In-class Lab 6

ECON 425 (Justin Heflin, West Virginia University)

February 20, 2023

The purpose of this in-class lab is to further practice your regression skills. The lab may be completed as a group, but each student should turn in their own work. To get credit, upload your .R script to the appropriate place on eCampus ("In-Class Labs'' folder).

For starters

Open up a new R script (named `ICL6_XYZ.R`, where `XYZ` are your initials)

Install a new R Package

```
install.packages("ggplot2")
```

```
library(ggplot2)
```

Explaining Wage (using a regression model)

For this lab, let's use data on wages during the year 1980 and other information on 935 people (observations). This data is located in the `wage2` data-set in the `wooldridge` R package. Each observation represents an individual human being.

```
library(wooldridge)
wage_data <- as.data.frame(wage2)
```

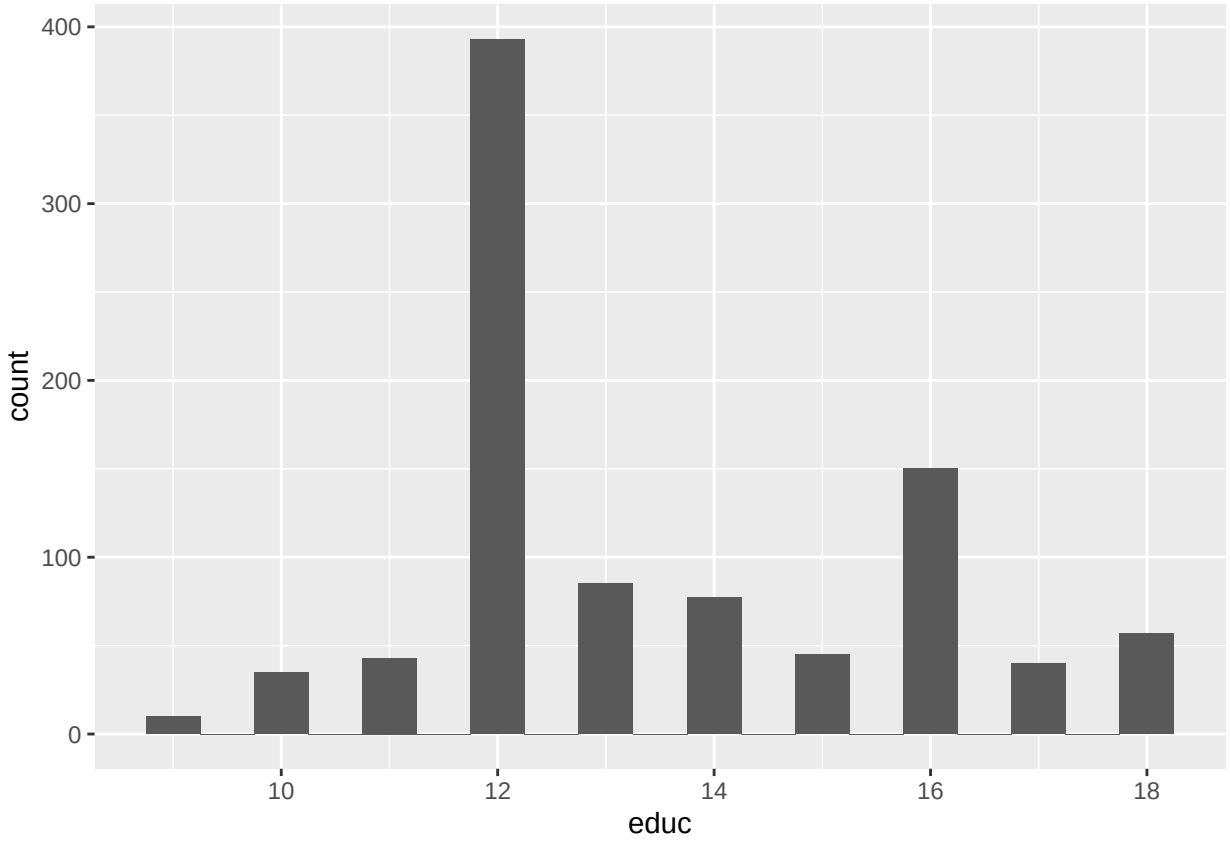
Variable Names and brief description:

- **wage**: monthly earnings
- **hours**: average weekly hours
- **IQ**: IQ score
- **educ**: years of education
- **exper**: years of work experience
- **tenure**: years with current employer
- **sibs**: number of siblings
- **brthord**: birth order

- **meduc**: mother's education
- **feduc**: father's education

First, let's plot a histogram of the years of education distribution and then calculate the expected value (average) of wage given a certain number of years of education:

```
ggplot(data = wage_data, aes(x=educ)) + geom_histogram(binwidth = 0.5)
```



