ANALYSIS OF PANEL PATENT DATA USING POISSON, NEGATIVE BINOMIAL AND GMM ESTIMATION

By

Qi Hu

B.E. Dalian University of Technology, 1996

M.Mgmt. Dalian University of Technology, 2000

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE IN MASTERS OF ARTS

In The Department Of Economics

SIMON FRASER UNIVERSITY

July 2002

APPROVAL

Name: **Qi Hu**

Degree: **M. A. (Economics)**

Title of Project: **Analysis Of Panel Patent Data Using Poisson, Negative Binomial And GMM Estimation**

Examining Committee:

Chair: **Brian Krauth**

Peter Kennedy
Senior Supervisor

Terry Heaps
Supervisor

Ken Kasa
Internal Examiner

Date Approved: Friday July 26, 2002

Title of Project **Analysis Of Panel Patent Data Using Poisson, Negative Binomial And GMM Estimation**

Author:

Qi *Hu*

ABSTRACT

The relationship between patent applications and R&D expenditure is a prominent example of a count data model with panel data. This paper examines the patents-R&D relationship by estimating three different panel count data models. The data includes 346 firms from 1970 to 1979. First is the Poisson model. Since we have panel data, fixed effects and random effects models are developed to allow individual firms to have their own average propensity to patent. Then a more commonly used model, the negative binomial model, is estimated to allow for overdispersion which typically characterizes data of this form. The results from these two models are compared using an overdispersion test and a Hausman test. The fixed effect negative binomial model is found to be the superior model. These two models assume strict exogeneity of the explanatory variables. Estimation via GMM is undertaken to address this problem.

DEDICATION

DEDICATED TO MY FAMILY

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude to my senior supervisor, Professor Peter Kennedy for his invaluable guidance and encouragement throughout the process of my graduate work. I also thank Professor Terry Heaps, from whom I learned much about how to write this project in a professional manner. Many thanks to Dr. Xiannian Sun for his many suggestions on the project. I am also grateful to the faculty and staff at the SFU Economics Department, for their assistance during my graduate studies.

TABLE OF CONTENTS

LISTS OF TABLES AND FIGURES

I INTRODUCTION

The main focus of this paper is to apply different panel count data models to conduct an empirical study on the relationship between Patent applications and R&D expenditures, and to find a better model. For the period of 1970 to 1979 there are 346 firms whose annual data are used for this study. A simple "Poisson model" is employed first given the non-negative and discrete nature of the number of patent applications. Since we have panel data, fixed and random effect Poisson models are applied as well. In addition, we applied the negative binomial model in order to offset the strong assumptions of the Poisson models. Our finding was that the negative binomial model with fixed effects outperforms the other models. To avoid some assumptions that may fail in the context, an empirical analysis by the generalized method of moments (GMM) is also carried out in the last part.

The rest of the paper is organized as follows: Section II reviews the study of patent-R&D expenditure relationship in the open literature. Section III provides a brief review of the assumptions of the basic Poisson model and applies this model to the data in order to obtain a benchmark result. Data description is also included in this part. Section IV takes into consideration the nature of the panel data. Fixed and random effect models are applied to allow each individual firm to have its own average propensity to patent. Section V tests for 'overdispersion' in the data as usually the assumption of equal mean and variance may be unreasonable. By letting each firm have its own individual Poisson parameter, we find that the negative binomial model is more suitable than the panel Poisson model. A Hausman test is conducted in Section VI to decide which model, the fixed effects or random effects model, is more appropriate. Consistency of the estimators of both Poisson and Negative Binomial models requires strict exogeneity of the explanatory variables; however, this assumption is likely to fail in the patent-R&D relationship. Therefore, an estimation procedure by the GMM is undertaken in Section VII. Finally, in Section VIII we draw up a conclusion.

II LITERATURE REVIEW

Pakes and Griliches (1984) present the first empirical work designed to study the patent-R&D relationship. It focuses mainly on the degree of correlation between the number of patent applications and R&D expenditures over a period of 6 years, as well as the lag structure relationship. The data used is from 121 U.S. firms over an 8-year period (1968~1975). Several different types of linear functional forms are examined from which a linear log-log functional form is then chosen as best. The results show that the relationship is very strong because there is a high fit (R^2=0.9) in the cross-sectional model and weaker but still significant (R^2=0.3) for the within-firm time-series dimension. The effects of the previous R&D expenditures are also significant but relatively small and the structure is not well defined, which indicates a lag-truncation bias. However this could not be distinguished from a fixed firm effect. A negative time trend is also found in the whole data set as well as in the subsets.

Hausman, Hall and Grilliches (1984) continue the work of Pakes and Griliches (1984) and develop models allowing for the fact that the patents data consists of nonnegative integers, in the context of panel data. A data set of 128 firms for a period of 7 years (1968-1974) is used. In this paper, the simple Poisson regression model is applied first. Given the nature of the panel data set, generalized Poisson models with fixed effects and random effects are developed which allow each firm to have its own average propensity to patent. A Hausman test is then carried out to examine whether the fixed effect model or the random effect model is more appropriate. To allow for 'overdispersion' into the data, the negative binomial model is also used, and the fixed effects and the random effects versions of this model are also developed. In addition to the current R&D expenditures, various other explanatory variables are used (none vs. 5 lagged R&D terms, a size variable, a science sector dummy variable and an R&D-time interaction variable). The current R&D in all the specifications exhibits a significant effect, but less so than that of the linear model as given by Pakes and Griliches's. Adding the firm's specific effects reduces the coefficients by a small amount. Compared by their loglikelihood functions, negative binomial models fit better than the ordinary Poisson models. The total coefficient of R&D expenditures of the different Poisson models falls to between 0.35 and 0.59 whereas for the negative binomial models it lies between 0.37 and 0.60, which is slightly higher than that of the Poisson model.

