Applied Statistics

Gaia Bertarelli, Sant'Anna School of Advanced Studies (Italy)

Statistics, Statistical learning, Computing and Data Analytics

March, 1, 2022

The presentation at a glance

Non-response or Missing data

Weighting methods for non-response

Imputation

Introduction to non-response

- The best way to deal with non-response (missing data) is to prevent it.
- Two type of non-responses:
  1. **unit non-response**: the person provides no information for the survey.
  2. **item non-response**: some information are present, the person does not respond to a particolar item.

Introduction to non-response

- Four approaches to dealing with nonresponse:
  1. Prevent it: design the survey so that the non-response is low.
  2. Take a representative subsample of the non-respondents: use the subsample to make inference about the other nonrespondents.
  3. Use a model to predict the values of the nonrespondents. Imputation often adjusts for item non-response, weighting class adjustment methods use a model to adjust for unit non-response. Parametric models may be used for unit and item non-response.
  4. Ignore the non-response (not correct even if common): non-response and undercoverage present serious problem for survey inference. The failure to obtain information from some units in the selected sample (nonresponse), or the failure to include parts of the population in the sampling frame (indercoverage) can lead to bias estimates of the population quantities.

Effects of Ignoring non-response

Moreover, increasing the sample size without targeting nonresponse does nothing to reduce nonresponse bias; a larger sample size merely provides more observations from the class of persons that would respond to the survey. Increasing the sample size may actually worsen the nonresponse bias, as the larger sample size may divert resources that could have been used to reduce or remedy the nonresponse, or it may result in less care in the data collection. Recall that the infamous

Effects of Ignoring non-response

The main problem caused by nonresponse is potential bias. Think of the population as being divided into two somewhat artificial strata of respondents and nonrespondents. The population respondents are the units that would respond if they were chosen to be in the sample; the number of population respondents, N_R, is unknown.

Effects of Ignoring non-response

Similarly, the N_M (M for missing) population nonrespondents are the units that would not respond. We then have the following population quantities:

Stratum	Size	Total	Mean	Variance
Respondents	N_R	t_R	$\overline{y}_{RU}$	S_R^2
Nonrespondents	N_M	t_M	$\overline{y}_{MU}$	S_M^2
Entire population	N	t	$\overline{y}_U$	S^2

Effects of Ignoring non-response

The population as a whole has variance $S^2 = \sum_{i=1}^{N} (y_i - \bar{y}_U)^2/(N-1)$, mean $\bar{y}_U$, and total t. A probability sample from the population will likely contain some respondents and some nonrespondents. But, of course, on the first call we do not observe y_i for any of the units in the nonrespondent stratum. If the population mean in the nonrespondent stratum differs from that in the respondent stratum, estimating the population mean using only the respondents will produce bias.[1]

Effects of Ignoring non-response

Let $\bar{y}_R$ be an approximately unbiased estimator of the mean in the respondent stratum, using only the respondents. As

