What’s next?

- Once the data are collected, we are stuck with the sample we have and the problems that come with it.
- Unfortunately, things in the data collection procedure aren’t always optimal.
  - Second-hand data sources.
  - Things come up you didn’t think about.
  - Sub-optimal planning.
- So what do we do afterwards? How do we get the most out of our data?

Data consistency

The data should not present contradictory information.

- Men should not have a response for “Ever been pregnant?”
- People who claim they never graduated from college should not have a “date of graduation” recorded.
- Percentages should be between 0-100 (valid range).
- Historical: subject cannot “undo” prior history; e.g., “married” in 1990, but “never married” in 2000.
- Valid skips: subjects currently “not working” should not report a wage.

Consistency vs Integrity

- Data consistency and integrity are not the same thing.
- i.e. People who claim they never graduated from college should not have a “date of graduation” recorded. (which is the right answer?)
- You can try to use auxiliary information to determine the truth, but this is not always available.
- Use of computer assisted entry systems “traps” errors before you see the data, but this requires having foresight as to which problems may arise.

Data restructuring

The format of the data collected is not always the format you want for analysis. For example:

- Need to assign numeric codes to “check box” answers. (What were you planning on doing with your “check your top three”?)
- You have open-ended questions, such as “what do you like about NYC?” To facilitate analysis, the text in the response can be recoded into a small number of categories.
    - The food, the people, the buildings, etc., and these choices can be restricted to a reasonable number.

Data restructuring

- A variant of the above is “what type of work do you do?” or “what was your major in college?”
- It is often useful to recode the response into a smaller list of occupations or college majors, in part to reduce redundancy (e.g., PoliSci=Gov’t).
- Can sometimes avoid this by restricting the answers in the input stage, but often the point of open-ended questions is to *not restrict* this process.

Data restructuring

- You ask for specific names (of employers, schools etc.), but you need to “de-identify” these using codes so that anonymity is preserved.
- Note from your friendly neighborhood statistician: Please please please please document ALL coding steps carefully. Otherwise, your data is completely illegible to your data analyst.

Use of human coders

- Make sure that your coders are properly trained and communicate with you and with each other.
- The skills needed for a good Coder are: good subjective comprehension of what is being said and clean objectivity when selecting the correct code.
- Think about coder-related variance and coder effects (like interviewer variance or cluster effects)
- Why does this happen?

Coding open-ended responses

1. Read through the responses
   - In order to get a good sense of the data, you first need to read through the responses.
   - Depending on *n*, you may read all or some.
   - As you read through them, you will begin to get a sense of the emerging themes.
   - It is also helpful to have another person review the responses independently and check their sense of the data against your own.
   - This will help minimize the influence of your own biases on the data, especially if you are closely involved with the program.

Coding open-ended responses

2. Develop categories
   - Next you will need to develop categories that include the themes that emerged in your initial review.
   - For example, if the survey question asked people for suggestions on ways to improve a program, your categories might include things like "changes to content", or "more group activities", or "no changes needed".

Coding open-ended responses

3. Assign each response to a category (or categories)
    - Once you have established your categories you will need to assign each comment to one or several categories (this is known as “coding”).
4. Check your categories
    - Now is a good time to check and see if your categories are actually appropriate.
    - You might find that most of your responses fall into one category and that the category could actually be broken into more specific subcategories.

Coding

Example: “What did you like best about this program?”

Non-response

- Nonresponse can be broken down into its two major forms: unit nonresponse and item nonresponse.
- Unit nonresponse occurs for three major reasons: non-contacts, inability to participate and refusals.
- These mechanisms are different and can be prevented and handled in different ways.

Noncontacts

- Noncontacts are those respondents who we failed to reach.
- This can be because of “access impediments” (i.e. Do Not Call lists, junk mail screening, locked exterior doors to apartment buildings.)
- Also because of partially random and partially predictable patterns (working people are not at home between M-F 9 AM - 5 PM. Best to send interviewers to these places on S-R 6 PM - 9 PM).
- These issues are obviously mode-dependent.

Noncontacts

- The more people who live in a house, the higher the chance of reaching *someone.*
- If you can reach someone, then you can get info for others.
- Which types of people live alone vs with other people?
- Remember there are issues both of lower sample sizes and nonresponse bias (when specific subgroups of people are less likely to respond or be contacted).

