

## Simple Linear Regression

Previously we've discussed the notion of inference – using sample statistics to find out information about a population. This was done in the context of the mean of a random variable. In addition we talked early on about correlation between two variables. The ANOVA analysis allowed us to examine the sources of variation for a particular outcome, but did not really get into specific causal relationships.

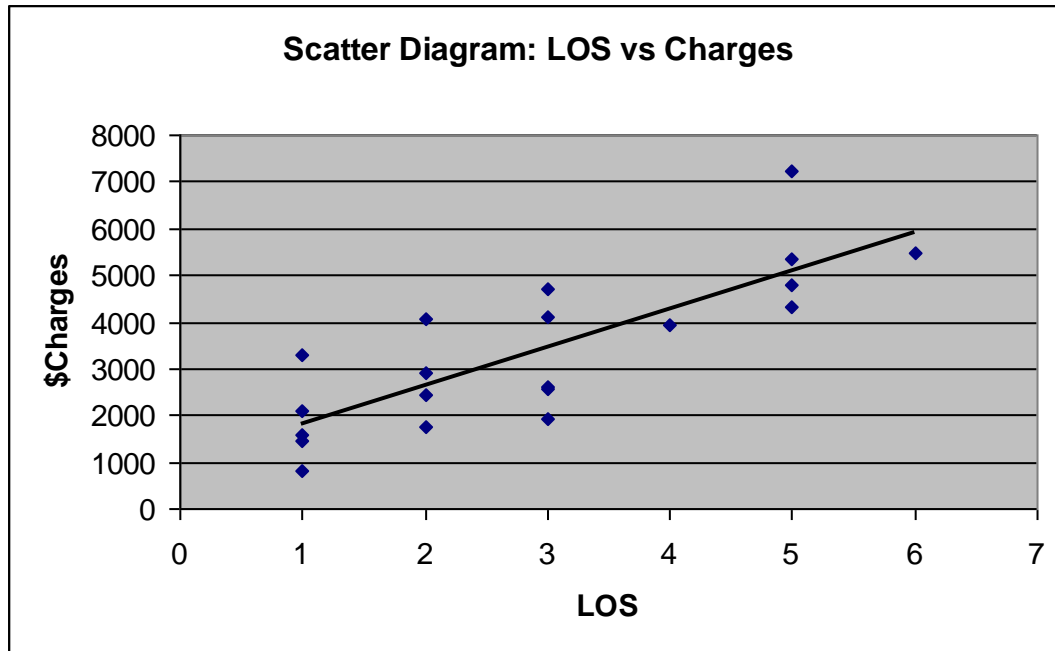
In this chapter we want to take this a step further and formalize the relationship between two variables, and then extend this to multivariate analysis. This is done through the concept of regression analysis.

Suppose we are interested in the relationship between two variables: length of stay and hospital charges. We think that LOS causes charges—a longer LOS results in higher charges. How would we go about testing this? If we take a sample of individual stays and measure their LOS and charge, we might get the following:

| Stay | LOS | Charges |
|------|-----|---------|
| 1    | 3   | 2614    |
| 2    | 5   | 4307    |
| 3    | 2   | 2449    |
| 4    | 3   | 2569    |
| 5    | 3   | 1936    |
| 6    | 5   | 7231    |
| 7    | 5   | 5343    |
| 8    | 3   | 4108    |
| 9    | 1   | 1597    |
| 10   | 2   | 4061    |
| 11   | 2   | 1762    |
| 12   | 5   | 4779    |
| 13   | 1   | 2078    |
| 14   | 3   | 4714    |
| 15   | 4   | 3947    |
| 16   | 2   | 2903    |
| 17   | 1   | 1439    |
| 18   | 1   | 820     |
| 19   | 1   | 3309    |
| 20   | 6   | 5476    |

Just looking at the data there looks to be a positive relationship, individuals with more LOS seem to have a higher charge; but not always.

Another way of looking at the data would be with a scatter diagram:



Ignore the line for now.

Charge is on the Y axis, LOS is on the X. It looks like there is a positive relationship between charges and LOS. Simply eyeballing the data, it looks like the dots go up and to the right.