Hall, Griliches and Hausman(1986) analyze two larger data sets to try to capture the lag structure of the patent-R&D relationship. The first has observations for 642 firms over an 8-year period (1972~1979). This data set includes almost all of the manufacturing firms who reported R&D expenditures. First the properties of the R&D expenditures are examined and strong evidence of a low order AR process is found. Then some basic estimations are carried out, with the current R&D, the three lagged R&Ds, the size variable and the science sector dummy as the explanatory variables. Different empirical methods, nonlinear least squares, Poisson, Negative binomial and quasi-generalized pseudo maximum likelihood methods are compared. All the results are qualitatively the same, showing a strong contemporaneous relationship. In order to distinguish whether the relatively large coefficient on the last lag is due to the correlation of the last R&D variable with the R&D expenditures of earlier period, or whether it might be caused by correlated fixed effects, data for a subset of the firms is used, for their is a longer time period (346 firms for 10 years). But due to the large standard error, the results are fairly unstable so there is still not a clear-cut conclusion about the effect of the long run R&D expenditures on patents.

However, for the patent-R&D relationship, the assumption of strict exogeneity, that is the current and lagged values of dependent variables do not explain the explanatory variables, is likely to fail since it is possible that patents may generate additional R&D expenditures to further develop or improve the embodied innovation. However if this assumption is violated, the estimation is no longer consistent. Therefore, Montalvo(1997) gets around the problem of inconsistency by employing the GMM estimation to examine the second data set used by Hall, et al (1986). The explanatory variables include the current and five lagged R&D expenditures and a time trend. Compared with the estimators of conditional maximum likelihood and pseudo maximum likelihood, the contemporaneous R&D and the first lag of the R&D are significantly positive; the total effect of the R&D on patents is larger than the alternative fixed-effect panel-data estimators, yet as well the standard errors are larger.

Crepon and Duguet (1997) used the GMM to estimate a patent-R&D relationship with fixed effects for European data. The data set used consisted of 698 French manufacturing firms over a 6-year period (1984~1989). Two mothods are used. The first uses the lagged explanatory variables and an intercept; the second set includes the first set plus all the independent cross-products of the previous explanatory variables. Results show that the estimates obtained from the two methods are similar. Also there is an efficiency gain when using the second set of instruments. When considering the fixed effects, the estimated return to the R&D is approximately 0.3. Moreover, introducing restrictions on the serial correlations as well as allowing for weak exogeneity of the R&D do not alter the results. For the rest of the paper, a dynamic effect is examined by including dummy variables for firms that have applied for at least one patent in the previous year. The finding is that for small positive numbers of the past innovations, there are positive effects on the production of innovation. However, this effect slowly vanishes as the number of innovations increases.

III BASIC POISSON REGRESSION MODEL

3.1 The basic Poisson regression model

The basic Poisson model is

$$\Pr(y_{it}) = f(y_{it}) = \frac{e^{-\lambda_{it}}\lambda_{it}^{y_{it}}}{y_{it}!}, \tag{1}$$

where i indexes firms and t indexes years and $\log\lambda_{it} = x_{it}^{\prime}\beta$, x_{it} is a vector of m regressors for unit i at time t.

This basic Poisson Model embodies some strong assumptions, e.g.

$E(y_{it} \mid x_{it}) = \lambda_{it} = V(y_{it} \mid x_{it})$.

In this model, it is also assumed that all observations occurred randomly and independently across firms and across time.

3.2 Data Description

For the period of 1970 to 1979 there are 346 firms in the data set. For each firm there are ten years of data on patents and logR&D expenditures (this data was obtained from Bronwyn H. Hall's website). In this paper, the specification used is chosen from one of the specifications in Hall, et al(1986)'s paper[1]. The R&D expenditures of the current year and the five previous years are used as the explanatory variables. In order to account for the differences in propensity to patent across these firms, a dummy variable for the scientific sector is added; to proxy the firm size, the book value of the firms in 1971 is used.

Table 1 and Table 2 show the definition and the descriptive statistics of the variables. The histogram of the number of patents per million dollar R&D expenditure can be found in appendix I.

Two specifications are used in this paper. So for the basic Poisson model,

$$\Pr(PAT_{it}) = f(PAT_{it}) = \frac{e^{-\lambda_{it}} \lambda_{it}^{PAT_{it}}}{PAT_{it}!}$$

where $\log \lambda_{it} = \alpha + \beta_0 LOGR_{it} + \sum_{M=1}^{5} \beta_m LOGR_{it,m} + \varphi_n \sum_{N=2}^{5} DYEAR_n$.

Or

$$\log \lambda_{it} = \alpha + \beta_0 LOGR_{it} + \sum_{M=1}^{5} \beta_m LOGR_{it,m} + \varphi_n \sum_{N=2}^{5} DYEAR_n + \delta SCISECT_i + \gamma LOGK_i$$

[1] Inexplicably the results in Hall(1986) could not be replicated.

