Modeling Demand for Outdoor Recreation with Choice-Based Samples

by

Kavita Sardana

(Under the direction of John C. Bergstrom)

Abstract

This dissertation consists of three essays on theoretical and empirical issues related to modeling choice-based samples. The research objectives include investigating the joint effect of endogenous and exogenous stratification in modeling recreational preferences, estimating recreational demand functions by relaxing the basic and naive assumption of randomness for a choice-based clustered sample, and modeling heterogeneous recreational preferences in a latent-class model. In this dissertation, it is shown that welfare estimates of willingness to pay are extremely sensitive to model specification when controlling for sampling related problems.

In the first essay, it is shown that estimating regional demand models by pooling different samples without correcting for such differences causes model misspecification when each sample belongs to a different population. Estimating a weighted regression using pseudolikelihood improves the efficiency of estimates after correcting for heteroskedasticity. The estimates still remain biased as the weights interact with covariates to explain part of model misspecification. The comparisons between weighted and unweighted models go unnoticed because results from both models are rarely reported. By reporting results from both models it is shown that it is best to use unweighted regression when the coefficient on interactions

with the weight variable are jointly insignificant. However, the model needs to be respecified if these interactions are jointly significant but the estimation still proceeds using an unweighted regression.

In the second essay, the dependence between individuals surveyed at a particular site in a choice-based sample is modeled. This dependence is due to some observed or unobserved site-specific effects. Individuals surveyed at a given site are most likely correlated rather than independent. The above argument is used to develop a mixture model where site specific random effects follow a standard normal distribution. When evaluating policy changes such as opening a new site, developing an existing site, or closing an old site, significant site effects show that the expected mean calculations which are used in deriving welfare estimates are sensitive to assumptions about the sampling procedure.

In the third essay, heterogeneous recreational preferences in a latent-class model are considered. This class is based on frequency of visits to a National Forest and is treated as latent due to arbitrariness in defining how many visits constitute high or low frequency. The results show different marginal effects for the two populations; high frequency visitors, who take frequent short duration visits mostly to general forest areas; and low frequency visitors who take less frequent, long duration trips mostly to developed sites. This information on market segregation between high and low frequency visitors can be of importance to the USDA Forest Service because differences in consumer surpluses across classes provide potential scope for differential pricing policies.

INDEX WORDS: Choice Based Samples, Recreational Demand Models, On-Site Samples, Stratification, Weighting, Finite Mixture Models, Infinite Mixture Models, Latent class Models, Dissertations, Theses (academic)

Modeling Demand for Outdoor Recreation with Choice-Based Samples

by

Kavita Sardana

B.A., Delhi University, India, 2001.

M.A., Jawaharlal Nehru University, India, 2004.

MPhil., Jawaharlal Nehru University, India, 2006.

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2010

Modeling Demand for Outdoor Recreation with Choice-Based Samples

by

Kavita Sardana

Approved:

Major Professor: John C. Bergstrom

Committee: J.M. Bowker
C.M. Cornwell

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2010

Acknowledgements

It is a pleasure to thank those who made this research possible. First of all, I would like to give my special thanks to my supervisor, Dr. Bergstrom, for patiently guiding and improving my work. Dr. Bergstrom provided timely and instructive comments and evaluations from the preliminary to the concluding level, allowing me to complete this research on schedule.

I am grateful to Dr. Bowker and Dr. Cornwell for serving on my committee. They have generously given their time and expertise to better my work.

I am grateful to the computer staff and all other members of staff of the Department of Agricultural and Applied Economics. I am thankful to my colleagues and friends at Conner Hall. Special credits to Oleksiy Tokovenko, who was always willing to help and give his best suggestions.

I am grateful to Thomas Montague for helping me proofread my dissertation. His help has been invaluable in improving the writing quality of this research.

I would also like to thank my parents, my brother and my fiancé for their support and encouragement.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Choice-based samples are nonrandom samples based on some kind of stratification. A subsample of the population consisting of subjects with one outcome is collected. For example, the outcome would be participation by the user population when modeling an on-site sample. Data is then collected within the subsample with different attributes of the user population varying the outcome. The primary reason for using choice-based samples is that the outcome variable is a rare event and using household survey data would require an immense amount of data collection effort, which is in most cases is implausible and expensive. Choice-based samples provide economies of scale, which are not available with household surveys. Choice based samples have been used often when modeling demand for mode of transport in the analysis of transportation behavior. For more examples and benefits of choice-based sampling refer to the introduction by Manski and Lerman (1977).

Some of the predominant problems with on-site samples that are choice-based are truncation, endogenous stratification, and non-negativity. These problems are briefly explained. Choice-based sampling has its drawbacks in that nonusers are not included in the sample which causes a truncated population. At this point it is important to clarify the distinction between censoring and truncation. An on-site sample is truncated but not censored. Censoring is when data is unobservable for a certain size of the population, whereas truncation is when the probability mass is cut off at a specific value. An example of censoring would be

data on wage income for the population not included in the labor force, which is by definition unavailable.

Econometric modeling therefore needs to account for truncation related problems within the distributional assumptions. Another problem with choice-based samples is stratification. However, when stratification is on the exogenous variable, the econometric modeling would proceed in a normal fashion. It is only when stratification is on the dependent variable that the econometric modeling needs to correct for such stratification to derive consistent estimates. The dependent variable in most cases in these samples is a non-negative integer. In this case, the outcome variable is modeled as a discrete and not a continuous process. The most frequently used distribution is the poisson which is derived from the binomial distribution. As the number of draws in a binomial distribution approach infinity, the binomial distribution approaches poisson. Therefore, poisson is an asymptotic outcome. As Hellerstein (1991) points out,

> "Count data distribution such as poisson is an asymptotic outcome i.e if the probability of taking a trip on any given day is small, constant and independent of earlier decisions",

1.2 Literature Review

Before discussing how the above problems with choice-based samples are resolved in the literature, frontier models used in modeling recreational preferences are briefly discussed. Two-part models include any kind of Tobit models and/or hurdle models. Any kind of two-part models which involve modeling the first stage as the participation equation cannot be applied to an on-site sample for the precise reason that on-site samples are truncated and not censored. Another class of prominently used models of recreational demand includes extreme corner solution models or continuous/discrete models and generalized corner solution models.

Extreme corner solution models (Hanemann, 1984) are derived by imposing mutual exclusivity on the consumption of commodities in the consumption basket. Therefore only one commodity is consumed, as the name extreme corner solution suggests. In the case of recreational demand models, this would mean that only one recreational site is visited. These models can be applied to on-site samples. The only problem is that the consumption quantity is modeled as a continuous process. Therefore, they are also called Discrete/Continuous models, where the first stage is modeled as a discrete process in a RUM setting and the second stage models consumption of the quantity as a continuous process. Two-stage budgeting models (Hausman et al., 1995) are similar to extreme corner solution models with the difference that the consumption quantity is modeled as a discrete process. Generalized corner solution models first applied by Phaneuf et al. (2000) to modeling recreational preferences can model zero consumption along with positive consumption across the consumption basket. These models can be used if information is available for different sites visited by the same individuals. Other advanced models can be used if recreational preferences of the same individuals are observed across different time periods, including multinomial poisson log normal models (Egan and Herriges, 2006), and seemingly unrelated negative binomial models (Winkelmann, 2000), in a panel settings by modeling correlation among individuals across different periods.

Now it will be shown how the problems related to choice-based samples have been resolved in the literature. The problem of non-negative integers is easily resolved by using count models. Modeling the number of visits to a National Forest results in endogenous stratification because more avid visitors staying close to the forest have a higher probability of being in the sample. Shaw (1988) derived the distribution correcting for the joint effects of truncation and endogenous stratification for the poisson distribution. Englin and Shonkwiler (1995) derived the distribution correcting for the joint effects of truncation and endogenous stratification for negative binomial distribution. The later is used predominantly in the literature in estimating demand for on-site recreational demand.

1.3 Objectives

This dissertation will go beyond the problems of non-negativity, truncation, and endogenous stratification in modeling choice-based samples. The specific objectives of this dissertation are to:

1. Investigate the joint effects of endogenous and exogenous stratification on the inference about visitors' recreational preferences by jointly estimating both weighted and unweighted regional demand models of recreation. In estimating these models, we assume that in an unweighted model, the population of interest is the region and is known. However, in a weighted regional model, we assume that the population of interest is unknown and can be approximated closely by using sample weights;
2. Model dependence among individuals sampled at the same site within a given stratum for a stratified choice-based sample using a finite mixture model.
3. Model heterogeneous recreational preferences by assuming different distributions for two latent classes of visitors using a latent class model and showing that the welfare measures differ for the two classes, suggesting a potential for differential pricing policies.

1.4 Organization of the Study

This dissertation consists of five chapters. The introductory chapter (Chapter 1), three essays on modeling choice-based samples (Chapters 2, 3, and 4), and a concluding chapter (Chapter 5) summarizing the results of this dissertation. The first appendix at the end of this dissertation include weighted and unweighted summary statistics for the data used in the first essay and the second appendix include detailed data documentation.

Chapter 2

Accounting for Stratification and Differences in Sampling Rates in a Regional Model of Demand for National Forests in the Southeastern U.S.

Abstract

We estimate regional demand for National Forest settings in the southeastern U.S using weighted and unweighted regression. In estimating these models, we assume that in an unweighted model, the population of interest is the region and is known. However, in a weighted regional model, we assume that the population of interest is unknown and can be approximated closely by using sample weights. Using estimation of demand for National Forests as a case study, we resolve problems relating to inference about the data generating process when different samples are pooled together. In estimating a regional demand model, we show that though efficiency of weighted estimates improves after correcting for heteroskedasticity, they still remain biased because the weights interact with covariates which explain part of model misspecification due to pooling of different populations. In this paper, we show that it is best to use unweighted regression when the coefficient on interactions with the weight variable are jointly insignificant. However, the model needs to be respecified if these interactions are jointly significant but the estimation still proceeds using an unweighted regression.

2.1 Introduction

Many if not all on-site samples are choice-based samples. In a choice-based sample, stratification is on the endogenous variable, directly affecting the kernel of the likelihood. Therefore econometric procedures used in estimation need to account for endogenous stratification

in order to obtain consistent parameter estimates (Manski and McFadden, 1981). This is achieved by deriving appropriate weights for the relevant distribution in a weighted regression. For a count outcome, Shaw (1988) derives weights to correct for the joint effect of truncation and endogenous stratification for poisson distributions, and Englin and Shonkwiler (1995) derive weights for the same correction for negative binomial distributions. However, estimation proceeds in a regular fashion when stratification is on the exogenous variable. In this case, the econometric correction amounts to adding a constant of proportionality which does not affect the kernel of the likelihood. Manski and McFadden(1981) point out that it is important for practitioners to understand that in exogenous stratification, distribution of strata is defined by the domain of exogenous variables. In that case, knowledge about the distribution of exogenous covariates alone is sufficient to know the distribution of strata, even when the distribution of strata affects the choice probability only trivially. Wooldridge (2001) shows that under the assumption of homoskedasticity, econometric procedures do not need to account for exogenous stratification. He further shows that weighted estimates that correct for exogenous stratification are consistent but less efficient than unweighted estimates in a linear specification.

For the purpose of inference about the relevant population, weighted regression is often used in empirical estimation to correct for the differences in sampling rates due to exogenous variables, such as race, age, gender, etc. For example, suppose there are 50 % females in the relevant population and the sample includes only 30 % females. In this case, weights usually derived from the U.S. Census are used to equate the sampling distribution to the population distribution. Even in that case, differences in the weighted and the unweighted results point to some form of model misspecification. Korn and Graubard (1995) give at least two reasons for the weighted estimates to be different from the unweighted estimates when stratification is on the exogenous variable: The model must be very misspecified, or an omitted variable must have a strong interaction with the independent variable and that omitted variable must be highly correlated with the weights. Winship and Radbil (1994) attribute the differences

in the weighted and unweighted results to the pooling of two different samples together. This is particularly relevant when insufficient data leads to the pooling of observations in regional models. Another possible reason for the difference results is that weights interacting with covariates account for the omitted variables in the regression. DuMouchel and Duncan(1983) apply a simple F-test to the latter reason.

The objective of this paper is to provide some insight on the reasons for differences in parameter estimates of weighted and unweighted regression, especially in estimating models where different samples are pooled together. It is shown how to obtain consistent and efficient parameter estimates if that is the case. This information is of relevance to federal agencies such as the USDA Forest Service. The USDA Forest Service conducts on-site samples of recreational visitor use on a regular basis for the purposes of projections and budget allocations. In order to keep survey costs low, regional estimates are of more interest than individual forest estimates. This paper empirically estimates demand for National Forest settings in the southeastern U.S. We also show how inference can be completely erroneous if incorrect specifications of standard errors are used.

