Accounting for persistence in panel count data models. An application to the number of patents awarded

Stefanos Dimitrakopoulos*

Department of Accounting, Finance and Economics, Oxford Brookes University, Oxford, OX33 1HX, UK

Summary

We propose a Poisson regression model that controls for three potential sources of persistence in panel count data; dynamics, latent heterogeneity and serial correlation in the idiosyncratic errors. We also account for the initial conditions problem. For model estimation, we develop a Markov Chain Monte Carlo algorithm. The proposed methodology is illustrated by a real example on the number of patents granted.

Keywords: dynamics, initial conditions, latent heterogeneity, Markov Chain Monte Carlo, panel count data, serial correlation

JEL classification: C1, C5, C11, C13

1 Introduction

There is a vast econometrics literature on the analysis of count data (Winkelmann, 2008; Cameron and Trivedi, 2013). In this paper we propose a Poisson model that

*Correspondence to: Stefanos Dimitrakopoulos, E-mail: sdimitrakopoulos@brookes.ac.uk.

accounts for three potential sources of the persistent behaviour of counts across economic units; true state dependence, spurious state dependence and serial error correlation.

True state dependence is modelled through a lagged dependent variable that controls for dynamic effects, spurious state dependence is captured by a latent random variable (Heckman, 1981) that controls for unobserved heterogeneity, while serial correlation in the idiosyncratic errors is assumed to follow a first-order stationary autoregressive process. The resulting model specification is a dynamic panel Poisson model with latent heterogeneity and serially correlated errors.

We also account for an inherent problem in our model, that of the endogeneity of the initial count for each cross-sectional unit (initial conditions problem). The assumption of exogenous initial conditions produces biased and inconsistent estimates (Fotouhi, 2005). To tackle this problem we apply the approach of Wooldridge (2005) that attempts to model the relationship between the unobserved heterogeneity and initial values.

In the context of Poisson regression analysis of event counts, researchers have proposed dynamic Poisson models with unobserved heterogeneity (Crépon and Duguet, 1997; Blundell et al., 2002) in order to disentangle true and spurious state dependence. Yet, the issue of persistence (true state dependence, spurious state dependence, serial error correlation) as well as the initial values problem have not been properly addressed in panel counts. This paper aspires to fill this gap.

To estimate the parameters of the proposed model, we develop a Markov Chain Monte Carlo (MCMC) algorithm, the efficiency of which is evaluated with a simulation study. We also conduct model comparison. Our methodology is illustrated with an empirical example on patenting.

The paper is organized as follows. In section 2 we set up the proposed model and in section 3 we describe the posterior analysis. The empirical results are presented in section 4. Section 5 concludes. An Online Appendix accompanies this paper.

2 Econometric framework

Let y_{it} be the observed count outcome for individual $i = 1, ..., N$ at time $t = 1, ..., T$, that follows the Poisson distribution with conditional mean λ_{it}

$$f(y_{it}; \lambda_{it}) = \frac{\lambda_{it}^{y_{it}} \exp(-\lambda_{it})}{y_{it}!}. \tag{1}$$

For λ_{it} we assume the following exponential mean function

$$\lambda_{it} = exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{it-1} + \varphi_i + \epsilon_{it}), \tag{2}$$

where $\mathbf{x}_{it}$=$(x_{1,it}, .., x_{k,it})'$ is a vector of exogenous covariates[1] that contains an intercept, φ_i denotes the individual-specific random effect that controls for spurious state dependence, whereas the coefficient on y_{it-1} measures the strength of true state dependence.

Since y_{it} is non-negative, a positive coefficient γ makes the model explosive as $\gamma y_{it-1} > 0$. To overcome this problem we replace y_{it-1} in (2) by its logarithm, $\ln y_{it-1}$, and then use a strictly positive transformation y^*_{it-1} of the y_{it-1} values, when $y_{it-1} = 0$. In particular, we rescale only the zero values of y_{it-1} to a constant c, that is, $y^*_{it-1} = max(y_{it-1}, c), c \in (0, 1)$; see also Zeger and Qaqish (1988). Therefore, expression (2) is replaced by

$$\lambda_{it} = exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma \ln y^*_{it-1} + \varphi_i + \epsilon_{it}). \tag{3}$$

For the idiosyncratic error terms ϵ_{it}, we assume the following first-order stationary ($|\rho| < 1$) autoregressive specification

$$\epsilon_{it} = \rho\epsilon_{it-1} + v_{it}, \quad -1 < \rho < 1, \quad v_{it} \overset{i.i.d}{\sim} N(0, \sigma_v^2). \tag{4}$$

[1]Addressing the issue of potential violation of the exogeneity assumption in the context of the proposed model is a changeling econometric task and thus is left for future research; see also Biewen (2009) for potential treatment.

The random variables v_{it} are independent and identically normally distributed across all i and t with mean zero and variance σ_v^2. We also assume that v_{it} and φ_i are mutually independent.

To tackle the initial values problem we follow the approach of Wooldridge (2005) and model φ_i as follows

$$\varphi_i = h_1 \ln y_{i0}^* + \overline{\mathbf{x}}_i' \mathbf{h}_2 + u_i, \quad u_i \sim N(0, \sigma_u^2), \quad i = 1, ..., N. \tag{5}$$

As before, if the first available count in the sample for individual i, y_{i0}, is zero, it is rescaled to a constant c, that is, $y_{i0}^* = max(y_{i0}, c), c \in (0, 1)$. Also, $\overline{\mathbf{x}}_i$ is the time average of $\mathbf{x}_{it}$ and u_i is a stochastic disturbance, which is assumed to be uncorrelated with y_{i0} and $\overline{\mathbf{x}}_i$. For identification reasons, time-constant regressors that maybe included in $\mathbf{x}_{it}$ should be excluded from $\overline{\mathbf{x}}_i$.

To conduct Bayesian analysis we impose priors over the parameters ($\boldsymbol{\delta}$, $\mathbf{h}, \rho, \sigma_v^2$, σ_u^2),

$$p(\boldsymbol{\delta}) \propto 1, \mathbf{h} \sim \mathbf{N}_{k+1}(\widetilde{\mathbf{h}}, \widetilde{\mathbf{H}}),$$

$$\rho \sim N(\rho_0, \sigma_\rho^2) I_{(-1,1)}(\rho), \; \sigma_v^{-2} \sim \mathcal{G}(\frac{e_1}{2}, \frac{f_1}{2}), \sigma_u^2 \sim \sim \mathcal{IG}(\frac{e_0}{2}, \frac{f_0}{2}),$$

where $\boldsymbol{\delta} = (\boldsymbol{\beta}', \gamma)'$, $\mathbf{h} = (h_1, \mathbf{h}_2)'$, $\mathcal{G}$ denotes the gamma distribution and $\mathcal{IG}$ denotes the inverse gamma distribution. The prior distribution for $\boldsymbol{\delta}$ is flat. A truncated normal is imposed on ρ.

3 Posterior analysis

3.1 MCMC algorithm

To estimate the model parameters, we follow closely the paper by Chib and Jeliazkov (2006) and develop a similar MCMC algorithm that augments the parameter space (Tanner and Wong, 1987) to include the latent variables $\{\lambda_{it}^*\}_{i \geq 1, t \geq 1}$, where $\lambda_{it}^* =$

$\mathbf{w}_{it}'\boldsymbol{\delta} + \varphi_i + \epsilon_{it}$ and $\mathbf{w}_{it}' = (\mathbf{x}_{it}', \ln y_{it-1}^*)$.

The details of the estimation method are given in the Online Appendix, where we also conduct a Monte Carlo experiment.

3.2 Model comparison

For model comparison we compute the marginal likelihood (ML). There are many ways to do that. One popular numerical method is the method of Chib (1995) and Chib and Jeliazkov (2001); see, also, Chib et al. (1998). In this paper we use the Bayesian Information Criterion (BIC)- (Schwarz, 1978). As an alternative model comparison criterion, we also calculate cross-validation (CV) predictive densities. Higher BIC and CV values indicate better in-sample fit. Both criteria are explained in the Online Appendix.

4 Empirical application

4.1 Data

As an empirical illustration of the proposed model, we focus on the number of patents awarded to firms and its relationship with research and development (R&D) expenditures. This topic has already been analyzed by various researchers (Hausman et al., 1984; Hall et al., 1986; Blundell et al., 1995, 1999, 2002; Montalvo, 1997; Crépon and Duguet, 1997; Cincera, 1997).

In particular, we use a balanced panel data set on 346 firms for the years 1975 – 1979. This data set has also been analyzed by Hall et al. (1986)[2]. Figure 1, which plots the dependent variable for all the firms over time, suggests that persistence is an issue.