Table 1 Definition of Variables

NAME	DESCRIPTION
PAT	The number of patents applied for during the year that were eventually granted
LOGR,	The logarithm of the R&D spending during the current year, (in 1972 dollars)
LOGR1,…,LOGR5	The logarithm of the R&D spending during the previous one year…the previous five years respectively (in 1972 dollars)
DYEAR2,...,DYEAR5	Year dummies, in ascending order from 1976. eg. DYEAR2=1 if it is 1976, =0 otherwise;… DYEAR5=1 if it is 1979, =0 otherwise
SCISECT	The dummy variable, equal to one for firms in the scientific sector
LOGK	The logarithm of the book value of capital in 1972

Table 2 Descriptive Statistics of Variables

Variable	Mean	Std.Dev	Skew.	Kurt.	Minimum	Maximum	Cases
PAT	34.772	70.875	3.455	17.109	0.000	515.000	1730
YEAR	3.000	1.415	0.000	1.699	1.000	5.000	1730
SCISECT	0.425	0.494	0.304	1.092	0.000	1.000	1730
LOGK	3.921	2.093	0.124	2.644	-1.770	9.666	1730
LOGR	1.256	2.006	0.170	2.575	-3.849	7.034	1730
LOGR1	1.234	1.984	0.206	2.555	-3.849	7.065	1730
LOGR2	1.219	1.967	0.229	2.562	-3.849	7.065	1730
LOGR3	1.206	1.952	0.244	2.576	-3.849	7.065	1730
LOGR4	1.197	1.942	0.244	2.585	-3.674	7.065	1730
LOGR5	1.204	1.934	0.228	2.596	-3.674	7.065	1730
DYEAR2	0.200	0.400	1.500	3.248	0.000	1.000	1730
DYEAR3	0.200	0.400	1.500	3.248	0.000	1.000	1730
DYEAR4	0.200	0.400	1.500	3.248	0.000	1.000	1730
DYEAR5	0.200	0.400	1.500	3.248	0.000	1.000	1730

3.3 Empirical Results of Basic Poisson Model

The following results for the basic Poisson model (Table 3) are obtained from Limdep 7.0 since it has the build in capability to deal with panel count data estimation.

The current R&D expenditure has a substantial influence on the number of patent applications. When time-invariant variables are included, both LOGK and SCISECT have significant effects on the expected number of patents. But all the LOGRs coefficients display a U-shape, that is, the coefficients of LOGR and LOGR5 are much higher than any of the other years' effects. This contradicts our intuition that the effect of the R&D expenditures decreases over time. The negative coefficients of the year dummies imply that the patent applications from the current year to the years that follow are decreasing in this period. When looking at the raw data, an overall pattern of decreasing numbers of patents can be found over time. One tentative explanation to these negative coefficients is that R&D investment was more risky because of the high inflation during the period, thus firms had less incentive to patent.

Table 3 Estimates of the Basic Poisson Model

Basic Poisson Model						
	Without time-invariant var.			With time-invariant var.		
	Coeff.	Std.Dev.	t-ratio	Coeff.	Std.Dev.	t-ratio
Constant	1.8283	0.0212	38.2239	0.8099	0.0212	38.2239
LOGR	0.1525	0.0307	4.3793	0.1345	0.0307	4.3793
LOGR1	0.0220	0.0428	-1.2371	-0.0529	0.0428	-1.2371
LOGR2	0.0437	0.0398	0.2069	0.0082	0.0398	0.2069
LOGR3	0.0827	0.0370	1.7879	0.0661	0.0370	1.7879
LOGR4	0.1040	0.0333	2.7052	0.0902	0.0333	2.7052
LOGR5	0.3011	0.0225	10.6652	0.2395	0.0225	10.6652
DYEAR2	-0.0440	0.0131	-3.3165	-0.0435	0.0131	-3.3165
DYEAR3	-0.0604	0.0133	-3.9466	-0.0524	0.0133	-3.9466
DYEAR4	-0.1892	0.0135	-12.5828	-0.1702	0.0135	-12.5828
DYEAR5	-0.2298	0.0135	-14.9170	-0.2019	0.0135	-14.9170
LOGK				0.2529	0.0044	57.2831
SCISECT				0.4543	0.0092	49.1552

IV PANEL POISSON REGRESSION MODEL

Unlike the basic count data models, the assumption of independent observations is no longer required to hold in the panel data models. One advantage of the panel data formulation over the cross-sectional data formulation is that it permits more general types of individual heterogeneity. For example, in the estimation of the relationship between the number of patent applications and the R&D expenditures, if a cross-section model is used, the only way to control for heterogeneity is to include firm-specific attributes such as industry, or firm size. When there happen to be other components affecting individual firm-specific propensity to patent, the estimates may become inconsistent. According to the nature of patent application, it is very likely that the differences in technological opportunities or operating skills may affect the observed number of patents. But these firm specific factors are not captured by the explanatory variables in the basic Poisson model. In this part, we estimate the effect of the R&D expenditures on patent applications by individual firms, controlling for individual firm-specific propensity to patent by including a firm-specific term an unobserved firm-specific propensity to patent. For this panel Poisson model, the data is assumed to be independent over individual units for a given year; but it is permitted to be correlated over time for a given individual firm. As in least square panel data model, two models can be used, fixed effects, where separate dummies are included for each individual firm; and random effects, where the individual specific term is drawn from a specified distribution.

4.1 Random Effects Poisson Model

For the Poisson model with intercept heterogeneity, the random effects model as developed by Hausman et al (1984) was:

$$\tilde{\lambda}_{it} = \exp(x_{it}{}' \beta + \varepsilon_i), \qquad i = 1, \ldots, N, t = 1, \ldots, T_i \tag{2}$$

$$= \lambda_{it} u_i$$

Where $u_i (= \exp(\varepsilon_i))$ is the firm specific random variable. x_{it} is a vector of regressors including the overall intercept. $\tilde{\lambda}_{it}$ and $\tilde{\lambda}_{it'}$ $(t \neq t')$ are correlated because of u_i, while $\tilde{\lambda}_{it}$ and $\tilde{\lambda}_{jt'}$ are uncorrelated by the assumption that u_i is independent of $u_j (i \neq j)$.

