NEW ESTIMATION METHODS FOR PANEL DATA MODELS

By

Valentin Verdier

A DISSERTATION

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

Economics - Doctor of Philosophy

2014

ABSTRACT

NEW ESTIMATION METHODS FOR PANEL DATA MODELS

By

Valentin Verdier

This dissertation is composed of three chapters that develop new estimation methods for several models of panel data. The first and third chapters are mainly concerned with understanding and aproximating the structure of optimal instruments for estimating dynamic panel data models with cross-sectional dependence in the case of the first chapter, and non-linear panel data models with strictly exogeneous covariates in the case of the third chapter. The second chapter is concerned with additional restrictions that can be used to estimate non-linear dynamic panel data models.

The first chapter considers the estimation of dynamic panel data models when data are suspected to exhibit cross-sectional dependence. A new estimator is defined that uses cross-sectional dependence for efficiency while being robust to the misspecification of the form of the cross-sectional dependence. I show that using cross-sectional dependence for estimation is important to obtain an estimator that is more accurate than existing estimators. This new estimator also uses nuisance parameters parsimoniously so that it exhibits good small sample properties even when the number of available moment conditions is large. As an empirical application, I estimate the effect of attending private school on student achievement using a value added model.

The second chapter considers the instrumental variable estimation of non-linear models of panel data with multiplicative unobserved effects where instrumental variables are predetermined as opposed to strictly exogenous. Existing estimators for these models suffer from a weak instrumental variable problem, which can cause them to be too inaccurate to be reliable. In this chapter I present additional sets of restrictions that can be used for more precise estimation. Monte Carlo simulations show that using these additional moment conditions improves the precision of the estimators significantly and hence should facilitate the use of these models.

In the third chapter I study the efficiency of the Poisson Fixed Effects estimator. The Poisson fixed effects estimator is a conditional maximum likelihood estimator and as such is consistent under specific distributional assumptions. It has also been shown to be consistent under significantly weaker restrictions

on the conditional mean function only. I show that the Poisson Fixed Effects estimator is asymptotically efficient in the class of estimators that are consistent under restrictions on the conditional mean function, as long as the assumptions of equal conditional mean and variance and zero conditional serial correlation are satisfied. I then define another estimator that is optimal under more general conditions. I use Monte Carlo simulations to investigate the small-sample performance of this new estimator compared to the Poisson fixed effects estimator.

ACKNOWLEDGEMENTS

I particularly thank Jeffrey Wooldridge who served as the chair of my dissertation committee. His teaching throughout my studies as Michigan State University shaped this dissertation and my current research. I also thank my other committee members, Peter Schmidt, Timothy Vogelsang and Robert Myers, whose help at various stages of my research had a large positive impact on the quality of my work.

I also thank graduate students in the department of economics and the department of agricultural economics for helpful conversations.

Finally I thank Margaret Lynch and Lori Jean Nichols, who work in the administrative staff of the department of economics, whose continuous support for five years helped a lot in completing this degree.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

ESTIMATION OF DYNAMIC PANEL DATA MODELS WITH CROSS-SECTIONAL DEPENDENCE

1.1 Introduction

In some econometric studies of panel data, researchers might want to account for the presence of feedback between the dependent variable and explanatory variables, i.e. for current values of the dependent variable to affect future values of the explanatory variables or even for both dependent and independent variables to be jointly determined. The simplest example of such models is the dynamic panel data model where lagged values of the dependent variable are used as covariates. In such cases, explanatory variables can not be treated as strictly exogenous. In virtually all panel data applications, researchers also want to control for unobserved heterogeneity that affects the dependent variable but might also be correlated with the covariates.

The presence of both non strictly exogenous covariates and unobserved heterogeneity in panel data models causes many estimation methods to be invalid (see for instance Wooldridge (2010)). In the context of cross-sectionally independent data, a valid estimator for dynamic panel data models that relies on first differencing and instrumental variables has been defined in early work by Anderson and Hsiao (1981) Anderson and Hsiao (1981). Additionally, an asymptotically efficient estimator is found in Arellano and Bond (1991)[1]. In the rest of the paper, we refer to this estimator as the AB estimator. These estimators often suffer from having a large variance because the instrumental variables that they use are weak.[2] In addition, inference for the AB estimator is often unsatisfactory when the number of time periods in the data set is relatively large because of problems due to using many moment conditions, as studied in Alvarez and Arellano (2003) or Windmeijer (2005) for the case of cross-sectional independence.

In this paper, we consider the estimation of panel data models with covariates that are not strictly exogenous when data also exhibit cross-sectional dependence. We will define a new estimator that is more

[1] The Arellano and Bond estimator is asymptotically efficient in the class of estimators using linear functions of the instruments.

[2] To address this problem, papers such as Ahn and Schmidt (1995), Arellano and Bover (1995), and Blundell and Bond (1998) considered using for estimation additional assumptions such as homoscedasticity, uncorrelation of the transitory shocks, or restrictions on initial conditions. Another approach to obtain efficiency gains by using additional assumptions can be found in the literature on First Difference Quasi-Maximum Likelihood estimation, as in Hsiao et al. (2002) for instance which relies on assumptions of homoscedasticity and serial uncorrelation. We do not consider these estimators here since we am interested in estimators that are consistent under the only assumption of mean independence of the transitory shock, without any other assumption holding.

efficient than the AB estimator and for which inference is significantly better in small samples. The main reason why our estimator is more efficient than previous estimators that were defined for data with cross-sectional independence is that it makes use of cross-sectional dependence to obtain stronger instruments.

In order to obtain an estimator with not only good properties in terms of point estimation, but also good properties for inference, we use an auxiliary model for optimal instruments. Optimal instruments are instruments that, once interacted with corresponding moment functions, provide an optimal set of exactly identifying moment conditions so that the resulting estimator achieves the asymptotic efficiency bound for estimating unknown parameters from the assumption of mean independence of the transitory shocks. Optimal instruments for estimating dynamic panel data models without cross-sectional dependence are found in Chamberlain (1992a) and they can be generalized to the case of cross-sectional dependence. In this paper, we propose auxiliary assumptions sufficient to model optimal instruments for panel data models with covariates that are not strictly exogenous and cross-sectional dependence. The advantage of such an approach is that it provides a systematic way of weighting many moment conditions while making use of few nuisance parameters. As a result, our estimator exhibits good small sample properties and inference while being robust to the misspecification of our model of optimal instruments.

Arellano (2003) and Alvarez and Arellano (2004) have previously considered modeling optimal instruments for dynamic panel data models in the special case of cross-sectional independence. We show that cross-sectional dependence can be particularly useful to obtain more accurate estimators. Previous work on dynamic panel data models that has considered cross-sectional dependence has not made use of this dependence to obtain stronger instruments. Mutl (2006), for instance, studied a GMM estimator based on the same moment conditions as in Anderson and Hsiao (1981) or Arellano and Bond (1991) and only uses an optimal weighting matrix based on a specific model of spatial dependence. Elhorst (2005) and Su and Yang (2013) generalized maximum likelihood estimators as in Hsiao et al. (2002) to the case of cross-sectional dependence but these estimators are not robust to heteroscedasticity, serial correlation of the transitory shocks or misspecification of the cross-sectional dependence.

In Section 1.2, we present the simplest example of the models we consider, the dynamic panel data model without covariates for data with cross-sectional dependence. In Section 1.3, we define our estimator and compare it to existing estimators. In Section 1.4, we generalize our estimator to general models with non strictly exogenous covariates. In Section 1.5, we present Monte-Carlo evidence that the efficiency gains from using cross-sectional dependence for estimation can be significant and that the estimator we propose

has superior small sample properties compared to existing estimators. In Section 1.6, we apply our estimator to the estimation of the effect of attending private school on student achievement using a value-added model and taking into account the possibility that student achievements are correlated within schools.

1.2 Dynamic Panel Data Models with Cross-Sectional Dependence

1.2.1 The Model

Throughout the paper we will consider large n, fixed T asymptotics.[3] Consider first the model for any observation i from a sample of n observations and any time period t from a fixed number T of time periods:

$$y_{it} = \rho_0 y_{it-1} + c_i + u_{it}\ t = 1, ..., T \tag{1.2.1}$$

$$E(u_{it}|Y_{t-1}) = 0\ t = 1, ..., T \tag{1.2.2}$$

where $Y_t = [Y_{1t}', ..., Y_{nt}']'$ and $Y_{it} = [y_{i0}, ..., y_{it}]'$ are random vectors that stack values of y_{it} across time and observations and c_i are time constant unobserved effects, also called unobserved heterogeneity. We also assume that $\rho_0 \neq 1$ so that ρ_0 is identified from differenced equations as seen in the next subsection.

In the case where there is no cross-sectional dependence, (1.2.1) and (1.2.2) correspond to the linear dynamic model for panel data as presented in Arellano and Bond (1991) for instance. When there is cross-sectional dependence, (1.2.1) and (1.2.2) impose the restriction that cross-sectional dependence does not cause Y_{t-1} to be endogenous.

For instance if contemporaneous spatial lags were omitted variables in (1.2.1), then (1.2.2) would be violated. Some papers such as Cizek et al. (2011), Elhorst (2005), Su and Yang (2013) and Baltagi et al. (2014) have considered models with both dynamic effects and contemporaneous spatial lag effects. Since estimators for such models rely on correct specification of the form of cross-sectional dependence, we do not consider them here and concentrate on models where cross-sectional dependence of some unknown form is present in the residuals.[4] Lagged values of the dependent variable of neighboring observations could also be

[3] Using a parsimonious number of nuisance parameters seems to grant the estimator we propose good properties with relatively large numbers of time periods but a formal derivation of results under large N, large T asymptotics is left for future research.

[4] It is also important to note that, with cross-sectional dependence, it is not likely for $E(u_{it}|Y_{it-1}) = 0$ to hold without (1.2.2) holding. If (1.2.2) is not satisfied, it is likely that both estimators for cross-sectionally independent data such as the Arellano and Bond estimator and the alternative estimator proposed in this chapter will be inconsistent. For instance suppose for simplicity that $n = 2$ and $E(u_{1t}|Y_{t-1}) = \alpha + \beta_1 y_{1t-1} + \beta_2 y_{2t-1} \neq 0$ so that $\beta_1 \neq 0$ or $\beta_2 \neq 0$. Then $E(u_{1t}|Y_{1t-1}) = \alpha + \beta_1 y_{1t-1} + \beta_2 E(y_{2t-1}|Y_{1t-1})$ and it likely that $E(y_{2t-1}|Y_{1t-1})$ be a function of $y_{10}, ..., y_{1t-2}$ in addition to y_{1t-1} so that, in general, $\alpha + \beta_1 y_{1t-1} \neq -\beta_2 E(y_{2t-1}|Y_{1t-1})$ and $E(u_{1t}|y_{1t-1}) \neq 0$.

included in the model as covariates to control for dynamic cross-sectional effects. We will discuss models with covariates in Section 1.4.

The objective of the next section is to characterize estimators for ρ_0 that are consistent when (1.2.1) and (1.2.2) hold under general conditions on the form of cross-sectional dependence in c_i and u_{it}.

1.2.2 Consistent Estimation

The presence of unobserved heterogeneity rules out estimating ρ_0 by a regression. Because (1.2.1) and (1.2.2) form a dynamic model, fixed effects estimation is also ruled out because explanatory variables are not strictly exogenous.

To estimate ρ_0, we will consider a first difference transformation. All of the derivations in this paper can be generalized to other transformations, such as the forward filtering transformation presented in Arellano and Bover (1995) for instance, which can be useful in the case of unbalanced panels. Define:

$$m_{it}(\rho) = \Delta y_{it} - \rho \Delta y_{it-1} \; \forall t = 2, ..., T \tag{1.2.3}$$

where Δ is the first difference operator. Therefore, $m_{it}(\rho_0) = u_{it} - u_{it-1}$ and (1.2.1) and (1.2.2) imply:

$$E(m_{it}(\rho_0)|Y_{t-2}) = 0 \; \forall t = 2, ..., T \tag{1.2.4}$$

Define $m_i(\rho) = [m_{it}(\rho)]_{t=2,...,T}$ to be the column vector with $m_{it+1}(\rho)$ as its t^{th} element. Sometimes we will also shorten notation by writing $m_i = m_i(\rho_0)$, $m_{it} = m_{it}(\rho_0)$ and $\Delta Y_{-1,i} = [\Delta y_{it-1}]_{t=2,...,T}$.

Define:

$$Z_i = [Z_{i2}, ..., Z_{iT}] \tag{1.2.5}$$

to be a matrix containing instruments for each time period so that Z_{it} is a function of Y_{t-2} and therefore $E(Z_{it} m_{it}(\rho_0)) = 0$ and $E(Z_i m_i(\rho_0)) = \sum_{t=1}^{T} E(Z_{it} m_{it}(\rho_0)) = 0$.[5]

Define Ξ to be some weighting matrix. Define $\hat{\rho}$ an estimator for ρ_0 as:

$$\hat{\rho} = argmin_{\rho} \Big(\sum_{i=1}^{n} Z_i m_i(\rho) \Big)' \Xi \sum_{i=1}^{n} Z_i m_i(\rho) \tag{1.2.6}$$

Consider first the case where cross-sectional dependence is captured by a large group of clusters with fixed numbers of observations so that observations within a cluster might be related but observations across

[5]Note that we need to assume $\rho_0 \neq 1$ for $E(Z_i m_i(\rho)) = 0$ to hold for $\rho = \rho_0$ only since if $\rho_0 = 1$ then $E(Z_i m_i(\rho)) = 0 \; \forall \rho$.

clusters are independent. Standard results on asymptotic properties of GMM estimators with clustering, found in White (2001) for instance, imply that $\hat{\rho}$ will be consistent for ρ_0 and asymptotically normal as the number of clusters grows unboundedly under standard regularity conditions.

For more general forms of cross-sectional dependence, Conley (1999), Jenish and Prucha (2009), Jenish and Prucha (2012) consider different sets of regularity conditions that guarantee that $\hat{\rho}$ is consistent and asymptotically normal as long as $E(Z_i m_i(\rho_0)) = 0$.

In this paper, we will assume that either set of regularity conditions holds so that the probability limits $D = plim(\frac{1}{n}\sum_{i=1}^{n} Z_i \Delta Y_{-1,i})$ and $\Upsilon = plim\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} Z_i m_i m_j' Z_j'$ exist and are finite, $D'\Xi D \neq 0$, $\hat{\rho} \xrightarrow{p} \rho_0$ and, as $n \to \infty$:[6]

$$\sqrt{n}(\hat{\rho} - \rho_0) \xrightarrow{d} N(0, V) \tag{1.2.7}$$

$$V = (D'\Xi D)^{-1} D'\Xi\Upsilon\Xi D (D'\Xi D)^{-1} \tag{1.2.8}$$

In the next sections we consider efficient feasible GMM estimation where the matrix of instruments Z_i and an estimator of the weighting matrix Ξ are chosen so that the resulting estimator of ρ_0 is efficient under some auxiliary assumptions. It is important to note that all of the estimators we propose will be asymptotically equivalent to estimators of the type defined by (1.2.6) so that they will be consistent as long as (1.2.1) and (1.2.2) hold, independently of whether the auxiliary models we specify are true or not.

1.3 Efficient Estimation under Clustering

In this section, we consider an auxiliary model for deriving optimal instruments that assumes that every observation belongs to one of a large number of clusters. Observations are treated as correlated within clusters but independent across clusters. While clustering only represents a specific form of cross-sectional dependence, it might be a good approximation for more general forms of dependence in many applications. In addition, the method outlined in this section for the special case of clustering can easily be extended to other forms of cross-sectional dependence. Therefore we restrict our attention in this paper to auxiliary models that make use of the clustering assumption.

For simplicity we will consider in this section the case where each observation belongs to the same cluster across all time periods but the results in this section can be generalized to clusters changing over

[6]Note that in the case of clustering we consider $\{n_g\}_{g=1,\ldots,G}$ to be a set of fixed values, where n_g denotes the number of observations in cluster g and G the number of clusters. Then $\sqrt{n}$-asymptotic normality or $\sqrt{G}$-asymptotic normality are equivalent since $n/min\{n_g\} \leq G \leq n/max\{n_g\}$.

time as shown in Section 1.6. Previous work that estimated dynamic models of panel data with clustered sampling generally used estimators developed for i.i.d. data such as the ones found in Anderson and Hsiao (1981), Arellano and Bond (1991), or Ahn and Schmidt (1995), and adjusted inference by using clustered standard errors. Such an analysis can be found for instance in de Brauw and Giles (2008) where farming households are treated as clustered by village or Andrabi et al. (2011) where students are clustered by school[7]. Topalova and Khandelwal (2010) and Balasubramanian and Sivadasan (2010) consider the case where firms are clustered by industry.

In this section, we show that there is much to gain in terms of efficiency by using a different estimator that takes into account correlation within cluster but is robust to misspecification of the form of this correlation. We will consider the case where the data is composed of a large number of clusters indexed by $g = 1, ..., G$, each with a fixed number of observations denoted n_g so that asymptotics will be performed for $G \to \infty$.

In the first subsection, we present the special case of two time periods since in this case the problem reduces to estimating ρ_0 from only one differenced equation using instrumental variables.

1.3.1 Special Case of Independent Disturbances and $T = 2$

For this simple special case, we derive an efficient estimator for the case where $\{u_{it}\}_{i=1,...,n,t=1,2}$ are independent both cross-sectionally and across time, where $T = 2$ and where we have conditional homoscedasticity so that:

$$Var(u_{it}|Y_{t-1}) = \sigma_u^2 \ \forall t = 1,2 \tag{1.3.1}$$

When $T = 2$, there is only one differenced equation that can be used for estimation:

$$\Delta y_{i2} = \rho_0 \Delta y_{i1} + \Delta u_{i2} \tag{1.3.2}$$

for which the available instruments are Y_0. Under the assumption of independence of disturbances and homoscedasticity, Δu_{i2} is also cross-sectionally independent and homoscedastic, so the optimal instrument for the differenced equation is the best prediction of Δy_{i1} based on all the available instruments, i.e. $E(\Delta y_{i1}|Y_0)$.

To find $E(\Delta y_{i1}|Y_0)$, note that under (1.2.1) and (1.2.2), $y_{i1} = \rho_0 y_{i0} + c_i + u_{i1}$ so that $E(\Delta y_{i1}|Y_0) = (\rho_0 - 1)y_{i0} + E(c_i|Y_0)$. Therefore the quality of the prediction of Δy_{i1} based on the instruments will depend on

[7]We will show in Section 1.6 however that the clustering used in Andrabi et al. (2011) is not appropriate to obtain robust standard errors due to observations moving across clusters during the period of observation. We will show robust standard errors that take this factor into consideration.

the quality of the prediction of c_i based on the instruments. In many applications, it is very likely that agents that belong to the same cluster will have levels of unobserved heterogeneity that are related. For instance, farmers that live in the same village might farm plots with with similar soil quality or develop similar farming practices over time. Firms that operate in the same industry might also face similar constraints such as for instance regulation or access to skilled labor force. Similarly, households that live in the same district might have been selected based on common characteristics such as wealth, income, family status or values. Therefore, in many applications, we can expect that using information from other observations in the same cluster in addition to one's own previous outcomes can provide a better predictor for one's level of unobserved heterogeneity.

For this simple case, we could derive an optimal predictor for c_i by using the assumption that for any observation i belonging to cluster g we have:

$$c_i = c_g + e_i \tag{1.3.3}$$

where $\{c_g\}_{g=1,...,G}$ forms a sequence of i.i.d. random variables, $\{e_i\}_{i=1,...,n}$ is an i.i.d. sequence of zero-mean random variables with e_i being mean independent of $\{y_{j0}\}_{j\neq i}$ conditional on y_{i0}. Then for any observation i in cluster g we have $E(c_i|Y_0) = E(c_g|Y_0) + E(e_i|y_{i0})$. To obtain a parsimonious model for the optimal instruments, we can postulate that conditional expectations are linear and that each observation within a cluster contributes in the same way to predict c_g. Then for any observation in cluster g, $E(c_i|Y_0) = \alpha_0 + \beta_0 \frac{1}{n_g} \sum_{j\in g} y_{j0} + \gamma_0 y_{i0}$ where n_g denotes the number of observations in cluster g.

Therefore the optimal instrument for (1.3.2) for an observation in cluster g is $z_i^\star = (\rho_0 - 1)y_{i0} + \alpha_0 + \beta_0 \frac{1}{n_g} \sum_{j\in g} y_{j0} + \gamma_0 y_{i0}$. A feasible version of this optimal instrument can be obtained from a consistent preliminary estimator of ρ_0, denote it $\ddot{\rho}$, since consistent estimators for α_0, β_0, γ_0 can be obtained from a pooled regression of $y_{it} - \ddot{\rho} y_{it-1}$ on an intercept, $\frac{1}{n_g} \sum_{j\in g} y_{j0}$ and y_{i0}. Using the information contained in past outcomes for other observations in the cluster will presumably yield a much better predictor of c_i and hence a much better instrument, which can lead to sizable gains in efficiency. Even though we derived this efficient estimator by using very strong auxiliary assumption, it is consistent as long as (1.2.1) and (1.2.2) hold and one can use inference that is robust to all of our auxiliary assumptions being violated as shown in the next sub-section.

1.3.2 General Case

In this sub-section, we consider efficient estimation with T being any fixed integer equal or greater than two and disturbances being potentially correlated within clusters. Here we will generalize the idea developed in the previous sub-section of using other observations from a cluster to predict one's level of unobserved heterogeneity. We will start with the same auxiliary assumption of clustering as in the previous subsection:

Auxiliary Assumption 1: Clusters of observations are independent and identically distributed.

With Auxiliary Assumption 1, we can derive the optimal estimator for ρ_0 by generalizing the work on optimal instruments for cross-sectionally independent data in Chamberlain (1992a) to the case of cluster-sampling.

In this section we will index observations by cluster so that for any i, g_i denotes the cluster to which observation i belongs and j_g denotes the j^{th} observation of cluster g so that for any observation i in g, there is j such that $j_g = i$ and $\{\{x_{j_g}\}_{j=1,...,n_g}\}_{g=1,...,G} = \{x_i\}_{i=1,...,n}$ for any sequence of variables $\{x_i\}_{i=1,...,n}$. Consider stacking all observations by cluster and define $m_t^g(\rho) = [m_{1_g,t}(\rho),...,m_{n_{gg},t}(\rho)]'$, $m^g(\rho) = [m_2^{g'}(\rho),...,m_T^{g'}(\rho)]'$, $m_t^g = m_t^g(\rho_0)$ and $m^g = m^g(\rho_0)$. Similarly, define $u_t^g = [u_{1_g,t},...,u_{n_{gg},t}]'$, $u^g = [u_1^{g'},...,u_T^{g'}]'$, $c^g = [c_{1_g},...,c_{n_{gg}}]'$ $y_t^g = [y_{1_g,t},...,y_{n_{gg},t}]'$, $Y_t^g = [y_0^{g'},...,y_t^{g'}]'$, and $\Delta Y_{-1}^g = [\Delta y_1^{g'},...,\Delta y_{T-1}^{g'}]'$.

Appendix A.1.1 shows that the optimal estimator for ρ_0 is defined by:

$$\sum_{g=1}^{G} Z_{opt}^g m^g(\hat{\rho}_{opt}) = 0 \tag{1.3.4}$$

where $Z_{opt}^g = L^{\star g'}(\Phi^g)^{-1/2}$ where $\Phi^g = [Cov(m_t^g, m_s^{g'}|Y_{max\{t,s\}-2}^g)]_{t=2,...,T}^{s=2,...,T}$, $(\Phi^g)^{-1/2}$ is the upper diagonal matrix such that $(\Phi^g)^{-1/2'}(\Phi^g)^{-1/2} = (\Phi^g)^{-1}$, $L^{\star g} = [L_t^{\star g'}]_{t=2,...,T}$ and $L_t^{\star g} = E((\Phi_t^g)^{-1/2}\Delta Y_{-1}^g|Y_{t-2}^g)$ where $(\Phi_t^g)^{-1/2}$ is the $(t-1)^{th}$ $n_g \times n_g(T-1)$ matrix composing $(\Phi^g)^{-1/2}$.

One could estimate these optimal instruments non-parametrically by using series of instruments that include lagged values of the dependent variable for an observation but also lagged values of the dependent variable for neighboring observations. A similar estimator has been studied for the case of cross-sectionally independent data in Donald et al. (2009) for static models and Hahn (1997) for dynamic models. However such an approach would not be practical here since there are too many possible terms to consider as instruments. Also, it would involve using many nuisance parameters which can cause poor small sample properties for the estimator, as is discussed later. Instead, we propose two auxiliary auxiliary assumptions

that will allow us to model optimal instruments and drastically reduce the number of nuisance parameters needed. The resulting estimator will be consistent as long as (1.2.1) and (1.2.2) hold and efficient when these auxiliary assumptions are satisfied. Because the estimator we propose makes use of few nuisance parameters, it will have good small sample properties even when the auxiliary assumptions do not hold, as evidenced in Section 1.5.

The second auxiliary assumption we will use is the assumption of conditional homoscedasticity as well as conditional serial uncorrelation and conditional equi-correlation within clusters:

Auxiliary Assumption 2a: For any $i, j \in g, t, s = 1, ..., T, t \geq s$:

$$\begin{aligned} Cov(u_{it}, u_{js} | c^g, Y_{t-1}^g) &= \sigma_u^2 \; if \; i = j, t = s \\ &= \tau_u \sigma_u^2 \; if \; i \neq j, t = s \\ &= 0 \; otherwise \end{aligned}$$

Under Auxiliary Assumption 2a, Appendix A.1.2 shows that the optimal instrument for m^g, Z_{opt}^g, is now a linear function of $\{E(\Delta y_{t-1}^g | Y_{t-s})\}_{t=2,...,T, s=2,...,t}$. This corresponds to the intuition developed in the previous section where we found that, for the special case $T = 2$, optimal instruments were simply $E(\Delta y_1^g | Y_0)$.

From (1.2.1) and (1.2.2):

$$E(\Delta y_{t-1}^g | Y_{t-s}) = (\rho_0 - 1)\rho_0^{s-1} y_{t-s}^g + \sum_{r=0}^{s-2} \rho_0^r E(c^g | Y_{t-s}) \tag{1.3.5}$$

$$= (\rho_0 - 1)\rho_0^{s-1} y_{t-s}^g + \frac{1 - \rho_0^{s-1}}{1 - \rho_0} E(c^g | Y_{t-s}) \tag{1.3.6}$$

Under Auxiliary Assumption 1:

$$E(\Delta y_{t-1}^g | Y_{t-s}) = (\rho_0 - 1)\rho_0^{s-1} y_{t-s}^g + \frac{1 - \rho_0^{s-1}}{1 - \rho_0} E(c^g | Y_{t-s}^g) \tag{1.3.7}$$

Therefore in order to obtain a model for optimal instruments, one needs to make additional assumptions so that there exists a parametric model for the mean of unobserved heterogeneity conditional on lagged values of the dependent variable. In order to keep the number of nuisance parameters low, it is useful to use the assumption that unobserved heterogeneity follows the simple cluster correlation structure:

$$Corr(c_i, c_j) = \tau_c \; if \; i \neq j, i, j \in g \tag{1.3.8}$$

$$= 0 \; otherwise \tag{1.3.9}$$

Also we use the assumption that disturbances $\{u_t^g\}_{t=1,...,T}$ are independent from unobserved heterogeneity, that both have a joint normal distribution and that the initial values of the dependent variable are in the stationary state associated with (1.2.1), i.e.:

$$y_0^g = \frac{c^g}{1-\rho_0} + \tilde{u}_0^g \tag{1.3.10}$$

where $\tilde{u}_0^g$ is independent of c^g and $\{u_t^g\}_{t=1,...,T}$, follows normal distribution with zero mean, variance equal to $\sigma_u^2/(1-\rho_0^2)$ and has a within cluster correlation of τ_u.[8] Let the variance-covariance matrix of u_t^g for $t = 1, ..., T$ be denoted by Σ_u^g:

$$\Sigma_u^g = \sigma_u^2 \begin{bmatrix} 1 & & & \\ \tau_u & 1 & & \\ ... & & ... & \\ \tau_u & ... & \tau_u & 1 \end{bmatrix} \tag{1.3.11}$$

Let the variance-covariance matrix of c^g be denoted by Σ_c^g:

$$\Sigma_c^g = \sigma_c^2 \begin{bmatrix} 1 & & & \\ \tau_c & 1 & & \\ ... & & ... & \\ \tau_c & ... & \tau_c & 1 \end{bmatrix} \tag{1.3.12}$$

The last auxiliary assumption of our model for optimal instruments is:

Auxiliary Assumption 3a: Suppose that for any cluster $g = 1, ..., G$:

$$\begin{bmatrix} c^g \\ y_0^g \\ u^g \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_c \iota_{n_g} \\ \frac{1}{1-\rho_0}\mu_c \iota_{n_g} \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_c^g & & \\ \frac{1}{1-\rho_0}\Sigma_c^g & \frac{1}{(1-\rho_0)^2}\Sigma_c^g + \frac{1}{1-\rho_0^2}\Sigma_u^g & \\ 0 & 0 & I_T \otimes \Sigma_u^g \end{bmatrix} \right) \tag{1.3.13}$$

[8] The auxiliary assumption of stationary initial conditions can easily be generalized, at the expense of introducing three additional nuisance parameters, by assuming:

$$\begin{aligned} y_0^g &= \alpha + \beta c^g + \tilde{u}_0^g \\ \tilde{u}_0^g | c^g &\sim N(0, \tilde{\Sigma}_0) \\ Var(\tilde{u}_{i0}) &= \tilde{\sigma}_0 \\ Corr(\tilde{u}_{i0}, \tilde{u}_{j0}) &= \tau_u \, if \, i \neq j \, but \, g_i = g_j \end{aligned}$$

where Σ_c^g and Σ_u^g have been defined previously and ι_{n_g} is a column vector of ones of dimension $n_g \times 1$.

