
Munich Personal RePEc Archive

Welfare Estimation Using Aggregate and Individual-Observations Models: A Comparison Using Monte Carlo Techniques

Hellerstein, Daniel

Economic Research Service USDA

1995

Online at https://mpra.ub.uni-muenchen.de/25267/
MPRA Paper No. 25267, posted 23 Sep 2010 16:09 UTC

Welfare Estimation Using Aggregate and Individual-Observation Models: A Comparison Using Monte Carlo Techniques**

Daniel Hellerstein
Final Submission: 4/17/95

Please include following disclaimer.

** Daniel Hellerstein is a natural resource economist with the Natural Resources and Environment Division, Economic Research Service, USDA. The views expressed here are not necessarily those of the Economic Research Service.

Title: Welfare Estimation Using Aggregate and Individual-Observation Models: A Comparison Using Monte Carlo Techniques

Author: Daniel Hellerstein

Abstract

Due to the weak behavioral foundations of aggregate demand models, zonal travel cost models have been largely abandoned in favor of models based on individual observations. However, sample selection difficulties in individual -observation models often require the use of distribution sensitive limited-dependent variables estimators. This paper uses Monte-Carlo simulations to investigate whether the bias from aggregation is worse than possible bias from these narrowly specified estimators. Somewhat surprisingly, the results indicate that zonal models often outperform the individual-observation models, especially when using an aggregate model that incorporates intra-zonal variance of the explanatory variables.

Key words: aggregation bias, zonal models, individual-observation models

Welfare Estimation Using Aggregate and Individual-Observation Models: A Comparison Using Monte Carlo Techniques

At its conception by Hotelling, and in its earliest applications by Trice and Wood, and Clawson, travel cost analysis used aggregate data to compute the contributions to social welfare flowing from peoples' use of recreational sites. This use of aggregate data, whether it be "concentric zones" or political units (such as counties), is best justified as a compromise between theory (of consumer behavior) and practice (the relative abundance and low cost of aggregate data). Although in some cases this compromise is consistent with economic principles, it is often simply a statistical convenience.

In recognition of this theoretic amorphousness, and in conjunction with the increased availability of micro-level data, aggregate models have fallen out of favor in microeconometric analysis (of which travel cost analysis is a branch). In their place, a variety of individual-observation models have been proposed. These models often require the use of limited-dependent variables estimators, in conjunction with assumptions about the conditional probability of behavior, to yield consistent estimates of the determinants of behavior. The classic example is the use of the TOBIT estimator (which assumes an additive and normally distributed random shock) to control for censoring in the linear demand model.

Unfortunately, many of these limited-dependent variables estimators require strong assumptions about the probability distribution of demand. Should these assumptions be untrue, the models will generate biased results. Since aggregate models can frequently be estimated using relatively robust estimators, it is an empirical question as to which modeling strategy is best.

In other words, under ideal circumstances, models based on individual observations models should dominate aggregate models. However, with less than perfect sampling schemes,

it is uncertain as to which class of models is best. On the one hand, use of limited- dependent variables estimators with individual observations can produce bias due to minor misspecifications of the probability distribution of observed demand. On the other hand, aggregate models often introduce an aggregation bias.

Given the complexity of limited-dependent variables estimators, an analytic solution to the question of "which is better" would be exceedingly difficult. As an alternative, this paper adopts an empirical approach, and gauges the relative performance of aggregate and individual-observation models using Monte Carlo simulation. While the results of this examination are strictly true only for the cases evaluated, some general tendencies can be discerned.

To maintain simplicity, a simple Poisson model of demand is used, with demand a function of travel cost and income. A large artificial dataset is generated, and used as the "population" upon which several limited- dependent variables models based on individual observations, and several aggregate models based on zonal measures, are estimated. Coefficient estimates from these models are then used to compute consumer surplus measures, and then these consumer surplus measures are compared to the true consumer surplus for this "population". The degree of deviation from true consumer surplus is used to rank the relative quality of each estimator.

The unexpected result of this exercise is that the aggregate models often dominate the individual-observation models. This is true over a range of coefficient values (representing a range of elasticities), and over several different sampling schemes. Furthermore, this result is enhanced when a misspecification is introduced. Lastly, a modified aggregate model, that uses information on the covariance of the explanatory variables, is shown to provide noticeable improvements over the simpler aggregate model.