To estimate the coefficients, u_i is assumed to distribute as a gamma random variable with parameter (δ,δ) (normalized so that the mean is 1, and the variance is 1/δ), so

$$f(y_{i1},\ldots,y_{iT} \mid x_{i1},\ldots,x_{iT}) = \frac{\Gamma(\sum_t y_{it} + \delta)}{\Gamma(\delta)} \left(\frac{\delta}{\delta + \sum_t \lambda_{it}}\right)^{\delta} \times \left(\sum_t \lambda_{it} + \delta\right)^{-\sum_t y_{it}} \prod_t \left(\frac{x_{it}^{y_{it}}}{y_{it}!}\right) \quad (3)$$

Estimated by maximum likelihood estimation, the results are given in table 4.

Table 4 Estimates of Panel Poisson Model with Random Effects

Panel model with random effects						
	Without time-invariant var.			With time-invariant var.		
Variable	coefficient	std.dev	t-ratio	coefficient	std.dev	t-ratio
Constant	1.40286	0.05330	26.31900	0.41079	0.13163	3.12100
LOGR	0.47658	0.02536	18.79200	0.40345	0.02786	14.47900
LOGR1	-0.00771	0.03720	-0.20700	-0.04618	0.03991	-1.15700
LOGR2	0.13641	0.03579	3.81100	0.10792	0.03608	2.99100
LOGR3	0.05919	0.02567	2.30600	0.02977	0.02615	1.13800
LOGR4	0.02752	0.03204	0.85900	0.01070	0.03161	0.33800
LOGR5	0.08255	0.02331	3.54100	0.04061	0.02316	1.75300
DYEAR2	-0.04688	0.01177	-3.98200	-0.04496	0.01205	-3.73200
DYEAR3	-0.05609	0.00858	-6.53500	-0.04839	0.00866	-5.59000
DYEAR4	-0.19031	0.00695	-27.39000	-0.17416	0.00708	-24.61200
DYEAR5	-0.25268	0.00797	-31.70400	-0.22590	0.00815	-27.70900
LOGK				0.29169	0.03096	9.42300
SCISECT				0.25700	0.11776	2.18200
Delta	0.86591	0.06655	13.01100	0.85493	0.06675	12.80900

4.2 Fixed effects Poisson Model

For the Poisson model, the specification of fixed effects is:

$$\tilde{\lambda}_{it} = \exp(d_i + x'_{it}\,\beta) \qquad i = 1,\dots,N, \qquad t = 1,\dots,T_i \tag{4}$$
$$= \alpha_i \lambda_{it}$$

where d_i are firm specific dummies, $\alpha_i (= \exp(d_i))$ are the individual specific effect, x_{it} is a vector of regressors.

In order to estimate the fixed effects model, the conditional maximum likelihood method was developed by Hausman et al(1984). Since y_{it} follows the Poisson distribution, the sum of patent $\sum_t y_{it}$ also follows the Poisson distribution.

The distribution of y_{it} conditional on $\sum_t y_{it}$ follows a multinomial distribution:

$$f(y_{i1},\dots,y_{iT} \mid \sum_t y_{it}) = \frac{(\sum_t y_{it})!}{\prod_t (y_{it}!)} \prod_t \left[\frac{\tilde{\lambda}_{it}}{\sum_t \tilde{\lambda}_{it}}\right]^{y_{it}} \tag{5}$$

Compared with the fixed effects model, one advantage of the random effects model is that it is more efficient when correctly specified, that is, relative to the fixed effects model, it has n additional degrees of freedom. In addition it allows for the use of time-invariant variables, which in the fixed effects model are absorbed into the individual-specific effect α_i, and are not identified. In contrast, the advantages of the fixed effects model are that the population distribution of α_i need not to be specified, which avoids the inconsistency that might happen in the misspecified random effects model. Rather than having to assume that the individual effects are uncorrelated with the other regressors, the estimators of the fixed effects give consistent estimation under all circumstances.

4.3 Empirical Results for the Panel Poisson Model

Table 4 and Table 5 are the results of the estimation of the two models above. According to the random effects model, the coefficients for LOGR(0.47 and 0.40 respectively) are much higher than those for the basic model(0.15 and 0.13 respectively). The firm specific variables, the firm size and the science sector variables still show a strong influence on patent application, which is consistent with our assumptions that large firms or firms in the science sectors have more incentive or are more efficient in getting the R&D output patented. The U-shaped pattern of R&D expenditure coefficients is attenuated. Yet all year dummies are still negative and significant.

Looking at the results from the panel Poisson model with fixed effects, the current R&D expenditure also shows a strong influence on the patent application (0.32) both statistically and economically. However, the first year and fourth year's coefficients are negative with the first year displaying a significantly negative value, which contradicts our assumption that previous R&D expenditures should have a diminishing but positive effect. As for the year dummies, the results are similar to the results of random effect model.

Table 5 Estimates of Panel Poisson Model with Fixed Effects

Panel Poisson model with fixed effects			
Variable	coefficient	std.dev	t-ratio
LOGR	0.32221	0.02846	11.32300
LOGR1	-0.08713	0.04037	-2.15800
LOGR2	0.07858	0.03624	2.16900
LOGR3	0.00106	0.02621	0.04000
LOGR4	-0.00464	0.03078	-0.15100
LOGR5	0.00261	0.02297	0.11300
DYEAR2	-0.04261	0.01217	-3.50200
DYEAR3	-0.04005	0.00864	-4.63400
DYEAR4	-0.15712	0.00744	-21.12300
DYEAR5	-0.19803	0.00832	-23.79700

V NEGATIVE BINOMIAL MODEL AND OVERDISPERSION TEST

The Poisson model has the strong restriction that the variance and mean are equal. However, this assumption is often violated in the real count data sets, that is, the data is overdispersed. Overdispersion occurs when the conditional variance exceeds the conditional mean. This may be caused by unobserved individual heterogeneity, or excessive zeros in the count data, which is quite common in the real world. When we examine the statistics of the data, 'PAT' varies from 0 to 515, with distribution skewed to the left together with a long right tail. This is a common feature of overdispersion, which shifts the mean towards the origin.

