Homework3

Net ID: ghosh17

Subhankar Ghosh

Question 1

```r
library(faraway)
head(cornnit)
```

```
##   yield nitrogen
## 1   115        0
## 2   128       75
## 3   136      150
## 4   135      300
## 5    97        0
## 6   150       75
```

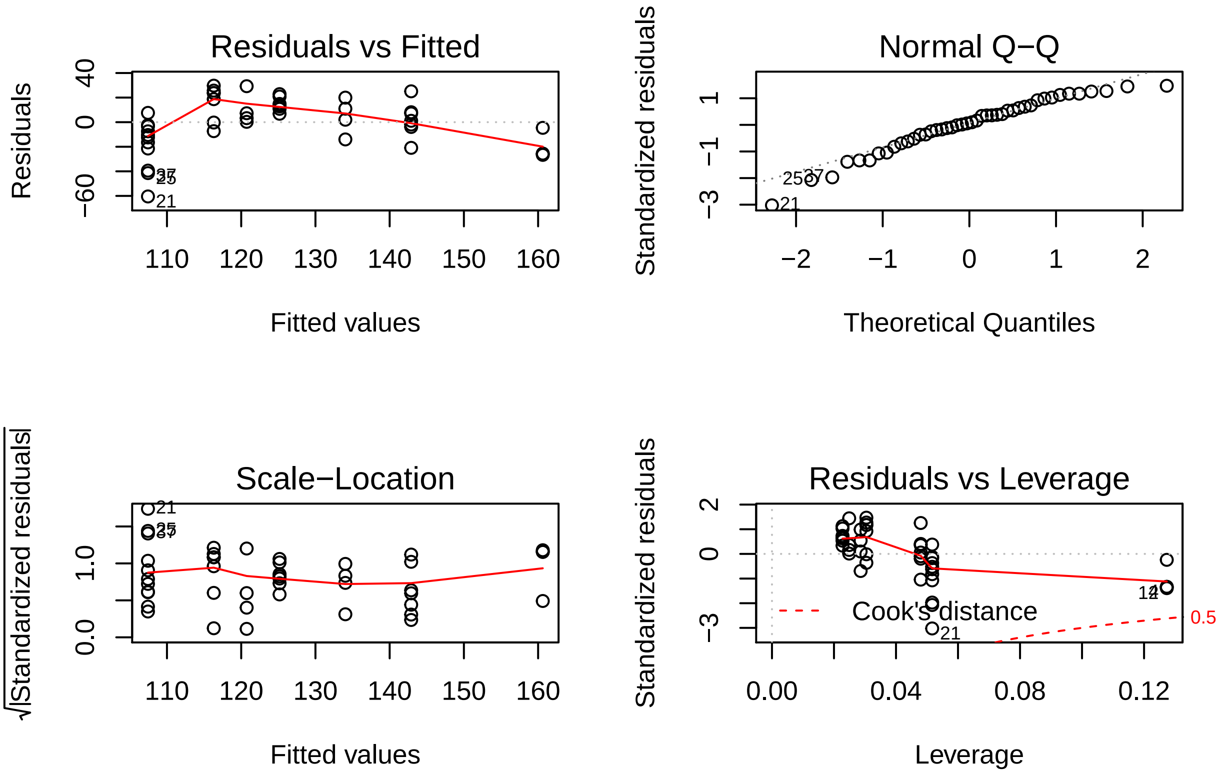
First we fit a simple model and check the diagnostic plots of this model

```r
simple_mod = lm(yield ~ nitrogen, data = cornnit)
summary(simple_mod)
```

```
##
## Call:
## lm(formula = yield ~ nitrogen, data = cornnit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.439 -10.939   1.534  14.082  29.697
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 107.43864    4.66622   23.02  < 2e-16 ***
## nitrogen      0.17730    0.03377    5.25 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.53 on 42 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3818
## F-statistic: 27.56 on 1 and 42 DF,  p-value: 4.713e-06
```

We get a Multiple R-squared of 0.3962 whcih is quite low. Lets see what transformation we can apply to get a better Multiple R-squared value.

```r
par(mfrow=c(2,2))
plot(simple_mod)
```

