GESG seminar, 19 February 2016
Universidad Francisco Marroquín, Guatemala

Dynamic conditional score patent count panel data models

SZABOLCS BLAZSEK, UNIVERSIDAD FRANCISCO MARROQUÍN
ALVARO ESCRIBANO, UNIVERSIDAD CARLOS III DE MADRID

Motivation

The main interest of this paper is to suggest new dynamic score-driven models for measuring the relationships between R&D expenses and patent application activity.

There is an extensive body of literature on this issue (e.g., Hausman, Hall and Griliches, 1984; Blundell, Griffith and Windmeijer, 2002; Wooldridge, 2005; and many more).

All these works suggest different specifications and use different estimation methods to identify R&D-patent relationships.

Motivation

Recently, **score-driven dynamic econometric models** appeared.

This family of models is named **dynamic conditional score** (DCS) (Harvey and Chakravarty, 2008; Harvey, 2013) or **generalized autoregressive score** (GAS) models (Creal, Koopman and Lucas, 2008).

These are observation-driven models that can be relatively simply estimated, typically they fit well to data, and for many DCS model the asymptotic properties of maximum likelihood estimator are derived.

Different DCS time-series models were suggested for location, scale and association (Harvey, 2013). For count data, however, there is only a suggestion in Harvey (2013, Ch. 5.11).

Contribution

We suggest a new class of **score-driven patent count panel data models**, and investigate if the performance of these models is superior to previous patent count panel data models of the body of literature.

Structure

1. Data and variables

2. Maximum likelihood estimation (MLE)

3. Quasi maximum likelihood estimation (QMLE)

4. Competing count panel data models

5. Model diagnostics and results

6. References

Data and variables

Panel data

We use the U.S. patent and firm specific variable dataset of Blazsek and Escribano (2010 JoE, 2016 JoE).

This panel includes firm level annual data of 4476 U.S. companies for period 1979 to 2000.

Variables in this panel are indexed by firm $i = 1, \ldots, N$ and year $t = 1, \ldots, T$.

Observable variables

n_{it} : patent application count for firm i in year t

rd_{it} : log of inflation-adjusted R&D expenditure

D_i : hi-tech dummy, it takes the value one for drug, computer, scientific instrument, chemical and electronic components industries, and zero otherwise

b_i : firm size, log of the inflation adjusted book value for the simple midpoint year

Observable variables

n_{it} : patent application count for firm i in year t

This is the dependent variable of the model.

This is a non-negative integer number, hence the econometric model to be applied is a non-linear **count data model**.

In the literature several models were proposed for the **conditional expectation** of n_{it}, denoted by λ_{it}, in order to study the determinants of patent application activity.

Maximum likelihood estimation

Hausman, Hall and Griliches (HHG) (1984, Econometrica)

$E(n_{it}|X_{it},\alpha_i) = \lambda_{it} = \exp(X_{it}\theta)\,\alpha_i$

$X_{it}\theta = \mu_0 + \gamma_1 t + \gamma_2(t \times rd_{it}) + \gamma_3 D_i + \gamma_4 b_i + \beta_0 rd_{it} + \cdots + \beta_5 rd_{it-5}$

$X_{it}\theta$ is used to simplify notation (observable variables)

$\alpha_i > 0$ represents *unobserved and time-constant firm-specific effects*. α_i captures time-constant omitted variables.

α_i can be considered for panel data models; which is one of their main advantage.

Assumptions of HHG (1984)

(ML1) $n_{it}|(X_{i1},\dots,X_{iT},\alpha_i)\sim\mathrm{Poisson}(\lambda_{it})$. This assumption implies strict exogeneity of all explanatory variables conditional on α_i.

(ML2) λ_{it} is modelled by the exponential function to ensure positivity, i.e. $\lambda_{it}=\exp[X_{it}\theta]\alpha_i=\exp[X_{it}\theta+\ln\alpha_i]$. Multiplicative count data model.

