Functional Form and Heterogeneity in Models for Count Data

Functional Form and Heterogeneity in Models for Count Data

William Greene

Department of Economics
Stern School of Business
New York University
New York
NY 10012
USA
wgreene@stern.nyu.edu

the essence of knowledge

Boston – Delft

Foundations and Trends® in Econometrics

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is W. Greene, Functional Form and Heterogeneity in Models for Count Data, Foundations and Trends® in Econometrics, vol 1, no 2, pp 113–218, 2005

ISBN: 978-1-60198-054-0

Foundations and Trends® in Econometrics

Volume 1 Issue 2, 2005

Editorial Scope

Foundations and Trends® in Econometrics will publish survey and tutorial articles in the following topics:

- Identification
- Model Choice and Specification Analysis
- Non-linear Regression Models
- Simultaneous Equation Models
- Estimation Frameworks
- Biased Estimation
- Computational Problems
- Microeconometrics
- Treatment Modeling
- Discrete Choice Modeling
- Models for Count Data
- Duration Models
- Limited Dependent Variables
- Panel Data
- Dynamic Specification
- Inference and Causality
- Continuous Time Stochastic Models
- Modeling Non-linear Time Series
- Unit Roots
- Cointegration
- Latent Variable Models
- Qualitative Response Models
- Hypothesis Testing
- Interactions-based Models
- Duration Models
- Financial Econometrics
- Measurement Error in Survey Data
- Productivity Measurement and Analysis
- Semiparametric and Nonparametric Estimation
- Bootstrap Methods
- Nonstationary Time Series
- Robust Estimation

Information for Librarians

Foundations and Trends® in Econometrics, 2005, Volume 1, 4 issues. ISSN paper version 1551-3076. ISSN online version 1551-3084. Also available as a combined paper and online subscription.

Foundations and Trends® in
Econometrics
Vol. 1, No. 2 (2005) 113–218

DOI: 10.1561/0800000008

Functional Form and Heterogeneity in Models for Count Data

William Greene

Department of Economics, Stern School of Business, New York University, 44 West 4th St., New York, NY 10012, USA, wgreene@stern.nyu.edu

Abstract

This study presents several extensions of the most familiar models for count data, the Poisson and negative binomial models. We develop an encompassing model for two well-known variants of the negative binomial model (the NB1 and NB2 forms). We then analyze some alternative approaches to the standard log gamma model for introducing heterogeneity into the loglinear conditional means for these models. The lognormal model provides a versatile alternative specification that is more flexible (and more natural) than the log gamma form, and provides a platform for several "two part" extensions, including zero inflation, hurdle, and sample selection models. (We briefly present some alternative approaches to modeling heterogeneity.) We also resolve some features in Hausman, Hall and Griliches (1984, Economic models for count data with an application to the patents–R&D relationship, *Econometrica* **52**, 909–938) widely used panel data treatments for the Poisson and negative binomial models that appear to conflict with more familiar models of fixed and random effects. Finally, we consider a bivariate Poisson model that is also based on the lognormal heterogeneity model. Two recent applications have used this model.

We suggest that the correlation estimated in their model frameworks is an ambiguous measure of the correlation of the variables of interest, and may substantially overstate it. We conclude with a detailed application of the proposed methods using the data employed in one of the two aforementioned bivariate Poisson studies.

Keywords: Poisson regression; negative binomial; panel data; heterogeneity; lognormal; bivariate poisson; zero inflation; two part model; hurdle model.

JEL codes: C14, C23, C25

Contents

1

Introduction*

Models for count data have been prominent in many branches of the recent applied literature, for example, in health economics (e.g., in numbers of visits to health facilities[1]) management (e.g., numbers of patents[2]), and industrial organization (e.g., numbers of entrants to markets[3]). The foundational building block in this modeling framework is the Poisson regression model.[4] But, because of its implicit restriction on the distribution of observed counts — in the Poisson model, the variance of the random variable is constrained to equal the mean — researchers routinely employ more general specifications, usually the negative binomial (NB) model which is the standard choice for a basic count data model.[5] There are also many applications that extend the

* This study has benefited from the helpful comments of Andrew Jones on an earlier version. Any remaining errors are the author's responsibility.

