Spring 2021 - Final Examination

Abhijit Shamkant Gokhale

Instructions

Your goal for this final exam is to conduct the necessary analyses of vaccination rates in California school districts and then write up a technical report for a scientifically knowledgeable staff member in a California state legislator's office. You should provide sufficient numeric and graphical detail that the staff member can create a comprehensive briefing for a legislator (see question 10 for specific points of interest). You can assume that the staff member understands the concept of statistical significance and other basic concepts like mean, standard deviation, and correlation.

For this exam, the report writing is very important: Your responses will be graded on the basis of clarity; conciseness; inclusion and explanation of specific and appropriate statistical values; inclusion of both frequentist and Bayesian inferential evidence (i.e., it is not sufficient to just examine the data); explanation of any included tabular material and the appropriate use of graphical displays when/if necessary. It is also important to conduct a thorough analysis, including both data exploration and cleaning and appropriate diagnostics. Bonus points will be awarded for work that goes above expectations.

In your answer for each question, make sure you write a narrative with complete sentences that answers the substantive question. You can choose to put important statistical values into a table for readability, or you can include the statistics within your narrative. Be sure that you not only report what a test result was, but also what that result means substantively. Make sure to include enough statistical information so that another analytics professional could review your work. Your report can include graphics created by R, keeping in mind that if you do include a graphic, you will have to provide some accompanying narrative text to explain what it is doing in your report. Finally, be sure to proofread your final knitted submission to ensure that everything is included and readable.

You may not *receive assistance, help, coaching, guidance, or support from any human except your instructor at any point during this exam. Your instructor will be available by email throughout the report writing period if you have questions, but don't wait until the last minute!*

Data

You have an RData file available on Blackboard area that contains two data sets that pertain to vaccinations for the U.S. as a whole and for Californian school districts. The U.S. vaccine data is a time series and the California data is a sample of end-of-year vaccination reports from n=700 school districts. Here is a description of the datasets:

```
# Load the data from the below location

load("C:/Users/abhij/Desktop/Lectures/SEM 1/IST 772/Final Exam/datasets6.RData")
```

usVaccines – Time series data from the World Health Organization reporting vaccination rates in the U.S. for five common vaccines

```
Time-Series [1:38, 1:5] from 1980 to 2017:
 - attr(*, "dimnames")=List of 2
  ..$ : NULL
  ..$ : chr [1:5] "DTP1" "HepB_BD" "Pol3" "Hib3" "MCV1"...
```

(Note: DTP1 = First dose of Diphtheria/Pertussis/Tetanus vaccine (i.e., DTP); HepB_BD = Hepatitis B, Birth Dose (HepB); Pol3 = Polio third dose (Polio); Hib3 – Influenza third dose; MCV1 = Measles first dose (included in MMR))

districts – A sample of California public school districts from the 2017 data collection, along with specific numbers and percentages for each district:

```
'data.frame':   700 obs. of  14 variables:
 $ DistrictName    : Name of the district
 $ WithDTP         : Percentage of students in the district with the DTP vaccine
 $ WithPolio       : Percentage of students in the district with the Polio vaccine
 $ WithMMR         : Percentage of students in the district with the MMR vaccine
 $ WithHepB        : Percentage of students in the district with Hepatitis B vaccine
 $ PctUpToDate     : Percentage of students with completely up-to-date vaccines
 $ DistrictComplete: Boolean showing whether or not district's reporting was complete
 $ PctBeliefExempt : Percentage of all enrolled students with belief exceptions
 $ PctMedicalExempt: Percentage of all enrolled students with medical exceptions
 $ PctChildPoverty : Percentage of children in district living below the poverty line
 $ PctFamilyPoverty: Percentage of families in district living below the poverty line
 $ PctFreeMeal     : Percentage of students in the district receiving free or reduced cost meals
 $ Enrolled        : Total number of enrolled students in the district
 $ TotalSchools    : Total number of different schools in the district
```

As might be expected, the data are quite skewed: districts range from 1 to 582 schools enrolling from 10 to more than 50,000 students. Further, while most districts have low rates of missing vaccinations, a handful are quite high. Be sure to note problems the data cause for the analysis and address any problems you can.

Univariate Exploration

```
# inspecting dataset
library(psych)
describe(districts)
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
```

```
##                  vars   n   mean      sd median trimmed    mad min   max range
## DistrictName*       1 700 416.38  245.76  410.5  415.34 312.09   1   846   845
## WithDTP             2 700  89.98   10.84   93.0   91.97   7.41  23   100    77
## WithPolio           3 700  90.40   10.79   94.0   92.42   5.93  23   100    77
## WithMMR             4 700  89.92   11.22   94.0   92.04   5.93  23   100    77
## WithHepB            5 700  92.40    9.76   96.0   94.32   4.45  23   100    77
## PctUpToDate         6 700  88.59   14.84   92.0   90.30   7.41  23   201   178
## DistrictComplete    7 700    NaN      NA     NA     NaN     NA Inf  -Inf  -Inf
```

```
## PctBeliefExempt     8 700   5.51    8.70    2.0    3.58   2.97   0    77    77
## PctMedicalExempt    9 452   0.17    0.63    0.0    0.00   0.00   0     5     5
## PctChildPoverty    10 700  22.35   11.99   21.0   21.31  11.86   2    72    70
## PctFamilyPoverty   11 700  11.54    8.07   10.0   10.53   7.41   0    47    47
## PctFreeMeal        12 700  48.60   24.66   50.0   49.16  29.65   0   100   100
## Enrolled           13 700 633.59 2226.60  224.0  362.15 286.14  10 54238 54228
## TotalSchools       14 700   7.27   24.04    3.0    4.37   2.97   1   582   581
##                     skew kurtosis    se
## DistrictName*       0.03    -1.20  9.29
## WithDTP            -2.37     7.55  0.41
## WithPolio          -2.47     8.11  0.41
## WithMMR            -2.29     6.78  0.42
## WithHepB           -3.05    12.66  0.37
## PctUpToDate         0.65    17.23  0.56
## DistrictComplete      NA       NA    NA
## PctBeliefExempt     3.43    15.73  0.33
## PctMedicalExempt    5.13    30.28  0.03
## PctChildPoverty     0.77     0.35  0.45
## PctFamilyPoverty    1.20     1.54  0.31
## PctFreeMeal        -0.14    -0.93  0.93
## Enrolled           20.24   478.00 84.16
## TotalSchools       19.91   465.07  0.91
```

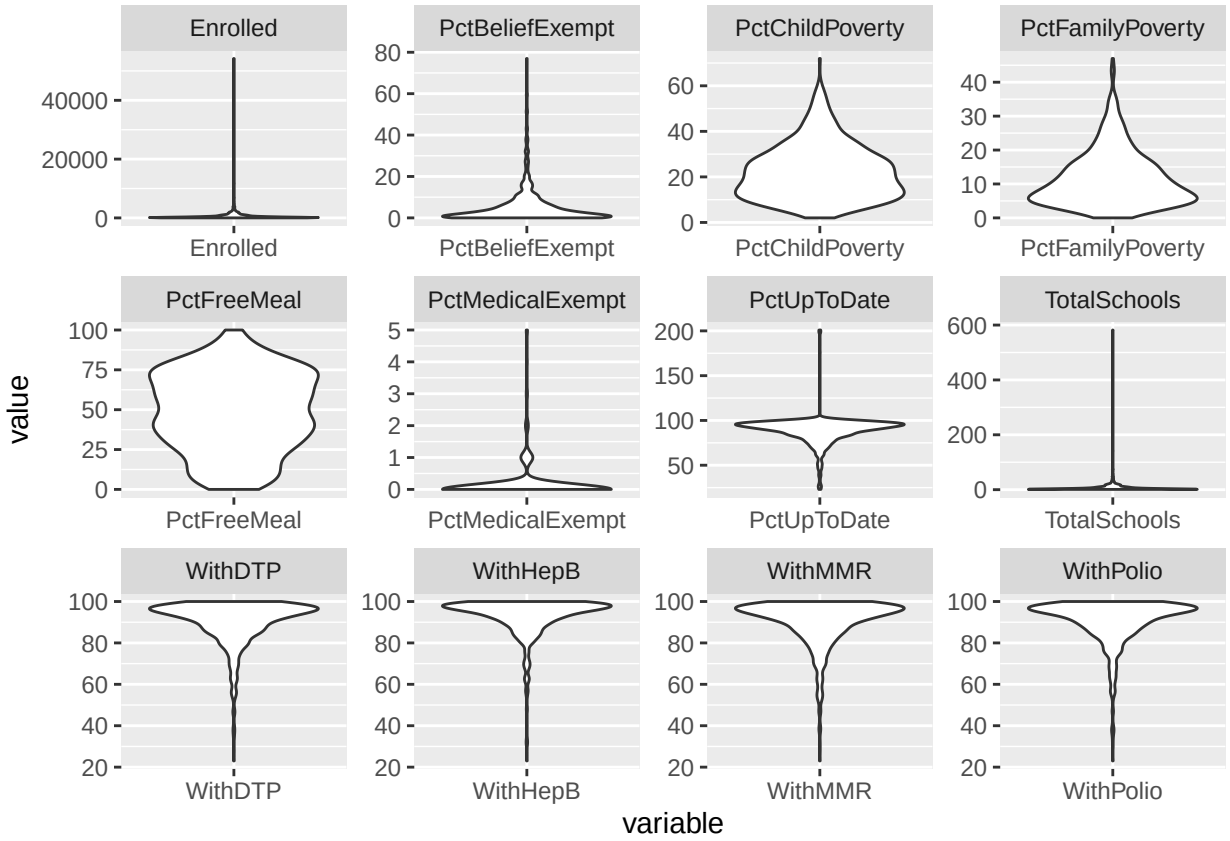
_In the above results, we can see that there are missing values in PctMedicalExempt because n for PctMedicalExempt is 452 whereas for rest of the columns it is 700.

Let's check the skewness in the variables

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'dplyr' was built under R version 4.0.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
districts %>% pivot_longer(cols=-c("DistrictName","DistrictComplete"), names_to="variable",
                           values_to="value", values_drop_na = TRUE) %>%
ggplot(aes(x=variable, y=value)) + geom_violin() + facet_wrap( ~ variable, scales="free")
```



In the above plots, we can see that apart from PctFreeMeal and PctChildPoverty there is a lot of skewness either to the left side or to the right side. In the PctUpToDate plot we have some values greater than 100 and as the percentage values can't be greater than 100 we should remove them.

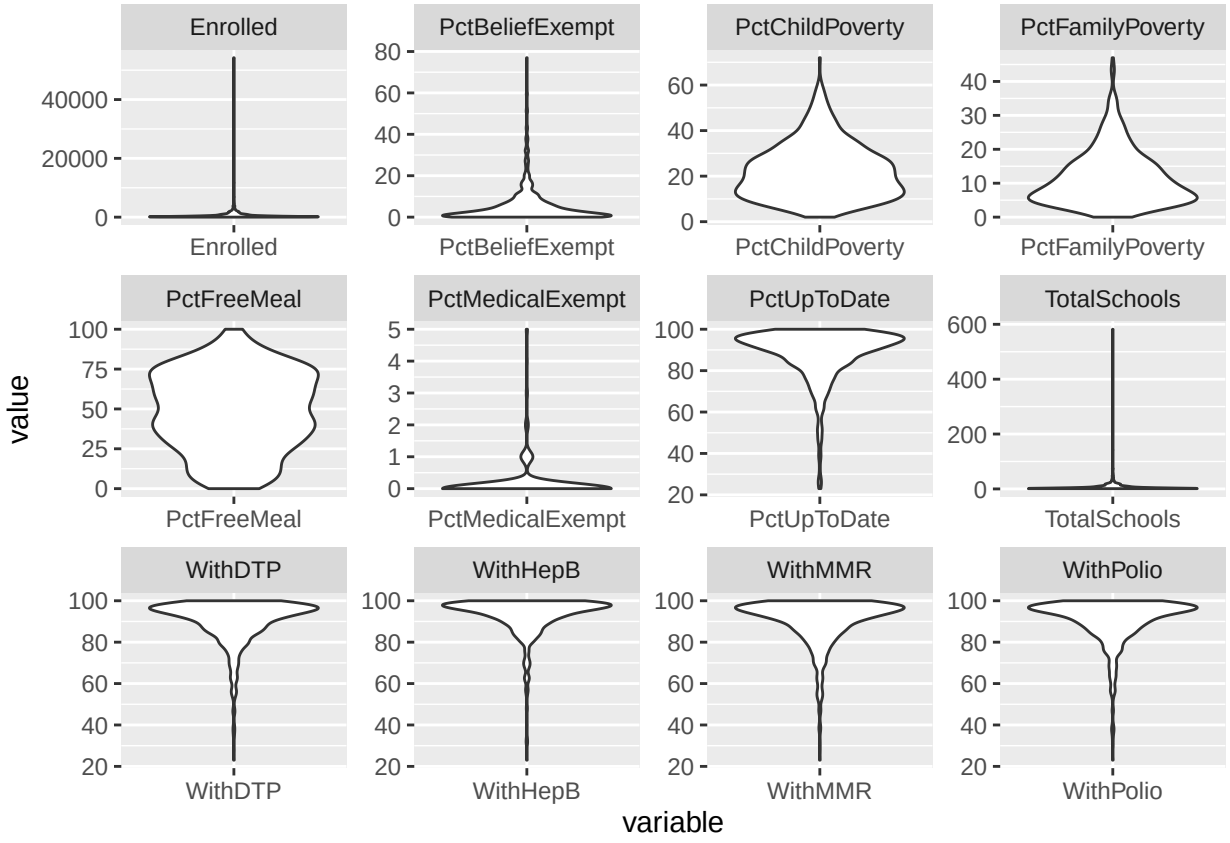
```
#Checking for values(Outliers) greater than 100
districts[districts$PctUpToDate > 100,]
```

```
##                DistrictName WithDTP WithPolio WithMMR WithHepB PctUpToDate
## 472 Redwood City Elementary      85        88      84       96         176
## 334      Hanford Elementary      98        98      99       99         199
## 345 Burrel Union Elementary     100       100     100      100         201
## 328    Farmersville Unified      98        99      99       99         199
##     DistrictComplete PctBeliefExempt PctMedicalExempt PctChildPoverty
## 472             TRUE               1                0              18
## 334             TRUE               1               NA              30
## 345             TRUE               0                0              30
## 328             TRUE               0                0              55
##     PctFamilyPoverty PctFreeMeal Enrolled TotalSchools
## 472                8          47     1062           14
## 334               17          72      762            9
## 345               29          78       19            1
## 328               32          86      171            1
```

From the above results, PctUpToDate has 4 values greater than 100 so we should remove them.

```
# using subset function to remove the 4 outliers from the dataset districts
districts_new <- subset(districts, PctUpToDate <= 100 )
```

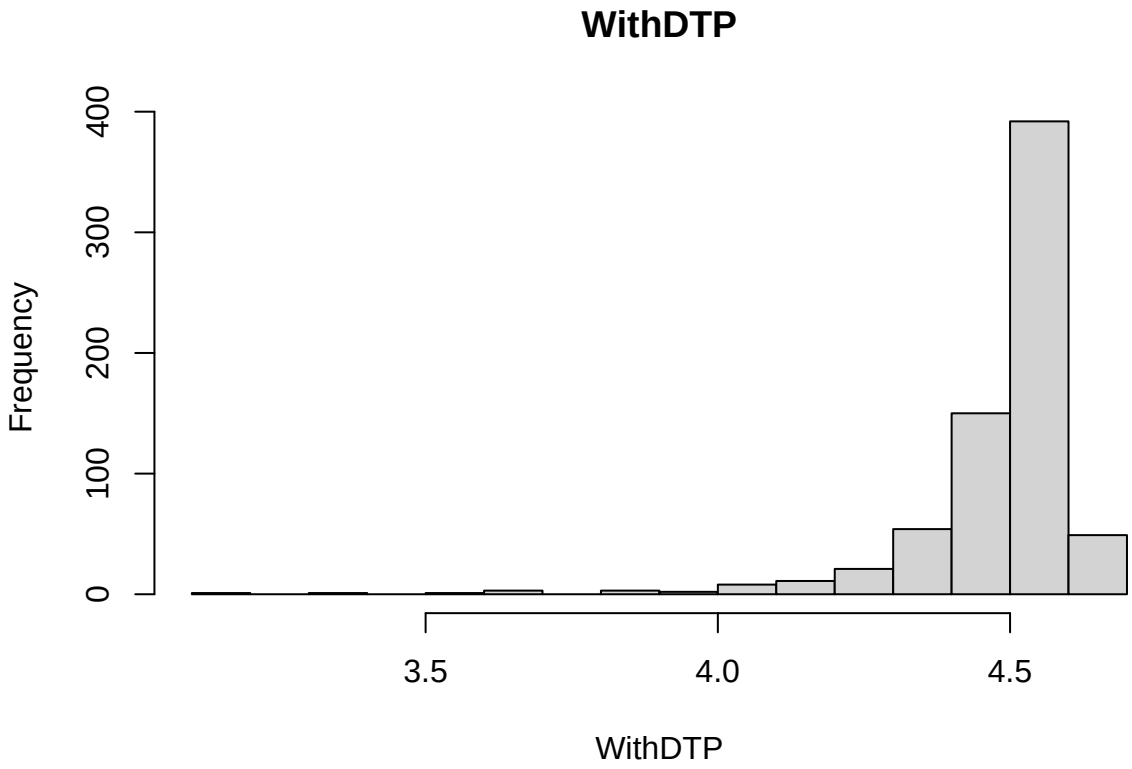
```
library(tidyverse)
districts_new %>% pivot_longer(cols=-c("DistrictName","DistrictComplete"), names_to="variable",
                        values_to="value", values_drop_na = TRUE) %>%
ggplot(aes(x=variable, y=value)) + geom_violin() + facet_wrap( ~ variable, scales="free")
```

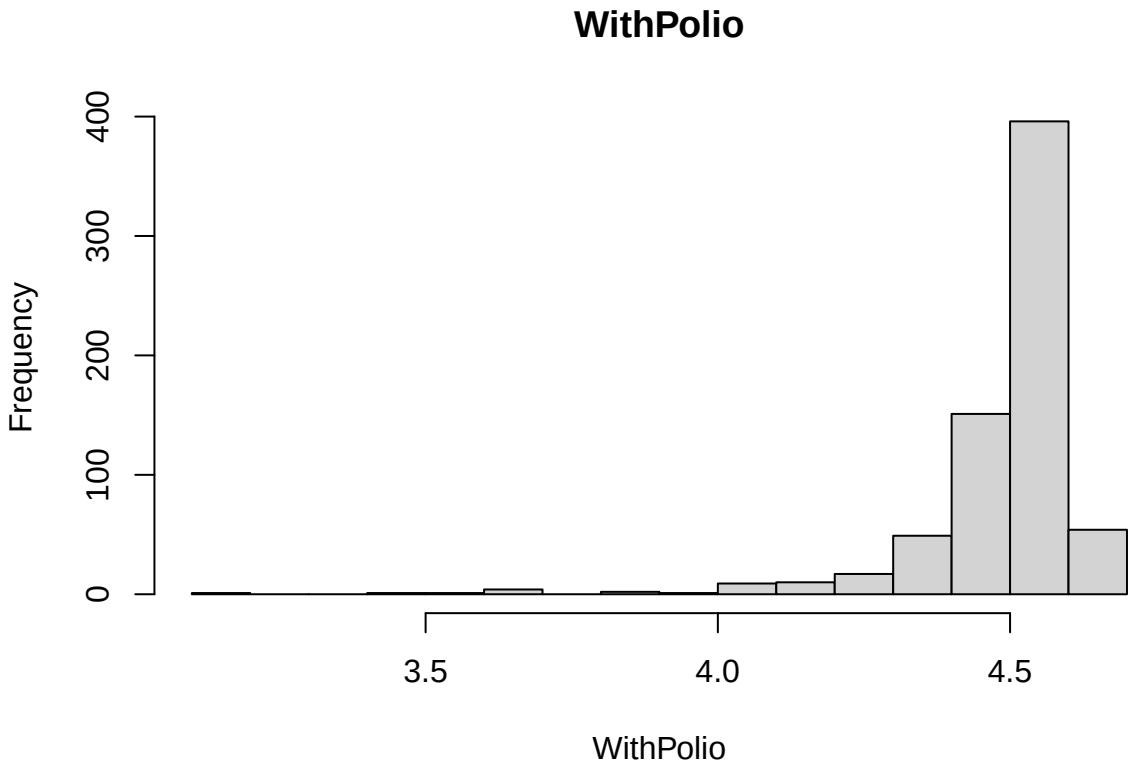


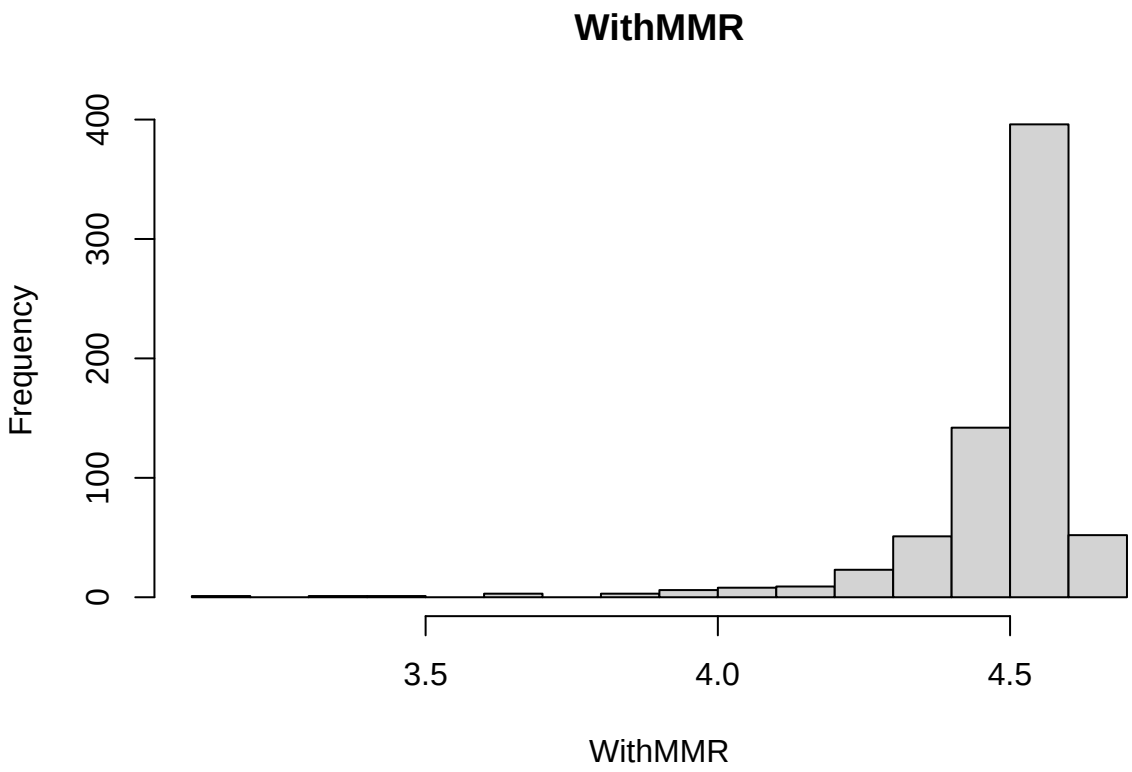
As we can see from above plot, we have not hampered other variables plot by removing 4 outliers from PctUpToDate column. Also, there still exists signficant skewness in the variables.

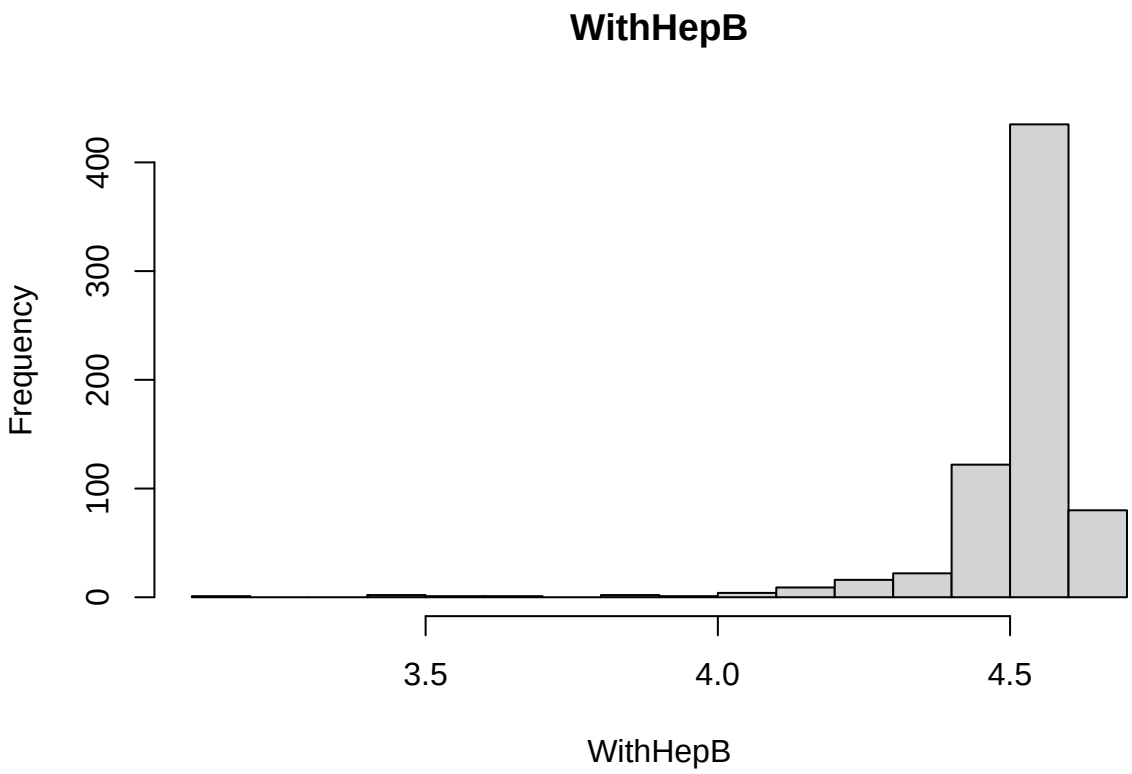
We have to apply transformations to verify whether we can remove the skewness from the above plots

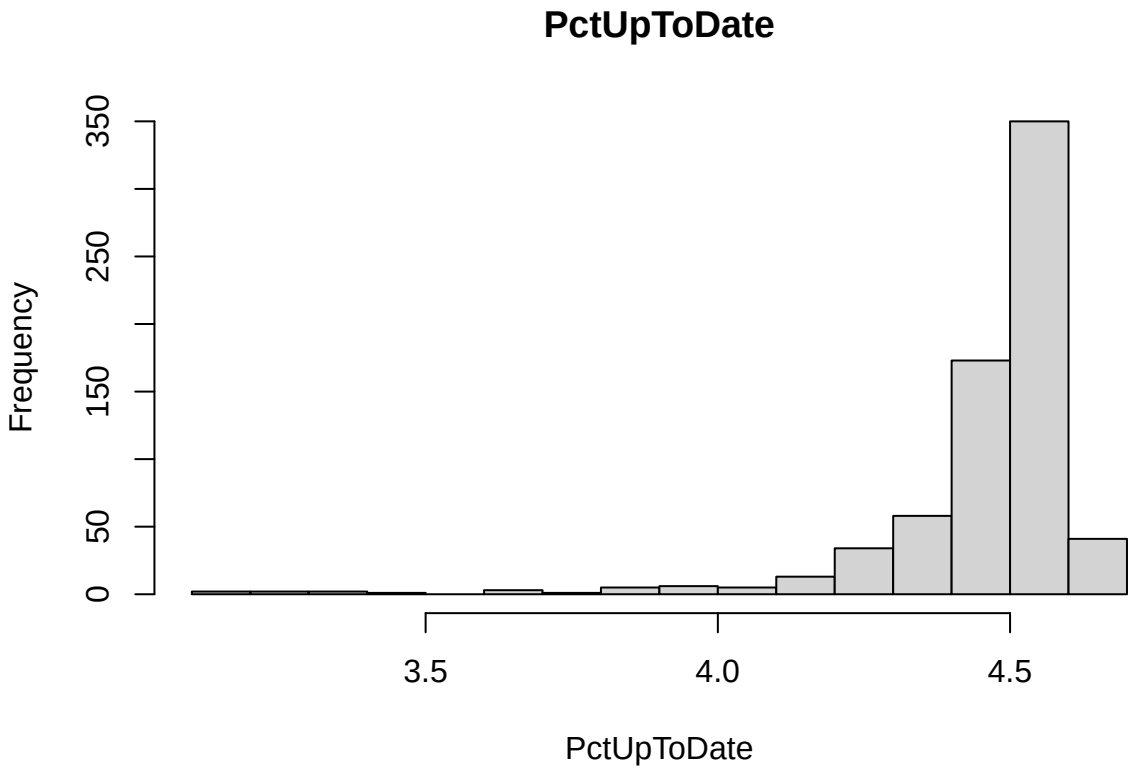
```
# Separating Numerical Data and storing it into "districts_nmr"
districts_nmr <- districts_new[,!colnames(districts_new) %in% c("DistrictName","DistrictComplete")]


