STAT 840 Linear Regression

Final Project Proposal

Juwuk Kim

05/07/2021

Introduction

Introduction to the topic and motivation

Prostate Cancer is the most common cancer for men. According to the American cancer society, about 1 man in 9 will be diagnoed with prostate cacncer in his lifetime. It is more likely to occur to African-American and about 6 cases in 10 are diagnosed at the age of 65 or older, and it is uncommon under 40. In addition, prostate cancer is the second leading cause of cancer death for men in America, behind lung cancer. About 1 man in 41 will die of prostate cancer.

Prostate Specific Antigen or PSA, is an enzyme found in men's boold produced exclusively by prostate cells. An abnormal rise in PSA, can indicate Prostate Cancer. Higher levels of PSA can be found in the blood as prostate cancer cells begin to proliferate in an uncontrolled way.

In this project, PSA level is predicted using multiple linear regression as well as association between PSA level and predictor varia1ble is discovered.

Introduction to the dataset

The dataset was obtained from one of the published datasets by The Elements of Statistical Learning. There are 97 observations(men) who have prostate cancer. The predictor variable are below. data (https://web.stanford.edu/~hastie/ElemStatLearn/)

- LPSA: Log PSA level
- LProWeight: log prostate weight
- age: age of patient
- weight: log prostate weight
- age: age of patient
- LBPH: Log of the amount of benign prostatic hyperplasia
- SemVeInv: seminal vesicle invasion
- LCanVol: log of cancer volume
- Gleason: Gleason score
- PerGG: percent of Gleason scores 4 or 5

Aims

The purpose of this data analysis is to predict the PSA level with multiple predictor variables. And the second goal is to estimate the β_i coefficients to discover association between PSA level and other predictor variables.

Statistical Model

A multiple linear regression model is utilized.
Y_i Log PSA level
X_i1 log prostate weight

X_i2 age of patient
X_i3 log prostate weight
X_i4 age of patient
X_i5 Log of the amount of benign prostatic hyperplasia
X_i6 seminal vesicle invasion
X_i7 log of capsular penetration
X_i8 Gleason score
X_i9 percent of Gleason scores 4 or 5

This is the model is considered for the first time.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \epsilon_i$$

Where $\epsilon_i \sim iidN(0, \sigma^2)$, This is a 97x9 matrix.

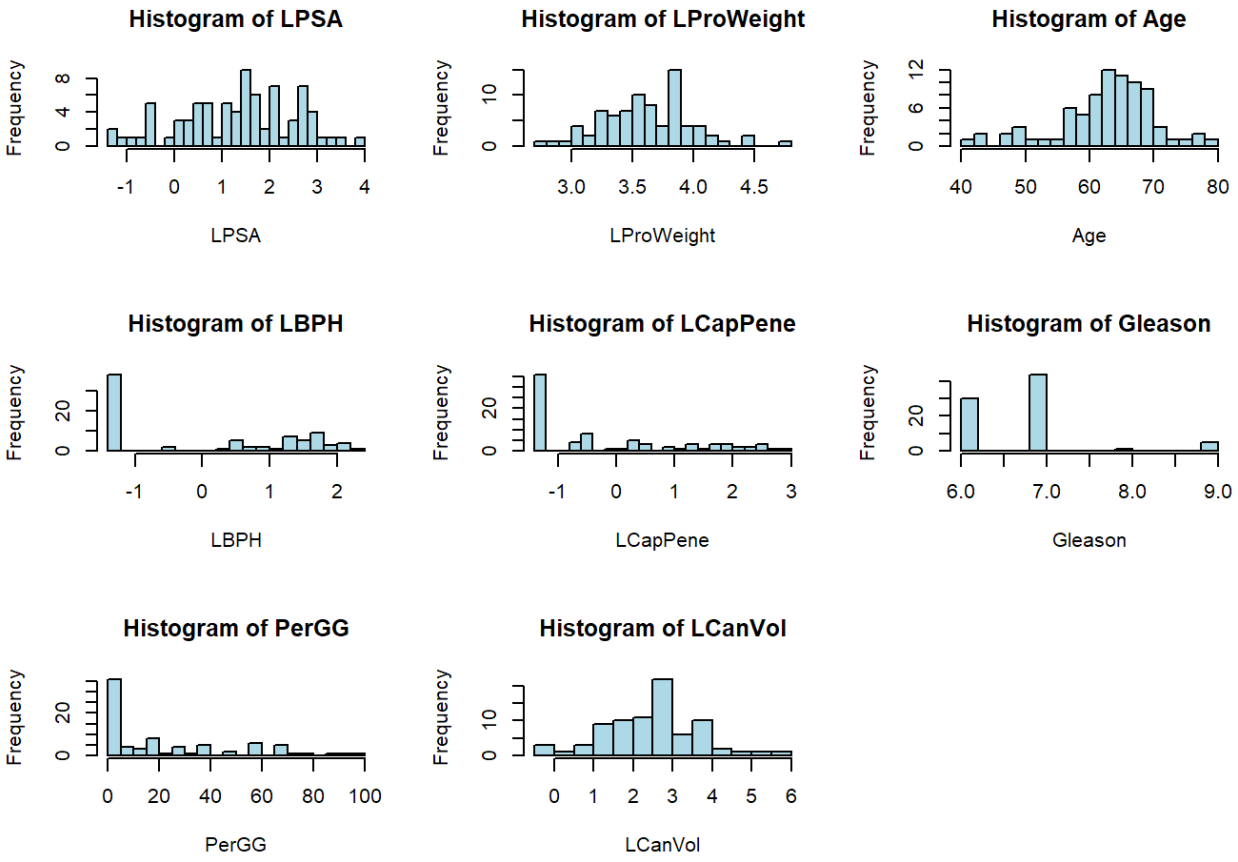
Explanatory Data Analysis

Exploring predictor variables and the response variable using plots and tables to discover any association between the predictor varialbes and the reponse variables. Even any relationship between the predictor variables is also important to the model.

