Slides for project B: Treatment of Nonresponse in Surveys

David Haziza
Département de mathématiques et de statistique
Université de Montréal

adapted by Anne Ruiz-Gazen
for M2 Statistics and Econometrics
TSE-UT1C / UT3

November 22, 2024

Definitions

- Nonresponse: failure to obtain usable responses from sampled units
- Response rates have been steadily decreasing
- Concern: nonresponse can significantly affect the quality of survey statistics
- Nonresponse occurs in sample surveys, censuses and administrative data

Levels of nonresponse

- We distinguish between two types of nonresponse:

  1. Unit nonresponse:
     - Complete lack of information on a given unit.
  2. Item nonresponse:
     - Some (but not all) variables are observed.

Levels of nonresponse

$$
\begin{array}{cc}
 & \begin{array}{cccccccc} y_1 & y_2 & y_3 & y_4 & \cdots & \cdots & & ..y_p \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ n \end{array} &
\left[\begin{array}{cccccccc}
\surd & \surd & \surd & \surd & \surd & \surd & \surd & \surd \\
\surd & \surd & \surd & \surd & \surd & \surd & \surd & \surd \\
\surd & \surd & \surd & \surd & \surd & \surd & \surd & \surd \\
\text{x} & \surd & \text{x} & \surd & \surd & \surd & \text{x} & \text{x} \\
\surd & \text{x} & \surd & \text{x} & \text{x} & \text{x} & \surd & \surd \\
\text{x} & \text{x} & \surd & \surd & \surd & \surd & \surd & \surd \\
\text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} \\
\text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x} & \text{x}
\end{array}\right]
\begin{array}{l}
\left.\begin{array}{l} \vphantom{\surd} \\ \vphantom{\surd} \\ \vphantom{\surd} \end{array}\right\} \text{Full response} \\
\left.\begin{array}{l} \vphantom{\surd} \\ \vphantom{\surd} \\ \vphantom{\surd} \end{array}\right\} \text{Item nonresponse} \\
\left.\begin{array}{l} \vphantom{\surd} \\ \vphantom{\surd} \end{array}\right\} \text{Unit nonresponse}
\end{array}
\end{array}
$$

Methods of treatment

- Treatment of unit nonresponse: typically, weight adjustment procedures methods are used. They consist of:
  - eliminating the nonrespondents from the data file
  - increasing the sampling weight of the respondents to compensate for the elimination of the nonrespondents
- Item nonresponse: typically, single imputation is used. It consists of creating a single 'artificial' value to fill in the hole.

Effects of nonresponse

- Nonresponse bias: due to the fact that respondents and nonrespondents do not have the same characteristics with respect to the variables of interest.
- Sample size is smaller than expected ⇒ Variance of estimators is generally greater than that of estimators that would have been obtained if there were no nonresponse. This additional variability is called the nonresponse variance.
- The main objective of all the methods (weighing procedures or imputation) is to reduce the nonresponse bias and possibly control the variance due to nonresponse.

Reasons for unit nonresponse

- Reasons for unit nonresponse may be classified into two broad categories:
  - accessibility issues: refers to the ability to make contact with the sampled unit. For example, accessibility issues may be the inability to find anyone at home for an in-person or telephone survey, or the inability to trace a unit successfully from a list.
  - amenability issues: refers to the unit's willingness to cooperate with the survey request after contact has been made.

Factors affecting accessibility and amenability

- Mode of data collection
  - Important factor related to amenability, especially in household surveys:
    - Face-to-face surveys result in the highest response rates
    - Telephone survey response rates are usually lower than in-person response rates
    - Mail surveys generally have the lowest response rates
  - Also related to accessibility:
    - Households without telephones are not accessible in telephone surveys.

Factors affecting accessibility and amenability

- Type of sampled unit
  - Establishment surveys generally have lower response rates than surveys of individuals
- New technologies
  - In telephone surveys, answering machines and caller ID devices are technological barriers to obtaining high response rates.
  - Studies have found that households with answering machines or caller ID devices were harder to reach, but once contacted, they were as likely to respond as those without answering machines.
- Importance of the topic of the survey to the sampled unit
  - People who are interested in or have characteristics related to the survey topic may be more likely to cooperate.

Factors affecting accessibility and amenability

- Type of survey: compulsory vs. optional.
- Reputation of the agency in charge of the survey.
- Experience, aptitudes and workload of the staff and interviewers in charge of data collection has an impact on amenability.

