

New Developments in Nonresponse Adjustment Methods

Fannie Cobben*

January 23, 2009

1 Introduction

In this paper, we describe two relatively new techniques to adjust for (unit) nonresponse bias: The sample selection model (Heckman, 1979) and the propensity score method (Rosenbaum and Rubin, 1983). We show how the sample selection model can be applied to the situation of nonresponse in section 2. In section 3, we describe a number of methods that use response propensities to adjust for nonresponse bias. In addition, we discuss the application of these methods to the POLS 2002 survey, in section 4. We end this paper with a discussion in section 5.

2 Sample selection model

2.1 The response selection model

The sample selection model was first proposed by Heckman (1979). In the context of survey nonresponse, the sample selection arises due to self-selection of respondents, either explicit (refusal) or implicit (not able, non-contact) and in some cases also due to actual sample selection (unprocessed cases). We refer to both types as *response selection*.

Let $i = 1, 2, \dots, N$ be the population of interest. The first order inclusion probabilities are denoted by π_i for $i = 1, \dots, n$. The selection indicator is denoted by δ_i which is 1 if element i is selected in the sample, otherwise it is 0. Let $d_i = 1/\pi_i$ be the design weight. A sample of size n is selected from the population.

*Statistics Netherlands, e-mail: F.Cobben@cbs.nl

For every sample element i we have a vector of auxiliary variables, denoted by $\mathbf{X}_i = (X_{1i}, X_{2i}, \dots, X_{Ji})'$. This information comes, for instance, from the population register. In addition, we have a survey item Y that is only observed for respondents. We assume that every element i in the population has a nonzero, unknown response probability, denoted by ρ_i . This corresponds to the Random Response model. If element i is selected in the sample, a random mechanism is activated that results with probability ρ_i in response and with probability $(1 - \rho_i)$ in nonresponse.

In this section, we describe the sample selection model for survey nonresponse in the simplified form where we only consider the final response and the outcome for the survey item. We will refer to this model as the *response selection model*. The generalisation to more response stages is discussed in section 2.2. The response selection model consists of two equations. The first equation models the response probability. The outcome of the first equation determines whether the survey item is observed. This equation is therefore referred to as the *selection equation*. The second equation accounts for the censoring of sample elements that do not respond, and models the survey item. Therefore the second equation is referred to as the *regression equation*.

The response selection model consists of two equations

$$\begin{aligned} \rho_i^* &= \mathbf{X}_i^R \boldsymbol{\beta}^R + \epsilon_i^R, \\ Y_i^* &= \mathbf{X}_i^Y \boldsymbol{\beta}^Y + \epsilon_i^Y \end{aligned} \quad (1)$$

for $i = 1, \dots, n$. The response probability of sample element i is denoted ρ_i^* , and its value for survey item Y is denoted Y_i^* . \mathbf{X}_i is a vector of auxiliary variables for sample element i that is available for all elements in the sample, i.e. for $i = 1, \dots, n$. The superscript R is used to denote variables that are related to response behaviour. The superscript Y is used to denote auxiliary variables that have a relationship with the survey item Y . The terms ϵ_i^R and ϵ_i^Y are error terms. Both equations in (1) are latent variable regression equations. We do not observe ρ_i^* , but instead we observe the binary response indicator R_i which is either 0 (nonresponse) or 1 (response). If $R_i = 1$, we observe Y_i^* , otherwise Y_i^* is missing, i.e. censored. Thus, we can define Y_i as

$$Y_i = \begin{cases} Y_i^*, & \text{if } R_i = 1 \\ \text{missing}, & \text{if } R_i = 0 \end{cases} \quad (2)$$

The bivariate probit model with sample selection (Van de Ven and Van Praag, 1981) can be regarded as a sample selection model for discrete outcomes. In case of a binary categorical survey item, for instance having a job

yes/no, the bivariate probit model with sample selection can be applied to handle selective nonresponse.

