OutliersDetection_StepwiseRegression

Marta Fajlhauer

18 April 2018

Outliers detection, Model preparation and Stepwise regression.

Univariate outliers detection

The data used here comes from UCI Machine Learning repository and is the same that I have used in the Missing Values Imputation Methods, so I already finished the first process of cleaning messy data. The next step is outliers detection that is very important to detect any unusual observations that can bias our result. We could argue that the explanatory variables are real-life cases coming from an observational study and they should not be treated as outliers or that the explanatory variables have high leverages rather than outliers but in this case, variables are coming from the population we might think should be normally distributed. From the dataset description, we know that: *all values more than 3 SD above the mean are normalized to 1.00; all values more than 3 SD below the mean are normalized to 0.00*[1] . Due to the decision of removing all observations that were 0 or 1 we expect that the problem with outliers is minimalised.

There may be many reasons to detect outliers including:

- It can mislead us with pattern interpretation in the plot.
- Elimination of outliers from a regression model can bring a different result. Observations that are both outliers and high leverage impact on both the slopes and intercept model.

We can also recognise different types of unusual Observations:

- **Univariate outlier** is not necessarily a regression outlier, but it has some observations that are unusual in one of the variables in the dataset
- **Regression outliers** If we get variable Y conditional on X there may be some unusual values but will have large residual, however, may not affect the coefficient slope.
- **high leverage values** are values with unusual x-value. The more it differs from the mean of X the more leverage has on the regression fit.
- **Influential observation** is the observation that has high leverage and is an outlier in the term Y conditional on X.

The first method of univariate outlier detection that I decided to use is based on F-statistics from ANOVA table. Different states may have different tendencies, like different population rate, urban areas, poverty level, ethnicity and so on. Looking at all states as one country and looking for outliers in this way may consequence in excluding full state that has unusual tendencies comparing to the USA as a whole or to removing important observations to that particular state. To minimalise this problem I have decided to test first how the distribution of numerical variables differ by categorical variable: states using ANOVA. This is important as big F-value suggest that there is a big difference in the mean between the states so outliers for given states are not necessarily outliers at all. The variables with high F-test >10 are:

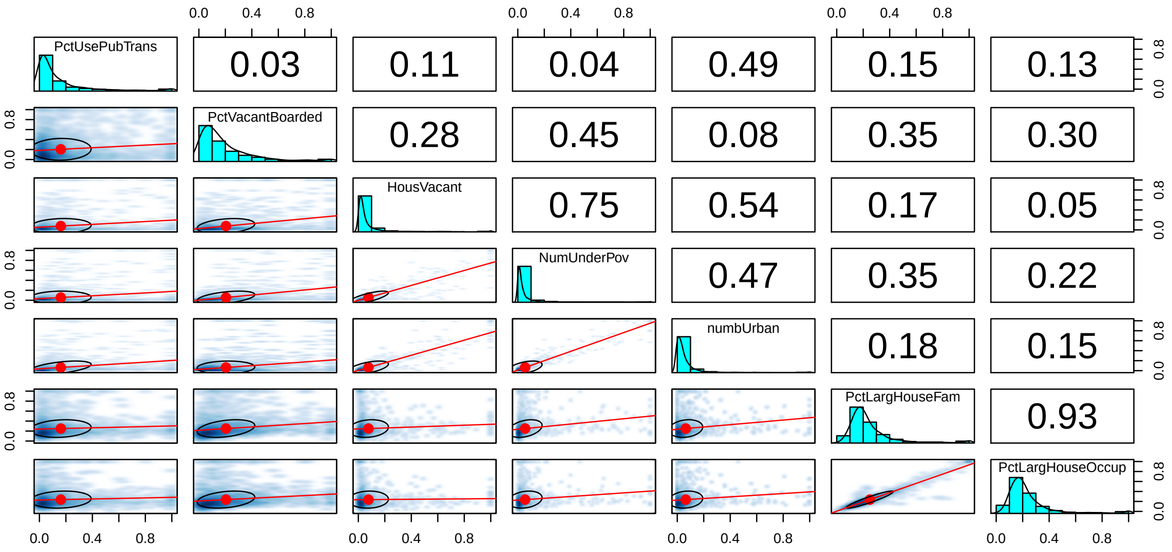
```
   StatsF                Vars
1  10.015          HispPerCap
2  10.319       PctSameCity85
3  10.681          pctWRetire
4  10.742      PctPersOwnOccup
5  11.228        PctImmigRec10
6  12.572          PctTeen2Par
7  12.647   ViolentCrimesPerPop
```

```
8  12.784           PctEmplManu
9  12.871               PopDens
10 12.900      PctYoungKids2Par
11 12.914           blackPerCap
12 13.008              PctIlleg
13 13.300           whitePerCap
14 13.757            PctWorkMom
15 13.852     MedRentPctHousInc
16 13.873   PctWorkMomYoungKids
17 14.286             perCapInc
18 14.758      PersPerOccupHous
19 15.851           PctKids2Par
20 16.099     PersPerOwnOccHous
21 16.763        MalePctDivorce
22 16.775            PersPerFam
23 19.154        MedYrHousBuilt
24 19.190        PctPopUnderPov
25 19.381            pctWInvInc
26 20.009        PctSameState85
27 20.013          pctWFarmSelf
28 21.609          PctHousOccup
29 21.914             medIncome
30 22.196        PctVacMore6Mos
31 22.311   PctNotSpeakEnglWell
32 22.800          racePctWhite
33 22.869    PersPerRentOccHous
34 23.247       PctLargHouseFam
35 23.326     PctLargHouseOccup
36 23.846          FemalePctDiv
37 26.623        PctRecentImmig
38 26.907        PctSameHouse85
39 27.725          racepctblack
40 29.544          PctRecImmig5
41 30.173        PctHousNoPhone
42 31.950      PctSpeakEnglOnly
43 32.465      PctPersDenseHous
44 32.487          PctRecImmig8
45 32.487          PctRecImmig8
46 33.787         PctRecImmig10
47 34.465           racePctHisp
48 38.622          racePctAsian
49 45.082        PctForeignBorn
50 47.706 MedOwnCostPctIncNoMtg
51 53.166               MedRent
52 57.106          OwnOccMedVal
53 65.377      PctBornSameState
54 68.629      MedOwnCostPctInc
```

For those variables with large F-test, I remove outliers for different states and use mean substitution method by calculating mean for this particular state and substituting missing value in the state with the mean value for this state.

I decided to use Tukey method for outliers detection for rest of the variables that the method can improve data significantly. Tukey method uses quartiles which are resistant to extreme values. The method assumes that a value between the inner and outer fences in boxplot is a possible outlier and an extreme value beyond the outer fences is a probable outlier. The Tukey's method is good for skewed data as it makes no assumptions based on distribution and it does not depend on a mean or standard deviation. The drawback is that although the boxplot may be applicable for both symmetric and skewed data, the more skewed the data, the more observations may be detected as outliers. This results from the fact that this method is based on robust measures such as lower and upper quartiles and the IQR without considering the skewness of the data. Hence, the decision of removing outliers using Tukey method is based on 2 cases: Case 1 is when the total number of removed observations is not greater than 10% and case 2 is when there is more than 10% of missing values but outliers are nicely cleaned without too many repetitions of outlier detection.

There were some variables such as: "PctUsePubTrans", "PctVacantBoarded", "HousVacant", "NumUnderPov", 'numbUrban', 'PctLargHouseFam', 'PctLargHouseOccup' that Tukey method outlier detection didn't work well. For these observations, I have decided to plot scatterplot and density plots to understand them further.



