JOURNAL OF APPLIED ECONOMETRICS, VOL. 12, 203–223 (1997)

# COUNT DATA REGRESSION USING SERIES EXPANSIONS: WITH APPLICATIONS

A. COLIN CAMERON[a]* AND PER JOHANSSON[b]

[a] *Department of Economics, University of California, Davis, CA 95616, USA. E-mail: accameron@ucdavis.edu*
[b] *Department of Economics, Umea University, S-901 87 Umea, Sweden*

SUMMARY

A new class of parametric regression models for both under- and overdispersed count data is proposed. These models are based on squared polynomial expansions around a Poisson baseline density. The approach is similar to that for continuous data using squared Hermite polynomials proposed by Gallant and Nychka and applied to financial data by, among others, Gallant and Tauchen. The count models are applied to underdispersed data on the number of takeover bids received by targeted firms, and to overdispersed data on the number of visits to health practitioners. The models appear to be particularly useful for underdispersed count data. © 1997 by John Wiley & Sons, Ltd.

*J. Appl. Econ.*, **12**, 203–223 (1997)

No. of Figures: 5. No. of Tables: 7. No. of References: 20.

## 1. INTRODUCTION

Count data regression models are models for the special case where the dependent variable takes only non-negative integer values or counts. Overviews of standard models include Cameron and Trivedi (1986), Winkelmann (1994), and Gurmu and Trivedi (1994).

The benchmark Poisson model for count data imposes the restriction that the conditional variance equals the conditional mean. This restriction is usually rejected in economic applications. In the common case of overdispersion, i.e. the conditional variance exceeds the conditional mean, the negative binomial is widely used. For underdispersion, i.e. the conditional variance is less than the mean, the preferred treatment is less well established. The Katz system or GECk model (see King, 1989, and Winkelmann and Zimmermann, 1991) and the generalized Poisson (see Consul and Famoye, 1992) have the theoretical weakness that a restriction is placed on the range of values that the dependent variable can take. Furthermore, this range is determined by the parameter values, a departure from the usual assumptions made in establishing consistency. The double Poisson model of Efron (1986) involves an approximation so that the probabilities do not sum to exactly one. The hurdle model (see Mullahy, 1986) is another possible model for underdispersed data, but is not parsimonious as in typical applications the number of parameters to be estimated is doubled.

We present a new class of parametric models for count data, based on a squared polynomial expansion around any given discrete density. For underdispersed data these models provide a

* Correspondence to: Colin Cameron, Department of Economics, University of California, Davis, CA 95616-8578, USA.

Contract grant sponsor: Swedish Research Council for the Humanities and Social Sciences.

CCC 0883–7252/97/030203–21$17·50

*Received 4 May 1996*
*Revised 15 December 1996*

parsimonious model without restrictions on the range of the dependent count, while for overdispersed data these models provide an alternative to the negative binomial.

The approach is similar to that for continuous data, using a squared Hermite series expansion from a baseline normal density, developed by Gallant and Nychka (1987) and applied to finance data in many applications, beginning with Gallant and Tauchen (1989). Applications to discrete data include Gabler, Laisney, and Lechner (1993) for binary data and Gurmu, Rilstone, and Stern (1994) for count data. Gurmu *et al.* (1994) use an orthogonal series expansion from a baseline gamma density to model an unobserved heterogeneity term in a Poisson mixture model. This provides a sequence of models for overdispersed data which nests the Poisson and negative binomial but does not permit underdispersion. Here we consider series expansion for the count variable, rather than the heterogeneity term, and obtain a model that can fit both over- and underdispersed data.

The approach is particularly attractive when the data are underdispersed. An application studied here is the number of takeover bids (after the initial bid that made it a takeover target) received by firms that have been targeted for takeover. Jaggia and Thosar (1993) modelled a sample of 126 targeted firms. Their Poisson regression analysis confirmed *a priori* beliefs that the number of bids decreased the more attractive was the initial offer, and first increased and then decreased with firm size. There is no support, however, for the view that defensive actions taken by management are associated with a decrease in the number of bids. Re-analysis of the Jaggia and Thosar data reveals that the data are underdispersed, albeit mildly so. More importantly, the fitted Poisson model greatly overpredicts the probability of a firm receiving zero bids. The series expansion model proves to be capable of accommodating the underdispersion and predicting quite well the probability of zero bids.

A very common application of count regression models is to measures of health utilization such as number of doctor visits, with explanatory variables including various socio-economic variables, health status, and type of health insurance. Health-utilization counts are usually (if not always) overdispersed. We use data from Cameron *et al.* (1988) on the number of visits to non-doctor health professionals for a large sample of 5190 individuals. The data are quite overdispersed, and are very well fit by a negative binomial model. This provides a very competitive benchmark against which to compare the series expansion model. In particular, in applications to continuous time-series data, models based on squared polynomial expansions are not always parsimonious. For our application, the preferred series expansion model is a fifth-order model. This outperforms the negative binomial model, as does a fourth-order model. For this application successful modelling does not require too high an order expansion, though it is more parsimonious to use the negative binomial.

In Section 2 we present the model and its properties, with additional details provided in Appendices A and B at the end of the paper. A simulation study is presented in Section 3. Applications to both under- and overdispersed data are presented in Section 4. Both simulation and application use the Poisson density as the baseline density. Conclusions are presented in Section 5.

## 2. MODEL BASED ON SQUARED POLYNOMIAL EXPANSION

### 2.1. General Results

We begin with a general presentation for any type of data and baseline density, before specializing to count data with a Poisson density as a baseline. Consider a scalar random variable

 

$y$ with baseline density $f(y \mid \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is possibly a vector. Define the $p$th-order polynomial

$$h_p(y \mid \mathbf{a}) = \sum_{k=0}^{p} a_k y^k \tag{1}$$

where $\mathbf{a} = (a_0, a_1, \ldots, a_p)'$ and normalize $a_0 = 1$. The density based on a squared polynomial series expansion is

$$g_p(y \mid \boldsymbol{\lambda}, \mathbf{a}) = f(y \mid \boldsymbol{\lambda}) \cdot \frac{h_p^2(y \mid \mathbf{a})}{\eta_p(\boldsymbol{\lambda}, \mathbf{a})} \tag{2}$$

where $\eta_p(\boldsymbol{\lambda}, \mathbf{a})$ is a normalizing constant term that ensures that the density $g_p(y \mid \boldsymbol{\lambda}, \mathbf{a})$ sums to unity, and squaring the polynomial ensures that the density is non-negative. In Appendix A it is shown that

$$\eta_p(\boldsymbol{\lambda}, \mathbf{a}) = \sum_{k=0}^{p} \sum_{l=0}^{p} a_k a_l m_{k+l} \tag{3}$$

where $m_r \equiv m_r(\boldsymbol{\lambda})$ denotes the $r$th non-central moment of the baseline density $f(y \mid \boldsymbol{\lambda})$.

The moments of the random variable $y$ with density $g_p(y \mid \boldsymbol{\lambda}, \mathbf{a})$ are readily obtained from those of the baseline density $f(y \mid \boldsymbol{\lambda})$ as

$$E[y^r] = \frac{\sum_{k=0}^{p} \sum_{l=0}^{p} a_k a_l m_{k+l+r}}{\eta_p(\boldsymbol{\lambda}, \mathbf{a})} \tag{4}$$

(see Appendix A). The $r$th moment of $y$ will generally differ from the $r$th moment of the baseline density. In particular, this is the case for the mean.

### 2.2. Baseline Density Poisson (PPp model)

For the Poisson baseline density, $\lambda$ is a scalar and $f(y \mid \lambda)$ in equation (2) is given by

$$f(y \mid \lambda) = \frac{\mathrm{e}^{-\lambda} \lambda^y}{y!} \qquad y = 0, 1, 2, \ldots \tag{5}$$

The normalizing constant $\eta_p(\lambda, \mathbf{a})$ defined in equation (3) and the moments $E[y^r]$ defined in equation (4) are evaluated at the moments $m_r(\lambda)$ of the Poisson, which can be obtained from the moment-generating function using $m_r(\lambda) = \partial^r \exp(-\lambda + \lambda \mathrm{e}^t)/\partial t^r \mid_{t=0}$.

