Linear Regression Primer

Prepared by Tim Murray, PhD[†]

January 2022

- Linear regression is one of the most commonly used statistical tools used in economics
- A regression allows researchers to estimate the impact of some variable, x, on an outcome, y, while holding other factors constant (i.e., treating them as if they do not change)
- Typically, a research question is framed as: "I want to know the impact of x on y, or, how does y change as x changes"

1 Simple Linear Regression

- In the most basic case, a regression takes the following form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- The regression equation models the relationship between x and y for the population
- i denotes the unit of observation
  - This could be an individual, a house, a city, a state, a country, a school, etc.
- For every research question, there are other factors besides x that influence y, sometimes these factors are unobservable to researchers
  - The impact of these factors on y are captured in what is called the error term, which is represented by ε_i
  - ε_i can be thought of as the "margin of error" or, the variation in y that x does not explain
  - For a regression model to generate unbiased results, it is assumed that, on average, $E(\varepsilon_i) = 0$
- If we were to graph the data with y on the vertical axis and x on the horizontal axis, linear regression estimates the linear relationship between x and y generating a trend line for the data
- The slope of the line can be interpreted as follows: Holding everything else constant, for a given population, for each 1 unit change in x, there is a β_1 unit change in y
  - Alternatively, β_1 can be interpreted as the derivative of y with respect to x, so $\beta_1 = \frac{dy_i}{dx_i}$

[†] Assistant Professor of Economics, Virginia Military Institute. Email: murrayta@vmi.edu.

- However, we usually do not have data for the entire population, so we will have a data set that is a sample of $i = 1, 2, 3, \cdots, N$ units of observation from a larger population
- Linear regression generates an estimate for β_0 and β_1 using the sample of N units
- The estimated values are represented with a "hat", $\hat{\beta}_0$ and $\hat{\beta}_1$
- The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are calculated by drawing a line that has the smallest vertical distance away from each point on the graph

Example

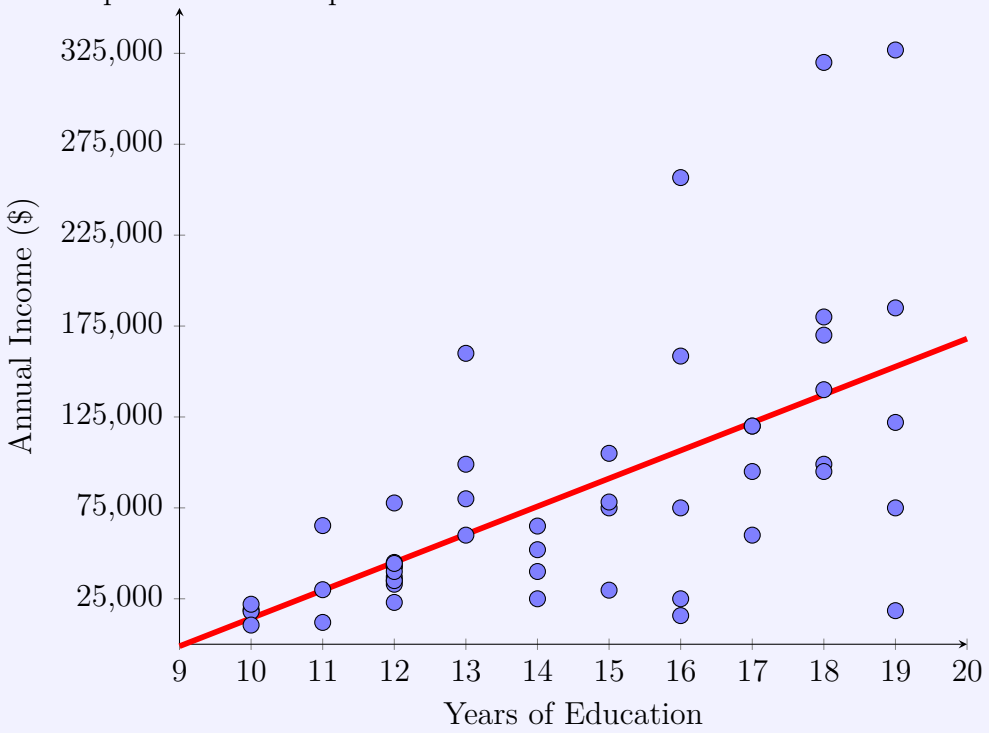
Suppose our research question is as follows: what is the impact of years of education on income?

