AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Discussion Paper

CHUNG-HUA INSTITUTION FOR ECONOMIC RESEARCH
75 Chang Hsing St., Taipei, Taiwan
Republic of China

Discussion Paper

CHUNG-HUA INSTITUTION FOR ECONOMIC RESEARCH
75 Chang Hsing St., Taipei, Taiwan
Republic of China

DISCUSSION PAPER

NO. 8701

On-site Samples' Regression: Problems of Nonnegative Integers, Truncation, and Endogenous Stratification*

Daigee Shaw

蕭 代 基

January 1987

* Part of the paper is based on part of the author's doctoral dissertation in the School of Natural Resources, the University of Michigan. I would like to thank a number of people for their helpful advice and comments: Dr. Robert Mendelsohn of Yale University, Drs. Carol Jones, Roger Gordon, Alvin Jensen, Gary Solon, and especially Saul Hymans of the University of Michigan. The other part of the paper is done in CHIER. I would like to thank Ms. Irene Chang for her excellent job in programming. I am also indebted to Prof. Takeshi Amemiya and two referees for helpful comments and suggestions.

Abstract

The paper corrects an estimation problem that has not yet been recognized in previous estimates of demand functions using on-site samples. There are three kinds of problems that one faces with in on-site samples, namely, nonnegative integers, truncation and endogenous stratification. Two theoretically correct maximum likelihood methods are developed based on two different assumptions about the variable distribution: the normal distribution and the Poisson distribution. A simulation are performed to compare the two methods using generated data sets of known models. We should not use OLS and instead should use the maximum likelihood methods developed here to estimate demand functions that use on-site samples. The choice between the normal and the Poisson ML methods depends on whether or not one believes that consumers have a negative desired quantity demanded. We cannot say which one is not correct. However, if forecasting is the purpose of estimation, then the simulation indicates that the Poisson ML method may be better.

1. Introduction

This paper arose out of the analysis of recreational site demands. The recreational site demand is the relationship between the number of trips of a consumer in a given period and the price (travel cost) of the trip. The recreational site demand function is widely used to forecast the number of visits to sites and to measure the social benefit of proposed changes of sites. Both the demand forecasts and the social benefit measures can be very useful for making decisions about the uses of natural resources. Unfortunately, because of three intrinsic data problems, all previous recreational site demand function estimations using on-site samples are biased. The three problems (or characteristics) with on-site sampling are:

(a) Nonnegative integers: The dependent variable, the number of trips of a consumer in a given period, is a nonnegative integer. This characteristic exists in population-wide samples too.

(b) Truncation: Only those people who have taken at least one trip are sampled, and all information about non-users is truncated from the sample.

(c) Endogenous stratification: People who go to the site frequently are more likely to be sampled than people who go to the site only occasionally. This is endogenous stratification because the selection of persons to be surveyed is based on a stratified sample design, with

random sampling within strata. The stratum is defined by the endogenous variable, the number of trips.

The purpose of the paper is to develop econometric methods to solve the three problems of on-site samples simultaneously. In addition to recreational site demands, a number of other economic and sociological data should apply the methodology developed here, for instance, the number of days of lost work and the number of hospital visits sampled on hospitals, transportation demand studies sampled on transportation modes, and household demand studies sampled on supermarkets. However, as in recreational site demand studies, the coexistence of the three problems has not been recognized in those other studies, and previous estimations may be biased too.

The three intrinsic problems of on-site samples have been studied separately. The first problem, nonnegative integers, was studied in depth by Hausman et al.(1984). They applied first the simple Poisson regression and then generalized it to allow the Poisson parameter to be randomly distributed, and compounded the two distributions. They derived maximum likelihood estimators (MLE). Hausman, Ostro and Wise (1984), Portney and Mullahy(1986) have applied the simple and the compounded Poisson models to health data. Lee (1986) has derived specification test for the Poisson regression model proposed by Hausman et al.(1984). Mullahy (1986) modified the simple Poisson and Geometric regression to permit the relative probabilities of zero and non-zero realizations to differ from those implied by the parent distributions. The second problem, truncated samples, was first presented by Hausman and Wise (1976,1977), who derived the MLEs for the parameters β and σ^2. Later on, it was studied in

the context of more general limited dependent variable models by Olsen (1980), Greene (1981, 1983), Goldberger (1981), and Chung and Goldberger (1984), who derived the estimators by the method of moments using a non-iterative correction to the ordinary least squares estimator. The third problem, endogenous stratification, was first recognized by Manski and Lerman (1977) and then Hausman and Wise (1982). They derived two conceptually parallel MLEs and a weighted least squares estimator. The complicated coexistence of these three or two problems has not yet been studied. This paper explores this issue by developing and evaluating various methods that can solve these three problems together.

The plan of this paper is as follows:

Section 2: A general demand model and the density function of an on-site sample are developed.

Section 3: Existing estimation methods for this model are reviewed and general inadequacies are noted.

Section 4: Two new estimation methods for the model are presented.

Section 5: A numerical simulation is explored in order to evaluate the two methods.

Section 6: The overall results of the paper are discussed and several recommendations are given.

2. On-site Samples' Probability Density Function

A general demand model for recreation, durable goods, or transportation must be able to explain an obvious phenomenon: Some portion of the population has zero demand for these goods or services in a

given season. Tobin (1958) noted that this kind of data destroys the assumptions of the multiple regression model so that the least squares method is clearly inappropriate.

