DATA 606: Statistical Methods in Data Science

—— Nonresponse

Wenjun Jiang

Department of Mathematics & Statistics
The University of Calgary

Lecture 6

General idea

The best way to deal with nonresponse is to prevent it!

General idea

The best way to deal with nonresponse is to prevent it!

- *Unit nonresponse*: the entire observation unit is missing.

Example 1

In a survey of breeding ducks, for example, some birds will not be found by the researchers.

- *Item nonresponse:* some characteristics of the unit are missing.

Example 2

The nest may be raided by predators before the investigator can determine how many eggs were laid.

General idea

How to deal with nonresponse?

- Prevent it. Design the survey so that nonresponse is low. This is by far the best method.
- Take a representative subsample of the nonrespondents; use that subsample to make inferences about the other nonrespondents.
- Use a model to predict values for the nonrespondents.
- Ignore the nonresponse (not recommended, but unfortunately common in practice).

The effects of ignoring nonresponse

The main problem caused by nonresponse is potential bias.

Example 3 (Bias)

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

The effects of ignoring nonresponse

- $\bar{y}_R$ is an approximately unbiased estimator of the mean in the respondent stratum.
- $\bar{y}_U = \frac{N_R}{N} \cdot \bar{y}_{RU} + \frac{N_M}{N} \cdot \bar{y}_{MU}$.
- The bias is approximately

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

  The bias is small if
  - The mean for the nonrespondents is close to the mean fo the respondents.
  - The fraction $\frac{N_M}{N}$ is small.

Survey design

Background: Many persons new to surveys (and some, unfortunately, not new) simply jump in and start collecting data without considering potential problems in the data collection process; they mail questionnaires to everyone in the target population and analyze those that are returned.

Survey design

Background: Many persons new to surveys (and some, unfortunately, not new) simply jump in and start collecting data without considering potential problems in the data collection process; they mail questionnaires to everyone in the target population and analyze those that are returned.

Example 4

The 1990 U.S. decennial census attempted to survey each of the over 100 million households in the United States. The response rate for the mail survey was 65%; households that did not mail in the questionnaire needed to be contacted in person, adding millions of dollars to the cost of the census. Increasing the mail response rate for future censuses would result in tremendous savings.

Survey design

Some factors which may influence response rates and data accuracy.

- *Survey content*. E.g., Asurvey on drug use or financial matters may have a large number of refusals. Sometimes the response rate can be increased for sensitive items by careful ordering of the questions.

Survey design

Some factors which may influence response rates and data accuracy.

- *Survey content*. E.g., Asurvey on drug use or financial matters may have a large number of refusals. Sometimes the response rate can be increased for sensitive items by careful ordering of the questions.

- *Time of survey*. Some calling periods or seasons of the year may yield higher response rates than others. The vacation month of August, for example, would be a bad time to take a one-time household survey in Germany.

Survey design

Some factors which may influence response rates and data accuracy.

- *Survey content*. E.g., Asurvey on drug use or financial matters may have a large number of refusals. Sometimes the response rate can be increased for sensitive items by careful ordering of the questions.

- *Time of survey*. Some calling periods or seasons of the year may yield higher response rates than others. The vacation month of August, for example, would be a bad time to take a one-time household survey in Germany.

- *Data-collection method*. Generally, telephone and mail surveys have a lower response rate than in-person surveys. Computer-Assisted Telephone Interviewing (CATI) and Computer-Assisted Personal Interviewing (CAPI) have been demonstrated to improve accuracy of data collected.

Survey design

- *Questionnaire design*. We have already seen in Lecture 1 that question wording has a large effect on the responses received; it can also affect whether a person responds to an item on the questionnaire.

Survey design

- *Questionnaire design*. We have already seen in Lecture 1 that question wording has a large effect on the responses received; it can also affect whether a person responds to an item on the questionnaire.

- *Incentives and disincentives.* Incentives, financial or otherwise, may increase the response rate. Disincentives may work as well: Physicians who refused to be assessed by peers after selection in a stratified sample from the College of Physicians and Surgeons of Ontario registry had their medical licenses suspended.

Callbacks and two-phase sampling

Callbacks

- Virtually all good surveys rely on callbacks to obtain responses from persons not at home for the first try.
- Analysis of callback data can provide some information about the biases that can be expected from the remaining nonrespondents.

Drawback: the follow-up calls use a more expensive method such as a personal interview.

Callbacks and two-phase sampling

- The population is divided into two strata, respondents and nonrespondents.
- Total number of units in the sample: n, in which n_r respond and n_M do not respond.
- Make a second call on a randomly subsampled $100v\%$ of the n_M nonrespondents.
- Let $\bar{y}_R$ be the sample average of the original respondents, and $\bar{y}_M$ be the average of the subsampled nonrespondents.

Callbacks and two-phase sampling

- The population is divided into two strata, respondents and nonrespondents.
- Total number of units in the sample: n, in which n_r respond and n_M do not respond.
- Make a second call on a randomly subsampled $100v\%$ of the n_M nonrespondents.
- Let $\bar{y}_R$ be the sample average of the original respondents, and $\bar{y}_M$ be the average of the subsampled nonrespondents.

