Brown, Nicholas; Wooldridge, Jeffrey M.

Working Paper

More efficient estimation of multiplicative panel data models in the presence of serial correlation

Queen's Economics Department Working Paper, No. 1497

Provided in Cooperation with:
Queen's University, Department of Economics (QED)

Suggested Citation: Brown, Nicholas; Wooldridge, Jeffrey M. (2023) : More efficient estimation of multiplicative panel data models in the presence of serial correlation, Queen's Economics Department Working Paper, No. 1497, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
https://hdl.handle.net/10419/281101

Queen's Economics Department Working Paper No. 1497

More Efficient Estimation of Multiplicative Panel Data Models in the Presence of Serial Correlation

Nicholas Brown
Queen's University

Jeffrey Wooldridge
Michigan State University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

2-2023

More Efficient Estimation of Multiplicative Panel Data Models in the Presence of Serial Correlation

Nicholas L. Brown
Department of Economics
Queen's University

Jeffrey M. Wooldridge*
Department of Economics
Michigan State University

* Corresponding author. E-mail: wooldri1@msu.edu.

Abstract: We provide a systematic approach in obtaining an estimator asymptotically more efficient than the popular fixed effects Poisson (FEP) estimator for panel data models with multiplicative heterogeneity in the conditional mean. In particular, we derive the optimal instrumental variables under appealing "working" second moment assumptions that allow underdispersion, overdispersion, and general patterns of serial correlation. Because parameters in the optimal instruments must be estimated, we argue for combining our new moment conditions with those that define the FEP estimator to obtain a generalized method of moments (GMM) estimator no less efficient than the FEP estimator and the estimator using the new instruments. A simulation study shows that the overidentfied GMM estimator behaves well in terms of bias and it often delivers nontrivial efficiency gains – even when the working second-moment assumptions fail. We apply the new estimator to modeling firm patent filings and spending on R&D, and find nontrivial reductions in standard errors using the new estimator.

Keywords: Fixed effects Poisson; serial correlation; optimal instruments; generalized method of moments

JEL Classification Code: C23

1. Introduction

The fixed effects Poisson (FEP) estimator was originally developed by Hausman, Hall, and Griliches (1984) (hereafter, HHG) in their study of the effects of firm-level R&D spending on patent filings. HHG used the method of conditional maximum likelihood estimation (CMLE) to estimate the parameters in the conditional mean. In deriving the CMLE, HHG assumed that, conditional on the unobserved heterogeneity and the history of the covariates, the outcome variable is independent over time with a Poisson distribution. HHG showed that, conditional on the covariates and the sum of the counts over time, the joint distribution of the counts is multinomial and does not depend on the heterogeneity. Therefore, standard maximum likelihood theory applies, and the asymptotic theory assuming a fixed number of time periods is standard. Hahn (1997) verified that the FEP estimator achieves the semiparametric efficiency bound under the full distributional and conditional independence assumptions.

Wooldridge (1999) showed that the consistency of the FEP estimator only requires correct specification of the conditional mean function up to a multiplicative heterogeneity term. In particular, any kind of variance is allowed along with any kind of serial dependence. In fact, the outcome variable need not even be a count variable: it can be any nonnegative outcome, including a continuous outcome or corner solution response. Thus, the FEP estimator is to multiplicative panel data models what the linear FE estimator is to linear models with additive heterogeneity.

Under the weak assumption that the conditional mean function is differentiable in the parameters, Wooldridge (1999) established Fisher consistency of the FEP. Specifically, Wooldridge showed that the score has a zero conditional mean (evaluated at the true parameter value) when the structural conditional mean – that is, conditioned on unobserved heterogeneity

– is correctly specified. The zero conditional mean property of the score leads to additional moment conditions that can be exploited in generalized method of moments (GMM) estimation to obtain estimators asymptotically more efficient than the FEP estimator. Unfortunately, the extra moment conditions proposed by Wooldridge (1999) are essentially ad hoc: they are not based on any notion of optimality. Consequently, the GMM approach to estimating multiplicative panel data models has not caught on: FEP estimation with the fully robust standard errors derived in Wooldridge (1999) is much more common. Some recent examples include McCabe and Snyder (2014, 2015), Schlenker and Walker (2016), Krapf, Ursprung, and Zimmermann (2017), Castillo, Mejia, and Restrepo (2018), and Williams, Burnap, Javed, Liu, and Ozalp (2020).

Given that the FEP estimator is fully robust to distributional misspecification and serial independence, it is natural to wonder about its asymptotic efficiency under assumptions weaker than the full set of assumptions used by Hahn (1997). Recently, Verdier (2018) showed that the Poisson distributional assumption and conditional independence are not necessary for the FEP estimator to achieve Chamberlain's (1987, 1992) efficiency bound. In particular, Verdier (2018) showed that it is sufficient to impose the Poisson assumption that the variance equals the mean and that the outcomes are serially uncorrelated conditional on heterogeneity and the covariates. While weaker than the HHG assumptions, they are still restrictive. The assumption that the variance equals the mean, even after conditioning on unobserved heterogeneity, is very special. For example, the most common parameterization of the gamma distribution violates equality of the variance and mean. Moreover, serial correlation in the idiosyncratic errors of linear unobserved effects models is pervasive, and it is known how to exploit serial correlation in fixed effects versions of generalized least squares (GLS) to

improve efficiency over the usual fixed effects estimator – see, for example, Im, Ahn, Schmidt, and Wooldridge (1999). It seems natural to search for analogous improvements over the FEP estimator in the presence of serial correlation and more flexible variance-mean relationships.

In this paper, we relax the second moment assumptions that are implied by the traditional HHG assumptions and derive the optimal instruments, thereby showing how to obtain an estimator that achieves Chamberlain's (1992) lower bound. Our efficiency result is new, and includes the Verdier (2018) result as a special case. The variance assumption we use to derive the optimal instruments is appealing because, conditional on the observed covariates and unobserved heterogeneity, it allows for underdispersion (relative to the Poisson) or overdispersion. In the spirit of the popular generalized estimating equations (GEE) approach – see Liang and Zeger (1986) – we assume constant conditional correlations, but allow for any pattern of serial correlation. One important difference from the GEE literature is that our assumptions are more "structural" in the sense that we state the second moment assumptions conditional on the unobserved heterogeneity, consistent with the idea that in the conditional expectation we want to control for unobserved heterogeneity. This is analogous to the linear model with an additive, unobserved effect when the working correlation matrix of the idiosyncratic errors is assumed to be constant but is otherwise unrestricted.

In order to obtain parametric forms for the optimal instruments, we supplement the flexible second moment assumptions for the response variable with moment assumptions about the multiplicative heterogeneity. These parametric assumptions are fairly flexible and are commonly used in the literature, particularly in traditional and correlated random effects environments when one needs to impose distributional assumptions on the heterogeneity in order to obtain consistent estimators. Here, we impose first and second moment assumptions in

order to obtain the optimal instruments.

We must emphasize that the estimator based on the optimal instruments – which we refer to as the "generalized FEP (GFEP) estimator" – does not require any assumptions for consistency and asymptotic normality beyond those used by the FEP estimator. That our new estimator is just as robust as the FEP estimator in terms of consistency is important, as it is unfair to claim efficiency improvements if the new estimator is not as robust as the popular, robust FEP estimator. In order to emphasize the robustness of our estimator, we use the term "working" assumptions when referring to assumptions used only to obtain the optimal instruments. If the working assumptions are correct, then we have a just identified estimator that is more efficient than the FEP estimator.

If any of the working assumptions are incorrect, the "optimal" instrumental variables (IVs) are no longer optimal, and so the GFEP no longer achieves Chamberlain's lower bound. Therefore, we have two estimators that are consistent under the same assumptions but efficient under different working assumptions. To ensure that we have an estimator that is at least as efficient than both the FEP estimator and the GFEP estimator, we combine the two sets of moment conditions. With K parameters this gives K overidentifying restrictions. The overidentifying restrictions are useful for testing the conditional mean specification – not the working assumptions, as those are not being used for consistency.

To summarize, this paper has three primary contributions. First, we relax the second moment assumptions implied by the traditional fixed effects Poisson setting and obtain optimal instruments under an appealing set of second moment working assumptions, including allowing for general patterns of serial correlation. Second, we operationalize the estimator by imposing additional working assumptions on moments of the heterogeneity distribution,

resulting in a GMM estimator that is computationally simple and is guaranteed to be asymptotically more efficient than both the FEP estimator and the GFEP estimator. Third, we significantly relax the conditions under which the FEP estimator achieves the asymptotic variance lower bound, allowing for both underdispersion and overdispersion in the variance conditional on observed covariates and unobserved heterogeneity.

The underlying asymptotic theory in this paper is for the microeconometric setting that treats the number of time periods, T, as fixed, and lets the cross section dimension, N, increase without bound. We assume random sampling in the cross section dimension but impose no restrictions on the time series dependence. We do not provide formal regularity conditions because the asymptotic theory is standard. We do assume smoothness so that certain derivatives – in particular, that of the conditional mean function – exist and are continuous.

The rest of the paper is organized as follows. Section 2 presents the conditional mean model and summarizes the consistency result for the FEP estimator. Section 3 derives the optimal instruments under two working variance assumptions, including an unrestricted (but constant) conditional correlation matrix. Section 4 shows how to implement the GFEP estimator and the GMM estimator that combines the two sets of moment conditions. Section 5 provides promising simulation evidence comparing the FEP, GFEP, and GMM estimators under serial correlation with both underdispersion and overdispersion in the variance. In Section 6 we apply the new estimators to a firm-level data set on patent filings and R&D spending. Section 7 contains concluding remarks.

2. Model and Background

We consider a balanced panel data setting where, for each i, $\{(y_{it}, \mathbf{x}_{it}, c_i) : t = 1, 2, \ldots, T\}$ is a random draw from the population. We observe the nonnegative response variable $y_{it} \geq 0$

and $\mathbf{x}_{it}$, a $1 \times K$ vector. The scalar c_i is the unobserved heterogeneity. As is usual in fixed effects environments, the elements of $\mathbf{x}_{it}$ must have variation across t for at least some population units. Typically $\mathbf{x}_{it}$ would include dummy variables indicating different time periods to allow for flexible aggregate time effects. The entire observed history of the covariates is $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$. As mentioned in the introduction, we are treating T as fixed in the asymptotic analysis. Therefore, because we assume random sampling in the cross section, relevant assumptions can be stated for a random draw i from the population.

The substantive assumptions that we make throughout the paper are that the model of the conditional mean is correctly specified, the heterogeneity is multiplicative, and the covariates are strictly exogenous conditional on c_i. These are all captured by the following.

Assumption Conditional Mean (CM): For $t = 1, \ldots, T$ and some $\boldsymbol{\beta}_o \in \mathbb{R}^P$,

$$\mathrm{E}(y_{it}|\mathbf{x}_i, c_i) = \mathrm{E}(y_{it}|\mathbf{x}_{it}, c_i) = c_i m_t(\mathbf{x}_{it}, \boldsymbol{\beta}_o), \tag{2.1}$$

where $m_t(\mathbf{x}_t, \cdot) \geq 0$ is continuously differentiable on $\mathbb{R}^P$ for all $\mathbf{x}_t \in \mathcal{X}_t$, the support of $\mathbf{x}_{it}$. □

As discussed in Wooldridge (1999), for consistency of the FEP estimator one can get by with continuity over the parameter space, but we impose assumptions that imply asymptotic normality and easy calculation of asymptotic efficiency bounds. See Newey and McFadden (1994) or Wooldridge (2010, Chapter 12) for formal regularity conditions. In terms of smoothness, assuming $m_t(\mathbf{x}_{it}, \cdot)$ is twice continuously differentiable is sufficient and is almost always true in practice.