Tobin (1958) formulated the following model known as the Tobit (censored normal regression) model which generates data having the observed phenomenon of the demand for recreation, durable goods and transportation:

$$y_i^* = X_i\beta + u_i \qquad i=1,\ldots,N \qquad (1)$$

$$y_i = y_i^* \quad \text{if } y_i^* > 0 \qquad (2)$$

$$= 0 \quad \text{if } y_i^* \leq 0$$

where y_i^* can be interpreted as the i-th person's desired quantity demanded; N is the population; X_i is a vector of k independent variables such as prices, travel costs, wages, and income; y_i can be interpreted as the observed quantity demanded; u_i can be interpreted as the collection of all the unobservable variables which affect the utility function, and can be assumed as being distributed with mean 0 and variance σ^2. Then, y_i^* distributes $(X_i\beta,\sigma^2)$ and its density function is denoted by $f(y_i^* \mid X_i)$.

Equation (1) can be estimated by using data sampled population-wide or on-site. The more expensive population-wide random sample is a censored sample which has been studied in depth (Tobin, 1958; Amemiya, 1973; Heckman, 1976, 1979; Hausman et al.,1984). While an on-site random sample is cheaper to survey, it has two intrinsic problems as defined above: truncation and endogenous stratification. The other problem of nonnegative integers is common to both kinds of samples. The on-site sample can be viewed as being sampled randomly from the on-site population which is truncated at zero and is stratified endogenously.

If we know the density function of the ith person in the population is $f(y_i^* | X_i)$, then our problem becomes what the density function of the same person in an on-site population (or sample) is, i.e. what the probability of being sampled on-site is.[1] Suppose that n person in the population of size N have the same X devoted by X^0. Then we can generate the desired quantity demanded $y_1^*,\ldots,y_n^*$ according to $f(y_i^* | X^0)$, and the observed quantity demanded $y_1,\ldots,y_n$ according to equation (2), i.e. according to a censored density function $g(y_i | X^0)$ where $g=f$ if $y_i^*>0$, $g=\int_{-\infty}^{0} f(y_i^*)dy^*$ if $y_i^* \leq 0$. We can reorder y_i such that those $y_i>0$ are the first m. And then we can generate on-site observation y according to the probability distribution:

$$y = \begin{cases} y_1 & \text{with probability } y_1 / \sum_{i=1}^{n} y_i \\ y_2 & \text{with probability } y_2 / \sum_{i=1}^{n} y_i \\ \\ y_n & \text{with probability } y_n / \sum_{i=1}^{n} y_i \end{cases} \qquad (3)$$

Thus, y is truncated at zero and stratified endogenously. What we observe on site is y and X^0 of m persons, $m \leq n$. Then the probability of y equal to a specific value, say t, is

---

1 The following derivation is indebted to the discussion with Prof. T. Amemiya.

$$P_m = P(y_1 = t, y_2,\cdots,y_m, \& y = y_1) + P(y_1, y_2 = t, y_3,\cdots,y_m, \& y = y_2) + \cdots$$

$$P(y_1,\cdots, y_{m-1}, y_m = t \& y = y_m)$$

$$=g(y_1=t\,|\,X^0)\cdot E[t/(t+y_2+\cdots+y_m)]+g(y_2=t\,|\,X^0)\cdot E[t/(y_1+t+y_3+\cdots+y_m)]+\cdots+$$

$$g(y_m=t\,|\,X^0)\cdot E[t/(y_1+\cdots+y_{m-1}+t)]$$

$$\{\text{since } E[t/(t+y_2+\cdots+y_m)]=\sum_{y_m}\cdots\sum_{y_2}[t/(t+y_2+\cdots+y_m)]g(y_2)\cdot g(y_3)\cdots g(y_m)\}$$

$$=m\cdot g(t\,|\,X^0)E[t/(t+y_2+\cdots+y_m)]$$

$$=t\cdot g(t\,|\,X^0)E[1/y] \qquad (4)$$

If we assume m to be large, we can take P_m to infinity

$$\lim P_m=t\cdot g(t\,|\,X^0)E[1/Ey]=t\cdot g(t\,|\,X^0)\cdot(1/Ey)$$

$$= \frac{t\cdot g(t\,|\,X^0)}{\sum_{t=1}^{\infty} t\cdot g(t\,|\,X^0)} \qquad \text{(if y is an integer variable)}$$

$$= \frac{t\cdot f(t\,|\,X^0)}{\sum_{t=1}^{\infty} t\cdot f(t\,|\,X^0)} \qquad \text{(since f=g when t>0)} \qquad (5)$$

This is the density function of an observation y=t and $X=X^0$ in the on-site population. In the general form, it can be written as

$$h(y_i\,|\,X_i) = \frac{y_i\cdot f(y_i\,|\,X_i)}{\sum_{y_i=1}^{\infty} y_i\cdot f(y_i\,|\,X_i)} \qquad (6)$$

If y is a continuous variable, then[2]

$$h(y_i|X_i) = \frac{y_i \cdot f(y_i|X_i)}{\int_0^{\infty} y_i \cdot f(y_i|X_i)\, dy} \qquad (y_i>0) \qquad (7)$$

and u_i given X_i has a density h:

$$h(u_i|X_i) = \frac{(X_i\beta + u_i) \cdot f(u_i)}{\int_{-X_i\beta}^{\infty} (X_i\beta + u_i) \cdot f(u_i)du} \qquad (u_i > -X_i\beta) \qquad (8)$$

y_i^* can be continuous or integer variable in different case. If y_i^* is a continuous variable, we usually assume $f(y_i^*|X_i) \sim IN(X_i\beta,\sigma^2)$. If y_i^* is an integer variable, we can assume $f(y_i^*|X_i) \sim$ I. Poisson(λ_i) following Hausman et al. (1984).

