Identifying Unreliable Respondents: Can Missed Interviews Proxy for Validation?*

Christopher R. Bollinger, Martin H. David, and Kent Marquis

M. H. David, 1180 Observatory Dr., University of Wisconsin, Madison WI 53706

Keywords **Response error, missing interviews**

Structure of the problem

Groves (1989) exhorts survey experts to consider total survey errors. Total error includes sampling error, processing error, and response error. Attention to response error has been the least of the three. Fortunately panel surveys offer some insight into the relationship between error and difficulties in contacting respondents because each respondent is to provide data at several points in time. As it were, the enumerator has several trials in this measurement experiment, and the respondent may or may not provide data at each wave of interviewing.

Panel surveys are characterized by increasing levels of non-response at successive waves. In the *1984 Survey of Income and Program Participation* which is studied here, unit non-response in the first wave disqualifies households from further contacts. Additional wave non-response may be random or it may reflect attrition at the time of subsequent contacts (Zabel 1996). Lansing *et al.* *(*1961) conjectured that any non-response may be correlated with failure to give correct information. Our objective is to determine whether propensity to miss interviews and propensity to give erroneous data are related.

The conceptual structure of the problem we investigate includes four random variables and two time points. Y^*_{ti} is the vector of four behaviors pertaining to use of Food Stamps at time t for individual i. X_{ti} is a vector of J conditioning variables used to model behavior. A probability sample of households yields interview data for N individuals in H households. Matching those individuals to administrative records yields data on *use* of Food Stamps for the sample, y^*_{ti0}. In principle, those data can be modelled by probit analysis to find maximum likelihood estimates for the linear index, $X_{ti0}\beta_0$,that describes the $\Pr[y^*_{ti0} = 1] = F(X_{ti0}\beta_0)$ where $F(.)$ is the standard normal distribution function. Food Stamps are administered to households; thus we could also refer to $y^*_{1h0} = 1$ as the appropriate measure of use. Because of response errors and missing waves in the panel sample, this simple procedure must be elaborated when estimates are based on survey measures of Food Stamp use, y_{th0}.

y_{Th1} is an indicator for interviews missed by members of households in the first interview over the life of the panel, T waves.

Interviews encompass response errors. Omissions, false negative responses, to the screening question: "Are you certified to receive Food Stamps?" are indicated by y_{1i2} which may be 1 only for Food Stamp users. Commission errors, false positive responses, are indicated by y_{1i3}. y_{1i3} may be 1 only for non-users of Food Stamps. The indicators for survey response, administrative data, and errors are linked

$$y_{ti0} \equiv y^*_{ti0}(1 - y_{ti2}) + (1 - y^*_{ti0})y_{ti3} \qquad (1)$$

The expectation of (1) gives

$$\Pr[y_{ti0} = 1] = \Pr[y^*_{ti0} = 1] \cdot (1 - q_{ti} - p_{ti}) + p_{ti} \qquad (2)$$

where $q_{ti} = \Pr[y_{ti2} = 1|X_{ti2}]$ and

$p_{ti} = \Pr[y_{ti3} = 1|X_{ti3}]$.

Recall that Y^*_{1i} corresponds to data that are available from *validation* data (survey measurements matched to administrative records for the same units) For validation data, the correct model for Food Stamp participation can be estimated either from Y^*_{ti} or the corresponding vector Y_{ti},where the survey response replaces the administrative record data. *Primary* data obtained through conventional sample survey methods do not obtain y^*_{1i0} and imply that (2) must be used in estimating a probit on participation (Bollinger and David 1996).

We assume that Y^*_{1i} can be modelled by a multi-variate probit, where $F(BX_{1i}, \Omega_4)$ encompasses the four indicators for participation, missing waves and errors. B $\equiv$ $(\beta_0, \beta_1, \beta_2, \beta_3)$ estimates all behavioral responses to conditioning variables. Ω_4 describes interaction of the four random variates. Deleting the first row and column from Ω_4, gives $\Omega \equiv \begin{pmatrix} 1 & \rho_o & \rho_c \\ \rho_o & 1 & * \\ \rho_c & * & 1 \end{pmatrix}$. The random error ε^*_{1i0}, relating to participation, is assumed independent of the other random errors. The random error associated with missing the second, or subsequent, waves to T, is not independent of the random errors associated with erroneous responses, as indicated by ρ_o and ρ_c.