In this empirical example, we take into account the three potential sources of

[2] It can also be downloaded from
http://faculty.econ.ucdavis.edu/faculty/cameron/racd2/RACD2programs.html.

persistence in the number of patents granted. The true state dependence implies the past decisions of the patent offices that issue the patent documents have a direct impact on their current patent decisions. Spurious state dependence entails that the decisions of the patent offices are entirely attributed to firm-specific unobserved components. Serial error correlation could be justified by the fact that the firms operate in an economic environment, which is subject to shocks that affects over time their R&D output measured through patents.

Our set of regressors contains the logarithm of current and up to five past years' research and development expenditures $(\ln R_0, \ln R_1, \ln R_2, \ln R_3, \ln R_4, \ln R_5)$, the logarithm of the book value of capital in 1972, which is a measure of firm size $(\ln SIZE)$, an indicator variable that equals 1 if the firm belongs to the science sector (SS), as well as time dummies $(YEAR)$. $\ln SIZE$ and SS are time-invariant covariates and therefore are excluded from Wooldridge's (2005) equation. The same holds for the year dummies.

In our empirical analysis, the proposed model (model 1) is compared against three competing panel Poisson models that have already been used by the literature on panel count data. The first competing model is a panel Poisson model with dynamics and Wooldridge (2005)'s-type latent heterogeneity (model 2), the second one is a panel Poisson with only latent heterogeneity (model 3) and the third one is a panel Poisson model with only dynamics (model 4). Models 2-4 are described in the Online Appendix, along with their MCMC algorithms that draw heavily upon the algorithm of Chib et al. (1998).

The empirical results (posterior means and standard deviations) are presented in Table 1. These results were obtained after running the MCMC algorithm for 80000 iterations with a burn-in phase of 50000 cycles. The fixed quantity c was set equal to 0.5. Alternative values, such as 0.1 or 0.8, did not affect the results.

4.2 Results

The set of the common statistically significant variables across the four models includes the $\ln y^*_{it-1}, \ln R_0, YEAR = 1978$ and $YEAR = 1979$. Our goal is to identify the potential sources of inertia in the number of patents awarded to firms.

For the Poisson models that control for dynamics (models 1,2 and 4), we observe that the estimated coefficients on $\ln y^*_{it-1}$ are positive and statistically significant; the number of patents granted in the previous period is a valid determinant of the number of patents granted in the current period. The positive sign implies that the number of patents granted in the previous period is less likely to affect downwards the number of patents granted in the current period. It is also worth noting that the coefficient on $\ln y^*_{it-1}$ is close to one in model 4 but as we move to models 2 and 1, it decreases towards zero.

Due to the nonlinear nature of the Poisson model, we also calculated the average partial effects (APEs) for y_{it-1}, which is the main covariate of interest[3]. The APEs for y_{it-1} reflect the strength of true state dependence. In the proposed model, the (statistically significant) APEs is 0.1005 with a standard deviation of 0.0586; given the number of patents in the previous period, the probability of a firm having a larger number of patents awarded in the current period increases by 10.05% . For models 2 and 4, the corresponding (statistically significant) APEs are 0.2597 (0.1239) and 0.9432 (0.0921), respectively. Standard deviations are in the parentheses. So, true state dependence is weak in model 2, weaker in model 1 and strong in model 4.

Also, there is evidence of strong dynamic dependence in the counts through the serial correlation in the idiosyncratic errors; the autoregressive parameter ρ is positive, significant and high in magnitude (0.8311). Furthermore, as can be seen from Table 2, the current counts are conditioned on the initial counts but not on the mean of explanatory variables; the coefficient h_1 is significant but the coefficients in $\mathbf{h}_2$ are not in models 1 and 2.

[3]For the calculation of the APEs, see the Online Appendix.

Across the models of Table 1 that account for unobserved heterogeneity (models 1, 2 and 3) the error variance σ_u^2 is significant. This implies that the persistence in the counts is not only the result of serially correlated errors and true state dependence but also of the firm-related unobserved heterogeneity (spurious state dependence).

Model 1, which controls for dynamics, latent heterogeneity and serially correlated errors, has the best fit to the data set, as it produces the largest BIC value (-1390.21) and the largest CV value(0.2095). Controlling only for dynamics and latent heterogeneity, model 2 delivers worse BIC and CV values, an indication that serial correlation in the idiosyncratic errors should not be ignored. Goodness of fit deteriorates even further, when we control only for latent heterogeneity (model 3) or only for dynamics (model 4), signalling the importance of accounting for both true and spurious state dependence. Hence, the most (least) preferred model is model 1 (model 4).

For robustness check, we re-estimated the proposed model (model 1) without the mean variables $\overline{\mathbf{x}}$ (model 1a), without the initial counts $\ln y_{i0}^*$ (model 1b) and with an AR(2) error structure($1_{AR(2)}$ model). The results obtained from these models are the same with those of model 1, in terms of the significance of the covariates and the sources of persistence; see Online Appendix.

5 Conclusion

In this paper we proposed a Poisson panel data model with dynamics, latent heterogeneity and serial error correlation. We also accounted for the initial conditions problem. Our Bayesian methodology was illustrated by a real data set on the number of patents awarded. We found that all three sources of persistence are present in the data set, with dynamics being weak and with serial error correlation being strong.

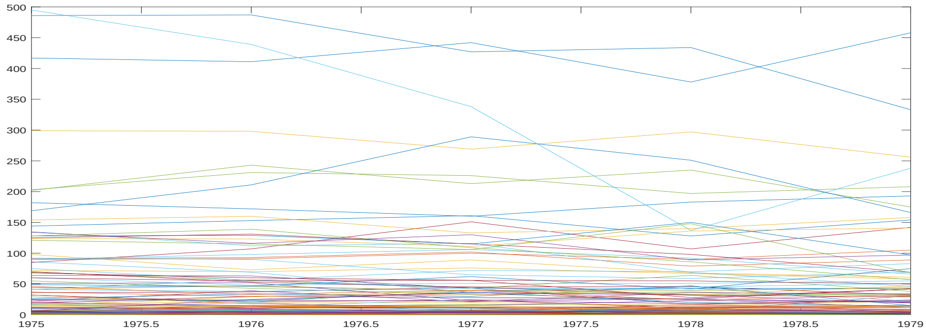

Figure 1: Empirical results. Plot of the dependent variable for all firms over time

Table 1: Empirical results for the competing Poisson models

	model 1	model 2	model 3	model 4
constant	0.1249	0.0632	-0.1350	0.0294
	(0.1072)	(0.1153)	(0.2030)	(0.0303)
$\ln y^{*}_{it-1}$	0.0936*	0.2448*		0.9311*
	(0.0325)	(0.0248)		(0.0082)
SS	-0.0173	0.0264	0.4325*	0.0312*
	(0.0689)	(0.0657)	(0.1218)	(0.0120)
$\ln SIZE$	-0.0369	-0.0012	0.2843*	0.0205*
	(0.0291)	(0.0316)	(0.0511)	(0.0059)
$\ln R_0$	0.2998*	0.3504*	0.4205*	0.2427*
	(0.0697)	(0.0637)	(0.0588)	(0.0487)
$\ln R_1$	-0.0720	-0.0777	-0.0380	-0.1659*
	(0.0706)	(0.0718)	(0.0701)	(0.0681)
$\ln R_2$	0.0396	0.0670	0.1157	-0.0514
	(0.0641)	(0.0661)	(0.0660)	(0.0646)
$\ln R_3$	0.0096	0.0090	0.0373	-0.0294
	(0.0624)	(0.0608)	(0.0597)	(0.0599)
$\ln R_4$	0.0281	0.0151	0.0142	0.0062
	(0.0579)	(0.0541)	(0.0538)	(0.0540)
$\ln R_5$	-0.0183	0.0285	0.0488	0.0337
	(0.0503)	(0.0443)	(0.0421)	(0.0361)
YEAR=1976	-0.0384	-0.041*	-0.0457*	-0.0222
	(0.0227)	(0.0177)	(0.0179)	(0.0176)
YEAR=1977	-0.0327	-0.0372*	-0.0501*	0.0059
	(0.0273)	(0.0181)	(0.0182)	(0.0177)
YEAR=1978	-0.1457*	-0.1611*	-0.1776*	-0.1129*
	(0.0294)	(0.0192)	(0.0189)	(0.0182)
YEAR=1979	-0.2002*	-0.1774*	-0.2316*	-0.0453*
	(0.0341)	(0.0213)	(0.0199)	(0.0185)
σ_u^2	0.1091*	0.1481*	0.9942*	
	(0.0386)	(0.0208)	(0.0963)	
σ_v^2	0.0355*			
	(0.0037)			
ρ	0.8311*			
	(0.0751)			
BIC	-1390.21	-1411.47	-1432.98	-1439.74
CV	0.2095	0.1748	0.1744	0.1612

*Significant based on the 95% highest posterior density interval. Standard deviations in parentheses.

