FLEXIBLE MATCHING IMPUTATION IN THE AMERICAN HOUSING SURVEY

Kimberly Long, Bureau of the Census
Statistical Research Division, Room 3134-4, Washington, D.C. 20233

Keywords: Imputation, linear regression, hierarchy, matching, respondent, nonrespondent

Introduction

The American Housing Survey (AHS) is designed to provide a current series of information on the size and composition of housing units , the characteristics of its occupants, the changes in the housing stock unit resulting from new construction and from losses, the indicators of housing and neighborhood quality, the characteristics of recent movers, and the characteristics of urban and rural housing units. The AHS is done for both a Metropolitan Sample (MS) and a National Sample (N). The MS is conducted in 44 selected metropolitan areas which are divided into four groups of 11 each; each group is interviewed once every four years on a rotating basis. The National Sample covers the United States, inside and outside metropolitan areas, urban and rural, and in the four census regions. Prior to 1984, the AHS was called the Annual Housing Survey. The name was changed to American Housing Survey since the National sample is no longer conducted annually but every other year in odd numbered years. The sample size in 1989 consisted of approximately 49,400 housing units located throughout the United States. The Metropolitan Sample consisted of approximately 35,200 housing units.

The problem of missing survey items occurs when some or all of the responses are not collected for a sample housing unit. Item nonresponse occurs when some but not all of the required responses are collected; in contrast, total nonresponse occurs when none of the responses are collected. Item nonresponse occurs because of item refusals, such as for the income questions, "don't know" responses, or omission of questions by the interviewer. Also, artificial item nonresponse occurs when an item fails edit. Total nonresponse occurs because of unwillingness to participate in the survey, incapacity of the respondent to participate, or "not at home's". The standard method for handling item nonresponse is imputation, that is, assigning values for the missing responses and the standard method for handling total nonresponse is weighting procedures.

Imputation has two major advantages. First, complete data methods of analysis can be done on the imputed data set. Second, the time required to impute large data sets will only need to be spent once, by the data producer. As with any process, there are disadvantages to imputation. Most data users treat the imputed data set as complete data; therefore, the variability due to imputing missing values is lost. Also, correlations could be extremely biased. Some alternatives to imputation are to discard the incomplete data and to analyze the complete data only or to model the incomplete data based on maximum likelihood methods. Each of these methods has advantages as well as disadvantages. The first, discarding the incomplete data, is easy to do, but if a large amount of the data is missing, this method may not be very efficient. The modeling approach allows flexibility in analyzing the data and avoids the use of ad hoc methods, but it requires a large amount of work to find an appropriate model.

Each housing unit surveyed in the AHS is defined as a record, where a respondent record is a record with all items answered and a nonrespondent record is a record with one or more items, but not all, missing for a specified group of items. Total nonrespondents are discarded. The current procedure in the AHS for handling item nonresponse is a sequential hot deck imputation, where hot deck refers to a procedure in which the value assigned for the missing response is taken from a respondent record to the current survey and sequential refers to the order of filling the missing responses with previous records. A new procedure being studied is a modified hot deck procedure called flexible matching imputation, where information contained in the variables is used to match records. This procedure stratifies the records into imputation groups, where each of the groups has a group of variables associated with it; this group consists of explanatory variables that are always present, such as personal characteristics. Also, the imputation variables, variables that need to be imputed, are used as explanatory variables. The nonrespondent records are matched with respondent records using a hierarchical approach, in the sense that if a nonrespondent record cannot be matched with a respondent record, then a variable is dropped and the match is tried again. This procedure is currently used in the March Supplement of the Current Population Survey (CPS) conducted by the Census Bureau.

The purpose of this research was to find an imputation method that will work for both the National Sample and the Metropolitan Sample. The variables with nonresponse rates greater than 3% were targeted for the new imputation procedure. The following sections describe the flexible matching procedure and discuss the advantages and disadvantages of the flexible matching procedure compared to the current sequential hot deck procedure.

Data Preparation

The data used for this research, the 1987 AHS National file, were stratified by urban / rural and the four Census regions, Northeast, Midwest, South, and West. Within each stratified data set, the variables were divided into imputation groups, such as mortgage related variables, equipment failure variables, neighborhood / housing condition variables, and housing cost variables. Each of the imputation groups was assigned a group of variables to help in imputation.

