STAT 504: Linear Regression

Homework 3

Pratima K C

Question 1a (1 point) (1 point) Look at the output of summary(). What conclusion can you draw with respect to the investment strategy of this FoHF when you consider the estimated coefficients, the p-values, the global F-test and the R-squared (the small p-values should indicate the indices that a FoHFs invests in)? What does a large R-squared value indicate?

Answer: Based on the output summary four predictors FIA, CTA, RV, and CA are the important variables because they have smaller p-value at significant level α=0.05. Among these variable, the coefficient of CA, FIA, & CTA are poistive, therefore the increase in these variable will increase the expected average change in FoHF. However, RV has a negative coefficient, that mean an increase in this variable will decrease in the value FoHF. Therefore FoHFs can invests in increasing CA, FIA, and CTA variables and lower the RV. The global F-test showed that larger model is better than the empty model because p-values (2.2e-16) is very small. Lagre R-squared value mean we are able to explain the variance in the response is explained by the model. Also, lager R-squared value shows how good is the model. In this case the R-squared is 0.807 which mean it is good at explaining the variability in the response variable in the model.

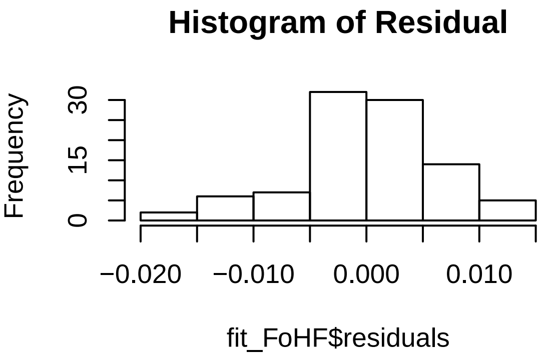
```
##
## Call:
## lm(formula = FoHF ~ RV + CA + FIA + EMN + ED + DS + MA + LSE +
##     GM + EM + CTA + SS, data = FoHF)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0185186 -0.0031189  0.0004069  0.0035469  0.0148925
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.002256   0.001299  -1.736   0.0862 .
## RV          -0.388854   0.171151  -2.272   0.0257 *
## CA           0.238653   0.104522   2.283   0.0250 *
## FIA          0.363010   0.087832   4.133 8.51e-05 ***
## EMN          0.184766   0.197475   0.936   0.3522
## ED           0.314914   0.215792   1.459   0.1482
## DS          -0.007699   0.124324  -0.062   0.9508
## MA          -0.028413   0.169406  -0.168   0.8672
## LSE          0.153636   0.099548   1.543   0.1266
## GM           0.127093   0.086897   1.463   0.1474
## EM           0.049183   0.035065   1.403   0.1645
## CTA          0.159225   0.037304   4.268 5.20e-05 ***
## SS           0.032630   0.023424   1.393   0.1673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006563 on 83 degrees of freedom
## Multiple R-squared:  0.8076, Adjusted R-squared:  0.7798
## F-statistic: 29.03 on 12 and 83 DF,  p-value: < 2.2e-16
```

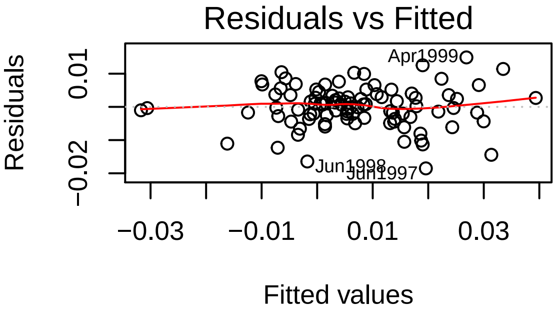
Questions 1b (2 points) Check whether any assumptions are violated (TA and QQ plot). Also check whether there are problems with respect to multicollinearity.

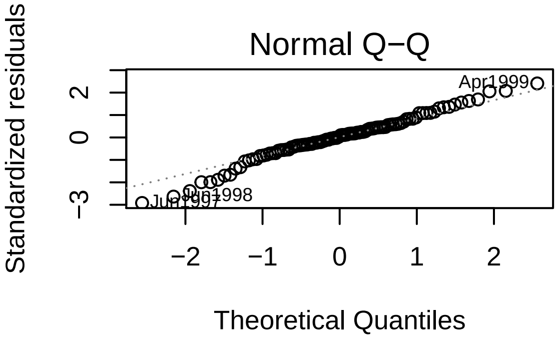
Answer: The normality assumption is not necessarly voilatied based on the histogram of residuals and QQ plot. These plots show the residuals are kind of normally distributed with little tail on left side. Since, there is not much curveture in the mean, the error of mean assumption is not violated either. Similarly, the constant variance assumption is also not violated as we can see in TA plots. Some of the point in the lower fitted value has less variance but overall it seems to have constant variance.

Yes, there are multicollinearity problems, as we can see many predictors variables have vif values larger than 5 such as: EM, ED, DS, MA, LSF, and GM. The vif value above 5 suggest there is collinearity.

```
##        RV        CA       FIA       EMN        ED        DS        MA       LSE
##  6.387024  2.982646  2.271113  3.672017 29.973694  9.404810  8.001994 10.046374
##        GM        EM       CTA        SS
##  5.699120  4.255477  2.232320  4.972861
```







Question 1c (1 point) If you have solved the previous subproblem correctly, you will have found some issues. Formulate a strategy how those can be fixed in order to obtain a valid and interpretable result. Hint: Creating new predictors is not helpful.

Answer: There was a multicollinearity issues with above model. In order to obtain a valid and interpretalble result we can remove some of the predicitors with higher vif values from the model.

When we dropped the predictor variables such as ED, MA, LSE, then rest of the predictors vif values came below 5 in the new model. The RV seems to correlated with MA more than

DS because when we included MA the vif value of RV goes up. Therefore, we droppoed MA from the model. The vif values for each predictors in new model is given below:

```
##       RV       CA      FIA      EMN       DS       GM       EM      CTA
## 3.879437 2.625715 2.071173 2.592558 3.366653 4.203217 4.103678 2.009650
##       SS
## 3.125200
```

Question 1d (3 points) Perform variable selection using the BIC criterion. Implement the following search strategies, identify the best/final model and compare:

(i) Stepwise variable selection, starting with the full model.

(ii) Stepwise variable selection, starting with the empty model.

(iii) All Subsets variable selection.

