DEUTSCHE BUNDESBANK
EUROSYSTEM

Discussion Paper

Deutsche Bundesbank
No 25/2013

Estimation of linear dynamic panel data models with time-invariant regressors

Sebastian Kripfganz
(Goethe University Frankfurt)

Claudia Schwarz
(Goethe University Frankfurt and Deutsche Bundesbank)

Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main,
Postfach 10 06 02, 60006 Frankfurt am Main

Tel +49 69 9566-0

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax +49 69 9566-3077

Internet http://www.bundesbank.de

ISBN 978–3–86558–931–6 (Printversion)
ISBN 978–3–86558–932–3 (Internetversion)

Non-technical summary

Panel data comprises of cross-sectional units, e.g. countries or households, observed at different points in time. The combination of cross-sectional and time series data allows for more accurate conclusions and reduces statistical problems. In addition, dynamic adjustment processes can be analyzed for a broad base of cross-sectional units. A model is called dynamic if past observations of the variable of interest can influence the current value. Macroeconomic growth regressions and microeconomic wage regressions are examples where dynamic panel data models are used.

This paper analyzes the identification of effects of time-invariant regressors in dynamic panel data models as the methods currently used can be very imprecise. Time-invariant regressors play an important role in many empirical applications but estimation of the effects is non-trivial because there are various statistical problems that may arise. We propose a two-stage estimation procedure. A major advantage of the two-stage approach is that misspecified assumptions on the time-invariant regressors do not influence the estimation results for the time-varying variables. In extensive simulation studies we show that the currently most widely used estimation method, the generalized method of moments, can be quite biased whereas our method provides more precise and robust results. Furthermore, we develop an error correction term for the standard errors of the second stage. Neglecting the correction term can generate misleading implications.

To illustrate these methods we estimate a dynamic model with data from the Panel Study of Income Dynamics, a US household survey. As explanatory variables, among others, the past realization of income, work experience and level of education are used. The latter is a time-invariant variable. The results demonstrate the importance of choosing an adequate estimation method and of using the standard error correction term developed in this paper.

Nicht-technische Zusammenfassung

Paneldatensätze enthalten Daten für verschiedene Beobachtungseinheiten, z.B. Länder oder Haushalte, und für mehrere aufeinanderfolgende Zeitpunkte. Durch die Kombination von Querschnitts- und Zeitreihendaten lassen sich präzisere Aussagen über ökonomische Zusammenhänge machen und statistische Probleme verringern. Anhand von Paneldatensätzen kann man vor allem Modelle dynamischer Anpassungsprozesse besonders gut empirisch überprüfen. Ein Modell wird als dynamisches Modell bezeichnet, wenn vergangene Werte der analysierten Variable Einfluss auf den gegenwärtigen Wert haben können. Anwendungsbeispiele umfassen die makroökonomische Analyse dynamischer Wachstumsprozesse sowie die mikroökonomische Untersuchung von Einkommensentwicklungen.

In dieser Arbeit befassen wir uns mit der Identifizierung der Effekte von im Zeitablauf konstanten Einflussfaktoren in dynamischen Panelmodellen, da die momentan gängigen Schätzmethoden diese nur sehr ungenau schätzen können. Durch die Zeitinvarianz nehmen diese Variablen eine Sonderstellung ein, da es bei der Schätzung zu vielfältigen Arten von statistischen Problemen kommen kann. Wir entwickeln ein zweistufiges Schätzverfahren, mit dem sich sicherstellen lässt, dass fehlerhafte Annahmen bezüglich dieser zeitinvarianten Modellkomponenten nicht zu Verzerrungen bei der Bestimmung der Effekte von zeitvariierenden Einflussfaktoren führen. Anhand umfangreicher Simulationen zeigen wir dabei auf, dass das derzeit am weitesten verbreitete Schätzverfahren, die verallgemeinerte Momentenmethode, große Verzerrungen zur Folge haben kann. Die von uns vorgeschlagene Methode liefert hingegen genauere und robustere Ergebnisse. Außerdem bestimmen wir für das zweistufige Verfahren einen Korrekturterm für die Standardfehler der zweiten Stufe, dessen Nichtberücksichtigung zu falschen statistischen Schlussfolgerungen führen kann.

Als Anwendungsbeispiel schätzen wir ein dynamisches Modell zur Erklärung von Gehaltsunterschieden auf der Grundlage US-amerikanischer Haushaltsumfragedaten. Als erklärende Variablen werden neben dem Gehalt in der Vorperiode insbesondere die bisherige Berufserfahrung sowie der Bildungsgrad der Arbeitnehmer berücksichtigt. Der Bildungsgrad ist dabei innerhalb unserer Stichprobe eine zeitinvariante Variable. Die Ergebnisse unterstreichen die Bedeutung einer geeigneten Methodenauswahl sowie der korrekten Berechnung der Standardfehler im zweistufigen Schätzansatz.

Bundesbank Discussion Paper No 25/2013

Estimation of Linear Dynamic Panel Data Models with Time-Invariant Regressors*

Sebastian Kripfganz† Claudia Schwarz‡

This Version: May 6, 2013

Abstract

This paper considers estimation methods and inference for linear dynamic panel data models with unit-specific heterogeneity and a short time dimension. In particular, we focus on the identification of the coefficients of time-invariant variables in a dynamic version of the Hausman and Taylor (1981) model. We propose a two-stage estimation procedure to identify the effects of time-invariant regressors. We first estimate the coefficients of the time-varying regressors and subsequently regress the first-stage residuals on the time-invariant regressors to recover the coefficients of the latter. Standard errors are adjusted to take into account the first-stage estimation uncertainty. As potential first-stage estimators we discuss generalized method of moments estimators and the transformed likelihood approach of Hsiao, Pesaran, and Tahmiscioglu (2002). Monte Carlo experiments are used to compare the performance of the two-stage approach to various system GMM estimators that obtain all parameter estimates simultaneously. The results are in favor of the two-stage approach. We provide further simulation evidence that GMM estimators with a large number of instruments can be severely biased in finite samples. Reducing the instrument count by collapsing the instrument matrices strongly improves the results while restricting the lag depth does not. Finally, we estimate a dynamic Mincer equation with data from the Panel Study of Income Dynamics to illustrate the approach.

Keywords: System GMM; Instrument proliferation; Maximum likelihood; Two-stage estimation; Monte Carlo simulation; Dynamic Mincer equation

JEL Classification: C13; C23; J30

---

*We thank Michael Binder, Horst Entorf, Angela Fiedler, Georgios Georgiadis, Mehdi Hosseinkouchack, and Christian Schlag for support and helpful discussions. We are also grateful for comments from Richard Blundell, Maurice Bun, Jan Kiviet, Frank Windmeijer, and other participants at the Econometric Society Australasian and European meetings 2012 and the Warsaw International Economic Meeting 2012, as well as seminar participants at Goethe University Frankfurt, the World Bank, and Deutsche Bundesbank. Conference travel support from *Vereinigung von Freunden und Förderern der Goethe-Universität*, University of Warsaw, and German Academic Exchange Service is gratefully acknowledged.

†Corresponding author. Goethe University Frankfurt, Department of Money and Macroeconomics, Grueneburgplatz 1, 60323 Frankfurt am Main, Germany. kripfganz@wiwi.uni-frankfurt.de

‡Goethe University Frankfurt, Department of Finance, Grueneburgplatz 1, 60323 Frankfurt am Main, Germany; and Deutsche Bundesbank, Makroprudenzielle Analysen, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main. claudia.schwarz@bundesbank.de

1 Introduction

This paper considers estimation methods and inference for linear dynamic panel data models with a short time dimension. In particular, we focus on the identification of coefficients of time-invariant variables in the presence of unobserved unit-specific effects. In many empirical applications time-invariant variables play an important role in structural equations. In labor economics researchers are interested in the effects of gender, nationality, ethnic and religious background, or other time-invariant characteristics on the evolution of wages but would still like to control for unobserved time-invariant individual-specific effects such as worker's ability. As a recent example, Andini (2010b) estimates a dynamic version of the Mincer equation controlling for a rich set of time-invariant characteristics. In macroeconomic cross-country studies institutional features or group-level effects play a role in explaining economic development. For example, Hoeffler (2002) studies the growth performance of Sub-Saharan Africa countries by introducing a regional dummy variable in her dynamic panel data model. Cinyabuguma and Putterman (2011) focus on within Sub-Saharan differences by adding socio-economic and geographic factors to the analysis.

If there is unobserved unit-specific heterogeneity, it is often hard to disentangle the effects of the observed and the unobserved time-invariant heterogeneity. Standard fixed and random effects estimators cannot be used because of multicollinearity problems and, when the time dimension is short, the familiar Nickell (1981) bias in dynamic panel data models first discovered by Hurwicz (1950) for time series models. Therefore, it is common practice in empirical work to apply the generalized method of moments (GMM) framework proposed by Arellano and Bond (1991), Arellano and Bover (1995), and Blundell and Bond (1998), amongst others. However, as Binder et al. (2005) and Bun and Windmeijer (2010) emphasize, GMM estimators might suffer from a weak instruments problem when the autoregressive parameter approaches unity or when the variance of the unobserved unit-specific effects is large. Moreover, the number of instruments can rapidly become large relative to the sample size. The consequences of instrument proliferation, summarized by Roodman (2009), range from biased coefficient and standard error estimates to weakened specification tests.

In order to overcome the weak instruments problem in the context of estimating the effects of time-varying regressors, Hsiao et al. (2002) propose a transformed likelihood approach that is based on the model in first differences. A shortcoming of this approach is the inability to estimate the coefficients of time-invariant regressors. In this paper, we propose a two-stage estimation procedure to identify the latter. In the first stage, we estimate the coefficients of the time-varying regressors. Subsequently, we regress the first-stage residuals on the time-invariant regressors.[1] We achieve identification by us-

[1]For a static model, Plümper and Troeger (2007) propose a similar three-stage approach that they label fixed effects vector decomposition (FEVD). Their first stage is a classical fixed effects regression. In a recent symposium on the FEVD method, Breusch et al. (2011) and Greene (2011) show that the first two stages can be characterized by an instrumental variable estimation with a particular choice of

ing instrumental variables in the spirit of Hausman and Taylor (1981), and adjust the second-stage standard errors to account for the first-stage estimation error. Our methodology applies to any first-stage estimator that consistently estimates the coefficients of the time-varying variables without relying on coefficient estimates for the time-invariant regressors. As potential first-stage candidates we discuss the quasi-maximum likelihood (QML) estimator of Hsiao et al. (2002) as well as GMM estimators. A major advantage of the two-stage approach is the invariance of the first-stage estimates to misspecifications regarding the model assumptions on the correlation between the time-invariant regressors and the unobserved unit-specific effects.[2]

We perform Monte Carlo experiments to evaluate the finite sample performance in terms of bias, root mean square error (RMSE), and size statistics of our two-stage procedure relative to various GMM estimators that estimate all coefficients simultaneously. The results suggest that the two-stage approach is to be preferred when the researcher is interested both in the coefficients of time-varying and time-invariant variables. However, the quality of the second-stage estimates depends crucially on the precision of the first-stage estimates. Among our first-stage candidates the two-stage QML estimator performs very well when the time-varying regressors (besides the lagged dependent variable) are strictly exogenous. GMM estimators can be an alternative if effective measures are taken to avoid instrument proliferation. Our Monte Carlo analysis unveils sizable finite sample biases when the GMM instruments are based on the full set of available moment conditions. An easy way of reducing the instrument count is a restriction of the lag depth in the formation of the instrument matrices. However, our simulation results suggest that this does not solve the problem because the efficiency loss of disregarding relevant information outweighs the benefits of a more parsimonious instrument set. In contrast, collapsing the instrument matrices by forming linear combinations of the initial instruments improves the finite sample results considerably. Finally, in contrast to conventionally computed standard errors our adjusted second-stage standard errors can account remarkably well for the first-stage estimation error.

To illustrate these methods we estimate a dynamic Mincer equation with data from the Panel Study of Income Dynamics (PSID). We use Hausman and Taylor (1981)-type instruments to deal with the endogeneity of the schooling variable that is assumed to be correlated with unobserved individual-specific ability. In our sample of salaried workers, education is a time-invariant variable. To identify the return to schooling we use the level of the time-varying variables as instruments, in particular the industry dummy variables. Compared with the non-instrumented case, the return to schooling is sizably reduced. Moreover, the correct adjustment of the second-stage standard errors proves to be important for valid inference.

instruments, and that the third stage is essentially meaningless. Because the FEVD is widely associated with the original three-stage approach of Plümper and Troeger (2007), we do not adopt this name here.

[2]Hoeffler (2002) and Cinyabuguma and Putterman (2011) argue similarly. They apply GMM estimation in the first stage, and ordinary least squares estimation in the second stage. However, they do not correct the second-stage standard errors.

The paper is organized as follows: Section 2 explains the model and the identification strategy. Section 3 lays out the two-stage estimation procedure to identify the coefficients of time-invariant regressors. Section 4 describes one- and two-stage GMM estimation, while Section 5 briefly describes two-stage QML estimation. Section 6 provides simulation evidence on the performance of the two-stage estimators in comparison to several one-stage GMM estimators under different scenarios. In Section 7 we discuss the empirical application of the methods discussed in this paper, and Section 8 concludes.

2 Model

Consider the dynamic panel data model with units $i = 1, 2, \ldots, N$, and a fixed number of time periods $t = 1, 2, \ldots, T$:[3]

$$y_{it} = \lambda y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{f}_i'\boldsymbol{\gamma} + e_{it}, \tag{1}$$

$$e_{it} = \alpha_i + u_{it}, \tag{2}$$

where $\mathbf{x}_{it}$ is a $K_x \times 1$ vector of time-varying variables. The initial observations of the dependent variable, y_{i0}, and the regressors, $\mathbf{x}_{i0}$, are assumed to be observed. $\mathbf{f}_i$ is a $K_f \times 1$ vector of observed time-invariant variables that includes an overall regression constant,[4] and α_i is an unobserved unit-specific effect of the i-th cross section. In a strict sense, α_i is called a fixed effect if it is allowed to be correlated with all of the regressor variables $\mathbf{x}_{it}$ and $\mathbf{f}_i$,[5] and it is a random effect if it is independently distributed. In this paper we look at a hybrid (or intermediate case) of the dynamic fixed and random effects models where some of the regressors are correlated with α_i but not all of them.[6] Throughout the paper we maintain the following assumptions:[7]

Assumption 1: The disturbances u_{it} are independently and identically distributed for all i and t with $E[u_{it}] = 0$, $E[u_{is}u_{it}] = 0 \; \forall s \neq t$, and $E[u_{it}^2] = \sigma_u^2$.

Assumption 2: The unit-specific effects α_i are independently distributed from the disturbances u_{it} with $E[\alpha_i] = 0$ and $E[\alpha_i^2] = \sigma_\alpha^2$, and $E[\alpha_i u_{it}] = 0$.

Identification of the (structural) parameters λ, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ now crucially hinges on the assumptions about the dependencies between the regressors and the unit-specific effects.

[3]The model can be generalized to the inclusion of higher order lags of the dependent variable as well as distributed lags of the time-varying exogenous regressors.

[4]Without loss of generality, set the first entry of $\mathbf{f}_i$ equal to 1 for all i.

[5]Note that α_i is correlated with the lagged dependent variable by construction.

[6]Compare Arellano and Bover (1995) and Greene (2011) in the context of a static panel data model.

[7]The assumptions on homoscedasticity of u_{it} and α_i can be relaxed but we stick to it for ease of exposition.

Assumption 3: The explanatory variables can be decomposed as $\mathbf{x}_{it} = (\mathbf{x}'_{1it}, \mathbf{x}'_{2it})'$ and $\mathbf{f}_i = (\mathbf{f}'_{1i}, \mathbf{f}'_{2i})'$ such that $E[\alpha_i|\mathbf{x}_{1it}, \mathbf{f}_{1i}] = 0$, $E[\alpha_i|\mathbf{x}_{2it}] \neq 0$ and $E[\alpha_i|\mathbf{f}_{2i}] \neq 0$.

The resulting model is the dynamic counterpart of the Hausman and Taylor (1981) model. For further reference, the lengths of the subvectors are K_{x1}, K_{x2}, K_{f1}, and K_{f2}, respectively.[8] Accordingly, the parameter vectors are partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}'_1, \boldsymbol{\gamma}'_2)'$. If $K_{x2} = K_{f2} = 0$ the model collapses to the dynamic random effects model. Contrarily, $K_{x1} = 0$ and $K_{f1} = 1$ (the constant term) leads to the dynamic fixed effects model.

For the static model ($\lambda = 0$) with strictly exogenous regressors $\mathbf{x}_{it}$, Hausman and Taylor (1981) propose an instrumental variable estimator that uses deviations from their within-group means, $\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$, as instruments for the regressors $\mathbf{x}_{it}$, and the within-group means $\bar{\mathbf{x}}_{1i}$ as instruments for $\mathbf{f}_{2i}$. The full set of instruments is $\mathbf{z}_{it} = (\ddot{\mathbf{x}}'_{it}, \bar{\mathbf{x}}'_{1i}, \mathbf{f}'_{1i})'$. To improve on the efficiency of the estimator, Amemiya and MaCurdy (1986) propose to use all time periods of $\mathbf{x}_{1it}$ separately as instruments instead of the within-group means such that $\mathbf{z}_{it} = (\ddot{\mathbf{x}}'_{it}, \mathbf{x}'_{1i1}, \ldots, \mathbf{x}'_{1iT}, \mathbf{f}'_{1i})'$. Breusch et al. (1989) additionally suggest to use each individual deviation from the within-group means as a separate instrument. Thus, $\mathbf{z}_{it} = (\ddot{\mathbf{x}}'_{i1}, \ldots, \ddot{\mathbf{x}}'_{iT}, \mathbf{x}'_{1i1}, \ldots, \mathbf{x}'_{1iT}, \mathbf{f}'_{1i})'$. Furthermore, excluded exogenous instruments might be available. This approach requires $K_{f2} \leq (K_z - K_x - K_{f1})$ for the parameters $\boldsymbol{\gamma}_2$ to be at least just identified, where K_z is the total number of instruments. With appropriate instruments for the lagged dependent variable, this approach directly extends to the dynamic model. We then need $K_{f2} \leq (K_z - 1 - K_x - K_{f1})$ to achieve identification of γ_2.

In some applications, time-invariant regressors do not emerge directly from a theoretical model but from an attempt to obtain a dynamic random effects model without assuming $K_{x2} = 0$ from the outset. Mundlak (1978) proposes to model the latent effects as an affine function of the within-group means of the time-varying regressors:

$$\alpha_i = b + \bar{\mathbf{x}}'_i \boldsymbol{\pi} + \eta_i, \tag{3}$$

with $E[\eta_i|\mathbf{x}_{it}] = 0$. Similarly, Chamberlain (1982) proposes to project the unobserved effects on all elements of the time-varying regressors $\mathbf{x}_{it}$ instead of the within-group means:

$$\alpha_i = b + \sum_{s=0}^{T} \mathbf{x}'_{is} \boldsymbol{\pi}_s + \eta_i. \tag{4}$$

Consequently, we obtain a representation of model (1) with $\mathbf{f}_i = (1, \bar{\mathbf{x}}'_i)$ in case of projection (3), and $\mathbf{f}_i = (1, \mathbf{x}'_{i0}, \mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iT})$ for projection (4). Moreover, for the transformed model $K_{x2} = K_{f2} = 0$. This approach, however, usually does not work if time-invariant regressors are already present in the structural model equation. To illustrate this case, assume that projection (3) is applied when the model includes regressors $\mathbf{f}_{1i}$. As a consequence, $E[\eta_i|\mathbf{f}_{1i}] \neq 0$ although $E[\alpha_i|\mathbf{f}_{1i}] = 0$ unless another restrictive condition holds.

[8] Consequently, $K_{x1} + K_{x2} = K_x$ and $K_{f1} + K_{f2} = K_f$.

