ESTIMATING THE DEMAND FOR AND VALUE OF RECREATION ACCESS TO NATIONAL FOREST WILDERNESS: A COMPARISON OF TRAVEL COST AND ON-SITE COST DAY MODELS

by

JIANPING ZHU

(Under the Direction of Lynne Seymour and J. M. Bowker)

ABSTRACT

The Travel Cost Method (TCM) is the dominant method for recreation demand analysis. The number of trips to a site is assumed to relate to travel cost, travel time and other demographics. One disadvantage of the conventional TCM is that it does not recognize potential spatial limitations. Bell and Leeworthy (1990) developed an alternative model to deal with the spatial problems associated with visitors coming from significant distances to use principally beach resources. Instead of modeling annual trip numbers to the recreational site, the number of days spent at the site is modeled. This model is called on-site cost model (OSCM) or expenditure day model. In both TCM and OSCM, the response variables are count data based on an on-site survey. In this thesis, the OSCM, introduced by Bell and Leeworthy (1990), is applied to visitation at National Forest Wilderness areas. In both cases, the estimators need to account for the fact that the dependent variables are non-negative integers under truncation, overdispersion, and endogenous stratification. Additionally, the more conventional TCM is applied to these same data. Results from both models are compared and the models are tested with respect to derived welfare estimates. Visitors' total Wilderness days are modeled as a dependent variable on several demographic and socioeconomic factors. This model includes visitors who are from a significant distance and those from a short distance. A statistical justification about which assumptions are violated is provided by applying Ordinary Least Square (OLS) regression model to recreation count data models. Then instead of OLS, negative binomial regression is applied to build the models. Structured weights were incorporated into the models to reflect the nature endogenous stratification.

The consumer surplus estimates for different models are derived by using the bootstrap method, and statistical tests are used to examine their congruency. The findings of this research in general suggest that there is no significant difference in net economic value on per individual per day basis between TCM and OSCM.

INDEX WORDS: Wilderness Recreation, Zero Truncated Negative Binomial Distribution

ESTIMATING THE DEMAND FOR AND VALUE OF RECREATION ACCESS TO NATIONAL FOREST WILDERNESS: A COMPARISON OF TRAVEL COST AND EXPENDITURE DAY MODELS

by

JIANPING ZHU

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2007

ESTIMATING THE DEMAND FOR AND VALUE OF RECREATION ACCESS TO NATIONAL FOREST WILDERNESS: A COMPARISON OF TRAVEL COST AND EXPENDITURE DAY MODELS

by

JIANPING ZHU

Major Professors: Lynne Seymour

J.M. Bowker

Committee: Gary Green

Abhyuday Mandal

Electronic Version Approved:

Maureen Grasso

Dean of the Graduate School

The University of Georgia

May 2007

DEDICATION

To my parents, Guangrong Zhu and Ruying Gao, who supported my education.

To my mother-in-law, Yumin Wang, who encourages me.

To my wife, Xiongfei Wang, who trusts and loves me.

To my daughter, Chichi Zhu, who makes the world so meaningful to me.

ACKNOWLEDGEMENTS

I am extremely thankful to my major professors, Dr. J.M. Bowker and Dr. Lynne Seymour for their invaluable guidance and encouragement throughout my entire research. I have benefited tremendously from their enormous support.

I would like to thank my committee members Dr. Abhyuday Mandal and Dr. Gary Green for serving in my committee and for their advice and guidance.

I would also like to thank Dr. Don English, National Visitor Use Monitoring Program Director for the USDA Forest Service and Dr. Meghan Starbuck in the Department of Economics and International Business at New Mexico State University for their assistance with data used in my analysis.

My appreciation also goes to Dr. Mingliang Wang for his insightful suggestions and invaluable help.

My sincere gratitude is extended to my mother and mother-in-law, who have been very supportive and encouraging.

Last but certainly not least, I wish to express my sincere thanks to my wife and my daughter for their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The Travel Cost Method (TCM) is the dominant method for recreation demand analysis. The number of trips to a site is assumed to relate to travel cost, travel time and other demographics. One disadvantage of conventional TCM is that it does not recognize potential spatial limitations such as tourists from a significant distance. Bell and Leeworthy (1990) developed an alternative model to deal with visitors coming from significant distances to use principally beach resources. After that, a series of papers (Shaw 1991, Hof and King 1992, Kerkvliet and Nowell 1999) attempted to accommodate the diversity of visitors by focusing on differences in the distance they travel. Instead of modeling annual trip numbers to the recreational site, the number of days spent at the site was modeled. The model was called the on-site cost model (OSCM). In both TCM and OSCM, the response variables are count data based from on on-site surveys. In earlier research, OLS was used to model the data (Ovaskainen et al. 2001) but did not respect the fact that the dependent variables were non-negative integers.

An on-site sampling of visits is a convenient way to collect data for recreation modeling. However, the nature of the sampling and the nature of the raw data exhibit a series of features which give rise to many issues related to estimation and interpretation of the recreation demand parameters (Ovaskainen et al. 2001). First, in both individual TCM and OSCM, the only values the response variable can take are nonnegative integers. Second, all observed visitors have taken at least the current trip, and non-visitors are not observed, so the sample is truncated at zero. Third, sampling on-site often is an example of what is known as choice-based sampling, because frequent visitors have a better chance to be sampled than occasional visitors. This means that the data will be endogenously stratified (Ovaskainen et al. 2001). Fourth, the data is usually overdispersed which means that the variance is greater than the mean, because a few visitors may take many trips and spend lots of days on-site but most visitors may take only a few trips and stay few days on-site.

Truncated count data models have been now routinely applied in single-site recreation demand studies after since their evolution in the late eighties (Shaw 1988, Creel and Loomis 1990, Grogger and Carson 1991, Hellerstein and Mendelsohn 1993). Englin and Shonkwiler (1995) made great progress by developing and empirically applying a truncated, endogenously stratified negative binomial model. However, empirical applications for OSCM with on-site sampling are very few, if any. In this thesis, truncated count data models are applied to a Wilderness recreation visit and valuation problem. In the next section, a brief introduction to the National Wilderness Preservation System (NWPS) and Wilderness related issues will be given.

The nation's Wilderness is an important natural resource and composes a huge recreational asset. The NWPS contained 63 areas in 1994, an 86% increase from the original 54 Wilderness areas designated by Congress in the Wilderness Act of 1964. Today, . Four federal agencies including the Bureau of Land Management, the U.S. Fish and Wildlife Service, the U.S. Department of Agriculture Forest Service, and the National Park Service are involved in the management of Wilderness (Cole 1996, Hendee and Dawson 2004) and there is more than 103 million acres Wilderness.

Wilderness is valued for many reasons and many kinds of Wilderness use exist. The economic value of Wilderness was classified into seven categories by Morton (1999), which includes on-site recreation, community, scientific, off-site, biodiversity conservation, ecological service and passive use benefits. The benefit from on-site recreation use of Wilderness is referred by Morton (1999) as direct use benefit; it derives from recreational activities in a Wilderness area such as fishing, hunting, birdwatching, rafting, backpacking, hiking and camping (Bowker et al. 2005). Bowker et al. (2005) also provided estimates for some of these values based on studies done over the past 30 years. The total recreational use of Wilderness has increased due to population increase and is projected to continue to increase (Freimund and Cole 2000).

The growing demand for Wilderness recreation has led to academic and managerial involvement in estimating social economic impact and benefits. Much has been written about estimating the economic effect of Wilderness recreational service (Bowker et al. 2005). Bowker et al. (2005) identified fourteen published studies that estimated individual consumer surplus for on-site Wilderness recreation. Either the TCM or contingent valuation method (CVM) is employed to estimate consumer surplus and net economic value. The CVM uses a survey as an instrument to elicit an individual's stated willingness to pay for recreational access or other values. The TCM estimates recreational visits based on actual or reported travel behavior and associated actual expenditure (Bowker et al. 2005)..

In the TCM model, the individuals perceive and respond to changes in the travel-related component of the cost of a visit to a recreation site in the same way as they would respond to a change in price (Freeman 1993). One disadvantage of this conventional TCM is that it does not recognize potential spatial limitations (Smith and Kopp 1980) because people from significant distance may have different consumption manners. One-day trips as used by TCM are inapplicable to those coming from significant distance. The often used TCM assumes that trip is intended for the use of recreational site only and not to serve multiple objectives (Bell and Leeworthy 1990). Recreationists who travel a long distance may behave differently from recreationists who travel a short distance in the TCM model. The visitors from great distances quite often are deleted from samples in recreational demand analysis (Morey and Shaw 1990). Bell and Leeworthy (1990) developed a new model, the OSCM, to deal with visitors coming from significant distances to use principally beach resources for extended visits. Instead of modeling the number of trips to the recreational site, the number of days spent at the site is modeled as the dependent variable.

In this thesis, first, the statistical and economical background for estimating over dispersed, truncated, and endogenously stratified count data models are described. Bell and Leeworthy's method is applied to model visitor's total Wilderness days and compared to the TCM on the same data. In the OSCM, visitors' total Wilderness days are modeled as a dependent variable on several demographic and social economical factors. This model includes visitors who are from a significant distance and those from a short distance. Bell and Leeworthy's (1990) Ordinary Least Square (OLS) estimator is problematic, because their model does not respect the fact that the data are integers. In this thesis, a statistical justification about which assumptions are violated is provided by applying OLS regression model to recreation count data models. Then instead of OLS, negative binomial regression is applied to build the models and structured weights were incorporated into the models to reflect the nature of endogenous stratification. The consumer surplus estimates for different models will be derived by using the bootstrap method, and statistical tests will be used to examine their congruency.

CHAPTER 2

THEORETICAL BACKGROUND

This chapter presents first, a description of the statistical theory for estimating recreation demand models. Next, the theoretical concepts necessary to estimate the net economic benefits of Wilderness recreational trips will be given. The last section explains some theoretical background for the TCM and the OSCM.

2.1 Statistical Background

Count data models have been used extensively to model recreation demand (Creel and Loomis 1990, Englin and Shonkwiler 1995, Gurmu and Trivedi 1996, Bowker and Leeworthy 1998, Shrestha et al. 2002, Bowker et al. 2006). These models attempt to explain site visitation as a function of cost, site characteristics, socioeconomic characteristics, and other factors. In count data regression, the response variable is an event count (i.e. a nonnegative integer random variable) and the number of events is assumed to be independent and identically distributed. Hellerstein and Mendelsohn (1993) provided a theoretical basis for using the Poisson distribution to model recreational demand.

2.1.1 The Poisson Distribution

The Poisson probability distribution function is given by:

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad 2.1$$

Where:

y = the number of events (a non-negative integer)

λ = the intensity rate parameter

The intensity rate is the expected number of occurrences that occur during the given interval; i.e. E(y) =λ.

In fact, $E(Y) = Var(Y) = \lambda$. This is an important property of the Poisson distribution and is called equidispersion. Equidipersion is a strong restriciton that often is not met with recreation demand data. Data on the number of trips are often overdispersed relative to the Poisson distribution. As long as the conditional mean is correctly specified, the Poisson maximum likelihood estimator with overdispersion is still consistent, but it underestimates the standard errors and inflates the *t*-statistics in the usual maximum-likelihood output (Ovsskaimen et al. 2001, Cameron and Trivedi 1998). For cases where the overdispersion problem is serious, in order to accommodate this variance in the dependent variable, the negative binomial distribution is widely used to model the data (Ovaskainen et al. 2001, Englin and Shonkwiler 1995).

2.1.2 Negative Binomial Distribution

The negative binomial probability distribution can is given by:

$$P(Y = y) = \frac{\Gamma(y + \alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}(\alpha\lambda)^y (1+\alpha\lambda)^{-(y+\alpha^{-1})} \qquad 2.2$$
$$\alpha \geq 0, y = 0,1,2...$$

Where:

λ =the intensity parameter

Γ = the gamma function which is defined as:

$$\Gamma(z)=\int_0^\infty x^{z-1}e^{-x}dx$$

where z > 0. When z is a positive integer (1, 2, …), the gamma function reduces to the $\Gamma(z)=(z-1)!$.

α = the over-dispersion parameter

A significant α indicates the presence of overdispersion, making the negative binomial model more appropriate than the Poisson. When the overdispersion parameter α is approaching 0 (Appendix A), both $E(Y)$ and $Var(Y)$ are equal to λ, and the Poisson model is appropriate (Cameron and Trivedi 1998, Yen and Adamowicz 1993). Liston-Heyes and Heyes (1999), Shrestha et al. (2002), and Bowker et al. (2006) applied the untruncated negative binomial distribution to model recreational demand based on household survey.

