Imputation of Rental Equivalence in the Consumer Expenditure Survey

Barry Steinberg
Bureau of Labor Statistics

Abstract

Data for the rental equivalence of an owned home is collected quarterly by the Consumer Expenditure Interview Survey. The question asked is "If someone were to rent this home today, how much do you think it would rent for monthly, unfurnished and without utilities"? Historically, response rates for this important item have been low. We designed an estimator that will impute rental equivalence values where missings are recorded. After testing several different types of models, we chose a multiple level linear regression model to replace the existing hotdeck method. This paper will focus on a description of the final model that was chosen to be implemented.

KEY WORDS: Rental Equivalence, Owned Home, Imputation

Any opinions expressed in this paper are those of the authors, and do not constitute policy of the Bureau of Labor Statistics.

Introduction

The Consumer Price Index (CPI) aims to be a "cost of living index" rather than a "price index" in spite of its name. That means it attempts to measure the change in the cost of achieving a certain standard of living rather than the change in the cost of purchasing a certain market basket of goods and services. "Rental equivalence" is an important part of the CPI's cost-of-living framework in which homeowners are asked to estimate how much they think their homes would rent for on the open market. These estimates are considered to be the value of the shelter services provided by the dwellings to their homeowners, which are equal to their contributions to the homeowners' standard of living. These estimated "rental equivalence" values are collected by the Consumer Expenditure Survey for the CPI and are used in the weights of the CPI's rental equivalence index.

The specific question asks:

> *If someone were to rent your home today, how much do you think it would rent for monthly, unfurnished and without utilities?*

Currently, approximately 75% of homeowners participating in the CE interview provide an answer to this question while the other 25% do not. Some respondents may be reluctant to provide an answer to this question and some simply may not know what value to report.

The objective of this paper is to review the models that were recently considered to impute values for missing or invalid rental equivalence values. One of these models, the geographic dimension linear regression model, was selected for implementation beginning with April 2007 data. The motivation underlying the study was to reduce the bias in the mean rental equivalence value, thereby providing a more accurate cost weight for the single largest item in the CPI.

Background

The Consumer Expenditure Survey (CE) is a nationwide household survey conducted by the U.S. Bureau of Labor Statistics to find out how Americans spend their money. One of the primary uses of the data is to provide expenditure weights for the CPI.

The CE Survey consists of two separate surveys, the Diary and Interview surveys. The Diary Survey collects detailed expenditure data on small, frequently purchased items such as food and apparel. The Interview Survey collects detailed expenditure data on large items such as property, automobiles, and major appliances; and on expenses that occur on a regular basis such as rent, utility bills, and insurance premiums. Approximately 3,000 households are visited each quarter of the year in the Diary Survey and approximately 13,500 households are visited

each quarter of the year in the Interview Survey. This paper will focus on Rental Equivalence, which comes from the Interview Survey.

Historically, imputation rates for rental equivalence have been quite high in the CE Survey. For example, the imputation rate was 44 percent in 2003 quarter one. It fell to approximately 25 percent when the survey's method of data collection changed from paper-and-pencil to laptop computers in 2003 quarter two. Even though the need for imputation is less, concern regarding the reasonableness of the imputed values remains. Rental Equivalence accounts for nearly 25% of the CPI's weight, so its accuracy is extremely important to us.

Nonresponse's Effect on Survey Estimates

It is well known that nonresponse has the potential to introduce bias into any survey estimate when the responders and nonresponders differ in terms of the characteristic being measured. Imputation reduces this bias by inserting artificial values that are thought to be close to the true but unreported values.

Each treatment of missing values has its own advantages and disadvantages. For example, case deletion avoids the insertion of artificial values into the database, but it increases the survey's variance by reducing the survey's "effective sample size." A substantial bias can also occur if the responders and nonresponders differ in terms of the characteristic being measured. Given these negative consequences, imputation is often preferred over case deletion by data users.

However, imputation has its own set of problems. Treating imputed values as observed values generally results in downwardly biased variance estimates of the survey estimates. This occurs because the mean value is frequently used as a substitute for the unreported values, which artificially deflates the variability of the data. In addition, it artificially inflates the number of observations in the data, which can lead to further under-estimation of the variance.

