PROJECT

**Pre Processing**

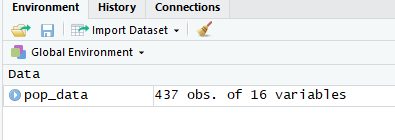
1.

#Loading data in R

library(readxl)

setwd("C:\\Users\\sanji\\Documents\\MS-USF\\Semester 2\\AMB\\Project")

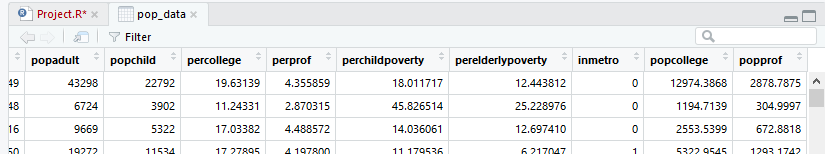
pop\_data = read\_excel("6304 Regression Project Data.xlsx")



2.

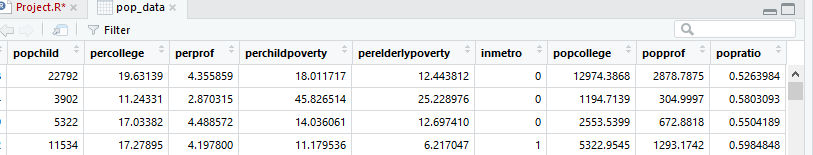
pop\_data$popcollege = (pop\_data$percollege \* pop\_data$poptotal)/100

pop\_data$popprof = (pop\_data$perprof \* pop\_data$poptotal)/100



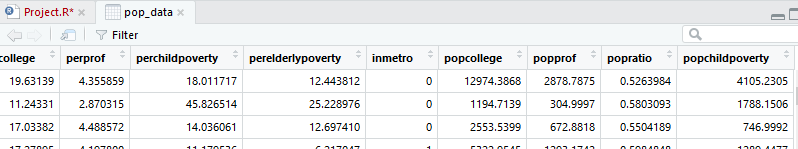
3.

pop\_data$popratio = (pop\_data$popchild / pop\_data$popadult)



4.

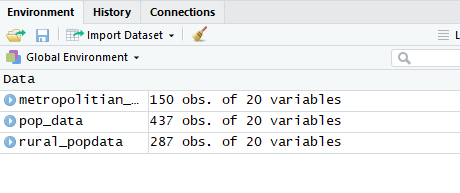
pop\_data$popchildpoverty = (pop\_data$popchild \* pop\_data$perchildpoverty)/100



5.

rural\_popdata=subset(pop\_data,pop\_data$inmetro==0)

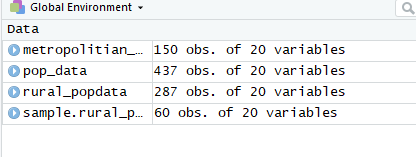
metropolitian\_popdata=subset(pop\_data,pop\_data$inmetro==1)



6.

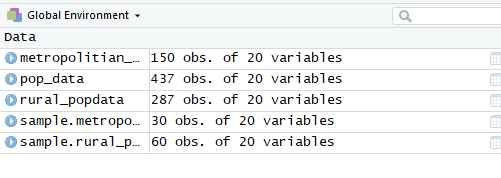
set.seed(44317266)

sample.rural\_popdata=rural\_popdata[sample(1:nrow(rural\_popdata),60,replace=FALSE),]



7.

sample.metropolitian\_popdata=metropolitian\_popdata[sample(1:nrow(metropolitian\_popdata),30,replace=FALSE),]

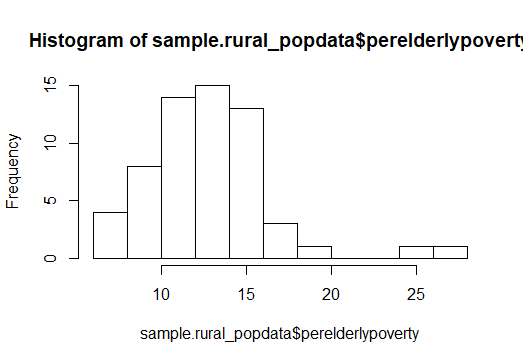


**Analysis:**

1.