Inability to participate

- Assuming eligibility (otherwise its a coverage problem), some people may be unable to participate because of physical or mental incapability or if language barriers prevent comprehension of the survey.
- (How many of you thought about training F-T-F interviewers in ASL?)
- Last week, we talked about an example where parents with small children were unable to answer the survey as directed because they were taking care of their kids. This applies also to the inability to participate.

Inability to participate

- Establishments may also be unable to participate if records or information is unavailable (i.e. a survey of schools seeking to measure the extent of contributions of private donors will be very difficult if no records are kept, even if the schools want to participate).
- Timing of surveys can affect the ability to participate as well. (How many of you would take time to answer a survey at the beginning of the semester vs during finals?)

Refusals

- Probably the best studied area of unit nonresponse is refusals.
- There is research to suggest that since survey requests are relatively rare, interviewers must act fast to prevent automatic refusals as a result of erroneously believe the survey to be a sales pitch.
- (How many of you have ignored those people on the street corner who “just want to know if you have a minute..”)

Recontacting refusals

- In general, potential respondents will decide to decline quickly (within the first 30 seconds), so the initial contact script plays an important role.
- Though, somewhere around 1/3 of people in a typical survey who initially refuse will accept if contacted later (refusal conversion), so don't give up after the first attempt.
- Approaches of "I'm not selling anything" have worked well.

Why respondents refuse

- Social/environment factors (e.g. levels of trust and reciprocity): for example, there is usually a higher refusal rate in large urban areas than in suburban or rural locales.
- Personal attributes: In general, males tend to refuse at a higher rate than females.
- Experience of interviewers: more experience = better at avoiding refusals (and better at converting).
- Survey aspects: incentives and presentation strategies may play a role in limiting nonresponse.

Why respondents refuse

- Opportunity cost: i.e. “but i could be doing all those other things instead.” Busy people and high earners might be reluctant to lose several hundred dollars in billable time providing free data to some researcher.
- Social isolation: This theory attempts to explain low response rates among the lowest and highest SES levels by positing that those far from the mean consider themselves isolated from major social institutions that conduct surveys. (i.e. universities, large corporations, government)

Why respondents refuse

- Interest in the topic: A major thread through the research on nonresponse is the common sense theory that people are more likely to respond on issues they care about.
- Oversurveying: The overall “burden” on respondents due to surveying has increased to the point that they now refuse.

Why respondents refuse

Leverage-Salience theory

The textbook tries to unify these theories and empirical results by explaining that different aspects of the survey (length, topic, sponsor, incentives) and the respondent's salience for each of these factors determines propensity for answering. This implies:

- People have many different considerations they weigh when deciding to participate.
- No one introduction or approach will work for all people, or even the same person over time.
- As a result, interviewers must be skilled at "selling" their survey in different ways to different respondents.

Reducing non-response

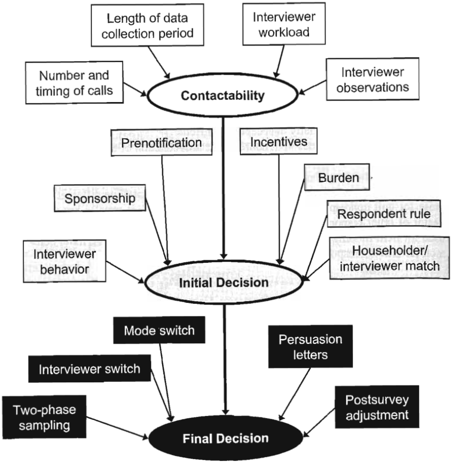


Figure 6.13 Tools for reducing unit nonresponse rates.

Still open questions

- When efforts to interview reluctant respondents succeed, do they provide responses more tainted by measurement error?
- When do efforts to increase response rates affect nonresponse errors and when do they not?
- How should efforts to reduce non contact vs refusal rates be balanced against one another?
- Considering both sampling error and nonresponse error, for a fixed budget, when can the researcher justify less than full effort to reduce nonresponse rates?

Weighting

- We’ve been dancing around the topic of weighting survey data all semester, now its time to get serious.
- Complex sampling designs may have unequal probabilities of selection:
  - Oversampling a particular stratum to ensure we get enough observations.
  - Imperfect stratification of clusters.
  - Mistakes in the procedures.
- Survey researchers also often use weights to adjust for unit non-response and to adjust the survey data to known population figures.