We call the model with density (2) and baseline density (5) the PPp model (for *Poisson Polynomial of order p*). As an example the PP2 model is

$$g_2(y \mid \lambda, \mathbf{a}) = \frac{\mathrm{e}^{-\lambda} \lambda^y}{y!} \cdot \frac{(1 + a_1 y + a_2 y^2)^2}{\eta_2(\lambda, \mathbf{a})} \tag{6}$$

where

$$\eta_2(\lambda, \mathbf{a}) = 1 + 2a_1 m_1 + (a_1^2 + 2a_2) m_2 + 2a_1 a_2 m_3 + a_2^2 m_4 \tag{7}$$

The first two moments of $y$ are

$$\begin{aligned} E[y] &= (m_1 + 2a_1 m_2 + (a_1^2 + 2a_2) m_3 + 2a_1 a_2 m_4 + a_2^2 m_5)/\eta_2(\lambda, \mathbf{a}) \\ E[y^2] &= (m_2 + 2a_1 m_3 + (a_1^2 + 2a_2) m_4 + 2a_1 a_2 m_5 + a_2^2 m_6)/\eta_2(\lambda, \mathbf{a}) \end{aligned} \tag{8}$$

 

Table I. Means ($E$) and dispersion ratios ($R$) of PP1 model for various parameter values

| | $E = 0{\cdot}5$ | | $E = 1$ | | $E = 5$ | |
|---|---|---|---|---|---|---|
| | $\lambda$ | $a_1$ | $\lambda$ | $a_1$ | $\lambda$ | $a_1$ |
| $R = 1{\cdot}0$ | 0·500 | 0·000 | 1·000 | 0·000 | 5·000 | 0·000 |
| $R = 0{\cdot}7$ | 0·075 | 2·148 | 0·342 | 1·182 | 3·290 | 6·115 |
| $R = 2{\cdot}0$ | 0·373 | −1·279 | 1·900 | −0·299 | 3·572 | −0·299 |

where evaluation of (7) and (8) requires the first six moments of the Poisson density

$$
\begin{aligned}
m_1 &= \lambda \\
m_2 &= \lambda + \lambda^2 \\
m_3 &= \lambda + 3\lambda^2 + \lambda^3 \\
m_4 &= \lambda + 7\lambda^2 + 6\lambda^3 + \lambda^4 \\
m_5 &= \lambda + 15\lambda^2 + 25\lambda^3 + 10\lambda^4 + \lambda^5 \\
m_6 &= \lambda + 31\lambda^2 + 90\lambda^3 + 65\lambda^4 + 15\lambda^5 + \lambda^6
\end{aligned}
$$

The PPp model permits a wide range of models for count data, including multimodal densities and densities with either under- or overdispersion. These possibilities are illustrated using the PP1 model.

Table I presents values of $\lambda$ and $a_1$ for the PP1 model that produce densities with combinations of mean ($E$) equal to 0·5, 1·0, and 5·0 and variance/mean ratio ($R$) equal to 0·7, 1·0, and 2·0. This illustrates the ability to model both over- and underdispersion, with equidispersion when $a_1 = 0$, and that the mean of $y$ differs from the mean of the baseline density when $a_1 \neq 0$.

These values of $\lambda$ and $a_1$, for given $E$ and $R$, are not unique. For example, when $E = 1$ and $R = 1$, possible values of $(\lambda, a_1)$ are the Poisson value of $(1, 0)$, and the PP1 value of approximately $(0{\cdot}276, -2{\cdot}532)$. This poses no identification problem as while the first two moments are the same, other moments such as the third will differ and the distribution will differ. This is illustrated in Figure 1, which presents the two different probability densities for these two

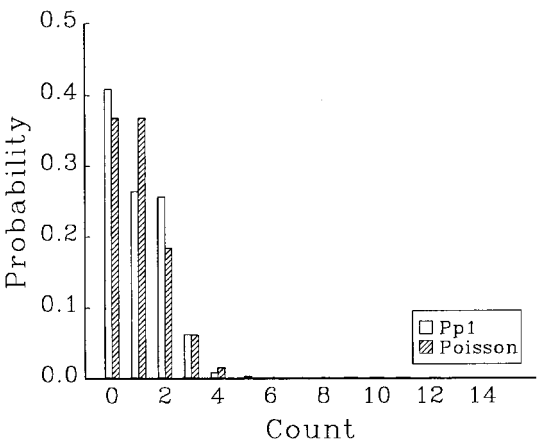


Figure 1. Frequency distributions for Poisson and PP1 ($\lambda = 0{\cdot}276, a_1 = -2{\cdot}525$) models with same mean $E = 1$ and variance–mean ratio $R = 1$

 

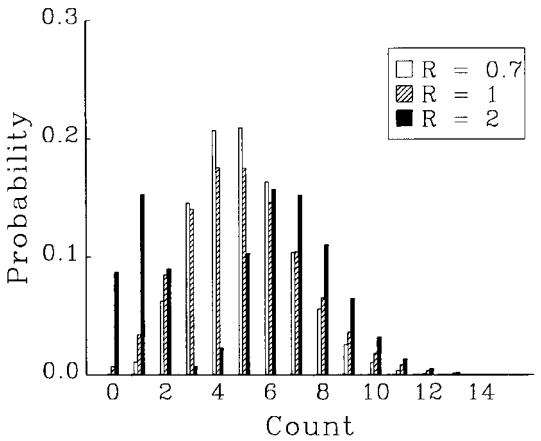


Figure 2. Frequency distributions for PP1 models with mean $E = 5$ and different variance–mean ratios

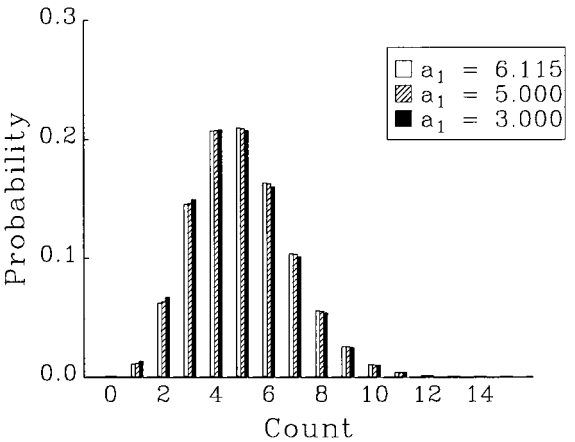


Figure 3. Frequency distributions for PP1 models with $\lambda = 3{\cdot}290$ and different values of $a_1$

cases with mean and variance equal to unity. The distribution is uniquely determined by a particular value of $(\lambda, a_1)$.

Figure 2 shows changes in the density as $(\lambda, a_1)$ changes to accommodate different departures from equidispersion, holding the mean fixed at 5 ($E = 5$). As expected, underdispersed data are more centred around the mean than the Poisson while the overdispersed data are more scattered. The overdispersion case $R = 2$ shows the possibility of a bimodal density, though overdispersed PP1 densities are not necessarily bimodal.

Figure 3 shows that the densities can at times change little with changes in $a_1$. The values $\lambda = 3{\cdot}290$ and $a_1 = 6{\cdot}115$ produce $R = 0{\cdot}7$ and $E = 5{\cdot}0$. Decreasing $a_1$ from 6·115 to 5·0 or even 3·0 makes little change to the density. In fact holding $\lambda = 3{\cdot}290$, letting $a_1 = 5{\cdot}0$ produces $R = 0{\cdot}703$ and $E = 4{\cdot}987$, while letting $a_1 = 3{\cdot}0$ produces $R = 0{\cdot}714$ and $E = 4{\cdot}943$. If in application such a flat spot in the density is encountered, one can expect imprecision in the separate estimation of $a_1$ and $\lambda$. This point is discussed further in Section 3.

 

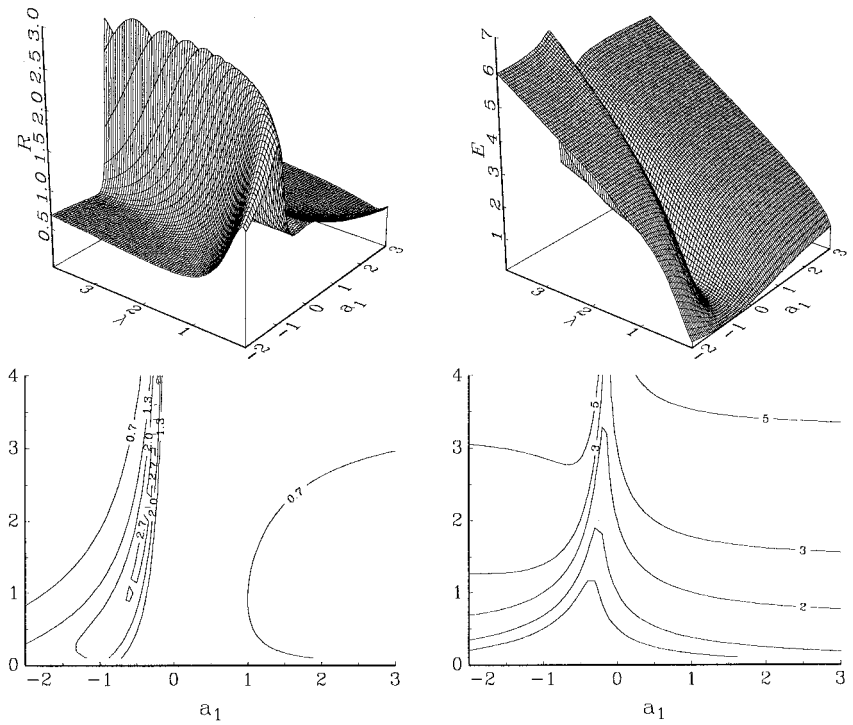


Figure 4. Variance–mean ratio ($R$) and mean ($E$) for PP1 models with $\lambda$ in [0·1, 4·0] and $a_1$ in $[-2, 3]$. (In the third panel, the first contour gives regions with $R < 0{\cdot}7$, the second gives $0{\cdot}7 \leqslant R < 1{\cdot}3$, etc. In the fourth panel, the first contour gives regions with $E < 0{\cdot}5$, the second gives $0{\cdot}5 \leqslant E < 1$, etc.)