Together, (1) and (2) describe the response selection model. In the response selection model the error term distribution is assumed to be bivariate normal with variance σ^2 , which leads to following the joint distribution of the error terms

$$\begin{pmatrix} \epsilon_i^R \\ \epsilon_i^Y \\ \epsilon_i \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} 1 & \zeta\sigma \\ \zeta\sigma & \sigma^2 \end{bmatrix}\right) \quad (3)$$

with ζ the correlation between the two error terms, i.e. the standardised covariance $\frac{\zeta\sigma}{\sqrt{\sigma^2}}$.

We are interested in the population mean of survey item Y . The population consists of N elements, from which we select a sample of size n . The Horvitz-Thompson estimator \bar{y}_{ht} is an unbiased estimator for the population mean of a survey item Y , see Horvitz and Thompson (1952)

$$\bar{y}_{ht} = \frac{1}{N} \sum_{i=1}^N \delta_i \frac{Y_i}{\pi_i}$$

where δ_i is the sample selection indicator, which equals 1 in case element i is selected in the sample, and 0 otherwise. The first order inclusion probability for sample element i is denoted π_i . Due to nonresponse, this estimator becomes biased. The modified Horvitz-Thompson estimator \bar{y}_{ht}^r deals with nonresponse bias by adjusting the inclusion probabilities for the response probability, see for example Bethlehem (1988)

$$\bar{y}_{ht}^r = \frac{1}{N} \sum_{i=1}^n \frac{R_i Y_i}{\pi_i \rho_i}$$

In terms of the response selection model, we can modify the Horvitz-Thompson estimator by replacing Y_i with the expected value of Y_i conditional on $R_i = 1$. We denote by \bar{y}_{ht}^{sel} the modified Horvitz-Thompson estimator for the mean of survey item Y based on the response selection model

$$\bar{y}_{ht}^{sel} = \frac{1}{N} \sum_{i=1}^n \frac{E\left[Y_i | R_i = 1, \mathbf{X}_i^R, \mathbf{X}_i^Y\right]}{\pi_i} \quad (4)$$

We can calculate (4) by estimating the parameters of the response selection model. Estimation methods are described in Cobben (2009).

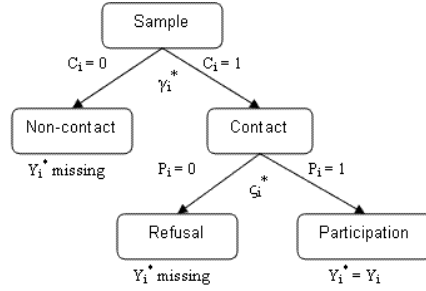


Figure 1: *The sample selection model with multiple selection equations*

2.2 Different types of respondents

It is often argued in the survey literature (Groves and Couper, 1998, Stoop, 2005) that we should distinguish between different types of response, for example contact and participation. This leads to a sequential representation of the response process, see figure 1. The selection equation in the response selection model described by (1) and (2) can be modified to account for the sequential nature of the response process. The resulting response selection model then consists of multiple selection equations. We illustrate this approach by regarding the response stages contact and participation, see figure 1. The generalisation to more response types, e.g. able to participate, is straightforward but cumbersome in notation.

The response selection model in (1) and (2) has to be extended to include two selection equations; one for contact and one for participation. Consequently, the selectivity bias is defined as a function of these two processes. The modified response selection model becomes

$$\begin{aligned}
 \gamma_i^* &= \mathbf{X}_i^C \boldsymbol{\beta}^C + \epsilon_i^C, \\
 \varrho_i^* &= \mathbf{X}_i^P \boldsymbol{\beta}^P + \epsilon_i^P, \\
 Y_i^* &= \mathbf{X}_i^Y \boldsymbol{\beta}^Y + \epsilon_i^Y
 \end{aligned} \tag{5}$$