The leading case of the conditional mean function is

$$\mathrm{E}(y_{it}|\mathbf{x}_{it}, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}_o), \tag{2.2}$$

where $\mathbf{x}_{it}$ can include time period dummies to allow different intercepts inside the exponential

function. Naturally, $\mathbf{x}_{it}$ can also include nonlinear functions of underlying explanatory variables, including squares and interactions. Given the choice in (2.2), $P = K$, but we also allow more general mean functions. Because we want to allow arbitrary dependence between c_i and $\mathbf{x}_{it}$, we need time variation in the latter for at least some units in the population. This permits, for example, interactions among variables that have some time variation and others that do not.

Strict exogeneity conditional on the unobserved effect c_i is implied by the first equality in (2.1). This assumption is restrictive – for example, it rules out lagged dependent variables – but it is much less restrictive than the strict exogeneity assumption typically used in the GEE literature because of conditioning on c_i. In the typical GEE approach the strict exogeneity assumption is stated as $\mathrm{E}(y_{it}|\mathbf{x}_i) = \mathrm{E}(y_{it}|\mathbf{x}_{it})$. [For a discussion of GEE from an econometrics perspective, see Wooldridge (2010, Section 13.11.4).] Using iterated expectations, if (2.1) holds then

$$\mathrm{E}(y_{it}|\mathbf{x}_i) = \mathrm{E}(c_i|\mathbf{x}_i)m_t(\mathbf{x}_{it},\boldsymbol{\beta}_o),$$

and the latter expression is not $\mathrm{E}(y_{it}|\mathbf{x}_{it})$ if $\mathrm{E}(c_i|\mathbf{x}_i) \neq \mathrm{E}(c_i)$.

The multiplicative formulation using the exponential function in (2.2) can be obtained from

$$\mathrm{E}(y_{it}|\mathbf{x}_{it},a_i) = \exp(a_i + \mathbf{x}_{it}\boldsymbol{\beta}_o)$$

where $c_i \equiv \exp(a_i)$. In applications where $\mathrm{P}(y_{it} = 0) > 0$, it is important to use (2.2) to allow for the possibility that $c_i = 0$, which then implies $y_{it} = 0$, $t = 1,2,\ldots,T$. Remember, we are only assuming $y_{it} \geq 0$; no other restrictions are imposed on the support of y_{it}. A model such as (2.2) is sensible when y_{it} has no natural upper bound.

In FEP estimation, the following residual function, first studied by HHG, plays an

important role:

$$u_{it}(\boldsymbol{\beta}) \equiv y_{it} - n_i p_t(\mathbf{x}_i, \boldsymbol{\beta}), \tag{2.3}$$

where $n_i \equiv \sum_{r=1}^{T} y_{ir}$ and

$$p_t(\mathbf{x}_i, \boldsymbol{\beta}) \equiv \frac{m_t(\mathbf{x}_{it}, \boldsymbol{\beta})}{\sum_{r=1}^{T} m_r(\mathbf{x}_{ir}, \boldsymbol{\beta})}. \tag{2.4}$$

As convenient shorthand, we write $m_{it}(\boldsymbol{\beta}) = m_t(\mathbf{x}_{it}, \boldsymbol{\beta})$ and $p_{it}(\boldsymbol{\beta}) = p_t(\mathbf{x}_i, \boldsymbol{\beta})$. We can stack the $p_{it}(\boldsymbol{\beta})$ into the $T \times 1$ vector $\mathbf{p}(\mathbf{x}_i, \boldsymbol{\beta})$ and write

$$\mathbf{u}_i(\boldsymbol{\beta}) = \mathbf{y}_i - \mathbf{p}(\mathbf{x}_i, \boldsymbol{\beta}) n_i = \mathbf{y}_i - \mathbf{p}(\mathbf{x}_i, \boldsymbol{\beta}) \mathbf{1}_T' \mathbf{y}_i = [\mathbf{I}_T - \mathbf{p}(\mathbf{x}_i, \boldsymbol{\beta}) \mathbf{1}_T'] \mathbf{y}_i, \tag{2.5}$$

where $\mathbf{u}_i(\boldsymbol{\beta})$ is the $T \times 1$ vector with t^{th} element $u_{it}(\boldsymbol{\beta})$ and $\mathbf{1}_T$ is the $T \times 1$ vector with all elements unity. As shown in Wooldridge (1999) under Assumption CM.1,

$$\mathrm{E}[\mathbf{u}_i(\boldsymbol{\beta}_o) | \mathbf{x}_i] = \mathbf{0}. \tag{2.6}$$

Further, the score of the quasi-log-likelihood function for random draw i can be written as

$$\mathbf{s}_i(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \mathbf{p}(\mathbf{x}_i, \boldsymbol{\beta})' \mathbf{W}(\mathbf{x}_i, \boldsymbol{\beta}) \mathbf{u}_i(\boldsymbol{\beta}) \tag{2.7}$$

where

$$\mathbf{W}(\mathbf{x}_i, \boldsymbol{\beta}) = \mathrm{diag}\left\{[p_{i1}(\boldsymbol{\beta})]^{-1}, [p_{i2}(\boldsymbol{\beta})]^{-1}, \ldots, [p_{iT}(\boldsymbol{\beta})]^{-1}\right\} \tag{2.8}$$

is $T \times T$. It follows immediately that

$$\mathrm{E}[\mathbf{s}_i(\boldsymbol{\beta}_o) | \mathbf{x}_i] = \mathbf{0}, \tag{2.9}$$

and this translates, under standard regularity conditions, into the consistency and $\sqrt{N}$-asymptotic normality of the FEP estimator. For emphasis, only Assumption CM is needed for consistency and asymptotic normality, and fully robust inference using a sandwich estimator is essentially trivial.

Wooldridge (1999) also notes that the conditional moment restrictions in (2.6) leads to uncountably many unconditional moment restrictions beyond those used by the FEP estimator, which are given by

$$\mathrm{E}[\mathbf{s}_i(\boldsymbol{\beta}_o)] = \mathbf{0}.$$

In the next section we derive the optimal instruments under a set of second moment assumptions.

3. Optimal Instruments under Second Moment Assumptions

Given the moment conditions in (2.6), we can apply Chamberlain's (1992) semiparametric efficiency bound to obtain an asymptotically efficient estimator. Define

$$\mathbf{D}_o(\mathbf{x}_i) \equiv \mathrm{E}[\nabla_{\boldsymbol{\beta}}\mathbf{u}_i(\boldsymbol{\beta}_o)|\mathbf{x}_i] \tag{3.1}$$

and

$$\mathbf{V}_o(\mathbf{x}_i) \equiv \mathrm{Var}[\mathbf{u}_i(\boldsymbol{\beta}_o)|\mathbf{x}_i]. \tag{3.2}$$

Under regularity conditions of the kind found in Newey and McFadden (1994), Newey (2001) extended Chamberlain (1992) by allowing $\mathbf{V}_o(\mathbf{x}_i)$ to be singular and showed that the efficient estimator that uses only (2.6) has asymptotic variance

$$\left\{\mathrm{E}\left[\mathbf{D}_o(\mathbf{x}_i)'\mathbf{V}_o(\mathbf{x}_i)^{-}\mathbf{D}_o(\mathbf{x}_i)\right]\right\}^{-1}, \tag{3.3}$$

where $\mathbf{V}_o(\mathbf{x}_i)^{-}$ denotes any generalized inverse (*g*-inverse), which means $\mathbf{V}_o(\mathbf{x}_i)\mathbf{V}_o(\mathbf{x}_i)^{-}\mathbf{V}_o(\mathbf{x}_i) = \mathbf{V}_o(\mathbf{x}_i)$. Because $\mathbf{V}_o(\mathbf{x}_i)$ is symmetric, a symmetric *g*-inverse always exists, and it simplifies notation to take $\mathbf{V}_o(\mathbf{x}_i)^{-}$ to be symmetric. Below we obtain an explicit formula for a symmetric *g*-inverse. Given a random sample of size *N* and knowledge of $\mathbf{D}_o(\mathbf{x}_i)$ and $\mathbf{V}_o(\mathbf{x}_i)$, an estimator $\hat{\boldsymbol{\beta}}_{OPT}$ that achieves this lower bound solves the exactly identified

moment equations

$$\sum_{i=1}^{N} \mathbf{D}_o(\mathbf{x}_i)' \mathbf{V}_o(\mathbf{x}_i)^{-} \mathbf{u}_i\left(\hat{\boldsymbol{\beta}}_{OPT}\right) = \mathbf{0}. \tag{3.4}$$

Of course, this estimator is infeasible because $\mathbf{D}_o(\mathbf{x}_i)$ and $\mathbf{V}_o(\mathbf{x}_i)$ are generally unknown. In principle, both can be nonparametrically estimated. However, especially given the often large dimension of $\mathbf{x}_i$, nonparametric estimation of many conditional means, variances, and covariances hardly seems worth it just to improve asymptotic efficiency over the FEP estimator. Plus, the finite-sample properties of the the resulting estimator could be poor. Our goal here is to obtain simple formulas for the optimal IVs, $\mathbf{Z}^*(\mathbf{x}_i) \equiv \mathbf{V}_o(\mathbf{x}_i)^{-}\mathbf{D}_o(\mathbf{x}_i)$. under reasonably flexible parametric second moment assumptions that have antecedents in the count data literature.

To find $\mathbf{D}_o(\mathbf{x}_i)$, note that

$$\nabla_{\boldsymbol{\beta}} \mathbf{u}_i(\boldsymbol{\beta}) = -\nabla_{\boldsymbol{\beta}} \mathbf{p}(\mathbf{x}_i, \boldsymbol{\beta}) n_i, \tag{3.5}$$

where, for each t, we can write

$$\nabla_{\boldsymbol{\beta}} p_{it}(\boldsymbol{\beta}) = \left[\sum_{r=1}^{T} m_{ir}(\boldsymbol{\beta})\right]^{-1} \left\{\nabla_{\boldsymbol{\beta}} m_{it}(\boldsymbol{\beta}) - \left[\sum_{r=1}^{T} \nabla_{\boldsymbol{\beta}} m_{ir}(\boldsymbol{\beta})\right] p_{it}(\boldsymbol{\beta})\right\}.$$

Therefore,

$$\begin{aligned} \nabla_{\boldsymbol{\beta}} \mathbf{p}_i(\boldsymbol{\beta}) &= \left[\sum_{r=1}^{T} m_{ir}(\boldsymbol{\beta})\right]^{-1} \{\nabla_{\boldsymbol{\beta}} \mathbf{m}_i(\boldsymbol{\beta}) - \mathbf{p}_i(\boldsymbol{\beta})[\mathbf{1}_T' \nabla_{\boldsymbol{\beta}} \mathbf{m}_i(\boldsymbol{\beta})]\} \\ &= \left[\sum_{r=1}^{T} m_{ir}(\boldsymbol{\beta})\right]^{-1} [\mathbf{I}_T - \mathbf{p}_i(\boldsymbol{\beta})\mathbf{1}_T'] \nabla_{\boldsymbol{\beta}} \mathbf{m}_i(\boldsymbol{\beta}) \end{aligned} \tag{3.6}$$

Further, because

$$\mathrm{E}(n_i|\mathbf{x}_i, c_i) = c_i\left[\sum_{r=1}^{T} m_{ir}(\boldsymbol{\beta}_o)\right]$$

we have

$$\mathrm{E}[\nabla_{\boldsymbol{\beta}}\mathbf{u}_i(\boldsymbol{\beta}_o)|\mathbf{x}_i, c_i] = -c_i[\mathbf{I}_T - \mathbf{p}_i(\boldsymbol{\beta}_o)\mathbf{1}_T']\nabla_{\boldsymbol{\beta}}\mathbf{m}_i(\boldsymbol{\beta}_o)$$

Now, let

$$\mu_c(\mathbf{x}_i) \equiv \mathrm{E}(c_i|\mathbf{x}_i).$$

Then we have shown

$$\mathbf{D}_o(\mathbf{x}_i) = -\mu_c(\mathbf{x}_i)[\mathbf{I}_T - \mathbf{p}_i(\boldsymbol{\beta}_o)\mathbf{1}_T']\nabla_{\boldsymbol{\beta}}\mathbf{m}_i(\boldsymbol{\beta}_o), \tag{3.7}$$

which is the first piece needed to derive the optimal instruments. The unknown function in $\mathbf{D}_o(\mathbf{x}_i)$, $\mu_c(\mathbf{x}_i)$, is the conditional mean in the heterogeneity distribution; all other functions are known up to $\boldsymbol{\beta}_o$.

