Lab Report 01

Linear Regression in R

Submitted To
Kisan Kumar Ganguly
Lecturer
Institute of Information Technology
University of Dhaka

Submitted By
Tulshi Chandra Das
BSSE 0811

Date**:** 23 September 2019

Lab Report 1

Introduction

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. [1] The regression model is used widely for data mining projects and machine learning fields. The mathematical equation for a single predictor variable can be generalized as follows:

$Y = \beta 1 + \beta 2X + \epsilon$

where, $\beta 1$ is the intercept and $\beta 2$ is the slope. Collectively, they are called regression coefficients. ϵ is the error term, the part of Y the regression model is unable to explain.

Objective

- Getting experiences in making regression model in R.
- Getting concept of concept of application of linear regression in real life problem.

Methodology

Selecting data set

The first step is selection of a dataset

Exploratory data analysis

I should analyze the variables before building the regression model. The graphical view of dataset provides good overall understanding of dataset.

Graphical Analysis

The aim of this exercise is to build a simple regression model that I can use to predict Distance (dist) by establishing a statistically significant linear relationship with Speed (speed). But before jumping in to the syntax, let’s try to understand these variables graphically. Typically, for each of the independent variables (predictors), the following plots are drawn to visualize the following behavior:

1. Scatter plot: Visualize the linear relationship between the predictor and response

2. Box plot: To spot any outlier observations in the variable. Having outliers in my predictor can drastically affect the predictions as they can easily affect the direction/slope of the line of best fit.
3. Density plot: To see the distribution of the predictor variable. Ideally, a close to normal distribution (a bell-shaped curve), without being skewed to the left or right is preferred. Let us see how to make each one of them.

.

Finding correlation

Correlation is a statistical measure that suggests the level of linear dependence between two variables, that occur in pair. Correlation can take values between -1 to +1. If I observe for every instance where predictor increases, the response also increases along with it, then there is a high positive correlation between them and therefore the correlation between them will be closer to 1. The opposite is true for an inverse relationship, in which case, the correlation between the variables will be close to -1.
A value closer to 0 suggests a weak relationship between the variables. A low correlation ($-0.2 < x < 0.2$) probably suggests that much of variation of the response variable (Y) is unexplained by the predictor (X), in which case, I should probably look for better explanatory variables.

Build Linear Model

After finding the linear relationship pictorially in the scatter plot and by computing the correlation, the next step is building the linear model.

Linear Regression Diagnostics

After the linear model is built and I have a formula that I can use to predict the response value if a corresponding predictor is known, but it is not enough to use this model. Before using a regression model, I need to be ensured that it is statistically significant.

The p Value: Checking for statistical significance

The p-Values are very important because, I can consider a linear model to be statistically significant only when both these p-Values are less that the pre-determined statistical significance level, which is ideally 0.05. This is visually interpreted by the significance stars at the end of the row. The more the stars beside the variable’s p-Value, the more significant the variable.

Null and alternate hypothesis

When there is a p-value, there is a hull and alternative hypothesis associated with it. In Linear Regression, the Null Hypothesis is that the coefficients associated with the variables is equal to zero. The alternate hypothesis is that the coefficients are not equal to zero (i.e. there exists a relationship between the independent variable in question and the dependent variable).

T-value

I can interpret the t-value something like this. A larger t-value indicates that it is less likely that the coefficient is not equal to zero purely by chance. So, higher the t-value, the better.

Predicting Linear Models

After building the model, I need to test how the model performs with the new data. So the preferred practice is to split the dataset into a 80:20 sample (training: test), then, build the model on the 80% sample and then use the model thus built to predict the dependent variable on test data. Doing it this way, I will have the model predicted values for the 20% data (test) as well as the actuals (from the original dataset). By calculating accuracy measures (like minimax accuracy) and error rates (MAPE or MSE), I can find out the prediction accuracy of the model.