for $i = 1, \dots, n$. The contact probability of sample element i is denoted γ_i^* and the participation probability is denoted by ϱ_i^* . Both γ_i^* and ϱ_i^* are latent variables. We do not observe them, but instead we observe the sign of the underlying processes. If $\gamma_i^* > 0$, the corresponding indicator C_i equals 1 and else $C_i = 0$. The same holds for P_i and ϱ_i^* . Hence C_i respectively $P_i = 1$ if element i is contacted respectively participates, and zero otherwise. Furthermore, P_i is only observed when $C_i = 1$, otherwise it is censored. For the outcome equation holds that Y_i^* is a latent variable that is only observed

when $C_i = 1$ and $P_i = 1$. In this situation we define Y_i as

$$Y_i = \begin{cases} Y_i^*, & \text{if } C_i = 1; P_i = 1 \\ \text{missing}, & \text{if } C_i = 1; P_i = 0 \text{ or } C_i = 0 \end{cases}$$

The vector of error terms is assumed to follow a multivariate $N_3(\mathbf{0}, \mathbf{\Sigma})$ distribution. A more detailed description of this model is given in Cobben (2009).

3 Response propensity methods

The idea of response probabilities has received much attention in survey methodological literature lately, see for example Lee (2006), due to the introduction of the propensity score method (Rosenbaum and Rubin, 1983) in treatment evaluation. Here, we describe how it can be used to adjust for nonresponse bias.

3.1 Response propensity

The response probabilities can be estimated based on the sample. By using an appropriate model based on auxiliary information, we can compute sample-based estimates of the response probabilities, i.e.

$$\hat{\rho}_i = \rho(\mathbf{X}_i) = P(R_i = 1 | \delta_i = 1, \mathbf{X}_i) \quad (6)$$

for $i = 1, 2, \dots, n$. We refer to $\rho(\mathbf{X}_i)$ as the *response propensity*. The response propensity is the estimated response probability conditional on the sample and the individual characteristics \mathbf{X}_i . The response propensity can be used to adjust for nonresponse in different ways. We distinguish between two approaches. The first approach uses the response propensities directly in the estimation of the survey item. This approach can be applied in two distinct ways, *response propensity weighting* and *response propensity stratification*. In this direct approach, the available auxiliary information is used to model the relationship between the response and the auxiliary variables. The second approach involves the relationship between the survey item and the auxiliary information as well. This approach is an extension of the GREG-estimator. We refer to it as the *response propensity GREG-estimator*. We outline these three estimators that make use of the response propensities to adjust for nonresponse.

3.2 Reponse propensity weighting

Following the suggestion of Särndal (1981) the response propensity can be inserted in the modified Horvitz-Thompson estimator which then becomes

$$\bar{y}_{ht}^r = \frac{1}{N} \sum_{i \in r} \frac{d_i Y_i}{\hat{\rho}(\mathbf{X}_i)} \quad (7)$$

we refer to this estimator as the *response propensity weighting* estimator.

3.3 Response propensity stratification

Another possibility of directly using the response propensities in the estimation of survey item Y is to stratify the sample based on the response propensities. Suppose we stratify the sample into $F = 5$ strata based on the response propensities. The strata are denoted by s_1, s_2, \dots, s_F . The sample size of stratum f is denoted by n_f ¹. Post-stratification assigns the same weight to all elements in the same stratum. The correction weight g_i for element i in stratum f is defined as

$$g_i = \frac{n_f}{n_{r,f}} \quad (8)$$

where $n_{r,f}$ is the number of responding sample elements in stratum f . Consequently, the post-stratification estimator can be expressed as

$$\begin{aligned} \hat{\bar{y}}_{ps}^{\hat{\rho}(\mathbf{X})} &= \frac{1}{N} \left(\frac{n_1}{n_{r,1}} \sum_{i \in s_1} Y_i + \frac{n_2}{n_{r,2}} \sum_{i \in s_2} Y_i + \dots + \frac{n_F}{n_{r,F}} \sum_{i \in s_F} Y_i \right) \\ &= \frac{1}{N} \sum_{f=1}^F n_{r,f} \bar{y}_r^{(f)} \end{aligned} \quad (9)$$

where $\bar{y}_r^{(f)}$ is the (unweighted) response mean for the survey item in stratum f . This estimator is referred to as the *response propensity stratification* estimator.

