**THE SURVEY OF INCOME AND PROGRAM PARTICIPATION**

# An Investigation of Model-Based Imputation Procedures Using Data from the Income Survey Development Program

No. 11

Vicki J. Huggins and Lynn Weidman

U.S. Census Bureau

June 1986

U.S. Department of Commerce U.S. CENSUS BUREAU

# THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

## AN INVESTIGATION OF MODEL-BASED IMPUTATION PROCEDURES USING DATA FROM FROM THE INCOME SURVEY DEVELOPMENT PROGRAM

No. 8603 11

by
Vicki J. Huggins and Lynn Weidman
Bureau of the Census

SIPP Working Paper #11

June 1986

Acknowledgements

This paper was prepared by Vicki J. Huggins and Lynn Weidman of the Statistical Research Division Bureau of the Census.

---

Suggested Citation

Huggins, Vicki J. and Lynn Weidman. "An Investigation of Model-based Imputation Procedures Using Data from the Income Survey Development Program," SIPP Working Paper Series No. 8603. Washington, D.C.: U.S. Bureau of the Census, 1986.

# Survey of Income and Program Participation

AN INVESTIGATION OF MODEL-BASED
IMPUTATION PROCEDURES USING DATA FROM THE
INCOME SURVEY DEVELOPMENT PROGRAM

No. 8603

by
Vicki J. Huggins and Lynn Weidman
Bureau of the Census

June 1986

Acknowledgements

This paper was prepared by Vicki J. Huggins and Lynn Weidman of the Statistical Research Division Bureau of the Census.

---

Suggested Citation

Huggins, Vicki J. and Lynn Weidman. "An Investigation of Model based Imputation Procedures Using Data from the Income Survey Development Program," SIPP Working Paper Series No. 8603. Washington, D.C.: U.S. Bureau of the Census, 1986.

TABLE OF CONTENTS


Introduction . . . . . 1

Creation of Estimation Files . . . . . 3

Model Estimation . . . . . 4

Overview . . . . . 4

Discussion of Estimation Results . . . . . 4

Receipt of Wages . . . . . 5

Medicaid Receipt . . . . . 6

Earnings Amounts . . . . . 6

Weeks with Pay . . . . . 7

Missing Wage and Salary Records . . . . . 7

Imputation Results . . . . . 8

Conclusions . . . . . 9

Recommendations for Further Study . . . . . 9

Appendixes A. Models Fit to the Data . . . . . 11

B. Variable Transformation Used in Fitting Models . . . . . 22

C. Earnings Amounts . . . . . 26

## INTRODUCTION

The purpose of this study is to investigate the feasibility of using model-based imputation methods for record nonresponse in a longitudinal survey. Record nonresponse means that the responses to an entire set of questions (record type) are missing for a wave. In this study we have selected four variables to model and impute: (i) rcpt = receipt of earnings; (ii) wpay = weeks worked with pay; (iii) earn = earnings amount; and (iv) maid = Medicaid coverage. Maid is on the person (P) record and the others on the wage and salary (WS) record. For any wave, a person may respond to neither record type, to P, or to both. So the first three variables are reported or missing simultaneously and maid may or may not be missing at the same time as the others.

In order to reduce the amount of data manipulation required in this study, we want to select a subset of the available ISDP waves. The methods we envision will impute months in their order of occurrence, so that all previous months of data are available at the time a given month is imputed. Thus, we will use three waves of data--waves 1 and 2 will be complete data and the variables in the months of wave 2 will be modeled. Wave 3 will include missing record types so that we may model the relationship of missing variables to responses in wave 2. We will use only one rotation group in order to reduce the amount of data manipulation required and any complications which would be caused by waves overlapping for different rotation groups; i.e., all data will cover the same three waves and 9 months.

Previous study of the relationship between demographic and employment-related variables has shown that the race (white, nonwhite) and sex status of a person is an important factor. For this reason we will attempt to put the data into four race-sex cells and model each one separately. This, in effect, models the interaction of race-sex with all the other variables in the model. Because of the small number of records available for use after fulfilling the data requirements introduced in the previous paragraph, we may not be able to fit models for all four race-sex cells. Or we may have to reduce the number of variables in some of the models.

For the data in each cell, we must estimate models and evaluate imputations which use these models. The imputations are done by month and within month a specified order of variables is used. When imputing a variable, the current month value of all previously

imputed variables on the same and other record types are available, as are observed variables from other record types. All previous month variables are available as are all following month variables that are observed.

Of the four variables we are modeling, two of them will be treated as continuous (weeks with pay and earnings) and two of them as categorical (receipt of earnings and Medicaid coverage). Each month for each variable will be modeled separately. The explanatory variables will include those shown in Table 1 and values of some demographic variables in wave 2. For the categorical variables we will fit logit models and for the continuous variables, linear regression models.

**Table 1:** Months of Variables Used in Fitting Models

| | | Variable in Model | | | |
|---|---|---|---|---|---|
| Month Modeled | Variable Modeled | rcpt | wpay | earn | maid |
| 4 | rcpt | 1,2,3,7 | 3,7 | 2,3,7 | 3,4,5 |
| | wpay | - | 1,2,3,7 | 2,3,7 | - |
| | earn | - | 1,2,3,7 | 1,2,3,7 | - |
| | maid | 3,4,5 | 3,4,5 | 3,4,5 | 1,2,3,7 |
| 5 | rcpt | 2,3,4,7,8 | 4,7 | 3,4,7 | 4,5,6 |
| | wpay | - | 2,3,4,7,8 | 3,4,7 | - |
| | earn | - | 2,3,4,7,8 | 2,3,4,7,8 | - |
| | maid | 4,5,6 | 4,5,6 | 4,5,6 | 2,3,4,7,8 |
| 6 | rcpt | 3,4,5,7,8,9 | 5,7 | 4,5,7 | 5,6,7 |
| | wpay | - | 3,4,5,7,8,9 | 4,5,7 | - |
| | earn | - | 3,4,5,7,8,9 | 3,4,5,7,8,9 | - |
| | maid | 5,6,7 | 5,6,7 | 5,6,7 | 3,4,5,7,8,9 |

The numbers are the months for which the variable at the top of the column is used in modeling the variable at the left.

We will discuss three major stages in this study:

1. Creation of data files that include nonresponse to be used for estimating model parameters.
2. Estimating models and searching for those most applicable.
3. Imputing values onto a data file for comparison with originally reported values.

Following that, we will present conclusions and recommendations for further study.

## CREATION OF ESTIMATION FILES

A file of records to be used for model estimation was created for each of white males, white females, and nonwhites. Because of the small number of records of nonwhites available in our selected data set, we were not able to separate them by sex. When estimating models for variables in wave 2, we must allow for record types WS, WS and P, or neither being missing in each of waves 2 and 3. The records of complete respondents for wave 1 were separated into two sets.

i) Both record types reported in wave 2. The following response patterns occurred for wave 3. (R = reported, M = missing)

| record P | type WS | number |
|---|---|---|
| R | R | 1217 |
| R | M | 169 |
| M | M | 16 |

ii) One or both record types missing in wave 2. The following response patterns occurred for waves 2 and 3.

| wave 2 | | wave 3 | | |
|---|---|---|---|---|
| P | WS | P | WS | number |
| R | M | R | R | 22 |
| R | M | R | M | 220 |
| M | M | M | M | 24 |
| R | M | M | M | 7 |
| M | M | R | R | 2 |
| M | M | R | M | 1 |

We will not simulate records with the last three patterns because of their small frequencies of occurrence.

For each demographic group, each record in (i) is assigned one of the first three patterns from (ii) or not used, according to a set of probabilities. The records selected for use are written out to form the estimation file for that group.