#Finding the nature of the data

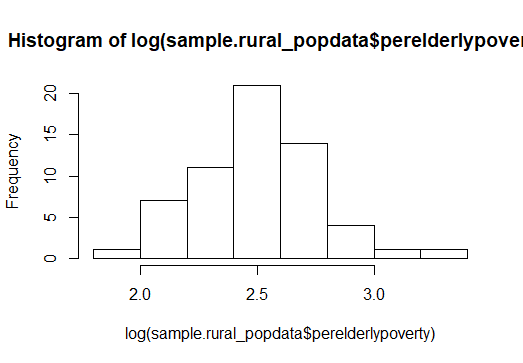
hist(sample.rural\_popdata$perelderlypoverty)



**As seen above the data is near to normal.**

**However after applying log transformation to it, it becomes more closer to normal.**

hist(log(sample.rural\_popdata$perelderlypoverty))



#Applying the Kitchen sink model

m1<-lm(sample.rural\_popdata$perelderlypoverty~sample.rural\_popdata$area+sample.rural\_popdata$poptotal+sample.rural\_popdata$popdensity+sample.rural\_popdata$popwhite+sample.rural\_popdata$popblack+sample.rural\_popdata$popasian+sample.rural\_popdata$popadult+sample.rural\_popdata$popchild+sample.rural\_popdata$popcollege+sample.rural\_popdata$popprof+sample.rural\_popdata$popratio+sample.rural\_popdata$popchildpoverty)

summary(m1)

Call:

lm(formula = sample.rural\_popdata$perelderlypoverty ~ sample.rural\_popdata$area +

sample.rural\_popdata$poptotal + sample.rural\_popdata$popdensity +

sample.rural\_popdata$popwhite + sample.rural\_popdata$popblack +

sample.rural\_popdata$popasian + sample.rural\_popdata$popadult +

sample.rural\_popdata$popchild + sample.rural\_popdata$popcollege +

sample.rural\_popdata$popprof + sample.rural\_popdata$popratio +

sample.rural\_popdata$popchildpoverty)

Residuals:

Min 1Q Median 3Q Max

-5.7665 -1.2933 -0.0196 1.2895 8.3109

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10.8855552 7.0066850 1.554 0.1269

sample.rural\_popdata$area -0.0027471 0.0022469 -1.223 0.2274

sample.rural\_popdata$poptotal -0.0008124 0.0010419 -0.780 0.4394

sample.rural\_popdata$popdensity -0.0700437 0.0302475 -2.316 0.0249 \*

sample.rural\_popdata$popwhite 0.0000816 0.0007815 0.104 0.9173

sample.rural\_popdata$popblack 0.0004287 0.0009716 0.441 0.6610

sample.rural\_popdata$popasian 0.0053612 0.0050874 1.054 0.2972

sample.rural\_popdata$popadult 0.0011206 0.0009495 1.180 0.2437

sample.rural\_popdata$popchild NA NA NA NA

sample.rural\_popdata$popcollege -0.0011697 0.0006222 -1.880 0.0662 .

sample.rural\_popdata$popprof 0.0001564 0.0015840 0.099 0.9218

sample.rural\_popdata$popratio 10.6851421 12.8514993 0.831 0.4098

sample.rural\_popdata$popchildpoverty 0.0034961 0.0006827 5.121 5.34e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.529 on 48 degrees of freedom

Multiple R-squared: 0.5934, Adjusted R-squared: 0.5003

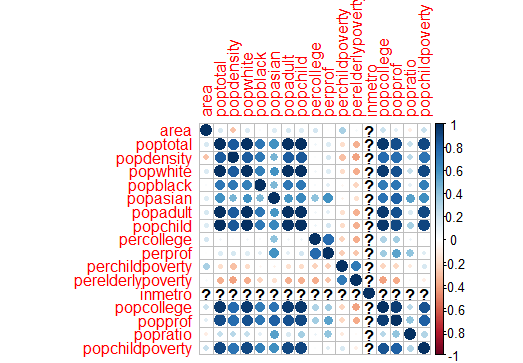
F-statistic: 6.369 on 11 and 48 DF, p-value: 2.272e-06

