AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

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

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

Help ensure our sustainability.

Give to AgEcon Search

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

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

Aggregation in Recreation Economics: Issues of Estimation and Benefit Measurement

K. E. McConnell and N. E. Bockstael

Problems of aggregation plague applications of microeconomics. The theory is derived from postulates of behavior of individuals, but we often have data only for groups of individuals. The economics of outdoor recreation is no exception. This paper addresses the aggregation issue for estimating the demand for outdoor recreation. What are the estimation and welfare implications of using individual vs. aggregated observations, if the latter is all we have?

The question of aggregation has only recently begun to receive attention, though it is an important issue. Brown and associates (Brown and Nawas, Brown et al.) in several papers have focused our attention on the practical consequences of aggregation. Disaggregation, the use of individual obserations, may increase our ability to make inferences about the coefficients of highly correlated exogenous variables such as the cost of time and income, but it may also increase the measurement error in exogenous variables. Other researchers have explored the implication that aggregation makes the variance of the error term nonconstant (e.g. Bowes and Loomis, Christiansen and Price, Vaughan, Russell and Hazilla). This paper addresses the question of aggregation in terms of estimation of behavioral parameters and calculation of aggregate welfare measures.

The evolution of the travel cost method gives some insight into the aggregation issue and how it developed. The travel cost method was initially proposed as an approach using aggregate data: "Let concentric zones be defined around each park so that the cost of travel to the park from all points in one of these zones is approximately constant. . . . If we assume that the benefits are the same no matter what the distance, we have, for those living near the park, consumer's surplus consisting of the differences in transportation costs. The comparison of the cost of coming from a zone with the number of people who do come from it, together with a count of the population of the zone, enables us to plot one point for each zone on a demand curve for the service of the park." (Hotelling). In fact the development of methods of estimating the demand for recreation so closely paralleled the use of zonal models that the travel cost method is often considered synonymous with the use of zones. However, as the need for benefit estimates in the decision process became more pressing, individual researchers looked for better data sources and closer connections to welfare economics to ground the valuation techniques.

Recreation economics is a product of two legacies: one is derived from the analogy of markets and uses average behavior to gain plausible measures of the value of recreation sites; the other is derived from axions of optimizing behavior and attempts to develop exact welfare measures based on individual behavior. It is over the issue of aggregation where the two legacies come into conflict.

The basic theme of this paper is that regardless of how they are estimated, individual behavioral parameters should be used for welfare measurement. There are two parts to this paper. The first part treats appropriate estimation techniques for data sets which are aggregated over individuals, that is, traditional zonal travel cost models. The second part explores models of individual behavior which incorporate changes in participation as well as changes in the number of visits. Dealing with the aggregation problem involves specifying a framework in which decisions about whether to participate are not confounded with decisions by participants to change their level of

The authors are, respectively, Professor and Associate Professor of Agricultural and Resource Economics, University of Maryland.

Scientific Article No. A-1234, Contribution No. 1234 of the Maryland Agricultural Experiment Station.

participation. Estimation and welfare measurement using individual observations in frameworks which distinguish between the level of participation and decisions to participate are well developed, having been explored in Bockstael, Strand, and Hanemann and in Wetzstein and Ziemer. The second part of the paper is essentially a review and elaboration of these extant models.

The intent of the paper is to explore the issues suggesting the kinds of topics that may warrant further research. One topic worth mentioning at the beginning is the expansion of estimates of trips or benefit estimates to the population. Economists have devoted most of their efforts to issues surrounding the specification and estimation of individual demand curves. A cursory look at the aggregation issue suggests that the plausibility of benefit estimates could be substantially improved if greater care were given to the extrapolation of benefits calculated from a sample to aggregate benefit estimates for the population. While this topic is not addressed in much detail in the paper, it is clearly the direction in which work in this area must proceed.