2.1.3 Truncation

If the response variable is the number of days during a recreational visit to a Wilderness area, or the number of trips per year, it must be a nonnegative integer. Second, the distribution of number of Wilderness days is truncated at zero (i.e., zero is never observed), because all surveyed visitors must have taken at least one trip and non-participants are not observed. Therefore, in the empirical models, these factors need to be addressed. Failing to account for truncation leads parameter estimates to be biased and inconsistent because the conditional mean is misspecified (Shaw 1988, Grogger and Carson 1991, Cameron and Trivedi 1998). Following Grogger and Carson (1991), a zero truncated negative binomial distribution takes the form of

$$P(Y=y)=\frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})P(Y>0)}(\alpha\lambda)^{y}(1+\alpha\lambda)^{-(y+\alpha^{-1})} \qquad 2.3$$

$$\alpha\geq 0,\quad y=1,\ 2,\ 3...$$

Where:

$$P(Y = 0) = (1 + \alpha\lambda)^{-(1/\alpha)}$$

$$P(Y > 0) = 1 - P(Y = 0).$$

Yen and Adamowicz (1993) compared welfare measures obtained from truncated and untruncated regressions based on the data collected from bighorn hunting license holders in Alberta, Canada. Zawacki et al. (2000) did a similar comparison by using data from the 1991 National Survey of the Fishing, Hunting and Wildlife Association (FHWAR).

2.1.4 Endogenous Stratification

If the data are collected on-site, the sample is typically endogenously stratified. This is because a visitor's likelihood of being sampled is positively related to the number of trips they make to the site. That is, frequent visitors are more likely to be sampled. This problem (sometimes referred to as choice-based sampling) was first addressed by Shaw (1988) with a camping study, while Englin and Shonkwiler (1995) extended Shaw's analysis with an application of the truncated and endogenously stratified negative binomial model to Wilderness hikers in Washington. If the assumption of equidispersion holds, standard regression packages can be used to estimate a Poisson model adjusted for both truncation and endogenous stratification Shaw (1988).

For the case where overdispersion is significant, the density of the negative binomial distribution truncated at zero and adjusted for endogenous stratification for the count (y) was derived by Englin and Shonkwiler (1995) as:

$$P(Y=y)=\frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\alpha^{y}\lambda^{y-1}(1+\alpha\lambda)^{-(y+\alpha^{-1})}$$
$$\alpha\geq 0,\quad y=1,\ 2,\ 3... \tag{2.4}$$

Except for 2.4, all the other models previously described can be readily estimated using LIMDEP econometric software package (Greene, 2002). Model 2.4 needs to be estimated by a user defined 2-stage routine. In an application to forest recreation in Finland, Ovaskainen et al. (2001) showed that the stratified fit is only slightly better than the non-stratified fits and there are only minor differences in the estimated coefficients and *t*-statistics between the two models. In this research, the model specified in the equation 2.3 will be applied. In all previous studies, few if any incorporated stratified weights into the models. In this thesis, structured weights are applied to incorporate the endogenous stratification into the models.

2.2 Econometric Background

2.2.1 Travel Cost Model

The basic premise of the TCM is that the time and travel cost expenses that people incur to visit a site represent the "price" of access to the site. Thus, the Willingness-To-Pay to visit the site can be estimated based on the number of trips that are made at different travel costs. This is analogous to estimating the Willingness-To-Pay for demand to a marketed good based on the quantity demanded at different prices.

The TCM is among the most popular methods for recreation demand analysis (Peterson and Arnold 1987, Bowker and Leeworthy 1998).The TCM allows for the construction of a demand curve where the number of trips to a site is assumed to relate to cost, time and other demographics (Parsons 2003). If a demand curve can be estimated, the value of site access can be measured. The measure of this value is net Willingness-To-Pay (WTP), or consumer surplus. According to Freeman (1993), the TCM operates on the following six assumptions:

1. Individuals will respond to changes in travel costs in the way they would respond to a change in the access fee;
2. Each trip to the site is primarily for site use;
3. All visits entail the same amount of time on-site;
4. There is no utility/disutility derived from time spent traveling to the site;
5. The wage rate represents the opportunity cost of time;
6. There are no alternative recreation sites available.

The TCM is a demand model for trips to a recreation site by a person over a period of time. The quantity demanded is the number of trips a person takes to the site. The price is the cost of reaching the site, including a person's travel expenses and time cost necessary to make the trip possible. The trip cost alone will not explain an individual's demand for recreation trips. Demand will also depend on factors such as income, age, experience in the recreation activities available at the site, and proximity to other recreation sites. The TCM can take the general form of

$$E(Y_{NT}) = f(X_{FULLTC},\ X_{SUBST},\ X_{SOC},\ X_{INC},\ X_{QUAL}) \qquad 2.5$$

Where:

Y_{NT}= number of trips to recreation sites;

X_{FULLTC} = cost of travel expenses plus opportunity cost;

X_{SUBST} = a vector of trip costs to other recreation sites;

X_{SOC} = a vector of demographic variables;

X_{INC} = yearly income; and

X_{QUAL} = site quality perceived by visitors.

In any demand function, one expects a negative relationship between quantities demanded (Y_{NT}) and price (X_{FULLTC}). Figure 1 shows the conceptualization of consumer surplus. Y_{NT}^{0} is the number of trips demanded by a typical visitor to a recreation site. The area A is individual consumer surplus for accessing the site during the period; i.e., the difference between total Willingness-To-Pay for the trips (area $A+B$) and total trip cost (area B). If the recreation sites are closed, an individual would lose access to the site and consequently the area A. A more general expression for consumer surplus, which is applied to any functional form, is the area under demand curve between individual's current price and the choke price at which trips fall to zero in the model. Mathematically, the consumer surplus is (Parsons 2003):

$$Y_{CS}=\int_{X_{FULLTC^{0}}}^{X_{FULLTC^{choke}}} f(u|\quad X_{SUBS},\quad X_{SOC},\quad X_{QUAL})du \qquad 2.6$$

Where:

$X_{FULLTC^{0}}$ = the individual's trip cost; and

$X_{FULLTC^{choke}}$ = the choke price.

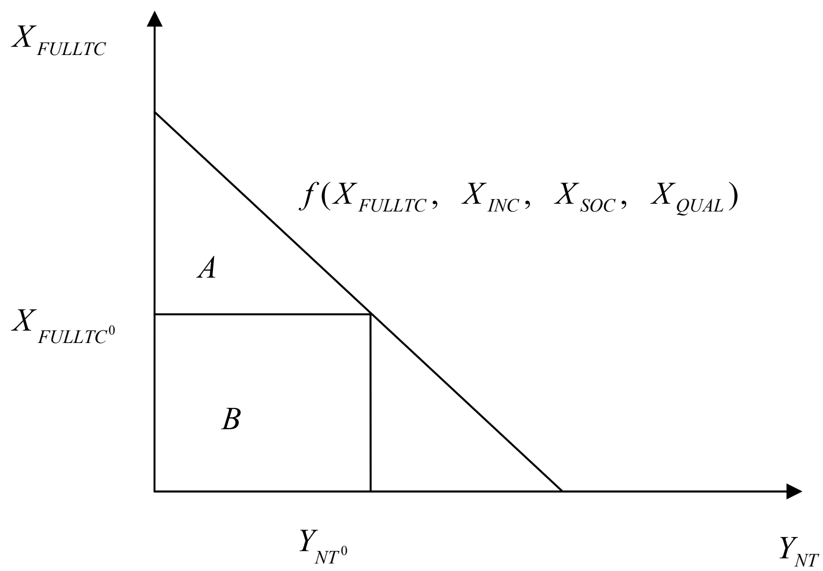

Figure 1. Access Value in TCM Model

2.2.2 On-site cost model

In an OSCM (Bell and Leeworthy 1990), the visitors face two distinct types of cost. It is assumed that the visitors need to pay a certain charge before the consumption of recreation service on site (at the price of X_{OSC}). It can be considered a payment for privilege of purchasing the on-site service.

The OSCM takes the form of

$$E(Y_{DAYS}) = f(X_{OSC},\ X_{TC},\ X_{SUBST},\ X_{INC},\ X_{SOC},\ X_{QUAL}) \tag{2.7}$$

Where:

Y_{DAYS} = days stay on site;

X_{OSC} = actual on-site cost per day (per person);

X_{TC} = total travel cost;

X_{SUBST} = a vector of trip costs to other recreation sites;

X_{INC} = yearly income;

X_{SOC} = a vector of demographic and social economics variables; and

X_{QUAL} = site quality perceived by visitors;

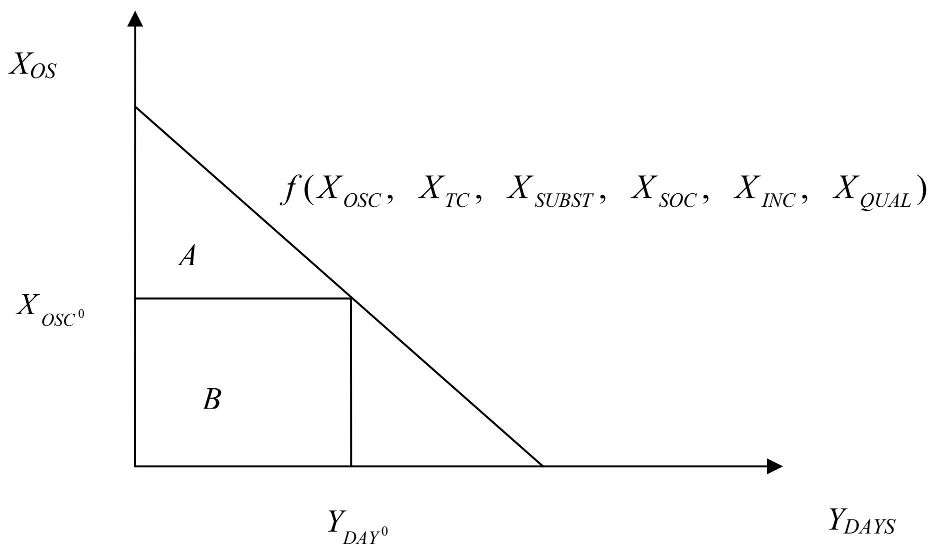

Figure 2. Access in OSCM Model

If a person faces on-site expenses X_{OSC^0} in this model, she or he spends Y_{DAYS^0} days at the recreation site (Figure 2). Y_{DAYS^0} is the number of days demanded by a typical visitor to a recreation site. The area *A* is individual consumer surplus for stays at the site during the period, or the difference between total Willingness-To-Pay for the trips (area *A*+*B*) and total trip cost (area *B*). If the recreation sites are closed, an individual would lose access to the site and consequently to area *A*. Mathematically, the consumer surplus is

$$Y_{CS} = \int_{X_{OSC^0}}^{X_{OSC^{choke}}} f(u| \quad X_{TC}, \quad X_{SUBS}, \quad X_{SOC}, \quad X_{QUAL})du \qquad 2.8$$

Where:

X_{OSC^0} = the individual's on-site expenses; and

$X_{OSC^{choke}}$ = the choke price for expenditures per day.

CHAPTER 3

DATA AND METHODS

In this chapter, a description of the data used for this study is provided first. A summary of variables is given in the second section. The weight variable is described in the third section. The last section describes the models.

3.1 Data

3.1.1 Data source

The USDA Forest Service (FS) began developing the National Visitor Use Monitoring Program (NVUM) in 1998. The purpose of the NVUM is to meet the need for improved information on the recreational use of National Forest System land. From 2000 to 2003 (the four year as a cycle period), data from 120 National Forests and Grasslands (hereafter referred to as National Forests or NFs) are collected via the NVUM program on a continuous basis by using a stratified random sampling procedure (English et al. 2002). While providing a scientific basis to estimate visitation to the NF system and to individual NFs, the NVUM program also puts efforts on an on-site survey to obtain visitor information on the number of annual visits, primary activity, local area expenditures, satisfaction with facilities, and limited demographic information. There are in total 90,542 individual recreation visitor observations in the preliminary dataset for the first cycle of on-site surveying which are from 7,532 different sites aggregated from 120 NFs. The dataset includes more than 200 variables per observation (English et al. 2002).

Several adjustments were made to the preliminary dataset for both theoretical and empirical reasons. Observations from Alaska were deleted, because Alaska visitation is characterized by large numbers of tour groups including cruise liners and buses, as well as large numbers of locals who take numerous daily visits to the NFs, because the physical characteristics of the NFs in Alaska are different from other NFs, and because the majority of nonresidents are not typically visiting the state solely to visit either of the two NFs. Likewise, observations from Puerto Rico were deleted from the master sample due to the significantly dissimilar patterns of recreation relative to the rest of the NFs.

Observations from visitors whose primary trip purpose was not NF recreation were also deleted from the sample in order to avoid violating a basic assumption of the TCM. Observations from incidental visitors were also deleted from the estimation sample because adequate methods have not been developed to apportion the costs of the total trip to the incidental (non-primary purpose) visit. Foreign visitors were not included in the master sample because of the high likelihood of being on a multi-purpose trip and because of the intractability of accurately measuring travel costs.

