Exercise Sheet 2

2024-04-08

2.1

Consider the data set PSID1982 from the R-library AER. Have a look at the corresponding help file of this data set. (If you have/can find a more recent data set including wages and education, you are welcome to use that.)

```
library(AER)
```

```
## Lade nötiges Paket: car
```

```
## Lade nötiges Paket: carData
```

```
## Lade nötiges Paket: lmtest
```

```
## Lade nötiges Paket: zoo
```

```
##
## Attache Paket: 'zoo'
```

```
## Die folgenden Objekte sind maskiert von 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Lade nötiges Paket: sandwich
```

```
## Lade nötiges Paket: survival
```

```
data("PSID1982")
```

(a)

What is n for this data set?

n = number of observations

```
n <- nrow(PSID1982)
```

n = 595

(b)

Regress the variable wage on an intercept as well as the variable educ, meaning that you compute the least-squares estimators when wage is the dependent and educ the explanatory variable. This can e.g. be done in R through the following commands. data(PSID1982) lm(wage educ, data = PSID1982) Use the summary command to list the output.

General information:

wage (dependent variable): The wage varies depending on the educ.

educ (explanatory variable): If you change the educ score the wage should also change.

least-square estimators: $\widehat{\beta}_1$, $\widehat{\beta}_2$ They are used to create the least-square regression line which is a ‘mean’ line between all values.

The $\widehat{\beta}_1$ is calculated by $\bar{y}$-$\widehat{\beta}_2\bar{x}$. Which equals the distance on the y-axis.

The $\widehat{\beta}_2$ is calculated by $\frac{S_{xy}}{S_{xx}}$ which is equal to $\frac{cov(x_i,y_i)}{var(x_i)}$ and is the slope of the least-square regression line.

```
model <- lm(wage ~ education, data = PSID1982)
model_summary <- summary(model)
print(model_summary)
```

```
##
## Call:
## lm(formula = wage ~ education, data = PSID1982)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1038.6  -311.6   -48.1   222.3  3603.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   70.467     92.231   0.764    0.445
## education     83.888      7.017  11.955   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 477.1 on 593 degrees of freedom
## Multiple R-squared:  0.1942, Adjusted R-squared:  0.1929
## F-statistic: 142.9 on 1 and 593 DF,  p-value: < 2.2e-16
```

$\widehat{\beta}_1 = 70.467$

$\widehat{\beta}_2 = 83.888$

(c)

Comment on the sign of the estimated slope parameter.

General: The sign of the estimated slope parameter in a linear regression model indicates the direction of the relationship between the independent variable and the dependent variable.

If the slope parameter is positive, it means that as the independent variable increases, the dependent variable also increases. This is known as a positive correlation.

If the slope parameter is negative, it means that as the independent variable increases, the dependent variable decreases. This is known as a negative correlation.

Answer:

The slope parameter is equal to $\widehat{\beta}_2$, so we have a positive slope parameter.

That means the higher the education value is, the higher is the wage.

(d)

Sum up the residuals in this model, i.e., compute

$\sum_{i=1}^{n} \hat{u}$

What do you get?

Generell:

Residuals: are the difference between the line and the actual points $\widehat{u}_i = y_i - \widehat{y}_i$

```r
residuals <- resid(model)
sum_residuals <- sum(residuals)
```

sum of Residuals = 3.08233438772731e-11 ~ 0 -> close to 0

If the sum of residuals is close to 0 in a regression analysis, it suggests that the model's predictions are, on average, fairly accurate and that there isn't any systematic bias in the model's predictions.

(e)

Compute the quantity

$R^2 = \frac{S_{\hat{y}\hat{y}}}{S_{yy}}$

for this data example. Can you find it in your R-output? (We will still discuss R2 in class.)

```r
r_squared <- model_summary$r.squared
```

$R^2 = 0.194216581545736$

2.2

(a)

Now regress log(wage) on a constant and the variable educ and look again for the multiple R-squared. How does it compare to the one from the level-level model?