The regression equation for this research question would look as follows:

$$Income_i = \beta_0 + \beta_1 YearsOfEducation_i + \varepsilon_i$$

For this question, the unit of observation, i, is the individual. So you collect data from 50 people on the number of years they went to school and their annual income. Since there are 50 individuals, our sample size, $N = 50$.

The data that you collected is a sample of the population. By plotting the sample data with income on the vertical axis and years of education on the horizontal axis, the regression will estimate a "line of best fit" that minimizes the vertical distance between each point of the sample and the line:



- The equation for the above regression line is $Income_i = -139,464 + 15,376 YearsOfEducation_i$

- $\hat{\beta}_0$ is the estimated y-intercept. In this example, $\hat{\beta}_0 = -139,464$
- $\hat{\beta}_1$ is the estimated slope of the regression line. In this example, $\hat{\beta}_1 = 15,376$
    - Alternatively, if we think about this in terms of the derivative, then $\hat{\beta}_1 = \frac{dIncome_i}{dYearsOfEducation_i} = 15,376$
- We can interpret $\hat{\beta}_1$ as follows: if we hold everything else constant, for every 1 additional year of education, there is a \$15,376 increase in annual income

2 Multiple Linear Regression

- Generally there is more than one variable that influences changes in y
- Using the previous example, while years of education does effect income, so does ability, years of experience, where you live, the type of job you have, and other factors
- Because of this, we can extend the simple linear regression model to include multiple variables
    - Recall that anything that effects the outcome, y, that is not included in the regression model is assumed to be in the error term, ε_i
    - Since we assume that on average, $E(\varepsilon_i) = 0$, then we must include all variables that we can observe that effect the outcome, y, in our regression model, otherwise the results of the model will be biased
- A multiple regression model takes the following form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 m_i + \varepsilon_i$$

- Linear regression will generate estimated values for $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$
- $\hat{\beta}_0$ is still called the intercept term
- $\hat{\beta}_1 = \frac{\partial y_i}{\partial x_i}$ and can be interpreted as holding everything else constant, a 1 unit increase in x, leads to $\hat{\beta}_1$ unit increase in y
- $\hat{\beta}_2 = \frac{\partial y_i}{\partial z_i}$ and can be interpreted as holding everything else constant, a 1 unit increase in z, leads to $\hat{\beta}_2$ unit increase in y
- $\hat{\beta}_3 = \frac{\partial y_i}{\partial m_i}$ and can be interpreted as holding everything else constant, a 1 unit increase in m, leads to $\hat{\beta}_3$ unit increase in y

3 Testing Significance of Estimated Results

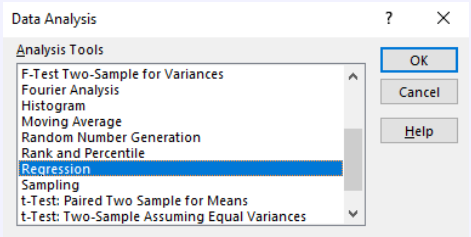
- Because linear regression is estimating the effect of x on y using a sample of data generated from the whole population, we need to be confident that we can trust the result of the estimation
- For each coefficient estimate, linear regression also estimates a standard error
- The standard error of the coefficient tells us how precise our estimate is
- The smaller the standard error, the more precise our estimate is
- We can use the standard error to conduct a hypothesis test to determine if our estimated coefficient is statistically significant or not
    - In order to be statistically significant, we want to know if there is enough evidence from our model and data to suggest that the individual regression estimate is statistically different from zero
- To determine if an individual regression estimate is significant, we use the t-test
    - For the t-test, our null hypothesis is, $H_0 : \hat{\beta}_i = 0$ and the alternative hypothesis is, $H_1 : \hat{\beta}_i \neq 0$
- We can calculate the t-statistic using the following equation: $t = \dfrac{\hat{\beta}_i}{\text{Standard Error}}$
- As a rule of thumb, if the t-statistic > 1.96, then we can reject the null hypothesis and say that $\hat{\beta}_i \neq 0$
    - This means that the regression estimate is significant at the 95% confidence level
- Alternatively, we can use the t-statistic to generate a p-value
- If the p-value for a given estimate is less than 0.05, then we can say that the estimated value, $\hat{\beta}_i$ is statistically significant

