6304 Regression Project

Jashwanth Gottipati

**Preprocessing**

**Code:**

*rm(list=ls())*

*library(car)*

*library(readxl)*

*library(rio)*

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

*colnames(project)=tolower(make.names(colnames(project)))*

*project$popcollege = (project$poptotal \* project$percollege ) / 100*

*project$popprof = (project$poptotal \* project$perprof ) / 100*

*project$childadultratio = project$popchild / project$popadult*

*project$popchildpoverty = (project$popchild \* project$perchildpoverty) / 100*

*attach(project)*

*ruraldata <- subset(project,inmetro==0)*

*metrodata <- subset(project,inmetro==1)*

*set.seed(73175862)*

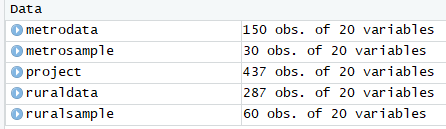
*ruralsample=ruraldata[sample(1:nrow(ruraldata),60,replace=FALSE),]*

*metrosample=metrodata[sample(1:nrow(metrodata),30,replace=FALSE),]*

*attach(ruralsample)*

*attach(metrosample)*

**Output:**



**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.

**Solution:**

I have selected **“step”** function to predict independent variables that need to be used in the model.

**Code:**

*step(lm(ruralsample$perelderlypoverty~ruralsample$area+ruralsample$poptotal+ruralsample$popdensity+ruralsample$popwhite+ruralsample$popblack+ruralsample$popasian+ruralsample$popadult+ruralsample$popchild+ruralsample$percollege+ruralsample$perprof+ruralsample$perchildpoverty+ruralsample$perelderlypoverty+ruralsample$popcollege+ruralsample$popprof+ruralsample$childadultratio+ruralsample$popchildpoverty, data=ruralsample,direction = "both"))*