(ML3) $n_{it}|(X_{i1},\dots,X_{it},\alpha_i)$ and $n_{is}|(X_{i1},\dots,X_{is},\alpha_i)$ are independent

(ML4) α_i is i.i.d. with $\mathrm{Gamma}(1,\delta)$ distribution

RE MLE and FE MLE of HHG (1984)

Under (ML1), (ML2), (ML3) and (ML4), *random effects MLE (RE MLE)*.

For RE MLE α_i is independent of X_{it}.

Under (ML1), (ML2) and (ML3), *fixed effects MLE (FE MLE)*.

For FE MLE α_i may be associated with X_{it}.

Hence, FE MLE is more general than RE MLE.

Disadvantages of MLE

(a) exact conditional distribution is assumed for $n_{it}|(X_{i1}, \dots, X_{iT}, \alpha_i)$ for both RE MLE and FE MLE

(b) exact distribution is assumed for α_i for RE MLE

(c) strict exogeneity of all explanatory variables is assumed.

Strict exogeneity implies that (i) lags of n_{it} cannot be considered as explanatory variables within λ_{it}; (ii) n_{it} cannot have impact on future values of any explanatory variables, for example R&D.

Pooled Patent count data models

We can drop the **strict exogeneity assumption** by excluding the unobserved effects term α_i from the model.

This is going to permit lags of n_{it} as explanatory variables and also we can include such explanatory variables (e.g. R&D) for which future values (rd_{it+1},…, rd_{iT}) are influenced by n_{it}.

Nevertheless, we will still need the **contemporaneous exogeneity** of explanatory variables.

Of course, dropping α_i from the model is a price that we have to pay for dropping the strict exogeneity assumption.

Pooled Patent count data models

For the remainder, we shall see patent count data models for which α_i is excluded. The general name for this kind of panel data models is **pooled count panel data models**.

Quasi maximum likelihood estimation

QMLE method (Gourieroux, Monfort, Trognon, GMT, 1984 a, b, Econometrica)

The assumptions on exact distribution and strict exogeneity question the validity of MLE for patent count panel data models.

We use an alternative estimation method that is more robust since it does not require assumptions on the exact distribution of $n_{it}|(X_{i1},\ldots,X_{iT},\alpha_i)$ and α_i. Furthermore, the strict exogeneity assumption is also relaxed and lags of n_{it} will be permitted as explanatory variables within λ_{it}.

This alternative estimation method is named in the literature '**pseudo MLE**' (**PMLE**) or '**quasi MLE**' (**QMLE**).

GMT (1984 a, b)

QMLE, similar to MLE, provides parameter estimates by maximizing an objective function.

For MLE this objective function is the log of the joint density of $(n_{i1}, \ldots, n_{iT})$, and it is maximized with respect to the parameters.

MLE involves an assumption about the conditional distribution of the dependent variable.

GMT (1984 a, b)

GMT (1984 a, b) suggests objective functions to be maximized however it is not needed to assume a specific conditional distribution for the dependent variable.

The objective functions of QMLE are log density functions of a certain class of probability distributions, named the **linear exponential family (LEF)**.

LEF density functions have a certain functional form.

GMT (1984 a, b)

Some examples of discrete and continuous LEF distributions:

A) $Y \sim \text{Binomial}(n, p)$ distribution, with given n

B) $Y \sim \text{Poisson}(\lambda)$ distribution

C) $Y \sim \text{Negative binomial}(\delta, \lambda)$ distribution, with given δ

D) $Y \sim \text{Normal}(\mu, \sigma^2)$ distribution, with given σ

E) $(Y_1, \dots, Y_K) \sim \text{Multinomial}(n, p_1, \dots, p_K)$ distribution, with given n

F) $(Y_1, \dots, Y_K) \sim \text{Multivariate normal } (\mu, \Sigma)$, with given Σ

GMT (1984 a, b)

For each of the LEF there is the corresponding objective function to be maximized. This objective function is given by the log joint density of the LEF distribution.

However, this is a *pseudo* log likelihood, since we do not assume that data are in fact generated from the LEF, i.e. QMLE uses an auxiliary objective function in order to estimate parameters.

GMT (1984 a, b) show that the QMLE method provides consistent parameter estimates.

The QMLE method is a special case of the more general **M-estimator** (see Wooldridge, 2002, Ch. 12).

