MA717: Applied Regression and Experimental Data Analysis

Assignment template

Meghana Dhongadi Ashoka - 2310246

Task 1: Data reading and simple exploration (25%)

1.1. Read "College.csv" file into R with following command and use dim() and head() to check if you read the data correct. You should report the number of observations and the number of variables. **(5 %)**

```
#read data
mydata<-read.csv("College.csv", header=T, stringsAsFactors=TRUE)
dim(mydata)
```

```
## [1] 775  17
```

```
head(mydata)
```

```
##   Private Apps Accept Enroll F.Undergrad P.Undergrad Outstate Room.Board Books
## 1     Yes 1660   1232    721        2885         537     7440       3300   450
## 2     Yes 2186   1924    512        2683        1227    12280       6450   750
## 3     Yes 1428   1097    336        1036          99    11250       3750   400
## 4     Yes  417    349    137         510          63    12960       5450   450
## 5     Yes  193    146     55         249         869     7560       4120   800
## 6     Yes  587    479    158         678          41    13500       3335   500
##   Personal PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate Elite
## 1     2200  70       78      18.1          12   7041        60    No
## 2     1500  29       30      12.2          16  10527        56    No
## 3     1165  53       66      12.9          30   8735        54    No
## 4      875  92       97       7.7          37  19016        59   Yes
## 5     1500  76       72      11.9           2  10922        15    No
## 6      675  67       73       9.4          11   9727        55    No
```

Number of observations - 775

Number of variables - 17

1.2. Use your registration number as random seed, generate a random subset of College data with sample size 700, name this new data as mynewdata. Use summary() to output the summarized information about mynewdata. Please report the number of private and public university and the number of Elite university and non-Elite university in this new data. **(12 %)**

```
##using my registration number as random seed
set.seed(2310246)

#generating a random subset of College data with sample size 700
```

```r
index<-sample(775,size=700)
mynewdata<-mydata[index, ]

dim(mynewdata)
```

```
## [1] 700  17
```

```r
#summarize mynewdata
summary(mynewdata)
```

```
##  Private        Apps           Accept            Enroll        F.Undergrad   
##  No :196   Min.   :   81   Min.   :   72.0   Min.   :  35.0   Min.   :  139  
##  Yes:504   1st Qu.:  779   1st Qu.:  599.8   1st Qu.: 242.0   1st Qu.:  991  
##            Median : 1600   Median : 1193.5   Median : 439.0   Median : 1722  
##            Mean   : 3089   Mean   : 2066.5   Mean   : 802.4   Mean   : 3798  
##            3rd Qu.: 3820   3rd Qu.: 2536.0   3rd Qu.: 925.5   3rd Qu.: 4292  
##            Max.   :48094   Max.   :26330.0   Max.   :6392.0   Max.   :31643  
##   P.Undergrad          Outstate       Room.Board       Books       
##  Min.   :    1.00   Min.   : 2340   Min.   :1780   Min.   :  96.0  
##  1st Qu.:   97.25   1st Qu.: 7248   1st Qu.:3580   1st Qu.: 475.0  
##  Median :  352.50   Median : 9912   Median :4180   Median : 521.5  
##  Mean   :  849.94   Mean   :10423   Mean   :4340   Mean   : 551.3  
##  3rd Qu.:  971.50   3rd Qu.:12906   3rd Qu.:5004   3rd Qu.: 600.0  
##  Max.   :21836.00   Max.   :21700   Max.   :7425   Max.   :2340.0  
##     Personal          PhD            Terminal       S.F.Ratio    
##  Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50  
##  1st Qu.: 875   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50  
##  Median :1210   Median : 75.00   Median : 82.0   Median :13.60  
##  Mean   :1353   Mean   : 72.79   Mean   : 79.9   Mean   :14.11  
##  3rd Qu.:1700   3rd Qu.: 85.25   3rd Qu.: 92.0   3rd Qu.:16.50  
##  Max.   :6800   Max.   :100.00   Max.   :100.0   Max.   :39.80  
##   perc.alumni        Expend        Grad.Rate       Elite    
##  Min.   : 0.00   Min.   : 3365   Min.   : 10.00   No :628  
##  1st Qu.:13.00   1st Qu.: 6790   1st Qu.: 53.00   Yes: 72  
##  Median :21.00   Median : 8412   Median : 65.00            
##  Mean   :22.77   Mean   : 9729   Mean   : 65.26            
##  3rd Qu.:31.00   3rd Qu.:10872   3rd Qu.: 77.00            
##  Max.   :64.00   Max.   :56233   Max.   :100.00            
```

```r
#reporting the number of private and public university
private<-table(mynewdata$Private)
cat('Number of Private universities: ',private['Yes'],'\n')
```

```
## Number of Private universities:  504
```

```r
cat('Number of Public universities: ',private['No'],'\n')
```

```
## Number of Public universities:  196
```

```r
#reporting the number of Elite university and non-Elite university
elite<-table(mynewdata$Elite)
cat('Number of Elite universities: ',elite['Yes'],'\n')
```

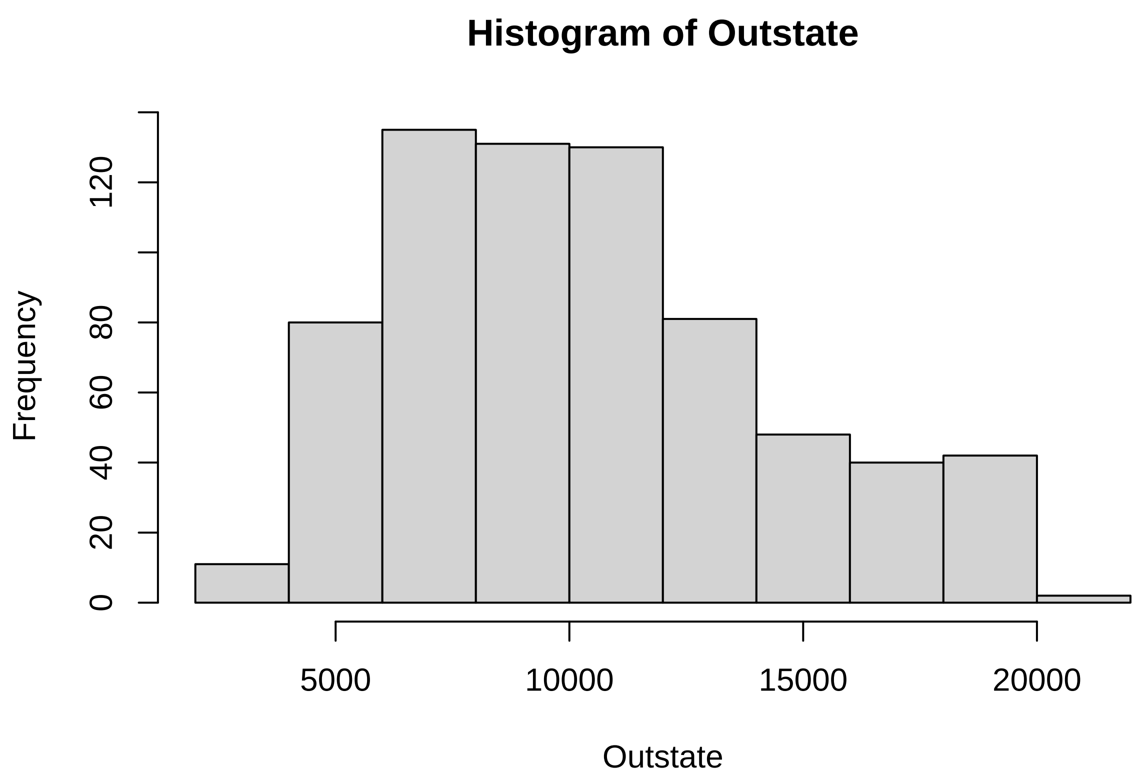
```
## Number of Elite universities:  72
```

```r
cat('Number of non-Elite universities: ',elite['No'],'\n')
```

