Considering the prostate cancer dataset consisting of data on 97 men with advanced prostate cancer. We would like to understand how PSA level is related to the other predictors in the dataset.

#read prostate\_cancer data

**cancer <- read.table("prostate\_cancer.csv", sep=",", header=T)**

**str(cancer)**

'data.frame': 97 obs. of 9 variables:

$ subject : int 1 2 3 4 5 6 7 8 9 10 ...

$ psa : num 0.651 0.852 0.852 0.852 1.448 ...

$ cancervol: num 0.56 0.372 0.601 0.301 2.117 ...

$ weight : num 16 27.7 14.7 26.6 30.9 ...

$ age : int 50 58 74 58 62 50 64 58 47 63 ...

$ benpros : num 0 0 0 0 0 ...

$ vesinv : int 0 0 0 0 0 0 0 0 0 0 ...

$ capspen : num 0 0 0 0 0 0 0 0 0 0 ...

$ gleason : int 6 7 7 6 6 6 6 6 7 6 ...

table(cancer$cancervol) table(cancer$weight) table(cancer$age) table(cancer$benpros) table(cancer$vesinv) table(cancer$capspen) table(cancer$gleason)

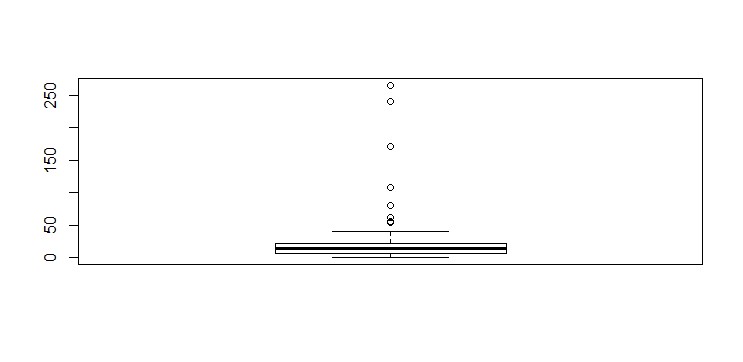
**boxplot(cancer$psa)**

#mean – 23.730

#median – 13.330

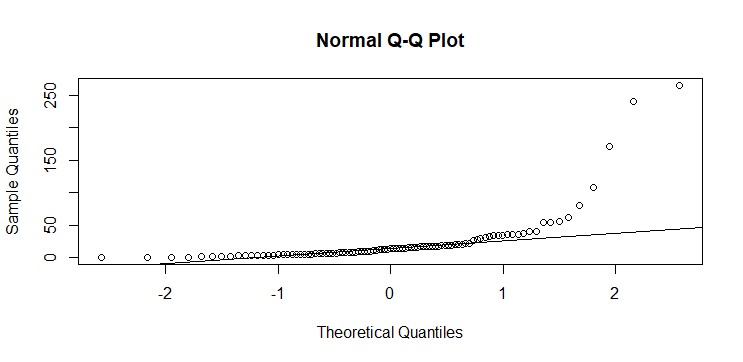
#Also it is right skewed

Outliers are found in the boxplot.



**qqnorm(cancer$psa)**

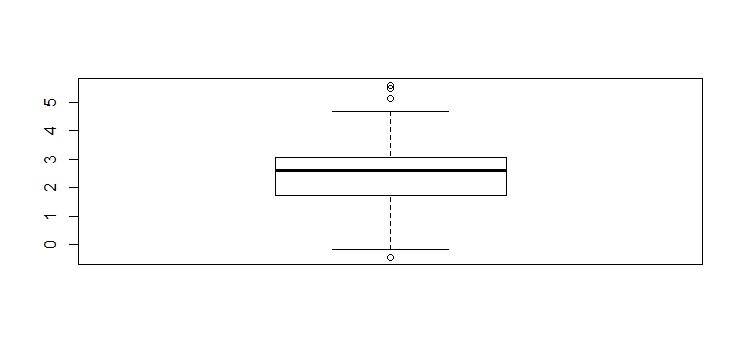
**qqline(cancer$psa)**



# Not approximately normal.

**boxplot(log(cancer$psa))**

For the model assumption to hold, we need to apply transformation in response variable. Here, natural log transformation is used.



