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.

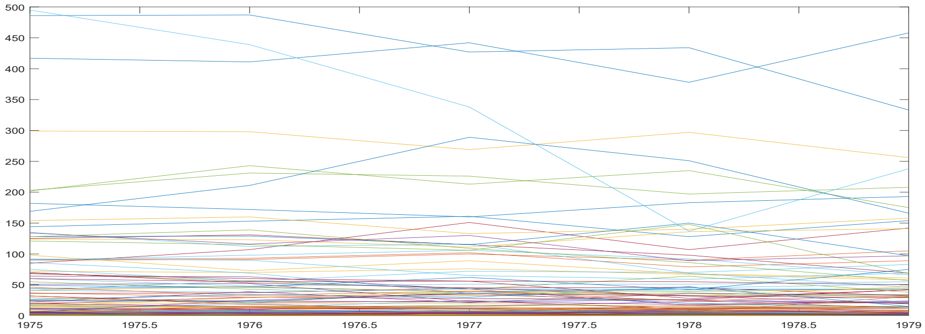

Figure 1: Empirical results. Plot of the dependent variable for all firms over time

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^2_u	0.1091*	0.1481*	0.9942*	
	(0.0386)	(0.0208)	(0.0963)	
σ^2_v	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)
$h21_{(\ln R_0)}$	-0.0799	-0.1460
	(0.3551)	(0.3254)
$h22_{(\ln R_1)}$	0.0490	0.0524
	(0.5988)	(0.5646)
$h23_{(\ln R_2)}$	-0.0432	-0.0557
	(0.6324)	(0.5941)
$h24_{(\ln R_3)}$	0.0809	0.0168
	(0.6028)	(0.5535)
$h25_{(\ln R_4)}$	-0.2022	-0.1171
	(0.5466)	(0.4891)
$h26_{(\ln R_5)}$	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.