Note that $E(c^g|Y_t^g) = E(c^g|y_0^g, c^g + u_1^g, ..., c^g + u_t^g)$. Define V^g as:

$$V^g = \begin{bmatrix} \Sigma_c^g & & \\ \frac{1}{1-\rho_0}\Sigma_c^g & \frac{1}{(1-\rho_0)^2}\Sigma_c^g + \frac{1}{1-\rho_0^2}\Sigma_u^g & \\ 0 & 0 & I_T \otimes \Sigma_u^g \end{bmatrix} \tag{1.3.14}$$

Under Auxiliary Assumption 3a:

$$\begin{bmatrix} c^g \\ y_0^g \\ c^g + u_1^g \\ ... \\ c^g + u_T^g \end{bmatrix} \sim N(\mu^g, A^g V^g A^{g'}) \tag{1.3.15}$$

where $\mu^g = A^g \begin{bmatrix} \mu_c \iota_{n_g} \\ \frac{1}{1-\rho_0}\mu_c \iota_{n_g} \\ 0 \end{bmatrix}$ and A^g is the deterministic matrix of ones and zeros so that $A^g \begin{bmatrix} c^g \\ y_0^g \\ u^g \end{bmatrix} = \begin{bmatrix} c^g \\ y_0^g \\ c^g + u_1^g \\ ... \\ c^g + u_T^g \end{bmatrix}$.

Therefore, using the properties of the multivariate normal distribution, $E(c^g|Y_t)$ can be obtained as a linear function of $y_0^g, c^g + u_1^g, ..., c^g + u_t^g$ with coefficients given by the elements of V^g. The exact form of $E(c^g|Y_t)$ under Auxiliary Assumptions 1, 2a, 3a is given in Appendix A.1.3.

Only five nuisance parameters compose V^g and can be consistently estimated if a consistent preliminary estimator of ρ_0 is available, denote it $\ddot{\rho}$. Let $r_{it}(\rho) = y_{it} - \rho y_{it-1}$. Consistent estimators for the nuisance

parameters in V are:

$$\hat{\sigma}_u^2 = \frac{1}{2}\frac{1}{T-1}\frac{1}{n}\sum_{t=2}^{T}\sum_{i=1}^{n} m_{it}(\ddot{\rho})^2$$

$$\hat{\tau}_u = \frac{1}{\hat{\sigma}_u^2}\frac{1}{2}\frac{1}{T-1}\frac{1}{n}\sum_{t=2}^{T}\sum_{i=1}^{n}\frac{1}{n_{g_i}-1}\sum_{j=1}^{n} 1[i\neq j, g_i = g_j] m_{it}(\ddot{\rho}) m_{jt}(\ddot{\rho})$$

$$\hat{\sigma}_c^2 = \frac{1}{T(T-1)}\frac{1}{n}\sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{i=1}^{n} 1[t\neq s] r_{it}(\ddot{\rho}) r_{is}(\ddot{\rho}) - \hat{\mu}_c^2$$

$$\hat{\mu}_c = \frac{1}{T}\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} r_{it}(\ddot{\rho})$$

$$\hat{\tau}_c = \frac{1}{\hat{\sigma}_c^2}\frac{1}{T(T-1)}\frac{1}{n}\sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{i=1}^{n}\frac{1}{n_{g_i}-1}\sum_{j=1}^{n} 1[t\neq s, g_i = g_j; i\neq j] r_{it}(\ddot{\rho}) r_{js}(\ddot{\rho})$$

Let $\hat{\Phi}^g$ be the consistent estimator for the variance-covariance matrix $\Phi^g = Var(m^g(\rho_0))$ composed of $\hat{\sigma}_u$ and $\hat{\tau}_u$ from the formula derived in Appendix A.1.2. Let $\hat{\Phi}^{g-1/2}$ be the upper-diagonal matrix such that $\hat{\Phi}^{g-1/2\prime}\hat{\Phi}^{g-1/2} = \hat{\Phi}^{g-1}$. Denote $\hat{\Phi}_t^{g-1/2}$ the t^{th} $n_g \times n_g(T-1)$ matrix composing $\hat{\Phi}^{g-1/2}$. Let $\hat{\mu}_t^{gc}$ be a consistent estimator of $E(c^g|Y_t)$ from the formula given in the Appendix A.1.3.

A consistent estimator for the optimal instrument for $m^g(\rho)$ under (1.2.1) and (1.2.2) and Assumptions 1, 2a, 3a is:

$$\hat{Z}_{opt}^g = [\hat{\Phi}_1^{g-1/2}\begin{bmatrix}(\ddot{\rho}-1)y_0^g + \hat{\mu}_0^{gc}\\ \dots \\ \ddot{\rho}^{T-2}(\ddot{\rho}-1)y_0^g + \frac{1-\ddot{\rho}^{T-1}}{1-\ddot{\rho}}\hat{\mu}_0^{gc}\end{bmatrix}, \dots, \hat{\Phi}_{T-1}^{g-1/2}\begin{bmatrix}0\\ \dots \\ (\ddot{\rho}-1)y_{T-2}^g + \hat{\mu}_{T-2}^{gc}\end{bmatrix}]\hat{\Phi}^{g-1/2} \tag{1.3.16}$$

and the estimator obtained from using this instrument matrix is defined by:

$$\sum_{g=1}^{G}\hat{Z}_{opt}^g m^g(\hat{\rho}^\star) = 0 \tag{1.3.17}$$

So that:

$$\hat{\rho}^\star = \frac{\sum_{g=1}^{G}\hat{Z}_{opt}^g \Delta y^g}{\sum_{g=1}^{G}\hat{Z}_{opt}^g \Delta y_{-1}^g} \tag{1.3.18}$$

$$= \rho_0 + \frac{\sum_{g=1}^{G}\hat{Z}_{opt}^g \Delta u^g}{\sum_{g=1}^{G}\hat{Z}_{opt}^g \Delta y_{-1}^g} \tag{1.3.19}$$

where $\Delta y^g = [\Delta y_2^{g\prime}, \dots, \Delta y_T^{g\prime}]'$, $\Delta y_{-1}^g = [\Delta y_1^{g\prime}, \dots, \Delta y_{T-1}^{g\prime}]'$ and $\Delta u^g = [\Delta u_2^{g\prime}, \dots, \Delta u_T^{g\prime}]'$.

Let $\ddot{Z}^g_{opt}$ to be the random vector defined as in (1.3.16) but where $\ddot{\rho}$, $\hat{\sigma}^2_u$, $\hat{\sigma}^2_c$, $\hat{\tau}_u$, $\hat{\tau}_c$, $\hat{\mu}_c$ are replaced by $plim(\ddot{\rho})$, $plim(\hat{\sigma}^2_u)$, $plim(\hat{\sigma}^2_c)$, $plim(\hat{\tau}_u)$, $plim(\hat{\tau}_c)$, $plim(\hat{\mu}_c)$. When (1.2.1), (1.2.2) and Auxiliary Assumption 1 hold, $\hat{\rho}^\star$ is asymptotically normal:

$$\sqrt{G}(\hat{\rho}^\star - \rho_0) \xrightarrow{d} N(0, V_\rho) \tag{1.3.20}$$

$$V_\rho = E(\ddot{Z}^g_{opt}\Delta y^g_{-1})^{-2} Var(\ddot{Z}^g_{opt}\Delta u^g) \tag{1.3.21}$$

Standard errors for $\hat{\rho}^\star$ that are consistent as long as (1.2.1), (1.2.2) and Auxiliary Assumption 1 hold are given by:

$$s.e. = ((\sum_{g=1}^{G} \hat{Z}^g_{opt}\Delta y^g_{-1})^{-2} \sum_{g=1}^{G} (\hat{Z}^g_{opt} m^g(\hat{\rho}^\star))^2)^{1/2} \tag{1.3.22}$$

The estimator defined by (1.3.17) is consistent and asymptotically normal even when the Auxiliary Assumption 1 of cluster sampling is not satisfied, as long as some regularity conditions hold on the strength of cross-sectional dependence. As in section 1.2.2, cross-sectional dependence has to be weak enough so that asymptotic theorems can be applied:

$$\frac{1}{G}\sum_{g=1}^{G} \ddot{Z}^g_{opt}\Delta u^g \xrightarrow{p} 0 \tag{1.3.23}$$

$$\frac{1}{G}\sum_{g=1}^{G} \ddot{Z}^g_{opt}\Delta y^g_{-1} \xrightarrow{p} a \tag{1.3.24}$$

$$\frac{1}{\sqrt{G}}\sum_{g=1}^{G} \ddot{Z}^g_{opt}\Delta u^g \xrightarrow{d} N(0, v) \tag{1.3.25}$$

where $a = plim(\frac{1}{G}\sum_{g=1}^{G} \ddot{Z}^g_{opt}\Delta y^g_{-1}) \neq 0$ and $v = plim(\frac{1}{G}(\sum_{g=1}^{G} \ddot{Z}^g_{opt}\Delta u^g)^2)$. In this case:

$$\sqrt{G}(\hat{\rho}^\star - \rho_0) \xrightarrow{d} N(0, a^{-2}v) \tag{1.3.26}$$

a can simply be estimated by $\frac{1}{G}\sum_{g=1}^{G} \hat{Z}^g_{opt}\Delta y^g_{-1}$ and non-parametric estimators for $plim(\frac{1}{G}(\sum_{g=1}^{G} Z^g_{opt}\Delta u^g)^2)$ as well as statistical tests with general forms of spatial dependence are available and have been discussed in Conley (1999), Bester et al. (2011b), Kim and Sun (2011) and Bester et al. (2011a).

In situations where available preliminary estimators might have poor small sample properties, one can also use an iterated version of the feasible optimal estimator. Denote $\hat{Z}^g_{opt}(\rho)$ to be the value of the estimated optimal instruments for a preliminary estimator (previously denoted $\ddot{\rho}$) evaluated at ρ. The iterated optimal estimator is defined by:

$$\sum_{g=1}^{G} \hat{Z}^g_{opt}(\hat{\rho}_{iter}) m^g(\hat{\rho}_{iter}) = 0 \tag{1.3.27}$$

This estimator has the same $\sqrt{n}$-asymptotic properties as the two step estimator defined by (1.3.17) but its small sample properties will not depend on the small sample properties of a preliminary estimator.

1.3.3 Comparison to Existing Estimators

The estimator defined by (1.3.17) can be rewritten as $\hat{\rho}^*$ that satisfies the equation:

$$\sum_{g=1}^{G} w^{g*}(\hat{\eta}) Z^g m^g(\hat{\rho}^*) = 0 \tag{1.3.28}$$

where $\hat{\eta} = [\hat{\sigma}_u^2, \hat{\tau}_u, \hat{\sigma}_c^2, \hat{\mu}_c, \hat{\tau}_c]$, Z^g is the matrix containing all valid instruments for m^g:

$$Z^g = \begin{bmatrix} I_{n_g} \otimes Y_0^g & & & \\ 0 & I_g \otimes Y_1^g & & \\ \dots & & \dots & \\ 0 & \dots & 0 & I_{n_g} \otimes Y_{T-2}^g \end{bmatrix} \tag{1.3.29}$$

and $w^{g*}()$ is the row vector function such that $w^{g*}(\hat{\eta}) Z^g = \hat{Z}_{opt}^g$.

The Arellano and Bond estimator can also be written as exactly identified from:

$$\sum_{g=1}^{G} \hat{w}_{AB}^g Z^g m^g(\hat{\rho}_{AB}) = 0 \tag{1.3.30}$$

where:

$$\hat{w}_{AB}^g = \sum_{i=1}^{n} (\Delta Y_{-1,i}^{'} Z_i^{'})(\sum_{i=1}^{n} Z_i m_i(\tilde{\rho}) m_i(\tilde{\rho})^{'} Z_i^{'})^{-1} S^g \tag{1.3.31}$$

where $\tilde{\rho}$ is a preliminary consistent estimator and S^g is the matrix of zeros and ones such that $S^g Z^g m^g(\rho) = \sum_{i \in g} Z_i m_i(\rho)$ where:

$$Z_i = \begin{bmatrix} Y_{i0} & & & \\ 0 & Y_{i1} & & \\ \dots & & \dots & \\ 0 & \dots & 0 & Y_{iT-2} \end{bmatrix} \tag{1.3.32}$$

In the presence of cross-sectional dependence, it is likely that our estimator will perform better than the Arellano and Bond estimator even when some of the Auxiliary Assumptions 1, 2a, 3a are violated because our estimator gives non-zero weights to moment conditions obtained from using instruments from neighboring observations. As discussed in previous sections, these instruments may have significant predictive

power for the covariates in the differenced equations so that these additional moment conditions might be useful to improve the accuracy of the estimator.

In addition, our estimator relies on the estimation of only five nuisance parameters to compute weights for all $n_g^2 \times T \times (T-1)/2$ moment conditions available per cluster, whereas the Arellano and Bond estimator relies on the estimation of $T \times (T-1)/2$ weights. When T is relatively large, estimating that many nuisance parameters causes the Arellano and Bond estimator to suffer from poor small sample properties in terms of bias, precision and inference, which was studied in the context of cross-sectional independence in Alvarez and Arellano (2003) and Windmeijer (2005). Because our estimator makes use of few nuisance parameters, it will have good properties in finite samples even when T is relatively large. A formal derivation of the asymptotic properties of our estimator when both n and T grow unboundedly is left for future research.

As a result of both using non-zero weights for useful moment conditions and using nuisance parameters parsimoniously, the results from Monte Carlo simulations presented in Section 1.5 show that our estimator has significantly better small sample properties than the Arellano and Bond estimator in terms of efficiency and quality of inference, particularly in cases with cross-sectional dependence but also without cross-sectional dependence.

So-called system GMM estimators presented in Ahn and Schmidt (1995), Arellano and Bover (1995), and Blundell and Bond (1998) are similar to the Arellano and Bond estimator but use additional moment conditions based on additional assumptions of homoscedasticity, no serial correlation, or stationary initial conditions. Since our estimator is only based on the mean independence of transitory shocks conditional on past outcomes, it is more robust than the estimators presented in Ahn and Schmidt (1995), Arellano and Bover (1995) or Blundell and Bond (1998).

1.4 Models with Covariates

Similar auxiliary assumptions as in the previous section can be listed to model optimal instruments for models with covariates. In this section we consider a model that allows for some of the covariates to be strictly exogenous (w_{it}) and some of the covariates to be sequentially exogenous or contemporaneously

endogenous (x_{it}):

$$y_{it} = x_{it}\beta_0 + w_{it}\gamma_0 + c_i + u_{it}\ t = 1,...,T \tag{1.4.1}$$

$$E(u_{it}|Z_t, W) = 0 \tag{1.4.2}$$

where $W = [W_1,...,W_n]$ and $W_i = w_{i1},...,w_{iT}$ and $Z_t = [z_{i1},...,z_{it}]$ and for every random variable $x_{it}^{(j)}$ in x_{it}, either $x_{it}^{(j)}$ or $x_{it-1}^{(j)}$ is in z_{it}.[9] $x_{it}^{(j)}$ is said to be sequentially exogenous if it is in z_{it}. If $x_{it-1}^{(j)}$ only is in z_{it}, x_{it}^{j} is said to be contemporaneously endogenous. Such a model specification is flexible enough to allow for complex interactions between unobserved factors and covariates of interest, an example will be given in Section 1.6. The estimation method presented in this section can be generalized to the case where neither x_{it} nor x_{it-1} are part of z_{it} but where some other instruments are available, which is also treated in the example given in Section 1.6. As a notational matter, we generalize the notation from the previous section by denoting by x^g the vector $[x_{1_g},...,x_{n_{gg}}]'$ for any sequence of variables $\{x_i\}_{i=1,...,n}$.

A consistent estimator of β_0, γ_0 is obtained from the differenced equation:

$$\Delta y_{it} = \Delta x_{it}\beta_0 + \Delta w_{it}\gamma_0 + \Delta u_{it}\ t = 1,...,T \tag{1.4.3}$$

$$E(\Delta u_{it}|Z_{t-1}, W) = 0 \tag{1.4.4}$$

To model optimal instruments for estimating β_0 and γ_0 from (1.4.3) and (1.4.4), we will make use of the same auxiliary assumption of clustering, i.e. we maintain the use of Auxiliary Assumption 1a. We also generalize Auxiliary Assumption 2a so that homoscedasticity and serial correlation are specified conditional on the relevant instruments:

Auxiliary Assumption 2b: For any $i, j \in g, t, s = 1,...,T, t \geq s$:

$$\begin{aligned} Cov(u_{it}, u_{js}|c^g, Z_t^g, W^g) &= \sigma_u^2\ if\ i = j, t = s \\ &= \tau_u\sigma_u^2\ if\ i \neq j, t = s \\ &= 0\ if\ t > s \end{aligned}$$

As in the previous section, this assumption guarantees that the optimal instruments will be known linear functions of $\{E(\Delta x_{it}|Z_s^g, W^g)\}_{t,s=1,...,T,s\leq t-1}$ and W_i (up to the unknown nuisance parameters σ_u^2

[9] Note that a special case of this model is the dynamic model we considered in the previous section where $x_{it} = y_{it-1}$, $x_{it} = z_{it}$, and $w_{it}\gamma_0 = 0$. In most applications, even if x_{it} includes other covariates than lagged values of the dependent variable, it is expected that y_{it-1} will be included in x_{it} in order to identify the effect of x_{it} on y_{it} separately from the dynamic effects in y_{it} and x_{it}.

and τ_u). Therefore, we need to generalize Auxiliary Assumption 3a to obtain a parsimonious model for $\{E(\Delta x_{it}|Z_s^g, W^g)\}_{t,s=1,...,T,s\leq t-1}$. To do so, we can model z_t^g as a VAR process conditional on W^g:

Auxiliary Assumption 3b: Suppose that for any observation $i = 1, ..., n$:

$$z_{it} = \Gamma z_{it-1} + w_{it}\eta + d_i + v_{it} \tag{1.4.5}$$

and:

$$\begin{bmatrix} d^g \\ z_0^g \\ v^g \end{bmatrix} |W^g \sim N(\begin{bmatrix} \mu_d(W^g) \\ \mu_{z_0}(W^g) \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_d^g & & \\ \Sigma_{dz_0} & \Sigma_{z_0} & \\ 0 & 0 & \Sigma_v \end{bmatrix} \tag{1.4.6}$$

where $v^g = [v_1^{g\prime}, ..., v_T^{g\prime}]'$.

In particular applications, one will impose auxiliary restrictions on $\mu_d(.)$, $\mu_{z_0}(.)$, Σ_d^g, Σ_{dz_0}, Σ_{z_0}, Σ_v so that they can be estimated with few enough nuisance parameters.

Auxiliary Assumption 3b implies:

$$E(z_{it}|Z_{t-s}^g, W^g) = \Gamma^s z_{it-s} + \sum_{r=0}^{s-1} \Gamma^r (w_{it-r}\eta + E(d_i|Z_{t-s}^g, W^g)) \tag{1.4.7}$$

and:

$$E(d_i|Z_t^g, W^g) = E(d_i|z_0^g, W^g, d^g + v_1^g, ..., d^g + v_t^g) \tag{1.4.8}$$

which can be derived from Auxiliary Assumption 3b as was done in the previous section. For any covariate $x_{it}^{(j)}$, either $x_{it}^{(j)}$ is in z_{it} or $x_{it-1}^{(j)}$ is in z_{it}, therefore Auxiliary Assumption 3b yields a model for $E(x_{it}^{(j)}|Z_s^g, W^g)\ \forall s \leq t$ and hence $E(\Delta x_{it}|Z_s^g, W^g)\ \forall s \leq t-1$ as a function of the nuisance parameters in Auxiliary Assumption 3b.

Therefore, under (1.4.1), (1.4.2) and Auxiliary Assumptions 1a, 2b and 3b, one can find a parametric model for the optimal instruments for estimating β_0 and γ_0. A feasible version of these instruments can be obtained from a preliminary estimator of (β_0, η_0) as in the previous section. One can also use an iterated version of this feasible estimator in order to obtain an estimator with better performances in small samples.

1.5 Monte Carlo Simulations

In this section, we will study the small sample properties of the estimator we propose using Monte Carlo simulations. Consider the simple data generating process for a model with cluster correlation and without covariates:

$$\begin{aligned} n_g &\sim Poisson(\alpha)+1 \\ c^g &\sim F_c \\ y_0^g|c^g &\sim Normal(\mu_0(c^g),\Sigma_0(c^g)) \\ y_t^g|c^g,y_{t-1}^g,...,y_0^g &\sim Normal(c^g+\rho y_{t-1}^g,\Sigma_u(c^g,y_{t-1}^g,...,y_0^g)) \end{aligned}$$

We compare the properties of three estimators of ρ: The estimator defined in Arellano and Bond (1991) which we call the AB estimator, the estimator defined by (1.3.27), which we denote by Estimator 1 and the estimator defined by (1.3.27) but with estimated within-cluster correlations replaced by zero which we denote by Estimator 2.[10] As a benchmark for comparison, we also show the results from using an unfeasible optimal estimator (UO) which is optimal in the class of estimators that use linear functions of the instruments. This estimator weights optimally all available moment conditions that use linear instruments using the true unobserved optimal weights so that it is defined by:

$$\sum_{g=1}^{G} w^g Z^g m^g(\hat{\rho}_{UO}) = 0 \tag{1.5.1}$$

$$w^g = \Delta^{g'}(W^g)^{-1} \tag{1.5.2}$$

$$\Delta^g = E(Z^g \frac{\partial m^g}{\partial \rho}) \tag{1.5.3}$$

$$W^g = E(Z^g m^g m^{g'} Z^{g'}) \tag{1.5.4}$$

When Auxiliary Assumptions 1, 2a, 3a hold, the UO estimator is the same as the estimator defined by (1.3.4) and will be efficient in the class of estimators using any function of the instruments. When these assumptions hold, Estimator 1 and the unfeasible optimal estimator will also be asymptotically equivalent so that, for small samples, the difference in their performances is due to the extra noise in Estimator 1

[10] In most of the scenarios we simulate, transitory shocks will be homoscedastic, serially uncorrelated and the dependent variable will be stationary so that additional moment conditions presented in Arellano and Bover 1995, Ahn and Schmidt 1995 or Blundell and Bond 1998 hold. We do not present estimators that use these moment conditions however since we are interested in studying the properties of estimators that are robust to these moment conditions being false.

due to estimating the nuisance parameters needed. When Auxiliary Assumptions 2a or 3a are violated, the unfeasible optimal estimator is asymptotically more efficient than Estimator 1.

Estimator 1 is asymptotically more efficient than the AB estimator or than Estimator 2 when there exists cross-sectional dependence and Auxiliary Assumptions 1, 2a, 3a hold. When Auxiliary Assumptions 1, 2a, 3a hold and there is no cross-sectional dependence, the AB Estimator, Estimator 1 and Estimator 2 have the same asymptotic variance. When Auxiliary Assumptions 2a or 3a are violated and there is no cross-sectional dependence, the AB estimator has a smaller asymptotic variance than Estimator 1 and 2 but, in finite samples, Estimator 1 or Estimator 2 might still have better properties than the AB estimator because they make use of less nuisance parameters. When Auxiliary Assumptions 2a or 3a are violated and there is cross-sectional dependence, which of the AB estimator or Estimator 1 has smallest asymptotic variance depends on the data generating process but we expect Estimator 1 to perform better since, by making use of instruments from other observations in the cluster, it should use a weighted sum of moment conditions that is closer to optimal that the sum used for the AB estimator.

For inference for the AB estimator, we will consider GMM robust standard errors with clustered standard errors with and without the finite sample correction proposed by Windmeijer (2005). For inference for Estimators 1 and 2, we use the standard errors defined in (1.3.22) that only require (1.2.1), (1.2.2) and Auxiliary Assumption 1 to hold in order to be consistent.

We will study the small sample properties of the estimators in three different scenarios: within cluster equi-correlation, cross-sectional independence and general within cluster correlation with unobserved heterogeneity that does not have a Normal distribution. Scenario 1 and 2 will correspond to Auxiliary Assumptions 1, 2a, 3a holding. In Scenario 1 there is cross-sectional dependence and in Scenario 2 there is no cross-sectional dependence. Scenario 3 corresponds to only Auxiliary Assumption 1 holding.

More precisely, Scenario 1 uses the following parameterization:

$$F_c = Normal(0, \begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ ... & & ... & \\ 0.5 & ... & 0.5 & 1 \end{bmatrix})$$

$$\Sigma_u(c^g, y^g_{t-1}, ..., y^g_0) = \begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ ... & & ... & \\ 0.5 & ... & 0.5 & 1 \end{bmatrix}$$

$$\mu_0(c^g) = \frac{c^g}{1-\rho_0}$$

$$\Sigma_0(c^g) = \frac{1}{1-\rho_0^2}\Sigma_u(c^g, y^g_{t-1}, ..., y^g_0)$$

Scenario 2 uses:

$$F_c = Normal(0, \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ ... & & ... & \\ 0 & ... & 0 & 1 \end{bmatrix})$$

$$\Sigma_u(c^g, y^g_{t-1}, ..., y^g_0) = \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ ... & & ... & \\ 0 & ... & 0 & 1 \end{bmatrix}$$

$$\mu_0(c^g) = \frac{c^g}{1-\rho_0}$$

$$\Sigma_0(c^g) = \frac{1}{1-\rho_0^2}\Sigma_u(c^g, y^g_{t-1}, ..., y^g_0)$$

And Scenario 3 uses:

$$F_c = LogNormal(0, \begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ ... & & ... & \\ 0.5 & ... & 0.5 & 1 \end{bmatrix})$$

$$\Sigma_u(c^g, y^g_{t-1}, ..., y^g_0) = \begin{bmatrix} u^2_{i_1 t-1} & & & \\ 0.5u_{i_1 t-1}u_{i_2 t-1} & u^2_{i_2 t-1} & & \\ ... & & ... & \\ 0.5u_{i_1 t-1}u_{i_{n_g} t-1} & ... & 0.5u_{i_{n_g-1} t-1}u_{i_{n_g} t-1} & u^2_{i_{n_g} t-1} \end{bmatrix}$$

$$\mu_0(c^g) = \frac{c^g}{1-\rho_0}$$

$$\Sigma_0(c^g) = \frac{1}{1-\rho_0^2} \begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ ... & & ... & \\ 0.5 & ... & 0.5 & 1 \end{bmatrix}$$

All Monte Carlo results were obtained using 1,000 replications. Because Estimators 1 and 2 are iterated versions of our estimator, we present results from simulations conditional on Estimators 1 and 2 converging. Table 1.1 shows the number of observations where all estimators converged, which represents all or almost all draws except when $T = 5$, $G = 100$ and $\rho = 0.8$. In this case Estimator 1 or Estimator 2 did not converge in 15%-22% of the replications depending on the scenario. In particular applications, convergence of the iterated Estimators 1 and 2 will depend on the particular numerical algorithm chosen and properties of the data. For instance in the application presented in Section 1.6, convergence was achieved in just a few iterations even though $T = 3$.