What is a weight?

- Think of it like the number of units in the population that the respondent represents.
- SRS is simplest: Imagine a population of size $N = 2,000,000$ from which we take a sample of size $n = 1,000$.
- How many other units does each unit in my sample represent?

What is a weight?

- Then the probability of selecting any one unit is:

$$p = n/N = 0.0005$$

- Each unit in the sample gets a weight of:

$$1/p = N/n = 2,000$$

- That is, each unit in the sample represents 2,000 units in the population

Weighting

- Today we will talk about four types of weights:
  - First-stage ratio adjustments.
  - Differential selection probabilities.
  - Weighting adjustment for Unit Nonresponse.
  - Poststratification Weighting.
- These weights are all applied on the individual unit level (people, firms, etc).

First-stage ratio adjustments

- In complex, multistage designs, the primary sampling units (PSU's - usually decent-sized geographical areas) are usually sampled with probability proportional to size (like county-level population counts from the most recent Census).
- A "problem" that arises is that primary sampling units (PSU; e.g., counties within regions) do not accurately represent important strata in the population.

First-stage ratio adjustments

- A first-stage adjustment is made so that selected PSUs can be used to calculate means, etc. that reflect the proper proportions of known demographics (the strata) in the population.
- This will increase the precision of any measurement that is sensitive to the strata in the population.
- Example: Target population is of size 1000; 700 identify as White, 100 as Black, 200 identify as something else. (Suppose you know this from the Census).

First-stage ratio adjustments

- Suppose the population is split into 10 equally sized PSU's (of size 100).
- Each has 1/10 of the total population, so you expect 70 White and 10 Black respondents in each PSU.
- You sample one PSU, but the #White = 60, while the #Black = 20.
- What's the naive estimate of the number of people that identify as Black in the population?

First-stage ratio adjustments

- Ratio adjustment suggests building a weight:

$$W_{i1} = \frac{\text{Stratum Pop Total from Frame}}{\text{Pop Total for Selected PSU/Prob of Selecting PSU}}$$

- In this case, thats: W_{i1} = (Total # Black in Population)/(Estimated Total # Black based on PSU sampled).
- Or:

$$W_{i1} = \frac{100}{\frac{20}{\frac{1}{10}}} = \frac{1}{2} \text{ for Black}$$

$$W_{i1} = \frac{700}{\frac{60}{\frac{1}{10}}} = \frac{7}{6} \text{ for White}$$

Differential Selection Probabilities

- The second type of weight is to adjust for differential selection probabilities of units.
- Example 1: Suppose you are running a survey on MPH students, and want to include both first year and graduating year students in your sample. While first year students are easy to find in classes, it is very hard to find graduating students. Therefore, you oversample (i.e. more than proportionate allocation would suggest, for example) those students to ensure a large enough representation of those students in your sample.

Oversampling

- Oversampling hard to reach sub-populations (women in math, graduating students, native American 4th graders) allows for a large enough *n* in that subpopulation to get reasonable standard errors for stratum level statistics.
- If your only goal is make inference within subgroups, then there's no problem and nothing needs to be done.
- However, we usually also want to aggregate (make inference about all students, all mathematicians, all 4th graders, etc) as well.
- How do you combine your sample?

Differential Selection Probabilities

- To aggregate data, you need to account for the intentional disproportional presence of certain subgroups (compared to an SRS or proportionate stratification, or the population).
- If you forget this step, your estimates will be off.
- So how do you fix it? Weight by the inverse probability of selection.

$$w_{i2} = \frac{1}{p_{i,selection}}$$

Weighting for Unit Nonresponse

- As we talked about earlier, nonresponse comes in two major forms: unit nonresponse and item nonresponse.
- There's also another distinction: what is the mechanism for missingness?
  - MCAR: Missing completely at random
  - MAR: Missing at random
  - NMAR: Not missing at random.
- MCAR and MAR are also referred to as ignorable missingness while NMAR is non-ignorable missingness.

Categorizing missingness

Consider the response y_i for person i. Define:

$$R_i = \begin{cases} 1 & \text{if } y_i \text{ is observed.} \\ 0 & \text{if } y_i \text{ is missing.} \end{cases}$$

The real question is: What does the probability of R_i depend on?