3.1.2 Dependent Variables

The dependent variable Y_{NFV}, the number of NF visits, for TCM model is created based on the variable $Y_{NFV12MO}$ from the survey, which represents how many times a visitor came to the current National Forest for recreation in the past 12 months (not including the current one). Thus $Y_{NFV}=Y_{NFV12MO}+1$ so that it accounts for the total NF recreation visit times in the past one year. Hereafter, the TCM using Y_{NFV} as the dependent variable will be referred to as the NFV model.

For the OSCM, the dependent variable Y_{WD}, on-site Wilderness days per year, is based on $Y_{TIMESITE}$ and Y_{NFV}. $Y_{TIMESITE}$ represents how long a visitor stays in a Wilderness recreational area during one trip. The result is derived from the answer to the question "On this visit to this national forest, did you go or do you plan to go to any places for recreation other than this one?" If the answer is yes, $Y_{NWILDDAYS}$, which is the total days spent in Wilderness, is used as $Y_{TIMESITE}$. If the answer is no, the day intervals between starting date and ending date is used as $Y_{TIMESITE}$.

The $Y_{TIMESITE}$ is a non-negative integer number. So all negative numbers which most probably are caused by recording error from the dataset are removed. All observed users must have taken at least one trip. Since it is an on-site interview, non-participants cannot get a chance to be observed. Therefore the observations where $Y_{TIMESITE}$ equals to 0 which is most probably caused by recording error are also dropped. Prior knowledge and NF regulations make it very rare for people to stay in Wilderness more than two weeks for a visit. Therefore all records with $Y_{TIMESITE}$ more than 14 are removed. A total of only sixobservations are dropped for these reasons.

Y_{WD} is constructed as $Y_{WD} = Y_{TIMESITE} * Y_{NFV}$ based on the assumption that all reported visits are the same in terms of length of stay and destination. Similar to the NFV model, this model will be referred to as WD model in following sections.

3.1.3 Explanatory Variables

Both the WD and the NFV models have 12 explanatory variables. In this section, a description is given on how some of the independent variables are obtained. The variable X_{EXP50} is the on-site expenditures representing the summation of all on-site expenditures within 50 miles of the Wilderness site on recreation services including government-owned lodging, food/drink at restaurants and bars, gasoline and oil, activities (including guide fee and equipment rental), souvenirs/clothing, private-owned lodging, other food and beverage, other transportation, entry park and or creation use fees, and any other expenses.

The variable $X_{AVGEXPV}$ was created from X_{EXP50} to develop the OSCM ($X_{AVGEXPV} = X_{EXP50} / Y_{TIMESITE}$). Thus, $X_{AVGEXPV}$ represents average expenditure per vehicle per day. The response to the question "On this visit to this National Forest, have you gone or do you plan to go to any places for recreation other than this one?" is $X_{OTHSITE}$. If the answers is yes, $X_{OTHSITE}$ takes the value of 1; otherwise, it takes the value of 0. The response to the question "How many people (including yourself) traveled here in the same vehicle with you?" is $X_{PEOPVEH}$. The response to the question "How many of those people are less than 16 years old?" is $X_{UNDER16}$.

Most Wilderness sites are located in remote areas. It is rare that one can spend more than one thousand dollars per day on-site. By talking with the project manager who designed the survey (English, 2006), some input errors in on-site expenditure data were identified. For example, there are several cases which show that people spent more than $1,000 in a restaurant on a per person per day basis. Those cases are likely to be input errors. Thus, 88 observations, which have a per person per day expenditure greater than $1,000, were discarded.

The variable X_{DIST} represents one way distance from home. The distance is calculated based on the ZIP code of home and the ZIP and latitude-longitude combination of the destination.

The service quality vector $X_{QUALITY}$ has four entries: $X_{SCNERYR}$ which describes the scenery of the visited NF; $X_{PARKAVR}$ which describes availability of parking; $X_{PARKLOTR}$ which describes parking lot condition and X_{CROWD} describes the visitor's feeling about the crowding situation of the visited NF. Because both $X_{PARKAVR}$, $X_{PARKLOTR}$ describe the status of parking lot, a new variable X_{PARK} is formed by averaging them. In this survey, questions about the on-site expenditure and about site quality are taken from different people because of time restrictions.

This means that people who were asked the questions about on-site expenditure were not asked questions about site quality. Therefore, those observations which have on-site expenditure data lack site quality information. To solve this problem, site average for quality by visitors of the same site (with same region, forest, and site number) are plugged into the corresponding recreation with on-site expenditure information.

The variable $X_{PEOPVEH}$ represents number of people per vehicle. People in a large group may behave differently from people in a small group in their on-site consumption. The main purpose for this research is to model the economic behavior of a small group and an individual in a small group. Prior knowledge tells a group larger than nine behaves differently from a group of nine or less in terms of Wilderness consumption. Hence, twenty observations which have $X_{PEOPVEH}$ greater than 9 are removed.

The variable X_{SUBST} records the response to the question “If for some reason you had been unable to go to this national forest for this visit, would you have: (1) Gone somewhere else for the same activity; (2) Gone somewhere else for a different activity; (3) Come back another time; (4) Stayed home; (5) Gone to work at your regular job; or (6) None of them”. The value is 1 if the answer is (1) or (2), and 0 otherwise because (1) and (2) indicate that substitutions exist.

The variable X_{COST} represents total traveling cost ($X_{COST} = 2 * X_{DIST} * X_{COSTPMILE}$). Where $X_{COSTPMILE}$ is the AAA average of average operating cost for an average car from 2000-2003. The information of the average operating cost for an average car from 2000-2003 is not available but the average operating cost for an average car per mile for 2005 (AAA1 2005) is available (14.10 cents) and AAA's annual driving cost estimates (AAA2 2005) from 2000-2005 are also available. Assuming the ratio of average operating cost to annual driving cost across years stays the same, the average operating costs for average cars are estimated over 2000-2003. Then the average of average operating cost for average cars from 2000-2003 is obtained, which is 12.69 cents per mile.

The variable X_{OPCOST} represents opportunity cost ($X_{OPCOST} = (2 * X_{DIST} / X_{MPH}) * X_{WAGE}$). X_{WAGE} is average federal minimum wage per hour from year 2000 to year 2003, which is \$5.15 (Inforplease), and X_{MPH} is average driving speed (Miles Per Hour) which is 50.

X_{FULLTC} is the addition of total travel cost X_{TC} and X_{OPCOST}.

The estimation of recreation benefits based on the TCM are known to be highly sensitive to the magnitude of opportunity cost used (Bockstael et al. 1991). In this research, travel cost and time cost are combined as total cost (Zawacki et al. 2000).

In the NFV model, the variable X_{DHIUSE} is a binary variable. By trial and error, for this data, 18 is chosen as the cut point for Y_{NFV} to make NFV model converge stably. Also by prior knowledge, it is reasonable to classify people who visit NF recreation sites more than 18 times a year as high frequency users. Thus, X_{DHIUSE} takes 1 if Y_{NFV} >18, and 0 otherwise.

The NVUM survey did not collect any income data from respondents. Economic theory suggests that income and substitute prices should be included in travel cost demand models. To provide a proxy for income, U.S. Internal Revenue Service (IRS) data on adjusted gross income, tax returns, and ZIP Code for Tax Year 2002 were used. Thus, income (X_{INC}) is represented by the average after tax income as reported by the IRS for the ZIP Code in which the individual resides.

Table 1. Variables Used in the Empirical Model (By subscript)

Variables	Definitions
EXP50	Total on-site expenditure
DIST	One way travel distance from home to Wilderness site
TC	Traveling cost for two way trip
OPCOST	Opportunity cost
OTHSITE	1 if visit other site, 0 otherwise
INC	Visitor's annual income
GENDER	1 if male, 0 otherwise
AGEGROUP	1. 16-19 2. 20-29 3. 30-39 4. 40-49 5. 50-59 6. 60 and over
PARKAVR	Parking availability 2. Very good 1. Good 0. Average -1. Fair -2 Poor
PARKLOTR	Parking lot condition 2. Very good 1. Good 0. Average -1. Fair -2 Poor
ENVTC	Condition of natural environment 2. Very good 1. Good 0. Average -1. Fair -2 Poor
CROWD	Ranged from 1 to 10 with 1: hardly anyone, 10: overcrowded
PEOPVEH	People who travel in the same vehicle
UNDER16	Number of people in a vehicle whose age is under 16
NFV	National forest visits in past year including current visit
DIUSE	1 if NFV>18, 0 Otherwise
SV	Weighting variable
TIMESITE	Days staying on-site
WD	Total days spent in Wilderness (on per vehicle base) annually

After deleting observations from the original sample (10002 records for NF Wilderness) as described above and deleting all observations with missing values of variables used in final models, an adjusted sample consisting of 1,782 observations remained, referred to as the 'ALL' sample.

3.2 Variable summary

In this survey there are more than 200 variables. The variables either directly obtained or derived from the original data are used in this thesis (Table 1). The variables can be divided into groups such as travel cost variables (X_{DIST}), on-site expenditure variables (X_{EXP50}), demographic variables (X_{GENDER}, X_{AGE}), quality variables ($X_{PARKAVR}$, $X_{PARKLOTR}$, X_{ENVTC}, and X_{CROWD}), and others (X_{INC}, $X_{PEOPVEH}$, and $X_{UNDER16}$).

3.3 Weight

A stratified on-site sampling methodology was developed by English et al. (2002). This method is used to precisely estimate visitation to all NFs in the NVUM project. The purpose of stratification is to help to minimize variance in the estimate of visits up to the NF level. The sampling process can be divided into two stages. In the first stage, a stratified random sample of times and locations where recreation visitors can be counted as they exited the sites was selected to create a set of potential sites and times at which to survey. The sampling methods include four types of site and three user levels which form 12 sampling strata by combination. In the second stage, random draws were selected from the population of sampling days from within these strata and from across the forests to be sampled. Traffic counts were conducted simultaneously with interviews of visitors to adjust traffic counts to the number of unique visits for each sampling time and location. Thus, each sample day site visit estimates were obtained and averaged by strata, and then expanded according to classical random sampling methodology (Cochran 1977).

In this project, the site expansion weights to expand the observations to the stratum is given by

$$SV_h = N_h \sum_{i=1}^{Nh} \frac{C_{hi} P_{hi} V_{hij}}{n_h} \qquad 3.1$$

Where $i = 1, 2, 3, ..., n_h$ is the sampled site-day; $h = 1, 2, 3, ...H$ is the stratum; C_{hi} is the total car count; V_{hij} is the number of persons in the j^{th} sampled vehicle on site-day i;

$$P_{hi} = \sum_{j=1}^{J} \frac{LR_{hij}}{J} \qquad 3.2$$

is the proportion of vehicles on site-day i that were last exiting, with LR_{hij} an indicator variable that equals 1 if the j^{th} vehicle sampled on the ith site-day is a last exiting recreation vehicle and zero otherwise ; and N_h is the total number of site-days in stratum h.

The weight SV is good for WD and NFV models if each of visitors has the same probability to be interviewed. As indicated above, the sampling method for this research is an endogenously stratified sampling method. This means that frequent users have larger possibilities to be selected for interview. Therefore observations are weighted according to the following: $W_h = SV_h /(Y_{NFV_h} * X_{PEOPVEH_H})$ (Zarnoch 2006).

3.4 Models

In this section, detailed information on how to construct the two models for this study is presented. The dependent variables for the two models are Y_{WD} and Y_{NFV} respectively.

Traditionally, the amount of demand for a good or service is a function of price of the good or service, the availability or price of its substitute, the consumer's income, quality factors and other socio-economic factors. The demand function for recreational service has the same structure. Here both models are based on the same available data. The first approach is to model total Wilderness days for people coming in group. Equation 2.3 is the general form. In this thesis, the equation can be expressed as:

$$E(\ln(Y_{WD})) = f(X_{AVGEXPV},\ X_{DIST},\ X_{SUBST},\ X_{INC},\ X_{SOC},\ X_{QUALITY},\ X_{OTHERS}) \qquad 3.3$$

Where:

$E(ln(Y_{WD}))$ is the expected value of $ln(Y_{WD})$;

Y_{WD} = number of days spent at Wilderness area for the group of people who traveled together per year (i.e. group days);

$X_{AVGEXPV}$ = on-site cost per day per vehicle;

X_{DIST} = one way distance of miles traveled from home to site;

X_{SUBST} = 1 if there are substitutes, 0 otherwise;

X_{INC} = average after tax income in observation's zip code;

X_{SOC} = a vector of age, and gender;

$X_{QUALITY}$ = a vector of environmental condition or parking conditions; and

X_{OTHERS} = all variables which are not in above category such as $X_{OTHSITE}$, $X_{PEOPVEH}$, and $X_{UNDER16}$.

