Homework2.R

Jingyi Wang

2021-10-07

# HOMEWORK 2  
# BAN250- JINGYI WANG  
  
  
# 1.(a)  
Sigma = matrix(c(1,-2,0,-2,5,0,0,0,2),nrow = 3,byrow = T)  
Sigma

## [,1] [,2] [,3]  
## [1,] 1 -2 0  
## [2,] -2 5 0  
## [3,] 0 0 2

## Since the cov(X1,X2) = (-2,-2)', we can get the conclusion that:  
## X1 and X2 are not independent.  
  
# 1.(b)  
## Since the cov(X2,X3) = (0,0)', we can get the conclusion that:  
## X2 and X3 are independent.  
  
# 1.(c)  
## In order to test the independence of (X1+X2) and X3, let's set a matrix A as followed:  
A = matrix(c(1,1,0,0,0,1), nrow = 2, byrow = T)  
t(A)

## [,1] [,2]  
## [1,] 1 0  
## [2,] 1 0  
## [3,] 0 1

## Set the covariance matrix of (X1+X2) and X3 as COV\_c:  
COV\_c = A %\*% Sigma %\*% t(A)  
COV\_c

## [,1] [,2]  
## [1,] 2 0  
## [2,] 0 2

## Since the COV((X1+X2),X3) = (0,0)', they are independent.

# 1.(d)  
## Since COV((X1+X2)/2,X3)= 1/2 COV((X1+X2),X3) = 0, We can get the conclusion that:The variable (X1+X2)/2 and X3 are also independent.

## We can prove this conclusion by conducting the same calculation as followed: In order to test the independence of (X1+X2)/2 and X3, let's set a matrix B as followed:  
B = matrix(c(1/2,1/2,0,0,0,1),nrow = 2,byrow = T)  
## Set the covariance matrix of (X1+X2)/2 and X3 as COV\_d:  
COV\_d = B %\*% Sigma %\*% t(B)  
COV\_d

## [,1] [,2]  
## [1,] 0.5 0  
## [2,] 0.0 2

## Since the COV((X1+X2)/2,X3) = (0,0)', we proved that they are independent.  
  
# 1.(e)  
## In order to test the independence of X2 and (X2-5/2X1-X3), let's set a matrix C as followed:  
C = matrix(c(0,1,0,-5/2,1,-1),nrow = 2, byrow = T)  
COV\_e = C %\*% Sigma %\*% t(C)  
COV\_e

## [,1] [,2]  
## [1,] 5 10.00  
## [2,] 10 23.25

## Since the COV(X2,(X2-5/2X1-X3)) = (10,10)', they are not independent.  
  
  
# 2.(a)  
## Set a 2x2 matrix for A and B:  
A = matrix(c(2,1,3,3), nrow = 2, byrow = T)  
B = matrix(c(0,4,1,2), nrow = 2, byrow = T)  
Com\_AOBO = matrix(c(2,1,0,0,3,3,0,0,0,0,0,4,0,0,1,2), nrow = 4, byrow = T)

A

## [,1] [,2]  
## [1,] 2 1  
## [2,] 3 3

B

## [,1] [,2]  
## [1,] 0 4  
## [2,] 1 2

Com\_AOBO

## [,1] [,2] [,3] [,4]  
## [1,] 2 1 0 0  
## [2,] 3 3 0 0  
## [3,] 0 0 0 4  
## [4,] 0 0 1 2

as.integer(det(Com\_AOBO))==as.integer(det(A)\*det(B))

## [1] TRUE

# 2.(b)  
## Set a 2X2 matrix for C:  
C = matrix(c(1,6,4,3), nrow = 2, byrow = T)  
Com\_ACOB = matrix(c(2,1,1,6,3,3,4,3,0,0,0,4,0,0,1,2), nrow = 4, byrow = T)  
A

## [,1] [,2]  
## [1,] 2 1  
## [2,] 3 3

B

## [,1] [,2]  
## [1,] 0 4  
## [2,] 1 2

C

## [,1] [,2]  
## [1,] 1 6  
## [2,] 4 3

Com\_ACOB

## [,1] [,2] [,3] [,4]  
## [1,] 2 1 1 6  
## [2,] 3 3 4 3  
## [3,] 0 0 0 4  
## [4,] 0 0 1 2

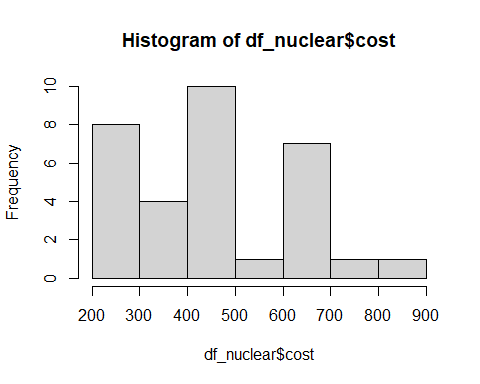
as.integer(det(Com\_ACOB))==as.integer(det(A)\*det(B))

## [1] TRUE

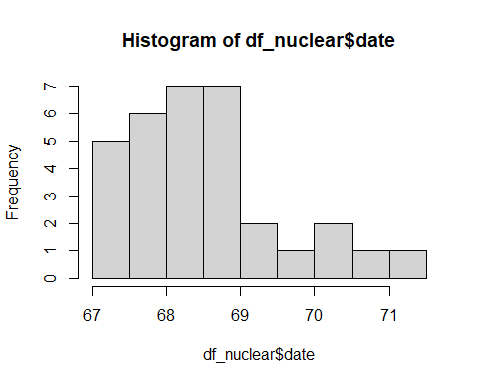
# 3.  
## import the Nuclear data set  
nuclear\_raw <- read.csv("C:/Users/jwang/OneDrive/Desktop/R\_BAN250/Homework/HW2/nuclear.csv")  
## select columns 2 to 6, and set a new dataframe named df\_nclear  
df\_nuclear = nuclear\_raw[,c(2:6)]  
## summary() can provide the mean for each column  
summary(df\_nuclear)

## cost date t1 t2   
## Min. :207.5 Min. :67.17 Min. : 7.00 Min. :44.00   
## 1st Qu.:310.3 1st Qu.:67.90 1st Qu.:11.75 1st Qu.:56.50   
## Median :448.1 Median :68.42 Median :13.00 Median :62.50   
## Mean :461.6 Mean :68.58 Mean :13.75 Mean :62.38   
## 3rd Qu.:612.0 3rd Qu.:68.92 3rd Qu.:15.25 3rd Qu.:70.25   
## Max. :881.2 Max. :71.08 Max. :22.00 Max. :85.00   
## cap   
## Min. : 457.0   
## 1st Qu.: 745.0   
## Median : 822.0   
## Mean : 825.4   
## 3rd Qu.: 947.2   
## Max. :1130.0

