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 β_0, β_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 $\dfrac{\frac{SSR_r - SSR_u}{r}}{\frac{SSR_u}{N-K-1}}$. 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

Module 8

(Co)Variance functions

- **var(x)**

Computes the variance of ***x,*** which is a vector, matrix or dataframe.

- **covar(x,y)**

Computes the covariance of ***x*** and ***y,*** where both arguments are vectors, matrices or dataframes with comparable dimensions to each other.

- **anova(object)**

Computes the analysis of variance of ***object***, which is a variable holding the results of a model fit (such as a linear model fit).

Linear model fitting etc.

- **lm(formula, data, subset, weights, na.action, method = 'qr', model = TRUE, x = FALSE, y = FALSE, qr = FALSE, ..)**

Fits a linear model to the given data and is used for linear regression. Returns the coefficients of the fit. The arguments are:

- **formula –** an object of class 'formula', which is a symbolic description of the model to be fitted (essentially, the model description in mathematical terms)
- **data –** an optional dataframe or list. If not specified, the arguments specified in ***formula*** are taken as variables by default
- **subset –** an optional vector specifying the subset of data values to be used in the fitting
- **weights –** an optional vector of weights to be used in the fitting process. Defaults to NULL, but if specified, uses a weighted least squares process to fit the model
- **na.action –** a function that indicates what should happen to NA values in the fitting process. The ***action*** values are:
    - **na.fail –** the regression fails
    - **na.omit –** excludes NA values
    - **na.exclude –** similar to na.omit, but behaves differently only when used with other functions computing residuals and predictions. It corrects for the vector lengths when these operations are conducted
    - **NULL**
- **method –** the fitting method ***'qr'*** is the default and is widely applicable
- **model, x, y, qr –** If TRUE, the function returns these components of the fit
- **linearHypothesis(model,…)**

Generic function for testing a linear hypothesis for a variety of linear models. (NOTE: For mixed effects models, the default test is the Chi-Square test for testing fixed effects).

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:

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

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

0.0/1.0 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

Answer: 0.09290

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
```

Show answer

Question 2

0.0/1.0 point (graded)

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

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

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
```

Show answer

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Question 3

0.0/1.0 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 $\hat{\beta}_1$?

- 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* ✔

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 $y\backslash.Similarly, thedenominatoristhesamplevarianceof\backslash(x$. Then, if we have $cov\left(x, y\right)$ and $var\left(x\right)$, 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).

Show answer

Submit   You have used 0 of 2 attempts

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

Answer: 0.652578

Explanation

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

```
# simple 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
```

Show answer

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Question 5

0.0/1.0 point (graded)

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

- True ✔
- False

Explanation

The statement is true and we can show this by the closed form solution $\hat{\beta}_0$, which is $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$. 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 $\hat{\beta}_0 + \hat{\beta}_1 \overline{x}$. Then from the closed form expression for $\hat{\beta}_0$, we have that:

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

Show answer

Submit

You have used 0 of 1 attempt

Answers are displayed within the problem

Question 6

0.0/1.0 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?

- Yes

No

Explanation

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

Show answer

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Now, we are interested in estimating the following model:

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

Question 7

0.0/1.0 point (graded)

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):

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

Researcher B claims that Researcher A is wrong, and that in this second model, it is not possible to separately identify $\beta_0,\ \beta_1,\ \text{and}\ \beta_2$. Who is correct?

- ◯ Researcher A
- ◯ Researcher B ✔

Explanation

The model proposed by Researcher A has the problem of multicollinearity. In particular we have that

$other\ race + black = 1$ which is the vector we use to estimate the intercept β_0. Thus, Researcher B is right -- it is not possible to separately identify $\beta_0,\ \beta_1$, and β_2.

Show answer

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Question 8

0.0/1.0 point (graded)

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 using OLS?

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

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

Explanation
This was discussed in the lecture. In general, we have that since $\mathbb{E}\varepsilon_i = 0$

$$\mathbb{E}\left[lwage|black = 0\right] = \beta_0 + \beta_1 \times 0 + 0 = \beta_0$$

$$\mathbb{E}\left[lwage|black = 1\right] = \beta_0 + \beta_1 \times 1 + 0 = \beta_0 + \beta_1$$

Thus, the sample analogues must satisfy:

$$\bar{y}_{other} = \hat{\beta}_0$$

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

Show answer

Submit You have used 0 of 2 attempts

0.0/2.0 points (graded)

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!