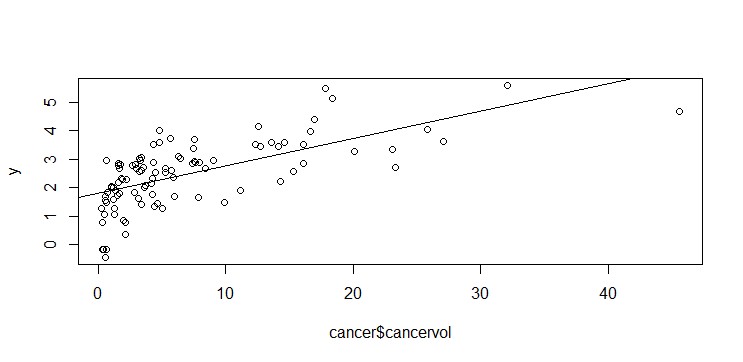
**y<-log(cancer$psa)**

Relationship between response variable **psa** and other variables are seen below:

plot(cancer$cancervol,y)

fit1 <- lm(y ~ cancervol, data = cancer)

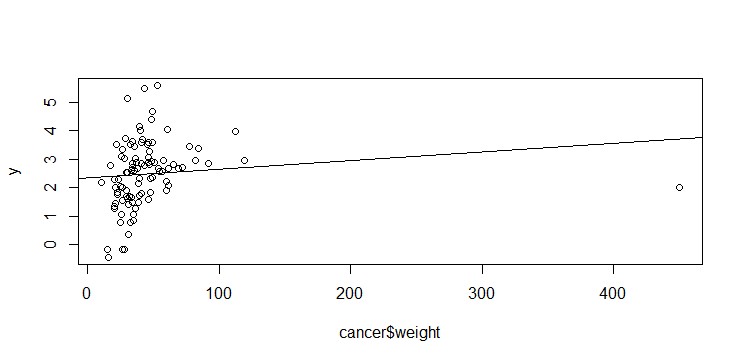
abline(fit1)



plot(cancer$weight,y)

fit2 <- lm(y ~ weight, data = cancer)

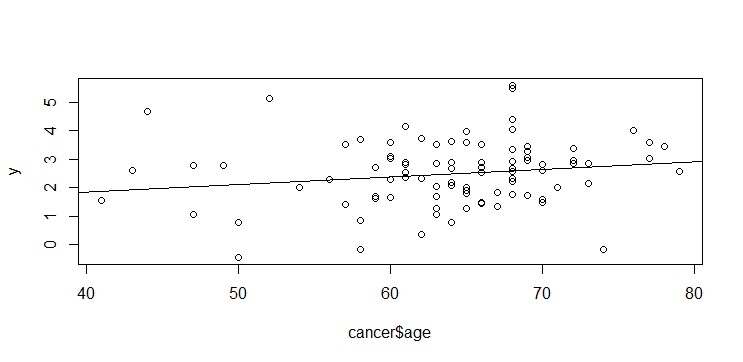
abline(fit2)



plot(cancer$age,y)

fit3 <- lm(y ~ age, data = cancer)

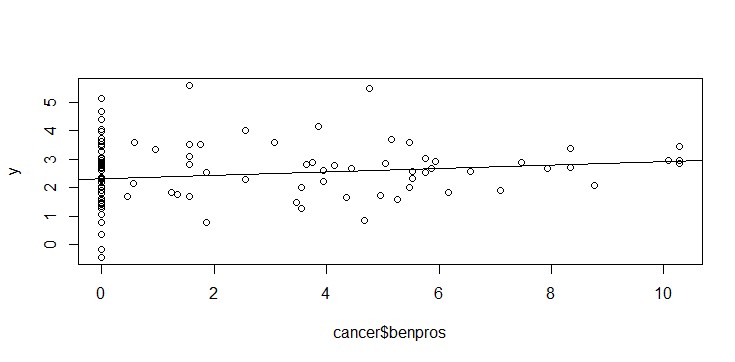
abline(fit3)



plot(cancer$benpros,y)

fit4 <- lm(y ~ benpros, data = cancer)