# Applying log transformations to see the effect of skewness
for (i in c(1:ncol(districts_nmr))) hist(log(districts_nmr[[i]]),main = colnames(districts_nmr)[i],xlab
```

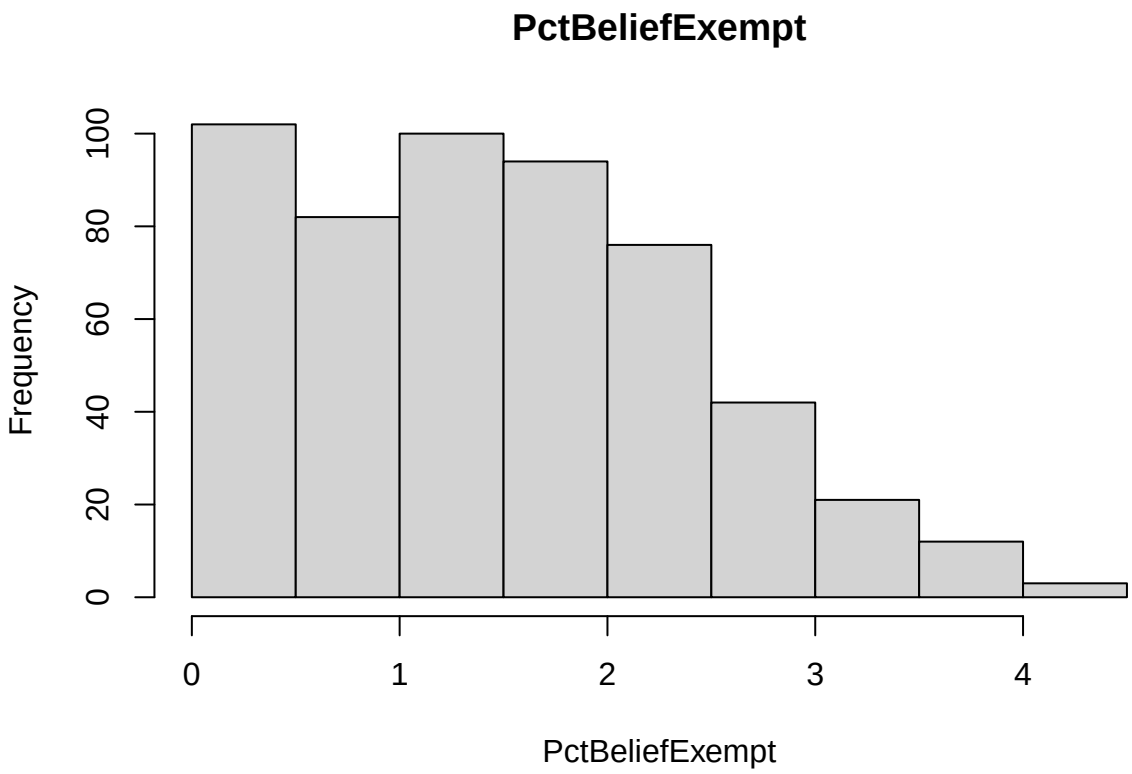

WithDTP
Frequency
0
100
200
300
400
3.5
4.0
4.5
WithDTP

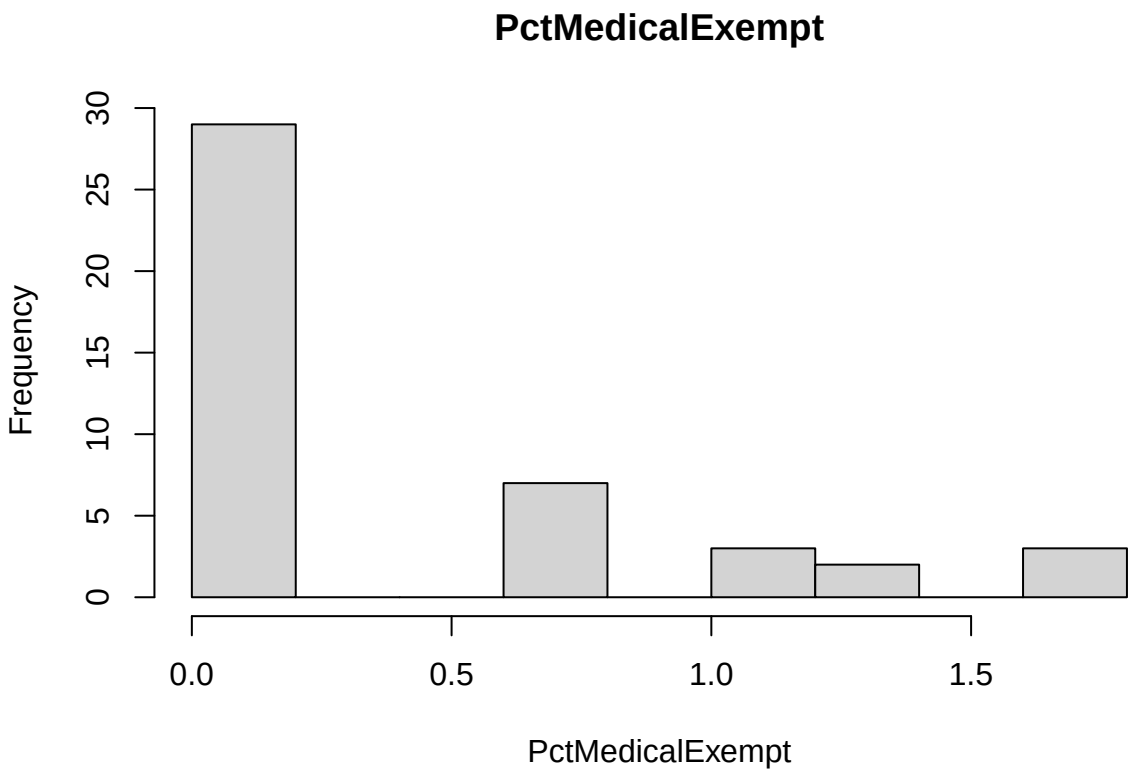

WithPolio
Frequency
0
100
200
300
400
3.5
4.0
4.5
WithPolio

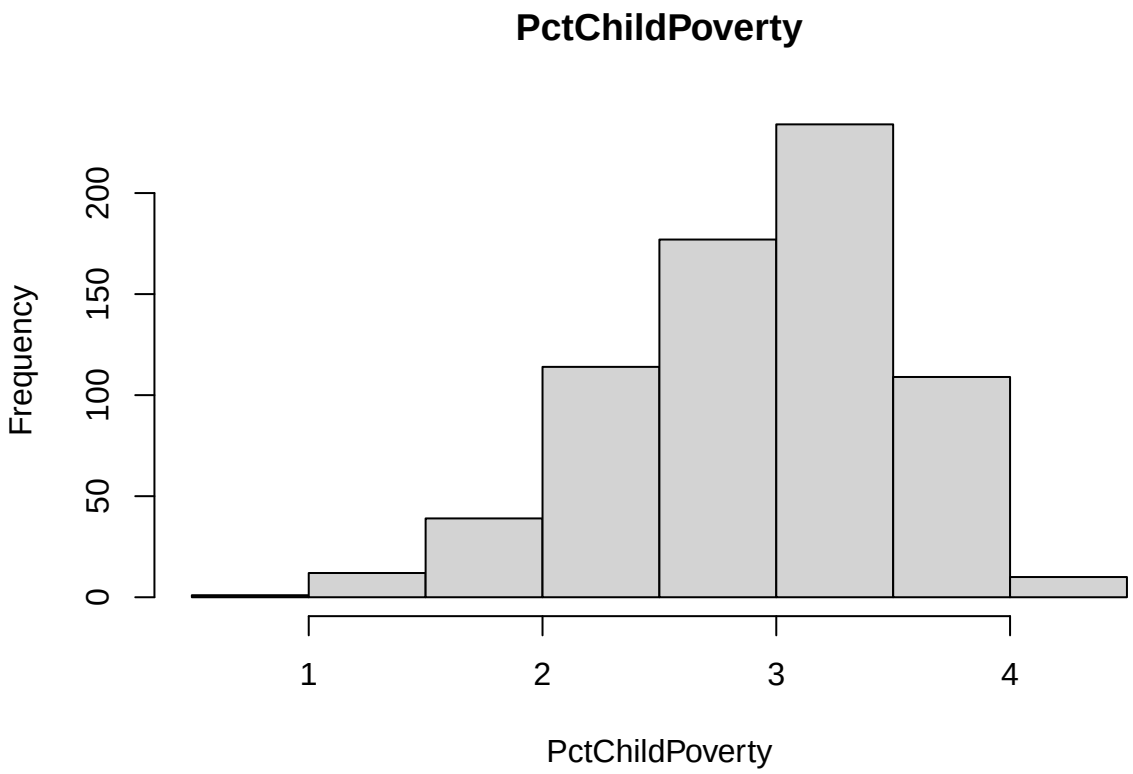

WithMMR
Frequency
0
100
200
300
400
3.5
4.0
4.5
WithMMR

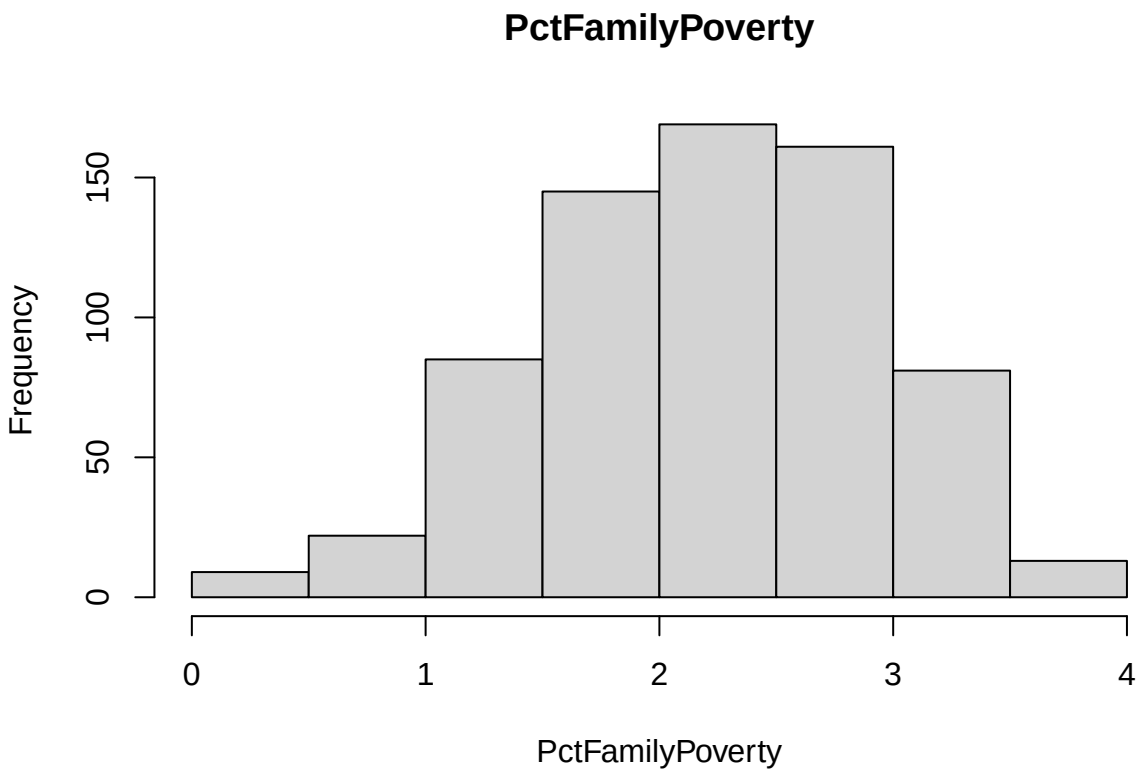

WithHepB
Frequency
0
100
200
300
400
3.5
4.0
4.5
WithHepB

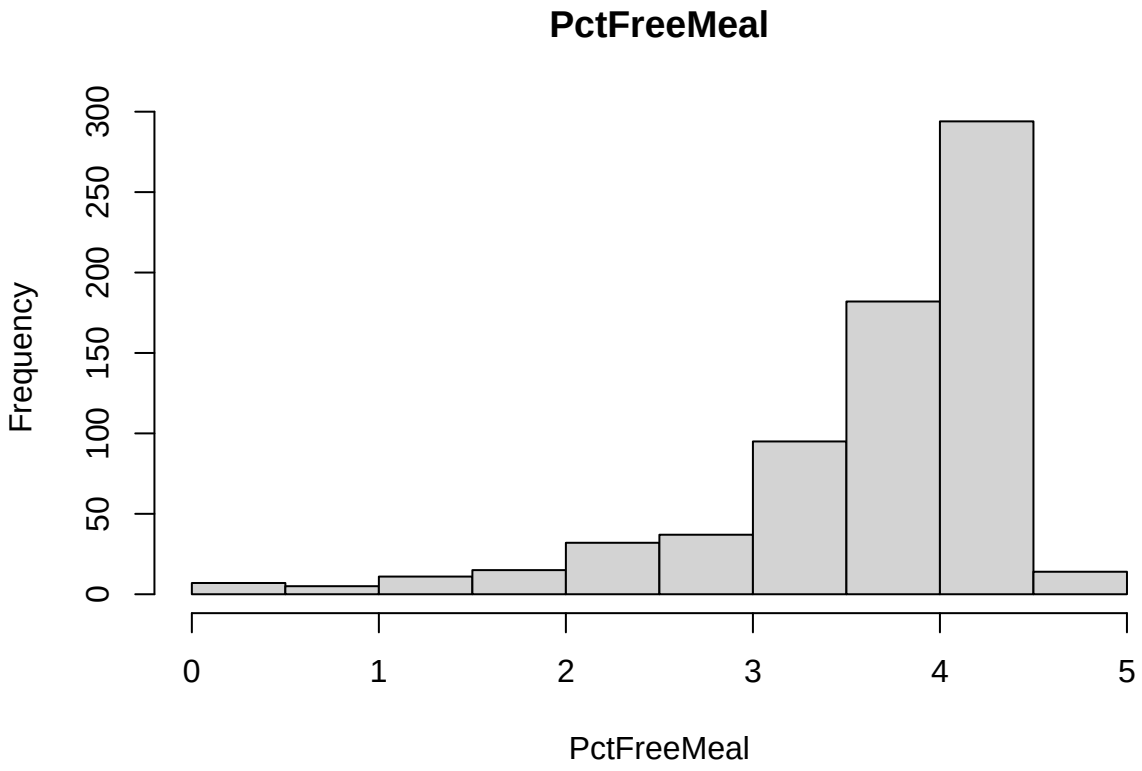

PctUpToDate
Frequency
0
50
150
250
350
3.5
4.0
4.5
PctUpToDate

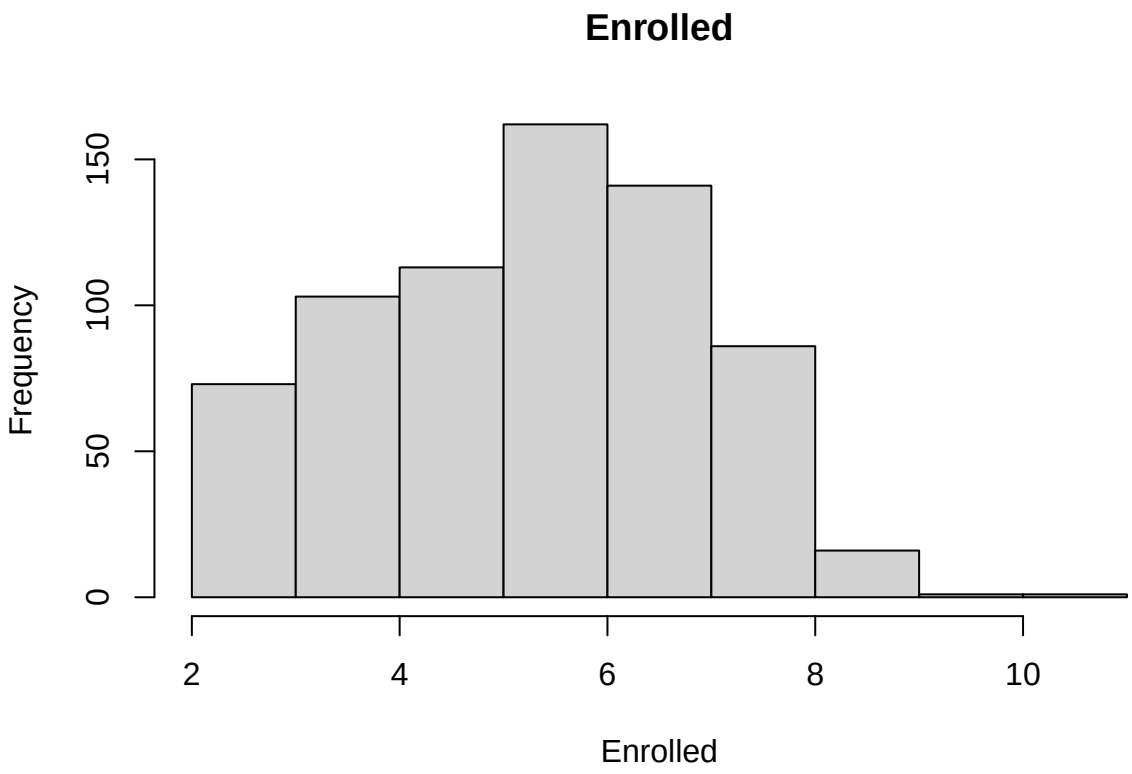

PctBeliefExempt
Frequency
0
20
40
60
80
100
0
1
2
3
4
PctBeliefExempt

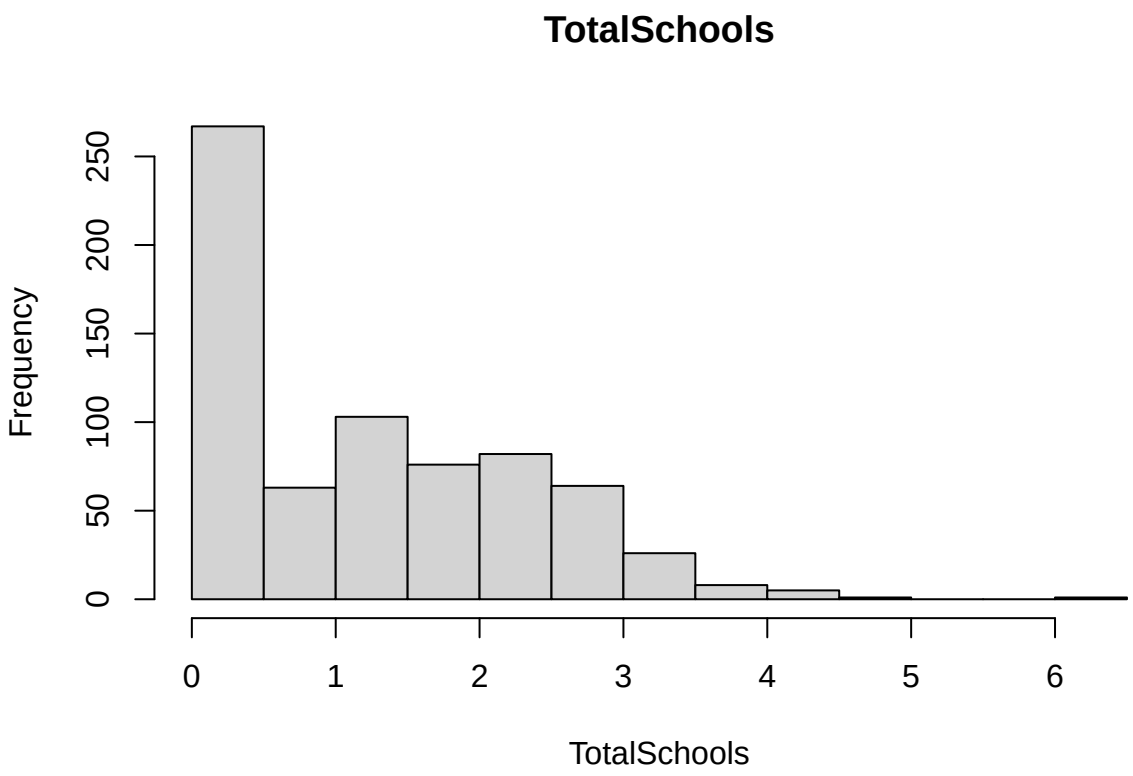

PctMedicalExempt
Frequency
0
5
10
15
20
25
30
0.0
0.5
1.0
1.5
PctMedicalExempt


PctChildPoverty
Frequency
0
50
100
150
200
1
2
3
4
PctChildPoverty


PctFamilyPoverty
Frequency
0
50
100
150
0
1
2
3
4
PctFamilyPoverty


PctFreeMeal
Frequency
0
50
100
150
200
250
300
0
1
2
3
4
5
PctFreeMeal


Enrolled
Frequency
0
50
100
150
2
4
6
8
10
Enrolled



In the above plots, we can see that the skewness has been decreased; but there still exists significant skewness in most variables except PctBeliefExempt, PctChildPoverty, PctFamilyPoverty, Enrolled and TotalSchools. In addition to this, we can also see that PctFreeMeal column has been affected inveresly and the skewness got induced which was not present earlier.

Apart from these NA values let us check the outliers and outlier's plots

```
library(dlookr)
```

```
## Warning: package 'dlookr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dlookr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
## The following object is masked from 'package:psych':
##
##     describe
```

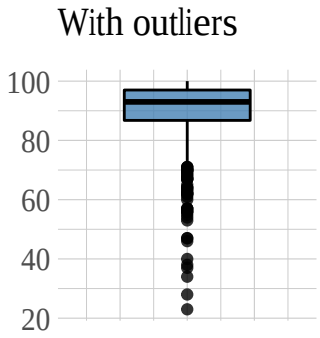
```
## The following object is masked from 'package:base':
##
##     transform
```

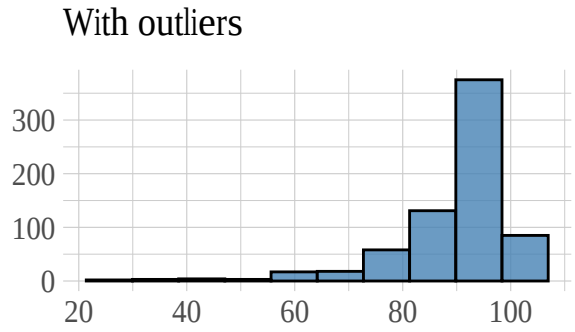
```
diagnose_outlier(districts_new)
```

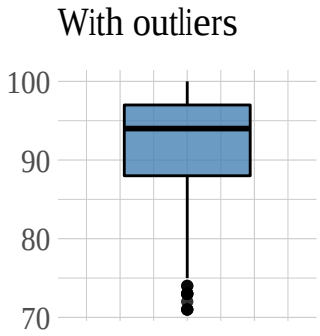
```
##           variables outliers_cnt outliers_ratio outliers_mean   with_mean
## 1           WithDTP           43       6.178161     57.790698  89.9525862
## 2         WithPolio           44       6.321839     58.409091  90.3620690
## 3           WithMMR           40       5.747126     55.725000  89.8879310
## 4          WithHepB           43       6.178161     62.186047  92.3663793
## 5       PctUpToDate           38       5.459770     49.052632  87.9813218
## 6   PctBeliefExempt           55       7.902299     29.436364   5.5344828
## 7  PctMedicalExempt           44       6.321839      1.704545   0.1670379
## 8   PctChildPoverty           11       1.580460     58.181818  22.2887931
## 9  PctFamilyPoverty           16       2.298851     37.437500  11.4784483
## 10      PctFreeMeal            0       0.000000           NaN  48.4755747
## 11         Enrolled           63       9.051724   3630.047619 634.3390805
## 12     TotalSchools           56       8.045977     41.625000   7.2758621
##    without_mean
## 1     92.070444
## 2     92.518405
## 3     91.971037
## 4     94.353752
## 5     90.229483
## 6      3.483619
## 7      0.000000
## 8     21.712409
## 9     10.867647
## 10    48.475575
## 11   336.187994
## 12     4.270313
```

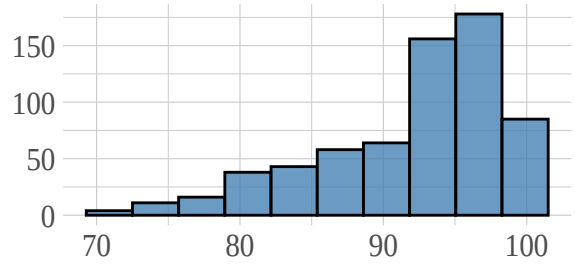
```
plot_outlier(districts_new)
```

Outlier Diagnosis Plot (WithDTP)

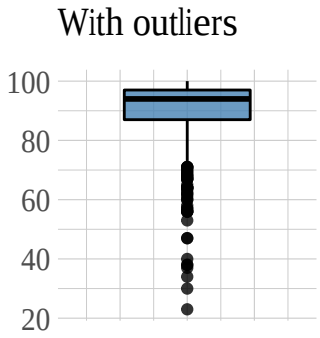


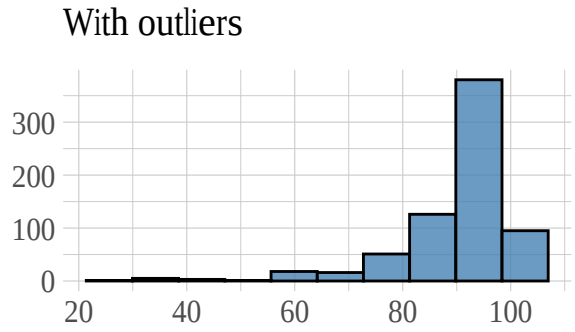


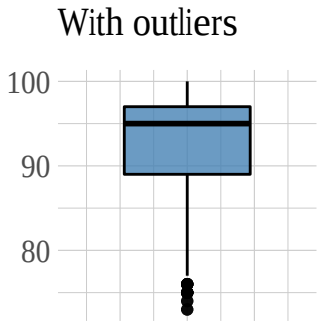


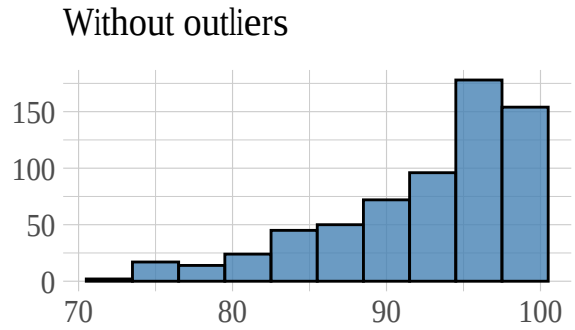


Outlier Diagnosis Plot (WithPolio)



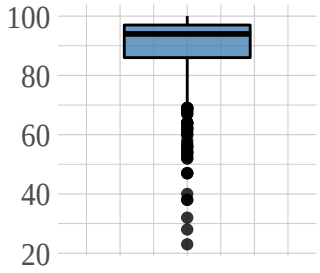




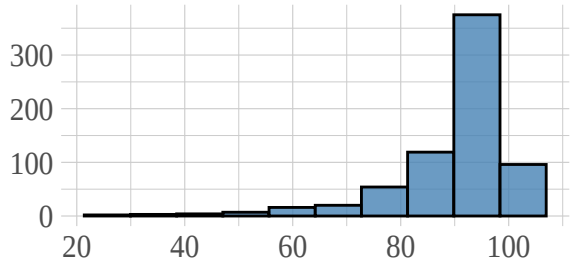


Outlier Diagnosis Plot (WithMMR)

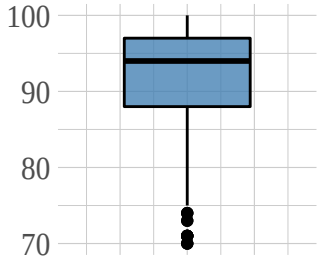
With outliers



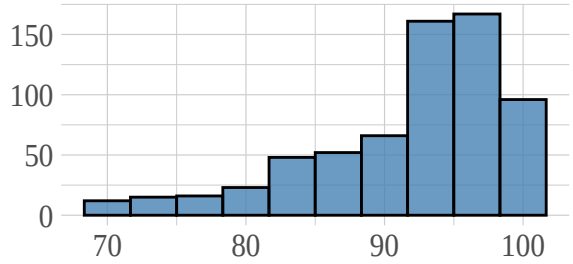
With outliers



With outliers

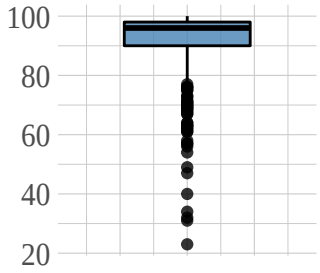


Without outliers

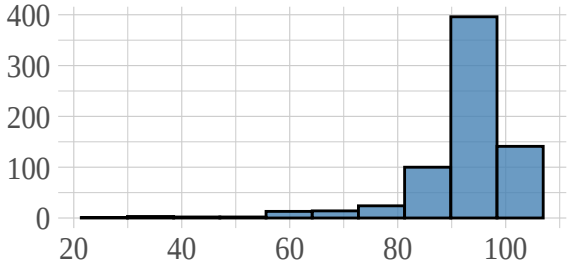


Outlier Diagnosis Plot (WithHepB)

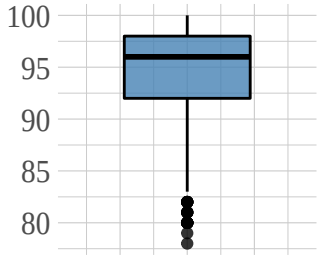
With outliers



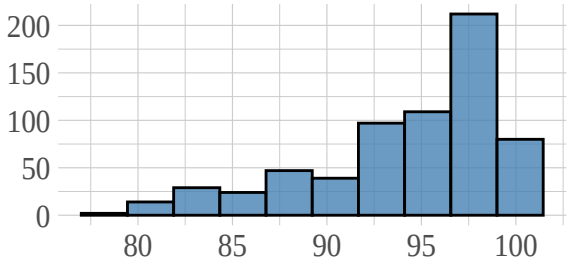
With outliers



With outliers

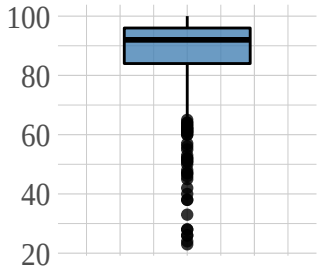


Without outliers

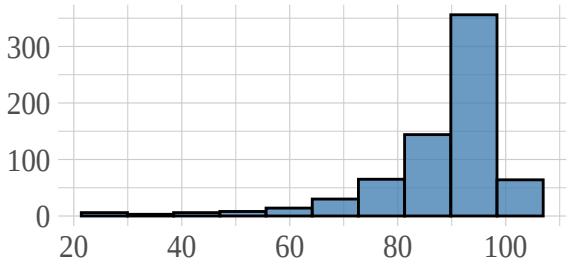


Outlier Diagnosis Plot (PctUpToDate)

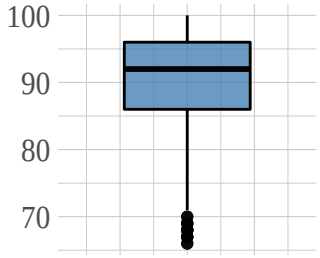
With outliers



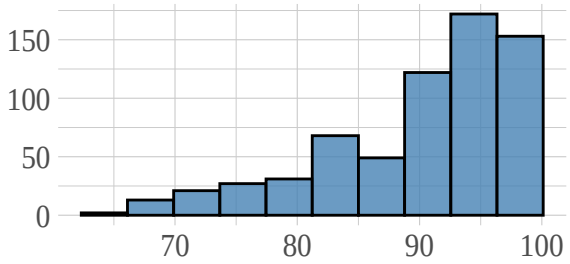
With outliers



With outliers

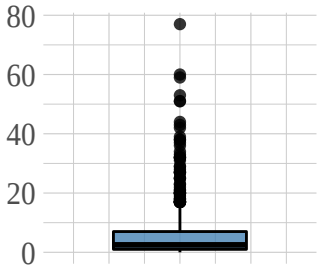


Without outliers

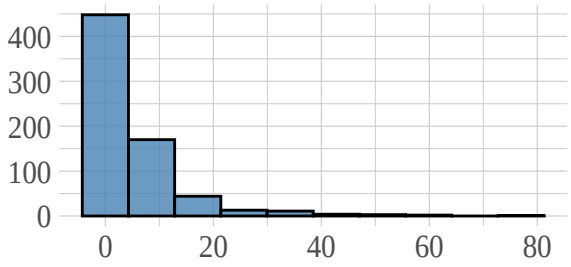


Outlier Diagnosis Plot (PctBeliefExempt)

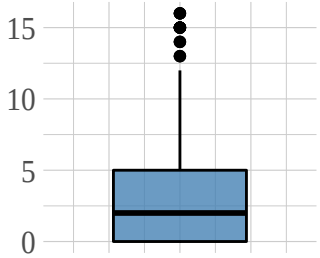
With outliers



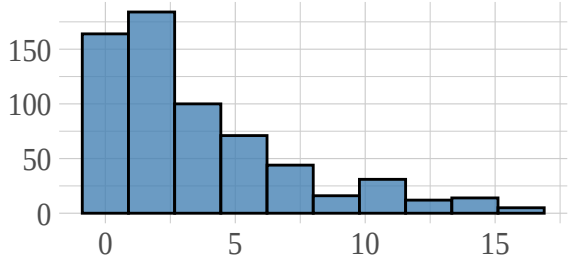
With outliers



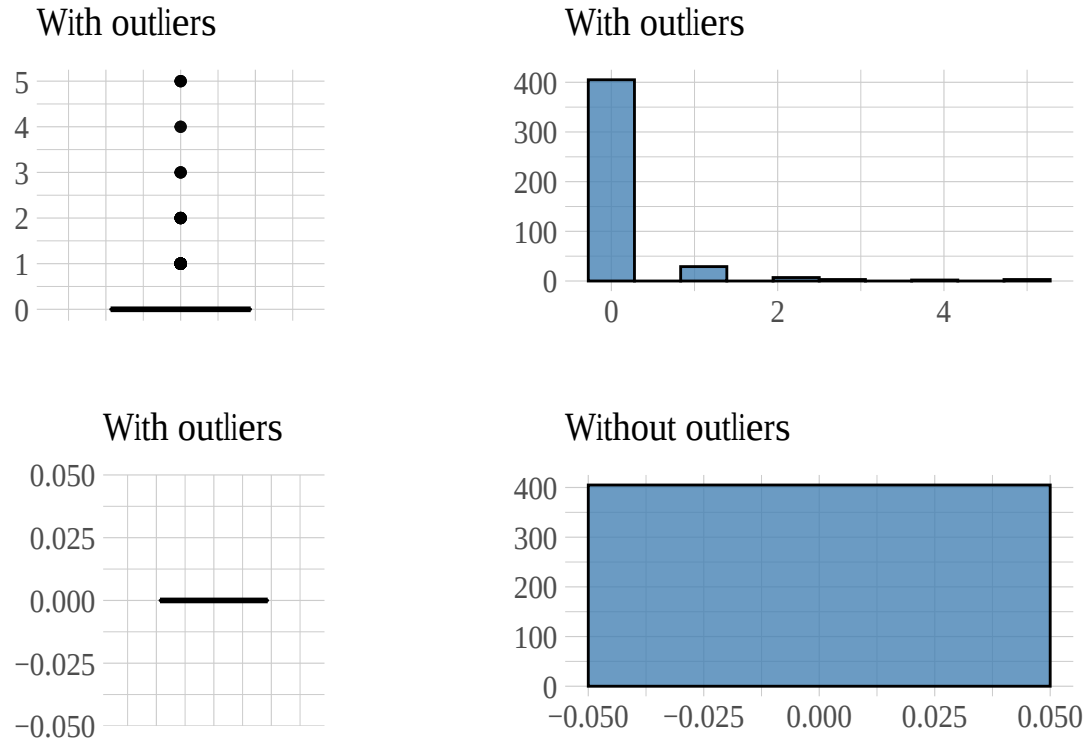
With outliers



Without outliers

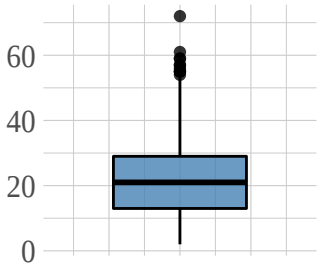


Outlier Diagnosis Plot (PctMedicalExempt)

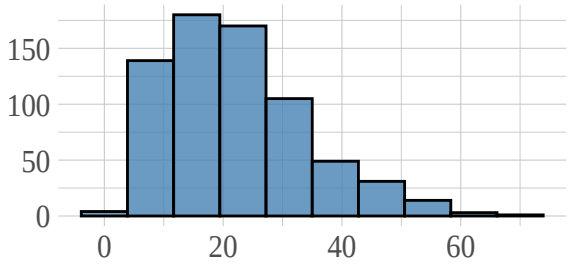


Outlier Diagnosis Plot (PctChildPoverty)

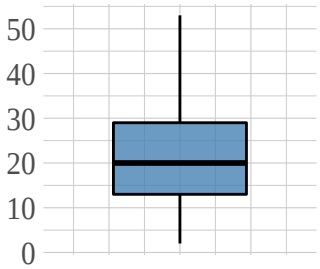
With outliers



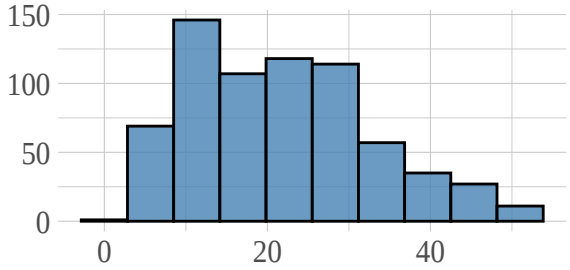
With outliers



With outliers

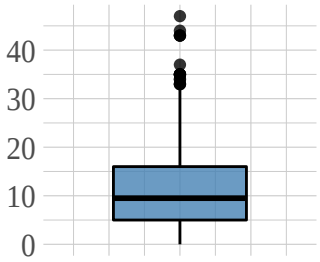


Without outliers

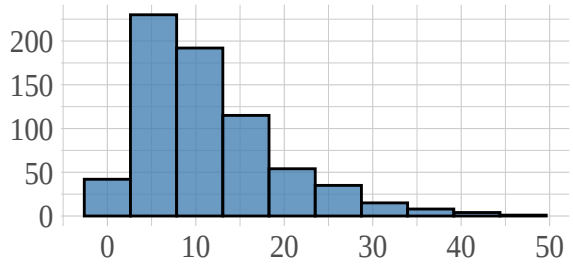


Outlier Diagnosis Plot (PctFamilyPoverty)

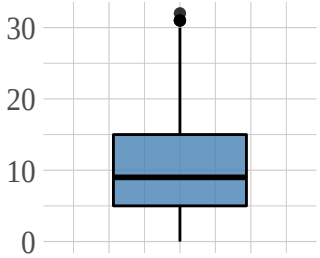
With outliers



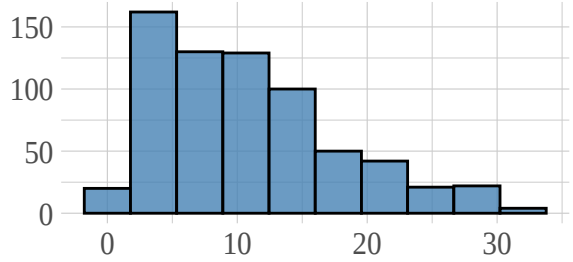
With outliers



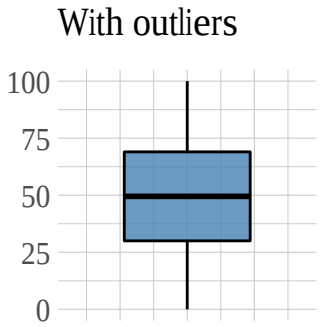
With outliers

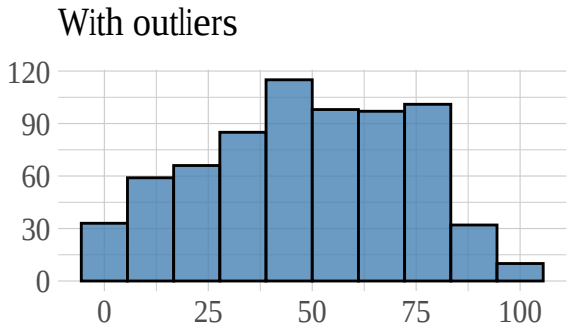


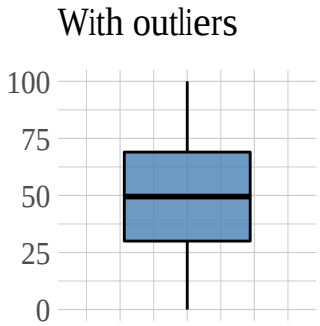
Without outliers

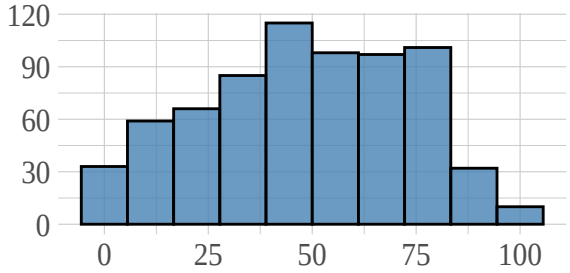


Outlier Diagnosis Plot (PctFreeMeal)



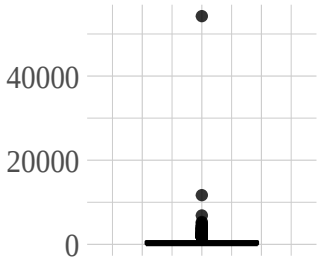




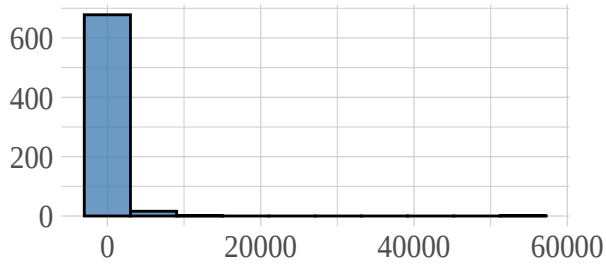


Outlier Diagnosis Plot (Enrolled)

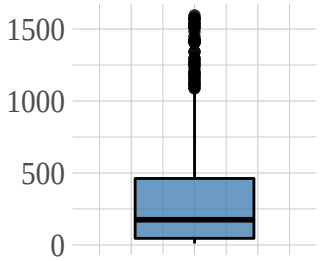
With outliers



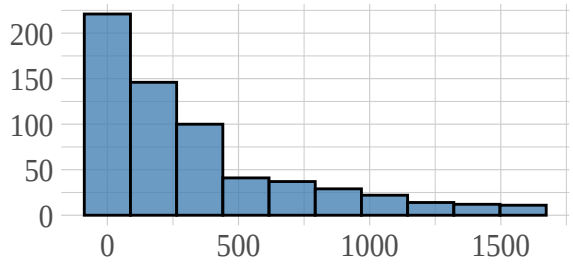
With outliers

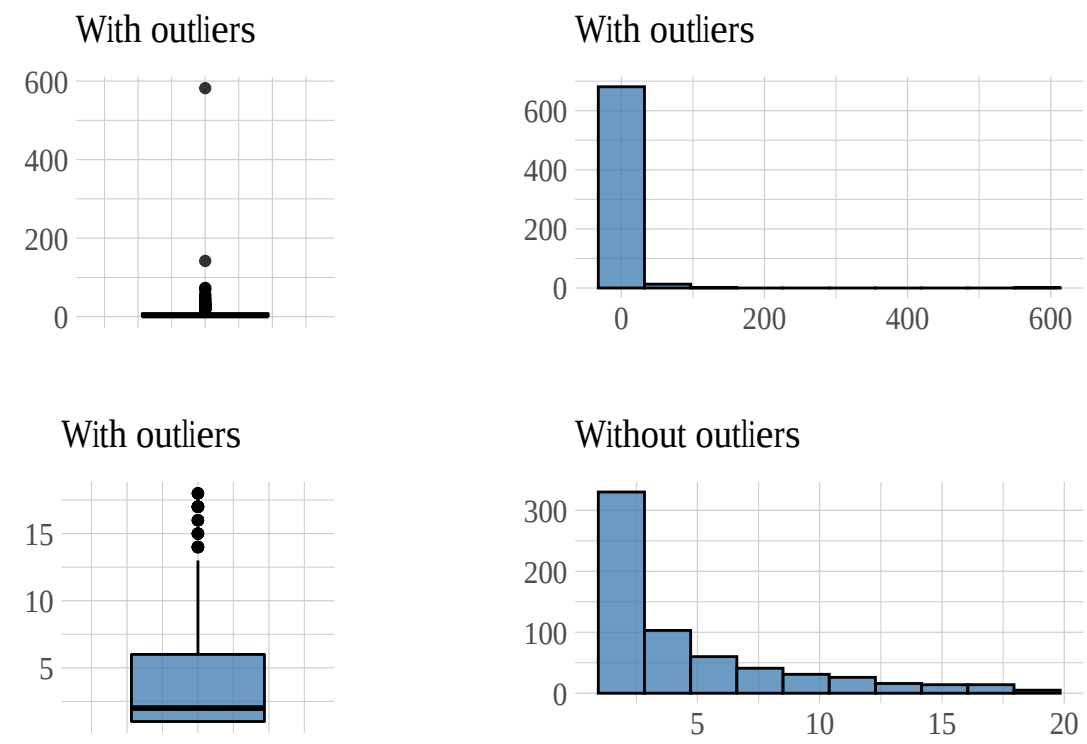


With outliers



Without outliers





In the above results, we can see that there exists many outliers spread across the dataset. There is only single column PctFreeMeal which doesn't have any outlier. Even though we have lot of outliers, we should keep them to the data unbiased

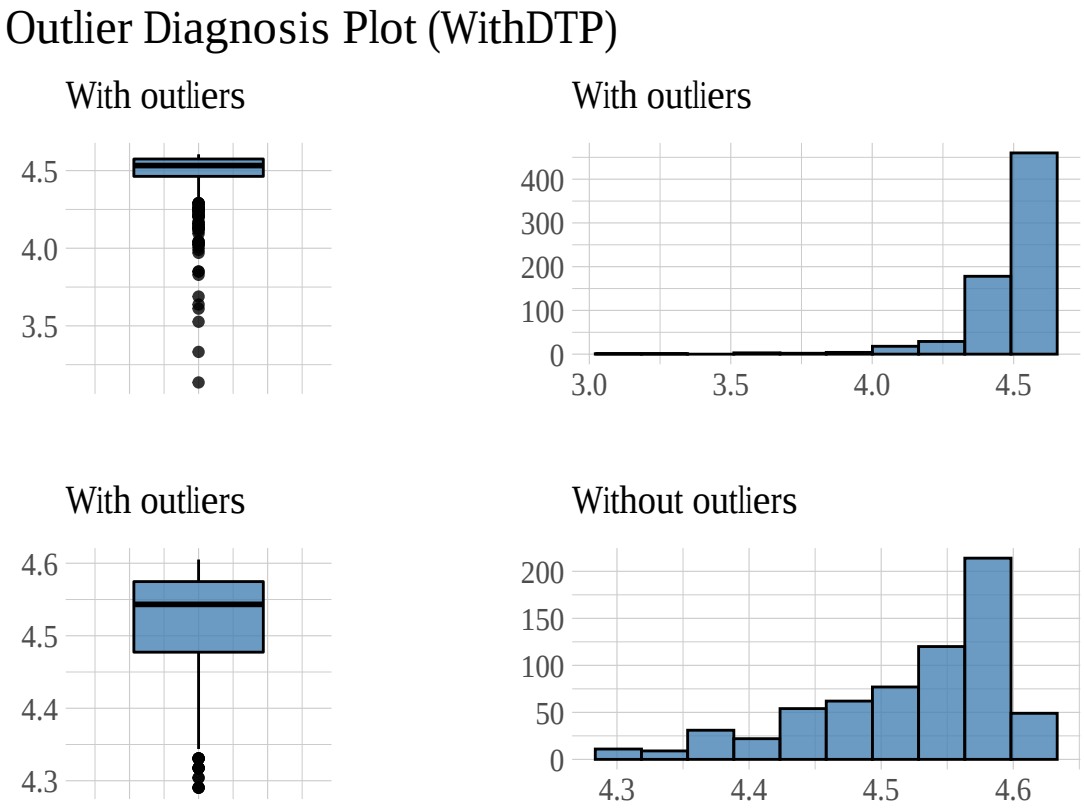
Now checking outliers on log transformed data