Using Aggregate Data

Estimation

The travel cost method grew up as a method which utilized zonal averages. The approach is intuitive—the idea that when people must travel further they will incur higher per trip costs and hence be expected to visit a site on average less often. The data demands for the travel cost method in its simple form are not especially great. But the aggregation over individuals, both users and nonusers, makes it somewhat difficult to reconcile the traditional travel cost method with models of individual choice. In this section we analyze aggregate data resulting from discrete decisions in a framework discussed in Maddala (see pp. 182–185).

In this model, we assume that zones are distinct and well-defined. A zone is shorthand for any geographical area, whether determined by political boundaries such as counties or by distance from site as originally conceived by Hotelling. We suppose that there are M such zones, and in each zone i (i = 1, M) there are P_i people (the level of population), n_i of whom visit the site at least once.

In keeping with the idea that we are attempting to estimate the parameters of individual behavioral functions, we hypothesize the functions which lead to aggregate data. Suppose that data are only available from zones, but that these data were in reality generated by the following individual choice model:

$$(1)\qquad q_{ij} = \begin{array}{ll} x_i\beta + \epsilon_{ij}, & x_i\beta + \epsilon_{ij} > 0 \\ 0 & , x_i\beta + \epsilon_{ij} \leq 0 \end{array}$$

where i is the distance zone, j is a user. Thus we assume that individuals in zone i differ by the random error term ϵ_{ij} but have the same arguments x_i. This is a very strong assumption, implying that all people within a given zone have the same income, time costs, etc., which suggests small and homogeneous zones. The error term is distributed as

$$\epsilon_{ij} \sim N(0, \sigma^2).$$

From the above definitions, n_i/P_i is the proportion of the population who visit the site at least once. We shall occasionally denote this rate as π_i, where $\pi_i = n_i/P_i$. Both n_i and n_i/P_i are random because n_i is the realization of random drawings of the disturbance terms described in expression (1). While the individual's model in (1) generates the data, we observe only zonal averages. By convention, let us suppose that the first n_i people are participants and the last $P_i - n_i$ are not. Then when we observe zonal averages, we have

$$\begin{aligned}(2)\qquad q_i &= \sum_{j=1}^{P_i} q_{ij}/P_i \\ &= \sum_{j=1}^{n_i} q_{ij}/P_i + \sum_{j=n_i+1}^{P_i} 0 \\ &= \sum_{j=1}^{n_i} (x_i\beta + \epsilon_{ij})/P_i \\ &= \frac{n_i}{P_i} x_i\beta + \sum_{j=1}^{n_i} \epsilon_{ij}/P_i\end{aligned}$$

This expression gives the nature of the data observed when expression (1) describes the decision process.

Before we examine the stochastic properties and implications for estimation of (2), let us show how we can obtain the traditional zonal travel cost model. Let us ignore the decision process and error structure in (1) and assume, contrary to (1), that the participation rate is constant: $n_i/P_i = \pi$ for all i. Then we can write (2) as

$$(3) \qquad q_i = \pi x_i\beta + \sum_{j=1}^{n_i} \gamma_{ij}/P_i.$$

where γ_{ij} is random component. Since π and P_i are not random in this model, n_i cannot be random either. To make the model consistent with what has been done in the past, the error term must have the following properties:

$$\gamma_{ij} \sim N(0), \sigma^2); E\left(\sum_{j=1}^{n_i} \gamma_{ij}/P_i\right) = 0,$$

which implies

$$E\left(\left(\sum_{j=1}^{n_i} \gamma_{ij}/P_i\right)^2\right) = \frac{\pi\sigma^2}{P_i}.$$