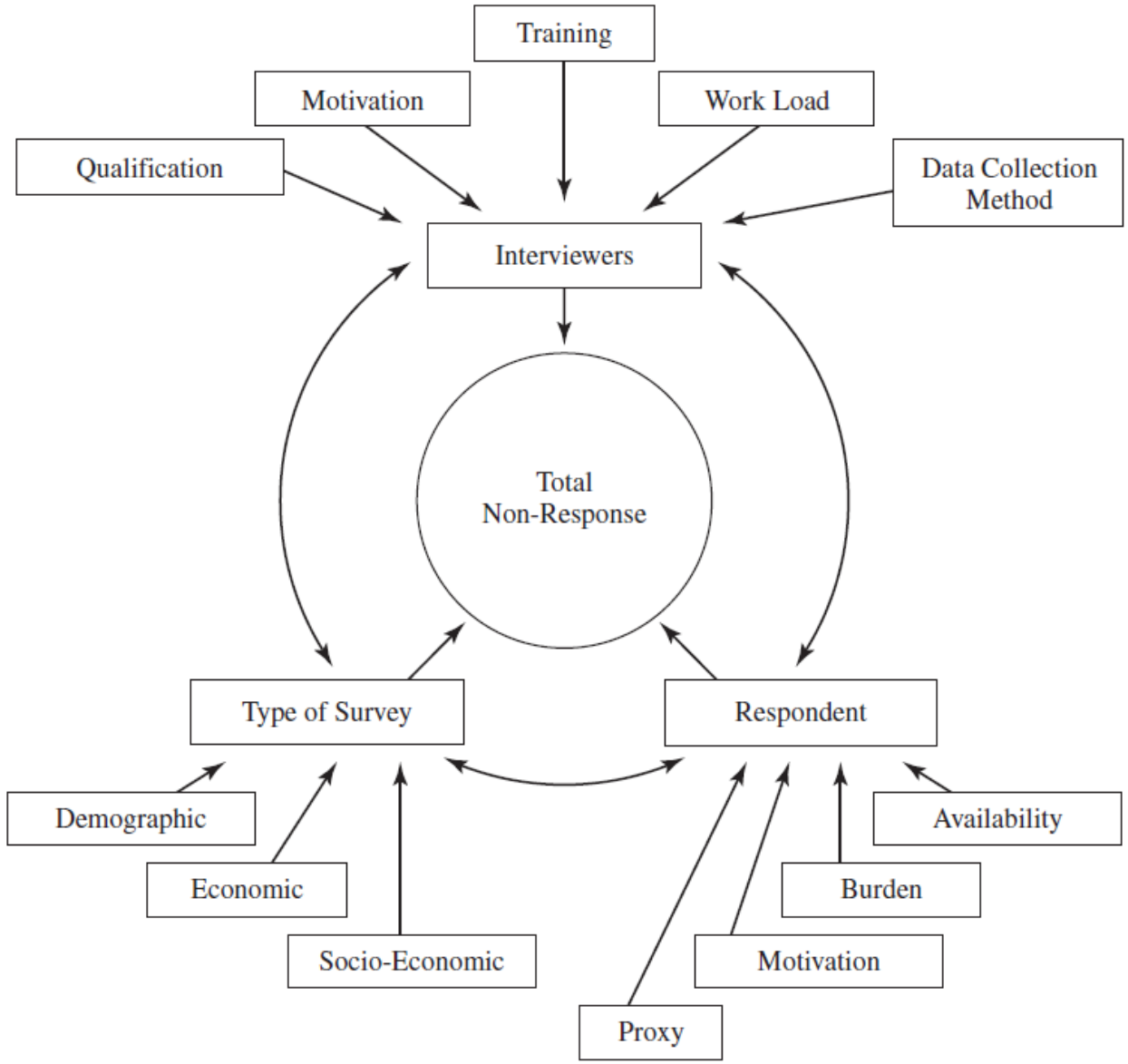
$$\bar{y}_U = \frac{N_R}{N}\bar{y}_{RU} + \frac{N_M}{N}\bar{y}_{MU},$$

the bias is approximately

$$E[\bar{y}_R] - \bar{y}_U \approx \frac{N_M}{N}(\bar{y}_{RU} - \bar{y}_{MU}).$$

The bias is small if either (1) the mean for the nonrespondents is close to the mean for the respondents, or (2) N_M/N is small—there is little nonresponse. But we can never be assured of (1), as we generally have no data for the nonrespondents. Minimizing the nonresponse rate is the only sure way to control nonresponse bias.

SOURCE: "Some Factors Affecting Non-Response," by R. Platek, 1977, *Survey Methodology, 3,* 191–214.



