REGRESSION ANALYSIS

As the “Hello World” of machine learning algorithms, regression analysis is a simple supervised learning technique used to find the best trendline to describe a dataset.

The first regression analysis technique that we will examine is linear regression, which uses a straight line to describe a dataset. To unpack this simple technique, let’s return to the earlier dataset charting Bitcoin values to the US Dollar.

Date	Bitcoin Price	No. of Days Transpired
19-05-2015	234.31	1
14-01-2016	431.76	240
09-07-2016	652.14	417
15-01-2017	817.26	607
24-05-2017	2358.96	736

Imagine you’re back in high school and it's the year 2015 (which is probably much more recent than your actual year of graduation!). During your senior year, a news headline piques your interest in Bitcoin. With your natural tendency to chase the next shiny object, you tell your family about your cryptocurrency aspirations. But before you have a chance to bid for your first Bitcoin on Coinbase, your father intervenes and insists that you try paper trading before you go risking your life savings. “Paper trading” is using simulated means to buy and sell an investment without involving actual money.

So over the next twenty-four months, you track the value of Bitcoin and write down its value at regular intervals. You also keep a tally of how many days have passed since you first started paper trading. You never anticipated to still be paper trading after two years, but unfortunately, you never got a chance to enter the cryptocurrency market. As suggested by your father, you waited for the value of Bitcoin to drop to a level you could afford. But instead, the value of Bitcoin exploded in the opposite direction. Nonetheless, you haven’t lost hope of one day owning Bitcoin. To assist your decision on

whether you continue to wait for the value to drop or to find an alternative investment class, you turn your attention to statistical analysis. You first reach into your toolbox for a scatterplot. With the blank scatterplot in your hands, you proceed to plug in your x and y coordinates from your dataset and plot Bitcoin values from 2015 to 2017. However, rather than use all three columns from the table, you select the second (Bitcoin price) and third (No. of Days Transpired) columns to build your model and populate the scatterplot (shown in Figure 1). As we know, numerical values (found in the second and third columns) are easy to plug into a scatterplot and require no special conversion or one-hot encoding. What's more, the first and third columns contain the same variable of "time" and the third column alone is sufficient.

