BEBR

FACULTY WORKING PAPER NO. 1335

Exploring Income Nonresponse: A Logit Model Analysis

Kjell Grønhaug
Mary C. Gilly
Ben M. Enis

College of Commerce and Business Administration
Bureau of Economic and Business Research
University of Illinois, Urbana-Champaign

BEBR

FACULTY WORKING PAPER NO. 1335

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

February 1987

Exploring Income Nonresponse:
A Logit Model Analysis.

Kjell Grønhaug, Visiting Professor
Department of Business Administration

Mary C. Gilly
Univeristy of California

Ben M. Enis
University of Southern California

EXPLORING INCOME NONRESPONSE: A LOGIT MODEL ANALYSIS

Abstract

This article examines income nonresponse in a large-scaled survey study among approximately 2,500 households. Income nonresponse was found to be a predominant problem, more so for older than younger people, more frequent among respondents low than high in education, and for whites more than non-white. The conclusion is that income item nonresponses should be estimated rather than ignored or eliminated from the data base.

The purpose of this article is to explore non-reporting of household income and income non-reporting correlates. Incomplete data are of major concern in survey research. Madow et al. (1983), in their comprehensive discussion of this problem, contend:

> The main problem caused by incomplete data in sample surveys is that estimates of population characteristics and relations must be assumed to be biased unless very convincing evidence to the contrary is provided......As a result, means, variances, covariances, and other statistical functions will be biased and have distributions affected by incompleteness (p. 15-16).

Data incompleteness can be classified as follows: undercoverage, unit nonresponse, and item nonresponse. Undercoverage occurs when units that should be in the frames from which a sample is selected are not in those frames, or units in the sample are mistakenly classified as ineligible or are omitted from the sample or from the units interviewed. Unit nonresponse is present when units in the selected sample and eligible for the survey do not provide the requested information, or the provided information is unusable. Item nonresponse is present when eligible units in the selected sample provide some, but not all, of the requested information, or information provided for some items is unusable. A variety of problems related to unit nonresponse, such as magnitude of unit nonresponse, and techniques to increase response rates have extensively been dealt with in past research (for overviews see Kanuk & Berenson 1975; Houston & Ford 1976; Yu & Cooper 1983). Biases caused by unit nonresponse and methods for adjustments of the information gathered have also been examined (cf. Kanuk & Berenson 1975; Houston & Ford 1976; Madow et al. 1983). Less research has been directed towards undercoverage (cf. Madow et al. 1983) and item nonresponse (cf. Ferber 1966; Houston & Ford 1976).

ITEM NONRESPONSE

In 1966 Ferber (1966) wrote, "Item nonresponse, ..., has virtually received no attention" (p. 399). Ten years later Houston & Ford (1976) stated: "Reviews of methodological research on mail surveys have yet to touch on item omission as a dependent variable, ..." (p. 400). And even ten additional years later, item nonresponse has only received modest attention among researchers (see Peterson & Kerin 1981, Madow et al. 1983).

Ferber (1966) has offered the most penetrating look into the item nonresponse problem based on 14,600 usable questionnaires obtained from a mailing to 40,000 members of Consumers Union, i.e., a response rate of 36.5 percent. A disturbing finding reported in Ferber's study is that only 37.5 percent of the returned questionnaires were filled out completely, indicating the magnitude of the item nonresponse problem. In Ferber's study item nonresponse was found to be much higher for females compared to men, higher for older compared to younger people, and higher for people low compared with those high in education. Questions requiring more thought on the part of the respondent were more subject to nonresponse than those of a factual nature. In addition to Ferber's study some studies examining correlates between item nonresopnse and demographics (see Peterson & Kerin 1981) and how various devices such as advance contact, removal of anonymity and cash gift influence item nonresponse have been reported (for overview see Houston & Ford 1976; Hornik 1982).

Income is one of the indicators most commonly used in marketing and social research to characterize individuals and households. The literature reveals that income is believed to reflect economic resources, buying power, as well as status and social class (cf. Sheth 1977; Smelser 1976). It has been recognized, however, that questions regarding income can be perceived as "threatening" or difficult to answer, resulting in under- and overreporting of income (cf. Sudman & Bradburn 1982) and income nonresponse as well (Madow et al. 1983).