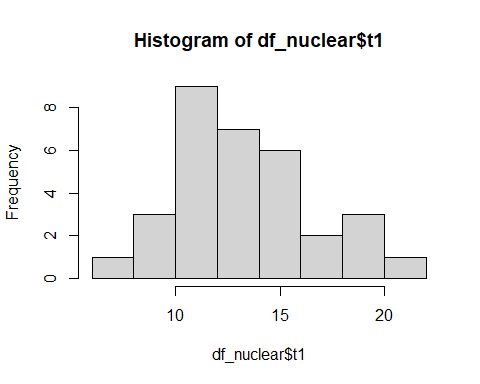
## We can use histogram to visualize the distribution of each variable:  
hist(df\_nuclear$cost)



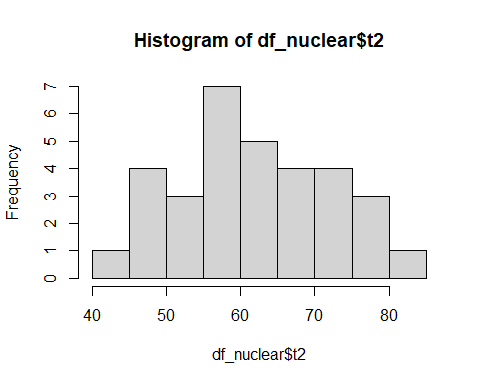
hist(df\_nuclear$date)



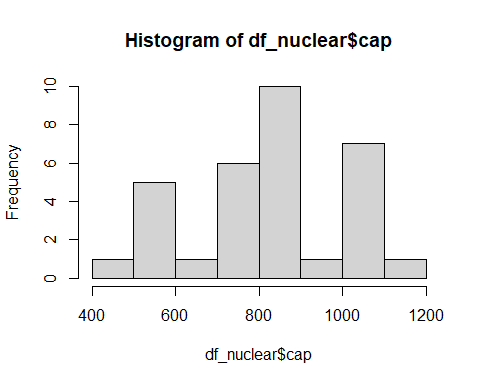
hist(df\_nuclear$t1)



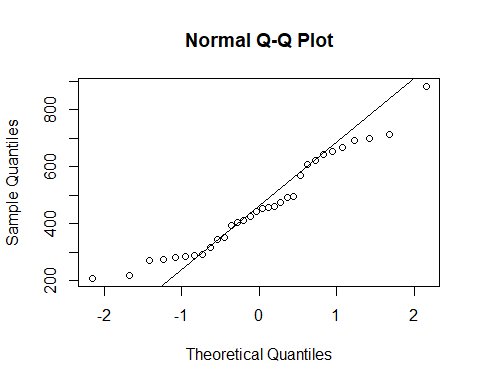
hist(df\_nuclear$t2)



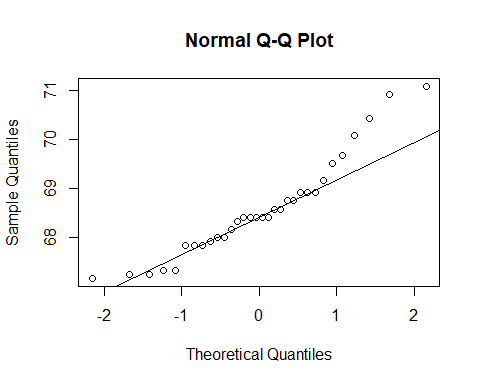
hist(df\_nuclear$cap)



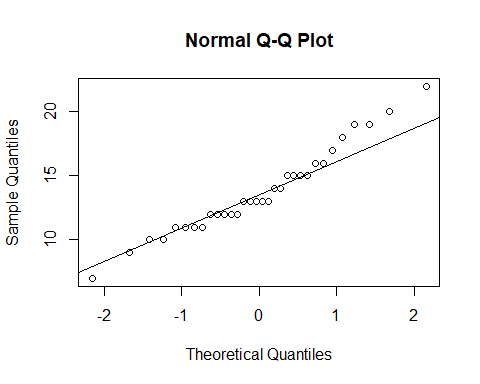
## From the histogram and the difference between the mean and median of the data in each column, I think we can tentatively identify the variable t2 is normally distributed, the variable t1 also looks like normally distributed, but the rest are not.  
  
## Conduct qqplot for each of these variables:  
## For the first column, variable: cost, Not Normally Distributed.  
## The dots curve and deviate from the straight diagonal at both ends.  
qqnorm(df\_nuclear$cost)  
qqline(df\_nuclear$cost)



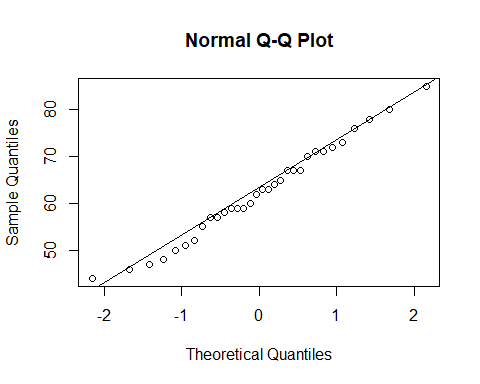
## For the second column, variable: date, Not Normally Distributed.  
## The dots deviate far away from the straight line as the value become bigger, which also  
## confirms the right-skewed distribution that we saw from the histogram.  
qqnorm(df\_nuclear$date)  
qqline(df\_nuclear$date)



## For the third column, variable: t1, Not Normally Distributed.  
## We can see clearly that the dots deviate far away from the straight line as the value become bigger.  
qqnorm(df\_nuclear$t1)  
qqline(df\_nuclear$t1)



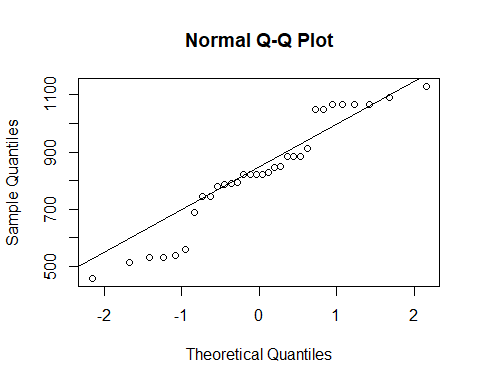
## For the forth column, variable: t2, Normally Distributed.  
## The dots overlap closely with the diagonal straight line.  
qqnorm(df\_nuclear$t2)  
qqline(df\_nuclear$t2)