1 Jones (2000), Munkin and Trivedi (1999), Riphahn et al. (2003). See, as well, Cameron and Trivedi (2005).

2 Hausman et al. (1984) and Wang et al. (1998).

3 Asplund and Sandin (1999).

4 Hausman et al. (1984), Cameron and Trivedi (1986, 1998), and Winkelmann (2003).

5 The NB model is by far the most common specification. See Hilbe (2007). The latent class (finite mixture) and random parameters forms have also been employed. See, e.g., Wang et al., op. cit., Deb and Trivedi (1997) and Bago d'Uva (2006).

Poisson and NB models to accommodate special features of the data generating process, such as hurdle effects,[6] zero inflation,[7] and sample selection.[8] The basic models for panel data, fixed and random effects, have also been extended to the Poisson and NB models for counts.[9] Finally, there have been several proposals for extending the Poisson model to bivariate and multivariate settings.[10] This list includes a substantial fraction of the received extensions of the basic Poisson and NB models. There have, however, been scores of further refinements and extensions that are documented in a huge literature and several book length treatments such as Cameron and Trivedi (CT) 1998, Winkelmann (2003), and Hilbe (2007).

This paper will survey some practical extensions of the Poisson and NB models that practitioners can employ to refine the specifications or broaden their reach into new situations. We will also resolve some apparent inconsistencies of the panel data models with other more familiar results for the linear regression model.

- There are two well known, nonnested forms of the negative binomial model, denoted NB1 and NB2 in the literature. (See CT (1986)). Researchers have typically chosen one form or the other (typically NB2), but not generally formed a preference for one or the other. We propose an encompassing model that nests both of them parametrically and allows a statistical test of the two functional forms against a more general alternative.
- The NB model arises as the result of the introduction of log gamma distributed unobserved heterogeneity into the loglinear Poisson mean. A lognormal model provides a suitable alternative specification that is more flexible than the

[6] See, e.g., Mullahy (1986), Rose et al. (2006) and Yen and Adamowicz (1994) on separately modeling participation and usage.

[7] See, e.g., Heilbron (1992) and Lambert (1992) on industrial processes, Greene (1994) on credit defaults and Zorn (1998) on Supreme Court Decisions.

[8] See, e.g., Greene (1995) on derogatory credit reports and Terza (1998).

[9] See, again, Hausman et al. (1984) on the relationship between patents and research and development.

[10] See King (1989), Munkin and Trivedi (1999) and Riphahn et al. (2003).

log gamma form, and provides a platform for several useful extensions, including hurdle, zero inflation, and sample selection models.[11] We will develop this alternative to the NB model, then show how it can be used to accommodate in a natural fashion, e.g., sample selection, hurdle effects, and a new model for zero inflation.

- The most familiar panel data treatments, fixed effects (FE) and random effects (RE), for count models were proposed by Hausman et al. (HHG) (1984). The Poisson FE model is particularly simple to analyze, and has long been recognized as one of a very few known models in which the incidental parameters problem (see Neyman and Scott (1948) and Lancaster (2000)) is, in fact, not a problem. The same is not true of the NB model. Researchers are sometimes surprised to find that the HHG formulation of the FE NB model allows an overall constant — a quirk that has also been documented elsewhere. (See Allison (2000) and Allison and Waterman (2002), for example.) We resolve the source of the ambiguity, and consider the difference between the HHG FE NB model and a "true" FE NB model that appears in the familiar index function form. The true FE NB model has not been used by applied researchers, probably because of the absence of a computational method. We have developed a method of computing the true FE NB model that allows a comparison to the HHG formulation.
- The familiar RE Poisson model using a log gamma heterogeneity term produces the NB model. We consider the lognormal model as an alternative, again, as a vehicle for more interesting specifications, and compare it to the HHG formulation. The HHG RE NB model is also unlike what

[11] The Poisson lognormal mixture model has a long history, apparently beginning with Pielou (1969) and Bulmer (1974), both of whom build on a Gaussian model proposed by Preston (1948). Hinde (1982) describes the context of "generalized linear models." It appears in the econometrics literature with Greene (1994, 1995), Terza (1998), Million (1998), Geil et al. (1997), and several recent applications including, e.g., Winkelmann (2003), Van Ourti (2004), and Riphahn et al. (2003).

might seem the natural application in which the heterogeneity term appears as an additive common effect in the conditional mean. Once again, this was a practical solution to the problem. The lognormal model provides a means of specifying the RE NB model in a natural index function form. We will develop this model, and, once again, compare it to the HHG formulation.

