HW1_Complete by Sudhanshu Raj Singh

Chapter 1: Exercise 1

Problem Statement:The dataset teengamb concerns a study of teenage gambling in Britain. Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data.

Solution:

I started with loading the data set and understanding it by using help. Following output is generated from help:

Study of teenage gambling in Britain

Description

The teengamb data frame has 47 rows and 5 columns. A survey was conducted to study teenage gambling in Britain.

Usage

data(teengamb) Format

This data frame contains the following columns:

sex 0=male, 1=female

status Socioeconomic status score based on parents' occupation

income in pounds per week

verbal verbal score in words out of 12 correctly defined

gamble expenditure on gambling in pounds per year

Source

Ide-Smith & Lea, 1988, Journal of Gambling Behavior, 4, 110-118

Then I loaded the data and summarised it.

```r
library('faraway')
```

```
## Warning: package 'faraway' was built under R version 3.4.4
```

```r
data(teengamb,package='faraway')
summary(teengamb)
```

```
##       sex             status          income           verbal
##  Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00
##  1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00
##  Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00
##  Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66
##  3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00
##  Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00
##      gamble
##  Min.   :  0.0
##  1st Qu.:  1.1
##  Median :  6.0
```

```
##  Mean   : 19.3
##  3rd Qu.: 19.4
##  Max.   :156.0
```

Making Sex as factor

We can see from the documentation that Sex is a factor variable where 0 means Male and 1 means Female. So I changed the Sex column of teengamb dataset.Again looking at summary, data looks more organized.

```
teengamb$sex<-factor(teengamb$sex)
levels(teengamb$sex)<-c("Male","Female")
summary(teengamb)
```

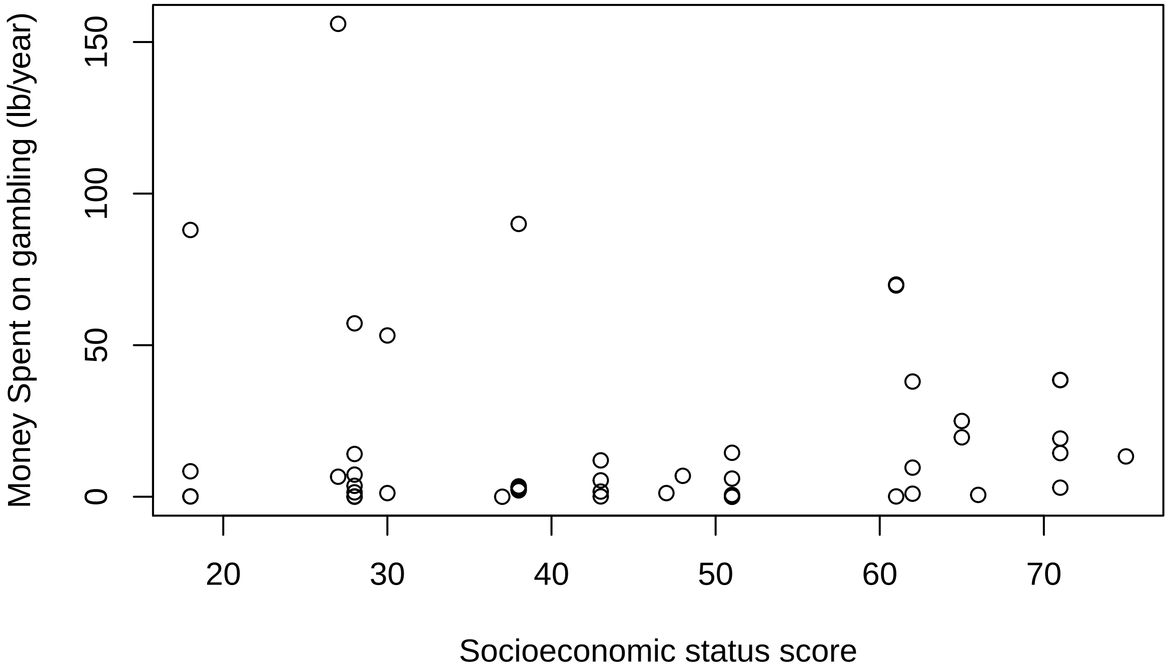
```
##      sex          status          income           verbal     
##  Male  :28   Min.   :18.00   Min.   : 0.600   Min.   : 1.00  
##  Female:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00  
##              Median :43.00   Median : 3.250   Median : 7.00  
##              Mean   :45.23   Mean   : 4.642   Mean   : 6.66  
##              3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00  
##              Max.   :75.00   Max.   :15.000   Max.   :10.00  
##      gamble     
##  Min.   :  0.0  
##  1st Qu.:  1.1  
##  Median :  6.0  
##  Mean   : 19.3  
##  3rd Qu.: 19.4  
##  Max.   :156.0
```

After this, I tried to see how gambling tendence varies with sex.The following box plot indicates that male are more likely to spend money on gambling than females.Also there is more variation in spending among males than females. As the income rises, males spend more money on gambling than women.

```
boxplot(teengamb$gamble~teengamb$sex,ylab="Money Spent on gambling (lb/year)",xlab="Sex")
```

