Econ 103 Week 1

Manu Navjeevan

January 5, 2020

1 Introduction

I am Manu Navjeevan, a second year PhD student in the Economics department. My research interests are in Econometrics and Industrial Organization. I will be your TA this quarter.

If you have questions you can email me at mnavjeevan@g.ucla.edu or come to office hours which are (tentatively) on Tuesdays and Thursdays from 11am - 12pm (Noon) in the Alper Room.

2 Linear Regression

We have all probably encountered linear regression in the past or at least have some idea of what linear regression is. The goal of this course will be to formalize this intuition and gain a better sense of what *exactly* linear regression is. Along the way we will expose both the strengths and drawbacks of linear regression.

To begin, suppose we have an explanatory variable X and a dependent variable Y. For example, let X be age and Y be earnings.[1] The data on age vs. earnings can be plotted in a scatter plot which could look like the one below in Figure 1. A linear regression model relates the explanatory variable X to the dependent variable Y by the formula below. For a number of reasons (prediction, etc.) we may be interested in estimating this model.

$$\begin{aligned} Y_i &= \alpha + X_i \cdot \beta + \epsilon_i \\ \epsilon_i &\overset{i.i.d}{\sim} (0, \sigma^2) \\ 0 &= E[\epsilon_i | X_i] \end{aligned}$$

The subscript i denotes a member of the population. Here $\epsilon_i \overset{i.i.d}{\sim} (0, \sigma^2)$ simply means that the errors are independently and identically distributed with mean 0 and constant

[1] In general X can have multiple dimensions; e.j X could contain both age and education. For the expository purpose of this example, we will only consider a unidimensional X

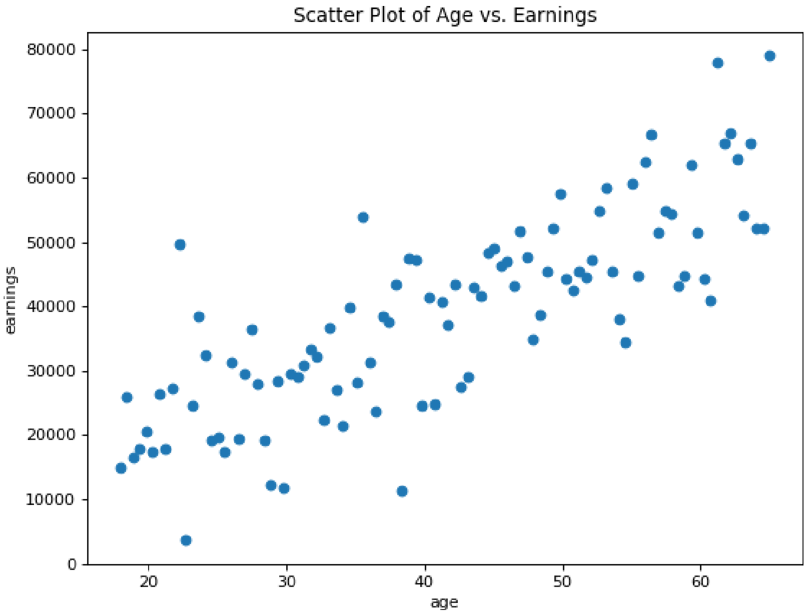


Figure 1: Scatter plot of age against earnings. We may be interested in the linear relationship between age and earnings.

variance σ^2. It is important to note that the the linear regression model specified above is specified for the *whole population*. That is it describes the "true" linear relationship between X and Y in the population. Since we only have a random and finite sample of the population, we can only try and estimate α and β. The logical way of doing this is to find the line of best fit through our data. Heuristically, the best estimate of the linear relationship between X and Y in the population is the linear relationship (line of best fit) between X and Y in our data. We now turn to the problem of estimating this line of best fit

2.1 Estimating the Linear Regression Model

The regression coeffecients are estimated by choosing $(\hat{\alpha}, \hat{\beta})$ to minimize the sum of squared errors between Y and it's predicted value $\hat{\alpha} + \hat{\beta} \cdot X_i$. Formally:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{a,b} \sum_{i=1}^{n} (Y_i - a - X_i \cdot b)^2$$

This is a bit complicated to solve so thankfully someone has done the math before us to get closed form expressions for $\hat{\alpha}$ and $\hat{\beta}$. These expressions are given below. Even more thankfully, computers now estimate $\hat{\alpha}$ and $\hat{\beta}$ for us.