This paper is organized as follows. The first section briefly discusses the theory of weighted and unweighted regression. The second section specifies the empirical models of demand for four settings: Day-Use Developed Sites (DUDS), Overnight-Use Developed Sites (OUDS), General Forest Area (GFA) and Wilderness(WILD). The third section explains the results and draws conclusions. The appendix contains four tables of summary statistics for each setting and the last table in the appendix contains weighted means for all settings.

2.2 Data Description

Data for estimating the empirical model specified above were obtained from the National Visitor Use Monitoring Program (NVUM). NVUM began collecting visitor use information for a stratified on-site sample in the year 2000. In its first four-year cycle (2000-2003), NVUM

collected information on the annual number of visits to National Forest for the primary purpose of outdoor recreation, primary activity for an individual, and other socio-economic variables. The ZIP code was used to pull income information from IRS data (which is available according to ZIP code)as a proxy for the income variable. The ZIP code was also used in the calculation of the implicit price variable, Travel Cost. The original master dataset has information on 10 RPA regions and 120 National Forests across the U.S. For further information on adjustments made in the original dataset, refer to Bowker et al (2009).

NVUM is based on a stratified sample technique suggested by English (2002). Every National Forest within the sample is divided into 12 strata according to site type and site usage. Site types or settings include Day Used Developed Sites (DUDS), Overnight Used Developed Sites (OUDS), General Forest Area (GFA) and Wilderness (WILD). Site usage includes Low (L), Medium (M) and, High (H). Random samples are drawn from each stratum.

Data for the Southeastern U.S., or U.S. Forest Service Region 8, is used for analysis. The data is collected for 14 National Forests including the Chattahoochee-Oconee National Forest, George Washington-Jefferson National Forest, Croatan National Forest, Daniel Boone National Forest, Cherokee National Forest, Francis Marion National Forest, Conecuh National Forest, Ozark National Forest, Apalachicola National Forest, DeSoto National Forest, Ouachita National Forest, Bienville National Forest, Kisatchie National Forest, Davy Crockett National Forest, and Land Between The Lakes National Forest. The NVUM survey sampled 25% of total National Forests in its 2000 cycle and 20% in its 2004 cycle. The dataset for the southeastern region included 7473 sample observations.

2.3 Theoretical Model

When different samples are pooled together, estimation can proceed using pseudolikelihood, first used by Besag (1975; 1977). A pseudolikelihood estimation is based on the assumption that each random process is independent. In the case of regional demand models, demand for various samples across the region are indeed independent of each other.

We briefly explain the methodology below from Wang et al. (2004).

let,

$$X = (X_1, X_2, ..., X_n) \tag{2.1}$$

be random variables with probability density functions

$$f_1, f_2, ..., f_n \tag{2.2}$$

The density of interest is

$$f(., \theta), \theta \in \Theta \tag{2.3}$$

of a study variable X. At least in some qualitative sense, the

$$f_1, f_2, ..., f_n$$

is thought to be like

$$f(., \theta)$$

We assume that each independent distribution is related to the distribution of interest through relevant weights. Pseudolikelihood or what is popularly known as Power likelihood is therefore given by,

$$\prod_{j=1}^{m} \prod_{i=1}^{nj} f^{\lambda_j}(x_{ij}, \theta) \tag{2.4}$$

where, $j = 1, 2, ...m$ are the number of independent random samples, and $i = 1, 2, ...nj$ are the number of individuals in each sample. Therefore the concept of pseudolikelihood is used to estimate the parameter of interest. It is important to understand that the weights, though constructed based on exogenous variables, do not enter the likelihood as a constant of proportionality. Therefore weights in this case affect the kernel of the likelihood.

In our model, it is assumed that the data generating process follows a negative binomial distribution to correct for endogenous stratification and truncation.The log-likelihood for a

negative binomial distribution, accounting for truncation and endogenous stratification, is given by,

$$log(y)+log\Gamma(y+\alpha^{-1})+ylog(\alpha)+(y-1)(x\beta)-(y+\alpha^{-1})log(1+\alpha exp(x\beta))-log\Gamma(\alpha^{-1}) \quad (2.5)$$

and the expected mean is given by,

$$E(y \mid x) = EXP(X\beta) + 1 + \alpha EXP(X\beta) \quad (2.6)$$

The score function is given by,

$$\frac{dlogL}{d\beta} = \sum_i (y_i - 1)X_i - (y_i + \alpha^{-1})\frac{\alpha X_i EXP(X_i\beta)}{1 + \alpha X_i EXP(X_i\beta)} \quad (2.7)$$

The information matrix is given by the inverse of the second derivative,

$$\frac{dlogL}{d\beta'\beta} = \sum_i \frac{-(y_i + \alpha^{-1})\alpha X_i' X_i EXP(X_i\beta)}{(1 + \alpha X_i EXP(X_i\beta))'(1 + \alpha X_i EXP(X_i\beta))} \quad (2.8)$$

The score function for weighted regression is given by,

$$\frac{dlogL}{d\beta} = \sum_j \lambda_j \sum_i (y_{ij} - 1)X_{ij} - (y_{ij} + \alpha^{-1})\frac{\alpha X_{ij} EXP(X_{ij}\beta)}{1 + \alpha X_{ij} EXP(X_{ij}\beta)} \quad (2.9)$$

If we make an assumption of a pseudolikelihood.

The information matrix is given by the inverse of the second derivative,

$$\frac{dlogL}{d\beta'\beta} = \sum_j \lambda_j \sum_i \frac{-(y_{ij} + \alpha^{-1})\alpha X_{ij}' X_{ij} EXP(X_{ij}\beta)}{(1 + \alpha X_{ij} EXP(X_{ij}\beta))'(1 + \alpha X_{ij} EXP(X_{ij}\beta))} \quad (2.10)$$

2.4 Empirical Model

In a stratified sample, sampling weights are used to expand each individual to be representative of the proper population. It is given by,

$$\frac{N_j}{n_j}$$

where, N_j are the number of individuals in stratum j in the population and n_j are the number of individuals sampled in stratum j. In many cases the numerator is known. But in cases where it is not known, the numerator must be estimated. In the case of NVUM,N_j is not observed directly and is estimated by,

$$NVEXPAND_{jk} = ExitingTraffic_{jk} * PropLastExit_{jk} \tag{2.11}$$

where ExitingTraffic$_{jk}$ is the average exiting traffic count per day for the stratum, and PropLastExit$_{jk}$ is the ratio of last exiting recreation vehicles to total count of vehicles.
The NVUM survey sample collects sufficient data to allow computation of weights. Its computation is based on the proportion last exited visitors in a given stratum j in a forest k. These weights are used in weighted regression. For further information on NVUM survey samples refer to Appendix B in Bowker et al(2009).

Empirical Model Specification

Visits to a National Forest are modeled as a truncated negative binomial model correcting for endogenous stratification. Both weighted and unweighted regional demand models for settings are estimated using the following empirical specification:

$$\begin{aligned} NFV12MO^i_j &= f(PEOP^i_j, INCE^i_j, GEND^i_j, AGE^i_j, TC^i_{jk}, HF^i_j \\ &= OSITE^i_j, OVERNT^i_j, ECO_k, SUPPLYVAR_{jk}) \end{aligned} \tag{2.12}$$

where,

$$i = 1, 2, ...N$$

are the number of individuals

$$j = 1, 2, ...4$$

are the strata in the sample

$$k = 1, 2, ...13$$

are the number of forests in southeastern U.S.

The dependent variable is the number of annual recreation visits to a National Forest per

vehicle group. Demand for visits is a function of seven variables: own price (TC) [1] number of people in the vehicle (PEOPVEH), annual income (INCOME), gender (GENDER1), age (AGE), an indicator for staying overnight (ONITE), an indicator if an individual visited any other site (OSITES), a dummy variable if a forest belongs to a subtropical ecoregion (SUBTROP), a dummy variable if a forest belongs to a hot continental ecoregion (HOT-CONT) and a dummy variable if a forest belongs to a mountain ecoregion (MOUNTAIN). In the model we drop the dummy variable for a subtropical ecoregion. An additional term, HF, has been incorporated to capture the differences between high and low frequency users, where HF=1 if the number of annual visits was greater than 15, else $HF = 0$. The supply variables for the General Forest Area setting include the percentage of forest area within a radius of 100 miles of a visitor's origin (FORESTP), and miles of trails in a National Forest as a proxy for access to general forest areas (TRAILS). Supply variables for the Overnight Use Developed settings include the total number of tent camping sites in a National Forest (TENTC), and the total number of establishments in recreation and vacation camps within 100 miles of origin (SUMCAMPS). Supply variables in Day Use Developed setting include total number of recreation areas in a National Forest with picnic tables as a proxy for total number of day use sites (PICNICTAB), total number of recreation areas in a National Forest with swimming areas as a proxy for high-attraction day use sites (SWIMMING) and total number of establishments in nature parks and similar institutions within 50 miles of a visitor's origin as a proxy for private day use sites (SUMNATPARK). Supply variables for the Wilderness setting include acres of designated wilderness area in a given National Forest (DESIG).

[1] Bowker et al. (2009),

$$TC = 2 * (0.12 * PRACDIS) + 2 * (0.33 * \frac{INCE}{2000} * PRACTIME)$$

where PRACDIS is the one way distance to the site. A per mile cost of 0.12 dollar was used. Income forgone is calculated as one third of the wage rate, where wage rate is calculated as the proxy of annual income divided by 2000 work hours. PRACTIME is the time spent at the site.

2.5 Results

Tables 1 through 4 include results for the settings of GFA, DUDS, OUDS and WILD. The first row gives the coefficient for unweighted[2] and weighted[3] models referred to as Model 1 and Model 2, respectively. The second row gives the standard errors computed using the Newton-Raphson algorithm assuming homoskedasticity, and the third row includes White's standard errors corrected for heteroskedasticity. The purpose of including heteroskedasticity corrected standard errors is to show that though in the unweighted regression, the assumption of homoskedasticity can be maintained,in the case of weighted regression, however the same cannot be assumed as claimed by Winship and Radbill(1994). This is because covariates in the weighted regression become correlated with the error term. It is therefore important to correct weighted standard errors for heteroskedasticity. We will explain this later when we discuss our results in Table 5 .

Results in Table 2.1 show that in explaining demand for trips to General Forest Area settings, the standard errors in Model 1 with heteroskedasticity correction are bigger but do not change inference in terms of significance of the coefficients. Such is not the case with unweighted regression. The reason for a change in significance of coefficients is twofold. Not only are the heteroskedasticity corrected standard errors significantly different but the coefficient estimates become inconsistent due to significant interactions of some important variables with the weights. This can be seen from Table 5. These interacting variables include dummies for hot continental and mountain ecoregions, the income variable, a dummy for overnight stay, and the trails supply variable. Differences in signs of weighted and unweighted models can be attributed to inconsistencies in the weighted model.

[2]Negative binomial correcting for truncation and endogenous stratification

[3]Negative binomial correcting for truncation and endogenous stratification and weighted to account for differences in sampling rates using NVEXPAND as the weight

Table 2.1: Estimation Results of Outdoor Recreation for GFA Settings in the Southeastern U.S. (2000-2003)

	Model 1[a]	Model 2[b]
Intercept	1.282	0.968
	(0.130)*	(0.001)*
	(.160)*	(.258)*
HOTCON	0.197	-0.077
	(.0006)*	(.006)*
	(.058)*	(.093)
MOUNTN	0.113	0.131
	(.0006)***	(.0006)*
	(.057)**	(.082)
FOREST	0.003	0.007
	(.002)	(0.00002)*
	(.002)	(.004)***
TRAIL	-0.0004	0.0001
	(.0001)*	(0.000001)*
	(.0001)*	(.0001)
INCE	-0.000009	-0.000004
	(0.00002)	(0.00000003)*
	(.000003)*	(.000004)
AGE	0.002	0.003
	(0.001)	(0.00001)*
	(.001)	(.0021)***
GENDER	-0.164	-0.113
	(0.046)*	(.0005)*
	(.053)*	(.101)
PEOPVEH	-0.027	-0.042
	(0.013)**	(.0001)*
	(.013)**	(.021)**
OSITE	-0.067	-0.026
	(0.040)***	(0.0005)*
	(.042)***	(.071)
OVERNTE	0.039	0.193
	(0.041)	(0.0005)*
	(.042)	(.063)*
TC	-0.003	-0.003
	(0.0003)*	(0.000003)*
	(.0004)*	(.0007)*
HF	1.816	1.707
	(0.034)*	(0.0003)*
	(.031)*	(.050)*
ALPHA	0.561	0.425
	(0.047)*	(0.0003)*
	(.046)*	(.054)*
NOBS	1979	
LOGL	54920.7	480453000
BIC	-54867.5	-480453000

[a]Unweighted model with coefficient, standard errors without heteroskedasticity correction, and heteroskedasticity robust standard errors reported in the first, second, and third rows respectively, where *, **, and *** represent statistical significance at 1%, 5% and 10% level respectively.