## For the last column, variable: cap, Not Normally Distributed.  
## The dots deviate from the diagonal straight line at both sides.  
qqnorm(df\_nuclear$cap)  
qqline(df\_nuclear$cap)  
  
## MVN Test  
library(MVN)

## Warning: package 'MVN' was built under R version 4.0.5

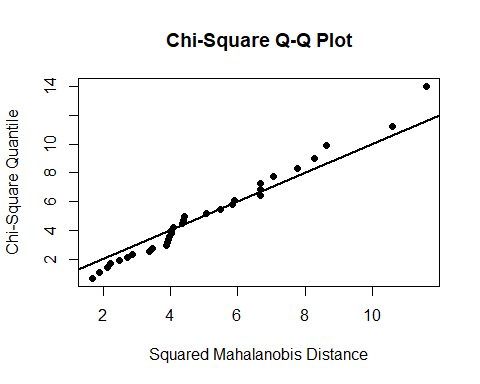
## Registered S3 methods overwritten by 'tibble':  
## method from   
## format.tbl pillar  
## print.tbl pillar



## -- Mardia Test & Anderson-Darling Test  
mvn(df\_nuclear,mvnTest="mardia",multivariatePlot = "qq")

## $multivariateNormality  
## Test Statistic p value Result  
## 1 Mardia Skewness 34.6879839256377 0.483072755539514 YES  
## 2 Mardia Kurtosis -1.40651537396082 0.15957114264546 YES  
## 3 MVN <NA> <NA> YES  
##   
## $univariateNormality  
## Test Variable Statistic p value Normality  
## 1 Anderson-Darling cost 0.5402 0.1531 YES   
## 2 Anderson-Darling date 0.7528 0.0449 NO   
## 3 Anderson-Darling t1 0.5078 0.1854 YES   
## 4 Anderson-Darling t2 0.1461 0.9633 YES   
## 5 Anderson-Darling cap 0.8128 0.0317 NO   
##   
## $Descriptives  
## n Mean Std.Dev Median Min Max 25th 75th  
## cost 32 461.56031 170.120670 448.105 207.51 881.24 310.3225 611.9625  
## date 32 68.58125 1.015267 68.420 67.17 71.08 67.8975 68.9200  
## t1 32 13.75000 3.369694 13.000 7.00 22.00 11.7500 15.2500  
## t2 32 62.37500 10.394633 62.500 44.00 85.00 56.5000 70.2500  
## cap 32 825.37500 189.359148 822.000 457.00 1130.00 745.0000 947.2500  
## Skew Kurtosis  
## cost 0.4543829 -0.71909410  
## date 0.8022647 0.06535129  
## t1 0.4937134 -0.22859170  
## t2 0.1371768 -0.77431764  
## cap -0.2402936 -0.88120435

## -- Henze-Zirkler Test & Anderson-Darling Test  
mvn(df\_nuclear,mvnTest="hz",multivariatePlot = "qq")



## $multivariateNormality  
## Test HZ p value MVN  
## 1 Henze-Zirkler 0.8574048 0.2428964 YES  
##   
## $univariateNormality  
## Test Variable Statistic p value Normality  
## 1 Anderson-Darling cost 0.5402 0.1531 YES   
## 2 Anderson-Darling date 0.7528 0.0449 NO   
## 3 Anderson-Darling t1 0.5078 0.1854 YES   
## 4 Anderson-Darling t2 0.1461 0.9633 YES   
## 5 Anderson-Darling cap 0.8128 0.0317 NO   
##   
## $Descriptives  
## n Mean Std.Dev Median Min Max 25th 75th  
## cost 32 461.56031 170.120670 448.105 207.51 881.24 310.3225 611.9625  
## date 32 68.58125 1.015267 68.420 67.17 71.08 67.8975 68.9200  
## t1 32 13.75000 3.369694 13.000 7.00 22.00 11.7500 15.2500  
## t2 32 62.37500 10.394633 62.500 44.00 85.00 56.5000 70.2500  
## cap 32 825.37500 189.359148 822.000 457.00 1130.00 745.0000 947.2500  
## Skew Kurtosis  
## cost 0.4543829 -0.71909410  
## date 0.8022647 0.06535129  
## t1 0.4937134 -0.22859170  
## t2 0.1371768 -0.77431764  
## cap -0.2402936 -0.88120435

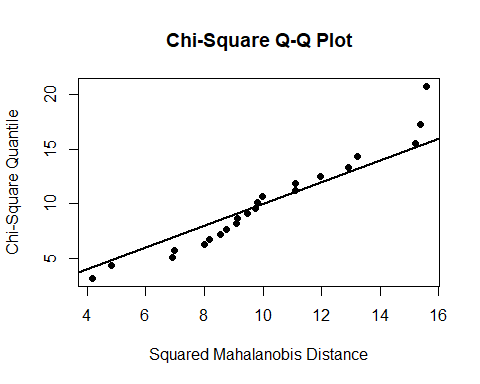
## -- Royston Test & Anderson-Darling Test  
mvn(df\_nuclear,mvnTest="royston",multivariatePlot = "qq")

## $multivariateNormality  
## Test H p value MVN  
## 1 Royston 10.35908 0.05044789 YES  
##   
## $univariateNormality  
## Test Variable Statistic p value Normality  
## 1 Anderson-Darling cost 0.5402 0.1531 YES   
## 2 Anderson-Darling date 0.7528 0.0449 NO   
## 3 Anderson-Darling t1 0.5078 0.1854 YES   
## 4 Anderson-Darling t2 0.1461 0.9633 YES   
## 5 Anderson-Darling cap 0.8128 0.0317 NO   
##   
## $Descriptives  
## n Mean Std.Dev Median Min Max 25th 75th  
## cost 32 461.56031 170.120670 448.105 207.51 881.24 310.3225 611.9625  
## date 32 68.58125 1.015267 68.420 67.17 71.08 67.8975 68.9200  
## t1 32 13.75000 3.369694 13.000 7.00 22.00 11.7500 15.2500  
## t2 32 62.37500 10.394633 62.500 44.00 85.00 56.5000 70.2500  
## cap 32 825.37500 189.359148 822.000 457.00 1130.00 745.0000 947.2500  
## Skew Kurtosis  
## cost 0.4543829 -0.71909410  
## date 0.8022647 0.06535129  
## t1 0.4937134 -0.22859170  
## t2 0.1371768 -0.77431764  
## cap -0.2402936 -0.88120435

