Travel Cost Models, Heteroskedasticity, and Sampling

Donald H. Rosenthal and Jana C. Anderson

Using theoretical derivations, it is shown that collecting data on individuals' visitation rates to a recreation site by each of these methods: (1) on-site sampling of visits; (2) sampling individuals surrounding the recreation site; and (3) sampling license holders, results in three unique heteroskedasticity problems. A different weighted least squares approach is offered in each case when estimating the visits per capita-travel cost relationship in zonal travel cost models. Furthermore, to the extent that individuals within an origin zone face different prices, there is an inherent aggregation bias when estimating consumer surplus.

The travel cost model (TCM) is probably the most widely used technique for deriving economic estimates of the value of recreation sites. This situation is likely to continue despite alternative, and perhaps better, methods for valuing recreation sites. These alternative methods, notably the contingent valuation technique, require more and/or different information to use. Such information is often difficult or costly to obtain. Because of its widespread use, it is important that econometric issues associated with the TCM model be clearly understood.

Recent articles have discussed the issue of heteroskedasticity in the zonal TCM (Bowes and Loomis; Christensen and Price; Vaughan *et al.*). Collectively, these articles indicate that heteroskedasticity (1) is directly related to the number of visits from the zone of origin and inversely related to the zone population, and (2) can easily be confused with misspecification of functional form. Because of these findings the use of weighted least squares to correct for heteroskedasticity is widely recommended for estimating demand curves in zonal travel cost models. Failure to use the weighted least squares approach when heteroskedasticity exists results in estimates of regression coefficients that are unbiased and consistent, but not efficient. Also, variance estimates of the regression coefficients will be biased, thereby affecting hypothesis testing (Kementa, p. 255). In contrast, the weighted least squares estimates have the desirable properties listed above, are efficient, and produce unbiased variance estimates.

The purpose of this paper is to show that there is yet another factor in the heteroskedasticity problem—the manner in which the data is collected. A discussion of heteroskedasticity will follow a review of the TCM. This review will illustrate the utility theoretic foundations of the TCM and show that there is probably an aggregation bias in estimates of total consumer surplus when estimates are based on the zonal TCM.

Potential Aggregation Bias in the Zonal TCM

By assuming that individuals maximize utility subject to both time and budget

Donald H. Rosenthal is an Economist and Jana C. Anderson is a Biometrician with the USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO.

The authors would like to thank Rudy King, John Hof, Dan Stynes, John Loomis and anonymous reviewers for helpful comments on this manuscript. Any remaining errors are, of course, the responsibility of the authors.

Western Journal of Agricultural Economics, 9(1): 58–65

constraints, the demand for visits to a recreation site becomes dependent on a variety of factors (Wilman). These factors include distance to the site, entry fee, required travel time, marginal (dis)utility of travel time, opportunity cost of time, wage rate, and others. However, for the purpose of discussing the relationship of the individual TCM to the zonal TCM, such detailed formulations are not needed.

To be specific, assume that individuals behave as if they are maximizing a utility function $U = U(v, x)$ subject to:

$$pv + cx \leqq M \tag{1}$$

where:

- v = number of visits to a recreation site,
- x = quantity of a composite good,
- p = price per recreation trip, expressed to include "cost of time," entry fees, and travelling costs,
- c = price of composite good, and
- M = income.

Solution of the above maximization problem yields the following ordinary demand functions:

$$v = v(p, c, M) \tag{2}$$

$$x = x(p, c, M) \tag{3}$$

By definition, the Marshallian consumer surplus associated with optimal consumption of visits is:

$$\int_{p}^{p^*} v(p, c, M)\, dp \tag{4}$$

where:

- p = existing price per trip, and
- p^* = price per trip for which $v = 0$.

The total value of the recreation site would be (4) summed across all recreationists using the site.

