Using Linear Regression to predict Poverty among elderly people in the American Midwest

The dataset contains 16 columns with information on the population and other factors of 437 counties in the American Midwest. 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. I am including new variables called "popcollege" and "popprof". These are the population in each county with a college degree, and the population with a professional job. Using the "popchild" and "popadult" variables I am calculating a new variable "ratioca" which will be the ratio of children to adults in each county's population. Using the "inmetro" variable, I am subdividing the full data set to create two smaller data frames which include only rural and metropolitan counties, respectively. Finally, using a random seed I am taking a random sample of 60 counties from the rural poverty dataset and 30 counties from the metro poverty dataset.

Set the working directory and loading the data onto R for analysis.

### pre-processing ######  
#install.packages("data.table")  
library(corrplot)

## corrplot 0.84 loaded

setwd("D:/USF/Semester 2/Qmb")  
library(readxl)  
popdata=read\_excel("6304 Regression Project Data.xlsx")  
popdata$popcollege = (popdata$percollege\*popdata$poptotal)/100  
popdata$popprof = (popdata$perprof\*popdata$poptotal)/100  
popdata$ratioca = (popdata$popchild/popdata$popadult)  
popdata$popchildpoverty = (popdata$perchildpoverty\*popdata$popchild)/100  
##subset  
metro = subset(popdata, inmetro == 1)  
rural = subset(popdata, inmetro == 0)  
##Random seed  
set.seed(93827161)  
some.rural.poverty = rural[sample(1:nrow(rural),60,replace=FALSE),]  
some.metro.poverty = metro[sample(1:nrow(metro),30,replace=FALSE),]

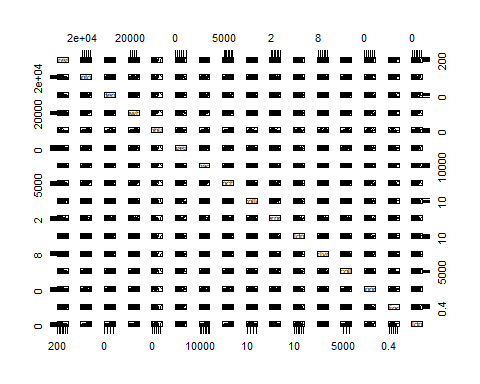
Lets view the data to understand what it looks like

head(popdata)

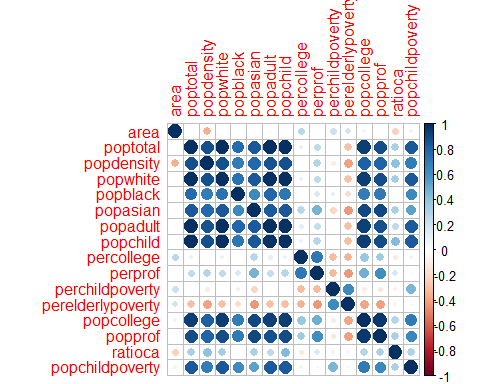
## # A tibble: 6 x 20  
## ID county state area poptotal popdensity popwhite popblack popasian  
## <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 ADAMS IL 857. 66090 77.2 63917 1702 249  
## 2 2 ALEXA~ IL 236. 10626 45.0 7054 3496 48  
## 3 3 BOND IL 380. 14991 39.4 14477 429 16  
## 4 4 BOONE IL 281. 30806 110. 29344 127 150  
## 5 5 BROWN IL 306. 5836 19.1 5264 547 5  
## 6 6 BUREAU IL 869. 35688 41.1 35157 50 195  
## # ... with 11 more variables: popadult <dbl>, popchild <dbl>, percollege <dbl>,  
## # perprof <dbl>, perchildpoverty <dbl>, perelderlypoverty <dbl>,  
## # inmetro <dbl>, popcollege <dbl>, popprof <dbl>, ratioca <dbl>,  
## # popchildpoverty <dbl>