Reasons for item nonresponse

- The respondents may not understand the question, resulting in a missing value.
- The respondents may not recall the relevant information.
- The respondent may refuse to answer to a sensitive question.
- A question may be skipped.
- Poorly constructed questionnaires.
- Answer inconsistent or unusable: edit rules not satisfied.
- see de Leeuw, Hox and Husman (2003)

Reducing unit nonresponse

- Number and timing of attempts to access the sample person
  - Vary visiting (or calling) hours
  - Nights and weekends normally lead to higher response rates.
- Data collection period
  - Length of the collection period should adequate.
  - In surveys using interviews in person or telephone interviews, 1-2 weeks is often appropriate.
  - Mail surveys need longer data collection periods.
- Interviewer workload
  - The number of cases assigned to an interviewer should be reasonable.

Reducing unit nonresponse

- Incentives
  - Tend to increase cooperation rates.
  - Cash incentives tend to be more powerful than other types of incentives
- Respondent burden
  - The length of interviews should be limited.
  - Repeated/supplementary surveys can discourage participation.
- Prenotification
  - Advance letter can generate higher response rates.
  - Standard practice in statistical agencies.
- Persuasion letter
  - Send person who initially refused to be interviewed a letter reenforcing the serious purpose of the survey.

Reducing unit nonresponse

- Interviewers
    - Adequate and ongoing interviewer training program.
    - Introductory behaviour: the first few seconds in telephone surveys can affect cooperation rates.
    - Matching interviewers with sample persons may increase cooperation.
    - Switching the interviewer: in case of a refusal, replace the initial interviewer with another interviewer whose characteristics are more acceptable to the sample person
- Mode switches
    - Many surveys start with a cheap method of data collection (e.g., mail questionnaire) and then use more expensive modes (telephone of in person interviews) for nonrespondents of the first mode.

Reducing item nonresponse

- Questionnaire
  - Short, simple and easily understood questions.
  - Avoid open questions.
  - Avoid abbreviations.
  - Good translation.
  - Minimal instructions on the questionnaire.
  - Take into account the data collection method.
  - The questionnaire should be developed by all parties (subject matter people, methodologists, collection experts, computer scientists,respondents, data users, etc.)

Reducing item nonresponse

Questionnaire

- The presentation and general look of the questionnaire (for mail surveys) has an impact on the response rate.
- Assess previous surveys.
- Pilot surveys and tests.
- Apply cognitive research (discussions groups, etc.).
- In case of sensitive questions:
  - Explain reasons for the question.
  - Develop fairly broad response categories.
  - Reassure respondent with respect to confidentiality of information.

Auxiliary information

- In the absence of nonresponse, survey statisticians try to avoid using estimation procedures whose validity depends on the validity of a given model (as it is typically the case in classical statistics)
- The use of models is unavoidable in the presence of nonresponse. In this case, the quality of statistics depends on the quality of the assumed model, which in turns, depends on the availability of auxiliary information.
- Auxiliary information plays an important role in surveys because it allows for the use of more efficient sampling and estimation procedures. It is also used for reducing coverage and nonresponse errors.

Auxiliary information

- Several types of auxiliary variables:
  - The design variables: they are used at the design stage and hence, must be available for all the population units prior to sampling. They are typically used to stratify the population or use some form of proportional-to-size sampling designs.
  - The calibration (or benchmark) variables: they are used are the estimation stage. They are minimally required for all the sampled units and the corresponding population totals must be known (e.g., census counts)
  - The variables used in the nonresponse treatment procedures (weighting and imputation). Minimally, they need to be known for all the sampled units.

Choosing the auxiliary variables

- A given auxiliary variable may be used at more than one stage.
- To reduce the nonresponse bias effectively, we need to choose the auxiliary variables that:
  - are related to the variables of interest

  and

  - are related to the probability of response

Examples of auxiliary variables

- Auxiliary variables that are often used for nonresponse treatment:
  - Geographical information (typical in household surveys)
  - Variables on the sampling frame: for example, in business surveys, variables such as Province, NAICS and Size of the business are available
  - Tax data (mainly used in business surveys)
  - Demographic data (e.g., age, sex)
  - Paradata: data about the survey process

  Examples of paradata:
  - Interviewer call record data
    - Date and time of call, time between calls, contact strategy used, outcome of call, language of the interview
  - Interviewer observations about each household
    - Type of accommodation, physical barriers, security device, indications if children present, observations about neighbourhood

Nonresponse mechanism

- Let r_i be the response indicator attached to unit i such that $r_i = 1$ if unit i is a respondent and $r_i = 0$, otherwise.
- In the case of unit nonresponse $r_i = 1$ if unit i is a respondent to the survey
- In the case of item nonresponse, $r_i = 1$ if unit i responded to a given item y
- Let p_i be the response probability for unit i; that is, $p_i = P(r_i = 1 \mid s;\ i \in s)$.

