A new approach for modeling economic count data

Winkelmann, R; Zimmermann, K F


Postprint available at:
http://www.zora.unizh.ch

Posted at the Zurich Open Repository and Archive, University of Zurich.
http://www.zora.unizh.ch

Originally published at:
A new approach for modeling economic count data

Abstract

A new parametric model for the econometric analysis of non-negative integers is proposed. Its distinguishing feature is that it allows for more flexible variance-mean relationships than the models used hitherto. Estimation with maximum likelihood is illustrated using a dataset on ship damage incidents.
A New Approach for Modeling Economic Count Data

Forthcoming: Economics Letters

Rainer Winkelmann
University of Munich

Klaus F. Zimmermann
University of Munich and CEPR

July 1991

Abstract
A new parametric model for the econometric analysis of non-negative integers is proposed. Its distinguishing feature is that it allows for more flexible variance-mean relationships than the models used hitherto. Estimation with maximum likelihood is illustrated using a dataset on ship damage incidents.

*Financial support from the German Science Foundation (DFG) is gratefully acknowledged.

Correspondence to:
Prof. Dr. Klaus F. Zimmermann
SELAPO, University of Munich
Ludwigstr. 28 RG
D-8000 München 22
Germany
1 Introduction

In this paper we propose a new model for the analysis of non-negative integers, i.e. count data. Many economic issues give rise to count data and can therefore be usefully studied in this framework. Examples are the number of job changes of an individual, the number of patents obtained by a firm or the number of company takeovers in a specific country.

In Section 2 we describe the benchmark model for count data, the Poisson model. It has been recognized for long that this model often does not fit the data well, mainly due to the restrictive assumption that the conditional mean is equal to the conditional variance. We therefore discuss a more general canonical variance function and derive in Section 3 a generalized event count model based on this variance function. As an illustration, we apply the new model in Section 4 to a well studied data set on ship damages. Section 5 concludes.

2 Poisson model

As several researchers have pointed out, the Poisson regression model is a natural first choice for modeling count data. To assure a non-negative expectation, the explanatory variables are introduced like in a log-linear model:

\[ E(y_i | x_i) = \exp(x_i \beta) \quad i = 1, \ldots, n \]  
(1)

\( x_i \) is a (1xk) vector of covariates and \( \beta \) a (kx1) vector of coefficients. The observed values \( y_i \) are assumed to be drawings from a Poisson distribution with parameter \( \lambda_i = E(y_i | x_i) = \text{Var}(y_i | x_i) = \exp(x_i \beta) \), where the probability function is given by

\[ f(y_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \quad \lambda_i > 0, \ y_i = 0, 1, 2, \ldots \]  
(2)

Estimation with maximum likelihood is straightforward. Unlike in OLS, a closed form solution is not available, but since the log-likelihood function is globally con-
cave, standard numerical algorithms will converge rapidly to a unique maximum of the log-likelihood function.

Our focus is on the restrictive assumption of the equality of conditional mean and conditional variance. If the conditional mean is greater than the conditional variance, ‘extra Poisson variation’ is present, a situation often referred to as overdispersion. The opposite case is called underdispersion. There exists an extensive literature how to test for and how to model violations of equidispersion using parametric and semi-parametric alternatives. In both cases, the variance is modeled as a function of the mean and the assumptions about the specific form of the variance function (VF) are crucial.

We will introduce a canonical VF to discuss the existing proposals and to use it as a building block for our own model:

\[
\text{Var}(y_i|x_i) = (\sigma^2 - 1)[E(y_i|x_i)]^{k+1} + E(y_i|x_i). \tag{3}
\]

\(\sigma^2\) is called dispersion parameter and \(k\) non-linearity parameter. Both parameters are assumed to be scalars and especially not to depend on \(\beta\). This VF contains most of the cases previously studied in the literature on count data. \(\sigma^2 = 1\) gives the VF of the Poisson model. \(0 < \sigma^2 < 1\) and \([E(y_i|x_i)]^k \leq 1/(1 - \sigma^2)\) indicates underdispersion and corresponds to the VF of the continuous parameter binomial model\(^1\). \(\sigma^2 > 1\), a situation of overdispersion, corresponds to the VF of the negative binomial model\(^2\). In all applications of these models, a specific value of \(k\), either 0 or 1, has been assumed a priori. \(k = 1\) implies that the variance-mean ratio is a linear function of the mean, \(k = 0\) that it is a constant.

The value of \(k\) also plays a role in two semi-parametric results. If the true data generation process has a mean function given by (1) and a VF (3) with \(k = 1\) and \(\sigma^2 > 1\), Gourieroux, Monfort and Trognon (1984) show that consistent estimates of the parameters can be obtained with four linear exponential families

---

\(^1\)King (1989) defines this model for \(k = 0\).

\(^2\)See e.g. Lawless (1987) for a discussion of the negative binomial model with \(k=1\).
including the Poisson. For \( k = 0 \), the Poisson model will give consistent parameter estimates regardless of the value of \( \sigma^2 \) (see e.g. McCullagh and Nelder 1989). These results, however, heavily rely on the assumption of a specific \( k \). Also the standard errors of the estimates are generally biased.

