BLS WORKING PAPERS

U.S. Department of Labor
U.S. Bureau of Labor Statistics
Office of Prices and Living Conditions

Exponential panel models with coefficient heterogeneity

Robert S. Martin, U.S. Bureau of Labor Statistics

Working Paper 503
August 2018

Exponential panel models with coefficient heterogeneity

Robert S. Martin*

August 6, 2018

Abstract

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.

Keywords: Fixed effects Poisson model; Panel data; Quasi-conditional maximum likelihood; Random coefficients

JEL Codes: C23

*Division of Price and Index Number Research, Bureau of Labor Statistics, 2 Massachusetts Ave, NE, Washington, DC 20212, USA. Email: Martin.Robert@bls.gov

1 Introduction

The fixed effects Poisson (FEP) estimator, also known as the multinomial quasi conditional maximum likelihood estimator (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 paper 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 paper is an extension to the unobserved effects exponential model that allows for additional heterogeneity in the form of random coefficients. While several previous studies have considered conditional Poisson random 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 paper 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 test that is applicable to broader types of data. One novel contribution of this paper is to treat random coefficients and the traditional multiplicative effect separately.[1] This is particularly important for testing, as one may want to test for slope heterogeneity without taking a stand on the dependence between the multiplicative effect and the explanatory variables.

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

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. I also provide estimators of average partial effects. In an application to the patent R&D relationship among U.S. manufacturing firms, I find that the classical LM test strongly rejects the null of constant slopes. However, a test that restricts attention to the conditional mean function (as I propose), fails to reject the null of constant slopes. Accordingly, fitting parametric models of heterogeneous slopes tends to offer insignificant estimates.

The rest of this paper 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 paper'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 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, 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.

Generalizations of the Poisson distribution that allow for heterogeneity are, naturally, aimed at analyzing counts like health outcomes. 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).[2] 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 full MLE with 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. Such an approach might be possible in a correlated random effects (CRE) setting, but computational difficulty increases substantially as more coefficients are allowed to be random.

Wang, Cockburn, and Puterman (1998)) allow dependence between the heterogeneity and explanatory variables in the 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

[2] Jochmann and León-González (2004) also propose semiparametric Bayesian approach where a subset of coefficients are allowed to be random effects.

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 previous studies by proposing a method to flexibly allow continuous random slopes in exponential panel models while not limiting analysis to counts or fully specifying the joint distribution of the responses. One benefit of my approach is that as with FEP, I can allow an unrestricted relationship between the explanatory variables and the multiplicative effect.

3 Theory

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

$$E(y_{it}|\boldsymbol{x}_i, c_i) = E(y_{it}|\boldsymbol{x}_{it}, c_i) = c_i \exp(\boldsymbol{x}_{it}\boldsymbol{\beta}_0) \tag{1}$$

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.[3] Eq. 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 as Poisson with mean given by Eq. 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

$$\ell_i^M(\boldsymbol{\beta}) = \sum_{t=1}^{T} y_{it} \log\left[p_t(\boldsymbol{x}_i, \boldsymbol{\beta})\right], \tag{2}$$

where

$$p_t(\boldsymbol{x}_i, \boldsymbol{\beta}) \equiv \frac{\exp(\boldsymbol{x}_{it}\boldsymbol{\beta})}{\sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta})}. \tag{3}$$

The feature that c_i enters conditional mean function multiplicatively means it cancels out of

[3] 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.

$p_t(\boldsymbol{x}_i, \boldsymbol{\beta})$ 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 Eq. 2 as long as Eq. 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).

3.1 The fixed effects Poisson model with coefficient heterogeneity

Eq. 1 is generally untrue, however, if the coefficients in the conditional mean function vary by individual i, as in the following:

$$E(y_{it}|\boldsymbol{x}_i, c_i, \boldsymbol{b}_i) = E(y_{it}|\boldsymbol{x}_{it}, c_i, \boldsymbol{b}_i) = c_i \exp(\boldsymbol{x}_{it}\boldsymbol{b}_i), \tag{4}$$

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 Eq. 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

$$\frac{\partial \log\left[E(y_{it}|\boldsymbol{x}_i, c_i, \boldsymbol{b}_i)\right]}{\partial x_{itj}} = b_{ij}, \tag{5}$$

so the model 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 Eq. 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 Eq. 4 and the Law of Iterated Expectations

(LIE) that

$$E(y_{it}|\boldsymbol{x}_i, c_i) = c_i \exp\left[\boldsymbol{x}_{it}\boldsymbol{\beta}_0 + g_t(\boldsymbol{x}_i, \boldsymbol{x}_{it}, c_i; \boldsymbol{\psi}_0)\right], \tag{6}$$

where

$$g_t(\boldsymbol{x}_i, \boldsymbol{x}_{it}, c_i; \boldsymbol{\psi}_0) = \log\left\{E\left[\exp(\boldsymbol{x}_{it}\boldsymbol{d}_i)|\boldsymbol{x}_i, c_i\right]\right\} = \log\left\{\iint_{\mathbb{R}^K} \exp(\boldsymbol{x}_{it}\boldsymbol{\mathfrak{d}}_i) f(\boldsymbol{\mathfrak{d}}_i|\boldsymbol{x}_i, c_i)\, \mathrm{d}\boldsymbol{\mathfrak{d}}_i\right\}, \tag{7}$$

assuming the expectation exists. The exponential function now contains an unknown term that is generally nonzero and varies over time.[4] 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 Eq. 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.

The key contributions of this paper follow from focusing attention on $g_t(\boldsymbol{x}_i, c_i; \boldsymbol{\psi}_0)$ in the conditional mean function instead of modeling the entire conditional distribution of $\{y_{i1}, \ldots, y_{iT}\}$. As the next subsections describe, testing using the full conditional likelihood is likely to be over-sensitive. On the other hand, a parametric assumption for $D(\boldsymbol{b}_i|c_i, \boldsymbol{x}_i)$ yields a simple estimation strategy and has some robustness properties in terms of testing.

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

[4] 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 $\boldsymbol{x}_{it}$.