Mechanisms for Non-response

- Even if the survey is well defined non-responses can occur.
- The methods for try to fix the non-responses are model-based.
- If we want to make inference on non-respondents we must assume that they are related to respondents in some way.

$$R_i = \begin{cases} 1 & \text{if unit } i \text{ responds} \\ 0 & \text{if unit } i \text{ does not respond.} \end{cases}$$

- $\phi_i = P(R_i = 1)$ (unknow and assumed positive) is the probability that a unit selected for the sample will respond (propensity score of unit i.

Mechanisms for Non-response

- **Missing completely at random**(MCAR): Missing data are missing completely at random if ϕ_i does not depend on $\mathbf{x}_i, y_i$ or the survey design.
- **Missing at random given covariates** (MAR): If ϕ_i depends on $\mathbf{x}_i$ but not on y_i. The nonresponse depends only on observed variables.
- **Not missing at random** (NMAR): If the probability of nonresponse depends on the value of a missing response variable and cannot be completely explained by values in the observed data.

Weighting methods for non-response

- Weights can be used to adjust for non-response.
- Z_i is the indicator variable for presence in the selected sample:

$$P(Z_i = 1)\pi_i$$

- If R_i indep. of Z_i

$$P(\text{ unit i selected in sample and responds}) = \pi_i\phi_i$$

where $\phi_i = P(R_i = 1)$.

Weighting methods for non-response

- The probability of responding ϕ_i is estimated for each unit in the sample using auxiliary information that is known for all units in the selected sample.
- The final weight for a respondent is

$$\frac{1}{\pi_i \hat{\phi}_i}$$

- **Weighting methods assume that the response probabilities can be estimated from variables known for all units.**
- Weighting methods assume MAR data.

Weighting Class Adjustment

- w_i: number of units in the population represented by unit i in the sample.
- Weighting methods extend this approach to compensate for nonsampling errors.
- Variables known for all units in the selected sample are use to form weighting adjustment classes and it is supposed that respondents and nonrespondents in the same class are similar.
- Weights are increased so that the respondents represent the nonrespondents' share of the population as well as their own.

Example 8.4 - Lohr

Suppose the age is known for every member of the selected sample and that person i in the selected sample has sampling weight $w_i = 1/\pi_i$. Then weighting classes can be formed by dividing the selected sample among different age classes, as Table 8.2 shows.

We estimate the response probability for each class by

$$\hat{\phi}_c = \frac{\text{sum of weights for respondents in class } c}{\text{sum of weights for selected sample in class } c}.$$

Then the sampling weight for each respondent in class c is multiplied by $1/\hat{\phi}_c$, the weight factor in Table 8.2. The weight of each respondent with age between 15 and 24, for example, is multiplied by 1.622. Since there was no nonresponse in the over-65 group, their weights are unchanged. ■

Example 8.4 - Lohr

TABLE 8.2
Illustration of Weighting Class Adjustment Factors

	Age					
	15–24	25–34	35–44	45–64	65+	Total
Sample size	202	220	180	195	203	1000
Respondents	124	187	162	187	203	863
Sum of weights for sample	30,322	33,013	27,046	29,272	30,451	150,104
Sum of weights for respondents	18,693	28,143	24,371	28,138	30,451	
$\hat{\phi}_c$	0.6165	0.8525	0.9011	0.9613	1.0000	
Weight factor	1.622	1.173	1.110	1.040	1.000	

- The probability of response is assumed to be the same in each weighting class.
- The probability of response does not depend on y (MAR).
- As already said, the weight of respondent in weighting class c is

$$\frac{1}{\pi_i \hat{\phi}_c}$$

- To estimate the population total:

$$\hat{w}_i = w_i \sum_c \frac{x_{ci}}{\hat{\phi}_c}$$

where $x_{ci} = 1$ if unit i is in class c, 0 otherwise, w_i is the sampling weight for unit i.