A variety of "missing data" problems affect estimation of B and Ω_4. First, few households are simultaneously users and non-users of Food Stamps within the four-month reference period. This precludes reliable estimation of the parameter "*". Second, the universe of households *eligible* for Food Stamps is defined by z_{ti}. z_{1i} is missing, precluding simultaneous estimation of B at $t = 1$. $\beta_1, \beta_2, \beta_3$, and Ω can be estimated using survey data from wave 1 and the record of missed interviews to T. The estimation of the household error process encompassed by $\varepsilon_{\cdot hT} \equiv (\varepsilon_{\cdot Th1}, \varepsilon_{\cdot 1h2}, \varepsilon_{\cdot 1h3})$ is explained below.

At t the vector Y^*_{th} is unobserved. Records of Food Stamp use are unavailable, and response errors can not be identified. Survey data provide y_{th0} that is measured with error on an attritted sample. z_{th} is observed. Because the sample of households *eligible* for Food Stamps is identified, β_0 can be identified using the $\hat{q}_{th}, \hat{p}_{th}$ predicted from models estimated here.

Estimating errors of the survey design

The log likelihood of the validation sample is

$$\begin{aligned} \mathcal{L} \;=\; & \sum_{h=1}^{H} y^*_{1h0}[y_{Th1} \cdot y_{1h2} \cdot PO_{11} + \\ & y_{Th1} \cdot (1 - y_{1h2}) \cdot PO_{10} + \\ & (1 - y_{Th1}) \cdot y_{1h2} \cdot PO_{1} + \\ & (1 - y_{Th1}) \cdot (1 - y_{1h2}) \cdot PO_{00}] + \\ & [1 - y^*_{1h0}][y_{Th1} \cdot y_{1h3} \cdot PC_{11} + \\ & y_{Th1} \cdot (1 - y_{1h3}) \cdot PO_{10} + \\ & (1 - y_{Th1}) \cdot y_{1h3} \cdot PO_{01} + \\ & (1 - y_{Th1}) \cdot (1 - y_{1h3}) \cdot PC_{00}] \qquad (3) \end{aligned}$$

$\mathrm{PO}_{r,s}$ denotes the probability of observing $y_{1h1} = r$ and $y_{1h2} = s$. and $\mathrm{PC}_{r,s}$ denotes the probability of realizing $y_{1i1} = r$ and $y_{1i3} = s$. $r, s = 0, 1$ are binary indexes.

Data and preliminary findings

We deal with an aggregate of individuals located in the household entering the panel in wave 1. The reason for this is that the presence of one or more cooperative household members may allow us to obtain correct data on household use of Food Stamps. Also use of Food Stamps reported by a person other than the administratively certified person is irrelevant, so long as all household members are benefiting (nearly all households (Martini 1992). Dissolution of the household following wave 1 is irrelevant

Table 1: Household FS reports by record status for 4-months

	FS record		
SIPP FS	No	Yes	Total
No	2469	29	2498
Yes	9	178	187
Total	2478	207	2685

to the errors made in wave 1. We also deal with an aggregate of time – a four-month reference period. Failure to respond affirmatively to a screening question for that reference period accounts for 90% of the errors of omission. These considerations lead to the counts of errors shown in Table 1.

The result of estimating β_2, β_3 without considering the multi-variate model is given in Bollinger and David (1993). Systematic effects of income per capita on both errors of omission and commission were estimated. The data are taken from Marquis and Moore's (1990) validation study of Food Stamps and other income maintenance programs. Measures of household missing waves come from Bollinger and David (1995). These measures differ from historic analyses of attrition in several important ways. First individuals are aggregated. Secondly, we consider missing waves that are not attrition patterns. While analysis of attrition dominates the literature (Zabel 1996) non-attrition patterns are almost as common; 12.1% of the sample show other patterns (Lepkowski *et al.* 1989).