```
library(dlookr)

diagnose_outlier(log(districts_nmr))
```

```
##           variables outliers_cnt outliers_ratio outliers_mean with_mean
## 1           WithDTP           47      6.7528736      4.049286  4.489517
## 2         WithPolio           46      6.6091954      4.049923  4.494173
## 3           WithMMR           52      7.4712644      4.054124  4.488110
## 4          WithHepB           45      6.4655172      4.111455  4.517966
## 5       PctUpToDate           49      7.0402299      3.930615  4.462786
## 6   PctBeliefExempt          164     23.5632184          -Inf      -Inf
## 7  PctMedicalExempt          405     58.1896552          -Inf      -Inf
## 8   PctChildPoverty            4      0.5747126      0.997246  2.944462
## 9  PctFamilyPoverty           11      1.5804598          -Inf      -Inf
## 10      PctFreeMeal           48      6.8965517          -Inf      -Inf
## 11         Enrolled            1      0.1436782     10.901137  5.287066
## 12     TotalSchools            1      0.1436782      6.366470  1.166135
##    without_mean
## 1     4.5213977
## 2     4.5256125
```

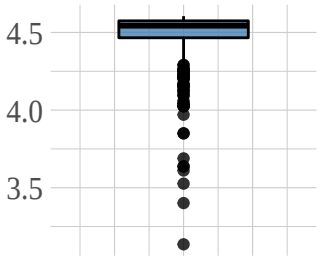
```
## 3     4.5231523
## 4     4.5460658
## 5     4.5030898
## 6     1.4268563
## 7     0.3579266
## 8     2.9557174
## 9     2.2161042
## 10    3.8246118
## 11    5.2789887
## 12    1.1586524
```

```
plot_outlier(log(districts_nmr))
```

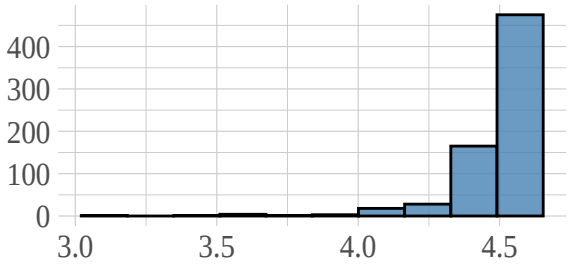


Outlier Diagnosis Plot (WithPolio)

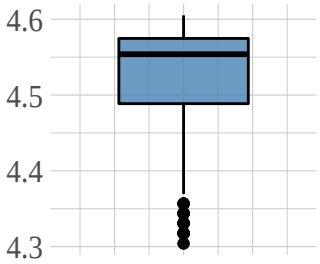
With outliers



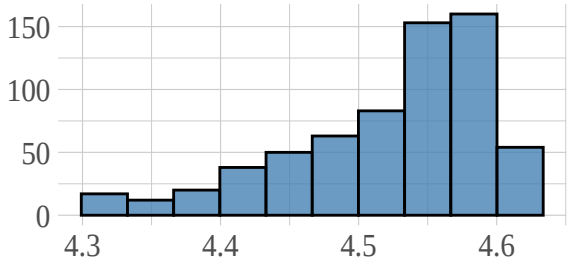
With outliers



With outliers

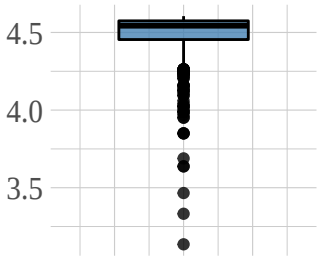


Without outliers

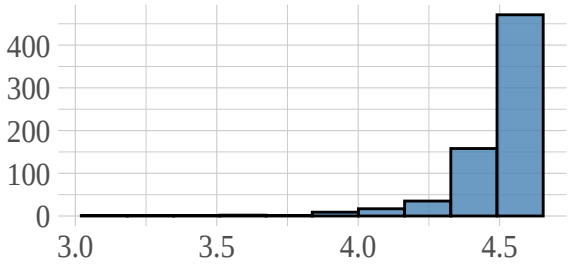


Outlier Diagnosis Plot (WithMMR)

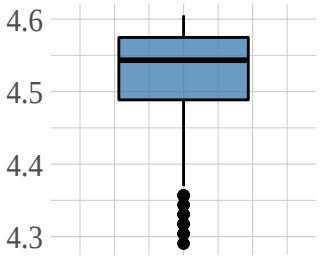
With outliers



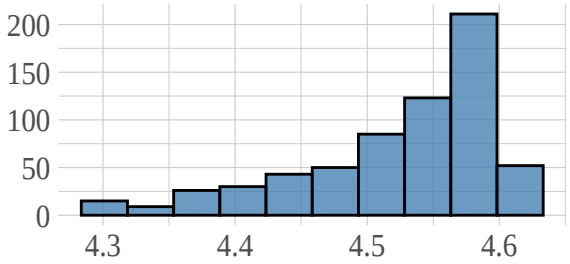
With outliers



With outliers

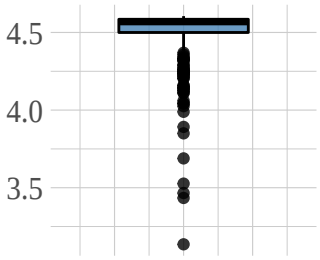


Without outliers

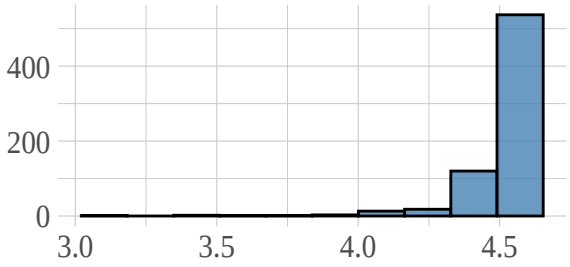


Outlier Diagnosis Plot (WithHepB)

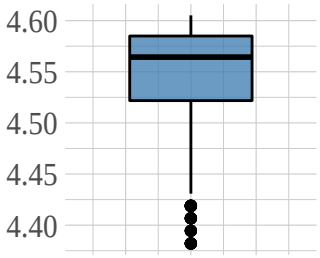
With outliers



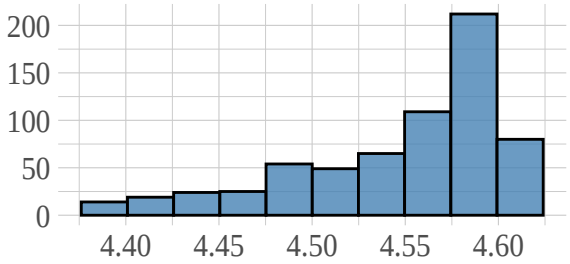
With outliers



With outliers

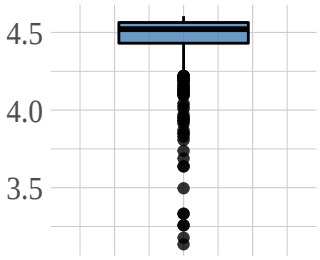


Without outliers

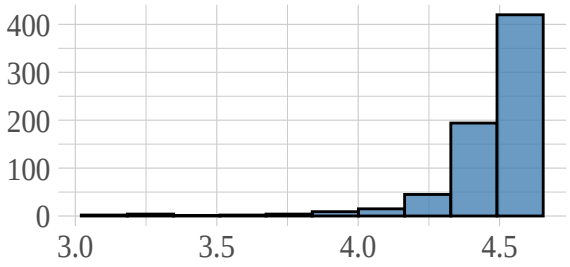


Outlier Diagnosis Plot (PctUpToDate)

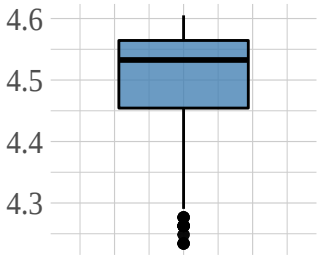
With outliers



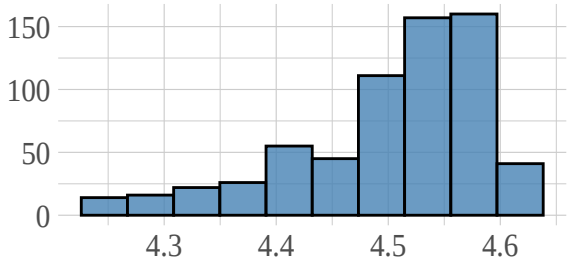
With outliers



With outliers

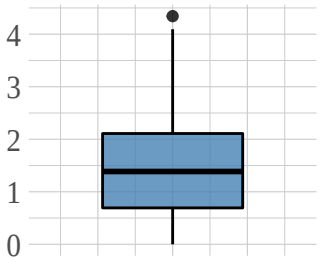


Without outliers

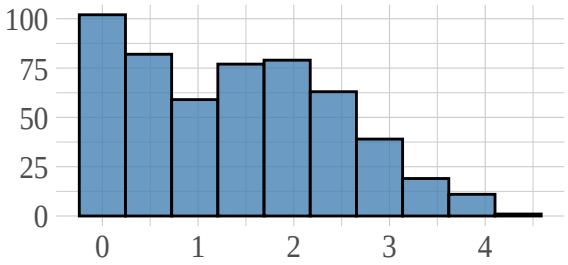


Outlier Diagnosis Plot (PctBeliefExempt)

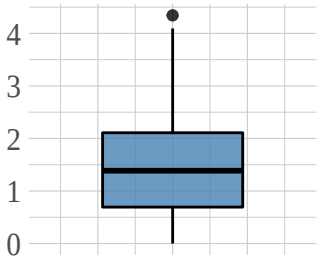
With outliers



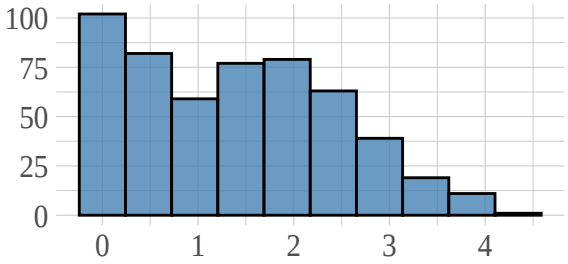
With outliers



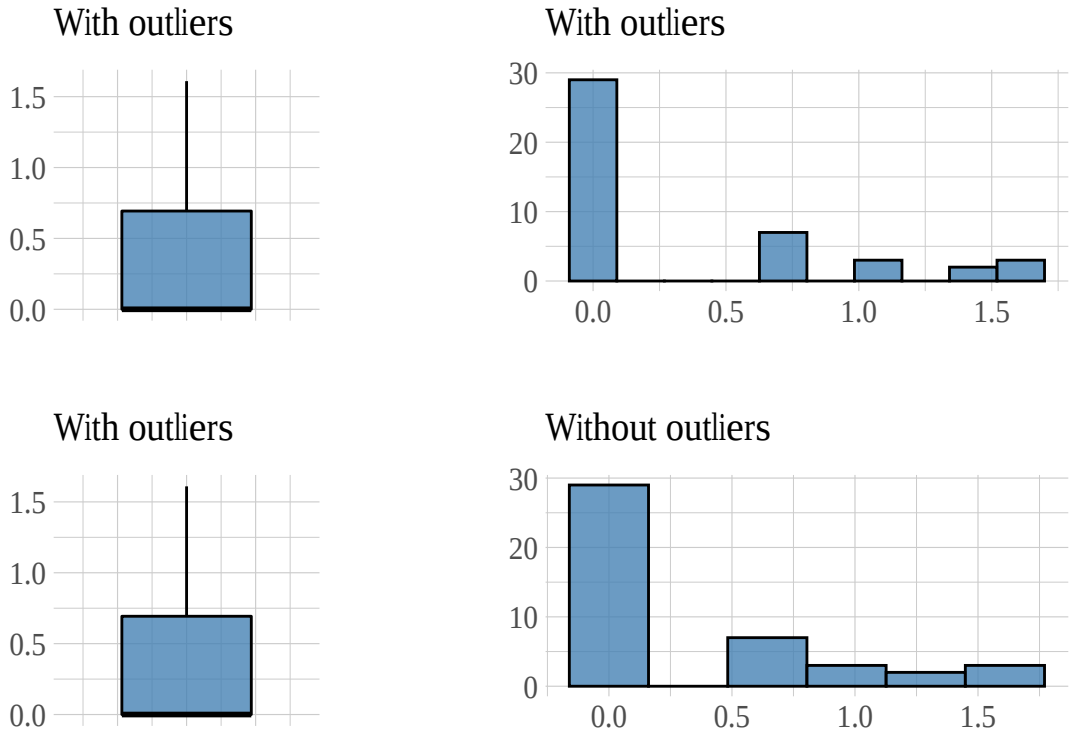
With outliers



Without outliers

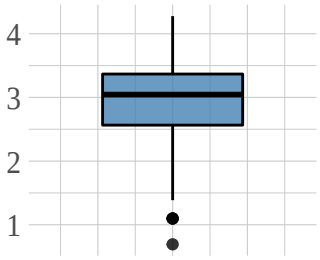


Outlier Diagnosis Plot (PctMedicalExempt)

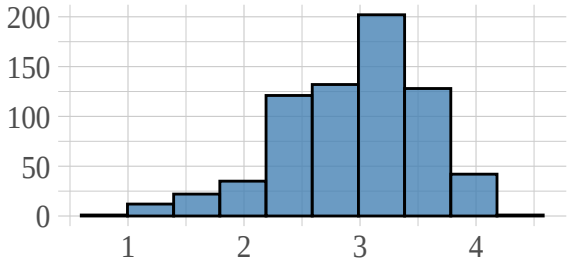


Outlier Diagnosis Plot (PctChildPoverty)

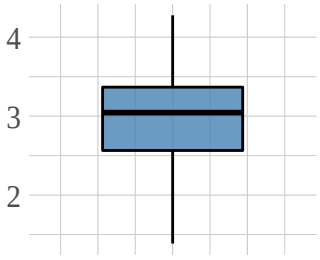
With outliers



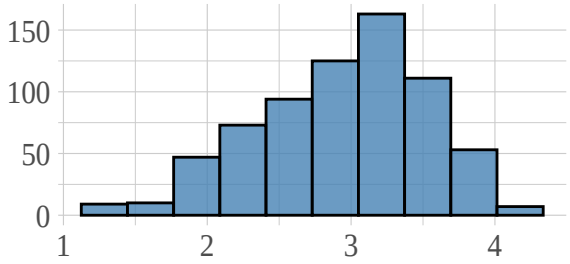
With outliers



With outliers

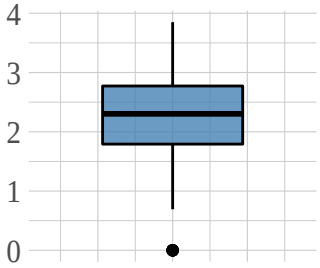


Without outliers

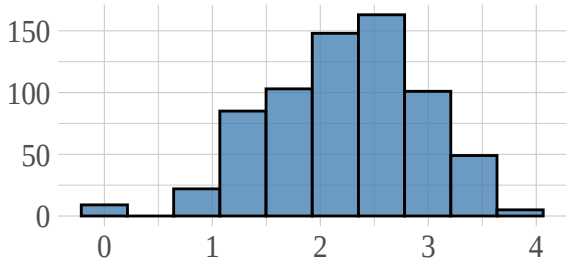


Outlier Diagnosis Plot (PctFamilyPoverty)

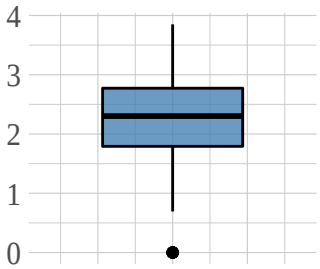
With outliers



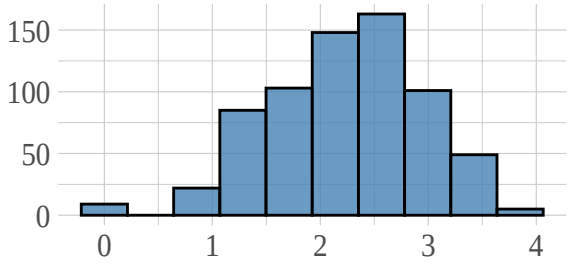
With outliers



With outliers

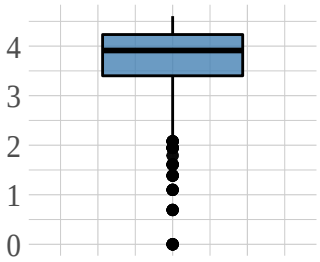


Without outliers

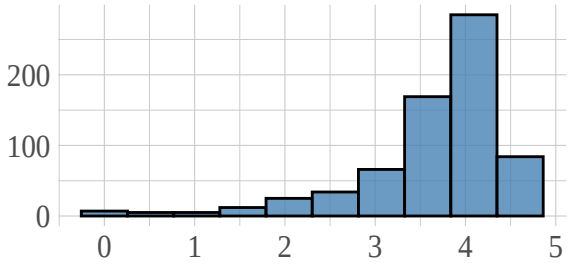


Outlier Diagnosis Plot (PctFreeMeal)

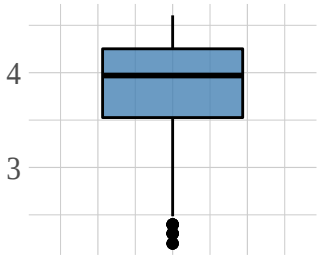
With outliers



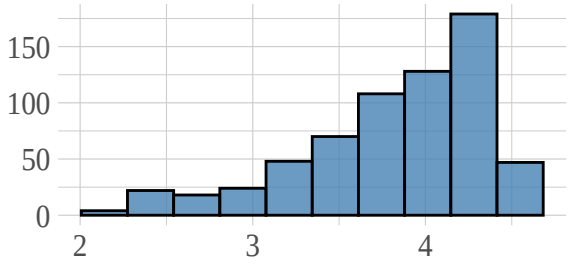
With outliers



With outliers

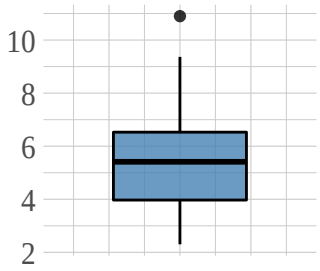


Without outliers

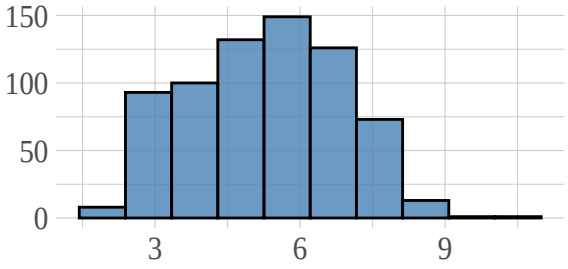


Outlier Diagnosis Plot (Enrolled)

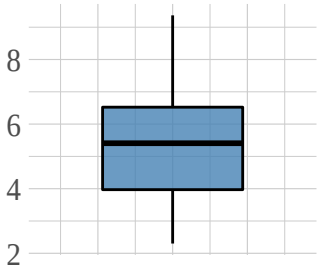
With outliers



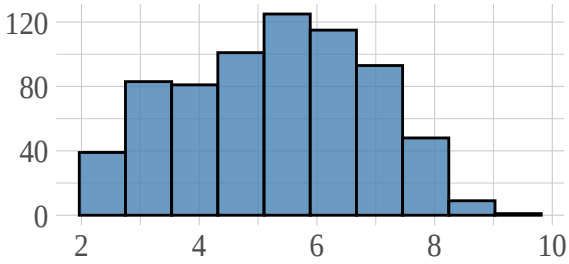
With outliers

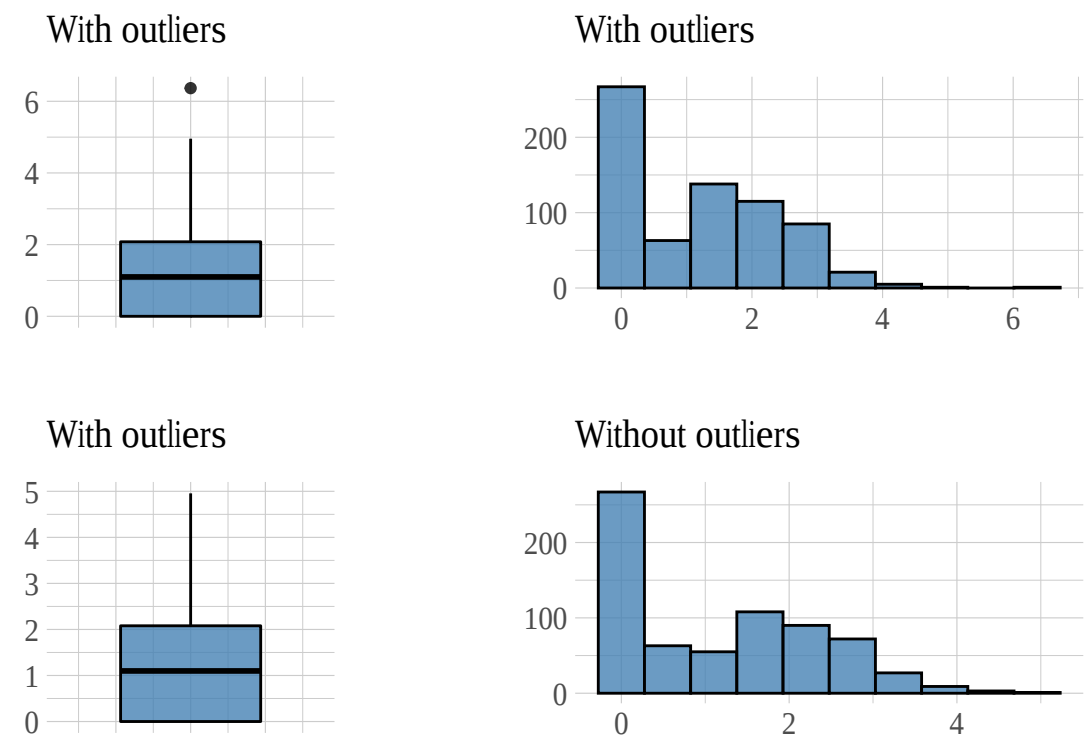


With outliers



Without outliers





In the above results, we can see that outlier count has been drastically increase for PctBeliefExempt and PctMedicalExempt. Whereas for rest of the columns outliers are of similar range.

After all this analysis let's diagnose further to check for missing values

```
library(dlookr)
diagnose(districts_new)
```

```
## # A tibble: 14 x 6
##    variables       types  missing_count missing_percent unique_count unique_rate
##    <chr>           <chr>          <int>           <dbl>        <int>       <dbl>
##  1 DistrictName    factor             0               0          696     1
##  2 WithDTP         numer~             0               0           53     0.0761
##  3 WithPolio       numer~             0               0           50     0.0718
##  4 WithMMR         numer~             0               0           51     0.0733
##  5 WithHepB        numer~             0               0           48     0.0690
##  6 PctUpToDate     numer~             0               0           60     0.0862
##  7 DistrictComple~ logic~             0               0            2     0.00287
##  8 PctBeliefExempt numer~             0               0           44     0.0632
##  9 PctMedicalExem~ numer~           247            35.5            7     0.0101
## 10 PctChildPoverty numer~             0               0           59     0.0848
## 11 PctFamilyPover~ numer~             0               0           40     0.0575
## 12 PctFreeMeal     numer~             0               0           99     0.142
## 13 Enrolled        numer~             0               0          455     0.654
## 14 TotalSchools    numer~             0               0           44     0.0632
```

The above results shows that there are 247 NA values in PctMedicalExempt. Missing values will create a

problem because most analyses will drop the entire row if a value is missing in any variable. So, I will prefer to keep the missing values column aside. 247 out of 696 records accounts to about 35% missing values in PctMedical Exempt. It is better to remove this column from further analysis.