Answer: The BIC backward, forward and subset selection, all method choose the same model. The AIC vaulue for the best model is -940.78 for both BIC backward and forward setpwise selection method. The plot of all subsets variable selection method showed that the mallow cp and adjusted R^2 choose the larger model as a good model. However, the best final model selected using the BIC backward & forward stepwise and the subset variable selection is smaller and is given as:

$FoHF \sim CA + FIA + ED + GM + CTA$

```
# BIC backward stepwise variable selection:
fit.a <- lm(FoHF ~ ., data = FoHF)
scp <- list(lower = ~ 1, upper = ~ .) # . is shorthand for all
fit.b <- step(fit.a, scope = scp, direction = "backward", k = log(96))
```

```
## Start:  AIC=-919.68
## FoHF ~ RV + CA + FIA + EMN + ED + DS + MA + LSE + GM + EM + CTA +
##     SS
##
##        Df  Sum of Sq       RSS     AIC
## - DS    1 0.00000017 0.0035753 -924.24
## - MA    1 0.00000121 0.0035763 -924.21
## - EMN   1 0.00003771 0.0036128 -923.24
## - SS    1 0.00008358 0.0036587 -922.03
## - EM    1 0.00008474 0.0036598 -922.00
## - ED    1 0.00009173 0.0036668 -921.81
## - GM    1 0.00009214 0.0036672 -921.80
## - LSE   1 0.00010260 0.0036777 -921.53
## <none>              0.0035751 -919.68
## - RV    1 0.00022234 0.0037974 -918.45
## - CA    1 0.00022456 0.0037996 -918.40
## - FIA   1 0.00073576 0.0043108 -906.28
## - CTA   1 0.00078472 0.0043598 -905.20
##
## Step:  AIC=-924.24
## FoHF ~ RV + CA + FIA + EMN + ED + MA + LSE + GM + EM + CTA +
##     SS
##
##        Df  Sum of Sq       RSS     AIC
## - MA    1 0.00000108 0.0035763 -928.78
## - EMN   1 0.00003761 0.0036129 -927.80
## - SS    1 0.00008344 0.0036587 -926.59
## - EM    1 0.00008500 0.0036603 -926.55
## - GM    1 0.00009213 0.0036674 -926.36
## - LSE   1 0.00010811 0.0036834 -925.95
## <none>                0.0035753 -924.24
## - RV    1 0.00022710 0.0038024 -922.89
## - CA    1 0.00022724 0.0038025 -922.89
## - ED    1 0.00023020 0.0038055 -922.82
## - FIA   1 0.00073934 0.0043146 -910.76
## - CTA   1 0.00079410 0.0043693 -909.55
##
## Step:  AIC=-928.78
## FoHF ~ RV + CA + FIA + EMN + ED + LSE + GM + EM + CTA + SS
##
##        Df  Sum of Sq       RSS     AIC
## - EMN   1 0.00003909 0.0036154 -932.30
## - SS    1 0.00008398 0.0036603 -931.11
## - EM    1 0.00008759 0.0036639 -931.02
## - GM    1 0.00010058 0.0036769 -930.68
## - LSE   1 0.00010832 0.0036846 -930.48
## <none>                0.0035763 -928.78
## - CA    1 0.00024101 0.0038173 -927.08
## - RV    1 0.00026057 0.0038369 -926.59
## - ED    1 0.00035144 0.0039278 -924.34
## - CTA   1 0.00079349 0.0043698 -914.11
## - FIA   1 0.00079685 0.0043732 -914.03
##
## Step:  AIC=-932.3
## FoHF ~ RV + CA + FIA + ED + LSE + GM + EM + CTA + SS
##
##        Df  Sum of Sq       RSS     AIC
## - EM    1 0.00007430 0.0036897 -934.91
## - SS    1 0.00010742 0.0037228 -934.05
## - GM    1 0.00012492 0.0037403 -933.60
## <none>                0.0036154 -932.30
## - LSE   1 0.00018537 0.0038008 -932.06
## - RV    1 0.00025580 0.0038712 -930.30
## - ED    1 0.00035729 0.0039727 -927.82
## - CA    1 0.00040461 0.0040200 -926.68
## - FIA   1 0.00075873 0.0043741 -918.57
## - CTA   1 0.00088516 0.0045006 -915.84
##
## Step:  AIC=-934.91
## FoHF ~ RV + CA + FIA + ED + LSE + GM + CTA + SS
##
##        Df  Sum of Sq       RSS     AIC
## - SS    1 0.00005369 0.0037434 -938.09
## - LSE   1 0.00014269 0.0038324 -935.83
## <none>                0.0036897 -934.91
## - GM    1 0.00024280 0.0039325 -933.36
## - RV    1 0.00026091 0.0039506 -932.92
## - CA    1 0.00037752 0.0040672 -930.12
## - ED    1 0.00058092 0.0042706 -925.44
## - CTA   1 0.00081755 0.0045073 -920.26
## - FIA   1 0.00083132 0.0045210 -919.97
##
## Step:  AIC=-938.09
## FoHF ~ RV + CA + FIA + ED + LSE + GM + CTA
##
##         Df  Sum of Sq       RSS     AIC
## - LSE   1 0.00009111 0.0038345 -940.34
## <none>               0.0037434 -938.09
## - RV    1 0.00024437 0.0039878 -936.58
## - GM    1 0.00027617 0.0040196 -935.82
## - CA    1 0.00038539 0.0041288 -933.24
## - ED    1 0.00052919 0.0042726 -929.96
## - CTA   1 0.00083822 0.0045816 -923.25
## - FIA   1 0.00092714 0.0046706 -921.41
##
## Step:  AIC=-940.34
## FoHF ~ RV + CA + FIA + ED + GM + CTA
##
##         Df  Sum of Sq       RSS     AIC
## - RV    1 0.00016854 0.0040031 -940.78
## <none>               0.0038345 -940.34
## - CA    1 0.00031176 0.0041463 -937.40
## - GM    1 0.00072123 0.0045558 -928.36
## - CTA   1 0.00074886 0.0045834 -927.78
## - ED    1 0.00083966 0.0046742 -925.90
## - FIA   1 0.00085130 0.0046858 -925.66
##
## Step:  AIC=-940.78
## FoHF ~ CA + FIA + ED + GM + CTA
##
##         Df  Sum of Sq       RSS     AIC
## <none>               0.0040031 -940.78
## - CA    1 0.00019591 0.0041990 -940.76
## - GM    1 0.00066084 0.0046639 -930.67
## - ED    1 0.00071917 0.0047222 -929.48
## - FIA   1 0.00073423 0.0047373 -929.18
## - CTA   1 0.00089226 0.0048953 -926.03
```

```
summary(fit.b)
```

```
##
## Call:
## lm(formula = FoHF ~ CA + FIA + ED + GM + CTA, data = FoHF)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.017656 -0.003736  0.000617  0.003476  0.016531
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0018567  0.0009089  -2.043 0.043984 *
## CA           0.1756651  0.0837020   2.099 0.038645 *
## FIA          0.2984918  0.0734666   4.063 0.000103 ***
## ED           0.2654424  0.0660132   4.021 0.000120 ***
## GM           0.2469422  0.0640654   3.855 0.000217 ***
## CTA          0.1535033  0.0342726   4.479  2.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006669 on 90 degrees of freedom
## Multiple R-squared:  0.7846, Adjusted R-squared:  0.7726
## F-statistic: 65.55 on 5 and 90 DF,  p-value: < 2.2e-16
```

```
# BIC forward stewise variable selection:
fit.a <- lm(FoHF ~ 1, data = FoHF)
scp <- list(lower = ~ 1, upper = ~ RV + CA + FIA + EMN + ED + DS + MA + LSE + GM + EM + CTA + SS) # . i
fit.b <- step(fit.a, scope = scp, direction = "forward", k = log(96))
```

```
## Start:  AIC=-816.23
## FoHF ~ 1
##
##         Df Sum of Sq       RSS     AIC
## + GM     1 0.0116463 0.0069343 -906.29
## + ED     1 0.0072492 0.0113314 -859.15
## + EMN    1 0.0071500 0.0114306 -858.31
## + DS     1 0.0066117 0.0119689 -853.89
## + EM     1 0.0065705 0.0120101 -853.56
## + FIA    1 0.0059473 0.0126333 -848.70
## + LSE    1 0.0058787 0.0127020 -848.18
## + RV     1 0.0055859 0.0129947 -846.00
## + CA     1 0.0047059 0.0138748 -839.71
## + MA     1 0.0043282 0.0142525 -837.13
## + CTA    1 0.0027357 0.0158449 -826.96
## + SS     1 0.0020455 0.0165351 -822.87
## <none>             0.0185806 -816.23
##
## Step:  AIC=-906.29
## FoHF ~ GM
##
##         Df  Sum of Sq       RSS     AIC
## + CA     1 0.00134397 0.0055904 -922.41
## + FIA    1 0.00128668 0.0056477 -921.43
## + DS     1 0.00081028 0.0061241 -913.66
## + ED     1 0.00072625 0.0062081 -912.35
## + RV     1 0.00049097 0.0064434 -908.78
## + MA     1 0.00046432 0.0064700 -908.38
## + EMN    1 0.00041949 0.0065148 -907.72
## <none>              0.0069343 -906.29
## + EM     1 0.00027079 0.0066636 -905.55
## + CTA    1 0.00003184 0.0069025 -902.17
## + LSE    1 0.00002796 0.0069064 -902.11
## + SS     1 0.00000094 0.0069334 -901.74
##
## Step:  AIC=-922.41
## FoHF ~ GM + CA
##
##         Df  Sum of Sq       RSS     AIC
## + FIA   1 0.00049445 0.0050959 -926.73
## + CTA   1 0.00037299 0.0052174 -924.47
## <none>              0.0055904 -922.41
## + DS    1 0.00013763 0.0054527 -920.24
## + ED    1 0.00010212 0.0054882 -919.61
## + EM    1 0.00004467 0.0055457 -918.61
## + MA    1 0.00003925 0.0055511 -918.52
## + SS    1 0.00003299 0.0055574 -918.41
## + EMN   1 0.00001913 0.0055712 -918.17
## + RV    1 0.00000254 0.0055878 -917.89
## + LSE   1 0.00000236 0.0055880 -917.88
##
## Step:  AIC=-926.73
## FoHF ~ GM + CA + FIA
##
##        Df  Sum of Sq       RSS     AIC
## + CTA   1 0.00037369 0.0047222 -929.48
## <none>              0.0050959 -926.73
## + ED    1 0.00020059 0.0048953 -926.03
## + MA    1 0.00015050 0.0049454 -925.05
## + DS    1 0.00014523 0.0049507 -924.95
## + EMN   1 0.00014113 0.0049548 -924.87
## + EM    1 0.00007042 0.0050255 -923.51
## + LSE   1 0.00003826 0.0050577 -922.89
## + SS    1 0.00000555 0.0050904 -922.27
## + RV    1 0.00000552 0.0050904 -922.27
##
## Step:  AIC=-929.48
## FoHF ~ GM + CA + FIA + CTA
##
##        Df  Sum of Sq       RSS     AIC
## + ED    1 0.00071917 0.0040031 -940.78
## + DS    1 0.00049338 0.0042289 -935.51
## + LSE   1 0.00039821 0.0043240 -933.37
## + EM    1 0.00037089 0.0043513 -932.77
## + MA    1 0.00035241 0.0043698 -932.36
## <none>              0.0047222 -929.48
## + SS    1 0.00015897 0.0045633 -928.20
## + EMN   1 0.00014474 0.0045775 -927.91
## + RV    1 0.00004805 0.0046742 -925.90
##
## Step:  AIC=-940.78
## FoHF ~ GM + CA + FIA + CTA + ED
##
##        Df  Sum of Sq       RSS     AIC
## <none>              0.0040031 -940.78
## + RV    1 1.6854e-04 0.0038345 -940.34
## + EMN   1 4.9124e-05 0.0039539 -937.40
## + EM    1 2.7889e-05 0.0039752 -936.88
## + LSE   1 1.5281e-05 0.0039878 -936.58
## + SS    1 1.0194e-05 0.0039929 -936.46
## + DS    1 8.0550e-06 0.0039950 -936.41
## + MA    1 1.5360e-06 0.0040015 -936.25
```

```r
summary(fit.b)
```

```
##
## Call:
## lm(formula = FoHF ~ GM + CA + FIA + CTA + ED, data = FoHF)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.017656 -0.003736  0.000617  0.003476  0.016531
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0018567  0.0009089  -2.043 0.043984 *
## GM           0.2469422  0.0640654   3.855 0.000217 ***
## CA           0.1756651  0.0837020   2.099 0.038645 *
## FIA          0.2984918  0.0734666   4.063 0.000103 ***
## CTA          0.1535033  0.0342726   4.479  2.2e-05 ***
## ED           0.2654424  0.0660132   4.021 0.000120 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006669 on 90 degrees of freedom
## Multiple R-squared:  0.7846, Adjusted R-squared:  0.7726
## F-statistic: 65.55 on 5 and 90 DF,  p-value: < 2.2e-16
```

