OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 59, 3 (1997)
0305-9049

PRACTITIONERS CORNER

Estimating Nested Count Data Models

*Atanu Saha and Diansheng Dong**

I. INTRODUCTION

In count data models the endogenous variable takes only non-negative integer values corresponding to the number of events occurring in a given interval of time or space. Examples of count data model applications include number of patents applied for by firms (Hausman, Hall, and Griliches, 1984), number of visits to physicians (Cameron and Trivedi, 1986), number of trips to a recreational area (Hellerstein, 1991), number of defective products in a manufacturing process (Lambert, 1992), and number of takeover bids received by a target firm after an initial bid (Jaggia and Thosar, 1993). Gurmu and Trivedi (1994) provide an excellent survey of the relevant literature.

The benchmark model for count data is the Poisson model. In the Poisson regression model, however, the conditional mean of the endogenous variable given the exogenous variables is equal to its conditional variance. To overcome this limitation several generalizations have been proposed. Among these, negative binomial (NB) models, in which the conditional variance can exceed the conditional mean (i.e., allow overdispersion), have been widely used. Within NB models, specifications differ in their implied relationship between the conditional mean and variance of the dependent variable. The purpose of this study is (a) to propose tests for selection among the Poisson and NB models by formally demonstrating that the loglikelihood function (LLF) of the general NB model nests the LLF of the Poisson and the two most widely used NB models as special cases, and (b) to propose estimation of the general NB model since it allows greater flexibility in the relationship between the mean and variance of the dependent variable than the widely used NB specifications. An application to micro-level data on the number of recreational boating trips illustrates the results.

*The authors thank Teofilo Ozuna for providing the data set used in this study.

II. PRELIMINARIES

Let y_i, $i=1, \ldots, n$ denote observations of an integer-valued discrete variable. Let x_i denote the ith row of the matrix of k regressors, and let β be the $k \times 1$ vector of parameters. The Poisson model assumes that y_i is independently distributed as a Poisson variate with $\lambda_i = \exp(x_i\beta) > 0$ being the parameter of the distribution. In this model:

$$E[y_i|x_i] = \text{Var}(y_i|x_i) = \lambda_i. \tag{1}$$

The negative binomial (NB) model's a probability density function (pdf) is:

$$\frac{\Gamma(y_i+\gamma_i)}{\Gamma(1+y_i)\Gamma(\gamma_i)} \cdot \left(\frac{\gamma_i}{\gamma_i+\theta_i}\right)^{\gamma_i} \left(\frac{\theta_i}{\gamma_i+\theta_i}\right)^{y_i}. \tag{2}$$

Cameron and Trivedi (1986) have proposed, without loss of generality, the parameterization $\theta_i = \exp(x_i\beta) = \lambda_i$ and $\gamma_i = 1/\alpha \exp(k \cdot (x_i\beta)) = \lambda_i^k/\alpha$, where α and k are non-negative parameters. Under this parameterization, the relationship between the conditional variance and mean of y_i becomes:

$$\text{Var}(y_i|x_i) = E[y_i|x_i] + \alpha E[y_i|x_i]^{2-k} = \lambda_i + \alpha \cdot \lambda_i^{2-k} \tag{3}$$

where α is the overdispersion parameter. The loglikelihood function (LLF) of the NB model is:

$$\sum_{i=1}^{n} \left\{ \ln \Gamma\left(y_i + \frac{Z_i}{\alpha}\right) - \ln \Gamma\left(\frac{Z_i}{\alpha}\right) - \ln \Gamma(1+y_i) \right.$$

$$\left. - \left(y_i + \frac{Z_i}{\alpha}\right) \ln\left(\frac{Z_i}{\alpha} + \exp(x_i\beta)\right) + y_i \cdot (x_i\beta) + \frac{Z_i}{\alpha} \cdot (k(x_i\beta) - \ln \alpha) \right\} \tag{4}$$

where $Z_i = \lambda_i^k$. The two most commonly used specifications in the econometrics literature are the NBI and NBII models obtained by setting $k=1$ and $k=0$, respectively. Thus, the alternative models imply the following relations between the conditional mean and variance of y_i:

Poisson ($\alpha=0$): $\text{Var}(y_i|x_i) = E[y_i|x_i] = \lambda_i$

NBI ($k=1$): $\text{Var}(y_i|x_i) = E[y_i|x_i] \cdot (1+\alpha) = \lambda_i \cdot (1+\alpha)$

NBII ($k=0$): $\text{Var}(y_i|x_i) = E[y_i|x_i] \cdot (1+\alpha E[y_i|x_i]) = \lambda_i \cdot (1+\alpha\lambda_i)$.