```
# removing missing values column

mydistricts <- districts_new[,!colnames(districts_new) %in% c("PctMedicalExempt")]

diagnose(mydistricts)
```

```
## # A tibble: 13 x 6
##    variables       types  missing_count missing_percent unique_count unique_rate
##    <chr>           <chr>          <int>           <dbl>        <int>       <dbl>
##  1 DistrictName    factor             0               0          696     1
##  2 WithDTP         numer~             0               0           53     0.0761
##  3 WithPolio       numer~             0               0           50     0.0718
##  4 WithMMR         numer~             0               0           51     0.0733
##  5 WithHepB        numer~             0               0           48     0.0690
##  6 PctUpToDate     numer~             0               0           60     0.0862
##  7 DistrictComple~ logic~             0               0            2     0.00287
##  8 PctBeliefExempt numer~             0               0           44     0.0632
##  9 PctChildPoverty numer~             0               0           59     0.0848
## 10 PctFamilyPover~ numer~             0               0           40     0.0575
## 11 PctFreeMeal     numer~             0               0           99     0.142
## 12 Enrolled        numer~             0               0          455     0.654
## 13 TotalSchools    numer~             0               0           44     0.0632
```

```
describe(mydistricts)
```

```
## # A tibble: 11 x 26
##    variable          n    na   mean     sd se_mean   IQR skewness kurtosis   p00
##    <chr>         <int> <int>  <dbl>  <dbl>   <dbl> <dbl>    <dbl>    <dbl> <dbl>
##  1 WithDTP         696     0  90.0  1.09e1   0.412  10.2   -2.38     7.62     23
##  2 WithPolio       696     0  90.4  1.08e1   0.409  10     -2.47     8.17     23
##  3 WithMMR         696     0  89.9  1.12e1   0.426  11     -2.30     6.85     23
##  4 WithHepB        696     0  92.4  9.77e0   0.370   8     -3.05    12.7      23
##  5 PctUpToDate     696     0  88.0  1.25e1   0.475  12     -2.28     6.61     23
##  6 PctBeliefExe~   696     0   5.53 8.72e0   0.330   6      3.43    15.8       0
##  7 PctChildPove~   696     0  22.3  1.20e1   0.453  16      0.769    0.368     2
##  8 PctFamilyPov~   696     0  11.5  8.03e0   0.304  11      1.21     1.64      0
##  9 PctFreeMeal     696     0  48.5  2.47e1   0.934  39     -0.137   -0.918     0
## 10 Enrolled        696     0 634.   2.23e3  84.6   631.    20.3    480.       10
## 11 TotalSchools    696     0   7.28 2.41e1   0.914   7     20.0    467.        1
## # ... with 16 more variables: p01 <dbl>, p05 <dbl>, p10 <dbl>, p20 <dbl>,
## #   p25 <dbl>, p30 <dbl>, p40 <dbl>, p50 <dbl>, p60 <dbl>, p70 <dbl>,
## #   p75 <dbl>, p80 <dbl>, p90 <dbl>, p95 <dbl>, p99 <dbl>, p100 <dbl>
```

```
summary(mydistricts)
```

```
##                    DistrictName    WithDTP         WithPolio
##  ABC Unified             :  1   Min.   : 23.00   Min.   : 23.00
##  Ackerman Charter        :  1   1st Qu.: 86.75   1st Qu.: 87.00
```

```
##  Acton-Agua Dulce Unified:  1   Median : 93.00   Median : 94.00
##  Adelanto Elementary     :  1   Mean   : 89.95   Mean   : 90.36
##  Alameda Unified         :  1   3rd Qu.: 97.00   3rd Qu.: 97.00
##  Albany City Unified     :  1   Max.   :100.00   Max.   :100.00
##  (Other)                 :690
##     WithMMR          WithHepB       PctUpToDate     DistrictComplete
##  Min.   : 23.00   Min.   : 23.00   Min.   : 23.00   Mode :logical
##  1st Qu.: 86.00   1st Qu.: 90.00   1st Qu.: 84.00   FALSE:38
##  Median : 94.00   Median : 96.00   Median : 92.00   TRUE :658
##  Mean   : 89.89   Mean   : 92.37   Mean   : 87.98
##  3rd Qu.: 97.00   3rd Qu.: 98.00   3rd Qu.: 96.00
##  Max.   :100.00   Max.   :100.00   Max.   :100.00
##
##  PctBeliefExempt  PctChildPoverty PctFamilyPoverty  PctFreeMeal
##  Min.   : 0.000   Min.   : 2.00   Min.   : 0.00    Min.   :  0.00
##  1st Qu.: 1.000   1st Qu.:13.00   1st Qu.: 5.00    1st Qu.: 30.00
##  Median : 2.500   Median :21.00   Median : 9.50    Median : 49.50
##  Mean   : 5.534   Mean   :22.29   Mean   :11.48    Mean   : 48.48
##  3rd Qu.: 7.000   3rd Qu.:29.00   3rd Qu.:16.00    3rd Qu.: 69.00
##  Max.   :77.000   Max.   :72.00   Max.   :47.00    Max.   :100.00
##
##     Enrolled        TotalSchools
##  Min.   :   10.0   Min.   :  1.000
##  1st Qu.:   53.0   1st Qu.:  1.000
##  Median :  224.0   Median :  3.000
##  Mean   :  634.3   Mean   :  7.276
##  3rd Qu.:  684.2   3rd Qu.:  8.000
##  Max.   :54238.0   Max.   :582.000
##
```

From the above results, we can see that there are no missing values now and we can proceed further.

Descriptive Reporting

1. *Basic Introductory Paragraph*

In your own words, write about three sentences of introduction addressing the staff member in the state legislator's office. Frame the problem/topic that your report addresses.

The data for vaccination rates in the Californian school districts are available for DTP1, HepB, Pol3 and MMR, which is going help further analysis on understanding how these rates are varying over the years using time series data. Whereas we have DTP1, HepB, Pol3, Hib3 and MCV1 rates in the time series, which makes us little difficult to understand how MMR rate is varying over the years from 1980:2017. Also as there are no rates for Hib3 and MCV1 district wise, we won't be able to consider the effects of these on other data variables such as percentage of students with completely up-to-date vaccines. Apart from this, with the available data we can result out the analysis to understand which factors are affecting the vaccination rates and provide some suggestions to improve them. Also, we can enhance the process of district's reporting so that all districts would be able to complete their reporting.

The time series data is helping to understand the growth trends in Usvaccines, which could be used further to understand which factors are causing to keep the trend high and if somewhere it is steep down then what was the cause for that.

2. *Descriptive Overview of U.S. Vaccinations*

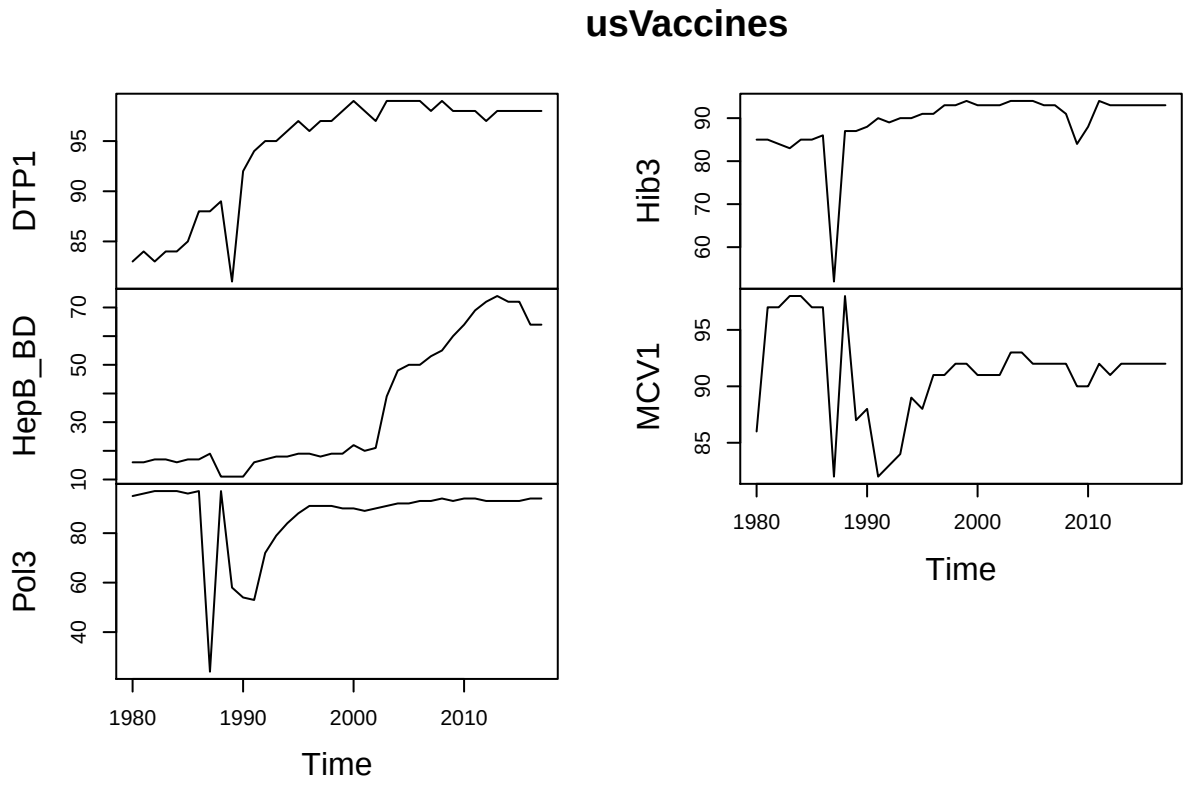
You have U.S. vaccination data going back 38 years, but the staff member is only interested in recent vaccination rates as a basis of comparison with California schools.

```
# we can get the latest vaccination rates with below code:

vccn_latest_rates <- window(usVaccines, start = 2012, end = 2017)

vccn_latest_rates
```

```
## Time Series:
## Start = 2012
## End = 2017
## Frequency = 1
##      DTP1 HepB_BD Pol3 Hib3 MCV1
## 2012   97      72   93   93   91
## 2013   98      74   93   93   92
## 2014   98      72   93   93   92
## 2015   98      72   93   93   92
## 2016   98      64   94   93   92
## 2017   98      64   94   93   92
```

a. ***How have U.S. vaccination rates varied over time?***

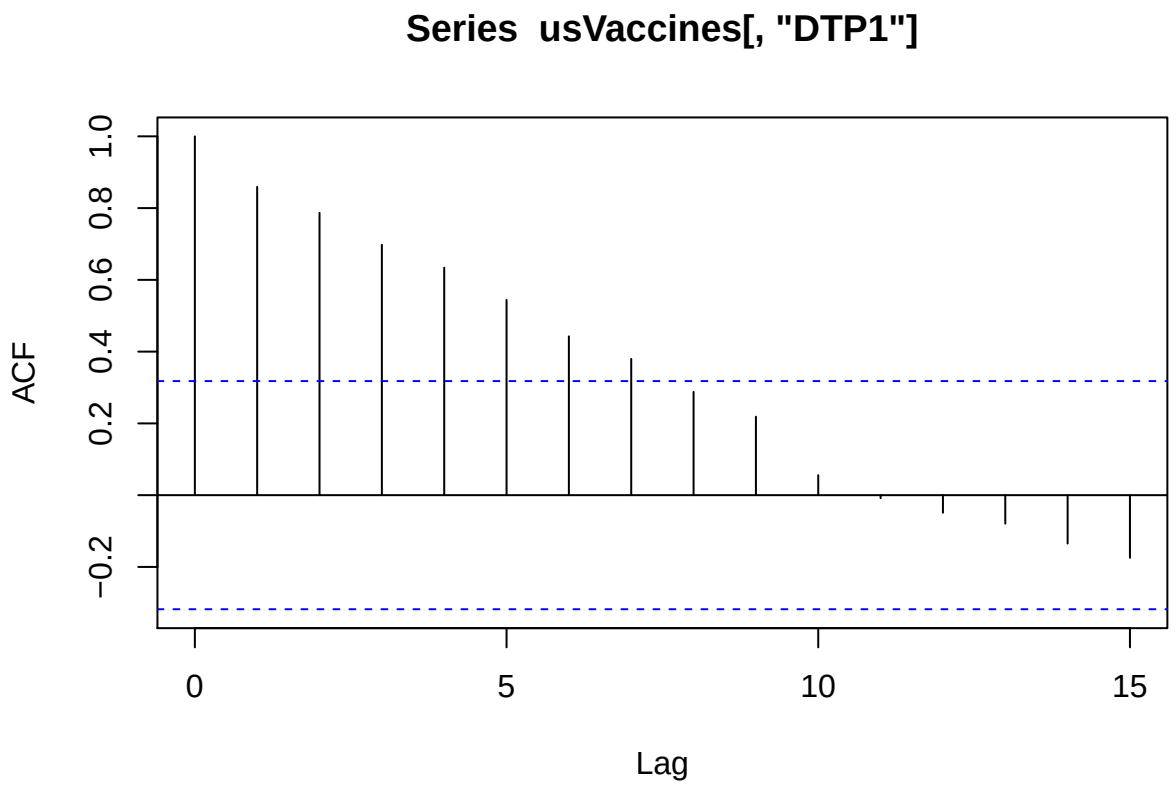
```
# running time series plots
plot.ts(usVaccines)
```



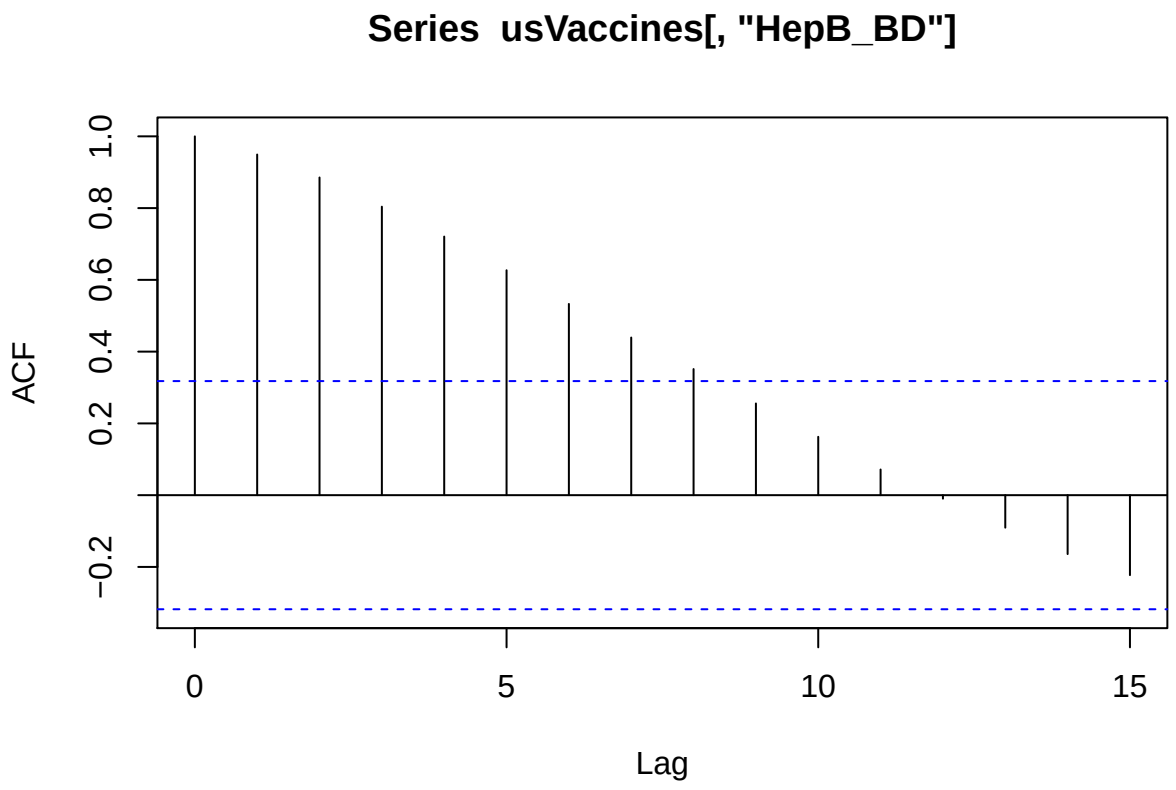
In the above plots, as we can see that in the start at 1980 only HepB_BD vaccine's rate was lower around 20; for all other vaccines the rate were around 80. Also we can see that, around year 1987 to 1989 the rates were decresed almost for all vaccines. The same thing happened for MCV1 vaccine rate, rates decreased with steep down curve again in the year of 1990 and 1991; and Pol3 has decreased to some extent again in the year of 1992. After these years we can see that almost all vaccine rates are increasing after or at year 2000 with small jitters in the rate values. We can see that ther rates are almost constant at around 2016 and 2017.

b. ***Are there notable trends or cyclical variation in U.S. vaccination rates?***

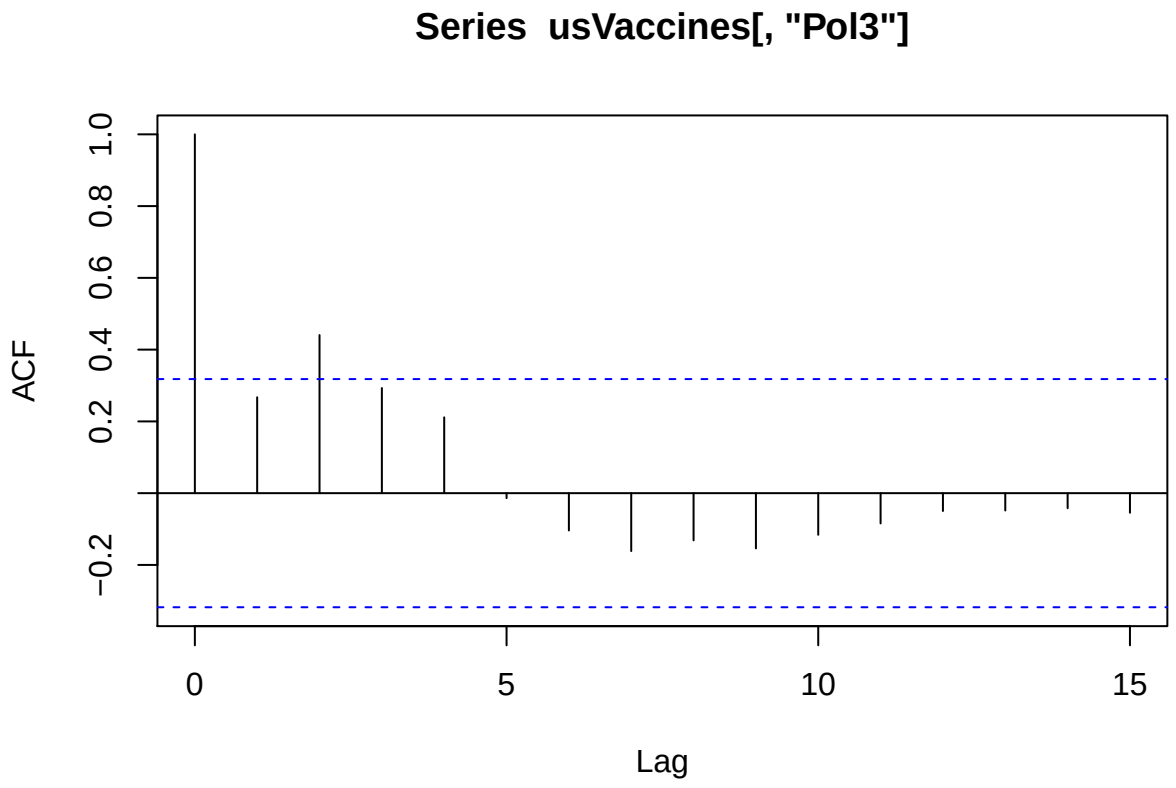
```
acf(usVaccines[,"DTP1"])
```



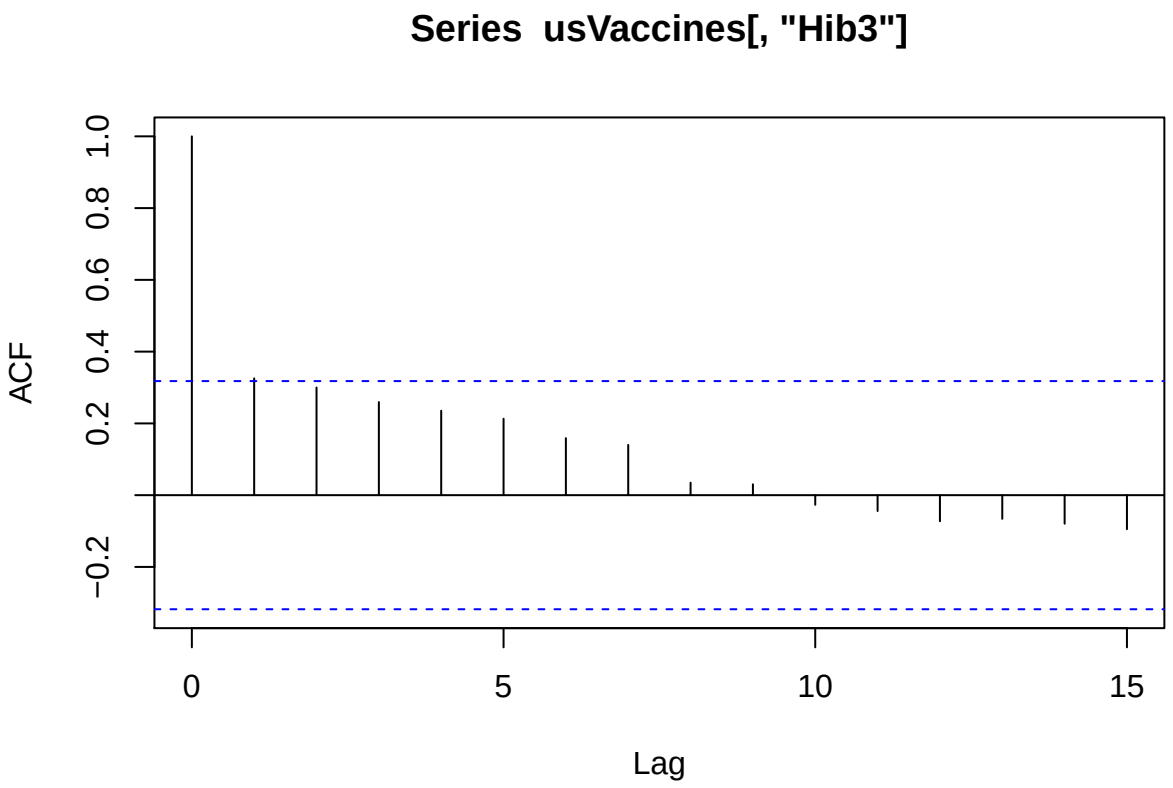

```
acf(usVaccines[,"HepB_BD"])
```



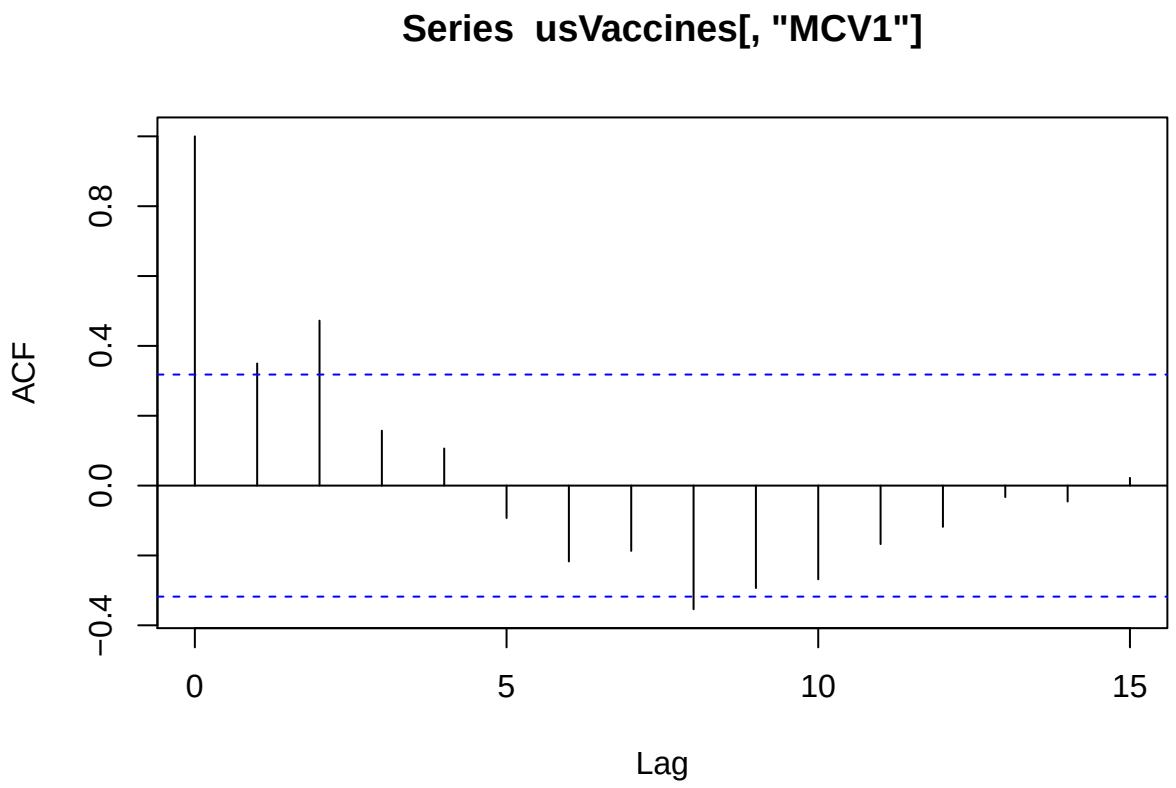

```
acf(usVaccines[,"Pol3"])
```




```
acf(usVaccines[,"Hib3"])
```




```
acf(usVaccines[,"MCV1"])
```



The auto correlation function, often abbreviated as the ACF, correlates a variable with itself at a later time period. The height of each little line shows the sign and magnitude of the correlation of the original variable correlated with itself at different amounts of lag. The blue horizontal dotted lines show the threshold of statistical significance for positive and negative correlations. For a stationary time-series process, all of the lagged correlations (other than zero lag) should be nonsignificant and there should be no pattern to the size of the correlations or to the variations between positive and negative correlations.

In the above plots, exists 7 autocorrelations that are statistically significant for DTP1, 8 autocorrelations that are statistically significant for HepB_BD, 1 autocorrelation which is statistically significant for Pol3, 1 autocorrelations that are statistically significant for Hib3 and 3 autocorrelations that are statistically significant for MCV1. The height of the bar pokes out above or below the horizontal dotted lines shows that there is significantness and not counting the ever-present perfect autocorrelation at lag = 0.

With these autocorrelations, the problem exists in the the overall pattern. It is evident from the pattern of positive and negative autocorrelations for DTP1 and HepB_BD is that Triangular patter is present, there is only single peak pattern for Pol3 and Hib3, and that the sinusoidal pattern is present in the time series of the data for MCV1. Thus, the process of whitening is imperfect. In order to confirm these analysis, we can also perform an inferential test about whether or not this is a stationary process by using the augmented Dickey–Fuller test and adf.test().

Conducting adf test

```
#install.packages("tseries")
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'quantmod':
```

```
##   method            from
##   as.zoo.data.frame zoo
```

```
adf.test(usVaccines[,"DTP1"])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, "DTP1"]
## Dickey-Fuller = -0.87963, Lag order = 3, p-value = 0.943
## alternative hypothesis: stationary
```

```
adf.test(usVaccines[,"HepB_BD"])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, "HepB_BD"]
## Dickey-Fuller = -1.9729, Lag order = 3, p-value = 0.5839
## alternative hypothesis: stationary
```

```
adf.test(usVaccines[,"Pol3"])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, "Pol3"]
## Dickey-Fuller = -2.3918, Lag order = 3, p-value = 0.4202
## alternative hypothesis: stationary
```

```
adf.test(usVaccines[,"Hib3"])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, "Hib3"]
## Dickey-Fuller = -2.3377, Lag order = 3, p-value = 0.4414
## alternative hypothesis: stationary
```

```
adf.test(usVaccines[,"MCV1"])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, "MCV1"]
## Dickey-Fuller = -2.5324, Lag order = 3, p-value = 0.3652
## alternative hypothesis: stationary
```

From the above results of adf test we can see that p value for all is greater than assumed threshold value of 0.05. Thus we failed to reject the null hypothesis, which says that the process is non stationary and there exists trends and cyclicality.

c. ***What are the mean U.S. vaccination rates when including only recent years in the calculation of the mean (examine your answers to the previous question to decide what a reasonable recent period is, i.e., a period during which the rates are relatively constant)?***

```
# Filtering out different vaccines rates over recent periods i.e. 2016 and 2017 as they have constant r
vccn_latest_con_rts <- window(usVaccines, start = 2016, end = 2017)

vccn_latest_con_rts
```

```
## Time Series:
## Start = 2016
## End = 2017
## Frequency = 1
##      DTP1 HepB_BD Pol3 Hib3 MCV1
## 2016   98      64   94   93   92
## 2017   98      64   94   93   92
```

```
# Converting different vaccines rates over recent periods i.e. 2016 and 2017 as they have constant rate
usvaccine_df <- data.frame(vccn_latest_con_rts)

usvaccine_df
```

```
##   DTP1 HepB_BD Pol3 Hib3 MCV1
## 1   98      64   94   93   92
## 2   98      64   94   93   92
```

```
# calculating mean of different vaccines over recent periods i.e. 2016 and 2017 as they have constant r
mean(usvaccine_df$DTP1)
```

```
## [1] 98
```

```
mean(usvaccine_df$HepB_BD)
```

```
## [1] 64
```

```
mean(usvaccine_df$Pol3)
```

```
## [1] 94
```

```
mean(usvaccine_df$Hib3)
```

```
## [1] 93
```

```
mean(usvaccine_df$MCV1)
```

```
## [1] 92
```

As we can see from above results, 98 is the mean rate for DTP1, 64 is the mean rate for HepB_BD, 94 is the mean rate for Pol3, 93 is the mean rate for Hib3 and 92 is the mean rate for MCV1. HepB_BD has lowest mean rate among other vaccines in recent years.

3. *Descriptive Overview of California Vaccinations*

Your districts dataset contains four variables that capture the individual vaccination rates by district: WithDTP, WithPolio, WithMMR, and WithHepB.

a. ***What are the mean levels of these variables across districts?***

```
# calculating the mean of available vaccine rates

mean(districts_new$WithDTP)
```

```
## [1] 89.95259
```

```
mean(districts_new$WithPolio)
```

```
## [1] 90.36207
```

```
mean(districts_new$WithMMR)
```

```
## [1] 89.88793
```

```
mean(districts_new$WithHepB)
```

```
## [1] 92.36638
```

The above results shows that from districts_new data WithDTP has mean 89.95259, WithPolio has mean 90.36207, WithMMR has mean of 89.88793 and WithHepB 92.36638.

b. ***Among districts, how are the vaccination rates for individual vaccines related? In other words, if there are students with one vaccine, are students likely to have all of the others?***

```
# Creating correlation matrix for all vaccine rates available in the Districts_new dataset
cor(districts_new[2:5])
```

```
##             WithDTP WithPolio   WithMMR  WithHepB
## WithDTP   1.0000000 0.9817596 0.9796205 0.8902101
## WithPolio 0.9817596 1.0000000 0.9669604 0.9028865
## WithMMR   0.9796205 0.9669604 1.0000000 0.8892113
## WithHepB  0.8902101 0.9028865 0.8892113 1.0000000
```

From the above results of correlation matrix we can see that all the vaccine rates are highly correlated with each other. Because of this high correlation we can say that if there are students with one vaccine, there is higb chance that students likely to have all of the others vaccines.

c. ***How do these Californian vaccination levels compare to U.S. vaccination levels (recent years only)? Note any patterns you notice.***

As we can see from the above analysis in question 2 c) and 3 a) we can see that

```r
mtrx_vaccine_rates <- matrix(data = c(98,89.95259,64,92.36638,94,90.36207), nrow = 2, ncol = 3)

rownames(mtrx_vaccine_rates) <- c("Us Vaccine Rates","Californian Vacine Rates")

colnames(mtrx_vaccine_rates) <- c("DTP1", "HepB_BD", "Pol3")

mtrx_vaccine_rates
```

```
##                              DTP1  HepB_BD     Pol3
## Us Vaccine Rates         98.00000 64.00000 94.00000
## Californian Vacine Rates 89.95259 92.36638 90.36207
```

In the above results, we can see that the DTP1 and Pol3 Californian rates are less compared to Us Vaccine Rate less, but for HepB_BD Californian rates are greater compared to Us Vaccine Rates.

4. *Conclusion Paragraph for Vaccination Rates*

Provide one or two sentences of your professional judgment about where California school districts stand with respect to vaccination rates and in the larger context of the U.S.

The mean of the vaccination rates for Tetanus, Hepatitis B, Polio and MMR in all of the Californian districts are about 90. Also, some districts have 100% vaccination rates while some are still completing with all vaccinations. From the available datasets, The vaccine rates shows upward trend in the recent years and that trend is getting reflected in the Californian vaccination rates. After comparing with vaccination rates in the the United States and Californian states we can say that the Californian districts are not doing bad in getting done with completely up-to-date vaccines.

Inferential Reporting

For every item below except 7, use PctChildPoverty, PctFamilyPoverty, Enrolled, and TotalSchools as the four predictors. Explore the data and transform variables as necessary to improve prediction and/or interpretability. Be sure to include appropriate diagnostics and modify your analyses as appropriate.

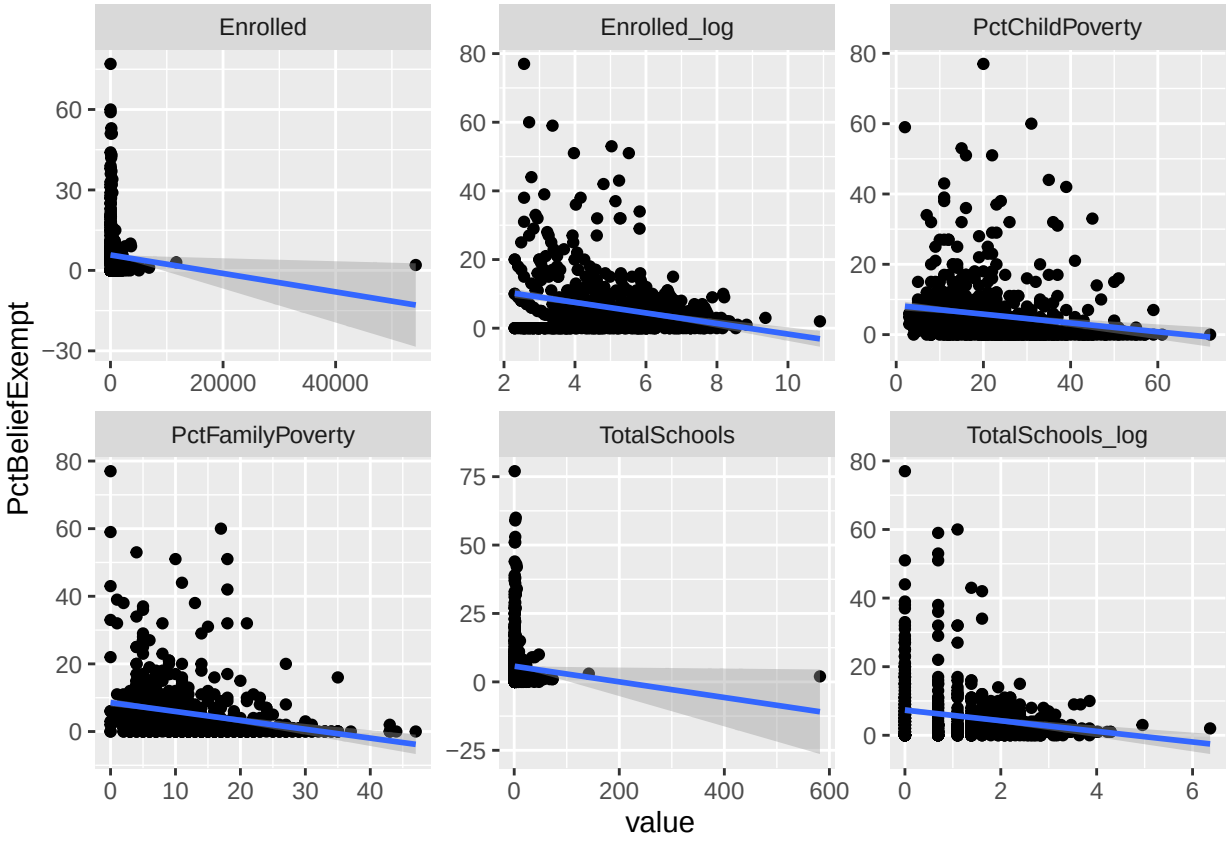
5. *Which of the four predictor variables predicts the percentage of all enrolled students with belief exceptions?*

Creating new dataset to check the effect of PctChildPoverty, PctFamilyPoverty, Enrolled, and TotalSchools as the four predictors on PctBeliefExempt

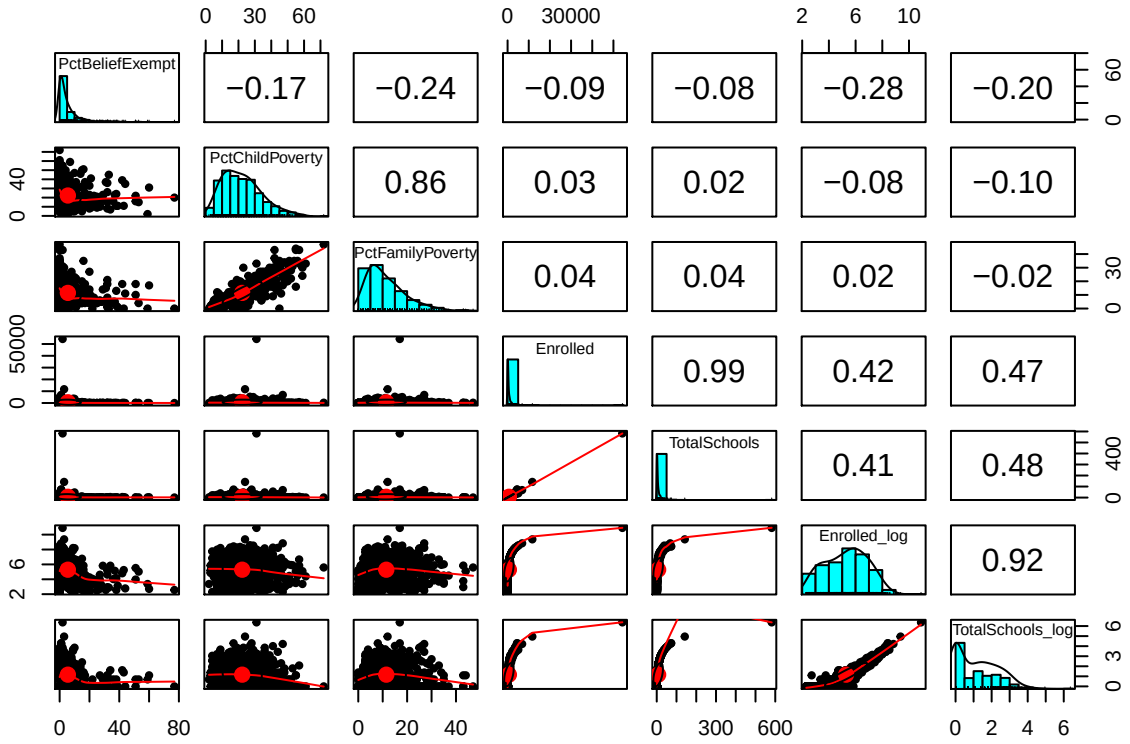
```r
belief_exempt <- subset(districts_new, select = c("PctBeliefExempt","PctChildPoverty",
                                                  "PctFamilyPoverty","Enrolled","TotalSchools"))

# Applying log transformations on Enrolled and TotalSchools columns
belief_exempt$Enrolled_log <- log(belief_exempt$Enrolled)

belief_exempt$TotalSchools_log <- log(belief_exempt$TotalSchools)


# checking the scatter plots for each variables with respect to the PctBeliefExempt
belief_exempt %>% pivot_longer(-PctBeliefExempt, names_to="variable", values_to="value", values_drop_na
```

```
            ggplot(aes(x=value, y=PctBeliefExempt)) + geom_point() +
                 geom_smooth(method = "lm") + facet_wrap( ~ variable, scales="free")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# checking the histograms, scatterplots and correlation betwen each variables in the plots
pairs.panels(belief_exempt)
```



The above plots shows that how log tranformations have helped Enrolled and TotalSchools variables to get a better linear relationship with PctBeliefExempt compared to when they were not log transformed, as well as in reducing the skewness in the right tail. We can see that accumulation of points across blue line is increased in the scatter plots of Enrolled_log and TotalSchools_log variables. In addition to this, in other scatter plots, we can see sufficient accumulation of points across blue line.

Also in above calculations, I have not applied log transformations to percentage columns in newly created belief_exempt dataset, because usually the percentage values varies from 0 to 100 and log transforming them can certainly hamper the dataset meaning. We can also verify from the outliers_plot that outlier values are increasing and varying drastically on log transformed data.

