Analysis 3 – Baldridge Data
Yagna Venkitasamy

Part 1 First run a multiple regression of y1pccust on inf1pcma with sector and period dummies (you will need to create a period variable that divides 1990-2006 into 3 periods: before 1995, 1995-1998, and 1999-2006). In this model also include other covariates such as ld1pcla+ st1pcdev + cs1pccmk + hr3pcsat + pm1pcvc; and include an interaction effect between inf1pcma and education sector, and also an interaction effect between inf1pcma and period 1999-2006.

Now assess this model for the extent to which it satisfies the following six regression assumptions discussed in class.

```
rm(list=ls())
library(rio)
library(moments)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(stargazer)

##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary
Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

library(ggplot2)
library(broom)
library(car)

## Loading required package: carData

##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
setwd("C:/Users/yagna/Documents/R/R workings")
df = import("Balridge_data_prep_25Jan2020.csv")
str(df)
```

```
## 'data.frame':    1099 obs. of  13 variables:
##  $ slnoskm17mar11: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ permanentid   : chr  "B-179" "" "" "" ...
##  $ y1pccust      : num  68.1 28.1 74.6 18.9 62.7 ...
##  $ inf1pcma      : num  71.7 23.3 73.3 15 55 ...
##  $ sector        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ year          : int  1990 1990 1990 1990 1990 1990 1990 1990 1990 1990
...
##  $ ld1pcla       : num  76.2 10 73.8 11.2 86.2 ...
##  $ st1pcdev      : num  80 22.9 80 17.1 45.7 ...
##  $ cs1pccmk      : num  80 40 80 26 70 66 40 66 60 66 ...
##  $ hr3pcsat      : num  80 0 60 0 70 0 0 50 0 0 ...
##  $ pm1pcvc       : num  66.3 22.1 72.6 24.2 66.3 ...
##  $ iirtotal      : int  686 147 704 124 612 284 140 609 253 269 ...
##  $ icat7total    : int  204 46 243 29 184 86 36 196 73 84 ...
```

```
getwd()
```

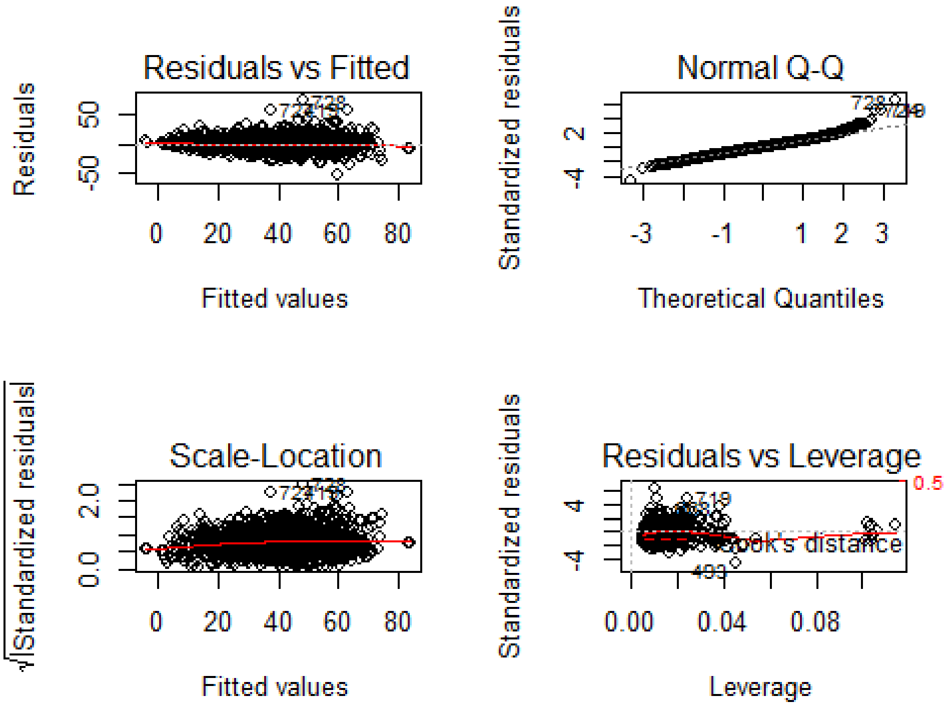
```
## [1] "C:/Users/yagna/Documents/R/R workings"
```

```
df$time_period <- ifelse(df$year < 1995, 0 ,ifelse(df$year < 1999, 1, 2))
m1 <- lm(y1pccust ~ inf1pcma +as.factor(sector) + as.factor(time_period) +
              ld1pcla + st1pcdev + cs1pccmk + hr3pcsat + pm1pcvc +
              I(inf1pcma*(sector==4)) + I(inf1pcma*(time_period==2)),data =
df)
stargazer(m1,type = "text")
```

```
##
## ============================================================
##                                     Dependent variable:
##                                ---------------------------
##                                          y1pccust
## ------------------------------------------------------------
## inf1pcma                                 0.304***
##                                           (0.052)
##
## as.factor(sector)2                         1.707
##                                           (1.181)
##
## as.factor(sector)3                         1.318
##                                           (1.073)
##
## as.factor(sector)4                       9.164***
##                                            (3.280)
##
## as.factor(sector)5                        4.857***
##                                            (1.318)
##
## as.factor(sector)6                          2.363
##                                            (3.862)
##
## as.factor(time_period)1                   -4.647***
##                                            (1.080)
##
## as.factor(time_period)2                     2.429
##                                            (2.504)
##
## ld1pcla                                     0.067
##                                            (0.045)
##
## st1pcdev                                  0.156***
##                                            (0.038)
##
## cs1pccmk                                  0.289***
##                                            (0.040)
##
## hr3pcsat                                   -0.055
##                                            (0.035)
##
## pm1pcvc                                   0.275***
##                                            (0.041)
##
## I(inf1pcma * (sector == 4))                0.157**
##                                            (0.071)
##
## I(inf1pcma * (time_period == 2))          -0.159***
##                                            (0.051)
##
## Constant                                  -8.806***
##                                            (1.751)
##
## ------------------------------------------------------------
## Observations                                1,098
## R2                                          0.634
## Adjusted R2                                 0.629
## Residual Std. Error                  11.634 (df = 1082)
## F Statistic                      125.052*** (df = 15; 1082)
## ============================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

(1) MLR 1: Population or true model is linear in parameters