There is a strong correlation between PctLargHouseFam and PctLargHouseOccup (0.92). As PctLargHouseFam is a subset of PctLargHouseOccup decision of removing the first one where the second one is cleaned already. NumbUrban and population are hard to specify as outliers as different places can have different urbanisation rate. For example, small villages have small urbanisation and population rate where big cities will have big. Without knowledge of the geographic area, we cannot say if the specific record is an outlier or not hence the decision of leaving the variables without cleaning. From the metadata, we know that the variables:

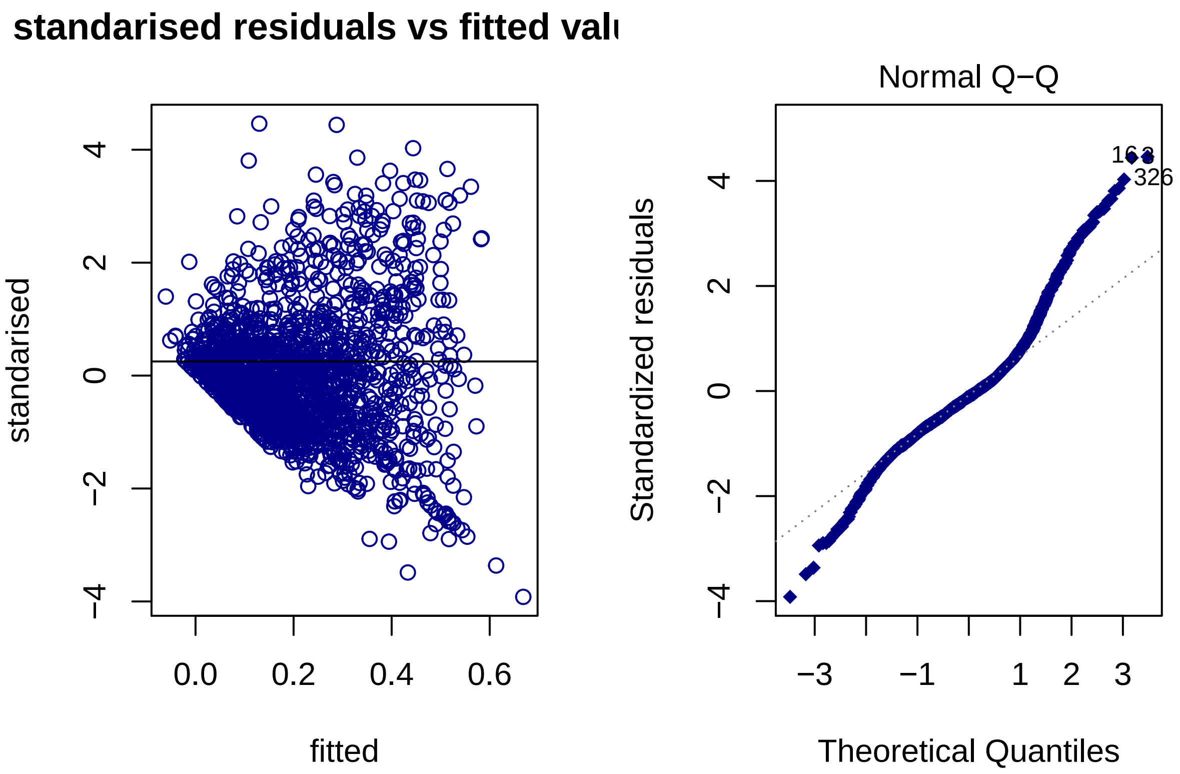
- PctUsePubTrans: percent of people using public transit for commuting.
- PctVacantBoarded: percent of vacant housing that is boarded up.
- HousVacant: number of vacant households.
- NumUnderPov: number of people under the poverty level (numeric - decimal).

As the first 3 (HousVacant, PctVacantBoarded and PctUsePubTrans) may represent poverty level in the population I have decided to remove them and leave variable NumUnderPov.

Multivariate outliers detection.

After cleaning outliers in univariate case I will concentrate on cleaning multivariate outliers. Before doing that I decided to decode comunitynames and states variables on real geographic regions using Google and GoogleMap and plot in geospatial position on a map. To perform Multivariate outliers detection first I need to fit a linear model (ordinary least squares) for all numerical variables in the dataset. We are simply putting everything into one huge model to check the most important variables, colinearity, normality assumption of the model and so on. From graphs produced we can find that:

- The graph for Residuals vs fitted shows that there is an increase of variance within the residuals which suggests the problem with heteroscedasticity and variables 3, 16 and 326 are potential outliers.
- Normal QQ plot shows that the data is heavily skewed and we should not preserve normality assumption.



All of these suggests that there are many potential multivariate outliers and we need to clean the data from them.