Biases in a survey's point estimates and variance estimates tend to increase with the nonresponse rate, and the biases can be substantial when the nonresponse rate is high. Therefore, extreme caution is required when selecting the proper imputation method.

CE's Previous Imputation Method

Homeowners are asked to provide rental equivalence values for their homes each quarter.

The rental equivalence question is asked for the homeowner's primary residence, and if owned, the homeowner's vacation home or recreational property.

Prior to April, 2007, a hot deck method was used to impute missing rental equivalence values in the CE survey[1] In this method, a "recipient's" missing rental equivalence value was replaced with the reported rental equivalence value from a "donor" household. Recipients and donors were grouped together in cells defined by PSU[2] (Primary Sampling Unit) and building type (e.g., single family detached, town home, mobile home, etc.). There can be many potential donors in a given cell, and recipients were matched to donors within the cell according to prescribed matching criteria (number of rooms, bathrooms, half baths, presence or type of air-conditioning, and age of structure). Finally, the exact donor was randomly selected from the set of all possible matched donors.

Previous investigations found that this hot deck procedure occasionally imputed values that were inconsistent with other data the household reported, such as the buildings' estimated market values, mortgage payments, and property taxes. Additionally, it was possible for an extremely large or small donor value to be used multiple times.

Data and Methods

As a result of these problems, we examined several alternative methods to improve the imputed values. The goal was to produce more reasonable (ie, more consistent with other housing unit data) than are currently produced. Housing unit characteristics, location, property

[1] The name originates from the use of decks of computer punch cards used in processing data files with the term "hot" referring to the same data file.

[2] In the CE survey, a PSU (primary sampling unit) is essentially a metropolitan area. The CE survey collects data in approximately 100 PSUs across the United States.

type, and socio-demographic variables were considered as explanatory variables in the imputation models. Socio-demographic characteristics are not used in the current Hot Deck Model.

We analyzed five alternative models: four regression models and an alternative hot deck model. These models are described in more detail later in the paper. The model selected for implementation is a linear regression that uses a combination of socio-demographic and housing unit characteristics that are thought to influence the market value of rents and a consumer's decision making in terms of the amount they are willing to spend on rent.

Data

The data for the study were collected in 2003 quarter 2 through 2004 quarter 3 using the CAPI instrument. Each quarterly report was treated as an independent observation in the analysis. Approximately 5,600 homeowners were in the survey during this time period. As noted earlier, 75 percent reported rental equivalence. Thus, 25 percent of the homeowner sample needed to have rental equivalence values imputed.

Before testing the five alternative imputation methods, the data were examined for correlations, outliers, and missing values of independent variables for the regression models and matching criteria in the case of the alternative Hot Deck method. The model samples were further reduced to meet a minimum cell size criterion. Each step is described below.

Stratification: Preliminary analysis of the data showed that property values are correlated with rental equivalence values, but the regression coefficients varied across property codes (e.g., primary home versus vacation home) and geographic areas. Therefore, stratification by these variables was deemed necessary.

The ideal stratification should be "internally homogeneous and externally heterogeneous" (Carrington, Eltinge, and McCue (2000)). Since the surveyed units in the same property code (primary vs. vacation homes) tend to be more alike than survey units across all property groups, ignoring this information may result in increasing the variance of the estimated population means and their standard errors, and on linear regression coefficients leading to faulty inference.

In CE, the smallest geographic areas are primary sample units (PSUs), which is roughly equivalent to the Census Bureau's "metropolitan statistical areas". Each PSU is classified by its population size. (A, B, C, and D with "A" PSUs having the largest population, and "D" PSUs having the smallest population) and its geographic regions (Northeast, Midwest, South, and West. In this model, the bottom-to-top hierarchical geographic dimension structure is CPI Areas (in essence "A" PSUs) followed by Region-Size Class, Region-Urban Class, Population Size Class, and national level.

If counts within a cell are not sufficient (less than 50 observations), a higher level of geography is evaluated for sufficiency until a valid regression can be obtained. Once a geographic level is identified in which sufficiency is obtained, regression parameters for that level of geography are used in the regression model of imputed Rental Equivalence. The geographic hierarchy in which sufficiency is determined is defined below (from lowest level to highest level):

Geographic Level	**Number of Cells**
CPIAREA	42 cells
REGION*AREATYPE	16 cells
REGION*URBAN	8 cells
AREATYPE	4 cells
NATIONAL	1 cell

The data in each cell is characterized as "reported" or "missing" based on whether a Consumer Unit reported a Rental Equivalence value. Cell size sufficiency is determined by a count of reported observations within a cell. If counts are sufficient to run a valid regression, the regression parameter estimates for that cell are used to calculate imputed rental equivalence values.

