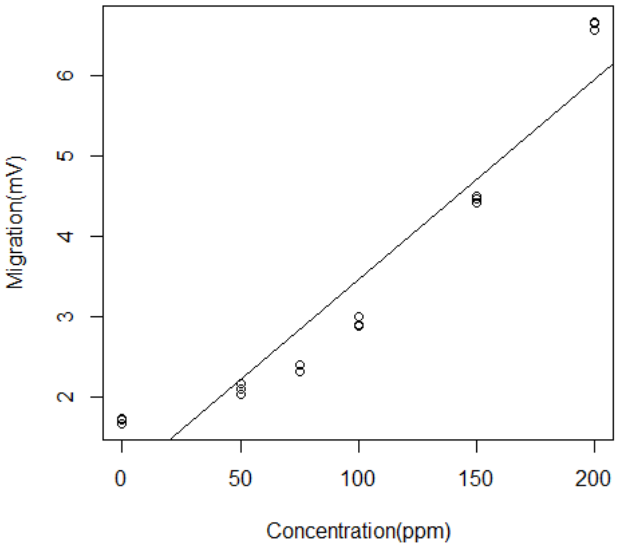
Problem 1

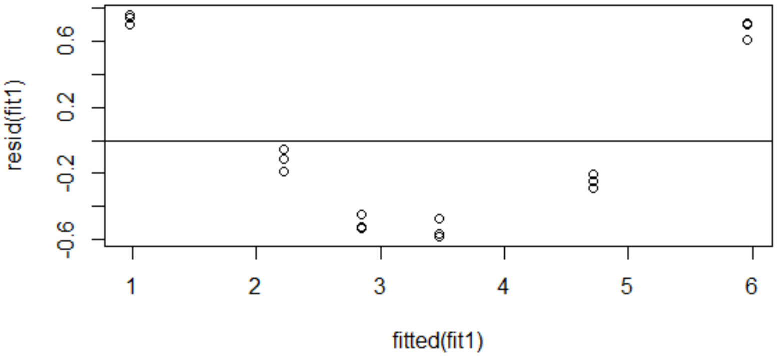
```
> ISEs=read.table(file="HW10_ISEs.dat",header=TRUE)
> colnames(ISEs)=c("concentration","migration")
> head(ISEs,n=3)
  concentration migration
1             0      1.72
2             0      1.68
3             0      1.74
```

(a)

```
> fit1=lm(migration~concentration,data = ISEs)
> plot(ISEs$concentration,ISEs$migration,xlab = "Concentration(ppm)",ylab = "
Migration(mV)")
> abline(fit1)
```



```
> plot(fitted(fit1),resid(fit1))
> abline(h=0)
```



From both plots above, we can conclude that the linear model does NOT seem to be appropriate.

(b)

Quadratic model: $migration_i = ß_0 + ß_1 * concentration_i + ß_2 * (concentration_i \text{^} 2) + e_i$

