**Assignment 1 – Linear Regression**

**Athira Beena Thulaseedharan**

**Std ID:300202270, STAT271**

**Consider the following Regression model:**

**TPrice = β0 + β1 Mileage + β2 Cyl + β3 Liter + β4 Doors+ β5Cruise + β6 Sound + β7 Leather + ε**

**Apply the techniques including variable selection techniques and model assumptions checking.**

**(a) determine the best fitted regression model.**

**(b) interpret the coefficients from the final regression model in a.**

Read csv file into variable. cars.data is a data frame.

**cars.data <- read.csv** ("C:\\Users\\Athira Beena Thulasi\\OneDrive\\Documents\\UFV\\DAC\\Fall Sem\\Stat\_271\_Data Analysis and Modelling\\UFV\_Dataset\\Cars.csv")

Summary fun in R returns the results of basic statistical calculations (min, 1st qt, median, mean, 3rd qt, and max) for a numerical vector

For categorical vector - it returns Class, Mode and Length.

**summary(cars.data)**

Structure of the cars returns the type, number of observations & variables, response variable & explanatory variable, their datatypes.

**str(cars.data)**

The following model is the transformation of the full model with all predictor variables.

Applied log10 to the response variable to make it the best fit.

**TPrice <- log10(cars.data$Price)**

**model.trf <- lm (TPrice ~ Mileage + Cyl + Liter + Doors + Cruise + Sound + Leather, data = cars.data)**

**summary(model.trf)**

Forward and backward stepwise regression

**forward <- step (model.trf, direction = "forward")**

**summary(forward)**

**backward <- step (model.trf, direction = "backward")**

**summary(backward)**

Best Subset selection method

leaps () perform an exhaustive search for the best subsets of the variables in x for predicting y in linear regression, using an efficient branch-and-bound algorithm.

**library(leaps)**

**data.pred <-cbind(cars.data$Mileage,cars.data$Cyl,cars.data$Liter,cars.data$Doors,cars.data$Cruise,cars.data$Sound,cars.data$Leather)**

Apply adjR2 to choose the best fitted model

**lm. adjr2 <- leaps (data.pred,TPrice, method = "adjr2",**

**names = c('Mileage','Cyl','Liter','Doors','Cruise','Sound','Leather'), nbest=3)**

WHICH gives the TRUE indices of a logical object, allowing for array indices.

**cb<-cbind (lm.adjr2$which, lm.adjr2$adjr2)**

Then the suggested model is :

**Price ~ Mileage + Cyl + Litre + Doors + Cruise + Sound + Leather + Mileage**

Mallow's Cp to choose the best fitted model

**lm.cp <- leaps (data.pred,TPrice, method = "Cp", names = c('Mileage','Cyl','Liter','Doors','Cruise','Sound','Leather'), nbest=3)**

**cbind (lm.cp$which, lm.cp$Cp)**

The model with all predictor variables has Cp = 8, which is close to the number of parameters in the regression model. So, the suggested model is :

**Price ~ Mileage + Cyl +Litres + Doors + Cruise + Sound + Leather + Mileage**

**Assumptions Check**

**Diagnostic Plot / Residual Plot**

**Residual plots** are often used to assess whether or not the [residuals](https://www.statology.org/residuals/) in a regression analysis are normally distributed and whether or not they exhibit [heteroscedasticity](https://www.statology.org/understanding-heteroscedasticity-in-regression-analysis/).

**Par ()** is used to set/query graphical parameters

**mfrow ()** allows to split the screen in several panels.

**plot** () generic fun for plotting graphs – scatter plot

**Par (mfrow =c (2,2))**

**plot(model.trf)**

yDiagram, schematic

Description automatically generated

**Residual Vs Fitted Plot:** The red lines representing the mean of the residuals are all basically horizontal and centred around zero. This means there are no outliers or biases in the data that would make a linear regression invalid.

**QQ Plot:** for the most part this sample data appears to be normally distributed.

**Fitted values vs. Residuals plot**

**Plot (fit, residual)**

**abline (0,0, col="red")**

**abline ()** add straight lines to a plot.

Chart, scatter chart

Description automatically generated

**Histogram of residuals**

To check whether the variance is normally distributed

**hist(residual)**

**hist(residual.trf) hist ()** computes histogram of the given data values

Chart

Description automatically generated

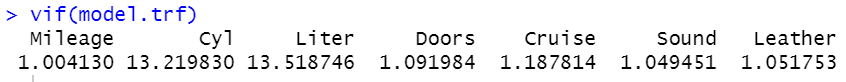
**The observations are roughly bell-shaped, so can proceed with the linear regression.**

**Multicollinearity**

Multicollinearity can be assessed by computing a score called VIF which measures how much the variance of a regression Coeff is inflated due to multicollinearity in the model VIF () calculates variance Inflation for regression model

**model.trf <- lm (TPrice ~ Mileage + Cyl + Liter + Doors + Cruise + Sound+ Leather, data = cars.data)**

**vif (model.trf)**

****

Here, **Cyl and Liter contain VIF > 5,** which indicates these **two variables are multicollinear**. But as the **goal is to create a model for predicting car price,** multicollinearity is not a problem and there is no need to eliminate one of the variables.

**Model for price prediction**

**Price = 4018 - .000003236 Milage + .3360 Cyl + .03050 litre - .01305 doors + .01361cruise - .03906 sound + .05187 leather**

**Interpretation of Regression coefficients**

Milage: For every unit increase in Milage, the Price is expected to decrease by 0.000003236 units, holding all the variables constant.

Cyl: For every unit increase in the number of Cylinders, the Price is expected to increase by 0.3360 units, holding all other predictor variables constant

litre: For every unit increase in the litre, the Price is expected to increase by 0.03050 units, holding all other predictor variables constant

Doors: For every unit increase in the number of doors, the Price is expected to decrease by 0.01305 units, holding all other predictor variables constant

Cruise: For every unit increase in the presence of cruise control, the Price is expected to increase by 0.01361 units, holding all other predictor variables constant

Sound: For every unit increase in the sound feature, the Price is expected to decrease by 0.03906 units, holding all other predictor variables constant

Leather: On average, price is estimated to increase by .05187 if the car has leather seats holding all other predictor variables constant.