In the previous section we looked at why a regression approach is better to identify the relationship between variables than using something like a visualization or a correlation. Remember there are many methods of data analysis. We can create graph and charts to assess patterns in the data. We can look at summary statistics to know about the range, the spread, and the central tendency of the data. We can look at correlations between variables to understand how they change with one another.

Regression analysis is one type of data analysis method and it is used for very specific instances. To understand relationship between factors when one suspect that the casual relationship is excess between the factors and there are multiple factors that influence a variable that we are interested in.

Let's start with understanding the simplest type of a regression model which is a linear regression model.

>>

In the simplest kind of linear regression model, we are going to assume that there is only one variable that influences the target variable. Remember the target variable is the variable that we are trying to understand. In this case baby birth weight. Mathematically we are saying that baby birth weight is sum function of gestation weeks.

Remember this is a simple regression. So we are assuming that there is only one variable that

influences birth weight. Obviously this is an assumption and we are going to relax it very soon but just to understand the logic of the regression model.

Birthweight = f (gestation weeks)

What we are trying to determine here in other words is, what is the functional form of that relationship between gestation week and birth weight.

Currently we are assuming that it is a linear function because we are working on a linear regression model. What is the linear relationship mean? A linear relationship between two variables is essentially depicted as a straight line relationship.

>>

If you remember mathematics from high school, the equation that denotes the straight line relationship between two variables, x and y is

$$Y = mx + c$$

Where, m = Slope and c = intercept.

What is slope? The rate of change of Y when X changes or the magnitude of the impact of changes in X on Y. The slope is the rate at which there is changes in Y because of changes in X. What is the intercept? The intercept is the value of Y when X = 0. Let's understand this visually.

>>

This is a straight line relationship. In fact the straight line equation for this relationship is Y = 2 + 3X. Here 2 is the intercept and 3 is the slope. What is that mean? Supposing X changes by one unit, what will be the change in Y. It will be 3 units. Let's calculate that. If X = 1; Y = 5, if X = 2; Y = 8. So when X changes from one to two what happens to the change in Y? It moves from 5 to 8. So the change in Y is 3 relative to a unit change in X and that is the slope.

When you have a straight line relationship, irrespective at the level of which X changes whether X changes from 1 – 2, 2 – 3, 3 – 4, or 20 – 21. The rate of change of Y is always constant. In this example it is 3. If you have a non-linear relationship, then you have a curved line. The rate of change of Y relative to X is not constant. What is that mean? The rate at which Y changes, let's say between x = 1, and x = 2 may not be the same at the rate at which Y changes when X = 10 to X = 11. Because you will have a curved line. So the slope is not going to be constant. So essentially a linear relationship is when we believe that there is a straight line relationship between X and Y. every time X changes by one unit, Y changes by the same amount, which is the slope.

>>

What happens if the slope is zero then Y = 2. In other words this is a straight line that is parallel to the X axis. Why it does not depend on X. because X = 0. When X is zero, Y is zero. When X is one Y is

also zero, when X is 2 Y is also zero. In other words Y does not depend on X.

What happens when they intercept a zero? If they intercept a zero, you have a straight line that now passes through the 0,0 point. In other words there is no intercept. Remember the intercept is the value of Y when X = 0. If X = 0; Y = 2. Intercept for the Y = 2 + 3X line is 2. But if we make the intercept zero, then you have a line with the same slope but it now passes through the origin. That is what the straight line relationship is.

>>

In a simple linear regression model, we are assuming that there is a straight line relationship between the X and the Y. Remember Y is the target variable, the variable that we are interested in understanding the birthweight. X is the independent variable or influencing variable. In this example, it is the number of weeks of gestation. Since we are looking at a straight line relationship in a linear regression model, the linear regression equation is typically displayed like this.

$$Y = \beta_0 + \beta_1 X + e$$

where, β_0 = Intercept

β_1 = Slope

e = Error

If you are simply look at

$$Y = \beta_0 + \beta_1 X$$

This is the same as saying y= mx + c. Here m is β_1 and c is β_0. We also have another term called e which is the error. Error very simply and we will come back to it is random variation. So for now we are going to ignore random variation. We are going to assume that the average of random variation is zero. Because it is random and we are going to say that the linear regression model is essentially

$$Y = \beta_0 + \beta_1 X$$

In other words we think that the straight line relationship between the number of weeks of gestation and birthweight and that straight line equation is captured as

$$Y = \beta_0 + \beta_1 X$$

In order to understand that equation, that relationship, we need to know what is the value of β_0 and value of β_1. In a linear regression model our essential effect is directed at estimating what are values of β_0 and β_1. Because once I have the values of β_0 and β_1, I know exactly what is the relationship between my X variable and my Y variable.