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 could be less of a concern in nonlinear models. The following statements formalize the assumptions:

$$y_{it}|(\boldsymbol{x}_i, c_i, \boldsymbol{b}_i) \sim Poisson\left[c_i \exp(\boldsymbol{x}_{it}\boldsymbol{b}_i)\right],\ i = 1, \ldots, N;\ t = 1, \ldots, T, \tag{8}$$

$$\{y_{i1}, \ldots, y_{iT}\} \text{ are independent conditional on } \{\boldsymbol{x}_i, c_i, \boldsymbol{b}_i\} \tag{9}$$

$$\boldsymbol{b}_i = \boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_0 \boldsymbol{u}_i, \text{where } \boldsymbol{u}_i|(\boldsymbol{x}_i, c_i) \sim F(\boldsymbol{0}, \boldsymbol{I}_K), \tag{10}$$

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

From Chesher (1984), assuming Eq. 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

$$\boldsymbol{y}_i|(n_i, \boldsymbol{x}_i, c_i, \boldsymbol{b}_i) \sim Multinomial(n_i, p_1(\boldsymbol{x}_i, \boldsymbol{b}_i), \ldots, p_T(\boldsymbol{x}_i, \boldsymbol{b}_i)), \tag{11}$$

where

$$p_t(\boldsymbol{x}_i, \boldsymbol{b}_i) \equiv \frac{\exp(x_{it}\boldsymbol{b}_i)}{\sum_{r=1}^{T} \exp(x_{ir}\boldsymbol{b}_i)}. \tag{12}$$

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

$$\ell_i(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \log\left\{\iint_{\mathbb{R}^K} \frac{n_i!}{\prod_{t=1}^{T} y_{it}!} \prod_{t=1}^{T} \left[p_t(\boldsymbol{x}_i, \boldsymbol{b}_i)^{y_{it}}\right] f(\boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u}_i\right\}, \tag{13}$$

where the integral is of K dimensions.

An LM test of $H_0 : \boldsymbol{\Lambda}_0 = \boldsymbol{0}$ 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.[5] Chesher (1984) proposed re-parameterizing the scale assumption and restricting the correlation among the heterogeneity allowed under the alternative.[6]

$$\boldsymbol{\Lambda}_0 = diag\left\{\sqrt{\lambda_{1,0}}, \ldots, \sqrt{\lambda_{K,0}}\right\} \tag{14}$$

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.[7] Under Eq. 14, the restricted score has the 0/0 form, but the limits follow from L'Hopital's rule. The algebraic details are collected in the appendix. Collecting the λ_j in the $K \times 1$ vector $\boldsymbol{\lambda}$, the restricted score is:

$$\boldsymbol{s}_i(\boldsymbol{\beta}, \mathbf{0}) \equiv \lim_{\boldsymbol{\lambda} \downarrow 0} \left\{\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\beta}, \boldsymbol{\lambda})'\right\} = \sum_{i=1}^{N} \left\{\begin{array}{c} \sum_{t=1}^{T} y_{it} \left[\nabla_{\beta} p_t(\boldsymbol{x}_i, \boldsymbol{\beta})' / p_t(\boldsymbol{x}_i, \boldsymbol{\beta})\right] \\ \frac{1}{2} a_1(\boldsymbol{x}_i, \boldsymbol{\beta}) \\ \vdots \\ \frac{1}{2} a_K(\boldsymbol{x}_i, \boldsymbol{\beta}) \end{array}\right\}, \tag{15}$$

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

$$\begin{aligned} &\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}) \\ &\equiv \sum_{t=1}^{T} \nabla^2_{\boldsymbol{\beta}} \ell^M_{it}(\boldsymbol{\beta}) + \left(\sum_{t=1}^{T} \nabla_{\boldsymbol{\beta}} \ell^M_{it}(\boldsymbol{\beta})\right)' \left(\sum_{t=1}^{T} \nabla_{\boldsymbol{\beta}} \ell^M_{it}(\boldsymbol{\beta})\right). \end{aligned} \tag{16}$$

In this last expression, ℓ^M_{it} 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 Eq. 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.

[5] See the appendix for the derivation.

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

[7] 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). See Gu (2016) for an approach to the one-sided test.

The downside of this approach concerns robustness to failure of Eq. 8 or Eq. 9. Chesher (1984) notes that statistics derived using this approach resemble White's (1982) information matrix test for general model misspecification, as $E[\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})] = \mathbf{0}$ 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.

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:

$$E(y_{it}|\boldsymbol{x}_i, c_i, \boldsymbol{b}_i) = E(y_{it}|\boldsymbol{x}_{it}, c_i, \boldsymbol{b}_i) = c_i \exp(\boldsymbol{x}_{it}\boldsymbol{b}_i) \tag{17}$$

and

$$\boldsymbol{b}_i = \boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_0 \boldsymbol{u}_i, \text{where } \boldsymbol{u}_i|(\boldsymbol{x}_i, c_i) \sim F(\mathbf{0}, \boldsymbol{I}_K), \tag{18}$$

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

Similar to before, these conditions imply:

$$E(y_{it}|\boldsymbol{x}_i, c_i) = c_i \exp\left[\boldsymbol{x}_{it}\boldsymbol{\beta}_0 + m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda}_0)\right], \tag{19}$$

where

$$m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda}_0) = \log\left\{E\left[\exp(\boldsymbol{x}_{it}\boldsymbol{\Lambda}_0\boldsymbol{u}_i)|\boldsymbol{x}_i, c_i\right]\right\} = \log\left\{\iint_{\mathbb{R}^K} \exp(\boldsymbol{x}_{it}\boldsymbol{\Lambda}_0\boldsymbol{u}_i) f(\boldsymbol{u}_i)\,\mathrm{d}\boldsymbol{u}_i\right\}. \tag{20}$$

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 $\boldsymbol{H_0} : \boldsymbol{\Lambda}_0 = 0$ 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 multinomial with the augmented conditional mean function.[8]

$$\ell_i(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \sum_{t=1}^{T} y_{it} \log \left[p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})\right], \tag{21}$$

where

$$p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \equiv \frac{\exp\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda})\right)}{\sum_{r=1}^{T} \exp\left(\boldsymbol{x}_{ir}\boldsymbol{\beta} + m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda})\right)}. \tag{22}$$

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 Eq. 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 the appendix. The score evaluated at the parameter restriction is:

$$\boldsymbol{s}_i(\boldsymbol{\beta}, \mathbf{0}) = \left\{ \begin{array}{c} \sum_{t=1}^{T} y_{it} \left[\nabla_{\boldsymbol{\beta}} p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \mathbf{0})' / p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \mathbf{0})\right] \\ \frac{1}{2} \sum_{t=1}^{T} y_{it} \left[\sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta}) \left(x_{it1}^2 - x_{ir1}^2\right)\right] / \sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta}) \\ \vdots \\ \frac{1}{2} \sum_{t=1}^{T} y_{it} \left[\sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta}) \left(x_{itK}^2 - x_{irK}^2\right)\right] / \sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta}) \end{array} \right\}. \tag{23}$$

The last K elements are equal to $\frac{1}{2}$ times the restricted scores for testing an exclusion of $\left\{x_{it1}^2, \ldots x_{itK}^2\right\}$ from the FEP model. As the next subsection illustrates, these correspond to exclusion restrictions one would test when the slopes are assumed to have conditional normal distributions under the alternative hypothesis, though normality was not used in the derivation of this test.