- Two recent applications, Munkin and Trivedi (1999) and RWM (2003), have used a form of the bivariate Poisson model in which the correlation is introduced through additive correlated variables in the conditional mean functions. Both of these studies have misinterpreted (and overstated) the correlation coefficient estimated in their model frameworks. What they have specified is correlation between the logs of the conditional mean functions. How this translates to correlation between the count variables themselves is quite unclear. We will examine this in detail.

The study is organized as follows: Section 2 will detail the basic modeling frameworks for count data, the Poisson and NB models and will propose models for observed and unobserved heterogeneity in count data. This section will suggest a parameterization of the NB model that introduces measured heterogeneity into the scaling parameter. We then develop the NBP model to encompass NB1 and NB2. Finally, we propose the lognormal model as an alternative to the log gamma model that produces the NB specification. Section 3 will extend the lognormal model to several two part models. Section 4 will examine the fixed and random effects models for panel data. Section 5 will consider applications of the Bivariate Poisson model. The various model extensions proposed are applied to the RWM panel data on health care utilization in Section 6. Some conclusions are drawn in Section 7.

As documented in a vast literature, there are many aspects of modeling count data. This study is focused on two large issues, first, the accommodation of overdispersion and heterogeneity in the basic count framework and, second, the functional form of the conditional mean and the extension of models of heterogeneity to models for panel data and sources of correlation across outcomes. The first of these is more straightforward. In principle, these are elements of the conditional variance of the distribution of counts that can be analyzed apart from the conditional mean. Robust inference methods for basic models can be relied upon to preserve the validity of estimation and inference procedures. The second feature motivates the development of more intricate models such as the two part, panel and bivariate models presented in what follows.

References

Abramovitz, M. and I. Stegun (1971), *Handbook of Mathematical Functions*. New York: Dover Press.

Aigner, D., K. Lovell, and P. Schmidt (1977), 'Formulation and estimation of stochastic frontier production function models'. *Journal of Econometrics* **6**, 21–37.

Allison, P. (2000), 'Problems with fixed-effects negative binomial models'. Manuscript, Department of Sociology, University of Pennsylvania.

Allison, P. and R. Waterman (2002), 'Fixed-effects negative binomial regression models'. *Sociological Methodology* **32**, 247–265.

Amid, A. (1978), 'A note on the moments of the generalized negative binomial distribution and on certain properties of the distribution'. *SIAM Journal of Applied Mathematics* **34**(2), 223–224.

Asplund, M. and R. Sandin (1999), 'The number of firms and production capacity in relation to market size'. *The Journal of Industrial Economics* **47**(1), 69–85.

Bago d'Uva, T. (2006), 'Latent class models for utilisation of health care'. *Health Economics* **15**, 329–343.

Bulmer, M. (1974), 'On fitting the Poisson lognormal distribution to species-abundance data'. *Biometrics* **30**, 101–110.

Butler, J. and R. Moffitt (1982), 'A computationally efficient quadrature procedure for the one factor multinomial probit model'. *Econometrica* **50**, 761–764.

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

Cameron, C. and P. Trivedi (1998), *Regression Analysis of Count Data*. New York: Cambridge University Press.

Cameron, C. and P. Trivedi (2005), *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Clark, A. and F. Etile (2006), 'Don't give up on me baby: Spousal correlation in smoking behaviours'. *Journal of Health Economics* **25**, 958–978.

Clark, A., F. Etile, F. Postel-Vinay, C. Senik, and K. van Straeten (2005), 'Heterogeneity in reported well being: Evidence from twelve European countries'. *The Economic Journal* **115**, C118–C132.

Deb, P. and P. Trivedi (1997), 'Demand for medical care by the elderly: A finite mixture approach'. *Journal of Applied Econometrics* **12**(3), 313–336.

Econometric Software Inc. (2003), 'LIMDEP Version 8.0 and NLOGIT Version 3.0'. Plainview, New York.