We can show that truncation and endogenous stratification of on-site samples can cause population-wide samples' ordinary estimators to be biased and inconsistent in any case. If $f(y_i^*|X_i) \sim IN(X_i\beta,\sigma^2)$, then the density function of on-site samples is equation (7) and the OLS estimators are biased and inconsistent. This bias occurs because these two problems imply that the expected value of disturbances is not zero and is a function of $X_i\beta/\sigma$. Indeed both the expected value and variances of disturbances are

---

2 The density functions (6) and (7) can be made clearer and don't change the equality by multiplying both the numerator and the denominator of the RHS by N_i, the number of persons having the same X_i. Then, the numerator $y_i \cdot f(y_i|X_i) \cdot N_i$ becomes the total number of trips of those people having X_i and taking y_i number of trips; the denominator $\sum_{y_i=1}^{\infty} y_i \cdot (y_i|X_i) \cdot N_i$ becomes the total number of trips taken by the people having X_i. Thus, $h(y_i|X_i)$ is the sampling probability of a trip taken by those people having X_i and taking y_i number of trips when we sample randomly on site.

functions of $X_i\beta/\sigma$ and are give by

$$E(u_i \mid X_i) = \sigma \cdot Q(d_i) \qquad (9)$$

$$Var(u_i \mid X_i) = \sigma^2 [2 - d_i \cdot Q(d_i) - Q(d_i)^2] \qquad (10)$$

where $d_i=X_i\beta/\sigma$; $q_i=\varphi(d_i)/\Phi(d_i)$; $Q(d_i)=1/(d_i+q_i)$; $\varphi(d_i)$ and $\Phi(d_i)$ are, respectively, the standard normal density and distribution function evaluated at d_i. As a direct consequence of (9), the population regression function for the number of trips, given that $h(y_i \mid X_i)$ is the density function of y_i in the on-site population, is

$$E(y_i \mid X_i) = X_i\beta + E(u_i \mid X_i) \qquad (11)$$
$$= X_i\beta + \sigma \cdot Q(d_i)$$

Since data points for an on-site sample are distributed according to $h(y_i \mid X_i)$, an ordinary least squares (OLS) regression of y on X will closely approximate equation (11). Hence the OLS estimates of β will be biased and inconsistent.

If $f(y_i^* \mid X_i) \sim$ I. Poisson(λ_i), then $y_i^*=y_i$ and

$$f(y_i \mid X_i)=\exp(-\lambda_i)\lambda_i^{y_i}/y_i! \qquad y_i=0,1,2,\cdots \qquad (12)$$

where λ_i is the Poisson parameter, i indexes individuals. We specify $\ln\lambda_i=X_i\beta$ following Hausman et al. (1984). They pointed out that this specification is analogous to the familiar econometric regression specification; in particular, $E(y_i \mid X_i) = \lambda_i$ and estimation of unknown parameters can be done either by an iterative nonlinear weighted least squares (WLS) method or by a maximum likelihood (ML) method. Under this assumption, on-site samples' density function is

$$h(y_i \mid X_i)=\exp(-\lambda_i)\lambda_i^{y_i-1}/(y_i-1)! \qquad (y_i \geq 1) \qquad (13)$$

Its expected value and variance are

$$E(y_i \mid X_i)=\lambda_i+1=\exp(X_i\beta)+1 \qquad (14)$$

$$Var(y_i \mid X_i)=\lambda_i \qquad (15)$$

Thus, if we view on-site samples as population-wide samples and estimate accordingly using the WLS or the ML methods, the estimates of β and λ will be biased and inconsistent. In the following section, I shall consider three commonly used methods of estimation of the recreational site demand functions using on-site samples and demonstrate their inadequacy in solving the problems.

3. Review of Previous Estimation Methods

Clawson's OLS using Zone-average Observations

First let us consider the traditional OLS method using zone-average observations, pioneered by Clawson (1959). Under the normal distribution assumption the method gives biased results.[3] On the other hand, under the Poisson distribution assumption, the OLS method is not appropriate, however, the WLS and ML methods of population-wide samples are appropriate.

First, our underlying model (1)+(2) is modified slightly to bring in circular zones:

$$y_{ij}^* = X_j\beta + u_{ij} \qquad i=1,\ldots,N_j \qquad j=1,\ldots,J \qquad (16)$$

$$y_{ij} = y_{ij}^* \quad \text{if } y_{ij}^* > 0 \qquad (17)$$

$$= 0 \quad \text{otherwise}$$

where y_{ij}^* is the desired number of trips taken by individual i living in zone j, N_j is the population in zone j. The zone-average method implicitly

[3] This is the problem of aggregation in the Tobit (censored normal regression) model which has been studied by Maddala and Nelson (1975), and Maddala (1983, p.183).

assumes that the values of X_j are the same for each i in zone j. The dependent variable becomes each zone's visitation rate

Y_j = total number of trips to the site from the population in zone j

/ population in zone j, (18)

which is aggregated over the individual observable number of trips. Then visitation rates Y_j are regressed on X_j.

If $f(y_{ij}^* | X_j) \sim IN(X_j\beta, \sigma^2)$, then the expected value of Y_j is

$$E(Y_j) = \Phi(X_j\beta/\sigma) \cdot X_j\beta + \sigma \cdot \varphi(X_j\beta/\sigma).^4 \quad (19)$$

which is actually the expectation function of the censored normal regression model. This result is clearly due to the fact that Y_j is the incomplete aggregation over the observable y_{ij} instead of the complete aggregation over the desired, but partly unobservable, y^*_{ij} in the beginning. Thus, a straightforward OLS regression of Y_j on X_j will give biased estimates of β.