To see this, take the conditional expectation of (3) with respect to $\mathbf{f}_{1i}$:

$$\begin{aligned} E[\alpha_i|\mathbf{f}_{1i}] &= b + E[\bar{\mathbf{x}}_i|\mathbf{f}_{1i}]'\boldsymbol{\pi} + E[\eta_i|\mathbf{f}_{1i}] \\ &= (E[\bar{\mathbf{x}}_i|\mathbf{f}_{1i}] - E[\bar{\mathbf{x}}_i])'\,\boldsymbol{\pi} + E[\eta_i|\mathbf{f}_{1i}]. \end{aligned} \tag{5}$$

Consequently, $E[\eta_i|\mathbf{f}_{1i}]$ can only be zero for any $\boldsymbol{\pi}$ if $E[\bar{\mathbf{x}}_i|\mathbf{f}_{1i}] = E[\bar{\mathbf{x}}_i]$.[9]

Another example of time-invariant variables that emerge from econometric considerations are cluster-specific effects. Without loss of generality, let us define clusters $C_j = \{i|N_{j-1} < i \leq N_j\}, j = 1, 2, \ldots, J$, with $N_0 = 0$, $1 \leq N_1 < N_2 < \ldots < N_J$, and $N_J = N$ for appropriately ordered units i. The affiliation of the units to the clusters is non-random and known. Moreover, $J \ll N$ and $J/N \to 0$ as $N \to \infty$ to avoid the incidental parameters problem.[10] Then, we decompose the unit-specific effects α_i into cluster-specific effects c_j and a random component η_i:[11]

$$\alpha_i = c_j + \eta_i, \quad i \in C_j, \tag{6}$$

such that $E[\eta_i|c_j] = 0$. When the size of the cluster effects is of interest to get a sense of the cross-cluster heterogeneity that is unexplained by the remaining regressors, we add $J-1$ cluster dummy variables to the set of time-invariant regressors.[12]

In the remaining sections, we distinguish between weakly and strictly exogenous regressors $\mathbf{x}_{it}$ with respect to the disturbance term u_{it}.[13]

Assumption 4.1: The time-varying regressors $\mathbf{x}_{it}$ are strictly exogenous with respect to the disturbances u_{it}: $E[u_{it}|\mathbf{x}_{i0}, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}] = 0$.

Assumption 4.2: The time-varying regressors $\mathbf{x}_{it}$ are weakly exogenous with respect to the disturbances u_{it}: $E[u_{it}|\mathbf{x}_{i0}, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}] = 0$ and $E[u_{it}|\mathbf{x}_{is}] \neq 0\ \forall s > t$.

In addition, we assume:

Assumption 5: $E[u_{it}|\mathbf{f}_i] = 0$.

[9] A trivial example for this condition being satisfied is when the only time-invariant variable is the regression constant, that is $f_i = 1$. Note that its coefficient is $\gamma + b$.

[10] When $J/N \to \kappa$ with $\kappa \neq 0$, the number of parameters to be estimated increases at the same rate as the sample size. This leads to the familiar incidental parameters problem discussed by Neyman and Scott (1948).

[11] This is also a stylized way of introducing spatial dependence into the model.

[12] Recall that there is already an intercept term in $\mathbf{f}_{1i}$. Hoeffler (2002) investigates the slow growth performance of economies in Sub-Sahara Africa by including a regional dummy variable in an augmented Solow model. Implicitly, she applies projection (6) with two clusters, namely Sub-Saharan Africa on the one side and the remaining countries in her sample on the other side.

[13] We do not explicitly treat the case of a combination of strictly and weakly exogenous regressors as the necessary adjustments are straightforward.

To facilitate the subsequent derivations we introduce the following notation. Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, $\mathbf{y}_{i,-1} = (y_{i0}, y_{i1}, \ldots, y_{i,T-1})'$, $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iT})'$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})'$, $\mathbf{W}_i = (\mathbf{y}_{i,-1}, \mathbf{X}_i)$, $\mathbf{F}_i = \mathbf{f}_i' \otimes \boldsymbol{\iota}_T$, where $\otimes$ denotes the Kronecker product and $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones, and $\boldsymbol{\theta} = (\lambda, \boldsymbol{\beta}')'$. Then, model (1) can be written in stacked form as:

$$\mathbf{y}_i = \mathbf{W}_i\boldsymbol{\theta} + \mathbf{F}_i\boldsymbol{\gamma} + \alpha_i\boldsymbol{\iota}_T + \mathbf{u}_i. \tag{7}$$

The corresponding model in first differences is:

$$\Delta\mathbf{y}_i = \Delta\mathbf{W}_i\boldsymbol{\theta} + \Delta\mathbf{u}_i, \tag{8}$$

where $\Delta\mathbf{y}_i = \mathbf{D}\mathbf{y}_i$, $\Delta\mathbf{W}_i = \mathbf{D}\mathbf{W}_i$, and $\Delta\mathbf{u}_i = \mathbf{D}\mathbf{u}_i$ for the $(T-1) \times T$ transformation matrix $\mathbf{D} = [(\mathbf{0}, \mathbf{I}_{T-1}) - (\mathbf{I}_{T-1}, \mathbf{0})]$, where $\mathbf{I}_{T-1}$ is the identity matrix of dimension $(T-1)$. Obviously, this transformation removes all time-invariant components. To further ease the notational burden, let $\mathbf{y}_i^* = (y_{i0}, \mathbf{y}_i')'$, $\mathbf{X}_i^* = (\mathbf{x}_{i0}, \mathbf{X}_i')'$, and accordingly $\Delta\mathbf{y}_i^* = (\Delta y_{i1}, \Delta\mathbf{y}_i')'$, and $\Delta\mathbf{X}_i^* = (\Delta\mathbf{x}_{i1}, \Delta\mathbf{X}_i')'$. Unit-specific time means are denoted with a bar, for example $\bar{\mathbf{x}}_i^* = (T+1)^{-1}(\mathbf{X}_i^*)'\boldsymbol{\iota}_{T+1}$. When we stack the data for all units below each other, for example $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_N')'$, we leave out the subscripts.

Finally, under assumptions 1 and 2, the asymptotic variance-covariance matrix of the residuals $\mathbf{e}_i = \alpha_i\boldsymbol{\iota}_T + \mathbf{u}_i$ is $\Omega^l = \sigma_\alpha^2\boldsymbol{\iota}_T\boldsymbol{\iota}_T' + \sigma_u^2\mathbf{I}_T$. Consequently, the asymptotic variance-covariance matrix of the first-differenced disturbances $\Delta\mathbf{u}_i$ is:

$$\Omega^d = \sigma_u^2\mathbf{D}\mathbf{D}' = \sigma_u^2 \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & & \\ 0 & -1 & 2 & & \\ \vdots & & & \ddots & -1 \\ 0 & & & -1 & 2 \end{pmatrix}.$$

3 Two-Stage Estimation

The contribution of this paper is to focus on the estimation of the coefficients $\boldsymbol{\gamma}$ of the time-invariant variables. In the next section, we show how GMM estimators can be adjusted to the assumed correlation structure. However, since all regression coefficients are recovered simultaneously, a misspecification of the moment conditions might lead to a biased estimation of all coefficients including λ and $\boldsymbol{\beta}$. We therefore lay down a two-stage estimation procedure. In a first stage, we subsume the time-invariant variables $\mathbf{f}_i$ under the unit-specific effects, $\alpha_i^* = \alpha_i + \mathbf{f}_i'\boldsymbol{\gamma}$, and consistently estimate the coefficients λ and $\boldsymbol{\beta}$ independent of the assumptions on the correlation structure between $\mathbf{f}_i$ and α_i. In the second stage, we recover γ.

The first-stage model is

$$y_{it} = \alpha^* + \lambda y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + e_{it}^*, \tag{9}$$

$$e_{it}^* = \alpha_i^* - \alpha^* + u_{it}, \tag{10}$$

where $\alpha^* = E[\mathbf{f}_i']\boldsymbol{\gamma}$. To obtain the first-stage estimates $\hat{\lambda}$ and $\hat{\boldsymbol{\beta}}$ we can apply a transformation that eliminates the time-invariant unit-specific effects α_i^*. In particular, the GMM estimator of Arellano and Bond (1991) and the QML estimator of Hsiao et al. (2002) are based on the first-differenced model (8) while Arellano and Bover (1995) propose a GMM estimator based on forward orthogonal deviations. Alternatively, system GMM estimators that also make use of the level relationship can be applied taking into account that the error term of the first-stage model is e_{it}^* instead of e_{it}. This distinction is important if $K_{x1} > 0$ but some or all of the variables in $\mathbf{x}_{1it}$ are correlated with $\mathbf{f}_i$. These variables are uncorrelated with α_i but not α_i^*.

In the second stage, we estimate the coefficients $\boldsymbol{\gamma}$ of the time-invariant variables based on the cross-sectional relationship in the level equation:

$$\hat{r}_{iT} = \mathbf{f}_i'\boldsymbol{\gamma} + v_{iT}, \quad i = 1, 2, \ldots, N, \tag{11}$$

where

$$\hat{r}_{iT} = y_{iT} - \hat{\lambda} y_{i,T-1} - \mathbf{x}_{iT}'\hat{\boldsymbol{\beta}}, \tag{12}$$

$$v_{iT} = \alpha_i + u_{iT} - (\hat{\lambda} - \lambda) y_{i,T-1} - \mathbf{x}_{iT}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{13}$$

In particular, note the two additional terms in (13) that are due to the first-stage estimation error. Under assumption 3, the regressors $\mathbf{f}_{2i}$ are endogenous in the regression model (11). Hausman and Taylor (1981) suggest to use the within-group means $\bar{\mathbf{x}}_{1i}^*$ as instruments for these endogenous time-invariant regressors. Consequently, the $K_{x1} + K_{f1}$ instruments for the second-stage regression would be $\tilde{\mathbf{z}}_i = ((\bar{\mathbf{x}}_{1i}^*)', \mathbf{f}_{1i}')'$. Following Amemiya and MaCurdy (1986) and Breusch et al. (1989), a more efficient instrumental variable estimator makes use of all individual time periods such that $K_{x1}(T+1) + K_{f1}$ instruments are available, namely $\tilde{\mathbf{z}}_i = (\mathbf{x}_{1i0}', \ldots, \mathbf{x}_{1iT}', \mathbf{f}_{1i}')'$. Note that under assumption 4.2 both sets of instruments are only valid when we base the second-stage model (11) on the cross-section in period T. For any other generic time period t the available instrument set shrinks to $\tilde{\mathbf{z}}_i = (\mathbf{x}_{1i0}', \ldots, \mathbf{x}_{1it}', \mathbf{f}_{1i}')'$ and the within-group averages are no longer valid instruments.[14] Identification of $\boldsymbol{\gamma}_2$ requires that the number of instruments is greater or equal than K_f. The 2SLS estimator is given by:

$$\hat{\hat{\boldsymbol{\gamma}}} = \left[\left(\sum_{i=1}^N \mathbf{f}_i \tilde{\mathbf{z}}_i' \right) \left(\sum_{i=1}^N \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i' \right)^{-1} \left(\sum_{i=1}^N \tilde{\mathbf{z}}_i \mathbf{f}_i' \right) \right]^{-1} \left(\sum_{i=1}^N \mathbf{f}_i \tilde{\mathbf{z}}_i' \right) \left(\sum_{i=1}^N \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i' \right)^{-1} \sum_{i=1}^N \tilde{\mathbf{z}}_i \hat{r}_{iT}. \tag{14}$$

It is easily seen that $E[\alpha_i + u_{iT} | \tilde{\mathbf{z}}_i] = 0$ together with consistency of the first-stage

[14] Another choice might be to regress the within-group means $\hat{\bar{r}}_i$ on $\mathbf{f}_i$. However, under assumption 4.2 this attempt shrinks the available instrument set to $\tilde{\mathbf{z}}_i = (\mathbf{x}_{1i0}', \mathbf{x}_{1i1}', \mathbf{f}_{1i}')'$.

estimator $\hat{\boldsymbol{\theta}}$ imply consistency of the second-stage estimator:

$$\begin{aligned}\operatorname{plim} \hat{\hat{\boldsymbol{\gamma}}} &= \boldsymbol{\gamma} + (\mathbf{A}_1\mathbf{A}_2^{-1}\mathbf{A}_1')^{-1}\mathbf{A}_1\mathbf{A}_2^{-1} \operatorname{plim} \frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{z}}_i v_{iT} \\ &= \boldsymbol{\gamma} + (\mathbf{A}_1\mathbf{A}_2^{-1}\mathbf{A}_1')^{-1}\mathbf{A}_1\mathbf{A}_2^{-1} \left[\operatorname{plim} \frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{z}}_i(\alpha_i + u_{it}) - \mathbf{A}_3 \operatorname{plim}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right] \\ &= \boldsymbol{\gamma}, \end{aligned} \tag{15}$$

where

$$\mathbf{A}_1 = \operatorname{plim} \frac{1}{N}\sum_{i=1}^{N} \mathbf{f}_i\tilde{\mathbf{z}}_i', \quad \mathbf{A}_2 = \operatorname{plim} \frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i', \quad \mathbf{A}_3 = \operatorname{plim} \frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{z}}_i\mathbf{w}_{iT}',$$

with $\mathbf{w}_{iT} = (y_{i,T-1}, \mathbf{x}_{iT}')'$.

Importantly, the error term v_{iT} of the second-stage regression is cross-sectionally correlated and exhibits heteroscedasticity due to the presence of the estimation error in the first-stage coefficients. Therefore, the conventional standard errors obtained from OLS or 2SLS estimation are inconsistent and lead to invalid inferences. However, this does not affect the consistency of the second-stage estimator $\hat{\hat{\boldsymbol{\gamma}}}$. The asymptotic distribution of the second-stage estimator (14) is determined by the components of v_{iT}:

$$\begin{aligned}\sqrt{N}(\hat{\hat{\boldsymbol{\gamma}}} - \boldsymbol{\gamma}) &= \left[\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{f}_i\tilde{\mathbf{z}}_i'\right)\left(\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{z}}_i\mathbf{f}_i'\right)\right]^{-1} \\ &\quad \times \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{f}_i\tilde{\mathbf{z}}_i'\right)\left(\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i'\right)^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \tilde{\mathbf{z}}_i v_{iT}, \end{aligned} \tag{16}$$

where

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \tilde{\mathbf{z}}_i v_{iT} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \tilde{\mathbf{z}}_i(\alpha_i + u_{iT}) - \left(\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{z}}_i\mathbf{w}_{iT}'\right)\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{17}$$

By applying the central limit theorem, we can now establish the joint asymptotic distribution of the first-stage and the second-stage estimators:

$$\sqrt{N}\begin{pmatrix}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \\ \hat{\hat{\boldsymbol{\gamma}}} - \boldsymbol{\gamma}\end{pmatrix} \overset{a}{\sim} \mathcal{N}\left(\begin{pmatrix}\mathbf{0} \\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\Sigma_\theta & \Sigma_{\theta\gamma} \\ \Sigma_{\theta\gamma}' & \Sigma_\gamma\end{pmatrix}\right), \tag{18}$$

with

$$\Sigma_{\theta\gamma} = (\Sigma_\theta\mathbf{A}_3' + \mathbf{C})\mathbf{B}_2', \tag{19}$$

$$\Sigma_\gamma = \mathbf{B}_2(\tilde{\mathbf{V}} + \mathbf{A}_3\Sigma_\theta\mathbf{A}_3' - \mathbf{C}'\mathbf{A}_3' - \mathbf{A}_3\mathbf{C})\mathbf{B}_2', \tag{20}$$

where $\mathbf{B}_1 = \mathbf{A}_1\mathbf{A}_2^{-1}\mathbf{A}_1'$, $\mathbf{B}_2 = \mathbf{B}_1^{-1}\mathbf{A}_1\mathbf{A}_2^{-1}$, and $\tilde{\mathbf{V}} = \text{plim}\, N^{-1}\sum_{i=1}^N \tilde{\mathbf{z}}_i v_{iT}^2 \tilde{\mathbf{z}}_i'$. Moreover, $\Sigma_\theta = Avar[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})]$ is the asymptotic variance-covariance matrix of the first-stage estimator, and $\mathbf{C} = Acov[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), N^{-1/2}\sum_{i=1}^N(\alpha_i + u_{iT})\tilde{\mathbf{z}}_i']$. Under homoscedasticity of α_i and u_{it} equation (20) simplifies to

$$\Sigma_\gamma = (\sigma_\alpha^2 + \sigma_u^2)\mathbf{B}_1^{-1} + \mathbf{B}_2(\mathbf{A}_3\Sigma_\theta\mathbf{A}_3' - \mathbf{C}'\mathbf{A}_3' - \mathbf{A}_3\mathbf{C})\mathbf{B}_2'. \tag{21}$$

We can estimate the asymptotic variance-covariance matrix of the second-stage estimator by calculating the respective sample analogs of the matrices $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$. An estimate for σ_u^2 is typically available from the first-stage regression but not for σ_α^2. However, a consistent estimate for $\sigma_e^2 = \sigma_\alpha^2 + \sigma_u^2$ can be obtained as follows:

$$\hat{\hat{\sigma}}_e^2 = \frac{1}{NT - (1 + K_x + K_f)}\sum_{i=1}^N\sum_{t=1}^T \hat{\hat{v}}_{it}^2, \tag{22}$$

with $\hat{\hat{v}}_{it}$ given by

$$\hat{\hat{v}}_{it} = y_{it} - \hat{\lambda}y_{i,t-1} - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}} - \mathbf{f}_i'\hat{\hat{\boldsymbol{\gamma}}}. \tag{23}$$

Note that $\hat{\hat{v}}_{it} = \hat{\hat{v}}_{iT} - \sum_{s=t+1}^T \widehat{\Delta u}_{is},\ t = 1, 2, \ldots, T-1$, where $\widehat{\Delta u}_{it}$ are the estimated first-stage residuals of the first-differenced model. An estimate of Σ_θ is readily available from the first-stage estimation. Estimation of the asymptotic covariance matrix $\mathbf{C}$ requires knowledge of the closed-form solution of the first-stage estimator $\hat{\boldsymbol{\theta}}$. We will derive an expression of $\mathbf{C}$ for the GMM and QML first-stage estimators that we discuss in the next sections.[15]

4 GMM Estimation

In this section, we discuss generalized method of moments estimation for linear dynamic panel data models that is based on moment conditions for the model in first differences, $E[(\mathbf{Z}_i^d)'\Delta\mathbf{u}_i] = \mathbf{0}$, and the model in levels, $E[(\mathbf{Z}_i^l)'\mathbf{e}_i] = \mathbf{0}$. In the following subsection, we discuss the moment conditions that are implied by the model assumptions and the resulting design of the instrument matrices $\mathbf{Z}_i^d$ and $\mathbf{Z}_i^l$. Subsection 4.2 derives the system GMM estimator that combines all moment conditions, and Subsection 4.3 discusses transformations of the instrument matrices to reduce the number of overidentifying restrictions. Finally, Subsection 4.4 derives the two-stage GMM estimator as a robust alternative to one-stage GMM estimators that obtain all coefficient estimates simultaneously.

[15]In practice, ignoring the variance components involving $\mathbf{C}$ should still yield a good approximation of the asymptotic variance-covariance matrix. In Monte Carlo simulations we find that this approximation works very well.