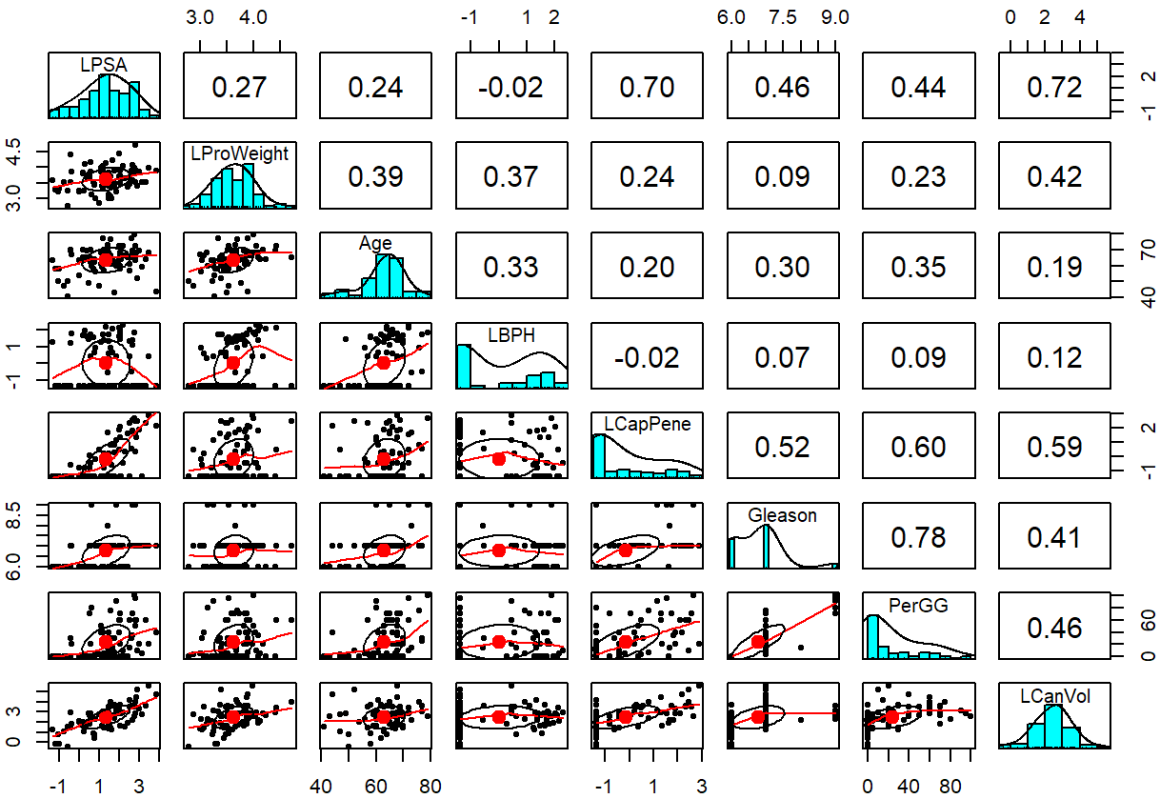
The histograms display that some of the variables such as LBPH, LCapPene and PerGG are skewed data and a SemVeinv variable would be considered a catergorical variable.

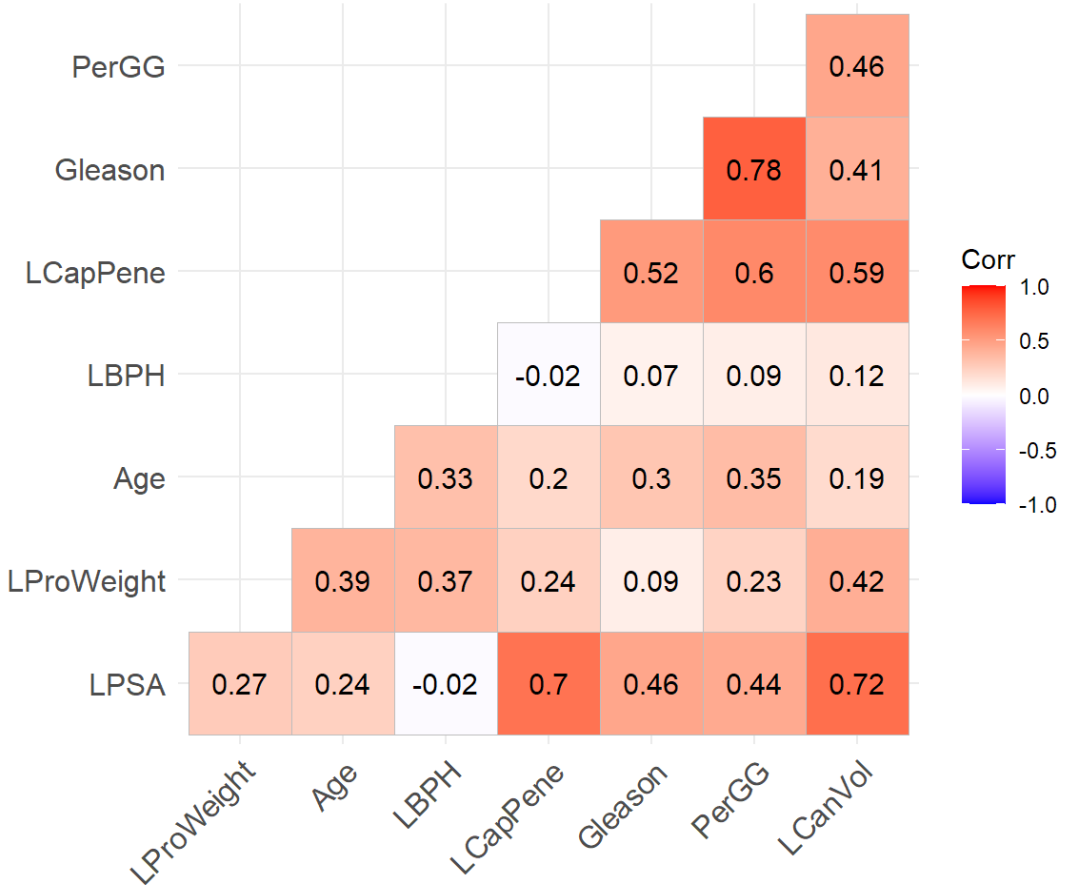
Summary statistic

LPSA	LProWeight	Age	LBPH	SemVeInv	LCapPene	Gleason	PerGG	LCanVol
Min. :-1.3471	Min. :2.769	Min. :41.00	Min. :-1.38629	0:62	Min. :-1.3863	Min. :6.000	Min. : 0.00	Min. :-0.4308
1st Qu.: 0.5349	1st Qu.:3.389	1st Qu.:60.00	1st Qu.:-1.38629	1:18	1st Qu.:-1.3863	1st Qu.:6.000	1st Qu.: 0.00	1st Qu.: 1.7133
Median : 1.4528	Median :3.628	Median :64.00	Median :-0.11376	NA	Median :-0.6982	Median :7.000	Median : 12.50	Median : 2.5688
Mean : 1.3365	Mean :3.623	Mean :63.14	Mean : 0.01833	NA	Mean :-0.1614	Mean :6.763	Mean : 23.75	Mean : 2.4467
3rd Qu.: 2.1045	3rd Qu.:3.866	3rd Qu.:68.00	3rd Qu.: 1.47930	NA	3rd Qu.: 0.9029	3rd Qu.:7.000	3rd Qu.: 40.00	3rd Qu.: 3.0421
Max. : 3.8210	Max. :4.718	Max. :79.00	Max. : 2.30757	NA	Max. : 2.9042	Max. :9.000	Max. :100.00	Max. : 5.5829



The correlation plot indicates that the predictor variable has strong linear associations with CapPene and LCanVol. In addition, emVelnv and CapPene and Gleason and PerGG have strong linear ssociations, which may cause a multicollinearity issue.