Figure 4 plots both the variance–mean ratio and the mean for values of $\lambda$ in $[0{\cdot}1, 4]$ and $a_1$ in $[-2, 3]$. Overdispersion is obtained for values of $a_1$ less than zero but greater than a value that ranges from approximately $-0{\cdot}5$ for $\lambda = 4$ to $-2$ for $\lambda = 0{\cdot}1$. The mean is generally increasing in $\lambda$ and increasing in the absolute value of $a_1$. There is great variation in both $R$ and $E$ for values of $a_1$ close to zero. By contrast, for values of $a_1$ a considerable distance from zero and a fixed value of $\lambda$, both $R$ and $E$ are relatively invariant to changes in $a_1$. An example of this behaviour has already been presented in Figure 3.

We conclude that even the simplest generalization of the Poisson model, the PP1, is a quite flexible model for counts.

### 2.3. Estimation

We consider estimation based on a sample of independent observations $\{(y_1, X_1), \ldots, (y_N, X_N)\}$ of size $N$. Regressors are introduced by allowing the parameter to vary with regressors, while the polynomial coefficients $\mathbf{a}$ are unknown parameters that do not vary with regressors.

The parameter $\lambda_i$ is determined by a known function of regressors $\mathbf{X}_i$ and an unknown parameter vector $\boldsymbol{\beta}$

$$\lambda_i = \lambda(\mathbf{X}_i, \boldsymbol{\beta}) \tag{9}$$

 

The log-likelihood function is then

$$\ln \mathcal{L}(\boldsymbol{\beta}, \mathbf{a}) = \sum_{i=1}^{N} \{\ln f(y_i \mid \lambda(\mathbf{X}_i, \boldsymbol{\beta})) + \ln h_p(y_i \mid \mathbf{a})^2 - \ln \eta_p(\lambda(\mathbf{X}_i, \boldsymbol{\beta}), \mathbf{a})\} \tag{10}$$

with first-order conditions

$$\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ \frac{\partial \ln f(y_i \mid \lambda_i)}{\partial \lambda_i} - \frac{\partial \eta_p(\lambda_i, \mathbf{a})}{\partial \lambda_i} \frac{1}{\eta_p(\lambda_i, \mathbf{a})} \right\} \frac{\partial \lambda_i}{\partial \boldsymbol{\beta}}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \mathbf{a}} = \sum_{i=1}^{n} \left\{ 2 \frac{\partial \ln h_p(y_i \mid \mathbf{a})}{\partial \mathbf{a}} - \frac{\partial \eta_p(\lambda_i, \mathbf{a})}{\partial \mathbf{a}} \frac{1}{\eta_p(\lambda_i, \mathbf{a})} \right\}$$

which given equation (3) become

$$\begin{aligned} \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{N} \left\{ \frac{\partial \ln f(y_i \mid \lambda_i)}{\partial \lambda_i} - \frac{\sum_{k=0}^{p} \sum_{l=0}^{p} a_k a_l \partial m_{k+l,i} / \partial \lambda_i}{\sum_{k=0}^{p} \sum_{l=0}^{p} a_k a_l m_{k+l,i}} \right\} \frac{\partial \lambda_i}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ln \mathcal{L}}{\partial a_j} &= \sum_{i=1}^{N} 2 \left\{ \frac{y^j}{\sum_{k=0}^{p} a_k y^k} - \frac{\sum_{k=0}^{p} a_k m_{k+j,i}}{\sum_{k=0}^{p} \sum_{l=0}^{p} a_k a_l m_{k+l,i}} \right\} \qquad j = 1, \ldots, p \end{aligned} \tag{11}$$

Consider the PPp model, i.e. the baseline density is specified as the Poisson, with the usual Poisson regression parameterization of the mean

$$\lambda_i = \exp(\mathbf{X}_i' \boldsymbol{\beta}) \tag{12}$$

Then the first term in equation (11) simplifies to

$$\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \left\{ y_i - \frac{\sum_{k=0}^{p} \sum_{l=0}^{p} a_k a_l m_{k+l+1,i}}{\sum_{k=0}^{p} \sum_{l=0}^{p} a_k a_l m_{k+l,i}} \right\} \mathbf{X}_i \tag{13}$$

(see Appendix A) while the second term does not simplify.

Using equation (4) with $r = 1$ and equation (3), (13) can be re-expressed as

$$\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} (y_i - E[y_i \mid \mathbf{X}_i]) \mathbf{X}_i \tag{14}$$

Thus the residual is orthogonal to the regressors, and the residuals sum to zero if an intercept term is included in the model.[1] These properties do not hold for other generalizations of the Poisson such as the negative binomial.

[1] This result holds more generally when the baseline density is a linear exponential family density with conditional mean function corresponding to choosing the canonical link function.

 

By standard results for ML estimation the MLE for $\boldsymbol{\beta}$ and $\mathbf{a}$ is asymptotically normal distributed with variance matrix the inverse of the information matrix, under the assumption that the data are generated by equations (2) and (12). Note that we do not consider the technically more difficult question of whether by letting $p \to \infty$ as $N \to \infty$ the PPp model can approximate any model arbitrarily well.

As is common for many non-linear models, the likelihood function can have multiple optima. To increase the likelihood that a global maximum is obtained we follow Horowitz (1992) and use fast simulated annealing (Szu and Hartley, 1987), a variation on simulated annealing (cf. Goffe *et al.*, 1994), to obtain parameter estimates close to the global optima which are used as starting values for standard gradient methods. The computational methods used are detailed in Appendix B.

### 2.4. Testing and Model Evaluation

For continuous time-series data, Hall (1990) proposed Lagrange multiplier (LM) tests of normality against a squared Hermite polynomial expansion. The null hypothesis of normality implies that all components of $\mathbf{a}$ except $a_0$ equal zero, i.e. $\mathbf{a} = \mathbf{e}$, where $\mathbf{e} = (1\ 0\ 0 \ldots 0)'$. Hall (1990, p. 419) noted, however, that under this null hypothesis there are linear dependencies among components of the score vector.

A similar situation arises here in the count data setting where we consider LM tests of the null hypothesis that the data are Poisson, against the alternative of a squared polynomial expansion. For the PPp model with $\lambda_i = \exp(\mathbf{X}_i'\boldsymbol{\beta})$, under $H_0$:$\mathbf{a} = \mathbf{e}$, equations (11) reduce to

$$\begin{aligned} \left.\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}}\right|_{\mathbf{a}=\mathbf{e}} &= \sum_{i=1}^{N} (y_i - \lambda_i)\mathbf{X}_i \\ \left.\frac{\partial \ln \mathcal{L}}{\partial a_j}\right|_{\mathbf{a}=\mathbf{e}} &= \sum_{i=1}^{N} 2(y^j - m_{j,i}) \qquad j = 1, \ldots, p \end{aligned} \tag{15}$$

There is clearly a problem when $\mathbf{X}_i$ includes an intercept term with coefficient $\beta_0$. Then for the derivative with respect to the first term in the polynomial series expansion we have

$$\left.\frac{\partial \ln \mathcal{L}}{\partial a_1}\right|_{\mathbf{a}=\mathbf{e}} = \left.\frac{\partial \ln \mathcal{L}}{\partial \beta_0}\right|_{\mathbf{a}=\mathbf{e}} = \sum_{i=1}^{N} (y_i - \lambda_i)$$

so that these derivatives with respect to different parameters are identical. Following Hall (1990) we therefore drop this first term. The second and higher terms are simply tests of the second and higher raw moments. Thus, for example, the test based on $\partial \ln \mathcal{L}/\partial a_2 \,|_{\mathbf{a}=\mathbf{e}}$ is a test of whether $E[y_i^2 \,|\, \mathbf{X}_i] = \lambda_i^2 + \lambda_i$ and is clearly related to the usual test of overdispersion or underdispersion which is a test of $E[((y_i - \lambda_i)^2 - y_i) \,|\, \mathbf{X}_i] = 0$.

Wald and LR tests of statistical significance of $\mathbf{a}$ and $\boldsymbol{\beta}$ can also be performed. For Wald tests three different estimates of the variance matrix might be used. Let $\mathbf{A} = \Sigma_{i=1}^{N} \partial^2 g(y)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}' \,|_{\hat{\boldsymbol{\theta}}}$ and $\mathbf{B} = \Sigma_{i=1}^{N} (\partial g(y)/\partial\boldsymbol{\theta}) \cdot (\partial g(y)/\partial\boldsymbol{\theta}') \,|_{\hat{\boldsymbol{\theta}}}$, where $g(y)$ is the PPp density and $\boldsymbol{\theta} = (\boldsymbol{\beta}'\mathbf{a}')'$. Then the simplest estimate of the variance matrix is the outer-product (OP) form $\mathbf{B}^{-1}$, another standard estimate is the Hessian (H) form $-\mathbf{A}^{-1}$, and a measure robust to model misspecification is the sandwich (S) form $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$.

 

A standard model-selection criterion is the Bayesian information criterion, BIC $= -2 \ln \mathcal{L} + \ln(N) \cdot \dim(\boldsymbol{\theta})$, which is used by, for example, Gallant and Tauchen (1989).[2] We also consider the closeness between the actual and fitted distributions. Let $\hat{p}_{ji} = g(y_j \mid \mathbf{X}_i, \hat{\boldsymbol{\theta}})$ denote the predicted probability that $y_i$ equals $j$. It is useful to compare the average of these predicted probabilities, $\overline{\hat{p}_j} = (1/N)\Sigma_{i=1}^{N} \hat{p}_{ji}$, to $\bar{p}_j$, the fraction of the observations $y_i$ that take value $j$. A summary statistic is the sum of absolute differences $\Sigma_{j=0}^{\max(y_j)} \mid \bar{p}_j - \overline{\hat{p}_j} \mid$. These measures are particularly useful when comparing non-nested models.