Answer:

```r
model_log <- lm(log(wage) ~ education, data = PSID1982)
summary_log <- summary(model_log)
print(summary_log)
```

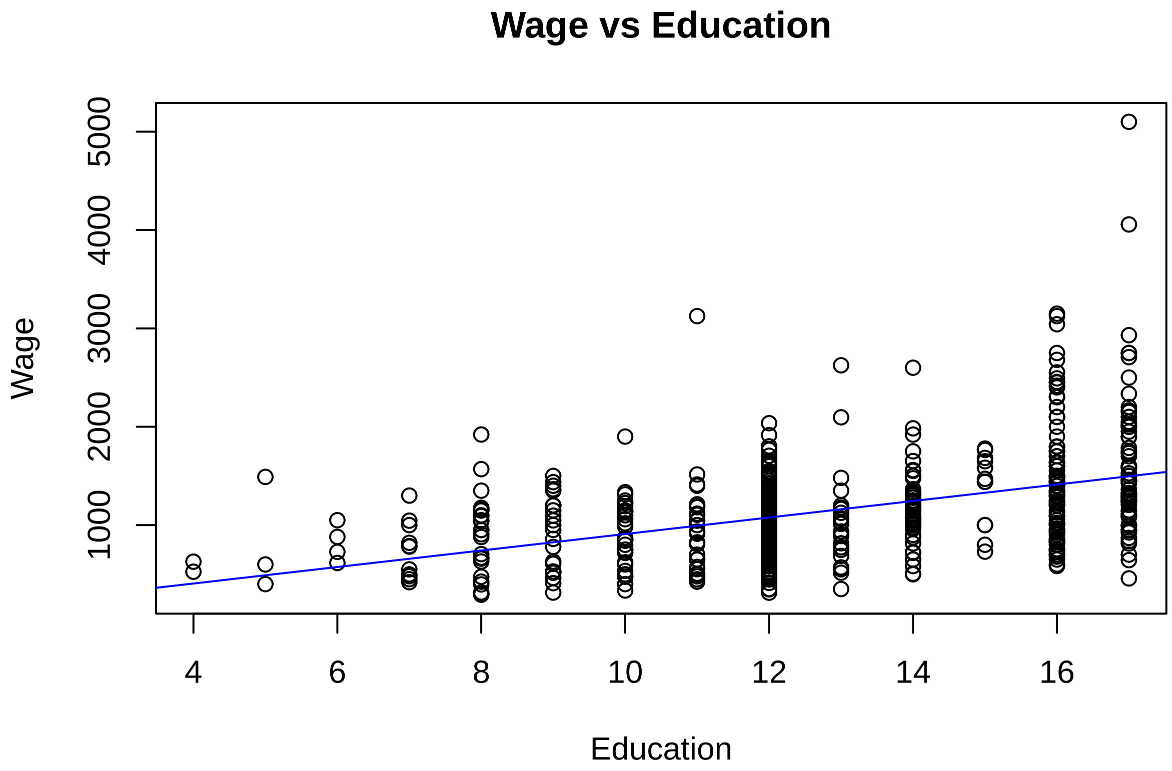
```
## 
## Call:
## lm(formula = log(wage) ~ education, data = PSID1982)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.14389 -0.25317  0.03051  0.26343  1.28819 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 6.029191   0.075460    79.9   <2e-16 ***
## education   0.071742   0.005741    12.5   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.3904 on 593 degrees of freedom
## Multiple R-squared:  0.2085, Adjusted R-squared:  0.2071 
## F-statistic: 156.2 on 1 and 593 DF,  p-value: < 2.2e-16
```