## Analyzing the data, plotting to check correlation between variables with a correlation matrix

continuous.rural.poverty = subset(some.rural.poverty,select=c("area","poptotal","popdensity","popwhite","popblack","popasian","popadult","popchild","percollege","perprof","perchildpoverty","perelderlypoverty","popcollege","popprof","ratioca","popchildpoverty"))  
plot(continuous.rural.poverty)



xx=cor(continuous.rural.poverty)  
corrplot(xx,method="circle")



We can see that correlation exits between poptotal and popchild, poptotal and popdensity, popcollege, popchildpoverty and so on.

regout1=lm(perelderlypoverty~.-ID-county-state,data=some.rural.poverty)  
summary(regout1)

##   
## Call:  
## lm(formula = perelderlypoverty ~ . - ID - county - state, data = some.rural.poverty)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6834 -0.7881 -0.0655 0.9936 3.8881   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.0787348 6.4159153 0.480 0.6337   
## area 0.0017233 0.0016083 1.072 0.2896   
## poptotal -0.0021941 0.0010328 -2.124 0.0392 \*  
## popdensity -0.0097412 0.0213050 -0.457 0.6497   
## popwhite 0.0013139 0.0009105 1.443 0.1559   
## popblack 0.0012119 0.0010076 1.203 0.2353   
## popasian 0.0036146 0.0064161 0.563 0.5760   
## popadult 0.0012014 0.0009237 1.301 0.2000   
## popchild NA NA NA NA   
## percollege -0.0699152 0.2859046 -0.245 0.8079   
## perprof -0.3893469 0.8438411 -0.461 0.6467   
## perchildpoverty 0.1795608 0.0923500 1.944 0.0581 .  
## inmetro NA NA NA NA   
## popcollege 0.0002372 0.0011014 0.215 0.8305   
## popprof 0.0003850 0.0027135 0.142 0.8878   
## ratioca 16.8244793 10.0155876 1.680 0.0999 .  
## popchildpoverty 0.0004816 0.0006185 0.779 0.4403   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.736 on 45 degrees of freedom  
## Multiple R-squared: 0.605, Adjusted R-squared: 0.4821   
## F-statistic: 4.922 on 14 and 45 DF, p-value: 2.079e-05

regout2=lm(perelderlypoverty~perchildpoverty+popchildpoverty+area+poptotal+ratioca,data=some.rural.poverty)  
summary(regout2)

##   
## Call:  
## lm(formula = perelderlypoverty ~ perchildpoverty + popchildpoverty +   
## area + poptotal + ratioca, data = some.rural.poverty)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.0314 -1.0023 -0.1732 1.3074 4.0385   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.085e+00 2.442e+00 2.902 0.00536 \*\*  
## perchildpoverty 2.313e-01 7.469e-02 3.097 0.00310 \*\*  
## popchildpoverty 2.854e-04 4.259e-04 0.670 0.50566   
## area 1.138e-03 9.449e-04 1.204 0.23379   
## poptotal -6.175e-05 3.159e-05 -1.954 0.05584 .   
## ratioca 3.481e+00 3.103e+00 1.122 0.26692   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.744 on 54 degrees of freedom  
## Multiple R-squared: 0.5212, Adjusted R-squared: 0.4769   
## F-statistic: 11.76 on 5 and 54 DF, p-value: 1.031e-07

The first model (regout1) is a kitchen sink model, with almost all the variables included in it. Here only the popchildpoverty was significant having a p-value of 0.0340 The second model (regout2), has only a few variables and gives better results than model one, with a DF of 54, and two significant variables.