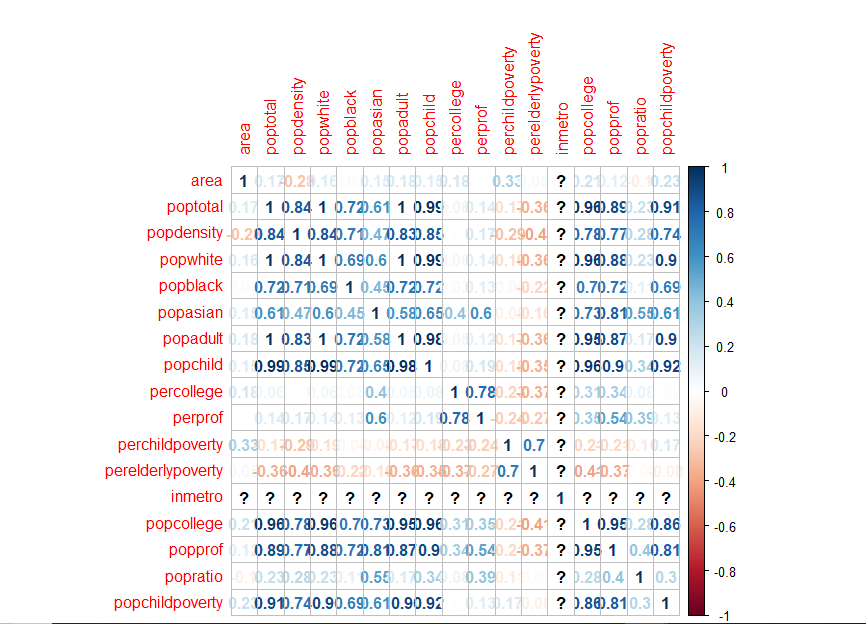
**Interpretation: The kitchen sink model has all the X variables to predict the Y variable(perelderlypoverty).This model explains 59% of variance of X variables on Y.As seen in the above model popdensity and popchildpoverty are significant predictors of perelderlypoverty.As per the beta coefficients, when the population density increases by one square unit, the percentage of elderly poverty decreases by 0.07% .Also, when child poverty population increases by one, the percentage of elderly poverty increases by 0.003%.**

#Finding correlations between X variables

x\_cor<- cor(cbind(sample.rural\_popdata[,4:20]))

corrplot(x\_cor)





**As seen from the above correlation matrix, the X variables considered in the kitchen sink model are correlated as per below:**

**Poptotal:-**

**poptotal and popadult are completely correlated, hence I have not included it in the analysis.**

**Popwhite:-**

**Popwhite and popadult are completely correlated, hence I have not included it in the analysis.**

**Popchild:-**

**Popchild and popadult are correlated by 0.98 , hence I have not included it in the analysis.Also, I want to focus the analysis on adult/elders as these parameter will be more significant to determine perelderlypoverty.**

**Popprof:-**

**Popprof and popcollege are highly correlated with each other(0.95), hence I have not included it in the analysis.**

**Popratio :-**

**I want to focus the analysis on adult/elders as these parameter will be more significant to determine perelderlypoverty.I think this ratio of child/adult will not have a significant effect on perelderlypoverty.**

#Applying model 2 with removal of redundant variables

m2<-lm(sample.rural\_popdata$perelderlypoverty~sample.rural\_popdata$popdensity+sample.rural\_popdata$popblack+sample.rural\_popdata$popasian+sample.rural\_popdata$popadult+sample.rural\_popdata$popcollege+sample.rural\_popdata$popchildpoverty)

summary(m2)

Call:

lm(formula = sample.rural\_popdata$perelderlypoverty ~ sample.rural\_popdata$popdensity +

sample.rural\_popdata$popblack + sample.rural\_popdata$popasian +

sample.rural\_popdata$popadult + sample.rural\_popdata$popcollege +

sample.rural\_popdata$popchildpoverty)