It is hypothesized that the expected Y_{WD} will be inversely related to $X_{AVGEXPV}$ because if people have to spend more to access the site, it is more likely that they will spend fewer days there. The expected Y_{WD} will be inversely related to X_{DIST} because local users more likely to use Wilderness more often annually. $X_{AVGEXPV}$ and X_{DIST} are two main factors that affect the Y_{WD}.

The second model is to model total Wilderness visits per year. Similar to equation 3.3, the model takes the form of:

$$E(\ln(Y_{NFV})) = f(X_{FULLTC},\ X_{SUBST},\ X_{INC},\ X_{SOC},\ X_{QUALITY},\ X_{OTHERS}) \qquad 3.4$$

Where:

Y_{NFV} = the total Wilderness visits per year;

X_{FULLTC} = the addition of total travel cost and time opportunity cost; and the rest variables in the model are as same as the ones in the previous model.

One of the main objectives of this research is to test if there is a significant difference in consumer surplus estimation for these two empirical models. The hypothesis to test for consumer surplus from two models is:

H_0: $CS_{wd} = CS_{nfv}$

H_a: $CS_{wd} \neq CS_{nfv}$

Where:

CS_{wd} = average consumer surplus on a per day per person basis for the WD model.

CS_{nfv} = average consumer surplus on a per day per person basis for the NFV model.

CHAPTER 4

RESULTS

4.1 Descriptive Summary

Table 2 contains sample means, both weighted and unweighted for variables used in the WD and NFV models. Large differences between the weighted and unweighted results are observed for some variables. This is not surprising because the weights were built in such a way that different observations have different weights associated with them. For some variables, WD and NFV for example, the weighted mean become smaller because the observations with higher values have lower weights according to the data, therefore the weighted mean become smaller. For others, AVGEXPV for example; the records with higher values have higher weighted value according to the data, so the weighted value is inflated.

Count data distributions are normally characterized by a concentration of values on a few small discrete values (Cameron. and Trivedi, 1998). Histograms of Y_{WD} and Y_{NVF} (Figures 3 and 4) show that the distributions skew to the right. Most visitor groups spend less than 15 days per year on Wilderness sites and most respondents visited the site less than 25 times per year.

Table 2. Means for Explanatory Variable for WD and NFV Variables, N=1782

(By Subscript)

Variable	*Weighted*	*Unweighted*
WD	4.093281	18.998317
NFV	2.565013	15.356902
AVGEXPV	174.530281	91.376353
COST	144.975945	276.833951
FULLTC	266.378565	68.101152
DIST	589.333109	125.128946
INC	47992.222300	44920.259100
SUBST	0.594844	0.396184
GENDER	0.643636	0.654882
OTHSITES	0.426455	0.315937
AGEGROUP	3.459087	3.558923
PEOPVEH	2.634474	2.516835
NUNDER16	0.405703	0.391134
CROWD	4.064438	4.149984
PARK	1.209163	1.154104
ENVT	1.650957	1.596279
DHIUSE	0.017298	0.220

4.2 OLS Estimator

Either visit times to the Wilderness or days staying in Wilderness are count data. Several studies mention that Ordinary Least Square regression (OLS) is not proper for count data models because count data violate the assumptions of OLS (Ovaskainen et al. 2001, Shaw 1988, Cameron and Trivedi 1998) without providing a standard test. Bell and Leeworthy (1990) applied OLS to their original OSCM without any justification of the model although OLS was the standard practice at that time. In this section, OLS regressions are run on both models to test which assumption is violated on the simple linear format and semi-log format. Because the main purpose is to check the validity of the model, the model estimation results from OLS will not be reported.

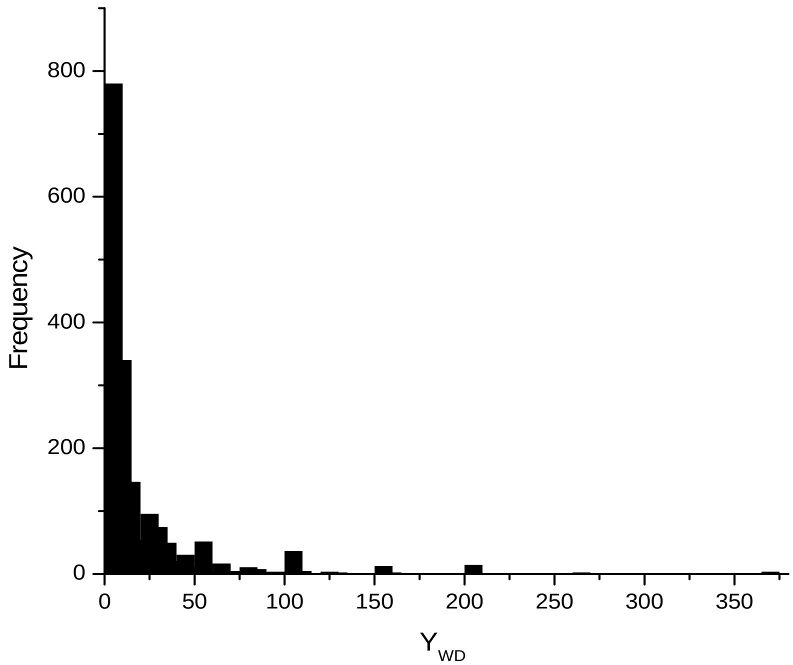

Figure 3. Frequency Distribution of Y_{WD}

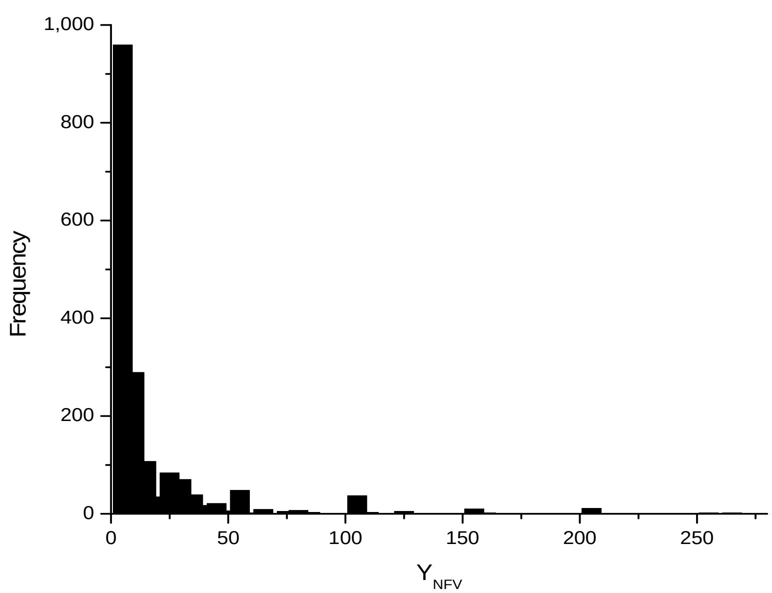

Figure 4. Frequency Distribution of Y_{NVF}

The dependent variable of demand equations 3.3 and 3.4 are changed to WD and NFV (Equations 4.1 and 4.2) respectively as simple linear format , the same format for semi-log model (Equations 4.3 and 4.4) and square root format for square root linear model (Equations 4.5 and 4.6). The square root model is a common variance stabilizing transformation method which can potentially correct the problem that the variance increase as the predicted value increases. Four new models are estimated by using OLS regression.

$$E(Y_{WD}) = f(X_{AVGEXPV}, \ X_{DIST}, \ X_{SUBST}, \ X_{INC}, \ X_{SOC}, \ X_{QUALITY}, \ X_{OTHERS}) \tag{4.1}$$

$$E(Y_{NFV}) = f(X_{FULLTC}, \ X_{SUBST}, \ X_{INC}, \ X_{SOC}, \ X_{QUALITY}, \ X_{OTHERS}) \tag{4.2}$$

$$E(\ln(Y_{WD})) = f(X_{AVGEXPV}, \ X_{DIST}, \ X_{SUBST}, \ X_{INC}, \ X_{SOC}, \ X_{QUALITY}, \ X_{OTHERS}) \tag{4.3}$$

$$E(\ln(Y_{NFV}) = f(X_{FULLTC}, \ X_{SUBST}, \ X_{INC}, \ X_{SOC}, \ X_{QUALITY}, \ X_{OTHERS}) \tag{4.4}$$

$$E(\sqrt{Y_{WD}}) = f(X_{AVGEXPV}, \ X_{DIST}, \ X_{SUBST}, \ X_{INC}, \ X_{SOC}, \ X_{QUALITY}, \ X_{OTHERS}) \tag{4.5}$$

$$E(\sqrt{Y_{NFV}}) = f(X_{FULLTC}, \ X_{SUBST}, \ X_{INC}, \ X_{SOC}, \ X_{QUALITY}, \ X_{OTHERS}) \tag{4.6}$$

One assumption for an OLS regression model to be valid, which needs to be tested, is the normality of the residuals. The quantile-quantile (q-q) plot is a standard graphical technique for determining if two data sets come from populations with a common distribution.

If the two sets come from the same population, the quantiles should fall approximately along a line, the slope of the line equals to ε of underlining normal distribution. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. For simple linear models, the plot of the standard normal distribution quantiles vs. our observed data quantiles shows that it is not approximately a line (Figures 5 and 6), which violates the normality assumption.

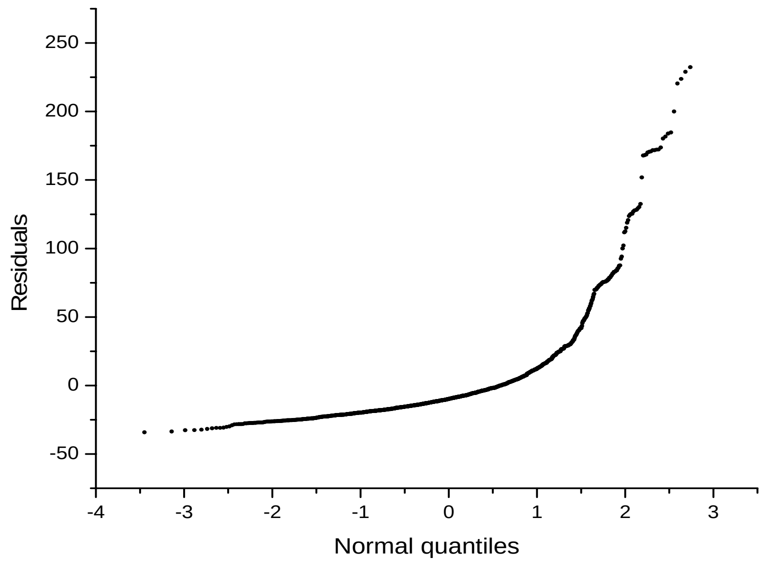

Figure 5. q-q Plot of the Quantiles for the Simple WD model

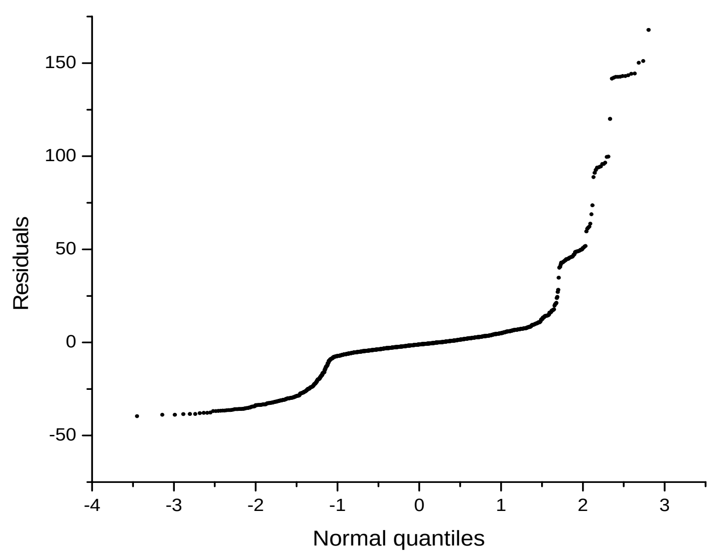

Figure 6. q-q Plot of the Quantiles for the Simple NFV Model

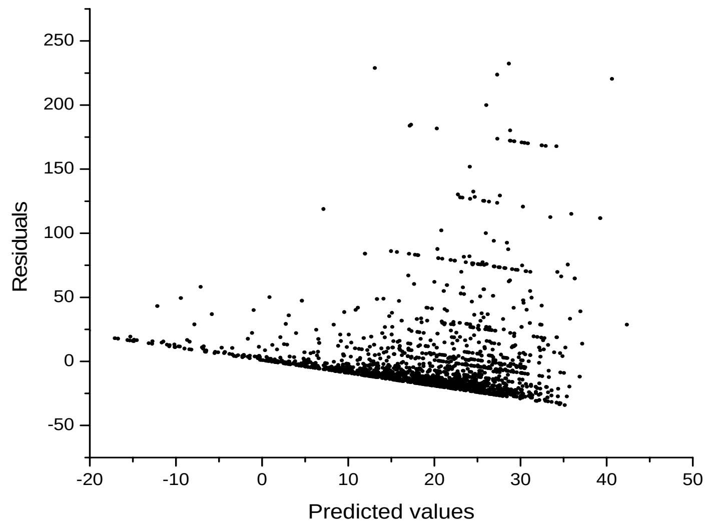

Figure 7. Plot of Residuals vs. Predicted Values Responsible Variable for the Simple WD Model