Before imputation can begin, the data must be sorted by applicable variables and missing variables within each imputation group. Since there are skip patterns on the questionnaires, not every item gets a response. For example in the mortgage related variable imputation group, new and assumed mortgages are distinguished and therefore different questions are asked for each, such as the number of years assumed for assumed mortgages and the number of years mortgaged for the new mortgages. This is what is meant by applicable variables. Next, the data set is divided into nonrespondent records and respondent records. Again, a nonrespondent record is a record that has at least one variable missing within the imputation group but not all missing. Now, the nonrespondent records are sorted by missing patterns, that is, they are sorted according to which variables are missing - variable one, variable two, variable three, etc. After sorting, the records are coded into a matching file. For example, if a variable has a value of x and the 10^{th} percentile is y where $0<x\leq y$ then the coded value is 1. If the 20^{th} percentile is z and $y<x\leq z$ then the coded value is 2. This is done for the percentiles 10, 20, ..., 100. The data collected are rounded to the nearest whole amount, such as the price of a home; the term " continuous" will be used to define the variables that are continuous but are rounded off. Only "continuous" variables are coded; categorical variables are left

alone. This matching file is used in the imputation procedure.

Hierarchy of variables

The hierarchy of matching variables for continuous imputation variables is determined by a forward stepwise regression procedure. Stepwise regression is a search procedure that develops a sequence of regression models, including or deleting a variable at each step in order to produce a model with a subset of "good" variables. The standard univariate forward stepwise regression starts by fitting a simple regression model for each of the potential variables. The variable with the largest F value is the variable first entered. Next the procedure fits all two variable models that contain the first variable entered; then, the variable with the largest F^* value (partial F test) is selected as the next variable to be entered. The next stage would compute all three variable models containing the first two variables entered and then the third variable would be selected. This pattern continues until the F^* values do not exceed a predetermined significance level or no more variables remain. The first variable entered is labeled the most important. The multivariate procedure is done similarly to the univariate procedure except for the function used to determine the entering variable. Here, the function is the absolute value of the trace of the matrix ϵ where $\epsilon = e_f - e_r$. e_f is the matrix of the sum of squares of error for the full model, the model that contains all the variables in the imputation group, and e_r is the matrix of the sum of squares of error for the reduced model, the model that contains the variables selected by each step of the stepwise routine. The model which has the smallest trace value is selected as the most important variable. The regression models contained only the linear effects, no interactions, since the individual values are used to match.

The univariate procedure is used for finding matching variables for patterns having one variable missing and the multivariate procedure is used for patterns having more than one variable missing. Five has arbitrarily been chosen as the maximal number of matching variables to use since the probability of finding a match decreases as the number of variables used for matching increases. The forward stepwise procedure will select the five variables using the information contained in the respondent records.

Imputation

After the hierarchy of variables is established, the nonrespondent records are matched with the respondent records. Using the matching variables for each missing record, a match is found among the respondent records. That is, variable x's coded value in the nonrespondent record matches variable x's coded value in the respondent record. There is a successful match at level 1 if all the matching variables can be matched. If no match can be found, the least significant variable is dropped and a match is tried at level 2 -- all variables except the least significant one. This pattern continues, i.e., dropping the least significant variable if no match is made, until a match is ultimately made at level 5, the number of matching variables. When one match has been made, the search for the next match starts on the next respondent record, that is, the one after the record used for imputation. If the procedure fails and no match is found, a simple imputation procedure is preformed where predetermined explanatory variables are used to match. The variables used to match will always be present; some examples are race, age, and sex of the reference person.

After a match is found the matched file refers back to the original data file and substitutes the missing responses of the nonrespondent record with the original responses of the matching record.

