TRUNCATION BIAS IN ENERGY DEMAND EQUATIONS: EFFECTS OF MISSING SURVEY DATA

Lisa A. Skumatz, Pacific Gas and Electric Company
Suzanne Holt, University of California, Santa Cruz
Richard S. Barnes, Pacific Gas and Electric Company
Paul Ong, University of California, Los Angeles

ABSTRACT

Survey data traditionally suffers from two types of non-response: systematic failure to return surveys, and non-response to selected questions by respondents (item non-response). The bias from item non-response was the focus of this analysis. Item non-response can cause problems in two ways. First, estimates of group means, proportions (saturations), and other simple statistics will be biased if the item non-response is not random (that is, if certain types of customers systematically leave items blank). Second, regression and other analyses based on the data will produce biased coefficients, especially since most statistical packages delete the observation if any variable in the analysis is missing. The estimation is based on a truncated data set.

Exploratory work on the extent and consequences of the bias was performed at PGandE using a large, residential customer database containing responses on structural characteristics, appliance holdings, and demographic information. The results of a probit analysis showed that there were systematic, predictable differences between those respondents who are informed about energy characteristics and those who are not. An analysis of the extent of bias introduced in using truncated samples and samples augmented via "hot deck" techniques for saturations and model-building was performed. The work led to suggestions on the most effective methods for "imputing" information for missing data.

TRUNCATION BIAS IN ENERGY DEMAND EQUATIONS: EFFECTS OF MISSING SURVEY DATA

Lisa A. Skumatz, Pacific Gas and Electric Company
Suzanne Holt, University of California, Santa Cruz
Richard S. Barnes, Pacific Gas and Electric Company
Paul Ong, University of California, Los Angeles

ABSTRACT

Missing data is an important problem in survey databases. Some standard statistical packages exclude observations with missing data. Using truncated samples for saturations or model-building can lead to bias. Exploratory work on the extent and consequences of the bias was performed at PGandE using a large, residential customer database containing responses on structural characteristics, appliance holdings, and demographic information. The results of a probit analysis showed that there were systematic, predictable differences between those respondents who are informed about energy characteristics and those who are not. An analysis of the extent of bias introduced in using truncated samples and samples augmented via "hot deck" techniques for saturations and model-building was performed. The work led to suggestions on the most effective methods for "imputing" information for missing data.

INTRODUCTION

Survey data traditionally suffers from two types of non-response: systematic failure to return surveys, and non-response to selected questions by respondents. The bias from item non-response can cause problems in two ways. First, estimates of group means, proportions (saturations), and other simple statistics will be biased if the item non-response is not random (that is, if certain types of customers systematically leave items blank). Second, regression and other analyses based on the data will produce biased coefficients, especially since most statistical packages delete the observation if any variable in the analysis is missing. The estimation is based on a truncated data set.

The purpose of this research effort was to examine the presence of this problem in an existing large residential survey data base collected by Pacific Gas and Electric Company (PGandE), and assess ways of correcting for this truncation in energy analyses either through imputation or other methods. The data base contains approximately 15,000 responses to a questionnaire on structural characteristics, appliance holdings, and demographic information weighted to represent PGandE's 3.5 million residential customers. The data were collected in 1983. Several key variables important in energy analyses have traditionally suffered from high item non-response. These are:

- o whether the housing unit has ceiling insulation (18% non-response in the sample)
- o whether the unit has wall insulation (25% sample non-response)
- o the space heating fuel (12% non-response)
- o the water heating fuel (10% non-response)
- o the square footage of the dwelling (36% non-response)
- o the age of the dwelling (17% non-response)
- o the household income. (21% non-response)

These are important analytical variables affecting energy usage, customer energy demand modeling and conservation market research.

The first part of the project determined the extent to which failure to report building energy-related variables might be related to the costs and benefits to the resident of learning the information. Survey respondents may be uninformed about the energy characteristics of their home because the acquisition of information is costly or benefits of the information are low. The decision to be uninformed may be a rational response to these conditions. Other occupants may only become informed about energy-related factors of their residence through experience which is not available to others.

