Journal of Development and Agricultural Economics Vol. 2(1), pp. 293-303, August 2010
Available online at http://www.academicjournals.org/jdae
ISSN 2006- 9774 ©2010 Academic Journals

Full Length Research Paper

Truncation and endogenous stratification in various count data models for recreation demand analysis

Tomoaki Nakatani[1]* and Kazuo Sato[2]

[1]Department of Agricultural Economics, Hokkaido University, Kita9 Nishi9, Sapporo-shi, 060-8589, Hokkaido, Japan.
[2]Department of Dairy Science, Rakuno Gakuen University, Midorimachi 582, Bunkyodai, Ebetsu-shi, 069-8501, Hokkaido, Japan.

Accepted 29 July, 2010

The aim of this paper is to extend the truncated and endogenously stratified Poisson and negative binomial models to three alternative discrete distributions, namely the generalized Poisson, geometric and Borel distributions. Our primary intention here is to demonstrate how the improper treatments of the data generate divergent outcomes by applying those results to the recreation trip data surveyed from the visitors to an indigenous horse park in Japan. Our empirical application shows that failure to account for overdispersion, truncation and endogenous stratification leads to substantial changes in parameter estimates and their standard errors. The parameter on the travel cost tends to be underestimated in absolute value in the standard setups. This induces serious overestimation of the economic benefit that the recreation site offers to the society. Even when the endogenous stratification is incorporated, ignoring the overdispersion estimates the per capita per trip consumer surplus over 7 times larger than the one obtained under the endogenous stratification and overdispersion.

Key words: Count data models, endogenous stratification, overdispersion, recreation demand analysis, consumer surplus.

INTRODUCTION

The aim of this paper is to extend the truncated and endogenously stratified Poisson and negative binomial regression models to various alternative count data models. Truncated count data models have long been applied in many research fields.

In agricultural economics, the recreation demand analysis, or the travel cost method (TCM) is a major area of applications where the number of trips of individuals or households to a particular site, which usually takes non-negative integers, is described in count data models. The data sets for recreation demand analyses are usually built up through survey sampling since there rarely exist ready-made statistics that represent consumers' behavior towards recreation sites as well as their socio-economic variables. It is often the case that a recreation demand analysis is aimed at evaluating in monetary term the social value that the focused recreation site implicitly offers to the public, so that samples are supposed to be collected over a reasonably wide range of population.

Due to the difficulty in implementing such a large scale survey, an on-site sampling is preferred in many empirical studies. An on-site survey, however, involves two possibly serious disadvantages unless correctly treated. One is that the observed counts are truncated at zero[1]. This occurs because the survey obviously excludes the possibility for researchers to gather information from non-visitors.

As a result, researchers do not take a sample from the population but from a subset of the population. In this situation, the statistical inference has to be carried out

*Corresponding author. E-mail: naktom2@gmail.com. Tel / Fax: +81-11-706-3879.

[1] Although there are many forms of truncation, the discussions in this paper are limited to the truncation at zero, or simply referred to as "truncation" hereafter. General treatments for the various forms of truncation can be found, for instance, in Johnson et al. (1992).

based on a truncated distribution. The other is a problem of an endogenously stratified sampling. An on-site survey is likely to contain individuals who come to the site more frequently than those who come less. Shaw (1988) first pointed out this circumstance in the context of recreation demand analysis and called it "endogenous stratification" (or endogenous sampling). Endogenous stratification "may cause the sample probability of observations to differ from the corresponding population probabilities" (Cameron and Trivedi, 1998: 326). Therefore, in addition to truncation, endogenous stratification should also be taken into account in making an inference on the population demand function if on-site surveyed data are used[2].

Since the economic theory does not suggest a particular distribution[3] for recreation trips, it is an empirical question which distribution to be chosen from candidate distributions. There exists another motivation for considering alternative distributions in that the equidispersion, that is, the mean-variance equality in the Poisson distribution, is often turned out to be unrealistic. Although, there are many discrete distributions that may be applicable to recreation trip data, the Poisson and negative binomial distributions have been the common choices in the past literature when truncation is considered (Creel and Loomis, 1990; Gurmu, 1991; Grogger and Carson, 1991). For further references, see Cameron and Trivedi (1998) and Winkelmann (2003) for count data models in general.

An exception is Santos (1997a) who employs the generalized Poisson distribution of Consul and Jain (1973) and Consul (1989) to the truncated fishing trip data in Alaska. Another example is Sarker and Surry (2004) who propose six alternative truncated count data models, namely, the geometric, logarithmic, Borel, Yule, generalized negative binomial, and generalized Poisson distributions to estimate the demand for moose hunting in Ontario.

In the endogenously stratified setting, Shaw (1988) develops the theoretical motivations and applies them to the Poisson distribution. Santos (1997b) argues from a different perspective the same idea as Shaw (1988) in relation to unobserved heterogeneity. Englin and Shonkwiler (1995) later extend the results of Shaw (1988) to the negative binomial distribution for describing the hiking behavior in Washington. However, empirical applications are quite limited. To the best of our knowledge, no author seems to have attempted to extend it from the negative binomial to alternative distributions.

In what follows, the next section outlines the theoretical backgrounds for the truncation and endogenous stratification in a count data distribution, and applies them to miscellaneous distributions. Then the empirical application is carried out and the outcomes of ignoring truncation and/or endogenous stratification are discussed. Finally some concluding remarks and directions for future research are stated.