Table 2: Empirical results for Wooldridge's (2005) regression

	model 1	model 2
h_1	0.7376*	0.6010*
	(0.0407)	(0.0349)
$h_{21_{(\ln R_0)}}$	-0.0799	-0.1460
	(0.3551)	(0.3254)
$h_{22_{(\ln R_1)}}$	0.0490	0.0524
	(0.5988)	(0.5646)
$h_{23_{(\ln R_2)}}$	-0.0432	-0.0557
	(0.6324)	(0.5941)
$h_{24_{(\ln R_3)}}$	0.0809	0.0168
	(0.6028)	(0.5535)
$h_{25_{(\ln R_4)}}$	-0.2022	-0.1171
	(0.5466)	(0.4891)
$h_{26_{(\ln R_5)}}$	0.1320	0.01851
	(0.2912)	(0.2621)

*Significant based on the 95% highest posterior density interval. Standard deviations in parentheses.

References

Biewen M. 2009. Measuring state dependence in individual poverty histories when there is feedback to employment status and household composition. *Journal of Applied Econometrics* **24**: 1095–1116.

Blundell R, Griffith R, Reenan JV. 1995. Dynamic count models of technological innovation. *Economic Journal* **105**: 333—-344.

Blundell R, Griffith R, Reenan JV. 1999. Market share, market value and innovation in a panel of british manufacturing firms. *Review of Economic Studies* **66**: 529—-554.

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

Cameron A, Trivedi P. 2013. *Regression Analysis of count data*. Cambridge University Press, second edition.

Chib S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**: 1313–1321.

Chib S, Greenberg E, Winkelmann R. 1998. Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics* **86**: 33 – 54.

Chib S, Jeliazkov I. 2001. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association* **96**: 270–281.

Chib S, Jeliazkov I. 2006. Inference in semiparametric dynamic models for Binary longitudinal data. *Journal of the American Statistical Association* **101**: 685–700.

Cincera M. 1997. Patents, R&D and technological spillovers at the firm level: Some evidence from econometric count models for panel data. *Journal of Applied Econometrics* **12**: 265—-280.

Crépon B, Duguet E. 1997. Estimating the innovation function from patent numbers: GMM on count data. *Journal of Applied Econometrics* **12**: 243–263.

Fotouhi A. 2005. The initial conditions problem in longitudinal Binary process: A simulation study. *Simulation Modelling Practice and Theory* **13**: 566–583.

Hall B, Griliches Z, Hausman J. 1986. Patents and R and D: Is there a lag? *International Economic Review* **27**: 265–283.

Hausman J, Hall B, Griliches Z. 1984. Econometric models for count data with an application to the patents– R and D relationship. *Econometrica* **52**: 909–938.

Heckman J. 1981. Heterogeneity and state dependence. *Studies in labor markets, University of Chicago Press* : 91–140.

Montalvo J. 1997. GMM estimation of count-panel-data models with fixed effects and predetermined instruments. *Journal of Business and Economic Statistics* **15**: 82—-89.

Schwarz G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6**: 461—-464.

Tanner M, Wong W. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**: 528–540.

Winkelmann R. 2008. *Econometric analysis of count data*. Springer-Verlag Berlin, fifth edition.

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

Zeger S, Qaqish B. 1988. Markov regression models for time series: A quasi-likelihood approach. *Biometrics* **44**: 1019–1031.

Online Appendix for: Accounting for persistence in panel count data models. An application to the number of patents awarded

Stefanos Dimitrakopoulos*

Department of Accounting, Finance and Economics, Oxford Brookes University, Oxford, OX33 1HX, UK

1 MCMC algorithm for the proposed Poisson model

If we stack the latent equation $\lambda^*_{it} = \mathbf{w}'_{it}\boldsymbol{\delta} + \varphi_i + \epsilon_{it}$ over t within i we get

$$\boldsymbol{\lambda}^*_i = \mathbf{W}_i\boldsymbol{\delta} + \mathbf{i}_T\varphi_i + \boldsymbol{\epsilon}_i, \tag{A.1}$$

where $\mathbf{W}_i = (\mathbf{w}_{i1}, ..., \mathbf{w}_{iT})'$, $\mathbf{i}_T$ is a $T \times 1$ vector of ones and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{iT})'$ follows a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\sigma^2_v\Omega_i$, which is symmetric and positive definite with

$$\Omega_i = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{pmatrix}.$$

The MCMC algorithm works as follows:

- We sample $\sigma_v^{-2}, \boldsymbol{\delta}|\{\boldsymbol{\lambda}^*_i\}, \{\Omega_i\}, \{\varphi_i\}, e_1, f_1$ in one block by sampling

(a) $\sigma_v^{-2}|\{\boldsymbol{\lambda}^*_i\}, \{\Omega_i\}, \{\varphi_i\}, e_1, f_1 \sim \mathcal{G}(\frac{\overline{e_1}}{2}, \frac{\overline{f_1}}{2})$, where $\overline{e_1} = e_1 + NT - k - 1$, $\overline{f_1} = f_1 + (\widetilde{\boldsymbol{\lambda}}^* - \mathbf{W}\widehat{\boldsymbol{\delta}})'\Omega^{-1}(\widetilde{\boldsymbol{\lambda}}^* - \mathbf{W}\widehat{\boldsymbol{\delta}})$, $\widetilde{\boldsymbol{\lambda}}^*$ contains the elements $\widetilde{\lambda_{it}}^* = \lambda^*_{it} - \varphi_i$, $i = 1, ...N$, $t = 1, ..., T$ that have been stacked over i and t, $\mathbf{W} = (\mathbf{W}'_1, ..., \mathbf{W}'_N)'$, $\widehat{\boldsymbol{\delta}}$ is the OLS estimator of $\boldsymbol{\delta}$ given by $\widehat{\boldsymbol{\delta}} = (\mathbf{W}'\Omega^{-1}\mathbf{W})^{-1}\mathbf{W}'\Omega^{-1}\widetilde{\boldsymbol{\lambda}}^*$ and Ω is a block diagonal matrix,

$$\Omega = \begin{pmatrix} \Omega_1 & & & \\ & \Omega_2 & & \\ & & \ddots & \\ & & & \Omega_N \end{pmatrix}.$$

*Correspondence to: Stefanos Dimitrakopoulos, E-mail: sdimitrakopoulos@brookes.ac.uk.

(b) $\boldsymbol{\delta}|\{\boldsymbol{\lambda}_i^*\}, \sigma_v^2, \{\Omega_i\}, \{\varphi_i\} \sim N\left(\widehat{\boldsymbol{\delta}}, (\frac{1}{\sigma_v^2}\mathbf{W}'\Omega^{-1}\mathbf{W})^{-1}\right)$.

• We sample $\varphi_i|\boldsymbol{\lambda}_i^*, \mathbf{h}, \boldsymbol{\delta}, \Omega_i, \sigma_v^2, \sigma_u^2 \sim N(d_0, D_0)$, $i = 1, ..., N$, where $D_0 = \left(\frac{1}{\sigma_u^2} + \sigma_v^{-2}\mathbf{i}_T'\Omega_i^{-1}\mathbf{i}_T\right)^{-1}$ and $d_0 = D_0\left(\frac{\mathbf{k}_i'\mathbf{h}}{\sigma_u^2} + \sigma_v^{-2}\mathbf{i}_T'\Omega_i^{-1}(\boldsymbol{\lambda}_i^* - \mathbf{W}_i\boldsymbol{\delta})\right)$ with $\mathbf{k}_i' = (\ln y_{i0}^*, \overline{\mathbf{x}}_i')$.

• We sample $\mathbf{h}|\{\varphi_i\}, \widetilde{\mathbf{H}}, \widetilde{\mathbf{h}}, \sigma_u^2 \sim \mathbf{N}(d_\mathbf{h}, D_\mathbf{h})$, where $d_\mathbf{h} = D_\mathbf{h}\left(\widetilde{\mathbf{H}}^{-1}\widetilde{\mathbf{h}} + \frac{\mathbf{k}'\boldsymbol{\varphi}}{\sigma_u^2}\right)$ and $D_\mathbf{h} = \left(\widetilde{\mathbf{H}}^{-1} + \frac{\mathbf{k}'\mathbf{k}}{\sigma_u^2}\right)^{-1}$, where $\mathbf{k}$ is the matrix that consists of all $\mathbf{k}_i$ and $\boldsymbol{\varphi}$ is the vector of all φ_i.

• We sample $\boldsymbol{\lambda}_i^*$, $i = 1, ..., N$, from the posterior distribution of $\boldsymbol{\lambda}_i^*|\boldsymbol{\delta}, \sigma_v^2, \Omega_i, \varphi_i, \boldsymbol{y}_i$, which is proportional to $N(\boldsymbol{\lambda}_i^*|\mathbf{W}_i\boldsymbol{\delta} + \mathbf{i}_T\varphi_i, \sigma_v^2\Omega_i)Poisson(\boldsymbol{y}_i|exp(\boldsymbol{\lambda}_i^*))$, where $\boldsymbol{y}_i = \{y_{it}\}_{t\geq 1}$. This density does not have closed form. Therefore we use an independence Metropolis-Hastings (MH) algorithm [see, for example, Chib and Greenberg (1995)] to update each $\boldsymbol{\lambda}_i^*$. In this paper, we orthogonalize the correlated errors so that the elements within each $\boldsymbol{\lambda}_i^*$ can be sampled independently of one another (Chib and Jeliazkov, 2006).