3.4 The response propensity GREG-estimator

In the response propensity weighting approach, the design weights are updated for the second phase in the sampling process by dividing them by the response propensities. The available information is used to model the relationship between the response and the auxiliary variables. We can extend this approach, by regarding the relationship between the survey item and

¹These sample sizes are random variables and not fixed numbers.

the auxiliary information as well. One possibility to do this, is by updating the weights in the GREG-estimator (Bethlehem, 1988). The g -weights based on this estimator then can be expressed as

$$g_{i\hat{\rho}(\mathbf{x})} = 1 + \left(\bar{\mathbf{X}} - \bar{\mathbf{x}}_{ht}^r\right)' \left(\sum_{i=1}^{n_r} \frac{d_i}{\rho(\mathbf{X}_i)} \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \mathbf{X}_i \quad (10)$$

in this expression, the design weights d_i are modified for the second phase which involves nonresponse. This use of the response propensities is indirect, and referred to as the *response propensity GREG-estimator*.

4 Application to POLS 2002

4.1 Introduction

The aim of the analysis is to compare the response propensity methods to the regular nonresponse adjustment method, i.e. the GREG-estimator. We use data from the Integrated Survey on Living Conditions, denoted by its Dutch acronym POLS; ‘Permanent Onderzoek LeefSituatie’. The data that we use in the analysis is obtained by aggregating the monthly POLS surveys for the year 2002. The CAPI survey sample consists of 35, 594 sample elements. For more information about POLS 2002, see Cobben (2009).

There are two types of variables that we use in the analysis: auxiliary variables and survey items. Auxiliary variables are available for both respondents and nonrespondents. These variables come from registers like the population register and the Centre for Work and Income (CWI). The survey items are the answers to the survey questions; these are only available for respondents. The auxiliary variables are: Age, marital status, ethnic group, region, degree of urbanization, household type, having a listed land-line telephone, receiving a disability insurance, receiving a social insurance, average house value, and % of non-natives. The survey items are: employment status, educational level and religion.

Schouten (2004) proposes a weighting model for POLS:

$$\begin{aligned} &Age_{15} + Housevalue_{14} + \%non - natives_8 \\ &+ Ethnicgroup_7 + Region_{15} + Householdtype_4 + Telephone_2 \end{aligned} \quad (11)$$

The subscripts denote the number of categories.

The model proposed by Schouten (2004) is based on the relationship with \mathbf{R} , but also with some of the survey items. The response propensities are calculated based on the relationship with \mathbf{R} alone, therefore the variables

in the two models will not exactly be the same. For reasons of comparison, we have performed the direct response propensity methods using the two different sets of response propensities; one based on the variables in (11) and one based on the response propensity model. The response propensity GREG-estimator is only calculated with response propensities based on the response model, because the GREG-model already uses model (11).

4.2 Response propensity model

To construct a response propensity model, we used all the available auxiliary variables. First, we determined the strength of the bivariate relationship with the response indicator. Next, we constructed a multivariate response model by starting with the most significant variable and stepwise including less significant variables until no more significant relationships with \mathbf{R} remain. Following this strategy, the final model for the response propensities becomes

$$Age_{15} + Housevalue_{14} + Degreeurbanization_5 + Maritalstatus_4 + Ethnicgroup_7 + Region_{15} + Householdtype_4 + Telephone_2 \quad (12)$$

The response model does not differ much from model (11). The variable % non-natives has been excluded from the model, and the variables degree of urbanization and marital status have been added to the model. The strongest contributions to the model come from the variables with the largest χ^2 -values, which are age, region and telephone. However, based on the values for the pseudo R^2 , the model fit in general is very low. The final model has a pseudo R^2 of 2.2%.