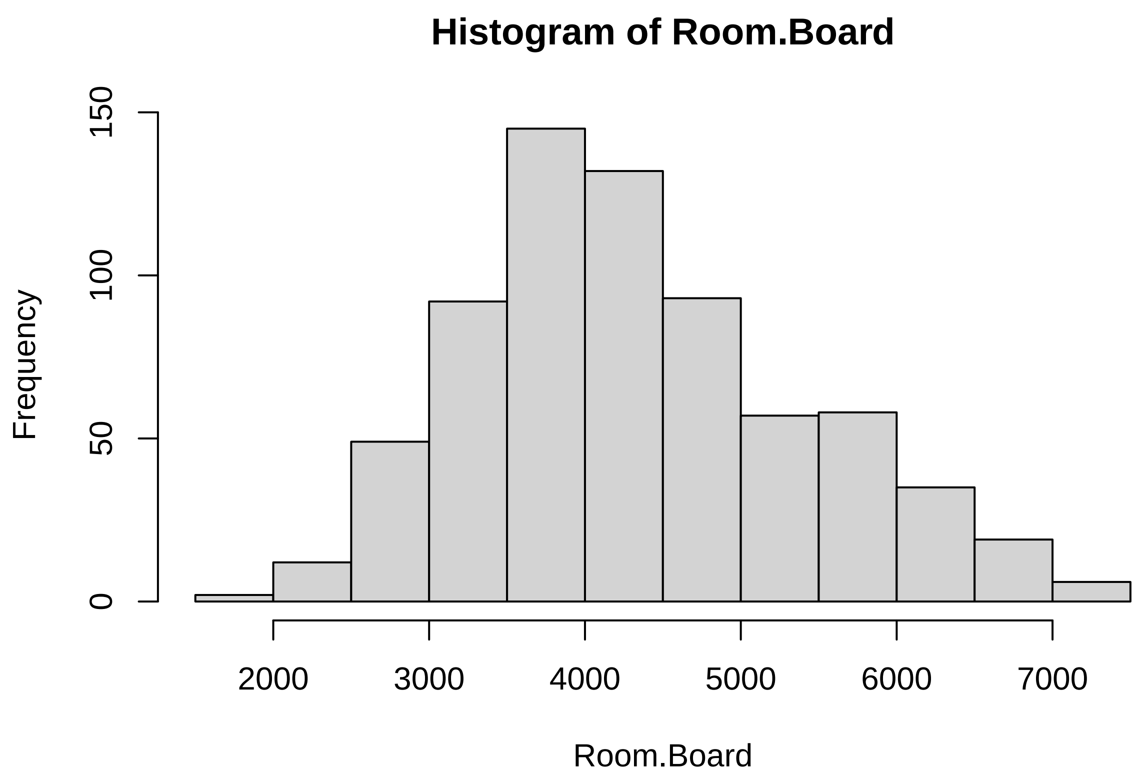
```
## Number of non-Elite universities:  628
```

1.3. Use mynewdata, plot histogram plots of four variables "Outstate", "Room.Board", "Books" and "Personal". Give each plot a suitable title and label for x axis and y axis. (**8**%)

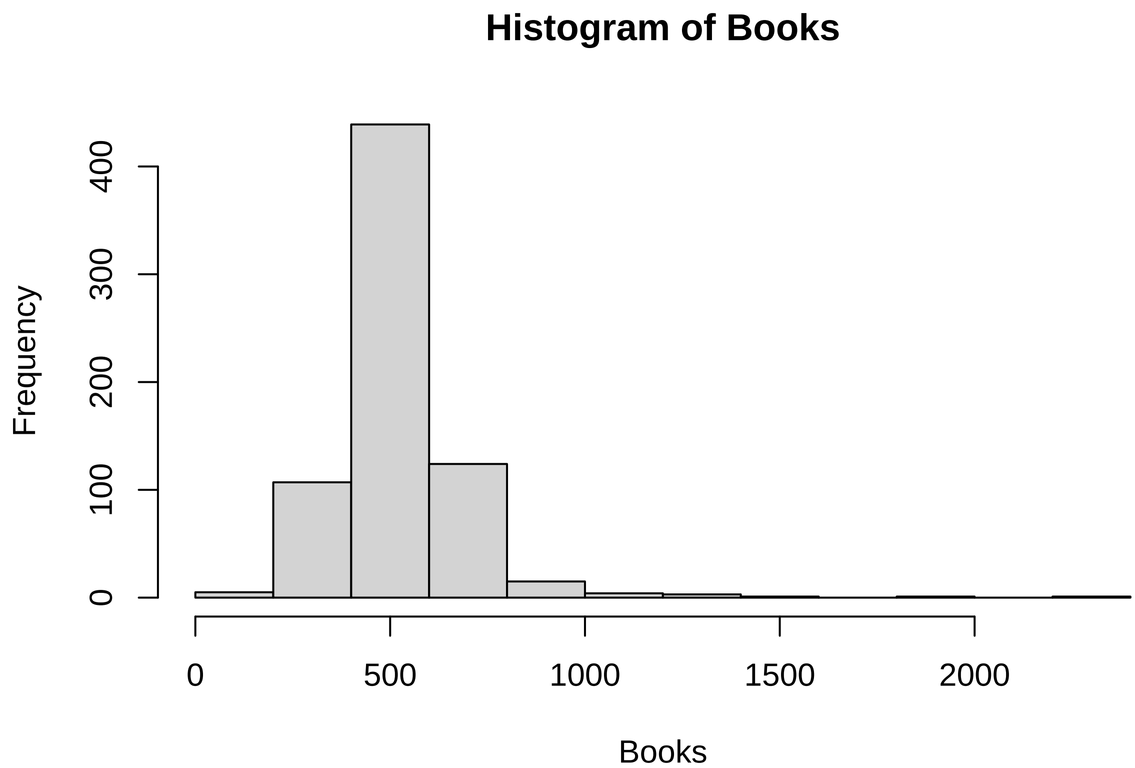
```r
#Let w is the variable 'Outstate', x is 'Room.Board', y is 'Books', and z is 'Personal'
#of mynewdata
w<-mynewdata$Outstate
x<-mynewdata$Room.Board
y<-mynewdata$Books
z<-mynewdata$Personal

#histogram plot
hist(w,main="Histogram of Outstate",xlab="Outstate")
```



