**Regression Project**

**QMB-6304 Analytical Methods for Business**

Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

**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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.

**Answer:**

##Shravya Katukuri

#6304 Regression Project

setwd("E:/Sem II/AMB/Project")

rm(list=ls())

library(readxl)

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

colnames(midwest) = tolower(make.names(colnames(midwest)))

attach(midwest)

midwest$popcollege=poptotal\*percollege/100

midwest$popcollege=round(midwest$popcollege, digits = 0)

midwest$popprof=poptotal\*perprof/100

midwest$popprof=round(midwest$popprof, digits = 0)

midwest$ratio=popchild/popadult

midwest$popchildpoverty=popchild\*perchildpoverty/100

midwest$popchildpoverty=round(midwest$popchildpoverty, digits = 0)

rural=midwest[which(inmetro==0),]

metropolitan=midwest[which(inmetro==1),]

set.seed(59033316)

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

set.seed(59033316)

random.metropolitan=metropolitan[sample(1:nrow(metropolitan),30,replace=FALSE),]

**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.

**Answer:**

**Code:**

perelderlypoverty.out=lm(perelderlypoverty~. -id -county -state -inmetro, data=random.rural)

##Using kitchen sink for only numerical variables; excluded the others

summary(perelderlypoverty.out)

AIC(perelderlypoverty.out)

vif(perelderlypoverty.out)

**Output:**

summary(perelderlypoverty.out)

Call:

lm(formula = perelderlypoverty ~ . - id - county - state - inmetro,

data = random.rural)

Residuals:

Min 1Q Median 3Q Max

-3.2992 -1.6167 -0.0426 1.1878 3.8957

Coefficients: (1 not defined because of singularities)

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

(Intercept) -7.5947204 7.0688712 -1.074 0.288375

area 0.0035181 0.0021508 1.636 0.108877

poptotal -0.0051666 0.0013249 -3.900 0.000318 \*\*\*

popdensity 0.0220275 0.0343851 0.641 0.525023

popwhite 0.0028366 0.0009257 3.064 0.003680 \*\*

popblack 0.0026397 0.0009028 2.924 0.005392 \*\*

popasian 0.0057798 0.0082002 0.705 0.484543

popadult 0.0031273 0.0010797 2.896 0.005810 \*\*

popchild NA NA NA NA

percollege -0.5647050 0.1893334 -2.983 0.004603 \*\*

perprof 1.9904492 0.7808959 2.549 0.014286 \*

perchildpoverty 0.2457012 0.0779073 3.154 0.002870 \*\*

popcollege 0.0024937 0.0008036 3.103 0.003303 \*\*

popprof -0.0064629 0.0024904 -2.595 0.012722 \*

ratio 30.3286513 10.4204242 2.911 0.005593 \*\*

popchildpoverty 0.0019186 0.0009346 2.053 0.045931 \*

---

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

Residual standard error: 2.008 on 45 degrees of freedom

Multiple R-squared: 0.7517, Adjusted R-squared: 0.6745

F-statistic: 9.731 on 14 and 45 DF, p-value: 2.029e-09

> AIC(perelderlypoverty.out)

[1] 268.6553

> vif(perelderlypoverty.out)

Error in vif.default(perelderlypoverty.out) :

there are aliased coefficients in the model

Since this model is a kitchen sink of all numerical variables, it has high multicollinearity (as shown by vif command that gives aliased coefficients indicating perfect multicollinearity between variables). **Popchild** variable in the model also shows NA due to multicollinearity.

We now plot a stepwise regression model to determine which variables might have an influence over perelderlypoverty variable by comparing different models.

**Code:**

step(lm(perelderlypoverty~. -id -county -state -inmetro, data=random.rural), direction="both")

**Output:**

step(lm(perelderlypoverty~.-id -county -state -inmetro, data=random.rural),direction="both")