[b]Weighted Model

Table 2.2: Estimation Results for Day Use Settings in the Southeastern U.S. (2000-2003)

	Model 1[a]	Model 2[b]
Intercept	0.673	-0.585
	(0.176)*	(0.007)*
	(.237)*	(.846)
HOTCONT	0.095	0.353
	(.073)	(0.002)*
	(.069)	(.125)*
MOUNTAIN	-0.257	0.178
	(.105)**	(.003)*
	(.103)**	(.201)
PICNICTAB	0.001	-.0004
	(.00008)	(0.00002)*
	(.00008)	(.0002)**
NATPARK	.004	0.0001
	(.003)	(0.000005)*
	(.004)	(.007)
SWIMMING	-0.019	-.066
	(.014)	(.0003)*
	(.014)	(.025)*
INCE	-0.0001	-0.000009
	(0.00002)*	(0.00000007)*
	(.000004)*	(.0000056)***
AGE	-0.003	0.005
	(0.001)**	(0.00003)*
	(.002)***	(.003)***
GENDER	-0.012	-0.039
	(0.047)	(.001)*
	(.050)	(.096)
PEOPVEH	-0.053	-0.039
	(0.014)*	(.0003)*
	(.0144)*	(.0308)
OSITE	-0.259	-0.017
	(0.0409)*	(0.001)*
	(.055)*	(.111)
OVERNTE	-0.193	-0.439
	(0.073)*	(0.002)*
	(.078)*	(.154)*
TC	-0.003	-0.003
	(0.0003)*	(0.000005)*
	(.001)*	(.001)*
HF	2.217	2.248
	(0.050)*	(0.001)*
	(.037213)*	(.072)*
ALPHA	2.592	6.466
	(0.476)*	(0.047)*
	(.652)*	(5.557)
NOBS	2394	
LOGL	36023.8	87280200

[a]Unweighted model with coefficient, standard errors without heteroskedasticity correction, and heteroskedasticity robust standard errors reported in the first, second, and third rows respectively

[b]Weighted Model

Results in Table 2.2 show that for the weighted regression for Day Use Developed sites, the intercept and the dispersion parameters both become insignificant. An insignificant dispersion parameter points to the failure of an important theoretical assumption of the model, i.e. the difference in mean and variance of the population. This points to inconsistencies of parameter estimates of weighted regression.

Results in Table 2.3 show that in explaining demand for trips to Overnight Use Developed sites, six of the variables become insignificant in the weighted model: age, gender, number of people in the vehicle, dummy for hotcontinental and mountain ecoregion, and number of camping sites.

Table 2.4 gives the coefficient and standard errors for the wilderness model. In Table 5 only the intercept and income variables have significant interactions with the weight variable, and the interactions with the other covariates of the model are insignificant. In these results, unlike the previous models the signs for weighted and unweighted models stay the same.

In Table 2.5, we have only included the covariates interacted with weights since the remaining coefficient remains the same in the unweighted regression. Table 2.5 shows that weights that are constructed to provide correction for the differences in sample rates have strong interactions with covariates included in the model. We have also conducted the likelihood ratio test for all four settings. In the constrained model for all settings, coefficients on interactions terms with the weight are zero. For general forest area and day use developed sites model, the chi-square statistic is given by 38.2 and 39.6 respectively with 13 degrees of freedom. The p values from the chi-square table are given by 34.53, 27.69 and 22.36 for 1%, 5%, and 10% significance levels respectively. So we reject the null hypothesis that the coefficients of interactions with the weight variable are zero. Therefore, weights interact with covariates to partially or fully explain the variables omitted from the model and the model needs to be respecified. On the other hand, for the wilderness and overnight use developed

Table 2.3: Estimation Results of Outdoor Recreation for Overnight Use Developed in the Southeastern U.S. (2000-2003)

	Model 1[a]	Model 2[b]
Intercept	0.630	1.503
	(0.165)*	(0.003)*
	(.194)*	(.330)*
HOTCONT	-0.370	0.271
	(.118)*	(0.002)*
	(.116)*	(.209)
MOUNTAIN	-0.234	-0.162
	(.081)*	(.002)*
	(.083)*	(.200)
TENTC	0.0003	0.0001
	(.0001)**	(0.00002)*
	(.0001)**	(.0002)
SUMCAMPS	.001	0.0005
	(.001)	(0.00002)*
	(.001)	(.002)
INCE	-0.0001	-0.00003
	(0.000003)*	(0.00000008)*
	(.000003)*	(.00001)*
AGE	0.005	0.004
	(0.002)*	(0.00004)*
	(.002)*	(.003)
GENDER	-0.106	-0.115
	(0.052)**	(.001)*
	(.0589)**	(.119)
PEOPVEH	-0.043	-0.010
	(0.017)*	(.0004)*
	(.0167)*	(.036)
OSITE	-0.196	-0.245
	(0.050)*	(0.001)*
	(.0520)*	(.101)*
OVERNTE	0.009	-0.305
	(0.049)	(0.001)*
	(.050)	(.097)*
TC	-0.003	-0.002
	(0.0004)*	(0.00001)*
	(.001)*	(.001)*
HF	2.188	1.885
	(0.059)*	(0.001)*
	(.0412)*	(.127)*
ALPHA	1.461	.407
	(0.215)*	(0.001)*
	(.240)*	(.096)*
NOBS	1707	
LOGL	18949.7	31155200
BIC	-18897.6	-31155200

[a]Unweighted model with coefficient, standard errors without heteroskedasticity correction, and heteroskedasticity robust standard errors reported in the first, second, and third rows respectively, where *, **, and *** represent statistical significance at 1%, 5% and 10% level respectively.

[b]Weighted Model

Table 2.4: Estimation Results of Outdoor Recreation for Wilderness Settings in the Southeastern U.S. (2000-2003)

	Model 1[a]	Model 2[b]
Intercept	0.481	2.113
	(0.446)	(0.009)*
	(.526)	(.430)*
HOTCONT	0.521	0.374
	(.102)*	(0.004)*
	(.103)*	(.200)***
MOUNTAIN	0.110	-0.258
	(.200)	(.006)*
	(.190)	(.335)
DESIGW	0.000005	0.000007
	(.000003)	(0.0000001)*
	(.000003)	(.000005)
SUMWILDERN	.007	0.007
	(.004)	(0.00007)*
	(.002)*	(.002)*
INCE	-0.00002	-0.00003
	(0.000004)*	(0.0000002)*
	(.000007)**	(.00001)*
AGE	-0.004	-0.002
	(0.004)	(0.0001)*
	(.004)	(.006)
GENDER	-0.077	-0.362
	(0.101)	(.003)*
	(.100)	(.172)**
PEOPVEH	-0.068	-0.075
	(0.038)***	(.001)*
	(.035)**	(.060)
OSITE	-0.325	-0.161
	(0.105)***	(0.004)*
	(.099)*	(.166)
OVERNTE	-0.192	-0.545
	(0.112)***	(0.004)*
	(.114)***	(.173)*
TC	-0.003	-0.004
	(0.0004)*	(0.00002)*
	(.001)*	(.001)*
HF	2.092	2.337
	(0.145)*	(0.004)*
	(.001)*	(.225)*
ALPHA	3.711	0.873
	(1.735)**	(0.006)*
	(1.869)**	(.381)**
NOBS	618	
LOGL	4059.37	7875890
BIC	-4014.39	-7875890

[a]Unweighted model with coefficient, standard errors without heteroskedasticity correction, and heteroskedasticity robust standard errors reported in the first, second, and third rows respectively, where *, **, and *** represent statistical significance at 1%, 5% and 10% levels respectively.

[b]Weighted Model

Table 2.5: Estimation Results for Weights Interaction Terms for all Four Settings

	GFA	DUDS	OUDS	WILD
INTERCEPT	-.146E-04*	-.952E-04**	.315E-04	.572490E-03*
HOTCONT	-.901E-05*	.281E-04	.743E-04*	.228631E-04
MONUNTAIN	.779E-05***	.766E-04**	.794E-05	-.257495E-04
INCE	.715E-09 *	.414E-09	-.217E-08*	-.156295E-07*
AGE	.104E-06	.143E-05*	.370E-06	.224673E-05
GENDER	.200E-06	-.821E-05	-.213E-04***	.224673E-05
PEOPVEH	-.161E-05	.275E-05	.482E-06	-.341661E-04
OSITE	.670E-05	.218E-04***	-.127E-04	.220476E-04
OVERNTE	.127E-04**	-.687E-04	-.353E-04*	-.938949E-04
TC	-.404E-07	.893E-07	.108E-06	.177425E-07
HF	-.571E-05**	.535E-05	-.197E-04	.416311E-04
FORESTP	-.490619E-07	-	-	-
TRAILS	.203004E-07***	-	-	-
PICNICTAB	-	.203004E-07	-	-
SUMNATPARK	-	.159524E-05	-	-
SWIMMIMG	-	-.734884E-05***	-	-
TENTC	-	-	-.228780E-07	-
SUMCAMPS	-	-	-.323342E-06	-
SUMWILD	-	-	-	.235090E-05
DESIGW	-	-	-	-.390465E-08

sites models, the chi-square statistic are given by 14.64 and 21.4 respectively with 13 degrees of freedom. The p values from the chi-square table are given by 34.53, 27.69 and 22.36 for 1%, 5%, and 10% significance level respectively. So we fail to reject the null hypothesis that the coefficients of interactions with the weight variable are zero. Therefore, the model is correctly specified.

2.6 Conclusions and Implications

Insufficient data from each National Forest in the NVUM sample necessitate the pooling of observations for forests in the same region. This encourages analysts to use weighted regression to equate the sampling distribution with the population distribution for the purpose of

inference about the relevant population which in this case is National Forest recreation visitors. However, differences in the coefficient estimates of weighted and unweighted regressions point to model misspecification resulting from the pooling of different samples. This can be seen from the significant interactions of weights with the covariates included in the model. Correcting standard errors for heteroskedasticity increases the efficiency of the estimates but they are still biased. Therefore, the model needs to be respecified by identifying missing variables and interactions in an unweighted regression if the coefficients on interaction terms are large and jointly significant. The reasons for this (Winship and Radbill, 1994) could be that the weight variable is a function of covariates omitted from the model or interactions of covariates with the weight variable act as a proxy for interactions with the covariates omitted from the model. However, if the coefficients on interaction terms are jointly insignificant, the model does not need to be respecified.

Chapter 3

Modeling Site Specific Heterogeneity in an On-Site Stratified Random Sample of Recreational Demand

Abstract

Using estimation of demand for the George Washington/Jefferson National Forest as a case study, it is shown that in a stratified/clustered on-site sample, latent heterogeneity needs to be accounted for twice: first to account for dispersion in the data caused by unobservability of the process that results in low and high frequency visitors in the population, and second to capture unobservable heterogeneity among individuals surveyed at different sites according to a stratified random sample (site specific effects). It is shown that both of the parameters capturing latent heterogeneity are statistically significant. It is therefore claimed in this paper, that the model accounting for site-specific effects is superior to the model without such effects. Goodness of fit statistics show that our empirical model is superior to models that do not account for latent heterogeneity for the second time. Though the price coefficient for the travel cost variable remains the same, the expected mean changes across different models. This information is of importance to the USDA Forest Service for the purpose of projections for budget allocation and resource utilization.

3.1 Introduction

In order to reduce survey costs, on-site survey samples are either clustered or stratified. Random samples are drawn within these clusters to make inferences about the relevant populations. According to Cameron and Trivedi (1986), survey data are usually dependent. This may be due to the use of cluster samples to reduce survey costs. In such cases the data

may be correlated within a cluster owing to a presence of a common unobserved cluster-specific term. According to Pepper (2002), whenever a group of sample observations share a common factor, any theoretical and empirical analysis not accounting for clustering effects would give inconsistent parameter estimates. This points to the need to account for cluster-specific effects in the modeling data generated from on-site samples, where individuals are surveyed at various sites in a given stratum across the National Forest.

In NVUM surveys, individuals are sampled at various sites within a National Forest which are stratified according to site type and site use. A group of individuals surveyed at a particular site share common factors, the observed and unobserved site specific attributes. For example, individuals surveyed at a fishing site have a common recreational use-value for fishing. Statistically, there is a strong reason to believe that individuals intercepted at the same site are somehow correlated rather than independent. According to Galwey (2006), the relationship of the outcome variable, which is visits to a recreation site, may be perfectly replicated for each site, but most likely there will be some differences in this relationship. These differences, or between-site variations, could be ascribed to chance or to some observed or unobserved characteristics or attributes. Therefore to capture the within-site correlation, it is important to introduce site-specific heterogeneity.

Count outcomes are modeled as discrete outcomes and not continuous quantities using a poisson or a negative binomial distribution. The latter is a more flexible and reasonable assumption for empirical data because it drops the assumption of equidispersion. A negative binomial distribution is derived by introducing heterogeneity resulting from unobserved individual taste and preference. Greene (2005) points out that heterogeneity can be introduced the second time if a negative binomial is the base model. We exploit this idea to introduce heterogeneity for the second time. But unlike Greene, we introduce site-specific heterogeneity instead of individual-specific heterogeneity, to explain correlation among individuals sampled at the same site.