QMLE for count data (Wooldridge, 1997a)

Wooldridge (1997a) uses the QMLE method for cross section and panel count data. QMLE assumption:

(QMLE1) $\mathrm{E}[n_{it}|X_{i1},\ldots,X_{it}]=\lambda_{it}$ where X_{it} can be any observed variable. *Correct specification of the conditional expectation.*

(QMLE1) also means that contemporaneous explanatory variables are exogenous.

Result of Wooldridge (1997a) that is related to GMT (1984 a, b):

Under (QMLE1), the pooled Poisson QMLE and also the pooled negative binomial QMLE provide consistent estimates of parameters; since both distributions are LEF.

QMLE for count data (Wooldridge, 1997a)

The pooled negative binomial QMLE is **more efficient** than the pooled Poisson QMLE since the negative binomial QMLE objective function has an additional parameter (δ), hence it is more flexible.

However, the negative binomial QMLE assumes that δ is already given.

Thus, for all patent count data models we apply the **two-step negative binomial QMLE procedure** suggested by Wooldridge (1997a). First step: estimate δ. Second step: use $\hat{\delta}$ in the negative binomial QMLE objective function to estimate the remaining parameters.

QMLE for count data (Wooldridge, 1997a)

This two-step negative binomial QMLE estimation procedure is directly related to the **two-step quasi-generalized pseudo MLE (QGPMLE) procedure** suggested by GMT (1984 a,b) in which the given parameter of the LEF is estimated in a first step, and remaining parameters are estimated in the second step.

GMM estimation

In the body of literature of count data models there exists an alternative estimation method that drops the strict exogeneity assumption: **generalized method of moments**, GMM (Hansen, 1982). (For count data: Chamberlain, 1992; Wooldridge, 1997b.)

In these models α_i is also included, and it may be correlated with the explanatory variables (i.e. it is a fixed effect, FE).

We had (i) computation problems (slow iterations); (ii) for the DCS count data models suggested in this paper α_i cannot be considered as it is unobservable. *Thus, we do not use GMM.*

Competing patent count data models

Pooled Hausman, Hall and Griliches (1984) *finite distributed lag (FDL) model*

$$E(n_{it}|X_{it}) = \lambda_{it} = \exp(X_{it}\theta)$$

$$X_{it}\theta = \mu_0 + \gamma_1 t + \gamma_2(t \times rd_{it}) + \gamma_3 D_i + \gamma_4 b_i + \beta_0 rd_{it} + \cdots + \beta_5 rd_{it-5}$$

Multiplicative count data model

Pooled Wooldridge (2005) *exponential feedback model (EFM)*

$E(n_{it}|X_{it}) = \lambda_{it} = \exp[X_{it}\theta + \gamma_5 n_{i1} + \varphi_1 n_{it-1}]$

This model considers AR(1) dynamics and also controls for the initial condition of the dynamic process ($\gamma_5 n_{i1}$).

Multiplicative count data model

Pooled Blundell, Griffith and Windmeijer (2002) *linear feedback model (LFM)*

$E(n_{it}|X_{it}\ ) = \lambda_{it} = \varphi_1 n_{it-1} + \exp(Y_{it}\tilde{\theta})$ with $0 < \varphi_1 < 1$

$Y_{it}\tilde{\theta} = X_{it}\theta + \gamma_5 n_{i1}$

Additive count data model

Multiplicative DCS (MDCS) Poisson model

Davis, Dunsmuir and Streett (2003, 2005), and Harvey (2013) suggest MLE for a dynamic model for time-series count data, updated by the score of the Poisson distribution.

Poisson is member of LEF, hence QMLE can also be used.

The score of n_{it} is computed with respect to λ_{it} as follows:

$$f(n_{it}|X_{i1},\dots,X_{it}) = \frac{\exp(-\lambda_{it})\lambda_{it}{}^{n_{it}}}{n_{it}!} \qquad \text{(density of Poisson)}$$

$$\frac{\partial \ln f(n_{it}|X_{i1},\dots,X_{it})}{\partial \lambda_{it}} = \frac{n_{it}}{\lambda_{it}} - 1 \equiv u_{it} > -1 \qquad \text{(conditional score)}$$

Multiplicative DCS (MDCS) Poisson model

Davis, Dunsmuir and Streett (2003, 2005); Harvey (2013) suggest dynamic models for count data in which only lags of u_{it} update the dynamic equation (i.e. MA-type terms), but they do not consider AR terms.