```
par(mfrow=c(2,2))
plot(m1)
```

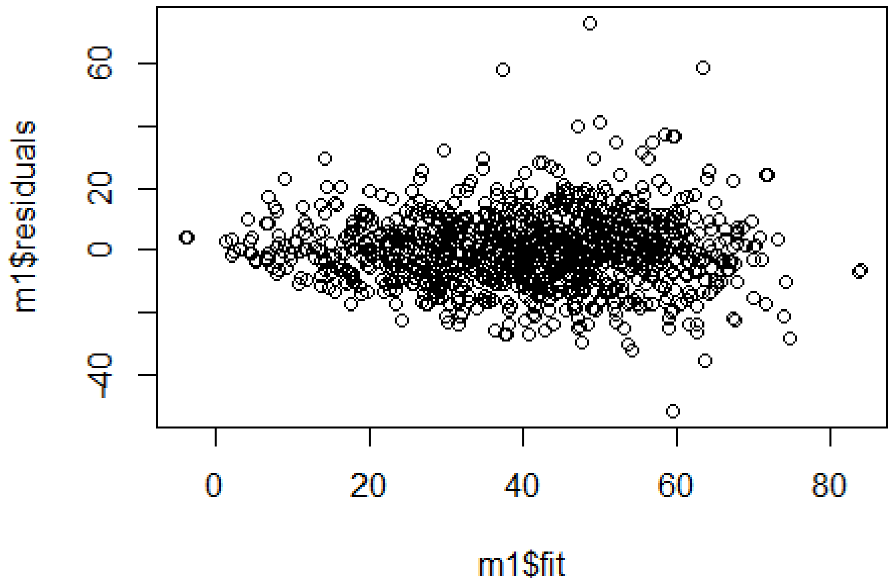


Interpretation:

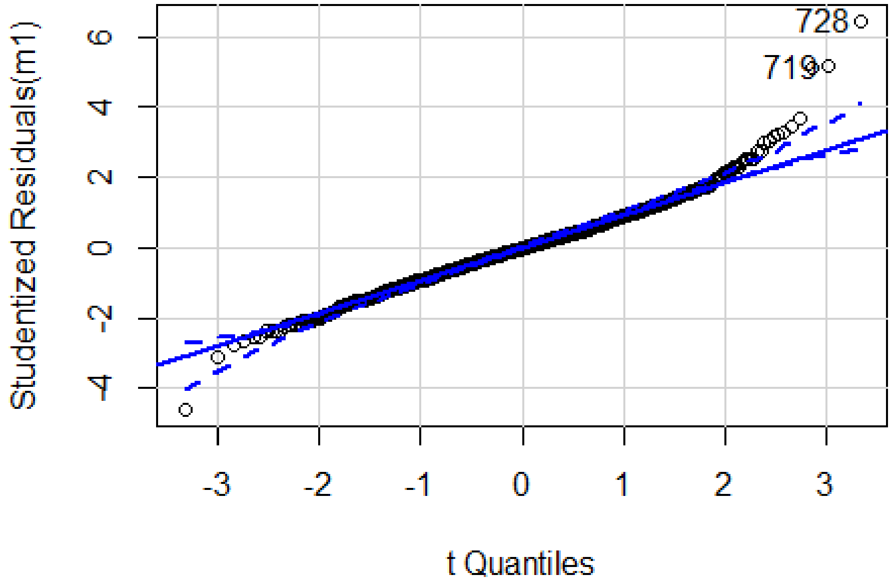
The residuals and the fitted plots shows that the data is fairly linear and centered around zero.

(2) MLR 2: Random sampling

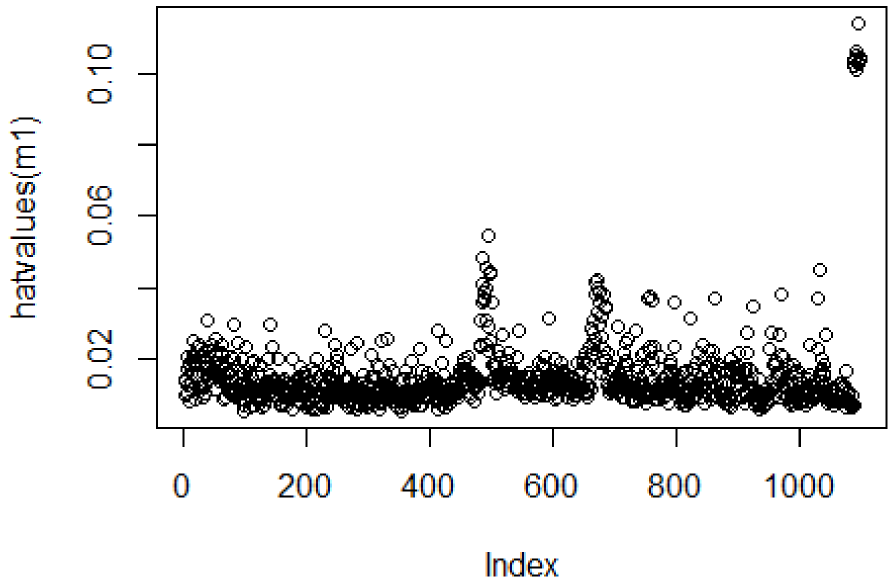
```
par(mfrow=c(1,1))
plot (m1$residuals ~ m1$fit)
```



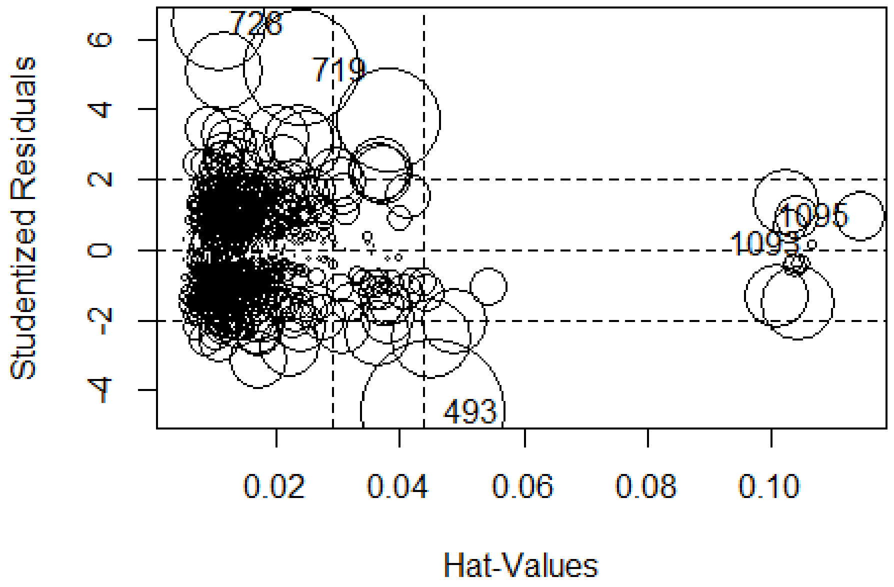
```
library(Rcpp)
qqPlot(m1, simulate=T, labels=row.names(df))
```