>>

Let's just quickly review some terminology for simple linear regression models. We have use the word, dependent: Y: predicted; the variable whose behaviour we hypothesized can be explained or influenced by other factors is called dependant

variable, the target variable, the Y variable or the predicted variable. In our example, the target variable are the dependent variable is infant birth weight.

Independent variables are X(s), also called predictors are the factors that we think influence the dependant variable. In our current example gestation period also the education of the mother, the smoking status, and the race of the mother. All of these are independent variables.

The Beta Coefficients are the estimates of magnitude of the impact of changes in the predictors on the predicted variable. In other words, the Beta Coefficients tell us what is the rate at which changes in X influence Y.

The error also called e or u is the impact of unobserved variables on the dependent variable. Remember we are dealing with random samples. So they will be some random variation. For now let's understand error as random sampling variation

>>

Coming back to the birth weight example, the linear regression model says

$$Y = \beta_0 + \beta_1 X + e$$

Essentially we are saying

Birthweight = β_0 + β_1 *Gestation period + e

What we need to do now is look at the data and use the data to understand what should be the right values for β_0 and β_1.

>>

The Ordinary Least Square regression technique is what we are going to use to estimate these coefficient values. The Ordinary Least Square regression, also called OLS estimates coefficients on the variables by identifying the line that minimize the sum of squared differences between point on the estimated line and the actual values of the independent variable. This is a mouthful.

Let's understand what exactly an ordinary least square regression does? Remember one way to look at a relationship between two variables is to do a visualization. If we look at birthweight and gestation and I simply have done a scatter plot. We can see that they are exist some relationship. In this example, we are saying that this is a straight line relationship that may or may not always be the case.

This is approximately straight perhaps. If we assume that there is a straight line relationship. I should be able to draw a line that captures the relationship between birthweight and gestation. But if you look at this relationship, how many lines are possible? Only one, many lines are possible. If we want to capture a relationship between a birthweight and gestation and do that using a straight line. I can draw many straight lines through this data points.

Which is the best possible straight line? Remember that they will not be a single straight line that will capture all the data points. Because there is random variation. We know that other factors influence birthweight but what is the best possible line. That best captures a relationship between birthweight and gestation. How do I choose this best possible line?

That is essentially what an ordinary least square regression estimation does. It tries to identify the best possible line, the slope, and the intercept on the best possible line are essentially are Beta coefficients.

>>

Now a problem becomes. How do we find one line that best captures the relationship between the X variable and the Y variable. Let's think about it. What you think is a good method for us to identify the best possible line. Lot of us say that the best possible line is a line that covers as many data points as possible. Is that valid? The best possible line is a line that goes through as many data points as possible. That's not necessarily true.

You could have a line that goes through a lot of data points but it could be very far from a lot of the other data points. In fact if you think about it a simple way of identifying the best possible line is to say the best possible line through all these points is a line that is as close as possible as many points as possible. It sounds simple.

Let's just say that again. The best possible line is the line that is as close as possible to as many points as possible. How do we find that out? What is it mean to say as close as possible to as many points as possible?

>>

Imagine that you drew one line through the data points and then you started calculating the distance of each point from the line. Now this line will go throw some of the points. Some of the points are going to be on the line. So the distance will be zero. But it will not go through all the data points.

Supposing you have a data point here and the line is here. We know what the distance is. We can get back from looking at the vertical distance on the Y axis. Similarly for the second point we know what this distance is. This could be a negative distance and this could be a positive distance. Negative meaning below the line and positive meaning above the line. So we could draw this line, any line and then calculate what is the sum of the differences of each point from the line.

