Article

On

“Linear Regression Model”

Submitted by

Swami Prem Pranav Kayashyap

Linear Regression Model (part-1)

In this article, we dive into machine learning models, more specific to linear regression models. We will build a simple linear regression model in R.

Introduction

Model: A model is a transformation engine that helps us to express dependent variable as a function of independent variables. There are mainly three machine learning techniques.

- Regression
- Classification, and
- Clustering.

One of the most important regression models in machine learning is linear regression. It is said that “All models are wrong. Some are useful.”

Linear regression model tries to approximate the relationship between dependent and independent variables in a straight line.

$y = \beta_0 + \beta_1 x + \varepsilon$

β_0 and β_1 are two unknown constants that represent the intercept and slope. They are the parameters. Parameters are ingredients added to the model for estimating the output and ε is the error term.

Model Creation

We will look into it with TV marketing example. Where we need to build a model which takes input as marketing budget and model can forecast the sales.

First, we want to evaluate if indeed we can predict sales based on marketing budget. The first set of analysis seeks the answers to the following questions:

- Is sale of TV related with marketing budget? How strong is the relationship?
- Is the relationship linear?
- Can we predict/estimate TV sales price based on marketing budget?

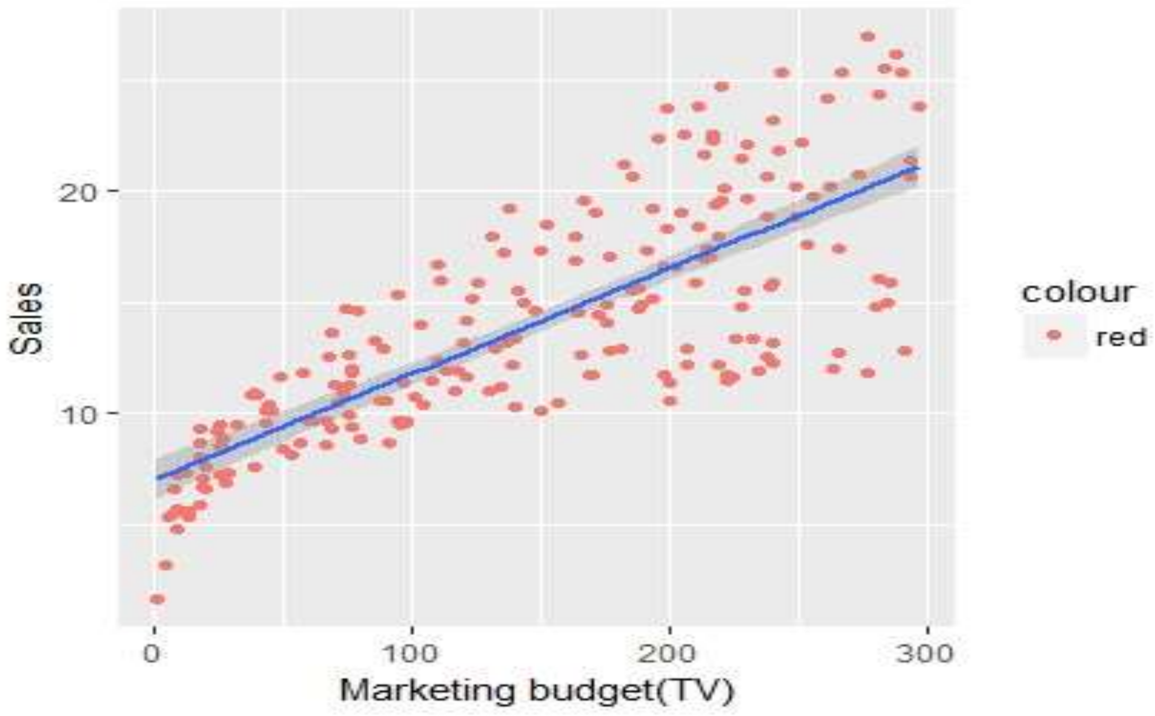
We can do correlation analysis using ggplot() function and cor() in R. It can be figured out from the following plot that there is a correlation between marketing budget and sales. Correlation coefficient is 0.78 which means quite strong relationship.

As you can see a straight line can fit which mean relationship is linear.

