Problem Set Assignment No. 2

Aria Muchhal

06 February, 2023

```
## Keep this line always
knitr::opts_chunk$set(echo = TRUE,
                      collapse = TRUE,
                      warning = FALSE, message = FALSE,
                      fig.align = 'center')
```

1. Setup your document

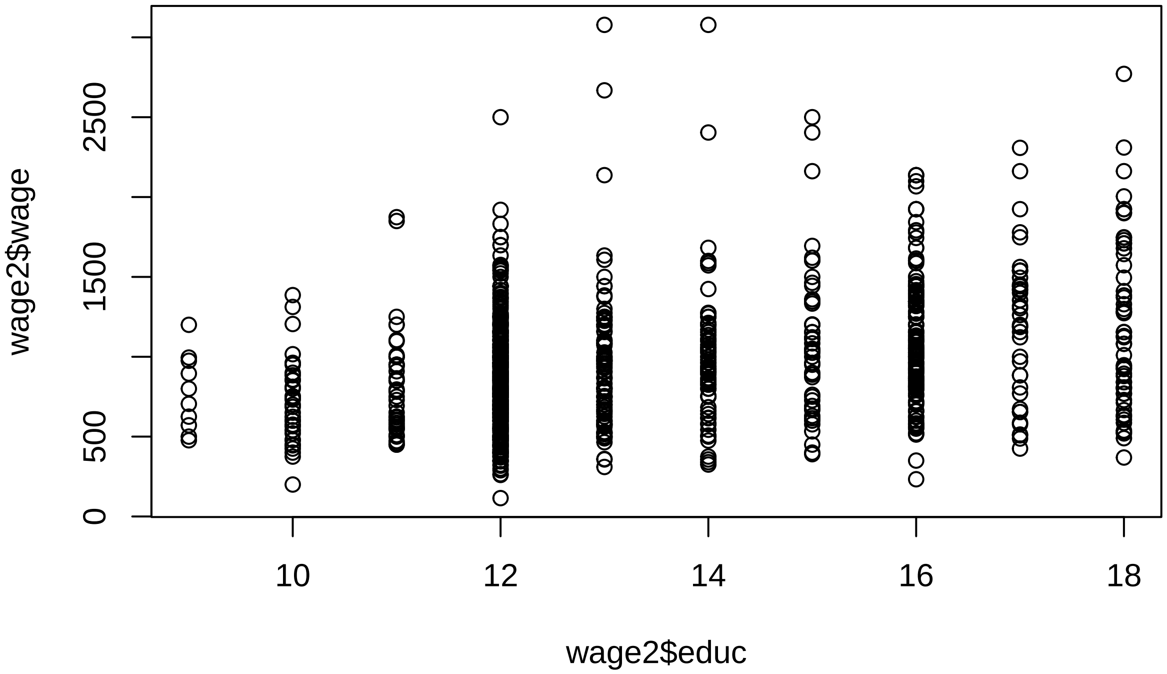
A. Load Libraries

```
library(knitr)
library(wooldridge)
wage2=wooldridge::wage2
```

2. Single Variable Regression.

A. Plot the relationship

```
plot(wage2$educ,wage2$wage)
```



Does there seem to be a relationship in the data between education and wage? If we were to "eyeball" a line, would it be upward sloping or downward sloping? What does economic theory tell us we should expect?

There seems to be a roughly positive linear relationship between education in wage. An eyeballed line would be upward sloping, which follows economic theory that years of education is an effective signal to employers of a worker's ability, thus corresponding to a higher wage.

B. Calculate wage-bar and educ-bar.

```
wagebar=mean(wage2$wage)
educbar=mean(wage2$educ)
```

What is the mean of wage?

957.9455.

C. Calculate the sample variance of educ

```
a=wage2$educ-educbar
wage2$a<-a
n=nrow(wage2)-1
svar_educ=(1/n)*sum(a^2)
print(svar_educ)
## [1] 4.825288
```

What is the sample variance of educ?

4.825288

D. Repeat 2B and 2C, but for wage

```
b=wage2$wage-wagebar
wage2$b<-b
svar_wage=(1/n)*sum(b^2)
print(svar_wage)
## [1] 163507.7
```

What is the sample variance of wage?

163507.7

E. Calculate Cov(wage, educ) using the results from 2B and 2C

```
cov_wage_educ=sum(a*b)/n
```

Report the covariance. Is the covariance between educ and wage positive or negative?

290.5513: positive.

F. Calculate B1

```
beta1hat=cov_wage_educ/svar_educ
```

What is your B1? What is the interpretation of the coefficient in terms of the population regression function?

60.21428. This means that one additional year of education is associated with a $60.21 increase in expected monthly earnings, all else held equal.

G. Calculate B0

```
beta0hat=wagebar-beta1hat*educbar
```

What is your B0? What is the interpretation of the coefficient in terms of the population regression function? When does yhat= B0?

146.9524.

Beta0 is the intercept. It is the y value when x=0 on the regression function. ybar=beta0 if xbar=0–the mean of x must be 0.