Simple linear regression is used to discover beta coefficients for each model. This is a univariate approach that helps to visualize any linear relationship existing between the reponse and the prector variables, before utilizing the mutiple linear model.

Coefficient of LProWeight

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.8358	1.2784	-1.44	0.1550
LProWeight	0.8756	0.3511	2.49	0.0147

Coefficient of Age

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.0115	1.0850	-0.93	0.3541
Age	0.0372	0.0171	2.18	0.0323

Coefficient of LBPH

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3368	0.1338	9.99	0.0000
LBPH	-0.0198	0.0941	-0.21	0.8340

Coefficient of SemVeInv

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9873	0.1271	7.77	0.0000
SemVeInv1	1.5518	0.2679	5.79	0.0000

Coefficient of LCapPene

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.4315	0.0966	14.82	0.0000
LCapPene	0.5887	0.0686	8.59	0.0000

Coefficient of Gleason

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.4660	1.0617	-3.26	0.0016
Gleason	0.7102	0.1560	4.55	0.0000

Coefficient of PerGG

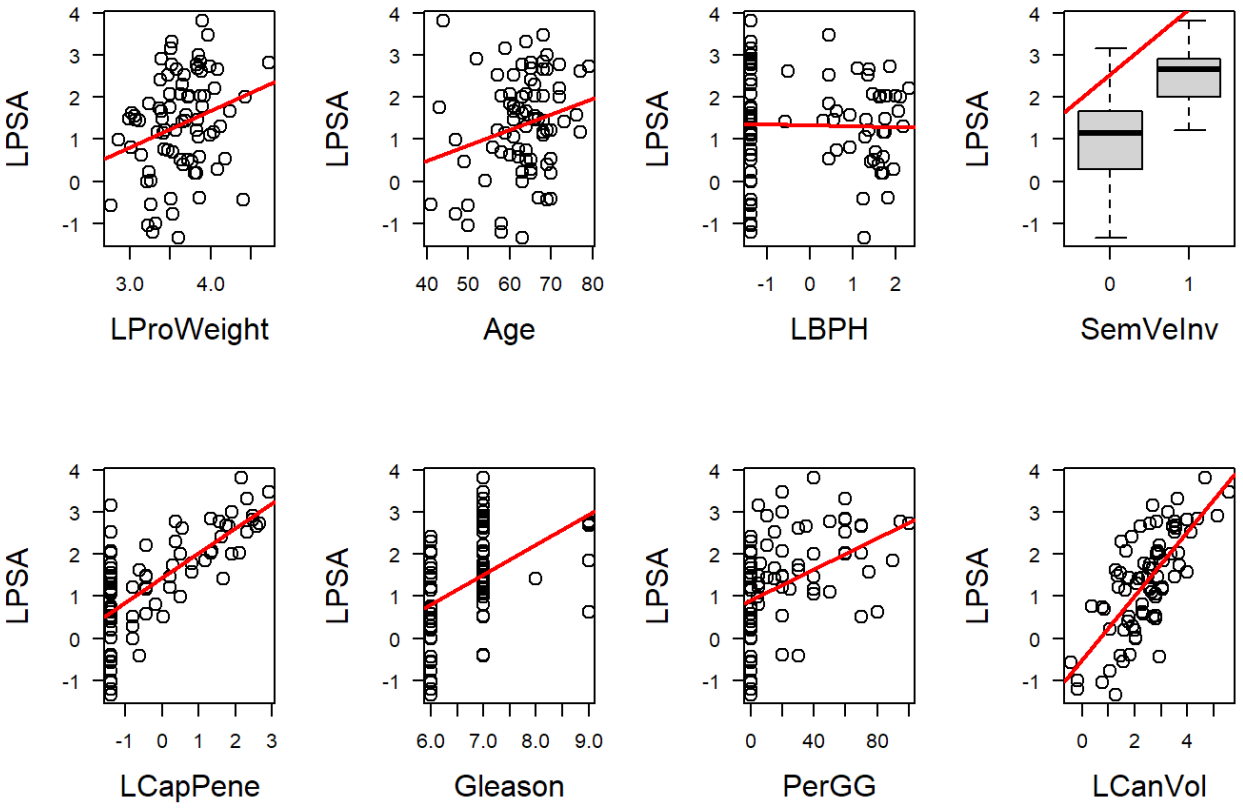
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.8953	0.1574	5.69	0.0000
PerGG	0.0186	0.0043	4.34	0.0000

Coefficient of LCanVol

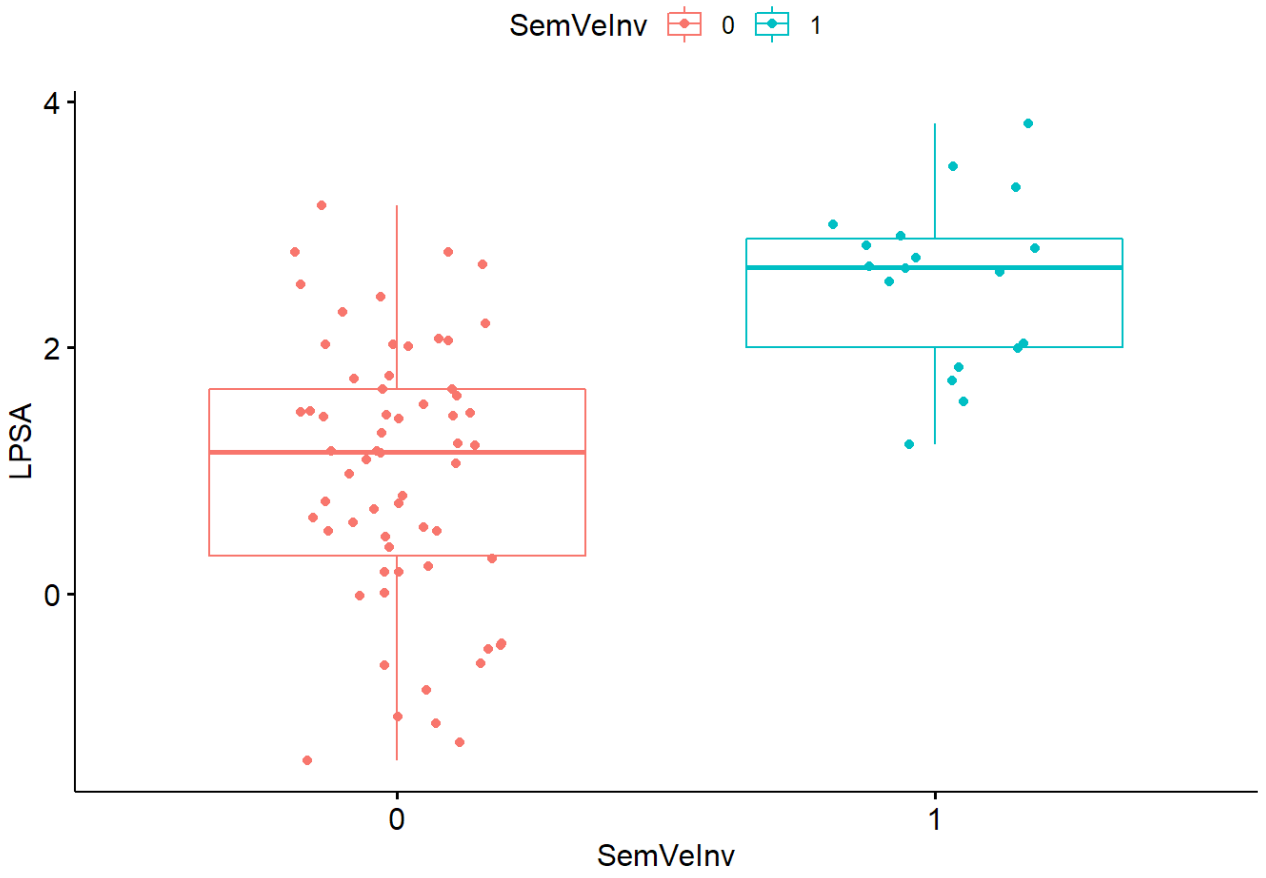
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.5155	0.2196	-2.35	0.0214
LCanVol	0.7569	0.0814	9.29	0.0000