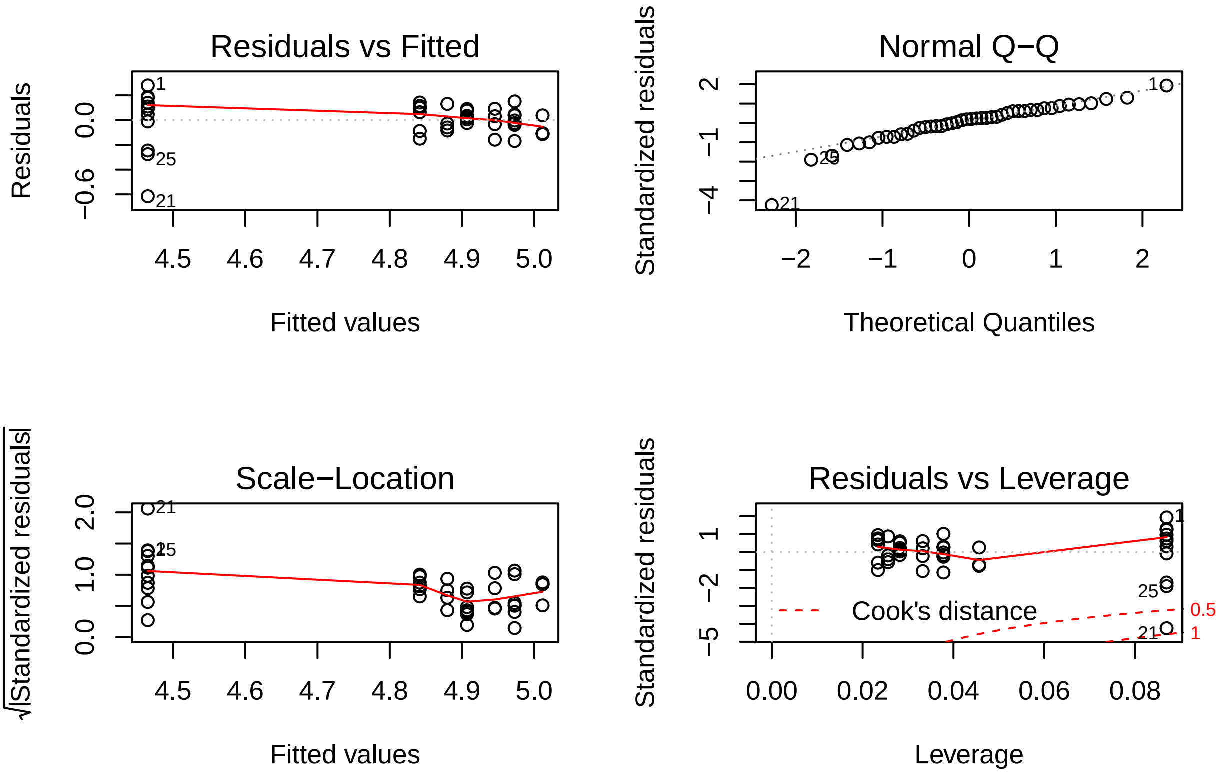


Clearly from the diagnostic plots we can see that the **linearity assumption is violated**. Since the relationship is non linear we will apply log transform on both resonse variable as well as the predictor variable.

```
log_mod = lm(log(yield) ~ log(nitrogen + 1), data = cornnit)
summary(log_mod)
```

```
##
## Call:
## lm(formula = log(yield) ~ log(nitrogen + 1), data = cornnit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61487 -0.06563  0.02932  0.09357  0.27992
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.46502    0.04468   99.94  < 2e-16 ***
## log(nitrogen + 1)  0.09577    0.01075    8.91 3.13e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1515 on 42 degrees of freedom
## Multiple R-squared:  0.654,  Adjusted R-squared:  0.6458
## F-statistic: 79.39 on 1 and 42 DF,  p-value: 3.128e-11
```

```
par(mfrow=c(2,2))
plot(log_mod)
```



We can see an increase in Multiple R-squared value from 0.39 to 0.654.