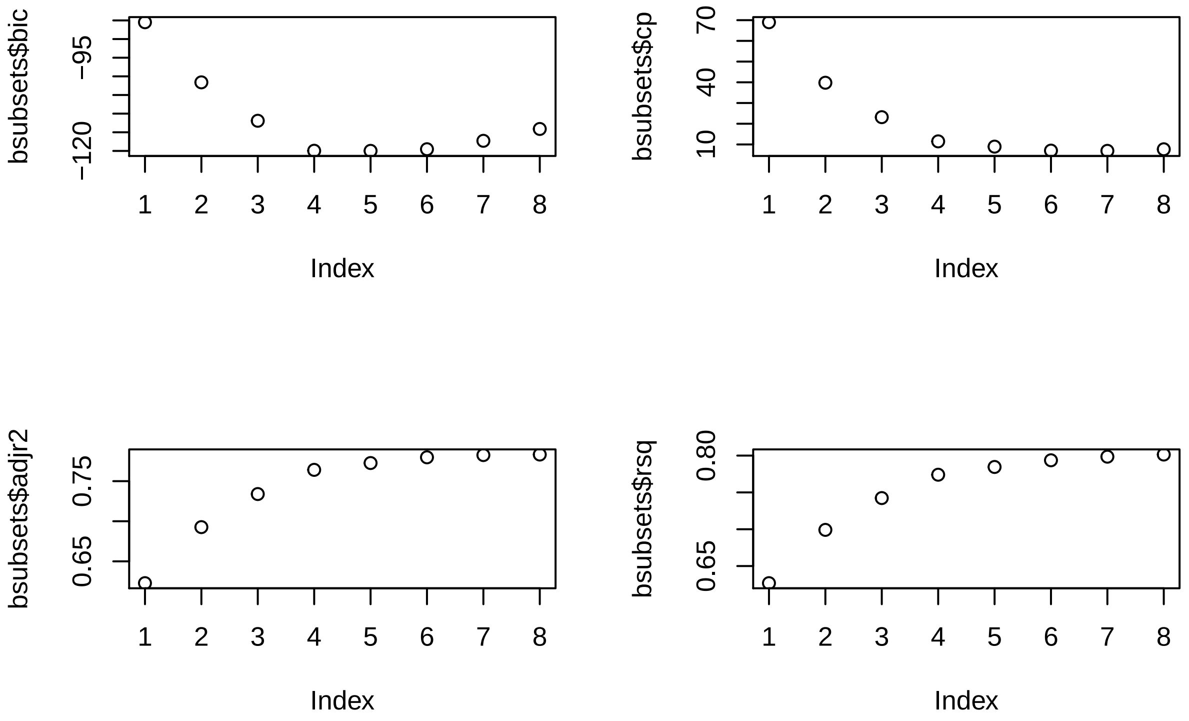
```r
# Best subset selection
library(leaps)
subsets <- regsubsets(FoHF ~ RV + CA + FIA + EMN + ED + DS + MA + LSE + GM + EM + CTA + SS,
                      data = FoHF)
summary(subsets)
```

```
## Subset selection object
## Call: regsubsets.formula(FoHF ~ RV + CA + FIA + EMN + ED + DS + MA +
##     LSE + GM + EM + CTA + SS, data = FoHF)
## 12 Variables  (and intercept)
##     Forced in Forced out
## RV      FALSE      FALSE
## CA      FALSE      FALSE
## FIA     FALSE      FALSE
## EMN     FALSE      FALSE
## ED      FALSE      FALSE
## DS      FALSE      FALSE
## MA      FALSE      FALSE
## LSE     FALSE      FALSE
## GM      FALSE      FALSE
## EM      FALSE      FALSE
## CTA     FALSE      FALSE
## SS      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          RV  CA  FIA EMN ED  DS  MA  LSE GM  EM  CTA SS
## 1  ( 1 ) " " " " " " " " " " " " " " " " "*" " " " " " "
## 2  ( 1 ) " " "*" " " " " " " " " " " " " "*" " " " " " "
## 3  ( 1 ) " " " " "*" " " "*" " " " " " " " " " " "*" " "
## 4  ( 1 ) " " " " "*" " " "*" " " " " " " "*" " " "*" " "
## 5  ( 1 ) " " "*" "*" " " "*" " " " " " " "*" " " "*" " "
## 6  ( 1 ) "*" "*" "*" " " "*" " " " " " " "*" " " "*" " "
## 7  ( 1 ) "*" "*" "*" " " "*" " " " " "*" "*" " " "*" " "
## 8  ( 1 ) "*" "*" "*" " " "*" " " " " "*" "*" " " "*" "*"
```

```
bsubsets <- summary(subsets)
bsubsets$bic
```

```
## [1]  -85.49217 -101.61005 -111.91826 -119.95704 -119.97951 -119.54463 -117.28892
## [8] -114.11136
```

```
par(mfrow = c(2,2))
plot(bsubsets$bic) # BIC, AIC not implemented :(
plot(bsubsets$cp) # Mallow's Cp
plot(bsubsets$adjr2) # adjusted R squared
plot(bsubsets$rsq)
```



