14.310x: Data Analysis for Social Scientists - Homework 9

Welcome to your ninth homework assignment! You will have about one week to work through the assignment. We have provided this PDF copy of the assignment so that you can print and work through the assignment offline. You can also go online directly to complete the assignment. If you choose to work on the assignment using this PDF, please go back to the online platform to submit your answers based on the output produced.

Good luck :)!

For the following questions, you will need the data set: nlsw88.csv. The data has information on labor market outcomes of a representative sample of women in the US. It contains the following variables: the logarithm of the wage (*lwage*), total years of schooling (*yrs_school*), total experience in the labor markets (*ttl_experience*), and a dummy variable that indicates whether the woman is black or not. Since we are going to work with this data throughout this homework, please load it into R using the command `read.csv`

As a first step, we are interested in estimating the following linear model.

$$log(wage_i) = \beta_0 + \beta_1 yrs\_school_i + \varepsilon_i \tag{1}$$

Estimate this equation by OLS using the command `lm`. Please go to the documentation in R to understand the syntax of the command. Now, based on your results answer the following questions.

1. According to this model, what is the estimate of β_1?

2. What is the 90% CI of $\hat{\beta}_1$ according to this model?

   (a) It is given by $[0.08174972, 0.1040900]$

   (b) It is given by $[0.08736549, 0.09847428]$

   (c) It is given by $[0.08442308, 0.1014167]$

   (d) It is given by $[0.08579005, 0.1000497]$

3. Assume that instead of having all the data you just know that the covariance between the logarithm of the wage and the years of schooling is 0.6043267. What other piece of information would you need to be able to find $\hat{\beta}_1$?

   (a) The sample variance of the variable *yrs_school*

   (b) The sample variance of the variable *lwage*

   (c) The sample variance of the error term

   (d) The sample covariance between the error term and *yrs_school*

4. After running your code what is the value you found for $\hat{\beta}_0$?

5. True or False: For any single linear regression model it is true that the predicted value when $x = \overline{x}$ is $\overline{y}$?

(a) True

(b) False

6. After running your model, use the command `residuals` to calculate the residuals of the regression. Calculate the sum of the residuals. Should we be surprised that the sum is so close to zero?

   (a) Yes

   (b) No

   Now, we are interested in estimating the following model:

$$log(wage_i) = \beta_0 + \beta_1 black + \varepsilon_i \tag{2}$$

7. Researcher A says that this model is not correctly specified. Researcher A suggests that the correct model should estimate the following equation (where `other race` is a dummy variable equal to 1 when the person is not black):

$$log(wage_i) = \beta_0 + \beta_1 black + \beta_2 other\ race + \varepsilon_i$$

   Researcher B claims that researcher A is wrong and that in this model is not possible to separately identify β_0, β_1, and β_2. Would you agree with researcher A or with researcher B?

   (a) Researcher A

   (b) Researcher B

8. Assume that you don't have all the data. However, you know that the sample mean of the log wage for women who are not black is $\overline{y}_{other}$ and the sample mean of the log wage for black women is $\overline{y}_{black}$. What are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ if we run this model by OLS?

   (a) We have that $\hat{\beta}_0 = \overline{y}_{black}$ and $\hat{\beta}_1 = \overline{y}_{other} - \overline{y}_{other}$

   (b) We have that $\hat{\beta}_0 = \overline{y}_{black}$ and $\hat{\beta}_1 = \overline{y}_{black} - \overline{y}_{other}$

   (c) We have that $\hat{\beta}_0 = \overline{y}_{other}$ and $\hat{\beta}_1 = \overline{y}_{black} - \overline{y}_{other}$

   (d) We have that $\hat{\beta}_0 = \overline{y}_{other}$ and $\hat{\beta}_1 = \overline{y}_{other} - \overline{y}_{other}$

9. Now, estimate this model by yourself using both the sample means approach or the regression approach with the command `lm`. (You should get the same results!)

   (a) What value did you find for $\hat{\beta}_0$?

   (b) What value did you find for $\hat{\beta}_1$?

10. A critic is claiming that this doesn't prove that there are differences in the wage of black women and women of other race. You decide to conduct a test on the parameter β_1, where the null hypothesis is $\beta_1 = 0$. What is the value of the statistic of this test?

11. Would you reject this null hypothesis using a 99% level of confidence?

(a) Yes

(b) No

Labor economists have estimated Mincer equations that include not only total years of schooling, but also total experience as explanatory variables of the wage. Assume now that you want to estimate the following model?

$$log(wage_i) = \beta_0 + \beta_1 yrs\_school_i + \beta_2 total\ experience + \varepsilon_i \tag{3}$$

12. If you run this model in R what would be the value of the R^2?

Some young folks are claiming that they prefer to drop-out from school since each additional year of schooling changes the log of the wage in the same amount as half year of experience. A group of parents are really worried. They ask you to conduct a formal test over this sample.

13. What would be the null hypothesis of this test?

(a) The null hypothesis of this test is $\beta_1 = 2\beta_2$

(b) The null hypothesis of this test is $\beta_1 = \beta_2 + \beta_1$

(c) The null hypothesis of this test is $\beta_1 + \beta_2 = \beta_2$

(d) The null hypothesis of this test is $2\beta_1 = \beta_2$

14. Which of the following would correspond to the restricted model under this null hypothesis?

(a) The model $log(wage_i) = \beta_0 + \beta_2(yrs\_school_i + 2total\ experience_i) + \varepsilon_i$

(b) The model $log(wage_i) = \beta_0 + \beta_1(\frac{1}{2}yrs\_school_i + total\ experience_i) + \varepsilon_i$

(c) The model $log(wage_i) = \beta_0 + \beta_1(yrs\_school_i + 2total\ experience_i) + \varepsilon_i$

(d) The model $log(wage_i) = \beta_0 + (\beta_1 + 2\beta_2)yrs\_school_i + \varepsilon_i$

(e) The model $log(wage_i) = \beta_0 + (2\beta_1 + \beta_2)yrs\_school_i + \varepsilon_i$

(f) The model $log(wage_i) = \beta_0 + \beta_2\left(\frac{1}{2}yrs\_school_i + total\ experience_i\right) + \varepsilon_i$

15. Estimate the restricted model in R. What is the value that you obtain for $\hat{\beta}_1$ in the restricted model? For the restricted model, use the first correct model in the list from question 15, i.e. if both (a) and (d) are correct, use (a) or if both (b) and (c) are correct, use (b).

16. Use the `anova` command in R to calculate the test $\frac{SSR_r - SSR_u}{r} / \frac{SSR_u}{N-K-1}$, what is the value of the test?

17. Do you reject or not reject this null hypothesis at a confidence level of 95%?

(a) Reject

(b) Do not reject