**Output:**

Start: AIC=131.44

ruralsample$perelderlypoverty ~ ruralsample$area + ruralsample$poptotal +

ruralsample$popdensity + ruralsample$popwhite + ruralsample$popblack +

ruralsample$popasian + ruralsample$popadult + ruralsample$popchild +

ruralsample$percollege + ruralsample$perprof + ruralsample$perchildpoverty +

ruralsample$perelderlypoverty + ruralsample$popcollege +

ruralsample$popprof + ruralsample$childadultratio + ruralsample$popchildpoverty

Step: AIC=131.44

ruralsample$perelderlypoverty ~ ruralsample$area + ruralsample$poptotal +

ruralsample$popdensity + ruralsample$popwhite + ruralsample$popblack +

ruralsample$popasian + ruralsample$popadult + ruralsample$popchild +

ruralsample$percollege + ruralsample$perprof + ruralsample$perchildpoverty +

ruralsample$popcollege + ruralsample$popprof + ruralsample$childadultratio +

ruralsample$popchildpoverty

Step: AIC=131.44

ruralsample$perelderlypoverty ~ ruralsample$area + ruralsample$poptotal +

ruralsample$popdensity + ruralsample$popwhite + ruralsample$popblack +

ruralsample$popasian + ruralsample$popadult + ruralsample$percollege +

ruralsample$perprof + ruralsample$perchildpoverty + ruralsample$popcollege +

ruralsample$popprof + ruralsample$childadultratio + ruralsample$popchildpoverty

Df Sum of Sq RSS AIC

- ruralsample$popasian 1 0.271 325.68 129.50

- ruralsample$poptotal 1 1.235 326.65 129.67

- ruralsample$popwhite 1 1.384 326.80 129.70

- ruralsample$popprof 1 1.474 326.89 129.72

- ruralsample$perprof 1 2.254 327.67 129.86

- ruralsample$popblack 1 2.412 327.82 129.89

- ruralsample$popadult 1 3.233 328.65 130.04

- ruralsample$perchildpoverty 1 4.595 330.01 130.29

<none> 325.41 131.44

- ruralsample$childadultratio 1 13.134 338.55 131.82

- ruralsample$popchildpoverty 1 13.895 339.31 131.95

- ruralsample$area 1 14.553 339.96 132.07

- ruralsample$popdensity 1 19.566 344.98 132.95

- ruralsample$popcollege 1 25.319 350.73 133.94

- ruralsample$percollege 1 56.353 381.76 139.03

Step: AIC=129.49

ruralsample$perelderlypoverty ~ ruralsample$area + ruralsample$poptotal +

ruralsample$popdensity + ruralsample$popwhite + ruralsample$popblack +

ruralsample$popadult + ruralsample$percollege + ruralsample$perprof +

ruralsample$perchildpoverty + ruralsample$popcollege + ruralsample$popprof +

ruralsample$childadultratio + ruralsample$popchildpoverty

Df Sum of Sq RSS AIC

- ruralsample$poptotal 1 1.003 326.69 127.68

- ruralsample$popwhite 1 1.144 326.83 127.70

- ruralsample$popprof 1 1.229 326.91 127.72

- ruralsample$perprof 1 2.011 327.69 127.86

- ruralsample$popblack 1 2.143 327.83 127.89

- ruralsample$popadult 1 3.187 328.87 128.08

- ruralsample$perchildpoverty 1 4.400 330.08 128.30

<none> 325.68 129.50

- ruralsample$childadultratio 1 13.667 339.35 129.96

- ruralsample$popchildpoverty 1 14.777 340.46 130.16

- ruralsample$area 1 15.864 341.55 130.35

- ruralsample$popdensity 1 19.924 345.61 131.06

- ruralsample$popcollege 1 25.204 350.89 131.97

- ruralsample$percollege 1 56.166 381.85 137.04

Step: AIC=127.68

ruralsample$perelderlypoverty ~ ruralsample$area + ruralsample$popdensity +

ruralsample$popwhite + ruralsample$popblack + ruralsample$popadult +

ruralsample$percollege + ruralsample$perprof + ruralsample$perchildpoverty +

ruralsample$popcollege + ruralsample$popprof + ruralsample$childadultratio +

ruralsample$popchildpoverty

Df Sum of Sq RSS AIC

- ruralsample$popwhite 1 0.150 326.84 125.71

- ruralsample$popprof 1 1.789 328.47 126.01

- ruralsample$popblack 1 2.029 328.72 126.05

- ruralsample$perprof 1 2.281 328.97 126.10

- ruralsample$popadult 1 3.767 330.45 126.37

- ruralsample$perchildpoverty 1 3.809 330.50 126.38

<none> 326.69 127.68

- ruralsample$area 1 15.268 341.95 128.42

- ruralsample$childadultratio 1 17.252 343.94 128.77

- ruralsample$popchildpoverty 1 18.349 345.04 128.96

- ruralsample$popdensity 1 19.220 345.91 129.11

- ruralsample$popcollege 1 25.048 351.73 130.11

- ruralsample$percollege 1 55.978 382.66 135.17

Step: AIC=125.71

ruralsample$perelderlypoverty ~ ruralsample$area + ruralsample$popdensity +

ruralsample$popblack + ruralsample$popadult + ruralsample$percollege +

ruralsample$perprof + ruralsample$perchildpoverty + ruralsample$popcollege +

ruralsample$popprof + ruralsample$childadultratio + ruralsample$popchildpoverty

Df Sum of Sq RSS AIC

- ruralsample$popprof 1 1.665 328.50 124.01