We therefore advocate an alternative full parametric approach that gives efficient estimates and correct asymptotic standard errors for both the parameters of the VF(3), \( k \) and \( \sigma^2 \), and the parameters \( \beta \), using maximum likelihood estimation. The most important models used previously, the Poisson and the negative binomial models, as well as the continuous parameter binomial model, are nested within this more general model. Our model is based on a contribution by King (1989) who developed a generalized event count model (GEC) for \( k = 0 \). We generalize his model by deriving a generalized event count model with endogeneous and continuous \( k \) (GECk) according to the canonical VF(3).

3 GENERALIZED EVENT COUNT MODEL

The derivation along the lines of King (1989) is based on the properties of the Katz family of distributions, which is implicitly defined by a recursive formula for the probabilities \( f(y) \):

\[
\frac{f(y+1)}{f(y)} = \frac{\theta + \gamma y}{1 + y} \quad \text{for} \quad y = 0, 1, 2, \ldots \quad \text{and} \quad \theta + \gamma y \geq 0.
\] (4)

Using recursive substitution, (4) can be rewritten as

\[
f(y|\theta, \gamma) = f(0) \prod_{j=1}^{y_i} \left[ \frac{\theta + \gamma(j - 1)}{j} \right], \quad y_i = 1, 2, \ldots
\] (5)

where \( f(0) \) is determined by the fact that the probabilities have to sum up to one. Mean and variance are given by

\[
E(y) = \frac{\theta}{(1 - \gamma)}, \quad \text{Var}(y) = \frac{\theta}{(1 - \gamma)^2}.
\] (6)
It is easily seen that this family produces equidispersion for $\gamma = 0$, overdispersion for $0 < \gamma < 1$, and underdispersion for $\gamma < 0$. To obtain the canonical VF(3) we parameterize as follows:

$$
\gamma = \frac{(\sigma^2 - 1)\lambda_i^{k}}{(\sigma^2 - 1)\lambda_i^{k} + 1}, \quad \theta = \frac{\lambda_i}{(\sigma^2 - 1)\lambda_i^{k} + 1}, \quad \lambda_i = \exp(x_i\beta) \quad . \quad (7)
$$

The complete distribution is then given by:

$$
f_{\text{geck}}(y_i|\lambda_i, \sigma^2, k) = f(0|\lambda_i, \sigma^2, k) \times \begin{cases} 
\prod_{j=1}^{\nu_i} \left[ \frac{\lambda_i^{+ (\sigma^2 - 1)\lambda_i^{k}(j-1)}}{(\sigma^2 - 1)\lambda_i^{k} + 1} \right] & \text{for } y_i = 1, 2, \ldots \\
1 & \text{for } y_i = 0 
\end{cases} \quad (8)
$$

where

$$
f(0|\lambda_i, \sigma^2, k) = \begin{cases} 
(1 + (\sigma^2 - 1)\lambda_i^{k})^{\nu_i} & \text{for } \sigma^2 \geq 1 \\
(1 + (\sigma^2 - 1)\lambda_i^{k})^{\nu_i}D_i^{-1} & \text{for } 0 < \sigma^2 < 1, \lambda_i^k \leq 1/(1 - \sigma^2) \\
0 & \text{and } y_i \leq \text{int}^*(\nu_i) \\
\text{otherwise} 
\end{cases} 
$$

$$
\nu_i = \frac{\lambda_i^{1-k}}{(1 - \sigma^2)} 
$$

$$
D_i = \sum_{m=0}^{\text{int}^*(\nu_i)} f_{\text{binomial}}(m|\lambda_i, \sigma^2, k) 
$$

and $\text{int}^*(y) = \begin{cases} 
\text{int}(y)+1 & \text{for } \text{int}(y)<y \\
y & \text{for } \text{int}(y)=y 
\end{cases}$.

The limit of $f(0|\lambda_i, \sigma^2, k)$ for $\sigma^2 \to 1$ is $e^{-\lambda_i}$ and (8) converges to the Poisson model. The log-likelihood has the following form:

$$
\ln L(\beta, \sigma^2, k|y) = \sum_{i=1}^{n} \left\{ \ln f(0|\lambda_i, \sigma^2, k) + \sum_{j=1}^{y_i} \ln \left[ \frac{\lambda_i^{+ (\sigma^2 - 1)\lambda_i^{k}(j-1)}}{(\sigma^2 - 1)\lambda_i^{k} + 1} \right] \right\} \quad (9)
$$

for $f(0|\lambda_i, \sigma^2, k) \neq 0$. The maximizing values for $\beta$, $\sigma^2$ and $k$ can be found by using a numerical optimization algorithm. For our data we had no problems with convergence using own procedures written in GAUSS. The usual asymptotic
theory for maximum likelihood estimation holds. The hypothesis that the data are Poisson distributed can be tested with $H_0 : \sigma^2 = 1$. The hypothesis that the data follow the negative binomial model with quadratic VF as used by Lawless (1987) corresponds to the joint hypothesis $H_0 : \sigma^2 > 1, k = 1$. The hypothesis for the negative binomial model with linear VF is given by $H_0 : \sigma^2 > 1, k = 0$. One major gain of introducing the additional parameter $k$ is that it allows to discriminate between these two competing hypotheses, both of which have been separately assumed in the literature without the possibility to test the assumption in a parametric framework.