Few studies have investigated income nonresponse. In March of each year the Bureau of the Census's Current Population Survey questionnaire includes a supplement in which detailed data on income are requested for each member aged 14 and over of the sample households. Madow et al. (1983) in quoting the findings from the 1978 survey, show that the nonresponse rate for one or more of 11 questions of income varies from 14 percent for people 21-24 years of age to 27 percent for people in the 55-64 age bracket (p. 24). Craig and McCann (1978) in their analysis of six data bases found income nonresponses related to type of product. Sheth et al. (1980) reported approximately 10 percent income nonresponse in a mail survey examining the impact of asking race information. Peterson et al. (1981) in a telephone survey conducted among approximately 6200 adult consumers found that 12.6 did not report income. Respondents high more than low in education were found to report income; older more than younger and females more often than men refused to answer the income question. The present knowledge about reporting of income is, however, rather modest.

Income is often measured as the "household/family" income and one respondent is usually used as "key" informant to represent the household. An important, but still unanswered question, is whether the respondent's position in the family, e.g., husband or wife may influence income item response.

DATA AND MEASUREMENT

This research is based on a survey study conducted in Houston, Texas, by a market research company. This company conducts an annual survey covering a variety of topics such as buying plans, reading habits, use of financial services, etc., and the results are sold to several clients.

Data

The survey is based on sequential sampling, aiming at a total sample size of 4,000 respondents. The unit of observation is the household, and one person is used as "key informant." Here it was decided to explore the reporting of income of "complete" families, i.e., wife and husband plus children if any. Of the 4,000 respondents, 2,505 were classified as "complete" families, e.g., 62.6 percent of the total sample.

The sampling procedure was as follows: Central addresses were randomly drawn within the various section of the city. Around each central address a given number of personal interviews was conducted in a specific order. Thus the sampling procedure is comparable to stratified, single-stage cluster sampling. Sampling of respondents continued until the intended number of interviews had been reached. No report on number of households the market research firm had tried to contact exists, and thus the response rate cannot be estimated. The persons interviewed was the female/male "head" of the household selected on an odd-even rotation basis.

Measurement

The variables to be included in the analysis were operationalized as reported below:

INC: "And what is the approximate arrival income for all members in your family?" (Coded 1 = income not reported, 2 = income reported)

AGE: "And what is your age?" (Coded: 1 = 34 years or less, 2 = 35 years or more).

RACE: "What is your ethnic background?" (Coded 1 = non-white, 2 = white)

EDUC: "What was the last grade of school completed?" (Coded 1 = high school graduate or less, 2 = some college or more)

SEX: (Coded: 1 = female, 2 = male)

FINDINGS

Item Nonresponse

The item nonresponses for the variables included in this study are shown in Table 1:

Insert Table 1 about here

From Table 1 it is seen that the majority of the item nonresponses is related to income. Fourteen percent of the respondents included in this sample failed to report income, which is very close to the income non-response rate reported by Peterson et al. (1980). Table 1 shows a total of (9 + 6 + 6 + 1 =) 22 item nonresponses for variables other than income, i.e., the non-response rates for these variables are negligible. Almost no overlap was observed across variables for these item non-responses. The nonresponses for the other variables than income were removed from the sample due to its modest magnitude and computational reasons reducing the sample size from 2505 to 2484, e.g., a reduction of 21 respondents--less than one percent. Compared to the item non-response rates for the same variables reported by Sheth et al. (1980), the item nonresponse in the present study is somewhat higher for age and lower for education, race and sex. Also in the study reported by Sheth et al., (1980) the highest nonresponse rate was found for income.

Bivariate Results

Table 2 reports income nonresponses across the various predictor variables included in this study.

Insert Table 2 about here

Inspection of Table 2 reveals that income nonresponse is by far more predominant for older than younger people, somewhat more frequent among people low compared to those high in education, somewhere higher for non-white compared to white, and slightly higher among females than males. The finding for age is in accordance with the results reported by Madow et al. (1983), and the findings for education and sex are in the same direction as reported by Ferber (1966).

Logit Analysis

Several of the independent variables are interrelated, so multi-variate analysis was conducted to explore the impact of the various variable on income nonresponse. Logit analysis treating income as the dependent variable was deemed appropriate for further examination. The logit model, which is a special case of the general loglinear model, is based on odds ratios, or as stated by Knoke & Burke (1980): "... odd ratio is the workhorse of log-linear models" (p. 10).

In such models expected cell frequences are represented by η, a baseline from which effects are measured, τ_i^j, the effect for each of the i categories of variable j, as well as various interaction effects of two or more variables.