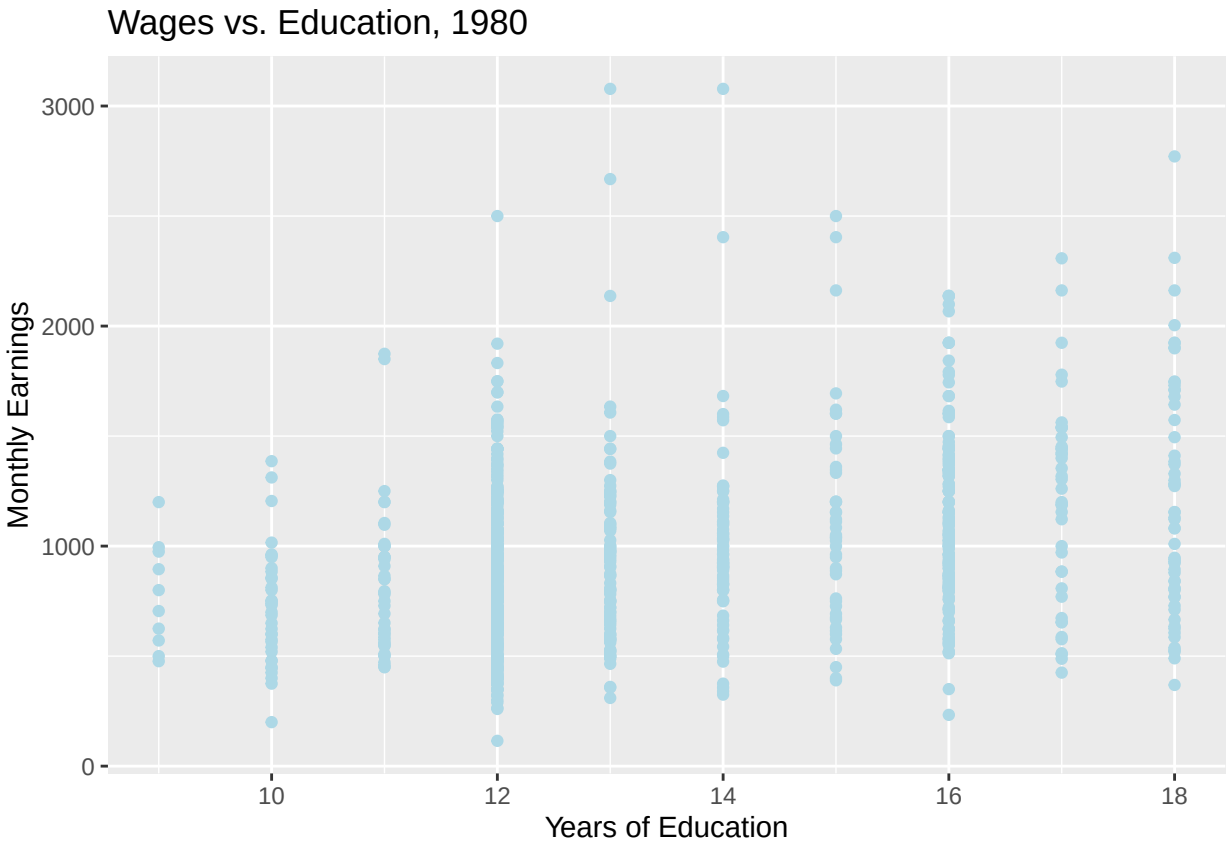

```
#frequency of each value for years of education
table(unlist(wage_data$educ))
```

```
##
##   9  10  11  12  13  14  15  16  17  18
##  10  35  43 393  85  77  45 150  40  57
```

```
#Calculate the expected value
Wage_10 <- weighted.mean(wage_data$wage, wage_data$educ=="10") #$719.48
Wage_12 <- weighted.mean(wage_data$wage, wage_data$educ=="12") #$862.67
Wage_14 <- weighted.mean(wage_data$wage, wage_data$educ=="14") #$990.01
Wage_16 <- weighted.mean(wage_data$wage, wage_data$educ=="16") #$1,108.71
Wage_18 <- weighted.mean(wage_data$wage, wage_data$educ=="18") #$1,200.82
```

