Review questions and exercises: Topic 5

Before you start

Make sure you are working in the correct directory: `pwd` . Change to another directory if necessary. Also, you may wish to open a command log and a result log:

```
cmdlog using logname.do
log using logname.log
```

1 About missings

For these exercises, use the hints from the last slide.

1. `display` whether `. > 5` is true or false.
2. Find out what . + . is.
3. Find out what . + .a is.
4. Find out what .a * .b is.

2 Missing values in regression analysis

Use the labour market survey data: `webuse nlsw88, clear`

1. Investigate the missing values in this dataset. Are there any? Can you detect certain patterns of missings?

2. Think about whether the data could be MAR. Describe in words which assumption have to be met for the data to be MAR. Do you believe they are MAR?

3. Disregarding dummy variables, calculate a few bivariate correlations between variables which have missing values. Then calculate the same correlations all in one command. Does the sample size change? And do the results change?

4. Run a regression of wage on age, married, i2.race, south, smsa, collgrad, ttl_exp and union. Store the estimates (`estimates store`). Then re-run the regression but dropping the variable union. Store the estimates again. Compare the two sets of results. (You could use `estimates table` if you wanted to.) Are there any big changes in sign, magnitude or significance of coefficients?

5. (optional) Follow the guidance in the lecture notes to perform multiple imputation on the variable union. Use the command
   `mi impute logit union wage age married i2.race south smsa collgrad ttl_exp, add(20)`
   at the imputation stage. Compare the results to the first model above.