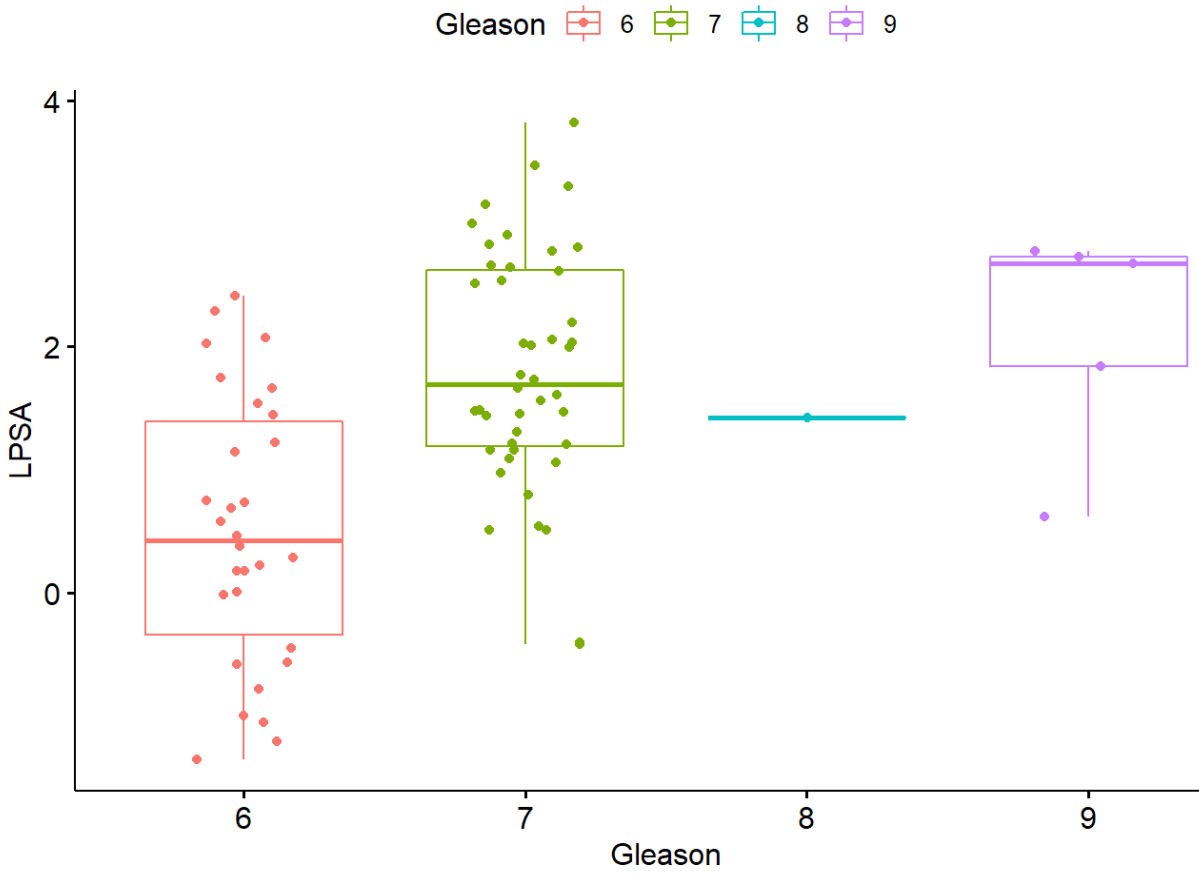
Coefficient of LProWeight

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.8358	1.2784	-1.44	0.1550
LProWeight	0.8756	0.3511	2.49	0.0147



Box plots for SemVeInv and Gleason variables.





Define the model

Model selection

Automatic variable selection is employed to identify the best parsimonious model which should be considered. Three criterion are considered for the model selection. They consist of $C_p, BIC, adjR^2$.

