Applied Quantitative Analysis II
Lab 9

Eric G. Zhou

Spring 2020

NYU Wagner School of Public Service
NYU Grossman School of Medicine

Cotents today

Missing data

Multiple imputations

Missing data

Missing data is one common issues of quantitative data analysis, also called **attrition**

There are two general types of missing data

(1) data that exists however not include in the analytical sample for some reasons

e.g., respondent refuses to answer the question, unrealistic values, incomplete answers

(2) data that does not easily exist or may not be accessible (less to worry about)

e.g., discrepancies between accounting and internal data, crime motives, aspirations, etc.

Missing data

Reasons for missing data/ attrition also vary

- honest mistakes in responding to a survey (recall bias)
- refusals due to privacy, religious, political reasons
- mobilities, death, health issues

Missing data

President Trump wants to add a question for citizenship to the US Census.

What impact do you think this will have on the quality of Census data?

Missing data

To deal with the missingness, we have to make some assumptions about their missing mechanism. Assume we have a model of $Y_i = \beta' X + \epsilon$, where $X = (x_1, x_2, ..., x_k)$.

Missing at random (MAR)

This means that the missingness in x_i does not depend on its value but may depend on the values of $x_j (j \neq i)$. Formally, we have

$$\Pr[x_i \text{ is missing } | x_i, x_j \forall j \neq i] = \Pr[x_i \text{ is missing } | x_j \forall j \neq i] \tag{1}$$

In this case, if we ignore the missing values and do listwise deletion, our results will not be biased, but will be inefficient.

Question for you all, is the MAR assumption testable? If so, how?

Missing data

Missing completely at random (MCAR)

This is a special case of MAR, which means Y_{obs} is a simple random sample of all potentially observable data values. Formally, we have

$$\Pr[x_i \text{ is missing } | x_i, x_j \forall j \neq i] = \Pr[x_i \text{ is missing }] \tag{2}$$

MACR is violated if those who do not report income tend to be richer, younger individuals than those who do. If satisfied, listwise deletion won't cause biased estimators, would still be less efficient than the original data.

MACR is a strong assumption and is not testable.