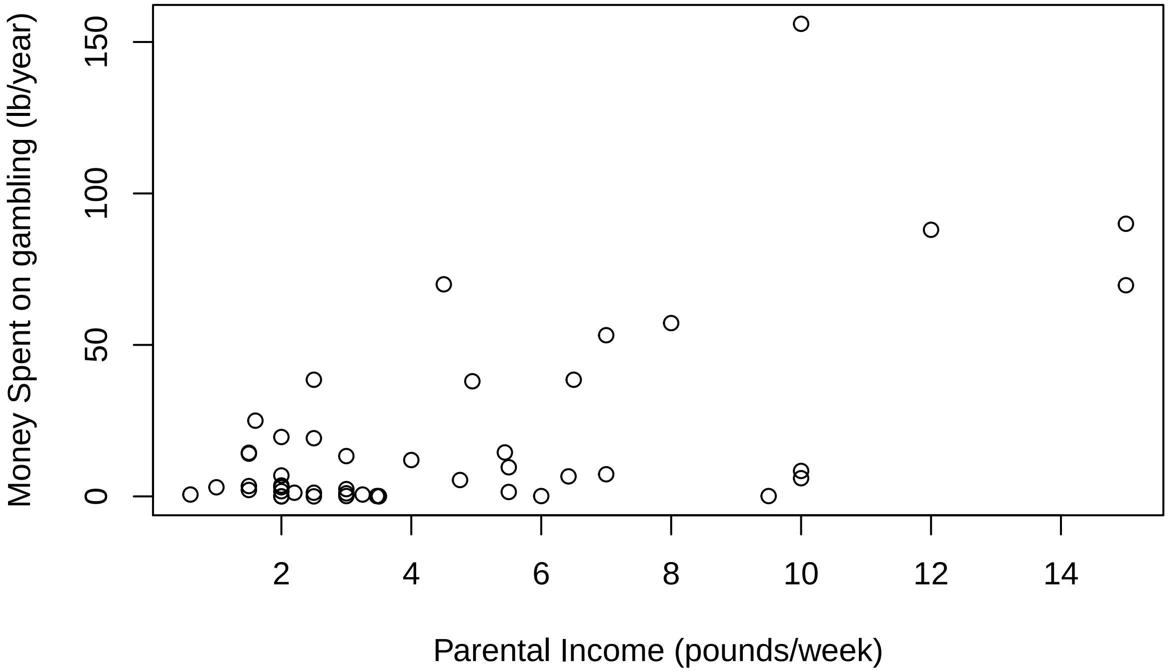
After this, I tried to see how gambling varies with the status. The following plot reveals that there seems to be a weak negative correlation between status and money spent on gambling, that is, money spent on gambling decreases as the socioeconomic status of teenagers increases.

```
plot(teengamb$gamble~teengamb$status,ylab="Money Spent on gambling (lb/year)",xlab="Socioeconomic statu
```



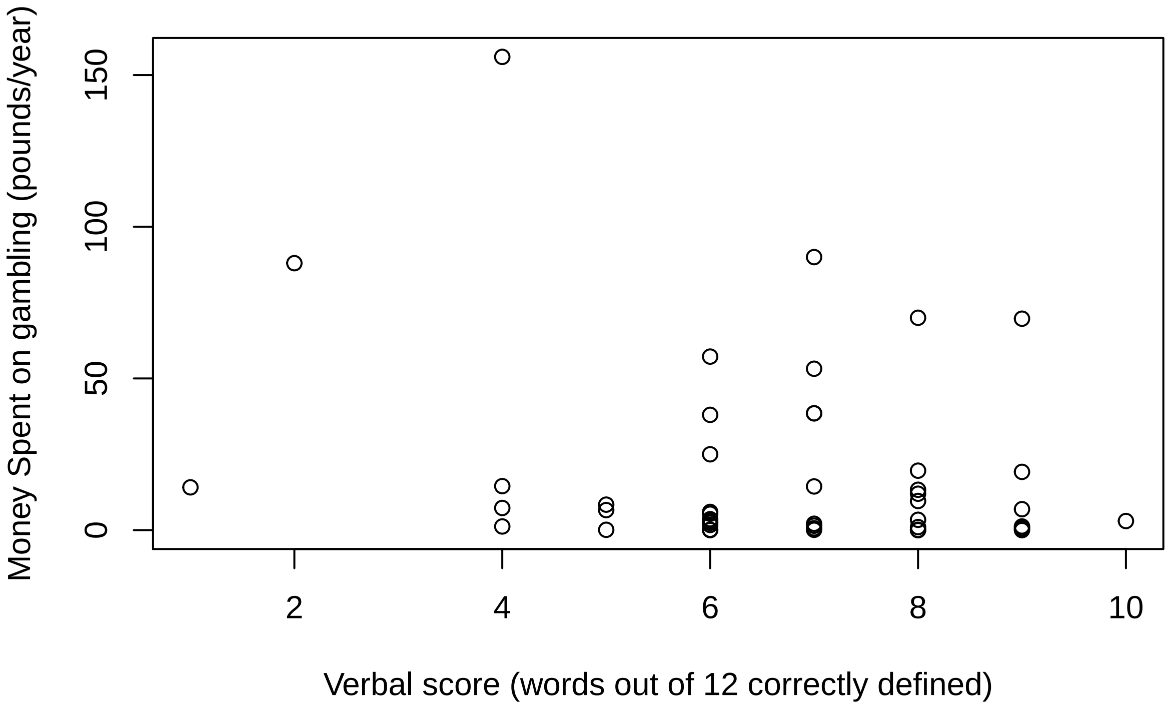
After this, I tried to see how gambling varies with the parental income. The following plot reveals that there seems to be a positive correlation between parental income and money spent on gambling, that is, money spent on gambling increases as the parental income of teenagers increases.

```
plot(teengamb$gamble~teengamb$income,ylab="Money Spent on gambling (lb/year)",xlab="Parental Income (po
```