## 3. SIMULATION

The following simulation demonstrates the major points that emerged from a much wider range of simulations. The data are generated from a PP1 model with one regressor. Specifically, $\lambda_i = \exp(\beta_0 + \beta_1 x_i)$ where $x_i$ is generated from the uniform distribution on the unit interval and the same draw $x_1, \ldots, x_N$ is used in all simulations. The parameters $a_1$ and $\beta_0$ are chosen so that $y_i$ has mean 1·0 and variance 1·0, 0·7 or 2·0 in the i.i.d. model (i.e. $\beta_1 = 0$). These values are, respectively, $(a_1, \beta_0) = (0, 0)$, $(1{\cdot}182, -1{\cdot}074)$ and $(-0{\cdot}299, 0{\cdot}642)$, and correspond to, respectively, equidispersion, underdispersion and overdispersion. Simulations from the model with $a_1 = 0$ (so $R = 1$) can be used to test size properties of tests for Poisson while other values are useful for power. The parameter $\beta_1$ is set either to 0, to investigate size, or 0·5, to investigate power. All simulations use the sample size $N = 200$ and are performed $s = 1000$ times. The parameters to be estimated are $a_1$, $\beta_0$ and $\beta_1$, where $\beta_0 = \ln \lambda$ when $\beta_1 = 0$ since then $\lambda = \exp(\beta_0)$.

The bias and MSE for the PP1 MLE in the i.i.d. case ($\beta_1 = 0$) are presented in the first three rows of Table II. There is very little bias in estimating the slope coefficient $\beta_1$ which is found to be very close to its true value of zero. When the data-generating process (dgp) also includes a regressor ($\beta_1 = 0{\cdot}5$), from rows four to six of Table II the bias in estimating $\beta_1$ increases but is still small, while the bias in estimating $a_1$ and $\beta_0$ falls in the overdispersed case. The last two rows consider Poisson estimation when the dgp is the Poisson ($a_1 = 0$). Compared to the PP1 estimates (in rows one and four) we see that for the slope coefficient $\beta_1$ there is no penalty in estimating the

Table II. Bias and MSE of PP1 MLE (first six rows) and Poisson MLE (last two rows) for Poisson ($R = 1$) and PP1 ($R = 0{\cdot}7$ or 2) dgp's. ($N = 200, s = 1000$)

| | True values | | | Bias | | | MSE | | |
|---|---|---|---|---|---|---|---|---|---|
| $R$ | $a_1$ | $\beta_0$ | $\beta_1$ | $a_1$ | $\beta_0$ | $\beta_1$ | $a_1$ | $\beta_0$ | $\beta_1$ |
| 1·0 | 0·000 | 0·000 | 0·000 | −0·055 | 0·044 | 0·008 | 0·092 | 0·154 | 0·067 |
| 0·7 | 1·182 | −1·074 | 0·000 | 0·163 | −0·064 | −0·008 | 0·921 | 0·247 | 0·085 |
| 2·0 | −0·299 | 0·642 | 0·000 | −0·087 | 0·210 | −0·007 | 0·028 | 0·162 | 0·065 |
| 1·0 | 0·000 | 0·000 | 0·500 | −0·055 | 0·065 | −0·030 | 0·080 | 0·133 | 0·060 |
| 0·7 | 1·182 | −1·074 | 0·500 | 0·173 | −0·081 | 0·014 | 1·078 | 0·287 | 0·112 |
| 2·0 | −0·299 | 0·642 | 0·500 | −0·020 | 0·068 | −0·079 | 0·006 | 0·051 | 0·077 |
| 1·0 | 0·000 | 0·000 | 0·000 | — | 0·004 | 0·009 | — | 0·020 | 0·061 |
| 1·0 | 0·000 | 0·000 | 0·500 | — | 0·008 | −0·002 | — | 0·020 | 0·051 |

[2] This is a variant of the Akaike information criteria (AIC) that has a penalty for additional parameters greater than the AIC $= -2 \ln \mathcal{L} + \dim(\boldsymbol{\theta})$ but not as high as some other variants such as the consistent AIC which uses $-2 \ln \mathcal{L} + (1 + \ln N) \cdot \dim(\boldsymbol{\theta})$.

 

Table III. Size and power using outer-product (OP), Hessian (H) and sandwich (S) variance estimates of PP1 MLE (first six rows) and Poisson MLE (last two rows) for Poisson ($R = 1$) and PP1 ($R = 0{\cdot}7$ or 2) dgp's. ($N = 200, s = 1000$)

| Dgp | | Test $a_1 = 0$ | | | Test $\beta_0 = 0$ | | | Test $\beta_1 = 0$ | | | Selection | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $R$ | $\beta_1$ | OP | H | S | OP | H | S | OP | H | S | LR | BIC |
| 1·0 | 0·0 | 0·205 | 0·227 | 0·248 | 0·164 | 0·192 | 0·226 | 0·043 | 0·047 | 0·044 | 0·047 | 0·025 |
| 0·7 | 0·0 | 0·986 | 0·985 | 0·984 | 0·977 | 0·992 | 0·991 | 0·039 | 0·040 | 0·043 | 0·969 | 0·940 |
| 2·0 | 0·0 | 0·711 | 0·711 | 0·947 | 0·708 | 0·736 | 0·733 | 0·121 | 0·080 | 0·065 | 0·711 | 0·711 |
| 1·0 | 0·5 | 0·212 | 0·227 | 0·247 | 0·161 | 0·187 | 0·224 | 0·647 | 0·650 | 0·642 | 0·054 | 0·018 |
| 0·7 | 0·5 | 0·992 | 0·994 | 0·992 | 0·980 | 0·989 | 0·994 | 0·992 | 0·994 | 0·992 | 0·959 | 0·980 |
| 2·0 | 0·5 | 0·930 | 0·930 | 1·000 | 0·999 | 0·999 | 0·999 | 0·930 | 0·946 | 0·936 | 0·930 | 0·930 |

overfitted PP1 model, but there is much less precision in estimating $\beta_0$ when in addition one attempts to estimate $a_1$.

In Table III the precision of estimation of parameter estimation and the difference between various estimates of precision are investigated. This table presents the proportion of times that the null hypothesis of a zero coefficient is rejected when performing a Wald test at 5%. Three different estimates of the variance matrix presented in Section 2.4 are used, namely outer-product (OP), Hessian (H) and sandwich (S). Comparing the columns OP, H and S in Table III there is relatively little difference in rejection rates using these different measures. From the first three rows of Table III where the dgp sets $\beta_1 = 0$ the size of tests of $\beta_1 = 0$ is quite good. The actual size is with one exception, the OP version in the overdispersed case, in the range 0·039 to 0·080 compared to a nominal size of 0·050. From rows one and four, however, where the dgp sets $a_1 = 0$ there is clearly a problem in the size of tests of $a_1 = 0$ with the nominal size ranging between 0·205 and 0·248. Similar problems arise in this case for the Wald test of $\beta_0 = 0$. This problem is closely related to problems discussed in Section 2.4 of performing an LM test of $a_1 = 0$. Under the null hypothesis two of the first-order conditions are identical. The problem may potentially be solved by an LR test as it does not attempt to separate the role of $\beta_0$ and $a_1$.

The tenth column (LR) in Table III gives the proportion of times the PP1 model is rejected on the basis of a likelihood ratio test at the 5% significance level. From rows one and four this has very good size properties with actual size of 0·047 and 0·054 very close to nominal size. The eleventh column of the table (BIC) gives the proportion of times the PP1 model is rejected on the basis of BIC presented in Section 2.4. We expect rejection of the Poisson model less often using BIC than LR, since BIC rejects Poisson if the difference in $-2 \ln \mathcal{L}$ exceeds $\ln(200) = 5{\cdot}30$, whereas the critical value for LR is $\chi^2_{0{\cdot}05}(1) = 3{\cdot}84$. This is the case for all but the overdispersed model results in rows three and six where the rejection rates are the same. Finally the fitted distribution is much closer to the actual distribution for PP1 and Poisson where $a_1 \neq 0$, and also closer when $a_1 = 0$, in which case the dgp is the Poisson. This closeness is measured by computing the average over the simulations of the absolute differences in predicted probabilities as discussed in Section 2.4. For space reasons this statistic is not included in the table.

In summary there appears to be a gain to fitting a PP1 model over fitting a Poisson when the dgp is PP1 or Poisson. Care needs to be taken in estimating $a_1$ and $\beta_0$ and their standard errors, and it is best to use the LR test rather than LM or Wald test in performing a formal test of Poisson against PP1. No real problems arise in estimating the slope coefficient and its standard error.