## Conclusion:  
### For MVN Test:  
### -- Ho: Data follows a multivariate normal distribution  
### -- Ha: Data is not MVN  
### -- The P-value for all three MVN tests is > 0.05, we do no reject Ho, which means we have evidence to support the null hypothesis. In other words, the distribution of data are not significantly different from multivariate normal distribution.  
### For Univariate Test (Anderson-Darling Test):  
### -- Ho: Data is normal distributed  
### -- Ha: Data is not normal distributed  
### -- For variables: cost, t1, and t2, P-value > 0.05,we do not reject Ho, which means we have evidence to support that the three variables: cost, t1, and t2 are normally distributed.   
### -- For variables: date, cap, P-value < 0.05, we reject Ho, which means  
there is less than 5% probability that the null hypothesis is true, therefore, we believe they are not normal distributed.  
  
  
# 4.(a)  
## Import the data set  
propval <- read.table("C:/Users/jwang/OneDrive/Desktop/R\_BAN250/Homework/HW2/propval.txt", header=TRUE, quote="\"")  
## Test the normality for each variable by running Anderson-Darling Test:  
mvn(propval)$univariateNormality

## Test Variable Statistic p value Normality  
## 1 Anderson-Darling y 0.4953 0.1922 YES   
## 2 Anderson-Darling x1 0.7776 0.0367 NO   
## 3 Anderson-Darling x2 4.1878 <0.001 NO   
## 4 Anderson-Darling x3 0.1985 0.8701 YES   
## 5 Anderson-Darling x4 0.3974 0.3383 YES   
## 6 Anderson-Darling x5 1.5305 4e-04 NO   
## 7 Anderson-Darling x6 1.3684 0.0011 NO   
## 8 Anderson-Darling x7 2.7522 <0.001 NO   
## 9 Anderson-Darling x8 0.2838 0.5979 YES   
## 10 Anderson-Darling x9 4.9002 <0.001 NO

### Conclusion:  
### -- Ho: Data is normal distributed  
### -- Ha: Data is not normal distributed  
### -- For variables: y(Sale price of the house),x3(Lot size),x4(Living space),and x8(Age of the home),  
### -- P-value > 0.05,we do not reject Ho, which means we have evidence to support the Ho, they are normally distributed.  
### -- For the rest of variables (x1,x2,x5,x6,x7, and x9), P-value < 0.05, we reject Ho, which means  
### -- there is less than 5% probability that the Ho is true, therefore, we believe they are not normal distributed.  
  
## Test the MVN   
## -- Mardia Test   
mardia = mvn(propval,mvnTest="mardia")$multivariateNormality  
## -- Dornik-Haansen  
dh = mvn(propval,mvnTest="dh")$multivariateNormality  
## -- Royston Test   
royston = mvn(propval,mvnTest="royston")$multivariateNormality  
mvn(propval,multivariatePlot = "qq")



## $multivariateNormality  
## Test HZ p value MVN  
## 1 Henze-Zirkler 0.9742683 0.1210746 YES  
##   
## $univariateNormality  
## Test Variable Statistic p value Normality  
## 1 Anderson-Darling y 0.4953 0.1922 YES   
## 2 Anderson-Darling x1 0.7776 0.0367 NO   
## 3 Anderson-Darling x2 4.1878 <0.001 NO   
## 4 Anderson-Darling x3 0.1985 0.8701 YES   
## 5 Anderson-Darling x4 0.3974 0.3383 YES   
## 6 Anderson-Darling x5 1.5305 4e-04 NO   
## 7 Anderson-Darling x6 1.3684 0.0011 NO   
## 8 Anderson-Darling x7 2.7522 <0.001 NO   
## 9 Anderson-Darling x8 0.2838 0.5979 YES   
## 10 Anderson-Darling x9 4.9002 <0.001 NO   
##   
## $Descriptives  
## n Mean Std.Dev Median Min Max 25th 75th  
## y 22 34.9954545 6.1327751 36.40000 25.9000 45.8000 29.925000 38.650000  
## x1 22 6.5394455 1.5529777 6.03625 4.5429 9.1416 5.119850 8.051025  
## x2 22 1.1818182 0.2461830 1.00000 1.0000 1.5000 1.000000 1.500000  
## x3 22 6.0755909 2.0218375 5.68500 2.2750 9.8900 4.991225 7.270150  
## x4 22 1.4106364 0.2817822 1.49400 0.9750 1.8900 1.226750 1.627000  
## x5 22 1.2954545 0.6106026 1.00000 0.0000 2.0000 1.000000 2.000000  
## x6 22 6.5454545 0.9116846 6.00000 5.0000 8.0000 6.000000 7.000000  
## x7 22 3.1818182 0.5884899 3.00000 2.0000 4.0000 3.000000 3.750000  
## x8 22 37.5000000 14.6116392 40.00000 3.0000 62.0000 30.000000 49.500000  
## x9 22 0.2727273 0.4558423 0.00000 0.0000 1.0000 0.000000 0.750000  
## Skew Kurtosis  
## y 0.15253176 -1.3225218  
## x1 0.35282191 -1.4591677  
## x2 0.52873390 -1.7959711  
## x3 0.19297969 -0.8407037  
## x4 -0.06078948 -1.2461199  
## x5 -0.42573149 -0.5674873  
## x6 0.23200852 -0.9961664  
## x7 -0.02211853 -0.4965961  
## x8 -0.40158314 -0.5906288  
## x9 0.95182981 -1.1397211

### Conclusion:   
### -- Ho: Data is MVN  
### -- Ha: Data is not MVN  
### -- Two of the four tests considered the data to be consistent with the MVN distribution. Therefore, I think we have sufficient evidence to support Ho, the distribution of data is not significantly different from MVN.  
  