Unfortunately, data about individual recreationists are often not available, hence (4) cannot be estimated. Because of this, the zonal TCM model is frequently used to estimate an aggregate demand function. In the zonal TCM an aggregate function of the form

$$T_i = T(C_i, P_i, Z_i) \tag{5}$$

is estimated where:

- T_i = number of visits to the recreation site from origin i ($i = 1 \ldots m$),
- C_i = price of composite commodity for origin i,
- P_i = price per trip from origin i, and
- Z_i = characteristics of origin i, including population and income.

By making the dependent variable trips from a geographically defined area, information about the quantity of recreation consumed by individuals is not needed.

Based on equation (5) total consumer surplus for the site is:

$$\sum_{i=1}^{m} \int_{P_i}^{P_i^*} T(C_i, P_i, Z_i)\, dP_i \tag{6}$$

where: P_i^* = value of P_i which relates to zero predicted visits.

An obvious question is, under what conditions will consumer surplus calculated from equation (4) summed across all individuals equal the aggregate measure of consumer surplus (6)? The two measures will be identical if the aggregate demand function is properly specified and all individuals within an origin zone face identical values for any variable within the aggregate function.

The following example will clarify this point. Assume that all individuals have linear demand functions of the form

$$v_{ij} = a_{ij} - b_{ij}p_{ij} \tag{7}$$

where the subscripts refer to the jth ($j = 1 \ldots N_i$) individual from the ith ($i = 1 \ldots m$) origin zone. The value of N_i is the pop-

ulation of the ith origin. In this case, total consumer surplus summed across all individuals is

$$\sum_{i=1}^{m}\sum_{j=1}^{N_i}\int_{p_{ij}}^{p_{ij}^*}(a_{ij}-b_{ij}p_{ij})\,dp_{ij}. \tag{8}$$

This is equivalent to

$$\sum_{i=1}^{m}\sum_{j=1}^{N_i}(a_{ij}p_{ij}^* - \tfrac{1}{2}b_{ij}p_{ij}^{*2} - a_{ij}p_{ij} + \tfrac{1}{2}b_{ij}p_{ij}^2) \tag{9}$$

which is the true consumer surplus for the site.

The estimated linear per-capita demand function corresponding to the zonal TCM, equation (5), is

$$\text{vcap}_i = \hat{A} - \hat{B}P_i + \epsilon_i \tag{10}$$

where:

$$\text{vcap}_i = \sum_{j=1}^{N_i} v_{ij}/N_i,$$

and

$$P_i = \sum_{j=1}^{N_i} p_{ij}/N_i.$$

In the zonal TCM, total site consumer surplus is calculated as

$$\sum_{i=1}^{m} N_i(\hat{A}\hat{P}^* - \tfrac{1}{2}\hat{B}\hat{P}^{*2} - \hat{A}P_i + \tfrac{1}{2}\hat{B}P_i^2). \tag{11}$$

Equation (11) is the sum across zones of the per capita consumer surplus multiplied by the zone population.

Sufficient conditions for equation (11) to be equivalent to the true consumer surplus shown in equation (9), are:

$$a_{ij} \equiv \hat{A},$$
$$b_{ij} \equiv \hat{B},$$
$$p_{ij}^* \equiv P^*, \text{ and}$$
$$p_{ij} \equiv P_i.$$

The first three conditions are met if and only if all individuals have identical demand functions. Clearly, the assumption of identical demand functions is quite restrictive. It would be nice if the zonal TCM would give unbiased estimates of consumer surplus under less stringent assumptions.

If the first three conditions are violated then, in general, (10) will not be linear. However, the consumer surplus estimates from the zonal TCM will still be unbiased if the distribution of tastes, preferences, and incomes is constant across origin zones. If the effect of substitute recreation opportunities is not accounted for in the per-capita demand curve, these recreation opportunities must be assumed equally available to all origins (Cicchetti *et al.*). If any one of these factors varies from origin to origin, then variables must be included in the per-capita demand curve to reflect the differences.