Next, consider $\mathbf{V}_o(\mathbf{x}_i)^-$. First, we can write

$$\begin{aligned}\mathbf{V}_o(\mathbf{x}_i) &\equiv \mathrm{Var}[\mathbf{u}_i(\boldsymbol{\beta}_o)|\mathbf{x}_i] = \mathrm{Var}\{[\mathbf{I}_T - \mathbf{p}_i(\boldsymbol{\beta}_o)\mathbf{1}_T']\mathbf{y}_i|\mathbf{x}_i\} \\ &\equiv (\mathbf{I}_T - \mathbf{P}_i)\boldsymbol{\Omega}_i(\mathbf{I}_T - \mathbf{P}_i')\end{aligned} \tag{3.8}$$

where

$$\boldsymbol{\Omega}_i \equiv \mathrm{Var}(\mathbf{y}_i|\mathbf{x}_i) \tag{3.9}$$

is assumed to be nonsingular (with probability one) and $\mathbf{P}_i \equiv \mathbf{p}_i(\boldsymbol{\beta}_o)\mathbf{1}_T'$ is $T \times T$. Because the $p_{it}(\boldsymbol{\beta}_o)$ sum to unity across t, it is easily shown that $\mathbf{P}_i$ is an idempotent (but not symmetric) matrix with $\mathrm{rank}(\mathbf{P}_i) = 1$.

In establishing that the FEP estimator is asymptotically efficient under the Poisson first and second moment assumptions, Verdier (2018) shows that the symmetric matrix

$$\begin{aligned}\mathbf{V}_o(\mathbf{x}_i)^- &= \boldsymbol{\Omega}_i^{-1} - \boldsymbol{\Omega}_i^{-1}\mathbf{p}_i(\boldsymbol{\beta}_o)\left[\mathbf{p}_i(\boldsymbol{\beta}_o)'\boldsymbol{\Omega}_i^{-1}\mathbf{p}_i(\boldsymbol{\beta}_o)\right]^{-1}\mathbf{p}_i(\boldsymbol{\beta}_o)'\boldsymbol{\Omega}_i^{-1} \\ &= \boldsymbol{\Omega}_i^{-1} - \boldsymbol{\Omega}_i^{-1}\mathbf{m}_i(\boldsymbol{\beta}_o)\left[\mathbf{m}_i(\boldsymbol{\beta}_o)'\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i(\boldsymbol{\beta}_o)\right]^{-1}\mathbf{m}_i(\boldsymbol{\beta}_o)'\boldsymbol{\Omega}_i^{-1} \qquad (3.10)\end{aligned}$$

is a generalized inverse of $\mathbf{V}_o(\mathbf{x}_i)$. The second equality in (3.10) follows by the definition of $\mathbf{p}_i(\boldsymbol{\beta}_o)$ and by cancelling terms. We can use this expression to find a simple formula for the optimal instruments under Assumption CM. By simple multiplication it is easily seen that

$$\mathbf{p}_i(\boldsymbol{\beta}_o)'\mathbf{V}_o(\mathbf{x}_i)^- = \mathbf{0}$$

and so

$$\mathbf{D}_o(\mathbf{x}_i)'\mathbf{V}_o(\mathbf{x}_i)^- = -\mu_c(\mathbf{x}_i)\nabla_{\boldsymbol{\beta}}\mathbf{m}_i(\boldsymbol{\beta}_o)'\mathbf{V}_o(\mathbf{x}_i)^-. \qquad (3.10)$$

The expression for the optimal instruments in (3.10) is not directly applicable because $\mu_c(\cdot)$ and $\mathbf{V}_o(\cdot)$ are unknown, with the latter depending on the unknown $\boldsymbol{\Omega}_i$. We now impose assumptions on the structural variance-covariance matrix, $\text{Var}(\mathbf{y}_i|\mathbf{x}_i, c_i)$, that lead to useful simplifications. The first restriction is on the diagonal elements.

Assumption Working Variance 1 (WV.1): For $t = 1, \ldots, T$, there exists $\alpha > 0$ such that

$$\text{Var}(y_{it}|\mathbf{x}_i, c_i) = \text{Var}(y_{it}|\mathbf{x}_{it}, c_i) = \alpha \text{E}(y_{it}|\mathbf{x}_{it}, c_i) = \alpha c_i m_{it}(\boldsymbol{\beta}_o). \qquad (3.11)$$

Assumption WV.1 is motivated by the count data literature, where the assumption that the variance is proportional to the mean is commonly used in generalized linear models (GLM) and GEE settings; see, for example, McCullagh and Nelder (1989), Liang and Zeger (1986), Hardin and Hilbe (2012), and Wooldridge (2010, Section 13.11). Again, one important difference between our setting and the standard GEE setting is that we state the first and second moments conditional on the unobserved heterogeneity, c_i, in addition to the observable variables, $\mathbf{x}_i$. Once the population is effectively partitioned on the basis of $(\mathbf{x}_i, c_i)$, the so-called

"GLM variance assumption" is more appealing. We do not restrict the value of $\alpha = \text{Var}(y_{it}|\mathbf{x}_{it}, c_i)/\text{E}(y_{it}|\mathbf{x}_{it}, c_i)$, and so the y_{it} can exhibit underdispersion or overdispersion relative to the Poisson distribution. This variance-mean relationship also holds for one popular parameterization of the negative binomial distribution (which implies overdispersion), and can hold for continuous outcomes as well, such as a common parameterization of the gamma distribution.

The second working assumption is on the conditional correlation matrix.

Assumption Working Variance 2 (**WV**.**2**): For a $T \times T$ symmetric, positive definite matrix $\mathbf{R}$ (with unity down the diagonal),

$$\text{Corr}(\mathbf{y}_i|\mathbf{x}_i, c_i) = \mathbf{R}. \ \square \tag{3.12}$$

Assumption WV.2 is motivated by the GEE literature, where a constant conditional correlation matrix is the leading example of a working correlation assumption. We do not put restrictions on the elements of $\mathbf{R}$, $\rho_{ts} = \text{Corr}(y_{it}, y_{is}|\mathbf{x}_i, c_i)$, other than those that ensure $\mathbf{R}$ is a valid correlation matrix. The special case of no serial correlation conditional on $(\mathbf{x}_i, c_i)$ is $\mathbf{R} = \mathbf{I}_T$. One could impose an exchangeability restriction on $\mathbf{R}$, as is common in the GEE literature, but that is less attractive here because we are conditioning on c_i (which would often be assumed to be an explanation for an exchangeable structure without conditioning on c_i). With large N and small T, there is little reason to impose restrictions on $\mathbf{R}$. Again, an important difference with the GEE literature is we condition the correlation matrix on c_i as well as $\mathbf{x}_i$ – which makes $\mathbf{R} = \mathbf{I}_T$ more tenable (but still unnecessary).

We can combine Assumptions WV.1 and WV.2 into a working variance-covariance matrix conditional on $(\mathbf{x}_i, c_i)$:

$$\text{Var}(\mathbf{y}_i|\mathbf{x}_i, c_i) = \alpha c_i \mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2}, \tag{3.13}$$

where $\mathbf{M}_i \equiv \text{diag}\{m_{i1}(\boldsymbol{\beta}_o), m_{i2}(\boldsymbol{\beta}_o), \ldots, m_{iT}(\boldsymbol{\beta}_o)\}$ and $\mathbf{M}_i^{1/2} = \text{diag}\left\{\sqrt{m_{i1}(\boldsymbol{\beta}_o)}, \sqrt{m_{i2}(\boldsymbol{\beta}_o)}, \ldots, \sqrt{m_{iT}(\boldsymbol{\beta}_o)}\right\}$ is the obvious matrix square root. If not for conditioning on the unobserved heterogeneity c_i, (3.13) has a structure very familiar from the GEE literature on estimating conditional means of count variables with longitudinal data.

In stating Assumptions WV.1 and WV.2, we have opted not to include a "*o*" subscript on α or $\mathbf{R}$. This decision requires a brief explanation. For deriving the optimal instruments, we are assuming the existence of "true values." However, when we discuss implementation of our new estimator in Section 4, we do not assume Assumptions WV.1 or WV.2 are in force. To ensure that the focus is on estimating $\boldsymbol{\beta}_o$, and to simplify the notation, we omit the "*o*" subscripts on the parameters in the working assumptions.

Before deriving the optimal instruments, we first obtain $\boldsymbol{\Omega}_i = \text{Var}(\mathbf{y}_i|\mathbf{x}_i)$ and provide a useful expression for its inverse. As shorthand, let $\mathbf{m}_i$ be the $T \times 1$ vector of $m_{it}(\boldsymbol{\beta}_o)$, and define $\mathbf{M}_i^{1/2}$ as above. We use $\sqrt{\mathbf{m}_i}$ to denote the $T \times 1$ vector containing the square roots of the $m_{it}(\boldsymbol{\beta}_o)$. In stating the next lemma, let

$$\sigma_c^2(\mathbf{x}_i) = \text{Var}(c_i|\mathbf{x}_i).$$

Lemma 3.1: Under Assumptions CM, WV.1, and WV.2,

$$\text{Var}(\mathbf{y}_i|\mathbf{x}_i) = \boldsymbol{\Omega}_i = \alpha\mu_c(\mathbf{x}_i)\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2} + \sigma_c^2(\mathbf{x}_i)\mathbf{m}_i\mathbf{m}_i', \tag{3.14}$$

which is positive definite. Further,

$$\boldsymbol{\Omega}_i^{-1} = \frac{1}{[\alpha\mu_c(\mathbf{x}_i)]}\mathbf{M}_i^{-1/2}\left\{\mathbf{R}^{-1} - \frac{\sigma_c^2(\mathbf{x}_i)}{\left[\alpha\mu_c(\mathbf{x}_i) + \sigma_c^2(\mathbf{x}_i)\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\right]}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\right\}\mathbf{M}_i^{-1/2}. \square$$

Proofs of all results are given in the appendix. Establishing the formula for $\mathbf{\Omega}_i$ uses the law of total variance (for matrices). Positive definiteness of $\mathbf{\Omega}_i$ follows because the first term in (3.14) is positive definite under WV.1 and WV.2 and the second is always positive semi-definite. As shown in the appendix, the formula for $\mathbf{\Omega}_i^{-1}$ applies a result due to Sherman and Morrison (1950).

Now we can state the main optimal instrument result.