The trendline basically connects the dots as best as possible. Note that the slope of this line will tell you the marginal impact of LOS on charges: what happens to charges if LOS increases by 1 unit. This line intersects the Y axis just above zero. This would be the predicted charge if LOS was zero.

If the correlation coefficient between LOS and charges was 1 then all the dots would be on this line. Note that for some observations the line is real close to the dot, while for others it is pretty far away. Let the distance between the dot and the line be the error in the trendline, The trendline is drawn so that the error term is minimized. Since errors above the line will be positive, while errors below the line will be negative we have to be careful – positive errors will tend to wash out negative errors. Thus a strategy in estimating this line would be to draw the line such that we minimize the squared error term.

This is known as **The Least Squares Method**.

### I. The Logic

The idea of Least Squares is as follows:

In theory our relationship is:

$$Y = \beta_0 + \beta_2 X + \varepsilon$$

Y is the dependent variable – the thing we are trying to explain.  
X is the independent variable – what is doing the explaining.

$\beta_0$  and  $\beta_1$  are population parameters that we are interested in knowing

In our case Y is the charge, X is LOS.  $\beta_0$  is the intercept (where the line crosses the Y axis), and  $\beta_1$  is the slope of the line. This coefficient  $\beta_1$  is the marginal impact of X on Y.

These are population parameters that we do not know. From our sample we estimate the following:

$$Y = b_0 + b_1X + e$$

Note I've switched to non-Greek letters since we are now dealing with the sample. So  $b_0$  is an estimator for  $\beta_0$  and  $b_1$  is an estimator for  $\beta_1$ .  $e$  is the error term reflecting the fact that we will not always be exactly on our line. If we wanted to get a predicted value for Y (charges) we would use:

$$\hat{Y}_i = b_0 + b_1X_i \quad \{\text{the } \hat{\text{ means predicted value}}\}$$

Note the error term is gone. So we are looking at the line. So suppose that  $b_0=3$  and  $b_1 = 2$ , then someone with a LOS of 5 days would be predicted to have  $3+2*5=\$13$  in charges, etc.

Least squares finds  $b_0$  and  $b_1$  by minimizing the sum of the squared difference between the actual and predicted value for Y:

## II. Specifics

$$\text{Sum of squared difference} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Substituting:

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n [Y_i - (b_0 + b_1X_i)]^2$$

Thus least squares finds  $b_0$  and  $b_1$  to minimize this expression. We are not going to go into the details here of how this is done, but we will focus on the intuition of what is going on.

The easiest way to think about it is to go back to the scatter diagram: least squares draws the trend line to connect the dots the best way possible. We choose the parameters to minimize the size of our mistakes or errors.

### III. How do we do this in Excel?

Excel can do both simple (one dependent variable) and multiple (more than one) regression.

You need the Data Analysis Toolpack to do it.

#### SUMMARY OUTPUT

| <i>Regression Statistics</i> |          |
|------------------------------|----------|
| Multiple R                   | 0.807337 |
| R Square                     | 0.651794 |
| Adjusted R Square            | 0.632449 |
| Standard Error               | 995.4983 |
| Observations                 | 20       |

#### ANOVA

|            | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-----------|-----------|----------|-----------------------|
| Regression | 1         | 33390795  | 33390795  | 33.69347 | 1.69E-05              |
| Residual   | 18        | 17838305  | 991016.9  |          |                       |
| Total      | 19        | 51229100  |           |          |                       |

|           | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| Intercept | 997.4659            | 465.7353              | 2.141701      | 0.046147       | 18.99164         | 1975.94          |
| LOS       | 818.8394            | 141.0671              | 5.804607      | 1.69E-05       | 522.4681         | 1115.211         |

What does all this mess mean?

Skip the first two sections for now and just look at the bottom part. The numbers under Coefficients are our coefficient estimates:

$$\text{Charge}_i = 997.5 + 818.8 \cdot \text{LOS}_i + e_i$$

So we would predict that someone with zero LOS would have charges of \$997.5 (whatever that means) and each day adds 818.8 to your charge.

The least squares estimate of the effect of LOS on charges is \$818.8 per day. So someone with a LOS of 5 years is predicted to have:  $997.5 + 818.8 \cdot 5 = \$5091.5$  in charges.

