**Simple Linear Regression: Introduction**

Simple linear regression is a statistical procedure that models the relationship between two variables. The two variables are the dependent variable (DV) called y, and the independent variable (IV), called x.

A statistical model is a simplified picture of the relationship between variables that emphasizes key features and regular patterns. In simple linear regression, we model the relationship between two variables as a straight line.

Why do we build statistical models? It is to be able to explain what is going on with the dependent variable, and to make predictions about the future based on data in the past. For example, we can model the relationship between a person’s level of education (x) and salary (y) to predict how much money they will make – in fact we could use regression on this problem and predict how much more people will make for each additional year of education they have. We can model the durability of a painted coating (y) based on the thickness of the coating (x). How thick does the coating have to be to hit our durability target? Statistical modeling can help with that. Regression in particular is for modeling linear relationships. Other statistical models exist to model other kinds of relationships.

The **Regression Model** is a straight line defined by the following equation:

where

The regression model describes the relationship between x and y in the population – so and are **population parameters.** Since they are population parameters, they are not directly measured. Rather, we use sample data to estimate these population values, and draw conclusions about them using statistical inference, like confidence intervals and hypothesis tests.

In regression, a random sample is taken from a population, and the data in the sample is used to calculate a linear equation, called the **Estimated Regression Equation (ERE)**:

The Estimated Regression Equation gives us a way to calculate the **expected value of y at a given value of x.** The expected value of y at a given x is **the long-run mean of y: the mean value of y at a given x if many samples were taken from the same population.** To state this more formally, is a point estimate for E(y), the expected value of y at a given x. It is often called the **predicted value of y**.

**A Refresher on the Equation of a Line**

The equation of a line is:

The slope of the line:

When graphing a line, the vertical axis is the y axis, and the horizontal axis is the x axis. Hence, the slope is often described as because it describes how much the line rises up the y axis as it runs along the x axis.

***Example 1: Refresher 1.***Graph the following equation. What is the slope? Interpret the slope by stating in words how much y changes (and in what direction) for a one unit change in x.



***Example 2: Refresher 2.***Graph the following equation. What is the slope? Interpret the slope by stating in words how much y changes (and in what direction) for a one unit change in x.



***Example 3: Refresher 3.***Graph the following equation. What is the slope? Interpret the slope in this equation by stating in words how much y changes (and in what direction) for a one unit change in x.



**And now back to regression…**

Now, data do not typically fall exactly on a nice straight line when graphed. Real-world data looks more like a cloud of points. Linear regression can only be used in cases where there is a plausible straight-line relationship between x and y. If the relationship between x and y is not linear, then you have to use some other technique to model it. Scatterplots which graph each data point are an important tool to recognize the relationships in your data.

Here is how data that would typically be modeled with regression might look. Can you visualize the line that could model this data? Would it have a positive or a negative slope?

**Figure 1.**

Here is another example of data that could be modeled with linear regression. Can you visualize the line that could model this data? Would it have a positive or a negative slope?

**Figure 2.**

What about this one?

**Figure 3.**

Now let’s look at some examples of data you would **NOT** model with linear regression. Do you see the nonlinear pattern?

**Figure 4.**

If you modeled the data in **Figure 4** with linear regression, the regression procedure would give you a straight line with a positive slope. But that straight line would misrepresent the actual relationship, which would cause your predictions to be way off! You as the researcher are responsible for avoiding this mistake – no computer is going to tell you not to make it. (NOTE: there are ways to transform data like this so that it can be modeled – if you continue studying statistics, you will find out about those). For now, protect yourself against such errors by graphing scatterplots of your data.

**Figure 5.**

This is a dangerous pattern, because regression would model this as a horizontal line, or something close to that. And a horizontal line (which means the slope is zero) indicates there is **no relationship** between x and y. But there clearly is a relationship here – it is just not a linear relationship.

Let’s look at some examples:

Remember, the **regression model** is the relationship between x and y in the population:

The **estimated regression equation (ERE)**is the relationship between x and y in the sample:

Remember, is the **average** for a given , and is called the **expected** or **predicted** value of .

***Example 4.*** Regression could be used to investigate the relationship between weekly sales and weekly advertising.

To use regression to model this relationship, first designate the DV and the IV.

Then take a random sample of weeks and measure sales and advertising in the prior week for each. The graph of the data looks like this. Do you expect a positive or a negative slope on the regression line?:

The sample data is used to calculate the following values:

a) Write down the Estimated Regression Equation (ERE).

b) If $235 is spent on advertising, what would be the predicted sales? Interpret this value.

c) What is the interpretation of the slope,

d) Interpret the 95% confidence interval for the population slope,

***Example 5.*** The owners of a restaurant think that as the temperature of a certain drink increases, customer satisfaction will decrease. A random sample of customers who order this drink is taken, and the temperature of the drink along with the customer satisfaction rating (0 to 100) is measured.

To use regression to model this relationship, first we designate the DV and the IV.

Here is the scatterplot of the data collected:

The sample data is used to calculate the following values:

a) Write down the Estimated Regression Equation (ERE)

b) If the drink is served at , what would be the predicted customer satisfaction score? Interpret this value.

c) What is the interpretation of the slope,

d) Interpret the 95% confidence interval for the population slope,