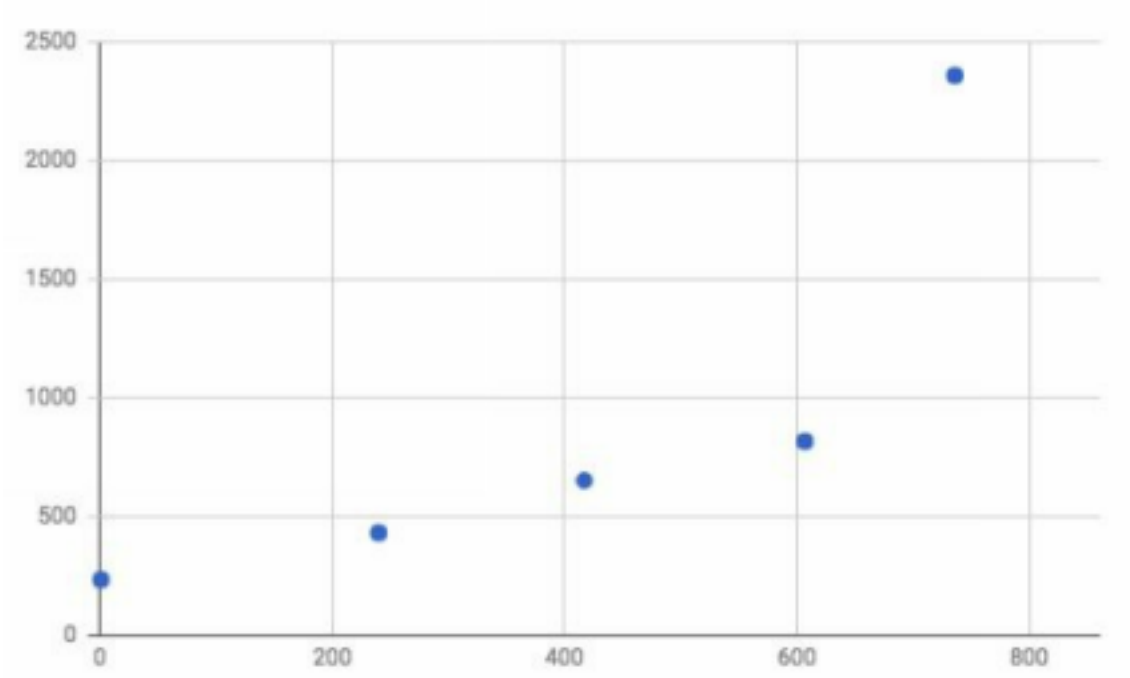

Figure 1: Bitcoin values from 2015-2017 plotted on a scatterplot

As your goal is to estimate what Bitcoin will be valued at in the future, the y axis plots the dependent variable, which is "Bitcoin Price." The independent variable (X), in this case, is time. The "No. of Days Transpired" is thereby plotted on the x-axis.

After plotting the x and y values on the scatterplot, you can immediately see a trend in the form of a curve ascending from left to right with a steep increase between day 607 and day 736. Based on the upward trajectory of the curve, it

might be time to quit hoping for a drop in value.

However, an idea suddenly pops up into your head. What if instead of waiting for the value of Bitcoin to fall to a level that you can afford, you instead borrow from a friend and purchase Bitcoin now at day 736? Then, when the value of Bitcoin rises further, you can pay back your friend and continue to earn asset appreciation on the Bitcoin you fully own.

In order to assess whether it's worth borrowing from your friend, you will need to first estimate how much you can earn in potential profit. Then you need to figure out whether the return on investment will be adequate to pay back your friend in the short-term.

It's now time to reach into the third compartment of the toolbox for an algorithm. One of the simplest algorithms in machine learning is regression analysis, which is used to determine the strength of a relationship between variables. Regression analysis comes in many forms, including linear, non linear, logistic, and multilinear, but let's take a look first at linear regression.

Linear regression comprises a straight line that splits your data points on a scatterplot. The goal of linear regression is to split your data in a way that minimizes the distance between the regression line and all data points on the scatterplot. This means that if you were to draw a vertical line from the regression line to each data point on the graph, the aggregate distance of each point would equate to the smallest possible distance to the regression line.

