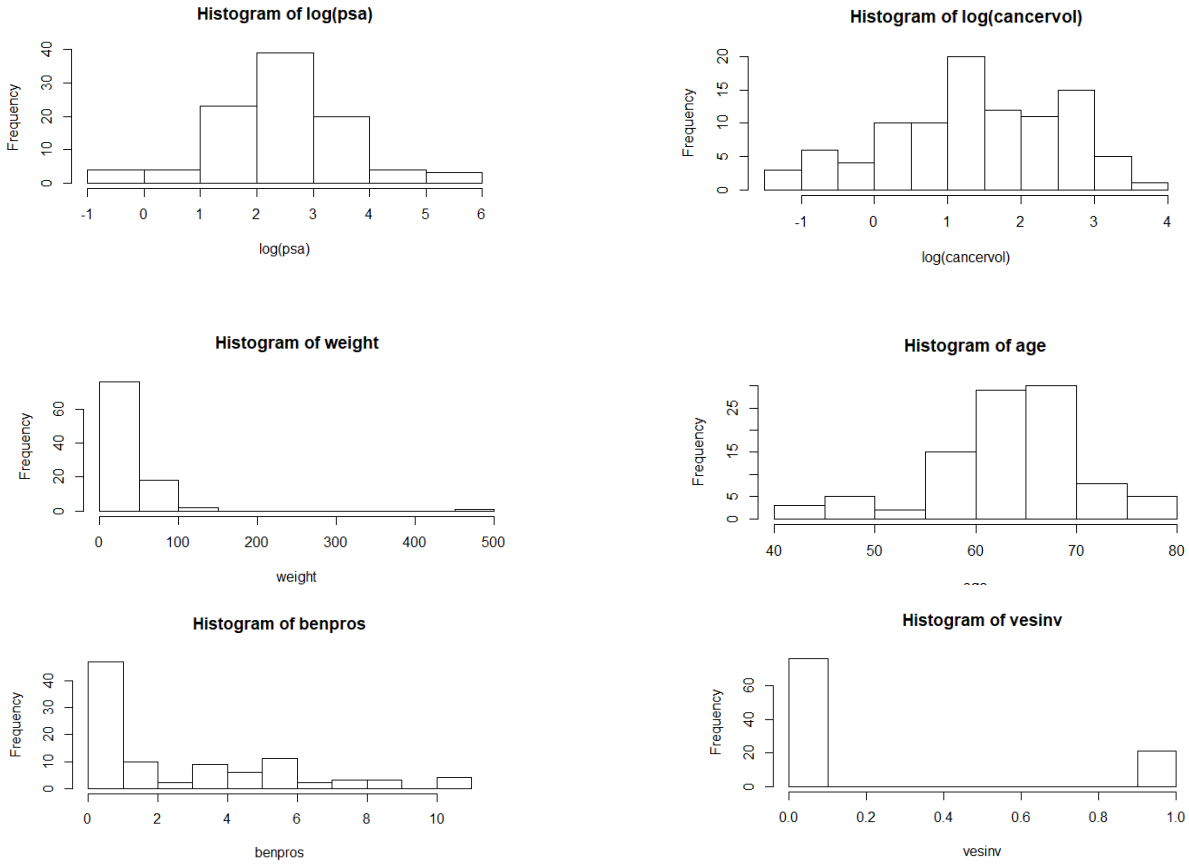
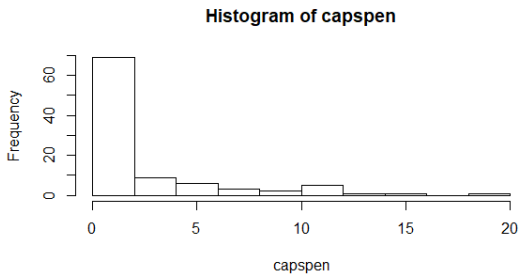
Section 1

The prostate cancer dataset consists of data on 97 men with advanced prostate cancer. The goal is to understand how PSA level is related to the other variables in the dataset. Before constructing a “reasonably good” linear model, an initial exploration of data was done.

The distributions of the prostate cancer variables were examined with histograms. Since the distributions of psa, and cancervol were highly skewed, log transformations were done to transform this data to approximately conform to normality. This makes patterns in the data more interpretable. Additionally, one of the assumptions of regression is that residuals are normally distributed. It is often easier to meet this assumption if the variables in the analysis are normally distributed.

