Worksheet week # 10

1. Use data *wage.gdt* to estimate the returns to education equation.

   (a) Estimate the baseline model of the impact of education and experience on wages:

   $$ln(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \varepsilon_i \ .$$

   Interpret the meaning of the coefficient β_1.

   (b) Reestimate the model using robust standard errors, comment on the differences.

   (c) Test for heteroskedasticity in the model in part (a). Is it necessary to use robust standard errors in this case?

   (d) Estimate the model with quadratic specification of experience:

   $$ln(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + \varepsilon_i \ .$$

   Comment on how and why the coefficient β_2 changed with respect to part (a). Did the coefficient β_1 change as well? Why or why not?

   (e) Do you believe that the coefficient β_1 is correctly estimated? Is there anything that could create a bias in this equation? If yes, how would you solve for this problem?

   (f) Include in the model the education of the mother and of the father of the observed individuals:

   $$ln(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + \beta_4 motheduc_i + \beta_5 fatheduc_i + \varepsilon_i \ .$$

   i. Is there an impact on the coefficient β_1? Does this signal there was a bias in the model from part (d)? Comment on the sign of this bias.

   ii. Are both *motheduc* and *fatheduc* individually significant? Are they jointly significant?

   iii. What happens if you exclude one these variables from the regression? Which one would you keep?

   (g) Instead of the education of parents, include the variable measuring ability in the model:

   $$ln(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper^2 + \beta_4 abil + \varepsilon_i \ .$$

   Is there an impact on the coefficient β_1? Does this signal there was a bias in the model from part (d)? Comment on the sign of this bias.

2. Use data *bwght1.csv* to estimate the model describing the impact of smoking on birth weight. The file contains data on births to women in the US.

bwght ... birth weight in ounces
cigs ... cigarettes smoked per day while pregnant
packs ... packs of cigarettes smoked per day while pregnant

(a) Import data into Gretl from csv file.

(b) Generate log of birth weight (using $logbwght = log(bwght)$).

(c) Define and estimate the model describing the impact of smoking on birth weight (in logarithm).

   i. Should we include both *cigs* and *packs* into the model?
   ii. Which assumption might be violated if we do?
   iii. How do we test for this assumption?

(d) Interpret the results of *cigs* and *packs* coefficients in the two separate models. Are they significant? What is the interpretation of the coefficients magnitude?