Multiple Linear Regression

Subham Roy
Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY 14260
subhamro@buffalo.edu

1 Overview

This report is about linear regression. Linear regression is a useful tool for predicting a quantitative response. Though it may not be particularly a modern, linear regression is still useful and widely used as a statistical learning method. Moreover, it serves as a good starting point for newer approaches. In this report, we review the key ideas of Multiple Linear regression model, as well as implement the same on a Dataset and analyze the outcome.

2 About Multiple Linear Regression

Simple linear regression was predicting a response based on one single predictor variable or one of many at a given time. However, in reality we have more than one predictor variable. For example, we will be using in this report, the Fish Market data. Here, we will examine the relationship between Weight of the fish and data for the Species and the various dimensions of the fishes, and then we want to know whether these variables are associated with the weight of the fish. Here one option that we have is to run multiple separate simple linear regressions, each of which uses a different attribute as a predictor. Though, we the chosen dataset one can argue it to be reasonable, but if the dataset is large, it won't be an option. Secondly, some of the attributes may not be useful when taken individually and may provide unclear results that might not be helpful. Instead of fitting separate simple linear regression models for individual predictors, a better approach is to extend the simple linear regression model so that it can directly accommodate multiple predictors. We can do this by giving each predictor a separate slope coefficient in a single model. In general, suppose that we have n distinct predictors. Then the multiple linear regression model takes the form -

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

where Xj represents the jth predictor, βj quantifies the association between that predictor and the response and ϵ some error in the prediction. We interpret βj as the average effect on Y of a one unit increase in Xj , holding all other predictors fixed. In the Fish Market example, becomes –

$$\text{Weight} = \beta_0 + \beta_1 \text{ x Species} + \beta_2 \text{ x Length1} + \cdots + \epsilon$$

3 Experiment

The Dataset used in this experiment was taken from Kaggle <https://www.kaggle.com/datasets/aungpyaeap/fish-market>

Initially, we start with loading the Dataset and checking the dimensions and the number of variables within the Dataset. In this case the dataset has 7 variables and a dimension of 159 rows and 7 columns.

Next step is we perform data preprocessing. We perform the following steps before we can be certain that we can apply the regression model. This is almost similar for all the techniques we apply. First, we check for missing data, then check for the spread of the data that means where the Mean, Median of the data, are there any outliers, we take care of that. Now that we have cleaned the dataset, we check for categorical data, and if needed we format that to binary using One Hot Encoding or simply sequentially replacing it with factors. This helps the regression model evaluate all the parameters without any bias.

In this case study, the dataset had no particular Null values or outliers. One thing we noticed in our particular dataset is that it had a variable which was Categorical containing values

such as ['Roach', 'Pike', 'Perch', 'Bream', 'Parkki', 'Whitefish', 'Smelt']. These are the species of the fishes within the dataset. So, we decided to turn this into binary with One Hot Encoding. Here, we completed our data cleaning and preprocessing before implementing the regression model. Now we check how each predictor are correlated to the response variable. This helps us identify important variables and ignore the ones that might be redundant, because we already know more data always doesn't equate to good data and can lead to very misleading estimates of the association. Below is the Correlation matrix over the dataset –

	Weight	Length1	Length2	Length3	Height	Width
Weight	1.000000	0.915712	0.918618	0.923044	0.724345	0.886507
Length1	0.915712	1.000000	0.999517	0.992031	0.625378	0.867050
Length2	0.918618	0.999517	1.000000	0.994103	0.640441	0.873547
Length3	0.923044	0.992031	0.994103	1.000000	0.703409	0.878520
Height	0.724345	0.625378	0.640441	0.703409	1.000000	0.792881
Width	0.886507	0.867050	0.873547	0.878520	0.792881	1.000000

We see particularly Lengths are highly correlated to Weight (i.e., over 90%) while Height and Width are lower compared to that. Here we can ignore Height and Width if required (But in our analysis, we did not remove any columns for predictions as it did not affect the fit by a significant margin).

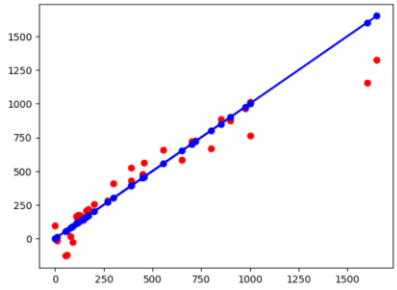
We now separate the dataset in two parts – first we choose x (i.e., the independent variable) and second y that we plan on predicting. Here x has many independent variables such as ['Species','Length1','Length2','Length3','Height','Width'] and y is Weight of the fishes. Then we split the dataset into two parts – Training set and Test set. Training set data will help our model to build a linear relationship model between independent and dependent variables more precisely finding the values of $\hat{\beta}_0$, $\hat{\beta}_1$,$\hat{\beta}_2$, $\hat{\beta}_3$ and so on.

Now in the next step, by using x-test dataset we predict y-test.

4 Analysis

We find out the Intercept and the Coefficient values i.e., $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ and so on from our analysis. In this case we got – Intercept approximately -749.593 and Coefficient value as -180.896, -57.0365, 48.799 among a few.

Now, we plot the for x-test dataset we predict y-test –



Considering it is Multiple Linear Regression, the fit looks close enough to all the actual datapoints.

Overall, it has an RMSE of 6. 7165629 which is on the higher side for a Linear Regression model but still acceptable. Now, evaluating the overall fit of the prediction, Rsquared came out to be approximately 91.02%, which is great considering it is a Multiple Linear Regression.

References

[1] Gareth James [at], Daniela Witten [aut], Trevor Hastie [aut, cre], Rob Tibshirani [aut], Balasubramanian Narasimhan [ctb], *Introduction to Statistical Learning, Second Edition*.

[2] Tom M. Mitchell, *Machine Learning: A multistrategy approach, 1997 Edition*.