Statistical Method of Data Science Project #6

Name: Prashant Prakash

Net Id: pxp141730

P**roblem 1:**

One variable is fixed as Cancer volume and then all other variables are taken as second variable and the model are compared to each others in all cases.

# since one variable is already picked in project 5 so we will check with different combinations of this variable with other variable available   
  
# read data from csv file   
data = read.csv("/Users/Prashant/Downloads/prostate\_cancer.csv",header = TRUE , sep=",")   
attach(data)  
# check with all other variables keeping one variable as CancerVolume   
  
# combination of cancer Volume and weight   
psa.reg1 <- lm(psa ~ cancervol + weight)  
summary(psa.reg1)

##   
## Call:  
## lm(formula = psa ~ cancervol + weight)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.368 -8.739 -1.584 3.386 181.233   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.19442 5.45761 0.036 0.972   
## cancervol 3.22933 0.41685 7.747 1.08e-11 \*\*\*  
## weight 0.02055 0.07188 0.286 0.776   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.19 on 94 degrees of freedom  
## Multiple R-squared: 0.3901, Adjusted R-squared: 0.3771   
## F-statistic: 30.06 on 2 and 94 DF, p-value: 8.079e-11

anova(psa.reg1)

## Analysis of Variance Table  
##   
## Response: psa  
## Df Sum Sq Mean Sq F value Pr(>F)   
## cancervol 1 62202 62202 60.0404 1.072e-11 \*\*\*  
## weight 1 85 85 0.0817 0.7756   
## Residuals 94 97385 1036   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# combination of cancer Volume and age   
psa.reg2 <- lm(psa ~ cancervol + age)  
summary(psa.reg2)

##   
## Call:  
## lm(formula = psa ~ cancervol + age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.835 -8.903 -1.695 3.015 181.331   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.63788 28.43833 0.128 0.898   
## cancervol 3.23139 0.41732 7.743 1.1e-11 \*\*\*  
## age -0.03951 0.44175 -0.089 0.929   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.2 on 94 degrees of freedom  
## Multiple R-squared: 0.3896, Adjusted R-squared: 0.3766   
## F-statistic: 30 on 2 and 94 DF, p-value: 8.382e-11

anova(psa.reg2)

## Analysis of Variance Table  
##   
## Response: psa  
## Df Sum Sq Mean Sq F value Pr(>F)   
## cancervol 1 62202 62202 59.993 1.088e-11 \*\*\*  
## age 1 8 8 0.008 0.9289   
## Residuals 94 97461 1037   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# combination of cancer volume and benpros   
  
psa.reg3 <- lm(psa ~ cancervol + benpros)  
summary(psa.reg3)

##   
## Call:  
## lm(formula = psa ~ cancervol + benpros)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -60.069 -9.505 -0.918 4.404 178.647   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.5168 5.3861 -0.282 0.779   
## cancervol 3.2767 0.4192 7.816 7.74e-12 \*\*\*  
## benpros 0.9130 1.0899 0.838 0.404   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.08 on 94 degrees of freedom  
## Multiple R-squared: 0.3941, Adjusted R-squared: 0.3812   
## F-statistic: 30.57 on 2 and 94 DF, p-value: 5.933e-11

anova(psa.reg3)

## Analysis of Variance Table  
##   
## Response: psa  
## Df Sum Sq Mean Sq F value Pr(>F)   
## cancervol 1 62202 62202 60.4361 9.484e-12 \*\*\*  
## benpros 1 722 722 0.7017 0.4043   
## Residuals 94 96747 1029   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# combination of cancer volume and vesinv   
  
psa.reg4 <- lm(psa ~ cancervol + factor(vesinv))   
summary(psa.reg4)

##   
## Call:  
## lm(formula = psa ~ cancervol + factor(vesinv))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -55.145 -7.535 -1.129 4.256 170.018   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.060 4.231 0.251 0.8027   
## cancervol 2.477 0.495 5.003 2.62e-06 \*\*\*  
## factor(vesinv)1 24.647 9.423 2.616 0.0104 \*   
## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 31.09 on 94 degrees of freedom  
## Multiple R-squared: 0.431, Adjusted R-squared: 0.4189   
## F-statistic: 35.6 on 2 and 94 DF, p-value: 3.098e-12

anova(psa.reg4)

## Analysis of Variance Table  
##   
## Response: psa  
## Df Sum Sq Mean Sq F value Pr(>F)   
## cancervol 1 62202 62202 64.3543 2.869e-12 \*\*\*  
## factor(vesinv) 1 6613 6613 6.8414 0.01038 \*   
## Residuals 94 90857 967   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# combination of cancer volume and capspen  
  
psa.reg5 <- lm(psa ~ cancervol + capspen)  
summary(psa.reg5)