The behavior by which occupants gain information is not likely to be random, but to vary systematically with factors that influence the costs and benefits of being informed. Therefore, informed and uninformed respondents might be expected to differ significantly in these factors. When they differ, customer profiles and energy demand forecasts based on the truncated data set are likely to be biased, reflecting only the distinctive characteristics of informed respondents. In essence, uninformed respondents select themselves out of the data base.

PROBIT ANALYSES OF DIFFERENCES BETWEEN INFORMED AND UNINFORMED RESPONDENTS

A probit approach[1] was used to account for differences in factors influencing the costs and benefits of obtaining building energy information. The analysis found that there are systematic, predictable differences between those respondents who are informed about energy characteristics and those who are not, indicating significant bias results from using truncated samples for saturations and model-building.

---

[1] Probit analysis was selected over logit techniques for several reasons. Results from probit are easily translated into Mill's ratios for use in the analyses presented later in this paper. Also, we are not aware of work that shows equivalent inverse Mill's ratios can be derived from logit analyses. Finally, SAS probit procedure provides the calculation of inverse Mill's ratios as an option.

The dependent variables for the probit analysis were defined as:[2]

KNOWC - does the respondent know whether the unit has ceiling insulation;
KNOWW - does the respondent know whether the unit has wall insulation;
WTRFU - does the respondent know the type of fuel used to heat water;
SPFU - does the respondent know the type of fuel used for space heating;
SQFT - does the respondent know the unit's square footage.

Each dependent variable is coded as a 1 if the respondent is informed, and a 0 if uninformed. The probit analysis predicts the likelihood that a respondent is informed on the basis of the values of the explanatory variables. The explanatory variables used fall into several basic categories: information benefits, information costs, and experience.

Information Benefits

Information about the energy characteristics of housing is valuable because it creates opportunities for savings on fuel bills. It follows that the larger an occupant's fuel bill, the more information on energy characteristics would be valued. Therefore, it is expected that factors correlated with higher fuel bills would lead to an increase in the value of information, and an increased likelihood that the respondent is informed.

The explanatory factors fall into several groups:

- o household size related. Variables affecting the fuel bill through the size and energy appliance characteristics of the unit include number of bedrooms, whether or not the home is air-conditioned, and whether the unit is a single-family type. Number of residents is expected to influence the water-heating use.

- o climate zone related. Households located in more adverse climate zones are also expected to have a greater stake in obtaining information about the energy characteristics of the housing unit. These effects are captured in variables measuring the number of heating degree days and cooling degree days.

- o income related. A measure of income is included to reflect the higher demand for comfort and the fact that higher income households tend to generate higher fuel bills.

---

[2] Note: Probit analyses were also conducted on the two other variables, income and age of dwelling (INC, and AGE). The following discussion of benefits and costs of energy-related information does not apply directly to these variables. A discussion of the techniques applied for these variables is included in the probit results sections for these variables.

A special variable noting whether the respondent was a PGandE employee was included, because it was expected that Company employees might have stronger preferences for energy conservation and might be more likely to be informed. Measures of energy price were not generally included as explanatory variables because of the strong relationship with size of household and climate variables.

Information Costs

Information about the energy characteristics of housing can be costly to obtain. Information costs can be thought of as investments that are amortized over the time the household occupies a housing unit - the longer the amortization period, the lower the amortized information costs. In addition, the lower the discount rate used to annualize information costs, the lower the amortized information costs. For the household, it becomes worthwhile to acquire information as long as amortized information costs are lower than the expected savings in fuel bills made possible by the information.

Information costs are expected to be lower, and households are more likely to be informed about energy characteristics if they are PGandE employees, owner occupants, or have higher incomes. Owner occupants are assumed to have learned some characteristics in the process of house hunting, and are likely to stay in the unit longer, lowering the amortized costs. Economic evidence from studies of consumer durables purchases[3] suggests that people with higher incomes have lower discount rates. In terms of the information hypothesis, these households would face lower amortized information costs, and would therefore be more likely to be informed.

Other factors affecting information costs include the age of the dwelling. Newer housing units were built under stricter energy code standards, decreasing the cost of obtaining energy-characteristic information. Single-family dwellings provide easier access to insulation, making it easier for a household to acquire information about the presence of insulation. This is also true of space and water heating fuels, and may also be true for age of dwelling.