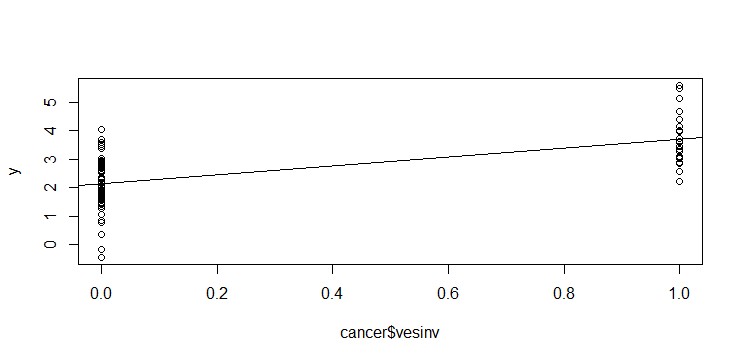
abline(fit4)



plot(cancer$vesinv,y)

fit5 <- lm(y ~ vesinv, data = cancer)

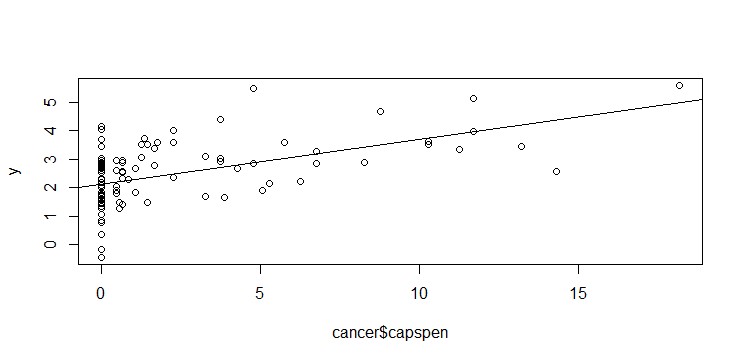
abline(fit5)



plot(cancer$capspen,y)

fit6 <- lm(y ~ capspen, data = cancer)

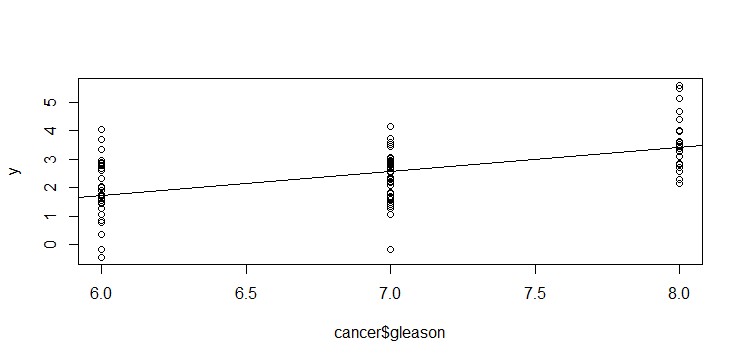
abline(fit6)



plot(cancer$gleason,y)

fit7 <- lm(y ~ gleason, data = cancer)

abline(fit7)



From the plots we can infer that building the linear model can be started with two variables

, vesinv and gleason.

fit8 <- lm(y ~ vesinv+gleason, data = cancer)

summary(fit8)

Call:

lm(formula = y ~ vesinv + gleason, data = cancer)

Residuals:

Min 1Q Median 3Q Max

-2.46103 -0.56567 0.01435 0.58912 2.29527

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.6621 0.9091 -1.828 0.0707 . vesinv 1.1448 0.2405 4.759 7.01e-06 \*\*\* gleason 0.5661 0.1346 4.205 5.95e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8815 on 94 degrees of freedom

Multiple R-squared: 0.4283, Adjusted R-squared: 0.4161

F-statistic: 35.21 on 2 and 94 DF, p-value: 3.861e-12

**Slope of the predictors is 0 -- H0 and atleast one predictor is not 0 -- H1**

(If p-value less than 0.05 reject null hypothesis, if greater than 0.05 then accept null hypoth esis)

Now adding another variable,

update(fit8,. ~ . +capspen) (or)

fit9 <- lm(y ~ vesinv+gleason+capspen, data = cancer)

summary(fit9)

Call:

lm(formula = y ~ vesinv + gleason + capspen, data = cancer)

Residuals:

Min 1Q Median 3Q Max

-2.40890 -0.49416 0.04112 0.56127 2.30142

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) -1.39234 0.92972 -1.498 0.13762 vesinv 0.91167 0.30015 3.037 0.00310 \*\* gleason 0.52015 0.13881 3.747 0.00031 \*\*\* capspen 0.04316 0.03345 1.290 0.20017

