**Linear Regression**

**Name:** Saqif Abrar

**Table of Contents:**

Introduction

History

Linear Regression Using the Least Squares Method & Derivations

Advantages and Disadvantages

Python Code & Example

References

**Introduction**

Linear Regression is a very power mathematical tool used in many areas of application. It is a core topic in many data-driven applications such as mathematics, statistics, biology, geology, business analytics and machine learning/data science. It is a very important topic found almost everywhere and it is very intuitive! There are many kinds of regression methods such as logistic regression or polynomial regression, but we will be focusing on linear regression.

**History**

Many have heard the famous “prince of mathematicians”, Carl Friedrich Gauss. Maybe the Gaussian distribution was discussed in a statistics class or Gaussian elimination from a linear algebra unit. Well, the discovery of linear regression is often associated with Gauss as well!

The 19th century was an age of exploration, which fixed many mathematicians to observe the celestial body as discerning the heavens aided navigation throughout the world. Gauss was one of the few key mathematicians who studied the planetary orbits. During these astronomical observations, he applied an earlier form of linear regression, known as the least squares method. While he performed these mathematical calculations on his own to determine the planets’ and comets’ orbital paths, another mathematician, known as Adriene Marie Legendre, had published this method earlier. While many say that Gauss had taken claim of this method for himself, he did discuss this on a higher technical level in his publications. This is one of the reasons why Gauss is often associated as the founder of the least squares method and the father of linear regression.

**Linear Regression Using the Least Squares Method & Derivations**

Suppose we have a few observations in a 2-dimensional plane. We would like to find a linear function, f(x), that fits through the different observations. Let’s say we have an observation (x, y). The distance between our observation and the coordinate at the same x-value that lies on the line, (x, f(x)), would be our error (refer to graph X.X). The objective of the least-square method is to determine a line that would result in the lowest total errors, also called residuals, from all the actual observations. Essentially, we want to minimize the distances between our observations and the line.

Chart, scatter chart

Description automatically generated

Plot 1.1 – The distance between the plot to the line in the same x-values are called the errors or residuals displayed as red arrows in this plot. Linear regressions minimizes the total residuals.

The objective of linear regression is to regression line which is simply a linear function that tries to estimate your dependant variable, . The idea of the regression line is to minimize the number of residuals/errors between your observations (all points in the graph) and the line. We must find the magnitude of the slope and the y intercept of our line of best fit which we will call and respectively. This is simply the equation of a line in the end.

First, let’s calculate . We derive it from the mean of all the x-coordinates, , and y-coordinates, . If we subtract the respective means from and of the coordinate, we can minimize the errors/residuals in each axis. In the least squares method, we find the product of the differences we found for the x-coordinates and the y-coordinates and we find the sum of all the products from coordinates. Here is the representation of the least square formula so far where n represents the total number of observations/coordinates.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Plot 1.2 & 1.3 – A representation of the distances between the observations and the mean of the x values and y values from left to right. The residuals are found from the differences.

Now we complete the squares method and solve for by taking our residuals from the x-axis and squaring them. Hence, the least “squares” method. We sum up all our squares and divide it with what we have so far to complete the calculation. This resembles the idea of a “rise over run” in a slope.

Now that calculated our , we must find the intercept of the line, . A universal truth about a regression line is that the line must intersect with the mean of all x-coordinates and y-coordinates in the plane. With this knowledge, we can plug in () and into our linear function and solve for .

Now we have found the equation for our regression line for the data so we can graph it (found in the blue line of Plot 1.1)!

**Advantages and Disadvantages**

One of the reasons why linear regression was so widely accepted in the mathematical community because it was easy to understand the advantages of its application and usage widely. From the words “linear”, we can understand that this method would require a relatively “linear” data set. This means if the independent and dependant variables correlate in a linear relationship, it would potentially lead to easier and more accurate estimates even beyond the actual observations.

However, when there are outliers or scrambled data, it results in drastic changes of slope at various magnitudes which can result in a relatively inaccurate regression line. Additionally, you may find that your observations may not have a linear relationship which invalidates the usage of this method. Data must be normally distributed and independent from one another. Distributions are further discussed in statistics. There are different regression models that may better suite your observations such as polynomial regression or logistic regression.

**Python Code & Example**

Linear regression may be very intuitive and a relatively easy method, but the difficulty comes in when programming this. Specifically, managing the data. The big processes in computer science and machine learning are input, processing and output. Input tends to be a very lengthy process that can get complicated. This is because there are many forms of data that is organized in different ways. It is critical to understand how the data will be organized in your program. After you organize the input, it may be very easy to output. The process can vary in difficulty however depending on the language, dimension of your data and other many other factors. Here I have designed an example of linear regression in action with an organized data set from the UCI Machine Learning Repository. You can see the code and data set in this this private GitHub repository at: [www.github.com](http://www.github.com). There will also be MATLAB code and graph for the plot from above. This can be opened in the Jupyter notebook online platform at ease.

Text

Description automatically generated

Graphical user interface, chart

Description automatically generated with medium confidence

**References**

<https://priceonomics.com/the-discovery-of-statistical-regression/>

<https://www.youtube.com/watch?v=zPG4NjIkCjc>

<https://www.youtube.com/watch?v=JvS2triCgOY>

<https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data#introduction-to-trend-lines>

<https://www.mathworks.com/help/matlab/creating_plots/using-high-level-plotting-functions.html>

<https://medium.com/future-vision/plotting-equations-in-python-d0edd9f088c8>

<https://www.youtube.com/watch?v=1BYu65vLKdA>