 

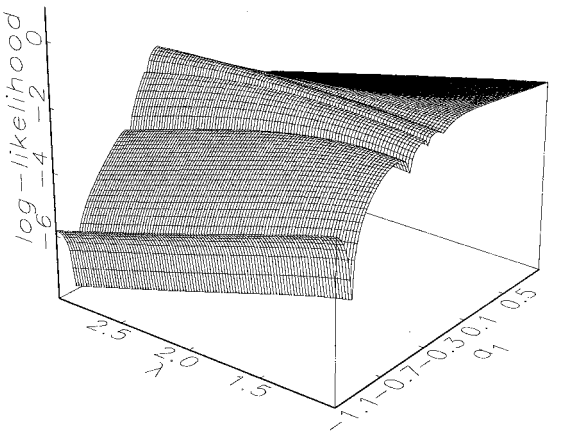


Figure 5. Log-likelihood function of $\lambda$ and $a_1$ for the sample from a PP1 model with $\lambda = 1{\cdot}9$ and $a_1 = -0{\cdot}2999$

We additionally investigated the advantage of using fast simulated annealing (FSA). Figure 5 presents the log-likelihood function for different values of $a_1$ and $\lambda = \exp(\beta_0)$ for one randomly drawn sample from the PP1 model with $(\lambda, a_1) = (1{\cdot}9, -0{\cdot}299)$.[3] Clearly there are several local optima. We then performed a simulation analysis where 1000 samples were drawn from this PP1 dgp. The PP1 model for these samples was estimated using as the starting value $(\lambda, a_1) = (\bar{y}, -0{\cdot}7)$, where the starting value for $\lambda$ corresponds to using the Poisson MLE. The hybrid FSA/gradient method detailed in Appendix B was used and compared to standard Newton–Raphson (NR). In 992 cases out of 1000 the FSA method gives higher ln $\mathcal{L}$ and in the other eight the same ln $\mathcal{L}$ as NR. Clearly there is a problem with NR and the FSA method is better. The implementation of the FSA method we use (see Appendix B) does have the advantage of using several different starting values. When we instead use just the one starting value the FSA gives higher ln $\mathcal{L}$ than NR in 949 out of 1000 cases, and the same ln $\mathcal{L}$ in the other 51 cases.

This simulation indicates that there are considerable computational advantages to using FSA over NR. But because we use a hybrid FSA/gradient method to reduce computational time, rather than pure FSA, the method is not guaranteed to converge to the global optimum and it is advisable to use several different starting values. Also it should be clear from Appendix B that there are quite a number of decisions to be made in implementing the hybrid FSA/gradient method and it is possible that one could make sufficiently poor choices that FSA might be no better than NR.

## 4. APPLICATIONS

### 4.1. Takeover Bids

Jaggia and Thosar (1993) model the number of bids received by 126 US firms that were targets of tender offers during the period 1978–85, and were actually taken over within 52 weeks of the

[3] The log-likelihood was calculated as $\ln \mathcal{L}(\lambda, a_1) = \Sigma_{i=1}^{N} \ln g_1(y_i \mid \lambda, a_1)$ where $g_1(y \mid \lambda, a_1)$ denotes the PP1 density and $y_i, i = 1, \ldots, N$, is a random sample of size $N$ from $g_1(y \mid 1{\cdot}9, -0{\cdot}299)$.

 

initial offer. The dependent count variable is the number of bids after the initial bid (*NUMBIDS*) received by the target firm. The fitted model is Poisson with regressors that measure:

(1) Defensive actions taken by management of the target firm — indicator variables for legal defence by lawsuit (*LEGLREST*), proposed changes in asset structure (*REALREST*), proposed change in ownership structure (*FINREST*) and management invitation for friendly third-party bid (*WHITEKNT*). These are expected to decrease the number of bids, aside from *WHITEKNT*, which may increase bids as it is itself a bid.
(2) Firm-specific characteristics — bid price divided by price 14 working days before bid (*BIDPREM*), percentage of stock held by institutions (*INSTHOLD*), total book value of assets in billions of dollars (*SIZE*) and book value squared (*SIZESQ*). A high value of *BIDPREM* indicates a bid so attractive that additional bids are unlikely. The greater the institutional holdings, the more likely outside offers are to be favourably received, which will encourage more bids. As the size of the firm increases there are expected to be more bids, up to a point where the firm gets so large that few others are capable of making a credible bid.
(3) Intervention by federal regulators — an indicator variable for Department of Justice intervention (*REGULATN*). Regulator intervention is likely to discourage bids.

The data have two interesting features — underdispersion and relatively few zeroes. The amount of underdispersion is relatively modest. The sample mean of $y$ is 1·738 and sample variance is 2·050. This is only a small amount of overdispersion ($2{\cdot}050/1{\cdot}738 = 1{\cdot}18 \simeq 1$), which can be expected to disappear as regressors are added. In fact attempts at ML estimation of the negative binomial (Negbin 2) model yield dispersion parameter equal to its boundary value of zero. A regression-based test of under- or overdispersion yields a coefficient of $-0{\cdot}0683$ with a $t$-statistic of 1·18. This indicates some underdispersion, though not enough to reject the null hypothesis of equidispersion. Jaggia and Thosar accordingly used only the Poisson model.

The relatively few zeroes are quite striking. The frequencies for 0, 1, 2, ..., 10 bids are, respectively, 9, 63, 31, 12, 6, 1, 2, 1, 0, 0, 1. After inclusion of regressors in the Poisson model, the average predicted frequencies for 0, 1, 2, ..., 5 bids are, upon rounding and using results presented below, respectively, 25, 38, 30, 19, 8, 3, so that the Poisson model greatly overpredicts the probability of 0 counts and underpredicts the probability of 1 count. The problem is that while the sample average is only 1·7 bids received (after the first), virtually all target firms do receive at least one bid.

Using the LR test the Poisson model is rejected at 5% when testing against the PP1 model, while the PP1 model is not rejected when tested against a PP2 model. The PP1 model is therefore preferred. Coefficient estimates, $t$-statistics and mean marginal effects for the PP1 model are given in Table IV, along with those for the Poisson model. The mean marginal effect of a one unit change in each of the regressors is computed as $(1/N)\Sigma_i \partial E[y_i \mid X_i]/\partial X_{ij}$ for all regressors, including indicator variables.[4] We discuss the PP1 estimates. The only defensive action variable statistically significant at 5% is *LEGLREST*, which has a surprising positive sign. From the last column of Table IV legal defence by lawsuit leads to, on average, an increase of one-half bid. The effect of *REALREST* is quite large and negative, but is statistically insignificant. The coefficient

[4] Strictly speaking for indicator variables one should instead evaluate the conditional mean of $y$ at indicator variable values of zero and one, and subtract. But taking the derivative is a reasonable approximation and is just as informative if making comparisons across models.

 

Table IV. Takeover bids: parameter estimates, *t*-statistics, and mean marginal effects for Poisson and PP1 models

| Variable | Estimate | | *t*-statistic | | Mean deriv. | |
|---|---|---|---|---|---|---|
| | Poisson | PP1 | Poisson | PP1 | Poisson | PP1 |
| *ONE* | 0·986 | 0·210 | 2·39 | 0·28 | | |
| *LEGLREST* | 0·260 | 0·522 | 2·09 | 2·38 | 0·452 | 0·466 |
| *REALREST* | −0·196 | −0·372 | −1·08 | −1·39 | −0·341 | −0·332 |
| *FINREST* | 0·074 | 0·138 | 0·28 | 0·47 | 0·129 | 0·124 |
| *WHITEKNT* | 0·481 | 1·013 | 4·54 | 3·88 | 0·837 | 0·906 |
| *BIDPREM* | −0·678 | −1·334 | −2·29 | −2·46 | −1·178 | −1·192 |
| *INSTHOLD* | −0·362 | −0·757 | −1·13 | −1·23 | −0·629 | −0·677 |
| *SIZE* | 0·179 | 0·329 | 2·87 | 3·99 | 0·310 | 0·294 |
| *SIZESQ* | −0·008 | −0·014 | −2·74 | −3·30 | −0·013 | −0·013 |
| *REGULATN* | −0·029 | −0·081 | −0·21 | −0·36 | −0·051 | −0·073 |
| $a_1$ | | 3·382 | | 3·02 | | |

of *WHITEKNT* indicates that a management invitation for a friendly third-party bid merely adds another bidder, as it increases the number of bids by, on average, 0·9 of a bid. *BIDPREM* is statistically significant though has a relatively modest effect with an increase in the bid premium of 0·2, which is approximately one standard deviation of *BIDPREM*, leading to an increase on average of 0·24 in the number of bids. Larger institutional holdings, if anything, are associated with a decrease in the number of bids, though this effect is statistically insignificant. The size of the firm matters, with bids first increasing and then decreasing as size increases. The effect of intervention by government regulators is very small in magnitude and statistical significance. The model results are generally in accord with *a priori* beliefs, except that they provide no support for the view that defensive measures by management are associated with a decrease in the number of bids.

Comparing the PP1 and Poisson model estimates, the PP1 estimates are more precise, with *t*-statistics around 10% higher on average than Poisson *t*-statistics, where the Poisson *t*-statistics are based on the sandwich (or Eicker–White) estimate of the variance matrix to control for the underdispersion. The coefficients of the Poisson and PP1 model are scaled differently and not directly comparable. To check the reasonableness of the PP1 parameter estimates we compare the mean marginal effects of PP1 with those from Poisson. There is relatively little difference, with the mean effects for all variables being within 10% of each other. The model differences lie in predicting probabilities and other aspects of the distribution aside from the mean.