Also the diagnostic plots look linear. Therefore log transform is a good transform for this model.

We will now perform **Lack of fit test**

```
logfactor_mod = lm(log(yield) ~ factor(log(nitrogen + 1)), data = cornnit)
anova(log_mod, logfactor_mod)
```

```
## Analysis of Variance Table
##
## Model 1: log(yield) ~ log(nitrogen + 1)
## Model 2: log(yield) ~ factor(log(nitrogen + 1))
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     42 0.96447
## 2     37 0.92202  5   0.04245 0.3407 0.8849
```

Null Hypothesis of this F-test is H_0 : our model after transforms is a linear model. Since the $p-value = 0.8849$ is greater than significance level of 0.05 we cannot reject the null hypothesis. And thus conclude that the model after applying log transform is a linear model.

Question 2

Lets look at the ozone data

```
head(ozone)
```

```
##   O3   vh wind humidity temp  ibh dpg ibt vis doy
## 1  3 5710    4       28   40 2693 -25  87 250  33
```

```
## 2   5 5700    3       37   45   590 -24 128 100   34
## 3   5 5760    3       51   54  1450   25 139   60   35
## 4   6 5720    4       69   35  1568   15 121   60   36
## 5   4 5790    6       19   45  2631  -33 123 100   37
## 6   4 5790    3       25   55   554  -28 182 250   38
```

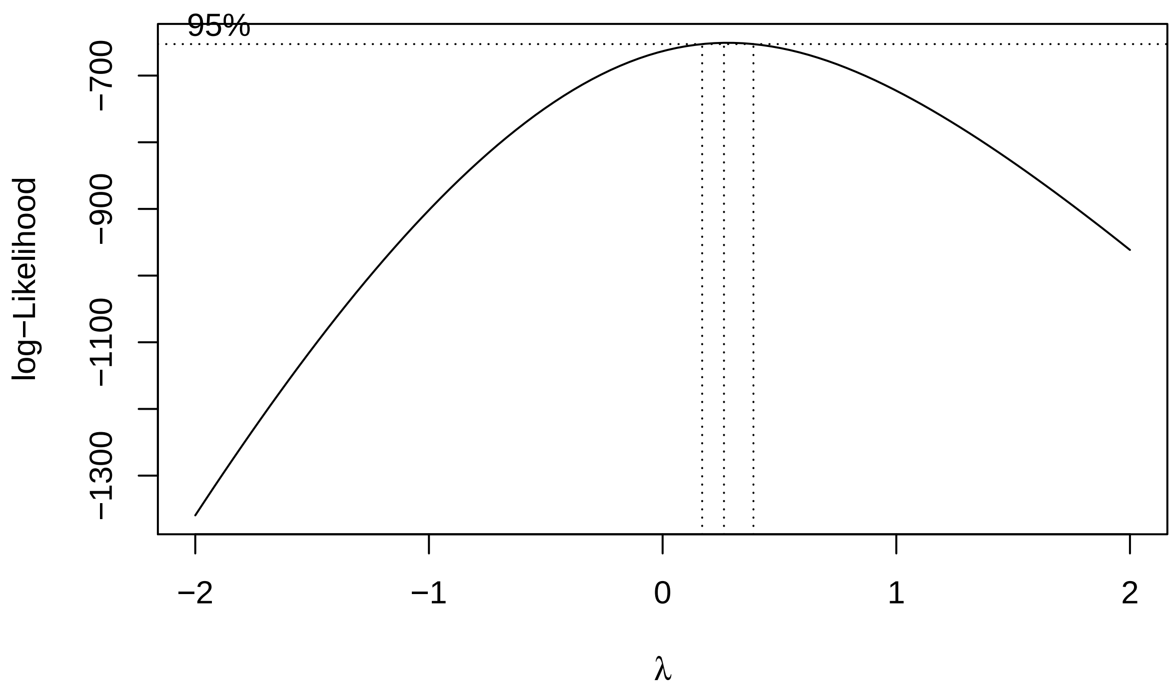
We will now fit a very simple model with temp, humidity and ibh as predictors

```r
mdl = lm(O3 ~ temp + humidity + ibh, data = ozone)
summary(mdl)
```

```
##
## Call:
## lm(formula = O3 ~ temp + humidity + ibh, data = ozone)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5291  -3.0137  -0.2249   2.8239  13.9303
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.049e+01  1.616e+00  -6.492 3.16e-10 ***
## temp         3.296e-01  2.109e-02  15.626  < 2e-16 ***
## humidity     7.738e-02  1.339e-02   5.777 1.77e-08 ***
## ibh         -1.004e-03  1.639e-04  -6.130 2.54e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.524 on 326 degrees of freedom
## Multiple R-squared:  0.684,  Adjusted R-squared:  0.6811
## F-statistic: 235.2 on 3 and 326 DF,  p-value: < 2.2e-16
```