Experience

Independent of the value or cost of information, households may become informed about the energy characteristics of their housing unit by experience. Factors that increase the opportunity to observe energy

---

[3] See for example Hausman (1979). "Individual Discount Rates and the Purchase and Utilization of Energy-Using Durables", Bell Journal of Economics, 10, 33-54., and Train (1985). "Discount Rates in Consumers' Energy-Related Decisions: A Review of the Literature", Energy, 10, 1243-1253.

characteristics increase the likelihood of being informed. Several variables reflecting experience were included in the analysis. In general, the longer the household has occupied the residence, the more likely it is that they have become informed about the unit's energy characteristics. However, this is likely not true in gathering information on square footage - that information is generally collected at the time of first occupancy, and becomes less important as time goes on. Households that have installed new heating or cooling systems, or have weatherized their units in the last 12 months are more likely to have observed and become informed about insulation and space-heating fuels. Those having installed a new water-consuming appliance in the last 12 months are also more likely to have observed or become informed about water-heating fuel.

Results of the Probit Analysis of Energy-Related Variables

Table 1 defines the explanatory variables and gives a summary of the results of the probit analysis. A plus sign indicates the estimated coefficient was significant and positive. A minus sign indicates a significant negative coefficient. A zero indicates an insignificant estimated coefficient.

In every equation, the chi square statistic gives 99.9% confidence that the set of explanatory variables distinguish the informed from uninformed respondents. Several variables show overall strong performance in the equations. The variable indicating single family residence performs well in the shell and space heating equations. The indicator of owner-occupied households is strong in the building shell and both fuel equations. The variable indicating a new house is strongly related to the probability of knowing information about insulation. The length of time the household has occupied the dwelling is significant and positive in the equations as expected, and has the expected negative sign in the equation for knowing the square footage. The variable indicating the installation of weatherization or heating/cooling systems is significant and of the expected sign in all equations in which it is included. The indicator of whether the household has a Company employee shows a significant positive coefficient. The variables indicating household size (either number of bedrooms or number of residents) have significant positive effects for every characteristic but space heating fuel type.

Probit analysis on the age of dwelling

The dependent variable for this equation is defined as 1 if the respondent reports the unit's age and 0 if not. This variable is treated differently than the other energy characteristics. A prospective occupant may use the apparent age of a housing unit as an initial signal of its quality. However, even brief experience occupying the unit provides more direct and informative signals of housing quality in general or energy efficiency in particular. Therefore, age is a relatively low-valued type of information.

TABLE 1 Results of the Probit Analysis

VARIABLE/DEFINITION	EXPECTED SIGN	KNOWC[a]	KNOWW	WTRFU	SPFU	SQFT	AGE	INC
CBEDROOM - Number of bedrooms	+	+	0		-			
NUMRES - Number of occupants	+			+				
H65 - Avg. Daily Htg. degree days	+	0	+	0	+	+		
C75B - Avg. Daily Cooling DD	+		0			+		
AIR - '1' if have air-conditioning	+	0	0					
SINGLFAM - '1' if single family unit	+	+	+		+			
INCOME - Total Annual Household Income ($5000's)	+	0	+	0	0	+	+	
SAMPEMP - '1' if household member is PGandE employee	+	+	+	+	+	+		+
OWNERS - '1' if owner-occupant	+	+	+	+	+	+	+	-
NEWHOUSE - '1' if unit built since 1974	+	+	+		-			
MAXAGE - Age of oldest household member	-							-
RESTIME - Years at current address	+	+	+	+	+	-	+	
ENMOD - '1' if weatherization or new heating or cooling system installed within last year	+	+	+		+	+	+	
WATERAPP - '1' if new water-consuming appliance in last 12 months	+			0				
CSQFT - square footage in feet								-
SPA - '1' if household has private spa								0
POOL - '1' if have private pool								0
MICROWAVE- '1' if have microwave								0
RESPTYPE - '1' if response obtained:								
DMAIL - during mail phase								-
DPHONE- during phone follow-up								-
DINTERV-during interview follow-up								-

Key: '+' indicates positive coefficient significant at 95% level
'-' indicates negative coefficient significant at 95% level
'0' indicates insignificant coefficient
'blank' indicates variable not included in equation

[a] Variables KNOWC, KNOWW, WTRFU, SPFU, SQFT, AGE, and INC defined in the text.