4.3 Methods applied to POLS 2002

Table 1 gives the response means and the adjusted estimates when applying the regular nonresponse adjustment; the GREG-estimator with weighting model (11). The fourth column then gives the estimates for the survey items when applying the response propensity GREG estimator. This estimator uses weighting model (11) as \mathbf{X} -variables; the inclusion probabilities are adjusted for the response propensities estimated by model (12). In table 2 the results from applying the direct response propensity methods are given. Both response propensity stratification and response propensity weighting are applied with weighing model (11) and response model (12).

Table 1: *Unweighted and weighted response means for POLS 2002 (in %).*

<i>Survey item</i>	<i>Response mean</i>	<i>GREG</i>	<i>Response propensity GREG</i>
<i>Employment</i>			
12 hours or more	52.4 (0.35)	53.7 (0.18)	53.8 (0.27)
Unemployed	6.7 (0.18)	6.4 (0.11)	6.4 (0.16)
Less than 12 hours	40.9 (0.35)	39.9 (0.17)	39.9 (0.26)
<i>Education</i>			
Primary	7.2 (0.18)	6.3 (0.09)	6.3 (0.13)
Junior general secondary	12.1 (0.23)	12.0 (0.15)	12.0 (0.22)
Pre-vocational	19.7 (0.28)	19.9 (0.18)	19.9 (0.28)
Senior general secondary	7.1 (0.18)	7.2 (0.12)	7.2 (0.18)
Secondary vocational	30.8 (0.32)	31.1 (0.21)	31.0 (0.32)
Higher professional	16.7 (0.26)	16.9 (0.17)	16.9 (0.26)
University	6.3 (0.17)	6.4 (0.11)	6.5 (0.17)
Other	0.2 (0.03)	0.2 (0.02)	0.2 (0.03)
<i>Religion</i>			
None	37.7 (0.34)	38.5 (0.21)	38.5 (0.33)
Roman-Catholic	33.5 (0.33)	32.4 (0.19)	32.5 (0.29)
Protestant	21.0 (0.29)	20.4 (0.17)	20.4 (0.26)
Islamic	2.5 (0.11)	3.3 (0.06)	3.2 (0.08)
Other	5.2 (0.16)	5.4 (0.01)	5.4 (0.16)

Table 2: *Direct response propensity methods applied to POLS 2002 (in %).*

<i>Survey item</i>	<i>Response propensity stratification</i>		<i>Response propensity weighting</i>	
	<i>Weighting model</i>	<i>Response model</i>	<i>Weighting model</i>	<i>Response model</i>
<i>Employment</i>				
12 hours or more	53.5 (0.38)	53.7 (0.38)	53.7 (0.34)	53.8 (0.34)
Unemployed	6.4 (0.17)	6.4 (0.17)	6.4 (0.17)	6.4 (0.18)
Less than 12 hours	40.0 (0.35)	39.9 (0.37)	39.9 (0.34)	39.8 (0.33)
<i>Education</i>				
Primary	6.3 (0.20)	6.3 (0.18)	6.3 (0.18)	6.3 (0.20)
Junior general secondary	12.0 (0.23)	12.0 (0.22)	12.0 (0.25)	12.0 (0.24)
Pre-vocational	19.9 (0.27)	19.9 (0.26)	19.9 (0.27)	19.9 (0.29)
Senior general secondary	7.2 (0.17)	7.2 (0.19)	7.2 (0.18)	7.2 (0.21)
Secondary vocational	31.0 (0.31)	31.1 (0.32)	31.1 (0.32)	31.0 (0.32)
Higher professional	16.8 (0.26)	16.9 (0.25)	16.9 (0.25)	16.9 (0.26)
University	6.4 (0.16)	6.5 (0.18)	6.4 (0.19)	6.5 (0.18)
Other	0.2 (0.03)	0.2 (0.03)	0.2 (0.03)	0.2 (0.03)
<i>Religion</i>				
None	38.3 (0.35)	38.4 (0.34)	38.4 (0.37)	38.5 (0.37)
Roman-Catholic	32.7 (0.34)	32.6 (0.35)	32.5 (0.35)	32.5 (0.35)
Protestant	20.6 (0.28)	20.5 (0.30)	20.5 (0.35)	20.4 (0.29)
Islamic	3.1 (0.18)	3.1 (0.11)	3.2 (0.12)	3.2 (0.12)
Other	5.4 (0.16)	5.3 (0.16)	5.4 (0.15)	5.4 (0.15)