If $f(y_{ij}^* | X_j) \sim$ I. Poisson (λ_{ij}), then

$$E(Y_j) = \lambda_{ij} = \exp(X_{ij}\beta) \quad (20)$$

Thus, the expected values of Y_j and y_{ij}^* are the same. This is because $y_{ij}^* = y_{ij}$ in the Poisson distribution, i.e., the desired quantity demanded is the same as the observed. As a consequence, the WLS and ML methods of population-wide samples are appropriate, and the OLS method is not appropriate still.

In addition to bias, there are other theoretical and empirical disadvantages of aggregate demand analysis. First, the basic unit of observation in the aggregated model is a zone which can be inconsistent

---

4 The proof is in Maddala (1983, p.184).

with the theory of consumer behavior where the decision-maker is an individual, not a zone. Second, the aggregation causes losses of efficiency due to multicollinearity, insufficient dispersion among aggregated independent variables, and the loss of information about the variation of observations within each zone (Kmenta, 1971, pp. 322-29; Brown and Nawas, 1973). Thus, the zone-average method is far from being a perfect method to estimate recreational site demand functions using on-site samples.

The Twin Linear Probability Method

The term twin linear probability method was first used by Goldberger (1964) and was applied to recreation demand studies by Davidson, Adam and Seneca (1966), Cicchetti, Seneca and Davidson (1969), and Russell and Vaughan (1982). The method is based on the equation

$$E(y_i \mid X_i)=Prob(y_i>0 \mid X_i)\cdot E(y_i \mid y_i>0,X_i) \qquad (21)$$

The method is also based on a two-stage decision framework in which individuals first decide whether or not to participate, $Prob(y_i>0 \mid X_i)$, and then decide on how frequently to participate, $E(y_i \mid y_i>0,X_i)$. Therefore, the method has two steps. In the first step, $Prob(y_i>0 \mid X_i)$ is estimated by fitting a linear probability model, Probit model or Logit model for the (0, 1) dichotomous variable, where 0 represents non-users and 1 users. Thus, this method needs a population-wide sample. In the second step, $E(y_i \mid y_i>0, X_i)$ is obtained from an OLS regression on the user subsample of the population-wide sample.

This method's problem occurs in the second step. If $f(y_i^* \mid X_i)\sim IN(X_i\beta,\sigma^2)$, then the expectation function of the truncated normal

regression function is

$$E(y_i \mid y_i>0, X_i) = X_i\beta + \sigma \cdot q(d_i) \qquad (22)$$

where $d_i=X_i\beta/\sigma$ and $q(d_i)=\varphi(d_i)/\Phi(d_i)$. Given a person's X_i, we can predict his/her participating probability from the first step result, $\hat{\Phi}(d_i)$, and his/her expected number of trips as a user from the second step result

$$\hat{y}_i = X_i\hat{\beta}. \qquad (23)$$

Consequently, his/her expected number of trips as an individual (user or nonuser) in the population is

$$\hat{\hat{y}}_i = X_i\hat{\beta}\cdot\hat{\Phi}(d_i) \qquad (24)$$

$$\approx X_i\beta(1-d_i\cdot q(d_i)-q^2(d_i))\cdot\hat{\Phi}(d_i)$$

which is obviously biased away from true $y_i^*=X_i\beta$.

If $f(y_i^* \mid X_i)\sim$ I. Poisson(λ_i), then

$$E(y_i^* \mid X_i)=Prob(y_i>0 \mid X_i)\cdot E(y_i \mid y_i>0,X_i)$$

$$=(1-\exp(-\lambda_i))\cdot(\lambda_i/(1-\exp(-\lambda_i))) \qquad (25)$$

where $\lambda_i=\exp(X_i\beta)$. It is very obvious that the use of Probit model and OLS method to estimate the two steps is inappropriate. Actually it is not necessary to do a Poisson two-stage estimation instead of one-step.

The correct two-stage estimation of the Tobit model under normal distribution should be the famous procedure proposed by Heckman (1976). After estimating β and σ, given a person's X_i, we can get predictions of his/her desired number of trips from equation (1), or of his/her observable number of trips as a user from equation (22), or of his/her observable number of trips as an individual (user or nonuser) in the population from equation (19).

Brown et al.'s OLS using Adjusted Individual Observations

Brown et al. (1983) recommend adjusting the dependent variable of an on-site sample, the number of trips per user, to a per capita basis which they call individual observed visits per capita. This adjusted dependent variable is then regressed on individual unadjusted independent variables as in

$$(y_{ij} \mid y_{ij}>0, X_{ij}) \cdot \text{Prob}(y_{ij}>o \mid X_{ij}) = X_{ij}\beta+u_{ij}. \qquad (26)$$

They illustrate that this method will produce the same recreational site demand functions as Clawson's OLS using zone average observations.