III. MODEL SELECTION AND ESTIMATION ISSUES

It is clear from the expressions for conditional mean and variance that in both NB models $\text{Var}(y_i|x_i) = E[y_i|x_i] = \lambda_i$, as in the Poisson, if the over-

dispersion parameter α is zero. This observation is central to the model selection tests in the existing literature (see Lee, 1986, and the references therein). For example, to choose between Poisson and NBII, Cameron and Trivedi (1996) propose to perform the following auxiliary OLS regression without a constant:

$$\frac{(y_i-\hat{\lambda}_i)^2-y_i}{\hat{\lambda}_i}=\alpha\hat{\lambda}_i+u_i \tag{5}$$

where $\hat{\lambda}_i=\exp(x_i\hat{\beta})$, with $\hat{\beta}$ being the estimate of β from the Poisson model, and u_i denotes the error term. The t-statistic for α is asymptotically normal under the null hypothesis of no overdispersion against the alternative of overdispersion of the NBII form. To test the Poisson against the NBI model, (5) is replaced by:

$$\frac{(y_i-\hat{\lambda}_i)^2-y_i}{\hat{\lambda}_i}=\alpha+u_i \tag{5′}$$

Alternatively, one may estimate the NBI and NBII models and use an asymptotic t-test on the estimate of the overdispersion parameter α.

The test procedures discussed above have been widely used and warrant comments. Both tests are based on the null $H_0:\alpha=0$ which implies a specification wherein the conditional mean and variance of y_i are equal. Clearly, Poisson is not the only choice that meets this criterion. There are numerous distributions that exhibit equality of first two moments under appropriate parameter values. For the non-rejection of the null $H_0:\alpha=0$ to imply that *Poisson* is the preferred model, it is necessary to show that the general NB pdf collapses to the Poisson pdf when α is zero.

Another feature of the Poisson versus NB model selection tests outlined above is their lack of generality. The values of k (1 and 0) that yield the NBI and NBII models are arbitrary. For example, k can take a value of 2, in which case, the relationship between the moments becomes (see equation (3)): $\mathrm{Var}(y_i|x_i)=\lambda_i+\alpha$, a relationship no less plausible *a priori* than that in NBI or NBII. More generally, k can take a range of values (not necessarily integers) to yield valid NB model specifications. But the model selection tests discussed above hold only for the two values of k, zero and one, when, in fact, neither NBI or NBII may be the preferred model.

The last point also brings up the issue of selection among NB model specifications. In the existing literature, tests based on auxiliary regressions similar to (5) and (5′) have been proposed for choosing between NBI and NBII (see, for example, Cameron and Trivedi, 1986; Ozuna and Gomez, 1995). The regression-based tests involve estimating a set of four equations; they are not reproduced here in the interest of brevity. A far

more direct approach to selection among NB models would be to estimate the parameter k by maximizing a general NB likelihood function that nests the NBI and NBII likelihoods as special cases. One can then apply a standard hypothesis test on the maximum likelihood (ML) estimate of k to decide which NB specification is preferred. In addition to being more direct, this procedure is more general since it allows the possibility of rejecting both, NBI and NBII.

The foregoing considerations motivate the following results:

(i) For all finite values of k, the LLF of the general NB model given in (4) approaches the LLF of the Poisson model as α approaches zero.

(ii) For all finite values of α, the LLF of the general NB model becomes identical to the LLF of the NBI and NBII when $k=1$ and $k=0$, respectively.

Part (ii) of the result is obvious from (4) and is presented here in the interest of completeness. The proof of part (i) is available from the authors on request. It is not unexpected that the general NB pdf nests the Poisson pdf as a limiting case when $\alpha=0$ because both the NB and Poisson belong to the Katz (1963) family of distributions. However, despite a thorough search, we failed to find a formal statement and proof of the result in the existing literature.

Resulting (i) and (ii), in light of the arguments leading up to it, suggest that k, α, and β should be jointly estimated by maximizing the likelihood function in (4), which we will denote as the LLF of the general NB(NBG) model. Because (4) is highly non-linear in parameters, convergence in an iterative procedure to maximize this LLF can be a problem unless appropriate starting values are chosen. We propose the following steps. Step 1: the parameters α and k are estimated through non-linear least squares using the following auxiliary regression equation:

$$\frac{(y_i-\hat{\lambda}_i)^2-y_i}{\hat{\lambda}_i}=\alpha(\hat{\lambda}_i)^{1-k}+u_i, \tag{6}$$

where $\hat{\lambda}_i$ has the same definition as in (5). Step 2: holding the parameters α and k at their estimated values, maximize the LLF in (4) with respect to the βs. Step 3: using the new estimate of β, denoted by $\hat{\hat{\beta}}$, form $\hat{\hat{\lambda}}_i=\exp(x_i\hat{\hat{\beta}})$; use $\hat{\hat{\beta}}_i$ to re-estimate (6), getting a second round estimate of α and k. These estimates of α and k, and $\hat{\hat{\beta}}$ provide the starting values for the maximization of (4).