We extend their dynamic score-driven count data model and (i) we consider AR terms too in the dynamic equation; (ii) we suggest DCS count data models for panel data instead of time-series; (iii) we estimate the count data model by QMLE and not by MLE, hence Poisson distribution is not assumed for $(n_{it}|X_{i1},\ldots,X_{it})$.

Multiplicative DCS (MDCS) Poisson model

$E(n_{it}|X_{it}) = \lambda_{it} = \exp(\Psi_{it} + Y_{it}\tilde{\theta})$

MDCS-QMA(q) (Davis, Dunsmuir and Streett 2003, 2005; Harvey 2013):

$\Psi_{it+1} = \theta_0 u_{it} + \cdots + \theta_q u_{it-q}$

MDCS-QAR(1):

$\Psi_{it+1} = \varphi_1 \Psi_{it} + \theta_0 u_{it}$

MDCS-QARMA(p, q):

$\Psi_{it+1} = \varphi_1 \Psi_{it} + \cdots \varphi_p \Psi_{it-p} + \theta_0 u_{it} + \cdots + \theta_q u_{it-q}$

Additive DCS (ADCS) Poisson model

$E(n_{it}|X_{it}) = \lambda_{it} = \Psi_{it} + \exp(Y_{it}\tilde{\theta})$

ADCS-QMA(q):

$\Psi_{it+1} = \theta^* + \theta_0 u_{it} + \cdots + \theta_q u_{it-q}$ where $\theta^* = \theta_0 + \cdots + \theta_q$

ADCS-QAR(1):

$\Psi_{it+1} = \theta_0 + \varphi_1 \Psi_{it} + \theta_0 u_{it}$

ADCS-QARMA(p, q):

$\Psi_{it+1} = \varphi_1 \Psi_{it} + \cdots \varphi_p \Psi_{it-p} + \theta_0 u_{it} + \cdots + \theta_q u_{it-q}$ where $\theta^* = \theta_0 + \cdots + \theta_q$