We will now use Box-Cox method to find the best transformation of the response variable

```r
boxcox_mdl = MASS::boxcox(mdl)
```



Lets get the exact value of λ for which log-likelihood is maximum. That is the best transformation of the response variable.

```
#boxcox_mdl$y
boxcox_mdl$x[which(boxcox_mdl$y == max(boxcox_mdl$y))]
```

```
## [1] 0.2626263
```

Best transformation: $\lambda = 0.2626$

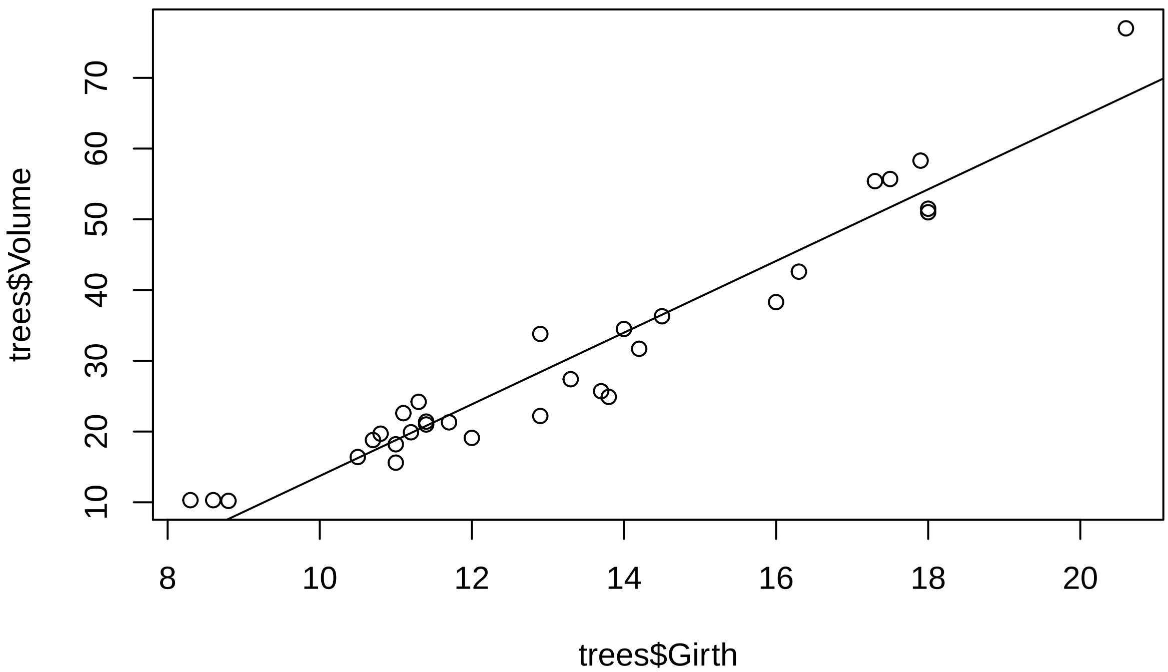
Question 3

```
par(mfrow=c(1,1))
head(trees)
```