Aggregate Models

An aggregate demand model uses summary statistics, such as means and totals, to examine the correlation between the quantity of a good demanded and explanatory factors such as price and income. These summary statistics are from distinct subpopulations, so that a matrix of such measures, with each row of the matrix derived from a distinct subpopulation, comprises an aggregate dataset. A typical example would be aggregating visits to a recreational site at the county level, where total number of visits per county is the dependent variable, and countywide per capita income, average travel cost, and similar "census" measures are the independent variables.

The use of aggregate models can be justified in two fashions. First, aggregate models provide a compact description of the data. Trends can be detected, providing a sense of how demand for the commodity in question varies as the characteristics of the subpopulations vary.

Second, economic analysis requires that the observed data be the result of some underlying demand process that operates at the individual (or household) level. If aggregate data is to be used for formal demand analysis, there must be some process by which the aggregation of individual demands can be consistently modeled using aggregate data. More precisely, one requires that the underlying individual coefficients (the coefficients that condition each individual's demand behavior) be recoverable from coefficients that are estimated with the aggregate model (Stoker, 1993).

The "representative consumer model" is a common set of assumptions used to provide such a basis for the use of aggregate data. The representative consumer model specifies that the aggregate measures used are equivalent to demand, and characteristics, by a mythical representative consumer (Deaton and Muelbauer). This representative consumer is defined so

that the estimation of a model using representative consumers yields results that are equal to the same model estimated using the actual micro-level (individual specific) data. As a simple example, in the linear model with additive error term, a representative consumer model which regresses demand (say, trips from a county) on average characteristics (say, per capita income and average travel cost) can be used to obtain the underlying individual coefficients.

The use of the representative consumer assumptions in demand models has been heavily criticized (Kirman). Proper use of representative consumer models requires that the aggregate measures used be consistent with the underlying (individual) demand curves. In many cases, such as when per capita measures are used in non-linear functional forms, this consistency is not present (Stoker, 1993).

The shortcomings of the representative consumer model suggest that the uses of aggregate data are limited. In particular, if only averages and sums are available, only linear models can be estimated consistently. However, with additional information on the distribution of the aggregate data, the set of models from which underlying (individual) parameters are recoverable can be expanded to include non-linear functional forms. These modified models incorporate additional distributional information, such as on the variance and covariance of subpopulation characteristics, into the function to be estimated.

To illustrate these points, consider the Poisson travel cost model (Hellerstein, 1991). The Poisson travel cost model assumes that the distribution of trip demand for individual i is a non-negative integer, with distribution:

1) $$F_i(n; n=0,..,) = \frac{\exp(-\lambda_i)\lambda_i^n}{n!} \qquad 0$$

where $\lambda_i=\exp(X_i\beta)$, X_i is a kx1 vector of individual i's characteristics (i.e.; travel cost and income), and β is a kx1 vector of coefficients. Note that λ equals the mean, and the variance, of demand. Also, the functional form of λ, $(\exp(x\beta))$, guarantees that the mean and variance are

greater than zero. Lastly, if X includes a constant term, the sum of observed demands will equal the sum of predicted demands.

An additional advantage of the Poisson model is that it yields consistent coffiicient estimates under mean preserving misspecification (Goureiroux, Montfort and Trognonl). For this reason, and for the sake of brevity, in this paper we focus on the Poisson model, and do not discuss other estimators (such as the linear functional forms estimated with TOBIT models).

If there are Iz individuals in a subpopulation z, z=1,..,Z (for example, one of Z counties), the distribution of the total number of trips taken by all Iz individuals in subpopulation z is (Mood, Graybill and Boes, p. 193):

(2) $F_z(N; N=0,..,) = \frac{\exp(-\lambda_{Iz})\lambda_{Iz}^{N}}{N!}$ 0

where $\lambda_{Iz} = \sum_{i=1}^{Iz} \lambda_i$ 0.

If data on individuals (X_i) are not available, but *per capita* data for all z=1,..,Z "counties" are available, the following aggregate model may be estimated:

(3) **Error!** 0

Equation 3 can be justified as an approximation to a representative consumer model. Ideally, the χ_z term would be calculated so that the identity $\lambda_{Iz} = \lambda_z$ is maintained for all z. In this example, since λ_i is convex in X_i, and λ_z is derived from *per capita* data, the identity will not be maintained.[i] In a sense, equation (3) suffers from an "errors in variables" bias.