```
# checking the correlation with below correlation matrix
round(cor(belief_exempt),3)
```

```
##                  PctBeliefExempt PctChildPoverty PctFamilyPoverty Enrolled
## PctBeliefExempt            1.000          -0.172           -0.242   -0.088
## PctChildPoverty           -0.172           1.000            0.864    0.026
## PctFamilyPoverty          -0.242           0.864            1.000    0.044
## Enrolled                  -0.088           0.026            0.044    1.000
## TotalSchools              -0.079           0.021            0.038    0.994
## Enrolled_log              -0.279          -0.080            0.018    0.416
## TotalSchools_log          -0.205          -0.099           -0.020    0.474
##                  TotalSchools Enrolled_log TotalSchools_log
## PctBeliefExempt        -0.079       -0.279           -0.205
## PctChildPoverty         0.021       -0.080           -0.099
## PctFamilyPoverty        0.038        0.018           -0.020
```

```
## Enrolled               0.994        0.416           0.474
## TotalSchools           1.000        0.406           0.480
## Enrolled_log           0.406        1.000           0.917
## TotalSchools_log       0.480        0.917           1.000
```

From the above results we can see that there is high positive correlation [0.864] between PctChildPoverty and PctFamilyPoverty. In addition to this there exists high positive correlation between Enrolled and TotalSchools [0.994]; and Enrolled_log and TotalSchools_log [0.917]. In case of PctBeliefExempt, there exists medium negative correlation with Enrolled_log [-0.279] which shows r value, and tells us that if there is 1 unit increase in Enrolled_log then there is a likely possibility of PctBeliefExempt getting reduced by -0.279.

```
belief_lm_out <- lm(PctBeliefExempt~PctChildPoverty+PctFamilyPoverty+Enrolled_log+TotalSchools_log, data

# checking the effects of multicolinearity in the model predictors
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

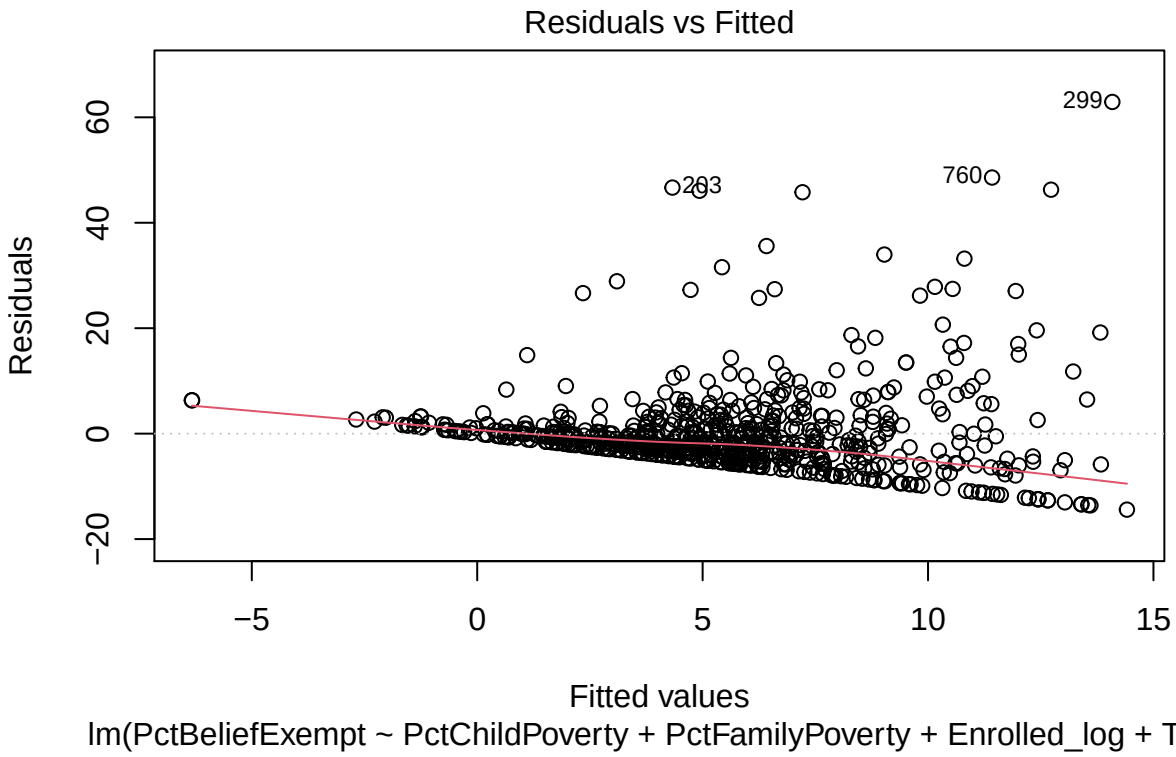
```
vif(belief_lm_out)
```

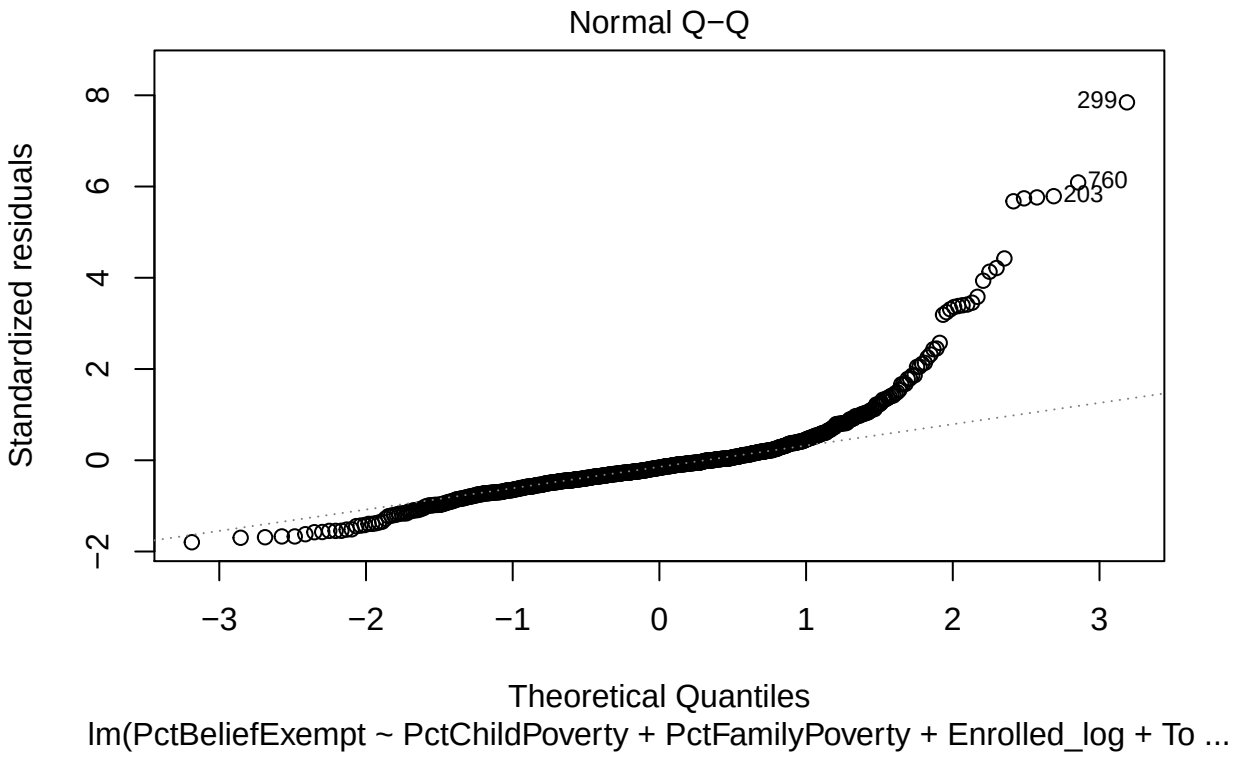
```
##  PctChildPoverty PctFamilyPoverty     Enrolled_log TotalSchools_log
##         4.109105         4.102388         6.407923         6.341736
```

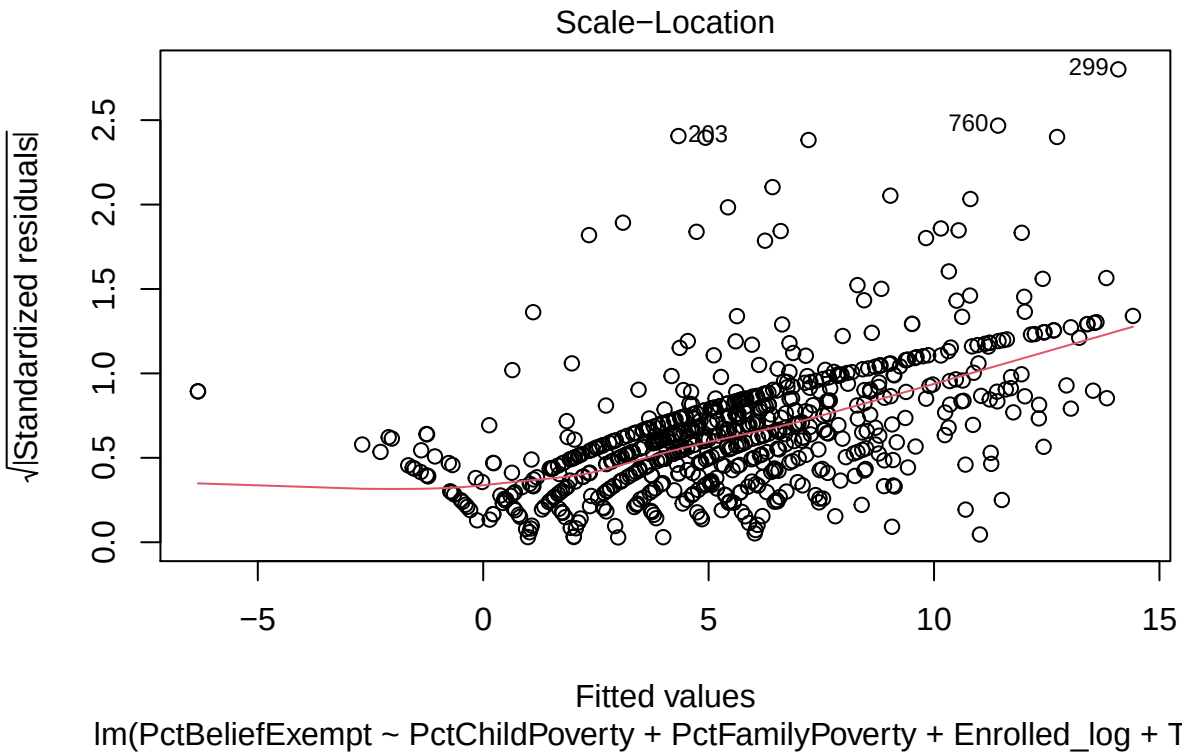
The multicolinearity check is passed as all the variance inflation factors for predictor values are less than 10, but we need to careful for variance inflation factors > 5 for Enrolled_log and TotalSchools_log.

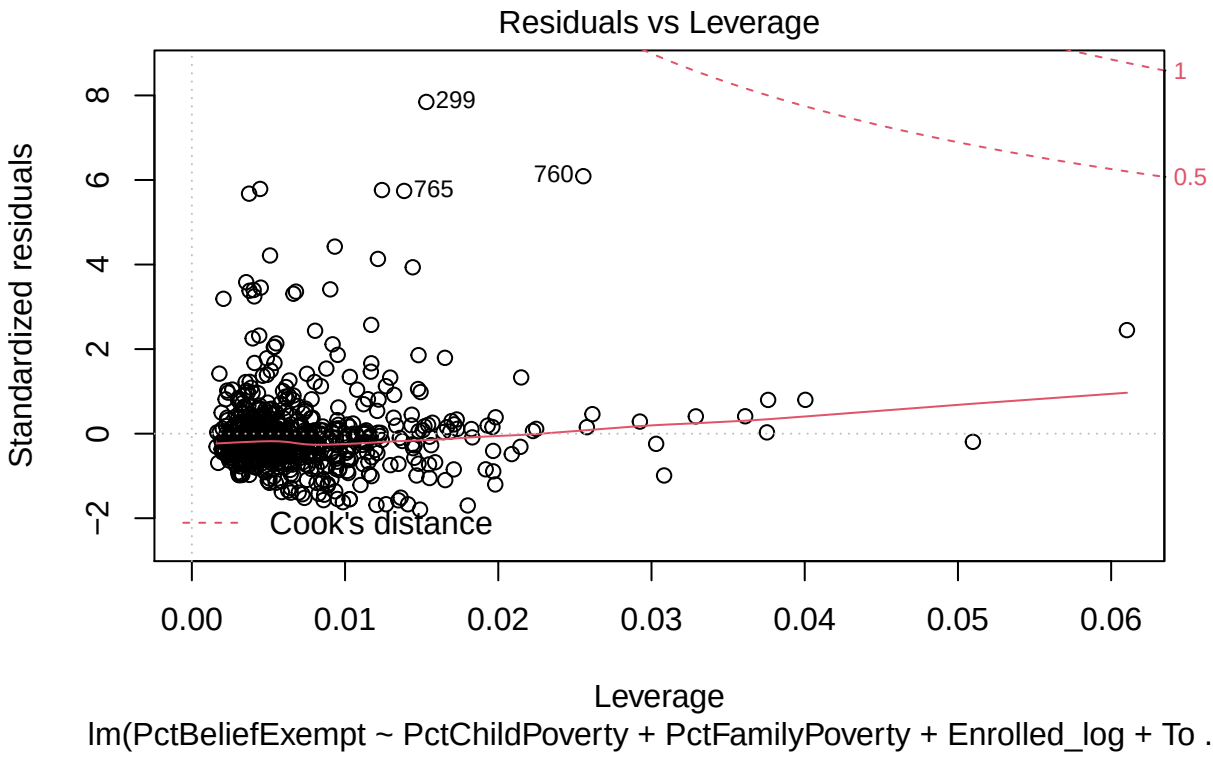
Analyze the model plots and model's residual histogram plot to check the linearity

```
# plotting model plots
plot(belief_lm_out)
```

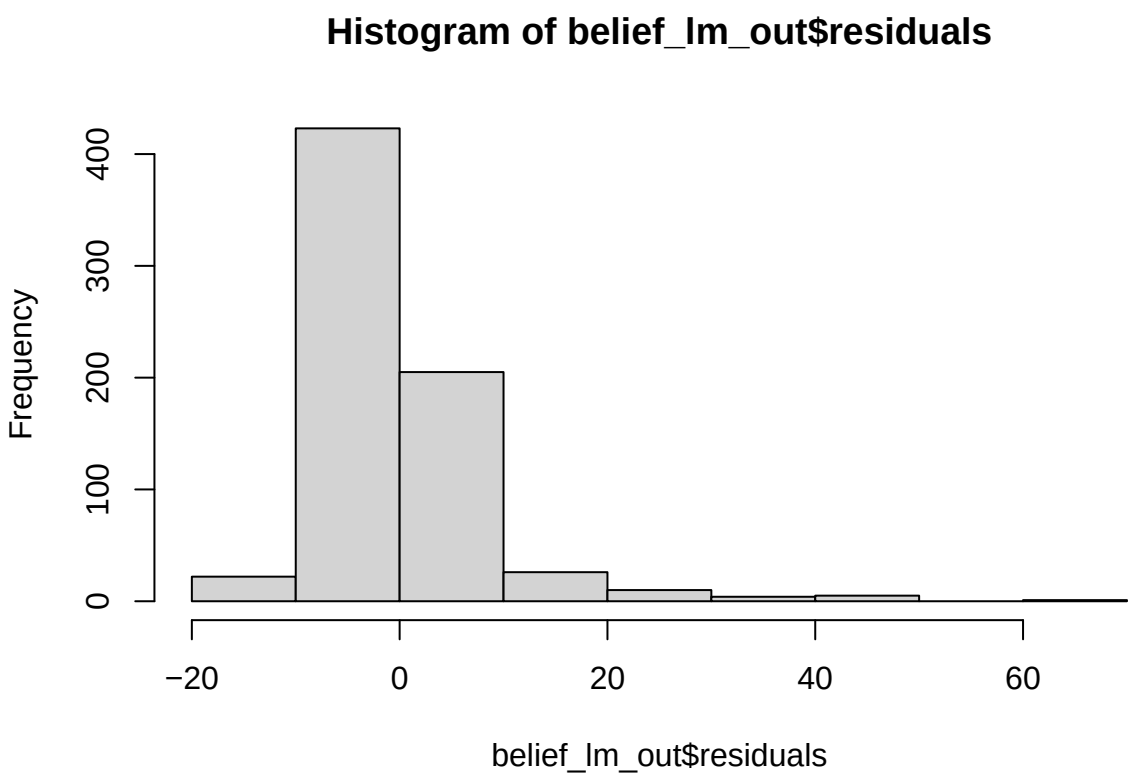

Residuals vs Fitted
Residuals
Fitted values
lm(PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty + Enrolled_log + To ...
299
203
760
60
40
20
0
−20
−5
0
5
10
15


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty + Enrolled_log + To ...
299
760
203
−2
0
2
4
6
8
−3
−2
−1
0
1
2
3


Scale−Location
299
203
760
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
−5
0
5
10
15
Fitted values
lm(PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty + Enrolled_log + To ...




```
# Plotting histograms of residuals
hist(belief_lm_out$residuals)
```




```
# checking mean and median for residuals
mean(belief_lm_out$residuals)
```

```
## [1] 5.767706e-17
```

```
median(belief_lm_out$residuals)
```

```
## [1] -1.337174
```

In above plots we can see that, there is some skewness in the right tail of the residual histogram plot. This result is also reflected in the Q-Q plot of the model. The Q-Q plot is having a slight curve on the higher side. Also, it is visible that in all plots the accumulation of points is more along the red line and even though there is no outlier effect, it is not clearly having strong linear relationship with dependent variable.

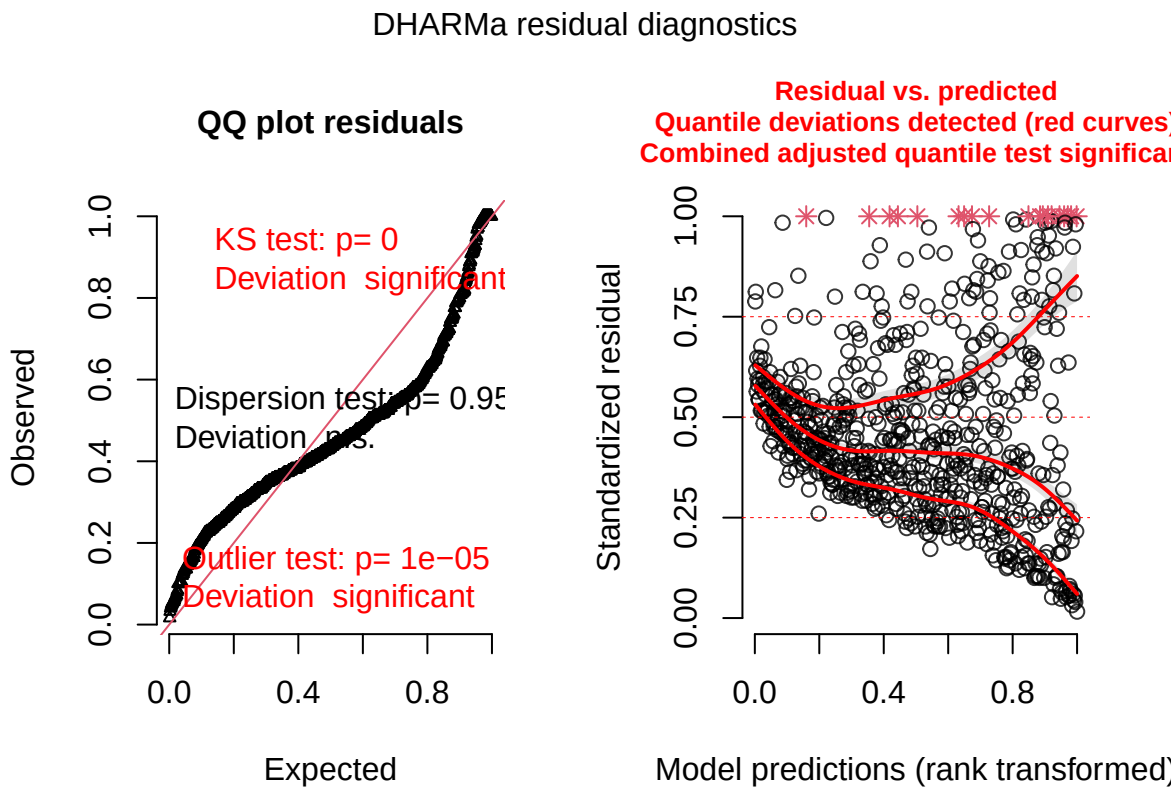
```
library(DHARMa)
```

```
## Warning: package 'DHARMa' was built under R version 4.0.5
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

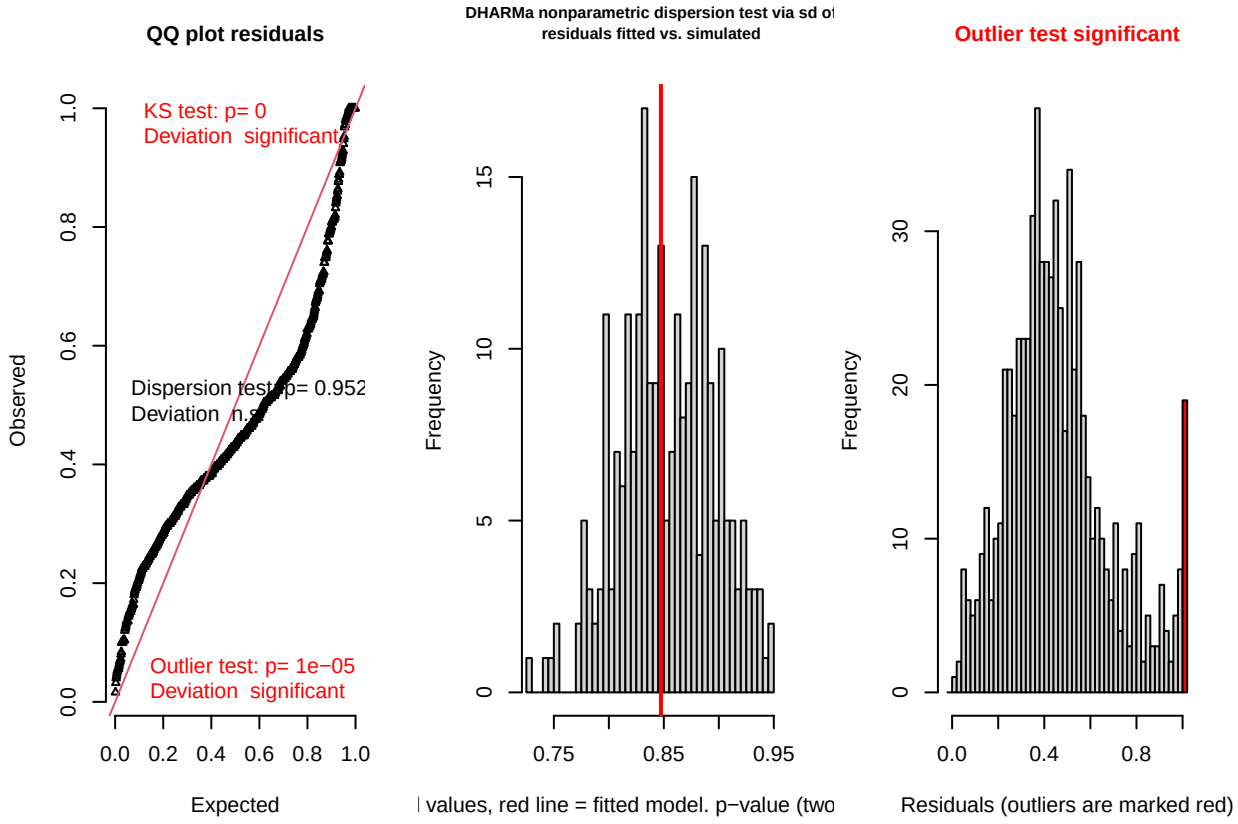
```
## This is DHARMa 0.4.1. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa')
```

```
simulationOut <- simulateResiduals(fittedModel = belief_lm_out, n = 250)

plot(simulationOut)
```



The 'DHARMa' package uses a simulation-based approach to create readily interpretable scaled (quantile) residuals for fitted (generalized) linear mixed models. The resulting residuals are standardized to values between 0 and 1 and can be interpreted as intuitively as residuals from a linear regression. The package also provides a number of plot and test functions for typical model misspecification problems, such as over/underdispersion, zero-inflation, and residual spatial and temporal autocorrelation. In case of above plots, as Q-Q plot is not exactly along the red line, we can say that there is no strong liner relationship with dependent variable.

```
# The test are run as follows:
invisible(testResiduals(simulationOut))
```




```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.18643, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.99311, p-value = 0.952
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 19, observations = 696, p-value =
```

```
## 5.315e-06
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.01651396 0.04230330
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.02729885
```

The above results we can see that Q-Q plot has a small curve and Dispersion plot is normally distributed with some skewness

```
summary(belief_lm_out)
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty +
##     Enrolled_log + TotalSchools_log, data = belief_exempt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.412  -3.707  -1.337   1.369  62.912
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      20.86116    2.02877  10.283  < 2e-16 ***
## PctChildPoverty   0.02642    0.05199   0.508 0.611417
## PctFamilyPoverty -0.28061    0.07735  -3.628 0.000307 ***
## Enrolled_log     -2.84671    0.49054  -5.803 9.91e-09 ***
## TotalSchools_log  2.02038    0.66974   3.017 0.002650 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.081 on 691 degrees of freedom
## Multiple R-squared:  0.1457, Adjusted R-squared:  0.1408
## F-statistic: 29.47 on 4 and 691 DF,  p-value: < 2.2e-16
```

In the above experiment the model belief_lm_out has, Multiple R-squared = 0.1457 and Adjusted R-squared = 0.1408 represents the proportion of about 14% variation in PctBeliefExempt (about its mean) explained by the multiple linear regression model with predictors in the model.

Calculating beta weights below to verify how the standardized variations have been changed for all predictors

```
# checking beta weights to see standardized deviation
#install.packages("lm.beta")
library(lm.beta)
summary(lm.beta(belief_lm_out))
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty +
##     Enrolled_log + TotalSchools_log, data = belief_exempt)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -14.412  -3.707  -1.337   1.369  62.912
##
## Coefficients:
##                   Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)      20.86116      0.00000    2.02877  10.283  < 2e-16 ***
## PctChildPoverty   0.02642      0.03623    0.05199   0.508 0.611417
## PctFamilyPoverty -0.28061     -0.25836    0.07735  -3.628 0.000307 ***
## Enrolled_log     -2.84671     -0.51651    0.49054  -5.803 9.91e-09 ***
## TotalSchools_log  2.02038      0.26711    0.66974   3.017 0.002650 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.081 on 691 degrees of freedom
## Multiple R-squared:  0.1457, Adjusted R-squared:  0.1408
## F-statistic: 29.47 on 4 and 691 DF,  p-value: < 2.2e-16
```

As we can see in the above experiment, Standardized deviations have been reduced to greater extent from std.error.

Now conducting a Bayesian linear regression analysis, using the facilities in the BayesFactor package.