- ruralsample$perprof 1 2.132 328.97 124.10

- ruralsample$popblack 1 2.729 329.57 124.21

- ruralsample$perchildpoverty 1 3.716 330.55 124.39

<none> 326.84 125.71

- ruralsample$area 1 15.124 341.96 126.42

- ruralsample$popdensity 1 19.366 346.20 127.16

- ruralsample$popchildpoverty 1 22.198 349.03 127.65

- ruralsample$popcollege 1 25.201 352.04 128.16

- ruralsample$childadultratio 1 31.823 358.66 129.28

- ruralsample$popadult 1 35.362 362.20 129.87

- ruralsample$percollege 1 55.865 382.70 133.18

Step: AIC=124.01

ruralsample$perelderlypoverty ~ ruralsample$area + ruralsample$popdensity +

ruralsample$popblack + ruralsample$popadult + ruralsample$percollege +

ruralsample$perprof + ruralsample$perchildpoverty + ruralsample$popcollege +

ruralsample$childadultratio + ruralsample$popchildpoverty

Df Sum of Sq RSS AIC

- ruralsample$perprof 1 0.495 329.00 122.10

- ruralsample$popblack 1 2.787 331.29 122.52

- ruralsample$perchildpoverty 1 4.139 332.64 122.76

<none> 328.50 124.01

- ruralsample$popchildpoverty 1 20.533 349.03 125.65

- ruralsample$area 1 21.492 349.99 125.81

- ruralsample$childadultratio 1 30.405 358.91 127.32

- ruralsample$popadult 1 35.210 363.71 128.12

- ruralsample$popdensity 1 37.102 365.60 128.43

- ruralsample$popcollege 1 50.283 378.78 130.56

- ruralsample$percollege 1 70.670 399.17 133.70

Step: AIC=122.1

ruralsample$perelderlypoverty ~ ruralsample$area + ruralsample$popdensity +

ruralsample$popblack + ruralsample$popadult + ruralsample$percollege +

ruralsample$perchildpoverty + ruralsample$popcollege + ruralsample$childadultratio +

ruralsample$popchildpoverty

Df Sum of Sq RSS AIC

- ruralsample$popblack 1 2.845 331.84 120.62

- ruralsample$perchildpoverty 1 3.751 332.75 120.78

<none> 329.00 122.10

- ruralsample$area 1 21.861 350.86 123.96

- ruralsample$popchildpoverty 1 22.017 351.01 123.99

- ruralsample$childadultratio 1 29.910 358.91 125.32

- ruralsample$popdensity 1 36.654 365.65 126.44

- ruralsample$popadult 1 39.411 368.41 126.89

- ruralsample$popcollege 1 53.289 382.28 129.11

- ruralsample$percollege 1 78.984 407.98 133.01

Step: AIC=120.62

ruralsample$perelderlypoverty ~ ruralsample$area + ruralsample$popdensity +

ruralsample$popadult + ruralsample$percollege + ruralsample$perchildpoverty +

ruralsample$popcollege + ruralsample$childadultratio + ruralsample$popchildpoverty

Df Sum of Sq RSS AIC

- ruralsample$perchildpoverty 1 2.968 334.81 119.15

<none> 331.84 120.62

- ruralsample$area 1 21.889 353.73 122.45

- ruralsample$popchildpoverty 1 29.956 361.80 123.80

- ruralsample$popdensity 1 34.123 365.96 124.49

- ruralsample$childadultratio 1 35.175 367.02 124.66

- ruralsample$popadult 1 38.206 370.05 125.16

- ruralsample$popcollege 1 50.464 382.30 127.11

- ruralsample$percollege 1 77.443 409.28 131.20

Step: AIC=119.15

ruralsample$perelderlypoverty ~ ruralsample$area + ruralsample$popdensity +

ruralsample$popadult + ruralsample$percollege + ruralsample$popcollege +

ruralsample$childadultratio + ruralsample$popchildpoverty

Df Sum of Sq RSS AIC

<none> 334.81 119.15

- ruralsample$area 1 21.058 355.87 120.81

- ruralsample$popdensity 1 39.487 374.30 123.84

- ruralsample$childadultratio 1 55.709 390.52 126.39

- ruralsample$popcollege 1 76.817 411.63 129.55

- ruralsample$popadult 1 88.893 423.70 131.28

- ruralsample$percollege 1 113.871 448.68 134.72

- ruralsample$popchildpoverty 1 162.547 497.36 140.90

Call:

lm(formula = ruralsample$perelderlypoverty ~ ruralsample$area +

ruralsample$popdensity + ruralsample$popadult + ruralsample$percollege +

ruralsample$popcollege + ruralsample$childadultratio + ruralsample$popchildpoverty,

data = ruralsample, direction = "both")

Coefficients:

(Intercept) ruralsample$area ruralsample$popdensity ruralsample$popadult

3.753e+01 -4.028e-03 -6.753e-02 -4.875e-04

ruralsample$percollege ruralsample$popcollege ruralsample$childadultratio ruralsample$popchildpoverty

-7.541e-01 1.389e-03 -1.590e+01 2.494e-03

**Inclusion from the output:**

From the result of step function, we can consider that the independent variables we need for models are area, popdensity, popadult, percollege, popcollege, childadultratio and popchildpoverty.

The AIC value we got after considering all the independent variables is 131.44, after not considering few variables, we have got an AIC value of 119.15.