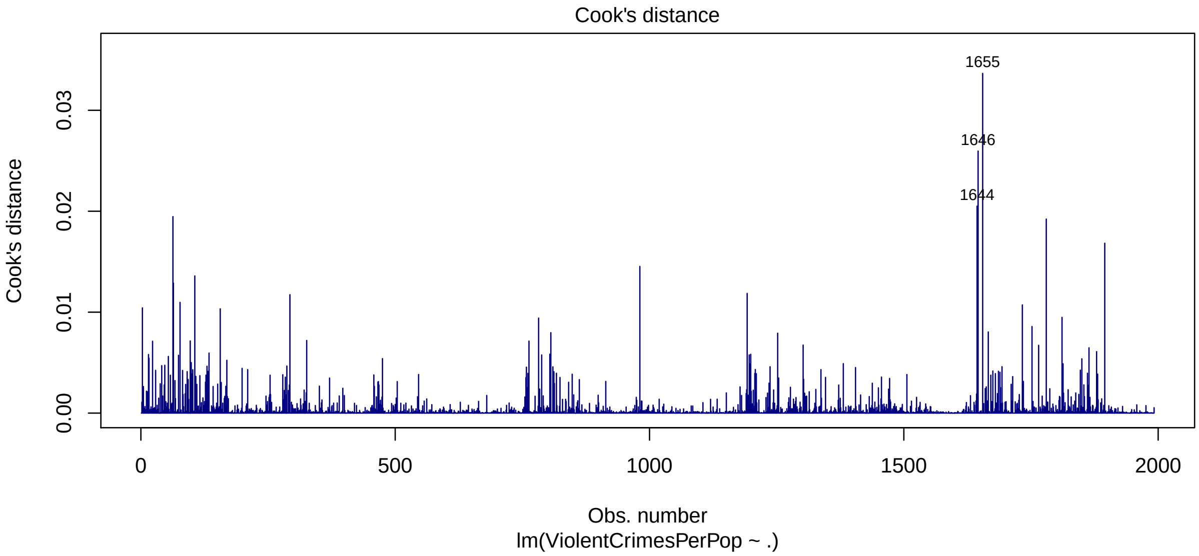
From the produced OLS result is:

```
Call:
lm(formula = ViolentCrimesPerPop ~ ., data = CrimeData[, c(-1,
    -2, -3, -4)])

Residuals:
     Min       1Q   Median       3Q      Max
-0.44648 -0.06690 -0.01241  0.04840  0.52000

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.5280790  0.1049241   5.033 5.28e-07 ***
population             0.0107442  0.0535420   0.201 0.840979
householdsize          0.0115247  0.0443373   0.260 0.794945
racepctblack           0.0900084  0.0257064   3.501 0.000473 ***
racePctWhite          -0.0620623  0.0278296  -2.230 0.025858 *
```

```
racePctAsian          0.0026793  0.0283991   0.094 0.924845
racePctHisp           0.0134615  0.0359580   0.374 0.708174
agePct12t21           0.0872148  0.0491631   1.774 0.076225 .
agePct12t29          -0.0407327  0.0638847  -0.638 0.523813
agePct16t24          -0.0481621  0.0753436  -0.639 0.522748
agePct65up            0.0244164  0.0485574   0.503 0.615138
numbUrban             0.0160976  0.0357431   0.450 0.652495
medIncome             0.0233290  0.0403384   0.578 0.563109
pctWWage              0.0711707  0.0531108   1.340 0.180391
pctWFarmSelf          0.0106819  0.0206731   0.517 0.605421
pctWInvInc           -0.1544048  0.0372046  -4.150 3.47e-05 ***
pctWSocSec            0.0923468  0.0595105   1.552 0.120883
pctWPubAsst           0.0720064  0.0290288   2.481 0.013205 *
pctWRetire            0.0018970  0.0267437   0.071 0.943458
perCapInc             0.0296941  0.0654850   0.453 0.650277
whitePerCap           0.0323499  0.0621787   0.520 0.602934
blackPerCap          -0.0328804  0.0303808  -1.082 0.279267
indianPerCap         -0.0209408  0.0227782  -0.919 0.358037
AsianPerCap          -0.0039927  0.0273234  -0.146 0.883835
OtherPerCap           0.0404851  0.0249735   1.621 0.105156
HispPerCap            0.0038909  0.0248171   0.157 0.875432
NumUnderPov           0.0422339  0.0529634   0.797 0.425308
PctPopUnderPov        0.0133902  0.0313584   0.427 0.669424
PctLess9thGrade      -0.0163198  0.0348683  -0.468 0.639809
PctNotHSGrad         -0.0086603  0.0389415  -0.222 0.824032
PctBSorMore          -0.0326014  0.0426173  -0.765 0.444377
PctUnemployed        -0.0192619  0.0282851  -0.681 0.495959
PctEmploy             0.0632254  0.0425905   1.484 0.137843
PctEmplManu          -0.0343755  0.0224905  -1.528 0.126567
PctEmplProfServ      -0.0190075  0.0281465  -0.675 0.499563
PctOccupManu         -0.0025524  0.0357126  -0.071 0.943030
PctOccupMgmtProf     -0.0204289  0.0440425  -0.464 0.642811
MalePctDivorce       -0.0210606  0.0401234  -0.525 0.599716
MalePctNevMarr        0.0101602  0.0357168   0.284 0.776085
FemalePctDiv         -0.0036276  0.0448050  -0.081 0.935478
PersPerFam           -0.0843212  0.0595086  -1.417 0.156659
PctKids2Par          -0.2256539  0.0432328  -5.220 1.99e-07 ***
PctYoungKids2Par      0.0401947  0.0311199   1.292 0.196650
PctTeen2Par          -0.0663565  0.0314775  -2.108 0.035156 *
PctWorkMomYoungKids   0.0108238  0.0329317   0.329 0.742438
PctWorkMom           -0.0444670  0.0336360  -1.322 0.186324
NumIlleg              0.0060716  0.0314258   0.193 0.846819
PctIlleg              0.1847377  0.0309383   5.971 2.80e-09 ***
NumImmig             -0.0482988  0.0225232  -2.144 0.032127 *
PctImmigRecent        0.0444251  0.0290523   1.529 0.126395
PctImmigRec5         -0.0216919  0.0343749  -0.631 0.528090
PctImmigRec8          0.0257389  0.0415749   0.619 0.535927
PctImmigRec10        -0.0541855  0.0312398  -1.735 0.082991 .
PctRecentImmig       -0.0113748  0.0461865  -0.246 0.805493
PctRecImmig5          0.0261465  0.0604594   0.432 0.665453
PctRecImmig8          0.0051372  0.0633552   0.081 0.935382
PctRecImmig10        -0.0231475  0.0523003  -0.443 0.658114
PctSpeakEnglOnly     -0.0582034  0.0404985  -1.437 0.150833
PctNotSpeakEnglWell   0.0988044  0.0466508   2.118 0.034308 *
```

```
PctLargHouseOccup     -0.0079749  0.0505659  -0.158 0.874699
PersPerOccupHous       0.0780510  0.0579109   1.348 0.177890
PersPerOwnOccHous     -0.0594206  0.0487281  -1.219 0.222832
PersPerRentOccHous    -0.1041858  0.0336155  -3.099 0.001968 **
PctPersOwnOccup       -0.0689765  0.0483205  -1.427 0.153605
PctPersDenseHous       0.0584925  0.0408000   1.434 0.151839
PctHousLess3BR        -0.0196342  0.0323844  -0.606 0.544398
PctHousOccup          -0.0954194  0.0274303  -3.479 0.000515 ***
PctHousOwnOcc          0.0854575  0.0502288   1.701 0.089037 .
PctVacMore6Mos        -0.0735618  0.0213600  -3.444 0.000586 ***
MedYrHousBuilt         0.0091041  0.0216676   0.420 0.674407
PctHousNoPhone         0.0646698  0.0244153   2.649 0.008146 **
PctWOFullPlumb        -0.0175267  0.0273338  -0.641 0.521463
OwnOccMedVal           0.0372283  0.0327904   1.135 0.256374
MedRent                0.0122735  0.0338004   0.363 0.716557
MedRentPctHousInc      0.0226964  0.0271748   0.835 0.403710
MedOwnCostPctInc       0.0353104  0.0258296   1.367 0.171770
MedOwnCostPctIncNoMtg -0.0867387  0.0215409  -4.027 5.88e-05 ***
PctForeignBorn        -0.0673042  0.0387760  -1.736 0.082775 .
PctBornSameState      -0.0529976  0.0279766  -1.894 0.058330 .
PctSameHouse85         0.0295917  0.0370013   0.800 0.423957
PctSameCity85         -0.0068775  0.0281227  -0.245 0.806829
PctSameState85         0.0037908  0.0318099   0.119 0.905153
LandArea               0.1641585  0.1097774   1.495 0.134981
PopDens                0.0003491  0.0255522   0.014 0.989101
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1179 on 1908 degrees of freedom
Multiple R-squared:  0.5528,    Adjusted R-squared:  0.5333
F-statistic: 28.42 on 83 and 1908 DF,  p-value: < 2.2e-16
```

we know that multiple regression r-sq is 0.5528 so only 55% of the variability in the data is explained by the model. We can produce Cook,s distance graph to see values with the highest Cook value in comparison to their observations and to decide if we should keep the observation or remove them from our dataset. The highest Cook's distance for 1644, 1648, 1655.



Cook's distance depends on the size of standarised residuals and leverage. *Definition of the Cook's distance for the general linear model is given by*

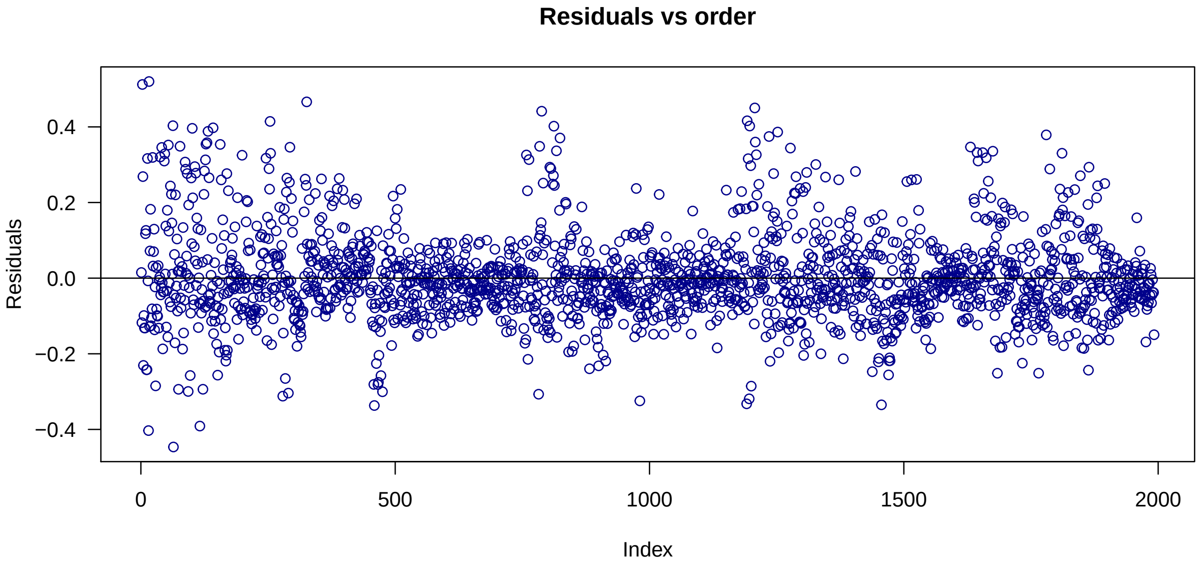
$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T (X^T X)(\hat{\beta} - \hat{\beta}_{(i)})}{ps^2}$$

