Report on multiple imputation for annual survey of hours and earnings - resident analysis in Britain from 2002-2010

Adi Pierewan

7 March 2011

Missing data is ubiquitous in social survey and research. To deal with the problem, one of the promising solutions is multiple imputation. This process consists of two initial steps: imputing datasets and analysing datasets. The advantage of multiple imputation is the entirely process can be separated into those process, it means that analysts can impute datasets in different time with different software. This report aims to provide the process of imputing datasets for two datasets on annual survey of hours and earnings provided by Office for National Statistics. Although the number of missing data in those datasets is relatively small, but we need to impute the datasets to cover entire districts in the UK. This reports is organised by software I used: STATA, Mplus and Amelia II, explaining the throughout process including the problems and solutions. Lastly, I notice some lesson learned from the process.

1 STATA

At the first time I used STATA especially mi procedure to impute the missing data, using multiple imputation approach. I follow the standard procedure provided by STATA reference that is MULTIPLE IMPUTATION reference. From beginning until the end of the process, there is no problem. But, when I checked again on the imputed datasets, I found that there seem to be missing data. The missing data is not in each

percentile or element in every district, but in some entirely districts. The districts that have a complete data from percentile 10 to percentile 90 tend to be dissapeared. Then I used another package that provides multiple imputation that is ice. When using ice, STATA can generate more completed datasets. I generate complete datasets for data of earnings in 2002-2010, but the missing data is still exist for data of earnings in 2006-2006. There are 20 missing data of 1900 units, because of the unit missingness. In addition, I have checked the tendency of increasing monotone using assert command in STATA. The result is all data are in satisfactorily conditions.

2 Mplus

To solve the problem occured in STATA, I used another software that is Mplus. This software provides Bayesian estimation for imputing data. In addition, Mplus allows us to impute data without model, therefore I choose this software to overcome the problem occured in the previous step. Then, I started imputing the datasets. As expected, the imputing process generates five complete imputed datasets. In addition, when checking the entire datasets, I found the some imputed data do not make sense. Because the software produces the data in percentile 20 is larger than that in percentile 40, for example. Because the results are realtively odds, I try another estimator and specification. Using ML estimator and treating mean as outcome, Mplus can produce completed datasets with increasing monotonone when I test using assert command.

3 Amelia II

By experiencing some softwares that provide multiple imputation procedure, I try to use Amelia II package which works under R statistical software. As mentioned on the Amelia II website, this software is never crashed. In additon, some users recommend to use this package to impute the missing data. This package simply works by using both GUI and R Console. First, I try to impute by using GUI and the results are satisfactorily. The five imputed datasets are compeletely filled. Moreover,

unlike Mplus results, this package generates make sense datasets where the results are increasing monotone, the larger percentile, the bigger number. For checking these results, I used assert comman. Second, I try to impute using R Console, but error occurs whereas the code for imputing procedure is relatively simple. I try to find solution to use Amelia via R console. The additional feature that is beneficial for users is the flexibility of either input data we have or output data we need, ranging from excel, tab-delimited to STATA 10.

4 Lessons learned

This is a useful process which is insightful about missing data, multiple imputation and MI software. I have learned that the problem of missing data can be solved by identifying the structure of datasets. When we need to impute the simple imputation, we need a software that provides a simple model as well. It means that not all of the softwares are suitable for particular case of missing data. We need to be careful to decide which software that more suitable for our need.