Variable	Mean	Std.Dev	Skew.	Kurt.	Minimum	Maximum	Cases
PAT	34.772	70.875	3.455	17.109	0.000	515.000	1730

If this violation is true then the Poisson model would be inappropriate. It may lead to very erroneous and overly optimistic conclusions concerning the statistical significance of the regressors. To deal with overdispersion, a distribution that permits more flexible modeling of the variance than the Poisson model should be used, the negative binomial distribution is such a distribution.

5.1 PANEL NEGATIVE BINOMIAL MODEL

The negative binomial model allows each firm's Poisson parameter to have its own random distribution. Like in the panel Poisson model, two different models, the fixed effects and random effects model are developed and was done by Hausman et al(1984).

5.1.1 Negative binomial model with fixed effects

When we add firm specific effects to the negative binomial model, we get:

$$f(y_{it}) = \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it}+1)} \left(\frac{1}{1+\theta_i}\right)^{\lambda_{it}} \left(\frac{\theta_i}{1+\theta_i}\right)^{y_{it}} \tag{6}$$

with $E(y_{it} \mid \theta_i) = \lambda_{it}\theta_i$

and $Var(y_{it} \mid \theta_i) = \lambda_{it}(\theta_i + \theta_i^2)$

therefore,

$$f(y_{it},\dots,y_{iT} \mid \sum_t y_{it}) = \frac{\Gamma(\sum \lambda_{it})\Gamma(\sum_t y_{it}+1)}{\Gamma(\sum_t \lambda_{it} + \sum_t y_{it})} \prod_t \frac{\Gamma(\lambda_{it}+y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it}+1)} \qquad (7)$$

all terms involving θ_i are cancelled out.

5.1.2 Negative Binomial random effects model

For the random effects model, starting from equation (6), by assuming $1/(1+\ \theta_i)$ is distributed as beta(a,b), the joint density for the i[th] individual is :

$$f(y_{it},\dots y_{iT} \mid x_{iq},\dots x_{iT}) = \frac{\Gamma(a+b)\Gamma(a+\sum_t \lambda_{it})\Gamma(b+\sum_t y_{it})}{\Gamma(a\ )\Gamma(b)\Gamma(a+b+\sum_t \lambda_{it} + \sum_t y_{it})} \prod_t \frac{\Gamma(\lambda_{it}+y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it}+1)} \qquad (8)$$

5.1.3 Empirical results of the panel negative binomial model

The results of this negative binomial model can be found in table 6. The current R&D expenditure shows a strong positive influence. All coefficients of the previous years' LOGR expenditure are positive and decreasing over time, which coincides with our expectations, yet all of them are insignificant. For the fixed effects model, the first two-year dummies are positive, but the latter two years are negative, which can be explained that sometimes the number of the patent applications varies over years in a pattern similar to a wave. We can also notice that the standard error of these models are greater than the counterparts in the panel Poisson model, this is because the negative binomial model allows for an additional source of variance.

5.2 Test for Overdispersion

There are several methods of testing for overdispersion. For example, there are regression-based tests, the Wald test and the LR test. In view of the fact that the log-likelihood function of both the panel Poisson model and the negative binomial model can

Table 6 Estimates of Negative Binomial Model

Negative binomial model with fixed and random effects						
	Random effects			Fixed effects		
Variable	Coefficient	std.dev	t-ratio	Coefficient	std.dev	t-ratio
Constant	1.39944	0.09081	15.41100			
LOGR	0.38762	0.06982	5.55200	0.36324	0.11511	3.15600
LOGR1	0.01833	0.10030	0.18300	0.15553	0.15503	1.00300
LOGR2	0.12909	0.09604	1.34400	0.17401	0.15493	1.12300
LOGR3	0.02469	0.07347	0.33600	0.01485	0.11750	0.12600
LOGR4	0.05775	0.07785	0.74200	0.02876	0.11681	0.24600
LOGR5	0.10003	0.05446	1.83700	0.13606	0.07877	1.72700
DYEAR2	-0.04671	0.03092	-1.51100	0.02646	0.04276	0.61900
DYEAR3	-0.06512	0.02360	-2.75900	0.00211	0.03553	0.05900
DYEAR4	-0.20006	0.02115	-9.45700	-0.13782	0.03218	-4.28300
DYEAR5	-0.25096	0.02083	-12.04800	-0.19296	0.03202	-6.02700
a	2.64396	0.28263	9.35500			
b	2.02183	0.19617	10.30600			

be obtained without difficulty, the LR test is used to test for overdispersion (Cameron, 1998). Here we test the results of the models with fixed effects.

H_0: $E(y_{it}) = Var(y_{it})$, which means the Negative Binomial model reduces to the Poisson;

H_1: $E(y_{it}) < Var(y_{it})$, which implies overdispersion.

$LR = -2(LLFr - LLFu)$,

where $LLFr$ is the log-likelihood function of the Poisson model, $LLFu$ is the log-likelihood function of the negative binomial model,

It follows the χ^2 distribution. Since there is only one constraint, the degree of freedom is one.