An improvement to this model is possible if additional information on the distribution of X is available. In general, if the distribution of characteristics is a member of the "exponential family", and the appropriate moments are known, then the underlying (individual) coefficients can be recovered regardless of the functional form of the underlying demand model (Stoker, 1984).

For example, if X is jointly normally distributed, and the mean and covariance matrix of X are known, then the following model can be used to recover β:[ii]

$$F_z(N; N=0,..,) = \frac{\exp(-\Lambda_z)\Lambda_z^N}{N!}$$

where:

(4) 0.

$$\Lambda_z = I_z \times \exp(\chi_z \beta + \frac{\beta' \Omega_z \beta}{2}),$$

$\Omega_z = E[(X_i - \chi_z)'(X_i - \chi_z)]$=covariance of $X_i, \forall i \in z$

Although equation (4) provides consistent estimates of β, it imposes distributional restrictions on the independent variables, and it requires information that may not be readily obtained from typical sources of aggregate data (such as the U.S. Census). Instead of further tinkering with aggregate data, it might be wiser to obtain a sample of micro (individual) data, and directly estimate the underlying demand curves using limited-dependent variables estimators.

Individual Observations using Limited-Dependent Variables Estimators

Individual (or household) data are often combined with limited-dependent variables estimators to examine the correlation between the quantity of a good demanded and explanatory factors such as price and income. These models explicitly recognize that the necessity of gathering a sample (rather than a complete population survey), and intrinsic bounds on the quantity demanded per individual, affects one's estimates. Disregarding these factors can lead to biased estimators (Maddala).

In particular, censoring, truncation, and endogenous stratification are important sources of bias that need to be addressed (Hellerstein, 1992). Censoring arises when the potential quantity demanded is physically limited (e.g., it is impossible to consume less than zero trips). Truncation arises when individuals are only observed when their demand falls within a limited range (e.g., when only participants are sampled, people with zero demand will not be represented). Endogenous stratification occurs when the probability of being sampled is a function of the

quantity demanded (e.g., frequent visitors are more likely to be interviewed on site than infrequent visitors).

When one of these sources of bias is present, an appropriate limited-dependent variable estimator should be used. In general, these limited-dependent variables estimators use information about the probability distribution of an individual's demand to control for such biases. Maximum likelihood estimation is then conducted, based on this probability information, the specifics of the model to be estimated, and the sampling strategy used.

To illustrate these points, consider equation (1), the Poisson travel cost model. The Poisson travel cost model is attractive in that it automatically controls for censoring, since the Poisson distribution is non-zero only over the non-negative integers.

The Poisson assumes that "non-participants" are just as likely to be sampled as "participants", ceteris paribus. For goods, such as visits to a recreational site, that are consumed by a small fraction of the population, the costs of a usable sample (one containing sufficient variation in the dependent variable) may be prohibitive. To circumvent this problem, the sample is often limited to participants.

Since participants are, by definition, non-zero demanders, truncation bias must be corrected. For example, the truncated version of the Poisson is (Grogger and Carson):

$$(5) \qquad F_i(n; n=1,..) = \frac{\lambda_i^n}{(\exp(\lambda_i)-1)n!} \qquad 0.$$

$$\text{with expected value} = \frac{\lambda}{(1-e^{-\lambda})}$$

Equation 5 is valid when each participant has an equal probability of being sampled. In many cases, such as when an on-site sample is collected, frequent visitors are more likely to be sampled, and an endogenous stratification bias must be corrected. For example, the endogenous stratification version of the Poisson is (Shaw):

(6) $$F_i(n; n=1,..) = \frac{\exp(-\lambda_i)\lambda_i^{(n-1)}}{(n-1)!} \qquad 0$$

with expected value=$\lambda+1$

These models, as with many limited-dependent variables estimators, make strong distributional assumptions. If these assumptions are incorrect, the estimators will often be biased. For example, equations 5 and 6 are based on equation 1, which assumes equality of the mean and variance. If this should prove false, equations 5 and 6 will produce biased estimates.[iii]

Comparing Aggregate and Individual-Observation Models

To summarize the preceding discussion, individual-observation models (often analyzed with limited-dependent variables estimators) would seem to provide a better basis for accurate estimation. However, aggregate data is often cheaper and easier to obtain, and aggregate models can often be estimated using more robust functional forms (i.e., with estimators that require fewer assumptions about the structure of the sample or the distribution of random shocks). In short, it is an empirical question as to which source of bias is worse: the error-in-variables type of bias that may occur with aggregation, or bias from model misspecification that individual-observation models may induce.