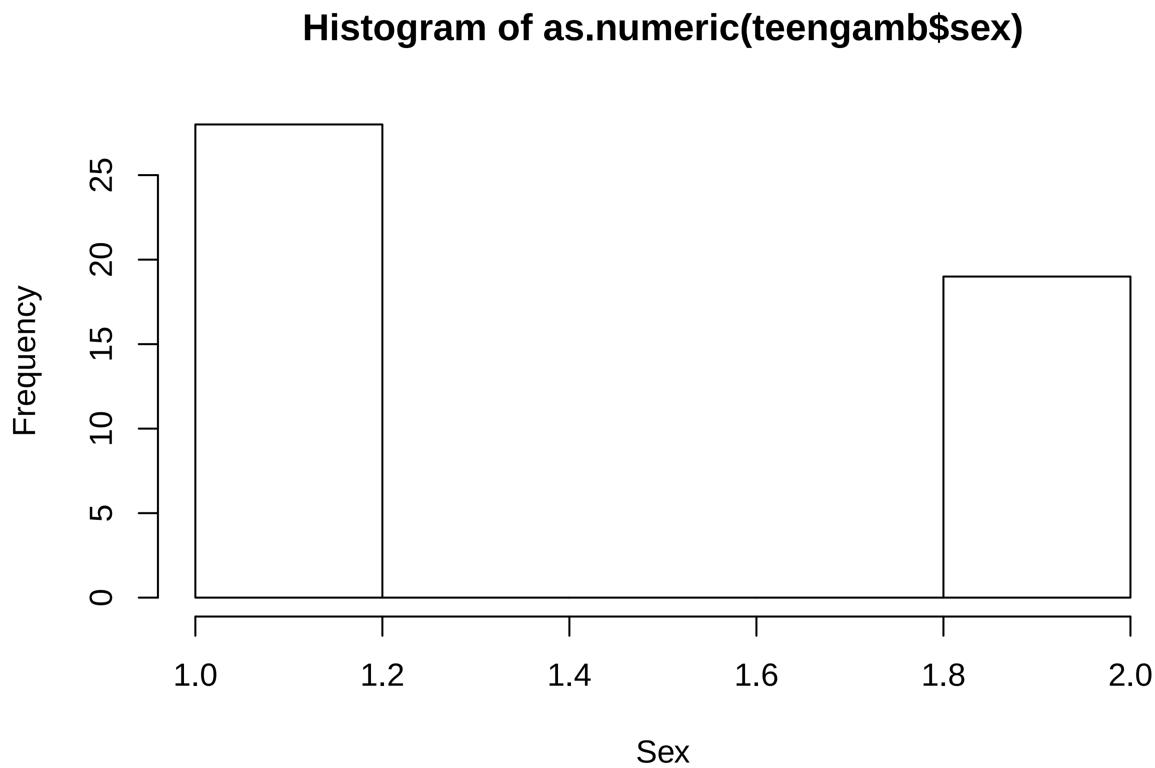
After this, I tried to see how gambling varies with the verbal scores of teenagers. The following plot reveals that there seems to be a strong negative correlation between verbal scores and money spent on gambling, that is, money spent on gambling by a teenager decreases as the verbal scores of teenagers increases.

```
plot(teengamb$gamble~teengamb$verbal,ylab="Money Spent on gambling (pounds/year)",xlab="Verbal score (w
```



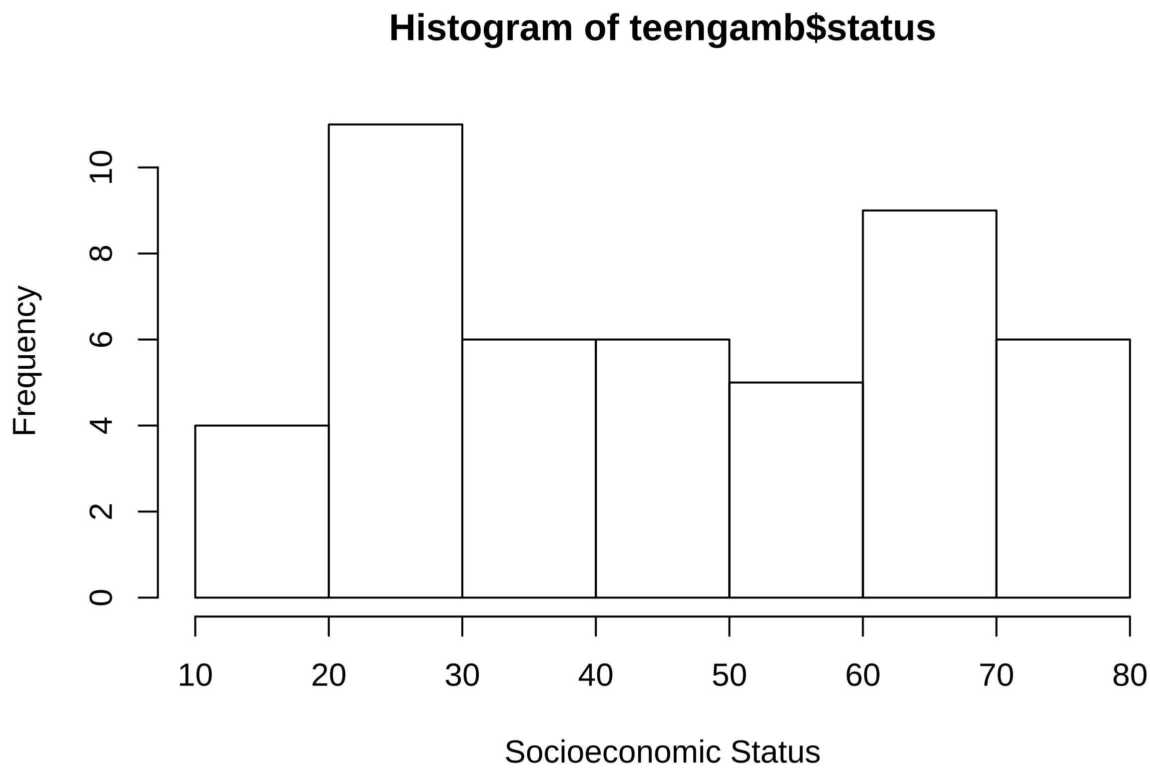
Lastly, I tried to see if there is skewness in the sample by plotting the count of observations for each variable. Without any knowledge about composition of British society, I cannot comment on all but it can be easily said that regarding the sex variable, the sample is skewed towards male (60% of observations) which is not the representative of population.

```
hist(as.numeric(teengamb$sex),xlab="Sex")
```

