Assignment 2

STAT2008

u6408405

Question 1 part (a)

load the table and read the first 3 rows

In [1]:

```
senic = read.table('senic.txt', header=TRUE)
head(senic, 3)
```

Id	Length	Age	Risk	Culture	Xray	Beds	Affiliation	Region	Patients	Nurses	Facilities
1	7.13	55.7	4.1	9.0	39.6	279	2	4	207	241	60
2	8.82	58.2	1.6	3.8	51.7	80	2	2	51	52	40
3	8.34	56.9	2.7	8.1	74.0	107	2	3	82	54	20

The dataframe has 113 rows, each row representing a hospital; and 12 variables.

In [2]:

```
dim(senic)
```

113 12

Fit the multilinear regression model

In [3]:

```
m1 <- lm(Risk ~ Length + Age + Culture + Xray + Beds + Affiliation
         + Region + Patients + Nurses + Facilities, senic)
```

The estimated regression coefficient of the model is given by the following

In [4]:

```
coef(m1)
```

(Intercept)
-3.23999167606266
Length
0.242239872622796
Age
0.0100770347894278
Culture
0.0534439329430643
Xray
0.0126320790759262
Beds
-0.00317319552798139
Affiliation
0.562186183939178
Region
0.297558456037447
Patients
0.00282888832146191
Nurses
0.00206398797509475
Facilities
0.0232722329099752

The anova table gives information about the degree of freedom of SSR and SSE

In [5]:

```
anova(m1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Length	1	57.3051098	57.3051098	67.4590207	7.027380e-13
Age	1	2.0750596	2.0750596	2.4427401	1.211672e-01
Culture	1	31.4573472	31.4573472	37.0312847	2.055320e-08
Xray	1	3.8476488	3.8476488	4.5294149	3.572596e-02
Beds	1	6.5164189	6.5164189	7.6710652	6.667769e-03
Affiliation	1	1.1547556	1.1547556	1.3593671	2.463662e-01
Region	1	4.4202912	4.4202912	5.2035241	2.461983e-02
Patients	1	0.7631874	0.7631874	0.8984168	3.454468e-01
Nurses	1	2.6178349	2.6178349	3.0816899	8.218037e-02
Facilities	1	4.5751780	4.5751780	5.3858553	2.229105e-02
Residuals	102	86.6469916	0.8494803	NA	NA

Use F test to ascertain if the model is significant. The test statistic used in the F test is denoted by the variable F_star and is equal to F_star = MSR / MSE

The null hypothesis and the alternative hypothesis are as follows: H0: all regression coefficients are equal to zero. Ha: not all regression coefficients are equal to zero

When the null hypothesis is true, the test statistic (F_star) is going to follow an F distribution with 10 and 102 degrees of freedom

I start by calculating SSR, MSE, SSE and MSE

In [6]:

```
SSR <- with(senic,
            sum((m1$fitted.values - mean(Risk)) ** 2))
MSR <- SSR / 10
paste('The value of MSR is ', MSR)

e <- m1$residuals
SSE <- t(e) %*% e
MSE <- SSE / m1$df.residual
paste('The value of MSE is ', MSE)
```

'The value of MSR is 11.4732831413103'

'The value of MSE is 0.849480309762219'

I then calculate F_star

In [7]:

```
F_star <- MSR / MSE
paste('The value of F_star is ', F_star)
```

'The value of F_star is 13.5062378838679'

To control type I error, I specify alpha to be 0.05. And calculate the cutoff point for an F test

In [8]:

```
alpha <- 0.05
F_cutoff <- qf(1 - alpha, 10, m1$df.residual)
paste('The cutoff point for the F test is', F_cutoff)
```

'The cutoff point for the F test is 1.92477794141672'

Since the value of F_star (13.5) is greater than the cutoff value of 1.925, I reject H0 and conclude that not all regression parameters are equal to zero and that the model is significant

----------end of question 1 part a----------

question 1 part b

The coefficient estimates and associated standard error are given by summary(m1)

In [9]:

```
summary(m1)
```

```
Call:
lm(formula = Risk ~ Length + Age + Culture + Xray + Beds + Affiliation + 
    Region + Patients + Nurses + Facilities, data = senic)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.86974 -0.56269 -0.02893  0.51925  2.32390 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.239992   1.413934  -2.291 0.023992 *  
Length       0.242240   0.070668   3.428 0.000879 ***
Age          0.010077   0.021500   0.469 0.640280    
Culture      0.053444   0.010545   5.068  1.8e-06 ***
Xray         0.012632   0.005280   2.392 0.018568 *  
Beds        -0.003173   0.002660  -1.193 0.235605    
Affiliation  0.562186   0.319765   1.758 0.081727 .  
Region       0.297558   0.104486   2.848 0.005324 ** 
Patients     0.002829   0.003430   0.825 0.411415    
Nurses       0.002064   0.001688   1.223 0.224125    
Facilities   0.023272   0.010028   2.321 0.022291 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9217 on 102 degrees of freedom
Multiple R-squared:  0.5697,    Adjusted R-squared:  0.5276 
F-statistic: 13.51 on 10 and 102 DF,  p-value: 7.974e-15
```

The estimated regression paramter of one of the covariates captures the marginal effect of a per unit increase in that covariate on the mean risk of nosocomial infection, holding other covariates constant.

However, when the covariates are highly correlated, predicting mean response outside of the range of observation could give imprecise results.

I use the body fat dataset from Wattle as an example.

In [313]:

```
bodyfat <- read.table('..\\bodyfat.txt', header=TRUE)
head(bodyfat, 2)
```

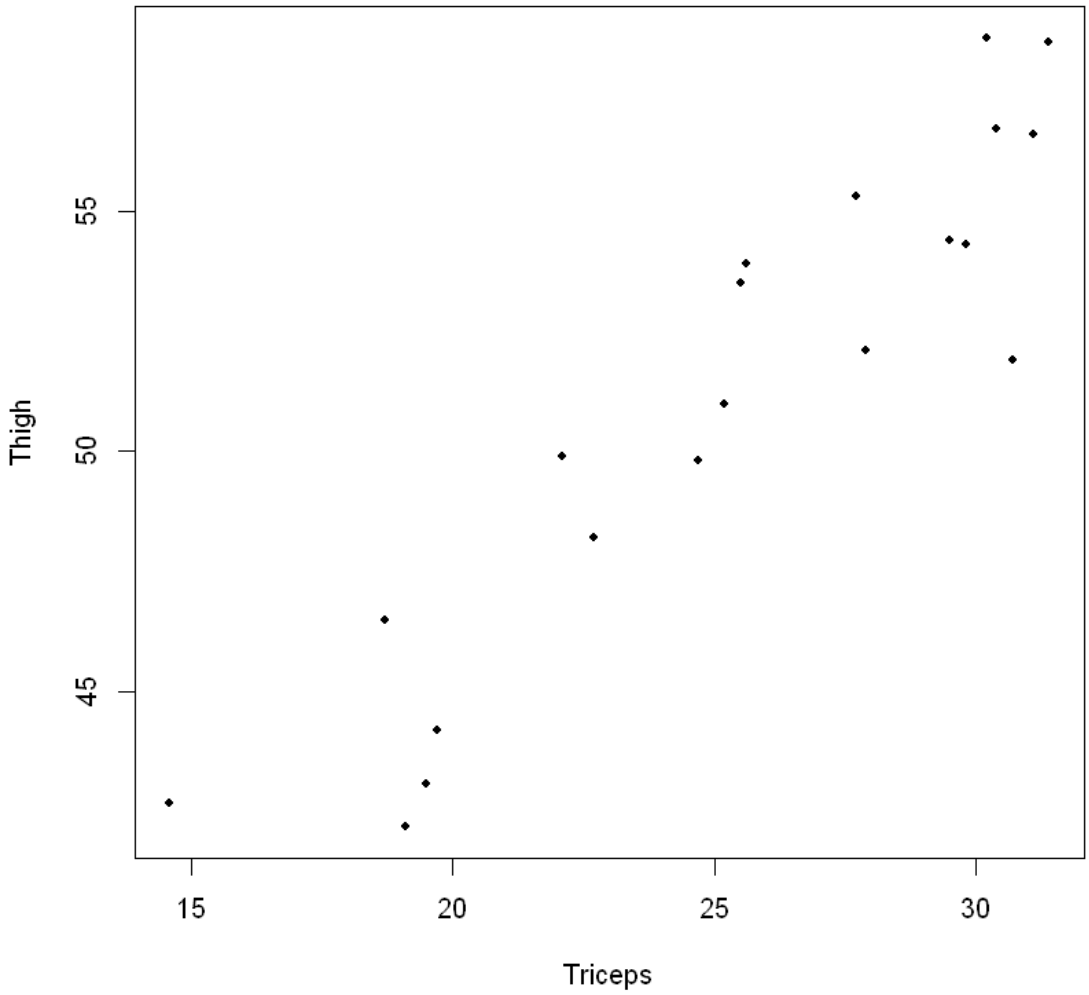
Triceps	Thigh	Midarm	Fat
19.5	43.1	29.1	11.9
24.7	49.8	28.2	22.8

The triceps circumference and thigh circumference are highly correlated.

In [11]:

```
paste('the correlation betweem Thigh and Triceps is ', with(bodyfat,
    cor(Triceps, Thigh)))
with(bodyfat, plot(Triceps, Thigh, pch=20, col='black'))
```

'the correlation betweem Thigh and Triceps is 0.92384251264551'



In [12]:

```
fat_triceps_thigh <- lm(Fat ~ Triceps + Thigh, bodyfat)
with(bodyfat,
     paste('the observed range of Triceps size is between', min(Triceps), 'and', max(Triceps)))
with(bodyfat,
     paste('the observed range of Thigh size is between', min(Thigh), 'and', max(Thigh)))
```

'the observed range of Triceps size is between 14.6 and 31.4'

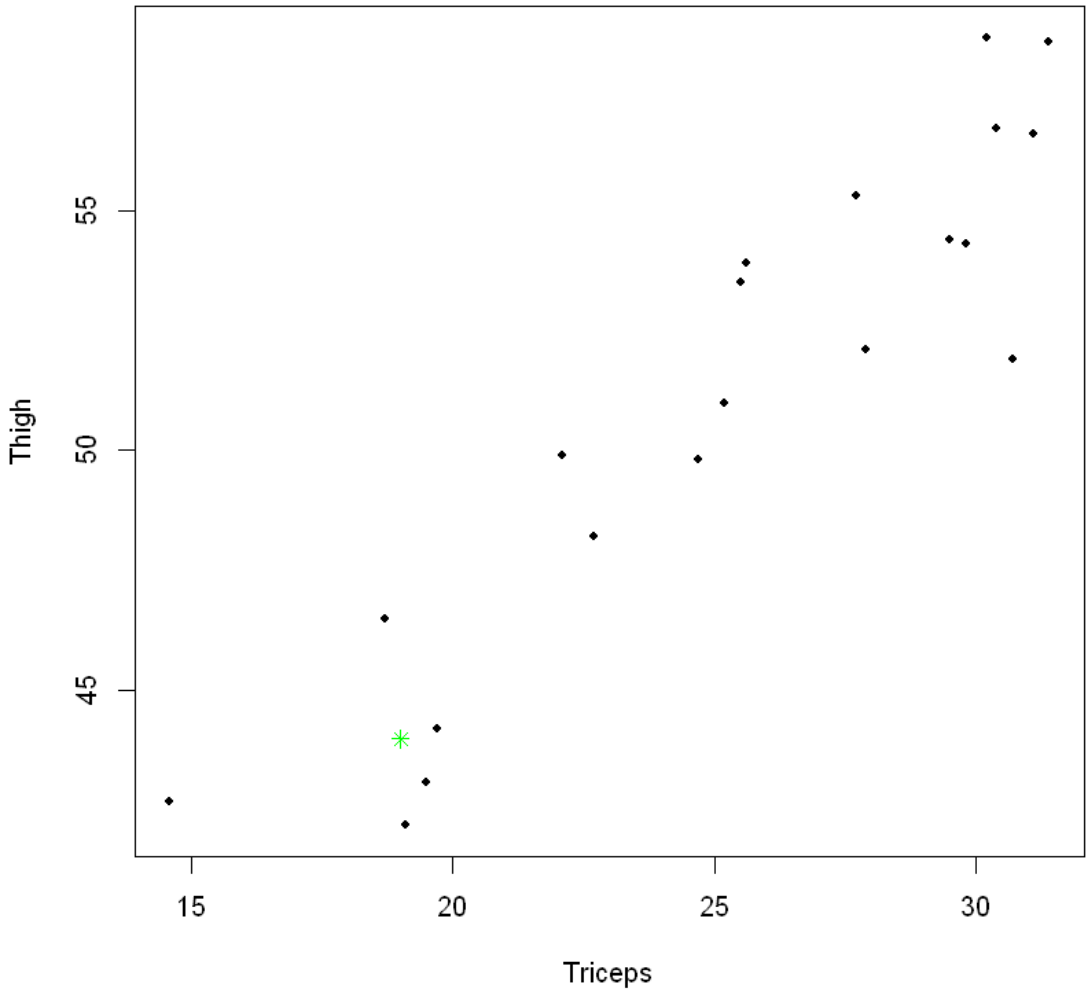
'the observed range of Thigh size is between 42.2 and 58.6'

I now find the fitted value using predictor values that lie within the observed range (i.e. Triceps=19, Thigh=44), I highlight the point to make it green on the scatterplot below

In [13]:

```
predict_within_the_range = predict(fat_triceps_thigh, data.frame(Triceps=19, Thigh=44), se.fit=TRUE)
paste('The predicted value is', predict_within_the_range$fit, ' and the standard error of the fitted value is',
      predict_within_the_range$se.fit)
with(bodyfat,
    plot(Triceps, Thigh, pch=20, col='black'))
points(data.frame(Triceps=19, Thigh=44), col='green', pch=8)
```

'The predicted value is 14.0650126324497 and the standard error of the fitted value is 0.980959891045791'

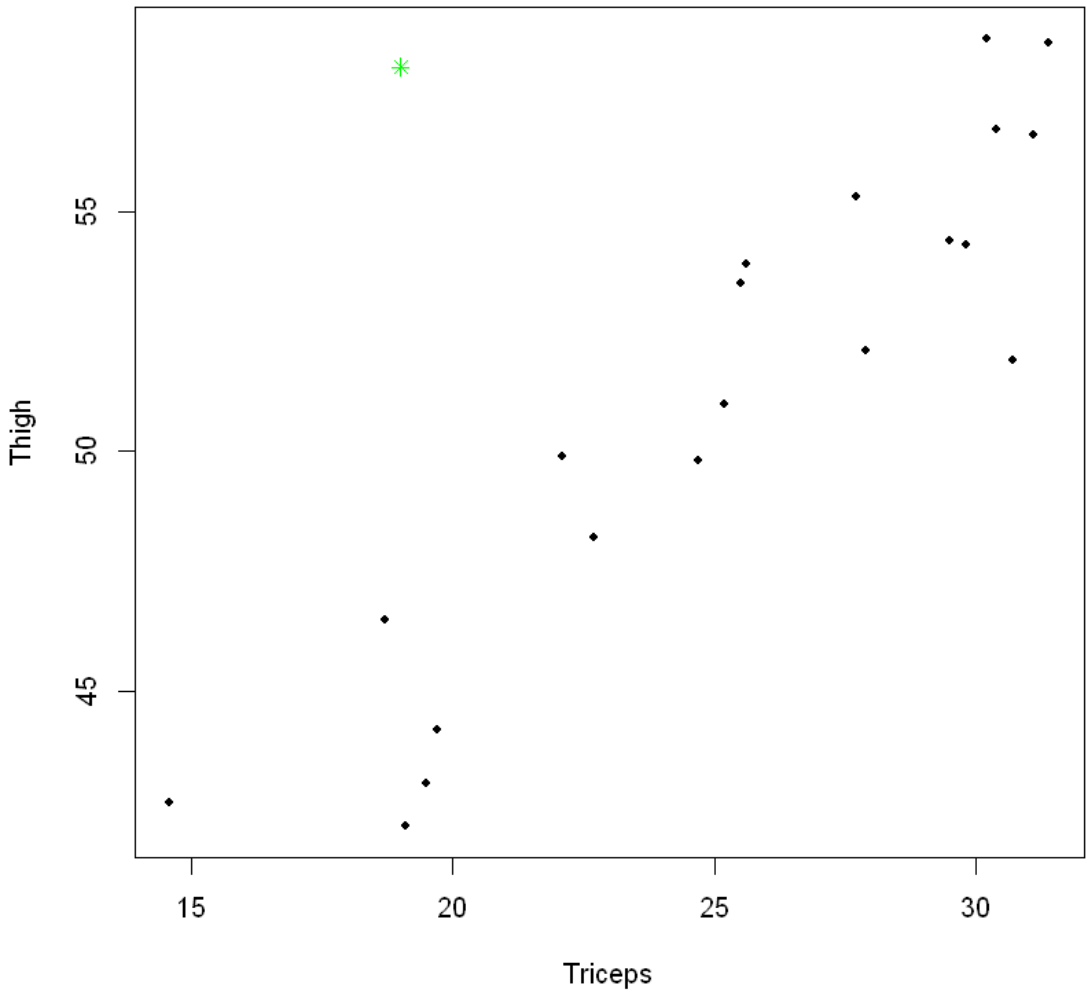


Now I find the fitted value using predictor values outside the observed range (i.e. Triceps=19, Thigh=58)

In [14]:

```
predict_outside_the_range = predict(fat_triceps_thigh, data.frame(Triceps=19, Thigh=58), se.fit=
TRUE)
paste('The predicted value is', predict_outside_the_range$fit, ' and the standard error of the f
itted value is',
      predict_outside_the_range$se.fit)
with(bodyfat,
     plot(Triceps, Thigh, pch=20, col='black'))
points(data.frame(Triceps=19, Thigh=58), col='green', pch=8)
```

'The predicted value is 23.2969177828607 and the standard error of the fitted value is 3.86902708544415'



The precision of the fitted value decreases drastically as I use predictor values outside the range of observation. This is a result of multicolinearity. In this question, while it is fair to claim 'the estimated regression paramter of one of the covariates captures the marginal effect of a per unit increase in that covariate on the mean risk of nosocomial infection, holding other covariates constant.', this interpretation may be of little use when trying to predict mean risk of infection using data outside of the range of observation.

----------end of question 1 part b----------

question 1 part c

A multiple linear regression model can be expressed as: E{Y} = b0 + b1X1 + ... + b(p-1)X(p-1). Each time we make an measurement, the estimated regression coefficients (b0, b1 ... b(p-1)) are going to change, these estimated regression coefficients are therefore variables. As variables, they follow a normal distribution with mean equal to the true model parameter denoted by β followed by its corresponding subcript number, and its corresponding variance given by its variance-covariance matrix. When we estimate the variance-covariance matrix using MSE, instead of true standard deviation of the error term, the estimated regression coefficients follows a t distribution with n - p degrees of freedom (where n is the number of observations taken and p is the number of estimated regression coefficients in the model).

The t-test associated with a estimated coefficient (bk) can be used to test whether we should reject the null hypothesis; the true regression parameter βk = 0; and accept the alternative hypothesis: the true regression parameter βk != 0. In a singple linear regression, the t-test tests whether there is a relationship between the covariate and the response. In a multiple linear regression however, the t-test associated with an estimated coefficient tests whether there is significant contribution to the model's ability in reducing the variability of the response by adding the covariate, when there are already other covariates in the model. In other words, in a MLR, the t-test on any coefficient is a marginal test, not a 'full test'.

In [314]:

```
summary(m1)
```

```
Call:
lm(formula = Risk ~ Length + Age + Culture + Xray + Beds + Affiliation + 
    Region + Patients + Nurses + Facilities, data = senic)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.86974 -0.56269 -0.02893  0.51925  2.32390 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.239992   1.413934  -2.291 0.023992 *  
Length       0.242240   0.070668   3.428 0.000879 ***
Age          0.010077   0.021500   0.469 0.640280    
Culture      0.053444   0.010545   5.068  1.8e-06 ***
Xray         0.012632   0.005280   2.392 0.018568 *  
Beds        -0.003173   0.002660  -1.193 0.235605    
Affiliation  0.562186   0.319765   1.758 0.081727 .  
Region       0.297558   0.104486   2.848 0.005324 ** 
Patients     0.002829   0.003430   0.825 0.411415    
Nurses       0.002064   0.001688   1.223 0.224125    
Facilities   0.023272   0.010028   2.321 0.022291 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9217 on 102 degrees of freedom
Multiple R-squared:  0.5697,    Adjusted R-squared:  0.5276 
F-statistic: 13.51 on 10 and 102 DF,  p-value: 7.974e-15
```

----------end of question 1 part c----------

question 1 part d

H0: βage = βbeds = 0 Ha: not both βage and βbeds = 0

I then fit the full model including beds and age as covariates, and the reduced model excluding the two covariats.

In [316]:

```
full_model <- lm(Risk ~ Length + Age + Culture + Xray + Beds + Affiliation
        + Region + Patients + Nurses + Facilities, senic)
reduced_model <- lm(Risk ~ Length + Culture + Xray + Affiliation
        + Region + Patients + Nurses + Facilities, senic)
```

I use SSR.F to denote the SSR under the full model and SSR.R to denote the SSR under the reduced model.

I use SSE.F to denote the SSE under the full model

In [317]:

```
SSR.F <- with(senic,
                sum((full_model$fitted.values - mean(Risk)) ** 2))
print(SSR.F)
SSR.R <- with(senic,
                 sum((reduced_model$fitted.values - mean(Risk)) ** 2))
print(SSR.R)
SSE.F <- with(senic,
                 sum(full_model$residuals ** 2))
print(SSE.F)
```

```
[1] 114.7328
[1] 113.3864
[1] 86.64699
```

I then calculate the test statistic F_star

In [318]:

```
F_star <- ((SSR.F - SSR.R) / 2) / (SSE.F / full_model$df.residual)
paste('F_star is equal to', F_star)
```

'F_star is equal to 0.792509496937781'

let the level of significance be alpha = 0.05

In [319]:

```
alpha = 0.05
cutoff <- qf(1-alpha, 2, full_model$df.residual)
cutoff
```

3.08546503257048

In [320]:

```
F_star > cutoff
```

FALSE

The test statistic does not fall in the rejection region. Therefore, do not reject H0, that beds and age are not significant contributors to the model.

anova function gives similar result.

In [321]:

```
anova(reduced_model, full_model)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
104	87.99343	NA	NA	NA	NA
102	86.64699	2	1.346442	0.7925095	0.4554748

----------end of question 1 part d----------

question 1 part e

The model suggested by my colleage would be the same as that fitted in part a.

In [322]:

```
suggested_model <- m1
summary(suggested_model)
```

```
Call:
lm(formula = Risk ~ Length + Age + Culture + Xray + Beds + Affiliation + 
    Region + Patients + Nurses + Facilities, data = senic)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.86974 -0.56269 -0.02893  0.51925  2.32390 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.239992   1.413934  -2.291 0.023992 *  
Length       0.242240   0.070668   3.428 0.000879 ***
Age          0.010077   0.021500   0.469 0.640280    
Culture      0.053444   0.010545   5.068  1.8e-06 ***
Xray         0.012632   0.005280   2.392 0.018568 *  
Beds        -0.003173   0.002660  -1.193 0.235605    
Affiliation  0.562186   0.319765   1.758 0.081727 .  
Region       0.297558   0.104486   2.848 0.005324 ** 
Patients     0.002829   0.003430   0.825 0.411415    
Nurses       0.002064   0.001688   1.223 0.224125    
Facilities   0.023272   0.010028   2.321 0.022291 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9217 on 102 degrees of freedom
Multiple R-squared:  0.5697,    Adjusted R-squared:  0.5276 
F-statistic: 13.51 on 10 and 102 DF,  p-value: 7.974e-15
```

The estimated intercept term for this model would be as follows:

In [323]:

```
paste('The estimated intercept term is', coef(suggested_model)[1])
```

'The estimated intercept term is -3.23999167606266'

This model is problematic because the 'Region' variable should be divided into four dummy variable each taking only 2 values (0 and 1). This is because Region variable takes categorical data, not ordinal data. In other words, there is no ordinal relationship between the regions. The appropriate model should be as follows

Risk ~ Length + Age + Culture + Xray + Beds + Affiliation + (NE + NC + S) + Patients + Nurses + Facilities (Assuming region 'W' is the baseline level to avoid dummy variable trap).