Start: AIC=96.38

perelderlypoverty ~ (id + county + state + area + poptotal +

popdensity + popwhite + popblack + popasian + popadult +

popchild + percollege + perprof + perchildpoverty + inmetro +

popcollege + popprof + ratio + popchildpoverty) - id - county -

state - inmetro

Step: AIC=96.38

perelderlypoverty ~ area + poptotal + popdensity + popwhite +

popblack + popasian + popadult + percollege + perprof + perchildpoverty +

popcollege + popprof + ratio + popchildpoverty

Df Sum of Sq RSS AIC

- popdensity 1 1.654 183.06 94.927

- popasian 1 2.003 183.41 95.041

<none> 181.40 96.383

- area 1 10.786 192.19 97.848

- popchildpoverty 1 16.988 198.39 99.754

- perprof 1 26.191 207.59 102.474

- popprof 1 27.148 208.55 102.751

- popadult 1 33.816 215.22 104.639

- ratio 1 34.148 215.55 104.731

- popblack 1 34.468 215.87 104.820

- percollege 1 35.861 217.26 105.206

- popwhite 1 37.852 219.25 105.754

- popcollege 1 38.822 220.22 106.018

- perchildpoverty 1 40.095 221.50 106.364

- poptotal 1 61.307 242.71 111.851

Step: AIC=94.93

perelderlypoverty ~ area + poptotal + popwhite + popblack + popasian +

popadult + percollege + perprof + perchildpoverty + popcollege +

popprof + ratio + popchildpoverty

Df Sum of Sq RSS AIC

- popasian 1 4.972 188.03 94.535

<none> 183.06 94.927

+ popdensity 1 1.654 181.40 96.383

- area 1 14.220 197.28 97.416

- popchildpoverty 1 17.401 200.46 98.376

- perprof 1 24.658 207.72 100.510

- popprof 1 26.162 209.22 100.942

- popblack 1 33.312 216.37 102.959

- percollege 1 34.246 217.30 103.217

- ratio 1 36.638 219.70 103.874

- popadult 1 37.584 220.64 104.132

- popwhite 1 37.743 220.80 104.175

- perchildpoverty 1 39.269 222.33 104.588

- popcollege 1 41.300 224.36 105.134

- poptotal 1 62.043 245.10 110.439

Step: AIC=94.54

perelderlypoverty ~ area + poptotal + popwhite + popblack + popadult +

percollege + perprof + perchildpoverty + popcollege + popprof +

ratio + popchildpoverty

Df Sum of Sq RSS AIC

<none> 188.03 94.535

+ popasian 1 4.972 183.06 94.927

+ popdensity 1 4.624 183.41 95.041

- area 1 11.726 199.76 96.165

- popchildpoverty 1 13.756 201.79 96.772

- perprof 1 21.711 209.74 99.092

- popprof 1 23.554 211.58 99.617

- popadult 1 32.615 220.64 102.132

- percollege 1 32.677 220.71 102.149

- popblack 1 32.686 220.72 102.152

- ratio 1 32.915 220.94 102.214

- popwhite 1 35.698 223.73 102.965

- perchildpoverty 1 44.650 232.68 105.319

- popcollege 1 46.793 234.82 105.869

- poptotal 1 57.137 245.17 108.456

Call:

lm(formula = perelderlypoverty ~ area + poptotal + popwhite +

popblack + popadult + percollege + perprof + perchildpoverty +

popcollege + popprof + ratio + popchildpoverty, data = random.rural)

Coefficients:

(Intercept) area poptotal popwhite popblack

-6.046804 0.002155 -0.004825 0.002745 0.002557

popadult percollege perprof perchildpoverty popcollege

0.002826 -0.521199 1.756354 0.255825 0.002358

popprof ratio popchildpoverty

-0.005487 29.018862 0.001662

From the above output, we can see thatthe last model has the lowest AIC of 94.54 i.e., best fit. Let’s call this model **perelderlypoverty.out.2.**

**Code:**

perelderlypoverty.out.2=lm(perelderlypoverty ~ area + poptotal + popwhite +

popblack + popadult + percollege + perprof + perchildpoverty +

popcollege + popprof + ratio + popchildpoverty, data = random.rural)

summary(perelderlypoverty.out.2)

**Output:**

|  |
| --- |
| summary(perelderlypoverty.out.2)  Call:  lm(formula = perelderlypoverty ~ area + poptotal + popwhite +  popblack + popadult + percollege + perprof + perchildpoverty +  popcollege + popprof + ratio + popchildpoverty, data = random.rural)  Residuals:  Min 1Q Median 3Q Max  -3.038 -1.495 0.127 1.084 3.969  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -6.0468039 6.9363041 -0.872 0.387770  area 0.0021553 0.0012589 1.712 0.093477 .  poptotal -0.0048247 0.0012767 -3.779 0.000443 \*\*\*  popwhite 0.0027446 0.0009188 2.987 0.004464 \*\*  popblack 0.0025571 0.0008946 2.858 0.006329 \*\*  popadult 0.0028264 0.0009899 2.855 0.006382 \*\*  percollege -0.5211988 0.1823677 -2.858 0.006335 \*\*  perprof 1.7563538 0.7539394 2.330 0.024176 \*  perchildpoverty 0.2558254 0.0765767 3.341 0.001644 \*\*  popcollege 0.0023582 0.0006895 3.420 0.001304 \*\*  popprof -0.0054871 0.0022614 -2.426 0.019137 \*  ratio 29.0188616 10.1169594 2.868 0.006161 \*\*  popchildpoverty 0.0016621 0.0008963 1.854 0.069971 .  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 2 on 47 degrees of freedom  Multiple R-squared: 0.7426, Adjusted R-squared: 0.6769  F-statistic: 11.3 on 12 and 47 DF, p-value: 3.689e-10  > vif(perelderlypoverty.out.2)  area poptotal popwhite popblack popadult  1.311974 11951.978995 5523.251124 26.411915 2853.912688  percollege perprof perchildpoverty popcollege popprof  5.204516 6.899025 4.969891 117.273043 89.430681  ratio popchildpoverty  7.009061 24.020921  > AIC(perelderlypoverty.out.2)  [1] 266.8079 |
| This model has the best fit with an R2 of around 68% (of all the possible models).  We draw scatterplots for all numerical independent variables and after checking these scatterplots, the ones that sort  of have a curvilinear relationship and could better be explained with a squared term are – percollege, popcollege and popprof.  Below are the scatterplots:  **Code: (In R)**  plot(random.rural$percollege, random.rural$perelderlypoverty, pch=19, xlab = "percollege",  ylab = "perelderlypoverty")  abline(lm(perelderlypoverty~percollege,data=random.rural),col="red",pch=19,lwd=3)  plot(random.rural$popcollege, random.rural$perelderlypoverty, pch=19, xlab = "popcollege", ylab = "perelderlypoverty")  abline(lm(perelderlypoverty~popcollege,data=random.rural),col="red",pch=19,lwd=3)  plot(random.rural$popprof, random.rural$perelderlypoverty, pch=19, xlab = "popprof", ylab = "perelderlypoverty")  abline(lm(perelderlypoverty~popprof,data=random.rural),col="red",pch=19,lwd=3)  **Output: (From Minitab)**  I took the plots from Minitab as they were clearer and were showing the change in R2 values too.  **Percollege:** Increase in R2 from 11.3% to 15.4%      **Popcollege:** Increase in R2 from 15.9% to 16.6%    **Popprof:** Increase in R2 from 15.5% to 16.8%    We compute 7 different models using combinations of the above squared terms and get the following table.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Terms added** | **R2 value (in %)** | **Number of terms with high vif (>10)** | **AIC** | **Main term** | **Squared term** | | percollege^2 | 67 | 9 | 268.7897 | non significant | non significant | | popcollege^2 | 67.37 | 8 | 268.1213 | significant | non significant | | popprof^2 | 66.99 | 9 | 268.8078 | non significant | non significant | | percollege+popcollege | 66.64 | 10 | 270.1204 | only popcollege is significant | non significant | | popcollege^2+popprof^2 | 68.86 | 10 | 265.99 | both are significant | non significant | | percollege^2+popprof^2 | 66.27 | 11 | 270.7897 | both are non significant | non significant | | percollege^2+popcollege^2+popprof^2 | 68.16 | 12 | 267.9695 | popcollege, popprof are significant | non significant |   Our best fit model now is the one with an R2 of 68.86%. Let’s call it perelderlypoverty.out.3.  **Code:**  perelderlypoverty.out.3=lm(perelderlypoverty ~ area + poptotal + popwhite +  popblack + popadult + percollege +perprof + perchildpoverty +  popcollege + I(popcollege^2)+popprof +I(popprof^2)+ratio + popchildpoverty, data = random.rural)  summary(perelderlypoverty.out.3)  vif(perelderlypoverty.out.3)  AIC(perelderlypoverty.out.3)  **Output:**  summary(perelderlypoverty.out.3)  Call:  lm(formula = perelderlypoverty ~ area + poptotal + popwhite +  popblack + popadult + percollege + perprof + perchildpoverty +  popcollege + I(popcollege^2) + popprof + I(popprof^2) + ratio +  popchildpoverty, data = random.rural)  Residuals:  Min 1Q Median 3Q Max  -3.9434 -1.3481 -0.3223 1.3329 3.9755  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -3.202e+00 8.702e+00 -0.368 0.714614  area 1.638e-03 1.287e-03 1.273 0.209424  poptotal -4.656e-03 1.264e-03 -3.684 0.000614 \*\*\*  popwhite 2.796e-03 9.061e-04 3.086 0.003465 \*\*  popblack 2.442e-03 8.893e-04 2.746 0.008649 \*\*  popadult 2.384e-03 1.032e-03 2.309 0.025586 \*  percollege -8.837e-01 2.592e-01 -3.410 0.001382 \*\*  perprof 2.893e+00 1.215e+00 2.381 0.021569 \*  perchildpoverty 2.860e-01 8.255e-02 3.464 0.001178 \*\*  popcollege 5.232e-03 1.630e-03 3.210 0.002452 \*\*  I(popcollege^2) -1.116e-07 5.753e-08 -1.940 0.058702 .  popprof -1.413e-02 5.978e-03 -2.364 0.022449 \*  I(popprof^2) 1.485e-06 8.289e-07 1.791 0.080022 .  ratio 2.499e+01 1.086e+01 2.300 0.026127 \*  popchildpoverty 1.431e-03 9.173e-04 1.560 0.125730  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1.964 on 45 degrees of freedom  Multiple R-squared: 0.7625, Adjusted R-squared: 0.6886  F-statistic: 10.32 on 14 and 45 DF, p-value: 8.095e-10  > AIC(perelderlypoverty.out.3)  [1] 265.99 |