- Population Total

$$\hat{t}_{wc} = \sum_{i \in S} \hat{w}_i y_i$$

- Population Mean

$$\hat{\bar{y}}_{wc} = \frac{\hat{t}_{wc}}{\sum_{i \in S} \hat{w}_i}$$

Construction of Weighting Class

Construction of Weighting Classes Weighting adjustment classes should be constructed as though they were strata; as shown in the next section, weighting adjustment is similar to poststratification. If we could construct weighting classes so that in each weighting class c (a) the response variable y_i is constant in class c, (b) the response propensity ϕ_i is the same for every unit in class c, or (c) the response y_i is uncorrelated with the response propensity ϕ_i in class c, then we would largely eliminate nonresponse bias for estimating population means and totals (see Exercise 17).

Consequently, the weighting classes should be formed so that units within each class are as similar as possible with respect to the major variables of interest, and so that the response propensities vary from class to class but are relatively homogeneous within a class. At the same time, it is desirable to avoid very large weight adjustments. Eltinge and Yansaneh (1997) discuss methods for choosing the number of weighting classes to use.

Post-Stratification

Poststratification was introduced in Section 4.4; it is a form of ratio adjustment. To use poststratification to try to compensate for nonresponse, we modify the weights so that the sample is calibrated to population counts in the poststrata. Poststratification is similar to weighting class adjustment, except that population counts are used to adjust the weights. Suppose an SRS is taken. After the sample is collected, units are grouped into H different poststrata, usually based on demographic variables such as race or sex. The population has N_h units in poststratum h; of these, n_h were selected for the sample and n_{hR} responded. The poststratified estimator for $\bar{y}_U$ is

$$\bar{y}_{\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{hR};$$

the weighting class estimator for $\bar{y}_U$ is

$$\bar{y}_{wc} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_{hR}.$$

Post-Stratification

The two estimators are similar in form; the only difference is that in poststratification, the N_h are known while in weighting class adjustments the N_h are unknown and estimated by Nn_h/n. A variance estimator for poststratification will be given in Exercise 17 of Chapter 9.

Post-Stratification

Poststratification Using Weights

In a general survey design, the sum of the weights in a subgroup, $\sum_{i \in \mathcal{S}_h} w_i$, is supposed to estimate the population count for that subgroup, N_h. Poststratification uses the ratio estimator within each subgroup to adjust by the true population count.

Let $x_{hi} = 1$ if unit i is a respondent in poststratum h, and 0 otherwise. Then let

$$w_i^* = w_i \sum_{h=1}^{H} x_{hi} \frac{N_h}{\sum_{j \in \mathcal{R}} w_j x_{hj}},$$

where $\mathcal{R}$ is the set of respondents in the sample. Using the modified weights,

$$\sum_{i \in \mathcal{R}} w_i^* x_{hi} = N_h,$$

and the poststratified estimator of the population total is

$$\hat{t}_{\text{post}} = \sum_{i \in \mathcal{R}} w_i^* y_i.$$

Note that the modified weights w_i^* depend on the particular sample selected.

Poststratification adjusts for undercoverage as well as nonresponse if the population count N_h includes individuals not in the sampling frame for the survey. As shown in Chapter 4, poststratification can reduce the variance of estimated population quantities by calibrating the survey to the known population counts.

Pros and Cons

Weighting class adjustments and poststratification can both help reduce nonresponse bias. The models for weighting adjustments for nonresponse are strong: In each weighting cell or poststratum, the respondents and nonrespondents are assumed to be similar, or each individual in a weighting class is assumed equally likely to respond to the survey or have a response propensity that is uncorrelated with y. These models never exactly describe the true state of affairs, and you should always consider their plausibility and implications. It is an unfortunate tendency of some survey practitioners to treat the weighting adjustment as a complete remedy and then act as though there was no nonresponse. Weights may improve many of the estimates, but they rarely eliminate all nonresponse bias. If weighting adjustments are made (and remember,

- Weighting adjustments are usually used for unit non-response, not for item non-response.

