CHAPTER VII- TN 19: OBTAINING EFFICIENT ESTIMATES OF PARK USE AND TESTING FOR THE STRUCTURAL ADEQUACY OF MODELS

By J. Beaman, J.L. Knetsch, H.K. Cheung

(see a version in 1977 Canadian Journal of Statistics, Section C: Applications, 5(1):57-92)

ABSTRACT

This paper provides an examination of the problem of heteroscedasticity as it relates to estimating park use, although the results can also be applied to a wide variety of flow problems involving traffic, people or commodities. The major issue is that estimates of flows obtained using ordinary least squares, OLS, often yield statistically significant results while still giving rise to large differences between observed and predicted flows.

The paper presents results which show that for the flow estimation problem of concern, more accurate use estimates may be obtained by using generalized least squares, GLS, rather than using OLS. Weights to use in a GLS regression are derived. These are presented in a covariance matrix which is developed taking in to account the variance to be expected in origin-destination flows.

It is shown that deriving the correct weights, estimates of variances, to use in a regression analysis results in an "absolute" test for the structural appropriateness of the regression model. Tests related to the "absolute" adequacy test are introduced and their use to identify specific structural problems with a model is illustrated.

INTRODUCTION

Increasingly, more formal methods of estimating attendance at proposed parks and recreation areas are being used in the planning and justification of such areas. Ordinary least squares regression (OLS) models are characteristically developed to estimate the relationship between (B) measures of the use of parks and recreation areas, usually specified as the volume of origin-destination flows for a number of existing sites, and (2) various independent variables influencing a park's use. These latter variables are usually chosen to reflect the characteristics of the sites under consideration and to be measures of the size and proximity of populations from which visitors come (Boyet & Tolley 1966; Ellis & VanDoren 1966; Cheung 1972; Pankey & Johnston 1969).

While the "relations" established by these kinds of regression are statistically significant, there is typically an undesirable lack of precision in the prediction of actual origin-destination visitor flows (attendance figures). Estimates often differ from observed values on which they are based by several hundred percent" thus there is a problem with the accuracy of estimates (Ellis & VanDoren 1966; Elsner 1971). The difficulty can be illustrated by examining the origins of visitors to a provincial park in Saskatchewan (Rowan's Ravine) and employing a simple model to explain the variation in flows as a function of the distance from the visitors, origins to the park (see Table 1) and the sizes of the populations at the origins. The total observed 1969 attendance at this park was 9,828 parties, coming from fifteen different origin areas. However, of the total, 5,868 parties (or about sixty percent) came from a single origin area located fairly close to the park and containing Regina, the largest population-centre in the region.

Ordinary least squares regression was used to estimate the parameters in Equation 1. The relationship between various flows from the different origins to Rowan's Ravine was obtained:

(1) $\log ((V(o,d)+1)/ P(o)) = 2.811 - 0.0241\ D(o,d)$

WHERE V(o,d) = number of visiting parties plus 1 coming from an origin, o, to a park, p,

P(o) is the population in thousands of o; and

D(o,d) is the distance from o to the destination, d, in road miles.

The constant 1.0 was added to visit numbers to avoid the problem created if computation involved taking the logarithm of zero. Logarithms used were 'base 10' (subsequently natural, base e, logarithms are specified as “ln”).

Equation 1 provided a reasonable explanation of the variation in the dependent variable. The R^2 for the regression is 0.77. The regression coefficients are highly significant according to the usual F-test. Also, the standard error of the regression coefficient of D(o,d) is only 0.00783. Nevertheless, the explanation of the use of the park is not particularly good, as can be seen from Table 1. Even though the fitting was done using the data shown, the estimate of a total of 5,887 visiting parties from all origins is nearly 4,000 below the actual total visitor flow.

In using OLS regression to estimate relationship between visits and distance, each observation point was treated as if it were as important as any other. In the example, the use of OLS regression treats the observation for origin unit 4 (which contributed 60 percent of the observed total use) as the equal of the observation for origin unit 13 (which contributed approximately O.3 percent of the total use) or the same as observation 16 which involved no visits and therefore contributed nothing to total use.

While errors in prediction are partly due to omission of causal factors and 'measurement errors' (Pankey & Johnston 1969), a further and major cause of poor predictions using the model is clearly the heteroscedasticity among the observations, and this is not properly dealt with when estimating the parameters. The preceding statement is made without applying a test for heteroscedasticity (Goldfield & Quandt 1965; Goodchild TN 35) because, in the following, the nature of the variance in observations is derived.