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.

**Answer:**

Our best fit model is the one with the least AIC and the highest R2 of 68% (having the popcollege^2 + popprof^2 terms). Plotting graphs for this model, we get:

**Code:**

qqnorm(perelderlypoverty.out.3$residuals,pch=19)

qqline(perelderlypoverty.out.3$residuals,col="red",lwd=3)

stdresid=rstandard(perelderlypoverty.out.3)

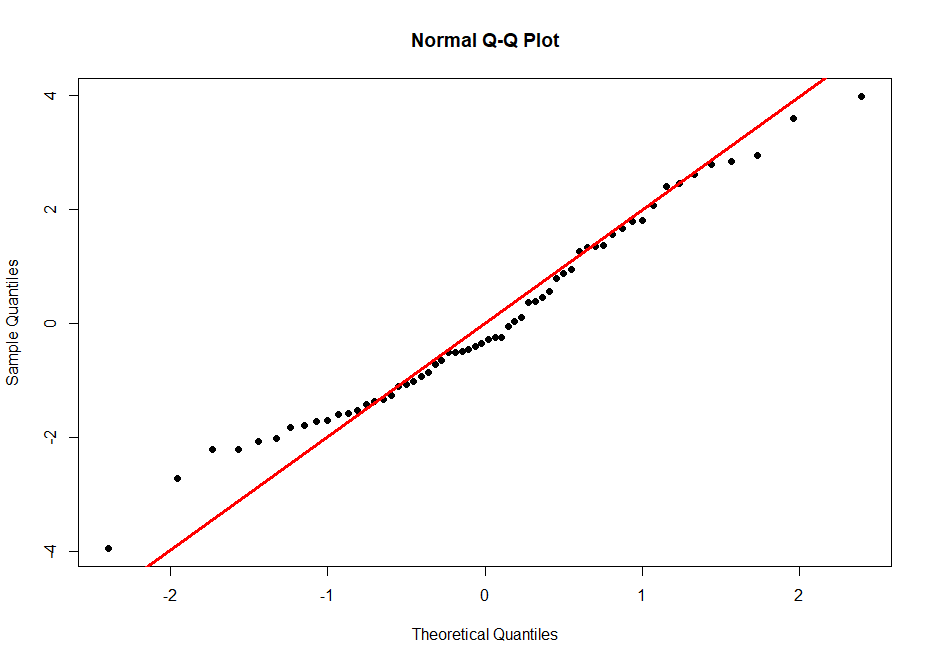
plot(random.rural$perelderlypoverty,stdresid, pch=19,xlab = "perelderlypoverty", ylab = "std residuals")