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8784 on 93 degrees of freedom

Multiple R-squared: 0.4384, Adjusted R-squared: 0.4202

F-statistic: 24.2 on 3 and 93 DF, p-value: 1.164e-11

> anova(fit8,fit9)

Analysis of Variance Table

Model 1: y ~ vesinv + gleason

Model 2: y ~ vesinv + gleason + capspen

Res.Df RSS Df Sum of Sq F Pr(>F)

1 94 73.044

2 93 71.760 1 1.2845 1.6647 0.2002

P-value 0.2002 > 0.05 accept null hypothesis.Capspen is not a significant

> fit10 <- lm(y ~ vesinv+gleason+benpros, data = cancer)

> summary(fit10)

Call:

lm(formula = y ~ vesinv + gleason + benpros, data = cancer)

Residuals:

Min 1Q Median 3Q Max

-2.24378 -0.50263 0.09248 0.50253 2.48252

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) -1.66930 0.88178 -1.893 0.06145 . vesinv 1.23478 0.23579 5.237 1.01e-06 \*\*\* gleason 0.53613 0.13108 4.090 9.16e-05 \*\*\* benpros 0.07656 0.02911 2.630 0.00999 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.855 on 93 degrees of freedom

Multiple R-squared: 0.4679, Adjusted R-squared: 0.4507

F-statistic: 27.26 on 3 and 93 DF, p-value: 9.747e-13

> anova(fit8,fit10)

Analysis of Variance Table

Model 1: y ~ vesinv + gleason

Model 2: y ~ vesinv + gleason + benpros

Res.Df RSS Df Sum of Sq F Pr(>F)

1 94 73.044

2 93 67.987 1 5.0575 6.9182 0.009986 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

P-value 0.009986 < 0.05 reject null hypothesis. benpros is a significant f actor

> fit11 <- lm(y ~ vesinv+gleason+cancervol, data = cancer)

> summary(fit11)

Call:

lm(formula = y ~ vesinv + gleason + cancervol, data = cancer)

Residuals:

Min 1Q Median 3Q Max

-2.16928 -0.44558 0.08431 0.60719 1.64082

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) -0.72120 0.85749 -0.841 0.4025 vesinv 0.62117 0.24962 2.488 0.0146 \* gleason 0.38491 0.12966 2.969 0.0038 \*\*

F-statistic: 34.64 on 3 and 93 DF, p-value: 4.022e-15

> anova(fit8,fit11)

Analysis of Variance Table

Model 1: y ~ vesinv + gleason

Model 2: y ~ vesinv + gleason + cancervol

Res.Df RSS Df Sum of Sq F Pr(>F)

1 94 73.044

2 93 60.340 1 12.704 19.58 2.62e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

P-value 2.62e-05 < 0.05 reject null hypothesis. cancervol is a significant factor

> fit12 <- lm(y ~ vesinv+gleason+weight, data = cancer)

> summary(fit12)

Call:

lm(formula = y ~ vesinv + gleason + weight, data = cancer)

Residuals:

Min 1Q Median 3Q Max

-2.40331 -0.54588 0.04837 0.58015 2.25028

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.852463 0.906953 -2.043 0.0439 \* vesinv 1.141209 0.238141 4.792 6.23e-06 \*\*\* gleason 0.571961 0.133336 4.290 4.38e-05 \*\*\* weight 0.003321 0.001949 1.704 0.0918 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8727 on 93 degrees of freedom

Multiple R-squared: 0.4456, Adjusted R-squared: 0.4277

F-statistic: 24.92 on 3 and 93 DF, p-value: 6.414e-12

> anova(fit8,fit12)

Analysis of Variance Table

Model 1: y ~ vesinv + gleason

Model 2: y ~ vesinv + gleason + weight

Res.Df RSS Df Sum of Sq F Pr(>F)

1 94 73.044

2 93 70.834 1 2.2108 2.9026 0.09178 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

P-value 0.09178 > 0.05 accept null hypothesis. Weight is not a significant

> fit13 <- lm(y ~ vesinv+gleason+age, data = cancer)

> summary(fit13)

Call:

lm(formula = y ~ vesinv + gleason + age, data = cancer)

Residuals:

Min 1Q Median 3Q Max

-2.52551 -0.52693 -0.00436 0.59444 2.25562

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.978374 1.097209 -1.803 0.074612 . vesinv 1.141813 0.241537 4.727 8.07e-06 \*\*\* gleason 0.552141 0.137820 4.006 0.000124 \*\*\* age 0.006469 0.012457 0.519 0.604758

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.885 on 93 degrees of freedom

Multiple R-squared: 0.43, Adjusted R-squared: 0.4116

F-statistic: 23.38 on 3 and 93 DF, p-value: 2.301e-11

> anova(fit8,fit13)

Analysis of Variance Table

Model 1: y ~ vesinv + gleason

Model 2: y ~ vesinv + gleason + age

Res.Df RSS Df Sum of Sq F Pr(>F)

1 94 73.044

2 93 72.833 1 0.21123 0.2697 0.6048

P-value 0.6048 > 0.05 accept null hypothesis. age is not a significant fac tor.

> finalfit <- lm(y ~ vesinv+gleason+benpros+cancervol

, data=cancer)

> summary(finalfit)

Call:

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

Residuals:

Min 1Q Median 3Q Max

-1.88531 -0.50276 0.09885 0.53687 1.56621

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) -0.65013 0.80999 -0.803 0.424253 vesinv 0.68421 0.23640 2.894 0.004746 \*\* gleason 0.33376 0.12331 2.707 0.008100 \*\* benpros 0.09136 0.02606 3.506 0.000705 \*\*\*

F-statistic: 32.21 on 4 and 92 DF, p-value: < 2.2e-16

> anova(fit8,finalfit)

Analysis of Variance Table

Model 1: y ~ vesinv + gleason

Model 2: y ~ vesinv + gleason + benpros + cancervol

Res.Df RSS Df Sum of Sq F Pr(>F)

1 94 73.044

2 92 53.229 2 19.816 17.124 4.762e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

P-value 4.762e-07 < 0.05 reject null hypothesis. benpros and cancervol are significant factors.

**β0=** -0.65013 **β1=** 0.68421 **β2=** 0.33376 **β3=** 0.09136 **β4** 0.06488

**Forward selection based on AIC**

> fitforward <- step(lm(y ~ 1, data = cancer),scope = list(upper = ~ vesin v + gleason + capspen + benpros + cancervol + weight + age),direction = "f orward")

Start: AIC=28.72 y ~ 1

Df Sum of Sq RSS AIC

+ cancervol 1 55.164 72.605 -24.0986

+ vesinv 1 40.984 86.785 -6.7944

+ gleason 1 37.122 90.647 -2.5707

+ capspen 1 34.286 93.482 0.4169

+ age 1 3.688 124.080 27.8831

+ benpros 1 3.166 124.603 28.2911

<none> 127.769 28.7246

+ weight 1 1.893 125.876 29.2767

Step: AIC=-24.1 y ~ cancervol

Df Sum of Sq RSS AIC

+ gleason 1 8.2468 64.358 -33.794

+ benpros 1 7.8034 64.802 -33.128

+ vesinv 1 6.5468 66.058 -31.265

+ age 1 2.6615 69.944 -25.721

+ weight 1 1.7901 70.815 -24.520

<none> 72.605 -24.099

+ capspen 1 0.9673 71.638 -23.400

Step: AIC=-33.79

y ~ cancervol + gleason

Df Sum of Sq RSS AIC

+ benpros 1 6.2827 58.075 -41.758

+ vesinv 1 4.0178 60.340 -38.047

+ weight 1 2.0334 62.325 -34.908

<none> 64.358 -33.794

+ age 1 0.9611 63.397 -33.253

+ capspen 1 0.1685 64.190 -32.048

Step: AIC=-41.76

y ~ cancervol + gleason + benpros

Df Sum of Sq RSS AIC

+ vesinv 1 4.8466 53.229 -48.211

<none> 58.075 -41.758

+ weight 1 0.4006 57.675 -40.429

+ capspen 1 0.1863 57.889 -40.069

+ age 1 0.0059 58.070 -39.768

Step: AIC=-48.21

y ~ cancervol + gleason + benpros + vesinv

Df Sum of Sq RSS AIC

<none> 53.229 -48.211

+ capspen 1 0.39230 52.837 -46.928

+ weight 1 0.33060 52.898 -46.815

+ age 1 0.02497 53.204 -46.256

Backward selection based on AIC

> fitbackward <- step(lm(y ~ vesinv + gleason + capspen + benpros + cancer vol + weight + age, data = cancer), scope = list(lower = ~1), direction =" backward")

