Economics Bulletin

Volume 31, Issue 1

Semiparametric estimation of on-stie count data models

Masaki Narukawa
Tohoku University

Katsuhito Nohara
Tohoku University

Abstract

This article proposes semiparametric estimation of on-site count data models based on a series expansion approach of Gurmu, Rilstone and Stern (1999, Journal of Econometrics 88, 123-150), which is flexible and adaptable for a form of overdispersion as long as the exponential mean parameterization is given. We also provide the empirical illustration of demand for a recreation site. The result suggests that the existing parametric approaches will cause wrong statistical inference for the on-site count data.

The authors are grateful to an anonymous referee for the helpful comments on an earlier version. The research of the second author was supported by Grant-in-Aid for Research Activity Start-up from the Japan Society for the Promotion of Science.
Citation: Masaki Narukawa and Katsuhito Nohara, (2011) "Semiparametric estimation of on-stie count data models", *Economics Bulletin*, Vol. 31 no.1 pp. 584-590.
Submitted: Jul 26 2010. **Published:** February 15, 2011.

1. Introduction

Count data collected by on-site sampling are often employed in the analysis of demand for a recreation site using travel cost methods, and there is a large literature, for instance, Shaw (1988), Englin and Shonkwiler (1995), Loomis (2003), Martínez-Espiñeira and Amoako-Tuffour (2008), and Nohara (2010). They, however, applied only parametric count regression models including the Poisson and the negative binomial to the analysis of their on-site count data. Therefore, if there occurs overdispersion which means the variance of the underlying random variable exceeds the mean, or its misspecification, the standard errors are grossly underestimated in the usual maximum likelihood estimation. This underestimation implies that the t-statistics are over-inflated (see, e.g., Cameron and Trivedi, 1998, ch.3). Moreover, as shown by Santos Silva (1997), unlike the results of Gourieroux et al. (1984), in the case of on-site sampling the estimates of parameters of interest by the Poisson pseudo likelihood are no longer consistent, even if the conditional mean is correctly specified. We can obtain the consistent estimates only under the restrictive conditions where the correct specifications of the first two conditional moments are given. These facts would cause wrong statistical inference for on-site count data and lead to incorrect conclusions of empirical studies.

In order to avoid such defects in the parametric estimation, this article proposes more flexible estimation of on-site count data models based on a series expansion. This extends the semiparametric approach of Gurmu et al. (1999) to on-site sampling count data by use of Shaw's (1988) correction for its characteristic problems. The proposed semiparametric approach nests the existing parametric ones such as the Poisson and negative binomial models in the case of on-site sampling as a special case. It should be noted that the series expansion approach does not cover underdispersed count data, as pointed out by Cameron and Johansson (1997, p.204). However, this range seems to be sufficient from an empirical point of view, because it is well known that on-site count data such as the number of trips is often observed with overdispersed (see, e.g., Martínez-Espiñeira and Amoako-Tuffour, 2008, p.1322). We also provide the illustrative application of the proposed approach to an analysis of demand for a recreation site. The result reveals that the usefulness is statistically supported by various criteria for model selection.

2. Semiparametric estimation of on-site models

In this section, we describe the semiparametric estimation of on-site count data models by use of Gurmu et al.'s (1999) Laguerre series expansion approach. Suppose that a random variable y_i is a count data with mean parameter θ_i and x_i is a vector of covariates with p linearly independent variables including a constant. Then by the standard exponential mean parameterization,

$$\theta_i = \exp(\mathrm{x}_i'\boldsymbol{\beta}), \quad i = 1, \cdots, N, \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients. The parameterization (1) ensures the non-negativity of θ_i and is also regarded as the generalized linear models with the log-link function. When y_i has the Poisson distribution, (1) is called the Poisson regression model. This is a basic parametric model but too restrictive for modeling count data because of equality of the conditional mean and variance, namely the equidispersion property of the Poisson distribution. As discussed by Cameron and Trivedi (1998, p.96), for instance, it is well recognized that count data is usually

overdispersed, which may be due to unobserved heterogeneity caused by misspecification of the Poisson parametric models.