##   
## Call:  
## lm(formula = psa ~ cancervol + capspen)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -60.346 -8.324 -1.205 4.159 183.843   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.3276 4.2861 0.310 0.757   
## cancervol 2.4139 0.5655 4.269 4.69e-05 \*\*\*  
## capspen 2.4533 1.1779 2.083 0.040 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 31.48 on 94 degrees of freedom  
## Multiple R-squared: 0.4165, Adjusted R-squared: 0.4041   
## F-statistic: 33.55 on 2 and 94 DF, p-value: 1.01e-11

anova(psa.reg5)

## Analysis of Variance Table  
##   
## Response: psa  
## Df Sum Sq Mean Sq F value Pr(>F)   
## cancervol 1 62202 62202 62.757 4.654e-12 \*\*\*  
## capspen 1 4300 4300 4.338 0.03999 \*   
## Residuals 94 93170 991

## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# combination of cancer volume and gleason   
  
psa.reg6 <- lm(psa ~ cancervol + gleason)  
summary(psa.reg6)

##   
## Call:  
## lm(formula = psa ~ cancervol + gleason)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -65.191 -10.410 -1.681 5.553 175.299   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -59.6587 32.9439 -1.811 0.0733 .   
## cancervol 2.8113 0.4673 6.016 3.4e-08 \*\*\*  
## gleason 9.2657 4.9788 1.861 0.0659 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 31.62 on 94 degrees of freedom  
## Multiple R-squared: 0.4113, Adjusted R-squared: 0.3987   
## F-statistic: 32.83 on 2 and 94 DF, p-value: 1.537e-11

anova(psa.reg6)

## Analysis of Variance Table  
##   
## Response: psa  
## Df Sum Sq Mean Sq F value Pr(>F)   
## cancervol 1 62202 62202 62.1985 5.518e-12 \*\*\*  
## gleason 1 3464 3464 3.4634 0.06587 .   
## Residuals 94 94006 1000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

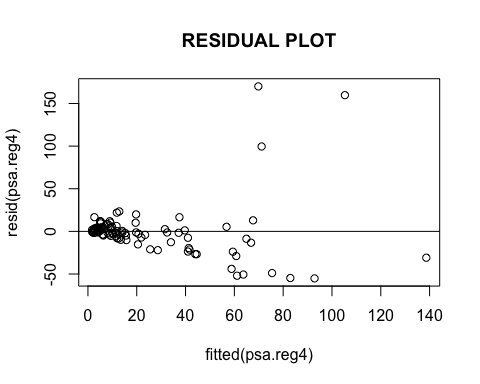
**Observations:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Combination** | **Residual Standard Error** | **R Square** | **F-Statistic** |
| **Cancer Volume and weight** | 32.19 | 0.3901 | 30.06 |
| **Cancer Volume and weight** | 32.2 | 0.3896 | 30 |
| **Cancer Volume and benpros** | 32.08 | 0.3941 | 30.57 |
| **Cancer Volume and vesinv** | 31.09 | 0.431 | 35.6 |
| **Cancer Volume and capspen** | 31.48 | 0.4165 | 33.55 |
| **Cancer Volume and Gleason** | 31.6 | 0.4113 | 31.62 |

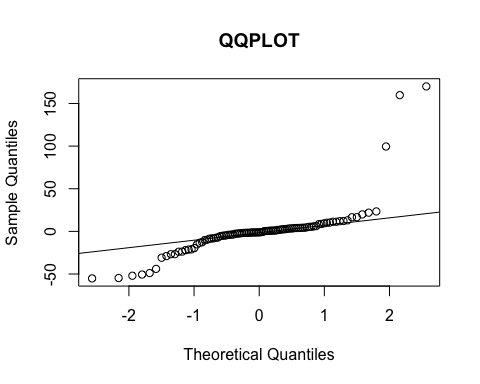
**Clearly from above table we can say that Cancer Volume and vesinv are best among all for prediction of PSA level.**

**More Regression Diagnostics**

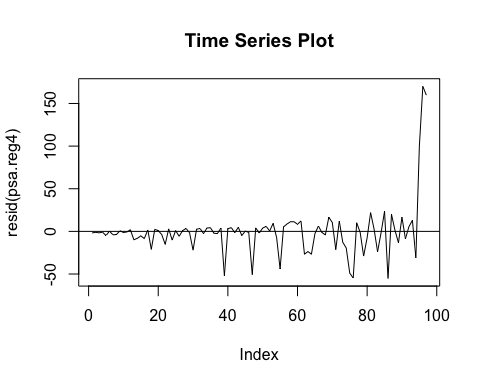
# clearly after comparsion of all regression models we find that mode with cancervol and vesinv is best among all   
  
#Check for errors whether they have mean zero constant variance,normality and independence of residual error  
# Residual Plot  
plot(fitted(psa.reg4),resid(psa.reg4), main="RESIDUAL PLOT")  
abline(h=0)



# QQ Plot  
qqnorm(resid(psa.reg4),main="QQPLOT")  
qqline(resid(psa.reg4))



# Time Series plot  
  
plot(resid(psa.reg4), type="l",main="Time Series Plot")  
abline(h=0)



**Observation:**

As we can see from different plots. We can assume following.