Nonresponse mechanism

- We assume that
$$0 < p_i \leqslant 1$$
for all i.
- We assume that $p_i > 0$ for all i, which may not be realistic in most surveys because a fraction of sampled units are hardcore nonrespondents (Kott, 1994).
- We also assume that the units respond independently of one another so
$$p_{ij} = P(r_i = 1,\ r_j = 1 \mid s,\ i \in s,\ j \in s,\ i \neq j) = p_i p_j$$
- The nonresponse mechanism, $q(s_r \mid s)$, may be described as follows:
$$r_i \overset{\text{ind.}}{\sim} B(1, p_i).$$
- This distribution is unknown since the p_i's are unknown.

Nonresponse mechanism

- If the probability of response is the same for all the population units, the nonresponse mechanism is said to be uniform.
- Typically, the probability of response varies from one unit to another. It may depend on age, sex, income, etc.
- The probability of response may depend on auxiliary variables (that are observed for both the respondents and the nonrespondents) but it may also depend on a variable of interest (which is only observed for the respondents).

Uniform Nonresponse Mechanism

- A nonresponse mechanism is said to be uniform if $p_i = p$ for all the units in the population.
- When the nonresponse mechanism is uniform, the data are said to be Missing Completely at Random (MCAR).
- The data are said to be MCAR if the response probability does not depend on any variable (variables of interest or auxiliary variable), which is the case for a uniform nonresponse mechanism.
- A uniform nonresponse mechanism is unrealistic in practice. However, it is customary to assume a uniform nonresponse mechanism within classes.

Ignorable vs. Nonignorable Mechanism

- A nonresponse mechanism is said to be ignorable if

$$P(r_i = 1 \mid \mathbf{y}, \mathbf{z}) = P(r_i = 1 \mid \mathbf{z})$$

- In other words, a nonresponse mechanism is ignorable if, after having taken all the appropriate auxiliary information, **z**, into account, the response probability does not depend on the variables of interest.
- When a nonresponse mechanism is ignorable, the distribution of the *y*-variable among the respondents is the same as the distribution of the *y*-variable among the nonrespondents, after conditioning on **z**.
- When the reponse mechanism is ignorable, the data are said to be Missing At Random (MAR).

Ignorable vs. Nonignorable Mechanism

- A nonresponse mechanism which is not ignorable is called nonignorable.
- When the reponse mechanism is nonignorable, the data is said to be Not Missing At Random (NMAR).
- In general, one does not know if the ignorability assumption holds. Only in some situations, it is known that the nonresponse mechanism is ignorable. This type of situations occurs in the case of planned nonresponse. For example, split questionnaires: the questionnaire is divided in sections. Sets of sections are randomly assigned to the sampled units.

Ignorable vs. Nonignorable Mechanism: Illustration

- Suppose that the probability of response is related to a variable z_1 and that the y-variable is related to z_2. In addition, we know that z_2 is not related to y. In this case the data are MCAR since the response probability and the y variable are unrelated.
- Suppose that a common variable z is related to both the probability of response and the y-variable. It follows that the y-variable and the response probability are related through a common link to z. However, if we account for z at the nonresponse treatment stage, the data are MAR since there will no residual relationship between the probability of response and the y-variable after accounting for z.

Ignorable vs. Nonignorable Mechanism: Illustration

- Suppose that there is a direct relation between the response probability and the y-variable. That is, the y-variable is itself the cause of nonesponse. In this case, the data are NMAR. While the use of an appropriate set of auxiliary variables will help in weakening the relationship between y and the response probability, it won't completely eliminate it.
- When the data ar MAR, there would be no nonresponse bias. When the data are NMAR, estimators will remain biased even after accounting for the appropriate auxiliary information but we expect to achieve a good bias reduction if the auxiliary variables are highly related to the y-variable.

Total error

- We want to estimate the population total, $Y = \sum_{i \in U} y_i$.
- We select a sample, s, of size n, according to a given sampling design $p(s)$.
- We observe some amount of nonresponse (unit or item nonresponse)
- Let $\widehat{Y}_{NR}$ be the estimator after treatment (e.g., weight adjustment procedure, imputation, etc.)
- The total error of $\widehat{Y}_{NR}$, $\widehat{Y}_{NR} - Y$, can be decomposed as:

$$\widehat{Y}_{NR} - Y = \left(\widehat{Y}_F - Y\right) + \left(\widehat{Y}_{NR} - \widehat{Y}_F\right), \tag{7.1}$$

where $\widehat{Y}_F$ denotes the complete data or (full sample) estimator.

Total error

- The complete data estimator, $\widehat{Y}_F$, is generally chosen so it is
  - p-unbiased; e.g., the expansion estimator, $\widehat{Y}_F = \widehat{Y}_\pi$.
  - asymptotically p-unbiased; e.g., a calibration estimator, $\widehat{Y}_F = \widehat{Y}_C$.
- The term $\widehat{Y}_F - Y$ denotes the sampling error.
- The term $\widehat{Y}_{NR} - \widehat{Y}_F$ denotes the nonresponse error.
- To assess the properties of the estimators after treatment, different approaches (or modes) of inference can be used.