The statistics behind all this is pretty complicated, but the interpretation is easy. And note that we can now add as many variables as we want, and the coefficient estimate for each variable is calculated holding constant the other right-hand-side variables.

#### IV. Measures of Variation:

We now want to talk about how well the model predicted the dependent variable. Is it a good model or not? This will then allow us to make inferences about our model.

##### The Sum of Squares

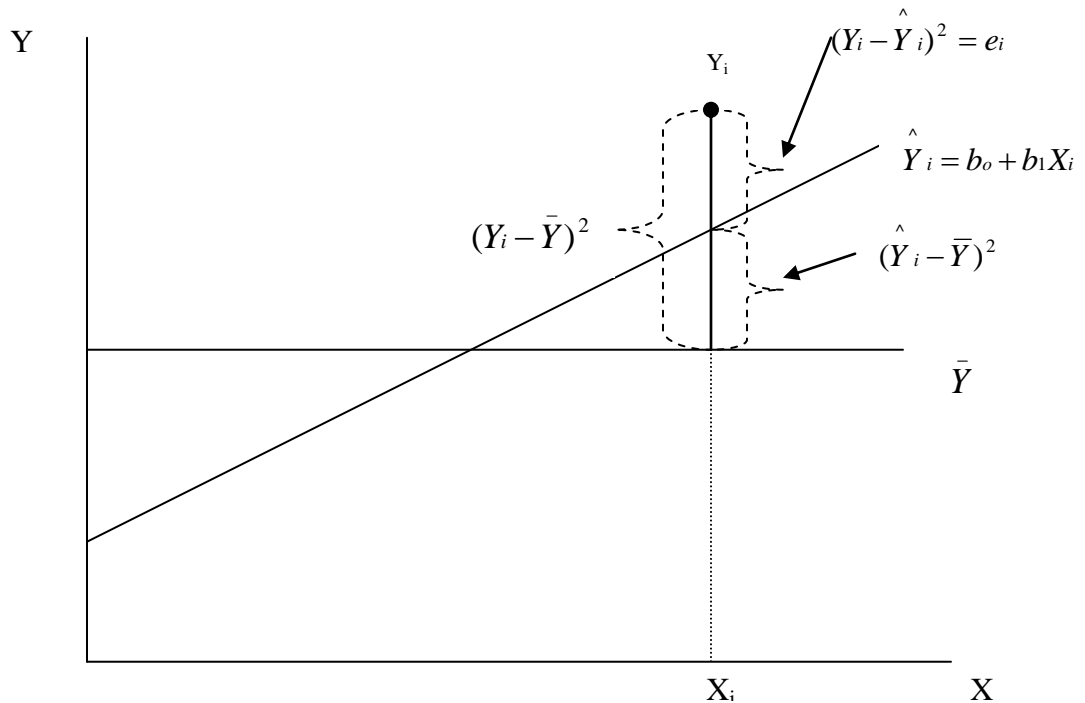
The total sum of squares (SST) is a measure of variation of the Y values around their mean. This is a measure of how much variation there is in our dependent variable.

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

[Note that if we divide by n-1 we would get the sample standard deviation for Y]

The sum of squares can be divided into two components

Explained variation, or the Regression Sum of Squares (SSR) – that which is attributable to the relationship between X and Y, and unexplained variation, or Error Sum of Squares (SSE) – that which is attributable to factors other than the relationship between X and Y.



The dot, represents one particular actual observation  $(X_i, Y_i)$ , the horizontal line represents the mean of Y ( $\bar{Y}$ ), the upward sloping line represents the estimated regression line. The distance between  $Y_i$  and  $\bar{Y}$  is the total variation (SST). This is broken into two parts, that explained by X and that not explained. The distance between the predicted value of Y and the mean of Y is that part of the variation that is explained by X. This is SSR. The distance between the predicted value of Y and the actual value of Y is the unexplained portion of the variation. This is SSE.

Suppose that X had no effect whatsoever on Y. Then the best regression line would simply be the mean of Y. So the predicted value of Y would always be the Mean of Y no matter what X is. So X is doing nothing in helping us to explain Y. Then all the variation in Y will be unexplained.

