BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-88/09

REPORT ON MODELLING OF ELECTRIC UTILITY AVERAGE MONTHLY COSTS FOR 12 MONTH OWNERS IN LITTLE ROCK, ARKANSAS

by

Linda K. Schultz
Statistical Research Division
Bureau of the Census
Room 3128A, F.O.B. #4
Washington, D.C. 20233 U.S.A.

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233.

Recommended by: Nash J. Monsour

Report completed: March 1988

Report issued: March 1988

Report on Modelling of Electric Utility Average Monthly Costs for 12 Month Owners in Little Rock, Arkansas

by

Linda K. Schultz

I. Introduction

This report investigates the relationship between decennial census reported housing and occupancy characteristics and the cost of electricity. Housing Division began investigating this relationship upon becoming aware of a great deal of interest in the utility cost data. They learned of the interest from participants at Local Public Meetings as well as from members of the Interagency Working Group on Housing for the 1990 Census.

In regard to the census, through its investigation, Housing Division has learned of a problem with people overreporting their utility costs. As a result, in 1980 there was a major evaluation study done to investigate the effect of providing a sample of respondents with actual utility cost data. This study is described in Preliminary Evaluation Results Memorandum, PERM Report No. 59.

Since utility cost data is becoming of increasing interest, if it were possible to more accurately represent it with a statistical model an improvement in the accuracy of the census will have been made. As a result of an improvement in the accuracy of utility costs, improvements in shelter costs and gross figures will also be realized since utility cost is a component of these variables.

In PERM Report No. 59 seven cities were examined. For reasons of expedience and cost we arbitrarily elected to examine Little Rock, Arkansas more closely. With help from members of Housing Division a group of 18 explanatory variables were selected based on subject matter appeal. These variables were then used in arriving at a number of statistical models that could be used to predict electric utility cost. The work described in this report pertains only to households in Little Rock, Arkansas in which the occupants were owners that had lived in the house at least 12 months.

II. Methodology

Several different issues and considerations have been investigated in the course of the work described in this report.

A. Edits

Specifically the data were examined and were required to pass an extra edit subsequent to Housing Division edits. In the original data set of 2565 households in Little Rock, 519 households were found to have reported house values of $0 or property tax payments of $0. After consulting with subject matter specialists in Housing Division it was determined that these cases constituted problems. Since it was impossible to go back to the original transcription records the households were dropped from the study. Two other households were dropped because the reported number of people living in the households were considered to be erroneous, 130 in one case, 50 in the other. This left us with 2044 households to use in our modelling study.

B. Bias and Validation

Based on papers by Alan J. Miller and Ronald D. Snee potential bias problems occurring in the estimation of model parameters were addressed along with the importance of validation of regression models. Miller's paper, presented to the Royal Statistical Society January 1984, discussed issues in the selection of subsets of regression variables. One of his major points has to do with potential biasses in the parameters of the regression models. This can occur when the regression variables are chosen and then subsequently estimated with the same data set. Miller points out that estimating the parameters using the same data set used for model selection may result in the parameter estimates being off by 2-3 standard deviations.

Snee, in his paper, emphasizes the importance of model validation. He presents a program developed by R.W. Kennard. The program splits a data set into two equivalent groups. This allows one group to be used after model selection for model validation when it is not possible to select a new data set. In the work presented in this report Statistical Research Division personnel modified the DUPLEX program allowing it to work with a much larger data set. The original program was also modified to split the available data into three groups, one data set for model selection, another for parameter estimation and a third for model validation. In the following, we refer to each group as a partition. Splitting the original data set in this manner allowed both the potential bias problem and the validation issue to be addressed.

C. Residual Analysis

A third possibility was also examined and that addressed concern generated upon the examination of some normal probability plots. Potential transformations of the dependent variable as well as using the Least Absolute Values Regression Procedure found in SAS were both considered. It was our determination that neither improved our straight least squares fit, therefore our final models are straight normal least squares fits.

III. Results

The goal in the work presented below was to develop a model that can accurately predict average monthly electric cost using census information. Housing Division, in PERM No. 59 report, has documented that respondents overreport utility costs. The concern is that this overreporting may distort the shelter costs as well as the gross rent figures of which one component is electric utility cost.

Using the first partition of the 2044 data points the explanatory variables that explained the largest proportion of the variation of the data were selected. Average monthly cost as reported by the utility company was used as the dependent variable. A list of the explanatory variables considered in this stage of the modelling is provided in Appendix A. The average monthly electric bill as reported by the respondent, ELEC$, was without question the most important explanatory variable. When an indicator variable identifying notified versus not notified cases is added to the model containing ELEC$, the proportion of the variance explained by the model does not increase. Examination of residual plots also showed no difference between notified and not notified groups once reported electric cost was in the model. Therefore, an indicator variable was not used. The other two variables that seemed to be possibilities for the model were value of the house, VALUE, and the number of persons living in the residence, PERSONS.

