A consistent estimator for the binomial distribution in the presence of "incidental parameters": an application to patent data

Matilde P. Machado*

Department of Economics, Universidad Carlos III de Madrid, Calle Madrid, 126, 28903 Getafe, Madrid, Spain

Accepted 19 March 2003

Abstract

In this paper a consistent estimator for the Binomial distribution in the presence of incidental parameters, or fixed effects, when the underlying probability is a logistic function is derived. The consistent estimator is obtained from the maximization of a conditional likelihood function in light of Andersen's work. Monte Carlo simulations show its superiority relative to the traditional maximum likelihood estimator with fixed effects also in small samples, particularly when the number of observations in each cross-section, T, is small. Finally, this new estimator is applied to an original dataset that allows the estimation of the probability of obtaining a patent.

JEL classification: C15; C25; O34

Keywords: Count data; Fixed effects; Conditional likelihood; Binomial; Patents

1. Introduction

"Incidental parameters" is a terminology first introduced by Neyman and Scott (1948). It refers to those parameters that are not common to the distribution function of all observations in the data as opposed to structural parameters. The most frequent form of incidental parameters in panel data models are "fixed effects," which have

* Corresponding author. Tel.: +34-91-624-9571; fax: +34-91-624-9875.
E-mail address: mmachado@eco.uc3m.es (M.P. Machado).

become a common instrument to control for time-invariant omitted variables.[1,2] The introduction of fixed effects, however, may cause inconsistent estimates of the slope structural parameters.[3] It is this source of asymptotic bias that is commonly known as "the incidental parameters bias" (see Neyman and Scott, 1948; Andersen, 1973; Lancaster, 2000).

This paper has three main contributions to this literature. First, it derives a consistent, asymptotically normal estimator of the structural parameters of a binomial distribution when the probability of success is a logistic function with fixed effects. This particular binomial distribution is a generalization of the work by Andersen (1973) and Chamberlain (1980) for the case of $N \geqslant 1$ Bernoulli trials. Second, it provides evidence, by way of Monte Carlo simulations, that the small sample performance of this new estimator is superior to the conventional maximum likelihood estimators with fixed effects (m.l.e.f.e). Notice that the asymptotic properties of the new estimator do not render the Monte Carlo simulations superfluous. On the contrary, the simulations are crucial because the new estimator entails some information loss that had unknown consequences in its small sample performance vis a vis the m.l.e.f.e. The simulations are also crucial on giving an indication on the minimum depth of the panel (e.g. number of periods T) required to rend the incidental parameter bias negligible. Third, the paper ends with an application of the new estimator to an original data set drawn from the European Patent Office (EPO) database that, by combining information on the number of patent applications and patents granted at the firm level, allows the estimation of the probability of obtaining a patent.

The derivation of a consistent estimator for this specific distribution[4] is based on the maximization of a conditional (on the sufficient statistic) likelihood function that is free of the incidental parameters. The paper proves the consistency and asymptotic normality of the conditional maximum likelihood estimator (c.m.l.e.) under certain conditions. The proofs are based on Andersen (1970) although Pfanzagl (1993)[5] derives less restrictive sufficient conditions for consistency of conditional maximum likelihood estimators. The main reasons for relying on Andersen's approach are its greater simplicity and the fact

[1] Incidental parameters are also associated with the error-in-variables literature (see Aigner et al., 1984 and the references therein). For a nice review on the "incidental parameters" literature read Lancaster (2000).

[2] Random effects are also a common way of controlling for unobservable heterogeneity. This technique, however, requires independence between the unobservable heterogeneity terms and the other regressors. See, for example, the discussion on the difference between fixed effects and random effects by Lancaster (2000).

[3] Researchers are usually not interested in the fixed effects estimates per se as their interpretation is difficult, but rather on obtaining consistent estimates of slope parameters. Arellano (2000) alerts to the fact that slope parameters alone have a very limited use. In the context of a logit model, for example, the ratio β_1/β_2 is only informative about the relative impact of explanatory variables x_1 and x_2 on the probability of success. On the other hand, a more interesting measure is the marginal impact of a given explanatory variable on the probability of success of the average individual in the sample. The latter, however, begs for an estimate of the fixed effects or knowledge of their conditional distribution.

[4] Reid (2000) and specially Reid (1995) review the roles of conditional inference.

[5] Pfanzagl (1993) Section 4 works out an example with the logistic distribution that is similar to the example in Andersen (1973). He derives a sufficient condition on the sequence of incidental parameters to guarantee consistency of the conditional maximum likelihood estimator.

that for the specific distribution analyzed here some of the restrictions that Pfanzagl points out are naturally satisfied.

Monte Carlo simulations allowed the comparison of the mean square error performances of the c.m.l.e. and the m.l.e.f.e relative to the standard maximum likelihood estimator with a common constant term (hereafter m.l.e.), in relatively small samples. Apart from showing that the c.m.l.e. outperforms the standard m.l.e.f.e. in small samples, the simulations revealed that the biggest gain from using c.m.l.e. is when the depth of the panel, T, and the number of Bernoulli trials N is small. An additional advantage from using the c.m.l.e. for very small T is that the estimation time is much shorter than that of m.l.e.f.e. due to the big reduction in the number of parameters to be estimated.

The Monte Carlo simulations have also shed light on the minimum size of T that makes the incidental parameter bias negligible. Previous studies on this topic are Monte Carlo experiments of the logit model with fixed effects from Wright and Douglas (1976) and the probit model with fixed effects from Heckman (1995). Wright and Douglas (1976) conclude that T has to be around 20 periods for the standard maximum likelihood approach to give results as good as alternative consistent estimators. Heckman, on the other hand, finds that for T as small as 8 (and 100 individuals) the probit model with fixed effects performs reasonably well. For the binomial, the "minimum" T depends not only on the number of individuals in the panel but crucially on the number of Bernoulli trials (i.e. the parameter N on the binomial distribution $B(p,N)$). The simulations show, for example, that for T as small as 5, $N=5$ and a total of 100 observations, the c.m.l.e. reduces the mean square error (MSE) relative to the m.l.e. by only 1.8% more than the m.l.e.f.e. This difference may be small enough to justify the use of the m.l.e.f.e when estimation time is an important issue.[6] For smaller N's, however, the "minimum" T has to be bigger. For $N=2$ and $T=5$, for example, the c.m.l.e. reduces the MSE by 37% more than the m.l.e.f.e. The "minimum" T for $N=2$ is around 10 periods i.e. the double than for $N=5$. The intuition behind this result is that bigger N works as if there were more data points and, therefore, more information in the data, leading to smaller biases in general.

Section 2 introduces the Binomial distribution with a logistic probability of success and presents a simple example of inconsistency of the maximum likelihood estimator with fixed effects. Section 3 derives the c.m.l.e. Section 4 presents the Monte Carlo simulations. Finally, Section 5 applies this estimator to the patents dataset. The appendix provides proofs as well as a short description of the patent data set.

2. The standard MLE for the binomial

The binomial distribution can be applied to numerous data sets. Machado (2001a) for example, uses the binomial distribution to model the treatment outcome of substance abusers. Other examples are the production of homogenous products along an assembly

[6] The estimation time of the c.m.l.e. grows exponentially with T and exceeds the estimation time of the m.l.e.f.e. for relatively small T. The estimation time also increases with the average number of successes in the data set.

line or the number of patents applications that are granted. Suppose there are T periods of data for each firm i, $i = 1,\ldots I$. Denote by N_{it} and K_{it} the number of patent applications and the number of patents granted, respectively, from and to firm i in period t. The probability of observing K_{it} patents granted out of N_{it} applications, for $K_{it} = 0,\ldots,N_{it}$, from firm i, in period t, follows a binomial distribution with parameters N_{it} and p_{it}. The probability of obtaining a patent depends on firm's characteristics, such as the investment in R&D, represented by the vector X_{it}. Assume that the probability of success p_{it} is a logistic function of the data, as follows:

$$p_{it} = \frac{\exp(X'_{it}\beta + \tau_i)}{1 + \exp(X'_{it}\beta + \tau_i)}, \tag{1}$$

where β is a vector of structural parameters while $\tau_1,\ldots,\tau_I$ are incidental parameters representing time-invariant firm characteristics.

Under the hypothesis of independence across periods and firms the log likelihood function is

$$\log L = \sum_{i=1}^{I}\sum_{t=1}^{T} \log l_{it} = \sum_{i=1}^{I}\sum_{t=1}^{T} \log C_{K_{it}}^{N_{it}} + \sum_{i=1}^{I}\sum_{t=1}^{T} [K_{it}(X'_{it}\beta + \tau_i) - N_{it}\log(1 + \exp(X'_{it}\beta + \tau_i))], \tag{2}$$

where

$$C_{K_{it}}^{N_{it}} = \frac{N_{it}!}{K_{it}!(N_{it} - K_{it})!}.$$

The traditional maximum likelihood estimators with fixed effects (m.l.e.f.e.) are the values of β and $\tau_1,\ldots,\tau_I$ that solve the system of first-order conditions (FOC):

$$\begin{cases} \dfrac{\partial \log L}{\partial \tau_i} = 0, \\ \dfrac{\partial \log L}{\partial \beta} = 0, \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{t=1}^{T} K_{it} = \sum_{t=1}^{T} N_{it} \frac{\exp(X'_{it}\beta + \tau_i)}{1 + \exp(X'_{it}\beta + \tau_i)}, \\ \displaystyle\sum_{i=1}^{I}\sum_{t=1}^{T} K_{it}X_{it} = \sum_{i=1}^{I}\sum_{t=1}^{T} N_{it}X_{it} \frac{\exp(X'_{it}\beta + \tau_i)}{1 + \exp(X'_{it}\beta + \tau_i)}, \end{cases} \tag{3}$$

where the first equation of (3) equates the total number of successes and the number of expected successes, as in the case of estimating a single probability.

The m.l.e.f.e. is inconsistent. To prove it, consider the simplest case. Let $T = 2$, $N_{i1} = 1$, $N_{i2} = 2$ $\forall i$, and suppose X_{it} represents a time dummy, i.e. $X_{i1} = 1$ and $X_{i2} = 0$, for all i, similar to Andersen's example (Andersen, 1973). For these values of T, N's and X's, the FOC simplify to:

$$\sum_{t=1}^{2} K_{it} = 1\,\frac{\exp(\beta + \tau_i)}{1 + \exp(\beta + \tau_i)} + 2\,\frac{\exp(\tau_i)}{1 + \exp(\tau_i)},$$

$$\sum_{i=1}^{I} K_{i1} = \sum_{i=1}^{I} \left((\exp \tau_i)\,\frac{\exp(\beta)}{1 + \exp(\beta + \tau_i)} \right). \tag{4}$$

The possible values of $\tilde{K}_i = \sum_{t=1}^{2} K_{it}$ are $\{0,1,2,3\}$. If $\tilde{K}_i$ is 3 or 0, then $\tau_i = +\infty$ and $-\infty$, respectively. If $\tilde{K}_i = 1$, the FOC for τ_i implies:

$$\exp(\tau_i) = \frac{-1 + \sqrt{1 + 8\exp(\beta)}}{4\exp(\beta)} \tag{5}$$

and for $\tilde{K}_i = 2$, the solution is

$$\exp(\tau_i) = \frac{\exp(\beta) + \sqrt{\exp(2\beta) + 8\exp(\beta)}}{2\exp(\beta)}. \tag{6}$$

By replacing (5) and (6) in the second equation of (4), the FOC for β becomes:

$$\sum_{i=1}^{I} K_{i1} = n_1 \left(\frac{-1 + \sqrt{1 + 8\exp(\beta)}}{3 + \sqrt{1 + 8\exp(\beta)}} \right) + n_2 \left(\frac{\exp(\beta) + \sqrt{\exp(2\beta) + 8\exp(\beta)}}{2 + \exp(\beta) + \sqrt{\exp(2\beta) + 8\exp(\beta)}} \right) + n_3, \tag{7}$$

where n_K are the number of firms with $\tilde{K}_i = K$.