In the following cell, I fit the improved model.

In [324]:

```
n <- length(senic$Id)

NE <- data.frame(NE=rep(0, n))
NC <- data.frame(NC=rep(0, n))
S <- data.frame(S=rep(0,n))

new_senic <- cbind(senic, NE, NC, S)

for (i in 1:n){
    region = new_senic$Region
    if (region[i] == 1){
        new_senic$NE[i] <- 1
    } else if (region[i] == 2){
        new_senic$NC[i] <- 1
    } else if (region[i] == 3){
        new_senic$S[i] <- 1
    }
}

new_model <- lm(Risk ~ Length + Age + Culture + Xray + Beds + Affiliation
                + Patients + Nurses + Facilities + NE + NC + S, new_senic)
summary(new_model)
```

```
Call:
lm(formula = Risk ~ Length + Age + Culture + Xray + Beds + Affiliation +
    Patients + Nurses + Facilities + NE + NC + S, data = new_senic)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8362 -0.4945 -0.0606  0.5284  2.5225

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.024958   1.358265  -1.491 0.139151
Length       0.242403   0.070184   3.454 0.000812 ***
Age          0.013231   0.021771   0.608 0.544745
Culture      0.054485   0.010550   5.165 1.23e-06 ***
Xray         0.011553   0.005264   2.195 0.030504 *
Beds        -0.003489   0.002677  -1.303 0.195420
Affiliation  0.660823   0.321392   2.056 0.042375 *
Patients     0.003865   0.003451   1.120 0.265477
Nurses       0.001787   0.001695   1.055 0.294165
Facilities   0.020570   0.010059   2.045 0.043492 *
NE          -1.149535   0.339173  -3.389 0.001004 **
NC          -0.724026   0.297823  -2.431 0.016835 *
S           -0.782774   0.288954  -2.709 0.007941 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9138 on 100 degrees of freedom
Multiple R-squared:  0.5854,    Adjusted R-squared:  0.5356
F-statistic: 11.76 on 12 and 100 DF,  p-value: 1.963e-14
```

----------end of question 1 part e----------

question 1 part f

In [325]:

```
Xh <- data.frame(Length=11, Age=45, Culture=18, Xray=100, Beds=400, NC=1, NE=0, S=0,
                 Patients=400, Nurses=340, Facilities=52, Affiliation=1)
predict(new_model, Xh, level=0.99, interval='prediction')
```

fit	lwr	upr
5.137175	2.522474	7.751876

Therefore, the predicted infection rate is 5.14%. Prediction interval is from 2.52% to 7.75%

----------end of question 1 part f----------

question 2 part a

In [23]:

```
share <- read.table('marketshare.txt', header=TRUE)
attach(share)
```

In [24]:

```
m2 <- lm(Share ~ Price + Exposure + Discounted + Promoted, share)
summary(m2)
```

```
Call:
lm(formula = Share ~ Price + Exposure + Discounted + Promoted,
    data = share)

Residuals:
      Min        1Q    Median        3Q       Max
-0.284946 -0.102265 -0.001004  0.103386  0.240284

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.158e+00  4.405e-01   7.168 4.67e-08 ***
Price       -3.439e-01  1.767e-01  -1.946   0.0607 .
Exposure     1.993e-05  1.714e-04   0.116   0.9081
Discounted   3.999e-01  5.246e-02   7.623 1.35e-08 ***
Promoted     1.165e-01  5.394e-02   2.160   0.0386 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1522 on 31 degrees of freedom
Multiple R-squared:  0.7066,    Adjusted R-squared:  0.6688
F-statistic: 18.67 on 4 and 31 DF,  p-value: 6.641e-08
```

H0: all coefficients β0, β1 ... β(p-1) are zero

Ha: not all coefficients are zero

From the output above, the p-value is equal to 6.641 * 10 ^ (-8), which is smaller than 0.05 Reject the H0 and accept the alternative hypothesis. The model is significant in reducing the variation in market share.

The coefficients for the categorical variable indicates the increase in the expected market share when the categorical variable takes the value of 1, holding all other variables constant. For instance, the model indicates that the expected market share will increase by 0.35 percent when discounting the price, all else equal.

I perform an F test to test if discounting the price increases the expected market share, all else equal. H0: β(discount) = 0 Ha: β(discount) != 0

In [25]:

```
reduced_model <- lm(Share ~ Price + Exposure + Promoted + Month + Year, share)
full_model <- lm(Share ~ Price + Exposure + Discounted + Promoted + Month + Year, share)
```

In [26]:

```
SSE.F <- sum(full_model$residuals ** 2)
paste('SSE.F:', SSE.F)
SSE.R <- sum(reduced_model$residuals ** 2)
paste('SSE.R:', SSE.R)

F_star <- (SSE.R - SSE.F) / 1 / (SSE.F / m2$df.residual)
paste('The test statistic is', F_star)

alpha = 0.05
paste('the level of significance is ', 1-alpha)
p_value = 1 - pf(F_star, 1, full_model$df.residual)
paste('p-value is', p_value)
```

'SSE.F: 0.323460606063188'

'SSE.R: 1.02197481741182'

'The test statistic is 66.9445989586052'

'the level of significance is 0.95'

'p-value is 1.19807671783434e-07'

The p-value is smaller than 0.05, and it is equal to the Summary function output. Therefore, reject the null hypothesis and accept the alternative hypothesis: all else equal, discounting the price will increase the expected market share. In other words, the categorical variable 'Discounted' is a significant contributor to the model.

I can also perform the above F-test using an anova table instead.

In [27]:

```
anova(reduced_model, full_model)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
20	1.0219748	NA	NA	NA	NA
19	0.3234606	1	0.6985142	41.03056	3.836752e-06

From the ANOVA table above, the p-value for the Discounted variable gives the same value as I previously calculated.

----------end of question 2 part a----------

question 2 part b

H0: βdiscounted = βpromoted = 0

Ha: not both βdiscounted and βpromoted are equal to 0

I start by fitting a full model (fm) and a reduced model (rm) excluding Discounted and Promoted

In [28]:

```
fm <- lm(Share ~ Price + Exposure + Month + Year + Promoted + Discounted)
rm <- lm(Share ~ Price + Exposure + Month + Year)
```

I then calculate the F-statistic (f_star)

In [29]:

```
sse.f <- sum(fm$residuals ** 2)
sse.r <- sum(rm$residuals ** 2)

f_star = (sse.r - sse.f) / 2 / (sse.f / fm$df.residual)
paste('F-statistic: ', f_star)

p_value = 1 - pf(f_star, 2, fm$df.residual)
paste('p-value: ', p_value)
```

'F-statistic: 22.6616332151661'

'p-value: 9.30384250596994e-06'

The p-value is very small. Therefore, reject the null hypothesis and accept the alternative hypothesis.

Equivalently, I can also use the anova table.

In [30]:

```
anova(rm, fm)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
21	1.0950549	NA	NA	NA	NA
19	0.3234606	2	0.7715943	22.66163	9.303843e-06

The anova table performs the F test and returns the same p-value as I previously calculated.

----------end of question 2 part b----------

question 2 part c

All else being equal, the expected difference in market share if the management decides to discount and promote their product is given by the sum of coefficients of Discounted and Promoted

When the covariate Discounted and Promoted both take the value of 1 : E{share} = β0 + β1 + β2 + β3 *Price* + *β4* Exposure where β1 and β2 are coefficients of Discounted and Promoted\

When the covariate Discounted and Promoted both take tha value of 0: E{share} = β0 + β3 *Price* + *β4* Exposure.

The expected difference in market share would change by β1 + β2

In [290]:

```
as.numeric(coef(m2)['Discounted']) + as.numeric(coef(m2)['Promoted'])
```

0.516380201936669

Therefore, the market share is expected to increase by 0.52 percent if the management decides to discount and promote the product at the same time.

----------end of question 2 part c----------

question 2 part d

H0: coefficients of discounted and promoted are the same (βdiscounted - βpromoted = 0)

Ha: coefficients of discounted and promoted are not the same (βdiscounted - βpromoted != 0)

The appropriate restricted model to use to test the hypothesis is as follows

E{Share} = b0 + b1 *Price* + *b2* Exposure + b3 * (Discounted + Promoted)

The full model to use is the model obtained in part a

In [32]:

```
restricted_model <- lm(Share ~ Price + Exposure + I(Discounted + Promoted), share)
full_model <- m2
```

If the alternative hypothesis is true, we should see significant reduction in SSE by the full model where Promoted is allowed to have its own distinct coefficient, compared to the restricted model, where the coefficient of Promoted is 'forced' to be identical to that of Discounted. In other words, if Promoted and Discounted have different coefficients, the full model should do a statistically significant better job in reducing the SSE compared to the restricted model.

This test can be achieved by using the anova table.

In [33]:

```
anova(restricted_model, full_model)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
32	0.9963188	NA	NA	NA	NA
31	0.7176404	1	0.2786784	12.0381	0.001554535

The p-value is smaller than 0.05. Therefore, reject the null hypothesis and accept the alternative hypothesis. The impact of discounting the price on market share is different to that of promoting the product.

Alternatively, there is a function in package 'car' that can perform the above test.

In [327]:

```
library('car')

linearHypothesis(m2, 'Discounted = Promoted')
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
32	0.9963188	NA	NA	NA	NA
31	0.7176404	1	0.2786784	12.0381	0.001554535

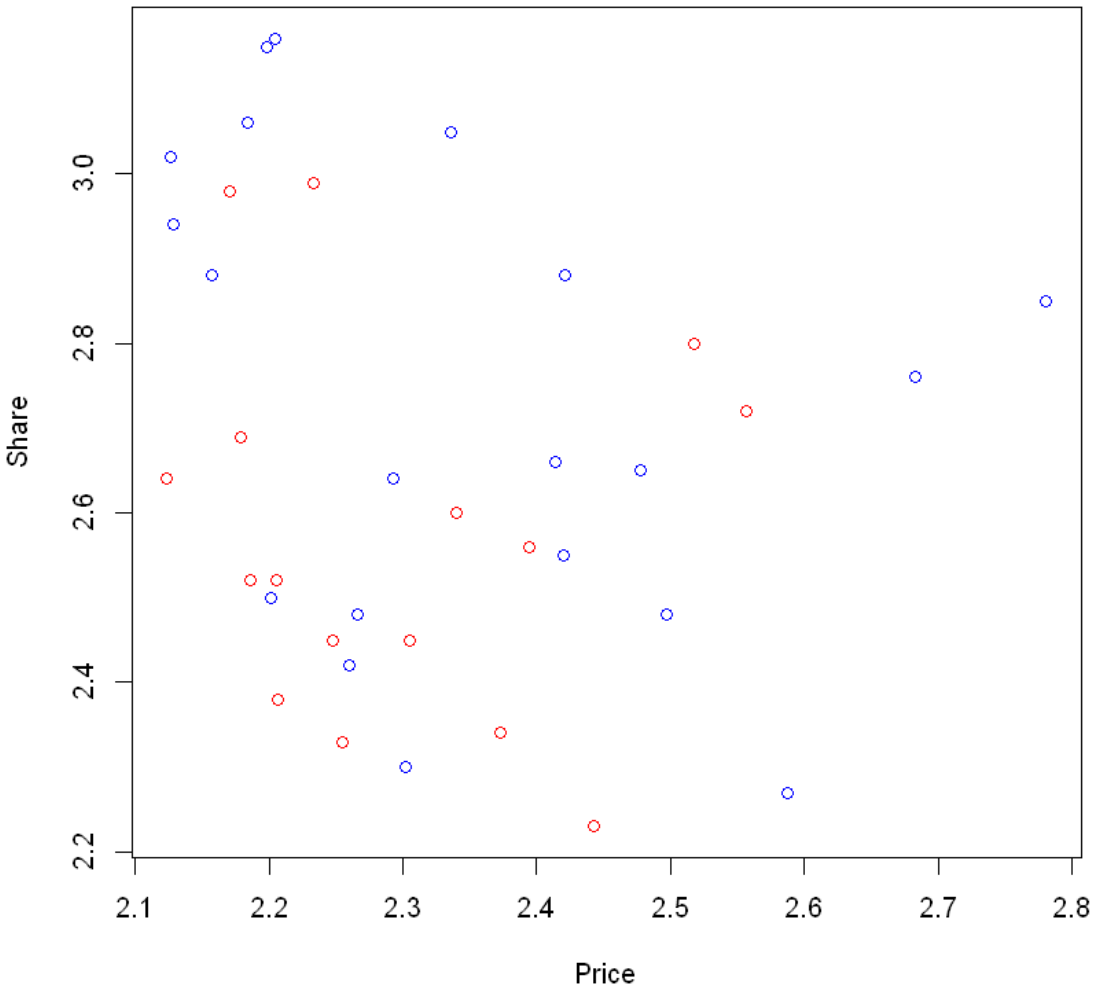
The package function gives the same p-value.