Predicted probabilities of the Poisson and PP1 models are compared in Table V. The second column gives the fraction of the sample taking the particular value given in the first column. Columns three and four give the average predicted probabilities from estimated Poisson and PP1 models for that count. The first of these columns displays the already discussed inability of the Poisson model to fit the empirical distribution mentioned earlier. The PP1 model does quite well in fitting the empirical distribution, with the predicted probability of 0 bids of 0·0794 being close to the actual frequency of 0·0714, whereas the Poisson model predicts 0 bids with probability 0·2132. For the PP1 model there is still some underprediction of counts of 1 and overprediction of counts of 2 and 3.

The commonly used negative binomial model for counts cannot be applied to underdispersed data. The simplest to implement and most commonly used model in this case is the Poisson

 

Table V. Takeover bids: predicted probabilities from Poisson, PP1, Hurdle, double Poisson, and GECk models

| Counts | Actual | Poisson | PP1 | Hurdle | DP | GECk |
|---|---|---|---|---|---|---|
| 0 | 0·0714 | 0·2132 | 0·0794 | 0·0718 | 0·1437 | 0·1793 |
| 1 | 0·5000 | 0·2977 | 0·4313 | 0·4916 | 0·3616 | 0·3151 |
| 2 | 0·2460 | 0·2327 | 0·2864 | 0·2382 | 0·2760 | 0·2598 |
| 3 | 0·0952 | 0·1367 | 0·1252 | 0·1079 | 0·1395 | 0·1429 |
| 4 | 0·0476 | 0·0680 | 0·0482 | 0·0486 | 0·0575 | 0·0630 |
| 5 | 0·0079 | 0·0305 | 0·0182 | 0·0221 | 0·0214 | 0·0249 |
| 6 | 0·0159 | 0·0128 | 0·0070 | 0·0103 | 0·0075 | 0·0094 |
| 7 | 0·0079 | 0·0052 | 0·0027 | 0·0049 | 0·0025 | 0·0036 |
| 8 | 0·0000 | 0·0020 | 0·0010 | 0·0024 | 0·0008 | 0·0013 |
| 9 | 0·0000 | 0·0007 | 0·0004 | 0·0011 | 0·0002 | 0·0004 |
| 10 | 0·0079 | 0·0003 | 0·0002 | 0·0005 | 0·0001 | 0·0001 |
| $-\ln L$ | | 185·0 | 172·4 | 160·0 | 177·9 | 181·5 |
| BIC | | 418·3 | 398·1 | 416·7 | 409·0 | 416·2 |

hurdle model presented in Mullahy (1986). This introduces a regression model for zero counts in addition to and different from the Poisson regression for positive counts. It should do very well for the takeover bids data, as it is particularly the zeros that are poorly predicted by the Poisson model. At the same time, one should only use a hurdle model if indeed there is strong theoretical reason for treating zero counts differently from positive counts. Otherwise its use is a data-mining exercise similar to putting an observation-specific dummy variable in a least squares regression whenever an observation is poorly predicted. Predicted probabilities from an estimated Poisson hurdle are presented in Table V (for space reasons the parameter estimates are not given). The Poisson hurdle model does quite well in predicting probabilities and has lower $\ln \mathcal{L}$ than PP1. But it has almost twice as many parameters (20 versus 11) as the PP1 model. Allowing for this using BIC leads to preference for the PP1 model.

An alternative model for underdispersed data is the double Poisson model proposed by Efron (1986). This introduces one additional parameter, so is quite parsimonious compared to the hurdle model. The regression parameter estimates (not reported) were all within 5% of the Poisson estimates, and the additional dispersion parameter was highly significant with a $t$-statistic of 3·92. The predicted probabilities from this model are presented in Table V, in the column labelled DP. These probabilities sum to 1·0108, illustrating the theoretical weakness that probabilities for the double Poisson model do not sum to one, though the difference here is not great. For the most problematic 0 and 1 counts the predicted probabilities for the double Poisson model are roughly half-way between those for Poisson and the actual probabilities. The PP1 model is clearly preferred with predicted probabilities closer to actual, considerably higher $\ln \mathcal{L}$ and lower BIC.

Yet another model that can be applied to underdispersed data is the Katz system or GECk model proposed by King (1989) and Winkelmann and Zimmermann (1991). This model has the attraction of nesting the Poisson and, for overdispersed data, the negative binomial model. In principle, this model can accommodate a variance of the form $\lambda_i + \alpha\lambda_i^k$, where $\alpha$ and $k$ are dispersion and non-linearity parameters to be estimated. Underdispersion can arise when $-1 < \alpha < 0$ and $k \leqslant 1$. In practice, however, computational problems arise if at any stage of estimation and for any observation $\lambda_i + \alpha\lambda_i^k < 0$. Such problems were experienced here, and we set $k = 1$. To avoid convergence problems the regressors were rescaled so that Poisson regression

 

coefficients, used as starting values, differed in order of magnitude by no more than 10. The GECk regression parameter estimates (not reported) differed from Poisson estimates by, on average, 10%, and the dispersion parameter $\hat{\alpha} = -0{\cdot}258$ was highly significant with a $t$-statistic of 3·74. The predicted probabilities from this model, presented in the column labelled GECk in Table V, are roughly as close to the actual probabilities as those from double Poisson. Again the PP1 model is clearly preferred with predicted probabilities closer to the actual, considerably higher ln $\mathcal{L}$ and lower BIC.

### 4.2. Health Service Utilization

Several health service utilization measures are analysed by Cameron *et al.* (1988) using data from the 1977–8 Australian Health Survey. Here we model the number of health professional visits (*HPVISITS*), defined as the number of consultations in the past four weeks with non-doctor health professionals (chemist, optician, physiotherapist, social worker, district community nurse, chiropodist, or chiropractor). The regressors are:

(1) Socio-economic variables — an indicator variable for whether female (*SEX*), age in years (*AGE*), age-squared (*AGESQ*), annual income in hundreds of dollars (*INCOME*).
(2) Health insurance status indicator variables — private insurance cover (*LEVYPLUS*), free government insurance cover due to low income (*FREEPOOR*) and free government cover due to old age, disability or veteran status (*FREEREPA*). The omitted category is the default government Medibank insurance cover paid for by an income levy (*LEVY*).
(3) Recent health-status measures — number of illnesses in past two weeks (*ILLNESS*) and number of days of reduced activity in past two weeks due to illness or injury (*ACTDAYS*).
(4) Long-term health status measures — general health questionnaire score using Goldberg's method with high score indicating bad health (*HSCORE*), indicator variable for chronic condition not limiting activity (*CHCOND1*), and indicator variable for chronic condition limiting activity (*CHCOND2*).

The most notable feature of the data is overdispersion. The sample mean of $y$ is 0·215 and sample variance is 0·932. This is a considerable amount of overdispersion ($0{\cdot}932/0{\cdot}215 = 4{\cdot}335$), which only partially disappears as regressors are added. Statistical tests strongly reject the null hypothesis of no overdispersion. The frequencies for $0, 1, 2, \ldots, 11$ visits in the sample of size 5190 are, respectively, 0·909, 0·054, 0·016, 0·003, 0·005, 0·001, 0·002, 0·007, 0·001, 0·002, 0·000, 0·001.

The PP5 model is preferred to PP1–PP4 and PP6. Table VI presents parameter estimates, $t$-statistics, and mean marginal effects for the PP5 model, along with those for the Poisson and negative binomial (NB). For NB we use the Negbin 2 version with variance equal to $\lambda_i + \delta\lambda_i^2$ where the dispersion parameter $\delta$ is reported in the row labelled $\delta$ and $a_1$. The PP5 estimates reveal that the most important determinant of health professional visits is health status, with all but *ILLNESS* statistically significant at 5%. The coefficient of the insurance indicator variable *LEVYPLUS* is consistent with the view that more generous health insurance is associated with greater use of health services, while the positive coefficient of *FREEREPA* most likely reflects health problems by people in this group not fully picked up by health-status regressors. The coefficient of *FREEPOOR* is essentially zero, indicating that those who receive free government insurance cover due to low income use the same amount of services as those who receive the same government health insurance by paying the Medibank levy. The socio-economic effects are of the expected signs though generally statistically insignificant. Particularly striking is the small coefficient and strong statistical insignificance of *INCOME*.