TRUNCATION AND ENDOGENOUS STRATIFICATION

Baseline distribution

Basic assumptions for the population distribution we assume are that the number of trips (Y) taken by person i, $i = 1, \ldots, n$, in a society (or in the population in the statistical sense) follows a certain probability distribution defined on non-negative integers $\mathbb{Z}^+$. Also assumed are the independence and identical distribution over the population. Let y_i denote the person i's realization of Y. The socio-economic attributes of the person i are represented as a $(k \times 1)$ vector $\mathbf{x}_i$ that may include a constant and the site-specific characteristics. The travel cost to the site is incorporated in $\mathbf{x}_i$. As usual, the conditional mean $\mathsf{E}[Y|\mathbf{X}]$ is defined as a nonlinear function of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ and $\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of parameters to be estimated. Let $f(y_i|\mathbf{x}_i, \boldsymbol{\beta})$ and $F(Y \leq y_i|\mathbf{x}_i, \boldsymbol{\beta})$, $y_i \in \mathbb{Z}^+$, denote the probability density function (pdf) and the cumulative distribution function (cdf) conditional on $\mathbf{x}_i$ and $\boldsymbol{\beta}$ of the underlying population distribution, respectively. In what follows, we suppress the condionality on $\mathbf{x}_i$ or $\mathbf{X}$ and $\boldsymbol{\beta}$ unless otherwise required.

Suppose that we have a sample of size n collected from the population. Then we do not need to adjust the distribution. The log-likelihood function in this case is constructed such that:

$$\ell(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \ln f(y_i) \tag{1}$$

Where $\mathbf{y}$ is an $(n \times 1)$ vector of observed number of trips. Assuming that the regularity conditions (Cameron and Trivedi, 1998, pp. 23-24, for example) are satisfied and that the conditional mean $\mathsf{E}[Y|\mathbf{X}]$ is correctly specified, the maximum likelihood estimator of $\boldsymbol{\beta}$ is efficient, consistent and asymptotically normal. An on-site sampling, however, is not a sampling from the population. As we will see later, the conditional mean is no longer given by $\mathsf{E}[Y|\mathbf{X}]$. Therefore, the inference based on (1) in general does not produce consistent results.

Truncated distribution

If the survey was conducted on the site, we were not able to have an observation with zero visit. This means that we do not sample from the population but from a subset of the population. In this environment, the conditional

[2] For simplicity, we refer to the truncation and endogenous stratification as the "endogenous stratification" hereafter. Note that when we mention "the truncation and endogenous stratification", it means, to be precise, that "the truncation" and "the truncation and endogenous stratification".

[3] In this paper, we use the terms "distribution" and "model" interchangeably.

mean is no longer $\mathsf{E}[Y|\mathbf{X}]$, so that the estimator based on (1) are not consistent. Given that the sampling is random, an appropriate procedure has to be established on the truncated distribution. When an underlying distribution is truncated, the resulting pdf is obtained by:

$$f(y|Y>0) = f_{Tr}(y) = \frac{f(y)}{\mathsf{Pr}(Y>0)} = \frac{f(y)}{1-F(0)}. \tag{2}$$

It follows from the basic probability theory that the mean and variance of the truncated distribution are linked to those of the underlying distribution in the following manner. By definition of the expectation of a random variable, the mean of the truncated distribution is derived as:

$$\mathsf{E}[Y|Y>0] = \sum_{j=1}^{\infty} j f_{Tr}(j) = \frac{\mathsf{E}[Y]}{1-F(0)}. \tag{3}$$

Obviously (3) differs from $\mathsf{E}[Y|\mathbf{X}]$. By the similar manipulation, we have

$$\mathsf{E}[Y^2|Y>0] = \sum_{j=1}^{\infty} j^2 f_{Tr}(j) = \frac{\mathsf{E}[Y^2]}{1-F(0)}. \tag{4}$$

Hence the variance of the truncated distribution is connected to the first and second moments of the original distribution in such a way that

$$\begin{aligned}\mathsf{Var}[Y|Y>0] &= \mathsf{E}[Y^2|Y>0] - (\mathsf{E}[Y|Y>0])^2 \\ &= \frac{1}{1-F(0)}\left\{\mathsf{E}[Y^2] - \frac{1}{1-F(0)}(\mathsf{E}[Y])^2\right\}\end{aligned} \tag{5}$$

Now the log likelihood function for the truncated sample is equal to

$$\ell_{Tr}(\boldsymbol{\beta};\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n} \ln f_{Tr}(y_i). \tag{6}$$

Given that the regularity conditions hold and the conditional mean is correctly specified, the estimator based on (6) are efficient, consistent and asymptotically normal. Nevertheless, an on-site sampling invokes yet another point of interest, that is, an endogenously stratified sampling that is explained in the following subsection.

Endogenously stratified distribution

Although, the sample is selected randomly within same strata, the fact that the selection of strata is dependent on the number of trips makes the sample probability of observations differ from the corresponding population probabilities. As was first pointed out by Shaw (1988), an on-site survey tends to have, in addition to the truncation, a stratified sampling scheme because the frequent visitors are more likely to be included in the sample. Here, we only reproduce the results shown by Shaw (1988) and Santos (1997b). Shaw (1988) shows that the density function for the endogenously stratified sample is given by

$$h(y) = \frac{y f(y)}{\sum_{t=1}^{\infty} t f(t)} = \frac{y f(y)}{\mathsf{E}[Y]} \tag{7}$$