Using these results, we can write the model given in (3) as

$$(4) \qquad q_i = \pi(x_i\beta + \mu_i)$$

where $\mu_i \equiv \sum_{j=1}^{n_i} \gamma_{ij}/P_i$ and $\mu_i \sim N(0, \sigma_o^2/P_i)$. The model in this form is the traditional zonal travel cost model with heteroscedastic errors. (Compare with Bowes and Loomis, for example.) If we know the participation rate, we can correct by dividing by π before we estimate the model. The intuition of such a correction is to convert per capita data to per user data, which is quite acceptable when all users are identical. If we do not know the participation rate, as is usually the case, the π will fall out anyway in computing benefits, at least for the linear demand curve in (1). Thus we see that only when the participation rate is constant across zones can zonal averages be used to estimate parameters of individual behavior. We could complete the zonal travel cost model as traditionally used by computing per capita consumer's surplus from the function in (4):

$$cs^*_i = -(2\beta_1\pi)^{-1}(\pi x_i\beta)^2$$

where β_1 is the slope of the individual's demand function in price space. This expression is accurate only so long as π is constant.

Of course, the constancy of π is a fiction. It violates the model given in (1), which provides an explanation for why some people are, in fact, nonparticipants. Thus the traditional travel cost model outlined above cannot hold when the decision process is given in (1). The participation rate cannot be constant and non-random, because it is determined in part by random errors and in part by systematic variation in factors such as travel cost. Hence, because of this the error in (3) violates Gauss-Markov assumptions and we must find an alternative to OLS.

There are two problems which stem from aggregate data and from the decision process of (1). First, n_i is not observed and, because of (2), is random. Second, the expectation of $\sum_{j=1}^{n_i} \epsilon_{ij}/P_i$ is not zero. In order to estimate the relationship, we can try to put (2) in the form of

$$(5) \qquad q_i = Eq_i + \theta_i.$$

This expression is an identity, created by finding the deterministic part of q_i, Eq_i, and then finding the error by subtraction, $\theta_i \equiv q_i - Eq_i$. The advantage of writing the expression as (5) is that it allows the application of nonlinear least squares. Note that by definition $E\theta_i = E(q_i - Eq_i) = 0$. Following Maddala we can show that

$$(6) \qquad Eq_i = F(x_i\beta/\sigma)\, x_i\beta + \sigma\, f(x_i\beta/\sigma)$$

where $F(\cdot)$ is the cumulative distribution function and $f(\cdot)$ is the density function of a variate distributed N(0, 1). Equation (6) is the nonstochastic part of the nonlinear regression (5). Combining (5) and (6) yields

$$(7) \quad q_i = F(x_i\beta/\sigma)\, x_i\beta + \sigma\, f(x_i\beta/\sigma) + \theta_i$$

where by definition $E\theta_i = 0$, $E\theta_i\theta_j = 0$. There is, however, a problem of heteroscedasticity in the θ_i. Nevertheless, expression (7) can be used to estimate the parameters of the individual behavioral model: β and σ^2. The estimates of these parameters based on (7) will be consistent but not efficient unless some account is made for the heteroscedasticity.

From the perspective of recreation economics, we see that (7) differs from the traditional model (estimated with distance zones) by two quantities. The factor $F(x_i\beta/\sigma)$ is the probability that an individual with arguments x_i will have positive trips. This is the participation rate and will arise whether we use the truncated model (1) or simply assume that the participation rate is constant. The term $\sigma\, f(x_i\beta/\sigma)$ enters only when the decision process is outlined as in (1). If the participation rate were constant across zones, then the second term would disappear, but as we see from (4), the first term would stay.

The troubles with the classical zonal approach to estimating the demand for recreation are clear from the developments of this section. First we assume that individuals within each zone are identical except for a random error. This assumption is violated by what we know, for example, about the distribution of income, the cost of time and other influences on individual behavior. Second, to use OLS in the traditional way on aggregates of zones, we must be assured that the participation rate is constant across zones. If we have reason to believe, as seems likely, that participation rates vary across zones, then we must choose carefully an estimation approach which accounts for this variation.

Using Individual Observations

In this section we deal with the problem of estimating demand functions and computing benefits for a site using samples of individuals. We consider two kinds of samples: first, a sample of the population which contains users and non-users; second, an on-site sample of users only. These problems have been considered in detail elsewhere. Bockstael, Strand and Hanemann estimate equations for both kinds of data. Wetzstein and Ziemer treat the problem of having only data gathered on-site. The on-site data problem is also explored in Smith, Desvousges, and McGivney. The analysis that follows is a generalization on and elaboration of these works. Both kinds of samples are handled by Heckman's analysis of sample selection bias as specification error.

