OEconomica Regression Worksheet

Johnny Ma

February 22, 2017

```
knitr::opts_chunk$set(echo = TRUE)
#install.packages("ggplot2")
#install.packages("AER")
#install.packages("stargazer")
library(ggplot2)
library(AER)
library(stargazer)
```

Linear Regression

Today we will run a simple linear regression using R and work to interpret the results. We will also learn about the assumptions we make when running such a regression.

As always, please ask a board member if you have any questions!

What is a Regression?

Statisticians use regressions to predict one variable (dependent variable) from other variables (independent variables). For example, one might want to a multivariate regression to predict current income from the number of years of education, age, and gender. The equation that coincides with this regression is shown below.

$$Wage = \beta_0 + \beta_1 * (educ) + \beta_2 * (age) + \beta_3 * (gender) + \epsilon$$

Where the βs are constants and the ϵ is the error term. In a regression we try to find the values for these constants, called coefficients, that give us the smallest error. The data we would want is some big individual-level numeric data with income, education level, age, and gender. The regression is can be linear if it is in the form of

$$y = mx + b$$

where y is a dependent variable (in this case, wage), x is a variable of interest (in this case, education, age, or gender) and m is the corresponding β, with b as the intercept. We'll go into more detail as to what these variables mean later on.

For now, please write the equation that would coincide with a simple linear regression that predicts diamond *price* (dependent) using *carat* (independent) using β_0 and β_1 with ϵ.

____

____

Though y or income is considered the dependent variable, in that values of x should determine y, we need to find some values for β_0 and β_1 first. To do this, we need to use our data, with values of y and x, to give us these β values using some equation. What equation makes sense for us to use?

OLS (Ordinary Least Squares) and BLUE (Best Linear Unbiased Estimator)

When we run the simple linear regression in R on these two variables, we will be asking the program to give us the values of our coefficients β_0 and β_1 with the smallest ϵ. Economists typically use OLS regression, which uses a equation that *minimizes the sum of the squares of the error (ϵ) term.* So, we are trying to find the coefficient values that minimize the difference between the *predicted value of the independent variable (from the equation) and the known value of the independent variable.* The "difference" is the error term, and can be seen explicitly in this image:

So, we are trying to minimize the sum of the squares of the 'error', or the differences between observed responses and predicted values. This gives us our fitted line! There are a number of assumptions in this (heteroskedascity, $E[\epsilon|X] = 0$, etc.), but we use this method (OLS) because the Estimators for β are Unbiased and Consistent. Look up what these mean if you're interested!

Mathematically, if we have a simple linear regression:

$$y = \beta_0 + \beta_1 * x + \epsilon$$

We are trying to solve for βs in the following optimization problem:

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \epsilon^2 = \min_{\hat{\beta}_0,\hat{\beta}_1} \sum (y - \hat{y})^2 = \min_{\hat{\beta}_0,\hat{\beta}_1} \sum (y - (\hat{\beta}_1 * x + \hat{\beta}_0))^2$$

With

$$\hat{\beta}_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$$

With $\hat{y}$ being the predicted values generated from the $\hat{\beta}$ values, which come from the averages $(\bar{y}, \bar{x})$ and observed values of x and y. So, given some data, we can use this model to predict values of y with our estimators, β as BLUE.

But you don't need to know what any of this means or what's happening with the math to do a linear regression in R. For further mathematical explaination, take Econometrics!