Where $h(y)$ denotes the sample density function. It is straightforward to obtain the mean and variance of $h(y)$ in terms of the lower moments of the population distribution. For the mean,

$$\begin{aligned}\mathsf{E}^*[Y] &= \sum_{j=1}^{\infty} j h(j) \\ &= \sum_{j=1}^{\infty} \frac{j^2 f(j)}{\mathsf{E}[Y]} \\ &= \mathsf{E}[Y] + \frac{\mathsf{Var}[Y]}{\mathsf{E}[Y]}\end{aligned} \tag{8}$$

that is shown by Santos (1997b), while for the variance we find that

$$\begin{aligned}\mathsf{Var}^*[Y] &= \mathsf{E}^*[Y^2] - \{\mathsf{E}^*[Y]\}^2 \\ &= \sum_{j=1}^{\infty} j^2 h(j) - \left\{\frac{\mathsf{E}[Y^2]}{\mathsf{E}[Y]}\right\}^2 \\ &= \frac{\mathsf{E}[Y^3]}{\mathsf{E}[Y]} - \left\{\frac{\mathsf{E}[Y^2]}{\mathsf{E}[Y]}\right\}^2\end{aligned} \tag{9}$$

The log-likelihood function for the endogenously stratified sample is thus given by

$$\ell_{ES}(\boldsymbol{\beta};\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n} \ln h(y_i) \tag{10}$$

Again, given that the regularity conditions are fulfilled and the model is correctly specified, the estimator based on (10) is efficient, consistent and asymptotically normal.

Application to various models

In the previous two subsections, we have derived from the population distribution the general expressions of the pdfs, means, variances and log likelihood functions for the truncated or endogenously stratified distributions. We now apply these general expressions to particular distributions. Besides the conventional distributions

Table 1. Pdfs of standard, truncated and endogenously stratified distributions and their parameterizations.

Distribution	$f(y)$	$f(y\|Y>0)$	$h(y)$	Parameterization	Remark
Poisson	$\frac{\lambda^y e^{-\lambda}}{y!}$	$\frac{\lambda^y}{y!(e^\lambda-1)}$	$\frac{\lambda^{y-1}e^{-\lambda}}{(y-1)!}$	$\lambda=\exp(\mathbf{x}'\boldsymbol{\beta})$	
NegBin II	$\frac{\Gamma(y+\frac{1}{\alpha})}{\Gamma(y+1)\Gamma(\frac{1}{\alpha})}\left(\frac{1}{1+\alpha\lambda}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\lambda}{1+\alpha\lambda}\right)^y$	$\frac{\Gamma(y+\frac{1}{\alpha})}{\Gamma(y+1)\Gamma(\frac{1}{\alpha})}\left(\frac{\alpha\lambda}{1+\alpha\lambda}\right)^y\frac{1}{(1+\alpha\lambda)^{\frac{1}{\alpha}}-1}$	$\frac{y\Gamma(y+\frac{1}{\alpha})}{\Gamma(y+1)\Gamma(\frac{1}{\alpha})}\frac{\alpha^y\lambda^{y-1}}{(1+\alpha\lambda)^{(y+\frac{1}{\alpha})}}$	$\lambda=\exp(\mathbf{x}'\boldsymbol{\beta})$	$\alpha>0$ Poisson when $\alpha\to 0$ Geometric when $\alpha=1$
Generalized Poisson	$\frac{A^y(1+By)^{y-1}}{y!e^{A(1+By)}}$	$\frac{A^y(1+By)^{y-1}}{y!(e^A-1)e^{ABy}}$	$\frac{(1-AB)[A(1+By)]^{y-1}}{(y-1)!e^{A(1+By)}}$	$A=\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\theta_0\exp[(1+\theta_1)\mathbf{x}'\boldsymbol{\beta}]}$ $B=\theta_0\exp(\theta_1\mathbf{x}'\boldsymbol{\beta})$	$\theta_0\geq 0$ Poisson when $\theta_0=0$ Borel when $\theta_0=1$ and $\theta_1=0$
Geometric	$\left(\frac{1}{1+\lambda}\right)\left(\frac{\lambda}{1+\lambda}\right)^y$	$\frac{\lambda^{y-1}}{(1+\lambda)^y}$	$\frac{y\lambda^{y-1}}{(1+\lambda)^{y+1}}$	$\lambda=\exp(\mathbf{x}'\boldsymbol{\beta})$	
Borel	$\frac{(y+1)^{y-1}\lambda^y e^{-\lambda(y+1)}}{y!}$	$\frac{(y+1)^{y-1}\lambda^y e^{-\lambda y}}{y!(e^\lambda-1)}$	$\frac{\{\lambda(y+1)\}^{y-1}(1-\lambda)e^{-\lambda(y+1)}}{(y-1)!}$	$\lambda=\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}$	

(Poisson and negative binomial), the ones considered in this paper are the generalized Poisson, geometric and Borel distributions[4]. The analytical expressions of the pdfs, means and variances for the candidate distributions in the standard (Std), truncated (Tr) and endogenously stratified (ES) cases are derived and summarized in Tables 1 - 3, respectively. Table 1 also contains parameterizations, due to Sarker and Surry (2004), that bring $\mathrm{E}[Y|\mathbf{x}]=\exp(\mathbf{x}'\boldsymbol{\beta})$, and remarks that explain parameter restrictions and the relationship among distributions.

The assumption of the equidispersion, that is the mean is equal to the variance, is common criticism for the standard Poisson distribution. Indeed, many previous applications find the overdispersion, that is, the variance being greater than the mean, in real data and conclude that the equidispersion is an unrealistic assumption. Integrating the overdispersion into a model constitutes a main motivation for the use of alternative count data models. For this purpose, Cameron and Trivedi (1986) propose a version of the negative binomial distribution with an additional parameter α to control the degree of the overdispersion in the population, and name it the negative binomial II or NegBin II. The geometric distribution is nested in the NegBin II with the restriction of $\alpha=1$. The positive α indicates the overdispersion while the Poisson distribution results as a limit when $\alpha\to 0$.