Introducing heterogeneity in a poisson model to derive a negative binomial distribution causes heteroskedasticity in estimation of standard errors. Espiñeira and Tuffour(2008) use a more flexible specification for the overdispersion parameter to correct for heteroskedasticity in modeling demand for Gros Morne National Park. They make the overdispersion parameter a function of individual characteristics and show that doing so improves the goodness of fit. Greene (2005) also recommends this specification to correct for heteroskedasticity.

In this paper it has been hypothesized that in a stratified on-site sample, there is a strong reason to believe that individuals sampled at the same site are correlated rather than independent. The hypothesis is tested by modeling demand for outdoor recreation at the George Washington/Jefferson National Forest, where individuals are sampled at 88 sites clustered under four settings types. It is shown that the site-specific effects are significant and there is a strong theoretical and empirical reason to introduce such site-specific effects. By estimating design effects, it is shown that the asymptotic standard errors for the travel cost variable are significantly different under the assumption of clustered sampling rather than random sampling. It is also shown that the expected mean estimates, which are often used for the purpose of projections, is significantly different in each model and so is the estimate of the overdispersion parameter.

This paper is organized as follows. In the second section we give details about the data used for our analysis. In the third section we explain our theoretical model. In the fourth section we specify our empirical model and summary statistics. In the fifth section we estimate six models: a poisson model accounting for stratification and truncation(TSP); a negative binomial model accounting for stratification and truncation(TNB); A poisson model accounting for truncation, stratification and site-specific effects(TSP2); a negative binomial model accounting for stratification, truncation, and an overdispersion parameter to vary by individual characteristics(TNB1); a negative binomial accounting for stratification, truncation, and site-specific effects(TNB2); and finally a negative binomial model accounting

for stratification, truncation, and accounting for site-specific effects and an overdispersion parameter to vary by individual characteristics(TNB3). Conclusions are presented at the end of the chapter.

3.2 Data

The empirical model will be estimated using NVUM data collected for the George Washington/Jefferson National Forest in the southeastern region of the U.S. The NVUM was conducted at 88 sites stratified by settings within the National Forest. The settings include Wilderness (WILD), Day Use Developed Sites (DUDS), Overnight Use Developed Sites (OUDS) and General Forest Area (GFA). There are 781 sample observations. The data was collected for four sample years, 2000-2003. More detail was provided on NVUM in the previous chapter. For our analysis, we have only included observations for which recreational trips to the National Forest are less than 52. Following Bowker et al. (2009) we also deleted observations with travel cost greater than 720 and people in the vehicle greater than 10.

3.3 Theoretical Model

According to Haab and McConell (1996),

> "estimation of single site demand models begins with an assessment of the data generating process which is governed by the assumed stochastic structure of the demand functions and the sampling procedure."

In this chapter we discuss modeling the stochastic structure of the demand functions. The stochastic structure of demand depends on whether the dependent variable, which is an individual's trips to a site, is assumed to be distributed continuously or as a count variable. For the travel cost model the dependent variable is a count variable. Count data for number of visits to a recreational site is not available in continuous quantities.Under this scenario poisson distribution results in an asymptotic outcome, according to Hellerstein (1996). This

is because a binomial distribution approaches a poisson distribution as the number of draws approaches infinity. However, when the dependent variable is a count outcome, equidispersion of data is rarely a realistic empirical assumption. A negative binomial distribution is statistically derived by introducing an unobserved individual specific effect in the poisson distribution. The effect is random and each effect is independent of each other and follows a gamma distribution with a dispersion parameter.

Unobserved individual effects are consistent with utility theory. These unobserved effects are attributed to an individual's taste and preferences which are known by the individual but are unobserved by the analyst. One common phenomenon with any travel cost study is that the high frequency visitors who live close to the site make numerous low cost visits, whereas the low frequency visitors who live far away from the site make a few high cost visits. Combining high frequency and low frequency visitors does not account for differences in these individuals, leading to observed over-dispersion in the data. Therefore, we have reasons to believe that the base model for travel cost is a negative binomial with the introduction of unobserved individual-specific effects in the poisson model. We use a negative binomial with a quadratic variance function (NB2) as our base model which is a good approximation in many empirical situations. Also, maximum likelihood estimation of NB2 is robust to misspecification of the conditional mean (Cameron and Trivedi, 1986).

In this chapter, it has been hypothesized that there are reasons to believe that the stochastic process includes unobserved site-specific effects which account for the differences across various sites where the on-site sampling is conducted. Therefore, according to our hypothesis, unobserved effects are introduced in the model. But these are not individual specific unobserved effects but site-specific effects. In the previous chapter issues of weighting to control for choice-based survey design were discussed. In this chapter limitations of the independence assumption in survey data is discussed and econometric techniques are suggested to correct for such limitations.

In microeconometrics, an individual's choice between various sites is treated as a separate estimation equation, logit or nested logit. Because it is conditional on choice, the dependent variable is estimated as a count process. Various applications include site-specific effects in the choice equation. However, the sampled site data for all the sites is extremely costly and in most cases is not available. In this case it becomes even more important to introduce site-specific effects in the count equation. This model can be used to estimate demand for a given National Forest where a random sample is selected at various sites within a setting. When non-negativity, stratification and truncation are included this model would also account for correlation in the variance parameter among various individuals going to the same settings.

The random negative binomial model (RNBM) used by Greene (2005) in a panel data setting is generalized to a cluster of sites in the George Washington/Jefferson National Forest to capture intra-cluster correlation in the variance-covariance matrix. Greene (2005) shows that heterogeneity can be introduced twice if a negative binomial is the base model. A random model is chosen over a fixed effect model to capture the intra-cluster correlation which implies from relaxing the independence assumption within a given cluster.
The log-likelihood of poisson correcting for truncation and stratification is given by (TSP),

$$logl = (y_{ij} - 1)(X'_{ij}\beta) - Exp(Xij'\beta) - log(\Gamma(y_{ij})) \tag{3.1}$$

and expected mean is given by,

$$E(y \mid x) = EXP(x\beta) + 1 \tag{3.2}$$

Site-specific effects are added in the mean statement, to derive the poisson distribution model correcting for truncation and stratification with site-specific effects(TSP2). In recreational demand models these site-specific effects could be attributes about a particular site which are unobserved. According to Murdock (2006),

> "one obvious way to address unobserved heterogeneity is to simply include a full set of alternative specific constants. The proposed approach will be useful when

there are important characteristics that only vary across recreation locations and not also across time or individuals."

He mentions such site characteristics for fishing such as regulations, water quality, fish consumption advisories, physical characteristics, adjacent land use, and the presence of facilities.

$$Xij'\beta + \sigma b_j \tag{3.3}$$

where,

$$b_j \sim N(0,1) \tag{3.4}$$

The negative binomial correcting for truncation and endogenous stratification can be derived by introducing individual-specific heterogeneity which follows a one parameter gamma distribution (TNB),

$$log(y_{ij})+log\Gamma(yij+\alpha^{-1})+y_{ij}log(\alpha)+(y_{ij}-1)(x'ij\beta)-(y_{ij}+\alpha^{-1})log(1+\alpha exp(x'_{ij}\beta))-log\Gamma(\alpha^{-1}) \tag{3.5}$$

and the expected mean is given by,

$$E(y \mid x) = EXP(x\beta) + 1 + \alpha EXP(x\beta) \tag{3.6}$$

Subject-specific effects in 3.6 are similar to 3.3 and 3.4.

The overdispersion parameter in the TNB1 and TNB3 models is specified as,

$$\alpha = \frac{1}{exp(Z'_i\gamma)} \tag{3.7}$$

The nlmixed procedure in SAS is used to maximize the unconditional likelihood given by,

$$Prob[Y = y_{ij}|x_{ij}] = \int_{b_j} Prob[Y = y_{ij}|x_{ij}, b_j]f(b_j)db_j \tag{3.8}$$

where,

$$Prob[Y = y_{ij}|x_{ij}, b_j]$$

is given by 3.5, and $f(b_j)$ is given by 3.4,

3.4 Empirical Model

The empirical model is specified as,

$$NFV12MO_j^i = f(INCE^i, AGE^i, PEOPVEH^i, GENDER^i, TC^i, HF^i, OVERNTE^i) \tag{3.9}$$

where,

$$i = 1, 2, ...N$$

are the number of individuals

$$j = 1, 2, ...88$$

are the number of sites in the sample The dependent variable in the empirical model is the number of annual recreation visits to the George Washington/Jefferson National Forest per group. Demand for visits is a function of six variables: own price or cost of the trip (TC), number of people in the vehicle (PEOPVEH), annual income (INCOME), gender (GENDER1), age (AGE), and an indicator for staying overnight (ONITE). An additional term has been incorporated to capture the differences between high and low frequency users (HF), where HF=1 if number of annual visits was greater than 15, and otherwise $HF = 0$. Site-specific effects are included in the mean statement additively.

The following example illustrates the motivation of the introduction of site-specific effects to capture the correlation between individuals sampled at the same site. In the Chattahoochee National Forest, Brasstown Bald is a popular visitor attraction. Rising 4,784 feet above sea level, Georgia's highest mountain allows clear views of four southern states, Georgia, Tennessee, North Carolina and South Carolina. This site has four hiking trails:

Table 3.1: Summary Statistics for George Washington/Jefferson National Forest NVUM Data, 2000-2003

	Mean1	Min	Max
INCE[a]	23964.61	11618.19	111898.27
AGE[b]	41.541	18	75
GENDER[c]	0.216	0	1
PEOPVEH[d]	2.549	1	9
OVERNTE[e]	0.235	0	1
TC[f]	49.481	0.469	1103.84
HF[g]	0.243	0	1
NFV12MO1[h]	11.282	1	51
STYPE1[i]	0.146	1	0
STYPE2[j]	0.200	1	0
STYPE3[k]	0.366	1	0
STYPE4[l]	0.288	1	0
NOBS	781		

[a]IRS reported average after tax income for an individual's ZIP Code
[b]Age
[c]A dummy for Gender equals 1 if female
[d]No.of people in the vehicle
[e]Indicator for overnight stay
[f]As a function of one way travel distance and income foregone (Refer to footnote 1, Chapter2)
[g]HF=1 if no. of annual visits greater than 15, else $HF = 0$
[h]Number of annual recreation visits per group
[i]An indicator for Wilderness visits
[j]An indicator for Day Used Developed Site visits
[k]An indicator for Overnight Used Developed Site visits
[l]An indicator for General Forest Area visits

Brasstown Bald Trail, the Arkaquah Foot Trail, Jack Knob Foot Trail, and Wagon Train Foot Trail. The observatory also provides facilities for picnicking and nature viewing. The view from the 4,784 feet peak is a popular attraction and most visits to the site are of short duration and usually involve nature viewing and relaxing as the primary activities.

The 5.5 mile long Arkaquah Foot Trail near the observatory is a wilderness trail that attracts a wide variety of hikers and nature viewers. The duration of visits to this trail is usually longer than the duration of visits to the observatory and the site draws both locals and non-locals. The trailhead connects to Track Rock Gap, one of the best known of the petroglyph, or marked stone sites, in Georgia. The Jack Knob Foot Trail is about 4.5 miles and leads to the famous Appalachian Trail. The Wagon Train Foot Trail is 5.8 miles and leads to the Wagon Train Road which ends at Young Harris College. The trail is traditionally hiked by graduating students and their families, the evening before graduation. Thus, it mostly draws locals for a short duration of time.

The above example suggests dependence between individuals surveyed at a particular site due to some observed or unobserved site-specific effects. In a survey sample of this nature, random sampling might not be the most reasonable assumption about the data. Dependence between individuals visiting the same site to estimate the demand for a single National Forest is modeled. It is most likely that individuals surveyed at a given site are correlated rather than independent. The above argument is used to motivate a mixture model where site-specific random effects follow a standard normal distribution. In this model, the overdispersion parameter is modeled as,

$$\alpha = f(INCOME, AGE, OVERNT, STYPE) \tag{3.10}$$

where,

Site types (STYPE) or settings is a dummy variable for each settings type. Settings include Day Used Developed Sites (DUDS), Overnight Used Developed Sites (OUDS), General Forest

Area (GFA) and Wilderness (WILD). For estimation, the dummy for Overnight Used Developed Sites is dropped. Thus, the OUDS setting serves as base.

3.5 Results

Similar to Pepper (2002), design effects for the variables are constructed in the mean statement for two models, TNB1 and TNB3. Design effects are defined as the ratio of asymptotic variance under the assumption of random sampling to asymptotic variance under the assumption of clustered sampling. The sampling scheme has negligible effects on the asymptotic variance for most of the variables. However, for the indicator variable for an overnight stay, the design effect is 1.104, implying almost a 5 % difference. For the travel cost variable,the design effect is huge, around 1.667, implying that the estimated standard error in the clustered sample exceeds that of random sample by 29%. Also, the sample size of our data is fairly small and number of clusters are fairly large(88 sites) where the survey was conducted. The design effects tend to grow as more observations are made within a cluster.