Nonresponse bias

- Using the decomposition (7.1), the bias of $\widehat{Y}_{NR}$ under the NM approach can be written as

$$\begin{aligned} Bias\left(\widehat{Y}_{NR}\right) \equiv E\left(\widehat{Y}_{NR} - Y\right) &= E_pE_q\left(\widehat{Y}_{NR} - Y \mid s\right) \\ &= E_p\left(\widehat{Y}_F - Y\right) + E_pE_q\left(\widehat{Y}_{NR} - \widehat{Y}_F \mid s\right) \\ &= E_pE_q\left(\widehat{Y}_{NR} - \widehat{Y}_F \mid s\right) \\ &= E_p\left(B_q\right), \end{aligned}$$

  where $B_q = E_q\left(\widehat{Y}_{NR} - \widehat{Y}_F \mid s\right)$ is the conditional nonresponse bias.

- Here, $E_q(.)$ denotes the expectation with respect to the nonresponse model.

Nonresponse bias

- The estimator $\widehat{Y}_{NR}$ will be asymptotically unbiased if $B_q \approx 0$. for all samples s. In this case, we say that $\widehat{Y}_{NR}$ is asymptotically *pq*-unbiased.

- In practice, it is generally not possible to determine if the nonresponse bias is zero. As a result, in practice, it is customary to assume that the bias is small. But this assumption is justified only if all the appropriate auxiliary information was used for treating nonresponse.

Nonresponse variance

- Under the nonresponse model approach, assuming that $B_q = 0$, it follows from (7.1) that the total variance of $\widehat{Y}_{NR}$ can be expressed as:

$$\begin{aligned} V\left(\widehat{Y}_{NR}\right) &= E\left(\widehat{Y}_{NR} - Y\right)^2 = E_p E_q \left(\widehat{Y}_{NR} - Y\right)^2 \\ &= E_p E_q \left(\widehat{Y}_F - Y\right)^2 + E_p E_q \left(\widehat{Y}_{NR} - \widehat{Y}_F\right)^2 \\ &= V_p\left(\widehat{Y}_F\right) + E_p V_q\left(\widehat{Y}_{NR} \mid s\right) \\ &= V_{SAM} + V_{NR}. \end{aligned} \tag{7.2}$$

- The term V_{SAM} denotes the sampling variance.
- The term V_{NR} denotes the nonresponse variance.

The Mean of the respondents

- The mean of the respondents, often called unadjusted estimator, is given by:

$$\widehat{\overline{Y}}_r = \frac{\sum_{i \in s_r} d_i y_i}{\sum_{i \in s_r} d_i}.$$

- Using a first-order Taylor expansion, the nonresponse bias of $\widehat{\overline{Y}}_r$ can be approximated by:

$$Bias\left(\widehat{\overline{Y}}_r\right) \equiv E_p E_q \left(\widehat{\overline{Y}}_r - \widehat{\overline{Y}}_\pi \mid s\right) \approx \frac{1}{\overline{P}} \sum_{i \in U} \left(p_i - \overline{P}\right)\left(y_i - \overline{Y}\right) \tag{7.3}$$

where $\overline{P} = N^{-1} \sum_{i \in U} p_i$.

The Mean of the respondents

- The asymptotic relative bias of $\widehat{\overline{Y}}_r$ can be expressed as:

$$RB\left(\widehat{\overline{Y}}_r\right) = \frac{Bias\left(\widehat{\overline{Y}}_r\right)}{\overline{Y}} \approx \rho_{py} CV(p) CV(y), \qquad (7.4)$$

  where $\rho_{py} = \frac{\sum_{i \in U}(p_i - \overline{P})(y_i - \overline{Y})}{S_p S_y}$, $CV(p) = S_p/\overline{P}$ and $CV(y) = S_y/\overline{Y}$.

- Noting that $|\rho_{py}| \leq 1$, we obtain an expression of the maximal relative bias:

$$|RB\left(\widehat{\overline{Y}}_r\right)| \leq CV(p)CV(y).$$

The Mean of the respondents

- The nonresponse bias of $\widehat{\overline{Y}}_r$ in (7.4) vanishes if the probability of response is unrelated to the y-variable (which corresponds to MCAR) or if $CV(p) = 0$ (which corresponds to uniform nonresponse), assuming $CV(y) > 0$.
- Absence or presence of bias in some situations:
  - Suppose that the probability of response is related to a z-variable but y and z are unrelated. In this case, we have $\rho_{py} = 0$ and the unadjusted estimator is unbiased.
  - Suppose that y and z are related but the probability of response is not related to z. In this case, we have $\rho_{py} = 0$ and the unadjusted estimator is unbiased.
  - Suppose that the z-variable is related to both the y-variable and the response probability. In this case, the unadjusted estimator is biased and the bias is large if ρ_{py} is large and/or $CV(p)$ is large.