The important point is that an unbiased estimate of the total site consumer surplus will still be obtained if the per-capita demand curve is properly specified. This is true because a properly specified per-capita demand function traces out the average individual demand function for an origin zone. At a fixed price, the per-capita demand function predicts the expected value of the number of trips to the site that are taken by any individual within an origin. Individuals who do not visit the site, i.e., quantity equals zero, are included in this expected value computation. The area under this average function but above price, i.e., average consumer surplus, times the zone population equals (4) summed across all individuals within a zone.

The fourth condition is that the price of a trip to a site is the same for all individuals within a zone. To the extent that this is not true there is an aggregation bias in the zonal TCM. This effect, which holds even if all individuals have identical demand functions, is apparent in the last term of equation (11). If all individuals have identical demand functions, then, using the Jensen inequality and noting that $E(p_{ij}) = P_i$, it follows that $\tfrac{1}{2}\hat{B}E(p_{ij}^2) \geqq$

$\frac{1}{2}\hat{B}P_i^2$ (the expectation is computed over j) with the equality holding if $p_{ij} \equiv P_i$. In the linear case shown in equation (11), the effect of aggregation bias is to underestimate consumer surplus. More generally, if any variable included in the aggregate per capita demand equation varies among individuals in the same origin zone there will be aggregation bias. To minimize this bias, zones should be kept as small and homogeneous as possible.

In the zonal TCM the dependent variable used in regression analysis is visits per capita from a specific origin. The value of this variable is usually estimated from a sample. Because of random sampling variation, the precision of the estimate varies for different origin zones. This measurement error in the dependent variable creates heteroskedasticity when regression analysis is applied in the zonal TCM. The remainder of the paper will now discuss this heteroskedasticity and methods for correcting the problem.

Heteroskedasticity When Visits On-Site Are Sampled

The nature of heteroskedasticity in the zonal TCM depends on the manner in which the data is collected. The three major ways in which data can be collected for travel cost models are through (1) random samples of visits at the recreation site, i.e., on-site data collection, (2) random samples of individuals surrounding the recreation site, i.e., household survey, and (3) random samples of individuals holding licenses to participate in certain activities on the site, such as hunting or fishing. Data collected in each of these ways results in a different heteroskedasticity problem. Therefore, corrections for heteroskedasticity problems must be specific concerning which type of data and model they apply to.

When data are collected by randomly sampling visits on-site, the process can be viewed as a stationary Bernoulli process. It is stationary because the probability of each visit being sampled is constant throughout the sampling period. For notational purposes, let:

T_i = number of visits from origin i (i = 1 . . . m) to the recreation site;
N_i = population of zone i;
$vcap_i = \frac{T_i}{N_i}$ visits per capita from zone i;
t_i = number of visits (trips) sampled at the recreation site from zone i;
$\hat{T}_i$ = estimated number of visits from zone i;
p = the sampling rate or the probability of sampling any visit to the site;
q = 1 − p; and
P_i = average travel cost from zone i to the recreation site.

The number of trials in this Bernoulli sampling process is the number of visits from origin i, T_i, and the number of successes is the number of visits actually sampled, t_i. In other words, origins send visits to the site which each have known probability p of encountering a sampler. The probability of t_i visits being sampled from T_i total visits is,

$$p(t_i) = \binom{T_i}{t_i} p^{t_i} q^{T_i - t_i}$$

with mean $T_i p$ and variance $T_i pq$. Using the method of moments the estimated number of visits from zone i, $\hat{T}_i$, is

$$\hat{T}_i = t_i/p \text{ with variance}$$

$$\text{Var}(\hat{T}_i) = \frac{\text{Var}(t_i)}{p^2} = \frac{T_i pq}{p^2} = \frac{T_i q}{p}.$$

Using the above information, the esti-

mated visits per capita from zone i, $\widehat{vcap}_i$, is

$$\widehat{vcap}_i = \hat{T}_i/N_i.$$

For large T_i, $\widehat{vcap}_i$ is approximately distributed as

$$\widehat{vcap}_i \sim N[vcap_i, T_iq/pN_i^2].$$

Inspection of the variance term for $\widehat{vcap}_i$, reveals a heteroskedasticity problem. The larger N_i or the smaller T_i, the less the variance.