Example 1. In the limit as $I \to \infty$, the maximum likelihood estimator, i.e. the solution to Eq. (7), is always inconsistent unless, just as in the logit case, the true parameter β_0 equals zero. Moreover, the m.l.e. will underestimate the true parameter β_0 whenever $\beta_0 < 0$ and overestimates the true parameter whenever $\beta_0 > 0$. Furthermore, it is possible to show that $1 < \hat{\beta}/\beta_0 < 2$, i.e the bias is strictly smaller than in the logit case.

Proof. See appendix.

3. A consistent estimator for the binomial distribution

3.1. The conditional maximum likelihood estimator

Let K_{it}, where $i = 1,\ldots,I$ and $t = 1,\ldots,T$, be a binomial random variable with the underlying probability of success p_{it} given by (1). X_{it} is a vector of regressors associated with the vector of structural parameters β and τ_i for $i = 1,\ldots,I$ are firm-specific fixed effects. Define $K = (K_1,\ldots,K_I)$, and $N = (N_1,\ldots,N_I)$, where $K_i = (K_{i1},\ldots,K_{iT})$ and $N_i = (N_{i1},\ldots,N_{iT})$, $i = 1,\ldots,I$. Under the assumption that observations are independent across time, the joint probability distribution for firm i is

$$f_i(K_i|\beta,\tau_i,N_i) = C_{K_{i1}}^{N_{i1}} p_{i1}^{K_{i1}} (1 - p_{i1})^{N_{i1}-K_{i1}} \ldots C_{K_{iT}}^{N_{iT}} p_{iT}^{K_{iT}} (1 - p_{iT})^{N_{iT}-K_{iT}} \tag{8}$$

which after replacing p_{it} with (1) for $t = 1,\ldots,T$, becomes

$$f_i(K_i|\beta,\tau_i,N_i) = C_{K_{i1}}^{N_{i1}} \cdots C_{K_{iT}}^{N_{iT}} \frac{e^{K_{i1}X_{i1}'\beta + \cdots + K_{iT}X_{iT}'\beta}}{(1 + e^{X_{i1}'\beta+\tau_i})^{N_{i1}} \ldots (1 + e^{X_{iT}'\beta+\tau_i})^{N_{iT}}} e^{\tau_i \sum_{t=1}^{T} K_{it}}. \tag{9}$$

The joint distribution function for all firms under the assumption of independence across both time and firms, expression (10), is not free of the incidental parameters and, given the nonlinearities in the FOC, results in "incidental parameter bias."

$$f(K|\beta,\tau_1,\ldots,\tau_I,N_1,\ldots,N_I)=\prod_{i=1}^{I}\frac{(\prod_{t=1}^{T}C_{K_{it}}^{N_{it}})e^{(K_{i1}X_{i1}'\beta+\cdots+K_{iT}X_{iT}'\beta)}e^{\tau_i\sum_{t=1}^{T}K_{it}}}{(1+e^{X_{i1}'\beta+\tau_i})^{N_{i1}}\ldots(1+e^{X_{iT}'\beta+\tau_i})^{N_{iT}}}. \quad (10)$$

Notice, however, that expression (10) can be decomposed into two parts as Neyman's factorization (11):

$$f(K|\beta,\tau_1,\ldots,\tau_I,N)=u(K,\beta,N)\prod_{i=1}^{I}v_i\left(\sum_{t=1}^{T}K_{it},\tau_i,\beta,N\right). \quad (11)$$

It follows, by definition, that $\mathcal{T}_i=\sum_{t=1}^{T}K_{it}$, $i=1,\ldots,I$ are joint sufficient statistics for $\tau_1,\ldots,\tau_I$. It turns out that the existence of a set of sufficient statistics allows the construction of a joint conditional likelihood function that is independent of the incidental parameters and, therefore, solves the "incidental parameter bias" problem.

Given the distribution function of the sufficient statistic $\mathcal{T}_i$ for firm i:

$$\begin{aligned}P(\mathcal{T}_i=\bar{K}_i|\beta,\tau_i)&=\sum_{z_1+\cdots+z_T=\bar{K}_i}C_{z_1}^{N_{i1}}..C_{z_T}^{N_{iT}}p_{i1}^{z_1}(1-p_{i1})^{N_{i1}-z_1}\ldots p_{iT}^{z_T}(1-p_{iT})^{N_{iT}-z_T}\\&=\frac{\sum_{z_1+\cdots+z_T=\bar{K}_i}C_{z_1}^{N_{i1}}\ldots C_{z_T}^{N_{iT}}e^{z_1X_{i1}'\beta+\cdots+z_TX_{iT}'\beta}e^{\tau_i(z_1+\cdots+z_T)}}{(1+e^{X_{i1}'\beta+\tau_i})^{N_{i1}}\ldots(1+e^{X_{iT}'\beta+\tau_i})^{N_{iT}}}.\end{aligned} \quad (12)$$

One can easily derive the conditional distribution of K_i given $\sum_{t=1}^{T}K_{it}$:

$$\begin{aligned}\phi_i\left(K_i|\beta,\sum_{t=1}^{T}K_{it}=\bar{K}_i\right)&=\frac{f_T(K_i|\beta,\tau_i,N_i)}{P(\mathcal{T}_i=\bar{K}_i|\beta,\tau_i)}\\&=\frac{C_{K_{i1}}^{N_{i1}}\ldots C_{K_{iT}}^{N_{iT}}e^{K_{i1}X_{i1}'\beta+\cdots+K_{iT}X_{iT}'\beta}}{\sum_{z_1+\cdots+z_T=\bar{K}_i}C_{z_1}^{N_{i1}}\ldots C_{z_T}^{N_{iT}}e^{z_1X_{i1}'\beta+\cdots+z_TX_{iT}'\beta}}.\end{aligned} \quad (13)$$

As claimed, ϕ_i does not depend on τ_i and, therefore, is free of the "incidental parameter bias". [7]

The likelihood function under the assumption of independence across i's is simply

$$L(\beta|K,N,\bar{K}_1,\ldots,\bar{K}_I)=\phi=\prod_{i=1}^{I}\phi_i \quad (14)$$

[7] For examples on other distribution functions see Lancaster (2000).

Finally, define the c.m.l.e. as the vector β that maximizes:

$$\log L(\beta|K,N,\bar{K}_1,\ldots,\bar{K}_I) = \sum_{i=1}^{I} \log l_i$$

$$= \sum_{i=1}^{I} \log \frac{C_{K_{i1}}^{N_{i1}} \ldots C_{K_{iT}}^{N_{iT}} e^{K_{i1}X_{i1}'\beta+\cdots+K_{iT}X_{iT}'\beta}}{\sum_{z_1+\cdots+z_T=\bar{K}_i} C_{z_1}^{N_{i1}} \ldots C_{z_T}^{N_{iT}} e^{z_1X_{i1}'\beta+\cdots+z_TX_{iT}'\beta}}. \quad (15)$$

3.2. Consistency and asymptotic normality

Theorems 1 and 2 give sufficient conditions for consistency of the c.m.l.e. for the single and the multiple parameter case, respectively.

Theorem 1. *The single parameter case. Take a data set* (K_{it},X_{it}), $t=1,\ldots,T$ *and* $i=1,\ldots,I$ *where* K_{it} *are independent draws from Binomial distributions with a logistic probability of the form of Eq.* (1) *and* X_{it} *are exogenous non-random independent variables. Then, if the following conditions hold, there will exist a unique c.m.l.e. and it is consistent*: (1) *the true structural parameter* $|\beta_0|<\infty$, (2) *the incidental parameters,* τ_i, *belong to a compact set* Ω_0, *for all* i, (3) *there is at least one* (i,j) *(where* i *may equal* j*) such that* $K_i'X_i \neq \max_{z\in Z_i} z'X_i$ *and* $K_j'X_j \neq \min_{z\in Z_j} z'X_j$ *where* Z_s *is the set of all possible vectors* z *that satisfy* $\sum_{t=1}^{T} z_t = \bar{K}_s$ *where* z_t *is an integer such that* $0 \leqslant z_t \leqslant N_{st}$ *for* $s=i,j$.

Note 1. A necessary condition for (3) to be satisfied is that for at least one i the variable X_i is not constant for all t and the set Z_i has at least three elements. A necessary condition for Z_i to have at least three elements is, of course, that $N_{it}>1$ or $T>2$.

Theorem 2. *The multiparameter case. Take a data set* (K_{it},X_{it}), $t=1,\ldots,T$ *and* $i=1,\ldots,I$ *where* K_{it} *are independent draws from Binomial distributions with a logistic probability of the form of Eq.* (1) *and* X_{it} *are exogenous non-random independent vectors. Then, if the following conditions hold, there will exist a unique c.m.l.e. and it is consistent*: (1) *the true structural vector of parameters* $|\beta_{0q}|<\infty$, *for all* $q=1,\ldots,m$, (2) *the incidental parameters,* τ_i, *belong to a compact set* Ω_0, *for all* i, (3) *there is at least one* (i,j) *(where* i *may equal* j*) such that* $K_i'X_{iq} \neq \max_{z\in Z_i} z'X_{iq}$ *and* $K_j'X_{jq} \neq \min_{z\in Z_j} z'X_{jq}$ *for each* $q=1,\ldots,m$, *where* Z_s *is the set of all possible vectors* z, *where* z_t *is an integer such that* $0 \leqslant z_t \leqslant n_{st}$ *and* $z_t \leqslant N_{st}$ *for* $s=i,j$. (4) *The rank*$(X)=m$.

Note 1 from Theorem 1 also applies by changing X_{it} for X_{iqt}.

Proof. [8] The conditions on the parameters β_0 and τ_i's are equivalent to requiring enough variation in the dependent variable across and within firms i. For example, if

[8] Note that Theorems 1 and 2 lay out sufficient conditions for the uniqueness and consistency of the c.m.l.e. These are not, however, necessary. Chamberlain's conditional logit case for $T=2$ (Chamberlain, 1980) is an example where condition (3) is never satisfied. In that case one needs to have enough variation in the sample in order to obtain an interior solution for the FOC.