For the following answers, please round to the third decimal place, i.e. if the solution is 0.23412, please round to 0.234, and if it is 0.23498, please round to 0.235.

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

Answer: 1.911614

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

Answer: -0.1655357

Explanation

If we run the following code:

```
meanother <- mean(nlsw88$lwage[nlsw88$black == 0])
meanblack <- mean(nlsw88$lwage[nlsw88$black == 1])
meanother
meanblack - meanother
```

This is the output we get:

```
> #dummy variables
> meanother <- mean(nlsw88$lwage[nlsw88$black == 0])
> meanblack <- mean(nlsw88$lwage[nlsw88$black == 1])
> meanother
[1] 1.911614
> meanblack - meanother
[1] -0.1655357
>
> dummymodel <- lm(lwage ~ black, data = nlsw88)
> summary(dummymodel)

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.90667 -0.40290 -0.03418  0.37105  1.96129

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.91161    0.01398 136.739  < 2e-16 ***
black       -0.16554    0.02744  -6.033 1.88e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5701 on 2244 degrees of freedom
Multiple R-squared:  0.01596,	Adjusted R-squared:  0.01552
F-statistic: 36.39 on 1 and 2244 DF,  p-value: 1.88e-09
```

Show answer

Question 10

0.0/1.0 point (graded)

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 t-statistic?

Please round to the third decimal place, i.e. if the solution is 0.23412, please round to 0.234, and if it is 0.23498, please round to 0.235.

Answer: -6.033

Explanation

As Sara discussed in lecture, we use a t-statistic to perform this test. The t-statistic is defined as:

$$\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{-0.1655357}{0.02744} = -6.033.$$

Show answer

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Question 11

0.0/1.0 point (graded)

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

- Yes ✔
- No

Explanation

According to the R output, the p-value associated with this test is 1.88e-09. This is less than 0.01, so we can reject the null hypothesis at a 99% level of confidence.

Show answer

Submit You have used 0 of 1 attempt

Homework due Nov 2, 2020 18:30 EST Past Due

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:

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

Question 12

0.0/1.0 point (graded)

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

Please round your answer to the third decimal place, i.e. if your answer is 0.7283, please round to 0.728 and if it is 0.7289, round to 0.729.

Answer: 0.2671

Explanation

If we run the following code:

```
#multivariable regression
multi1 <- lm(lwage ~ yrs_school + ttl_exp, data = nlsw88)
summary(multi1) # show results
```

This is the output that we get:

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

Residuals:
     Min       1Q   Median       3Q      Max 
-2.09807 -0.29945 -0.00571  0.25158  2.49949 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.336944   0.057308    5.88 4.73e-09 ***
yrs_school  0.079148   0.004150   19.07  < 2e-16 ***
ttl_exp     0.039559   0.002296   17.23  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4921 on 2243 degrees of freedom
Multiple R-squared:  0.2671,    Adjusted R-squared:  0.2664 
F-statistic: 408.7 on 2 and 2243 DF,  p-value: < 2.2e-16
```

From there, we know that the R^2 is 0.2671. This implies that 26.71% of the total variance in the logarithm of the wage is explained by the years of schooling and the total experience.

Show answer

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

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

0.0/1.0 point (graded)

What would be the null hypothesis of this test?

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

Explanation
If the effect of one year of experience is equivalent to two years of education over the log of the wage. Then, the null hypothesis of this test is that $2\beta_1 = \beta_2$.

Question 14

0.0/1.0 point (graded)

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

- [ ] $log\left(wage_i\right) = \beta_0 + \beta_2\left(yrs\_school_i + 2total\ experience_i\right) + \varepsilon_i$
- [ ] $log\left(wage_i\right) = \beta_0 + \beta_1\left(\frac{1}{2}yrs\_school_i + total\ experience_i\right) + \varepsilon_i$
- [ ] $log\left(wage_i\right) = \beta_0 + \beta_1\left(yrs\_school_i + 2total\ experience_i\right) + \varepsilon_i$ ✔
- [ ] $log\left(wage_i\right) = \beta_0 + \left(\beta_1 + 2\beta_2\right)yrs\_school_i + \varepsilon_i$
- [ ] $log\left(wage_i\right) = \beta_0 + \left(2\beta_1 + \beta_2\right)yrs\_school_i + \varepsilon_i$
- [ ] $log\left(wage_i\right) = \beta_0 + \beta_2\left(\frac{1}{2}yrs\_school_i + total\ experience_i\right) + \varepsilon_i$ ✔

Explanation

If we substitute the null hypothesis $\left(2\beta_1 = \beta_2\right)$ in the equation
$log\left(wage_i\right) = \beta_0 + \beta_1 yrs\_school_i + \beta_2 total\ experience + \varepsilon_i$, then we have that:

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

$$log\left(wage_i\right) = \beta_0 + \beta_1 yrs\_school_i + 2\beta_1 total\ experience + \varepsilon_i$$

$$log\left(wage_i\right) = \beta_0 + \beta_1\left(yrs\_school_i + 2 total\ experience_i\right) + \varepsilon_i$$

Analogously we have that:

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

$$log\left(wage_i\right) = \beta_0 + \frac{\beta_2}{2} yrs\_school_i + \beta_2 total\ experience + \varepsilon_i$$

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

Show answer

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

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.

Please round your answer to the fourth decimal place, i.e. if your answer is 0.78244, please round to 0.7824, and if it is 0.78247, please round to 0.7825.

Answer: 0.026292

Explanation

If we run the following code:

```
#Restricted model
nlsw88$newvar <- nlsw88$yrs_school + 2*nlsw88$ttl_exp
restricted <- lm(lwage ~ newvar, data = nlsw88)
summary(restricted) # show results
```

This is the output that we get:

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.79637 -0.32172 -0.02268  0.27505  2.39896
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.865430   0.042395   20.41   <2e-16 ***
newvar      0.026292   0.001075   24.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5106 on 2244 degrees of freedom
Multiple R-squared:  0.2106,    Adjusted R-squared:  0.2102
F-statistic: 598.6 on 1 and 2244 DF,  p-value: < 2.2e-16
```

Show answer

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Question 16

0.0/1.0 point (graded)

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?

Please round your answer to the second decimal places, i.e. if your answer is 89.28397, please round to 89.28 and if it is 89.28997, round to 89.29

Answer: 172.9599

Explanation

If we run the following code:

```
#multivariable regression
multi <- lm(lwage ~ yrs_school + ttl_exp, data = nlsw88)
summary(multi) # show results
anova_unrest <- anova(multi)
#Restricted model
nlsw88$newvar <- nlsw88$yrs_school + 2*nlsw88$ttl_exp
restricted <- lm(lwage ~ newvar, data = nlsw88)
summary(restricted) # show results
anova_rest <- anova(restricted)
#Test
statistic_test <- (((anova_rest$`Sum Sq`[2]-anova_unrest$`Sum Sq`[3]/1)/((anova_unrest`Sum Sq`[3]/anova_unrest$Df[
statistic_test
pvalue <- df(statistic_test, 1, anova_unrest$Df[3])
pvalue
```

This is the output that we get:

```
> statistic_test
```

```
[1] 172.9599
> pvalue <- df(statistic_test, 1, anova_unrest$Df[3])
> pvalue
[1] 1.930469e-38
```

Show answer

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Question 17

0.0/1.0 point (graded)

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

- ( ) Reject ✔
- ( ) Do not reject

Explanation

The p-value associated with this test is less than 0.05. Then, we can reject the null hypothesis at this confidence level. You can also use the **car** package in R and the following code to perform the test directly:

```
matrixR <- c(0, -2, 1)
linearHypothesis(multi, matrixR)
```

```
Hypothesis:
- 2 yrs_school  + ttl_exp = 0

Model 1: restricted model
Model 2: lwage ~ yrs_school + ttl_exp

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2244 585.09
2   2243 543.20  1    41.887 172.96 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Show answer

Submit You have used 0 of 1 attempt

Answers are displayed within the problem