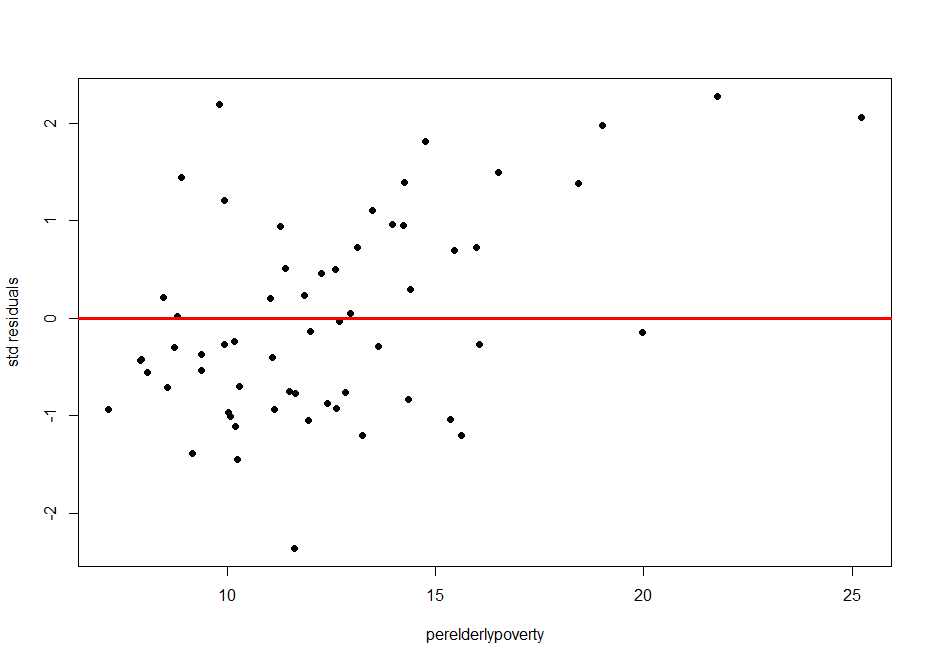
abline(0,0,col="red",lwd=3)

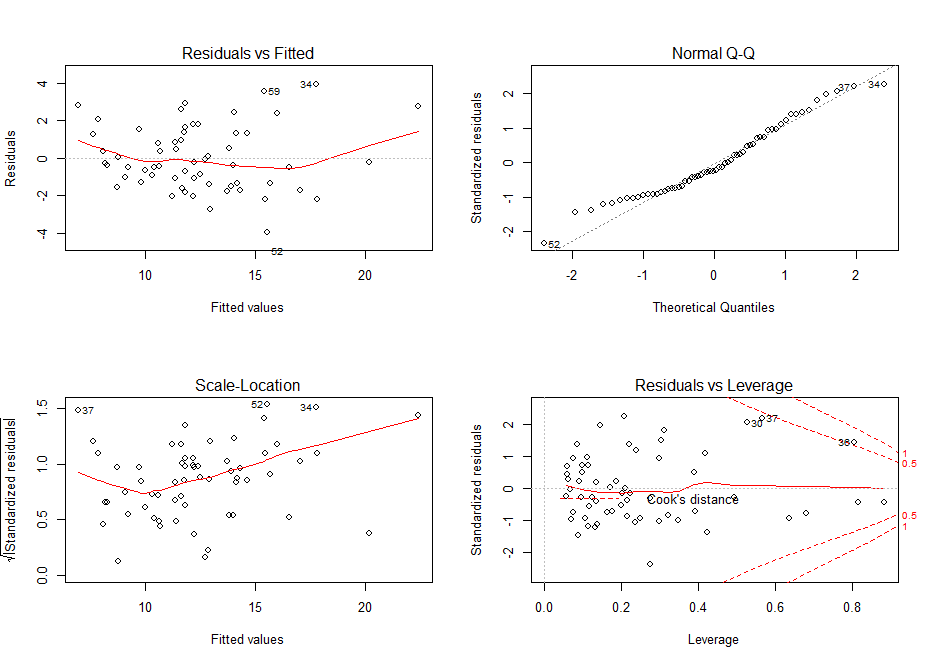
par(mfrow=c(2,2))

plot(perelderlypoverty.out.3)

**Output:**







**LINE Assumptions:**

**Linearity:** From the residuals vs fitted values plot,we can see that the observations are somewhat equally distributed around 0, except for a few extreme values (outliers). This model can be called fairly linear. So, this assumption is satisfied.

**Independence:** This is not time series data, so independence assumption is satisfied.

**Normality:** From the QQ plot, we can see that most of the observations fall roughly on the QQ line or are around it. There are a few extreme values (outliers) that deviate away from it. So, this assumption is satisfied and the model can be called as fairly normal.

**Equality of Variances:** From the standard residuals vs fitted values plot, we can see that the observations are scattered all over and are not equally distributed around zero, thus resulting in heteroscedasticity. So, this assumption fails.

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.

**Answer:**

vif(perelderlypoverty.out.3)

area poptotal popwhite popblack popadult percollege

1.422023 12153.267652 5572.856654 27.078629 3220.084301 10.905214

perprof perchildpoverty popcollege I(popcollege^2) popprof I(popprof^2)

18.595327 5.991664 679.928190 267.560833 648.322636 278.395329

ratio popchildpoverty

8.383079 26.103940

We can see that there’s high multicollinearity between a few variables in the model. So, we eliminate variables with high vif values from the above to try and reduce multicollinearity.From the above vif command, we decide to eliminate poptotal, popwhite, popadult and perprof variables from the model. Let’s call the model perelderlypoverty.out.4.