Intuitively, there is also another way of testing that the impact of discouting the price is differenct than promoting the product.

By ploting market share against price and separating the data points into two groups according to 'Discounted' and 'Promoted', it is clear that the including Discounted in the model helps explain more variance in market share than does Promoted.

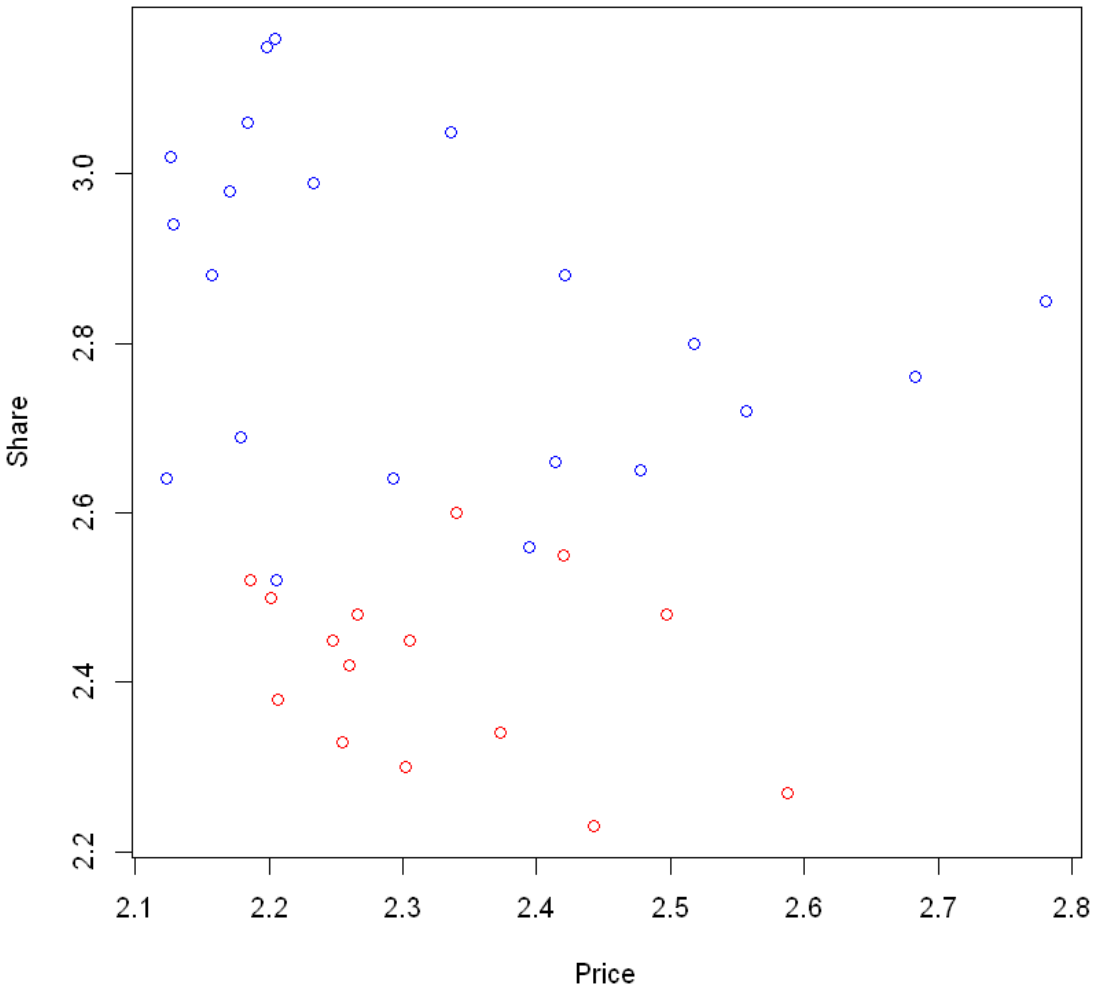
In [35]:

```
# separating the data into two groups according to Promoted
plot(Price,Share, col=c('red', 'blue')[Promoted + 1])
```



In [36]:

```
# separating the data according to Discounted
plot(Price, Share, col=c('red', 'blue')[Discounted + 1])
```



Therefore, intuitively we can also see that the coefficient of Discounted and Promoted are different.

----------end of question 2 part d----------

question 2 part e

Assess model assumptions

Examine the distribution of covariates to see if the values of the two quantitative covariates (Price, Exposure) makes sense.

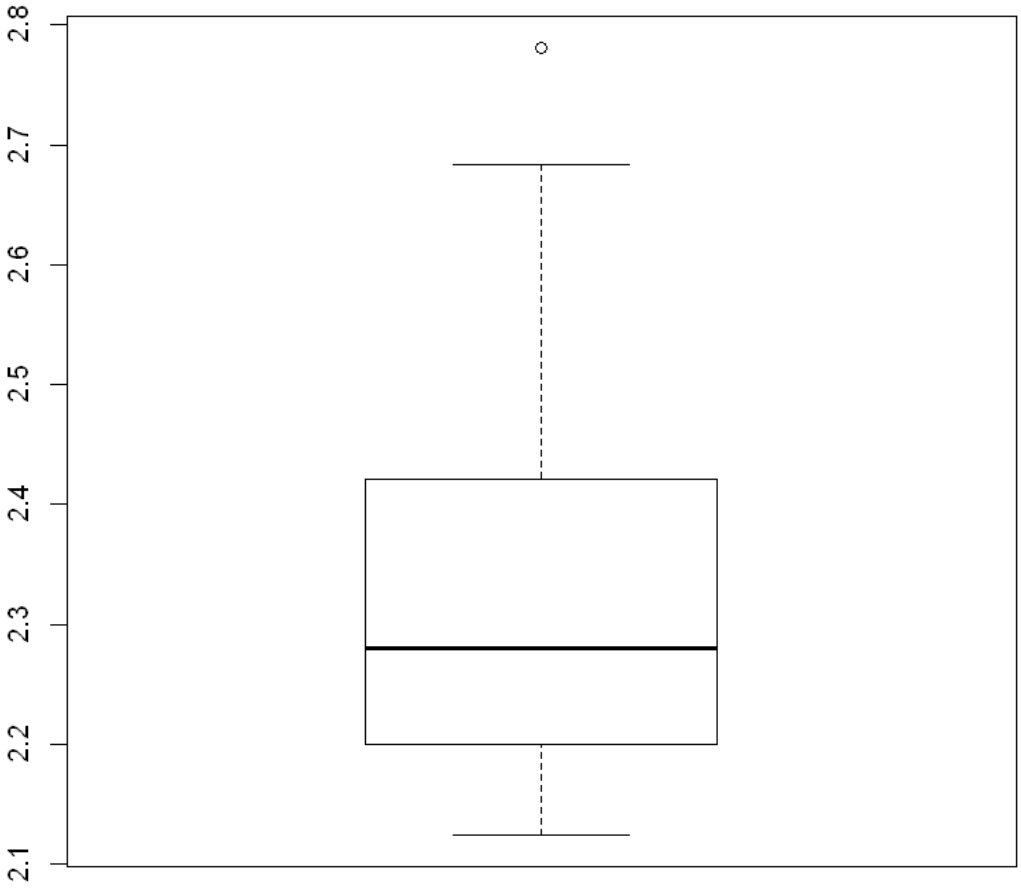
In [122]:

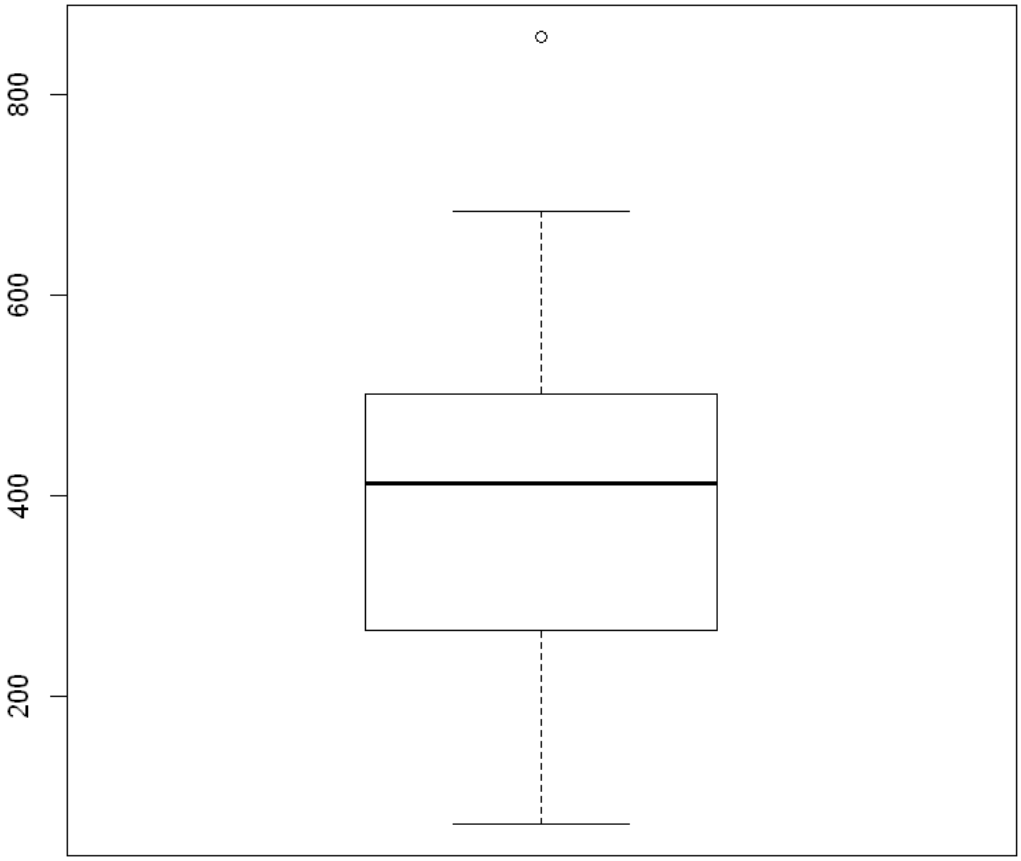
```
stem(Price)
stem(Exposure)
```

```
The decimal point is 1 digit(s) to the left of the |

21 | 23367889
22 | 00111356679
23 | 014479
24 | 12248
25 | 0269
26 | 8
27 | 8


The decimal point is 2 digit(s) to the right of the |

0 | 78
1 | 558
2 | 13567
3 | 1222259
4 | 023455779
5 | 00122479
6 | 8
7 |
8 | 6
```

In [38]:

```
boxplot(Price)
boxplot(Exposure)
```


2.8
2.7
2.6
2.5
2.4
2.3
2.2
2.1


800
600
400
200

There does not seem to be any abnormally large values or negative values in the quantitative covariates in the data.