4.1 Moment Conditions

Following Arellano and Bond (1991) and Blundell et al. (2001), assumptions 1 and 2 imply the following $T(T-1)/2$ moment conditions for the model in first differences:

$$E[y_{i,t-s}\Delta u_{it}] = 0, \quad t = 2, 3, \ldots, T, \quad 2 \leq s \leq t. \tag{24}$$

Under the strict exogeneity assumption 4.1 we have another $K_x(T+1)(T-1)$ moment conditions:

$$E[\mathbf{x}_{is}\Delta u_{it}] = \mathbf{0}, \quad t = 2, 3, \ldots, T, \quad 0 \leq s \leq T. \tag{25}$$

In the case of weakly exogenous regressors, assumption 4.2, there are only the following $K_x(T+2)(T-1)/2$ moment conditions available:

$$E[\mathbf{x}_{i,t-s}\Delta u_{it}] = \mathbf{0}, \quad t = 2, 3, \ldots, T, \quad 1 \leq s \leq t. \tag{26}$$

At this stage, we do not need to make a distinction between regressors that are correlated and those that are uncorrelated with α_i. Following Arellano and Bover (1995), the presence of time-invariant regressors provides another $K_f(T-1)$ moment conditions:

$$E[\mathbf{f}_i\Delta u_{it}] = \mathbf{0}, \quad t = 2, 3, \ldots, T. \tag{27}$$

Under assumption 1, the disturbances u_{it} are homoscedastic through time. Following Ahn and Schmidt (1995), this implies another $(T-2)$ moment conditions:

$$E[y_{i,t-2}\Delta u_{i,t-1} - y_{i,t-1}\Delta u_{it}] = 0, \quad t = 3, \ldots, T. \tag{28}$$

We can combine these moment conditions for the first-differenced equation:

$$E[(\mathbf{Z}_i^d)'\Delta \mathbf{u}_i] = \mathbf{0}, \tag{29}$$

where $\mathbf{Z}_i^d = (\mathbf{Z}_{y,i}^d, \mathbf{Z}_{x,i}^d, \mathbf{I}_{T-1} \otimes \mathbf{f}_i', \mathbf{Z}_{u,i}^d)$ with

$$\mathbf{Z}_{y,i}^d = \begin{pmatrix} (\mathbf{z}_{y,i2}^d)' & 0 & \cdots & 0 \\ 0 & (\mathbf{z}_{y,i3}^d)' & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & (\mathbf{z}_{y,iT}^d)' \end{pmatrix}, \quad \mathbf{Z}_{x,i}^d = \begin{pmatrix} (\mathbf{z}_{x,i2}^d)' & 0 & \cdots & 0 \\ 0 & (\mathbf{z}_{x,i3}^d)' & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & (\mathbf{z}_{x,iT}^d)' \end{pmatrix},$$

$$\mathbf{Z}_{u,i}^d = \begin{pmatrix} y_{i1} & 0 & \cdots & 0 \\ -y_{i2} & y_{i2} & & \vdots \\ 0 & -y_{i,3} & \ddots & 0 \\ \vdots & & \ddots & y_{i,T-2} \\ 0 & \cdots & 0 & -y_{i,T-1} \end{pmatrix}$$

and $\mathbf{z}_{y,it}^d = (y_{i0}, y_{i1}, \ldots, y_{i,t-2})'$. The instruments $\mathbf{z}_{x,it}^d$ differ according to the assumption about the regressor variables. We have $\mathbf{z}_{x,it}^d = (\mathbf{x}_{i0}', \mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')'$ under strict exogeneity and $\mathbf{z}^d_{x,it} = (\mathbf{x}'_{i0}, \mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{i,t-1})'$ under weak exogeneity of the regressors. Based on the moment conditions (29) we can set up a first estimator that is often called "difference GMM" estimator. However, this estimator will generally be inefficient as it does not exploit all available information, and it cannot identify $\boldsymbol{\gamma}$. To add further moment conditions for the model in levels we will partly rely on the following assumption:

Assumption 6.1: $E[\Delta y_{i1}\alpha_i] = 0$, and $E[\Delta \mathbf{x}_{2it}\alpha_i] = 0,\ t = 1, 2, \ldots, T$.[16]

Under the additional assumption 6.1, Blundell and Bond (1998) establish the following $(T-1)$ linear moment conditions for the model in levels:

$$E[\Delta y_{i,t-1} e_{it}] = 0, \quad t = 2, 3, \ldots, T. \tag{30}$$

For the regressors $\mathbf{x}_{1it}$, Arellano and Bond (1991) introduce the following $K_{x1}(T+1)$ level moment conditions:

$$E[\mathbf{x}_{1i0} e_{i1}] = \mathbf{0}, \quad \text{and} \quad E[\mathbf{x}_{1it} e_{it}] = \mathbf{0}, \quad t = 1, 2, \ldots, T. \tag{31}$$

Moreover, Arellano and Bover (1995) and Blundell et al. (2001) introduce another $K_{x2}T$ moment conditions for the regressors $\mathbf{x}_{2it}$ under assumption 6.1:

$$E[\Delta \mathbf{x}_{2it} e_{it}] = \mathbf{0}, \quad t = 1, 2, \ldots, T. \tag{32}$$

Note that assumption 6.1 implies $E[(\mathbf{x}_{2it} - \bar{\mathbf{x}}^*_{2i})\alpha_i] = 0,\ t = 0, 1, \ldots, T$. Exploiting this relationship, we have K_{x2} additional moment conditions:[17]

$$E\left[(\mathbf{x}_{2iT} - \bar{\mathbf{x}}^*_{2i}) e_{iT}\right] = \mathbf{0}. \tag{33}$$

All remaining moment conditions for the model in levels are redundant for these variables. Arellano and Bover (1995) further suggest K_{f1} moment conditions for the time-invariant regressors $\mathbf{f}_{1i}$ that are uncorrelated with the unit-specific effects α_i:

$$E\left[\mathbf{f}_{1i} \sum_{t=1}^{T} e_{it}\right] = \mathbf{0}. \tag{34}$$

The additional level moment conditions $E[\mathbf{f}_{1i} e_{it}] = 0,\ t = 1, 2, \ldots, T$, are again redundant. Finally, we combine the level moment conditions:

$$E[(\mathbf{Z}^l_i)' \mathbf{e}_i] = \mathbf{0}, \tag{35}$$

[16]To guarantee that Δy_{it} and $\Delta \mathbf{x}_{2it}$ are uncorrelated with α_i a restriction on the initial conditions has to be satisfied. Deviations of y_{i0} and $\mathbf{x}_{2i0}$ from their long-run means must be uncorrelated with α_i. A sufficient but not necessary condition for assumption 6.1 to hold is joint mean stationarity of the processes y_{it} and $\mathbf{x}_{it}$. Moreover, $E[\Delta y_{it}\alpha_i] = 0,\ t = 2, 3, \ldots, T$, is implied by assumption 6.1. See Blundell and Bond (1998), Blundell et al. (2001), and Roodman (2009) for a discussion.

[17]We choose e_{iT} to make these moment conditions valid both under assumptions 4.1 and 4.2.

where $\mathbf{Z}_i^l = (\mathbf{Z}_{y,i}^l, \mathbf{Z}_{x,i}^l, \mathbf{F}_{1i})$, with

$$\mathbf{Z}_{y,i}^l = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \Delta y_{i1} & 0 & \cdots & 0 \\ 0 & \Delta y_{i2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \Delta y_{i,T-1} \end{pmatrix},$$

and

$$\mathbf{Z}_{x,i}^l = \begin{pmatrix} \mathbf{x}_{1i0}' & \mathbf{x}_{1i1}' & 0 & \cdots & 0 & \Delta\mathbf{x}_{2i1}' & 0 & \cdots & 0 & 0 \\ 0 & 0 & \mathbf{x}_{1i2}' & & \vdots & 0 & \Delta\mathbf{x}_{2i2}' & & \vdots & \vdots \\ \vdots & \vdots & & \ddots & 0 & \vdots & & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \mathbf{x}_{1iT}' & 0 & \cdots & 0 & \Delta\mathbf{x}_{2iT}' & \mathbf{x}_{2iT} - \bar{\mathbf{x}}_{2i}^* \end{pmatrix}.$$

Ahn and Schmidt (1995) derive further non-linear moment conditions under assumptions 1 and 2, namely $E[u_{it}\Delta u_{i,t-1}] = 0,\ t = 3, 4, \ldots, T$, and $E[\bar{u}_i \Delta u_{i2}] = 0$. The latter again results from homoscedasticity of u_{it}. In this paper, we restrict our attention to the linear moment conditions above.[18]

4.2 System GMM Estimator

The moment conditions for the two equations can be combined by considering a system of equations:

$$\mathbf{y}_i^+ = \mathbf{W}_i^+ \boldsymbol{\theta} + \mathbf{F}_i^+ \boldsymbol{\gamma} + \mathbf{e}_i^+, \tag{36}$$

where $\mathbf{y}_i^+ = \mathbf{T}\mathbf{y}_i$, $\mathbf{W}_i^+ = \mathbf{T}\mathbf{W}_i$, and $\mathbf{F}_i^+ = \mathbf{T}\mathbf{F}_i$, with the $(2T-1) \times T$ transformation matrix $\mathbf{T} = (\mathbf{D}', \mathbf{I}_T)'$. The residuals are $\mathbf{e}_i^+ = \mathbf{T}\mathbf{e}_i = (\Delta\mathbf{u}_i', \mathbf{e}_i')'$. The instruments for the full equation system are combined in the following block-diagonal matrix:

$$\mathbf{Z}_i^+ = \begin{pmatrix} \mathbf{Z}_i^d & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_i^l \end{pmatrix}.$$

Based on the sample moments $N^{-1}\sum_{i=1}^N (\mathbf{Z}_i^+)'\mathbf{e}_i^+ = N^{-1}(\mathbf{Z}^+)'\mathbf{e}^+$ we can derive the GMM estimator $(\hat{\boldsymbol{\theta}}', \hat{\boldsymbol{\gamma}}')$ as a minimum distance estimator:

$$\begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \arg\min_{\boldsymbol{\theta}, \boldsymbol{\gamma}} \left[(\mathbf{Z}^+)'\mathbf{e}^+\right]' \mathbf{V}_N (\mathbf{Z}^+)'\mathbf{e}^+,$$

where $\mathbf{V}_N$ is a weighting matrix that might depend on the data, with $\operatorname{plim} \mathbf{V}_N = \mathbf{V}$ for a positive definite matrix $\mathbf{V}$. It is readily seen that this minimization problem can be

[18] The $(T-2)$ non-linear moment conditions that are not a consequence of the homoscedasticity assumption are implied by the conditions (30) to (32). See Blundell and Bond (1998) in the absence of exogenous regressors.

rewritten in terms of the level residuals only with the transformed instrument matrix $\mathbf{Z}_i = \mathbf{T}'\mathbf{Z}_i^+ = (\mathbf{D}'\mathbf{Z}_i^d, \mathbf{Z}_i^l)$:

$$\begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \arg\min_{\boldsymbol{\theta},\boldsymbol{\gamma}} (\mathbf{Z}'\mathbf{e})'\mathbf{V}_N\mathbf{Z}'\mathbf{e}.$$

Consequently,

$$\begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = (\dot{\mathbf{W}}'\mathbf{Z}\mathbf{V}_N\mathbf{Z}'\dot{\mathbf{W}})^{-1}\dot{\mathbf{W}}'\mathbf{Z}\mathbf{V}_N\mathbf{Z}'\mathbf{y}, \tag{37}$$

where $\dot{\mathbf{W}} = (\mathbf{W}, \mathbf{F})$. Notably, the first block of the transformed instrument matrix, $\mathbf{D}'\mathbf{Z}_i^d$, is a set of instruments that are orthogonal to any time-invariant variable.

We can consistently estimate the asymptotic variance-covariance matrix Σ of the GMM estimator as follows:

$$\hat{\Sigma} = (\dot{\mathbf{W}}'\mathbf{Z}\mathbf{V}_N\mathbf{Z}'\dot{\mathbf{W}})^{-1}\dot{\mathbf{W}}'\mathbf{Z}\mathbf{V}_N\hat{\Xi}\mathbf{V}_N\mathbf{Z}'\dot{\mathbf{W}}(\dot{\mathbf{W}}'\mathbf{Z}\mathbf{V}_N\mathbf{Z}'\dot{\mathbf{W}})^{-1}, \tag{38}$$

where $\hat{\Xi}$ is a consistent estimate of $\Xi = \operatorname{plim} N^{-1}\sum_{i=1}^N \mathbf{Z}_i'\mathbf{e}_i\mathbf{e}_i'\mathbf{Z}_i$. Under assumptions 1 and 2, $\Xi = \operatorname{plim} N^{-1}\sum_{i=1}^N \mathbf{Z}_i'\Omega^l\mathbf{Z}_i$. The GMM estimator (37) is efficient for a given instrument matrix $\mathbf{Z}$ if $\mathbf{V} = c\,\Xi^{-1}$ for any constant scalar $c > 0$. Then, a consistent estimate of the asymptotic variance-covariance matrix (38) is given by:

$$\hat{\Sigma} = (\dot{\mathbf{W}}'\mathbf{Z}\hat{\Xi}^{-1}\mathbf{Z}'\dot{\mathbf{W}})^{-1}. \tag{39}$$

With prior knowledge of the ratio $\tau = \sigma_\alpha^2/\sigma_u^2$, an optimal weighting matrix is:

$$\mathbf{V}_N = \left[\frac{1}{N}\sum_{i=1}^N \mathbf{Z}_i'(\tau\boldsymbol{\iota}_T\boldsymbol{\iota}_T' + \mathbf{I}_T)\mathbf{Z}_i\right]^{-1}, \tag{40}$$

such that $\mathbf{V} = \sigma_u^2\Xi^{-1}$. In general, however, there exists no asymptotically efficient one-step GMM estimator[19] since τ is unknown.[20] In this case, it is common practice to use

$$\mathbf{V}_N = \left[\frac{1}{N}\sum_{i=1}^N (\mathbf{Z}_i^+)'\mathbf{H}_j\mathbf{Z}_i^+\right]^{-1}, \quad j \in \{1,2,3\}, \tag{41}$$

with either

$$\mathbf{H}_1 = \mathbf{I}_{2T-1}, \quad \mathbf{H}_2 = \begin{pmatrix} \mathbf{D}\mathbf{D}' & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_T \end{pmatrix}, \quad \text{or} \quad \mathbf{H}_3 = \begin{pmatrix} \mathbf{D}\mathbf{D}' & \mathbf{D} \\ \mathbf{D}' & \mathbf{I}_T \end{pmatrix}.$$

[19] In this paper, we call an estimator a one-*step* estimator if it is not based on prior estimates. In contrast, a two-step estimator is a feasible efficient estimator that makes use of consistent initial estimates. The denotation of a one-*stage* estimator is used for estimators that obtain all coefficient estimates simultaneously (potentially in two steps) while a two-stage estimator first obtains the coefficients of the time-varying regressors and second the coefficients of the time-invariant regressors given the first-stage estimates.

[20] Compare Blundell and Bond (1998), Windmeijer (2000), and Blundell et al. (2001).

$\mathbf{H}_1$ is used, among others, by Arellano and Bover (1995) and Blundell and Bond (1998), while Blundell et al. (2001) take the first-order serial correlation in the first-differenced residuals into account by choosing $\mathbf{H}_2$. When σ_α^2 is small, Windmeijer (2000) suggests to reduce the potential efficiency loss by using $\mathbf{H}_3$. In fact, when $\tau = 0$, the weighting matrix (41) based on $\mathbf{H}_3$ equals the optimal weighting matrix (40) under assumptions 1 and 2 since $\mathbf{H}_3 = \mathbf{T}\mathbf{T}'$.

We can obtain a two-step GMM estimator with optimal weighting matrix by choosing $V_N = \hat{\Xi}^{-1}$ given consistent initial estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$. A consistent unrestricted estimate of Ξ is obtained as $\hat{\Xi} = N^{-1}\sum_{i=1}^N \mathbf{Z}_i'\hat{\mathbf{e}}_i\hat{\mathbf{e}}_i'\mathbf{Z}_i$, with $\hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{W}_i\hat{\boldsymbol{\theta}} - \mathbf{F}_i\hat{\boldsymbol{\gamma}}$. Under assumptions 1 and 2, we can obtain a restricted estimate as $\hat{\Xi} = N^{-1}\sum_{i=1}^N \mathbf{Z}_i'\hat{\Omega}^l\mathbf{Z}_i$ with either $\hat{\Omega}^l = N^{-1}\sum_{i=1}^N \hat{\mathbf{e}}_i\hat{\mathbf{e}}_i'$ or $\hat{\Omega}^l = \hat{\sigma}_\alpha^2\boldsymbol{\iota}_T\boldsymbol{\iota}_T' + \hat{\sigma}_u^2\mathbf{I}_T$. The variance estimates $\hat{\sigma}_u^2$ and $\hat{\sigma}_\alpha^2$ can be obtained as follows:

$$\hat{\sigma}_u^2 = \frac{1}{2[N(T-1)-(1+K_x)]}\sum_{i=1}^N\sum_{t=2}^T \widehat{\Delta u}_{it}^2, \tag{42}$$

$$\hat{\sigma}_e^2 = \frac{1}{NT-(1+K_x+K_f)}\sum_{i=1}^N\sum_{t=1}^T \hat{e}_{it}^2, \tag{43}$$

$$\hat{\sigma}_\alpha^2 = \hat{\sigma}_e^2 - \hat{\sigma}_u^2, \tag{44}$$

where $\widehat{\Delta u}_{it} = \Delta y_{it} - \Delta\mathbf{W}_{it}\hat{\boldsymbol{\theta}}$. The importance of choosing an appropriate first-step weighting matrix should not be underestimated in applied work. Although the two-step GMM estimator is asymptotically unaffected, its finite sample performance still depends on the choice of $\mathbf{V}_N$ in the first step. Windmeijer (2005) shows that asymptotic standard error estimates of the two-step GMM estimator can be severely downward based in finite samples. He derives a finite sample variance correction.

When the estimator only involves moment conditions for the model in first differences, such that $\mathbf{Z}_i = \mathbf{D}'\mathbf{Z}_i^d$, the optimal weighting matrix boils down to $\mathbf{V}_N = (N^{-1}\sum_{i=1}^N \mathbf{Z}_i'\mathbf{Z}_i)^{-1}$, as suggested by Arellano and Bond (1991). In this case, the one-step estimator is as efficient as the two-step estimator under assumption 1 because the optimal weighting matrix does not depend on τ any more.

4.3 Reduction of the Number of Instruments

If all the assumptions are met, using the full instrument matrix based on the moment conditions set out in Subsection 4.1 is asymptotically efficient. In finite samples, however, severe distortions can result from having too many instruments relative to the sample size. Roodman (2009) stresses four main symptoms of instrument proliferation. First, the coefficient estimates can be biased towards the non-instrumented estimates because a large set of instruments potentially overfits the model. Second, the optimal weighting matrix of two-step GMM estimators might be poorly estimated because its dimension increases with the number of the instruments. Third, as a result of the imprecisely estimated weighting matrix the estimated standard errors of two-step GMM estimators tend to be downward biased. This issue is addressed by Windmeijer (2005) who provides a finite sample correction for the variance of two-step GMM estimators. Fourth, specification tests for two-step GMM estimators that are also based on an estimate of the optimal weighting matrix, as the Hansen (1982) J-test for the validity of the overidentifying restrictions, become weak. This might lead to a false indication that the overidentifying restrictions are valid when in fact they are not.[21]

When the full instrument matrix $\mathbf{Z}_i$ is used these problems can become severe already for time dimensions that are usually still considered as being small. It is easily seen that the number of moment conditions for the first-differenced equation grows with rate T^2 when the time dimension increases, and those of the level equation with rate T. In applied work, researchers often restrict the number of lags that are used to construct the instrument matrix for the first-differenced equation. With a fixed lag depth the number of instruments becomes linear in T. Alternatively, reducing the instrument count from quadratic to linear in T can also be achieved by systematically using linear combinations instead of all moment conditions separately.[22] Both procedures are effectively deterministic transformations of the instrument matrix. As discussed by Mehrhoff (2009), a transformation of the instrument matrix $\mathbf{Z}_i$ such that $\mathbf{Z}_i^* = \mathbf{Z}_i\mathbf{R}$ for any deterministic matrix $\mathbf{R}$ also leads to a valid set of moment conditions, $E[(\mathbf{Z}_i^*)'\mathbf{e}_i] = \mathbf{0}$. The corresponding GMM estimator is obtained by replacing $\mathbf{Z}_i$ with $\mathbf{Z}_i^*$ in the previous subsection. Appendix A provides practically relevant examples of the transformation matrix $\mathbf{R}$.

