Covariates of Unit Nonresponse Error Components Based on Proxy Household Information

H. Öztaş AYHAN

Department of Statistics
Middle East Technical University
Ankara, Turkey

OUTLINE OF PRESENTATION

1. INTRODUCTION
2. SURVEY METHODOLOGY
3. COVARIATES OF NONRESPONSE
4. PROPOSED MODELS AND TESTING
5. CONCLUSIONS

1. INTRODUCTION

Covariates of unit nonresponse error components has been a concern of survey researchers as a major part of the total survey error.

Components of unit nonresponse error is basically associated with the factors related to the reasons of survey non-participation.

In order to have logical causality measures, one has to identify the direct and indirect factors affecting such relations. In many cases, information on such ideal factors may not be available as a survey variable, due to the current objectives of such a survey.

Alternative information can be derived from the other existing survey variables which are naturally available due to the survey objective. Consequently, the researchers have to make sense out of such information, because the ideal information which will explain the causality may not be available.

With a limited research budget, one can obtain information only on a reasonably small scale. On the other hand, for a large scale survey, additional questions will also bring extra cost, which may not be tolerable by the survey management. Under the circumstances, another alternative may be to utilize the best of the available information.

2. SURVEY METHODOLOGY

2.1. Sample Design and Implementation

The sample design and sample size of the *Turkish Demographic and Health Survey* (TDHS) – 2003 makes it possible to perform analyses for Turkey as a whole, for urban and rural areas and for the five demographic regions of the country. A weighted, multistage, stratified cluster sampling approach was used in the selection of the survey sample.

The results of the household and individual questionnaire executions are summarized in *Table 1*. Information is provided on the overall coverage of the sample, including household and individual nonresponse rates.

Table 1. ***Results of the Household and Individual Interviews in 2003 Turkish Demographic and Health Survey***

Results	***Urban***	***Rural***	**Total**
Household interviews:			
Dwellings sampled	**8718**	**2941**	**11659**
Households interviewed	**7956**	**2880**	**10836**
Household nonresponse rate	***0.087***	***0.021***	***0.071***
Individual interviews:			
Eligible person selected	**6259**	**2188**	**8447**
Eligible person interviewed	**5976**	**2099**	**8075**
Individual nonresponse rate component	***0.045***	***0.041***	***0.044***
Individual person nonresponse rate *	***0.128***	***0.061***	***0.112***

***: IPNRR = [1 – HHRR * IRRC]**

2.2. Questionnaire Design

The data collection for household sample surveys have been executed in two stages; the completion of the household schedule, and the individual survey. The household schedule is completed by a selected adult member of the household, as a proxy respondent for the other members of the household, and a self respondent for him/herself.

For the individual survey, data are only collected from the individual respondent as a self respondent, and no information is available for the non-respondents. On the other hand, household schedule also contains some more additional information about other characteristics of the responders and non-respondents of the individual survey.

For the responding households, generally the household schedule contains full information on all household members. On the other hand, the selected household member for the individual survey may or may not respond to the individual person's interview. Consequently, we will have two possible groups for the individual survey; respondents and non-respondents.

This study combines the household based proxy information for selected variables, and response-nonresponse outcome information of the individual person's survey from the same household.

3. COVARIATES OF NONRESPONSE

The following household information are obtained from the household schedule by proxy interviews;

A. <u>Independent Survey Variables</u>: (*Household based*)

(1). Stratification / Survey Variables:

- Region
- Type of place of residence

(2). Household Based Proxy Individual Variables:

- Gender
- Age groups
- Place of birth
- Maternal and paternal survival
- Migration and mobility
- Literacy and education status
- Work status
- Marital status

(3). Housing Characteristics:

- Household ownership
- Safe water access
- Sanitary toilet
- Number of rooms
- Household durability
- Household facilities
- Household income

B. <u>Dependent Survey Variable</u>: (*Indv. Survey based*)

- Binary nonresponse information

Table 2. *Current and Generated Variables, Options and Their Frequencies*

Name of variables	Explanation	Weighted percent
Response–Nonresponse	1 (Nonresponse)	4.7
	0 (Response)	95.3
hv017- Number of visits to household	1	79.7
	2	14.9
	3	5.4
hv024 – Regions	1 West	40.7
	2 South	12.7
	3 Central	23.1
	4 North	7.3
	5 East	16.2
hv025 - Type of place of residence	1 Urban	71.2
	2 Rural	28.8
hv270 - Wealth index	1 Poorest	15.6
	2 Poorer	18.1
	3 Middle	20.2
	4 Richer	22.4
	5 Richest	23.6
hv102 - Usual resident	0 No	3.6
	1 Yes	96.4
sh26 - Currently working	0 No	75.1
	1 Yes	24.9
SANITATE- Sanitary toilet	0 No	90.7
	1 Yes	9.3
SAFEWAT – Safewater	0 No	92.4
	1 Yes	7.6
CROWD – Number of persons per room	0 less than 3	80.5
	1 more than 3 and over	19.5
Educ – Education level	1 No education/ Primary incomplete	22.1
	2 Primary complete/ secondary incomplete	60.7
	3 Secondary +	17.2
hv116 - Marital status	1 Currently married	94.7
	2 Formerly/ever married	5.3
agegroup(*) – Age groups	1 15-19	3.0
	2 20-24	12.9
	3 25-29	18.2
	4 30-34	18.3
	5 35-39	17.5
	6 40-44	16.5
	7 45-49	13.5