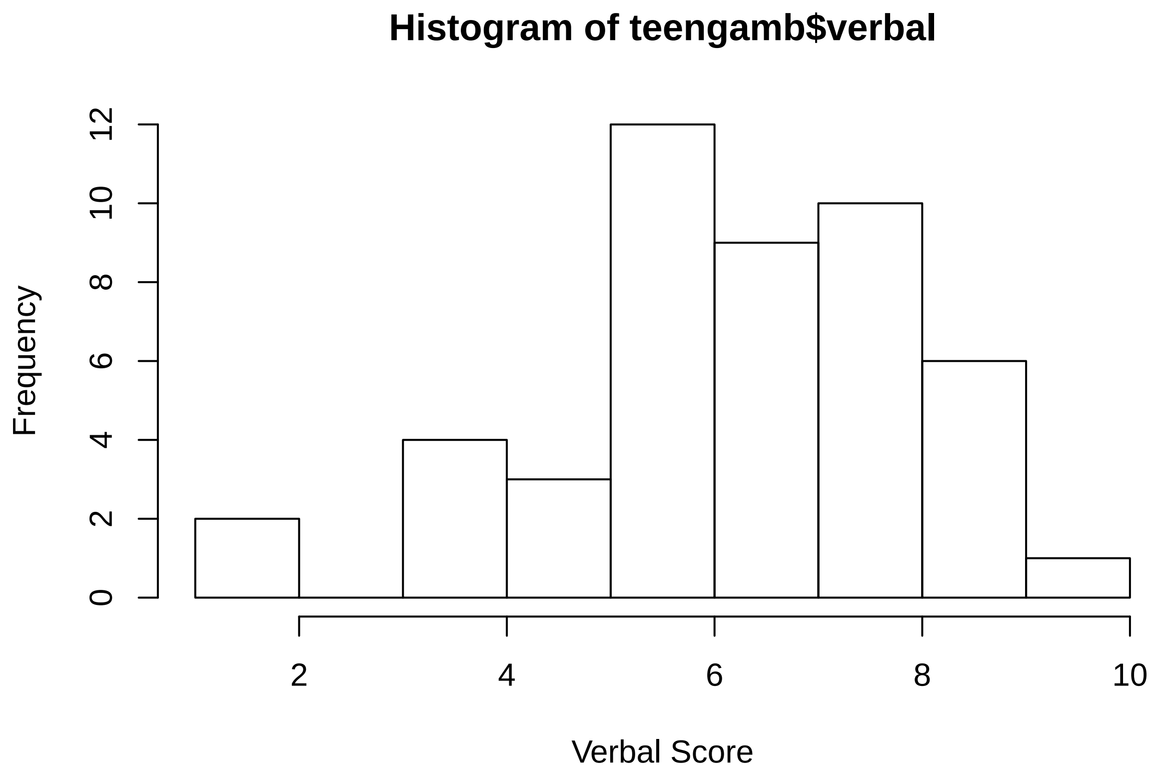



```r
hist(teengamb$income,xlab="Parental Income")
```


```
hist(teengamb$status,xlab="Socioeconomic Status")
```




```r
hist(teengamb$verbal,xlab="Verbal Score")
```



Chapter 1: Exercise 3 The dataset prostate is from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Make a numerical and graphical summary of the data as in the first question.

Solution:

Following output is generated from help about this data.

Prostate cancer surgery

Description

The prostate data frame has 97 rows and 9 columns. A study on 97 men with prostate cancer who were due to receive a radical prostatectomy.

Usage

data(prostate) Format

This data frame contains the following columns:

lcavol log(cancer volume)

lweight log(prostate weight)

age age

lbph log(benign prostatic hyperplasia amount)

svi seminal vesicle invasion

lcp log(capsular penetration)

gleason Gleason score

pgg45 percentage Gleason scores 4 or 5

lpsa log(prostate specific antigen)

Source

Andrews DF and Herzberg AM (1985): Data. New York: Springer-Verlag

```r
data(prostate,package="faraway")
head(prostate,n=5)
```

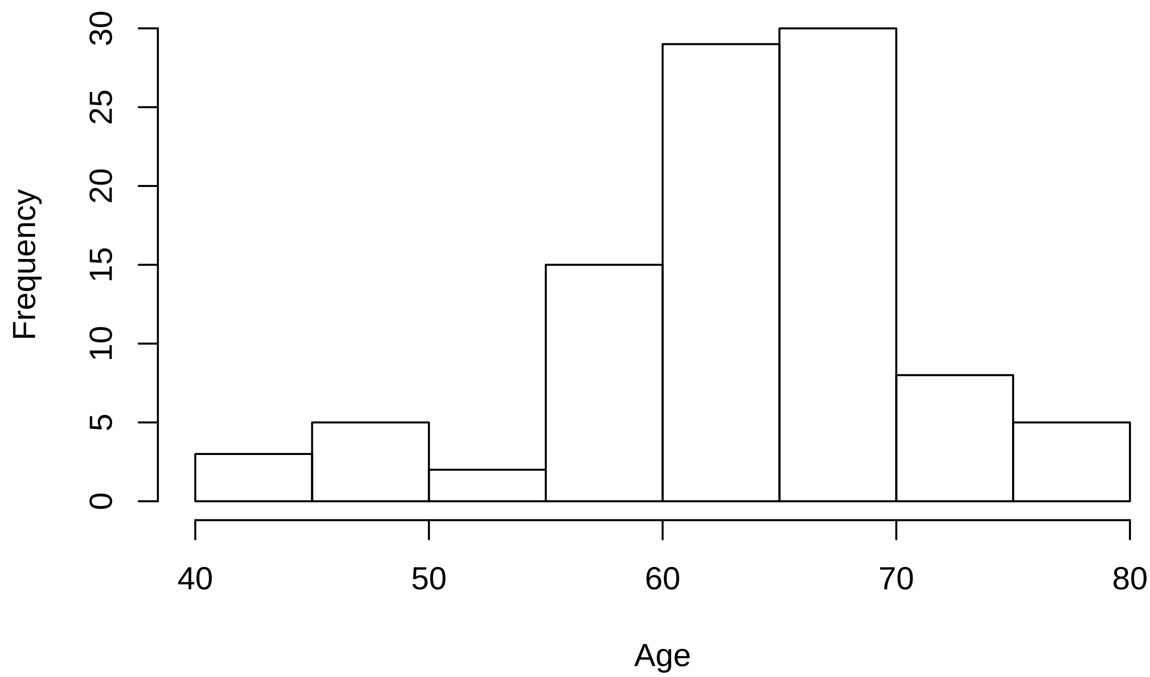
```
##       lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
## 1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
## 2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
## 3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
## 4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
## 5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
```

```r
summary(prostate)
```

```
##      lcavol           lweight           age             lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004  
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581  
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263  
##       svi              lcp             gleason          pgg45       
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
##  3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00  
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
##       lpsa        
##  Min.   :-0.4308  
##  1st Qu.: 1.7317  
##  Median : 2.5915  
##  Mean   : 2.4784  
##  3rd Qu.: 3.0564  
##  Max.   : 5.5829  
```