To remove the restrictive property of the Poisson model, we introduce an unobserved heterogeneity variable to the mean parameterization (1) as follows:

$$\mu_i = \exp(\mathrm{x}_i'\boldsymbol{\beta} + \varepsilon_i) = \theta_i \nu_i, \quad i = 1, \cdots, N, \tag{2}$$

where $\nu_i = \exp(\varepsilon_i)$ and $E(\nu_i) = 1$. Suppose that the distribution of y_i is Poisson with the mean parameter μ_i and the unobserved heterogeneity ν_i is independent of the covariates x_i. Let $g(\nu_i)$ denote the probability density function of ν_i. The conditional mixture density of y_i given x_i can be written as

$$f(y_i \mid \mathrm{x}_i) = \int \frac{\exp(-\mu_i)\mu_i^{y_i}}{\Gamma(y_i+1)} \cdot g(\nu_i)\, d\nu_i = \int \frac{\exp(-\theta_i\nu_i)(\theta_i\nu_i)^{y_i}}{\Gamma(y_i+1)} \cdot g(\nu_i)\, d\nu_i,$$

which is called a mixed Poisson distribution (see, e.g., Cameron and Trivedi, 1998, pp.98-99). This expression depends on the specification of $g(\nu_i)$ but is a natural generalization of the Poisson regression models. For example, assuming that $g(\nu_i)$ is a gamma density, we obtain the negative binomial as a mixed Poisson distribution, and then (2) is the negative binomial model.

Next, considering that the data are collected by on-site sampling, there are problems of truncation at zero and oversampling or endogenous stratification as a special case of choice-based sampling. Therefore, we shall employ the correction of Shaw (1988, pp.213-215), who modified the probability density function to adjust these problems. An on-site mixed Poisson distribution is given by

$$\begin{aligned} f^s(y_i \mid \mathrm{x}_i) &= \frac{y_i}{E(y_i \mid \mathrm{x}_i)} \cdot f(y_i \mid x_i) = \frac{y_i}{\theta_i} \cdot \int \frac{\exp(-\theta_i\nu_i)(\theta_i\nu_i)^{y_i}}{\Gamma(y_i+1)} \cdot g(\nu_i)\, d\nu_i \\ &= \frac{\theta_i^{(y_i-1)}}{\Gamma(y_i)} \cdot \int \nu_i^{y_i} \exp(-\theta_i\nu_i) \cdot g(\nu_i)\, d\nu_i =: \frac{\theta_i^{(y_i-1)}}{\Gamma(y_i)} \cdot M_\nu^{(y)}(\theta_i), \end{aligned} \tag{3}$$

where the second equality results from $E(\mu_i \mid \theta_i) = \theta_i$ because $E(\nu_i) = 1$, and $M_\nu^{(y)}(\theta_i)$ is the y-th order derivative of the moment generating function $M_\nu(\theta_i)$ with respect to ν_i. Since in order to implement (3) it is necessary to explicitly approximate the unknown density $g(\nu_i)$, we apply the Laguerre series expansion approach to $M_\nu^{(y)}(\theta_i)$ proposed by Gurmu et al. (1999). The Laguerre polynomial approximation provides a flexible modeling of a mixed Poisson distribution, including the geometric and negative binomial models as well as the basic Poisson model. The approach is semiparametric in the sense that we do not assume that $g(\nu_i)$ has a specific parametric form but the degree of polynomial K employed in the infinite series expansion is allowed to increase with the sample size (see Gurmu and Trivedi, 1996, p.472).