Estimation

We begin with a model of individual behavior slightly more general than the model examined in the previous section. Let

$$\pi^*_j = x_{1j}\beta_1 + \epsilon_{1j} \qquad (8)$$

be an indicator of individual participation. (The index j now refers to individual j. It will be dropped unless it is needed for clarity.) If $\pi^*_j > 0$, then the individual participates; if $\pi^*_j = 0$, the individual does not participate. We may think of π^* as a latent variable measuring desire to visit the site. Let trips to the site be determined by

$$q_j = x_{2j}\beta_2 + \epsilon_{2j}. \qquad (9)$$

Again, the observation index j stands for individual j. The random terms ϵ are assumed to be distributed $N(0, \Sigma)$ where Σ is not necessarily diagonal.

Note that if $x_{1j} = x_{2j}$, $\beta_1 = \beta_2$ and $\epsilon_{1j} = \epsilon_{2j}$, we have the same model as in equation 1. This is a more general model allowing for different factors to affect the two decisions. For most utility theoretic models, however, the x_{1j} would appear as elements in the x_{2j} vector. Several interesting behavioral and statistical characteristics are embodied in this model. The characteristics of this model stem from the distributions of π^* and q. Because π^* is an index indicating participation, q is observed only when $\pi^* > 0$. Hence the model is not a standard linear model. Instead (Heckman, Sections I and II):

$$E(q|x_2, \pi^* > 0) = x_2\beta_2 + E(\epsilon_2|\epsilon_1 > -x_1\beta_1)$$

and

$$E(\pi|x_1, \pi^* > 0) = x_1\beta_1 + E(\epsilon_1|\epsilon_1 > -x_1\beta_1)$$

Let $\lambda \equiv f(x_1\beta_1/\sigma_{11})/F(x_1\beta_1/\sigma_{11})$. Then using Heckman's results we can show that

$$\pi^* = x_1\beta_1 + \sigma_{11}\lambda + \theta_1 \qquad (10)$$

$$q = x_2\beta_2 + \frac{\sigma_{12}}{\sigma_{22}}\lambda + \theta_2 \qquad (11)$$

where $E\theta_1 = 0$.

The structure in (10) and (11) reduces to a variety of special cases used in recreation economics. First, when Σ is diagonal, ($\sigma_{12} = 0$), we have a modified two step model first popularized by Davidson, Adams, and Seneca. In this approach two equations are estimated, one for participation and one for days per user. Consider (10) and (11) when Σ is diagonal:

$$\pi^* = x_1\beta_1 + \sigma_{11}\lambda + \theta_1$$

$$q = x_2\beta_2 + \theta_2$$

This is a modified version of the Davidson, Adams and Seneca approach because of the additional argument $\sigma_{11}\lambda$ in the probability of participation equation.

A second special case occurs when $x_1 = x_2$, $\beta_1 = \beta_2$ and the distribution of the ϵ's is singular. These assumptions yield the Tobit model:

$$E(q|x_2, q > 0) = x_2\beta_2 + E(\epsilon_2|\epsilon_2 > -x_2\beta_2). \qquad (12)$$

This model has been explored in Bockstael, Strand and Hanemann, and in Wetzstein and Ziemer.

Let us examine the general case of (10) and (11) further. The denominator of λ, $F(x_1\beta_1/\sigma_{11})$, is the probability that an individual participates at the site which we denote π. If the probability is constant across individuals, then OLS applied to (11) will not suffer from sample selection through the participation process. Further, if there is a very high rate of participation among the population, λ will be small, and OLS estimates not so bad. The sample selection problem is most severe when there is a very low participation rate and λ is very high. The presence of λ allows for the possibility of considerable misspecification. A variable which actually belongs in the participation equation but which is put in the trips equation may appear significant if λ is improperly omitted. The omission of λ will cause the estimates of β_2 to be biased where λ is correlated with any dimension of x_2.