The following are the counts of these patterns for the three estimation files:

| wave 2 | | wave 3 | | white | white | non- |
|---|---|---|---|---|---|---|
| P | WS | P | WS | male | female | white |
| R | M | R | R | 15 | 18 | 28 |
| R | M | R | M | 181 | 166 | 84 |
| M | M | M | M | 10 | 7 | 22 |

MODEL ESTIMATION

Procedure

There are 36 cases in this study for which models can be estimated--3 sex/race groups x 4 variables x 3 months. Because of previously determined prevalance of change in response to questions from wave to wave, more models were fit for month 1 of wave 2 than for months 2 and 3. We have not had time to examine in detail all the models estimated. These include:

| | |
|---|---|
| month 1, wave 2: | rcpt - white female, nonwhite |
| | earn - white female |
| | wpay - white female, nonwhite |
| | maid - nonwhite |
| month 2, wave 2: | earn - white female |
| month 3, wave 2: | wpay - white female, nonwhite |

also missingness for WS in wave 3 for all records combined.

Table 1 lists the months of data for each of these variables used when estimating a model for one of these variables in a specific month. The actual terms in the models are given in appendix A and their definitions in appendix B.

The statistical package GLIM (Generalized Linear Interactive Modeling) was used for modeling. It will estimate both linear regression and logit models, as well as many others. There are two main reasons it was selected: 1) it tell- -e user when there are linear dependencies among the independent variables and leave -ne linearly dependent variables out of the model; and 2) it is easy to add terms to ( -elete terms from an existing model interactively. It also performs transformations i calculations with variables and arrays.

For each case estimated, several models were fit by adding to d subtracting from independent variables used in a prior fit. This was done to find m -ıels that used fewer terms without significantly decreasing the closeness of the model fit. In the case of

linear regression we can actually perform F-tests to determine the effect of an increase or decrease in the number of terms included. For the logit models there are only asymptotically approximate chi-square tests (see appendix A), so we use our judgement to decide on a model to use for imputation. The measure of fit given by GLIM is the scaled deviance, which is the residual sum of squares for linear regression models.

Appendix A includes tables of models fit that include terms in the model, scaled deviance, and degrees of freedom. Some of the cases were modeled extensively to get a good idea of how the different variables affected the fit, but only a few models were tried for most cases.

## Discussion of Estimation Results

### Receipt of Wages

Logit models were fit in order to estimate the probability that a person did or did not receive wages in a given month. A difficulty encountered was that only a small percentage of persons reported no receipt of wages. For wave 2, month 1, the counts are

i. 10 of 191 white females
ii. 2 of 206 white males
iii. 7 of 134 nonwhites.

Models for white females and nonwhites were estimated. It is difficult to determine if any individual variables significantly affect receipt. The variances of parameter estimates are fairly large for most cases, especially for nonwhites. The numbers of nonreceipt are really too small to base any conclusions on them, but there are indications that the models are somewhat useful.

Seven white females of the 10 nonreceipt cases have probability of nonreceipt ranging from .3433 to .8927; .0866 is the smallest. Only 12 of the 181 receipt cases have a probability as large as .1. Five of these have probability greater than .3433 with .7060 the largest. An additional 40 cases have probability between .01 and .1.

For the seven nonwhite nonreceipt cases we estimated P(no receipt) as .0523, .1048, .2607, .8811, .9965, .9988, 1.0.

Of the 127 cases with receipt, only 11 have P(no receipt) ≥ .1 and 44 have probability essentially 0.

These results suggest that there are sets of variables highly correlated with no receipt of wages. Further examination with more data should be done.

**Medicaid Receipt**

Only the nonwhites had enough cases of Medicaid receipt to attempt modeling. If Medicaid receipt was reported in a wave for a person, it was reported in all months of the wave. No one reported receiving Medicaid after not receiving Medicaid in a previous wave. Thus, we were essentially modeling the probability of discontinuing Medicaid receipt for the first month in a wave. Of the eight cases that remained on Medicaid in wave 2, seven have P(Medicaid) = 1.0 and the other P(Medicaid) = .3333. Of the 6 cases that went off Medicaid, two have P(Medicaid) = .3333 and the others, less than .0002. All those not on Medicaid in wave 3 have very small P(Medicaid) in wave 2.

This indicates some success in modeling discontinuance of Medicaid, but more data is required for further investigation.

**Earnings Amounts**

There are some problems that become apparent from examination of the data.

1. Some people report amounts that fluctuate with the number of pay periods or weeks in a month; others don't. (See figures C.1 to C.4 in appendix C.)
2. Do "weeks with pay" correspond directly to "monthly amounts", or can "amounts" be from the previous month's work while "weeks" is for the current month?
3. There are lots of fluctuations in earnings for some people but not for others. We can't expect to get good models by grouping them together. We suggest breaking down records into four types that can be easily identified.

   a. constant earnings
   b. deterministic fluctuations (e.g., due to number of weeks)
   c. random fluctuations
   d. severe fluctuations

Types (a) and (b) are easily imputed. Type (c) can be modeled; (d) can be modeled but some imputes will have large errors. These cases can be modeled together with (c) after editing extreme values.

When using the residual sum of squares to measure model goodness of fit, a few very large residuals can distort this measure. For our longitudinal data large residuals will occur when a person has earnings for a single month that are much higher or lower than in other months. In fact, for month five one residual contributes a very large percentage of the total deviance for all cases. This problem can be tackled by the use of data editing. In appendix A models are included for two types of editing for month 4 earnings: (1) not using Ø earnings when modeling; (2) editing all months according to month-to-month ratios. It is apparent that these procedures improve the overall fit, especially (2).

**Weeks with Pay**

Weeks pay were scaled by dividing by the maximum number of work weeks in the month before modeling. Imputes would be made by determining the appropriate fraction from the model, multiplying by the maximum weeks, and rounding to the nearest integer.

The results for both white females and nonwhites followed the same general pattern in going from month 4 to month 6. The fit for month 4 was not significant, but was for months 5 and 6. This can be seen by looking at the F-statistics in appendix A. An examination of residuals from these models gives the same story. In month 4 only one of the records with fewer than the maximum weeks reported was fitted correctly, while about 50 percent were fitted correctly for 3 of the 4 cases in months 5 and 6. The reason for this fit pattern is probably the increase in information available for use as successive months are modeled. A reason that it is difficult in general to model wpay is that there are not many cases of fewer than maximum weeks reported (less than 10 percent for white females). Separately estimating models for people whose wpay are "frequently" less than the maximum may improve this fit.

**Missing Wage and Salary Records**

We wanted to see if there was any information that would indicate when a person would not respond in wave 3. That is, does one's response to questions in wave 2 tell us anything about the propensity to respond in wave 3? New estimation data sets for white

males and females were created by selecting subsets directly from records of type (i). The fits from this modeling were very poor, especially for those missing in wave 3.

## IMPUTATION RESULTS

The imputation of variables onto a data file is performed by a FORTRAN program that uses the model parameters estimated by GLIM. Each month that is imputed requires a different modification of this program because different months of the independent variables are used. A version for imputing month 4 was prepared and used to impute rcpt, wpay, and earn for white females. This imputation was done for all the appropriate records with complete wave 1 and wave 2 responses. The distributions of imputed and observed values are compared below.

| | rcpt | |
|---|---|---|
| | yes | no |
| observed | 549 | 36 |
| imputed | 580 | 5 |

| | wpay | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 |
| observed | 2 | 8 | 15 | 10 | 36 | 514 |
| imputed | 0 | 0 | 0 | 0 | 3 | 582 |

Earnings were arbitrarily placed into categories for the purpose of this comparison.

| | earnings | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| upper bound | 200 | 400 | 600 | 800 | 1000 | 1200 | 1500 | 2000 | 2500 | 3000 | 4000 | + |
| observed | 107 | 62 | 99 | 102 | 85 | 49 | 30 | 26 | 14 | 4 | 5 | 2 |
| imputed | 79 | 75 | 109 | 108 | 85 | 48 | 30 | 31 | 12 | 1 | 4 | 3 |

The results for rcpt and wpay are not very good. They follow the patterns expected from the model fits as discussed previously. The agreement for earnings is very close, especially for amounts above $400. From our examination of the earnings models and residuals, we expect that there are some reported amounts close to zero that will not be imputed accurately by this model. This definitely shows up on the lower tail of the above distributions.

Additional comparisons for uncategorized earnings are shown in appendix C.