$\beta_0 = \infty$ then all firms would register a success rate of 100 percent, i.e. $\bar{K}_i = \sum_{t=1}^{T} N_{it}$ for all i. The same would happen with unbounded values of τ_i. These, however, do not constitute a problem. The reason being that all firms with sufficient statistics equal to either 0 or $\sum_{t=1}^{T} N_{it}$ are dropped from the estimation since their conditional probability distribution contains no information on β_0.[9] See proof in the appendix. □

Let $\beta \in \Re^m$ and $B^2(\beta, \tau)$ be the matrix of elements $b^2_{jp}(\beta, \tau)$:

$$b^2_{jp}(\beta,\tau) = E_{\beta,\tau}\left\{\left(\frac{\partial \log \phi(K|\beta,\mathcal{T})}{\partial \beta_j}\right)\left(\frac{\partial \log \phi(K|\beta,\mathcal{T})}{\partial \beta_p}\right)\right\} \quad \text{for } p,q = 1,\ldots,m \tag{16}$$

Denote by $B^2_I = \sum_{i=1}^{I} B^2(\beta_0, \tau_i)$. Since B^2_I is a positive semidefinite matrix there is a B_I such that $B^2_I = B'_I B_I$.

Theorem 3. *If the conditions in Theorem* 1 *for* $m = 1$ (*or Theorem* 2 *for* $m > 1$) *hold then the c.m.l.e.* $\hat{\beta}_I$ *is asymptotically normal distributed with mean* β_0 *and variance-covariance matrix* B_I^{-2}, *i.e.* $(\hat{\beta}_I - \beta_0)B'_I$ *converges in distribution to a m-dimensional standard normal distribution.*

Proof. See the appendix.[10]

4. Monte Carlo simulations

The Monte Carlo simulations intend to: (1) compare the estimation results obtained with the standard m.l.e.f.e. and the results obtained with the c.m.l.e. for small samples; and (2) examine how big T must be for the "incidental parameter bias" to become negligible. One may think that because the c.m.l.e. is consistent (1) is a superfluous exercise. The comparison of the two estimators for small samples, however, is not obvious because the conditioning on sufficient statistics entails some efficiency loss. Efficiency is only guaranteed if the sufficient statistic for τ_i is also ancillary for β, in which case, by definition, the conditioning does not involve any loss of information

[9] Pfanzagl (1993) for instance does not require that τ_i's belong to a compact set. In the setting of the binomial distribution with the logistic probability, however, the requirement that τ_i's belong to a compact set is not restrictive because if for some j, $|\tau_j| = \infty$ then this observation does not contribute to the inference of the structural parameter(s) (or in Pfanzagl terminology, its sufficient statistic is contracting).

[10] Chamberlain (1980) footnote 6 discusses the fact that the conditional maximum likelihood estimator for the logistic distribution does not attain the Cramer–Rao lower bound. This is not surprising since the sufficient statistic is not ancillary. Mantel and Godambe (1993) say that when the sufficient statistic is ancillary then the conditional score is a globally optimal estimating function. In that paper they derive optimal estimating functions belonging to the set of linear functions in the presence of incidental parameters for cases where the sufficient statistic is not ancillary. The derivation of the optimal linear estimating functions is, however, beyond the scope of this paper.

about the structural parameter.[11] The sufficient statistic $\mathscr{T}$ is not ancillary for β, as can be seen from (12).

4.1. The data generating process

Two sets of experiments are shown. In the first set, the number of Bernoulli trials $N_{it}=N=5$, for all t and i, while in the second set N_{it}'s are set equal to 2. In both sets, the fixed effects, τ_i, are correlated with the single regressor X through the relationship:

$$X_{it} = \tau_i + \varepsilon_{it}, \tag{17}$$

where $\varepsilon_{it} \sim \mathrm{N}(0,1)$. At each simulation a different τ_i is drawn from an *i.i.d.* $\mathrm{N}(0,1)$, and given τ_i, a new X_{it} is randomly drawn from $\mathrm{N}(\tau_i,1)$.[12] The true value of the structural parameter β is 0.5. The probability of success p_{it} is given by the logistic function:

$$p_{it} = \frac{\exp(\tau_i + 0.5X_{it})}{1+\exp(\tau_i + 0.5X_{it})}. \tag{18}$$

In both sets of experiments, the total number of observations $I \times T$ is first kept fixed while the length of the panel T varies. Keeping $I \times T$ fixed allows to distinguish the impact of changes in T from changes in $I \times T$ on the estimated $\hat{\beta}$. Changing T is also important because the bias from leaving the fixed effects out of the estimation changes with T.[13]

The results are normalized by comparing the bias and the MSE obtained from the estimation with fixed effects (m.l.e.f.e. and c.m.l.e.) with the bias and MSE obtained with the m.l.e. This normalization enables the comparison of the results when T changes. The idea is to see what fraction of the MSE obtained with the m.l.e. is eliminated by using the c.m.l.e. instead of the standard m.l.e.f.e., as well as to see how this fraction changes with T.

---

[11] A statistic $\mathscr{E}$ is ancillary for the structural parameter β, if the distribution of $\mathscr{E}$ is independent of β. For proof that ancillary is a sufficient condition to attain efficiency, see Theorem 3.4 in Andersen (1973).

[12] Note that the data generating process is quite general and can be interpreted as a random effects model in the sense that the τ's are not fixed but are drawn from a distribution.

[13] Note that in the case of a linear model $y_{it} = \tau_i + \beta X_{it} + u_{it}$ the asymptotic bias on $\hat{\beta}$ from estimating the model $y_{it} = \tau + \beta X_{it} + u_{it}$ does not depend on T:

$$\frac{\sum_i \sum_t \tau_i X_{it}}{\sum_i \sum_t X_{it}^2} = \frac{T\sum_i \tau_i^2 + \sum_i \sum_t \tau_i \varepsilon_{it}}{T\sum_i \tau_i^2 + \sum_i \sum_t \varepsilon_{it}^2 + 2\sum_i \sum_t \tau_i \varepsilon_{it}}$$

$$= \frac{(\sum_i \tau_i^2)/I + (\sum_i \sum_t \tau_i \varepsilon_{it})/I}{((\sum_i \tau_i^2)/I) + ((\sum_i \sum_t \varepsilon_{it}^2)/IT) + ((2\sum_i \sum_t \tau_i \varepsilon_{it})/IT)} \overset{I\to\infty}{\to} \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}.$$

The best intuition for this result is the fact that the fixed effects can be easily eliminated by subtracting the sample averages:

$$y_{it} - y_i = \beta(X_{it} - X_i) + u_{it} - u_i,$$

where y_i, X_i, and u_i are the sample averages for individual i.

Table 1
Comparison of estimators when total observations = 100

$\beta = 0.5$	Standard m.l.e.				c.m.l.e.	
	Without fixed effects		With fixed effects			
	$N = 5$	$N = 2$	$N = 5$	$N = 2$	$N = 5$	$N = 2$
$T = 2$	$\bar{\hat{\beta}} = 0.758$ $\sigma_{\hat{\beta}} = 0.101$	$\bar{\hat{\beta}} = 0.524$ $\sigma_{\hat{\beta}} = 0.145$	$\bar{\hat{\beta}} = 0.568$ $\sigma_{\hat{\beta}} = 0.184$	$\bar{\hat{\beta}} = 0.712$ $\sigma_{\hat{\beta}} = 0.390$	$\bar{\hat{\beta}} = 0.509$ $\sigma_{\hat{\beta}} = 0.164$	$\bar{\hat{\beta}} = 0.527$ $\sigma_{\hat{\beta}} = 0.285$
$T = 4$	$\bar{\hat{\beta}} = 0.856$ $\sigma_{\hat{\beta}} = 0.119$	$\bar{\hat{\beta}} = 0.739$ $\sigma_{\hat{\beta}} = 0.160$	$\bar{\hat{\beta}} = 0.531$ $\sigma_{\hat{\beta}} = 0.140$	$\bar{\hat{\beta}} = 0.592$ $\sigma_{\hat{\beta}} = 0.256$	$\bar{\hat{\beta}} = 0.503$ $\sigma_{\hat{\beta}} = 0.132$	$\bar{\hat{\beta}} = 0.514$ $\sigma_{\hat{\beta}} = 0.219$
$T = 5$	$\bar{\hat{\beta}} = 0.872$ $\sigma_{\hat{\beta}} = 0.124$	$\bar{\hat{\beta}} = 0.784$ $\sigma_{\hat{\beta}} = 0.163$	$\bar{\hat{\beta}} = 0.572$ $\sigma_{\hat{\beta}} = 0.238$	$\bar{\hat{\beta}} = 0.572$ $\sigma_{\hat{\beta}} = 0.238$	$\bar{\hat{\beta}} = 0.506$ $\sigma_{\hat{\beta}} = 0.131$	$\bar{\hat{\beta}} = 0.511$ $\sigma_{\hat{\beta}} = 0.211$
$T = 10$	$\bar{\hat{\beta}} = 0.886$ $\sigma_{\hat{\beta}} = 0.146$	$\bar{\hat{\beta}} = 0.855$ $\sigma_{\hat{\beta}} = 0.186$	$\bar{\hat{\beta}} = 0.517$ $\sigma_{\hat{\beta}} = 0.122$	$\bar{\hat{\beta}} = 0.544$ $\sigma_{\hat{\beta}} = 0.208$	$\bar{\hat{\beta}} = 0.501$ $\sigma_{\hat{\beta}} = 0.127$	$\bar{\hat{\beta}} = 0.515$ $\sigma_{\hat{\beta}} = 0.195$
$T = 20$	$\bar{\hat{\beta}} = 0.860$ $\sigma_{\hat{\beta}} = 0.189$	$\bar{\hat{\beta}} = 0.856$ $\sigma_{\hat{\beta}} = 0.225$	$\bar{\hat{\beta}} = 0.513$ $\sigma_{\hat{\beta}} = 0.123$	$\bar{\hat{\beta}} = 0.529$ $\sigma_{\hat{\beta}} = 0.201$	—	—
$T = 50$	$\bar{\hat{\beta}} = 0.745$ $\sigma_{\hat{\beta}} = 0.266$	$\bar{\hat{\beta}} = 0.750$ $\sigma_{\hat{\beta}} = 0.298$	$\bar{\hat{\beta}} = 0.504$ $\sigma_{\hat{\beta}} = 0.121$	$\bar{\hat{\beta}} = 0.511$ $\sigma_{\hat{\beta}} = 0.193$	—	—

The simulations are repeated for the number of cross sectional units (I) fixed while T varies. The purpose of this experiment is to measure the value of an additional year/month/week of data.

4.2. Results

All tables show the average value of the structural parameter β estimates over 4000 simulations and the standard deviation across those simulations, $\sigma_{\hat{\beta}} = (\frac{1}{4000} \sum_{j=1}^{4000} (\hat{\beta}^j - \bar{\hat{\beta}})^2)^{1/2}$ where $\bar{\hat{\beta}} = \frac{1}{4000} \sum_{j=1}^{4000} \hat{\beta}^j$. The results from Table 1 show that, for the range of T considered, the bias and the standard deviation of the c.m.l.e. are always smaller than those for the m.l.e.f.e. for both $N = 5$ and 2 when the number of observations is kept constant at 100.[14] The table also shows that the difference between the two estimators decreases as T grows. Furthermore, with the exception of $N = 2$, $T = 2$, the m.l.e.f.e. is always preferable to the standard m.l.e. without fixed effects.