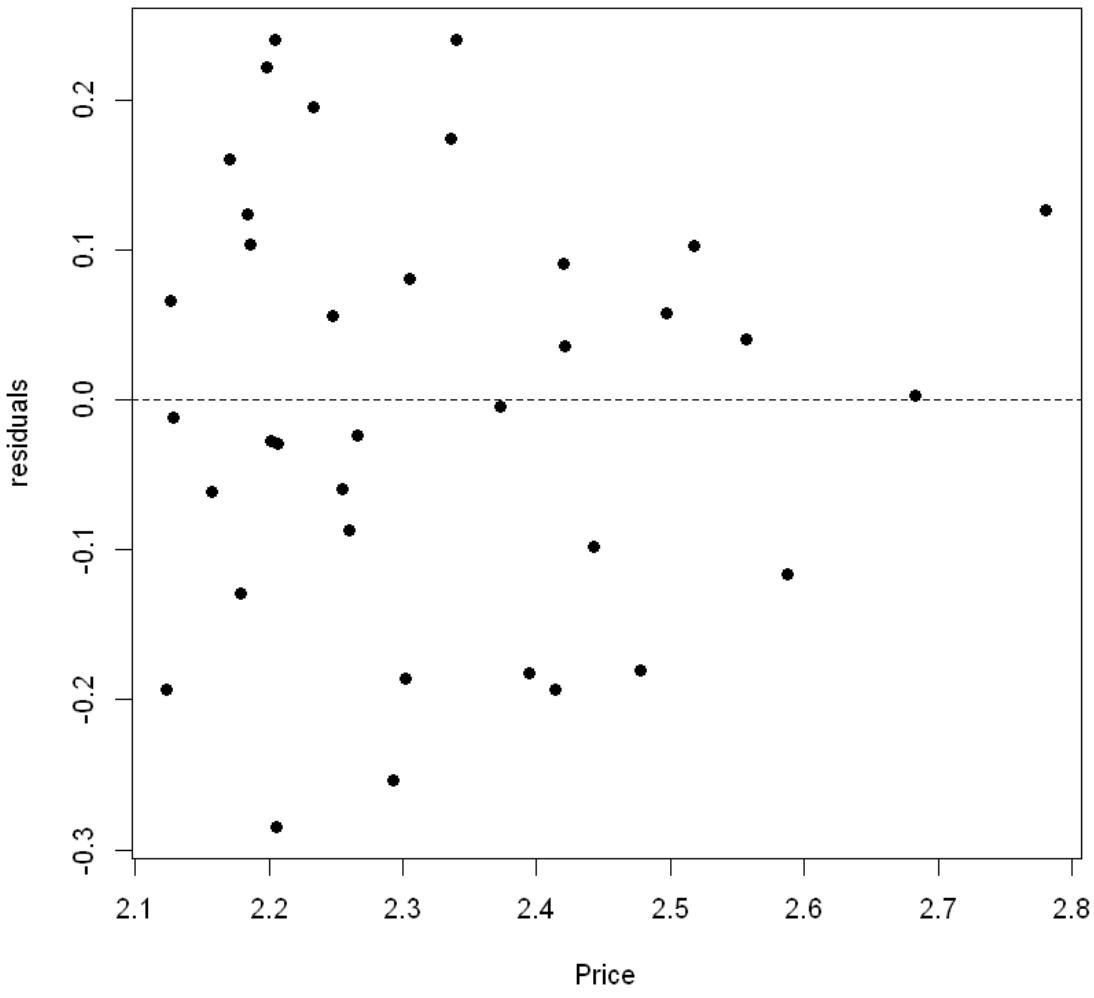
The model assumes that the error terms should have the following 4 properties: 1. its variance does not depend on the level of covariates. (Constancy) 2. there should not be any pattern in relation to the covariates (linearity) 3. The error term should be normally distributed around the mean = 0. (Normality) 4. The error term should be independent. (Independence)

Test of constancy and linearity

Plot of residuals against price

In [40]:

```
plot(Price, residuals(m2), pch=19, ylab='residuals')
abline(0, 0, lty=2)
```

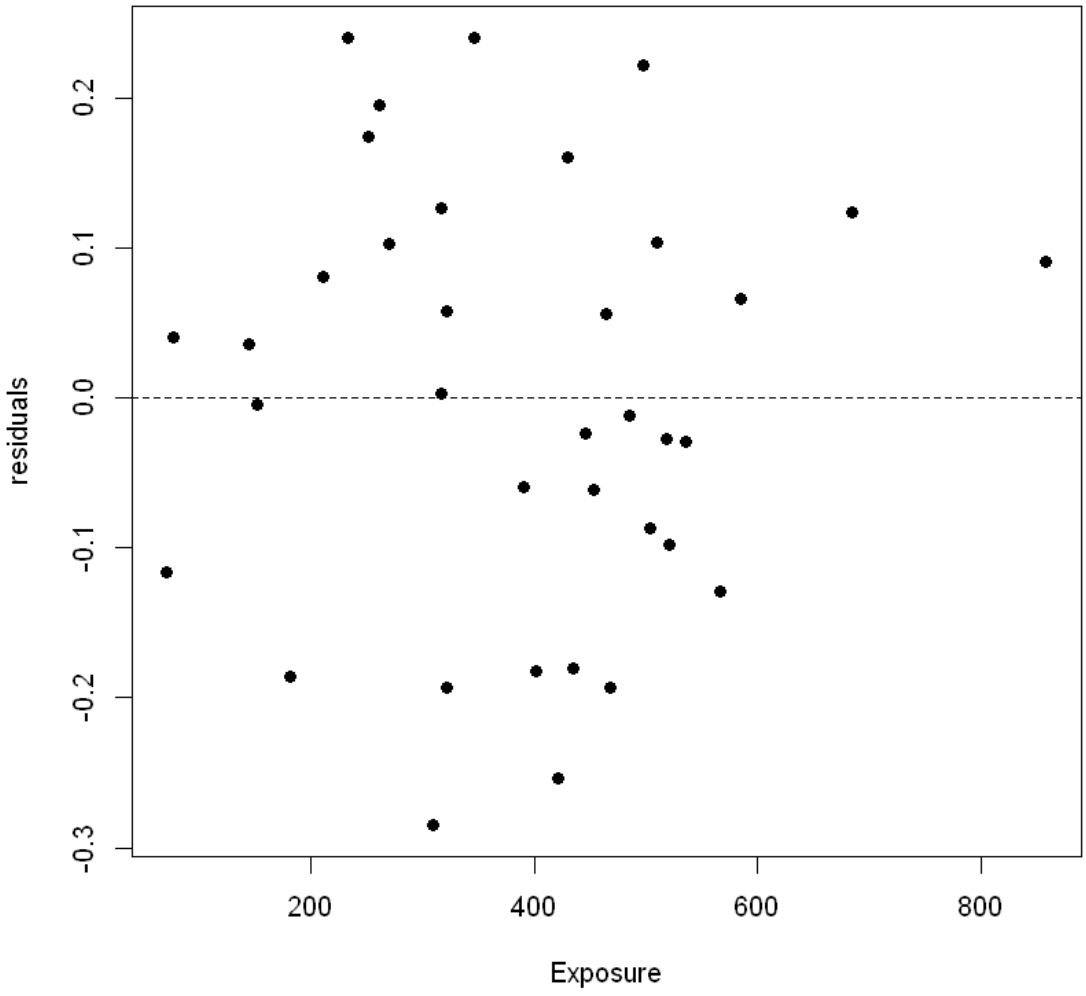


The plot above does not raise the concern of residual variance being inconstant, and there is no deviation from linearity. Most of the scatter points are clustered to the left, this is due to the fact that most prices are clustered on the lower side.

Plot of residuals against exposure

In [41]:

```
plot(Exposure, residuals(m2), pch=19, ylab='residuals')
abline(0,0,lty=2)
```

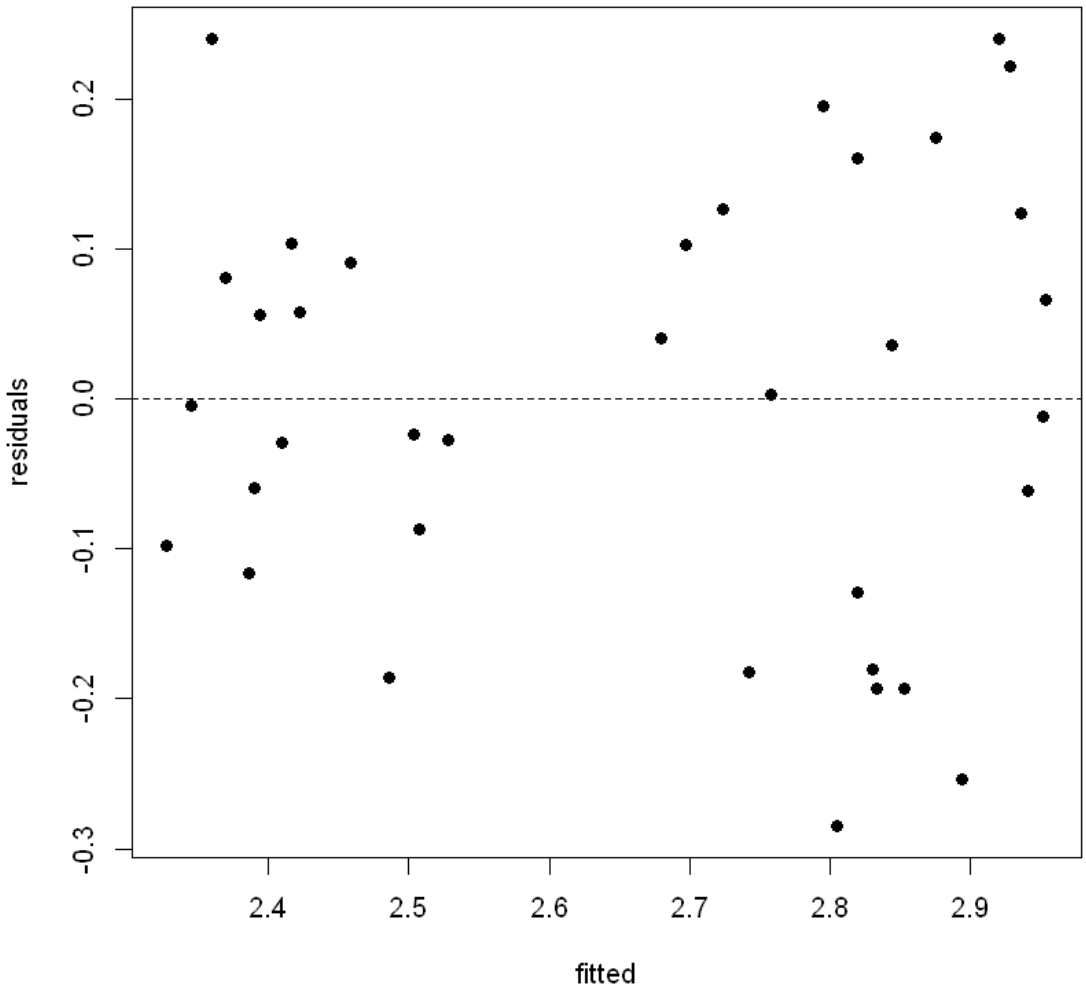


The scatterplot of residuals against exposure does to seem to have any pattern.

