**Regression Project**

**Arpit Srivastava**

**Preprocessing**

1. **Load the file “6304 Regression Project Data.xlsx” into R. This file contains information on the population and other factors of 437 counties in the American Midwest. This is your full data set. Variable names are self-explanatory with those beginning with a “pop” prefix being numbers of population and those with a “per” prefix being percentages of the total population.**

rm(list=ls())  
library(readxl)  
library(car)

## Warning: package 'car' was built under R version 3.5.3

## Loading required package: carData

library(corrplot)

## corrplot 0.84 loaded

midwest=read\_excel("6304 Regression Project Data.xlsx",sheet="midwest")  
colnames(midwest) = tolower(make.names(colnames(midwest)))

1. **Using the “poptotal” variable in combination with “percollege” and “perprof” calculate new variables “popcollege” and “popprof”. These of course are the population in each county with a college degree, and the population with a professional job. Add these variables to the data frame.**

midwest$popcollegedegree=midwest$poptotal\*(midwest$percollege/100)  
midwest$popprofessionaljob=midwest$poptotal\*(midwest$perprof/100)

1. **Using the “popchild” and “popadult” variables calculate a new variable which will be the ratio of children to adults in each county’s population. Add this variable to the data frame.**

midwest$childtoadult=midwest$popchild/midwest$popadult

1. **Using the “popchild” and “perchildpoverty” variables calculate a new variable which will be the number of children living in poverty in each county. Add this variable to the data frame.**

midwest$popchildpoverty=(midwest$popchild\*(midwest$perchildpoverty/100))  
midwest$popasian2=midwest$popasian^2  
midwest$popadult2=midwest$popadult^2  
midwest$childtoadult2=midwest$childtoadult^2

1. **Subdivide the full data set to create two smaller data frames which include only rural and metropolitan counties, respectively. Use the “inmetro” variable for this.**

midwest.metro=subset(midwest, inmetro==1,)  
midwest.rural=subset(midwest, inmetro==0,)

1. **Using the numerical portion of your U number as a random number seed and the random selection method presented in class, take a random sample of 60 counties from the rural poverty data set.**
2. **Using the numerical portion of your U number as a random number seed and the random selection method presented in class, take a random sample of 30 counties from the metro poverty data set.**

set.seed(13660244)  
midwest.rural.sample=midwest.rural[sample(1:nrow(midwest.rural),60,replace=FALSE),]  
midwest.metro.sample=midwest.metro[sample(1:nrow(midwest.metro),30,replace=FALSE),]  
attach(midwest.rural.sample)  
attach(midwest.metro.sample)

## The following objects are masked from midwest.rural.sample:  
##   
## area, childtoadult, childtoadult2, county, id, inmetro,  
## perchildpoverty, percollege, perelderlypoverty, perprof,  
## popadult, popadult2, popasian, popasian2, popblack, popchild,  
## popchildpoverty, popcollegedegree, popdensity,  
## popprofessionaljob, poptotal, popwhite, state

**Analysis**

1. **Using the “perelderlypoverty” as the dependent variable apply any or all of the remaining numerical variables (except “id”) to parameterize the best possible fit multiple regression model. Use the “some.rural.poverty” data frame for this and apply only main-effects variables. Where needed feel free to apply any data transforms to improve this fit. Show the results of this best fit model using the summary(df.out) command. Describe the methodology you used to arrive at the selection of independent variables you used in your model.**

ruralreg.out=lm(perelderlypoverty ~popdensity+percollege+perchildpoverty+childtoadult+perprof+childtoadult2,data = midwest.rural.sample)  
summary(ruralreg.out)

##   
## Call:  
## lm(formula = perelderlypoverty ~ popdensity + percollege + perchildpoverty +   
## childtoadult + perprof + childtoadult2, data = midwest.rural.sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.1473 -1.4310 -0.0084 1.1060 7.4458   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.61207 11.14365 -1.042 0.30213   
## popdensity -0.02916 0.01106 -2.638 0.01094 \*   
## percollege -0.37576 0.13464 -2.791 0.00729 \*\*   
## perchildpoverty 0.24259 0.05179 4.684 2e-05 \*\*\*  
## childtoadult 81.62896 39.58695 2.062 0.04412 \*   
## perprof 1.12965 0.55418 2.038 0.04651 \*   
## childtoadult2 -68.97523 36.83374 -1.873 0.06664 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.382 on 53 degrees of freedom  
## Multiple R-squared: 0.523, Adjusted R-squared: 0.469   
## F-statistic: 9.685 on 6 and 53 DF, p-value: 3.467e-07

The model uses perelderlypoverty as the dependent variable where finding out the independent variables was the biggest challenge.

a) To get the perfect fit, I went through the sample and chose to act on instinct first choosing parameters of education and poverty to find out what variables had a pattern with the perelderlypoverty. The results were close but the significance of variables and the multiple R-squared values were not satisfactory.

b)The next step was to look at the scatterplots of every variable against the dependent variable individually to get an overview of which factors could be used in the final multiple regression model.