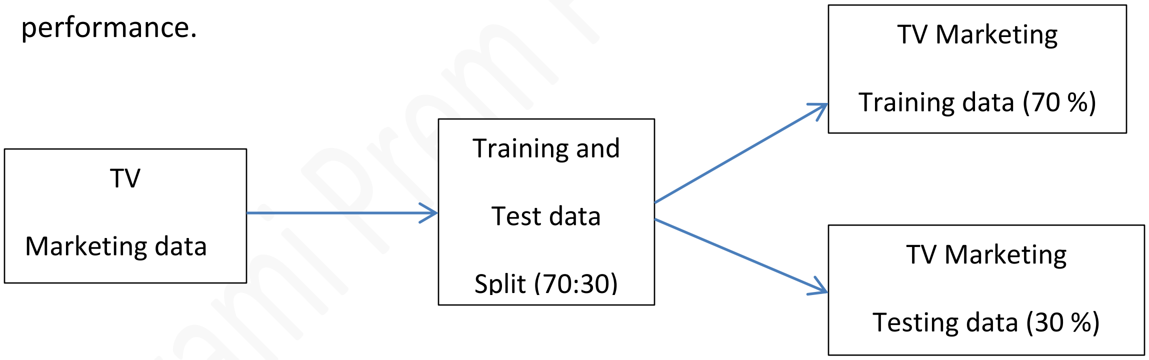
Equation of the model is as follow: Sales = 7.220933 + 0.047 x marketing budget

Linear Regression Model (part-1)

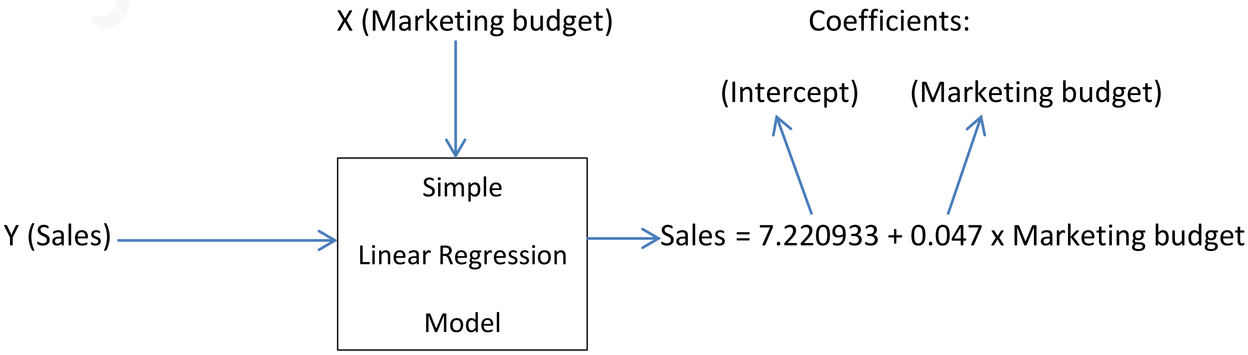
As we can write an equation which means sales can be estimated based on marketing budget.



We need to splits the data into training and test set. 70% of data can be used for training. Remaining can be used for the test. The training data is used to learn about the data. The training data is used to create the model. The testing data is used to evaluate the model performance.



We can builds a linear regression model using lm() function in R. The model creates a linear equation that expresses Sales price of TV as a function of marketing budget.



Congratulations! The model is built. One unit increase in marketing budget will increase the average sales of the TV by 0.047 units.

Model evaluation

The evaluation is done in two parts.

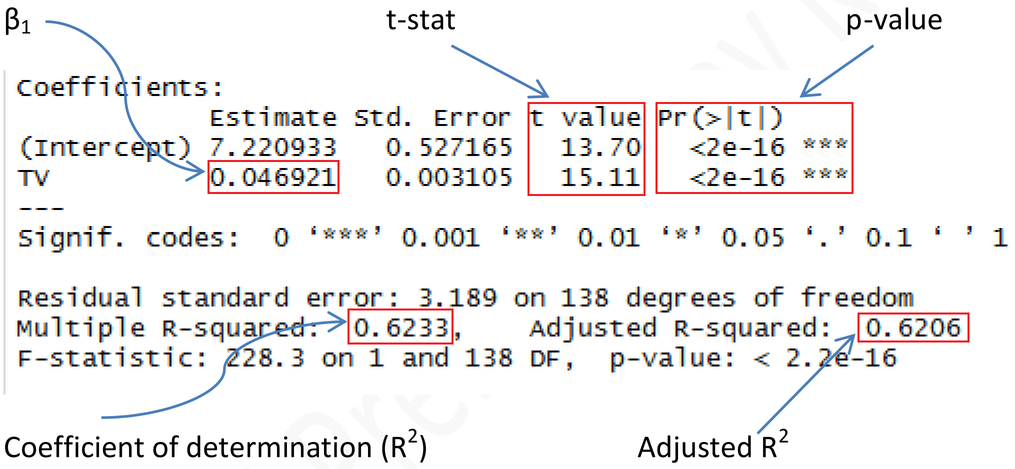
1. Test to establish the robustness of the model.
2. Test to evaluate the accuracy of the model.