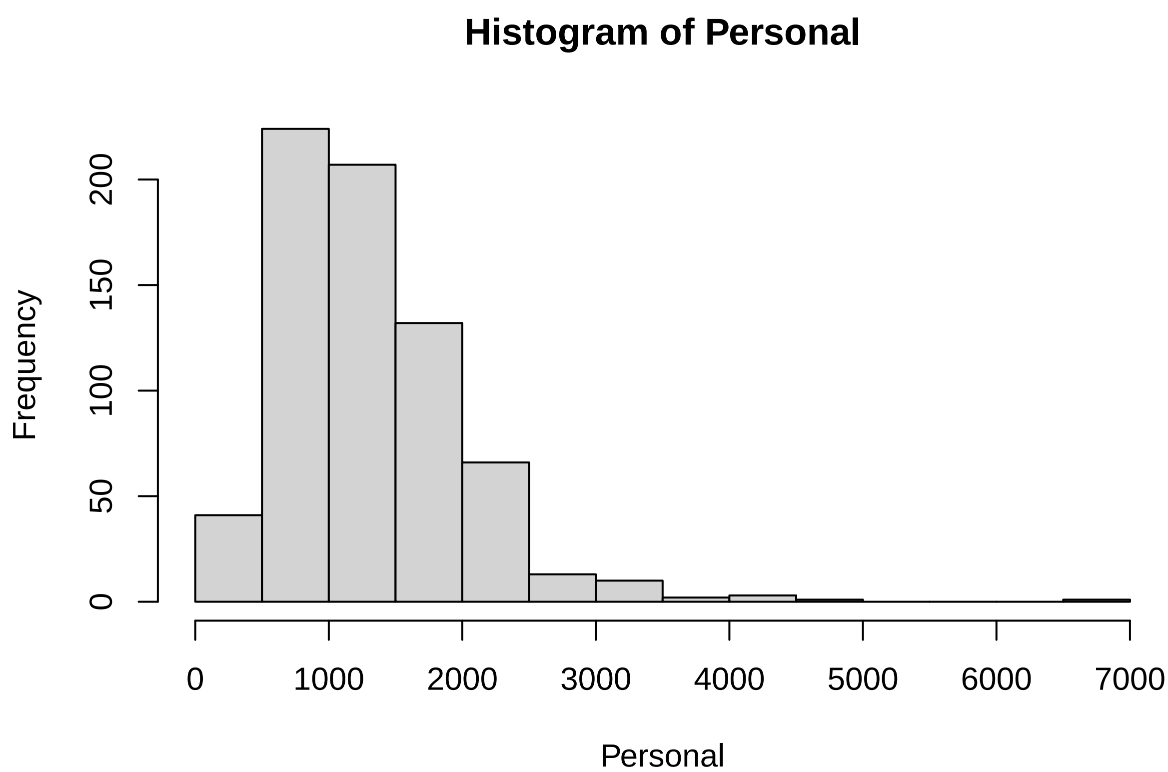
```r
hist(x,main="Histogram of Room.Board",xlab="Room.Board")
```




```r
hist(y,main="Histogram of Books",xlab="Books")
```




```r
hist(z,main="Histogram of Personal",xlab="Personal")
```



Task 2: Linear regression (45%)

2.1. Use mynewdata, do a linear regression fitting when outcome is "Grad.Rate" and predictors are "Private" and "Elite". Show the R output and report what you have learned from this output (you need to discuss significance, adjusted R-squared and p-value of F-statistics). **(6%)**.

```
#using linear regression to fit the data and summary the output
fitting<-lm(Grad.Rate~Private+Elite, data=mynewdata)
summary(fitting)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private + Elite, data = mynewdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.269  -9.439   0.731   9.731  44.561
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   55.439      1.066  52.012   <2e-16 ***
## PrivateYes    10.830      1.251   8.654   <2e-16 ***
## EliteYes      19.687      1.850  10.644   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 14.82 on 697 degrees of freedom
## Multiple R-squared:  0.2257, Adjusted R-squared:  0.2235
## F-statistic: 101.6 on 2 and 697 DF,  p-value: < 2.2e-16
```

Significance of coefficients - Both PrivateYes and EliteYes are highly significant with p-value <2e-16 *** which is smaller than 0.05.

Adjusted R-squared - The adjusted R-squared is 0.2235. This explains about 22.35 percent of the variability in Grad.Rate.

p-value of F-statistics - The F-statistics is 101.6 and p-value is <2.2e-16 which is smaller than 0.05. So, we reject the null hypothesis. The fitting model is better than the null model.

2.2. Use the linear regression fitting result in 2.1, calculate the confidence intervals for the coefficients. Also give the prediction interval of "Grad.Rate" for a new data with Private="Yes" and Elite="No". **(4%)**

```r
#confidence interval for the coefficients
confint(fitting)
```

```
##                 2.5 %   97.5 %
## (Intercept) 53.346410 57.53191
## PrivateYes   8.372823 13.28654
## EliteYes    16.055093 23.31794
```

```r
#prediction value and prediction interval of "Grad.Rate" for a new data with Private="Yes"
#and Elite="No"
predict(fitting, newdata=data.frame(Private="Yes", Elite="No"), interval="prediction")
```

```
##        fit      lwr      upr
## 1 66.26884 37.13309 95.40459
```