# 4.(b)  
library(car)

## Loading required package: carData

library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

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

library(nortest)  
## Run a regression model with all the variables  
md1 = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9, data = propval)   
summary(md1)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9,   
## data = propval)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8504 -1.4017 0.0929 1.7541 3.7206   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.11351 5.88549 2.908 0.0131 \*  
## x1 2.39009 1.05740 2.260 0.0432 \*  
## x2 5.74422 4.35113 1.320 0.2114   
## x3 0.12998 0.52530 0.247 0.8087   
## x4 2.63623 4.34493 0.607 0.5553   
## x5 2.32382 1.46160 1.590 0.1378   
## x6 -1.62471 2.40137 -0.677 0.5115   
## x7 -0.09723 3.38794 -0.029 0.9776   
## x8 -0.04445 0.06212 -0.716 0.4879   
## x9 2.03656 1.97372 1.032 0.3225   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.841 on 12 degrees of freedom  
## Multiple R-squared: 0.8774, Adjusted R-squared: 0.7854   
## F-statistic: 9.539 on 9 and 12 DF, p-value: 0.0003125

## The utility of this model is good, since the P-value of F-test is less than 0.05,we do not reject the Ho that our model is utility;  
## However, only one variable is significant in our model:  
## -- Ho: Beta\_n is 0; Ha : Beta\_n is not 0  
## -- Only variable X1 has P-value < 0.05, we reject Ho; in other words, we have evidence to support the coefficient of X1 is not 0;  
## -- Other variables are all non significant, which means those variables are not significant in the model and therefore should not be included.  
## The model has Adjusted R-squared: 78.54%, in other words, 78.54% of the variance in Y is explained by this regression model.   
  
vif(md1)

## x1 x2 x3 x4 x5 x6 x7 x8   
## 7.015543 2.985189 2.934683 3.899826 2.072176 12.469747 10.341931 2.143457   
## x9   
## 2.105980

# 4.(c) Let's check 4 assumptions!  
## No.1 Mean of the residuals is 0.  
mean(md1$residuals)

## [1] 4.979445e-17

### Conclusion: mean of the residuals is almost 0, therefore, md1 meets this assumption  
  
## No.2 Residuals are normally distributed.  
shapiro.test(md1$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: md1$residuals  
## W = 0.9679, p-value = 0.6625

ad.test(md1$residuals)

##   
## Anderson-Darling normality test  
##   
## data: md1$residuals  
## A = 0.2543, p-value = 0.6975

### Conclusion: md1 meets this assumption  
### -- Ho: Residuals are normally distributed.  
### -- Ha: Residuals are not normally distributed.  
### -- P-value > 0.05, we have evidence to support Ho, DO NOT REJECT Ho.  
  
## No.3 Residuals have constant variance.  
bptest(md1)

##   
## studentized Breusch-Pagan test  
##   
## data: md1  
## BP = 13.341, df = 9, p-value = 0.1478

### Conclusion: md1 meets this assumption  
### -- Ho: Residuals have constant variance.  
### -- Ha: Residuals does not have constant variance.  
### -- P-value > 0.05, we have evidence to support Ho, DO NOT REJECT Ho.  
  
## No.4 Residuals are independent of each other.  
durbinWatsonTest(md1)

## lag Autocorrelation D-W Statistic p-value  
## 1 -0.1865815 2.227783 0.682  
## Alternative hypothesis: rho != 0

### Conclusion: md1 meets this assumption  
### -- Ho: Residuals are independent.  
### -- Ha: Residuals are not independent  
### -- P-value > 0.05, we have evidence to support Ho, DO NOT REJECT Ho.  
#### All 4 assumptions are validated.  
  
# 4.(d)   
## We can test for multicollinearity:  
vif(md1)

## x1 x2 x3 x4 x5 x6 x7 x8   
## 7.015543 2.985189 2.934683 3.899826 2.072176 12.469747 10.341931 2.143457   
## x9   
## 2.105980

### The numbers for variable X1, X6, and X7 are all bigger than 6, which indicates we have multicollinearity problem.  
# 5.  
library(MVN)  
## Import the data  
elementart <- read.csv("C:/Users/jwang/OneDrive/Desktop/R\_BAN250/Homework/HW2/elementart(1).csv")  
typeof(elementart)

## [1] "list"

elementart = as.data.frame(elementart,row.names = NULL)  
is.data.frame(elementart)

## [1] TRUE

## Select the following variables and make a new data frame:  
## -- the number of English language learners: $ell ~ 8(As Dependent Variable)  
## -- the percentage of free meals: $meals ~ 7  
## -- year Round school: $yr\_rnd ~ 9  
## -- mobility: $mobility ~ 10  
## -- average class size in k-3: $acs\_k3 ~ 11  
## -- average class size in 4-6: $acs\_46 ~ 12  
## -- pct of full credential: $full ~ 19  
## -- pct of emer credentials: $emer ~ 20  
## -- the number of students enrolled: $enroll ~ 21  
rawdata = elementart[, c(7,9:12,19:21,8)]  
  
## Review our new data set  
## -- set variable yr\_rnd as dummy variable (as.factor)  
rawdata$yr\_rnd = as.factor(rawdata$yr\_rnd)  
summary(rawdata)

## meals yr\_rnd mobility acs\_k3 acs\_46   
## Min. : 0.00 No :308 Min. : 2.00 Min. :14.00 Min. :20.00   
## 1st Qu.: 31.00 Yes: 92 1st Qu.:13.00 1st Qu.:18.00 1st Qu.:27.00   
## Median : 67.50 Median :17.00 Median :19.00 Median :29.00   
## Mean : 60.31 Mean :18.25 Mean :19.16 Mean :29.69   
## 3rd Qu.: 90.00 3rd Qu.:22.00 3rd Qu.:20.00 3rd Qu.:31.00   
## Max. :100.00 Max. :47.00 Max. :25.00 Max. :50.00   
## NA's :1 NA's :2 NA's :3   
## full emer enroll ell   
## Min. : 37.00 Min. : 0.00 Min. : 130.0 Min. : 0.00   
## 1st Qu.: 76.00 1st Qu.: 3.00 1st Qu.: 320.0 1st Qu.: 9.75   
## Median : 88.00 Median :10.00 Median : 435.0 Median :25.00   
## Mean : 84.55 Mean :12.66 Mean : 483.5 Mean :31.45   
## 3rd Qu.: 97.00 3rd Qu.:19.00 3rd Qu.: 608.0 3rd Qu.:50.25   
## Max. :100.00 Max. :59.00 Max. :1570.0 Max. :91.00   
##

## -- check missing data, no suspicious outliers  
clean\_data = rawdata[complete.cases(rawdata),]  
levels(clean\_data$yr\_rnd) # double-check the dummy variable

## [1] "No" "Yes"

summary(clean\_data)