Consul and Jain (1973) generalize the Poisson distribution to incorporate the overdispersion in a different manner. The generalized Poisson distribution is attractive because it includes both the Poisson and Borel distributions as special cases with explicit parameter constraints. It is reduced to the Poisson distribution with $\theta_0=1$ and to the Borel distribution with $\theta_0=1$ and $\theta_1=0$, respectively. The degree of the overdispersion in the population is controlled by $(1-AB)^{-2}$ in our notation found in Table 1. The monograph by Consul (1989) provides the intensive summary of the properties of the generalized Poisson distribution. The log-likelihood functions (1), (6) and (10) are formulated according to the pdfs appeared in table 1. As mentioned earlier, if the regularity conditions are satisfied and if the moments of distributions are correctly specified, the ML estimation attains consistency, efficiency and asymptotic normality of the estimators. While $f(y)$ and $f(y|Y>0)$ in Table 1 are confirmed to satisfy the regularity conditions see (Cameron and Trivedi, 1998, for the Poisson and negative binomial, Santos Silva, 1997a, for the generalized Poisson, and Sarker and Surry, 2004, for the geometric and Borel distributions, respectively), it is yet to be proved whether $h(y)$ also satisfies them. For the current application, we assume that the regularity conditions hold in $h(y)$ appeared in, Table 1 and that the conditional means are correctly specified.

Empirical illustration

Data

The data set was collected by on-site interview in November 2003 for the visitors to the Noma highland indigenous horse park in Ehime prefecture, Japan. The

[4] Sarker and Surry (2004) also investigated the Logarithmic and Yule distributions. Nonetheless, we exclude these distributions because the consumers' surplus is not easily evaluated in the Logarithmic distribution, and because $\mathrm{E}[Y]\in(0,1)$ is necessary in the Yule distribution for the existence of the mean. The former may be an obstacle to a practical application, while the latter is violated in our data set.

Table 2. Means of standard, truncated and endogenously stratified distributions.

Distribution	$E[Y]$	$E[Y\|Y>0]$	$E^*[Y]$
Poisson	λ	$\frac{\lambda e^{\lambda}}{e^{\lambda}-1}$	$\lambda+1$
NegBin II	λ	$\frac{\lambda(1+\alpha\lambda)^{\frac{1}{\alpha}}}{(1+\alpha\lambda)^{\frac{1}{\alpha}}-1}$	$1+(1+\alpha)\lambda$
Generalized Poisson	$\frac{A}{1-AB}$	$\frac{Ae^{A}}{(e^{A}-1)(1-AB)}$	$\frac{A(1-AB)+1}{(1-AB)^2}$
Geometric	λ	$1+\lambda$	$1+2\lambda$
Borel	$\frac{\lambda}{1-\lambda}$	$\frac{\lambda e^{\lambda}}{(e^{\lambda}-1)(1-\lambda)}$	$\frac{\lambda(1-\lambda)+1}{(1-\lambda)^2}$

Table 3. Variances of standard truncated and endogenously stratified distributions.

Distribution	$Var[Y]$	$Var[Y\|Y>0]$	$Var^*[Y]$
Poisson	λ	$\frac{\lambda e^{\lambda}}{e^{\lambda}-1}\left\{1-\frac{\lambda}{e^{\lambda}-1}\right\}$	λ
NegBin II	$\lambda(1+\alpha\lambda)$	$\frac{\lambda(1+\alpha\lambda)^{\frac{1}{\alpha}}}{(1+\alpha\lambda)^{\frac{1}{\alpha}}-1}\left\{(1+\alpha\lambda)-\frac{\lambda}{(1+\alpha\lambda)^{\frac{1}{\alpha}}-1}\right\}$	$\lambda+\alpha\lambda(1+\lambda+\alpha\lambda)$
Generalized Poisson	$\frac{A}{(1-AB)^3}$	$\frac{Ae^{A}}{(e^{A}-1)(1-AB)^2}\left\{\frac{A(1-AB)+1}{1-AB}-\frac{Ae^{A}}{e^{A}-1}\right\}$	$\frac{A\{1-(A+2)B\}}{(1-AB)^4}$
Geometric	$\lambda(1+\lambda)$	$\lambda(1+\lambda)$	$2\lambda(1+\lambda)$
Borel	$\frac{\lambda}{(1-\lambda)^3}$	$\frac{\lambda e^{\lambda}}{(e^{\lambda}-1)(1-\lambda)^2}\left\{\frac{1}{1-\lambda}-\frac{\lambda}{(e^{\lambda}-1)}\right\}$	$\frac{\lambda(3-\lambda)}{(1-\lambda)^4}$

Table 4. Description of variables.

Name	Description	Mean	S.E.
Trips	The number of trips to the site for the past three year period	9.065	11.215
TC	The travel cost to the site in 10^3 JPY. Constructed by monetary expenses plus 1/3 of the minimum wage rate.	0.913	1.430
Kids	A binary dummy equal to 1 if accompanying a child/children	0.801	0.400
Male	A binary dummy equal to 1 if a respondent is male	0.392	0.490
Age	The natural logarithm of the age divided by 10 of a respondent	1.148	0.301
Horse	A binary dummy equal to 1 if a respondent is in favor having contact with horses on the site	0.720	0.450

questionnaire includes the travel cost and time, the frequency of visits to the site for the past three-year period, and the age, gender and other attributes of respondents. The travel cost consists of actual expenditures from home to the site that are surveyed through questionnaire, plus the time or opportunity cost of travel. The opportunity cost is estimated as the 1/3 of the minimum wage rate[5]. A total of 409 respondents participated in the survey. After eliminating incomplete responses, the sample consists of 186 observations[6].

The data set is originally analyzed in Sato and Kasubuchi (2005)

[5] Cesario (1976) advocates using the 1/3 of the actual wage rate of a respondent. However, difficulty in enquiring into private matters prevents us from including such a question in the questionnaire. Thus, as a moderate figure, the 1/3 of the minimum wage rate is used.

