14.310x Data Analysis for Social Scientists
Single and Multivariate Linear Models

Welcome to your eighth homework 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 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:

$$\log(wage_i) = \beta_0 + \beta_1 yrs\_school_i + \varepsilon_i$$

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

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

Question 2
What is the 90% confidence interval (CI) of $\hat{\beta}_1$ according to this model?

- [0.08579005, 0.1000497]
- [0.08736549, 0.09847428]
- [0.08442308, 0.1014167]
- [0.08174972, 0.1040900]

Question 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 information would you need to be able to find $\hat{\beta}_1$?

- The sample covariance between the error term and *yrs_school*
- The sample variance of the variable *lwage*
- The sample variance of the error term
- The sample variance of the variable *yrs_school*

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

Question 5
True or False: For any simple bivariate linear regression model, the predicted value when $x = \bar{x}$ is $\bar{y}$.

- True
- False

Question 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?

- Yes
- No

Now, we are interested in estimating the following model:

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

Question 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 second model, it is not possible to separately identify $\beta_0, \beta_1,$ and β_2.Who is correct?

- Researcher A
- Researcher B

Question 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 $\bar{y}_{other}$, and the sample mean of the log wage for black women is $\bar{y}_{black}$. What are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ if we run this model using OLS?

- $\hat{\beta}_0 = \bar{y}_{other}$ and $\hat{\beta}_1 = \bar{y}_{other} - \bar{y}_{black}$
- $\hat{\beta}_0 = \bar{y}_{black}$ and $\hat{\beta}_1 = \bar{y}_{black} - \bar{y}_{other}$
- $\hat{\beta}_0 = \bar{y}_{other}$ and $\hat{\beta}_1 = \bar{y}_{black} - \bar{y}_{other}$
- $\hat{\beta}_0 = \bar{y}_{black}$ and $\hat{\beta}_1 = \bar{y}_{other} - \bar{y}_{black}$

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

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

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

Question 10
A critic is claiming that this doesn't prove that there are differences in the wage of black women and women of other races. 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 the t-statistic?

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

- Yes
- 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:

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

Question 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 one half year of experience. A group of parents is really worried. They ask you to conduct a formal test over this sample.

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

- $2\beta_1 = \beta_2$
- $\beta_1 = \beta_2 + \beta_1$
- $\beta_1 + \beta_2 = \beta_2$
- $\beta_1 = 2\beta_2$

Question 14
Which of the following would correspond to the restricted model under this null hypothesis? (Select all that apply)

- ☐ $\log(wage_i) = \beta_0 + \beta_2(yrs\ school_i + 2total\ experience_i) + \varepsilon_i$
- ☐ $\log(wage_i) = \beta_0 + \beta_1\left(\frac{1}{2} yrs\ school_i + total\ experience_i\right) + \varepsilon_i$
- ☐ $\log(wage_i) = \beta_0 + \beta_1(yrs\ school_i + 2\ total\ experience_i) + \varepsilon_i$
- ☐ $\log(wage_i) = \beta_0 + (\beta_1 + 2\beta_2) yrs\ school_i + \varepsilon_i$
- ☐ $\log(wage_i) = \beta_0 + (2\beta_1 + \beta_2) yrs\ school_i + \varepsilon_i$

- $\log(wage_i) = \beta_0 + \beta_2 \left(\frac{1}{2} yrs\ school_i + total\ experience_i\right) + \varepsilon_i$

Question 15
Estimate the restricted model in R. What is the value that you obtain for $\hat{\beta}_1$ in the restricted model?

Note: use the model from Question 14 that defines the restricted model ONLY in terms of β_1.

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

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

- Reject
- Do not reject