UNIVERSIDAD CARLOS III DE MADRID working papers

UC3M Working papers
Economics
16-10
July, 2016
ISSN 2340-5031

Departamento de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 916249875

Score-driven dynamic patent count panel data models

Szabolcs Blazsek[a] and Alvaro Escribano[b],[*]

Abstract

This paper suggests new Dynamic Conditional Score (DCS) count panel data models. We compare the statistical performance of static model, finite distributed lag model, exponential feedback model and different DCS count panel data models. For DCS we consider random walk and quasi-autoregressive formulations of dynamics. We use panel data for a large cross section of United States firms for period 1979 to 2000. We estimate models by using the Poisson quasi-maximum likelihood estimator with fixed effects. The estimation results and diagnostics tests suggest that the statistical performance of DCS-QAR is superior to that of alternative models.

Keywords: research and development; count panel data; dynamic conditional score; quasi-maximum likelihood.
JEL classification codes: C33, C35, C51, C52, O3.
Word count: 1,955

[a] *School of Business, Universidad Francisco Marroquín, Guatemala. Address: School of Business, Universidad Francisco Marroquín, Calle Final 6, 01010, Ciudad de Guatemala, Guatemala. E-mail: sblazsek@ufm.edu.*

[b] *Department of Economics, Universidad Carlos III de Madrid, Spain. Address: Department of Economics, Universidad Carlos III de Madrid, Calle Madrid 126, 28903, Getafe (Madrid), Spain. E-mail: alvaroe@eco.uc3m.es*

[*] *Corresponding author*

1. Introduction

Gourieroux et al. (1984a, 1984b) and Wooldridge (1997a, 2002) motivate the use of the Quasi-Maximum Likelihood Estimator (QMLE) for count panel data models. For QMLE, a pseudo Log-Likelihood (LL) objective function is maximized, for which the pseudo probability distribution is within the Linear Exponential Family (LEF). An example of LEF is the Poisson distribution. In this paper we use Poisson QMLE for patent count panel data models, hence we use $n_{it}|\mathcal{F}_t \sim \text{Poisson}(\lambda_{it})$ as a pseudo distribution for the patent count variable n_{it}. For this distribution (i) $E(n_{it}|\mathcal{F}_t) = \lambda_{it}$, (ii) the log of the conditional probability mass function is

$$\ln f(n_{it}|\mathcal{F}_t) = -\lambda_{it} + n_{it} \ln \lambda_{it} - \ln(n_{it}!) \tag{1}$$

(iii) the conditional score of n_{it} with respect to λ_{it} is

$$\frac{\partial \ln f(n_{it}|\mathcal{F}_t)}{\partial \lambda_{it}} = \frac{n_{it}}{\lambda_{it}} - 1 = s_{it} \tag{2}$$

(iv) under correct specification of the conditional mean of n_{it}, $(s_{i1}, \ldots, s_{iT})$ is a martingale difference sequence with respect to $\mathcal{F}_t$. In this paper, we suggest count panel data models for which the error term e_{it} is possibly serially correlated. We introduce serial correlation into e_{it} by the dynamic variable Ψ_{it} that is updated by the pseudo conditional score s_{it-1}. We name these models as Dynamic Conditional Score (DCS) count panel data models.

2. DCS patent count panel data models

In the body of literature Davis et al. (2003, 2005) and Harvey (2013) suggest dynamic time-series models for Poisson dependent variables updated by the conditional score. In this paper we extend those works since (i) we use panel data models with unobserved effects, (ii) we consider autoregressive dynamics for the impact of conditional score, and (iii) we use robust Poisson QMLE for statistical inference. The DCS count panel data model is

$$n_{it} = \exp(X_{it}'\beta) v_i e_{it} = \exp(X_{it}'\beta) v_i h(\Psi_{it}) \epsilon_{it} \tag{3}$$