Also, in the TSP2, TNB2, and TNB3 models, significant site-specific effects are found. This can be seen from the significance of the variance parameter, given by sigma in the results. The parameter estimates are 0.475, 0.373, and 0.317 for the TSP2, TNB2, and TNB3 models respectively, each significant at the 1% significance.

Model comparison shows that the TSP2 model which accounts for site-specific effects performs better than the TSP model without site-specific effects. The log-likelihood and BIC model fitness criterion for TSP model are -2327.53 and 2353.68 respectively and for the TSP2 model, they are -2263.5 and 2283.8 respectively. The log-likelihood for the TSP2 model is higher than the TSP model. In terms of the BIC criterion, smaller values of BIC suggest a better model. Therefore, clearly the TSP2 model accounting for site-specific effects performs better.

Table 3.2: Design Effects for TNB1 and TNB3 model

TNB1[1]	TNB3[2]
Intercept	1.022
Income	0.1
Age	1
Overnight	1.104
TC	1.667
HF	0.922

In comparing four negative binomial models, the simple negative binomial model accounting for stratification and truncation (TNB) and the negative binomial model (TNB2) additionally accounting for site-specific effects give more or less the same results, with log-likelihoods of 13765.7 and 13768.5 respectively. Parameters and standard errors are slightly different but not enough to change the statistical significance or therefore the inference.

Now we compare the two negative binomial models including overdispersion as a function of individual characteristics: TNB1 does not account for site-specific effects, while TNB3 does. The log-likelihood for TNB3 is slightly higher than TNB1 (13781.5 and 13780.5 respectively). TNB3 also does better than TNB1 in terms of the BIC criterion. The BIC criterions are -13734.8 and -13748 respectively for TNB1 and TNB3. In modeling mean and overdispersion, the parameter estimates for the intercept are very different in the two models. We can see this in Table 3. The expected mean for TNB1 and TNB3 are 4.66 and 4.90 respectively. Estimates of overdispersion parameters are 0.527 and 0.491 for TNB1 and TNB3 respectively.

3.6 Conclusions and Implications

It is shown that there is a theoretical and empirical reason to account two times for heterogeneity in modeling recreational demand for National Forests, where individuals are

sampled at various sites which are stratified or clustered according to their use. The first time, heterogeneity accounts for dispersion in the data due to unobservability of the process which results in existence of two different types of visitors in the population, high-frequency and low-frequency. The second time heterogeneity is accounted for in order to capture dependence between individuals sampled at similar sites according to a stratified random sample. Positive results for our hypothesis are found. Both in the poisson and the negative binomial model, the model accounting for site-specific effects performs better than the one not accounting for site-specific effects, with statistically significant results. The results are not of particular interest in deriving consumer surplus per person trip, since the coefficient for the price variable is the same across each model. However, model differences would be important for the purpose of deriving future projections of demand. This is because the expected mean changes across different models. This can be clearly seen from the TNB1 and TNB3 models. The TNB3 model that accounts for site-specific effects has a statistically different intercept in modeling mean and dispersion than the TNB1 model as we allow the intercept to vary randomly across the sample according to a standard normal distribution. Therefore, in this paper a case for treating individuals within a given stratum as dependent rather than independent is made.

Table 3.3: Estimation Results of Outdoor Recreation Demand for George Wahington/Jefferson National Forest: NVUM DATA: 2000-2003

	TSP[a]	TNB[b]	TSP2[c]	TNB1[d]	TNB2[e]	TNB3[f]
Covariates[g]						
Intercept	1.781	1.320	1.778	1.782	1.314	1.795
	(.142)*	(.184)*	(0.084)*	(0.232)*	(0.152)*	(0.227)*
INCE	-0.000004	-0.000003	-.00001	-0.00002	-0.00003	-0.00002
	(0.000004)	(0.000006)	(.000003)**	(.000009)**	(0.00004)	(0.00009)**
AGE	0.003	0.003	0.005	0.003	0.004	0.004
	(0.002)***	(0.002)***	(.001)*	(0.002)***	(0.002)**	(0.002)*
GENDER	-0.050	-0.113	-0.042	-.105	-0.107	-0.105
	(0.071)	(.079)	(0.040)	(0.077)	(0.076)	(0.075)
PEOPVEH	-0.041	-0.048	-0.034	-.034	-0.040	-0.031
	(0.020)**	(.0215)**	(0.010)*	(.021)	(0.022)***	(0.021)
OVERNTE	-0.263	-0.225	-0.189	-0.478	-0.196	-0.465
	(0.071)*	(0.085)***	(0.053)*	(0.212)**	(0.082)**	(0.192)**
TC	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004
	(0.001)*	(0.001)*	(0.0004)*	(0.001)*	(0.0006)*	(0.0006)*
HF	1.715	1.722	1.647	1.740	1.695	1.714
	(0.055)*	(0.055)*	(0.032)*	(0.059)*	(0.065)*	(0.064)*
ALPHA	-	.548	-		0.499	-
	-	(0.092)*	-		(0.076)*	-
sigma	-	-	0.475		0.373	0.317
	-	-	(0.032)*		(0.055)*	(0.071)*
log(a)						
Intercept				1.695		1.768
				(0.340)*		(0.345)*
Income				-.00004		-0.00004
				(0.00001)**		(0.00001)*
OVERNIGHT				-0.963		-0.968
				(0.390)**		(0.382)*
WILD				-.109		-0.059
				(.188)		(0.246)
DUDS				0.653		0.644
				(0.253)***		(0.273)**
GFA				-0.040		-0.048
				0.190		(0.201)
NOBS	781					
LOGL	-2325.67		-2263.5	1376		13780.5
BIC	2355.07	-13733.4	2283.8	-13734.8		-13748.5

*1% significance
**5% significance
***10% significance

[a]Truncated Stratified Poisson
[b]Truncated Stratified Negative Binomial
[c]Truncated Stratified Poisson Accounting For Site-Specific Effects
[d]Truncated Stratified Negative Binomial; Modeling Overdispersion Parameter
[e]Truncated Stratified Negative Binomial Accounting For Site Specific Effect
[f]Truncated Stratified Negative Binomial; Modeling Overdispersion Parameter and Accounting For Site Specific Effects
[g]Coefficient estimates reported in the first row and standard error reported in parentheses

Table 3.4: Estimation Results of Expected Mean and Overdispersion Parameter

	TSP[a]	TNB[b]	TSP2	TNB1	TNB2	TNB3
E(Y)	7.086	4.437	6.155	4.664	4.693	4.909
alpha	0	0.548	0	0.527	0.500	0.491

[a]For all TSP models,

$$E(y \mid x) = EXP(X'\beta) + 1$$

[b]For all TNB models,

$$E(y \mid x) = EXP(X'\beta) + 1 + \alpha EXP(X'\beta)$$

Chapter 4

Estimating Recreational Demand for an On-Site-Sample: A Latent Class Poisson Model

Abstract

In this paper, a recreational demand model for George Washington/Jefferson National Forest using a Latent Class Poisson Model (LCPM) is estimated. We show that the visitor population can be segregated into different classes, each having a different distribution. This class membership for each individual is treated as latent, due to arbitrariness in defining how many visits constitute high or low frequency, and is generated by a multinomial distribution. Conditioned on class membership, visits to a National Forest are modeled as a poisson process. It is shown that there are two classes of visitors, low and high frequency, each visitors with different price estimates. The results are important in deriving consumer surplus measurements. This is because the price coefficient of frequent visits inflates the overall consumer surplus when the two populations are treated as coming from the same distribution.

4.1 Introduction

In previous studies, researchers have dropped high frequency visitors because they constitute a small percentage of the data. Englin and Shonkwiler (1995) dropped the high frequency visitors because of the complexity of modeling high count outcomes. Bowker et al. (2009) controls for high frequency visitors by introducing a dummy variable, which takes a value of 1, when the number of visits exceed 15 visits per year. This approach incorporates heterogeneity in the intercept as a fixed effect. Another approach is to allow the intercept to vary randomly across the sample according to some distribution, in a random effect

approach. Examples include a poisson log normal model and a poisson log gamma model popularly known as the negative binomial model. All these models incorporate heterogeneity based on some definition of a high frequency visitor. There are at least two problems with this approach. First, the dummy variable for high frequency visitors is defined on some assumption about how many visits constitute high frequency - every other day, once in a weekend, or once in two weeks. There is too much arbitrariness in defining high versus low frequency visits. This definition is not the same as having an indicator variable take a value of 1 for participation in recreation and 0 otherwise. Second, high frequency visitors and low frequency visitors come from different populations with different marginal effects. The fixed effect or the random effect treatment does not account for the differences in marginal effects across different populations.

In this paper, class membership is treated as unobserved data. In estimating recreational demand models, two classes are hypothesized. One is a low frequency visitors who make less frequent and longer duration trips with higher travel costs and stronger marginal effects. The other is high frequency visitors who make short duration trips with lower travel costs and weaker marginal effects. The class membership to a particular population is treated as latent and is estimated through posterior probabilities using multinomial distribution. A latent class poisson model is estimated using EM algorithm to estimate demand for the National Forest for the two different populations, low and high frequency visitors.

In the literature,Wedel, M. et al. (1993) use latent class models for count outcomes to identify potential customer market potential customer market segments with different reactivity to direct mail variables. Greene (2005), and Deb and Trivedi (1997), have used latent class models for analyzing health data with the latent classes specifying the unobserved or latent health status. In recreational demand, latent class models are used in a RUM setting for recreational site choice. Examples include mixed site choice models (Train, 2008; Boxall and Adamowicz, 2002). We are not aware of any analysis of on-site recreational samples using latent class models.

4.2 Data

The empirical model will be estimated using NVUM data collected for the George Washington/Jefferson National Forest in the southeastern region of the U.S. The NVUM survey was conducted at 88 sites stratified on the basis of settings within the National Forest. The settings include Wilderness (WILD), Day Use Developed Sites (DUDS), Overnight Use Developed Sites (OUDS), and General Forest Area (GFA). There are 750 sample observations. The data was collected for four sample years, 2000-2003. More detail was provided on NVUM in the previous chapter. For the analysis in this paper, observations for which recreational trips to the National Forest that are less than 160 per year are included. The reason for modeling visits less than 160 is to avoid numerical errors in estimating posterior probabilities. Following Bowker et al. (2009) we also deleted observations with travel costs greater than 720 and observations with people in the vehicle greater than 10.

Table 4.1: Summary Statistics for George Washington/Jefferson National Forest NVUM Data, 2000-2003

	Mean1	Min	Max
INCE[a]	23499.58	11618.19	111898.27
AGE[b]	41.572	18	75
GENDER[c]	0.181	0	1
PEOPVEH[d]	2.459	1	9
OVERNTE[e]	0.223	0	1
TC[f]	46.632	0.469	1103.84
HF[g]	0.316	0	1
NFV12MO1[h]	17.880	1	145
NOBS	750		

[a]IRS reported average after tax income for an individual's ZIP Code
[b]Age
[c]A dummy for Gender equals 1 if female
[d]No.of people in the vehicle
[e]Indicator for overnight stay
[f]As a function of one way travel distance and income foregone
[g]HF=1 if no. of annual visits greater than 15, else $HF = 0$
[h]Number of annual recreation visits per individual or group

4.3 Theoretical Model

The theoretical Model used in this paper is based on Wedel, M. et al. (1993),
In explaining the theoretical model we ignore the time variable as an offset used by Wedel, M. et al. (1993). The estimation is based on the EM algorithm,

$s = 1...S$ denotes class

$l = 1...L$ denotes number of parameters in the model

$i = 1...N$ denotes number of individuals in the sample.

Class membership is unobserved data. Class membership can be estimated via the expectation step,

where,

$$u_{is} \sim MN \tag{4.1}$$

The expectation step is given by,

$$E(u_{is} \mid y_i, a_s, \beta_s) = \frac{a_s P_i \mid s(y_i \mid \beta_s)}{\sum_{s=1}^{S} a_s P_i \mid s(y_i \mid \beta_s)} \tag{4.2}$$

which calculates the posterior probabilities as,

$$\theta_{is} = \frac{\hat{a_s}\hat{P_i} \mid s(y_i \mid \beta_s)}{\sum_{s=1}^{S} \hat{a_s}\hat{P_i} \mid s(y_i \mid \beta_s)} \tag{4.3}$$

Once the expectation is calculated, the likelihood is maximized, keeping the posterior probabilities constant. Maximization step is given by,

$$\sum_{i=1}^{N}\sum_{s=1}^{S} \theta_{is}^{*} ln P_i \mid s(y_i \mid \beta_s) + \sum_{i=1}^{N}\sum_{s=1}^{S} \theta_{is}^{*} ln \alpha_s \tag{4.4}$$

The estimate of a_s^* is given by,

$$a_s^* = \sum_{i=1}^{N} \theta_{is}^{*} / N \tag{4.5}$$

where 'a' is the proportion of individuals belonging to a particular class, s.
The probability distribution for the dependent variable in our model is the poisson distribution which corrects for endogenous stratification and truncation (Shaw, 1988),

$$lnP_i \mid s(y_i \mid \beta_s) = -exp(x_{il}\beta_s) + (y_i - 1)x_{il}\beta_s - log(\Gamma(y_i)) \tag{4.6}$$

Estimates of β_{ls} are given by,

$$\sum_{i=1}^{N} \theta^*_{is}((y_i - 1) - \lambda_{i|s})x_{il} = 0 \tag{4.7}$$

Expectation and the maximization step is iterated until convergence.