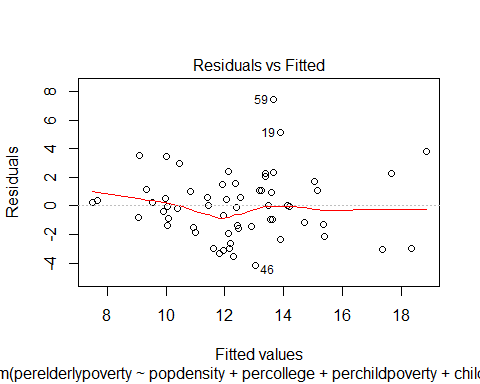
c)After using the shortlisted variables, I created the model and experimented with other variables to get the best fit.

d)after reaching a perticular fit, I realised that the perelderlypoverty seemed to have a non-linear realtionship with childtoadult. So, I added childtoadult2 which contains sqaured values of childtoadult.

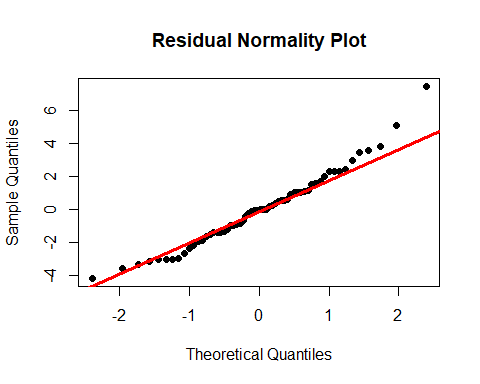
There were some options in which the multiple r squared value was slightly higher than the chosen best fit model, but the p values were ridiculouly high, which,to me, was not a good option.

1. **Assess your best fit model’s conformity to the LINE assumptions of regression. State your conclusions and show appropriate graphs and/or analytical output to support those conclusions.**

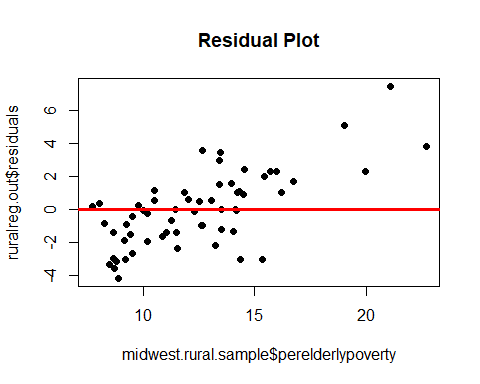
#residuals vs fitted  
plot(ruralreg.out,1)



#Normality  
qqnorm(ruralreg.out$residuals,pch=19,main = "Residual Normality Plot")  
qqline(ruralreg.out$residuals,col="red",lwd=3)



#Equality of variances  
plot(midwest.rural.sample$perelderlypoverty,ruralreg.out$residuals,pch=19,main="Residual Plot")  
abline(0,0,lwd=3,col="red")

 As it can be seen that there is no pattern in the residual plot, it is safe to assume that the model is not in violation of Linearity.

Since it is not a time series problem, Violation of independence is not an issue.

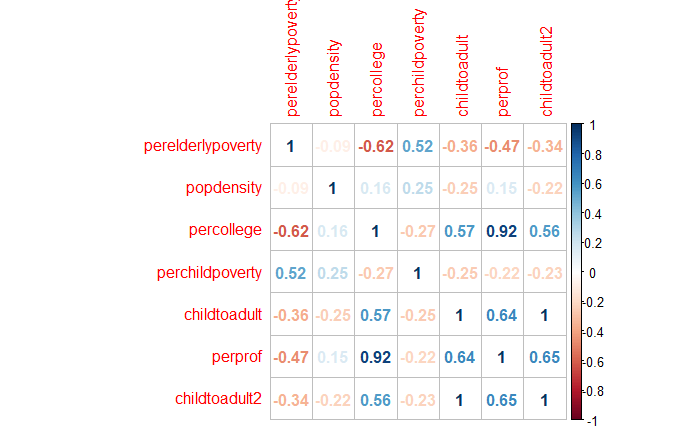
Other than a few outliers, it is pretty close when it comes to Normality, as it can be seen in the Residual Normality Plot.