Table 1.2, 1.3 and 1.4 show the results for the four estimators considered in terms of bias, standard deviation and root mean squared error for a value of ρ of 0.8. Table 1.2 shows results for the case where there is equi-correlation within clusters (Scenario 1), Table 1.3 the case where there is no cross-sectional correlation (Scenario 2) and Table 1.4 the case where there is heteroscedasticity and cross-sectional correlation (Scenario 3). The first conclusion from these three tables is that Estimator 1 and 2 exhibit virtually no bias compared to the AB estimator. Estimator 1 also has significantly smaller standard deviations when there is cross-sectional correlation (Scenarios 1 and 3). Both of these features of our estimator result in significantly

smaller values for mean squared error. The smaller standard deviations of our estimator are due to the use of instruments from other observations in the cluster that are relevant in the presence of cross-sectional dependence. The low bias is attributable to our estimators using very few nuisance parameters compared to the AB estimator. The improvement of Estimator 1 over the AB estimator is particularly striking when T is large and G is small, which is when the AB estimator uses the most nuisance parameters compared to the sample size. When there is no within cluster correlation (Scenario 2), Estimators 1 and 2 have standard deviations only slightly lower than the AB estimator so that the decrease in rmse of Estimators 1 and 2 compared to the AB estimator is mostly due to the elimination of the bias. In Scenario 3 where the unfeasible optimal estimator is asymptotically more efficient than Estimator 1, Estimator 1 performs very closely to the unfeasible optimal estimator, which shows that the approximation of the optimal weighted sum of moment conditions used by Estimator 1 is good in this case.

Table 1.5, Table 1.6 and Table 1.7 show results in terms of bias in standard errors (captured by the ratio of the mean of the standard errors over the standard deviations of the estimators), coverage of the 95% confidence interval and average length of 95% confidence intervals. All three tables show that standard errors for the AB estimator without the Windmeijer correction are seriously downward biased, particularly when T is large, resulting is very low coverage of 95% confidence intervals (as low as 48%). The Windmeijer correction yields unbiased standard errors for the AB estimator but the resulting confidence intervals still have low coverage because of the bias in the AB estimator of ρ. The standard errors for Estimators 1 and 2 are unbiased and the resulting confidence intervals have the correct coverage of 95%. Because our estimators have smaller standard deviations that the AB estimator, the average length of their 95% confidence intervals is also smaller than that of the AB estimator so that our estimators have confidence intervals that are both tighter and have the correct coverage.

Tables 1.8-1.13 show the same results for $\rho_0 = 0.5$. Estimators 1 and 2 show similar improvements over the AB estimator but slightly less markedly since, with this lower level of persistence, the instruments used by the AB estimator are not as weak as when $\rho_0 = 0.8$ so that there is less to gain compared to the unfeasible optimal estimator.

Table 1.1: Number of replications where all estimators converged (out of 1,000)

	$\rho = 0.8$			$\rho = 0.5$		
	Scenario 1	Scenario 2	Scenario 3	Scenario 1	Scenario 2	Scenario 3
T=5						
G=100	802	854	781	1000	999	1000
G=200	906	935	867	1000	1000	1000
G=400	977	976	942	1000	1000	1000
T=10						
G=100	998	992	989	1000	999	1000
G=200	1000	999	1000	1000	1000	1000
G=400	1000	1000	1000	1000	1000	1000
T=15						
G=100	1000	999	1000	1000	999	1000
G=200	1000	1000	1000	1000	1000	1000
G=400	995	1000	1000	1000	1000	1000

Table 1.2: Bias and RMSE, $\rho = .8$, equi-correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	Estimator 1	Estimator 2
T=5					
G=100	bias	−0.031	−0.158	−0.037	−0.045
	sd	0.121	0.176	0.137	0.191
	rmse	0.125	0.236	0.142	0.196
G=200	bias	−0.018	−0.087	−0.018	−0.025
	sd	0.089	0.127	0.093	0.134
	rmse	0.091	0.154	0.095	0.136
G=400	bias	−0.001	−0.033	−0.002	0.000
	sd	0.064	0.092	0.065	0.097
	rmse	0.064	0.098	0.065	0.097
T=10					
G=100	bias	0.001	−0.060	0.001	0.001
	sd	0.047	0.067	0.048	0.068
	rmse	0.047	0.090	0.048	0.068
G=200	bias	−0.001	−0.033	−0.001	−0.003
	sd	0.034	0.048	0.034	0.047
	rmse	0.034	0.058	0.034	0.047
G=400	bias	0.001	−0.016	0.001	0.000
	sd	0.024	0.035	0.024	0.034
	rmse	0.024	0.038	0.024	0.034
T=15					
G=100	bias	−0.001	−0.041	−0.000	−0.003
	sd	0.028	0.038	0.028	0.037
	rmse	0.028	0.056	0.028	0.037
G=200	bias	0.000	−0.022	0.000	−0.001
	sd	0.018	0.027	0.018	0.026
	rmse	0.018	0.034	0.018	0.026
G=400	bias	0.000	−0.010	0.000	0.000
	sd	0.013	0.019	0.013	0.018
	rmse	0.013	0.021	0.013	0.018

Table 1.3: Bias and RMSE, $\rho = .8$, no correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	Estimator 1	Estimator 2
T=5					
G=100	bias	−0.028	−0.097	−0.030	−0.034
	sd	0.121	0.119	0.135	0.127
	rmse	0.124	0.153	0.138	0.131
G=200	bias	−0.013	−0.046	−0.013	−0.014
	sd	0.093	0.091	0.097	0.094
	rmse	0.093	0.102	0.098	0.095
G=400	bias	−0.001	−0.020	−0.001	−0.002
	sd	0.063	0.063	0.066	0.065
	rmse	0.063	0.066	0.066	0.065
T=10					
G=100	bias	0.000	−0.033	0.000	−0.000
	sd	0.046	0.050	0.047	0.047
	rmse	0.046	0.060	0.047	0.047
G=200	bias	−0.001	−0.018	−0.001	−0.002
	sd	0.033	0.035	0.033	0.033
	rmse	0.033	0.039	0.033	0.033
G=400	bias	0.001	−0.008	0.001	0.001
	sd	0.024	0.025	0.024	0.024
	rmse	0.024	0.026	0.024	0.024
T=15					
G=100	bias	−0.001	−0.023	−0.001	−0.001
	sd	0.028	0.031	0.028	0.028
	rmse	0.028	0.038	0.028	0.028
G=200	bias	0.000	−0.011	0.000	0.000
	sd	0.018	0.020	0.018	0.018
	rmse	0.018	0.023	0.018	0.018
G=400	bias	0.000	−0.005	0.000	0.000
	sd	0.013	0.013	0.013	0.013
	rmse	0.013	0.014	0.013	0.013

Table 1.4: Bias and RMSE, $\rho = .8$, heteroscedasticity and correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	Estimator 1	Estimator 2
T=5					
G=100	bias	−0.027	−0.226	−0.065	−0.049
	sd	0.183	0.218	0.271	0.367
	rmse	0.185	0.314	0.279	0.370
G=200	bias	−0.025	−0.140	−0.027	−0.033
	sd	0.125	0.166	0.135	0.203
	rmse	0.128	0.217	0.137	0.206
G=400	bias	−0.007	−0.069	−0.008	−0.010
	sd	0.084	0.121	0.085	0.131
	rmse	0.085	0.139	0.086	0.131
T=10					
G=100	bias	0.001	−0.075	−0.000	0.000
	sd	0.055	0.075	0.056	0.074
	rmse	0.055	0.106	0.056	0.074
G=200	bias	0.000	−0.042	0.000	−0.000
	sd	0.039	0.055	0.039	0.054
	rmse	0.039	0.069	0.039	0.054
G=400	bias	0.001	−0.019	0.001	0.001
	sd	0.027	0.039	0.027	0.038
	rmse	0.027	0.043	0.027	0.038
T=15					
G=100	bias	−0.001	−0.046	−0.000	−0.002
	sd	0.031	0.041	0.030	0.040
	rmse	0.031	0.062	0.030	0.040
G=200	bias	0.001	−0.024	0.001	0.000
	sd	0.020	0.029	0.020	0.028
	rmse	0.020	0.037	0.020	0.028
G=400	bias	0.001	−0.012	0.000	0.000
	sd	0.015	0.021	0.014	0.020
	rmse	0.015	0.024	0.014	0.020

Table 1.5: Inference, ρ = .8, equi-correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	AB w/ Windmeijer correction	Estimator 1	Estimator 2
T=5						
G=100	ratio	1.121	0.788	1.051	1.016	1.002
	coverage	0.969	0.763	0.895	0.964	0.959
	length	0.539	0.550	0.749	0.552	0.760
G=200	ratio	1.063	0.834	1.045	1.024	0.978
	coverage	0.956	0.796	0.917	0.951	0.956
	length	0.375	0.417	0.525	0.376	0.518
G=400	ratio	1.073	0.863	1.029	1.042	0.977
	coverage	0.976	0.888	0.939	0.969	0.954
	length	0.268	0.313	0.376	0.268	0.372
T=10						
G=100	ratio	0.987	0.661	1.004	0.962	0.943
	coverage	0.948	0.653	0.873	0.952	0.949
	length	0.184	0.176	0.270	0.184	0.254
G=200	ratio	0.983	0.734	1.003	0.975	0.961
	coverage	0.949	0.760	0.905	0.949	0.944
	length	0.130	0.139	0.194	0.130	0.179
G=400	ratio	0.974	0.766	0.975	0.968	0.943
	coverage	0.953	0.835	0.922	0.952	0.939
	length	0.092	0.104	0.134	0.091	0.127
T=15						
G=100	ratio	0.941	0.561	1.015	0.939	0.983
	coverage	0.941	0.480	0.815	0.943	0.948
	length	0.105	0.084	0.140	0.105	0.144
G=200	ratio	1.025	0.688	1.053	1.022	1.006
	coverage	0.961	0.696	0.916	0.958	0.951
	length	0.074	0.072	0.110	0.074	0.102
G=400	ratio	1.012	0.763	1.024	1.013	1.006
	coverage	0.952	0.823	0.925	0.957	0.944
	length	0.052	0.057	0.078	0.052	0.072

Table 1.6: Inference, $\rho = .8$, no correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	AB w/ Windmeijer correction	Estimator 1	Estimator 2
T=5						
G=100	ratio	1.140	1.042	1.142	1.013	1.060
	coverage	0.978	0.883	0.923	0.959	0.960
	length	0.548	0.491	0.558	0.542	0.533
G=200	ratio	1.038	1.000	1.053	0.984	1.004
	coverage	0.961	0.909	0.932	0.954	0.948
	length	0.379	0.359	0.383	0.377	0.372
G=400	ratio	1.082	1.052	1.073	1.040	1.041
	coverage	0.975	0.947	0.951	0.968	0.965
	length	0.269	0.262	0.271	0.268	0.267
T=10						
G=100	ratio	1.006	0.820	1.007	0.974	0.975
	coverage	0.958	0.836	0.910	0.951	0.952
	length	0.185	0.163	0.204	0.182	0.182
G=200	ratio	0.990	0.890	0.971	0.981	0.979
	coverage	0.950	0.884	0.918	0.951	0.949
	length	0.130	0.122	0.142	0.129	0.129
G=400	ratio	0.976	0.920	0.975	0.964	0.965
	coverage	0.952	0.914	0.932	0.950	0.950
	length	0.092	0.089	0.097	0.091	0.091
T=15						
G=100	ratio	0.949	0.686	0.959	0.935	0.937
	coverage	0.948	0.730	0.866	0.943	0.944
	length	0.106	0.084	0.105	0.105	0.105
G=200	ratio	1.027	0.846	1.021	1.022	1.020
	coverage	0.962	0.860	0.925	0.957	0.958
	length	0.074	0.066	0.082	0.074	0.074
G=400	ratio	1.011	0.935	1.033	1.008	1.009
	coverage	0.952	0.902	0.945	0.955	0.956
	length	0.052	0.050	0.058	0.052	0.052

Table 1.7: Inference, $\rho = .8$, heteroscedasticity and correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	AB w/ Windmeijer correction	Estimator 1	Estimator 2
T=5						
G=100	ratio	1.027	0.773	1.088	0.834	0.812
	coverage	0.963	0.666	0.872	0.949	0.940
	length	0.747	0.669	0.943	0.897	1.181
G=200	ratio	1.028	0.801	1.033	0.943	0.854
	coverage	0.945	0.736	0.884	0.955	0.934
	length	0.507	0.523	0.680	0.501	0.684
G=400	ratio	1.078	0.845	1.027	1.036	0.919
	coverage	0.969	0.831	0.901	0.960	0.947
	length	0.358	0.402	0.491	0.348	0.474
T=10						
G=100	ratio	0.973	0.652	1.009	0.937	0.954
	coverage	0.948	0.608	0.857	0.942	0.954
	length	0.214	0.193	0.301	0.208	0.280
G=200	ratio	0.968	0.711	0.982	0.951	0.936
	coverage	0.934	0.729	0.899	0.932	0.929
	length	0.150	0.155	0.218	0.146	0.199
G=400	ratio	0.980	0.772	0.988	0.968	0.942
	coverage	0.953	0.829	0.926	0.947	0.943
	length	0.106	0.117	0.152	0.103	0.141
T=15						
G=100	ratio	0.955	0.537	1.002	0.939	0.965
	coverage	0.940	0.478	0.800	0.940	0.951
	length	0.117	0.088	0.149	0.113	0.152
G=200	ratio	1.016	0.681	1.028	1.006	0.996
	coverage	0.964	0.700	0.897	0.952	0.947
	length	0.082	0.077	0.118	0.079	0.109
G=400	ratio	0.977	0.746	1.004	0.988	0.978
	coverage	0.945	0.810	0.914	0.943	0.945
	length	0.058	0.061	0.085	0.056	0.077

Table 1.8: Bias and RMSE, $\rho = .5$, equi-correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	Estimator 1	Estimator 2
T=5					
G=100	bias	−0.001	−0.022	0.000	0.002
	sd	0.063	0.087	0.063	0.085
	rmse	0.063	0.090	0.063	0.085
G=200	bias	−0.002	−0.015	−0.002	−0.004
	sd	0.045	0.064	0.045	0.064
	rmse	0.045	0.066	0.045	0.064
G=400	bias	−0.000	−0.006	0.000	0.000
	sd	0.031	0.044	0.031	0.044
	rmse	0.031	0.045	0.031	0.044
T=10					
G=100	bias	0.001	−0.015	0.001	0.002
	sd	0.029	0.041	0.029	0.040
	rmse	0.029	0.044	0.029	0.040
G=200	bias	−0.001	−0.008	−0.000	−0.001
	sd	0.021	0.029	0.021	0.028
	rmse	0.021	0.030	0.021	0.028
G=400	bias	0.000	−0.004	0.000	−0.000
	sd	0.014	0.021	0.014	0.020
	rmse	0.014	0.021	0.014	0.020
T=15					
G=100	bias	0.000	−0.014	0.000	−0.001
	sd	0.020	0.028	0.020	0.027
	rmse	0.020	0.031	0.020	0.027
G=200	bias	0.000	−0.007	0.000	−0.000
	sd	0.013	0.019	0.013	0.018
	rmse	0.013	0.020	0.013	0.018
G=400	bias	0.000	−0.003	0.000	0.000
	sd	0.010	0.014	0.010	0.013
	rmse	0.010	0.014	0.010	0.013

Table 1.9: Bias and RMSE, $\rho = .5$, no correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	Estimator 1	Estimator 2
T=5					
G=100	bias	−0.001	−0.013	−0.000	−0.001
	sd	0.063	0.063	0.062	0.063
	rmse	0.063	0.065	0.062	0.063
G=200	bias	−0.002	−0.008	−0.002	−0.002
	sd	0.045	0.045	0.045	0.045
	rmse	0.045	0.046	0.045	0.045
G=400	bias	−0.000	−0.003	0.000	−0.000
	sd	0.031	0.031	0.031	0.031
	rmse	0.031	0.031	0.031	0.031
T=10					
G=100	bias	0.001	−0.008	0.001	0.001
	sd	0.029	0.031	0.029	0.029
	rmse	0.029	0.032	0.029	0.029
G=200	bias	−0.001	−0.005	−0.000	−0.001
	sd	0.021	0.022	0.021	0.021
	rmse	0.021	0.022	0.021	0.021
G=400	bias	0.000	−0.002	0.000	0.000
	sd	0.014	0.015	0.014	0.014
	rmse	0.014	0.015	0.014	0.014
T=15					
G=100	bias	0.000	−0.007	0.000	0.000
	sd	0.020	0.022	0.020	0.020
	rmse	0.020	0.023	0.020	0.020
G=200	bias	0.000	−0.003	0.000	0.000
	sd	0.013	0.014	0.013	0.013
	rmse	0.013	0.015	0.013	0.013
G=400	bias	0.000	−0.002	0.000	0.000
	sd	0.010	0.010	0.010	0.010
	rmse	0.010	0.010	0.010	0.010

Table 1.10: Bias and RMSE, $\rho = .5$, heteroscedasticity and correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	Estimator 1	Estimator 2
T=5					
G=100	bias	0.001	−0.035	0.001	0.003
	sd	0.079	0.109	0.076	0.103
	rmse	0.079	0.114	0.076	0.103
G=200	bias	−0.003	−0.022	−0.002	−0.004
	sd	0.058	0.080	0.055	0.078
	rmse	0.058	0.083	0.055	0.078
G=400	bias	−0.001	−0.012	−0.001	−0.002
	sd	0.039	0.057	0.038	0.054
	rmse	0.039	0.058	0.038	0.054
T=10					
G=100	bias	0.001	−0.017	0.001	0.002
	sd	0.033	0.045	0.032	0.043
	rmse	0.033	0.048	0.032	0.043
G=200	bias	0.000	−0.009	0.000	0.000
	sd	0.023	0.032	0.023	0.031
	rmse	0.023	0.034	0.023	0.031
G=400	bias	0.001	−0.004	0.001	0.000
	sd	0.016	0.023	0.015	0.022
	rmse	0.016	0.023	0.016	0.022
T=15					
G=100	bias	0.000	−0.015	0.000	−0.001
	sd	0.022	0.030	0.021	0.028
	rmse	0.022	0.033	0.021	0.028
G=200	bias	0.001	−0.006	0.001	0.001
	sd	0.014	0.021	0.014	0.020
	rmse	0.014	0.022	0.014	0.020
G=400	bias	0.000	−0.004	0.000	0.000
	sd	0.011	0.015	0.010	0.014
	rmse	0.011	0.015	0.010	0.014

Table 1.11: Inference, ρ = .5, equi-correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	AB w/ Windmeijer correction	Estimator 1	Estimator 2
T=5						
G=100	ratio	1.009	0.818	1.034	1.004	1.014
	coverage	0.959	0.875	0.948	0.957	0.951
	length	0.251	0.282	0.362	0.250	0.344
G=200	ratio	0.985	0.805	0.968	0.978	0.956
	coverage	0.945	0.875	0.943	0.944	0.942
	length	0.175	0.204	0.249	0.175	0.242
G=400	ratio	1.030	0.851	1.000	1.024	0.993
	coverage	0.961	0.907	0.950	0.960	0.950
	length	0.124	0.148	0.175	0.124	0.172
T=10						
G=100	ratio	0.969	0.682	0.986	0.969	0.974
	coverage	0.931	0.795	0.933	0.932	0.939
	length	0.112	0.111	0.165	0.112	0.153
G=200	ratio	0.973	0.746	0.977	0.975	0.987
	coverage	0.944	0.841	0.935	0.942	0.948
	length	0.079	0.086	0.117	0.079	0.110
G=400	ratio	0.997	0.790	0.979	0.997	0.969
	coverage	0.952	0.873	0.946	0.950	0.946
	length	0.056	0.064	0.081	0.056	0.077
T=15						
G=100	ratio	0.964	0.561	1.002	0.961	0.983
	coverage	0.945	0.666	0.925	0.943	0.950
	length	0.076	0.062	0.104	0.076	0.104
G=200	ratio	1.016	0.696	1.052	1.020	1.018
	coverage	0.952	0.795	0.951	0.955	0.950
	length	0.053	0.053	0.079	0.053	0.074
G=400	ratio	0.993	0.765	1.008	0.994	0.996
	coverage	0.949	0.862	0.939	0.951	0.942
	length	0.038	0.041	0.056	0.038	0.052

Table 1.12: Inference, $\rho = .5$, no correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	AB w/ Windmeijer correction	Estimator 1	Estimator 2
T=5						
G=100	ratio	1.015	0.960	1.022	1.007	1.004
	coverage	0.960	0.928	0.947	0.952	0.953
	length	0.252	0.242	0.264	0.249	0.249
G=200	ratio	0.990	0.960	0.990	0.978	0.980
	coverage	0.944	0.936	0.938	0.943	0.943
	length	0.176	0.172	0.180	0.175	0.175
G=400	ratio	1.032	1.010	1.027	1.026	1.026
	coverage	0.960	0.954	0.955	0.962	0.963
	length	0.125	0.123	0.126	0.124	0.124
T=10						
G=100	ratio	0.976	0.828	0.985	0.968	0.968
	coverage	0.938	0.880	0.942	0.928	0.932
	length	0.113	0.102	0.125	0.111	0.111
G=200	ratio	0.973	0.885	0.948	0.977	0.975
	coverage	0.945	0.917	0.927	0.943	0.942
	length	0.079	0.075	0.087	0.079	0.079
G=400	ratio	0.999	0.950	0.991	0.994	0.996
	coverage	0.952	0.938	0.952	0.951	0.951
	length	0.056	0.054	0.059	0.056	0.056
T=15						
G=100	ratio	0.970	0.701	0.970	0.961	0.962
	coverage	0.952	0.811	0.931	0.941	0.942
	length	0.076	0.061	0.077	0.076	0.076
G=200	ratio	1.018	0.852	1.017	1.020	1.018
	coverage	0.956	0.906	0.948	0.955	0.954
	length	0.054	0.048	0.059	0.053	0.053
G=400	ratio	0.996	0.930	1.012	0.995	0.995
	coverage	0.950	0.926	0.955	0.949	0.950
	length	0.038	0.036	0.041	0.038	0.038

Table 1.13: Inference, $\rho = .5$, heteroscedasticity and correlation within clusters

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	AB w/ Windmeijer correction	Estimator 1	Estimator 2
T=5						
G=100	ratio	1.019	0.798	1.025	1.016	0.992
	coverage	0.955	0.857	0.935	0.959	0.951
	length	0.320	0.345	0.448	0.306	0.407
G=200	ratio	0.983	0.809	0.977	0.994	0.943
	coverage	0.948	0.864	0.936	0.953	0.944
	length	0.223	0.255	0.313	0.215	0.290
G=400	ratio	1.018	0.828	0.985	1.020	0.982
	coverage	0.961	0.883	0.928	0.962	0.946
	length	0.157	0.185	0.220	0.152	0.207
T=10						
G=100	ratio	0.956	0.680	0.983	0.957	0.968
	coverage	0.938	0.795	0.929	0.930	0.949
	length	0.126	0.120	0.179	0.122	0.164
G=200	ratio	0.962	0.739	0.971	0.961	0.971
	coverage	0.937	0.840	0.931	0.935	0.942
	length	0.088	0.094	0.128	0.086	0.117
G=400	ratio	0.997	0.793	0.984	0.995	0.969
	coverage	0.952	0.877	0.945	0.953	0.942
	length	0.062	0.070	0.089	0.061	0.083
T=15						
G=100	ratio	0.957	0.542	0.988	0.956	0.974
	coverage	0.945	0.650	0.915	0.939	0.944
	length	0.082	0.064	0.109	0.080	0.108
G=200	ratio	1.022	0.678	0.999	1.026	0.988
	coverage	0.961	0.807	0.935	0.960	0.940
	length	0.058	0.055	0.083	0.056	0.077
G=400	ratio	0.971	0.758	1.002	0.980	0.983
	coverage	0.946	0.858	0.939	0.950	0.942
	length	0.041	0.043	0.059	0.040	0.055

1.6 Application: Estimation of Persistence in Student Achievement

In this section, we are interested in estimating the effect of attending private schools on student achievement in the province of Punjab, Pakistan. In a non-experimental framework, estimating the causal effects of some factors on student achievement requires accounting for factors that affected student achievements in previous time periods since these factors might affect students' learning ability in the future but also be correlated across time. A model for studying the effect of some factor x on student achievement y can be written as in Andrabi et al. (2011) in its summary of the work in Todd and Wolpin (2003):

$$y_{it} = \sum_{j=0}^{t-1} \alpha_j x_{it-j} + \sum_{j=0}^{t-1} \theta_j \mu_{it-j} \tag{1.6.1}$$

where μ_t are unobserved shocks to student achievement. If one assumes that both $\{\alpha_j\}_{j=1,...,T}$ and $\{\theta_j\}_{j=1,...,T}$ form geometric series such that $\alpha_j = \rho\alpha_{j-1}$ and $\theta_j = \rho\theta_{j-1}$, we can write:

$$y_{it} = \alpha x_{it} + \rho y_{it-1} + \mu_{it} \tag{1.6.2}$$

where θ_0 was normalized to one.

In order to account for the possibility that students have unobserved characteristics that affect their ability to learn and are related to other educational inputs in x_{it}, we can decompose μ_{it} between time constant unobserved factors (also called unobserved heterogeneity) and transitory shocks:

$$\mu_{it} = c_i + u_{it} \tag{1.6.3}$$

where c_i is arbitrarily related to $x_i = [x_{i1}, ..., x_{iT}]$ and x_{it} is either strictly exogenous:

$$E(u_{it}|X_i, Y_{it-1}) = 0 \tag{1.6.4}$$

or sequentially exogenous:

$$E(u_{it}|X_{it}, Y_{it-1}) = 0 \tag{1.6.5}$$

or contemporaneously endogenous:

$$E(u_{it}|X_{it-1}, Y_{it-1}) = 0 \tag{1.6.6}$$

where $X_i = [x_{i1}, ..., x_{iT}]$ and $X_{it-1} = [x_{i1}, ..., x_{it-1}]$.

In this section we will use the data analyzed in Andrabi et al. (2011) to estimate the effect of attending private schools on student achievement in three districts of Punjab in Pakistan so that the input of interest

is attendance of private school. The other covariates included are wealth and variables indicating whether each parent lives with the student. We treat all of these inputs as contemporaneously endogenous since it is likely that they follow a dynamic process with unobserved transitory shocks that are correlated with shocks to student achievement. For instance, an unobserved and unexpected increase in income might result in a student enrolling in private school but also benefiting from better study conditions at home so that $Cov(u_{it}, privateschool_{it}) \neq 0$ while it is still possible that $Cov(u_{it}, privateschool_{it-1}) = 0$.

It is likely that transitory shocks are correlated within schools since there are school or class-level unobserved shocks, such as changes in infrastructure, staff or teachers, that will affect all students within a school or class. The data-set we use collected between 0 and 25 students per school in each year with most schools being represented by less than 10 students, which is too small to estimate time-varying school fixed effects accurately. Instead, we prefer treating u_{it} as cross-sectionally correlated within schools. It is likely that unobserved heterogeneity will also be correlated across students within schools since students might attend specific schools based on unobserved characteristics, such as residential location, socio-economic characteristics or past achievements, that relate to their performance. As described in the rest of the paper, using this cross-sectional correlation for estimation can result in significant efficiency gains.