Example

For research purposes, a subset of the mortgage imputation group was selected to test. The questionnaire distinguishes between assumed mortgages and new mortgages; the records used in this example were new mortgages, since the skip patterns are different for the two. Appendix 1 lists the complete mortgage imputation group variables and the group of variables chosen as explanatory variables. The subset is 1) *AMTMORT*, the original amount mortgaged for the home, 2) *MMP*, the current monthly mortgage payment for the home, 3) *LEN*, the length of the original mortgage, 4) *INTRATE*, the current interest rate of the mortgage and 5) *PRICE*, the original price of the home. Also, a subset of explanatory variables was selected; they are as follows: 1) *TYPEM*, the type of mortgage, 2) *DOWNP*, the source of the downpayment, 3) *UNIT*, the type of housing unit, 4) *TAXES*, the real estate taxes, 5) *HHINC*, the household income, 6) *POVPERC*, the household income as a percent of the poverty level, 7) *ROOMS*, the number of rooms in the home, 8) *AGE*, the age of the reference person, 9) *SEX*, the sex of the reference person, and 10) *RACE*, the race of the reference person. *UNIT, TAXES, HHINC, POVPERC, ROOMS, AGE, SEX,* and *RACE* are the explanatory variables that are always present. *LEN, AMTMORT, INTRATE, MMP, PRICE, TYPEM* and *DOWNP* are imputation variables that will also be used as explanatory variables. In order to obtain the respondent data set which is used in the stepwise regression procedure, the records must be sorted by missing patterns. Since there are 7 imputation variables, the number of missing patterns is 2^7-1=127. This count includes the pattern where all variables are present in the imputation group and the patterns where at least one variable is missing in the imputation group, but excludes the pattern where all variables in an imputation group are missing. A 127x7 matrix will be constructed containing 0's and 1's representing if a variable is present (1) or missing (0). This matrix represents the missing patterns. Each record is checked and assigned the appropriate missing pattern, then the frequency of the patterns are computed.

Next, the records will be coded into a matching file where the values in the matching file are the variables coded by percentiles. For example, if the variable *AMTMORT* has the value 20,000 and the 10th percentile is 21,000 then the coded value is 1. If *AMTMORT* equals 25,000 and the 20th percentile is 30,000 then the coded value will be 2. This is done for the percentiles 10, 20, ... , 100.

Consider the pattern where AMTMORT is missing. *AMTMORT* will be regressed on the variables in the explanatory group using the univariate/multivariate forward regression procedure where the five most important variables will be chosen. The hierarchy of the variables is given in Table 1. The table lists all of the stratified data sets and the rankings obtained in each data set. There is not a considerable difference in the rankings between the different regions and urban/rural areas. Generally, each data set has a core set of variables selected. The one variable most related to *AMTMORT* is *MMP* , the current monthly mortgage payment, which is ranked 1 or 2 in all the data sets; likewise, *PRICE* , the original price of the home, is ranked 1, 2 or 3 in all the data sets.

With the hierarchy completed for the patterns, flexible matching imputation can be done. Using the matching file, the nonrespondent records are matched with the respondent records. A match is found among the respondent records using the explanatory variables in each missing pattern. For example, in the South Urban data set, the first variable to be dropped will be *DOWNP* when no

match can be found in the first pass of the respondent data, the second to be dropped will be *TYPEM*, the third, *LEN*, etc. Table 2 lists the percentage of variables matched in the data sets. More of the patterns, 32%, matched on all five variables than matched on any fewer number of variables.

Appendix 2 shows the variables after imputation and Appendix 3 shows which variables were used to match with and which variables were dropped if any. In Appendix 3, the label *MATCHED* signifies that the record was matched successfully on all five matching variables. If the label is *MATCHED* with the labels *Variables Matched* and *Variables Dropped*, then the record was matched on only the variables specified. For example, Record 5 has *MMP, PRICE,* and *LEN* as matched variables and *TYPEM* and *DOWNP* as dropped variables. This concurs with the hierarchy given in Table 1 where *DOWNP* is the lowest ranked and *TYPEM* is the second lowest. To examine the imputed data set, the plots of *AMTMORT vs. PRICE, AMTMORT vs. MMP,* and *MMP vs. PRICE* were constructed. These three variables have pairwise linear relationships. The plot of the respondent records and the plot of nonrespondent records were studied; similar patterns were seen in both sets of plots. These similarities are expected results because of matching; the results would be different if only records with unusual combinations of matching variables need to be imputed. Appendices 4 and 5 show the respective plots for *AMTMORT vs. PRICE*.

A reasonable impute would be a value of *AMTMORT* that is similar to the *PRICE*, functionally related to *MMP, LEN,* and *HHINC*, etc. Likewise, the record that needs *LEN* imputed should receive an imputed value related to *INTRATE* and *MMP*. In Appendix 2, record 1 shows *PRICE*=$26,500, *MMP*=$352, and *LEN*=30 yrs; the imputed value for *AMTMORT* is $27,500 which is similar to the original price of the home, $26,500 and functionally related to *MMP*. An example of a bad impute, not included in Appendix 2, is an extreme case with *AMTMORT*=$114000, *MMP*=$975 and *PRICE*=$3000. If these extreme cases are present in the data set then some bad imputes are expected.