4 Software to Conduct Regression Estimations

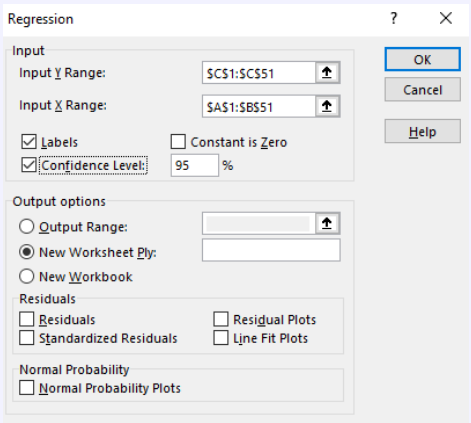
- Regression estimates, standard errors, t-statistics, and p-values are estimated using statistics software
- Some of the most popular software that can estimate regressions are Microsoft Excel, Stata, and R
- Each software will generate regression output in a slightly different format
- While Microsoft Excel can do basic regression analysis, it is recommended using a more advanced statistical software such as Stata or R

Regression in Microsoft Excel

- **Step 1**: Under the Data tab, select Data Analysis (if you do not see the Data Analysis button, you need to enable the Analysis TookPak found under File→Options)
- **Step 2**: Select Regression



- **Step 3**: Select your the range of cells for the y variable and select the range(s) of cells for any x variables



- **Step 3**: The regression output will be generated on a new worksheet

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.607622939
R Square	0.369205636
Adjusted R Square	0.342363322
Standard Error	59363.34025
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	96942682839	4.85E+10	13.75461311	1.98319E-05
Residual	47	1.65628E+11	3.52E+09		
Total	49	2.62571E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-145137.0126	46717.73973	-3.10668	0.003205297	-239120.9823	-51153.043	-239120.9823	-51153.04286
Years of Experience	302.1458958	927.4623284	0.325777	0.746039113	-1563.667645	2167.95944	-1563.667645	2167.959437
Years of Education	15421.42498	2940.558823	5.244386	3.66488E-06	9505.783658	21337.0663	9505.783658	21337.06629

Regression in R

To estimate linear regression in R, it is best practice to use the `feols` command.

In order to use this command, you will need to load the `fixest` library. You can install the `fixest` library using the following command: `install.packages("fixest")`.

```
# Install the fixest package, you only need to do this once
install.packages("fixest")

# Load the fixest library
library(fixest)

# FEOLS command syntax: feols(y ~ x1 + x2 + ··· + xN, data=FileName)
feols(Income ~ Education + Experience, data=income)
```

The regression output will be displayed as follows:

```
OLS estimation, Dep. Var.: Income
Observations: 50
Standard-errors: IID
                Estimate Std. Error   t value   Pr(>t)
(Intercept) -145137.013  46717.740 -3.106679 3.2053e-03 **
Education     15421.425   2940.559  5.244386 3.6649e-06 ***
Experience      302.146    927.462  0.325777 7.4604e-01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 57,554.9   Adj. R2: 0.342363
```

Regression in Stata

To estimate the linear regression in Stata, use the `regress` command. The first variable is your y variable, followed by any x variables

```
* Regression command syntax: reg y x1 x2 ··· xN
reg Income Education Experience
```

The regression output will be displayed as follows:

```
      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F(2, 47)        =     13.75
       Model |  9.6943e+10         2  4.8471e+10   Prob > F        =    0.0000
    Residual |  1.6563e+11        47  3.5240e+09   R-squared       =    0.3692
-------------+----------------------------------   Adj R-squared   =    0.3424
       Total |  2.6257e+11        49  5.3586e+09   Root MSE        =     59363

------------------------------------------------------------------------------
      Income | Coefficient  Std. err.      t    P>|t|     [95\% conf. interval]
-------------+----------------------------------------------------------------
   Education |   15421.42   2940.559     5.24   0.000     9505.784    21337.07
  Experience |   302.1459   927.4623     0.33   0.746    -1563.668    2167.959
       _cons |    -145137   46717.74    -3.11   0.003      -239121   -51153.04
------------------------------------------------------------------------------
```