Figure 1. Distributions of prostate cancer variables

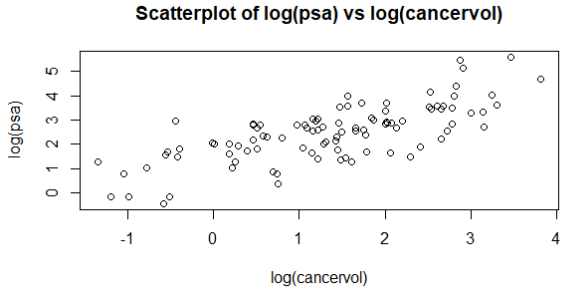


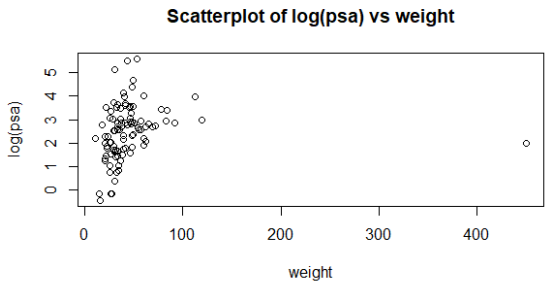


From the histograms we can see that log(psa) is approximately normally distributed, but the distribution of capsen is highly skewed. The variable gleason takes integer values higher than 6. Additionally, vesinv is binary and hence it can be directly entered as a predictor in a multiple regression model. Recoding as dummy variables is not necessary since there are only two categories, 0 and 1.

Next, scatterplots were made for each quantitative variable against psa to see if there appears to be a relationship between predictors and the response variable. Based on the scatterplots below, cancervol and capspen seems to have a weak positive relationship with psa. The relationship of the response variable with other predictors is not clear given the small correlation coefficients (r) and the random scatter in the plots.

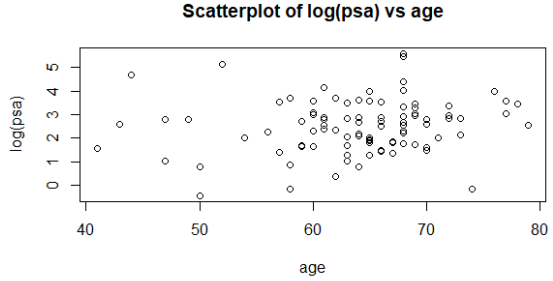
Figure 2. Scatterplots of prostate cancer variables against log(psa)

