Linear Regression

Yahia Mesharaf
Electronic Engineering
Hamm-Lippstadt University of Applied Sciences
Summer Semester 2022
yahia.mesharaf@stud.hshl.de

***Abstract*—Linear Regression is a linear approach and one of the well-known algorithms used in Deep Learning as well as in Statistics. It focuses on having linearity with respect to the certain data and predicting based on it what will be the output of the following data entries. In this paper, Linear Regression will be discussed with main focus on its involvement in Deep Learning and its implementation as code in Python using the scikit library.**

I. Introduction

Linear Regression is one of the widely known and used algorithms in statistics and deep (machine) learning. It is mainly built for the purpose of handling datasets with linear relationship, so how does it work?
Let's start our analysis of regression models by defining the context we're working with. A regression is a model that associates an input vector, , with one or more continuous dependent variables (for simplicity, we're going to refer to single outputs), . In a general scenario, there's no explicit dependence on time, even if regression models are often employed to model time series. The main difference is that, in the latter, the order of the data points cannot be changed, because there are often inter-dependencies. On the other hand, a generic regression can be used to model time-independent phenomena, and, in the context of GLMs, we're initially assuming that we work with stateless associations where the output value depends only on the input vector. In such cases, it's also possible to shuffle the dataset without changing the final result (of course, this is not true if the output at time t depends, for example, on yt-1, which is a function of , and so on).

Imagine having a dataset, , containing N m-dimensional observations drawn from the same data generating process, pdata. Each observation is associated with the corresponding continuous label contained in . A GLM models the relationship between y and as:

The values are called regressors, and we say that y has been regressed on the set of variables. The noise term models the intrinsic uncertainty of a specific phenomenon and it's a fundamental element that cannot be discarded unless the relationship is purely linear (in other words, all the points lie on the same hyperplane). However, there are two possible scenarios associated with the noise term, , which we always considered as conditioned to X for example, while we generally don't know the value of . This means that we can never estimate the moments of the noise directly, but always through the conditioning on an input sample. However more about that will be discussed later.

II. Main Body

A. Definition

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation shown below.

B. Application

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

If the goal is prediction, forecasting, or error reduction,[clarification needed] linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response. Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the ”lack of fit” in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms ”least squares” and ”linear model” are closely linked, they are not synonymous.

C. Implementation

The implementation of Linear Regression in Machine Linear is done in Python using the scikit library. It basically follows the mathematical concept mentioned earlier.

What happens first is that we must import a set of Data in our code to train our algorithm on and test how good our model is. The main mathematical formula the model follows is ”y=mx+c” where Y is the dependent variable (Output), x is the independent variable (input), m is the slope of the straight line representing the relation between x and y which could be either positive in case of directly proportional relation or negative in case of inversely proportional relation. And c is the y-intercept.

The main challenge always is always to calculate m or the slope. This is done by dividing the standard deviation of y values by the standard deviation of x values and then multiply this by the correlation between x and y [1]. When the slope is positive it looks like in figure 1. On the other hand when m is negative it looks like in figure 2.

Whether the Slope is positive or negative it is still considered a Linear Regression, so basically the linearity is independent of the slope value.

After calculating the slope, c has to be calculated as well. It is calculated by subtracting the mean value of x multiply

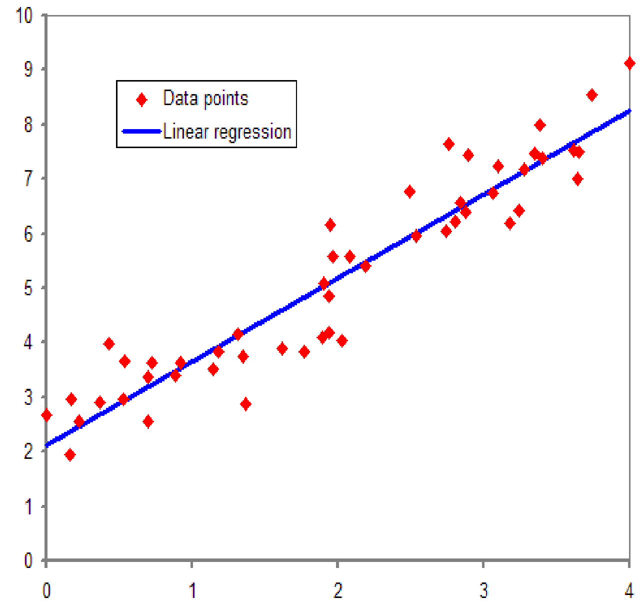

Fig. 1. Positive Linear Regression [2]

