Chapter 11

Simple Linear Regression Tutorial

Linear regression is a very simple method but has proven to be very useful for a large number of situations. In this chapter you will discover exactly how linear regression works step-by-step. After reading this chapter you will know:

- How to calculate a simple linear regression step-by-step.
- How to make predictions on new data using your model.
- A shortcut that greatly simplifies the calculation.

Let's get started.

11.1 Tutorial Data Set

The data set we are using is completely made up. Below is the raw data.

```
x   y
1   1
2   3
4   3
3   2
5   5
```

Listing 11.1: Tutorial Data Set.

The attribute x is the input variable and y is the output variable that we are trying to predict. If we got more data, we would only have x values and we would be interested in predicting y values. Below is a simple scatter plot of x versus y.

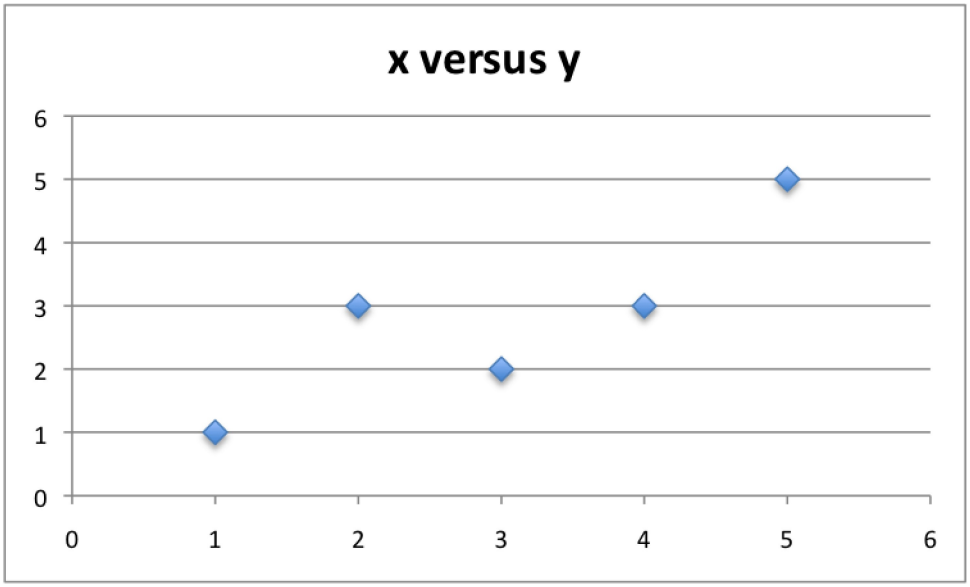

Figure 11.1: Simple Linear Regression Dataset.

We can see the relationship between x and y looks kind-of linear. As in, we could probably draw a line somewhere diagonally from the bottom left of the plot to the top right to generally describe the relationship between the data. This is a good indication that using linear regression might be appropriate for this little dataset.

11.2 Simple Linear Regression

When we have a single input attribute (x) and we want to use linear regression, this is called simple linear regression. If we had multiple input attributes (e.g. $X1$, $X2$, $X3$, etc.) This would be called multiple linear regression. The procedure for linear regression is different and simpler than that for multiple linear regression, so it is a good place to start. In this section we are going to create a simple linear regression model from our training data, then make predictions for our training data to get an idea of how well the model learned the relationship in the data. With simple linear regression we want to model our data as follows:

$$y = B0 + B1 \times x \tag{11.1}$$

This is a line where y is the output variable we want to predict, x is the input variable we know and $B0$ and $B1$ are coefficients that we need to estimate that move the line around. Technically, $B0$ is called the intercept because it determines where the line intercepts the y-axis. In machine learning we can call this the bias, because it is added to offset all predictions that we make. The $B1$ term is called the slope because it defines the slope of the line or how x translates into a y value before we add our bias.