 

Table VI. Health professional visits: parameter estimates, *t*-statistics, and mean marginal effects for Poisson, NB, and PP5 models

| Variable | Estimate | | | *t*-statistic | | | Mean effect | | |
|---|---|---|---|---|---|---|---|---|---|
| | Poisson | NB | PP5 | Poisson | NB | PP5 | Poisson | NB | PP5 |
| *ONE* | −2·444 | −2·784 | −1·199 | −5·32 | −6·40 | −7·81 | | | |
| *SEX* | 0·332 | 0·231 | 0·093 | 2·20 | 1·85 | 2·47 | 0·071 | 0·063 | 0·076 |
| *AGE* | −3·308 | −2·676 | −0·693 | −1·46 | −1·10 | −1·08 | −0·710 | −0·725 | −0·571 |
| *AGESQ* | 4·390 | 3·854 | 0·949 | 1·75 | 1·47 | 1·43 | 0·942 | 1·044 | 0·783 |
| *INCOME* | −0·035 | −0·062 | −0·015 | −0·16 | −0·33 | −0·24 | −0·008 | −0·017 | −0·012 |
| *LEVYPLUS* | 0·328 | 0·299 | 0·111 | 1·84 | 1·89 | 1·90 | 0·070 | 0·081 | 0·092 |
| *FREEPOOR* | 0·015 | −0·197 | 0·030 | 0·04 | −0·56 | 0·18 | 0·003 | −0·053 | 0·019 |
| *FREEREPA* | 0·482 | 0·588 | 0·144 | 2·41 | 2·69 | 2·20 | 0·104 | 0·159 | 0·119 |
| *ILLNESS* | 0·055 | 0·144 | 0·014 | 1·31 | 3·09 | 1·35 | 0·012 | 0·039 | 0·012 |
| *ACTDAYS* | 0·098 | 0·137 | 0·020 | 6·16 | 8·03 | 6·92 | 0·021 | 0·037 | 0·017 |
| *HSCORE* | 0·045 | 0·074 | 0·010 | 1·80 | 2·65 | 1·73 | 0·010 | 0·020 | 0·008 |
| *CHCOND1* | 0·519 | 0·412 | 0·199 | 3·28 | 2·88 | 3·64 | 0·111 | 0·111 | 0·164 |
| *CHCOND2* | 1·079 | 1·124 | 0·322 | 5·12 | 6·14 | 5·45 | 0·232 | 0·111 | 0·266 |
| $\delta$ and $a_1$ | | 8·909 | 2·740 | | 13·19 | 4·32 | | | |
| $a_2$ | | | −6·351 | | | −5·35 | | | |
| $a_3$ | | | 3·869 | | | 5·26 | | | |
| $a_4$ | | | −0·939 | | | −5·08 | | | |
| $a_5$ | | | 0·078 | | | 4·91 | | | |

Comparing statistical significance across models, the PP5 model has *t*-statistics that are quite similar to the sandwich *t*-statistics for Poisson. Comparing *t*-statistics instead to the NB model, the PP5 model has *t*-statistics that are marginally lower on average, aside from *ILLNESS*, which has much larger mean effect in the NB model compared to the PP5.

Table VI also compares the predicted effects of changes in regressors on the mean. Generally the marginal effects are greater in the NB model than in the Poisson model. For the PP5 and other PPp models there is no such general tendency with marginal effects closer to those of the Poisson. This is most likely a consequence of residuals summing to zero for the Poisson and PPp models but not for the NB model, so that for the NB model the average of the fitted means differs from $\bar{y}$.

Table VII reveals that the PPp models do very well in fitting the distribution, especially compared to the Poisson. The major gain in log-likelihood is going from Poisson to PP1, though the data support adding additional terms and the preferred model is a PP5 model. Both the PP4 and PP5 models have higher log-likelihood than the NB model, though are considerably less parsimonious. Using BIC to discriminate between non-nested models with a different number of parameters, both PP4 and PP5 are preferred to NB.

## 5. CONCLUSIONS

The simulations and applications demonstrate that the new class of models proposed can be estimated, even up to the sixth order in the health application. The distributional fit is quite good on average, and reasonable predictions of effects of regressors on the conditional mean are obtained.

Like other series expansion models, such as expansions around the normal for continuous data, the PPp model has log-likelihood which is not globally concave and hence multiple optima

 

Table VII. Health professional visits: predicted probabilities from Poisson, NB, and PPp models

| Counts | Empirical | Poisson | NB | PP1 | PP2 | PP3 | PP4 | PP5 |
|---|---|---|---|---|---|---|---|---|
| 0 | 0·9087 | 0·8311 | 0·9088 | 0·8955 | 0·8898 | 0·9067 | 0·9093 | 0·9087 |
| 1 | 0·0536 | 0·1377 | 0·0521 | 0·0573 | 0·0759 | 0·0558 | 0·0543 | 0·0536 |
| 2 | 0·0162 | 0·0222 | 0·0167 | 0·0091 | 0·0058 | 0·0142 | 0·0152 | 0·0162 |
| 3 | 0·0027 | 0·0056 | 0·0075 | 0·0211 | 0·0070 | 0·0022 | 0·0024 | 0·0026 |
| 4 | 0·0050 | 0·0020 | 0·0040 | 0·0114 | 0·0093 | 0·0017 | 0·0041 | 0·0055 |
| 5 | 0·0012 | 0·0008 | 0·0024 | 0·0040 | 0·0069 | 0·0072 | 0·0024 | 0·0008 |
| 6 | 0·0019 | 0·0003 | 0·0016 | 0·0012 | 0·0034 | 0·0068 | 0·0027 | 0·0018 |
| 7 | 0·0071 | 0·0001 | 0·0011 | 0·0003 | 0·0013 | 0·0036 | 0·0034 | 0·0051 |
| 8 | 0·0012 | 0·0000 | 0·0008 | 0·0000 | 0·0004 | 0·0013 | 0·0029 | 0·0037 |
| 9 | 0·0015 | 0·0000 | 0·0006 | 0·0000 | 0·0001 | 0·0004 | 0·0018 | 0·0015 |
| 10 | 0·0004 | 0·0000 | 0·0005 | 0·0000 | 0·0000 | 0·0000 | 0·0009 | 0·0004 |
| 11 | 0·0006 | 0·0000 | 0·0004 | 0·0000 | 0·0000 | 0·0000 | 0·0003 | 0·0001 |
| $-\ln L$ | | 3109·4 | 2160·5 | 2425·3 | 2297·6 | 2192·6 | 2142·6 | 2136·4 |
| BIC | | 6329·9 | 4440·8 | 4970·4 | 4723·5 | 4522·0 | 4430·5 | 4426·9 |

can arise. In some simulations there was advantage to using a variant of fast simulated annealing, but in the applications such problems did not arise. At the least, however, it is advisable to try a range of starting values.

The PPp model was particularly useful in application to underdispersed takeover bids data, where it clearly outperformed the double Poisson and GECk models and was more parsimonious than the hurdle model. For application to overdispersed health-utilization data, fitted very well by the negative binomial model, the PPp model was able to outperform the negative binomial model but was not as parsimonious. In this latter example the PPp model had estimates of the average marginal effect of regressors on the conditional mean somewhat different from negative binomial and similar to Poisson, most likely a consequence of the PPp model property that, like Poisson regression residuals, sum to zero. On the basis of these applications the PPp model is as useful as standard parametric models for counts. It is not necessarily as parsimonious so that, for example, if the negative binomial fits an overdispersed data model very well there is unlikely to be great advantage in moving to the PPp model.

The particular model considered here was chosen in part for its relative simplicity of use and ease of exposition. There are clearly many possible variants.

First, the weights $a_j$ in the polynomial function $h(\cdot)$ can be permitted to be a function of regressors, in which case $a_j$ is replaced by $a_{ji} = \mathbf{X}_i'\boldsymbol{\alpha}_j$. This is done in applications to continuous time-series data, where the regressors are lags of the dependent variable being modelled. In the cross-section case considered here such an extension runs the risk of introducing many more parameters, unless attention is restricted to the one or two regressors thought to be most important.

Second, baseline densities other than the Poisson might be chosen. In particular the negative binomial is an obvious choice. This offers the prospect of more parsimonious models for overdispersion, while still being applicable to underdispersed data as it nests the PPp model as a special case (the Poisson is a special case of the negative binomial).

Third, the polynomial could be a function of a transformation of $y$ rather than $y$ itself, i.e. the polynomial function $h(y \mid \mathbf{a})$ is instead $h(t(y) \mid \mathbf{a})$ for specified function $t(y)$. One possibility is to centre around the mean using $t(y) = (y - \lambda)$. A second possibility is to standardize to

 

approximately constant variance using $t(y) = (y - \lambda)/\sqrt{\lambda}$. This complicates the analysis considerably, and like Hall (1990) we do not take this approach here. Note also that such standardizations will not produce a variable with mean zero and variance one since, for example, $E[y] \neq \lambda$ from Section 2.

Extension of the single-equation cross-section model in this paper to more complicated types of count data is of particular interest. One example is multivariate data, such as bivariate counts for use of two different but related types of health service. Existing models generally place restrictions on the correlation coefficient, including restricting it to be non-negative.

## APPENDIX A: DERIVATION OF RESULTS

### Derivation of Equations (3) and (4)

While proof is for the discrete case, the same result holds for the continuous case. We first derive equation (4).

$$\begin{aligned}
E[y^r] &= \sum_y y^r g_p(y \mid \boldsymbol{\lambda}, \mathbf{a}) \\
&= \sum_y y^r f(y \mid \boldsymbol{\lambda}) h_p^2(y \mid \mathbf{a}) / \eta_p(\boldsymbol{\lambda}, \mathbf{a}) \\
&= \sum_y \sum_{k=0}^{p} \sum_{l=0}^{p} y^r f(y \mid \boldsymbol{\lambda}) a_k a_l y^k y^l / \eta_p(\boldsymbol{\lambda}, \mathbf{a}) \\
&= \sum_{k=0}^{p} \sum_{l=0}^{p} a_k a_l \left\{ \sum_y y^{k+l+r} f(y \mid \boldsymbol{\lambda}) \right\} / \eta_p(\boldsymbol{\lambda}, \mathbf{a}) \\
&= \sum_{k=0}^{p} \sum_{l=0}^{p} a_k a_l m_{k+l+r} / \eta_p(\boldsymbol{\lambda}, \mathbf{a})
\end{aligned}$$

Derivation of equation (3) is obtained by setting $r = 0$ in equation (4) and using $E[y^0] = 1$.