For dichotomous variables as used here, the tau effect parameters for each variable's categories are reciprocals: $\tau^j = \tau_1^j = 1/\tau_2^j$ (the numerical subsamples 1, 2 refer to the category of the variables to which the tau values applies).

The subroutine LOGLINEAR in the SPSSX program package was used to estimate the model parameters. The complete table is shown in Appendix A. The saturated model, which includes all possible interactions and represents the table completely, was chosen as a starting point. Several models were computed. In applying the criteria goodness of fit and simplicity, the following model was finally chosen:

Insert Table 3 about here

The fit of the chosen model as measured by likelihood chi square, $L^2 = 10.96$, d.f. = 10, p = .361. The large sample size and Type I error being within the recommended range for such models indicate that the true relationships are captured (see Knoke & Burke 1980, pp. 30-33 for further discussion). The number in parentheses are the standard deviations, and estimates equivalent to t-values are obtained by dividing the coefficients estimated by their standard deviations.

The log-odds coefficients estimated by the program are shown in column (1). Column (1) shows that AGE is the variable possessing the highest descriptive and predictive power as was demonstrated in Table 2 (for thorough discussion of interpretation of the effects, see Aldrich & Nelson 1984, pp. 40-44). "Regression-like" coefficients of the estimated log-odds coefficients are obtained by multiplying by 2 (column (2)) due to the corresponding log-linear model. This means that the age effect of age on income will take the value -.431 when AGE = 1, and .4131 when AGE = 2, and so on. By calculating the antilog of the "regression-like" coefficients, the model is translated into odds rather log odds (column (3)). The findings will of course remain the same, but by doing this transformation, the findings will be easier to interpret. A multiplicative model has to be used, when the model is translated into odds, e.g.:

$$(Fijk_1/Fijk_2) = T * T(AGE)_i * T(EDUC)_j * T(RACE)_k * T(AGE,RACE)_{ik} * T(EDUC,RACE)_{jk}$$

where:

F is an expected frequency, and

$T = .1439$

$T(AGE)_i = .6562$ for $i = 1$

$1/.6526$ for $i = 2$

$T(EDUC)_j = 1.314$ for $j = 1$

$1/1.314$ for $j = 2$

$T(RACE)_k$ = .9667 for k = 1
1/.9667 for k = 2

$T(AGE,RACE)_{ik}$ = .8571 for i = k
1/.8571 for i = not k

$T(EDUC,RACE)_{jk}$ = 1.1654 for j = k
1/1.1654 for j = not k

For example, consider a person young of age, low in education and of non-white race. For this individual i = j = k = 1 because of the coding of the variables as discussed above. The expected odds given by the model is .1041, which is close to the observed odds, .1153. The model decomposes these expected odds into components:

$$.1041 = (.1439)(.6562)(1.1314)(.9667)(.8577)(1.1654)$$

where the effects are:

- .1439 is the mean or overall effect.
- .6562 is the age effect, indicating that individuals young of age are much <u>less</u> inclined to income nonresponse than are their old counterparts.
- 1.314 is the effect of education, showing that individuals low in education <u>more</u> frequent than those high in education do not report income.
- .9667 is the effect of race. Whites somewhat <u>more</u> than non-whites fail to report income.
- .8577 is the interaction effect of age and race. This effect-means that persons being young and non-white will be <u>more</u> inclined to report income than indicated by combining the main effect of being young with the main effect of being non-white.
- 1.1654 is the interaction effect of education and race, indicating that persons being low in education and non-white will be <u>less</u> inclined to report income than observed from combining the main effect of being low in education with the main effect of being non-white.

Table 4 reports the observed and estimated number of income nonresponses. By adding the numerical values of all residuals, the total

---
Insert Table 4 about here
---

number amounts to 47.83, e.g., our model has classified 13.9 percent of the cases incorrectly, or the other way around; 84.1 percent of the cases are correctly classified, indicating a rather good fit to the data. The same number of misclassifications, as reported in Table 4, but with opposite signs will appear for those who reported income. Inspection of Table 4 reveals that the highest number of misclassified cases is found for older, white males, low in education (-9.82), counting for 20.5 percent of all misclassified cases in spite of the fact that this group only counts for 6.7 percent of all income nonresponses.

CONCLUSIONS

Consistent with previous findings (cf. Ferber 1966, Madow et al. 1983; Peterson & Kerin 1981), the present study has shown that item nonresponse occurs, and that item nonresponse is not randomly distributed. Indeed, item nonresponse is likely to yield biased results, particularly with respect to income data.