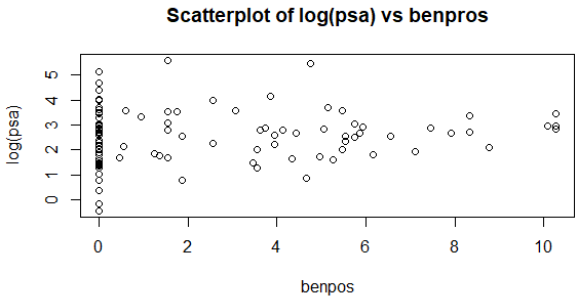




r = 0.624

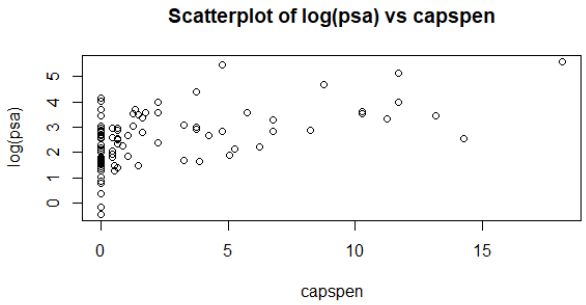
r = 0.026

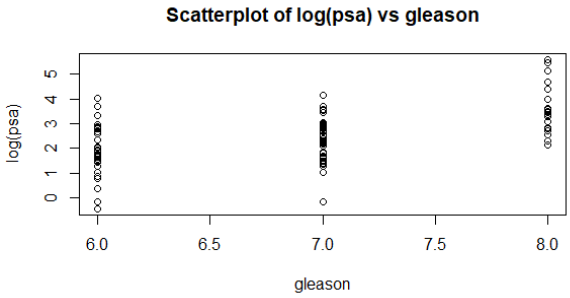




r = 0.017

r = -0.016

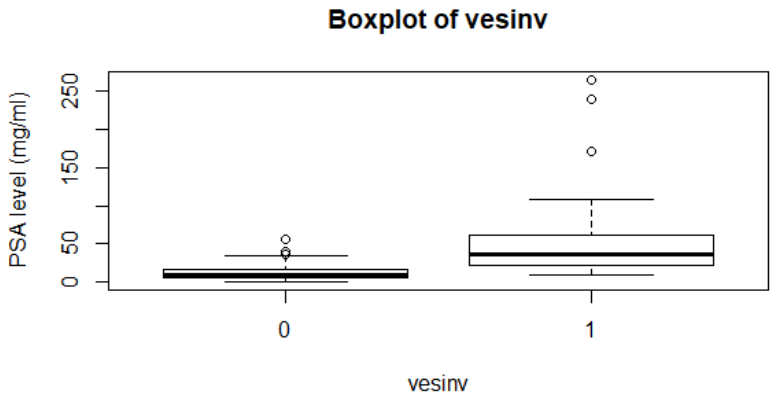




r = 0.551

r = 0.429

Boxplots were constructed for the categorical variable vesinv. Vesinv category 0 seems to have slightly lower PSA levels compared to category 1.



Linear regression

Based on the exploratory data analysis above, cancervol has the strongest relationship with psa. If we wanted to predict psa from other cancer variables it seems that we should at least include cancervol as a predictor. To begin with, a simple regression model was fitted with log(psa) as the response and log(cancervol) as the predictor. The null hypothesis is H_0: $\beta_1 = 0$ and the alternative hypothesis is $H_a = \beta_1 \neq 0$. The F statistic was used to compute the significance of the model. From the scatterplot of log(psa) vs. log(cancervol) it seems that these two are linearly correlated and the intercept passes through the origin. Therefore, the intercept was set as 0 in the model. Following are the results from the lm() function in R:

```
Call:
lm(formula = log(psa) ~ 0 + log(cancervol))

Residuals:
    Min      1Q  Median      3Q     Max
-1.6142 -0.0628  0.5538  1.4664  3.5701
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
log(cancervol)  1.35553    0.07193   18.85   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.267 on 96 degrees of freedom
Multiple R-squared:  0.7872,   Adjusted R-squared:  0.785
F-statistic: 355.2 on 1 and 96 DF,  p-value: < 2.2e-16
```

A high value for the F statistic and a very low p-value ($< 2.2e^{-16}$) implies that the null hypothesis can be rejected. The coefficient estimate is highly statistically significant with a p-value $<2.2e^{-16}$. This implies that there is a relationship between log(cancervol) and log(psa) as was expected from the exploratory data analysis above. The adjusted R-squared value of 0.785 indicates that 78.5% of the variability in the data can be explained by the model.

Multivariate linear regression:

It is possible that including more information and using multiple predictors can improve our model. The first step was fitting a multiple regression model using all the predictors. The null hypothesis in this case states that there is no relation between any of the predictors and the response variable. This was tested by computing the F statistic. The lm() function in R gives the following results:

```
Call:
lm(formula = log(psa) ~ 0 + log(cancervol) + weight + age + benpros +
    vesinv + capspen + gleason)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55930 -0.43197  0.03849  0.50033  1.99311

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
log(cancervol)  0.535066   0.082390   6.494 4.48e-09 ***
weight          0.002250   0.001679   1.341 0.183433
age            -0.013723   0.008918  -1.539 0.127370
benpros         0.064268   0.026666   2.410 0.017981 *
vesinv          0.731290   0.245693   2.976 0.003745 **
capspen        -0.021996   0.028873  -0.762 0.448154
gleason         0.328248   0.082614   3.973 0.000143 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7098 on 90 degrees of freedom
Multiple R-squared:  0.9374,   Adjusted R-squared:  0.9325
F-statistic: 192.4 on 7 and 90 DF,  p-value: < 2.2e-16
```

A high value for the F statistic and a very low p-value ($< 2.2e^{-16}$) implies that the null hypothesis can be rejected. There is a potential relationship between the predictors and the response variable. The adjusted R^2 value increased compared to the simple linear model. 93.25% of the variance in the data can be explained by our new model. The log(cancervol), vesinv and gleason coefficient estimates are highly statistically significant with p-values < 0.05. Benpros coefficient is statistically significant with p-value <0.05. The coefficient estimates of weight, age and capspen have high p values (> 0.05) which implies that they are insignificant in predicting psa.