Conclusion

In contrast to the sequential hot deck method which does not allow for any measure of closeness in regard to matches, the flexible matching procedure uses as many variables as possible to match and allows for some measure of closeness. In conclusion, the flexible matching procedure is a simple and fast way of obtaining imputes and the results of the procedure show overall reasonable imputes. Most of the nonrespondent records were matched with respondent records using three or more variables. This procedure captures the timely changes in the respondent data which is used to select the variables used for matching; different variables might be selected each year based on the collected responses. Also, the required knowledge of the user is minimal; for example, stratification of the data set is required, the distinction of the continuous and categorical variable is required, and some idea of the relationship of variables. All of these things are easily attainable by a quick scan of the data. Based on these results, the application of flexible matching imputation in the AHS looks promising.

Present/Future Research

Presently, this procedure will handle imputation of continuous variables only. Since categorical variables can not be predicted with a linear regression model, a new approach has to be taken. Instead of a linear regression model, a log linear model is used. Iterative proportional fitting of a log linear model to a contingency table is being studied to establish the hierarchy of the matching variables for categorical imputation variables. The statistic used is the χ^2 value which corresponds to the F value for the linear regression model. One of the problems encountered is that the size of the contingency table is too large; a suggested solution is to break the groups into small sets.

Imputes based on extreme values, values near the upper and lower end of the range of the variable, of matching variables seem to suffer from this procedure since all matches are based on percentiles. A supplemental method is available that will help these imputes; the method uses ratio of a highly correlated matching variable and the matching record as follows: $x_a = \frac{y_a}{y_r} * x_r$

where y is a matching variable for x. The imputed value is x_a, the value of the matching variable for the nonrespondent is y_a, the value of the matching variable for the respondent is y_r, and x_r is the respondent value used to impute for x_a. This supplemental procedure is only available for the continuous variables that have proven relationships and only one variable is used in the ratio.

Other topics being studied are 1) the possibility of using incomplete records to impute, (e.g. If a nonrespondent record is missing variables x and y, then a nonrespondent record missing variable z could be used to impute), and 2) the effect of the correlation of variables between the different imputation groups since all of the variables are not independent.

References

Bailar, J.C. and Bailar, B., (1982), "Comparison of two procedures for imputing missing survey values," *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 462-467.

Kalton, G. and Kasprzyk, D. (1982), "Imputing for missing survey responses," *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 22-31.

Kalton, G. and Kasprzyk, D. (1986), "The treatment of missing survey data," *Survey Methodology,* 12, 1-16.

Little, R. and Rubin, D. (1987), *Statistical analysis with missing data*, New York: John Wiley and Sons.

Rubin, D. (1987), *Multiple imputation for nonresponse in surveys*, New York: John Wiley and Sons.

U.S. Department of Commerce, U.S. Department of Housing and Urban Development, Economic and Statistics Administration, Bureau of the Census, Office of Policy Development and Research; *American Housing Survey for the US in 1989.* Current Housing Reports, Series H150/89, Washington, D.C.

This paper reports the general results of research undertaken by the Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

Table 1
Data Set vs. Hierarchy of Explanatory Variables for Imputing AMTMORT
Key: * = imputed variable; 1 = most important variable; 5 = least important variable; NE = Northeast; S = South; R = Rural; MW = Midwest; W = West; U = Urban

Variable	NE-R	MW-R	S-R	W-R	NE-U	MW-U	S-U	W-U
loa	3	3	3		3	3	3	4
amtmort	*	*	*	*	*	*	*	*
intrate	5			2		4		2
mmp	1	1	2	1	1	1	1	1
typem			5	3	5		4	5
price	2	2	1		2	2	2	3
downp		5	4	5			5	
unit								
taxes								
hhinc		4						
povperc								
rooms								
age					4	5		
race	4							
sex				4				

Table 2
Data Set vs. Percentage of variables matched per record
Key: NE = Northeast; S = South; R = Rural; MW = Midwest; W = West; U = Urban

DATA SET	1	2	3	4	5
NE-U	1	15	22	14	48
MW-U	3	13	30	19	35
S-U	2	12	15	14	57
W-U	1	16	15	18	49
NE-R	5	15	6	22	52
MW-R	4	23	28	13	32
S-R	3	16	25	14	42
W-R	5	26	11	13	45

Appendix 1

The mortgage imputation group is listed below. The variables highlighted are the ones selected for the example.