**Performing regression from output we have got from step function:**

**Code:**

*reg.out = lm(formula = ruralsample$perelderlypoverty ~*

*ruralsample$popadult + ruralsample$percollege +*

*ruralsample$popcollege + ruralsample$childadultratio + ruralsample$popchildpoverty,*

*data = ruralsample)*

*summary(reg.out)*

**Output:**

> summary(reg.out)

lm(formula = ruralsample$perelderlypoverty ~ ruralsample$area +

ruralsample$popdensity + ruralsample$popadult + ruralsample$percollege +

ruralsample$popcollege + ruralsample$childadultratio + ruralsample$popchildpoverty,

data = ruralsample)

Residuals:

Min 1Q Median 3Q Max

-5.3322 -1.9061 0.2633 1.4692 6.1939

Coefficients:

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

(Intercept) 3.753e+01 4.261e+00 8.808 6.88e-12 \*\*\*

ruralsample$area -4.028e-03 2.227e-03 -1.808 0.076320 .

ruralsample$popdensity -6.753e-02 2.727e-02 -2.476 0.016560 \*

ruralsample$popadult -4.875e-04 1.312e-04 -3.716 0.000496 \*\*\*

ruralsample$percollege -7.541e-01 1.793e-01 -4.205 0.000103 \*\*\*

ruralsample$popcollege 1.389e-03 4.022e-04 3.454 0.001107 \*\*

ruralsample$childadultratio -1.590e+01 5.404e+00 -2.941 0.004868 \*\*

ruralsample$popchildpoverty 2.494e-03 4.963e-04 5.024 6.32e-06 \*\*\*

---

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

Residual standard error: 2.537 on 52 degrees of freedom

Multiple R-squared: 0.6221, Adjusted R-squared: 0.5712

F-statistic: 12.23 on 7 and 52 DF, p-value: 4.058e-09

***Note: Even though the p value for area is higher, removing it from the model is showing a huge change in the R-Squared and Adjusted R-Squared value and as the data we are considering is very small amount, we can ignore.***

**Detective work:**

Let try to use squared terms in the model,

**Code:**

*reg.out = lm(formula = ruralsample$perelderlypoverty ~ ruralsample$area +*

*ruralsample$popdensity + ruralsample$popadult + ruralsample$percollege +*

*ruralsample$popcollege + ruralsample$childadultratio + ruralsample$popchildpoverty+I(area^2)*

*+I(popdensity^2) + I(popadult^2),*

*data = ruralsample)*

**Output:**

Call:

lm(formula = ruralsample$perelderlypoverty ~ ruralsample$area +

ruralsample$popdensity + ruralsample$popadult + ruralsample$percollege +

ruralsample$popcollege + ruralsample$childadultratio + ruralsample$popchildpoverty +

I(area^2) + I(popdensity^2) + I(popadult^2), data = ruralsample)

Residuals:

Min 1Q Median 3Q Max

-4.6455 -1.7938 0.0953 1.4921 6.3011

Coefficients:

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

(Intercept) 4.085e+01 4.980e+00 8.203 9.38e-11 \*\*\*

ruralsample$area -1.834e-02 9.534e-03 -1.923 0.060288 .

ruralsample$popdensity -7.848e-02 9.560e-02 -0.821 0.415656

ruralsample$popadult -4.265e-04 2.909e-04 -1.466 0.148986

ruralsample$percollege -6.955e-01 1.920e-01 -3.622 0.000692 \*\*\*

ruralsample$popcollege 1.337e-03 4.498e-04 2.973 0.004568 \*\*

ruralsample$childadultratio -1.560e+01 6.784e+00 -2.299 0.025793 \*

ruralsample$popchildpoverty 2.682e-03 5.503e-04 4.874 1.19e-05 \*\*\*

I(area^2) 8.807e-06 4.970e-06 1.772 0.082622 .

I(popdensity^2) -3.817e-06 3.258e-04 -0.012 0.990699

I(popadult^2) -4.296e-10 2.225e-09 -0.193 0.847705

---

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

Residual standard error: 2.528 on 49 degrees of freedom

Multiple R-squared: 0.6465, Adjusted R-squared: 0.5743

F-statistic: 8.961 on 10 and 49 DF, p-value: 3.631e-08

**Conclusion:**

Output: Even after all the detective work, the adjusted R-squared value is almost same as that of the model we got from step function and even all the p values for the squared terms are greater than 0.05. Hence, I would like to stick with model that was the output from step function.