An alternative to maximum likelihood estimation using the EM algorithm is iteratively reweighted least Squares. The dependent variable and weights are updated at each iteration until convergence. For estimation using iteratively reweighted least Squares (Wedel et al., 1993),
The standard errors are computed as,

$$F_{ll'} = -E(\frac{\partial^2 l_s}{\partial\beta_{ls}\partial\beta_{l's}}) = \sum_{i=1}^{N} \theta_{is}\lambda_{i|s}x_{il}xil' \tag{4.8}$$

The consistent akaike Information Criterion (Bozdogan, 1987) used in Wedel et al. (1993) is used to determine the appropriate number of classes.

$$CAIC = -2\sum_{s=1}^{S} l_s + [(L+1)S - 1](ln(N) + 1) \tag{4.9}$$

4.4 Empirical Model

The empirical model is specified as,

$$NFV12MO^i = f(INCE^i, AGE^i, GENDER^i, TC^i, OVERNTE^i, PEOPVEH^i) \tag{4.10}$$

The dependent variable in the empirical model is the number of annual recreation visits per group to the George Washington/Jefferson National Forest. Demand for visits is a function of six variables: IRS reported average after tax income for an individual's ZIP Code (INCE), own price or cost of the trip (TC)[1] , number of people in the vehicle (PEOPVEH), gender (GENDER1), age (AGE), and an indicator for staying overnight (OVERNTE).

4.5 Results

4.5.1 Model Selection

Initially, a mix of two negative binomial distributions accounting for stratification and truncation was tried. The model becomes highly unstable as the probability of each individual being a low frequency visitor approaches one and the probability of each individual being a high frequency visitor approaches zero. Therefore, the results must be based on poisson probabilities.

Figure 4.1 compares the empirical data with two negative binomial distributions each with different means and overdispersion parameters. It can be easily seen that high frequency visitors are not represented using a negative binomial model even with varying overdispersion. The empirical density in Figure 4.1 suggests three latent classes: low frequency visitors with an average of 10 visits in a year, medium frequency visitors with an average of 50 visits, and high frequency visitors with an average of about 100 visits. Respectively these three classes can be identified as: visitors who make few long duration visits perhaps once a month; visitors who make frequent short duration visits, perhaps once in a weekend; and visitors who visit a national forest as a part of their exercise schedule, perhaps every other day. However, the

[1] Bowker et al. (2009),

$$TC = 2 * (0.12 * PRACDIS) + 2 * (0.33 * \frac{INCE}{2000} * PRACTIME)$$

where PRACDIS is the one way distance to the site. A per mile cost of 0.12 dollar was used. Income forgone is calculated as one third of the wage rate, where wage rate is calculated as the proxy of annual income divided by 2000 work hours. PRACTIME is the time spent at the site.

model fitness criterion suggests that the two-class model performs better than the three-class model, implying the existence of only two classes, i.e. high and low frequency visitors. The purpose of using a latent class model is to show heterogeneity, not just in the intercept, but also in the travel cost (price) variable for different classes. The travel cost variable is expected to be of less importance for the high frequency visitor. In the second section, it is also shown how the consumer surplus measures vary across different classes and also across models.

In estimating the two class latent poisson model in Table 4.2, the EM algorithm converges in about 32 iterations. Parameter estimates of 'a' suggest that 72% of visitors come from a low frequency population and 28% come from a high frequency population. The travel cost variable is significant for both classes. However, the magnitude differs in the two classes. The travel cost variable for low frequency visitors is -0.014, and for high frequency visitors it dampens to -0.006. The heterogeneity in the intercept is clearly shown with 2.935 for low frequency visitors and 5.251 for high frequency visitors. Also, the coefficient for overnight stay has a different sign for two classes. For low frequency visitors, who take few long duration trips, the sign of overnight visits is negative. An overnight stay at a national forest correlates with a reduced number of visits per year for the precise reason that they make few such trips in a year. For high frequency visitors, the opposite is true: Visitors who take frequent short duration visits have more visits per year.

In estimating the three class latent poisson model, the EM algorithm converges in about 50 iterations. The estimates of 'a' in Table 4.3 suggest that 57% of visitors come from a low frequency population, 33% from a medium frequency population, and 10% from a high frequency population. The travel cost variable is significant for all three classes. However, the magnitude differs in three classes. These results are consistent with the two-class model. Also, the magnitude of the travel cost variable is consistent with the two-class model. As with the two-class model, heterogeneity in the intercept is clearly shown with 1.919 for low frequency visitors, 3.650 for medium frequency visitors, and 4.473 for high frequency visitors.

Table 4.2: Estimation Results of Two Class Latent Poisson Model of Outdoor Recreation for George Washington/Jefferson National Forest : NVUM Data, 2000-2003

Explanatory Variables[a]	TSP	TSNB	Class1	Class2
Intercept	2.477	1.395	2.935	5.251
	(0.195)*	(.199)*	(0.093)*	(0.064)*
INCE	-.00002	-.000005	-0.00001	-0.00003
	(0.000007)*	(.000006)	(0.000003)*	(0.000002)*
AGE	-.0018	.00120	0.001	-0.006
	(.002)	(.0021)	(0.001)	(0.0008)*
GENDER	-.189	-.159	-0.760	-0.622
	(.0778)**	(.083)**	(0.056)*	(0.043)*
TC	-.004	-.004	-0.014	-0.006
	(.0010)*	(.001)*	(0.0007)*	(0.0004)*
PEOPVEH	-.081	-.068	-0.129	-0.096
	(.0281)*	(.023)*	(0.014)*	(0.009)*
OVERNTE	-.254	-.219	-0.186	0.851
	(.131)***	(.092)**	(0.044)*	(0.102)*
a	-	-	0.72	0.28
HF	2.027	2.075	-	-
	(.066)*	(.065)*	-	-
alpha	-	.790	-	-
	-	(.103)*	-	-
NOBS	750			
LOGL	-3906	3217	-4.8800e+003	-2.6188e+004
CAIC	-	-		6.2235e+004
BIC	3932.95	-32145.2	-	-
*1% significance				
**5% significance				
***10% significance				

[a]Coefficient estimates reported in the first row and standard error reported in parentheses

Table 4.3: Estimation Results of Three Class Latent Poisson Model of Outdoor Recreation for George Washington/Jefferson National Forest : NVUM Data, 2000-2003

Explanatory Variables[a]	Class1	Class2	Class3
Intercept	1.919	3.650	4.473
	(0.142)*	(0.009)*	(0.087)*
INCE	-0.000001	-0.00001	-0.000008
	(0.000005)	(0.000003)*	(0.000003)**
AGE	0.005	0.005	0.0026
	(0.002)*	(0.001)*	(0.001)**
GENDER	-0.515	-0.608	-0.604
	(0.076)*	(0.053)*	(0.059)*
TC	-0.014	-0.009	-0.006
	(0.001)*	(0.0006)*	(0.0005)*
PEOPVEH	-0.061	-0.019	0.058
	(0.021)*	(0.010)*	(0.017)***
OVERNTE	-0.566	-0.702	0.396
	(0.079)*	(0.049)*	(0.102)*
a	0.5715	0.3351	0.0934
NOBS	750		
LOGL	-881.7038	-1.2730e+004	-1.8551e+004
CAIC		6.4478e+004	
*1% significance			
**5% significance			
***10% significance			

[a]Coefficient estimates reported in the first row and standard error reported in parentheses

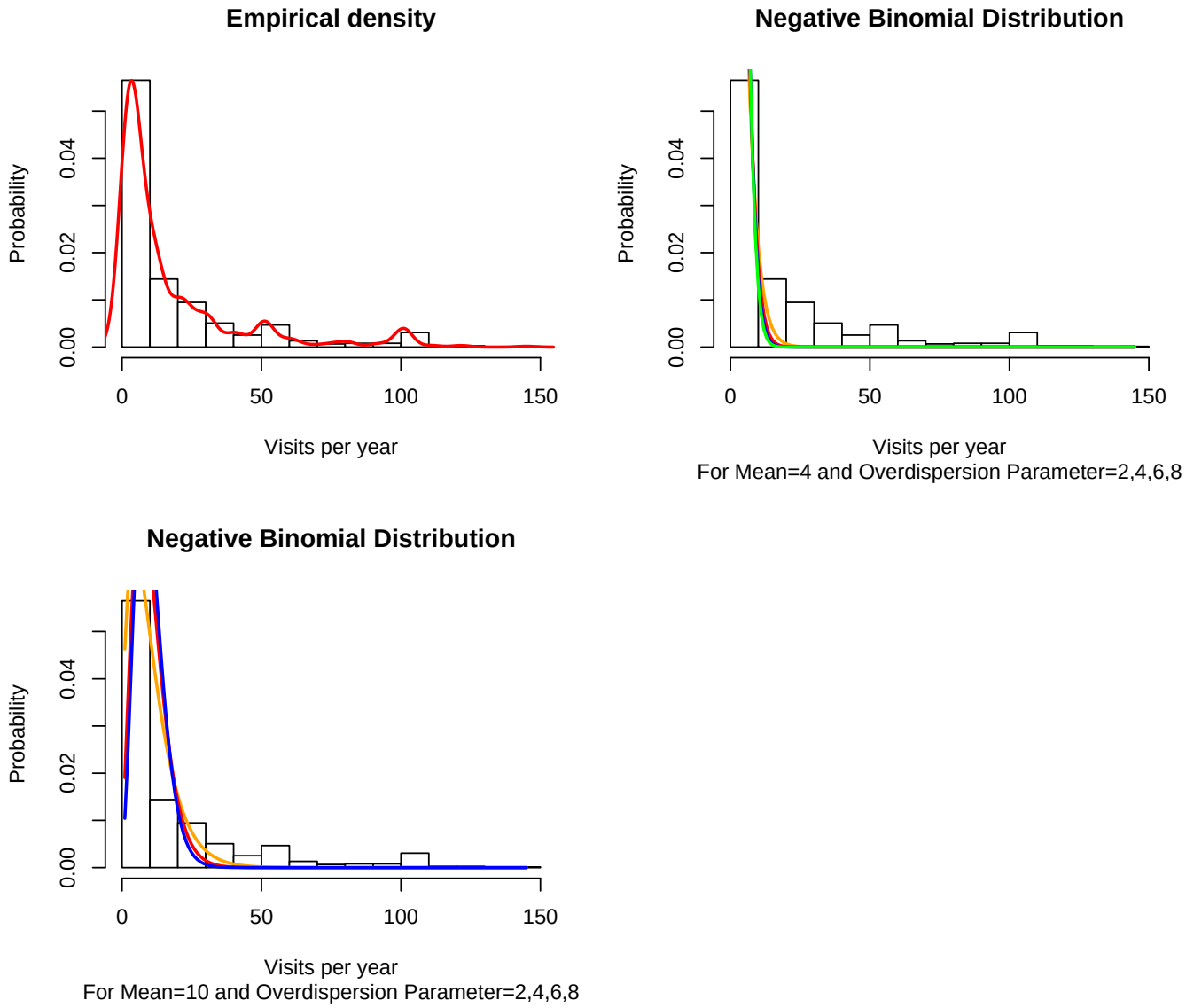

Figure 4.1: Empirical Density for George Washington/ Jefferson National Forest: NVUM Data, 2000-2003

Consistent with the results of the two-class model, the three-class model has different signs for the coefficient for the overnight stay across the three classes. For low and medium frequency visitors, the sign is negative, again showing that an overnight stay at a national forest reduces

Table 4.4: Estimation Results of Goodness of Fit for Latent Class Poisson Model

	Class2	Class3
CAIC	6.2235e+004	6.4478e+004

their number of visits per year. For high frequency visitors, the opposite is true:Visitors who take frequent short duration trips visits have more visits per year.

Table 4.4 gives the goodness of fit criteria for the two and three-class latent poisson model. The CAIC goodness of fit criterion is calculated using the formula described in the theoretical section of this paper. Based on the CAIC criterion, the two-class model performs better than the three-class model. Therefore, the existence of only two classes: high and low frequency visitors to a national forest is claimed in this paper. There is no need for the additional class of medium frequency visitors.