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)

[8] Note that in general, this would not be the true log-likelihood function even if the dependent variables were conditional Poisson.

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 Eq. 15, but fails to reject based on Eq. 23, then he or she can proceed in estimating the model based on Eq. 1 with some peace of mind.

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:

$$\begin{aligned} \boldsymbol{b}_i &= \boldsymbol{\alpha}_0 + \boldsymbol{\Gamma}_0 \bar{\boldsymbol{x}}_{\boldsymbol{i}}' + \boldsymbol{d}_i, \\ \boldsymbol{d}_i|(\boldsymbol{x}_i, c_i) &\sim Normal(\mathbf{0}, \boldsymbol{\Omega}_0), \end{aligned} \tag{24}$$

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 $\boldsymbol{\Gamma}_0 = \mathbf{0}$, then Eq. 24 amounts to a stronger version of Eq. 10 where then $\boldsymbol{\alpha}_0 = \boldsymbol{\beta}_0$. Note that Eq. 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. As in FEP, the relationship between $\boldsymbol{x}_i$ and c_i is left completely unrestricted. Eq. 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 alternatives where the two are dependent. The two sources of heterogeneity are still allowed, through $\boldsymbol{x}_i$, to be correlated unconditionally. Allowing for conditional dependence between c_i and $\boldsymbol{b}_i$ requires some additional structure, but APE are still identified, as shown in Appendix B.

Under Eq.'s 4 and 24, it follows from properties of the lognormal distribution and the LIE that

$$
\begin{aligned}
E(y_{it}|\boldsymbol{x}_i, c_i) =& E(y_{it}|\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, c_i) \\
=& c_i \exp\left(\boldsymbol{x}_{it}\boldsymbol{\alpha}_0 + \boldsymbol{x}_{it}\boldsymbol{\Gamma}_0\bar{\boldsymbol{x}}_i' + \frac{1}{2}\boldsymbol{x}_{it}\boldsymbol{\Omega}_0\boldsymbol{x}_{it}'\right) \\
=& c_i \exp\left(\boldsymbol{x}_{it}\boldsymbol{\alpha}_0 + (\bar{\boldsymbol{x}}_i \otimes \boldsymbol{x}_{it})\text{vec}(\boldsymbol{\Gamma}_0) + \frac{1}{2}\left(\sum_{j=1}^{K}\omega_j x_{itj}^2 + 2\sum_{j=1}^{K-1}\sum_{h\neq j}^{K}\rho_{jh}x_{itj}x_{ith}\right)\right) \\
\equiv& c_i \exp\left(\boldsymbol{x}_{it}\boldsymbol{\alpha}_0 + (\bar{\boldsymbol{x}}_i \otimes \boldsymbol{x}_{it})\boldsymbol{\gamma}_0 + \frac{1}{2}\check{\boldsymbol{x}}_{it}\boldsymbol{\omega}_0\right),
\end{aligned}
\tag{25}
$$

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 $\rho_{jh} = Cov(b_j, b_h)$.

Eq. 25, along with regularity conditions, implies that FEP of y_{it} on $\boldsymbol{x}_{it}$, $\bar{\boldsymbol{x}}_i \otimes \boldsymbol{x}_{it}$, $x_{it1}^2, \ldots x_{itK}^2$, and $x_{it1}x_{it2}, \ldots x_{it,K-1}x_{itK}$ 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 Eq. 25, the unconditional means of the $\boldsymbol{b}_i$ are easy to estimate using the following implication of the LIE:

$$
\boldsymbol{\beta}_0 \equiv E(\boldsymbol{b}_i) = \boldsymbol{\alpha}_0 + \boldsymbol{\Gamma}_0\boldsymbol{\mu}_{\bar{\boldsymbol{x}}}', \tag{26}
$$

where $\boldsymbol{\mu}_{\bar{\boldsymbol{x}}} = E(\bar{\boldsymbol{x}}_i)$.

Using the lognormal distribution in the FEP setting in this way is novel, to my knowledge. It also offers the advantage of still allowing one source of heterogeneity to be correlated with $\boldsymbol{x}_i$ in an unrestricted fashion.[9] This procedure is easy to implement, as the FEP estimator is available in software packages like Stata, 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

[9] 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 $E(y_{it}|\boldsymbol{x}_i)$.

absorbed by c_i when conditioning on n_i.

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.

Finally, under this approach, binary variables and some nonlinearities prevent separate identification of elements of the structural parameter $\boldsymbol{\alpha}_0$ and elements of the coefficient variance matrix $\boldsymbol{\Omega}_0$. For example, for a binary element k of $\boldsymbol{x}_{it}$, FEP only identifies $\alpha_k + \frac{1}{2}\omega_k$. Likewise, if one allows covariance between c_i and $\boldsymbol{b}_i$, then $\boldsymbol{\alpha}_0$ is not separately identified from this covariance. This is a drawback if the structural parameters are of primary interest, but it does not matter for estimating APE, as I show in Section 3.6 for the case of binary regressors.

3.5 Adding second moment assumptions

While under our assumptions, FEP is consistent under correct specification of the conditional mean in Eq. 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 separately identify the parameters relating to binary variables.

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

$$Var\left[y_{it}|\boldsymbol{x}_i, c_i, \boldsymbol{b}_i\right] = \sigma_0 c_i \exp(\boldsymbol{x}_{it}\boldsymbol{b}_i) \tag{27}$$

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

$$\log(c_i)|\boldsymbol{x}_i, \boldsymbol{b}_i \sim Normal(\psi_1 + \bar{\boldsymbol{x}}_i\boldsymbol{\xi}_1, \sigma_a^2) \tag{28}$$

Under assumptions 4, 24, 27, and 28, it follows from the properties of the lognormal distribution, the LIE, and the Law of Total Variance that

$$E(y_{it}|\boldsymbol{x}_i) = E(y_{it}|\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i) = \exp\left[h(\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + \frac{1}{2}v(\boldsymbol{x}_{it}, \boldsymbol{\tau}_0)\right] \tag{29}$$

and

$$\begin{aligned} Var(y_{it}|\boldsymbol{x}_i) =& Var(y_{it}|\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i) \\ =& \sigma_0 \exp\left[h(\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + \frac{1}{2}v(\boldsymbol{x}_{it}, \boldsymbol{\tau}_0)\right] \\ &+ \exp\left[2h(\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + v(\boldsymbol{x}_{it}, \boldsymbol{\tau}_0)\right]\left\{\exp\left[v(\boldsymbol{x}_{it}, \boldsymbol{\tau}_0)\right] - 1\right\}, \end{aligned} \tag{30}$$