1. Mortgage number (MORTNUM)
2. Bought mortgage/home in same year (SAMEYR)
3. Assumed or new mortgage (NEW_ASS)
4. Amount assumed (AMTASS)
5. Years on assumed mortgage (YRSASS)
6. Year of mortgage (YRMORT)
7. **Length of original mortgage (LEN)**
8. Number of years of pay loan (YRSPAY)
9. **Original amount mortgaged (AMTMORT)**
10. Mortgage include other homes (OTHERH)
11. Mortgage include farm land (INCFARM)
12. Mortgage include a business (INCBUS)
13. Amount of mortgage for own home (AMTOWNH)
14. **Current interest rate (whole value) (INTRATE)**
15. Current interest rate (fraction value) (INTFRAC)
16. **Current monthly mortgage payment (MMP)**
17. Payment include property tax (INCTAX)
18. Payment include insurance (INCINS)
19. Payment include anything else (INCANY)
20. Amount of other charges (OTHCHAR)
21. **Type of mortgage (TYPEM)**
22. Source of loan (SCLOAN)
23. Source was former owner (FORMOWN)
24. Same payment for loan duration (SAMEPAY)
25. Reason for change in payment leadin (LEADIN)
26. Reason for change in payment (CHGREAS)
27. % of loan in last payment (PERC)
28. Year purchased home (YRPURC)
29. Year received home (YRREC)
30. **Original price of home (PRICE)**
31. **Source of downpayment (DOWNP)**
32. Government program mortgage (GPM)
33. Number of mortgages (NUMMORT)
34. Value of house and property (VALUE)

The following variables have been chosen as explanatory variables for mortgage imputation group. The variables highlighted are the ones used in the example.

1. **type of structure/unit (UNIT)**
2. **age of reference person (AGE)**
3. **sex of reference person (SEX)**
4. **race of reference person (RACE)**
5. **number of rooms in unit (ROOMS)**
6. **household income (HHINC)**
7. **income as a percent of the poverty level (POVPERC)**
8. number in the household (HHNUM)
9. urbanized area (URBAN)
10. **real estate taxes (TAXES)**

Appendix 2

PATTERN RECORD	LEN	AMTMORT	INTRATE	MMP	TYPEM	PRICE	DOWNP	UNIT	TAXES	HHINC	POVPERC	ROOMS	AGE	SEX	RACE
2 1 IMPUTED RECORD	30	27500	12	352	1	26500	2	1	272	24100	2092	5	54	1	2
COMPLETE RECORD MATCHED	30	27500	10	378	1	28000	2	1	644	29400	2552	5	28	1	1
2 2 IMPUTED RECORD	30	49000	11	641	1	55000	2	1	707	49864	6556	6	30	2	1
COMPLETE RECORD MATCHED	30	49000	12	624	1	49000	2	2	750	43000	4704	6	36	1	1
2 3 IMPUTED RECORD	15	99000	10	1337	4	16000	2	1	1500	55000	4775	8	29	1	1
COMPLETE RECORD MATCHED	30	99000	12	1138	2	99000	8	1	1100	84000	7292	8	42	1	1
2 4 IMPUTED RECORD	30	62000	13	643	4	65000	2	1	236	28000	3063	5	27	1	1
COMPLETE RECORD MATCHED	30	62000	10	698	4	65400	2	1	450	44000	5785	6	25	1	1
2 5 IMPUTED RECORD	15	25000	14	577	2	32000	1	4	345	23600	2579	6	54	1	1
COMPLETE RECORD MATCHED	15	25000	9	500	4	35000	1	1	296	32000	2687	6	55	1	1
2 6 IMPUTED RECORD	30	29000	9	370	4	32000	2	1	1200	75500	8497	9	51	1	1
COMPLETE RECORD MATCHED	30	29000	9	366	4	32900	2	1	995	58950	6635	6	49	1	1
2 7 IMPUTED RECORD	30	80000	9	896	1	92650	2	1	1243	53680	2915	6	59	1	4
COMPLETE RECORD MATCHED	30	80000	9	872	1	89000	2	1	1435	58000	7626	9	34	1	1
2 8 IMPUTED RECORD	15	30000	7	299	4	75000	2	1	1000	71000	9335	5	35	1	1
COMPLETE RECORD MATCHED	15	30000	10	327	4	70000	2	3	600	47000	6179	4	42	2	1
2 9 IMPUTED RECORD	15	25000	10	330	4	29500	8	1	100	25000	2735	4	23	1	1
COMPLETE RECORD MATCHED	15	25000	10	300	4	25000	8	1	330	20000	1736	6	29	1	1
2 10 IMPUTED RECORD	30	93000	11	1700	4	16000	1	1	2100	95000	9997	9	51	1	1
COMPLETE RECORD MATCHED	30	93000	9	1000	4	150000	2	1	1800	95000	6837	10	45	1	1
2 11 IMPUTED RECORD	30	41520	9	540	4	56000	1	2	980	20000	3385	6	44	2	1
COMPLETE RECORD MATCHED	30	41520	9	550	4	51900	1	1	1454	54100	6089	7	42	2	1
2 12 IMPUTED RECORD	30	80000	10	950	4	109000	2	1	1200	55000	9308	7	28	1	1
COMPLETE RECORD MATCHED	30	80000	12	1139	4	96000	2	1	927	125000	9997	9	34	1	2
2 13 IMPUTED RECORD	15	78000	10	850	1	87000	1	2	1200	55500	7297	6	43	2	1
COMPLETE RECORD MATCHED	15	78000	9	811	1	95000	1	1	900	70000	7657	6	37	1	1
2 14 IMPUTED RECORD	25	31500	5	369	4	32500	2	1	228	1	1	8	61	1	2
COMPLETE RECORD MATCHED	25	31500	12	367	4	37500	4	1	102	36350	3976	5	30	1	1

