

Homework # 3

Biostatistics 210

DUE October 28, 2014 to Barbara Grimes Box in Mission Hall

This assignment will give you practice in examining repeated measures data with missing values.

This is based on the HERS example data in Chapter 11 of VGSM.

The dataset has 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. This assignment will have you explore the nature of the missing data and you will compare mixed model and GEE analyses.

This is an invented dataset and one which give the complete data but also includes a variable “missing”. A value of one for “missing” is used to flag observations for which SBP at year 1 (sbp2) would be missing. This data allows you to explore how analyses on a complete dataset with no missing values for sbp2 would compare to analysis based on the available data which is the dataset for which missing == 0.

1. DOWNLOAD

Download the datasets `hwk3_wide.dta` from the course website. It is in the “wide format” with one person per line. The variables are

- `pptid` (participant ID)
- `sbp1` (systolic blood pressure at baseline)
- `sbp2` (systolic blood pressure at year 1 -- possibly missing)
- `missing` (indication of constructed in Year1 SBP missingness, 1=missing, 0=not miss)
- `bmi1` (body mass index at baseline)
- `bmi2` (body mass index at year 1 -- possibly missing)
- `baseline_dm` (diabetes reported at baseline visit)

2. ASSUMPTIONS

Question 1: (a) Using the wide format data, `hwk3_wide.dta` examine the association between missing data at year1 (`missing=1`) with the predictors collected at the baseline visit. Summarize what you find. Are baseline factors associated with a missing blood pressure at year 1? (b) Does year 1 BMI appear to be associated with a missing value of blood pressure at year 1?

Question 2: Classify the following scenarios as leading to missingness which is MCAR, MAR or NMAR (and explain your answer).

- (a) missingness only depends on baseline SBP
- (b) missingness only depends on baseline diabetes
- (c) missingness only depends on BMI at year 1
- (d) missingness only depends on SBP at year 1

Question 3: Fill in the following table. Recall, you have the complete data and can take out the missing values by analyzing the subset with non-missing value (if `missing==0`)

	Mean SBP
Baseline	
Year One (All Data)	
Year One (Non-Missing Data)	

What is the mean change in SBP from baseline to year 1 with the complete data? What is it using only the non-missing values at year 1?

How do you explain the difference?

3. DOWNLOAD

Download the datasets `hwk3_long.dta` from the course website. The variables are

- `pptid` (participant ID)
- `sbp` (systolic blood pressure)
- `visit` (1=baseline, 2 = year 1)
- `missing` (indicator of constructed missingness, see Ch 11 of VGSM, pp 386-391)
(1=missing, 0 = not missing)
- BMI (body mass index)
- `baseline_dm` (diabetes reported at baseline visit)

This is the same data but in the “long” format rather than the “wide” format which makes it more amenable to GEE and mixed models.

4 . GEE

Using the long format data, “`hwk3_long.dta`” fit a GEE model with a predictor for `visit`. and an exchangeable correlation. 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 missing==0, i(pptid) corr(indep) robust
```

Question 4: 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 with non-missing values? Why are they different?

5 . LINEAR MIXED MODEL

Using the long format data, "hwk3 long.dta" fit a linear mixed model with a predictor for visit and an exchangeable correlation. Fit first a model to the complete data

```
xtmixed sbp i.visit || pptid:, ml
```

Then fit a model to the non-missing data

```
xtmixed sbp i.visit if missing==0, || pptid:, ml
```

Question 5: How to the mixed model results compare between missing and non-missing data?

How do the GEE and mixed models compare in the presence of the missing data? Why do they differ?