Weighting procedures for unit nonresponse

- In this chapter, we consider the commonly used weighting procedure to treat unit nonresponse: weighting using the inverse of the estimated response probability:

$$\underbrace{d_i}_{\text{design weights}} \longrightarrow \underbrace{w_i^*}_{\text{weights adjusted for nonresponse}} \longrightarrow \underbrace{w_i}_{\text{Calibrated (or final) weights}}$$

Weighting by the inverse of the estimated response probability

- The idea behind this type of weighting procedure is the following.
- Let p_i be the response probability to the survey associated with unit i.
- For now, let us assume that the p_i's are known. In this case, an unbiased estimator (called Propensity Score Adjusted estimator) of a population total, $Y = \sum\limits_{i \in U} y_i$ would be:

$$\tilde{Y}_{PSA} = \sum_{i \in s_r} \frac{d_i}{p_i} y_i = \sum_{i \in s_r} \frac{1}{\pi_i p_i} y_i. \tag{7.5}$$

Weighting by the inverse of the estimated response probability

- In practice, the p_i's are unknown ⇒ they need to be estimated.
- Suppose that the response probabilities can be parametrically modeled. That is,
$$p_i = f(\mathbf{z}_i, \gamma), \tag{7.6}$$
for some known function $f(\mathbf{z}_i, .)$, where $\mathbf{z}_i$ is a vector of auxiliary variables available for both respondents and nonrespondents and γ is a vector of unknown parameters.
- The model (7.6) is the so-called nonresponse model.
- Let $\widehat{p}_i \equiv f(\mathbf{z}_i, \widehat{\gamma})$ be an estimate of p_i, where $\widehat{\gamma}$ is an estimate of γ (for example, the maximum likelihood estimator).

Weighting by the inverse of the estimated response probability

- We obtain a weighting system adjusted for nonresponse $\{w_i^*; \in s_r\}$, where

$$w_i^* = \frac{d_i}{\widehat{p}_i}.$$

- The factor $\widehat{p_i}^{-1}$ is often called the nonresponse adjustment factor attached to unit i.
- A Propensity Score Adjusted Estimator (PSA) estimator of Y is then given by

$$\widehat{Y}_{PSA} = \sum_{i \in s_r} w_i^* y_i = \sum_{i \in s_r} \frac{d_i}{\widehat{p}_i} y_i. \qquad (7.7)$$

Weighting by the inverse of the estimated response probability

- If $\widehat{p}_i = p_i$ for all i, $\widehat{Y}_{PSA}$ is unbiased for Y. The nonresponse bias is given by

$$Bias\left(\widehat{Y}_{PSA} \mid s\right) = E_q\left(\widehat{Y}_{PSA} - \widehat{Y}_{\pi} \mid s\right) = 0,$$

where $\widehat{Y}_{\pi}$ denotes the expansion estimator $\widehat{Y}_{\pi}$.

Weighting by the inverse of the estimated response probability

- If the nonresponse model (7.6) is correctly specified (that is, both the form of the function $f(\mathbf{z}_i;.)$ and the vector $\mathbf{z}$ are correctly specified), then the PSA estimator is asymptotically unbiased and consistent for Y; Kim and Kim (2007).
- If the nonresponse model is correctly specified, the PSA estimator, $\widehat{Y}_{PSA}$, is asymptotically unbiased for Y, regardless of the y-variable being estimated.
- However, if the nonresponse model is incorrectly specified, $\widehat{Y}_{PSA}$ could be seriously biased.

Weighting by the inverse of the estimated response probability

Estimation of p_i:

- Parametric estimation: we are willing to make assumptions about the form of the function $f(\mathbf{z}_i; .)$.
  - Example: Logistic regression model

$$p_i = \frac{e^{\mathbf{z}_i'\gamma}}{1 + e^{\mathbf{z}_i'\gamma}}. \tag{7.8}$$

The estimated response probabilities are obtained as:

$$\widehat{p}_i = \frac{e^{\mathbf{z}_i'\widehat{\gamma}}}{1 + e^{\mathbf{z}_i'\widehat{\gamma}}},$$

where $\widehat{\gamma}$ is the maximum likelihood estimator of γ.

Weighting by the inverse of the estimated response probability

- Nonparametric estimation: we are not willing to make assumptions about the form of the function $f(\mathbf{z}_i; .)$.
  - Weighting classes formed on the basis of the estimated response probabilities. The idea is to order the estimated response probabilities and to cluster the units in a small number of clusters (around 5) according to these probabilities. The clusters are called response homogeneity groups. The nonresponse adjustment factor considered in the estimator definition is obtained by using the response rate in each cluster.
  - Kernel regression (Da Silva and Opsomer, 2006) or local regression (DaSilva and Opsomer, 2008) are other possibilities.
- More flexible and robust than parametric methods if the form of the function $f(\mathbf{z}_i; .)$ is misspecified.