The standard errors are given in parentheses².

The results for the different estimation techniques are not significantly different. Especially for the variables employment and education the estimated values based on the different methods are very similar. However, the estimated values are all quite different from the response mean. This result is especially remarkable for the response propensity stratification method, taking into account the bad fit of the response model.

5 Discussion

The two models that we used to compute response propensities are very similar. Only three out of eight variables differ between the models and the most influential variables are included in both models. Moreover, the pseudo R^2 for the response model is low, only 2.2%.

In general, the different estimation methods produce estimates for the survey items that are very much alike. Furthermore, although the results do not differ much between the methods, they are significantly different from the response mean. Also, when the same variables are used for the GREG-estimator and the propensity weighting estimator, the estimates for these two methods are similar for two out of three survey items. It seems that the set of variables that is used is of larger influence than the method.

For future research, we intend to also apply the response selection method in section 2 to the data from POLS 2002 and compare the results to the ‘regular’ GREG-estimator and the response propensity methods presented in section 3. Furthermore, it would be very informative if we could determine which method produces the best estimates. One way to determine this, is to use an auxiliary variable as survey item, so that we know the real value in the sample. Another way to achieve this would be to simulate different missing-data-mechanisms.

²For the response means, the standard errors are calculated by $\sqrt{p(1-p)/n}$, with p the percentage of the survey item category and $n = 20,168$ the number of respondents. For the GREG estimates in table 1, column 3 and 4, the standard errors are calculated by a first order Taylor approximation of the linear regression estimator. For the direct response propensity methods in table 2 a non-parametric bootstrap estimator is applied, based on 200 replications.

References

- BETHLEHEM, J. (1988): “Reduction of Nonresponse Bias through Regression Estimation,” *Journal of Official Statistics*, 4(3), 251–260.
- COBBEN, F. (2009): “How to deal with nonresponse in sample surveys. Methods for analysis and adjustment, *forthcoming*,” Ph.D. thesis, University of Amsterdam.
- GROVES, R., AND M. COUPER (1998): *Nonresponse in Household Interview Surveys*. Wiley series in probability and statistics. Survey methodology section.
- HORVITZ, D., AND D. THOMPSON (1952): “A Generalization of Sampling Without Replacement from a Finite Universe,” *Journal of the American Statistical Association*, 47, 663–685.
- LEE, S. (2006): “Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys,” *Journal of Official Statistics*, 22(2), 329–349.
- ROSENBAUM, P., AND D. RUBIN (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70(1), 41–55.
- SÄRNDAL, C. (1981): “Frameworks for Inference in Survey Sampling with Application to Small Area Estimation and Adjustment for Non-Response,” *Bulletin of the International Statistical Institute*, (49), 494 – 513.
- SCHOUTEN, B. (2004): “Adjustment for Bias in the Integrated Survey on Living Conditions (POLS) 1998,” Discussion paper 04001, Statistics Netherlands, Available at www.cbs.nl.
- STOOP, I. (2005): “The Hunt for the Last Respondent. Nonresponse in Sample Surveys,” Ph.D. thesis, University of Utrecht.
- VEN, W. V. D., AND B. V. PRAAG (1981): “The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection,” *Journal of Econometrics*, 17(2), 229 – 252.