$R^2 = 0.2085$

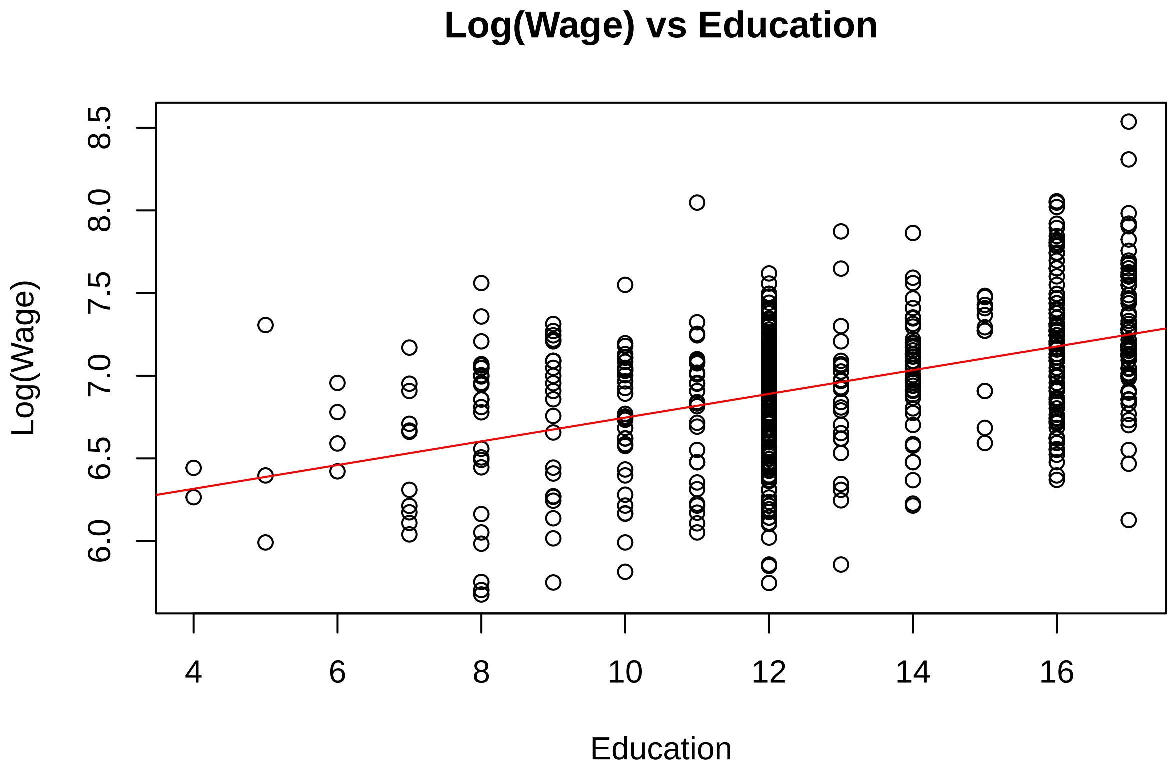
(b)

For both models, plot the data together with the estimated regression line. Comment.

```r
plot(PSID1982$educ, PSID1982$wage, main = "Wage vs Education", xlab = "Education", ylab = "Wage")
abline(model, col = "blue")
```



```
plot(PSID1982$educ, log(PSID1982$wage), main = "Log(Wage) vs Education", xlab = "Education", ylab = "Lo
abline(model_log, col = "red")
```



In these plots, the blue line represents the estimated regression line for the first model (wage regressed on education), and the red line represents the estimated regression line for the second model (log(wage) regressed on education). The slope of the regression line represents the estimated effect of education on wage (or log(wage) in the second model).

If the slope is positive, it means that as education increases, wage (or log(wage)) also increases.

If the slope is negative, it means that as education increases, wage (or log(wage)) decreases.

The steepness of the slope indicates the strength of this relationship.

(c)

In both models, quantitatively intepret the estimated coefficient $\widehat{\beta}_2$.

Answer:

$\widehat{\beta}_2 = 83.888$ of wage

$\widehat{\beta}_2 = 0.0717$ of log(wage)

Wage: This means that for a one-unit increase in the independent variable associated with $\widehat{\beta}_2$ the dependent variable (wage) is expected to increase by 83.888 units, holding all other variables constant.

Log(wage): Here, the interpretation is in terms of percentage change. Specifically, for each one-unit increase, the logarithm of wage is expected to increase by 0.071742 units. Since the logarithm of wage is used, the interpretation isn't in terms of the absolute change in wage but rather the proportional change. This means for each additional year of education, the wage is expected to increase by approximately 7.17%.