The first model with only log(cancervol) as predictor was compared with this model using the anova function in R.

```
Analysis of Variance Table

Model 1: log(psa) ~ 0 + log(cancervol)
Model 2: log(psa) ~ 0 + log(cancervol) + weight + age + benpros + vesinv +
    capspen + gleason
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     96 153.997
2     90  45.341  6    108.66 35.947 < 2.2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The low p-value of <2.2e-06 indicates that the inclusion of more predictors did lead to a significant improvement over just using log(cancervol). However, the complex model can be improved because it contains several insignificant variables. In order to improve the model, the insignificant variables were removed, and a new multiple regression model was constructed. The results obtained are below:

```
Call:
lm(formula = log(psa) ~ 0 + log(cancervol) + gleason + vesinv +
    benpros)

Residuals:
     Min       1Q   Median       3Q      Max
-1.65073 -0.35337  0.02895  0.52542  1.90098

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
log(cancervol)  0.51410    0.07612   6.754 1.23e-09 ***
gleason         0.21506    0.01921  11.196  < 2e-16 ***
vesinv          0.67773    0.21166   3.202  0.00187 **
benpros         0.06394    0.02437   2.624  0.01016 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7137 on 93 degrees of freedom
Multiple R-squared:  0.9345,   Adjusted R-squared:  0.9317
F-statistic:   332 on 4 and 93 DF,  p-value: < 2.2e-16
```

The regression coefficients are all highly statistically significant with p-values <0.01 or statistically significant with p-values <0.05. The small p-value of the F statistic implies that this model is significant. The high value for the adjusted R^2 of 0.9317 indicates that the model is a good fit. 93.17% of the variability in the data can be explained by our model. This model was compared with the full model that included all the predictors.

```
Analysis of Variance Table

Model 1: log(psa) ~ 0 + log(cancervol) + gleason + vesinv + benpros
Model 2: log(psa) ~ 0 + log(cancervol) + weight + age + benpros + vesinv +
    capspen + gleason
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     93 47.371
2     90 45.341  3    2.0303 1.3433 0.2654
```

The high p-value of 0.2654 above indicates that the additional predictors in the full model did not lead to a significantly improved fit over the partial model. Therefore, the partial model is preferable because a simple model is better than a complex model.

Comparing the 3 models constructed gives the following results:

```
Analysis of Variance Table

Model 1: log(psa) ~ 0 + log(cancervol)
Model 2: log(psa) ~ 0 + log(cancervol) + gleason + vesinv + benpros
Model 3: log(psa) ~ 0 + log(cancervol) + weight + age + benpros + vesinv +
    capspen + gleason
  Res.Df     RSS Df Sum of Sq       F Pr(>F)
1     96 153.997
2     93  47.371  3    106.63 70.5500 <2e-16 ***
3     90  45.341  3      2.03  1.3433 0.2654
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The low p-value of <2e-16 indicates that including gleason, vesinv and benpros in Model 2 leads to a significantly improved fit over the model with just log(cancervol). However, the additional predictors in Model 3 does not lead to a significant improvement over Model 2. Since Model 2 is simpler than Model 3 and all its regression coefficients are statistically significant, we use it as a “reasonably good” linear model for predicting psa.

To confirm results, the step function in R was also used to perform forward stepwise selection with AIC. This method starts with no predictors in the model and then iteratively adds the most contributive predictors. It stops when improvement is not statistically significant. See below for results.