For any subject j among English, Urdu and Mathematics (denoted E, U, M), denote by y_{it}^j the grade obtained by student i in year t and subject j and denote by x_{it} the variable indicating whether student i attended a private school in year t. Let w_{it} be the vector containing other predetermined explanatory variables for student i at time period t. Denote by u_{it}^j transitory shocks in achievement in subject j and denote measurement error in that achievement by ε_{it}^j. Also denote by g_{it} the school attended by student i in year t. We will assume clustering so that, in a given year, transitory shocks are independent across schools. We consider a model with measurement error and contemporaneously endogenous covariates. As in Andrabi et al. (2011), we assume that measurement errors are independent across subjects. We can write such a model as:

$$y_{it}^j = d_t^j + \alpha_0^j x_{it} + \rho_0^j y_{it-1}^j + w_{it}\beta_0^j + c_i^j + u_{it}^j + \varepsilon_{it}^j - \rho_0^j \varepsilon_{it-1}^j \tag{1.6.7}$$

$$E(u_{it}^j | X_{t-1}, Y_{t-1}, W_{t-1}, g_{t-1}) = 0 \tag{1.6.8}$$

$$E(\varepsilon_{it}^j | X_t, Y_t^{-j}, W_{t-1}, g_{t-1}) = 0 \tag{1.6.9}$$

where $Y_t = \{Y_t^k\}_{k=E,U,M}$, $Y_t^{-j} = \{Y_t^k\}_{k\neq j,k=E,U,M}$ and d_t^j are time specific intercepts. The first difference with the model used in Andrabi et al. (2011) is that we use predetermined instruments instead

of sequentially exogenous instruments. This is a more suitable assumption since, as explained previously, the covariates used in this model are likely jointly determined. The second difference is that we include as potential instruments the lagged values of the covariates for all observations instead of only $X_{it-1}, Y_{it-1}^{-j}, W_{it-1}$. As pointed out in Section 1.2, since u_{it}^{j} is correlated cross-sectionally, it is unlikely that $E(u_{it}^{j}|X_{t-1}, Y_{it-1}, W_{it-1}) = 0$ holds without $E(u_{it}^{j}|X_{t-1}, Y_{t-1}, W_{t-1}) = 0$ holding. It could be interesting to introduce peer effects in the model but we do not consider it here for simplicity and comparability with results in Andrabi et al. (2011).

In this application, clusters (school membership) are not time constant and, as pointed out previously, not strictly or sequentially exogenous. Therefore it is possible that:

$$E(u_{it}|g_{it}, X_{t-1}, Y_{t-1}, W_{t-1}) \neq 0 \tag{1.6.10}$$

even though:

$$E(u_{it}|g_{it-1}, X_{t-1}, Y_{t-1}, W_{t-1}) = 0 \tag{1.6.11}$$

Hence we can use as instruments lagged values of achievements of students from schools where an observation was previously enrolled but not from schools where it is currently enrolled.

There are three time periods $t = 0, 1, 2$ available in the data-set used so the only transformed equation for each subject that can be used for estimation is:

$$\Delta y_{i2}^{j} = \delta_0^{j} + \alpha_0^{j}\Delta x_{i2} + \rho_0^{j}\Delta y_{i1}^{j} + \Delta w_{i2}\beta_0^{j} + \Delta u_{i2}^{j} + \Delta\varepsilon_{i2}^{j} - \rho\Delta\varepsilon_{i1}^{j} \tag{1.6.12}$$

$$E(\Delta u_{i2}^{j}|X_0, Y_0^{-j}, W_0, g_0) = 0 \tag{1.6.13}$$

$$E(\Delta\varepsilon_{i2}^{j}|X_0, Y_0^{-j}, W_0, g_0) = 0 \tag{1.6.14}$$

$$E(\Delta\varepsilon_{i2}^{j}|X_0, Y_0^{-j}, W_0, g_0) = 0 \tag{1.6.15}$$

where $\delta_0^{j} = d_2^{j} - d_1^{j}$.

Let $\phi^{j} = [\delta^{j}, \alpha^{j}, \rho^{j}, \beta^{j\prime}]^{\prime}$, $m_i^{j}(\phi^{j}) = \Delta y_{i2}^{j} - (\delta^{j} + \alpha^{j}\Delta x_{i2} + \rho^{j}\Delta y_{i1}^{j} + \Delta w_{i2}\beta^{j})$ and $m_i^{j} = m_i^{j}(\phi_0^{j})$. The Arellano and Bond estimator for this model is defined by:

$$\hat{\phi}_{AB}^{j} = argmin_{\phi^{j}}\Big(\sum_{i=1}^{n} Z_i^{jAB} m_i^{j}(\phi^{j})\Big)^{\prime}\Big(\sum_{i=1}^{n} Z_i^{jAB} m_i^{j}(\tilde{\phi}^{j}) m_i^{j}(\tilde{\phi}^{j})^{\prime} Z_i^{jAB\prime}\Big)^{-1}\Big(\sum_{i=1}^{n} Z_i^{jAB} m_i^{j}(\phi^{j})\Big) \tag{1.6.16}$$

where $\tilde{\phi}^{j}$ is a preliminary estimator of ϕ_0^{j} and $Z_i^{jAB} = [1, y_{i0}^{-j}, x_{i0}, w_{i0}]^{\prime}$.

This estimator is inefficient because it ignores cross-sectional dependence[11]. Using the previous results in this paper, we can specify auxiliary assumptions so that an estimator can be derived which will be consistent as long as the identifying assumptions defined above hold and efficient if the auxiliary assumptions also hold.

The first auxiliary assumption we can use is of conditional homoscedasticity and cluster equi-correlation. For $j = U, E, M$:

$$Cov(u_{it}^j, u_{ls}^j | Y_0^{-j}, X_0, W_0, g_0) = \sigma_{uj}^2 \, if \, i = l, t = s \tag{1.6.17}$$

$$= \tau_j \sigma_{uj}^2 \, if \, i \neq j, \, g_{it} = g_{ls}, t = s \tag{1.6.18}$$

$$= 0 \, otherwise \tag{1.6.19}$$

and:

$$Cov(u_{it}^j, \varepsilon_{ls}^j | Y_0^{-j}, X_0, W_0, g_0) = 0 \, \forall \, i, l, j, k, t, s \tag{1.6.20}$$

$$Cov(\varepsilon_{it}^j, \varepsilon_{ls}^j | Y_0^{-j}, X_0, W_0, g_0) = \sigma_{\varepsilon j}^2 \, if \, i = l, \, j = k, t = s \tag{1.6.21}$$

$$= 0 \, otherwise \tag{1.6.22}$$

Under this assumption:

$$Cov(m_i^j, m_l^j | Y_0^{-j}, X_0, W, g_0) = 2\sigma_{uj}^2 + 2\sigma_{\varepsilon j}^2 (1 + \rho + \rho^2) \, if \, i = j \tag{1.6.23}$$

$$= \tau_j \sigma_{uj}^2 (1[g_{i1} = g_{l1}] + 1[g_{i2} = g_{l2}]) \, if \, i \neq j \tag{1.6.24}$$

Under the previous auxiliary assumption, the optimal instruments for $m_i^j(\phi^j)$ will be linear functions of $E(\Delta y_{i1}^j | Y_0^{-j}, X_0, W_0, g_0)$, $E(\Delta x_{i2} | Y_0^{-j}, X_0, W_0, g_0)$ and $E(\Delta w_{i2} | Y_0^{-j}, X_0, W_0, g_0)$. Since we have $T = 2$ so that there is only one transformed equation available for estimation, we can use the simple second auxiliary

[11] Without measurement error, it would also be possible to use correlation of transitory shocks across outcomes to obtain an efficient joint estimator of $\{\phi^j\}_{j=U,E,M}$. However because of measurement error, the sets of instruments across subjects are non-overlapping, so that optimal instruments cannot be derived. Since there is no restriction in the parameters across equations, weighting of optimally weighted moment conditions or minimum distance methods cannot be used either.

assumption:

$$E(\Delta y_{i1}^j|Y_0^{-j},X_0,W_0,g_0) = a_0^j + \sum_{k\neq j} a_{01k}^j y_{i0}^k + \sum_{k\neq j} a_{02k}^j \frac{1}{\#g_{i0}} \sum_{l\in g_{i0}} y_{l0}^k + a_{03}^j x_{i0}$$
$$+ a_{04}^j w_{i0} + a_{05}^j \frac{1}{\#g_{i0}} \sum_{l\in g_{i0}} w_{l0} \tag{1.6.25}$$

$$E(\Delta z_{i2}|Y_0^{-j},X_0,W_0,g_0) = b_{z0}^j + \sum_{k\neq j} b_{z01k}^j y_{i0}^k + \sum_{k\neq j} b_{z02k}^j \frac{1}{\#g_{i0}} \sum_{l\in g_{i0}} y_{l0}^k + b_{z03}^j x_{i0}$$
$$+ b_{z04}^j w_{i0} + b_{z05}^j \frac{1}{\#g_{i0}} \sum_{l\in g_{i0}} w_{l0} \tag{1.6.26}$$

where $z \in \{x,w\}$and consistently estimate these unknown parameters by OLS regression.

Define $\hat{E}_i^{\Delta y^j}$, $\hat{E}_i^{\Delta x^j}$ and $\hat{E}_i^{\Delta w^j}$ to be the estimated conditional expectations defined by (1.6.25) and (1.6.26). Define:

$$D_i^{j'} = \begin{bmatrix} \hat{E}_i^{\Delta y^j} \\ \hat{E}_i^{\Delta x^j} \\ \hat{E}_i^{\Delta w^j} \end{bmatrix} \tag{1.6.27}$$

and $D^{j'} = [D_1^{j'},...,D_n^{j'}]$. Define $\hat{\Sigma}^j(\phi^j) = [\hat{Cov}(m_i^j(\phi^j), m_l^j(\phi^j))]_{i=1,...,n}^{l=1,...,n}$ and

$$m^j(\phi^j) = [m_1^{j'}(\phi^j),...,m_n^{j'}(\phi^j)]' \tag{1.6.28}$$

The efficient estimator for ϕ_0^j under the auxiliary assumptions is $\hat{\phi^j}_{opt}$ defined by:

$$D^{j'}\hat{\Sigma}(\hat{\phi}_{opt}^j)^{-1} m(\hat{\phi}_{opt}^j) = 0 \tag{1.6.29}$$

Let $M_i^{j'} = \begin{bmatrix} \Delta y_1^j \\ \Delta x_{i2} \\ \Delta w_{i2} \end{bmatrix} = (\frac{\partial m_i^j(\phi)}{\partial \phi'})'$. Both the Arellano and Bond estimator and our optimal estimator can be written as:

$$\sum_{i=1}^n Z_i^j m_i^j(\hat{\phi}^j) = 0 \tag{1.6.30}$$

where for the Arellano and Bond estimator:

$$Z_i^j = (\sum_{l=1}^n M_l^{j'} Z_l^{jAB'}) \hat{\Theta}^{jAB-1} Z_i^{jAB} \tag{1.6.31}$$

with

$$\hat{\Theta}^{jAB} = \sum_{i=1}^n Z_i^{jAB} m_i^j(\tilde{\phi}^j) m_i^j(\tilde{\phi}^j)' Z_i^{jAB'} \tag{1.6.32}$$

For our optimal estimator, Z_i^j is the i^{th} column of $D^{j'}\hat{\Sigma}^j(\hat{\phi}_{opt}^j)^{-1}$.

Under the assumption that transitory shocks are independent across schools, $\hat{\phi}^j$ is consistent for ϕ_0^j and asymptotically normal. The asymptotic variance-covariance matrix of both estimators is[12]:

$$AVar(\hat{\phi}) = A^{'}BA \tag{1.6.33}$$

$$A = plim(\frac{1}{n}\sum_{i=1}^{n} Z_i^j M_i^j)^{-1} \tag{1.6.34}$$

$$B = plim(\frac{1}{n}\sum_{i=1}^{n} Z_i^j m_i^j \sum_{i=1}^{n} m_i^{j'} Z_i^{j'}) \tag{1.6.35}$$

$$= plim(\frac{1}{n}\sum_{i=1}^{n} Z_i^j m_i^j \sum_{l=1}^{n} 1[\{g_{it} = g_{ls}\}_{t,s=1,2}] m_l^{j'} Z_l^{j'}) \tag{1.6.36}$$

which can be estimated consistently since there is a small number of observations in each school.

The students' achievement in each subject was measured by the results obtained by students on a test administed by the authors of Andrabi et al. (2011) and graded using the Item Response Theory so that scores can be compared across students and years and the standard deviation of scores in the first year (third grade) is one. Table 1.14 shows the average and standard deviations of scores by subject and grade. Table 1.15 reports the estimated degree of persistence and the estimated effect of attending private schools on performance for the three subjects considered. We also show the associated standard errors and 95% confidence intervals. Similarly as in Andrabi et al. (2011), we find that there is significant persistence in scores except for Mathematics. We estimate effects of attending private school that are smaller than in Andrabi et al. (2011), which can be attributed to Andrabi et al. (2011) treating the covariates as sequentially exogenous instead of contemporaneously endogenous while it is likely that unobserved factors simultaneously affect performance and school attendance, as explained previously. The optimal estimator we presented in this section yields smaller standard errors compared to the Arellano and Bond estimator both for estimating persistence in student achievements and for estimating the effect of attending private school, with particularly significantly smaller standard errors for the latter.

[12]Note that clustering standard errors by the first school attended, which is used in Andrabi et al. (2011), is not justified since transitory shocks should be correlated within a school that a child is currently attending and not necessarily only across students who attended the same school in the first time period.

Table 1.14: Averages and standard deviations of scores per subject and per grade

	English		Math		Urdu	
	Average	s.d.	Average	s.d.	Average	s.d.
Grade 3	0	1	0	1	0	1
Grade 4	0.18	1.04	0.18	1.11	0.24	1.10
Grade 5	0.68	0.89	0.81	1.04	0.82	0.94

Table 1.15: Effects of Attending Private Schools on Student Achievement

	Optimal Estimator			Arellano and Bond Estimator			Andrabi et al. 2010		
	English	Urdu	Math	English	Urdu	Math	English	Urdu	Math
Persistence	0.31	0.30	0.04	0.34	0.53	0.23	0.19	0.35	0.12
	(0.14)	(0.12)	(0.11)	(0.17)	(0.18)	(0.14)	(0.10)	(0.11)	(0.12)
	[0.04,0.58]	[0.06,0.54]	[-0.18,0.26]	[0.01,0.67]	[0.18,0.88]	[-0.04,0.50]	[-0.01,0.39]	[0.13,0.57]	[-0.12,0.36]
Private School	0.44	0.89	0.30	0.40	0.81	0.43	1.15	0.90	0.46
	(0.38)	(0.41)	(0.31)	(0.55)	(0.59)	(0.54)	(0.39)	(0.48)	(0.50)
	[-0.30,1.18]	[0.09,1.69]	[-0.31,0.91]	[-0.68,1.48]	[-0.35,1.97]	[-0.63,1.49]	[0.39,1.91]	[-0.04,1.84]	[-0.52,1.44]

Numbers in parenthesis are standard errors and intervals are 95% confidence intervals. Standard errors and confidence intervals in Andrabi et al. 2010 do not take into account changes in school attendance across time. Covariates are treated as sequentially exogenous instead of contemporaneously endogenous in Andrabi et al. 2010.

1.7 Conclusion

We have presented an estimation method that used cross-sectional dependence to improve the accuracy with which dynamic models of panel data are estimated while making use of few nuisance parameters and being robust to the misspecification of the form of the cross-sectional dependence. This method can be generalized to models with covariates that are strictly exogenous, sequentially exogenous or contemporaneoulsy endogenous.

Monte Carlo simulations and an application to the estimation of a value-added model show that, when there is cross-sectional dependence, this method dominates existing estimators in terms of accuracy and quality of inference.

Extensions of this work that are the subject of ongoing research consider the generalization of the results in this paper to non-linear panel data models, the use of other forms of cross-sectional dependence than clustering in the auxiliary restrictions, and the asymptotic properties of our estimator with large numbers of time period and of observations within clusters.

CHAPTER 2

ESTIMATION OF UNOBSERVED EFFECTS PANEL DATA MODELS UNDER SEQUENTIAL EXOGENEITY

2.1 Introduction

Time constant unobserved effects are now routinely introduced in models of panel data to address endogeneity issues that are due to time constant unobserved variables. A first group of estimators for such models uses iterated conditioning by specifying an auxiliary model for unobserved effects conditional on the covariates. Such models are commonly called Correlated Random Effects (CRE) models. A second group of estimators implements instrumental variable estimation methods on transformed data as long as some specific functional form assumptions can be made.

In the case of linear models with strictly exogenous covariates, CRE estimators have first been proposed in Mundlak (1978) and Chamberlain (1982). When covariates are strictly exogenous, Wooldridge (2010) contains many examples of the generalization of this approach to non-linear models of panel data. For linear models with strictly exogenous covariates, the instrumental variable estimators are the well-known Fixed Effects estimator and the First Difference estimator. These estimators have also been generalized to non-linear models that are linear in random coefficients in Chamberlain (1992b).

The estimators mentioned above cannot be used in applications where there are sequentially exogenous covariates or, more generally, instruments. Sequentially exogenous instruments arise when transitory shocks to the dependent variable are independent of past and current values of the instruments but affect future values of the instruments. This scenario is particularly plausible in a dynamic optimization framework. The simplest example is when lagged values of the dependent variables are used as covariates and instruments. Dynamic models are frequently used to analyze panel data. A review of linear dynamic models for panel data can be found in Bond (2002). Sometimes instruments other than lagged values of the dependent variable can be sequentially exogenous. This is the case, for instance, in Clerides et al. (1998) which investigates the causal effect of exporting on firm efficiency but recognizes that shocks to firm efficiency will affect current and future exports as well. Another well know example is Blundell et al. (1995) which investigates the causal effect of Research and Development spending on innovation knowing that current success in

deposing patents will affect future Research and Development spending.

CRE models encounter strong limitations when instruments are sequentially exogenous. Dynamic CRE models can be used for special cases where lagged dependent variables are included in the list of covariates but all other covariates are strictly exogenous. Such models and the corresponding estimation methods are discussed in Wooldridge (2005). One application can be found in Browning et al. (2010). In more general cases, however, one would need a very large set of auxiliary assumptions in order to use a CRE model to analyze panel data with sequential exogeneity.

Instrumental variable estimation can be used for estimation of panel data models with sequential exogeneity as long as there exists a transformation of the data so that the method of moments can be applied, which makes it a much more flexible method. For models with additive unobserved effects, such transformations of the data are presented in Arellano and Bond (1991). Chamberlain (1992a) and Wooldridge (1997) discuss transformations of the data for models with multiplicative unobserved effects. Once the transformed equations are obtained, these papers advocate for a two-step GMM estimation of the unknown parameters using the transformed equations at each time period and the corresponding available sets of instruments. These estimators are efficient given the set of unconditional moment conditions that are used, but they are still known for suffering from a weak instrumental variable problem that can hinder their use in practice.

In this paper, we consider using additional assumptions to derive useful additional moment conditions and hence obtain a more precise estimator. The additional moment conditions that we present in this paper are generalized versions of the additional moment conditions for the linear dynamic model with additive unobserved effects presented in Arellano and Bover (1995), Ahn and Schmidt (1995) and Blundell and Bond (1998). Windmeijer (2000) considered some of the additional moment conditions we present here, namely uncorrelation of the transitory shocks, for a special case of the group of models we define. However, the chosen set of assumption in Windmeijer (2000) is actually too weak to support the moment conditions that are used for estimation. Hence it seems useful to present these moment conditions here as part of a unifying framework.

In Section 2.2 we will present the model and assumptions we use. In Section 2.3 we will discuss the estimator that is currently used. In Section 2.4 we will present additional sets of restrictions that can be used for estimation when instruments are stationary or when transitory shocks are serially uncorrelated. In Section 2.5 we will show using Monte Carlo simulations that the propositions to address the weak instrumental variable problem of these models result in significant improvements in accuracy and hence effectively mit-

igate the weak instrumental variable problem. In Section 2.6 we will show how to estimate and perform inference on measures of interest of the effect of covariates on the mean of the dependent variable.

2.2 Model and Assumptions

The models we consider are such that for each observation i of a random sample of large size n and each time period t of a fixed number of time periods T we can specify:

$$E(y_{it}|x_i^t, u_i^t, z_{it}) = h_0(x_{it}, \beta_0) + h_1(x_{it}, \beta_0)u_{it} \tag{2.2.1}$$

$$E(u_{it}|z_{it}) = E(u_{it+1}|z_{it}) \; \forall t \leq T-1 \tag{2.2.2}$$

for known functions h_0, h_1. In this model x_{it} are observed covariates and u_{it} captures the effect of unobserved covariates. $x_i^t = \{x_{is}\}_{s=1,...,t}$ contains all values of the covariates up to the current time period and similarly we denote $u_i^t = \{u_{is}\}_{s=1,...,t}$. z_{it} are observed instruments that do not belong to the mean equation for y_{it} once we condition on the observed and unobserved covariates x_{it}, u_{it}. We consider cases where $z_{i1} \subseteq z_{i2} \subseteq ... \subseteq z_{iT}$ so that we have sequential exogeneity, also called predetermined instruments. (2.2.1) was specified in terms of a conditional expectation instead of simply in terms of y_{it} in order to allow for dependent variables with discrete supports as we will see later in this section. (2.2.2) requires that at each time period the effects of unobserved covariates have the same mean conditional on the instruments as the effects of unobserved covariates at future time periods. Hence it requires that the source of endogeneity of the instruments be time constant. For simplicity we will consider the case where y_{it}, $h_0(.,.)$, $h_1(.,.)$ and c_i are scalars but all the results can be generalized to systems of equations if needed.

Dynamic linear models with additive heterogeneity are a special case of the group of models described by (2.2.1) and (2.2.2) with $x_{it} = y_{it-1}$, $z_{it} = [y_0, ..., y_{it-1}]$, $h_0(x, \beta) = \beta x$, $h_1(.,.) = 1$:

$$y_{it} = \beta_0 y_{it-1} + c_i + \nu_{it} \tag{2.2.3}$$

$$E(c_i + \nu_{it}|y_{it-1}, ..., y_0) = E(c_i|y_{it-1}, ..., y_0) \tag{2.2.4}$$

Here we wrote $u_{it} = c_i + \nu_{it}$. Traditionally unobserved effects have been explicitely decomposed between a time constant part, sometimes called unobserved heterogeneity, and a transitory part. In this paper we keep a more general notation as in (2.2.1) and (2.2.2) for more flexibility.

Other special cases of the models we consider have been used to model count dependent variables, such as the linear feedback model presented in Blundell et al. (2002):

$$E(y_{it}|y_{it-1},...,y_{i0},x_{it},...,x_{i1},c_i) = \gamma_0 y_{it-1} + exp(x_{it}\theta_0)c_i \; \forall t = 1,...,T \tag{2.2.5}$$

In this case we simply have $u_{it} = c_i$.

Our specification also includes models for count dependent variables where covariates cannot be used as instruments, i.e. $x_{it} \notin z_{it}$, but where enough instruments are available to identify the parameters of the model. An example where instruments available are lagged covariates is presented in Windmeijer (2000)[1]:

$$y_{it} = exp(x_{it}\beta_0)c_i\nu_{it} \tag{2.2.6}$$

$$E(c_i\nu_{it}|x_{it-1},...,x_{i0}) = E(c_i|x_{it-1},...,x_{i0}) \tag{2.2.7}$$

Multiplicative unobserved effects models have first been used to analyze count data and a description of the state of the literature on dynamic models of count data with unobserved heterogeneity is given in Windmeijer (2008). However the class of models we consider in this paper is very appropriate for the analysis of any data that require the specification of a non-linear response function like binary, fractional, ordered, non-negative, corner solution response data and so on. For a binary dependent variable for instance, a dynamic probit model with sequential exogeneity in the explanatory variables could be specified:

$$E(y_{it}|y_{it-1},...,y_{i0},c_i,x_{it},...,x_{i1}) = c_i\Phi(\gamma_0 y_{it-1} + x_{it}\theta_0) \tag{2.2.8}$$

Here $c_i/2$ is the conditional probability of being in state $y_{it} = 1$ when $y_{it-1} = 0, x_{it} = 0$ and also captures a time constant unobserved propensity to be in state $y_{it} = 1$.

It is also important to note that the generality of our chosen specification also allows us to use models where some of the explanatory variables are endogenous but where instruments are available:

$$E(y_{it}|y_{it-1},...,y_{i0},x_{it},...,x_{i1},u_{it},...,u_{i1},z_{it}) = \Phi(\gamma_0 y_{it-1} + x_{it}\theta_0)u_{it} \tag{2.2.9}$$

$$E(u_{it}|z_{it}) = E(u_{it+1}|z_{it}) \tag{2.2.10}$$

where u_{it} is a random variable between zero and one and captures the effect on the mean of y_{it} of unobserved explanatory variables which are not independent from x_{it} but have the same mean conditional on instruments as effects on the mean in future time periods.

[1] We present slightly different assumptions here than in Windmeijer (2000) since Windmeijer (2000) goes to great lengths to avoid making assumptions on conditional means and only consider assumptions of uncorrelation but does so at the expense of making two mistakes. One of which being that assuming that x_{it-1} is uncorrelated with ν_{it} does not imply $E(x_{it-1}c_i(\nu_{it+1} - \nu_{it})) = 0$ as is claimed in Windmeijer (2000). The other one will be mentioned in Section 2.4.2.

2.3 Estimation without Additional Assumptions

Following the argument made in Chamberlain (1992a), the model described by (2.2.1), (2.2.2) is statistically indistinguishable from:

$$E(\frac{y_{iT}-h_0(x_{iT},\beta_0)}{h_1(x_{iT},\beta_0)}|z_{iT}) = E(u_{iT}|z_{iT}) \tag{2.3.1}$$

$$E(\Delta\frac{y_{it}-h_0(x_{it},\beta_0)}{h_1(x_{it},\beta_0)}|z_{it-1}) = 0 \; \forall t = 2,...,T \tag{2.3.2}$$

where Δ denotes the difference operator. Since $E(u_{iT}|z_{iT})$ is unknown and unrestricted, (2.3.1) does not participate in estimating β_0. Therefore we can restrict our attention to estimating β_0 from (2.3.2). For notation we will write:

$$\rho_t(w_i,\beta) \equiv \Delta\frac{y_{it}-h_0(x_{it},\beta_0)}{h_1(x_{it},\beta)} \; \forall t = 2,...,T \tag{2.3.3}$$

where $w_i \equiv \{y_{it},x_{it}\}_{t=1,...,T}$. So the conditional moment restrictions available for estimation are:

$$E(\rho_t(w_i,\beta_0)|z_{it-1}) = 0 \; \forall t = 2,...,T \tag{2.3.4}$$

Chamberlain (1992a) has shown that an optimal estimator would be $\hat{\beta}_{opt}$ that solves:

$$\sum_{i=1}^{n} \tilde{D}_{it}^{'}\tilde{\Sigma}_{it}^{-1}\tilde{\rho}_t(w_i,z_i,\hat{\beta}_{opt}) = 0 \tag{2.3.5}$$

Such an estimator would achieve the asymptotic information bound for estimating β_0 from these conditional moment restrictions which is $J = E(\sum_{t=2}^{T}\tilde{D}_{it}^{'}\tilde{\Sigma}_{it}^{-1}\tilde{D}_{it})$ where $\tilde{D}_{it} \equiv E(\frac{\partial\tilde{\rho}_t}{\partial\beta'}(w_i,z_i,\beta_0)|z_{it-1})$, $\tilde{\Sigma}_{it} \equiv Var(\tilde{\rho}_t(w_i,z_i,\beta_0)|z_{it-1})$, $\tilde{\rho}_t()$ is defined by:

$$\tilde{\rho}_T(w_i,z_i,\beta) = \rho_T(w_i,\beta) \tag{2.3.6}$$

$$\tilde{\rho}_t(w_i,z_i,\beta) = \rho_t(w_i,\beta) - \Gamma_{it,t+1}\tilde{\rho}_{t+1}(w_i,z_i,\beta) - ... - \Gamma_{it,T}\tilde{\rho}_T(w_i,z_i,\beta) \; \forall t = 2,...,T-1 \tag{2.3.7}$$

where $z_i = \{z_{it-1}\}_{t=2,...,T}$, $\Gamma_{it,s} \equiv Cov(\rho_{it},\tilde{\rho}_{is}|z_{is-1})Var(\tilde{\rho}_{is}|z_{is-1})^{-1} \; \forall s > t$ where $\rho_{it} = \rho_{it}(\beta_0)$, $\rho_{it}(\beta) = \rho_t(w_i,\beta)$ and $\tilde{\rho}_{it} = \tilde{\rho}_{it}(\beta_0)$, $\tilde{\rho}_{it}(\beta) = \tilde{\rho}_t(w_i,z_i,\beta)$. The intuition behind this result is that the asymptotic information bound from all the sequential conditional moment restrictions is the sum of the information bounds for each conditional moment restriction once these restrictions have been orthogonalized.