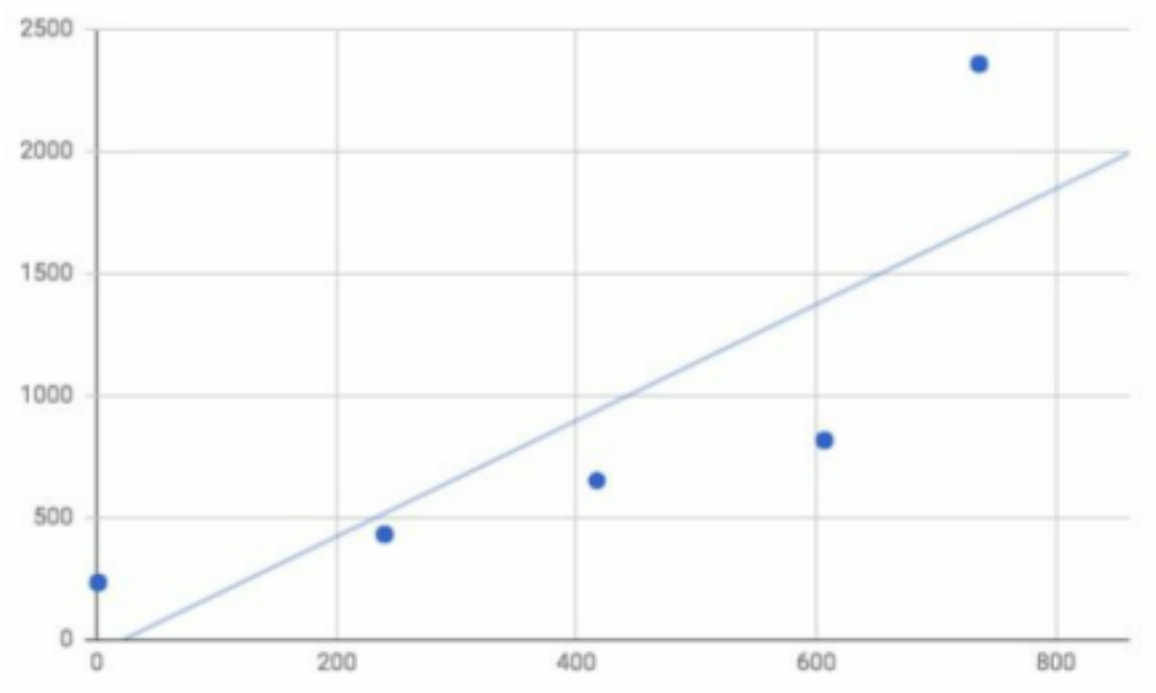

Figure 2: Linear regression line

The regression line is plotted on the scatterplot in Figure 2. The technical term for the regression line is the *hyperplane*, and you will see this term used throughout your study of machine learning. A hyperplane is practically a trendline—and this is precisely how Google Sheets titles linear regression in its scatterplot customization menu.

Another important feature of regression is *slope*, which can be conveniently calculated by referencing the hyperplane. As one variable increases, the other variable will increase at the average value denoted by the hyperplane. The slope is therefore very useful in formulating predictions. For example, if you wish to estimate the value of Bitcoin at 800 days, you can enter 800 as your x coordinate and reference the slope by finding the corresponding y value represented on the hyperplane. In this case, the y value is USD $1,850.

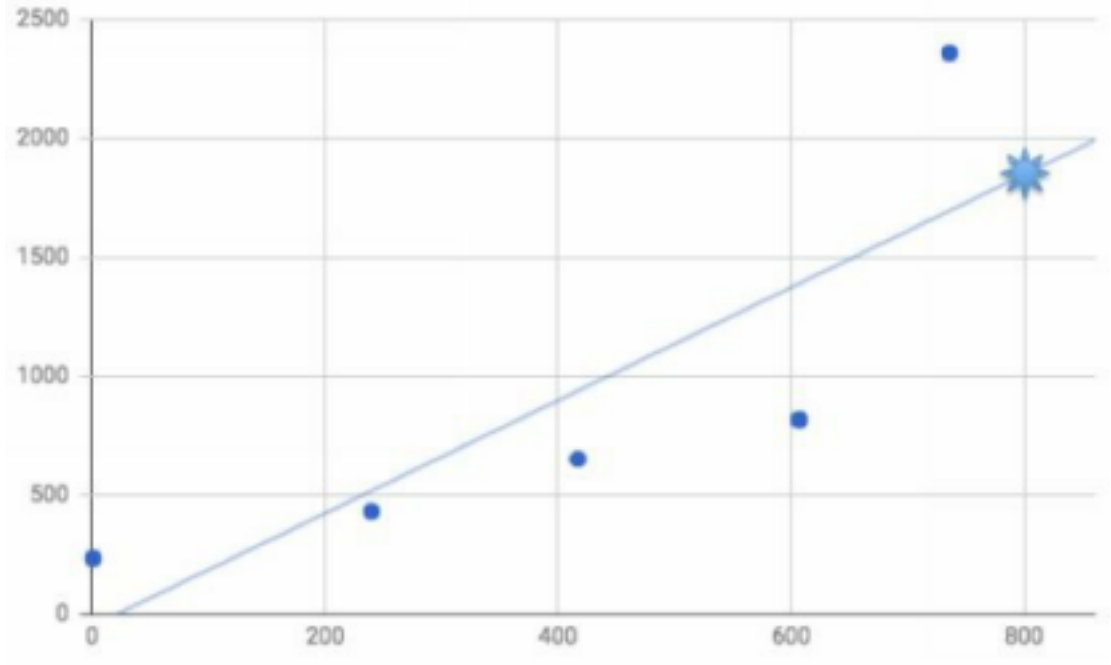

Figure 3: The value of Bitcoin at day 800

As shown in Figure 3, the hyperplane reveals that you actually stand to lose money on your investment at day 800 (after buying on day 736)! Based on the slope of the hyperplane, Bitcoin is expected to depreciate in value between day 736 and day 800—despite no precedent in your dataset for Bitcoin ever dropping in value.

