**Simple Linear Regression: What it all means**

Simple linear regression is a statistical modeling procedure that models the relationship between one Dependent Variable (DV), and one Independent Variable (IV), A model is a simplified representation that captures important characteristics but leaves out many details. The model of the relationship between and in linear regression is a **straight line**.

The linear relationship between and in the whole **population of interest** is represented in the **Regression Model**:

where

The betas are population parameters, and as such, we cannot observe them directly. Instead, we take a random sample from the population and use the data from the sample to **estimate** the betas. That is what the regression procedure does: it analyzes all the data in the sample and uses it to calculate the **Estimated Regression Equation**:

Regression output contains many different quantities, all of which are calculated from the data: that is, from the individual values of and in the observations in the data set.

The data set will look like this, although the columns do not have to be in this order:

|  |  |  |
| --- | --- | --- |
| **Observation**  **()** | **Dependent**  **Variable**  **()** | **Independent**  **Variable**  **()** |
| 1 |  |  |
| 2 |  |  |
| … | … | … |
|  |  |  |

The Observation column – which may be labeled by Case or Subject, or whatever is appropriate for the study – is a unique identifier for each observation in the dataset. Every observation in the dataset consists of measurements for each variable. For example, if this dataset was for a regression investigating income and years of education, each observation would be an individual person (), and for each person, income () and years of education () would be measured and recorded. From the values of and for each observation , we can calculate all of the values in the regression output.

Linear regression calculates the estimated regression equation by employing something called the **Least Squares Criterion**, which minimizes the squared vertical distance between each observed value of in the sample (each ) and the predicted from the estimated regression equation (each ) at every value of . This vertical distance between observed and predicted () is a very important quantity in regression, called the Residual. Mathematically, the Least Squares Criterion is stated:

So, the Least Squares Criterion minimizes the sum of the squared residuals. **Conceptually, the estimated regression equation that linear regression calculates is the line that is as close as possible to all the points in the data set at once: that is what it means to satisfy the Least Squares Criterion.**

For linear regression with more than one independent variable, matrix algebra or calculus is required to calculate the estimated regression equation. The equations simplify in the case of Simple Linear Regression, and so we could easily calculate the equation, and in fact all the regression output, by hand – if we only had the time.

For now, let’s concentrate on learning what the numbers in the regression output mean.

**Interpreting Linear Regression Output**

NOTE: in many of the interpretations, I refer to the DV as and the IVs as etc. When actually interpreting a regression, these variables have meaning and the meaning should be substituted in for these placeholder variables.

**In the *Regression Statistics* table (in order of importance/informativeness):**

1. The **Coefficient of Determination,** (called **R Square** in Excel) gives the proportion of the variability in the dependent variable that is explained by the independent variable or variables. varies from 0 to 1, and when interpreted, it is converted to a percentage.
   1. In Simple Linear Regression: measures the proportion of the variability in that is explained by .

Example: If then the **interpretation** is: *72.31% of the variability in is explained by*

* 1. In Multiple Regression: measures the proportion of the variability in that is explained by all of the independent variables, .

Example: If then the **interpretation** is: *49.79% of the variability in is explained by*

1. The **Standard Error of the Estimate, s,** often called the **Root Mean Square Error (RMSE)**,measures the accuracy of the predictions made by the Estimated Regression Equation. It is the average distance that the observed values in the sample fall from the regression line.
   1. **Interpretation:** the average error we would make using the estimated regression equation to predict . In other words: if we used the estimated regression equation to predict , we would be off by **s** on average.

Example: If then the **interpretation** is:  *is the average error we would make if we used the estimated regression equation to predict*

Another way to state the same thing is: *If we used the estimated regression equation to predict y, we would be off by 4 units on average.*

* 1. The Standard Error of the Estimate is in the same units as , which makes it very easy to understand and interpret.
  2. The lower the Standard Error of the Estimate, the more accurate the predictions, so the lower the better!
  3. Given two regression models that predict the same DV measured in the same units, **the Standard Error of the Estimate can be used to choose between models**: **whichever one has the lower Standard Error of the Estimate is the better model**, because the predictions made with it will be more accurate.