## meals yr\_rnd mobility acs\_k3 acs\_46   
## Min. : 0.00 No :303 Min. : 2.00 Min. :14.00 Min. :20.00   
## 1st Qu.: 31.00 Yes: 92 1st Qu.:13.00 1st Qu.:18.00 1st Qu.:27.00   
## Median : 67.00 Median :17.00 Median :19.00 Median :29.00   
## Mean : 60.17 Mean :18.29 Mean :19.16 Mean :29.68   
## 3rd Qu.: 89.50 3rd Qu.:22.00 3rd Qu.:20.00 3rd Qu.:31.00   
## Max. :100.00 Max. :47.00 Max. :25.00 Max. :50.00   
## full emer enroll ell   
## Min. : 37.0 Min. : 0.00 Min. : 130.0 Min. : 0.00   
## 1st Qu.: 76.0 1st Qu.: 3.00 1st Qu.: 321.0 1st Qu.: 9.00   
## Median : 88.0 Median :10.00 Median : 436.0 Median :25.00   
## Mean : 84.6 Mean :12.67 Mean : 485.4 Mean :31.16   
## 3rd Qu.: 97.0 3rd Qu.:19.00 3rd Qu.: 611.0 3rd Qu.:49.50   
## Max. :100.0 Max. :59.00 Max. :1570.0 Max. :91.00

## -- check the normality of numerical variables: Unfortunately, they are all not norally distributed  
num\_data = clean\_data[,c(1,3:9)]  
mvn(num\_data)$univariateNormality

## Test Variable Statistic p value Normality  
## 1 Anderson-Darling meals 10.9125 <0.001 NO   
## 2 Anderson-Darling mobility 3.9263 <0.001 NO   
## 3 Anderson-Darling acs\_k3 11.4590 <0.001 NO   
## 4 Anderson-Darling acs\_46 9.7124 <0.001 NO   
## 5 Anderson-Darling full 13.0269 <0.001 NO   
## 6 Anderson-Darling emer 10.6969 <0.001 NO   
## 7 Anderson-Darling enroll 6.4546 <0.001 NO   
## 8 Anderson-Darling ell 10.1844 <0.001 NO

## Run a regression model with all the variables

model1 = lm(ell ~ meals + yr\_rnd + mobility + acs\_k3 + acs\_46 + full + emer + enroll, data = clean\_data)   
summary(model1)

## Call:  
## lm(formula = ell ~ meals + yr\_rnd + mobility + acs\_k3 + acs\_46 +   
## full + emer + enroll, data = clean\_data)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -41.281 -8.382 -1.115 8.407 38.743   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -23.705929 15.002823 -1.580 0.1149   
## meals 0.579360 0.028720 20.173 < 2e-16 \*\*\*  
## yr\_rndYes 4.813394 2.223826 2.164 0.0310 \*   
## mobility -0.666933 0.099792 -6.683 8.20e-11 \*\*\*  
## acs\_k3 1.325880 0.540156 2.455 0.0145 \*   
## acs\_46 -0.074814 0.192875 -0.388 0.6983   
## full -0.025368 0.114974 -0.221 0.8255   
## emer -0.042633 0.146279 -0.291 0.7709   
## enroll 0.021804 0.003903 5.586 4.39e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.72 on 386 degrees of freedom  
## Multiple R-squared: 0.6991, Adjusted R-squared: 0.6929   
## F-statistic: 112.1 on 8 and 386 DF, p-value: < 2.2e-16

### Passed F-test, with Adjusted R-squared: 69.29%; however, there are 3 non significant coefficients.  
  
## Remove non significant coefficients (terms) but keep "acs\_46":  
## It's interesting that the coefficient of acs\_46 is non significant, while the coefficient of acs\_k3 is significant; However, since they are all stands for "class size", it doesn't make sense if we only remove one of them.   
model2 = lm(ell ~ meals + yr\_rnd + mobility + acs\_k3 + acs\_46 + enroll, data = clean\_data)   
summary(model2)

##   
## Call:  
## lm(formula = ell ~ meals + yr\_rnd + mobility + acs\_k3 + acs\_46 +   
## enroll, data = clean\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41.681 -8.360 -1.044 8.304 38.653   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -26.113097 10.809810 -2.416 0.0162 \*   
## meals 0.578012 0.025774 22.426 < 2e-16 \*\*\*  
## yr\_rndYes 4.704816 2.183350 2.155 0.0318 \*   
## mobility -0.670884 0.096646 -6.942 1.64e-11 \*\*\*  
## acs\_k3 1.314647 0.532837 2.467 0.0140 \*   
## acs\_46 -0.070911 0.191812 -0.370 0.7118   
## enroll 0.021801 0.003823 5.703 2.34e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.68 on 388 degrees of freedom  
## Multiple R-squared: 0.699, Adjusted R-squared: 0.6944   
## F-statistic: 150.2 on 6 and 388 DF, p-value: < 2.2e-16

### The utility of model2 is good, since the P-value of F-test is less than 0.05, we do not reject the Ho that our model is utility; The Adjusted R-squared of model2 is 69.44%, which is improved relative to model1.  
  
## Let's test Mulicollinearity:  
vif(model2)

## meals yr\_rnd mobility acs\_k3 acs\_46 enroll   
## 1.422067 1.796599 1.107551 1.127502 1.147349 1.577261

### which is good, all smaller than 6, we don't have mulicollinearity problem

## Let's test 4 regression assumptions:  
## - No.1 Mean of the residuals is 0.  
mean(model2$residuals)

## [1] 2.256315e-16

### Conclusion: mean of the residuals is almost 0, therefore, model2 meets this assumption  
  
## - No.2 Residuals are normally distributed.  
shapiro.test(model2$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: model2$residuals  
## W = 0.99376, p-value = 0.1036

ad.test(model2$residuals)

##   
## Anderson-Darling normality test  
##   
## data: model2$residuals  
## A = 0.82658, p-value = 0.03266

### Conclusion: the result of shapiro test support that model2 meets this assumption.  
### -- Ho: Residuals are normally distributed.  
### -- Ha: Residuals are not normally distributed.  
### -- P-value > 0.05, we have evidence to support Ho, DO NOT REJECT Ho.