TABLE 1: STATISTICS PERTAINING TO OBSERVED AND ESTIMATED DAY VISITS TO ROWAN'S RAVINE PROVINCIAL PARK, SASKATCHEWAN

Observation Unit	Distance to Park in Miles	Population	Observed Visits Vehicles	Estimated Visits Un-Weighted Regression	Estimated Visits Weighted Regression	Percent Error 100*(Obsd.-Est.)/Est. Weighted Regression	Weighted Regression
1	133	32,489	36	11	22	227	64
2	126	51,923	0	29	55	-100	-100
3	104	17,813	63	34	66	85	-5
4	61	132,432	5,868	2,890	5,731	103	2
5	34	11,594	720	1,133	2,315	-36	-69
6	14	B,632	1,980	483	1,011	310	96
7	21	3,871	378	778	1,614	-51	-77
8	67	2,829	36	43	86	-16	-58
9	110	36,889	99	51	98	94	1
10	107	3,271	0	4	9	-100	-100
11	109	6,181	18	8	16	125	13
12	40	4,237	414	296	601	40	-31
13	139	21,104	27	5	10	440	170
14	84	16,284	63	98	190	-36	-67
15	154	117,405	126	21	38	500	232
16	117	4,456	0	3	7	-100	-100
17	129	2,729	0	0	1	-100	-100
Total			9,828	5,887	11,870		

A FIRST STEP TOWARD 'PROPER' ESTIMATION: DERIVATION OF A COVARIANCE MATRIX FOR THE OBSERVED VISITOR FLOWS

When the parameters of a model are estimated using generalized least squares (GLS), it is necessary to know the covariance matrix of the observations (matrix with variances of observations on the diagonal and correlations between them off diagonal). Subsequently, a covariance matrix for observed visitor flow, $\sum_{v()}$ is derived which is critical in obtaining the covariance matrix of log/ln transformed observation, $\sum_{L}$. Because regression is for Equation 1 the covariance matrix for ln((V(o,d)+1)/ P(o)) is eventually used in a regression analysis. To facilitate the discussion involved in deriving $\sum_{L}$ the following notations and definitions are used:

v(o,d,t,g)=observed number of vehicles, with parties in collectivity g, going from origin o to destination d on day t;

WHERE the collectivity g is a set of parties which tend to have similar behaviour in terms of their probability of participating in a given package of recreation activities at destination d on a day t with its given weather, park crowding and traffic, conditions, etc.

v_p(o,d,t,g)=the predicted value of v(o,d,t,g), 'an estimate' of it,

p(o,d,t,g)=the probability that a party in collectivity g would go to d from o on day t,

V(o,d,t,g)=the random variable that generates the observed values v(o,d,t,g),

E(V(o,d,t,g))=expected value of V(o,d,t,g)

VAR(V(o,t,d,g))=variance of V(o,d,t,g).

Additionally, when some subscripts are removed from v(o,d,t,g), the resulting expression implies a sum over the given subscript(s). The number of parties going from origin o to a destination d on day t is $v(o,d,t) = \sum_g v(o,d,t,g)$; the number over all days for all g is $v(o,d) = \sum_t \sum_g v(o,d,t,g)$; and the total attendance at the park is $v(d) \sum_o \sum_t \sum_g v(o,d,t,g) = \sum_o v(o,d)$.

The most usual assumption in regression analysis (the OLS or homoscedasticity assumption) is that for flows v( ), $\sum_{v()}$ is a diagonal matrix of the form:

$$\sum_{v()} = \sigma^2 \begin{vmatrix} 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \vdots \\ \vdots & & 1 & & \vdots \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{vmatrix}$$

WHERE σ^2 is the variance that applies to all observations, v( ).

But, now consider that visitor flows from o to d on day t for group g depend on N(o,g), the number of people in g available to participate, and p(o,d,t,g), their probability of participating. The very nature of the definition of V(o,d,t,g) in terms of N(o,g) and p(o,d,t,g) implies that V(o,d,t,g) is a binomial random variable with mean and variance as follows:

(4) E(V(o,d,t,g)) = N(o,g)p(o,d,t,g)

(5) VAR(v(o,d,t,g)) = N(o,g)p(o,d,t,g)(1-p(o,t,g,d))

(6) and, if p(o,d,t,g) is small, Equation 4 holds so that Equations 5 and 6 follow.