Barring 2 or 3 outliers,the model does not seem to be in violation of equality of variances as it can be seen in the Residual Plot.

1. **Determine whether you believe multicollinearity exists in your best fit model. State your conclusions and show appropriate graphs and/or analytical output to support those conclusions.**

some.of.rural=subset(midwest.rural.sample,select=c( "perelderlypoverty","popdensity","percollege","perchildpoverty","childtoadult","perprof","childtoadult2"))

corrplot::corrplot(xx,method="number")



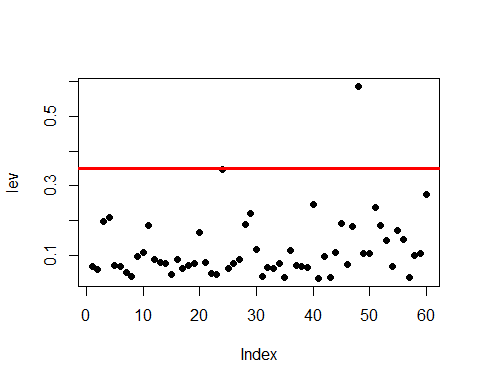
vif(ruralreg.out)

## popdensity percollege perchildpoverty childtoadult   
## 1.424276 3.315205 1.401154 91.256646   
## perprof childtoadult2   
## 3.731812 90.528715

The correlation plot shows that there is high correlation between two variable i.e childtoadult and childtoadult2, which was expected because childtoadult2 is squared value of childtoadult. Other than that, there is high correlation between perprof and percollege. But this is not an exact measure of multicollinearity because it only considers pairwise effect. To get a better understanding, VIF has been used. So, it is safe to assume that multicollinearity exists in the model because of childtoadult and childtoadult2 but it can be ignored because it is cause by a powered value.

1. **Determine if any of the counties in your “some.rural.poverty” data set have an outsized leverage in influencing your best fit model. If so, state which counties (county name and state) have this outsized influence.**

lev=hat(model.matrix(ruralreg.out))  
plot(lev,pch=19)  
abline(3\*mean(lev),0,col="red",lwd=3)



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

## [1] "PORTAGE"

midwest.rural.sample[lev>(3\*mean(lev)),3,4]

## [1] "WI"

Portage county in Wisconsin has an outsized influence

1. **Assess how well your best fit model predicts “perelderlypoverty” when applied to the “some.metro.poverty” data frame. Tell whether you believe the fit is better or worse than when the model is used with the “some.rural.poverty” data. Show appropriate graphs and/or analytical output to support your conclusions.**

metroreg.out=lm(perelderlypoverty ~popdensity+percollege+perchildpoverty+childtoadult+perprof+childtoadult2,data = midwest.metro.sample)  
summary(metroreg.out)

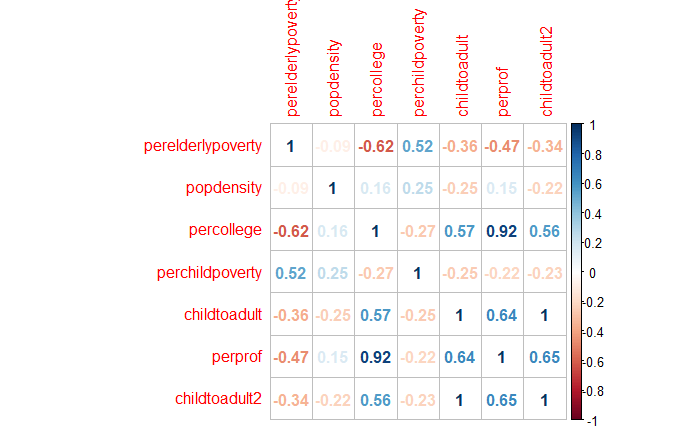
##   
## Call:  
## lm(formula = perelderlypoverty ~ popdensity + percollege + perchildpoverty +   
## childtoadult + perprof + childtoadult2, data = midwest.metro.sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8874 -0.9009 -0.0408 0.7177 5.5382   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.7944605 15.7709373 0.684 0.50053   
## popdensity -0.0006964 0.0006920 -1.006 0.32473   
## percollege -0.3937476 0.1327649 -2.966 0.00692 \*\*  
## perchildpoverty 0.1760274 0.0638130 2.758 0.01119 \*   
## childtoadult 6.7116258 48.8212164 0.137 0.89185   
## perprof 0.7477504 0.3807175 1.964 0.06172 .   
## childtoadult2 -8.9373230 36.9020727 -0.242 0.81078   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.81 on 23 degrees of freedom  
## Multiple R-squared: 0.6033, Adjusted R-squared: 0.4999   
## F-statistic: 5.831 on 6 and 23 DF, p-value: 0.0008213