for a panel of $i = 1, \ldots, N$ firms and $t = 1, \ldots, T$ years, where X_{it} is a vector of explanatory variables, v_i represents unobserved effects, e_{it} is a possibly serially correlated error term with $E(e_{it}) = 1$, Ψ_{it} is a possibly serially correlated term with $E(\Psi_{it}) = 0$, and ϵ_{it} is an i.i.d. term with $E(\epsilon_{it}) = 1$. For Ψ_{it} we consider two alternatives. First, the Random Walk (RW) specification is

$$\Psi_{it} = \Psi_{it-1} + \gamma_1 s_{it-1} \tag{4}$$

Second, the first-order Quasi-Autoregressive (QAR) specification (Harvey, 2013) is

$$\Psi_{it} = \alpha_1 \Psi_{it-1} + \gamma_1 s_{it-1} \quad \text{with} \quad |\alpha_1| < 1 \tag{5}$$

We initialize both filters by parameter Ψ_0. For $h(\Psi_{it})$ we use a function for which $E[h(\Psi_{it})] = 1$ if $E(\Psi_{it}) = 0$. Some examples of $h(\Psi_{it})$ are

$$h(\Psi_{it}) = \tanh(\Psi_{it}) + 1 \tag{6}$$

$$h(\Psi_{it}) = \frac{1 - \exp(-\Psi_{it})}{1 + \exp(-\Psi_{it})} + 1 \tag{7}$$

$$h(\Psi_{it}) = 2F(\Psi_{it}) \tag{8}$$

where $\tanh(\cdot)$ is the hyperbolic tangent function and $F(\cdot)$ is the distribution function of any continuous symmetric probability distribution centered at zero.

3. Statistical inference

We estimate the parameters of DCS patent count panel data models by using QMLE with fixed effects. We maintain the following assumptions:

(A1) (pre-sample data) Pre-sample data $(n_{it} : t = 1, \ldots, P)$ and $(X_{it} : t = 1, \ldots, P)$ are available. Let $\mathcal{F}_P$ denote the information set generated by pre-sample data.

(A2) (fixed effects) Replace v_i by $p_i(\mathcal{F}_P) > 0$, where $p_i(\mathcal{F}_P)$ includes averages of n_{it} and X_{it} that are computed for the pre-sample data period.

(A3) (correct specification of the mean) $E(n_{it}|X_{it}, \Psi_{it}, \mathcal{F}_P) = \exp(X'_{it}\beta)p_i(\mathcal{F}_P)h(\Psi_{it})$.

(A4) (martingale difference sequence) $(s_{it} : t = 1, \ldots, T)$ is a martingale difference sequence with respect to $\mathcal{F}_t = (X_{it}, \Psi_{it}, \mathcal{F}_P)$.

(A5) (exogeneity) All variables in X_{it} are predetermined (Blundell et al., 2002) (alternatively, all variables in X_{it} satisfy the sequential moment restrictions; Chamberlain, 1992 and Wooldridge, 1997a, 1997b, 2002).

We estimate the parameters consistently by using the pooled Poisson QMLE method with $\lambda_{it} = \exp(X'_{it}\beta)p_i(\mathcal{F}_P)h(\Psi_{it})$, by solving the maximization problem

$$\arg\max_{\Theta} LL(\Theta) = \arg\max_{\Theta} \sum_{i=1}^{N}\sum_{t=1}^{T} -\lambda_{it} + n_{it}\ln\lambda_{it} - \ln(n_{it}!) \tag{9}$$

For this estimation the pseudo score is $s_{it} = n_{it}/[\exp(X'_{it}\beta)p_i(\mathcal{F}_P)h(\Psi_{it})] - 1$. For the pooled Poisson QMLE, we use the asymptotic distribution and robust covariance matrix of parameter estimates of Wooldridge (1997a, 2000).

4. Data

The source of the United States (US) utility patent dataset of this work is MicroPatent LLC. The patent database includes the US Patent and Trademark Office (USPTO) patent number, application date, publication date, USPTO patent number of cited patents, three-digit US technological class and company name for each patent. We perform all data procedures according to the recommendations of Hall et al. (2001). We count the number of successful patent applications n_{it} for each firm and year. We measure spillovers of knowledge among firms by the log of the number of citations made to past patents of other firms of the same industry IA_{it} and to past patents of other firms of other industries IE_{it}. Company specific information is from the Standard & Poor's Compustat data files. We use inflation-corrected log R&D expenses r_{it} to measure R&D investment. We created a match file and crossed the patent and firm datasets.