Several estimation techniques are available for (11) and (12). (See Maddala for details.) Here we outline a mixed maximum likelihood-least squares model because it gives some insight into the deficiencies of OLS. Full ML estimates of (10) are obtained in Bockstael, Strand and Hanemann.

In the full sample case, a two step procedure can be used. First, one can estimate β_1 by maximizing the likelihood function

$$L = \prod_{j \notin S} \left[1 - F\left(\frac{-x_1\beta_1}{\sigma_{11}}\right)\right] \prod_{k \in S} F\left(\frac{-x_1\beta_1}{\sigma_{11}}\right)$$

where S is the set of nonparticipants ($\pi^* = 0$). Note that this likelihood function is based only on the participation decision and requires a sample of the population. Using the estimates of β_1, compute λ for each observation and use it as a regressor with the parameter $\frac{\sigma_{12}}{\sigma_{11}}$ in the least squares equation (11).

For the regression

$$(13) \qquad q = x_2\beta_2 + \frac{\sigma_{12}}{\sigma_{11}}\lambda + \theta_2$$

the variance of θ_2 is heteroscedastic, so that OLS estimates are not efficient.

To complete this section on individual observations, we address the common case of a sample of users only. This case is perhaps the most typical for recreation research. For this case, we are forced to abandon the hypothesis that participation decisions are motivated differently from decisions about the level of activity.

For an on-site sample of users only, the basic structure is

$$(14) \qquad \begin{array}{ll} q = x\beta + \epsilon, & x\beta + \epsilon > 0 \\ \quad\; 0 & x\beta + \epsilon \leq 0 \end{array}$$

where ϵ is distributed $N(0, \sigma^2)$. This is the same behavioral model analyzed in the previous section. Bockstael, Strand and Hanemann give the ML estimates. OLS estimates are biased. We cannot show the sign of the bias but we can show (from Maddala, p. 153) that:

$$\beta_{ML} = \beta_{OLS} - \sigma(x'x)^{-1}x'\lambda$$

where λ is now the column vector of the individual λ_j. The ML estimates equal the OLS estimates plus a constant times the regression coefficient of λ on x. Again we see the importance of the decision to participate. For someone unlikely to participate ($x_i\beta$ very small) the truncation of ϵ is large, and the λ term is also large. If everyone sampled is quite likely to come, then λ will be small and OLS estimates will have good properties. Olsen has devised a correction factor for OLS estimates, which has been applied by Smith, Desvousges and McGivney and by Wetzstein and Ziemer.

Computing Benefits

There are many issues relating simply to the computation of benefits which could be explored (see Bockstael and Strand, for example.) Here we ask a simple question about the expansion of benefits per individual to the population as a whole. Suppose we estimate the parameters of the individual model, given in (9) or (14), but we do not have a census of users. Instead, we have means of the independent variables by geographical or political subdivisions which we will denote zones.

Consumer's surplus for user j, zone i can be written

$$cs_{ij} = cz_{ij}^2$$

where c is the constant $-(2\beta_1)^{-1}$, $z_{ij} \equiv x_{ij}\beta$ where β is the vector of coefficients relating to trips per user, and x_{ij} is the vector of exogenous variables for individual j in zone i. The coefficient β_1 is the demand slope in price space. If we had a census of users in zone i, then benefits for the zone would be

$$(15) \qquad B_i = c \sum_{j=1}^{n_i} (z_{ij})^2$$

where n_i is the number of users. Suppose we have only mean data for users in zone i. Let $\bar{z}_i$

$= \sum_{j=1}^{n_i} z_{ij}$ be the mean. Then benefits for zone i would be

$$B_i^o = cn_i\bar{z}_i^2 \tag{16}$$

The difference between (15) and (16) is

$$B_i - B_i^o = c[\Sigma z_{ij}^2 - n_i\bar{z}_i^2]$$

which is proportional to the variance of the $z_{ij}(= x_{ij}\beta)$. Thus, when individuals have substantially different values of variables influencing their trips, computing consumer's surplus from means will undervalue a site. This result is not new (for example it is found in Dwyer, Kelly and Bowes), but it shows clearly the importance of the homogeneity of zones, which was apparent from the analysis of the previous section.