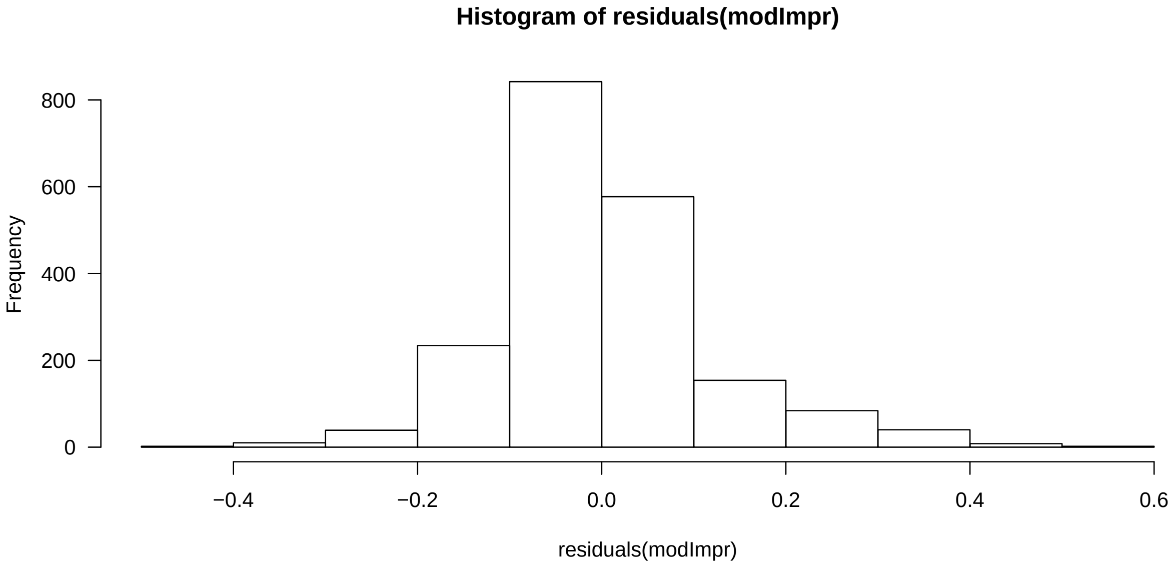
. This can be thought of as a scaled distance between $\hat{\beta}$ and $\hat{\beta}_{(i)}$ which is the estimate of β omiting the ith observation. [2]

On the plot, we can see that the spikes for 1653 and 1644 differ in height from other data points however for the observation 1642 the spike is a similar height to the rest of the data points so it is not an influential observation. These observations are:

```
       state communityname
1 California     San Dimas
2 California  Santa Monica
```

We can plot residual plot pattern to see if the residuals are uncorrelated as any pattern in the residual plot suggests otherwise.





We can check using VIF if there is a problem with multicollinearity in the model. Since there are variables with VIF larger than 10 then we have the problem with multicollinearity. Below are attached only the variables with VIF more than 10.

```
  VIF.pctWWage VIF.pctWSocSec VIF.perCapInc VIF.whitePerCap
1     11.80939       13.82963      13.69074        11.27949
  VIF.PctRecImmig5 VIF.PctRecImmig8 VIF.PctRecImmig10 VIF.PersPerOccupHous
1         14.58736         16.45493          11.65711             10.16978
  VIF.PctPersOwnOccup
1            11.55608
```