Missing Completely at Random (MCAR)

- MCAR is the “best” case, but just often isn’t the truth.
- For MCAR to hold:

$$Pr(R_i = 1 | x_i, y_i) = Pr(R_i = 1)$$

- That is, R_i is independent of known covariates x_i, the measurements of interest y_i, as well as the survey design.
- Under the MCAR assumption, if we start with a simple random sample from the population, then the units with missingness would be a simple random subsample of the original sample.

Examples of MCAR in surveys

- The post-office just lost the mail survey (it happens!)
- I completely forgot to answer a question.
- Related to nothing else, I just decided I didn't want to talk to the FTF interviewer.

The implication of MCAR is simply a loss of efficiency, which isn't great - but isn't horrible.

Weighting for Unit Nonresponse under MCAR

If the missingness is really MCAR, then the appropriate third stage weight is simply:

$$w_{i3} = \frac{n_{total}}{\#(R_i = 1)}$$

the total number of people surveyed (including those that are missing) over the number of responses obtained - or the inverse of the response rate.

Missingness at Random

- One step weaker than MCAR is MAR.
- Under MAR, we assume that the probability of response is related to observed covariates (or demographic characteristics), but has nothing to do with the survey questions asked.
- When MAR holds:

$$Pr(R_i = 1|x_i, y_i) = Pr(R_i = 1|x_i) = f(X_i\beta)$$

- In this case, you can fit a logistic regression of R_i on the demographic info to get an estimate of response propensity, $Pr(R_i = 1|x_i)$.

Adjusting for MAR

- Can only adjust for nonresponse using variables known on everyone in sample.
- Has to be based on external (not survey) data - often these are demographic types of variables.
- Estimate $Pr(R_i = 1|x_i)$ from a logistic regression, and then:

$$w_{i3} = \frac{1}{Pr(R_i = 1|x_i)}$$

Not Missing at Random

- This is the hardest and, often, most realistic situation.
- In this case $Pr(R_i = 1 | x_i, y_i)$ cannot be simplified.
- However, you don't have y_i because its missing.
- Why might this happen?
- What do we do about it?
    - This, unfortunately is beyond the scope of this course.
    - Though, take APSTA-GE 2013: Advanced Topics in Quantiative Methods: Missing Data. Meets Spring 2015: F 9:30-12:15 (Second 7 weeks of term) with Professor Jennifer Hill.

Post-stratification weighting

- The last type of weight we will discuss is poststratification - which adjusts the sample to an external sources of highly reliable data about the population (like the Census).
- For example, if the total population by age and gender is known due to a recent Census, then after all other weights are computed, survey researchers will compare their estimated totals to those Census figures. (why?)

Post-stratification

- Example: Population is known to be 52% female and 48% male
- However, survey results differ and show 50% female and 50% male
- Then adjust female weight by 0.52/0.50 = 1.04 and male weight by 0.48/0.50 = 0.96

Putting the weights together

- So there are four stages of weights. How do you mix them?

  $$w_{i,final} = w_{i1} w_{i2} w_{i3} w_{i4}$$

  multiply them!
- What do you do with them? Calculate a weighted mean:

  $$\bar{y} = \frac{\sum_i w_{i,final} y_i}{\sum_i w_{i,final}}$$

Example

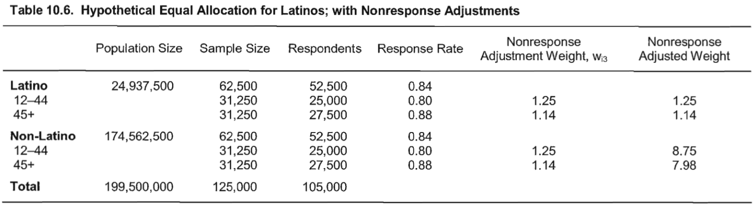


Table 10.6. Hypothetical Equal Allocation for Latinos; with Nonresponse Adjustments