Outliers Identification

One of the most challenging tasks in our modeling efforts was to detect outliers. Both univariate and bivariate tests were used to detect outliers. A general 1% top-coding (winsorizing) of the extreme values was applied to the input data for primary residences. For vacation homes,

we used studentized residuals for top-coding due to small sample sizes.

Filtering: As mentioned earlier, separate regressions were run for each geographic area. It was decided that a minimum of 50 valid data records were needed to run a regression in each geographic area. Here a "valid data record" is the information for an individual property in which the reported rental equivalence value is greater than \$0 and at least 50% of the regression's independent variables have reported values. When a geographic area has fewer than 50 valid data records, broader geographic areas are examined until one is found that has at least 50 valid data records. Regression parameters from the broader area are then applied to the reported data in original lower-level geographic area. We called this the "50/50 rule."

Imputations for missing values of independent variables

After processing the 50/50 rule, a conditional mean method was used for imputing the missing values of independent variables. For this step, the sample was divided into two sub-samples: (1) those with valid rental equivalence data, (a recorded rental equivalence value greater than \$0) and (2) those with invalid rental equivalence data, (a value for rental equivalence that is missing or less than or equal to \$0). For the valid rental equivalence sub-sample, the missing values of the independent/matching variables were replaced with the mean of the reported values. For the invalid rental equivalence sub-sample, the missing values of the independent/matching variables were replaced with the mean of reported values from both the valid and invalid rental equivalence subsample. The nonresponse mechanism for the missing independent/matching variables was assumed to be missing completely at random (i.e. the probability of these missing values were to be independent to the reported feature values and values from other variables) in our dataset.[3]

[3] Rubin (1976) classified the missing value mechanism into three subgroups: Missing Completely At Random (MCAR) where the probability of a value to be missing is independent of that feature value, and independent of the values of all other features, Missing At Random (MAR) where the probability of a value to be missing independent of that feature value, but may be depended upon other features' values, Non-ignorable where the probability of a value to be missing may depend up the actual feature of the value. This classification has been widely discussed since the eighties and becomes more and more popular in statistical model estimation.

Models examined

The following models were analyzed to replace the current Hot Deck Model:

1. Geographic Dimension Regression Model
2. Geographic Dimension Double Log Regression Model
3. Capitalization Rate Model
4. Branched Regression Model
5. Alternative Hot Deck Model

Geographic Dimension Linear Regression Model

This model is used to impute the missing rental equivalence values for each property type at the geographic area level. Reported rental equivalence was regressed on selected homeowner socio-demographic characteristics, housing characteristics, and owned-home property value. See Appendix A for variable descriptions.

Two forms of the model were proposed: one for the household's primary residence and one for vacation homes and recreational properties.

(a) The primary residence model at each CPI Area level is defined as follows:

$$RNTEQVX_{jk} = \beta_0 + \beta_1(EDUINDEX_{jk}) + \beta_2(AGE\_REF_{jk}) + \beta_3(KIDGT11_{jk}) + \beta_4(OCSPO910_{jk}) + B_5(PROPVALX_{jk}) + \beta_6(BATHRMS_{jk}) + B_7(BEDROOMQ_{jk}) + \beta_8(AGEBUILD_{jk}) + B_9(CENTAC_{jk}) + \varepsilon_{jk},\ \varepsilon_{jk} \sim (0, \sigma^2),$$

where *RNTEQVX* is the rental equivalence value, *j* denotes a geographic area, *k* denotes individual records *(k = 1, .., K).* $\beta_0, \beta_1, \ldots, \beta_9$ are the regression parameters. The independent variables were examined and found to be independent of each other and uncorrelated with the disturbances ε, meeting the underlying assumptions of ordinary least squares regression (OLS).