[6] We also drop samples with the frequency of trips greater than or equal to 60. Our experience in the numerical optimizations showed that the log-likelihood functions became numerically unstable if those samples were incorporated. It is rather arbitrary, though Englin and Shonkwiler (1995) also eliminate samples with more than 12 trips.

Table 5. Estimation results of the various recreation demand models.

	Poisson			NegBin II			Generalized Poisson		
Estimate	**Std**	**Tr**	**ES**	**Std**	**Tr**	**ES**	**Std**	**Tr**	**ES**
Const.	1.619 (0.135)	1.632 (0.137)	1.442 (0.145)	1.245 (0.395)	0.499 (0.607)	-3.387 (0.884)	1.675 (0.380)	1.751 (0.294)	-0.912 (1.817)
$\widehat{\beta}_{TC}$	-0.527 (0.045	-0.580 (0.048)	-0.663 (0.051)	-0.193 (0.057)	-0.228 (0.073)	-0.279 (0.086)	-0.335 (0.103)	-1.950 (0.002)	-4.696 (1.314)
$\widehat{\beta}_{Kids}$	0.238 (0.067)	0.241 (0.068)	0.272 (0.072)	0.222 (0.192)	0.271 (0.276)	0.252 (0.202)	0.165 (0.182)	0.438 (0.236)	1.477 (1.121)
$\widehat{\beta}_{Male}$	-0.185 (0.053)	-0.181 (0.054)	-0.204 (0.057)	-0.284 (0.157)	-0.371 (0.227)	-0.320 (0.166)	-0.173 (0.153)	0.036 (0.139)	0.129 (0.393)
$\widehat{\beta}_{Age}$	0.484 (0.086)	0.493 (0.087)	0.552 (0.091)	0.599 (0.277)	0.828 (0.416)	0.739 (0.297)	0.442 (0.238)	0.636 (0.002)	1.357 (0.742)
$\widehat{\beta}_{Horse}$	0.360 (0.061)	0.360 (0.061)	0.406 (0.066)	0.427 (0.168)	0.559 (0.242)	0.503 (0.177)	0.257 (0.166)	0.197 (0.190)	0.604 (0.588)
$\widehat{\alpha}$				0.900 (0.099)	2.397 (0.700)	74.695 (58.815)			
$\widehat{\theta}_0$							1.336 (0.581)	3.406 (0.002)	1.703 (0.140)
$\widehat{\theta}_1$							-0.748 (0.194)	-1.107 (0.002)	-0.974 (0.014)
psd-R^2	0.787	0.805	0.844	0.119	0.090	0.153	0.117	0.260	0.191
-logL	1153.7	1145.2	1232.9	594.1	561.9	570.2	583.5	542.4	552.5
AIC	2319.4	2302.3	2477.9	1202.3	1137.8	1154.4	1182.9	1100.8	1120.9
BIC	2338.7	2321.7	2497.2	1224.9	1160.3	1176.9	1208.7	1126.6	1146.7
CAIC	2344.7	2327.7	2503.2	1231.9	1167.3	1183.9	1216.7	1134.6	1154.7

Notes: “Std”, “Tr” and “ES” stand for the standard, truncated and endogenously stratified models, respectively. The numbers in parenthesis are ML standard errors. The psd-R^2 is the pseudo-R^2 measure suggested by Maddala (1983, p.39). –logL denotes the negative of the log-likelihood at the estimates. AIC, BIC and CAIC denote the Akaike, Bayesian and consistent Akaike information criteria, respectively.

who consider only the Std and Tr NegBin II models. We follow Sato and Kasubuchi (2005) regarding the choice of variables. The definition, sample mean and standard deviation of the variables used in the empirical analysis are summarized in Table 4. The mean of the number of trips is about 9.1 that is relatively large compared to the figures found in the previous researches. Its standard deviation takes about 11.2. The variance-mean ratio in the sample is more than 13. The analytical expressions for the mean and variance in Tables 2 and 3 show that the Poisson distribution is unable to capture this overdispersion in the sample both in the truncated and endogenously stratified models, whereas the other four distributions may capture it depending on the value of λ. This fact suggests that even if the truncation or endogenous stratification is considered, the Poisson distribution is not adequate when a sample from an on-site survey exhibits the overdispersion.

RESULTS

Numerical optimizations are carried out by the BFGS algorithm in the “optim” function of the statistical software

Table 5. Contd.

	Geometric			Borel		
	Std	**Tr**	**ES**	**Std**	**Tr**	**ES**
Const.	1.240 (0.414)	0.939(0.423)	0.322(0.318)	0.935(1.235)	0.118(0.818)	-0.323(0.332)
$\widehat{\beta}_{TC}$	-0.188(0.056)	-0.277(0.085)	-0.408(0.086)	-0.122(0.082)	-0.165(0.069)	-0.256(0.078)
$\widehat{\beta}_{Kids}$	0.222(0.201)	0.252(0.203)	0.249(0.153)	0.177(0.599)	0.202(0.390)	0.185(0.158)
$\widehat{\beta}_{Male}$	-0.286(0.164)	-0.321(0.166)	-0.285(0.126)	-0.321(0.494)	-0.372(0.320)	-0.230(0.129)
$\widehat{\beta}_{Age}$	0.599(0.291)	0.739(0.299)	0.748(0.223)	0.832(0.979)	1.008(0.652)	0.573(0.234)
$\widehat{\beta}_{Horse}$	0.428(0.176)	0.503(0.178)	0.493(0.135)	0.483(0.503)	0.567(0.327)	0.378(0.139)
$\widehat{\alpha}$						
$\widehat{\theta}_0$						
$\widehat{\theta}_1$						
psd-R^2	0.114	0.155	0.279	0.013	0.043	0.153
-logL	594.6	569.7	612.3	676.9	572.9	567.0
AIC	1201.2	1151.5	1236.6	1365.8	1157.9	1145.9
BIC	1220.6	1170.8	1255.9	1385.2	1177.2	1165.3
CAIC	1226.6	1176.8	1261.9	1391.2	1183.2	1171.3