## CONCLUSIONS

1. Not enough cases with no receipt of wages, Medicaid coverage, or weeks with pay less than the maximum occurred to be able to model them well.

2. We should try to improve the fit for wpay in the first month of a wave. Part of our difficulty might be that month 4 can have 5 weeks, but months 2,3,5,6,7 and 8 all have 4 weeks. Another type of scaling than the one we used might be needed.

3. Imputes for rcpt are based on Prob(rcpt). Most of the nonreceipt cases have Prob(rcpt) ≤ .6567, and a small percentage of the receipt cases have probabilities that are small. The distribution of imputed rcpt would better match that of observed rcpt if we adjusted the imputation probabilities to make use of this information. One reason for this result is the very small number of nonreceipt cases.

4. Before modeling earn, the records should be separated into groups according to variability of amount reported. For the most variable groups, data editing may also be needed to improve the model fit.

5. Our attempt to model probability of nonresponse in wave 3 failed completely. If this continues to be true with other data sets, it would tell us that there are no identifiable differences between respondents and nonrespondents for this record type. This would support the application of models fit to respondents to imputation of nonrespondents.

## RECOMMENDATIONS FOR FURTHER STUDY

In the current study we have accumulated knowledge about the longitudinal behavior of the variables we attempted to model, including the frequency of different responses. Much of this came about from examining the data in order to see if there were reasons for the estimated models to look as they did. Much of this knowledge is summarized in the previous section. Based on what we have learned, we suggest our work continue along the following lines.

1. Use as our data set three consecutive waves from the Survey of Income and Program Participation.

2. Construct our imputation file more carefully so that it has more records with infrequently occurring responses. (See (1) under Conclusions.)

3. Look into ways for improving the estimated models. For example, including more response variables, different functions of previously used response variables, and interactions.

4. Determine ways of classifying longitudinal patterns of observed values for earn and wpay in order to fit more accurate models.

5. Investigate the feasibility of using prob(rcpt = yes) differently for the imputation of rcpt.

6. Look further into estimating the probability of WS nonresponse. This can give more information about the nonresponse mechanism or lack thereof.

7. Fit models for all months and investigate the longitudinal consistency of the imputations.

## APPENDIX A

The models fit to the data are summarized here. Each model is fit for a particular dependent variable, month, and demographic group. The exception is the last table for missing record type in wave 3.

Each table has four columns containing information about the model being fit. Under variables are listed the explanatory variables in the model. For model 1, this is a list of the variables. For other models, a line beginning with a "+" gives variables added to the preceding model and a line beginning with a "-" gives variables removed from the preceding model. Occasionally there will be a listing of the form "(5) + ___, - ___;" (5) is the model which is being altered at this step, not the preceding model.

Column 2 gives the scaled deviance for each model. If $\ell_f$ is the likelihood of the full model (using all the information in the observations) and $\ell_c$ is the likelihood of the current model, then scaled deviance is defined by

$$S(c, f) = -2 \log (\ell_c / \ell_f) .$$

For the linear regression models fitted, this is the same as the residual sum of squares.

Column 3 gives the degrees of freedom (number of observations minus number of parameters estimated) for each model. For wpay, column 4 has F-tests for the significance of the regression. For other models, this column has comments concerning the correlation matrix of the estimated parameters.

In order to determine whether adding terms to a model improves or deleting terms from a model degrades the fit, we can use an asymptotic test similar to those of analysis of variance. Let model 2 with $r_2$ degrees of freedom be nested within model 1 with $r_1$ degrees of freedom. If the full model f has n degrees of freedom, then

$$S(1,f) \sim X^2_{n - r_1} \text{ and } S(2,f) \sim X^2_{n - r_2}$$

where the distribution is exact for normal error models and approximate for others. For comparing models 1 and 2, we can then look at

$$S(2,f) - S(1,f) = S(2,1) \sim X^2_{r_1 - r_2} .$$

RCPT - white females - month 4

| variables | deviance | df | comments |
|---|---|---|---|
| 1. rm3, rm2, rm1, rp3, mm1, m0, wpm1, em1n, mp1, wpp3, ep3 | 41.55 | 180 | lots of aliasing<br>2 high correlations |
| 2. +age, ed, mars, rel | 33.93 | 172 | |
| 3. +smsa<br>-rp3, mm1, mp1, wpp3 | 32.32 | 172 | no high correlations<br>used for imputation |

MEDICAID - nonwhites - month 4

| | variables | deviance | df | comments |
|---|---|---|---|---|
| 1. | rm1, r0, rp3, mm3, mm2, mm1, mp3, wpm1, wp0, mwk3, em1, e0, me3, em1r, e0r | 3.567 | 120 | lots of aliasing high correlations |
| 2. | -gm, mm3, mm2, mwk3, me3, em1r, e0r | 3.567 | 122 | one high correlation |
| 3. | -wpm1 | 3.567 | 123 | |
| 4. | -wp0, -em1 | 3.824 | 125 | |
| 5. | -e0 | 3.824 | 126 | used for imputation |

## EARNINGS - white females

### all cases month 4

| | | | | |
|---|---|---|---|---|
| 1. | em3, em2, em1, ep3, me3 | 1821 + 04 | 185 | |
| 2. | +age, ed, mars, rel, smsa | 1798 + 04 | 175 | |

### 0 earnings omitted month 4

| | | | | |
|---|---|---|---|---|
| 1. | em3, em3a, e3, em2, em2a, e2, em1, em1a, e1, ep3, ep3a, e3p, me3 | 1268 + 04 | 168 | |
| 2. | -em3a, em2a, em1a, ep3a | 1372 + 04 | 172 | used for imputation |
| 3. | (1)-e2, em1 | 1311 + 04 | 170 | |
| 4. | -em3a, em2a, e3p | 1316 + 04 | 172 | |
| 5. | -ep3a | 1342 + 04 | 173 | |
| 6. | +ep3a, age, ed, mars, rel, smsa | 1285 + 04 | 162 | mse increased over (4) |

### earnings edited month 4

| | | | | |
|---|---|---|---|---|
| 1. | em3, em2, em1, ep3, me3 | 622 + 04 | 170 | |
| 2. | log(earn) dependent | 1435 + 05 | 170 | much worse |

### all cases month 5

| | | | | |
|---|---|---|---|---|
| 1. | em3, em2, em1, ep3 | 3348 + 04 | 184 | one very large residual |
| 2. | +age, ed, mars, rel, smsa | 3067 + 04 | | |

### all cases month 6

| | | | | |
|---|---|---|---|---|
| 1. | em3, em2, em1, ep3 | 8714 + 03 | 183 | |
| 2. | +age, ed, mars, rel, smsa | 8517 + 03 | 173 | |

WPAY - white females - month 4

| | variables | deviance | df | F-test |
|---|---|---|---|---|
| 1. | wpm3, wpm2, wpm1, wpp3, ep3, em1, mep3, age, ed, mar, eam1, rel, cnt, smsa, region, em1, ern, em2, em3, mwp3 | 1.414 | 165 | $\frac{1509/25}{1.414/165} = 1.17$ <br> $F_{25,165} = 1.90$ <br> Cannot reject hypothesis that regression coefficients are 0. |
| 2. | -mep3 | 1.414 | 165 | |
| 3. | -mwp3 | 1.414 | 166 | |
| 4. | -em2, -em3 | 1.415 | 168 | |
| 5. | +em2, +em3, -wpp3, -mep3 +mep3, +mwp3 | 1.414 | 166 | |
| 6. | -em2, -em3, -ent0, -mep3, -mwp3 | 1.419 | 169 | model used for imputation |

zero cases out of 11 where # of wpay < max were estimated correctly by model 6.