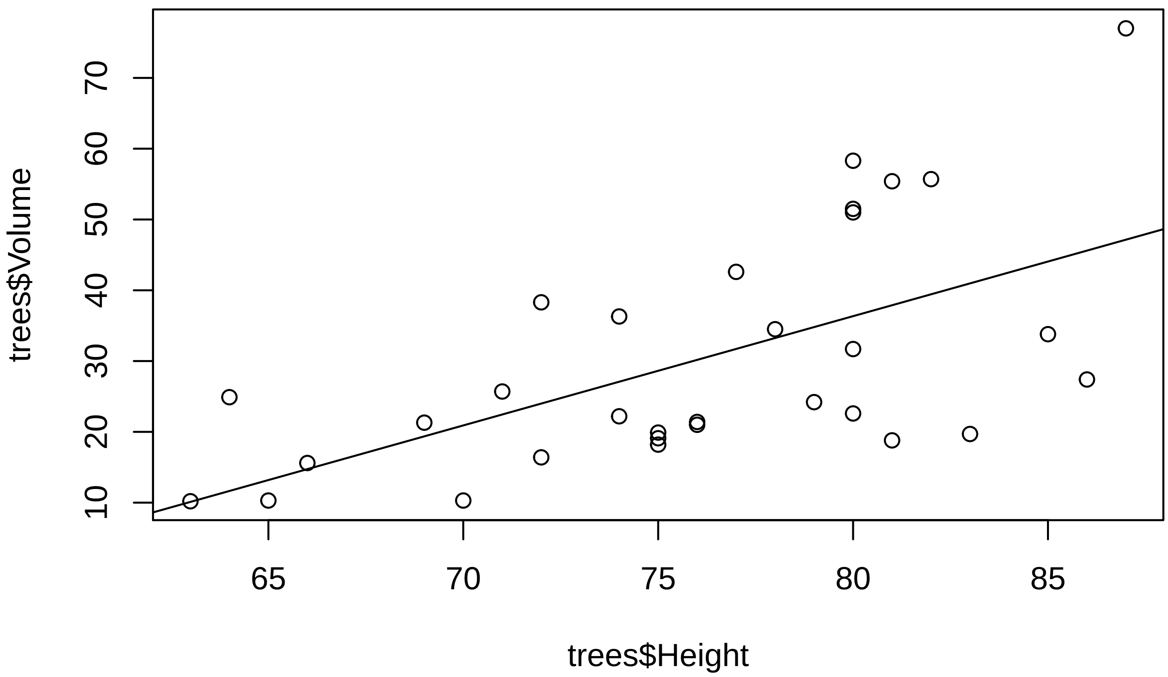
```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

Let us look at the plots of Volume vs Girth and Volume vs Height

```
plot(trees$Girth, trees$Volume)
abline(lsfit(trees$Girth, trees$Volume))
```

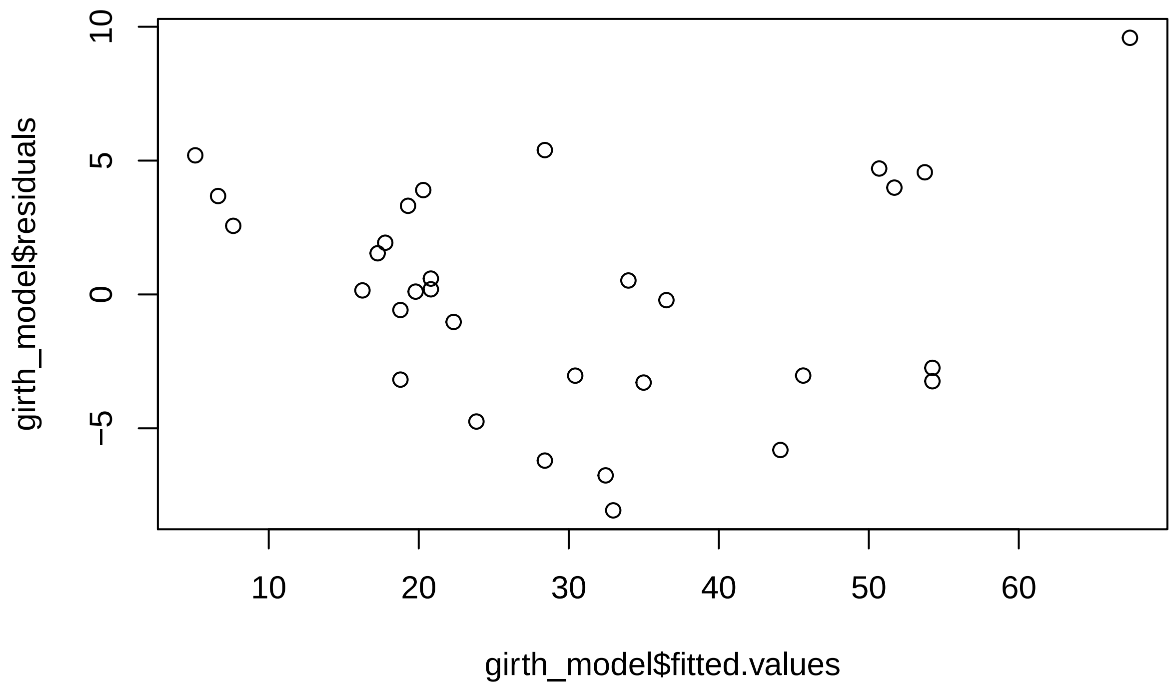