3. Goodness of Fit

A. Calculate the Sum of Squares Total (SST)

```
SST=sum((wage2$wage-wagebar)^2)
```

B. Calculate the residuals u from the regression in 2

```
residuals=wage2$wage-beta0hat-beta1hat*wage2$educ
wage2$residuals<-residuals
mean(residuals)
## [1] 4.412083e-14
```

Is this a strange result, or did you expect this?

The mean was almost 0, which makes sense because E(u)=0 because the residuals are measures of deviations from the regression function, and the regression function is a line of best fit, so the deviations should average to 0.

C. Calculate the SSR

```
SSR=sum(wage2$residuals^2)
```

What is the SSR? Is this larger or smaller than the SST? Can it ever be larger than the SST?

136375524, which is smaller than the SST, because it fundamentally has to be. It only measures a part of the deviation, while SST looks at total deviation. Another way to think about this is SST=SSR+SSE where none of these elements are <0. Thus, SSR is always less than or equal to SST.

D. Calculate the R2

```
Rsquare=1-(SSR/SST)
```

Interpret that R2 in the context of our regression

R2=0.1070001, which measures the fraction of variance in wage explained by the model. As R2 is close to 0, this indicates that the SSE is close to 0, so the model isn't explaining much of the variance in wage.

4. Simple regression

A. Running our regression

```
myRegression<-lm(wage2$wage~wage2$educ,data = wage2)
summary(myRegression)
##
## Call:
## lm(formula = wage2$wage ~ wage2$educ, data = wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -877.38 -268.63  -38.38  207.05 2148.26
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  146.952     77.715   1.891   0.0589 .
## wage2$educ    60.214      5.695  10.573   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 382.3 on 933 degrees of freedom
## Multiple R-squared:  0.107,  Adjusted R-squared:  0.106
## F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```

How do the coefficients from 4A compare to the estimates you did “by hand” in Q2?

I was exactly correct on both.

How does the R2 compare to the R2 you calculated in Q3?

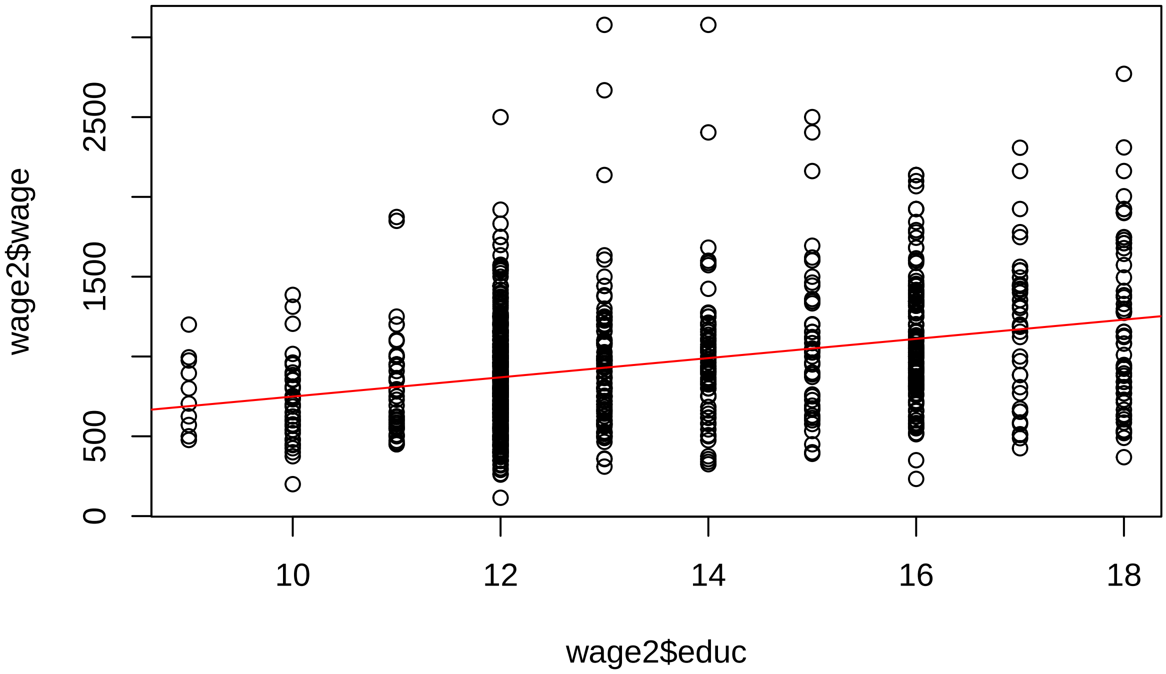
I was also correct on this.

The output gives the degrees of freedom. How was this calculated?

This was the sample size(935) minus the number of parameters needed to calculate during the analysis(2), resulting in DF=933.

B. Plotting the Regression

```
plot(wage2$educ,wage2$wage)
abline(myRegression, col = 'red')
```



C. Last Question

I spent 5 continuous hours on this problem set, but it took around half a day with breaks.