To investigate the relative performance of these models, this paper uses Monte Carlo simulation. Although analytic results would be more general, they are difficult to obtain.[iv] Simulation, while not offering final answers, can suggest some overall rules.

The basis of each simulation is a known population, with each individual in the population possessing a unique demand for "trips to a park". The population characteristics, and functional form and parameters of the demand curve, are set by the analyst. A variety of samples are drawn from this population, with each sample designed for a particular estimator. Coefficient estimates

for each sample are computed, and they are then used to compute an expected value of consumer surplus. Since all characteristics of the population are known, a "true" expected value of consumer surplus can also be computed. Comparisons of estimator quality are then obtained by comparing the predicted expected value of consumer surplus with the true value, for each sample, over a number of populations.

Each simulation is constructed as follows.[v]

(a)A population of I "simulated" individuals is generated, with each individual randomly assigned to one of Z zones. Using zone specific average non-wage income R_z, average wage income W_z, and average distance to park D_z, an individual non-wage income R_i, wage income W_i, and distance D_i are then generated for each of I_z individuals in a zone (z). Based on W_i and D_i, a travel cost to the park (P_i) is computed (using "gas cost" per mile, and a "time cost" per mile based on a fraction of the wage rate), and a total income (Y_i) is computed from R_i and W_i. The net effect is to create a population that is heterogeneous, but clusters around "zonal" means.

(b)Demand for trips to the park is computed for each individual using a Poisson demand curve (equation 1). Specifically, demand (Q_i) is assumed to be Poisson distributed with $\lambda_i=\exp(X_i\beta)g(\varepsilon_i)=\exp(\beta_0 + \beta_p P_i + \beta_y Y_i)^* (\exp(\varepsilon_i)/\kappa)$; with $\kappa=\exp(\sigma_\varepsilon^2/2)$, which insures that $E[g(\varepsilon)]=1.0$. Note that each simulation uses a different value of β, with each value representing a different kind of park. Also note that ε_i is a normally distributed "misspecification term", so that when $\sigma_\varepsilon=0$, the model is correctly specified.[vi]

(c)All observations are used, and a full information model is estimated:

(ALL OBSERVATIONS) β_{ao} is estimated using equation 1 with all I observations.

(d) For each zone z, the mean vector (X_z) and covariance matrix (Ω_z) of P and Y, and the total demand ($Q_z=\Sigma_{i=1..Iz}q_i$), are computed. Using X_z, Ω_z, and Q_z three "aggregate models" are

estimated:

(SIMPLE AGGREGATE) β_{sa} is estimated using equation 3 and Z zonal aggregates, Ω_z is not used.

(CORRECTED AGGREGATE) A "corrected model" that uses Ω_z. The paramter vector β_{ca} is estimated using equation 4 and Z zonal aggregates.

(PARTIALLY CORRECTED AGGREGATE) Similar to corrected aggregate, but with non-diagonal elements of Ω_z set to zero. The parameter vector β_{pa} is estimated using equation 4 and Z zonal aggregates. This model is used to investigate the effect of using limited information on the intra-zonal variance of characteristics (since covariance terms may not be as readily available as variance terms).

To investigate the effects of incomplete information on aggregate models, two variants of these three models are estimated: one using a complete count of visit per zone, (%PERMITS=100%), and the other using a partial count (%PERMITS=25%).[vii]

e)Three separate samples are drawn from the population, and three individual-observation models are estimated:

(POPULATION SURVEY) A sample of S individuals is drawn at random from the population of I observations. The parameter vector β_{ps} is estimated using equation 1.

(USER SURVEY) A sample of S users are drawn, yielding a sample with truncation at one. The parameter vector β_{us} is estimated using equation 5.

(ON-SITE SURVEY) A sample of S users are drawn "on-site", yielding a sample with endogenous stratification. The parameter vector β_{os} is estimated using equation 6.

It should be noted that the POPULATION SURVEY models will yield consistent coefficient estimates under a mean preservering misspecification. However, the USER SURVEY and ON-SITE SURVEY estimators do not possess this desirable feature.