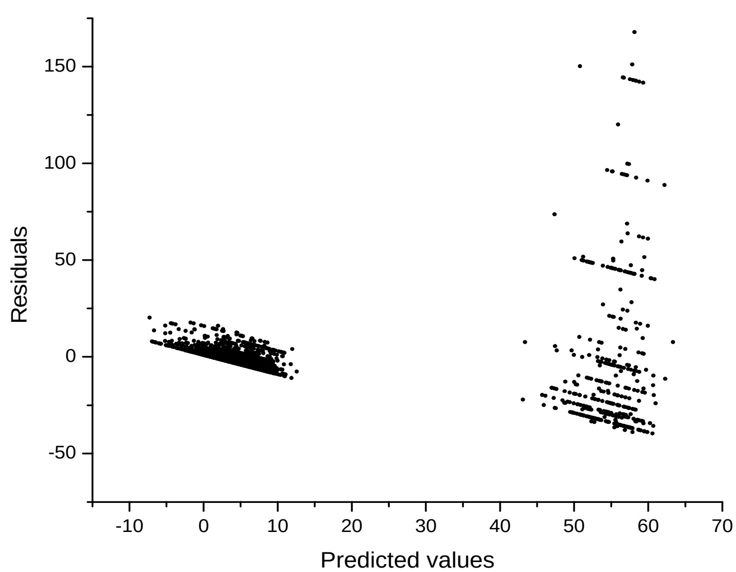

Figure 8. Plot of Residuals vs. Predicted Values of Responsible Variable for the Simple NFV Model

Another assumption to be checked is that the residuals have constant variance. Figures 7 and 8 show the plots of residuals vs. predicted value of response variables. Apparently the residuals do not have constant variance. Therefore this assumption is violated.

Equations 4.3 and 4.4 are semi-log models, with the response variable taking the form of a logarithm. The plot of the standard normal quantiles vs. the observed data quantiles shows approximately a line, implying that the normality assumption holds (Figures 9 and 10). Figures 11 and 12 show the plots of residuals vs. predicted values of the response variable. Here the residuals do not have constant variance. Therefore the assumption is violated.

Equations 4.5 and 4.6 are square root models, with the response variable taking the form of square root. The plot of the standard normal quantiles vs. the observed data quantiles does not show a line, implying that the normality does not hold (Figures 13 and 14). Figures 15 and 16 show the plots of residuals vs. predicted values of the response variable. Here the residuals do not have constant variance. Therefore the assumption is also violated.

The results form simple linear regression, semi-log regression and square root regression indicate that OLS regression is not appropriate for the dataset in this research. The plots of residuals (Figure 8, 10 and16) have "holes" simply because of lacking of predicted values for certain range in the WD related OLS models.

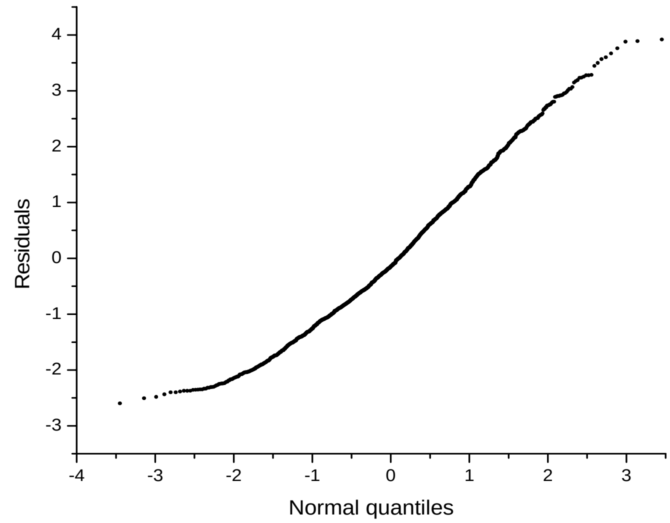

Figure 9. q-q Plot of the Quantiles for the Semi-log WD Model

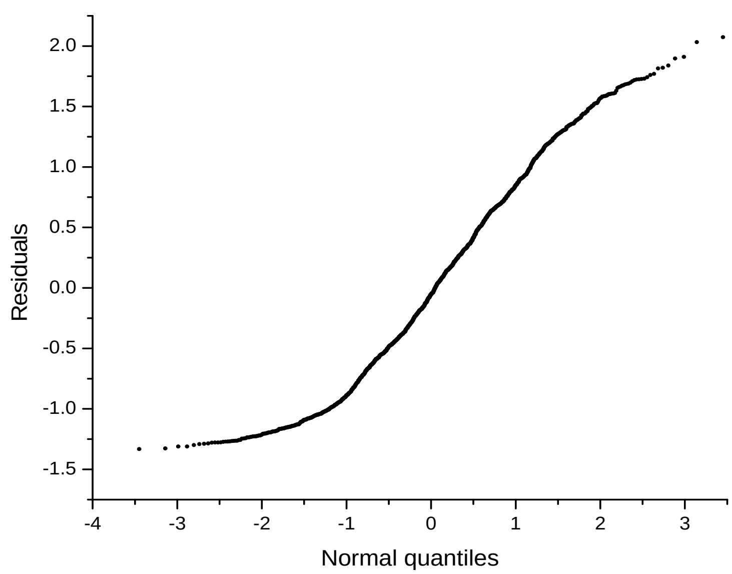

Figure 10. q-q Plot of the Quantiles for Semi-log NFV Model

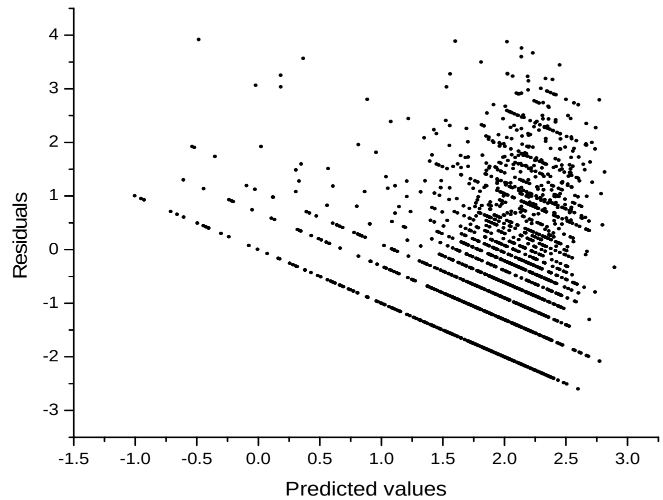

Figure 11. Plot of Residuals vs. Predicted Values of Response Variable of the Semi-log WD Model

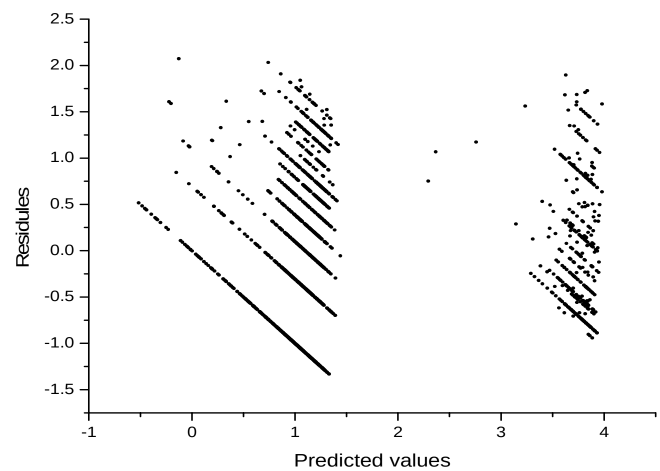

Figure 12. Plot of Residuals vs. Predicted Values of Response Variable of the Semi-log NFV Model

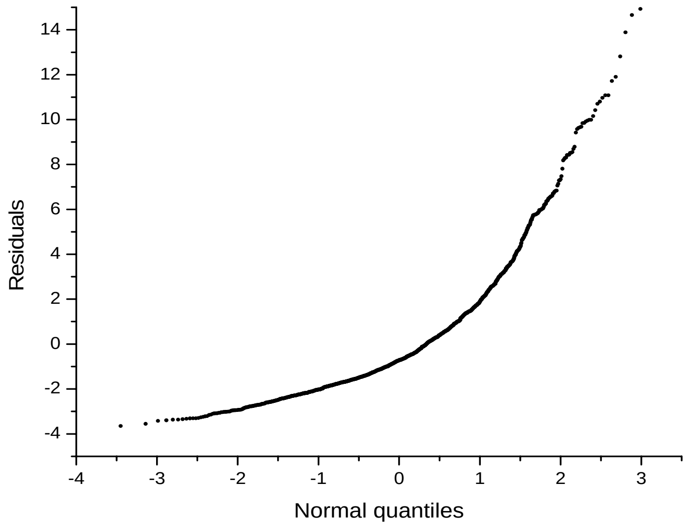

Figure 13. Plot of Residuals vs. Predicted Values of Response Variable of the Square Root WD Model

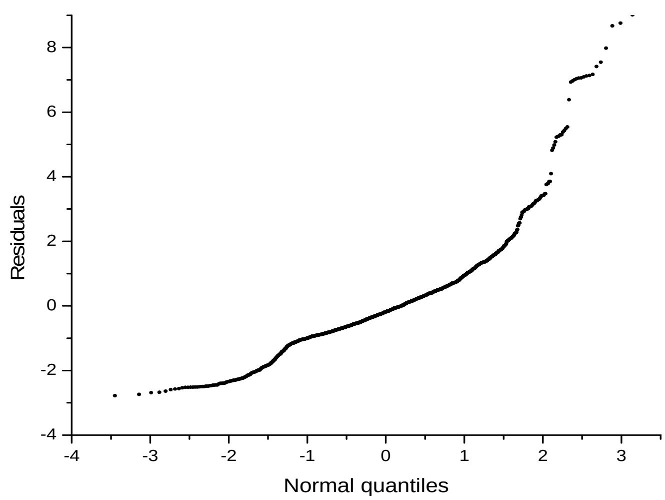

Figure 14. Plot of Residuals vs. Predicted Values of Response Variable of the Square Root NFV Model

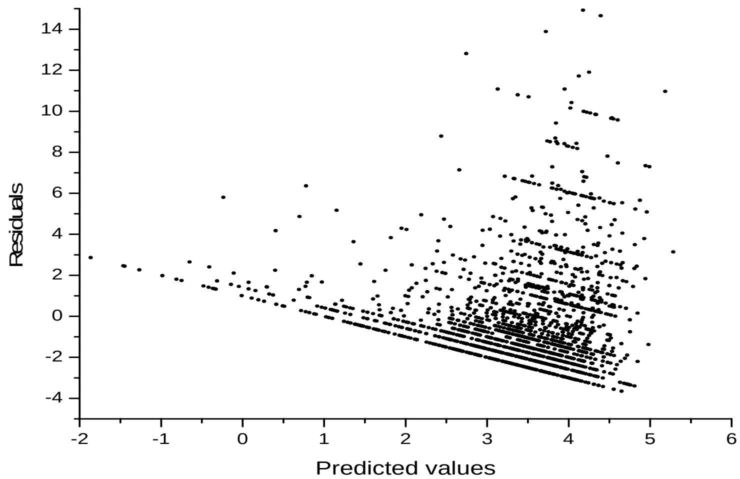

Figure 15. Figure 12. Plot of Residuals vs. Predicted Values of Response Variable of the Square Root WD Model

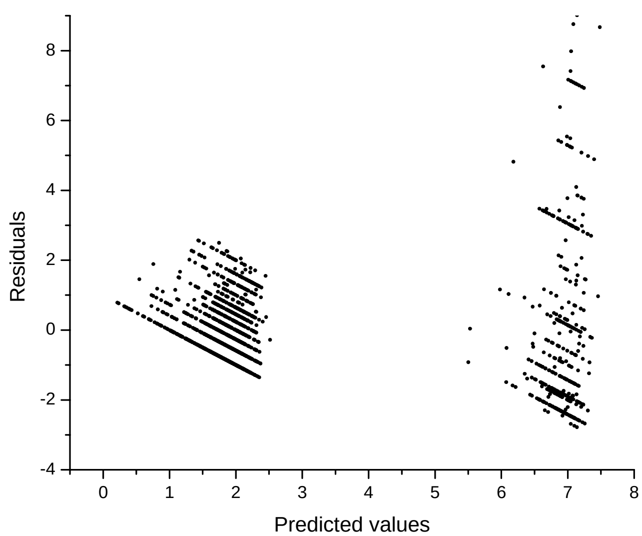

Figure 16. Plot of Residuals vs. Predicted Values of Response Variable of the Square Root NFV Model

4.3 Model Estimation

Both WD and NFV models (Equations 3.3 and 3.4) were estimated using LIMDEP 8.0 (Greene 2002). Results are reported in Tables 3 and 4. Because the negative binomial distribution model is a semi-log model, with the response variable taking the form of a logarithm, the interpretation of parameter estimates is straightforward and can be explained as the percentage change in expected days or visit times per unit change in the explanatory variables.