Notes: “Std”, “Tr” and “ES” stand for the standard, truncated and endogenously stratified models, respectively. The numbers in parenthesis are ML standard errors. The psd-R2 is the pseudo-R2 measure suggested by Maddala (1983, p.39). –logL denotes the negative of the log-likelihood at the estimates. AIC, BIC and CAIC denote the Akaike, Bayesian and consistent Akaike information criteria, respectively.

R[7]. The variance covariance matrices of the estimates in various models are calculated from the inverse of the negative of the numerical Hessian. To ensure the global maximum of the likelihood functions, we use ten randomly chosen initial values of parameters to see if the convergence is achieved around the same estimates. All the versions of the Poisson, geometric and Borel, and the Std and Tr NegBin II distributions achieve the convergence without difficulty. In the ES NegBin II, the estimate of α is unstable while other parameters and the maximum of the log-likelihood function remain almost identical regardless of the value of $\widehat{\alpha}$. The Tr and ES generalized Poisson models are highly sensitive to starting values[8]. For these two models, we increase the number of different initial values, limit the range of initial values and finally obtain the optimum.

Our primary intention here is to demonstrate how the results differ in various models under miscellaneous assumptions. Similar to the findings in Grogger and Carson (1991), we see that failure to account for overdispersion, truncation and endogenous stratification leads to substantial changes in parameter estimates and their standard errors. The estimation results of the five count data models under the three different distributional assumptions are tabulated in Table 5. The table contains the parameter estimates, standard errors, pseudo-R^2 of Maddala (1983, p. 39), negative values of the log likelihood function at the maximum, and three information criterion (AIC, BIC and the consistent AIC).

The parameter estimates are similar in their signs across

[7] **R** is downloadable from <www.r-project.org> at free of charge.

[8] This is due to the property of the “optim” function of **R** in which the function value of the first iteration must be finite. The infinite values are ignored if happened during the course of the optimization.

Table 6. Actual and predicted frequencies.

		Poisson			NegBin II			Gen. Poisson			Geometric			Borel		
Frequency	**Actual**	**Std**	**Tr**	**ES**	**Std**	**Tr**	**ES**	**Std**	**Tr**	**ES**	**Std**	**Tr**	**ES**	**Std**	**Tr**	**ES**
1	31	2	1	1	1	1	1	1	65	1	1	1	1	1	0	1
2	33	4	5	5	0	0	0	0	44	0	0	0	0	0	1	0
3	19	5	6	5	2	1	2	3	27	0	1	2	4	2	2	3
4	11	10	10	11	7	9	8	4	28	1	8	8	9	11	10	9
5	19	9	8	8	12	13	14	9	16	9	12	14	15	15	18	15
6	6	19	22	23	20	18	21	17	4	23	20	21	21	12	12	23
7	4	17	17	16	21	25	21	19	2	59	21	21	20	25	26	19
8	1	19	22	25	22	19	22	30	0	19	21	22	23	19	18	24
9	1	15	10	9	24	27	21	23	0	9	27	21	14	20	19	13
10	19	17	16	16	23	16	20	24	0	17	21	20	22	11	9	23
11-15	7	65	63	61	52	52	51	55	0	35	52	51	53	63	60	50
16-20	13	4	6	6	2	5	5	1	0	11	2	5	4	7	11	6
21-25	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
26-30	11	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
31+	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

models and setups. In particular, $\widehat{\beta}_{TC}$ is estimated all negative, which agrees with the demand theory. Inaddition, the standard errors are small enough to reject by the *t*-test the null of $\beta_{TC} = 0$ in almost all the models and setups. $\widehat{\beta}_{Kids}$ and $\widehat{\beta}_{Horse}$ have positive sign as expected. No information is available *a priori* for the other two. $\widehat{\beta}_{Kids}$ and $\widehat{\beta}_{Male}$ are not significant based on the *t*-statistics other than the Poisson models while $\widehat{\beta}_{Age}$ has the significant positive relation with the number of trips in most of models. The levels of significance for $\widehat{\beta}_{Horse}$ are mixed. $\widehat{\beta}_{Horse}$ is not significant in all versions of the generalized Poisson and the Std and Tr Borel distribution, whereas it is highly significant in other models. All distributional parameters ($\widehat{\alpha}$, $\widehat{\theta}_0$ and $\widehat{\theta}_1$) are highly significant except in the ES NegBin II, suggesting that the NegBin II and generalized Poisson models are preferred over the other three models.

It should be emphasized that $\widehat{\beta}_{TC}$ in absolute value is systematically underestimated in the standard models, followed by the truncated models, compared to the endogenously stratified models. This finding is of great importance in empirical researches. Ignoring the truncation and endogenous stratification, or how the sample under consideration is collected, would seriously mislead arguments regarding, for instance, how the marginal change in travel cost affects the trips taken even if the population distribution is correctly specified. Another serious consequence can be found when one wants to evaluate the monetary value of a recreational site. This point is investigated in more detail later.

Turning our attention to the overall performance of the models, it is not meaningful to directly compare the log likelihood values over the models or across the different setups because, except for some cases remarked in Table 1, models are not nested in one another. For those nested pairs, one can easily construct the likelihood ratio (LR) test statistics.