(f) For each of these seven coefficient vectors, an aggregate expected value of consumer surplus is computed as:

$$ECS_{xx} = \sum_{i=1}^{I} -\exp(,X_i,\beta_{xx},) / \beta_{p_{xx}} \quad 0 \tag{7}$$

where xx refers ao,sa,etc.

As a measure of accuracy, the deviation is computed for each sample, where deviation is defined using the "true expected value of consumer surplus" (ECS*):

$$DEV_{xx} = \text{ABS}(,ECS^*,-,ECS_{xx},) / ECS^* \quad 0 \tag{8}$$

(g) Steps b through f are repeated for R replications. The average (over R replications) of DEV is used to compare estimators (note that lower values of DEV signal better results).

Summarizing, the simple aggregate model is estimated with robust estimators, but might suffer from an errors-in-variables type of aggregation bias due to the use of per capita measures as approximations to measures that would be obtained from a "representative consumer". The corrected and partially corrected models control for aggregation bias by using the intrazonal covariances of the independent variables. Since this information (especially the variances) may be readily obtainable from census sources, this model has practical implications for applied analysis. The user and on-site samples are estimated with models that are sensitive to minor misspecifications in the distribution (i.e., errors in the second moment), but have a strong

theoretical foundation (i.e., based on individual behavior). Lastly, the population survey model uses a robust estimator, and has a strong theoretical foundation, but it may suffer from a lack of variability in the dependent variable due to low overall visitation rates in the population at large.

For purposes of this paper, five basic simulations are reported, with each simulation corresponding to a variation in the sampling scheme and a value for the β vector. These simulations were chosen to cover a range of price elasticities, income elasticities, extent of market, and overall attractiveness.[viii] Table 1 describes these simulations; with averages defined over the entire population (I=50,000). Note that the descriptive simulation names used in table 1 are not to be taken literally; they are only meant to suggest the type of park (in terms of elasticities, attractiveness, and market area) that the simulation might correspond to.

Several variants of each simulation are investigated, representing combinations of the following control parameters:

(a) Degree of misspecification: σ_ε=0, σ_ε=0.5, and σ_ε=1.5. This affects all models.

(b) Sample size for individual-observation models: S=1200 and S=500. This affects the "individual-observation" models only.

(c) Intensity of sampling in aggregate models: %PERMITS=100% and %PERMITS=25%. This affects the "aggregate" models only.

A priori, we would expect the individual-observation models (that use limited-dependent variables estimators) to improve as S increases, all models to worsen as σ_ε increases (with perhaps greater problems for the individual-observation models), and the aggregate models to improve as %PERMITS approaches 100%.

Each simulation is replicated R=100 times, with DEV (equation 8) computed for each model at each replication. The average DEV (over the R replications) is reported in tables 2a through 2c; each table corresponds to a different value of σ_ε. The basic comparison is between

cells of a column, with each cell representing a different method of measuring the welfare contribution of a particular "park".[ix]

Discussion of Results

Examining the $\sigma_\varepsilon=0$ scenarios first (table 2a), the most striking result is that in many cases the aggregate models clearly dominate the individual-observation models. This is especially prevalent for scenarios (i.e., the NATIONAL PARK) in which the average value of λ is low, corresponding to cases where most individuals have zero demand, and the most of the (few) non-zero demanders have a demand equal to one. For example, the simple aggregate model for the NATIONAL PARK has a DEV value less then 0.07. Compare this to the on-site survey (the best of the individual observation models), with DEV values of 1.5 and 1.3 (for S=1200 and S=500 respectively). In such circumstances, where the variation of the dependent variable is very small, it would be surprising if the individual-observation models were able to return accurate estimates. In contrast, the aggregate models will still possess a fair degree of variation in the dependent variable (with nearby zones having larger demand than more distant zones).

Conversely, when demand is relatively large per individual, and when demand is elastic (i.e, the LOCAL NATURE PARK), the individual-observation models outperform the simple aggregate models. In these cases, the intra-zonal variability is high relative to the inter-zonal variability, and the use of per capita models introduces substantial bias. For example, for the LOCAL NATURE PARK, the DEV values for the simple aggregate model and on-site survey model are 0.78 and 0.048 respectively.