2.3 Use mynewdata, do a multiple linear regression fitting when outcome is "Grad.Rate", all other variables as predictors. Show the R output and report what you have learned from this output (you need to discuss significance, adjusted R-squared and p-value of F-statistics). Is linear regression model in 2.3 better than linear regression in 2.1? Use ANOVA to justify your conclusion. **(14%)**

```r
#using linear regression to fit the data and summary the output
fitting.full<-lm(Grad.Rate~., data=mynewdata)
summary(fitting.full)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ ., data = mynewdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.735  -7.320  -0.325   6.998  52.561
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.1684605  4.9454715   7.111 2.91e-12 ***
## PrivateYes   4.2822942  1.7360181   2.467 0.013880 *
## Apps         0.0016922  0.0004268   3.965 8.12e-05 ***
```

```
## Accept      -0.0014746  0.0008267  -1.784 0.074900 .
## Enroll       0.0019190  0.0023585   0.814 0.416128
## F.Undergrad -0.0001757  0.0004185  -0.420 0.674809
## P.Undergrad -0.0016754  0.0004024  -4.163 3.54e-05 ***
## Outstate     0.0010423  0.0002373   4.392 1.30e-05 ***
## Room.Board   0.0013343  0.0006220   2.145 0.032292 *
## Books       -0.0005698  0.0029118  -0.196 0.844908
## Personal    -0.0017331  0.0007890  -2.197 0.028382 *
## PhD          0.1958691  0.0577048   3.394 0.000728 ***
## Terminal    -0.0880281  0.0644621  -1.366 0.172521
## S.F.Ratio    0.0280500  0.1635404   0.172 0.863868
## perc.alumni  0.3151392  0.0500106   6.301 5.29e-10 ***
## Expend      -0.0004068  0.0001530  -2.659 0.008017 **
## EliteYes     5.3375437  2.0523590   2.601 0.009505 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.52 on 683 degrees of freedom
## Multiple R-squared:  0.4585, Adjusted R-squared:  0.4459
## F-statistic: 36.15 on 16 and 683 DF,  p-value: < 2.2e-16
```

```r
#anova for simple and full model
anova(fitting,fitting.full)
```

```
## Analysis of Variance Table
##
## Model 1: Grad.Rate ~ Private + Elite
## Model 2: Grad.Rate ~ Private + Apps + Accept + Enroll + F.Undergrad +
##     P.Undergrad + Outstate + Room.Board + Books + Personal +
##     PhD + Terminal + S.F.Ratio + perc.alumni + Expend + Elite
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    697 153153
## 2    683 107096 14     46057 20.98 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significance of coefficients - In this fitting, all predictors except "Enroll", "F.Undergrad", "Books", "Terminal" and "S.F.Ratio" are significantly associated with "Grad.Rate".

Adjusted R-squared - The adjusted R-squared is 0.4459. This explains about 44.59 percent of the variability in Grad.Rate.

p-value of F-statistics - The p-value is < 2.2e-16 which is smaller than 0.05. So, we reject Null hypothesis.The fitting model is better than the null model.

My interpretation from ANOVA table:

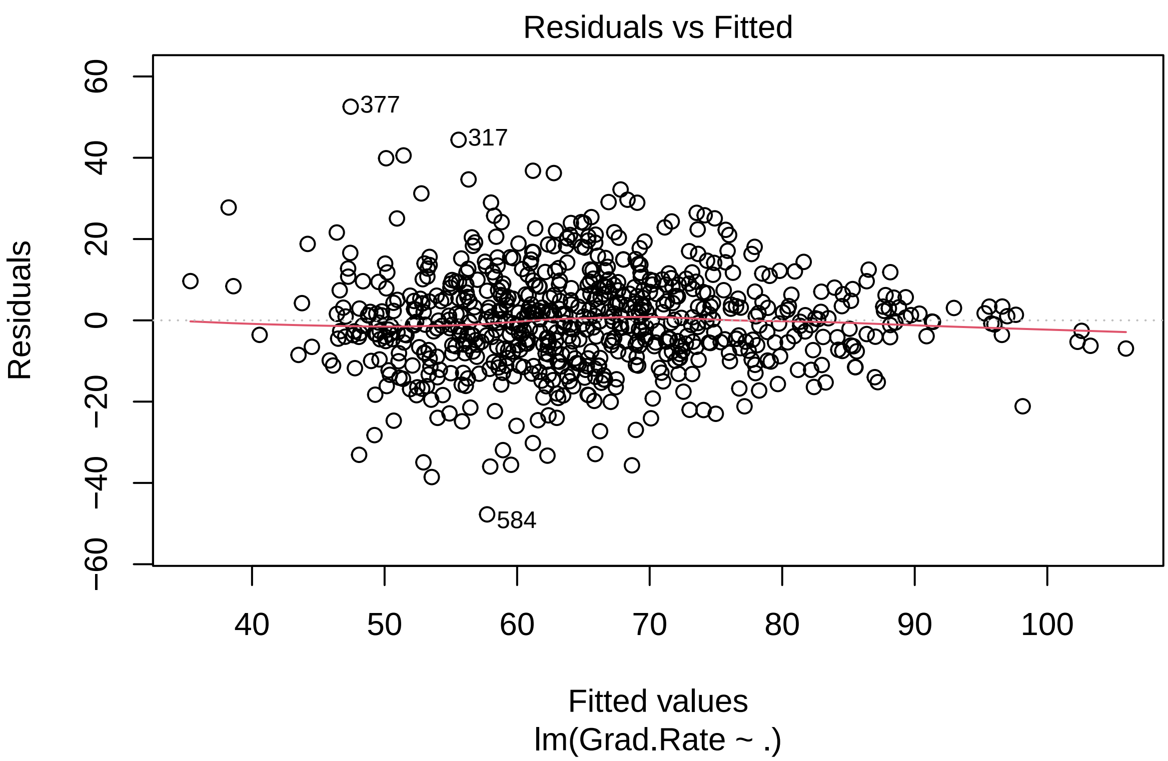
- **Degrees of Freedom** - Model 2 has more degrees of freedom than Model 1. Model 2 is therefore more complicated.
- **Residual Sum of Squares (RSS)** - Model 2 has a smaller residual sum of squares (RSS), which suggests a better fit to the data.
- **F-statistic** - A model that fits the data better has a higher F-statistic.

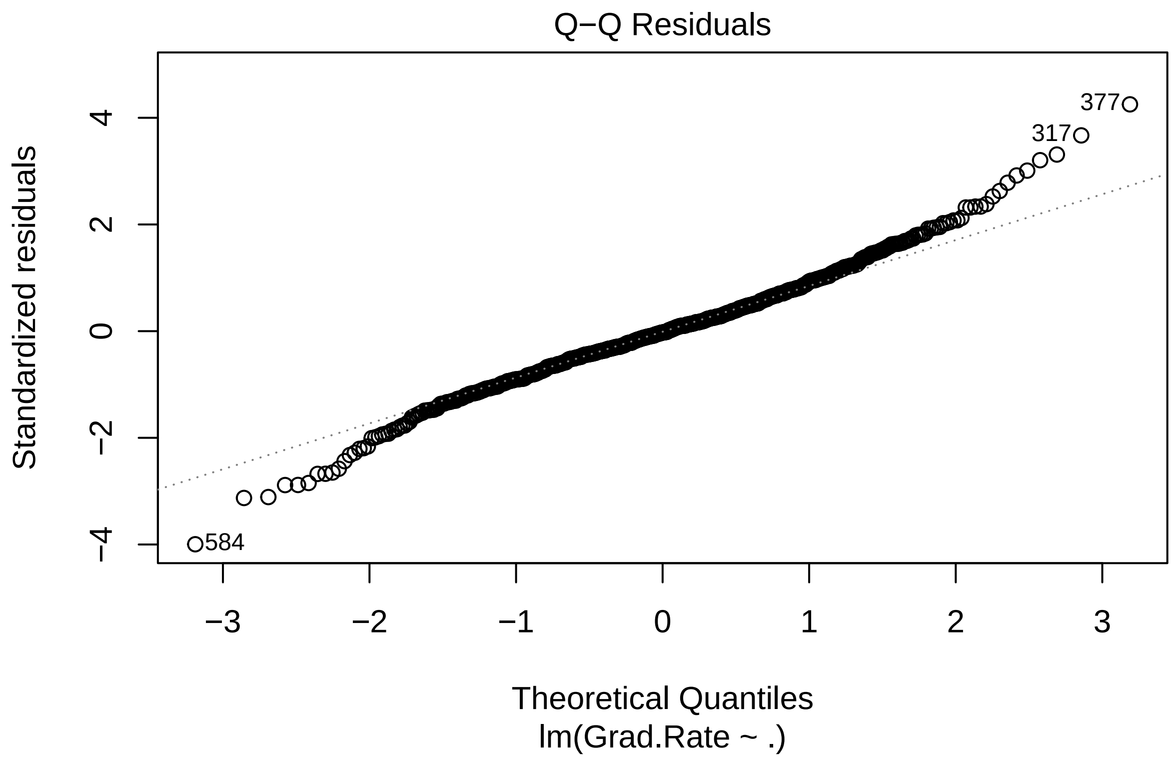
- **p-value[Pr(>F)]** - The p-value approaches zero quite closely. Model 2 is statistically significant as a result.

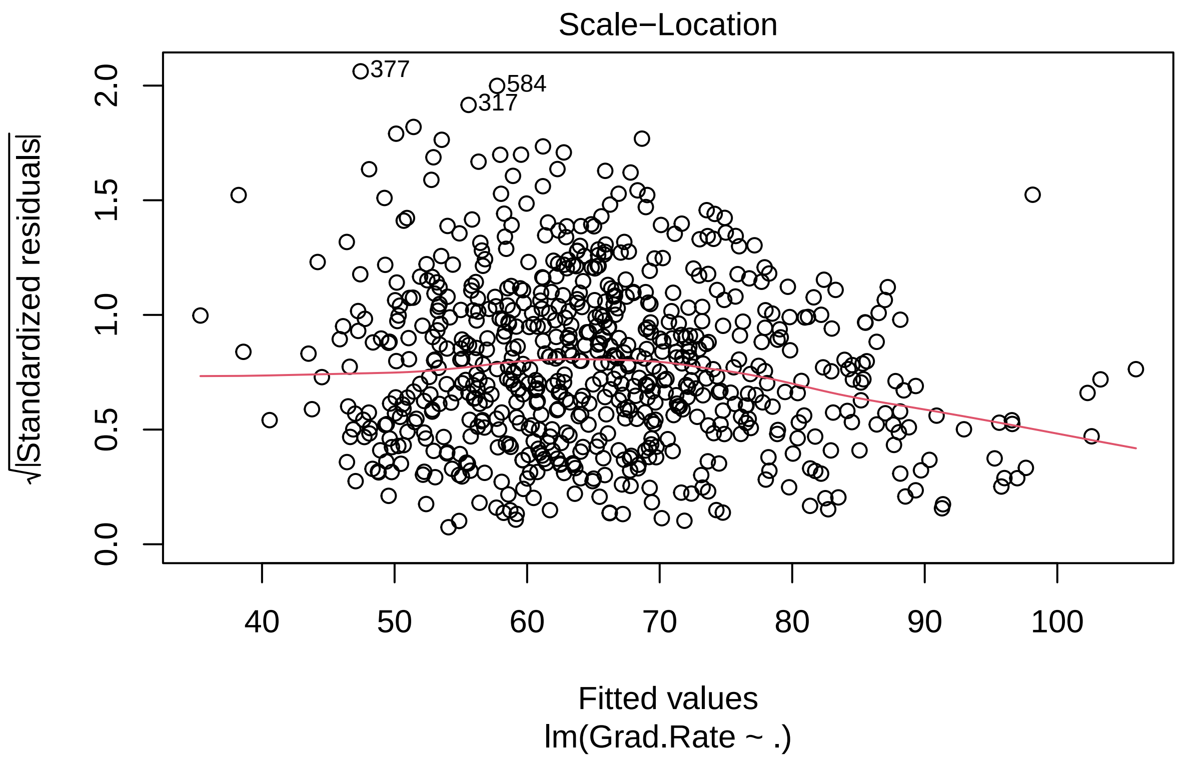
Conclusion - Model 2 includes more set of predictor variables, which is significantly better at explaining the variation in the Grad.Rate compared to Model 1. The adjust R-squared is 0.4459 for multiple regression model, which is higher than 0.2235 in simple linear regression model. The higher R-squared, the more variability of outcome "Grad.Rate" explained so the multiple regression model is better than simple linear regression model. Therefore, based on the ANOVA results, Model 2 is better than Model 1.

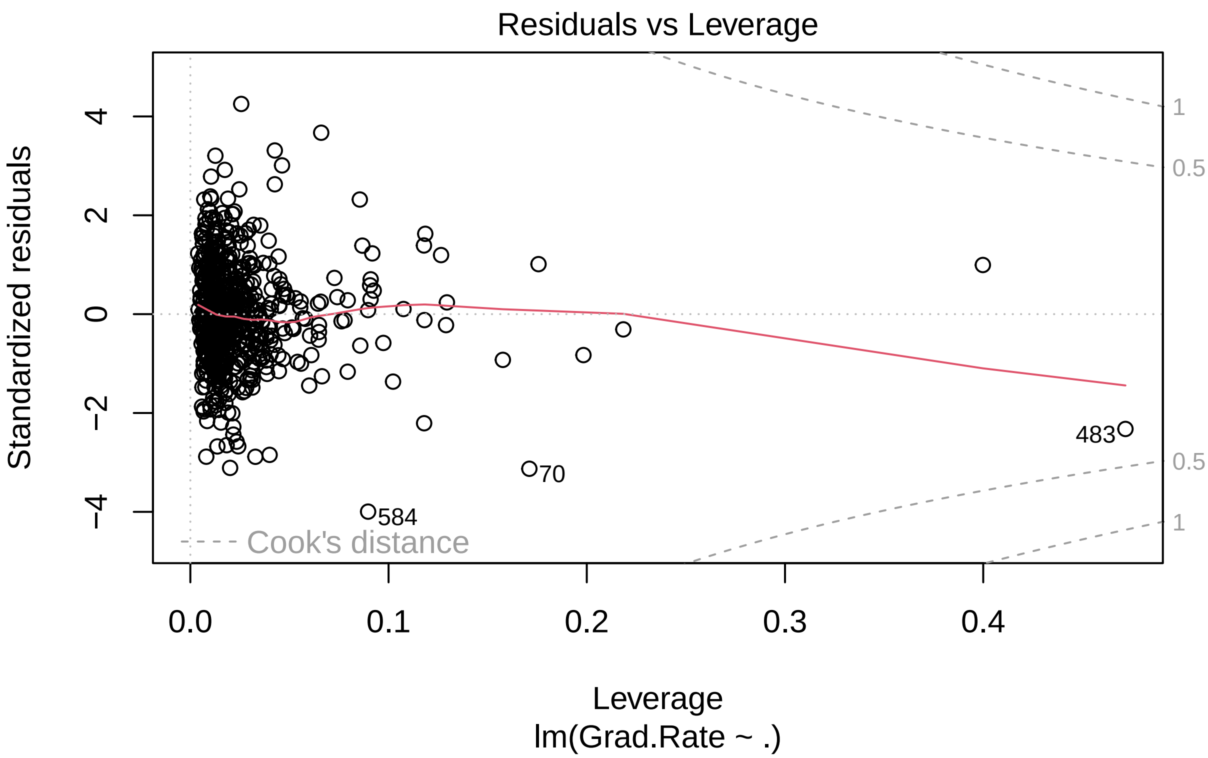
2.4. Use the diagnostic plots to look at the fitting of multiple linear regression in 2.3. Please comment what you have seen from those plots. **(7%)**