In particular, we decompose the covariance matrix Ω_i as $\Omega_i = \xi\boldsymbol{I}_T + \widetilde{R}_i$, where $\boldsymbol{I}_T$ is the $T \times T$ identity matrix, ξ is an arbitrary constant that satisfies the constraint $\bar{\xi} > \xi > 0$, where $\bar{\xi}$ is the minimum eigenvalue of Ω_i and $\widetilde{R}_i$ is a symmetric positive definite matrix. The algorithm becomes stable by setting $\xi = \bar{\xi}/2$ [see, also, Chib and Jeliazkov (2006)]. $\widetilde{R}_i$ can be further decomposed into $\widetilde{R}_i = C_i'C_i$ (Cholesky decomposition). Hence, $\Omega_i = C_i'C_i + \xi\boldsymbol{I}_T$.

Using this decomposition, the latent regression for $\boldsymbol{\lambda}_i^*$, $i = 1, ..., N$ can be written as

$$\boldsymbol{\lambda}_i^* = \mathbf{W}_i\boldsymbol{\delta} + \mathbf{i}_T\varphi_i + C_i'\boldsymbol{\eta}_i + \boldsymbol{e_i}, \qquad (A.2)$$

where $\boldsymbol{\eta}_i \sim N(0, \sigma_v^2\boldsymbol{I}_T)$ and $\boldsymbol{e_i} \sim N(0, \xi\sigma_v^2\boldsymbol{I}_T)$. Using (A.2), the (intractable) full conditional distribution of each λ_{it}^*, $i = 1, ..., N$, $t = 1, ..., T$ is given by

$$p(\lambda_{it}^*|\boldsymbol{\delta}, \sigma_v^2, \rho, \varphi_i, y_{it}) \propto exp\left(-exp(\lambda_{it}^*) + \lambda_{it}^* y_{it} - exp\left(\frac{1}{2\xi\sigma_v^2}(\lambda_{it}^* - \mathbf{w}_{it}'\boldsymbol{\delta} - \varphi_i - q_{it})^2\right)\right),$$

where q_{it} is the $t - th$ element of $q_i = C_i'\boldsymbol{\eta}_i$. Let $St(\lambda_{it}^*|\hat{\lambda}_{it}^*, c_1V_{\lambda_{it}^*}, v_1)$ denote a Student-t distribution, where $\hat{\lambda}_{it}^*$ is defined as the modal value of the $logp(\lambda_{it}^*|\boldsymbol{\delta}, \sigma_v^2, \rho, \varphi_i, y_{it})$, $V_{\lambda_{it}^*} = (-H_{\lambda_{it}^*})^{-1}$ is defined as the inverse of the negative Hessian of the of the $logp(\lambda_{it}^*|\boldsymbol{\delta}, \sigma_v^2, \rho, \varphi_i, y_{it})$ evaluated at $\hat{\lambda}_{it}^*$, v_1 is the degrees of freedom and c_1 is a positive-valued scale parameter. Both v_1 and c_1 are essentially tuning parameters which are determined prior to the main MCMC loop. To obtain the modal value we use a few Newton-Raphson rounds implemented via the gradient

$$\hat{\lambda}_{it}^* = -exp(\lambda_{it}^*) + y_{it} - \frac{1}{\xi\sigma_v^2}(\lambda_{it}^* - \mathbf{w}_{it}'\boldsymbol{\delta} - \varphi_i - q_{it}),$$

and the Hessian

$$H_{\lambda_{it}^*} = -exp(\lambda_{it}^*) - \frac{1}{\xi\sigma_v^2}.$$

Then, sample a proposal value $\lambda_{it}^{*(p)}$ from the density $St(\lambda_{it}^{*}|\hat{\lambda}_{it}^{*}, c_1 V_{\lambda_{it}^{*}}, v_1)$ and move to $\lambda_{it}^{*(p)}$ given the current point $\lambda_{it}^{*(c)}$ with probability of move

$$min\left(\frac{p(\lambda_{it}^{*(p)}|\boldsymbol{\delta},\sigma_v^2,\rho,\varphi_i,y_{it})St(\lambda_{it}^{*}|\hat{\lambda}_{it}^{*(c)},c_1V_{\lambda_{it}^{*}},v_1)}{p(\lambda_{it}^{*(c)}|\boldsymbol{\delta},\sigma_v^2,\rho,\varphi_i,y_{it})St(\lambda_{it}^{*(p)}|\hat{\lambda}_{it}^{*},c_1V_{\lambda_{it}^{*}},v_1)}, 1\right).$$

- To update $q_i = C_i'\boldsymbol{\eta}_i$, $i = 1, ..., N$ in each iteration we sample $\boldsymbol{\eta}_i$ from $\boldsymbol{\eta}_i|\boldsymbol{\lambda}_i^*, \boldsymbol{\delta}, \varphi_i, \sigma_v^2 \sim N(p_1, P_1)$, where $p_1 = P_1\left(\frac{C_i\left(\boldsymbol{y}_i^* - \mathbf{W}_i\boldsymbol{\delta} - \mathbf{i}_T\varphi_i\right)}{\xi\sigma_v^2}\right)$ and $P_1 = \left(\frac{I_T}{\sigma_v^2} + \frac{C_iC_i'}{\xi\sigma_v^2}\right)^{-1}$.

- We sample $\rho|\boldsymbol{\epsilon}, \sigma_v^2, \rho_0, \sigma_\rho^2 \propto \Psi(\rho)\times N(d_2, D_2)I_{(-1,1)}(\rho)$, where $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1', ..., \boldsymbol{\epsilon}_N')'$, $\epsilon_{it} = \lambda_{it}^* - \mathbf{w}_{it}'\boldsymbol{\delta} - \varphi_i$, $\Psi(\rho) = \sqrt{(1-\rho^2)^N}\times\exp\left(-\frac{(1-\rho^2)}{2\sigma_v^2}\sum_{i=1}^N \epsilon_{i1}^2\right)$, $d_2 = D_2\left(\frac{\rho_0}{\sigma_\rho^2} + \sigma_v^{-2}\sum_{i=1}^N\sum_{t=2}^T \epsilon_{it}\epsilon_{it-1}\right)$ and $D_2 = \left(\frac{1}{\sigma_\rho^2} + \sigma_v^{-2}\sum_{i=1}^N\sum_{t=2}^T \epsilon_{it-1}^2\right)^{-1}$. We use an independence Metropolis-Hastings algorithm in order to simulate ρ. A candidate value ρ' is generated from the density $N(d_2, D_2)I_{(-1,1)}(\rho)$ and is accepted as the next value in the chain with probability $\min(\Psi(\rho')/\Psi(\rho), 1)$; otherwise, the current value ρ is taken to be the next value in the sample.

- We obtain deterministically the errors u_i from $u_i = \varphi_i - h_1 \ln y_{i0} - \overline{\mathbf{x}}_i'\mathbf{h}_2$,$i = 1, ..., N$.

- We update σ_u^2 from

$$\sigma_u^2|\{u_i\}, e_0, f_0 \sim \mathcal{IG}(\frac{\overline{e_0}}{2}, \frac{\overline{f_0}}{2}),$$

where

$$\overline{e_0} = e_0 + N, \overline{f_0} = f_0 + \sum_{i=1}^N u_i^2.$$

2 A simulation study for the proposed Poisson model

In this section we evaluate the efficiency of the MCMC scheme for the proposed Poisson model, utilizing a simulated data set.

