Lab #2

Biostatistics 210

NOT FOR HANDIN

0. BACKGROUND

This lab is meant to give you further understanding of the missing at random assumption and how it can be handled by multiple imputation.

1. DOWNLOAD

•Download the dataset lab2.dta. The data here is faked so you can have the experience of look at different scenarios for missing data and seeing how different patterns of missingness affect the complete case analysis

•dead (1=died, 0 = survived)

•copper_50 (serum cooper in 50 mg/dL units)

•bilir_2 (bilirubin in 2 mg/dL units)

•hepatomegaly (1=present, 0 = absent)

•spiders (presence of spider angioma, 1=present, 0 = absent)

•spiders_mar (a version of spiders with missing values generated by MAR missingness)

•miss_mar = (1 = spiders is missing at random, 0 = spiders is observed)

2. EXPLORING MISSINGNESS

What fraction of spider values are missing?

1. What fraction of non-missing data has spiders present: how does that compare with the actual spiders data

2. Can you see any association between other data points and whether the spiders is missing?

When you do this look at hepatomegaly, death and their interaction

3. COMPLETE CASE ANALYSIS

The underlying data has been generated to fit a logistic regression model with OR: 1.5 for copper_50, 1.5 for bilir_2, and 2.0 for spiders and hepatomegaly. Verify this on the full data

logistic dead spiders hepat bilir copper

Examine how the complete case analysis performs:

logistic dead spider_mar hepat bilir copper

How do the results compare to the true odds ratios which generated the data? Can you explain this?

4. MULTIPLE IMPUTATION

The missingness in the dataset was generated to be a function of the interaction of hepatom and death. You can verify this by looking at a logistic model

gen prod = hepat*dead

xi: logistic miss_mar hepat dead prod

Now, let's look at how MI performs when we fully understand the MAR scenarios versus when we don't

First, declare the data as an mi set

```
mi set mlong
```

Then we must declare the variables to be regular or imputed. Since spiders_mar is the only missing one we are focused on, we declare it as needing imputing

```
mi register regular copper_50 bilir_2 hepatom dead prod

mi register imputed spider_mar
```

Now, let's try imputation based on knowing the real model for missingness which involves only death, hepatomegaly and the interaction

```
mi impute logit spider_mar = dead hepat prod, add(50) rseed(333) replace
```

Now, fit the logistic regression model with imputed values of spiders

```
mi estimate, or: logistic dead hepat spider_mar copper_50 bilir_2
```

Make sure to note these results for a comparison later.

The cleanest way to perform a second imputation is the `mi exact` command. First, I suggest saving this dataset.

```
save imputeddata1, replace
```

Now, clear the data. Type

```
mi extract 0, clear
```

Start over with declarations

```
mi set mlong

mi register regular copper_50 bilir_2 hepatom dead

mi register imputed spider_mar
```

In the previous imputation, we had an ideal imputation model because I told you to the real model used to generate missingness and we used only those variables to generate imputations. In practice, we would not know this model. Now, try imputations based on putting in all the other outcome and predictors (but forgetting the interaction).

```
mi impute logit spider_mar = dead hepat copper_50 bilir, add(50) rseed(333)
```

And fit the model on the imputed data

```
mi estimate, or: logistic dead hepat spider_mar copper_50 bilir_2
```

5. IMPUTATION RESULTS

Does the first imputation model appear to give results which match the complete case analysis? Does it approximate true odds ratios (hepat & spider = 2, copper_50 & bilir_2 = 1.5)?

Does the second imputation model appear to give results which match the complete case analysis? Does it approximate true odds ratios (hepat & spider = 2, copper_50 & bilir_2 = 1.5)?