Problem 2

```
> library(faraway)
> data(prostate)
> nrow(prostate)
[1] 97
> save(prostate,file="prostate.Rdata")
```

(a)

```
> fit = lm(lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45)
> summary(fit)

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
    gleason + pgg45)

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 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886
---
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
```

(b)

Null hypothesis Ho:

Alternative hypothesis Ha:

```
> fit0 = lm(lpsa~1)   # Null model
> anova(fit0,fit)     # anova(Nullmodel,Fullmodel)
Analysis of Variance Table

Model 1: lpsa ~ 1
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     96 127.918
2     88  44.163  8    83.755 20.861 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F test statistic = 20.861, and p-value < $2.2*10^{-16}$ < α =0.05 .

Therefore, Reject Ho at α = 5% significance level.

(c)

90% Confidence Interval:

```
> confint(fit,"age",level=0.90)
             5 %          95 %
age -0.0382102 -0.001064151
```

95% Confidence Interval:

```
> confint(fit,"age",level=0.95)
           2.5 %      97.5 %
age -0.04184062 0.002566267
```

For regression summary: Ho: Parameter(age) = 0

90% CI does not cover 0, Reject Ho, so p-value < 0.1

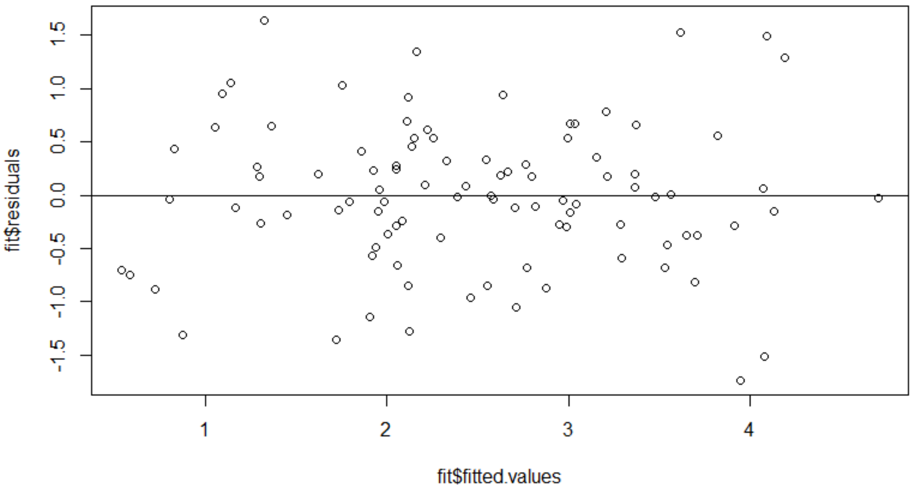
95% CI covers 0, Do Not Reject Ho, so p-value > 0.05

Therefore, we can deduce that 0.05 < p-value < 0.1

Indeed, the regression summary shows that the parameter associated with age has p-value = 0.08229.

(d)

```
> plot(fit$fitted.values,fit$residuals)
> abline(h = 0)
```



```
> library(lmtest)
> bptest(fit)

