Data Imputation

Bolun Zhang

5/2/2021

What about missing data?

It is almost certain when we find a data set, there are many missing information across the data set. How should we deal with such a situation?

I will give some basic intuitions here, and we will google together some solutions to the missing data problem.

BTW: Missing data is THE PROBLEM in current big data practice in the industry. It is complicated, and we will only touch the very surface of it.

We can delete it all?

If our model is simple, for example, when we only have a dependent variable, an independent variable and 2-3 control variables in the model, it will probably be fine to delete all cases with any missing.

However:

- If your model is complicated, for example, you have 10 or more variables, you will probably end with far fewer observations than you expected.
- You want to use information from non-missing variables. Sometimes for a observation, there is only one variable missing, but without imputation you will have to delete it
- If the missing is not completely random, your result will be biased. (For example, the missing on certain problem is related to class status)

The principle of dealing with missing data

If it is the case, you will probably need to impute the missing data.

- Do not impute your dependent variable, unless you have 1000% confidence in doing that and persuade your mentor.
- You can try logic imputation first. It means that you can use data from other variables or data from other sources to impute them.
- Most importantly, the goal of data imputation is not the recover the lost true value of the case. The goal is to achieve a more accurate standard error.

Logic imputation

Logic imputation uses information from other variables to impute the variable we care about. It's safer to do that for core independent variables as well.

Sometimes we can also use other data set to impute the missing values, since related values are originally from outside data set like census data. We will discuss how to merge data sets next week.

Examples of logic imputation

Example 1: education attainment

If we have two variables concerning the education. One is the year of education and the other is the highest degree, and we plan to add the highest degree as categorical variable (we recoded it into two cats: No BA & BA) in the model. There is a possibility that we can use the year of education to impute the missing. For example, we might assume that those with 18 or more years of education have BA degree.

Example 2: location variable

We want to explore the influence of geographic features of the county on individual level outcome, however, some countycode information is missing, while we have the zipcode data. Thus, we can use the zipcode data to retrive the countycode and then find related geographic features.

Mean/mode imputation

Mean/mode two simple methods for everyone to begin with.

- If the variable is continuous, impute the missing with the mean.
- If the variable is categorical, impute the missing with the mode.

As our in class example shows, you need to consider the scope of cases that you need to impute.

Multiple imputation

The package for this task in R is MICE. This package can also be used for mean imputation.

Intuition: We can approximate the missing value using statistic model (Remember, even here we are not interested in the true value). Then we can impute multiple datasets with slightly different values, estimate our model in them, take the average to get a more robust standard error estimation. Usually, imputed more than 3 datasets should be fine.

- for the example, please refer to the code.

Further information

For more information about missing data, you can refer to:

- https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/
- https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/
- https://www.youtube.com/watch?v=gY12FJryF7k (This could also be helpful, but using another dataset)