Question 2a (1 point) Some predictors in this data are probably colinear or multicolinear. Based on the description of the data, which predictors are those? Print a correlation matrix to and comment on the output.

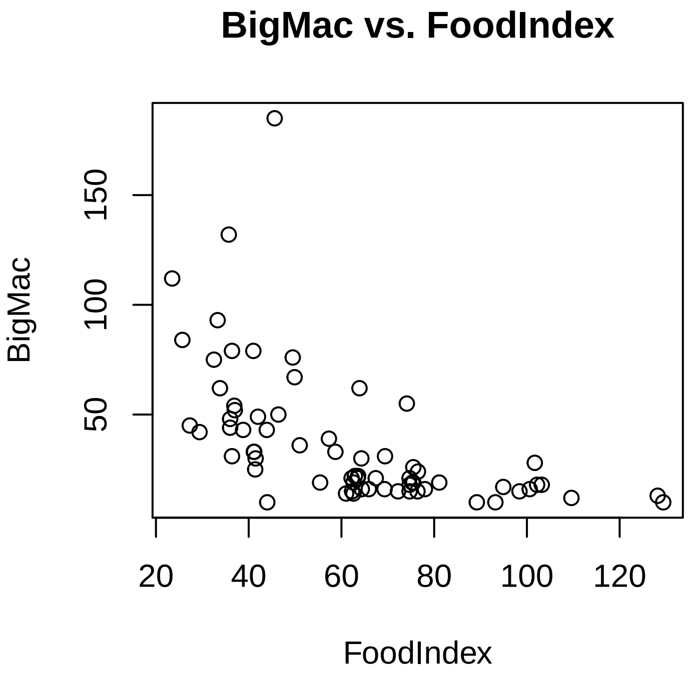
Answer: Based on the description of the data, the predictors the TeachGI and TeachNI is correlated with each other, apt could be correlated with TeachNI, TeachGI. Similarly, I expected there should be correlation between teachHours and teachNI & teachGI but the corrleation

