# Implementation of Linear and Logistic Regression (Packages: glmnet, earth)

Expt No: 5 April 3,2019
Author: Subalakshmi Shanthosi S (186001008)

## Aim

Implementation of Linear and Logistic Regression using R program.

## Description

1. Linear Regression:
   - Linear regression is used to predict the value of an outcome variable Y based on one or more input predictor variables X.
   - Linear regression aims at establishing a linear relationship (a mathematical formula) between the predictor variable(s) and the response variable, so that, we can use this formula to estimate the value of the response Y, when only the predictors (Xs) values are known.
2. Logistic Regression:
   - Extension of Linear Regression where the dependent variable is categorical and not continuous. It predicts the probability of the outcome variable.
   - Logistic regression can be binomial or multinomial.
   - Binomial Logistic regression :
     - Value= Yes
     - Value= No
   - Multinomial Logistic Regression : Outcome having 3 possibilities:
     - Value= Best
     - Value= Very Good
     - Value= Best

## Equation

1. Linear Regression Formula:

   $$Y = \beta_1 + \beta_2 X + \epsilon \tag{1}$$

   where the Regression Coefficients are:

   $\beta_1$ is the intercept
   $\beta_2$ is the slope
   $\epsilon$ is the error time period for which there is no prediction of value Y

2. Logistic Regression Formula:

   $$P(y=1) = \frac{1}{(1 + e^{(}-1*(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)))} \tag{2}$$

   where the Logistic Regression Coefficients are:

   $\beta$ is the coefficient selected to maximize the likelihood of predicting a high probability for observations.

# Tools and Packages

1. Tools
   - RStudio.
   - R Version 1.1.463
2. Linear and Logistic Regression Packages
   - glmnet
   - earth

# Procedure

1. Split the data set as:
   - Training dataset.
   - Testing dataset.
2. Develop the model on the training data and use it to predict the distance on test data.
3. Review diagnostic measures.
4. Calculate prediction accuracy and error rates
5. Metric Calculation

# How to know if the model is best fit for your data?

1. R-Squared: R-Squared tells us the proportion of variation in the dependent variable that has been used.

$$R^2 = 1 - \frac{SSE}{SST} \tag{3}$$

2. Adj R-Squared: Adj R-Squared penalizes total value for the number of terms (read predictors) in your model. Therefore when comparing nested models, it is a good practice to look at adj-R-squared value over R-squared.

$$R^2_{adj} = 1 - \frac{((1 - R^2)(n-1))}{(n-q)} \tag{4}$$

3. Standard Error and F-Statistic: Both standard errors and F-statistic are measures of goodness of fit.

$$Std.Error = \sqrt{MSE} = \sqrt{\frac{SSE}{n-q}} \tag{5}$$

$$F - statistic = \frac{MSR}{MSE} \tag{6}$$

where, n is the number of observations, q is the number of coefficients and MSR is the mean square regression, calculated as,

$$MSR = \frac{\sum_i^n \left(y_i \hat{-} \bar{y}\right)}{q-1} = \frac{SST - SSE}{q-1} \tag{7}$$

4. AIC and BIC: The Akaikes information criterion - AIC and the Bayesian information criterion - BIC are measures of the goodness of fit of an estimated statistical model and can also be used for model selection. Both criteria depend on the maximized value of the likelihood function L for the estimated model. The AIC is defined as:

$$AIC = (2)ln(L) + (2k) \tag{8}$$

where, k is the number of model parameters and the BIC is defined as:

$$BIC = (2)ln(L) + kln(n) \tag{9}$$

Table 1: Output: Statistical Features

| STATISTIC | VALUE | CRITERION |
|---|---|---|
| R-Squared | 15.38 | Higher the better and greater 0.70 |
| Adj R-Squared | 0.6438 | Higher the better |
| F-Statistic | 23.113231 | Higher the better |
| Std. Error | 0.4869 | Closer to zero the better |
| t-statistic | 8.8634 | Should be greater 1.96 for p-value to be less than 0.05 |
| AIC | 338.4489 | Lower the better |
| BIC | 343.5155 | Lower the better |
| Mallows cp | 10.1 | Should be close to the number of predictors in model |
| MAPE (Mean absolute percentage error) | 0.6995 | Lower the better |
| MSE (Mean squared error) | 23.113231 | Lower the better |
| Min Max Accuracy | 0.38004 | Higher the better |

## Coding

```
# Create Training and Test data -
set.seed(100)  # setting seed to reproduce results of random sampling
trainingRowIndex <- sample(1:nrow(cars), 0.8*nrow(cars))  # row indices for training data
trainingData <- cars[trainingRowIndex, ]  # model training data
testData  <- cars[-trainingRowIndex, ]   # test data

# Build the model on training data -
lmMod <- lm(dist ~ speed, data=trainingData)  # build the model
distPred <- predict(lmMod, testData)  # predict distance

summary (lmMod)  # model summary

AIC (lmMod)  # Calculate akaike information criterion

actuals_preds <- data.frame(cbind(actuals=testData$dist, predicteds=distPred))  # make act
correlation_accuracy <- cor(actuals_preds)  # 82.7%
head(actuals_preds)

min_max_accuracy <- mean(apply(actuals_preds, 1, min) / apply(actuals_preds, 1, max))
mape <- mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals
```

## Result

Thus the implementation of Linear and Logistic Regression is executed successfully using R program.