WPAY - white females - month 5

| | variables | deviance | df | F-test |
|---|---|---|---|---|
| 1. | wpm3, wpm2, wpm1, wpp2, ep2, em1, mep2, age, ed, mar, rel, cnt, smsa, region, eap1, wpp3, mwp2, mwp3, ep3, mep3, em2, em3 | 2.461 | 164 | $\frac{1.693/26}{2.461/164} = 4.66$ <br> $F_{28,164} = 1.90$ <br> Reject hypothesis that regression coefficients are 0. |
| 2. | -mep2 | 2.461 | 164 | |
| 3. | -wpp2 | 2.528 | 165 | |
| 4. | -ep3 | 2.530 | 166 | |

Five cases out of 12 where # of wpay < max were estimated correctly by model 4.

WPAY - white females - month 6

| | variables | deviance | df | comments |
|---|---|---|---|---|
| 1. | wpm3, wpm2, wpm1, wpp1, ep1, em1, mep1, age, ed, mar, rel, cnt, smsa, region, enp, wpp2, mwp2, ep2, mep2, wpp3, ep3, mwp1, mwp3, mep3, em2, em3 | 2.102 | 162 | $\frac{5.310/28}{2.102/162} = 14.64$ <br> $F_{28,162} = 1.85$ <br> Reject hypothesis that regression coefficients are 0. |
| 2. | -eap1, -wpp3 | 2.104 | 164 | |
| 3. | -mep1 | 2.104 | 164 | |
| 4. | -wpp2 | 2.104 | 165 | |
| 5. | -wpp1 | 2.105 | 166 | |

Five out of 12 cases where # of wpay < max were estimated correctly by model 5.

WPAY - nonwhites - month 4

| | variables | deviance | df | comments |
|---|---|---|---|---|
| 1. | wpm3, wpm2, wpp3, ep3, em1, mep3, ale, ed, mar, eam1, rel0, cnt, smsa, region, em1, ern, em2, em3, mwp3 | 1.895 | 108 | $\frac{1.737/25}{1.895/108} = 99$<br>$F_{25,108} = 1.97$<br>Reject hypothesis that regression coefficients are 0. |
| 2. | -em2, -em3 | 1.904 | 110 | |
| 3. | -ep3, -em1 | 1.939 | 112 | |

One case out of 20 where # of wpay < max was estimated correctly by model 3.

WPAY - nonwhites - month 5

| | variables | deviance | df | comments |
|---|---|---|---|---|
| 1. | wpm3, wpm2, wpm1, wpp2, ep2, em1, mep2, age, ed, mar, rel, cnt, smsa, region, eap1, wpp3, mwp2, wmp3, ep3, mep3, em2, em3 | 2.239 | 107 | $\frac{3.377/26}{2.239/107} = 6.21$ <br> $F_{26,107} = 1.98$ <br> Reject hypothesis that regression coefficients are 0. |
| 2. | -ep3 | 2.327 | 108 | |

Eight cases out of 15 cases where # wpay < max were estimated correctly by model 2.

WPAY - nonwhites - month 6

| | variables | deviance | df | comments |
|---|---|---|---|---|
| 1. | wpm3, wpm2, wpm1, wpp1, ep1, em1, mep1, age, ed, mar, rel, cnt, smsa, region, eap1, wpp2, mwp2, ep2, mep2, wpp3, ep3, mwp1, mwp3, wep3, em2, em3 | 5.486 | 105 | $\frac{5.945/28}{5.486/105} = 4.02$ <br> $F_{28,105} = 1.90$ <br> Reject hypothesis that regression coefficients are 0. |
| 2. | -mep1 | 5.486 | 105 | |
| 3. | -mwp2 | 5.486 | 105 | |
| 4. | -ep2 | 5.540 | 160 | |

Four cases out of 20 where # of wpay < max were estimated correctly by model 4.

# MISSING WAGE & SALARY RECORD IN WAVE 3

## white males

| | variables | deviance | df | |
|---|---|---|---|---|
| 1. | rm3, rm2, rm1, em2, em1, wpm2, wpm1 | 112.1 | 157 | |
| 2. | -wpm2, -wpm1 | 112.9 | 159 | |
| 3. | -em1 | 121.4 | 160 | much worse |
| 4. | +em1, age, ed, mars, rel, hhnum, smsa | 101.3 | 148 | |

## white females

| | variables | deviance | df |
|---|---|---|---|
| 1. | rm3, rm2, rm1, em2, em1, wpm2, wpm1 | 106.3 | 129 |
| 2. | -wpm2 | 108.0 | 130 |
| 3. | +age, ed, mars, rel, hhnum, smsa | 96.37 | 118 |

# APPENDIX B

Variable transformations used in fitting models

Definitions of variables used in models

The month for which a model is being estimated has the designation 0 [illegible] month previous is m1, etc.; one month in future is p1, etc.

r = receipt of wages
wp = weeks with pay
e = earnings amount
m = medicaid coverage
wm = maximum weeks in a month

Variables that are computed as functions of these variables will be defined. rm3, rm2, rm1, r0, rp1, rp2, rp3, mm3, mm2, mm1, m0, mp1, mp2, mp3 always have the obvious meaning described above.

<u>transformed variables used in modeling receipt of wages</u>

$$wpm1 = wp1/wmm1$$

$$wpp3 = \begin{cases} wpp3/wmp3 & \text{if wpp3 observed} \\ 0 & \text{otherwise} \end{cases}$$

$$em1n = \min\ (ep3+.0005)/(em2+.0005),\ 5$$

$$ep3 = \begin{cases} \min\ (ep3+.0005)/(em1+.0005),\ 5 & \text{if ep3 observed} \\ 0 & \text{otherwise} \end{cases}$$

<u>transformed variables used in modeling Medicaid coverage</u>

$$em1 = \min\ (e0+.0005)/(em1+.0005),\ 5$$

$$e0 = \begin{cases} \min\ (ep3+.0005)/(e0+.0005),\ 5 & \text{if ep3 observed} \\ 0 & \text{otherwise} \end{cases}$$

$$e0r = \min\ (e0)(wm0)/wmp3,\ 5$$

$$em1r = \min\ (em1) * (wmm1)/wm1,\ 5$$

$$me3 = \begin{cases} 1 & \text{if ep3 missing} \\ 0 & \text{otherwise} \end{cases}$$

transformed variables used in modeling earnings

$$em3a = \begin{cases} (em3)\ \dfrac{wp0}{wm0} \Big/ \dfrac{wpm3}{wmm3} & \text{if } wpm3 \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$e3 = \begin{cases} 0 & \text{if } wpm3 \neq 0 \\ (em3 + em2 + em1)/3 & \text{otherwise} \end{cases}$$

$$em2a = \begin{cases} (em2)\ \dfrac{wp0}{wm0} \Big/ \dfrac{wpm2}{wmm2} & \text{if } wpm2 \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$e2 = \begin{cases} 0 & \text{if } wpm2 \neq 0 \\ (em3 + em2 + em1)/3 & \text{otherwise} \end{cases}$$

$$em1a = \begin{cases} (em1)\ \dfrac{wp0}{wm0} \Big/ \dfrac{wpm1}{wmm1} & \text{if } wpm1 \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$e1 = \begin{cases} 0 & \text{if } wpm1 \neq 0 \\ (em3 + em2 + em1)/3 & \text{otherwise} \end{cases}$$

$$ep3a = \begin{cases} (ep3)\ \dfrac{wp0}{wm0} \Big/ \dfrac{wpp3}{wmp3} & \text{if } wpp3 \neq 0 \text{ and not missing} \\ 0 & \text{otherwise} \end{cases}$$

$$e3p = \begin{cases} 0 & \text{if } wpp3 \neq 0 \\ (em3 + em2 + em1)/3 & \text{otherwise} \end{cases}$$

$$me3 = \begin{cases} 0 & \text{if } ep3 \text{ not missing} \\ (em3 + em2 + em1)/3 & ep3 \text{ missing} \end{cases}$$

transformed variables used in modeling weeks with pay

month 4

em3 = min em3/(wpm3+.005), 5000
em2 = min em2/(wpm2+.005), 5000
em1 = min em1/(wpm1+.005), 5000
eam1 = min em1/(em2+.005), 50

$$ep3 = \begin{cases} \min\ ep3/(wpp3+.005),\ 5000 & \text{if ep3 observed} \\ 0 & \text{missing} \end{cases}$$

$$mep3 = \begin{cases} 0 & \text{if ep3 observed} \\ 1 & \text{missing} \end{cases}$$

mwp3 = mep3

$$ern = \begin{cases} \min\ ep3/(em1+.005),\ 50 & \text{if ep3 observed} \\ 0 & \text{missing} \end{cases}$$

month 5

$$ep2 = \begin{cases} \min\ ep2/(wpp2+.005),\ 5000 & \text{if ep2 observed} \\ 0 & \text{missing} \end{cases}$$

$$eap2 = \begin{cases} \min\ ep2/(em1+.005),\ 50 & \text{if ep2 observed} \\ 0 & \text{missing} \end{cases}$$

$$mep2 = \begin{cases} 0 & \text{if ep2 observed} \\ 1 & \text{missing} \end{cases}$$

mwp2 = mep2

month 6

$$ep1 = \begin{cases} \min\ ep1/(wpp1+.005),\ 5000 & \text{if ep1 observed} \\ 0 & \text{missing} \end{cases}$$