        studentized Breusch-Pagan test

data:  fit
BP = 10.08, df = 8, p-value = 0.2594
```

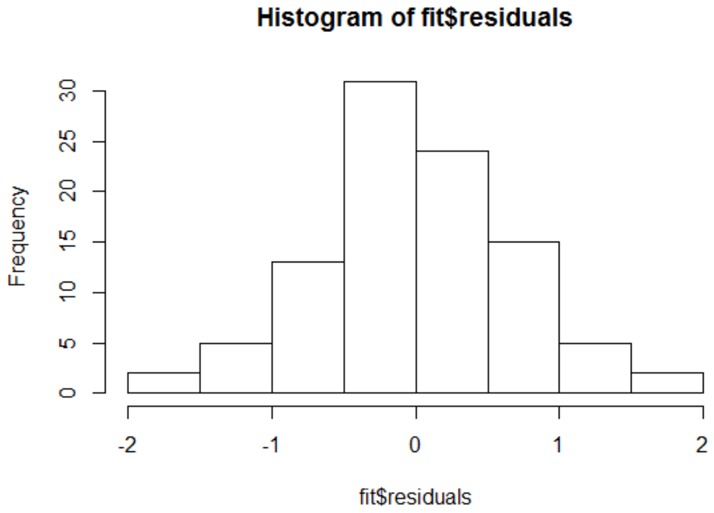
Result and Comment:

Fitted vs. Residual Plot does not seem to show any obvious pattern in the variance of the residuals, thus not showing clear evidence for a non-constant variance.

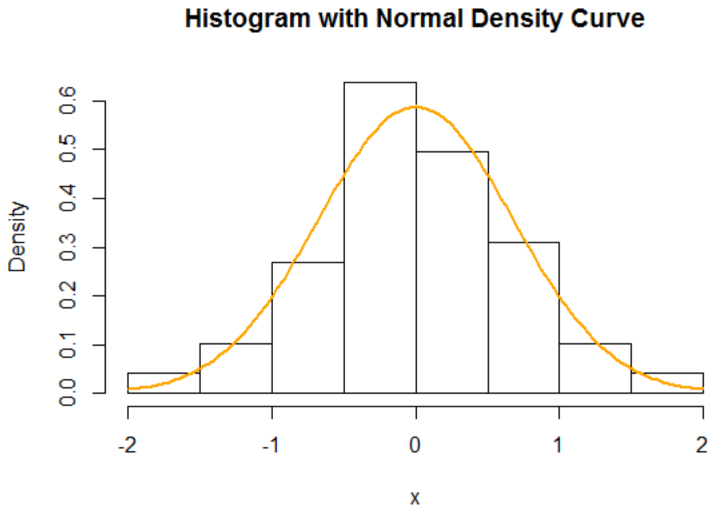
And the Breusch-Pagan test has a p-value=0.2594 > 0.1, Do Not Reject Null Hypothesis (homoscedasticity) at α = 10% significance level or smaller. Therefore, the residuals' variance is basically constant.

(e)

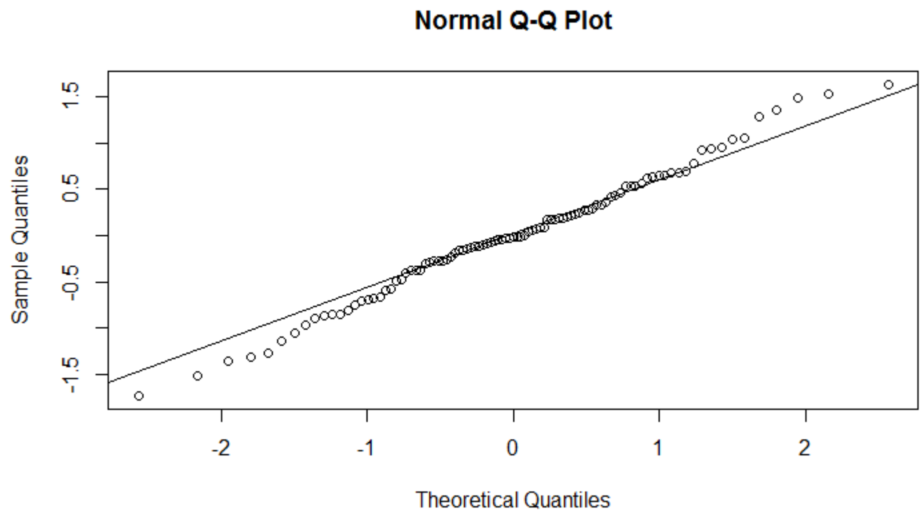
```
> hist(fit$residuals)
```



```
> histNorm <- function(x, densCol = "darkblue"){
+   m <- mean(x)
+   std <- sqrt(var(x))
+   h <- max(hist(x,plot=FALSE)$density)
+   d <- dnorm(x, mean=m, sd=std)
+   maxY <- max(h,d)
+   hist(x, prob=TRUE,
+        xlab="x", ylim=c(0, maxY),
+        main="Histogram with Normal Density Curve")
+   curve(dnorm(x, mean=m, sd=std),
+         col=densCol, lwd=2, add=TRUE)
+ }
> histNorm(fit$residuals, "orange")
```