The solution to heteroskedasticity problem in general, is to minimize

$$\sum_{i=1}^{m} \frac{1}{Var(Y_i)}[Y_i - f(X_i)]^2 \qquad (12)$$

where: $f(X)$ = a function of X.

In the case of sampling visits at the recreation site the weight that should be used, i.e., $1/Var(\widehat{vcap}_i)$, is shown in Table 1. Therefore, the visits per capita-travel cost relationship should be found by minimizing

$$\sum_{i=1}^{m} \frac{N_i^2}{\hat{T}_i}[\widehat{vcap}_i - f(P_i)]^2 \qquad (13)$$

The weighted least squares results in error terms having approximately constant variance, q/p.

The form of the function $f(P_i)$ in equation (13) can be quite general. Depending on what the researcher feels is appropriate, a linear, quadradic, cubic, or logarithmic form can be used. Recent research indicates that a semi-log model (log dependent variable) might be the appropriate functional form (Ziemer *et al.*; Sutherland; Vaughan *et al.*; Strong). However, if the dependent variable is transformed in this manner the weights given in Table 1 no longer hold.

For it to be appropriate to take the log of the dependent variable and fit the equation using ordinary least squares the original function should be of the form

TABLE 1. Regression Weights by Sampling Method.

Visits	Individuals	License
$\frac{N_i^2}{\hat{T}_i}$	$\frac{n_i^*}{\sigma_i^2}$	$\frac{\ell_i N_i^2}{\sigma_i^{2*} L_i^2}$

$$vcap_i = \exp(A - BP_i + \epsilon_i) \qquad (14)$$

where ϵ_i is a normally distributed random disturbance term with zero mean and constant variance. Theoretically, the error should enter in the exponential term so it becomes additive in the log-linear regression.

However, from the derivations presented above it is apparent there is a heteroskedastic and normally distributed error component due to sampling that enters in an additive manner of the form

$$vcap_i = \exp(A - BP_i + \epsilon_i) + \Delta_i \qquad (15)$$

If the ϵ_i term in (15) is assumed negligible, it can be estimated using a non-linear regression program such as BMDP P3R (Dixon). If such a program is used then the weights in Table 1 still hold because the dependent variable is no longer transformed.

The point of this discussion is that further research into the pros and cons of log-linear versus non-linear regression would be worthwhile. As the sampling rate, p, decreases the additive error term in (15) becomes increasingly significant. Past research has indicated that log-linear ordinary least squares gives surprisingly robust coefficient estimates even when the error term is misspecified (Barr and Horrell).

Heteroskedasticity When Individuals Surrounding the Recreation Site Are Sampled

For notational purposes, let:

v_{ij} = number of visits to the site

from the ith origin by the jth ($j = 1 \ldots N_i$) person,

$vcap_i = \sum_{j=1}^{N_i} v_{ij}/N_i$ = visits per capita from the ith zone,

$\sigma_i^2 = \sum_{j=1}^{N_i} (v_{ij} - vcap_i)^2/N_i$ = variance of visitation rates at origin i,

n_i^* = number of individuals sampled in ith origin, and

$\widehat{vcap}_i = \sum_{j=1}^{n_i^*} v_{ij}/n_i^*$ = estimated visits per capita from the ith zone.

Early reviews of this paper indicated there is confusion about what σ_i^2 is. While each value of v_{ij} if fixed, the population of v_{ij} values from any origin varies about $vcap_i$ with variance σ_i^2. This variance is of concern when trying to estimate the demand curve for a particular site.

