PO 7005: Assignment 3

Thomas Chadefaux

NOTE: Always justify your answer. Show R code when relevant. Late submissions will be penalized (5 points per day). Each question is equally weighted.

1. The dataset 'uswages.dta', included in the same folder as this homework, provides an extract on weekly wages for US male workers sampled from the Current Population Survey in 1988.[1] Import that dataset into R. Hint: use 'library(foreign)' and 'read.dta'. Report the number of observations. Plot a matrix of scatterplots of wage, education and experience (hint: use the function 'pairs()').

2. Estimate the following model

$$Wage_i = \beta_0 + \beta_1 education_i + \varepsilon_i. \tag{1}$$

   (a) Interpret your estimate of the coefficient on education (β_1). Does it have the expected sign?

   (b) Is β_1 statistically significant? And at what level? What does it mean?

   (c) Report R^2 and interpret it.

3. Now estimate

$$Wage_i = \beta_0 + \beta_1 education_i + \beta_2 experience_i + \beta_3 race_i + \beta_4 smsa_i + \beta_5 pt_i + \varepsilon_i \tag{2}$$

   (a) Interpret β_3, the coefficient on race.

[1] A description of the data and variables is at `http://www.joselkink.net/wp-content/uploads/2013/01/uswages.txt`.

(b) How has β_1 changed compared to model 1? Why?

(c) How would you test the hypothesis that $\beta_2 = 50$?

4. What is model 2's R^2. How does it compare to model 1's? Does it mean that model 2 is a better model?

5. Compute an F-test to determine whether the non-restricted model (2) improves upon the restricted model (1). You may wish to compute it manually, or to use R's command 'anova(lm1, lm2)'. Interpret your results.

6. Determine which is a better model by comparing their out-of-sample predictions. I.e., split the sample in two, estimate the model on the first half, and make predictions for the second half. How do your results compare?

7. Report model 1 and model 2's AIC and BIC. Which one is preferable?

8. Suppose now that you want to test the hypothesis that geography has no impact on wage (i.e., that where people live does not affect their income). How exactly would you test this hypothesis? I.e., which variables would you include, and what test would you conduct?

9. How would you test the hypothesis that the effect of education on wage is lower for 'black' individuals than for 'white' ones. Run the test, report your code and results, and interpret the results. Given your modified model, what is the effect of one additional year of education on wage?

10. You hypothesize that the effect of experience on wage is quadratic. How would you modify model (2) to reflect it? Was the modification justified? Explain why or why not.

11. Now estimate

$$log(Wage_i) = \beta_0 + \beta_1 education_i + \beta_2 experience_i + \beta_3 race_i + \beta_4 smsa_i + \beta_5 pt_i + \varepsilon_i \quad (3)$$

12. Interpret the coefficient on education, β_1

13. Plot a histogram of the residuals of model 2 and model 3. What is the difference? Which one better satisfies the Gauss-Markov assumptions?

14. Test whether the variance of model 3's residuals is homoskedastic. In particular, perform three tests.

    - The Goldfeld-Quandt test (perform this test for the variable education).
    - The Breusch-Pagan test
    - The White test

    Do you conclude that you have a problem of heteroskedasticity? Justify.

15. Assuming you want to correct for heteroskedasticity, you want to implement White's robust standard errors. Compute and report them for models 1, 2 and 3. How did the standard errors change, and why?

16. Report all your results for model 1, 2 and 3 in a single table that would be suitable for publication, both in content and in appearance. I.e., include coefficients, standard errors, significance level and relevant explanations needed to understand what you did. Format it as you would for submission to your favourite social science journal (say, APSR or AER).