[14] For $T \geqslant 10$ the estimation time for the c.m.l.e. becomes excessive even when using sub-routines written in C language. Therefore, the c.m.l.e. results are only partially reported for $T = 10$ and $N = 5$ (185 simulations) and no results are reported for $T > 10$. The partial results for $T = 10$ and $N = 5$ show a somewhat bigger standard deviation for the c.m.l.e. in comparison with the m.l.e.f.e. which may be due to the reduced number of simulations.

Table 2
MSE when total observations = 100

$\beta = 0.5$	Standard m.l.e.						c.m.l.e.			
	Without fixed effects		With fixed effects							
	MSE		MSE		Reduction (%)		MSE		Reduction (%)	
	$N = 5$	$N = 2$	$N = 5$	$N = 2$	$N = 5$	$N = 2$	$N = 5$	$N = 2$	$N = 5$	$N = 2$
$T = 2$	0.0768	0.0216	0.0385	0.1970	49.87	−812.20	0.0230	0.0820	64.86	−279.40
$T = 4$	0.1409	0.0827	0.0206	0.074	85.41	10.54	0.0174	0.0482	87.63	41.78
$T = 5$	0.1538	0.1072	0.0196	0.0618	87.28	42.34	0.0172	0.0446	88.82	58.37
$T = 10$	0.1703	0.1606	0.0152	0.0452	91.09	71.86	0.0161	0.0383	90.53	76.19
$T = 20$	0.1653	0.1774	0.0153	0.0412	90.75	76.75	—		—	
$T = 50$	0.1308	0.1513	0.0147	0.0374	88.79	75.30	—		—	

Table 2 shows the MSE corresponding to the results in Table 1 and the reduction in the MSE that the standard m.l.e.f.e and the c.m.l.e. bring with respect to m.l.e. First, notice that the MSE of both c.m.l.e. and m.l.e.f.e. decreases with T. Second, the MSE of the c.m.l.e. is always smaller than the MSE for the m.l.e.f.e.,[15] and the difference between the two decreases with T. For T as small as 5 and $N = 5$, the reduction in the MSE of the c.m.l.e. and m.l.e.f.e. are almost equivalent, 88.82% and 87.28%, respectively. For $T = 5$ and $N = 2$ the difference between c.m.l.e. and m.l.e.f.e. is larger. The Monte Carlo runs have also shown that the estimation time of the c.m.l.e. increases exponentially with T although it is lower than the m.l.e.f.e.'s for very low T. Taken as a whole, the standard m.l.e.f.e. is a reasonable alternative to the c.m.l.e. for relatively large values of T and N as it delivers similar results faster.

Lastly, the first column of Table 2 reveals that the MSE for the m.l.e. increases with T for relative small values of T. It is plausible that the constant term has a bigger standard variation when the number of τ_i's is smaller, i.e. when T is large. In other words, the constant term $\tilde{\tau}$ is capturing something related to the mean of the fixed effects and since the τ_i's are normally distributed, the variance of $\tilde{\tau}$ decreases with I (increases in T). This effect dominates at least up to $T = 10$ for $N = 5$ and up to $T = 20$ for $N = 2$.

Tables 3 and 4 show simulation results when I is fixed and T varies. The biggest gain of the c.m.l.e. vis-a-vis the m.l.e.f.e. is again when T is relatively small. In all cases, the c.m.l.e. is preferable to the m.l.e.f.e. Again, with the exception of $N = 2$ and $T = 2$, the m.l.e.f.e. is preferable to the standard m.l.e. without fixed effects.

Comparing both sets of experiments, it is clear that when N decreases, the MSE performance of the c.m.l.e. and the m.l.e.f.e decreases while that of the m.l.e. increases. The advantage of the c.m.l.e. over the m.l.e.f.e., however, increases when N decreases.

[15] Again the exception for $T = 10$ and $N = 5$ is likely due to the reduced number of simulations conducted.

Table 3
Comparison of estimators when $I = 50$

$\beta = 0.5$	Standard m.l.e.				c.m.l.e.	
	No fixed effects		With fixed effects			
$I = 50$	$N = 5$	$N = 2$	$N = 5$	$N = 2$	$N = 5$	$N = 2$
$T = 2$	$\bar{\hat{\beta}} = 0.758$ $\sigma_{\hat{\beta}} = 0.101$	$\bar{\hat{\beta}} = 0.524$ $\sigma_{\hat{\beta}} = 0.145$	$\bar{\hat{\beta}} = 0.568$ $\sigma_{\hat{\beta}} = 0.184$	$\bar{\hat{\beta}} = 0.712$ $\sigma_{\hat{\beta}} = 0.390$	$\bar{\hat{\beta}} = 0.509$ $\sigma_{\hat{\beta}} = 0.164$	$\bar{\hat{\beta}} = 0.527$ $\sigma_{\hat{\beta}} = 0.285$
$T = 3$	$\bar{\hat{\beta}} = 0.826$ $\sigma_{\hat{\beta}} = 0.090$	$\bar{\hat{\beta}} = 0.664$ $\sigma_{\hat{\beta}} = 0.122$	$\bar{\hat{\beta}} = 0.540$ $\sigma_{\hat{\beta}} = 0.124$	$\bar{\hat{\beta}} = 0.619$ $\sigma_{\hat{\beta}} = 0.231$	$\bar{\hat{\beta}} = 0.503$ $\sigma_{\hat{\beta}} = 0.115$	$\bar{\hat{\beta}} = 0.511$ $\sigma_{\hat{\beta}} = 0.188$
$T = 4$	$\bar{\hat{\beta}} = 0.857$ $\sigma_{\hat{\beta}} = 0.081$	$\bar{\hat{\beta}} = 0.735$ $\sigma_{\hat{\beta}} = 0.108$	$\bar{\hat{\beta}} = 0.533$ $\sigma_{\hat{\beta}} = 0.10$	$\bar{\hat{\beta}} = 0.586$ $\sigma_{\hat{\beta}} = 0.175$	$\bar{\hat{\beta}} = 0.505$ $\sigma_{\hat{\beta}} = 0.094$	$\bar{\hat{\beta}} = 0.509$ $\sigma_{\hat{\beta}} = 0.150$
$T = 5$	$\bar{\hat{\beta}} = 0.873$ $\sigma_{\hat{\beta}} = 0.077$	$\bar{\hat{\beta}} = 0.779$ $\sigma_{\hat{\beta}} = 0.100$	$\bar{\hat{\beta}} = 0.525$ $\sigma_{\hat{\beta}} = 0.085$	$\bar{\hat{\beta}} = 0.566$ $\sigma_{\hat{\beta}} = 0.145$	$\bar{\hat{\beta}} = 0.503$ $\sigma_{\hat{\beta}} = 0.081$	$\bar{\hat{\beta}} = 0.507$ $\sigma_{\hat{\beta}} = 0.128$
$T = 6$	$\bar{\hat{\beta}} = 0.881$ $\sigma_{\hat{\beta}} = 0.072$	$\bar{\hat{\beta}} = 0.806$ $\sigma_{\hat{\beta}} = 0.093$	$\bar{\hat{\beta}} = 0.518$ $\sigma_{\hat{\beta}} = 0.076$	$\bar{\hat{\beta}} = 0.553$ $\sigma_{\hat{\beta}} = 0.129$	$\bar{\hat{\beta}} = 0.501$ $\sigma_{\hat{\beta}} = 0.073$	$\bar{\hat{\beta}} = 0.505$ $\sigma_{\hat{\beta}} = 0.117$
$T = 7$	$\bar{\hat{\beta}} = 0.889$ $\sigma_{\hat{\beta}} = 0.068$	$\bar{\hat{\beta}} = 0.828$ $\sigma_{\hat{\beta}} = 0.087$	$\bar{\hat{\beta}} = 0.517$ $\sigma_{\hat{\beta}} = 0.066$	$\bar{\hat{\beta}} = 0.547$ $\sigma_{\hat{\beta}} = 0.111$	$\bar{\hat{\beta}} = 0.502$ $\sigma_{\hat{\beta}} = 0.065$	$\bar{\hat{\beta}} = 0.506$ $\sigma_{\hat{\beta}} = 0.102$

Table 4
MSE when $I = 50$

$\beta = 0.5$	Standard m.l.e.						c.m.l.e.			
	Without fixed effects		With fixed effects							
$I = 50$	MSE		MSE		Reduction (%)		MSE		Reduction (%)	
	$N = 5$	$N = 2$	$N = 5$	$N = 2$	$N = 5$	$N = 2$	$N = 5$	$N = 2$	$N = 5$	$N = 2$
$T = 2$	0.0768	0.0216	0.0385	0.1970	49.87	−812.20	0.0270	0.0820	64.86	−279.40
$T = 3$	0.1144	0.0418	0.0170	0.0675	85.16	−61.61	0.0132	0.0355	88.43	15.11
$T = 4$	0.1340	0.0669	0.0111	0.0380	91.73	43.16	0.0089	0.0226	93.39	66.24
$T = 5$	0.1451	0.0878	0.0079	0.0254	94.59	71.11	0.0066	0.0164	95.47	81.29
$T = 6$	0.1503	0.1023	0.0061	0.0195	95.94	80.98	0.0053	0.0137	96.45	86.59
$T = 7$	0.1559	0.1152	0.0046	0.0145	97.02	87.38	0.0042	0.0104	97.29	90.93

Lastly, for smaller N, the minimum number of periods needed to render the "incidental parameter bias" acceptable increases. [16]

[16] The increase in the "minimum" T as N falls was to be expected as it approaches the logit case where the results by Wright and Douglas (1976) apply.

5. Application to patents data

The innovation literature has, in the past, relied on datasets that either contained information on *patents granted* or on *patent applications*. It was not possible then, to model and estimate the probability of obtaining a patent. The subsample from the EPO explored in this section is innovative since it includes data on both *patents granted* and *patent applications* at the firm level.[17]

The purpose of this section is mainly to provide an example where to apply the c.m.l.e. estimator rather than a serious attempt at modelling the issue. A more serious model would require modelling not only the probability of success but also the number of applications as a function of R&D expenditures. This example only deals with the former. The number of applications is assumed to be exogenous.

The variables N_{it} and K_{it} were defined in Section 2. Table 5 presents results for several different specifications of the matrix of regressors X_{it}.[18] It shows that, in general, contemporaneous R&D expenditures have a positive impact although only in three out of six specifications it is statistically significantly different from zero. On the other hand, lagged values of this variable contribute negatively (often statistically significantly) to the probability of success. When the number of lags increases, the strongest negative effects concentrate on the medium lag and then fade away at the extremes.[19] It is likely that the higher R&D expenditures contribute to an inflation of applications that are rejected by the EPO.[20,21] Finally, two ad hoc measures of the sum of the R&D expenditures over the sample period were constructed using a discount rate of 15%. The coefficients on these variables are significantly negative. The coefficients on the year dummies in columns 5–7 of Table 5 show that years 1988–1991 are statistically different from year 1992. The trend coefficient is always negative and for most cases significantly different from zero.

[17] Hausman et al. (1984) (HHG) on the relationship between patents granted and R&D expenditures is the main reference for this section. For more details on the application to patents data and the EPO subsample used here and its similarities with HHG sample, refer to the working paper version of this paper (Machado, 2001b).

[18] R&D expenditures are divided by 1000 and normalized to have mean zero and standard deviation equal to 1.