Imputation

Missing items may occur in surveys for several reasons: An interviewer may fail to ask a question; a respondent may refuse to answer the question or cannot provide the information; a clerk entering the data may skip the value. Sometimes, items with responses are changed to missing when the data set is edited or cleaned—a data editor may not be able to resolve the discrepancies for an individual 3-year-old who voted in the last election, and may set both values to missing.

Imputation is commonly used to assign values to the missing items. A replacement value, often from another person in the survey who is similar to the item nonrespondent on other variables, is imputed (filled in) for the missing value. When imputation is used, an additional variable should be created for the data set that indicates whether the response was measured or imputed.

Imputation

1. Deductive Imputation
2. Cell Mean Imputation
3. Hot deck Imputation
4. Regression Imputation
5. Cold-Deck Imputation
6. Multiple Imputation

Deductive Imputation

Some values may be imputed in the data editing, using logical relations among the variables. Person 9 is missing the response for whether she was a victim of violent crime. But she had responded that she was not a victim of any crime, so the violent crime response should be changed to 0.

Deductive imputation may sometimes be used in longitudinal surveys. If a woman has two children in year 1 and two children in year 3, but is missing the value for year 2, the logical value to impute would be two.

Cell Mean Imputation

- Similar as in weighting class methods: respondents are divided in class based of their known variables.
- The average of the values for the responding units in cell c, $\hat{y}_{Rc}$ is substituted for each missing value.
- **Cell eman imputation assumes that missing items are MCAR within the cells.**

TABLE 8.3
Small Data Set Used to Illustrate Imputation Methods

Person	Age	Sex	Years of Education	Crime Victim?	Violent Crime Victim?
1	47	M	16	0	0
2	45	F	?	1	1
3	19	M	11	0	0
4	21	F	?	1	1
5	24	M	12	1	1
6	41	F	?	0	0
7	36	M	20	1	?
8	50	M	12	0	0
9	53	F	13	0	?
10	17	M	10	?	?
11	53	F	12	0	0
12	21	F	12	0	0
13	18	F	11	1	?
14	34	M	16	1	0
15	44	M	14	0	0
16	45	M	11	0	0
17	54	F	14	0	0
18	55	F	10	0	0
19	29	F	12	?	0
20	32	F	10	0	0

The four cells for our example are constructed using the variables age and sex. (In practice, of course, you would want to have many more individuals in each cell.)

		Age	
		≤ 34	≥ 35
Sex	M	Persons 3, 5, 10, 14	Persons 1, 7, 8, 15, 16
	F	Persons 4, 12, 13, 19, 20	Persons 2, 6, 9, 11, 17, 18

Persons 2 and 6, missing the value for years of education, would be assigned the mean value for the four women aged 35 or older who responded to the question: 12.25. The mean for each cell after imputation is the same as the mean of the respondents. The imputed value, however, is not one of the possible responses to the question about education. ■

Cell Mean Imputation

- Mean imputation method fails to reflect the variability of the non-respondents.
- The distribution of y will be distorted because of a spike at the value of the sample mean of the respondents $\rightarrow$ estimation variance too small.
- It is possible to adopt a Stochastic cell mean imputation: if the response variables were approx. normally distributed the missing values can be imputed randomly from a normal distribution with mean equal to $\bar{y}_{cR}$ and std.dev s_{cR}.
- **Mean imputation, sthocastic imputation, ..., distorts relationships among different variables because imputation is done separately for each missing item.**

Hot-Deck Imputation

In *hot-deck imputation*, as in cell mean imputation and weighting adjustment methods, the sample units are divided into classes. The value of one of the responding units in the class is substituted for each missing response. Often, the values for a set of related missing items are taken from the same donor, to preserve some of the multivariate relationships. The name *hot deck* is from the days when computer programs and data sets were punched on cards—the deck of cards containing the data set being analyzed was warmed by the card reader, so the term *hot deck* was used to refer to imputations made using the same data set. Fellegi and Holt (1976) discuss methods for data editing and hot-deck imputation with large surveys.