```r
#diagnostic plots
plot(fitting.full)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Grad.Rate ~ .)
377
317
584
−4
−2
0
2
4
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
377
584
317
40
50
60
70
80
90
100
Fitted values
lm(Grad.Rate ~ .)



Residual plot - The red line in the plot shows a slight curve indicating non-linearity between outcome and predictors.

Q-Q plot - The residuals are not fully normal as all points do not align with the diagonal line.

Scale-Location plot - The red line on the plot is roughly horizontal across the plot which satisfy homoscedasticity assumption (constant variance).

Leverage plot - The plot shows that there are some high influence points.

2.5. Use mynewdata, do a variable selection to choose the best model. You should use plots to justify how do you choose your best model. Use the selected predictors of your best model with outcome "Grad.Rate", do a linear regression fitting and plot the diagnostic plots for this fitting. You can use either exhaustive, or forward, or backward selection method. **(14%)**

```r
#load library(leaps)
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.3.2
```

```r
#Use regsubsets to get exhaustive search
myfit.regsub<-regsubsets(Grad.Rate~., data=mynewdata, nvmax=16)
myfit.regsub.sum<-summary(myfit.regsub)

par(mfrow=c(2,2))

#plot rss
```

```r
plot(myfit.regsub.sum$rss, xlab="Variable number", ylab="RSS", type="l")

#plot adjusted R-square and maximum point
plot(myfit.regsub.sum$adjr2, xlab="Variable number", ylab="Adjust R-squared", type="l")
#check which model gives the maximum adjusted R-squared
which.max(myfit.regsub.sum$adjr2)
```

```
## [1] 12
```