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$
$$\hat{\alpha} = \bar{Y} - \hat{\beta} \cdot \bar{X}$$

Once we have found $\hat{\alpha}$ and $\hat{\beta}$, we can use them to define a predicted value for Y for any value of X. We denote this predicted value $\hat{Y}$ and note that

$$\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X_i$$

Using this we can also estimate the residuals of the model (ϵ_i), which we will later look at to see if our linear regression model is a good fit for the data. For now, we will just define the estimated residuals

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

Because $\hat{\alpha}$ and $\hat{\beta}$ are chosen to minimize the squared error between Y and its linear predictor $\hat{Y}$, the line $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$ gives us a line of best fit through our data. We can see this by plotting the estimated regression line from our earlier scatter plot (Figure 1). This is seen below in Figure 2.

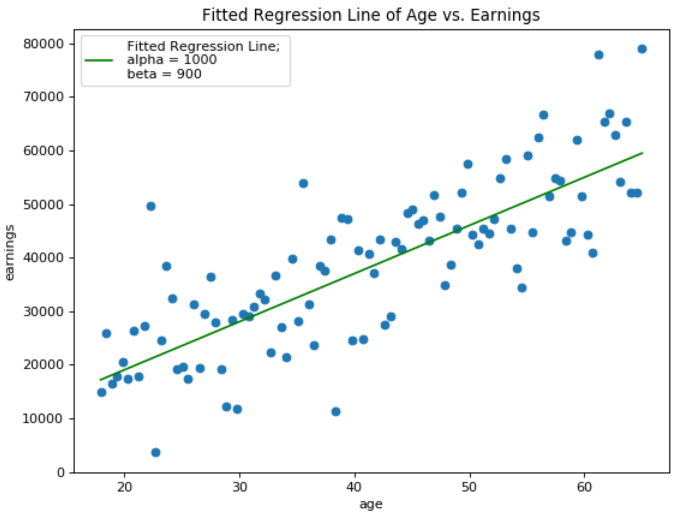


Figure 2: The estimated regression line gives us a line of best fit for the data

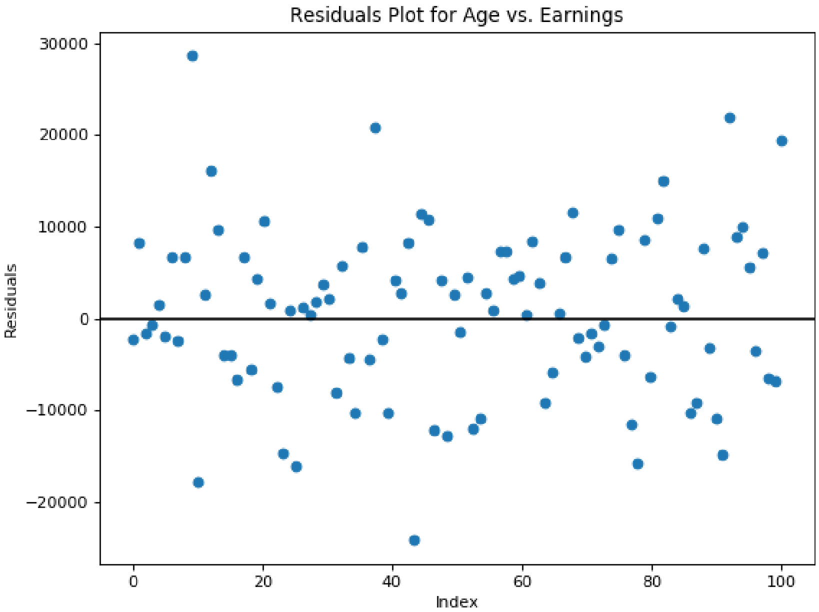


Figure 3: Residual Plot of the Age vs. Earnings regression. The residuals plot does not give us reason to think the assumptions of the model are violated

3 Evaluating the Regression Model

3.1 Looking at the Estimated Errors

It is important at this point to remember that we are interested in the linear relationship between X and Y specified by the linear model $Y_i = \alpha + X_i \cdot \beta + \epsilon_i$ where our error terms ϵ_i are distributed independently with mean 0 and finite variance. If any of these assumptions on our error terms ϵ_i are violated, then our model is not a good fit for the data and we cannot say that our $\hat{\beta}$ is a good estimate of the true β. To make sure that these assumptions are not being violated, it is important to look at the plot of our estimated residuals. Want we want to look for is relationships between errors or evidence that the variance of the residuals are not the same for all residuals (i.e the variance increases with age). We check the residuals of our regression of age against earnings in Figure 3

Thankfully, upon looking at the resudials plot of our regression, we do not find evidence that the assumptions of our model are violated.