Any pattern of wave non-response in waves 2-9 in the household is labelled *anymiss.* A missing wave pattern that encompasses all members of the household is *fammis. Twomis* and *fammis2* define a pattern ending in two missed interviews for at least one household member, and all household members respectively. *Pctmis* reflects the proportion of interviews missed. All of these measures reflect the survey design: Some sample elements were censored after wave 4 and others after wave 8. Table 2 shows the probability associated with these differing definitions of household missing interviews and other descriptive statistics for the validation sample. Later we will introduce one more measure, defined for the first year of the panel; *Miss2-4* is an indicator for any household containing an individual who missed wave 2, wave 3, or wave 4.

The principal comment on regressors used, is that per capita income includes Food Stamp vouchers. Multi-variate analyses of surveys show that Food Stamp participation is inversely related to earnings. We need to find a consistent estimator for the coefficient on earnings in future work. The task in this paper is to determine whether wave 1 response errors and subsequent wave missingness are linked to a common, unobserved source and systematically related to observed attributes of households and their members.

Results

Tables 3 and 4 give the estimates obtained from (3). Table 3 displays estimates of the coefficient vector β_1; Table 4 displays β_2, β_3,followed by ρ_o, ρ_c. The results for panel missingness are an artifact of the aggregated nature of the variable and the small size of the sample, 2685 households. (Zabel 1996 has a much larger set of regressors for *individual* panel missingness.) When the panel missingness variable is defined by the absence of the last two interviews, individuals attrit from larger households at a higher rate and larger households attrit at a lower rate than small households. The probability of false negative Food Stamp reports increases as income increases (for every measure of panel missingness except the proportion of interviews missed). The covariance of panel missingness and omissions is significantly positive, e.g. $\rho_o > 0$. This last finding suggests several possibilities: a random event, such as illness, provokes both response error and later missing waves. An omitted attribute of the respondent may also lead to poor cooperation (Eisenhower *et al.* 1991). For example, the respondent may fear strangers or is unable to comprehend questions asked. Some aspect of the interviewer could also lead to this finding.

The last column of Tables 3 and 4 show a remarkably similar response to regressors for the first year and the panel as a whole. The error models are identical and the level of missing waves is lower, as one would expect from the shorter period. We conclude that timing of missing waves has little to do with error, and that any missing wave gives some evidence of a propensity to errors of omission.

We conclude that missing waves in a panel and response error can not be separated. Weighting techniques are inadequate tools for selection bias in the estimation of models. Furthermore, estimation of nonlinear models, such as probits, must be specified to include information on predicted response errors.

ACKNOWLEDGMENT

*Supported by NSF grant #SBR9511935 and #SBR-952174. We thanks the Census Bureau for indispensable assistance. The views expressed are the authors', and do not represent the views of the Census Bureau.

REFERENCES

Bollinger, Christopher R. and Martin H. David. 1993. Modeling Food Stamps in the presence of reporting errors. *Proceedings of the Bureau of the Census Annual Research Conference.* Washington DC: Bureau of the Census. 289-310.

Bollinger, Christopher R. and Martin H. David. 1995. Sample attrition and response error: Do two wrongs make a right? *Proceedings of the Bureau of the Census Third Annual Research Conference.* Washington DC: Bureau of the Census.

Bollinger , Christopher R. and Martin H. David. 1996 Modeling discrete choice with response error. (unpublished working paper) Madison WI: Institute for Research on Poverty.

Eisenhower, Donna, Nancy A. Mathiowetz, and David Morganstein. 1991. Recall error: Sources and bias reduction techniques. In Paul P. Biemer *et al. Measurement errors in surveys* New York NY: John Wiley, 127-144.

Groves, Robert M. 1989. *Survey Errors and Survey Costs.* New York NY: John Wiley

Lansing, John B., Gerald P. Ginsburg, and Kaisa Braaten. 1961. *An Investigation of Response Error.* Urbana IL: bureau of Economic and Business Research, University of Illinois.

Marquis, Kent and Jeffrey Moore. 1990b. Measurement Errors in SIPP Program Reports. *Proceedings of the Bureau of the Census 1990 Annual Research Conference.* Washington DC: Bureau of the Census721-745.