```
> qqnorm(fit$residuals)
> qqline(fit$residuals)
```



```
> shapiro.test(fit$residuals)

        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.99113, p-value = 0.7721
```

Result and Comment:

Histogram and Normal Q-Q Plot shows some evidence that the residuals could be from a normal distribution, but does not fit a normal distribution very well.

And the Shapiro-Wilk test has a p-value=0.7721 > 0.1, Do Not Reject Null Hypothesis at α = 10% significance level or smaller. Therefore, the residuals' data basically follow a normal distribution.

(f)

Two ways to calculate and find large leverages:

```
> diag_H = hatvalues(fit)
> sum(diag_H)
[1] 9
> 2*mean(diag_H)
[1] 0.185567
> sum(diag_H>2*mean(diag_H))
[1] 5
> diag_H[which(diag_H>2*mean(diag_H))]
       32        37        41        74        92 
0.3304757 0.2184392 0.2410079 0.1912109 0.2092421
```

OR

```
> X = cbind(rep(1,97), prostate[,1:8])
> X = as.matrix(X)
> H = X %*% solve(t(X)%*%X) %*% t(X)
> sum(diag(H))
[1] 9
> 2*mean(diag(H))
[1] 0.185567
> sum(diag(H)>2*mean(diag(H)))
[1] 5
> diag(H)[which(diag(H)>2*mean(diag(H)))]
       32        37        41        74        92 
0.3304757 0.2184392 0.2410079 0.1912109 0.2092421
```

Result:

There are five observations with large leverage, they are the 32th, 37th, 41th, 74th, and 92th points.

And their values are shown above.

(g)

Two ways to find potential outliers:

```
> stu_r = rstudent(fit)
> hist(stu_r)
> max(abs(stu_r))
[1] 2.61698
> n = length(stu_r)
> p = length(fit$coefficients)
> df = n-p-1
> alpha = 0.05 ## Here we use 5% significance level to perform the t-test
```

Without Bonferroni adjustment:

```
> t = qt(1-alpha/2,df)
> sum(abs(stu_r)>t)
[1] 6
> stu_r[abs(stu_r)>t]
       39        47        69        81        95        97 
-2.616980 -2.376671  2.553530  1.987947  2.385070  2.293279 
```

With Bonferroni adjustment:

```
> t_B = qt(1-(alpha/2)/n,df)
> sum(abs(stu_r)>t_B)
[1] 0
```

OR

```
> library(car)
> outlierTest(fit)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
39 -2.61698            0.01046           NA
```

Result:

There are 6 potential outliers without Bonferroni adjustment, and their values are shown above.

After Bonferroni adjustment, there are no outliers at α = 5% significance level.

(h)

```
> cook = cooks.distance(fit)
> sum(cook>4/n)
[1] 7
> cook[cook>4/n]
        32         39         47         69         95         96         97 
0.12269771 0.05201916 0.10574362 0.10053751 0.09873809 0.05593862 0.07377558 
```

Result:

There are seven influential observations with a large Cook's Distance.

And their Cook's Distances are shown above.

(i)

In the full model, only the parameters associated with predictors "lcavol", "lweight", and "svi" .

have p-value < 0.05, which are significant at α = 5% significance level. Therefore, set up the new, smaller model only with these tree predictors.

```
> fit_new = lm(lpsa~lcavol+lweight+svi)
> anova(fit_new,fit)
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lweight + svi
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + 
    pgg45
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     93 47.785                           
2     88 44.163  5    3.6218 1.4434 0.2167
```

F test statistic = 1.4434, and p-value = 0.2167 > 0.1

Therefore, Do not Reject Ho (Null Model) at α = 10% or smaller significance level.

The new, smaller model is preferred, which is only explained by predictors “lcavol”, “lweight”, and “svi” .