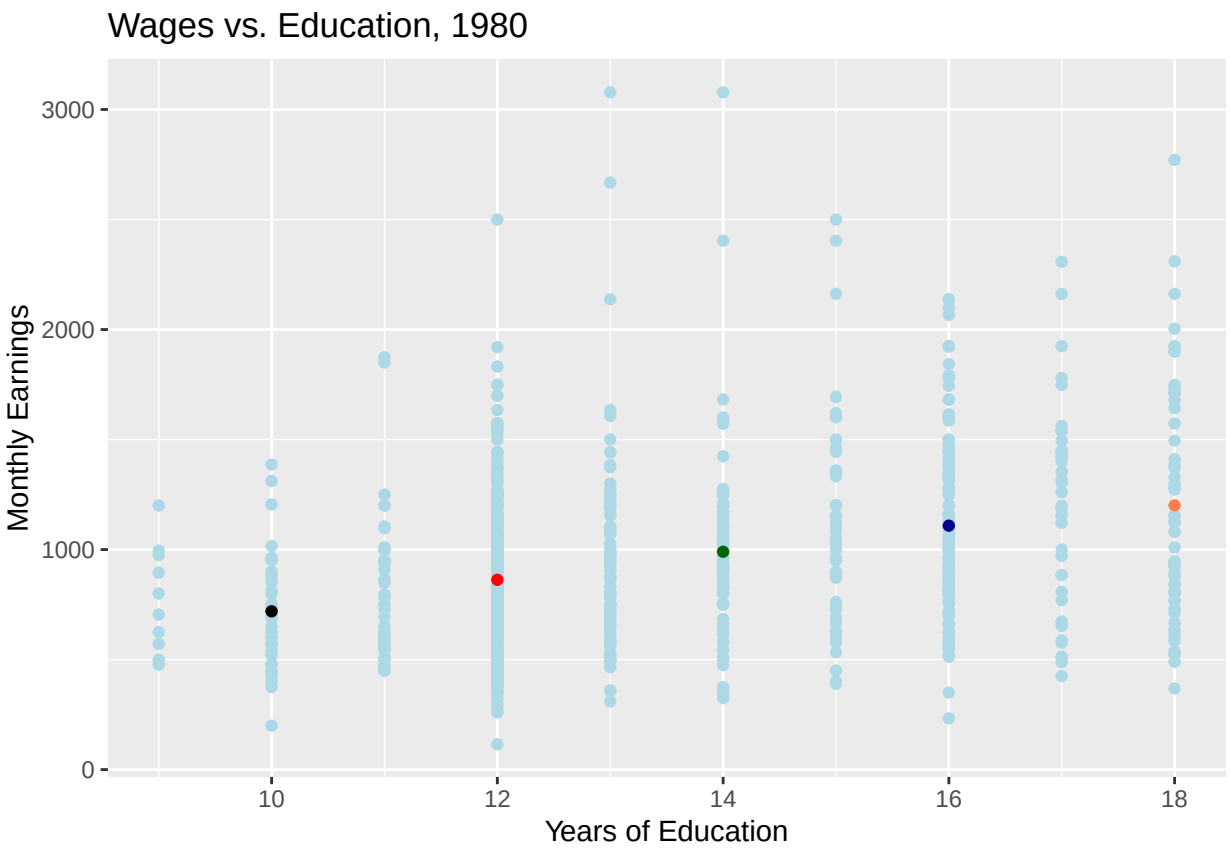
Now, let's make a scatter-plot of the two variables (monthly earnings and years of education) and look for possible patterns in the relationship between them.

```
ggplot() + geom_point(data = wage_data, mapping = aes(x = educ, y = wage),
  color = 'lightblue') +labs(y = "Monthly Earnings",
  x = "Years of Education", title = "Wages vs. Education, 1980")
```



Let's add in our expected value of wages that we calculated earlier to visually see how more education can lead to increases in wages

```
ggplot() + geom_point(data = wage_data, mapping = aes(x = educ, y = wage),
  color = 'lightblue') + labs(y = "Monthly Earnings",
  x = "Years of Education", title = "Wages vs. Education, 1980") +
  geom_point(mapping = aes(y = Wage_10, x = 10), color = 'black') +
  geom_point(mapping = aes(y = Wage_12, x = 12), color = 'red') +
  geom_point(mapping = aes(y = Wage_14, x = 14), color = 'darkgreen') +
  geom_point(mapping = aes(y = Wage_16, x = 16), color = 'darkblue') +
  geom_point(mapping = aes(y = Wage_18, x = 18), color = 'coral')
```



A simple linear regression model explaining wage is:

$$wage_i = \beta_0 + \beta_1 educ_i + \epsilon_i$$

Before we run our regression, what "signs'' do you believe our coefficients will have once we estimate our model? Do you expect the intercept term (β_0) to be positive or negative? What about the slope coefficient (β_1) for education, do you expect it to be positive or negative?

