**Mini Project 6 (Solution)**

**Mini Project Duo Group # 12**

**Contribution of each group member**

Chetan Siddappareddy – 50%

Ankit Sahu – 50%

Both of us have contributed equally to the project. We learnt R through collaboration and then write the R scripts for the corresponding and report all the findings.

**Section 1**

**Problem 1**

Setting the work environment and loading the csv data.

#setting the env to the location of the project

setwd("/Users/sahuankit010/Desktop/Repo/CS-6313-Stats/Mini Projects/MP6")

#checks whether it is set or not

getwd()

#loading the data into the cancer\_data variable

cancer\_data = read.csv("prostate\_cancer.csv")

#printing the data

cancer\_data

Table

Description automatically generated

In Order to verify the correlation among the variables, we are plotting correlation matrix.

#installing the corrplot, it provides a visual exploratory tool on correlation matrix that supports automatic variable reordering to help detect hidden patterns among variables.

install.packages("corrplot")

#loading the corrplot library

library(corrplot)

#finding correlation

cor.data = cor(cancer\_data)

corrplot(cor.data)

#attaching the data so that we can use the variables

attach(cancer\_data)

Chart, bubble chart

Description automatically generated

Using different plots:

plot(psa)

hist(psa)

boxplot(psa)

Chart, line chart

Description automatically generated

Chart, histogram

Description automatically generatedEngineering drawing

Description automatically generated

From the box plot, we can see there are many outliers in the data. We need transformation to the data to fit our linear model. Here we are applying logarithmic transformation and analyzing using the plot and box plot.

psa.log = log(psa)

plot(psa.log)

boxplot(psa.log)

Chart, line chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

As vesinv is a qualitative variable we use as\_factor converts a variable into a factor and preserves the value and variable label attributes.

cancer\_data$vesinv = as.factor(cancer\_data$vesinv)

Now fitting the models.

**Model 1**

Null Hypothesis-> H0: None of the predictors are usefuk for predicting response.

Alternate Hypothesis -> H1: Atleast one of the predictors is useful for predicting the response.

Code:

fit1 = lm(psa.log ~ cancervol + vesinv + capspen + gleason + weight + age + benpros)

summary(fit1)

Output:

Table

Description automatically generated

From the results we can see that the cancervol which is \*\*\*, vesinv, gleason, benpros which has \*\* are the significant predictors. Hence, we reject the null hypothesis.

**Model 2:**

For the above hypothesis now we will only consider the significant predictors.

Code:

fit2 = update(fit1,.~. - capspen - age - weight)

summary(fit2)

Output:

Table

Description automatically generated

From the correlation matrix we know that capsen is also important so for the model 3 we can add capsen for model 2.

**Model 3:**

Code**:**

fit3 = update(fit2,.~. + capspen)

summary(fit3)

Output:

Table

Description automatically generated

We can see that the adjusted R-squared value decreases saying that capsen is not an optimal predictor for predicting the response variable.

**Comparing all the three models**

Code

anova(fit1)

anova(fit2)

anova(fit3)

anova(fit2,fit3)

anova(fit1,fit2,fit3)

Table

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

From the above results we can say that model 2 is the best linear model.

Now checking the best model using AIC

Code:

Text

Description automatically generated

Output:

Text, letter

Description automatically generated

Comparing the AIC scores of the three fitted models.

Code:

A picture containing text

Description automatically generated

Output:

Text

Description automatically generated

We can see from the above results l1 or fit2 linear model has the lowest aic score saying that it’s the best model among all the models.

**Model Evaluation**

Code:

plot(fitted(fit2),resid(fit2), main = "Residual Plot")

abline(h=0)

Output:

Chart, scatter chart

Description automatically generated

The points are scattered around zero and there is not pattern. So, we can say the errors have mean zero and constant variance.

Code:

qqnorm(resid(fit2))

qqline(resid(fit2))

Output:

Chart

Description automatically generated

Errors are normally distributed.