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 Eq. 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 paper. 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

$$Cov\left[y_{it}, y_{ir}|\boldsymbol{x}_i, c_i, \boldsymbol{b}_i\right] = 0, t \neq r. \tag{31}$$

Equation 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 Eq.'s 4, 24, 27, 31, and 28,

$$
\begin{aligned}
&Cov(y_{it}, y_{ir}|\boldsymbol{x}_i) = \\
&\qquad \exp\left[h(\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + h(\boldsymbol{x}_{ir}, \bar{\boldsymbol{x}}_i, \boldsymbol{\theta}_0) + \frac{1}{2}\left(v(\boldsymbol{x}_{it}, \boldsymbol{\tau}_0) + v(\boldsymbol{x}_{ir}, \boldsymbol{\tau}_0)\right)\right] \\
&\qquad\qquad \times \left\{\exp(\boldsymbol{x}_{it}\boldsymbol{\Omega}_0\boldsymbol{x}'_{ir} + \sigma_a^2) - 1\right\}. \qquad (32)
\end{aligned}
$$

3.6 Estimating average partial effects

Even though the coefficients in Eq. 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[10]

$$\frac{\partial E(y_t|\boldsymbol{x}_t, c, \boldsymbol{b})}{\partial x_{tj}} = c_i \exp(\boldsymbol{x}_t\boldsymbol{b})b_j. \qquad (33)$$

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.

$$
\begin{aligned}
&E(y_t|\boldsymbol{x}_{t\not{k}}, x_{tk} = 1, c, \boldsymbol{b}) - E(y_t|\boldsymbol{x}_{t\not{k}}, x_{tk} = 0, c, \boldsymbol{b}) \\
&\qquad = c\exp(\boldsymbol{x}_{t\not{k}}\boldsymbol{b}_{\not{k}} + b_k) - c\exp(\boldsymbol{x}_{t\not{k}}\boldsymbol{b}_{\not{k}}). \qquad (34)
\end{aligned}
$$

In this and expressions to follow, the subscript $\not{k}$ signifies that x_{tk}, $\bar{x}_k$, or their associated coefficients have been omitted from the vector or matrix.

Of course, direct study is infeasible as we do not observe c or $\boldsymbol{b}$. Therefore, this section focuses

[10] Assuming $\boldsymbol{x}_t\boldsymbol{b}_i$ is linear in $\boldsymbol{x}_t$.

on ways to average out the heterogeneity. I consider two choices as to how to proceed. The first is to maintain assumptions 24, 27, and 28 and take derivatives and differences of $E(y_{it}|\boldsymbol{x}_i)$ directly using Eq. 29, as is often shown in Wooldridge (2010). I leave these derivations for the appendix. The second option is to estimate an Average Structural Function (ASF), as proposed by Blundell and Powell (2003),. Loosely speaking, $\bar{\boldsymbol{x}}$ proxies for unobserved heterogeneity and is averaged out before taking derivatives and differences. One advantage of using the ASF is that it turns out we do not need to observe or restrict c_i to average it out.

Estimation of the Average Structural Function

The ASF is defined as:

$$ASF(\boldsymbol{x}_t) \equiv E_{\boldsymbol{v}_i}\left[c_i \exp(\boldsymbol{x}_t \boldsymbol{b}_i)\right], \tag{35}$$

where again, $\boldsymbol{x}_t$ is a fixed argument. Under Eq. 24, the L.I.E. implies

$$ASF(\boldsymbol{x}_t) = E_{c_i, \bar{\boldsymbol{x}}}\left[c_i \exp\left(\boldsymbol{x}_t \boldsymbol{\alpha}_0 + (\bar{\boldsymbol{x}}_i \otimes \boldsymbol{x}_t)\boldsymbol{\gamma}_0 + \frac{1}{2}\check{\boldsymbol{x}}_t \boldsymbol{\omega}_0\right)\right], \tag{36}$$

where the difference from Eq. 25 is that x_t is a fixed argument, and the expectation is with respect to the distribution of $(c_i, \bar{\boldsymbol{x}}_i)$.

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

$$\delta_j(\boldsymbol{x}_t) = E_{\bar{\boldsymbol{x}}}\left[\exp\left(\boldsymbol{x}_t \boldsymbol{\alpha}_0 + (\bar{\boldsymbol{x}}_i \otimes \boldsymbol{x}_t)\boldsymbol{\gamma}_0 + \frac{1}{2}\check{\boldsymbol{x}}_t \boldsymbol{\omega}_0\right)\left(\alpha_j + \bar{\boldsymbol{x}}_i \boldsymbol{\gamma}_j' + \omega_j x_{tj} + \sum_{h \neq j}^{K} \rho_{jh} x_{th}\right)\right], \tag{37}$$

where $\boldsymbol{\gamma}_j$ is the jth row and $\boldsymbol{\gamma}^j$ is the jth column of $\boldsymbol{\Gamma}_0$. For a binary x_{tk}, the APE is:

$$\begin{aligned}
\delta_k(\boldsymbol{x}_t) = & \\
& E_{c_i, \bar{\boldsymbol{x}}}\left[c_i \exp\left(\boldsymbol{x}_{t\not{k}} \boldsymbol{\alpha}_{\not{k}} + \boldsymbol{x}_{t\not{k}} \boldsymbol{\Gamma}_{\not{k}} \bar{\boldsymbol{x}}_{\not{k}} + \boldsymbol{x}_{t\not{k}} \bar{x}_{ik} \boldsymbol{\gamma}_{\not{k}}^{k} + \alpha_k + \bar{\boldsymbol{x}}_i \boldsymbol{\gamma}_k' + \frac{1}{2}\check{\boldsymbol{x}}_{\not{k}} \boldsymbol{\omega}_{\not{k}} + \frac{1}{2}\omega_k + \sum_{h \neq k}^{K} \rho_{kh} x_{th}\right)\right] \\
& - E_{c_i, \bar{\boldsymbol{x}}}\left[c_i \exp\left(\boldsymbol{x}_{t\not{k}} \boldsymbol{\alpha}_{\not{k}} + \boldsymbol{x}_{t\not{k}} \boldsymbol{\Gamma}_{\not{k}} \bar{\boldsymbol{x}}_{\not{k}} + \boldsymbol{x}_{t\not{k}} \bar{x}_{ik} \boldsymbol{\gamma}_{\not{k}}^{k} + \frac{1}{2}\check{\boldsymbol{x}}_{\not{k}} \boldsymbol{\omega}_{\not{k}}\right)\right]
\end{aligned} \tag{38}$$