Residuals:

Min 1Q Median 3Q Max

-5.2464 -1.5792 -0.3479 1.2175 9.1458

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.469e+01 6.609e-01 22.223 < 2e-16 \*\*\*

sample.rural\_popdata$popdensity -4.285e-02 1.734e-02 -2.471 0.01672 \*

sample.rural\_popdata$popblack 2.947e-04 3.413e-04 0.864 0.39174

sample.rural\_popdata$popasian 7.244e-03 4.170e-03 1.737 0.08817 .

sample.rural\_popdata$popadult 3.988e-05 1.299e-04 0.307 0.76003

sample.rural\_popdata$popcollege -1.441e-03 4.483e-04 -3.214 0.00223 \*\*

sample.rural\_popdata$popchildpoverty 3.003e-03 6.038e-04 4.973 7.3e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.504 on 53 degrees of freedom

Multiple R-squared: 0.5601, Adjusted R-squared: 0.5103

F-statistic: 11.25 on 6 and 53 DF, p-value: 4.602e-08

**As seen from the above analysis the model has an r square of 0.56 i.e it explains about 56% of variance in the y variable. The model m1 is a slightly better model than m2 with respect to the r square.However, more meaningful X variables have been considered in m2 model, hence according to me m2 is a better model.**

**Also,the intercept, popdensity ,popcollege and popchildpoverty are significant variables in the model(more number of variables are significant compared to m1). Model m2(53) has more degrees of freedom than m1(48), making it a better model.**

**When all the x variables are zero, the value of percentage elderly poverty is 1.469e+01.**

**When popdensity increases by 1 square unit, the percentage elderly poverty decreases by 4.285e-02 .**

**When population having college degree increases by 1, the percentage elderly poverty decreases by 1.441e-03 .**

**When population of children poverty increases by 1, the percentage elderly poverty increases by 3.003e-03 .**

#Applying model 3 with log transformation on dependant variable(perelderlypoverty)

m3<-lm(log(sample.rural\_popdata$perelderlypoverty)~sample.rural\_popdata$popdensity+sample.rural\_popdata$popblack+sample.rural\_popdata$popasian+sample.rural\_popdata$popadult+sample.rural\_popdata$popcollege+sample.rural\_popdata$popchildpoverty)

summary(m3)

Call:

lm(formula = log(sample.rural\_popdata$perelderlypoverty) ~ sample.rural\_popdata$popdensity +

sample.rural\_popdata$popblack + sample.rural\_popdata$popasian +

sample.rural\_popdata$popadult + sample.rural\_popdata$popcollege +

sample.rural\_popdata$popchildpoverty)

Residuals:

Min 1Q Median 3Q Max

-0.46584 -0.10031 -0.00369 0.09723 0.50924

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.670e+00 4.568e-02 58.455 < 2e-16 \*\*\*

sample.rural\_popdata$popdensity -3.834e-03 1.199e-03 -3.199 0.002328 \*\*

sample.rural\_popdata$popblack 2.468e-05 2.359e-05 1.046 0.300103

sample.rural\_popdata$popasian 6.206e-04 2.882e-04 2.153 0.035881 \*

sample.rural\_popdata$popadult 3.590e-06 8.976e-06 0.400 0.690812

sample.rural\_popdata$popcollege -1.098e-04 3.099e-05 -3.542 0.000837 \*\*\*

sample.rural\_popdata$popchildpoverty 2.272e-04 4.173e-05 5.444 1.37e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.173 on 53 degrees of freedom

Multiple R-squared: 0.6186, Adjusted R-squared: 0.5754

F-statistic: 14.32 on 6 and 53 DF, p-value: 1.268e-09

**Interpretation:The model m3 is the best fit model, as its R square is 0.61 (highest among all the three) i.e this model explains 61% of variance of X variables on Y.According to this model, the intercept,popdensity,popasian,popcollege and popchildpoverty are significant factors in determining the perelderlypoverty.**

**When all the x variables are zero, the value of percentage elderly poverty is 2.670e+00 % .**

**When popdensity increases by 1 square unit, the percentage elderly poverty decreases by 3.834e-03% .**

**When population of asian increases by 1, the percentage elderly poverty decreases by 6.206e-04 %.**

**When population having college degree increases by 1, the percentage elderly poverty decreases by 1.098e-04% .**

**When population of children poverty increases by 1, the percentage elderly poverty increases by 2.272e-04 % .**

2.