We set $N = 800$ and $T = 5$. For the parameters of interest, we assume the following true values:

$$\beta_1 = 0(intercept), \beta_2 = -0.1, \beta_3 = 0.2, \beta_4 = 0.3, \gamma = 0.1, \rho = 0.8,$$

$$\sigma_v^2 = \sigma_u^2 = 0.1, h_1 = 0.3, h_{21} = 0.1, h_{22} = 0.5, h_{23} = 0.8.$$

Also, x_{2it}, x_{3it} and x_{4it} are generated independently from 0.5N(0,1), 0.2+ N(0,1) and 0.6N(0,1), respectively. We also use the following priors:

$$\sigma_v^{-2} \sim \mathcal{G}(4.2/2, 0.5/2), \rho \sim N(0, 10)I_{(-1,1)}(\rho), \boldsymbol{\delta} \sim \mathbf{N}(0, 100\boldsymbol{I}),$$
$$\mathbf{h} \sim \mathbf{N}(\mathbf{0}, 100 \times \boldsymbol{I}_4), \sigma_u^2 \sim I\mathcal{G}(4.2/2, 0.5/2).$$

Notice that the intercept is excluded from Wooldridge's (2005) regression. We run the algorithm 20000 times after throwing away the first 25000 iterations. The posterior means and standard deviations of the parameters in question are presented in Table 1. To monitor

Table 1: Simulated data. Estimation results

True values	Mean	IF	CD
$\beta_1 = 0$	0.0525	25.3605	0.7729
	(0.0540)		
$\beta_2 = -0.1$	-0.1127	26.5100	-1.7495
	(0.0233)		
$\beta_3 = 0.2$	0.1902	43.8098	-2.5536
	(0.0123)		
$\beta_4 = 0.3$	0.3112	49.5420	2.0279
	(0.0237)		
$\gamma = 0.1$	0.1036	47.3089	-1.6519
	(0.0165)		
$\sigma_v^2 = 0.1$	0.1058	55.3660	-0.2752
	(0.0105)		
$\rho = 0.8$	0.7724	116.6448	0.9027
	(0.0676)		
$\sigma_u^2 = 0.1$	0.1843	118.0805	-1.5325
	(0.0943)		
$h_1 = 0.3$	0.2652	80.66	0.5460
	(0.0394)		
$h_{21} = 0.1$	0.0374	35.065	-1.9768
	(0.1450)		
$h_{22} = 0.5$	0.4485	33.708	0.8700
	(0.0586)		
$h_{23} = 0.8$	0.8182	64.41	2.6057
	(0.1233)		

Standard deviations in parentheses.

convergence and mixing, we also report the CD statistics of Geweke (1992) and the inefficiency factor (IF); see, for example, Chib (2001). As can be seen from Table 1, the estimated parameters are close to their true values.

3 Model comparison

The marginal likelihood is used to measure the fit of the model to the data in hand. For model $\mathcal{M}$ with likelihood $p(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\theta})$, where $\boldsymbol{y}$ is the data vector, and prior $p(\boldsymbol{\theta}|\mathcal{M})$, the marginal likelihood (ML) is defined as

$$p(\boldsymbol{y}|\mathcal{M}) = \int p(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}, \qquad (A.3)$$

where $\boldsymbol{\theta} = (\boldsymbol{\delta}, \mathbf{h}, \rho, \sigma_u^2, \sigma_v^2)$.

Expression (A.3), though, is intractable. Using the Bayesian Information Criterion

(Schwarz, 1978), the marginal likelihood for model $\mathcal{M}$ can be approximated by

$$p(\boldsymbol{y}|\mathcal{M}) \approx n^{\frac{-d_k}{2}} L(\hat{\boldsymbol{\theta}}|\mathcal{M}, \boldsymbol{y}), \tag{A.4}$$

where $L(\boldsymbol{\theta}|\mathcal{M}, \boldsymbol{y})$ is the likelihood function and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$. By taking the logarithm of both sides we have

$$\log p(\boldsymbol{y}|\mathcal{M}) \approx \log L(\hat{\boldsymbol{\theta}}|\mathcal{M}, \boldsymbol{y}) - \log(n)\frac{d_k}{2} = BIC, \tag{A.5}$$

where d_k is the dimension of $\boldsymbol{\theta}$ and n is the total sample size.

Based on the second-order Taylor series expansion, we use the MCMC draws to approximate the maximum log-likelihood function, $\log L(\hat{\boldsymbol{\theta}}|\mathcal{M}, \boldsymbol{y})$, from the posterior log-likelihood score (LLS), $\log L(\tilde{\boldsymbol{\theta}}|\mathcal{M}, \boldsymbol{y})$, where $\tilde{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$. The LLS is calculated as the posterior expectation of the log-likelihood function, $LLS \approx E(\log L(\tilde{\boldsymbol{\theta}}|\mathcal{M}, \boldsymbol{y}))$.

An alternative model comparison criterion is based on cross-validation predictive densities. In particular, we apply the leave-one-out cross validation (CV) method that requires the calculation of the conditional predictive ordinate (CPO),

$$CPO_{it} = f(y_{it}|y_{-it}) = \int f(y_{it}|\boldsymbol{\Theta})f(\boldsymbol{\Theta}|y_{-it}) = \mathbf{E}_{\boldsymbol{\Theta}|y_{-it}}[f(y_{it}|\boldsymbol{\Theta})],\ i = 1, ..., N, t = 1, ..., T, \tag{A.6}$$

where $y_{-it} = \mathbf{y} \setminus \{y_{it}\}$ and $\boldsymbol{\Theta} = (\boldsymbol{\delta}, \{\varphi_i\}, \{q_{it}\}, \xi, \sigma_v^2)$. Gelfand and Dey (1994) and Gelfand (1996) proposed a Monte Carlo integration of CPO. More specifically,

$$C\hat{P}O_{it} = \hat{f}(y_{it}|y_{-it}) = \left(\frac{1}{L}\sum_{l=1}^{L}\left(f(y_{it}|y_{-it}, \boldsymbol{\Theta}^{(l)})\right)^{-1}\right)^{-1}, \tag{A.7}$$

where L is the number of iterations after the burn-in period. Then, for each model we calculate the average of the estimated CPO values, $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{f}(y_{it}|y_{-it})$. Higher values of this average imply better "goodness of fit" of a model.

4 Average marginal effects

For the proposed model, the marginal effect (ME) for the it-th component with respect to the k-th continuous regressor is

$$ME_{kit} = \frac{\partial E(y_{it}|\mathbf{w}_{it}, \boldsymbol{\delta}, \varphi_i, q_{it})}{\partial x_{k,it}} = \beta_k exp(\mathbf{w}_{it}'\boldsymbol{\delta} + \varphi_i + q_{it}). \tag{A.8}$$

By integrating out all the unknowns (including the random effects), the posterior distribution of ME_{kit} is

$$\pi(ME_{kit}|data) = \int \pi(ME_{kit}|\boldsymbol{\delta}, \varphi_i, q_{it}, data)d\pi(\boldsymbol{\delta}, \varphi_i, q_{it}|data). \tag{A.9}$$

Using the composition method, we can produce a sample of ME_{kit} values, using the posterior draws of $\boldsymbol{\delta}, \varphi_i, q_{it}$ from the MCMC algorithm. Chib and Hamilton (2002) also used this method to calculate average treatment effects. Given a posterior sample of ME_{kit} values obtained from $\pi(ME_{kit}|data)$, which we denote by $\{ME_{kit}^{(l)}\}$, the average marginal effect (AME) can be defined as

$$AME_k = \frac{\sum_{l=1}^{L}\sum_{i=1}^{N}\sum_{t=1}^{T} ME_{kit}^{(l)}}{L \times N \times T}, \tag{A.10}$$

where $ME_{kit}^{(l)} = \beta_k^{(l)} exp(\mathbf{w}_{it}'\boldsymbol{\delta}^{(l)} + \varphi_i^{(l)} + q_{it}^{(l)})$ and L is the total number of iterations after the burn-in period.

If $x_{k,it}$ is binary, the marginal effect is

$$\Delta_j(x_{k,it}) = exp((\mathbf{w}_{it}{'}\boldsymbol{\delta} - x_{k,it}\beta_k) + \beta_k + \varphi_i + q_{it}) - exp((\mathbf{w}_{it}{'}\boldsymbol{\delta} - x_{k,it}\beta_k) + \varphi_i + q_{it}). \tag{A.11}$$

5 Empirical analysis

5.1 Additional empirical models

Model 2: Dynamic panel Poisson model with (Wooldridge, 2005)'-type latent heterogeneity

$$\begin{aligned}
&y_{it}|\lambda_{it} \sim Poisson(\lambda_{it}),\\
&\lambda_{it} = exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma \ln y_{it-1}^{*} + \varphi_i),\\
&\varphi_i = h_1 \ln y_{i0}^{*} + \overline{\mathbf{x}}_i'\mathbf{h}_2 + u_i,\\
&u_i \sim N(0, \sigma_u^2).
\end{aligned}$$

Model 3: panel Poisson model with latent heterogeneity

$$\begin{aligned}
&y_{it}|\lambda_{it} \sim Poisson(\lambda_{it}),\\
&\lambda_{it} = exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \varphi_i),\\
&\varphi_i \sim N(0, \sigma_u^2).
\end{aligned}$$

Model 4: Dynamic panel Poisson model

$$\begin{aligned}
&y_{it}|\lambda_{it} \sim Poisson(\lambda_{it}),\\
&\lambda_{it} = exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma \ln y_{it-1}^{*}).
\end{aligned}$$

5.1.1 MCMC algorithm for Model 2

Drawing upon the algorithm of Chib et al. (1998), we present the MCMC algorithm for model 2 as the algorithms for models 3 and 4 are straightforward. It is also worth noting that the MCMC algorithm that we used for the static model (model 3) is exactly the same as that in the Chib et al. (1998) paper.