Since the α_k and ω_k enter Eq. 38 as the linear combination $\alpha_k + \frac{1}{2}\omega_k$, they do not need to be separately identified. To estimate the APE, choose the fixed value $\boldsymbol{x}_t$, replace the expectations with the sample average operator over i, and insert FEP parameter estimates. Finally, a proxy exists for c_i that is sufficient for averaging it from the APE. It corresponds to the Poisson QMLE when treating the c_i as parameters to estimate, and is given in Eq. 39.

$$\hat{c}_i \equiv \frac{\sum_{t=1}^{T} y_{it}}{\sum_{t=1}^{T} \exp\left(\boldsymbol{x}_{it}\hat{\boldsymbol{\alpha}} + (\bar{\boldsymbol{x}}_i \otimes \boldsymbol{x}_{it})\hat{\boldsymbol{\gamma}} + \frac{1}{2}\check{\boldsymbol{x}}_{it}\hat{\boldsymbol{\omega}}\right)} \tag{39}$$

It is well-known that Poisson QMLE that treats c_i as parameters is not only consistent, but equivalent to FEP. Martin (2017) shows that even with T fixed, sample averages where c_i enters multiplicatively (as in Eq.'s 36, 37, and 38) can be consistently estimated with $\hat{c}_i$ standing in for c_i. If researchers are interested in a single APE, Eq.'s 37 and 38 might be averaged again across the distribution of $\boldsymbol{x} + it$. Standard errors are available via the delta method, but the panel bootstrap may also be convenient. Finally, it is easy to see that if assuming that $\boldsymbol{b}_i$ is independent of $\boldsymbol{x}_i$ (conditional on c_i), estimation simplifies considerably, as terms depending on $\bar{\boldsymbol{x}}_i$ drop out. In this case, if the target statistic is a single APE, the simplest method is to use the direct approach and average once over sample version of $c_i \exp(\boldsymbol{x}_{it}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{x}_{it}\boldsymbol{\Omega}\boldsymbol{x}'_{it})$.

4 Monte Carlo

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:

$$y_{it}|(\boldsymbol{x}_i, \boldsymbol{w}_i, c_i, b_{i1}, b_{i2}) \sim \text{Poisson}\left[c_i \exp(b_{i1}x_{it} + b_{i2}w_{it})\right], \tag{40}$$

$$\log(c_i) \sim Normal(0, 1/16) \tag{41}$$

$$\begin{aligned} x_{it} &= \log(c_i) + .5x_{i,t-1} + v_{it},\ t > 1 \\ x_{i1} &= \log(c_i)_i + v_{i1},\ v_{it} \sim N(0, 1/2) \end{aligned} \tag{42}$$

$$w_{it} = \mathbf{1}\left[x_{it} + h_{it} > 0\right], \; h_{it} \sim N(0, 1/2) \tag{43}$$

$$\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim Normal\left[\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} \omega_1^2 & \rho \\ \rho & \omega_2^2 \end{pmatrix}\right] \tag{44}$$

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 Eq.'s 27 and 28.[11] I set $\omega_1 = \omega_2 = \omega$ but do not assume equal variance in estimation. In each case, one thousand replications were used.

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

Table 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: 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 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 Eq. 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 the FEP, FEP2 or NQML parameter estimates and averaging once over $\{\boldsymbol{x}_{it}, \hat{c}_i\}$. Table 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.

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 paper to work well. To explore this, I generate the responses using random coefficients of different distributions.

$$b_{ij2} = 1 + \omega\left((u_{j2} - 0.5)/\sqrt{1/12}\right),\, u_{j2} \sim U(0,1) \tag{45}$$

$$b_{ij3} = 1 + \omega\left((u_{j3} - 4)/\sqrt{8}\right),\, u_{j3} \sim \chi^2_4 \tag{46}$$

$$b_{ij4} = 1 + \omega\left(u_{j4}/\sqrt{5/3}\right),\, u_{j4} \sim t_5 \tag{47}$$

$$b_{ij5} = 1 + \omega\left(u_{j5} - 1\right), u_{j5} \sim \textit{Exponential}\ (1) \tag{48}$$

$$b_{ij6} \sim \textit{Gamma}\ (1/\omega^2, \omega^2) \tag{49}$$

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_{itj} is generated as in Eq. 42.

$$y_{it}|(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, c_i, b_{i1h}, b_{i2h}) \sim \text{Poisson}\left[c_i \exp(b_{i1h}x_{it1} + b_{i2h}x_{it2})\right] \tag{50}$$

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 for significance was then performed on x_{t1}^2, x_{t2}^2, and $x_{t1}x_{t2}$.[12] 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 3 shows that as expected, rejection probabilities increase with ω when the coefficients are normal, and are quite high when ω is large. 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.

[12]The Wald and LM tests are asymptotically equivalent in this case (Wooldridge, 2010).

Table 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

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, 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.

5.1 Description of the data and CRC model

This paper's contribution is to look at an updated panel and consider a new specification motivated by random coefficients. The population of interest 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 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 4 and 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,

[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.

and Hausman (1986) in terms of net sales, though they achieve 90% coverage of total R&D.

Table 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 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 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.

Table 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

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

interest are the logs of current R&D and up to three lags. I include year dummies, but assume their coefficients are constant.

$$E\left[patents_{it}|\log(R_{i1}),\ldots,\log(R_{iT}),\delta_t,c_i,\boldsymbol{b}_i\right]=c_i\exp\left(\sum_{s=0}^{\tau}b_{i,s}\log(R_{i,t-s})+\delta_t\right), \tag{51}$$

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

$$\boldsymbol{b}_i|(\log(R_{i,t-0}),\ldots,\log(R_{i,t-\tau}),\delta_t,c_i)\sim Normal(\boldsymbol{\alpha}+\boldsymbol{\gamma}'\overline{\log(R)}_i,\boldsymbol{\Omega}), \tag{52}$$

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.