Under this sampling arrangement, the sampling unit is individuals. Therefore, in each zone there are as many sample units as there are individuals. Data is collected by randomly sampling individuals in an origin zone and asking them how many trips they made (will make) to the site during a specified time period. The estimated visits per capita is simply the average of the indicated number of trips. The heteroskedasticity correction, therefore, follows along the usual lines for grouped data (Maddala, p. 268).

At this point it should be noted that knowledge of the variables listed above is almost sufficient to use the individual TCM instead of the zonal TCM. If the price per trip for each individual could be determined, then the individual TCM could be used. That might be the preferable approach. The option of assuming $P_i \equiv p_{ij}$, as is done in the zonal TCM, and then using the individual TCM creates its own set of problems related to measurement error which are beyond the scope of this paper (Brown *et al.*).

For large samples, or small samples when v_{ij} is normally distributed, $\widehat{vcap}_i$, has the following distribution:

$$\widehat{vcap}_i \sim N(vcap_i, \sigma_i^2/n_i^*).$$

Inspection of the variance term reveals a heteroskedasticity problem caused by grouping. The larger n_i^* or the smaller σ^{2i}, the smaller the variance. The weight to be used in this case is shown in Table 1.

The critical issue in determining the actual weights to use for each case is whether or not σ_i^2 is constant across origin zones, i.e., $\sigma_i^2 = \sigma^2$. Christensen and Price discuss this issue and interested readers are referred to their paper. However, it should be noted that no assumptions about σ_i^2 were needed to derive the weights for the previously described case of sampling visits on-site. If the variance is constant then n_i^* can be used as the weight. Assuming an equal proportion of all individuals in each zone have been sampled, weighting by n_i^* is equivalent to weighting by N_i (they differ only by a constant proportion) as originally suggested by Bowes and Loomis (1980).[1]

Heteroskedasticity When License Holders Are Sampled

In some recreation activities, notably fishing and hunting, there are good records of individuals who hold licenses to participate in certain activities. If these license holders are sampled, then a slightly different heteroskedasticity problem emerges than when individuals are sampled. For notational purposes, let:

[1] Bowes and Loomis state the weighting factor is $\sqrt{N_i}$. Minimizing equation (12) using N_i as the weight results in exactly the same solution as multiplying all variables in the regression by $\sqrt{N_i}$ and using ordinary least squares to estimate a regression which is forced through the origin. Therefore, the Bowes and Loomis solution is exactly the same as suggested here, but the terminology differs.

L_i = number of license holders in origin i,
ℓ_i = number of license holders sampled, and
σ_i^{*2} = variance of visitation rates for license holders in origin i.

Under this sampling scheme, the objective is still to derive an estimate of $vcap_i$; only now people not holding licenses are assumed to have a zero value for v_{ij}. As before, $vcap_i$ is estimated by dividing the total number of trips from origin i by the population of origin i. This procedure takes into account both the changing number of trips per year by license holders and the changing percentage of the population that holds licenses.

The estimated number of trips per sampling period by license holders has a mean

$$\widehat{vcap}_i(\text{license}) = \frac{\sum_{j=1}^{\ell_i} v_{ij}}{\ell_i} \text{ and variance } \frac{\sigma_i^{*2}}{\ell_i}.$$

It should be noted that $\widehat{vcap}_i$(license) is not a good dependent variable to use in the TCM because it does not reflect how the percentage of population visiting the site falls off with distance. It only reflects how the number of trips per capita for those who own licenses varies with distance. To estimate $vcap_i$ for all people in the zone, not just license holders, $\widehat{vcap}_i$(license) should be multiplied by L_i and divided by N_i. If this is done, $\widehat{vcap}_i$ approaches the following distribution

$$\widehat{vcap}_i \sim N(vcap_i, \sigma_i^{*2}L_i^2/\ell_i N_i^2).$$

Inspection of the variance term for $vcap_i$ reveals a heteroskedasticity problem. Specifically, the larger σ_i^{*2} or L_i and the smaller ℓ_i or N_i the larger the variance. The weights to be used in this case are shown in Table 1. If it is assumed that $\sigma_i^{*2} = \sigma^{*2}$ and a constant proportion of license holders in each zone are sampled, then the Table 1 weight simplifies to N_i^2/L_i.