The dataset includes 488,149 US utility patents with application dates for period 1979 to 2000 (22 years) of 4,476 US firms ($N = 4,476$). We divide the full data window into two subperiods.

First, the pre-sample data window is for period 1979 to 1983 ($P = 5$ years). Second, the in-sample data window is for period 1984 to 2000 ($T = 17$ years). It is noteworthy that Blazsek and Escribano (2010, 2016) use the same dataset.

5. Competing patent count panel data models

We compare five alternative multiplicative patent count panel data models. The first specification is the Static Model (SM) for patent counts. For this model $\Psi_{it} = 0$ and $X_{it}'\beta$ is

$$X_{it}'\beta = c + \zeta_1 t + \zeta_2(t \times r_{it}) + \zeta_3 r_{it}^2 + \kappa_0 r_{it} + \nu_0 r_{it}\mathrm{IA}_{it} + \xi_0 r_{it}\mathrm{IE}_{it} \tag{10}$$

where r_{it}^2, IA_{it} and IE_{it} are motivated by Blazsek and Escribano (2010, 2016). The second specification is the Finite Distributed Lag (FDL) model (Hausman et al., 1984) for which

$$X_{it}'\beta = c + \zeta_1 t + \zeta_2(t \times r_{it}) + \zeta_3 r_{it}^2 + \sum_{k=0}^{5} \kappa_k r_{it-k} + r_{it}\sum_{k=0}^{5} \nu_k \mathrm{IA}_{it-k} + r_{it}\sum_{k=0}^{5} \xi_k \mathrm{IE}_{it-k} \tag{11}$$

and $\Psi_{it} = 0$. The third specification is the Exponential Feedback Model (EFM) (Wooldridge, 2005) for which $X_{it}'\beta$ is according to Equation 10 and $h(\Psi_{it}) = \exp[g(n_{it-1})]$ with $g(n_{it-1}) = \alpha_1 \mathbb{1}\{n_{it-1} > 0\}\ln(n_{it-1})$. The fourth and fifth specifications are DCS count panel data models with RW and QAR(1), respectively. For DCS $X_{it}'\beta$ is Equation 10 and $h(\Psi_{it}) = \tanh(\Psi_{it}) + 1$. We also considered alternatives of $h(\cdot)$, but estimation results were identical. We estimate all models by Poisson QMLE with fixed effects, and we use $p_i(\mathcal{F}_P) = \exp(\delta_1 \overline{n}_i + \delta_2 \overline{r}_i)$ where the averages are computed for pre-sample data (Blundell et al., 2002).

6. Empirical results

Table 1 presents the parameter estimates and robust standard errors for all models. Table 2 presents the Average Partial Effects (APE) of r_{it} for cross-section and time-series dimensions. Figure 1 presents the evolution of APE of r_{it} for the cross-section dimension. APE is interpreted as the average increase in n_{it} due to a 1% increase in R&D expenses. It is noteworthy that Ψ_{it} is averaged out by APE for DCS. Table 2 presents four model selection metrics: (i) mean LL; (ii) mean Akaike Information Criterion (AIC); mean Bayesian Information Criterion (BIC); (iv) mean Hannan-Quinn Criterion (HQC) (Hamilton, 1994). All criteria suggest that the in-sample statistical performance of DCS-QAR(1) is superior to the alternatives. Table 2 presents two tests for the serial correlation of residuals ϵ_{it}. For the first test, we estimate the AR(1) model

$$\epsilon_{it} = c^* + \rho\epsilon_{it-1} + u^*_{it} \tag{12}$$

by using robust System Ordinary Least Squares (SOLS) (Wooldridge, 2002). For the second test, we use the Arellano–Bond (1991) dynamic panel data model