(7) $VAR(V(o,d,t,g)) \approx N(o,g)p(o,d,t,g) = E(V(o,d,t,g))$

(8) $E(V(o,d,t)) \approx \sum_g VAR(V(o,d,t,g))$

(9) $VAR(V(o,d,t) \approx \sum_g VAR(V(o,d,t,g)) = \sum E(V(o,d,t,g)) = E(V(o,d,t))$

a. One should note that v(o,d,t) can be observed and it is an estimate of E(V(o,d,t)); an observation is an estimate of its expected value (if the expected value exists). So, from Equation 5 one can see that it is possible to obtain estimates of the variance of V(o,d,t).

b. Similarly, for total use from an origin to destination, one obtains:

(10) $\text{VAR}(V(o,d)) = \sum_g \text{VAR}(V(o,d,t) \approx \sum_g E(V(o,d,t)) = E(V(o,d))$

Thus the variances in V(o,d,t) and V(o,d) are approximately proportional to their respective expected values, meaning that either observations or predicted values can be used as estimates of the variance V(o,d) (see Equation 4). The merits of using observations to define weights to obtain estimates that can then be employed as variance estimates in a second cycle of estimation is not discussed.

Now it will be the exception, rather than the rule, that a change in vehicle flow to one site will be correlated with change in visitor flows to other sites. This is because a trip by a single visiting party of a given type on a given day from a given origin to one site or another is not expected to influence the decision of other parties in other vehicles. Obviously, parties may make decisions based on what they think other parties will do, but this is not the issue.

Assuming that the fluctuations in daily flows to one origin-destination pair for one type of user are not correlated with similar flows to another origin-destination pair, it follows that the estimated covariance matrices of v(o,d) for n flows can be written as follows (with the matrix for v(o,d,t) or other observations being written in a similar way):

(8) $\sum_{v()} = \sigma^2 \begin{vmatrix} v(o_1,d_1) & 0 & \cdots & 0 & 0 \\ \vdots & & & & \vdots \\ & & v(o_k,d_k) & & \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & \cdots & 0 & v(o_n,d_n) \end{vmatrix}$

This matrix is the appropriate covariance matrix of observations if non-linear regression were being used with the v(o,d) as the dependent variable (see TN 35).

OBTAINING MODEL PARAMETERS USING LINEAR REGRESSION

For the estimates of the parameters to be efficient, certain distributional properties of the error term e(o,d) must be assumed. The form of Equation 1 suggests that the ' fluctuations, in the random variable V(o,d) define the variance of the error term e(o,d). It is complicated to give an exact relationship that shows how the fluctuations in V(o,d) define fluctuations in e(o,d). However, a Taylor series expansion of ln((V(o,d)+ΔV)/P(o)) around E(V(o,d,)) results in Equation 1 taking the form shown in Equation 9 which, yields Equation 10.

(9) (9) $\log f|_{E(V,o,d)} + (\partial f/\partial V\ \Delta V \approx a + bD(o,d) + e(o,d)$

WHERE f = ln ((V(o,d)+1)/P(o)) and the derivative is evaluated at E(V(o,d))

(10) $\ln f + \Delta V /(E(V(o,d)) + 1) \approx a + bD(o,d)+e(o,d)$

a. Because the random fluctuations on the two sides of Equation 10 must be approximately equal:

(11) $\Delta V/(E(V(o,d))+1) \approx e(o,d)$

a. Because the series expansion is about E(V(o,d)), by definition ΔV is the fluctuation of V(o,d) around its expected value, so its variance is the same as the variance of V(o,d). What is more, the variance of V(o,d) is VAR(V(o,d)) which, as was shown earlier, is approximated by E(V(o,d)). So, by well known statistical theorems:

(12) $\text{VAR}(e(o,d)) \approx E(V(o,d))/(E(V(o,d)) + 1)^2$

Given that the v(o,d)'s are uncorrelated, an appropriate covariance matrix for GLS estimation of the parameters in Equation 1, $\sum_L$, is one that is defined using variance estimates given by Equation 12 and with zeros off the diagonal.

It should be noted that the preceding discussion has implied that all use of a park during a given period is monitored. Usually, however, the total traffic flow (or the components of that

flow) to a site from an origin is not observed but estimated. The dependent variable in a regression thus will likely be a function of vv(o,d), a weighted sum of observations. Yet the fact that in a particular regression vv(o,d) is the dependent variable presents no particular problem. One may simply use the variance in the vv(o,d)'s in a GLS regression by entering them in place of E(V(o,d)) in Equation 12. The weights w(o,d,t) used to multiply the flows v(o,d,t) to get vv(o,d) can be used to obtain the variance in vv(o,d). Using both the results presented in Equation 8 and the well known statistical theorems that deal with variance, it follows that if V is used as a notation to indicate survey observations with time in hours, days or some appropriate unit and if $vv(o,d) = \Sigma\, w\,(o,d,t)\, v(o,d,t)$ where the sum is over t, then:

(13) an estimate of variance in $vv(o,d) = \Sigma w^2\,(o,d,t) vv(o,d,t)$ summed over all sample times

Alternatively, if a survey design allowing variances in origin-destination visitor flow estimates to be calculated was used, the estimates obtained could be employed in defining the covariance matrix.

