Århus November 19 2013

Applied Statistical Analysis with Missing Data
Exercise 3 (Exercise 2 continued)

Consider again the dataset `ess2e03_scand.dta`, cf. Exercise 2, and the four variables:

cntry	Country of interviewee.
edulvl	Highest level of education
incmean	Annual household income
total_noncompl	Non-compliance at last prescription of a new medication

A first imputation model

Assume that income is Missing At Random, when education is observed.

Q6: Use the model for log-income developed in **Q3** to predict the expected mean and the standard deviation for the missing log income values. This done via the predict command (after the relevant regress command):

```
. predict meanlogincome if missing(incmean), xb
. predict sdlogincome   if missing(incmean), stdf
```

Browse or list the data to see what you have generated.

Q7: Use this to impute a random normally distributed log-income for subjects with a missing income value. (Hint: Use the `-rnomal-`, function)

Q8: Fit a logistic regression of non-compliance on both educational level and income (including the imputed values) ignoring their possible interaction. Interpret the coefficients. Compare with your neighbour.

Q9: Predict the odds of non-compliance for the observations which has missing values for non-compliance. Use this to impute the missing values for non-compliance.

Updating the imputations

Above we ignored the information on non-compliance, when we imputed income. We will now assume that income is related to non-compliance (why is that reasonable?).

Q10: Make a linear regression of log-income on educational level and non-compliance (including the imputed values) - is there an interaction?

Q11: Update your expected mean log-income and its standard deviation and compare this with what you got above in **Q6**. Impute the missing income values based on this.

Analysis of imputed datasets and stratified imputation (optional)

Q12 Repeat the imputations to create five completed datasets, each of which has no missing values. Obtain an estimate for each complete dataset of how non-compliance depends on income, and compute the

average estimate. Note: there is a small prize to the one who gets closest to the "true" value (as defined by HS :-)).

Q13: Should the imputation model depend on country? Does this impact your overall estimate?