It is interesting to note that the performance of the aggregate models are not terribly degraded when less than a full sample of visitors is used to compute the dependent variable. For example, the DEV values for the simple aggregate model equals 0.069 in the 100% PERMITS

case, and 0.080 in the 25% PERMITS case. This suggests that in some cases a partial count of permits, coupled with estimates of total visitation, can be used instead of a full count of permits.

Perhaps the most interesting result concerns the value of information on the covariance of the independent variables. In all cases, the use of the "corrected aggregate" models leads to results that are quite good. For example, the DEV value for the corrected aggregate (100% permits) LOCAL NATURE PARK is 0.026, which is substantially better than the DEV value of 0.78 model for the simple aggregate model. The semi-corrected aggregate models, which use a coarse approximation (diagonal elements only) to the covariance of the independent variables, also offer limited improvements over the simple aggregate models (i.e., 0.27 for the LOCAL NATURE PARK). In a sense, the use of this covariance information protects the aggregate model from the large errors that occur when intra-zonal variability grows (such as in the LOCAL NATURE PARK scenarios).

In the individual-observation models, the most striking result is the previously noted terrible performance when average per capita demand is small. Also of interest is that larger sample sizes did not help much in these "worst case" scenarios (i.e., the NATIONAL PARK), but did help for the other scenarios (i.e., the user survey model of the REGIONAL and STATE parks, where DEV values are approximately halved when S increases from 500 to 1200). Also, somewhat surprisingly, the user and on-site surveys do not seem to dominate the population surveys (even though population surveys can contain many non-participants).

The "misspecified" models (table 2b and table 2c) tell similar stories, except that the users-only models (user survey and on-site survey) are worse in all cases, especially in the highly misspecified (σ_ε=1.5) model. In comparison, the population survey model and the aggregate models are only moderately degraded by increasing levels of misspecification increases. Since the estimators used for users-only models are not robust to a mean preserving misspecification, this result is not surprising; although it is interesting to note the relative size of the bias.

Conclusions

Increased availability of micro data, coupled with well-known problems of aggregation bias, has led to abandonment of the zonal travel cost demand model. The use of individual data is certainly appealing from a theoretical perspective. However, potential problems can arise, especially if one's data collection scheme requires the use of estimators that are highly sensitive to misspecification. Therefore, given a limited data acquisition budget, it is an empirical question as to whether the use of relatively costly survey (micro-level) data will yield more accurate models than models based on more easily obtained aggregate data.

To explore this question, a series of Monte-Carlo simulations were performed. Several different demand simulations were examined, corresponding to a simulated population's demand for trips to several "parks". For each simulation, several individual-observation and zonal-aggregate models were estimated. A consumer surplus measure was then computed for each model, and compared to the known consumer surplus for the population (for that park).

Somewhat surprisingly, these simulations suggest that in many cases aggregate models clearly outperform the individual-observation based models. This is especially true when average per capita demand is small, and when a misspecification is introduced into individual demand curves. Individual-observation models perform well when average per capita demand is high, and when high demand elasticities cause large intra-zonal variation in demand.

The performance of aggregate models was further improved when estimated using the intra-zonal covariance of the explanatory variables (in addition to the intra-zonal means). This improvement was most noticeable in cases where the simpler aggregate models suffered from large biases. Furthermore, even with a limited version of this model, which incorporated only a subset of this covariance data (variances only), some improvement was usually obtained. These positive results are encouraging, since such measures may not be difficult to find. For example,

computing variances in income, and travel cost, from census level block data is one means by which such measures could be incorporated into an applied model.[x]

Although these results may be an artifact of the demand process used and the set of estimators employed, the results are robust enough to suggest that with limited data budgets, there can be an advantage to using aggregate data. This is especially true when the average and variance of demand (across individuals) are small, with little variation in the dependent variable even when truncated individual-observation datasets are used. In these cases, aggregate data may contain a high degree of variation in the dependent variable, which can offset flaws in the representative consumer models that are generally adopted.

Some specific cases are worth mentioning. For "local parks", the simulations suggest that the individual-observation population survey model is a good choice. Furthermore, since the "market area" for such parks is small, obtaining such a sample would be relatively inexpensive. In contrast, the "national parks" are modeled well with aggregate analysis. Since the market area for these parks is very large, and since permit data for these parks is often available, a population survey is not recommended.[xi] Of course, the importance of accurate measures should also be considered when deciding which method to use; for example, if a "state" park of potentially high value is under contention, then a large (expensive) population survey may be appropriate.