1. Residual Plot: we can see the graph has some pattern.
2. QQ Plot: It is not normally distributed, rather the points divert from around right tail.
3. Time Series Plot: There are some pattern and errors are dependent on each other.

**So we do log transformation**

# log transformation   
  
z1<-log(data$psa)  
z2<-log(data$cancervol)  
z3<-(data$vesinv)  
  
psa.reg4.log <- lm(z1 ~z2 + factor(z3))  
summary(psa.reg4.log)

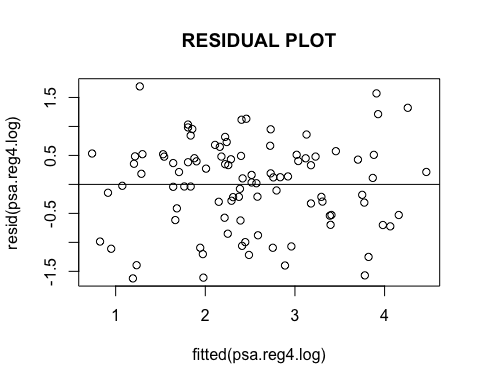
##   
## Call:  
## lm(formula = z1 ~ z2 + factor(z3))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.6217 -0.5281 0.1209 0.4840 1.6907   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.53531 0.11733 13.086 < 2e-16 \*\*\*  
## z2 0.59118 0.07767 7.611 2.07e-11 \*\*\*  
## factor(z3)1 0.67187 0.22113 3.038 0.00308 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7558 on 94 degrees of freedom  
## Multiple R-squared: 0.5797, Adjusted R-squared: 0.5708   
## F-statistic: 64.84 on 2 and 94 DF, p-value: < 2.2e-16

anova(psa.reg4.log)

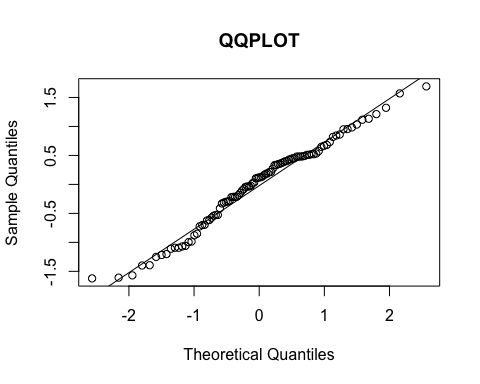
## Analysis of Variance Table  
##   
## Response: z1  
## Df Sum Sq Mean Sq F value Pr(>F)   
## z2 1 68.801 68.801 120.4445 < 2e-16 \*\*\*  
## factor(z3) 1 5.273 5.273 9.2313 0.00308 \*\*   
## Residuals 94 53.695 0.571   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**More Regression Diagnostics after Log Transformation**

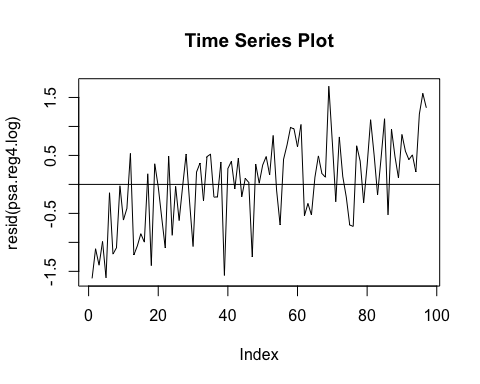
# Residual Plot  
plot(fitted(psa.reg4.log),resid(psa.reg4.log), main="RESIDUAL PLOT")  
abline(h=0)



# QQ Plot  
qqnorm(resid(psa.reg4.log),main="QQPLOT")  
qqline(resid(psa.reg4.log))



# Time Series plot  
  
plot(resid(psa.reg4.log), type="l",main="Time Series Plot")  
abline(h=0)



**Observation:**

1. Residual Plot: The points are scattered around zero and there is no pattern.
2. QQ Plot: The plot is considerably normal.
3. Time Series Plot: There is no patter and errors are independent.

**Comparison between before transformation and after transformation**

|  |  |  |
| --- | --- | --- |
| **Variable** | **R Square** | **F-statistics** |
| **Before Transformation** | 0.431 | 35.6 |
| **After Transformation** | 0.5797 | 64.84 |

**Problem 2.**

**Prediction:**

# prediction with model wothout any transformation  
x.new <- data.frame(cancervol = median(data$cancervol), vesinv = median(data$vesinv))  
predict(psa.reg4, newdata=x.new)

## 1   
## 11.61889

# prediction with model with log transformation   
log.predict<- exp(1)^predict(psa.reg4.log,data.frame(z2=log(median(data$cancervol)), z3 = median(data$vesinv)))  
log.predict

## 1   
## 10.94097