(b) The vacation home or recreational property regression model is defined as follows:

$$RNTEQVX_k = \beta_0 + \beta_1(NEWPROPX_k) + B_2(EDUINDEX_k) + \beta_3(TMSHAR1_k) + \beta_4(TMSHAR2_k) + B_5(TMSHAR3_k) + \beta_6(PRPVALD1_k) + B_7(EDUINDD1_k) + \beta_8(PRPVALD2_k) + B_9(EDUINDD2_k) + \varepsilon_k, \ \varepsilon_k \sim (0, \sigma^2),$$

where $RNTEQVX_k$ is the rental equivalence value, *k* denotes individual records *(k = 1, .., K).* $\beta_0, \beta_1, \ldots, \beta_9$ are the regression parameters. All imputations are done at the National level due to small sample sizes.

Geographic Dimension Double Log Regression Model

The imputation and thin cell adjustment methods in this model are similar to the previous Geographic Dimension Regression Model but with the following variations: (1) different geographical dimension structure, (2) different function form assigned to the dependent variable and asset price variables, and (3) no distinguishing among the property codes. In this model, the bottom-to-top hierarchical geographic dimension structure is CPIAREAS followed by Region-Size Class, Population Size Class, and national level.

Capitalization Rate Regression Model

In the economics literature, the user cost of owned housing can be interpreted as a capitalization rate.[4] In a simplified model, this is the rate at which rents (in a perfectly competive market), *R*, are discounted into asset prices, *V*.

$$V = \frac{R}{i}$$

Algebraic manipulation gives:

$$R = i * V$$

The capitalization rate *i* can be estimated either by a ratio of mean rents to mean market values, or as a regression parameter estimate. For this study, a regression model was estimated. The parameter value was multiplied by the market value of each owned home for which rental equivalence was missing or invalid.

Branched Regression Model

This model is a simple univariate regression based on property value. However, if the property value is missing, this method branches off to other regression models using other variables such as Annual Property Taxes, Total Rooms, Bathrooms, Half Bathrooms, Year Built, and level of Education.

Alternative Hot Deck Model

This model is very similar to the Current Hot Deck Model with the exception that Property Value is used instead of PSU.

Testing and Evaluation

The diverse composition of the models tested precluded us from using traditional statistical methods to compare their effectivenss. More specifically, the Alternative Hot Deck Model was not an ordinary least squares model. Therefore, we used rank tests.

Rank tests are valid for all types of populations whether continuous, discrete, or a mixture. The specific rank test chosen for this evaluation was the Kruskal Wallis test which can analyze several samples simultaneously, one from each of k different populations. In this project, each population represents the set of imputations from a particular model that we tested. We tested the null hypothesis that all of the imputation sets have the same distribution against the alternative hypothesis that some of the models tend to furnish greater observed values than other models.

[4] See, for example, Green, Richard K. and Stephen Malpezzi, *A Primer on U.S. Housing Markets and Housing Policy*, AREUEA Monograph Series No. 3,, The Urban Institute Press, Washington, DC, 2003, pp.58-59.

The value of interest for this analysis was the sum of the absolute values of the deviation (difference) between the "true" mean or average of the model means and the predicted (imputed) value. For each model/data quarter combination, the average of the absolute deviations is given a rank to be compared against the other models and data quarters. For example, if the test uses six models and six quarters of data, then there would be ranks with values 1 through 36. The smallest deviation would have a rank of 1, the second smallest, 2, and so on to the largest value of 36. In general, if a model consistently outperforms the rest, the sum of their ranks will be noticeably lower than the others.

Kruskal Wallis Test Results

MODEL	Sum of Ranks
Geographic Dimension Regression Model	23
Geographic Dimension Double Log Regression Model	63
Capitalization Rate Regression Model	103
Branched Regression Model	111
Alternative Hot Deck	165

Using the Kruskal Wallis Test, the **Geographic Dimension Regression Model** consistently performed at or near the top compared to the other models. This provided strong support for using this model to impute missing rental equivalence values. As a result of this finding, we recommended changing the imputation method for missing rental equivalence values in the CE from the current Hot Deck model to regression. This recommendation was accepted and was implemented beginning in April, 2007. It is expected that this new imputation approach will lead to more accurate rental equivalence values and will reduce the effect of outliers compared to the current Hot Deck method. The real goal was to reduce bias in the mean rental equivalence due to the Hot Deck imputation and thereby providing a more accurate cost weight for the single largest item in the CPI market.