To finalise we can look for the best model that can be fitted using backwards and forward elimination as well as the combination of both in the stepwise regression. In the backward elimination first we are fitting all variables in the model and we are taking into account the smallest value of AIC that accounts for the ranking what would happen if we take one variable out. The lower the AIC value is the better the model is. We are taking out one variable at the time as if you will take one variable out the variable that was colinear to it may have changed behaviour a bit. The forward substitution regression model selection works in a similar way but first, we are taking into consideration the null model. The null model that does not have any explanatory variables so we are starting only with the mean of violent crimes per population. We add one variable at the time with the smallest AIC that is lower than AIC of the model that we start with. Backward and forward elimination shows us, different models. In stepwise regression, we start with the null model that has only intercept. We are adding or removing variable at each step depending on the AIC result. Full models of backwards, forward and stepwise regression are:

```
Call:
lm(formula = ViolentCrimesPerPop ~ racepctblack + racePctWhite +
    agePct12t21 + pctWWage + pctWInvInc + pctWSocSec + pctWPubAsst +
    OtherPerCap + NumUnderPov + PctEmploy + PctEmplManu + PersPerFam +
    PctKids2Par + PctTeen2Par + PctWorkMom + PctIlleg + NumImmig +
    PctImmigRecent + PctImmigRec10 + PctSpeakEnglOnly + PctNotSpeakEnglWell +
    PersPerRentOccHous + PctPersDenseHous + PctHousOccup + PctHousOwnOcc +
    PctVacMore6Mos + PctHousNoPhone + OwnOccMedVal + MedOwnCostPctInc +
    MedOwnCostPctIncNoMtg + PctForeignBorn + PctBornSameState,
    data = CrimeData[, c(-1, -2, -3, -4)])

Residuals:
     Min       1Q   Median       3Q      Max 
-0.44267 -0.06808 -0.01192  0.04734  0.52884 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            0.46166    0.07193   6.418 1.73e-10 ***
racepctblack           0.09511    0.02411   3.945 8.26e-05 ***
racePctWhite          -0.07407    0.02517  -2.943 0.003291 ** 
agePct12t21            0.06166    0.03839   1.606 0.108393    
pctWWage               0.07527    0.04784   1.573 0.115846    
pctWInvInc            -0.13415    0.02909  -4.612 4.25e-06 ***
pctWSocSec             0.12040    0.04513   2.668 0.007698 ** 
pctWPubAsst            0.05708    0.02578   2.214 0.026955 *  
OtherPerCap            0.04363    0.02362   1.847 0.064892 .  
NumUnderPov            0.05545    0.02291   2.420 0.015603 *  
PctEmploy              0.07588    0.03466   2.189 0.028716 *  
PctEmplManu           -0.03729    0.01588  -2.348 0.018956 *  
PersPerFam            -0.07817    0.03670  -2.130 0.033308 *  
PctKids2Par           -0.20345    0.03728  -5.457 5.46e-08 ***
PctTeen2Par           -0.05998    0.03019  -1.986 0.047120 *  
PctWorkMom            -0.05099    0.02022  -2.522 0.011734 *  
PctIlleg               0.18351    0.02926   6.273 4.36e-10 ***
NumImmig              -0.04960    0.02087  -2.376 0.017588 *  
PctImmigRecent         0.04247    0.02439   1.741 0.081784 .  
PctImmigRec10         -0.04870    0.02080  -2.341 0.019328 *  
PctSpeakEnglOnly      -0.06011    0.03666  -1.640 0.101232    
PctNotSpeakEnglWell    0.10007    0.04253   2.353 0.018728 *  
PersPerRentOccHous    -0.07354    0.02778  -2.647 0.008175 ** 
PctPersDenseHous       0.06117    0.03488   1.754 0.079655 .  
PctHousOccup          -0.08923    0.02467  -3.616 0.000306 ***
PctHousOwnOcc          0.08699    0.02408   3.612 0.000311 ***
PctVacMore6Mos        -0.07201    0.01983  -3.631 0.000290 ***
```

```
PctHousNoPhone         0.06279    0.02132   2.945 0.003267 **
OwnOccMedVal           0.06804    0.02591   2.626 0.008714 **
MedOwnCostPctInc       0.04643    0.02226   2.085 0.037153 *
MedOwnCostPctIncNoMtg -0.09191    0.01987  -4.626 3.97e-06 ***
PctForeignBorn        -0.06285    0.03046  -2.063 0.039242 *
PctBornSameState      -0.05409    0.01892  -2.859 0.004296 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1171 on 1959 degrees of freedom
Multiple R-squared:  0.5473,    Adjusted R-squared:  0.5399
F-statistic: 74.01 on 32 and 1959 DF,  p-value: < 2.2e-16


Call:
lm(formula = ViolentCrimesPerPop ~ PctKids2Par + racePctWhite +
    PctBornSameState + PctIlleg + PctPersDenseHous + PctHousOccup +
    PctVacMore6Mos + pctWInvInc + PctHousOwnOcc + perCapInc +
    MedOwnCostPctIncNoMtg + PctSpeakEnglOnly + racepctblack +
    MedRentPctHousInc + PersPerRentOccHous + NumImmig + NumUnderPov +
    PctHousNoPhone + PctImmigRec10 + OwnOccMedVal + PctImmigRecent +
    PctTeen2Par + PctForeignBorn + PctNotSpeakEnglWell + OtherPerCap +
    PctEmplManu + PctBSorMore + LandArea + PersPerFam + pctWPubAsst +
    PersPerOccupHous + PctUnemployed, data = CrimeData[, c(-1,
    -2, -3, -4)])

Residuals:
     Min       1Q   Median       3Q      Max
-0.45242 -0.06875 -0.01186  0.04708  0.52989

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.59391    0.05815  10.213  < 2e-16 ***
PctKids2Par           -0.19872    0.03720  -5.342 1.02e-07 ***
racePctWhite          -0.07381    0.02499  -2.953 0.003185 **
PctBornSameState      -0.05564    0.01852  -3.004 0.002697 **
PctIlleg               0.18747    0.02934   6.390 2.06e-10 ***
PctPersDenseHous       0.05165    0.03480   1.484 0.137901
PctHousOccup          -0.08505    0.02493  -3.411 0.000660 ***
PctVacMore6Mos        -0.07004    0.01967  -3.561 0.000379 ***
pctWInvInc            -0.15208    0.03175  -4.790 1.79e-06 ***
PctHousOwnOcc          0.07097    0.02503   2.835 0.004625 **
perCapInc              0.06363    0.03465   1.836 0.066438 .
MedOwnCostPctIncNoMtg -0.07723    0.01895  -4.075 4.79e-05 ***
PctSpeakEnglOnly      -0.06718    0.03665  -1.833 0.066975 .
racepctblack           0.09193    0.02415   3.806 0.000145 ***
MedRentPctHousInc      0.03077    0.02348   1.310 0.190182
PersPerRentOccHous    -0.08306    0.02875  -2.889 0.003903 **
NumImmig              -0.04222    0.02099  -2.012 0.044386 *
NumUnderPov            0.05095    0.02296   2.219 0.026572 *
PctHousNoPhone         0.05788    0.02085   2.777 0.005546 **
PctImmigRec10         -0.05137    0.02075  -2.475 0.013391 *
OwnOccMedVal           0.07078    0.02727   2.596 0.009504 **
PctImmigRecent         0.04482    0.02423   1.850 0.064430 .
```

```
PctTeen2Par          -0.05511    0.03027  -1.820 0.068869 .
PctForeignBorn       -0.06595    0.03055  -2.159 0.030968 *
PctNotSpeakEnglWell   0.10411    0.04270   2.438 0.014857 *
OtherPerCap           0.04513    0.02361   1.911 0.056118 .
PctEmplManu          -0.03784    0.01630  -2.321 0.020385 *
PctBSorMore          -0.04345    0.02658  -1.635 0.102216
LandArea              0.15269    0.10141   1.506 0.132289
PersPerFam           -0.08942    0.04786  -1.868 0.061865 .
pctWPubAsst           0.06055    0.02636   2.297 0.021729 *
PersPerOccupHous      0.06098    0.04556   1.338 0.180962
PctUnemployed        -0.03992    0.02584  -1.545 0.122625
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1172 on 1959 degrees of freedom
Multiple R-squared:  0.5462,    Adjusted R-squared:  0.5388
F-statistic: 73.68 on 32 and 1959 DF,  p-value: < 2.2e-16


Call:
lm(formula = ViolentCrimesPerPop ~ PctKids2Par + racePctWhite +
    PctBornSameState + PctIlleg + PctPersDenseHous + PctHousOccup +
    PctVacMore6Mos + pctWInvInc + PctHousOwnOcc + perCapInc +
    MedOwnCostPctIncNoMtg + PctSpeakEnglOnly + racepctblack +
    PersPerRentOccHous + NumImmig + NumUnderPov + PctHousNoPhone +
    PctImmigRec10 + OwnOccMedVal + PctImmigRecent + PctTeen2Par +
    PctForeignBorn + PctNotSpeakEnglWell + OtherPerCap + PctEmplManu +
    PctBSorMore + PersPerFam + pctWPubAsst + PersPerOccupHous +
    MedOwnCostPctInc, data = CrimeData[, c(-1, -2, -3, -4)])

Residuals:
     Min       1Q   Median       3Q      Max
-0.44846 -0.06849 -0.01177  0.04877  0.52960

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.57201    0.05752   9.945  < 2e-16 ***
PctKids2Par          -0.19983    0.03694  -5.409 7.11e-08 ***
racePctWhite         -0.07234    0.02502  -2.891 0.003881 **
PctBornSameState     -0.05266    0.01872  -2.813 0.004962 **
PctIlleg              0.18707    0.02924   6.398 1.96e-10 ***
PctPersDenseHous      0.05515    0.03473   1.588 0.112510
PctHousOccup         -0.08258    0.02472  -3.340 0.000853 ***
PctVacMore6Mos       -0.07059    0.01961  -3.600 0.000326 ***
pctWInvInc           -0.13977    0.03186  -4.387 1.21e-05 ***
PctHousOwnOcc         0.07877    0.02463   3.198 0.001405 **
perCapInc             0.06432    0.03423   1.879 0.060384 .
MedOwnCostPctIncNoMtg -0.08431    0.01928  -4.373 1.29e-05 ***
PctSpeakEnglOnly     -0.06238    0.03667  -1.701 0.089099 .
racepctblack          0.09567    0.02400   3.986 6.98e-05 ***
PersPerRentOccHous   -0.09005    0.02898  -3.107 0.001917 **
NumImmig             -0.04492    0.02097  -2.142 0.032294 *
NumUnderPov           0.05446    0.02279   2.390 0.016954 *
PctHousNoPhone        0.06202    0.02108   2.942 0.003304 **
```

```
PctImmigRec10        -0.05024    0.02075  -2.420 0.015593 *
OwnOccMedVal          0.06071    0.02815   2.157 0.031141 *
PctImmigRecent        0.04714    0.02420   1.948 0.051551 .
PctTeen2Par          -0.05524    0.03023  -1.828 0.067774 .
PctForeignBorn       -0.07155    0.03058  -2.340 0.019372 *
PctNotSpeakEnglWell   0.10468    0.04263   2.456 0.014154 *
OtherPerCap           0.04491    0.02358   1.905 0.056945 .
PctEmplManu          -0.03838    0.01600  -2.398 0.016564 *
PctBSorMore          -0.03761    0.02639  -1.425 0.154352
PersPerFam           -0.09822    0.04775  -2.057 0.039831 *
pctWPubAsst           0.05417    0.02393   2.264 0.023671 *
PersPerOccupHous      0.06604    0.04553   1.451 0.147072
MedOwnCostPctInc      0.04092    0.02178   1.879 0.060435 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1172 on 1961 degrees of freedom
Multiple R-squared:  0.5456,    Adjusted R-squared:  0.5387
F-statistic:  78.5 on 30 and 1961 DF,  p-value: < 2.2e-16
```

The best substitution method is the one with the lowest AIC. In this case: Backwards(AIC=-8513.62) < Combination (AIC=-8510.48) < Forward(AIC = -8508.92). So the best selection method for the stepwise regression will be backwards selection.