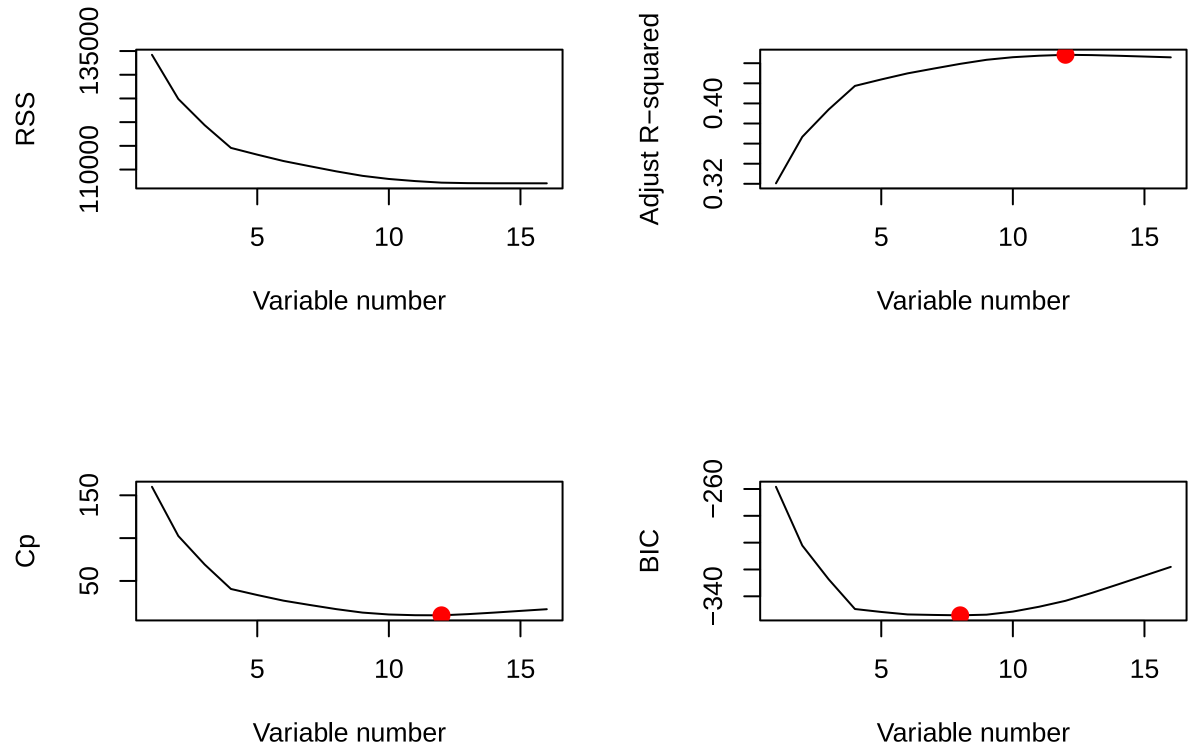
```r
points(12, myfit.regsub.sum$adjr2[12], col="red", cex=2, pch=20)

# plot cp and minimum point
plot(myfit.regsub.sum$cp, xlab="Variable number", ylab="Cp", type="l")
#check which model gives the minimum cp
which.min(myfit.regsub.sum$cp)
```

```
## [1] 12
```

```r
points(12, myfit.regsub.sum$cp[12], col="red", cex=2, pch=20)

#plot BIC and minimum point
plot(myfit.regsub.sum$bic, xlab="Variable number", ylab="BIC", type="l")
#check which model gives the minimum bic
which.min(myfit.regsub.sum$bic)
```

```
## [1] 8
```

```r
points(8, myfit.regsub.sum$bic[8], col="red", cex=2, pch=20)
```




```r
#check the coefficients of model 8
coef(myfit.regsub, 8)
```

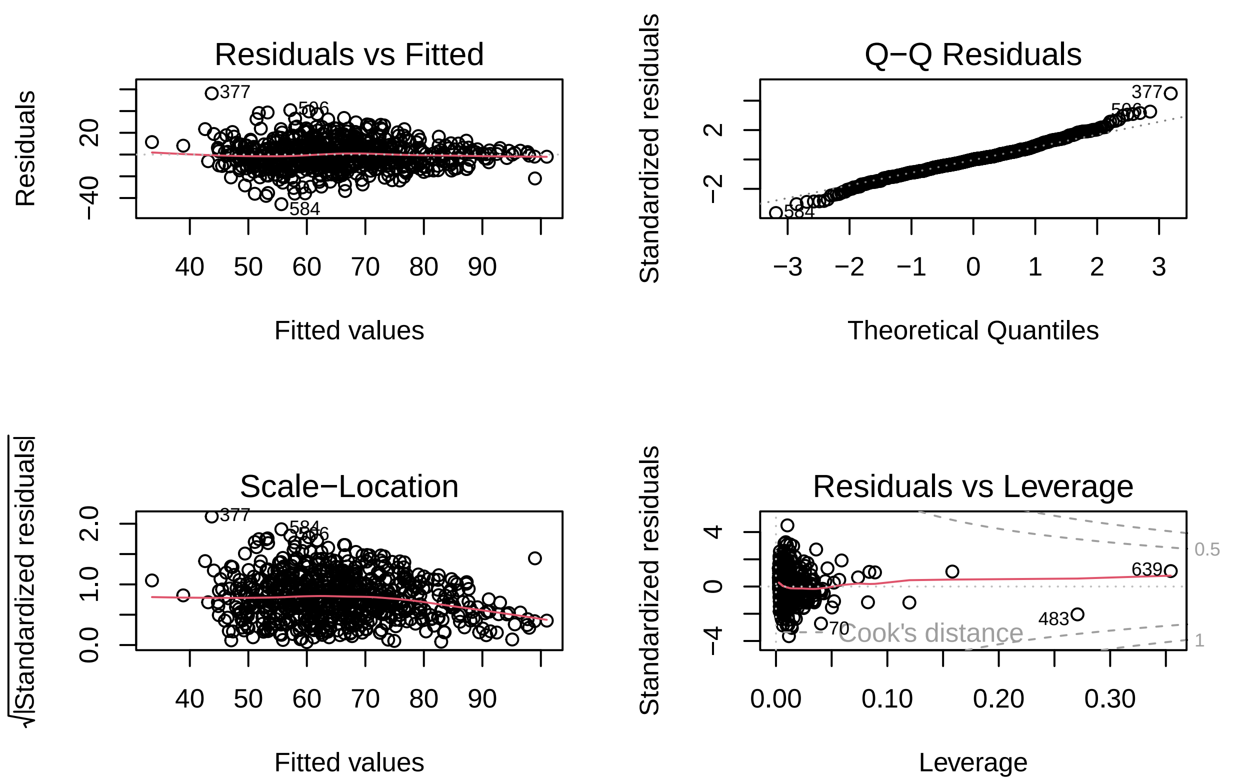
```
##   (Intercept)    PrivateYes          Apps   P.Undergrad      Outstate
## 32.8509962291  5.1856068639  0.0010215630 -0.0018070571  0.0012056953
##           PhD   perc.alumni        Expend      EliteYes
##  0.1416733754  0.3164279180 -0.0003799887  6.5206467688
```

```r
# use lm() and summary() to do linear regression model
myfit.lm.best=lm(Grad.Rate~Private+Apps+P.Undergrad+Outstate+PhD+perc.alumni
                 +Expend+Elite, data=mynewdata)
summary(myfit.lm.best)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private + Apps + P.Undergrad + Outstate +
##     PhD + perc.alumni + Expend + Elite, data = mynewdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.647  -7.719   0.033   6.868  56.255
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.8509962  2.6136528  12.569  < 2e-16 ***
```

```
## PrivateYes   5.1856069  1.6491045   3.144 0.001735 **
## Apps         0.0010216  0.0001595   6.405 2.78e-10 ***
## P.Undergrad -0.0018071  0.0003766  -4.799 1.96e-06 ***
## Outstate     0.0012057  0.0002104   5.730 1.50e-08 ***
## PhD          0.1416734  0.0383898   3.690 0.000242 ***
## perc.alumni  0.3164279  0.0482605   6.557 1.08e-10 ***
## Expend      -0.0003800  0.0001376  -2.761 0.005912 **
## EliteYes     6.5206468  1.9700651   3.310 0.000982 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.6 on 691 degrees of freedom
## Multiple R-squared:  0.4458, Adjusted R-squared:  0.4394
## F-statistic: 69.47 on 8 and 691 DF,  p-value: < 2.2e-16
```

```r
plot(myfit.lm.best)
```



The best model we tend to choose is the one that uses the least number of variables. So, here we can look at the model selected by BIC, is model with **8 variables**. Out of all predictors, our best model selected by **exhaustive search** includes - **Private, Apps, P.Undergrad, Outstate, PhD, perc.alumni, Expend, Elite**.