R step () function results:

```
Start:  AIC=194.94
log(psa) ~ 0

                 Df Sum of Sq    RSS     AIC
+ gleason         1    621.09 102.62   7.467
+ age             1    598.82 124.90  26.521
+ log(cancervol)  1    569.72 154.00  46.836
+ weight          1    332.66 391.06 137.233
+ capspen         1    307.50 416.22 143.280
+ benpros         1    291.38 432.34 146.967
+ vesinv          1    289.86 433.85 147.306
<none>                        723.72 194.940

Step:  AIC=7.47
log(psa) ~ gleason - 1

                 Df Sum of Sq     RSS     AIC
+ log(cancervol)  1    47.786  54.838 -51.322
+ vesinv          1    26.982  75.642 -20.124
+ capspen         1    20.634  81.990 -12.307
+ age             1     2.404 100.220   7.168
<none>                        102.624   7.467
+ benpros         1     1.950 100.674   7.607
+ weight          1     1.130 101.494   8.393

Step:  AIC=-51.32
log(psa) ~ gleason + log(cancervol) - 1

          Df Sum of Sq    RSS     AIC
+ vesinv   1    3.9594 50.878 -56.592
+ benpros  1    2.2443 52.593 -53.376
+ weight   1    1.7988 53.039 -52.558
<none>                 54.838 -51.322
+ capspen  1    0.3804 54.457 -49.998
+ age      1    0.3347 54.503 -49.916

Step:  AIC=-56.59
log(psa) ~ gleason + log(cancervol) + vesinv - 1

          Df Sum of Sq    RSS     AIC
+ benpros  1    3.5072 47.371 -61.520
+ weight   1    1.9310 48.947 -58.345
<none>                 50.878 -56.592
+ capspen  1    0.2569 50.621 -55.083
+ age      1    0.0771 50.801 -54.739

Step:  AIC=-61.52
log(psa) ~ gleason + log(cancervol) + vesinv + benpros - 1

          Df Sum of Sq    RSS     AIC
<none>                 47.371 -61.520
+ age      1   0.85937 46.512 -61.296
+ weight   1   0.67340 46.698 -60.909
+ capspen  1   0.15631 47.215 -59.841
```

The model suggested with the step function is the same as Model 2 above.

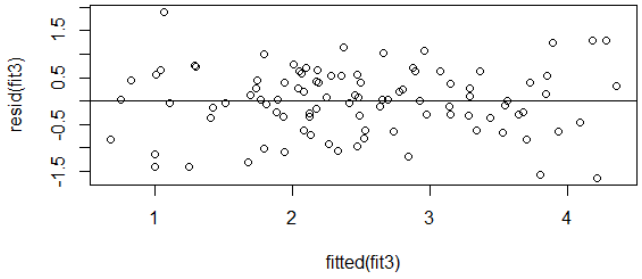
Final chosen model with adjusted R^2 of 0.9317:

$$log(psa) = \beta_0 + \beta_1 log(cancervol) + \beta_2 gleason + \beta_3 vesinv + \beta_4 benpros$$

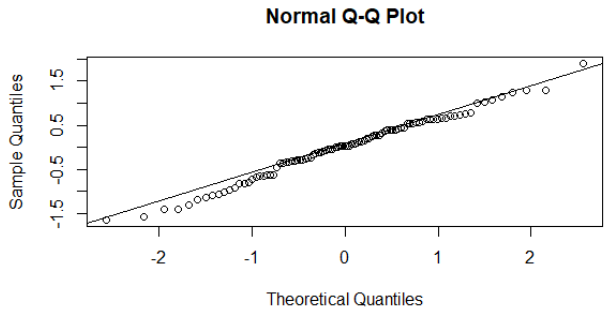
$$log(psa) = 0 + 0.51410\, log(cancervol) + 0.21506\, gleason + 0.67773\, vesinv + 0.06394\, benpros$$

Evaluation of final model:

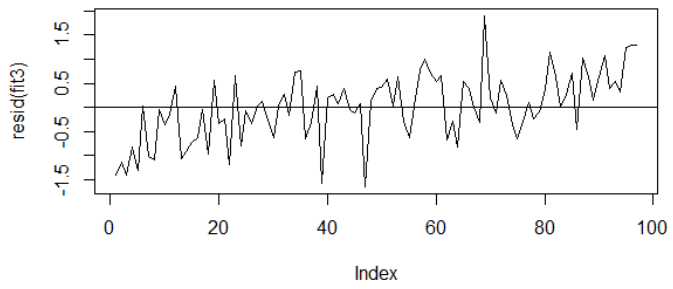
One of the assumptions of linear regression is that there is a linear relationship between the predictors and the response. If the underlying relationship is not linear then our model is not correct. The linearity of the model was verified via a residual plot of fitted values vs. the residuals. A non-random pattern would suggest that a linear model is not appropriate. Since the points appear randomly scattered with no pattern, the linearity assumption is confirmed. Additionally, the absence of a pattern also verifies the assumption that residuals have constant variance.