As a matter of fact, this method has several problems and will produce biased results. First, $\text{Prob}(y_{ij}>0 \mid X_{ij})$ in (26)— (total number of users given X_{ij})/(population given X_{ij})— is not available from on-site samples because of endogenous stratification. Second, even if the sample is collected population-wide and the $\text{Prob}(y_{ij} \mid X_{ij})$ is available, the regression still gives biased results. Under the normal distribution assumption, the expected value of the LHS of equation (26) is

$$E[(y_{ij} \mid y_{ij}>0, X_{ij}) \cdot \text{Prob}(y_{ij}>0 \mid X_{ij})] \qquad (27)$$
$$= \Phi(d_{ij}) \cdot X_{ij}\beta + \sigma \cdot \varphi(d_{ij}).$$

where $d_{ij}=X_{ij}\beta/\sigma$. This expectation function is exactly the same as equation (19). Thus, as illustrated by Brown et al. for a generated population-wide sample (which they think is an on-site sample), the OLS regression of $(y_{ij} \mid y_{ij}>0,X_{ij}) \cdot \text{Prob}(y_{ij}>o \mid X_{ij})$ on X_{ij} will produce the same results as the zone average method. However, both give biased estimates of β. The conclusion is the same under the Poisson distribution assumption. Third, in the case of an on-site sample, under the normal distribution assumption, the expected value of the LHS of equation (26) is

$$E[(y_{ij} \mid y_{ij}>0, X_{ij}) \cdot \text{Prob}(y_{ij}>o \mid X_{ij})] \qquad (28)$$

$$= \Phi(d_{ij})\, X_{ij}\beta + \sigma \cdot \Phi(d_{ij}) \cdot Q(d_{ij})$$

where $d_{ij}=X_{ij}\beta/\sigma$, $q(d_{ij})=\varphi(d_{ij})/\Phi(d_{ij})$, $Q(d_{ij})=1/(d_{ij}+q(d_{ij}))$, and assuming $\text{Prob}(y_{ij}>o \mid X_{ij})$ is available by other means. Thus, the OLS estimates of β are biased and will be different from the zone average method estimates, contrary to Brown et al. On the other hand, under the Poisson distribution assumption, its expect function is $(1-\exp(-\lambda_{ij}))\cdot(\lambda_{ij}+1)$. Thus, the OLS method is inappropriate and the appropriate MLE will be different from that of the zone average method.

In conclusion, all previous estimation methods using on-site samples result in biased recreational site demand equations and social benefit estimates. In the following section, I shall develop estimation methods using the on-site samples.

4. New Estimation Methods

Two ML methods are developed here to estimate parameters using on-site samples. In the first one, the dependent variable's distribution is assumed to be normal. Thus, only the problems of truncation and endogenous stratification are considered. While in the second one, the dependent variable's distribution is assumed to be Poisson, and the problems of truncation, endogenous stratification and nonnegative integers are considered.

Method One: The Normal ML Method

First let us consider the maximum likelihood method of normal distribution. We are concerned with the on-site sample, whose density function $h(y_i|X_i)$ is given in equation (7). Therefore, the log-likelihood function is

$$L=\ln \angle=-n\cdot\ln(\sigma^2\sqrt{2\pi})+\sum_{i=1}^{n}\ln y_i-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-X_i\beta}{\sigma}\right)^2-\sum_{i=1}^{n}\ln[d_i\cdot\Phi(d_i)+\varphi(d_i)] \qquad (29)$$

The parameter estimates can be obtained by maximizing L with respect to β and σ^2. We can write down first and second derivatives of L with respect to β and σ^2. Then the first order conditions are

$$\frac{\partial L}{\partial \beta}=\frac{1}{\sigma^2}\sum_{i}^{n}(y_i-X_i\beta)X_i-\frac{1}{\sigma}\sum_{i}^{n}Q(d_i)\,X_i=0 \qquad (30)$$

$$\frac{\partial L}{\partial \sigma^2}=-n\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{i}^{n}(y_i-X_i\beta)^2+\frac{1}{2\sigma^2}\sum_{i}^{n}Q(d_i)\,d_i=0 \qquad (31)$$

Postmultiplying equation (30) by $\beta/2\sigma^2$ and adding the result to equation (31), we get

$$\sigma^2=\frac{1}{2n}\mathbf{Y}'(\mathbf{Y}-\mathbf{X}\beta) \qquad (32)$$

where $\mathbf{Y}$ is a $n\times1$ vector of n observations on y_i; $\mathbf{X}$ is a $n\times k$ matrix of values of X_i, which is a $1\times k$ vector; and β is still a $k\times1$ vector. After multiplying throughout by σ, equation (30) can be written as

$$(1/\sigma)\,\mathbf{X}'(\mathbf{Y}-\mathbf{X}\beta)-\mathbf{X}'\mathbf{Q}=0$$

$$\beta_{MLE}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}-\sigma\cdot(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q} \qquad (33)$$

$$=\beta_{LS}-\sigma\cdot(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}$$

where $\mathbf{Q}$ is a n×1 vector of values of $Q(d_i)=1/(d_i+q(d_i))$, and β_{LS} is the OLS estimator for β obtained from the on-site sample. This equation shows that the OLS estimators of β are biased when on-site samples are used. In the more general case, the bias of the OLS estimator of individual parameters cannot be determined a priori.

Since σ^2 is a function of β in equation (32) and β is also a function of σ^2 in equation (33), there is no way to estimate β and σ^2 directly as in OLS. Of course, there are several standard iteration methods for obtaining the MLE's of β and σ^2, such as the Newton-Raphson method, the method of scoring, the BHHH method and the Fair's (1977) method. The Fair's method obtains estimates by an iteration using equations (32) and (33) and can be done fairly easily using some simple matrix operations and OLS.