Suppose, alternatively, that the predicted value was exactly correct – the dot is on the regression line. Then notice that all the variation in Y is being explained by the variation in X. In other words, if you know X you know Y exactly.

As shown above, some of the variation in Y is due to variation in X  $((\hat{Y}_i - \bar{Y})^2)$  and some of the variation is not explained by variation in X  $((Y_i - \hat{Y}_i)^2)$ .

So to get the SSR we calculate  $SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$

And to get the SSE we calculate:  $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

Referring back to our first regression output notice the middle table looks as follows:

| ANOVA      |           |           |           |          |                       |
|------------|-----------|-----------|-----------|----------|-----------------------|
|            | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
| Regression | 1         | 33390795  | 33390795  | 33.69347 | 1.69E-05              |
| Residual   | 18        | 17838305  | 991016.9  |          |                       |
| Total      | 19        | 51229100  |           |          |                       |

The third column is labeled SS (or sum of squares), then the first row is regression So  $SSR = 33390795$ . Residual is another word for error (or leftover) so  $SSE = 17838305$ , and the  $SST = 51229100$ . Notice that:  $51229100 = 33390795 + 17838305$

How do we use this information? In general, the method of Least Squares chooses the coefficients so to minimize SSE. So we want that to be as small as possible – or equivalently, we want SSR to be as big as possible.

Notice that the closer SSR is to SST the better our regression is doing. In a perfect world  $SSR = SST$ : or our model explains ALL the variation in Y. So if we look at the ratio of SSR to SST, this will tell us how our model is doing. This is known as the Coefficient of Determination or  $R^2$ .

$$R^2 = SSR/SST$$

for our example:  $R^2 = 33390795/51229100 = .652$

Is this bad? It depends.

Thus, 65% of the variation in charges can be explained by variation in LOS. Note that this is pretty low since there are many other things that determine charges.

### **Standard Error of the Estimate**

Note that for just about any regression all the data points will not be exactly on the regression line. We want to be able to measure the variability of the actual Y from the predicted Y. This is similar to the standard deviation as a measure of variability around a mean. This is called The Standard Error of the Estimate

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$$

Notice that this looks very much like the standard deviation for a random variable. But here we're looking at variation of actual values around a prediction.

For our example  $S_{YX} = \sqrt{\frac{17838305}{18}} = 995.5$

Note that the top table in the Excel output has the R-squared and the Standard Error listed, among other things.

This is a measure of the variation around the fitted regression line – a loose interpretation would be that on average the data points are about \$995 off of the regression line. We will use this in the next section to make inferences about our coefficients.

## **V. Inference**

We made our estimates above for the regression line based on our sample information. These are estimates of the (unknown) population parameters. In this section we want to make inferences about the population using our sample information.

t-test for the slope

Again, our estimate of  $\beta$  is  $b$ . We can show that under certain assumptions (to come in a bit) that  $b$  is an unbiased estimator for  $\beta$ . But as discussed above there will still be some sampling error associated with this estimate. So we can't conclude that  $\beta=b$  every time, only on average. Thus we need to take this sampling variability into account.

Suppose we have the following null and alternative hypothesis:

Ho:  $\beta_1=0$  (there is no relationship between X and Y)

H1:  $\beta_1 \neq 0$  (there is a relationship)

This can also be one tailed if you have some prior information to make it so.

Our test statistic will be:

$t = \frac{b_1 - \beta_1}{S_{b_1}}$ , where  $S_{b_1}$  is the standard error of the coefficient.

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}}$$

Where  $SSX = \sum(X_i - \bar{X})^2$

This follows a t-distribution with  $n-2$  degrees of freedom.

[NOTE: in general this test has  $n-k-1$  degrees of freedom, where  $k$  is the number of right-hand-side variables. In this case  $k=1$  so it is just  $n-2$ ]

So the Standard error of the coefficient is the standard error of the estimate divided by the squared deviation in X.