```
library(BayesFactor)
```

```
## Warning: package 'BayesFactor' was built under R version 4.0.4
```

```
## Loading required package: coda
```

```
## Warning: package 'coda' was built under R version 4.0.4
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richarddmorey
##
## Type BFManual() to open the manual.
## ************
```

```
# calculating Bayes Factor
belief_lmbf_out <- lmBF(PctBeliefExempt~PctChildPoverty+PctFamilyPoverty+Enrolled_log+TotalSchools_log,

belief_lmbf_out
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFamilyPoverty + Enrolled_log + TotalSchools_log : 2.752423e+19 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
# Running MCMC test on belief_lmbf_out using posterior distributions
belief_lmbf_out1 <- lmBF(PctBeliefExempt~PctChildPoverty+PctFamilyPoverty+Enrolled_log+TotalSchools_log

summary(belief_lmbf_out1)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                      Mean      SD  Naive SE Time-series SE
## mu                5.54123 0.30685 0.0030685      0.0030685
## PctChildPoverty   0.02587 0.05146 0.0005146      0.0005146
## PctFamilyPoverty -0.27313 0.07709 0.0007709      0.0007709
## Enrolled_log     -2.77225 0.48701 0.0048701      0.0048701
## TotalSchools_log  1.96239 0.66338 0.0066338      0.0066338
## sig2             65.37208 3.52483 0.0352483      0.0352483
## g                 0.09759 0.13044 0.0013044      0.0013044
##
## 2. Quantiles for each variable:
##
##                      2.5%       25%      50%     75%   97.5%
## mu                4.94617  5.333950  5.54115  5.7485  6.1348
## PctChildPoverty  -0.07707 -0.008576  0.02569  0.0605  0.1250
## PctFamilyPoverty -0.42588 -0.325116 -0.27252 -0.2217 -0.1210
## Enrolled_log     -3.73778 -3.096238 -2.76501 -2.4470 -1.8192
## TotalSchools_log  0.68104  1.510841  1.96169  2.4088  3.2751
## sig2             58.82250 62.897969 65.22202 67.6652 72.5734
## g                 0.02216  0.043234  0.06679  0.1082  0.3474
```

Result 5th A linear regression was performed to estimate the percentage of all enrolled students with belief exceptions with use of PctChildPoverty, PctFamilyPoverty, Enrolled , and TotalSchools as the four predictors.

Bi-variate exploratory data analysis noted that the variables were somewhat skewed with a hint of a non-linear relationship. As the distributions were highly skewed for Enrolled and TotalSchools, so the data were log transformed for analysis, which generally improved the skew and the linearity of the relationship. A linear regression found strong support for the relationship (F(4,691)=29.47, p-value<0.001, adjusted R2 = 0.1408). Among predictors, PctFamilyPoverty (b=-0.28061, t=-3.628, p<0.001), Enrolled_log (b=-2.84671, t=-5.803, p<0.001) and TotalSchools_log(b=2.02038, t=3.017, p<0.01) were significant. PctChildPoverty(b=0.02642,t=0.05199, p>0.05) is not significant because p value is greater than 0.05 and we failed to reject the null hypothesis.

A Bayesian regression also found overwhelming evidence in support of a model with significant predictors PctFamilyPoverty, Enrolled_log and TotalSchools_log. The BayesFactor analysis shows that Bayes Factor of 2.752423e+19:1 are very strong odds in the favor of alternative hypothesis. So we reject the null hypothesis which suggest that Intercept only model is better. The sampled coefficients had similar values, a mean of -0.27324 for PctFamilyPoverty with an 95% HDI of -0.42182 to -0.1232, a mean of -2.77446 for Enrolled_log with an 95% HDI of -3.72306 to -1.8306, and a mean of 1.96750 for TotalSchools_log with an 95% HDI of 0.67979 to 3.2556. Apart from this, we can see that, a mean of 0.02507 for PctChildPoverty with 95% HDI of -0.07373 to 0.1251 shows that HDI has 0, which tells us that PctChildPoverty is not a good predictor because there is chance that mean value is 0. This result is perfectly aligning with the traditional linear model analysis.

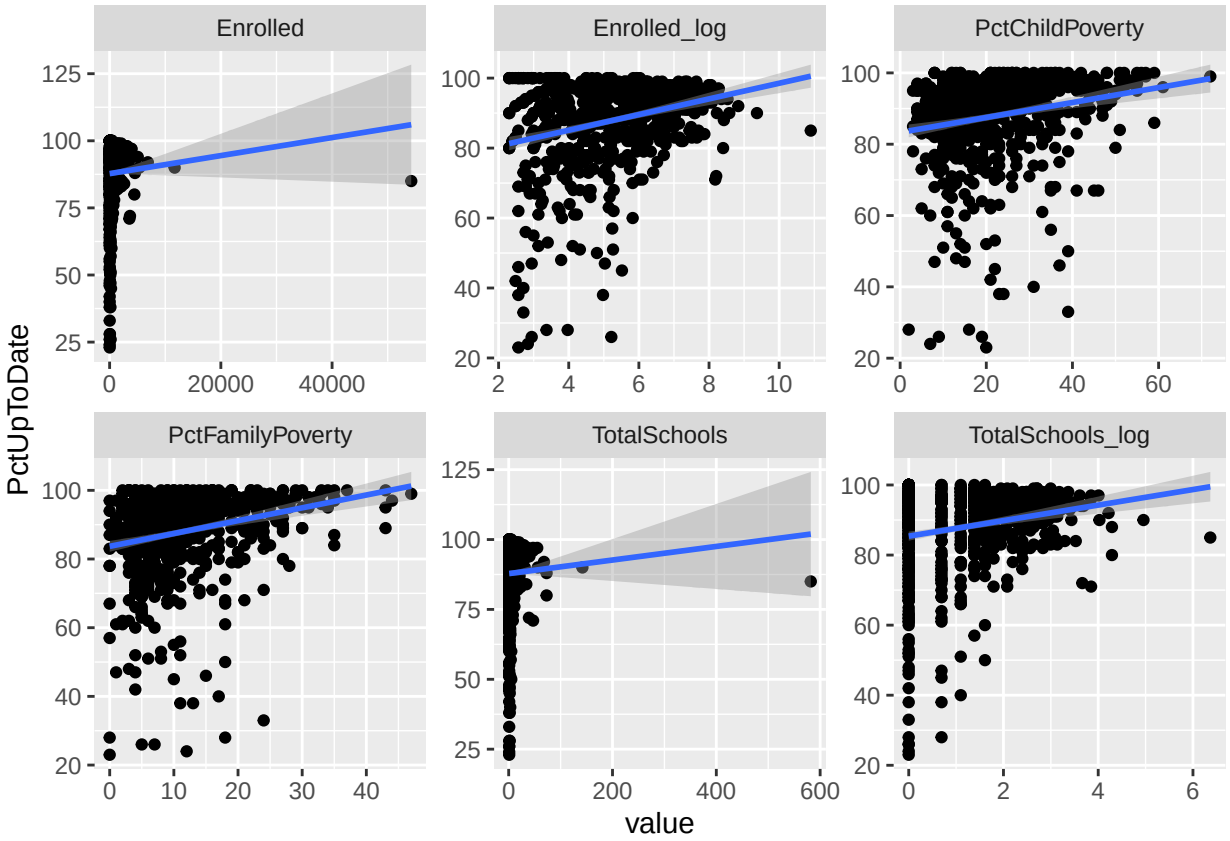
Overall, we can say that PctFamilyPoverty,Enrolled_log and TotalSchools_log provide an excellent estimate of the percentage of all enrolled students with belief exceptions.

6. *Which of the four predictor variables predicts the percentage of all enrolled students with completely up-to-date vaccines?*

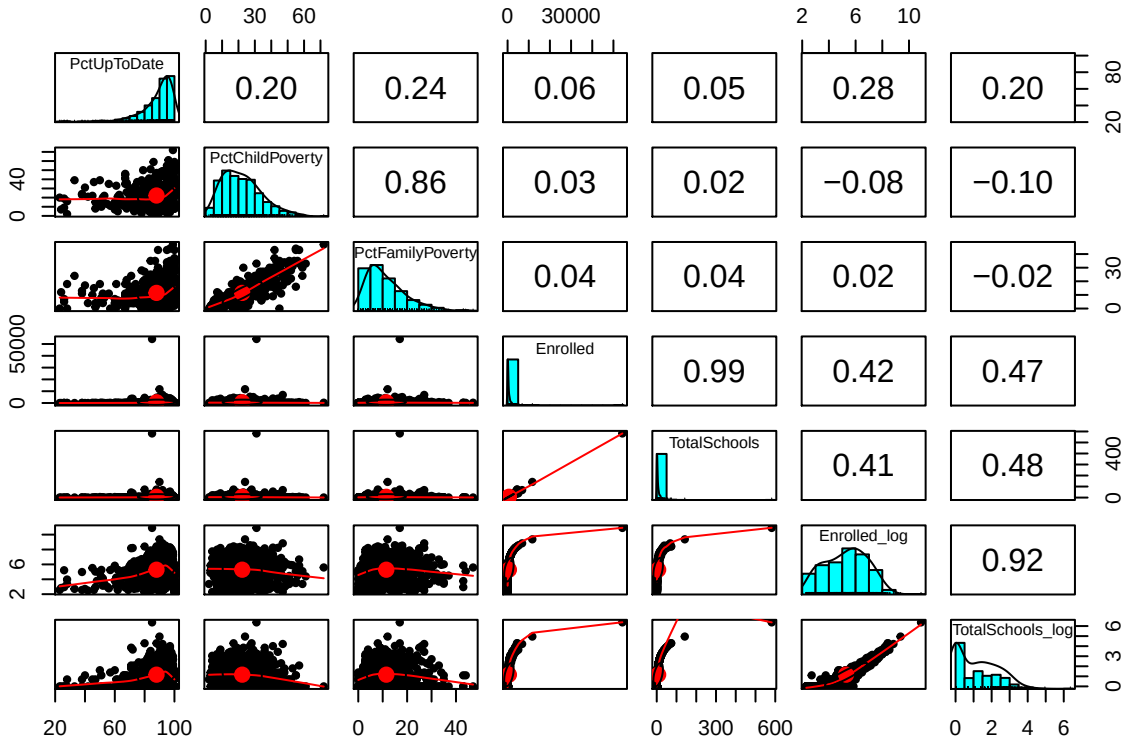
Creating new dataset to check the effect of PctChildPoverty, PctFamilyPoverty, Enrolled, and TotalSchools as the four predictors on PctUpToDate

```
uptodate <- subset(districts_new, select = c("PctUpToDate","PctChildPoverty",
                                             "PctFamilyPoverty","Enrolled","TotalSchools"))

# Applying log transformations on Enrolled and TotalSchools columns
uptodate$Enrolled_log <- log(uptodate$Enrolled)

uptodate$TotalSchools_log <- log(uptodate$TotalSchools)


# checking the scatter plots for each variables with respect to the PctBeliefExempt
uptodate %>% pivot_longer(-PctUpToDate, names_to="variable", values_to="value", values_drop_na = TRUE) %
             ggplot(aes(x=value, y=PctUpToDate)) + geom_point() +
                  geom_smooth(method = "lm") + facet_wrap( ~ variable, scales="free")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# checking the histograms, scatterplots and correlation between each variables in the plots
pairs.panels(uptodate)
```



The above plots shows that how log tranformations have helped Enrolled and TotalSchools variables to get a better linear relationship with PctUpToDate compared to when they were not log transformed, as well as in reducing the skewness in the right tail. We can see that accumulation of points across blue line is increased in the scatter plots of Enrolled_log and TotalSchools_log variables. In addition to this, in other scatter plots, we can see sufficient accumulation of points across blue line.

Also in above calculations, I have not applied log transformations to percentage columns in newly created uptodate dataset, because usually the percentage values varies from 0 to 100 and log transforming them can certainly hamper the dataset meaning. We can also verify from the outliers_plot that outlier values are increasing and varying drastically on log transformed data.

```
# checking the correlation with below correlation matrix
round(cor(uptodate),3)
```

```
##                 PctUpToDate PctChildPoverty PctFamilyPoverty Enrolled
## PctUpToDate           1.000           0.200            0.240    0.060
## PctChildPoverty       0.200           1.000            0.864    0.026
## PctFamilyPoverty      0.240           0.864            1.000    0.044
## Enrolled              0.060           0.026            0.044    1.000
## TotalSchools          0.047           0.021            0.038    0.994
## Enrolled_log          0.283          -0.080            0.018    0.416
## TotalSchools_log      0.204          -0.099           -0.020    0.474
##                 TotalSchools Enrolled_log TotalSchools_log
## PctUpToDate            0.047        0.283            0.204
## PctChildPoverty        0.021       -0.080           -0.099
## PctFamilyPoverty       0.038        0.018           -0.020
```

```
## Enrolled              0.994        0.416           0.474
## TotalSchools          1.000        0.406           0.480
## Enrolled_log          0.406        1.000           0.917
## TotalSchools_log      0.480        0.917           1.000
```

From the above results we can see that there is high positive correlation [0.864] between PctChildPoverty and PctFamilyPoverty. In addition to this there exists high positive correlation between Enrolled and TotalSchools [0.994]; and Enrolled_log and TotalSchools_log [0.917]. In case of PctUpToDate, there exists medium positive correlation with Enrolled_log [0.283] which shows r value, and tells us that if there is 1 unit increase in Enrolled_log then there is a likely possibility of PctUpToDate getting increased by 0.283

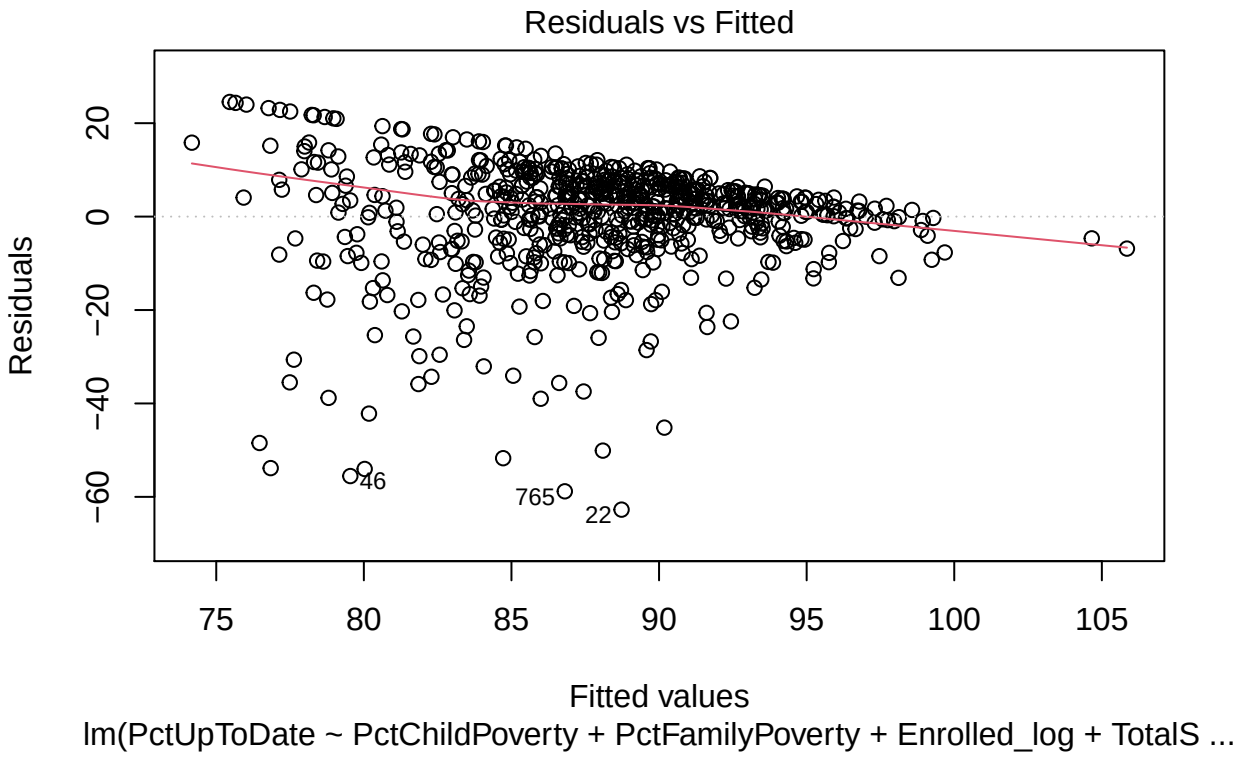
```
uptodate_lm_out <- lm(PctUpToDate~PctChildPoverty+PctFamilyPoverty+Enrolled_log+TotalSchools_log, data =

# checking the effects of multicolinearity in the model predictors
library(car)

vif(uptodate_lm_out)
```

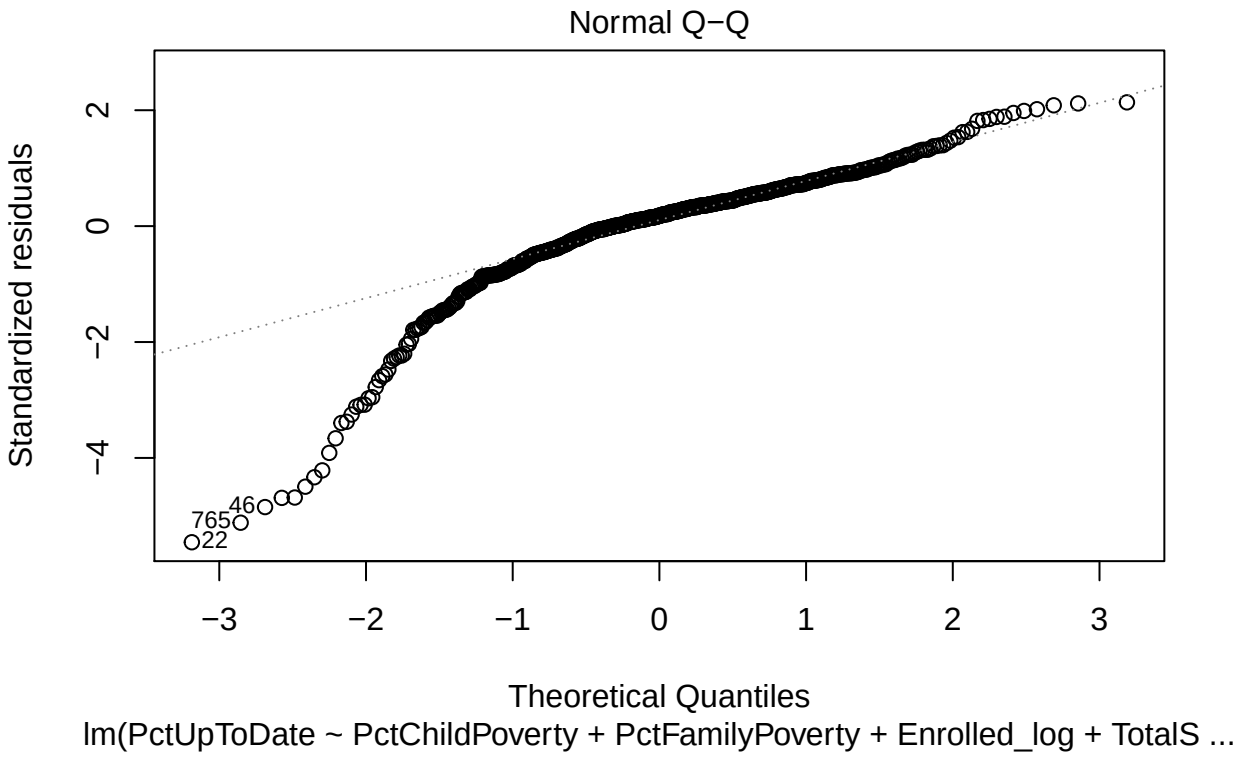
```
##  PctChildPoverty PctFamilyPoverty     Enrolled_log TotalSchools_log
##         4.109105         4.102388         6.407923         6.341736
```

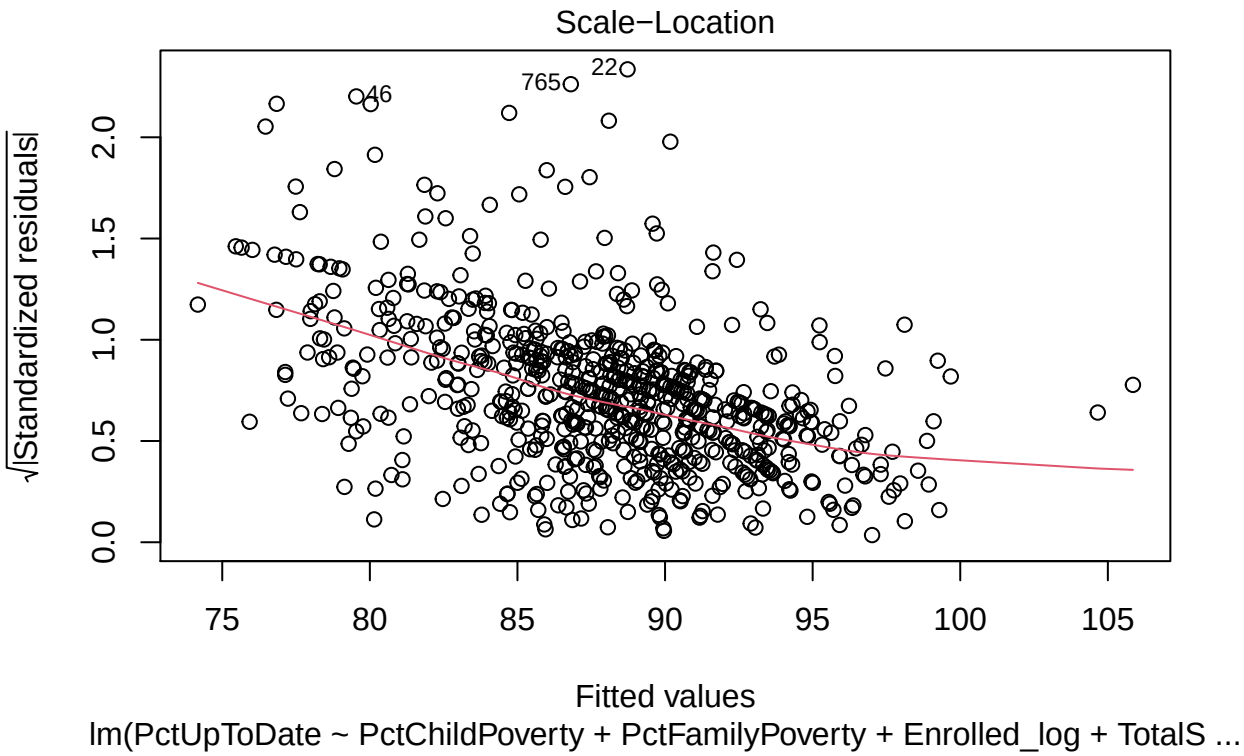
The multicolinearity check is passed as all the variance inflation factors for predictor values are less than 10, but we need to careful for variance inflation factors > 5 for Enrolled_log and TotalSchools_log.

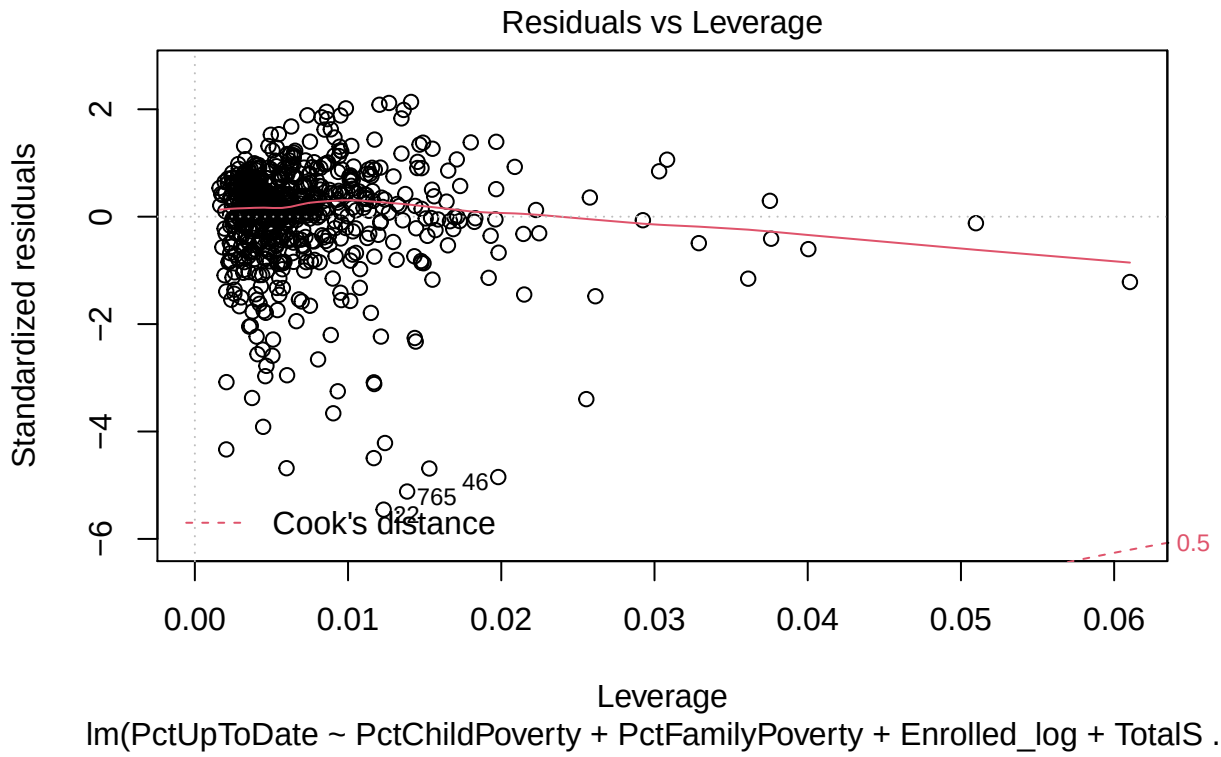
Analyze the linear model plots and linear model's residual histogram plot to check the linearity

```
# plotting model plots
plot(uptodate_lm_out)
```

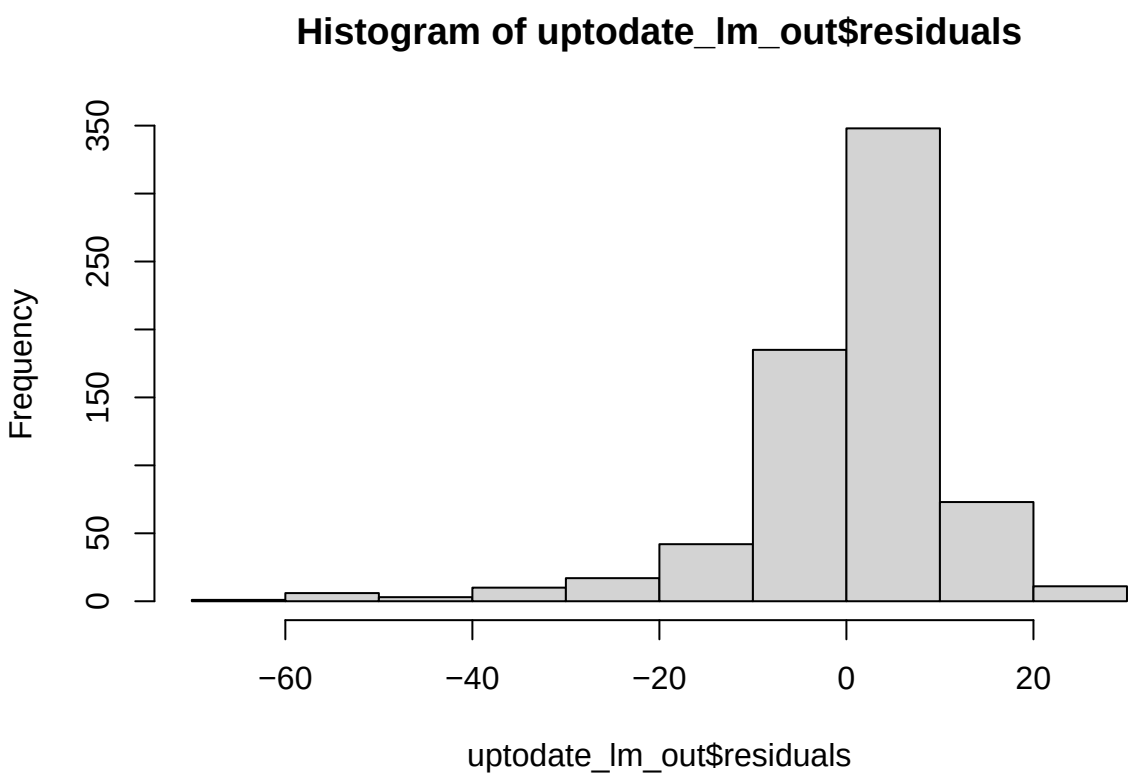

Residuals vs Fitted
Residuals
20
0
−20
−40
−60
46
765
22
75
80
85
90
95
100
105
Fitted values
lm(PctUpToDate ~ PctChildPoverty + PctFamilyPoverty + Enrolled_log + TotalS ...


Normal Q−Q
Standardized residuals
2
0
−2
−4
765
46
22
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(PctUpToDate ~ PctChildPoverty + PctFamilyPoverty + Enrolled_log + TotalS ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
46
765
22
75
80
85
90
95
100
105
Fitted values
lm(PctUpToDate ~ PctChildPoverty + PctFamilyPoverty + Enrolled_log + TotalS ...




```
# Plotting histograms of residuals
hist(uptodate_lm_out$residuals)
```




```
# checking mean and median for residuals
mean(uptodate_lm_out$residuals)
```

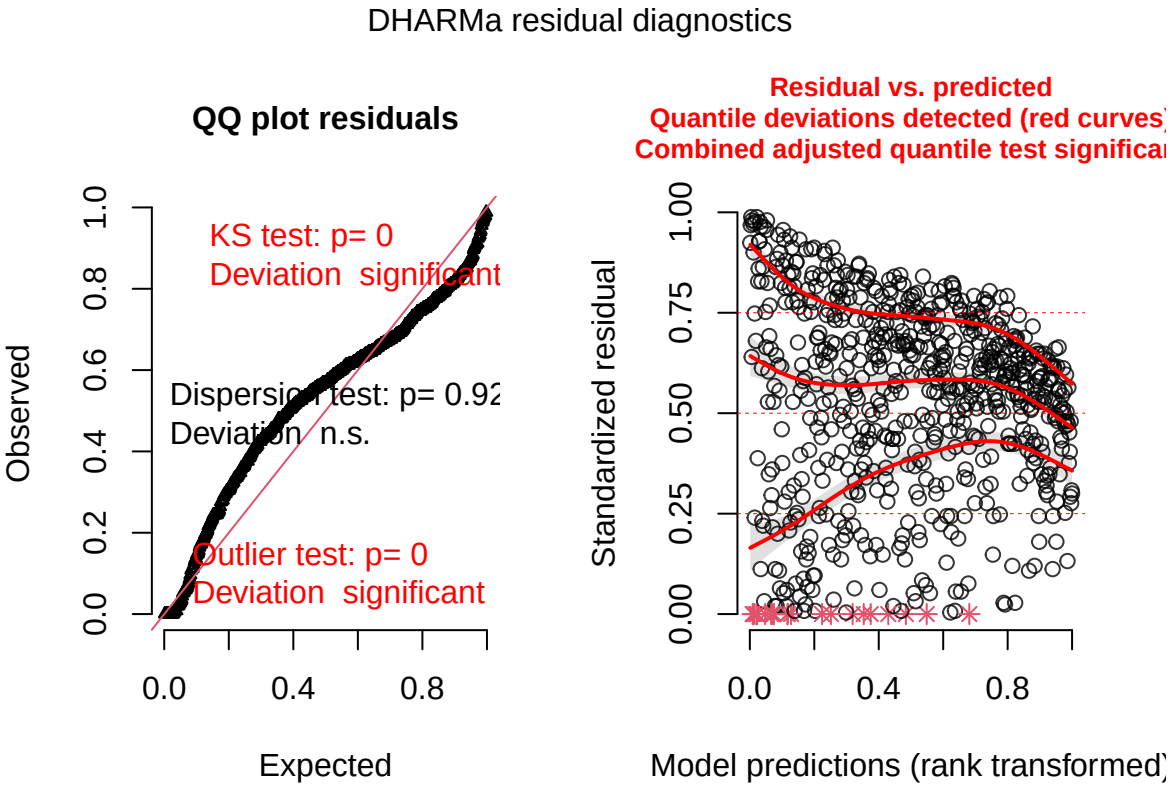
```
## [1] -4.516711e-16
```

```
median(uptodate_lm_out$residuals)
```

```
## [1] 2.072876
```

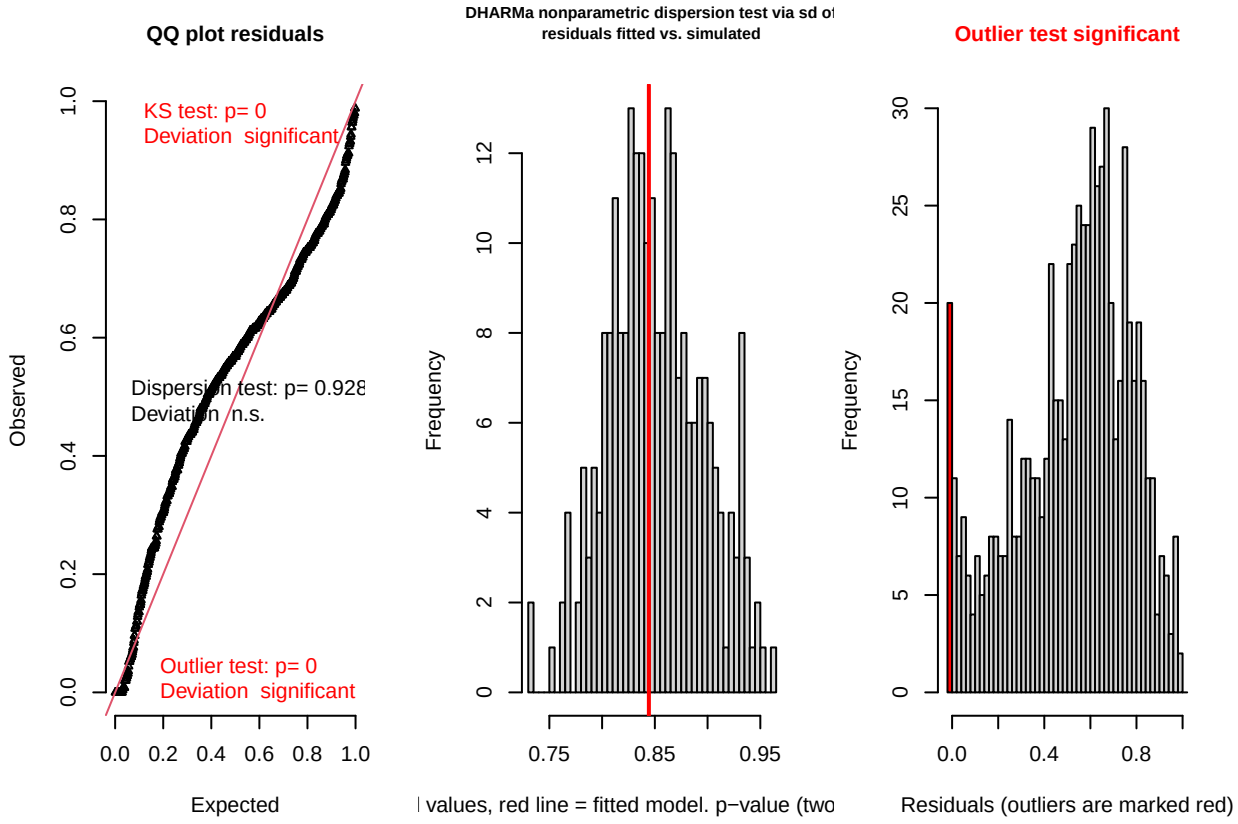
In above plots we can see that, there is skewness in the left tail of the residual histogram plot. This result is also reflected in the Q-Q plot of the model. The Q-Q plot is having a curve on the lower side. Also, it is visible that in all plots the accumulation of points is more along the red line and even though there is no outlier effect, it is not clearly having strong linear relationship with dependent variable.

```
library(DHARMa)

simulationOut <- simulateResiduals(fittedModel = uptodate_lm_out, n = 250)

plot(simulationOut)
```



The 'DHARMa' package uses a simulation-based approach to create readily interpretable scaled (quantile) residuals for fitted (generalized) linear mixed models. The resulting residuals are standardized to values between 0 and 1 and can be interpreted as intuitively as residuals from a linear regression. The package also provides a number of plot and test functions for typical model misspecification problems, such as over/underdispersion, zero-inflation, and residual spatial and temporal autocorrelation. In case of above plots, as Q-Q plot is not exactly along the red line, we can say that there is no strong liner relationship with dependent variable.

```
# The test are run as follows:
invisible(testResiduals(simulationOut))
```




```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.13377, p-value = 3.042e-11
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.99158, p-value = 0.928
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 20, observations = 696, p-value =
```

```
## 1.421e-06
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.01763894 0.04403216
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.02873563
```

The above results we can see that Q-Q plot has a small curve and Dispersion plot is normally distributed with some skewness. Also, DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated shows that it failed to reject the null hypothesis as p_value is greater than 0.05.

```
summary(uptodate_lm_out)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ PctChildPoverty + PctFamilyPoverty +
##     Enrolled_log + TotalSchools_log, data = uptodate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.730  -4.019   2.073   6.461  24.542
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      63.55849    2.90507  21.878  < 2e-16 ***
## PctChildPoverty   0.09397    0.07444   1.262  0.20725
## PctFamilyPoverty  0.22724    0.11076   2.052  0.04057 *
## Enrolled_log      4.44683    0.70243   6.331 4.39e-10 ***
## TotalSchools_log -3.25077    0.95903  -3.390  0.00074 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.57 on 691 degrees of freedom
## Multiple R-squared:  0.1507, Adjusted R-squared:  0.1458
## F-statistic: 30.66 on 4 and 691 DF,  p-value: < 2.2e-16
```

In the above experiment, Multiple R-squared = 0.1507 and Adjusted R-squared = 0.1458 represents the proportion of about 15% variation in PctUpToDate (about its mean) explained by the multiple linear regression model with predictors in the model

Calculating beta weights below to verify how the standardized variations have been changed for all predictors

```
# checking beta weights to see standardized deviation
#install.packages("lm.beta")
library(lm.beta)
summary(lm.beta(uptodate_lm_out))
```

```
##
## Call:
## lm(formula = PctUpToDate ~ PctChildPoverty + PctFamilyPoverty +
##     Enrolled_log + TotalSchools_log, data = uptodate)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.730  -4.019   2.073   6.461  24.542
##
## Coefficients:
##                  Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)      63.55849      0.00000    2.90507  21.878  < 2e-16 ***
## PctChildPoverty   0.09397      0.08971    0.07444   1.262  0.20725
## PctFamilyPoverty  0.22724      0.14569    0.11076   2.052  0.04057 *
## Enrolled_log      4.44683      0.56182    0.70243   6.331 4.39e-10 ***
## TotalSchools_log -3.25077     -0.29926    0.95903  -3.390  0.00074 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.57 on 691 degrees of freedom
## Multiple R-squared:  0.1507, Adjusted R-squared:  0.1458
## F-statistic: 30.66 on 4 and 691 DF,  p-value: < 2.2e-16
```

As we can see in the above experiment, Standardized deviations have been reduced only for Enrolled_log and TotalSchools_Log to some extent from std.error.

Now conducting a Bayesian linear regression analysis, using the facilities in the BayesFactor package.