These 8 variables are used in the model and we reuse lm() to fit the data to get the information. From the results, we can see that all 8 variables are significant (variables are not all the same significant variables as the linear regression result at the beginning) but the model fitting is not the best.

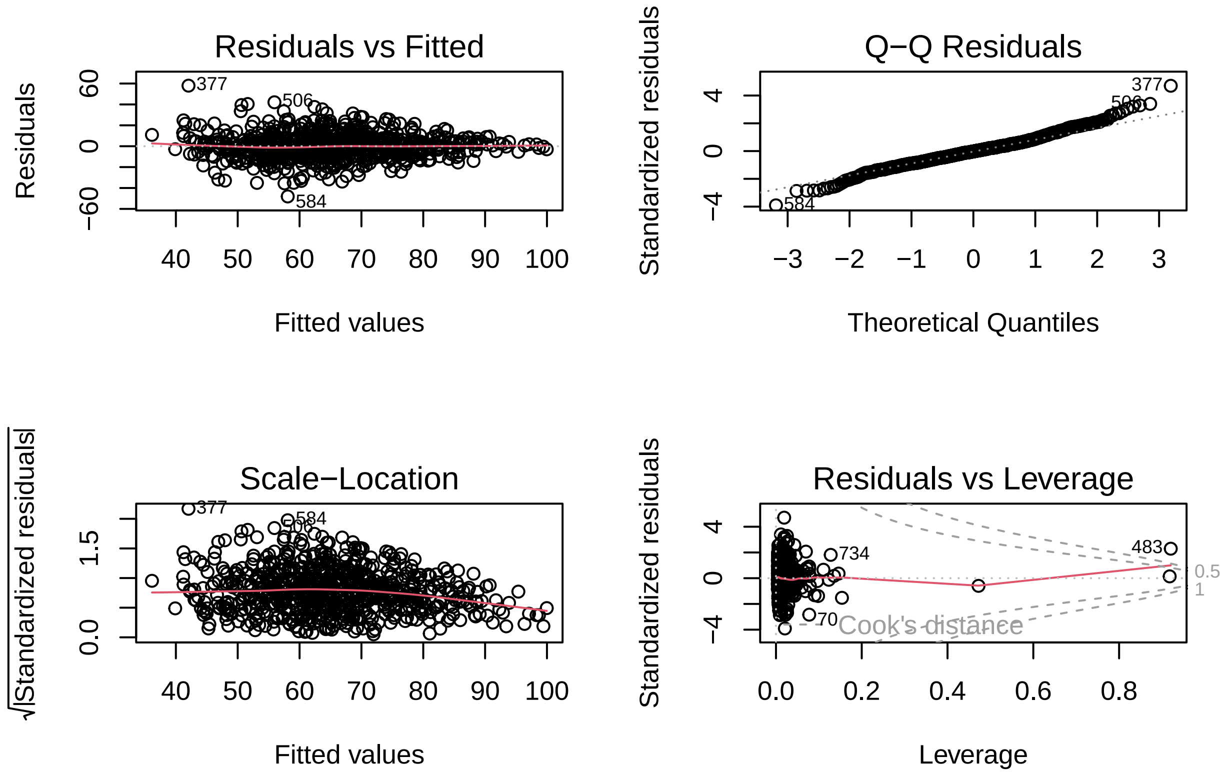
Task 3: Open question (30%)

Use mynewdata, discuss and perform any step(s) that you think that can improve the fitting in Task 2. You need to illustrate your work by using the R codes, output and discussion.

```
#using lm(), poly() and summary() to do polynomial linear regression with degree 2
myfit.pol<-lm(Grad.Rate~poly(Private,2,raw=T)+poly(Apps,2,raw=T)+poly(P.Undergrad,2,raw=T)
              +poly(Outstate,2,raw=T)+poly(PhD,2,raw=T)+poly(Expend,2,raw=T)+poly(Elite,2,raw=T)
              +poly(perc.alumni,2,raw=T),data=mynewdata)
summary(myfit.pol)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ poly(Private, 2, raw = T) + poly(Apps,
##     2, raw = T) + poly(P.Undergrad, 2, raw = T) + poly(Outstate,
##     2, raw = T) + poly(PhD, 2, raw = T) + poly(Expend, 2, raw = T) +
##     poly(Elite, 2, raw = T) + poly(perc.alumni, 2, raw = T),
##     data = mynewdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.087  -7.601  -0.284   6.663  57.956
##
## Coefficients: (2 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.926e+01  7.661e+00   2.514  0.01217 *
## poly(Private, 2, raw = T)1      5.754e+00  1.780e+00   3.233  0.00128 **
## poly(Private, 2, raw = T)2             NA         NA      NA       NA
## poly(Apps, 2, raw = T)1         1.756e-03  2.840e-04   6.182 1.08e-09 ***
## poly(Apps, 2, raw = T)2        -2.623e-08  8.689e-09  -3.019  0.00263 **
## poly(P.Undergrad, 2, raw = T)1 -2.761e-03  7.028e-04  -3.929 9.40e-05 ***
## poly(P.Undergrad, 2, raw = T)2  6.471e-08  4.522e-08   1.431  0.15287
## poly(Outstate, 2, raw = T)1     1.679e-03  7.109e-04   2.362  0.01847 *
## poly(Outstate, 2, raw = T)2    -1.419e-08  2.972e-08  -0.477  0.63332
## poly(PhD, 2, raw = T)1          1.388e-01  1.875e-01   0.741  0.45925
## poly(PhD, 2, raw = T)2          7.029e-05  1.467e-03   0.048  0.96179
## poly(Expend, 2, raw = T)1      -1.218e-03  3.722e-04  -3.273  0.00112 **
## poly(Expend, 2, raw = T)2       1.773e-08  7.119e-09   2.490  0.01300 *
## poly(Elite, 2, raw = T)1        6.393e+00  1.995e+00   3.204  0.00142 **
## poly(Elite, 2, raw = T)2               NA         NA      NA       NA
## poly(perc.alumni, 2, raw = T)1  5.804e-01  1.474e-01   3.939 9.04e-05 ***
## poly(perc.alumni, 2, raw = T)2 -4.945e-03  2.622e-03  -1.886  0.05971 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.44 on 685 degrees of freedom
## Multiple R-squared:  0.4638, Adjusted R-squared:  0.4529
## F-statistic: 42.33 on 14 and 685 DF,  p-value: < 2.2e-16
```

```
# diagnostic plots
par(mfrow=c(2,2)) # make 2 by 2 plots
plot(myfit.pol)
```

