ESSAYS ON NONLINEAR PANEL MODELS WITH UNOBSERVED HETEROGENEITY

By

Robert Martin

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics – Doctor of Philosophy

2017

ABSTRACT

ESSAYS ON NONLINEAR PANEL MODELS WITH UNOBSERVED HETEROGENEITY

By

Robert Martin

This dissertation concerns nonlinear panel data estimation relevant to the fields of econometrics and applied microeconomics. Panel data is attractive for estimating causal effects when unobserved heterogeneity in cross-sectional units is correlated with explanatory variables. For instance, well-known linear fixed effects and first difference estimators use within-group variation to achieve consistent estimation. However, nonlinear models often better represent limited dependent variables like binary outcomes or counts, and extending traditional panel techniques to these settings can be problematic. For instance, treating heterogeneity as parameters to be estimated usually leads to what is known as the incidental parameters problem. Furthermore, heterogeneous slopes in a conditional mean function can also confound estimation, but fewer remedies exist than do for additive effects. I aim to address these issues in my research with an emphasis on practical applicability.

Chapter 1: Finite sample properties of bias-corrected fixed effects estimators for panel binary response models

Maximum likelihood estimation (MLE) of nonlinear unobserved effects panel models is known to be generally inconsistent when treating the heterogeneity as parameters. Several authors have proposed corrections justified by large-T expansions of the inconsistency under conditions like dynamic completeness. Using Monte Carlo (MC) techniques, I find that failure of dynamic completeness can increase bias in slope and average partial effects (APE) estimates in shorter panels, but has little impact on APE for longer panels. I also compare bias-corrections to correlated random effects (CRE) and Conditional MLE using MC and welfare data from the Survey of Income and Program Participation (SIPP).

Chapter 2: Exponential panel models with coefficient heterogeneity

If heterogeneous slopes are ignored in exponential panel models, fixed effects Poisson may not estimate any quantity of interest. Existing estimation methods often involve treating only a small subset of the slopes as "random effects" and integrating from the likelihood, increasing computational difficulty. I propose a test to detect slope heterogeneity that, unlike the traditional approach, does not amount to testing for information matrix equality. Additionally, I present a correlated random coefficients approach to identification which allows for estimation of the coefficient means and average partial effects. I test these proposed methods using a Monte Carlo experiment and apply them to the patent-R&D relationship for U.S. manufacturing firms.

Chapter 3: Estimation of average marginal effects in multiplicative unobserved effects panel models

This chapter concerns estimation of average marginal effects in static multiplicative unobserved effects panel models for nonnegative dependent variables. While fixed effects Poisson (FEP) consistently estimates the parameters of the conditional mean function, marginal effects generally depend on the unobserved heterogeneity. They would therefore seem inestimable without either additional assumptions or some form of bias correction. I show, however, that Average Partial Effect (APE) and Average Treatment Effect (ATE) estimators that use estimated individual effects are consistent and asymptotically normal. This is in contrast with cases like fixed effects logit, where similar marginal effects estimators suffer from the incidental parameters problem.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the chair of my dissertation committee, Jeff Wooldridge, for all of his advice, encouragement, and helpful critiques. I would also like to thank Peter Schmidt, Kyooil Kim, and Nicole Mason for serving on my committee and providing valuable feedback and assistance. I also appreciate the comments of seminar participants at Michigan State University, the 2016 MEA Conference, and the 2016 Annual Meeting of the Midwest Econometrics Group.

I am especially grateful for the financial support I received from the Graduate School and the Department of Economics at Michigan State University, including the David Kelley Fellowship, Summer Research Fellowship, and Dissertation Completion Fellowship. I also appreciate the support and advice that Lori Jean Nichols, Steven Haider, Todd Elder, and Steve Woodbury all gave me as I navigated the graduate program and job market.

Finally, I cannot thank my family enough for their support and encouragement. I am especially grateful to my wife, Kara, for moving with me to East Lansing and then to Washington DC, as well as all the countless ways she has supported my endeavors over the years.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

FINITE SAMPLE PROPERTIES OF BIAS-CORRECTED FIXED EFFECTS ESTIMATORS FOR PANEL BINARY RESPONSE MODELS

1.1 Introduction

Nonlinear models are popular in economics in many settings. For instance, binary response models are common for analyzing outcomes like labor force participation, employment, or union membership. At the same time, panel data can be attractive when controlling for unobserved heterogeneity is necessary to identify causal effects. However, it is well-known that maximum likelihood estimation (MLE) that treats heterogeneity as parameters to estimate is inconsistent. For example, in the case of cross-section heterogeneity, the problem arises in the typical large-N, fixed-T microeconometric setting because only a handful of observations contribute to the estimation of each individual's fixed effect (Lancaster, 2000). This is known as the incidental parameters problem, first described by Neyman and Scott in 1948.

In the statistics and econometrics literature, there have been many approaches to estimation in the presence of incidental parameters. In some special cases, it is possible to re-parameterize the model or find a conditioning variable that removes the incidental parameters from the likelihood function (Lancaster, 2000). A leading example of this is the conditional logit model, where the conditioning variable is the number of successes observed for cross-sectional unit (Chamberlain, 1980). However, while conditional maximum likelihood in a case like this consistently estimates the slope parameters of the index of the logit function, conditioning usually does not identify partial effects, which depend on the heterogeneity (Wooldridge, 2010). Other approaches involve restricting the relationship between the heterogeneity and explanatory variables in some way. For instance, if we are willing to assume independence between the heterogeneity and the explanatory variables, then we can use a random effects approach. In many cases, however, correlation between heterogeneity and covariates is of concern. The correlated random effects (CRE) approach

of Chamberlain (1980, 1982) or Mundlak (1978), restricts the conditional distribution of the heterogeneity to have a mean that is linear function of the explanatory variables, but the restriction at least buys the researcher identification of APE and scaled slope parameters (Wooldridge, 2010). Assumptions restricting the nature of the heterogeneity are a potential drawback. For instance, Rabe-Hesketh and Skrondal (2013) explore a special case in the dynamic probit setting where misspecification of the heterogeneity causes significant bias . In general, however, we do not know the robustness of CRE is when the distributional assumption fails or when the researcher chooses the wrong conditional mean function.

If one prefers to leave the nature of the heterogeneity completely unrestricted, a linear probability model (LPM) estimated by fixed effects ordinary least squares is thought to do a reasonable job approximating, and even consistently estimates them under certain assumptions regarding the explanatory variables (Stoker, 1986). Nevertheless, often the index slope parameters are of interest, or the researcher wants to estimate partial effects at different values of the explanatory variables. In these cases it is tempting to use a nonlinear "fixed effects" estimator, whereby the heterogeneity are estimated as parameters alongside the index slopes in a MLE procedure, but this is problematic. Particularly when the number of time periods is small, fixed effects estimators often perform worse than simply ignoring the heterogeneity entirely (Greene, 2004). In the case of cross-sectional heterogeneity only, several studies have noted that inconsistency diminishes as the number of time periods increases, and that estimates of slope parameters are consistent with both N and T growing to infinity. However, the asymptotic distribution of fixed effects estimators is not centered around the true parameter values, so confidence intervals can still be misleading (Hahn and W. Newey, 2004).

I study bias corrections for models with cross-sectional heterogeneity that subtract the leading term of a large-T expansion of the bias from the uncorrected fixed effects MLE. Analytical bias corrections estimate this term from expressions specific to the parametric model. Jackknife corrections estimate it non-parametrically by generating variation in the uncorrected MLE by dropping some time periods. These techniques reduce the bias from $O_p(T^{-1})$ to $O_p(T^{-2})$, but they can

require significant restrictions on the underlying distribution of the data (Hahn and W. Newey, 2004). Both approaches assume at least that the explanatory variables are stationary and weakly dependent. The analytical and jackknife corrections developed by Hahn and Newey (2004) also require the dependent variables to be serially independent conditional on the heterogeneity and the explanatory variables. The analytical correction of Fernandez-Val (2009) and the split-panel jackknife of Dhaene and Jochmans (2015) relax conditional independence to accommodate models with lagged dependent variables, but still require dynamic completeness.

Either conditional independence, or dynamic completeness rule out serially correlated error terms, which is potentially a serious problem for static models. Serial correlation is certainly a concern in linear models, as demonstrated by widespread use of clustered standard errors and post-estimation testing. Extending that concern to nonlinear models is particularly prudent given that in cases like the probit or logit, serial correlation causes inconsistency in the estimators themselves, not just their standard errors. Without unobserved heterogeneity, APE are still identified in probit or logit models with serial correlation, so the problem is easily handled by using pooled MLE with cluster-robust standard errors (Wooldridge, 2010). To my knowledge, however, no researchers have simulated bias-corrected estimators in the presence of serial correlation.

This chapter aims to answer three questions. First, how robust are bias corrections when latent errors have serial correlation? Second, how do the bias corrections compare to the CRE approach when the heterogeneity does not satisfy the CRE conditional distribution assumption? Finally, the incidental parameters problem causes bias not only in slope estimates, but in APE estimates as well, but how severe is bias in APE estimates when the slopes are estimated consistently with a procedure like conditional logit?

The first goal is to inform practitioners who wish to account for unobserved heterogeneity while being agnostic about serial dependence. Using Monte Carlo techniques, I evaluate the impact of serially correlated errors on the analytical bias corrections of Hahn and Newey (2004) and Fernandez-Val (2009). I also evaluate the drop-one-period jackknife of Hahn and Newey (2004) and the split-panel jackknife of Dhaene and Jochmans (2015). I generate the error terms in the

latent variable model as first order autoregressive processes, but simulate estimators that use clustered standard errors to allow for general (weak) serial dependence. Since slope parameters are only identified up to scale in this setting, I focus primarily on estimation of APE, which are still identified (Wooldridge, 2010).

While simulation evidence from the aforementioned studies shows that bias-corrected estimators often have much more desirable finite sample properties than the uncorrected fixed effects MLE (at least for slope parameters), less work has been done to evaluate sensitivity of these properties to relaxation of the assumptions underlying the corrections. Dhaene and Jochmans (2015) examine departures from stationarity in dynamic models, particularly of initial observations and propose a Wald test for evaluating the validity of the split-panel approach overall. Alexander and Breunig (2014) simulate the performance of several bias corrections for the fixed effects probit estimator while varying parameters like the variance of the heterogeneity and correlation between heterogeneity and explanatory variables. but do not consider any departures from stationarity or conditional independence.

In addition to using clustered standard errors, many researchers will find it attractive to make a CRE assumption to avoid the issue of incidental parameters. In fact, in studying the issue of serial correlation, many of my simulation results show that the CRE estimator of APE tends to have better finite sample properties than the uncorrected or corrected fixed effects methods. This result is not surprising given the data generating process I employ. Therefore, my second contribution is to consider the relative performance of the CRE approach versus the fixed effects approach when the CRE conditional distribution assumption does not hold.

Finally, if researchers are willing to assume the dependent variables are conditionally independent, then a logit specification can be attractive because conditional maximum likelihood estimation (conditioning on the individual's sum of the dependent variables) allows for consistent estimation of slope parameters with only $N \to \infty$. However, partial effects are not identified because they depend on the heterogeneity terms that have been conditioned out of the likelihood function. Nevertheless, it is tempting to implement the following procedure: 1) Estimate slope

parameters by conditional MLE. 2) Estimate the heterogeneity parameters using logit MLE, while restricting the slopes to be equal to the estimates from stage 1), and then estimate partial effects. For instance, an empirical example in Greene (2012, Chapter 17) on German health care utilization follows this procedure in estimating partial effects evaluated at the average of the explanatory variables (PEA) (Greene, 2012). This procedure is likely to suffer from the incidental parameters problem because, although the slope parameters are consistent, the heterogeneity estimates still do not converge to anything with fixed T (and it is unclear if the sample average of the estimated heterogeneity converge to anything interesting as N gets large). Fernandez-Val (2009) uses this procedure to estimate a model of female labor force participation, but corrects the APE estimates for the incidental parameters problem in the second stage (Fernandez-Val, 2009). Therefore, this chapter's third contribution is to include Monte Carlo evidence that uncorrected APE estimates derived in this manner from conditional logit estimation can have significant bias.

Strictly speaking, any conclusions drawn from these simulations are valid only for the data generating processes I employ. However, the results presented are still useful in alerting empirical researchers to potential benefits and pitfalls when implementing one of the discussed estimation methods.

The rest of the paper is organized as follows. Section 2 reviews the incidental parameters problem in the panel binary response model, as well as the bias correction techniques considered here. Section 3 describes the Monte Carlo experiment. Section 4 presents and discusses results including the application to the SIPP data. Section 5 concludes. Additional tables, as well as descriptions of the analytical bias correction formulas, are collected in Appendices.

1.2 The panel binary response model with incidental parameters

I consider the following panel binary response model with unobserved heterogeneity.

$$y_{it} = \mathbf{1}\left[\alpha_i + \boldsymbol{x}_{it}\theta_0 + r_{it} > 0\right], \text{ for } i = 1,\ldots,N \text{ and } t = 1,\ldots,T. \tag{1.1}$$

where y_{it} is a scalar outcome variable, $\boldsymbol{x}_{it}$ is a vector of explanatory variables, α_i is an individual fixed effect, and r_{it} is a error term. In the probit (logit) case, r_{it} is distributed standard normal (standard logistic). $\mathbf{1}[\cdot]$ is the indicator function. The log-likelihood function for individual i in period t is

$$\ell_{it}(\theta, \alpha_i) = y_{it} \log[G(\alpha_i + \boldsymbol{x}_{it}\theta)] + (1 - y_{it}) \log[1 - G(\alpha_i + \boldsymbol{x}_{it}\theta)], \tag{1.2}$$

where G is either the standard normal CDF or standard logistic CDF. Following the notation of Hahn and Newey (2004) and Fernandez-Val (2009), the maximum likelihood estimator of θ_0 maximizes the profile log-likelihood, concentrating out the alphas:

$$\widehat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T} \ell_{it}(\theta, \widehat{\alpha}_i(\theta))/NT \tag{1.3}$$

where

$$\widehat{\alpha}_i(\theta) = \arg\max_{\alpha} \sum_{t=1}^{T} \ell_{it}(\theta, \alpha_i)/T \tag{1.4}$$

The incidental parameters problem arises because with T fixed, as $N \to \infty$,

$$\widehat{\theta} \xrightarrow{p} \theta_T, \text{ where } \theta_T = \arg\max_{\theta} \mathbb{E}_N \left[\sum_{t=1}^{T} \ell_{it}(\theta, \widehat{\alpha}_i(\theta))/T \right] \tag{1.5}$$

where $\mathbb{E}_N[m(Z_{it}, \alpha_i)] \equiv \lim_{N\to\infty} \sum_{i=1}^{N} m(Z_{it}, \alpha_i)/N$. For finite T, $\theta_T \neq \theta_0$ because $\widehat{\alpha}(\theta) \neq \alpha_i$, even when evaluated at the true θ_0. Hahn and Newey (2004) show that for smooth likelihoods like the probit and logit,

$$\theta_T = \theta_0 + \mathscr{B}/T + O(T^{-2}) \tag{1.6}$$

where $\mathscr{B} = \mathscr{I}^{-1} b$. In this expression, b represents a higher order expansion of the bias in $\widehat{\alpha}(\theta)$ as T gets large, while $\mathscr{I}$ is the information matrix of the profile log-likelihood. Both terms together capture the effect of estimation error in $\widehat{\alpha}(\theta)$ on $\widehat{\theta}$. While it is true that $\widehat{\theta}$ is consistent for θ_0 if both N and $T \to \infty$, the limiting distribution of $\sqrt{NT}(\widehat{\theta} - \theta_0)$ is centered around $\mathscr{B}\sqrt{\kappa}$, where $N/T \to \kappa$. Therefore, confidence intervals for coefficient estimates will likely have poor coverage (Hahn and W. Newey, 2004).

1.2.1 Bias correction techniques

Arellano and Hahn (2007) provide a thorough review of different approaches to mitigating bias from the incidental parameters problem. The techniques that I consider in this chapter involve estimating $\mathscr{B}$ and using it to construct an estimator with a bias of lower order. Analytical bias corrections use expressions for $\mathscr{B}$ (denoted for an arbitrary θ as $\mathscr{B}(\theta)$) derived from large-T expansion of the scores of the profile log-likelihood around the true α_i. I focus mainly on the "one-step" estimator $\mathscr{B}(\widehat{\theta})$, which is evaluated at the uncorrected MLE. Then the bias corrected estimator is formed as

$$\widetilde{\theta}_{bc} = \widehat{\theta} - \mathscr{B}(\widehat{\theta})/T \tag{1.7}$$

Previous simulations have shown that the one-step estimator performs reasonably well compared to an iterated procedure or related analytical corrections that solve modified scores (Hahn and W. Newey, 2004). I examine the methods of Hahn and Newey (2004) and Fernandez-Val (2009) for estimating $\mathscr{B}(\widehat{\theta})$. Full expressions for the analytical bias corrections can be found in Appendix A.

Jackknife corrections estimate $\mathscr{B}$ nonparametrically by using variation in $\widehat{\theta}$ when estimated over the full panel and shorter sub-panels. This approach is advantageous because it does not require an explicit characterization of $\mathscr{B}$, though it does require more computation. Hahn and Newey (2004) proposed a technique where the MLE is estimated over the T subpanels formed by dropping one period. Their corrected estimator is formed as

$$\widetilde{\theta}_{hnjk} = T\widehat{\theta} - \frac{T-1}{T}\sum_{s=1}^{T}\widehat{\theta}_s, \tag{1.8}$$

where $\widehat{\theta}_s$ is the uncorrected MLE estimated over the periods $\{1,\ldots,s-1,s+1,\ldots,T\}$.

Dhaene and Jochmans (2015) show that splitting the panel into equal, or almost-equal, length sub-panels minimizes the impact of imprecise estimation of $\mathscr{B}$ on the remaining bias and allows for dynamic models (Dhaene and Jochmans, 2015). To illustrate how the estimator is formed, suppose T is even for simplicity. Let $\widehat{\theta}_{S_1}$ and $\widehat{\theta}_{S_2}$ be the uncorrected MLE estimated over the

periods $\{1,2,\ldots,T/2\}$ and $\{T/2+1,\ldots,T\}$. Then, the jackknife corrected estimator is formed as

$$\widetilde{\theta}_{djjk} = 2\widehat{\theta} - (1/2)(\widehat{\theta}_{S_1} + \widehat{\theta}_{S_1}). \tag{1.9}$$

Researchers are often interested in estimating functions of the data and parameters, like the partial effect of the kth element of $\boldsymbol{x}_{it}$ on the probability that y_{it} equals one:

$$m_k(\theta, \alpha_i, \boldsymbol{x}_{it}) = \theta_k g(\alpha_i + \boldsymbol{x}_{it}\theta), \tag{1.10}$$

where $g()$ is the derivative of $G()$. Many past simulation and theoretical work has suggested that uncorrected MLE on static binary response models has a "small bias" property for estimates of APE. This means that the bias in APE estimates tends to be smaller than that of slope parameters, and in the probit case with no heterogeneity, is exactly zero (Fernandez-Val, 2009). This suggests that biases in $\widehat{\theta}_k$ and $\sum_{i=1}^{N}\sum_{t=1}^{T} g(\widehat{\alpha}_i + \boldsymbol{x}_{it}\widehat{\theta})$ move in opposite directions. Since APE and other functions of the data generally depend directly on the α's, correcting the slope parameters only (or using a consistent procedure like conditional logit) is insufficient to handle the incidental parameters problem as it only removes one source of the bias. In fact, $\widehat{\alpha}_i(\theta)$, even if evaluated at θ_0, does not converge to its true value with T fixed, or at a slower rate when T is allowed to grow (Fernandez-Val, 2009). APE estimates with consistent estimates of θ but no correction for imprecise estimation of the α's may have much larger biases than APE estimates derived from the uncorrected MLE, as section IV explores.

The analytical and jackknife corrections for APE are implemented in a similar fashion to their counterparts for slope estimates. In the analytical case (see Appendix A), a bias term is estimated and subtracted, while for the jackknife, APE are estimated for the full panel and the subpanels separately and then combined just like the slope estimates.

Under dynamic completeness for the Fernandez-Val case and conditional independence for the Hahn and Newey case, analytical bias-corrected estimators been shown to be consistent and asymptotically normal as long as T grows faster than $N^{1/3}$, and a similar property has been conjectured for the Hahn and Newey jackknife correction (Hahn and W. Newey, 2004). This makes

them reasonable procedures to implement when N is fairly large relative to T, as is typical in microeconometrics. The split-panel jackknife of Dhaene and Jochmans is only consistent with T and N growing at the same rate, but they find evidence it reduces bias with as few as six time periods. The analytical and jackknife corrections analyzed here allow explanatory variables to be only sequentially exogenous, but require the assumption of dynamic completeness, meaning that no additional lags of $\boldsymbol{x}$ or y affect the current y_{it} after $\boldsymbol{x}_{it}$ has been included. Dynamic completeness is written formally as

$$f(y_{it}|\alpha_i, \boldsymbol{x}_{it}, y_{i,t-1}, \boldsymbol{x}_{i,t-1} \ldots, y_{i1}, \boldsymbol{x}_{i1}) = f(y_{it}|\alpha_i, \boldsymbol{x}_{it}) \tag{1.11}$$

Either conditional independence or Assumption (1.11) imply that the scores of the log-likelihood are serially uncorrelated and rule out any serial dependence in the per-period shocks. For the many researchers interested in estimating static models, however, this assumption is less than ideal. Empirical researchers routinely encounter static models with neglected serial correlation in the linear case, and take care to conduct inference using clustered standard errors. Consequently, we would rather not assume that a static model has fully captured the dynamics in the nonlinear case either. One attractive point about the CRE approach with clustered standard errors is that for binary response models with unobserved heterogeneity, arbitrary serial correlation do not cause inconsistency in APE estimates (Wooldridge, 2010). Any complete comparison of bias corrections, therefore, should evaluate their robustness to this common problem.

1.3 Monte Carlo experiment

The data generating process I specify is similar to Greene (2004) and Fernandez-Val and Weidner (2016). The outcome is generated as

$$y_{it} = \mathbf{1}\left[\alpha_i + \beta_0 x_{it} + \gamma_0 d_{it} + r_{it} > 0\right] \tag{1.12}$$

where

$$x_{it} = \alpha_i + .5x_{i,t-1} + v_{it}, t > 1 \tag{1.13}$$

$$x_{i1} = \alpha_i + v_{i1},\, v_{it} \sim N(0, 1/2) \tag{1.14}$$

$$d_{it} = \mathbf{1}\left[x_{it} + h_{it} > 0\right],\, h_{it} \sim N(0, 1/2) \tag{1.15}$$

$$\alpha_i \sim N(0, 1/16) \tag{1.16}$$

I set $\beta_0 = \gamma_0 = 1$. In this model, d_{it} represents a policy or treatment variable of interest, while x_{it} is a continuous control variable that is both correlated with d_{it} and its own past values. Both x_{it} and d_{it} are generated to be strictly exogenous, though the Fernandez-Val and Dhaene and Jochmans corrections only require sequential exogeneity. Correlation between x_{it} and α_i is roughly 0.5, while correlation between d_{it} and α_i is roughly 0.3. Correlation between x_{it} and d_{it} is about 0.6.

Let μ_w be the population APE of w on the probability that y equals one, for $w \in \{x, d\}$. In general, this quantity varies by T, so for comparison, I report the estimated APE divided by their true value. For $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\alpha}$ (the uncorrected MLE),

$$\frac{\hat{\mu}_w}{\mu_w} = \frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m_w(\hat{\beta}, \hat{\gamma}, \hat{\alpha}_i, z_{it})}{E\left[\frac{1}{T}\sum_{t=1}^{T} m_w(\beta_0, \gamma_0, \alpha_i, z_{it})\right]}, \tag{1.17}$$

where $z_{it} = (x_{it}, d_{it})$ and

$$m_w(\beta, \gamma, \alpha_i, z_{it}) = \begin{cases} \beta\, g(\alpha_i + \beta x_{it} + \gamma d_{it}) & \text{for } w = x \\ G(\alpha_i + \beta x_{it} + \gamma) - G(\alpha_i + \beta x_{it}) & \text{for } w = d \end{cases} \tag{1.18}$$

where for the probit (logit) simulations, $G()$ and $g()$ are the CDF and PDF, respectively, for the standard normal (logistic) distribution. The expectation in the denominator is simulated with a single draw from a panel of $1,000,000$ individuals. Note that the sum in the numerator is divided by the entire sample size, NT. An individual j whose value of y_{jt} does not change over the length of the panel gets, the uncorrected MLE for the heterogeneity, $\hat{\alpha}_j$, is unbounded, so the individual is dropped from the estimation of the structural parameters. The estimate $m_w(\hat{\beta}, \hat{\gamma}, \hat{\alpha}_j, z_{jt})$ for that observation is zero (Alexander and Breunig, 2014). I will discuss practical issues this can cause when the panels are short and the data are highly persistent. Details on the analytical corrections can be found in Appendix A. The jackknife-corrected APE estimators are constructed analogously to the slope estimators in equations (1.8) and (1.9).

1.3.1 Evaluating the dynamic completeness assumption

I relax dynamic completeness in the panel probit case by introducing serial correlation into the error term r_{it} from the latent variable model. I use the following procedure:

$$r_{it} = \psi_{t,\rho} u_{it} \tag{1.19}$$

$$u_{it} = \rho u_{i,t-1} + e_{it}, t > 1 \tag{1.20}$$

$$u_{i1} = e_{i1}/\psi_{t,\rho}, e_{it} \sim i.i.d.N(0,1) \tag{1.21}$$

$$\psi_{t,\rho} \equiv \begin{cases} \sqrt{1-\rho^2} & \text{if } \rho < 1 \\ 1/\sqrt{t} & \text{if } \rho = 1 \end{cases} \tag{1.22}$$

Division of e_{i1} by $\psi_{t,\rho}$ ensures that each element of $\{u_{it}\}_{t=1}^{T}$ has the same variance, which otherwise would not hold because of finite length (Vamoş, Şoltuz, and Crăciun, 2007). Multiplication of u_{it} by $\psi_{t,\rho}$ is to give r_{it} unit variance. I maintain unit variance of the error terms to remove the coefficient scaling that would otherwise occur in probit MLE. This allows us to better compare slope estimates across estimators and values of ρ. In the logit case, I use a Gaussian copula based on these series of normal errors.

I present results from simulations that set ρ equal to 0, 0.4, 0.8 to represent cases of dynamic completeness, moderate serial correlation, and high serial correlation. While the copula is not guaranteed to maintain the exact serial correlation for the logit case, the autocorrelations were within two decimal points of the specified ρ. Consistent with the literature, I considered panel lengths of 6, 8, 12, and 20, and I set $N = 100$ in all cases for ease of computation. Previous work by Fernandez-Val (2009) and Alexander and Breunig (2014) has found that the larger N does not affect the relative performance of the different estimators in terms of bias, but does increase their overall precision. I find evidence of these findings, which can be found in Appendix B for the $N = 500$, $T = 6$ case. One important finding is that when estimators have finite sample bias, coverage of confidence intervals generally decreases with sample size as standard errors shrink.

I also estimate the probit slope coefficients and APE using the pooled MLE version of Mundlak's (1978) correlated random effects (CRE), and the APE using a LPM for comparison. Standard errors for each estimator are clustered by individual to account for serial dependence in the scores. For each pair of ρ and T, I run 1000 replications.

1.3.2 Comparing bias correction and CRE under more general forms of heterogeneity

A correlated random affects approach of Mundlak (1978) applied to the panel probit model with two strictly exogenous explanatory variables assumes that

$$D(c_i|\boldsymbol{x}_i,\boldsymbol{d}_i) = Normal(\psi + \xi_1\bar{x}_i + \xi_2\bar{d}_i, \sigma_a^2), \tag{1.23}$$

which implies

$$D(y_{it}|\boldsymbol{x}_i,\boldsymbol{d}_i) = Probit(\beta_a x_{it} + \gamma_a d_{it} + \psi_a + \xi_{1,a}\bar{x}_i + \xi_{2,a}\bar{d}_i) \tag{1.24}$$

where $\bar{x}_i$ and $\bar{d}_i$ denote time averages, and the "a" subscript indicates the coefficients are scaled by $1/\sqrt{1+\sigma_a^2}$. Therefore, pooled probit of y_{it} on x_{it}, d_{it}, $\bar{x}_i$ and $\bar{d}_i$ identifies β and γ up to scale. Since the APE depend on the scaled coefficients, they can be estimated consistently with no problem (Wooldridge, 2010).

Tables 1.1-1.4 in Section 4 show that CRE used on probit data generated with the above process (or similarly for the logit case) performs well because the heterogeneity enters the equation for x_{it} additively; therefore, the α_i can be written as a linear function of the time averages of x_{it}. Consequently, a natural question, is how much better do the fixed effects approaches perform when the CRE assumption fails?