**Code:**

perelderlypoverty.out.4=lm(perelderlypoverty ~ area +

popblack + percollege + perchildpoverty +

popcollege + I(popcollege^2)+popprof +I(popprof^2)+ratio + popchildpoverty, data = random.rural)

summary(perelderlypoverty.out.4)

vif(perelderlypoverty.out.4)

AIC(perelderlypoverty.out.4)

**Output:**

summary(perelderlypoverty.out.4)

Call:

lm(formula = perelderlypoverty ~ area + popblack + percollege +

perchildpoverty + popcollege + I(popcollege^2) + popprof +

I(popprof^2) + ratio + popchildpoverty, data = random.rural)

Residuals:

Min 1Q Median 3Q Max

-4.3471 -1.7532 0.0055 1.2699 5.5057

Coefficients:

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

(Intercept) 3.787e+00 4.204e+00 0.901 0.372

area 9.309e-04 1.512e-03 0.616 0.541

popblack 1.837e-04 3.918e-04 0.469 0.641

percollege 3.456e-02 1.572e-01 0.220 0.827

perchildpoverty 2.969e-01 6.763e-02 4.390 6.03e-05 \*\*\*

popcollege -7.150e-04 1.039e-03 -0.688 0.495

I(popcollege^2) 2.252e-08 5.073e-08 0.444 0.659

popprof 1.377e-03 3.084e-03 0.446 0.657

I(popprof^2) -3.492e-07 6.843e-07 -0.510 0.612

ratio 6.476e+00 5.498e+00 1.178 0.245

popchildpoverty 2.341e-04 7.048e-04 0.332 0.741

---

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

Residual standard error: 2.417 on 49 degrees of freedom

Multiple R-squared: 0.6083, Adjusted R-squared: 0.5284

F-statistic: 7.611 on 10 and 49 DF, p-value: 3.559e-07

> vif(perelderlypoverty.out.4)

area popblack percollege perchildpoverty popcollege I(popcollege^2)

1.296926 3.470541 2.649713 2.655653 182.478322 137.353273

popprof I(popprof^2) ratio popchildpoverty

113.939970 125.276292 1.417960 10.175966

> AIC(perelderlypoverty.out.4)

[1] 288.0019

There’s not much multi-collinearity now but we construct a correlation plot by taking a subset of only the above columns to see if any other variables have high correlation values and if we can eliminate any to increase R2.

m.num = random.rural[c(4,8,12,14,11,18,19,20)]

> cor(m.num)

area popblack percollege perchildpoverty popchild popprof

area 1.000000000 0.002512423 0.19916126 0.06890899 0.1643542 0.1306814

popblack 0.002512423 1.000000000 0.04470785 0.12493240 0.6304829 0.6579915

percollege 0.199161261 0.044707850 1.00000000 -0.32169900 0.2505139 0.3553808

perchildpoverty 0.068908988 0.124932403 -0.32169900 1.00000000 -0.2343872 -0.2611017

popchild 0.164354189 0.630482916 0.25051391 -0.23438718 1.0000000 0.9552111

popprof 0.130681410 0.657991498 0.35538078 -0.26110171 0.9552111 1.0000000

ratio -0.152400582 0.046890566 -0.09679065 -0.20985992 0.3795596 0.2801143

popchildpoverty 0.177439691 0.662519027 0.09214439 0.11914109 0.9030304 0.8267959

ratio popchildpoverty

area -0.15240058 0.17743969

popblack 0.04689057 0.66251903

percollege -0.09679065 0.09214439

perchildpoverty -0.20985992 0.11914109

popchild 0.37955961 0.90303040

popprof 0.28011432 0.82679594

ratio 1.00000000 0.28631117

popchildpoverty 0.28631117 1.00000000

Popchild, popblack, popprof and popchildpoverty have high correlations of around 0.9 or more. We remove only popchildpoverty and popblack as popprof has a squared term and cannot be removed. Popchild has already been removed.

**Code:**

perelderlypoverty.out.5=lm(perelderlypoverty ~ area +

percollege + perchildpoverty +

popcollege + I(popcollege^2)+popprof +I(popprof^2)+ratio, data = random.rural)

summary(perelderlypoverty.out.5)

vif(perelderlypoverty.out.5)

AIC(perelderlypoverty.out.5)

m.num = random.rural[c(4,12,14,18,19)]

View(m.num)

cor(m.num)

**Output:**

summary(perelderlypoverty.out.5)

Call:

lm(formula = perelderlypoverty ~ area + percollege + perchildpoverty +

popcollege + I(popcollege^2) + popprof + I(popprof^2) + ratio,

data = random.rural)

Residuals:

Min 1Q Median 3Q Max

-4.3359 -1.4259 -0.1317 1.2014 5.6855

Coefficients:

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

(Intercept) 3.978e+00 4.106e+00 0.969 0.337

area 7.521e-04 1.448e-03 0.519 0.606

percollege -9.677e-03 1.239e-01 -0.078 0.938

perchildpoverty 3.184e-01 4.730e-02 6.732 1.44e-08 \*\*\*

popcollege -3.515e-04 7.524e-04 -0.467 0.642

I(popcollege^2) 5.460e-09 3.938e-08 0.139 0.890

popprof 5.972e-04 2.694e-03 0.222 0.825

I(popprof^2) -1.188e-07 5.311e-07 -0.224 0.824

ratio 6.468e+00 5.338e+00 1.212 0.231

---

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

Residual standard error: 2.376 on 51 degrees of freedom

Multiple R-squared: 0.6059, Adjusted R-squared: 0.5441

F-statistic: 9.802 on 8 and 51 DF, p-value: 4.039e-08

> vif(perelderlypoverty.out.5)

area percollege perchildpoverty popcollege I(popcollege^2) popprof

1.230096 1.701023 1.343591 98.961308 85.634872 89.959055

I(popprof^2) ratio

78.053357 1.382567

> AIC(perelderlypoverty.out.5)

[1] 284.3727

> m.num = random.rural[c(4,12,14,18,19)]

> cor(m.num)

area percollege perchildpoverty popprof ratio

area 1.00000000 0.19916126 0.06890899 0.1306814 -0.15240058

percollege 0.19916126 1.00000000 -0.32169900 0.3553808 -0.09679065

perchildpoverty 0.06890899 -0.32169900 1.00000000 -0.2611017 -0.20985992

popprof 0.13068141 0.35538078 -0.26110171 1.0000000 0.28011432

ratio -0.15240058 -0.09679065 -0.20985992 0.2801143 1.00000000

This is our final best model fit with an R2 of around 54% after adjusting for multicollinearity. We can see from the correlation plot that there’s no multicollinearity in the remaining variables. Popcollege and popprof will have high values because of their squared terms. AIC is also the lowest, indicating that this model is the best possible fit for the chosen random rural sample.