Unfortunately the optimal estimator from equation (2.3.5) is usually not feasible without additional assumptions since $\tilde{D}_{it}$, $\tilde{\Sigma}_{it}$ and $\tilde{\rho}_{it}$ are not observed, i.e. they are not known functions of the data and of β_0. One could think about approximating such moment conditions arbitrarily well, as suggested in Chamberlain

(1992a) or partially studied in Hahn (1997), but this introduces several new problems and therefore is left for future research.

Under the conditional moment restrictions given in (2.3.4), any function of z_{it-1} can be used as instruments for $\rho_{it}(\beta)$ to estimate β_0. Windmeijer (2008), for instance, recommends the use of all available lags of the instruments in levels, in our notation this is just z_{it-1}. So the estimator that is commonly used to estimate β_0 from the model given in (2.2.1) and (2.2.2) is:

$$\hat{\beta} = \underset{\beta}{argmin} \sum_i (Z_i\rho_i(\beta))' (Z_i\rho_i(\tilde{\beta})\rho_i(\tilde{\beta})' Z_i')^{-1} \sum_i Z_i\rho_i(\beta) \tag{2.3.8}$$

where $\tilde{\beta}$ is a preliminary $\sqrt{n}$-consistent estimator of β_0, $\rho_i(\beta) = [\rho_{it}(\beta)]_{t=2,...,T}$,

$$Z_i = \begin{bmatrix} z_{i1}' & 0 & ... & 0 \\ 0 & z_{i2}' & & ... \\ ... & & ... & 0 \\ 0 & ... & 0 & z_{iT-1}' \end{bmatrix} \tag{2.3.9}$$

The asymptotic variance of of $\sqrt{n}(\hat{\beta} - \beta_0)$ is:

$$Avar = (E(Z_i \frac{\partial \rho_i(\beta_0)}{\partial \beta})' E(Z_i\rho_i(\beta_0)\rho_i(\beta_0)' Z_i')^{-1} E(Z_i \frac{\partial \rho_i(\beta_0)}{\partial \beta}))^{-1} \tag{2.3.10}$$

It is shown in Appendix B.1 that this asymptotic variance is equal to:

$$Avar = (E(\sum_{t=2}^{T} \tilde{D}_{it}' \tilde{\Sigma}_{it}^{-1} \tilde{D}_{it}) - E(\sum_{t=2}^{T} e_{it}' e_{it}))^{-1} \tag{2.3.11}$$

where e_t is the error term from the linear projection of $\tilde{D}_{it}' \tilde{\Sigma}_{it}^{-1/2}$ on $\tilde{z}_{it-1} = \{z_{is}\Gamma_{s+1,t}\tilde{\Sigma}_{it}^{1/2}\}_{s=1,...,t-1}$, so that $\hat{\beta}$ can be seen as the estimator resulting from a linear approximation of the optimal moment conditions.

This estimator is often quite imprecise. For dynamic linear models, the weak instrumental variable problem affecting the estimator described in this section has been documented in Arellano and Bover (1995), Ahn and Schmidt (1995) and Blundell and Bond (1998). The additional moment conditions that are proposed in these papers to address this issue are the ones we generalize to our more general set up in the next section. Blundell et al. (2002) have documented the weak instrumental variable problem for estimating the Linear Feedback Model for count data, which we will use in the next section as an example. The additional assumptions used by Blundell et al. (2002) in order to alleviate the weak instrumental variable problem are quite unrealistic compared to the additional sets of assumptions we present in the next section. In addition, Monte

Carlo simulations show that using the additional assumptions presented in this paper achieves efficiency gains that are similar to those obtained in Blundell et al. (2002) when both sets of additional assumptions hold.

2.4 Additional Assumptions

2.4.1 Estimation with Stationary Instruments

For the models described in (2.2.1), it is possible in some applications that part of the instruments, denote it z_{it}^{stat}, has a time constant covariance with the unobserved effects and a time constant mean so that in addition to (2.2.2) we can assume:

$$E(z_{it}^{stat}) = \mu_{z^{stat}} \tag{2.4.1}$$

$$Cov(z_{it}^{stat}, u_{it}) = \gamma \tag{2.4.2}$$

(2.4.1) and (2.4.2) imply that $E(z_{it}^{stat} u_{it})$ is time constant as well since $E(z_{it}^{stat} u_{it}) = \gamma + \mu_{z^{stat}} E(u_{it})$ and $E(u_{it})$ is time constant by the law of iterated expectations and $E(u_{i1}|z_{i1}) = ... = E(u_{iT}|z_{i1})$ which is implied (2.2.2). This in turn implies that $E(z_{it}^{stat} u_{it}) = E(z_{it-1}^{stat} u_{it})$ since $E(z_{it-1}^{stat} u_{it-1}) = E(z_{it-1}^{stat} u_{it})$ from (2.2.2). Let K^{stat} be the dimension of z_{it}^{stat}. We can use for estimation the $(T-1) \times K^{stat}$ additional moment conditions[2]:

$$E((z_{it}^{stat} - z_{it-1}^{stat}) \frac{y_{it} - h_0(x_{it}, \beta_0)}{h_1(x_{it}, \beta_0)}) = 0 \; \forall t = 2, ..., T \tag{2.4.3}$$

2.4.1.1 Example of the Linear Feedback Model

An example of a model where such additional moment conditions can be used but have not been exploited in previous studies is the linear feedback model (LFM) presented in Blundell et al. (2002). For $|\gamma_0| < 1$:

$$E(y_{it}|y_{it-1}, ..., y_{i0}, x_{it}, ..., x_{i1}, c_i) = \gamma_0 y_{it-1} + c_i \mu(x_{it}, \theta_0) \tag{2.4.4}$$

[2]Note that the moment conditions $E((z_{it-s}^{stat} - z_{it-s-1}^{stat}) u_{it}) = 0$ for $s \geq 1$ do not constitute useful additional moment conditions since they are implied by the moment conditions $E((z_{it-s}^{stat} - z_{it-s-1}^{stat}) u_{it-s}) = 0$ and $E(z_{it-s}^{stat}(u_{it-\tau} - u_{it-\tau-1})) = 0 \; \forall \tau = 0, ..., s-1$ since:

$$\begin{aligned}(z_{it-1}^{stat} - z_{it-2}^{stat}) u_{it} &= z_{it-1}^{stat} u_{it} - z_{it-2}^{stat} \Delta u_{it} - z_{it-2}^{stat} u_{it-1} \\ &= z_{it-1}^{stat} u_{it} - z_{it-2}^{stat} \Delta u_{it} + \Delta z_{it-1}^{stat} u_{it-1} - z_{it-1}^{stat} u_{it-1} \\ &= z_{it-1}^{stat} \Delta u_{it} - z_{it-2}^{stat} \Delta u_{it} + \Delta z_{it-1}^{stat} u_{it-1}\end{aligned}$$

and by iteration: $(z_{it-s}^{stat} - z_{it-s-1}^{stat}) u_{it} = z_{it-s}^{stat} \Delta u_{it} - z_{it-s-1}^{stat} \Delta u_{it} + \Delta z_{it-s}^{stat} u_{it-1}$ is a function of $(z_{it-s}^{stat} - z_{it-s-1}^{stat}) u_{it-s}$ and $\{z_{it-s}^{stat}(u_{it-\tau} - u_{it-\tau-1}). \; \forall \tau = 0, ..., s-1\}$

For estimation we can use the sequence of conditional moment conditions corresponding to the conditional moment conditions (2.3.2) considered in the previous section:

$$E(\frac{y_{it}-\gamma_0 y_{it-1}}{\mu(x_{it},\theta_0)}-\frac{y_{it-1}-\gamma_0 y_{it-2}}{\mu(x_{it-1},\theta_0)}|y_{it-2},...,y_{i0},x_{it-1},...,x_{i1})=0\,\forall t=2,...,T \tag{2.4.5}$$

so for this specific model we have $u_{it}=c_i$.

Blundell et al. (2002) also assumes that x_{it} is strictly stationary conditional on c_i.[3] This implies that $E(\mu(x_{it},\theta_0)|c_i)=g_1(c_i)$ for some arbitrary function $g_1(.)$. Consider the difference equation given by:

$$y_{it}=\gamma_0 y_{it-1}+\mu(x_{it},\theta_0)c_i\varepsilon_{it}. \tag{2.4.6}$$

where $\varepsilon_{it}=\frac{y_{it}-\gamma_0 y_{it-1}}{c_i\mu(x_{it},\theta_0)}$. The associated stationary process is defined by

$$s_{it}=\sum_{s=1}^{\infty}\gamma_0^s c_i\mu(x_{it-s},\theta_0)\varepsilon_{it-s} \tag{2.4.7}$$

Then $E(s_{it}|c_i)=\frac{c_i g_1(c_i)}{1-\gamma_0}$ since $E(\varepsilon_{it}|c_i,x_{it},x_{it-1},...)=E(\frac{y_{it}-\gamma_0 y_{it-1}}{c_i\mu(x_{it},\theta_0)})=1$. So if we simply assume that the deviation of y_{i0} from s_{i0} has mean zero conditional on c_i, we have $E(y_{i0}|c_i)=\frac{c_i g_1(c_i)}{1-\gamma_0}$ so that $E(y_{it}|c_i)=\frac{c_i g_1(c_i)}{1-\gamma_0}\;\forall t=1,...,T$. This assumption is the generalization of the restriction on initial conditions made in Blundell and Bond (1998) for dynamic linear models with additive unobserved effects. It results in the additional over-identifying moment conditions:

$$E((y_{it-1}-y_{it-2})\frac{y_{it}-\gamma_0 y_{it-1}}{\mu(x_{it},\theta_0)})=E((y_{it-1}-y_{it-2})c_i)\;\forall t=2,...,T \tag{2.4.8}$$

$$=0\;\forall t=2,...,T \tag{2.4.9}$$

Since for this specific model these conditions would not be plausible without the stationarity of x_{it}, we can also add the moment conditions:

$$E((x_{it}-x_{it-1})\frac{y_{it}-\gamma_0 y_{it-1}}{\mu(x_{it},\theta_0)})=0\;t=2,...,T-1 \tag{2.4.10}$$

Monte Carlo simulations show that these extra moment conditions improve the efficiency of estimators significantly, even tough they rely on assumptions that are more realistic than the assumptions imposed in Blundell et al. (2002).

[3] They do so in a different attempt to mitigate the weak IV problem of FE estimators for the LFM. Blundell et al. (2002) proposes a so called pre-sample mean estimator which attempts to control for unobserved heterogeneity by using the average of observations on the dependent variable for many periods before the rest of the sample started as a proxy for time constant unobserved heterogeneity. However this estimator suffers from two severe drawbacks which make it unusable in practice: it supposes one has many observations on the dependent variable before the start of the rest of the sample but most importantly the assumptions under which the pre-sample average is a good proxy for unobserved heterogeneity are highly unrealistic, in particular it supposes that the covariates x_{it} have a mean that is proportional to c_i and restricts $\mu()$ to be the linear index exponential function.

2.4.1.2 Time Demeaned Instruments

In some applications it might not be plausible to assume that some of the instruments are mean stationary. However similar additional moment conditions as (2.4.3) can be obtained after time demeaning of the instruments if $E((z_{it}^{stat} - E(z_{it}^{stat}))u_{it}) = Cov(z_{it}^{stat}, u_{it})$ is time constant. In this section we consider the conditions necessary for this to be true when z_{it}^{stat} is not itself mean stationary.

From (2.2.2) alone, $Cov(z_{it}^{stat}, u_{it}) = Cov(z_{it}^{stat}, u_{iT})$ since (2.2.2) implies $E(z_{it}^{stat} u_{it}) = E(z_{it}^{stat} u_{iT})$ and $E(u_{it}) = E(u_{iT})$. $Cov(z_{it}^{stat}, u_{iT})$ will be time constant if $\forall t, s \leq T-1$:

$$Cov(z_{it}^{stat}, u_{iT}) - Cov(z_{is}^{stat}, u_{iT}) = 0 \tag{2.4.11}$$

$$Cov(z_{it}^{stat} - z_{is}^{stat}, u_{iT}) = 0 \tag{2.4.12}$$

Hence $Cov(z_{it}^{stat}, u_{iT})$ will be time constant if the change in z_{it}^{stat} over time is uncorrelated with the unobserved effects at the last time period. To provide more intuition regarding what such an assumption mean, we can consider the unobserved heterogeneity decomposition of unobserved effects and write $u_{iT} = c_i \nu_{iT}$ with c_i and ν_{iT} such that $E(u_{iT}|z_{it}) = E(c_i|z_{it})$. Therefore:

$$Cov(z_{it}^{stat} - z_{is}^{stat}, u_{iT}) = Cov(z_{it}^{stat} - z_{is}^{stat}, c_i) \tag{2.4.13}$$

So for $Cov(z_{it}^{stat}, u_{iT})$ to be time constant, we need the change in z_{it}^{stat} over time to be uncorrelated with the time constant part of the unobserved effects.

This will be satisfied for instance if z_{it}^{stat} is composed of a deterministic time component f_t, a time constant component that is arbitrarily correlated with c_i, denote it d_i, and a time varying component that is uncorrelated with c_i, denote it ε_{it}:

$$z_{it}^{stat} = f_t + d_i + \varepsilon_{it} \tag{2.4.14}$$

Indeed in this case,

$$Cov(z_{it}^{stat} - z_{is}^{stat}, u_{iT}) = Cov(z_{it}^{stat} - z_{is}^{stat}, c_i) \tag{2.4.15}$$

$$= Cov(\varepsilon_{it} - \varepsilon_{is}, c_i) \tag{2.4.16}$$

$$= 0 \tag{2.4.17}$$

As long as $Cov(z_{it}^{stat}, u_{iT})$ is time constant, we can use for estimation the following additional moment conditions:

$$E((\tilde{z}_{it}^{stat} - \tilde{z}_{it-1}^{stat})\frac{y_{it} - h_0(x_{it}, \beta_0)}{h_1(x_{it}, \beta_0)}) = 0 \; \forall t = 2, ..., T \tag{2.4.18}$$

where $\tilde{z}_{it}^{stat} = z_{it}^{stat} - E(z_{it}^{stat})$. For estimation, $E(z_{it}^{stat})$ can be simply replaced by the sample average of z_{it}^{stat}. The asymptotic variance of the estimator of β_0 will not be affected by this preliminary estimation, following the results in Newey and McFadden (1994) for instance.

A simple informal test of whether the change in z_{it}^{stat} over time is uncorrelated with the time constant part of the unobserved effects could be to regress the change in z_{it}^{stat} over time on time period dummies and as many time constant explanatory variables as available and test the joint significance of the time constant covariates in the regression.

2.4.2 Serially Uncorrelated Transitory Shocks

In some applications it might be unlikely that instruments or functions of the instruments have a time constant covariance with unobserved effects. For instance consider the case of the linear feedback model where only time period dummy variables are used as covariates so that $x_{it} = D_t$. Then $E(y_{it}|y_{it-1},...,y_{i0},c_i) = \gamma_0 y_{it-1} + \mu_t c_i$ where μ_t is a deterministic constant that depends on t. Even if we assume that y_{i0} does not deviate from the stationary process $s_{i0} = \sum_{s=1}^{\infty} \gamma_0^s c_i \mu_{-s} \varepsilon_{i-s}$, $E(y_{i1} - y_{i0}|c_i) = \sum_{s=1}^{\infty} \gamma_0^s c_i (\mu_{-s+1} - \mu_{-s})$ so that in general $y_{it} - y_{it-1}$ will be correlated with c_i and therefore $y_{it-1} - y_{it-2}$ can not be used as an instrument for the equation in level even if it is time demeaned.

However in such cases other additional restrictions might be available that would come from restrictions on the variance covariance matrix of $u_i = [u_{i1},...,u_{iT}]'$. It is sometimes plausible to assume that the only source of serial correlation in the unobserved effects is time constant unobserved effects so that $Cov(u_{it},u_{is}) = Cov(u_{iq},u_{ir}) \; \forall s < t, q < r$[4]. In general such restrictions imply $T \times (T-1)/2 - 1$ additional overidentifying moment restrictions which can be written as:

$$E\left(\frac{y_{it} - h_0(x_{it},\beta_0)}{h_1(x_{it},\beta_0)} \frac{y_{is} - h_0(x_{is},\beta_0)}{h_1(x_{is},\beta_0)}\right) = \tau_0 \; \forall t,s = 1,...,T, \; s < t \tag{2.4.19}$$

where τ_0 is an additional parameter added to β_0 defined by $\tau_0 = Cov(u_{it},u_{is}) + E(u_{it})E(u_{is}) \; \forall t \neq s$ which doesn't depend on t or s since $E(u_{it})$ is constant by (2.2.2). This is however not true in the case of dynamic models since then some of these moment conditions are already implied by (2.2.1) and (2.2.2). For dynamic models, $u_{it} = (y_{it} - h_0(x_{it},\beta_0))/h_1(x_{it},\beta_0)$ and $y_{it-1}, x_{it} \in z_{it}$ so $u_{it-1} \in z_{it}$. Hence (2.2.1) and

[4] One could also consider weaker additional restrictions of the type $E(u_{it}u_{it-s}) = \tau(s)$ so that serial correlation in the unobserved effects only depends on the number of lags s separating these unobserved effects and not on the chosen time period t. We do not consider this possibility in this paper for simplicity, it would be straightforward to modify the derivations of this section to consider this case.

(2.2.2) imply $E(u_{it}u_{is}) = E(u_{it}u_{ir})\ \forall t < s,r$ so that $Cov(u_{it}u_{is}) = Cov(u_{it}u_{ir})\ \forall t < s,r$. Therefore assuming $Cov(u_{it},u_{is}) = Cov(u_{iq},u_{ir})\ \forall t < s, q < r$ in the case of dynamic models will only imply the additional $T-2$ over-identifying restrictions:

$$E\left(\frac{y_{iT} - h_0(x_{iT},\beta_0)}{h_1(x_{iT},\beta_0)}\frac{y_{iT-s} - h_0(x_{iT-s},\beta_0)}{h_1(x_{iT-s},\beta_0)}\right) = \tau_0\ \forall s = 1,...,T-1 \tag{2.4.20}$$

These moment restrictions are the generalization of the additional moment conditions derived for linear dynamic models in Ahn and Schmidt (1995).

In the case of dynamic models, an assumption such as $Cov(u_{it},u_{is}) = Cov(u_{iq},u_{ir})\ \forall t \neq s, q \neq r$ can be very plausible since it is possible to argue that modeling dynamics will also account for all serial correlation in unobserved effects other than serial correlation due to the time constant part of unobserved effects. For instance, the linear feedback model presented in (2.4.4) implies such additional moment restrictions even though they have not been exploited for estimation in previous studies. Indeed from 2.4.4[5]:

$$E\left(\frac{y_{iT} - \gamma_0 y_{iT-1}}{\mu(x_{iT},\theta_0)}\frac{y_{iT-s} - \gamma_0 y_{iT-s-1}}{\mu(x_{iT-s},\theta_0)}\right) = E(c_i^2)\ s = 1,...,T-1 \tag{2.4.21}$$

Windmeijer (2000) has also derived similar moment conditions for the model presented in (2.2.6) and (2.2.7) but under a set of assumptions that was too weak. Windmeijer (2000) only assumes that c_i^2 is uncorrelated with ε_{it} and that ε_{it} is uncorrelated with ε_{is} for $t \neq s$ which doesn't imply $E(c_i^2\varepsilon_{it}\varepsilon_{is}) = E(c_i^2)E(\varepsilon_{it})E(\varepsilon_{is}) = E(c_i^2)$ hence it seems that a specification of models in terms of conditional expectations and unobserved effects as in (2.2.1) and (2.2.2) is more straightforward than the specification of the model in terms of uncorrelation found in Windmeijer (2000).

[5]I was made aware after finishing a draft of this paper that, in unpublished work, Kitazawa (2007) also considers similar moment conditions for the LFM. Note however that the LFM is only one of the special cases considered by the group of models we defined.

2.5 Monte Carlo Evidence

To study the small sample performance of the estimators we present in this paper, we consider estimating the Linear Feedback model presented in Blundell et al. (2002):

$$y_{it} \sim Poisson(\gamma y_{it-1} + exp(\beta x_{it} + \eta_i)) \; \forall t = 1, ..., T \tag{2.5.1}$$

$$x_{it} = \rho x_{it-1} + \tau \eta_i + \varepsilon_{it} \tag{2.5.2}$$

$$x_{i0} = \frac{\tau}{1-\rho}\eta_i + \xi_i \tag{2.5.3}$$

$$y_{i0} \sim Poisson(\frac{exp(\beta x_{i0} + \eta_i)}{1-\gamma}) \tag{2.5.4}$$

$$\eta_i \sim N(0, \sigma_\eta^2) \tag{2.5.5}$$

$$\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2) \tag{2.5.6}$$

$$\xi_i \sim N(0, \frac{\sigma_\varepsilon^2}{1-\rho^2}) \tag{2.5.7}$$

The only difference with the data generating process of Blundell et al. (2002) is that we do not obtain y_{i0} as the last draw from fifty draws starting at $y_{i-49} \sim Poisson(exp(\beta x_{i-49} + \eta_i))$ but instead impose $E(y_{i0}|c_i) = c_i E(exp(\beta x_{i0}))/(1-\gamma)$. Since we will restrict our attention to $\gamma < 1$, both data generating processes will be very similar even though not exactly equivalent.

With this model, we will consider using for estimation the sequence of moment conditions:

$$E(z_{it}(\frac{y_{it} - \gamma_0 y_{it-1}}{h_1(x_{it}, \beta_0)} - \frac{y_{it-1} - \gamma_0 y_{it-2}}{h_1(x_{it-1}, \beta_0)})) = 0 \; t = 2, ..., T \tag{2.5.8}$$

where $z_{it} = (y_{it-2}, ..., y_{i0}, x_{it-1}, ..., x_{i1})$ or $z_{it} = (y_{it-2}, x_{it-1})$.

The additional conditions that arise from the restriction imposed on the initial conditions are:

$$E((y_{it-1} - y_{it-2})\frac{y_{it} - \gamma_0 y_{it-1}}{h_1(x_{it}, \beta_0)}) = 0 \; t = 2, ..., T \tag{2.5.9}$$

$$E((x_{it} - x_{it-1})\frac{y_{it} - \gamma_0 y_{it-1}}{h_1(x_{it}, \beta_0)}) = 0 \; t = 2, ..., T \tag{2.5.10}$$

The additional conditions that arise from serial uncorrelation of the transitory shocks are:

$$E(\frac{y_{iT} - \gamma_0 y_{iT-1}}{h_1(x_{iT}, \beta_0)}(\frac{y_{it} - \gamma_0 y_{it-1}}{h_1(x_{it}, \beta_0)} - \frac{y_{it-1} - \gamma_0 y_{it-2}}{h_1(x_{it-1}, \beta_0)})) = 0 \; t = 2, ..., T-1 \tag{2.5.11}$$

We will consider four groups of estimators: using no additional moment conditions, using the additional moment conditions from the restrictions on the initial conditions, using the additional moment conditions

from serially uncorrelated transitory shocks, and using both sets of additional moment conditions. Within each group we will consider the GMM estimator that uses all available lags of the instruments for the conditional moment conditions and the GMM estimator that uses only one lag of the instruments for the conditional moment conditions. For each estimator we will also consider the two-step GMM estimator with the identity matrix as initial weighting matrix and the iterated GMM estimator, which is a multiple step GMM estimator that takes as many steps as are needed for the estimates to converge[6]. Therefore we will be considering a total of sixteen estimators.

Table 2.1 and Table 2.2 report the bias and root Mean Squared Error (MSE) of the estimators of γ[7]. Table 2.3 and Table 2.4 report the ratio of the mean of standard errors of the estimators of γ over the standard deviations of these estimators. Therefore these tables capture the bias in the estimators of the variance of the estimators of γ. Table 2.5 and Table 2.6 report the coverage rate of the 95% confidence intervals created from the estimators of γ and their associated standard errors. All results are from 1,000 replications.

The first conclusion from Table 2.1 and Table 2.2 is that using the additional moment conditions presented in the previous section results in large efficiency gains with very sizable decreases in both bias and standard deviations. This gain is especially noticeable when either set of additional moment conditions is used compared to not using any set of additional moment conditions. The addition of a second set of additional moment conditions causes a more modest gain in efficiency. Bias is almost always smaller when using only one lag of the instruments instead of all available lags. When all available lags of the instruments are used, iterated GMM seems to perform better than two step GMM.

Table 2.3 and Table 2.4 show a severe downward bias in standard errors for small n and large T when all available lags of the instruments are used. This problem is alleviated by using iterated GMM, particularly when T is large. However, even using iterated GMM can result in standard deviations being significantly underestimated. This bias in standard errors is due to the use of many over-identifying moment conditions. The same problem of downward biased standard errors has been studied for the special case of linear models in Windmeijer (2005) and for models of count data in Windmeijer (2008). However these two papers concentrate on the bias originating from using a preliminary estimator to compute the optimal weighting matrix,

[6]We do not present the results of iterated GMM estimation for $n = 100$ because for this small sample size and with the convergence criterion we used for the other sample sizes, the iterated GMM algorithm failed to converge in less than 400 iterations in 25% of the bootstrap draws when $T = 4$ and 50% of the bootstrap draws when $T = 8$. Conditional on having the iterated GMM algorithm converging for $n = 100$, using iterated GMM instead of two step GMM seemed to provide some efficiency gain and significantly better inference when many moment conditions are used in a similar way as for larger sample sizes.

[7]We only show results for the estimation of γ here but results for estimation of β exhibit similar patterns.

whereas we see that using iterated GMM instead of two step GMM helps but does not solve completely the problem of downward biased standard errors. Asymptotic analysis under many moment conditions performed in separate work in progress seems to indicate that most of the bias comes from the correlation between the gradient of the moment functions and the moment functions themselves. This result has been presented in a more general setting in Newey and Windmeijer (2009). Bootstrapped standard errors might also be a solution.

Table 2.5 and Table 2.6 show the effect of both downward biased standard errors and bias in the estimator of γ on inference. For small n or large T the coverage of confidence intervals is significantly lower than the confidence level of 95%, particularly when all available lags of the instruments are used. This problem is alleviated by using iterated GMM but not completely solved. Corrected standard errors should participate in constructing better confidence intervals as could bias correction, particularly in the case where no additional moment condition is available. Similarly as for correction of the standard errors, bias correction could be based on higher order asymptotic analysis.

The first conclusion of this section is that using additional restrictions on the stationarity of the instruments or serial uncorrelation of transitory shocks can make a big difference in terms of the precision of the point estimates. It does not solve however the problem of inference which was already present with previous estimators and is due to the poor properties of GMM standard errors in cases where many over-identifying conditions are used.

Using iterated GMM can improve the quality of inference compared to two step GMM especially when T is relatively large without solving the problem completely. The results presented in this section also suggest that using only one lag of the instruments can result in much better inference especially when T is relatively large. Previous studies of instrumental variable estimation of models similar to the ones we consider in this paper, such as Arellano and Bond (1991) or Windmeijer (2008), recommended the use of all lags of the instruments in (2.5.8). However the Monte Carlo evidence we presented indicates that using only one lag of the instruments causes only a modest loss in accuracy, especially when additional moment conditions are available, but results in significantly lower bias and significantly better inference compared to using all available lags of the instruments.