We can check for the VIF of the selected models of the stepwise regression by seeing range of the VIF and what variables they are for: backwise substitution

```
[1] 1.106790 9.719698
```

```
$pctWWage
[1] 9.719698
```

```
$NumImmig
[1] 1.10679
```

Forward substituion:

```
[1] 1.107034 7.238674
```

```
$PctKids2Par
[1] 7.238674
```

```
$LandArea
[1] 1.107034
```

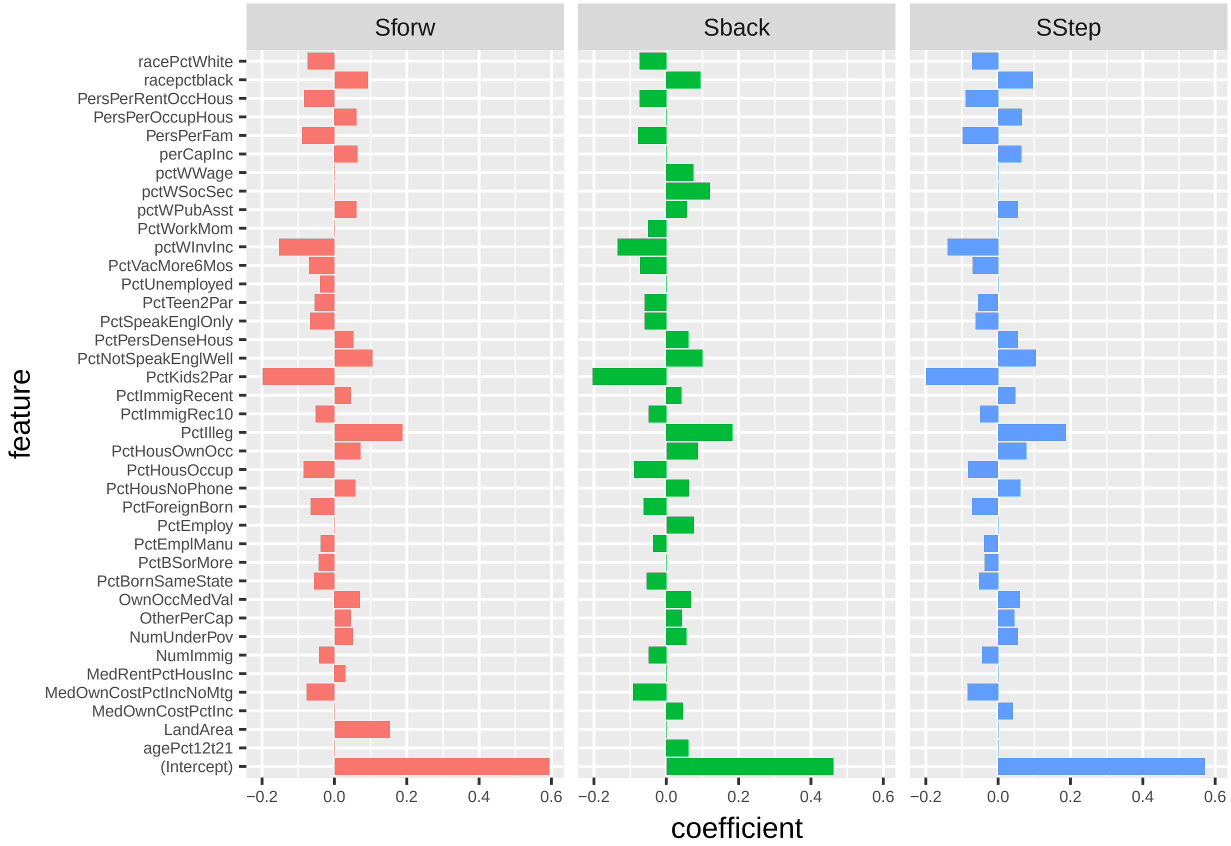
and Combination of forward and backwards:

```
[1] 1.113762 7.138986
```

```
$PctKids2Par
[1] 7.138986
```

```
$NumImmig
[1] 1.113762
```

No variables with higher than 10 VIF hence the problem with multicolinearity has been reduced. However, we can show the graph importance of features for the best models of different approach

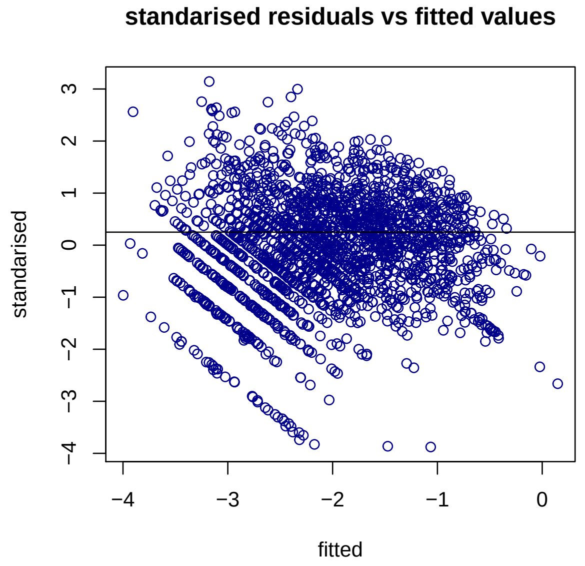


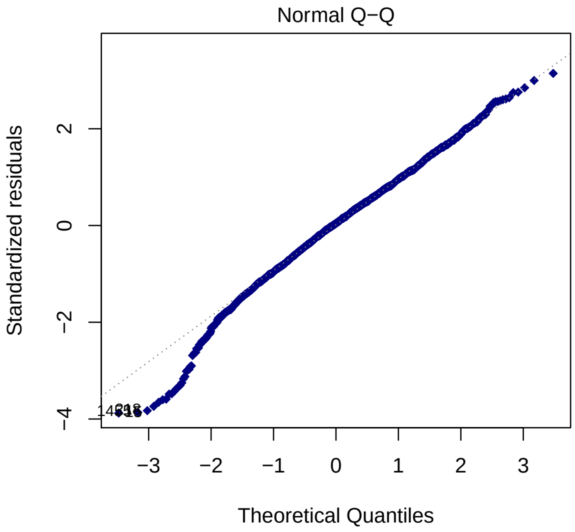
And we can compare the important measurements of the output such as R-squared, R-squared Adjusted, F-statistics and Residuals Standard error

```
      Method                 Output     Values
1  Backwards StandardErrorResiduals  0.1170512
2  Backwards              R-squared  0.5472802
3  Backwards           R-squaredAdj  0.5398851
4  Backwards            Fstatistics 74.0056295
5    Forward StandardErrorResiduals  0.1171919
6    Forward              R-squared  0.5461911
7    Forward           R-squaredAdj  0.5387782
8    Forward            Fstatistics 73.6810947
9   Stepwise StandardErrorResiduals  0.1172041
10  Stepwise              R-squared  0.5456335
11  Stepwise           R-squaredAdj  0.5386825
12  Stepwise            Fstatistics 78.4966481
```

However, distribution of residuals seems to be skewed. Problem with not constant variance and non-linearity we may try to overcome using log transformation. Summary of the result after fitting the regression model with logarithmic transformation of the response variable is available after running this code:

```
 CrimeData <- CrimeData[c(-324,-1644, -1648, -1655),]
 modImpr_log <- lm(log(ViolentCrimesPerPop) ~ ., data = CrimeData[,c(-1,-2, -3, -4)])
# summary(modImpr_log) # 0.5827
```