4.4 Two-Stage GMM Estimation

If some of the variables $\mathbf{f}_{2i}$ that are correlated with the unit-specific effects are mistakenly classified as variables $\mathbf{f}_{1i}$ that are supposed to be uncorrelated with the latent effects, all coefficient estimates will generally be biased including those of the time-varying regressors. However, there is an important exception. If $\boldsymbol{\gamma}$ is only just identified (or even underidentified), that is $\text{rk}(\mathbf{Z}^l) \leq K_f$, the identification of the coefficients λ and $\boldsymbol{\beta}$ does neither depend on $\mathbf{Z}_i^l$ nor on the covariance of the time-varying with the time-invariant regressors. This is a consequence of $\mathbf{DF}_i = \mathbf{0}$. Therefore, a bias in $\hat{\boldsymbol{\gamma}}$ does not translate into a bias in $\hat{\boldsymbol{\theta}}$ in this case. An example for this case is $K_{f2} = 0$ and $\mathbf{Z}_i^l = \mathbf{F}_{1i}$.

A brute force alternative would be to specify a GMM estimator that treats all variables as potentially correlated with α_i. While this procedure still allows to identify λ and $\boldsymbol{\beta}$, the coefficients $\boldsymbol{\gamma}$ are identified only technically by virtue of the overidentifying restrictions. Although unbiased, the GMM estimates $\hat{\boldsymbol{\gamma}}$ are not informative. The transformed instruments for the first-differenced equation are orthogonal to all time-invariant variables by construction. The remaining demeaned and first-differenced instruments also do not help identifying $\boldsymbol{\gamma}$ because it is unlikely that these instruments are correlated

[21]See Roodman (2009) and the references therein for an extensive discussion.
[22]See again Roodman (2009).

with the time-invariant regressors when assumption 6.1 holds.[23]

A robust alternative is based on a two-stage estimation strategy. In the first stage, the time-invariant variables (besides the regression constant) are subsumed under the unit-specific effects, and the corresponding moment conditions (34) are left disregarded. We thus require that assumption 6.1 still holds for $\alpha_i^* = \alpha_i + \mathbf{f}_i'\boldsymbol{\gamma}$. If the regressors $\mathbf{x}_{1it}$ are correlated with $\mathbf{f}_i$, the moment conditions (31) become invalid as well. We then have to treat $\mathbf{x}_{1it}$ equivalently to $\mathbf{x}_{2it}$ and use the corresponding moment conditions (32) and (33) instead. The resulting first-stage GMM estimator is given by equation (37) after adjusting the instrument matrix appropriately and replacing $\dot{\mathbf{W}}$ with $(\mathbf{W}, \boldsymbol{\iota}_{NT})$.[24] The first-stage variance-covariance matrix Σ is adjusted accordingly. The estimates $\hat{\boldsymbol{\theta}}$ are subsequently used to recover $\boldsymbol{\gamma}$ as described in Section 3.

The second-stage estimator $\hat{\hat{\boldsymbol{\gamma}}}$ is given in equation (14), and its asymptotic variance-covariance matrix Σ_γ in equation (20). Under assumptions 1 and 2, the asymptotic covariance matrix $\mathbf{C}$ is given by:[25]

$$\mathbf{C} = \tilde{\mathbf{J}}[\mathbf{S}_1\mathbf{V}\mathbf{S}_1']^{-1}[\mathbf{S}_1\mathbf{V}(\sigma_u^2\mathbf{S}_2 + \sigma_\alpha^2\mathbf{S}_3)], \tag{45}$$

with $\mathbf{V} = \operatorname{plim} \mathbf{V}_N$ as defined in Subsection 4.2, $\tilde{\mathbf{J}} = (\mathbf{I}_{1+K_x}, \mathbf{0}_{1+K_x\times 1})$, and

$$\mathbf{S}_1 = \operatorname{plim} \frac{1}{N}\sum_{i=1}^N \dot{\mathbf{W}}_i'\mathbf{Z}_i, \quad \mathbf{S}_2 = \operatorname{plim} \frac{1}{N}\sum_{i=1}^N \mathbf{Z}_i'\mathbf{s}_T\tilde{\mathbf{z}}_i', \quad \mathbf{S}_3 = \operatorname{plim} \frac{1}{N}\sum_{i=1}^N \mathbf{Z}_i'\boldsymbol{\iota}_T\tilde{\mathbf{z}}_i',$$

where $\mathbf{s}_T = (0, \ldots, 0, 1)'$ is of dimension $T \times 1$ with 1 at position T and zeros elsewhere.

This two-stage procedure is not restricted to a GMM estimator in the first stage. We can apply any first-stage estimator that consistently estimates the coefficients of the time-varying regressors without relying on estimates of the coefficients of time-invariant regressors. Thus, estimators that are solely based on the first-differenced model are natural first-stage candidates. An example for such an estimator is the QML estimator of Hsiao et al. (2002) that we discuss in the next section.

5 Two-Stage Quasi-Maximum Likelihood Estimation

If $K_{x2} = K_{f2} = 0$ we can immediately estimate model (1) with the random effects maximum likelihood estimator of Bhargava and Sargan (1983) and Hsiao et al. (2002). When this strong assumption does not hold, Hsiao et al. (2002) propose to estimate the coefficients of the time-varying regressors including the lagged dependent variable based on the first-differenced model (8). However, this procedure not only eliminates the incidental parameters α_i but also the time-invariant variables $\mathbf{f}_i$. The latter can be recovered in a second stage.

[23]Compare Arellano (2003), Chapter 8.5.4.

[24]The first-stage intercept is only needed for consistent estimation of the parameters of interest if $E[\mathbf{f}_i']\boldsymbol{\gamma} \neq 0$. We obtain a final estimate of the regression constant in the second stage.

[25]If all columns of $\mathbf{Z}_i$ are orthogonal to time-invariant variables the term $\sigma_\alpha^2\mathbf{S}_3$ drops out.

Under the strict exogeneity assumption 4.1 the joint density of $\Delta\mathbf{y}_i^*$ conditional on $\Delta\mathbf{X}_i^*$ is given by:

$$f(\Delta\mathbf{y}_i^*|\Delta\mathbf{X}_i^*) = f(\Delta y_{iT}|\Delta y_{i,T-1}, \Delta y_{i,T-2}, \ldots, \Delta y_{i1}, \Delta\mathbf{X}_i^*) \\ f(\Delta y_{i,T-1}|\Delta y_{i,T-2}, \Delta y_{i,T-3}, \ldots, \Delta y_{i1}, \Delta\mathbf{X}_i^*) \ldots \\ f(\Delta y_{i2}|\Delta y_{i1}, \Delta\mathbf{X}_i^*) f(\Delta y_{i1}|\Delta\mathbf{X}_i^*). \tag{46}$$

All but the last term can be easily derived from the first-differenced model (8). This is not the case for $f(\Delta y_{i1}|\Delta\mathbf{X}_i^*)$ because Δy_{i0} is not observed. Hsiao et al. (2002) apply linear projection techniques to derive the following initial observation condition:

$$\Delta y_{i1} = b + \sum_{s=1}^{T} \Delta\mathbf{x}_{is}'\boldsymbol{\pi}_s + \xi_{i1}, \tag{47}$$

based on the following assumption:

Assumption 6.2: $\mathbf{x}_{it}$ is generated either by a trend stationary or first-difference stationary process such that $\Delta\mathbf{x}_{it}$ is covariance stationary:

$$\Delta\mathbf{x}_{it} = \mathbf{g} + \sum_{s=0}^{\infty} \mathbf{B}_s\boldsymbol{\epsilon}_{i,t-s}, \quad \mathbf{B}_0 = \mathbf{I}_{K_x}, \quad \sum_{s=0}^{\infty} \mathbf{B}_s\mathbf{B}_s' < \infty, \tag{48}$$

where $E[\boldsymbol{\epsilon}_{it}] = \mathbf{0}$, $E[\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}'] = \boldsymbol{\Sigma}_\epsilon$ and $E[\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{is}'] = \mathbf{0}$ for $t \neq s$.

If this stationarity assumption is violated, b and $\boldsymbol{\pi} = (\boldsymbol{\pi}_1', \boldsymbol{\pi}_2', \ldots, \boldsymbol{\pi}_T')'$ might depend on i such that the incidental parameters problem is still present. The properties of the error term ξ_{i1} are $E[\xi_{i1}^2] = \sigma_\xi^2$, $E[\xi_{i1}|\Delta\mathbf{x}_i] = 0$, $E[\xi_{i1}\Delta u_{i2}] = -\sigma_u^2$, and $E[\xi_{i1}\Delta u_{it}] = 0$ for $t = 3, 4, \ldots, T$. Under the additional assumption that u_{it} and $\boldsymbol{\epsilon}_{it}$ are normally distributed we can now set up the transformed likelihood function:

$$\ln L_T = -\frac{NT}{2}\ln(2\pi) - \frac{N}{2}\ln|\Omega^*| - \frac{1}{2}\sum_{i=1}^{N} \Delta\mathbf{u}_i^{*'}(\Omega^*)^{-1}\Delta\mathbf{u}_i^*, \tag{49}$$

where $\Delta\mathbf{u}_i^* = (\xi_{i1}, \Delta\mathbf{u}_i')'$. Moreover,[26]

$$\Omega^* = \sigma_u^2\tilde{\Omega} = \sigma_u^2\begin{pmatrix} \omega & -\mathbf{s}_1' \\ -\mathbf{s}_1 & \mathbf{DD}' \end{pmatrix} = \sigma_u^2\begin{pmatrix} \omega & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & & \\ 0 & -1 & 2 & & \\ \vdots & & & \ddots & -1 \\ 0 & & & -1 & 2 \end{pmatrix},$$

[26] Hayakawa and Pesaran (2012) extend the transformed likelihood estimator to accommodate for heteroscedastic errors.

where $\mathbf{s}_1 = (1, 0, \dots, 0)'$ is a selection vector of compatible length, here $(T-1) \times 1$, with 1 at position 1 and zeros elsewhere, and $\omega = \sigma_\xi^2 / \sigma_u^2$ such that $\ln |\Omega^*| = T \ln \sigma_u^2 + \ln[1 + T(\omega - 1)]$, as demonstrated by Hsiao et al. (2002). With $\Delta \mathbf{u}_i^* = \Delta \mathbf{y}_i^* - \Delta \mathbf{W}_i^* \boldsymbol{\varphi}$, where $\boldsymbol{\varphi} = (b, \boldsymbol{\pi}', \lambda, \boldsymbol{\beta}')'$, and $\Delta \mathbf{W}_i^* = (\Delta \tilde{\mathbf{X}}_i, \Delta \tilde{\mathbf{W}}_i)$ with

$$\Delta \tilde{\mathbf{X}}_i = \begin{pmatrix} 1 & \Delta \tilde{\mathbf{x}}_i' \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \Delta \tilde{\mathbf{W}}_i = \begin{pmatrix} \mathbf{0}' \\ \Delta \mathbf{W}_i \end{pmatrix},$$

and $\Delta \tilde{\mathbf{x}}_i = (\Delta \mathbf{x}_{i1}', \Delta \mathbf{x}_{i2}', \dots, \Delta \mathbf{x}_{iT}')'$, we can rewrite the log-likelihood function as

$$\begin{aligned} \ln L_T = & -\frac{NT}{2} \ln(2\pi) - \frac{NT}{2} \ln \sigma_u^2 - \frac{N}{2} \ln[1 + T(\omega - 1)] \\ & - \frac{1}{2\sigma_u^2} \sum_{i=1}^{N} (\Delta \mathbf{y}_i^* - \Delta \mathbf{W}_i^* \boldsymbol{\varphi})' \tilde{\Omega}^{-1} (\Delta \mathbf{y}_i^* - \Delta \mathbf{W}_i^* \boldsymbol{\varphi}). \end{aligned} \tag{50}$$

The weak exogeneity assumption 4.2 requires the derivation of a joint density of $\Delta \mathbf{y}_i^*$ and the weakly exogenous regressors. According to Hsiao et al. (2002), the only difference in the likelihood function compared to the strict exogeneity case is the initial condition which they model as a function of the initial observations of the regressor variables only instead of using all available observations:[27]

$$\Delta y_{i1} = b + \Delta \mathbf{x}_{i1}' \boldsymbol{\pi} + \xi_{i1}, \tag{51}$$

such that $\Delta \tilde{\mathbf{x}}_i = \Delta \mathbf{x}_{i1}$.

Hsiao et al. (2002) derive the following first-order conditions:

$$\hat{\boldsymbol{\varphi}} = \left(\sum_{i=1}^{N} (\Delta \mathbf{W}_i^*)' \hat{\tilde{\Omega}}^{-1} \Delta \mathbf{W}_i^* \right)^{-1} \sum_{i=1}^{N} (\Delta \mathbf{W}_i^*)' \hat{\tilde{\Omega}}^{-1} \Delta \mathbf{y}_i^* \tag{52}$$

$$\hat{\sigma}_u^2 = \frac{1}{NT} \sum_{i=1}^{N} (\Delta \mathbf{y}_i^* - \Delta \mathbf{W}_i^* \hat{\boldsymbol{\varphi}})' \hat{\tilde{\Omega}}^{-1} (\Delta \mathbf{y}_i^* - \Delta \mathbf{W}_i^* \hat{\boldsymbol{\varphi}}) \tag{53}$$

$$\hat{\omega} = \frac{T-1}{T} + \frac{1}{NT^2 \hat{\sigma}_u^2} \sum_{i=1}^{N} (\Delta \mathbf{y}_i^* - \Delta \mathbf{W}_i^* \hat{\boldsymbol{\varphi}})' \boldsymbol{\vartheta} \boldsymbol{\vartheta}' (\Delta \mathbf{y}_i^* - \Delta \mathbf{W}_i^* \hat{\boldsymbol{\varphi}}), \tag{54}$$

where $\boldsymbol{\vartheta} = (T, T-1, T-2, \dots, 1)'$. By inserting the first-order conditions for $\hat{\boldsymbol{\varphi}}$ and $\hat{\sigma}_u^2$ back into the log-likelihood function, we get a concentrated log-likelihood function that depends only on ω. We can then apply an iterative procedure to derive the maximizing value $\hat{\omega}$ and subsequently recover the other parameter values. For a given value of ω the

[27] See Hsiao et al. (2002) for details on the derivation.

quasi-maximum likelihood estimator for λ and $\boldsymbol{\beta}$ is given by:

$$\begin{aligned}\hat{\boldsymbol{\theta}} = & \left[\sum_{i=1}^{N} \Delta\tilde{\mathbf{W}}_i'\tilde{\Omega}^{-1}\Delta\tilde{\mathbf{W}}_i - \left(\sum_{i=1}^{N} \Delta\tilde{\mathbf{W}}_i'\tilde{\Omega}^{-1}\Delta\tilde{\mathbf{X}}_i\right)\right. \\ & \left.\times \left(\sum_{i=1}^{N} \Delta\tilde{\mathbf{X}}_i'\tilde{\Omega}^{-1}\Delta\tilde{\mathbf{X}}_i\right)^{-1} \left(\sum_{i=1}^{N} \Delta\tilde{\mathbf{X}}_i'\tilde{\Omega}^{-1}\Delta\tilde{\mathbf{W}}_i\right)\right]^{-1} \\ & \times \left[\sum_{i=1}^{N} \Delta\tilde{\mathbf{W}}_i'\tilde{\Omega}^{-1}\Delta\mathbf{y}_i^* - \left(\sum_{i=1}^{N} \Delta\tilde{\mathbf{W}}_i'\tilde{\Omega}^{-1}\Delta\tilde{\mathbf{X}}_i\right)\right. \\ & \left.\times \left(\sum_{i=1}^{N} \Delta\tilde{\mathbf{X}}_i'\tilde{\Omega}^{-1}\Delta\tilde{\mathbf{X}}_i\right)^{-1} \left(\sum_{i=1}^{N} \Delta\tilde{\mathbf{X}}_i'\tilde{\Omega}^{-1}\Delta\mathbf{y}_i^*\right)\right], \end{aligned} \tag{55}$$

as derived by Hsiao et al. (2002).[28] The variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is the lower-right $(1 + K_x) \times (1 + K_x)$ block of the inverse negative Hessian matrix from the above maximization problem.

Analogously to the two-stage estimation discussed in the GMM Subsection 4.4 we can recover the coefficients $\boldsymbol{\gamma}$ of the time-invariant regressors in a second stage. The second-stage estimator $\hat{\hat{\boldsymbol{\gamma}}}$ is again given by equation (14), and the joint distribution of the first-stage and second-stage estimators by equations (18) to (20), where now Σ_θ is the asymptotic variance-covariance matrix of the QML estimator $\hat{\boldsymbol{\theta}}$ and

$$\mathbf{C} = \sigma_u^2[\mathbf{S}_4 - \mathbf{S}_5\mathbf{S}_6^{-1}\mathbf{S}_5']^{-1}[\mathbf{S}_7 - \mathbf{S}_5\mathbf{S}_6^{-1}\mathbf{S}_8], \tag{56}$$

with

$$\mathbf{S}_4 = \text{plim}\,\frac{1}{N}\sum_{i=1}^{N} \Delta\tilde{\mathbf{W}}_i'\tilde{\Omega}^{-1}\Delta\tilde{\mathbf{W}}_i, \quad \mathbf{S}_5 = \text{plim}\,\frac{1}{N}\sum_{i=1}^{N} \Delta\tilde{\mathbf{W}}_i'\tilde{\Omega}^{-1}\Delta\tilde{\mathbf{X}}_i,$$

$$\mathbf{S}_6 = \text{plim}\,\frac{1}{N}\sum_{i=1}^{N} \Delta\tilde{\mathbf{X}}_i'\tilde{\Omega}^{-1}\Delta\tilde{\mathbf{X}}_i, \quad \mathbf{S}_7 = \text{plim}\,\frac{1}{N}\sum_{i=1}^{N} \Delta\tilde{\mathbf{W}}_i'\tilde{\Omega}^{-1}\mathbf{s}_T\tilde{\mathbf{z}}_i',$$

$$\mathbf{S}_8 = \text{plim}\,\frac{1}{N}\sum_{i=1}^{N} \Delta\tilde{\mathbf{X}}_i'\tilde{\Omega}^{-1}\mathbf{s}_T\tilde{\mathbf{z}}_i',$$

and $\mathbf{s}_T$ defined in Section 4.4.

[28]Hsiao et al. (2002) provide a proof for consistency of the QML estimator for models without exogenous regressors. For a derivation of the consistency of the estimator including strictly exogenous variables see Appendix B.

6 Monte Carlo Simulation

6.1 Simulation Design

In our Monte Carlo experiments we focus on the case $K_{x1} = K_{f2} = 0$. That is, the time-varying regressor x_{it} is correlated with the unobserved fixed effects and the time-invariant regressor f_i is uncorrelated with them. For the ease of comparability of the different estimation methods we choose $K_x = K_f = 1$, even though we note that some problems of GMM estimators that result from too many overidentifying restrictions might aggravate with a larger number of time-varying regressors. We generate y_{it} and x_{it} according to the following processes:

$$y_{it} = \lambda y_{i,t-1} + \beta x_{it} + \gamma f_i + \alpha_i + u_{it}, \quad u_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2), \tag{57}$$

and

$$x_{it} = \phi x_{i,t-1} + \nu\rho f_i + \nu\sqrt{1-\rho^2}\eta_i + \epsilon_{it}, \quad \epsilon_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \tag{58}$$

such that x_{it} is strictly exogenous with respect to u_{it}.