I explore this question with the panel probit model through the following modifications:

$$y_{it} = \mathbf{1}\left[\alpha_{j,i} + \beta_0 x_{it} + \gamma_0 d_{it} + r_{it} > 0\right] \tag{1.25}$$

$$x_{it} = .5x_{i,t-1} + v_{it}, t > 1 \tag{1.26}$$

$$x_{i1} = v_{i1}, v_{it} \sim N(0, 1/2) \tag{1.27}$$

Where $\alpha_{j,i}$ is one of:

$$\alpha_{1,i} = -1 + \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_{it}^2 + a_i \tag{1.28}$$

$$\alpha_{2,i} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (x_{it} + x_{it}^2 + x_{it}^3) + a_i \tag{1.29}$$

$$\alpha_{3,i} \sim N\left(0, \exp\left[\frac{0.125}{\sqrt{T}} \sum_{t=1}^{T} (x_{it} + x_{it}^2 + x_{it}^3)\right]\right) \tag{1.30}$$

Where in the first two cases, $a_i \sim N(0, 1/4)$.

Table 1.5 compares the uncorrected fixed effects MLE, MLE with Fernandez-Val's analytical bias correction, and two estimators based on CRE. One adds $\bar{x}_i$ and $\bar{d}_i$ to the probit index, and a more flexible version (CRE2), where the index includes squares of $\bar{x}_i$ and $\bar{d}_i$ and interactions between the explanatory variables and time averages. I consider panels with $T = 6$ and $T = 12$, for the dynamically complete case.

1.3.3 Conditional logit and the importance of correcting APE estimates

To evaluate the finite sample properties of APE estimates derived from conditional logit slope estimates, I generate a panel of logit dependent variables using the process described in (12). I only consider the $\rho = 0$ case, as conditional logit is not valid when dynamic completeness fails. I estimate APE using the uncorrected logit MLE, and two conditional logit procedures which estimate the heterogeneity with a restricted MLE as described on the introduction. One procedure does not correct for the incidental parameters problem while the other uses Fernandez-Val's 2009 correction for the logit case.

1.4 Results

For brevity, I mainly report the bias corrections for the probit case. The logit case is qualitatively similar, though the effect of serial correlation on the Fernandez-Val correction is much less severe. I also report only the $T = 6$ and $T = 12$ results as they seem to be representative of the short panel

and long panel cases, respectively. Each of the tables lists the mean and standard deviation of the estimator, the coverage probability of a 95% confidence interval, and the ratio of the estimated (cluster-robust) standard error to the standard deviation. They show quite an interesting range of performance for both the uncorrected MLE and the different bias reduction techniques. Results for the

1.4.1 Evalauating the dynamic completeness assumption

Tables 1.1 and 1.2 show the performance of the probit slope estimators for different levels of serial correlation. In line with evidence from the literature, the uncorrected MLE can be severely biased for the index slopes in the presence of incidental parameters.

Table 1.1: Probit Estimates of β ($\beta_0 = 1$)

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$
T=6												
MLE	1.36	0.24	0.70	0.96	1.56	0.30	0.48	0.90	2.49	0.55	0.05	0.83
A-FV09	0.96	0.14	0.97	1.15	1.03	0.14	0.97	1.17	0.63	0.59	0.58	0.29
A-HN04	1.18	0.21	0.87	0.92	1.36	0.26	0.66	0.83	2.24	0.52	0.07	0.72
J-DJ15	0.85	0.34	0.64	0.46	0.73	0.50	0.49	0.35	0.80	1.03	0.39	0.28
J-HN04	0.87	0.16	0.82	0.95	0.99	0.20	0.90	0.82	1.43	0.45	0.49	0.50
CRE	1.01	0.14	0.95	0.99	1.01	0.15	0.94	0.98	1.02	0.15	0.93	0.95
T=12												
MLE	1.14	0.12	0.79	0.99	1.22	0.13	0.61	0.99	1.61	0.19	0.05	0.96
A-FV09	1.00	0.10	0.95	1.03	1.05	0.11	0.94	1.02	1.33	0.14	0.32	0.98
A-HN04	1.03	0.10	0.94	1.00	1.10	0.12	0.87	0.98	1.45	0.16	0.16	0.92
J-DJ15	0.94	0.12	0.82	0.78	0.90	0.16	0.69	0.62	0.75	0.32	0.39	0.31
J-HN04	0.96	0.09	0.93	1.03	1.02	0.10	0.95	1.01	1.30	0.14	0.38	0.95
CRE	0.99	0.09	0.95	1.01	0.99	0.10	0.94	0.98	1.00	0.10	0.95	1.01

Table 1.2: Probit Estimates of γ ($\gamma_0 = 1$)

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$
T=6												
MLE	1.31	0.26	0.79	0.98	1.52	0.30	0.59	0.95	2.49	0.77	0.10	0.65
A-FV09	0.95	0.16	0.98	1.26	1.02	0.16	0.99	1.32	0.49	0.82	0.62	0.29
A-HN04	1.14	0.22	0.91	1.00	1.33	0.27	0.78	0.93	2.25	0.76	0.13	0.54
J-DJ15	0.78	0.73	0.76	0.30	0.24	1.53	0.54	0.18	-1.68	2.58	0.17	0.21
J-HN04	0.87	0.17	0.93	1.18	0.95	0.32	0.96	0.65	0.85	1.79	0.59	0.17
CRE	0.98	0.17	0.95	0.99	0.99	0.16	0.96	1.01	1.00	0.15	0.95	0.99
T=12												
MLE	1.15	0.14	0.82	1.00	1.23	0.15	0.65	0.99	1.61	0.20	0.11	0.97
A-FV09	1.01	0.12	0.97	1.10	1.07	0.12	0.95	1.07	1.33	0.15	0.44	1.05
A-HN04	1.04	0.12	0.96	1.06	1.11	0.13	0.89	1.03	1.44	0.18	0.25	0.96
J-DJ15	0.95	0.13	0.91	0.93	0.91	0.16	0.83	0.82	0.64	0.70	0.48	0.21
J-HN04	0.97	0.11	0.96	1.13	1.03	0.12	0.97	1.09	1.30	0.15	0.52	1.06
CRE	1.00	0.11	0.95	1.00	1.00	0.11	0.95	0.98	1.00	0.11	0.94	0.97

In the dynamically complete case ($\rho = 0$), bias diminishes as T grows, there is still room for improvement even when $T = 12$. For instance, the uncorrected MLE for γ has a bias of 31% when $T = 6$, but only 15% when $T = 12$. As predicted by theory, coverage of the 95% confidence interval is still somewhat low at 0.82 when $T = 12$, meaning for a 5% significance level, one would expect to reject a true null hypothesis 18% of the time. As found in previously published simulations, the correction techniques reduce bias and generally increase coverage. In particular, Fernandez-Val's analytical correction performs better than the others in all panels, both in terms of bias and variance, particularly for the short panels. The split panel jackknife of Dhaene and Jochmans tends to have higher variance than the others.

If one is concerned primarily with estimating APE, however, the incidental parameters problem clearly has much less bite, as shown by Tables 1.3 and 1.4. For the dynamically complete case, bias in the uncorrected MLE for μ_x is less than 1% for either panel length, while the bias in that of μ_d is 4% or less. This supports the "small bias" property for APE estimators found by many previous studies of static models (Fernandez-Val, 2009). The bias corrected estimators perform well for the longer panels, but even in the dynamically complete case, many of them have higher bias than the uncorrected MLE for the short panels. Among the different bias correction techniques, both corrections from Hahn and Newey (2004) tend to have the smallest bias, while the split panel jackknife does worse. Additionally, while theory suggests that both corrections reduce bias without any change in variance, it appears that the jackknife corrections may increase variance, especially in shorter panels.

Table 1.3: Probit Estimates of $\widehat{\mu}_x/\mu_x$ (true value = 1)

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$
T=6												
MLE	1.00	0.14	0.94	0.96	0.99	0.14	0.93	0.92	0.94	0.14	0.86	0.86
A-FV09	0.96	0.13	0.93	0.96	0.94	0.13	0.90	0.94	0.57	0.44	0.37	0.28
A-HN04	1.05	0.15	0.89	0.88	1.06	0.16	0.88	0.81	1.05	0.16	0.82	0.72
J-DJ15	1.10	0.19	0.78	0.69	1.15	0.22	0.71	0.66	1.26	0.22	0.58	0.70
J-HN04	1.04	0.15	0.89	0.83	1.07	0.17	0.84	0.74	1.15	0.19	0.63	0.58
CRE	1.01	0.13	0.96	1.03	1.01	0.13	0.95	1.02	1.01	0.13	0.95	0.99
LPM	0.95	0.13	0.94	1.03	0.95	0.13	0.94	1.02	0.94	0.13	0.92	1.00
T=12												
MLE	1.00	0.09	0.94	0.96	0.99	0.09	0.93	0.93	0.99	0.09	0.90	0.89
A-FV09	0.99	0.09	0.93	0.94	0.99	0.09	0.93	0.91	0.98	0.09	0.89	0.86
A-HN04	1.00	0.09	0.93	0.93	1.00	0.10	0.93	0.90	1.01	0.10	0.90	0.85
J-DJ15	1.00	0.11	0.89	0.81	1.01	0.12	0.84	0.72	1.05	0.13	0.78	0.67
J-HN04	1.00	0.09	0.93	0.93	1.00	0.09	0.93	0.90	1.00	0.09	0.90	0.84
CRE	1.00	0.09	0.96	1.01	1.00	0.09	0.94	0.97	1.00	0.09	0.95	0.97
LPM	0.93	0.09	0.88	1.03	0.93	0.09	0.86	1.00	0.93	0.09	0.87	0.99

Table 1.4: Probit Estimates of $\widehat{\mu}_d/\mu_d$ (true value = 1)

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$
T=6												
MLE	0.96	0.19	0.93	0.95	0.97	0.19	0.92	0.92	0.96	0.18	0.88	0.84
A-FV09	0.93	0.18	0.93	1.01	0.91	0.17	0.92	1.01	0.42	0.56	0.37	0.31
A-HN04	1.00	0.19	0.93	0.92	1.01	0.19	0.92	0.89	1.00	0.18	0.89	0.83
J-DJ15	1.09	0.25	0.81	0.72	1.11	0.26	0.80	0.67	1.05	0.27	0.75	0.60
J-HN04	1.04	0.22	0.89	0.83	1.05	0.21	0.88	0.79	1.05	0.20	0.83	0.70
CRE	0.99	0.19	0.95	1.00	0.99	0.18	0.95	1.00	0.99	0.17	0.95	1.00
LPM	1.28	0.19	0.68	0.99	1.28	0.18	0.67	1.00	1.29	0.17	0.62	1.00
T=12												
MLE	1.01	0.13	0.94	0.99	1.01	0.13	0.94	0.97	1.00	0.12	0.94	0.95
A-FV09	1.00	0.13	0.94	0.99	1.00	0.13	0.95	0.97	0.99	0.12	0.93	0.95
A-HN04	1.01	0.13	0.94	0.98	1.01	0.13	0.94	0.96	1.01	0.12	0.93	0.94
J-DJ15	1.03	0.15	0.91	0.88	1.03	0.15	0.88	0.82	1.03	0.16	0.87	0.78
J-HN04	1.01	0.13	0.94	0.96	1.02	0.13	0.93	0.94	1.01	0.12	0.92	0.91
CRE	1.01	0.13	0.95	1.00	1.01	0.13	0.94	0.98	1.00	0.13	0.94	0.98
LPM	1.33	0.13	0.29	1.00	1.33	0.13	0.27	0.99	1.33	0.13	0.27	0.97

The simulation results for models where dynamic completeness fails reveal many interesting implications for the uncorrected and corrected fixed effects probit estimators. To begin with, higher levels of serial dependence in the error terms and y_{it} exacerbate a practical difficulty in performing MLE while treating heterogeneity as parameters to be estimated. The problem relates to the fact that the partial effect is not well-defined for an individual j whose value of y_{jt} is constant. In this case, the dummy variable for observation j perfectly predicts the outcome, so the estimate of α_i is technically unbounded (Fernandez-Val, 2009). These observations are therefore dropped from the estimation sample. These individual's contributions to the sample APE are equal to zero. The true α's in these cases tend to be larger in magnitude, and while this means $m(\beta_0, \gamma_0, \alpha, z_{it})$ will be smaller by the properties of the standard normal PDF and CDF, it should still be strictly positive. This explains the tendency of MLE to under-predict APE (Alexander and Breunig, 2014). Additionally, there may be distributional differences between the subpopulation that has a changing response and the population in general that could cause additional bias.

The probability of observing an individual with a constant y_{it} increases significantly in the shorter panels as serial dependence in the errors increases. To illustrate, for $T = 6$ and $\rho = 0$, across the 1000 replications, 21% of the individuals were dropped on average, while for $T = 6$ and $\rho = 0.8$, 32% were dropped on average. For comparison, with $T = 12$, this dropping rate was only 7.5% for $\rho = 0$ and 14.5% for $\rho = 0.8$. Splitting the panel for Dhaene and Jochman's jackknife makes this much worse, especially when the panel is only six periods long to begin with. Practically speaking, losing more observations makes it more likely that the numerical maximization algorithm will not converge (at least when N is relatively small). The worst case of this occurring in this study was for the split-panel jackknife in the $T = 6$, $\rho = 0.8$ case, in which 32% of replications had a failure to converge. Similar rates of non-convergence occurred as well for (unreported) runs of the uncorrected MLE and analytical corrections with high ρ and only three or four time periods.

The results show that, as expected, the failure of dynamic completeness significantly increases the bias and decreases the precision of all of the fixed effects slope estimators. By design of the

data generating process, this bias is separate from the scaling that would occur from the latent model errors having non-unit variance as a result of their autoregressive structure. In the worst of cases, the means of the split-panel jackknife estimates of γ for the shorter panels even have the wrong sign when $\rho = 0.8$. For small panels, the standard errors of the corrected estimators also do a poor job estimating the true standard deviations. The increased bias is not surprising given that in the presence of unobserved heterogeneity, a conditional independence assumption for $\{y_{i1}, y_{i2}, \ldots, y_{iT}\}$ is required to identify unscaled slope parameters in the panel probit model (Wooldridge, 2010). Fernandez-Val's analytical correction and Hahn and Newey's jackknife continue to mitigate the bias and perform relatively well when $\rho = 0.4$. While they still provide an improvement over the uncorrected MLE when $\rho = 0.8$, they are still severely biased.

The performance of the fixed effects estimators in estimating APE is much more relevant when dynamic completeness fails. In the case of the short panel ($T = 6$), the effect of higher serial correlation in the errors on the performance of the fixed effects estimators is quite mixed. Comparisons between estimators in Tables 1.3 and 1.4 suggest that the analytical correction proposed by Hahn and Newey seem fairly robust to serial correlation, with biases in APE estimates of 6% or less. Bias in the Fernandez-Val correction only increases slightly at low-to-moderate levels of serial correlation, but the combination of high autocorrelation and short panel length causes a substantial downward bias of 40% to 60%. With the longer panels, however, the effect of ρ on the bias of this and the other corrections is much smaller, 2% or less for the $T = 12$ case.

The effect of ρ on the jackknife APE corrections is different for each explanatory variable. For instance, the bias in the split-panel jackknife APE estimates for x increase with higher ρ in the $T = 6$ case, but those for d appear to be less affected. Hahn and Newey's jackknife shows a very similar pattern, but with much smaller variance of the estimators. Furthermore, the results for the split-panel jackknife illustrate that slope and APE estimates do not necessarily agree in sign. This is another drawback to using this procedure on short panels, since splitting the panel increases variance substantially. Perhaps larger N would mitigate this problem.

1.4.1.1 Comparison with uncorrected MLE

As in the dynamically complete case, it is important to note that the uncorrected APE estimators often have lower bias than either the analytical or jackknife corrected estimators, especially for the short panels. For longer panels, the uncorrected MLE, analytical corrections, drop-one-period jackknife, and CRE behave very similarly, while the split-panel jackknife has higher variance. For comparison, the CRE and LPM are not really affected by either failure of dynamic completeness or the length of the panel. The structure of the data is such that one would expect CRE to do well. As a side note, I found that a generalized estimating equations approach with either an exchangeable or AR(1) covariance matrix was not much more efficient than pooled MLE for the CRE model. In contrast to the CRE, the best linear approximation performs fairly well for the continuous variable (bias of $5-7\%$) but does not perform very well for the discrete variable (bias of $28-34\%$).

1.4.2 Comparing bias correction and CRE under more general forms of heterogeneity

Table 1.5 compares probit APE estimates for the continuous variable x using the uncorrected MLE, Fernandez-Val correction, and two Correlated Random Effects estimators, described in Section 3. I consider panels with $T = 6$ and $T = 12$, in the case of serially independent errors. The estimators are compared across three different forms of heterogeneity which do not satisfy the conditional distribution assumptions for either CRE estimator. The uncorrected MLE and Fernandez-Val correction, in contrast, place no restriction on the nature of the heterogeneity.

Table 1.5: Probit Estimates of $\widehat{\mu}_x/\mu_x$ Under Different Heterogeneity (true value = 1)

	α_1				α_2				α_3			
	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$
T=6												
MLE	1.00	0.16	0.92	0.90	0.98	0.23	0.91	0.86	0.99	0.15	0.92	0.90
A-FV09	0.96	0.15	0.92	0.91	0.93	0.21	0.89	0.87	0.95	0.14	0.90	0.90
A-HN04	1.04	0.16	0.88	0.83	1.03	0.24	0.87	0.78	1.03	0.16	0.88	0.82
CRE	0.76	0.15	0.59	0.96	0.75	0.20	0.74	1.03	0.95	0.15	0.92	0.99
CRE2	0.78	0.15	0.65	0.95	0.76	0.20	0.75	1.01	0.96	0.15	0.93	0.98
T=12												
MLE	0.99	0.12	0.91	0.89	0.98	0.18	0.85	0.76	1.00	0.10	0.92	0.91
A-FV09	0.98	0.12	0.91	0.88	0.97	0.18	0.84	0.75	0.99	0.10	0.92	0.90
A-HN04	1.00	0.12	0.91	0.87	0.99	0.18	0.84	0.74	1.01	0.10	0.91	0.89
CRE	0.68	0.11	0.22	1.01	0.79	0.17	0.72	0.98	0.95	0.10	0.92	1.02
CRE2	0.70	0.11	0.26	1.00	0.80	0.17	0.73	0.97	0.96	0.10	0.93	1.01

Since the pooled-MLE version of CRE only identifies slope parameters up to scale, I only report on the APE. The tables show that the bias in the CRE estimators is higher in all three specifications. For instance, in the second specification ($\alpha = \alpha_2$), CRE underestimates the APE of x by about 25% when $T = 6$, while the biases in the uncorrected MLE and the Fernandez-Val correction are only 2% and 9%, respectively. The results for the APE of d were comparatively similar, though the CRE estimators tended to have a positive bias. This illustrates the importance of the functional form assumption when specifying a CRE model, and suggests an advantage in the FE approaches as they place no restrictions on the α_i.

1.4.3 Conditional logit and the importance of correcting APE estimates

Table 1.6 explores a possible approach to handling unobserved cross-sectional heterogeneity in logit models where the response variables are conditionally independent. Using conditional logit to consistently estimate slope parameters does not allow for estimating average partial effects unless the researcher can somehow recover estimates of the α_i. One way is to estimate them by MLE, restricting the slope parameters to their conditional logit estimates, but this causes bias in APE estimates. The table shows the APE estimates (for the continuous variable x) from the uncorrected pooled logit MLE, conditional logit without correcting the APE estimates (denoted CLOG), and conditional logit where the APE have been corrected using Fernandez-Val's formula (CLOGC). Simulations for the APE of d showed a very similar pattern.

Table 1.6: Corrected and Uncorrected Logit Estimates of $\widehat{\mu}_x/\mu_x$ (true value = 1)

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$	Mean	SD	cv:.95	$\frac{SE}{SD}$
T=6												
MLE	1.01	0.19	0.95	1.01	1.01	0.18	0.94	0.99	0.99	0.18	0.92	0.88
CLOGIT	0.87	0.16	0.93	1.14	0.87	0.16	0.91	1.10	0.87	0.15	0.82	0.93
CLOGIT-C	1.00	0.18	0.94	1.00	0.99	0.18	0.93	0.97	0.98	0.17	0.90	0.83
T=12												
MLE	1.00	0.12	0.95	1.01	1.00	0.12	0.95	1.01	1.00	0.12	0.94	0.95
CLOGIT	0.94	0.11	0.94	1.06	0.94	0.11	0.94	1.06	0.94	0.11	0.91	0.99
CLOGIT-C	1.00	0.12	0.95	1.00	1.00	0.12	0.95	1.00	0.99	0.12	0.93	0.94

The table illustrates a couple of interesting points. First, the uncorrected conditional logit APE estimates have biases that are 5-13 percentage points higher than the corrected versions. This shows that inconsistent estimation of the α_i is a significant problem even when a consistent procedure is used to estimate the slope coefficients. Moreover, these suggest that the "small bias" property in the uncorrected MLE APE estimates observed earlier is the result of two competing biases. In the case of this chapter's data generating process, an upward bias in the slope estimate is being offset by a scale factor that is biased toward zero. Using a procedure like conditional logit (or any bias correction) while failing to correct APE estimates removes only one source of the problem and may increase the bias compared to doing no correction at all.

1.4.4 Empirical example: Welfare participation

As an additional demonstration of the relative performance of these fixed effects estimators, I apply them to a dataset on participation in Aid to Families with Dependent Children (AFDC), a U.S. welfare program. The data are by way of Chay and Hyslop (2014), who use the 1990 Survey of Income and Program Participation (SIPP). The panel consists of AFDC participation, age, race, marital status, number of children, and poverty level for $1,934$ women who either received benefits or had income below a certain threshold at some point during the sample period. As welfare participation is a binary response that is thought to be highly persistent over time, Chay and Hyslop differentiate between unobserved heterogeneity, and structural state dependence as sources of persistence, finding significant evidence for the latter using dynamic estimators under varying assumptions about the nature of the heterogeneity and initial conditions (Chay and Hyslop, 2014). Although their findings suggest that a dynamic model may be more appropriate, these data still provide an interesting and relevant setting for evaluating the bias-corrected fixed effects estimators in the static case. Table 1.7 lists slope parameter estimates for two key determinants of participation, marital status and number of children. Note that in addition to several control variables, these specifications include time period dummies. While technically, they are also incidental parameters under large-T bias corrections, it is customary to include them in this type of analysis. In (unre-

Table 1.7: Welfare Participation: Slope Estimates

	Full Sample	Sample with changing participation					
	CRE	MLE	A-FV09	A-HN04	J-DJ15	J-HN04	CRE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Marriage	-0.986	-1.908	-1.579	-1.730	-1.822	-1.565	-1.462
	(0.011)	(0.208)	(0.178)	(0.189)	(0.229)	(0.176)	(0.022)
Kids	0.162	0.481	0.409	0.437	0.447	0.380	0.358
	(0.001)	(0.104)	(0.098)	(0.100)	(0.121)	(0.096)	(0.006)
	N=1934	N*=494					
	T=8	T=8					

Controls include education, poverty level, a quadratic in age, a race dummy, and time period dummies. Standard errors were clustered by individual

ported) simulations with true time effects, I found that the additional bias caused by their inclusion to be smaller and that it did not change the relative performance of the different FE estimators. Table 1.8 lists estimated APE. Unlike the simulations, these tables include CRE and LPM estimates over the estimation subsample of the fixed effects estimators. This application highlights the problems that may arise when many individuals have responses that do not change. In this case, only 494, or roughly 25% of women in the sample had participation that changed over the 32 months of the survey. In the worst simulation case ($T = 6, \rho = 0.8$) 68% of the sample still had responses that changed. Practically speaking, not only does this increase variance of the estimators, but it potentially exacerbates any bias stemming from sample selection (which did not appear to be much of a problem in the simulations).

The bias-corrected slope estimates in both cases are smaller in magnitude than the uncorrected MLE, and are similar in magnitude to CRE estimates over the subsample of changing responses, though quite different from the CRE estimates over the whole sample. Probit slope estimates from the 1998 and 2014 versions of Chay and Hyslop range from -0.934 to -0.658 for the marriage variable, and 0.11 to 0.152 for the kids variable. Both are much smaller in magnitude than the non-linear fixed effects estimates suggesting that persistence, state dependence and/or sample selection are playing a significant role.

Table 1.8: Welfare Participation: Average Partial Estimates

	Full Sample		Sample with changing participation						
	CRE	LPM	MLE	A-FV09	A-HN04	J-DJ15	J-HN04	CRE*	LPM*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Marriage	-0.260	-0.271	-0.112	-0.110	-0.115	-0.162	-0.129	-0.112	-0.126
	(0.001)	(0.001)	(0.007)	(0.008)	(0.007)	(0.008)	(0.008)	(0.000)	(0.000)
Kids	0.047	0.052	0.034	0.033	0.035	0.049	0.034	0.034	0.033
	(0.000)	(0.000)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.000)	(0.000)
	N=1934		N*=494						
	T=8		T=8						

*Sum of partial effects divided by full sample size for comparison with FE estimators

The 1998 version of Chay and Hyslop contains several estimates of LPMs, including the static model estimated with fixed effects (column 2 of Table 2), which are compared to the bias-correct APE estimates in Table 1.8. The Chay and Hyslop estimates (that account for heterogeneity) range from -0.271 to -0.143 for marriage and from 0.029 to 0.068 for kids. The bias corrected estimates range from -0.162 to -0.110 for marriage and from 0.033 to 0.050 for kids, which seem more in line than the slope estimates, echoing previous research and the simulation evidence in this chapter for the "small bias" property.

1.5 Conclusion

The simulation evidence in this chapter suggests that these bias corrections continue to estimate APE fairly well when the level of serial correlation is low to moderate, but strong serial correlation may cause bias when the panel is short. As such, dynamic completeness may be a substantive requirement unless the researcher has access to many time periods of data. Estimation in shorter panels may also present sample selection or computational challenges. While it may seem unfair to evaluate a technique based on large-T asymptotic approximations using panels with only six time periods, others have suggested these techniques have desirable properties in large-N, small-T settings. Moreover, the results of this chapter suggest that if a researcher is primarily concerned with estimating APE in a static model, then the included bias correction techniques may offer little benefit relative to the uncorrected MLE while adding the cost of a more complicated estimation procedure. It should be noted, however, that the "small bias" property of APE does not hold in dynamic models, where correction techniques have been found to decrease bias substantially. Additionally, I find that the fixed effects approach (with or without a bias correction) may offer advantages over CRE when the heterogeneity does not satisfy the CRE assumption. I also find evidence that highlights the importance of correcting for inconsistent estimation of the heterogeneity terms when a consistent procedure is used to estimate the slopes.

There are many important avenues for future research. First and foremost, an interesting ques-

tion is how well the analytical bias correction of Hahn and Kuersteiner (2011) performs in this setting. It accommodates serial correlation in theory, but requires "moderately large T." Furthermore, in practical applications like a policy or program analysis, it is important to control for time effects, which I did not include in this set of simulations. The reason is that under the large-T asymptotics that justify these corrections, time dummies are also incidental parameters. I did run a set of simulations over the same values of ρ and T where time effects were estimated, but not part of the true data generating process for y_{it}. I found that the same relative patterns held across estimators as in this chapter, but the additional incidental parameters caused slightly higher bias in slope parameters and virtually no increase in bias for APE except for the short panels, where bias increased slightly. Fernandez-Val and Weidner (2016) allow for both time and cross-sectional heterogeneity in analytical and jackknife corrections. However, the results depend on N/T being constant in the limit. Therefore, unlike the wide and short panels included in this chapter, their application is intended for settings where N and T are of similar magnitude.

CHAPTER 2

EXPONENTIAL PANEL MODELS WITH COEFFICIENT HETEROGENEITY

2.1 Introduction

The fixed effects Poisson (FEP) estimator, also known as multinomial QCMLE, is an attractive choice for modeling nonnegative responses whose conditional means contain an unobserved individual effect that may be correlated with the explanatory variables. Unlike other conditional-ML estimators, notably the FE logit, FEP does not require assuming a full distribution or conditional independence (Wooldridge, 1999). This chapter considers the exponential conditional mean, which is logically consistent for nonnegative dependent variables and has the feature that coefficients on the regressors can be interpreted as semi-elasticities.