The conditional distributions of $\mathbf{h}$ and σ_u^2 are the same as those in the proposed model.

The posterior densities of φ_i and $\boldsymbol{\delta} = (\boldsymbol{\beta}', \gamma)'$ are intractable and therefore we use the independence Metropolis-Hastings algorithm to make draws.

• The posterior distribution of φ_i, $i = 1, ...N$ is given by

$$p(\varphi_i|\{y_{it}\}_{t\geq 1}, \mathbf{h}, \boldsymbol{\delta}, \sigma_u^2) \propto exp\left(-\frac{1}{2\sigma_u^2}(\varphi_i - \mathbf{k}_i'\mathbf{h})^2\right) \times \prod_{t=1}^{T} \frac{exp[-exp(\mathbf{w}_{it}'\boldsymbol{\delta}+\varphi_i)][exp(\mathbf{w}_{it}'\boldsymbol{\delta}+\varphi_i)]^{y_{it}}}{y_{it}!}.$$

A proposed draw $\varphi_i^{(p)}$ is generated from the Student-t distribution $St(\varphi_i^{(p)}|\hat{\varphi}_i, c_2 V_{\varphi_i}, v_2)$, where $\hat{\varphi}_i = argmax\ logp(\varphi_i|\{y_{it}\}_{t\geq 1}, \mathbf{h}, \boldsymbol{\delta}, \sigma_u^2)$ is the modal value of the logarithm of the posterior distribution of φ_i, $V_{\varphi_i} = (-H_{\varphi_i})^{-1}$ is the inverse of the negative Hessian of $logp(\varphi_i|\{y_{it}\}_{t\geq 1}, \mathbf{h}, \boldsymbol{\delta}, \sigma_u^2)$ evaluated at $\hat{\varphi}_i$, v_2 is the degrees of freedom and $c_2 > 0$ is a constant. To obtain the modal value we use the Newton-Raphson method that requires the calculation of the gradient

$$g_{\varphi_i} = -(\varphi_i - \mathbf{k}_i'\mathbf{h})/\sigma_u^2 + \sum_{t=1}^{T}[y_{it} - exp(\mathbf{w}_{it}'\boldsymbol{\delta} + \varphi_i)],$$

and the Hessian

$$H_{\varphi_i} = -\sigma_u^{-2} - \sum_{t=1}^{T} exp(\mathbf{w}_{it}'\boldsymbol{\delta} + \varphi_i).$$

Given the current value $\varphi_i^{(c)}$, we move to the proposed point $\varphi_i^{(p)}$ with probability

$$a_p(\varphi_i^{(c)}, \varphi_i^{(p)}) = min\left(\frac{p(\varphi_i^{(p)}|\{y_{it}\}_{t\geq 1}, \mathbf{h}, \boldsymbol{\delta}, \sigma_u^2) St(\varphi_i^{(c)}|\hat{\varphi}_i, c_2 V_{\varphi_i}, v_2)}{p(\varphi_i^{(c)}|\{y_{it}\}_{t\geq 1}, \mathbf{h}, \boldsymbol{\delta}, \sigma_u^2) St(\varphi_i^{(p)}|\hat{\varphi}_i, c_2 V_{\varphi_i}, v_2)}, 1\right).$$

• The target density of $\boldsymbol{\delta}$ is also intractable,

$$p(\boldsymbol{\delta}|\{y_{it}\}_{i\geq 1, t\geq 1}, \{\varphi_i\}) \propto \prod_{i=1}^{N}\prod_{t=1}^{T} \frac{exp[-exp(\mathbf{w}_{it}'\boldsymbol{\delta}+\varphi_i)][exp(\mathbf{w}_{it}'\boldsymbol{\delta}+\varphi_i)]^{y_{it}}}{y_{it}!}.$$

To generate $\boldsymbol{\delta}$ from its full conditional we use a multivariate Student-t distribution $MVt(\boldsymbol{\delta}|\widehat{\boldsymbol{\delta}}, c_3\widehat{\Sigma}_{\boldsymbol{\delta}}, v_3)$, where $\widehat{\boldsymbol{\delta}} = argmax\ logp(\boldsymbol{\delta}|\{y_{it}\}_{i\geq 1, t\geq 1}, \{\varphi_i\})$ is the mode of the logarithm of the right hand side of the above conditional distribution and $\widehat{\Sigma}_{\boldsymbol{\delta}} = [-H_{\boldsymbol{\delta}}]^{-1}$ is the negative inverse of the Hessian matrix of $p(\boldsymbol{\delta}|\{y_{it}\}_{i\geq 1, t\geq 1}, \{\varphi_i\})$ at the mode $\widehat{\boldsymbol{\delta}}$. The degrees of freedom v_3 and the scaling factor c_3 are, as before, adjustable parameters. The maximizer $\widehat{\boldsymbol{\delta}}$ is obtained by using the Newton-Raphson procedure with gradient vector

$$g_{\boldsymbol{\delta}} = \sum_{i=1}^{N}\sum_{t=1}^{T}[y_{it} - exp(\mathbf{w}_{it}'\boldsymbol{\delta} + \varphi_i)]\mathbf{w}_{it},$$

and Hessian matrix

$$H_{\boldsymbol{\delta}} = -\sum_{i=1}^{N}\sum_{t=1}^{T}[exp(\mathbf{w}_{it}'\boldsymbol{\delta} + \varphi_i)]\mathbf{w}_{it}\mathbf{w}_{it}'.$$

The algorithm to generate $\boldsymbol{\delta}$ works as follows:
1) Let $\boldsymbol{\delta}^{(c)}$ be the current value.
2) Generate a proposed value $\boldsymbol{\delta}^{(p)}$ from $MVt(\boldsymbol{\delta}|\widehat{\boldsymbol{\delta}}, c_3\widehat{\Sigma}_{\boldsymbol{\delta}}, v_3)$.

3) A move from $\boldsymbol{\delta}^{(c)}$ to $\boldsymbol{\delta}^{(p)}$ is made with probability

$$\min(\frac{p(\boldsymbol{\delta}^{(p)}|\{y_{it}\}_{i\geq 1,t\geq 1},\{\varphi_i\})\ \ MVt(\boldsymbol{\delta}^{(c)}|\widehat{\boldsymbol{\delta}},c_3\widehat{\Sigma}_{\boldsymbol{\delta}},v_3)}{p(\boldsymbol{\delta}^{(c)}|\{y_{it}\}_{i\geq 1,t\geq 1},\{\varphi_i\})\ \ MVt(\boldsymbol{\delta}^{(p)}|\widehat{\boldsymbol{\delta}},c_3\widehat{\Sigma}_{\boldsymbol{\delta}},v_3)},1).$$

5.1.2 Additional empirical results

Table 2: Empirical results for variants of model 1

	model 1a	model 1b	model $1_{AR(2)}$
constant	0.1344	0.4952*	0.1085
	(0.1039)	(0.1612)	(0.1044)
$\ln y^*_{it-1}$	0.1060*	0.2140*	0.1237*
	(0.0342)	(0.0807)	(0.0464)
SS	-0.0202	0.0627	-0.0112
	(0.0657)	(0.1006)	(0.0658)
$\ln SIZE$	-0.0362	0.0126	-0.0307
	(0.0271)	(0.0412)	(0.0284)
$\ln R_0$	0.2757*	0.3022*	0.3201*
	(0.0637)	(0.0732)	(0.0707)
$\ln R_1$	-0.0996	-0.0948	-0.0836
	(0.0677)	(0.0712)	(0.0695)
$\ln R_2$	0.0275	0.0405	0.0526
	(0.0627)	(0.0661)	(0.0656)
$\ln R_3$	0.0040	0.00347	0.0139
	(0.0597)	(0.0643)	(0.0631)
$\ln R_4$	0.0250	0.0256	0.0378
	(0.0599)	(0.0593)	(0.0590)
$\ln R_5$	-0.0123	-0.0033	-0.0094
	(0.0472)	(0.0493)	(0.0502)
YEAR=1976	-0.0366	-0.0368	-0.0386
	(0.0232)	(0.0242)	(0.0233)
YEAR=1977	-0.0298	-0.0305	-0.0351
	(0.0275)	(0.0272)	(0.0270)
YEAR=1978	-0.1409*	-0.1429*	-0.1490*
	(0.0302)	(0.0298)	(0.030)
YEAR=1979	-0.1888*	-0.1853 *	-0.202*
	(0.0335)	(0.0342)	(0.0343)
σ^2_u	0.1053*	0.1468*	0.1051*
	(0.0349)	(0.0421)	(0.0386)
σ^2_v	0.0353*	0.0361*	0.0367*
	(0.0036)	(0.0038)	(0.0042)
ρ	0.8213*	0.8696*	0.7165*
	(0.0793)	(0.1292)	(0.1407)
ρ_2			0.1192
			(0.1321)
BIC	-1402.55	-1398.73	1410.20
CV	0.2071	0.1995	0.17463

*Significant based on the 95% highest posterior density interval. Standard deviations in parentheses. The APEs for y_{it-1} is 0.1141 with a standard deviation of 0.0647 in model 1a and 0.2276 with a standard deviation of 0.1215 in model 1b. For the $1_{AR(2)}$ model, the APE is 0.1320 with a standard deviation of 0.0733.