In this sample, $LLFr$ is –3536, $LLFu$ is –3391,

LR	$\chi^2_{0.01}(1)$
290	6.63

As one can see there is strong evidence of overdispersion, thereby the negative binomial model is more suitable for this sample.

VI HAUSMAN TEST

Although the random effects model can include time-invariant variables and is more efficient when compared to the fixed effects model, it is consistent only when it is correctly specified. So it is necessary to test which one is better for our data. The Hausman test is used on the estimates of the negative binomial model (Hausman, 1984). Under the null hypothesis, the random effects model is correctly specified, so both the fixed and the random effects models are consistent, while under the alternative hypothesis, the random effects are correlated with the regressors, so the random effects model loses its consistency.

Thus, the Hausman test:

H_0: The random effects model is appropriate. The preferred estimator is $\hat{\beta}_{RE}$.

H_1: The fixed effects model is appropriate. The preferred estimator is $\hat{\beta}_{FE}$.

The Hausman test is based on the distance:

$$T_H = (\hat{\beta}_{RE} - \hat{\beta}_{FE})'[V(\hat{\beta}_{FE}) - V(\hat{\beta}_{RE})]^{-1}(\hat{\beta}_{RE} - \hat{\beta}_{FE})$$

It follows that the χ^2_α distribution, with a degree of freedom of q (the dimension of β_{FE}).

From the result,

T_H	$\chi^2_{0.05}(10)$	$\chi^2_{0.01}(10)$
109.2	18.31	23.21

So now we can reject the null hypothesis, and accept that the fixed effects model is more appropriate, since there is strong evidence that the firm specific effects in the random effects model are correlated with the explanatory variables.

VII GMM ESTIMATION

GMM estimation is a very general estimation method used in econometrics. In some sense, maximum likelihood and quasi-maximum likelihood can be considered as special cases of GMM. The generalized method of moments (GMM) has become popular in recent years since it requires fewer assumptions. The consistency of the estimators from the Poisson and the Negative Binomial models in this panel count data model requires the explanatory variables be strictly exogenous. But in the patent-R&D relationship, this assumption is likely to fail since once a firm succeeds in one patent application, it is likely that more R&D expenditures will be needed for full development or improvement of products embodying the patent. Consequently R&D expenditures should not be considered as strictly exogenous. Moreover, the GMM estimation procedure is not based on assumptions about the distribution of the error terms. In this part of the paper, the GMM method is applied to the fixed effects model and the results are compared to those obtained from the previous sections[2].

7.1 Introduction to GMM Method

One approach to get moment conditions for the GMM estimation is to find a set of instrumental variables (z_i) and residuals (ε_i) to satisfy the orthogonality conditions:

$$E[z_i \varepsilon_i] = 0$$

Since ε_i is a function of the parameters, the estimators can be obtained by solving these moment conditions. If the number of moment conditions (L) equals the number of parameters (K), it is identified; if $L>K$, it is overidentified.

Then the sample moment conditions will be:

$$m(\beta) = \sum_{i=1}^{n} m_i(\beta) = Z'e(X, \beta)$$

In case of overidentification, the GMM estimator $\hat{\beta}_G$ will be obtained by minimizing

$F = m(\beta)' W^{-1} m(\beta)$, where W^{-1} is the weight matrix.

---

[2] For a detailed introduction regarding GMM, please refer to Kennedy(2000).

7.1.1 Defining the residuals

The transformation method, developed by Chamberlain (1992), is used here to eliminate the unobservable effects in the context of sequential moment restrictions for models with multiplicative fixed effects. The transformation is as follows:

For the fixed effect model,

$$y_{it} = \exp(d_i + x_{it}'\beta) + u_{it} \qquad i = 1,...,N, t = 1,...,T_i \tag{9}$$

under weak exogeneity, u_{it}

$$E(u_{it} \mid x_{i1},...,x_{it},\alpha_i) = 0 \tag{10}$$

In order to eliminate the unobserved firm specific effects, rewrite (10) for t+1, and solving for α_i, then substitute back into (9), we obtain:

$$\begin{aligned} y_{it} &= \exp(x_{it}'\beta)\left(\frac{y_{it+1} - u_{it+1}}{\exp(x_{it+1}'\beta)}\right) + u_{it} \\ &= y_{it+1}\exp[(x_{it} - x_{it+1})'\beta] + v_{it} \end{aligned}$$

where

$$\begin{aligned} v_{it} &= u_{it} - u_{it+1}\exp[(x_{it} - x_{it+1})'\beta] \\ &= y_{it} - y_{it+1}\exp[(x_{it} - x_{it+1})\beta] \end{aligned} \tag{11}$$

The conditional expectation of (11) is

$$\begin{aligned} E(v_{it} \mid x_{i1},...,x_{it}) &= E[u_{it} - u_{it+1}\exp[(x_{it} - x_{it+1})'\beta] \mid x_{i1},...,x_{it}] \\ &= -E[E[u_{it+1} \mid x_{i1},...,x_{it}]\exp[(x_{i1},...,x_{it})'\beta] \mid x_{i1},...,x_{it}] \\ &= 0 \end{aligned}$$

So v_{it} is uncorrelated with past and current values of x.

Define:

$$v_i(\beta) = \begin{bmatrix} v_{i1}(\beta) \\ v_{i2}(\beta) \\ \dots \\ v_{iT-1}(\beta) \end{bmatrix} \tag{12}$$

This is the residual we are going to use.