The focus of this chapter is an extension to the unobserved effects exponential model that allows for additional heterogeneity in the form of random coefficients. While there is some literature considering Poisson variables in this setting, less insight exists into how to proceed for other nonnegative or non-count variables, or even what the consequences are of ignoring the heterogeneity. In the linear unobserved effects model with strictly exogenous regressors and random coefficients, for instance, it is straightforward to show that fixed effects OLS is consistent for the means of the coefficients so long as they are mean-independent of the time-demeaned regressors. This is not necessarily true for nonlinear models, as this chapter shows for the exponential case. Moreover, it is unknown whether other quantities of interest, like average partial effects (APE), can be consistently estimated while ignoring coefficient heterogeneity. Furthermore, much of the literature assumes all sources of heterogeneity are independent of covariates, which can cause inconsistent estimation of coefficient means as well as type II errors in tests for random coefficients

These potential complications motivate testing for neglected heterogeneity. An LM test in the style of Chesher (1984), however, is likely to reject when the Poisson distribution is misspecified or when conditional independence fails. Therefore, I extend this methodology specifically to the

FEP setting, deriving a simple variable addition test that is more broadly applicable. Furthermore, I propose a method for parametrically identifying the means of random coefficients that leads to estimators that are computationally simple related to existing approaches to random coefficients in this model. One novel contribution of this chapter is to treat random coefficients and the traditional multiplicative effect[1] separately, as the latter can be handled without restricting their dependence on explanatory variables. I also provide estimators of average partial effects. In an application to the patent R&D relationship among U.S. manufacturing firms, I find evidence of heterogeneous elasticities and lagged effects, though the results are not robust to changes in the estimation sample.

The rest of this chapter is organized as follows: Section 2 gives an overview of the existing literature, Section 3 reviews the FEP model and the classical test for the Fixed Effects Poisson case, before proposing this chapter's theoretical contributions. Section 4 contains a Monte Carlo experiment for the methods proposed, while Section 5 describes the empirical application. Section 6 consists of a brief conclusion and direction for future research.

2.2 Literature Review

Applying Andersen's (1970) conditional ML methodology, Hausman, Hall, and Griliches (1984) developed the FEP estimator for count data that allows arbitrary dependence between the unobserved effect and the regressors. They implemented their techniques to analyze the patent-R&D relationship in the U.S. manufacturing industry. Wooldridge (1999), showed that correct specification of the conditional mean and strict exogeneity of the regressors (conditional on the unobserved effect) were sufficient for consistency of FEP, broadening its application as a quasi-CMLE. Cameron and Trivedi (2013) considered the panel unobserved effects Poisson model with random coefficients in a "random effects" setting where all heterogeneity were assumed to be normally distributed and independent of the regressors. They concluded that "unlike for the linear model,

[1]The multiplicative effect can also be expressed as a random intercept inside the exponential conditional mean function.

the conditional mean for the random slopes model differs from that for the pooled and random effects models, making model comparison and interpretation more difficult."

Lagrange multiplier (LM) statistics are attractive in testing for coefficient heterogeneity because they use parameter estimates from a restricted model which can be simpler to estimate. In this case, the restricted model is FEP, for which built-in procedures exist in Stata and other programs. Moreover, LM tests are valid for null values on the boundary of the parameter space, unlike Wald tests, which is important because parameters (i.e. variances) associated with random coefficients should be nonnegative (Wooldridge, 2010). Random coefficients are an example of neglected heterogeneity that Chesher (1984) derived a test for in the ML setting. Chesher, as well as Lee and Chesher (1986), developed methodology for deriving test statistics in this and other settings where scores are identically zero under the parameter restriction. Greene and MacKenzie (2015) applied this methodology to random effects probit MLE. Hahn, Newey, and Smith (2014) extend Chesher's to moment condition estimators like Generalized Method of Moments (GMM). Hahn, Moon, and Snider (2015) allow for dependence between the heterogeneity and covariates when testing the likelihood setting, though they also find that tests that treat the heterogeneity and regressors as mean and second-moment independent still have power under alternatives where this is not true. A common feature of tests for neglected heterogeneity in the likelihood setting is that they have the interpretation of being either for information matrix (IM) equality or for overdispersion, making them less attractive for settings where researchers do not want to fully specify a distribution. I derive a test for slope heterogeneity in exponential models that does not have this drawback.

A Poisson-normal mixture model like the one described by Cameron and Trivedi is one of the "Generalized linear latent and mixed models" studied by Rabe-Hesketh and Skrondal (2004). The likelihood function consists of a multi-dimensional integral that must be numerically approximated, limiting its application to models where only a small number of coefficients are believed to be random. The authors used adaptive Gaussian quadrature to estimate a model of seizure counts for 236 subjects of (randomly assigned) epilepsy treatment trial, where both the intercept and the

coefficient on a variable for time of visit were allowed to be vary by individual. While a random effects approach makes sense for the experimental setting, treating the heterogeneity as independent of covariates can cause inconsistent estimation in many economic applications.

Wang, Cockburn, and Puterman (1998), do allow dependence between the heterogeneity and explanatory variables in the panel Poisson setting, assuming a parametric form for the dependence as well as a particular distribution for the heterogeneity. With the patent-R&D relationship in mind, they propose a mixed-Poisson regression approach which assumes that the coefficients follow a discrete distribution with finite support, modeling the probability mass at each point as multinomial logit. Their method involves using economic intuition or selection criteria to select the number of support points. Moreover, they suggest using a continuous model for the coefficients if model selection criteria indicate four or more points of support. My paper complements their work by proposing such a model. One benefit of my approach is that as in FEP, cases I can allow an unrestricted relationship between the explanatory variables and the multiplicative effect, as well as analyze non-counts.

2.3 Theory

2.3.1 The fixed effects Poisson model with coefficient heterogeneity

The standard fixed effects Poisson model with an exponential mean function assumes:

$$E(y_{it}|\boldsymbol{x}_i, c_i) = E(y_{it}|\boldsymbol{x}_{it}, c_i) = c_i \exp(\boldsymbol{x}_{it}\boldsymbol{\beta}_0) \tag{2.1}$$

for $i = 1,\ldots,N; t = 1,\ldots,T$. In this expression, $\boldsymbol{x}_{it}$ is a $1 \times K$ vector of time-varying explanatory variables, c_i is unobserved heterogeneity, and $\boldsymbol{\beta}_0$ is a $K \times 1$ unknown vector of coefficients.[2] Equation (2.1) implicitly assumes that $\boldsymbol{x}_{it}$ is strictly exogenous. Hausman, Hall, and Griliches (1984) showed that if conditional on $\boldsymbol{x}_i = \{\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{iT}\}$ and c_i, the y_{it} are independently distributed

[2] Wooldridge (1999) considered conditional mean functions of the form $c_i m(\boldsymbol{x}_i, \boldsymbol{\beta}_0)$ of which $m(\boldsymbol{x}_i, \boldsymbol{\beta}_0) = \exp(\boldsymbol{x}_{it}\boldsymbol{\beta}_0)$ is a special case.

as Poisson with mean given by (2.1), then conditioning on $n_i \equiv \sum_{t=1}^{T} y_{it}$ results in the multinomial distribution for $\{y_{i1}, \ldots, y_{iT}\}$.

The multinomial log-likelihood is

$$\ell_i^M(\boldsymbol{\beta}) = \sum_{t=1}^{T} y_{it} \log\left[p_t(\boldsymbol{x}_i, \boldsymbol{\beta})\right], \tag{2.2}$$

where

$$p_t(\boldsymbol{x}_i, \boldsymbol{\beta}) \equiv \frac{\exp(\boldsymbol{x}_{it}\boldsymbol{\beta})}{\sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta})}. \tag{2.3}$$

The feature that c_i enters conditional mean function multiplicatively means it cancels out of $p_t(\boldsymbol{x}_i, \boldsymbol{\beta})$ and therefore $\ell_i(\boldsymbol{\beta})$, meaning dependence between c_i and $\boldsymbol{x}_i$ may remain unrestricted. This structure also has the consequence that coefficients on time-constant regressors are not identified because these terms also cancel. This model is particularly attractive because as shown by Wooldridge (1999), $\boldsymbol{\beta}_0$ maximizes the expected value of 2.2 as long as (2.1) is true. Therefore, under additional regularity conditions, FEP consistently estimates $\boldsymbol{\beta}_0$ with N growing and T fixed. Notably, consistency does not require a distribution assumption for the responses and allows them to be arbitrarily serially correlated (Wooldridge, 1999).

Condition (2.1) generally fails, however, if the coefficients in the conditional mean function vary by individual i, as in the following:

$$E(y_{it}|\boldsymbol{x}_i, c_i, \boldsymbol{b}_i) = E(y_{it}|\boldsymbol{x}_{it}, c_i, \boldsymbol{b}_i) = c_i \exp(\boldsymbol{x}_{it}\boldsymbol{b}_i), \tag{2.4}$$

where now $\boldsymbol{b}_i$ is a $K \times 1$ vector of unobserved random variables such that $E(\boldsymbol{b}_i) = \boldsymbol{\beta}_0$. Defining $\boldsymbol{d}_i \equiv \boldsymbol{b}_i - \boldsymbol{\beta}_0$, the conditional mean in (2.4) is equivalent to $c_i \exp(\boldsymbol{x}_{it}\boldsymbol{\beta}_0 + \boldsymbol{x}_{it}\boldsymbol{d}_i)$, meaning one interpretation of the heterogeneity is unobserved interactions in the index of the mean function. There is a more practical, economic interpretation as well. Assuming element j is not functionally related to any other elements of $\boldsymbol{x}_{it}$, then

$$\frac{\partial \log\left[E(y_{it}|\boldsymbol{x}_i, c_i, \boldsymbol{b}_i)\right]}{\partial x_{itj}} = b_{ij}, \tag{2.5}$$

so model (2.4) implies semi-elasticities of the conditional mean of y_{it} that vary by individual. If x_{itj} is the log of another variable, as in some applications, then the b_{ij} are individually-varying elasticities.

An immediate consequence is that the heterogeneity likely causes specification error if we want to use FEP assuming (2.1). To see this, suppose for concreteness that $\boldsymbol{d}_i$ is continuous, and write its PDF conditional on $\boldsymbol{x}_i$ and c_i as $f(\cdot;\boldsymbol{\psi}_0)$, where $\boldsymbol{\psi}_0$ is an unknown parameter that is nonzero only if the coefficients are random. It follows under (2.4) and the Law of Iterated Expectations (LIE) that

$$E(y_{it}|\boldsymbol{x}_i,c_i) = c_i \exp\left[\boldsymbol{x}_{it}\boldsymbol{\beta}_0 + g_t(\boldsymbol{x}_i,\boldsymbol{x}_{it},c_i;\boldsymbol{\psi}_0)\right], \tag{2.6}$$

where

$$g_t(\boldsymbol{x}_i,\boldsymbol{x}_{it},c_i;\boldsymbol{\psi}_0) = \log\left\{E\left[\exp(\boldsymbol{x}_{it}\boldsymbol{d}_i)|\boldsymbol{x}_i,c_i\right]\right\} = \log\left\{\iint_{\mathbb{R}^K} \exp(\boldsymbol{x}_{it}\boldsymbol{d}_i) f(\boldsymbol{d}_i|\boldsymbol{x}_i,c_i)\,\mathrm{d}\boldsymbol{d}_i\right\}, \tag{2.7}$$

assuming the expectation exists. The exponential function now contains an unknown term that is generally nonzero and varies over time.[3] Depending on what we are willing to assume about the dependence between $\boldsymbol{b}_i$ and $\boldsymbol{x}_i$, we may not be able to distinguish between coefficients that are random and a more flexible functional form. The consequence of ignoring the coefficient heterogeneity is that now (2.1) is not correct, and so FEP of y_{it} on $\boldsymbol{x}_{it}$ can no longer be shown to be generally consistent for $\boldsymbol{\beta}_0$. This is true even under ideal conditions like independence between $\boldsymbol{b}_i$ and $\{\boldsymbol{x}_i,c_i\}$ In fact, simulation evidence from Section 4 suggests that substantial bias and inconsistency for FEP in this case. This is to contrast with the linear unobserved effects model with random coefficients, in which Fixed Effects OLS is consistent for the means of the coefficients so long as the coefficients are mean independent of the time-demeaned regressors (Wooldridge, 2010). In this case, the random coefficients cause a certain form of system heteroskedasticity in the idiosyncratic errors that is handled completely with robust inference.

[3] If $g_t(\boldsymbol{x}_i,c_i;\boldsymbol{\psi}_0)$ were time-constant, then it would also cancel from $p_t(\boldsymbol{x}_i,\boldsymbol{\beta},\boldsymbol{\psi})$ and FEP would be consistent, but there is no reason to think this should be the case with time-varying $\boldsymbol{x}_{it}$.

2.3.2 Testing under full distributional assumptions

If the y_{it} are count data and researchers are willing to take full distributional assumptions seriously, the approach of Chesher (1984) provides a simple LM test. The slopes are not allowed to depend on the covariates or c_i under the alternative, which avoids having to specify a particular joint distribution for $\boldsymbol{b}_i$ and $\boldsymbol{x}_i$. However, lack of power may be an issue in alternatives where $\boldsymbol{b}_i$ depends on $\boldsymbol{x}_i$. Findings of Hahn, Moon and Snider (2015), however, suggest that this is less of a concern in nonlinear models. The following statements formalize the assumptions:

$$y_{it}|(\boldsymbol{x}_i, c_i, \boldsymbol{b}_i) \sim Poisson\left[c_i \exp(\boldsymbol{x}_{it}\boldsymbol{b}_i)\right], i = 1, \ldots, N; t = 1, \ldots, T, \tag{2.8}$$

$$\{y_{i1}, \ldots, y_{iT}\} \text{ are independent conditional on } \{\boldsymbol{x}_i, c_i, \boldsymbol{b}_i\} \tag{2.9}$$

$$\boldsymbol{b}_i = \boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_0 \boldsymbol{u}_i, \text{where } \boldsymbol{u}_i|(\boldsymbol{x}_i, c_i) \sim F(\boldsymbol{0}, \boldsymbol{I}_K), \tag{2.10}$$

where $\boldsymbol{I}_K$ is the $K \times K$ identity matrix.

From Chesher (1984), assumption (2.10) does not assume a particular distribution for $\boldsymbol{b}_i$, but specifies that they follow a "location-scale generalization of the class of spherical distributions" described by Kelker (1970). Denote the PDF of $\boldsymbol{u}_i$ as $f()$.

It follows that

$$\boldsymbol{y}_i|(n_i, \boldsymbol{x}_i, c_i, \boldsymbol{b}_i) \sim Multinomial(n_i, p_1(\boldsymbol{x}_i, \boldsymbol{b}_i), \ldots, p_T(\boldsymbol{x}_i, \boldsymbol{b}_i)), \tag{2.11}$$

where

$$p_t(\boldsymbol{x}_i, \boldsymbol{b}_i) \equiv \frac{\exp(x_{it}\boldsymbol{b}_i)}{\sum_{r=1}^{T} \exp(x_{ir}\boldsymbol{b}_i)}. \tag{2.12}$$

Therefore, the log-likelihood for an observation i, integrating out the random part of the slopes, is

$$\ell_i(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \log\left\{ \iint_{\mathbb{R}^K} \frac{n_i!}{\prod_{t=1}^{T} y_{it}!} \prod_{t=1}^{T} \left[p_t(\boldsymbol{x}_i, \boldsymbol{b}_i)^{y_{it}}\right] f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i \right\}, \tag{2.13}$$

where the integral is of K dimensions.

An LM test of $H_0 : \mathbf{\Lambda}_0 = \mathbf{0}$ is attractive because in this case, $\boldsymbol{b}_i = \boldsymbol{\beta}_0$, and so the restricted model can be estimated using FEP. It also turns out that the restricted score does not depend on the unknown PDF $f()$.

However, the parameterization of this model causes a complication in deriving the restricted scores, as described by Chesher (1984) and Lee and Chesher (1986) for a more general class of models. It turns out the score of the unrestricted model evaluated at the parameter restriction is identically zero.[4] Chesher (1984) proposed re-parameterizing the scale assumption and restricting the correlation among the heterogeneity allowed under the alternative.[5]

$$\mathbf{\Lambda}_0 = diag\left\{\sqrt{\lambda_{1,0}}, \ldots, \sqrt{\lambda_{K,0}}\right\} \tag{2.14}$$

Allowing no covariance between coefficients may affect power under alternatives in which this does not hold, but at the same time, information about the covariances is only relevant if there is evidence that the variances are nonzero.[6] Under (2.14), the restricted score has the $0/0$ form, but the limits follow from L'Hopital's rule. The algebraic details are collected in Appendix C. Collecting the λ_j in the $K \times 1$ vector $\boldsymbol{\lambda}$, the restricted score is:

$$\boldsymbol{s}_i(\boldsymbol{\beta}, \mathbf{0}) \equiv \lim_{\boldsymbol{\lambda} \downarrow 0} \left\{\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\beta}, \boldsymbol{\lambda})'\right\} = \sum_{i=1}^{N} \begin{Bmatrix} \sum_{t=1}^{T} y_{it} \left[\nabla_{\beta} p_t(\boldsymbol{x}_i, \boldsymbol{\beta})' / p_t(\boldsymbol{x}_i, \boldsymbol{\beta})\right] \\ \frac{1}{2} a_1(\boldsymbol{x}_i, \boldsymbol{\beta}) \\ \vdots \\ \frac{1}{2} a_K(\boldsymbol{x}_i, \boldsymbol{\beta}) \end{Bmatrix}, \tag{2.15}$$

where $a_j(\boldsymbol{x}_i, \boldsymbol{\beta})$ is the (j, j)th element of

$$\begin{aligned} &\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}) \\ &\equiv \sum_{t=1}^{T} \nabla^2_{\boldsymbol{\beta}} \ell^M_{it}(\boldsymbol{\beta}) + \left(\sum_{t=1}^{T} \nabla_{\boldsymbol{\beta}} \ell^M_{it}(\boldsymbol{\beta})\right)' \left(\sum_{t=1}^{T} \nabla_{\boldsymbol{\beta}} \ell^M_{it}(\boldsymbol{\beta})\right). \end{aligned} \tag{2.16}$$

[4] See Appendix C for the derivation.

[5] Chesher's solution would be to assume $\mathbf{\Lambda}_0 = \sqrt{\lambda_0} I_K$

[6] The relevant alternative, strictly speaking, should be that at least one $\lambda_{j,0} \geq 0$, but for simplicity, the two-sided alternative is treated here, as in Chesher (1984).

In this last expression, ℓ_{it}^{M} is the multinomial log-likelihood for observation i in period t.

The outer product of the score version of the LM statistic is then N times the uncentered R-squared from the regression of 1 on $\widetilde{\boldsymbol{s}}_i'$, where for each observation i, $\widetilde{\boldsymbol{s}}_i$ is the appropriate summand in right hand side of (2.15) evaluated at $\widetilde{\boldsymbol{\beta}}_{FEP}$. The advantage to this approach is its relative simplicity. The unrestricted model may be even computationally infeasible to estimate, but a test of the null hypothesis of constant coefficients is relatively easy to implement.

The downside of this approach concerns robustness to failure of (2.8) or (2.9). Chesher (1984) notes that statistics derived using this approach resemble White's (1982) information matrix test for general model misspecification, as $E\left[\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})\right] = \boldsymbol{0}$ if the conditional multinomial distribution is correct. This means coefficient heterogeneity cannot be distinguished from failures of the model's other assumptions, such as the Poisson distribution or conditional independence.

2.3.3 Testing under weaker assumptions

In the previous section, I showed the classical test applicable to conditionally independent Poisson dependent variables. While the statistic is simple to calculate, the test is likely to reject in cases where the Poisson or conditional independence assumption fail regardless of the presence of random coefficients. This is similar to the case of a linear model where the presence random slopes (that are assumed to be independent of covariates) is indistinguishable from a certain form of system heteroskedasticity. In this section, I extend Chesher's approach to testing for neglected heterogeneity to the FEP setting where only the conditional mean of $\boldsymbol{y}_{it}$ is assumed to be correctly specified. I show that an LM test of exclusion restrictions on squared regressors is valid when the coefficients are allowed to belong to a location-scale family under the alternative.

As before, assume:

$$E(y_{it}|\boldsymbol{x}_i, c_i, \boldsymbol{b}_i) = E(y_{it}|\boldsymbol{x}_{it}, c_i, \boldsymbol{b}_i) = c_i \exp(\boldsymbol{x}_{it}\boldsymbol{b}_i) \tag{2.17}$$

and

$$\boldsymbol{b}_i = \boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_0 \boldsymbol{u}_i, \text{where } \boldsymbol{u}_i|(\boldsymbol{x}_i, c_i) \sim F(\boldsymbol{0}, \boldsymbol{I}_K), \tag{2.18}$$

where again the CDF $F()$ and the corresponding PDF $f()$ are left unspecified.

Similar to before, these conditions imply:

$$E(y_{it}|\boldsymbol{x}_i, c_i) = c_i \exp\left[\boldsymbol{x}_{it}\boldsymbol{\beta}_0 + m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda}_0)\right], \tag{2.19}$$

where

$$m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda}_0) = \log\left\{E\left[\exp(\boldsymbol{x}_{it}\boldsymbol{\Lambda}_0\boldsymbol{u}_i)|\boldsymbol{x}_i, c_i\right]\right\} = \log\left\{\iint_{\mathbb{R}^K} \exp(\boldsymbol{x}_{it}\boldsymbol{\Lambda}_0\boldsymbol{u}_i) f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i\right\}. \tag{2.20}$$

It is easy to see that $m_t(\boldsymbol{x}_i, \mathbf{0}) = 0$. In the multivariate normal case, $m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda}_0) = \frac{1}{2}\boldsymbol{x}_{it}\boldsymbol{\Omega}_0\boldsymbol{x}_{it}'$, where $\boldsymbol{\Omega}_0 = \boldsymbol{\Lambda}_0\boldsymbol{\Lambda}_0'$. Rejecting $H_0 : \boldsymbol{\Lambda}_0 = 0$ provides evidence against the null of constant coefficients.

I follow Chesher's derivation of the LM statistic as before, but unlike other methods, I only integrate $\boldsymbol{u}_i$ out of the conditional mean function, not the entire likelihood or score. The unrestricted quasi log-likelihood is

$$\ell_i(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \sum_{t=1}^{T} y_{it} \log\left[p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})\right], \tag{2.21}$$

where

$$p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \equiv \frac{\exp\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda})\right)}{\sum_{r=1}^{T} \exp\left(\boldsymbol{x}_{ir}\boldsymbol{\beta} + m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda})\right)}. \tag{2.22}$$

The first K elements of the unrestricted score evaluated at $\boldsymbol{\Lambda} = \mathbf{0}$ are just the usual FEP scores. The gradient with respect to $\boldsymbol{\Lambda}$ evaluated at $\boldsymbol{\Lambda} = \mathbf{0}$, however, presents a similar problem as before. I make the same re-parameterization as before, shown in equation (2.14), restricting the coefficients to be uncorrelated with each other under the alternative. The restricted scores have a $0/0$ form and are evaluated using L'Hopital's Rule. The details are collected in Appendix C.

The score evaluated at the parameter restriction is:

$$\boldsymbol{s}_i(\boldsymbol{\beta}, \mathbf{0}) = \begin{Bmatrix} \sum_{t=1}^{T} y_{it} \left[\nabla_{\boldsymbol{\beta}} p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \mathbf{0})' / p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \mathbf{0})\right] \\ \frac{1}{2}\sum_{t=1}^{T} y_{it} \left[\sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta}) \left(x_{it1}^2 - x_{ir1}^2\right)\right] / \sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta}) \\ \vdots \\ \frac{1}{2}\sum_{t=1}^{T} y_{it} \left[\sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta}) \left(x_{itK}^2 - x_{irK}^2\right)\right] / \sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta}) \end{Bmatrix}. \tag{2.23}$$

The last K elements are proportional to the restricted FEP scores for testing the exclusion of squared regressors from the model with constant slopes. Therefore, in the exponential case, we cannot distinguish random coefficients from the presence of quadratics in $E(y_{it}|\boldsymbol{x}_{it}, c_i)$. As an empirical matter, however, this test takes no stand on the (conditional) distribution, overdispersion, or serial correlation of y_{it}, so it may offer some advantages to the approach in Section 3.2. For example, if a researcher rejects the null using the test based on (2.15), but fails to reject based on (2.23), then he or she can proceed in estimating the model based on (2.1) with some peace of mind.

2.3.4 A correlated random coefficients approach to testing and estimation

When one wishes to allow more than one or two slopes to be random, "random effects" type estimation based on integrating out the heterogeneity is computationally difficult and may not be robust to misspecification of the response variable's distribution. A straightforward alternative, which is applicable not only to counts but also to other nonnegative responses, is to make a parametric, distributional assumption for $\boldsymbol{b}_i$ that allows us to derive $E\left[\exp(\boldsymbol{x}_{it}\boldsymbol{d}_i)|\boldsymbol{x}_i, c_i\right]$. Here, I assume correlated random coefficients (CRC) and (conditional) multivariate normality:

$$\begin{aligned} \boldsymbol{b}_i &= \boldsymbol{\alpha}_0 + \boldsymbol{\Gamma}_0 \bar{\boldsymbol{x}}_{\boldsymbol{i}}' + \boldsymbol{d}_i, \\ \boldsymbol{d}_i|(\boldsymbol{x}_i, c_i) &\sim Normal(\boldsymbol{0}, \boldsymbol{\Omega}_0), \end{aligned} \tag{2.24}$$

where $\bar{\boldsymbol{x}}_{\boldsymbol{i}} = \sum_{t=1}^{T} \boldsymbol{x}_{it}$, $\boldsymbol{\alpha}_0$ is an unknown $K \times 1$ vector, and $\boldsymbol{\Gamma}_0$ is an unknown $K \times K$ matrix. This assumption states that the dependence between $\boldsymbol{x}_i$ and the mean of $\boldsymbol{b}_i$ is captured entirely through the time averages of $\boldsymbol{x}_{it}$, and is the application of Mundlak (1978) to the current setup. Alternatively, one could allow the mean of $\boldsymbol{b}_i$ to depend on $\boldsymbol{x}_i$ in the style of Chamberlain (1980). If $\boldsymbol{\Gamma}_0 = \boldsymbol{0}$, then (2.24) amounts to a stronger version of (2.10) where then $\boldsymbol{\alpha}_0 = \boldsymbol{\beta}_0$. Note that (2.24) only requires multivariate normality of the coefficients conditional on $\boldsymbol{x}_i$; their unconditional distribution may not be normal, though logically speaking it should be continuous and have unbounded support. Condition (2.24) also implies $\boldsymbol{b}_i$ and c_i are independent, conditional on $\boldsymbol{x}_i$. This is less restrictive for testing purposes because $\boldsymbol{b}_i$ is constant under the null, but it could affect power under alterna-

tives where the two are dependent. The two sources of heterogeneity are still allowed, through $\boldsymbol{x}_i$, to be correlated unconditionally. As in FEP, the relationship between $\boldsymbol{x}_i$ and c_i is left completely unrestricted.

Under (2.4) and (2.24), it follows from properties of the lognormal distribution and the LIE that

$$\begin{aligned}
E(y_{it}|\boldsymbol{x}_i, c_i) =& E(y_{it}|\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, c_i) \\
=& c_i \exp\left(\boldsymbol{x}_{it}\boldsymbol{\alpha}_0 + \boldsymbol{x}_{it}\boldsymbol{\Gamma}_0\bar{\boldsymbol{x}}_i' + \frac{1}{2}\boldsymbol{x}_{it}\boldsymbol{\Omega}_0\boldsymbol{x}_{it}'\right) \\
=& c_i \exp\left(\boldsymbol{x}_{it}\boldsymbol{\alpha}_0 + (\bar{\boldsymbol{x}}_i \otimes \boldsymbol{x}_{it})\text{vec}(\boldsymbol{\Gamma}_0) + \frac{1}{2}\left(\sum_{j=1}^{K}\omega_j x_{itj}^2 + 2\sum_{j=1}^{K-1}\sum_{h\neq j}^{K}\rho_{jh}x_{itj}x_{ith}\right)\right) \\
\equiv& c_i \exp\left(\boldsymbol{x}_{it}\boldsymbol{\alpha}_0 + (\bar{\boldsymbol{x}}_i \otimes \boldsymbol{x}_{it})\boldsymbol{\gamma}_0 + \frac{1}{2}\check{\boldsymbol{x}}_{it}\boldsymbol{\omega}_0\right), \qquad (2.25)
\end{aligned}$$

where $\boldsymbol{\gamma}_0 = vec(\boldsymbol{\Gamma}_0)$, $\check{\boldsymbol{x}}_{it} = (x_{it1}^2, \ldots x_{itK}^2, x_{it1}x_{it2}, x_{it1}x_{it3}\ldots x_{it,K-1}x_{itK})$,
$\boldsymbol{\omega}_0 \equiv (\omega_1, \ldots \omega_K, 2\rho_{12}, 2\rho_{13}\ldots, 2\rho_{K-1,K})'$, $\omega_j = Var(b_j)$, and $\rho_{jh} = Cov(b_j, b_h)$.