#Checking if LINE asumptions are satistified for model m3(best fit model)

durbinWatsonTest(m3)

lag Autocorrelation D-W Statistic p-value

1 -0.2826734 2.558951 0.024

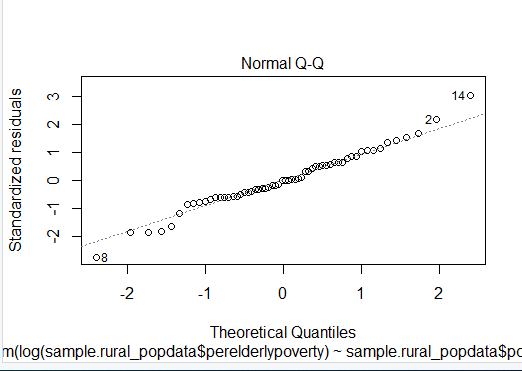
Alternative hypothesis: rho != 0

**The data has a negative autocorrelation of 0.28. Hence the data is not independent.**

plot(m3)

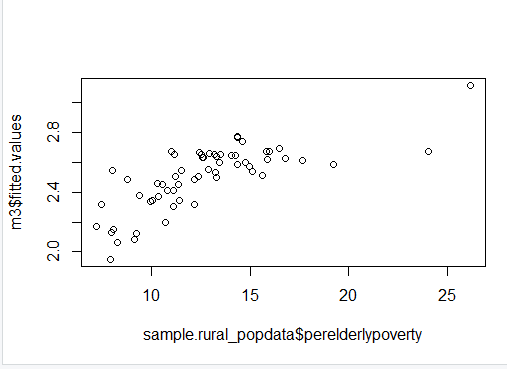
qqnorm(sample.rural\_popdata$perelderlypoverty)

qqline(sample.rural\_popdata$perelderlypoverty,col="red")



**As seen from the above graph, the data is near to normal for model m3.The data of the data set was near to normal, however after applying log transformation, the data became more near to normal and I got a better fit model.**

plot(sample.rural\_popdata$perelderlypoverty,m3$fitted.values )

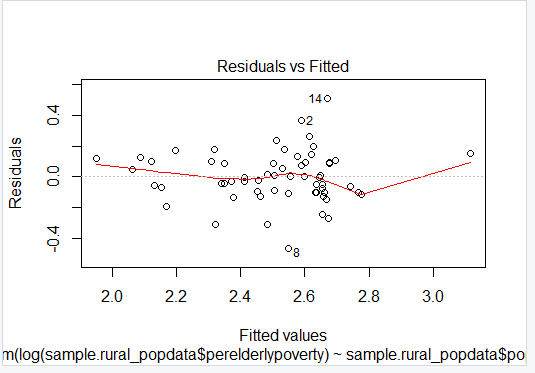


**The data is a non linear data as seen from above graph.**

plot(m3)

plot(m3$residuals,m3$fitted.values )

abline(0,0,lwd=3,col="Red")