3.2 Goodness of Fit

Once we have made sure that the assumptions of our model are satisfied, we may be interested in how well the model fits the data. Estimating the linear relationship between X and Y may not be of much use if there is not a very tight relationship between X and Y (i.e if the data points are not very close to the regression line). To do this, we generally use the R and R^2 coeffecients, which tell us about the strength of the relationship between X and Y in our data.

The R correlation coeffecient is bounded between -1 and 1, and tells us both about the strength and direction of the linear relationship between X and Y. A R coeffecient of -1 indicates a perfect negative relationship between X and Y, whereas a R coeffecient of 1 indicates a perfect positive relationship. A R coeffecient of 0 indicates no linear relationship between X and Y. The R coeffecient is calcualted

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y_i})}{(n-1)s_X s_Y}$$

where s_X and s_Y are the sample sample deviation of X and Y, respectively. You may notice at this point that the linear regression model is reminicent of the $\hat{\beta}$ coeffecient and indeed we can write $\hat{\beta}$

$$\hat{\beta} = R\frac{s_Y}{S_X}$$

Showing this is an exercise left to the reader. The other goodness of fit statistic we are interested in the time being is R^2. As the name would suggest, for simple (univariate) linear regression this is calculated by squaring the R coeffecient. For more advanced models, R^2 is calcualted using the estimated resudials of the model, but we will get to that in time. For now what is important to know si that the R^2 coeffecient can be interpreted as the fraction of the variance in the dependent variable Y that can be described bt the linear model of the X variable specified. For example a R^2 of 0.75 would mean that 75% of the variance in Y can be explained by the linear model with X.

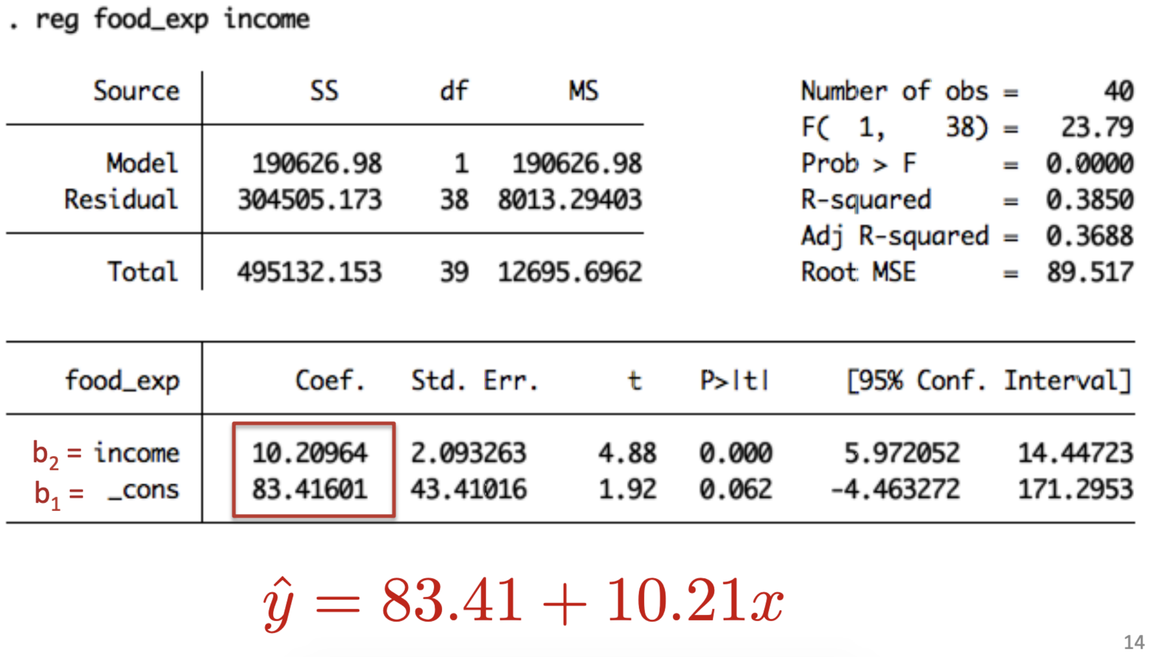
4 Linear Regression in Stata

As mentioned before, linear regression can now be done on computers. A popular software in economics to do linear regression (and other, more advanded, statistical analysis) is Stata. To run a linear regression of Y (dependent) against X (explanatory) in Stata simply load your data set (for this example named 'data.dta') into Stata using the command

```
use data.dta
```