We now check LINE assumptions for this.

**LINE:**

**Code:**

qqnorm(perelderlypoverty.out.5$residuals,pch=19)

qqline(perelderlypoverty.out.5$residuals,col="red",lwd=3)

stdresid=rstandard(perelderlypoverty.out.5)

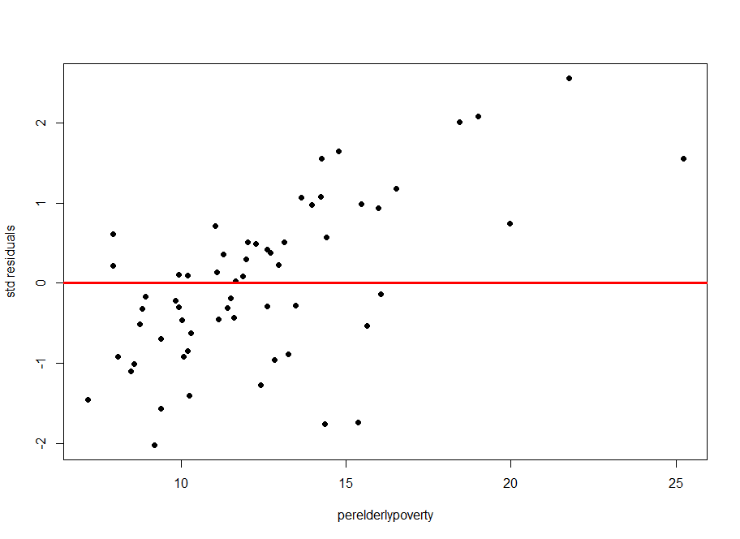
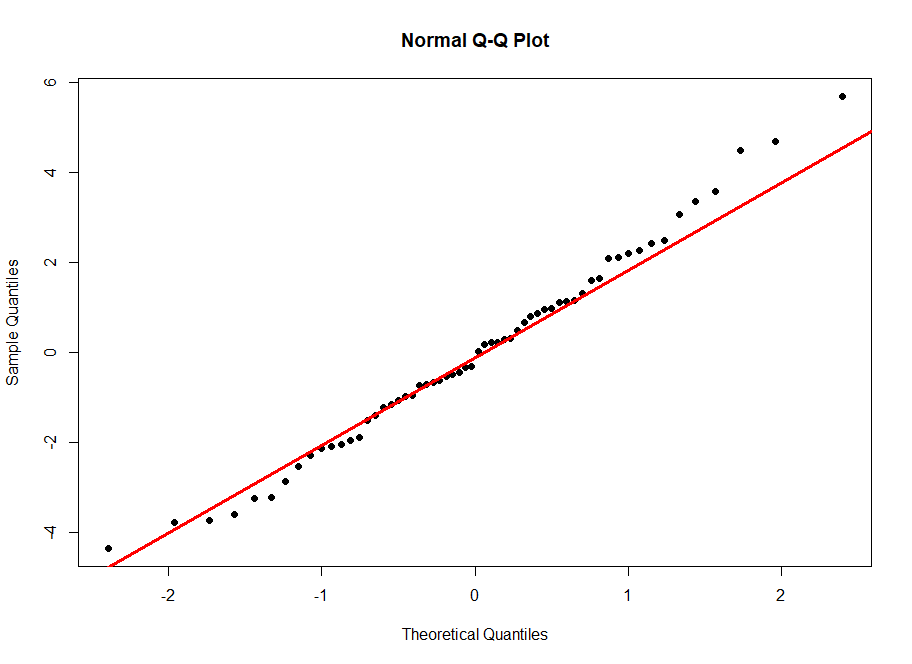
plot(random.rural$perelderlypoverty,stdresid, pch=19,xlab = "perelderlypoverty", ylab = "std residuals")

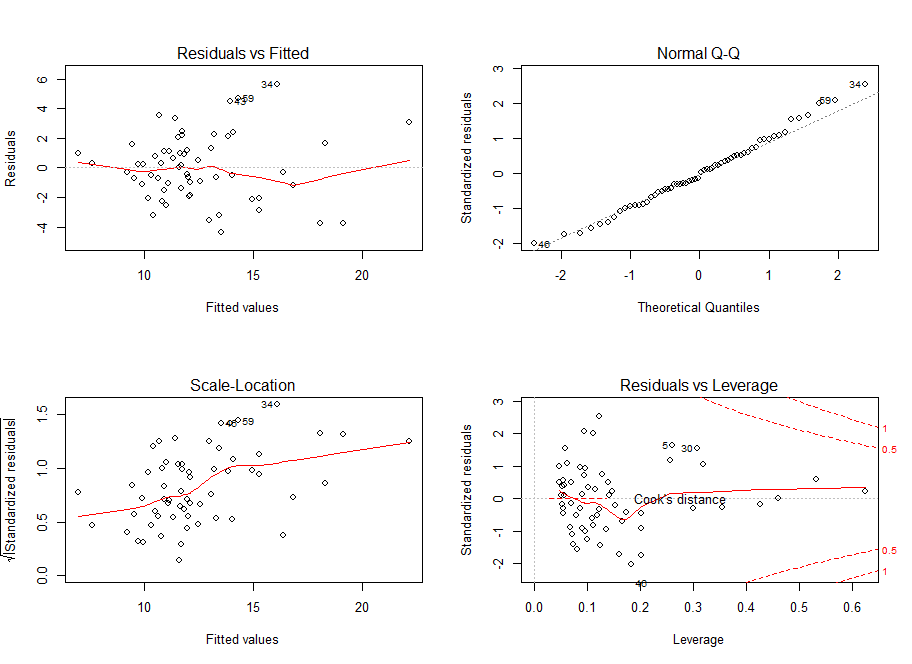
abline(0,0,col="red",lwd=3)

par(mfrow=c(2,2))

plot(perelderlypoverty.out.5)

**Output:**





**LINE:**

Linearity: From the residuals vs fitted values plot, the model looks to be linear except for a couple of values at the extremes.