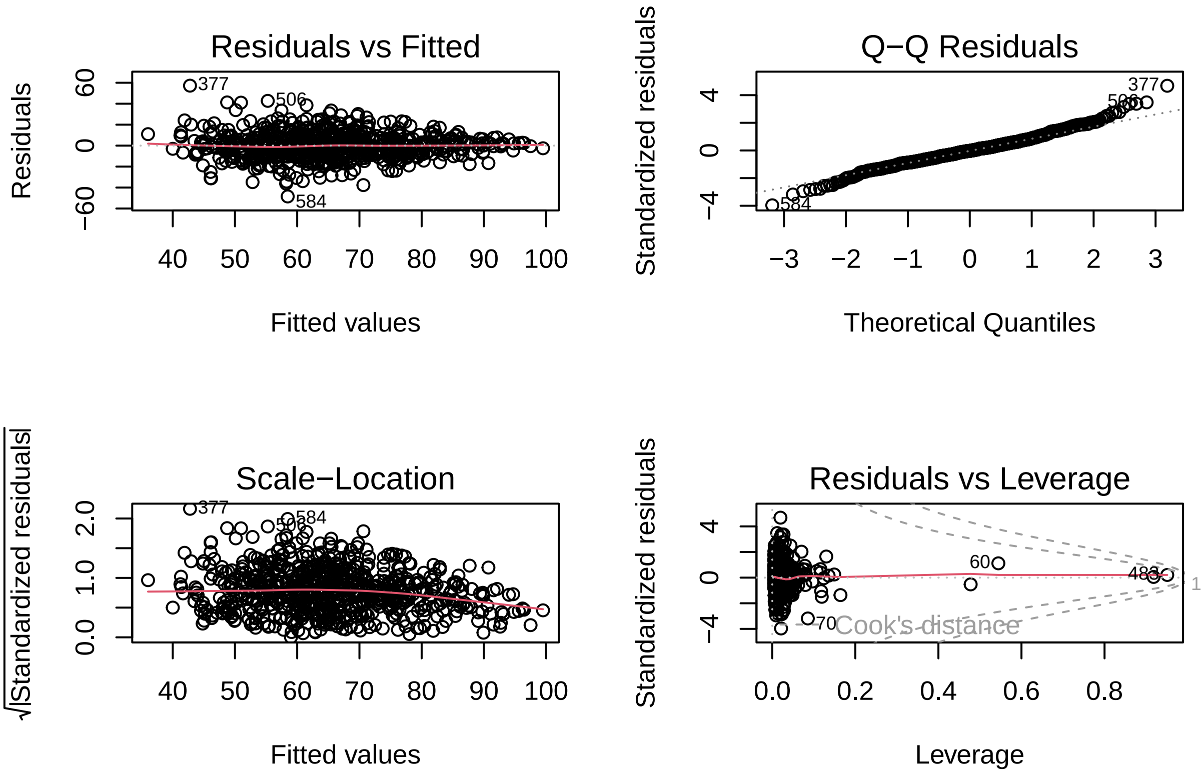



```r
#using lm(), poly() and summary() and add interaction
myfit.pol.int<-lm(Grad.Rate~poly(Private,2,raw=T)*poly(Apps,2,raw=T)+poly(P.Undergrad,2,raw=T)
                 +poly(Outstate,2,raw=T)+poly(PhD,2,raw=T)+poly(Expend,2,raw=T)+poly(Elite,2,raw=T)
                 +poly(perc.alumni,2,raw=T),data=mynewdata)
summary(myfit.pol.int)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ poly(Private, 2, raw = T) * poly(Apps,
##     2, raw = T) + poly(P.Undergrad, 2, raw = T) + poly(Outstate,
##     2, raw = T) + poly(PhD, 2, raw = T) + poly(Expend, 2, raw = T) +
##     poly(Elite, 2, raw = T) + poly(perc.alumni, 2, raw = T),
##     data = mynewdata)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -48.472  -7.659  -0.341   6.891  57.226
##
## Coefficients: (4 not defined because of singularities)
##                                                         Estimate Std. Error
## (Intercept)                                            2.351e+01  7.817e+00
## poly(Private, 2, raw = T)1                             1.648e+00  2.367e+00
## poly(Private, 2, raw = T)2                                    NA         NA
## poly(Apps, 2, raw = T)1                               -9.784e-04  9.054e-04
## poly(Apps, 2, raw = T)2                                1.360e-07  4.854e-08
```

```
## poly(P.Undergrad, 2, raw = T)1                       -2.723e-03  6.992e-04
## poly(P.Undergrad, 2, raw = T)2                        6.497e-08  4.497e-08
## poly(Outstate, 2, raw = T)1                           2.192e-03  7.518e-04
## poly(Outstate, 2, raw = T)2                          -4.183e-08  3.215e-08
## poly(PhD, 2, raw = T)1                                1.310e-01  1.867e-01
## poly(PhD, 2, raw = T)2                               -5.122e-05  1.462e-03
## poly(Expend, 2, raw = T)1                            -1.099e-03  3.711e-04
## poly(Expend, 2, raw = T)2                             1.516e-08  7.122e-09
## poly(Elite, 2, raw = T)1                              6.190e+00  1.982e+00
## poly(Elite, 2, raw = T)2                                     NA         NA
## poly(perc.alumni, 2, raw = T)1                        5.744e-01  1.463e-01
## poly(perc.alumni, 2, raw = T)2                       -4.653e-03  2.605e-03
## poly(Private, 2, raw = T)1:poly(Apps, 2, raw = T)1  2.373e-03  7.059e-04
## poly(Private, 2, raw = T)2:poly(Apps, 2, raw = T)1         NA         NA
## poly(Private, 2, raw = T)1:poly(Apps, 2, raw = T)2 -1.522e-07  4.575e-08
## poly(Private, 2, raw = T)2:poly(Apps, 2, raw = T)2         NA         NA
##                                                    t value Pr(>|t|)
## (Intercept)                                          3.008 0.002731 **
## poly(Private, 2, raw = T)1                           0.696 0.486570
## poly(Private, 2, raw = T)2                              NA       NA
## poly(Apps, 2, raw = T)1                             -1.081 0.280267
## poly(Apps, 2, raw = T)2                              2.802 0.005215 **
## poly(P.Undergrad, 2, raw = T)1                      -3.894 0.000108 ***
## poly(P.Undergrad, 2, raw = T)2                       1.445 0.148996
## poly(Outstate, 2, raw = T)1                          2.916 0.003657 **
## poly(Outstate, 2, raw = T)2                         -1.301 0.193659
## poly(PhD, 2, raw = T)1                               0.702 0.483154
## poly(PhD, 2, raw = T)2                              -0.035 0.972059
## poly(Expend, 2, raw = T)1                           -2.961 0.003170 **
## poly(Expend, 2, raw = T)2                            2.129 0.033587 *
## poly(Elite, 2, raw = T)1                             3.124 0.001862 **
## poly(Elite, 2, raw = T)2                                NA       NA
## poly(perc.alumni, 2, raw = T)1                       3.926 9.53e-05 ***
## poly(perc.alumni, 2, raw = T)2                      -1.786 0.074502 .
## poly(Private, 2, raw = T)1:poly(Apps, 2, raw = T)1   3.362 0.000818 ***
## poly(Private, 2, raw = T)2:poly(Apps, 2, raw = T)1      NA       NA
## poly(Private, 2, raw = T)1:poly(Apps, 2, raw = T)2  -3.327 0.000923 ***
## poly(Private, 2, raw = T)2:poly(Apps, 2, raw = T)2      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.35 on 683 degrees of freedom
## Multiple R-squared:  0.4731, Adjusted R-squared:  0.4608
## F-statistic: 38.33 on 16 and 683 DF,  p-value: < 2.2e-16
```

```
# diagnostic plots
par(mfrow=c(2,2)) # make 2 by 2 plots
plot(myfit.pol.int)
```



Comments on polynomial linear regression results: The fitting result given by polynomial (degree 2) regression is better than the fitting result given in Q2.5. We can see that comparing to the results of multiple linear model in Q2.5, the adjust R-squared is larger in polynomial fitting (0.4529>0.4394) and diagnostics plots have better performance.

After interaction: When new interaction terms included (given at the end of coefficients table), adjusted-R squared improved from 0.4529 to 0.4608, which indicating some improvement. Also most interaction terms are significant.