Conclusion

Under three different sampling schemes, three different heteroskedasticity problems were found. Each of these problems requires a different weighted least squares approach if the parameters of the visits per capita-travel cost relationship are to be efficiently estimated. Thus, corrections for heteroskedasticity problems suggested in the literature need to be specific concerning what type of data they are applicable to.

Additionally, when data are collected by sampling visits on site, no assumptions about σ_i^2 need to be made to derive the heteroskedasticity problem. This is a fortunate outcome, because one of the strengths of the zonal travel-cost model is that it is well suited to using on-site data.

Because of sampling errors in estimates of $vcap_i$, log-linear regression might not be appropriate when estimating per capita demand curves. Further research in this area is needed because choice of functional form can markedly affect estimates of consumer surplus (Sutherland; Strong; Ziemer).

The zonal travel cost model is consistent with individual recreationists acting as utility maximizers. However, there is apt to be some aggregation bias associated with estimates of consumer surplus based on the zonal model. This bias stems from the fact that it is unlikely all recreationists within an origin zone face the same price of visiting the recreation site. As the variation of individual prices within a zone increases, the potential aggregation bias becomes more of a concern.

References

Barr, T. N. and J. F. Horrell. "Least Squares Robustness when the Error Term is Misspecified in Cobb-Douglas Type Functions." *Agricultural Economics Research*, 28(1976): 4 136–45.

Bowes, M. D. and J. B. Loomis. "A Note on the Use

of Travel Cost Models with Unequal Zonal Population." *Land Economics*, 56(1980): 4 465–70.

Bowes, M. D. and J. B. Loomis. "A Note on the Use of Travel Cost Models with Unequal Zoneal Population: Reply." *Land Economics*, 58(1982): 3 408–10.

Brown, W. G., C. Sorhus, B. Chou-Yang, and J. A. Richards. "Using Individual Observations to Estimate Outdoor Recreation Demand Functions: A Caution." *American Journal of Agricultural Economics*, 65(1983): 1 154–57.

Christensen, J. B. and C. Price. "A Note on the Use of Travel Cost Models with Unequal Zonal Populations: Comment." *Land Economics*, 58(1982): 3 395–99.

Cicchetti, C. J., V. K. Smith, J. L. Knetsch, and R. A. Patton. "Recreation Benefit Estimation and Forecasting: Implications of the Identification Problem." *Water Resources Research*, 8(1972): 4 840–50.

Dixon, W. J. *BMDP Statistical Software 1982*. University of California Press, Berkeley, CA, 1981.

Kementa, J. *Elements of Econometrics*. Macmillan Publishing Co., Inc., New York, 1971.

Maddala, G. S. *Econometrics*. McGraw-Hill, New York, 1977.

Strong, E. J. "An Note on the Functional Form of Travel Cost Models with Zones of Unequal Populations." *Land Economics*, 59(1983): 3 247–54.

Sutherland, R. J. "The Sensitivity of Travel Cost Estimates of Recreation Demand to the Functional Form and Definition of Origin Zones." *Western Journal of Agricultural Economics*, 7(1982): 1 87–98.

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): 3 400–07.

Wilman, E. A. "The Value of Time in Recreation Benefit Studies." *Journal of Environmental Economics and Management*, 7(Sept. 1980): 272–86.

Ziemer, R. F., W. N. Musser, and R. C. Hill. "Recreation Demand Equations: Functional Form and Consumer Surplus." *American Journal of Agricultural Economics*, 62(1980): 1 136–41.