Weighting by the inverse of the estimated response probability

- If the p_i's were known, the total variance of $\tilde{Y}_{PSA}$ in (7.5) would be given by

$$V\left(\tilde{Y}_{PSA}\right) = \underbrace{V_p\left(\widehat{Y}_F\right)}_{\text{sampling variance}} + \underbrace{\sum_{i\in U} d_i \frac{(1-p_i)}{p_i} y_i^2}_{\text{nonresponse variance}} \qquad (7.9)$$

- If $p_i = 1$ in (7.9), the nonresponse variance is equal to zero, as expected.
- Since $(1-p_i)/p_i$ is a decreasing function of p_i, the nonresponse variance decreases as p_i increases.

Weighting by the inverse of the estimated response probability

- When the p_i's are estimated, it is not possible to obtain the exact variance of $\widehat{Y}_{PSA}$ in (7.7). A Taylor expansion procedure needs to be used; Kim and Kim (2007).
- If the p_i's are parametrically modeled and the design weights are not used for estimating the p_i's, Kim and Kim (2007) showed that
$$V\left(\widehat{Y}_{PSA}\right) \leq V\left(\tilde{Y}_{PSA}\right).$$

Imputation for item nonresponse

Why impute?

- Imputation leads to the creation of a complete data file.
- Unlike weighting adjustment methods, imputation permits the use of a single sampling weight.
- Once the data have been imputed, results of identical analysis performed by different analysts will be consistent.
- Complete data estimation methods can be readily applied by data users to obtain point estimates (but not variance estimates).

Imputation for item nonresponse

Warnings:

- Imputed data are artificial and could give a false impression of accuracy.
- Imputation distorts the relationships between variables.
- Treating imputed values as if they were observed could lead to a substantial underestimation of the variance, especially if the nonresponse rate is appreciable

An imputed estimator

- Suppose we want to estimate the population total $Y = \sum_{i \in U} y_i$.
- We select a random sample, s, of size n, according to a given sampling design $p(s)$.
- Some y-values are missing. Therefore, it is not possible to compute the complete data estimator $\widehat{Y}_\pi = \sum_{i \in s} d_i y_i$.
- Let s_r be the (random) set of respondents to item y, of size n_r and let s_m be the (random) set of nonrespondents to item y, of size n_m.
- We have $s = s_r \cup s_m$ and $n_r + n_m = n$.

An imputed estimator

- We define an imputed estimator of Y as

$$\widehat{Y}_I = \sum_{i \in s} d_i r_i y_i + \sum_{i \in s} d_i (1 - r_i) y_i^*,$$

  where r_i is a response indicator attached to unit i such that $r_i = 1$ is unit i responded to item y and $r_i = 0$, otherwise and y_i^* denotes the imputed value used to 'fill in' the missing value y_i.

- The value y_i^* depends on the imputation method used.

Imputation Methods

- The imputation methods may be classified into 2 groups:
    - Deterministic imputation methods: methods that yield a fixed imputed value given the sample and the set of respondents if the imputation process is repeated.
    - Random imputation methods: methods having a random component (and, as a result, that do not yield a fixed value if the imputation process is repeated).

Imputation methods

Deterministic imputation methods:

- Regression imputation
- Ratio imputation
- Mean imputation
- Previous value imputation
- Nearest-neighbour imputation

Random imputation methods:

- Random hot-deck imputation
- Imputation with added residuals (e.g., ratio and regression imputation with added residuals)

Imputation methods

An alternative classification, whereby we distinguish between the donor imputation methods and the predicted value imputation methods, can be used:

- Donor imputation methods: methods that use the observed, actual, value of a respondent to replace the missing value
- Predicted value imputation methods: methods that use functions of the respondent values to obtain an imputed value.

Imputation methods

Donor imputation methods:

- Random hot-deck imputation
- Nearest-neighbour imputation

Predicted value imputation methods

- Regression imputation
- Ratio imputation
- Mean imputation
- Previous value imputation

Imputation methods

- Except manual imputation, all the imputation methods used in practice may be motivated by the general model (Kalton and Kasprzyk, 1986):

$$y_i = f(\mathbf{z}_i; \boldsymbol{\beta}) + \varepsilon_i, \tag{7.10}$$

$$E(\varepsilon_i) = 0, \quad E(\varepsilon_i \varepsilon_j) = 0, \quad i \neq j, \quad E\left(\varepsilon_i^2\right) = \sigma^2 c_i,$$

where $\mathbf{z} = (z_1, \ldots, z_Q)^\top$ is a q-vector of auxiliary variables available for all the sampled units (respondents and nonrespondents), $\boldsymbol{\beta}$ is a vector of unknown parameters and c_i is a known quantity attached to unit i.