Independence: This assumption is satisfied as it’s not a time series data.

Normality: From the QQnorm plot, the model looks to be fairly normal as most of the values fall on the line and there are no extreme values. A few high values deviate a little from the line at the extreme end.

Equality of Variances: The model looks to be homoscedastic from the std residuals vs fitted values plot as it’s distributed around zero and doesn’t have any pattern.

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.

**Answer:**

**Code:**

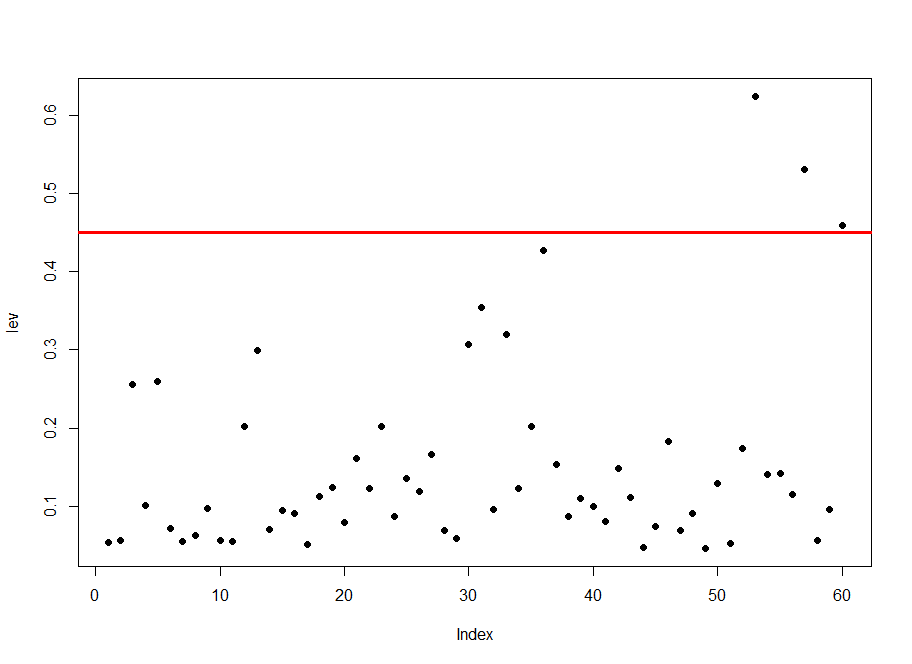
lev=hat(model.matrix(perelderlypoverty.out.5))

plot(lev,pch=19)

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

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

**Output:**



There are 3 values with high leverage influencing my best fit model. They are:

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

# A tibble: 3 x 20

id county state area poptotal popdensity popwhite popblack popasian popadult popchild

*<dbl>* *<chr>* *<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 148 LA PO~ IN 598. 107066 179. 96286 9580 431 70102 36964

2 309 HANCO~ OH 531. 65536 123. 63572 591 401 41492 24044

3 191 WAYNE IN 404. 71951 178. 67532 3795 296 46603 25348

# ... with 9 more variables: percollege *<dbl>*, perprof *<dbl>*, perchildpoverty *<dbl>*,

# perelderlypoverty *<dbl>*, inmetro *<dbl>*, popcollege *<dbl>*, popprof *<dbl>*, ratio *<dbl>*,

# popchildpoverty *<dbl>*

The three counties are

LA PORTE in Indiana,

WAYNE in Indiana and

HANCOCK in Ohio.

|  |
| --- |
|  |

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.

**Answer:**

**Code:**

perelderlypoverty.out.6=lm(perelderlypoverty ~ area +

percollege + perchildpoverty +

popcollege + I(popcollege^2)+popprof +I(popprof^2)+ratio, data = random.metropolitan)

summary(perelderlypoverty.out.6)

vif(perelderlypoverty.out.6)

AIC(perelderlypoverty.out.6)

**LINE:**

qqnorm(perelderlypoverty.out.6$residuals,pch=19)

qqline(perelderlypoverty.out.6$residuals,col="red",lwd=3)

par(mfrow=c(2,2))

plot(perelderlypoverty.out.6)

**Output:**

summary(perelderlypoverty.out.6)

Call:

lm(formula = perelderlypoverty ~ area + percollege + perchildpoverty +

popcollege + I(popcollege^2) + popprof + I(popprof^2) + ratio,

data = random.metropolitan)

Residuals:

Min 1Q Median 3Q Max

-3.3008 -0.8944 0.0411 0.7400 2.8271

Coefficients:

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

(Intercept) 7.236e+00 3.104e+00 2.331 0.029807 \*

area -1.452e-03 2.357e-03 -0.616 0.544443

percollege -2.489e-01 1.255e-01 -1.984 0.060512 .