Following the derivation given in Gurmu et al. (1999, pp.128-130), by use of a squared K-th degree Laguerre polynomial approximation to $g(\nu_i)$, we obtain

$$\begin{aligned} M_{\nu,K}^{*(y)}(\theta_i) = &\left(1 + \frac{\theta_i}{\lambda}\right)^{-\alpha} \left(\lambda + \theta_i\right)^{-y_i} \frac{\Gamma(\alpha)}{\phi_N} \sum_{j=0}^{K} \sum_{k=0}^{K} \sum_{l=0}^{j} \sum_{m=0}^{k} \eta_j \eta_k (h_j h_k)^{1/2} \\ &\times \binom{j}{l}\binom{k}{m} \frac{\Gamma(\alpha + l + m + y_i)}{\Gamma(\alpha + l)\Gamma(\alpha + m)} \left(-1 - \frac{\theta_i}{\lambda}\right)^{-(l+m)}, \end{aligned} \tag{4}$$

as the K-th degree approximation to $M_v^{(y)}(\theta_i)$, where $\alpha > 0$, $\eta_0 = 1$,

$$h_j := \frac{\Gamma(j+\alpha)}{\Gamma(\alpha)\Gamma(j+1)}, \qquad \phi_N := \sum_{j=0}^{K} \eta_j^2,$$

and λ is given by

$$\lambda = \frac{\Gamma(\alpha)}{\phi_N} \sum_{j=0}^{K} \sum_{k=0}^{K} \sum_{l=0}^{j} \sum_{m=0}^{k} \eta_j \eta_k (h_j h_k)^{1/2} \binom{j}{l} \binom{k}{m} (-1)^{-(l+m)} \frac{\Gamma(\alpha+l+m+1)}{\Gamma(\alpha+l)\Gamma(\alpha+m)}, \tag{5}$$

owing to $E(\nu_i) = 1$. Thus, given the degree of the polynomial K and replacing $M_v^{(y)}(\theta_i)$ with (4) in the on-site mixed Poisson distribution, the log-likelihood function is conducted as follows:

$$\log \mathcal{L}_N(\boldsymbol{\beta}, \alpha, \boldsymbol{\eta}) = \sum_{i=1}^{N} \left\{ (y_i - 1) \log \theta_i - \log \Gamma(y_i) + \log M_{v,K}^{*(y)}(\theta_i) \right\}, \tag{6}$$

where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_K)'$. We obtain the semiparametric estimators of the parameters $(\boldsymbol{\beta}, \alpha, \boldsymbol{\eta})$ by maximizing the log-likelihood function under the condition (5). The form of (6) comes to the same as that of Gurmu et al. (1999) except a little modification by allowing for on-site sampling, that is, to subtract 1 from all y_i observations of the first two terms in (6). It is similar to the basic Poisson model with on-site count data shown by Shaw (1988). On the other hand, the Laguerre polynomial approximation to $M_v^{(y)}(\theta_i)$ is not affected at all even if on-site sampling is employed. From these facts, we expect that the consistency of the above semiparametric estimators will follow from the result of Gurmu et al. (1999), though its asymptotic normality remains open as well as their original approach. The exponential mean parameterization (2) with the mixed Poisson distribution (3) is regarded as a semiparametric model for on-site count data. This modeling nests the Shaw's (1988) on-site Poisson model, if $\alpha^{-1} = \lambda^{-1} \to 0$ and $\eta_j = 0$ for $j \geq 1$, and the on-site negative binomial model considered and proposed by Englin and Shonkwiler (1995) and Martínez-Espiñeira and Amoako-Tuffour (2008), if $\alpha = \lambda$ and $\eta_j = 0$ for $j \geq 1$, as a special case, respectively. These nested structures make it possible to construct the likelihood ratio statistics for model selection in the following empirical analysis.