Diagnostics and results

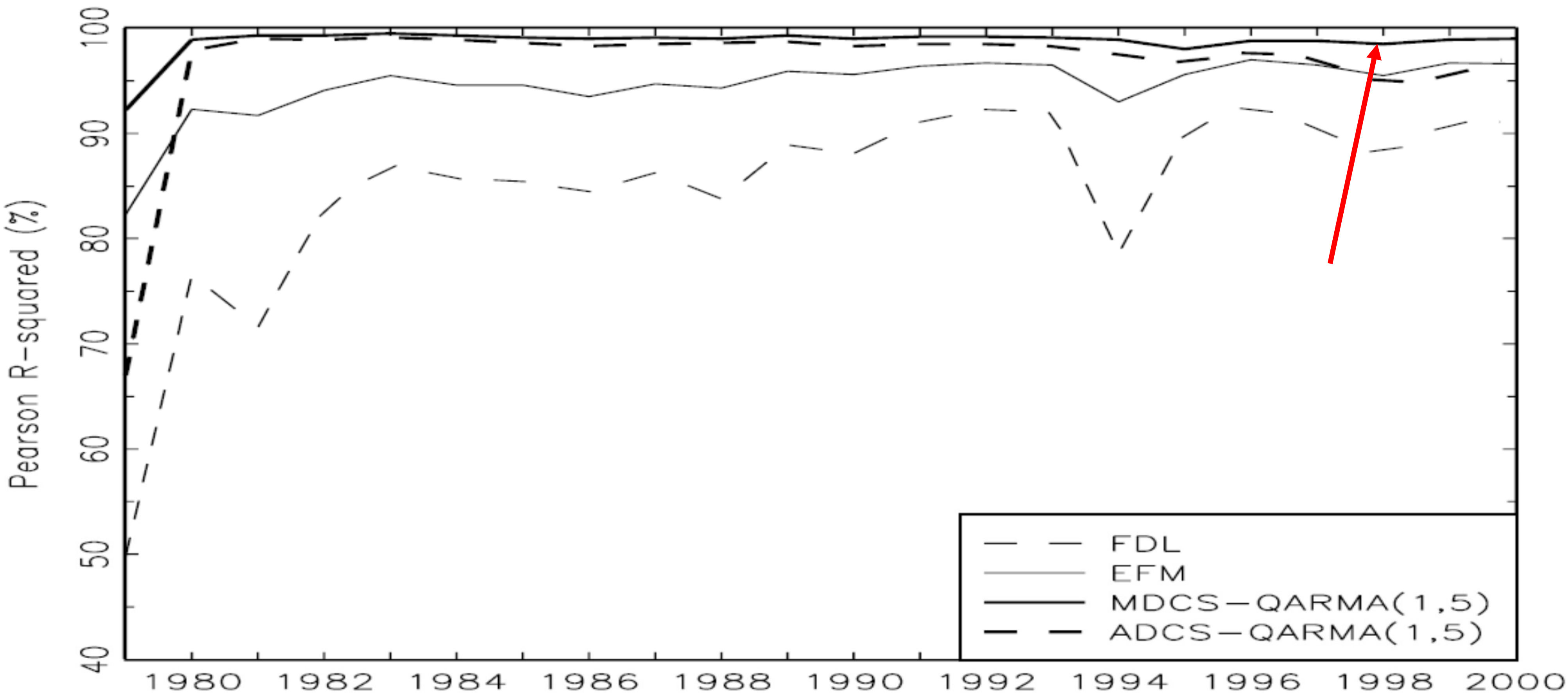
Diagnostics and results

Table 1 and 2 show the two-step negative binomial QMLE estimation results. (See Tables 1 and 2.)

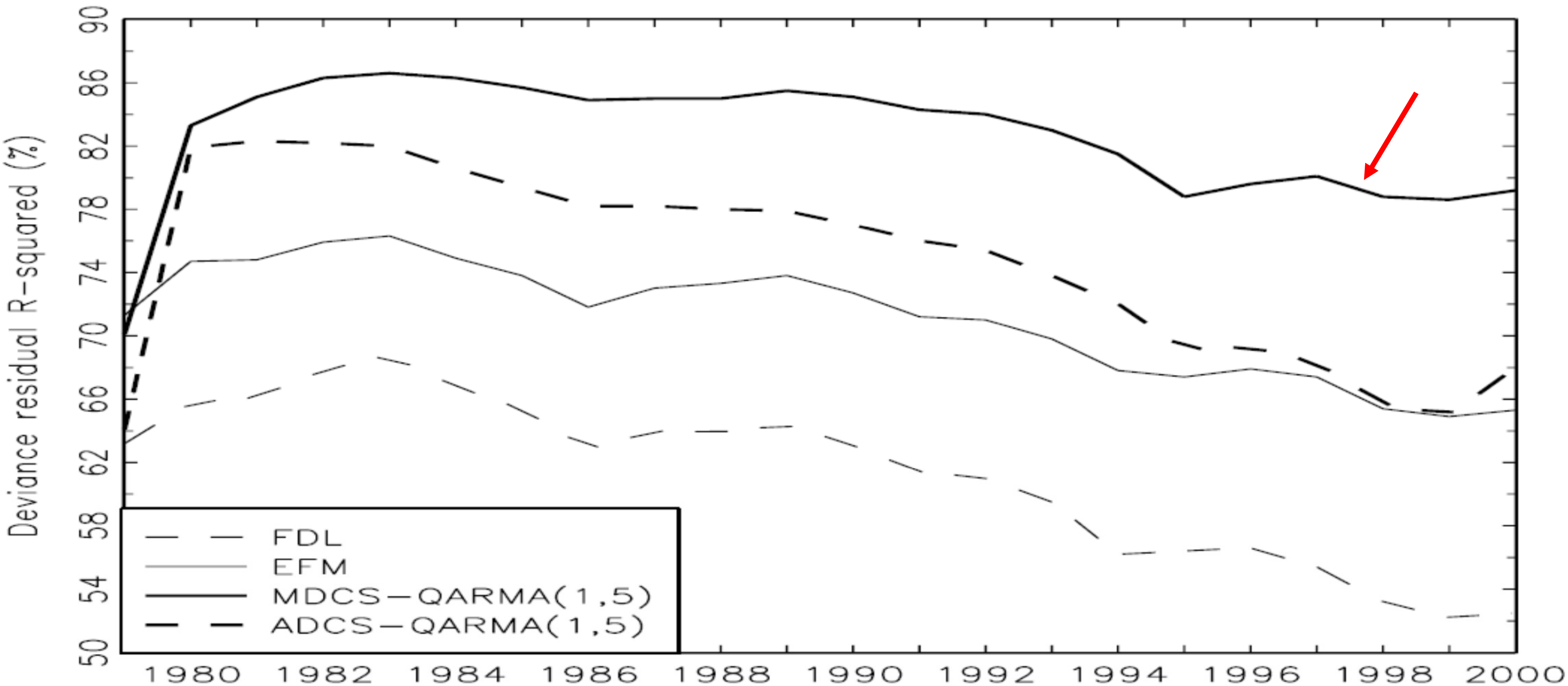
Table 3 Panels A and B, shows test results on contemporaneous exogeneity of R&D expenses; needed for the pooled negative binomial QMLE.