	Population Size	Sample Size	Respondents	Response Rate	Nonresponse Adjustment Weight, w_{i3}	Nonresponse Adjusted Weight
Latino	24,937,500	62,500	52,500	0.84		
12–44		31,250	25,000	0.80	1.25	1.25
45+		31,250	27,500	0.88	1.14	1.14
Non-Latino	174,562,500	62,500	52,500	0.84		
12–44		31,250	25,000	0.80	1.25	8.75
45+		31,250	27,500	0.88	1.14	7.98
Total	199,500,000	125,000	105,000			

$$w_2(Latino) = 24,937,500/62,500 = 399$$

$$w_2(Non - Latino) = 174,562,500/62,500 = 2,793$$

$$2793/399 = 7$$

Latino's were oversampled at a rate of 7 times that of Non-Latinos

Example

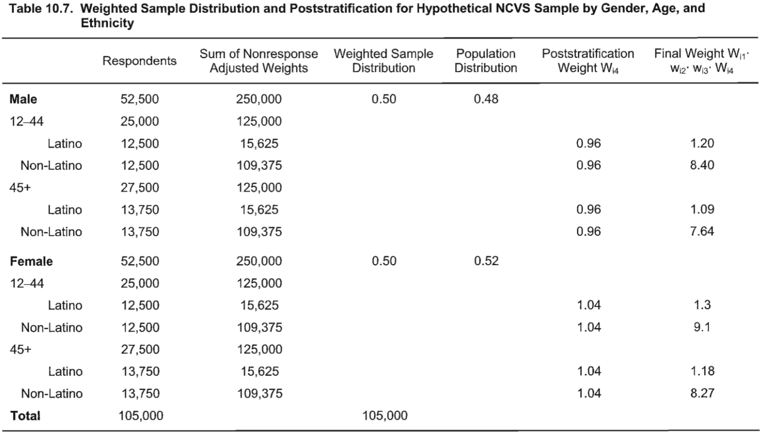


Table 10.7. Weighted Sample Distribution and Poststratification for Hypothetical NCVS Sample by Gender, Age, and Ethnicity

	Respondents	Sum of Nonresponse Adjusted Weights	Weighted Sample Distribution	Population Distribution	Poststratification Weight W_{i4}	Final Weight $W_{i1} \cdot w_{i2} \cdot w_{i3} \cdot W_{i4}$
Male	52,500	250,000	0.50	0.48		
12–44	25,000	125,000				
Latino	12,500	15,625			0.96	1.20
Non-Latino	12,500	109,375			0.96	8.40
45+	27,500	125,000				
Latino	13,750	15,625			0.96	1.09
Non-Latino	13,750	109,375			0.96	7.64
Female	52,500	250,000	0.50	0.52		
12–44	25,000	125,000				
Latino	12,500	15,625			1.04	1.3
Non-Latino	12,500	109,375			1.04	9.1
45+	27,500	125,000				
Latino	13,750	15,625			1.04	1.18
Non-Latino	13,750	109,375			1.04	8.27
Total	105,000		105,000			

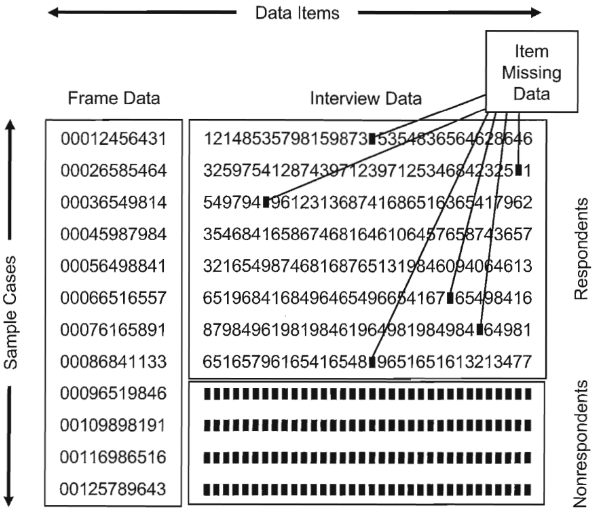
Weights can’t fix everything

- Weights are good for fixing some features of the dataset, but they can’t do everything.
- There’s nothing you can do to fix selection bias issues.
- No matter how good your weights can be, it is still better (you will get better estimates) if you just take a good sample to begin with.

Item Non-response

- Often times, you will receive an incomplete survey.
- In this case, only the responses to certain items will be missing.
- What are the options for handling this?
  - Option 1: Delete that respondent.
  - Option 2: Fill in a value for what they “likely” would have answered.