Since income is often used as a classificatory or predictor variable, biases in income data can seriously distort research results. Our conclusions therefore are twofold: researchers must acknowledge the possibility of bias due to income nonresponse, and take explicit steps to correct or at least mitigate the resulting distortions. We first briefly review the nature of possible distortions, and then offer suggestions for correction or mitigation.

When contrasting the reported findings with previous results, it is observed that the higher propensity to income nonresponse among older compared to younger people is in accordance with the findings reported by Madow et al. 1983 and Peterson et al. 1981. The reasons for higher income nonresponse among the older may be several, including a tendency to perceive such questions as more embarrassing and threatening among this group. Whatever the reason(s) might be, the reported finding is important, and precautions should be taken regarding how to cope with the mature higher income nonresponse problem when planning largescaled survey studies.

The somewhat higher income nonresponse among people low compared to those high in education is in accordance with the findings reported by Ferber (1966) and Peterson et al. (1981). In contrast to what was reported by Ferber (1966) and Peterson et al. (1981) <u>no</u> differences between females and males were found regarding item nonresponse. In this study sex was found to neither have any direct nor any interaction effect with other variables on income nonresponse.

This result seems to lend support to the use of either husband or wife as "key" informant in household studies. This conclusion would be welcomed by family researchers. But we advise caution, since the present data base does not permit husband/wife comparison (cf. Granbois and Willitt 1970, Haberman and Elinson 1967). Differences in responses between husbands <u>and their wives</u> should be studied before concluding that either could serve as key informant for a given household.

Researchers should begin to think explicitly about biases introduced by income nonresponse; too often, this issue has simply been ignored. Nor will eliminating all data from income nonrespondents solve the problem. Income item nonresponse is related, in non-random ways, to other data. Results may be even more distorted by simply eliminating all data when income item nonresponse is encountered.

Our recommendation, therefore, is that income item nonresponse be estimated, and that the estimates be incorporated into the data base.

While estimation based on logit or some other procedure is not a perfect solution to missing income data, it is, in our opinion, preferred to ignoring the problem of throwing data away. Further research is clearly warranted here. Until now, item nonresponse has received only modest attention. It is particularly important for marketing and social research that income item nonresponse by studied.

REFERENCES

Aldrich, J. N. and Nelson, F. N. (1984), Linear Probability, Logit and Probit Models, Beverly Hills, Sage Publications.

Craig, C. S. and McCann, J. M. (1978), "Item Nonresponse in Mail Surveys: Extent and Correlates," J. of Marketing Research, Vol. XV (May), 285-87.

Ferber, R. (1966), "Item Nonresponse For a Consumer Survey," Public Opinion Quarterly, Vol. XXX (Fall), No. 3, 399-415.

Granbois, D. H. and Willett, R. P. (1970), "Equivalence of Family Role: Measures Based on Husband and Wife Data," J. of Marriage and Family, Vol. 32, No.1 (February), 68-72.

Haberman, P. W. and Elinson, J. (1967), "Family Income Reported in Surverys: Husbands versus Wives," J. of Marketing Research, Vol. IV (May), 191-94.

Hornik, J. (1982), "Impact of Pre-Call Request Form and Gender Interaction on Response to a Mail Survey," J. of Marketing Research, Vol. XIX (February), 141-51.

Houston, M. J. and Ford, N. M. (1976), "Reconsidering the Scope of Methodological Research In Mail Survey," J. of Marketing Research, Vol. III (November), 397-403.

Kanuk, L. and Berenson, C. (1975), "Mail Survey and Response Rates: A Literature Review," J. of Marketing Research, Vol. 12 (November), 440-53.

Knoke, D. and Burke, P. J. (1980), Log-Linear Models, Beverly Hills, Sage Publications.

Madow, W. G., Nisselson, H. and Olkin, I. (1983), Incomplete Data in Sample Surveys, New York, Academic Press.

Peterson, R. A. and Kerin, R. A. (1981), "The Quality of Self-Report Data: Review and Synthesis," in B. M. Enis and K. J. Roering (eds.), Review in Marketing 1981, Chicago, American Marketing Association, 5-20.

Peterson, R. A., Leone, R. P. and Sabertehrani (1980), "Investigating 'Income Refusals' in a Telephone Survey by Means of Logit Analysis," in K. B. Monroe (ed.), Advances in Consumer Research, Vol. VIII, Ann Arbor, Association for Consumer Research, 287-291.