k- Fold Cross validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

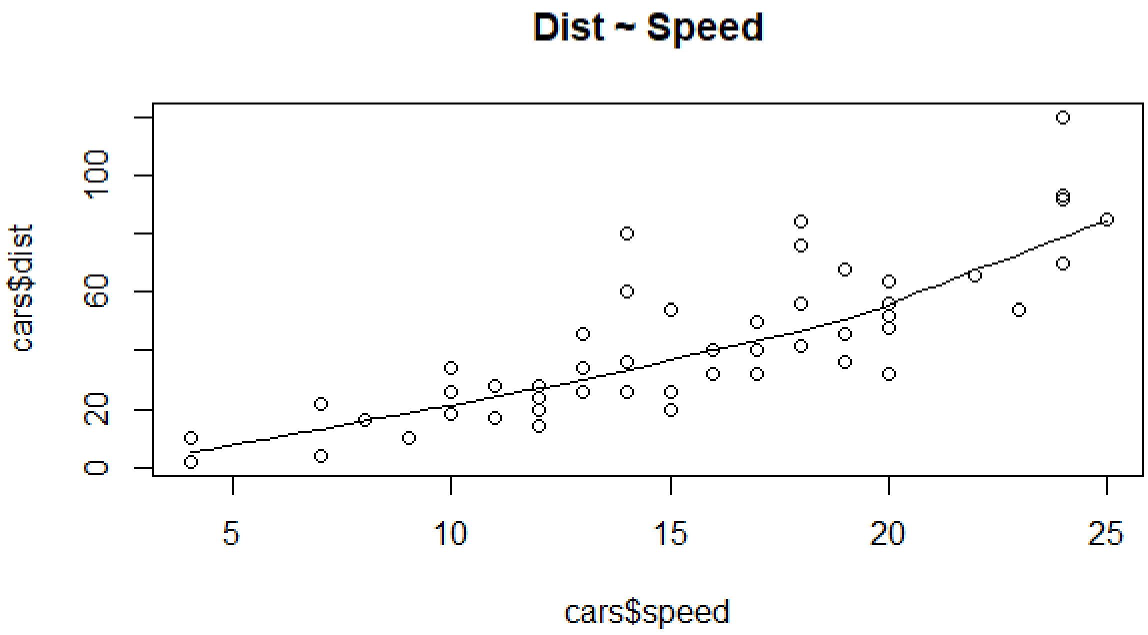
The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
    1. Take the group as a hold out or test data set
    2. Take the remaining groups as a training data set
    3. Fit a model on the training set and evaluate it on the test set
    4. Retain the evaluation score and discard the model