For instance, the LR test statistics of the ES geometric against the ES NegBin II, and of the ES Borel against the ES generalized Poisson are 84.2 and 29.0, respectively. Both have the χ^2 distribution with one and two degrees of freedom, with 5% critical values of 3.8 and 6.0, respectively, so that in both cases, the null of the ES geometric and ES Borel models are rejected.

Since no test is available for testing a particular model against all others, we rely on the information criteria to choose the one that describes the data set best[9]. The three criteria have similar variation, so that we focus on the values of AIC. Within each distribution, the endogenously stratified models do not necessarily have the smallest AIC. In fact, the truncated models attain the smallest except in the Borel distribution. However, because the theory encourages us to use it, we restrict our discussions to the endogenously stratified cases. Within the endogenously stratified settings, the generalized Poisson takes the minimum AIC, followed by the Borel, NegBin II, geometric and Poisson models. This result seems to suggest that the models with overdispersion are preferred to the Poisson model.

The goodness of fit of a model is measured by the pseudo-R^2. It varies greatly from 0.013 in the Std Borel model to 0.844 in the ES Poisson model. It is worth

[9] Santos Silva (2001) proposed a score test for non-nested alternatives. Testing all possible combinations by the score test may enable us to select the appropriate one.

Table 7. Estimated consumer surplus, their standard errors and the 95% confidence interval.

Distribution	Type	CS/Trip	S.E.	CI95L	CI95U
Poisson	Std	1.898	0.162	1.580	2.215
	Tr	1.724	0.142	1.446	2.002
	ES	1.508	0.116	1.281	1.736
NegBin II	Std	5.174	1.518	2.198	8.150
	Tr	4.393	1.411	1.628	7.158
	ES	3.587	1.105	1.420	5.754
Generalized Poisson	Std	2.982	0.912	1.195	4.769
	Tr	0.513	0.001	0.512	0.514
	ES	0.213	0.060	0.096	0.330
Geometric	Std	5.327	1.588	2.214	8.440
	Tr	3.615	1.116	1.427	5.802
	ES	2.449	0.516	1.439	3.460
Borel	Std	8.230	5.549	-2.645	19.106
	Tr	6.062	2.549	1.066	11.057
	ES	3.912	1.187	1.585	6.238

Notes: CS/Trip is expressed in 10^3 JPY. CI95L and CI95U denote the lower and upper bounds of the 95% confidence interval, respectively.

mentioning that, in spite of being constructed under unsuitable assumptions, the Poisson models show relatively high pseudo-R^2 values. Table 6 displays the actual and predicted numbers of trips from the various models. The predictions are made based on the mean equations listed in Table 2, and calculated by plugging the parameter estimates reported in Table 5 into the formulae.

The predicted numbers are rounded to the nearest integers. From this table, one can see that most of the models fail to predict the number of trips greater than 21 where we still have 22 actual observations. The only exception is the ES generalized Poisson model that predicts one in the range between 21 - 25, and the other in between 26 - 30. As long as the goodness of fit and the model predictability are concerned, the Poisson models perform well. If we further take into account the complexity of estimation in other models, the Std Poisson model may not be a bad choice at least for a prediction purpose. On the other hand, difference among the parameter estimates on the travel cost heavily affects the estimate of the economic benefit that a recreation site offers since the value is usually estimated though a function of the estimate of the coefficient on the travel cost. Therefore, for a purpose of an economic evaluation of a recreation site, ill-treatments for the overdispersion and the way how the data are collected would cause misleading results. The following subsection examines this aspect.

Welfare analysis

If we see $\mathrm{E}[Y|\mathbf{x}]$ as a "quantity demanded", the estimated models in the previous subsection can be regarded as the demand functions for the recreation site. Once we obtain a recreation demand function, the value that the recreation site provides is to be quantified through a welfare measure. The consumer surplus is typically used in TCM researches. Suppose that we have a population demand function for a particular site, and let the demand function be denoted by $D(TC, \mathbf{x}_{-TC})$ where TC is the price or the travel cost to the site, and $\mathbf{x}_{-TC}$ is the vector of the i th person's attributes without the element for the travel cost. Then, the per capita consumer surplus is defined, keeping other variables fixed, as the area under $D(TC, \mathbf{x}_{-TC})$ from TC_0 to ∞ where TC_0 is the current price. When the population mean is specified as $D(TC, \mathbf{x}_{-TC}) = \mathrm{E}[Y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$, the per capita consumer surplus is given by

$$\mathrm{CS} = \int_{TC_0}^{\infty} \exp(\mathbf{x}'\boldsymbol{\beta}) dTC = -\frac{1}{\beta_{TC}} \mathrm{E}[Y|\mathbf{x}] \tag{1}$$

Where β_{TC} is a coefficient on the travel cost. Consequently, the per capita per trip consumer surplus is calculated by

$$\frac{\mathrm{CS}}{\mathrm{E}[Y|\mathbf{x}]} = -\frac{1}{\beta_{TC}} \tag{2}$$

that is the negative of the reciprocal of β_{TC}. Since the models we consider in this paper all have the population means of the form $\mathrm{E}[Y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$, it is straightforward to construct the per capita per trip consumer surpluses, their standard errors and the confidence intervals from the coefficient estimates and variance covariance matrices found in the previous subsection.

Table 7 contains the estimates of the per capita per trip consumer surplus (denoted as CS/Trip), their standard errors, and 95% confidence intervals. CS/Trip is expressed in 10^3 JPY (1 USD $\approx$ 110 JPY at the time of survey). The standard errors are constructed from the ML standard errors of the estimates by the Delta method. The last two columns in Table 7 correspond to the lower and upper bounds of the 95% confidence interval, respectively. The estimates vary across different distributions and different setups. For the Std Borel model, the lower bound of the 95% confidence interval takes a negative value because the corresponding standard error of $\widehat{\beta}_{TC}$ is not small enough to reject the null of $\beta_{TC} = 0$.