### Derivation of Equation (13)

For the Poisson density,

$$\frac{\partial \log f(y \mid \lambda)}{\partial \lambda} = \frac{y - \lambda}{\lambda}$$

and

$$\begin{aligned}
\frac{\partial m_r(\lambda)}{\partial \lambda} &= \sum_y y^r \frac{\partial \exp(-\lambda + y \log \lambda - \log y!)}{\partial \lambda} \\
&= \sum_y y^r \left( -1 + \frac{y}{\lambda} \right) \exp(-\lambda + y \log \lambda - \log y!) \\
&= \frac{m_{r+1} - \lambda m_r}{\lambda}
\end{aligned}$$

 

For the mean function (12)

$$\frac{\partial \lambda(\mathbf{X}'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \lambda \mathbf{X}$$

Substituting these results into the first equation of (11):

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \left\{ (y_i - \lambda_i) - \frac{\sum_{k=0}^{p}\sum_{l=0}^{p} a_k a_l \{m_{k+l+1,i} - \lambda_i m_{k+l,i}\}}{\sum_{k=0}^{p}\sum_{l=0}^{p} a_k a_l m_{k+l,i}} \right\} \mathbf{X}_i$$

$$= \sum_{i=1}^{N} \left\{ y_i - \frac{\sum_{k=0}^{p}\sum_{l=0}^{p} a_k a_l m_{k+l+1,i}}{\sum_{k=0}^{p}\sum_{l=0}^{p} a_k a_l m_{k+l,i}} \right\} \mathbf{X}_i$$

## APPENDIX B: COMPUTATIONAL METHODS

The method to compute parameter estimates is a hybrid of fast simulated annealing (FSA) and the standard gradient methods, based on Horowitz (1992).

Simulated annealing differs from gradient methods in permitting at times movements that decrease rather than increase the objective function (which here we seek to maximize), so that one is not locked into moving steadily towards a local optimum. Key parameters of the algorithm are the step length $V$, and the temperature $T$. For a description of simulated annealing and explanation of the term temperature, see Goffe, Ferrier, and Rogers (1994, pp. 68–70). Fortran code for the program used by Goffe *et al.* (1994) is available at http://netec.mcc.ac.uk/~adnetec/CodEc/Fortran/SimAnnealing.si. A Gauss program by E. G. Tsionas for 'Global optimization of statistical functions with simulated annealing' is available at the Gauss archive at http://netec.mcc.ac.uk/~adnetec/CodEc/GaussAtAmericanU/index.html.

FSA, proposed by Szu and Hartley (1987), is a faster method that replaces the uniform $(-1, 1)$ random number $r$ in equation (1) of Goffe *et al.* (1994) by a Cauchy random variable $r_i$ scaled by the temperature. It also permits a fixed step length $V$ and a simpler adjustment of the temperature with equation (3) of Goffe *et al.* (1994) replaced by $T' = T/N_s$, where $N_s$ is the number of FSA iterations.

In principle, the models in this paper can be estimated using only FSA. This method can be shown to converge to the global maximum, but is computationally very expensive. Instead we reduce computation time by following Horowitz (1992) and using FSA to obtain starting values for gradient methods. The algorithm is:

(1) Search for an optimum using FSA. Stop when there is relatively little absolute difference between the average function values over the last ten iterations and the optimal function value to date.
(2) Check that, for the parameter values obtained from step 1, the Hessian is negative definite.
   (a) If this test is failed, then return to step 1 using these parameter values as starting values, and significantly decrease the temperature.

 

(b) If this test is passed, then begin BFGS iterations (cf. Fletcher, 1981) moving to Newton–Raphson iteration using these parameter values as starting values. The estimates obtained are the optimum estimates and estimation stops.

(3) If optimum estimates are not obtained after 250 FSA simulations, or if step 2(a) is encountered six times, then use these current parameter estimates as starting values for BFGS iterations.

This procedure was performed from ten different starting values. The starting values were random draws from the normal distribution, added to the Poisson parameter estimates in the case of the regression parameters $\boldsymbol{\beta}$.

For PP1 to PP5 analytical first and second derivatives were used in the Newton–Raphson algorithm. Results were checked against numerical derivatives. For PP6 only analytical first derivatives were used. The first derivatives are very easily programmed for the Poisson with exponential mean — see equation (13) and the second equation in (11). The second derivatives are not given in the paper but the computer code is available. The main reason for using analytical derivatives is for quicker computation, especially when many simulations are being performed.

There are several choice variables in implementing the algorithm. The above describes the method used in applications. For simulations we used three different starting values and in step 3 we used up to 50 FSA simulations or three encounters of step 2(a). Other choice variables are the temperature and step length. The Gauss code used is available at the journal web-site.

Once PPp parameter estimates were obtained their standard errors were computed using the Hessian, except in the Table III simulations where other estimates were additionally used. In the applications Tables IV and VI the Poisson standard errors and $t$-statistics used the sandwich form which is robust to departures from variance–mean equality, while other models, including PPp, used the Hessian.

ACKNOWLEDGEMENTS

The authors thank the Department of Statistics at The Australian National University for their hospitality. They also thank Kurt Brännäs, Adrian Pagan, George Tauchen, Pravin Trivedi, and two anonymous referees for helpful suggestions, Joel Horowitz for access to his program for maximum score estimation, Klaus Zimmermann for access to Gauss code for the GECk model, and Sanjiv Jaggia for providing the data on takeover bids. Financial support from the Swedish Research Council for the Humanities and Social Sciences is gratefully acknowledged. This work has benefited from seminar presentations at Monash and Umea Universities, the European University Institute, and the Universities of Sydney, New South Wales, Texas–Austin, and Munich.

REFERENCES

Cameron, A. C. and P. K. Trivedi (1986), 'Econometric models based on count data: comparisons and applications of some estimators and tests', *Journal of Applied Econometrics*, **1**, 29–54.

Cameron, A. C., P. K. Trivedi, F. Milne, and J. Piggott (1988), 'A microeconometric model of the demand for health insurance and health care in Australia', *Review of Economic Studies*, **55**, 85–106.

Consul, P. C. and F. Famoye (1992), 'Generalized Poisson regression model', *Communications in Statistics: Theory and Method*, **21**, 89–109.

Efron, B. (1986), 'Double exponential families and their use in generalized linear regression', *Journal of the American Statistical Association*, **81**, 709–21.

 

Fletcher, R. (1981), *Practical Methods of Optimization, Vol. 1: Unconstrained Optimization*, Wiley, Chichester.
Gabler, S., F. Laisney, and M. Lechner (1993), 'Semiparametric estimation of binary choice models with an application to labor force participation', *Journal of Business and Economic Statistics*, **11**, 61–70.
Gallant, A. R. and D. W. Nychka (1987), 'Seminonparametric maximum likelihood estimation', *Econometrica*, **55**, 363–90.
Gallant, A. R. and G. Tauchen (1989), 'Seminonparametric estimation of conditionally constrained heterogeneous processes: asset pricing applications', *Econometrica*, **57**, 1091–1120.
Goffe, W. L., Ferrier, G. D., and J. Rogers (1994), 'Global optimization of statistical functions with simulated annealing', *Journal of Econometrics*, **60**, 65–99.
Gourieroux, C., A. Montfort, and A. Trognon (1984), 'Pseudo maximum likelihood methods: applications to Poisson models', *Econometrica*, **52**, 681–700.
Gurmu, S., P. Rilstone, and S. Stern (1994), 'Semiparametric estimation of count regression models', Department of Economics, University of Virginia, Charlottesville.
Gurmu, S. and P. K. Trivedi (1994), 'Recent developments in models of event counts: a survey', Discussion Paper No. 261, Thomas Jefferson Center, University of Virginia, Charlottesville.
Hall, A. (1990), 'Lagrange multiplier tests for normality against seminonparametric alternatives', *Journal of Business and Economic Statistics*, **8**, 417–25.
Horowitz, J. (1992), 'A smoothed maximum score estimator for the binary response model', *Econometrica*, **60**, 505–31.
Jaggia, S. and S. Thosar (1993), 'Multiple bids as a consequence of target management resistance: a count data approach', *Review of Quantitative Finance and Accounting*, December, 447–57.
King, G. (1989), 'Variance specification in event count models: from restrictive assumptions to a generalized estimator', *American Journal of Political Science*, **33**, 762–84.
Mullahy, J. (1986), 'Specification and testing of some modified count data models', *Journal of Econometrics*, **33**, 341–65.
Szu, H. and R. Hartley (1987), 'Fast simulated annealing', *Physics Letters A*, **122**, 157–62.
Winkelmann, R. (1994), *Count Data Models: Econometric Theory and an Application to Labor Mobility*, Springer-Verlag, Berlin.
Winkelmann, R. and K. F. Zimmermann, (1991), 'A new approach for modeling economic count data', *Economics Letters*, **37**, 139–43.

 