Table 2.1: Bias and RMSE for estimating γ, $T = 4$

		$N = 100$		$N = 500$		$N = 1000$		$N = 2000$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Two step GMM									
no additional conditions	All Lags	−0.214	0.314	−0.070	0.117	−0.039	0.076	−0.022	0.053
	One Lag	−0.186	0.320	−0.059	0.137	−0.031	0.088	−0.019	0.065
initial conditions	All Lags	−0.036	0.167	−0.008	0.068	0.001	0.046	−0.001	0.033
	One Lag	−0.006	0.142	−0.001	0.067	0.003	0.047	−0.000	0.034
serial uncorrelation	All Lags	−0.108	0.196	−0.034	0.072	−0.015	0.046	−0.008	0.033
	One Lag	−0.070	0.168	−0.022	0.073	−0.007	0.049	−0.005	0.036
both sets of conditions	All Lags	−0.029	0.166	−0.007	0.070	−0.001	0.048	−0.002	0.033
	One Lag	−0.006	0.135	−0.003	0.064	0.002	0.043	−0.001	0.031
Iterated GMM									
no additional conditions	All Lags			−0.063	0.105	−0.036	0.074	−0.021	0.051
	One Lag			−0.060	0.132	−0.029	0.088	−0.018	0.063
initial conditions	All Lags			−0.001	0.062	0.004	0.046	−0.001	0.032
	One Lag			0.003	0.065	0.004	0.047	−0.000	0.034
serial uncorrelation	All Lags			−0.023	0.058	−0.010	0.043	−0.007	0.031
	One Lag			−0.017	0.061	−0.006	0.047	−0.004	0.034
both sets of conditions	All Lags			−0.010	0.086	−0.001	0.042	−0.002	0.029
	One Lag			−0.003	0.061	−0.000	0.043	−0.002	0.030

The values of parameters used for the simulations are: $\gamma = 0.5$, $\beta = 0.5$,$\rho = 0.5$, $\tau = 0.1$, $\sigma_\eta^2 = 0.5$, $\sigma_\varepsilon^2 = 0.5$

Table 2.2: Bias and RMSE for estimating γ, $T = 8$

		$N = 100$		$N = 500$		$N = 1000$		$N = 2000$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Two step GMM									
no additional conditions	All Lags	−0.244	0.319	−0.070	0.092	−0.037	0.050	−0.018	0.028
	One Lag	−0.126	0.183	−0.033	0.061	−0.019	0.041	−0.010	0.028
initial conditions	All Lags	−0.122	0.204	−0.030	0.061	−0.014	0.032	−0.005	0.017
	One Lag	−0.013	0.105	−0.002	0.041	−0.001	0.029	−0.000	0.020
serial uncorrelation	All Lags	−0.191	0.267	−0.052	0.075	−0.025	0.039	−0.011	0.020
	One Lag	−0.068	0.128	−0.017	0.041	−0.010	0.027	−0.005	0.018
both sets of conditions	All Lags	−0.107	0.192	−0.026	0.060	−0.012	0.032	−0.005	0.017
	One Lag	−0.002	0.101	−0.002	0.038	−0.002	0.026	−0.001	0.018
Iterated GMM									
no additional conditions	All Lags			−0.044	0.058	−0.026	0.038	−0.015	0.024
	One Lag			−0.030	0.057	−0.017	0.039	−0.009	0.027
initial conditions	All Lags			−0.009	0.035	−0.005	0.023	−0.002	0.015
	One Lag			−0.001	0.041	−0.001	0.029	−0.000	0.020
serial uncorrelation	All Lags			−0.019	0.034	−0.011	0.023	−0.007	0.015
	One Lag			−0.011	0.036	−0.008	0.026	−0.004	0.018
both sets of conditions	All Lags			−0.010	0.039	−0.006	0.022	−0.003	0.014
	One Lag			−0.004	0.037	−0.003	0.025	−0.002	0.018

The values of parameters used for the simulations are: $\gamma = 0.5$, $\beta = 0.5$,$\rho = 0.5$, $\tau = 0.1$, $\sigma_\eta^2 = 0.5$, $\sigma_\varepsilon^2 = 0.5$

Table 2.3: Ratio of standard errors over standard deviations of estimators of γ, $T = 4$

		$N = 100$	$N = 500$	$N = 1000$	$N = 2000$
Two step GMM					
no additional conditions	All Lags	0.612	0.929	1.033	1.029
	One Lag	0.772	0.883	1.025	0.967
initial conditions	All Lags	0.428	0.763	0.870	0.921
	One Lag	0.556	0.826	0.908	0.945
serial uncorrelation	All Lags	0.562	0.900	0.990	0.975
	One Lag	0.718	0.921	0.980	0.958
both sets of conditions	All Lags	0.368	0.724	0.844	0.887
	One Lag	0.520	0.794	0.914	0.939
Iterated GMM					
no additional conditions	All Lags		0.971	0.999	1.039
	One Lag		0.902	0.990	0.986
initial conditions	All Lags		0.758	0.827	0.922
	One Lag		0.809	0.899	0.939
serial uncorrelation	All Lags		0.976	0.971	0.997
	One Lag		1.022	0.983	0.976
both sets of conditions	All Lags		0.496	0.829	0.907
	One Lag		0.774	0.877	0.943

The values of parameters used for the simulations are: $\gamma = 0.5$, $\beta = 0.5$,$\rho = 0.5$, $\tau = 0.1$, $\sigma_\eta^2 = 0.5$, $\sigma_\varepsilon^2 = 0.5$

Table 2.4: Ratio of standard errors over standard deviations of estimators of γ, $T = 8$

		$N = 100$	$N = 500$	$N = 1000$	$N = 2000$
Two step GMM					
no additional conditions	All Lags	0.146	0.540	0.734	0.914
	One Lag	0.599	0.898	0.948	0.948
initial conditions	All Lags	0.081	0.403	0.615	0.837
	One Lag	0.353	0.736	0.834	0.899
serial uncorrelation	All Lags	0.099	0.415	0.591	0.832
	One Lag	0.435	0.823	0.907	0.955
both sets of conditions	All Lags	0.055	0.354	0.555	0.803
	One Lag	0.290	0.696	0.822	0.899
Iterated GMM					
no additional conditions	All Lags		0.749	0.854	0.942
	One Lag		0.915	0.954	0.963
initial conditions	All Lags		0.561	0.727	0.849
	One Lag		0.713	0.814	0.891
serial uncorrelation	All Lags		0.695	0.803	0.923
	One Lag		0.855	0.924	0.976
both sets of conditions	All Lags		0.440	0.683	0.848
	One Lag		0.684	0.823	0.897

The values of parameters used for the simulations are: $\gamma = 0.5$, $\beta = 0.5$,$\rho = 0.5$, $\tau = 0.1$, $\sigma_\eta^2 = 0.5$, $\sigma_\varepsilon^2 = 0.5$

2.6 Average Partial Effects

With multiplicative heterogeneity models, Average Partial Effects (APE) are very simple to compute. Average Partial Effects are defined by:

$$APE_{f_W} = E_{f_W}(\frac{\partial y}{\partial x}) \tag{2.6.1}$$

where f_W is some distribution over the domain of w which represents all the information observed for one observation and $\frac{\partial y}{\partial x}$ denotes the change in y caused by a small change in x.[8] Eliminating the subscripts,

[8]Here we use the notation for partial derivatives but in case of discrete changes in x of $\Delta_x$we could use the counterfactuals notation and use $\Delta_y(x) = y|(x+\Delta_x) - y|x$ interchangeably.

Table 2.5: Coverage of 95% confidence intervals for γ, $T = 4$

		$N = 100$	$N = 500$	$N = 1000$	$N = 2000$
Two step GMM					
no additional conditions	All Lags	0.628	0.835	0.895	0.924
	One Lag	0.788	0.875	0.930	0.929
initial conditions	All Lags	0.614	0.855	0.903	0.923
	One Lag	0.717	0.887	0.927	0.934
serial uncorrelation	All Lags	0.628	0.847	0.910	0.921
	One Lag	0.774	0.894	0.930	0.940
both sets of conditions	All Lags	0.538	0.832	0.906	0.918
	One Lag	0.699	0.868	0.919	0.936
Iterated GMM					
no additional conditions	All Lags		0.849	0.893	0.931
	One Lag		0.879	0.914	0.930
initial conditions	All Lags		0.842	0.891	0.921
	One Lag		0.874	0.918	0.926
serial uncorrelation	All Lags		0.884	0.918	0.933
	One Lag		0.911	0.934	0.951
both sets of conditions	All Lags		0.808	0.891	0.929
	One Lag		0.863	0.902	0.933

The values of parameters used for the simulations are: $\gamma = 0.5$, $\beta = 0.5$,$\rho = 0.5$, $\tau = 0.1$, $\sigma_\eta^2 = 0.5$, $\sigma_\varepsilon^2 = 0.5$

(2.2.1) can be written as:

$$y = h_0(x,\beta_0) + h_1(x,\beta_0)u \tag{2.6.2}$$

Therefore:

$$APE_{f_w} = E_{f_w}\left(\frac{\partial h_0(x,\beta_0)}{\partial x} + \frac{\partial h_1(x,\beta_0)}{\partial x}u\right) \tag{2.6.3}$$

$$= E_{f_w}\left(\frac{\partial h_0(x,\beta_0)}{\partial x} + \frac{\partial h_1(x,\beta_0)}{\partial x}\frac{y - h_0(x,\beta_0)}{h_1(x,\beta_0)}\right) \tag{2.6.4}$$

For notational simplicity in this section we will consider $h_0(.,.) = 0$ but this will not affect any of the results.

Table 2.6: Coverage of 95% confidence intervals for γ, $T = 8$

		$N = 100$	$N = 500$	$N = 1000$	$N = 2000$
Two step GMM					
no additional conditions	All Lags	0.098	0.434	0.649	0.821
	One Lag	0.590	0.862	0.898	0.913
initial conditions	All Lags	0.091	0.534	0.757	0.875
	One Lag	0.522	0.846	0.902	0.923
serial uncorrelation	All Lags	0.086	0.410	0.637	0.820
	One Lag	0.573	0.853	0.895	0.926
both sets of conditions	All Lags	0.066	0.501	0.724	0.863
	One Lag	0.457	0.832	0.902	0.924
Iterated GMM					
no additional conditions	All Lags		0.620	0.753	0.862
	One Lag		0.875	0.904	0.920
initial conditions	All Lags		0.729	0.834	0.895
	One Lag		0.836	0.888	0.920
serial uncorrelation	All Lags		0.731	0.822	0.897
	One Lag		0.876	0.911	0.933
both sets of conditions	All Lags		0.697	0.820	0.892
	One Lag		0.823	0.892	0.923

The values of parameters used for the simulations are: $\gamma = 0.5$, $\beta = 0.5$,$\rho = 0.5$, $\tau = 0.1$, $\sigma_\eta^2 = 0.5$, $\sigma_\varepsilon^2 = 0.5$

Many applications are interested in the average effect across an observed subset of the population, denote it A. This corresponds to using $f_w = f(w|A)$ so that $APE_A = E(\frac{\partial y}{\partial x}|A) = E(\frac{\partial h_1(x,\beta_0)}{\partial x}\frac{y}{h_1(x,\beta_0)}|A)$.

For instance we could be interested in the average effect of x on y across the entire population in some given time period t $APE_t = E(\frac{\partial h_1(x_{it},\beta_0)}{\partial x}\frac{y_{it}}{h_1(x_{it},\beta_0)})$. Or in the case of a binary explanatory variable x^1, with $x = (x^1, x^{-1})$, we could be interested in Average Treatment Effect on the Treated at a given time period:

$$ATET_t = E(y(1,x_{it}^{-1}) - y(0,x_{it}^{-1})|x_{it}^1 = 1) \tag{2.6.5}$$

$$= E((h_1((1,x_{it}^{-1}),\beta_0) - h_1((0,x_{it}^{-1}),\beta_0))\frac{y_{it}}{h_1(x_{it},\beta_0)}|x_{it}^1 = 1) \tag{2.6.6}$$

Estimation and inference are straightforward in this case once a consistent estimator $\hat{\beta}$ for β_0 is defined. Since $E(1(i \in A)(\frac{\partial h_1(x_{it},\beta_0)}{\partial x}\frac{y_{it}}{h_1(x_{it},\beta_0)} - APE_{At})) = 0$ where $1(.)$ is the indicative function, we can just

add this moment condition to the moment conditions used to estimate β_0 and obtain an additional estimator of APE_{At} as well an estimator for the asymptotic variance of $\hat{APE}_{At}$ and covariance between $\hat{APE}_{At}$ and $\hat{\beta}$ where $\hat{\beta}$ denotes the estimator of β_0 we will be using.

Since we are adding one moment condition for one new parameter, the estimator $\hat{\beta}$ will not be affected by estimation of Average Partial Effects. In addition the GMM estimator of APE will be given by:

$$\hat{APE}_{At} = \frac{1}{n_A} \sum_{i=1}^{n} 1(i \in A) \frac{\partial h_1(x_{it}, \hat{\beta})}{\partial x} \frac{y_{it}}{h_1(x_{it}, \hat{\beta})} \tag{2.6.7}$$

Where $n_A = \sum_{i=1}^{n} 1(i \in A)$.

If we think that APE should be equal across time periods, we can impose this restriction in the GMM estimation by adding the moment restrictions $\{E(1(i \in A)(\frac{\partial h_1(x_{it}, \beta_0)}{\partial x} \frac{y_{it}}{h_1(x_{it}, \beta_0)} - APE_A)) = 0\}_{t=1,...,T}$ which might affect estimation of β_0 or we can estimate average partial effects for each time period and combine them using Minimum Distance Estimation which will not affect estimation of β_0.

In other situations, if f_w can be consistently estimated by $f_w(\hat{\eta})$ where $\hat{\eta}$ is a vector of estimators of nuisance parameters η_0, then:

$$\hat{APE}_{f_w} = E_{f_w(\hat{\eta})}(\frac{\partial h_1(x, \hat{\beta})}{\partial x} \frac{y}{h_1(x, \hat{\beta})}) \tag{2.6.8}$$

is consistent for APE_{f_w}. If $(\hat{\beta}, \hat{\eta})$ are jointly asymptotically normal and a consistent estimator of their asymptotic variance-covariance matrix is available then inference can be performed using the delta-method.

2.7 Conclusion

These results hopefully provide useful new options for researchers who wish to use non-linear models of panel data with unobserved effects in applications where only sequential exogeneity is available. The problem of weak instrumental variables seems to be mitigated significantly by the use of additional moment conditions originating from additional restrictions of stationarity of the instruments or serial uncorrelation of the transitory shocks. Monte Carlo evidence also seems to suggest that it is preferable to use only one or a few lags of the instruments compared to all available lags since this results in much better inference at the expense of only small losses in efficiency.

Two directions are available in order to obtain estimators with better inference. One consists in studying the higher order properties of the GMM estimator with many over-identifying restrictions, the other consists in finding good exactly identifying moment conditions. Both of these approaches are left for future research.

CHAPTER 3

EFFICIENCY OF THE POISSON FIXED EFFECTS ESTIMATOR

3.1 Introduction

A commonly used estimator for models of count panel data with multiplicative heterogeneity and strictly exogenous explanatory variables is the Poisson fixed effects (PFE) estimator introduced by Hausman et al. (1984). This estimator is a conditional maximum likelihood estimator which takes advantage of the assumptions of Poisson distribution and independent draws over time to derive a conditional distribution of the dependent variable that does not depend on the distribution of unobserved heterogeneity. In many applications, these distributional assumptions are likely to be violated. Wooldridge (1999) showed that the PFE estimator is consistent as long as the restriction on the conditional mean function is correctly specified, independently of whether the rest of the assumptions of the PFE model hold.

In this paper I show that, as long as the conditional mean of the dependent variable is equal to its conditional variance and the conditional serial correlation of the dependent variable is zero, the PFE estimator is also asymptotically efficient in the class of estimators that are consistent under restrictions on the conditional mean function. I then define another estimator that is asymptotically efficient in the same class of estimators under more general conditions.

In Section 3.2, I present the model considered in this paper and study the asymptotically efficient estimator for this model. I show under which conditions the PFE estimator is asymptotically efficient and propose an alternative estimator that is asymptotically efficient under more general conditions. In Section 3.3, I use Monte Carlo simulations to investigate the small sample properties of the PFE estimator and of this new estimator.

3.2 The Model and Estimators

As in Wooldridge (1999), we consider panel data models that specify a conditional mean function with strictly exogenous explanatory variables and multiplicative heterogeneity:

$$E(y_{it}|c_i,x_i) = c_i\mu(x_{it},\beta_0)\ \forall\, i = 1,...,n, t = 1,...,T \tag{3.2.1}$$

where i indexes cross-sectional observations, t indexes time, and $x_i = \{x_{i1}, ..., x_{iT}\}$. This model is also a special case of the random coefficients model presented in Section 4 of Chamberlain (1992b). Throughout this paper we consider the case of i.i.d. cross-sectional draws and large n, fixed T asymptotics. Denote $\mu_{it}(\beta) = \mu(x_{it}, \beta)$ and $\mu_{it} = \mu_{it}(\beta_0)$.

Wooldridge (1999) showed that the parameters in this model can be estimated from the conditional moment conditions:

$$E(\rho_{it}(\beta_0)|x_i) = 0 \; \forall t = 1, ..., T \tag{3.2.2}$$

where $\rho_{it}(\beta) = y_{it} - \mu(x_{it}, \beta)\frac{\sum_{t=1}^{T} y_{it}}{\sum_{s=1}^{T} \mu(x_{is}, \beta)}$. Under (3.2.1), for any deterministic functions $g_t(.)$, the following unconditional moment conditions will hold:

$$E(g_t(x_{i1}, ..., x_{iT})\rho_{it}(\beta_0)) = 0 \; \forall t = 1, ..., T \tag{3.2.3}$$

Therefore, under standard regularity conditions, any estimator $\hat{\beta}$ of β_0 defined by:

$$\sum_{i=1}^{n} \sum_{t=1}^{T} g_t(x_{i1}, ..., x_{iT})\rho_{it}(\hat{\beta}) = 0 \tag{3.2.4}$$

will be consistent for β_0 and asymptotically normal. All of the estimators considered in this paper can be written as (3.2.4) so that they are consistent as long as (3.2.1) holds, independently of what other assumptions are considered to study efficiency.

3.2.1 Asymptotically Efficient Estimation

The conditional moment conditions written in (3.2.2) can be rewritten in the form:

$$E(\rho_i(\beta_0)|x_i) = 0 \tag{3.2.5}$$

where $\rho_i(\beta) = [\rho_{i1}(\beta), ..., \rho_{iT}(\beta)]'$. Similarly as in Chamberlain (1987), an optimal estimator for β_0 from (3.2.5) can be postulated to be $\hat{\beta}_{opt}$ from:

$$\sum_{i=1}^{n} D_i' \Sigma_i^+ \rho_i(\hat{\beta}_{opt}) = 0 \tag{3.2.6}$$

where $D_i = E(\frac{\partial \rho_i}{\partial \beta'}(\beta_0)|x_i)$, $\Sigma_i = Var(\rho_i(\beta_0)|x_i)$ and Σ_i^+ is some generalized inverse of Σ_i.[1] If $\hat{\beta}_{opt}$ is indeed optimal, $D_i' \Sigma_i^+$ can be called the optimal instruments for the vector of moment functions $\rho_i(\beta)$.

[1] That $\hat{\beta}_{opt}$ is optimal for estimating β_0 from (3.2.5) has to be proven since Chamberlain (1987) considers cases where $Var(\rho_i(\beta_0)|x_i)$ is non-singular a.s., but in our case $Var(\rho_i(\beta_0)|x_i)$ can be shown to be non-invertible.

Under standard regularity conditions:

$$\sqrt{n}(\hat{\beta}_{opt} - \beta_0) \xrightarrow{d} N(0, V_{opt})$$
$$V_{opt} = E(D_i' \Sigma_i^+ D_i)^{-1} E(D_i' \Sigma_i^+ \Sigma_i \Sigma_i^+ D_i) E(D_i' \Sigma_i^+ D_i)^{-1}$$
$$= E(D_i' \Sigma_i^+ D_i)^{-1}$$

Appendix C.1 shows that, when a specific generalized inverse of Σ_i denoted Σ_i^- is used, $\hat{\beta}_{opt}$ is asymptotically efficient for estimating β_0 from (3.2.1) by showing that V_{opt} is equal to the inverse of the asymptotic information bound for estimating β_0 from (3.2.1) derived in Chamberlain (1992b).[2]

3.2.2 Conditions for Efficiency of the Poisson FE estimator

As shown in Wooldridge (1999), the Poisson fixed effects estimator, $\hat{\beta}_{PFE}$, is defined by:

$$\sum_{i=1}^{n} \left(\frac{\partial p_i(\hat{\beta}_{PFE})}{\partial \beta'}\right)' W_i(\hat{\beta}_{PFE})^{-1} \rho_i(\hat{\beta}_{PFE}) = 0 \tag{3.2.7}$$

where $p_i(\beta)' = [p_{i1}(\beta), ..., p_{iT}(\beta)]$, $p_{it}(\beta) = \frac{\mu_{it}(\beta)}{\sum_{s=1}^{T} \mu_{is}(\beta)}$, $W_i(\beta) = diag(p_i(\beta))$, and $diag(a)$ is the diagonal matrix with a for diagonal.

Under standard regularity conditions, $\hat{\beta}_{PFE}$ is asymptotically equivalent to $\tilde{\beta}_{PFE}$ defined by:

$$\sum_{i=1}^{n} \left(\frac{\partial p_i(\beta_0)}{\partial \beta'}\right)' W_i(\beta_0)^{-1} \rho_i(\tilde{\beta}_{PFE}) = 0 \tag{3.2.8}$$

We show in Appendix C.2 that $D_i' \Sigma_i^- = -\left(\frac{\partial p_i(\beta_0)}{\partial \beta'}\right)' W_i(\beta_0)^{-1}$ if (3.2.1) holds as well as:

$$Var(y_{it} | c_i, x_i) = c_i \mu_{it} \tag{3.2.9}$$

$$Cov(y_{it}, y_{it-s} | c_i, x_i) = 0 \; \forall s = 1, ..., t \tag{3.2.10}$$

Therefore under these additional conditions, the PFE estimator is using the optimal instruments for $\rho_i(\beta)$ in order to estimate β_0 and is asymptotically efficient in the class of estimators that are consistent under (3.2.1).

[2] A corrollary of that result is that $\hat{\beta}_{opt}$ is indeed also optimal for estimating β_0 from (3.2.5) since (3.2.1) implies (3.2.5). Therefore, the optimal estimator from (3.2.5) corresponds to the optimal estimator from (3.2.1), i.e. no information is lost for estimating β_0 from transforming (3.2.1) to (3.2.5).

3.2.3 An Alternative Estimator

In this section we derive an optimal estimator for cases where one thinks there might be overdispersion, so that instead of (3.2.9) we have:

$$Var(y_{it}|c_i,x_i) = c_i\mu_{it} + \theta c_i^2\mu_{it}^2 \tag{3.2.11}$$

where θ is an unknown parameter, and serial correlation so that instead of (3.2.10) we have:

$$Cov(y_{it},y_{it-s}|c_i,x_i) = \gamma c_i^2\mu_{it}\mu_{it-1}\ for s = 1 \tag{3.2.12}$$

$$= 0\ for s > 1 \tag{3.2.13}$$

where γ is an unknown parameter. Note that (3.2.9) and (3.2.10) are a special case of assumptions (3.2.11) and (3.2.12) since both sets of assumptions are the same with $\theta = 0$ and $\gamma = 0$.

Appendix C.3 shows that, as long as $T \geq 3$, consistent estimators of θ and γ can be obtained under the assumptions (3.2.1), (3.2.11) and (3.2.12), denote these estimators $\hat{\theta}$ and $\hat{\gamma}$.

As seen in Section 2.1, the optimal instruments for $\rho_i(\beta)$ are $D_i'\Sigma_i^-$ where:

$$D_i = -E(c_i|x_i)(\sum_{t=1}^{T}\mu_{it})[\frac{\partial p_{it}}{\partial \beta'}]_{t=1,...,T} \tag{3.2.14}$$

and Σ_i^- is a specific generalized inverse of $\Sigma_i = Var(\rho_i|x_i)$.

Without (3.2.9) and (3.2.10), $D_i'\Sigma_i^-$ does depend on $E(c_i|x_i)$ and $Var(c_i|x_i)$. Therefore with assumptions (3.2.11) and (3.2.12) instead of (3.2.9) and (3.2.10), the conditional mean and variance of the unobserved heterogeneity term c_i are needed to compute the optimal instruments. One can model these as known functions h_1 and h_2 of a vector of unknown nuisance parameters η:

$$E(c_i|x_i) = h_1(x_i,\eta) \tag{3.2.15}$$

$$Var(c_i|x_i) = h_2(x_i,\eta) \tag{3.2.16}$$

and estimate η consistently since under (3.2.1), (3.2.11), (3.2.12), (3.2.15) and (3.2.16):

$$E(y_{it}|x_i) = h_1(x_i,\eta)\mu_{it} \tag{3.2.17}$$

$$E(y_{it}^2|x_i) = h_1(x_i,\eta)\mu_{it} + (\theta+1)(h_2(x_i,\eta) + h_1(x_i,\eta)^2)\mu_{it}^2 \tag{3.2.18}$$

Therefore a consistent estimator of η under (3.2.1), (3.2.11), (3.2.12), (3.2.15) and (3.2.16) can be obtained by pooled non-linear regression from (3.2.17) and (3.2.18) with μ_{it} replaced by $\mu_{it}(\ddot{\beta})$ and θ replaced by $\hat{\theta}$, where $\ddot{\beta}$ is a preliminary consistent estimator of β_0. Denote by $\hat{\eta}$ the resulting estimator of η.

The alternative estimator to Poisson fixed effects we propose in this paper is $\hat{\beta}_{alt}$ defined by:

$$\sum_{i=1}^{n} \hat{D}_i \hat{\Sigma}_i^{-} \rho_i(\hat{\beta}_{alt}) = 0 \tag{3.2.19}$$

where

$$\hat{D}_i = h_1(x_i, \hat{\eta})(\sum_{t=1}^{T} \ddot{\mu}_{it})[\frac{\partial p_{it}}{\partial \beta'}(\ddot{\beta})]_{t=1,...,T} \tag{3.2.20}$$

and:

$$\hat{\Sigma}_i^{-} = \hat{\Sigma}_{y,i}^{-1} - \hat{\Sigma}_{y,i}^{-1} \ddot{\mu}_i (\ddot{\mu}_i' \hat{\Sigma}_{y,i}^{-1} \ddot{\mu}_i)^{-1} \ddot{\mu}_i' \hat{\Sigma}_{y,i}^{-1} \tag{3.2.21}$$

where $\ddot{\mu}_{it} = \mu_{it}(\ddot{\beta})$, $\ddot{\mu}_i = \mu_i(\ddot{\beta}) = [\mu_{i1}(\ddot{\beta}), ..., \mu_{iT}(\ddot{\beta})]'$ and the $(t,s)^{th}$ element of $\hat{\Sigma}_{y,i}$ is:

$$\hat{Cov}(y_{it}, y_{is}|x_i) = 1[t=s]h_1(x_i, \hat{\eta})\ddot{\mu}_{it} + (1[|t-s| \leq 1]\theta^{1-|t-s|}\gamma^{|t-s|}(h_2(x_i, \hat{\eta}) + h_1(x_i, \hat{\eta})^2) + h_2(x_i, \hat{\eta}))\ddot{\mu}_{it}\ddot{\mu}_{is}$$

where $\ddot{\beta}$ can simply be defined to be the Poisson fixed effects estimator.

This estimator uses optimal instruments for $\rho_i(\beta)$ and is asymptotically efficient in the class of estimators of β_0 that are consistent under (3.2.1) as long as (3.2.11), (3.2.12), (3.2.15) and (3.2.16) hold. Appendix D shows that when (3.2.9) and (3.2.10) hold, so that (3.2.11) and (3.2.12) hold with $\theta = 0$, $\gamma = 0$, and that $\hat{\beta}_{PFE}$ is asymptotically efficient, $\hat{\beta}_{alt}$ and $\hat{\beta}_{PFE}$ are asymptotically equivalent, independently of whether (3.2.15) and (3.2.16) hold. Therefore, the estimator $\hat{\beta}_{alt}$ is indeed efficient under more general conditions than the Poisson fixed effects estimator.