One can clearly observe from Table 7 that ignorance of truncation or how the data are collected results in misleading outcomes. The fact that $|\widehat{\beta}_{TC}|$ is systematically underestimated is reflected in the estimates of the consumer surplus. Within each distribution, the standard case gives the highest estimate, followed by the truncated and endogenously stratified ones. The ratios of the estimate in the standard setup to the endogenously stratified one are approximately 14.0 in the generalized Poisson, about 2.1 in both the geometric and Borel, 1.4 in the NegBin II, and 1.3 in the Poisson models. If one uses the Std Poisson model in which the overdispersion, truncation and endogenous stratification are all disregarded, the resulting per capita per trip consumer surplus is nearly 8.9 times larger than the one derived in the ES generalized Poisson model. If one instead utilizes the ES Poisson, the discrepancy between them is still large around 7.1.

CONCLUDING REMARKS

In this article, we extend the truncated and endogenously stratified count data models, originally proposed by Shaw (1988), to three alternative discrete distributions, namely the generalized Poisson, geometric and Borel distributions. Analytical expressions of these models for the mean, variance and density function under the endogenous stratification are derived. Then as an illustration of how the improper treatments of the data generate divergent outcomes, we apply our theoretical results to the trip data that were collected on the recreation site.

Our major findings can be summarized as follows. Among models considered in this paper the ES generalized Poisson performs well in terms of the information criterion while the ES NegBin II, which is the conventional alternative to the Poisson, does not fit well to the data set. The parameter on the travel cost tends to be underestimated in absolute value in the standard setups. This in turn induces the serious overestimation of the economic benefit that the recreation site offers to the society. In the extreme case where we miss both the overdispersion and endogenous stratification, the per capita per trip consumer surplus in the standard setup is almost nine times larger than the one found in the truncation and endogenous stratification. Even when the endogenous stratification is incorporated, ignoring the overdispersion produces the per capita per trip consumer surplus over seven times larger.

Finally we state some directions and remaining tasks for further research. The data set we investigate here is a particular example of recreation behavior. There are miscellaneous types of recreations ranging for example from a day trip to a rural area, a camping in forests and an overnight stay in farmhouse to fishing, trekking and mountaineering. Accumulating empirical applications in different sorts of activities will help us to find a better model for the economic evaluation of recreation sites. The effect of the inclusion/exclusion of substitution sites, or more generally, the system of recreation demand will be necessary to investigate.

The economic theory does not specify a particular statistical distribution for recreation trip behavior. In addition, there is a long list of discrete distributions. Accordingly extension to miscellaneous discrete distributions is another important direction. Constructing statistical tests is yet another area for the future research. Given that the assumptions we made for the various models hold, it is uncomplicated to derive Lagrange multiplier (LM) or score types of statistical tests. Building the LM test of the ES Poisson against the ES generalized Poisson would be an interesting example. Or building a non-nested LM test of the ES NegBin II against the ES generalized Poisson would be attractive from a practical point of view. However, it seems a demanding work to establish the asymptotic theories for the alternative models under the endogenous stratification, which are required to construct the above-mentioned statistical tests.

REFERENCES

Cameron AC, Trivedi PK (1986). Econometric models based on count data: comparisons and applications of some estimators and tests. J. Appl. Econom., 1: 29-53.

Cameron AC, Trivedi PK (1998). Regression analysis of count data. Cambridge University Press, New York.

Cesario FJ (1976). Value of time in recreation benefit studies. Land Econ., 52: 32-41.

Consul P (1989). Generalized Poisson distributions. Marcel Dekker, New York.

Consul P, Jain G (1973). A generalization of the Poisson distribution. Technomet., 15: 791-799.

Creel MD, Loomis JB (1990). Theoretical and empirical advantages of truncated count data estimates for analysis of deer hunting in California. Am. J. Agr. Econ., 72: 434-443.

Englin J, Shonkwiler J (1995) Estimating social welfare using count data models: an application to long-run recreation demand under conditions of endogenous stratification and truncation. Rev. Econ. Stat., 77: 104-112.
Grogger J, Carson R (1991). Models for truncated counts. J. Appl. Econ., 6: 225-238.
Gurmu S (1991). Tests for detecting overdispersion in the positive Poisson regression model. J. Bus. Econ. Stat., 9: 215-222.
Johnson N, Kotz S, Kemp A (1992). Univariate discrete distributions 2nd edn. Wiley, New York.
Maddala G (1983). Limited-dependent and qualitative variables in econometrics. Cambridge University Press, New York.
Santos Silva J (1997a). Generalized Poisson regression for positive count data. Comm. Stat. -- Simulation Comput., 26: 1089-1102.
Santos Silva J (1997b). Unobservables in count data models for on-site samples. Econ. Lett., 54: 217-220.
Santos Silva J (2001). A score test for non-nested hypotheses with applications to discrete data models. J. Appl. Econom., 16: 577-597.
Sarker R, Surry Y (2004). The fast decay process in outdoor recreational activities and the use of alternative count data models. Am. J. Agr. Econ. 86: 701-715.
Sato K, Kasubuchi M (2005). Economic evaluation of benefit from breeding indigenous horse (in Japanese). J. Rural. Econ., pp. 391-396.
Shaw D (1988). On-site samples' regression: problems of non-negative integers, truncation, and endogenous stratification. J. Econ., 37: 211-223.
Winkelmann R (2003). Econometric analysis of count data 4th edn. Springer, Berlin.