Table 3: Empirical results for Wooldridge's (2005) regression

	model 1a	model 1b	model 1$_{AR(2)}$
h_1	0.7236*		0.7125*
	(0.0406)		(0.0499)
$h21_{(\ln R_0)}$		0.0888	-0.1071
		(0.4336)	(0.3545)
$h22_{(\ln R_1)}$		-0.0078	0.0418
		(0.6965)	(0.6047)
$h23_{(\ln R_2)}$		-0.3338	0.0142
		(0.7116)	(0.6215)
$h24_{(\ln R_3)}$		0.3100	-0.0057
		(0.6862)	(0.5859)
$h25_{(\ln R_4)}$		0.3527	-0.1982
		(0.6329)	(0.5216)
$h26_{(\ln R_5)}$		0.0566	0.1350
		(0.3781)	(0.2882)

*Significant based on the 95% highest posterior density interval. Standard deviations in parentheses.

5.2 A semiparametric extension of the proposed model

For robustness check, we also considered a semiparametric modification of the proposed model. In particular, Wooldridge (2005) acknowledges that a misspecified distribution for the latent heterogeneity generally results in inconsistent parameter estimates. Therefore, we decided to let this distribution be unspecified, by imposing a nonparametric structure on it, the Dirichlet Process (DP) prior (Ferguson, 1973).

This prior has been widely used in Bayesian nonparametric modelling and it is a powerful tool for modelling random unknown distributions. For a detailed description of the Dirichlet process prior the interested reader is referred to Navarro et al. (2006) and Ghosal (2010) .

It is worth noting that semiparametric Bayesian Poisson regression models based on DP priors have been considered by Jochmann and Len-Gonzlez (2004) and Zheng (2008).

In our analysis, we assume that u_i follows the Dirichlet process mixture (DPM) model, which is defined as

$$\begin{aligned}
&u_i|\mu_i, \sigma_i^2 \overset{iid}{\sim} N(\mu_i, \sigma_i^2),\\
&\binom{\mu_i}{\sigma_i^2}|G \overset{iid}{\sim} G,\\
&G|a, G_0 \overset{iid}{\sim} DP(a, G_0),\\
&G_0(\mu_i, \sigma_i^2) \equiv N(\mu_i; \mu_0, \tau_0\sigma_i^2)\mathcal{IG}(\sigma_i^2; \tfrac{e_0}{2}, \tfrac{f_0}{2}),\\
&a \overset{iid}{\sim} \mathcal{G}(\underline{c}, \underline{d}).
\end{aligned}$$

Conditional on the mean μ_i and variance σ_i^2, the u_i are independent and normally distributed. The parameters μ_i and σ_i^2 are generated from an unknown distribution G on which the Dirichlet process (DP) prior is imposed. The DP prior is defined by the prior baseline distribution G_0, which is a conjugate normal-inverse gamma distribution, and a nonnegative concentration parameter a that follows a gamma prior.

So, our full model specification is

$$\begin{aligned}
&y_{it}|\lambda_{it} \sim Poisson(\lambda_{it}),\\
&\lambda_{it} = exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma \ln y_{it-1}^* + \varphi_i + \epsilon_{it}),\\
&\epsilon_{it} = \rho\epsilon_{it-1} + v_{it}, \quad -1 < \rho < 1, \quad v_{it} \overset{iid}{\sim} N(0, \sigma_v^2),\\
&\varphi_i = h_1 \ln y_{i0}^* + \overline{\mathbf{x}}_i'\mathbf{h}_2 + u_i,
\end{aligned}$$

where u_i follows the DPM model that is given above. Note that now $\mathbf{x}_{it}$ does not contain an intercept.

5.2.1 MCMC algorithm for the semiparametric model

Our MCMC scheme contains two parts. In part I, we update in each iteration the parameters $(\{\lambda_{it}^*\}_{i\geq 1, t\geq 1}, \boldsymbol{\delta}, \{\varphi_i\}, \mathbf{h}, \sigma_v^2, \rho)$ and recover the errors $\{u_i\}$ deterministically, using the auxiliary regression of Wooldridge (2005). In part II, we update the Dirichlet process parameters $\vartheta_i = (\mu_i, \sigma_i^2), i = 1, ..., N$, and a.

Part I

We update the parameters $(\{\lambda_{it}^*\}_{i\geq 1,t\geq 1}, \boldsymbol{\delta}, \sigma_v^2, \rho, \{u_i\})$ in the same way that we did for the parametric proposed model. In addition, we update φ_i and $\mathbf{h}$ as follows:

- We sample $\varphi_i|\boldsymbol{\lambda}_i^*, \mathbf{h}, \boldsymbol{\delta}, \Omega_i, \sigma_v^2, \vartheta_i \sim N(d_0, D_0)$, $i = 1, ..., N$, where $D_0 = \left(\frac{1}{\sigma_i^2} + \sigma_v^{-2}\mathbf{i}_T'\Omega_i^{-1}\mathbf{i}_T\right)^{-1}$ and $d_0 = D_0\left(\frac{\mathbf{k}_i'\mathbf{h}+\mu_i}{\sigma_i^2} + \sigma_v^{-2}\mathbf{i}_T'\Omega_i^{-1}(\boldsymbol{\lambda}_i^* - \mathbf{W}_i\boldsymbol{\delta})\right)$ with $\mathbf{k}_i' = (\ln y_{i0}^*, \overline{\mathbf{x}}_i')$.

- We sample $\mathbf{h}|\{\varphi_i\}, \widetilde{\mathbf{H}}, \widetilde{\mathbf{h}}, \{\vartheta_i\} \sim \mathbf{N}(d_\mathbf{h}, D_\mathbf{h})$, where $d_\mathbf{h} = D_\mathbf{h}\left(\widetilde{\mathbf{H}}^{-1}\widetilde{\mathbf{h}} + \frac{\mathbf{k}'(\boldsymbol{\varphi}-\boldsymbol{\mu})}{\sigma^2}\right)$ and $D_\mathbf{h} = \left(\widetilde{\mathbf{H}}^{-1} + \frac{\mathbf{k}'\mathbf{k}}{\sigma^2}\right)^{-1}$, where $\mathbf{k}$ is the matrix that consists of all $\mathbf{k}_i$,$\boldsymbol{\varphi}$ is the vector of all φ_i, σ^2 is the vector of all σ_i^2 and $\boldsymbol{\mu}$ is the vector of all μ_i.

Part II

To improve efficiency of sampling from $\boldsymbol{\theta}|\{u_i\}, \mu_0, \tau_0, e_0, f_0$, we sample from the equivalent distribution $\boldsymbol{\theta}^*, \boldsymbol{\psi}|\{u_i\}, \mu_0, \tau_0, e_0, f_0$, where $\boldsymbol{\theta} = (\vartheta_1, ..., \vartheta_N)'$, $\boldsymbol{\theta}^* = (\vartheta_1^*, ..., \vartheta_M^*)'$, $M \leq N$ contains the set of unique values from the $\boldsymbol{\theta}$ with ϑ_m^*, $m = 1, ..., M$ representing a cluster location and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_N)'$ is the vector of the latent indicator variables such that $\psi_i = m$ iff $\vartheta_i = \vartheta_m^*$. Together $\boldsymbol{\theta}^*$ and $\boldsymbol{\psi}$ completely define $\boldsymbol{\theta}$ (MacEachern, 1994). Let also $\boldsymbol{\theta}^{*(i)} = (\vartheta_1^{*(i)}, ..., \vartheta_{M^{(i)}}^{*(i)})'$ denote the distinct values in $\boldsymbol{\theta}^{(i)}$, which is the $\boldsymbol{\theta}$ with the element ϑ_i deleted. Also, the number of clusters in $\boldsymbol{\theta}^{*(i)}$ is indexed from $m = 1$ to $M^{(i)}$. Furthermore, we define $n_m^{(i)} = \sum_j \mathbf{1}(\psi_j = m,\ j \neq i)$, $m = 1, ..., M^{(i)}$ to be the number of elements in $\boldsymbol{\theta}^{(i)}$ that take the distinct element $\vartheta_m^{*(i)}$.