$$\epsilon_{it} = c^* + \rho\epsilon_{it-1} + v^*_i + u^*_{it} \tag{13}$$

and estimate the first-differenced model $\Delta\epsilon_{it} = \rho\Delta\epsilon_{it-1} + \Delta u^*_{it}$ by using robust optimal System Generalized Method of Moments (SGMM) (Wooldridge, 2002), for which we use $(\epsilon_{it-2}, \ldots, \epsilon_{it-6})$ as instrumental variables. For SGMM the Over-Identification Test Statistics (OITS) (Wooldridge, 2002) suggest that all instrumental variables are exogenous for all models. Both SOLS and SGMM suggest significant first-order serial correlation of ϵ_{it} for SM, which motivates dynamic specifications for patent count panel data. For EFM the SOLS and SGMM results are mixed, but for DCS none of those estimates indicate significant first-order serial correlation.

Acknowledgments

Both authors acknowledge and give thanks for the helpful suggestions of Jeffrey Wooldridge. Szabolcs Blazsek acknowledges financial support from Universidad Carlos III de Madrid and Universidad Francisco Marroquín. Alvaro Escribano acknowledges financial support from the Spanish Ministry of Innovation and Science ECO2012-36559, ECO2015-68715-R, Consolider Grant #2006/04046/002 and María de Maeztu Grant MDM 2014-0431.

References

Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Rev. Econ. Stud. 58, 277–297. doi: 10.2307/2297968

Blazsek, S., Escribano, A., 2010. Knowledge spillovers in U.S. patents: A dynamic patent intensity model with secret common innovation factors. J. Econometrics. 159, 14–32. doi: 10.1016/j.jeconom.2010.04.004

Blazsek, S., Escribano, A., 2016. Patent propensity, R&D and market competition: Dynamic spillovers of innovation leaders and followers. J. Econometrics. 191, 145–163. doi: 10.1016/j.jeconom.2015.10.005

Blundell, R., Griffith, R., Windmeijer, F., 2002. Individual effects and dynamics in count data models. J. Econometrics. 108, 113–131. doi: 10.1016/S0304-4076(01)00108-7

Chamberlain, G., 1992. Efficiency bounds for semiparametric regression. Econometrica. 60, 567–596. doi: 10.2307/2951584

Davis, R., Dunsmuir, W., Streett, S., 2003. Observation-driven models for Poisson counts. Biometrika. 90, 777–790. doi: 10.1093/biomet/90.4.777

Davis, R., Dunsmuir, W., Streett, S., 2005. Maximum likelihood estimation for an observation driven model for Poisson counts. Methodol. Comput. Appl. 7, 149–159. doi: 10.1007/s11009-005-1480-4

Gourieroux, C., Monfort, A., Trognon, A., 1984a. Pseudo maximum likelihood methods: Theory. Econometrica. 52, 681–700. doi: 10.2307/1913471

Gourieroux, C., Monfort, A., Trognon, A., 1984b. Pseudo maximum likelihood methods: Applications to Poisson models. Econometrica. 52, 701–720. doi: 10.2307/1913472

Hall, B., Jaffe, A.B., Trajtenberg, M., 2001. The NBER patent citation data file: Lessons, insights and methodological tools. NBER Working Paper No. 8498.

Hamilton, J.D., 1994. Time Series Analysis, Princeton University Press, Princeton.

Harvey, A.C., 2013. Dynamic Models for Volatility and Heavy Tails, Cambridge University Press, Cambridge. doi: 10.1017/CBO9781139540933

Hausman, J., Hall, B., Griliches, Z., 1984. Econometric models for count data with an application to the patents-R&D relationship. Econometrica. 52, 909–938. doi: 10.2307/1911191

Wooldridge, J.M., 1997a. Quasi-likelihood methods for count data, in: Pesaran, M.H., Schmidt, P. (Eds.), Handbook of Applied Econometrics, Volume 2, Blackwell, Oxford, pp. 352–406.

Wooldridge, J.M., 1997b. Multiplicative panel data models without the strict exogeneity assumption. Economet. Theor. 13, 667–678. doi: 10.1017/S0266466600006125

Wooldridge, J.M., 2002. Econometric Analysis of Cross Section and Panel Data, The MIT Press, Cambridge.

