Lab 2: Simple linear regression

Data: reach.csv (see the reach.pdf file for data dictionary/useful information).

Practical objective: To gain familiarity with implementation of simple linear regression in Stata. Note that this lab differs from Lab 1 in that it is primarily exercise-based rather than tutorial-based.

Scientific objective: To investigate the association between baseline HbA1c and six-month HbA1c.

Noteworthy commands: Below is a list of Stata commands that will be particularly helpful for this lab.

- `loess`
- `regress`
- `lincom`
- `display`
- `local`
- `lvr2plot`
- `predict`
- `qnorm`

Exercises: Below is a set of exercises that we will go through individually, in small groups, and/or together as appropriate and as time permits.

Exercise 1: Load the REACH data set into Stata. In this lab, we will only analyze the REACH subjects (irrespective of receipt of FAMS). Therefore, drop the control subjects from the data set.

Exercise 2: Write down a simple linear regression model that characterizes the relationship between baseline HbA1c and six-month HbA1c among patients receiving REACH. Provide literal interpretations for each of its coefficients. Which coefficients possess a real-world interpretation?

Exercise 3: Construct a scatter plot of baseline HbA1c and six-month HbA1c with a LOESS curve and retain a copy of the scatter plot for later use. Use Stata to produce estimates and 95% confidence intervals for the most meaningful coefficient of the regression model you previously proposed in Exercise 2.

Exercise 4: Construct a point estimate and 95% confidence interval for the difference in mean six-month HbA1c between individuals differing in their baseline HbA1c by 0.4%. You may find the `lincom` command helpful, though you can also use the `display` command.

Exercise 5: Construct a point estimate and 95% confidence interval for mean six-month HbA1c among individuals with a baseline HbA1c of 8.0%.

Exercise 6: What assumptions are required for you to trust that your results of Exercise 3 are leading you to a scientifically meaningful conclusion? Use diagnostics to assess the degree to which any necessary assumptions appear to be reasonable in this problem.

Exercise 7: One diagnostic plot that we have not yet discussed is a leverage vs. squared residual plot. This can be constructed in Stata using the post-regression command lvr2plot. Identify the point on the plot with the highest leverage and map it to a point on the scatter plot of Exercise 3. Is it likely to be an influential point? Confirm your suspicion by running a sensitivity analysis in which you exclude the highest-leverage observation. You may find the following command particularly useful: `predict lev, leverage`.

Exercise 8: What assumptions are required for you to trust your ability to use the model of Exercise 3 to form prediction intervals/reference ranges for HbA1c? Use diagnostics to assess the degree to which any necessary assumptions appear to be reasonable in this problem.

Exercise 9: Irrespective of your answer to Exercise 8, form (and compare) prediction intervals for six-month HbA1c among REACH subjects with a baseline HbA1c of 4.5%, 6.0%, 12.0%, and 15.0%.