Four variables are used to explain why households may be informed about their unit's age: income, owner occupied, length of residency, and whether or not weatherization or heating/cooling systems have been modified. Households with higher incomes are assumed to have a greater demand for all kinds of information about their housing including its age. Owner-occupants are assumed to face lower costs of acquiring any information about their housing. Experience factors, including length of residency and weatherization or system modifications are expected to increase the likelihood that the respondent is informed.

The results in Table 1 show that all the explanatory variables have the expected positive effects on information and are statistically significant.

Probit analysis of whether or not respondent reports income

The dependent variable for this section is defined as 1 if the respondent reports household income and 0 if not. This variable is treated differently from the energy characteristics because it reflects respondents' concerns for privacy rather than the benefits and costs of information. An analysis of this variable was provided because it was one with consistently low response, and is an important variable in many types of energy analyses.

Ten variables are used to distinguish those who report income from those who do not. PGandE employees are expected to be more cooperative, while owner-occupants, older households, and households with larger dwelling units are expected to have stronger privacy concerns. Secondary measures of income were also included, including the presence of spas, pools, and microwave ovens. Dummy variables were also created to reflect the stage in the survey process at which the response was actually collected: the initial mail stage; the first follow-up telephone; or the latter in-person interview stage. These would allow differences based on the demonstrated willingness to respond to the survey.

The results of the probit analysis for reporting of income weakly confirm most of the hypotheses. All the significant variables have coefficients of the expected sign. The variables representing luxury items are not statistically significant. The percentage of respondents with spas, pools, or microwave ovens does not appear to vary substantially between those who report income and those who do not. The results indicate that although income suffers from a high degree of non-response, there is little evidence of a strong systematic element distinguishing the respondents from non-respondents. Therefore, income non-response may not be a significant source of bias.

Implications of the Probit Analysis Results

The probit results show that a truncated data base, consisting of sets of responses from informed respondents only, is unreliable in calculating saturations and will result in biased estimates. The populations of respondents and nonrespondents are significantly different. Also, truncated data, when used to fit regressions, will lead to biased coefficients that will confound the effects of end-use related variables with the influence of the variables explaining the likelihood of responding to the question. For example, occupants of single-family units are more likely to be selected into the sample because they are more likely to be informed about energy characteristics. But they are also expected to consume more energy. Therefore, demand estimation ought to distinguish the influence of single-family units on survey response from their influence on energy use.

TESTS FOR BIAS IN ENERGY ANALYSES

In order to examine the extent of bias resulting in estimated regression coefficients and the effectiveness of possible correction procedures, conditional demand equations were estimated using three methods:

- estimation over the truncated data set;
- estimation over a data set with "hot deck" imputations;
- estimation over the truncated data set using a "Heckman"[4] technique.

A "hot deck" imputation procedure has traditionally been used at PGandE to impute likely responses for customers with missing data. This technique begins by segmenting the customers, trying to group similar customers into "cells". The variables used for segmentation differ from question to question depending on preliminary data analysis. For example, the database might be segmented by dwelling type, climate zone, etc. Then, for each cell, the procedure would randomly assign responses for customers with missing data based on the distribution of responses from similar customers with non-missing data.

The "Heckman" technique is a bias-reducing technique that introduces an extra set of variables into the regression equations. These variables, called inverse Mill's ratios, are related to the probabilities that were estimated in the probit analysis, and have the effect of capturing the bias in the extra coefficients. Mill's ratios were derived for each of the seven probit variables. The inverses of these ratios serve as indexes of factors distinguishing informed from uninformed respondents, factors that might mistakenly be omitted from the demand equation. These terms are included as additional explanatory variables in the energy demand forecast

---

[4] The "Heckman" method was developed in Heckman (1979). "Sample Selection Bias as a Specification Error", Econometrica, 47, 153-161.

equations for the data set truncated to include only informed respondents. Significant coefficients on these variables provide confirmation of the bias problem, and the inclusion of these terms corrects for the bias in the coefficients on the end-use variables.