Plot of residuals against fitted values

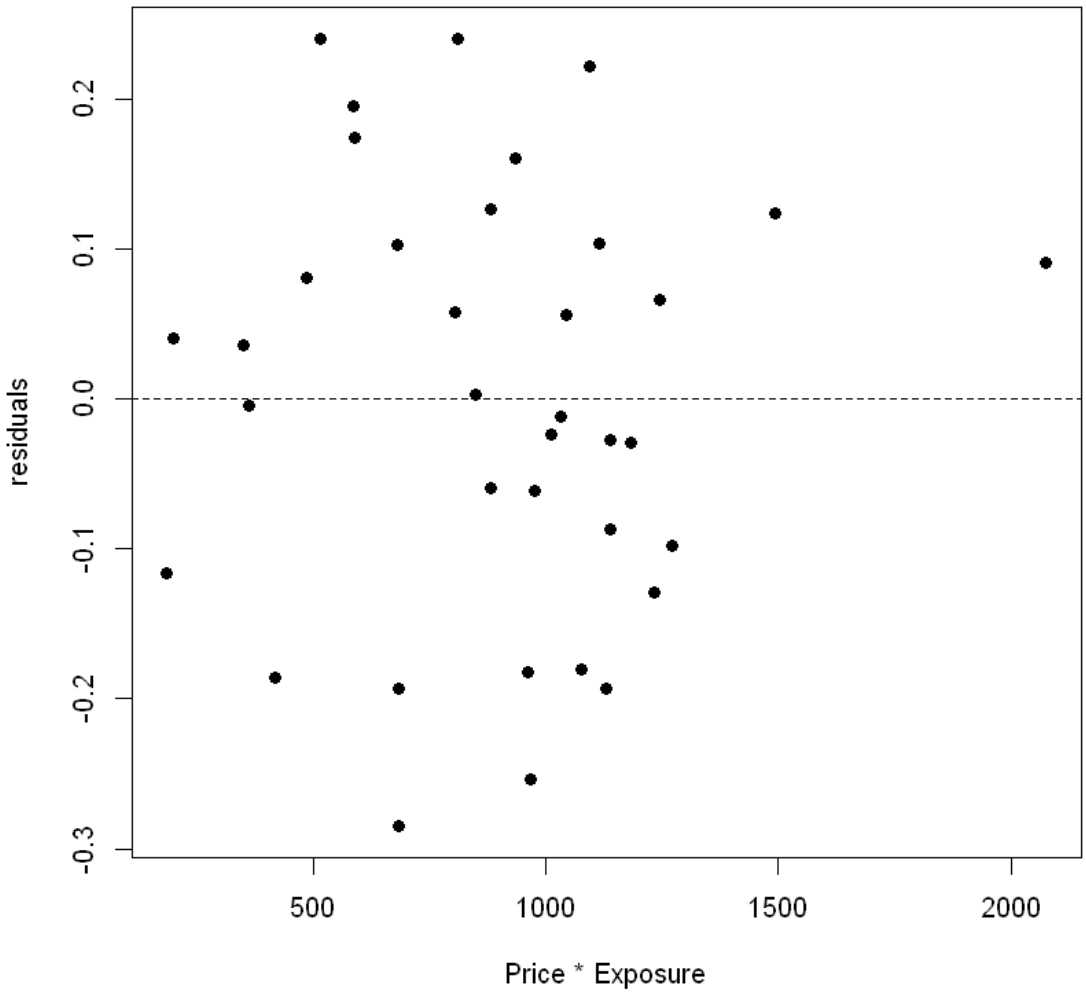
In [42]:

```
plot(fitted.values(m2), residuals(m2), pch=19, xlab='fitted', ylab='residuals')
abline(0,0,lty=2)
```



In [329]:

```
plot(Price * Exposure, residuals(m2), pch=19, xlab='Price * Exposure', ylab='residuals')
abline(0,0,lty=2)
```



Again, variance of residual seems constant.

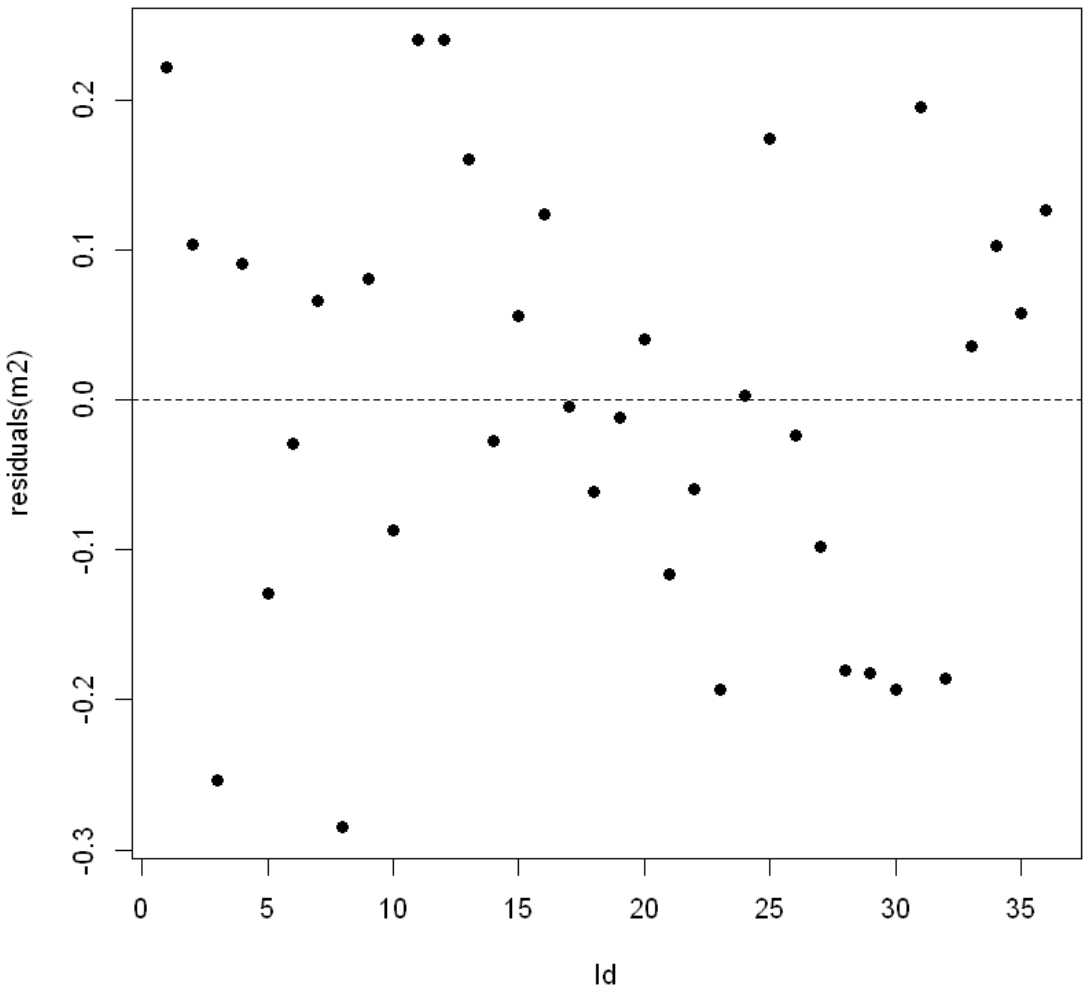
The plot of residual against fitted values suggests agreement with the linearity and constancy assumption, since there is no systematic patterns and the variance seems to be relatively constant.

Test of error term independence

Since the data is taken through time, it is sensible to consider a sequential plot of residualsx against time to examine whether the residuals are independent.

In [43]:

```
plot(Id, residuals(m2), pch=19)
abline(0,0,lty=2)
```



There does not seem to be any pattern in the sequential plot. Therefore the model upholds the independent residual assumption.

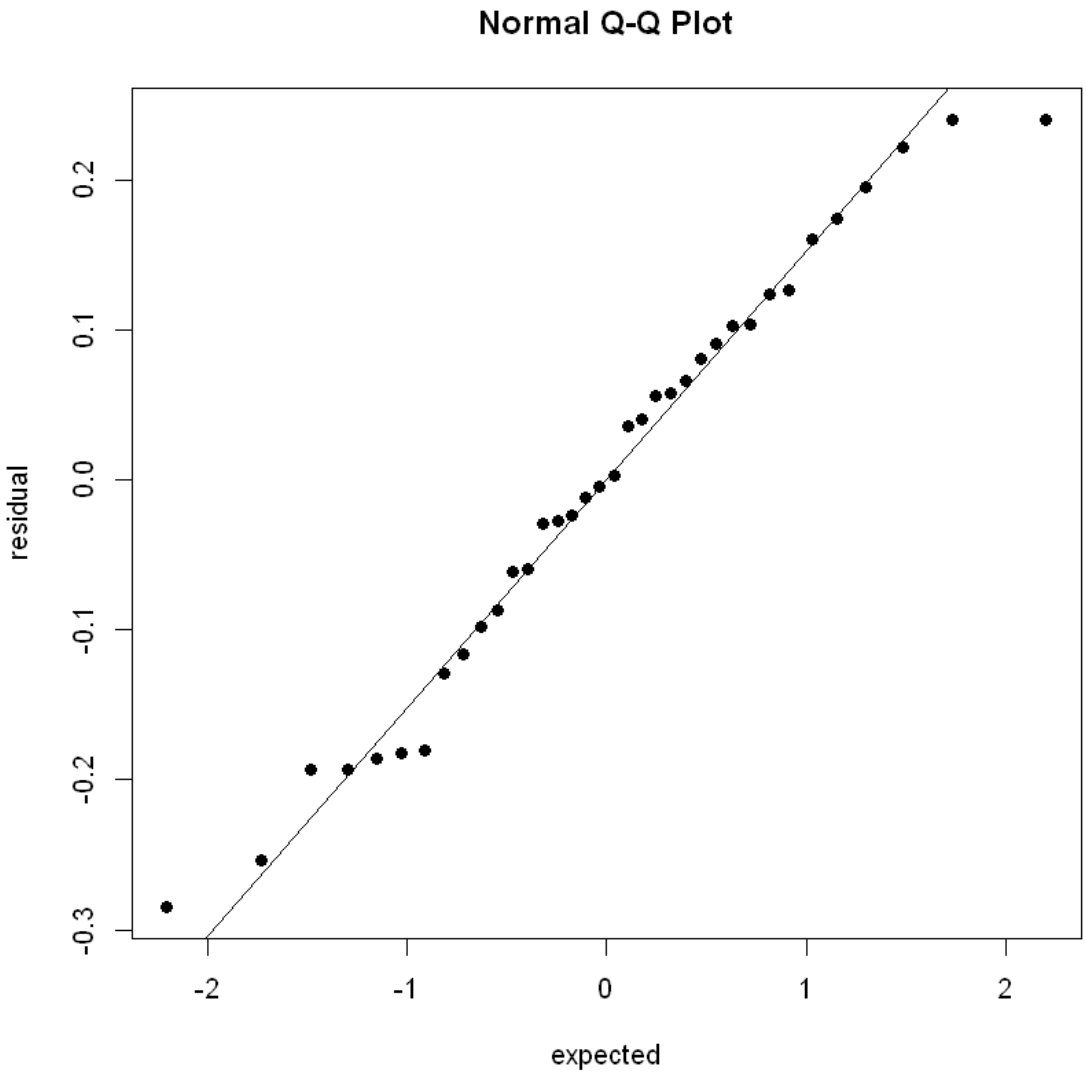
Test of normality of the error term

The error term in a MLR should be normally distributed.

This property is examined using the qqnorm function.

In [123]:

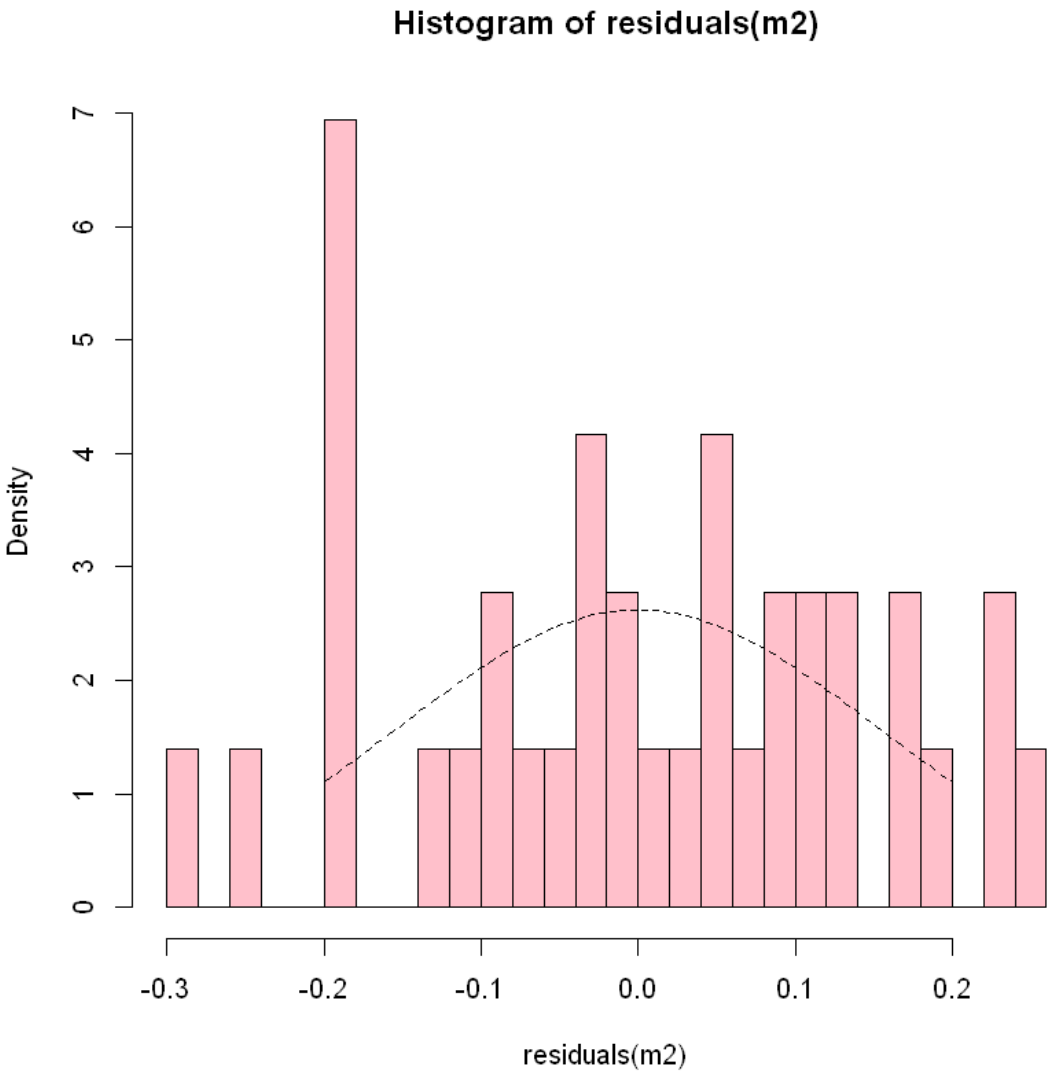
```
qqnorm(residuals(m2), pch=19, xlab='expected', ylab='residual')
qqline(residuals(m2))
```



A histogram of the residuals can also test if the residuals are normally distributed.

In [124]:

```
SSE <- sum(residuals(m2) ** 2)
MSE <- SSE / df.residual(m2)

hist(residuals(m2),20, freq=FALSE, col='pink')
plot(function(x)dnorm(x, sd=sqrt(MSE)), add=TRUE, from=-0.2, to=0.2, lty=2)
```



The histogram shows a lack of normality. The qqplot shows the residuals deviate from normality at both tails. Nonetheless, he residuals are roughly symmetrically distributed.

Influence diagnostics

A data point's influence is measure by its leverage and residual.

leverage

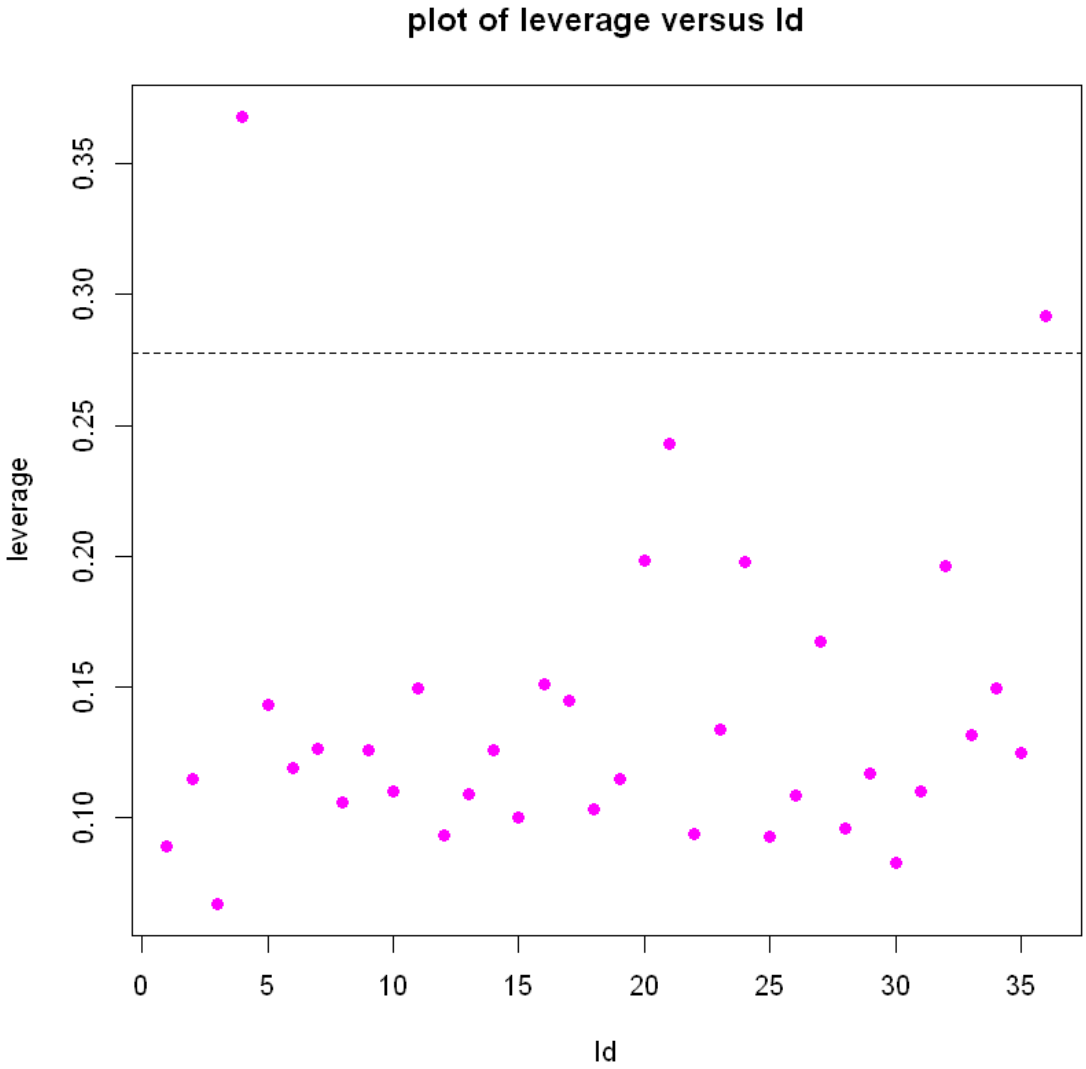
The leverage of a data point is obtained from the diagnal elements of the projection matrix.

A general rule of thumb is that for a dataset with n observations that is fitted to a model with p-1 explanatory variables, the cutoff leverage value is given by 2 * p / n. Hence, a data point whose its leverage level is high than the cutoff is deemed to have high leverage.

In [90]:

```
n <- length(share[,'Id'])
ones <- rep(1, n)
X <- cbind(ones, Price, Exposure, Discounted, Promoted)
leverage <- hat(X)

cutoff <- 2 * summary(m2)$df[1] / n
cat(paste('The cutoff leverage level is ', round(cutoff,3)))

plot(Id, leverage, pch=19, col='magenta', main='plot of leverage versus Id')
abline(cutoff, 0, lty=2)
```

```
The cutoff leverage level is  0.278
```


plot of leverage versus Id
leverage
0.10
0.15
0.20
0.25
0.30
0.35
0
5
10
15
20
25
30
35
Id

The scatterplot of leverage versus data id shows that points with id = 4 and 36 are high leverage points.

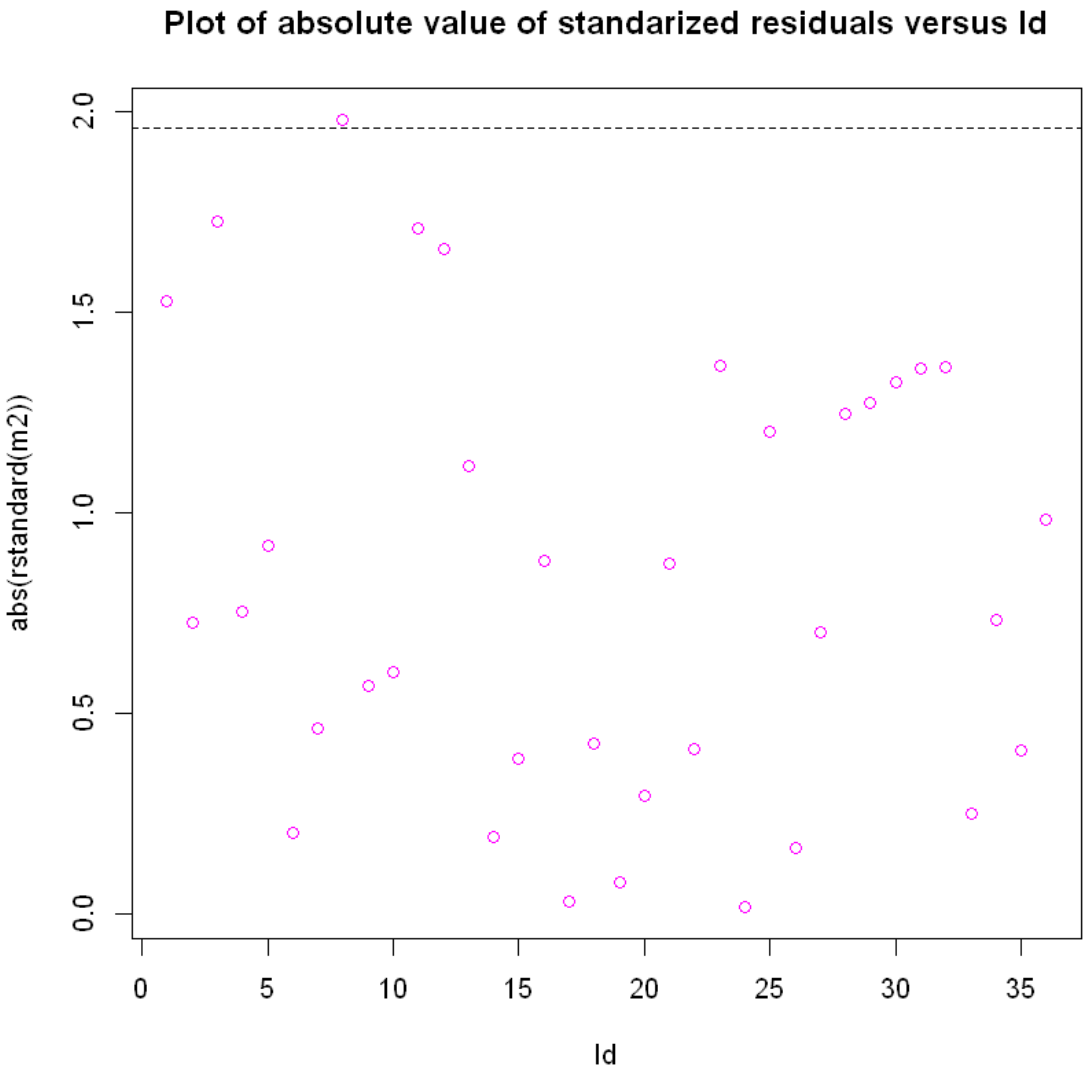
Residuals

The magnitude of residuals depends on unit of measurement, therefore it is important to standarize them by deviding an estimate of their standard deviation (or use rstandard()).

In [131]:

```
plot(Id, abs(rstandard(m2)), col='magenta', main='Plot of absolute value of standarized residuals versus Id')
abline(1.96, 0, lty=2)

paste('The data point with a standarized residual greater than 1.96 has id = ', Id[abs(rstandard(m2)) > 1.96])
```

'The data point with a standarized residual greater than 1.96 has id = 8'



According to the plot above, the data points with id of 3, 8, 11 and 12 have the highest resdiduals. As a general rule of thumb, if the standarized residual are greater than 1.96 (97.5 percentile of the normal distribution), then this data point is considered an outlier. Therefore, data point number 8 is an outlier, and it corresponds to April, 2000.

These data points corresponds to November 1999, April 2000, July 2000 and Aug, 2000, shown as follows.