Imputation methods

- In the case of deterministic imputation methods, the imputed value y_i^* for $i \in s_m$ is given by

$$y_i^* = f(\mathbf{z}_i; \widehat{\boldsymbol{\beta}}_r),$$

  where $\widehat{\boldsymbol{\beta}}_r$ is an estimator of $\boldsymbol{\beta}$ based of the responding units (e.g., the maximum likelihood estimator or the least square estimator).
- Random imputation may be seen as a deterministic imputation plus an added random residual.
- The residual may be drawn, for example, from a normal distribution with mean 0 and variance a.
- In practice, it is preferable to randomly select the residuals from the empirical distribution function of the respondent residuals. This method is nonparametric in nature.

Imputation methods

- Let

$$e_j = c_j^{-1/2}\left(y_j - f(\mathbf{z}_i;\widehat{\boldsymbol{\beta}}_r)\right).$$

- Let

$$\tilde{e}_j = e_j - \bar{e}_r,$$

  be the standardized residual for $j \in s_r$, where $\bar{e}_r$ is the mean of the residuals e_j based on the responding units.

- The missing value y_i, $i \in s_m$, is then replaced by

$$y_i^* = f(\mathbf{z}_i;\widehat{\boldsymbol{\beta}}_r) + c_i^{-1/2}e_i^*,$$

  where e_i^* is randomly selected (usually with replacement) from the set of the standardized residuals $E_r = \{\tilde{e}_j; j \in s_r\}$.

Imputation methods

- In practice, we may use some form of weighted or unweighted imputation.
- Weighted imputation (deterministic or random) uses the sampling weights d_i in the construction of the imputed values or to select the donors.

  Let

  $$\omega_i = \begin{cases} d_i & \text{for weighted imputation} \\ 1 & \text{for unweighted imputation} \end{cases}$$

- Other choices of ω_i can be used; e.g., Haziza (2009).

Deterministic regression imputation

We set $f(\mathbf{z}_i;\boldsymbol{\beta}) = \mathbf{z}_i^\top \boldsymbol{\beta}$ and $c_i = \boldsymbol{\lambda}^\top \mathbf{z}_i$, for some vector of constants $\boldsymbol{\lambda}$. This leads to

$$y_i^* = \mathbf{z}_i^\top \widehat{\mathbf{B}}_r, \tag{7.11}$$

where

$$\widehat{\mathbf{B}}_r = \left[\sum_{i\in s} \omega_i r_i c_i^{-1} \mathbf{z}_i \mathbf{z}_i^\top\right]^{-1} \sum_{i\in s} \omega_i r_i c_i^{-1} \mathbf{z}_i y_i$$

is the weighted least square estimator of $\boldsymbol{\beta}$ (with weights $\omega_i c_i^{-1}$).

- Appropriate when the relationship between y and $\mathbf{z}$ is linear and the vector $\mathbf{z}$ explains the y-variable well.

Deterministic ratio imputation

Ratio imputation is a special case of regression imputation. Suppose that a quantitative variable z is available. Let $f(\mathbf{z}_i; \boldsymbol{\beta}) = \beta z_i$ and $c_i = z_i$. Then, (7.11) reduces to

$$y_i^* = \widehat{B}_r z_i = \frac{\overline{y}_r}{\overline{z}_r} z_i,$$

where $(\overline{y}_r, \overline{z}_r) = \sum_{i \in s} \omega_i r_i (y_i, z_i) / \sum_{i \in s} \omega_i r_i$ denote the weighted means of the respondents corresponding to y and z, respectively.

- Appropriate when the relationship between y and z is linear, goes through the origin and the correlation between both variables is large.
- Could lead to substantially biased estimators if the relationship does not go through the origin.

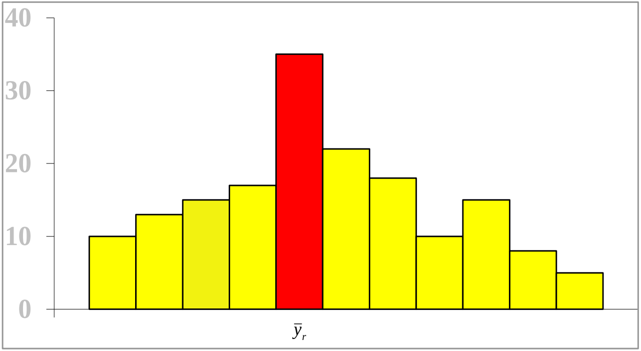
Mean imputation

Mean imputation is another special case of regression imputation. Each missing value is replaced by the mean of the respondents. Let $f(\mathbf{z}_i; \boldsymbol{\beta}) = \beta$ and $c_i = 1$. Then, (7.11) reduces to

$$y_i^* = \widehat{B}_r = \overline{y}_r = \frac{\sum_{i \in s} \omega_i r_i y_i}{\sum_{i \in s} \omega_i r_i}.$$

- May be useful when applied within imputation classes.
- Distorts the distribution of the variable being imputed considerably if the response rates are low.

