Lab 5-Write Up and R Code

library(xlsx)

library(car)

library(MASS)

install.packages('leaps')

library(leaps)

data<-read.xlsx('Lab\_5\_Dataset.xlsx',1)

attach(data)

data$Parish<-factor(data$Parish,levels=c('Kingston','Montego Bay','Clarendon'),labels=c(1,2,3))

summary(data)

scatterplotMatrix(data) #produces the matrix below

Diagram

Description automatically generated

* Increased research and development resulted in increased profits (uptrend in scatterplot).
* Profit also increased with marketing; this is observed from the plot.
* Fluctuations in the amount spent on R&D appear to be partially explained by the amount spent on marketing.
* Data is relatively evenly distributed over all parishes. R&D spent, Administration Cost, Marketing Cost and Profit are roughly normal in distribution.

**Multiple Linear Model Building:**

mlmod1<-lm(Profit~.,data)

summary(mlmod1)

Text, letter

Description automatically generated

* Significance testing (at 95% confidence) indicates that R&D is the only significant predictor variable (that is the estimator isn’t equal to 0). This implies that a linear would be best in this case.
* Note this model contains all predictor variables.

stepAIC(mlmod1,direction="backward") #used for feature selection, minimizing this value will point to the optimal model

Text

Description automatically generated

* Based on model selection AIC criterion, having the lowest AIC=822.88 indicates the best model from this selection algorithm.
* Method balances goodness of fit and a penalty for model complexity.

selectmlmod<- lm(Profit~ R.D.Spend+Administration, data=data) # stores the best model

mlmod2<-lm(Profit~1,data=data) #this creates a model in which the predicted values will be the mean value of y. In the forward direction, predictor variables will be added to the model based on the AIC criterion

stepAIC(mlmod2,direction="forward",scope=(~R.D.Spend+Administration+Marketing.Spend+Parish),trace=0) #forward stepping AIC model selection generator code

**Another Method:**

vs1<-regsubsets(Profit~.,data=data,nbest=1,method="exhaustive",really.big=T) #selects the best model by RSS (given multiple predictors)

plot(vs1,scale="adjr2",main="Adjusted R^2 Vs Predictors")

x11()

Chart

Description automatically generated

* Adding variables beyond “Administration Cost” does not influence the Adj-R2 value significantly enough to validate their addition to the final model. The best model according to this method contains the: intercept, Admin Cost and R&D Spent.

**Checking Model Validity:**

**Normality of Residuals:**

par(mfrow=c(2,2))

qqPlot(mlmod1,labels=row.names(data),id.method="identify",simulate=T,main="QQ Plot of Multiple Mod1") #Gives a QQ plot of the residuals(having point that fall on or that closely hug the straight line is ideal)

qqPlot(selectmlmod,labels=row.names(data),id.method="identify",simulate=T,main="QQ Plot of Selected Multiple Mod")

Chart, line chart, scatter chart

Description automatically generated

* Both plots show points varying significantly from the straight line with outliers 16 and 15 highlighted. Raises the question on true normality of residuals.

ncvTest(mlmod1) #conducts a chi square and hypothesis test for constant variance, the null hypothesis here is constant variance.

ncvTest(mlmod1)

Text

Description automatically generated

* At 95% confidence we cannot reject the null hypothesis for either model, both p values are >0.05.

spreadLevelPlot(mlmod1) #produces a plot of standardized residuals vs the values produced by the model.

spreadLevelPlot(selectmlmod)

Chart, scatter chart

Description automatically generated

* This plot shows greater randomness for the selected model from the AIC process. This improved randomness shows better normality of the residuals in the second model(one on the right).

**Auto-correlation of Errors:**

durbinwatsonTest(mlmod1) #test for correlation between the error terms

durbinwatsonTest(selectmlmod)

Text

Description automatically generated

* Both tests show some level of autocorrelation occurring between error terms, error terms are somewhat influenced by other error terms.

**Auto-correlation of Errors:**

outlierTest(mlmod1) #checks if there are any significant outlier affecting the model

outlierTest(selectmlmod)

Text

Description automatically generated

* The Bonferroni statistic has been used to identify entry 16 an outlier (that is data for the individual regressor variables). This corresponds to row 17, let us remove it and run model diagnostics once more.

**Revised Model:**

editdata<-data[-c(16),]

newmod<- lm(Profit~ R.D.Spend+Administration, data=editdata)

ncvTest(newmod)

Text

Description automatically generated with medium confidence

* The significance of any non-constant variance has decreased. At a 95% confidence level variance is constant for this model (this further validates it’s use).

durbinWatsonTest(newmod)

A picture containing text

Description automatically generated

* The significance of any correlation between the error terms has lessened by 0.614, in fact there is now no auto-correlation in comparison to “selectmlmod”(D-W Statistic is now >2).

**Influential Observation:**

influencePlot(newmod) #creates a Cook’s Distance plot using residuals and fitted values

Chart, diagram, bubble chart

Description automatically generated

* Observations above y=2 and below y=-2 are outliers (observations 36 and 15).
* Along x- axis are predicted values, those >2 indicate high leverage data points.
* The largest circle indicates the influential observation by Cook’s Distance (36, 15, 3 and 37).