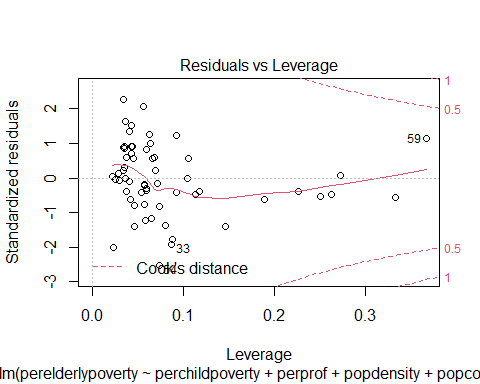
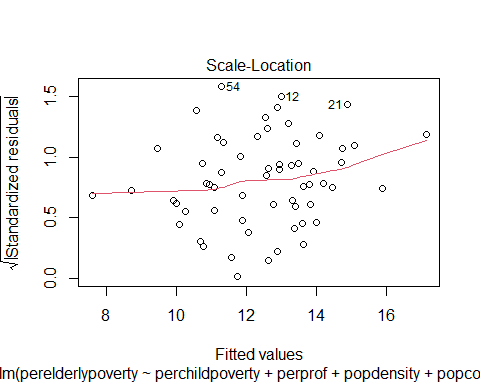
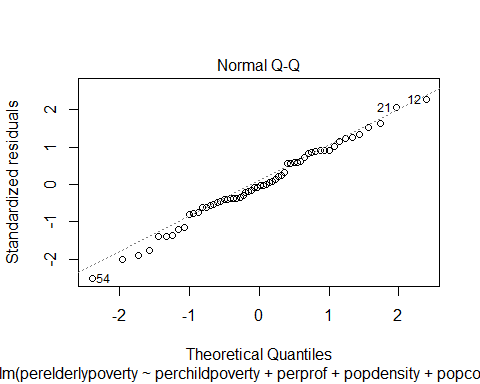
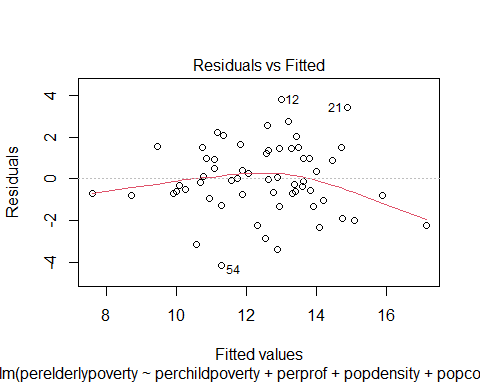
After adding and removing a few independent variables, I came up with the final model which outperformed the previous two, with an increased DF and better p-value, along with a good adusted R-squared score.

regout3=lm(perelderlypoverty~perchildpoverty+perprof+popdensity+popcollege,data=some.rural.poverty)  
summary(regout3)

##   
## Call:  
## lm(formula = perelderlypoverty ~ perchildpoverty + perprof +   
## popdensity + popcollege, data = some.rural.poverty)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.1140 -0.7701 -0.0915 1.2380 3.7887   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.052e+01 1.316e+00 7.988 9.12e-11 \*\*\*  
## perchildpoverty 2.388e-01 4.247e-02 5.623 6.49e-07 \*\*\*  
## perprof -3.376e-01 2.547e-01 -1.325 0.191   
## popdensity -1.500e-02 1.036e-02 -1.448 0.153   
## popcollege -5.364e-05 1.146e-04 -0.468 0.641   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.7 on 55 degrees of freedom  
## Multiple R-squared: 0.537, Adjusted R-squared: 0.5033   
## F-statistic: 15.95 on 4 and 55 DF, p-value: 1.002e-08

Potting the final regression equation to check its conformity to the LINE assumptions of regression

plot(regout3)



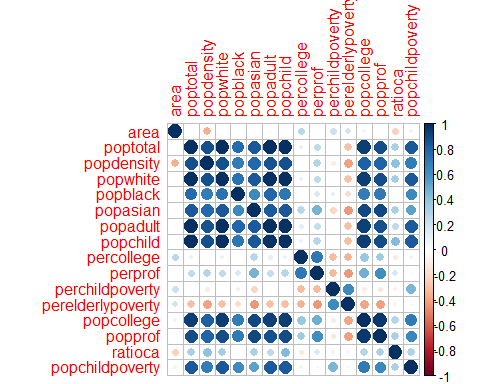
From the Residuals vs Fitted plot we can see that the model violates linearity assumption.

From the qqplot we can assume that the model is normal.

