Econ 453, Spring 2025, Problem Set 3+4

DUE: April 23 at 11:59pm, on D2L

Instructions

- Submit your R code along with your typed up answers.
- Upload your answers and code as a single PDF file to D2L. Please do not e-mail or submit hard copies.
- The necessary data files are available on D2L in an Excel file.

Question 1

The data are on D2L in pset34_data.xlsx, sheet "earnings". This is a cross-sectional data on earnings, education, age and region.

1. Create these new variables in your dataset: age squared, education squared, dummies for each region, natural log of wages.
2. Run a linear regression of log wages on age, education and dummies for region. (Hint: there should be 3 dummies with one excluded in the regression.) Which coefficients are statistically significant at the 1% level. Show the F-stat for the joint significance of all variables in the model. Is this F-stat significant at the 1% level?
3. Run a linear regression of wages on age, education and dummies for region. Which model fits better, the model where wages is the dependent variable or the model where log wages is the dependent variable?
4. Add age squared to the model where log wages is the dependent variable. Does adding age squared improve the fit? Run an F-test on whether the model with age squared is statistically different than the restricted model without age squared.
5. Add education squared to the model where log wages is the dependent variable. Does adding education squared improve the fit? Run an F-test on whether the model with education squared is statistically different than the restricted model without education squared.

Question 2

The data set has cross-sectional data on individual wages, a dummy for graduate degree and age of the individual. The data are on D2L in pset34_data.xlsx, sheet "wages".

1. Draw scatter plot for the data with age on x-axis and wages on y-axis. What can you infer from this chart?

2. There are 160 observations in the dataset. Partition the sample into two sub-samples: the first 120 observations (training set) and the last 40 observations (validation set). Using the first 120 observations, fit two regression models. In model 1, regress wages on graduate and age. In model 2, regress wages on graduate, age, and age-squared. Summarize estimated coefficients, standard errors, p-values, R^2, and Adjusted R^2 from both models in a well formatted table. Which model would you choose? Justify your choice.

3. Using estimated models, predict expected wages for a 30-year-old individual with a graduate degree.

4. According to estimated model 2, at what age are wages at the maximum?

5. Calculate the mean standard error and mean absolute error for both models over the training set and over the validation set. Which has the best in and out of sample fit?

Question 3

Complete Exercise 2 from Chapter 7.1. Show which equations/concepts you used to reach your answer.

Question 4

Complete Exercise 18 from Chapter 7.2. Show which equations/concepts you used to reach your answer.