3. An empirical illustration

This section provides the empirical comparison of the proposed semiparametric model with the other existing parametric ones for on-site sampling by the analysis of real data. The data employed in this section is obtained from Nohara (2010) and originally collected for his empirical analysis of recreation benefits of trips to Hokkaido by using travel cost methods. Following the modeling of Nohara (2010), the single demand function for the recreation site can be written as

$$\theta_i = \exp\left(\beta_0 + \beta_1 \left(\frac{p_i}{I_i}\right) + \beta_2 \left(\frac{t_i}{T_i}\right) + \beta_3 Q_i\right), \tag{7}$$

where p_i and t_i are the travel cost to Hokkaido and its travel time, respectively, I_i is the income, T_i is the available time except working hours, and Q_i is the environmental quality of the site,

Table 1: Estimates and t-statistics for Poisson, NB2 and SP2 ($N = 446$).

	Poisson		NB2		SP2	
Variable	Estimate	t-statistic	Estimate	t-statistic	Estimate	t-statistic
const.	0.1675	0.684	-0.9365	-2.151*	-0.3927	-1.219
p/I	-5.2520	-3.339**	-4.8718	-2.572**	-4.1623	-2.289*
t/T	-4.3987	-0.022	13.963	0.055	140.41	0.557
Q	0.0012	3.441**	0.0012	2.305*	0.0004	0.779
α	–	–	1.936	4.220**	1.7721	5.723**
η_1	–	–	–	–	0.5491	15.69**
η_2	–	–	–	–	0.2110	31.46**

Note 1: t-statistics for the SP2 estimates are calculated from the estimates of the asymptotic standard errors.
Note 2: * and ** indicate that the estimate is significantly different from zero at the 5% and 1% level.

Table 2: Log-likelihood and CAIC for Poisson, NB2 and SP.

Model	Log-likelihood	CAIC
Poisson	-736.85	1495.0
NB2	-670.39	1369.2
SP1	-670.39	1376.3
SP2	-651.11	1344.8
SP3	-651.21	1352.1
SP4	-649.57	1357.9
SP5	-649.41	1362.7

with the sample size $N = 446$. For details see Nohara (2010). It is easily seen that the expression (7) is specified by the exponential mean parameterization (1). We estimate the parameters by use of three on-site count data models: the Poisson, the negative binomial 2 (NB2) and the semiparametric (SP) approaches. As suggested by Gurmu and Trivedi (1996) and Gurmu et al. (1999), to select the degree of the Laguerre polynomial and also compare the performance of various models, we employ the consistent Akaike information criterion (CAIC) proposed by Bozdogan (1987). The criterion is defined as

$$CAIC = -2 \log \mathcal{L}_N(\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\boldsymbol{\eta}}) + q\,(\log(N) + 1),$$

where $\log \mathcal{L}_N(\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\boldsymbol{\eta}})$ is the estimated log-likelihood and q is the number of free parameters. CAIC is consistent, that is, correct selection of K, since the penalty term is monotonically increasing function of the sample size N. In the following, all of the computations was carried out in Ox (see Doornik, 2006).

Table 3: Likelihood ratio statistics for model selection.

	H_1	
H_0	NB2	SP2
Poisson	132.92**	171.49**
NB2	–	38.57**

Note: ** represents the rejection of H_0 at the 1% significance level.

Table 1 reports the estimates and the t-statistics for testing the null hypothesis that the coefficient is zero. We select the polynomial degree $K = 2$ for the SP model among $K = \{1, \cdots, 5\}$ based on CAIC, and their results are shown in Table 2 below with the estimated log-likelihood. It should be noted that we employ the estimates of the asymptotic standard errors based on the empirical Hessian of the log-likelihood function (6) to construct the t-statistics for the SP model. Gurmu et al. (1999) computed the nonparametric bootstrapped standard errors for their t-statistics, though in the on-site SP model, those of the Laguerre polynomial coefficients seem to be unstable and unreliable from a practical point of view. As discussed in Cameron and Trivedi (1998, p.361) or Gurmu et al. (1999, p.147), if the selected order K is treated as correct, the statistics based on the empirical Hessian or the outer product of the gradient estimator can be valid. Thus, we recommend the t-statistics based on the asymptotic standard errors in the on-site sampling context. Since the SP with $K = 2$ (SP2) model nests the Poisson and the NB2 models as referred in the previous section, the likelihood ratio (LR) statistics for testing H_0: the restricted model against H_1: the unrestricted model are conducted as a benchmark of model evaluation and reported in Table 3.