Now let's estimate this simple linear regression then add more independent variables to see how our model and the corresponding coefficients change. Before we run What signs'' do you believe

```
regression1 <- lm(wage ~ educ, data = wage_data)
summary(regression1)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = wage_data)
##
## 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 .
## 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
```

Before we add more independent variables to our regression model, let's interpret our estimated coefficients and decide if this model can be considered "good' '. First, what signs do our estimated coefficients have? Are they positive or negative? Do they match what you expected them to be *before* we ran the regression?

How do we interpret our intercept term (should be equal to 146.952)? Well, if an individual has no education, meaning educ = 0, then their monthly earnings would be equal to \$146.95/month.

What about our coefficient on education (should be equal to 60.214)? This coefficient implies for each additional year of education, they will earn \$60 more dollars a month.

For a sanity check we can use the expected values we calculated earlier to double check this coefficient. We know the expected value of wage (monthly earnings) given someone with 10 years of education is \$719.48. If they were to receive two more years of education, based on our $\hat{\beta}_1$ regression coefficient of 60.214 that would be equal to \$120.43 of additional monthly earnings, $\implies$ \$719.48 + \$120.43 = \$839.91, which is awfully close to the expected value of wage (monthly earnings) given someone with 12 years of education, \$862.67.

Now to determine if this model can be considered "good' ' let's take a look at our R^2, which for this model is equal to 0.106. Remember R^2 must lie on the interval between 0 and 1. We prefer a higher R^2 to a lower one. Based on the current R^2 of 0.106, our model only explains roughly 11% of the variation in our dependent variable, wage (monthly earnings).

We can increase our R^2 by adding more independent/explanatory variables. Let's first start off by adding one independent variable to see how our coefficients and R^2 change.

$$wage_i = \beta_0 + \beta_1 educ + \beta_2 exper + \epsilon_i$$

```
regression2 <- lm(wage ~ educ + exper, data = wage_data)
summary(regression2)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = wage_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -924.38 -252.74  -40.88  198.16 2165.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -272.528    107.263  -2.541   0.0112 *
## educ          76.216      6.297  12.104  < 2e-16 ***
## exper         17.638      3.162   5.578 3.18e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 376.3 on 932 degrees of freedom
## Multiple R-squared:  0.1359, Adjusted R-squared:  0.134
## F-statistic: 73.26 on 2 and 932 DF,  p-value: < 2.2e-16
```

Now our intercept coefficient is negative, not exactly what we would expect if both independent variables were equal to zero that someone would earn negative monthly earnings $272.53. In this case, our intercept term does not offer a great deal of insight.

Our estimated coefficient for education, $\hat{\beta}_1$, it is positive, which makes a great deal of sense, more education should lead to a higher wage. Now, our coefficient implies that for every additional year of education an individual will earn $76.22 more a month, holding constant their years of work experience. We could rationalize this by saying someone quit/left a job to go back to school. That happens all the time, when someone does this, they forego work experience to gain another year of education. And based on the coefficients where each additional year of education results in an additional $76.22/month relative to an additional year of work experience only increasing their monthly earnings by $17.64/month.

Moving onto our estimated coefficient for experience, $\hat{\beta}_2$, it is also positive, which again makes a lot of sense, the more experience a person has working, generally the more they earn. This coefficient implies that for every additional year of experience, they can expect an increase of $17.64/month, holding constant their years of education.

Finally, let's take a look at our R^2 to see if our model has "improved'' from our previous one. Our last model only explained roughly 11% of the variation of our dependent variable. Our current model (regression2) has an R^2 of 0.1359. This is an improvement but only marginally, we went from explaining roughly 11% to explaining 13.6%. Though this does demonstrate that by adding another independent variable our R^2 does increase how much of the variation of our dependent variable is explained by our model.

For our final regression of this lab, let's throw the kitchen sink at our dependent variable and see how much of the variation in our dependent variable can be explained by a model that has several independent/explanatory variables.

$$wage_i = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 age + \beta_4 IQ + \epsilon_i$$

```
regression3 <- lm(wage ~ educ + exper + age + IQ, data = wage_data)
summary(regression3)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + age + IQ, data = wage_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -871.69 -243.85  -43.11  182.09 2178.50
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -871.9720   159.7874  -5.457 6.21e-08 ***
## educ          52.0962     7.2994   7.137 1.92e-12 ***
## exper         11.2010     3.7181   3.013  0.00266 **
## age           14.1408     4.6634   3.032  0.00249 **
## IQ             5.2423     0.9384   5.586 3.04e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 369.1 on 930 degrees of freedom
## Multiple R-squared:  0.1702, Adjusted R-squared:  0.1666
## F-statistic: 47.68 on 4 and 930 DF,  p-value: < 2.2e-16
```

Comment out your opinion of this model along with your interpretation of the estimated coefficients.