3.3 Monte Carlo Simulations Study

To compare the small sample perfomance of the Poisson Fixed Effects estimator and the alternative estimator defined by (3.2.19), we use both estimators to estimate β_0 from the data generating process:

$$\begin{aligned}
x_{it} &\overset{i.i.d.}{\sim} Uniform(-a, a) \\
c_i|x_i &\sim F_c(x_i) \\
e_{it} &\overset{i.i.d.}{\sim} Uniform(a_e, b_e) \\
u_{it} &= \delta exp(e_{it-1}) + exp(e_{it}) \\
y_{it} &\overset{i.i.d.}{\sim} Poisson(c_i exp(\beta_0 x_{it}) u_{it})
\end{aligned}$$

where $Uniform(a,b)$ denotes the uniform distribution over the interval (a,b), $Poisson(\mu)$ denotes the Poisson distribution with mean. We set $\beta_0 = 1$. We also set a_e, b_e and δ so that $E(e_{it}) = \frac{1}{1+\delta}$, $Var(e_{it}) = \frac{\theta}{1+\delta^2}$ (so that $E(u_{it}|c_i,x_i) = 1$ and $Var(u_{it}|c_i,x_i) = \theta$) and $Cov(u_{it},u_{it-1}|c_i,x_i) = \delta Var(exp(e_{it})) = \gamma$. Therefore (3.2.1), (3.2.11) and (3.2.12) are satisfied.

Tables 3.1, 3.2 and 3.3 shows the performance of both estimators and of the unfeasible optimal estimator using $D_i\Sigma_i^-$ as instruments for $\rho_i(\beta)$ from Monte Carlo simulations. The results shown are measures of bias, standard deviation, and root MSE for sample sizes $N = 100$, $N = 500$ and $N = 1000$, with ten time periods and for the cases where $\{\theta = 0, \gamma = 0\}$ and $\{\theta = 1, \gamma = 0.5\}$. $F_c(c_i)$ is given by $c_i = exp(\lambda\bar{x}_i + Uniform(-a,a))$ where $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}$. We show results for $\lambda = 0$ and $\lambda = 1$. In both cases, as a increases, the variances of c_i and c_i^2 increase. We show results for $a = 1, 1.5, 2$. The model for the conditional mean and variance of c_i that we use is:

$$h_1(x_i,\eta) = \eta_1$$

$$h_2(x_i,\eta) = \eta_2$$

Therefore this model corresponds to the true data generating process when $\lambda = 0$ but not when $\lambda = 1$.

When $\{\theta = 0, \gamma = 0\}$, both the Poisson fixed effects estimator and our alternative estimator are asymptotically efficient in the class of estimators consistent under (3.2.1), independently of the distribution of c_i. When $\{\theta = 1, \rho = 0.5\}$, the Poisson fixed effects estimator is not asymptotically efficient and our alternative estimator is asymptotically efficient when $\lambda = 0$. When $\{\theta = 1, \rho = 0.5\}$ and $\lambda = 1$, neither the Poisson fixed effects estimator nor our alternative estimator are efficient.

Results for Tables 3.1, 3.2 and 3.3 show that significant gains in efficiency can be achieved by using the unfeasible optimal instruments but that in small samples, the additional noise originated from estimating η_1, η_2, θ and γ to compute a feasible estimator can overpower this gain in efficiency and result in the alternative estimator defined by (3.2.19) being signigicantly less accurate than the Poisson Fixed Effects estimator. The solution to this problem could be to derive a data-based criterion that captures the trade-off between asymptotic efficiency and finite sample noise from nuisance parameters and helps decide between different models of optimal instruments. This is left for future research. The conclusion of this section is that the Poisson Fixed Effects estimator performs well with small sample sizes compared to an alternative estimator that is asymptotically efficient under more general conditions. However with large enough sample sizes, significant gains in efficiency can be obtained from using a more general model of optimal instruments.

Table 3.1: $N = 100$: Bias, standard deviation and root mean squared error

	$c \sim exp(Uniform(-1,1))$			$c \sim exp(Uniform(-1.5,1.5))$			$c \sim exp(Uniform(-2,2))$		
	Bias	sd	rmse	Bias	sd	rmse	Bias	sd	rmse
$\theta = 0, \gamma = 0, \lambda = 0$									
Poisson FE	0.002	0.057	0.057	0.002	0.034	0.034	0.001	0.022	0.022
Feasible alternative	0.002	0.057	0.057	0.004	0.064	0.065	0.001	0.077	0.077
Unfeasible optimal	0.002	0.057	0.057	0.002	0.034	0.034	0.001	0.022	0.022
$\theta = 0, \gamma = 0, \lambda = 1$									
Poisson FE	0.001	0.055	0.055	−0.001	0.033	0.033	0.001	0.021	0.021
Feasible alternative	0.001	0.055	0.055	−0.002	0.051	0.051	−0.004	0.078	0.078
Unfeasible optimal	0.001	0.055	0.055	−0.001	0.033	0.033	0.001	0.021	0.021
$\theta = 1, \gamma = .5, \lambda = 0$									
Poisson FE	0.005	0.086	0.086	0.004	0.066	0.067	0.000	0.063	0.063
Feasible alternative	0.004	0.077	0.078	−0.003	0.102	0.102	−0.002	0.102	0.102
Unfeasible optimal	0.005	0.076	0.077	0.003	0.054	0.054	0.001	0.043	0.043
$\theta = 1, \gamma = .5, \lambda = 1$									
Poisson FE	0.007	0.088	0.089	0.003	0.069	0.069	−0.001	0.067	0.067
Feasible alternative	0.008	0.083	0.083	0.001	0.088	0.088	−0.004	0.128	0.128
Unfeasible optimal	0.007	0.078	0.079	0.004	0.053	0.053	−0.002	0.041	0.041

Table 3.2: $N = 500$: Bias, standard deviation and root mean squared error

	$c \sim exp(Uniform(-1,1))$			$c \sim exp(Uniform(-1.5,1.5))$			$c \sim exp(Uniform(-2,2))$		
	Bias	sd	rmse	Bias	sd	rmse	Bias	sd	rmse
$\theta = 0, \gamma = 0, \lambda = 0$									
Poisson FE	−0.000	0.025	0.025	−0.000	0.015	0.015	−0.000	0.010	0.010
Feasible alternative	−0.000	0.025	0.025	−0.000	0.015	0.015	−0.000	0.012	0.012
Unfeasible optimal	−0.000	0.025	0.025	−0.000	0.015	0.015	−0.000	0.010	0.010
$\theta = 0, \gamma = 0, \lambda = 1$									
Poisson FE	−0.001	0.024	0.024	0.000	0.015	0.015	0.000	0.010	0.010
Feasible alternative	−0.001	0.024	0.024	0.000	0.015	0.015	0.000	0.012	0.012
Unfeasible optimal	−0.001	0.024	0.024	0.000	0.015	0.015	0.000	0.010	0.010
$\theta = 1, \gamma = .5, \lambda = 0$									
Poisson FE	−0.002	0.040	0.040	−0.001	0.031	0.031	−0.001	0.028	0.028
Feasible alternative	−0.003	0.036	0.035	−0.002	0.030	0.030	−0.002	0.023	0.023
Unfeasible optimal	−0.003	0.036	0.035	−0.001	0.025	0.025	−0.001	0.019	0.019
$\theta = 1, \gamma = .5, \lambda = 1$									
Poisson FE	−0.002	0.040	0.040	−0.001	0.032	0.032	−0.000	0.030	0.030
Feasible alternative	−0.003	0.035	0.035	−0.001	0.024	0.024	−0.002	0.030	0.030
Unfeasible optimal	−0.003	0.035	0.035	−0.001	0.024	0.024	−0.000	0.019	0.019

Table 3.3: $N = 1000$: Bias, standard deviation and root mean squared error

	$c \sim exp(Uniform(-1,1))$			$c \sim exp(Uniform(-1.5,1.5))$			$c \sim exp(Uniform(-2,2))$		
	Bias	sd	rmse	Bias	sd	rmse	Bias	sd	rmse
$\theta = 0, \gamma = 0, \lambda = 0$									
Poisson FE	−0.001	0.017	0.017	−0.001	0.011	0.011	−0.000	0.007	0.007
Feasible alternative	−0.001	0.017	0.017	−0.001	0.011	0.011	−0.000	0.007	0.007
Unfeasible optimal	−0.001	0.017	0.017	−0.001	0.011	0.011	−0.000	0.007	0.007
$\theta = 0, \gamma = 0, \lambda = 1$									
Poisson FE	−0.000	0.017	0.017	−0.001	0.011	0.011	−0.000	0.007	0.007
Feasible alternative	−0.000	0.017	0.017	−0.001	0.011	0.011	−0.000	0.007	0.007
Unfeasible optimal	−0.000	0.017	0.017	−0.001	0.011	0.011	−0.000	0.007	0.007
$\theta = 1, \gamma = .5, \lambda = 0$									
Poisson FE	−0.000	0.028	0.028	0.000	0.022	0.022	0.001	0.020	0.020
Feasible alternative	0.000	0.025	0.025	0.001	0.017	0.017	0.000	0.014	0.014
Unfeasible optimal	0.000	0.024	0.024	0.001	0.017	0.017	−0.000	0.013	0.013
$\theta = 1, \gamma = .5, \lambda = 1$									
Poisson FE	−0.001	0.028	0.028	0.001	0.022	0.022	−0.000	0.021	0.021
Feasible alternative	−0.001	0.024	0.024	0.000	0.017	0.017	0.000	0.022	0.022
Unfeasible optimal	−0.001	0.024	0.024	0.000	0.016	0.016	−0.000	0.013	0.013

APPENDICES

APPENDIX A

ESTIMATION OF DYNAMIC PANEL DATA MODELS WITH CROSS-SECTIONAL DEPENDENCE

A.1 Efficient Estimation with Clustering

A.1.1 Unfeasible Optimal Instruments

Consider any GMM estimator of ρ_0 defined as in (1.2.6) for some set of valid instruments $\{Z_i\}_{i=1,...,n}$ of dimension $r \times (T-1)$ which can be rewritten as:

$$\hat{\rho} = argmin_{\rho}(\sum_{g=1}^{G} Z^g m^g(\rho))' \Xi \sum_{g=1}^{G} Z^g m^g(\rho) \tag{A.1.1}$$

where $Z^g = [Z_2^g, ..., Z_T^g]$ and $Z_t^g = [Z_{i_1 t}, ..., Z_{i_{n_g} t}]$.

From White (2001), $\hat{\rho}$ is consistent for ρ_0 and:

$$\sqrt{G}(\hat{\rho} - \rho_0) \xrightarrow{d} N(0, (D' \Xi D)^{-1} D' \Xi \Upsilon \Xi D (D' \Xi D)^{-1}) \tag{A.1.2}$$

with $D = plim(\frac{1}{G}\sum_{g=1}^{G} Z^g \Delta Y_{-1}^g)$ and $\Upsilon = plim(\frac{1}{G}\sum_{g=1}^{G} Z^g m^g m^{g'} Z^{g'})$.

$\Xi = \Upsilon^{-1}$ is the optimal weighting matrix for that estimator and with such weighting matrix:

$$\sqrt{G}(\hat{\rho} - \rho_0) \xrightarrow{d} N(0, (D' \Upsilon^{-1} D)^{-1}) \tag{A.1.3}$$

Therefore in this section we will show that the asymptotic variance of $\hat{\rho}_{opt}$ defined by (1.3.4) is smaller than $(D' \Upsilon^{-1} D)^{-1}$ for any set of valid matrices of instruments $\{Z_i\}_{i=1,...,n}$ as long as (1.2.1), (1.2.2) and Auxiliary Assumption 1 are satisfied.

Since $\Phi^{g\,-1/2}$ is upper triangular with its element in row j, column i being a function of $Y_{max\{i,j\}-1}^g$, for any valid set of instruments $\{Z^g\}_{g=1,...,G}$ we have:

$$E(Z^g (\Phi^g)^{-1/2} m^g) = 0 \tag{A.1.4}$$

because the j^{th} $r \times n_g$ component of $Z^g (\Phi^g)^{-1/2}$ is a function of Y_{j-1}^g.

In addition, we have:

$$Var(Z^g (\Phi^g)^{-1/2} m^g) = E(Z^g (\Phi^g)^{-1/2} \Phi^g (\Phi^g)^{-1/2\prime} Z^{g\prime}) \tag{A.1.5}$$

$$= E(Z^g Z^{g\prime}) \tag{A.1.6}$$

because the j^{th} $r \times n_g$ component of $Z^g(\Phi^g)^{-1/2}$ is a function of Y^g_{j-1} and

$$\Phi^g = [E(m^g_t m^{g'}_s | Y^g_{max\{t,s\}-2})]^{s=2,...,T}_{t=2,...,T} \tag{A.1.7}$$

Note that since $E((Z^g m^g m^{g'} Z^{g'}) - Var(Z^g m^g)) = 0$, then

$$plim \frac{1}{G} \sum_{g=1}^{G} (Z^g m^g m^{g'} Z^{g'}) = plim \frac{1}{G} \sum_{g=1}^{G} Var(Z^g m^g) \tag{A.1.8}$$

Define:

$$\Delta \tilde{Y}^g_{-1} = (\Phi^g)^{-1/2} \Delta Y^g_{-1} \tag{A.1.9}$$

Define $\Delta \tilde{Y}^g_{-1,t}$ the t^{th} block of n_g rows of $\Delta \tilde{Y}^g_{-1}$. Define:

$$L^{\star g}_t = E(\Delta \tilde{Y}^g_{-1,t} | Y_{t-2}) \tag{A.1.10}$$

Define $L^{\star g} = [L^{\star g'}_2, ..., L^{\star g'}_T]$ and:

$$Z^g_{opt} = L^{\star g} (\Phi^g)^{-1/2} \tag{A.1.11}$$

Define $D_{opt} = plim(\frac{1}{G} \sum_{g=1}^{G} Z^g_{opt} \Delta Y^g_{-1})$ and $\Upsilon_{opt} = plim(\frac{1}{G} \sum_{g=1}^{G} Z^g_{opt} m^g m^{g'} Z^{g'}_{opt})$.

$E(Z^g_{opt} \Delta Y^g_{-1} - L^{\star g} L^{\star g'}) = 0$ because the j^{th} $r \times n_g$ component of $Z^g \Phi^{g-1/2}$ is a function of Y^g_{j-1} and $L^{\star g}_t = E(\Delta \tilde{Y}^g_{-1,t} | Y_{t-2})$. Therefore:

$$D_{opt} = plim(\frac{1}{G} \sum_{g=1}^{G} Z^g_{opt} \Delta Y^g_{-1}) \tag{A.1.12}$$

$$= plim(\frac{1}{G} \sum_{g=1}^{G} L^{\star g} L^{\star g'}) \tag{A.1.13}$$

Since $Var(Z^g(\Phi^g)^{-1/2} m^g) = E(Z^g Z^{g'})$, we have in particular: $Var(L^{\star g}(\Phi^g)^{-1/2} m^g) = E(L^{\star g} L^{\star g'})$. Therefore:

$$\Upsilon_{opt} = plim(\frac{1}{G} \sum_{g=1}^{G} Z^g_{opt} m^g m^{g'} Z^{g'}_{opt}) \tag{A.1.14}$$

$$= plim(\frac{1}{G} \sum_{g=1}^{G} L^{\star g} L^{\star g'}) \tag{A.1.15}$$

$$= D_{opt} \tag{A.1.16}$$

so that $(D'_{opt} \Upsilon^{-1}_{opt} D_{opt})^{-1} = D^{-1}_{opt}$.

Therefore the estimator $\hat{\rho}_{opt}$ defined by:

$$\hat{\rho}_{opt} = argmin_{\rho}(\sum_{g=1}^{G} Z_{opt}^{g} m^{g}(\rho))' \sum_{g=1}^{G} Z_{opt}^{g} m^{g}(\rho) \tag{A.1.17}$$

is consistent for ρ_0 and $\sqrt{G}$-asymtotically normal with asymptotic variance:

$$V_{opt} = D_{opt}^{-1} \tag{A.1.18}$$

We can show that this variance-covariance matrix is smaller than $(D'\Upsilon^{-1}D)^{-1}$ no matter what set of instruments $\{Z^g\}_{g=1,...,G}$ is used. Denote Δ the difference between $D'\Upsilon^{-1}D$ and D_{opt}:

$$D = D'\Upsilon^{-1}D - D_{opt} \tag{A.1.19}$$

$$= D'\Upsilon^{-1}D - plim(\frac{1}{G}\sum_{g=1}^{G} L^{\star g} L^{\star g'}) \tag{A.1.20}$$

$$= D'\Upsilon^{-1}D - plim(\frac{1}{G}\sum_{g=1}^{G} Z_{opt}^{g} \Phi^{g} Z_{opt}^{g'}) \tag{A.1.21}$$

Since $(\Phi^g)^{-1/2'}(\Phi^g)^{-1/2} = (\Phi^g)^{-1}$ we also have $\Phi^g = \Phi^{g1/2}\Phi^{g1/2'}$ where $\Phi^{g1/2}$ is upper triangular and is composed of $n_g \times n_g$ matrices such that the $(j,k)^{th}$ matrix for $k > j$ is a function of Y_{k-1}^g. Therefore:

$$E(Z^g(\frac{\partial m^g(\rho_0)}{\partial \rho} - \Phi^g Z_{opt}^{g'})) = 0 \tag{A.1.22}$$

since the j^{th} $r \times n_g$ component of Z^g is a function of Y_{j-1}^g and $(\Phi^g)_t^{1/2} L_t^{\star g} = E((\Phi^g)^{1/2}\Delta\tilde{Y}_{-1,t}^g | Y_{t-2})$ where $(\Phi^g)_t^{1/2}$ is the $(t-1)^{th}$ $n_g \times n_g(T-1)$ matrix composing $(\Phi^g)^{1/2}$.

In addition we have

$$E(Z^g(m^g(\rho_0)m^g(\rho_0)' - \Phi^g)Z^{g'}) = 0 \tag{A.1.23}$$

We can then apply the WLLN to show that:

$$D = plim(\frac{1}{G}\sum_{g=1}^{G} Z^g \Phi^g Z_{opt}^{g'}) \tag{A.1.24}$$

$$\Upsilon = plim(\frac{1}{G}\sum_{g=1}^{G} Z^g \Phi^g Z^{g'}) \tag{A.1.25}$$

Define $D_n = \frac{1}{G}\sum_{g=1}^{G} Z^g \Phi^g Z_{opt}^{g'}$, $\Upsilon_n = \frac{1}{G}\sum_{g=1}^{G} Z^g \Phi^g Z^{g'}$ and $D_n^{\star} = \frac{1}{G}\sum_{g=1}^{G} Z_{opt}^{g} \Phi^g Z_{opt}^{g'}$. Define $Z^{\star} = [Z_{opt}^{1'}, ..., Z_{opt}^{G'}]'$, $Z = [Z_1', ..., Z_G']'$, $S = diag(\{\Phi^g\}_{g=1,...,G})$, then:

$$D_n'\Upsilon_n^{-1}D_n - G_n^{\star} = Z^{\star'}SZ(Z'SZ)^{-1}Z'SZ^{\star} - Z^{\star'}SZ^{\star} \tag{A.1.26}$$

$$= Z^{\star'}S^{1/2}(S^{1/2}Z(Z'SZ)^{-1}Z'S^{1/2} - I_{T\times n})S^{1/2}Z \tag{A.1.27}$$

Therefore $D_n^{'}\Upsilon_n^{-1}D_n - D_n^{\star}$ is positive semi-definite for any value of n. Therefore $D^{'}\Upsilon^{-1}D - D_{opt}$ is positive semi-definite by the continuous mapping theorem. A similar result was found in Chamberlain (1992a) for the case of cross-sectional independence.

A.1.2 Efficient Estimation with Auxiliary Assumptions

Under Auxiliary Assumptions 1-2a, the variance-covariance matrix of u_t^g is:

$$\Sigma_u^g = \sigma_u^2 \begin{bmatrix} 1 & & & \\ \tau_u & 1 & & \\ \dots & & \dots & \\ \tau_u & \dots & \tau_u & 1 \end{bmatrix} \tag{A.1.28}$$

and we have

$$E(m_t^g m_s^{g'} | Y_{max\{t,s\}-2}) = 2\Sigma_u^g \, if \, t = s \tag{A.1.29}$$

$$= -\Sigma_u^g \, if \, |t-s| = 1 \tag{A.1.30}$$

$$= 0 \, if \, |t-s| \geq 2 \tag{A.1.31}$$

Therefore:

$$\Phi^g = J^g (I_T \otimes \Sigma^g) J^{g'} \tag{A.1.32}$$

where:

$$J^g = \begin{bmatrix} -1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & -1 & 0 & \dots & 0 & 1 & \dots & 0 \\ & & & & \dots & & & \\ 0 & \dots & & 0 & -1 & \dots & 0 & 1 \end{bmatrix} \tag{A.1.33}$$

is the deterministic differencing matrix such that $J^g u^g = m^g$.

Therefore $L_t^{\star g} = \Phi^{g-1/2} E(\frac{\partial m^g}{\partial \rho} | Y_{t-2})$ and

$$Z_{opt}^g = [E(\Delta Y_{-1}^g | Y_0)^{'}, \dots, E(\Delta Y_{-1}^g | Y_{T-2})^{'}] \Psi^g \Phi^{g-1/2} \tag{A.1.34}$$

where:

$$\Psi^g = \begin{bmatrix} \Phi_1^{g-1/2'} & 0 & ... & 0 \\ 0 & \Phi_2^{g-1/2'} & & ... \\ ... & & ... & 0 \\ 0 & ... & 0 & \Phi_{T-1}^{g-1/2'} \end{bmatrix} \tag{A.1.35}$$

where $\Phi_j^{g-1/2}$ is the j^{th} $n_g \times n_g(T-1)$ matrix composing $(\Phi^g)^{-1/2}$.

A.1.3 Conditional Expectation of Unobserved Heterogeneity under Clustering

Under Auxiliary Assumption 1, 2a, 3a we have:

$$\begin{bmatrix} c^g \\ y_0^g \\ c^g + u_1^g \\ ... \\ c^g + u_T^g \end{bmatrix} \sim N(\mu^g, A^g V^g A^{g'}) \tag{A.1.36}$$

Therefore, using the properties of the multivariate normal distribution, we have:

$$E(c^g | \begin{bmatrix} y_0^g \\ c^g + u_1^g \\ ... \\ c^g + u_T^g \end{bmatrix}) = \mu_c \iota_{n_g} + A^g V^g A_{12}^{g'} A^g V^g A_{22}^{g'-1} (\begin{bmatrix} y_0^g \\ c^g + u_1^g \\ ... \\ c^g + u_T^g \end{bmatrix} - \begin{bmatrix} \mu_c \iota_{T \times n_g} \\ \frac{1}{1-\rho_0} \mu_c \iota_{n_g} \end{bmatrix}) \tag{A.1.37}$$

where $A^g V^g A_{12}^{g'} = Cov(c^g, \begin{bmatrix} y_0^g \\ c^g + u_1^g \\ ... \\ c^g + u_T^g \end{bmatrix})$ and $A^g V^g A_{22}^{g'} = Var(\begin{bmatrix} y_0^g \\ c^g + u_1^g \\ ... \\ c^g + u_T^g \end{bmatrix})$ and both matrices are components of $A^g V^g A^{g'}$.

$E(c^g | \begin{bmatrix} y_0^g \\ c^g + u_1^g \\ ... \\ c^g + u_t^g \end{bmatrix})$ can be obtained in a similar fashion by considering only the first $((t+2)n^g) \times ((t+2)n^g)$ block of $A^g V^g A^{g'}$

APPENDIX B

ESTIMATION OF UNOBSERVED EFFECTS PANEL DATA MODELS UNDER SEQUENTIAL EXOGENEITY

B.1 GMM Estimation and Efficiency Bound

Define $\Sigma_i = [Cov(\rho_{it}, \rho_{is}|z_{imax(t-1,s-1)})]_{t=1,...,T}^{s=1,...,T}$. define A_i and $\tilde{\Sigma}_i^{-1}$ to be the terms of the LDL decomposition of Σ_i^{-1}: $\Sigma_i^{-1} = A_i^{'}\tilde{\Sigma}_i^{-1}A_i$ where $\tilde{\Sigma}_i$ is diagonal and A_i is upper-triangular with only ones on the diagonal. We can show that $A_i = [1(s \geq t)(-1)^{1(s \neq t)}\Gamma_{it,s}]_{t=2,...,T}^{s=2,...,T}$ where $1(.)$ is the indicative function and $\tilde{\Sigma}_i = diag(\{Var(\tilde{\rho}_{it}|z_{it-1})\}_{t=2,...,T})$ so that:

$$J = E(\sum_{t=2}^{T} \tilde{D}_{it}^{'}\tilde{\Sigma}_{it}^{-1}\tilde{D}_{it}) \tag{B.1.1}$$

$$= E(\ddot{E}(A_i\frac{\partial \rho_i}{\partial \beta'}|z_i)^{'}\tilde{\Sigma}_i^{-1}\ddot{E}(A_i\frac{\partial \rho_i}{\partial \beta'}|z_i)) \tag{B.1.2}$$

where $\ddot{E}([g_t]_{t=2,...,T}|[x_1,...,x_{T-1}])$ is a matrix operator that returns $E(g_t|x_{t-1})$ as its $(t-1)^{th}$ row, $t = 2,...,T$, where g_t are row random vectors.

Using standard results of GMM estimation we can write:

$$\sqrt{n}(\hat{\beta}_{Lin} - \beta_0) = W^{-1}(\frac{1}{n}\sum_{i=1}^{n} Z_i\frac{\partial \rho_i}{\partial \beta'})^{'}(\frac{1}{n}\sum_{i=1}^{n} Z_i\rho_i\rho_i^{'}Z_i^{'})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i\rho_i + o_p(1) \tag{B.1.3}$$

$$W = (\frac{1}{n}\sum_{i=1}^{n} Z_i\frac{\partial \rho_i}{\partial \beta'})^{'}(\frac{1}{n}\sum_{i=1}^{n} Z_i\rho_i\rho_i^{'}Z_i^{'})^{-1}\frac{1}{n}\sum_{i=1}^{n} Z_i\frac{\partial \rho_i}{\partial \beta'} \tag{B.1.4}$$

where $\rho_i = \rho_i(\beta_0)$ and $\frac{\partial \rho_i}{\partial \beta'} = \frac{\partial \rho_i(\beta_0)}{\partial \beta'}$.

Applying the WLLN, $\frac{1}{n}\sum_{i=1}^{n} Z_i\frac{\partial \rho_i}{\partial \beta'} = O_p(1)$. Also $\frac{1}{n}\sum_{i=1}^{n} Z_i(\rho_i\rho_i^{'} - \Sigma_i)Z_i^{'} = o_p(1)$ since $E(Z_i(\rho_i\rho_i^{'} - \Sigma_i)Z_i^{'}) = 0$ from how Z_i and Σ_i where defined. Using the CLT, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i\rho_i = O_p(1)$. Using Slutsky's theorem, assuming $plim\frac{1}{n}\sum_{i=1}^{n} Z_i\rho_i\rho_i^{'}Z_i^{'}$ is p.d., we have

$$(\frac{1}{n}\sum_{i=1}^{n} Z_i\rho_i\rho_i^{'}Z_i^{'})^{-1} - (\frac{1}{n}\sum_{i=1}^{n} Z_i\Sigma_iZ_i^{'})^{-1} = o_p(1) \tag{B.1.5}$$

So $W = V + o_p(1)$ where $V = (\frac{1}{n}\sum_{i=1}^{n} Z_i\frac{\partial \rho_i}{\partial \beta'})^{'}(\frac{1}{n}\sum_{i=1}^{n} Z_i\Sigma_iZ_i^{'})^{-1}\frac{1}{n}\sum_{i=1}^{n} Z_i\frac{\partial \rho_i}{\partial \beta'}$ and using Slutsky's

theorem again, assuming $plim W$ is finite and p.d., $W^{-1} = V^{-1} + o_p(1)$. Therefore we can rewrite:

$$\sqrt{n}(\hat{\beta}_{Lin} - \beta_0) = V^{-1}(\frac{1}{n}\sum_{i=1}^{n} Z_i \frac{\partial \rho_i}{\partial \beta'})'(\frac{1}{n}\sum_{i=1}^{n} Z_i \Sigma_i Z_i')^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \rho_i + o_p(1) \tag{B.1.6}$$

$$V = (\frac{1}{n}\sum_{i=1}^{n} Z_i \frac{\partial \rho_i}{\partial \beta'})'(\frac{1}{n}\sum_{i=1}^{n} Z_i \Sigma_i Z_i')^{-1}\frac{1}{n}\sum_{i=1}^{n} Z_i \frac{\partial \rho_i}{\partial \beta'} \tag{B.1.7}$$

In addition:

$$n(\hat{\beta}_{Lin} - \beta_0)(\hat{\beta}_{Lin} - \beta_0)' = V^{-1} + o_p(1) \tag{B.1.8}$$

Since:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \rho_i (\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \rho_i)' = \frac{1}{n}\sum_{i=1}^{j}\sum_{j=1}^{n} Z_i \rho_i \rho_j' Z_j' \tag{B.1.9}$$

$$= \frac{1}{n}\sum_{i=1}^{j} Z_i \rho_i \rho_i' Z_i' + o_p(1) \tag{B.1.10}$$

$$= \frac{1}{n}\sum_{i=1}^{j} Z_i \Sigma_i Z_i' + o_p(1) \tag{B.1.11}$$

where the second equality follows from random sampling and the WLLN.