AN APPLICATION

The preceding discussion implies that it is appropriate (1) to accept the heteroscedasticity of variances in flows when using the ,'logarithmic additive' model, and (2) in determining regression coefficients, to give larger weights to origins contributing large flows of visitors than to those that contribute small visitor flows to the total use of an area.

For reasons cited earlier, the covariance matrix of the observations used in making GLS estimates of the parameters of Equation 1 is essentially given in Equation 12. When the parameters in Equation 14 were estimated using GLS for the same set of data as used to derive Equation B, the following equation was obtained:

(14) $\log (V(o,d)+1/P(o)) = 3.13701 - 0.260\, D(o,d)$

The predicted numbers of visits from each origin to Rowan's Ravine and the percent of error between predictions and observations are presented in Table 1 as they were for the OLS regression. The R^2 attained was 0.87, which was up from 0.77. As in the OLS regression, the regression coefficients are highly significant. And, in this case, the standard error of the regression coefficient of D(o,d) was 0.00246. Comparing Equations 15 and B, one sees that the coefficient of D(o,d) is relatively constant (at 0.02413 for OLS and 0.00246 for GLS). These coefficients also have small standard errors (O.00340 and 0.00246) in comparison to their actual values.

An obvious difference between the results of the two regressions is seen in Table B. The residuals obtained using GLS regression range from 1 to 1,595, as compared to the residuals of the unweighted regression analysis which range from 1 to 2,978. But, because of the bias involved when antilogarithms of predicted values are taken to obtain estimates of the individual flows, it is not clear to what extent the larger residuals for the OLS regression are due to model specification error, to measurement error and to "pure" logarithmic transformation bias.

One should recognize that GLS regression analysis resulted in an increase of percent error for some flows, such as those from observation units 5, 7, and 12. However, all of these flows are small compared to the flow from observation unit 4 to Rowan's Ravine. As one can see from Table 1 the average the percent error in the individual flows was greatly reduced by using weighted (GLS) regression. Had a weighted average been used to compute average error, GLS results would have appeared even better. Observation 4, which contributed about sixty percent of the total visitor flow, had its error reduced from 103 percent to two percent. Regardless, whether a percent RMS error measure or variances of parameters is considered, the GLS model is superior to the OLS model. (See Table 2.)

AN 'ABSOLUTE' MEASURE OF MODEL APPROPRIATENESS

From a critical perspective, the model described in this paper has been assumed to be structurally sound. The wary reader may be disturbed by this assumption. Actually, the theoretical error distributions developed can be used to see if the observed residuals are distributed as they should be if the model is structurally appropriate for the data. The results already presented suggest that one consider:

$$\chi^2_{M-N} \approx \sum (\text{residual})^2 / E(V(o,d)+1) \text{ over all o,d flows}$$

WHERE M number of flows observed and N number of parameters estimated.
For evaluating individual flows:

$$\chi^2_{M-N} \approx (\text{residual})^2 / E(V(o,d)+1)$$

The rationale for Equations 15 and 16 is that for the distribution being considered, a residual squared divided by its variance is approximately the square of a normal zero-one variable. This is by definition a chi-square with one degree of freedom. It is recognized that the residuals are not orthogonal to each other since degrees of freedom are lost when parameters are estimated. So, in Equation 15, degrees of freedom M-N are suggested with N being the number of regression parameters. The authors believe that using observations as GLS weights does not result in the loss of further degrees of freedom.

Using Equation 16, it is possible to see that, over all, there are structural problems with the model. The large χ^2 values in Table 3 actually make it clear that the model does not do as well as it should in explaining the observed flows. The χ^2_{M-N} having a highly sign value, is "absolute" proof that the structure of the model used is not totally adequate to explain the observed flows so its value is an "absolute" criterion for the structural adequacy of a model. Obviously, several models could be accepted as structurally adequate and, if this is the case, then it is possible that the methodology of Smith (1975) should be employed to select one model as the best.