These conclusions are somewhat contrary to conventional wisdom. One hesitates to recommend a procedure that is based on questionable variations of the representative consumer model. Nevertheless, these simulations suggest that one must consider the type of site being analyzed, and the relative cost of gathering individual versus zonal data. If these considerations suggest that zonal models may be preferable, the intra-zonal covariance data should be obtained (in addition to the usual average values) and incorporated into an appropriate model (such as equation 4).

Finally, many of these results may apply to other commodities. Although recreational site

visitation provides an interesting set of conditions (with zonal data often easy to obtain and individual data often subject to censoring and other sample selection problems) , there is no major peculiarity that limits these results strictly to travel cost models.

References:

Deaton, Angus and John Muellbauer. Economics and Consumer Behavior. Cambridge: Cambridge University Press, 1983.

Cameron, Colin and Pravin Trivedi. "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests". Journal of Applied Econometrics 1 (Jan. 1986),29-53

Clawson, Marion. Methods of Measuring the Demand for and Value of Outdoor Recreation. Reprint No. 10, Washington D.C.: Resources for the Future, 1959.

Gourieroux, C., A. Monfort and A. Trognon. "Pseudo Maximum Likelihood Methods: Applications". Econometrica 52 (May 1984),701-720.

Grogger, J.T. and R.T. Carson. "Models for Truncated Counts". Journal of Applied Econometrics, 6 (December 1991), 225-238

Hellerstein, Daniel. "Using Count Data Models in Travel Cost Analysis With Aggregate Data". American Journal of Agricultural Economics, 73 (August 1991), 860-866.

Hellerstein, Daniel. "The Treatment of Nonparticipant in Travel Cost Analysis and Other Demand Models". Water Resources Research, 28 (August, 1992), 1999-2004.

Hotelling, Harold. reproduced in An Economic Evaluation of the Oregon Salmon and Steelhead Fishery, Oregon Agricultural Experiment Station Technical Bulletin 74 (1964).

Kirman, Alan. "Whom or What Does the Representative Individual Represent". Journal of Economic Perspectives. 6 (Spring, 1992), 117-136.

Maddala, G.S. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge:Cambridge University Press, 1983

Mood, Alexander, Franklin Graybill, Duane Boes. Introduction to the Theory of Statistics. New York: McGraw Hill, 1974.

Shaw, Dai Gee. "On-Site Samples' Regression: Problems of Non-negative Integers, Truncation, and Endogenous Stratification". Journal of Econometrics 37 (Feb. 1988),211-223.

Stoker, Thomas. "Completeness, Distribution Restrictions, and The Form of Aggregate Functions". Econometrica 52 (July 1984), 887-907

Stoker, Thomas. "Empirical Approaches to the Problem of Aggregation over Individuals". Journal of Economic Literature, 31 (December, 1993), 1827-1876

Trice, A. and S. Wood, "Measuement of Recreation Benefits".Land Economics, 34, Aug. 1958.

Appendix: Derivation of the "Corrected" Aggregate Poisson Model.

Assume that individual i, possessing characteristics X_i, has demand (q_i) which is Poisson distributed with parameter λ_i:

(A1) $f_i(n; n=0,..,) = e^{-\lambda_i}\lambda_i^n/n!$ 0

with $\lambda_i = \exp(\mu_i)$; $\mu_i = X_i\beta$0.

Further assume that for i=1,..,I_j individuals in zone j, X_i is drawn from a multivariate normal distribution:

(A2) $X_i \overset{d}{\geq} N(\chi_j, \Omega_j); \forall_i \in j$0

χ_j = expected value of X_i for all i zone j

where

Ω_j = covariance of X_i for all i zone j 0

Given (A2), μ_i has a univariate normal distribution:

(A3) $\mu_i \overset{d}{\geq} N(M_j, \omega_j)$ *where* $M_j = \chi_j\beta$ $\omega_j = \beta'\Omega_j\beta$0

and $\lambda_i = \exp(\mu_i)$ is a random variable with a lognormal distribution:

(A4) $E(\lambda_i) = e^{M_j + \frac{\omega_j}{2}}$ $var(\lambda_i) = e^{2M_j + 2\omega_j} - e^{2M_j + \omega_j}$0

(note that equation A4 holds for all individuals in zone j).