This series of analyses tested whether using simple imputation procedures (hot deck methods), which eliminate the truncation problem, adequately reduce bias or lead to other problems which only more complicated procedures can address. Although the procedure estimates missing data for individual observations using values observed among those respondents who are most like the individual in some characteristics, the existing procedures do not use the systematic differences between informed and uninformed respondents, which may lead to bias in the energy demand coefficients. In addition, while synthetic estimation preserves a large data base, it introduces random error, potentially more than exists in the population. This would be reflected in a large root mean squared error term (MSE) for forecasts based on synthetic estimation.

Results of the Energy Demand Equations

The results of conditional demand equations for both electricity and gas are presented in Table 2. Columns 2 and 3 in Table 2 give estimated energy use factors for the hot deck and the Heckman-type estimations. These factors are calculated as the sum of the products of the coefficients of the explanatory variables included in the conditional demand equation and the mean level of the variable. Although similar to UECs,[5] they differ in that the coefficients are multiplied by the mean level of the variable for all households rather than the mean for only those households with the end-use. The units of the variables in Table 2 are kilowatt-hours per day and therms per day.

Bias in End-uses

A comparison of the factors presented in Columns 2 and 3 of Table 2, estimated via hot deck and Heckman methods, show that the coefficients on several end-use variables were significantly different. The results suggest that factors calculated from the hot deck coefficients are biased relative to those calculated from the Heckman coefficients. Generally, the results show upward bias when estimating the conditional demand equations on data imputed via PGandE's hot deck method.

---

[5] Unit Energy Coefficients (UECs) are generally defined as the average energy consumed by an end-use over the household that have that end-use. The energy use factors in this study refer to the energy consumed by an end-use taken over all households.

TABLE 2 Results of Demand Equations for KWH and Therms[a]

	TRUNCATED END-USE[b]	HOT DECK END-USE[b]	HECKMAN END-USE[b]
ELECTRIC END-USES			
Miscellaneous[c]	3.438	3.134	2.362
Refrigerators	4.659	4.806	4.315
Freezers	1.118	1.188	1.018
Other Single Term End-Uses[d]	5.694	5.863	5.479
Elec Central A/C	0.952	0.987	0.950
Window/Wall A/C	0.125	0.117	0.127
Elec Primary Htg: Heat Pump	0.163	0.188	0.157
Resistance Htg	0.659	0.670	0.628
Elec Aux. Htg.	0.333	0.288	0.314
Elec Water Htg.	1.709	1.593	1.654
Pool Pump	0.460	0.472	0.477
Spa or Hot Tub	0.126	0.135	0.118
Total	19.436	19.441	17.599
GAS END-USES			
Gas Space Htg	1.249	1.178	1.146
Aux. Gas Htg	0.015	0.012	0.011
Gas Water Htg	0.533	0.564	0.420
Gas Pool Htg	0.008	0.011	0.009
Miscellaneous[c]	0.046	0.061	0.039
Total	1.851	1.826	1.625
SUMMARY STATISTICS			
Electric:			
N	76721	142725	76721
MSE	49.75	52.93	49.43
F	2519.3	3842.8	2292.3
Adj. R^2	.617	.581	.622
Gas:			
N	37424	106580	37424
MSE	5.13	5.59	5.07
F	8517.2	18286.2	7053.7
Adj. R^2	.694	.623	.696

[a] Units for reported results are kilowatt-hours per day and therms per day. All results are significant at 95% confidence level.
[b] The estimates are derived using estimated coefficients multiplied by the means for the variables over the entire data set, not only those households including the end-use.
[c] The miscellaneous category for gas consumption consists mostly of cooking and clothes drying. Miscellaneous usage for electricity includes lighting and all other miscellaneous appliances.
[d] The term "other single term end-uses" is composed of: cooking, microwave oven, well water pump, waterbed, television, dishwasher, and clothes dryer end-uses.

The most affected electric end-uses are:

- o the miscellaneous category,[6]
- o refrigerators and freezers,
- o heat pumps, and
- o spas.

The most affected end-uses for gas are:

- o water heating, and
- o the miscellaneous category.[6]

These biases represent an estimated difference in average household consumption of 1.8 kilowatt-hours per day and .2 therms per day, a bias of approximately ten percent of average household consumption.