5.2 Estimation and testing under the null of constant coefficients

Table 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 previous findings, these estimates imply that correlation between patents and current-period 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 less than half of those estimated in previous studies (i.e. Hausman, Hall, and Griliches (1984) estimate elasticities of 0.35 and 0.43 in no-lag and five-lag models, respectively).[15]

[15] The choice of three lags may seem somewhat arbitrary. Shorter-lag models tend to show truncation bias of the sort described by Hall, Griliches, and Hausman (1986), while longer models tended to be imprecisely estimated. However, get very similar in magnitude results to Gurmu and Perez-Sebastian (2008) when I replicate their four-lag FEP model over 1982-1992, so it is possible that the nature of the patent-R&D relationship for large firms changed

Table 7: Results for traditional estimators

VARIABLES	(1) PQML 1	(2) PQML 2	(3) FEOLS 1	(4) FEOLS 2	(5) FEP 1	(6) 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 8: LM tests of neglected slope heterogeneity

Panel A: FEP 1 (No lag model)

	χ^2	d.f.	p value
Chesher	9.440	1	0.0021
QML	2.411	1	0.1205

Panel B: FEP 2 (Three lag model)

	χ^2	d.f.	p value
Chesher	36.621	4	2.16e-07
QML	7.585	4	0.1080

Table 8 displays the results of LM tests for neglected slope heterogeneity applied to the no-lag and the three-lag models following FEP estimation. The label "Chesher" denotes the test statistic derived under full distributional assumptions, while "QML" refers to the test statistic proposed in this paper. As previously discussed, Chesher's test amounts to testing diagonal elements of the information matrix. In both models, this test strongly rejects the null of constant coefficients. However, as previously discussed, failure of the conditional Poisson assumption and neglected serial correlation would lead to information matrix inequality without causing inconsistency in FEP. The test I propose for the QMLE setting restricts attention to the conditional mean function, which I show in Section 3.3 to be equivalent to testing the significance of squared regressors. In both the no-lag and three-lag models, we fail to reject the null of constant coefficients at the 10% level.

5.3 Estimating the CRC model

Section 3 and Section 5 imply that neglected slope heterogeneity could be a source of inconsistency in the FEP estimates reported in the previous subsection. While the prosed LM tests fail to reject the null of constant coefficients, I illustrate the proposed CRC estimation procedure by presenting results in Table 9. Models CRCFEP1 through CRCFEP 5 vary the lag length and assumptions about the covariance matrix of the random slopes. 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 $\boldsymbol{\Omega}$ is diagonal. I allow covariance between the random slopes in column (5).

As expected from the previous subsection, these data do not provide strong evidence of slope heterogeneity. In the no-lag models, none of the additional terms is statistically significant. The evidence is more 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 $\log(R)$ and its lags are jointly significant, none of the interactions and squares are individually significant. Neither the set of interactions or the set of squares is jointly significant in column (4). In column (5), the entire set of added heterogeneity terms (the interactions, squares, and cross-products) are jointly marginally significant ($p = 0.08$). However, the terms associated with $\boldsymbol{\Omega}$ and $\boldsymbol{\gamma}$ are each jointly insignificant. Therefore, while there may be marginal evidence of heterogeneity, I cannot parse it into its components. Multicollinearity

in the intervening decade.

Table 9: 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)
$\log(R_{-3}) \times \overline{\log(R_0)}$			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)
Observations	5,968	5,968	3,510	3,510	3,510
Number of pdpco	746	746	702	702	702

Clustered standard errors in parentheses. Year dummies included in all specifications.
*** p<0.01, ** p<0.05, * p<0.1

Table 10: 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 $\overline{\log(R)}$

is likely playing a role in reducing the precision of the individual estimates. $log(R)$ is highly correlated with its lags ($\rho > 0.9$), its time average ($\rho = 0.98$), the interactions with $\overline{\log(R)}_i$ ($\rho > 0.8$), and its current and lagged squares ($\rho > 0.8$).

As implied by Eq. 52, the estimator for the average elasticity with respect to R_{t-s} is given by

$$\widehat{\beta}_{-s} = \widehat{\alpha}_{s+1} + \widehat{\gamma}_{s+1}\overline{\log(R)}, \tag{53}$$

where $\overline{\log(R)} = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \log(R_{it})$. Model CRCFEP 3 (column 3 of 9) show slight evidence that the slopes depend the mean of $log(R)$ if this relationship is restricted to be deterministic. For illustration, I estimate the mean elasticities for this model and present them in Table 10.[16]

At face value, 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. However, given the multicollinearity between the $log(R)$ and $log(R) \times \bar{(}log(R)$ terms, the finding is suspect. Furthermore, 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

[16]The mean elasticities derived from columns (4) and (5) are statistically insignificantly different from zero.

if the were not granted before 2006. Moreover, the estimates are somewhat sensitive to including additional moment conditions as proposed in Wooldridge (1999).

Overall, the permutations of the CRC model considered here do not offer convincing evidence of slope heterogeneity, which is not surprising given the LM test in the previous subsection failed to reject the null of constant slopes. This contrasts starkly with the classical LM test, which rejects strongly in favor of coefficient heterogeneity. Given weakness of evidence for a more flexible conditional mean function, it seems plausible that the classical LM test is picking up some other feature of the data like conditional serial dependence or higher moments that differ from the Poisson's.

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 the proposed test to the U.S. manufacturing industry shows little evidence of slope heterogeneity when restricting attention to the conditional mean function, in contrast to the classical test, which rejects strongly. Accordingly, estimates of the CRC model corresponding to the slope heterogeneity tend to be jointly insignificant, with the exception of when the slopes are assumed to be deterministic functions of the covariates, in which case the added terms are marginally significant. 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.

References

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.

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.

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

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. 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.

Gu, Jiaying (2016). "Neyman's C(α) Test For Unobserved Heterogeneity". In: *Econometric Theory* 32.6, pp. 1483–1522.

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., H.R. Moon, and C. Snider (2017). "LM test of neglected correlated random effects and its application". In: *Journal of Business & Economic Statistics* forthcoming.

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.

Jochmann, Markus and Roberto León-González (2004). "Estimating the demand for health care with panel data: a semiparametric Bayesian approach". In: *Health Economics* 13.10, pp. 1003–1014.

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.

Martin, Robert S (2017). "Estimation of average marginal effects in multiplicative unobserved effects panel models". In: *Economics Letters* 160, pp. 16–19.

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

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.

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. (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.