Equation (2.25), along with regularity conditions, implies that FEP of y_{it} on $\boldsymbol{x}_{it}$, interactions between $\boldsymbol{x}_{it}$ and $\bar{\boldsymbol{x}}_i$, and squares and interactions of $\boldsymbol{x}_{it}$ will consistently estimate $\boldsymbol{\alpha}_0$, $\boldsymbol{\gamma}_0$, and $\boldsymbol{\omega}_0$ without assuming a distribution for y_{it} and while allowing arbitrary serial correlation (Wooldridge, 1999).

Following estimation of (2.25), the unconditional means of the $\boldsymbol{b}_i$ are easy to estimate using the following, where $\boldsymbol{\mu}_{\bar{\boldsymbol{x}}} = E(\bar{\boldsymbol{x}}_i)$:

$$\boldsymbol{\beta}_0 \equiv E(\boldsymbol{b}_i) = \boldsymbol{\alpha}_0 + \boldsymbol{\Gamma}_0\boldsymbol{\mu}_{\bar{\boldsymbol{x}}}', \qquad (2.26)$$

I believe that using the lognormal distribution in the FEP setting is novel and that it offers the crucial advantage of still allowing one source of heterogeneity to be correlated with $\boldsymbol{x}_i$.[7] This procedure is easy to implement, as the FEP estimator is available in software packages like Stata,

[7] A similar result appeared in Cameron and Trivedi (2013) for the case where $\boldsymbol{b}_i|\boldsymbol{x}_i, c_i \sim Normal(\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0)$ and $c_i|(\boldsymbol{x}_i, \boldsymbol{b}_i) \sim lognormal(0, \sigma_c^2)$ as a way of illustrating how random coefficients change $E(y_{it}|\boldsymbol{x}_i)$.

though practitioners should be careful to calculate cluster-robust standard errors to account for serial correlation and misspecification of the multinomial distribution. Another important note is if one believes that time constant variables $\boldsymbol{z}_i$ belong in the model and they also have random coefficients that are correlated with the coefficients on the $\boldsymbol{x}_{it}$, then the augmented FEP regression should also include interactions between $\boldsymbol{z}_i$ and $\boldsymbol{x}_{it}$ as these are not absorbed by c_i when conditioning on n_i.

One drawback to this approach is that for a binary element k of $\boldsymbol{x}_{it}$, FEP only identifies $\alpha_k + \frac{1}{2}\omega_k$. Similarly, some elements of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\Omega}_0$ are not separately identified when $\boldsymbol{x}_{it}$ contains both levels and higher order terms.

This model nests the traditional case of constant coefficients, which occurs when $\boldsymbol{\gamma}_0 = \mathbf{0}$ and $\boldsymbol{\omega}_0 = \mathbf{0}$). Rejection of the null that $\boldsymbol{\gamma}_0 = \mathbf{0}$ is perhaps most convincing evidence of that slopes vary by individual. Therefore, the primary contribution of this approach to random coefficients is to suggest the inclusion of interactions between time-varying regressors and time averages to see if more flexibility is necessary.

If there is no evidence that slopes are correlated with the $\bar{\boldsymbol{x}}_i$, then one should carefully consider how to interpret inference on $\boldsymbol{\omega}_0$. Statistically significant estimates may just indicate that squares and cross-products of $\boldsymbol{x}_{it}$ belong in the FEP regression. Clearly if the cross-products are significant while the squares are not, or if the coefficients on squared terms are negative and significant, then the random coefficient framework does not make sense, though the results may still have yielded useful insight into the what functions of the explanatory variables should be included in the analysis.

2.3.5 Adding second moment assumptions

While under our assumptions, FEP is consistent under correct specification of the conditional mean (2.25), it may be possible to achieve greater efficiency by adding assumptions about the conditional second moment of $\boldsymbol{y}_i$. Another reason may be to identify the coefficients on binary variables.

I assume a variance function that is proportional to the conditional mean.

$$Var[y_{it}|\boldsymbol{x}_i, c_i, \boldsymbol{b}_i] = \sigma_0 c_i \exp(\boldsymbol{x}_{it}\boldsymbol{b}_i) \tag{2.27}$$

Additionally, the following CRE assumption implies conditional mean and variance functions that do not depend on c_i.

$$\log(c_i)|\boldsymbol{x}_i, \boldsymbol{b}_i \sim Normal(\psi_1 + \bar{\boldsymbol{x}}_i\boldsymbol{\xi}_1, \sigma_a^2) \tag{2.28}$$

Under assumptions 2.4, 2.24, 2.27, and 2.28, it follows from the properties of the lognormal distribution, the LIE, and the Law of Total Variance that

$$E(y_{it}|\boldsymbol{x}_i) = E(y_{it}|\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i) = \exp\left[h(\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + \frac{1}{2}v(\boldsymbol{x}_{it}, \boldsymbol{\tau}_0)\right] \tag{2.29}$$

and

$$\begin{aligned} Var(y_{it}|\boldsymbol{x}_i) =& Var(y_{it}|\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i) \\ =& \sigma_0 \exp\left[h(\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + \frac{1}{2}v(\boldsymbol{x}_{it}, \boldsymbol{\tau}_0)\right] \\ &+ \exp\left[2h(\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + v(\boldsymbol{x}_{it}, \boldsymbol{\tau}_0)\right]\left\{\exp\left[v(\boldsymbol{x}_{it}, \boldsymbol{\tau}_0)\right] - 1\right\}, \end{aligned} \tag{2.30}$$

where $\boldsymbol{\theta} \equiv (\psi_1, \boldsymbol{\xi}_1', \boldsymbol{\alpha}', \boldsymbol{\gamma}')'$, $\boldsymbol{\tau} = (\boldsymbol{\omega}_0', \sigma_a^2)'$, $h(\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) \equiv \psi_1 + \bar{x}_i\boldsymbol{\xi}_1 + \boldsymbol{x}_{it}\boldsymbol{\alpha}_0 + (\bar{\boldsymbol{x}}_i \otimes \boldsymbol{x}_{it})\boldsymbol{\gamma}_0$, and $v(\boldsymbol{x}_{it}, \boldsymbol{\tau}_0) \equiv \check{\boldsymbol{x}}_{it}\boldsymbol{\omega}_0 + \sigma_a^2$.

Estimation of $\boldsymbol{\theta}_0$ and $\boldsymbol{\tau}_0$ can then proceed using pooled normal QMLE, specifying the mean and variance functions as above. As the normal distribution is a member of the quadratic exponential family, this procedure is consistent without the normal distribution being true (Gourieroux, Monfort, and Trognon, 1984) Once again, inference should be made cluster-robust to account for serial correlation and the true distribution being non-normal. Estimation of $\boldsymbol{\beta}_0$ can then proceed as before, and coefficients on binary or quadratic variables are now identified off of the nonlinearity in (2.30).

Normal QMLE in this case is straightforward to program in software like Stata using built-in maximum likelihood functions, and it had good finite sample properties in simulations run for this

chapter. Some researchers may wish to specify a conditional covariance structure for $\boldsymbol{y}_i$ as a way to get more efficiency. If so, one option is to assume

$$Cov\left[y_{it}, y_{ir} | \boldsymbol{x}_i, c_i, \boldsymbol{b}_i\right] = 0, t \neq r. \tag{2.31}$$

Equation (2.31) does not allow serial correlation when conditioning on $\boldsymbol{x}_i, c_i, \boldsymbol{b}_i$, but the presence of the time-constant heterogeneity ensures that the responses will be serially correlated when conditioning on $\boldsymbol{x}_i$ only. Under 2.4, 2.24, 2.27, 2.31, and 2.28,

$$\begin{aligned} &Cov(y_{it}, y_{ir} | \boldsymbol{x}_i) = \\ &\quad \exp\left[h(\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + h(\boldsymbol{x}_{ir}, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + \frac{1}{2}\left(v(\boldsymbol{x}_{it}, \boldsymbol{\tau}_0) + v(\boldsymbol{x}_{ir}, \boldsymbol{\tau}_0)\right)\right] \left\{\exp(\boldsymbol{x}_{it} \boldsymbol{\Omega}_0 \boldsymbol{x}_{ir}' + \sigma_a^2) - 1\right\}. \end{aligned} \tag{2.32}$$

2.3.6 Estimating average partial effects

Even though the coefficients in (2.4) have direct interpretations as semi-elasticities, it may still be desirable to estimate partial effects and APEs, perhaps to compare estimates between competing nonlinear models. Moreover, this sections shows that the average partial effects for a binary variable depend only on $\alpha_k + \frac{1}{2}\omega_k$, meaning that even though we cannot separately identify α_k and ω_k without second moment assumptions, we can still estimate average partial effects.

Let $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$, c, and $\boldsymbol{b} = \{b_1, b_2 \ldots, b_K\}$ denote fixed values of the variables. The partial effect of a continuous x_{tj} on the conditional mean of y_t is defined as[8]

$$\phi_j(\boldsymbol{x}_t, c, \boldsymbol{b}) \equiv \frac{\partial E(y_t | \boldsymbol{x}_t, c, \boldsymbol{b})}{\partial x_{tj}} = c_i \exp(\boldsymbol{x}_t \boldsymbol{b}) b_j. \tag{2.33}$$

For a binary x_{tk}, the partial effect is defined as the discrete difference in the conditional mean of y_t at each level of the binary variable. In the expressions to follow, the subscript $\not{k}$ signifies that x_{tk}, $\bar{x}_k$, or their associated coefficients have been omitted from the vector.

[8] I implicitly assume that x_{tj} is not functionally linked with any other element in $\boldsymbol{x}_t$.

$$\begin{aligned}\phi_k(\boldsymbol{x}_t, c, \boldsymbol{b}) \equiv & E(y_t|\boldsymbol{x}_{t\not{k}}, x_{tk} = 1, c, \boldsymbol{b}) - E(y_t|\boldsymbol{x}_{t\not{k}}, x_{tk} = 0, c, \boldsymbol{b}) \\ = & c\exp(\boldsymbol{x}_{t\not{k}}\boldsymbol{b}_{\not{k}} + b_k) - c\exp(\boldsymbol{x}_{t\not{k}}\boldsymbol{b}_{\not{k}})\end{aligned} \tag{2.34}$$

Of course, estimating features of the distributions of ϕ_j and ϕ_k is infeasible as we do not observe c or $\boldsymbol{b}$. Therefore, this section focuses mainly on APEs where the heterogeneity has been averaged out.

$$\delta_h(\boldsymbol{x}_t) \equiv E_{\boldsymbol{v}_i}\left[\phi_h(\boldsymbol{x}_t, c_i, \boldsymbol{b}_i)\right], \tag{2.35}$$

where $\boldsymbol{v} \equiv (c, \boldsymbol{b}')'$ and $h \in \{j, k\}$.

2.3.6.1 Approaches under the CRE assumption for c_i

To proceed, it is necessary to maintain the assumptions of correlated random coefficients (2.24). As c_i is unobserved, I also maintain (2.28). Later, I will discuss a possible "estimator" of c_i. For now, there are two choices as to how to proceed in estimating δ_j and δ_k. The first is to estimate an Average Structural Function (ASF), as proposed by Blundell and Powell (2003), where essentially $\bar{\boldsymbol{x}}$ proxies for $\boldsymbol{v}$ and is averaged out before taking derivatives and differences. The second is to use derivatives and differences of (2.29) directly (Wooldridge, 2010).

The ASF is defined as:

$$ASF(\boldsymbol{x}_t) \equiv E_{\boldsymbol{v}_i}\left[c_i\exp(\boldsymbol{x}_t\boldsymbol{b}_i)\right], \tag{2.36}$$

where again, $\boldsymbol{x}_t$ is a fixed argument. Under (2.24), (2.27), and (2.28) the L.I.E. implies

$$ASF(\boldsymbol{x}_t) = E_{\bar{\boldsymbol{x}}}\left[\exp\left[h(\boldsymbol{x}_t, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + \frac{1}{2}v(\boldsymbol{x}_t, \boldsymbol{\tau}_0)\right]\right] \tag{2.37}$$

Passing the derivative through the expectation, the APE for continuous x_{tj} is:

$$\delta_j(\boldsymbol{x}_t) = E_{\bar{\boldsymbol{x}}}\left[\exp\left(h(\boldsymbol{x}_t, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + \frac{1}{2}v(\boldsymbol{x}_t, \boldsymbol{\tau}_0)\right)\left(\alpha_j + \bar{\boldsymbol{x}}_i\boldsymbol{\gamma}_j' + \omega_j x_{tj} + \sum_{h\neq j}^{K}\rho_{jh}x_{th}\right)\right] \tag{2.38}$$

For a binary x_{tk}, the APE is:

$$\begin{aligned}\delta_k(\boldsymbol{x}_t) =& E_{\bar{\boldsymbol{x}}}\left[E\left(y_t|\boldsymbol{x}_{t\not k}, x_{tk}=1, \bar{\boldsymbol{x}}_i\right) - E\left(y_t|\boldsymbol{x}_{t\not k}, x_{tk}=0, \bar{\boldsymbol{x}}_i\right)\right] \\ =& E_{\bar{\boldsymbol{x}}}\left[\exp\left(h(\boldsymbol{x}_{t\not k}, 1, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + \frac{1}{2}v(\boldsymbol{x}_{t\not k}, 1, \boldsymbol{\tau}_0)\right) - \exp\left(h(\boldsymbol{x}_{t\not k}, 0, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + \frac{1}{2}v(\boldsymbol{x}_{t\not k}, 0, \boldsymbol{\tau}_0)\right)\right],\end{aligned} \tag{2.39}$$

where

$$\begin{aligned} h(\boldsymbol{x}_{t\not k}, 1, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) =& \psi_1 + \bar{x}_i\boldsymbol{\xi}_1 + \boldsymbol{x}_{t\not k}\boldsymbol{\alpha}_{\not k} + \boldsymbol{x}_{t\not k}\boldsymbol{\Gamma}_{\not k}\bar{\boldsymbol{x}}_{\not k} + \boldsymbol{x}_{t\not k}\bar{x}_{ik}\boldsymbol{\gamma}^k_{\not k} + \alpha_k + \bar{\boldsymbol{x}}_i\boldsymbol{\gamma}'_k, \\ h(\boldsymbol{x}_{t\not k}, 0, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) =& \psi_1 + \bar{x}_i\boldsymbol{\xi}_1 + \boldsymbol{x}_{t\not k}\boldsymbol{\alpha}_{\not k} + \boldsymbol{x}_{t\not k}\boldsymbol{\Gamma}_{\not k}\bar{\boldsymbol{x}}_{\not k} + \boldsymbol{x}_{t\not k}\bar{x}_{ik}\boldsymbol{\gamma}^k_{\not k}, \\ v(\boldsymbol{x}_{t\not k}, 1, \boldsymbol{\tau}_0) =& \check{\boldsymbol{x}}_{\not k}\boldsymbol{\omega}_{\not k} + \sigma_a^2 + \omega_k + 2\sum_{h\neq k}^{K}\rho_{kh}x_{th}, \\ \text{and } v(\boldsymbol{x}_{t\not k}, 0, \boldsymbol{\tau}_0) =& \check{\boldsymbol{x}}_{\not k}\boldsymbol{\omega}_{\not k} + \sigma_a^2. \end{aligned} \tag{2.40}$$

The direct approach consists of taking derivatives and differences of 2.29 directly. Note that since these expressions do not first average out $\bar{\boldsymbol{x}}$, the entire history of $\boldsymbol{x}$ is now a fixed argument. For a continuous variable x_{tj} the APE is:

$$\begin{aligned}\delta_j(\boldsymbol{x}) =& \frac{\partial E(y_t|\boldsymbol{x})}{\partial x_{tj}} \\ =& \exp\left(h(\boldsymbol{x}_t, \bar{\boldsymbol{x}}, \boldsymbol{\theta}_0) + \frac{1}{2}v(\boldsymbol{x}_t, \boldsymbol{\tau}_0)\right)\left(\xi_j/T + \alpha_j + \bar{\boldsymbol{x}}\boldsymbol{\gamma}'_j + \frac{1}{T}\boldsymbol{x}_t\boldsymbol{\gamma}^j + \omega_j x_{tj} + \sum_{h\neq j}^{K}\rho_{jh}x_{th}\right),\end{aligned} \tag{2.41}$$

where $\boldsymbol{\gamma}_j$ is the jth row and $\boldsymbol{\gamma}^j$ is the jth column of $\boldsymbol{\Gamma}_0$.

Define $z(\boldsymbol{x}_t, \bar{\boldsymbol{x}}, \boldsymbol{\theta}, \boldsymbol{\tau}) = h(\boldsymbol{x}_t, \bar{\boldsymbol{x}}, \boldsymbol{\theta}) + \frac{1}{2}v(\boldsymbol{x}_t, \boldsymbol{\tau})$. Then we have for a binary x_{tk},

$$\begin{aligned}\delta_k(\boldsymbol{x}) =& E\left(y_t|\boldsymbol{x}_{t\not k}, \{x_{sk}\}_{s\neq t}^T, x_{tk}=1\right) - E\left(y_t|\boldsymbol{x}_{t\not k}, \{x_{sk}\}_{s\neq t}^T, x_{tk}=0\right) \\ =& \exp\left(z(\boldsymbol{x}_{t\not k}, \bar{\boldsymbol{x}}_{\not k}, \boldsymbol{\theta}_{\not k}, \boldsymbol{\tau}_{\not k}) + \xi_j\bar{x}^{(1)}_{tk} + \alpha_k + \bar{\boldsymbol{x}}_{\not k}\boldsymbol{\gamma}'_{k\not k} + \gamma_{kk}\bar{x}^{(1)}_{tk} + \boldsymbol{x}_{t\not k}\bar{x}^{(1)}_{tk}\boldsymbol{\gamma}^k_{\not k} + \frac{1}{2}\omega_k + \sum_{h\neq k}^{K}\rho_{kh}x_{th}\right) \\ & - \exp\left(z(\boldsymbol{x}_{t\not k}, \bar{\boldsymbol{x}}_{\not k}, \boldsymbol{\theta}_{\not k}, \boldsymbol{\tau}_{\not k}) + \xi_j\bar{x}^{(0)}_{tk} + \boldsymbol{x}_{t\not k}\bar{x}^{(0)}_{tk}\boldsymbol{\gamma}^k_{\not k}\right),\end{aligned} \tag{2.42}$$

where γ_{kk} is the kth diagonal element of $\boldsymbol{\Gamma}_0$, $\bar{x}_{tk}^{(1)} \equiv \frac{1}{T}\left(1+\sum_{s\neq t}^{T} x_{sk}\right)$, and $\bar{x}_{tk}^{(0)} \equiv \frac{1}{T}\sum_{s\neq t}^{T} x_{sk}$.

Whichever approach is chosen, one can then estimate $\delta_j(\boldsymbol{x}_t)$ or $\delta_k(\boldsymbol{x}_t)$ by inserting the estimated parameters, replacing expectations over the distribution of $\bar{\boldsymbol{x}}$ with averages over i, and plugging in interesting values of $\boldsymbol{x}$. Many researchers will average over the distribution of $\boldsymbol{x}$ to get a single number. Asymptotic variances can be computed either via the delta method or using the panel bootstrap.

2.3.6.2 Estimation when the slopes are independent of covariates

The traditional case where $\boldsymbol{b}_i$ is independent of $\boldsymbol{x}_i$ (conditional on c_i) is one where the ASF is identified without placing any restriction on c_i or $Var(y_{it}|\boldsymbol{x}_i)$. The following summarizes the necessary condition.

$$\begin{aligned} \boldsymbol{b}_i &= \boldsymbol{\beta}_0 + \boldsymbol{d}_i, \\ \boldsymbol{d}_i|(\boldsymbol{x}_i, c_i) &\sim Normal(\boldsymbol{0}, \boldsymbol{\Omega}_0). \end{aligned} \tag{2.43}$$

The results of Section 3.4 continue to hold, but the time averages no longer enter $E(y_{it}|\boldsymbol{x}_{it}, c_i)$ (that is, $\boldsymbol{\Gamma}_0 = \boldsymbol{0}$).

The LIE implies that for a fixed $\boldsymbol{x}_t$,

$$ASF(\boldsymbol{x}_t) = E(c_i)\exp\left(\boldsymbol{x}_t\boldsymbol{\beta}_0 + \frac{1}{2}\boldsymbol{x}_t\boldsymbol{\Omega}_0\boldsymbol{x}_{it}'\right) \tag{2.44}$$

Passing the derivative through the expectation, the APE of a continuous variable x_{tj} is given by:

$$\delta_j(\boldsymbol{x}_t) = E(c_i)\exp\left(\boldsymbol{x}_t\boldsymbol{\beta}_0 + \frac{1}{2}\boldsymbol{x}_t\boldsymbol{\Omega}_0\boldsymbol{x}_{it}'\right)\left(\beta_j + +\omega_j x_{tj} + \sum_{h\neq j}^{K}\rho_{jh}x_{th}\right) \tag{2.45}$$

For a binary variable x_{tk}, the APE is:

$$\delta_k(\boldsymbol{x}_t) =$$
$$E(c_i)\left[\exp\left(\boldsymbol{x}_{tk}\boldsymbol{\beta}_k + \frac{1}{2}\check{\boldsymbol{x}}_{tk}\boldsymbol{\omega}_k + \alpha_k + \frac{1}{2}\omega_k + \sum_{h\neq k}^{K}\rho_{kh}x_{th}\right) - E(c_i)\exp\left(\boldsymbol{x}_{tk}\boldsymbol{\beta}_k + \frac{1}{2}\check{\boldsymbol{x}}_{tk}\boldsymbol{\omega}_k\right)\right]. \tag{2.46}$$

An estimator for $E(c_i)$ is conveniently available. Poisson QMLE using (2.25) and treating the c_i as (strictly positive) parameters is algebraically equivalent to multinomial QCMLE. [9]) In our current application, for a given $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}', \boldsymbol{\omega}')'$, the QMLE for c_i is:

$$c_i(\boldsymbol{\theta}) = \frac{n_i}{\sum_{t=1}^{T}\exp(\boldsymbol{x}_{it}\boldsymbol{\beta} + \check{\boldsymbol{x}}_{it}\boldsymbol{\omega})}, \tag{2.47}$$

where again, $n_i = \sum_{t=1}^{T} y_{it}$. Define $\widehat{c}_i = c_i(\widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}}$ is the FEP estimate of $(\boldsymbol{\beta}_0', \boldsymbol{\omega}_0')'$. The properties of $\widehat{c}_i$ are not well-known in either the constant or heterogeneous slope case. Though there is no incidental parameters problem for $\widehat{\boldsymbol{\theta}}$ in the FEP case, $c_i(\boldsymbol{\theta}) \neq c_i$, even when evaluated at $\boldsymbol{\theta}_0$. Viewing c_i as a parameter, there is no reason to think $\widehat{c}_i$ is unbiased and it cannot be consistent with T fixed.

However, the ASF in this case is proportional to $E(c_i)$. Strict exogeneity of $\boldsymbol{x}_{it}$ and (2.24) imply that

$$E(n_i|c_i, \boldsymbol{x}_i) = c_i \sum_{t=1}^{T}\exp\left(\boldsymbol{x}_{it}\boldsymbol{\beta}_0 + \frac{1}{2}\check{\boldsymbol{x}}_{it}\boldsymbol{\omega}_0\right) \tag{2.48}$$

It follows from the L.I.E. that

$$E(c_i) = E\left(\frac{n_i}{\sum_{t=1}^{T}\exp\left(\boldsymbol{x}_{it}\boldsymbol{\beta}_0 + \frac{1}{2}\check{\boldsymbol{x}}_{it}\boldsymbol{\omega}_0\right)}\right) \tag{2.49}$$

meaning $N^{-1}\sum_{i=1}^{N}\widehat{c}_i$ consistently estimates $E(c_i)$.

Many researchers are primarily interested in a single APE estimate (averaged across the sample of observables). In this case, it may be attractive to treat $\widehat{c}_i$ as the unobservable c_i and average

[9] See Wooldridge, 2010 or Cameron and Trivedi, 2013

across the distributions of $\widehat{c_i}$ and $\boldsymbol{x}_i$ at the same time. We would, generally expect such APE estimators for nonlinear FE models derived in such a way to suffer from the incidental parameters problem, even if the slopes are estimated consistently. [10] Given that $N^{-1}\sum_{i=1}^{N}\widehat{c_i}$ is consistent for $E(c_i)$, however, it may be that estimators including functions of c_i that are averaged across i have desirable properties. This appears to be true at least for the data generating process considered in this chapter. Simulation results in Section 4 indicate very small finite sample bias of overall APE estimators computed using $\widehat{c_i}$ in this way.

2.4 Monte Carlo

2.4.1 Comparing estimation methods

To illustrate the impact of ignoring random coefficients in the FEP setting, I simulate the performance of the different estimators in both the ideal case of constant coefficients and in the case where the coefficients vary by individual. I employed the following data generating process:

$$y_{it}|(\boldsymbol{x}_i,\boldsymbol{w}_i,c_i,b_{i1},b_{i2}) \sim \text{Poisson}\left[c_i\exp(b_{i1}x_{it}+b_{i2}w_{it})\right], \tag{2.50}$$

$$\log(c_i) \sim Normal(0,1/16) \tag{2.51}$$

$$\begin{aligned} x_{it} &= \log(c_i)+.5x_{i,t-1}+v_{it}, t>1 \\ x_{i1} &= \log(c_i)_i+v_{i1},\, v_{it}\sim N(0,1/2) \end{aligned} \tag{2.52}$$

$$w_{it} = \mathbf{1}\left[x_{it}+h_{it}>0\right], h_{it}\sim N(0,1/2) \tag{2.53}$$

$$\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim Normal\left[\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} \omega_1^2 & \rho \\ \rho & \omega_2^2 \end{pmatrix}\right] \tag{2.54}$$

[10] See, for example, Fernandez-Val, 2009.

For the above draws, $i = 1, \ldots, 1000$ and $t = 1, \ldots, 10$. The case where ω_1^2, ω_2^2, and ρ all equal zero corresponds to the constant coefficient case. For these simulations, the b_{ij} are generated to be independent of $\{\boldsymbol{x}_i, \boldsymbol{w}_i\}$, and this assumption is maintained in estimation. The b_{ij} are also generated to be independent of each other ($\rho = 0$) but this is not assumed in estimation.

In the following tables, FEP refers to the estimator that ignores the random coefficients. FEP2 refers to the estimator that adds the square of x and an interaction between x and w. Since this model's assumptions does not separately identify β_2 and ω_2^2 , the estimated coefficient on w is compared to $\beta_2 + \frac{1}{2}\omega_2^2$. NQML refers to the normal QML estimator that also assumes (2.27) and (2.28).[11] I set $\omega_1 = \omega_2 = \omega$ but do not assume equal variance in estimation. In each case, I used one thousand replications.

[11] APE estimates from NQML also plugged in $\widehat{c}_i$.