Another assumption is that residuals are normally distributed. To verify this assumption, a normal QQ plot was used. For normally distributed data, observations lie approximately on a straight line. Since the points in the QQ plot below are scattered roughly along the straight diagonal line, the normality assumption for residuals is satisfied.



To check for the assumption that residuals are independently distributed, a time series plot is used. In the plot below we can see a slight trend. A sequence of negative residuals from index 0-10 and a sequence of mostly positive residuals from index 80-100. Therefore, we cannot assume that residuals are truly independent, and our model is not perfect. However, it is "reasonably good" for predicting psa as demonstrated in the analysis above.



Prediction:

The final model

$$log(psa) = \beta_1 log(cancervol) + \beta_2\, gleason + \beta_3 vesinv + \beta_4\, benpros$$

with R^2 of 0.9317 is used 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.

```
> new.xvalues <-data.frame(cancervol=log(mean(cancervol)),gleason=mean(gleaso
n),vesinv=which.max(table(vesinv)), benpros = mean(benpros))
> log_psa = predict(fit3, newdata = new.xvalues)
> log_psa
       0
2.660839
> psalevel = exp(log_psa)
> psalevel
       0
14.30829
```

The predicted PSA level for the patient is 14.30829 mg/ml.

Section 2

R code:

```
#loading data
pc <- read.csv("prostate_cancer.csv")
attach(pc)

#drawing histograms for cancer variables
hist(log(cancervol))
hist(weight)
hist(log(psa))
hist(age)
hist(benpros)
hist(vesinv)
hist(capspen)
hist(gleason)

#drawing scatterplots and calculating r
plot(log(cancervol),log(psa),main = "Scatterplot of log(psa) vs
log(cancervol)", xlab = 'log(cancervol)', ylab = "log(psa)")
cor(cancervol,psa)
plot(weight,log(psa),main = "Scatterplot of log(psa) vs weight", xlab =
'weight', ylab = "log(psa)")
cor(weight,psa)
plot(age,log(psa),main = "Scatterplot of log(psa) vs age", xlab = 'age', ylab
= "log(psa)")
cor(age,psa)
plot(benpros,log(psa),main = "Scatterplot of log(psa) vs benpros", xlab =
'benpos', ylab = "log(psa)")
cor(benpros,psa)
plot(capspen,log(psa),main = "Scatterplot of log(psa) vs capspen", xlab =
'capspen', ylab = "log(psa)")
cor(capspen,psa)
plot(gleason,log(psa),main = "Scatterplot of log(psa) vs gleason", xlab =
'gleason', ylab = "log(psa)")
cor(gleason,psa)


#drawing boxplots for qualitative variable
boxplot(psa ~ vesinv, data = pc, main = "Boxplot of vesinv", ylab="PSA level
(mg/ml)")
boxplot(psa ~ gleason, data = pc, main = "Boxplot of gleason", ylab = "PSA
level (mg/ml)")

#simple linear regression model
fit1 <- lm(log(psa)~0+log(cancervol))
summary(fit1)

#full model using all predictors
fit2 <- lm(log(psa) ~ 0+log(cancervol) + weight + age + benpros+ vesinv +
capspen + gleason)
summary(fit2)
```

```r
#comparing models
anova(fit1,fit2)

#creating model after removing insignificant predictors
fit3 <- lm(log(psa) ~ 0+log(cancervol) + gleason+ vesinv + benpros)
summary(fit3)

#comparing models
anova(fit3,fit2)

#drawing residual plot and QQ plot
plot(fitted(fit3),resid(fit3))
abline(h=0)
qqnorm(resid(fit3))
qqline(resid(fit3))

#time series plot of residuals
plot(resid(fit3),type = "l")
abline(h=0)


#comparing all 3 models
anova(fit1,fit3,fit2)

#using forward stepwise selection
fitfull <- lm(log(psa) ~ 0+log(cancervol) + weight +age + benpros+ vesinv +
capspen +gleason)

fit_step3 <-step(fitnull,scope = list(upper=fitfull),direction = "forward",
trace = 1)
summary(fit_step3)


#prediction
new.xvalues <-data.frame(cancervol=
log(mean(cancervol)),gleason=mean(gleason),vesinv=which.max(table(vesinv)),
benpros = mean(benpros))
log_psa = predict(fit3, newdata = new.xvalues)
log_psa
psalevel = exp(log_psa)
psalevel
```