1. **Adjusted** is used as a model comparison statistic in multiple regression. When comparing two models that predict the same DV in the same units, the one with the higher **Adjusted**  is the better model.
   1. NOTE: it is **not** appropriate to use regular old to compare two models, because adding another IV to a regression model will increase , whether that IV explains anything or not. **Adjusted ,** on the other hand, will only increase if the additional IV adds explanatory power, and will actually decrease if it does not.
2. The **Correlation Coefficient** (called **Multiple R** in Excel) is denoted in simple linear regression and in multiple regression. It gives a descriptive measure of the strength of the linear association between and (in simple linear regression) and between and all the in multiple regression.
   1. The closer to 1, the stronger the association. Values close to 0 indicate that x and y are not linearly related.
   2. No rules are set in stone about what constitutes a “strong” or “weak” relationship. A common rule of thumb: < 0.25: weak linear association; between 0.25 and 0.75: moderate linear association; > 0.75: strong linear association

**The *ANOVA* table:**

* recall: the sample data is our best representation of the underlying population, so if our model is good at predicting the sample values, we can infer it will also be good at predicting population values
* The ANOVA table in the regression output compares two different methods of predicting the sample data in an effort to help us decide if our regression model is any good.
  + Premise: there are two alternative models you could use to predict . One model includes information about and one does not:
    - First, you could use the Estimated Regression Equation to predict (in other words, predict at each )
    - Second, you could use the next best option, which is to just use to predict (in other words, predict (the mean of y) at each )
      * This is also called the ‘Intercept only model’ because the equation of does not include any variables, thus being only an intercept, and representing a horizontal line at .
  + In the ANOVA table, you calculate the total errors you would make if you predicted the actual observed data using each of these models, square those errors, and sum them. In this way, we can compare the two possible models and decide whether our regression model is worth using.

1. The **Sum of Squares due to Error,** sometimes called the **Residual Sum of Squares,** is the total amount of prediction error we would make using the estimated regression equation to predict the observed sample data. It is the sum of the squared **residuals.**
   1. A **residual** is the **difference between the observed and predicted at a given** In other words:
   2. The , and has degrees of freedom
   3. Note: the sum of the squared residuals is the quantity that is minimized by the Least Squares Criterion!
2. The **Total Sum of Squares,** is the total amount of prediction error we would make if we used (the mean of ) to predict the observed sample data.
   1. The , and has degrees of freedom
3. The **Sum of Squares due to Regression,** is how much we will reduce prediction error by using the Estimated Regression Equation instead of the mean to predict the observed sample data.
   1. The , and has degrees of freedom
4. The **Mean Square Error,**  is the estimate of the variance of . Remember It is the random error term in the Regression Model:
   1. The

**The *Coefficients* table:**

1. The **, ,** quantify the magnitude and direction of the relationship between each IV and the DV. They are point estimates, respectively, of
   1. In Simple Linear Regression, the **interpretation** **of**  is:
      1. **If**

*For every one unit increase in , is predicted to increase by on average.*

* + 1. **If**

*For every one unit increase in , is predicted to decrease by on average.*

* 1. In Multiple Regression, each slope is interpreted. The interpretation for is given here, and the interpretation of the other slopes follows the same pattern. The **interpretation** **of**  is:
     1. **If**

*For every one unit increase in , is predicted to increase by on average, holding all other independent variables constant.*

* + 1. **If**

*For every one unit increase in , is predicted to decrease by on average, holding all other independent variables constant.*

1. The is the value of when This is not always interpretable in the context of a given regression. Interpretability depends on whether the value has substantive meaning.
2. The **, ,** and the are based on the significance level and are interpreted similarly to the other confidence intervals we have encountered:
   1. For a given **slope coefficient, say**  the **interpretation** is:

*We can be confident that the true value of the population slope coefficient for is between [lower bound] and [upper bound].*

* 1. For the **intercept, ,** the **interpretation** is:

*We can be confident that the true value of the population intercept is between [lower bound] and [upper bound].*