4. Summarize the skill of the model using the sample of model evaluation scores

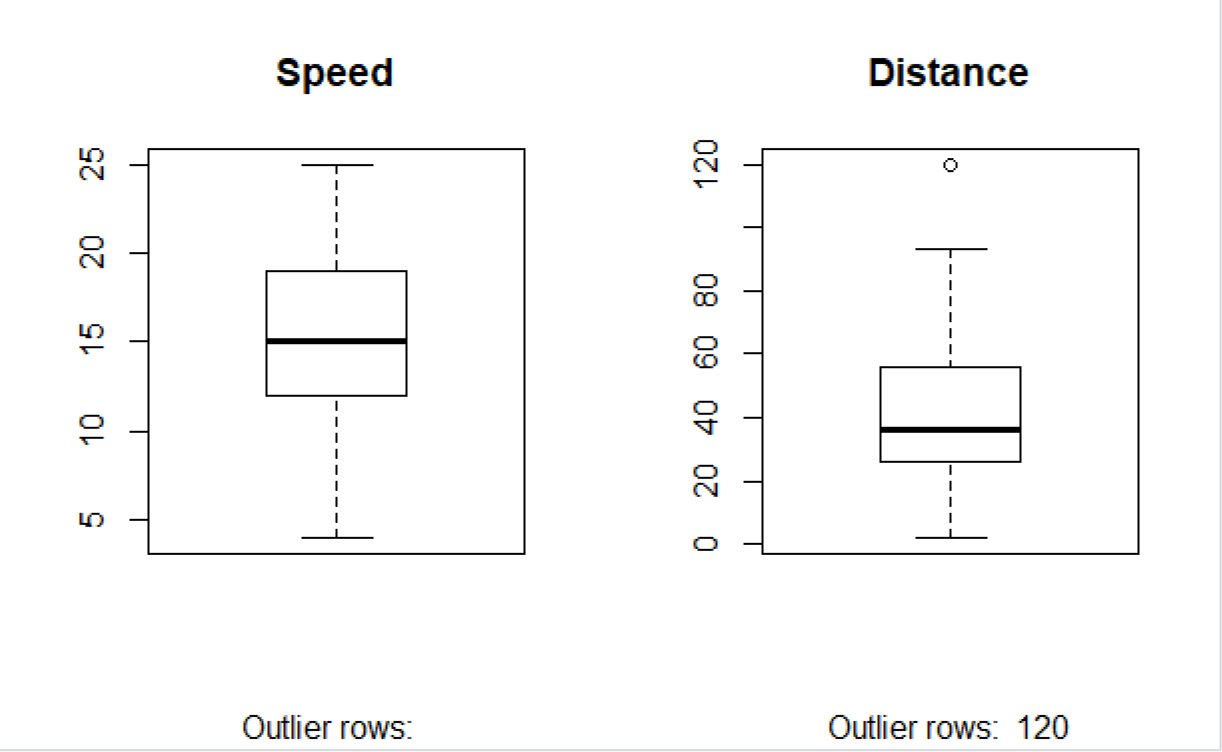
Result and Discussion

For this analysis, I used the cars dataset that comes with R by default. cars is a standard built-in dataset.
Scatter Plot

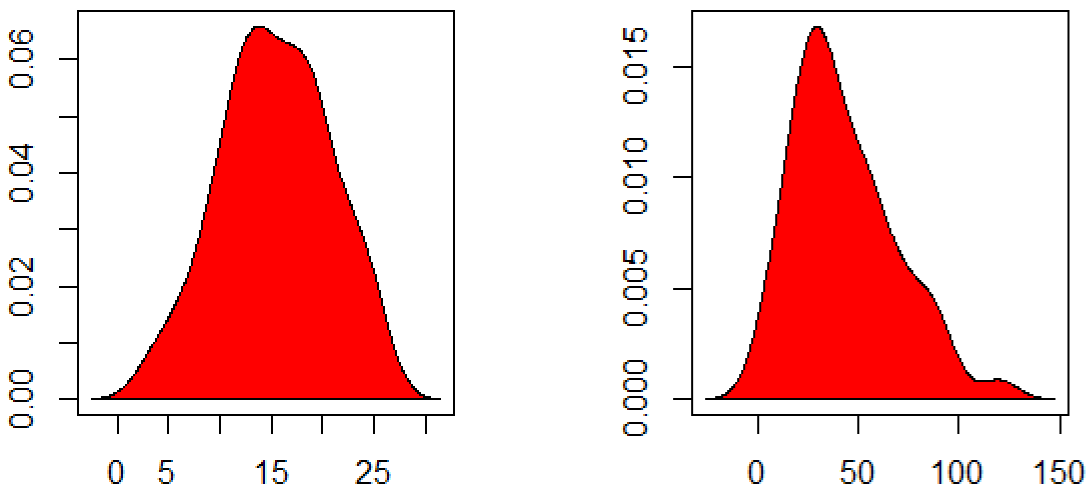


The scatter plot along with the smoothing line above suggests a linearly increasing relationship between the 'dist' and 'speed' variables. There is a linear relationship between the response and predictor variables.

BoxPlot – Check for outliers



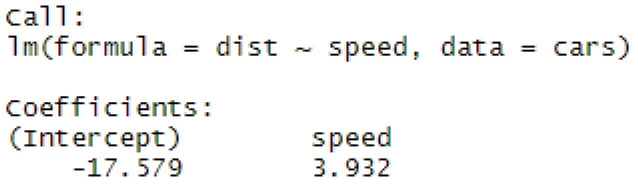
Density plot – Check if the response variable is close to normality