Appendix 3

Pattern : 2 Record : 1 MATCHED

Pattern : 2 Record : 2 MATCHED

Pattern : 2 Record : 3 MATCHED
Variable(s) Matched : mmp
Variable(s) Dropped : price len typem downp

Pattern : 2 Record : 4 MATCHED

Pattern : 2 Record : 5 MATCHED
Variable(s) Matched : mmp price len
Variable(s) Dropped : typem downp

Pattern : 2 Record : 6 MATCHED

Pattern : 2 Record : 7 MATCHED

Pattern : 2 Record : 8 MATCHED

Pattern : 2 Record : 9 MATCHED

Pattern : 2 Record : 10 MATCHED
Variable(s) Matched : mmp
Variable(s) Dropped : price len typem downp

Pattern : 2 Record : 11 MATCHED

Pattern : 2 Record : 12 MATCHED

Pattern : 2 Record : 13 MATCHED

Pattern : 2 Record : 14 MATCHED
Variable(s) Matched : mmp price len typem
Variable(s) Dropped : downp

Pattern : 2 Record : 15 MATCHED
Variable(s) Matched : mmp price len typem
Variable(s) Dropped : downp

Pattern : 2 Record : 16 MATCHED

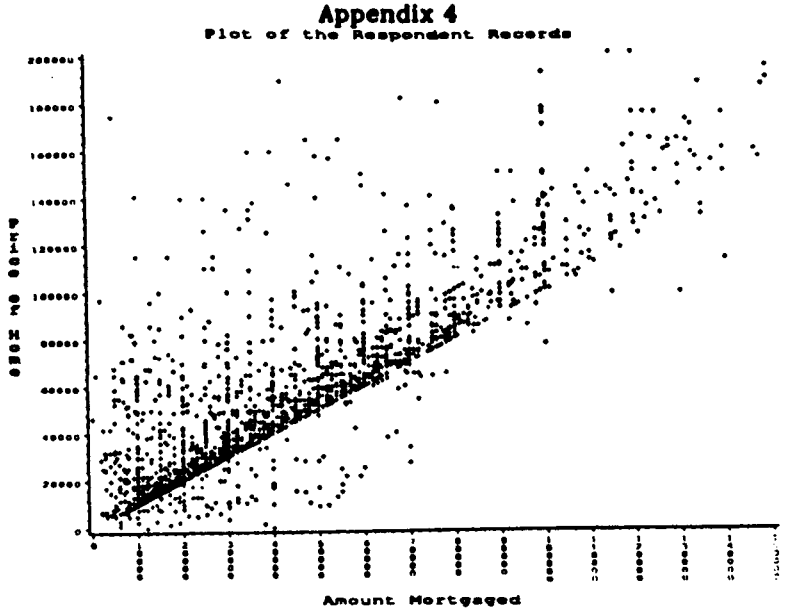
Pattern : 2 Record : 17 MATCHED
Variable(s) Matched : mmp price len typem
Variable(s) Dropped : downp

Pattern : 2 Record : 18 MATCHED

Pattern : 2 Record : 19 MATCHED
Variable(s) Matched : mmp price len typem
Variable(s) Dropped : downp

Pattern : 2 Record : 20 MATCHED

Appendix 4



Appendix 5