```
library(BayesFactor)

# Calculating Bayes Factor
uptodate_lmbf_out <- lmBF(PctUpToDate~PctChildPoverty+PctFamilyPoverty+Enrolled_log+TotalSchools_log, da

uptodate_lmbf_out
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFamilyPoverty + Enrolled_log + TotalSchools_log : 1.973708e+20 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
# Running MCMC test on uptodate_lmbf_out using posterior distributions
uptodate_lmbf_out1 <- lmBF(PctUpToDate~PctChildPoverty+PctFamilyPoverty+Enrolled_log+TotalSchools_log, d

summary(uptodate_lmbf_out1)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
```

```
##                       Mean      SD  Naive SE Time-series SE
## mu                 87.98258 0.43735 0.0043735      0.0042094
## PctChildPoverty     0.09123 0.07369 0.0007369      0.0007369
## PctFamilyPoverty    0.22326 0.10968 0.0010968      0.0010968
## Enrolled_log        4.33701 0.69449 0.0069449      0.0069449
## TotalSchools_log   -3.16945 0.95375 0.0095375      0.0105350
## sig2              134.08773 7.29164 0.0729164      0.0729164
## g                   0.09943 0.12585 0.0012585      0.0012585
##
## 2. Quantiles for each variable:
##
##                       2.5%       25%       50%      75%    97.5%
## mu                 87.12997  87.68522  87.98423  88.2816  88.8225
## PctChildPoverty   -0.05141   0.04170   0.09121   0.1421   0.2332
## PctFamilyPoverty   0.01599   0.14867   0.22159   0.2976   0.4404
## Enrolled_log       2.99840   3.86499   4.33747   4.8056   5.6962
## TotalSchools_log  -5.04916  -3.82028  -3.16594  -2.5338  -1.2868
## sig2             120.51691 129.11334 133.85668 138.7624 149.0447
## g                   0.02228   0.04441   0.06848   0.1122   0.3617
```

Result 6th A linear regression was performed to estimate the percentage of all enrolled students with completely up-to-date vaccines with use of PctChildPoverty, PctFamilyPoverty, Enrolled, and TotalSchools as the four predictors.

Bi-variate exploratory data analysis noted that the variables were somewhat skewed with a hint of a non-linear relationship. As the distributions were highly skewed for Enrolled and TotalSchools, so the data were log transformed for analysis, which generally improved the skew and the linearity of the relationship. A linear regression found strong support for the relationship (F(4,691)=30.66, p-value<0.001, adjusted R2 = 0.1458). Among predictors, PctFamilyPoverty (b=0.22724, t=2.052, p<0.05), Enrolled_log (b=4.44683, t=6.331, p<0.001) and TotalSchools_log(b=-3.25077, t=-3.390, p<0.001) were significant. PctChildPoverty(b=0.09397,t=1.262, p>0.05) is not significant because p value is greater than 0.05 and we failed to reject the null hypothesis.

A Bayesian regression also found overwhelming evidence in support of a model with significant predictors PctFamilyPoverty, Enrolled_log and TotalSchools_log. The BayesFactor analysis shows that Bayes Factor of 1.973708e+20:1 are very strong odds in the favor of alternative hypothesis. So we reject the null hypothesis which suggest that Intercept only model is better. The sampled coefficients had similar values, a mean of 0.2216 for PctFamilyPoverty with an 95% HDI of 0.004752 to 0.4339, a mean of 4.3326 for Enrolled_log with an 95% HDI of 3.007683 to 5.6782, and a mean of -3.25077 for TotalSchools_log with an 95% HDI of -4.998471 to -1.3202. Apart from this, we can see that, a mean of 0.0924 for PctChildPoverty with 95% HDI of -0.051296 to 0.2381 shows that HDI has 0, which tells us that PctChildPoverty is not a good predictor because there is chance that mean value is 0. This result is perfectly aligning with the traditional linear model analysis.

Overall, we can say that PctFamilyPoverty,Enrolled_log and TotalSchools_log provide an excellent estimate of the percentage of all enrolled students with with completely up-to-date vaccines.

7. *Using any set of predictors that you want to use, what's the best R-squared you can achieve in predicting the percentage of all enrolled students with completely up-to-date vaccines while still having an acceptable regression?*

Creating new dataset to check the effect of best predictors from to predict on PctUpToDate

```r
mydistricts_new <- mydistricts


# Applying log transformations on Enrolled and TotalSchools columns
mydistricts_new$Enrolled_log <- log(mydistricts_new$Enrolled)

mydistricts_new$TotalSchools_log <- log(mydistricts_new$TotalSchools)

# changing the DistrictComplete from logical to factor datatype

mydistricts_new$DistrictComplete_new <- as.factor(mydistricts_new$DistrictComplete)

# verifying the values using table command on old and new columns

table(mydistricts_new$DistrictComplete) # original logical data type column
```

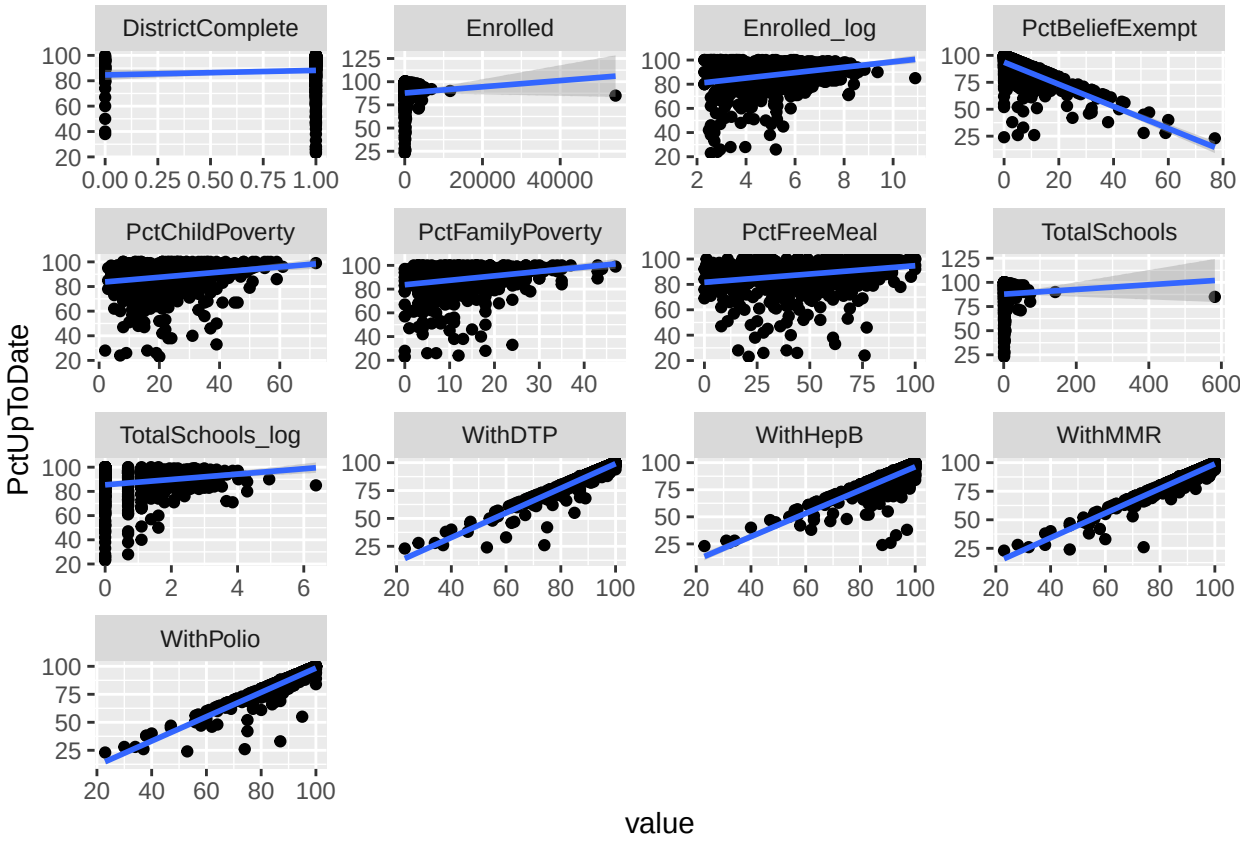
```
##
## FALSE  TRUE
##    38   658
```

```r
table(mydistricts_new$DistrictComplete_new) # new changed factor data type column
```

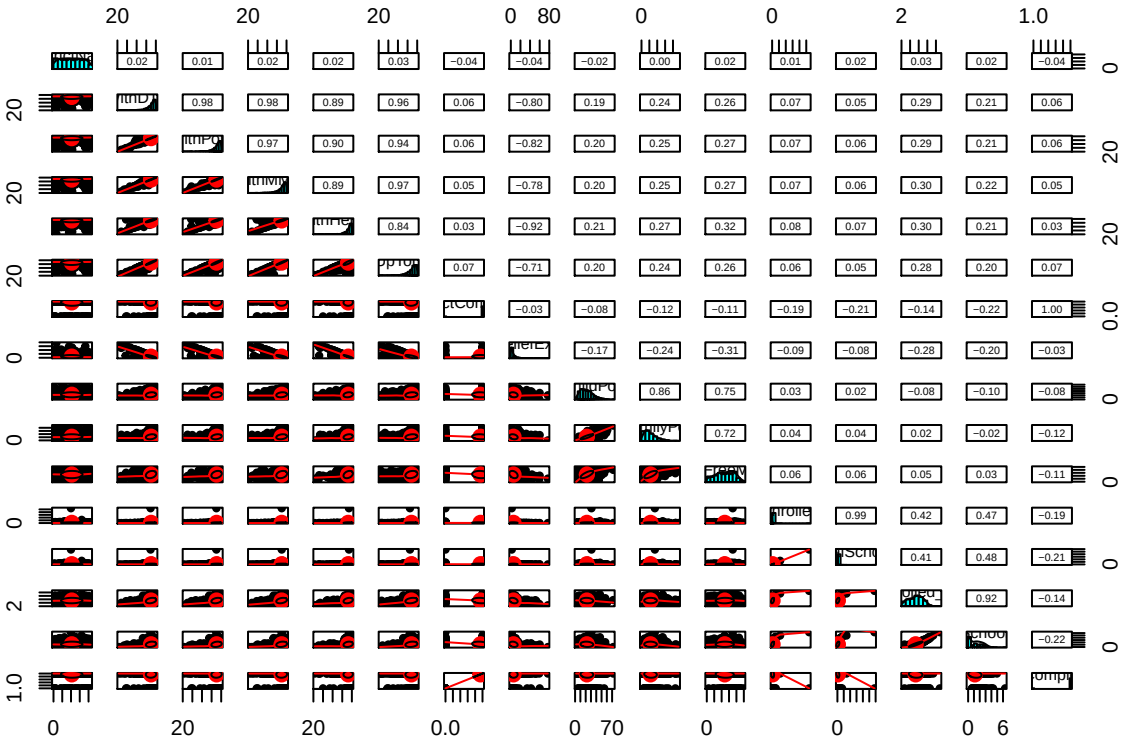
```
##
## FALSE  TRUE
##    38   658
```

```r
# checking the scatter plots for each variables with respect to the PctBeliefExempt
mydistricts_new %>% pivot_longer(-c("DistrictName","PctUpToDate","DistrictComplete_new"), names_to="vari
             ggplot(aes(x=value, y=PctUpToDate)) + geom_point() +
                  geom_smooth(method = "lm") + facet_wrap( ~ variable, scales="free")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# checking the histograms, scatterplots and correlation between each variables in the plots
pairs.panels(mydistricts_new)
```



The above plots shows that how log tranformations have helped Enrolled and TotalSchools variables to get a better linear relationship with PctUpToDate compared to when they were not log transformed, as well as in reducing the skewness in the right tail. Also, we can see that accumulation of points across blue line is increased in the scatter plots of Enrolled_log and TotalSchools_log variables. In addition to this, in other scatter plots, we can see sufficient accumulation of points across blue line.

Also in above calculations, I have not applied log transformations to percentage columns in newly created mydistricts_new dataset, because usually the percentage values varies from 0 to 100 and log transforming them can certainly hamper the dataset meaning. We can also verify from the outliers_plot that outlier values are increasing and varying drastically on log transformed data.

```
# checking the correlation with below correlation matrix


a <- subset(districts_new, select = -c(DistrictName, DistrictComplete))
round(cor(a),2)
```

```
##                  WithDTP WithPolio WithMMR WithHepB PctUpToDate PctBeliefExempt
## WithDTP             1.00      0.98    0.98     0.89        0.96           -0.80
## WithPolio           0.98      1.00    0.97     0.90        0.94           -0.82
## WithMMR             0.98      0.97    1.00     0.89        0.97           -0.78
## WithHepB            0.89      0.90    0.89     1.00        0.84           -0.92
## PctUpToDate         0.96      0.94    0.97     0.84        1.00           -0.71
## PctBeliefExempt    -0.80     -0.82   -0.78    -0.92       -0.71            1.00
## PctMedicalExempt      NA        NA      NA       NA          NA              NA
## PctChildPoverty     0.19      0.20    0.20     0.21        0.20           -0.17
## PctFamilyPoverty    0.24      0.25    0.25     0.27        0.24           -0.24
```

```
## PctFreeMeal         0.26      0.27    0.27     0.32        0.26          -0.31
## Enrolled            0.07      0.07    0.07     0.08        0.06          -0.09
## TotalSchools        0.05      0.06    0.06     0.07        0.05          -0.08
##                  PctMedicalExempt PctChildPoverty PctFamilyPoverty PctFreeMeal
## WithDTP                        NA            0.19             0.24        0.26
## WithPolio                      NA            0.20             0.25        0.27
## WithMMR                        NA            0.20             0.25        0.27
## WithHepB                       NA            0.21             0.27        0.32
## PctUpToDate                    NA            0.20             0.24        0.26
## PctBeliefExempt                NA           -0.17            -0.24       -0.31
## PctMedicalExempt                1              NA               NA          NA
## PctChildPoverty                NA            1.00             0.86        0.75
## PctFamilyPoverty               NA            0.86             1.00        0.72
## PctFreeMeal                    NA            0.75             0.72        1.00
## Enrolled                       NA            0.03             0.04        0.06
## TotalSchools                   NA            0.02             0.04        0.06
##                  Enrolled TotalSchools
## WithDTP              0.07         0.05
## WithPolio            0.07         0.06
## WithMMR              0.07         0.06
## WithHepB             0.08         0.07
## PctUpToDate          0.06         0.05
## PctBeliefExempt     -0.09        -0.08
## PctMedicalExempt       NA           NA
## PctChildPoverty      0.03         0.02
## PctFamilyPoverty     0.04         0.04
## PctFreeMeal          0.06         0.06
## Enrolled             1.00         0.99
## TotalSchools         0.99         1.00
```

From the above results we can see that there is high positive correlation [0.864] between PctChildPoverty and PctFamilyPoverty. In addition to this there exists high positive correlation between Enrolled and TotalSchools [0.994]; and Enrolled_log and TotalSchools_log [0.917]. In case of PctUpToDate, there exists medium positive correlation with Enrolled_log [0.283] which shows r value, and tells us that if there is 1 unit increase in Enrolled_log then there is a likely possibility of PctUpToDate getting increased by 0.283. Because of this we are deciding to select one predictor variable - WithMMR as it is highly correlated with PctUpToDate [0.967] and it is highly correlated with other vaccine rate variables such as WithDTP [0.980], WithPolio [0.967], WithHepB [0.889]. As mentioned earlier about high correlation between variables such as PctChildPoverty and PctFamilyPoverty, PctFreeMeal and PctFamilyPoverty, and Enrolled_log and TotalSchools_log; we will select only one variable among these considering the high colinearity.

```
# As we have high correlation among WithDTP, WithPolio, WithHepB and WithMMR we will select only one va

# We came across two different sets of models as shown below which were showing similar correlation as

# creating first linear model as below with WithMMR, PctBeliefExempt, Enrolled_log, PctFamilyPoverty an
updt_lm_out2 <- lm(PctUpToDate~WithMMR+PctBeliefExempt+Enrolled_log+PctFamilyPoverty+DistrictComplete_n

# checking the effects of multicolinearity in the model predictors
library(car)

# using vif function on updt_lm_out2
vif(updt_lm_out2)
```

```
##              WithMMR      PctBeliefExempt         Enrolled_log
##             2.691576             2.630762             1.144620
##     PctFamilyPoverty DistrictComplete_new
##             1.101266             1.054129
```

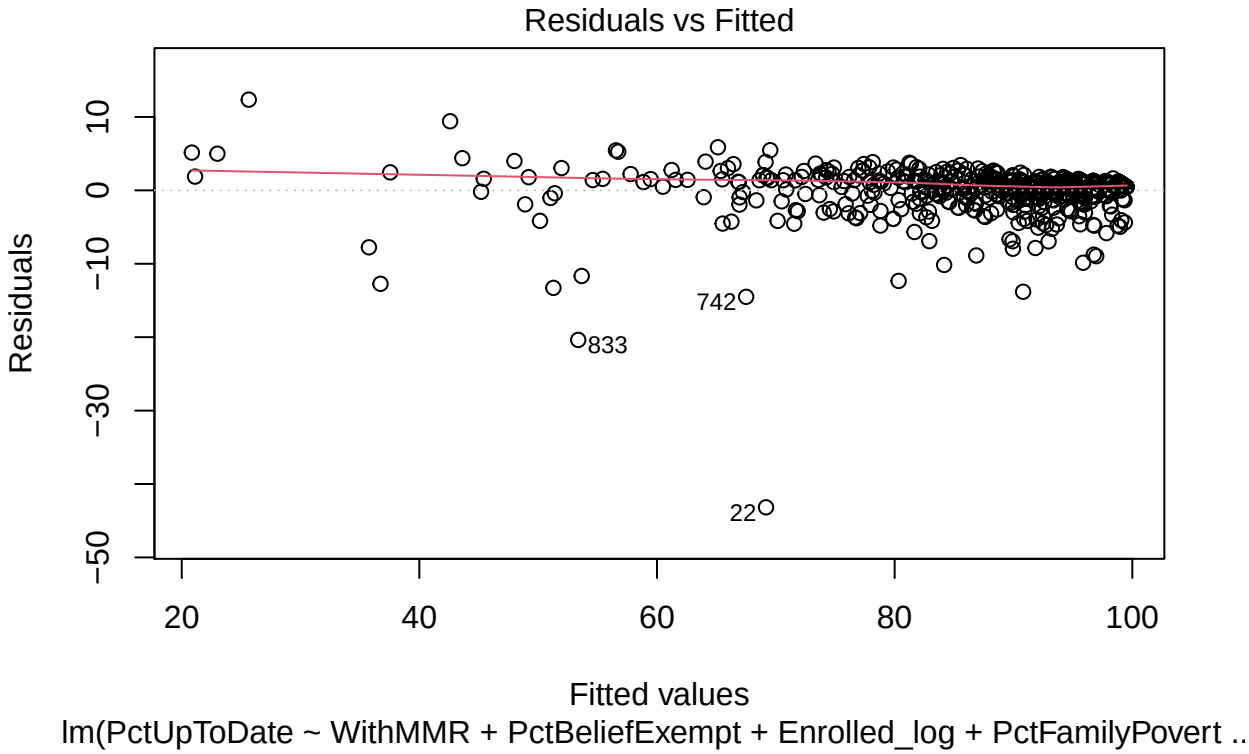
```
# creating second linear model as below with WithMMR, PctBeliefExempt, Enrolled_log, PctFreeMeal and Di
updt_lm_out3 <- lm(PctUpToDate~WithMMR+PctBeliefExempt+Enrolled_log+PctFreeMeal+DistrictComplete_new, da


# using vif function on updt_lm_out3
vif(updt_lm_out3)
```

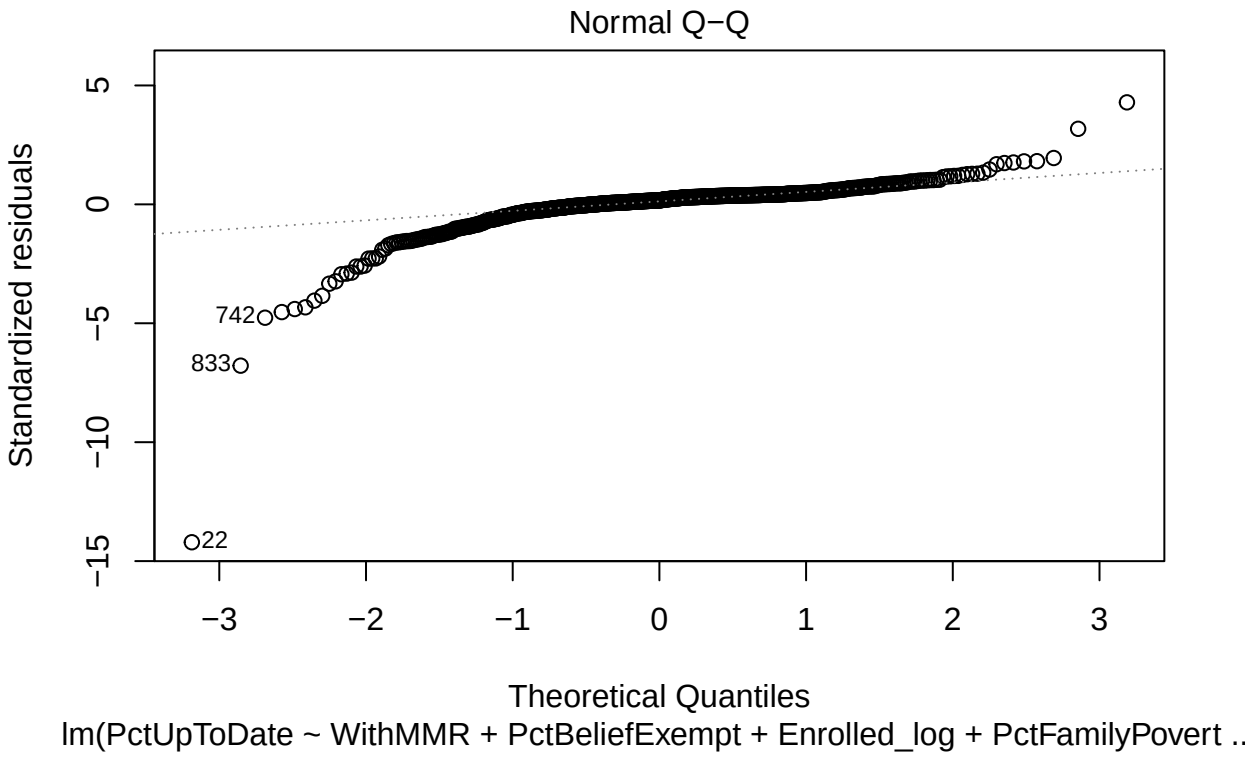
```
##              WithMMR      PctBeliefExempt         Enrolled_log
##             2.667730             2.685790             1.140729
##          PctFreeMeal DistrictComplete_new
##             1.129564             1.048599
```

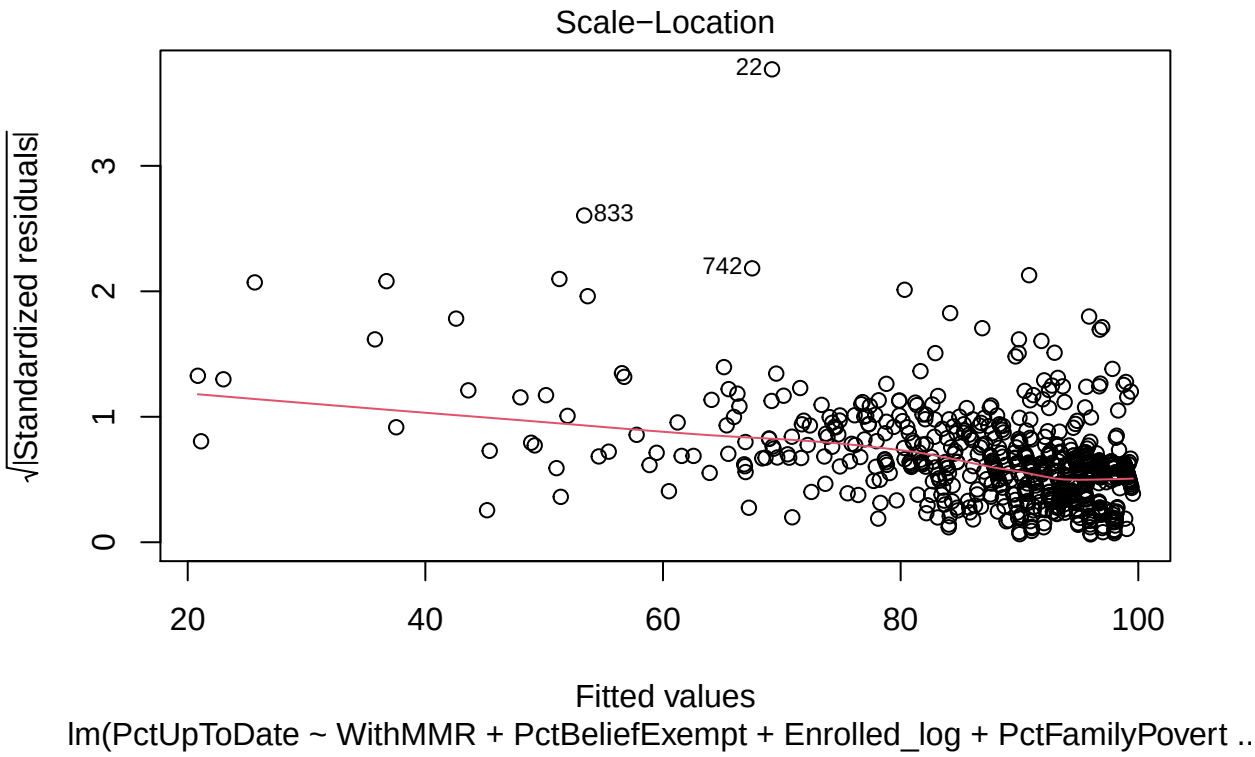
The multicolinearity check is passed as all the variance inflation factors for predictor values are less than 10.

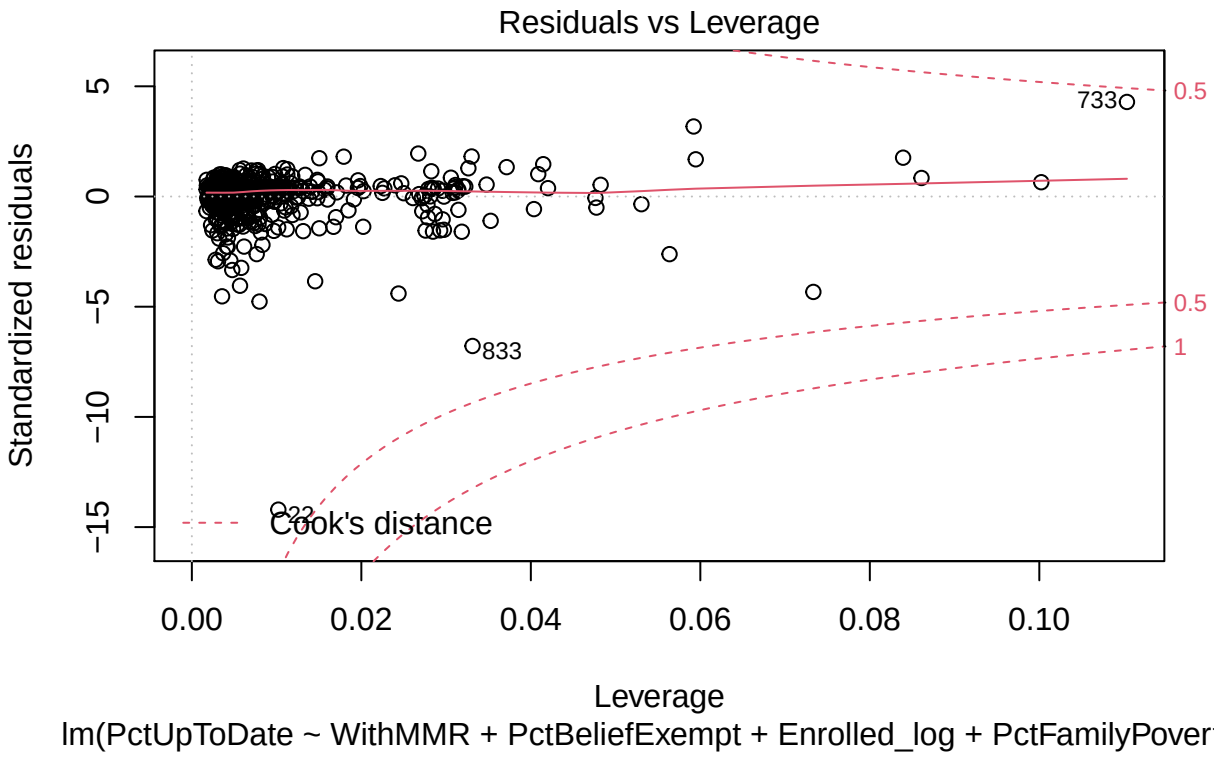
Analyze the linear model plots and linear model's residual histogram plot to check the linearity.

```
# plotting model plots for updt_lm_out2
plot(updt_lm_out2)
```

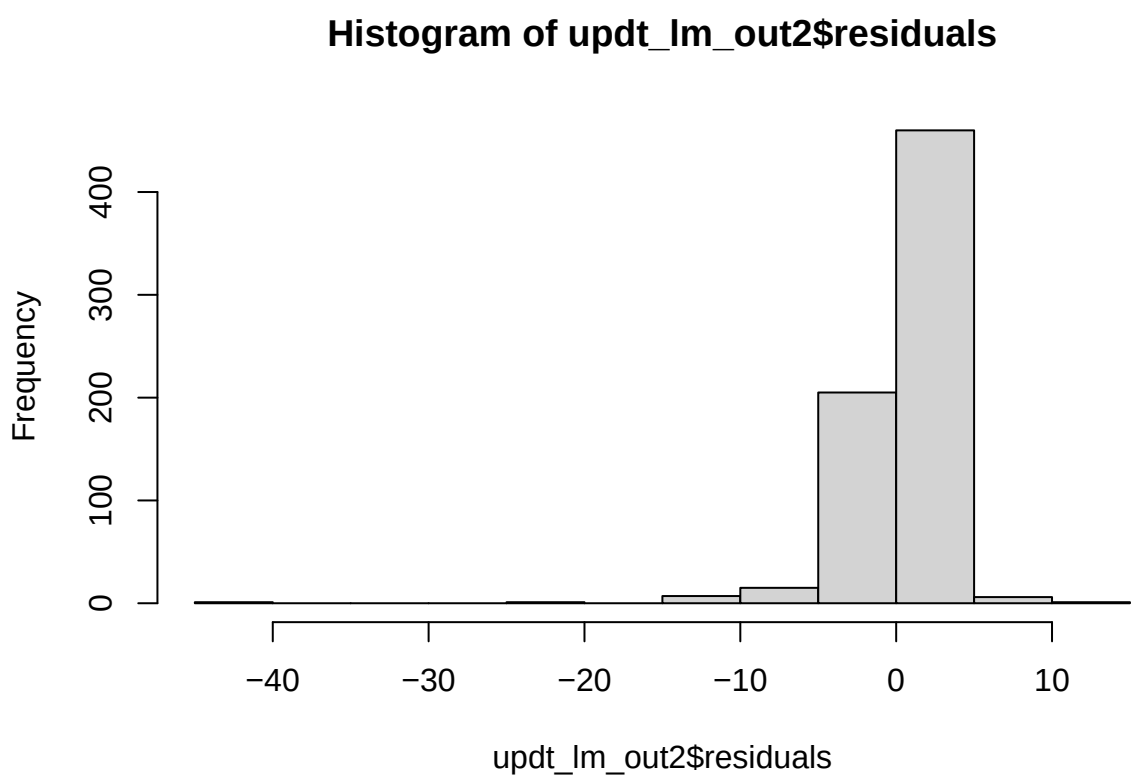



Normal Q−Q
Standardized residuals
5
0
−5
−10
−15
742
833
22
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(PctUpToDate ~ WithMMR + PctBeliefExempt + Enrolled_log + PctFamilyPovert ..


Scale−Location
22
833
742
√|Standardized residuals|
0
1
2
3
20
40
60
80
100
Fitted values
lm(PctUpToDate ~ WithMMR + PctBeliefExempt + Enrolled_log + PctFamilyPovert ..