The variance-covariance matrix of β and σ^2 can be estimated after the MLE's of β and σ^2 are obtained. The variance-covariance matrix is the inverse of the information matrix. The information matrix is

$$\begin{bmatrix} -E\left[\frac{\partial^2 L}{\partial\beta^2}\right] & -E\left[\frac{\partial^2 L}{\partial(\sigma^2)^2}\right] \\ -E\left[\frac{\partial^2 L}{\partial(\sigma^2)^2}\right] & -E\left[\frac{\partial^2 L}{\partial\beta^2}\right] \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{X}'(\mathbf{X}+\mathbf{P}_1\mathbf{X}) & -\frac{1}{2\sigma^3}\mathbf{X}'\mathbf{P}_2 \\ -\frac{1}{2\sigma^3}\mathbf{P}_2'\mathbf{X} & \frac{1}{4\sigma^4}\mathbf{P}_3'\mathbf{I} \end{bmatrix} \qquad (34)$$

where $\mathbf{P_1}$ is a n×n diagonal matrix whose value of the i-th element is $P_{1i}=[q(d_i)\cdot Q(d_i)-Q^2(d_i)]$; $\mathbf{P_2}$ is a n×1 vector of values of $P_{2i}=Q(d_i)+d_i\cdot P_{1i}$; $\mathbf{P_3}$ is a n×1 vector of values of $P_{3i}=3Q(d_i)\cdot d_i+d_i^2\cdot P_{1i}$; and $\mathbf{I}$ is a n×1 vector of constant 1. Then, the asymptotic variance of the MLE's of β and σ^2 are given by the diagonal elements of the inverse of the information matrix, i.e., the Cramer-Rao lower bound.

The MLE's are consistent and asymptotic normal. Amemiya (1973) has proved that the properties of the MLE's—consistency, asymptotic normality, and attainment of the Cramer-Rao lower bound—hold for the Tobit model. This proof can be easily altered to cover the present case.

Method Two: The Poisson ML Method

Equation (12) specifies the Poisson regression model already. Equations (13), (14), (15) are the density function, expected value and variance of on-site samples. Its log-likelihood function is

$$L = \sum_{i=1}^{n} [\ \ln y_i - e^{X_i\beta} + (y_i - 1)\ X_i\beta - \ln y_i!] \qquad (35)$$

The first and second derivatives are

$$\partial L/\partial\beta = -\mathbf{X'\lambda} + \mathbf{X'(Y-I)} \qquad (36)$$

$$\partial L/\partial\beta = -\mathbf{X'\lambda X} \qquad (37)$$

Where $\mathbf{X}$ is a n×k matrix of values of X_i, $\boldsymbol{\lambda}$ is a n×1 vector of λ_i, $\mathbf{Y}$ is a n×1 vector of y_i, $\mathbf{I}$ is a n×1 vector of constant 1. Its information matrix is $\mathbf{X'\lambda X}$.

The MLE of β can be obtained by the Newton-Raphson method, the method of scoring or the BHHH method.

5. The Comparison of Methods——Monte Carlo Simulation

The above two methods are different in their underlying assumptions of consumer behavior and forecasting abilities, even though both try to solve the same problems of on-site samples. Under the assumption of normal distribution, consumer's desired quantity demanded may be negative. On the other hand, the Poisson distribution requires that consumer's desired quantity demanded cannot be negative. Both consumer behavior assumptions have their microfoundations. However, if our purpose of estimating demand functions is to do forecasting, then we can expect the forecasting performance of the normal MLE may be weaker than the Poisson MLE because the negative desired quantity demanded will make the true regression line of the former assumption deviate away from the expected value of the observed quantity demanded which is the true regression line of the latter assumption. Thus, the estimated regression line of the normal MLE will underestimate the observed quantity demanded.

In order to evaluate their forecasting performances, I perform a Monte Carlo simulation (MC). First, for each distribution assumption, I create a particular recreational site demand model and generate 100 small size data sets representing on-site samples. For each data set I estimate parameters using the above two ML methods separately and calculate the mean and standard deviation for each parameter estimates. At last, I test the significance of the difference between both methods' forecasting performances using two large size out-of-samples and the test procedure of Ashley etal. (1980).

The normal distribution data sets are generated according to the model

$$y_i^* = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \qquad i=1,\ldots,N \qquad (38)$$

$$y_i = y_i^* \quad \text{if } y_i^* > 0 \qquad (39)$$

$$= 0 \quad \text{otherwise}$$

where x_{1i} are travel costs, x_{2i} are wages, y_i^* are the desired numbers of trips to a given site, y_i are the observable number of trips, and $u_i \sim N(0,\sigma^2)$. The parameters take values $\beta_0=25$, $\beta_1=-0.3$, $\beta_2=-0.1$, $\sigma^2=25$, These parameter values are chosen to roughly represent actual data with approximately 70% of the observations truncated at zero. x_{1i} and x_{2i} are distribute independently and uniformly between 0 and 300, and 0 and 30 respectively.

For each of the 100 data sets, using standard number generators, a sample of N = 100 for x_1, x_2, and u are generated based on the above specification. For each observation, I calculate y_i^* and y_i using equation (38) and (39) respectively. Then I round off the y_i to the nearest integer, to truncate the data set at 0, and to replicate each observation y_i times. This is the data set representing an on-site sample. The final sizes of the 100 data sets are around 400.

On the other hand, the Poisson distribution data sets are generated according to the model

$$y_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \qquad i = i,\cdots,N$$

where $\beta_0=0.5$, $\beta_1=-0.3$, $\beta_2=0.01$, N=300. y_i are distributed independently and Poissonly with parameter λ_i which is between 0.5 and 10.5. x_{1i} are distributed uniformly between 0 and 100. For each of the 100 data sets,

using standard number generators, a sample of N = 400 for y and x_{1i} are generated. Let $y_i' = y_i+1$. Then y_i' is the on-site sample's dependent variable. This can be easily seen from equation (14). For each observation, I calculate x_{2i} according to $x_{2i}=(\ln y_i' - \beta_0 - \beta_1 x_{1i})/\beta_2$. The resulting observations of y_i', x_{1i}, x_{2i} represents an on-site sample under the Poisson distribution.