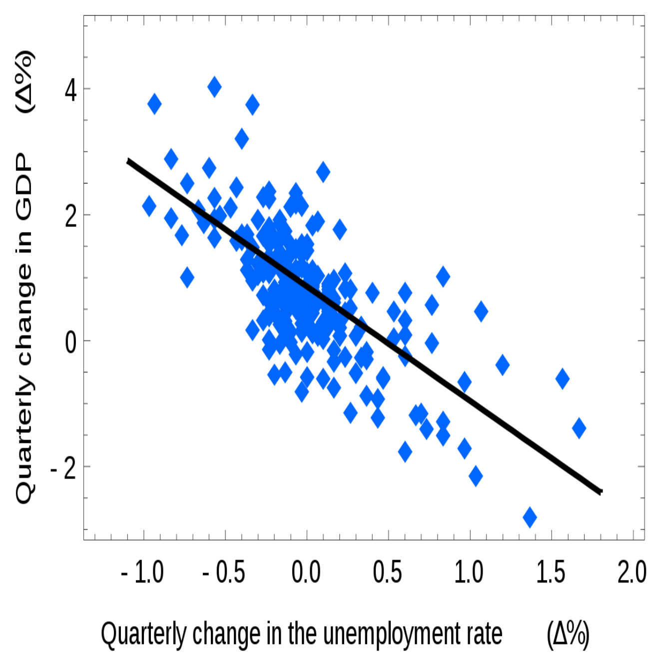

Fig. 2. Negative Linear Regression [3]

by the slope from the mean value of y.

Now after having all the elements of the equation available, the data must be mapped to the graph then the regression line is plotted. The main goal is to have the best fitting regression line with the least error values. This is done by the R-squared method which is also known as the Coefficient of Determination. The main purpose of this Coefficient is to measure the accuracy of the regression line by measuring the ”Expected Value” of the dependent variable y Vs the ”Actual Value”. This is done by the formula shown in figure 3.

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$

$$= 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}.$$

Fig. 3. R2 Formula

The sum squared regression is the sum of the residuals squared, and the total sum of squares is the sum of the distance the data is away from the mean all squared. The value of R2 is between 0 and 1, the nearer to 1 the better and more concrete your Regression line is.

Although Linear Regression is normally simple to apply, it could be the case that the data available has no linear relation to each other. This can be detected by just looking at the graph,nevertheless it will be even clearer after calculating the R2 as its value will be very low and much nearer to zero than to 1. This is one of the limitations of linear regression which will be discussed more in depth in the ”Limitations” section. More about the calculations will also be discussed in the ”Example” section.

D. Advantages

There are many advantages of Linear Regression [4]. Let’s just name few:

1) Simple in Implementation:
Linear Regression is a very simple algorithm that can be implemented very easily to give satisfactory results.Furthermore, these models can be trained easily and efficiently even on systems with relatively low computational power when compared to other complex algorithms.Linear regression has a considerably lower time complexity when compared to some of the other machine learning algorithms.The mathematical equations of Linear regression are also fairly easy to understand and interpret.Hence Linear regression is very easy to master.
2) Performance on linearly seperable datasets:
Linear regression fits linearly seperable datasets almost perfectly and is often used to find the nature of the relationship between variables.
3) Overfitting can be reduced by regularization:
Overfitting is a situation that arises when a machine learning model fits a dataset very closely and hence captures the noisy data as well.This negatively impacts the performance of model and reduces its accuracy on the test set.
Regularization is a technique that can be easily implemented and is capable of effectively reducing the complexity of a function so as to reduce the risk of overfitting.

E. Limitations

Despite its obvious advantages, Linear Regression also has some disadvantages or let’s call it ”limitations” that makes it not the best approach to use sometimes [4]. These limitations are:

1) Prone to underfitting:
Underfitting : ”A sitiuation that arises when a machine learning model fails to capture the data properly.This typically occurs when the hypothesis function cannot fit the data well.” [4] As shown in fig.4 .

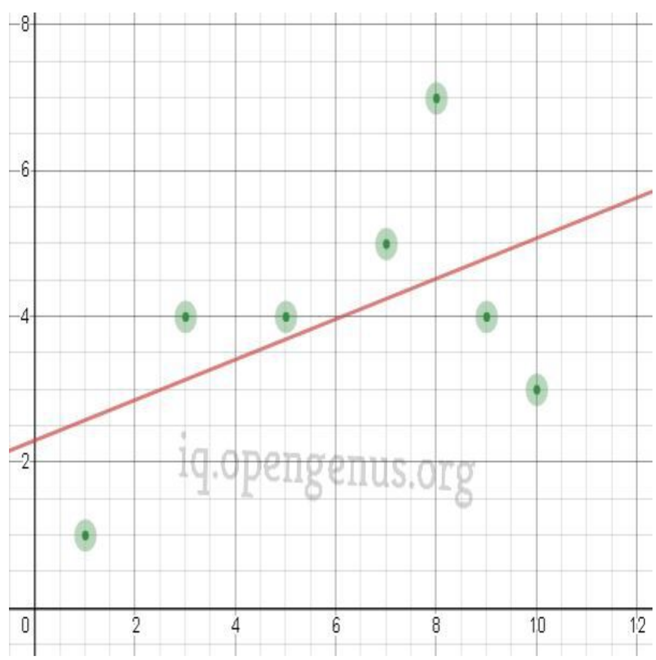

Fig. 4. Underfitting [4]