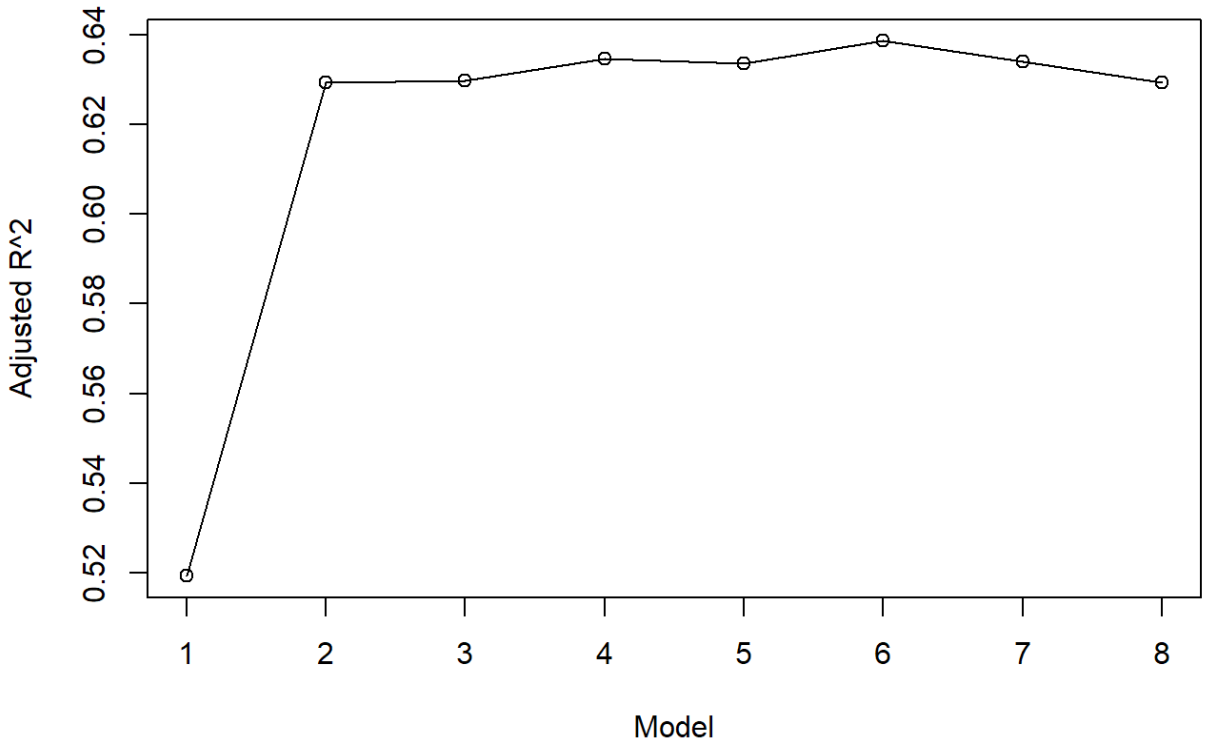
(Intercept)	LProWeight	Age	LBPH	SemVeInv	1	LCap	Pene	Gleason	PerGG	LCanVol	rss	adjr2	cp	bic
1	0	0	0	0	0	0	0				153.0060	0.5194	25.1157	-50.8704
1	0	0	0	0	1	0	0				140.3541	0.6294	2.9805	-68.3054
1	0	0	1	0	1	0	0				139.7863	0.6298	3.8974	-65.0569
1	0	1	1	0	1	0	0				138.7436	0.6347	3.9083	-62.7994
1	0	1	0	0	1	1	1				138.3403	0.6336	5.1390	-59.2545
1	0	1	1	0	1	1	1				137.3090	0.6385	5.1717	-57.0538
1	1	1	1	0	1	1	1				137.2635	0.6340	7.0848	-52.7696
1	1	1	1	1	1	1	1				137.2190	0.6293	9.0000	-48.4830

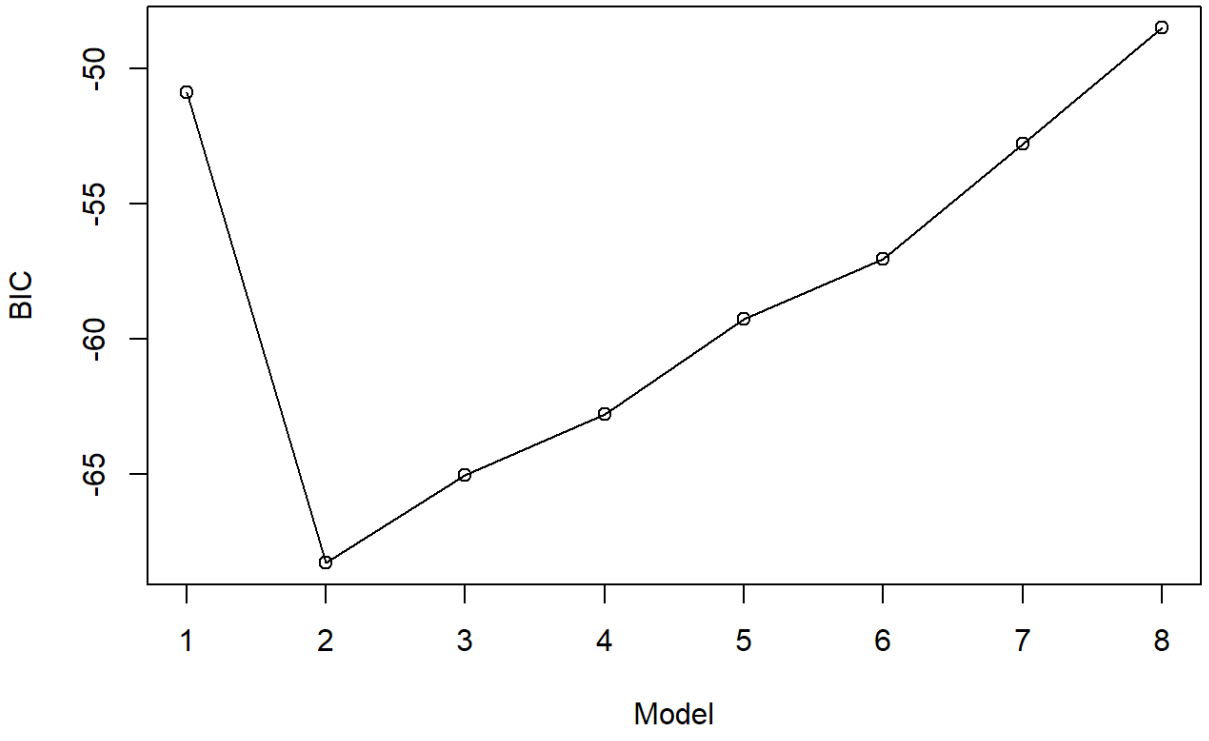
The plots indicate that Model 2 & 3 would be great for the predicting the response variable. And when Anova analysis is performed, the other variable such as Age, LBPH, SemVeInv, LCanVol and Gleason are not statisticall significant. Therefore, the final model is below.

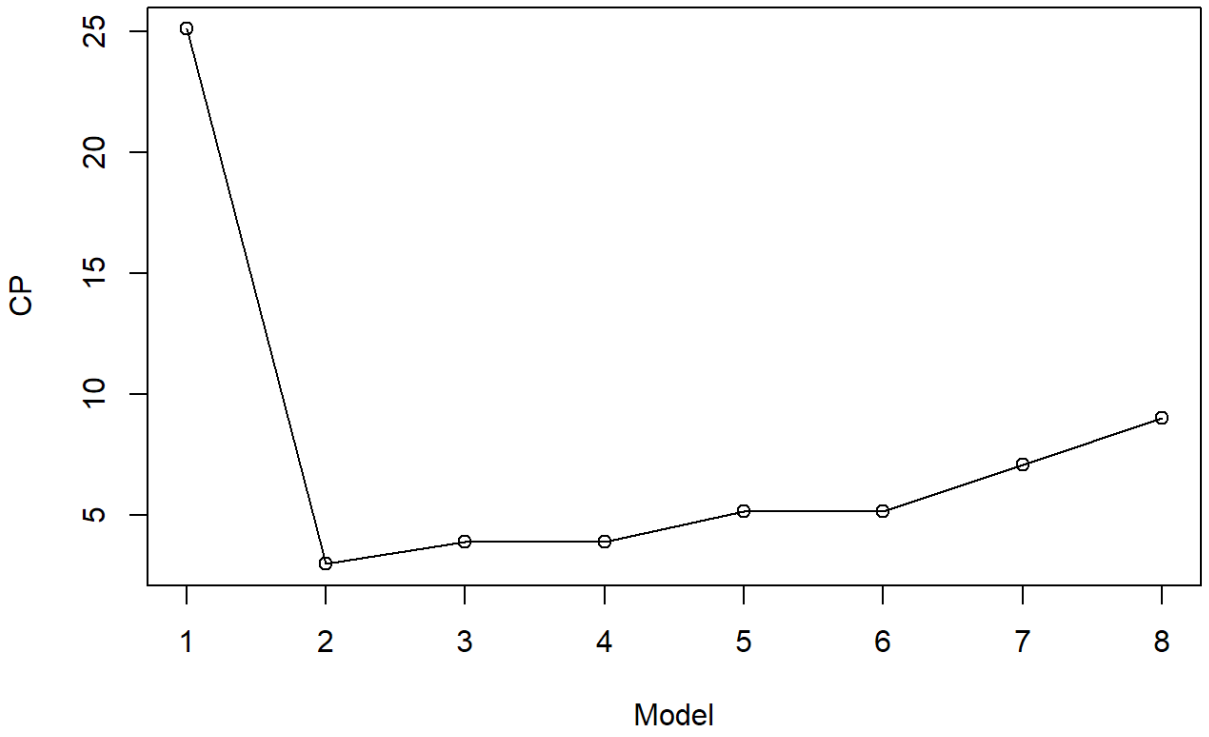
Y_i LPSA: Log PSA level
X_i1 LBPH: Log of the amount of benign prostatic hyperplasia
X_i2 SemVeInv: Seminal vesicle invasion