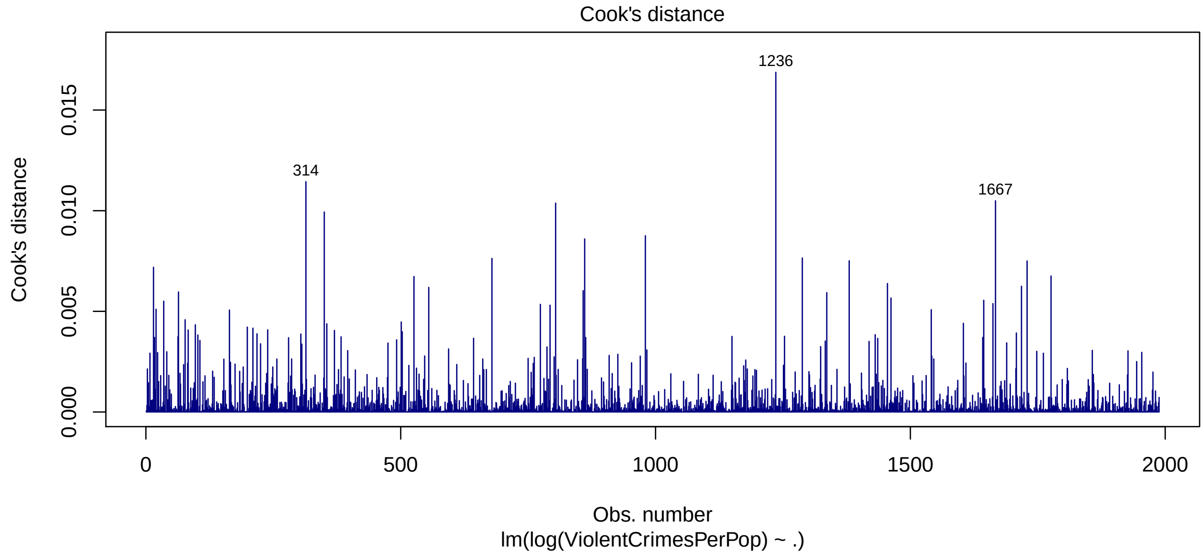




Normality plot and scale-location family plot has improved but the not constant variance problem still remains. To check normality assumption we can run Shapiro - Wilk test of normality. The null hypothesis of the Shapiro Wilk test is that data is normally distributed and we reject the hypothesis of normality if the p-value is less than 0.05

```
    Shapiro-Wilk normality test

data:  CrimeData$ViolentCrimesPerPop
W = 0.84868, p-value < 2.2e-16
```

Very small p-value suggests that the data is not normally distributed.



observation 312, 1234, 1667 have the highest Cook's distance but as for the observations 1234 and 1641 there is a big difference between the level of Cook's distance for those values and of the rest of the data and observation 312 seems to be at the same level of Cook's distance as rest of the data.

There are difficulties with logarithmic transformation approach such as there is just a slight improvement with the r-score and sqrt(MSE). *"In order to evaluate the performance of a statistical learning method on a given dataset, we need some way to measure how well its predictions actually match the observed data. That is, we need to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation. In the regression setting, the most commonly-used measure is the mean squared error (MSE), given by*

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

where ^ f(xi) is the prediction that ^ f gives for the ith observation. The MSE will be small if the predicted responses are very close to the true responses, and will be large if, for some of the observations, the predicted and true responses differ substantially." [3]

The good thing is that normality plot has improved significantly.

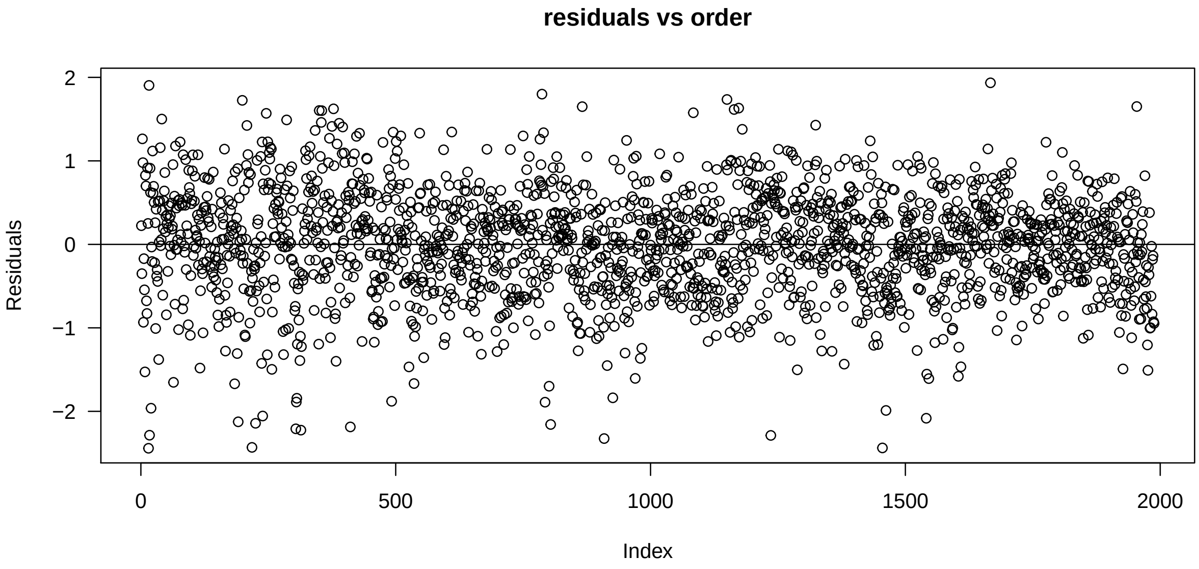
We can recheck standardised residuals plot. If the normal linear model holds, the standardised residuals have approximately standard normal distributions hence approximately 95 percent will be between 2 standard deviations out of the mean and 99.6 within 3 standard deviations. So the absolute value for most of the residuals would be smaller than:

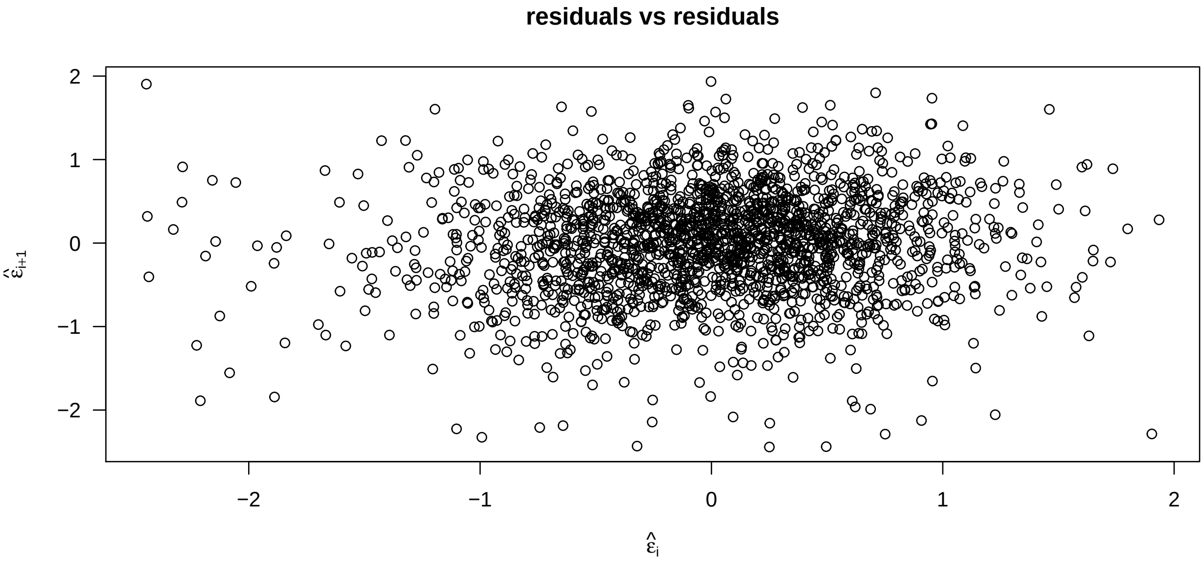
```
[1] 2.914912
```

However, in large data sets like this one, we should not assume that standardised residuals outside 3 standard deviations regime must be outliers – values a little outside can also occur by chance.

The first plot underneath is the plot of residuals vs order. To remind before cleaning multivariate outliers the residuals formed a pattern which now disappeared and the plot shows improvements.