Random Hot-Deck Imputation A donor is randomly chosen from the persons in the cell with information on all the missing items. To preserve multivariate relationships, usually values from the same donor are used for all missing items of a person.

In our small data set, person 10 is missing both variables for victimization. Persons 3, 5, and 14 in his cell have responses for both crime questions, so one of the three is chosen randomly as the donor. In this case, person 14 is chosen, and his values are imputed for both missing variables.

Nearest-Neighbor Hot-Deck Imputation Define a distance measure between observations, and impute the value of a respondent who is "closest" to the person with the missing item, where closeness is defined using the distance function.

If age and sex are used for the distance function, so that the person of closest age with the same sex is selected to be the donor, the victimization responses of person 3 will be imputed for person 10.

Selection bias still presents.

Regression Imputation

Regression imputation predicts the missing value using a regression of the item of interest on variables observed for all cases. A variation is *stochastic regression imputation*, in which the missing value is replaced by the predicted value from the regression model plus a randomly generated error term.

We only have 18 complete observations for the response crime victimization (not really enough for fitting a model to our data set), but a logistic regression of the response with explanatory variable age gives the following model for predicted probability of victimization, $\hat{p}$:

$$\log \frac{\hat{p}}{1-\hat{p}} = 2.5643 - 0.0896 \times \text{ age.}$$

The predicted probability of being a crime victim for a 17-year-old is 0.74; because that is greater than a predetermined cutoff of 0.5, the value 1 is imputed for Person 10.

Pros & Cons

Imputation creates a "clean," rectangular data set that can be analyzed by standard software. Analyses of different subsets of the data will produce consistent results. If the nonresponse is missing at random given the covariates used in the imputation procedure, imputation substantially reduces the bias due to item nonresponse. If parts of the data are confidential, the collector of the data can perform the imputation. The data collector has more information about the sample and population than is released to the public (for example, the collector may know the exact address for each sample member), and can often perform a better imputation using that information.

The foremost danger of using imputation is that future data analysts will not distinguish between the original and the imputed values. Ideally, the imputer should record which observations are imputed, how many times each nonimputed record is used as a donor, and which donor was used for a specific response imputed to a recipient. The imputed values may be good guesses, but they are not real data.

What is an acceptable response rate?

Often an investigator will say, "I expect to get a 60% response rate in my survey. Is that acceptable, and will the survey give me valid results?" As we have seen in this chapter, the answer to that question depends on the nature of the nonresponse: If the nonrespondents are MCAR, then we can largely ignore the nonresponse and use the respondents as a representative sample of the population. If the nonrespondents tend to differ from the respondents, then the biases in the results from using only the respondents may make the entire survey worthless.

Nonresponse and undercoverage present serious problems for survey inference. The main concern is that failure to obtain information from some units in the selected sample (nonresponse), or failure to include parts of the population in the sampling frame (undercoverage), can result in biased estimates of population quantities.

The survey design should include features to minimize nonresponse. Designed experiments can give insight into methods for increasing response rates. If possible, the survey frame should contain some information on everyone in the selected sample so that respondents and nonrespondents can be compared on those variables, and so that the auxiliary information can be used in adjusting for residual nonresponse.

Weighting adjustment methods and models can be used to try to reduce nonresponse bias. In weighting class methods, the weights of respondents in a grouping class are increased to compensate for the nonrespondents in that grouping class. In poststratification, the weights of respondents in a poststratum are increased so that they sum to an independent count of the population in that poststratum. The nonresponse mechanism can also be modeled explicitly.

Imputation methods create a "complete" data set by filling in values for data that are missing because of item nonresponse. You must be careful when analyzing imputed data sets to account for the imputation when estimating variances since the imputed values are usually derived from the data.

All surveys should report nonresponse rates. If imputation is used, the imputed values should be flagged so that data analysts know which values were observed and which values were imputed.

Gaia Bertarelli

Department of Economics and Management in the Era of Data Science

gaia.bertarelli@santannapisa.it