This is the final model is considered.

$$\hat{Y} = 0.16157 + 0.35051X_1 + 0.50331X_2 + Error$$


Adjusted R^2
0.52
0.54
0.56
0.58
0.60
0.62
0.64
1
2
3
4
5
6
7
8
Model


BIC
-50
-55
-60
-65
1
2
3
4
5
6
7
8
Model



Anova analysis is performed to display statistically significant variables using F-test. As it displays, Age, LBPH, SemVeInv, LProweight, Gleason and PerGG variables don't have statistically significant.

```
## Analysis of Variance Table
##
## Model 1: LPSA ~ LCapPene + LCanVol
## Model 2: LPSA ~ LProWeight + Age + LBPH + SemVeInv + LCapPene + Gleason +
##     PerGG + LCanVol
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     77 40.354
## 2     71 37.219  6     3.135 0.9967 0.4344
```

```
## Anova Table (Type III tests)
##
## Response: LPSA
##             Sum Sq Df F value    Pr(>F)
## (Intercept)  0.244  1  0.4648    0.4974
## LCapPene    12.652  1 24.1414 4.925e-06 ***
## LCanVol     17.064  1 32.5603 2.050e-07 ***
## Residuals   40.354 77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

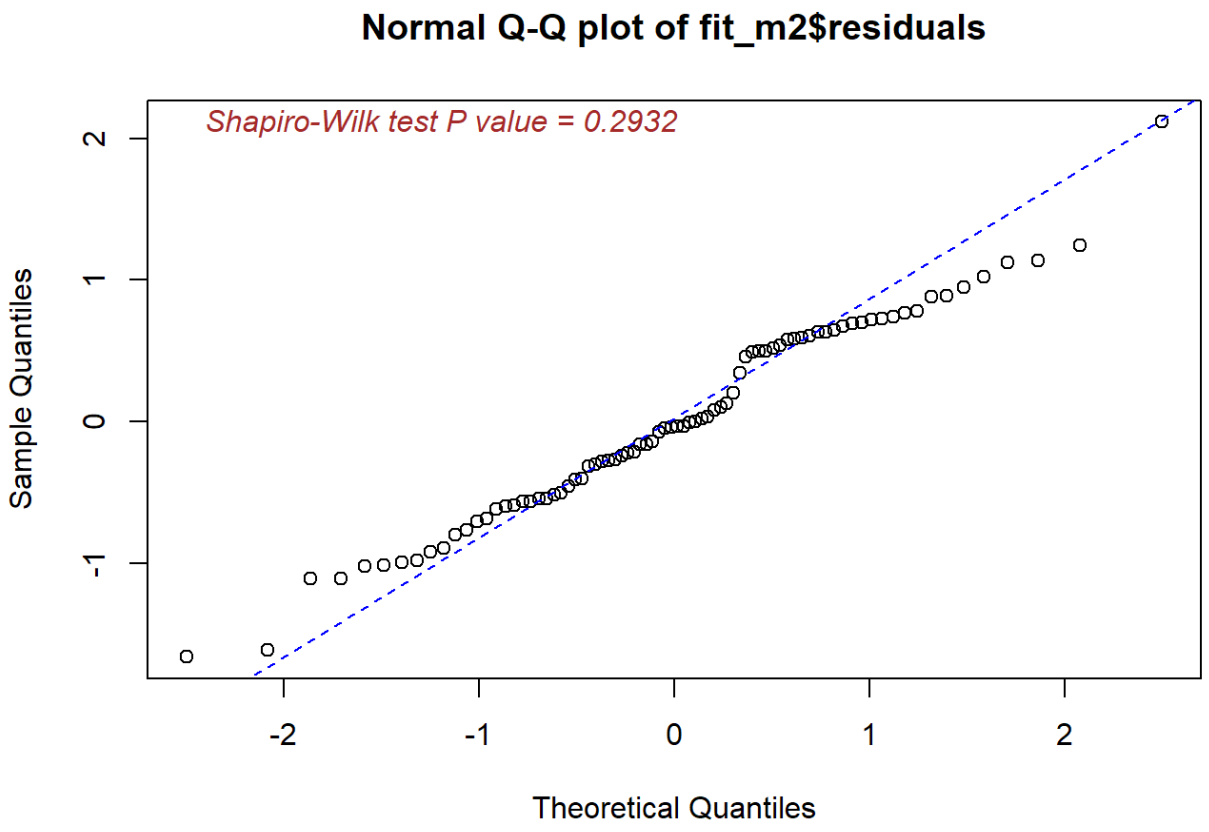
```