Since subject matter appeal is very important three models were estimated and checked to see how well they did with validation. From a straight statistical viewpoint electric cost as reported by the respondent as well as value of the house seemed the only necessary explanatory variables. Subject matter specialists in Housing Division indicated that they felt the number of persons in the household to be a very important variable.

The three possible models for the data are as follows. The parameters were estimated using the second partition to eliminate the possibility of a bias problem.

$$\text{AVEMCOST} = 9.84 + .662 * \text{ELEC\$} \quad (1)$$

$R^2 = .66 \quad S = 10.36$

$$\text{AVEMCOST} = 4.94 + .560 * \text{ELEC\$} + .796 * \text{VALUE} \quad (2)$$

$R^2 = .70 \quad S = 9.74$

$$\text{AVEMCOST} = 1.74 + .524 * \text{ELEC\$} + .798 * \text{VALUE} + 1.85 * \text{PERSONS} \quad (3)$$

$R^2 = .71 \quad S = 9.47$

(R^2 denotes the proportion of the variation explained by the model. S is the standard deviation around the regression line.)

Several measures were run to examine the validation results. (See Appendix B for definitions.) Using the third partition and the parameters as estimated in equations (1) - (3) as well as the census reported cost the results were as follows:

Table

	Model 1	Model 2	Model 3	Census
MSE	98.67	81.50	78.81	148.29
MSRE	2.98	2.28	2.14	4.79
MARE	.2192	.1989	.1878	.2235
MAE	6.85	6.40	6.15	7.04

Examining the results in the Table above one can see that all three models improve upon the census results, this illustrates that it is possible by using a model, to improve over the respondent reported electric costs. From examining the data it appears that respondents on the average are overreporting their actual electric costs on the census forms by approximately 12%. The mean average monthly electric cost as reported by the utility company was $36.82, as reported by respondents $41.19. The average value of the household was 11.3 which translates to a house valued between $40,000 and $50,000. The average number of people residing in a household was 2.5. One interesting, although unusual result is that when validating the results with the third partition it was found that the results were actually an improvement (smaller measures) over the results found when estimating the parameters.

Conclusion

The results presented above illustrate that it is feasible to use a model to improve upon census reported electric costs. It is left up to Housing Division to determine whether the improvements discussed in this work are substantial enough to warrant further consideration. The data is available to examine other measures of performance that may more closely represent interests and questions of Housing Division. It should be noted however that the work presented here only pertains to 12 month owners in Little Rock, Arkansas. A similar analysis could easily be completed for renters in Little Rock as well as for other cities within the United States.

References

1. Copas, J.B. (1983), Regression Prediction and Shrinkage, Journal of the Royal Statistical Society B, 45, No. 3, pp. 311-354.

2. Miller, Alan J., (1984), Selection of Subsets of Regression Variables, Journal of the Royal Statistical Society A, 147, Part 3, pp. 389-425.

3. Snee, Ronald D., (1977), Validation of Regression Models: Methods and Examples, Technometrics, Vol. 19, No. 4.

4. Tippett, Janet and Takei, Richard (1983), Evaluation of Reporting of Utility Costs for Selected Cities, Preliminary Evaluation Results Memorandum No. 59.

Appendix A

Variables Used in Utility Cost Study Analysis to Model the Cost of Electricity for 12 Month Owners

- average monthly cost of electricity as reported by the utility company
- whether or not respondents were notified of their utility cost
- number of persons in household
- amount of property taxes paid
- number of rooms in house
- reported value of house
- heating equipment
- heating fuel
- fuel used to heat water
- fuel used to cook
- cost of electricity as reported by respondent
- cost of gas as reported by respondent
- number of bathrooms
- cost of water for the year
- number of bedrooms
- type of air conditioning, if any
- household income
- number of teenagers
- number of persons greater than 60 years of age

Appendix B

Mean Square Error

Units

$$MSE(E_i) = \frac{1}{N} \sum_{i}^{N} (E_i - T_i)^2 \qquad \$^2$$

Mean Square Relative Error

$$MSRE(E_i) = \frac{1}{N} \sum_{i}^{N} \frac{(E_i - T_i)^2}{T} \qquad \$$$

Mean Absolute Relative Error

$$MARE(E_i) = \frac{1}{N} \sum_{i}^{N} \left|\frac{E_i - T_i}{T_i}\right|$$

Mean Absolute Error

$$MAE(E_i) = \frac{1}{N} \sum_{i}^{N} |E_i - T_i| \qquad \$$$

where

E_i the estimate of AVEMCOST generated from a particular model (or the census)

T_i the true AVEMCOST as reported by the utility company