

Homework # 4

Biostatistics 210

DUE November 3, 2015 to Barbara Grimes

This assignment will give you practice in doing multiple imputation (MI) for missing data in a repeated measures study.

This is based on the HERS example data which you used in the previous homework.

The dataset considers a measures of systolic blood pressure (SBP) measurements at a baseline visit and year 1. In truth, there is no average change in SBP from baseline to year 1. However, members of the cohort with a high baseline SBP are more likely to have a missing measurement at year 1. The missing data mechanism is missing at random (MAR) with respect to some baseline variables. The text showed that GEE performs poorly when applied to the dataset with missing data. You will try combining GEE with MI.

1. DOWNLOAD

Download the datasets `hwk4_long.dta` from the course website. The variables are

- `pptid` (participant ID)
- `sbp` (systolic blood pressure)
- `visit` (0=baseline, 1 = year 1)
- `miss_mar` (indication of constructed missingness, 1=missing, 0 = not missing)
- `basesbp` (baseline SBP)

2.ANALYSIS

Question 1:

Verify that the average baseline SBP is different for participants with missing values.

Use the command `ttest basesbp if visit==1, by(miss_mar)`.

- (a) What is the average baseline SBP (`basesbp`) for someone with a measured SBP at year one (`miss_mar==0`).
- (b) What is the average baseline SBP for someone with a missing SBP (`miss_mar==1`)?
- (c) Also, compare the year 1 SBP among missing and measured values. Use the command `ttest sbp if visit==1, by(miss_mar)`

Question 2:

Using the long format data, "hwk4_long.dta" fit a GEE model with a predictor for visit. Fit first a model to the complete data

```
xtgee sbp i.visit, i(pptid) corr(indep) robust
```

Then fit a model to the non-missing data

```
xtgee sbp i.visit if miss_mar==0, i(pptid) corr(indep) robust
```

What is the estimate of the average change (estimated by GEE) in SBP from baseline to year 1 in the *complete data*?

How does that compare with the average change (estimated by GEE) in SBP from baseline to year 1 using the data *which is not flagged as missing*.

3. DOWNLOAD

Download the datasets `hwk4_wide.dta` from the course website. It is based on the same dataset but it is the “wide format” with one person per line and includes some additional variables. The variables are

- `pptid` (participant ID)
- `sbp0` (systolic blood pressure at baseline)
- `sbp21` (systolic blood pressure at year 1 -- possibly missing)
- `miss_mar` (1= if `sbp1` missing, 0 = if `sbp1` is not missing)
- `bmi0` (systolic blood pressure at baseline)
- `bmi1` (systolic blood pressure at year 1 -- possibly missing)
- `baseline_dm` (diabetes reported at baseline visit)

4. MULTIPLE IMPUTATION

Types a series of command which will set your data up for multiple imputation

```
mi set mlong
mi register imputed sbp1 bmi1 bmi0 sbp0 baseline_dm
```

Note, above I registered all variables as “imputed”. There is no downside to registering a variable as needing imputation if it does not.

If you chose to impute sbp1 using all the available variables then you would type the command below.

```
mi impute chained (reg) sbp1 (reg) bmi0 (reg) bmi1 = sbp0
baseline_dm, add(100) rseed(668) dots
```

It is possible to examine summary of the first five multiple imputations using “mi xeq”

```
mi xeq 1/5: table miss_mar, c(mean sbp1)
```

Question 3: How do the values for SBP which have been imputed compared to the observed values? Do they seem right? Recall, that in question 1 you saw the mean of the actual values which have been artificially deleted.

You can fit a GEE model to the imputed data with the command

```
quietly mi reshape long sbp bmi, i(pptid) j(visit)

mi estimate, dots: xtgee sbp i.visit, i(pptid) corr(independent)
robust
```

Question 4: What is the estimate of the average change (estimated by GEE) in SBP from baseline to year 1 after using multiple imputation?

Question 5: Compare this to the MI results applied to a xtmixed model. (`mi estimate, dots: xtmixed sbp i.visit || pptid:`). Also try different models for the imputation (the “mi impute”) step. How do results change?