Theorem 3.1: Under Assumptions CM, WV.1, and WV.2, a symmetric generalized inverse of $\mathbf{V}_o(\mathbf{x}_i)$ is

$$\mathbf{V}_o(\mathbf{x}_i)^- = \frac{1}{[\alpha\mu_c(\mathbf{x}_i)]}\mathbf{M}_i^{-1/2}\left[\mathbf{R}^{-1} - \frac{1}{\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\,\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\right]\mathbf{M}_i^{-1/2}. \tag{3.15}$$

Further, the optimal $T \times K$ matrix of instruments, $\mathbf{Z}^*(\mathbf{x}_i)$, is

$$\mathbf{Z}^*(\mathbf{x}_i)' \equiv \nabla_{\boldsymbol{\beta}}\mathbf{m}_i(\boldsymbol{\beta}_o)'\mathbf{M}_i^{-1/2}\left[\mathbf{R}^{-1} - \frac{1}{\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\,\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\right]\mathbf{M}_i^{-1/2}, \tag{3.16}$$

where, again, $\mathbf{m}_i$ and $\mathbf{M}_i$ are evaluated at $\boldsymbol{\beta}_o$. We have dropped the minus sign in $\mathbf{D}_o(\mathbf{x}_i)$ as that does not affect the optimal choice. □

The optimal instrument matrix in (3.16) has a rather remarkable feature: it does not depend on the constant α measuring dispersion nor on the conditional first two moments of the heterogeneity distribution, $\mu_c(\mathbf{x}_i)$ and $\sigma_c(\mathbf{x}_i)$ – even though $\mathbf{\Omega}_i^{-1}$ depends on all of these quantities and $\mathbf{D}_o(\mathbf{x}_i)$ depends on $\mu_c(\mathbf{x}_i)$. Under the working variance matrix assumptions, the optimal instruments depend only on $\boldsymbol{\beta}_o$ and $\mathbf{R}$. We have a natural preliminary estimator of $\boldsymbol{\beta}_o$, namely, the FEP estimator. Estimating $\mathbf{R}$ is much more challenging, and for that we will introduce additional working assumptions – something we take up in the next section.

An interesting special case of Theorem 3.1 is when the $\{y_{it} : t = 1,2,\ldots,T\}$ are

conditionally uncorrelated, an assumption with a long history in linear and nonlinear unobserved effects models. Traditional treatments of linear unobserved effects models – often called "random effects" models – include the assumption that idiosyncratic shocks are serially uncorrelated, which implies that, conditional on $(\mathbf{x}_i, c_i)$, the $\{y_{it} : t = 1, 2, \ldots, T\}$ are uncorrelated. In using joint maximum likelihood to estimate popular nonlinear models with unobserved heterogeneity – random effects probit and ordered probit, random effects multinomial logit, random effects Tobit, random effects version of Poisson and negative binomial models, among others – it is almost always assumed that the $\{y_{it} : t = 1, 2, \ldots, T\}$ are independent conditional on $(\mathbf{x}_i, c_i)$; see Sections 13.9, 15.8, 17.8, and 18.7 in Wooldridge (2010).

Corollary 3.1: Under Assumptions CM, WV.1, and WV.2 with $\mathbf{R} = \mathbf{I}_T$, the FEP estimator is efficient among estimators that use only Assumption CM for consistency. □

Corollary 3.1 is a new result that shows the FEP estimator is asymptotically efficient for any $\alpha > 0$ in Assumption WV.1 provided there is no serial correlation. Conditional on $\mathbf{x}_i$ and c_i, any amount of constant underdispersion or overdispersion is allowed. Therefore, Corollary 3.1 improves on Verdier (2018), who imposed $\alpha = 1$ – the value that holds for the Poisson distribution. That FEP is asymptotically efficient for any α while allowing for arbitrary dependence between c_i and $\mathbf{x}_i$ is very satisfying and allows us to make an interesting connection with the cross-sectional GLM literature. As pointed out in Wooldridge (2010, Section 13.11.3), the cross-sectional version of Assumption WV.1 implies that the Poisson QMLE is asymptotically efficient among estimators that use only correct specification of the conditional mean function for consistency. We now have a panel data version of this result under the no serial correlation assumption $\mathbf{R} = \mathbf{I}_T$. Given that Corollary 3.1 allows

overdispersion and underdispersion, it seems very unlikely that there are weaker conditions under which the FEP estimator is asymptotically efficient.

4. Operationalizing Optimal IV Estimation

From Theorem 3.1, in order to obtain a feasible optimal IV estimator under Assumptions CM, WV.1, and WV.2, we need a preliminary consistent estimator of $\boldsymbol{\beta}_o$ and we either need to know $\mathbf{R}$ or have a consistent estimator of it. If we want to impose a specific structure on $\mathbf{R}$ – say, an AR(1) model with a known AR(1) parameter – then (3.16) can be used after replacing $\boldsymbol{\beta}_o$ with $\hat{\boldsymbol{\beta}}_{FEP}$ (the clear choice for a first-stage estimator of $\boldsymbol{\beta}_o$). Remember, imposing such a restriction when it is incorrect would not affect consistency of the method of moments estimator; but the estimator would not be asymptotically efficient. Generally, we want to estimate $\mathbf{R}$ without imposing any restrictions.

In order to ignore the first-stage estimation when obtaining the asymptotic variance of $\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{OPT} - \boldsymbol{\beta}_o\right)$, the first-stage estimators of $\boldsymbol{\beta}_o$ and $\mathbf{R}$ should be $\sqrt{N}$-consistent – a weak requirement because we are assuming random sampling and smooth moment and objective functions. See Wooldridge (2010, Chapter 14) for discussion. As mentioned earlier, it is very natural to use the FEP estimator as the initial estimator of $\boldsymbol{\beta}_o$. Estimation of $\mathbf{R}$ is more difficult because it is the (working) correlation matrix conditional on the unobserved heterogeneity, c_i, in addition to $\mathbf{x}_i$.

The key to estimating $\mathbf{R}$ is the relationship in (3.14). To see how (3.14) can be used, define a $T \times 1$ vector of errors

$$\mathbf{v}_i \equiv \mathbf{y}_i - \mathrm{E}(\mathbf{y}_i|\mathbf{x}_i) = \mathbf{y}_i - \mu_c(\mathbf{x}_i)\mathbf{m}_i. \tag{4.1}$$

Then

$$\mathrm{E}(\mathbf{v}_i\mathbf{v}_i'|\mathbf{x}_i) = \alpha\mu_c(\mathbf{x}_i)\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2} + \sigma_c^2(\mathbf{x}_i)\mathbf{m}_i\mathbf{m}_i', \tag{4.2}$$

which we can write in matrix error form as

$$\mathbf{v}_i\mathbf{v}_i' = \alpha\mu_c(\mathbf{x}_i)\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2} + \sigma_c^2(\mathbf{x}_i)\mathbf{m}_i\mathbf{m}_i' + \mathbf{S}_i$$

with

$$\mathrm{E}(\mathbf{S}_i|\mathbf{x}_i) = \mathbf{0}. \tag{4.3}$$

Next, define

$$\mathbf{k}_i \equiv \mathrm{E}(\mathbf{y}_i|\mathbf{x}_i) = \mu_c(\mathbf{x}_i)\mathbf{m}_i, \tag{4.4}$$

and let $\mathbf{K}_i$ be the diagonalized version of $\mathbf{k}_i$. Then

$$\mathbf{v}_i\mathbf{v}_i' - \sigma_c^2(\mathbf{x}_i)\mathbf{m}_i\mathbf{m}_i' = \alpha\sqrt{\mathbf{K}_i}\,\mathbf{R}\sqrt{\mathbf{K}_i} + \mathbf{S}_i \tag{4.5}$$

and so

$$\mathbf{K}_i^{-1/2}[\mathbf{v}_i\mathbf{v}_i' - \sigma_c^2(\mathbf{x}_i)\mathbf{m}_i\mathbf{m}_i']\mathbf{K}_i^{-1/2}/\alpha = \mathbf{R} + \mathbf{K}_i^{-1/2}\mathbf{S}_i\mathbf{K}_i^{-1/2}/\alpha. \tag{4.6}$$

By (4.3) and iterated expectations, the second term in (4.6), $\mathbf{K}_i^{-1/2}\mathbf{S}_i\mathbf{K}_i^{-1/2}/\alpha$, has a mean of zero. Therefore, we have shown

$$\mathbf{R} = \alpha^{-1}\mathrm{E}\left\{\mathbf{K}_i^{-1/2}[\mathbf{v}_i\mathbf{v}_i' - \sigma_c^2(\mathbf{x}_i)\mathbf{m}_i\mathbf{m}_i']\mathbf{K}_i^{-1/2}\right\}. \tag{4.7}$$

Combining (4.7) with (3.16) shows that α appears as a multiplicative factor in $\mathbf{Z}^*(\mathbf{x}_i)$, and therefore does not affect the optimal choice of instruments.

Equation (4.7) for $\mathbf{R}$ suggests simply computing the sample analog of the matrix inside the expected value. However, we must deal with the fact that the matrix depends on three unknown quantities: the parameter α, the conditional mean function $\mu_c(\cdot)$ (which appears in the definition of $\mathbf{v}_i$), and the conditional variance function $\sigma_c^2(\cdot)$.

There are different ways to approach estimation of $\mu_c(\cdot)$. For example, under Assumption CM,

$$\mathrm{E}(n_i|\mathbf{x}_i, c_i) = c_i\left[\sum_{r=1}^{T} m_{ir}(\boldsymbol{\beta}_o)\right] \tag{4.8}$$

and so

$$\mathrm{E}\left[\left.\frac{n_i}{\sum_{r=1}^{T} m_{ir}(\boldsymbol{\beta}_o)}\right|\mathbf{x}_i\right] = \mu_c(\mathbf{x}_i). \tag{4.9}$$

Alternatively, we can write

$$\mathrm{E}\left[\left. T^{-1}\sum_{t=1}^{T}\frac{y_{it}}{m_{it}(\boldsymbol{\beta}_o)}\right|\mathbf{x}_i\right] = \mu_c(\mathbf{x}_i). \tag{4.10}$$

Because we have available $\sqrt{N}$-consistent estimators of $\boldsymbol{\beta}_o$, expressions (4.9) and (4.10) show that $\mu_c(\cdot)$ is nonparametrically identified. In fact, we can use these expressions to motivate a nonparametric estimator. Using $\hat{\boldsymbol{\beta}}_{FEP}$ as the initial estimator of $\boldsymbol{\beta}_o$, we construct a dependent variable, $n_i/\left[\sum_{r=1}^{T}\hat{m}_{ir}\right]$, where $\hat{m}_{ir} = m_{ir}\left(\hat{\boldsymbol{\beta}}_{FEP}\right)$, and use it in a cross-sectional nonparametric regression to obtain $\hat{\mu}_c(\cdot)$.

For $\sigma_c^2(\cdot)$, the law of total variance implies

$$\begin{aligned}\mathrm{E}(v_{it}^2|\mathbf{x}_i) &= \mathrm{Var}(y_{it}|\mathbf{x}_i) = \mathrm{E}[\mathrm{Var}(y_{it}|\mathbf{x}_i, c_i)|\mathbf{x}_i] + \mathrm{Var}[\mathrm{E}(y_{it}|\mathbf{x}_i, c_i)|\mathbf{x}_i] \\ &= \mathrm{E}[\alpha c_i m_{it}(\boldsymbol{\beta}_o)|\mathbf{x}_i] + \mathrm{Var}[c_i m_{it}(\boldsymbol{\beta}_o)|\mathbf{x}_i] \\ &= \alpha\mu_c(\mathbf{x}_i)m_{it}(\boldsymbol{\beta}_o) + \sigma_c^2(\mathbf{x}_i)[m_{it}(\boldsymbol{\beta}_o)]^2,\end{aligned} \tag{4.11}$$

where we impose the working variance Assumption WV.1. Given that $\mu_c(\mathbf{x}_i)$ is identified from the previous argument, this expression identifies α and $\sigma_c^2(\cdot)$. In fact, after obtaining (semiparametric) residuals $\hat{v}_{it} = y_{it} - \hat{\mu}_c(\mathbf{x}_i)m_{it}\left(\hat{\boldsymbol{\beta}}_{FEP}\right)$, we can use the squared residuals, $\hat{v}_{it}^2$,

as the dependent variable in nonparametric estimation of $\sigma_c^2(\cdot)$. Therefore, a semiparametric approach to estimating the optimal IVs is available under Assumptions CM, WV.1, and WV.2.