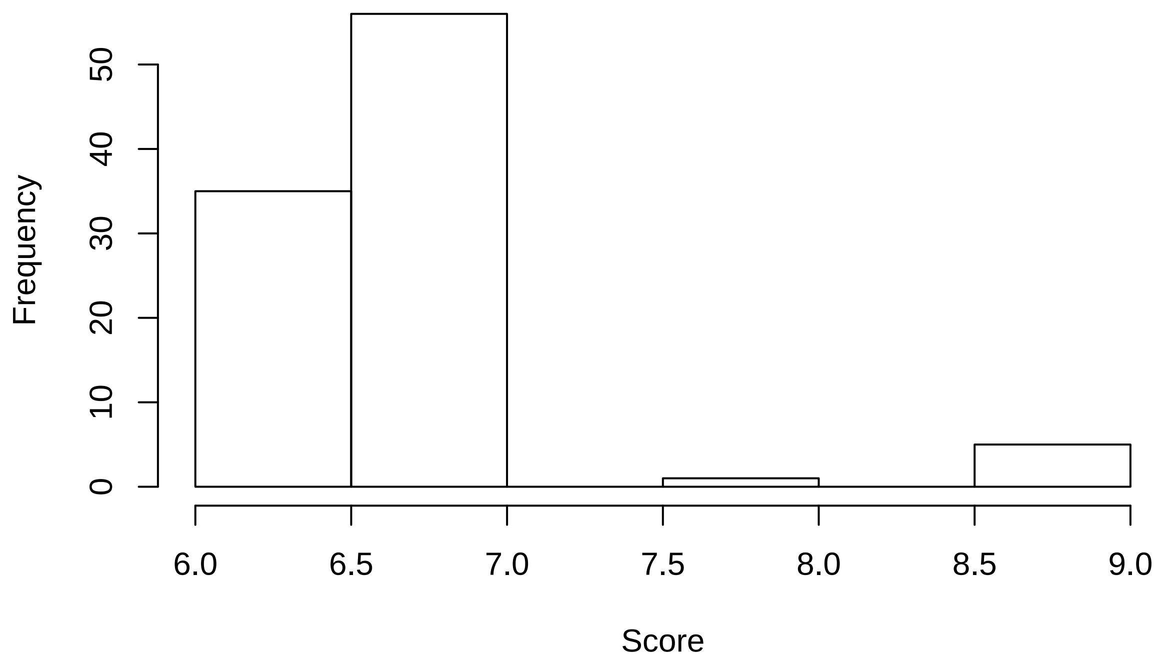
I started with plotting frequency of prostate cancer patients with age. It can be seen that people in age group 60-70 are most numerous in this group which is an indication that people of this age group are prone to having prostate cancer.

```r
 hist(prostate$age,xlab="Age",ylab="Frequency",main="")
```



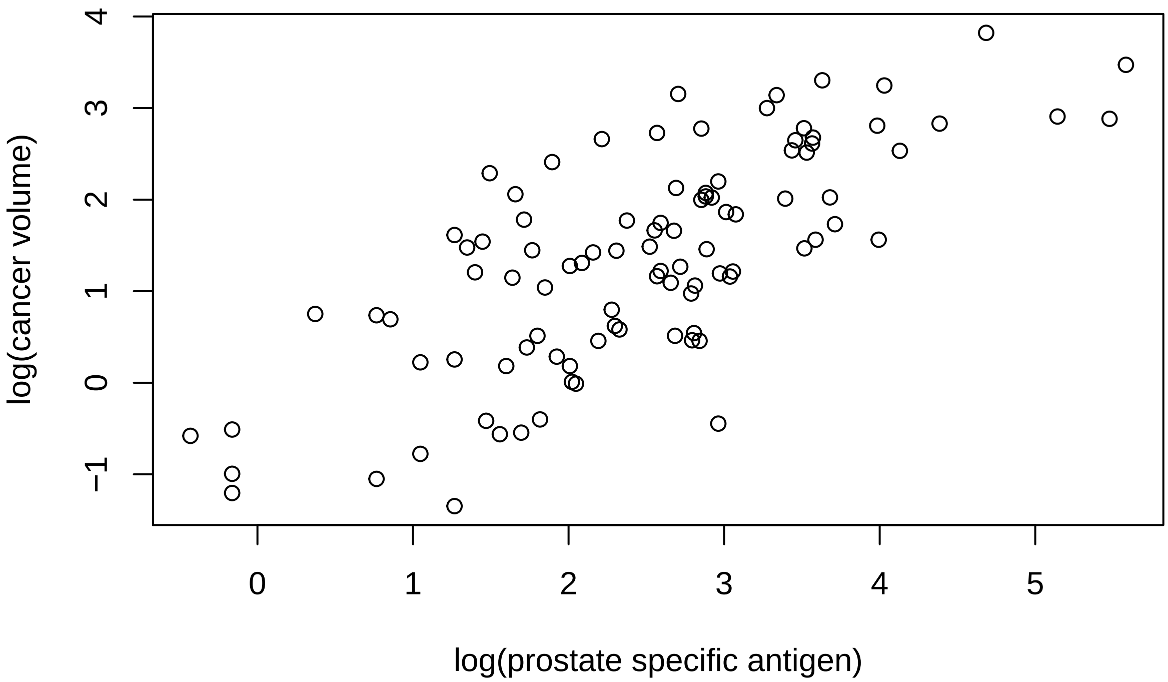
Next, I plotted frequency of gleason scores. It can be seen that most patient about to receive radial prostatectomy have gleason scores 6 or 7.

```
hist(prostate$gleason,xlab="Score",ylab="Frequency",main="")
```



Next, I plotted lcavol(log of cancer valume) vs lpsa(log(prostate specific antigen)). It is clearly evident that there is a positive correlation between the too. So prostate specific antigen can be a good indicator of cancer.Higher the lpsa, higher is the cancer volume.

```
plot(prostate$lpsa,prostate$lcavol,xlab="log(prostate specific antigen)",ylab="log(cancer volume)")
```