Table 2.1: Finite Sample Properties of Slope Estimators: $\beta_1 = 1, \beta_2 = -1$

	$\widehat{\beta}_1$						$\widehat{\beta}_2$				$\widehat{\beta_2 + \frac{1}{2}\omega_2^2}$		
	FEP		FEP2		NQML		FEP		NQML		FEP2		
ω	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Truth
0.00	1.00	0.02	1.00	0.03	1.00	0.02	-1.00	0.03	-1.00	0.04	-1.00	0.04	-1.00
0.05	1.00	0.02	1.00	0.03	1.00	0.02	-1.00	0.03	-1.00	0.04	-1.00	0.04	-1.00
0.10	1.01	0.02	1.00	0.03	1.00	0.02	-1.00	0.03	-1.00	0.04	-0.99	0.04	-1.00
0.15	1.02	0.02	1.00	0.03	1.00	0.03	-0.99	0.03	-1.00	0.04	-0.99	0.04	-0.99
0.20	1.03	0.02	1.00	0.03	1.00	0.03	-0.99	0.03	-1.00	0.04	-0.98	0.04	-0.98
0.25	1.05	0.03	1.00	0.03	1.00	0.03	-0.98	0.03	-1.00	0.04	-0.97	0.04	-0.97
0.30	1.07	0.03	1.00	0.03	1.00	0.03	-0.98	0.04	-1.00	0.04	-0.96	0.04	-0.96
0.35	1.10	0.04	1.00	0.04	1.00	0.04	-0.97	0.04	-0.99	0.05	-0.94	0.04	-0.94
0.40	1.14	0.06	1.00	0.04	1.00	0.04	-0.96	0.05	-0.99	0.05	-0.93	0.05	-0.92
0.45	1.18	0.07	1.00	0.04	0.99	0.05	-0.96	0.07	-0.99	0.05	-0.91	0.05	-0.90
0.50	1.23	0.09	1.00	0.04	0.99	0.05	-0.95	0.08	-0.98	0.06	-0.89	0.05	-0.88

Table 2.2: Finite Sample Properties of APE Estimators: $\beta_1 = 1, \beta_2 = -1$

	Est. APE of x							Est. APE of w						
		FEP		FEP2		NQML			FEP		FEP2		NQML	
ω	Truth	Mean	SD	Mean	SD	Mean	SD	Truth	Mean	SD	Mean	SD	Mean	SD
0.00	0.88	0.88	0.03	0.88	0.03	0.88	0.03	-1.12	-1.12	0.06	-1.12	0.08	-1.12	0.07
0.05	0.88	0.88	0.03	0.88	0.03	0.88	0.03	-1.12	-1.12	0.06	-1.12	0.08	-1.12	0.07
0.10	0.90	0.89	0.03	0.89	0.03	0.89	0.03	-1.13	-1.13	0.06	-1.13	0.09	-1.13	0.08
0.15	0.91	0.92	0.04	0.92	0.04	0.92	0.04	-1.14	-1.15	0.07	-1.14	0.10	-1.14	0.08
0.20	0.95	0.95	0.04	0.95	0.04	0.95	0.04	-1.15	-1.17	0.07	-1.15	0.10	-1.15	0.08
0.25	0.98	0.99	0.05	0.99	0.05	0.99	0.05	-1.16	-1.19	0.09	-1.17	0.12	-1.16	0.10
0.30	1.04	1.04	0.07	1.04	0.06	1.03	0.06	-1.19	-1.23	0.11	-1.19	0.16	-1.18	0.11
0.35	1.11	1.11	0.09	1.11	0.08	1.10	0.09	-1.22	-1.27	0.14	-1.22	0.20	-1.20	0.13
0.40	1.20	1.21	0.15	1.21	0.13	1.20	0.14	-1.26	-1.35	0.22	-1.26	0.29	-1.23	0.16
0.45	1.32	1.32	0.22	1.33	0.21	1.31	0.23	-1.30	-1.43	0.34	-1.30	0.47	-1.26	0.24
0.50	1.49	1.49	0.47	1.50	0.49	1.48	0.48	-1.36	-1.57	0.86	-1.35	0.60	-1.29	0.26

It appears from Table 2.1 that the standard deviation of the coefficients is positively related to the finite sample bias (in magnitude) in FEP slope estimates. This is not surprising given that (2.1) fails for $\omega > 0$. This is despite the fact that the coefficients are independent of the covariates and each other, a case in which random coefficients would not cause a problem in linear models. In contrast, the augmented FEP and the NQML estimators show much smaller bias at all levels of ω, with the exception of the FEP2 coefficient on w, which, as expected, appears to show small bias for $\beta_2 + \frac{1}{2}\omega^2$.

The APEs are estimated using expressions similar to (2.45) and (2.46) using the FEP2 and NQML parameter estimates. The difference is I treat $\widehat{c}_i$ as c_i and average over $\{\boldsymbol{x}_{it}, \widehat{c}_i\}$ only once. I followed an analogous procedure for the FEP case.

Table 2.2 suggests that this approach to estimating APEs has small bias for the FEP2 and NQML case, despite using estimates of incidental parameters. For FEP, bias in the APE of the binary variable increases as ω increases. Surprisingly, this is not the case for the continuous variable. Even though the simulation suggests a large bias in the FEP estimate of β_1. This warrants further investigation as it suggests there many be circumstances in which researchers can ignore random coefficients if all they care about is APEs of continuous variables, though it could also be an artifact of this data generating process.

2.4.2 Testing when coefficients are not normal

Section 3 shows that for slope heterogeneity in a location-scale family of spherical distributions (where the heterogeneity are independent of each other), an LM test for coefficient heterogeneity is equivalent to testing the coefficients on squares of the covariates, which suggests that the heterogeneity need not be normal for the approach of this chapter to work well. To explore this, I generate the responses using random coefficients of different distributions.

$$b_{ij2} = 1 + \omega\left((u_{j2} - 0.5)/\sqrt{1/12}\right), u_{j2} \sim U(0,1) \tag{2.55}$$

$$b_{ij3} = 1 + \omega\left((u_{j3} - 4)/\sqrt{8}\right), u_{j3} \sim \chi^2_4 \tag{2.56}$$

$$b_{ij4} = 1 + \omega\left(u_{j4}/\sqrt{5/3}\right), u_{j4} \sim t_5 \tag{2.57}$$

$$b_{ij5} = 1 + \omega\left(u_{j5} - 1\right), u_{j5} \sim \textit{Exponential}\ (1) \tag{2.58}$$

$$b_{ij6} \sim \textit{Gamma}\ (1/\omega^2, \omega^2) \tag{2.59}$$

These draws are made separately for $j = 1, 2$, and for simplicity, $Cov(b_{i1h}, b_{i2h}) = 0$ for each h. Each coefficient's data generating process ensures that it has a mean of 1 and variance of ω^2. Each of the first five coefficients falls into a location-scale family as they consist of a standardized random variable multiplied by ω to result in a variance of ω^2 and shifted to have a mean of one. The gamma coefficients, in contrast, are not drawn from a location-scale family, but are directly specified to have a mean of 1 and variance of ω^2.

Given the issue identifying parameters associated with binary regressors in the FEP2 setting, I generate the responses to depend on continuous regressors only, where each $x_{it\,j}$ is generated as in (2.52).

$$y_{it}|(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, c_i, b_{i1h}, b_{i2h}) \sim \text{Poisson}\left[c_i \exp(b_{i1h} x_{it1} + b_{i2h} x_{it2})\right] \tag{2.60}$$

After generating the data, β_1, β_2, ω_1^2, ω_2^2, and ρ were estimated using FEP of y_t on x_{t1}, x_{t2}, x_{t1}^2, x_{t2}^2, and $x_{t1}x_{t2}$. A Wald test was then performed on x_{t1}^2, x_{t2}^2, and $x_{t1}x_{t2}$. The results of Section 3.3 suggest that this test should perform well for the first five coefficient types, and I conjecture that it performs well for the Gamma coefficients as well. When testing for random slopes, is important to use a FE procedure if one is concerned that the multiplicative effect c_i is correlated with the explanatory variables. Otherwise, the omitted variable problem is likely to cause the test to be over-sized. In fact, in a simulation where Random Effects Poisson was used on the same set of

covariates, a Wald test rejected the null of constant slopes in 88% of replications when the true slopes were nonrandom.

Table 2.3: Testing when $\boldsymbol{b}_i$ is not normal

	Empirical Rejection Probability (Null value 0.05)					
ω	Normal	Uniform*	Chi2*	t_5*	Exp.*	Gamma
0.00	0.069	0.069	0.069	0.069	0.069	0.069
0.05	0.108	0.115	0.112	0.108	0.121	0.132
0.10	0.186	0.212	0.159	0.196	0.16	0.178
0.15	0.308	0.359	0.287	0.302	0.303	0.334
0.20	0.468	0.531	0.439	0.408	0.404	0.472
0.25	0.640	0.691	0.543	0.579	0.553	0.625
0.30	0.785	0.796	0.689	0.693	0.652	0.741
0.35	0.881	0.887	0.796	0.804	0.757	0.817
0.40	0.914	0.948	0.860	0.852	0.814	0.868
0.45	0.931	0.965	0.897	0.897	0.876	0.892
0.50	0.970	0.979	0.904	0.919	0.876	0.923

Table 2.3 shows that as expected, rejection probabilities increase with ω when the coefficients are normal, and are quite high when ω is large.[12] What is interesting is that there does not seem to be much change in either size or finite sample power when the coefficients are not normal, even when the coefficients are not drawn from a location-scale family.

2.5 Empirical application: the Patent-R&D relationship

There is a long history of economic inquiry into the relationship between a firm's research and development (R&D) expenditures and the number of patents for which it applies in a given year. Patent applications are viewed in the literature as an indicator of additions to the knowledge stock of a firm (Pakes and Griliches, 1980). Pakes and Griliches (1980) were among the first to focus on firm effects as a source of potential endogeneity in analyzing U.S. manufacturing firms. Hausman, Hall, and Griliches (1984) and Hall, Griliches, and Housman (1986) also look to firm effects to account for significant over-dispersion in the distribution of patent counts. In addition to FEP,

[12] I have not yet varied the cross-section size. I would expect these rejection probabilities to increase.

Negative Binomial models are also common as a way to introduce more dispersion. Nonlinear count models are not only attractive for logical reasons, but also because datasets can contain a nontrivial proportion of observations with zero patents. These observations must be eliminated or transformed in some ad hoc manner before estimating a linear log-log model(Hall, Griliches, and Hausman, 1986). Such observations seem to be more common in more recent datasets as well. While only 8% of observations were zero in Hall, Hausman, and Griliches 1968-1975 panel of 121 firms, 16.5% were zero in Gurmu and Perez-Sebastian's 1982-1992 panel of 391 firms (Gurmu and Pérez-Sebastián, 2008).

A common finding in the literature is that distributed lag models that do not account for any firm heterogeneity tend to have a U-shaped lag profile, and that after accounting for firm heterogeneity, only contemporaneous R & D expenditure tends to be significant (Hall, Griliches, and Hausman, 1986). In a cross-sectional analysis of the pharmaceutical industry, Wang, Cockburn, and Puterman (1998) use a Poisson model and allow for heterogeneity in both the multiplicative effect and coefficients. While the mixing distribution is allowed to depend on the regressors, they assume that the vector of heterogeneity has finite support, which in their analysis consisted of three or fewer points. This framework may be less palatable in studies with broader industry coverage.

The population of interest for this chapter is publicly-traded U.S. manufacturing firms in existence from 1996 to 2003. The patent data come from the United States Patent and Trademark Office by way of the National Bureau of Economic Research's Patent Data Project (PDP) and includes data through 2006. As patents are not recorded in the USPTO database until they are granted, the panel is truncated in 2003 to diminish the effect of the time-lag between application and granting.[13] Financial information on publicly-traded firms comes from the Compustat database, accessed through Wharton Research Data Services (WRDS) in September 2016. Hall, Jaffe, and Trajtenberg (2001) and Bessen (2009) thoroughly describe the patent data as well as matching information for the Compustat database. Matching patents to firms is not a trivial given

[13]The average lag over applications made in 1990-92 was 1.76 years, with 96.1% of patents granted in three years or less.

nonstandard naming in USPTO records, among other issues.

I mainly follow Bound, et. al (1982) and Hall, Griliches, and Hausman (1986) in assembling the panel dataset. The initial sample from the Compustat database consists of 3,126 firms in the U.S. manufacturing industry that were in existence in the year 2000. Following the literature, I require that data exist for patents and R&D expenditures for each year from 1996 to 2003, and that R&D expenditures be strictly positive since I take logs. I also eliminate firms that show large jumps in either gross capital or employment in a year. In the end, my sample consists of 848 firms over the period 1996-2003. I describe the selectivity of my sample in Tables 2.4 and 2.5. The tables show that although the sample covers only about a quarter of U.S. manufacturing firms in 2000, it covers nearly 70% of R&D expenditures. Coverage is generally poorer for smaller firms and higher for larger firms both in terms of net sales and R&D. Sample coverage is comparable to Hall, Griliches, and Hausman (1986) in terms of net sales, though they achieve 90% coverage of total R&D.

Table 2.4: Distribution of Net Sales in 2000

	Number in 2000 cross-section		Number in Sample	Coverage	
Net Sales	All	Pos. R&D		All	Pos. R&D
Less than $1M	332	207	49	0.15	0.24
$1M-10M	439	335	115	0.26	0.34
$10M-100M	900	672	242	0.27	0.36
$100M-1B	986	588	244	0.25	0.41
$1B-10B	402	271	157	0.39	0.58
More than $10B	67	52	41	0.61	0.79
Total	3,126	2,125	848	0.27	0.40

Table 2.5: R& D Expenditures in 2000

Firm R&D (2000 USD)	2000 Cross-section	Sample	Coverage
Less than $1M	170.15	55.32	0.33
$1M-10M	3695.48	1492.38	0.40
$10M-100M	21621.47	8765.10	0.41
$100M-1B	38160.81	25075.92	0.66
$1B-10B	67084.16	54007.14	0.81
Total	130732.08	89395.85	0.68

Table 2.6 shows summary statistics for the key variables over the sample of 848.[14] Consistent with the literature, this shows the distribution of patents to be right-skewed and over-dispersed with a thick right tail. Also noteworthy is that compared to previous studies, my sample contains a much higher proportion of zeros than previous studies. Compared to either Hall, Griliches, and Hausman (1986) or Gurmu and Perez-Sebastian (2008), the median number of patents is lower, and the maximum number of patents is higher in this sample.

[14]Note that firms with zero patents in all years drop from the multinomial log-likelihood.

Table 2.6: Summary of Key Variables in 2000

Variable	Mean	St.Dev.	Min	1st Q.	Med.	3rd Q.	Max
Net Sales (Millions of USD)	2506.28	12980.46	0.00	15.77	118.73	877.54	206083.00
R&D (Millions of USD)	105.42	490.95	0.01	2.22	7.53	31.71	6800.00
Patents	30.47	141.85	0.00	0.00	2.00	7.00	1811.00
Fraction with zero patents	0.35	0.48	–	–	–	–	–
Fraction in scientific sector	0.55	0.50	–	–	–	–	–

All dollars amounts are real 2000 USD.

The scientific sector is defined to include the drug, computer, electronic component, and scientific instrument industries.

I apply the exponential model introduced in Section 3 to patent counts where the regressors of interest are the logs of current R&D and up to three lags. I include year dummies, but assume their coefficients are constant.

$$E\left[patents_{it}|\log(R_{i1}),\ldots,\log(R_{iT}),\delta_t,c_i,\boldsymbol{b}_i\right]=c_i\exp\left(\sum_{s=0}^{\tau}b_{i,s}\log(R_{i,t-s})+\delta_t\right), \tag{2.61}$$

where R_{it} is real R&D expenditures by firm i in year t. The CRC assumption is:

$$\boldsymbol{b}_i|(\log(R_{i,t-0}),\ldots,\log(R_{i,t-\tau}),\delta_t,c_i)\sim Normal(\boldsymbol{\alpha}+\boldsymbol{\gamma}'\overline{\log(R)}_i,\boldsymbol{\Omega}), \tag{2.62}$$

where $\overline{\log(R)}_i=T^{-1}\sum_{t=1}^{T}\log(R_{it})$ is a scalar. Section 3 implies that FEP of patents on current and lagged $\log(R)$ terms, interactions between $\overline{\log(R)}$ and the $\log(R)$ terms, and squares and cross-products of the $\log(R)$ terms will be consistent under these assumptions.

Table 2.7: Results for traditional estimators

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	PQML 1	PQML 2	FEOLS 1	FEOLS 2	FEP 1	FEP 2
$\log(R_0)$	0.819***	0.423**	0.113***	0.0476**	0.318***	0.161***
	(0.0441)	(0.191)	(0.0198)	(0.0205)	(0.0682)	(0.0560)
$\log(R_{-1})$		0.234***		0.00784		0.0158
		(0.0637)		(0.0192)		(0.0378)
$\log(R_{-2})$		0.0845		0.00777		-0.0250
		(0.108)		(0.0180)		(0.0710)
$\log(R_{-3})$		0.0826		-0.00789		-0.00236
		(0.203)		(0.0204)		(0.0546)
Dum. for zero pat.			-0.543***	-0.442***		
			(0.0261)	(0.0301)		
Constant	-0.211	-0.228	1.091***	1.268***		
	(0.211)	(0.214)	(0.0440)	(0.0765)		
Sum of $\log(R)$ coeff.	0.819***	0.824***	0.113***	0.055	0.318***	0.1495
	(0.0441)	(0.045)	(0.0198)	(0.034)	(0.0682)	(0.1096)
Observations	6,784	4,240	6,784	4,240	5,968	3,510
Number of firms	848	848	848	848	746	702
R-squared			0.157	0.137		

Clustered standard errors in parentheses. Year dummies included in all specifications.

*** p<0.01, ** p<0.05, * p<0.1

Table 2.7 presents results from the six different specifications that assume constant coefficients. For all but columns (3) and (4), the dependent variable is the number of patents. Columns (1) and

(2) contains Poisson QMLE estimates where firm heterogeneity is ignored. Column (3) contains estimates from FE OLS where the dependent, variable is the log of patents. For this column only, zero patent counts are changed to 1, with a dummy variable added following Hall, Griliches, and Hausman (1986). Columns (5) and (6) contain FEP estimates.

Consistent with the literature, these estimates imply that correlation between patents and current R&D is strongest relative to lag effects, and that the total elasticity of patents with respect to R&D that is less than unity. I also find the estimated elasticities fall once I account for firm effects. For the Poisson specification, the total elasticity falls from 0.82 to 0.32 in the one-lag model and from 0.82 to 0.15 in the three-lag model. The three-lag FEP specification implies an elasticity with respect to current R&D that is only about half of those estimated in previous studies, and this estimate is sensitive to the time dimension of the panel and lag-length chosen. If I mimic Gurmu and Perez-Sebastian (2008) and estimate a four-lag FEP model over 1982-1992, I get very similar results to theirs. It is possible that the nature of the patent-R&D relationship changed in the intervening decade, but it may also be that the exponential model is incorrect, our specification neglects some dynamics or endogeneity, or that sample selection has had a different effect on the more current data.

Additionally, Section 3 and Section 5 imply that neglected slope heterogeneity could also be a source of bias in this model. Table 2.8 gives results from the CRC estimator proposed in this chapter, varying the lag length and assumptions about $\mathbf{\Omega}$. In columns (1) and (3), I impose that the $\boldsymbol{b}_i$ are deterministic linear functions of $\overline{\log(R)}_i$, while in column (4), I impose that $\mathbf{\Omega}$ is diagonal.

Given (2.61) and (2.62), these data do provide some evidence of slope heterogeneity. In the one-lag models, none of the additional terms are statistically significant. The evidence is mixed in the three-lag models. In column (3), the estimates of $\boldsymbol{\gamma}$ are jointly marginally significant ($p = 0.08$), with the interaction involving the second lag of $log(R)$ negative and significant at the 5% level. In column (4), while all terms involving $\log(R)$ are jointly significant, the interactions and squares are not. In column (5), the interactions, squares, and cross-products are jointly marginally signifcant ($p = 0.08$). The terms associated with $\mathbf{\Omega}$ are jointly insignificant, however, as are the interactions

Table 2.8: Results for CRC FEP estimators

VARIABLES	(1) CRCFEP 1	(2) CRCFEP 2	(3) CRCFEP 3	(4) CRCFEP 4	(5) CRCFEP 5
$\log(R_0)$	0.538***	0.548***	0.115	0.152	0.160
	(0.144)	(0.151)	(0.141)	(0.133)	(0.141)
$\log(R_{-1})$			0.0736	0.0604	0.111
			(0.0892)	(0.0951)	(0.0887)
$\log(R_{-2})$			0.444**	0.423***	0.360***
			(0.173)	(0.148)	(0.121)
$\log(R_{-3})$			-0.0384	-0.00633	0.0205
			(0.149)	(0.142)	(0.125)
$\log(R_0) \times \overline{\log(R_0)}$	-0.0394	0.165	0.00850	-0.182	-0.215
	(0.0285)	(0.183)	(0.0248)	(0.224)	(0.251)
$\log(R_{-1}) \times \overline{\log(R_0)}$			-0.0103	-0.118	0.0177
			(0.0167)	(0.195)	(0.294)
$\log(R_{-2}) \times \overline{\log(R_0)}$			-0.0844**	-0.556**	-0.167
			(0.0368)	(0.258)	(0.313)
$\log(R_{-4}) \times \overline{\log(R_0)}$			0.00672	-0.0775	-0.236
			(0.0284)	(0.159)	(0.262)
$[\log(R_0)]^2$		-0.102		0.0915	0.0921
		(0.0892)		(0.108)	(0.118)
$[\log(R_{-1})]^2$				0.0569	0.102
				(0.0978)	(0.108)
$[\log(R_{-2})]^2$				0.234**	0.309**
				(0.118)	(0.147)
$[\log(R_{-3})]^2$				0.0404	0.117
				(0.0735)	(0.0854)
$\log(R_0) \times \log(R_{-1})$					-0.0986
					(0.141)
$\log(R_0) \times \log(R_{-2})$					-0.0120
					(0.177)
$\log(R_0) \times \log(R_{-3})$					0.144
					(0.176)
$\log(R_{-1}) \times \log(R_{-2})$					-0.255
					(0.183)
$\log(R_{-1}) \times \log(R_{-3})$					0.123
					(0.129)
$\log(R_{-2}) \times \log(R_{-3})$					-0.266**
					(0.118)

Clustered standard errors in parentheses. Year dummies included in all specifications.

*** p<0.01, ** p<0.05, * p<0.1

Table 2.9: CRCFEP 3 estimated elasticities

Parameter	Estimate	S.E.	P-value	95% C.I.	
$\widehat{\beta}_0$	0.134	0.093	0.149	-0.048	0.315
$\widehat{\beta}_{-1}$	0.051	0.057	0.379	-0.062	0.163
$\widehat{\beta}_{-2}$	0.257	0.098	0.009	0.064	0.449
$\widehat{\beta}_{-3}$	-0.023	0.092	0.800	-0.205	0.158
$\widehat{\beta}_0+\widehat{\beta}_{-1}+\widehat{\beta}_{-2}+\widehat{\beta}_{-3}$	0.417	0.127	0.001	0.169	0.666

$\widehat{\beta}_{-\tau}=\widehat{\alpha}_{\tau+1}+\widehat{\gamma}_{\tau+1}\overline{\log(R)}$. Clustered S.E.'s ignore sampling error of $\overline{\log(R)}$

with the time average. Therefore, while there is marginal evidence of heterogeneity, I cannot parse it into its components.

Focusing on model (3), therefore, the results are quite interesting, at least at face value. The estimator for the average elasticity with respect to R_{t-s} is given by

$$\widehat{\beta}_{-s}=\widehat{\alpha}_{s+1}+\widehat{\gamma}_{s+1}\overline{\log(R)}, \tag{2.63}$$

where $\overline{\log(R)}=(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\log(R_{it})$. I give these estimates in Table 2.9.

This implied lag profile for the average elasticity is different from that previously observed in the literature, where typically the contemporaneous elasticity accounts for most of the total and the lags are much smaller in magnitude and often statistically insignificant. Model (3) estimates imply, however, that the highest estimated average elasticity is with respect to the second lag of $log(R)$, at 0.26 with a standard error of 0.098. Meanwhile, the contemporaneous and other lags are insignificantly different from zero. At face value, this seems to imply a delay in the benefit to R&D expenditures. Furthermore, the negative estimated coefficient on $\log(R_{-2})\times\overline{\log(R_0)}$ implies that the firms with larger R&D expenditures overall experience lower marginal returns. The correlation between $\overline{\log(R_0)}$ and the estimate of the multiplicative firm effect is 0.39, indicating that firms with a higher base rate of patenting tend to have lower marginal returns to R&D dollars, which echoes the findings of Wang, et. al. (1998) with regards to the pharmaceutical industry. Unfortunately, however, the results do not appear to be robust to changes in the estimation sample. If I construct a panel over 1994-2001, for instance, neither the lag-structure result or the finding of heterogeneous

slopes hold. It may be that there is still a sample selection problem caused by not observing any patent applications made through 2003 if the were not granted before 2006.

2.6 Conclusion

FEP analysis of count or other nonnegative response variables cannot generally be justified in the presence of heterogeneous slopes and may not lead to estimation of any quantity of interest. Given this, I extend Chesher's (1984) testing framework to the FEP setting and show that an LM test for neglected heterogeneity amounts to adding squares of regressors to the set of covariates. This procedure is more widely applicable than classical tests. Simulation evidence also suggests robustness to this approach when coefficients are neither normal nor belong to a location-scale family.

Identification via a correlated random coefficients assumption leads to FEP on a more flexible mean function as an estimation method. Under a proportional variance assumption and CRE assumption for the scalar, multiplicative effect, normal QMLE is another technique which may have advantages in cases of binary or time-constant regressors. Each of these options feasibly allows for higher dimensional random coefficients than estimators based on likelihoods with integrals, while also allowing for dependence between the heterogeneity and the regressors.

Application of these methods to the U.S. manufacturing industry may indicate firms may have heterogeneous elasticities of patenting with respect to R&D, and that in contrast to previous results, there may be a delay in the effect of R&D expenditures on patenting. results do not hold when estimating over different years of data. One immediate avenue for future research is to extend this type of correlated random coefficients model to cases where the regressors are not strictly exogenous, either because of feedback, contemporaneous endogeneity, or sample selection, as a way to explore robustness of these findings.

CHAPTER 3

ESTIMATION OF AVERAGE MARGINAL EFFECTS IN MULTIPLICATIVE UNOBSERVED EFFECTS PANEL MODELS

3.1 Introduction and Review

Nonlinear models often make logical sense for representing limited dependent variables like discrete choices and counts. Challenges can arise, however, in micro-econometric panel settings when one wishes to control for unobserved individual heterogeneity and has relatively few time periods of data. For static multiplicative effects models with strictly exogenous covariates, fixed effects Poisson (FEP) consistently estimates the parameters of a correctly-specified conditional mean function (Wooldridge, 1999). Researchers may also want to estimate quantities like Average Partial Effects (APE) and Average Treatment Effects (ATE), but as they depend on the unobserved heterogeneity, it is not immediately clear how to proceed.

I study an approach that estimates APE and ATE by combining FEP parameter estimates with estimates of the individual heterogeneity. The latter come from unconditional Poisson QMLE treating the heterogeneity as parameters to be estimated, a procedure that yields estimates of the conditional mean function parameters that are algebraically equivalent to FEP.[1] While easy to implement, such APE and ATE estimates potentially suffer from the incidental parameters problem (IPP) since the individual effect estimates are based on only T observations (Lancaster, 2000). However, I show that in multiplicative models, such APE and ATE estimators are consistent and asymptotically normal with only the cross-sectional dimension growing. The consistency result may not be surprising, but it is not implied by consistency of FEP for slope coefficients, and similar results do not hold for other nonlinear models. For instance, the IPP still biases APE estimates in fixed effects binary response models even if one knows the true values of the slope parameters or can estimate them consistently (Fernandez-Val, 2009).

[1] This result was derived independently by Lancaster (2002) and a version of Blundell, et. al. (2002).

To my knowledge, estimating APE and ATE using estimated incidental parameters has not been studied in multiplicative models specifically. Many authors have studied consistent slope parameter and marginal effect estimation using estimated incidental parameters in either general nonlinear models or in other specific settings. One solution is to employ bias corrections that are justified by large-T asymptotics. See, for example, Hahn and Newey (2004) for general nonlinear models estimated with unconditional MLE, or Fernandez-Val (2009) for the unobserved effects probit model. Although allowed to be much smaller than the number of individuals, the number of time periods needs to be sufficiently large for the asymptotic approximation of the bias to perform well. For static probit and logit models, Fernandez-Val, Greene (2004) and others have noted a "small bias" property for APE and ATE estimates from unconditional MLE . The multiplicative case, however, is special in that the average marginal effects estimators are actually consistent with only the cross-section size growing, a rare result outside of the linear model. This means they should perform well even with only two time periods.

Empirical researchers, of course, also have the option to focus on quantities that do not depend on unobserved heterogeneity. For instance, the exponential conditional mean function with a linear index gives the slope coefficients interpretations as semi-elasticities, and proportional treatment effects are also identified (M. Lee and Kobayashi, 2001). Another possibility is to make additional assumptions. For example, one could use a correlated random effects (CRE) approach by assuming a parametric form for the mean of the heterogeneity conditional on the explanatory variables. This is applicable in many nonlinear settings to estimate slope parameters as well as average partial effects (Wooldridge, 2010). Using estimated heterogeneity, however, avoids additional restrictions and allows the researcher to estimate average marginal effects in levels, which may be more meaningful than slope parameters and allows comparisons across models.

The rest of this chapter is organized as follows: Section 2 describes the multiplicative model and derives the asymptotic properties of the APE and ATE estimators that use estimated heterogeneity. I also discuss some interesting implications of using these estimators in exponential models. Section 3 evaluates the proposed estimators via Monte Carlo, and Section 4 concludes. Simu-

lation tables are collected in Appendix D.

