Aarhus, November 19, 2013

Applied Statistical Analysis with Missing Data
Exercise 5 (Continuation of Exercise 2 and 3)

Consider the dataset `ess2e03_scand.dta`, cf. Exercise 2 and 3.

First imputation model revisited:

Q1: Assume that data satisfies the MAR assumption and use `-mi impute-` to mimic the imputation you did in Exercise 3, Q6-Q9 with respect to non-compliance, education and income. You should explicitly define and record the relevant prediction equations.

Q2: Investigate the distribution of imputed log-income values (Hint: look at the imputed values as well as the residuals from the model, where log-income is the dependent variable).

Q3: Use a "black box" imputation approach and compare with your results above. Again you should look at the distribution of imputed log-income values.

Q4: Some researchers suggest that age is related to income, education level and non-compliance, as is gender. Incorporate this into your analysis.

Non-compliance as passive variable

Overall non-compliance is actually defined as being either *primary* or *secondary* non-compliant. To view this structure you could cross-tabulate the two variables `primary_noncompl` and `secondary_noncompl` with the missing option:

```
. tab primary_noncompl secondary_noncompl, missing

   Primary |
non-compli |
 ance (Did |
       not |     Secondary non-compliance
   collect |   (Collected, but did not use as
medication |            prescribed)
         ) |        No        Yes          . |     Total
-----------+---------------------------------+----------
        No |     5,761      1,041          0 |     6,802
       Yes |         0          0        371 |       371
         . |         0          0        623 |       623
-----------+---------------------------------+----------
     Total |     5,761      1,041        994 |     7,796
```

Q5: Constructthe generate statement that will define overall non-compliance from the two variables `primary_noncompl` and `secondary_noncompl`. Your resulting variable should be identical to the variable `total_noncompl`.

Q6: Use this statement and the following “trick” to modify your imputation such that primary and secondary non-compliance are imputed, and based on them the overall non-compliance is subsequently derived. The trick is that you may safely set secondary non-compliance to “Yes”, when primary non-compliance is “Yes”.

Q7: Analyze overall non-compliance as above based on this new imputed dataset, and compare with your previous findings.

Imputation stratified by country (optional)

Q8: Use country as a covariate in your analysis of non-compliance. Discuss whether this is reasonable when it was not included in the imputation model.

Q9: Stratify the imputation on country and update the analysis of Q8.