Sheth, J. N. (1977), "Demographics in Consumer Research," J. of Business Research, Vol. 5 (March), 129-138.

Sheth, J. N., LeClaire, A. and Wachspress, D. (1980), "Impact of Asking Race Information in Mail Survey," J. of Marketing, Vol. 44 (Winter), 67-70.

Sudman, S. and Bradburn, N. N. (1982), Asking Questions, San Francisco, Jossey-Bass Publishers.

Smelser, N. J. (1976), The Sociology of Economic Life, New Jersey, Prentice-Hall, Inc., Englewood Cliffs (2nd ed.).

Yu, J. and Cooper, H. (1983), "A Quantitative Review of Research Design Effects on Response Rates to Questionnaires," J. of Marketing Research, Vol. XX (February), 36-44.

TABLE 1

RESPONSES AND NONRESPONSES

	INC	AGE	EDUC	RACE	SEX
Not reported	351	9	6	6	1
Reported	2154	2496	2499	2499	2504
Total	2505	2505	3505	2505	2505

TABLE 2

INCOME NONRESPONSE BY AGE, EDUCATION, RACE AND SEX(%)

	AGE		EDUC		RACE		SEX		TOTAL
	≤34	≥35	≤ high school	≥ some college	Non-white	White	Female	Male	
Income Nonresponse (%)	9.7	17. 3	14.6	13.4	14.4	13.7	14.2	13.5	13.9
n =	(1112)	(1372)	(1076)	(1408)	(561)	(1923)	(1244)	(1240)	(2484)

TABLE 3

MODEL COEFFICIENTS

Effect	(1) Coefficient	(2) Coefficient * 2	(3) Antilog
INC by AGE	-.2065[a)] (.0377)	-.4131 (.054)	.6562
INC by EDUC	.0617 (.0376)	.1235 (.0752)	1.1314
INC by RACE	-.0169 (.0400)	-.0339 (.0800)	.9667
INC by AGE by RACE	-.0768 (.0377)	-.1535 (.0754)	.8577
INC by EDUC by RACE	.0766[b)] (.0376)	.1531 (.0752)	1.1654
INC	-.9692[b)] (.0400)	-1.9383 (.0800)	.1439

a) $p < .001$ b) $p < .05$

TABLE 4

OBSERVED AND ESTIMATED NUMBER OF INCOME NONRESPONSES

SEX	RACE	EDUC	AGE	Observed	Estimated	Residuals
Female	Non-white	≤ High School	≤34	9	8.20	.80
Female	Non-white	≤ High School	≥35	20	24.19	-4.19
Female	Non-white	≥ Some College	≤34	4	3.11	.89
Female	Non-white	≥ Some College	≥35	7	8.47	1.47
Female	White	≤ High School	≤34	24	18.66	5.34
Female	White	≤ High School	≥35	38	38.88	3.12
Female	White	≥ Some College	≤34	26	28.36	-2.36
Female	White	≥ Some College	≥35	49	45.45	3.55
Male	Non-white	≤ High School	≤34	6	6.79	-.79
Male	Non-white	≤ High School	≥35	23	18.82	4.18
Male	Non-white	≥ Some College	≤34	3	3.90	-.90
Male	Non-white	≥ Some College	≥35	9	7.53	1.57
Male	White	≤ High School	≤34	14	12.64	1.36
Male	White	≤ High School	≥35	23	32.82	-9.82
Male	White	≥ Some College	≤34	22	26.35	-4.34
Male	White	≥ Some College	≥35	68	64.85	3.15

APPENDIX A. THE DATA

		SEX																
		Female								Male								
		RACE								RACE								
		Non-White				White				Non-White				White				
		EDUC				EDUC				EDUC				EDUC				
		≤ High School		≥ Some College		≤ High School		≥ Some College		≤ High School		≥ Some College		≤ High School		≥ Some College		
		AGE		AGE		AGE		AGE		AGE		AGE		AGE		AGE		Total
		≤34	≥35	≤34	≥35	≤34	≥35	≤34	≥35	≤34	≥35	≤34	≥35	≤34	≥35	≤34	≥35	
NC	1	9	20	4	7	24	38	26	49	6	23	3	9	14	23	22	68	345
	2	78	79	51	47	162	183	242	225	66	54	66	39	112	185	227	323	1391
otal		87	99	55	54	186	221	268	274	72	77	69	48	126	208	249	391	2484