3.2 Theory

The multiplicative unobserved effects panel model assumes that for $i = 1,\ldots,N$; $T = 1,\ldots,T$,

$$E(y_{it}|\boldsymbol{x}_i, c_i) = E(y_{it}|\boldsymbol{x}_{it}, c_i) = c_i m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0), \tag{3.1}$$

where $m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)$ is a known, positive, continuous, differentiable function of a $1 \times K$ vector of explanatory variables $\boldsymbol{x}_{it}$ and an unknown $K \times 1$ parameter vector $\boldsymbol{\beta}_0$. The term c_i is unobserved heterogeneity that is assumed to be strictly positive. Equation (3.1) implicitly assumes that $\boldsymbol{x}_{it}$ is strictly exogenous, conditional on c_i. I assume that the vector $\{y_{i1},\ldots,y_{iT},\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{iT},c_i\}$ is independent and identically distributed across i, and that T is fixed.

A common choice in the empirical literature is $m(\boldsymbol{x}_{it}, \boldsymbol{\beta}) = \exp(\boldsymbol{x}_{it}\boldsymbol{\beta})$, but other forms are possible, and the responses need not even be counts. For example, under the restriction that $0 < c_i < 1$, y_{it} could be binary or fractional, in which case $m(\boldsymbol{x}_{it}, \boldsymbol{\beta})$ might be the logistic or normal cumulative distribution function. Another option for nonnegative responses is a panel version of Wooldridge's (1992) alternative to the Box-Cox transformation. In this case, with $\boldsymbol{\beta} = (\boldsymbol{\theta}', \lambda)'$, the specification would be:

$$m(\boldsymbol{x}_{it}, \boldsymbol{\beta}) = \begin{cases} [1 + \lambda \boldsymbol{x}_{it}\boldsymbol{\theta}]^{1/\lambda}, & \lambda \neq 0 \\ \exp(\boldsymbol{x}_{it}\boldsymbol{\theta}), & \lambda = 0. \end{cases} \tag{3.2}$$

The parameters are perhaps less interesting in these examples than in the exponential case, motivating the estimation of marginal effects. While most of the derivations in this section are for a generic $m(\boldsymbol{x}_{it}, \boldsymbol{\beta})$, I include a discussion of the exponential case at the end of this section.

Hausman, Hall, and Griliches (1984) showed that if conditional on $\boldsymbol{x}_i = \{\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{iT}\}$ and c_i, the y_{it} are independently distributed as Poisson with mean given by (3.1), then conditioning on $n_i \equiv \sum_{t=1}^{T} y_{it}$ results in the multinomial distribution for $\{y_{i1},\ldots,y_{iT}\}$. The resulting fixed effects

Poisson (FEP) estimator is given by:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmax} \sum_{i=1}^{N} \ell_i(\boldsymbol{\beta}), \tag{3.3}$$

$$\ell_i(\boldsymbol{\beta}) = \sum_{t=1}^{T} y_{it} \log \left[\frac{m(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\sum_{r=1}^{T} m(\boldsymbol{x}_{ir}, \boldsymbol{\beta})} \right]. \tag{3.4}$$

Wooldridge (1999) showed that $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$ under (3.1) only, making it a quasi conditional maximum likelihood estimator (QCMLE). Standard asymptotic theory for M-estimators yields that under regularity conditions:

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{A}_0^{-1} \boldsymbol{B}_0 \boldsymbol{A}_0^{-1}), \tag{3.5}$$

where $\boldsymbol{A}_0 = -E\left[\nabla_{\beta}^2 \ell_i(\boldsymbol{\beta}_0)\right]$, $\boldsymbol{B}_0 = Var[\boldsymbol{s}_i(\boldsymbol{\beta}_0)]$, and $\boldsymbol{s}_i(\boldsymbol{\beta}_0) = \nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta}_0)'$. The sandwich form of the asymptotic variance estimator should be used to account for the fact that without the stronger assumptions of Hausman, et. al., $\ell_i(\boldsymbol{\beta})$ is not the true log-likelihood for individual i.

Researchers are often interested in estimating marginal effects, as the β_j may not have an meaningful interpretation outside of the exponential case. I define the APE of a continuous variable x_j as:

$$\delta_{j,0} = E\left[\frac{\partial E(y_{it}|\boldsymbol{x}_{it}, c_i)}{\partial x_{it\,j}}\right] = E\left[c_i T^{-1} \sum_{t=1}^{T} \frac{\partial m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)}{\partial x_{it\,j}}\right] \equiv E\left[c_i T^{-1} \sum_{t=1}^{T} M_j(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)\right], \tag{3.6}$$

where $M_j(\boldsymbol{x}_{it}, \boldsymbol{\beta}) = \frac{\partial m(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial x_{it\,j}}$. I define the ATE for a binary x_k as:

$$\begin{aligned} \delta_{k,0} =& E\left[E(y_{it}|\boldsymbol{x}_{it(-k)}, x_{itk} = 1, c_i) - E(y_{it}|\boldsymbol{x}_{it(-k)}, x_{itk} = 0, c_i)\right] \\ =& E\left[c_i T^{-1} \sum_{t=1}^{T} \Big(m(\boldsymbol{x}_{it(-k)}, 1, \boldsymbol{\beta}_0) - m(\boldsymbol{x}_{it(-k)}, 0, \boldsymbol{\beta}_0)\Big)\right] \end{aligned} \tag{3.7}$$

where the subscript $(-k)$ indicates element k has been omitted, and where $m(\boldsymbol{x}_{it(-k)}, 1, \boldsymbol{\beta})$ and $m(\boldsymbol{x}_{it(-k)}, 1, \boldsymbol{\beta})$ correspond to a 1 or 0 being inserted for x_{itk} in $m(\boldsymbol{x}_{it}, \boldsymbol{\beta})$.

Both of these quantities depend on c_i, and so an additional assumption (i.e. correlated random effects) would seem necessary to proceed. However, unconditional QMLE that treats the c_i as

additional parameters offers algebraically equivalent estimates of $\boldsymbol{\beta}_0$ as FEP, as well as a closed-form estimate of c_i. The formula is:

$$c(\boldsymbol{w}_i, \widehat{\boldsymbol{\beta}}) = \frac{\sum_{t=1}^{T} y_{it}}{\sum_{t=1}^{T} m(\boldsymbol{x}_{it}, \widehat{\boldsymbol{\beta}})} \equiv \widehat{c}_i \tag{3.8}$$

where $\boldsymbol{w}_i \equiv \{y_{i1}, \ldots, y_{iT}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}\}$.

The analysis to follow hinges on studying the properties of this random function of the data, which I rewrite for a generic $\boldsymbol{\beta}$ as:

$$c(\boldsymbol{w}_i, \boldsymbol{\beta}) \equiv \frac{\sum_{t=1}^{T} y_{it}}{\sum_{t=1}^{T} m(\boldsymbol{x}_{it}, \boldsymbol{\beta})} \tag{3.9}$$

There is a practical reason to estimate $\boldsymbol{\beta}_0$ using FEP instead of unconditional QMLE (i.e. including N individual dummies in the exponential model). As pointed out by Cameron and Trivedi (2013), the econometrician may encounter computational or software limitations for large values of N. It is easier to just calculate $\widehat{c}_i$ following FEP estimation. The APE and ATE estimators I investigate are:

$$\widehat{\delta}_j = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{c}_i M_j(\boldsymbol{x}_{it}, \widehat{\boldsymbol{\beta}}) \tag{3.10}$$

$$\widehat{\delta}_k = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{c}_i \left[m(\boldsymbol{x}_{it(-k)}, 1, \widehat{\boldsymbol{\beta}}) - m(\boldsymbol{x}_{it(-k)}, 0, \widehat{\boldsymbol{\beta}}) \right] \tag{3.11}$$

Clearly $c(\boldsymbol{w}_i, \boldsymbol{\beta}) \neq c_i$, even if evaluated at $\boldsymbol{\beta}_0$, and with only N growing, $\widehat{c}_i$ cannot be consistent for c_i (under the view that c_i is one of N individual-specific parameters).[2] One should not generally expect marginal effects calculated from estimated incidental parameters to be consistent in nonlinear models, even if slope parameter estimates of are consistent. However, some sample averages involving $\widehat{c}_i$ are consistent in the FEP case due to the form of $c(\boldsymbol{w}_i, \boldsymbol{\beta})$ and the fact that c_i and $m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)$ are multiplicatively separable.

Theorem 1 *Suppose* $\widehat{\boldsymbol{\lambda}} \equiv N^{-1} \sum_{i=1}^{N} c(\boldsymbol{w}_i, \widehat{\boldsymbol{\beta}}) \boldsymbol{h}(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}})$ *is an estimator of* $\boldsymbol{\lambda}_0 \equiv E[c_i \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)]$. *Assume that* (3.1) *holds and that each element of the* $P \times 1$ *random vector* $\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta}) \equiv c(\boldsymbol{w}_i, \boldsymbol{\beta}) \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta})$

[2] Cameron and Trivedi (2013) assert $\widehat{c}_i \xrightarrow{p} c_i$ as $T \to \infty$, which is true if $\{y_{it}\}$ and $\{m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)\}$ are ergodic for the mean.

satisfies the regularity conditions on $q(\boldsymbol{w}_i, \boldsymbol{\beta})$ from Theorem 12.2 of Wooldridge (2010). Then

$$\widehat{\boldsymbol{\lambda}} \xrightarrow{p} \boldsymbol{\lambda}_0$$

Proof. Since $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$, then $N^{-1}\sum_{i=1}^{N} c(\boldsymbol{w}_i, \widehat{\boldsymbol{\beta}})\boldsymbol{h}(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) \xrightarrow{p} E\left[c(\boldsymbol{w}_i, \boldsymbol{\beta}_0)\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\right]$ by Lemma 12.1 in Wooldridge (2010). Furthermore, by the L.I.E.,

$$\begin{aligned} E\left[c(\boldsymbol{w}_i, \boldsymbol{\beta}_0)\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\right] =& E\left\{E\left[c(\boldsymbol{w}_i, \boldsymbol{\beta}_0)\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)|\boldsymbol{x}_i, c_i\right]\right\} \\ =& E\left[\frac{\sum_{t=1}^{T} E(y_{it}|\boldsymbol{x}_i, c_i)}{\sum_{t=1}^{T} m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)}\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\right] \\ =& E\left[\frac{c_i \sum_{t=1}^{T} m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)}{\sum_{t=1}^{T} m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)}\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\right] \\ =& E\left[c_i \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\right] \end{aligned} \tag{3.12}$$

■

Consistency of $N^{-1}\sum_{i=1}^{N} \widehat{c}_i$ for $E(c_i)$ follows from setting $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}) = 1$, while consistency of $\widehat{\delta}_j$ and $\widehat{\delta}_k$ follow from either setting

$$\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}) = T^{-1}\sum_{t=1}^{T} M_j(\boldsymbol{x}_{it}, \boldsymbol{\beta}) \text{ or} \tag{3.13}$$

$$\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}) = T^{-1}\sum_{t=1}^{T} \left[m(\boldsymbol{x}_{it(-k)}, 1, \boldsymbol{\beta}) - m(\boldsymbol{x}_{it(-k)}, 0, \boldsymbol{\beta})\right]. \tag{3.14}$$

Theorem (1) shows that unlike with other nonlinear fixed effects estimators, no bias correction is necessary to estimate the APE and ATE in this setting. One might expect, a priori, that $\widehat{\delta}_j$ and $\widehat{\delta}_k$ would perform well anyway as T grows and $\widehat{c}_i$ better approximates c_i. Nevertheless, Theorem (1) holds for an arbitrary T, so $\widehat{\delta}_j$ and $\widehat{\delta}_k$ should perform well even in panels with only two time periods (the minimum needed for FEP). The APE and ATE I consider are just two of many possible quantities of interest. Researchers might also want to know the average marginal effect for a specific time period, or for a specific subpopulation defined by the observables (i.e. the Average Treatment Effect on the Treated). One might also want to estimate the partial effect evaluated at the averages of the heterogeneity and covariates. As long as $\widehat{c}_i$ multiplies the relevant function

of the data, one need not worry about the difference between it and c_i when averaging over the cross-section.

As a caution, one cannot use $\widehat{c}_i$ to learn about other features of the distribution of c_i except in more restrictive cases. For instance, $Var(c_i)$ is identified only under additional assumptions about $Var(\boldsymbol{y}_i|\boldsymbol{x}_i, c_i)$. A simple example is when the Poisson variance assumption, $Var(y_{it}|\boldsymbol{x}_i, c_i) = E(y_{it}|\boldsymbol{x}_i, c_i)$, and zero conditional covariance, $Cov(y_{it}, y_{ir}|\boldsymbol{x}_i, c_i) = 0, t \neq r$, both hold. In this case, one can show that $Var(c_i) = Var\left[c(\boldsymbol{w}_i, \boldsymbol{\beta}_0)\right] - E\left[c_i / \sum_{t=1}^{T} m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)\right]$.

The asymptotic variance of $\widehat{\boldsymbol{\lambda}}$ can be derived similarly to the delta method, but making sure to account for the randomness in $\boldsymbol{w}_i$. [3]

Theorem 2 *Under the assumptions in Theorem* (1)*,*

$$\sqrt{N}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{D}_0),$$

where

$$\boldsymbol{D}_0 = Var\left[\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta}_0) - \boldsymbol{\lambda}_0 - \boldsymbol{G}_0 \boldsymbol{A}_0^{-1} \boldsymbol{s}_i(\boldsymbol{\beta}_0)\right],$$

$$\boldsymbol{G}_0 = E\left[\nabla_{\boldsymbol{\beta}} \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta}_0)\right] = E\left[c(\boldsymbol{w}_i, \boldsymbol{\beta}_0) \nabla_{\boldsymbol{\beta}} \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}_0) + \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta}_0) \nabla_{\boldsymbol{\beta}} c(\boldsymbol{w}_i, \boldsymbol{\beta}_0)\right],$$

$$\nabla_{\boldsymbol{\beta}} c(\boldsymbol{w}_i, \boldsymbol{\beta}) = -c(\boldsymbol{w}_i, \boldsymbol{\beta}) \left(\frac{\sum_{t=1}^{T} \nabla_{\boldsymbol{\beta}} m(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\sum_{t=1}^{T} m(\boldsymbol{x}_{it}, \boldsymbol{\beta})} \right),$$

$\nabla_{\boldsymbol{\beta}} \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta})$ *is the* $P \times K$ *Jacobian of* $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\beta})$*, and*

$\nabla_{\boldsymbol{\beta}} m(\boldsymbol{x}_{it}, \boldsymbol{\beta})$ *is the* $1 \times K$ *gradient of* $m(\boldsymbol{x}_{it}, \boldsymbol{\beta})$*.*

Proof. Define $\ddot{\boldsymbol{G}}_i$ as the $P \times K$ Jacobian of $\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta})$ evaluated at different mean values between $\widehat{\boldsymbol{\beta}}$

[3] The derivation here is essentially the same as the solution to Wooldridge (2010), Problem 12.17.

and $\boldsymbol{\beta}_0$. By a mean value expansion of each element of $\sqrt{N}\widehat{\boldsymbol{\lambda}} = N^{-1/2}\sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{w}_i, \widehat{\boldsymbol{\beta}})$ around $\boldsymbol{\beta}_0$,

$$N^{-1/2}\sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{w}_i, \widehat{\boldsymbol{\beta}}) = N^{-1/2}\sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta}_0) + \left(N^{-1}\sum_{i=1}^{N} \ddot{\boldsymbol{G}}_i\right)\sqrt{N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \tag{3.15}$$

$$= N^{-1/2}\sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta}_0) + \boldsymbol{G}_0\sqrt{N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) + o_p(1) \tag{3.16}$$

$$= N^{-1/2}\sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta}_0) - N^{-1/2}\sum_{i=1}^{N} \boldsymbol{G}_0\boldsymbol{A}_0^{-1}\boldsymbol{s}_i(\boldsymbol{\beta}_0) + o_p(1). \tag{3.17}$$

The second equality follows because consistency of $\widehat{\boldsymbol{\beta}}$ implies $N^{-1}\sum_{i=1}^{N} \ddot{\boldsymbol{G}}_i \xrightarrow{p} \boldsymbol{G}_0$ and because $\sqrt{N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = O_p(1)$. The third follows because $\sqrt{N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = -N^{-1/2}\sum_{i=1}^{N} \boldsymbol{A}_0^{-1}\boldsymbol{s}_i(\boldsymbol{\beta}_0) + o_p(1)$. Therefore,

$$\sqrt{N}\left(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\right) = N^{-1/2}\sum_{i=1}^{N}\left[\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta}_0) - \boldsymbol{\lambda}_0 - \boldsymbol{G}_0\boldsymbol{A}_0^{-1}\boldsymbol{s}_i(\boldsymbol{\beta}_0)\right] + o_p(1) \tag{3.18}$$

By the Asymptotic Equivalence Lemma, the limiting distribution of $\sqrt{N}\left(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\right)$ is the same as $N^{-1/2}\sum_{i=1}^{N}\left[\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta}_0) - \boldsymbol{\lambda}_0 - \boldsymbol{G}_0\boldsymbol{A}_0^{-1}\boldsymbol{s}_i(\boldsymbol{\beta}_0)\right]$, which is easily shown to be the scaled sample average of a mean-zero random vector. Therefore, by the Central Limit Theorem for i.i.d. sequences, the result follows. ■

Applying Theorem (2) for the APE of a continuous covariate x_j:

$$\sqrt{N}\left(\widehat{\delta}_j - \delta_{j,0}\right) \xrightarrow{d} N(0, D_{j,0}), \tag{3.19}$$

$$D_{j,0} = Var\left[T^{-1}\sum_{t=1}^{T} c(\boldsymbol{w}_i, \boldsymbol{\beta}_0)M_j(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0) - \delta_{j,0} - G_{j,0}\boldsymbol{A}_0^{-1}\boldsymbol{s}_i(\boldsymbol{\beta}_0)\right], \tag{3.20}$$

$$G_{j,0} = E\left[c(\boldsymbol{w}_i, \boldsymbol{\beta}_0)(T^{-1})\sum_{t=1}^{T}\left\{\nabla_{\boldsymbol{\beta}} M_j(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0) - M_j(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)\left(\frac{\sum_{t=1}^{T}\nabla_{\boldsymbol{\beta}} m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)}{\sum_{t=1}^{T} m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)}\right)\right\}\right] \tag{3.21}$$

For the ATE of the binary covariate x_k:

$$\sqrt{N}\left(\widehat{\delta}_k - \delta_{k,0}\right) \xrightarrow{d} N(0, D_{k,0}), \tag{3.22}$$

$$D_{k,0} = Var\left[T^{-1}\sum_{t=1}^{T} c(\boldsymbol{w}_i, \boldsymbol{\beta}_0)\left(m(\boldsymbol{x}_{it(-k)}, 1, \boldsymbol{\beta}_0) - m(\boldsymbol{x}_{it(-k)}, 0, \boldsymbol{\beta}_0\right) - \delta_{k,0} - G_{k,0}\boldsymbol{A}_0^{-1}\boldsymbol{s}_i(\boldsymbol{\beta}_0)\right], \tag{3.23}$$

$$G_{k,0} = E\left[c(\boldsymbol{w}_i, \boldsymbol{\beta}_0)(T^{-1})\sum_{t=1}^{T}\left\{\nabla_{\boldsymbol{\beta}} m_{it}(1) - \nabla_{\boldsymbol{\beta}} m_{it}(0) - (m_{it}(1) - m_{it}(0))\left(\frac{\sum_{t=1}^{T}\nabla_{\boldsymbol{\beta}} m_{it}}{\sum_{t=1}^{T} m_{it}}\right)\right\}\right], \tag{3.24}$$

where $m_{it} = m(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)$, $m_{it}(1) = m(\boldsymbol{x}_{it(-k)}, 1, \boldsymbol{\beta}_0)$, and $m_{it}(0) = m(\boldsymbol{x}_{it(-k)}, 0, \boldsymbol{\beta}_0)$. These asymptotic variances can be consistently estimated from the above expressions by plugging in $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ and forming the sample analogs to the expectation and variance operators.

3.2.1 Exponential Models

Since it is a common specification in empirical research, I include a few observations about the exponential conditional mean case. The form of the quasi log-likelihood means that one can estimate coefficients on time-varying $\boldsymbol{x}_{it}$ only. Nevertheless, $\delta_{j,0}$ and $\delta_{k,0}$ are still identified when the conditional mean function is exponential and includes time-constant observables. To see this, suppose the following:

$$E(y_{it}|\boldsymbol{x}_{it}, \boldsymbol{z}_i, v_i) = v_i \exp(\boldsymbol{x}_{it}\boldsymbol{\beta}_0 + \boldsymbol{z}_i\boldsymbol{\gamma}_0), \tag{3.25}$$

where now I use v_i to denote the unobserved heterogeneity. Define $c_i = v_i \exp(\boldsymbol{z}_i\boldsymbol{\gamma}_0)$. Then clearly

$$E(y_{it}|\boldsymbol{x}_{it}, c_i) = c_i \exp(\boldsymbol{x}_{it}\boldsymbol{\beta}_0). \tag{3.26}$$

The heterogeneity has absorbed the time-constant observables. Theorems (1) and (2) still hold, but the function $\widehat{c}_i$ now serves as a stand-in for the total contribution from all time-constant variables—observed and unobserved. Analogous to the linear case, $\boldsymbol{\gamma}_0$ is not identified, nor are the average partial effects of the $\boldsymbol{z}_i$, but given consistent estimates of $\boldsymbol{\beta}_0$, one can still consistently estimate the average partial effects of the time-varying regressors.

One alternative estimand studied by Lee and Kobayashi (2001) is the proportional treatment effect, which for a binary treatment and the simple index in (3.25) is: [4]

$$\xi_k \equiv \frac{E(y_{it}, \boldsymbol{x}_{it(-k)}, x_{itk} = 1, \boldsymbol{z}_i, v_i)}{E(y_{it}, \boldsymbol{x}_{it(-k)}, x_{itk} = 0, \boldsymbol{z}_i, v_i)} - 1 = \exp(\beta_k) - 1 \tag{3.27}$$

Of course, ξ_k may interesting in its own right, but my analysis shows that estimating the ATE in levels using (3.11) is another option, even when time-constant regressors belong in the model.

Furthermore, APE of a continuous variable simplifies in the exponential conditional mean case.

$$\delta_{j,0} = E\left[T^{-1}\sum_{t=1}^{T} c_i \exp(\boldsymbol{x}_{it}\boldsymbol{\beta})\right]\beta_{j,0} \tag{3.28}$$

$$= E\left[T^{-1}\sum_{t=1}^{T} E(y_{it}|\boldsymbol{x}_{it}, c_i)\right]\beta_{j,0} \tag{3.29}$$

$$= \left[T^{-1}\sum_{t=1}^{T} E(y_{it})\right]\beta_{j,0}, \tag{3.30}$$

where the last equality is by the L.I.E. Here, the population scale factor is analogous to the cross-section case and doesn't depend on the heterogeneity.

Moreover, an estimator that treats $\widehat{c}_i$ as the unknown c_i is equivalent to the sample analog of (3.30).

$$\widehat{\delta}_j = \left[(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} \widehat{c}_i \exp(\boldsymbol{x}_{it}\widehat{\boldsymbol{\beta}})\right]\widehat{\beta}_j = \left[(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}\right]\widehat{\beta}_j \tag{3.31}$$

Consistency of $\widehat{\delta}_j$ for δ_j is immediate given a consistent estimator of $\beta_{j,0}$. Since $\widehat{\delta}_j$ does not depend on $\widehat{c}_i$, one could even estimate $\boldsymbol{\beta}_0$ without assuming strict exogeneity of $\boldsymbol{x}_{it}$, using the GMM approach of either Chamberlain (1992) or Wooldridge (1997) based on sequential moment restrictions.

The asymptotic variance is simpler as well:

$$Avar\left[\sqrt{N}(\widehat{\delta}_j - \delta_{j,0})\right] = Var(\bar{y}_i\beta_j - \delta_j - \mu_y^T r_j A_0^{-1}\boldsymbol{s}_i(\boldsymbol{\beta}_0)), \tag{3.32}$$

[4] Lee and Kobayashi's model includes multi-valued treatment as well as interactions between the treatment and covariates, so the proportional treatment effect depends on $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_i$, but only involves coefficients on time-varying regressors and interactions.

where $\mu_y^T \equiv E(T^{-1}\sum_{t=1}^{T} y_{it})$, and r_j is a $1 \times K$-vector with jth element equal to 1 and all other elements equal to 0. The expression is similar if GMM is used to estimate $\boldsymbol{\beta}_0$.

3.2.2 A note about dropped observations

If the dependent variable for an observation l is zero in each time period, then observation l contributes nothing to the quasi log-likelihood, as can be seen in equation (3.4). Clearly, the terms in $\widehat{\delta}_j$ and $\widehat{\delta}_k$ corresponding to observation l's contribution are then equal to zero, since $c(\boldsymbol{w}_l, \boldsymbol{\beta}) = 0$. Nevertheless, if interested in an APE or ATE with respect to the entire population of interest, the sample size N in the formulas for $\widehat{\delta}_j$ and $\widehat{\delta}_k$ should correspond to the number of individuals in the entire the cross-section, not the number of individuals in the estimation sample (that is, with $n_i > 0$). Otherwise, the estimates will be conditional on this particular subsection of the population and be inflated by a factor of N/N_p, where $N_p = \sum_{i=1}^{N} \mathbf{1}[n_i > 0]$.

3.3 Monte Carlo

3.3.1 Design

I employ the following data generating process. For $i = 1, \ldots, N$ and $t = 1, \ldots, T$:

$$y_{it} | (\boldsymbol{x}_i, \boldsymbol{d}_i, c_i) \sim \text{Poisson}\left[c_i \exp(\beta_1 x_{it} + \beta_2 d_{it})\right], \tag{3.33}$$

$$\log(c_i) \sim Normal(0, \sigma^2) \tag{3.34}$$

$$x_{it} = \log(c_i) + \rho x_{i,t-1} + v_{it}, t > 1 \tag{3.35}$$

$$x_{i1} = \log(c_i)/(1-\rho) + v_{i1}/\sqrt{1-\rho^2}, v_{it} \sim N(0, 1/2), \tag{3.36}$$

$$\rho = 0.3 - 0.5\sigma \tag{3.37}$$

$$d_{it} = \mathbf{1}[x_{it} + \log(c_i) + h_{it} > 0], h_{it} \sim N(0, 1/2) \tag{3.38}$$

I study panels of dimensions $N \in \{500, 1000, 2000\}$ and $T \in \{2, 4, 10\}$. The conditional marginal distribution of y_t is Poisson with an exponential mean function. I set $\beta_1 = 0.5$ and $\beta_2 = -0.5$. I vary the degree of heterogeneity, with $\sigma \in \{0, 0.25.0.5, 0.75, 1\}$. The continuous covariate x_t and the binary covariate d_t are both correlated with the heterogeneity, and the strength of the correlation increases with σ. The scaling of x_{i1} is intended to keep $Var(x_t)$ constant across the different T.[5] That the autoregressive parameter in the equation for x_t depends on σ is an attempt to keep the autocovariance structure of x_t more consistent as σ increases.

I estimate β_1 and β_2 using FEP, and employ the APE and ATE estimators proposed in equations (3.10) and (3.11). In the tables to follow, FEP estimates are denoted with a "$\widehat{\ }$". For reference, I also estimate the slopes, APE, and ATE using pooled Poisson QMLE, which ignores c_i entirely. These estimates are denoted with a "$\widetilde{\ }$". Both FEP and Poisson QMLE are consistent when $\sigma = 0$, but only FEP is consistent when $\sigma > 0$. Reporting the results for Poisson QMLE is intended to give the reader a sense of how large a problem neglected heterogeneity causes under this particular DGP. For each estimator and parameter combination, I report the mean and standard deviation of the empirical distribution, the estimated bias, the ratio of the mean standard error to the empirical standard deviation, and the probability of rejecting a true null hypothesis at the 5 percent significance level. I use cluster robust asymptotic standard errors with the slope estimates, though they are technically not necessary with this DGP. For the APE and ATE estimates, I use the "unconditional" asymptotic standard errors derived in this chapter for the FEP case, as well as the analogous versions for Poisson QMLE. For each parameter combination, I draw 2000 replications.

3.3.2 Results

Full tables of simulation results can be found in Appendix D. I focus attention on the APE and ATE estimates, though the slope estimates are included for reference. As expected, across all values of N and T, there is virtually no finite sample bias in the Poisson QMLE and the FEP estimates in the absence of heterogeneity ($\sigma = 0$). In the presence of heterogeneity, however, Poisson QMLE slopes

[5] See Vamos, Soltuz, and Craciun 2007.

and APEs are biased. As heterogeneity increases, bias increases, and the probability of rejecting a true null hypothesis quickly approaches one. Therefore, this DGP succeeds in simulating settings where controlling for individual effects is important.