In the second plot, we are plotting on x-axis residual of observation i vs residual of observation i+1 that does not show an obvious problem with the correlation that we recheck by testing VIF. The test brings similar result as before so we still have a problem with collinearity





```
  VIF.pctWWage VIF.pctWSocSec VIF.perCapInc VIF.whitePerCap
1     11.82657       13.84513      14.31453         11.79558
  VIF.PctRecImmig5 VIF.PctRecImmig8 VIF.PctRecImmig10 VIF.PersPerOccupHous
1         15.38708         16.94831          11.68372             10.19607
  VIF.PctPersOwnOccup
1            11.71544
```

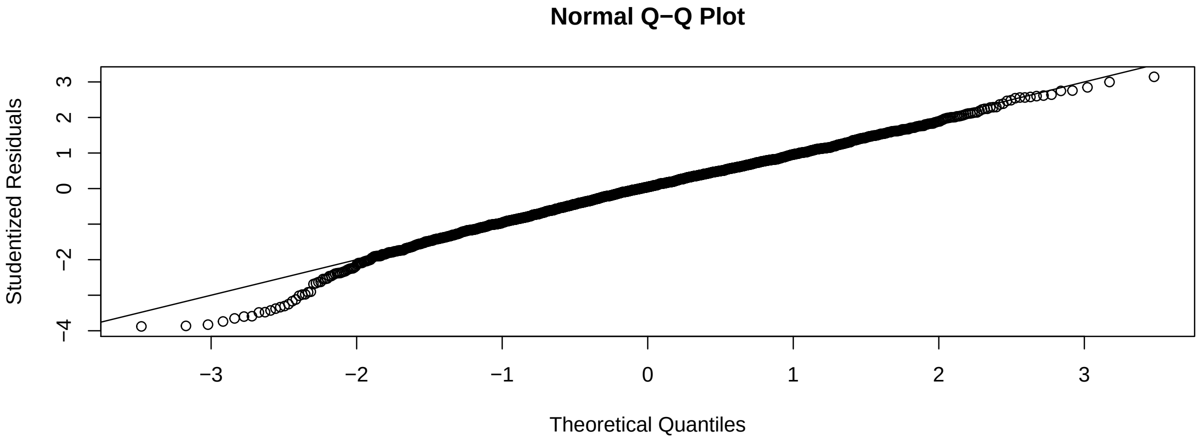
clearly,

- Assumption 1: The Y-values are independent is met
- Assumption 2: The Y-values can be expressed as a linear function of the X-variable is met
- Assumption 3: Variation of observations around the regression line is constant is met. For given values of X, Y values are normally distributed are close to the true value

We can look at leverages now. We regard an observation with 2p/n as high leverage and 3p/n as a very high leverage. There are 83 variables that are going into the regression model so observations with high leverage values are observations for which hat value is greater than 0.08350101 and very high leverage value is 0.1252515. The table below summarises distribution of hat values in the dataset. So we see that there are observations with high and very high leverage values.

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.009348 0.024130 0.033618 0.042254 0.050608 0.269917
```

Studentized residuals show the normal pattern as we would expect points to follow the diagonal line



The summary of stepwise regression models using backwards, forward or combination of both variables selection method after log transformation is included in Appendix 10. In case of choosing regression model before logarithmic transformation of response variable appeared to be backwise selection method. In case of choosing a model after the logarithmic transformation both the backwise and the combination of both are equally good models. Backwise (Step: AIC=-1750.03) = Combination (Step: AIC=-1750.03) = Forward (Step: AIC=-1748.63).

We can compare the important measurements of the output such as R-squared, R-squared Adjusted, F-statistics and Residuals Standard error

```
      Method                 Output     Values
1  Backwards StandardErrorResiduals  0.6381338
2  Backwards              R-squared  0.5753940
3  Backwards           R-squaredAdj  0.5686646
4  Backwards            Fstatistics 85.5039864
5    Forward StandardErrorResiduals  0.6382785
6    Forward              R-squared  0.5752014
7    Forward           R-squaredAdj  0.5684690
8    Forward            Fstatistics 85.4366272
9   Stepwise StandardErrorResiduals  0.6381338
10  Stepwise              R-squared  0.5753940
11  Stepwise           R-squaredAdj  0.5686646
12  Stepwise            Fstatistics 85.5039864
```

And like before we can check for multicolinearity in selected models by checking range of VIF values in backwise, forwards and combination of both seletion methods respecitvely:

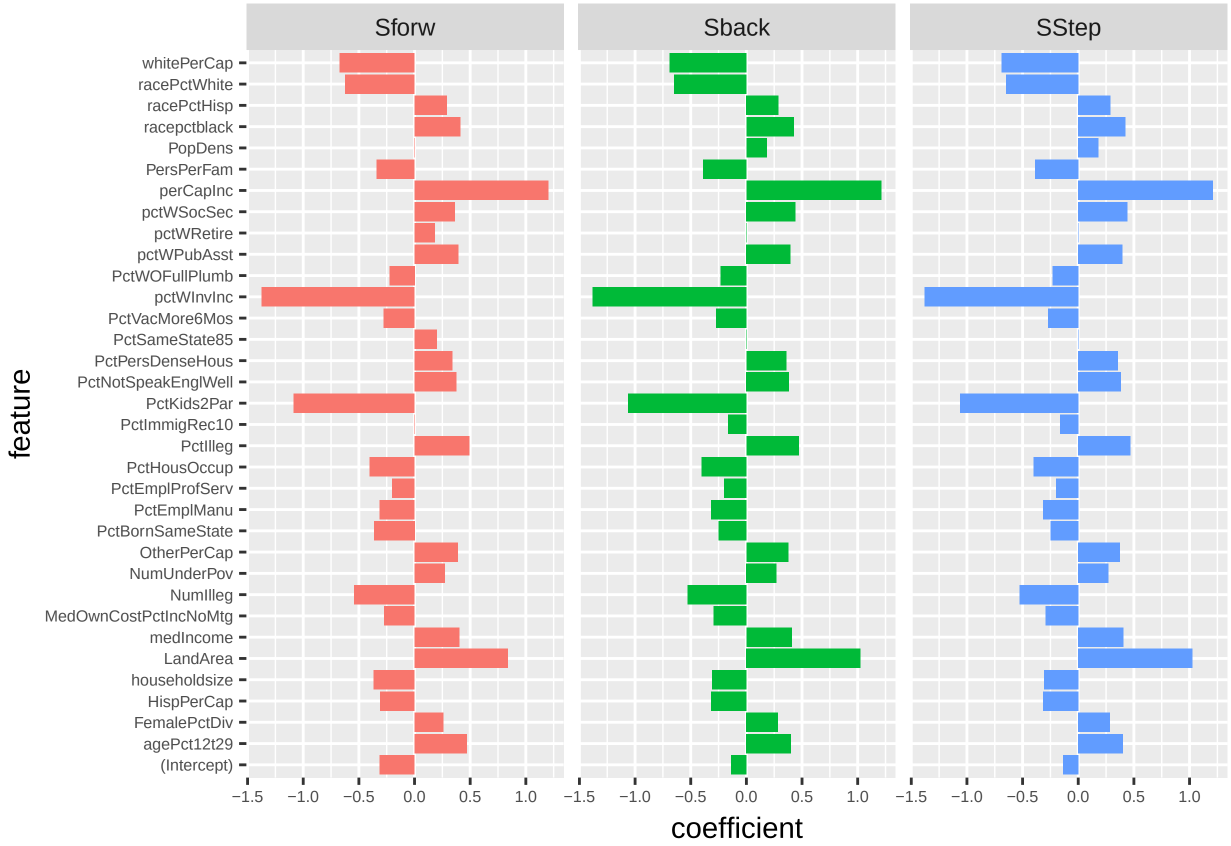
```
[1]  1.210993 13.252620
```

```
[1]  1.112525 13.267776
```

```
[1]  1.210993 13.252620
```

More than 10 so there is a problem with multicolinearity.

And we can show on the graph importance of features for the best models of different approach



References:

[1] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/communities+and+crime

[2] Statistical modelling 1 lecture notes 2016/2017. Chapter 3.9.4

[3] “An introduction to Statistical Learning with Applications in R” G.James, D. Witten, T. Hastie, R. Tibshirani, Springer Texts in Statistics, ISBN 978-1-4614-7137-0, page 29 - 30