Econometric Software Inc. (2007), 'LIMDEP Version 9.0 and NLOGIT Version 4.0'. Plainview, New York.

Freedman, D. (2006), 'On the so-called "Huber Sandwich Estimator" and "Robust Standard Errors"'. *The American Statistician* **60**(4), 299–302.

Geil, P., A. R. M. Rotte, and K. Zimmermann (1997), 'Economic incentives and hospitalization in Germany'. *Journal of Applied Econometrics* **12**(3), 295–312.

Gourieroux, C. and A. Monfort (1996), *Simulation-Based Methods Econometric Methods*. Oxford: Oxford University Press.

Gourieroux, C., A. Monfort, and A. Trognon (1984), 'Pseudo maximum likelihood methods: Applications to poisson models'. *Econometrica* **52**, 701–720.

Greene, W. (1994), 'Accounting for excess zeros and sample selection in Poisson and negative binomial regression models'. Working Paper

No. EC-94-10, Department of Economics, Stern School of Business, New York University.

Greene, W. (1995), 'Sample selection in the Poisson regression model'. Working Paper EC-95-06, Department of Economics, Stern School of Business, New York University.

Greene, W. (1997), 'FIML estimation of sample selection models for count data'. Working Paper EC-97-02, Department of Economics, Stern School of Business, New York University.

Greene, W. (2004), 'The behavior of the fixed effects estimator in nonlinear models'. *The Econometrics Journal* **7**(1), 98–119.

Greene, W. (2006), 'A general approach to incorporating "Selectivity" in a model'. Working Paper EC-06-10, Department of Economics, Stern School of Business, New York University.

Greene, W. (2007), 'LIMDEP 9.0 Reference Guide'. Econometric Software, Inc., Plainview.

Greene, W. (2008a), *Econometric Analysis*. 6th edition. Prentice Hall: Englewood Cliffs.

Greene, W. (2008b), 'The econometric approach to efficiency analysis'. In: H. Fried, C. A. K. Lovell, and S. Schmidt (eds.): *The Measurement of Productive Efficiency*. 2nd edition, Oxford University Press (forthcoming).

Greene, W., M. Harris, B. Hollingsworth, and P. Maitra (2007), 'A bivariate latent class correlated generalized ordered probit model with an application to modeling observed obesity levels'. Department of Econometrics, Monash University, Melbourne, manuscript.

Hall, D. B. (2000), 'Zero inflated poisson and binomial regression with random effects: A case study'. *Biometrics* **56**, 1030–1039.

Hausman, J., B. Hall, and Z. Griliches (1984), 'Economic models for count data with an application to the patents–R&D relationship'. *Econometrica* **52**, 909–938.

Heckman, J. (1979), 'Sample selection bias as a specification error'. *Econometrica* **47**, 153–161.

Heilbron, D. (1992), 'Zero altered and other regression models for count data with added zeros'. *Biometrical Journal* **36**, 531–547.

Hilbe, J. M. (2007), *Negative Binomial Regression*. Cambridge, UK: Cambridge University Press.

Hinde, J. (1982), 'Compound Poisson regression models'. In: R. Gilchrist (ed.): *GLIM, 1982: Proceedings of the International Conference on Generalized Linear Models*. Berlin: Springer.

Hur, K. (1998), 'A random effects zero inflated poisson regression model for clustered extra — zero counts'. Dissertation, Dartmouth Medical School.

Jain, C. and P. Consul (1971), 'A generalized negative binomial distribution'. *SIAM Journal of Applied Mathematics* **21**(4), 501–503.

Jochmann, M. and R. Leon-Gonzalez (2004), 'Estimating the demand for health care with panel data: A semiparametric Bayesian approach'. *Health Economics* **13**, 1003–1014.

Jones, A. (2000), 'Health econometrics'. In: A. J. Culyer and J. P. Newhouse (eds.): *Handbook of Health Economics*, Vol. 1. Elsevier, pp. 265–344.

Jung, R. and R. Winkelmann (1993), 'Two aspects of labor mobility: A bivariate Poisson regression approach'. *Empirical Economics* **18**, 543–556.

Karlis, D. and I. Ntzoufras (2003), 'Analysis of sports data by using bivariate poisson models'. *Journal of the Royal Statistical Society: Series D (The Statistician)* **52**(3), 381–393.