The robustness of the model is evaluated using hypothesis testing.

Ho (NULL hypothesis): $\beta_1 = 0$; There is no relationship between Sales and Market budget.

Ha (Alternate hypothesis): $\beta_1 \neq 0$; There exist a relationship between Sales and Market budget.

Based on the training dataset of 140 observation (t-distribution) following statistics we have:



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.220933   0.527165   13.70   <2e-16 ***
TV          0.046921   0.003105   15.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.189 on 138 degrees of freedom
Multiple R-squared: 0.6233,    Adjusted R-squared:  0.6206
F-statistic: 228.3 on 1 and 138 DF,  p-value: < 2.2e-16
```

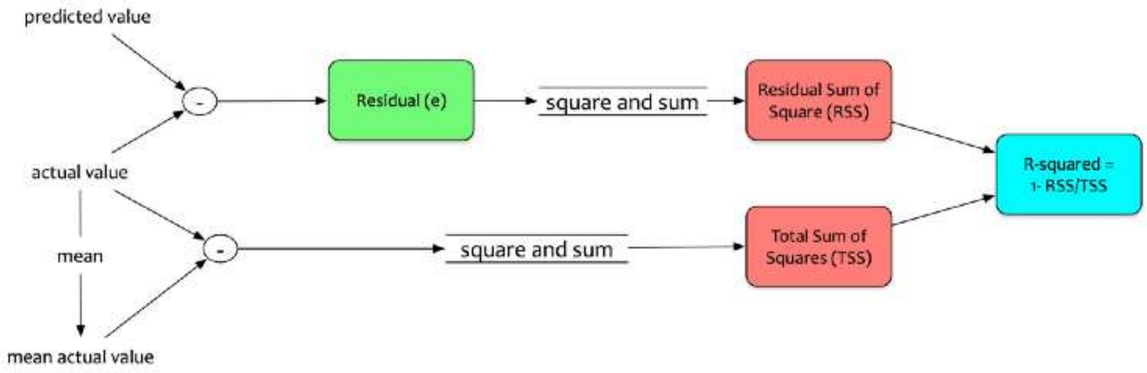
β_1: The value of β_1 determines the relationship between Sales and Market budget. β_1 represents the difference in the predicted value of Sales for each one-unit difference in Market budget. If $\beta_1 = 0$ then there is no relationship. In this case, β_1 is positive. It implies that there is some relationship between Sales and Market budget.

P-value: A low p-value (< 0.05) indicates that you can reject the null hypothesis ($\beta_1 = 0$;). β_1 having low p-value is likely to be a meaningful addition to our model.

T-state: The t-stat is the coefficient divided by its standard error. It means that how many standard deviations the coefficient estimate (β1) is far away from zero. Further, it is away from zero stronger the relationship between price and engine size. The coefficient is significant here, as in this case, t-stat is 15.11. It is far enough from zero.

Coefficient of determination (R^2): $R^2 = 1 – (RSS/TSS)$

This metric explains the fraction of variance between the values predicted by the model and the value as opposed to the mean of the actual. This value is between 0 and 1. The higher it is, the better the model can explain the variance.



- Error (ε) is the difference between the actual y and the predicted y by our model. These errors are also called as residuals.
- Residual Sum of Squares (RSS): Residuals is evaluated for each observation. Then all the residual values are squared and added. Lower the RSS, the better it is.
- Total sum of squares (TSS): Find the differences between the mean of actual values and actual value. These differences are then squared and added.

Adjusted R^2: $R^2_{adj} = 1 – [(1-R^2) * (n-1) / (n-k-1)]$; where n = size of sample and k is number of independent variables (in our case, it is only market budget).

We can test to evaluate the accuracy of the model using predict function in R. We need to input testing data (30%) into this predict function and compare those predicted sales value with actual Sales value.