```
## [1] 719 728

#it gives us a 95% confidence interval
outlierTest(m1)   #does a formal test, applies Bonferroni, identifies obs

##      rstudent unadjusted p-value Bonferroni p
## 728  6.440325         1.7903e-10   1.9658e-07
## 719  5.149960         3.0944e-07   3.3977e-04
## 724  5.095915         4.0940e-07   4.4952e-04
## 493 -4.599536         4.7347e-06   5.1987e-03

plot(hatvalues(m1))  #avg hatvalue h=(k+1)/n ; more than 2h or 3h problem
```



```
influencePlot(m1,id.method="identify")  #recheck if circles are prop to
cook's d

## Warning in plot.window(...): "id.method" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "id.method" is not a graphical
parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "id.method"
is not
## a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "id.method"
is not
## a graphical parameter
## Warning in box(...): "id.method" is not a graphical parameter

## Warning in title(...): "id.method" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.method" is not
a
## graphical parameter
```



```
##          StudRes        Hat       CookD
## 493  -4.5995358 0.04551866 0.061903385
## 719   5.1499596 0.02410426 0.039999344
## 728   6.4403253 0.01071487 0.027065159
## 1093  0.1874751 0.10646032 0.000261956
## 1095  0.9722442 0.11436809 0.007629645
```

Interpretation:

The studentized residuals graph shows us the regression outliers. Hat values tell us about potential influence of an observation. In the influence plot, the size of circle is proportional to Cook's distance in assumption #1(Residual vs Leverage graph), which tells us about actual influence of an observation for the model. When cases are outside of the Cook's distance / have high Cook's distance scores, the cases are influential to the regression results. In this model, there is no influential case. All cases are well inside of the Cook's distance lines which means that the model passes the test of random sampling.

(3) MLR 3: Zero conditional mean or exogeneous variables assumption (This assumption violated if we omit squared terms, use log x instead of x or vice versa, we omit some

variable, measurement error in some vars; Note: If xj is correlated with u for any reason then that xj is said to be endogenous.)