Null Hypothesis Ho: $ß_2 = 0$

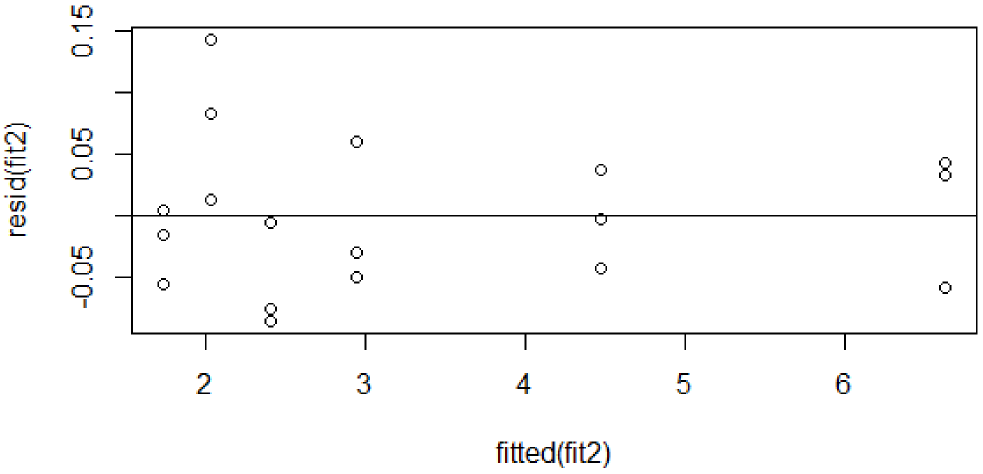
```
> fit2=lm(migration~concentration+I(concentration^2),data = ISEs)
> anova(fit1,fit2)
Analysis of Variance Table

Model 1: migration ~ concentration
Model 2: migration ~ concentration + I(concentration^2)
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)    
1     16 4.8323                                
2     15 0.0603  1     4.772 1187 1.073e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F test statistic = 1187, and under the null hypothesis, the test statistic has an F distribution with 1 and 15 degrees freedom.

p-value = 1.073e-15 << 0.05, Reject Ho (Null Model) at α = 0.05. Hence the new, quadratic model is much better at 5% or even much lower significance level.

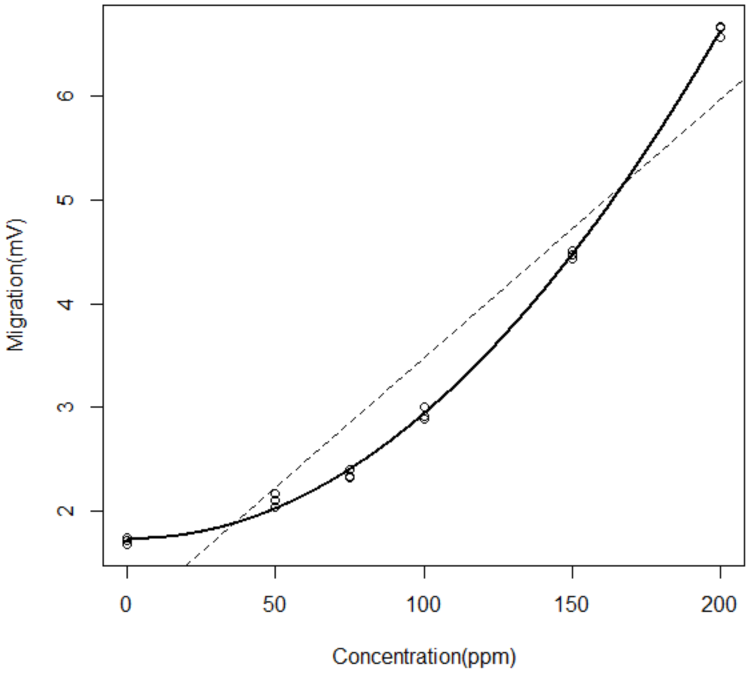
Also, we can see this result by looking at the plot of residuals vs. fitted values as shown below.

```
> plot(fitted(fit2),resid(fit2))
> abline(h=0)
```



(c)

```
> plot(ISEs$concentration,ISEs$migration,xlab = "Concentration(ppm)",ylab = "
Migration(mV)")
> abline(fit1,lty=2)
> xplot=seq(0,200,by=0.1)
> lines(xplot, predict(fit2,newdata=data.frame(concentration=xplot)),lwd=2)
```



Problem 2

```
> library(faraway)
> data(odor)
> nrow(odor)
[1] 15
> odor[1:3,]
  odor temp gas pack
1   66   -1  -1    0
2   39    1  -1    0
3   43   -1   1    0
```

(a)

```
> fit=lm(odor~temp+gas+pack+I(temp^2)+I(gas^2)+I(pack^2),data=odor)
> summary(fit)

Call:
lm(formula = odor ~ temp + gas + pack + I(temp^2) + I(gas^2) +
    I(pack^2), data = odor)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-20.625  -9.625  -1.375   4.021  28.875 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -30.667     10.840  -2.829   0.0222 * 
temp         -12.125      6.638  -1.827   0.1052   
gas          -17.000      6.638  -2.561   0.0336 * 
pack         -21.375      6.638  -3.220   0.0122 * 
I(temp^2)     32.083      9.771   3.284   0.0111 * 
I(gas^2)      47.833      9.771   4.896   0.0012 **
I(pack^2)      6.083      9.771   0.623   0.5509   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.77 on 8 degrees of freedom
Multiple R-squared:  0.8683,	Adjusted R-squared:  0.7695 
F-statistic: 8.789 on 6 and 8 DF,  p-value: 0.003616
```

F test statistic = 8.789, and under the null hypothesis, the test statistic has an F distribution with 6 and 8 degrees freedom.

p-value = 0.003616 < 0.05, hence Reject Ho at α = 0.05. Hence this quadratic model is significant at 5% significance level.

(b)