perchildpoverty 2.352e-01 6.053e-02 3.885 0.000855 \*\*\*

popcollege -2.680e-05 8.184e-05 -0.327 0.746571

I(popcollege^2) -1.353e-10 2.436e-10 -0.555 0.584572

popprof 9.982e-05 2.973e-04 0.336 0.740400

I(popprof^2) 1.897e-09 2.995e-09 0.633 0.533395

ratio 7.313e+00 6.446e+00 1.134 0.269383

---

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

Residual standard error: 1.63 on 21 degrees of freedom

Multiple R-squared: 0.6319, Adjusted R-squared: 0.4917

F-statistic: 4.506 on 8 and 21 DF, p-value: 0.002639

> vif(perelderlypoverty.out.6)

area percollege perchildpoverty popcollege I(popcollege^2) popprof

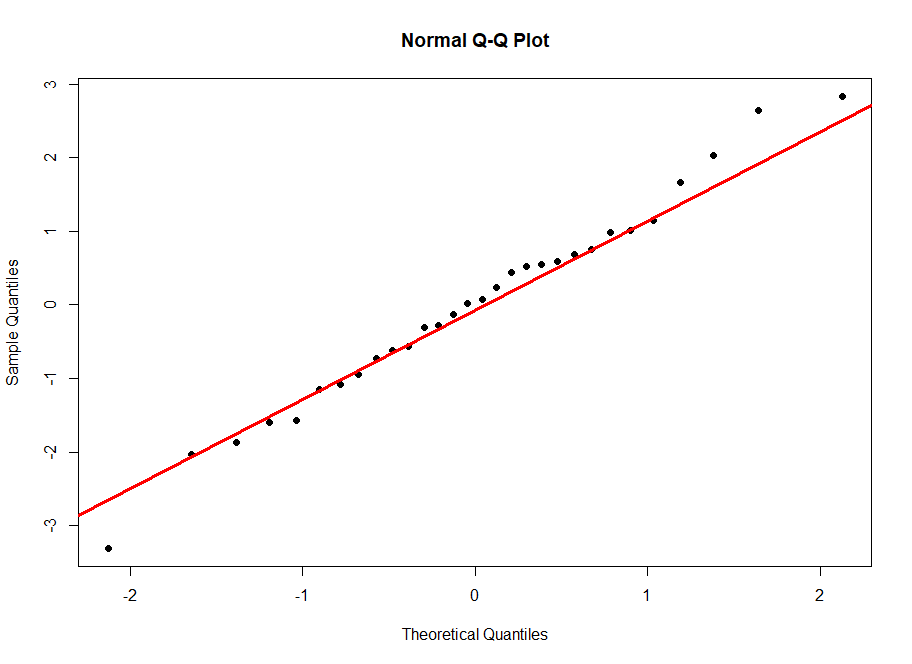
1.979564 9.972502 1.562939 465.379797 337.745032 514.780851

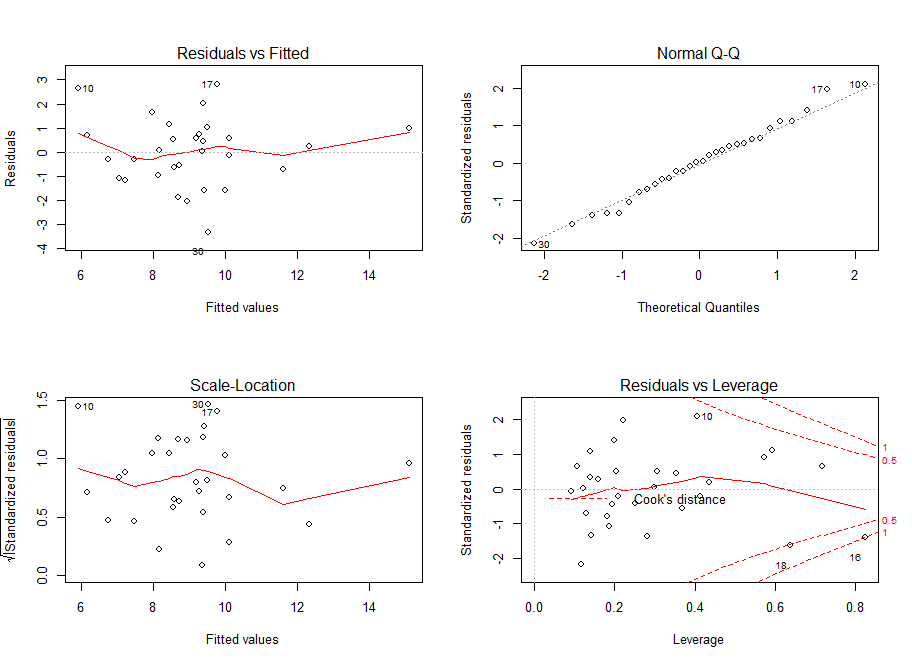
I(popprof^2) ratio

332.484695 3.388962

> AIC(perelderlypoverty.out.6)

[1] 123.7501





LINE:

Linearity: From the residuals vs fitted values plot, the model looks to be linear except for a couple of values at the extremes.

Independence: This assumption is satisfied as it’s not a time series data.

Normality: From the QQnorm plot, the model looks to be normal as most of the values fall on the line and there are no extreme values.

Equality of Variances: The model looks to be homoscedastic from the std residuals vs fitted values plot as it’s distributed around zero and doesn’t have any pattern.

The metro dataset has an R2 of 49% which is lower than that of the model for rural dataset. Comparing purely based on R2, we can say that the model fits the rural dataset better. But there are several other conditions and assumptions to be taken care of before coming to the above conclusion.

1. The lengths of both the datasets are different i.e, 60 and 30. So we cannot use AIC to compare which one is better as they’re different datasets.
2. Also, we have computed so many models for the rural dataset and picked the best one but computed only one model for the metro dataset. We might have to check other models too, to see if they are better or not, have better R2 and lower multicollinearity, etc along with the LINE assumptions for each model.

So, it needs to be investigated further.

Your deliverable will be a single MS-Word file created using R Markdown. Your file will show 1) the R script which executes the above instructions and 2) the results of those instructions. The first two lines of your deliverable will state this is the “Regression Project” of our course and your name as it appears in Canvas. Your code chunks and analysis results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.