We generate the observed unit-specific effects f_i as an independent binary variable from a Bernoulli distribution with success probability p. The unobserved unit-specific effects α_i and η_i are generated from a joint normal distribution:

$$\begin{pmatrix} \alpha_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\eta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\eta} \\ \sigma_{\alpha\eta} & p(1-p) \end{pmatrix} \right), \tag{59}$$

such that the variances of η_i and f_i coincide. The particular design of the process for x_{it} guarantees that the correlation between x_{it} and f_i can be altered while keeping the variance of x_{it} unchanged, because

$$Var(x_{it}) = \frac{1}{(1-\phi)^2}\left[\nu^2 p(1-p) + \frac{1-\phi}{1+\phi}\sigma_\epsilon^2\right] \tag{60}$$

is independent of ρ. $\nu \geq 0$ is introduced as a scale parameter. The correlation coefficient for x_{it} and f_i can be written as:

$$Corr(x_{it}, f_i) = \rho\sqrt{\frac{\nu^2 p(1-p)}{\nu^2 p(1-p) + \frac{1-\phi}{1+\phi}\sigma_\epsilon^2}}. \tag{61}$$

Since $\rho \in [-1, 1]$, it can be interpreted as a correlation coefficient net of the variation coming from ϵ_{it}.

We set the long-run coefficient $\beta/(1-\lambda) = 1$ and initialize the processes at $t = -50$ with their long-run means given the realizations of the unit-specific effects:

$$y_{i,-50} = x_{i,-50} + \frac{1}{1-\lambda}\left(\gamma f_i + \alpha_i\right), \tag{62}$$

$$x_{i,-50} = \frac{\nu}{1-\phi}\left(\rho f_i + \sqrt{1-\rho^2}\eta_i\right), \tag{63}$$

Table 1: Simulation Designs

Design	λ	β	ρ	σ^2_α
1	0.8	0.2	0.8	4
2	0.8	0.2	0.8	1
3	0.4	0.6	0.8	4
4	0.8	0.2	0	4

and discard the first 50 observations for the estimation. The covariance between the two unobserved fixed effects α_i and η_i is set to $\sigma_{\alpha\eta} = \frac{1}{2}\sigma_\alpha\sqrt{p(1-p)}$ which creates a positive correlation between x_{it} and α_i. We also fix $\gamma = 1$, $\phi = 0.8$, $\sigma^2_u = 1$, $\nu = 1$, $p = \frac{1}{2}$ and $\mu_\alpha = \mu_\eta = 0$. To ensure an adequate degree of fit, we fix the population value of the coefficient of determination for the first-differenced model, $R^2_{\Delta y}$, in a similar fashion as Hsiao et al. (2002). For the data generating process stated above it is given by:[29]

$$R^2_{\Delta y} = \frac{\beta^2\sigma^2_\epsilon}{\beta^2\sigma^2_\epsilon + (1+\phi)(1-\lambda\phi)\sigma^2_u}. \tag{64}$$

We fix $R^2_{\Delta y} = 0.2$ and determine σ^2_ϵ from the above equation:

$$\sigma^2_\epsilon = \frac{R^2_{\Delta y}}{1-R^2_{\Delta y}}\frac{(1+\phi)(1-\lambda\phi)}{\beta^2}\sigma^2_u. \tag{65}$$

Finally, we simulate the data with different combinations for the remaining parameters according to Table 1 with $T \in \{5, 10\}$ and $N \in \{50, 200, 500\}$. Our baseline calibration, design 1, features a relatively persistent process of y_{it}, a large variance of the unobserved unit-specific effects, and a high correlation between the strictly exogenous regressor x_{it} and the time-invariant variable f_i. In comparison to the baseline calibration, design 2 has a lower variance of α_i, in design 3 y_{it} is less persistent, and in design 4 the regressors x_{it} and f_i are uncorrelated. In total, we do 2500 repetitions for each simulation design.

The initial values for the QML optimization are obtained from a consistent system GMM estimation. Hsiao et al. (2002) report that the (first-stage) maximum likelihood estimator sometimes breaks down in their simulation. We face the same problem of getting an initial estimate of ω smaller than $(T-1)/T$ in some cases, especially when N is small. However, in contrast to Hsiao et al. (2002) we do not skip those replications but instead change the initial estimate of ω to $(T-1)/T + \delta$, where we choose $\delta = 0.01$.[30]

We compare the two-stage QML estimator, "2s-QML", to various GMM estimators that use different sets of instruments and recover the coefficient of the time-invariant regressor either in one or in two stages. To trace back the simulation results to specific

[29] The derivation of $R^2_{\Delta y}$ can be found in Appendix C.

[30] The particular choice of δ does not matter as long as it is small enough.

properties of the estimators, we report the results for the following five GMM specifications.[31] First, we set up a two-step system GMM estimator that exploits the full set of moment conditions and recovers all parameters jointly in one stage, "1s-sGMM-2 (full)".[32] Besides the moment conditions (27) and (34) that result from the presence of the time-invariant regressor, this estimator equals the one proposed by Blundell et al. (2001). To reduce the instrument count the most commonly applied method is restricting the lag depth of the instrumental variables. Therefore we set up an estimator using only a maximum number of two lags per variable, "1s-sGMM-2 (2 lags)". The remaining GMM estimators all use a collapsed set of instruments to deal with the problems resulting from too many instruments as discussed in Section 4.3.[33] For the collapsed one-stage system GMM estimator we report the results both for the one-step, "1s-sGMM-1 (collapsed)", and for the two-step version, "1s-sGMM-2 (collapsed)". Although the one-step system GMM estimator is generally inefficient due to the absence of an optimal weighting matrix, the results below reveal some interesting differences between both versions for the coefficient of the time-invariant regressor. Finally, we consider a GMM estimator that recovers the coefficients in two stages and is based in the first stage on a two-step system GMM estimator, "2s-sGMM-2 (collapsed)". For all GMM estimators we base the first-step weighting matrix (41) on the matrix $\mathbf{H}_2$. To compute the standard errors of the GMM estimators, we use the robust variance-covariance formula (38) with an unrestricted estimate of Ξ. For two-step estimators, we apply the Windmeijer (2005) correction.

6.2 Simulation Results

Table 2 summarizes the simulation results for our baseline design. First of all, the two-stage QML estimator shows a very small bias relative to the true parameter value for all three coefficients. For γ, the coefficient of our main interest, the two-stage QML estimator has a strong lead as its relative bias is only 0.5% compared to 3.2 % of the second-best estimator. The root mean square error also favors the QML approach. For the coefficients of the time-varying regressors, these results confirm the findings of Hsiao et al. (2002). For γ only the one-stage one-step system GMM estimator with collapsed instrument matrices has a slightly lower RMSE but the magnitudes are comparable. The actual rejection frequencies of the Wald tests that the estimated coefficients equal their true values also support the two-stage QML estimator. They are reasonably close to the nominal size of 5 or 10 percent, respectively, even though the estimated standard errors for the autoregressive parameter λ tend to be too small. On average they amount to 85% of the empirical standard deviation. For the second-stage parameter γ, the ratio

[31] We also did simulations for other transformations of the instrument matrices but they do not reveal additional insights in the behavior of the GMM estimators.

[32] We disregard the moment conditions (28) that are due to homoscedasticity and the additional level moment condition (33) since they are rarely used in applied work. For the regression constant we exploit only the moment conditions (34) but not the conditions (27).

[33] Appendix A describes the particular design of the transformation matrices.

Table 2: Summary Results for the Baseline Calibration

Coefficient	Estimator	Bias	RMSE	Size (5%)	Size (10%)	SE/SD
Design 1, $T = 10$, $N = 50$						
λ	2s-QML	-0.0002	0.0253	0.0600	0.1096	0.8460
	1s-sGMM-2 (full)	0.0637	0.0581	0.3704	0.4972	1.1600
	1s-sGMM-2 (2 lags)	0.0962	0.0822	0.7616	0.8312	0.9212
	1s-sGMM-1 (collapsed)	0.0062	0.0343	0.0936	0.1424	0.9119
	1s-sGMM-2 (collapsed)	0.0035	0.0352	0.0836	0.1396	0.9314
	2s-sGMM-2 (collapsed)	0.0040	0.0358	0.0892	0.1468	0.9234
β	2s-QML	-0.0007	0.0103	0.0480	0.1000	0.9853
	1s-sGMM-2 (full)	-0.0123	0.0147	0.0212	0.0528	1.2014
	1s-sGMM-2 (2 lags)	0.0066	0.0170	0.0568	0.1068	1.0042
	1s-sGMM-1 (collapsed)	-0.0024	0.0180	0.0624	0.1228	1.0115
	1s-sGMM-2 (collapsed)	0.0007	0.0159	0.0624	0.1140	1.0143
	2s-sGMM-2 (collapsed)	0.0012	0.0160	0.0588	0.1100	1.0196
γ	2s-QML	0.0050	0.6631	0.0524	0.1052	0.9908
	1s-sGMM-2 (full)	-0.4768	1.0565	0.1048	0.1680	1.2914
	1s-sGMM-2 (2 lags)	-0.7009	0.8383	0.4136	0.5068	0.9807
	1s-sGMM-1 (collapsed)	-0.0385	0.6478	0.0724	0.1264	0.9724
	1s-sGMM-2 (collapsed)	-0.1087	0.6770	0.0812	0.1376	0.9759
	2s-sGMM-2 (collapsed)	-0.0321	0.7024	0.0612	0.1164	0.9980

Note: We abbreviate the estimators as follows: "1s" and "2s" refer to one-stage and two-stage estimators, respectively. "QML" is the quasi-maximum likelihood estimator of Hsiao et al. (2002), and "sGMM" refers to system GMM estimators. The subsequent digit declares one-step and two-step GMM estimators. We follow Blundell et al. (2001) to form the first-step weighting matrix (41). In parenthesis, we refer to the set of instruments. The bias statistic measures the average bias relative to the true parameter value, e.g. $(\hat{\lambda} - \lambda)/\lambda$. RMSE is the root mean square error. The size statistics refer to the actual rejection rates of Wald tests that the parameter estimates equal their true values, given a nominal size of 5% and 10%, respectively. SE/SD is the average standard error relative to the standard deviation of the estimator for the 2500 replications. GMM standard errors are based on formula (38) with an unrestricted estimate of Ξ and the Windmeijer (2005) correction for two-step estimators. Two-stage standard errors account for the first-stage estimation error.

of the average standard error to the observed standard deviation is almost unity which supports our adjustment for the first-stage estimation error.

When we look at the different GMM specifications it is evident that the one-stage system GMM estimator with the full set of instruments strongly suffers from instrument proliferation.[34] With 6.4% for λ and -47.7% for γ its bias is far off from any acceptable range. For β the bias is smaller, only -1.2%, but largest among the estimators under consideration. To the contrary, it shows the lowest RMSE among the GMM estimators for the latter coefficient. As a consequence of the large biases on the one hand and too

[34]Discarding the instruments that result from the moment conditions (27) for the time-invariant regressors does not change the results much.

conservative standard errors, after applying the Windmeijer (2005) correction, on the other hand, this estimator also shows considerable size distortions. For λ and γ the Wald tests overreject the null hypothesis while there is underrejection for β.

As we discuss in Section 4.3 and Appendix A, the first choice to reduce the number of instruments might be a restriction of the lag depth in forming the instrument matrix for the model in first differences. Choosing a maximum lag depth of two lags reduces the instrument count from 174 to 65 in our example. However, our results reveal that this approach does even more harm as the bias for λ and γ increases sizably. The efficiency loss from disregarding a large amount of information seems to outweigh the benefits of a more parsimonious instrument set. This is even more pronounced when we have a look at the Wald tests. For λ the null hypothesis is rejected in 76% of the cases for a nominal size of 5%. For γ the rejection rate is 41%. Surprisingly, the lag depth restriction seems to work well for the coefficient β both in terms of bias and size statistics.

The second possibility to obtain a set of instruments that grows linearly in T instead of quadratically is to collapse the instrument matrices into smaller blocks. With this approach we retain the whole available information in a condensed form and at the same time reduce the instrument count further to 33. The simulation results clearly provide evidence in favor of this approach. The bias of all three parameters is reduced strongly. The RMSE also improves considerably. In particular for λ and β, the Wald tests are still oversized as a consequence of too small standard errors although the rejection rates are much closer to the nominal size than in the previous cases.

The comparison of one-step and two-step GMM estimators yields a noteworthy insight. As expected the feasible efficient two-step system GMM estimator tends to produce more precise estimates of λ and β than the one-step analog. The bias is almost cut in half for λ and decreases to less than one third for β. Interestingly, this picture turns upside down for the coefficient γ of the time-invariant regressor. The bias is almost three times larger for the two-step estimator and also the RMSE increases slightly.[35] This seems to be a consequence of lower weights that the estimated second-step weighting matrix puts on the time-invariant instruments that convey the relevant information for the identification of γ.

As an alternative to one-stage GMM estimation that obtains all coefficients simultaneously we consider a two-stage GMM estimator that recovers the coefficients of the time-invariant regressors in a second stage. Since the quality of the second-stage results depends crucially on the precision of the first-stage estimates and as a consequence of the above finding, we use a GMM estimator with a collapsed instrument set in the first stage. Moreover, the two-stage approach allows us to exploit the efficiency gains of two-step GMM estimation for the time-varying regressors and still to put full weight on

[35] For clarity, we only report the comparison between the one-step and the two-step system GMM estimator for the collapsed instrument set. The picture is similar for other instrument choices. For the GMM estimator with the full instrument set the RMSE of γ deteriorates even stronger for the two-step estimator (1.06) compared to the one-step estimator (0.69). However, the bias is slightly lower (in absolute value) for the two-step estimator (-48% versus -50%).

Table 3: Second-Stage Standard Errors for the Baseline Calibration

Coefficient	Estimator	Two-Stage	Conventional	Robust
Design 1, $T = 10$, $N = 50$				
γ	2s-QML	0.9908	0.9649	0.9433
	2s-sGMM-2 (collapsed)	0.9980	0.8975	0.8776

Note: See notes to Table 2 for a description of the estimators. We report the average standard error of $\hat{\gamma}$ relative to its standard deviation for the 2500 replications. Two-stage standard errors account for the first-stage estimation error, conventional standard errors assume homoscedastic error terms, and robust standard errors allow for heteroscedasticity through time.

the time-invariant instruments in the second stage. As a result, our two-stage estimator shows the lowest bias (in absolute value) among the group of GMM estimators. The cost of the two-stage approach is a moderately larger RMSE for γ that seems to be acceptable in the light of the improved size statistics. The improved size statistics seem to be a direct consequence of the more precise estimation of the standard errors. Table 3 shows that our correction of the second-stage standard errors to account for the first-stage estimation error performs well in comparison to uncorrected standard error estimates. The average ratio of the adjusted standard errors to the empirical standard deviation is close to unity for both two-stage estimators while the conventional standard errors and the heteroscedasticity-robust standard errors are too small. Moreover, the correction leads to standard error estimates that are more precise than the standard error estimates of the one-stage estimators.

The results for the other parameter designs broadly confirm the findings for the baseline calibration. However, a closer look at Tables 5 to 7 in Appendix D reveals some interesting further insights. Under design 2, when the variances of the unit-specific effects and the disturbances both equal unity, the GMM estimators can reduce the gap to the QML estimator although the latter retains its leading position. In particular, the system GMM estimator with the full instrument set benefits considerably from the reduction of the variance of the unit-specific effects. Yet, the weaknesses of one-stage two-step GMM estimators for the coefficients of time-invariant regressors are especially evident in this design. "1s-sGMM-2 (full)" produces a RMSE for γ that is 17 times as high as its one-step analog.[36] Also, the Wald tests now strongly underreject the null hypothesis that $\hat{\lambda}$ or $\hat{\gamma}$ equal their true value, respectively, while there is overrejection under the baseline calibration.

When the persistence parameter λ is reduced to 0.4, as in design 3, the average bias and the RMSE of the two-stage QML approach increase for all three parameters. We can observe the same effect for the GMM estimates of λ. For β and γ, the GMM estimators tend to produce less biased but more dispersed estimates. Noteworthy, although the two-

[36]Under design 2, the RMSE of "1s-sGMM-1 (full)" is 0.3565 for γ.

stage estimation procedures again have the smallest bias for γ, the two-stage two-step system GMM estimator now shows a smaller bias (in absolute terms) than the two-stage QML estimator. Also, the QML estimator now appears to suffer from an underestimation of the standard errors. The average ratio to the empirical standard deviation is only 0.46 for λ which is mainly a consequence of some outlying coefficient estimates close to unity. There are 39 estimates out of the 2500 replications that can be classified as outliers because $\hat{\lambda}$ lies in the interval $(0.99, 1.08)$ while none of the estimates lies between 0.56 and 0.99. This bias in the first-stage standard errors also carries over partly to the second-stage standard errors.

Design 4 sets the correlation between the time-invariant regressor and the exogenous time-varying regressor to zero. We first note that the first-stage estimates of the QML approach are not affected at all by this change since it is solely based on the model in first differences that does not involve the time-invariant regressor. Importantly, the estimates for γ improve when f_i is uncorrelated with x_{it}. We can observe the strongest gains in terms of bias reduction for the two-stage GMM estimator. For the latter the average bias is almost cut in half. Nevertheless, the QML procedure still shows the smallest bias.

Tables 8 to 10 in Appendix D show again simulation results for the baseline design 1 but now for different sample sizes. The ranking of the estimators remains mostly unchanged when the number of cross-sectional units increases. For $N = 500$ the one-stage two-step system GMM estimator with the full set of instruments still produces a bias of -5% for the coefficient of the time-invariant regressor and suffers from notable size distortions for the parameters λ and γ. When the number of time periods is reduced to $T = 5$ the bias tends to increase, in particular for γ. For the system GMM estimator with non-collapsed instruments the bias goes up to -68%. The two-stage QML estimator still produces by far the best results with an average bias of -2.7%. The GMM estimation procedure all reveal considerable biases of at least 15% in combination with large size distortions.

7 Empirical Application: Dynamic Wage Regression

Factors that influence the rates of labor income have long been studied in theoretical models and empirical applications. The seminal work of Mincer (1974) laid the ground for a vast strand of literature in modern labor economics analyzing the impact of human capital on wages often referred to as the return to schooling. Mincer (1974) derives an earnings function that depends on the number of years of education and experience, as well as the squared number of years of experience. Andini (2007) introduces a dynamic version of the Mincer equation that adds previous period's labor income as an additional explanatory variable. Andini (2010a) argues in favor of the dynamic approach "that observed earnings do not instantaneously adjust to net potential earnings". With our empirical application we take up this idea and estimate a dynamic version of the Mincer equation controlling for additional factors. We include several time-invariant factors to analyze their potential impact on wages with the dynamic panel data methods discussed in this paper.

We use data for 882 individuals from the PSID. The time span of our sample ranges from 1985 to 1992. We only include household heads and wifes that report salary income for each of the eight consecutive years to obtain a balanced panel. Our dependent variable is the natural logarithm of salary per hour. Besides the lagged dependent variable, and labor market experience and its square, we include as a set of 10 industry dummy variables as additional time-varying controls. All remaining regressors are considered time-invariant even though some of them actually show some variation over time. However, this variation is often very small compared to the cross-sectional variation which might lead to a weak instruments problem for GMM estimators, or poor identification with any estimator solely based on the first-differenced equation. These variables are our schooling variable, age and squared age, public sector employment, labor union membership, and geographic region. For each of these pseudo time-invariant regressors we take their realization in 1992 and assign it to all time periods. Finally, we add gender and race as truly time-invariant regressors.

We proxy education by the number of years of schooling. Because all individuals in our sample are employed in all years, most of them already reached their final level of education before the initial year of the sample. This justifies our approach of using only the cross-sectional variation to identify the return to schooling. We also consider the age and squared age of the individuals as time-invariant regressors to circumvent collinearity problems with labor market experience. In first differences, both age and experience would shrink to a constant without cross-sectional variation, and we would not be able to identify their coefficients separately with any estimator that is solely based on the first-differenced equation as it is the case for the QML estimator. With the latter, identification of the coefficient of experience is based on its deterministic rise over time, while the age effect is identified solely through the cross-sectional variation in our second stage.