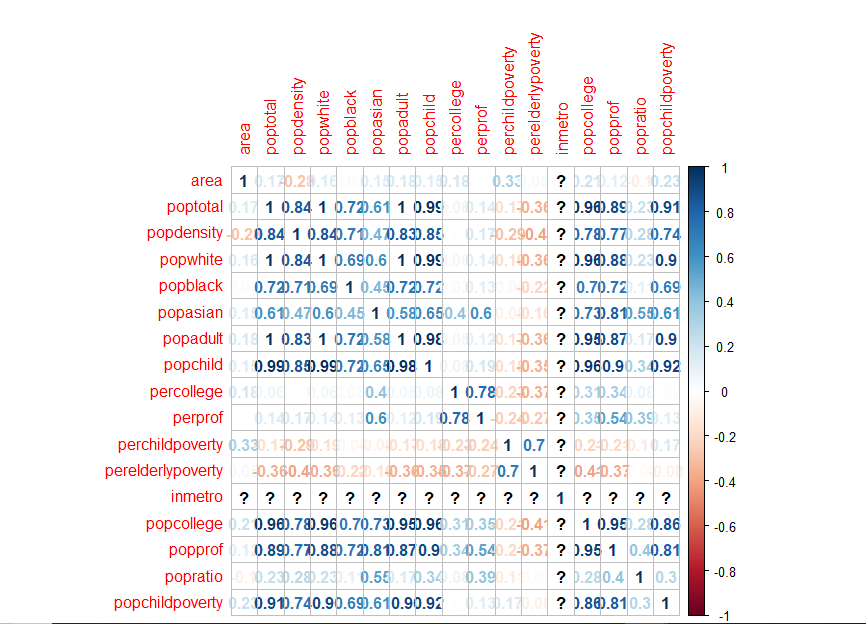


**As seen from the above graph, there is a pattern in the data, hence this model voilates the equality of variance assumption.**

3.

**Yes, as seen in answer 1, there is multicollinearity in the model. Multicollinearity is present when the predictor variables are correlated with each other.**

**As you can see from the below correlation plot some of the points with multi collinearity are,**



**The variables:**

1. **Poptotal and popdensity are highly correlated.**
2. **poptotal and popchild are almost correlated.**
3. **poptotal and popcollege are highly correlated.**
4. **poptotal and popprof are highly correlated.**
5. **poptotal and popchildpoverty are highly correlated.**
6. **Popdensity and popwhite are highly correlated**
7. **Popdensity and popchild are highly correlated**
8. **Popdensity and popadult are highly correlated**
9. **Poptotal and popwhite are highly correlated**
10. **Popwhite and popchild are highly correlated**
11. **Popwhite and popadult are highly correlated**
12. **Popasian and popprof are highly correlated**

# Variance Inflation Factors

Vif(m3)

sample.rural\_popdata$popdensity sample.rural\_popdata$popblack

3.627945 2.350736

sample.rural\_popdata$popasian sample.rural\_popdata$popadult

3.661851 30.217827

sample.rural\_popdata$popcollege sample.rural\_popdata$popchildpoverty

26.029803 6.094712

**Popcollege and popadult have high multicollinearity in the data as VIF > 10.**

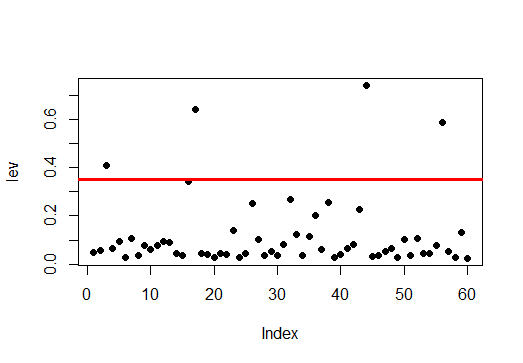
4.

#Finding the leverages on m3

lev=hat(model.matrix(m3))

plot(lev,pch=19)

abline(3\*mean(lev),0,col="red",lwd=3)



As seen from the above graph, there are 4 leverage points.

#Viewing the counties and states with leverages

sample.rural\_popdata[lev>(3\*mean(lev)),2:3]

# A tibble: 4 x 2

county state

*<chr>* *<chr>*

1 MUSKINGUM OH

2 LA SALLE IL

3 LA PORTE IN

4 MCDONOUGH IL

5.