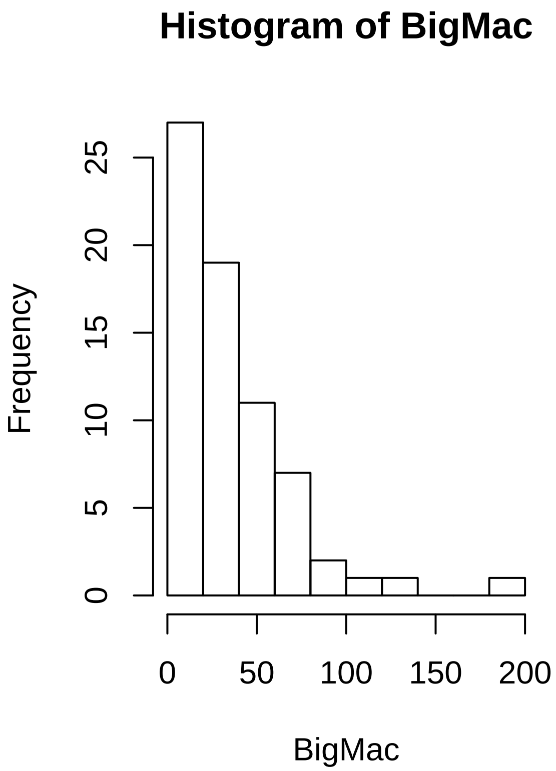
matrix shows there is very low correlation between these variable. Based on the correlation matrix, the foodIndex seems to be highly correlated with Apt, TeachNI, & TeachGI. There is also a good correlation between bus and other variables such as TeachGI, TeachNI. The larger correlation is between the TeachGI and TeachNI. The correlation matrix is given below:

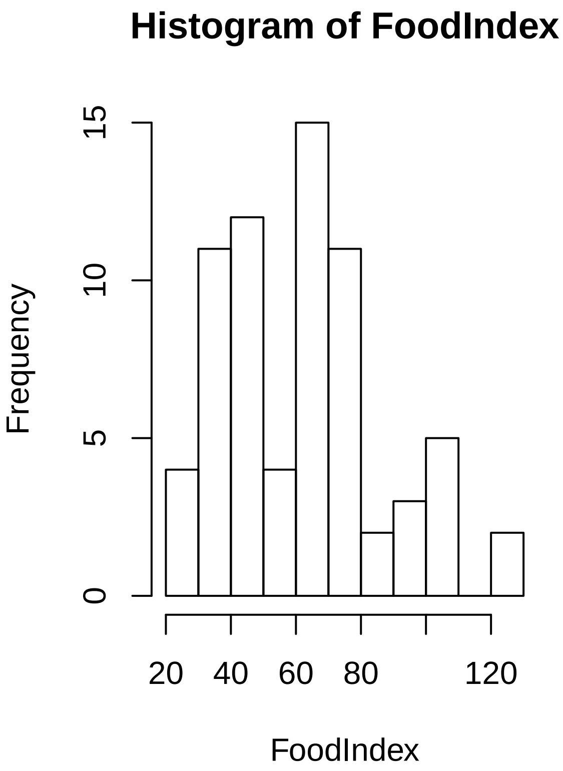
```
##                BigMac       Bread        Rice  FoodIndex        Bus        Apt
## BigMac      1.0000000  0.54458571  0.69614035 -0.5745753 -0.5430733 -0.5412829
## Bread       0.5445857  1.00000000  0.44604465 -0.2746116 -0.5036435 -0.4503529
## Rice        0.6961404  0.44604465  1.00000000 -0.3376461 -0.4496312 -0.4409093
## FoodIndex  -0.5745753 -0.27461157 -0.33764607  1.0000000  0.5457161  0.7167216
## Bus        -0.5430733 -0.50364354 -0.44963117  0.5457161  1.0000000  0.4693233
## Apt        -0.5412829 -0.45035290 -0.44090929  0.7167216  0.4693233  1.0000000
## TeachGI    -0.6178904 -0.54741960 -0.48776153  0.7735416  0.6970514  0.6606221
## TeachNI    -0.6106635 -0.52810694 -0.47931427  0.7903348  0.6503863  0.6674487
## TaxRate    -0.4535382 -0.43990916 -0.28810930  0.2910690  0.5766945  0.2925890
## TeachHours  0.0135003 -0.07643946  0.05920052  0.2394993  0.0173272  0.1216599
##                TeachGI    TeachNI    TaxRate  TeachHours
## BigMac     -0.61789041 -0.6106635 -0.4535382  0.01350030
## Bread      -0.54741960 -0.5281069 -0.4399092 -0.07643946
## Rice       -0.48776153 -0.4793143 -0.2881093  0.05920052
## FoodIndex   0.77354156  0.7903348  0.2910690  0.23949928
## Bus         0.69705144  0.6503863  0.5766945  0.01732720
## Apt         0.66062209  0.6674487  0.2925890  0.12165990
## TeachGI     1.00000000  0.9899063  0.4533588  0.09632324
## TeachNI     0.98990634  1.0000000  0.3561561  0.12698862
## TaxRate     0.45335882  0.3561561  1.0000000 -0.16987046
## TeachHours  0.09632324  0.1269886 -0.1698705  1.00000000
```

Question 2b (1 point) Draw the scatterplot with BigMac on the vertical axis and FoodIndex on the horizontal axis. Provide a qualitative description of this graph.

Answer: The scatterplot between BigMac vs. FoodIndex shows that they are negatively correlated. The lower FoodIndex has higher BigMac values. As the FoodIndex increase the BigMac values decreases. The plot doesnot seems to have linear relationship. The histogram plot showed that the BigMac is right skewed. This suggest that we should use transformation of the variable. The scatter plot between BigMac vs. FoodIndex and their histogram is given below:

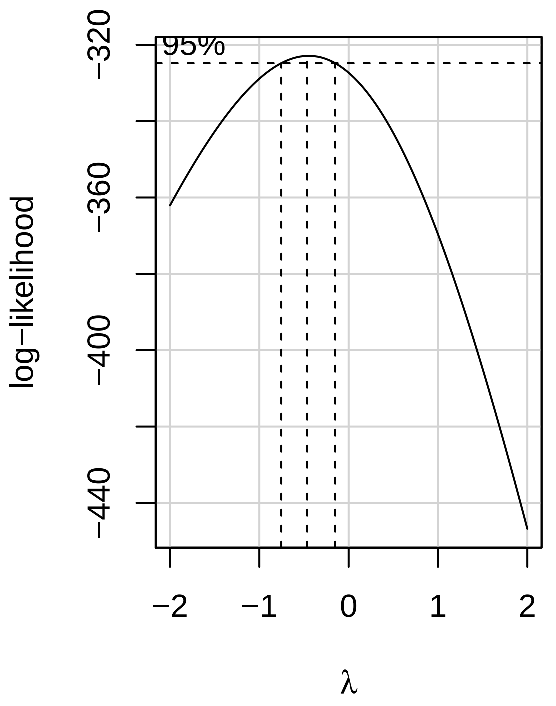


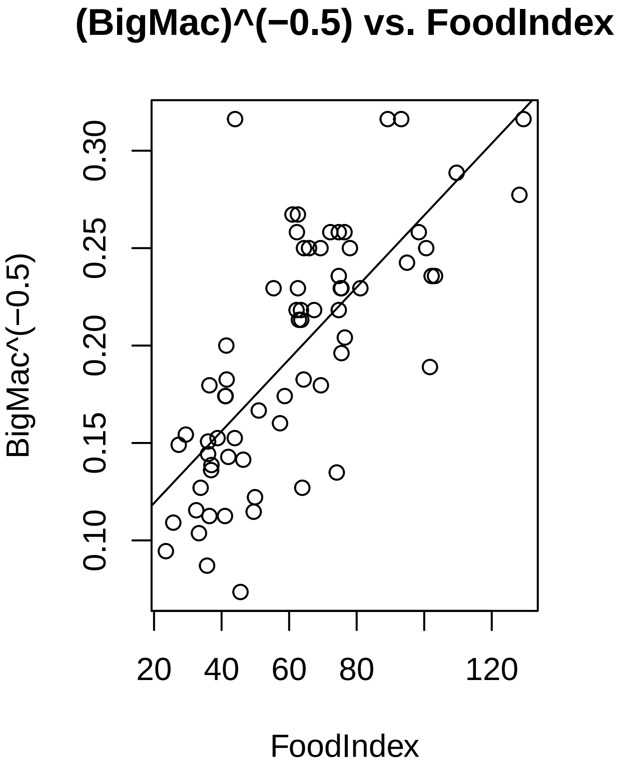




Question 2c (1 point) Use the Box-Cox method to find a transformation of BigMac so that the resulting *scatterplot has a linear mean function.*

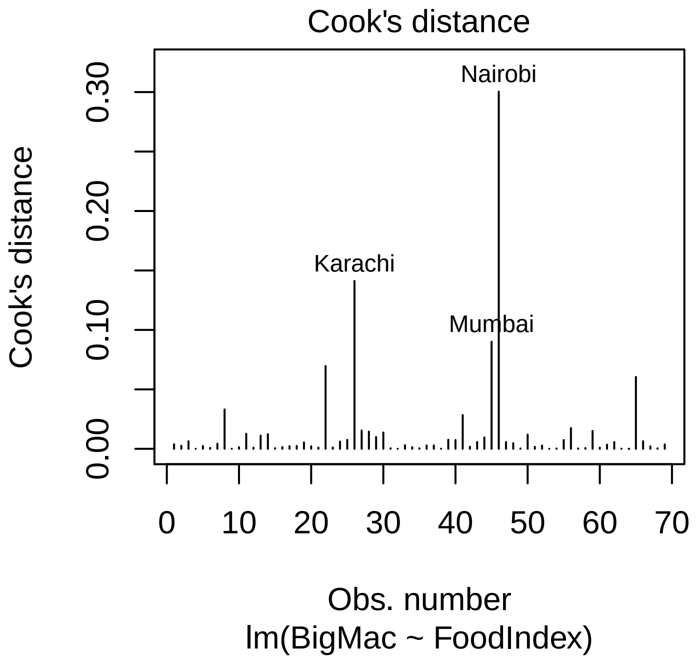
Answer: Using the Box-Cox function we found that $\lambda =$ ***nearly -0.5. Since, the*** $\lambda = -0.5$ ***the transformation used will be*** $BigMac^{-0.5}$***. The Box-cox plot and transformed scatter plot are shown below:***





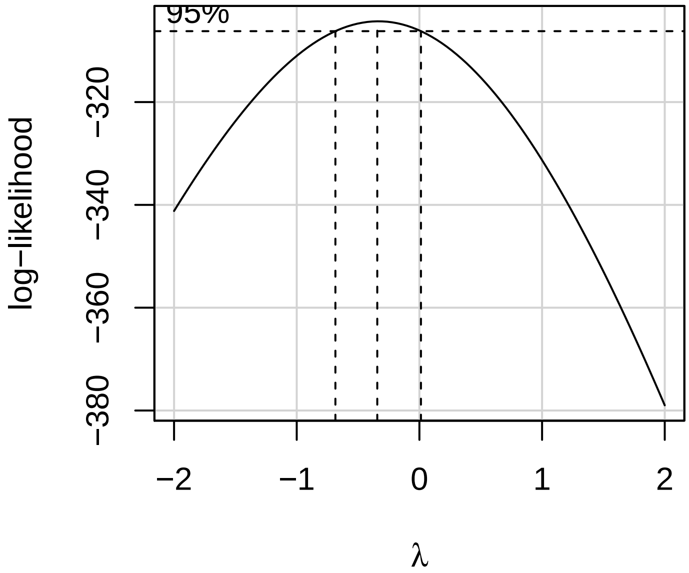
Question 2d (1 point) Two of the cities, with very large values for BigMac, are very infuential for selecting a transformation. What cities are those?

Answer: The two of the cities, with very large values of BigMac are: Nairobi and Karachi based on the cook's distance. The cook's plot is given below:



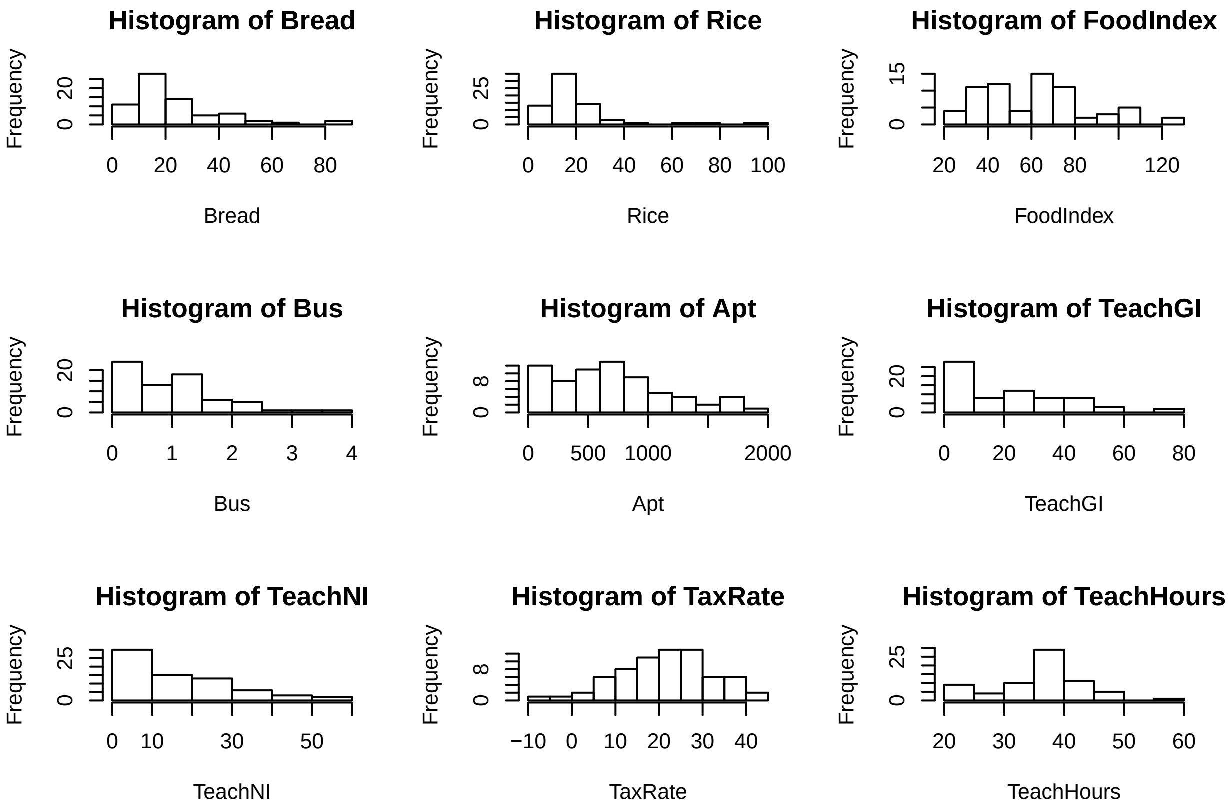
Question 2e (1 point) Remove the two cities you identified in the previous task and apply the Box-Cox method to the reduced data set. What has changed?