## - No.3 Residuals have constant variance.  
bptest(model2)

##   
## studentized Breusch-Pagan test  
##   
## data: model2  
## BP = 37.119, df = 6, p-value = 1.669e-06

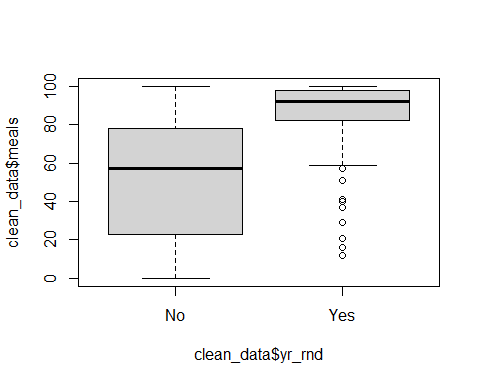
### Conclusion: model2 against this assumption  
### -- Ho: Residuals have constant variance.  
### -- Ha: Residuals does not have constant variance.  
### -- P-value < 0.05, we do not have evidence to support Ho, we REJECT Ho.  
  
## - No.4 Residuals are independent of each other.  
durbinWatsonTest(model2)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.2914791 1.410662 0  
## Alternative hypothesis: rho != 0

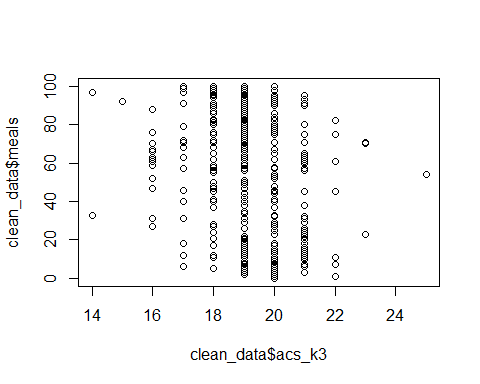
### Conclusion: model2 against this assumption  
### -- Ho: Residuals are independent.  
### -- Ha: Residuals are not independent  
### -- P-value < 0.05, we do not have evidence to support Ho, we REJECT Ho.  
  
#### For model2: Only 2 assumptions out of 4 are validated.

## Since the variables are not normally distributed, I want to visualize the correlations between variables:

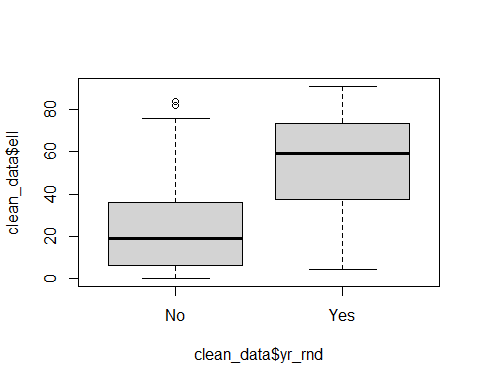
plot(clean\_data$meals ~ clean\_data$yr\_rnd)



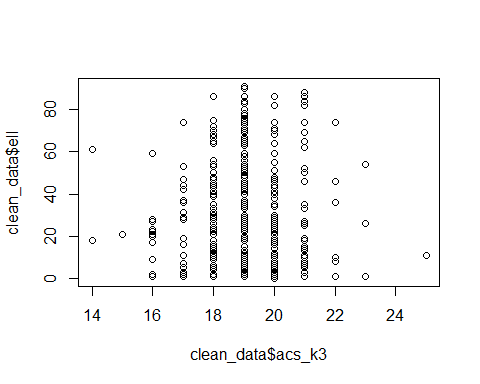
plot(clean\_data$meals ~ clean\_data$acs\_k3)



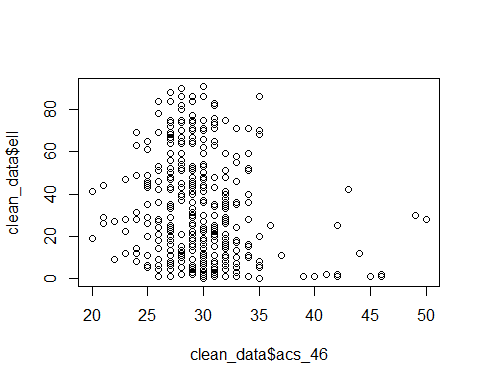
plot(clean\_data$ell ~ clean\_data$yr\_rnd)



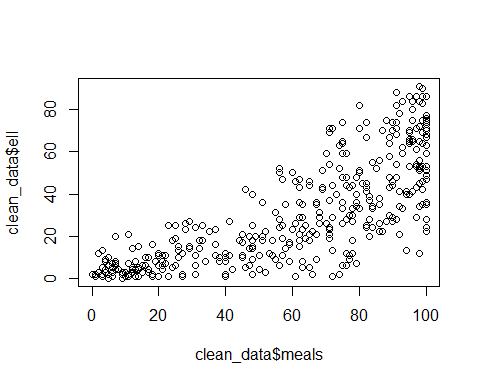
plot(clean\_data$ell ~ clean\_data$acs\_k3)



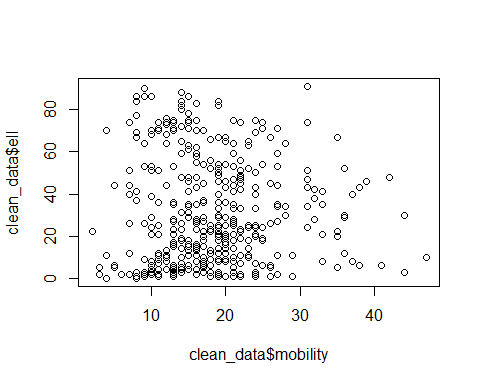
plot(clean\_data$ell ~ clean\_data$acs\_46)



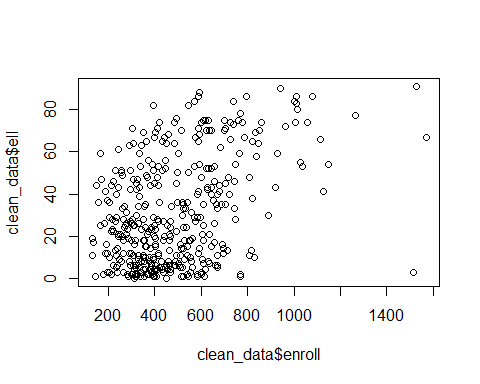
plot(clean\_data$ell ~ clean\_data$meals)



plot(clean\_data$ell ~ clean\_data$mobility)



plot(clean\_data$ell ~ clean\_data$enroll)