The own price elasticity of demand[2] are also computed for class 1 and class 2 in the Latent Class Poisson Model. This formula is derived from Wedel et al.(1993). The absolute own price elasticity of demand for the two classes are 0.653 and and 0.280, respectively. Demand for outdoor recreation is insensitive to price changes in both classes.

4.5.2 Welfare Calculations

In this section, the consumer surplus that users derive from recreating at a National Forest are derived. These consumer surplus measures apply to the recreation user population because the population density for our analysis is truncated at zero. This is different from Englin

[2] $$\varepsilon_{TC} = \beta_{TC} * \sum_i TC_i/N$$

where N are the total number of individuals in the sample and TC is the price variable.

and Shonkwiler (1995) where they derive their estimates from the entire general population instead of the recreation user population. where,

Table 4.5: Estimation Results of Consumer Surplus

	TSP	TSNB	Class 1(Low Frequency)	Class 2(High Frequency)
CS/persontrip/year	250.00	250.00	71.43	166.67
$E(y \mid x)$	3.10	1.45	2.55	4.85
CS/average person/year	772.51	363.00	182.22	809.33

$$E(y \mid x) = exp(x\beta) + 1 \tag{4.11}$$

for TSP and Latent Class TSP Model, and

$$E(y \mid x) = exp(x\beta) + 1 + \alpha exp(x\beta) \tag{4.12}$$

for TSNB model, based on Englin and Shonkwiler(1995)

$$CS/persontrip = -1/\beta_{TC} \tag{4.13}$$

$$CS/averageindividual = -1/\beta_{TC} E(y \mid x) \tag{4.14}$$

Model estimates for the price variable stay the same for the TSP and TSNB models. Therefore, in terms of the welfare measures, the consumer surplus per person trip remains the same for the two models. Accounting for heterogeneity with an introduction of the overdispersion parameter in the negative binomial model does not change the coefficient for the price variable. This is not true for the latent poisson model suggested in this paper. Not only is the coefficient different across two classes but it is different from the TSP model and TSNB models. Accounting for individual heterogeneity by introducing an overdispersion parameter in the negative binomial model fails to account for heterogeneity in the price variable due to the different trip frequencies of the visitors. However, the latent class poisson model does account for the individual heterogeneity.

Consumer surplus per average person per year gives the willingness to pay for the use of National Forest for the purpose of outdoor recreation. The willingness to pay for the high frequency visitor is much greater than the low frequency visitor.

4.6 Conclusion and Implications

Results from this study have several important implications. First, this study uses a superior model for recreational demand by treating class membership of a heterogeneous population as latent, thereby avoiding any ambiguity in defining classes based on frequency of visits. Also, treating class membership as observed by including a dummy for high and low frequency visitors will not account for the differences in coefficients across different classes. The technique used in this paper accounts for heterogeneity not only in the intercept but also in the coefficients. Second, in this paper it is shown that it is important to account for heterogeneity in price estimates when deriving welfare measures such as consumer surplus. It is shown that the high frequency visitors, who take frequent short duration visits, derive greater benefits from recreation at the national forest and therefore have a higher consumer surplus as compared to low frequency visitors who take less frequent long duration trips. This information on market segregation between high and low frequency visitors can be of importance to the USDA Forest Service because differences in consumer surplus across classes provide potential scope for differential pricing policies.

Future research can mix two negative binomial distributions to estimate an on site recreational demand model. Since the EM algorithm is slow to converge, maximum likelihood estimation can be performed by imposing restrictions on the probabilities of class membership.

Chapter 5

Conclusions and Implications

5.1 Summary and Conclusions

An overall goal of this dissertation is to address the modeling and estimation issues associated with choice-based samples that go beyond the problems of non-negativity, truncation, and endogenous stratification which have been resolved in the literature. Another overall goal is to show the sensitivity of welfare measurements when the basic and naive assumptions about the data generating process are dropped. This information is of relevance to government agencies such as the USDA Forest Service, which uses these estimates for evaluating policies such as developing a site, closing an old site, or opening a new site.
In the first essay (Chapter 2) empirical tests for comparing weighted and unweighted estimates of regional demand models of recreation are suggested. In the literature, weighting is used when regional demand models are misspecified by pooling different populations together when observations on a given population are not enough to correct misspecification. It is shown that in doing so the weights interact with the covariates in a weighted regression, resulting in heteroskedasticity. By estimating both weighted and unweighted models, we show that it is best to use unweighted regression when the coefficient on interactions with the weight variable are insignificant. However, the model needs to be respecified if these interactions are jointly significant but the estimation still proceeds using an unweighted regression.

Choice-based samples are nonrandom samples, especially if they are stratified. It is shown that it is important to account for dependence caused by sampling at different sites.

This dependence is caused by a common factor: the site where they are sampled. In the second essay (Chapter 3), statistically significant site-specific effects are obtained. In terms of the model criterion, the model with site-specific effects performs better. Because of the differences in the intercept estimates and other marginal coefficients, the expected mean changes across the models. This illustrates how willingness to pay measurements are sensitive to the assumptions that the sampling procedures are random or clustered.

High order integer values are dropped from the analysis of recreational preferences because they are difficult to model, or because these observations constitute a small percentage of the population. In the third essay (Chapter 4), heterogeneous preferences are modeled by assuming that there are two classes originating from different distributions of visitors based on frequency of visits. Due to the arbitrariness in defining what constitutes high and low frequency, these classes are treated as latent. The results from the latent class poisson models show the existence of two classes: low and high frequency visitors, each with different marginal effects. The price coefficient is of particular interest. For low frequency visitors, who have a less elastic demand, the sign remains the same but the magnitude is much stronger than for high frequency visitors. The consumer surplus measurements are different in these two classes.

5.2 Policy Implications

In this dissertation it is shown that welfare estimates are sensitive to assumptions about the stochastic process or the sampling procedure of the data generating process. These results can provide important information to policy makers such as the USDA Forest Service, who need these estimates for policy analysis. The first two essays points to differences in welfare estimates due to different assumption about the sampling procedure. The third essay points to the difference in welfare estimates resulting from different assumptions about the stochastic process of the data generating process. The welfare estimates derived in the third

essay give important information for potential differential pricing policies. The low frequency visitors with a relatively elastic demand who mostly visit developed sites would not be willing to pay a higher price because of the smaller consumer surplus per person trip and fewer trips per year. However, the high frequency visitors with a relatively inelastic demand would be willing to pay a higher price because of the larger consumer surplus and more visits per year.

Estimated recreation demand models are often used by natural resource management agencies to project current and future recreation visitation and estimate the per unit economic value of recreation (e.g., consumer surplus per person trip, per trip, or per day). Visitation projections and economic value estimates provide input into benefit-cost analysis and resource management decisions. The results of this research suggest that both recreation visitation estimates (e.g., expected trips) and per unit economic values (e.g., consumers surplus per trip) may be sensitive to specific econometric techniques used to estimate recreation demand functions from choice-based samples. Thus, resource managers should interpret and apply recreation demand functions estimated from choice-based samples with care and caution.

5.3 Limitations of the Study

The results in the second and third essay of this dissertation are subject to a common limitation. Because of a relatively small data set, different activities are aggregated together when estimating models of outdoor recreation. In doing so, the restrictive assumption of equal marginal effects across various outdoor recreational activities is imposed. A more realistic assumption would be to expect the marginal effects to vary across activities.

5.4 Recommendations for Future Research

Future research can focus on the missing variable problem in Chapter 2 in case the interaction terms are jointly significant.

In Chapter 4, future research can also focus on mixing two negative binomial distributions to model heterogeneous preferences in a latent class model. This would allow dispersion among individuals within a given latent class; an assumption that is relaxed in our study.

Bibliography

[1] Arnold, B.C., and D. Strauss, "Pseudolikelihood Estimation: Some Examples", *The Indian Journal of Statistics*, Series B, Vol. 53, No. 2, (Aug. 1991) 233-43.

[2] Besag, J.E., "Statistical Analysis of Non-Lattice Data", *The Statistician*, Vol. 24 (1975) 175-195.

[3] Besag, J.E., "Efficiency of Pseudolikelihood Estimators for Simple Gaussian fields", *Biometrika*, Vol. 64 (1977) 616-18.

[4] Bloom, D.E., and T.L. Idson, "The Practical Importance of Sample Weights", Proceedings of the Survey Research Methods Section. http://www.amstat.org/sections/SRMS/Proceedings/papers/1991_106.pdf (Accessed December 8, 2009).

[5] Bowker, J.M., C.M. Starbuck, D.B.K. English, J.C. Bergstrom, R.S. Rosenberger, and D.W. McCollum, "Estimating the Net Economic Value of National Forest Recreation: An Application of the National Visitor Use Monitoring Database" *Faculty Series Working Paper, FS 09-02, September 2009, Department of Agricultural and Applied Economics, The University of Georgia, Athens.*

[6] Boxall P.C., and W.L. Adamowicz, "Understanding Heterogeneous Preferences in Random Utility Models: A Latent Class Approach", *Environmental and Resource Economics* 23 (2002): 421-446.

[7] Bozdogan, H., "Model Selection and Akaike's Information Criterion (AIC) : The General Theory and its Analytical Extensions", *Psychometrika*52 (1987): 345-70.

[8] Cameron, A. C., and P. K. Trivedi, "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests ", *Journal of Applied Econometrics*, Vol 1(1) (Jan. 1986): 29-53.

[9] Cox, D.R., and N. Reid, "A Note on Pseudolikelihood Constructed from Marginal Densities", *Biometrika*, Vol. 91, No. 3 (Sep. 2004), pp. 729-737.

[10] Deb, P. and P.K. Trivedi "Demand for Medical Care by the Elderly: A Finite Mixture Approach.", *Journal of Applied Econometrics*, Vol. 12 (1997): 313-336.

[11] DuMouchel, W.H. and G.J. Duncan, "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples", *Journal of the American Statistical Association*, Vol. 78, No. 383 (Sep. 1983) 535-43.

[12] Egan, K., and J. A. Herriges, "Multivariate Count Data Regression Models with Individual Panel Data From an On-Site Sample", *Journal of Environmental Economics and Management*, Vol. 52 Issue 2 (Sep. 2006), 567-581.

[13] Englin, J., and J.S. Shonkwiler, "Estimating Social Welfare Using Count Data Models: An Application to Long Run Recreation Demand under Conditions of Endogenous Stratification and Truncation", *Review of Economics and Statistics* 77 (1995): 104-112.

[14] English, D.B.K, et al., "Forest Service National Visitor Use Monitoring Process: Research Method Documentation." *United States Department of Agriculture, Forest Service, General Technical Report SRS-57* 2002.

[15] Freedman D. A., "On The So-Called 'Huber Sandwich Estimator' and 'Robust Standard Errors'", *American Statistician*, Vol. 60(4) (2006) :299-302.

[16] Galwey, N.W., "Introduction to Mixed Modelling, Beyond Regression and Analysis of Variance", *John Wiley and Sons, Ltd.* 2006.

[17] Greene, W.H., "Functional Form and Heterogeneity in Models for Count Data, *Foundations and Trends in Econometrics"*, Vol. 1, No. 2 (2005) 113-218.

[18] Haab, T.C. and K.E. McConnell, "Count Data Models and the Problem of Zeros in Recreation Demand Analysis", *American Journal of Agricultural Economics*, Vol. 78 (Feb 1996): 89-102.

[19] Hanemann, W. N., "Discrete/Continuous Models of Consumer Demand" *Econometrica* Vol. 52, No.3 (May, 1978): 541-561.

[20] Hellerstein, D.M. "Using Count Data Models in Travel Cost Analysis with Aggregate Data", *American Journal of American Economics*, Vol. 73 (Aug 1991): 860-66.

[21] Jansen, R.C., "Maximum Likelihood in a Generalized Linear Finite Mixture Model by Using the EM Algorithm", *Biometrics*, Vol. 49, No. 1 (1993) 227-231.

[22] Karan, V., and D.P. Cram, "Review of Choice-Based and Matched Sample Studies in Management Accounting Research," July 2006.
http://aaahq.org/mas/MASPAPERS2007/research_forum/Karan%20and%20Cram.pdf
(Accessed July 15, 2010).

[23] Korn, E.L., and B.I. Graubard, "Examples of Differing Weighted and Unweighted Estimates from a Sample Survey", *The American Statistician*, Vol. 49, No. 3 (August, 1995): 291-95.

[24] Lawless, J.F., "Negative Binomial and Mixed Poisson Regression", *The Canadian Journal of Statistics*, Vol. 15, No. 3,(Sep. 1987), 209-25.

[25] Leamer, E.E.,"Specification Searches Ad Hoc Inference with Non Experimental Data" *Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, Inc.*, 1978.

[26] Longford, N.T., "Random Coefficient Models", *Oxford Science Publications, Clarendon Press*, 1993.