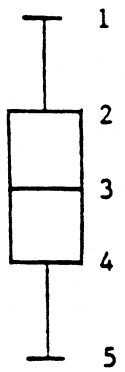
$$eap1 = \begin{cases} \min\ ep1/(em1+.005),\ 50 & \text{if ep1 observed} \\ 0 & \text{missing} \end{cases}$$

$$mep1 = \begin{cases} 0 & \text{if ep1 observed} \\ 1 & \text{missing} \end{cases}$$

mwp1 = mep1

## APPENDIX C
Earnings Amounts

Interpreting a box plot. There are five pieces of information given by each of the box plots.




1. Maximum value
2. Upper quartile
3. Median
4. Lower quartile
5. Minimum value

Not all box plots have these five components visible if two or more of them have the same value. Some plots have only a single horizontal line that indicates constant observed values.

Figure C.1 Each box summarizes 9 months of earnings for a white female. They show differences in variability of earnings.

Figures C.2. - C.4. Each box summarizes 3 months (one wave) of earnings. Plots with no median and one large value have two amounts at the lower edge of the box and one at the maximum.

Figure C.5. Scattergram of nonzero reported amounts vs. residuals (= observed-imputed).

Figure C.6. Histogram of imputed amounts for zero reported amounts.

Figure C.7. Histogram of percentage error of impute for nonzero reported amounts. Note that some values have been trimmed off each end.

Figure C.8. Scattergram for same data as C.7. This shows clearly that most large negative percentages are due to reported values of less than $500.