From the results of Table 2, we find that the estimated log-likelihood of the NB2 and SP models are clearly larger and their CAIC are smaller than those of the Poisson model, respectively. These values reflect the fact that there occurs overdispersion in the present data, as shown by each of the estimates of α that are significantly different from zero at the 5% level in Table 1. The LR statistics for H_0: Poisson against H_1: NB2 or SP2 in Table 3 reject the null at the 1% significance level, so that the existence of overdispersion is also supported. The NB2 model is inadequate for the specification of overdispersion, because the Laguerre polynomial coefficients η_1 and η_2 are significant at the 1% in Table 1, and moreover the SP2 model is said to be superior to it in terms of goodness-of-fit measures by CAIC and the LR statistics in Tables 2 and 3. Comparing the Poisson or NB2 with the SP2 models in Table 1, the existence or misspecification of overdispersion substantially affect the estimates and would lead some of the t-statistics over-inflate. These inflations make the estimated coefficient of Q significant at the 1% and 5% in the Poisson and NB2 models but not in the SP2 model. Since this obviously causes wrong statistical inference for on-site count data, it seems to be preferable to use the semiparametric modeling in practice. The proposed semiparametric approach is flexible and adaptable for a form of overdispersion as long as the mean parameterization is correctly specified by (2).

4. Conclusions

This article has proposed the semiparametric approach to on-site count data models by use of the Laguerre series expansion method of Gurmu et al. (1999). The proposed estimation

is flexible and adaptable for an unknown form of overdispersion. The existing parametric estimation of on-site count data such as the Poisson and negative binomial models are included as a special case. In the empirical illustration, we have shown that the proposed semiparametric model is statistically preferable to the above existing models in terms of CAIC and the LR statistics.

References

Bozdogan, H. (1987) "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions" *Psychometrika* **52**, 345-370.

Cameron, A.C. and P. Johansson (1997) "Count data regression using series expansions: With applications" *Journal of Applied Econometrics* **12**, 203-223.

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

Doornik, J.A. (2006) *Ox: An Object-Oriented Matrix Language* (5th eds.), Timberlake Consultants Press: London.

Englin, J. and J.S. Shonkwiler (1995) "Estimating social welfare using count regression models: an application under conditions of endogenous stratification and truncation" *Review of Economics and Statistics* **77**, 104-112.

Gurmu, S., P. Rilstone, and S. Stern (1999) "Semiparametric estimation of count regression models" *Journal of Econometrics* **88**, 123-150.

Gurmu, S. and P.K. Trivedi (1996) "Excess zeros in count models for recreational trips" *Journal of Business and Economic Statistics* **14**, 469-477.

Loomis, J. (2003) "Travel cost demand model based river recreation benefit estimates with on-site and household surveys: Comparative results and a correction procedure" *Water Resources Research* **39**, 1105.

Martínez-Espiñeira, R. and J. Amoako-Tuffour (2008) "Recreation demand analysis under truncation, overdispersion, and endogenous stratification: An application to Gros Morne National Park" *Journal of Environmental Management* **88**, 1320-1332.

Nohara, K. (2010) "An empirical study of recreation benefit using a truncated count data model (in Japanese)" Tohoku Economics Research Group, Discussion Paper No. 260, pp.1-20.

Santos Silva, J.M.C. (1997) "Unobservables in count data models for on-site samples" *Economics Letters* **54**, 217-220.

Shaw, D. (1988) "On-site sample's regression problems of non-negative integers, truncation, and endogenous stratification" *Journal of Econometrics* **37**, 211-223.