[19] Hausman et al. (1984) have also found an inverted u-shape relationship between the values of their coefficients and the lag structure of R&D which, they thought was evidence of truncation in the R&D lag structure. They corrected the truncation by introducing fixed effects which, caused the coefficients on lagged R&D to become very small and difficult to identify. Likewise, in Table 5, although lagged coefficients are sometimes large they seem not to be well identified due to correlation between the different R&D variables.

[20] Simple regressions of $\ln(N_{it}+1)$ and $\ln(K_{it}+1)$ on a trend and R&D expenditures of different lags show that the increase in R&D expenditures brings a bigger increase on the number of applications than on the number of patents accepted. It is likely that bigger firms can save on application costs to the point that is worth while applying for patents with lower probability of being accepted and, therefore, end up with a relative larger number of applications.

[21] Another explanation may be the endogeneity of the number of applications. All equations in Table 5 were reestimated with N_{it} as an additional regressor. Its coefficient estimate was between 0.0064 and 0.0088 and always very statistically significantly different from zero. Nevertheless, its inclusion did not affect any of the other coefficients.

Table 5
Results restricted to firms with less than 400 total patent applications

$n_{obs} = 405$	1	2	3	4	5	6	7	8
Year 1988					0.527	0.520		
					(0.082)	(0.079)		
Year 1989					0.528	0.516		
					(0.072)	(0.069)		
Year 1990					0.521	0.506		
					(0.067)	(0.066)		
Year 1991					0.474	0.462		
					(0.064)	(0.063)		
Trend (1988–1991)	−0.070	−0.070	−0.055	−0.022			−0.034	−0.014
	(0.023)	(0.022)	(0.023)	(0.025)			(0.024)	(0.025)
Year 1992	−0.785	−0.774	−0.711	−0.575			−0.600	−0.521
	(0.084)	(0.082)	(0.083)	(0.100)			(0.093)	(0.095)
$R\&D_t$	−0.456	1.013	0.258	1.269	0.012			0.813
	(0.327)	(0.484)	(0.307)	(0.676)	(0.428)			(0.386)
$R\&D_{t-1}$		−2.450		−1.689		−0.787		
		(0.613)		(0.689)		(0.346)		
$R\&D_{t-2}$			−2.879					
			(0.496)					
$R\&D_{t-3}$				−2.741	−2.609	−2.397		
				(0.469)	(0.423)	(0.427)		
$R\&D_{t-5}$				−0.252	−0.307	−0.108		
				(0.558)	(0.529)	(0.209)		
$R\&D_{t-7}$				−0.892				
				(0.436)				
$R\&D_{t-8}$				0.471	−0.582	−0.634		
				(0.442)	(0.273)	(0.248)		
$\sum_{s=t}^{t-T} (0.85)^{t-s} R\&D_s$							−2.952	
							(0.620)	
$\sum_{s=t-1}^{t-T} (0.85)^{t-1-s} R\&D_s$								−4.839
								(0.780)
Likelihood	−630.53	−622.65	−614.00	−604.25	−609.79	−608.24	−619.83	−611.42
R^2	0.946	0.950	0.954	0.959	0.956	0.957	0.952	0.955

After obtaining the estimates for the structural parameters, the estimates for the fixed effects can be obtained by maximizing a likelihood function using (12) and replacing β with its c.m.l.e. [22]

Table 6 recreates the Monte Carlo comparisons of last section for two of the specifications in Table 5 and two different values of T. First, observe that the differences between m.l.e.f.e. and the c.m.l.e. are much smaller than the ones found in the Monte Carlo simulations due to the much bigger values of N—the average number of applications in the EPOs subsample is around 50. [23] Second, notice that the difference between the c.m.l.e. and the m.l.e.f.e. decreases with T.

6. Conclusion

This paper derives a consistent and asymptotically normal estimator (c.m.l.e.) for the structural parameters of the binomial distribution when the probability of success is a logistic function with "incidental parameters." The c.m.l.e. is obtained by conditioning the likelihood function on sufficient statistics for the incidental parameters.

Results from Monte Carlo simulations show that c.m.l.e. is also superior in terms of means square error (MSE) to the traditional m.l.e.f.e. in relative small samples. This result was not obvious since the sufficient statistics are not ancillary. The bias, the variance, and the MSE of the maximum likelihood with fixed effects (m.l.e.f.e.) and the c.m.l.e. are shown to decrease with T as well as with the number of Bernoulli trials N in the Binomial distribution. Moreover, the advantage of the c.m.l.e. over the m.l.e.f.e. decreases with N and T. From the comparison of the MSEs it seems that for T as low as 5 (and $N = 5$) the values obtained with the two estimators are rather close. On the other hand, for $N = 2$ the value of T has to be at least 10 in order to render the difference between the two estimators negligible.

Lastly, the c.m.l.e. was applied to a new and original dataset that allowed the estimation of the probability of obtaining a patent. The results obtained with the c.m.l.e. and with the m.l.e.f.e. were very similar due to the big average number of patent applications (around 50).

Acknowledgements

I would like to thank Daniel A. Ackerberg, Manuel Arellano, Raquel Carrasco, Stephen Donald, Samuel Kortum, Kevin Lang, Michael H. Riordan, and Juan Ruiz, for their helpful comments. I also thank the detailed reports from two anonymous referees that contributed significantly to the improvement of the paper. I am also deeply grateful to Daniel Ackerberg for allowing me to use his workstation, and to Michele Cincera for

[22] The "R^2" corresponding to Table 5 is computed comparing the real number of successes in the data and the expected number of successes from the estimated model, i.e. $E(K_{it}) = (\widehat{p_{it}}) * N_{it}$ where $(\widehat{p_{it}})$ is the estimated logistic probability of success using both the c.m.l.e. $\hat{\beta}$ and the derived estimated fixed effects.

[23] The sample was restricted to firms with a total number of applications smaller than 400 in order to reduce the estimation time.

Table 6
Comparison between different estimators for different T

	Column 1			Column 1			Column 4			Column 4		
	$T = 5,\ n_{obs} = 405$			$T = 2,\ n_{obs} = 166$			$T = 5,\ n_{obs} = 405$			$T = 2,\ n_{obs} = 166$		
	m.l.e	m.l.e.f.e	c.m.l.e.	m.l.e	m.l.e.f.e	c.m.l.e.	m.l.e	m.l.e.f.e	c.m.l.e.	m.l.e	m.l.e.f.e	c.m.l.e.
Const.	0.225			0.479			0.2309			0.4683		
	(0.060)			(0.084)			(0.058)			(0.081)		
Trend	−0.035	−0.071	−0.070	−0.098	−0.116	−0.114	−0.0392	−0.0218	−0.022	−0.0946	0.0076	0.0069
	(0.021)	(0.022)	(0.023)	(0.052)	(0.059)	(0.059)	(0.021)	(0.020)	(0.025)	(0.05)	(0.076)	(0.078)
Year 92	−0.557	−0.791	−0.785				−0.5915	−0.5798	−0.575			
	(0.076)	(0.081)	(0.084)				(0.073)	(0.085)	(0.100)			
$R\&D_t$	−0.392	−0.456	−0.456	0.167	−1.939	−1.934	0.0551	1.2865	1.269	1.1136	0.3369	0.3070
	(0.074)	(0.312)	(0.327)	(0.053)	(0.653)	(0.648)	(0.424)	(0.522)	(0.676)	(0.427)	(1.054)	(1.110)
$R\&D_{t-1}$							0.2616	−1.6987	−1.689	−1.034	−1.144	−1.1241
							(0.580)	(0.499)	(0.689)	(0.500)	(0.784)	(0.296)
$R\&D_{t-3}$							−1.1496	−2.7625	−2.741	−1.254	−2.684	−2.635
							(0.382)	(0.327)	(0.469)	(0.480)	(0.813)	(0.556)
$R\&D_{t-5}$							0.7002	−0.2577	−0.252	1.629	1.017	0.9727
							(0.366)	(0.452)	(0.558)	(0.469)	(1.480)	(0.408)
$R\&D_{t-7}$							−0.1523	−0.9011	−0.892	0.0119	−0.8896	−0.8624
							(0.397)	(0.296)	(0.436)	(0.517)	(0.720)	(0.578)
$R\&D_{t-8}$							−0.1913	0.4785	0.471	−0.3300	−1.2037	−1.1779
							(0.275)	(0.294)	(0.442)	(0.375)	(0.746)	(0.832)
Likelihood	−1145.9	−810.0	−630.5	−621.0	−332.8	−167.5	−1134.2	783.5	−604.3	−604.9	−327.4	−162.2

providing me with the R&D expenditures data. I benefited from the financial support of a Alfred P. Sloan Dissertation fellowship, and from programs CIÊNCIA and PRAXIS (JNICT, Portugal). I also acknowledge the support from a grant from the Spanish Education and Culture Ministry grant (#PB98-0137), Fundación Empresa Pública and the National Institute of Drug Abuse (USA) (#1-RO1-DA08715-01). I would also like to thank the support from the staff at the Patents Office at the Universitat de Barcelona.

Appendix A.

A.1. Proof of Example 1

Rewrite (7) as

$$\bar{n}_1 + \bar{n}_2 + n_3 = n_1 - \frac{4n_1}{3 + \sqrt{1 + 8\exp(\beta)}} + n_2 - \frac{2n_2}{2 + \exp(\beta) + \sqrt{\exp(2\beta) + 8\exp(\beta)}} + n_3, \tag{A.1}$$

where $\bar{n}_s$ is the number of firms that had s successes and at least one occurred in the first period. By using the definition $n_K = \bar{n}_K + n_K^*$, $K = 1, 2$, rewrite (A.1) as

$$n_2^* + n_1^* - \frac{4n_1}{3 + \sqrt{1 + 8\exp(\beta)}} - \frac{2n_2}{2 + \exp(\beta) + \sqrt{\exp(2\beta) + 8\exp(\beta)}} = 0. \tag{A.2}$$

Denote by β_0 the true parameter. The following table gives the joint probability of all possible events for a given firm i:

K	$K_{i1} = 0$	$K_{i1} = 1$
$K_{i2} = 0$	$\frac{1}{(1 + \exp(\beta_0 + \tau_i))(1 + \exp(\tau_i))^2}$	$\frac{\exp(\beta_0 + \tau_i)}{(1 + \exp(\beta_0 + \tau_i))(1 + \exp(\tau_i))^2}$
$K_{i2} = 1$	$\frac{2\exp(\tau_i)}{(1 + \exp(\beta_0 + \tau_i))(1 + \exp(\tau_i))^2}$	$\frac{2\exp(\beta_0 + 2\tau_i)}{(1 + \exp(\beta_0 + \tau_i))(1 + \exp(\tau_i))^2}$
$K_{i2} = 2$	$\frac{\exp(2\tau_i)}{(1 + \exp(\beta_0 + \tau_i))(1 + \exp(\tau_i))^2}$	$\frac{\exp(\beta_0 + 3\tau_i)}{(1 + \exp(\beta_0 + \tau_i))(1 + \exp(\tau_i))^2}$