#Applying the model m3 on sample.metropolitian\_popdata

m4<-lm(log(sample.metropolitian\_popdata$perelderlypoverty)~sample.metropolitian\_popdata$popdensity+sample.metropolitian\_popdata$popblack+sample.metropolitian\_popdata$popasian+sample.metropolitian\_popdata$popadult+sample.metropolitian\_popdata$popcollege+sample.metropolitian\_popdata$popchildpoverty)

summary(m4)

Call:

lm(formula = log(sample.metropolitian\_popdata$perelderlypoverty) ~

sample.metropolitian\_popdata$popdensity + sample.metropolitian\_popdata$popblack +

sample.metropolitian\_popdata$popasian + sample.metropolitian\_popdata$popadult +

sample.metropolitian\_popdata$popcollege + sample.metropolitian\_popdata$popchildpoverty)

Residuals:

Min 1Q Median 3Q Max

-0.26166 -0.12924 -0.00146 0.11366 0.37871

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.328e+00 7.363e-02 31.620 <2e-16 \*\*\*

sample.metropolitian\_popdata$popdensity -5.876e-04 2.832e-04 -2.075 0.0494 \*

sample.metropolitian\_popdata$popblack 1.833e-06 3.341e-06 0.549 0.5885

sample.metropolitian\_popdata$popasian -2.954e-05 2.478e-05 -1.192 0.2454

sample.metropolitian\_popdata$popadult 3.882e-07 2.585e-06 0.150 0.8820

sample.metropolitian\_popdata$popcollege -8.004e-06 5.611e-06 -1.426 0.1672

sample.metropolitian\_popdata$popchildpoverty 4.095e-05 1.680e-05 2.437 0.0229 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1797 on 23 degrees of freedom

Multiple R-squared: 0.65, Adjusted R-squared: 0.5588

F-statistic: 7.12 on 6 and 23 DF, p-value: 0.0002216

**Comparing the two models m3 and m4:**

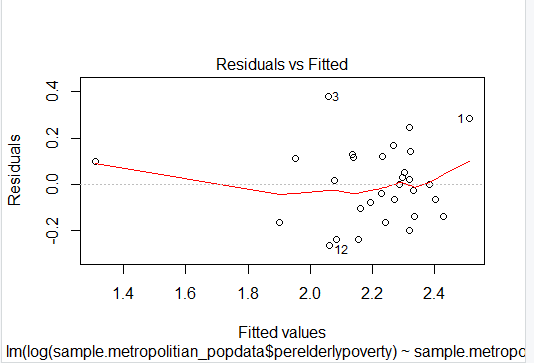
**1.Looking at the R square m4 is a slightly better fitting model than m3.**

**2.In m3, more number of variables are significant which is better than m4.**

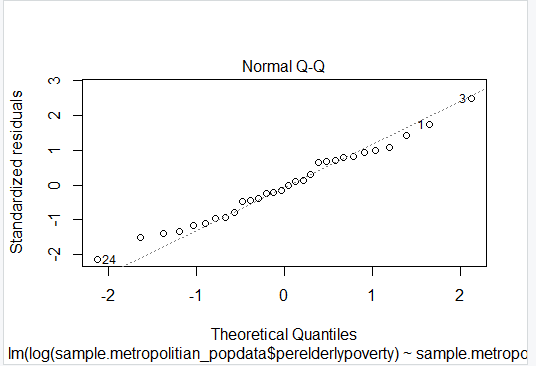
**3.Also, m3 has more degrees of freedom than m4.**

**Hence, m3 is better model than m4.**

#Checking the LINE assumptions of m4 model



We can see from above graph that this data satisfies the equality of variance.



This data satisfies normality assumption, as seen from the above graph.

durbinWatsonTest(m4)

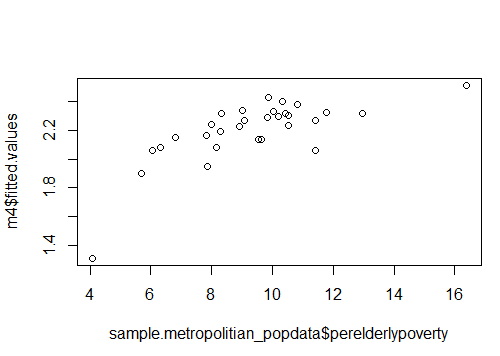
lag Autocorrelation D-W Statistic p-value

1 0.1860766 1.493217 0.108

Alternative hypothesis: rho != 0

The data has autocorrelation of 0.186, hence the data is not independent.

plot(sample.metropolitian\_popdata$perelderlypoverty,m4$fitted.values )



The data is not linear.