According to Spence (1973) workers choose their level of education to signal their ability to potential employers. Therefore, years of schooling are positively correlated with the unobservable ability of a worker. Not controlling for this endogeneity would lead to an upward bias in the estimation of the return to schooling when higher ability is associated with higher wages.[37] In the absence of excluded instruments, we follow the Hausman and Taylor (1981) identification strategy. Besides the lagged dependent variable and education we classify all regressors as uncorrelated with unobserved individual-specific ability. In particular, we want to use the level of the industry dummy for "professional and related services" in 1992 as an instrument for education. We notice that the sample correlation in our data set between this dummy variable and education is 0.31. The correlation conditional on the other time-invariant regressors, the partial R^2, is still 0.1134, and the F-statistic for significance of the industry dummy in the regression of

[37] See Boissiere et al. (1985).

the instruments on the endogenous variable is 111.445.[38] In this industry, education does not only serve as a signal of the workers to their potential employers but the firms themselves are interested in a high education of their workers to signal their expertise to potential clients. However, we argue that allocation of workers across industries is not a matter of ability but a matter of worker's preferences. Krueger and Summers (1988) and Blackburn and Neumark (1992) find that ability cannot explain the inter-industry wage differentials which supports our approach to treat the industry dummy variables as exogenous. We do not add the other industry dummys and labor market experience to the set of instruments because their correlation with education is weak.

We estimate this dynamic Mincer equation with a one-stage and a two-stage system GMM estimator, and the two-stage QML estimator. Both GMM estimators are two-step estimators with a collapsed set of instruments, and where the first-step weighting matrix (41) is formed with $\mathbf{H}_2$. Standard errors are computed with the Windmeijer (2005) correction and our two-stage variance formula (21). We present the results first under the assumption that education is exogenous, and second assuming that it is correlated with unobserved ability. For the two-stage GMM estimator we treat all time-varying regressors as endogenous in the first stage because the time-invariant regressors are now part of the individual-specific effect.[39]

Table 4 presents the estimation results. We recognize that the salary does not immediately adjust to changes in net potential earnings since previous period salary is a significant predictor of current salary. For labor market experience we find the hump-shaped profile that is consistent with the Mincerian theory. Among the other regressors we want to focus primarily on the return to schooling. When we ignore the potential endogeneity of education we obtain significantly positiv coefficients for the schooling variable. Due to our dynamic setting, this coefficient should be interpreted as the effect of one additional year of schooling on the annual change of salary holding all other factors fixed. This effect ranges between 5.3 and 6.8 percentage points. The differences are related to the diverse speeds of adjustment. The implied long-run effects of education on the level of salary are very similar and range from 10.3 to 10.5 percent.[40] When comparing the results with the second specification that accounts for the endogeneity of education, recall that the first-stage results for the time-varying regressors remain unaffected. The estimated return to schooling is reduced as expected. For the one-stage system GMM estimator the short-run effect goes down to 4.7 percentage points and is still significant at the 5 percent level. For the two-stage estimators the point estimates are only 0.4 to 1.4 percentage points. Moreover, they are not significant any more. The implied long-run effect from the two-stage estimators is about 0.6 to 2.7 percent, although insignificant

[38] Stock, Wright, and Yogo (2002) suggest that the F-statistic should exceed 10 for reliable inference.

[39] As this application shall have illustrative character, we ignore sample selection issues and refrain from an extensive discussion of the exogeneity assumption that we impose on the other regressors besides education.

[40] Let λ be the coefficient of the lagged dependent variable and β the coefficient of any other regressor. The corresponding long-run coefficient is given as β divided by the speed of adjustment $1 - \lambda$.

Table 4: Estimation Results: Dynamic Mincer Regression

	Education exogenous			Education endogenous		
ln (Salary)	1s-sGMM	2s-sGMM	2s-QML	1s-sGMM	2s-sGMM	2s-QML
ln $(\text{Salary})_{t-1}$	0.4094	0.4850	0.3481	0.4120	0.4850	0.3481
	(0.0493)***	(0.0545)***	(0.0162)***	(0.0494)***	(0.0545)***	(0.0162)***
Experience	0.0477	0.0556	0.0650	0.0477	0.0556	0.0650
	(0.0059)***	(0.0072)***	(0.0039)***	(0.0059)***	(0.0072)***	(0.0039)***
Experience^2	-0.0005	-0.0009	-0.0006	-0.0005	-0.0009	-0.0006
	(0.0001)***	(0.0001)***	(0.0001)***	(0.0001)***	(0.0001)***	(0.0001)***
Industry	base category: Agriculture, Mining			base category: Agriculture, Mining		
Construction	0.0116	-0.0421	-0.0405	0.0063	-0.0421	-0.0405
	(0.0553)	(0.0553)	(0.0434)	(0.0558)	(0.0553)	(0.0434)
Manufacturing	0.0008	-0.0704	-0.0541	0.0018	-0.0704	-0.0541
	(0.0450)	(0.0364)*	(0.0290)*	(0.0442)	(0.0364)*	(0.0290)*
Public Utilites	-0.0142	-0.0666	-0.0814	-0.0170	-0.0666	-0.0814
	(0.0459)	(0.0471)	(0.0363)**	(0.0452)	(0.0471)	(0.0363)**
Trade	-0.0532	-0.0853	-0.0549	-0.0550	-0.0853	-0.0549
	(0.0468)	(0.0397)**	(0.0337)	(0.0464)	(0.0397)**	(0.0337)
Financial Services	-0.0259	-0.0689	-0.0828	-0.0244	-0.0689	-0.0828
	(0.0480)	(0.0486)	(0.0374)**	(0.0474)	(0.0486)	(0.0374)**
Business Services	0.0045	-0.0456	-0.0545	0.0086	-0.0456	-0.0545
	(0.0440)	(0.0376)	(0.0318)*	(0.0436)	(0.0376)	(0.0318)*
Personal Services	-0.1014	-0.1760	-0.0829	-0.1044	-0.1760	-0.0829
	(0.1001)	(0.0636)***	(0.0774)	(0.0994)	(0.0636)***	(0.0774)
Entertainment	-0.1100	-0.1868	-0.1683	-0.1164	-0.1868	-0.1683
	(0.1113)	(0.2135)	(0.0846)**	(0.1104)	(0.2135)	(0.0846)**
Professional Services	-0.0798	-0.1046	-0.0741	-0.0691	-0.1046	-0.0741
	(0.0499)	(0.0469)**	(0.0317)**	(0.0511)	(0.0469)**	(0.0317)**
Public Administration	0.0270	-0.0562	-0.0170	0.0188	-0.0562	-0.0170
	(0.0502)	(0.0496)	(0.0352)	(0.0509)	(0.0496)	(0.0352)
Education	0.0623	0.0529	0.0675	0.0472	0.0141	0.0037
	(0.0071)***	(0.0074)***	(0.0060)***	(0.0214)**	(0.0221)	(0.0209)
Male	0.0714	0.1062	0.0821	0.0849	0.1320	0.1245
	(0.0213)***	(0.0256)***	(0.0241)***	(0.0298)***	(0.0341)***	(0.0298)***
White	0.0093	0.0179	0.0314	0.0172	0.0357	0.0607
	(0.0226)	(0.0261)	(0.0302)	(0.0235)	(0.0288)	(0.0333)*
Age	-0.0244	-0.0387	-0.0429	-0.0221	-0.0285	-0.0261
	(0.0107)**	(0.0128)***	(0.0127)***	(0.0112)**	(0.0135)**	(0.0142)*
Age^2	0.0001	0.0004	0.0002	0.0001	0.0003	-0.0000
	(0.0001)	(0.0001)***	(0.0001)	(0.0001)	(0.0001)*	(0.0002)
Government Work	-0.1346	-0.0940	-0.1396	-0.1228	-0.0635	-0.0895
	(0.0326)***	(0.0347)***	(0.0292)***	(0.0353)***	(0.0340)*	(0.0333)***
Labor Union	0.0179	0.0231	0.0229	0.0210	0.0367	0.0452
	(0.0259)	(0.0311)	(0.0358)	(0.0262)	(0.0322)	(0.0383)
Region	base category: South			base category: South		
Northeast	0.0272	-0.0055	-0.0112	0.0261	-0.0061	-0.0122
	(0.0301)	(0.0307)	(0.0355)	(0.0302)	(0.0316)	(0.0377)
North Central	-0.0560	-0.1034	-0.1234	-0.0524	-0.0977	-0.1141
	(0.0301)*	(0.0316)***	(0.0355)***	(0.0306)*	(0.0324)***	(0.0379)***
West	-0.0479	-0.0752	-0.0903	-0.0548	-0.0920	-0.1180
	(0.0283)*	(0.0300)**	(0.0330)***	(0.0295)*	(0.0335)***	(0.0362)***
Constant	3.5694	3.3715	4.3026	3.7097	3.6990	4.8413
	(0.3866)***	(0.4247)***	(0.2996)***	(0.4301)***	(0.5111)***	(0.3642)***
Observations	6,174	6,174	6,174	6,174	6,174	6,174
Individuals	882	882	882	882	882	882
Instruments (GMM)	178	167		177	167	

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Note: We abbreviate the estimators as follows: "1s" and "2s" refer to one-stage and two-stage estimators, respectively. "QML" is the quasi-maximum likelihood estimator of Hsiao et al. (2002), and "sGMM" refers to two-step system GMM estimators. We follow Blundell et al. (2001) to form the first-step weighting matrix (41), and use a collapsed set of instruments. Standard errors are in parenthesis. GMM standard errors are based on formula (38) with an unrestricted estimate of Ξ and the Windmeijer (2005) correction for two-step estimators. Two-stage standard errors account for the first-stage estimation error.

as well. The different estimates of the one- and two-stage system GMM estimator can be partly explained by the set of instruments used to identify the return to schooling. While the two-stage estimator is exactly identified as it uses a single industry dummy variable as instrument for education, the one-stage estimator makes use of a whole set of overidentifying restrictions. Many of those additional instruments are weakly correlated with education. As a consequence, the one-stage estimate of the return to schooling tends to be biased towards the non-instrumented estimate.

Furthermore, we want to highlight the importance of the second-stage variance correction. Ignoring this correction could result in misleading inference. As an example, take the coefficient for government work from the two-stage GMM estimator with endogenous education. The reported standard error with the appropriate two-stage correction is 0.034 which implies significance on the 10 percent level only, while without the adjustment it is 0.026 which would incorrectly signal significance on the 5 percent level.

8 Conclusion

Estimation of linear dynamic panel data models with unobserved unit-specific heterogeneity is challenging when the time dimension is short. The identification of the coefficients of time-invariant regressors poses additional complications. Yet, these parameters can be of special interest.

The identification of the coefficients of time-invariant regressors requires additional assumptions on the orthogonality of the regressors and the unobserved unit-specific effects. These orthogonality assumptions imply additional moment conditions that we can use to form a GMM estimator that estimates all parameters simultaneously. As an alternative we propose a two-stage estimation strategy. In the first stage, we subsume the time-invariant regressors under the unit-specific effects and estimate the coefficients of the time-varying regressors. In the second stage, we apply an instrumental variable regression of the first-stage residuals for the last period on the time-invariant regressors. Both time-varying and time-invariant variables that are assumed to be uncorrelated with the unit-specific effects qualify as instruments in the second stage.

We can base the first-stage regression on any estimator that consistently estimates the coefficients of the time-varying regressors without relying on estimates of the coefficients of time-invariant regressors. In this paper, we discuss GMM-type estimators and the transformed maximum likelihood estimator of Hsiao et al. (2002) as potential first-stage candidates. The latter is entirely based on the model in first differences and thus necessarily requires the two-stage approach to identify the coefficients of time-invariant regressors. In general, the two-stage approach is not restricted to models with a short time dimension. When the time span is large, a potential first-stage estimator might be the classical fixed effects estimator. The same is true for static models.

For GMM-type estimators the two-stage approach has three main advantages compared to the estimation of all parameters at once. First, the estimation of the coefficients of the time-varying regressors is robust to a model misspecification with regard to the time-invariant regressors. Second, too many overidentifying restrictions can bias the coefficients of endogenous time-invariant regressors towards their non-instrumented estimates. These overidentifying restrictions naturally arise from the presence of time-varying regressors in one-stage system GMM estimation, while in the second stage of our two-stage procedure the number of instruments can be easily reduced by selecting only the appropriate ones. Third, the estimated weighting matrix of the feasible efficient two-step system GMM estimator puts less weight on the time-invariant instruments than the inefficient one-step analog. Our Monte Carlo experiments confirm that this results in less precisely estimated coefficients of the time-invariant regressors because their identification hinges on the information that the time-invariant instruments convey. The two-stage approach circumvents this issue.

Our Monte Carlo analysis furthermore points out that GMM estimators that are based on the full set of available moment conditions suffer from instrument proliferation already at a modest time span. Reducing the number of instruments by collapsing the instrument matrices into smaller blocks improves the finite sample performance considerably. In contrast, reducing the instrument count by limiting the number of lags used to form the instrument matrices leads to a deterioration of the results. While the former approach still uses the whole information in a condensed form, the latter discards a large amount of the available information completely. This insight is also important for our two-stage approach because large first-stage estimation errors translate into poor second-stage estimates. In particular for the coefficients of time-invariant regressors only GMM estimators with collapsed instrument matrices produce reliable results. When we compare the various GMM specifications with the transformed likelihood approach, our simulation results provide strong evidence in favor of the latter in the presence of strictly exogenous time-varying regressors.

Importantly, the two-stage approach requires an adjustment of the second-stage standard errors due to the additional variation that is coming from the first-stage estimation error. We provide the asymptotic variance formula for the second-stage estimator. Our Monte Carlo results demonstrate that the adjustment works well and is quantitatively important. The relevance of the standard error correction is also demonstrated in our empirical application.

References

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

Amemiya, T. and T. E. MaCurdy (1986). Instrumental-Variable Estimation of an Error-Components Model. *Econometrica 54*(4), 869–880.

Andini, C. (2007). Returns to education and wage equations: a dynamic approach. *Applied Economics Letters 14*(8), 577–579.

Andini, C. (2010a). A dynamic Mincer equation with an application to Portuguese data. *Applied Economics 42*(16), 2091–2098.

Andini, C. (2010b). How Fast Do Wages Adjust to Human-Capital Productivity? Dynamic Panel-Data Evidence from Europe and the United States. CEEAplA Working Paper 01/2010, Centre of Applied Economics Studies of the Atlantic.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press.

Arellano, M. and S. R. Bond (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies 58*(2), 277–297.

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

Bhargava, A. and J. D. Sargan (1983). Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods. *Econometrica 51*(6), 1635–1659.

Binder, M., C. Hsiao, and M. H. Pesaran (2005). Estimation and Inference in Short Panel Vector Autoregressions With Unit Roots and Cointegration. *Econometric Theory 21*(4), 795–837.

Blackburn, M. and D. Neumark (1992). Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials. *Quarterly Journal of Economics 107*(4), 1421–1436.

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

Blundell, R., S. R. Bond, and F. Windmeijer (2001). Estimation in dynamic panel data models: Improving on the performance of the standard GMM estimator. *Advances in Econometrics 15*(1), 53–91.

Boissiere, M. X., J. B. Knight, and R. H. Sabot (1985). Earnings, Schooling, Ability, and Cognitive Skills. *American Economic Review 75*(5), 1016–1030.

Breusch, T. S., G. E. Mizon, and P. Schmidt (1989). Efficient Estimation Using Panel Data. *Econometrica 57*(3), 695–700.

Breusch, T. S., M. B. Ward, H. T. M. Nguyen, and T. Kompas (2011). On the Fixed-Effects Vector Decomposition. *Political Analysis 19*(2), 123–134.

Bun, M. J. G. and F. Windmeijer (2010). The weak instrument problem of the system GMM estimator in dynamic panel data models. *Econometrics Journal 13*(1), 95–126.

Chamberlain, G. (1982). Multivariate Regression Models for Panel Data. *Journal of Econometrics 18*(1), 5–46.

Cinyabuguma, M. M. and L. Putterman (2011). Sub-Saharan Growth Surprises: Being Heterogeneous, Inland and Close to the Equator Does not Slow Growth Within Africa. *Journal of African Economies 20*(2), 217–262.

Greene, W. H. (2011). Fixed Effects Vector Decomposition: A Magical Solution to the Problem of Time-Invariant Variables in Fixed Effects Models? *Political Analysis 19*(2), 135–146.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica 50*(4), 1029–1054.

Hausman, J. A. and W. E. Taylor (1981). Panel Data and Unobservable Individual Effects. *Econometrica 49*(6), 1377–1398.

Hayakawa, K. and M. H. Pesaran (2012). Robust Standard Errors in Transformed Likelihood Estimation of Dynamic Panel Data Models. IZA Discussion Paper 6583, Institute for the Study of Labor.

Hoeffler, A. E. (2002). The augmented Solow model and the African growth debate. *Oxford Bulletin of Economics and Statistics 64*(2), 135–158.

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

Hurwicz, L. (1950). Least-Squares Bias in Time Series. In T. C. Koopmans (Ed.), *Statistical Inference in Dynamic Economic Models*, Chapter 15, pp. 365–383. New York: Wiley.

Krueger, A. B. and L. H. Summers (1988). Efficiency Wages and the Inter-Industry Wage Structure. *Econometrica 56*(2), 259–293.

Mehrhoff, J. (2009). A solution to the problem of too many instruments in dynamic panel data GMM. Discussion Paper, Series 1: Economic Studies 31/2009, Deutsche Bundesbank.

Mincer, J. A. (1974). *Schooling, Experience, and Earnings*. Columbia University Press.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica 46*(1), 69–85.

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

Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica 49*(6), 1417–1426.

Plümper, T. and V. E. Troeger (2007). Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects. *Political Analysis 15*(2), 124–139.

Roodman, D. (2009). A Note on the Theme of Too Many Instruments. *Oxford Bulletin of Economics and Statistics 71*(1), 135–158.

Spence, M. (1973). Job Market Signaling. *Quarterly Journal of Economics 87*(3), 355–374.

Stock, J. H., J. H. Wright, and M. Yogo (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business & Economic Statistics 20*(4), 518–529.

Windmeijer, F. (2000). Efficiency Comparisons for a System GMM Estimator in Dynamic Panel Data Models. In R. D. H. Heijmans, D. S. G. Pollock, and A. Sattora (Eds.), *Innovations in Multivariate Statistical Analysis: A Festschrift for Heinz Neudecker*, Chapter 11, pp. 175–184. Dordrecht: Kluwer Academic Publishers.

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

A Transformations of GMM Instruments

This appendix provides examples of the transformation matrix $\mathbf{R}$ that are relevant in practical applications. For simplicity, we disregard the moment conditions (28) that are based on the homoscedasticity of u_{it}. In the following, we restrict our attention to block-diagonal versions of $\mathbf{R}$:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}^d & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^l \end{pmatrix},$$

such that $\mathbf{Z}_i^* = [\mathbf{D}'\mathbf{Z}_i^d\mathbf{R}^d, \mathbf{Z}_i^l\mathbf{R}^l]$. Similarly, we consider a block-diagonal partition of the transformation matrix for the first-differenced equation:

$$\mathbf{R}^d = \begin{pmatrix} \mathbf{R}_y^d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_x^d \otimes \mathbf{I}_{K_x} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_f^d \otimes \mathbf{I}_{K_f} \end{pmatrix}.$$

Often, the instrument count is reduced by restricting the number of lags used to construct the instrument matrix. This procedure is equivalent to the construction of a transformation matrix $\mathbf{R}^d$ that selects the appropriate columns of the full matrix $\mathbf{Z}_i^d$. As an example, the following matrices restrict the lag depth to $\kappa \geq 1$ for both $y_{i,t-1}$ and strictly exogenous $\mathbf{x}_{it}$ while also discarding future values of the latter:

$$\mathbf{R}_y^d = \begin{pmatrix} 1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{2,\kappa} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_{3,\kappa} & & \vdots \\ \vdots & \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{J}_{T-1,\kappa} \end{pmatrix}, \quad \mathbf{R}_x^d = \begin{pmatrix} \mathbf{J}_{1,\kappa}^* & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{2,\kappa}^* & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & & \vdots \\ \vdots & \vdots & & \mathbf{J}_{T-2,\kappa}^* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{J}_{T+1,\kappa} \end{pmatrix},$$

where $\mathbf{J}_{s,\kappa} = \mathbf{I}_s$ if $s \leq \kappa$, and $\mathbf{J}_{s,\kappa} = (\mathbf{0}, \mathbf{I}_\kappa)'$ with dimension $s \times \kappa$ if $s > \kappa$, and $\mathbf{J}_{s,\kappa}^* = (\mathbf{J}_{s+2,\kappa}', \mathbf{0})'$ with dimension $(T+1) \times \min\{s+2, \kappa\}$. We set $\mathbf{R}_f^d = \mathbf{I}_{T-1}$ in this case.

