Ordinary least squares

Table of Contents

> ***Linear regression might be the first "model" almost every online course teaches due to its simplicity, but its elegance can not be undermined. There are a set of conditions that a Linear regressor assumes to be true and these are usually misunderstood. The assumptions and any tests for checking the assumptions is mentioned in detail below.***

Assumptions

1. Functional form

Requirement

The response variable (dependent variable, y) should have a linear relation to the explanatory variables (independent variables, **X**).

$$y = \beta * \mathbf{X} + \epsilon$$

ϵ is the vector of residual errors (or simply residuals).

Failure

If there is not a linear correlation between the features and the label, Linear Regression can not model the dataset.

Tests for verifying linearity of data

1. Visual inspection.
   Visually inspect the plot of y against individual features (X_i's).
2. Calculate ***Pearson Correlation Coefficient (r).***
   Pearson Correlation Coefficient is a number between -1 and 1 denoting the strength and direction of linearity of the two variables. For Linear Regression we want the values to be as close to 1/-1 as possible. (r = 0 denotes no correlation, r>0.5 is usually good enough).

2. IID of residuals

- ## Independence of Residuals
  ### Requirement

  Residuals need to be independent random variables for our model's performance to be optimal. Since residual errors are a part of the data the model was unable to explain, non independent residuals mean that there is some underlying pattern in the data that the model could not learn during its training. Thus making the model sub-optimal.

  ### Failure

  If the residuals are not independent, it may be due to the following reasons:

  1. First assumption is false, i.e., data is not linear. If the features aren't linearly related to the response variable, the non-linearity shows up as a pattern in the residuals.
  2. One or more important features (explanatory variables, X_i's) are missing. That is, there was some relation between that missing variable and 'y' that appears in the residuals.
  3. Multicollinearity; If two of the features (explanatory variables, X_i's) are linearly related, the model's coefficients become unstable. For extreme multicollinearity, fitting the model becomes impossible as the model's least squares solver starts to throw infinities.

  ### Tests for verifying independence of residuals

  If the data is time-series use the **Dubin-Watson test** (a.k.a lag-1 autocorrelation test) to measure the degree of correlation between a residual and its previous residual.

If the data is non time-series then try the following tests:

1. Relation between residuals and predicted values

   Visual inspection of the plot of residuals against response variable. Any trend in some portion of the response variable's values shows less reliable predictions.

2. Relation between residuals and independent variables

   Visual inspection of the plot of residuals against individual independent variables. Usually, due to missing important explanatory variable(s).

- **Identical probability distributions of Residuals**

Desirable property. If not satisfied, some tests like the F-test might fail.

3. Normality of Residuals

Requirement

As stated above. Normality is *simply a desirable property* and even if not satisfied, the model may perform well. It's just that some statistically significant tests might fail. So, even with non-normal residuals one might have a good enough model, but it'd be difficult to quote the confidence intervals of the predictions.

We might often encounter data that are not even identically distributed, let alone normal. Normality only tells that a large proportion of our residuals have small errors.

> 💡 Special Case: Bi-modally distributed residuals: If the residual errors show two peaks, they are known to have bi-modal distribution. It usually happens due to missing an important explanatory variable (mostly binary explanatory variable). For e.g. if we miss an important feature having binary values, say 0 and 1. The model will fit the average of these two numbers (~0.5) and residuals will have two peaks, one around '0' and another around '1'.

Tests for normality

1. Histogram of residuals. (Easier to interpret, always perform and then test for conditions 2 or 3)
2. QQ-plot. (Harder to interpret, but more precise)
3. Skewness and Kurtosis (ideally 0 and 3). P-value of <= 0.05 on the **Jarque Bera Test** or **Omnibus K-squared test.**

4. Homoscedasticity

Homoscedasticity means constant variance. In the context of OLS Regression it means that the residuals must have a constant variance throughout the dataset.

Linear Regression from Scratch

After the assumptions are understood we can go on and make a Linear regressor of our own. This repository contains four different implementations of Linear Regression from scratch. Firstly using vanilla Python, then with the the help of Statmodels package of python, and then with the help of two different frameworks TensorFlow and Pytorch.