TABLE 2: PERCENT ROOT MEAN SQUARE (RMS) ERROR
AND RELATED STATISTICS OF OLS AND GLS ESTIMATES

			Improvement Factor
Error Measure	OLS	GLS	GLS/OLS
% RMSE Error*	200.00	95.91	0.48
S.D.of b**	O.00783	0.00567	0.72
Average	-	-	0.60

* % RMS Error - (B/17 Σ (% error of observation i)2)1/2 WHERE percent error of observation i =observed visits - estimated visits estimated visits of observation

** It is recognized that estimates of the Standard Division in the regression coefficient, S.D. of b, is biased when OLS is used with heteroscedastic data. One can of course use the data provided to calculate an unbiased OLS estimate of the S.D. of b but it is not relevant to the problem under consideration. This is because the concern is with comparing a procedure accepting the OLS model with a GLS result.

As well as recognizing that an overall structural adequacy test can be made, it should be noted that the very large χ^2 values associated with observation units 5 and 7 certainly reflect problems with the model used because those observed for the origin-destination flows have essentially zero probability of occurring. However, careful examination of how origin areas 5 and 7 were defined and how the distances from these areas to Rowan's Ravine were measured suggests that the poor agreement between predictions and observations is a result of D(o,d) being given a value that is smaller than the value it should have. Similar considerations allow one to understand other significant residuals. A table such as Table 3 has good information from which to improve a model and such a table should be computed as a routine part of analysis to

determine if a model is structurally adequate.

CONCLUSION

This paper has dealt with obtaining efficient estimates of the parameters in an equation that defines a relationship between visitor flows and other variables. The rather obvious conclusion that has been reached is that, in regression analysis of flows using an equation such as Equation B, greater weights should be given to the more accurate observations of visitor flows so that efficient flow and parameter estimates can be obtained.

A practical consequence of having more efficient estimates is that research costs can be reduced or planning accuracy improved without increasing existing data collection costs. The fact that, on the average, accuracy improvement was sixty percent (see Table 2) means that to achieve the GLS level of accuracy using OLS, about three times as much data, (1/0.60)2 =1.3, would be required. Since using GLS regression costs no more (or little more) than using OLS, using it makes sense.

Finally, regarding point 3, it is admitted that in deriving the covariance matrices for v(o,d) and v(o,d,t), a number of assumptions were important in reaching the expressions derived. The validity of these assumptions about the behaviour of recreators must be checked. The specific concern must be whether probabilities that are assumed to be small are small. However, one should not make too much of this. As noted for point 2, small or even moderate errors (30 or 40 percent) in the variance elements of covariance matrix - errors due to poor approximations - have less effect on the parameter values estimated using the covariance matrix than one might expect.

In conclusion, an example helps illustrate the importance of having both efficient estimates and an absolute measure of a model,' structural adequacy. The work of Ellis and VanDoren cited early in this paper compares the 'goodness' of a gravity model and a systems model to explain trip distribution in Michigan. They show that a systems model is 30 to 40 percent 'more accurate' than a gravity model. But they used OLS in estimating the parameters of the gravity model. If they had used GLS, they may well have achieved about 60 percent improvement in accuracy and it would have been concluded that both models were equally good or that the gravity model was slightly better.

It should also be noted that if a significant χ^2 value of 'absolute' fit were found for one or both Michigan models, one would be forced to face the fact that neither the system or gravity model had done that well in explaining behaviour. Certainly there is need to be more concerned with adequacy of model structure before comparing R^2's or related measures to show that a model is better.

TABLE 3 COMPARISON OF RESIDUALS, OBSERVED VALUES AND χ^2_1*

Origin area	Pred-icted Flow	Residual (observed-predicted)	Residual Squared	Approximate $\chi 2$
1	22	14	196	.39
3	66	-3	9	.006
4	5,731	137	18,769	.14
5	2,315	-1,595	2,544,025	49.11
6	1,001	969	938,961	41.50
7	1,614	-1,236	1,527,696	42.30
8	86	-50	2,500	1.30
9	98	1	1	.0004
11	16	2	4	.011
12	601	-187	34,969	2.60
13	10	17	289	1.32
14	190	-127	16,129	3.80
15	38	88	7,744	9.18

* The following give one an idea of the probability of χ^2_1 values occurring:

$P(\chi^2_1 > 3.84) = .05$ $\quad$ $P(\chi^2_1 < .0039) = .05$

$P(\chi^2_1 < .00063) = .02$ $\quad$ $P(\chi^2_1 < .00016) = .01$