Wooldridge, J.M., 2005. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. J. Appl. Econom. 20, 39–54. doi: 10.1002/jae.770

Table 1. Parameter estimates for Poisson QMLE with fixed effects

	SM	FDL	EFM	DCS-RW	DCS-QAR
c	$-1.065^{***}(0.1692)$	$-0.921^{***}(0.1451)$	$-1.052^{***}(0.0774)$	$2.449^{***}(0.0644)$	$3.454^{***}(0.0887)$
ζ_1	$0.068^{***}(0.0168)$	$0.054^{***}(0.0133)$	$0.053^{***}(0.0063)$	$0.067^{***}(0.0061)$	$-0.012^{**}(0.0060)$
ζ_2	$-0.009^{***}(0.0029)$	$-0.005^{*}(0.0031)$	$-0.012^{***}(0.0013)$	$-0.009^{***}(0.0015)$	$-0.009^{***}(0.0014)$
ζ_3	$-0.029^{***}(0.0081)$	$-0.033^{***}(0.0103)$	$-0.021^{***}(0.0015)$	$0.023^{***}(0.0028)$	$0.027^{***}(0.0028)$
κ_0	$1.014^{***}(0.1393)$	$0.609^{***}(0.1044)$	$0.455^{***}(0.0292)$	$0.186^{**}(0.0770)$	$0.175^{**}(0.0753)$
κ_1	NA	$0.140^{*}(0.0766)$	NA	NA	NA
κ_2	NA	$0.149^{***}(0.0499)$	NA	NA	NA
κ_3	NA	$0.134^{*}(0.0791)$	NA	NA	NA
κ_4	NA	$0.013(0.0197)$	NA	NA	NA
κ_5	NA	$0.016(0.0479)$	NA	NA	NA
ν_0	$0.007(0.0052)$	$0.008(0.0057)$	$0.002^{***}(0.0005)$	$0.008^{***}(0.0013)$	$0.008^{***}(0.0017)$
ν_1	NA	$0.005^{***}(0.0020)$	NA	NA	NA
ν_2	NA	$0.000(0.0051)$	NA	NA	NA
ν_3	NA	$-0.009^{**}(0.0039)$	NA	NA	NA
ν_4	NA	$-0.026^{***}(0.0079)$	NA	NA	NA
ν_5	NA	$0.042^{***}(0.0144)$	NA	NA	NA
ξ_0	$0.004(0.0102)$	$0.004(0.0111)$	$0.002^{**}(0.0010)$	$0.003(0.0027)$	$0.001(0.0033)$
ξ_1	NA	$-0.004(0.0056)$	NA	NA	NA
ξ_2	NA	$0.009(0.0150)$	NA	NA	NA
ξ_3	NA	$0.033^{***}(0.0115)$	NA	NA	NA
ξ_4	NA	$0.015(0.0093)$	NA	NA	NA
ξ_5	NA	$-0.107^{***}(0.0358)$	NA	NA	NA
δ_1	$0.001^{***}(0.0004)$	$0.002^{***}(0.0005)$	$0.000(0.0002)$	$0.002^{***}(0.0002)$	$0.002^{***}(0.0002)$
δ_2	$0.100(0.1453)$	$0.043(0.1119)$	$-0.003(0.0292)$	$0.340^{***}(0.0819)$	$0.310^{***}(0.0793)$
α_1	NA	NA	$0.887^{***}(0.0227)$	1.000(NE)	$0.970^{***}(0.0318)$
γ_1	NA	NA	NA	$0.099(0.0617)$	$0.121^{*}(0.0665)$
Ψ_0	NA	NA	NA	$-1.599^{***}(0.0623)$	$-2.061^{***}(0.0722)$

Notes: Not Available (NA); Not Estimated (NE). Robust standard errors are reported in parentheses. *, ** and *** indicate parameter significance at the 10%, 5% and 1% levels, respectively.

