

Lab #5

Biostatistics 210

0. BACKGROUND

This lab is meant to give you an introduction to multiple imputation. The data is the same data you worked with in Lab #4 last week.

1. DOWNLOAD

•Download the model building dataset lab4.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)

2. MISSINGNESS MODEL

Let's focus on the 3rd missingness scenario. Here

`spiders_miss3` the available spiders values for scenario 3

`miss3` (1 = spider data is missing, 0 = not missing)

The missingness here 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 miss3 hepat dead prod
```

Note, the outcome here is whether spiders is missing in scenario 3 and we see it is highly related to the interaction between death and hepatom.

Our interest is in looking at 4 predictors of death: hepatom, spiders (I include the fully observed variable), copper_50 and bilir_2

Examine the results if no data were missing

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

and the results of the complete case analysis under scenario #3

```
logistic dead spider_miss3 hepat bilir copper
```

Recall, the OR for hepatomegaly is greatly inflated.

3. IMPUTATION

First, declare the data as an mi set

```
mi set mlong
```

Then we must declare the variables to be regular or imputed. Since `spiders_miss3` is the only missing one we are focused on, we declare it as needing imputing

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

```
mi register imputed spider_miss3
```

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_miss3 = 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_miss3 copper_50 bilir_2
```

Make sure to note these results for a comparison later.

The cleanest way to perform a second imputation appears to be clearing memory and re-reading in the data.

Type

```
clear all
```

Now, click on lab4.dta and bring the data back into Stata. Start over with declarations

```
mi set mlong
```

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

```
mi register imputed spider_miss3
```

In the previous imputation, we had an ideal imputation model because I told you 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_miss3 = dead hepat prod, add(50) rseed(333) replace
```

And fit the model on the imputed data

```
mi estimate, or: logistic dead hepat spider_miss3 copper_50 bilir_2
```

3. IMPUTATION RESULTS

Q1: 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)?

Q2: 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)?

Q3: What is the relative efficiency of model 2 to the correct imputation model (the first imputation you did) and to the analysis with no predictors missing? What do you conclude about model strategies for imputation?

Variable	Rel Eff Model 2 v. "Correct Imputation" Model	Rel Eff Model 2 v. Complete Data
hepat		
spider_miss3		
copper_50		
bilir_2		