The goal is to find the best estimates for the coefficients to minimize the errors in predicting y from x. Simple regression is great, because rather than having to search for values by trial and error or calculate them analytically using more advanced linear algebra, we can estimate

them directly from our data. We can start off by estimating the value for $B1$ as:

$$B1 = \frac{\sum_{i=1}^{n}(x_i - mean(x)) \times (y_i - mean(y))}{\sum_{i=1}^{n}(x_i - mean(x))^2} \tag{11.2}$$

Where $mean()$ is the average value for the variable in our dataset. The x_i and y_i refer to the fact that we need to repeat these calculations across all values in our dataset and i refers to the i'th value of x or y. We can calculate $B0$ using $B1$ and some statistics from our dataset, as follows:

$$B0 = mean(y) - B1 \times mean(x) \tag{11.3}$$

Not that bad right? We can calculate these right in our spreadsheet.

11.2.1 Estimating The Slope (B1)

Let's start with the top part of the equation, the numerator. First we need to calculate the mean value of x and y. The mean is calculated as:

$$\frac{1}{n} \times \sum_{i=1}^{n} x_i \tag{11.4}$$

Where n is the number of values (5 in this case). You can use the `AVERAGE()` function in your spreadsheet. Let's calculate the mean value of our x and y variables:

$$\begin{aligned} mean(x) &= 3 \\ mean(y) &= 2.8 \end{aligned} \tag{11.5}$$

Now we need to calculate the error of each variable from the mean. Let's do this with x first:

```
x   mean(x)    x - mean(x)
1   3          -2
2              -1
4              1
3              0
5              2
```

Listing 11.2: Residual of each x value from the mean.

Now let's do that for the y variable.

```
y   mean(y)    y - mean(y)
1   2.8        -1.8
3              0.2
3              0.2
2              -0.8
5              2.2
```

Listing 11.3: Residual of each y value from the mean.

We now have the parts for calculating the numerator. All we need to do is multiple the error for each x with the error for each y and calculate the sum of these multiplications.

```
x - mean(x)    y - mean(y)    Multiplication
-2             -1.8           3.6
-1             0.2            -0.2
1              0.2            0.2
0              -0.8           0
2              2.2            4.4
```

Listing 11.4: Multiplication of the x and y residuals from their means.

Summing the final column we have calculated our numerator as 8. Now we need to calculate the bottom part of the equation for calculating $B1$, or the denominator. This is calculated as the sum of the squared differences of each x value from the mean. We have already calculated the difference of each x value from the mean, all we need to do is square each value and calculate the sum.

```
x - mean(x)    squared
-2             4
-1             1
1              1
0              0
2              4
```

Listing 11.5: Squared residual of each x value from the mean.

Calculating the sum of these squared values gives us a denominator of 10. Now we can calculate the value of our slope.

$$\begin{aligned} B1 &= \frac{8}{10} \\ B1 &= 0.8 \end{aligned} \tag{11.6}$$

11.2.2 Estimating The Intercept (B0)

This is much easier as we already know the values of all of the terms involved.

$$\begin{aligned} B0 &= mean(y) - B1 \times mean(x) \\ B0 &= 2.8 - 0.8 \times 3 \\ B0 &= 0.4 \end{aligned} \tag{11.7}$$

11.3 Making Predictions

We now have the coefficients for our simple linear regression equation.

$$\begin{aligned} y &= B0 + B1 \times x \\ y &= 0.4 + 0.8 \times x \end{aligned} \tag{11.8}$$

Let's try out the model by making predictions for our training data.

```
x   Predicted Y
1   1.2
2   2
4   3.6
```

```
3   2.8
5   4.4
```

Listing 11.6: Predicted y value for each x input value.

We can plot these predictions as a line with our data. This gives us a visual idea of how well the line models our data.