Item Missingness



Frame Data	Interview Data
00012456431	12148535798159873▮5354836564628646
00026585464	32597541287439712397125346842325▮1
00036549814	549794▮961231368741686516365417962
00045987984	3546841658674681646106457658743657
00056498841	3216549874681687651319846094064613
00066516557	6519684168496465496654167▮65498416
00076165891	8798496198198461964981984984▮64981
00086841133	65165796165416548▮9651651613213477
00096519846	▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮
00109898191	▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮
00116986516	▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮
00125789643	▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮

Figure 10.4 Unit and item nonresponse in a survey data file.

Option 1: Listwise (or case wise) deletion

- If someone is missing a response for ANY item, remove them from the sample (delete the whole row).
- This can be terribly inefficient (i.e. losing 90% of your sample) and biasing (if not MCAR).
- However, this is the default approach of statistical softwares, so be very very careful (in fact, you may not even notice that this has happened..)

Pro's and Con's of listwise deletion

Pro's:

- Its easy to do.
- Allows comparability of univariate statistics provided that only the complete responses are used for all analyses (that is, the sample doesn't vary over analyses.)
- If you have less than 5% missingness in your sample, then its not that bad.

Con's:

- Inefficient (loss of precision).
- Potential bias if the missingness mechanism is not MCAR.
- Let's be honest, we should be able to come up with *something* smarter to do.

Option 2: Fill in the Blank

- Imputation is a series of methods which fills in the missing data with “best” guesses for what it should be.
- Advantages:
  - Maximizes the use of all survey data
  - Univariate analysis of each variable will have same number of observations
- Disadvantages:
  - Some think of imputed data as ‘made up’ data
  - Statistical software often not designed to distinguish between real and imputed data
  - Results in improper standard error estimates, too narrow confidence intervals, etc.

Deductive imputation

- Some imputation may be done at the data cleaning stage
- Sometimes one can deduce quite accurately the value of one cell based on the values of other cells for a particular data record, just using the logical structure of the data
- If, for example, someone says that they report a positive status Type II diabetes at time 1 of a longitudinal survey, but then forget to fill that information in at a later date, then it makes sense to impute a positive diabetes value at time 2 (last observation carried forward)

Deductive imputation

- However, it may not make as much sense to impute a negative value from time 1 to time 2 in the absence of other information
- Deductive imputation is useful, but dangerous in that you are generally making assumptions that are not reflected in the sampled data, but rather some set of untestable assumptions

Unconditional (column) mean imputation

- Replace the missing value with the mean value in that column (the mean across all people for that item.)
- Highly underestimates standard errors (why? because there is no consideration for the uncertainty of these values, and you put everyone at the mean!) so more complicated standard error calculations need be be done.
- Think about how bad this would be under MNAR missingness.
- Conditional mean imputation methods also exist where the value imputed for you depends on some known information for you (like demographic info)

Regression imputation

- A refinement of this method is to use some sort of regression model to predict a given variable Y_j as a function of other observed variables.
- Assume that we are missing a fair amount of data on household income.
- We could estimate a conditional regression imputation model on the complete cases:

$$Income_i = \beta_0 + \beta_1 Education_i + \beta_2 Age_i + \ldots + \beta_k X_{ik} + \varepsilon_i$$

- Then, we use this regression model to predict income values for the missing cases.

Regression imputation

- This method relies on having complete information for the other variables as well.
- Still will result in underestimated standard errors.
- One option is: get a predicted value from this regression model, then add a random noise term (ε_i) to it (called stochastic regression imputation).
- This option fixes the “spiking” problem but not the standard error problem.

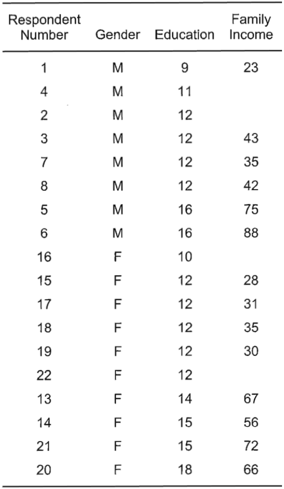
Biggest issue with Regression imputation

- The primary disadvantage is that the choice of an unbiased regression imputation model is not clear and incorrectly specifying the regression model for the missing values can lead to significant bias
- Hopefully, in your regression classes, you saw that using the wrong covariates in the regression model (including ones that don't belong or worse, not including ones that do belong) result in biased estimates and bad standard errors. The same issue exists here.