If you think about it the best possible line is a line that is as close as possible as many points as possible. So if I draw all possible lines through the point, I calculate this distances, I sum it up, and then the line that has the minimum distance is the best possible line. In fact we will go one step further. We will square the differences and then

sum them up. So when we calculate distance, remember some distances are positive and some distances are negative. We will square the differences and then sum up the distances.

If you think about it intuitively the line that has the minimum total sum of square distances has to be the best possible line. Because that is the line that is as close as possible to as many points as possible and that is essentially what we do in an ordinary least square regression technique. We draw lines through the data points and for every possible line, we can calculate the distance of every point from the line. The line that minimizes the sum of the square differences is the best possible line. If we look at the definition it makes a lot more sense.

>>

Identify the line that minimizes the sum of squared differences between points on the estimated line and in the actual values of the independent variable. This is an ordinary lest square regression. Now we understand the logic of calculating differences and squaring the differences. But do we want to really do this manually? No it turns out that there are mathematical ways of identifying the Beta coefficients that minimize the differences between the points and the line.

Mathematically if you remember calculus what we want to do is we want to minimize

$$Q = \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2$$

If you use differential calculus, you will get

$$b_o = \frac{\Sigma X_i^2 \Sigma Y_i - \Sigma X_i \Sigma X_i Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2} \qquad b_1 = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

This looks scary but they are all essential functions of X values and Y values. This is the big mouthful of formula. Remember this is only depend on X and Y and we don't even have to remember this. These are generated automatically by any tool that we use. But what we are saying is that if you calculate your β_0 and β_1 using this formula then the line that you generate $Y = \beta_0 + \beta_1 x$ automatically will be the possible best line. These estimates are called the ordinary least square estimates.

We can be sure that given this data, the ordinary least square estimated line minimizes the errors which is the difference between the points and the line more than any other line that we choose.

>>

Essentially this is what we do in a regression model. We use the data to estimate β_0 and β_1. How do we do that? We have formula for β_0 and β_1 that depend on the X and Y values. Once we do that we essentially have an equation that says

Birthweight = Intercept estimate + Beta coefficient * Gestation

This is the best fitted line. Remember what does that mean? This line will not cover every single point on the scattered plot. But we can be confident that this line will be as close as possible to as many points as possible on the scattered plot.

>>

That is what a simple linear regression model does. Let's look at how to actually generate this Beta coefficients using a tool.

>>

Let's now look at how to run a simple liner regression in excel. In order to create a regression model in excel, we have to use the data analysis option. In data analysis, we need to choose regression. This is the data that we have. Now I am going to run a simple linear regression. For that mean size I am going to have only one dependant variable obviously and one independent variable. I am going to choose gestation weeks as my independent variable and birthweight in grams is my dependant variable.

Now we go to data, data analysis, and choose regression. When we do that excel has a menu that pops up and asking for the Y and X range. Y is my dependent variable or my target variable which in this case is the grams variable. Since I have 1115 observations I am going to make this range 1116

because the first row is labels. This is my Y range. My X range is now gestation.

I have specified my Y range and X range. Remember here because we are dealing with the simple regression choosing only one independent variable but we could have multiple independent variables as well. In which case you would choose multiple variables in the X range. In this example we have labels. We are going to stick to a confidence level of 95% and this is pretty much all we need is to simply run the regression.

There are some other options as well. Output option, would you want the output in a same work sheet, in a new work sheet, or in a new work book. There are some options for generating additional output which is residuals or errors and some plots, line fit plots, normal probability plots etc. I am not going to include them for now. I am going to simply run a very simple regression model.

One of the things to remember with excel, you can have at the most 16 independent variables. If you have more than 16 variables then you can't run the regression in excel. Also in excel you can't have a missing values in the data. If you have missing values then excel will essentially show you an error that says invalid data type or invalid data in the data range. Make sure that we don't have any missing values.

Now I am going to click ok and because I am requesting the output in a new work sheet excel will generate regression output in a new

worksheet. So this is the standard regression output that excel generates when we run a simple regression using the regression option in the data analysis tab. We will look at how to interpret all of these output and understand what the regression model is telling us.

>>