```
plot(trees$Height, trees$Volume)
abline(lsfit(trees$Height, trees$Volume))
```



We create a simple model with only Girth as predictor

```
girth_model = lm(Volume ~ Girth, data = trees)
plot(girth_model$fitted.values, girth_model$residuals)
```



The plot has a U shape, thus we might include I(Girth ^ 2) term

```
summary(girth_model)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

Lets see what happens if we add Height to predictors

```
simple_model = lm(Volume ~ Girth + Height, data = trees)
summary(simple_model)
```

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Multiple R-squared has increased . Now let us add I(Girth ˆ 2) term and replace Height by I(heightˆ 0.5) term since the plot between Volume and Height had a square root shape

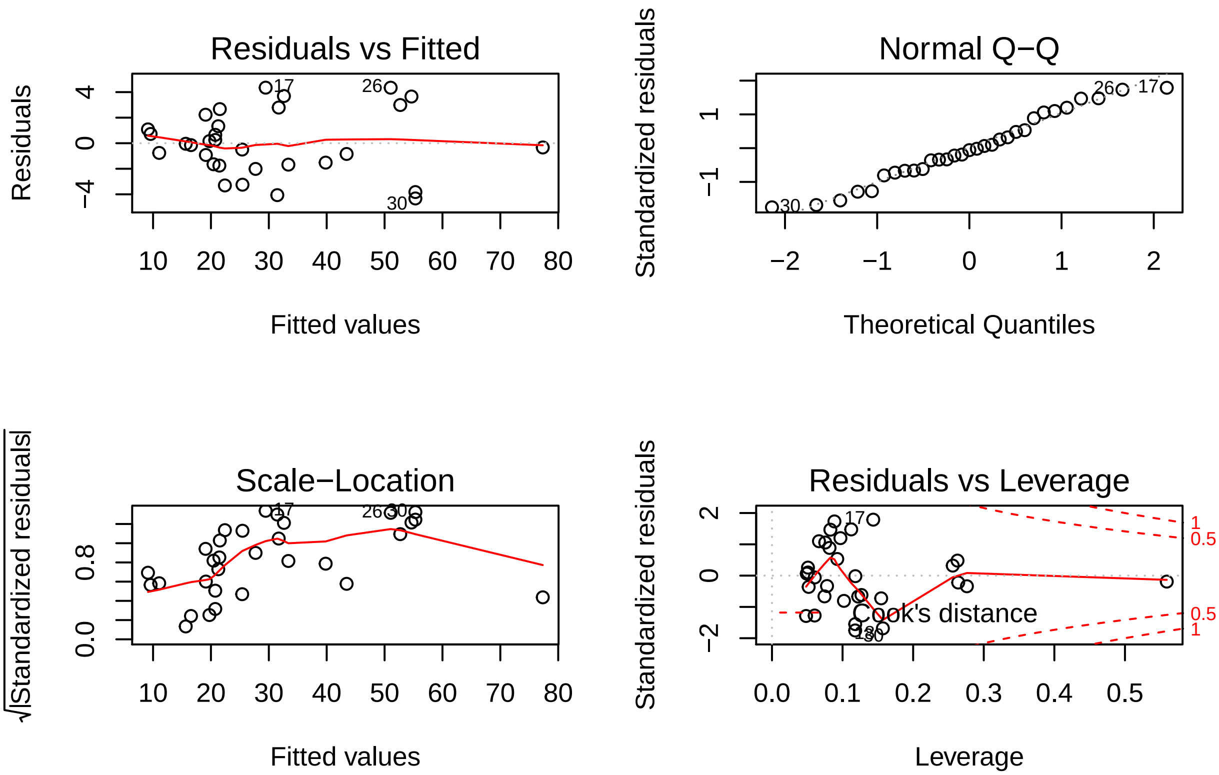
```r
trans2_model = lm(Volume ~ Girth + I(Girth ^ 2) + I(Height ^ 0.5), data = trees)
summary(trans2_model) # Multiple R-squared: 0.0.9771
```

```
##
## Call:
## lm(formula = Volume ~ Girth + I(Girth^2) + I(Height^0.5), data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3241 -1.6636 -0.1499  1.7799  4.3470
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -37.41967   14.36935  -2.604 0.014791 *
## Girth          -2.94786    1.31331  -2.245 0.033189 *
## I(Girth^2)      0.27090    0.04599   5.890 2.83e-06 ***
## I(Height^0.5)   6.48833    1.52458   4.256 0.000224 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.627 on 27 degrees of freedom
## Multiple R-squared:  0.977,  Adjusted R-squared:  0.9745
## F-statistic: 382.5 on 3 and 27 DF,  p-value: < 2.2e-16
```

