Review questions and exercises: Topic 3

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 Regression with Stata

1.1 Estimation

Estimate the following regression model using the `greeneF41_3` data:

$$\begin{aligned}\text{wife's earnings} =& \beta_0 + \beta_1\text{education} + \beta_2\text{experience}\\ &+ \beta_3\text{experience}^2 + \beta_4\text{kids dummy} + \beta_5\text{marginal tax rate}\\ &+ \beta_6\text{husband's education} + \beta_7\text{husband's mean hourly wage} + \text{ error}\end{aligned}$$

The variables have already been generated for you.

A constant term is added automatically unless you suppress it with an option. Find out what results are left behind for you by typing `ereturn list` after the regression.

1.2 Interpretation

Interpret the results. As a guide, consider the following questions:

1. Does the constant term have a meaningful interpretation?
2. What is the estimated effect of experience on earnings?
3. How can you interpret the dummy variable for kids?
4. Can you compare the magnitude of the coefficients?
5. Which coefficients are statistically significant?
6. Compare your results with the results from estimating the smaller model in the slides. Comment on the most obvious differences.
7. Re-play your regression to show standardized coefficients:

   ```
   regress, beta
   ```

   Standardized coefficients are identical to coefficients from a regression where all variables are *standardised* and no constant is included. How can you interpret the coefficients now?

1.3 Hypothesis testing

Typing `regress` on its own will replay your original (non-standardised) regression results.

For all hypothesis tests, state null and alternative hypotheses, the distribution of the test statistic under the null hypothesis, as well as your decision to reject/not reject.

1. Test the hypothesis that the coefficients on experience, experience squared and kids are jointly equal to zero.

2. Test the hypothesis that all coefficients apart from the constant term jointly equal zero.

3. Test a non-linear hypothesis of your choice using `nlcom`.

4. How can you test the null hypothesis that the marginal effect of experience on wife's earnings, evaluated at 30 years of experience, is equal to 0.1?

1.4 Variance inflation and correlations

Look at the variance inflation factors using `estat vif`. Also look at the correlations between the independent variables. Are you worried about any of them?

1.5 Predictions

Stata can calculate a number of predictions after regressions. Type `help regress` to find out more.

Generate predictions for residuals, studentised residuals, leverage and Cook's D with the following commands (be sure to re-run your regression first if you have estimated something else in between, or recall your stored results):

```
predict e, resid
```

```
predict t, rstudent
```

```
predict h, leverage
```

```
predict d, cooksd
```

Inspect the distribution of the residuals, e.

1.6 Heteroscedasticity

Create a Residual-versus-fitted plot (`rvfplot`). Do you find evidence for heteroscedasticity, or other unusual "behavior" of the residuals? Look at some Residual-versus-predictor plots, too.

Perform one or more tests for heteroscedasticity. State the null hypothesis, alternative hypothesis, distribution of the test statistic under the null and your conclusion.

1.7 Outliers and influential data

Take a look at the three variables t, h, and d. How are they distributed? Which observations have outstandingly large (absolute) values?

Investigate a plot of squared normalised residuals versus leverage: `lvr2plot` . Are you worried about one or more particular observation(s)?

Use `dfbeta` to find out by how much individual coefficient estimates are affected by specific observations.

1.8 Other tools for regression diagnostics

Type `help regress postestimation` to find out what other tools Stata offers out of the box. You can play around with added-variable plots and component-plus-residual plots. The latter help you in finding nonlinearities in your data.