```
> names(summary(fit))
 [1] "call"          "terms"         "residuals"     "coefficients"  "aliased
"     
 [6] "sigma"         "df"            "r.squared"     "adj.r.squared" "fstatis
tic"   
[11] "cov.unscaled" 
> summary(fit)$r.squared
[1] 0.8682799
```

86.83% of the observed variation of odor is explained by the model in part (a).

(c)

```
> summary(fit)$adj.r.squared
[1] 0.7694898

> extractAIC(fit)
[1]  7.00000 92.54619
```

Problem 3

```
> library(faraway)
> data(prostate)
> nrow(prostate)
[1] 97
> prostate[1:3,]
      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


> fit=lm(lpsa~.,data=prostate)
> summary(fit)

Call:
lm(formula = lpsa ~ ., data = prostate)

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
```

(a)

(a-i) Backward AIC

```
> n=length(resid(fit))
> fit_bac_AIC=step(fit,direction = "backward");fit_bac_AIC
Start:  AIC=-58.32
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + 
    pgg45

          Df Sum of Sq    RSS     AIC
- gleason  1    0.0412 44.204 -60.231
- pgg45    1    0.5258 44.689 -59.174
- lcp      1    0.6740 44.837 -58.853
<none>                 44.163 -58.322
- age      1    1.5503 45.713 -56.975
- lbph     1    1.6835 45.847 -56.693
- lweight  1    3.5861 47.749 -52.749
- svi      1    4.9355 49.099 -50.046
- lcavol   1   22.3721 66.535 -20.567
```

```
Step:  AIC=-60.23
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

          Df Sum of Sq    RSS     AIC
- lcp      1    0.6623 44.867 -60.789
<none>                 44.204 -60.231
- pgg45    1    1.1920 45.396 -59.650
- age      1    1.5166 45.721 -58.959
- lbph     1    1.7053 45.910 -58.560
- lweight  1    3.5462 47.750 -54.746
- svi      1    4.8984 49.103 -52.037
- lcavol   1   23.5039 67.708 -20.872

Step:  AIC=-60.79
lpsa ~ lcavol + lweight + age + lbph + svi + pgg45

          Df Sum of Sq    RSS     AIC
- pgg45    1    0.6590 45.526 -61.374
<none>                 44.867 -60.789
- age      1    1.2649 46.131 -60.092
- lbph     1    1.6465 46.513 -59.293
- lweight  1    3.5647 48.431 -55.373
- svi      1    4.2503 49.117 -54.009
- lcavol   1   25.4189 70.285 -19.248

Step:  AIC=-61.37
lpsa ~ lcavol + lweight + age + lbph + svi

          Df Sum of Sq    RSS     AIC
<none>                 45.526 -61.374
- age      1    0.9592 46.485 -61.352
- lbph     1    1.8568 47.382 -59.497
- lweight  1    3.2251 48.751 -56.735
- svi      1    5.9517 51.477 -51.456
- lcavol   1   28.7665 74.292 -15.871

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)

Coefficients:
(Intercept)       lcavol      lweight          age         lbph          svi  

    0.95100      0.56561      0.42369     -0.01489      0.11184      0.72095  
```

Best model: lpsa ~ lcavol + lweight + age + lbph + svi

(a-ii) Backward BIC

```
> fit_bac_BIC=step(fit,direction = "backward",k=log(n));fit_bac_BIC
Start:  AIC=-35.15
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + 
    pgg45

          Df Sum of Sq    RSS     AIC
- gleason  1    0.0412 44.204 -39.634
- pgg45    1    0.5258 44.689 -38.576
- lcp      1    0.6740 44.837 -38.255
- age      1    1.5503 45.713 -36.377
- lbph     1    1.6835 45.847 -36.095
<none>                 44.163 -35.149
- lweight  1    3.5861 47.749 -32.151
- svi      1    4.9355 49.099 -29.448
- lcavol   1   22.3721 66.535   0.030