Finite sample bias in $\widehat{\delta}_1$ is less than 0.01 for all values of N, T, and σ, which is not surprising given that in the exponential case, the APE scale factor does not even depend on c_i. Some ATE estimates at higher levels of σ are slightly biased away from zero when the panel is short and the sample is smaller. For instance, when $N = 500$ and $T = 2$, finite sample bias is between 2 and 2.5 percent of the true value when $\sigma \geq 0.5$. However, the magnitudes of these biases decrease to $1 - 1.5$ percent when $N = 1000$ and $0 - 1$ percent when $N = 2000$. In the $T = 4$ and $T = 10$ cases, the finite sample bias is less than 1 percent and quite small in the larger cross-sections.

The finite sample standard deviations behave in predictable ways, decreasing as either N or T increases. The variability in $\widehat{\delta}_2$ seems to be greater than that of $\widehat{\delta}_1$, and the spread between them increases with σ, which might be related to the fact that $\widehat{\delta}_1$ does not actually use $\widehat{c_i}$ in the exponential case. The standard errors derived in this chapter perform reasonably well, particularly with the largest cross-section, where at worst, their empirical mean underestimates the empirical standard deviation by about 4 percent. This occurs for the standard error of $\widehat{\delta}_1$ in the $T = 4, \sigma = 1$ case, where as a point of comparison, the mean standard error for $\widehat{\beta}_1$ also underestimates the finite sample standard deviation of $\widehat{\beta}_1$ by a similar amount. For the most part, the results suggest the approximations get better as N increases, though simulating more replications may be necessary to reduce sampling error. When σ is high, the apparent underestimation by the standard errors leads to slight over-rejection by about one or two percentage points, but larger N also mitigates this problem.

Overall, these simulations support this chapter's theoretical findings. The asymptotic properties derived in Section 2 for the APE and ATE estimators that use estimated incidental parameters seem to approximate their finite sample behavior very well.

3.4 Conclusion

It is already well-known that in static multiplicative panel models under strict exogeneity, estimating the heterogeneity still leads to consistent estimation of the parameters of a correctly-specified conditional mean function. This chapter adds the result that APE and ATE estimators that use estimated heterogeneity are also consistent and $\sqrt{N}$-asymptotically normal with T fixed. In fact, the results hold for estimating the mean of a wider class of random quantities where the heterogeneity is multiplicatively separable from functions of the data. I derive asymptotic standard errors for these estimators that perform well in simulations for a leading case in empirical research. One area for future research would be to use higher order expansions to derive standard errors that better approximate the standard deviation of the sampling distribution.

APPENDICES

APPENDIX A

ANALYTICAL BIAS CORRECTION EXPRESSIONS FROM CHAPTER 1

From Hahn and Newey (2004), and Fernandez-Val (2009), the one-step bias corrected estimator is formed as

$$\widetilde{\theta}_{bc} = \widehat{\theta} - \mathscr{B}(\widehat{\theta})/T, \tag{A.1}$$

where $\widehat{\mathscr{B}}(\theta) = \widehat{\mathscr{I}}(\theta)^{-1}\widehat{b}(\theta)$. Here θ denotes a generic coefficient vector, and $\widehat{\theta}$ is the uncorrected MLE.

A.1 Hahn and Newey's bias correction for M-estimators

With strictly exogenous regressors $\boldsymbol{x}_{it}$:

$$\widehat{\mathscr{I}}(\theta) = -\left\{ (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} [\widehat{u}_{it\theta}(\theta) - \widehat{u}_{it\alpha}(\theta)] \left(\sum_{t=1}^{T} \widehat{v}_{it\theta}(\theta) \right) / \left(\sum_{t=1}^{T} \widehat{v}_{it\alpha}(\theta) \right) \right\} \tag{A.2}$$

$$\widehat{b}(\theta) = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \widehat{u}_{it\alpha}(\theta) \left[\widehat{\beta}_i(\theta) + \widehat{\psi}_{it}(\theta) \right] + \widehat{u}_{it\alpha\alpha} \widehat{\sigma}_i^2(\theta)/2 \right\}, \tag{A.3}$$

where

$$\widehat{\beta}_i(\theta) = -\left(\sum_{s=1}^{T} \widehat{v}_{is\alpha}(\theta) \right)^{-1} \sum_{s=1}^{T} \left\{ \widehat{v}_{is\alpha}(\theta) \widehat{\psi}_{it}(\theta) + \widehat{v}_{is\alpha\alpha}(\theta) \widehat{\sigma}_i^2(\theta)/2 \right\}, \tag{A.4}$$

$$\widehat{\sigma}_i^2(\theta) = T^{-1} \sum_{s=1}^{T} \widehat{\psi}_{it}(\theta)^2. \tag{A.5}$$

In these expressions, $\widehat{u}_{it}(\theta)$ and $\widehat{v}_{it}(\theta)$ are derivatives of the log-likelihood with respect to θ and α_i, respectively, evaluated at $\alpha_i = \widehat{\alpha}_i(\theta) = \arg\max_{\alpha} \sum_{t=1}^{T} \ell_{it}(\theta, \alpha_i)/T$. Partial derivatives of $\widehat{u}_{it}(\theta)$ and $\widehat{v}_{it}(\theta)$ are denoted by the θ and α subscripts. The terms $\widehat{\psi}_{it}(\theta)$, $\widehat{\sigma}_i^2(\theta)$, $\widehat{\beta}_i(\theta)$ are estimators for the influence function, asymptotic variance, and higher order asymptotic bias, respectively, of $\widehat{\alpha}_i(\theta)$ as T grows.

A.2 Fernandez-Val's bias correction based on conditional expectations

Fernandez-Val (2009) simplifies the Hahn and Newey (2004) corrections by taking expectations conditional on $\{\boldsymbol{x}_i, \alpha_i\}$ and using the Law of Iterated Expectations. For static probit models with strictly exogenous regressors,

$$\widehat{\mathscr{I}}(\theta) = N^{-1}\sum_{i=1}^{N}\left[\left(T^{-1}\sum_{t=1}^{T}\widehat{G}_{it}(\theta)\boldsymbol{x}_{it}'\boldsymbol{x}_{it}\right) - \left(T^{-1}\sum_{t=1}^{T}\widehat{G}_{it}(\theta)\boldsymbol{x}_{it}'\right)\left(T^{-1}\sum_{t=1}^{T}\widehat{G}_{it}(\theta)\boldsymbol{x}_{it}\right)\widehat{\sigma}_i^2\right] \quad \text{(A.6)}$$

$$\widehat{b}(\theta) = N^{-1}\sum_{i=1}^{N}\left\{\left(-T^{-1}\sum_{t=1}^{T}\widehat{G}_{it}(\theta)\boldsymbol{x}_{it}'\right)\widehat{\eta}_i(\theta) + \left(T^{-1}\sum_{t=1}^{T}\widehat{G}_{it}(\theta)\widehat{\lambda}_i(\theta)\boldsymbol{x}_{it}'\right)\widehat{\sigma}_i^2/2\right\},$$

where

$$\widehat{G}_{it}(\theta) = \frac{[\phi(\widehat{\alpha}_i(\theta)+\boldsymbol{x}_{it}\theta)]^2}{\Phi(\widehat{\alpha}_i(\theta)+\boldsymbol{x}_{it}\theta)[1-\Phi(\widehat{\alpha}_i(\theta)+\boldsymbol{x}_{it}\theta)]},\ \widehat{\sigma}_i^2 = T\left(\sum_{t=1}^{T}\widehat{G}_{it}(\theta)\right)^{-1}, \quad \text{(A.7)}$$

$$\widehat{\eta}_i(\theta) = (1/2)\left(T^{-1}\sum_{t=1}^{T}\widehat{\lambda}_{it}(\theta)\widehat{G}_{it}(\theta)\right)\widehat{\sigma}_i^4, \quad \text{(A.8)}$$

$$\widehat{\lambda}_{it}(\theta) = \widehat{\alpha}_i(\theta)+\boldsymbol{x}_{it}\theta, \text{ and } \widehat{\alpha}_i(\theta) = \arg\max_{\alpha}\sum_{t=1}^{T}\ell_{it}(\theta,\alpha_i)/T \quad \text{(A.9)}$$

A.3 Average Partial Effects

As in equation (14) of Section II, we define the function $m(\beta,\gamma,\alpha,\boldsymbol{x}_{it})$ as the partial effect of w_{it} on the probability that $y_{it}=1$ for $w\in\{x,d\}$. Using one of the analytical bias-corrected slope estimators, $\widetilde{\theta}_{bc}$ and $\widetilde{\alpha}_{bc}=\widehat{\alpha}_i(\widetilde{\theta}_{bc})$, the bias-corrected estimator for the average partial effect is

$$\widetilde{\mu}_{w,bc} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}m_w(\widetilde{\beta}_{bc},\widetilde{\gamma}_{bc},\widetilde{\alpha}_{bc},\boldsymbol{x}_{it}) - \widehat{\Delta}/T. \quad \text{(A.10)}$$

Using Hahn and Newey's method:

$$\widehat{\Delta} = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \widetilde{m}_{\alpha} \widetilde{\beta}_i + (1/2) \widetilde{m}_{\alpha\alpha} \widetilde{\sigma}^2 \right\}, \tag{A.11}$$

where $\widetilde{\beta}_{it} = \widehat{\beta}_{it}(\widetilde{\theta}_{bc})$ $\widetilde{\sigma}_i^2 = \widehat{\sigma}_i^2(\widetilde{\theta}_{bc})$, and $\widetilde{m}_{\alpha}$ and $\widetilde{m}_{\alpha\alpha}$ denote partial derivatives with respect to α, evaluated at $\widetilde{\theta}_{bc}$ and $\widetilde{\alpha}_{bc}$.

Using Fernandez-Val's method:

$$\widehat{\Delta} = N^{-1} \sum_{i=1}^{N} \left\{ \left(T^{-1} \sum_{t=1}^{T} \widetilde{m}_{\alpha} \widetilde{\eta}_i \right) + (1/2) \left(T^{-1} \sum_{t=1}^{T} \widetilde{m}_{\alpha\alpha} \right) \left(T^{-1} \sum_{t=1}^{T} \widetilde{G}_{it} \right)^{-1} \right\} \tag{A.12}$$

where $\widetilde{\lambda}_{it} = \widehat{\lambda}_{it}(\widetilde{\theta}_{bc})$, $\widetilde{\eta}_i = \widehat{\eta}_i(\widetilde{\theta}_{bc})$ and $\widetilde{G}_{it} = \widehat{G}_{it}(\widetilde{\theta}_{bc})$.

APPENDIX B

SIMULATION RESULTS FOR BIAS CORRECTIONS ON A LARGER CROSS-SECTION

Table B.1: Probit Slope Estimates when $N = 500$, $T = 6$

	$\widehat{\beta}$ (true value = 1)				$\widehat{\gamma}$ (true value = 1)			
	Mean	SD	cv: .95	$\frac{SE}{SD}$	Mean	SD	cv: .95	$\frac{SE}{SD}$
$\rho = 0.0$								
MLE	1.33	0.10	0.08	0.99	1.32	0.10	0.15	1.06
A-FV09	0.95	0.06	0.92	1.13	0.97	0.07	0.99	1.34
A-HN04	1.15	0.09	0.57	0.94	1.15	0.09	0.71	1.09
J-DJ14	0.92	0.13	0.61	0.53	0.94	0.13	0.79	0.72
J-HN04	0.90	0.07	0.68	0.98	0.92	0.07	0.91	1.26
CRE	0.99	0.06	0.94	0.98	0.99	0.07	0.96	1.06
$\rho = 0.4$								
MLE	1.51	0.12	0.00	0.99	1.51	0.12	0.01	1.06
A-FV09	1.02	0.06	0.98	1.17	1.05	0.07	0.99	1.41
A-HN04	1.32	0.11	0.09	0.92	1.32	0.11	0.16	1.04
J-DJ14	0.85	0.19	0.41	0.36	0.87	0.18	0.58	0.50
J-HN04	1.02	0.08	0.92	0.90	1.03	0.08	0.97	1.14
CRE	0.99	0.06	0.94	1.01	0.99	0.07	0.95	1.05
$\rho = 0.8$								
MLE	2.36	0.20	0.00	1.00	2.37	0.22	0.00	0.99
A-FV09	0.79	0.21	0.41	0.32	0.76	0.24	0.50	0.40
A-HN04	2.12	0.19	0.00	0.87	2.14	0.21	0.00	0.86
J-DJ14	0.94	0.47	0.26	0.17	0.71	1.09	0.32	0.10
J-HN04	1.60	0.17	0.00	0.62	1.59	0.18	0.01	0.66
CRE	0.99	0.06	0.94	1.00	0.99	0.07	0.94	1.00

Table B.2: Probit APE Estimates when $N = 500$, $T = 6$

	$\widehat{\mu}_x/\mu_x$ (true value = 1)				$\widehat{\mu}_d/\mu_d$ (true value = 1)			
	Mean	SD	cv: .95	$\frac{SE}{SD}$	Mean	SD	cv: .95	$\frac{SE}{SD}$
$\rho = 0.0$								
MLE	0.99	0.06	0.93	0.94	0.98	0.08	0.95	1.02
A-FV09	0.95	0.06	0.85	0.94	0.94	0.08	0.91	1.08
A-HN04	1.03	0.07	0.87	0.86	1.01	0.08	0.94	1.00
J-DJ14	1.09	0.09	0.59	0.66	1.11	0.10	0.68	0.77
J-HN04	1.04	0.07	0.85	0.81	1.05	0.09	0.87	0.90
CRE	1.00	0.06	0.95	0.99	1.00	0.08	0.96	1.05
LPM	0.93	0.06	0.78	0.99	1.29	0.08	0.05	1.07
$\rho = 0.4$								
MLE	0.97	0.06	0.91	0.94	0.98	0.08	0.94	1.01
A-FV09	0.93	0.06	0.72	0.93	0.93	0.07	0.87	1.08
A-HN04	1.04	0.07	0.86	0.84	1.02	0.08	0.94	0.98
J-DJ14	1.13	0.09	0.41	0.59	1.13	0.11	0.56	0.71
J-HN04	1.05	0.07	0.78	0.77	1.06	0.09	0.84	0.87
CRE	1.00	0.06	0.95	1.01	1.00	0.08	0.96	1.04
LPM	0.93	0.06	0.80	1.01	1.29	0.08	0.04	1.05
$\rho = 0.8$								
MLE	0.92	0.06	0.66	0.91	0.96	0.07	0.89	0.94
A-FV09	0.72	0.15	0.02	0.33	0.64	0.18	0.01	0.39
A-HN04	1.04	0.07	0.82	0.77	1.00	0.07	0.93	0.93
J-DJ14	1.23	0.10	0.10	0.53	1.14	0.11	0.47	0.62
J-HN04	1.13	0.08	0.32	0.62	1.09	0.08	0.69	0.76
CRE	1.00	0.06	0.95	0.99	1.00	0.08	0.95	1.00
LPM	0.93	0.06	0.77	1.00	1.29	0.08	0.02	1.02

APPENDIX C

DERIVATIONS OF TEST STATISTICS FROM CHAPTER 2

C.1 Derivations from Section 2.3.2

From section 3.2, the score of (2.13) evaluated at $\boldsymbol{\Lambda} = \mathbf{0}$ is identically zero. Assuming we can pass the derivative through the integral, we can work out the following:

$$\nabla_{\Lambda}\ell_i(\boldsymbol{\beta},\boldsymbol{\Lambda}) = \frac{\int\int_{\mathbb{R}^K} h_{it}\left[\prod_{t=1}^{T} p_t(\boldsymbol{x}_i,\boldsymbol{b}_i)^{y_{it}}\right]\left[\sum_{t=1}^{T} y_{it}\boldsymbol{u}_i' \otimes q_t(\boldsymbol{x}_i,,\boldsymbol{b}_i)\right] f(\boldsymbol{u}_i)\,\mathrm{d}\boldsymbol{u}_i}{\int\int_{\mathbb{R}^K} f(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{u}_i,c_i,n_i) f(\boldsymbol{u}_i)\,\mathrm{d}\boldsymbol{u}_i} \tag{C.1}$$

where $h_{it} = \frac{n_i!}{\prod_{t=1}^{T} y_{it}!}$, $q_t(\boldsymbol{x}_i,\boldsymbol{b}_i) = \nabla_{\boldsymbol{b}_i} p_t(\boldsymbol{x}_i,\boldsymbol{b}_i)/p_t(\boldsymbol{x}_i,\boldsymbol{b}_i)$. Evaluating at $\Lambda = 0$, and pulling the terms that do not depend on $\boldsymbol{u}_i$ out of the integrals, we have:

$$\begin{aligned}
\nabla_{\Lambda}\ell_i(\boldsymbol{\beta},\boldsymbol{\Lambda})\Big|_{\boldsymbol{\Lambda}=\mathbf{0}} &= \frac{h_{it}\left[\prod_{t=1}^{T} p_t(\boldsymbol{x}_i,\boldsymbol{\beta})^{y_{it}}\right]\left[\sum_{t=1}^{T} y_{it}\int\int_{\mathbb{R}^K} \boldsymbol{u}_i' \otimes q_t(\boldsymbol{x}_i,\boldsymbol{\beta}) f(\boldsymbol{u}_i)\,\mathrm{d}\boldsymbol{u}_i\right]}{h_{it}\left[\prod_{t=1}^{T} p_t(\boldsymbol{x}_i,\boldsymbol{\beta})^{y_{it}}\right]\int\int_{\mathbb{R}^K} f(\boldsymbol{u}_i)\,\mathrm{d}\boldsymbol{u}_i} && \text{(C.2)}\\
&= \sum_{t=1}^{T} y_{it} E\left[\boldsymbol{u}_i' \otimes q_t(\boldsymbol{x}_i,\boldsymbol{b}_i)\right] && \text{(C.3)}\\
&= \mathbf{0}.
\end{aligned}$$

The second equality uses that $\int\int_{\mathbb{R}^K} f(\boldsymbol{u}_i)\,\mathrm{d}\boldsymbol{u}_i = 1$, while the third follows from independence of $\boldsymbol{x}_{it}$ and $\boldsymbol{u}_i$, as well as $E(\boldsymbol{u}_i) = \mathbf{0}$.

Following the re-parameterization shown in (2.14), stacking the λ_j into $K \times 1$ vector $\boldsymbol{\lambda}$, defining let $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}', \boldsymbol{\lambda}')'$, and following similar steps as before, we have:

$$\frac{\partial \ell_i(\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial \lambda_j}\Big|_{\boldsymbol{\lambda}=0} = \left\{\frac{1}{2\sqrt{\lambda_j}}\left[\sum_{t=1}^{T} y_{it} q_{tj}(\boldsymbol{x}_i,\boldsymbol{\beta})\right]\int\int_{\mathbb{R}^K} u_{ij} f(\boldsymbol{u}_i)\,\mathrm{d}\boldsymbol{u}_i\right\}_{\lambda_j=0} \tag{C.4}$$

where $q_{tj}()$ is the jth element of $q_t()$, The above has $0/0$ form since $E(\boldsymbol{u}_i) = \mathbf{0}$.

Using L'Hopital's rule, the limit, of $\frac{\partial \ell_i(\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial \lambda_j}$ as each element of $\boldsymbol{\lambda}$ approaches zero from above

is:

$$\frac{\frac{1}{2\sqrt{\lambda_j}} \iint_{\mathbb{R}^K} h_{it} \left[\prod_t p_t(\boldsymbol{x}_i, \boldsymbol{b}_i)^{y_{it}}\right] \left\{\sum_t y_{it} r_{tj}(\boldsymbol{x}_i, \boldsymbol{b}_i) + \left[\sum_t y_{it} q_{tj}(\boldsymbol{x}_i, \boldsymbol{b}_i)\right]^2\right\} u_{ij}^2 f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i}{2\left(\frac{1}{2\sqrt{\lambda_j}}\right) \iint_{\mathbb{R}^K} h_{it} \left[\prod_t p_t(\boldsymbol{x}_i, \boldsymbol{b}_i)^{y_{it}}\right] f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i}, \tag{C.5}$$

where $r_{tj}()$ is the (j,j)th element of $\nabla_{\boldsymbol{b}_i} q_t(\boldsymbol{x}_i, \boldsymbol{b}_i)$. The $\frac{1}{2\sqrt{\lambda_j}}$ terms cancel, as do the h_{it} the product terms when we evaluate at $\boldsymbol{\lambda} = \mathbf{0}$ ($\boldsymbol{b}_i = \boldsymbol{\beta}_0$). Then using $\iint_{\mathbb{R}^K} f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i = 1$ and $\iint_{\mathbb{R}^K} u_{ij}^2 f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i = E(u_{ij}^2) = 1$, we get the last K elements of (2.15).

C.2 Derivations from Section 2.3.3

As before, the restricted score of (2.21 is identically zero.

$$\begin{aligned}
\nabla_{\boldsymbol{\Lambda}} \ell_i(\boldsymbol{\beta}, \boldsymbol{\Lambda}) &= \sum_{t=1}^{T} y_{it} \left[\frac{\nabla_{\boldsymbol{\Lambda}} p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})}{p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})}\right] \\
&= \sum_{t=1}^{T} y_{it} \frac{\sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta} + m_r(\boldsymbol{x}_i, \boldsymbol{\Lambda})) \left[\nabla_{\boldsymbol{\lambda}} m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda}) - \nabla_{\boldsymbol{\lambda}} m_r(\boldsymbol{x}_i, \boldsymbol{\Lambda})\right]}{\sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta} + m_r(\boldsymbol{x}_i, \boldsymbol{\Lambda}))},
\end{aligned} \tag{C.6}$$

$$\nabla_{\boldsymbol{\lambda}} m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda}) = \frac{\iint_{\mathbb{R}^K} \exp\left(\boldsymbol{x}_{it}\boldsymbol{\Lambda}\boldsymbol{u}_i\right) (\boldsymbol{u}_i' \otimes \boldsymbol{x}_{it}) f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i}{\iint_{\mathbb{R}^K} \exp(\boldsymbol{x}_{it}\boldsymbol{\Lambda}_0\boldsymbol{u}_i) f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i}. \tag{C.7}$$

The complication arises because

$$\nabla_{\boldsymbol{\lambda}} m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda})\Big|_{\boldsymbol{\Lambda}=\mathbf{0}} = \frac{\iint_{\mathbb{R}^K} (\boldsymbol{u}_i' \otimes \boldsymbol{x}_{it}) f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i}{\iint_{\mathbb{R}^K} f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i} = \mathbf{0}, \tag{C.8}$$

which implies

$$\nabla_{\boldsymbol{\Lambda}} \ell_i(\boldsymbol{\beta}, \boldsymbol{\Lambda})\Big|_{\boldsymbol{\Lambda}=\mathbf{0}} = \mathbf{0}. \tag{C.9}$$

After the re-parameterization, for each of the λ_j, we have:

$$\nabla_{\lambda_j} m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda}) = \left\{\iint_{\mathbb{R}^K} \exp(\boldsymbol{x}_{it}\boldsymbol{\Lambda}_0\boldsymbol{u}_i) f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i\right\}^{-1} \frac{\iint_{\mathbb{R}^K} \exp\left(\boldsymbol{x}_{it}\boldsymbol{\Lambda}\boldsymbol{u}_i\right) x_{itj} u_{ij} f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i}{2\sqrt{\lambda_j}}. \tag{C.10}$$

When evaluated at $\boldsymbol{\lambda} = \mathbf{0}$, the second factor of (C.10) has the form $0/0$.

Using L'Hopital's rule, as each λ_j approaches zero from above, we have:

$$\begin{aligned}\lim_{\boldsymbol{\lambda}\downarrow 0}\left[\frac{\iint_{\mathbb{R}^K}\exp\left(\boldsymbol{x}_{it}\boldsymbol{\Lambda}\boldsymbol{u}_i\right)x_{itj}u_{ij}f(\boldsymbol{u}_i)\,\mathrm{d}\boldsymbol{u}_i}{2\sqrt{\lambda_j}}\right] &= \lim_{\boldsymbol{\lambda}\downarrow 0}\left[\frac{\frac{1}{2\sqrt{\lambda_j}}\iint_{\mathbb{R}^K}\exp\left(\boldsymbol{x}_{it}\boldsymbol{\Lambda}\boldsymbol{u}_i\right)x_{itj}^2u_{ij}^2f(\boldsymbol{u}_i)\,\mathrm{d}\boldsymbol{u}_i}{2(\frac{1}{2\sqrt{\lambda_j}})}\right] \\ &= \frac{x_{itj}^2\iint_{\mathbb{R}^K}u_{ij}^2f(\boldsymbol{u}_i)\,\mathrm{d}\boldsymbol{u}_i}{2} \\ &= \frac{1}{2}x_{itj}^2 \qquad \text{(C.11)}\end{aligned}$$

Plugging these limits in into the expression for $\nabla_{\boldsymbol{\Lambda}}\ell_i(\boldsymbol{\beta},\mathbf{0})$, we get (2.23).