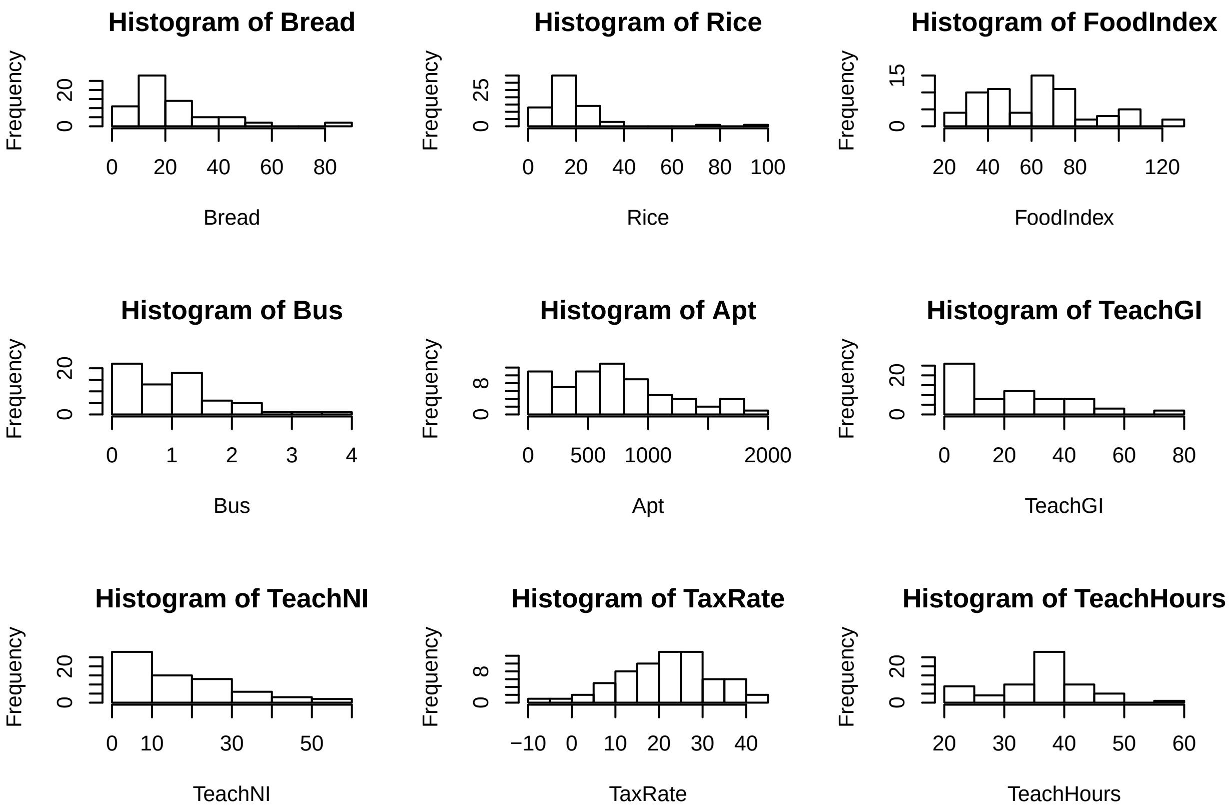
Answer: When two cities with lagerst BigMac values is removed and Box-Cox method was applied to the reduced data set than the λ value changed and shifted towards 0. Also, the center line of the λ value changed from -0.5 to -0.3. The Box-cox plot of reduced data set is given below:



Question 2f (2 points) Draw the histogram of each predictor (every variable except BigMac) in the data set. Do some of them appear right skewed? Which ones?

Answer: Based on the original and the removed dataset, yes, some of the predictor appears to be right skewed such as, Bread, Bus, Rice, TeachGI, and TeachNI. Histograms of all predictors from original and changed dataset is shown below:


Histogram of Bread
Frequency
0
20
0
20
40
60
80
Bread
Histogram of Rice
Frequency
0
25
0
20
40
60
80
100
Rice
Histogram of FoodIndex
Frequency
0
15
20
40
60
80
120
FoodIndex
Histogram of Bus
Frequency
0
20
0
1
2
3
4
Bus
Histogram of Apt
Frequency
0
8
0
500
1000
2000
Apt
Histogram of TeachGI
Frequency
0
20
0
20
40
60
80
TeachGI
Histogram of TeachNI
Frequency
0
25
0
10
30
50
TeachNI
Histogram of TaxRate
Frequency
0
8
−10
0
10
20
30
40
TaxRate
Histogram of TeachHours
Frequency
0
25
20
30
40
50
60
TeachHours



Question 2g (1 point) Use the data where you left out the two infuential points and fit the model with BigMac as the response (transformed using the Box-Cox suggested transformation) and the following predictors:log(Bread), log(Rice), log(Bus), Apt, log(TeachNI) Model 1: log(Bread), log(Rice); Model 2: log(Bread), log(Rice), Apt, log(Bus), Model 3: with all above predictors

Which of these model acheaves the best leave-one-out cross-validation score?