```
#3 conditional indep assumption--no endogeneity
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

resettest(m1)

##
##  RESET test
##
## data:  m1
## RESET = 3.5371, df1 = 2, df2 = 1080, p-value = 0.02944
```

Interpretation:

At alpha=0.05, p-value of 0.029 is the evidence of functional form misspecification. The assumption is violated. We might be omitting squared terms or some variable.

(4) MLR 4: No perfect Multicollinearity among Xs

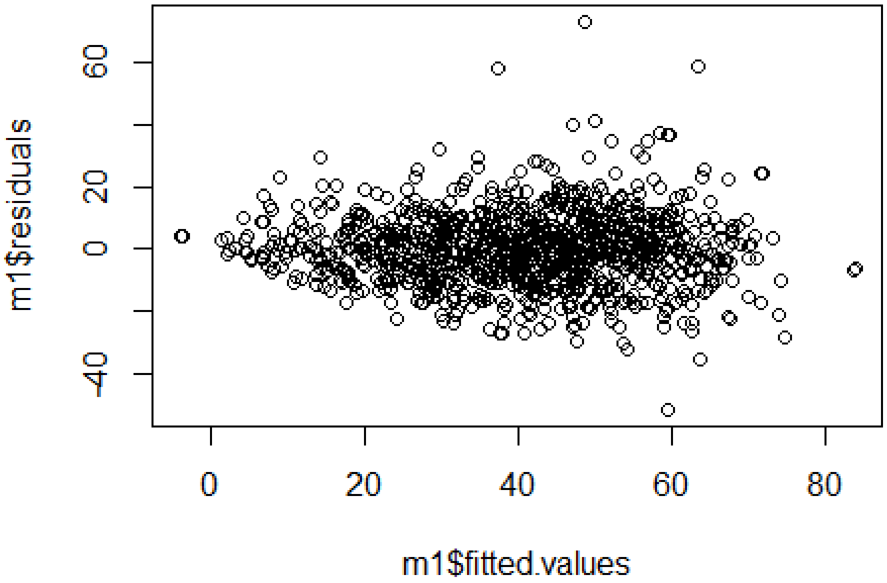
```
library(car)
vif(m1)   #vif=1/(1-rjsquared) where j=1...p; usually vif>10 problematic

##                                       GVIF Df GVIF^(1/(2*Df))
## inf1pcma                          6.161344  1        2.482206
## as.factor(sector)                16.586251  5        1.324265
## as.factor(time_period)           13.624688  2        1.921240
## ld1pcla                           4.977710  1        2.231078
## st1pcdev                          3.425576  1        1.850831
## cs1pccmk                          3.308256  1        1.818861
## hr3pcsat                          3.187092  1        1.785243
## pm1pcvc                           3.697927  1        1.923000
## I(inf1pcma * (sector == 4))       8.075124  1        2.841676
## I(inf1pcma * (time_period == 2)) 13.149551  1        3.626231
```

Interpretation:

After taking into account of the degree of freedoms for each variable, all VIF’s are low, indicating all predictor variables are not highly collinear with each other as the VIF values are very less than 5.

(5) MLR 5: Adding 5th assumption about homoskedasticity to make OLS estimator BLUE.

```
#Bartlett test is more sensitive to violations of normality than Levene test
plot(m1$residuals~m1$fitted.values)
```



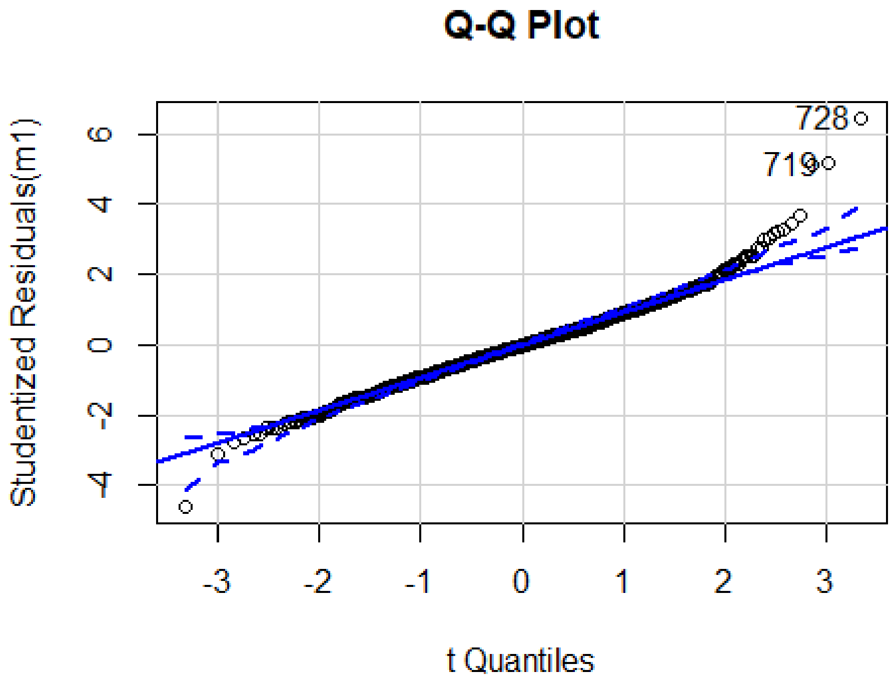
```
bartlett.test(list(m1$residuals,m1$fitted.values))    #here list coerces data objects into a dataframe which serves as input for Bartlett test

##
##  Bartlett test of homogeneity of variances
##
## data:  list(m1$residuals, m1$fitted.values)
## Bartlett's K-squared = 81.962, df = 1, p-value < 2.2e-16

#nonconstant variance score test--aka Breusch Pagan test
ncvTest(m1)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 42.17586, Df = 1, p = 8.3423e-11
```

Interpretation:

Looking at the Residuals vs fitted values graph, we can see a funnel/cone pattern indicating that there is heteroskedaticity in our model. After running the Bartlett test, since p-value is significant, we reject H0, concluding that the variances are not equal.

(6) MLR 6: Normality assumption for u (encompasses MLR 3 and MLR 5 also): Formally, U~ N (0, sigmasq).