4. PROPOSED MODELS AND TESTING

4.1. Search for Models

In this study, individual survey respondent's related household schedule characteristics are used as possible covariates for the non-response error. The possible covariates are evaluated under several alternative statistical models. For this purpose, several generalized linear models have been examined. As possible alternatives; *loglinear model, logit model, probit model,* and *logistic regression model*s have been evaluated. After the examination of the current available variables, *multiple logistic regression model* has been selected.

The present model takes non-response as the binary dependent variable which is associated with the other household covariates. In order to test our model, the latest TDHS – 2003 data is used. Questions and topics which are listed in Section 3 were asked during the household interviews.

The household survey and individual person's survey data sets are combined under the weighted, stratified cluster design, for the survey analysis. The *SPSS 13.0's "complex samples" feature* were used to perform *binary multiple logistic regression,* where the sample design was naturally taken into account.

4.2. Inference from Multiple Logistic Regression

A multiple logistic regression model has been proposed to explain the effect of covariates on survey unit nonresponse for this study. After the regression diagnosis as outlier detection and collinearly tests performed the following model and results were obtained. Some variables did not taken care of as work type since only a portion of women are working, only variables available for “all cases” were put into the model to increase the number of cases of the model.

The hypothesis to be tested will be;

$H_0 : \beta_i = 0$ versus $H_a : \beta_i \neq 0$

The multiple logistic regression prediction equation for an *S*–shaped curve for the desired probability *p* is;

$$p = \exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i \, x_i\right) \Bigg/ \left[1 + \exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i \, x_i\right)\right]$$

Within the *S*–shaped regression model, the probability *p* falls between 0 and 1 for all possible x values.

Test statistics for the regression model coefficients;

$$t_i = \left(\hat{\beta}_i - \beta_i\right) / se(\hat{\beta}_i)$$

4.3. The Odds Ratio

The odds ratio is used to interpret the computed coefficients of the multiple logistic regression prediction equation, in terms of relative comparative risks.

Table 3. ***The Data Layout Structure for Odds***

Variables	Nonresponse	Response	**Total**
Variable *Option A*	n_{11}	n_{12}	n_{1+}
Variable *Option A^c*	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

The desired (success) probabilities for the two groups are;

π_1 is estimated by $p_1 = n_{11} / n_{1+}$

π_2 is estimated by $p_2 = n_{21} / n_{2+}$

In 2x2 contingency tables, the *relative risk* is the ratio of the desired probabilities for the two groups.

The ***Relative Risk*** = π_1 / π_2

The ***ratio of odds*** from two rows;

$$\theta = \frac{\pi_1(1-\pi_1)}{\pi_2(1-\pi_2)} = \frac{\pi_{11}\ \pi_{22}}{\pi_{12}\ \pi_{21}}$$

Sample odds **(cross–product)** ***ratio*** is;

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}\ n_{22}}{n_{12}\ n_{21}}$$

The odds ratio can equal any nonnegative number.

The odds ratio can be interpreted as;

- When $1 < \theta < \infty$, the odds of success are higher in row 1 than in row 2.
- When *X* and *Y* are independent, $\pi_1 = \pi_2$ so that $\theta = \left[odds_1 / \ odds_2 \right] = 1$.
- When $0 < \theta < 1$, a success is likely in row 1 than in row 2, that is $\pi_1 < \pi_2$.

4.4. Model Based Survey Statistics and Outcomes

For the following proposed model is fitted to the TDHS 2003 data.

$$p = \Pr(Y_i = 1) = \exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i \, x_i \right) \Big/ \left[1 + \exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i \, x_i \right) \right]$$

Where,

$\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i \, x_i$ = –1.615 + 0.563*hv024(1) + 0.549* hv024(2) + 0.470* hv024(3) + 1.577*hv102(0) – 0.451*sh26(0) – 0.656*hv116(1) – 0.557*agegroup(2) – 0.433*agegroup(3) – 0.469*agegroup(4) – 0.448*agegroup(5)

Table 4. ***Several Pseudo R Square Values for the Model***

Cox and Snell	0.21
Nagelgerke	0.66
McFadden	0.56

The Nagelgerke R square is 0.66 so the power of the model is low but the model is significant (with a *p* value of 0.000, and Wald statistics value = 7.289, df 1 = 25, df 2 = 322).