With each generated data sets, the normal ML and the Poisson ML methods are used to estimate parameters. The two methods are estimated by the Newton-Raphson method and the BHHH method which are programmed in SAS's PROC MATRIX by the author. The mean and standard deviation of the 100 estimates of each parameter are presented in Table 1 and 2. The forecasting performances measured in MSE (mean squared error) are also presented.

The results of the MC generally indicate that the two estimation methods are correct, since the normal and the Poisson ML estimates are very close to the true value of the normal and the Poisson data sets respectively. On the other hand, as expected, the normal and the Poisson ML estimates are very different from the true value of the Poisson and the normal data sets respectively, since the methods and the data sets are setted up by two different distribution assumptions.

In order to compare forecasting performances, I generate an out-of-sample of size 2000 for each distribution model. These two out-of-samples are truncated data sets representing only the users' of two population-wide samples. Why population-wide samples? Because the two ML methods are used to estimate the population regression lines. Why

TABLE 1[a]

THE MEANS AND STANDARD DEVIATIONS OF THE MONTE-CARLO SIMULATION OF 100 ON-SITE SAMPLES UNDER THE NORMAL DISTRIBUTION

	True Value	Normal ML Method	Poisson ML Method
β_0	25	24.7882	3.9950
		(4.3638)	(0.5846)
β_1	-0.3	-0.2981	-0.0471
		(0.0386)	(0.0263)
β_2	-0.1	-0.0933	-0.0200
		(0.0851)	(0.0183)
σ^2	25	8.7039	
		(8.9906)	
MSE		25.8969	58.3135

[a] The standard deviations are in the parentheses.

TABLE 2[a]

THE MEANS AND STANDARD DEVIATIONS OF THE MONTE-CARLO SIMULATION OF 100 ON-SITE SAMPLES UNDER THE POISSON DISTRIBUTION

	True Value	Poisson ML Method	Normal ML Method
β_0	0.5	0.5864	2.1589
		(0.0577)	(0.6221)
β_1	-0.3	-0.2909	-1.0681
		(0.0046)	(0.0438)
β_2	-0.01	-0.0097	-0.0355
		(0.0002)	(0.0012)
σ^2			6.6291
			(0.1835)
MSE		5.4788	6.4860

[a] The standard deviations are in the parentheses.

users only? Because the comparison of forecasting non-users' behavior will result in too much disadvantage to the normal ML method where nonusers' demands are assumed to be negative.

Then, MSE's are calculated as in Table 1 and 2. Although the two pairs of MSE's are different, we cannot say anything sure without a formal test. With a finite data set, the test procedure proposed by Ashley et al. (1980) could be used to test if the two MSE's are significantly different. The result of the test indicates that, for the normal distribution assumption, the Poisson ML method cannot be regarded as a significant improvement over the normal ML method in forecasting performance.[5] However, for the Poisson distribution assumption, the Poisson ML method can be regarded as a significant improvement. This implies that we can support the argument that the forecasting performance of the normal MLE may be weaker than the Poisson MLE. This is hardly overwhelming evidence, but it does suggest that the Poisson ML method is more useful in forecasting than the normal ML.

6. Conclusion and Further Topics

The problems associated with regression analysis based on nonnegative integers, truncated or endogenously stratified samples have been studied separately in the last decade. Unfortunately, the coexistence of the three problems in on-site samples has not yet been recognized. As a result, all previous estimates of demand functions using on-site samples are subject to possible bias. In this paper, I have proved that this bias is

[5] In Appendix, I shall describe the test in detail.

likely to exist, and I have developed two ML methods to correct the three problems together based on two distribution assumptions.

The choice between the normal and the Poisson ML methods depends on whether or not we believe that consumers have a negative desired quantity demanded. We cannot say which one is not correct. However, if forecasting is the purpose of estimation, then the MC simulation indicates that the Poisson ML method may be better.

Although the Poisson regression does not have the problem of negative desired quantity demanded, its mean equalling variance does pose some empirical constraints. Thus, a possible topic is to improve the Poisson regression to avoid the problem. Although Hausman et al. (1984) have developed several ML based on non-negative integer distributions in addition to Poisson, the recent paper by Lee (1986) may provide a clue to use a general nonnegative integer distribution which includes Poisson, binomial, and negative binomial distributions as special cases. The parameter estimates can tell us which distribution the data belongs to.

At last, finding ways to expand the single equation ML method to deal with the estimation of a demand system using on-site samples is another interesting topic.

Appendix

Ashley et al. (1980) proposed an indirect test of the significance of improvements in MSE of forecasting. First, do the regression equation

$$\Delta_t=\beta_1+\beta_2[\Sigma_t-m(\Sigma_t)]+u_t$$

where $\Sigma_t=e_{1t}+e_{2t}$, $\Delta_t=e_{1t}-e_{2t}$, $m(\Sigma_t)$ denotes the mean of Σ_t, e_{1t} and e_{2t} are the forecast errors made by the two models under comparison, u_t is an error term with mean zero. If we can reject the null hypothesis $\beta_1=\beta_2=0$, than we can conclude that the second model outperforms the first model. If either of β_1 and β_2 is significantly negative, the second model clearly cannot be judged a significant improvement. If one estimate is negative but not significant, a one-tail t-test on the other estimated coefficient can be used.