For practical reasons, our suggestion is to avoid estimating either $\mu_c(\cdot)$ and $\sigma^2(\cdot)$ nonparametrically. Remember, we only need to estimate these conditional moments to obtain IVs more efficient than those used by the FEP estimator. The dimension of $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$ is often large. We can reduce the dimension by using a nonparametric Mundlak (1978) device, which would have $\mu_c(\cdot)$ and $\sigma^2(\cdot)$ depending only on time averages $\bar{\mathbf{x}}_i \equiv T^{-1}\sum_{r=1}^{T}\mathbf{x}_{ir}$. Nevertheless, estimating a conditional variance along with a conditional mean when K is even moderately large is still challenging, both theoretically and practically. It would involve choosing at least two tuning parameters. From a robustness perspective, we cannot improve over the FEP estimator because it is consistent under Assumption CM. High-dimensional nonparametric estimation seems unnecessary to improve over the usual FEP estimator in the presence of serial correlation and under- or overdispersion, especially if one factors in finite-sample considerations. Instead, we draw on the literature on models for nonnegative responses to suggest working assumptions for the conditional mean and variance of the heterogeneity – as summarized, for example, in Wooldridge (2010, Section 18.7.3).

For concreteness, and because it is by far the leading case, we now assume that $m_{it}(\boldsymbol{\beta}_o) = \exp(\mathbf{x}_{it}\boldsymbol{\beta}_o)$. Other forms of $m_{it}(\boldsymbol{\beta}_o)$ are easily handled, but the formulas and connections with other literatures is not as straightforward.

Assumption WH.1: For known $1 \times Q$ functions $\mathbf{h}(\mathbf{x}_i)$, a scalar η, and $\boldsymbol{\lambda}$ a $Q \times 1$ vector,

$$\mu_c(\mathbf{x}_i) \equiv \mathrm{E}(c_i|\mathbf{x}_i) = \exp[\eta + \mathbf{h}(\mathbf{x}_i)\boldsymbol{\lambda}]. \square \tag{4.12}$$

The leading case is to use the (nonredundant) time averages of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$, which is an

extension of the Mundlak (1978) device to the nonlinear case, so that $\mathbf{h}(\mathbf{x}_i) = \bar{\mathbf{x}}_i$. But we can also use Chamberlain's (1980) less restrictive version, or we can include, say, unit-specific second moments. It seems sensible to use something relatively simple as we are only using WH.1 to generate instruments.

When we combine Assumption WH.1 with the exponential conditional mean for $\mathrm{E}(y_{it}|\mathbf{x}_i, c_i)$, we obtain, by iterated expectations,

$$\mathrm{E}(y_{it}|\mathbf{x}_i) = \exp[\eta + \mathbf{h}(\mathbf{x}_i)\boldsymbol{\lambda}]\exp(\mathbf{x}_{it}\boldsymbol{\beta}_o) = \exp[\mathbf{x}_{it}\boldsymbol{\beta}_o + \eta + \mathbf{h}(\mathbf{x}_i)\boldsymbol{\lambda}].$$

The parameters in this conditional mean function can be consistently estimated using a variety of methods. A simple approach is to exploit equation (4.9) or (4.10) using exponential mean functions. After obtaining the FEP estimator $\hat{\boldsymbol{\beta}}_{FEP}$, estimate η and $\boldsymbol{\lambda}$ by a cross sectional Poisson regression with mean function $\exp[\eta + \mathbf{h}(\mathbf{x}_i)\boldsymbol{\lambda}]$ and one of the dependent variables

$$\frac{n_i}{\sum_{r=1}^{T}\exp\left(\mathbf{x}_{ir}\hat{\boldsymbol{\beta}}_{FEP}\right)} \quad \text{or} \quad T^{-1}\sum_{t=1}^{T}\frac{y_{it}}{\exp\left(\mathbf{x}_{it}\hat{\boldsymbol{\beta}}_{FEP}\right)}. \tag{4.13}$$

Even if the original y_{it} are count variables – and there is no presumption that they are – neither of the regressands in (4.13) would be a count variable. Of course, this is of no consequence because of the robustness of the Poisson QMLE for estimating the parameters of the conditional mean regardless of the nature of the dependent variable (provided it is nonnegative).

Alternatively, $\boldsymbol{\beta}_o$, η, and $\boldsymbol{\lambda}$ can be estimated jointly using the pooled Poisson QMLE. The pooled Poisson QMLE is completely robust to distributional misspecification and serial correlation. Of course, to preserve consistency of the resulting method of moments estimator we do not need Assumption WH.1 to hold; we are using it to estimate the optimal instruments

derived earlier.

The second working assumption on the heterogeneity distribution imposes a restriction on the variance-mean relationship.

Assumption WH.2: For $\delta > 0$,

$$\sigma_c^2(\mathbf{x}_i) \equiv \text{Var}(c_i|\mathbf{x}_i) = \delta[\mu_c(\mathbf{x}_i)]^2 = \delta\{\exp[\eta + \mathbf{h}(\mathbf{x}_i)\boldsymbol{\lambda}]\}^2. \square \tag{4.14}$$

Assumption WH.2 is very common in settings with nonnegative, continuous heterogeneity (including so-called random effects Poisson and negative binomial models). The condition that the variance is proportional to the square of the mean holds for the natural parameterizations of the gamma and lognormal distributions, and holds whenever

$$c_i = q_i\mu_c(\mathbf{x}_i) \tag{4.15}$$

for $q_i \geq 0$ and independent of $\mathbf{x}_i$, without any further restrictions on the distribution of q_i. Like Assumption WH.1, Assumption WH.2 is not needed for consistent estimation using the method of moments estimator but only to estimate the optimal instruments under the working Assumptions WV.1 and WV.2.

Using Assumptions CM, WV.1, WH.1, and WH.2 we can obtain estimating equations for α and δ. First, note that

$$\text{E}(v_{it}^2|\mathbf{x}_i) = \alpha k_{it} + \delta k_{it}^2 \tag{4.15}$$

where

$$k_{it} \equiv \text{E}(y_{it}|\mathbf{x}_i) = \exp[\mathbf{x}_{it}\boldsymbol{\beta}_o + \eta + \mathbf{h}(\mathbf{x}_i)\boldsymbol{\lambda}]$$

An immediate implication of equation (4.15) is

$$\mathrm{E}\left[\left(\frac{v_{it}}{\sqrt{k_{it}}}\right)^2 \middle| \mathbf{x}_i\right] = \alpha + \delta k_{it}, \tag{4.16}$$

which is the basis for estimating variance parameters in common cross-sectional models where heterogeneity is assumed independent of the covariates. A simple way to operationalize the conditional mean is

$$\hat{v}_{it} = y_{it} - \hat{k}_{it} = y_{it} - \exp\left[\mathbf{x}_{it}\hat{\boldsymbol{\beta}}_{FEP} + \hat{\eta} + \mathbf{h}(\mathbf{x}_i)\hat{\boldsymbol{\lambda}}\right], \tag{4.17}$$

where $\hat{\eta}$ and $\hat{\boldsymbol{\lambda}}$ are from one of the Poisson regressions described in equation (4.13). Then $\hat{\alpha}$ and $\hat{\delta}$ are, respectively, the intercept and slope in the pooled simple regression

$$\frac{\hat{v}_{it}^2}{\hat{k}_{it}} \text{ on } 1, \hat{k}_{it},\ t = 1,\ldots,T;\ i = 1,\ldots,N. \tag{4.18}$$

It is clear from equation (3.16) that $\hat{\alpha}$ does not appear in the optimal instruments, but we need to estimate α in order to obtain $\hat{\delta}$. In order to conclude the working assumptions are a reasonable approximation to reality, both $\hat{\alpha}$ and $\hat{\delta}$ should be nonnegative. If one of them is negative (most likely $\hat{\delta}$) then $\hat{\delta}$ should be set to zero. Because $\hat{\alpha}$ drops out of the optimal IVs, we need not estimate it when we set $\hat{\delta} = 0$. Nevertheless, one may be curious about the estimated amount of overdispersion when δ is set to zero. With $\delta = 0$, the estimate of α is simply

$$\hat{\alpha} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\hat{v}_{it}^2/\hat{k}_{it}\right), \tag{4.19}$$

and this is guaranteed to be nonnegative. However, as mentioned above, $\hat{\alpha}$ does not affect estimation of the optimal IVs when $\delta = 0$.

When we add Assumptions WH.1 and WH.2 to the previous assumptions, we obtain a

simple form for $\mathbf{R}$:

$$\mathbf{R} = \alpha^{-1}\mathrm{E}\left\{\mathbf{K}_i^{-1/2}[\mathbf{v}_i\mathbf{v}_i' - \delta\mathbf{k}_i\mathbf{k}_i']\mathbf{K}_i^{-1/2}/\alpha\right\},$$

which leads immediately to the method-of-moments/plug-in estimator

$$\hat{\mathbf{R}} = \left(\frac{1}{\hat{\alpha}}\right)N^{-1}\sum_{i=1}^{N}\hat{\mathbf{K}}_i^{-1/2}\left(\hat{\mathbf{v}}_i\hat{\mathbf{v}}_i' - \hat{\delta}\hat{\mathbf{k}}_i\hat{\mathbf{k}}_i'\right)\hat{\mathbf{K}}_i^{-1/2}. \tag{4.20}$$

By a standard application of the uniform weak law of large numbers [Wooldridge (2010, Lemma 12.1)], $\hat{\mathbf{R}} \xrightarrow{p} \mathbf{R}$. For each $t \neq s$, the correlations are estimated as

$$\hat{\rho}_{st} = \left(\frac{1}{\hat{\alpha}}\right)N^{-1}\sum_{i=1}^{N}\frac{\left(\hat{v}_{is}\hat{v}_{it} - \hat{\delta}\hat{k}_{is}\hat{k}_{it}\right)}{\sqrt{\hat{k}_{is}\hat{k}_{it}}}. \tag{4.21}$$

From the definition of $\hat{\alpha}$ and $\hat{\delta}$ obtained from (4.18), it is easily seen that $\hat{\rho}_{tt} = 1$ for $t = 1,\ldots,T$, and so this estimator imposes the logical requirement that a correlation matrix must have unity down its diagonal.

If we set $\delta = 0$, $\hat{\mathbf{R}}$ reduces to

$$\hat{\mathbf{R}} = \left(\frac{1}{\hat{\alpha}}\right)N^{-1}\sum_{i=1}^{N}\hat{\mathbf{K}}_i^{-1/2}(\hat{\mathbf{v}}_i\hat{\mathbf{v}}_i')\hat{\mathbf{K}}_i^{-1/2}. \tag{4.22}$$

With this choice of $\hat{\mathbf{R}}$, we can make a direct connection with the GEE literature by ignoring the presence of c_i and working off the first two conditional moments of $\mathbf{y}_i$ given $\mathbf{x}_i$ – see, for example, Liang and Zeger (1986) and Wooldridge (2010, Sections 13.11.4 and 18.7.3). Namely, under the full set of working assumptions with $\delta = 0$,

$$\mathrm{E}(y_{it}|\mathbf{x}_i) = \exp[\mathbf{x}_{it}\boldsymbol{\beta}_o + \eta + \mathbf{h}(\mathbf{x}_i)\boldsymbol{\lambda}] = k_{it}, t = 1,\ldots,T \tag{4.23}$$

$$\mathrm{Var}(y_{it}|\mathbf{x}_i) = \alpha\mathrm{E}(y_{it}|\mathbf{x}_i), t = 1,\ldots,T \tag{4.24}$$

$$\mathrm{Corr}(\mathbf{y}_i|\mathbf{x}_i) = \alpha\mathbf{K}_i^{1/2}\mathbf{R}\mathbf{K}_i^{1/2} \tag{4.25}$$

This collection of moment assumptions is precisely what is used in GEE applications of Poisson regression (whether or not y_{it} is a count variable), with the addition of the vector of functions $\mathbf{h}(\mathbf{x}_i)$. We emphasize that these are *all* working assumptions in the current context. Not even the conditional mean function in (4.23) is assumed to hold for consistency because (4.23) is obtained from Assumptions CM and WH.1, whereas we are only require Assumption CM for consistency. We impose Assumptions WH.1 and WH.2 in order to estimate $\mathbf{R}$ and then to estimate $\boldsymbol{\Omega}_i$. Provided it leads to a positive definite estimate, we prefer (4.20) because it is the correct expression under all of the working assumptions.