While it's needless to say that linear regression isn't a fail-proof method to picking investment trends, the trendline does offer a basic reference point to predict the future. If we were to use the trendline as a reference point earlier in time, say at day 240, then the prediction posted would have been more accurate. At day 240 there is a low degree of deviation from the hyperplane, while at day 736 there is a high degree of deviation. Deviation refers to the distance between the hyperplane and the data point.

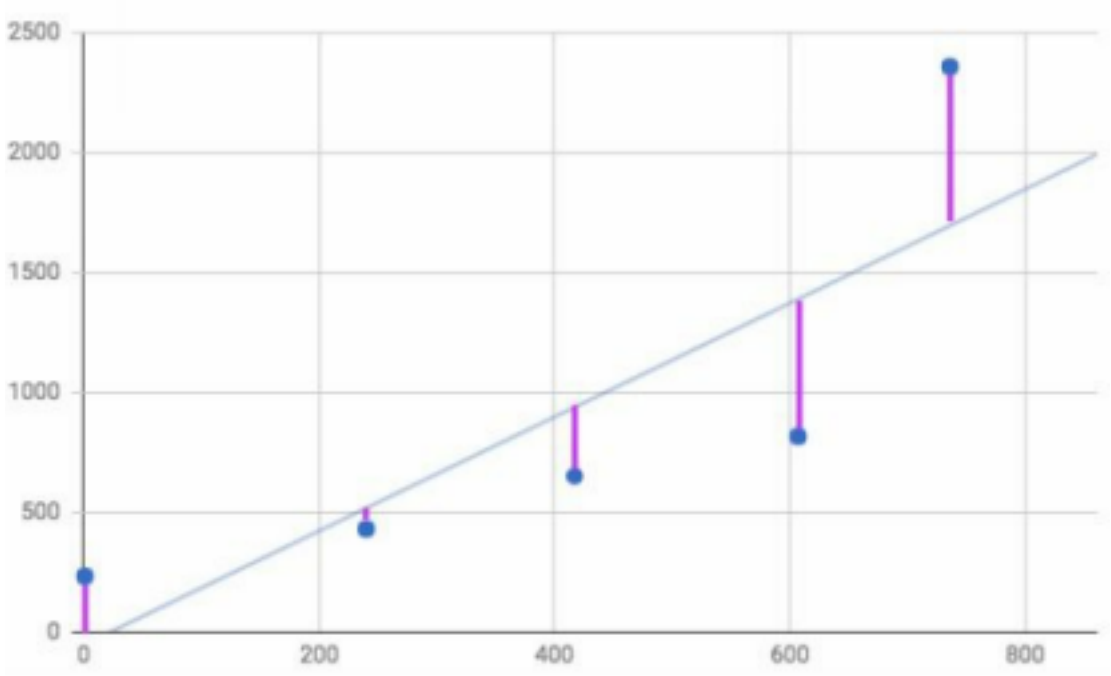

Figure 4: The distance of the data points to the hyperplane

In general, the closer the data points are to the regression line, the more accurate the final prediction. If there is a high degree of deviation between the data points and the regression line, the slope will provide less accurate predictions. Basing your predictions on the data point at day 736, where there is high deviation, results in poor accuracy. In fact, the data point at day 736 constitutes an outlier because it does not follow the same general trend as the previous four data points. What's more, as an outlier it exaggerates the trajectory of the hyperplane based on its high y-axis value. Unless future data points scale in proportion to the y-axis values of the outlier data point, the model's predictive accuracy will suffer.

Calculation Example

Although your programming language will take care of this automatically, it's useful to understand how linear regression is actually calculated. We will use the following dataset and formula to perform linear regression.