##
## Call:
## lm(formula = LPSA ~ LCapPene + LCanVol, data = train_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66041 -0.54263 -0.03626  0.59419  2.11661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.16157    0.23699   0.682    0.497
## LCapPene     0.35051    0.07134   4.913 4.93e-06 ***
## LCanVol      0.50331    0.08820   5.706 2.05e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7239 on 77 degrees of freedom
## Multiple R-squared:  0.6387, Adjusted R-squared:  0.6294
## F-statistic: 68.07 on 2 and 77 DF,  p-value: < 2.2e-16
```

Model diagnosis

There are not any abnormality found in the plots. There doesn't appear to be multicollinearity and other violation of linear regression assumptions.

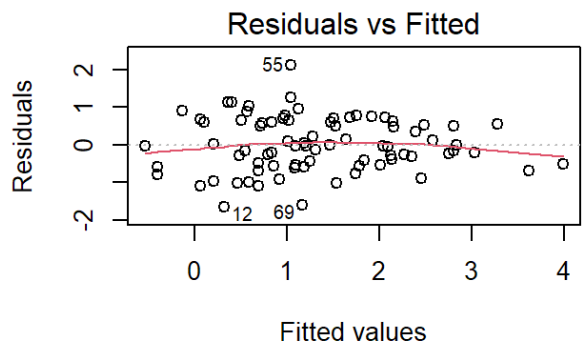
1. Normality and VIF index

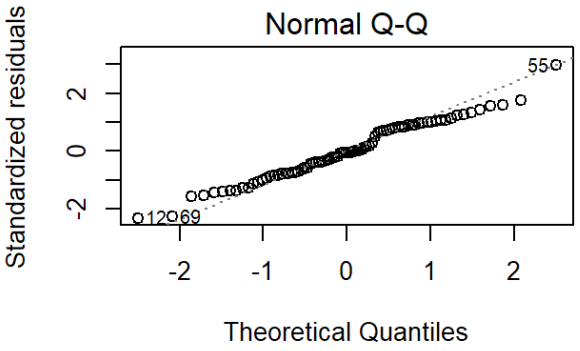
```
## LCapPene  LCanVol
## 1.520842 1.520842
```

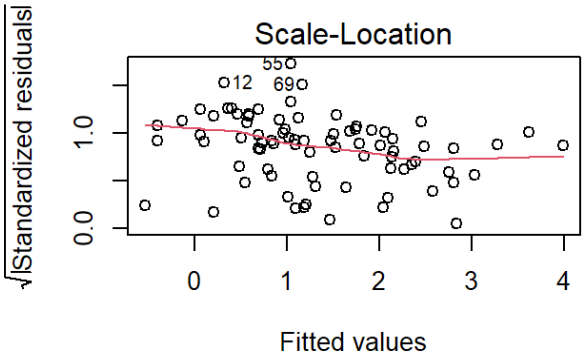


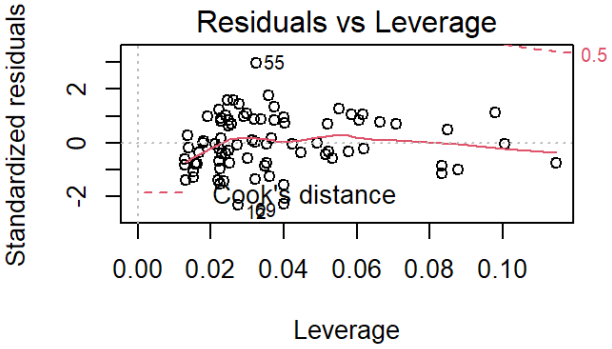
2. Plots for residual analysis

Nothing appears extreme or abnormal.









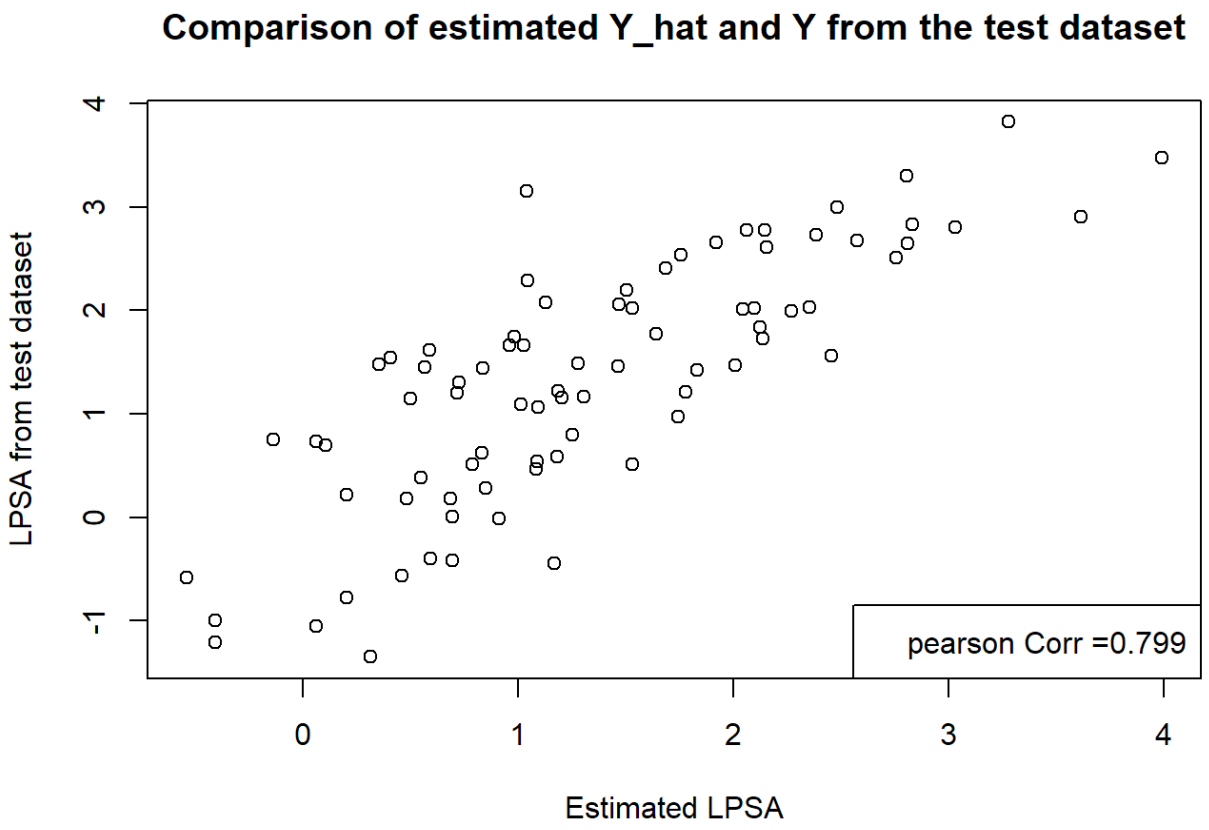
Fit the model to the test dataset for the validation

MSPR is used to evaluate the prediction.

$$MSPR = \sum (Y_i - \hat{Y}_i)^2 / n^*$$

n^* = the number of caeses in the validation dataset. The MSPR in this model is 1.264677e-29, which is quite small.