Alternatively, the dimension of the instrument matrix can be reduced by collapsing it into smaller blocks. The following transformation matrices linearly combine the columns of $\mathbf{Z}_i^d$, again for the case of strictly exogenous regressors $\mathbf{x}_{it}$:

$$\mathbf{R}_y^d = \begin{pmatrix} \tilde{\mathbf{J}}_{0,1,T-2} \\ \tilde{\mathbf{J}}_{0,2,T-3} \\ \vdots \\ \tilde{\mathbf{J}}_{0,T-2,1} \\ \tilde{\mathbf{I}}_{T-1} \end{pmatrix}, \quad \mathbf{R}_x^d = \begin{pmatrix} \tilde{\mathbf{J}}_{0,T+1,T-2} \\ \tilde{\mathbf{J}}_{1,T+1,T-3} \\ \vdots \\ \tilde{\mathbf{J}}_{T-3,T+1,1} \\ \tilde{\mathbf{J}}_{T-2,T+1,0} \end{pmatrix},$$

where $\tilde{\mathbf{J}}_{s_1,s_2,s_3} = (\mathbf{0}_{s_2 \times s_1}, \tilde{\mathbf{I}}_{s_2}, \mathbf{0}_{s_2 \times s_3})$ with dimension $s_2 \times (s_1 + s_2 + s_3)$, and $\tilde{\mathbf{I}}_{s_2}$ is the s_2-dimensional mirror identity matrix with ones on the antidiagonal and zeros elsewhere.

$\mathbf{Z}_{y,i}^d \mathbf{R}_y^d$ now corresponds to the collapsed matrix described by Roodman (2009). As a consequence, the $T(T-1)/2$ moment conditions (24) are replaced by the $(T-1)$ conditions $E\left[\sum_{t=s}^T y_{i,t-s}\Delta u_{it}\right] = 0$, $s = 2, 3, \ldots, T$. Similarly, the information contained in the $K_x(T+1)(T-1)$ moment conditions (25) is condensed into $K_x(2T-1)$ conditions. The instrument block containing $\mathbf{f}_i$ can be collapsed by setting $\mathbf{R}_f^d = \boldsymbol{\iota}_{T-1}$. The implied K_f moment conditions are $E[\mathbf{f}_i(u_{iT} - u_{i1})] = \mathbf{0}$ instead of the $K_f(T-1)$ conditions (27). The transformation matrices can be further adjusted to combine the collapsing approach with the lag depth restriction.

The instruments for the level equation, for clarity ignoring the moment conditions (33), can be collapsed into a set of standard instruments by applying the following transformation:

$$\mathbf{R}^l = \begin{pmatrix} \boldsymbol{\iota}_{T-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\iota}_T \otimes \mathbf{I}_{K_{x1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\iota}_T \otimes \mathbf{I}_{K_{x2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{K_{f1}} \end{pmatrix},$$

such that $\mathbf{Z}_i^l \mathbf{R}^l = [(0, \Delta\mathbf{y}_{i,-1}')', \mathbf{X}_{1i}, \Delta\mathbf{X}_{2i}^*, \mathbf{F}_{1i}]$.

B Consistency of the First-Stage QML Estimator

The following consistency proof for the estimated first-stage parameter vector $\hat{\boldsymbol{\varphi}}$ follows closely the lines in Hsiao et al. (2002) for their minimum distance estimator in the absence of additional regressors. In the generalized case with exogenous regressors, we have:

$$\hat{\boldsymbol{\varphi}} = \boldsymbol{\varphi} + \left(\sum_{i=1}^N \Delta\tilde{\mathbf{W}}_i' \hat{\tilde{\Omega}}^{-1} \Delta\tilde{\mathbf{W}}_i\right)^{-1} \sum_{i=1}^N \Delta\tilde{\mathbf{W}}_i' \hat{\tilde{\Omega}}^{-1} \Delta\mathbf{u}_i^*. \tag{66}$$

We will now show that the last term has a zero mean. Therefore, we make use of the matrix decomposition $\tilde{\Omega}^{-1} = \mathbf{A}'\tilde{\mathbf{A}}^{-1}\mathbf{A}$, proposed by Hsiao et al. (2002), where

$$\mathbf{A} = \begin{pmatrix} a_0 & 0 & 0 & \cdots & 0 \\ a_0 & a_1 & 0 & \cdots & 0 \\ a_0 & a_1 & a_2 & & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ a_0 & a_1 & a_2 & \cdots & a_{T-1} \end{pmatrix}, \quad \tilde{\mathbf{A}} = \begin{pmatrix} a_0 a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_1 a_2 & 0 & \cdots & 0 \\ 0 & 0 & a_2 a_3 & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & a_{T-1} a_T \end{pmatrix},$$

and $a_{s+1} - 2a_s + a_{s-1} = 0, s = 1, 2, \ldots, T$, with $a_0 = 1$ and $a_1 = \omega$. We can now show that the last term in the expression above equals:

$$\sum_{i=1}^{N} \Delta\tilde{\mathbf{W}}_i'\hat{\hat{\Omega}}^{-1}\Delta\mathbf{u}_i^* =$$

$$\sum_{i=1}^{N} \begin{pmatrix} a_0^2 \sum_{j=1}^{T}(a_{j-1}a_j)^{-1}\xi_{i1} + a_0 \sum_{j=2}^{T}(a_{j-1}a_j)^{-1} \sum_{k=1}^{j-1} a_k \Delta u_{i,k+1} \\ a_0 \sum_{j=1}^{T}(a_{j-1}a_j)^{-1}\Delta\tilde{\mathbf{x}}_i\xi_{i1} + a_0 \sum_{j=2}^{T}(a_{j-1}a_j)^{-1} \sum_{k=1}^{j-1} a_k \Delta\tilde{\mathbf{x}}_i \Delta u_{i,k+1} \\ \sum_{j=2}^{T}(a_{j-1}a_j)^{-1} \sum_{k=1}^{j-1} a_k \Delta y_{ik} \left(a_0\xi_{i1} + \sum_{l=1}^{j-1} a_l \Delta u_{i,l+1}\right) \\ \sum_{j=2}^{T}(a_{j-1}a_j)^{-1} \sum_{k=1}^{j-1} a_k \Delta\mathbf{x}_{i,k+1} \left(a_0\xi_{i1} + \sum_{l=1}^{j-1} a_l \Delta u_{i,l+1}\right) \end{pmatrix}. \quad (67)$$

Note that in the case of strictly exogenous regressors $\xi_{i1} = q_{i1} + \sum_{j=0}^{\infty} \lambda^j \Delta u_{i,1-j}$ under the assumption that the data generating process of y_{it} started in the infinite past, where q_{i1} is independently distributed of $\Delta\mathbf{x}_i$ with mean zero and constant variance. Under strict exogeneity of $\mathbf{x}_{it}$, the first, second and last entry of the above vector are obviously zero in expectations. It remains to show that the expected value of the third entry is zero as well. Therefore, we make use of the following relationships:

$$\Delta y_{it} = \lambda^{t-1}\Delta y_{i1} + \sum_{j=0}^{t-2} \lambda^j \Delta\mathbf{x}_{i,t-j}'\boldsymbol{\beta} + \sum_{j=0}^{t-2} \lambda^j \Delta u_{i,t-j}, \quad t = 2, 3, \ldots, T, \quad (68)$$

$$\Delta y_{i1} = \sum_{j=0}^{\infty} \lambda^j \Delta\mathbf{x}_{i,1-j}'\boldsymbol{\beta} + \sum_{j=0}^{\infty} \lambda^j \Delta u_{i,1-j}. \quad (69)$$

With $E[\Delta y_{i1}\xi_{i1}] = \sigma_\xi^2 = \omega\sigma_u^2 = a_1\sigma_u^2$ we get:

$$\begin{aligned} E[\Delta y_{it}\xi_{i1}] &= \lambda^{t-1}E[\Delta y_{i1}\xi_{i1}] + \lambda^{t-2}E[\Delta u_{i2}\Delta u_{i1}] \\ &= \lambda^{t-1}a_1\sigma_u^2 - \lambda^{t-2}\sigma_u^2 \\ &= \lambda^{t-2}(a_1\lambda - 1)\sigma_u^2, \quad t = 2, 3, \ldots, T, \end{aligned} \quad (70)$$

$$E[\Delta y_{it}\Delta u_{i,t+s}] = \begin{cases} 0, & s = 2, 3, \ldots, T - t \\ -\sigma_u^2, & s = 1 \\ (2-\lambda)\sigma_u^2, & s = 0 \end{cases}, \quad (71)$$

$$\begin{aligned} E[\Delta y_{it}\Delta u_{i,t-s}] &= E\left[\sum_{j=0}^{\infty} \lambda^j \Delta u_{i,t-j}\Delta u_{i,t-s}\right] \\ &= -\lambda^{s-1}\sigma_u^2 + 2\lambda^s\sigma_u^2 - \lambda^{s+1}\sigma_u^2 \\ &= -\lambda^{s-1}(1-\lambda)^2\sigma_u^2, \quad s = 1, 2, \ldots, \end{aligned} \quad (72)$$

and consequently:

$$
\begin{aligned}
E&\left[\sum_{k=1}^{j-1} a_k \Delta y_{ik}\left(a_0\xi_{i1} + \sum_{l=1}^{j-1} a_l \Delta u_{i,l+1}\right)\right] \\
&=\Bigg[a_1^2 + \sum_{k=2}^{j-1} a_k\lambda^{k-2}(a_1\lambda - a_0) + (2-\lambda)\sum_{k=2}^{j-1} a_{k-1}a_k \\
&\quad - (1-\lambda)^2\sum_{k=3}^{j-1} a_k \sum_{l=3}^{k} a_{l-2}\lambda^{k-l} - \sum_{k=1}^{j-1} a_k^2\Bigg]\sigma_u^2 \\
&= -\,\sigma_u^2 \sum_{k=2}^{j-1} a_k \sum_{l=2}^{k} \lambda^{k-l}(a_l - 2a_{l-1} + a_{l-2}) = 0. \qquad (73)
\end{aligned}
$$

C Derivation of the Coefficient of Determination

The two processes in first-differences are:

$$
\begin{aligned}
\Delta y_{it} &= \lambda \Delta y_{i,t-1} + \beta \Delta x_{it} + \Delta u_{it} \\
&= \beta \sum_{j=0}^{\infty} \lambda^j \Delta x_{i,t-j} + \sum_{j=0}^{\infty} \lambda^j \Delta u_{i,t-j}, \qquad (74) \\
\Delta x_{it} &= \phi \Delta x_{i,t-1} + \Delta\epsilon_{it} \\
&= \sum_{j=0}^{\infty} \phi^j \Delta\epsilon_{i,t-j}. \qquad (75)
\end{aligned}
$$

The unconditional variance of Δy_{it} can then be written as:

$$
\begin{aligned}
Var(\Delta y_{it}) &= Cov(\Delta y_{it}, \lambda\Delta y_{i,t-1} + \beta\Delta x_{it} + \Delta u_{it}) \\
&= \lambda Cov(\Delta y_{it}, \Delta y_{i,t-1}) + \beta Cov(\Delta y_{it}, \Delta x_{it}) + Cov(\Delta y_{it}, \Delta u_{it}). \qquad (76)
\end{aligned}
$$

We need to determine the individual components and start with the last term:

$$
\begin{aligned}
Cov(\Delta y_{it}, \Delta u_{it}) &= Cov\left(\beta \sum_{j=0}^{\infty} \lambda^j \Delta x_{i,t-j} + \sum_{j=0}^{\infty} \lambda^j \Delta u_{i,t-j}, \Delta u_{it}\right) \\
&= Cov(\Delta u_{it} + \lambda \Delta u_{i,t-1}, \Delta u_{it}) = (2-\lambda)\sigma_u^2, \qquad (77)
\end{aligned}
$$

since the disturbances are i.i.d. and uncorrelated with x_{it} due to the strict exogeneity assumption. For the remaining terms we need to calculate the variance and autocovariances of Δx_{it} first:

$$
\begin{aligned}
Var(\Delta x_{it}) &= Cov(\Delta x_{it}, \phi\Delta x_{i,t-1} + \Delta\epsilon_{it}) \\
&= \phi Cov(\Delta x_{it}, \Delta x_{i,t-1}) + Cov(\Delta x_{it}, \Delta\epsilon_{it}), \qquad (78)
\end{aligned}
$$

where

$$
\begin{aligned}
Cov(\Delta x_{it}, \Delta\epsilon_{it}) &= Cov\left(\sum_{j=0}^{\infty}\phi^j\Delta\epsilon_{i,t-j}, \Delta\epsilon_{it}\right) \\
&= Cov(\Delta\epsilon_{it} + \phi\Delta\epsilon_{i,t-1}, \Delta\epsilon_{it}) = (2-\phi)\sigma_\epsilon^2, \qquad (79)
\end{aligned}
$$

and

$$
\begin{aligned}
Cov(\Delta x_{it}, \Delta x_{i,t-1}) &= Cov(\phi\Delta x_{i,t-1} + \Delta\epsilon_{it}, \Delta x_{i,t-1}) \\
&= \phi Var(\Delta x_{i,t-1}) + Cov(\Delta x_{i,t-1}, \Delta\epsilon_{it}). \qquad (80)
\end{aligned}
$$

Also,

$$
\begin{aligned}
Cov(\Delta x_{i,t-1}, \Delta\epsilon_{it}) &= Cov\left(\sum_{j=0}^{\infty}\phi^j\Delta\epsilon_{i,t-1-j}, \Delta\epsilon_{it}\right) \\
&= Cov(\Delta\epsilon_{i,t-1}, \Delta\epsilon_{it}) = -\sigma_\epsilon^2. \qquad (81)
\end{aligned}
$$

Together, the above results for the variance and first-order autocovariance of Δx_{it} yield:

$$
\begin{aligned}
Var(\Delta x_{it}) &= \phi\left(\phi Var(\Delta x_{i,t-1}) - \sigma_\epsilon^2\right) + (2-\phi)\sigma_\epsilon^2 \\
&= \frac{2(1-\phi)}{1-\phi^2}\sigma_\epsilon^2 = \frac{2}{1+\phi}\sigma_\epsilon^2, \qquad (82)
\end{aligned}
$$

since $Var(\Delta x_{it}) = Var(\Delta x_{i,t-1})$ due to stationarity, and $1-\phi^2 = (1-\phi)(1+\phi)$. Consequently,

$$
Cov(\Delta x_{it}, \Delta x_{i,t-1}) = \left(\frac{2\phi}{1+\phi} - 1\right)\sigma_\epsilon^2 = \frac{2\phi-(1+\phi)}{1+\phi}\sigma_\epsilon^2 = -\frac{1-\phi}{1+\phi}\sigma_\epsilon^2. \qquad (83)
$$

For use below, the higher order autocovariances of Δx_{it} follow straightforwardly from the first-order autocovariance since $Cov(\Delta x_{i,t-j}, \Delta\epsilon_{it}) = 0\ \forall j \geq 2$:

$$
\begin{aligned}
Cov(\Delta x_{it}, \Delta x_{i,t-j}) &= Cov(\phi\Delta x_{i,t-1} + \Delta\epsilon_{it}, \Delta x_{i,t-j}) \\
&= \phi Cov(\Delta x_{i,t-1}, \Delta x_{i,t-j}) + Cov(\Delta x_{i,t-j}, \Delta\epsilon_{it}) \\
&= \phi Cov(\phi\Delta x_{i,t-2} + \Delta\epsilon_{i,t-1}, \Delta x_{i,t-1-j}) \\
&= \phi^2 Cov(\Delta x_{i,t-2}, \Delta x_{i,t-j}) \\
&= \dots \\
&= \phi^{j-1}(\Delta x_{i,t-(j-1)}, \Delta x_{i,t-j}) \\
&= -\phi^{j-1}\frac{1-\phi}{1+\phi}\sigma_\epsilon^2 \quad \forall j \geq 2, \qquad (84)
\end{aligned}
$$

since $Cov(\Delta x_{i,t-(j-1)}, \Delta x_{i,t-j}) = Cov(\Delta x_{it}, \Delta x_{i,t-1})$ again due to stationarity. Now, we can derive an expression for the second term in (76):

$$\begin{aligned}
Cov(\Delta y_{it}, \Delta x_{it}) &= Cov\left(\beta \sum_{j=0}^{\infty} \lambda^j \Delta x_{i,t-j} + \sum_{j=0}^{\infty} \lambda^j \Delta u_{i,t-j}, \Delta x_{it}\right) \\
&= \beta \left(Var(\Delta x_{it}) + \sum_{j=1}^{\infty} \lambda^j Cov(\Delta x_{it}, \Delta x_{i,t-j})\right) \\
&= \beta \left(\frac{2}{1+\phi} - \frac{1-\phi}{1+\phi} \sum_{j=1}^{\infty} \lambda^j \phi^{j-1}\right) \sigma_\epsilon^2 \\
&= \beta \left(\frac{2}{1+\phi} - \frac{1-\phi}{1+\phi} \frac{\lambda}{1-\lambda\phi}\right) \sigma_\epsilon^2 \\
&= \beta \frac{2(1-\lambda\phi) - \lambda(1-\phi)}{(1+\phi)(1-\lambda\phi)} \sigma_\epsilon^2 = \beta \frac{2-\lambda(1+\phi)}{(1+\phi)(1-\lambda\phi)} \sigma_\epsilon^2. \qquad (85)
\end{aligned}$$

It remains to determine the first-order autocovariance of Δy_{it}:

$$\begin{aligned}
Cov(\Delta y_{it}, \Delta y_{i,t-1}) &= Cov(\lambda \Delta y_{i,t-1} + \beta \Delta x_{it} + \Delta u_{it}, \Delta y_{i,t-1}) \\
&= \lambda Var(\Delta y_{it}) + \beta Cov(\Delta y_{i,t-1}, \Delta x_{it}) + Cov(\Delta y_{i,t-1}, \Delta u_{it}), \qquad (86)
\end{aligned}$$

where $Var(\Delta y_{i,t-1}) = Var(\Delta y_{it})$ again due to stationarity, and

$$\begin{aligned}
Cov(\Delta y_{i,t-1}, \Delta u_{it}) &= Cov\left(\sum_{j=0}^{\infty} \lambda^j \Delta u_{i,t-1-j}, \Delta u_{it}\right) \\
&= Cov(\Delta u_{i,t-1}, \Delta u_{it}) = -\sigma_u^2. \qquad (87)
\end{aligned}$$

Moreover,

$$\begin{aligned}
Cov(\Delta y_{i,t-1}, \Delta x_{it}) &= Cov\left(\beta \sum_{j=0}^{\infty} \lambda^j \Delta x_{i,t-1-j} + \sum_{j=0}^{\infty} \lambda^j \Delta u_{i,t-1-j}, \Delta x_{it}\right) \\
&= \beta \sum_{j=0}^{\infty} \lambda^j Cov(\Delta x_{it}, \Delta x_{i,t-1-j}) \\
&= \beta \sum_{j=1}^{\infty} \lambda^{j-1} Cov(\Delta x_{it}, \Delta x_{i,t-j}) \\
&= -\beta \frac{1-\phi}{1+\phi} \sum_{j=1}^{\infty} (\lambda\phi)^{j-1} \sigma_\epsilon^2 = -\beta \frac{1-\phi}{(1+\phi)(1-\lambda\phi)} \sigma_\epsilon^2. \qquad (88)
\end{aligned}$$