Chapter 2: Exercise 4

Loading data and applying linear model

```
data(teengamb,package='faraway')
attach(teengamb)
lmod<-lm(gamble~sex+status+income+verbal)
```

a. What percentage of variation in the response is explained by these predictors? Solution: It is equal to R-square *100

```
summary(lmod)$r.squared*100
```

```
## [1] 52.67234
```

b. Which observation has the largest (positive) residual? Give the case number? using which. max on residuals column of lmod

```
(which.max(lmod$residuals))
```

```
## 24
## 24
```

c. Compute the mean and median of the residuals? Using mean and median functions on residuals column of lmod

```
mean(lmod$residuals)
```

```
## [1] -3.065293e-17
```

```
median(lmod$residuals)
```

```
## [1] -1.451392
```

d. Compute the correlation of the residuals with the fitted values.

```
yhat<-fitted(lmod)
cor(lmod$residuals,yhat)
```

```
## [1] -1.070659e-16
```

e. Compute the correlation of the residuals with the income.

```
cor(lmod$residuals,income)
```

```
## [1] -7.242382e-17
```

f. For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

   Solution: As Male=0 and Female=1, the difference is simly the coefficient of the sex term in the linear model which comes out to be -22.12, that is to say that females, on average spend 22.12 pounds in a year less when compared to males.

```
summary(lmod)$coefficients[2]
```

```
## [1] -22.11833
```

Chapter 2 Exercise 4

The dataset prostate comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with lpsa as the response and lcavol as the predictor. Record the residual standard error and the R2. Now add lweight, svi, lbph, age, lcp, pgg45 and gleason to the model one at a time. For each model record the residual standard error and the R2. Plot the trends in these two statistics

```
data(prostate,package='faraway')
attach(prostate)
pros.mod<-lm(lpsa~lcavol)
summary(pros.mod)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67625 -0.41648  0.09859  0.50709  1.89673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.50730    0.12194   12.36   <2e-16 ***
## lcavol       0.71932    0.06819   10.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7875 on 95 degrees of freedom
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
## F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

```r
rsq<-rep(1,8)
rse<-rep(1,8)
rsq[1]=summary(pros.mod)$r.squared
res<-summary(pros.mod)$residuals
RSS<-c(crossprod(res))
MSE<-RSS/(length(res)-2)
rse[1]=sqrt(MSE)
```

R-Square is 0.5394 and Residual Standard Error is 0.7875.

Adding lweight to linear model and recording R2 and RSE.

```r
pros.mod<-lm(lpsa~lcavol+lweight)
summary(pros.mod)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61965 -0.50778 -0.02095  0.52291  1.89885
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.30262    0.56904  -0.532  0.59612
## lcavol       0.67753    0.06626  10.225  < 2e-16 ***
## lweight      0.51095    0.15726   3.249  0.00161 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7506 on 94 degrees of freedom
## Multiple R-squared:  0.5859, Adjusted R-squared:  0.5771
## F-statistic: 66.51 on 2 and 94 DF,  p-value: < 2.2e-16
```

```r
rsq[2]=summary(pros.mod)$r.squared
res<-summary(pros.mod)$residuals
RSS<-c(crossprod(res))
MSE<-RSS/(length(res)-3)
rse[2]=sqrt(MSE)
```

Adding svi to linear model and recording R2 and RSE.

```r
pros.mod<-lm(lpsa~lcavol+lweight+svi)
summary(pros.mod)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
## svi          0.66616    0.20978   3.176  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
rsq[3]=summary(pros.mod)$r.squared
res<-summary(pros.mod)$residuals
RSS<-c(crossprod(res))
MSE<-RSS/(length(res)-4)
rse[3]=sqrt(MSE)
```

Adding lbph and recording R2 and RSE.

```
pros.mod<-lm(lpsa~lcavol+lweight+svi+lbph)
summary(pros.mod)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82653 -0.42270  0.04362  0.47041  1.48530
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14554    0.59747   0.244  0.80809
## lcavol       0.54960    0.07406   7.422 5.64e-11 ***
## lweight      0.39088    0.16600   2.355  0.02067 *
## svi          0.71174    0.20996   3.390  0.00103 **
## lbph         0.09009    0.05617   1.604  0.11213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7108 on 92 degrees of freedom
## Multiple R-squared:  0.6366, Adjusted R-squared:  0.6208
## F-statistic: 40.29 on 4 and 92 DF,  p-value: < 2.2e-16
```

```
rsq[4]=summary(pros.mod)$r.squared
res<-summary(pros.mod)$residuals
RSS<-c(crossprod(res))
MSE<-RSS/(length(res)-5)
rse[4]=sqrt(MSE)
```

Adding age and recording R2 and RSE.

```
pros.mod<-lm(lpsa~lcavol+lweight+svi+lbph+age)
summary(pros.mod)
```

```
##
## Call:
```

```
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## svi          0.72095    0.20902   3.449 0.000854 ***
## lbph         0.11184    0.05805   1.927 0.057160 .
## age         -0.01489    0.01075  -1.385 0.169528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

```r
rsq[5]=summary(pros.mod)$r.squared
res<-summary(pros.mod)$residuals
RSS<-c(crossprod(res))
MSE<-RSS/(length(res)-6)
rse[5]=sqrt(MSE)
```

Adding age and recording R2 and RSE.

```r
pros.mod<-lm(lpsa~lcavol+lweight+svi+lbph+age+lcp)
summary(pros.mod)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age + lcp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82853 -0.40741  0.01695  0.47159  1.59040
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.93487    0.83577   1.119  0.26630
## lcavol       0.58765    0.08663   6.783  1.2e-09 ***
## lweight      0.41808    0.16792   2.490  0.01462 *
## svi          0.78256    0.24261   3.226  0.00175 **
## lbph         0.11381    0.05842   1.948  0.05452 .
## age         -0.01511    0.01081  -1.398  0.16546
## lcp         -0.04118    0.08135  -0.506  0.61392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7102 on 90 degrees of freedom
## Multiple R-squared:  0.6451, Adjusted R-squared:  0.6215
## F-statistic: 27.27 on 6 and 90 DF,  p-value: < 2.2e-16
```

```
rsq[6]=summary(pros.mod)$r.squared
res<-summary(pros.mod)$residuals
RSS<-c(crossprod(res))
MSE<-RSS/(length(res)-7)
rse[6]=sqrt(MSE)
```