On the basis of the signs and magnitudes of the coefficients, the different models appear stable. There are no sign changes across models for the main predictors. In both estimated models, the dispersion parameters αs (Tables 3 and 4) are significant confirming that overdispersion is a problem for the models and negative binomial distribution regression is desired for these models.

Failure to Poisson assumption of equidispersion has similar consequence to failure of assumption of heteroskedasticity in linear regression model ((Cameron and Trivedi 1998). But the magnitude of the effect on reported standard errors and t-statistics can be much larger. This problem can be handle by negative binomial distribution because negative binomial does not suppose equidispersion. It can handle overdispersion properly, and overdispersion is not a problem for negative binomial distribution.

In addition to the basic log-likelihood statistic, the pseudo-R^2 (or Likelihood Ratio Index, LRI) $R^2_{LRI} = 1 - LnL / LnL_0$, is reported. This is based on testing the improvement of the fit over a restricted model with only a constant term (with the restriction that all β=0). The R^2_{LRI} for both WD, and NFV models are 0.5306, and 0.2056 respectively (Tables 3 and 4), which indicates a good fit for the WD model and a fair fit for the NFV model.

4.3.1 The WD Model

The WD model can be expressed as:

$$E(Y_{WD} / X) = \exp(1.535424 - 0.001713 * X_{AVGEXPV} - 0.000821 * X_{DIST} - 0.000008 * X_{INC}$$

$$- 0.002594 * X_{SUBST} - 0.102501 * X_{GENDER} - 0.053726 * X_{OTHSITE} - 0.013456 * X_{AGEGROUP}$$

$$- 0.187053 * X_{PEOPVEH} + 0.082800 * X_{UNDER16} - 0.144278 * X_{CROWD}$$

$$- 0.153390 * X_{PARK} - 0.069437 * X_{ENVT})$$

From results for the WD model, one can observe a specific pattern for each explanatory variable in the estimated equations, enabling identification of factors that affect the Wilderness days demanded.

The coefficient of on-site expenditure for $X_{AVGEXPV}$ is negative and statistically significant at 95% level in WD models (Table 4). This means that more expenditure on-site leads to less Wilderness days in a certain period. Holding other factors constant, if a group in the same vehicle spent one more dollar on-site, their expected total Wilderness visiting days per year decrease 0.17% (Table 4).

The coefficient of the variable X_{DIST} is negative and significant, which indicates that the people who come from large distances tend to spend fewer Wilderness days in a vehicle group or in person per year. Holding other factors constant, if a group of persons in the same vehicle travels one more mile in order to get the site, the expected total Wilderness visiting days per year decrease by 0.082% (Table 4).

Theoretically and often empirically, income often has a positive effect in the TCM and the OSCM (Zawacki et al. 2000, Bowker et al. 2006, Bell and Leeworthy 1990). In this model,

X_{INC} has a negative sign and is significant. This counter-intuitive result has been encountered often in the recreation demand literature (Ovaskainen et al. 2001). This means that people with higher income are more likely to spend less Wilderness days per year. This may be because Wilderness is not a favorite place for people with high income to visit or Wilderness visitors tend to be younger or students who do not have high income.

X_{SUBST} does not have a significant effect in the WD model, which indicates that substitution does not have an effect on people's decision on how many Wilderness days they will stay in a vehicle group per year.

Table 3. Grouped Wilderness Day (WD) Negative Binomial Parameter Estimations, N=1782, Dependent Variable Y_{WD} (By Subscript)

Variable	Parameter estimation	Std Error	p-value
CONSTANT	1.535414	0.423206	0.000300
AVGEXPV	-0.001713	0.000049	0.000000
DIST	-0.000821	0.000036	0.000000
INC	-0.000008	0.000001	0.000000
SUBST	0.002594	0.063720	0.967500
GENDER	0.102501	0.061639	0.096300
OTHSITE	-0.053726	0.064444	0.404500
AGEGROUP	0.013456	0.023732	0.570700
PEOPVEH	-0.187053	0.032991	0.000000
UNDER16	0.082800	0.062449	0.184900
CROWD	-0.144278	0.021541	0.000000
PARK	-0.153390	0.060107	0.010700
ENVT	0.069437	0.086285	0.421000
ALPHA	11.361587	4.423931	0.010200
Log L	-3713.747		
Restricted Log L	-7913.087		
Pseudo-R^2	0.5306		

There are two variables in the X_{SOC} vector. One is X_{GENDER} and the other one is $X_{AGEGROUP}$. X_{GENDER} and $X_{AGEGROUP}$ are insignificant in the WD model, which means gender and age does not affect visitor's decision on how many days to stay on site. For X_{GENDER}, this finding is the same as in Kerkvliet and Nowell's (1999) model but different from Bowker et al.'s (2006) model, which has a positive significant X_{GENDER} coefficient. $X_{AGEGROUP}$ often appears in recreation demand models (Bell and Leeworthy 1990, Ovaskainen et al. 2001, Zawacki 2000, Bowker et al. 2006) and the sign and significance varies. This is maybe because people like to visit Wilderness in family group and a family can have both gender or different age group people. .

The $X_{QUALITY}$ vector includes three variables: X_{CROWD}, X_{PARK} and X_{ENVT}. X_{CROWD} is significant and has a negative sign in the WD model, which indicates that if people feel the Wilderness site they visited is crowded, they are more likely to spend fewer days there. This maybe because more Wilderness visitors in favor of quite places. X_{ENVT} is not significant, indicating that the quality of the environment does not affect the number of Wilderness days. This might because there are no significant difference in environment quality among different Wilderness sites. The coefficient for X_{PARK} is significant and has a negative sign, which indicates that if visitors feel that the Wilderness sites have good parking conditions, they are more likely to stay fewer days on-site per year. This counter-intuition result may parallel the fact that Wilderness sites which have good parking conditions are more likely state parks which are close to residential areas. Visitors who like real Wilderness may not like to stay long on these kinds of sites.

The coefficient for $X_{PEOPVEH}$ has a negative sign in the WD model, which indicates that a large group of people are more likely to spend fewer Wilderness days per year on-site. This maybe because large site groups are more likely for the scenery and small groups are more likely to engage in other activity like hunting. The variable $X_{UNDER16}$ is not significant in the WD model maybe because there are no too many groups having people under age 16.

The variable $X_{OTHSITE}$ is not significant in the WD model, which indicates that if the Wilderness site is not a visitor's only goal, she does not care how many days spent there per year.

4.3.2 The NFV model

The NFV model can be expressed as:

$$E(Y_{NFV} / X) = \exp(1.563074 - 0.002568 * X_{FULLTC} + 3.728603 * X_{DHIUSE} - 0.000005 * X_{INC}$$

$$- 0.099664 * X_{SUBST} - 0.035612 * X_{GENDER} - 0.0500577 * X_{OTHSITE} - 0.045972 * X_{AGEGROUP}$$

$$- 0.196094 * X_{PEOPVEH} + 0.081105 * X_{UNDER16} - 0.050457 * X_{CROWD}$$

$$- 0.244109 * X_{PARK} - 0.021276 * X_{ENVT})$$

In the NFV model, X_{FULLTC} is the cost variable and represents the travel cost and time opportunity cost in the traditional TCM. The variable coefficient is significant and has a negative sign, which indicates that there is a negative relationship between total national Wilderness visits and distance from home to the site (Table 5). Holding other factors constant, if a group in the same vehicle spends one more dollar on traveling, the expected total Wilderness visiting times per year decrease 0.26%.

In the NFV model, X_{INC} has a negative sign and is significant. This means that people with higher income are more likely to visit Wilderness sites fewer times per year. This maybe because Wilderness is not a favorite place for people with high income to visit or Lower income visitor more likely to come because of free admissions.

X_{SUBST} is not significant in the NFV model, which indicates that substitution does not have an effect on the decision of how many times to visit a NF per year.

The two variables in the X_{SOC} vector, X_{GENDER} and $X_{AGEGROUP}$, are not significant in the NFV model just as in the WD model.

Table 4. Negative binomial parameter estimates, N=1782, dependent variable Y_{NFV}

(By Subscript)

Variable	Parameter estimation	Std Deviation	p-value
CONSTANT	1.563074	0.316079	0.000000
FULLTC	-0.002568	0.000113	0.000000
DHIUSE	3.728603	0.221036	0.000000
INC	-0.000005	0.000002	0.003600
SUBST	-0.099664	0.077090	0.196100
GENDER	-0.035612	0.080824	0.659500
OTHSITE	-0.500577	0.078885	0.000000
AGEGROUP	-0.045972	0.032896	0.162300
PEOPVEH	-0.196094	0.046665	0.000000
UNDER16	0.081105	0.080099	0.311300
CROWD	-0.050457	0.025441	0.047300
PARK	-0.244109	0.069772	0.000500
ENVT	-0.021276	0.099242	0.830200
ALPHA	2.505852	0.584948	0.000000
Log L	-2399.16		
Restricted Log L	-3020.12		
Pseudo-R^2	0.205607		

X_{CROWD} is significant and has a negative sign in the NFV model, which indicates that if people feel the Wilderness site they visited is crowded, they are more likely to visit fewer times per year. X_{ENVT} is not significant in the NFV model just as in the WD model. And just as in the WD model, the coefficient for X_{PARK} is significant and has a negative sign, which indicates that if visitors feel that the Wilderness sites have good parking conditions, they are more likely to visit fewer times per year.

The variable $X_{PEOPVEH}$ has a negative sign in the NFV model which indicates that large group of people are more likely to visit the Wilderness fewer times per year. The coefficient for $X_{UNDER16}$ is not significant.

The variable $X_{OTHSITE}$ is significant and has a negative sign in NFV model, which indicates that if a Wilderness site is not a visitor's only goal, she is more likely to visit Wilderness sites fewer times per year. This is maybe because people often have plan on how may times or how long to spend for traveling for entire year.

The variable X_{DHIUSE} is significant and has a positive sign. High-frequency users should visit the Wilderness sites more times than low-frequency users. This is am arbitrary variable to differentiate high frequency user from low frequency user.

4.4 Model Diagnostic

Residuals measure the departure of fitted value from the actual value of the dependent variable. For count data, there is no one residual that has zero mean, constant variance and symmetric distribution. Pearson residual can used to correct this heteroskedasticity (Cameron and Trivedi 1998). It is defined as:

$$p_i = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\omega}_i}} \qquad 4.7$$

Where:

p_i=Pearson residual

y_i= The observation

$\hat{\mu}_i$ = The fitted mean

$\hat{\omega}_i$ = Estimation of variance of ω_i of y_i

In large sample, for Poisson and Negative Binomial models, this residual has approximately constant variance but is not asymmetrically distributed (Cameron and Trivedi 1998).