In [100]:

```
share[c(3, 8, 11, 12),]
```

	Id	Share	Price	Exposure	Discounted	Promoted	Month	Year
3	3	2.64	2.293	422	1	1	Nov	1999
8	8	2.52	2.206	310	1	0	Apr	2000
11	11	3.16	2.205	234	1	1	Jul	2000
12	12	2.60	2.340	347	0	0	Aug	2000

Influence diagnostics

The influence of a data point is measured in terms of cook's distance, which takes into account both a data point's residual and its leverage. Cook's distance measures the extent to which the regression coefficients change once a praticular data point is removed.

In [120]:

```
head(sort(cooks.distance(m2), decreasing=TRUE))
```

11
0.102780413479154
8
0.093256622755043
32
0.0907410524548266
36
0.0798866185647859
4
0.0660339735551045
23
0.0577790811593466

The output from the cell above shows the index of data points with high influence. These points are influencial in the sense that the regression coefficients change a great deal once they are removed from the data.

For instance, if I remove data point (id = 11) from the regression, I get the following coefficients.

In [115]:

```
coef(lm(Share ~ Price + Exposure + Discounted + Promoted, data=share[-11,]))
```

(Intercept)
2.95265452067565
Price
-0.265569201295907
Exposure
0.000100426770565391
Discounted
0.391813851140739
Promoted
0.0956880089596136

Compare this with the original coefficients of the full model

In [118]:

```
coef(m2)
```

(Intercept)
3.15760010905109
Price
-0.343932062179736
Exposure
1.99325840756103e-05
Discounted
0.399880934966916
Promoted
0.116499266969754

These points are influential in the sense that the parameter estimates changes a great deal when these points are removed from the regression.

Compare this with the effect of removing a data point with small cook's distance from the regression (e.g. Id = 24)

In [121]:

```
coef(lm(Share ~ Price + Exposure + Discounted + Promoted, data=share[-24,]))
```

(Intercept)
3.16068969420939
Price
-0.345203643384787
Exposure
1.96676104447524e-05
Discounted
0.399732553983244
Promoted
0.116445096502906

The coefficients have not changed much, if at all.

Hence, the changes in coefficients by removing an influential data point is more profound.

----------end of question 2 part e----------

question 2 part f

After adding all second-order terms involving Price and Exposure, the model becomes:

lm(Share ~ Price + Exposure + Discounted + Promoted + Price: Exposure + I(Price ^ 2) + I(Exposure ^ 2))

We want to test if adding these second-order terms can increase the model's ability to explain more variation in market share.

The appropriate test here would be an F-test with test statistic = ESS / k / MSE.F, where ESS is the extra sum of square brought about by adding the second-order terms, k is the number of second-order terms added and MSE.F is the full model's mean square of error.

It would be tedious and superfluous to calculate the test statistic by hand, so instead I just use the ANOVA function.

The null hypothesis is that all the true coefficient of the second-order terms are 0

H0: β5 = β6 = β7 = 0

Ha: not all β5, β6 and β7 are equal to 0

In [218]:

```
ma <- lm(Share ~ Price + Exposure + Discounted + Promoted)
mb <- lm(Share ~ Price + Exposure + Discounted + Promoted + Price: Exposure + I(Price ^ 2) + I(E
xposure ^ 2))
anova(ma, mb)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
31	0.7176404	NA	NA	NA	NA
28	0.6798107	3	0.03782973	0.5193762	0.6724098

The anova analysis reveals that adding these second-order terms decrease the residual sum of square by merely 0.038. Namely, the extra sum of square is only 0.038. The F-statistic associated with the test is equal to 0.52. p-value is quite large at 0.67. Therefore, it is not statistically significant that adding these second-order terms help the model better explain variation in market share (the response). Hence, do not reject the null hypothesis that all the second-order terms involving Price and Exposure can be dropped from the model given that Price and Discounted are already in the model.

A possible explanation of why these second-order terms do not add more information to the covariates that are already present in the model (i.e. Price, exposure, Discounted, Promoted) is that there exists multicolinearity.

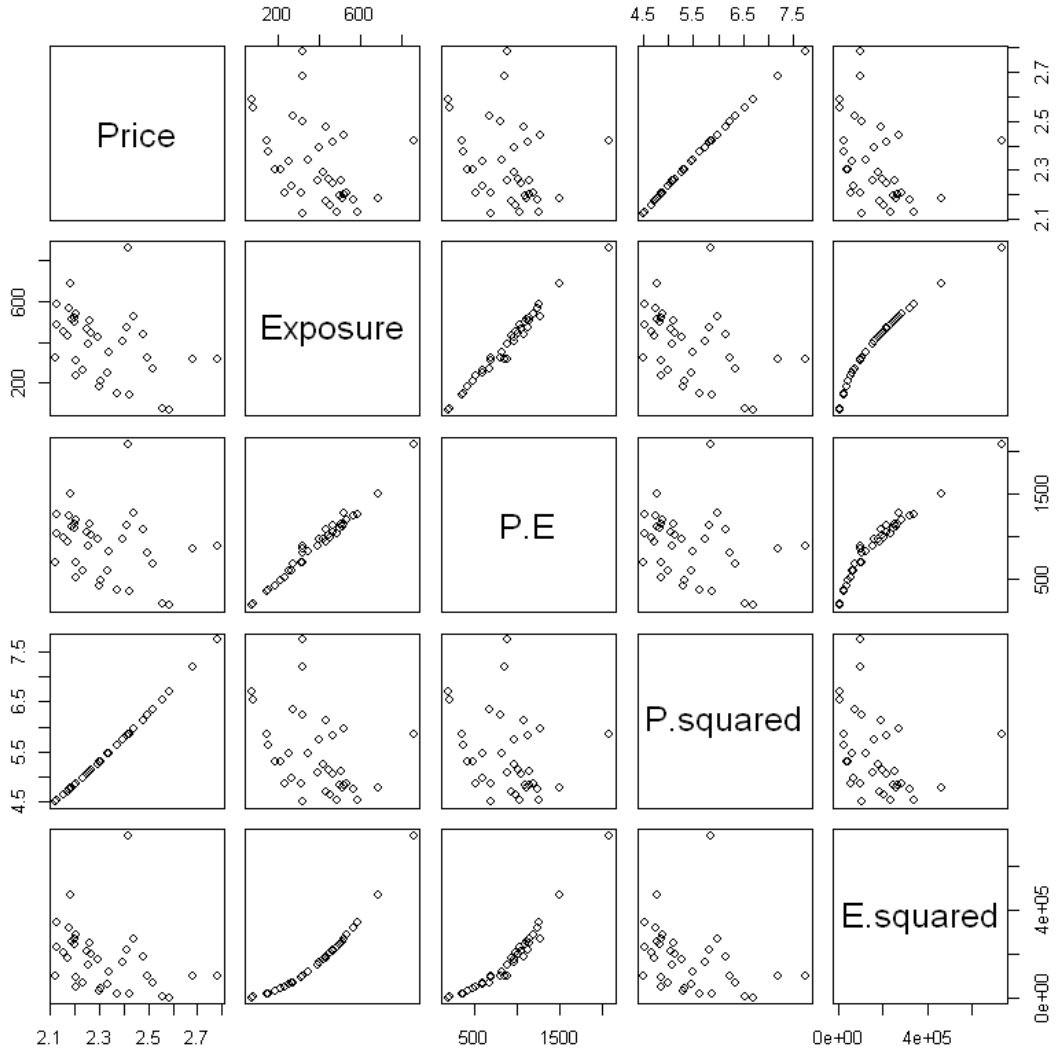
This fact is better demonstrated using a correlation matrix of the quantitative predictors and their corresonding second-order predictors.

In the cell below, I use P.E to denote the interaction term between Price and Exposure, P.squared to denote Price squared and E.squared to denote Exposure squared. It is clear from the output that P.E is highly correlated to Exposure (rho = 0.988), P.squared is highly correlated with Price (rho = 0.999), and E.squared is highly correlated with Exposure (rho = 0.957). When the newly added predictors are highly correlated with existing predictors, multicolinearity makes the newly added predictors superfluous. In other words, when predictors 'Price' and 'Exposure' are already in the model, adding one or more second-order models involving 'Price' and 'Exposure' will not contribute to the model's ability in reducing RSS.

In [284]:

```
P.E <- data.frame(P.E = Price * Exposure)
P.squared <- data.frame(P.squared = Price ** 2)
E.squared <- data.frame(E.squared = Exposure ** 2)

pairs(cbind(Price, Exposure, P.E, P.squared, E.squared))
cor(as.matrix(cbind(Price, Exposure, P.E, P.squared, E.squared)))
```

	Price	Exposure	P.E	P.squared	E.squared
Price	1.0000000	-0.3877765	-0.2512284	0.9992136	-0.2974865
Exposure	-0.3877765	1.0000000	0.9879616	-0.3833736	0.9568455
P.E	-0.2512284	0.9879616	1.0000000	-0.2476631	0.9546187
P.squared	0.9992136	-0.3833736	-0.2476631	1.0000000	-0.2957114
E.squared	-0.2974865	0.9568455	0.9546187	-0.2957114	1.0000000



Another explantion is that, the predictor Discounted does a better job in reducing variability in market share than do any of the second-order terms. Therefore, adding more second-order terms when Discounted is alreaady present is superfluous

This can be demonstrated using the example below. I drop Discounted and add Price ^ 2 as a predictor, the ANOVA tables shows Price ^ 2 is now a significant contributor to the model once Discounted has been dropped.

In [289]:

```
anova(lm(Share ~ Price + Exposure), lm(Share ~ Price + Exposure + I(Price ^ 2)))
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
33	2.359327	NA	NA	NA	NA
32	1.994837	1	0.3644904	5.846941	0.02147303

Some alternatives are

lm(Share ~ Price + Exposure + Discounted + Promoted + I(Price ^ 2))

lm(Share ~ Price + Exposure + Discounted + Promoted + I(Exposure ^ 2))

lm(Share ~ Price + Exposure + Discounted + Promoted + I(Price ^ 2) + I(Price ^ 2))

lm(Share ~ Price + Exposure + Discounted + Promoted + Price: Exposure)

lm(Share ~ Price + Exposure + Discounted + Promoted + Price: Exposure + I(Price ^ 2))

lm(Share ~ Price + Exposure + Discounted + Promoted + Price: Exposure + I(Exposure ^ 2))

In [260]:

```
alt1 <- lm(Share ~ Price + Exposure + Discounted + Promoted + I(Price ^ 2))
alt2 <- lm(Share ~ Price + Exposure + Discounted + Promoted + I(Exposure ^ 2))
alt3 <- lm(Share ~ Price + Exposure + Discounted + Promoted + I(Price ^ 2) + I(Price ^ 2))
alt4 <- lm(Share ~ Price + Exposure + Discounted + Promoted + Price: Exposure)
alt5 <- lm(Share ~ Price + Exposure + Discounted + Promoted + Price: Exposure + I(Price ^ 2))
alt6 <- lm(Share ~ Price + Exposure + Discounted + Promoted + Price: Exposure + I(Exposure ^ 2))
```

In [263]:

```
anova(m2, alt1)
anova(m2, alt2)
anova(m2, alt3)
anova(m2, alt4)
anova(m2, alt5)
anova(m2, alt6)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
31	0.7176404	NA	NA	NA	NA
30	0.7024115	1	0.0152289	0.6504265	0.4263077

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
31	0.7176404	NA	NA	NA	NA
30	0.7023699	1	0.01527046	0.6522401	0.4256704

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
31	0.7176404	NA	NA	NA	NA
30	0.7024115	1	0.0152289	0.6504265	0.4263077

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
31	0.7176404	NA	NA	NA	NA
30	0.7130567	1	0.004583671	0.192846	0.6637053

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
31	0.7176404	NA	NA	NA	NA
29	0.7019349	2	0.01570546	0.3244305	0.725529

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
31	0.7176404	NA	NA	NA	NA
29	0.6998005	2	0.01783991	0.3696463	0.6941875

According the above ANOVA tables, it is clear that p-values of all the alternative models are fairly large. This suggests that the second-order terms can all be dropped from the regression model because they does not contribute to the regression model. In other words, they are not significant contributor to the model.

The conclusion is that, when adding new higher-order of existing predictors as new predictors into the model, its significance depends on what predictors are already present in the model. If the existing predictors can do a better job in explaning the variability in the response, then the new predictors may not be significant, and adding them does not help improve the model.

----------end of question 2----------