```
qqPlot(m1, labels = FALSE,
       simulate = TRUE, main = "Q-Q Plot")
```



```
## [1] 719 728
```

```
shapiro.test(m1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m1$residuals
## W = 0.97401, p-value = 3.914e-13
```

Interpretation: The qq plot and the shapiro test show that the residual values are almost normally distributed. The points at the lower end and upper end are not lined up with qqline and there is a deviation.The mean of residuals is close to zero. The Shapiro test gives a p-value lesser than 0-05 and hence we reject the null hypothesis and the data is not normally distributed.

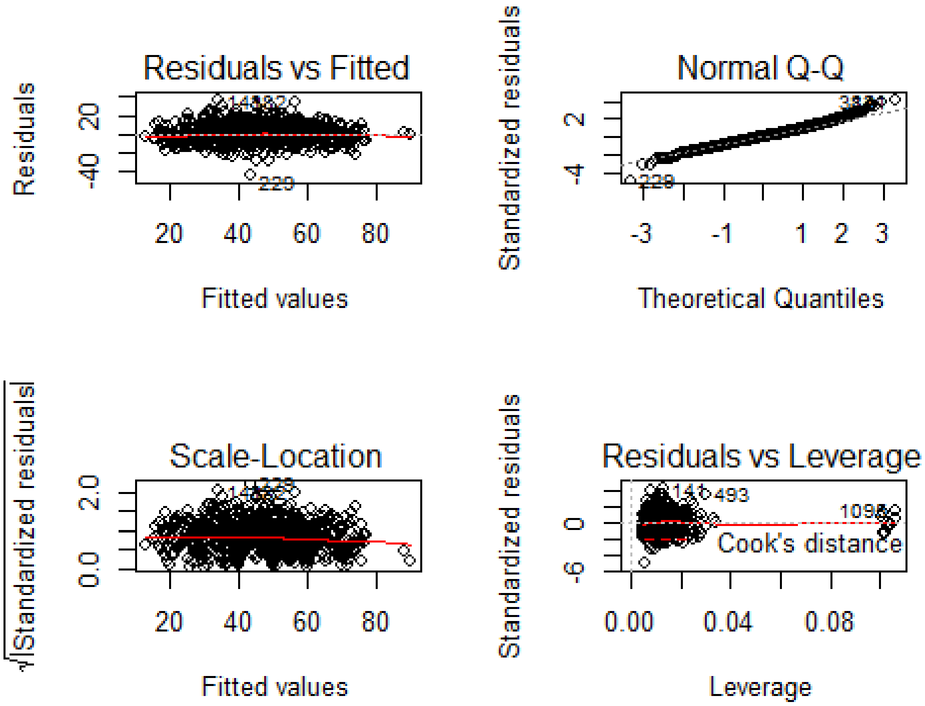
Part 2

Now run a multiple regression of cs1pccmk on inf1pcma with sector and period dummies, also include other covariates such as ld1pcla+ st1pcdev + hr3pcsat + pm1pcvc. Assess this model for the extent to which it satisfies the six regression assumptions discussed in class.

```
# Build and assess new model
m2 <- lm(cs1pccmk ~ inf1pcma + as.factor(sector) + as.factor(time_period) +
ld1pcla +
           st1pcdev + hr3pcsat + pm1pcvc, data = df)
stargazer(m2,type = "text")
```