**FIGURE C.1.a**

Nine months of earnings for 30 white females

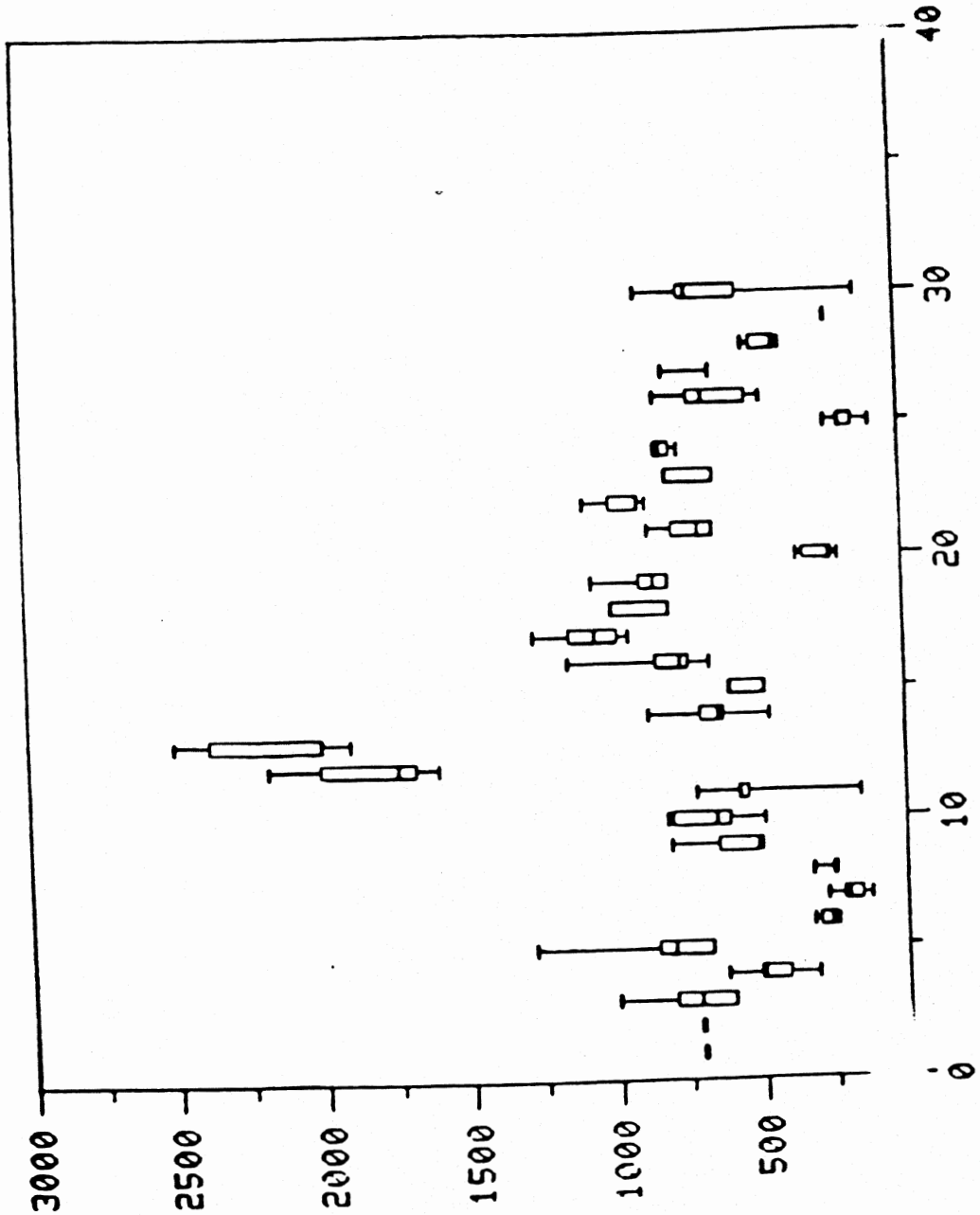


**Figure C.1.b**

Nine months of earnings for 30 white females

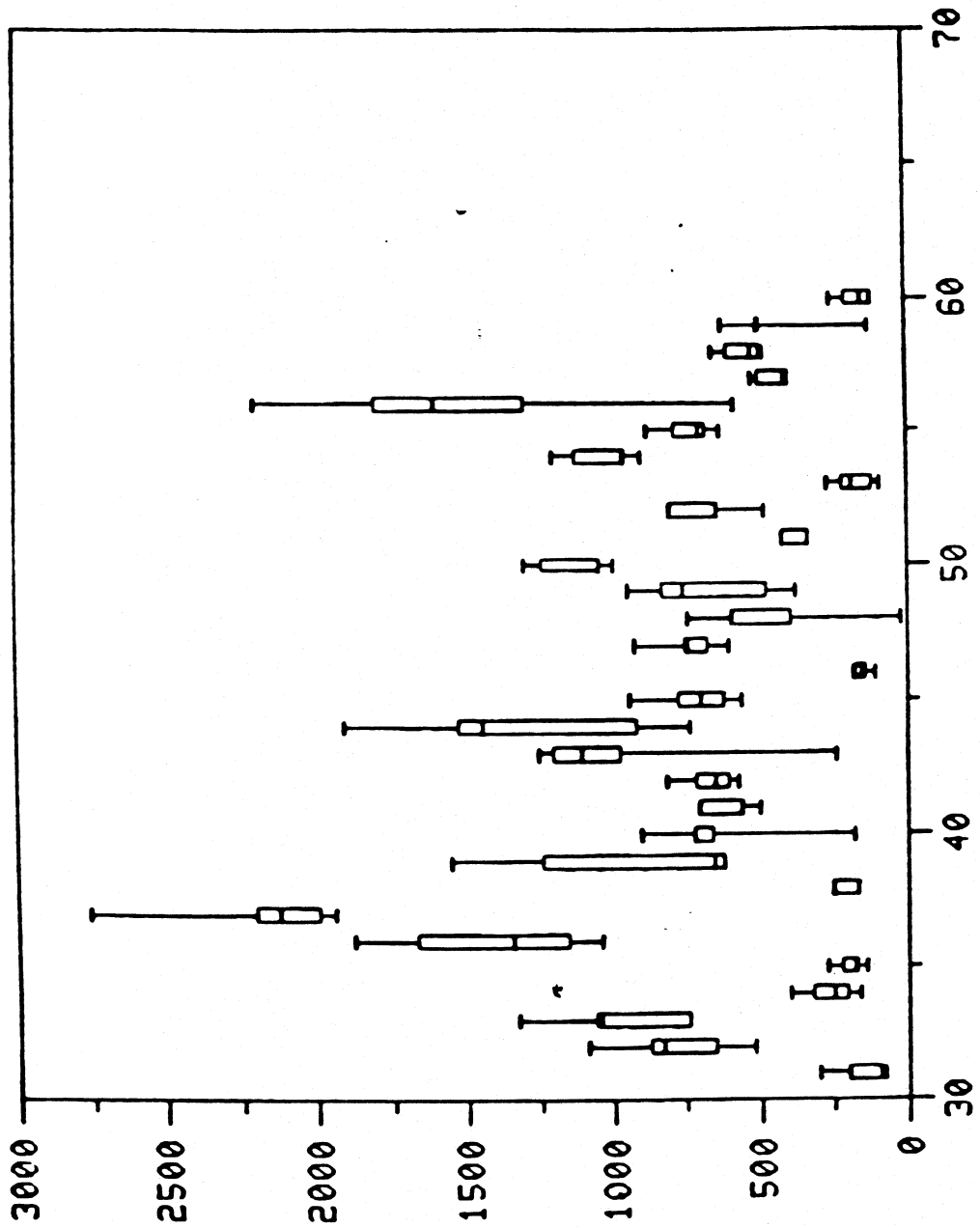


**Figure C.1.c**

Nine months of earnings for 30 white females

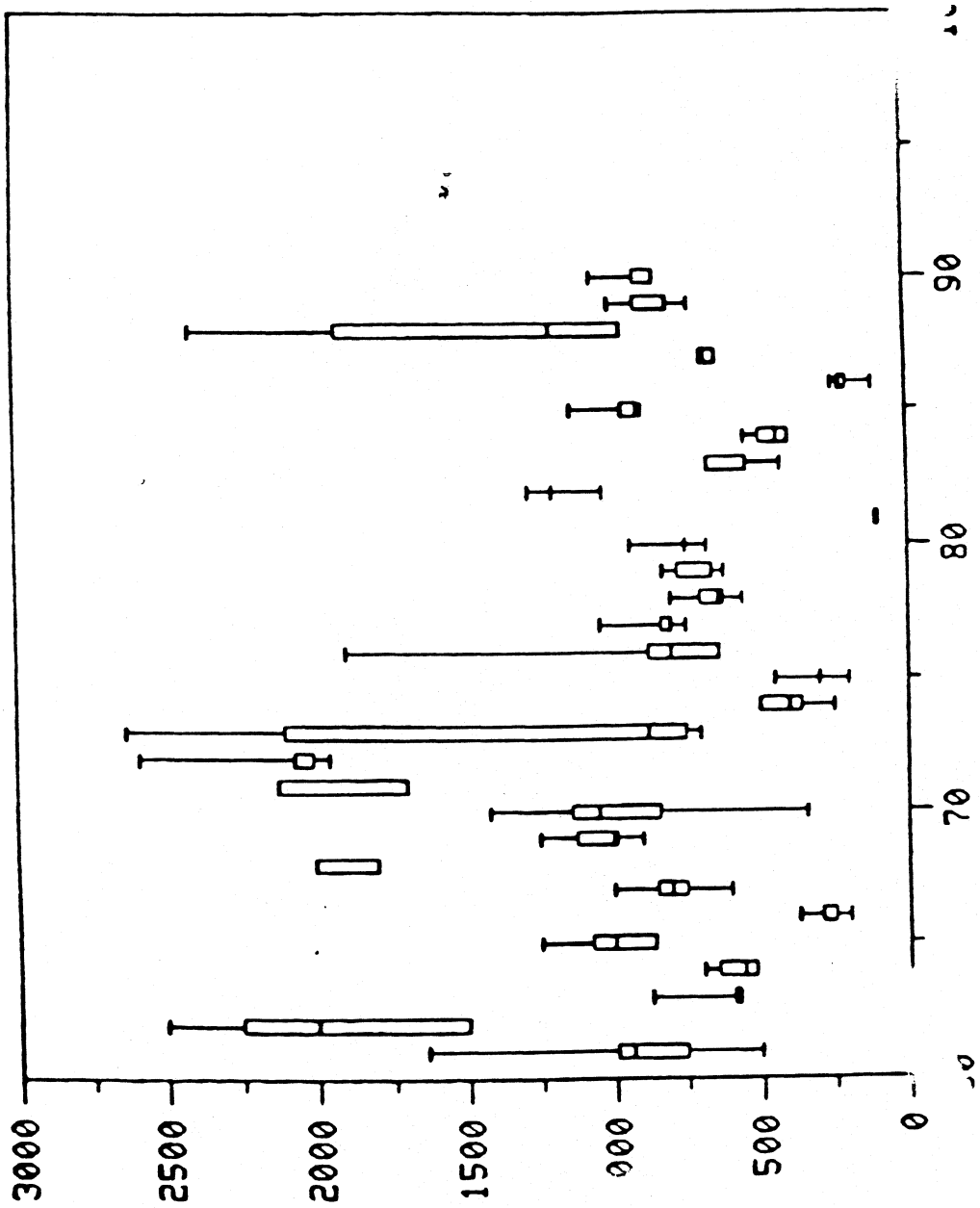


**Figure C.2**

Wave 1 earnings for 30 white females

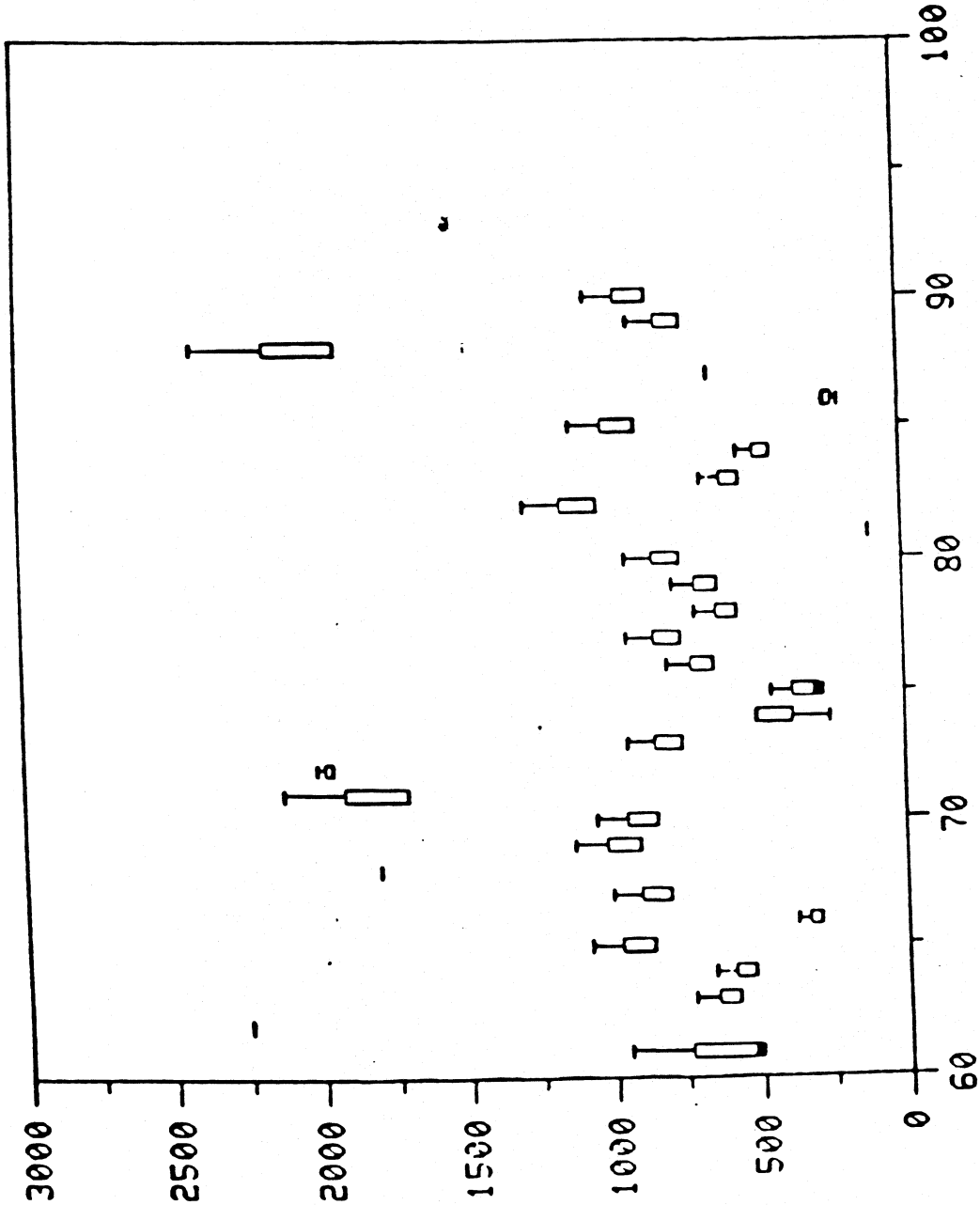


**Figure C.3**

Wave 2 earnings for 30 white females

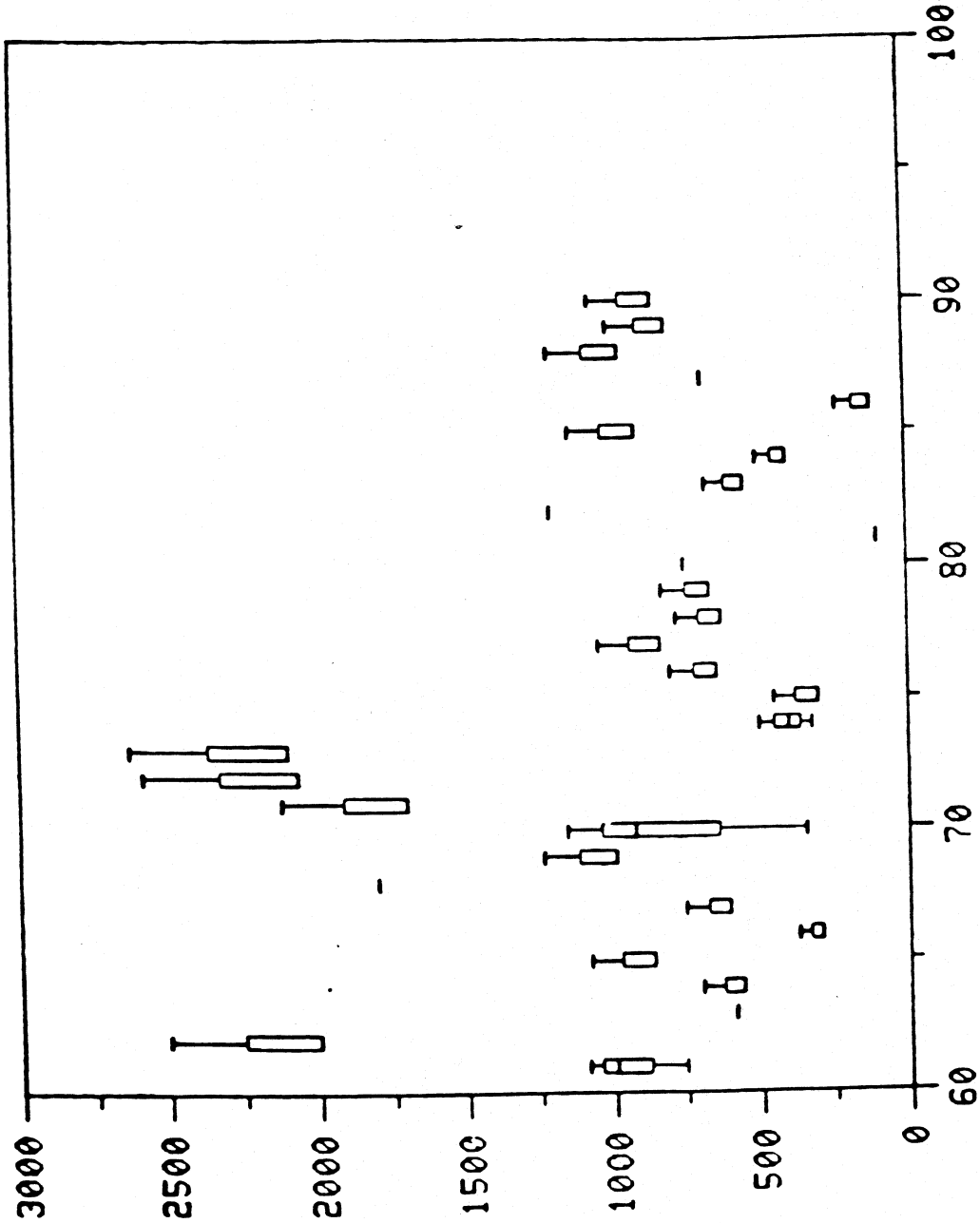


**Figure C.4**

Wave 3 earnings for 30 white females

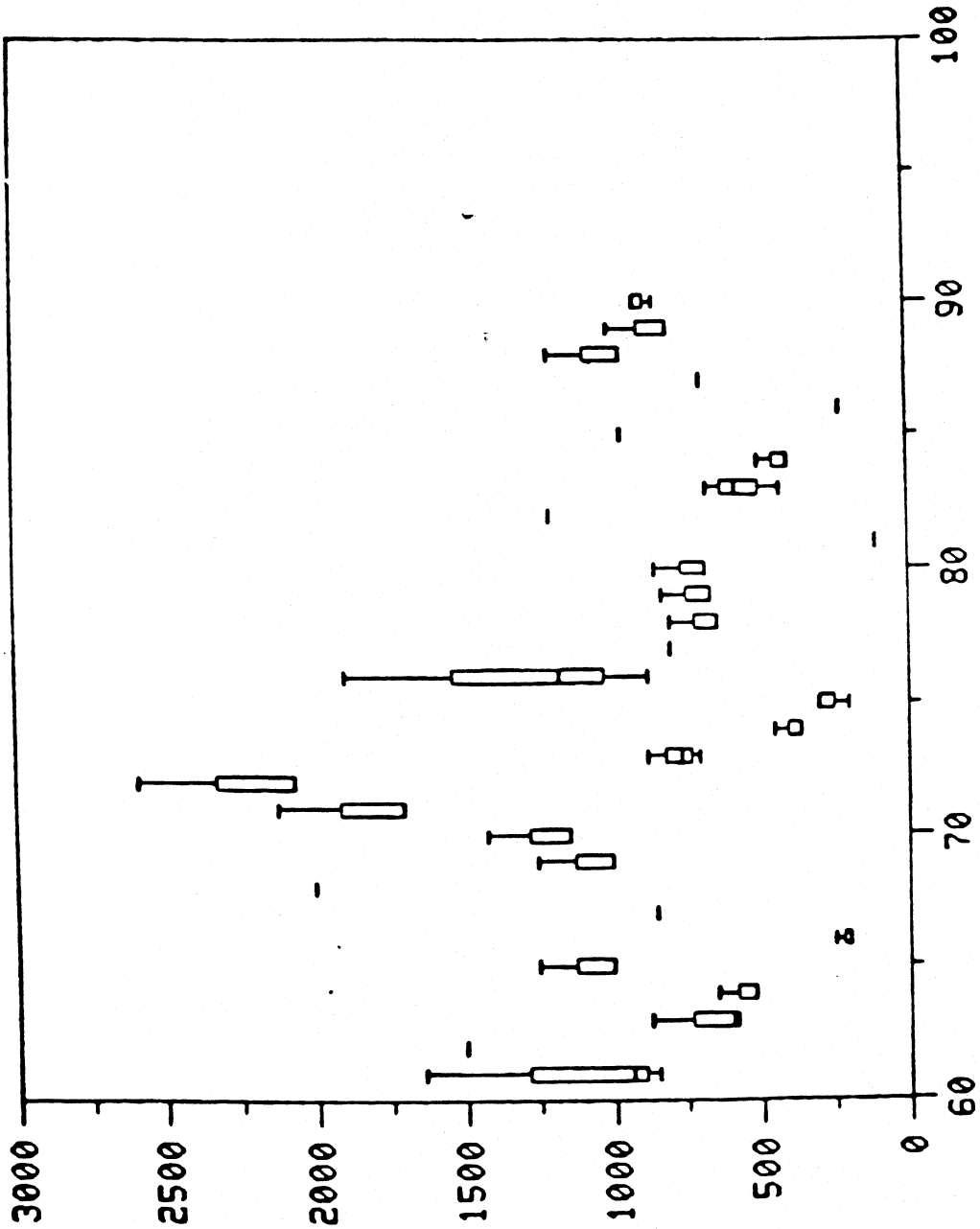


**Figure C.5**

Nonzero reported amounts vs. residuals

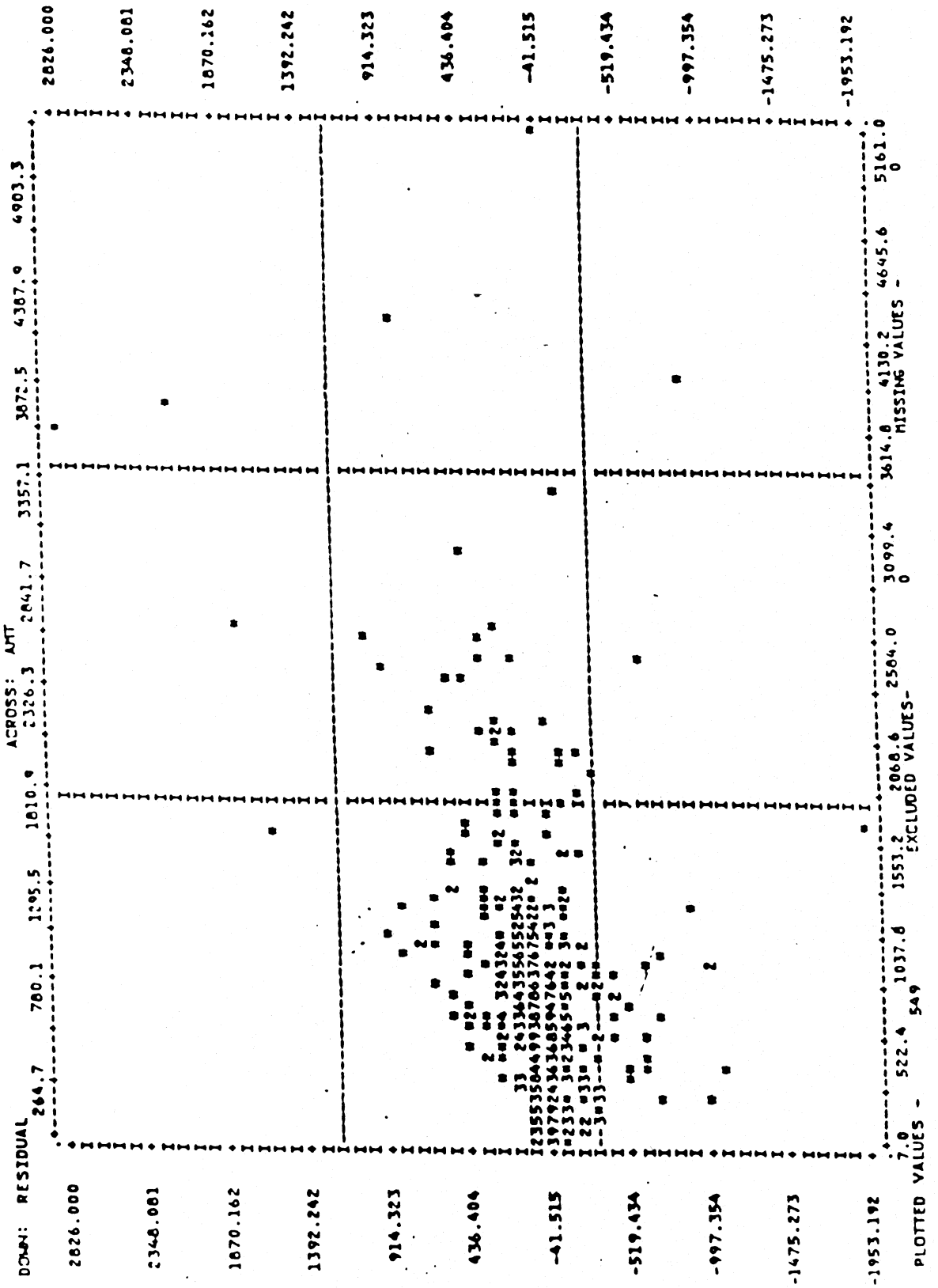


**Figure C.6**

Histogram of imputed amounts for amounts reported as zero

```
COUNT   MIDPOINT    ONE SYMBOL EQUALS APPROXIMATELY  1.50 OCCURRENCES

  4     -1.475  ***
  1     -1.425  *
  2     -1.375  *
  1     -1.325  *
  0     -1.275
  2     -1.225  *
  4     -1.175  ***
  1     -1.125  *
  1     -1.075  *