**2.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.**

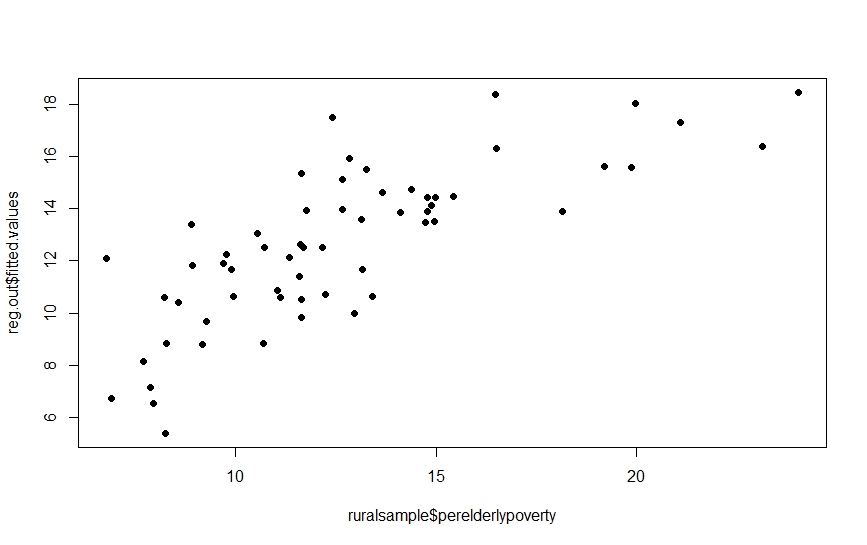
**Solution:**

**L-Linearity:**

**Code:**

*plot(ruralsample$perelderlypoverty, reg.out$fitted.values,pch=19)*

**Output:**

**

Observation: From the above plot, we can see that points are not linear. Hence assumption of linearity is not meet in the model.

**I – Independence:**

As this is not a time series data, we might not bother about assumption of independence.

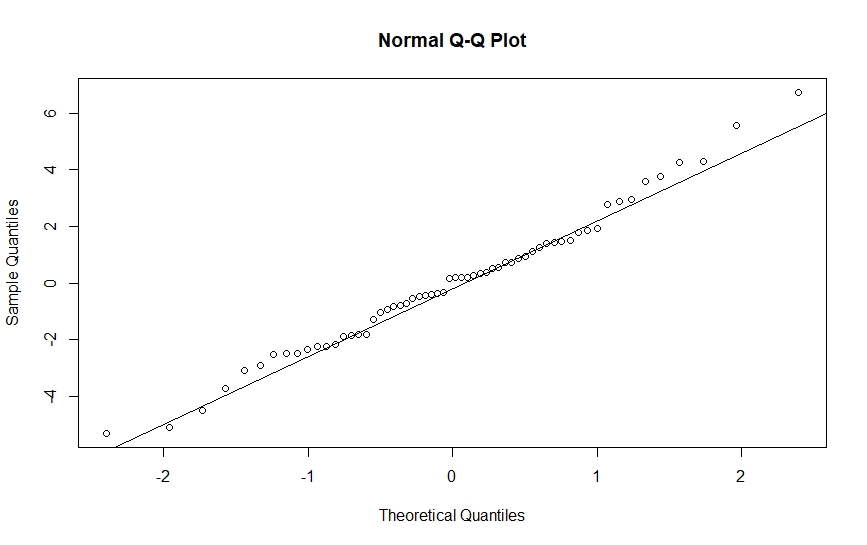
**N- Normality:**

**Code:**

*qqnorm(reg.out$residuals)*

*qqline(reg.out$residuals)*

**Output:**



Observation/conclusion: From the above Q-Q plot we can predict that all the residuals didn’t fall on the line. Many of them in top-right are far from the Q-Q line. Hence, we can assume that the assumption of normality may fail here.