We follow a two-step process in order to draw from $\boldsymbol{\theta}^*, \boldsymbol{\psi}|\{u_i\}, \mu_0, \tau_0, e_0, f_0$. In the first step , we sample $\boldsymbol{\psi}$ and M by drawing ϑ_i, $i = 1, ..., N$ from

$$\vartheta_i|\boldsymbol{\theta}^{(i)}, u_i, G_0 \sim c\frac{a}{a+N-1}q_{i0}p(\vartheta_i|u_i, \mu_0, \tau_0, e_0, f_0) + \sum_{m=1}^{M^{(i)}} \frac{c}{a+N-1}n_m^{(i)}q_{im}\delta_{\vartheta_m^{*(i)}}(\vartheta_i),$$

setting $\psi_i = M^{(i)} + 1$ and $\vartheta_i = \vartheta_{M^{(i)}+1}^*$ when $\vartheta_{M^{(i)}+1}^*$ is sampled from $p(\vartheta_i|u_i, \mu_0, \tau_0, e_0, f_0)$ or $\psi_i = m$, when $\vartheta_i = \vartheta_m^{*(i)}$, $m = 1, ..., M^{(i)}$. c is the normalizing constant and $\delta_{\vartheta_j}(\vartheta_i)$ represents a unit point mass at $\vartheta_i = \vartheta_j$. The new cluster value $\vartheta_{M^{(i)}+1}^*$ is sampled from $p(\vartheta_i|u_i, \mu_0, \tau_0, e_0, f_0)$, which is the posterior density of ϑ_i under the prior G_0. By conjugacy we have

$$\vartheta_i = (\mu_i, \sigma_i^2)|u_i, \mu_0, \tau_0, e_0, f_0 \sim N(\mu_i|\overline{\mu_0}, \overline{\tau_0}\sigma_i^2)\mathcal{IG}(\sigma_i^2|\frac{\overline{e_0}}{2}, \frac{\overline{f_0}}{2}),$$

where

$$\overline{\mu_0} = \frac{\mu_0 + \tau_0 u_i}{1+\tau_0}, \qquad \overline{\tau_0} = \frac{\tau_0}{1+\tau_0}, \overline{e_0} = e_0 + 1, \qquad \overline{f_0} = f_0 + \frac{(u_i - \mu_0)^2}{\tau_0 + 1}.$$

The probability of assigning ψ_i to a new cluster is proportional to the marginal density of u_i, $\tilde{q}_{i0} = \int f(u_i|\vartheta_i)dG_0(\vartheta_i) = q_t(u_i|\mu_0, (1+\tau_0)f_0/e_0, e_0)$, where q_t is the Student-t distribution, μ_0 is the mean, e_0 is the degrees of freedom and $(1+\tau_0)f_0/e_0$ is the scale factor. The probability of ψ_i equaling an existing cluster $m = 1, ..., M^{(i)}$ is proportional to $n_m^{(i)} q_{im}$, where $\tilde{q}_{im}$ is the normal distribution of u_i evaluated at $\vartheta_m^{*(i)}$; hence, $\tilde{q}_{im} = n_m^{(i)} \exp(-\frac{1}{2}\left(u_i - \mu_m^{*(i)}\right)^2 / \sigma_m^{*2(i)})$.

In the second step, given M and $\boldsymbol{\psi}$, we draw each ϑ_m^*, $m = 1, ..., M$ from

$$\vartheta_m^* = (\mu_m^*, \sigma_m^{*2})|\{u_i\}_{i \in F_m}, \mu_0, \tau_0, e_0, f_0 \sim N(\mu_m^*|\overline{\mu_m}, \overline{\tau_m}\sigma_m^{*2})\mathcal{IG}(\sigma_m^{*2}|\frac{\overline{e_m}}{2}, \frac{\overline{f_m}}{2}),$$

where

$$\overline{\mu_m} = \frac{\mu_0 + \tau_0 \sum\limits_{i \in F_m} u_i}{1 + \tau_0 n_m}, \qquad \overline{\tau_m} = \frac{\tau_0}{1 + \tau_0 n_m},$$

$$\overline{e_m} = e_0 + n_m, \overline{f_m} = f_0 + \frac{n_m(\frac{1}{n_m} \sum\limits_{i \in F_m} u_i - \mu_0)^2}{1 + \tau_0 n_m} + \sum_{i \in F_m} (u_i - \frac{1}{n_m} \sum_{i \in F_m} u_i)^2,$$

and $F_m = \{i : \vartheta_i = \vartheta_m^*\}$ is the set of individuals that share the same parameter ϑ_m^*.

-

To sample the precision parameter a we first sample $\tilde{\eta}$ from $\tilde{\eta}|a, N \sim Beta(a + 1, N)$, where $\tilde{\eta}$ is a latent variable and then sample a from a mixture of two gammas, $a|\tilde{\eta}, \underline{c}, \underline{d}, M \sim \pi_{\tilde{\eta}} \mathcal{G}(\underline{c} + M, \underline{d} - \ln(\tilde{\eta})) + (1 - \pi_{\tilde{\eta}}) \mathcal{G}(\underline{c} + M - 1, \underline{d} - \ln(\tilde{\eta}))$ with the mixture weight $\pi_{\tilde{\eta}}$ satisfying $\pi_{\tilde{\eta}}/(1 - \pi_{\tilde{\eta}}) = (\underline{c} + M - 1)/N(\underline{d} - \ln(\tilde{\eta}))$. For further details, see Escobar and West (1994).

5.2.2 Empirical results for the semiparametric model

Table 4: Empirical results for the semiparametric proposed model

$\ln y^*_{it-1}$	0.01062
	(0.0303)
SS	-0.1135
	(0.1064)
$\ln SIZE$	-0.0732
	(0.0414)
$\ln R_0$	0.3172*
	(0.0690)
$\ln R_1$	-0.0719
	(0.0677)
$\ln R_2$	0.0520
	(0.0660)
$\ln R_3$	0.0177
	(0.0624)
$\ln R_4$	0.0269
	(0.0573)
$\ln R_5$	-0.0097
	(0.0502)
YEAR=1976	-0.0408
	(0.0225)
YEAR=1977	-0.0408
	(0.0272)
YEAR=1978	-0.1550*
	(0.0309)
YEAR=1979	-0.2203*
	(0.0336)
σ^2_v	0.0354*
	(0.0037)
ρ	0.8523*
	(0.0459)
a	0.4288*
	(0.3150)

*Significant based on the 95% highest posterior density interval. Standard deviations in parentheses. The APE for y_{it-1} is 0.0111 with a standard deviation of 0.0070.

Table 5: Empirical results for Wooldridge's (2005) regression

h_1	0.7729*
	(0.0401)
$h_{21_{(\ln R_0)}}$	-0.0732
	(0.3983)
$h_{22_{(\ln R_1)}}$	0.0529
	(0.5987)
$h_{23_{(\ln R_2)}}$	-0.0444
	(0.6157)
$h_{24_{(\ln R_3)}}$	0.0854
	(0.6732)
$h_{25_{(\ln R_4)}}$	-0.1804
	(0.5125)
$h_{26_{(\ln R_5)}}$	0.1900
	(0.3677)

*Significant based on the 95% highest posterior density interval. Standard deviations in parentheses.

References

Chib S. 2001. *Markov chain Monte Carlo methods: computation and inference.* in J. J. Heckman and E. Leamer (Eds.), Handbook of Econometrics, Volume 5, pp. 3569-3649. Amsterdam: North-Holland.

Chib S, Greenberg E. 1995. Understanding the Metropolis-Hastings algorithm. *The American Statistician* **49**: 327–335.

Chib S, Greenberg E, Winkelmann R. 1998. Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics* **86**: 33 – 54.

Chib S, Hamilton B. 2002. Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics* **110**: 67 – 89.

Chib S, Jeliazkov I. 2006. Inference in semiparametric dynamic models for Binary longitudinal data. *Journal of the American Statistical Association* **101**: 685–700.

Escobar M, West M. 1994. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**: 577–588.

Ferguson T. 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**: 209–230.

Gelfand A. 1996. Model determination using sampling-based methods. In Gilks WR, Richardson S, Spiegelhalter DJ (eds.) *Markov Chain Monte Carlo in Practice*, chapter 9. London, UK: Chapman & Hall, 145–161.

Gelfand A, Dey D. 1994. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**: 501–514.

Geweke J. 1992.. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments.* In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), Bayesian Statistics, Volume 4, pp. 169-188. New York: Oxford University Press.

Ghosal S. 2010. *The Dirichlet process, related priors and posterior asymptotics.* in Hjort N, Holmes C, Mller P, Walker SG. (Eds.). Bayesian nonparametrics: principles and practice. Cambridge University Press.

Jochmann M, Len-Gonzlez R. 2004. Estimating the demand for health care with panel data: a semiparametric Bayesian approach. *Health Economics* **13**: 1003–1014.

MacEachern S. 1994. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics. Simulation and Computation* **23**: 727–741.

Navarro D, Griffiths T, Steyvers M, Lee M. 2006. Modeling individual differences using dirichlet processes. *Journal of Mathematical Psychology* **50**: 101 – 122.

Schwarz G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6**: 461–464.

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

Zheng X. 2008. Semiparametric Bayesian estimation of mixed count regression models. *Economics Letters* **100**: 435 – 438.