SEGMENTED REGRESSION ANALYSIS IN DEALING WITH CROSS-SECTION, QUALITATIVE DATA-- IN APPLICATION TO CONSUMER ATTITUDES TOWARD GASOLINE CONSERVATION MEASURES

Y. C. Chang and K. S. Kim
University of Notre Dame

INTRODUCTION

The purpose of this paper is to illustrate that in dealing with cross-sectional multiple regressions employing qualitative variables measured on a semantic differential scale, appropriate segmentation of the observations into separate regressions generally improves the fit of the regression equations. Within the context of marketing research, the problem of identifying market segments and the variables within each segment that are related to consumer behavior has extensively been discussed in the work by G. D. Hughes (1966) and H. L. Steele (1964). The effectiveness of using the technique of segmented regressions is further illustrated here using the recent survey data on public response to gasoline conservation measures in the United States (see Y. C. Chang and K. S. Kim [1976]).

Illustration

That segmented regression runs improve the fit of the equation for the case of qualitatively differentiable data is illustrated in Figure 1. Suppose that the respondents can be divided into two setments A and B. Let the cluster of observations encircled by points ABCDEF represent segment A; and that of observations by points abcde representing segment B. The equations estimated by least squares method are shown by line KK' and LL' for each respective segment; and by line MM' for the case in which the entire observations are treated as a single, homogeneous segment. This particular example shows that there is a noticeable improvement in the fit of the regression equation as a result of the segmentation of the respondents.

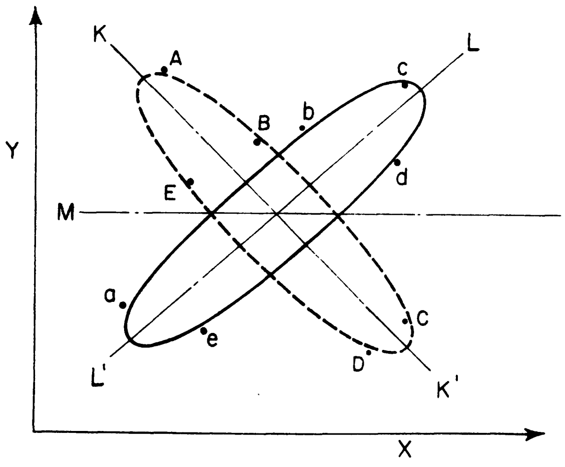

Figure 1. The scatter points and regression lines for the total and segmented population

THE DATA AND VARIABLES

The data analyzed in this study were drawn from a survey conducted jointly by Louis Harris Associates and the Center for the Study of Man in Contemporary Society at the University of Notre Dame. Responses were received from a random sample of 1665 men and women over eighteen years of age living in 100 different locations. The questionnaires included information on the personal backgrounds of the respondents and the usual demographic items. A special effort was made to obtain information on consumer attitudes toward alternative types of involuntary gasoline conservation measures. The particular data used here were obtained in February 1974 at the height of the national energy crisis.

The two questions in the survey relevant to this study were (1) Do you prefer to have a mandatory governmental rationing of gasoline (at the time 35 gallons per week) at current gasoline prices, or do you prefer to have no rationing and pay higher prices for gasoline? The respondents who indicated preferences for higher prices over a rationing system were then asked a second question: (2) How high would the price of gasoline have to go before you would prefer rationing at 35 gallons per week at current prices?

Our interest here is to identify relevant background factors that would explain differences in consumer response relative to the tolerable price level of gasoline. For this the analytical framework is set out simply by postulating a regression equation. The dependent variable is taken as the difference between the tolerable and the then current gasoline price per gallon. Thus the value of the dependent variable in the equation for a respondent opting for the rationing system would be zero.

This price differential is then assumed to be linearly related to such background and locational factors as income, age, occupation, education, sex, race, marital status, urban-rural environment, number of cars owned, as well as to the extent to which automobiles are used. Added to the regression equation are dummy variables for race (nonwhite = 0, white = 1), sex (female = 0, male = 1), urbanity (rural area = 0, urban area = 1), region (east and west coastal areas = 1, other regions =0) occupation (salesman = 1, others = 0), and marital status (single = 0, nonsingle = 1). The availability of public transportation is also included as an explanatory variable. Scores are assigned on the basis of the following four categories determined by the detree of availability: "Available" = 2, "Somewhat available" = 1, "Not available" = 0.

FINDINGS

The first experiment is to fit the regression equation to all the variable available to us using the entire sample. Where information is either missing or inadequate for a particular variable, such data have been assigned the value equal to the sample mean for the variable. The reliability of statistical estimate will not be affected by this procedure because least-squares regression always passes through the sample means of the variables in the equation. The first result shows Multiple R = .294. The F test for the regression, however, shows that it is significant at the .01 level. The low value of R has occurred because of the large amount of random noise in such a large sample in a cross-sectional study. The bulk of the variance explained by the regressions, however, is attributed to only a handful number of the variables. Table 1 reports the results of the first experiment, where the variables for which t values are at least greater than unity are reported.