  2     -1.025  *
  3      -.975  **
  1      -.925  *
  1      -.875  *
  3      -.825  **
  2      -.775  *
  4      -.725  ***
  4      -.675  ***
  1      -.625  *
  1      -.575  *
  2      -.525  *
  5      -.475  ***
  7      -.425  *****
  5      -.375  ***
  4      -.325  ***
  4      -.275  ***
 22      -.225  ***************
 13      -.175  *********
 27      -.125  ******************
 54      -.075  ************************************
 27      -.025  ******************
 37       .025  *************************
 68       .075  *********************************************
 55       .125  ************************************
 46       .175  ******************************
 25       .225  ****************
 21       .275  **************
 12       .325  ********
 10       .375  *******
 11       .425  *******
  8       .475  *****
  5       .525  ***
  4       .575  ***
  2       .625  *
  8       .675  *****
  2       .725  *
            I....+....I....+....I....+....I....+....I....+....I
            0        15        30        45        60        75
                   HISTOGRAM FREQUENCY

VALID CASES     549     MISSING CASES     0
```

**Figure C.7**

Histogram of percentage error of impute for Nonzero reported amounts

```
COUNT   MIDPOINT    ONE SYMBOL EQUALS APPROXIMATELY   .20 OCCURRENCES

    9     50.00   ##########################################
    3    150.00   ###############
    4    250.00   ####################
    1    350.00   #####
    7    450.00   ##################################
    1    550.00   #####
    3    650.00   ###############
    0    750.00
    0    850.00
    1    950.00   #####
    0   1050.00
    1   1150.00   #####
    0   1250.00
    0   1350.00
    0   1450.00
    0   1550.00
    0   1650.00
    3   1750.00   ###############
    0   1850.00
    0   1950.00
              I....+....I....+....I....+....I....+....I....+....I
              0         2         4         6         8        10
                      HISTOGRAM FREQUENCY

VALID CASES      36      MISSING CASES      0
```