We can then run a regression of Y against X using the command reg Y X

In the example below income is used to explain food expenditures



```
. reg food_exp income

      Source |       SS           df       MS      Number of obs   =        40
-------------+----------------------------------   F(1, 38)        =     23.79
       Model |   190626.98         1   190626.98   Prob > F        =    0.0000
    Residual |  304505.173        38  8013.29403   R-squared       =    0.3850
-------------+----------------------------------   Adj R-squared   =    0.3688
       Total |  495132.153        39  12695.6962   Root MSE        =    89.517

------------------------------------------------------------------------------
    food_exp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
b2 =  income |   10.20964   2.093263     4.88   0.000     5.972052    14.44723
b1 =   _cons |   83.41601   43.41016     1.92   0.062    -4.463272    171.2953
------------------------------------------------------------------------------
```

$$\hat{y} = 83.41 + 10.21x$$

Figure 4: Sample regression output from Stata. In this case income is being used to explain food expenditures

The coeffecients of the model ($\hat{\alpha}$ and $\hat{\beta}$) are found in the lower table under the "Coef." column. $\hat{\beta}$ is the coeffecient associated with the explanatory variable, in this case income, wheras $\hat{\alpha}$ is the coeffecient associated with the constant.

5 Practice Problem

1. **Interpretation** (Edition 4, Problem 2.6) A soda vendor at Louisiana State University football games observes that more sodas are sold the warmer the temperature at game time is. Based on 32 home games covering five years, the vendor estimates the relationship between soda sales and temperature to be $\hat{y} = -240 + 8x$ where $y =$ number of sodas she sells and $x =$ temperature in degrees Farenheit.

   (a) Q: Interpret the slope and intercept. Do the estimates make sense?
   A: The interecept is -240. This means that if there temperature is 0 degrees

Farenheit, our model predicts that - 240 sodas would be sold. This is clearly a nonsensical prediction and suggests our model is not a good fit for the data at the low end of the temperature distribution.
The slope is 8. This means that, holding all else constant, we expect a one degree increase in temperature to be associated with an 8 unit increase in sodas sold. This prediction is in line with prior beliefs.

(b) Q: On a day when the temperature at game time is forecast to be 80 degrees Farenheit,predict how many sodas the vendor will sell.
A: Using our estimated model $\hat{y} = -240 + 8 * 80 = 400$. Thus we would expect 400 sodas to be sold.

(c) Q: Below what temperature are the predicted sales zero?
A: To do this, we want to find the x at which $\hat{y}$ is 0, let's call this $\tilde{x}$. Since $\hat{y}$ is increasing in x, if x is lower that $\tilde{x}$ we would expect predicted sales to be zero (or more specifically negative, but we know this is impossible). To find $\tilde{x}$, set $\hat{y}$ equal to 0 in the estimated regression equation:

$$\begin{aligned} 0 &= -240 + 8\tilde{x} \\ 240 &= 8\tilde{x} \\ 30 &= \tilde{x} \end{aligned}$$

Thus, when the temperature is below 30 degrees, the predicted sales are 0.

(d) Q: Sketch a graph of the estimated regression line
A: See below:

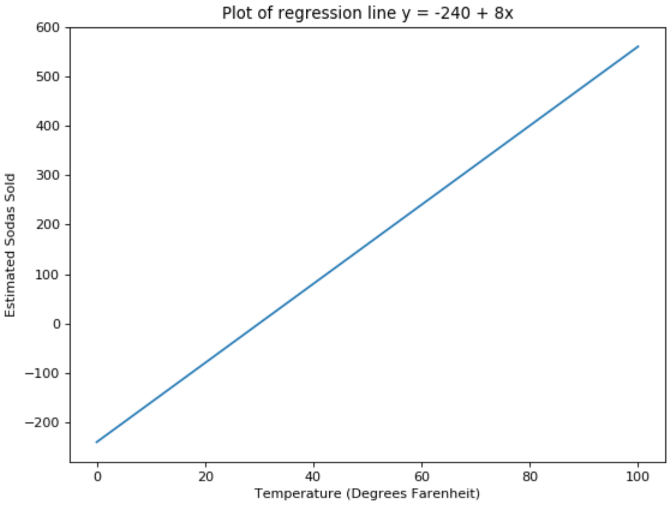

Plot of regression line y = -240 + 8x
600
500
400
300
200
100
0
−100
−200
Estimated Sodas Sold
0
20
40
60
80
100
Temperature (Degrees Farenheit)

Figure 5