```
## 
## ===================================================
##                              Dependent variable:   
##                          ---------------------------
##                                   cs1pccmk         
## ---------------------------------------------------
## inf1pcma                         0.246***          
##                                   (0.033)          
##                                                    
## as.factor(sector)2                1.565*           
##                                   (0.906)          
##                                                    
## as.factor(sector)3               -2.166***         
##                                   (0.815)          
##                                                    
## as.factor(sector)4               -2.744**          
##                                   (1.121)          
##                                                    
## as.factor(sector)5               -2.725***         
##                                   (1.000)          
##                                                    
## as.factor(sector)6               -9.137***         
##                                   (2.942)          
##                                                    
## as.factor(time_period)1           -0.493           
##                                   (0.827)          
##                                                    
## as.factor(time_period)2          -1.734**          
##                                   (0.819)          
##                                                    
## ld1pcla                           0.060*           
##                                   (0.035)          
##                                                    
## st1pcdev                         0.341***          
##                                   (0.027)          
##                                                    
## hr3pcsat                           0.016           
##                                   (0.027)          
##                                                    
## pm1pcvc                          0.181***          
##                                   (0.031)          
##                                                    
## Constant                         10.967***         
##                                   (1.221)          
##                                                    
## ---------------------------------------------------
## Observations                       1,098
## R2                                 0.697
## Adjusted R2                        0.694
## Residual Std. Error          8.937 (df = 1085)
## F Statistic             208.310*** (df = 12; 1085)
## ===================================================
## Note:                   *p<0.1; **p<0.05; ***p<0.01
```

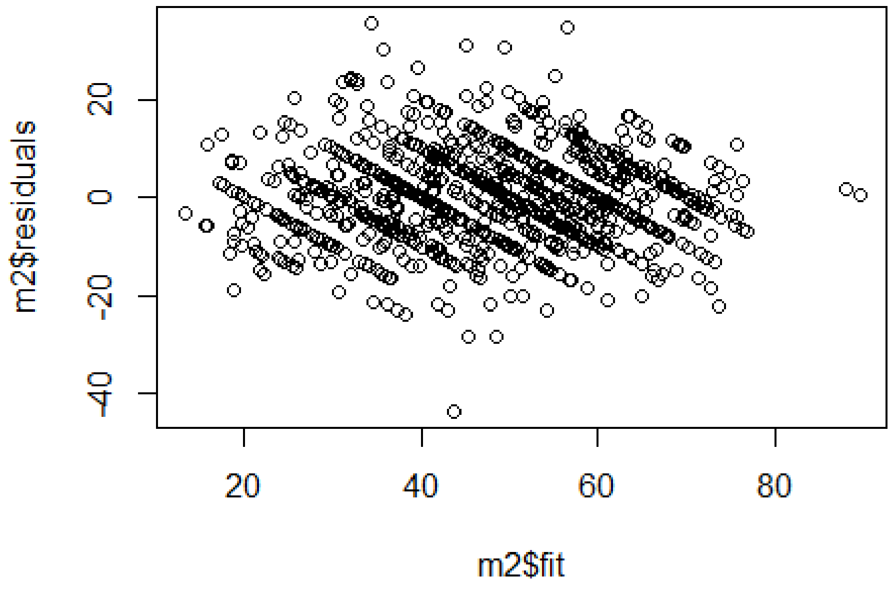
```
#1 Linearity
par(mfrow=c(2,2))
plot(m2)
```



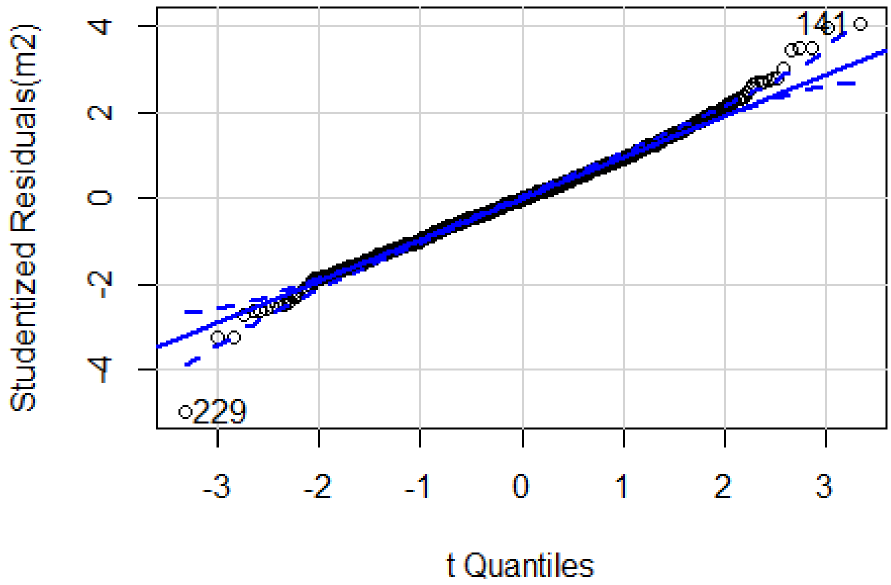
Interpretation:

The residuals and the fitted plots shows that the data is not linear as the red line shows a non linear pattern.

```
#2 Random sampling--Indep obs-- No autocorrelation--
par(mfrow=c(1,1))
plot (m2$residuals ~ m2$fit)
```

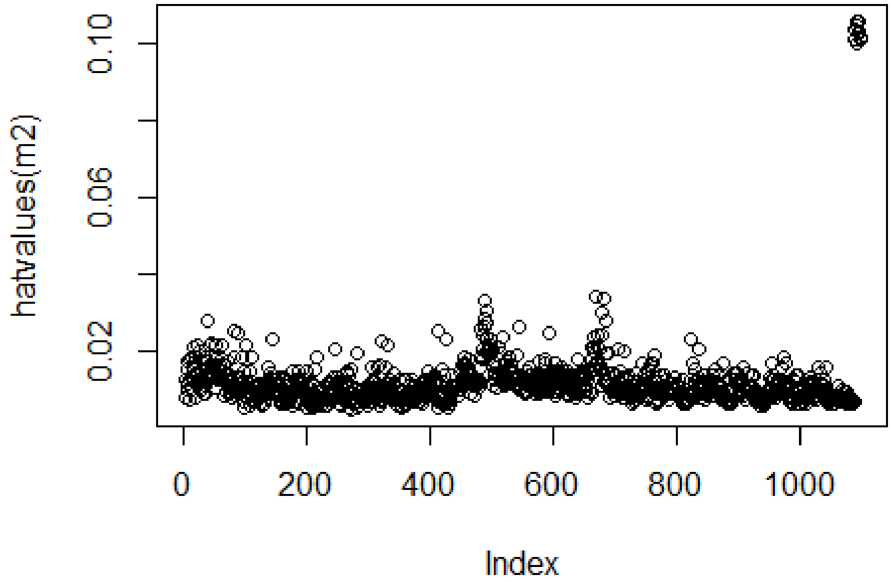