Start: AIC=-43.59

y ~ vesinv + gleason + capspen + benpros + cancervol + weight +

age

Df Sum of Sq RSS AIC

- age 1 0.0336 52.510 -45.529

- weight 1 0.3383 52.815 -44.968

- capspen 1 0.3841 52.861 -44.884

<none> 52.477 -43.591

- gleason 1 4.6180 57.095 -37.410

- vesinv 1 5.0155 57.492 -36.737

- benpros 1 5.1469 57.624 -36.516

- cancervol 1 13.2994 65.776 -23.680

Step: AIC=-45.53

y ~ vesinv + gleason + capspen + benpros + cancervol + weight

Df Sum of Sq RSS AIC

- weight 1 0.3264 52.837 -46.928

- capspen 1 0.3881 52.898 -46.815

<none> 52.510 -45.529

- gleason 1 4.6365 57.147 -39.322

- vesinv 1 4.9820 57.492 -38.737

- benpros 1 5.4873 57.998 -37.888

- cancervol 1 13.4654 65.976 -25.386

Step: AIC=-46.93

y ~ vesinv + gleason + capspen + benpros + cancervol

Df Sum of Sq RSS AIC

- capspen 1 0.3923 53.229 -48.211

<none> 52.837 -46.928

- gleason 1 4.4852 57.322 -41.025

- vesinv 1 5.0526 57.889 -40.069

- benpros 1 7.2024 60.039 -36.532

- cancervol 1 13.7311 66.568 -26.520

Step: AIC=-48.21

y ~ vesinv + gleason + benpros + cancervol

Df Sum of Sq RSS AIC

<none> 53.229 -48.211

- gleason 1 4.2389 57.468 -42.778

- vesinv 1 4.8466 58.075 -41.758

- benpros 1 7.1115 60.340 -38.047

- cancervol 1 14.7580 67.987 -26.473

**Both forward and backward selection**

> fitboth <- step(lm(y ~ 1, data = cancer),scope = list(lower = ~1, upper

= ~ vesinv + gleason + capspen + benpros + cancervol + weight + age),direc tion = "both")