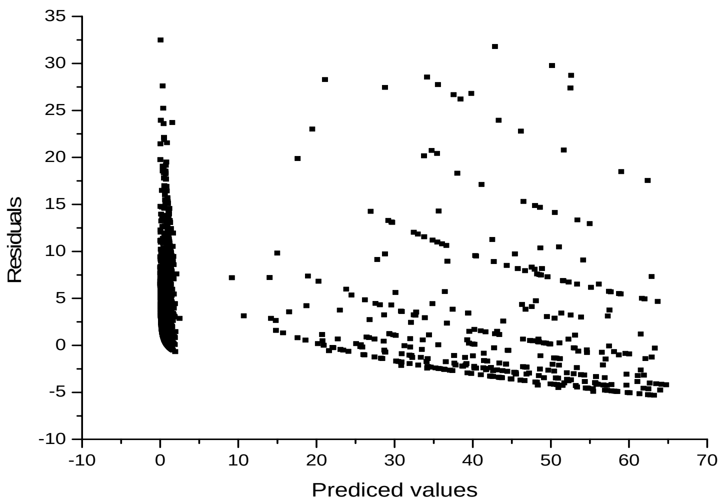

Figure 17.Plot of Pearson Residuals vs. Predicted Values of Response Variable of the NFV Model

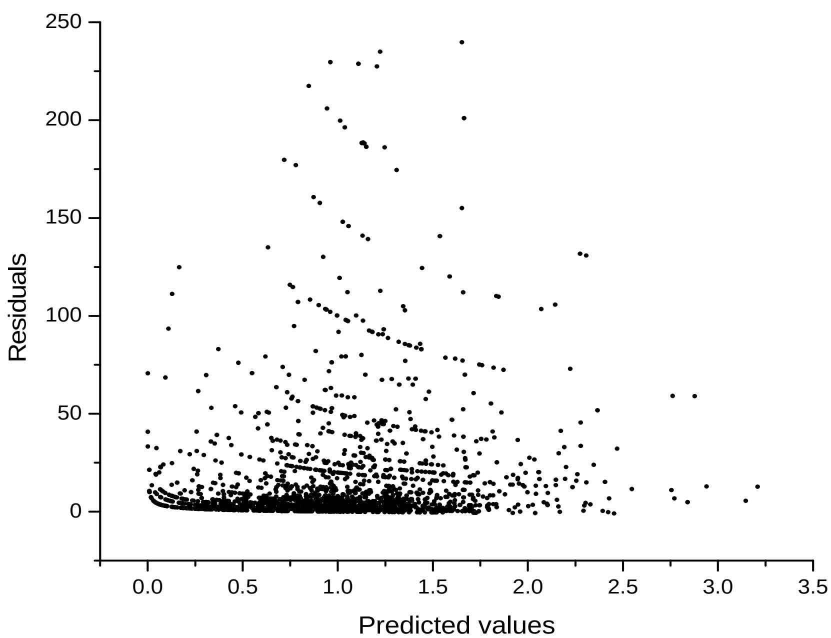

Figure 18. Plot of Pearson Residuals vs. Predicted Values of Response Variable of the WD Model

The plots of Pearson residuals vs. Predicted values show that the Pearson residuals for the NFV model has approximately constant variance (Figure 17). The WD model does not have an approximately constant variance indicating NFV model is a better fit the WD model.

4.5 Model Validation

Model validation is the final step in model-building process. Model validation usually involves checking the model against independent data. Three basic way of validating a regression model are (Neter et.al, 1996):

1. Collection of new data to check the model and it's prediction ability.
2. Comparison of results with theoretical expectation, earlier empirical results, and simulation result.
3. Use of holdout sample to check the model and it's prediction ability.

The preferable method to validate a regression model is through the collection of new data. However, this is neither practical nor feasible in most cases. An alternative when the dataset is large enough is to split the data into two sets. The first set call model-building, is used to develop the model. The model-building data set should be sufficiently large so that a reliable model can be developed. The second data set, called the validation data set, is used to reasonableness of the selected model.

There are total 1,782 observations. Thus this data set can be considered as a large sample. The data set sample was randomly split into two parts. Two thirds of the records (1,188 records) was used as model-building date and the rest was used as the validation data set (Neter et.al, 1996).

After parameter estimation by utilizing the model-building data set, The validation data was used to predict $\hat{Y}$. Residual as $r_i=y_i- \hat{y}_i$ is calculated for both models. The average residuals R was calculated as R= $\frac{\sum_i^n r_i}{n}$. R/ $\overline{Y}$ was also calculated to measure the relative change of average residuals to average response variable values. The NFV is better than the WD because it has a smaller R, Var(R), and R/ $\overline{Y}$ (Table 3).

Table 5. Statistics for model validation

Model	$\overline{Y}$	*R*	$R/\overline{Y}$	*Var(R)*
WD	16.51	15.10	91.46%	1085.30
NFV	13.63	7.78	57.08%	734.71

4.6 Consumer Surplus Estimation

4.6.1 Mean Consumer Surplus

Consumer surplus is a widely accepted measure of net social benefit. It is calculated from the difference between individual willingness to pay and actual payment for purchasing a good or service. Summing this over an entire population yields aggregate consumer surplus.

Across both models, the consumer surplus is the integral of the demand function from the beginning price to the choke price with zero trips. The general form for mean consumer surplus is (Ovaskainen et al. 2001)

$$Y_{cs} = -1/\beta . \quad 4.8$$

where:

Y_{CS} = mean consumer surplus; and

β = the coefficient of the cost variable ($X_{AVGEXPV}$ for the WD model and X_{FULLTC} for the NFV model).

Consumer surplus for recreation access to Wilderness for the WD model is

$$Y_{CS} = \int_{AVGEXPV^{0}}^{AVGEXPV^{CHOKE}} f(u| \; X_{DIST}, X_{SUBST}, X_{INC},$$

$$X_{SEC}, \; X_{QUALITY}, \; X_{OTHERS})du = -Y_{wd} / \beta_{AVGEXPV} \qquad 4.9$$

where:

$X_{AVGEXPV}{}^{0}$ = the actual on-site cost in terms of per person per group; and

$X_{AVGEXPV}{}^{choke}$ = the choke price at which number of days spent in Wilderness falls to zero

The formula for mean consumer surplus per vehicle group per day is

$$\overline{Y}_{CS} = -1/\beta_{AVGEXPV} \quad . \qquad 4.10$$

For the NFV model

$$Y_{CS} = \int_{FULLTC^{0}}^{FULLTC^{choke}} f(u| \; X_{SUBST}, \; X_{INC},$$

$$X_{SOC}, \; X_{QUALITY}, \; X_{OTHERS})du = -Y_{NFV} / \beta_{FULLTC} \qquad 4.11$$

where:

$X_{FULLTC}{}^{0}$ = the actual on site cost in terms of a vehicle group; and

$X_{FULLTC}{}^{choke}$ = the choke price at which trips falls to zero.

The formula for average consumer surplus on a per group per visit basis is

$$\overline{Y}_{CS} = -1/\beta_{FULLTC} \; . \qquad 4.12$$

The average consumer surplus for the WD model is on a per vehicle per day basis, but for the NFV model it is on a per vehicle per trip basis. To make the average consumer surplus from both models comparable, it is necessary to transform these averages $\overline{Y}_{CS}$ into per person per day basis ($\overline{Y}_{CS}^{P}$). For the WD model and the NFV model, average consumer surplus needs to be adjusted by a factor to get $\overline{Y}_{CS}^{P}$; hence,

$$\overline{Y}_{CS}^{P} = \frac{\overline{Y}_{CS}}{C} \qquad 4.13$$

Where C is the appropriate adjustment factor.

For the WD model, $C = X_{PEOPVEH}$, and for the NFV model, $C = X_{PEOPVEH} * X_{TIMESITE}$.

4.6.2 Confidence interval for average consumer surplus

The variance and confidence interval (CI) of $\overline{Y}_{CS}^{P}$ cannot be estimated directly. In the next section, the bootstrap method is used to estimate the CI.

Bootstrap is re-sampling with replacement from a given dataset. Bootstrap theory assumes that the data is the population, rather than a sample from the population. Bootstrap theory further says that an estimator constructed by bootstrapping is a good approximation of the estimate computed from the original data (Venables and Ripley 2002).

In the bootstrap method, large re-samples (greater than 100) need to be drawn in order to make valid conclusions (Venables and Ripley 2002). In this research, 1,000 re-samples were drawn from the original sample with replacement. Formula (4.5) is employed to obtain value of $\overline{Y}_{CS}$ s and formula (4.10) is employed to derive value of $\overline{Y}_{CS}^{P}$. The histograms of $\overline{Y}_{CS}^{P}$ for WD and NFV are shown in Figures 13 and 14. One can see that both figures skew to the right and are not normally distributed. Bootstrap theory allows this histogram to be used to get confidence intervals (at 95% level) for both models. The average, variance, maximum, minimum values and confidence intervals are listed in Table 6.

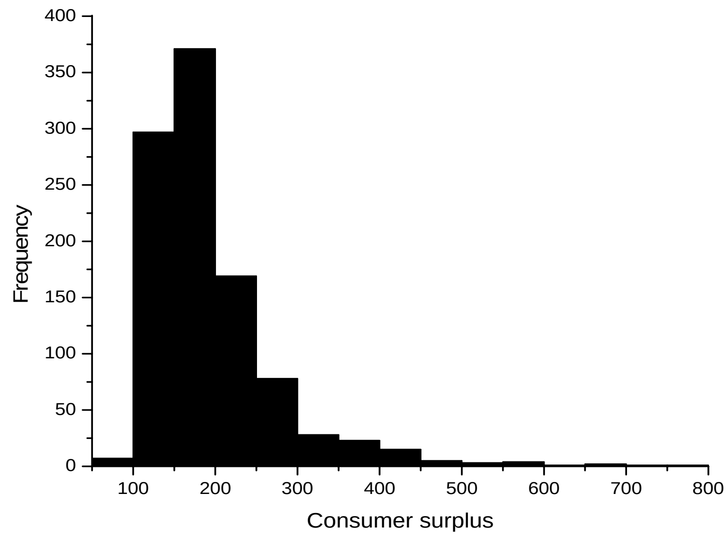

Figure 19. Frequency distribution of $\overline{Y}^{P}_{CS}$ for the WD model

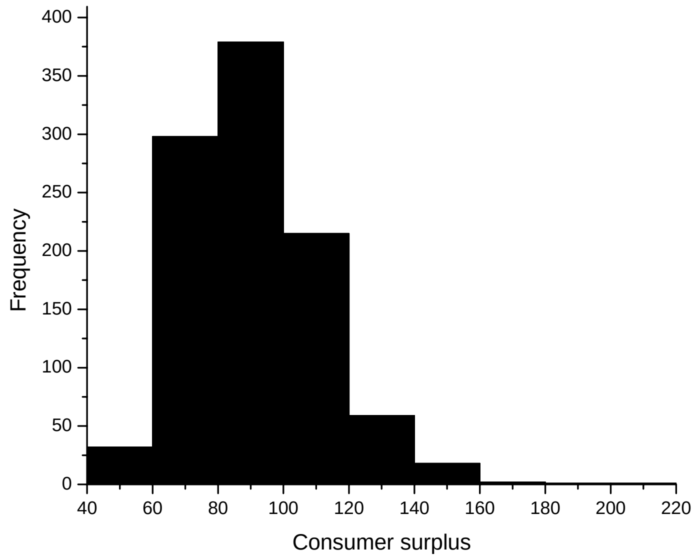

Figure 20. Frequency distribution of $\overline{Y}_{CS}^{P}$ for the NFV model

It can be observed that on an average per person per day basis the confidence intervals for WD and NFV model overlapped (Table 6), which means there are no statistically significant differences between the average $\overline{Y}_{CS}^{P}$ from the WD model and the NFV models. Therefore there is no sufficient evidence to reject H_0: $CS_{wd} = CS_{nfv}$. The significant level is less than 95%.

.

Table 6. Comparison of consumer surplus for the WD model and the NFV model from bootstrap

Model	*Mean* $\overline{Y}_{CS}^{P}$	*Var* $\overline{Y}_{CS}^{P}$	*Max* $\overline{Y}_{CS}^{P}$	*Min* $\overline{Y}_{CS}^{P}$	Confidence Interval($\overline{Y}_{CS}^{P}$
WD	194.7288	6297.18	798.2024	71.4396	[111.257, 416.78]
NFV	91.3794	404.71	205.4387	45.6578	[59.4636,138.055]

CHAPTER 5

CONCLUSION AND DISCUSSION

In this paper, the on-site survey data from NVUM is used to estimate and compare two zero truncated count data models of recreation demand. Bell and Leeworthy (1990) applied the OSCM to a Florida saltwater beach day problem. In this paper, this model is applied to National Wilderness visit days problems and compared to the conventional TCM. The bootstrap method is employed to estimate the distribution and construct confidence interval for average consumer surplus. In the bootstrap method, instead of using the average consumer surplus from the original sample as the estimate of the mean for $\overline{Y}_{CS}^{P}$, the average of $\overline{Y}_{CS}^{P}$ from 1,000 re-samples is used to build a sample distribution for average consumer surplus. The result shows that the average consumer surplus difference between the NFV model and the WD model is not significant. It should be noticed that $\overline{Y}_{CS}^{P}$ is estimated by random sampling without considering the stratified weight. If the stratified weights are incorporated into the random selection process, the bootstrap results from both models might change. Most probably, the results will follow normal distribution because the sample is from average number. If it is the case, the confidence intervals can be obtained based on normal distribution.

Both the variable of cost of travel expenses plus opportunity cost in the NFV model and the variable of actual on-site cost per day in the WD model have negative correlation with the dependent variables respectively as expected, and the variable of one way distance of miles traveled from home to site has a negative sign in the WD model as expected too.