Adding pgg45 and recording R2 and RSE.

```
pros.mod<-lm(lpsa~lcavol+lweight+svi+lbph+age+lcp+pgg45)
summary(pros.mod)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age + lcp +
##     pgg45)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73117 -0.38137 -0.01728  0.43364  1.63513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.953926   0.829439   1.150  0.25319
## lcavol       0.591615   0.086001   6.879 8.07e-10 ***
## lweight      0.448292   0.167771   2.672  0.00897 **
## svi          0.757734   0.241282   3.140  0.00229 **
## lbph         0.107671   0.058108   1.853  0.06720 .
## age         -0.019336   0.011066  -1.747  0.08402 .
## lcp         -0.104482   0.090478  -1.155  0.25127
## pgg45        0.005318   0.003433   1.549  0.12488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 89 degrees of freedom
## Multiple R-squared:  0.6544, Adjusted R-squared:  0.6273
## F-statistic: 24.08 on 7 and 89 DF,  p-value: < 2.2e-16
```

```
rsq[7]=summary(pros.mod)$r.squared
res<-summary(pros.mod)$residuals
RSS<-c(crossprod(res))
MSE<-RSS/(length(res)-8)
rse[7]=sqrt(MSE)
```

Adding gleason and recording R2 and RSE.

```
pros.mod<-lm(lpsa~lcavol+lweight+svi+lbph+age+lcp+pgg45+gleason)
summary(pros.mod)
```

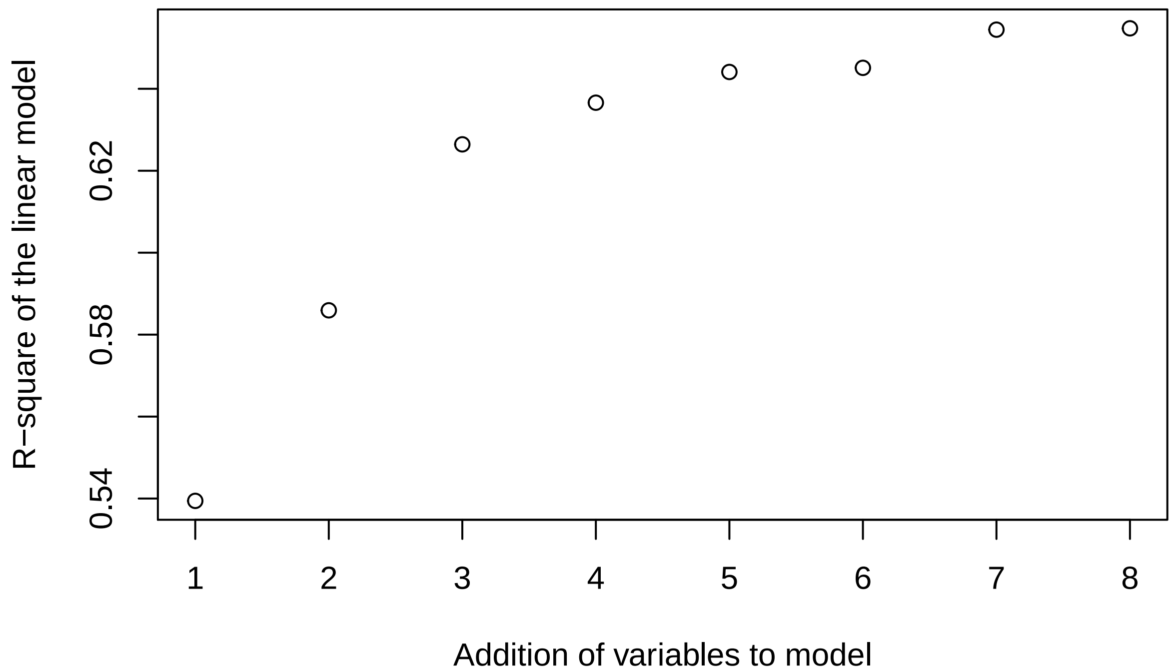
```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age + lcp +
##     pgg45 + gleason)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## svi          0.766157   0.244309   3.136  0.00233 **
## lbph         0.107054   0.058449   1.832  0.07040 .
## age         -0.019637   0.011173  -1.758  0.08229 .
## lcp         -0.105474   0.091013  -1.159  0.24964
## pgg45        0.004525   0.004421   1.024  0.30886
## gleason      0.045142   0.157465   0.287  0.77503
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

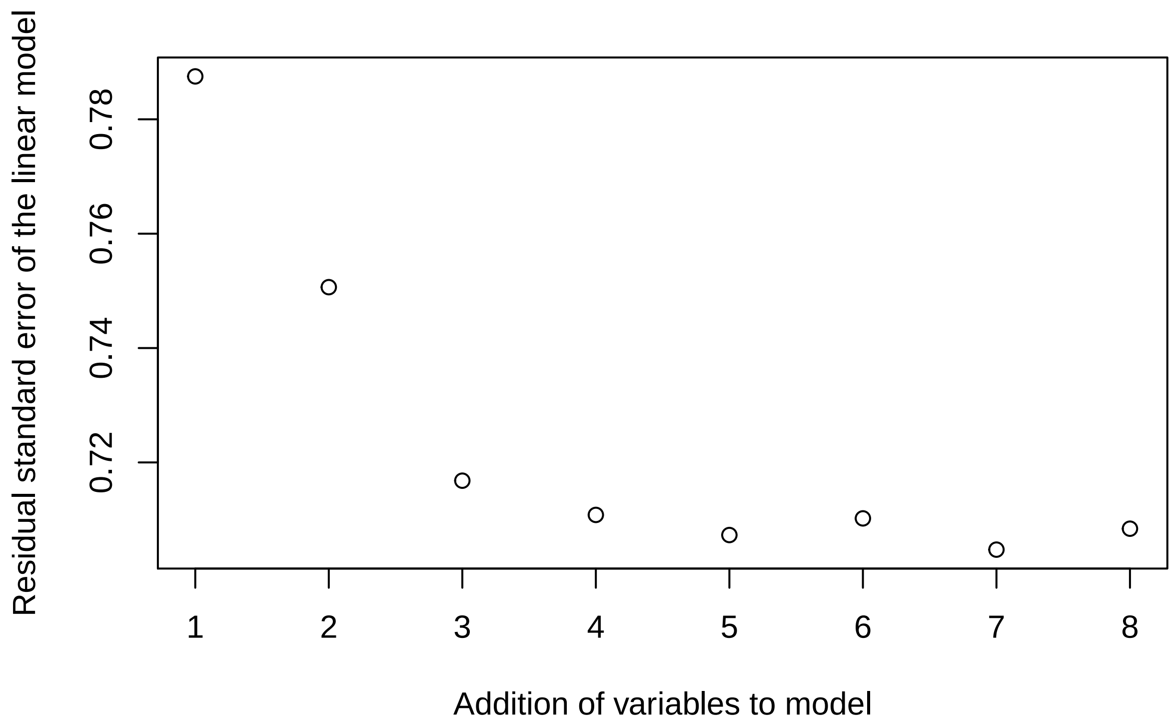
```
rsq[8]=summary(pros.mod)$r.squared
res<-summary(pros.mod)$residuals
RSS<-c(crossprod(res))
MSE<-RSS/(length(res)-9)
rse[8]=sqrt(MSE)
```

Plotting RSE and R2.

