 24 (5), 1254-1276.

Lewbel, A., Dong, Y. 2015. A simple estimator for binary choice models with endogenous regressors. Econometric Reviews 34, 82-105.

Mandic, P., Norton, E., Dowd, B., 2012. Interaction terms in nonlinear models. Health Services Research 47 (1), 255-274.

Manski, C., 1975. The maximum score estimator of the stochastic utility model of choice. Journal of Econometrics 3, 205-228.

Mourifie, I.,Meango, R., 2014. A note on the identification in two equations probit model with dummy endogenous regressor. Economics Letters 125, 360-363.

Mundlak, Y., 1978. On the pooling of time series and cross sectional data. Econometrica 56, 342-365.

Muris, c., 2017. estimation in the fixed-effects ordered logit model. Review of Economics and Statistics 99 (3), 465-477.

Nerlove, M., 1966. Pooling cross section and time series data in the estimation of a dynamic model: the demand for natural gas. Econometrica 34 (3), 585-612.

Neyman, J.,Scott, E., 1948. Consistent estimates based on partially consistent observations. Econometrica 16, 1-32.

Papke, L., Wooldridge, J., 1996. Econometric methods for fractional response variables with an application to 401(k) panel participation rates. Journal of Applied Econometrics 11 (6), 619-632.

Papke, L., Wooldridge, J., 2008. Panel data methods for fractional response variables with an application to test pass rates. Journal of Econometrics 145 (1-2), 121-133.

Pforr, K., 2011. Implementation of a multinomial logit model with fixed effects. Presented at Ninth German Stata Users Group Meeting, University of Mannheim, Bamberg. < https://core.ac.uk/download/pdf/6278780.pdf > (accessed 18.01.19.).

Pforr, K., 2014. Femlogit – implementation of the multinomial logit model with fixed effects. Stata Journal, 14 (4), 847-862.

Pinar, K., Norton, E., Dowd, B., 2012. Interaction terms in nonlinear models. Health Services Research 47 (1), 255-274.

Pitt, M., Lee, L., 1981. The measurement and sources of technical inefficiency in the Indonesian weaving industry. Journal of Development Economics 9, 43-64.

Raudebush, S., Bryk, A., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd. Ed. Sage, Thousand Oaks, CA.

Rasch, G., 1960. Probabilistic models for some intelligence and attainment tests. In Denmark Paedogiska, Copenhagen, Denmark.

Revelt, D., Train, K., 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. Review of Economics and Statistics 80, 647-658.

Rhine, S., Greene, W., 2013. Factors that contribute to becoming unbanked. Journal of Consumer Affairs 47 (1), 27-45.

Riphahn, R., Wambach, A., Million, A., 2003. Incentive effects in the demand for health care: a bivariate panel count data estimation. Journal of Applied Econometrics 18 (4), 387-405.

Schmidt, P., Sickles, R., 1984. Production frontiers and panel data. Journal of Business and Economic Statistics 2, 367-374.

Scott, A., Schurer, S., Jensen, P., Sivey, P., 2009. The effects of an incentive program on quality of care in diabetes management. Health Economics 18 (9), 1091-1108.

Semykina, A. and Wooldridge, J., 2010. Estimating panel data models in the presence of endogeneity and selection. Journal of Econometrics 157 (2), 375-380.

Stata, 2018. Stata Manual. Stata Press, College Station, TX.

Train, K., 2009. Discrete Choice Methods with Simulation. Cambridge University Press, New York.

Vella, F., 1998. Estimating models with sample selection bias: a survey. Journal of Human Resources 33, 439-454.

Vella, F., Verbeek, M., 1999. Two-step estimation of panel data models with censored endogenous variables and selection bias. Journal of Econometrics 90m 239-263.

Verbeek, M., 1990. On the estimation of a fixed effects model with selectivity bias. Economics Letters 34, 267-270.

Verbeek, M., Nijman, T., 1992. Testing for selectivity bias in panel data models. International Economic Review 33 (3), 267-270.

Wilde, J., 2000. Identification of multiple equation probit models with endogenous dummy regressors. Economics Letters 69, 309-312.

Willis, J., 2006. Magazine prices revisited. Journal of Applied Econometrics, 21 (3), 337-344.

Wooldridge, J., 1995. Selection Corrections for panel data models under conditional mean independence assumptions. Journal of Econometrics 68 (1), 115-132.

Wooldridge, J., 2010. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.

Wooldridge, J., 2005. Simple solutions to the initial conditions problem in dynamic nonlinear panel data models with unobserved heterogeneity. Journal of Applied Econometrics 20 (1), 39-54.

Wooldridge, J., 2003. Cluster sample methods in applied econometrics. American Economic Review 93, 133-138.

Wooldridge, J., 2002. Inverse probability weighted M estimators for sample stratification, attrition and stratification. Portuguese Economic Journal 1, 117-139.

Wu, D., 1973. Alternative tests of independence between stochastic regressors and disturbances. Econometrica 41, 733-750.

Zabel, J., 1992. Estimating fixed and random effects models with selectivity. Economics Letters 40, 269-272.