The population mean and total estimates are

$$\hat{\bar{y}} = \frac{n_R}{n}\bar{y}_R + \frac{n_M}{n}\bar{y}_M.$$

$$\hat{t} = N \cdot \hat{\bar{y}} = \frac{N}{n}\sum_{i \in S_R} y_i + \frac{N}{n}\frac{1}{v}\sum_{i \in S_M} y_i.$$

Why do not respond?

Suppose we have

$$\phi_i = \mathbf{P}(\text{unit } i \text{ responds}),$$

- y_i: the response of interest (annual income, medical treatment experience, weekly average working time).
- $\mathbf{x}_i = (x_1, x_2, \ldots, x_k)$: other information of unit i (e.g., age, gender, marriage status, etc.).

Question: does ϕ_i relate to y_i or $\mathbf{x}_i$?

Why do not respond?

- **Missing completely at random**
  This happens if ϕ_i does not depend on y_i, $\mathbf{x}_i$ or the survey design. A direct example is people forget. Another example from the textbook: someone at the laboratory drops a test tube containing the blood sample of one of the survey participants—there is no reason to think that the dropping of the test tube had anything to do with the white blood cell count.

Why do not respond?

- **Missing completely at random**
  This happens if ϕ_i does not depend on y_i, $\mathbf{x}_i$ or the survey design. A direct example is people forget. Another example from the textbook: someone at the laboratory drops a test tube containing the blood sample of one of the survey participants—there is no reason to think that the dropping of the test tube had anything to do with the white blood cell count.

- **Missing at random given covariates**
  This happens when ϕ_i depends on $\mathbf{x}_i$ but not on y_i. This is sometimes called the ignorable nonresponse. For example, people who work 10 hours a day is less willing to respond to a survey compared to those who only work 4 hours a day.

Why do not respond?

- **Not missing at random**
  This happens when ϕ_i depends on y_i. For example, in NCVS: It is suspected that a person who has been victimized by crime is less likely to respond to the survey than a nonvictim, even if they share the values of all known variables such as race, age, and gender.

Weighting method

- In a survey, we sample because we believe our sampled units can represent the population.

Weighting method

- In a survey, we sample because we believe our sampled units can represent the population.
- Similarly, to apply a weighting method, we believe those respondents could represent those nonrespondents. In other words, we assume the nonresponses are ignorable.

Weighting method

- *Sampling weights* w_i are the reciprocals of the inclusion probabilities π_i.
- In simple random sampling, weight $w_i = \frac{1}{\frac{n}{N}} = \frac{N}{n}$; in stratified sampling, weight $w_i = \frac{N_h}{n_h}$ for unit i in stratum h.
- In the presence of nonresponse,

$$\mathbf{P}(\text{unit } i \text{ is selected and responds}) = \pi_i \cdot \phi_i,$$

where the response probability is usually estimated as $\hat{\phi}_i$.

- The weight for a respondent is $\frac{1}{\pi_i \cdot \hat{\phi}_i}$.

Weighting method

- *Weighting adjustment class*: like stratum. It is hoped that respondents and nonrespondents in the same class are similar.

Weighting method

- *Weighting adjustment class*: like stratum. It is hoped that respondents and nonrespondents in the same class are similar.
- *Estimate of the response probability*:

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

Weighting method

- *Weighting adjustment class*: like stratum. It is hoped that respondents and nonrespondents in the same class are similar.
- *Estimate of the response probability*:

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

Example 5

	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	

Weighting method

Let the average for the respondents in class c be $\bar{y}_{cR}$ and the number of sample units in class c be n_c, then

$$\hat{t} = N \sum_c \frac{n_c}{n} \bar{y}_{cR}.$$

Weighting method

Let the average for the respondents in class c be $\bar{y}_{cR}$ and the number of sample units in class c be n_c, then

$$\hat{t} = N \sum_c \frac{n_c}{n} \bar{y}_{cR}.$$

Question: how to construct weighting adjustment classes?

Imputation

- Missing items may occur in surveys for several reasons: An interviewer may fail to ask a question; a respondent may forget the answer or cannot provide the information; a clerk entering the data may skip the value.

- **Imputation** is commonly used to assign values to the missing items[1].

[1] 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

Imputation

Example 6

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

Cell mean imputation

- Respondents are divided into classes (cells) based on known variables.
- The average of the values for the responding units in cell c is used to substitute each missing value.

Example 7

		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 values 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.

Hot-deck imputation

- **Sequential hot-deck imputation**: this procedure imputes the value in the same subgroup that was last read by the computer.

Example 8

Person 19 is missing the response for crime victimization. Person 13 had the last response recorded in her subclass, so the value 1 is imputed.

- **Random hot-deck imputation**: A donor is randomly chosen from the persons in the cell with information on all the missing items.

Example 9

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.

Hot-deck imputation

- **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.

Example 10

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.

Regression imputation

Regression imputation predicts the missing value using a regression of the item of interest on variables observed for all cases.

Regression imputation

Regression imputation predicts the missing value using a regression of the item of interest on variables observed for all cases.

In our example, a logistic regression could be applied

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

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.

Cold-deck imputation

In cold-deck imputation, the imputed values are from a previous survey or other information, such as from historical data.

Advantages and disadvantages

Advantages:

- 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.

Advantages and disadvantages

Advantages:

- 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.

Disadvantages:

- The foremost danger of using imputation is that future data analysts will not distinguish between the original and the imputed values.
- If you treat the imputed values as though they were observed in the survey, the estimated variance will be too small.