APPENDIX D

SIMULATION RESULTS FROM CHAPTER 3

Table D.1: Finite Sample Properties of Poisson QMLE: $\beta_1 = 0.5, \beta_2 = -0.5, N = 500$

	$\widetilde{\beta}_1$					$\widetilde{\beta}_2$				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.50	0.00	0.06	0.98	0.06	-0.50	0.00	0.09	0.99	0.05
$T = 4$	0.50	0.00	0.04	0.99	0.06	-0.50	0.00	0.06	1.00	0.05
$T = 10$	0.50	0.00	0.03	0.99	0.05	-0.50	0.00	0.04	0.98	0.05
$\sigma = 0.25$										
$T = 2$	0.58	0.08	0.06	0.99	0.29	-0.38	0.12	0.09	0.99	0.28
$T = 4$	0.58	0.08	0.04	1.00	0.52	-0.39	0.11	0.06	0.99	0.46
$T = 10$	0.58	0.08	0.03	1.01	0.87	-0.38	0.12	0.04	1.00	0.84
$\sigma = 0.50$										
$T = 2$	0.74	0.24	0.06	0.97	0.99	-0.19	0.31	0.10	0.99	0.90
$T = 4$	0.74	0.24	0.04	0.95	1.00	-0.19	0.31	0.07	0.95	0.99
$T = 10$	0.74	0.24	0.03	0.94	1.00	-0.19	0.31	0.05	1.00	1.00
$\sigma = 0.75$										
$T = 2$	0.92	0.42	0.07	0.86	1.00	-0.04	0.46	0.11	0.93	0.97
$T = 4$	0.92	0.42	0.06	0.84	1.00	-0.04	0.46	0.09	0.93	0.99
$T = 10$	0.91	0.41	0.05	0.82	1.00	-0.04	0.46	0.07	0.89	1.00
$\sigma = 1.00$										
$T = 2$	1.07	0.57	0.09	0.77	1.00	0.09	0.59	0.15	0.86	0.96
$T = 4$	1.08	0.58	0.09	0.70	1.00	0.08	0.58	0.13	0.78	0.96
$T = 10$	1.08	0.58	0.08	0.71	1.00	0.08	0.58	0.11	0.76	0.98

Table D.2: Finite Sample Properties of Fixed Effects Poisson: $\beta_1 = 0.5, \beta_2 = -0.5, N = 500$

	$\widehat{\beta}_1$					$\widehat{\beta}_2$				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.50	0.00	0.10	0.99	0.05	-0.50	0.00	0.13	0.99	0.05
$T = 4$	0.50	0.00	0.05	0.99	0.05	-0.50	0.00	0.07	1.00	0.05
$T = 10$	0.50	0.00	0.03	0.99	0.05	-0.50	0.00	0.04	0.98	0.05
$\sigma = 0.25$										
$T = 2$	0.50	0.00	0.09	0.98	0.06	-0.50	0.00	0.13	0.99	0.05
$T = 4$	0.50	0.00	0.05	1.01	0.05	-0.50	0.00	0.07	0.99	0.05
$T = 10$	0.50	0.00	0.03	1.00	0.05	-0.50	0.00	0.04	1.00	0.05
$\sigma = 0.50$										
$T = 2$	0.50	0.00	0.08	0.99	0.05	-0.50	0.00	0.14	1.00	0.05
$T = 4$	0.50	0.00	0.04	1.00	0.05	-0.50	0.00	0.08	0.98	0.06
$T = 10$	0.50	0.00	0.02	1.00	0.06	-0.50	0.00	0.04	1.02	0.05
$\sigma = 0.75$										
$T = 2$	0.50	0.00	0.06	1.01	0.05	-0.50	0.00	0.14	0.99	0.06
$T = 4$	0.50	0.00	0.03	0.99	0.06	-0.50	0.00	0.08	0.99	0.05
$T = 10$	0.50	0.00	0.02	0.99	0.06	-0.50	0.00	0.05	0.99	0.05
$\sigma = 1.00$										
$T = 2$	0.50	0.00	0.05	0.97	0.06	-0.50	0.00	0.15	0.97	0.06
$T = 4$	0.50	0.00	0.03	0.97	0.06	-0.50	0.00	0.08	1.01	0.05
$T = 10$	0.50	0.00	0.02	0.97	0.06	-0.50	0.00	0.05	1.01	0.05

Table D.3: Finite Sample Properties of Poisson QMLE: $\beta_1 = 0.5, \beta_2 = -0.5, N = 500$

	$\widetilde{\delta}_1$ (APE)					$\widetilde{\delta}_2$ (ATE)				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.41	0.00	0.05	0.99	0.05	-0.42	0.00	0.08	1.00	0.05
$T = 4$	0.41	0.00	0.04	0.99	0.05	-0.42	0.00	0.05	1.00	0.05
$T = 10$	0.41	0.00	0.02	0.99	0.05	-0.42	0.00	0.03	0.99	0.05
$\sigma = 0.25$										
$T = 2$	0.50	0.07	0.05	0.99	0.24	-0.34	0.12	0.08	1.00	0.30
$T = 4$	0.50	0.07	0.04	1.00	0.43	-0.34	0.11	0.06	0.99	0.49
$T = 10$	0.50	0.07	0.02	1.02	0.79	-0.34	0.12	0.04	1.00	0.85
$\sigma = 0.50$										
$T = 2$	0.74	0.24	0.07	0.97	0.94	-0.20	0.36	0.10	0.99	0.91
$T = 4$	0.74	0.24	0.06	0.95	1.00	-0.20	0.36	0.08	0.95	0.99
$T = 10$	0.74	0.24	0.05	0.97	1.00	-0.20	0.36	0.05	1.00	1.00
$\sigma = 0.75$										
$T = 2$	1.19	0.54	0.15	0.91	1.00	-0.06	0.70	0.16	0.92	0.96
$T = 4$	1.19	0.54	0.14	0.88	1.00	-0.06	0.71	0.12	0.92	0.98
$T = 10$	1.19	0.54	0.13	0.90	1.00	-0.06	0.71	0.10	0.88	0.99
$\sigma = 1.00$										
$T = 2$	2.00	1.07	0.36	0.83	1.00	0.13	1.28	0.28	0.84	0.95
$T = 4$	2.03	1.10	0.39	0.76	0.99	0.12	1.26	0.27	0.72	0.96
$T = 10$	2.02	1.09	0.35	0.82	1.00	0.14	1.28	0.22	0.73	0.97

Table D.4: Finite Sample Properties of Fixed Effects Poisson: $\beta_1 = 0.5, \beta_2 = -0.5, N = 500$

	$\widehat{\delta}_1$ (APE)					$\widehat{\delta}_2$ (ATE)				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.41	0.00	0.08	0.99	0.05	-0.43	0.00	0.11	0.99	0.05
$T = 4$	0.41	0.00	0.04	0.99	0.05	-0.42	0.00	0.06	1.00	0.05
$T = 10$	0.41	0.00	0.02	0.99	0.05	-0.42	0.00	0.04	0.99	0.05
$\sigma = 0.25$										
$T = 2$	0.43	0.00	0.08	0.98	0.05	-0.46	0.00	0.13	0.99	0.05
$T = 4$	0.43	0.00	0.04	1.00	0.05	-0.46	0.00	0.07	0.99	0.05
$T = 10$	0.43	0.00	0.02	1.01	0.05	-0.45	0.00	0.04	1.01	0.04
$\sigma = 0.50$										
$T = 2$	0.50	0.00	0.08	1.00	0.05	-0.57	-0.01	0.18	0.98	0.05
$T = 4$	0.50	0.00	0.05	0.99	0.06	-0.56	0.00	0.10	0.98	0.06
$T = 10$	0.50	0.00	0.03	1.02	0.05	-0.56	0.00	0.06	1.03	0.04
$\sigma = 0.75$										
$T = 2$	0.65	0.00	0.09	1.00	0.05	-0.79	-0.02	0.28	0.98	0.05
$T = 4$	0.65	0.00	0.06	0.98	0.06	-0.77	-0.01	0.16	0.99	0.06
$T = 10$	0.65	0.00	0.05	0.98	0.06	-0.77	-0.01	0.10	0.98	0.05
$\sigma = 1.00$										
$T = 2$	0.93	0.00	0.14	0.94	0.07	-1.17	-0.03	0.46	0.95	0.06
$T = 4$	0.93	0.00	0.12	0.90	0.08	-1.15	-0.01	0.28	0.98	0.06
$T = 10$	0.93	0.00	0.10	0.95	0.08	-1.15	0.00	0.19	0.97	0.06

Table D.5: Finite Sample Properties of Poisson QMLE: $\beta_1 = 0.5, \beta_2 = -0.5, N = 1000$

	$\widetilde{\beta}_1$					$\widetilde{\beta}_2$				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.50	0.00	0.04	1.00	0.05	-0.50	0.00	0.06	0.99	0.05
$T = 4$	0.50	0.00	0.03	1.00	0.05	-0.50	0.00	0.04	0.99	0.05
$T = 10$	0.50	0.00	0.02	0.98	0.05	-0.50	0.00	0.03	0.99	0.05
$\sigma = 0.25$										
$T = 2$	0.58	0.08	0.04	1.01	0.52	-0.38	0.12	0.06	1.01	0.46
$T = 4$	0.58	0.08	0.03	1.02	0.80	-0.38	0.12	0.04	0.99	0.78
$T = 10$	0.58	0.08	0.02	1.03	0.99	-0.38	0.12	0.03	1.03	0.99
$\sigma = 0.50$										
$T = 2$	0.74	0.24	0.04	0.98	1.00	-0.19	0.31	0.07	0.96	0.99
$T = 4$	0.74	0.24	0.03	0.98	1.00	-0.19	0.31	0.05	1.00	1.00
$T = 10$	0.74	0.24	0.02	0.97	1.00	-0.19	0.31	0.03	1.00	1.00
$\sigma = 0.75$										
$T = 2$	0.92	0.42	0.05	0.92	1.00	-0.04	0.46	0.08	0.96	0.99
$T = 4$	0.92	0.42	0.05	0.86	1.00	-0.05	0.45	0.07	0.90	0.99
$T = 10$	0.92	0.42	0.04	0.83	1.00	-0.04	0.46	0.05	0.88	0.99
$\sigma = 1.00$										
$T = 2$	1.09	0.59	0.08	0.78	1.00	0.07	0.57	0.11	0.85	0.98
$T = 4$	1.09	0.59	0.07	0.73	1.00	0.07	0.57	0.10	0.78	0.98
$T = 10$	1.09	0.59	0.07	0.75	1.00	0.07	0.57	0.09	0.77	0.99

Table D.6: Finite Sample Properties of Fixed Effects Poisson: $\beta_1 = 0.5, \beta_2 = -0.5, N = 1000$

	$\widehat{\beta}_1$					$\widehat{\beta}_2$				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.50	0.00	0.07	0.99	0.05	-0.50	0.00	0.09	1.01	0.05
$T = 4$	0.50	0.00	0.04	1.00	0.05	-0.50	0.00	0.05	0.98	0.05
$T = 10$	0.50	0.00	0.02	0.98	0.05	-0.50	0.00	0.03	0.99	0.05
$\sigma = 0.25$										
$T = 2$	0.50	0.00	0.06	0.98	0.05	-0.50	0.00	0.09	1.00	0.05
$T = 4$	0.50	0.00	0.03	1.00	0.05	-0.50	0.00	0.05	0.99	0.06
$T = 10$	0.50	0.00	0.02	1.04	0.04	-0.50	0.00	0.03	1.01	0.04
$\sigma = 0.50$										
$T = 2$	0.50	0.00	0.05	1.00	0.05	-0.50	0.00	0.10	1.01	0.05
$T = 4$	0.50	0.00	0.03	0.98	0.06	-0.50	0.00	0.06	1.00	0.06
$T = 10$	0.50	0.00	0.02	1.02	0.05	-0.50	0.00	0.03	0.99	0.05
$\sigma = 0.75$										
$T = 2$	0.50	0.00	0.04	1.02	0.04	-0.50	0.00	0.10	0.99	0.05
$T = 4$	0.50	0.00	0.03	0.99	0.05	-0.50	0.00	0.06	1.00	0.05
$T = 10$	0.50	0.00	0.01	1.01	0.05	-0.50	0.00	0.03	1.00	0.05
$\sigma = 1.00$										
$T = 2$	0.50	0.00	0.03	0.99	0.05	-0.50	0.00	0.11	0.99	0.05
$T = 4$	0.50	0.00	0.02	0.97	0.06	-0.50	0.00	0.06	1.01	0.05
$T = 10$	0.50	0.00	0.01	1.02	0.05	-0.50	0.00	0.03	1.00	0.05

Table D.7: Finite Sample Properties of Poisson QMLE: $\beta_1 = 0.5, \beta_2 = -0.5, N = 1000$

	$\widetilde{\delta}_1$ (APE)					$\widetilde{\delta}_2$ (ATE)				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.41	0.00	0.04	0.99	0.06	-0.42	0.00	0.05	0.99	0.05
$T = 4$	0.41	0.00	0.03	0.99	0.05	-0.42	0.00	0.04	0.99	0.05
$T = 10$	0.41	0.00	0.02	0.98	0.05	-0.42	0.00	0.02	1.00	0.05
$\sigma = 0.25$										
$T = 2$	0.50	0.07	0.04	1.00	0.45	-0.34	0.11	0.06	1.01	0.48
$T = 4$	0.50	0.07	0.03	1.00	0.71	-0.34	0.12	0.04	0.99	0.79
$T = 10$	0.50	0.07	0.02	1.01	0.98	-0.34	0.11	0.03	1.03	0.99
$\sigma = 0.50$										
$T = 2$	0.74	0.24	0.05	0.98	1.00	-0.20	0.36	0.08	0.96	0.99
$T = 4$	0.74	0.24	0.04	0.98	1.00	-0.20	0.36	0.05	1.00	1.00
$T = 10$	0.74	0.24	0.04	0.98	1.00	-0.20	0.36	0.04	1.00	1.00
$\sigma = 0.75$										
$T = 2$	1.19	0.54	0.11	0.95	1.00	-0.06	0.70	0.11	0.96	0.99
$T = 4$	1.19	0.54	0.10	0.91	1.00	-0.06	0.70	0.09	0.89	0.99
$T = 10$	1.19	0.54	0.10	0.90	1.00	-0.06	0.70	0.07	0.87	0.99
$\sigma = 1.00$										
$T = 2$	2.04	1.10	0.28	0.84	1.00	0.12	1.26	0.22	0.83	0.97
$T = 4$	2.03	1.10	0.29	0.80	1.00	0.12	1.26	0.21	0.73	0.97
$T = 10$	2.04	1.10	0.26	0.85	1.00	0.12	1.27	0.17	0.75	0.98

Table D.8: Finite Sample Properties of Fixed Effects Poisson: $\beta_1 = 0.5, \beta_2 = -0.5, N = 1000$

	$\widehat{\delta}_1$ (APE)					$\widehat{\delta}_2$ (ATE)				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.41	0.00	0.06	0.99	0.05	-0.42	0.00	0.08	1.01	0.05
$T = 4$	0.41	0.00	0.03	0.99	0.06	-0.42	0.00	0.05	0.98	0.05
$T = 10$	0.41	0.00	0.02	0.98	0.05	-0.42	0.00	0.03	0.99	0.05
$\sigma = 0.25$										
$T = 2$	0.43	0.00	0.06	0.98	0.05	-0.46	0.00	0.09	1.00	0.05
$T = 4$	0.43	0.00	0.03	1.00	0.05	-0.45	0.00	0.05	0.99	0.06
$T = 10$	0.43	0.00	0.02	1.02	0.05	-0.46	0.00	0.03	1.02	0.04
$\sigma = 0.50$										
$T = 2$	0.50	0.00	0.06	1.00	0.05	-0.57	-0.01	0.13	1.01	0.04
$T = 4$	0.50	0.00	0.03	0.99	0.05	-0.56	0.00	0.07	1.01	0.05
$T = 10$	0.50	0.00	0.02	1.01	0.05	-0.56	0.00	0.04	1.00	0.05
$\sigma = 0.75$										
$T = 2$	0.65	0.00	0.06	1.02	0.04	-0.78	-0.01	0.19	0.99	0.05
$T = 4$	0.65	0.00	0.04	0.98	0.06	-0.77	0.00	0.11	1.00	0.05
$T = 10$	0.65	0.00	0.03	1.00	0.05	-0.77	0.00	0.07	1.01	0.05
$\sigma = 1.00$										
$T = 2$	0.93	0.00	0.10	0.95	0.06	-1.16	-0.01	0.31	0.98	0.05
$T = 4$	0.93	0.00	0.08	0.93	0.07	-1.15	0.00	0.19	0.99	0.05
$T = 10$	0.93	0.00	0.07	0.96	0.07	-1.15	0.00	0.13	0.99	0.05

Table D.9: Finite Sample Properties of Poisson QMLE: $\beta_1 = 0.5, \beta_2 = -0.5, N = 2000$

	$\widetilde{\beta}_1$					$\widetilde{\beta}_2$				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.50	0.00	0.03	1.00	0.05	-0.50	0.00	0.04	0.99	0.05
$T = 4$	0.50	0.00	0.02	1.01	0.05	-0.50	0.00	0.03	1.00	0.05
$T = 10$	0.50	0.00	0.01	0.97	0.06	-0.50	0.00	0.02	0.99	0.05
$\sigma = 0.25$										
$T = 2$	0.58	0.08	0.03	1.02	0.81	-0.38	0.12	0.04	1.02	0.75
$T = 4$	0.58	0.08	0.02	1.01	0.98	-0.38	0.12	0.03	1.00	0.96
$T = 10$	0.58	0.08	0.01	0.98	1.00	-0.38	0.12	0.02	0.97	1.00
$\sigma = 0.50$										
$T = 2$	0.74	0.24	0.03	0.98	1.00	-0.19	0.31	0.05	1.00	1.00
$T = 4$	0.74	0.24	0.02	1.00	1.00	-0.19	0.31	0.03	1.04	1.00
$T = 10$	0.74	0.24	0.02	0.97	1.00	-0.19	0.31	0.02	0.99	1.00
$\sigma = 0.75$										
$T = 2$	0.92	0.42	0.04	0.92	1.00	-0.04	0.46	0.06	0.97	1.00
$T = 4$	0.92	0.42	0.04	0.88	1.00	-0.05	0.45	0.05	0.95	1.00
$T = 10$	0.92	0.42	0.03	0.89	1.00	-0.05	0.45	0.04	0.92	1.00
$\sigma = 1.00$										
$T = 2$	1.09	0.59	0.06	0.83	1.00	0.07	0.57	0.09	0.84	0.99
$T = 4$	1.09	0.59	0.06	0.79	1.00	0.06	0.56	0.08	0.82	0.99
$T = 10$	1.09	0.59	0.05	0.80	1.00	0.07	0.57	0.07	0.80	0.99

Table D.10: Finite Sample Properties of Fixed Effects Poisson: $\beta_1 = 0.5, \beta_2 = -0.5, N = 2000$

	$\widehat{\beta}_1$					$\widehat{\beta}_2$				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.50	0.00	0.05	0.99	0.06	-0.50	0.00	0.06	0.96	0.06
$T = 4$	0.50	0.00	0.03	1.00	0.05	-0.50	0.00	0.04	1.00	0.05
$T = 10$	0.50	0.00	0.01	0.97	0.06	-0.50	0.00	0.02	0.99	0.05
$\sigma = 0.25$										
$T = 2$	0.50	0.00	0.04	1.01	0.05	-0.50	0.00	0.07	0.99	0.06
$T = 4$	0.50	0.00	0.02	1.02	0.05	-0.50	0.00	0.04	1.00	0.05
$T = 10$	0.50	0.00	0.01	0.98	0.06	-0.50	0.00	0.02	0.98	0.05
$\sigma = 0.50$										
$T = 2$	0.50	0.00	0.04	0.99	0.05	-0.50	0.00	0.07	0.99	0.05
$T = 4$	0.50	0.00	0.02	1.02	0.05	-0.50	0.00	0.04	1.01	0.05
$T = 10$	0.50	0.00	0.01	0.99	0.05	-0.50	0.00	0.02	0.98	0.05
$\sigma = 0.75$										
$T = 2$	0.50	0.00	0.03	0.95	0.06	-0.50	0.00	0.07	0.97	0.05
$T = 4$	0.50	0.00	0.02	0.99	0.06	-0.50	0.00	0.04	1.03	0.04
$T = 10$	0.50	0.00	0.01	0.98	0.05	-0.50	0.00	0.02	0.98	0.05
$\sigma = 1.00$										
$T = 2$	0.50	0.00	0.02	0.98	0.05	-0.50	0.00	0.08	0.98	0.05
$T = 4$	0.50	0.00	0.01	0.96	0.06	-0.50	0.00	0.04	0.99	0.05
$T = 10$	0.50	0.00	0.01	1.01	0.05	-0.50	0.00	0.02	1.01	0.05

Table D.11: Finite Sample Properties of Poisson QMLE: $\beta_1 = 0.5, \beta_2 = -0.5, N = 2000$

	$\widetilde{\delta}_1$ (APE)					$\widetilde{\delta}_2$ (ATE)				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.41	0.00	0.02	0.99	0.05	-0.42	0.00	0.04	0.99	0.05
$T = 4$	0.41	0.00	0.02	1.01	0.05	-0.42	0.00	0.03	1.00	0.05
$T = 10$	0.41	0.00	0.01	0.96	0.05	-0.42	0.00	0.02	0.98	0.05
$\sigma = 0.25$										
$T = 2$	0.50	0.07	0.03	1.02	0.75	-0.34	0.11	0.04	1.02	0.77
$T = 4$	0.50	0.07	0.02	1.00	0.95	-0.34	0.11	0.03	1.00	0.97
$T = 10$	0.50	0.07	0.01	0.99	1.00	-0.34	0.11	0.02	0.97	1.00
$\sigma = 0.50$										
$T = 2$	0.74	0.24	0.04	0.98	1.00	-0.20	0.36	0.05	1.00	1.00
$T = 4$	0.74	0.24	0.03	0.99	1.00	-0.20	0.36	0.04	1.04	1.00
$T = 10$	0.74	0.24	0.03	0.99	1.00	-0.20	0.36	0.03	0.99	1.00
$\sigma = 0.75$										
$T = 2$	1.19	0.54	0.08	0.95	1.00	-0.06	0.71	0.08	0.96	1.00
$T = 4$	1.19	0.55	0.08	0.93	1.00	-0.06	0.70	0.06	0.95	1.00
$T = 10$	1.19	0.55	0.07	0.95	1.00	-0.06	0.70	0.05	0.92	1.00
$\sigma = 1.00$										
$T = 2$	2.03	1.10	0.19	0.91	1.00	0.11	1.26	0.16	0.84	0.99
$T = 4$	2.04	1.11	0.20	0.86	1.00	0.11	1.25	0.15	0.80	0.99
$T = 10$	2.04	1.11	0.19	0.90	1.00	0.11	1.26	0.13	0.77	0.99

Table D.12: Finite Sample Properties of Fixed Effects Poisson: $\beta_1 = 0.5, \beta_2 = -0.5, N = 2000$

	$\widehat{\delta}_1$ (APE)					$\widehat{\delta}_2$ (ATE)				
	Mean	Bias	SD	SE/SD	RP(0.05)	Mean	Bias	SD	SE/SD	RP(0.05)
$\sigma = 0.00$										
$T = 2$	0.41	0.00	0.04	0.99	0.05	-0.42	0.00	0.06	0.96	0.06
$T = 4$	0.41	0.00	0.02	0.99	0.05	-0.42	0.00	0.03	1.00	0.05
$T = 10$	0.41	0.00	0.01	0.97	0.06	-0.42	0.00	0.02	0.98	0.05
$\sigma = 0.25$										
$T = 2$	0.43	0.00	0.04	1.01	0.05	-0.46	0.00	0.07	0.98	0.05
$T = 4$	0.43	0.00	0.02	1.01	0.04	-0.45	0.00	0.04	1.00	0.05
$T = 10$	0.43	0.00	0.01	0.98	0.05	-0.46	0.00	0.02	0.99	0.05
$\sigma = 0.50$										
$T = 2$	0.50	0.00	0.04	0.98	0.05	-0.56	0.00	0.09	0.99	0.05
$T = 4$	0.50	0.00	0.02	1.03	0.04	-0.56	0.00	0.05	1.01	0.05
$T = 10$	0.50	0.00	0.02	0.99	0.05	-0.56	0.00	0.03	0.97	0.05
$\sigma = 0.75$										
$T = 2$	0.65	0.00	0.05	0.96	0.06	-0.77	0.00	0.14	0.97	0.06
$T = 4$	0.65	0.00	0.03	0.98	0.06	-0.77	0.00	0.08	1.03	0.05
$T = 10$	0.65	0.00	0.02	0.99	0.06	-0.77	0.00	0.05	0.99	0.05
$\sigma = 1.00$										
$T = 2$	0.93	0.00	0.07	0.99	0.06	-1.16	-0.01	0.22	0.98	0.05
$T = 4$	0.93	0.00	0.06	0.96	0.06	-1.15	0.00	0.14	0.99	0.06
$T = 10$	0.93	0.00	0.05	1.00	0.05	-1.15	0.00	0.09	0.99	0.05

REFERENCES

REFERENCES

Alexander, B. and R. Breunig (2014). "A Monte Carlo study of bias corrections for panel probit models". In: *Journal of Statistical Computation and Simulation* 86.1, pp. 74–90. DOI: 10.1080/00949655.2014.994516.

Andersen, E.B. (1970). "Asymptotic Properties of Conditional Maximum-Likelihood Estimators". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 32.2, pp. 283–301. ISSN: 00359246. DOI: 10.2307/2984535. URL: http://www.jstor.org/stable/2984535.

Arellano, M. and J. Hahn (2007). *Understanding Bias in Nonlinear Panel Models: Some Recent Developments. In Advances in Economics and Econometrics, Blundell R, Newey W, Persson T (eds)*. Cambridge: Cambridge University Press.

Bessen, J. (2009). "Matching patent data to compustat firms". In: *NBER working paper*.

Blundell, R. and J.L. Powell (2003). "Endogeneity in Nonparametric and Semiparametric Regression Models". In: *Advances in Economics and Econometrics: Theory and Applications: Eighth World Congress Vol II*, pp. 312–357. DOI: 10.1017/ccol0521818737.010.

Bound, J. et al. (1982). "Who does R&D and who patents?" In:

Cameron, A.C. and P.K. Trivedi (2013). *Regression analysis of count data*. 2nd ed. Cambridge University Press.

Chamberlain, G. (1980). "Analysis of Covariance with Qualitative Data". In: *Review of Economic Studies* 47, pp. 225–238. DOI: 10.2307/2297110.

— (1982). "Multivariate Regression Models For Panel Data". In: *Journal of Econometrics* 18, pp. 5–46. DOI: 10.1016/0304-4076(82)90094-x.

— (1992). "Comment: Sequential moment restrictions in panel data". In: *Journal of Business & Economic Statistics* 10.1, pp. 20–26.

Chay, K.Y. and D.R. Hyslop (2014). "Identification and Estimation of Dynamic Binary Response Panel Data Models: Empirical Evidence Using Alternative Approaches". In: *Safety Nets and Benefit Dependence (Research in Labor Economics)*, pp. 1–39. DOI: 10.1108/s0147-9121_2014_0000039001.

Chesher, A. (1984). "Testing for Neglected Heterogeneity". In: *Econometrica* 52.4, p. 865. DOI: 10.2307/1911188.

Dhaene, G. and K. Jochmans (2015). “Split-panel Jackknife Estimation of Fixed-effect Models”. In: *Review of Economic Studies* 82.3, pp. 991–1030. DOI: 10.1093/restud/rdv007.

Fernandez-Val, I. (2009). “Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models”. In: *Journal of Econometrics* 150, pp. 71–85. DOI: 10.1016/j.jeconom.2009.02.007.

Fernández-Val, Iván and Martin Weidner (2016). “Individual and time effects in nonlinear panel models with large N, T”. In: *Journal of Econometrics* 192.1, pp. 291–312.

Gourieroux, C., A. Monfort, and A. Trognon (1984). “Pseudo Maximum Likelihood Methods: Theory”. In: *Econometrica* 52.3, p. 681. DOI: 10.2307/1913471.

Greene, W.H. (2004). “The Behavior of the Fixed Effects Estimator in Nonlinear Models”. In: *The Econometrics Journal* 7, pp. 98–119. DOI: 10.1111/j.1368-423x.2004.00123.x.

— (2012). *Econometric analysis*. Prentice Hall.

Greene, W.H. and C. Mckenzie (2015). “An LM test based on generalized residuals for random effects in a nonlinear model”. In: *Economics Letters* 127, pp. 47–50. DOI: 10.1016/j.econlet.2014.12.031.

Gurmu, Shiferaw and Fidel Pérez-Sebastián (2008). “Patents, R&D and lag effects: evidence from flexible methods for count panel data on manufacturing firms”. In: *Empirical Economics* 35.3, pp. 507–526.

Hahn, J. and G. Kuersteiner (2011). “Bias reduction for dynamic nonlinear panel models with fixed effects”. In: *Econometric Theory* 27.06, pp. 1152–1191.

Hahn, J., H.R. Moon, and C. Snider (2015). “LM test of neglected correlated random effects and its application”. In: *Journal of Business & Economic Statistics* forthcoming.

Hahn, J. and W. Newey (2004). “Jackknife and Analytical Bias Reduction for Nonlinear Panel Models”. In: *Econometrica* 72, pp. 1295–1319. DOI: 10.1111/j.1468-0262.2004.00533.x.

Hahn, J., W.K. Newey, and R.J. Smith (2014). “Neglected heterogeneity in moment condition models”. In: *Journal of Econometrics* 178, pp. 86–100.

Hall, B., Z. Griliches, and J. Hausman (1986). “Patents and R and D: Is There a Lag?” In: *International Economic Review*, pp. 265–283.

Hall, B., A. Jaffe, and M. Trajtenberg (2001). *The NBER patent citation data file: Lessons, insights and methodological tools*. Tech. rep. National Bureau of Economic Research.

Hausman, J., B. Hall, and Z. Griliches (1984). "Econometric Models for Count Data with an Application to the Patents-R&D Relationship". In: *Econometrica* 52.4, p. 909. DOI: 10.2307/1911191.

Lancaster, T. (2000). "The Incidental Parameters Problem since 1948". In: *Journal of Econometrics* 95, pp. 391–413. DOI: 10.1016/s0304-4076(99)00044-5.

— (2002). "Orthogonal parameters and panel data". In: *The Review of Economic Studies* 69.3, pp. 647–666.

Lee, L. and A. Chesher (1986). "Specification testing when score test statistics are identically zero". In: *Journal of Econometrics* 31.2, pp. 121–149. DOI: 10.1016/0304-4076(86)90045-x.

Lee, M. and S. Kobayashi (2001). "Proportional treatment effects for count response panel data: effects of binary exercise on health care demand". In: *Health Economics* 10.5, pp. 411–428.

Mundlak, Y. (1978). "On the pooling of Time Series and Cross Section Data". In: *Econometrica* 46, pp. 69–85.

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

Pakes, A. and Z. Griliches (1980). "Patents and R and D at the firm level: A first look". In:

Rabe-Hesketh, S. and A. Skrondal (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. English. GB: CRC Press.

— (2013). "Avoiding biased versions of Wooldridge's simple solution to the initial conditions problem". In: *Economics Letters* 120.2, pp. 346–349. DOI: 10.1016/j.econlet.2013.05.009.

Stoker, T. (1986). "Consistent Estimation of Scaled Coefficients". In: *Econometrica* 54.6, pp. 1461–1481. DOI: 10.2307/1914309.

Vamoş, C., Ş. Şoltuz, and M. Crăciun (2007). "Order 1 autoregressive process of finite length". In: *Rev. Anal. Numér. Théor. Approx.* 36.2, pp. 199–214.

Wang, P., I.M. Cockburn, and M.L. Puterman (1998). "Analysis of patent data—a mixed-Poisson-regression-model approach". In: *Journal of Business & Economic Statistics* 16.1, pp. 27–41.

White, H. (1982). "Maximum likelihood estimation of misspecified models". In: *Econometrica: Journal of the Econometric Society*, pp. 1–25.

Wooldridge, J.M. (1992). "Some alternatives to the Box-Cox regression model". In: *International Economic Review*, pp. 935–955.

Wooldridge, J.M. (1997). “Multiplicative panel data models without the strict exogeneity assumption”. In: *Econometric Theory* 13.05, pp. 667–678.

— (1999). “Distribution-free estimation of some nonlinear panel data models”. In: *Journal of Econometrics* 90.1, pp. 77–97. DOI: 10.1016/s0304-4076(98)00033-5.

— (2010). *Econometric analysis of cross section and panel data*. MIT Press.