7.1.2 Minimization criterion

Thus the GMM estimator of β can be obtained by minimizing

$$F = \left[\sum_{i=1}^{n} \mathbf{v}_i^{'}\mathbf{Z}_i\right]\hat{\mathbf{W}}_n^{-1}\left[\sum_{i=1}^{n} \mathbf{Z}_i^{'}\mathbf{v}_i\right] \tag{13}$$

where $\hat{\mathbf{W}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i^T \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T \mathbf{Z}_i$, $\hat{v}_i$ is $\hat{v}$ estimated using some predetermined $\hat{\beta}$.

The matrix of instruments Z_i of this panel count model should look like:

$$Z_i = \begin{bmatrix} z_{i1} & 0 & \dots & 0 \\ 0 & z_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & z_{iT-1} \end{bmatrix}$$

where z_{it} is some function of the x variables.

Under mild regularity conditions, $\hat{\beta}$ is consistent and normally distributed with the asymptotic variance-covariance matrix:

$$\hat{V}(\hat{\beta}) = \frac{1}{N}\left[D(\hat{\beta})'\Omega(\hat{\beta})^{-1} D(\hat{\beta})\right]^{-1} \tag{14}$$

where $D(\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N} Z_i'\nabla v_i(\hat{\beta})$ (15)

and $\Omega(\hat{\beta}) = W(\hat{\beta})$

7.2 Empirical Result Using GMM Estimation

In this paper, we have n=346, T=5, so the residual $v_i(\beta)$ is:

$$\mathbf{v}_i = \begin{bmatrix} y_{i1} - y_{i2}[(x_{i1} - x_{i2})'\beta] \\ y_{i2} - y_{i3}[(x_{i2} - x_{i3})'\beta] \\ y_{i3} - y_{i4}[(x_{i3} - x_{i4})'\beta]. \\ y_{i4} - y_{i5}[(x_{i4} - x_{i5})'\beta] \end{bmatrix}_{4*1}$$

In order to choose the instrumental variables, we need to check the moment conditions. The orthogonalitiy conditions of the year dummies with the residuals are same as the moment conditions of the residual itself, so we can eliminate the redundancy. Since the residual is uncorrelated with past and present value of the R&D expenditures, z_{it} includes the intercept and all previous year's *LOGR*. So there are 34 (7+8+9+10) moment conditions, which are contrary to the case of strict exogeneity, where a common set of instruments are used. Here the number of instruments increases with the number of periods.

The estimates obtained from the negative binomial model with fixed effects are used as the starting values. Compared with the results of the Poisson Model, the contemporaneous R&D still has a significant effect on the patent application and the first year's lagged R&D is no longer negative. The total effect of R&D on patents is larger than that of the Poisson model, but not as large as the one derived from the negative binomial model. The standard errors are also larger than that of the Poisson model, indicating less efficiency in estimation. However, the last lag is negative and significant and all coefficients of the year dummies are negative.

Table 7 The Result of The GMM Estimation

GMM Estimation			
Variable	coefficient	std.dev	t-ratio
LOGR	0.35437	0.15521	2.28312
LOGR1	0.06183	0.05830	1.06059
LOGR2	0.11762	0.05307	2.21617
LOGR3	0.08297	0.05220	1.58923
LOGR4	-0.06742	0.04851	-1.38984
LOGR5	-0.12877	0.04611	-2.79291
DYEAR2	-0.02976	0.01372	-2.16939
DYEAR3	-0.02705	0.01879	-1.43921
DYEAR4	-0.12712	0.02770	-4.58974
DYEAR5	-0.17532	0.03572	-4.90832

7.3 Overidentification Test

Another advantage of the GMM is that it provides an easy specification test for the validity of the overidentifying restrictions. If it is exactly identified, the criterion for the GMM estimation (F) should be zero since we can find the estimates to exactly satisfy the moment conditions. When in the situation of overidentificaiton, the moment conditions imply substantive restrictions. So if the model we apply to derive the moment condition is incorrect, some of the sample moment conditions will be systematically violated. Following Winkelmann(2000), the overidentification test:

H_0: $E[v_i^{'}Z_i] = 0$, that is the restrictions are valid.

H_1: The specification is invalid.

The overidentification test is based on the minimum criterion function evaluated at $\beta = \hat{\beta}$ and divided by the sample size, that is: $F_t = F/n$, and the test statistic follows the χ^2_α distribution, with a degree of freedom of *L-K*, where *L* is the number of moment conditions (34 in this case) and *K* is the number of parameters (10) in this case.

F_t	$\chi^2_{0.05}(24)$	$\chi^2_{0.01}(24)$
33.132	36.415	42.980

So the result shows we should accept the null hypothesis that all the moment conditions are valid at a 5% significant level.

VIII CONCLUSION

Several models were applied to study the relationship between patent applications and the R&D expenditures. Firstly we used the Poisson model with cross-sectional data, which overlooks the panel nature of the data and controls for heterogeneity only by including firm specific effects such as firm size or as to whether it belongs to the science sector. The result is then used for comparison. Then the panel Poisson models with unobservable fixed effects and random effects are used to control for firm specific effects. One of the assumptions of the Poisson model is that the mean and the variance should be equal, which does not usually hold true. From the data description, we also find that the distribution of the patent application skews to the left together with a long right tail, which indicates that there might be overdispersion. So the alternative negative binomial models are applied for overdispersion. Finally the GMM estimation is carried out with the relaxed assumption of strict exogeneity, since additional R&D expenditures is very likely to happen when the firm succeeds in a new patent.