Under Assumption CM and the full set of working assumptions, we can estimate the optimal IVs, for each i, as

$$\nabla_{\boldsymbol{\beta}}\hat{\mathbf{m}}_i'\hat{\mathbf{M}}_i^{-1/2}\left[\hat{\mathbf{R}}^{-1} - \frac{1}{\sqrt{\hat{\mathbf{m}}_i}'\hat{\mathbf{R}}^{-1}\sqrt{\hat{\mathbf{m}}_i}}\hat{\mathbf{R}}^{-1}\sqrt{\hat{\mathbf{m}}_i}\sqrt{\hat{\mathbf{m}}_i}'\hat{\mathbf{R}}^{-1}\right]\hat{\mathbf{M}}_i^{-1/2}, \tag{4.26}$$

where "^" means the quantity is evaluated at a first-round estimator, most likely $\hat{\boldsymbol{\beta}}_{FEP}$, and $\hat{\mathbf{R}}$ is from (4.20) or, if necessary, (4.22). This results in a just identified set of equations. However, without the full set of working assumptions, this choice of IVs is *not* guaranteed to improve over the FEP estimator because of its dependence on $\hat{\mathbf{R}}$. A somewhat subtle point is that (4.26) is not even optimal under Assumptions CM, WV.1, and WV.2 because consistency of $\hat{\mathbf{R}}$ for $\mathbf{R}$ generally requires correct specification of the heterogeneity mean and variance – that is, Assumptions WH.1 and WH.2. As mentioned previously, if we did not have to estimate $\mathbf{R}$, we could use (4.26) with $\hat{\mathbf{R}}$ replaced by $\mathbf{R}$. Naturally, we want to use the data to provide an estimator of $\mathbf{R}$ better than just guessing. Incidentally, expression (4.26) shows that the estimator $\hat{\alpha}$ has no direct effect on the optimal IVs because it factors out as a constant.

In order to ensure improvements over FEP, our recommendation is to stack the FEP and the

new "optimal" IVs to form an expanded IV matrix, and use GMM with an optimal weighting matrix. The resulting estimator, which we simply call the "GMM estimator," is guaranteed to be asymptotically at least as efficient as the FEP and GFEP estimators; usually it is strictly more efficient than both. In other words, the $T \times 2K$ matrix of IVs is $\hat{\mathbf{Z}}_i$, written in transposed form as

$$\hat{\mathbf{Z}}_i' = \begin{pmatrix} \nabla_{\boldsymbol{\beta}} \hat{\mathbf{m}}_i' \hat{\mathbf{M}}_i^{-1/2} \left[\mathbf{I}_T - \sqrt{\hat{\mathbf{p}}_i} \sqrt{\hat{\mathbf{p}}_i}' \right] \hat{\mathbf{M}}_i^{-1/2} \\ \nabla_{\boldsymbol{\beta}} \hat{\mathbf{m}}_i' \hat{\mathbf{M}}_i^{-1/2} \left[\hat{\mathbf{R}}^{-1} - \frac{1}{\sqrt{\hat{\mathbf{m}}_i}' \hat{\mathbf{R}}^{-1} \sqrt{\hat{\mathbf{m}}_i}} \hat{\mathbf{R}}^{-1} \sqrt{\hat{\mathbf{m}}_i} \sqrt{\hat{\mathbf{m}}_i}' \hat{\mathbf{R}}^{-1} \right] \hat{\mathbf{M}}_i^{-1/2} \end{pmatrix} \tag{4.27}$$

Given this choice of $\hat{\mathbf{Z}}_i$, the mechanics of GMM are straightforward. After obtaining $\hat{\boldsymbol{\beta}}_{FEP}$, obtain the $T \times 1$ residual vectors

$$\tilde{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{p}\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{FEP}\right) n_i. \tag{4.28}$$

Then, given the estimators of η, $\boldsymbol{\lambda}$, α, δ, and $\mathbf{R}$ described above, obtain the $2K \times 2K$ matrix,

$$\hat{\boldsymbol{\Psi}} = N^{-1} \sum_{i=1}^{N} \hat{\mathbf{Z}}_i' \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i' \hat{\mathbf{Z}}_i. \tag{4.29}$$

Assuming $\hat{\boldsymbol{\Psi}}$ is positive definite (which generally holds with probability approaching one), the optimal GMM estimator, $\hat{\boldsymbol{\beta}}_{GMM}$, solves

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^K} \left(\sum_{i=1}^{N} \mathbf{u}_i(\boldsymbol{\beta})' \hat{\mathbf{Z}}_i \right) \hat{\boldsymbol{\Psi}}^{-1} \left(\sum_{i=1}^{N} \hat{\mathbf{Z}}_i' \mathbf{u}_i(\boldsymbol{\beta}) \right) \tag{4.30}$$

Because we have chosen very smooth mean, variance, and correlation functions, the consistency and $\sqrt{N}$-asymptotic normality are standard; see, for example, Wooldridge (2010, Chapter 14). Remember, $\hat{\boldsymbol{\Psi}}^{-1}$ is an (estimated) optimal weighting matrix given the choice of instruments; the standard GMM inference does not require that $\hat{\mathbf{Z}}_i$ is optimal. A nice byproduct

of the GMM estimation is we can use the overidentification test, which has K overidentification restrictions, to test Assumption CM.

It may be helpful to summarize the estimation steps, which also serves to illustrate the relatively simplicity of the estimator.

Procedure 4.1 (**GMM Estimation**):

1. Use FEP estimation to obtain $\hat{\boldsymbol{\beta}}_{FEP}$.

2. Use $\hat{\boldsymbol{\beta}}_{FEP}$ to construct one of the dependent variables in (4.13). Given a choice of $\mathbf{h}(\mathbf{x}_i)$, with the leading case being $\mathbf{h}(\mathbf{x}_i) = \bar{\mathbf{x}}_i$, use cross-sectional Poisson regression to obtain $\hat{\eta}$ and $\hat{\boldsymbol{\lambda}}$.

3. Compute the fitted values, $\hat{k}_{it} = \exp\left[\mathbf{x}_{it}\hat{\boldsymbol{\beta}}_{FEP} + \hat{\eta} + \mathbf{h}(\mathbf{x}_i)\hat{\boldsymbol{\lambda}}\right]$, and the residuals $\hat{v}_{it}$ in (4.17). Run the simple regression in (4.18) to obtain $\hat{\alpha}$ and $\hat{\delta}$. If $\hat{\delta} < 0$, set $\hat{\delta} = 0$.

4. Compute the estimated correlation matrix, $\hat{\mathbf{R}}$, as in (4.20).

5. Construct the "optimal" IVs as in (4.27).

6. Use the IVs from step (5) in an overidentified GMM estimation with optimal weighting matrix. □

5. A Small Simulation Study

We now present the results of a small Monte Carlo simulation to demonstrate the efficacy of the improved GMM estimator. The conditional mean model, which has an exponential form, includes three time-varying explanatory variables and multiplicative heterogeneity. We consider two conditional distributions for the outcome variable, y_{it}. In the first case, y_{it} is a count variable generated as

$$y_{it}|\mathbf{x}_i, c_i, \mathbf{e}_i \sim \text{Poisson}[c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta} + e_{it})], \tag{5.1}$$

where $\mathbf{e}_i = (e_{i1}, e_{i2}, \ldots, e_{iT})'$ is distributed as multivariate normal with unit variances. In order to generate serial dependence in $\{y_{it} : t = 1, \ldots, T\}$ conditional on $(\mathbf{x}_i, c_i)$, $\{e_{it} : t = 1, 2, \ldots, T\}$ follows an AR(1) process with first-order correlation $\phi \in \{0, 0.25, 0.75\}$. This autoregressive process generates no conditional dependence when $\phi = 0$ and fairly strong time series dependence when $\phi = 0.75$. Because of the inclusion of e_{it}, the conditional distribution $\mathrm{D}(y_{it}|\mathbf{x}_i, c_i)$ is not Poisson; in fact, it exhibits overdispersion because $\exp(e_{it})$ is integrated out in obtaining $\mathrm{D}(y_{it}|\mathbf{x}_i, c_i)$. However, consistency of all estimators requires only that that $\mathrm{E}(y_{it}|\mathbf{x}_i, c_i)$ has the exponential form with multiplicative c_i.

The strictly exogenous explanatory variables, $\mathbf{x}_{it}$, are generated as a trivariate, stationary vector autoregression, where the stochastic term is an independent multivariate standard normal distribution with autocorrelation parameter 0.125. The processes $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ and $\mathbf{e}_i$ are independent. The vector $\boldsymbol{\beta}$ is set to $\boldsymbol{\beta}' = (0.15, 0.25, 0.35)$ (where we drop the o subscript to make the tables easier to read).

To generate correlation between c_i and $\mathbf{x}_i$, we first use an exponential version of the Mundlak (1978) device and an exponential distribution:

$$c_i|\mathbf{x}_i \sim \text{Exponential}[\exp(\eta + \bar{\mathbf{x}}_i\boldsymbol{\lambda})]. \tag{5.2}$$

Under this specification, the working assumptions WH.1 and WH.2 are both satisfied with $\mathbf{h}(\mathbf{x}_i) = \bar{\mathbf{x}}_i$ and, in the case of WH.2, $\delta = 1$.

We estimate the parameters in the heterogeneity moments using a two-step pooled Poisson QMLE with the FEP estimator as the first-stage estimator of $\boldsymbol{\beta}$. The estimates $\hat{\alpha}$ and $\hat{\delta}$ are estimated via the pooled OLS regression in equation (4.18) and $\hat{\mathbf{R}}$ is estimated as in (4.20). When $\hat{\mathbf{R}}$ is not positive definite for a particular draw, we set $\hat{\delta} = 0$ and estimate $\hat{\mathbf{R}}$ as in (4.22) (in which case the value of $\hat{\alpha}$ plays no role in the estimation of $\boldsymbol{\beta}$). This situation occurs

between 60% and 80% of the simulations.

We use $N = 300$, $T \in \{4, 8\}$, and $1,000$ replications in the simulations. The findings are reported in Table 1.