King, G. (1989), 'A seemingly unrelated Poisson regression model'. *Sociological Methods and Research* **17**(3), 235–255.

Kocherlakota, S. and K. Kocherlakota (1992), *Bivariate Discrete Distributions*. New York: Marcel Dekker.

Krinsky, I. and L. Robb (1986), 'On approximating the statistical properties of elasticities'. *Review of Economics and Statistics* **68**(4), 715–719.

Lambert, D. (1992), 'Zero-inflated poisson regression, with an application to defects in manufacturing'. *Technometrics* **34**(1), 1–14.

Lancaster, T. (2000), 'The incidental parameters problem since 1948'. *Journal of Econometrics* **95**(2), 391–414.

McLachlan, G. and D. Peel (2000), *Finite Mixture Models*. New York: John Wiley and Sons.

Million, A. (1998), 'Models for correlated count data'. Unpublished dissertation, University of Munich.

Mullahy, J. (1986), 'Specification and testing of some modified count data models'. *Journal of Econometrics* **33**, 341–365.

Munkin, M. and P. Trivedi (1999), 'Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application'. *Econometrics Journal* **2**, 29–49.

Murphy, K. and R. Topel (1985), 'Estimation and inference in two step econometric models'. *Journal of Business and Economic Statistics* **3**, 370–379. Reprinted, 20, 2002, pp. 88–97.

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

Pielou, E. (1969), *An Introduction to Mathematical Ecology*. New York: John Wiley and Sons.

Preston, F. (1948), 'The commonness and rarity of species'. *Ecology* **9**, 254–283.

Riphahn, R., A. Wambach, and A. Million (2003), 'Incentive effects in the demand for health care: A bivariate panel count data estimation'. *Journal of Applied Econometrics* **18**(4), 387–405.

Rose, C., W. Martin, K. Wannemuehler, and B. Plikaytis (2006), 'On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data'. *Journal of Biopharmaceutical Statistics* **16**(4), 463–481.

Stata (2006), Stata User's Guide, Version 9. College Station, TX: Stata Press.

Terza, J. (1985), 'A tobit type estimator for the censored poisson regression model'. *Economics Letters* **18**, 361–365.

Terza, J. (1998), 'Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects'. *Journal of Econometrics* **84**(1), 129–154.

Train, K. (2003), *Discrete Choice Models with Simulation*. Cambridge: Cambridge University Press.

Van Ourti, T. (2004), 'Measuring horizontal inequity in belgian health care using a gaussian random effects two part count model'. *Health Economics* **13**, 705–724.

Vuong, Q. (1989), 'Likelihood ratio tests for model selection and non-nested hypotheses'. *Econometrica* **57**, 307–344.

Wang, P., I. Cockburn, and L. Puterman (1998), 'Analysis of patent data — A mixed-Poisson-regression-model approach'. *Journal of Business and Economic Statistics* **16**(1), 27–41.

Winkelmann, R. (2003), *Econometric Analysis of Count Data*. 4th edition. Heidelberg: Springer Verlag.

Winkelmann, R. and K. Zimmermann (1991), 'A new approach for modeling economic count data'. *Economics Letters* **37**, 139–143.

Winkelmann, R. and K. Zimmermann (1995), 'Recent developments in count data modelling: Theory and application'. *Journal of Economic Surveys* **9**(1), 1–36.

Wong, W. and F. Famoye (1997), 'Modeling household fertility decisions with generalized poisson regression'. *Journal of Population Economics* **10**, 273–283.

Xie, H., K. Hur, and G. McHugo (2006), 'Using random-effects zero-inflated Poisson model to analyze longitudinal count data with extra zeros'. Manuscript, Dartmouth Medical School, http://isi.cbs.nl/iamamember/CD2/pdf/582.PDF.

Yen, S. and V. Adamowicz (1994), 'Participation, trip frequency and site choice: A Multinomial-Poisson hurdle model of recreation demand'. *Canadian Journal of Agricultural Economics* **42**(1), 65–76.

Zorn, C. (1998), 'An analytic and empirical examination of zero-inflated and hurdle Poisson specifications'. *Sociological Methods and Research* **26**(3), 368–400.