The following two regressions are obtained for the simulated comparison studies. (Standard errors are in parenthesis)

$$\Delta_t=0.28809 - 0.25789\,[\,\Sigma_1-m(\Sigma_2)] \quad \text{(Normal distribution data sets)}$$
$$(0.12036)\ (0.01087)$$

$$\Delta_t=-0.03263 + 0.04352\,[\Sigma_1-m(\Sigma_1)] \quad \text{(Poisson distribution data sets)}$$
$$(0.02018)\ (0.00420)$$

The first regression clearly indicates that, for the normal distribution data sets, the Poisson ML method cannot be regarded as a significant improvement over the normal ML method in forecasting performance, since β_2 is significantly negative. The second regression clearly indicates that, for the Poisson distribution data sets, the Poisson

ML method outperforms the normal ML method since β_1 is negative but not significant, β_2 is significantly positive. We can reject the null hypothesis $\beta_1=\beta_2=0$.

References

Ashley, R., C. W. J. Granger, and R. Schmalensee, 1980, Advertising and aggregate consumption: an analysis of causality, Econometrica 48(5), 1149-1167.

Amemiya, Takeshi, 1973, Regression analysis when the dependent variable is truncated normal, Econometrica 41, 997-1016.

Berndt, E., B. H. Hall, R. Hall, and J. Hausman, 1974, Estimation and inference in nonlinear structural models, Annals of Economic and Social Measurement 3, 653-665.

Brown, William and Farid Nawas, 1973, Impact of aggregation on the estimation of outdoor recreation demand functions, American Journal of Agricultural Economics 55,246-49.

Brown, William, C. Sorhus, B. Chou-Yang, and Journal A. Richards, 1983, Using individual observations to estimate recreation demand functions: a caution, American Journal of Agricultural Economics 65,154-57.

Chung, Ching-Fan and Arthur Goldberger, 1984, Proportional projections in limited dependent variable models, Econometrica 52(2),531-34.

Cicchetti, C. J., J. J. Seneca, and P. Davidson, 1969, The Demand and Supply of Outdoor Recreation (Bureau of Economic Research, Rutgers University, New Brunswick, N. J.).

Clawson, Marion, 1959, Methods of Measuring the Demand for and Value of Outdoor Recreation (Reprints No. 10, Resources for the Future, Inc., Washington, D. C.).

Davidson, Paul, F. Gerard Adam, and Joseph Seneca, 1966, The social value of water recreation facilities resulting from an improvement in water quality: the Delaware estuary, in: Allen Kneese and Stephen Smith eds., Water Research (Resources for the Future, Inc., Washington, D. C.).

Fair, R.C., 1977, A note on the computation of the Tobit estimator, Econometrica 45(7),1723-7.

Goldberger, Arthur, 1964, Econometric theory (Wiley, N. Y.).

Goldberger, Arthur, 1981, Linear regression after selection, Journal of Econometrics 15,357-66.

Greene, William, 1981, On the asymptotic bias of the ordinary least squares estimator of the Tobit model, Econometrica 49(2),505-13.

Greene, William, 1983, Estimation of limited dependent variable models by ordinary least squares and the method of moments, Journal of Econometrics 21(2),195-212.

Hausman, J. A., B. H. Hall, and Zvi Griliches, 1984, Econometric models for count data with an application to the patents-R & D relationship, Econometrica 52, 909-938.

Hausman, J. A., B. D. Ostro, and D. A. Wise, 1984, Air Pollution and Lost Work, (NBER Working Paper No. 1263).

Hausman, J. A., and D. A. Wise, 1976, The evaluation of results from truncated samples: the New Jersey Negative Income Tax Experiment, Annals of Economic and Social Measurement 5, 421-445.

Hausman, J. A., and D. A. Wise, 1977, Social experimentation, truncated distributions and efficient estimation, Econometrica 45, 919-939.

Hausman, J. A., and D. A. Wise, 1982, Stratification on endogenous variables and estimation: the Gary Experiment, In C. Manski and D. McFadden (eds.), Structural Analysis of Discrete Data: with Econometric Applications (M.I.T. Press, Cambridge).

Heckman, J., 1976, The common structure of statistical models of truncation, sample selection and limited dependent variables and

a simple estimator for such models, Annals of Economic and Social Measurement 5,475-92.

Heckman, J., 1979, Sample selection bias as a specification error, Econometrica 47,153-61.

Kmenta, J., 1971, Elements of Econometrics (MacMillan, N.Y.).

Lee, Lung-Fei, 1986, Specification test for Poisson regression models, International Economic Review 27(3), 689-706.

Maddala, G. S., 1983, Limited-dependent and Qualitative Variables in Econometrics (Cambridge University Press, Cambridge).

Maddala, G. S., and F. Nelson, 1975, Switching regression models with exogenous and endogenous switching, Proceedings of the American Statistical Association (Business and Economics Section), 423-6.

Manski, C., and S. Lerman, 1977, The estimation of choice probabilities from choice-based samples, Econometrica 45,1977-88.

Mullahy, John, 1986, Specification and testing of some modified count data models, Journal of Econometrics 33, 341-365.

Olsen, Randall, 1980, Approximating a truncated normal regression with the method of moments, Econometrica 48(5),1099-1105.

Russell, Clifford S., and William J. Vaughan, 1982, the national recreational fishing benefits of water pollution control, Journal of Environmental Economics and Management 9,328-54.

Tobin, J., 1958, Estimation of relationships for limited dependent variables, Econometrica 26(1),24-36.