

Lab #2

Biostatistics 210

0. BACKGROUND

This lab is meant to give you practice using multiple imputation to analyze data using the missing at random assumption.

1. DOWNLOAD

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

- dead (1=died, 0 = survived) — the outcome variable for the data analysis

- copper_50 (serum copper 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)

The analytic objective is to relate the factors above the risk of death

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 for death 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 (data with no missing values)

```
logistic dead spiders hepat bilir copper
```

Examine how the complete case analysis performs for the data with missing value

```
logistic dead spider_mar hepat bilir copper
```

How do the results compare to the true odds ratios which generated the data (given above)? 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
```

```
logistic miss_mar i.hepat#i.dead
```

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 imputation.

```
mi register regular copper_50 bilir_2 hepatom dead
```

```
mi register imputed spider_mar
```

Now, let's try imputation excluding the outcome — using only the predictors and the outcome (the variable “dead”)

```
mi impute logit spider_mar = hepat copper_50 bilir, 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 note of these result we will compare them to imputations from another model.

The cleanest way to perform a second set of multiple imputations is the `mi extract` command. First, I suggest saving this dataset.

```
save imputeddata1.dta, 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 excluded the outcome. Now, try imputations based on putting in all the other outcome (the variable “dead”) and predictors.

```
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)?