What is Regression Analysis

Edris Safari

Bellevue University, Nebraska U.S.A.

Abstract

The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean) (Wikipedia). In the 20th century, “regression” is defined in very much the same way, but with statistical tools and methods. It has evolved into becoming a process whereby we can make predictions and choices with high confidence. We will examine the statistical concept of regression and introduce logistic regression. We will examine the statistical concept of regression and introduce logistic regression.

Key Words: Regression, Logistic Regression

What is Regression Analysis

Introduction

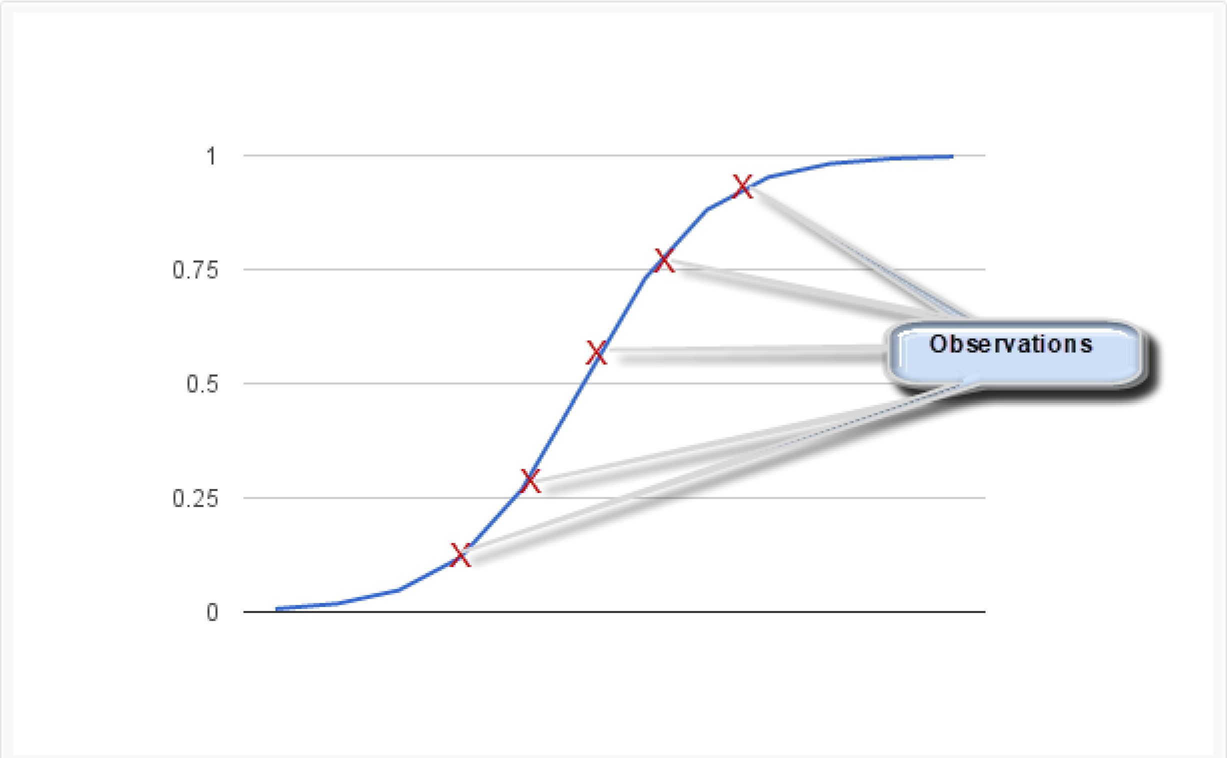
The basic concept of regression is to estimate the relationship or correlation among variables. Much in the same was as Francis Galton's description, variables that relate to each other, tend to contribute to one another in positive or negative way. For example, tall fathers tend to have tall sons, or tall mothers tend to have tall daughters. In these cases, the height of the ancestor can be used to reliably predict the height of the descendant. The variables involved in this use case are the parent's height, and the height of their descendants. And the data is based on a sample of a population which consists of a parent's height and their descendant's height. Since we want to estimate the descendant's height, we will call it the dependent variable. We will call the parent's height the independent variable or the predictor since it will be used to predict the descendant's height. The relationship between these two variables can be shown with the classic linear equation:

$$Y = mX + b$$

Where X is the independent variable and Y is the dependent variable. For example, the values of x could be age, and those of y's salary. We could use this equation to estimate the salary of a person based on their age. Both age and salary are what are called continuous variables where the numerical distance between two values is infinite. This model fails when the the dependent variable Y can only have a binary value such as yes/no or stay/leave, like/dislike, etc. Logistic regression addresses this limitation. Just as Linear Regaression, logistic regression is modeled by a function. Below function is an example:

$$\mathrm{F(x)} = \mathrm{Y} = 1/(1 + \mathrm{e}^{-\mathrm{X}})$$

Where ‘e’ is the exponent function. Other logistic regression equations basically follow the equation above which is called the “Sigmoid” function. The value of Y in this function can only vary from 0 to 1 as shown below:



Logistic Function

As shown, we can’t reliably predict the value of Y to be 0 or 1, but rather somewhere in between or better described as a probability of how close an observation is to 1. This function is represented as below:

$$P(X) = P(Y = 1|X)$$

The function reads as “the predicted value of X is the probability of Y being 100% given X. This equation is the basis of the Naïve Bayes method. It applies the Bayes theorem with the “Naïve” assumption of conditional independence between the pairs of variables. This sort of calculation returns values between 0 and 1 but not explicit 0 or 1 as we require in logistic regression. Sometimes we can only go by probability and sometimes we need explicit 0 or 1 answer, and sometimes this probability is mapped to more than two classes such as “Disagree”, “Strong Disagree”, ‘Agree”, “Strongly Agree”- Logistic Regression analysis and methods available to us can be used.

Conclusion

Logistic regression and the modeling and computations that go into it are all used in machine learning in mostly applications where classification of outcome is of interest, It’s usage in online retail is quite evident. For example, amazon uses customer comments to create customer rating. The spam filters are another example. The email messages are sent to machine learned algorithms to determine if the email is a spam and if so, are redirected to junk mail.

References

1. https://en.wikipedia.org/wiki/Regression_analysis
2. Regression I: What is regression? | SSE, SSR, SST | R-squared | Errors (e vs. e) - https://www.youtube.com/watch?v=aq8VU5KLmkY
3. Logistic Regression for Machine Learning - https://machinelearningmastery.com/logistic-regression-for-machine-learning/
4. Introduction to Machine Learning Algorithms: Logistic Regression - https://hackernoon.com/introduction-to-machine-learning-algorithms-logistic-regression-cbdd82d81a36
5. Logistic Regression - https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html