Step:  AIC=-39.63
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

          Df Sum of Sq    RSS     AIC
- lcp      1    0.6623 44.867 -42.766
- pgg45    1    1.1920 45.396 -41.627
- age      1    1.5166 45.721 -40.936
- lbph     1    1.7053 45.910 -40.537
<none>                 44.204 -39.634
- lweight  1    3.5462 47.750 -36.723
- svi      1    4.8984 49.103 -34.014
- lcavol   1   23.5039 67.708  -2.849

Step:  AIC=-42.77
lpsa ~ lcavol + lweight + age + lbph + svi + pgg45

          Df Sum of Sq    RSS     AIC
- pgg45    1    0.6590 45.526 -45.926
- age      1    1.2649 46.131 -44.644
- lbph     1    1.6465 46.513 -43.844
<none>                 44.867 -42.766
- lweight  1    3.5647 48.431 -39.925
- svi      1    4.2503 49.117 -38.561
- lcavol   1   25.4189 70.285  -3.800

Step:  AIC=-45.93
lpsa ~ lcavol + lweight + age + lbph + svi

          Df Sum of Sq    RSS     AIC
- age      1    0.9592 46.485 -48.478
- lbph     1    1.8568 47.382 -46.623
<none>                 45.526 -45.926
- lweight  1    3.2251 48.751 -43.862
- svi      1    5.9517 51.477 -38.583
- lcavol   1   28.7665 74.292  -2.997

Step:  AIC=-48.48
lpsa ~ lcavol + lweight + lbph + svi

          Df Sum of Sq    RSS     AIC
- lbph     1    1.3001 47.785 -50.377
<none>                 46.485 -48.478
- lweight  1    2.8014 49.286 -47.377
- svi      1    5.8063 52.291 -41.636
- lcavol   1   27.8298 74.315  -7.542

Step:  AIC=-50.38
lpsa ~ lcavol + lweight + svi

          Df Sum of Sq    RSS     AIC
<none>                 47.785 -50.377
- svi      1    5.1814 52.966 -44.966
- lweight  1    5.8924 53.677 -43.673
- lcavol   1   28.0445 75.829 -10.160

Call:
lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
```

```
Coefficients:
(Intercept)       lcavol      lweight          svi
    -0.2681       0.5516       0.5085       0.6662
```

Best model: lpsa ~ lcavol + lweight + svi

(b)

(b-i) Forward AIC

```
> n=length(resid(fit))
> fit_for_AIC=step(fit,direction = "forward");fit_for_AIC
Start:  AIC=-58.32
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45


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

Coefficients:
(Intercept)       lcavol      lweight          age         lbph          svi
        lcp      gleason
   0.669337     0.587022     0.454467    -0.019637     0.107054     0.766157
  -0.105474     0.045142
      pgg45
   0.004525
```

Best model: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45

(b-ii) Forward BIC

```
> fit_for_BIC=step(fit,direction = "forward",k=log(n));fit_for_BIC
Start:  AIC=-35.15
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45


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

Coefficients:
(Intercept)       lcavol      lweight          age         lbph          svi
        lcp      gleason
   0.669337     0.587022     0.454467    -0.019637     0.107054     0.766157
  -0.105474     0.045142
      pgg45
   0.004525
```

Best model: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45

(c)

```
> library(leaps)
> all_fits=regsubsets(lpsa~.,data=prostate)
> all_fits_sum=summary(all_fits)
> all_fits_sum$which
  (Intercept) lcavol lweight   age  lbph   svi   lcp gleason pgg45
1        TRUE   TRUE   FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
2        TRUE   TRUE    TRUE FALSE FALSE FALSE FALSE   FALSE FALSE
3        TRUE   TRUE    TRUE FALSE FALSE  TRUE FALSE   FALSE FALSE
4        TRUE   TRUE    TRUE FALSE  TRUE  TRUE FALSE   FALSE FALSE
5        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE FALSE
6        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE  TRUE
7        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE   FALSE  TRUE
8        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE    TRUE  TRUE
```

(c-i) “leaps” AIC