Our result on the convergence of NBG's LLF to that of Poisson also has implications for model selection. The result implies that one can simply use an asymptotic t-test using the ML estimate of α from the NBG model to test the null $H_0:\alpha=0$. Although the null hypothesis for the test is not different from the tests in the existing literature, there are import-

ant differences in terms of inference. In light of Result (i), non-rejection of the null implies that the true model is Poisson, and not merely any model wherein the conditional mean and variance of y_i are equal.

IV. AN APPLICATION

The empirical application belongs to the category of recreational demand models. The data, collected through a survey of registered boat owners in East Texas, have been used by Sellar, Stoll and Chauds (1985) and Ozuna and Gomez (1995). The Sellar *et al.* paper contains detailed description of the data set. The data on the dependent variable is the number of trips made by survery respondents to Lake Sommerville in East Texas. Summary statistics on the exogenous variables (except the intercept) with a brief explanation are presented in Table 1. The travel costs to the lakes can be viewed as a proxy for 'price' of receational boating.

All four models, Poisson, NBI, NBII, and NBG were estimated by maximizing their respective LLFs. In the NBG model, the null hypothesis of Poisson, $H_0: \alpha = 0$, is clearly rejected both by the asymptotic *t*-test and the likelihood ratio test. The Poisson model is also rejected by NBI and NBII. Within NB models, both NBI and NBII are also unambiguously rejected. The *P*-values associated with the asymptotic *t*-statistic for $H_0: k = 0$ (null for NBII) and $H_0: k = 1$ (null for NBI) are 0.0005 and

TABLE 1
Summary Statistics

Variable name	*Explanation*	*Mean*	*Standard deviation*
Quality	Respondents' rating score for Lake Sommerville; (0 = worst, 5 = best)	1.4188	1.8120
Ski	Dummy variable, equals one if respondent skied at Lake Sommerville	0.3672	0.4824
Income	Income of respondent's household head	3.8528	1.8519
CostS	Cost of travelling to and from Lake Sommerville ($)	60.038	46.339
CostC	Cost of travelling to and from Lake Conroe ($)	55.424	46.683
CostH	Cost of travelling to and from Lake Houston ($)	55.869	45.900
Trip	Dependent Variable; number of trips respondent took to Lake Sommerville; max. = 88; min. = 0; % of zeros = 63.3	2.2443	6.2925
	Total number of observations	659	

0.0009, respectively. This finding supports our contention that NBI and NBII are arbitrary specifications and may be rejected in many applications.

The inferential consequences of incorrect model specification are evident from the marginal effects of regressors in different models. In our model all explanatory variables were normalized by their mean; hence the jth regressor's elasticity evaluated at the sample mean is simply β_j. In the interest of brevity, we have compared only quality and own price (proxied by travel cost) elasticities of recreational demand across the four models. The results are presented in Table 3. We have taken the elasticity estimate from the NBG model as a benchmark and have computed its percentage difference from the corresponding elasticity in the other three

TABLE 2
Estimation Results

		*Models**			
Coefficient	*Variable*	*Poisson*	*NBI*	*NBII*	*NBG*
β_1	Quality	0.7047 (28.800)	0.8372 (16.958)	1.0321 (7.433)	0.9787 (16.945)
β_2	Ski	0.2322 (11.786)	0.1389 (3.370)	0.2244 (4.098)	0.1912 (3.667)
β_3	Income	−0.2966 (4.180)	−0.0723 (0.593)	−0.0865 (0.184)	−0.1343 (0.909)
β_4	CostS	−4.7770 (31.715)	−3.3340 (10.099)	−5.8455 (11.233)	−4.5657 (8.654)
β_5	CostC	1.2059 (4.698)	0.6346 (1.121)	2.6299 (2.154)	1.1868 (1.669)
β_6	CostH	2.6111 (11.756)	2.6500 (4.963)	2.6858 (2.005)	2.7326 (4.962)
β_7	Intercept	−0.0016 (0.017)	−0.6646 (3.317)	−1.1408 (1.509)	−0.9916 (6.818)
α			5.7421 (8.485)	1.3118 (7.606)	2.7962 (4.341)
k					0.5137 (3.620)
LLF value		−1367.53	−816.27	−821.42	−813.03
Asymptotic t-stat. for $H_0: k=1$					3.3072 (0.0009)**

Notes:
* Absolute value of asymptotic t-ratios in parentheses.
** Denotes P-value.