Using the law of large numbers:

$$\frac{n_1^*}{I} \to \lim_{I\to\infty} \frac{1}{I} \sum_{i=1}^{I} \Pr(K_{i1} = 0 \cap K_{i2} = 1) = \lim_{I\to\infty} \frac{1}{I} \sum_{i=1}^{I} \frac{2\exp(\tau_i)}{(1 + \exp(\beta_0 + \tau_i))(1 + \exp(\tau_i))^2} = A_1, \tag{A.3}$$

$$\frac{n_2^*}{I} \to \lim_{I\to\infty} \frac{1}{I}\sum_{i=1}^{I} \Pr(K_{i1}=0 \cap K_{i2}=2)$$

$$= \lim_{I\to\infty} \frac{1}{I}\sum_{i=1}^{I} \frac{\exp(2\tau_i)}{(1+\exp(\beta_0+\tau_i))(1+\exp(\tau_i))^2} = A_2, \tag{A.4}$$

$$\frac{n_1}{I} \to \lim_{I\to\infty} \frac{1}{I}\sum_{i=1}^{I} (\Pr(K_{i1}=0 \cap K_{i2}=1)$$

$$+ \Pr(K_{i1}=1 \cap K_{i2}=0)) = \exp(\beta_0)\frac{A_1}{2} + A_1, \tag{A.5}$$

$$\frac{n_2}{I} \to \lim_{I\to\infty} \frac{1}{I}\sum_{i=1}^{I} (\Pr(K_{i1}=1 \cap K_{i2}=1)$$

$$+ \Pr(K_{i1}=0 \cap K_{i2}=2)) = 2\exp(\beta_0)A_2 + A_2. \tag{A.6}$$

Putting everything together (A.2) converges to

$$\underbrace{A_1\left[\frac{-2\exp(\beta_0)-1+\sqrt{1+8\exp(\hat{\beta})}}{3+\sqrt{1+8\exp(\hat{\beta})}}\right]}_{P1}$$

$$\underbrace{+A_2\left[\frac{\exp(\hat{\beta})+\sqrt{\exp(2\hat{\beta})+8\exp(\hat{\beta})}-4\exp(\beta_0)}{2+\exp(\hat{\beta})+\sqrt{\exp(2\hat{\beta})+8\exp(\hat{\beta})}}\right]}_{P2} = 0. \tag{A.7}$$

It can easily be proved that $\hat{\beta} = \beta_0$ is never a solution to Eq. (A.7) unless $\beta_0 = 0$. Suppose the opposite is true, i.e. $\hat{\beta} = \beta_0 \neq 0$, then the sign of $P1$ and $P2$, given by their numerators, coincide always, which implies that expression (A.7) can never be zero:

$$P1 \gtrless 0 \;\Leftrightarrow\; -2\exp(\beta_0) - 1 + \sqrt{1+8\exp(\beta_0)} \gtrless 0 \;\Leftrightarrow\; \beta_0 \lesseqgtr 0, \tag{A.8}$$

$$P2 \gtrless 0 \;\Leftrightarrow\; -3\exp(\beta_0) + \sqrt{\exp(2\beta_0)+8\exp(\beta_0)} \gtrless 0 \;\Leftrightarrow\; \beta_0 \lesseqgtr 0. \tag{A.9}$$

If $\beta_0 = 0$ then $\hat{\beta} = \beta_0 = 0$ is the unique solution.[24] The MLE is, therefore, inconsistent unless the true parameter is $\beta_0 = 0$. Furthermore, because both $P1$ and $P2$ are increasing in $\hat{\beta}$ the MLE underestimates the true β_0 whenever $\beta_0 < 0$ and overestimates β_0 whenever $\beta_0 > 0$. In fact, it is straightforward to show that the bias obtained in this example is $1 < \hat{\beta}/\beta_0 < 2$, which is smaller than the bias obtained in the logit case $\hat{\beta}/\beta_0 = 2$ (Andersen, 1973; Chamberlain, 1980).

A.2. Proofs of the c.m.l.e. asymptotic properties

A.2.1. Consistency

Andersen (1970) postulates three sufficient conditions for consistency of a conditional maximum likelihood estimator (Assumptions A.1–A.3 below). In order to prove the consistency of the c.m.l.e. it suffices, therefore, to show that under the conditions stated on Theorems 1 and 2 the conditional likelihood function satisfies the three assumptions.

Call $\mathcal{T}$ the vector of sufficient statistics for the incidental parameters $\tau_1, \ldots, \tau_I$. Andersen starts by assuming that $\mathcal{T}$ is independent of β. This assumption is satisfied as it is shown in Section 3. Denote by $\mathbf{t}$ a realization of the vector $\mathcal{T}$. β_0 is the true value of the structural parameter. Assume that all (relevant) τ_i's $\in \Omega_0$ and denote by Θ the range space of β. Denote by $P_{\beta_0,\tau}$ the probability measure of K for the true parameters β_0 and τ.

Assumption A.1. $\log \phi(K|\beta, \mathbf{t})$ is a differentiable function of β and there exists a set $\mathbf{B}$ of values $\mathbf{t}$ with $P_{\beta_0,\tau}(\mathcal{T} \in \mathbf{B}) > 0$ for all τ and an open cube Θ_0 containing the true parameter β_0, such that for any $\mathbf{t} \in \mathbf{B}$ the functions $\phi(K|\beta, \mathbf{t})$ and $\phi(K|\beta', \mathbf{t})$ are not identical for any pair $\beta \in \Theta_0$, $\beta' \in \Theta_0$.

Assumption A.2 includes a slight modification to Andersen's (1970) Assumptions 1.2 and 2.2. but, as it can be easily confirmed, his consistency proof carries through with this modification. To be precise, Andersen's 1.2. and 2.2. hold for every I while A.2 has to hold only for $I \geqslant m$, i.e. for I big enough.

Assumption A.2. The maximum-likelihood equation

$$\sum_{i=1}^{I} \frac{\partial \log \phi_i(K_i|\beta, \mathbf{t}_i)}{\partial \beta} = 0 \tag{A.12}$$

[24] Notice that if $\beta_0 = 0$, $P1$ and $P2$ are always strictly positive for $\hat{\beta} > 0$ and strictly negative for $\hat{\beta} < 0$. Therefore, the only way the equation $P1 + P2 = 0$ is when both terms are equal to zero, which happens for $\hat{\beta} = 0$.

Proof. The sign of $P1$ in the case where $\beta_0 = 0$ depends on

$$-3 + \sqrt{1 + 8\exp(\hat{\beta})} \gtrless 0 \quad \Leftrightarrow \quad 1 + 8\exp(\hat{\beta}) \gtrless 9 \quad \Leftrightarrow \quad \hat{\beta} \gtrless 0, \tag{A.10}$$

the sign of $P2$ depends on the following:

$$\exp(\hat{\beta}) + \sqrt{\exp(2\hat{\beta}) + 8\exp(\hat{\beta})} - 4 \gtrless 0 \quad \Leftrightarrow \quad 16\exp(\hat{\beta}) \gtrless 16 \quad \Leftrightarrow \quad \hat{\beta} \gtrless 0, \tag{A.11}$$

has for all $I \geqslant m$ and almost all values of $(\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_I)$ a unique solution $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_m) \in \Theta$.

Assumption A.3. Assume that

$$\sum_{i=1}^{\infty} \frac{\sigma^2(\delta, \tau_i)}{i^2} < \infty \tag{A.13}$$

for all δ in an open interval enclosing 0 where[25]

$$\sigma^2(\delta, \tau_i) = var_{\beta_0, \tau}(\log \phi_i(K_i|\beta_0 + \delta, \mathscr{T}_i) - \log \phi_i(K_i|\beta_0, \mathscr{T}_i)). \tag{A.14}$$

The validity of Assumptions A.1 and A.2 is shown in (1) below and the validity of A.3 is shown in (2).

(1) Assumption A.1 requires that the conditional density ϕ is not essentially equal for different β's in a neighborhood of the true parameter value β_0.[26] Steps 1.1 and 1.2 prove that under the conditions of Theorems 1 and 2, $\log \phi$ is strictly concave for the single and the multiparameter case, respectively. The strict concavity property insures the validity of A.1 and A.2.

(1.1) Strict concavity, the single parameter case.

Write the individual conditional likelihood function for some individual i as a function of the single parameter β as

$$l_i = \frac{A_i \, e^{c_i \beta}}{\sum_{z \in Z_i} A_{iz} e^{c_{iz} \beta}} = \frac{w_i}{\sum_{z \in Z_i} w_{iz}}, \tag{A.15}$$

where $A_i = \prod_{t=1}^{T} C_{K_{it}}^{N_{it}}$, $c_i = \sum_{t=1}^{T} K_{it} X_{it}$ and $A_{iz} = \prod_{t=1}^{T} C_{z_t}^{N_{it}}$, $c_{iz} = \sum_{t=1}^{T} z_t X_{it}$ with $z = (z_1, \ldots, z_T) \in Z_i$. Z_i represents the set of combinations of integers $z_1, \ldots, z_T$ that satisfy $\sum_{t=1}^{T} z_t = \sum_{t=1}^{T} K_{it}$ and $0 \leqslant z_t \leqslant N_{it}$, $\forall t$. l_i lies in the interval $[0, 1]$. Note that if $|\beta_0| = \infty$, l_i's would be constant and equal to 1 for all β and all i, therefore, invalidating A.1.

Now consider the case of the joint conditional likelihood L. In the limit L will converge to:

$$L = \prod_{i}^{I} l_i = \prod_{i}^{I} \frac{1}{\sum_{z \in Z_i} \frac{w_{iz}}{w_i}} \overset{\beta \to +\infty}{\to} \left\{ \begin{array}{ll} 1 & \text{if } c_i = \max_{z \in Z_i} c_{iz} \text{ for all } i \\ 0 & \text{otherwise} \end{array} \right\} \tag{A.16}$$

and

$$L = \prod_{i}^{I} l_i \overset{\beta \to -\infty}{\to} \to \left\{ \begin{array}{ll} 1 & \text{if } c_i = \min_{z \in Z_i} c_{iz} \text{ for all } i \\ 0 & \text{otherwise} \end{array} \right\}. \tag{A.17}$$

[25] Notice that Andersen (1970) provides a stronger and simpler alternative to Assumption A.3 which simply requires that all τ_i's belong to a compact set Ω_0 and the continuity of function $\sigma^2(\delta, \tau_i)$. This alternative version is also verified in this case under the conditions of Theorems 1 and 2. The proof is simple. Since both ϕ_i and $\log \phi_i$ are continuous in τ_i for all $\tau_i \in \Omega_0$ and K is a discrete variable, then these variances are weighted sums of continuous functions and, therefore, also continuous.

[26] Under certain conditions, however, ϕ can attain the same value irrespective of β. This is the case when K can be perfectly inferred from $\mathscr{T}$, i.e. when $\mathscr{T}_i = \sum_{t=1}^{T} N_{it}$ or $\mathscr{T}_i = 0$ for every i. Observations with these characteristics are left out of the estimation. The other circumstance in which ϕ is constant in β is when the variables in X are constant.

Hence, a sufficient condition for the existence of at least one maximum (or a necessary condition for a single maximum) is that there is at least one i and j, where i and j may be equal, for which $c_i \neq \min_{z\in Z} c_{iz}$ and $c_j \neq \max_{z\in Z_i} c_{jz}$.[27] Next it will be proved that

$$\frac{\partial^2 \log L}{\partial \beta^2} = \sum_{i=1}^{I} \frac{\partial^2 \log l_i}{\partial \beta^2} < 0, \tag{A.18}$$

i.e. the function $\log L$ is strictly concave and, therefore, if there is a maximum, it is unique.