Table 1. **Conditional Poisson Distribution**

			Bias			SD			RMSE		
			FEP	GEFP	GMM	FEP	GEFP	GMM	FEP	GEFP	GMM
$\phi = 0$	$T = 4$	$\beta_1 = 0.15$	0.002	–0.004	0.000	0.082	0.075	0.072	0.082	0.075	0.072
		$\beta_2 = 0.25$	0.001	–0.011	–0.003	0.083	0.078	0.072	0.083	0.079	0.072
		$\beta_3 = 0.35$	–0.001	–0.016	–0.005	0.083	0.079	0.075	0.083	0.081	0.075
	$T = 8$	β_1	0.001	–0.010	–0.005	0.052	0.044	0.041	0.052	0.045	0.041
		β_2	0.000	–0.020	–0.011	0.053	0.044	0.042	0.053	0.049	0.044
		β_3	0.001	–0.027	–0.014	0.051	0.045	0.042	0.052	0.052	0.045
$\phi = 0.25$	$T = 4$	β_1	–0.007	–0.016	0.008	0.081	0.074	0.072	0.081	0.076	0.073
		β_2	–0.003	–0.014	0.004	0.082	0.075	0.070	0.082	0.077	0.070
		β_3	0.002	–0.015	0.003	0.079	0.075	0.070	0.079	0.077	0.070
	$T = 8$	β_1	–0.001	–0.014	–0.007	0.051	0.045	0.042	0.051	0.047	0.043
		β_2	0.000	–0.021	–0.010	0.048	0.044	0.040	0.048	0.049	0.042
		β_3	–0.001	–0.029	–0.015	0.051	0.046	0.043	0.051	0.054	0.046
$\phi = 0.75$	$T = 4$	β_1	–0.001	–0.007	–0.003	0.057	0.054	0.051	0.057	0.055	0.051
		β_2	0.005	–0.008	0.001	0.060	0.058	0.052	0.061	0.059	0.052
		β_3	0.001	–0.014	–0.002	0.060	0.059	0.053	0.060	0.060	0.053
	$T = 8$	β_1	0.001	–0.012	–0.004	0.043	0.035	0.034	0.043	0.037	0.034
		β_2	–0.001	–0.023	–0.011	0.044	0.036	0.034	0.044	0.043	0.036
		β_3	–0.002	–0.032	–0.015	0.047	0.038	0.036	0.047	0.050	0.039

Some general patterns emerge from Table 1. First, the FEP estimator shows very little bias, and its bias is almost always smaller than the GFEP and GMM estimators. The GFEP estimator generally shows the most bias – as high as nine percent in some cases. Still, we only have $N = 300$, which is not especially large. Interestingly, the bias in the GMM estimator – which combines both sets of moment conditions – is well below that of the GFEP estimator. The bias in both the GFEP and GMM estimators appears to increase with T. Overall, the bias in the GMM estimator seems acceptable, especially given the small N.

The GMM estimator always has the smallest sampling standard deviation, sometimes being about 80% of the FEP standard error. The SD of the GFEP estimator falls in between that of the FEP and GMM estimators. In a few cases the FEP estimator has smaller root mean squared error (RMSE) than the GFEP estimator. The asymptotic theory of GMM estimation implies that the GMM estimator is asymptotically more efficient than FEP or GFEP because, in the setting of the simulation, the entire set of working assumptions does not hold, and so GFEP does not use the optimal IVs. The ranking of the estimators in terms of the root mean squared error favors the GMM estimator in every case.

To see how the estimators perform when y_{it} is a continuous outcome, we generated y_{it} as

$$y_{it}|\mathbf{x}_i, c_i, \mathbf{e}_i \sim \text{Gamma}[\exp(\mathbf{x}_{it}\boldsymbol{\beta} + e_{it}), c_i], \tag{5.3}$$

where the gamma distribution is parameterized so that $\text{E}(y_{it}|\mathbf{x}_{it}, c_i, \mathbf{e}_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta} + e_{it})$, as before. The conditional variance is $\text{Var}(y_{it}|\mathbf{x}_{it}, c_i, \mathbf{e}_i) = c_i^2 \exp(\mathbf{x}_{it}\boldsymbol{\beta} + e_{it})$. We use the same process in (5.2) to generate c_i. The simulation findings are reported in Table 2.

Table 2. Conditional Gamma Distribution

			Bias			SD			RMSE		
			FEP	GEFP	GMM	FEP	GEFP	GMM	FEP	GEFP	GMM
$\phi = 0$	$T = 4$	$\beta_1 = 0.15$	0.000	–0.006	–0.002	0.090	0.087	0.081	0.090	0.087	0.081
		$\beta_2 = 0.25$	0.003	–0.008	0.003	0.089	0.085	0.080	0.089	0.085	0.080
		$\beta_3 = 0.35$	0.001	–0.014	0.000	0.090	0.088	0.083	0.090	0.089	0.083
	$T = 8$	β_1	0.000	–0.012	–0.006	0.056	0.049	0.048	0.056	0.051	0.048
		β_2	–0.001	–0.019	–0.009	0.052	0.050	0.047	0.052	0.054	0.048
		β_3	–0.001	–0.027	–0.014	0.054	0.051	0.048	0.054	0.058	0.050
$\phi = 0.25$	$T = 4$	β_1	0.002	–0.007	0.002	0.086	0.082	0.078	0.086	0.082	0.078
		β_2	–0.003	–0.016	–0.004	0.085	0.082	0.077	0.085	0.084	0.078
		β_3	0.002	–0.014	–0.001	0.086	0.084	0.081	0.086	0.085	0.081
	$T = 8$	β_1	0.000	–0.013	–0.006	0.057	0.050	0.048	0.057	0.052	0.048
		β_2	0.000	–0.019	–0.009	0.055	0.050	0.048	0.055	0.053	0.049
		β_3	–0.001	–0.033	–0.017	0.058	0.053	0.051	0.058	0.062	0.053
$\phi = 0.75$	$T = 4$	β_1	0.001	–0.006	0.000	0.069	0.067	0.063	0.069	0.067	0.063
		β_2	0.000	–0.012	–0.001	0.074	0.072	0.067	0.074	0.073	0.067
		β_3	0.000	–0.016	–0.001	0.070	0.072	0.064	0.070	0.074	0.064
	$T = 8$	β_1	0.001	–0.014	–0.005	0.049	0.041	0.040	0.049	0.044	0.040
		β_2	0.000	–0.023	–0.008	0.048	0.042	0.039	0.048	0.048	0.040
		β_3	–0.001	–0.034	–0.013	0.050	0.046	0.043	0.050	0.057	0.045

The general pattern found in Table 1 continues to hold in Table 2. The FEP estimator generally has the lowest bias, although the GMM estimator also does well with bias. The GFEP estimator, which uses only the "optimal" IVs, shows more bias – again, sometimes on the order of more than nine percent. In terms of precision and RMSE, the GMM estimator outperforms FEP and GFEP in all scenarios, although the gains are modest in some cases.

We tried several additional scenarios, including cases where Assumption WH.2 is violated – by drawing c_i from a Poisson distribution – and cases where, conditional on $(\mathbf{x}_i, c_i)$ – y_{it} is an underdispersed gamma random variable. In the former case, we found only minor differences among the estimators, although sometimes the FEP estimator outperformed the other two in terms of RMSE. In the latter case, where we did not allow serial correlation, the estimators perform very similarly. As a final set of simulations, we misspecified the conditional mean $\mathrm{E}(c_i|\mathbf{x}_i)$ in (5.2) by letting the mean depend on the average of the first and last time periods rather than $\bar{\mathbf{x}}_i$. In other words, Assumption WH.1 is violated. The GMM estimator uniformly performed the best based on RMSE and exhibited biases on the order of those reported in Tables 1 and 2. These simulations are available upon request from the authors.

6. Empirical Example

In this section we apply the FEP, GFEP, and GMM estimators to the patents-R&D data set used in Martin (2017), who updated the HHG (1984) data to include $N = 848$ firms for the $T = 8$ years 1996 to 2003. Martin (2017) estimates a static model for the patents-R&D relationship using the FEP estimator. Here we estimate a model that includes two lags of the natural log of R&D, which means in estimation we use data for 1998 through 2003. The estimated second lag is small and statistically insignificant in all cases, but we include it to make precision comparisons even when an estimated coefficient is not statistically significant.

We include a full set of year dummies.

The results for the three estimation methods are given in Table 3, with the coefficients on the year dummies suppressed. In addition to the estimated elasticities for the contemporaneous effect and each of the two lags, the long-run elasticity estimates and standard errors are also provided.

Table 3. **Estimates of the Patents-R&D Relationship**

Outcome Variable: Number of Patents Assigned			
	(1)	(2)	(3)
	FEP	GFEP	GMM
log(*rnd*)	0.1604 (0.0468)	0.1288 (0.0500)	0.1763 (0.0418)
log(*rnd*_1)	0.0596 (0.0386)	0.1053 (0.0537)	0.0705 (0.0333)
log(*rnd*_2)	−0.0009 (0.0706)	0.0191 (0.0599)	0.0022 (0.0350)
Long-Run Elasticity	0.2191 (0.0976)	0.2532 (0.0881)	0.2490 (0.0679)

As is well known in the empirical literature on the patents-R&D relationship, estimating the distributed lag coefficients precisely is challenging because R&D spending tends to move slowly over time (after removing aggregate trends). This is especially true with fixed effects methods, which rely on within-firm variation. Each of the three estimation methods produces a positive and statistically significant estimate of the impact effect, with the GMM estimator providing the most precise estimate (0.176 with $\text{SE} = 0.042$). Of the three estimates of the coefficient on the first lag, only the GMM estimate is statistically significant at the usual 5% significance level ($t = 2.12$). The GMM standard error is notably below the GFEP standard error and also less than the FEP standard error. That the FEP standard errors on log(*rnd*) and log(*rnd*_1) are below those of the corresponding GFEP standard errors suggests that the

working assumptions used to generate the "optimal" IVs are violated. By contrast, the preferred GMM estimator that combines the two sets of moment conditions has notably more precision than the other two estimators.

The long-run (LR) elasticity is of some interest in these studies. The estimated LR elasticities for FEP, GFEP, and GMM are 0.219, 0.253, and 0.249, respectively, which are reasonably close when accounting for sampling error. Notably, the standard error for the GMM estimate is only about 70% of the FEP standard error and about 77% of the GFEP standard error. This represents a substantial improvement in precision by using the new GMM estimator compared with the FEP estimator for estimating the LR elasticity. Moreover, the efficiency gains are in line with the simulations in Section 5 – and actually a bit better in the application, perhaps reflecting a more complicated conditional variance or pattern of serial correlation, or more complicated conditional moments for the heterogeneity.

7. Summary and Conclusion

We have characterized the optimal instruments in a multiplicative panel model under a general set of working assumptions. The variance-mean relationship, conditional on unobserved heterogeneity as well as covariates, is allowed to be any positive number. The conditional correlation matrix is assumed to be constant but is otherwise unrestricted. Under these assumptions, the optimal IVs depends only on the unknown correlation matrix, $\mathbf{R}$ (and the value of the conditional mean parameters, $\boldsymbol{\beta}_o$). In the special case that $\mathbf{R} = \mathbf{I}_T$, we show that the FEP estimator achieves the asymptotic efficiency bound for any amount of overdispersion or underdispersion. Even by itself, this result represents an important improvement in our understanding of the efficiency properties of the popular FEP estimator. When $\mathbf{R}$ is not the identity matrix, it is possible to improve on the FEP estimator.

To operationalize the optimal IVs in order to exploit serial correlation, we add working first and second moment assumptions on the conditional heterogeneity distribution. These assumptions are common in literatures that allows nonnegative heterogeneity in cross-sectional and panel data models. We show that estimating the optimal IVs is straightforward, and suggest a GMM approach that is guaranteed to improve asymptotic efficiency whether or not serial correlation is present. Our simulations show that the GMM estimator that combines the FEP moment conditions and the new "optimal" moment conditions has very good bias properties and provides nontrivial efficiency gains – even when the cross-sectional sample size is only $N = 300$. In our empirical example, we find the GMM estimator produces a standard error of the long-run elasticity 30% lower than the FEP estimator – a nontrivial improvement.

Our results and new estimator are appealing for cases where N is substantially larger than T, as we have used the standard microeconometric setting where T is fixed in the asymptotic analysis. Fernández-Val and Weidner (2017) and Chen, Fernández-Val, and Weidner (2020) have proposed quasi-MLEs that allow more heterogeneity. However, consistency requires $T \to \infty$ along with $N \to \infty$, and necessarily restricts the amount of time series heterogeneity and dependence.

References

Castillo, J.C., D. Mejia, and P. Restrepo (2020), "Scarcity without Leviathan: The Violent Effects of Cocaine Supply Shortages in the Mexican Drug War," *Review of Economics and Statistics* 102, 269–286.

Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data," *Review of Economic Studies* 47, 225–238.

Chamberlain, G. (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics* 34, 305–334.

Chamberlain, G. (1992), "Efficiency Bounds for Semiparametric Regression." *Econometrica* 60, 567–596.