In the second experiment the entire respondents have been divided into the four different regions in the United States - East, Midwest, South and West. A linear equation is fitted to each of these segments. The results are impressive. In all cases Multiple R's have greatly improved. The proportion of the variance explained by the explanatory variables is, in particular, larger for East and West. The coastal areas were in general more acutely conscious of the energy shortages during the 1974 oil crisis (Louis Harris & Associates Inc. and University of Notre Dame Report, 1974).

Comparisons of the regressions reported in Table 1, and Tables 3 and 4 show an additional interesting fact. Tables 3 and 4 indicate the results of regressions where the respondents are divided into subgroups by sex and the degree of "availability of public transportaiton" in the respondent's residence locality. These two variables have already been included as the dummy variables in the regression equation shown in Table 1. Thus, even when these dummies are included in the regression, the results indicate that Multiple R is uniformly lower in the entire sample case than in the segmented experiments. In theory, this does not have to be the case. The result of segmentation of the respondents would be particularly encouraging if these segmented groups each tend to be more homogeneous with respect to the variables examined.

Finally, variations in the coefficient value and explanatory power of the variables with the further segmentation of the sample are to be noted. Of course, differences in the value and significance of the coefficients identify the characteristics of the behavior of different subgroups. It is interesting to note that in all the cases of segmented regressions where the income and age variables are significant, income is directly related to option for higher gasoline prices while age is inversely related to it.

REGRESSION EQUATIONS INDICATING THE FACTORS DETERMINING THE OPTION FOR HIGHER GASOLINE PRICES

Note: * figures denote estimates significant

TABLE 1

THE ENTIRE SAMPLE

Variable	Reg. Coeff.	F-value
Income	.00005	7.991 *
Age	-.09961	6.877 *
Occupation	.09865	5.447 *
No. of cars owned	-.21767	3.561 *

Multiple R = .294

TABLE 2

REGRESSION EQUATIONS BY REGION

A: EAST

Variable	Reg. Coeff.	F-value
Income	.00006	8.967 *
Age	-.11236	1.006
Occupation	-.02656	.063
No. of cars owned	-.32305	1.406

Multiple R = .486

B: MIDWEST

Variable	Reg. Coeff.	F-value
Income	.00002	1.322
Age	-.07733	1.006
Occupation	.09732	1.572
No. of cars owned	.18632	.635

Multiple R = .377

C: SOUTH

Variable	Reg. Coeff.	F-value
Inc.	-.00000	.003
Age	-.24919	5.962 *
Occupation	.14954	1.427
No. of cars owned	.06737	.004

Multiple R = .343

D: WEST

Variable	Reg. Coeff.	F-value
Income	.00003	.740
Age	-.03729	.103
Occupation	.28496	5.386 *
No. of cars owned	-.08071	.058

Multiple R = .406

TABLE 3

REGRESSION EQUATIONS BY SEX

A: MALE

Variable	Reg. Coeff.	F-value
Income	.00007	3.974 *
Age	-.26030	8.998 *
Occupation	.11411	1.670
No. of cars owned	-.14590	.260

Multiple R = .374

B: FEMALE

Variable	Reg. Coeff.	F-value
Income	.00011	3.285 *
Age	-.15907	1.674
Occupation	.07034	.252
No. of cars owned	-.58154	2.559 *

Multiple R = .370

TABLE 4

REGRESSION EQUATIONS BY
AVAILABILITY OF PUBLIC TRANSPORTATION

A: AVAILABLE

Variable	Reg. Coeff.	F-value
Income	.00011	4.954 *
Age	-.08587	.464
Occupation	-.05296	.149
No. of cars owned	-.32814	.703

Multiple R = .403

B: SOMEWHAT AVAILABLE

Variable	Reg. Coeff.	F-value
Income	.00005	.928
Age	-.20248	3.091 *
Occupation	.10762	.871
No. of cars owned	-.60645	2.369 *

Multiple R - .344

C: NOT AVAILABLE

Variable	Reg. Coeff.	F-value
Income	.00007	1.476
Age	-.39740	8.443 *
Occupation	.24144	2.525 *
No. of cars owned	.05690	.032

Multiple R = .655

REFERENCES

1. Hughes, G. D. 1966 "Developing Marketing Strategy Through Multiple Regression," Journal of Marketing Research, 3, 412-415.

2. Steele, H. L. 1964 "On the Validity of Projective Questions," Journal of Marketing Research, 1, 46-49.

3. Louis Harris Associates Inc., and University of Notre Dame, 1974. Public Response to Gasoline Shortage, Report.

4. Chang, Y. C. & K. S. Kim, 1976 "The Socio-Economic Determinants of Gasoline Conservation Measures: Rationing or Higher Price?" MIDWEST AIDS Proceedings, pp. 423-426.