```
qqPlot(m2, simulate=T, labels=row.names(df))
```

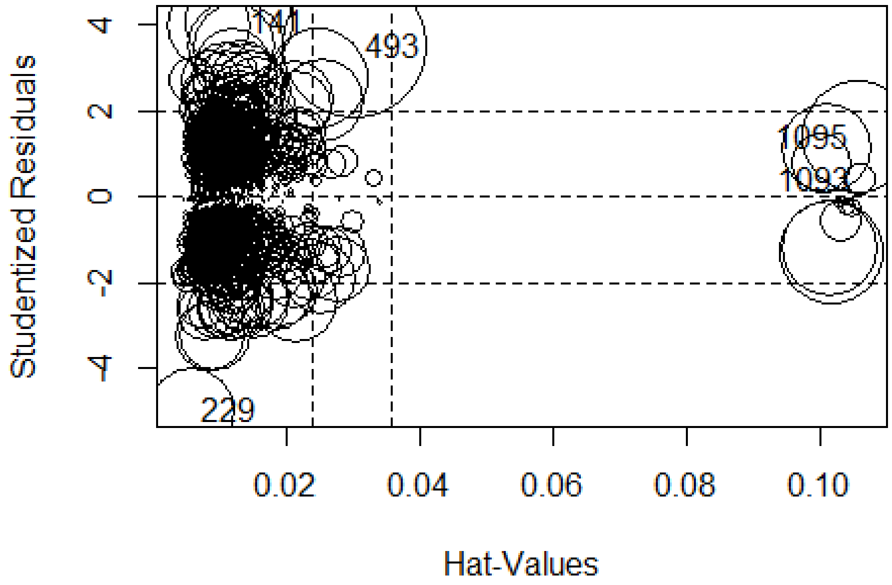


```
## [1] 141 229
```

```
#it gives us a 95% confidence interval
outlierTest(m2)   #does a formal test, applies Bonferroni, identifies obs

##     rstudent unadjusted p-value Bonferroni p
## 229 -4.95836         8.2461e-07   0.00090542

plot(hatvalues(m2))  #avg hatvalue h=(k+1)/n ; more than 2h or 3h problem
```



```
influencePlot(m2,id.method="identify")  #recheck if circles are prop to
cook's d

## Warning in plot.window(...): "id.method" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "id.method" is not a graphical
parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "id.method"
is not
## a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "id.method"
is not
## a graphical parameter

## Warning in box(...): "id.method" is not a graphical parameter

## Warning in title(...): "id.method" is not a graphical parameter
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.method" is not
a
## graphical parameter
```



```
##         StudRes         Hat       CookD
## 141   4.0492265 0.012992280 0.016369909
## 229  -4.9583601 0.005675061 0.010564175
## 493   3.5007128 0.030335975 0.029189351
## 1093  0.3856041 0.106051622 0.001357956
## 1095  1.3879749 0.105846333 0.017527211
```

Interpretation:

The studentized residuals graph shows us we have one regression outlier in model 2. However, the outlier may or may not be influential. Hat values tell us about potential influence of an observation. In the influence plot, the size of circle is proportional to Cook's distance in assumption #1(Residual vs Leverage graph), which tells us about actual influence of an observation for the model. When cases are outside of the Cook's distance / have high Cook's distance scores, the cases are influential to the regression results. In model 2, the regression outlier is not influential because it is not outside of the Cook's distance lines.

```
#3 conditional indep assumption--no endogeneity
resettest(m2)

##
##  RESET test
```

```
## 
## data:  m2
## RESET = 4.3685, df1 = 2, df2 = 1083, p-value = 0.01289
```

Interpretation:

At alpha=0.05, p-value of 0.012 is the evidence of functional form misspecification. The assumption is violated. We might be omitting squared terms or some variable.