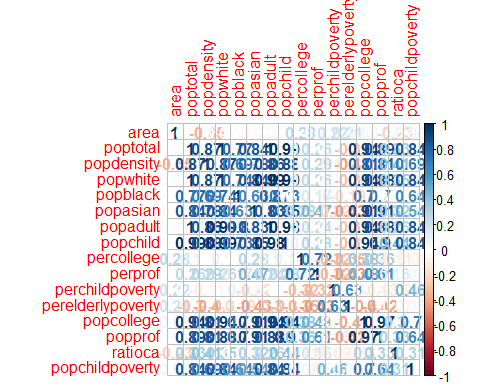
Also, from the Scale-Location plot, we can see that we have leverage points 13 and 46

From the Residuals vs Leverage plot we can see that the model is heteroscadastic

#putting a correlation matrix into object  
#install.packages('corrplot')  
#library(corrplot)  
xx=cor(continuous.rural.poverty)  
corrplot(xx,method="circle")

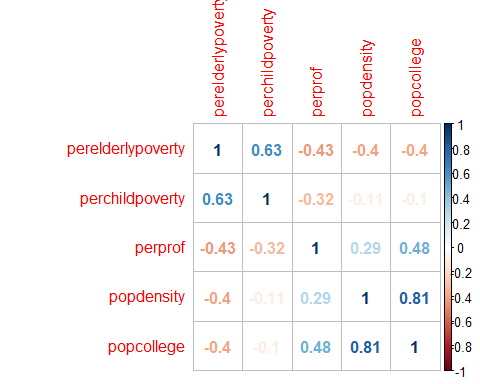


corrplot(xx,method="number")



Subsetting few columns to find correlation

d1=subset(some.rural.poverty,select=c("perelderlypoverty","perchildpoverty","perprof","popdensity","popcollege","popprof"))  
d2=subset(some.rural.poverty,select=c("perelderlypoverty","perchildpoverty","perprof","popdensity","popcollege"))  
  
yy=cor(d2)  
corrplot(yy,method="number")



The above plot was used to check correlation between the subsetted variables.

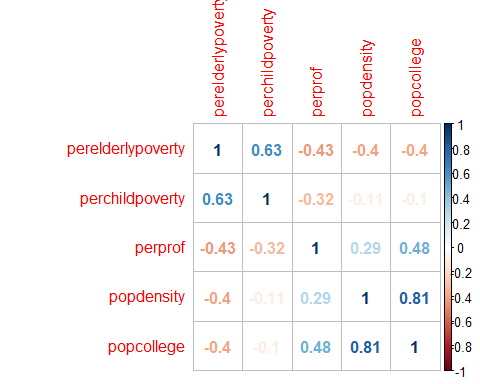
#Correlation matrix with p values.  
#install.packages('Hmisc')  
#library(Hmisc)  
#xx=rcorr(as.matrix(d2))  
#xx

### Using Variation Inflation Factors (VIF) to check if multicollinearity exits in the best fit model.

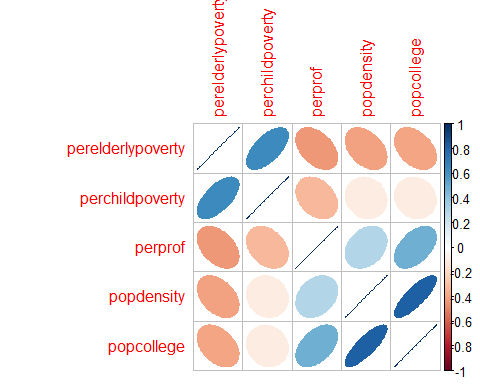
Multicollinearity occurs when independent variables in a regression model are correlated.

plot(d2)

xx=cor(d2)  
corrplot(xx,method="number")



corrplot(xx,method="ellipse")



library(car)

## Loading required package: carData

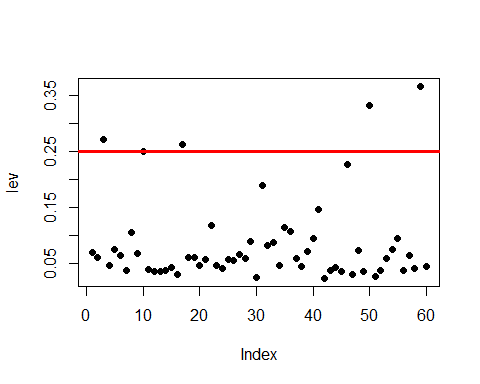
vif(regout3)

## perchildpoverty perprof popdensity popcollege   
## 1.131683 1.510722 3.066577 3.664088

Since the VIF values for all the variables are close to 1 and less than 5, we can conclude that there is a moderate correlation, but it is not severe enough to warrant corrective measures

### Determining if any of the counties in "some.rural.poverty" data set have an outsized leverage in influencing the best fit model.

