Simple Linear Regression

Subham Roy
Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY 14260

1 Overview

This report is about linear regression, a simple approach for supervised learning. Linear regression is a useful tool for predicting a quantitative response. It is comparatively an old technique. Though it may not be particularly a modern technique compared to some of the more advanced statistical learning approaches, linear regression is still useful and widely used as a statistical learning method. Moreover, it serves as a good starting point for newer approaches. Consequently, it is important to have a good understanding of linear regression before working on more complex learning methods. In this report, we review the key ideas of Simple Linear regression model, as well as implement the same on a Dataset and analyze the outcome.

2 About Simple Linear Regression

Simple linear regression is a very simple approach for predicting a quantitative response Y based on a single independent variable X. It will assume that there is approximately a linear relationship between X and Y . Mathematically, the linear relationship can be written as -

$$Y \approx \beta_0 + \beta_1 X$$

Here, "$\approx$" is shown because it "is approximately modeled as" . Here, we will sometimes describe the above relationship by saying that we are regressing Y on X. For TV-Marketing dataset, X may represent TV advertising information and Y may represent the corresponding sales. Then we can regress sales onto TV by fitting the Simple linear regression model -

$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV.}$$

In the above equation, β_0 and β_1 are two unknown constants that represent the intercept and slope terms for the linear model. Once we use our training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we can predict future sales based on the particular value of TV advertising data by computing the below equation -

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{y}$ indicates the prediction of Y based on the value of X = x. Here we used a hat symbol, ˆ , to denote the predicted or estimated value for an unknown parameter.

3 Experiment

The Dataset used in this experiment was taken from Kaggle <https://www.kaggle.com/datasets/devzohaib/tvmarketingcsv>

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 2 variables and a dimension of 200 rows and 2 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.

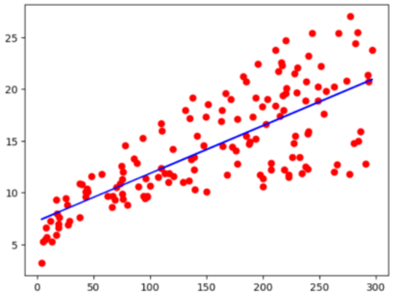
In this case study, since this was an experiment for Simple linear regression, the dataset chosen is pretty simple. Here there are no missing data, data is scaled, no significant outliers and lastly, no categorical data. So, we need not to apply any data preprocessing for this particular dataset.

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 is TV advertising information and y is Sales data. 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 between independent and dependent variables more precisely the value of $\hat{\beta}_0$ and $\hat{\beta}_1$.

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

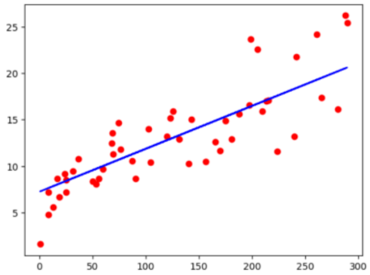
4 Analysis

Below plot shows the relationship between x_train and y_train dataset –



We can also find out the Intercept and the Coefficient values i.e., $\hat{\beta}_0$ and $\hat{\beta}_1$ from our analysis. In this case we got – Intercept approximately 7.24891 and Coefficient value as 0.04614341.

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



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

Overall, it has an RMSE of 2.954699 which respectable for Simple Linear Regression. Now, evaluating the overall fit of the prediction, Rsquared came out to be approximately 70%, which is good considering it is a Simple Linear Regression with just two variables.

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*.