Again note the bottom part of the Excel output:

|           | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| Intercept | 997.4659            | 465.7353              | 2.141701      | 0.046147       | 18.99164         | 1975.94          |
| LOS       | 818.8394            | 141.0671              | 5.804607      | 1.69E-05       | 522.4681         | 1115.211         |

So our LOS coefficient is 818.8. Is this statistically different from zero?

Our test statistic is:  $t = \frac{818.8 - 0}{141.1} = 5.8$ . We can use the reported p-value: .0000169 to conclude that we would reject the null hypothesis and say that there is evidence that  $\beta_1$  is not zero. That is, LOS has a significant effect on charges.

The t-test can be used to test each individual coefficient for significance in a multiple regression framework. The logic is just the same.

One could also test other hypothesis: Suppose it used to be the case that each day in the hospital resulted in \$1000 charge, is there evidence that it has changed?

Ho:  $\beta_1=1000$

Ha:  $\beta_1 \neq 1000$

$t = (818.8 - 1000) / 141.06 = -1.28$  the pvalue associated with this is .215 – so there is a 21.5% chance we could get a coefficient of 818 or further away from 1000 if the null is true. Thus, we would fail to reject the null and conclude that there is no evidence that the slope is different from 1000.

### F-test for the Slope

An alternative is the F-test for the Slope. In a simple (one variable) regression framework this is equivalent to the t-test. But as we'll see next time it is asking a more general question about the joint significance of all the right-hand-side variables.

Here the F statistic for the null hypothesis that the slope is zero is:

$$F = MS_R / MS_E$$

Where  $MS_R = SS_R / k$

And  $MS_E = SS_E / n - k - 2$

And k is the number of right-hand-side variables in the model (one in our case here)

This follows the F distribution with k and n-k-1 degrees of freedom. Looking at the middle section of the Excel output:

| ANOVA      |           |           |           |          |                       |
|------------|-----------|-----------|-----------|----------|-----------------------|
|            | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
| Regression | 1         | 33390795  | 33390795  | 33.69347 | 1.69E-05              |
| Residual   | 18        | 17838305  | 991016.9  |          |                       |
| Total      | 19        | 51229100  |           |          |                       |

Looking at the fourth column labeled MS:

$$MSR = 33390795 / 1 = 33390795$$

$$MSE = 17838305 / 18 = 991016.9$$

$$\text{So } F = 33390795 / 991016.9 = 33.69$$

The “Significance F” is the p-value. So we'd reject  $H_0$  and conclude there is evidence the slope is not zero.

Notice that here  $t^2 = F$

We will see this test again in the next chapter and it will be a more general test about all the coefficient estimates.

We could also estimate a confidence interval for the slope:

$$b_1 \pm t_{n-2} S_{b_1}$$

Where  $t_{n-2}$  is the appropriate critical value for  $t$ . You can get excel to spit this out for you as well. Just click the confidence interval box and type in the level of confidence and it will include the upper and lower limits in the output.

For my example we are 95% confident that the population parameter  $\beta_1$  is between 522 and 1115.

## VI. Assumptions of Least Squares

In order for the least squares estimate to be an unbiased estimate for  $\beta$  we have to make the following assumptions:

- i. Normality of Errors – the error terms are distributed normally around the regression line.
- ii. Homoscedasticity – this just means that the variance of the error term is constant across all observations. The errors vary the same amount when  $X$  is low as when  $X$  is high. If this is not true then we say there is heteroskedasticity on the data. Suppose we are interested in the average height across cities. We have lots of observations on say New York but relatively few for San Antonio. So the average in the city will have a much tighter variance in NY, than SA. We would want to put more weight or confidence in the NY number than the SA number. Least squares assumes they are equally weighted.
- iii. Independence of Errors—this says that the errors around the regression line must be independent for each value of  $X$ . That is if the error is high for one observation, that does not affect the error term for the next observation. This is especially a big problem for time-series data – if the temperature is higher than average today, it is likely to be higher than average tomorrow as well. When the error terms are not independent we have what is called autocorrelation.

For our purposes we are just going to assume that these are all good assumptions. There are tests to see if they are reasonable. The book talks some about this in the section on residual analysis, but it gets much trickier. If these assumptions are not valid, then Least Squares will not be the best way to go and more sophisticated methods are required.