Consider then the case of the individual conditional likelihood function (A.15) for individual i. Taking logs and differentiating with respect to β one obtains

$$\frac{\partial \log l_i}{\partial \beta} = c_i - \frac{\sum_{z\in Z_i} A_{iz} c_{iz} e^{c_{iz}\beta}}{\sum_{z\in Z_i} A_{iz} e^{c_{iz}\beta}}. \tag{A.19}$$

The second derivative of $\log l_i$ with respect to β is

$$\frac{\partial^2 \log l_i}{\partial \beta^2} = -\frac{(\sum_{z\in Z_i} A_{iz} c_{iz}^2 e^{c_{iz}\beta})(\sum_{z\in Z_i} A_{iz} e^{c_{iz}\beta}) - (\sum_{z\in Z_i} A_{iz} c_{iz} e^{c_{iz}\beta})^2}{(\sum_{z\in Z_i} A_{iz} e^{c_{iz}\beta})^2}. \tag{A.20}$$

The sign of (A.20) depends on the sign of its numerator which after some calculus simplifies to an expression that is always non-positive:

$$-\frac{1}{2}\sum_{z\in Z_i}\sum_{y\in Z_i} (c_{iz} - c_{iy})^2 A_{iz} A_{iy} e^{c_{iz}\beta} e^{c_{iy}\beta} \leqslant 0. \tag{A.21}$$

So far it was proven that $\partial^2 \log L/\partial\beta^2 \leqslant 0$. For strict concavity of the function $\log L$ it is enough to have $\partial^2 \log l_i/\partial\beta^2$ strictly negative for at least one i. Next, conditions under which (A.21) (and, therefore, also $\partial^2 \log l_i/\partial\beta^2$) is strictly negative are derived. For (A.21) to be strictly negative it is clear that it cannot be the case that $c_{iz} = c_{iy} \Leftrightarrow z'X_i = y'X_i$ for all $z, y \neq z \in Z_i$. Three necessary conditions to exclude these cases are: (1) that the variable X_{it} is not a constant. (2) that the set Z_i is not singleton (which holds for $\bar{K}_i \neq \{0, \sum_t N_{it}\}$). And (3) $N_{it} > 0$ for at least two values of t where the X_{it} differ. But are these also sufficient conditions? The proof follows.[28]

Assume w.l.o.g. that $X_{i1} \neq X_{i2}$, it has to be shown that there exist two vectors $z, y \in Z_i$ such that $c_{iz} \neq c_{iy}$. To prove that, assume for now that a vector $z' = (z_1, z_2, \ldots, z_T) \in Z_i$ with $z_1 < N_{i1}$ and $z_2 > 0$ exists. Then take $y' = (z_1 + 1, z_2 - 1, \ldots, z_T)$ such that only the first two components differ from z. It is clear that if $z \in Z_i$ then y must also be an element of Z_i. It follows that $c_{iz} - c_{iy} = z'X_i - y'X_i = X_{i2} - X_{i1} \neq 0$. To complete the proof it is necessary to show that $z \in Z_i$. Proceeding by contradiction, suppose such a vector does not exist. Then, it must be that all vectors belonging to Z_i are of one (or all) of the following types: $a = (a_1 < N_{i1}, a_2 = 0, \ldots, a_T)$, $b = (b_1 = N_{i1}, b_2 = 0, \ldots, b_T)$, or $d = (d_1 = N_1, d_2 > 0, \ldots, d_T)$. Take a vector such as a. Because $\bar{K}_i > 0$ (by condition (2) above) and $N_{i1}, N_{i2} > 0$ (by condition (3) above) it must be that $\exists t_0 \neq 2$ (but

[27] Note that this condition is satisfied by Assumption (3) in Theorem 1 and, therefore L has a maximum for finite β.

[28] I am indebted to an anonymous referee for the shape of this proof.

could equal 1) such that a vector $\bar{z}$ can be constructed with $\bar{z}_{t_0} = a_{t_0} - 1$ and $\bar{z}_2 = 1$. It turns out that the new constructed vector $\bar{z}$ also belongs to Z_i which is a contradiction since $\bar{z}_1 < N_{i1}$ and $\bar{z}_2 > 0$. Now take a vector of type b. Given that $N_{i1}, N_{i2} > 0$ then $\bar{z}$ can be constructed such that $\bar{z}_1 = b_1 - 1$ and $\bar{z}_2 = 1$, $\bar{z}$ is, therefore, also a member of Z_i which is again a contradiction. Lastly, take a vector such as d, since $\bar{K}_i < \sum_t N_{it}$ (by condition (2) above) and $N_{it} > 0$ for at least two t's then $\exists t_1 \neq 1$ (which could be 2) such that a vector $\bar{z}$ with $\bar{z}_1 = d_1 - 1$ and $\bar{z}_{t_1} = d_{t_1} + 1$ can be constructed. Again, it is clear that $\bar{z} \in Z_i$ which is a contradiction. Note that Assumption (3) in Theorem 1 guarantees not only that there is at least one maximum but also that conditions (1)–(3) above hold for at least one i. □

It was proved that under the conditions of Theorem 1, Assumptions A.1 and A.2 are valid for the single parameter case.

(1.2) Strict concavity, the multiparameter case

In the multiparameter case $\beta \in \Re^m$ and c_i is a vector where the qth element is $\sum_{t=1}^{T} K_{it} X_{itq}$. Just like in the single parameter case, the $\log l_i$ function is monotone in β_q if $c_{iq} \in \{\min_{z \in Z_i} c_{iqz}, \max_{z \in Z_i} c_{iqz}\}$. So a sufficient condition to have at least one maximum is that there exists at least one (i, j) where j may equal i such that $c_{iq} \neq \min_{z \in Z_i} c_{iqz}$, and $c_{jq} \neq \max_{z \in Z_i} c_{jqz}$ for all $q = 1, \ldots, m$. Assuming that this condition holds, i.e. that there is at least one maximum, the proof of strict concavity of the function $\log L$ will consist in showing that under the assumptions of Theorem 2, in particular $rank(X) = m$, the Hessian matrix is definite negative. The Hessian matrix is given by the expression below:

$$\frac{\partial^2 \log L}{\partial \beta \partial \beta'} = \sum_i \frac{\partial^2 \log l_i}{\partial \beta \partial \beta'}$$

$$= -\frac{1}{2} \sum_i \frac{\sum_{z \in Z_i} \sum_{y \in Z_i} (X_i'(z-y)(z-y)'X_i) A_{iz} A_{iy} e^{c_{iz}\beta} e^{c_{iy}\beta}}{(\sum_{z \in Z_i} A_{iz} e^{c_{iz}\beta})^2} \tag{A.22}$$

and one can easily see that the individual Hessians are negative semidefinite. Now, pick, for each i, a combination of vectors z, y belonging to Z_i such that $y = (z_1 + 1, z_2 - 1, \ldots, z_T)$. As it was proven for the single parameter case one can always find such a combination in any set Z_i. Notice that although the particular vectors z, y are different for every i, its difference is common to all i, i.e. $z - y = (1, -1, 0, \ldots, 0)$. This means that summation over the elements of Z_i can be split into two terms. The first term consists of a combination z, y such that $z - y = (1, -1, 0, \ldots, 0)$ for each i and the second term consists of all other elements of Z_i. Denoting by $b_{iwv} = A_{iw} A_{iv} e^{c_{iw}\beta} e^{c_{iv}\beta} / (\sum_{z \in Z_i} A_{iz} e^{c_{iz}\beta})^2$ one can write the Hessian (A.22) as

$$\frac{\partial^2 \log L}{\partial \beta \partial \beta'} = -\frac{1}{2} \sum_i X_i'(z-y)(z-y)'X_i b_{izy}$$

$$-\frac{1}{2} \sum_i \sum_{w \in Z_i} \sum_{v \in Z_i} (X_i'(w-v)(w-v)'X_i) b_{iwv}. \tag{A.23}$$

Note that the first term in (A.23) is equivalent to

$$-\frac{1}{2}X'\tilde{Z}\tilde{Z}'X, \tag{A.24}$$

where $\tilde{Z}'$ is a $I \times (I \times T)$ matrix with elements:

$$\begin{bmatrix} (z-y)'\sqrt{b_{1zy}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (z-y)'\sqrt{b_{2zy}} & \mathbf{0} & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (z-y)'\sqrt{b_{Izy}} \end{bmatrix}. \tag{A.25}$$

It is clear that $rank(\tilde{Z}) = I$, which implies $rank(\tilde{Z}\tilde{Z}') = I$. On the other hand, the $rank(X) = m$ by assumption. It follows that for $m \leqslant I$ the matrix $X'\tilde{Z}\tilde{Z}'X$ is definite positive, i.e. (A.24) definite negative. The condition that $m \leqslant I$[29] means, in other words, that the number of structural parameters should not exceed the number of individuals in the sample. In reality, most panel datasets where one may want to apply the c.m.l.e. probably have $m \leqslant I$. Nevertheless, since Theorem 2 refers to the asymptotic behavior when $I \to \infty$, this condition is always met for large enough I and fixed m.

The second term in (A.23) $-\frac{1}{2}\sum_i \sum_{w \in Z_i} \sum_{y \in Z_i} (X_i'(w-v)(w-v)'X_i) b_{iwv}$ is obviously negative semidefinite. Since the Hessian is the sum of a negative definite matrix and a negative semi-definite matrix, it is negative definite. This proves strict concavity in the multiparameter case.

(2) Proof that Assumption A.3 is satisfied. Notice that $\sigma^2(\delta, \tau_i)$ satisfies the following inequality:

$$\begin{aligned} \sigma^2(\delta, \tau_i) &= var_{\beta_0, \tau_i}(\log \phi_i(K_i|\beta_0 + \delta, \mathcal{T}_i) - \log \phi_i(K_i|\beta_0, \mathcal{T}_i)) \\ &< 4 \max\{var_{\beta_0, \tau_i}(\log \phi_i(K_i|\beta_0 + \delta, \mathcal{T}_i)), \\ &\quad var_{\beta_0, \tau_i}(\log \phi_i(K_i|\beta_0, \mathcal{T}_i))\}. \end{aligned} \tag{A.27}$$

Furthermore, under the conditions of Theorem 1 (or 2), one knows from (13) that the function ϕ is well defined:

$$0 < \phi_i(K_i|\beta_0, \mathcal{T}_i = \bar{K}_i) < 1 \Rightarrow -\infty < \log \phi_i(K_i|\beta_0, \mathcal{T}_i) < 0. \tag{A.28}$$

[29] Alternatively, it is easy to see that the rank of the individual Hessians H_i is at least 1. Since $rank(H) \leqslant \sum_i rank(H_i) \in [I, \min\{mI, \frac{1}{2}\sum_i (\#Z_i \times \#Z_i - \#Z_i)\}]$, $I \geqslant m$ is a necessary condition for full rank (and therefore negative definiteness) of the Hessian matrix H regardless of the dimension of the sets Z_i's. For example in the logit case $\#Z_i = 2$ (for all relevant I's) this means that:

$$rank(H) \leqslant \sum_i rank(H_i) = I, \tag{A.26}$$

so that $I \geqslant m$ is a necessary condition for strict concavity in the logit case.