```
## [1] 1.264677e-29
```


Comparison of estimated Y_hat and Y from the test dataset
LPSA from test dataset
Estimated LPSA
pearson Corr =0.799
-1
0
1
2
3
4
0
1
2
3
4

Appendix

```
        knitr::opts_chunk$set(
            tidy = FALSE, # display code in tidy format (FALSE= as typed)
            echo = TRUE)
        rm(list = ls())
        # rm(list = ls(all.names = TRUE)) # removes hidden objects also
        pacman::p_load('tidyverse', 'ggpubr', 'epiDisplay', 'kableExtra', 'xtable', 'ggcorrplot', 'ca
r', 'leaps', 'caret')
        set.seed(0)

setwd("D:/KU/stat840/final-project")
df <- read.table("pros.dat.txt")
col <- colnames(df)
lcolvol <- col[1]
lpsa <- col[9]
colnames(df)[1] <- lpsa
colnames(df)[9] <- lcolvol

df$SemVeInv <- as.factor(df$SemVeInv)
train_idx <- createDataPartition(df$LPSA, p = 0.8, list = F)
train_df <- df[train_idx, ]
test_df <- df[train_idx,]
train_df %>% summary %>% kbl(caption='Summary statistic') %>% kable_classic(full_width=F)

par(mfrow = c(3, 3))
for (i in 1:ncol(train_df)) {
        if (is.numeric(train_df[, i]) == TRUE) {
                hist(
                        train_df[, i],
                        xlab = colnames(train_df)[i],
                        main = paste("Histogram of", colnames(train_df)[i]),
                        breaks = 20,
                        col = 'lightblue'
                )
        }
}
train_df %>% select_if(is.numeric) %>% psych::pairs.panels()
train_df %>% select_if(is.numeric) %>% cor %>% ggcorrplot(type='lower', lab=TRUE)

fit_LProWeight <- lm(LPSA ~ LProWeight, data=train_df)
fit_Age <- lm(LPSA ~ Age, data=train_df)
fit_LBPH <- lm(LPSA ~ LBPH, data=train_df)
fit_SemVeInv <- lm(LPSA ~ SemVeInv, data=train_df)
fit_LCapPene <- lm(LPSA ~ LCapPene, data=train_df)
fit_Gleason <- lm(LPSA ~ Gleason, data=train_df)
fit_PerGG <- lm(LPSA ~ PerGG, data=train_df)
fit_LCanVol <- lm(LPSA ~ LCanVol, data=train_df)
print(xtable(summary(fit_LProWeight), caption = "Coefficient of LProWeight"), type='html', comment =
F, caption.placement = 'top')

print(xtable(summary(fit_Age), caption = "Coefficient of Age"), type='html', comment = F, caption.pla
cement = 'top')

print(xtable(summary(fit_LBPH), caption = "Coefficient of LBPH"), type='html', comment = F, caption.p
lacement = 'top')

print(xtable(summary(fit_SemVeInv), caption = "Coefficient of SemVeInv"), type='html', comment = F, c
```

```
aption.placement = 'top')

print(xtable(summary(fit_LCapPene), caption = "Coefficient of LCapPene"), type='html', comment = F, c
aption.placement = 'top')

print(xtable(summary(fit_Gleason), caption = "Coefficient of Gleason"), type='html', comment = F, cap
tion.placement = 'top')

print(xtable(summary(fit_PerGG), caption = "Coefficient of PerGG"), type='html', comment = F, captio
n.placement = 'top')

print(xtable(summary(fit_LCanVol), caption = "Coefficient of LCanVol"), type='html', comment = F, cap
tion.placement = 'top')

print(xtable(summary(fit_LProWeight), caption = "Coefficient of LProWeight"), type='html', comment =
F, caption.placement = 'top')


par(mfrow=c(2,4))
attach(train_df)

plot(LPSA ~ LProWeight, cex = 1.5, cex.lab=1.5, las=1, cex.main=1.5)
abline(fit_LProWeight, lwd = 2, col='red')

plot(LPSA ~ Age, cex = 1.5, cex.lab=1.5, las=1, cex.main=1.5)
abline(fit_Age, lwd = 2, col='red')

plot(LPSA ~ LBPH, cex = 1.5, cex.lab=1.5, las=1, cex.main=1.5)
abline(fit_LBPH, lwd = 2, col='red')

plot(LPSA ~ SemVeInv, cex = 1.5, cex.lab=1.5, las=1, cex.main=1.5)
abline(fit_SemVeInv, lwd = 2, col='red')

plot(LPSA ~ LCapPene, cex = 1.5, cex.lab=1.5, las=1, cex.main=1.5)
abline(fit_LCapPene, lwd = 2, col='red')

plot(LPSA ~ Gleason, cex = 1.5, cex.lab=1.5, las=1, cex.main=1.5)
abline(fit_Gleason, lwd = 2, col='red')

plot(LPSA ~ PerGG, cex = 1.5, cex.lab=1.5, las=1, cex.main=1.5)
abline(fit_PerGG, lwd = 2, col='red')

plot(LPSA ~ LCanVol, cex = 1.5, cex.lab=1.5, las=1, cex.main=1.5)
abline(fit_LCanVol, lwd = 2, col='red')

detach(train_df)
ggboxplot(train_df, x='SemVeInv', y="LPSA", color = 'SemVeInv', add='jitter')
ggboxplot(train_df, x='Gleason', y="LPSA", color = 'Gleason', add='jitter')
fit <- regsubsets(LPSA ~ LProWeight + Age + LBPH + SemVeInv + LCapPene + Gleason + PerGG + LCanVol, d
ata=train_df)
smm_fit <- summary(fit)
kableExtra::kable(with(smm_fit, cbind(which, rss, adjr2, cp,bic)), digits = 4)
fit_subset_criteria <- data.frame(variable=round(smm_fit$which), adj_r2=smm_fit$adjr2, BIC=smm_fit$bi
c, RSS=smm_fit$rss, CP=smm_fit$cp)


plot(fit_subset_criteria$adj_r2 ~ rownames(fit_subset_criteria), xlab='Model', ylab='Adjusted R^2')
```

```
lines(fit_subset_criteria$adj_r2 ~ rownames(fit_subset_criteria))


plot(fit_subset_criteria$BIC ~ rownames(fit_subset_criteria), xlab='Model', ylab='BIC')
lines(fit_subset_criteria$BIC ~ rownames(fit_subset_criteria))

plot(fit_subset_criteria$CP ~ rownames(fit_subset_criteria), xlab='Model', ylab='CP')
lines(fit_subset_criteria$CP ~ rownames(fit_subset_criteria))


fit_m2 <- lm(LPSA ~ LCapPene + LCanVol, data=train_df)
fit_full <- lm(LPSA ~., data=train_df)



anva_fit_m2_fit_full <- anova(fit_m2, fit_full)
anva_fit_m2_type_3 <- Anova(fit_m2, type=3)



anva_fit_m2_fit_full
anva_fit_m2_type_3


summary(fit_m2)


## VIF
car::vif(fit_m2) ## VIF
shapiro.qqnorm(fit_m2$residuals)
par(mfrow=c(2,2))
plot(fit_m2)
est <- predict(fit_m2, newdata = test_df)
## MPSE
(sum(est - test_df$LPSA)^2)/length(test_df$LPSA) ## MPSE

plot(test_df$LPSA ~ est, xlab="Estimated LPSA", ylab = "LPSA from test dataset", main='Comparison of
 estimated Y_hat and Y from the test dataset')
legend(x='bottomright', legend="pearson Corr =0.799")
```