```
plot(rsq,xlab="Addition of variables to model",ylab="R-square of the linear model")
```




```
plot(rse,xlab="Addition of variables to model",ylab="Residual standard error of the linear model")
```

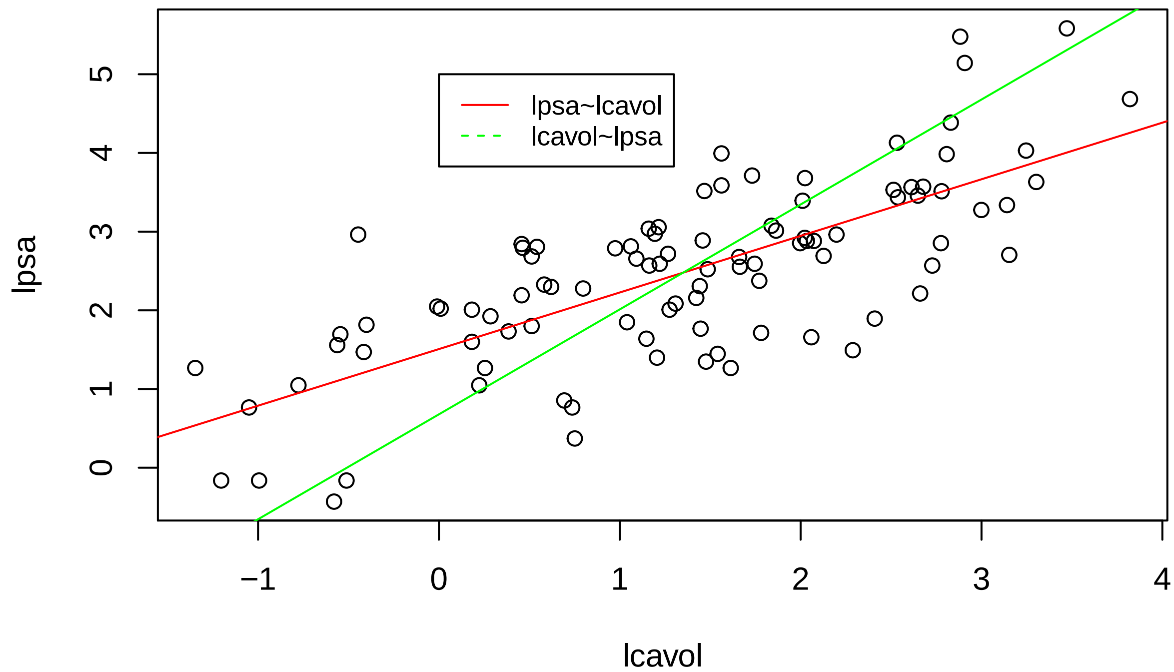


Chapter 2 Exercise 5 Using the prostate data, plot lpsa against lcavol. Fit the regressions of lpsa on lcavol and lcavol on lpsa. Display both regression lines on the plot. At what point do the two lines intersect?

Solution:

This is solving for the intesection of two lines: (y=m1x+c1) and(x=m2y+c2). The intersection will be (xmean,ymean) since both linear models will necessarily pass through this points.

```
data(prostate,package='faraway')
plot(lpsa~lcavol)
lmod1<-lm(lpsa~lcavol)
lmod2<-lm(lcavol~lpsa)
abline(lmod1,col="red")
abline(-lmod2$coefficients[1]/lmod2$coefficients[2],1/lmod2$coefficients[2],col="green")
legend(0, 5, legend=c("lpsa~lcavol", "lcavol~lpsa"),
       col=c("red", "green"), lty=1:2, cex=0.8)
```




```
if (!require("matlib")){
  install.packages("matlib")
}
```

```
## Loading required package: matlib
```

```
## Warning: package 'matlib' was built under R version 3.4.4
```

```
library(matlib)
A <- matrix(c(-lmod1$coefficients[2],1,1,-lmod2$coefficients[2]), 2, 2)
b<-c(lmod1$coefficients[1],lmod2$coefficients[1])
Solve(A,b)
```

```
## x1    =  1.35000958
##   x2  =  2.47838701
```

```
sprintf("mean_lcavol is %f",mean(lcavol))
```

```
## [1] "mean_lcavol is 1.350010"
```

```
sprintf("meanlpsa is %f",mean(lpsa))
```

```
## [1] "meanlpsa is 2.478387"
```

As can be seen x1=mean(lcavol) and x2=mean(lpsa)