Since K_i is a discrete random variable, $var_{\beta_0,\tau_i}(\log \phi_i(K_i|\beta_0 + \delta, \mathscr{T}_i))$ and $var_{\beta_0,\tau_i}(\log \phi_i(K_i|\beta_0, \mathscr{T}_i))$ are finite sums. Therefore, the following can be established:

$$\sigma^2(\delta, \tau_i) < \infty \quad \text{for all } i \tag{A.29}$$

$$\sum_{i=1}^{I} \frac{\sigma^2(\delta, \tau_i)}{i^2} < \max_i \{\sigma^2(\delta, \tau_i)\} \sum_{i=1}^{I} \frac{1}{i^2} \tag{A.30}$$

and since from condition (2) in both Theorems 1 and 2 the sequence of τ_i's is bounded, it can be established that:

$$\Rightarrow \sum_{i=1}^{\infty} \frac{\sigma^2(\delta, \tau_i)}{i^2} < \lim_{I\to\infty} \left\{ \max_i \{\sigma^2(\delta, \tau_i)\} \sum_{i=1}^{I} \frac{1}{i^2} \right\}$$

$$< \max_{i\to\infty} 2\{\sigma^2(\delta, \tau_i)\} < \infty. \tag{A.31}$$

A.2.2. Asymptotic normality

Andersen's proof of asymptotic normality relies also on a set of sufficient conditions. In addition to Assumptions A.1–A.3, Andersen requires the validity of Assumptions A.4 and A.5 that follow.

Assumption A.4. The set of first, second and third partial derivatives of $\log \phi(K|\beta, \mathscr{T}(K))$ with respect to $\beta_1, \ldots, \beta_m$ exist for all β in an open cube Θ_0 enclosing β_0, and for all τ and $\beta \in \Theta_0$ the following holds: [30]

$$E_{\beta,\tau}\left(\frac{\partial \log \phi(K|\beta, \mathscr{T})}{\partial \beta_j}\right) = 0, \quad j = 1, \ldots, m \tag{A.32}$$

and

$$E_{\beta,\tau}\left(\frac{\partial^2 \log \phi(K|\beta, \mathscr{T})}{\partial \beta_j \partial \beta_p}\right) = -b^2_{jp}(\beta, \tau), \quad j, p = 1, \ldots, m. \tag{A.33}$$

There further exist positive integrable functions $h_{pq}(K)$ such that for all $\beta \in \Theta_0$

$$\left|\frac{\partial^3 \log \phi(K|\beta, \mathscr{T}(K))}{\partial \beta_j \partial \beta_p \partial \beta_q}\right| \leqslant h_{pq}(K), \quad j, p, q = 1, \ldots, m, \tag{A.34}$$

such that $E_{\theta_0,\tau}[h_{pq}(K)]$ and $var_{\beta_0,\tau}[h_{pq}(K)]$ are for all p and q $(p, q = 1, \ldots, m)$ continuous functions of τ.

Assumption A.5. For all K, $f(K|\beta, \tau)$ is continuous in τ, and for all p, $j = 1, \ldots, m$, $var_{\beta_0,\tau}[\partial^2 \log \phi(K|\beta_0, \mathscr{T})/\partial \beta_j \partial \beta_p]$, and $b^2_{jp}(\beta_0, \tau)$ are continuous functions of τ. In addition, $B^2(\beta_0, \tau)$ with elements $b^2_{jp}(\beta_0, \tau)$, $p, j = 1, \ldots, m$ is non-singular for all τ.

Proof. Assumption A.4 is valid under the assumptions of Theorems 1 and 2:

Eqs. (A.32) and (A.33) are the usual regularity conditions which are trivially satisfied because K is a finite, discrete random variable. The third cross derivatives

[30] The single parameter case is a special case for $m = 1$.

$\partial^3 \log L/\partial\beta_j\partial\beta_p\partial\beta_q$ are integrable because they are functions of a finite random vector K. Denote, by Θ the smallest compact set that contains Θ_0. Since $|\partial^3 \log L/\partial\beta_j\partial\beta_p\partial\beta_q|$ is a continuous function of β then one knows that there is at least one maximum of this continuous function in this compact set. Therefore, there exists a positive function h_{pq}:

$$\left|\frac{\partial^3 \log L}{\partial\beta_j\partial\beta_p\partial\beta_q}\right|_{\beta\in\Theta_0} \leqslant \max_{\beta\in\Theta}\left|\frac{\partial^3 \log L}{\partial\beta_j\partial\beta_p\partial\beta_q}\right| = h_{pq}(K), \tag{A.35}$$

where h_{pq} is an integrable function because K is a finite random variable. $E_{\beta_0,\tau}[h_{pq}(K)]$ and $var_{\beta_0,\tau}[h_{pq}(K)]$ are continuous functions of τ because they are finite sums of continuous functions.

Proof. Assumption A.5 is valid under the assumptions of Theorems 1 and 2:

$$f(K|\beta,\tau) = \prod_i C_{K_{i1}}^{N_{i1}} \cdots C_{K_{iT}}^{N_{iT}} \times \frac{e^{K_{i1}X_{i1}'\beta+\cdots+K_{iT}X_{iT}'\beta}}{(1+e^{X_{i1}'\beta+\tau_i})^{N_{i1}}\ldots(1+e^{X_{iT}'\beta+\tau_i})^{N_{iT}}} e^{\tau_i(K_{i1}+\cdots+K_{iT})} \tag{A.36}$$

is obviously a continuous function of τ for all K. $var_{\beta_0,\tau}[\partial^2 \log \phi(K|\beta_0,\mathcal{T})/\partial\beta_j\partial\beta_p]$ and $b_{jp}^2(\beta_0,\tau)$, are continuous and finite because they are finite sums of continuous functions for $|\beta_0| < \infty$, $\tau_i \in \Omega_0$ and the identifiability condition on X.

Regarding the matrix B^2, notice that $B^2(\beta_0,\tau)=\text{cov}_{\beta_0,\tau}(\partial \log \phi(K|\beta_0,\mathcal{T})/\partial\beta_j)$ because the regularity conditions are satisfied (Assumption A.4). This implies that $B^2(\beta_0,\tau)$ is semi-positive definite or $|B^2(\beta_0,\tau)| \geqslant 0$. The necessary and sufficient conditions for B^2 to be positive definite are the same as the ones discussed in step 1.2 of the consistency proof above and, therefore, they are satisfied under the conditions of Theorem 2.

A.3. Data appendix—the EPO data

The sample of firms, their sector of activity, and their R&D expenditures from 1980 to 1992 was kindly made available by Michele Cincera and is closely related to the sample used by Cincera (1997). Data on the number of patent applications and patents granted by firm was added for the years 1980 to 1992 from the EPO ESPACE Bulletin Vol. 1997/006 CD-ROM.[31] The 188 American firms that compose the clean data set operate in different activity sectors. The most represented sector is the computer industry with 17.02% of the observations and the least represented sector is Metals with only 0.53% of the observations.

The number of patent applications reveals an increasing trend, which is probably due to the proximity of the beginning of the sample period with the foundation of the EPO. The last 3 years of data, however, show a halt on this trend revealing the maturity of the European patent system. On the other hand, the percentage of successes declines during the last 3 years of data which may be a symptom of data truncation,[32] or a

[31] The EPO was founded in June 1, 1978 (Straus, 1997).

[32] Patents granted here means patents granted within at least 5 years after its application date. According to Cincera (1998), the EPO claims to take an average of 3 years to grant a patent.

symptom of the decreasing quality of applications due to decreasing application costs over time. In spite of these drawbacks the EPO data has advantages over the US Patent Office data. First, the availability of data on both the number of patent applications and patents granted at the firm level. Second, the EPO uses date of application instead of date of issue, which proxies better the innovation date. Third, the application cost difference (Straus, 1997) seems enough to avoid the "domestic market effect" (Archibugi and Pianta, 1992; Cincera, 1998), which consists in applying for more patents than the number of inventions justify in order to reduce foreign competition. For this reason, the applications to foreign Patent Offices may be a better indicator of technological progress.

References

Aigner, D., Hsiao, C., Kapteyn, A., Wansbeek, T., 1984. Latent variable models in econometrics. In: Griliches, Z. (Eds.), Handbook of Econometrics, Vol. II, Elsevier, Amsterdam (Chapter 23).

Andersen, E.B., 1970. Asymptotic properties of conditional maximum likelihood estimators. Journal of the Royal Statistical Society, Series B (32), 283–301.

Andersen, E.B., 1973. Conditional Inference and Models for Measuring. PhD dissertation, Mentalhygiejnisk Forlag, Copenhagen.

Archibugi, D., Pianta, M., 1992. Specialization and size of technological activities in industrial countries: the analysis of patent data. In: Scherer, F.M., Perlman, M. (Eds.), Entrepreneurship, Technological Innovation, and Economic Growth: Studies in the Schumpeterian Tradition. University of Michigan Press, Ann Arbor, pp. 65–85.

Arellano, M., 2000. Discrete Choices with Panel Data. Mimeo, CEMFI.

Chamberlain, G., 1980. Analysis of covariance with qualitative data. Review of Economic Studies 47, 225–238.

Cincera, M., 1997. Patents, R&D and technological spillovers at the firm level: some evidence from econometric count data. Journal of Applied Econometrics 12, 265–280.

Cincera, M., 1998. Technological and economic performance of international firms. Ph.D. Dissertation, Université Libre de Bruxelles.

Hausman, J., Hall, H.B., Zvi, G., 1984. Econometric models for count data with an application to the patents-R&D relationship. Econometrica 52 (4), 909–938.

Heckman, J.J., 1995. The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA.

Lancaster, T., 2000. The incidental parameter problem since 1948. Journal of Econometrics 95, 391–413.

Machado, M.P., 2001a. Dollars and performance: alcohol misuse in Maine. Journal of Health Economics 20 (4), 645–672.

Machado, M.P., 2001b. A consistent estimator for the binomial distribution in the presence of incidental parameters: an application to patent data, Fundación Empresa Pública, Documento de Trabajo 0102.

Mantel, H.J., Godambe, V.P., 1993. Estimating functions for conditional inference: many nuisance parameter case. Annals of Institute of Statistical Mathematics 45 (1), 55–67.

Neyman, J., Scott, E.L., 1948. Consistent estimates based on partially consistent observations. Econometrica 16 (1), 1–32.

Pfanzagl, J., 1993. On the consistency of conditional maximum likelihood estimators. Annals of Institute of Statistical Mathematics 45 (4), 703–719.

Reid, N., 1995. The roles of conditioning in inference. Statistical Science 10 (2), 138–199.

Reid, N., 2000. Likelihood. Journal of the American Statistical Association 95, 1335–1340.

Straus, J., 1997. The present state of the patent system in the European Union—as compared to the situation in the United States of America and Japan. European Commission EUR 17014 EN.

Wright, B.D., Douglas, G., 1976. Better procedures for sample-free item analysis. Research Memorandum 20, Statistical Laboratory, Department of Education, University of Chicago.