Aarhus November 19 2013

Applied Statistical Analysis with Missing Data

Exercise 4: The drug study

The drug study involves 106 patients assigned to one of two drugs (DrugE and DrugN). Each patient had a baseline health-score just before assignment and follow up scores at week 13, 14,…, 24 and 38.

The data is available in a "long" format (`DrugStudyLong.dta`) as well as in a "wide" format (`DrugStudyWide.dta`).

```
. use DrugStudyWide, clear
. codebook, compact

Variable   Obs Unique      Mean  Min  Max  Label
-----------------------------------------------------------------------
pid        106    106      53.5    1  106  group(id)
drug       106      2  .6698113    0    1
sex        106      2  1.632075    1    2
age        106     41  43.16038   20   69
duration   106     39  22.26415    3   80  no week of last episode
episodes   106      3  1.877358    1    3  no of episodes
score0     106     32  16.99421    3   36  0 score
score13     95     30  15.54222    0   34  13 score
score14     87     33  14.88085    0   35  14 score
score15     94     34   14.1357    0   29  15 score
score16     83     31  13.80675    1   29  16 score
score17     73     27  12.95373    1   29  17 score
score18     79     27   11.7889    0   27  18 score
score19     72     24  10.59194    0   26  19 score
score20     75     27  11.48658    0   29  20 score
score21     66     27  10.63152    0   27  21 score
score22     66     28  10.27713    0   28  22 score
score23     60     24  10.45725    0   26  23 score
score24     67     27  10.32657    0   33  24 score
score38     39     18  6.105641    0   19  38 score
-----------------------------------------------------------------------
. use DrugStudyLong, clear
. codebook, compact

Variable    Obs Unique      Mean  Min  Max  Label
-----------------------------------------------------------------------
pid        1484    106      53.5    1  106  group(id)
week       1484     14  18.57143    0   38
score      1062     64  12.69841    0   36  health score
drug       1484      2  .6698113    0    1
sex        1484      2  1.632075    1    2
age        1484     41  43.16038   20   69
duration   1484     39  22.26415    3   80  no week of last episode
episodes   1484      3  1.877358    1    3  no of episodes
-----------------------------------------------------------------------
```

Part A – missing values

We will first look at the missing values and their association with what we have observed.
One way to do this, out of many, is to generate a binary variable taking the value 1 if the next score is recorded (observed) and 0 if it is missing:

```
. use DrugStudyLong, clear

. bysort pid (week):  ///
  generate NextRecorded = !missing(score[_n + 1])
```

List the data for patient number 6 to see what this command did.

Q1: Is missing score associated with the score the previous week, sex, age or type of drug?

Another way to look at the association between missing and the observations is to examine the number of observed scores for each person:

```
. use DrugStudyWide, clear
. egen NMiss = rowmiss(score*)
. egen NRecorded = rownonmiss(score*)
```

List the data for patient number 6 to examine what you have done.

Q2: Is the number of missing scores associated with the baseline score, sex, age or type of drug?

Part B – Single value imputation

Single value imputation based on a regression model of change from baseline to average of last five scores:

```
. use DrugStudyWide, clear
. generate changelast5            ///
      = (score21+score22+score23+score24+score38)/5 - score0
. regress changelast5 score0 duration i.sex i.drug age episodes
. predict predch, xb
. predict sdch  , stdf
. generate imputed = changelast5
. replace imputed = rnormal(predch, sdch) if missing(changelast5)
```

List the imputed values for the first 9 patients.

Q3: What is the unadjusted difference between the two drugs based on the imputed data?

Q4: What is the difference between the two drugs based on the imputed data, when you adjust for sex and duration?

Note that the results you just found will not properly account for the random variation in the data as they will have too small standard errors.

Part C – Declaring the data as a MI-dataset and making "black box" multiple imputations

First you declare the dataset to be a missing dataset, where you want the imputed values added as extra rows and that missing values are in the score variables

```
. use  DrugStudyWide, clear
. mi set mlong
. mi register imputed score*
. mi register regular pid drug sex age duration episodes
```

Now you can you use the tools for describing the patterns of the missing values:

```
. mi describe
. mi misstable pattern, freq
. mi misstable summarize
```

We want to generate 20 imputations of the missing scores based on a regression model on age, sex duration, episodes, drug and the other scores:

```
. mi impute chained ///
   (regress) score??= age i.sex duration episodes b0.drug, ///
  add(20) dots rseed(DDMM)
```

Read the output generated by Stata, and list some variables for patient 6:

```
list score15 score16 drug sex age _mi_id _mi_miss _mi_m if pid==6
```

If we do not want negative health scores we might set those to zero:

```
. foreach j of numlist 13/24 38 {
     replace score`j'=0 if score`j'<0
 }
```

Now let us save the data set with imputed values in both wide and long format:

```
. save ImputedWide, replace
. mi reshape long score, i(pid) j(week)
. save ImputedLong, replace
```

Part D Analysing the imputed data using the complete data model

The primary end point was the change from baseline to the average of the last five scores and the complete data analysis was `regress changelast5 b0.drug i.sex duration`

```
. use  ImputedWide, clear
. mi passive: generate changelast5=
      (score21+score22+score23+score24+score38)/5-score0
. mi estimate, dots: regress changelast5 b0.drug i.sex
duration
```