The multiple R Squared value increased from 0.948 to 0.977 which is quite significant.

Also note that the significance level of I(Height ˆ 0.5) is much higher than that of Height inthe previous model.

```r
par(mfrow=c(2,2))
plot(trans2_model)
```



The diagnostic plots also look quite good. Thus our final model is Volume ~ Girth + I(Girth ^ 2) + I(Height ^ 0.5).

Question 4

Looking at the data

```
head(teengamb)
```

```
##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

(a) Backward Elimination – we will set the elimination criteria to be F-statistic should be less than 2 only then we will eliminate a variable.

```
#Backward elimiation
backward_model = lm(gamble ~ ., data = teengamb)
summary(backward_model)
```

```
##
## Call:
## lm(formula = gamble ~ ., data = teengamb)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -51.082 -11.320  -1.451   9.452  94.252 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  22.55565   17.19680   1.312   0.1968    
## sex         -22.11833    8.21111  -2.694   0.0101 *  
## status        0.05223    0.28111   0.186   0.8535    
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816 
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

```
drop1(backward_model, test = "F")
```

```
## Single term deletions
## 
## Model:
## gamble ~ sex + status + income + verbal
##        Df Sum of Sq   RSS    AIC F value    Pr(>F)    
## <none>                21624 298.18                      
## sex     1    3735.8 25360 303.67  7.2561   0.01011 *  
## status  1      17.8 21642 296.21  0.0345   0.85349    
## income  1   12056.2 33680 317.00 23.4169 1.792e-05 ***
## verbal  1     955.7 22580 298.21  1.8563   0.18031    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Status is having the smallest F-statistic so we eliminate status