Correlation

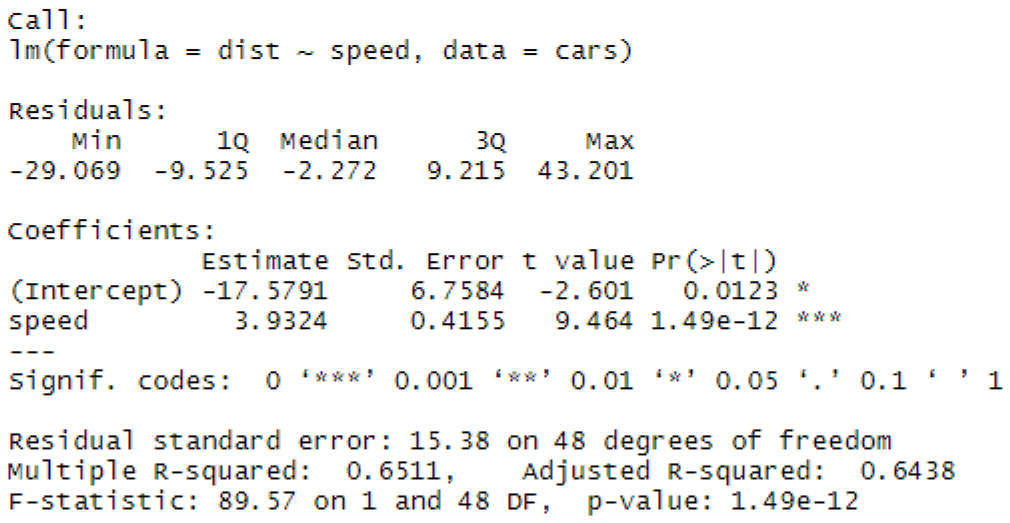
Correlation between the response and the predictor variable is 0.8068949.

Building Linear Model



```
Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
    -17.579        3.932
```



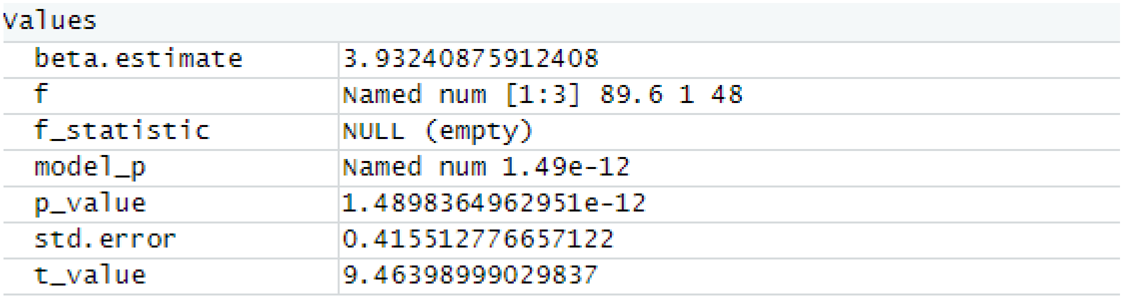
```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

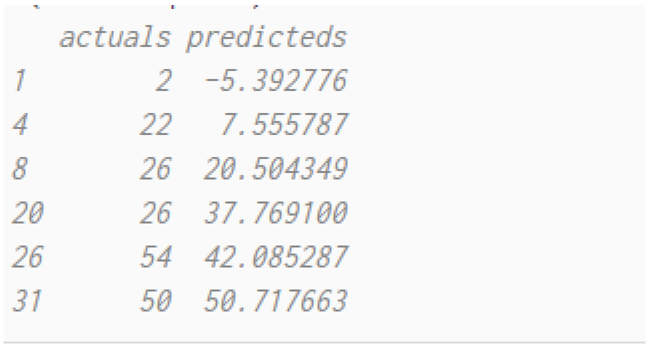
The small p-value refers to statistical significant impact on response variable distance of the speed.



Values

beta.estimate	3.93240875912408
f	Named num [1:3] 89.6 1 48
f_statistic	NULL (empty)
model_p	Named num 1.49e-12
p_value	1.4898364962951e-12
std.error	0.415512776657122
t_value	9.46398999029837

Calculate prediction accuracy and error rates



```
   actuals predicteds
1        2  -5.392776
4       22   7.555787
8       26  20.504349
20      26  37.769100
26      54  42.085287
31      50  50.717663
```

Conclusion

In this experiment, I have worked with the fundamental things of linear regression model in R. I have learnt how to perform exploratory analysis on my dataset to find out the relationship between the predictor and response, build linear regression model, measure statistical significance of the predictor and predict test data.