A Derivations of test statistics

Derivations from Section 3.2

From section 3.2, the score of Eq. 13 evaluated at $\mathbf{\Lambda} = \mathbf{0}$ is identically zero. Assuming we can pass the derivative through the integral, we can work out the following:

$$\nabla_\Lambda \ell_i(\boldsymbol{\beta}, \mathbf{\Lambda}) = \frac{\iint_{\mathbb{R}^K} h_{it} \left[\prod_{t=1}^T p_t(\boldsymbol{x}_i, \boldsymbol{b}_i)^{y_{it}}\right] \left[\sum_{t=1}^T y_{it} \boldsymbol{u}_i' \otimes q_t(\boldsymbol{x}_i, , \boldsymbol{b}_i)\right] f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i}{\iint_{\mathbb{R}^K} f(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{u}_i, c_i, n_i) f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i} \tag{54}$$

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:

$$\begin{aligned}
\nabla_\Lambda \ell_i(\boldsymbol{\beta}, \mathbf{\Lambda})\Big|_{\mathbf{\Lambda}=\mathbf{0}} &= \frac{h_{it} \left[\prod_{t=1}^T p_t(\boldsymbol{x}_i, \boldsymbol{\beta})^{y_{it}}\right] \left[\sum_{t=1}^T y_{it} \iint_{\mathbb{R}^K} \boldsymbol{u}_i' \otimes q_t(\boldsymbol{x}_i, \boldsymbol{\beta}) f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i\right]}{h_{it} \left[\prod_{t=1}^T p_t(\boldsymbol{x}_i, \boldsymbol{\beta})^{y_{it}}\right] \iint_{\mathbb{R}^K} f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i} && (55) \\
&= \sum_{t=1}^T y_{it} E\left[\boldsymbol{u}_i' \otimes q_t(\boldsymbol{x}_i, \boldsymbol{b}_i)\right] && (56) \\
&= \mathbf{0}.
\end{aligned}$$

The second equality uses that $\iint_{\mathbb{R}^K} f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i = 1$, 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 Eq. 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:

$$\frac{\partial \ell_i(\boldsymbol{\beta}, \boldsymbol{\lambda})}{\partial \lambda_j}\Big|_{\boldsymbol{\lambda}=0} = \left\{ \frac{1}{2\sqrt{\lambda_j}} \left[\sum_{t=1}^T y_{it} q_{tj}(\boldsymbol{x}_i, \boldsymbol{\beta})\right] \iint_{\mathbb{R}^K} u_{ij} f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i \right\}_{\lambda_j = 0} \tag{57}$$

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:

$$\frac{\frac{1}{2\sqrt{\lambda_j}} \iint_{\mathbb{R}^K} h_{it} \left[\prod_t p_t(\boldsymbol{x}_i, \boldsymbol{b}_i)^{y_{it}}\right] \left\{\sum_t y_{it} r_{tj}(\boldsymbol{x}_i, \boldsymbol{b}_i) + \left[\sum_t y_{it} q_{tj}(\boldsymbol{x}_i, \boldsymbol{b}_i)\right]^2\right\} u_{ij}^2 f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i}{2\left(\frac{1}{2\sqrt{\lambda_j}}\right) \iint_{\mathbb{R}^K} h_{it} \left[\prod_t p_t(\boldsymbol{x}_i, \boldsymbol{b}_i)^{y_{it}}\right] f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i}, \tag{58}$$

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 $\boldsymbol{\lambda} = \boldsymbol{0}$ ($\boldsymbol{b}_i = \boldsymbol{\beta}_0$). 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 Eq. 15.

Derivations from Section 3.3

As before, the restricted score of Eq. 21 is identically zero.

$$\begin{aligned}
\nabla_{\boldsymbol{\Lambda}} \ell_i(\boldsymbol{\beta}, \boldsymbol{\Lambda}) &= \sum_{t=1}^{T} y_{it} \left[\frac{\nabla_{\boldsymbol{\Lambda}} p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})}{p_t(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})}\right] \\
&= \sum_{t=1}^{T} y_{it} \frac{\sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta} + m_r(\boldsymbol{x}_i, \boldsymbol{\Lambda})) \left[\nabla_{\boldsymbol{\lambda}} m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda}) - \nabla_{\boldsymbol{\lambda}} m_r(\boldsymbol{x}_i, \boldsymbol{\Lambda})\right]}{\sum_{r=1}^{T} \exp(\boldsymbol{x}_{ir}\boldsymbol{\beta} + m_r(\boldsymbol{x}_i, \boldsymbol{\Lambda}))},
\end{aligned} \tag{59}$$

$$\nabla_{\boldsymbol{\lambda}} m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda}) = \frac{\iint_{\mathbb{R}^K} \exp\left(\boldsymbol{x}_{it}\boldsymbol{\Lambda}\boldsymbol{u}_i\right) (\boldsymbol{u}_i' \otimes \boldsymbol{x}_{it}) f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i}{\iint_{\mathbb{R}^K} \exp(\boldsymbol{x}_{it}\boldsymbol{\Lambda}_0 \boldsymbol{u}_i) f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i}. \tag{60}$$

The complication arises because

$$\nabla_{\boldsymbol{\lambda}} m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda})\Big|_{\boldsymbol{\Lambda}=\boldsymbol{0}} = \frac{\iint_{\mathbb{R}^K} (\boldsymbol{u}_i' \otimes \boldsymbol{x}_{it}) f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i}{\iint_{\mathbb{R}^K} f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i} = \boldsymbol{0}, \tag{61}$$

which implies

$$\nabla_{\boldsymbol{\Lambda}} \ell_i(\boldsymbol{\beta}, \boldsymbol{\Lambda})\Big|_{\boldsymbol{\Lambda}=\boldsymbol{0}} = \boldsymbol{0}. \tag{62}$$

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

$$\nabla_{\lambda_j} m_t(\boldsymbol{x}_i, \boldsymbol{\Lambda}) = \left\{\iint_{\mathbb{R}^K} \exp(\boldsymbol{x}_{it}\boldsymbol{\Lambda}_0 \boldsymbol{u}_i) f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i\right\}^{-1} \frac{\iint_{\mathbb{R}^K} \exp\left(\boldsymbol{x}_{it}\boldsymbol{\Lambda}\boldsymbol{u}_i\right) x_{itj} u_{ij} f(\boldsymbol{u}_i) \, \mathrm{d}\boldsymbol{u}_i}{2\sqrt{\lambda_j}}. \tag{63}$$

When evaluated at $\boldsymbol{\lambda} = \boldsymbol{0}$, the second factor of Eq. 63 has the form 0/0.

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

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

$$=\frac{x_{itj}^2\iint_{\mathbb{R}^K}u_{ij}^2f(\boldsymbol{u}_i)\,\mathrm{d}\boldsymbol{u}_i}{2}$$

$$=\frac{1}{2}x_{itj}^2 \tag{64}$$

Plugging these limits in into the expression for $\nabla_{\boldsymbol{\Lambda}}\ell_i(\boldsymbol{\beta},\mathbf{0})$, we get Eq. 23.