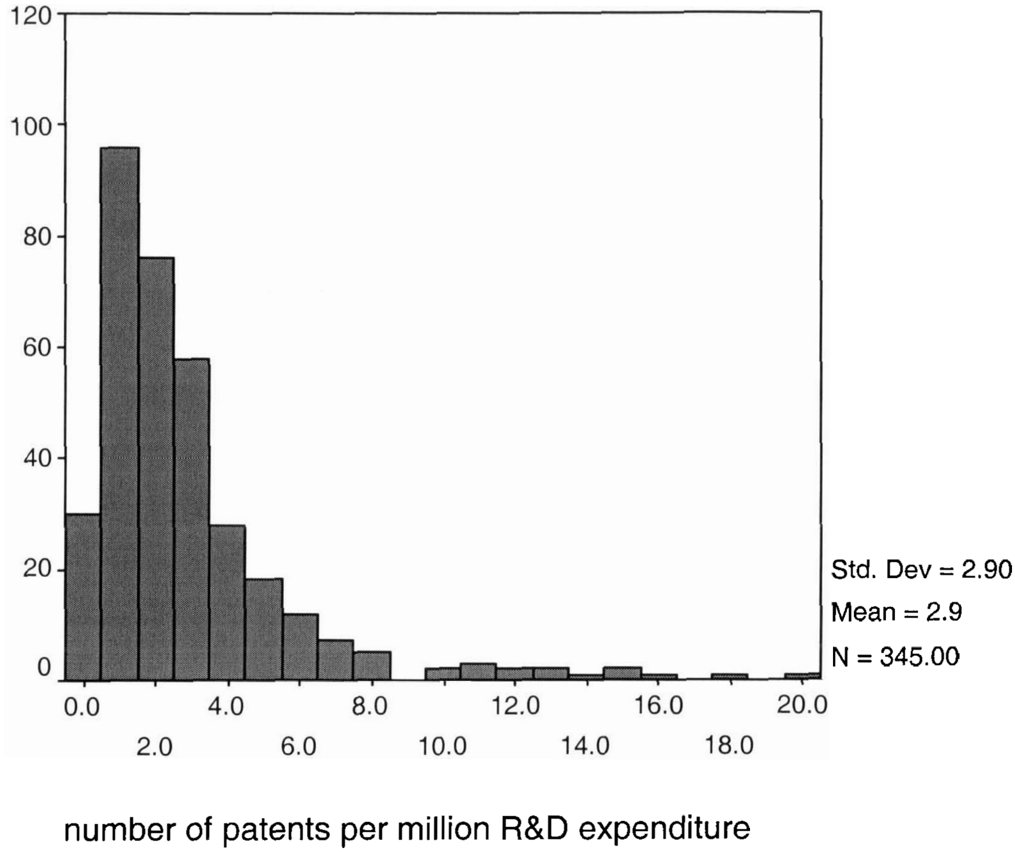
Table 8 is the summary of estimators from different models with only time-invariant variables. Two major results are compared. First is the coefficient of the current year's R&D expenditure. Since β_0's are from different generalizations of the Poisson models, they have the same interpretation and are found to be significantly important in all the models. The second result that is compared is the sum of all the coefficients of the current R&D expenditure and previous years' R&D expenditures. It is approximately equal to the product of the coefficients of R&D expenditures which is the long run effect of R&D expenditures on patent applications. When comparing the results of different models, I prefer the Negative Binomial model with fixed effects. According to the overdispersion test and Hausman test, the Negative Binomial model with fixed effects is more appropriate compared to other Poisson Models and Negative Binomial models. The result of the estimation, that is, all coefficients of R&D expenditures are positive and the long run effect is the highest, is in line with the assumption of the R&D-patent relationship. Though theoretically only the estimates from the GMM estimation is consistent, it is less efficient. The result is not very stable because of the large standard error. The only thing we can conclude from the GMM estimation is that the contemporaneous R&D expenditure has important effect on the number of patent applications.

Table 8 Comparison of Alternative Estimatiors

	β_0	t-ratio	$\sum_{i=0}^{5}\beta_i$
Basic Poisson Model	0.15	4.3793	0.69
Panel Poisson with fixed effects	0.32	11.323	0.31
Panel Poisson with random effects	0.47	18.792	0.75
Negative Binomial with fixed effects	0.36	3.156	0.84
Negative Binomial with random effects	0.38	5.552	0.69
GMM	0.35	2.283	0.42

APPENDIX I

Figure 1 Number of Patents Per Million R&D Expenditure



REFERENCES

Cameron, A.C, Trivedi, P.K. (1998), *Regression analysis of count data* **.** New York : Cambridge University Press.

Crepon, B. and Duguet, E., "Estimating the Innovation Function From Patent Numbers: GMM on Count Panel Data" *Journal of Applied Econometrics* 12(1997), 243-263.

Chamberlain, G. (1992), "Comment: Sequential Moment Restrictions in Panel Data," *Journal of Business and Economic Statistics* 10:20-26.

Greene H. William, *LIMDEP User's manual and reference guide,* version 7.0.

Hausman, J.A., Hall, B. H., and Griliches, Z., "Econometric Models for Count Data With an Application to the Patents-R&D Relationship", *Econometrica* 52(1984), 909-938.

Hall, B. H., Griliches, Z., and Hausman, J. "Patents and R&D: Is There a Lag?" *International Economic Review* 27(1986), 265-283.

Kennedy, P. (1998), *A guide to econometrics.* Cambridge: MIT press.

Montalvo, J.G., "GMM Estimation of Count-Panal-Data Models With Fixed Effects and Predetermined Instruments", *Journal of Business & Economic Statistics* 15 (1997), 82-89.

Pakes, A., and Griliches, Z., "Patents and R&D at the Firm Level: A First Look", in *R&D, Patents and Productivity*, ed. Z. Griliches, Chicago: University of Chicago Press, (1984) pp. 55-72.

Winkelmann, R. (2000), *Econometric analysis of count data***.** New York : Springer.