## Remove variable 'acs\_k3','acs\_46' from model2:  
model3 = lm(ell ~ meals + yr\_rnd + mobility + enroll, data = clean\_data)   
summary(model3) # Pass F-test, Adjusted R-squared: 0.6912

##   
## Call:  
## lm(formula = ell ~ meals + yr\_rnd + mobility + enroll, data = clean\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -42.895 -8.699 -0.805 8.206 40.015   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.050849 2.512470 -1.214 0.225   
## meals 0.565658 0.024670 22.929 < 2e-16 \*\*\*  
## yr\_rndYes 4.899937 2.192461 2.235 0.026 \*   
## mobility -0.657771 0.095432 -6.893 2.21e-11 \*\*\*  
## enroll 0.022795 0.003819 5.969 5.38e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.76 on 390 degrees of freedom  
## Multiple R-squared: 0.6943, Adjusted R-squared: 0.6912   
## F-statistic: 221.4 on 4 and 390 DF, p-value: < 2.2e-16

## repeat the tests for 4 assumptions:  
mean(model3$residuals)# PASS, Mean of residuals is 0

## [1] 6.256083e-16

shapiro.test(model3$residuals)# Failed for normal distribution

##   
## Shapiro-Wilk normality test  
##   
## data: model3$residuals  
## W = 0.99248, p-value = 0.04421

ad.test(model3$residuals)# Failed for normal distribution

##   
## Anderson-Darling normality test  
##   
## data: model3$residuals  
## A = 1.0483, p-value = 0.00926

bptest(model3)# Failed for constant variance

##   
## studentized Breusch-Pagan test  
##   
## data: model3  
## BP = 33.557, df = 4, p-value = 9.185e-07

durbinWatsonTest(model3)# Failed for constant variance

## lag Autocorrelation D-W Statistic p-value  
## 1 0.3008709 1.390064 0  
## Alternative hypothesis: rho != 0

### Model 3 failed at normality test, as well as the bptest and DurbinWatson Test.Only one assumption, mean of the residuals is 0, is validated.

## Remove variable 'yr\_rnd' from model3:  
model4 = lm(ell ~ meals + mobility + enroll, data = clean\_data)   
summary(model4) # Pass F-test, with Adjusted R-squared: 0.688

##   
## Call:  
## lm(formula = ell ~ meals + mobility + enroll, data = clean\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -45.894 -8.409 -0.953 7.609 38.738   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.022423 2.364490 -2.124 0.0343 \*   
## meals 0.586303 0.022991 25.501 < 2e-16 \*\*\*  
## mobility -0.682841 0.095253 -7.169 3.81e-12 \*\*\*  
## enroll 0.027594 0.003174 8.693 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.83 on 391 degrees of freedom  
## Multiple R-squared: 0.6904, Adjusted R-squared: 0.688   
## F-statistic: 290.6 on 3 and 391 DF, p-value: < 2.2e-16

mean(model4$residuals) # PASS, Mean of residuals is 0

## [1] 4.161206e-16

shapiro.test(model4$residuals) # Failed for normal distribution

##   
## Shapiro-Wilk normality test  
##   
## data: model4$residuals  
## W = 0.99185, p-value = 0.02908

ad.test(model4$residuals) # Failed for normal distribution

##   
## Anderson-Darling normality test  
##   
## data: model4$residuals  
## A = 1.1498, p-value = 0.005203

bptest(model4) # Failed for constant variance

##   
## studentized Breusch-Pagan test  
##   
## data: model4  
## BP = 28.208, df = 3, p-value = 3.285e-06

durbinWatsonTest(model4) # Failed for independency

## lag Autocorrelation D-W Statistic p-value  
## 1 0.2829442 1.425744 0  
## Alternative hypothesis: rho != 0

### No better than Model 3, failed 3 assumptions as model 3 did.  
  
  
## Based on the plot of ell and meals, I try to squared-root the variable “meals” based on model3:  
model5 = lm(ell ~ I(meals^1/2) + yr\_rnd + mobility + enroll, data = clean\_data)  
summary(model5) # Adjusted R-squared: 0.6912, Same as Model3

##   
## Call:  
## lm(formula = ell ~ I(meals^1/2) + yr\_rnd + mobility + enroll,   
## data = clean\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -42.895 -8.699 -0.805 8.206 40.015   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.050849 2.512470 -1.214 0.225   
## I(meals^1/2) 1.131317 0.049339 22.929 < 2e-16 \*\*\*  
## yr\_rndYes 4.899937 2.192461 2.235 0.026 \*   
## mobility -0.657771 0.095432 -6.893 2.21e-11 \*\*\*  
## enroll 0.022795 0.003819 5.969 5.38e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.76 on 390 degrees of freedom  
## Multiple R-squared: 0.6943, Adjusted R-squared: 0.6912   
## F-statistic: 221.4 on 4 and 390 DF, p-value: < 2.2e-16

mean(model5$residuals) # PASS, Mean of residuals is 0

## [1] 6.256083e-16

shapiro.test(model5$residuals)# Failed for normal distribution

##   
## Shapiro-Wilk normality test  
##   
## data: model5$residuals  
## W = 0.99248, p-value = 0.04421

ad.test(model5$residuals) # Failed for normal distribution

##   
## Anderson-Darling normality test  
##   
## data: model5$residuals  
## A = 1.0483, p-value = 0.00926

bptest(model5) # Failed for constant variance

##   
## studentized Breusch-Pagan test  
##   
## data: model5  
## BP = 33.557, df = 4, p-value = 9.185e-07

durbinWatsonTest(model5) # Failed for independency

## lag Autocorrelation D-W Statistic p-value  
## 1 0.3008709 1.390064 0  
## Alternative hypothesis: rho != 0

#### Conclusion:

- We should notice at the beginning that our dependent variables are not normally distributed.

Together with our relatively small sample size, we were able to anticipate that our regression models would most likely fail to meet the regression assumptions.

- I will go with Model3 (ell ~ meals + yr\_rnd + mobility + enroll)  
Since Model3 explains 69.12% of the variance with 4 variables compared to 69.44% with 6 variables in Model2.The rest models are no better than model3.

- Although all models have passed the F-test, they all not good at tests for regression assumptions:

One of the possible reasons is that we may miss some important factors, such as age. Recall that the coefficient of acs\_k3 is significant, while the coefficient of acs\_46 is non-significant. These two variables actually split our sample into two subsamples based on grade level, and there is evidence that these two subsamples are different, at least in our model.

Another possible reason is that there are other correlations between the variables, only that we have not found them yet.