Running Regression

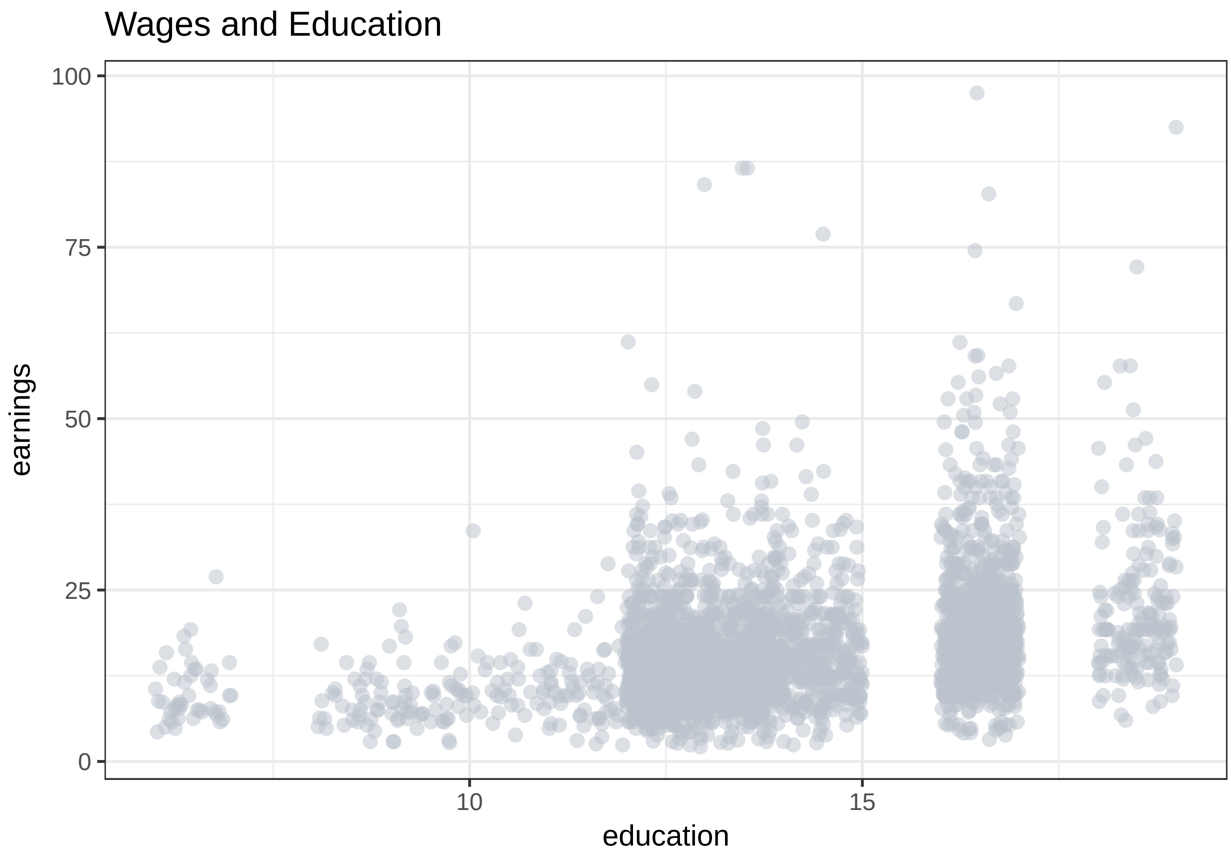
Let's look at some data. This is AER wage data from some year, with earnings per hour, years of education, age, and gender.

```
data("CPSSWEducation")

summary(data)
```