Distribution of the variable of interest after mean imputation



Previous value or historical imputation

The missing value $y_{i,t}$ at time t is replaced by the value observed on the same unit at a previous occasion, $y_{i,t-1}$. Here, $z_i \equiv y_{i,t-1}$.

We have $f(\mathbf{z}_i;\beta) = y_{i,t-1}$, which leads to

$$y^*_{i,t} = y_{i,t-1}.$$

- Useful in repeated surveys.
- Appropriate when the variables being imputed are stable over time. That is, it is appropriate when the relationship between y_t and y_{t-1} goes through the origin with a slope close to 1.
- Requires file matching.

Nearest-neighbour imputation

The missing value is replaced with the nearest-neighbour respondent value (based on a certain distance with respect to one or more auxiliary variables). We have

$$y_i^* = y_j \ \text{ for } \ j \in s_r \ \ s.t. \ \ dist(\mathbf{z}_j, \mathbf{z}_i) \text{ is minimum}$$

Example of distance function:

$$dist(\mathbf{z}_j, \mathbf{z}_i) = \left(\sum_{k=1}^{q} a_k |z_{kj} - z_{ki}|^b \right)^{1/b},$$

where a_k is a weight expressing the importance of the variable z_k, $k = 1, \ldots, q$ and $b \geq 1$.

- The case $b = 2$ leads to an Euclidian type distance.
- Letting b tend to infinity leads to the so-called minimax distance.

Nearest-neighbour imputation

- Nearest-neighbour is a nonparametric imputation method. We do not need to specify the form of $f(\mathbf{z}_i; \boldsymbol{\beta})$, nor the variance structure σ_i^2.
- Appropriate when the relationship between y and $\mathbf{z}$ is not linear.
- The resulting imputed value is an actual, observed, value.
- The distance measure needs to be chosen (Euclidian, minimax, ...).
- In case of more than one auxiliary variable, the latter should be standardized (for example by subtracting the mean and dividing by the standard deviation).

Random hot-Deck imputation

The missing value is replaced by the value of a donor selected randomly from the set of respondents, i.e.,

$$y_i^* = y_j \text{ for } j \in s_r.$$

Random hot-deck imputation can be seen as mean imputation plus an added random residual, i.e.,

$$y_i^* = \overline{y}_r + e_i^*,$$

where e_i^* is randomly selected from the set of standardized residuals, $\tilde{e}_j = y_j - \overline{y}_r,\ j \in s_r$.

- It leads to actual, observed, values.
- Unlike mean imputation, it tends to preserve the distribution of the variable being imputed.

Ratio imputation with added residuals

Ratio imputation with added residuals can be seen as ratio imputation plus a random residual e^*, that is,

$$y_i^* = \widehat{B}_r z_i + \sqrt{z_i} e_i^*,$$

where e_i^* is randomly selected from the set of standardized residuals:

$$\tilde{e}_j = \frac{1}{\sqrt{z_j}} \left(y_j - \widehat{B}_r z_j \right), \; j \in s_r.$$

- Appropriate when the relationship between y and z is linear, goes through the origin and the correlation between both variables is large.
- Unlike ratio imputation, it tends to preserve the distribution of the variable being imputed.

Properties of imputation methods

- Behind each imputation method, there is a set of assumption. It is important to perform model validation to make sure that the underlying assumptions hold.
- A careful modeling exercise should be performed in order to select the appropriate auxiliary variables.
- For each variable being imputed, response flags should be provided in the data file. Response flags are needed at the variance estimation stage and at the analysis stage.

Properties of imputation methods

- Deterministic methods lead to asymptotically unbiased estimators of population totals (or means) if the underlying models are valid.
- Deterministic methods (except nearest-neighbour imputation), unlike random imputation methods, tend to distort the distribution of the variables being imputed. Therefore, estimators of other types of parameters (e.g., quantiles) could be considerably biased. Random imputation methods are thus preferable in this case.
- Unlike deterministic methods, random imputation methods suffer from an additional component of variance (i.e., the variance due to imputation), due to the selection of random residuals.

Properties of imputation methods

- Unlike predicted value imputation methods, donor imputation methods (such as random hot deck imputation) yield actual (observed) data.
- Donor imputation methods are useful when the goal is to impute several variables at a time because a single donor can be used to impute all the missing values of a given nonrespondent. This helps preserving the relationships between variables.
- When the variable being imputed is categorical, donor imputation is preferable to avoid the possibility of impossible values in the data file.