Answer: Here we use the log transformation on the response variable based on the Box-cox plot from the question 2e. The third model achieves the best leave-one-out cross validation score. The LOOCV score is small for the third and largest model (LOOCV score= 0.1008) therefore we opt to use this model. The selected model is given below:

log(BigMac) ~ log(Bread) + log(Rice) + log(Bus) + Apt + log(TeachNI)

```
loocv.lm<-function(mdl){
  return(mean((residuals(mdl)/(1-hatvalues(mdl)))^2))
}

# The LOOCV score of Model 1
loocv.lm(fit1)
```

```
## [1] 0.198431
```

```
# The LOOCV score of  Model 2
loocv.lm(fit2)
```

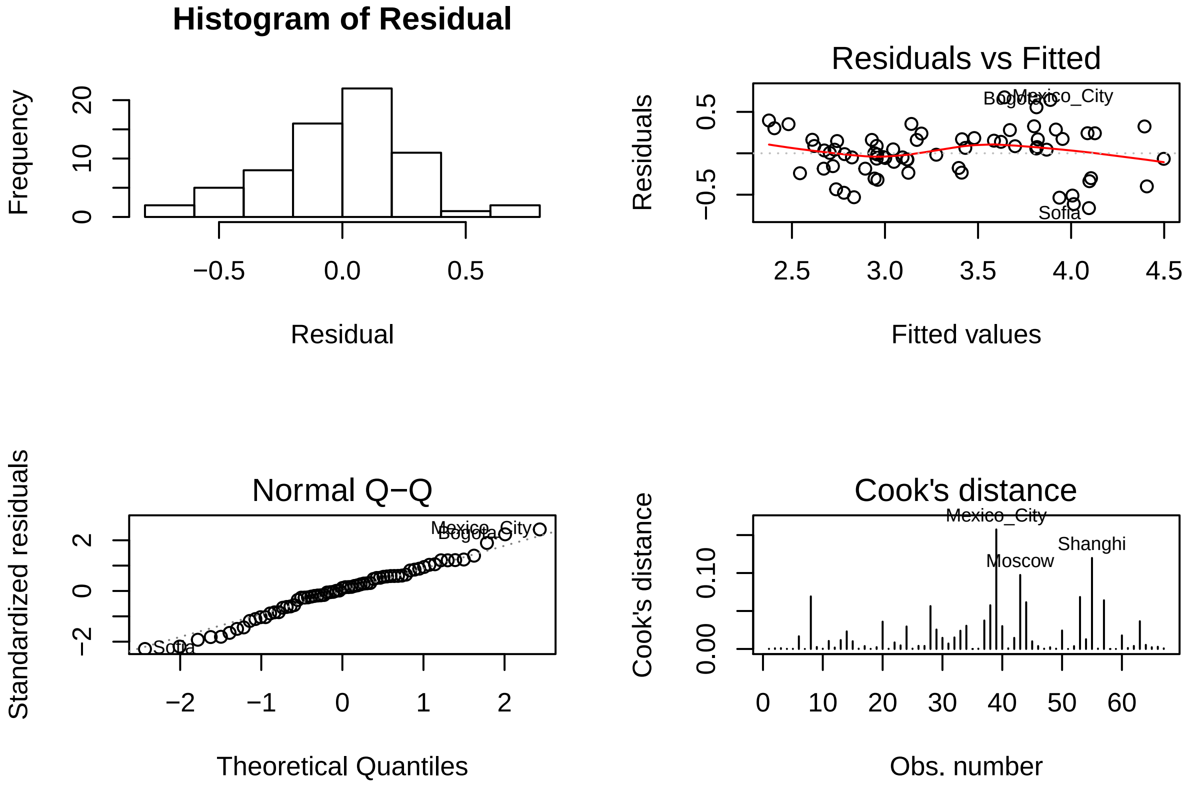
```
## [1] 0.1353951
```

```r
# The LOOCV score of Model 3
loocv.lm(fit3)
```

```
## [1] 0.1008867
```

Question 2h (2 points) For the model selected in the previous task, check the model diagnostic plots (TA, QQ, Cook's distance etc.). Do you notice any model assumption violations or any unusual points?

Answer: Based on the following plots non of the model assumption seems to be violated. The histogram of residual shows that distribution of residuals is symmetric so it holds the normality assumption, also shown by normal QQ plot. The TA plot shows that there is constant variance so it holds the constant variance assumption. Since there is no curveture in the mean in TA plot, it seems to hold the zero mean assumption of the error E[Y/X=x]=0. Based on the Cook's distance Mexico city has the higher Cook's distance that is also less than 0.5 (Cook's threshold distance) so there is no significant outliers.



Question 3a (1 point) Using backward elimination with p-values for $\alpha_{crit} = 0.05$ and the principle of hierarchy, which variable should be removed from the model next?

Answer a) X_1, because X_1 does not have other interaction effect in the model and have second largest p-value after X_4. (X_4 has interactive effect so cannot be removed.)

Question 3b (1 point) Following Preetam's suggestion, Marco decides to use the AIC criterion and the step() function in R for his variable selection. Below you are given a partial R output of the step() function where the current model is $Y \sim X_2 + X_3 + X_6$.

Answer: d) No variable will be added or removed in the next step, since the smallest AIC value is at none parameter.

Question 3c (1 point) Consider the R output in sub task b). Let modelAIC be the AIC of the model $Y \sim X_2 + X_6$ and let modelBIC be the BIC of that same model. Which of the following is true:

Answer: a) $model_{AIC} < model_{BIC}$***, because BIC uses log(n), assuming sample size is greater than 9 in above model, log(9)>2.***

Queston 3d (1 point) The plot below shows AIC scores of all different models obtained by applying the stepwise model search in both directions to the empty model.

Answer: c) The worst model in the plot does not contain variable X_7***.***

Question 3e (1 point) Which of the following is true:

Answer: d) The R^2 ***value of the model*** $Y \sim X_1 + X_2 + X_3$ ***is smaller than the R2 value of*** $Y \sim X_1 + X_2 + X_3 + X_4$***.***

Question 4a (1 point) Which of the following statements is false in the context of multiple linear regressions?

Answer: a) If the global F-test is significant, then we can conclude that $\beta_j \neq$ ***0 for all predictors*** x_j***.***

Question 4b (1 point) Which of the following statements is true?

Answer: b) If a 95%-confidence interval for a regression coefficient β_j ***contains the value 0, then the p-value for the test*** $H_0 : \beta_j =$ ***0 must be greater than 0.05***

Question 4c (1 point) Given the following plot, what is the most obvious model violation?

Answer: b) Non-constant variance of the errors.

Question 4d (1 point) Which of the following statements is true for a multiple linear regression model?

Answer: The normality assumption of the errors can be checked by verifying whether the (standardized) residuals and the corresponding quantiles of a standard normal distribution have a linear relationship.

Question 4e (1 point) Which of the following statements is false for mutiple linear regressions?

Answer: b) Assuming that the sample size increases and predictors stay the same the individual p-values in the R summary will grow with the sample size.