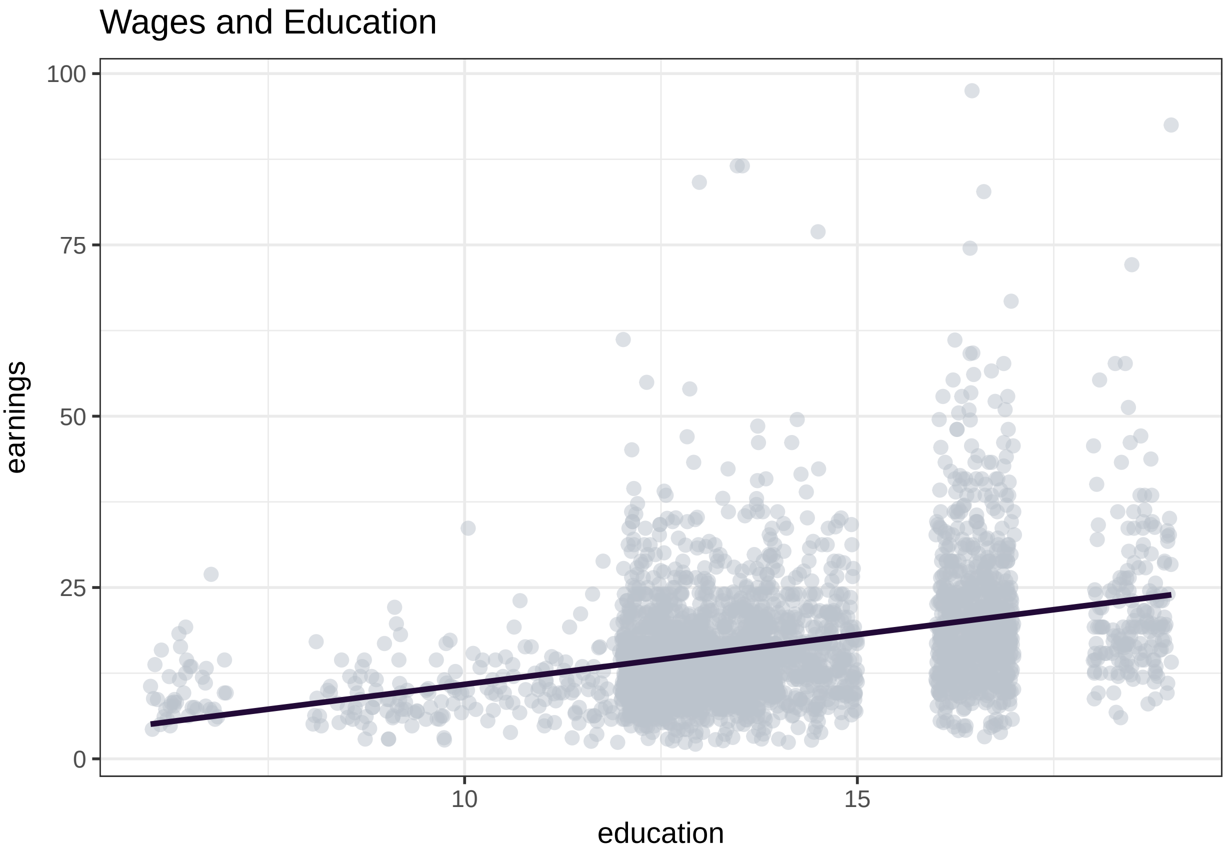
```
##       age          gender        earnings        education
##  Min.   :29.0   female:1202   Min.   : 2.137   Min.   : 6.002
##  1st Qu.:29.0   male  :1748   1st Qu.:10.577   1st Qu.:12.541
##  Median :29.0                 Median :14.615   Median :13.622
##  Mean   :29.5                 Mean   :16.743   Mean   :14.044
##  3rd Qu.:30.0                 3rd Qu.:20.192   3rd Qu.:16.260
##  Max.   :30.0                 Max.   :97.500   Max.   :18.997
##    education.c.V1          age.c.V1              rand
##  Min.   :-7.550169   Min.   :-0.4976271   Min.   :0.0003347
##  1st Qu.:-1.550169   1st Qu.:-0.4976271   1st Qu.:0.2427316
##  Median :-0.550169   Median :-0.4976271   Median :0.4915559
##  Mean   : 0.000000   Mean   : 0.0000000   Mean   :0.4939479
##  3rd Qu.: 2.449831   3rd Qu.: 0.5023729   3rd Qu.:0.7404450
##  Max.   : 4.449831   Max.   : 0.5023729   Max.   :0.9996788
```

```
qplot(education, earnings, data = data, size = I(2), alpha = I(0.5), colour = I("#bbc3cc")) + theme_bw(
```



If we look at this sample income-education data found in R, we see that there seems to be some positive relationship between education and earnings. Let's fit a line through this.

```
qplot(education, earnings, data = data, size = I(2), alpha = I(0.5), colour = I("#bbc3cc")) + theme_bw(
```


Wages and Education

Let’s try to run the regression that created this OLS line using the data. Regressions are run using the *lm()* function in R.

```
model <- lm(earnings ~ education.c, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = earnings ~ education.c, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.270  -5.355  -1.513   3.194  77.164
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.74272    0.16146  103.70   <2e-16 ***
## education.c  1.46693    0.06978   21.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.769 on 2948 degrees of freedom
## Multiple R-squared:  0.1304, Adjusted R-squared:  0.1301
## F-statistic: 441.9 on 1 and 2948 DF,  p-value: < 2.2e-16
```

(I sneakily centered the education data so the intercept makes more sense).