We can rewrite V as:

$$V = \frac{\partial \ddot{\rho}}{\partial \beta_0}' \ddot{Z}'(\ddot{Z}\ddot{Z}')^{-1}\ddot{Z}\frac{\partial \ddot{\rho}}{\partial \beta_0} \tag{B.1.12}$$

where $\frac{\partial \ddot{\rho}}{\partial \beta'} = [\frac{\partial \ddot{\rho}_1}{\partial \beta'}', ..., \frac{\partial \ddot{\rho}_n}{\partial \beta'}']'$, $\frac{\partial \ddot{\rho}_i}{\partial \beta'} = \tilde{\Sigma}^{-1/2} A \frac{\partial \rho_i}{\partial \beta'}$, $\ddot{Z} = [\ddot{Z}_1, ..., \ddot{Z}_n]'$, $\ddot{Z}_i = Z_i A^{-1}\tilde{\Sigma}^{1/2}$.

Consider the matrix linear projection of $\frac{\partial \ddot{\rho}_i}{\partial \beta'}$ on $\ddot{Z}_i$, $LP(\frac{\partial \ddot{\rho}_i}{\partial \beta'}|\ddot{Z}_i) = \ddot{Z}_i' C$, where C is a $dim(Z_i) \times dim(\beta)$ deterministic matrix defined by the moment conditions:

$$E(\ddot{Z}_i(\frac{\partial \ddot{\rho}_i}{\partial \beta'} - \ddot{Z}_i' C)) = 0 \tag{B.1.13}$$

It is a standard result that as long as $E(\ddot{Z}_i \ddot{Z}_i')$ is finite and p.d. and $E(\ddot{Z}_i \frac{\partial \ddot{\rho}_i}{\partial \beta'})$ exists, this linear projection is consistently estimated by:

$$\hat{LP}(\frac{\partial \ddot{\rho}}{\partial \beta'}|\ddot{Z}_i) = \ddot{Z}_i'(\ddot{Z}_i'\ddot{Z}_i)^{-1}\ddot{Z}_i'\frac{\partial \ddot{\rho}_i}{\partial \beta_0} \tag{B.1.14}$$

$$= LP(\frac{\partial \ddot{\rho}_i}{\partial \beta'}|\ddot{Z}_i) + o_p(1) \tag{B.1.15}$$

Define the stacked estimated linear projections by $\hat{LP}(\frac{\partial \ddot{\rho}}{\partial \beta'}|\ddot{Z}) = \ddot{Z}(\ddot{Z}'\ddot{Z})^{-1}\ddot{Z}'\frac{\partial \ddot{\rho}}{\partial \beta_0}$, denote $P_{\ddot{Z}} = \ddot{Z}(\ddot{Z}'\ddot{Z})^{-1}\ddot{Z}'$. Since $P_{\ddot{Z}}$ is idempotent, we have:

$$V = \frac{\partial \ddot{\rho}}{\partial \beta_0}' P_{\ddot{Z}} \frac{\partial \ddot{\rho}}{\partial \beta_0}/n \tag{B.1.16}$$

$$= \frac{\partial \ddot{\rho}}{\partial \beta_0}' P_{\ddot{Z}} P_{\ddot{Z}} \frac{\partial \ddot{\rho}}{\partial \beta_0}/n \tag{B.1.17}$$

$$= \hat{LP}(\frac{\partial \ddot{\rho}}{\partial \beta'}|\ddot{Z})' \hat{LP}(\frac{\partial \ddot{\rho}}{\partial \beta'}|\ddot{Z})/n \tag{B.1.18}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \hat{LP}(\frac{\partial \ddot{\rho}}{\partial \beta'}|\ddot{Z}_i)' \hat{LP}(\frac{\partial \ddot{\rho}}{\partial \beta'}|\ddot{Z}_i) \tag{B.1.19}$$

$$= E(LP(\frac{\partial \ddot{\rho}}{\partial \beta'}|\ddot{Z}_i)' LP(\frac{\partial \ddot{\rho}}{\partial \beta'}|\ddot{Z}_i)) + o_p(1) \tag{B.1.20}$$

where the last equality follows from Newey and McFadden (1994) for instance.

In addition, the matrix linear projection of $\frac{\partial \ddot{\rho}}{\partial \beta'}$ on $\ddot{Z}_i$ is the same as the matrix linear projection of $\ddot{E}(\frac{\partial \ddot{\rho}}{\partial \beta'}|Z_i)$ on $\ddot{Z}_i$ defined by:

$$E(\ddot{Z}_i(\ddot{E}(\frac{\partial \ddot{\rho}_i}{\partial \beta'}|Z_i) - \ddot{Z}_i'C)) = 0 \tag{B.1.21}$$

Since the t^{th} vector of $\ddot{Z}_i$, $\ddot{Z}_{it}$, is a function of Z_{it} since z_{it} contains $z_{i1}, ..., z_{it-1}$. Therefore:

$$V = E(LP(\ddot{E}(\frac{\partial \ddot{\rho}_i}{\partial \beta'}|Z_i)|\ddot{Z}_i)' LP(\ddot{E}(\frac{\partial \ddot{\rho}_i}{\partial \beta'}|Z_i)|\ddot{Z}_i)) + o_p(1) \tag{B.1.22}$$

So using the standard results on linear projection:

$$V = E(\ddot{E}(\frac{\partial \ddot{\rho}_i}{\partial \beta'}|Z_i)' \ddot{E}(\frac{\partial \ddot{\rho}_i}{\partial \beta'}|Z_i)) - E(e_i'e_i) + o_p(1) \tag{B.1.23}$$

$$= E(\sum_{t=2}^{T} \tilde{D}_t'\tilde{\Sigma}_t\tilde{D}_t) - E(\sum_{t=2}^{T} e_{it}'e_{it}) + o_p(1) \tag{B.1.24}$$

where $e_i = E(\frac{\partial \ddot{\rho}_i}{\partial \beta'}|Z_i) - LP(\ddot{E}(\frac{\partial \ddot{\rho}_i}{\partial \beta'}|Z_i)|\ddot{Z}_i)$ and $e_{it} = E(\frac{\partial \ddot{\rho}_{it}}{\partial \beta'}|Z_{it}) - LP(E(\frac{\partial \ddot{\rho}_{it}}{\partial \beta'}|Z_{it})|\ddot{Z}_{it})$.

APPENDIX C

EFFICIENCY OF THE POISSON FIXED EFFECTS ESTIMATOR

C.1 Efficient Estimation under Conditional Mean Restrictions

Chamberlain (1992b), page 581, showed that the asymptotic information bound for estimating β_0 from (3.2.1) is:

$$V_{0,\beta_0}^{-1} = E(h_i \partial\mu_i^{'}(\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1})\partial\mu_i h_i) \tag{C.1.1}$$

where $h_i = E(c_i|x_i)$, $x_i = \{x_{i1},...,x_{iT}\}$ $\partial\mu_i^{'} = [\partial\mu_{i1},...,\partial\mu_{iT}]$, $\partial\mu_{it} = \frac{\partial\mu_{it}}{\partial\beta}$, $\Sigma_{y,i} = Var(y_i|x_i)$, $y_i^{'} = [y_{i1},...,y_{iT}]$, $\mu_i^{'} = [\mu_{i1},...,\mu_{iT}]$.

$\rho_i = \rho_i(\beta_0)$ can be rewritten as:

$$\rho_i = (I - P_i)y_i \tag{C.1.2}$$

where

$$P_i = \frac{1}{\sum_{t=1}^{T}\mu_{it}} \begin{bmatrix} \mu_{i1} & ... & \mu_{i1} \\ & ... & \\ \mu_{iT} & ... & \mu_{iT} \end{bmatrix} \tag{C.1.3}$$

Therefore:

$$\begin{aligned} D_i &= E((\frac{1}{\sum_{t=1}^{T}\mu_{it}} \begin{bmatrix} \partial\mu_{i1} & ... & \partial\mu_{i1} \\ & ... & \\ \partial\mu_{iT} & ... & \partial\mu_{iT} \end{bmatrix} - \frac{\sum_{t=1}^{T}\partial\mu_{it}}{(\sum_{t=1}^{T}\mu_{it})^2} \begin{bmatrix} \mu_{i1} & ... & \mu_{i1} \\ & ... & \\ \mu_{iT} & ... & \mu_{iT} \end{bmatrix})y_i|x_i) \\ &= h_i(\frac{1}{\sum_{t=1}^{T}\mu_{it}} \begin{bmatrix} \partial\mu_{i1} & ... & \partial\mu_{i1} \\ & ... & \\ \partial\mu_{iT} & ... & \partial\mu_{iT} \end{bmatrix} - \frac{\sum_{t=1}^{T}\partial\mu_{it}}{(\sum_{t=1}^{T}\mu_{it})^2} \begin{bmatrix} \mu_{i1} & ... & \mu_{i1} \\ & ... & \\ \mu_{iT} & ... & \mu_{iT} \end{bmatrix})\mu_i \\ &= h_i(\partial\mu_i - \frac{\sum_{t=1}^{T}\partial\mu_{it}}{\sum_{t=1}^{T}\mu_{it}}\mu_i) \end{aligned}$$

Note that:

$$\mu_i^{'}(\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1}) = 0 \tag{C.1.4}$$

Therefore:

$$D_i^{'}(\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1})D_i = h_i\partial\mu_i^{'}(\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1})\partial\mu_i h_i \tag{C.1.5}$$

Therefore the only thing left to show is that $(\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1})$ is a generalized inverse of Σ_i.

$$\begin{aligned}\Sigma_i &= (I - P_i)\Sigma_{y,i}(I - P_i^{'}) \\ &= \Sigma_{y,i} - P_i\Sigma_{y,i} - \Sigma_{y,i}P_i^{'} + P_i\Sigma_{y,i}P_i^{'}\end{aligned}$$

Note that:

$$P_i\mu_i = \mu_i \tag{C.1.6}$$

and:

$$P_i^{'}(\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1}) = 0 \tag{C.1.7}$$

Therefore:

$$\begin{aligned}(\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1})\Sigma_i &= (\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1})\Sigma_{y_i} - 0 \\ &\quad - (\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1})\Sigma_{y,i}P_i^{'} + 0 \\ &= I - P_i^{'}\end{aligned}$$

Let $Mat_i = (\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1})$. Therefore:

$$\begin{aligned}Mat_i\Sigma_i Mat_i &= (\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1}) - 0 \\ &= (\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1})\end{aligned}$$

Note that:

$$P_iP_i = P_i \tag{C.1.8}$$

Therefore:

$$\begin{aligned}\Sigma_i(\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1})\Sigma_i &= \Sigma_i - \Sigma_iP_i^{'} \\ &= \Sigma_i - \Sigma_{y,i}P_i^{'} + P_i\Sigma_{y,i}P_i^{'} + \Sigma_{y,i}P_i^{'} - P_i\Sigma_{y,i}P_i^{'} \\ &= \Sigma_i\end{aligned}$$

So $(\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i^{'}\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i^{'}\Sigma_{y,i}^{-1})$ is indeed a generalized inverse of Σ_i. Therefore:

$$V_{0,\beta}^{-1} = E(D_i^{'}\Sigma_i^{-}D_i) = V_{opt}^{-1} \tag{C.1.9}$$

where $\Sigma_i^- = (\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i'\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i'\Sigma_{y,i}^{-1})$.

In addition, we can characterize $(\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i'\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i'\Sigma_{y,i}^{-1})$ in an alternative way that will be useful for future results. Denote $B_i = (\Sigma_{y,i}^{-1} - \Sigma_{y,i}^{-1}\mu_i(\mu_i'\Sigma_{y,i}^{-1}\mu_i)^{-1}\mu_i'\Sigma_{y,i}^{-1})$. We have shown that:

$$B_i\Sigma_i = I - P_i' \tag{C.1.10}$$

Since B_i and Σ_i are symmetric:

$$\Sigma_i B_i = I - P_i \tag{C.1.11}$$

B_i is the unique matrix X that satisfies:

$$X\Sigma_i X = X \tag{C.1.12}$$

$$\Sigma_i X \Sigma_i = \Sigma_i \tag{C.1.13}$$

$$X\Sigma_i = I - P_i' \tag{C.1.14}$$

$$\Sigma_i X = I - P_i \tag{C.1.15}$$

We have already shown that $X = B_i$ satisfies all of these requirements. This solution is unique since for any X, Y satisfying these requirements[1]:

$$X = X\Sigma_i X = X(I - P_i) = X\Sigma_i Y = (I - P_i')Y = Y\Sigma_i Y = Y \tag{C.1.16}$$

C.2 Efficient Estimation under the Poisson FE Assumptions

Under (3.2.9) and (3.2.10) we have $\Sigma_{y,i} = h_i diag(\mu_i) + v_i\mu_i\mu_i'$ where $v_i = Var(c_i|x_i)$. Therefore:

$$\begin{aligned}\Sigma_i &= Var(\rho_i|x_i) \\ &= (I - P_i)(h_i diag(\mu_i) + v_i\mu_i\mu_i')(I - P_i') \\ &= (I - P_i)h_i diag(\mu_i)(I - P_i')\end{aligned}$$

where the last equality follows from $\mu_{it}\mu_{is} - \mu_{is}p_{it}\sum_{r=1}^T \mu_{ir} = 0$.

Define:

$$X_i = h_i^{-1}(diag(\frac{1}{\mu_i}) - \frac{1}{\sum_{t=1}^T \mu_{it}}J) \tag{C.2.1}$$

[1] This proof of uniqueness is identical to the proof of uniqueness for the Moore-Penrose pseudo inverse found in Penrose (1955).

where $J = \begin{bmatrix} 1 & \dots & 1 \\ & \dots & \\ 1 & \dots & 1 \end{bmatrix}$ and, by an abuse of notation, $(\frac{1}{\mu_i})' = [\frac{1}{\mu_{i1}}, ..., \frac{1}{\mu_{iT}}]$.

Note that:

$$\begin{aligned} JP_i &= \frac{\sum_{t=1}^{T} \mu_{it}}{\sum_{t=1}^{T} \mu_{it}} J \\ &= J \end{aligned}$$

and:

$$diag(\frac{1}{\mu_i})P_i = \frac{1}{\sum_{t=1}^{T} \mu_{it}} J \tag{C.2.2}$$

Therefore:

$$\begin{aligned} \Sigma_i X_i &= (I - P_i) diag(\mu_i)(I - P_i')(diag(\frac{1}{\mu_i}) - \frac{1}{\sum_{t=1}^{T} \mu_{it}} J) \\ &= (I - P_i) diag(\mu_i)(diag(\frac{1}{\mu_i}) - \frac{1}{\sum_{t=1}^{T} \mu_{it}} J - \frac{1}{\sum_{t=1}^{T} \mu_{it}} J + \frac{1}{\sum_{t=1}^{T} \mu_{it}} J) \\ &= (I - P_i)(I - P_i) \\ &= I - P_i \end{aligned}$$

Since both Σ_i and X_i are symmetric:

$$X_i \Sigma_i = I - P_i' \tag{C.2.3}$$

So in order to show that $\Sigma_i^- = X_i$, there only remains to show that X_i satisfies (C.1.12) and (C.1.13).

For (C.1.12):

$$\begin{aligned} X_i \Sigma_i X_i &= X_i (I - P_i) \\ &= X_i - X_i P_i \\ &= X_i - h_i^{-1} (\frac{1}{\sum_{t=1}^{T} \mu_{it}} J - \frac{1}{\sum_{t=1}^{T} \mu_{it}} J) \\ &= X_i \end{aligned}$$

For (C.1.13):

$$\begin{aligned}\Sigma_i X_i \Sigma_i &= (I - P_i)X_i \\ &= \Sigma_i - P_i\Sigma_i \\ &= \Sigma_i - P_i(I - P_i)h_i diag(\mu_i)(I - P_i^{'}) \\ &= \Sigma_i\end{aligned}$$

Therefore we have shown that in this case:

$$\Sigma_i^{-} = h_i^{-1}(diag(\frac{1}{\mu_i}) - \frac{1}{\sum_{t=1}^{T}\mu_{it}}J) \tag{C.2.4}$$

Therefore:

$$\begin{aligned}D_i^{'}\Sigma_i^{-} &= \frac{h_i}{h_i}\partial\mu_i^{'}(diag(\frac{1}{\mu_i}) - \frac{1}{\sum_{t=1}^{T}\mu_{it}}J) \\ &= (\frac{\partial\mu_i}{\mu_i})^{'} - \frac{\sum_{t=1}^{T}\partial\mu_{it}}{\sum_{t=1}^{T}\mu_{it}}j^{'}\end{aligned}$$

where $j^{'} = [1, ..., 1]$ and, by an abuse of notation, $(\frac{\partial\mu_i}{\mu_i})^{'} = [\frac{\partial\mu_{i1}}{\mu_{i1}}, ..., \frac{\partial\mu_{iT}}{\mu_{iT}}]$.

Note that:

$$\frac{\partial p_{it}}{\partial\beta} = \frac{\partial\mu_{it}}{\sum_{t=1}^{T}\mu_{it}} - \frac{\sum_{s=1}^{T}\partial\mu_{is}}{(\sum_{s=1}^{T}\mu_{is})^2}\mu_{it} \tag{C.2.5}$$

so that:

$$\frac{\partial p_{it}}{\partial\beta}\frac{1}{p_{it}} = \frac{\partial\mu_{it}}{\mu_{it}} - \frac{\sum_{s=1}^{T}\partial\mu_{is}}{\sum_{s=1}^{T}\mu_{is}} \tag{C.2.6}$$

Therefore:

$$(\frac{\partial p_i(\beta_0)}{\partial\beta^{'}})^{'}W_i(\beta_0)^{-1} = (\frac{\partial\mu_i}{\mu_i})^{'} - \frac{\sum_{t=1}^{T}\partial\mu_{it}}{\sum_{t=1}^{T}\mu_{it}}j^{'} \tag{C.2.7}$$

Hence we have shown that under (3.2.9) and (3.2.10):

$$D_i^{'}\Sigma_i^{-} = (\frac{\partial p_i(\beta_0)}{\partial\beta^{'}})^{'}W_i(\beta_0)^{-1} \tag{C.2.8}$$

C.3 Consistent Estimation of θ and ρ

Under (3.2.1), (3.2.11) and (3.2.12):

$$E(y_{it}^2|c_i,x_i) = c_i\mu_{it} + (\theta+1)c_i^2\mu_{it}^2$$

$$E(y_{it}y_{it-s}|c_i,x_i) = (\gamma+1)c_i^2\mu_{it}\mu_{it-s}\ for s=1$$

$$= c_i^2\mu_{it}\mu_{it-s}\ for s>1$$

Therefore:

$$\theta = \frac{E(\frac{y_{it}^2-y_{it}}{\mu_{it}^2})}{E(\frac{y_{it}y_{it-2}}{\mu_{it}\mu_{it-2}})} - 1 \tag{C.3.1}$$

and:

$$\gamma = \frac{E(\frac{y_{it}y_{it-1}}{\mu_{it}\mu_{it-1}})}{E(\frac{y_{it}y_{it-2}}{\mu_{it}\mu_{it-2}})} - 1 \tag{C.3.2}$$

Therefore a consistent estimator for θ under the assumptions (3.2.1), (3.2.11) and (3.2.12) is:

$$\hat{\theta} = \frac{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{y_{it}^2-y_{it}}{\ddot{\mu}_{it}^2}}{\frac{1}{n(T-2)}\sum_{i=1}^{n}\sum_{t=3}^{T}\frac{y_{it}y_{it-2}}{\ddot{\mu}_{it}\ddot{\mu}_{it-2}}} - 1 \tag{C.3.3}$$

A consistent estimator of γ is:

$$\hat{\gamma} = \frac{\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=2}^{T}\frac{y_{it}y_{it-1}}{\ddot{\mu}_{it}\ddot{\mu}_{it-1}}}{\frac{1}{n(T-2)}\sum_{i=1}^{n}\sum_{t=3}^{T}\frac{y_{it}y_{it-2}}{\ddot{\mu}_{it}\ddot{\mu}_{it-2}}} - 1 \tag{C.3.4}$$

C.4 Asymptotic equivalence of Poisson fixed effects and our alternative estimator when $\theta = 0, \gamma = 0$

When (3.2.9) and (3.2.10) hold, so that (3.2.11) and (3.2.12) hold with $\theta = 0$ and $\gamma = 0$, $\hat{\theta} \xrightarrow{p} 0$ and $\hat{\gamma} \xrightarrow{p} 0$ independently of whether (3.2.15) and (3.2.16) hold. Therefore:

$$\hat{D}_i \xrightarrow{p} h_1(x_i,\eta_p)(\sum_{t=1}^{T}\mu_{it})[\frac{\partial p_{it}}{\partial \beta'}]_{t=1,\ldots,T})'$$

$$\hat{\Sigma}_i \xrightarrow{p} [1(t=s)-p_{it}]_{t=1,\ldots,T}^{s=1,\ldots,T}[1[t=s]h_1(x_i,\eta_p)\sqrt{\mu_{it}\mu_{is}} + h_2(x_i,\eta_p)\mu_{it}\mu_{is}]_{t=1,\ldots,T}^{s=1,\ldots,T}([1(t=s)-p_{it}]_{t=1,\ldots,T}^{s=1,\ldots,T})'$$

where $\eta_p = plim(\hat{\eta})$. Appendix B, replacing $E(c_i|x_i)$ by $h_1(x_i, \eta_p)$ and $Var(c_i|x_i)$ by $h_2(x_i, \eta_p)$, shows that:

$$\begin{aligned} plim(\hat{D}_i')plim(\hat{\Sigma}_i)^- &= -(\frac{\partial p_i(\beta_0)}{\partial \beta'})' W_i(\beta_0)^{-1} \\ &= D_i' \Sigma_i^- \end{aligned}$$

Therefore when (3.2.9) and (3.2.10) hold, $\hat{\beta}_{PFE}$ and $\hat{\beta}_{alt}$ are asymptotically equivalent.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ahn, S. C. and Schmidt, P. (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics*, 68(1):5–27.

Alvarez, J. and Arellano, M. (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica*, 71(4):1121–1159.

Alvarez, J. and Arellano, M. (2004). Robust likelihood estimation of dynamic panel data models. *CEMFI Working Paper 0421*.

Anderson, T. W. and Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association*, 76(375):598–606.

Andrabi, T., Das, J., Ijaz Khwaja, A., and Zajonc, T. (2011). Do value-added estimates add value? accounting for learning dynamics. *American Economic Journal. Applied Economics*, 3(3):29–54.

Arellano, M. (2003). Modelling optimal instrumental variables for dynamic panel data models. *CEMFI Working Paper*.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2):277–297.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, 68(1):29–51.

Balasubramanian, N. and Sivadasan, J. (2010). What happens when firms patent? new evidence from U.S. economic census data. *Review of Economics and Statistics*, 93(1):126–146.

Baltagi, B. H., Fingleton, B., and Pirotte, A. (2014). Estimating and forecasting with a dynamic spatial panel data model*. *Oxford Bulletin of Economics and Statistics*, 76(1):112–138.

Bester, A. C., Conley, T. G., and Hansen, C. B. (2011a). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2):137–151.

Bester, A. C., Conley, T. G., Hansen, C. B., and Vogelsang, T. J. (2011b). Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators. *Working Paper*.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1):115–143.

Blundell, R., Griffith, R., and Van Reenen, J. (1995). Dynamic count data models of technological innovation. *The Economic Journal*, 105(429):333–344.

Blundell, R., Griffith, R., and Windmeijer, F. (2002). Individual effects and dynamics in count data models. *Journal of Econometrics*, 108(1):113–131.

Bond, S. R. (2002). Dynamic panel data models: a guide to micro data methods and practice. *Portuguese Economic Journal*, 1(2):141–162.

Browning, M., Ejraes, M., and Alvarez, J. (2010). Modelling income processes with lots of heterogeneity. *The Review of Economic Studies*, 77(4):1353–1381.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, 18(1):5–46.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334.

Chamberlain, G. (1992a). Comment: Sequential moment restrictions in panel data. *Journal of Business & Economic Statistics*, 10(1):20–26.

Chamberlain, G. (1992b). Efficiency bounds for semiparametric regression. *Econometrica*, 60(3):567–596.

Cizek, P., Jacobs, J. P., Ligthart, J. E., and Vrijburg, H. (2011). GMM estimation of fixed effects dynamic panel data models with spatial lag and spatial errors. Discussion Paper 2011-134, Tilburg University, Center for Economic Research.

Clerides, S. K., Lach, S., and Tybout, J. R. (1998). Is learning by exporting important? micro-dynamic evidence from colombia, mexico, and morocco. *The Quarterly Journal of Economics*, 113(3):903–947.

Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, 92(1):1–45.

de Brauw, A. and Giles, J. (2008). Migrant labor markets and the welfare of rural households in the developing world: Evidence from china. 2008 Annual Meeting, July 27-29, 2008, Orlando, Florida 6085, American Agricultural Economics Association.

Donald, S. G., Imbens, G. W., and Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics*, 152(1):28–36.

Elhorst, P. J. (2005). Unconditional maximum likelihood estimation of linear and log-linear dynamic models for spatial panels. *Geographical Analysis*, 37(1):85–106.

Hahn, J. (1997). Efficient estimation of panel data models with sequential moment restrictions. *Journal of Econometrics*, 79(1):1–21.

Hausman, J., Hall, B. H., and Griliches, Z. (1984). Econometric models for count data with an application to the patents-r & d relationship. *Econometrica*, 52(4):909–938.

Hsiao, C., Pesaran, H. M., and Tahmiscioglu, K. A. (2002). Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics*, 109(1):107–150.

Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics*, 150(1):86–98.

Jenish, N. and Prucha, I. R. (2012). On spatial processes and asymptotic inference under near-epoch dependence. *Journal of Econometrics*, 170(1):178–190.

Kim, M. S. and Sun, Y. (2011). Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix. *Journal of Econometrics*, 160(2):349–371.

Kitazawa, Y. (2007). Some additional moment conditions for a dynamic count panel data model.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85.

Mutl, J. (2006). Dynamic panel data models with spatially correlated disturbances. *University of Maryland Theses and Dissertations*.

Newey, W. K. and McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden, editor, *Handbook of Econometrics*, volume Volume 4, pages 2111–2245. Elsevier.

Newey, W. K. and Windmeijer, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719.

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(03):406–413.

Su, L. and Yang, Z. (2013). QML estimation of dynamic panel data models with spatial errors. *Research Collection School of Economics (Open Access)*.

Todd, P. E. and Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485):F3–F33.

Topalova, P. and Khandelwal, A. (2010). Trade liberalization and firm productivity: The case of india. *Review of Economics and Statistics*, 93(3):995–1009.

White, H. (2001). *Asymptotic theory for econometricians*. Academic Press, San Diego.

Windmeijer, F. (2000). Moment conditions for fixed effects count data models with endogenous regressors. *Economics Letters*, 68(1):21–24.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics*, 126(1):25–51.

Windmeijer, F. (2008). GMM for panel data count models. In Matyas, L. and Sevestre, P., editors, *The Econometrics of Panel Data*, number 46 in Advanced Studies in Theoretical and Applied Econometrics, pages 603–624. Springer Berlin Heidelberg.

Wooldridge, J. M. (1997). Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory*, 13(5):667–678.

Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics*, 90(1):77–97.

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

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