Table 3 Panels C and D, shows mean R-squared model fit metrics for the panel (Cameron and Windmeijer, 1996). An R-squared is computed for each year, then averaged over $t = 1, \dots, T$

Pearson R-squared over $t = 1, \ldots, T$ for models where contemporaneous R&D is exogenous



Deviance residual R-squared over $t = 1, \ldots, T$ for models where contemporaneous R&D is exogenous



References

References

Blazsek S, Escribano A. 2010. Knowledge spillovers in U.S. patents: A dynamic patent intensity model with secret common innovation factors. *Journal of Econometrics* **159**: 14-32.

Blazsek S, Escribano A. 2016. Patent propensity, R&D and market competition: Dynamic spillovers of innovation leaders and followers. *Journal of Econometrics* **191**: 145-163.

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

Cameron AC, Windmeijer FAG. 1996. R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics* **14**: 209-220.

References

Chamberlain G. 1992. Efficiency bounds for semiparametric regression. *Econometrica* **60**: 567-596.

Creal D, Koopman SJ, Lucas A. 2008. A general framework for observation driven time-varying parameter models. Tinbergen Institute Discussion Paper, TI 2008-108/4, Amsterdam.

Davis R, Dunsmuir W, Streett S. 2003. Observation-driven models for Poisson counts. *Biometrika* **90**: 777-790.

Davis R, Dunsmuir W, Streett S. 2005. Maximum likelihood estimation for an observation driven model for Poisson counts. *Methodology & Computing in Applied Probability* **7**: 149-159.

References

Gourieroux C, Monfort A, Trognon A. 1984a. Pseudo maximum likelihood methods: Theory. *Econometrica* **52**: 681-700.

Gourieroux C, Monfort A, Trognon A. 1984b. Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica* **52**: 701-720.

Hansen LP. 1982. Large sample properties of generalized method of moments estimators. Econometrica 50: 1029-1054.

Harvey AC. 2013. *Dynamic Models for Volatility and Heavy Tails*. Cambridge University Press: Cambridge, UK.

Harvey, A. C. and Chakravarty, T. (2008) Beta-t-(E)GARCH, Cambridge Working Papers in Economics 0840, Faculty of Economics, University of Cambridge, Cambridge.

References

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.

Wooldridge JM. 1997a. Quasi-likelihood methods for count data. In *Handbook of Applied Econometrics*, Pesaran MH, Schmidt P (eds). Volume 2, Blackwell, 352-406: Oxford.

Wooldridge JM. 1997b. Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory* **13**: 667-678.

Wooldridge JM. 2002. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press: Cambridge, MA.

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

Thank you for your attention!

SBLAZSEK@UFM.EDU