How can we interpret the table we see now? Let’s first look at the “estimate” line.

The estimate line provides the values of the βs that we have been talking about. The "(Intercept)" is the β_0 and the "education.c" is the β_1. So now we have

$$y = \beta_0 + \beta_1 * (educ) = 16.74272 + 1.46693 * (educ)$$

The interpretation is as follows: β_0 is the estimated starting wage for someone of no education. For each year of education, the estimate wage increases by 1.46693$ per hour. This is why we go to college!

Keep in mind these values are *estimations* from the data. We are attempting to draw a relationship between education and earnings assuming there is a real, casual relationship. Causality versus Correlation is important to keep in mind (check out this website).

What are the other stats we get from the linear regression? Though we get an R-squared value, standard errors, f-statistics, and other information, the one that we're most interested in as Social Scientists (I might get flak from Sylvia for this) is the famous "p-value", whichis "Pr(>|t|)". The p-value helps you decide whether or not to accept the null hypothesis. In other words, whether or not the variable you're looking at has a 'significant' impact on the dependent variable (y). The cut-off point economists generally use is called the significance level, and is usually $\alpha = 0.05$. So it looks like education is significant in predicting earnings values. Who would have thought?

Multivariate Regression

How can we add more explanatory variables into the regression? Recall we had this model for wages.

$$Wage = \beta_0 + \beta_1 * (educ) + \beta_2 * (age) + \beta_3 * (gender) + \epsilon$$

To add more variables to the regression, we simply add a '+' to the lm() function.

```r
model <- lm(earnings ~ education.c + age.c^2 + gender, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = earnings ~ education.c + age.c^2 + gender, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.135  -5.334  -1.415   3.125  75.146
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.73450    0.25015   58.90   <2e-16 ***
## education.c  1.57873    0.06935   22.76   <2e-16 ***
## age.c        0.72289    0.31704    2.28   0.0227 *
## gendermale   3.38915    0.32659   10.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.609 on 2946 degrees of freedom
## Multiple R-squared:  0.1623, Adjusted R-squared:  0.1615
## F-statistic: 190.3 on 3 and 2946 DF,  p-value: < 2.2e-16
```

The interpretation is the same as before. What do the values of the Estimates mean? Don't forget that gender is a binary variable. (Also of consideration is how we use age^2, which is standard in the literature as earnings grow, peak, then go down as one gets too old).

One thing to note is that this larger regression “controls” for the variables included. So, the value of 1.578 for education is “taking into account” the relationships between education and gender or age. This is not exactly correct, but its easy to understand. This doesn’t mean you should control for everything by adding unrelated variables to the regression, but R will not call you out on these problems. Computers just do linear algebra, humans are still needed to interpret!

Try It with Diamond Data

Now it’s your turn! With the Diamond data, run some regressions that make sense. For example, high carrat diamonds should have higher prices, according to intuition. Write out a few potential models, then run the regressions! Keep in mind that the data has to be numerical; so for categories, we have to make them numerical somehow…

After you run these regressions, think about interpretations. Consider what other data you might want to add. Play around with adding more categories and watching the coefficients change! You can also play around with transformations of the data (adding 100 to the price of all diamonds. Does this change anything?) and use the

After all this, try to graph some of the relationships using ggplot and the code from before as a template.

If you’re not quite out of time, you can explore exporting the regressions into LaTeX. Use the “stargazer” and “xtable” packages.

If you need any help, don’t hesitate to ask a board member!