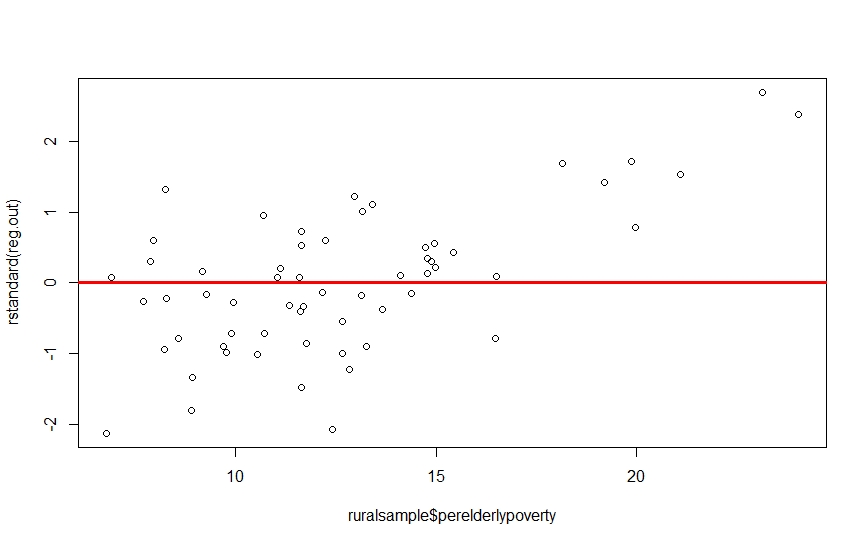
**E- Equality of variance of residuals**

**Code:**

*plot(ruralsample$perelderlypoverty,rstandard(reg.out))*

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

**Output:**



Observation/Conclusion: From the above plot we can deduct that the variance is not equally distributed from the horizontal line, there is lot of variance in the top right area of the plot. Hence this assumption can’t be meet.

**3.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**

**Code:**

*vif(reg.out)*

**Output:**

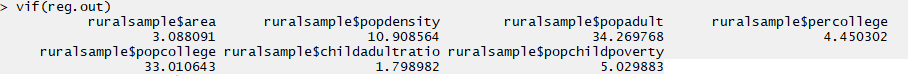
> vif(reg.out)

ruralsample$area ruralsample$popdensity ruralsample$popadult ruralsample$percollege

3.088091 10.908564 34.269768 4.450302

ruralsample$popcollege ruralsample$childadultratio ruralsample$popchildpoverty

33.010643 1.798982 5.029883



As the VIF value for few variables are high, that mean there is some multicollinearity in this model.

Let’s remove some of the variables and run the regression again.

**Code:**

*reg.out2 = lm(formula = ruralsample$perelderlypoverty ~ ruralsample$area+*

*ruralsample$popdensity + ruralsample$percollege + ruralsample$childadultratio + ruralsample$popchildpoverty,*

*data = ruralsample)*

*vif(reg.out2)*

**Output:**

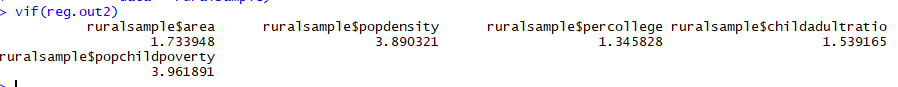
> vif(reg.out2)

ruralsample$area ruralsample$popdensity ruralsample$percollege ruralsample$childadultratio

1.733948 3.890321 1.345828 1.539165

ruralsample$popchildpoverty

3.961891



All the VIF values were reduced by a significant value, that mean, multicollinearity was reduced by a good amount.

***Lets run regression again with variables from above output and consider this as our best fit***

**Summary of new regression:**

> summary(reg.out2)

Call:

lm(formula = ruralsample$perelderlypoverty ~ ruralsample$area +

ruralsample$popdensity + ruralsample$percollege + ruralsample$childadultratio +

ruralsample$popchildpoverty, data = ruralsample)

Residuals:

Min 1Q Median 3Q Max

-5.1940 -1.7424 -0.4762 1.7286 7.4375

Coefficients:

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

(Intercept) 27.0422323 3.2284838 8.376 2.45e-11 \*\*\*

ruralsample$area -0.0068883 0.0018499 -3.724 0.000470 \*\*\*

ruralsample$popdensity -0.0885992 0.0180498 -4.909 8.85e-06 \*\*\*

ruralsample$percollege -0.2365861 0.1092927 -2.165 0.034849 \*

ruralsample$childadultratio -8.5278151 5.5404251 -1.539 0.129596

ruralsample$popchildpoverty 0.0017673 0.0004882 3.620 0.000651 \*\*\*

---

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

Residual standard error: 2.812 on 54 degrees of freedom

Multiple R-squared: 0.5179, Adjusted R-squared: 0.4732

F-statistic: 11.6 on 5 and 54 DF, p-value: 1.232e-07

Even though the R squared value was reduced by a lot compared to our reg.out’s R squared values and few of the variables have p-values greater than 0.05 but, in this model, there is no multicollinearity, hence let’s finalize reg.out2 model to be our final model.