The energy use factors estimated from the hot deck data overestimate demand for some end-uses for both gas and electricity. Further, the sum of these factors across all end-uses for the Heckman method does not add to the average kilowatt-hours or average household therms consumption, indicating bias in the underlying synthetically estimated observations. The coefficients on the inverse Mill's ratios (not presented) were found to be significant, indicating systematic self-selection.

Treatment Effects

Because bias was evident, an additional analysis was done to try to identify the source of the bias. The hot deck data was used to estimate special conditional demand equations that included "treatment effects". Two treatment effects were estimated with the purpose of breaking down the bias evident into its possible sources: self-selection bias (represented by the introduction of probability variables into the equation); and bias introduced by the hot deck imputation methodology (represented by the synthetic estimation dummy variables). The first treatment introduced the predicted probabilities that the respondents are informed as additional explanatory variables in the equation, and are included for the entire data set. The second treatment included the values of the synthetically estimated variables for those respondents who are not informed as additional variables.
+ The results showed that the first treatment effect was significant, especially for the income term, and the second treatment effect showed significant, but small, impacts. These results imply that the source of bias is related to self-selection, and not to biases in the imputation procedure itself.

---

[6] The miscellaneous category for gas consumption consists mostly of cooking and clothes drying. Miscellaneous usage for electricity includes lighting and all other miscellaneous appliances.

SUMMARY AND CONCLUSIONS

The exploratory work described in this paper examines the bias caused by missing data in survey data bases and its implications for imputing data and bias in empirical analyses. Seven variables with low response rates were examined: whether the unit has ceiling or wall insulation, the space and water heating fuels, the square footage and age of the dwelling, and household income. The research was conducted in two parts.

Probit analysis techniques were used to examine whether there were systematic differences between respondents and non-respondents to a series of energy-related variables, and whether the differences were related to the costs and benefits of being informed about household energy characteristics. The second part of the work examined the effects of the differences on a sample energy-related analysis: conditional energy demand equations. The results of conditional demand equations with truncated and hot deck imputations were compared to a Heckman specification to check for bias. Implications for data imputation were noted.

The conclusions derived from the work include the following.

- The probit analyses indicated there is significant bias inherent in the non-reported variables we examined. There are systematic, predictable differences between survey respondents who are informed about energy characteristics of their housing and those who are not.

- The truncated data set produces biased saturations and analytical coefficients.

- The "hot deck" method "fills in" the data base without introducing significant additional bias and decreases variability around saturations.

- In the conditional demand analysis, the estimates over the truncated sample show that additional correction is needed.

- Further correction beyond PGandE's existing hot deck method is needed for the conditional demand equations. Significant truncation bias remains after the hot deck imputation. The hot deck imputation can also introduce random error, although the evidence shows this to be small in this case.

- The Heckman estimation method theoretically corrects the bias problems. However, the technique requires multiple estimation procedures, and the treatment demand equations estimated in this study do not dramatically improve on goodness of fit over the synthetic demand equation.

- Statistically significant differences between respondents and non-respondents were noted. These differences are statistically significant for predicted demand for electricity and natural gas for certain end-uses: miscellaneous, refrigerators, freezers, heat pumps, electric spas, and gas water heating.

As a follow-up to this project, PGandE staff will be performing a final comparison of the difference in saturation estimates derived from the hot deck and discrete choice techniques. This analysis is expected to show little difference between the estimates. Then PGandE will expand the imputation variables list to include other variables with significant item non-response. The results of the probit analyses will be used to improve synthetic estimation procedures. However, the simpler hot deck procedures will be used in most cases because the procedures are easier to implement and the benefits to the extra work needed for the probit analysis may not be justified.

Conditional demand equation estimation would not use the imputed data for coefficient estimates. Instead, the Heckman technique will likely be applied to future demand estimation to derive unbiased coefficients, and the "corrected" saturations would be used in calculating the final UEC estimates.

The approaches and conclusions in this paper have applications to other data bases. However, the specific results concerning the size and direction of bias in conditional demand equations are of course specific to this test case. Until more analysis is done it cannot be determined whether other energy analyses might be significantly more or less affected by the use of truncated or synthetically estimated data.