4 Example

We illustrate the working of the GECk with an application to a well studied data-set on ship damage incidents. The data are given in McCullagh and Nelder (MN) (1989) who apply a quasi-likelihood approach. Lawless (1987) analyzes the same data using a negative binomial model. We will proceed analogous to both studies and use a log-linear specification $E(y_i|x_i) = \lambda_i = \exp(x_i\beta)N_i, i = 1, \ldots, 35,$ for the expected number of damage incidents. The explanatory variables $x_i$ are dummy variables for ship type, year of construction, and period of operation. The factor $N_i$ is used to adjust for varying aggregate months of service. We give the estimates for the negative binomial model with $k = 1$ and $k = 0$ respectively and for the GECk in Table 1. The overall impression is that the estimates for both coefficients and asymptotic t-values are broadly similar. However the results deserve further comments. The results for the negative binomial model for $k = 1$ are identical to those in Lawless(1987). Since the estimated $\sigma^2$ is equal to 1, this is the limiting case of the Poisson model and the estimated coefficients are identical to those given in MN for the latter model. Comparing this result to those for the negative binomial model with $k = 0$ and for the GECk, it becomes apparent, that this corner estimate for $\sigma^2$ is mainly caused by the restriction on $k$ and that
relaxing the restriction, the estimated $\sigma^2$ goes up to 1.85 in the GECk. In fact, we can reject the negative binomial model with $k = 1$ against the more general GECk with an estimated $k$ of $-0.74$, at the 5% confidence level ($t_{(24,0.025)} = 2.064$). The Poisson hypothesis ($\sigma^2 = 1$), however, cannot be rejected by any of the alternatives confirming the remarks in Lawless (1987, p.220) for the lack of precision due to the modest sample size. As to the estimated standard errors, it should be noted, that the GECk standard errors are only slightly larger than the Poisson standard errors, indicating that the correction by a moment estimator $\hat{\sigma}^2 = 1.69$ proposed by MN unduly underestimates the asymptotic t-values.

5 Concluding remarks

The GECk is a flexible model that incorporates several of the approaches used in recent econometric work on count data as special cases. Instead of choosing a priori one of these models, the researcher can now let the data determine the most appropriate. This gain in flexibility has been achieved by introducing one additional parameter. The likelihood function takes a relatively simple form and an application to a dataset on ship damage incidents has demonstrated the advantages of the GECk.

References


### TABLE 1: Estimates in the ship damage example*

<table>
<thead>
<tr>
<th></th>
<th>Neg.Bin. ($k = 1$)</th>
<th>Neg.Bin. ($k = 0$)</th>
<th>GEC$k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-6.4061 (29.46)</td>
<td>-6.3921 (26.20)</td>
<td>-6.3694 (27.89)</td>
</tr>
<tr>
<td>Ship type:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-0.5433 (3.060)</td>
<td>-0.5521 (2.776)</td>
<td>-0.5672 (3.031)</td>
</tr>
<tr>
<td>C</td>
<td>-0.6874 (2.089)</td>
<td>-0.7362 (1.954)</td>
<td>-0.7328 (1.916)</td>
</tr>
<tr>
<td>D</td>
<td>-0.0759 (0.261)</td>
<td>-0.1155 (0.343)</td>
<td>-0.1194 (0.358)</td>
</tr>
<tr>
<td>E</td>
<td>0.3256 (1.380)</td>
<td>0.3036 (1.136)</td>
<td>0.2603 (0.973)</td>
</tr>
<tr>
<td>Year of</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>construction:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60-64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65-69</td>
<td>0.6971 (4.659)</td>
<td>0.6895 (4.073)</td>
<td>0.6674 (4.288)</td>
</tr>
<tr>
<td>70-74</td>
<td>0.8184 (4.821)</td>
<td>0.8117 (4.218)</td>
<td>0.7750 (4.254)</td>
</tr>
<tr>
<td>75-79</td>
<td>0.4534 (1.945)</td>
<td>0.4584 (1.735)</td>
<td>0.4535 (1.845)</td>
</tr>
<tr>
<td>Service period:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70-74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75-79</td>
<td>0.3845 (3.251)</td>
<td>0.3886 (2.891)</td>
<td>0.4048 (3.191)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1.0000 (-)**</td>
<td>1.2832 (0.902)</td>
<td>1.8501 (0.937)</td>
</tr>
<tr>
<td>$k$</td>
<td></td>
<td>-0.7424 (2.469)</td>
<td></td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>-68.29</td>
<td>-67.67</td>
<td>-67.00</td>
</tr>
</tbody>
</table>

*The dependent variable is the number of reported ship damage incidents. Asymptotic absolute t-values in parentheses ($H_0$: Coeff=0, exception: $\sigma^2 = 1$, $k = 1$).

**No meaningful standard error could be calculated due to the boundary estimate for $\sigma^2$.**