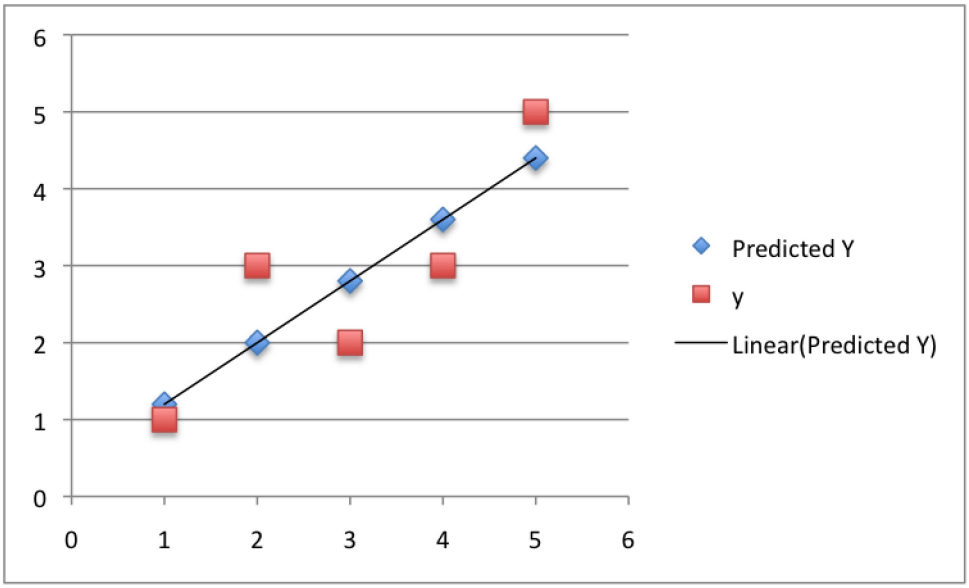

Figure 11.2: Simple Linear Regression Predictions.

11.4 Estimating Error

We can calculate an error score for our predictions called the Root Mean Squared Error or RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - y_i)^2}{n}} \tag{11.9}$$

Where you can use `SQRT()` function in your spreadsheet to calculate the square root, p is the predicted value and y is the actual value, i is the index for a specific instance, because we must calculate the error across all predicted values. First we must calculate the difference between each model prediction and the actual y values.

```
Predicted  y   Predicted - y
1.2        1   0.2
2          3   -1
3.6        3   0.6
2.8        2   0.8
4.4        5   -0.6
```

Listing 11.7: Error for predicted values.

We can easily calculate the square of each of these error values ($error \times error$ or $error^2$).

```
Predicted - y     squared error
0.2               0.04
-1                1
0.6               0.36
0.8               0.64
-0.6              0.36
```

Listing 11.8: Squared error for predicted values.

The sum of these errors is 2.4 units, dividing by 5 and taking the square root gives us:

$$RMSE = 0.692820323 \tag{11.10}$$

Or, each prediction is on average wrong by about 0.692 units.

11.5 Shortcut

Before we wrap up I want to show you a quick shortcut for calculating the coefficients. Simple linear regression is the simplest form of regression and the most studied. There is a shortcut that you can use to quickly estimate the values for $B0$ and $B1$. Really it is a shortcut for calculating $B1$. The calculation of $B1$ can be re-written as:

$$B1 = corr(x, y) \times \frac{stdev(y)}{stdev(x)} \tag{11.11}$$

Where $corr(x, y)$ is the correlation between x and y an $stdev()$ is the calculation of the standard deviation for a variable. Correlation (also known as Pearson's correlation coefficient) is a measure of how related two variables are in the range of -1 to 1. A value of 1 indicates that the two variables are perfectly positively correlated, they both move in the same direction and a value of -1 indicates that they are perfectly negatively correlated, when one moves the other moves in the other direction.

Standard deviation is a measure of how much on average the data is spread out from the mean. You can use the function `PEARSON()` in your spreadsheet to calculate the correlation of x and y as 0.852 (highly correlated) and the function `STDEV()` to calculate the standard deviation of x as 1.5811 and y as 1.4832. Plugging these values in we have:

$$\begin{aligned} B1 &= 0.852802865 \times \frac{1.483239697}{1.58113883} \\ B1 &= 0.8 \end{aligned} \tag{11.12}$$

11.6 Summary

In this chapter you discovered how to implement simple linear regression step-by-step in a spreadsheet. You learned:

- How to estimate the coefficients for a simple linear regression model from your training data.
- How to make predictions using your learned model.