```
backward_model = update(backward_model, . ~ . - status)
```

```
drop1(backward_model, test = "F")
```

```
## Single term deletions
## 
## Model:
## gamble ~ sex + income + verbal
##        Df Sum of Sq   RSS    AIC F value    Pr(>F)    
## <none>                21642 296.21                      
## sex     1    5787.9 27429 305.35 11.5001  0.001502 ** 
## income  1   13236.1 34878 316.64 26.2990 6.644e-06 ***
## verbal  1    1139.8 22781 296.63  2.2646  0.139667    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All the variables have an F-value greater than 2 thus we cannot eliminate any more variable. Therefore following is our final model.

```
summary(backward_model)
```

```
## 
```

```
## Call:
## lm(formula = gamble ~ sex + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.639 -11.765  -1.594   9.305  93.867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.1390    14.7686   1.634   0.1095
## sex         -22.9602     6.7706  -3.391   0.0015 **
## income        4.8981     0.9551   5.128 6.64e-06 ***
## verbal       -2.7468     1.8253  -1.505   0.1397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.43 on 43 degrees of freedom
## Multiple R-squared:  0.5263, Adjusted R-squared:  0.4933
## F-statistic: 15.93 on 3 and 43 DF,  p-value: 4.148e-07
```

(b) AIC method

```r
#AIC backward
full_model = lm(gamble ~ ., data = teengamb)
model.aic = step(full_model, direction = "backward", trace = 1)
```

```
## Start:  AIC=298.18
## gamble ~ sex + status + income + verbal
##
##          Df Sum of Sq   RSS    AIC
## - status  1      17.8 21642 296.21
## <none>                21624 298.18
## - verbal  1     955.7 22580 298.21
## - sex     1    3735.8 25360 303.67
## - income  1   12056.2 33680 317.00
##
## Step:  AIC=296.21
## gamble ~ sex + income + verbal
##
##          Df Sum of Sq   RSS    AIC
## <none>                21642 296.21
## - verbal  1    1139.8 22781 296.63
## - sex     1    5787.9 27429 305.35
## - income  1   13236.1 34878 316.64
```

```r
summary(model.aic)
```

```
##
## Call:
## lm(formula = gamble ~ sex + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.639 -11.765  -1.594   9.305  93.867
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.1390    14.7686   1.634   0.1095
## sex         -22.9602     6.7706  -3.391   0.0015 **
## income        4.8981     0.9551   5.128 6.64e-06 ***
## verbal       -2.7468     1.8253  -1.505   0.1397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.43 on 43 degrees of freedom
## Multiple R-squared:  0.5263, Adjusted R-squared:  0.4933
## F-statistic: 15.93 on 3 and 43 DF,  p-value: 4.148e-07
```

Here also we get the same model as Backward elimination method

(c) Adjusted R-squared method

```r
#Adjusted R2
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.4.2
```

```r
x = teengamb[,-5]
y = teengamb$gamble
best_model_r2 = leaps(x, y, nbest = 1, method = 'adjr2')
best_model_r2
```

```
## $which
##       1     2    3     4
## 1 FALSE FALSE TRUE FALSE
## 2  TRUE FALSE TRUE FALSE
## 3  TRUE FALSE TRUE  TRUE
## 4  TRUE  TRUE TRUE  TRUE
##
## $label
## [1] "(Intercept)" "1"           "2"           "3"           "4"
##
## $size
## [1] 2 3 4 5
##
## $adjr2
## [1] 0.3733570 0.4787240 0.4932879 0.4816495
```

Max Adjusted R^2 is given by 3rd model therefore we select the 3rd row in the which matrix to select the variables.

```r
colnames(x)[c(1, 3, 4)]
```

```
## [1] "sex"    "income" "verbal"
```

Same as (a) and (b)

(d) Mallows C_p

```r
#Cp
best_model_cp = leaps(x, y, nbest = 1, method = 'Cp')
best_model_cp
```

```
## $which
##       1     2    3     4
## 1 FALSE FALSE TRUE FALSE
```

```
## 2  TRUE FALSE TRUE FALSE
## 3  TRUE FALSE TRUE  TRUE
## 4  TRUE  TRUE TRUE  TRUE
##
## $label
## [1] "(Intercept)" "1"           "2"           "3"           "4"
##
## $size
## [1] 2 3 4 5
##
## $Cp
## [1] 11.401283  3.248323  3.034526  5.000000
```

Min Cp is given by 3rd model therefore we select the 3rd row in the which matrix to select the variables.

```r
colnames(x)[c(1, 3, 4)]
```

```
## [1] "sex"    "income" "verbal"
```

We get the same variables in all the four methods.