```
> p=length(coef(fit))
> n=length(resid(fit))
> AIC=n*log(all_fits_sum$rss/n) + 2*(2:p)
> AIC
[1] -44.36608 -52.69043 -60.67621 -61.35179 -61.37439 -60.78867 -60.23130 -5
8.32184
> which.min(AIC)
[1] 5
```

Best model #5: lpsa ~ lcavol + lweight + age + lbph + svi

(c-ii) “leaps” BIC

```
> BIC=n*log(all_fits_sum$rss/n) + log(n)*(2:p)
> BIC
[1] -39.21666 -44.96629 -50.37736 -48.47824 -45.92613 -42.76570 -39.63361 -3
5.14944
> which.min(BIC)
[1] 3
```

Best model #3: lpsa ~ lcavol + lweight + svi

(c-iii) “leaps” Adjusted R^2

```
> R_adj=all_fits_sum$adjr2
> R_adj
[1] 0.5345838 0.5771246 0.6143899 0.6208036 0.6245476 0.6258707 0.6272521 0.6
233681
> which.max(R_adj)
[1] 7
```

Best model #7: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

(d)

The seven best models from above (a)-(c), some of them are the same:

Method	Model	Criterions
Forward AIC	lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45	AIC=-58.32
Forward BIC	lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45	BIC=-35.15
"leaps" Adjusted R^2	lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45	62.73%
Backward AIC	lpsa ~ lcavol + lweight + age + lbph + svi	AIC=-61.37
"leaps" AIC	lpsa ~ lcavol + lweight + age + lbph + svi	AIC=-61.37
Backward BIC	lpsa ~ lcavol + lweight + svi	BIC=-50.38
"leaps" BIC	lpsa ~ lcavol + lweight + svi	BIC=-50.38

Hence we only to compare four models:

Model 1: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, AIC=-58.32, BIC=-35.15

Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45 , Adjusted R^2=62.73%

Model 3: lpsa ~ lcavol + lweight + age + lbph + svi , AIC=-61.37

Model 4: lpsa ~ lcavol + lweight + svi , BIC=-50.38

```
> fit1=lm(lpsa~.,data=prostate)
> fit2=lm(lpsa~.-gleason,data=prostate)
> fit3=lm(lpsa~lcavol+lweight+age+lbph+svi,data=prostate)
> fit4=lm(lpsa~lcavol+lweight+svi,data=prostate)
```

(1) AIC

```
> rbind(
+   extractAIC(fit1),
+   extractAIC(fit2),
+   extractAIC(fit3),
+   extractAIC(fit4)
+ )
     [,1]      [,2]
[1,]    9 -58.32184
[2,]    8 -60.23130
[3,]    6 -61.37439
[4,]    4 -60.67621
```

(2) Adjusted R^2

```
> rbind(
+   summary(fit1)$adj.r.squared,
+   summary(fit2)$adj.r.squared,
+   summary(fit3)$adj.r.squared,
+   summary(fit4)$adj.r.squared
+ )
```

```
          [,1]
[1,] 0.6233681
[2,] 0.6272521
[3,] 0.6245476
[4,] 0.6143899
```

(3) PRESS

```
> rbind(
+   sum( (resid(fit1) / (1 - hatvalues(fit1)))^2 ),
+   sum( (resid(fit2) / (1 - hatvalues(fit2)))^2 ),
+   sum( (resid(fit3) / (1 - hatvalues(fit3)))^2 ),
+   sum( (resid(fit4) / (1 - hatvalues(fit4)))^2 )
+ )
         [,1]
[1,] 54.23246
[2,] 53.27402
[3,] 52.67252
[4,] 52.84734
```

Overall Analysis:

The criterion “AIC” and “PRESS” both shows that the Model 3 is the best model.

On the other hand, the criterion “Adjusted R^2” shows that the Model 2 is the best model. However, the adjusted R squared value of Model 3 is similar with Model 2.

Hence, considering the overall effectiveness, the Model 3 is the best model as shown below:

Model 3:   lpsa ~ lcavol + lweight + age + lbph + svi