Martini, Alberto. 1992. Participation in the Food Stamp Program: A multivariate analysis *Current Perspectives on Food Stamp Participation,* Alexandria VA: USDA, Food and Nutrition Service, Office of Analysis and Evaluation.

Zabel, Jeffrey E. (forthcoming). A comparison of attrition in the *Panel Survey of Income Dynamics* and the *Survey of Income and Program Participation. Survey Methodology.*

Table 2: Description of the *validation sample*

Variable	Definition	mean	std dev
	Dependent – missingness		
Anymis	At least 1 **individual** missed at least 1 interview	0.28	0.45
Fammis	The **household** missed at least 1 interview	0.22	0.41
Twomis	At least 1 **individual** missed at least 2 interveiw	0.22	0.41
Fammis2	The **household** missed at least 2 interview	0.16	0.37
Pctmis	**Percentage** of missed interviews in household	0.14	0.28
Miss2-4	At least 1 **individual** missed **wave 2, 3, or 4**		
	Regressors		
PerInc	Total HH **income** / household size ($\$/mo.$)	949.	1431.
Gender F	**Female** householder	0.29	0.45
Married	**Marrried** householder	0.61	0.49
Ed	Education of householder	12.1	3.2
Size	Number of individuals in Household	2.65	1.54
Earn	Total HH **earnings**/ size ($\$/mo.$)	529.	720.
Proxy	Other person reports for respondent	0.256	0.436
M*Gender	Married - Gender interaction	.0399	0.196

Table 3: Missingness in the tri-variate model

	Missingness variable					
Variables	Anymis	Fammis	Twomis	Fammis2	Pctmis	Miss2-4
Constant	-0.68**	-0.86**	-0.99**	-1.17**	0.12**	-1.26
	(0.16)	(0.17)	(0.17)	(0.18)	(0.04)	(0.18)
PerInc	0.007	-0.01	0.003	-0.02	-0.002	0.008
	(0.02)	(0.02)	(0.02)	(0.02)	(0.004)	(0.02)
Sex	0.06	0.11*	0.13*	0.18**	0.02*	0.054
	(0.07)	(0.07)	(0.07)	(0.08)	(0.01)	(0.076)
MS	-0.21**	-0.08	-0.12	0.03	-0.009	-0.17**
	(0.07)	(0.08)	(0.08)	(0.08)	(0.015)	(0.08)
Ed	-0.0005	0.005	-0.004	0.006	-0.0009	0.012
	(0.008)	(0.008)	(0.009)	(0.009)	(0.002)	(0.009)
Size	0.06**	-0.02	0.06**	-0.05**	-0.004	0.06**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.004)	(0.02)

** significant at the 5% level; *significant at the 10% level

Table 4: Response error in the tri-variate model

	Missingness variable					
Variables	Anymis	Fammis	Twomis	Fammis2	Pctmis	Miss2-4
			Omission			
Constant	-1.30**	-1.29**	-1.28**	-1.29**	-1.27**	-1.27**
	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)	(0.18)
Earn ($\$ \cdot 10^{-3}$)	1.24**	1.35**	1.28**	1.31**	1.29**	1.25**
	(0.30)	(0.32)	(0.33)	(0.32)	(0.32)	(0.32)
			Commission			
Constant	-2.45**	-2.45**	-2.44**	-2.44**	-2.46**	-2.45**
	(0.15)	(016)	(0.15)	(0.15)	(0.15)	(0.16)
Earn ($\$ \cdot 10^{-3}$)	-0.90	-0.90	-0.94	-0.92	-0.88	-0.88
	(0.73)	(0.78)	(0.74)	(0.74)	(0.73)	(0.79)
			Correlation			
Ro	0.28**	0.35**	0.34**	0.30*	0.28**	0.35**
	(0.13)	(0.14)	(0.14)	(0.16)	(0.10)	(0.15)
Rc	0.083	0.06	-0.06	0.02	0.06	0.094
	(0.16)	(0.18)	(0.17)	(0.13)	(0.09)	(0.17)
Likelihood/N	-0.6423	-0.5690	-0.5677	-0.4881	-.4881	-0.4960

** significant at the 5% level; *significant at the 10% level