The test statistic T is the sum of the ranks for the measurements in sample i after the combined sample measurements have been ranked. The Chi Square distribution with k-1 degrees of freedom is used as an approximation to the null distribution of T. This value will be compared to the critical value of the Chi Square to see if there is statistical significance. If and only if the null hypothesis is rejected, there is a procedure to determine which pairs of populations (models) tend to differ.

Results

Results from the Kruskal Wallis Rank test are presented in the table below:

Reference

Carrington, W., J. Eltinge, and K. McCue (2000), "An Economist's Primer on Survey Samples." Center for Economics Studies, U.S. Bureau of the Census, Working Paper 00-15, October.

Conover, W. J., Practical Nonparametric Statistics (3rd edition), John Wiley & Sons, Inc., 1999, pp. 288-290

Duan, Naihua (1983), "Smearing Estimate: A Nonparametric Retransformation Method", Journal of the American Statistical Association, September 1983, Volume 78, Number 383.

Eberwein, C., Olsen, R. and Reagan, P. (2004), "Intracluster Correlation and Complex Sampling: Do Geographic Data Lessen the Problem?" Mimeo, Ohio State University, September.

Garner, Thesia (2004), "Incorporating the Value of Owner-Occupied Housing in Poverty Measurement", Discussion Paper, Workshop on Experimental Poverty Measures, Committee on National Statistics The National Academies, Washington, DC.

Johannessen, Randi (2004), "Owner-occupied Housing in Norway: Why the rental equivalent approach is preferred", Paper presented at the Eight Meeting of International Working Group on Price Indices, Helsinki, Finland, August.

Rubin, D.B. (1976), "Inference and missing data", Biometrika vol. 63,581-90.

Sande, I. G. (1983), Hot-Deck imputation procedures in Incomplete data in sample surveys. Vol.3. Proceedings of the symposium. W.G.Medow, and I.Olkin (eds). New-York: Academic press. 334-350.

Schafer, J.L., & Graham, J.W. (2002), Missing data: Our view of the state of the art. Psychological Methods, 7, 147-177.

Sharpe, **William** (1964), Capital asset prices: A theory of market equilibrium under conditions of risk,
Journal of Finance 19, p.425 – p.442.

Verbrugge, Randy (2004), "The Puzzling Divergence of Aggregate Rents and User Costs, 1978-2001." Paper presented to the International Conference on Index Number Theory and the Measurement of Prices and Productivity, Vancouver, British Columbia, June 30-July 3, 2004

White, H. (1998), "Artificial Neural Network and Alternative Methods for Assessing Naval Readiness," San Diego: NRDA Technical Report.

.Appendix A: Variable Description

Dependent Variable	Description	Data Type
RNTEQVX	**Rental equivalent of property**	**Numeric**
Independent Variable		
EDUINDEX	Education attainment of the reference person. 1 = Less than High School 2 = High School Graduate 3 = Some College 4 = Bachelors degree 5 = Post graduate	Numeric
AGE_REF	Age of reference person	Numeric
KIDGT11	Number of children whose age is 12-17	Numeric
OCSPO910	Is Spouse not working but taking care of home/family 1 = Yes 0 = No	Binary
PROPVALX	Estimated market value of the property	Numeric
BATHRMS	Number of baths + halfbaths in this unit	Numeric
BEDROOMQ	Number of bedrooms	Numeric
AGEBUILD	Age of the Building	Numeric
CENTAC	Does this unit have Central Air Conditioning? 1 = Yes 0=No	Binary
TMSHAR1	Dummy Variable to indicate timeshares in the regression model	Binary
TMSHAR2	Dummy Variable to indicate vacation homes in the regression model	Binary
TMSHAR3	Dummy Variable to indicate whether utilities are included in the estimate of rental equivalence	Binary
PRPVALD1	Interaction variable for PROPVALX using TMSHAR1	Numeric
EDUINDD1	Interaction Variable for EDUINDEX using TMSHAR1	Numeric
PRPVALD2	Interaction variable for PROPVALX using TMSHAR2	Numeric
EDUINDD2	Interaction Variable for EDUINDEX using TMSHAR2	Numeric