The conclusion, that we add up individual benefits, is only a useful rule so long as we have individual data. The difficulty is that in computing site benefits, having individual data requires a census of site users that is frequently unavailable. Often we have only fairly gross aggregated data on the arguments of individual demand functions. And, unfortunately it is not clear how to correct benefit measures when only aggregate data are available.

Conclusion

This paper has explored some issues in using individual and aggregate observations for estimating the demand for outdoor recreation. We have introduced an approach which gives some insight into the use of aggregate observations. Whether we are using individual or aggregate observations, it is imperative to keep track of the decision to participate. It provides the connection between individual behavior and aggregate data.

One aspect of the paper touched on but not explored concerns the expansion of estimates. There are considerable sample selection problems in expanding estimates. Economists have devoted most of their efforts to issues surrounding the specification and estimation of individual demand curves from a sample of observations. A cursory look at the aggregation issue suggests that the plausibility of benefit estimates could be substantially improved if greater care were given to the extrapolation of benefits calculated for the sample to aggregate benefit estimates for the population. In the process of thinking through precisely how total benefits are computed, we can learn a lot about the individual vs. aggregate observation issue.

References

Bockstael, N. E. and I. E. Strand. "The Nature of Error in Recreation Benefit Demand Analysis and its Effect on Estimates." Manuscript. 1984.

Bockstael, N. E., I. E. Strand and W. M. Hanemann. "Time and Income Constraints in Recreation Demand Analysis," Manuscript. 1984.

Bowes, M. David and J. B. Loomis. "A Note on the Use of Travel Cost Models with Unequal Zonal Populations." *Land Econ.* 56(1980):465–470.

Brown, W. G. and F. Nawas. "Impact of Aggregation on the Estimation of Outdoor Recreation Demand Functions." *AJAE* 55(1973):246–249.

Brown, W. G., C. Sorhus, B. Chou-Yang, and S. Richards. "Using Individual Observations to Estimate Recreation Demand Functions." *AJAE* 65: 154–157.

Christiansen, J. B. and C. Price. "A Note on the Use of Travel-Cost Models with Unequal Zonal Populations: Comment," *Land Economics* 58(1982):395–99.

Davidson, P., Adams, F. G., and Seneca, J. J. "The Social Value of Water Recreational Facilities Resulting from an Improvement in Water Quality: The Delaware Estuary" in *Water Research,* eds. Allen V. Kneese and Stephen Smith. Baltimore. Johns Hopkins University Press. 1966.

Dwyer, J. F., J. R. Kelly, and M. D. Bowes. *Improved Procedures for Valuation of the Contribution of Recreation to National Economic Development,* Urbana-Champaign: University of Illinois. 1977.

Heckman, J. J. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables," *Annals of Social and Economic Measurement* 5(1976):475–492.

Hotelling, Harold. Letter to the National Park Services in "Economics of Outdoor Recreation"—The Prewitt Report. 1948.

Maddala, G. S. *Limited-dependent and Qualitative Variables in Econometrics.* New York. Cambridge University Press. 1983.

Smith, V., Kerry, W. H. Desvousges, and Mathew P. McGivney. "The Opportunity Cost of Travel Time in Recreation Demand Models," *Land Econ.* 59 (1983):259–78.

Vaughan, W. J., C. S. Russell and M. Hazilla. "A Note on the Use of Travel-Cost Models with Unequal Zonal Populations: Comment," *Land Economics* 58(1982):400–07.

Wetzstein, Michael and Rod Ziemer. "An Application of a Truncated Regression Model to Recreation Demand" paper presented at AAEA meetings. 1982.