Code:

plot(resid(fit2),type = "l")

abline(h=0)

Output:

A picture containing antenna

Description automatically generated

Use the final model to predict the PSA level for a patient whose quantitative predictors are at the sample means of the variables and qualitative predictors are at the most frequent category.

lm(formula = y~ cancervol + vesinv + gleason + benpros)

Code:

summary(fit2)

Output:

Table

Description automatically generated

Now, predicting PSA with the model lm(formula = y~cancervol + vesinv + gleason + benpros)

Code:

table(gleason)

table(vesinv)

Text

Description automatically generated with medium confidence

Calculatng the mean of the gleason and vesinv.

Code:

mean(cancervol)

mean(benpros)

Output:

Text

Description automatically generated

From the above result we can see that gleason value 7 is being dominated in the data, vesinv value 0 is being dominated in the data and the mean of cancervol and benpros are 6.998 and 2.534 respectively.

Predicted value is equal to:

-0.65013 + 6.998682 \* (0.06488) + 7\*(0.33376) + 0.09136\* (2.534725) = 2.371837

Thus, the actual value of PSA is exp(2.371837) which is equal to 10.71706.

**Section 2**

#setting the env to the location of the project

setwd("/Users/sahuankit010/Desktop/Repo/CS-6313-Stats/Mini Projects/MP6")

#checks whether it is set or not

getwd()

#loading the data into the cancer\_data variable

cancer\_data = read.csv("prostate\_cancer.csv")

#printing the data

cancer\_data

#installing the corrplot, it provides a visual exploratory tool on correlation matrix that supports automatic variable reordering to help detect hidden patterns among variables.

install.packages("corrplot")

#loading the corrplot library

library(corrplot)

#finding correlation

cor.data = cor(cancer\_data)

corrplot(cor.data)

#attaching the data so that we can use the variables

attach(cancer\_data)

#plotting with plot, histogram and boxplot for psa.

plot(psa)

hist(psa)

boxplot(psa)

psa.log = log(psa)

plot(psa.log)

boxplot(psa.log)

cancer\_data$vesinv = as.factor(cancer\_data$vesinv)

#Fitting the linear models

#model 1

fit1 = lm(psa.log ~ cancervol + vesinv + capspen + gleason + weight + age + benpros)

summary(fit1)

#model 2

#reduced model

fit2 = update(fit1,.~. - capspen - age - weight)

summary(fit2)

#model 3

fit3 = update(fit2,.~. + capspen)

summary(fit3)

#analyzing all the three models

anova(fit1)

anova(fit2)

anova(fit3)

anova(fit2,fit3)

anova(fit1,fit2,fit3)

#checking the best model using the AI

fit.full = fit1 = lm(psa.log ~ cancervol + vesinv + capspen + gleason + weight + age + benpros)

for.aic = step(lm(psa.log ~ 1), direction = "forward", scope = formula(fit.full), k = 2, trace = 0) # forward AIC

for.bic = step(lm(psa.log ~ 1), direction = "forward", scope = formula(fit.full), k = log(32), trace = 0) # forward BIC

back.aic = step(fit.full, direction = "backward", k = 2, trace = 0) # backward AIC

back.bic = step(fit.full, direction = "backward", k = log(32), trace = 0) # backward BIC

(Adjusted\_R.square = data.frame("Method"=c("for.aic", "for.bic", "back.aic", "back.bic"),"Adj.r.square"=c(summary(for.aic)$adj.r.square,

summary(for.bic)$adj.r.square,

summary(back.aic)$adj.r.square, summary(back.bic)$adj.r.square)))

l1 <- glm(fit2)

l2 <- glm(fit1)

l3 <- glm(fit3)

l1$aic

l2$aic

l3$aic

#model evaluation

plot(fitted(fit2),resid(fit2), main = "Residual Plot")

abline(h=0)

qqnorm(resid(fit2))

qqline(resid(fit2))

plot(resid(fit2),type = "l")

abline(h=0)

#summary of fit2

summary(fit2)

#print the table of various variables

table(gleason)

table(vesinv)

mean(cancervol)

mean(benpros)

**Output:**

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

**Text

Description automatically generated**

**Text

Description automatically generated**