Finally, we can insert all results into (76) and obtain:

$$
\begin{aligned}
Var(\Delta y_{it}) =& \lambda \left(\lambda Var(\Delta y_{it}) - \beta^2 \frac{1-\phi}{(1+\phi)(1-\lambda\phi)} \sigma_\epsilon^2 - \sigma_u^2 \right) \\
&+ \beta^2 \frac{2-\lambda(1+\phi)}{(1+\phi)(1-\lambda\phi)} \sigma_\epsilon^2 + (2-\lambda)\sigma_u^2 \\
=& \frac{1}{1-\lambda^2} \left(\beta^2 \frac{2-\lambda(1+\phi)-(1-\phi)}{(1+\phi)(1-\lambda\phi)} \sigma_\epsilon^2 + 2(1-\lambda)\sigma_u^2 \right) \\
=& \frac{1}{1-\lambda^2} \left(\beta^2 \frac{2(1-\lambda)}{(1+\phi)(1-\lambda\phi)} \sigma_\epsilon^2 + 2(1-\lambda)\sigma_u^2 \right) \\
=& \frac{2}{1+\lambda} \left(\frac{\beta^2}{(1+\phi)(1-\lambda\phi)} \sigma_\epsilon^2 + \sigma_u^2 \right). \qquad (89)
\end{aligned}
$$

The conditional variance of Δy_{it} given the realizations of current and past Δx_{it} is simply:

$$
Var(\Delta y_{it} | \Delta x_{it}, \Delta x_{i,t-1}, \ldots) = \frac{2}{1+\lambda} \sigma_u^2. \qquad (90)
$$

Taking everything together we get the coefficient of determination for the first-differenced model:

$$
\begin{aligned}
R^2_{\Delta y} &= 1 - \frac{Var(\Delta y_{it} | \Delta x_{it}, \Delta x_{i,t-1}, \ldots)}{Var(\Delta y_{it})} \\
&= 1 - \frac{\sigma_u^2}{\frac{\beta^2}{(1+\phi)(1-\lambda\phi)} \sigma_\epsilon^2 + \sigma_u^2} = \frac{\beta^2 \sigma_\epsilon^2}{\beta^2 \sigma_\epsilon^2 + (1+\phi)(1-\lambda\phi)\sigma_u^2}. \qquad (91)
\end{aligned}
$$

D Tables with Simulation Results

Table 5: Summary Results for Alternative Calibrations

Coefficient	Estimator	Bias	RMSE	Size (5%)	Size (10%)	SE/SD
Design 2, $T = 10$, $N = 50$						
λ	2s-QML	-0.0005	0.0236	0.0596	0.1092	0.9043
	1s-sGMM-2 (full)	0.0090	0.0265	0.0284	0.0620	1.2357
	1s-sGMM-2 (2 lags)	0.0321	0.0374	0.2048	0.2904	0.9729
	1s-sGMM-1 (collapsed)	-0.0033	0.0286	0.0676	0.1176	0.9457
	1s-sGMM-2 (collapsed)	-0.0036	0.0299	0.0680	0.1224	0.9520
	2s-sGMM-2 (collapsed)	-0.0027	0.0304	0.0720	0.1236	0.9485
β	2s-QML	-0.0006	0.0103	0.0480	0.1000	0.9850
	1s-sGMM-2 (full)	0.0006	0.0131	0.0200	0.0464	1.2218
	1s-sGMM-2 (2 lags)	0.0021	0.0144	0.0528	0.0976	1.0146
	1s-sGMM-1 (collapsed)	-0.0022	0.0134	0.0556	0.1088	0.9948
	1s-sGMM-2 (collapsed)	0.0016	0.0136	0.0592	0.1092	1.0040
	2s-sGMM-2 (collapsed)	0.0022	0.0139	0.0588	0.1140	1.0059
γ	2s-QML	0.0049	0.4457	0.0492	0.0996	0.9834
	1s-sGMM-2 (full)	-0.0663	6.0548	0.0132	0.0308	0.8904
	1s-sGMM-2 (2 lags)	-0.2418	0.4377	0.1492	0.2140	1.0029
	1s-sGMM-1 (collapsed)	0.0275	0.3993	0.0576	0.1128	0.9800
	1s-sGMM-2 (collapsed)	-0.0233	0.4204	0.0712	0.1232	0.9764
	2s-sGMM-2 (collapsed)	0.0133	0.4867	0.0532	0.1052	1.0058

Note: We abbreviate the estimators as follows: "1s" and "2s" refer to one-stage and two-stage estimators, respectively. "QML" is the quasi-maximum likelihood estimator of Hsiao et al. (2002), and "sGMM" refers to system GMM estimators. The subsequent digit declares one-step and two-step GMM estimators. We follow Blundell et al. (2001) to form the first-step weighting matrix (41). In parenthesis, we refer to the set of instruments. The bias statistic measures the average bias relative to the true parameter value, e.g. $(\hat{\lambda} - \lambda)/\lambda$. RMSE is the root mean square error. The size statistics refer to the actual rejection rates of Wald tests that the parameter estimates equal their true values, given a nominal size of 5% and 10%, respectively. SE/SD is the average standard error relative to the standard deviation of the estimator for the 2500 replications. GMM standard errors are based on formula (38) with an unrestricted estimate of Ξ and the Windmeijer (2005) correction for two-step estimators. Two-stage standard errors account for the first-stage estimation error.

Table 6: Summary Results for Alternative Calibrations

Coefficient	Estimator	Bias	RMSE	Size (5%)	Size (10%)	SE/SD
Design 3, $T = 10$, $N = 50$						
λ	2s-QML	0.0186	0.0900	0.0708	0.1196	0.4616
	1s-sGMM-2 (full)	0.2034	0.1013	0.2060	0.3020	1.1702
	1s-sGMM-2 (2 lags)	0.3485	0.1536	0.6348	0.7336	0.9131
	1s-sGMM-1 (collapsed)	-0.0124	0.0598	0.0580	0.1100	0.9724
	1s-sGMM-2 (collapsed)	-0.0148	0.0637	0.0604	0.1128	0.9874
	2s-sGMM-2 (collapsed)	-0.0128	0.0648	0.0664	0.1172	0.9729
β	2s-QML	-0.0070	0.0616	0.0708	0.1140	0.8155
	1s-sGMM-2 (full)	0.0048	0.0741	0.0164	0.0440	1.3121
	1s-sGMM-2 (2 lags)	0.0299	0.0890	0.0540	0.1068	1.0100
	1s-sGMM-1 (collapsed)	0.0016	0.0761	0.0564	0.1180	0.9808
	1s-sGMM-2 (collapsed)	0.0066	0.0736	0.0552	0.1116	1.0061
	2s-sGMM-2 (collapsed)	0.0088	0.0739	0.0580	0.1112	1.0052
γ	2s-QML	-0.0214	0.7465	0.0692	0.1200	0.9051
	1s-sGMM-2 (full)	-0.4728	0.7733	0.1200	0.1936	1.0943
	1s-sGMM-2 (2 lags)	-0.8602	1.0601	0.3516	0.4492	0.9771
	1s-sGMM-1 (collapsed)	0.0303	0.7137	0.0628	0.1144	0.9820
	1s-sGMM-2 (collapsed)	-0.0242	0.7410	0.0696	0.1208	0.9889
	2s-sGMM-2 (collapsed)	0.0082	0.7625	0.0532	0.1016	0.9998

Note: We abbreviate the estimators as follows: "1s" and "2s" refer to one-stage and two-stage estimators, respectively. "QML" is the quasi-maximum likelihood estimator of Hsiao et al. (2002), and "sGMM" refers to system GMM estimators. The subsequent digit declares one-step and two-step GMM estimators. We follow Blundell et al. (2001) to form the first-step weighting matrix (41). In parenthesis, we refer to the set of instruments. The bias statistic measures the average bias relative to the true parameter value, e.g. $(\hat{\lambda} - \lambda)/\lambda$. RMSE is the root mean square error. The size statistics refer to the actual rejection rates of Wald tests that the parameter estimates equal their true values, given a nominal size of 5% and 10%, respectively. SE/SD is the average standard error relative to the standard deviation of the estimator for the 2500 replications. GMM standard errors are based on formula (38) with an unrestricted estimate of Ξ and the Windmeijer (2005) correction for two-step estimators. Two-stage standard errors account for the first-stage estimation error.

Table 7: Summary Results for Alternative Calibrations

Coefficient	Estimator	Bias	RMSE	Size (5%)	Size (10%)	SE/SD
Design 4, $T = 10$, $N = 50$						
λ	2s-QML	-0.0002	0.0253	0.0600	0.1096	0.8460
	1s-sGMM-2 (full)	0.0634	0.0575	0.3736	0.5016	1.1819
	1s-sGMM-2 (2 lags)	0.0957	0.0815	0.7748	0.8460	0.9212
	1s-sGMM-1 (collapsed)	0.0066	0.0343	0.0936	0.1484	0.9121
	1s-sGMM-2 (collapsed)	0.0037	0.0353	0.0860	0.1396	0.9307
	2s-sGMM-2 (collapsed)	0.0039	0.0358	0.0864	0.1384	0.9240
β	2s-QML	-0.0007	0.0103	0.0480	0.1000	0.9853
	1s-sGMM-2 (full)	-0.0108	0.0145	0.0248	0.0564	1.2111
	1s-sGMM-2 (2 lags)	0.0075	0.0169	0.0588	0.1076	1.0018
	1s-sGMM-1 (collapsed)	-0.0021	0.0179	0.0652	0.1192	1.0142
	1s-sGMM-2 (collapsed)	0.0011	0.0159	0.0604	0.1084	1.0108
	2s-sGMM-2 (collapsed)	0.0014	0.0161	0.0580	0.1128	1.0171
γ	2s-QML	0.0039	0.6396	0.0492	0.1056	1.0070
	1s-sGMM-2 (full)	-0.3108	1.2915	0.0548	0.0876	0.8194
	1s-sGMM-2 (2 lags)	-0.3912	0.5542	0.2316	0.3160	1.0031
	1s-sGMM-1 (collapsed)	-0.0220	0.5884	0.0556	0.1096	0.9854
	1s-sGMM-2 (collapsed)	-0.0782	0.6180	0.0632	0.1188	0.9852
	2s-sGMM-2 (collapsed)	-0.0170	0.6460	0.0476	0.1028	1.0136

Note: We abbreviate the estimators as follows: "1s" and "2s" refer to one-stage and two-stage estimators, respectively. "QML" is the quasi-maximum likelihood estimator of Hsiao et al. (2002), and "sGMM" refers to system GMM estimators. The subsequent digit declares one-step and two-step GMM estimators. We follow Blundell et al. (2001) to form the first-step weighting matrix (41). In parenthesis, we refer to the set of instruments. The bias statistic measures the average bias relative to the true parameter value, e.g. $(\hat{\lambda} - \lambda)/\lambda$. RMSE is the root mean square error. The size statistics refer to the actual rejection rates of Wald tests that the parameter estimates equal their true values, given a nominal size of 5% and 10%, respectively. SE/SD is the average standard error relative to the standard deviation of the estimator for the 2500 replications. GMM standard errors are based on formula (38) with an unrestricted estimate of Ξ and the Windmeijer (2005) correction for two-step estimators. Two-stage standard errors account for the first-stage estimation error.

Table 8: Summary Results for Alternative Sample Sizes

Coefficient	Estimator	Bias	RMSE	Size (5%)	Size (10%)	SE/SD
Design 1, $T = 10$, $N = 200$						
λ	2s-QML	-0.0001	0.0106	0.0496	0.0980	0.9986
	1s-sGMM-2 (full)	0.0281	0.0267	0.3996	0.5108	0.9204
	1s-sGMM-2 (2 lags)	0.0316	0.0315	0.3032	0.4180	0.9296
	1s-sGMM-1 (collapsed)	0.0020	0.0164	0.0596	0.1140	0.9876
	1s-sGMM-2 (collapsed)	0.0001	0.0155	0.0544	0.1012	1.0067
	2s-sGMM-2 (collapsed)	0.0002	0.0156	0.0524	0.1032	1.0046
β	2s-QML	-0.0001	0.0050	0.0444	0.0988	0.9992
	1s-sGMM-2 (full)	-0.0034	0.0072	0.0520	0.1024	0.9995
	1s-sGMM-2 (2 lags)	-0.0023	0.0082	0.0468	0.1028	1.0017
	1s-sGMM-1 (collapsed)	0.0004	0.0099	0.0592	0.1136	0.9918
	1s-sGMM-2 (collapsed)	0.0019	0.0067	0.0480	0.0964	1.0091
	2s-sGMM-2 (collapsed)	0.0020	0.0067	0.0476	0.0976	1.0091
γ	2s-QML	-0.0068	0.3288	0.0524	0.1016	0.9936
	1s-sGMM-2 (full)	-0.2035	0.3503	0.1288	0.2040	0.9998
	1s-sGMM-2 (2 lags)	-0.2333	0.3897	0.1588	0.2420	0.9544
	1s-sGMM-1 (collapsed)	-0.0213	0.3205	0.0580	0.0972	1.0084
	1s-sGMM-2 (collapsed)	-0.0619	0.3346	0.0644	0.1220	0.9854
	2s-sGMM-2 (collapsed)	-0.0115	0.3471	0.0508	0.1020	1.0020

Note: We abbreviate the estimators as follows: "1s" and "2s" refer to one-stage and two-stage estimators, respectively. "QML" is the quasi-maximum likelihood estimator of Hsiao et al. (2002), and "sGMM" refers to system GMM estimators. The subsequent digit declares one-step and two-step GMM estimators. We follow Blundell et al. (2001) to form the first-step weighting matrix (41). In parenthesis, we refer to the set of instruments. The bias statistic measures the average bias relative to the true parameter value, e.g. $(\hat{\lambda} - \lambda)/\lambda$. RMSE is the root mean square error. The size statistics refer to the actual rejection rates of Wald tests that the parameter estimates equal their true values, given a nominal size of 5% and 10%, respectively. SE/SD is the average standard error relative to the standard deviation of the estimator for the 2500 replications. GMM standard errors are based on formula (38) with an unrestricted estimate of Ξ and the Windmeijer (2005) correction for two-step estimators. Two-stage standard errors account for the first-stage estimation error.

Table 9: Summary Results for Alternative Sample Sizes

Coefficient	Estimator	Bias	RMSE	Size (5%)	Size (10%)	SE/SD
Design 1, $T = 10$, $N = 500$						
λ	2s-QML	-0.0001	0.0067	0.0472	0.1060	0.9962
	1s-sGMM-2 (full)	0.0074	0.0100	0.1372	0.2012	0.9644
	1s-sGMM-2 (2 lags)	0.0093	0.0129	0.1116	0.1924	0.9725
	1s-sGMM-1 (collapsed)	0.0007	0.0102	0.0508	0.0956	1.0055
	1s-sGMM-2 (collapsed)	-0.0001	0.0095	0.0508	0.0908	1.0050
	2s-sGMM-2 (collapsed)	0.0000	0.0095	0.0524	0.0928	1.0042
β	2s-QML	0.0004	0.0031	0.0412	0.0944	1.0189
	1s-sGMM-2 (full)	-0.0008	0.0038	0.0508	0.0964	1.0177
	1s-sGMM-2 (2 lags)	-0.0022	0.0047	0.0464	0.0996	1.0124
	1s-sGMM-1 (collapsed)	0.0001	0.0063	0.0528	0.0960	1.0138
	1s-sGMM-2 (collapsed)	0.0009	0.0041	0.0460	0.0924	1.0098
	2s-sGMM-2 (collapsed)	0.0010	0.0041	0.0444	0.0944	1.0117
γ	2s-QML	-0.0046	0.2104	0.0584	0.1048	0.9807
	1s-sGMM-2 (full)	-0.0535	0.2109	0.0724	0.1296	0.9675
	1s-sGMM-2 (2 lags)	-0.0675	0.2203	0.0920	0.1532	0.9304
	1s-sGMM-1 (collapsed)	-0.0084	0.2093	0.0540	0.1028	0.9803
	1s-sGMM-2 (collapsed)	-0.0286	0.2093	0.0632	0.1152	0.9729
	2s-sGMM-2 (collapsed)	-0.0053	0.2217	0.0544	0.1104	0.9852

Note: We abbreviate the estimators as follows: "1s" and "2s" refer to one-stage and two-stage estimators, respectively. "QML" is the quasi-maximum likelihood estimator of Hsiao et al. (2002), and "sGMM" refers to system GMM estimators. The subsequent digit declares one-step and two-step GMM estimators. We follow Blundell et al. (2001) to form the first-step weighting matrix (41). In parenthesis, we refer to the set of instruments. The bias statistic measures the average bias relative to the true parameter value, e.g. $(\hat{\lambda} - \lambda)/\lambda$. RMSE is the root mean square error. The size statistics refer to the actual rejection rates of Wald tests that the parameter estimates equal their true values, given a nominal size of 5% and 10%, respectively. SE/SD is the average standard error relative to the standard deviation of the estimator for the 2500 replications. GMM standard errors are based on formula (38) with an unrestricted estimate of Ξ and the Windmeijer (2005) correction for two-step estimators. Two-stage standard errors account for the first-stage estimation error.

Table 10: Summary Results for Alternative Sample Sizes

Coefficient	Estimator	Bias	RMSE	Size (5%)	Size (10%)	SE/SD
Design 1, $T = 5$, $N = 50$						
λ	2s-QML	0.0021	0.0559	0.0588	0.1124	0.9120
	1s-sGMM-2 (full)	0.0918	0.0843	0.5360	0.6152	0.9194
	1s-sGMM-2 (2 lags)	0.0999	0.0921	0.5172	0.6048	0.9393
	1s-sGMM-1 (collapsed)	0.0171	0.0619	0.1088	0.1644	0.9700
	1s-sGMM-2 (collapsed)	0.0146	0.0649	0.1148	0.1752	0.9617
	2s-sGMM-2 (collapsed)	0.0173	0.0655	0.1240	0.1828	0.9525
β	2s-QML	0.0007	0.0185	0.0548	0.1084	0.9752
	1s-sGMM-2 (full)	0.0047	0.0216	0.0656	0.1172	0.9856
	1s-sGMM-2 (2 lags)	0.0002	0.0249	0.0612	0.1116	0.9967
	1s-sGMM-1 (collapsed)	0.0037	0.0281	0.0668	0.1188	1.0359
	1s-sGMM-2 (collapsed)	0.0070	0.0243	0.0684	0.1104	0.9938
	2s-sGMM-2 (collapsed)	0.0080	0.0243	0.0644	0.1172	0.9983
γ	2s-QML	-0.0270	0.8277	0.0828	0.1376	0.8855
	1s-sGMM-2 (full)	-0.6765	0.8666	0.3632	0.4528	0.9810
	1s-sGMM-2 (2 lags)	-0.7285	0.9301	0.3628	0.4608	0.9829
	1s-sGMM-1 (collapsed)	-0.1482	0.7992	0.1116	0.1644	0.9944
	1s-sGMM-2 (collapsed)	-0.1767	0.8389	0.1156	0.1712	0.9881
	2s-sGMM-2 (collapsed)	-0.1622	0.8500	0.0920	0.1440	1.0028

Note: We abbreviate the estimators as follows: "1s" and "2s" refer to one-stage and two-stage estimators, respectively. "QML" is the quasi-maximum likelihood estimator of Hsiao et al. (2002), and "sGMM" refers to system GMM estimators. The subsequent digit declares one-step and two-step GMM estimators. We follow Blundell et al. (2001) to form the first-step weighting matrix (41). In parenthesis, we refer to the set of instruments. The bias statistic measures the average bias relative to the true parameter value, e.g. $(\hat{\lambda} - \lambda)/\lambda$. RMSE is the root mean square error. The size statistics refer to the actual rejection rates of Wald tests that the parameter estimates equal their true values, given a nominal size of 5% and 10%, respectively. SE/SD is the average standard error relative to the standard deviation of the estimator for the 2500 replications. GMM standard errors are based on formula (38) with an unrestricted estimate of Ξ and the Windmeijer (2005) correction for two-step estimators. Two-stage standard errors account for the first-stage estimation error.