Bell and Leeworthy (1990) utilized the simple OLS estimator to fit the model without any validation for the model. In this research, simple OLS estimation, semi-log OLS estimation and square root OLS estimation were tried, and it is found that the data violate the normality and constant variance assumptions, the log-transformed data violate the constant variance assumption and the square root data violate the constant variance and normality consumption. Therefore, OLS is not appropriate for count data in recreation service demand modeling. Zero truncated negative binomial regression is used to model the demand equation and stratified weights are applied to handle the stratification of sampling.

Bell and Leeworthy (1990) only modeled visitors from significant distance. Conventional TCM only deals with visitors from short distance. The reason is that heterogeneous visitors have long been suspected as a root cause of parameter instability in travel cost analysis (Kerkvliet and Nowell 1999). In this research, all people including those from long and short distance are modeled. The introduction of visitors from both long and short distance increases heterogeneity of visitors in the models and thus may not solve the parameter instability problem. The coefficients of the main predictors show good stability over different models which are estimated, which may be attributable to the use of stratified weights that may help to mitigate the visitor heterogeneity. It will be a challenge for future researchers to investigate this phenomenon further.

OSCM and TCM are applied to the same problem in this work. Which model is better? If the purpose is to carry out a Benefit/Cost analysis, OSCM should be more useful in explaining visitor's behavior, because it uses spending information and travel distance whereas the TCM only uses travel distance information. On the other hand, the statistical validation indicates both TCM and OSCM model do not have very strong predictive ability and the TCM predicted better than the OSCM did. The model diagnosis process also indicates TCM is better than OSCM. The consumer surplus estimated from TCM has lower variance and narrower confidence interval than the consumer surplus estimated from OSCM. Thus, statistically, the TCM is better than OSCM.

One potential problem with this analysis is that a complete dataset with all variables is not available. A compromise of imputing part of variables in the dataset used by simply taking an average from other datasets has to be employed. An alternative is to use a statistical estimation method like the Expectation Maximization (EM) algorithm to handle missing values. In the future, new datasets with all variables might be available. If it is the case, rebuilding the models will surely generate more accurate results.

The prediction abilities of the two fitted models need to be improved. The TCM and OSCM in the research are built based on empirical evidence. For example, all variables are based on literature and experience. The predictive ability might become better if we can combine standard statistical model selection method with empirical evidence.

REFERENCES

AAA1. 2005. Your driving cost 2005.
http://www.aaawa.com/news_safety/pdf/Driving_Costs_2004.pdf.

AAA2. 2005. Despite Higher Gas Prices, 2005 Driving Costs Nearly Unchanged From 2004, AAA Says. http://www.aaanewsroom.net/Main/Default.asp?CategoryID=4&ArticleID=361

Bell, W. F. and V. R. Leeworthy. 1990. Recreational demand by tourists for saltwater beach days. Journal of Environmental Economics and Management. 18:189-205.

Bockstael, N. E., K. E., McConnel, and I.E., Strand. 1991. “Recreation.” in Braden. J.B. and C.D. Kolstad editors:Measuring the Demand for Environmental Quality. Elsevier Science Publishers(North Holland). 227-270.

Bowker, J.M., J.E. III, Harvard, J.C. Bergstrom, H.K. Cordell, D.B.K. English, and J.B. Loomis. 2005. The Net Economic Value of Wilderness. In: Cordell, Bergstrom, Bowker, eds. The Multiple Values of Wilderness. State College, PA: Venture Publishing

Bowker, J.M., and V. R. Leeworthy. 1998. Accounting for ethnicity in recreation demanding: a flexible counting data approach. Journal of Leisure Research.30:64-78.

Bowker, J.M., D. Murphy, H.K. Cordell, D.B.K. English, J.C.Bergstrom, C.M. Starbuck, C.J. Bets, and G.T.Green. 2006. Wilderness and primitive area recreation participation and consumption: an examination of demographic and spatial factors. Journal of Agricultural and Applied Economics. 38:317-326.

Cameron, C.A. and P.K. Trivedi. 1998. Econometric Society Monograph: Regression analysis of count data. New York, NY: Cambridge University Press.

Cochran, W.G. 1977. Sampling Techniques. 3rd Edition. New York, NY: John Wiley.

Cole, D.N. 1996. Wilderness recreation trend, 1965 through 1994. Research paper INT-RP-488. Ogden UT:USDA, Forest Service, Intermountain Research Station.

Creel, M. and J. B. Loomis. 1990. Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California. American Journal of Agricultural Economics. 72:434–441.

English, D.B.K. 2006. Personal communication. USDA Forest Service, Southern Research Station. Economist. September.

English, D.B.K., S. M. Kocis. S.J. Zarnoch, and J.R. Arnoid. 2002. Forest Service National Use Monitoring Process: Research Method Documentation. General technical report SRS-57. Washington DC: U.S. Department of Agriculture. Forest Service.

Englin, J. and J. Shonkwiler. 1995. Estimating social welfare using count data models: An application under conditions of endogenous stratification and truncation. Review of Economics and Statistics 77:104–112.

Freeman, A. M. 1993. The measurement of Environmental and Resources Values: Theory and methods. Washington DC: Resource for future.

Freimund, W.A., and D.N. Cole. 2000. Visitor use density and Wilderness experience. Proceeding of rock mountain research station. W.A. Missoula, MT. Proceedings RMRS-P-20. Ogden, UT: U.S. Department of Agriculture, Forest Service.

Green, W.H. 2002. Limdep Version 8.0 Econometric model guide. New York, NY: Econometric Software, Inc.

Grogger, J. T. and R. T. Carson .1991. Models for truncated counts. Journal of Applied Econometrics 6: 225–238.

Gurmu, S. and P. Trivedi. 1996. Excess zeros in count models for recreational trips. Journal of Business and Economic Statistics 14:469–477.

Hellerstein, D. and R. Mendelsohn. 1993. A theoretical foundation for count data models. American Journal of Agricultural Economics 75 :604–611.

Hendee, J.C. and C.P. Dawson. 2004. Wilderness progress after forty years under the U.S. Wilderness Act. International Journal of Wilderness. 10:4-7.

Hof, G. J. and D.A. King. 1992. Recreational demand by tourists for saltwater beach days: comment. Journal of Environmental Economics and Management. 22:281-291.

Infoplease Federal Minimum Wage Rates, 1955–2006, http://www.infoplease.com/ipa/A0774473.html.

Kerkvliet, J. and C. Nowell. 1999. Heterogeneous visitors and spatial limits of the travel cost model. Journal of Leisure Research. 31:404-419.

Liston-Heyes, C. and A. Heyes. 1999. Recreational benefits from the dartmoor national park. Journal of Environmental Management. 55:69–80.

Morey, E. R. and W. D. Shaw. 1990. An economic model to assess the impact of acid rain, in "Advances in Applied Microeconomics" (V. K. Smith and A. N. Link, Eds.). Jai Press, Greenwich, CN. 5:195-216.

Morton, P. 1999. The economics benefit of Wilderness: Theory and practice. Denver Law Review. 76: 465-518

Neter, J., M.H. Kuter, C.J. Nachtsheim, W. Wasserman. 1996. Applied linear statistical models. Boston: MA. WCB/McGraw-Hill.

Ovaskainen, V., J. Mikkola, and E. Pouta. 2001. "Estimating Recreation Demand with On-Site Data: An Application of Truncated and Endogenously Stratified Count Data Models."

Journal of Forest Economics. 7: 125-144.

Parsons, G. 2003. The travel cost model. In: Patricia A.C., J. B. Kevin, and C.B. Thomas. A premier on non-market valuation. Dordrecht, Netherland: Kluwer Academic Publisher.

Peterson, G. L. and J. R. Arnold. 1987. The economic benefits of mountain running the Pike's Peak Marathon. Journal of Leisure Research.19:84-100.

Shaw, D. 1988. On-site sample regression: Problems of non-negative integers, truncation, and endogenous stratification. Journal of Econometrics 37:211–223.

Shaw, W. D. 1991. Recreational demand by tourists for saltwater beach days: comment. Journal of Environmental Economics and Management 20:284-289.

Shrestha, K. R., A. F. Seidl, and A. S. Moraes. 2002. Value of recreational fishing in the Brazilian Pantanal: A travel cost analysis using count data models. Ecological Economics 42: 289–299.

Smith, V. K., and R. J. Kopp. 1980. The spatial limits of the travel cost recreational demand model, Land Economics. 56:64-72.

Venables, W. N and B. D. Ripley. 2002. Modern Applied Statistics with S. New York, NY:Springer

Yen, S.T., and W.L. Adamowicz. 1993. Statistical properties of welfare measures from count-data models of recreation demand. Review of Agricultural Economics. 15:203–215.

Zarnoch, S. 2006. Personal communication. USDA Forest Service, Southern Research Station. Biometrician. November.

Zawacki, W. T., A. Marsinco, and J. M. Bowker. 2000. A travel cost analysis of nonconsumptive wildlife-associated recreation in the United States. Forest Science 46 :496–506.

Appendix A.

It is not immediately clear that the Poisson is a special case of the negative binomial. To show this, gamma recursive should be used (Cameron and Trivedi 1998).

Let a=1/α the density function (equation (2)) is re-written as:

$$P(Y=y)=\frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}(\alpha\lambda)^{y}(1+\alpha\lambda)^{-(y+\alpha^{-1})}$$

$$=\frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}(\frac{\alpha\lambda}{1+\alpha\lambda})^{y}(1+\alpha\lambda)^{-\alpha^{-1}}$$

$$=\frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)}(\frac{\lambda}{\alpha^{-1}+\lambda})^{y}\left(\frac{1}{1+\alpha\lambda}\right)^{a}$$

$$=\frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)}(\frac{\lambda}{a+\lambda})^{y}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda}\right)^{a}$$

$$=\frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)}(\frac{\lambda}{a+\lambda})^{y}\left(\frac{a}{a+\lambda}\right)^{a}$$

$$=\frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)}(\frac{\lambda}{a+\lambda})^{y}\left(1+\frac{\lambda}{a}\right)^{-a} \qquad \text{(a)}$$

Note that $\Gamma(y+1)=y!$, and that, $\Gamma(z+1)=z\Gamma(z)$, equation (a) can be re-written as:

$$P(Y=y)=\frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)}(\frac{\lambda}{a+\lambda})^{y}\left(1+\frac{\lambda}{a}\right)^{-a}$$

$$=\frac{(a+y-1)(a+y-2)...(a+1)a}{y!}(\frac{\lambda}{a+\lambda})^{y}\left(1+\frac{\lambda}{a}\right)^{-a}$$

$$=\frac{(a+y-1)(a+y-2)...(a+1)a}{(a+\lambda)^{y}}\frac{1}{y!}\lambda^{y}\left(1+\frac{\lambda}{a}\right)^{-a}$$

$$=(\prod_{j=0}^{y-1}\frac{a+j}{a+\lambda})\left(1+\frac{\lambda}{a}\right)^{-a}\lambda^{y}\frac{1}{y!} \qquad \text{(b)}$$

As a→∞, we have:

$$\lim_{n\to\infty}\left(\prod_{j=0}^{y-1}\frac{a+j}{a+\lambda}\right)\left(1+\frac{\lambda}{a}\right)^{-a}\lambda^{y}\frac{1}{y!}$$

$$=\lambda^{y}\frac{1}{y!}\lim_{n\to\infty}\left(\prod_{j=0}^{y-1}\frac{a+j}{a+\lambda}\right)\lim_{n\to\infty}\left(1+\frac{\lambda}{a}\right)^{-a}$$

$$=\lambda^{y}\frac{1}{y!}\lim_{n\to\infty}\left(1+\frac{\lambda}{a}\right)^{-a}$$

$$=\lambda^{y}\frac{1}{y!}\lim_{n\to\infty}\frac{1}{\left(1+\frac{\lambda}{a}\right)^{a}}$$

$$=\lambda^{y}\frac{1}{y!}\frac{1}{e^{\lambda}}$$

$$=\frac{e^{-\lambda}\lambda^{y}}{y!}$$

Because $\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^{n}=e^{x}$

Appendix B.

Source code:

```
?For bootstrap:
sample; all$
draw; n=1782;rep$

?For model WD:
Negbin;
    Truncation;
    lhs=wd;
    rhs=one,avgexpv,dist,inc,subst,gender,othsites, agegroup, peopveh,
        under16, crowd, park,envt;
    wts=wt;
    limit=0;
    model=n
    keep=yfit$

?For model NFV:

Negbin;
    Truncation;
    lhs=nfv;
    rhs=one,fulltc,dhiuse,inc,subst,gender,othsites, agegroup, peopveh,under16,
        crowd, park,envt;
    wts=wt;
    limit=0;
    model=n;
    keep=yfit$
```

Appendix C.

Table: Operating cost for average car by year

Year	Annual driving cost estimates	Annual operation cost estimates
2003	51.7	12.99
2002	50.2	12.62
2001	51.0	12.82
2000	49.1	12.34
Average		12.69