Start: AIC=28.72

y ~ 1

Df Sum of Sq RSS AIC

+ cancervol 1 55.164 72.605 -24.0986

+ vesinv 1 40.984 86.785 -6.7944

+ gleason 1 37.122 90.647 -2.5707

+ capspen 1 34.286 93.482 0.4169

+ age 1 3.688 124.080 27.8831

+ benpros 1 3.166 124.603 28.2911

<none> 127.769 28.7246

+ weight 1 1.893 125.876 29.2767

Step: AIC=-24.1 y ~ cancervol

Df Sum of Sq RSS AIC

+ gleason 1 8.247 64.358 -33.794

+ benpros 1 7.803 64.802 -33.128

+ vesinv 1 6.547 66.058 -31.265

+ age 1 2.662 69.944 -25.721

+ weight 1 1.790 70.815 -24.520

<none> 72.605 -24.099

+ capspen 1 0.967 71.638 -23.400

- cancervol 1 55.164 127.769 28.725

Step: AIC=-33.79

y ~ cancervol + gleason

Df Sum of Sq RSS AIC

+ benpros 1 6.2827 58.075 -41.758

+ vesinv 1 4.0178 60.340 -38.047

+ weight 1 2.0334 62.325 -34.908

<none> 64.358 -33.794

+ age 1 0.9611 63.397 -33.253

+ capspen 1 0.1685 64.190 -32.048

- gleason 1 8.2468 72.605 -24.099

- cancervol 1 26.2887 90.647 -2.571

Step: AIC=-41.76

y ~ cancervol + gleason + benpros

Df Sum of Sq RSS AIC

+ vesinv 1 4.8466 53.229 -48.211

<none> 58.075 -41.758

+ weight 1 0.4006 57.675 -40.429

+ capspen 1 0.1863 57.889 -40.069

+ age 1 0.0059 58.070 -39.768

- benpros 1 6.2827 64.358 -33.794

- gleason 1 6.7262 64.802 -33.128

- cancervol 1 29.9589 88.034 -3.407

Step: AIC=-48.21

y ~ cancervol + gleason + benpros + vesinv

Df Sum of Sq RSS AIC

<none> 53.229 -48.211

+ capspen 1 0.3923 52.837 -46.928

+ weight 1 0.3306 52.898 -46.815

+ age 1 0.0250 53.204 -46.256

- gleason 1 4.2389 57.468 -42.778

- vesinv 1 4.8466 58.075 -41.758

- benpros 1 7.1115 60.340 -38.047

- cancervol 1 14.7580 67.987 -26.473

Forward selection, backward selection picks the following model - benpros, vesinv, gleason, cancervol, Both forward and backward selection picks the following model - gleason, vesinv, benpros, cancervol

**Final fit is the reduced model.**

**vesinv, gleason, benpros, cancervol are the best predictors for this model. The prediction results of the final fit are better than any other predictors.**

> fullfit <- lm(y ~ vesinv+gleason+cancervol+benpros+capspen+weight+age, d ata = cancer)

> summary(fullfit)

Call:

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

weight + age, data = cancer)

Residuals:

Min 1Q Median 3Q Max

-1.88309 -0.46629 0.08045 0.47380 1.53219

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.685796 0.998754 -0.687 0.49409 vesinv 0.782623 0.268339 2.917 0.00448 \*\* gleason 0.358153 0.127976 2.799 0.00629 \*\* cancervol 0.069454 0.014624 4.749 7.77e-06 \*\*\*

benpros 0.087470 0.029605 2.955 0.00401 \*\* capspen -0.026521 0.032860 -0.807 0.42177 weight 0.001380 0.001822 0.757 0.45079

age -0.002799 0.011724 -0.239 0.81186

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7679 on 89 degrees of freedom

Multiple R-squared: 0.5893, Adjusted R-squared: 0.557

F-statistic: 18.24 on 7 and 89 DF, p-value: 7.694e-15

**Adjusted R-squared**

For fullfit 0.557

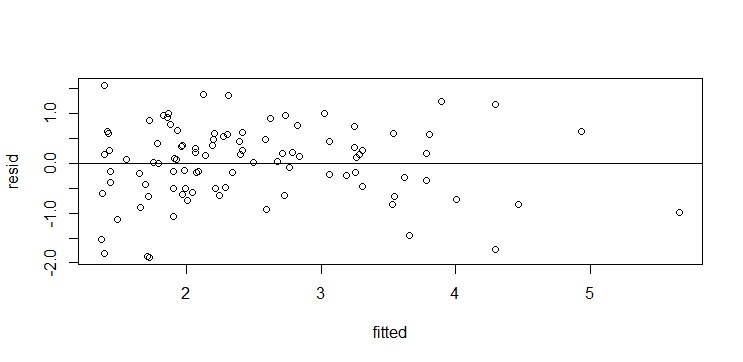
For finalfit 0.5653

Adjusted R-squared value for finalfit is higher than fullfit.

> plot(fitted(finalfit),resid(finalfit),xlab="fitted",ylab="resid")

> abline(h=0)

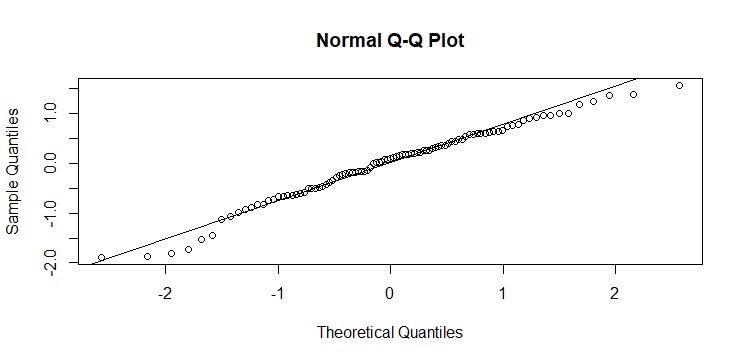
#Residual plot



The points are rightly scattered around 0. No pattern and no change in the vertical scatter.

> qqnorm(resid(finalfit))

> qqline(resid(finalfit))



The final fit is approximately normal, normality assumption holds.

> predict(finalfit, newdata = data.frame(vesinv=0,gleason=mean(cancer$glea son),benpros=mean(cancer$benpros),cancervol=mean(cancer$cancervol))

+ )

1

2.330541