[27] Manski, C.F., and D.L. McFadden, "Chapter 1: Alternative Estimators and Sample Designs for Discrete Choice Analysis, Structural Analysis of Discrete Data and Econometric Applications", *Cambridge: The MIT Press*, 1981.

[28] Manski, C.F., and S.R. Lerman, "The Estimation Of Choice Probabilities From Choice Based Samples", *Econometrica*, Vol.45, No.8,(Nov 1977): 1977-1988.

[29] Martínez-Espiñeira R., and J. Amoako-Tuffour, "Recreation Demand Analysis Under Truncation, Overdispersion and Endogenous Stratification: An Application to Gros Morne National Park", *Journal of Environmental Management* 88, 2008, 1320-32.

[30] Morey, E.R., "The Demand for Site-Specific Recreational Activities: A Characteristic Approach", *Journal of Environmental Economics and Management*, 8 (1981): 345-71.

[31] Murdock, J., "Handling Unobserved Site Characteristics in Random Utility Models of Recreation Demand", *Journal of Environmental Economics and Management*, 51 (2006): 1-25.

[32] Pepper, J.V., "Robust Inferences From Random Clustered Samples: An Application Using Data From Panel Data Study of Income Dynamics", *Economic Letters*, 75 (2002): 341-345.

[33] Phaneuf, D. J., C. L. Kling, and J. A. Herriges, "Estimation and Welfare Calculations in a Generalized Corner Solution Model with an Application to Recreation Demand", *Review of Economics and Statistics*, Vol 82(2) (2000): 83-92.

[34] Shaw, D., "On-Site Samples' Regression: Problems of Non-Negative Integers, Truncation, and Endogenous Stratification", *Journal of Econometrics*, Vol. 37 No. 2,(1988): 211-223.

[35] Train, K., E.,"EM Algorithms for Nonparametric Estimation of Mixing Distributions", *Journal of Choice Modeling*, 1(1),(2008):40-69.

[36] Wang, X., C.V. Eden, and J.V. Zidek, "Asymptotic Properties of Estimators in Maximum Likelihood Estimation", *Journal of Statistical Planning and Inference*, 119, (2004) 37-54.

[37] Wedel, M., et al., "A Latent Class Poisson Regression Model for Hetereogeneous Count Data", *Journal of Applied Econometrics*, Vol. 8, No. 4 (1993): 397-411.

[38] Winklemann, R., "Seemingly Unrelated Negative Binomial Regression," *Oxford Bull. Econ. Statist.*, 62(4), (2000): 553-560.

[39] Winship, C., and L. Radbill, "Sampling Weights and Regression Analysis", *Sociological Methods and Research*, Vol. 23 No. 2, Nov. 1994, 230-257.

[40] Woolodridge J.M., "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples," *Economic Theory*, Vol. 17, (2001): 451-70.

Appendix A

Appendix to Chapter 2

Table A.1: Summary Statistics for General Forest Area Settings, NVUM Data, 2000-2003

	Mean1	Min	Max
HOTCONT	0.457	0	1
MOUNTAIN	0.190	0	1
SUBTROP	0.352	0	1
FORESTP	44.22	0.085	85.468
TRAILS	386.078	20	704.700
INCE	21619.06	9910.434	90831.38
AGE	43.430	17.5	75
GENDER	0.160	0	1
PEOPVEH	2.254	1	10
OSITE	0.235	0	1
OVERNTE	0.225	0	1
TC	45.108	0	1221.672
HF	0.328	0	1
NFV12MO1	13.793	1	53
NOBS	1979		

Table A.2: Summary Statistics for Day Used Developed Settings, NVUM Data, 2000-2003

	Mean1	Min	Max
HOTCONT	0.373	0	1
MOUNTAIN	0.281	0	1
SUBTROP	0.346	0	1
PICNICTAB	163.485	1	1258
SUMNATPARK	10.927	0	204
SWMMING	5.619	0	9
INCE	22808.02	8006.103	105597.6
AGE	44.063	17.5	75
GENDER	0.328	0	1
PEOPVEH	2.835	1	10
OSITE	0.325	0	1
OVERNTE	0.111	0	1
TC	64.339	0.024	1150.758
HF	0.328	0	1
NFV12MO1	8.533	1	53
NOBS	2394		

Table A.3: Summary Statistics for Overnight Used Developed Settings, NVUM Data, 2000-2003

	Mean1	Min	Max
HOTCONT	0.374	0	1
MOUNTAIN	0.307	0	1
SUBTROP	0.319	0	1
SUMCAMPS	34.934	1	247
TENTC	452.149	22	1254
INCE	22570.38	9033.333	106902
AGE	42.693	17.5	75
GENDER	0.273	0	1
PEOPVEH	2.656	1	10
OSITE	0.331	0	1
OVERNTE	0.574	0	1
TC	42.405	0.296	728.2
HF	0.139	0	1
NFV12MO1	7.461	1	53
NOBS	1707		

Table A.4: Summary Statistics for Wilderness Settings, NVUM Data, 2000-2003

	Mean1	Min	Max
HOTCONT	0.412	0	1
MOUNTAIN	0.071	0	1
SUBTROP	0.517	0	1
SUMWILDERN	1.008	0	245
DESIGW	35187.1	13812	118337
INCE	26142.53	13052.6	111898.3
AGE	38.355	17.5	75
GENDER	0.276	0	1
PEOPVEH	2.754	1	9
OSITE	0.294	0	1
OVERNTE	0.297	0	1
TC	62.588	1.466	634.357
HF	0.075	0	1
NFV12MO1	5.442	1	53
NOBS	622		

Table A.5: Weighted Means of all Four Settings, NVUM Data, 2000-2003

	GFA	DUDS	OUDS	WILD
HOTCONT	0.334	0.300	0.506	0.252
MOUNTAIN	0.124	0.255	0.207	0.076
SUBTROP	0.542	0.444	0.287	0.671
FORESTP	44.425	-	-	-
TRAILS	301.166	-	-	-
PICNICTAB	-	152.148	-	-
SUMNATPARK	-	11.571	-	-
SWIMMING	-	5.082	-	-
SUMCAMPS	-	-	39.090	-
TENTC	-	-	565.453	-
SUMWILDERN	-	-	-	6.391
DESIGW	-	-	-	36006.0446
INCE	20893.5	22514.06	22213.99	26564.59
AGE	45.364	46.913	46.598	41.318
GENDER	0.135	0.272	0.260	.205
PEOPVEH	2.112	2.733	2.427	2.629
OSITE	0.139	0.215	0.202	0.269
OVERNTE	0.162	0.047	0.437	0.237
TC	40.059	69.535	47.9986	99.323
HF	0.372	0.165	0.170	0.143
NFV12MO1	14.638	8.483	8.008	8.537

Appendix B

Data Documentation

B.1 Input File: Supply Variable Creation

Generates a file that contains supply variables for all settings and merges it with the NVUM data for the Southeastern U.S. This .XLS file named trimmeddata can be found at the following directory on the external hard drive; G:/Chapter2_Data. This data is used for estimation in Chapter 2. This supply variable creation file is in SAS format and can be found at the following directory on the external hard drive; G:/CHAPTER 2/Supply Variable Creation.

B.1.1 Zonal Database

Source: RPA Recreation Supply Database, RWU 4953 - Greatest Good for the 21st Century : A Program for Pioneering Research on Changing Forest Values in the South and Nation, Southern Research Station, U.S.D.A Forest Service.

These are four SAS data files: Zone 30, Zone 50, Zone 100, and Zone 200. For our analysis, we used Zone 50 and Zone 100 SAS data files. These files contain information on counties within a 50 and 100 mile radius for the Southeastern U.S. These data files are in .XLS format and can be found at the following directory on the external hard drive; G:/Chapter2_Data/zone100.

B.1.2 Land-use Database

Source: RPA Recreation Supply Database, RWU 4953 - Greatest Good for the 21st Century: A Program for Pioneering Research on Changing Forest Values in the South and Nation, Southern Research Station, U.S.D.A Forest Service.

This is a .XLS file named "landarea" with 15 Worksheets and can be found at the following directory on the external hard drive; G:/Chapter2_Data/landarea. For our analysis, we have used three Worksheets: CBP07, RECGOV and LANDUSE. The variables SUMCAMPS, SUMWILDERN, and NATPARK in our analysis are created from the following variables.

(1.) Camps under the CBP07 worksheet:

(2.) Natpark under the CBP07 worksheet

(3.) Wildern under the RECGOV worksheet

B.1.3 Southern National Forest Recreation Site Characteristic Database

Source: John C. Bergstrom and Angela Boothe, Department of Agriculture and Applied Economics, The University of Georgia, Athens.

This is a .XLS file named site characteristics and can be found at the following directory on the external hard drive; G:/Chapter2_Data/Site characteristics. The variables used from the database include:

(1.) TentC

(2.) PicnicTab

(3.) Swimming

(4.) DesigW

(5.) Trails

B.1.4 Zipfips Database

Source : sashelp.zipcode

This is a .XLS datafile and can be found at the following directory on the external hard drive; G:/Chapter2_Data/zipfips. It includes the following variables:

(1.) Zipcode

(2.) Fips

for the U.S.

We converted this sas file in to a .XLS file and is named zipfips for our analysis.

B.1.5 NVUM 4 Database

Source : Donald B. K. English, 2000-2003 NVUM Raw Database, National Visitor Use Monitoring Program, Washington D.C.

This is a .XLS file and can be found at the following directory on the external hard drive; G:/Chapter2_Data/nvum4.

B.1.6 IRS Database

Source: J. M. Bowker et al., "Estimating the Net Economic Value of National Forest Recreation: An Application of the National Visitor Use Monitoring Database", Appendix A, Faculty Series Working Paper, FS 09-02, September 2009, The University of Georgia, Department of Agricultural and Applied Economics, Athens, GA.

We have used the SAS data file named irsdata under the Income Data folder. This is a .XLS datafile and can be found at the following directory on the external hard drive; G:/Chapter2_Data/irsdata.

B.1.7 SUBDIST. Database

Source: J. M. Bowker et al., "Estimating the Net Economic Value of National Forest Recreation: An Application of the National Visitor Use Monitoring Database", Appendix A, Faculty Series Working Paper, FS 09-02, September 2009, Department of Agricultural and Applied Economics, The University of Georgia, Athens.

We have used the SAS data file named subdist under the NAC4 GENERATION folder. This is a .XLS datafile and can be found at the following directory on the external hard drive; G:/Chapter2_Data/subdist.

B.1.8 Ecoregion Database

Source: J.M. Bowker and J. C. Bergstrom, Department of Agricultural and Applied Economics, The University of Georgia, Athens.

This is a .XLS file containing data on the ecoregion of National Forest in the Southeastern U.S. and can be found at the following directory on the external hard drive; G:/Chapter2_Data/ecoregion.

B.2 Input File : Chapter3_data_creation

Generates a file that contains observations from George Washington/Jefferson National Forest from NVUM Data: 2000-2003. This is a SAS input file and can be found at the following directory on the external hard drive; G:/Chapter3 _data_creation_prog. This input file creates a .XLS data file named georgeSimul that can be found at the following directory on the external hard drive; G:/Chapter3_Data. This data is used for estimation in Chapter 3.

B.2.1 NVUM 4 Database

Source: English, D.B.K., 2000-2003 NVUM Raw Database, National Visitor Use Monitoring Program, Washington D.C.

This is a .XLS file and can be found at the following directory on the external hard drive; G:/Chapter2_Data/nvum4.

B.3 Data for Chapter 4

The data file, data_chapter4 contains observations from George Washington/Jefferson National Forest with visits less than 160 from NVUM Data: 2000-2003 and can be found at the following directory on the external hard drive; G:/Chapter4_data. This data is used for estimation in Chapter 4.

B.3.1 NVUM 4 Database

Source: English, D.B.K., 2000-2003 NVUM Raw Database, National Visitor Use Monitoring Program, Washington D.C.

This is a .XLS file and can be found at the following directory on the external hard drive; G:/Chapter2_Data/nvum4.

Appendix C

Estimation Programs

C.1 Input files: Estimation Programs

This section contains information on the statistical software used in the estimation and the names of input files.

(1.) Chapter 2 is estimated using the ml procedure in TSP software. The input files include stype1_WILD_Table2.4, stype2_OUDS_Table2.3, stype3_DUDS_Table2.2, and stype4_GFA_Table2.1 under the folder Chapter2_estimation_prog and can be found at the following directory on the external hard drive; G:/Chapter2_estimation_prog

(2.) Chapter 3 is estimated using the proc nlmixed procedure in SAS software. The name of the input file is chapter3_estimation_prog and can be found at the following directory on the external hard drive; G:/Chapter3_estimation_prog

(3.) Chapter 4 is estimated using the fminsearch optimization routine in Matlab. The input files include chapter4_em_2class, chapter4_em_3class, and em under the folder Estimation_prog_cha4 and can be found at the following directory on the external hard drive; G:/Estimation_prog_cha4.