Chen, M., I. Fernández-Val, and M. Weidner (2020), "Nonlinear Factor Models for Network and Panel Data" forthcoming, *Journal of Econometrics*.

Fernández-Val, I. and M. Weidner (2016), "Individual and Time Effects in Nonlinear Panel Models with Large N, T," *Journal of Econometrics* 192, 291–312.

Hahn, J. (1997), "A Note on the Efficient Semiparametric Estimation of Some Exponential Panel Models," *Econometric Theory* 13, 583–588.

Hausman, J.A., B.H. Hall, and Z. Griliches (1984), "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica* 52, 908–938.

Hardin, J.W. and J.M. Hilbe (2012), *Generalized Estimation Equations*, second edition. London: Chapman and Hall.

Krapf, M., H.W. Ursprung, and C. Zimmermann (2017), "Parenthood and Productivity of Highly Skilled Labor: Evidence from the Groves of Academe," *Journal of Economic Behavior & Organization* 140, 147–175.

Liang, Y.-K. and S.L. Zeger (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika* 73, 13-22.

Martin, R.S. (2017), "Estimation of Average Marginal Effects in Multiplicative Unobserved Effects Panel Models," *Economics Letters* 160, 16-19.

McCabe, M.J. and C.M. Snyder (2015), "Does Online Availability Increase Citations? Theory and Evidence from a Panel of Economics and Business Journals," *Review of Economics and Statistics* 97, 144–165.

McCabe, M.J. and C.M. Snyder (2014), "Identifying The Effect of Open Access on Citations Using a Panel of Science Journals," *Economic Inquiry* 52, 1284-1300.

McCullagh, P. and J.A. Nelder (1989), *Generalized Linear Models*, second edition. London: Chapman and Hall.

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

Newey, W.K. (2001), "Conditional Moment Restrictions in Censored and Truncated Regression Models," *Econometric Theory* 17, 863-888.

Newey, W.K. and D.L. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Volume 4, R.F. Engle and D.L. McFadden (eds.). Amsterdam: North-Holland, 2111-2245.

Schlenker, W., and W.R. Walker (2016), "Airports, Air Pollution, and Contemporaneous Health," *Review of Economic Studies* 83, 768-809.

Sherman, J., and W.J. Morrison (1950), "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix," *Annals of Mathematical Statistics* 21, 124-127.

Verdier, V. (2018), “Local Semi-Parametric Efficiency of the Poisson Fixed Effects Estimator,” *Journal of Econometric Methods* 7, 1-10.

Williams, M.L., P. Burnap, A. Javed, H. Liu, and S. Ozalp (2020), “Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime,” *British Journal of Criminology* 60, 93-117.

Wooldridge, J.M. (1999), “Distribution-Free Estimation of Some Nonlinear Panel Data Models,” *Journal of Econometrics* 90, 77-97.

Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, second edition. Cambridge, MA: MIT Press.

Appendix

This appendix collects together proofs of the formal results stated in the text.

Proof of Lemma 3.1

From equation (3.13), Assumptions WV.1 and WV.2 imply

$$\mathrm{Var}(\mathbf{y}_i|\mathbf{x}_i, c_i) = \alpha c_i \mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2}.$$

By the law of total variance,

$$\begin{aligned}\mathrm{Var}(\mathbf{y}_i|\mathbf{x}_i) &= \mathrm{E}[\mathrm{Var}(\mathbf{y}_i|\mathbf{x}_i, c_i)|\mathbf{x}_i] + \mathrm{Var}[\mathrm{E}(\mathbf{y}_i|\mathbf{x}_i, c_i)|\mathbf{x}_i] \\ &= \mathrm{E}\left(\alpha c_i \mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2}\,\middle|\,\mathbf{x}_i\right) + \mathrm{Var}(c_i\mathbf{m}_i|\mathbf{x}_i) \\ &= \alpha\mu_c(\mathbf{x}_i)\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2} + \sigma_c^2(\mathbf{x}_i)\mathbf{m}_i\mathbf{m}_i'. \end{aligned} \tag{A.1}$$

To simplify notation in what follows, write $\mu_i \equiv \mu_c(\mathbf{x}_i)$, $\sigma_i^2 \equiv \sigma_c^2(\mathbf{x}_i)$. To derive $\mathbf{\Omega}_i^{-1}$, we apply an implication of Sherman and Morrison (1950): For a nonsingular $T \times T$ matrix $\mathbf{A}$ and $T \times 1$ vector $\mathbf{b}$,

$$(\mathbf{A} + \mathbf{b}\mathbf{b}')^{-1} = \mathbf{A}^{-1} - \frac{1}{1 + \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}}\mathbf{A}^{-1}\mathbf{b}\mathbf{b}'\mathbf{A}^{-1}, \tag{A.2}$$

which can be verified by direct multiplication. Take $\mathbf{A} \equiv \alpha\mu_i\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2}$ and $\mathbf{b} \equiv \sigma_i\mathbf{m}_i$ in (A.2) and note that $\left[\alpha\mu_i\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2}\right]^{-1} = \mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}/(\alpha\mu_i)$ and $\mathbf{M}_i^{-1/2}\mathbf{m}_i = \sqrt{\mathbf{m}_i}$. Therefore,

$$
\begin{aligned}
\boldsymbol{\Omega}_i^{-1} &= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} \\
&\quad - \frac{1}{1+\left[\sigma_i^2\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\right]/(\alpha\mu_i)}\sigma_i^2\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}/(\alpha\mu_i)^2 \\
&= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} \\
&\quad - \frac{\sigma_i^2}{\alpha\mu_i+\sigma_i^2\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}}\sigma_i^2\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}/(\alpha\mu_i) \\
&= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\left\{\mathbf{R}^{-1} - \frac{\sigma_i^2}{\left[\alpha\mu_i+\sigma_i^2\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\right]}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\right\}\mathbf{M}_i^{-1/2}.
\end{aligned}
$$

Proof of Theorem 3.1

Simplify the notation by defining $\mathbf{D}_i \equiv \mathbf{D}_o(\mathbf{x}_i)$, $\mathbf{V}_i \equiv \mathbf{V}_o(\mathbf{x}_i)$, $\mu_i \equiv \mu_c(\mathbf{x}_i)$, $\sigma_i^2 \equiv \sigma_c^2(\mathbf{x}_i)$, and drop dependences on $\boldsymbol{\beta}_o$. With this simplified notation,

$$
\mathbf{V}_i^- = \boldsymbol{\Omega}_i^{-1} - \boldsymbol{\Omega}_i^{-1}\mathbf{m}_i(\mathbf{m}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i)^{-1}\mathbf{m}_i'\boldsymbol{\Omega}_i^{-1}
$$

and, from Lemma 3.1,

$$
\boldsymbol{\Omega}_i^{-1} = \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} - \frac{\sigma_i^2}{\alpha\mu_i(\alpha\mu_i+a_i\sigma_i^2)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}
$$

where $a_i \equiv \sqrt{\mathbf{m}_i}'\mathbf{R}_i^{-1}\sqrt{\mathbf{m}_i}$. Therefore, because $\mathbf{M}_i^{-1/2}\mathbf{m}_i = \sqrt{\mathbf{m}_i}$,

$$
\begin{aligned}
\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i &= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} - \frac{\sigma_i^2}{\alpha\mu_i(\alpha\mu_i+a_i\sigma_i^2)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} \\
&= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} - \frac{a_i\sigma_i^2}{\alpha\mu_i(\alpha\mu_i+a_i\sigma_i^2)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} \\
&= \left[\frac{1}{\alpha\mu_i} - \frac{a_i\sigma_i^2}{\alpha\mu_i(\alpha\mu_i+a_i\sigma_i^2)}\right]\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} \\
&= \frac{[(\alpha\mu_i+a_i\sigma_i^2)-a_i\sigma_i^2]}{\alpha\mu_i(\alpha\mu_i+a_i\sigma_i^2)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} \\
&= \frac{1}{(\alpha\mu_i+a_i\sigma_i^2)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}
\end{aligned}
$$

Also,

$$\mathbf{m}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i = \frac{1}{(\alpha\mu_i + a_i\sigma_i^2)}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} = \frac{a_i}{\alpha\mu_i + a_i\sigma_i^2}.$$

It follows that

$$\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i(\mathbf{m}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i)^{-1}\mathbf{m}_i'\boldsymbol{\Omega}_i^{-1} = \frac{1}{a_i(\alpha\mu_i + a_i\sigma_i^2)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}$$

Plugging into $\mathbf{V}_i^-$ gives

$$\begin{aligned}
\mathbf{V}_i^- &= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} - \frac{\sigma_i^2}{\alpha\mu_i(\alpha\mu_i + a_i\sigma_i^2)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} \\
&\quad - \frac{1}{a_i(\alpha\mu_i + a_i\sigma_i^2)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} \\
&= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} - \left[\frac{\alpha\mu_i + a_i\sigma_i^2}{a_i\alpha\mu_i(\alpha\mu_i + a_i\sigma_i^2)}\right]\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} \\
&= \frac{1}{\alpha\mu_i}\left[\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} - \frac{1}{a_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}\right],
\end{aligned}$$

which completes the result for $\mathbf{V}_i^-$. From (3.10), the optimal IVs are

$$\mathbf{D}_i'\mathbf{V}_i^- = -\mu_i\nabla_{\boldsymbol{\beta}}\mathbf{m}_i'\mathbf{V}_i^- = -\frac{1}{\alpha}\nabla_{\boldsymbol{\beta}}\mathbf{m}_i'\left[\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} - \frac{1}{a_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}\right],$$

and we can drop $-1/\alpha$ and factor out $\mathbf{M}_i^{-1/2}$ to get the result. □

Proof of Corollary 3.1

Putting $\mathbf{R} = \mathbf{I}_T$ into (3.16) and using simple algebra gives the optimal IVs as

$$\mathbf{Z}^*(\mathbf{x}_i)' = \nabla_{\boldsymbol{\beta}}\mathbf{m}_i(\boldsymbol{\beta}_o)'\left(\mathbf{M}_i^{-1} - \frac{1}{\sum_{r=1}^{T} m_{ir}}\mathbf{1}_T\mathbf{1}_T'\right).$$

We show that this choice of instruments leads to the FEP first order condition, as expressed by Wooldridge (1999), using the definition of $\mathbf{W}_i$ given in Section 2:

$$\nabla_{\boldsymbol{\beta}}\mathbf{p}_i(\boldsymbol{\beta}_o)'\mathbf{W}_i = \nabla_{\boldsymbol{\beta}}\mathbf{m}_i(\boldsymbol{\beta}_o)'\left[\mathbf{I}_T - \mathbf{1}_T\mathbf{p}_i(\boldsymbol{\beta}_o)'\right]\mathbf{M}_i^{-1}$$

To see the equivalence, note that

$$\mathbf{1}_T \mathbf{p}_i(\boldsymbol{\beta}_o)' \mathbf{M}_i^{-1} = \frac{1}{\left(\sum_{r=1}^{T} m_{ir}\right)} \begin{pmatrix} \mathbf{m}_i \\ \mathbf{m}_i \\ \vdots \\ \mathbf{m}_i \end{pmatrix} \mathbf{M}_i^{-1} = \frac{1}{\sum_{r=1}^{T} m_{ir}} \mathbf{1}_T \mathbf{1}_T'$$

and so

$$\nabla_{\boldsymbol{\beta}} \mathbf{p}_i(\boldsymbol{\beta}_o)' \mathbf{W}_i = \nabla_{\boldsymbol{\beta}} \mathbf{m}_i(\boldsymbol{\beta}_o)' \left(\mathbf{M}_i^{-1} - \frac{1}{\sum_{r=1}^{T} m_{ir}} \mathbf{1}_T \mathbf{1}_T' \right) = \mathbf{Z}^*(\mathbf{x}_i)'. \square$$