**4. 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.**

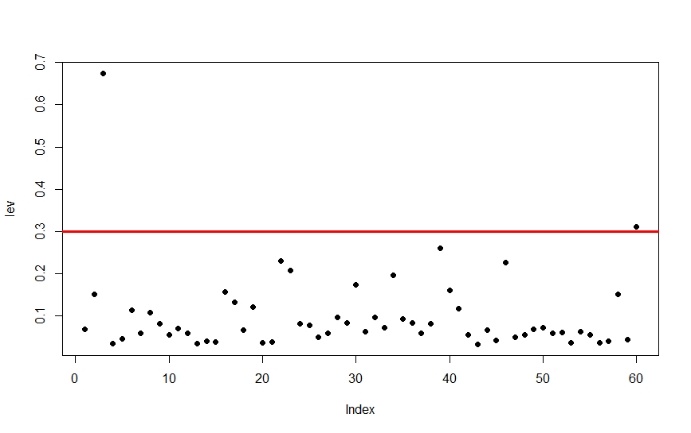
**Code:**

*lev=hat(model.matrix(reg.out2))*

*plot(lev,pch=19)*

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

**Output:**



From the above output, we can interpret that, there are few (2) high leverages that are affecting the model of the regression which are greater than three times the mean leverage.

2 observations that have a leverage higher than 3 times the mean leverage.

**Code:**

*high.leverage.points = ruralsample[lev>(3\*mean(lev)),]*

*high.leverage.points*

**Output:**

id county state area poptotal popdensity popwhite popblack popasian popadult popchild percollege perprof

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

1 248 MECOS~ MI 556. 37308 67.1 35739 978 187 19005 18303 25.0 7.85

2 222 GRAND~ MI 465. 64273 138. 63019 259 318 41094 23179 31.0 7.03

|  |  |
| --- | --- |
| County | State |
| MECOS | MI |
| GRAND | MI |

**5.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**

**Solution:**

**Code:**

*reg.outrural = lm(ruralsample$perelderlypoverty ~ ruralsample$area+*

*ruralsample$popdensity + ruralsample$percollege + ruralsample$childadultratio + ruralsample$popchildpoverty,*

*data = metrosample)*

*summary(reg.outrural)*

**Output:**

Call:

lm(formula = ruralsample$perelderlypoverty ~ ruralsample$area +

ruralsample$popdensity + ruralsample$percollege + ruralsample$childadultratio +

ruralsample$popchildpoverty, data = metrosample)

Residuals:

Min 1Q Median 3Q Max

-5.1940 -1.7424 -0.4762 1.7286 7.4375

Coefficients:

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

(Intercept) 27.0422323 3.2284838 8.376 2.45e-11 \*\*\*

ruralsample$area -0.0068883 0.0018499 -3.724 0.000470 \*\*\*

ruralsample$popdensity -0.0885992 0.0180498 -4.909 8.85e-06 \*\*\*

ruralsample$percollege -0.2365861 0.1092927 -2.165 0.034849 \*

ruralsample$childadultratio -8.5278151 5.5404251 -1.539 0.129596

ruralsample$popchildpoverty 0.0017673 0.0004882 3.620 0.000651 \*\*\*

---

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

Residual standard error: 2.812 on 54 degrees of freedom

Multiple R-squared: 0.5179, Adjusted R-squared: 0.4732

F-statistic: 11.6 on 5 and 54 DF, p-value: 1.232e-07

**Conclusion:**

I believe that the best fit model we used on **“sampleruraldata”,** when used on **“samplemetrodata”,** has really shown almost the same results. Multiple R-squared and Adjusted R-squared values are almost same compared to what we got in for **“sampleruraldata”**. Hence, I believe this model is same fit for **“samplemetrodata”** as compared to that of **“sampleruraldata”.**