Hot Deck Imputation

- The general idea underlying hot deck imputation is to replace missing values with values from “donors” who are similar in some way to the unit with missing data.
- The term comes from the era of punch card computers: the hot deck was the deck of punch cards currently being processed (vs cold deck imputation where you take a value from a different data set).
- Basic idea:
  - Sort data by important variables
  - Start at the top and replace any missing data with value of the immediately preceding observation
  - If first one is missing, replace with appropriate mean value

Example: Hotdeck



Respondent Number	Gender	Education	Family Income
1	M	9	23
4	M	11	
2	M	12	
3	M	12	43
7	M	12	35
8	M	12	42
5	M	16	75
6	M	16	88
16	F	10	
15	F	12	28
17	F	12	31
18	F	12	35
19	F	12	30
22	F	12	
13	F	14	67
14	F	15	56
21	F	15	72
20	F	18	66

Example: Hotdeck

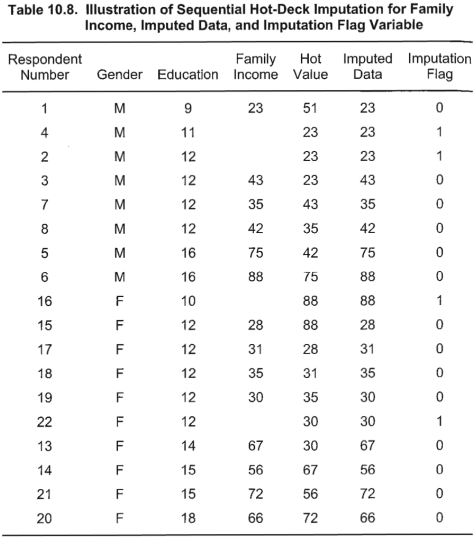


Table 10.8. Illustration of Sequential Hot-Deck Imputation for Family Income, Imputed Data, and Imputation Flag Variable

Respondent Number	Gender	Education	Family Income	Hot Value	Imputed Data	Imputation Flag
1	M	9	23	51	23	0
4	M	11		23	23	1
2	M	12		23	23	1
3	M	12	43	23	43	0
7	M	12	35	43	35	0
8	M	12	42	35	42	0
5	M	16	75	42	75	0
6	M	16	88	75	88	0
16	F	10		88	88	1
15	F	12	28	88	28	0
17	F	12	31	28	31	0
18	F	12	35	31	35	0
19	F	12	30	35	30	0
22	F	12		30	30	1
13	F	14	67	30	67	0
14	F	15	56	67	56	0
21	F	15	72	56	72	0
20	F	18	66	72	66	0

Single imputation vs multiple imputation

- All of these methods are single imputation methods - only one value is used for the missing case.
- The flaw with all of these methods is the issue with standard errors.
- Multiple imputation addresses these issues by creating many copies of the dataset, each with its own set of imputed values, then combining information from these datasets later.
- Although multiple imputation has founds its way into most areas of methodological statistical research, it actually originated from the area of survey sampling.

Basics of multiple imputation

- Multiple imputation can be divided into three steps
  1. Generate M copies of the missing data (Y^m_{mis}) from the a model for the predictive distribution of the missing data, $Pr(Y_{mis}|Y_{obs}, X)$
  2. Perform your usual complete-data analysis of each completed dataset (using the original observed data Y_{obs} and completing it with each Y^m_{mis}) to obtain estimates of the parameters of your model and associated variances of those estimates
  3. Summarize the results of the M complete data analyses by using a t-distribution that combines two sources of variability: variability within copies and variability between copies

Multiple Imputation

- MI depends strongly on the imputation model ($Pr(Y_{mis}|Y_{obs}, X)$) and getting this wrong (and there's no easy way to make sure you get it right!) can strongly affect your results.
- Level of difficulty to do MI: medium (you can do it, but you'll need to take a course first).
- Though, its worth learning because its one of the best options we have for missing data.

For next time

- Read Chapter 11 in the textbook.
  - Lecture 13: Chapter 11, how to analyze survey data, concluding remarks.
  - Lecture 14: Poster presentations.
- Assignment 4 was due on Wednesday - but I haven't received it from everyone :(
- Check Point 4 is due on Monday, November 24th, at 10 AM but you may hand it in as late as Friday, November 28th at 5 pm without lateness penalty.
- A5 is now posted and due on Dec 8.
- No class next week (Thanksgiving), see you Dec 4th.