```
#4 Multicollinerity
vif(m2)   #vif=1/(1-rjsquared) where j=1...p; usually vif>10 problematic
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## inf1pcma                4.292637  1        2.071868
## as.factor(sector)       2.169012  5        1.080504
## as.factor(time_period) 2.054809  2        1.197272
## ld1pcla                 4.953870  1        2.225729
## st1pcdev                2.964008  1        1.721629
## hr3pcsat                3.124945  1        1.767751
## pm1pcvc                 3.561494  1        1.887192
```

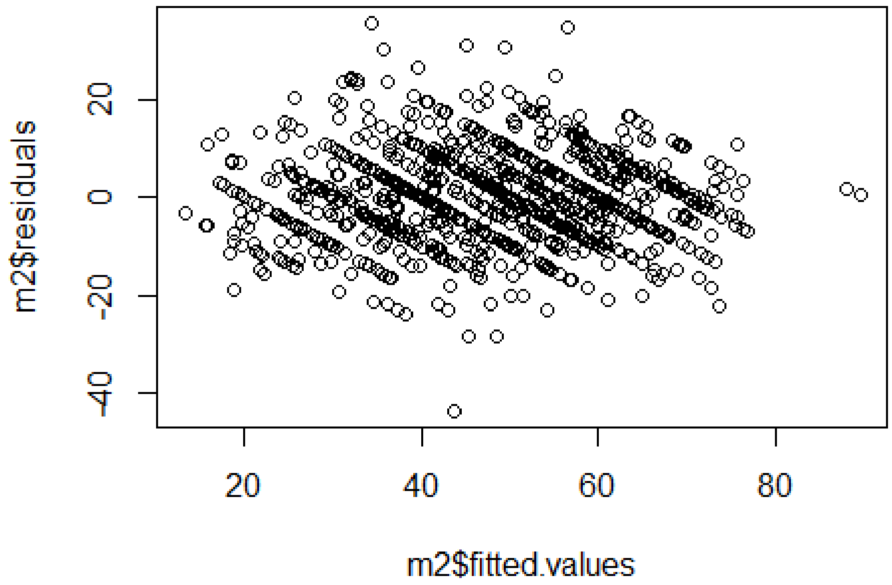
```
sqrt(vif(m2)) > 2   #flags coef that have high vif values say more than 4
```

```
##                          GVIF    Df GVIF^(1/(2*Df))
## inf1pcma                 TRUE FALSE           FALSE
## as.factor(sector)       FALSE  TRUE           FALSE
## as.factor(time_period) FALSE FALSE           FALSE
## ld1pcla                  TRUE FALSE           FALSE
## st1pcdev                FALSE FALSE           FALSE
## hr3pcsat                FALSE FALSE           FALSE
## pm1pcvc                 FALSE FALSE           FALSE
```

Interpretation:

After taking into account of the degree of freedoms for each variable, all VIF's are low (less than 5), indicating all predictor variables are not highly collinear with each other.

```
#5 Homoskedasticity
#Bartlett test is more sensitive to violations of normality than Levene test
plot(m2$residuals~m2$fitted.values)
```

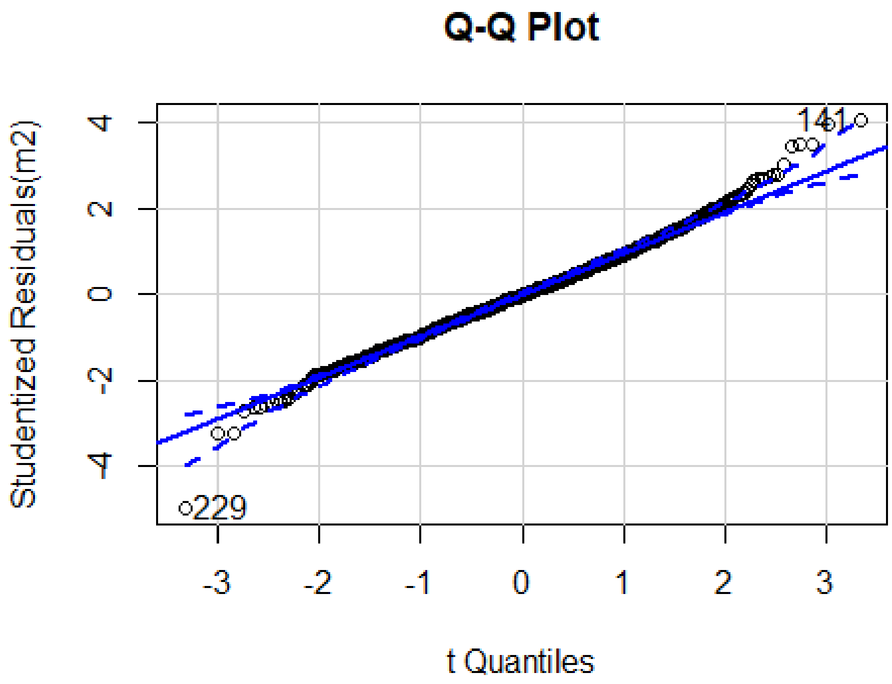


```
bartlett.test(list(m2$residuals,m2$fitted.values))    #here list coerces data objects into a dataframe which serves as input for Bartlett test

##
##  Bartlett test of homogeneity of variances
##
## data:  list(m2$residuals, m2$fitted.values)
## Bartlett's K-squared = 185.65, df = 1, p-value < 2.2e-16

#nonconstant variance score test--aka Breusch Pagan test
ncvTest(m2)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 9.817488, Df = 1, p = 0.0017286
```

Interpretation:

After running the Bartlett test, we can see that p-value is significant. Therefore we reject H0, concluding that the variances are not equal.

```
#6 Multivariate Normality
qqPlot(m2, labels = FALSE,
       simulate = TRUE, main = "Q-Q Plot")
```



```
## [1] 141 229

shapiro.test(m2$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  m2$residuals
## W = 0.99168, p-value = 7.333e-06
```

Interpretation:

The qq plot and the shapiro test show that the residual values are almost normally distributed. The points at the lower end and upper end are not lined up with qqline and there is a deviation.The mean of residuals is close to zero. The Shapiro test gives a p-value lesser than 0-05 and hence we reject the null hypothesis and the data is not normally distributed.