#boxplot(some.rural.poverty$perelderlypoverty)  
#max(some.rural.poverty$perelderlypoverty)  
#which(some.rural.poverty$perelderlypoverty==18.27736)  
  
#Leverage of Points  
lev=hat(model.matrix(regout3))  
plot(lev,pch=19)  
abline(3\*mean(lev),0,col="red",lwd=3)



some.rural.poverty[lev>(3\*mean(lev)),]

## # A tibble: 5 x 20  
## ID county state area poptotal popdensity popwhite popblack popasian  
## <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 123 FAYET~ IN 215. 26015 121. 25462 435 69  
## 2 148 LA PO~ IN 598. 107066 179. 96286 9580 431  
## 3 430 WALWO~ WI 555. 75000 135. 72747 454 494  
## 4 350 SCIOTO OH 612. 80327 131. 77253 2458 126  
## 5 15 COLES IL 508. 51644 102. 50177 925 341  
## # ... with 11 more variables: popadult <dbl>, popchild <dbl>, percollege <dbl>,  
## # perprof <dbl>, perchildpoverty <dbl>, perelderlypoverty <dbl>,  
## # inmetro <dbl>, popcollege <dbl>, popprof <dbl>, ratioca <dbl>,  
## # popchildpoverty <dbl>

#Leverage of Points  
#lev=hat(model.matrix(regout3))  
#plot(lev,pch=19)  
#abline(3\*mean(lev),0,col="red",lwd=3)  
#some.rural.poverty[lev>(3\*mean(lev)),]

The counties and states above, have an outsized leverage in influencing my best fit model

### Assessing how well my best fit model predicts "perelderlypoverty" when applied to the "some.metro.poverty" data frame

predict(regout3,some.metro.poverty,interval="confidence")

## fit lwr upr  
## 1 5.5276971 2.843899 8.211496  
## 2 2.5301778 -3.134210 8.194565  
## 3 -6.4400343 -18.139419 5.259351  
## 4 5.0723315 2.206721 7.937942  
## 5 9.4421501 8.349735 10.534566  
## 6 13.0751976 11.821291 14.329104  
## 7 11.2688859 10.250740 12.287032  
## 8 7.5797333 3.262061 11.897405  
## 9 -0.0704781 -8.946807 8.805851  
## 10 2.2190998 -4.213918 8.652117  
## 11 8.4627646 6.007678 10.917851  
## 12 7.5331834 5.227290 9.839076  
## 13 8.5119103 5.809252 11.214569  
## 14 3.5743975 -2.885722 10.034517  
## 15 14.0557647 12.158812 15.952718  
## 16 7.4263372 5.255033 9.597642  
## 17 2.3765947 -2.479987 7.233177  
## 18 11.2362234 10.227957 12.244490  
## 19 5.8377902 1.593759 10.081821  
## 20 1.4592702 -10.400462 13.319003  
## 21 8.3502867 4.396148 12.304425  
## 22 -12.8310071 -29.201390 3.539376  
## 23 11.6574723 10.936004 12.378940  
## 24 -0.5666199 -12.012639 10.879399  
## 25 -0.8397223 -14.184996 12.505552  
## 26 -57.5654415 -107.075946 -8.054937  
## 27 8.3543422 5.825888 10.882796  
## 28 8.5353169 4.959909 12.110725  
## 29 9.7476610 8.469909 11.025413  
## 30 -12.5948874 -28.543238 3.353463

The percentage of elderly poverty is predicted to be 2.98% in Floyd, IN based on the independent variables used in regout3 (final regression equation). And, we are 95% confident that the percentage of perelderlypoverty would fall between -5.6227978 - 11.592168 %. In a similar way rest of the 29 data entries can be interpreted.