#multicolinearity

*some.of.metro=subset(midwest.metro.sample,select=c( "perelderlypoverty","popdensity","percollege","perchildpoverty","childtoadult","perprof","childtoadult2"))*

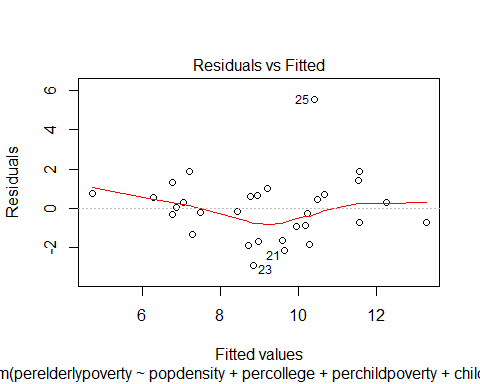
*corrplot::corrplot(xx,method="number")*

vif(metroreg.out)

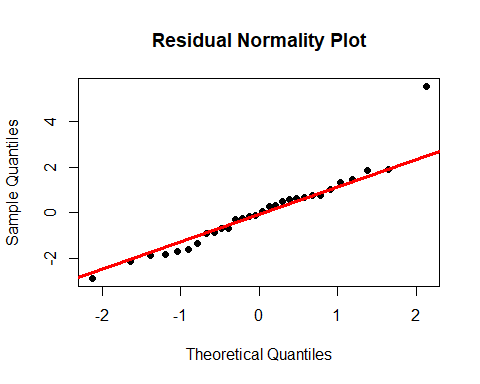


## popdensity percollege perchildpoverty childtoadult   
## 1.455084 8.384496 1.255085 152.025920   
## perprof childtoadult2   
## 9.521198 150.682699

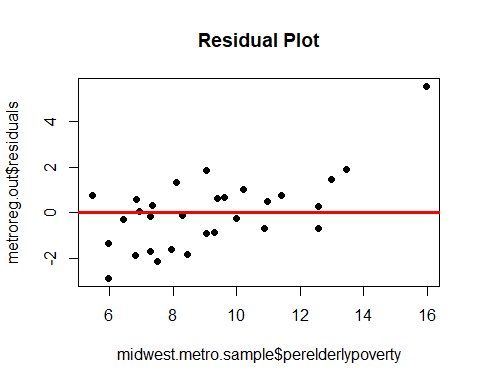
#residuals vs fitted  
plot(metroreg.out,1)



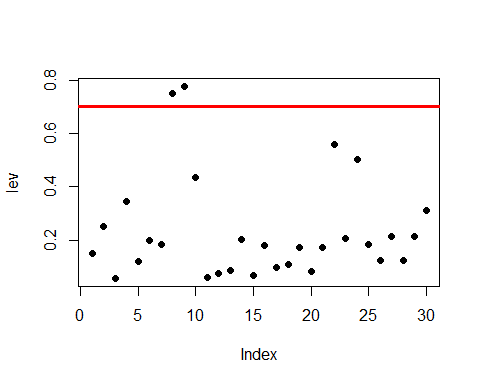
#Normality  
qqnorm(metroreg.out$residuals,pch=19,main = "Residual Normality Plot")  
qqline(metroreg.out$residuals,col="red",lwd=3)



#Equality of variances  
plot(midwest.metro.sample$perelderlypoverty,metroreg.out$residuals,pch=19,main="Residual Plot")  
abline(0,0,lwd=3,col="red")



lev=hat(model.matrix(metroreg.out))  
plot(lev,pch=19)  
abline(3\*mean(lev),0,col="red",lwd=3)



The model has a better r-squared value but it also has less number of significant variables, where the p-values are ridiculously high. This is not exactly a deal breaker but can be an indication an insufficient number of observations .

The multicolliearity in this case tells the same story which also seems to be true for LINE assumptions and leverage.

In my opinion, althought the multiple R-squared value suggests that this model is a better fit, the story is different. The significance of variables puts a big question mark on the superiority of this model.