B APE when intercept and slope heterogeneity are conditionally dependent

Here I sketch a model that, in contrast to the CRC model presented in Section 3, allows covariance between the random intercept and random slopes conditional on the covariates. For notational simplicity, I consider the case where the mean of the $\boldsymbol{b}_i$ does not depend on $\boldsymbol{x}_i$, and suppress the "0" subscripts indicating true population values. [17]. Defining $v_i=\ln(c_i)$, write the entire $(K+1)$-vector of heterogeneity as

$$\begin{pmatrix} v_i \\ \boldsymbol{b}_i \end{pmatrix}=\begin{pmatrix} g_1(\boldsymbol{x}_i)+a_i \\ \boldsymbol{\beta}+\boldsymbol{d}_i \end{pmatrix}, \tag{65}$$

and assume

$$\begin{pmatrix} a_i \\ \boldsymbol{d}_i \end{pmatrix}|\boldsymbol{x}_i\sim Normal\left(\mathbf{0},\begin{pmatrix} \omega_a & \boldsymbol{\omega}_{\boldsymbol{b}a}' \\ \boldsymbol{\omega}_{\boldsymbol{b}a} & \boldsymbol{\Omega} \end{pmatrix}\right). \tag{66}$$

It follows that:

$$E(y_{it}|\boldsymbol{x}_i)=\exp\left(g_1(\boldsymbol{x}_i)+\frac{1}{2}\omega_a\right)\exp\left(\boldsymbol{x}_{it}\left[\boldsymbol{\beta}+\boldsymbol{\omega}_{\boldsymbol{b}a}\right]+\frac{1}{2}\boldsymbol{x}_{it}\boldsymbol{\Omega}\boldsymbol{x}_{it}'\right) \tag{67}$$

$$=w_i\exp\left(\boldsymbol{x}_{it}\left[\boldsymbol{\beta}+\boldsymbol{\omega}_{\boldsymbol{b}a}\right]+\frac{1}{2}\boldsymbol{x}_{it}\boldsymbol{\Omega}\boldsymbol{x}_{it}'\right), \tag{68}$$

[17] Introducing the conditional covariance comes at the cost of modeling the conditional distribution of c_i, but its conditional mean, at least, is still fully general

where $w_i \equiv \exp\left(g_1(\boldsymbol{x}_i) + \frac{1}{2}\omega_a\right)$. It is immediately clear that FEP of y_{it} on $\boldsymbol{x}_{it}$ and $(\boldsymbol{x}_{it} \otimes \boldsymbol{x}_{it})$ does not require assuming the form of $g_1()$, but precludes separate estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\omega_{ba}}$ unless we assume $cov(a, \boldsymbol{b}) = 0$. However, APE are still identified, as they depend only on the sum $\boldsymbol{\beta} + \boldsymbol{\omega_{ba}}$. For example, given a fixed value of the regressors $\boldsymbol{x}_t$, the APE of a continuous variable x_{tj} is

$$\frac{\partial E(y_{it}|\boldsymbol{x}_t)}{\partial x_{tj}} = w_i \exp\left(\boldsymbol{x}_t\left[\boldsymbol{\beta} + \boldsymbol{\omega_{ba}}\right] + \frac{1}{2}\boldsymbol{x}_t\boldsymbol{\Omega}\boldsymbol{x}_t'\right)\left([\beta_j + \omega_{\boldsymbol{b}aj}] + \omega_j + \sum_{h\neq j}^{K}\omega_{jh}x_{th}\right) \tag{69}$$

where $\omega_{jh} = cov(b_j, b_h)$, and $\omega_{\boldsymbol{b}aj} = cov(b_j, a)$. I emphasize with square brackets the fact that this quantity depends only on the sum $\boldsymbol{\beta} + \boldsymbol{\omega_{ba}}$. Applying Martin (2017), estimation of the ASF or a single APE can proceed as discussed in Section 3.

C Derivations of APEs using the direct approach

The direct approach consists of taking derivatives and differences of Eq. 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:

$$\begin{aligned}
\delta_j(\boldsymbol{x}) =& \frac{\partial E(y_t|\boldsymbol{x})}{\partial x_{tj}} \\
=& \exp\left(h(\boldsymbol{x}_t, \bar{\boldsymbol{x}}, \boldsymbol{\theta}_0) + \frac{1}{2}v(\boldsymbol{x}_t, \boldsymbol{\tau}_0)\right)\left(\xi_j/T + \alpha_j + \bar{\boldsymbol{x}}\boldsymbol{\gamma}_j' + \frac{1}{T}\boldsymbol{x}_t\boldsymbol{\gamma}^j + \omega_j x_{tj} + \sum_{h\neq j}^{K}\rho_{jh}x_{th}\right),
\end{aligned} \tag{70}$$

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},

$$\begin{aligned}
\delta_k(\boldsymbol{x}) =& E\left(y_t|\boldsymbol{x}_{\not k}, \{x_{sk}\}_{s\neq t}^{T}, x_{tk} = 1\right) - E\left(y_t|\boldsymbol{x}_{\not k}, \{x_{sk}\}_{s\neq t}^{T}, x_{tk} = 0\right) \\
=& \exp\left(z(\boldsymbol{x}_{t\not k}, \bar{\boldsymbol{x}}_{\not k}, \boldsymbol{\theta}_{\not k}, \boldsymbol{\tau}_{\not k}) + \xi_j\bar{x}_{tk}^{(1)} + \alpha_k + \bar{\boldsymbol{x}}_{\not k}\boldsymbol{\gamma}_{k\not k}' + \gamma_{kk}\bar{x}_{tk}^{(1)} + \boldsymbol{x}_{t\not k}\bar{x}_{tk}^{(1)}\boldsymbol{\gamma}_{\not k}^{k} + \frac{1}{2}\omega_k + \sum_{h\neq k}^{K}\rho_{kh}x_{th}\right) \\
&- \exp\left(z(\boldsymbol{x}_{t\not k}, \bar{\boldsymbol{x}}_{\not k}, \boldsymbol{\theta}_{\not k}, \boldsymbol{\tau}_{\not k}) + \xi_j\bar{x}_{tk}^{(0)} + \boldsymbol{x}_{t\not k}\bar{x}_{tk}^{(0)}\boldsymbol{\gamma}_{\not k}^{k}\right),
\end{aligned} \tag{71}$$

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.