Since linear regression assumes a linear relationship between the input and output varaibles, it fails to fit complex datasets properly. In most real life scenarios the relationship between the variables of the dataset isn't linear and hence a straight line doesn't fit the data properly. In such situations a more complex function can capture the data more effectively.Because of this most linear regression models have low accuracy.

2) Sensitive to outliers:
As shown in fig. 5, Outliers of a data set are anomalies or extreme values that deviate from the other data points of the distribution. Data outliers can damage the performance of a machine learning model drastically and can often lead to models with low accuracy. Outliers can have a very big impact on linear regression's performance and hence they must be dealt with appropriately before linear regression is applied on the dataset.

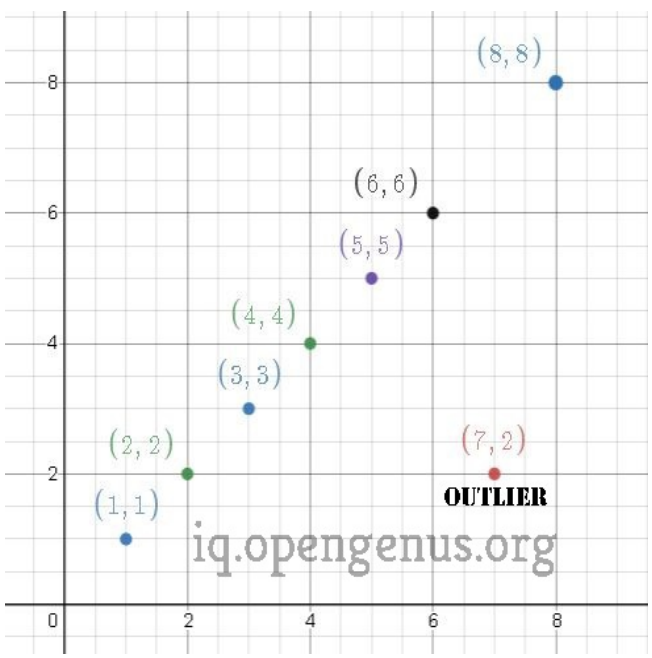

Fig. 5. Outlier [4]

3) Linear Regression assumes that the data is independent:
Very often the inputs aren't independent of each other and hence any multicollinearity must be removed before applying linear regression.

F. Example

The example that will be discussed in this paper is the Python implementation of the Linear Regression algorithm using scikit library. The building of the code can be split on multiple steps [5], let's break it down:

1) Import packages and classes:

2) Provide data:

3) Create a model and fit it:

4) Get results:

5) Predict response:

The full code can be found here:
https://github.com/MikeBlackbeard/AutonomousSystems
/tree/main/Deep_Learning_Seminar/Yahia/Algorithm%20Implementation

III. Conclusion

- To be Updated Shortly -

- Initial Bibliography -
Source [6]
Source [7]
Source [8]
Source [9]

References

[1] "How to calculate a regression line — gocardless," https://gocardless.com/guides/posts/how-to-calculate-a-regression-line/, (Accessed on 05/20/2022).

[2] "Simple linear regression - wikipedia," https://en.wikipedia.org/wiki/Simple$_linear_regression, (Accessedon 05/20/2022).$

[3] "Regression analysis - wikipedia," https://en.wikipedia.org/wiki/Regression$_analysis, (Acces$

[4] "Advantages and disadvantages of linear regression," https://iq.opengenus.org/advantages-and-disadvantages-of-linear-regression/, (Accessed on 05/26/2022).

[5] "Linear regression in python – real python," https://realpython.com/linear-regression-in-python/, (Accessed on 05/26/2022).

[6] B. Sravani and M. M. Bala, "Prediction of student performance using linear regression," in *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1–5.

[7] "Introduction to machine learning algorithms: Linear regression — by rohith gandhi — towards data science," https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a, (Accessed on 04/07/2022).

[8] D. Montgomery, E. Peck, and G. Vining, *Introduction to Linear Regression Analysis*, ser. Wiley Series in Probability and Statistics. Wiley, 2015. [Online]. Available: https://books.google.de/books?id=27kOCgAAQBAJ

[9] G. Bonaccorso, *Mastering Machine Learning Algorithms: Expert techniques to implement popular machine learning algorithms and fine-tune your models*. Packt Publishing, 2018. [Online]. Available: https://books.google.de/books?id=2HteDwAAQBAJ