As the sum of Poisson random variables (q_i), the sum of demand across all individuals in zone j (Q_j) will be Poisson distributed with parameter Λ_j, where:

(A5) $\Lambda_j = \sum_{i=1}^{I_j} \lambda_i$0

The parameter Λ_j, as a sum of random variables, will also be a random variable. Since λ_i are independent and identically distributed:

(A6) $E(\Lambda_j) = I_j \times \left(e^{M_j + \frac{\omega_j}{2}}\right)$ $var(\Lambda_j) = I_j \times \left(e^{2M_j + 2\omega_j} - e^{2M_j + \omega_j}\right)$0

Summarizing, Q_j will be Poisson distributed with parameter Λ_j. The parameter Λ_j is a random variable with mean and variance a function of the expected value and variance of X_i (χ_i and Ω_i, respectively).

Therefore, Q_i is distributed as a "compound" Poisson.

The exact distribution of Q_i requires jointly integrating over the distribution of Λ_i and over the Poisson distribution. Rather than attempt this, the robustness of count models, given that a consistent estimator of the mean is available, can be used (Gourieroux, Montfort and Trognon, Cameron and Trivedi). For the zonal model this mean is $E(\Lambda_i)$, which can be consistently estimated (given A1 and A2 are true). Thus, a consistent (but not efficient) estimate of β can be obtained by using equation A6 in a standard Poisson estimation (that is, use $E(\Lambda_i)$ instead of λ), with Q_i as the dependent variable, and χ_i and Ω_i used as independent variables (in equation A6). This yields equation 4 in the text.

Footnotes

i.. This can be shown using Jensen's inequality for convex functions g: $E(g(x)) \geq g(E(x))$.

ii.. This model is readily derived using the attributes of the log-normal distribution. A complete derivation of the model is provided in the appendix. GAUSS software for estimating this model can be obtained from the author upon request.

iii.. Although more general models can be used (i.e., those based on the negative binomial), some form of distribution is required, leaving open the possibility of bias should reality fail to conform to the analyst's assumptions.

iv.. That is, the exercise of linking unusual probability distributions (i.e., truncated count models) and simpler models based on summary statistics, ideally under some global model, is daunting.

v.. A GAUSS computer program (and documentation) used for these simulation can be obtained from the authors. This program will, given user selected rules, do the following: generate the population, create samples, estimate coefficients for each sample, and compare consumer surplus estimates.

vi.. More precisely, ε_i introduces unobservable individual heterogeneity into the constant term.

vii.. Consider the following case. Aggregate models frequently use "permit" information gathered from all visitors to a site. These (N) permits often contain the town, county, or zip code

of the visitor's home. Zones (z=1,..,Z) are then formed corresponding to these towns, counties, or zip codes; and Q_z is computed by appropriate aggregation. Suppose that only a subset of N_s permits are available, say only the permits from one month (albeit a typical month) of a 3 month season; yielding a %PERMITS of about 33%. These N_s permits account for Q_s ($Q_s=\Sigma_l q_l$; (l=1,..,N_s)) visits. If an accurate estimate of total visits ($Q^*=\Sigma_n q_n$, n=1,..,N) is available, then an approximation to Q_z can be computed by summing over the permits, in this subset, that are from individuals originating in a zone z (N_{sz}), and scaling the number of visits by a correction factor (Cf):

$Q_z \approx (\Sigma_j q_j) * Cf$; j=1,..,$N_{sz}$ and $Cf = Q^*/ Q_s$.

viii.. Many other scenarios, with different β, differences in population distribution, and differences in sampling strategy were also investigated. In the interests of brevity, and since the results did not vary drastically for similar scenarios, only variants of these five are reported.

ix.. There are 15 unique simulations; identified by the 5 types of parks times the three levels of misspecification. Thirteen models are computed for each unique simulation, consisting of:

1) The "full population" model,

2-7) two %permits for each of three aggregate models, and

8-13) two sampling intensities for each of three individual-observation models.

x.. Furthermore, when these models were run (against the simulated data) using values of Ω estimated from a small fraction of the observations, no noticeable changes occured. This suggests that even approximate measures of the variances and covariances would be helpful.

xi.. I thank an anonymous reviewer for pointing this out.