Table 5. ***Results of Test Statistics for Model Effects***

Sources	***df 1***	***df 2***	***Wald F***	***Significance***	***Indicator***
(Corrected model)	25	322	7.29	0.00	*
(Intercept)	1	346	54.61	0.00	*
hv017 - Number of visits	2	345	3.12	0.05	*
hv024 – Region	4	343	2.63	0.03	*
hv025 - Type of place of residence	1	346	0.97	0.33	
hv270 - Wealth index	4	343	1.03	0.39	
hv102 - Usual resident	1	346	63.59	0.00	*
sh26 - Currently working	1	346	7.28	0.01	*
SANITATE - Sanitary toilet	1	346	1.09	0.30	
SAFEWAT - Safewater	1	346	0.00	0.96	
CROWD - No of persons per room	1	346	0.30	0.58	
Educ - Education level	2	345	5.43	0.00	*
hv116 – Marital status	1	346	10.35	0.00	*
Age groups	6	341	1.88	0.08	

Table 6. *Multiple Logistic Regression Model Parameter Estimates*

Variable	Category	$\hat{\beta}_i$	$se(\hat{\beta}_i)$	t_i	*df*	*p-value*	*deff*	$\hat{\theta}$	Ind
	(Intercept)	-1.615	0.560	-2.885	346	0.00	1.54	0.20	*
hv017- Number of visits	1	-0.284	0.282	-1.004	346	0.32	1.69	0.75	
	2	0.192	0.296	0.650	346	0.52	1.74	1.21	
	3	0	.	.	.	.	.	1.00	
hv024 - Region	1 West	0.563	0.201	2.803	346	0.01	1.11	1.76	*
	2 South	0.549	0.238	2.309	346	0.02	1.21	1.73	*
	3 Central	0.470	0.224	2.098	346	0.04	1.19	1.60	*
	4 North	0.190	0.284	0.671	346	0.50	1.01	1.21	
	5 East	0	.	.	.	.	.	1.00	
hv025 - Type of place of residence	1 Urban	0.170	0.173	0.983	346	0.33	1.43	1.19	
	2 Rural	0	.	.	.	.	.	1.00	
hv270 - Wealth index	1 Poorest	-0.238	0.277	-0.859	346	0.39	1.76	0.79	
	2 Poorer	-0.358	0.206	-1.735	346	0.08	1.24	0.70	
	3 Middle	-0.264	0.210	-1.258	346	0.21	1.50	0.77	
	4 Richer	-0.343	0.197	-1.739	346	0.08	1.49	0.71	
	5 Richest	0	.	.	.	.	.	1.00	
hv102 - Usual resident	0 No	1.577	0.198	7.974	346	0.00	1.30	4.84	*
	1 Yes	0	.	.	.	.	.	1.00	
sh26 - Currently working	0 No	-0.451	0.167	-2.699	346	0.01	1.83	0.64	*
	1 Yes	0	.	.	.	.	.	1.00	
SANITATE- Sanitary toilet	0 No	-0.280	0.268	-1.042	346	0.30	1.69	0.76	
	1 Yes	0	.	.	.	.	.	1.00	
SAFEWAT - Safewater	0 No	-0.011	0.243	-0.045	346	0.96	1.53	0.99	
	1 Yes	0	.	.	.	.	.	1.00	
CROWD – no of persons per room	0 less than 3	-0.114	0.208	-0.548	346	0.58	1.68	0.89	
	1 more than 3 and over	0	.	.	.	.	.	1.00	
Educ – education level	1 No education/ Primary incomplete	0.335	0.245	1.366	346	0.17	1.58	1.40	
	2 Primary complete/ secondary incomplete	-0.198	0.178	-1.114	346	0.27	1.42	0.82	
	3 Secondary +	0	.	.	.	.	.	1.00	
hv116 - marital status	1 Currently married	-0.656	0.204	-3.217	346	0.00	1.25	0.52	*
	2 Formerly/ever marr.	0	.	.	.	.	.	1.00	
Age Group	1 15-19	0.136	0.369	0.368	346	0.71	1.63	1.15	
	2 20-24	-0.557	0.234	-2.384	346	0.02	1.26	0.57	*
	3 25-29	-0.433	0.192	-2.253	346	0.02	1.15	0.65	*
	4 30-34	-0.469	0.197	-2.374	346	0.02	1.20	0.63	*
	5 35-39	-0.448	0.215	-2.083	346	0.04	1.47	0.64	*
	6 40-44	-0.379	0.216	-1.754	346	0.08	1.55	0.68	
	7 45-49	0	.	.	.	.	.	1	

5. CONCLUSIONS

For the coefficients of this model, the following results can be summarized in terms of odds ratios;

- The probability of being a "non-responder" women is 1.76, 1.73 and 1.60 times higher for women who is in West, South and Central regions when compared to women in East region.

- Temporary members of the household are 4.84 times more "non-responders" than the usual members of the household.

- Non-working women are 1.56 (1 / 0.64) times better responders than working women.

- Similarly, currently married women are 2 (1 / 0.52) times better responders.

- People in middle age groups of 20-24 and 35-39, are also 1.5 times better responders than the oldest age group.