Table 2. Statistical performance and residual diagnostics

	SM	FDL	EFM	DCS-RW	DCS-QAR
Cross-section and time-series average partial effects:					
	3.5621	1.2803	0.6012	1.9681	2.1436
Statistical performance metrics:					
mean LL	−4.2907	−4.0618	−2.0681	−2.1143	−**2.0600**
mean AIC	8.5817	8.1242	4.1365	4.2290	**4.1204**
mean BIC	8.5818	8.1245	4.1366	4.2291	**4.1205**
mean HQC	8.5817	8.1242	4.1365	4.2290	**4.1204**
AR(1) model for residuals, robust SOLS:					
ρ	0.0493	0.0461	0.0275	0.0022	0.0001
p-value ρ	0.0860	0.2211	0.0068	0.5052	0.3355
Arellano–Bond model for residuals, robust optimal SGMM:					
ρ	−0.0552	−0.0062	0.0022	−0.0013	0.0000
p-value ρ	0.0128	0.3626	0.2318	0.2930	0.1223
OITS	5.4054	4.8676	4.8216	5.2005	1.3606
p-value OITS	0.2482	0.3011	0.3061	0.2673	0.8510

Notes: Bold numbers indicate superior statistical performance.

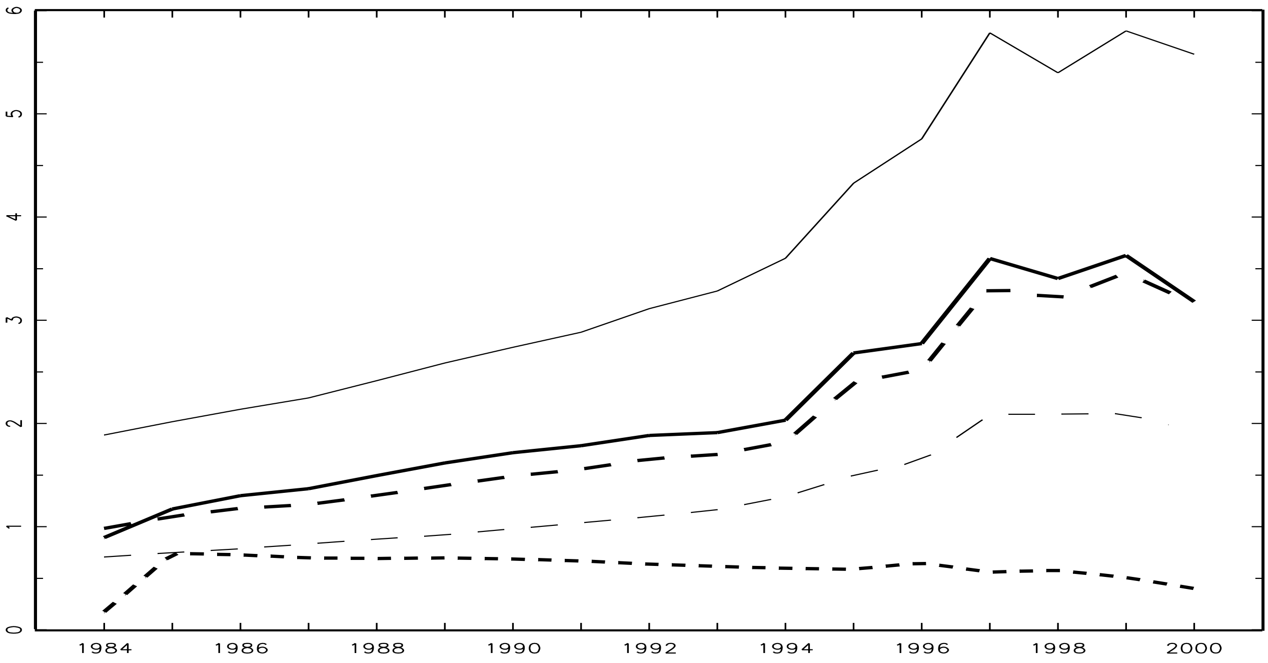

Figure 1. Cross-section average partial effects of r_{it}

Notes: SM (solid thin); FDL (dashed thin); EFM (short dashes thick); RW (dashed thick); QAR (solid thick)