TABLE 3
Elasticity Comparison Across Models

Elasticity of:	*NBG estimate*	*% difference from NBG model estimate*		
		Poisson	*NBI*	*NBII*
Quality	0.9787	−28.0	−14.5	+5.5
Own price (CostS)	−4.5657	−4.3	+25.3	−26.3

models. For example, the estimated quality elasticity in the Poisson model is 28 percent larger than the corresponding estimate in NBG. That estimates of elasticity would differ across models is to be anticipated; but the magnitude of the differences is somewhat unexpected and it, once again, highlights the need for a general model specification.

V. CONCLUDING COMMENTS

Count data models have found a wide variety of applications not only in applied economics and finance but also in diverse fields ranging from biometrics to political science. Poisson and negative binomial (NB) are two most extensively used model specifications in count data analysis. Unlike the Poisson, NB models allow overdispersion, that is, allow the conditional variance of the dependent variable to be larger than the conditional mean. Since overdispersion is frequently encountered in count data, two particular NB model specifications, NBI and NBII have been especially popular. However, these models impose arbitrary restrictions on the relationship between the conditional mean and variance of the dependent variable, limiting their generality.

The existing model selection tests (for Poisson versus NB) are framed only in terms of the alternatives of NBI or NBII when, in fact, the preferred model may be neither. Our contribution lies in formally demonstrating that the loglikelihood function (LLF) of a general NB model, that nests the LLFs of NBI and NBII, collapses to the Poisson model's LLF when the overdispersion parameter approaches zero. Thus, the LLF of all three model specifications, Poisson, NBI, and NBII, are parametrically nested within the general NB model's LLF as special cases; these parameter restrictions can be tested by simple asymptotic *t*-tests.

The empirical application, which uses micro-level data on recreational boating, provided support for the paper's main theme. Tests clearly rejected not only the Poisson, but also NBI and NBII, in favour of a different NB model, underscoring the importance of the general model specification.

Micronomics Inc., Los Angeles,
Texas A&M University

Date of Receipt of Final Manuscript: November 1996

REFERENCES

Cameron, A. Colin and Trivedi, P. (1986). 'Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests', *Journal of Applied Econometrics*, Vol. 1, pp. 29–54.

Cameron, A. Colin and Trivedi, P. (1996). 'Count Data Models for Financial Data', in Maddala, G. S. and Rao, C. R. (eds), *Handbook of Statistics: Statistical Methods in Finance*, North-Holland, Amsterdam.

Gourieroux, C., Monfort, A. and Trognon, A. (1984). 'Pseudo Maximum Likelihood Methods: Application to Poisson Models', *Econometrica*, Vol. 52, pp. 701–20.

Gurmu, S., Trivedi, P. K. (1994). 'Recent Developments in Models of Event Counts: A Survey', *Discussion Paper No. 261, Thomas Jefferson Center*, University of Virginia, Charlottesville.

Hausman, J. A., Lo, A. W., MacKinlay, A. C. (1992). 'An Ordered Probit Analysis of Transaction Stock Prices', *Journal of Financial Economics*, Vol. 31, pp. 319–79.

Hausman, J. A., Hall, B. and Griliches, Z. (1984). 'Economic Models for Count Data with an Application to the Patents–R&D Relationship', *Econometrica*, Vol. 52, pp. 909–38.

Hellerstein, D. M. (1991). 'Using Count Data Models in Travel Cost Analysis with Aggregate Data', *American Journal of Agricultural Economics*, Vol. 73, pp. 860–66.

Jaggia, S. and Thosar, S. (1993). 'Multiple Bids as a Consequence of Target Management Resistance: A Count Data Approach', *Review of Quantitative Finance and Accounting*, Vol. 00, pp. 447–57.

Katz, L. (1963). 'Unified Treatment of a Broad Class of Discrete Probability Distributions', *Proceedings of the International Symposium on Discrete Distributions*, Montreal, pp. 172–82.

Lambert, D. (1992). 'Zero Inflated Poisson Regression with and Application to Defects in Manufacturing', *Technometrics*, Vol. 34, pp. 1–14.

Lee, Lung-Fei (1986). 'Specification Test for Poisson Regression Models', *International Economic Review*, Vol. 26, pp. 689–706.

Ozuna, T. R., Gomez, I. A. (1995). 'Specification and Testing of Count Data Recreation Demand Functions', *Empirical Economics*, vol. 20, pp. 43–55.

Sellar, C., Stoll, J. R., Chavas, J. P. (1985) 'Validation of Empirical Measures of Welfare Change: Comparison of Nonmarket Techniques', *Land Economics*, Vol. 61, pp. 156–75.