Questions 1 - 6

Bookmark this page

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

1.0/1.0 point (graded)

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

Please round your answer to the third decimal point, i.e. if it is 0.12494, please round to 0.125 and if it is 0.1233, please round to 0.123

0.093

✔ **Answer:** 0.093

0.093

Explanation

The command that we should run in R after uploading the data is:

```
#simple linear regression
single <- lm(lwage ~ yrs_school, data = nlsw88)
summary(single) # show results
```

The output that you get after running this code is:

The Multivariate Linear Model
due Nov 28, 2016 05:00 IST

Module 9: Homework
due Nov 21, 2016 05:00 IST

Exit Survey

```
Call:
lm(formula = lwage ~ yrs_school, data = nlsw88)

Residuals:
     Min       1Q   Median       3Q      Max
-2.29340 -0.32611 -0.00807  0.29471  2.20496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.652578   0.057771   11.30   <2e-16 ***
yrs_school  0.092920   0.004333   21.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5236 on 2244 degrees of freedom
Multiple R-squared:  0.1701,    Adjusted R-squared:  0.1697
F-statistic: 459.9 on 1 and 2244 DF,  p-value: < 2.2e-16
```

Submit You have used 1 of 2 attempts

Question 2

1.0/1.0 point (graded)

What is the 90% confidence interval (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 $[0.08579005, 0.1000497]$ ✔

Explanation

The command in R to find the confidence interval is **confint**. If we run the following code:

```
#simple linear regression
single <- lm(lwage ~ yrs_school, data = nlsw88)
summary(single) # show results
coefficients(single) # model coefficients
ci <- confint(single, level=0.9)
ci
```

This is the output that we get:

```
                   5 %      95 %
(Intercept) 0.55751337 0.7476421
yrs_school  0.08579005 0.1000497
```

Submit You have used 1 of 2 attempts

Question 3

1.0/1.0 point (graded)

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 $\hat{\beta}_1$?

- 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*

Explanation

From the lecture we know that:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum(x_i - \overline{x})^2}$$

The numerator of this expression is just the sample covariance between x and y multiplied by $\frac{n}{n-1}$. Similarly, the denominator is the sample variance of x multiplied by $\frac{n}{n-1}$. Then, if we have $cov(x, y)$ and $var(x)$, then we are able to calculate $\hat{\beta}_1$. In this case y corresponds to the log of the wage and x to the total years of schooling. Then, the correct answer is (a).

Submit

You have used 1 of 2 attempts

Question 4

1.0/1.0 point (graded)

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

Please round your answer to the third decimal point, i.e. if it is 0.12494, please round to 0.125 and if it is 0.1233, please round to 0.123

0.653

Answer: 0.653

0.653

Explanation

The command that we should run in R after uploading the data is:

```
#simple linear regression
single <- lm(lwage ~ yrs_school, data = nlsw88)
summary(single) # show results
```

The output that you get after running this code is:

```
Call:
lm(formula = lwage ~ yrs_school, data = nlsw88)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.29340 -0.32611 -0.00807  0.29471  2.20496 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.652578   0.057771   11.30   <2e-16 ***
yrs_school  0.092920   0.004333   21.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5236 on 2244 degrees of freedom
Multiple R-squared:  0.1701,	Adjusted R-squared:  0.1697 
F-statistic: 459.9 on 1 and 2244 DF,  p-value: < 2.2e-16
```

Submit You have used 1 of 2 attempts

Question 5

1/1 point (graded)

True or False: For any single linear regression model, the predicted value when $x = \overline{x}$ is $\overline{y}$

- a. True ✔
- b. False

Explanation

The statement is true and we can show this by the definition of $\hat{\beta}_0$. In general, the predicted value of the model is given by $\hat{\beta}_0 + \hat{\beta}_1 x$. When $x = \overline{x}$ then we have that this is $\hat{\beta}_0 + \hat{\beta}_1 \overline{x}$. From the definition of $\hat{\beta}_0$, we have that:

$$\hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 \overline{x} = \overline{y}$$

Submit You have used 1 of 1 attempt

✔ Correct (1/1 point)

Question 6

1.0/1.0 point (graded)

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 ✔

Explanation

As Sara mentioned in lecture, one of the assumptions of the linear model is that $\mathbb{E}\left[\varepsilon_i\right] = \mathbf{0}$. By construction, the sum of the residuals that correspond to the sample analogue of ε should be very close to zero.

Submit You have used 1 of 1 attempt