```
# Plotting histograms of residuals
hist(updt_lm_out2$residuals)
```



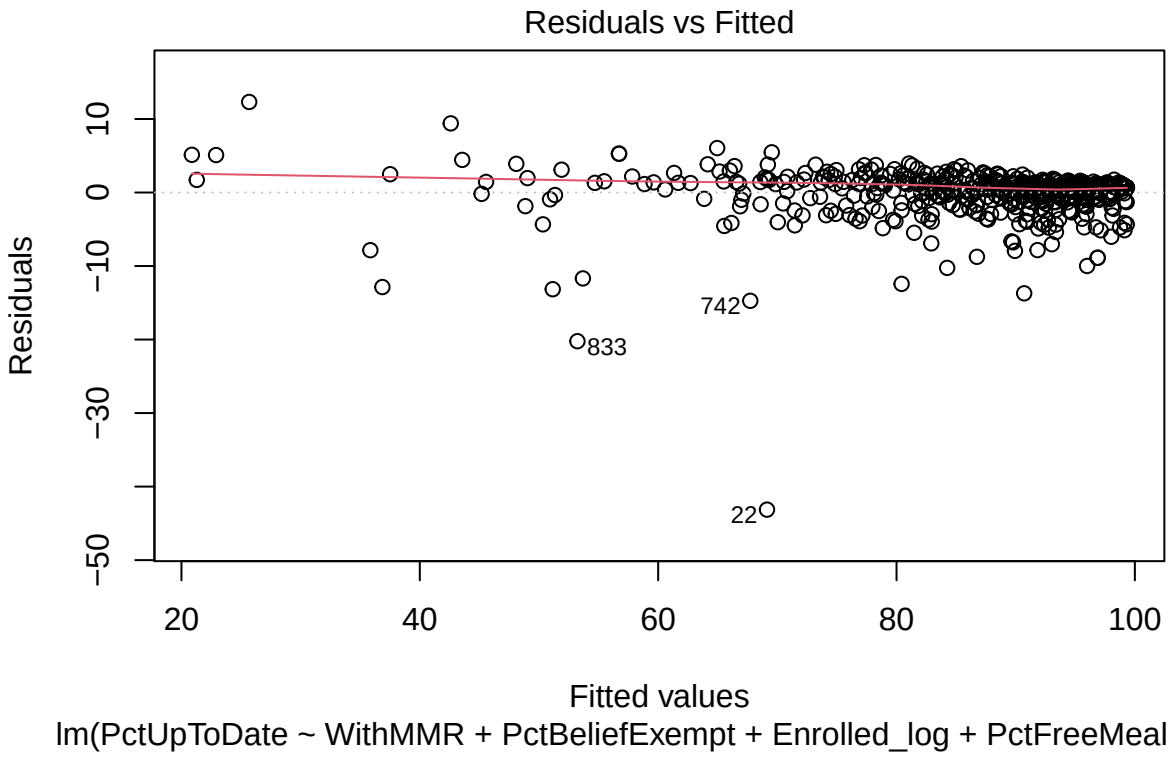

```
# checking mean and median for residuals
mean(updt_lm_out2$residuals)
```

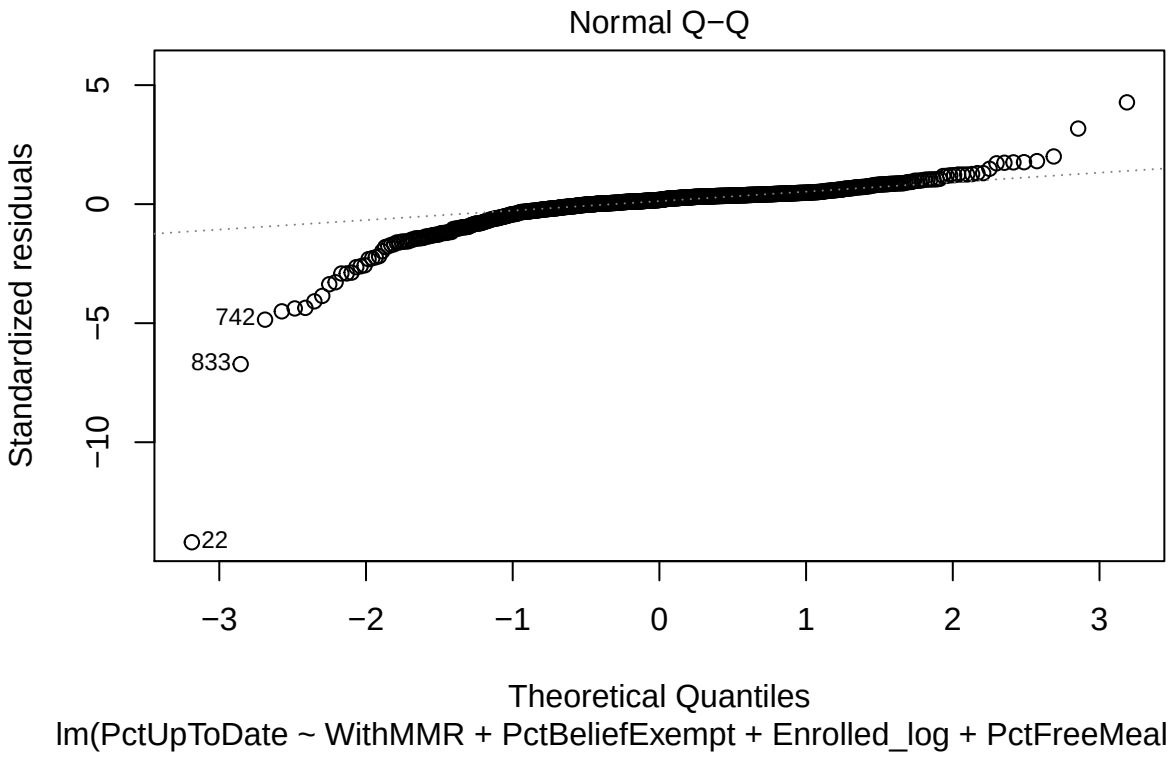
```
## [1] 3.997513e-17
```

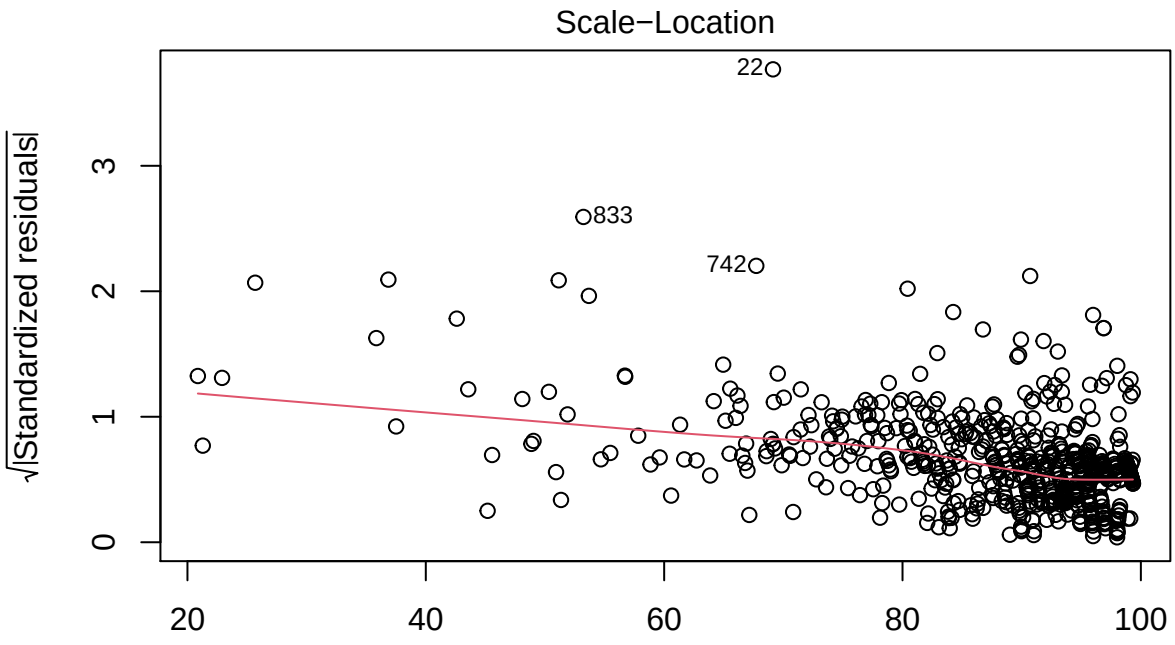
```
median(updt_lm_out2$residuals)
```

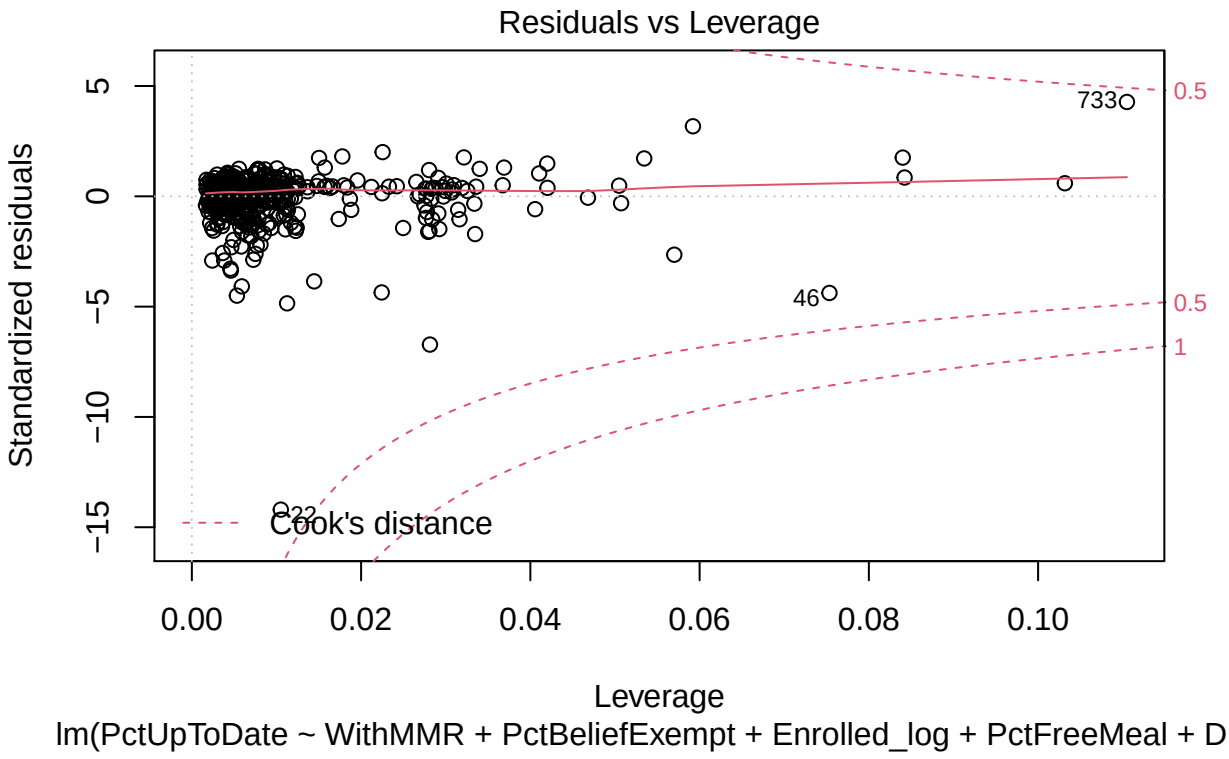
```
## [1] 0.5059719
```

```
#-------------------------------------------------

# plotting model plots for updt_lm_out3
plot(updt_lm_out3)
```

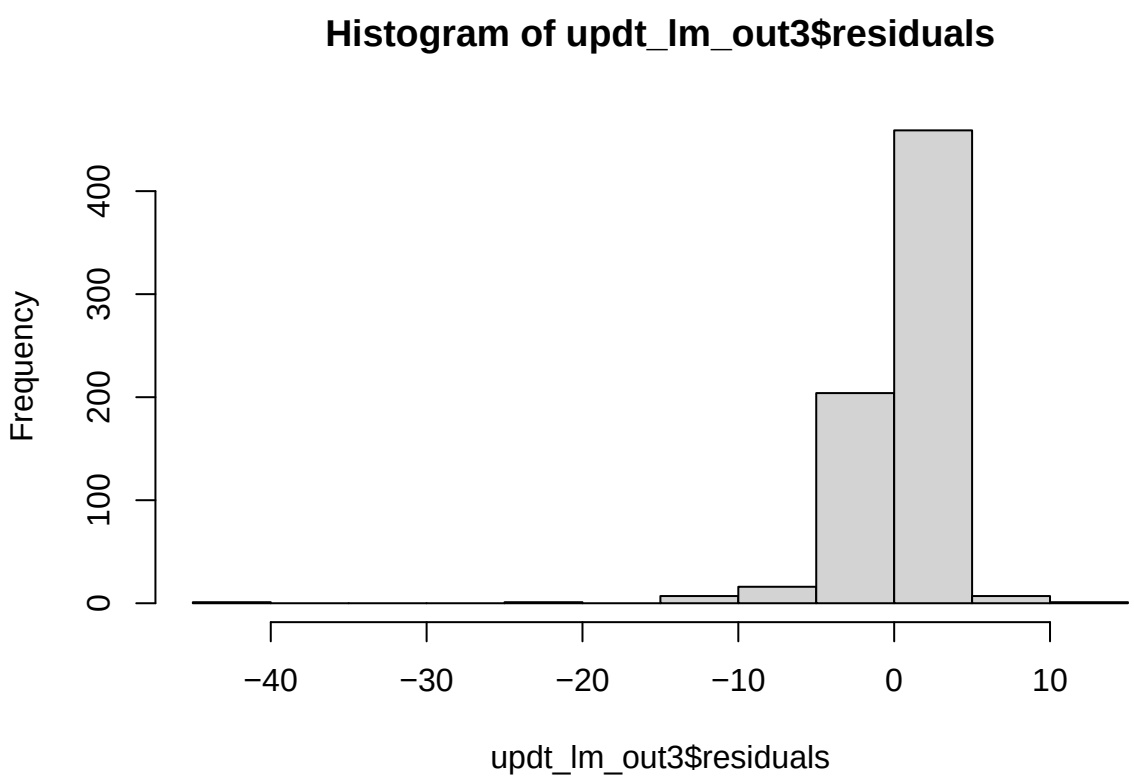

Residuals vs Fitted
Residuals
10
0
−10
−30
−50
742
833
22
20
40
60
80
100
Fitted values
lm(PctUpToDate ~ WithMMR + PctBeliefExempt + Enrolled_log + PctFreeMeal + D .


Normal Q−Q
Standardized residuals
5
0
−5
−10
742
833
22
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(PctUpToDate ~ WithMMR + PctBeliefExempt + Enrolled_log + PctFreeMeal + D .


Scale−Location
22
833
742
√|Standardized residuals|
0
1
2
3
20
40
60
80
100
Fitted values
lm(PctUpToDate ~ WithMMR + PctBeliefExempt + Enrolled_log + PctFreeMeal + D .




```
# Plotting histograms of residuals
hist(updt_lm_out3$residuals)
```




```
# checking mean and median for residuals
mean(updt_lm_out3$residuals)
```

```
## [1] 5.261734e-17
```

```
median(updt_lm_out3$residuals)
```

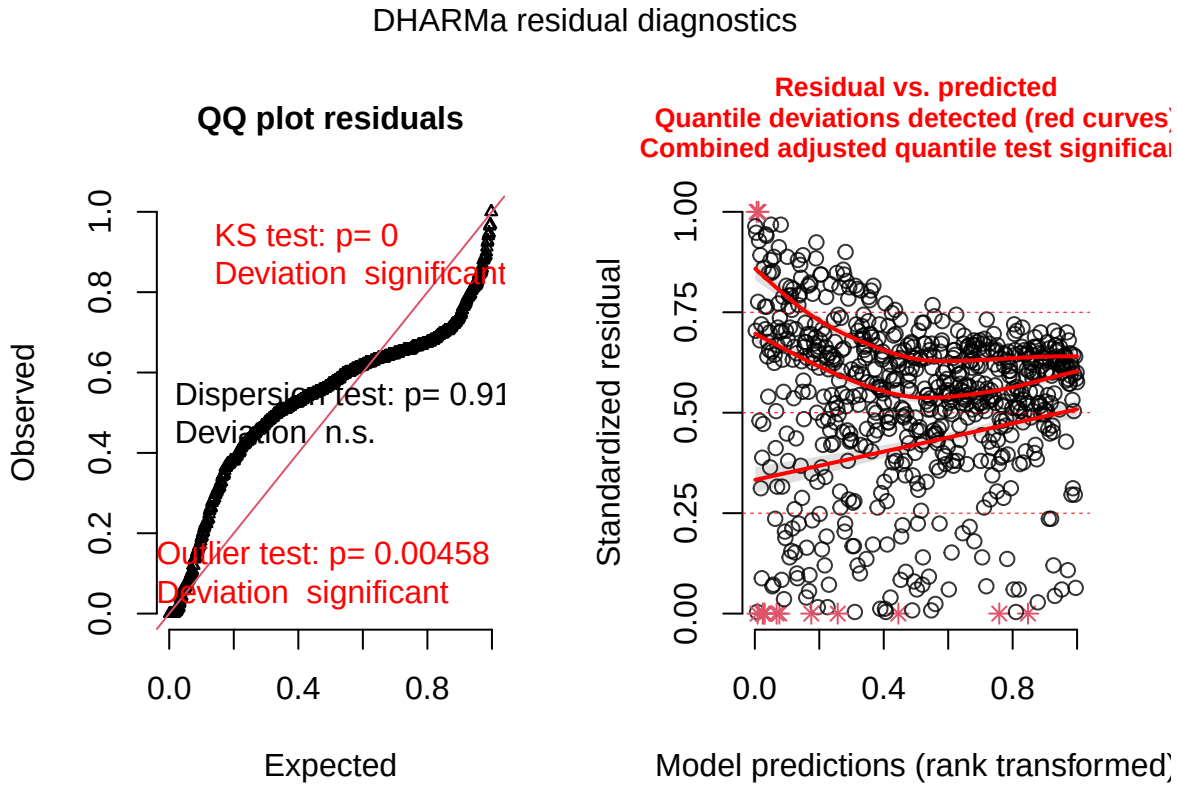
```
## [1] 0.5319157
```

In above plots we can see that, there is skewness in the left tail of the residual histogram plots of both models. This result is also reflected in the Q-Q plot of the model. The Q-Q plot is having a curve on the lower side. Also, it is visible that in all plots the accumulation of points is more along the red line and even though there is no outlier effect, it is not clearly having strong linear relationship with dependent variable.

```
library(DHARMa)

# plotting residual plots
simulationOut2 <- simulateResiduals(fittedModel = updt_lm_out2, n = 250)

plot(simulationOut2)
```

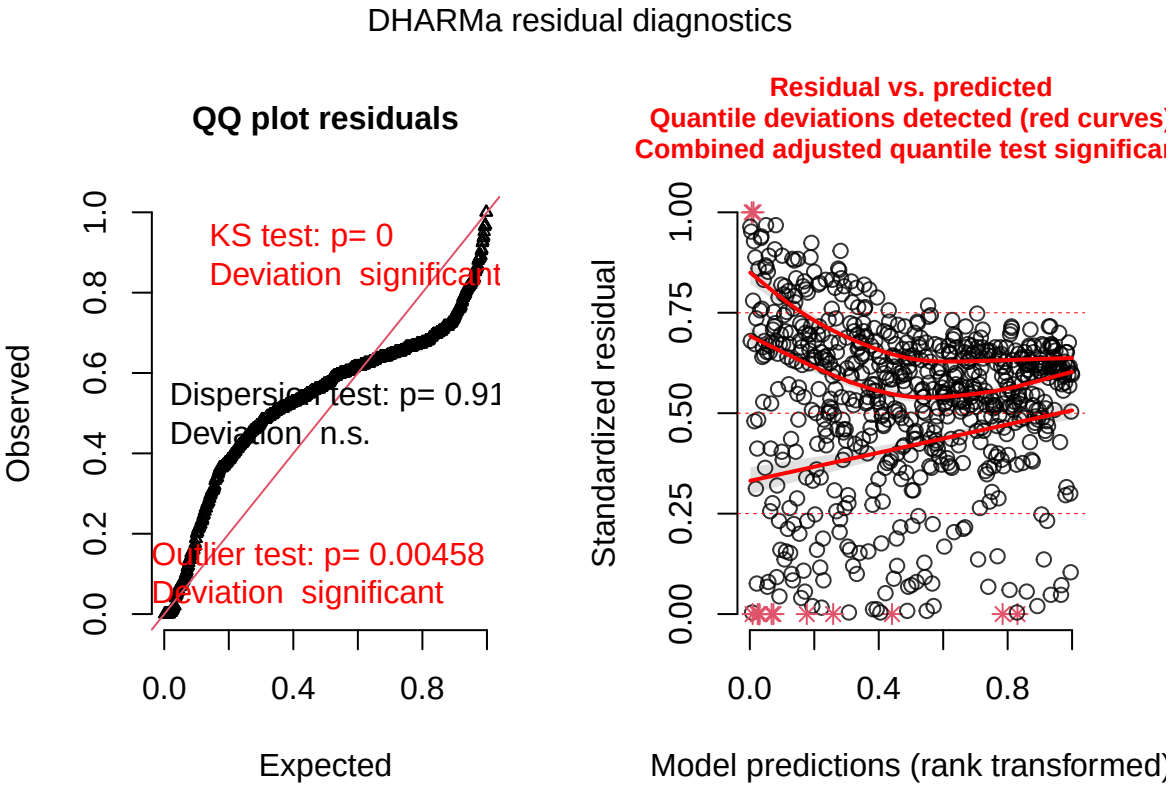
```
## Warning in asinh(z): NaNs produced

## Warning in asinh(z): NaNs produced
```

```
## Warning in asinh(z): NaNs produced
```



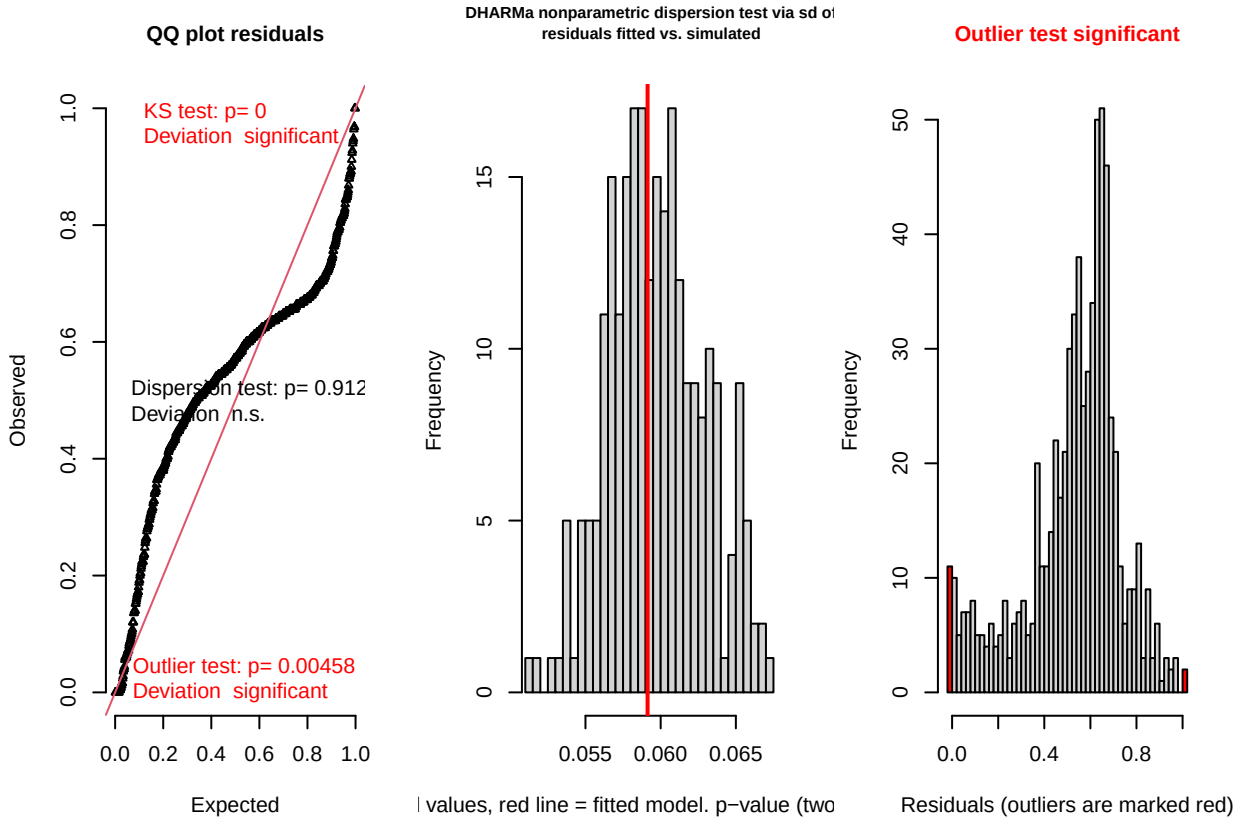

```
simulationOut3 <- simulateResiduals(fittedModel = updt_lm_out3, n = 250)

plot(simulationOut3)
```

```
## Warning in asinh(z): NaNs produced

## Warning in asinh(z): NaNs produced

## Warning in asinh(z): NaNs produced
```



The 'DHARMa' package uses a simulation-based approach to create readily interpretable scaled (quantile) residuals for fitted (generalized) linear mixed models. The resulting residuals are standardized to values between 0 and 1 and can be interpreted as intuitively as residuals from a linear regression. The package also provides a number of plot and test functions for typical model misspecification problems, such as over/underdispersion, zero-inflation, and residual spatial and temporal autocorrelation. In case of above plots, as Q-Q plot is not exactly along the red line, we can say that there is no strong liner relationship with dependent variable.

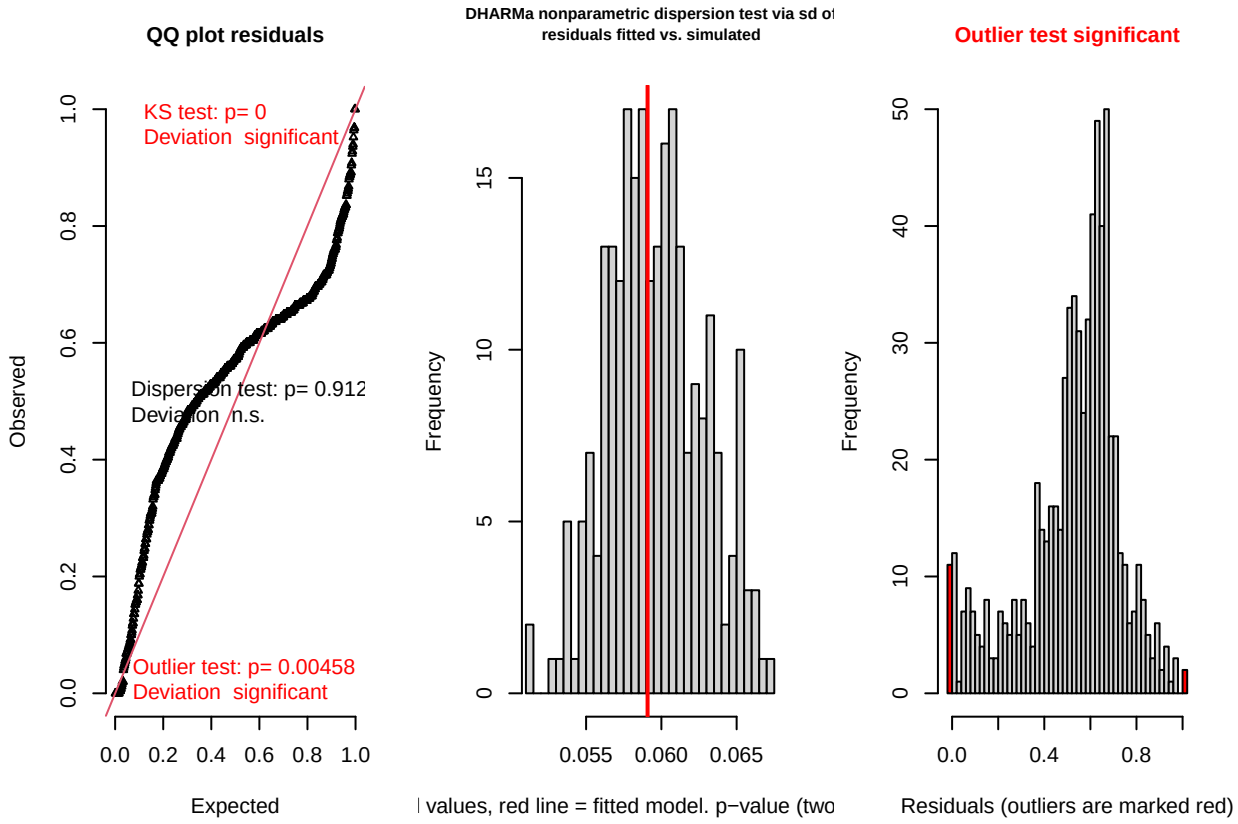
```
# The test are run as follows:
invisible(testResiduals(simulationOut2))
```




```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.19443, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.99014, p-value = 0.912
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 13, observations = 696, p-value = 0.004583
```

```
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.009981883 0.031728442
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                              0.01867816
```

```
invisible(testResiduals(simulationOut3))
```




```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.19302, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.99014, p-value = 0.912
```

```
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 13, observations = 696, p-value = 0.004583
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.009981883 0.031728442
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.01867816
```

The above results we can see that Q-Q plot has a small curve and Dispersion plot is normally distributed with some skewness. Also, DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated shows that it failed to reject the null hypothesis as p_value is greater than 0.05.

```
summary(updt_lm_out2)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ WithMMR + PctBeliefExempt + Enrolled_log +
##     PctFamilyPoverty + DistrictComplete_new, data = mydistricts_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.173  -0.437   0.506   1.199  12.358
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -19.726332   1.673817 -11.785  < 2e-16 ***
## WithMMR                    1.174561   0.016919  69.421  < 2e-16 ***
## PctBeliefExempt            0.166178   0.021555   7.709 4.42e-14 ***
## Enrolled_log               0.008933   0.078362   0.114   0.9093
## PctFamilyPoverty           0.017453   0.015147   1.152   0.2496
## DistrictComplete_newTRUE   1.017080   0.523221   1.944   0.0523 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.054 on 690 degrees of freedom
## Multiple R-squared:  0.9409, Adjusted R-squared:  0.9405
## F-statistic:  2198 on 5 and 690 DF,  p-value: < 2.2e-16
```

```
summary(updt_lm_out3)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ WithMMR + PctBeliefExempt + Enrolled_log +
```

```
##     PctFreeMeal + DistrictComplete_new, data = mydistricts_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.133  -0.430   0.532   1.205  12.316
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -19.940649   1.689274 -11.804  < 2e-16 ***
## WithMMR                    1.175350   0.016838  69.804  < 2e-16 ***
## PctBeliefExempt            0.168964   0.021771   7.761 3.05e-14 ***
## Enrolled_log               0.008159   0.078199   0.104   0.9169
## PctFreeMeal                0.006792   0.004993   1.360   0.1742
## DistrictComplete_newTRUE   1.020383   0.521649   1.956   0.0509 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.053 on 690 degrees of freedom
## Multiple R-squared:  0.941,  Adjusted R-squared:  0.9405
## F-statistic:  2199 on 5 and 690 DF,  p-value: < 2.2e-16
```

In the above experiment, the 1st model updt_lm_out2 has , Multiple R-squared = 0.9409 and Adjusted R-squared = 0.9405 represents the proportion of about 94% variation in PctUpToDate (about its mean) explained by the multiple linear regression model with predictors in the model. Similarly, the 2nd model updt_lm_out3 has, Multiple R-squared = 0.941 and Adjusted R-squared = 0.9405 represents the proportion of about 94% variation in PctUpToDate (about its mean) explained by the multiple linear regression model with predictors in the model

Calculating beta weights below to verify how the standardized variations have been changed for all predictors

```
# checking beta weights to see standardized deviation
#install.packages("lm.beta")
library(lm.beta)
summary(lm.beta(updt_lm_out2))
```

```
##
## Call:
## lm(formula = PctUpToDate ~ WithMMR + PctBeliefExempt + Enrolled_log +
##     PctFamilyPoverty + DistrictComplete_new, data = mydistricts_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.173  -0.437   0.506   1.199  12.358
##
## Coefficients:
##                             Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)              -19.726332     0.000000   1.673817 -11.785  < 2e-16
## WithMMR                    1.174561     1.053935   0.016919  69.421  < 2e-16
## PctBeliefExempt            0.166178     0.115713   0.021555   7.709 4.42e-14
## Enrolled_log               0.008933     0.001129   0.078362   0.114   0.9093
## PctFamilyPoverty           0.017453     0.011189   0.015147   1.152   0.2496
## DistrictComplete_newTRUE   1.017080     0.018469   0.523221   1.944   0.0523
##
## (Intercept)              ***
```

```
## WithMMR                  ***
## PctBeliefExempt          ***
## Enrolled_log
## PctFamilyPoverty
## DistrictComplete_newTRUE .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.054 on 690 degrees of freedom
## Multiple R-squared:  0.9409, Adjusted R-squared:  0.9405
## F-statistic:  2198 on 5 and 690 DF,  p-value: < 2.2e-16
```

```
summary(lm.beta(updt_lm_out3))
```

```
##
## Call:
## lm(formula = PctUpToDate ~ WithMMR + PctBeliefExempt + Enrolled_log +
##     PctFreeMeal + DistrictComplete_new, data = mydistricts_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.133  -0.430   0.532   1.205  12.316
##
## Coefficients:
##                            Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)              -19.940649     0.000000   1.689274 -11.804  < 2e-16
## WithMMR                    1.175350     1.054643   0.016838  69.804  < 2e-16
## PctBeliefExempt            0.168964     0.117653   0.021771   7.761 3.05e-14
## Enrolled_log               0.008159     0.001031   0.078199   0.104   0.9169
## PctFreeMeal                0.006792     0.013374   0.004993   1.360   0.1742
## DistrictComplete_newTRUE   1.020383     0.018529   0.521649   1.956   0.0509
##
## (Intercept)              ***
## WithMMR                  ***
## PctBeliefExempt          ***
## Enrolled_log
## PctFreeMeal
## DistrictComplete_newTRUE .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.053 on 690 degrees of freedom
## Multiple R-squared:  0.941,  Adjusted R-squared:  0.9405
## F-statistic:  2199 on 5 and 690 DF,  p-value: < 2.2e-16
```

As we can see in the above experiment, Standardized deviations have been reduced only for Enrolled_log and TotalSchools_Log to some extent from std.error.

```
library(BayesFactor)

# checking Baye's Factor using lmBF function uptodate_lmbf_out2
uptodate_lmbf_out2 <- lmBF(PctUpToDate~WithMMR+PctBeliefExempt+Enrolled_log+PctFamilyPoverty+DistrictCo

uptodate_lmbf_out2
```

```
## Bayes factor analysis
## --------------
## [1] WithMMR + PctBeliefExempt + Enrolled_log + PctFamilyPoverty + DistrictComplete_new : 2.814014e+4
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
# running MCMC test on the model uptodate_mcmc_out2 to understand the boundaries of 95% HDI regions of
uptodate_mcmc_out2 <- lmBF(PctUpToDate~WithMMR+PctBeliefExempt+Enrolled_log+PctFamilyPoverty+DistrictCo

summary(uptodate_mcmc_out2)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                                      Mean       SD  Naive SE Time-series SE
## mu                                87.565525  0.25432 0.0025432      0.0025432
## WithMMR-WithMMR                    1.173965  0.01697 0.0001697      0.0001697
## PctBeliefExempt-PctBeliefExempt    0.165781  0.02184 0.0002184      0.0002184
## Enrolled_log-Enrolled_log          0.006307  0.07829 0.0007829      0.0007829
## PctFamilyPoverty-PctFamilyPoverty  0.016996  0.01519 0.0001519      0.0001519
## DistrictComplete_new-FALSE        -0.465017  0.25229 0.0025229      0.0025890
## DistrictComplete_new-TRUE          0.465017  0.25229 0.0025229      0.0025890
## sig2                               9.364492  0.50875 0.0050875      0.0050875
## g_DistrictComplete_new             1.551162 17.01818 0.1701818      0.2025707
## g_continuous                       5.227610  6.38503 0.0638503      0.0638503
##
## 2. Quantiles for each variable:
##
##                                      2.5%       25%      50%      75%    97.5%
## mu                                87.06121 87.396723 87.56750 87.74149 88.05577
## WithMMR-WithMMR                    1.14005  1.162663  1.17413  1.18550  1.20693
## PctBeliefExempt-PctBeliefExempt    0.12249  0.150907  0.16623  0.18043  0.20828
## Enrolled_log-Enrolled_log         -0.14733 -0.046260  0.00702  0.05946  0.15757
## PctFamilyPoverty-PctFamilyPoverty -0.01322  0.006899  0.01700  0.02728  0.04690
## DistrictComplete_new-FALSE        -0.97056 -0.629202 -0.46481 -0.29584  0.02169
## DistrictComplete_new-TRUE         -0.02169  0.295839  0.46481  0.62920  0.97056
## sig2                               8.41535  9.012491  9.35000  9.69674 10.39807
## g_DistrictComplete_new             0.04019  0.109603  0.22796  0.56304  6.65537
## g_continuous                       1.24207  2.390865  3.60630  5.91819 18.06226
```

```
# checking Baye's Factor using lmBF function on uptodate_lmbf_out3
uptodate_lmbf_out3 <- lmBF(PctUpToDate~WithMMR+PctBeliefExempt+Enrolled_log+PctFamilyPoverty+DistrictCo

uptodate_lmbf_out3
```

```
## Bayes factor analysis
## --------------
## [1] WithMMR + PctBeliefExempt + Enrolled_log + PctFamilyPoverty + DistrictComplete_new : 2.608073e+4
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
# running MCMC test on the model uptodate_mcmc_out3 to understand the boundaries of 95% HDI regions of
uptodate_mcmc_out3 <- lmBF(PctUpToDate~WithMMR+PctBeliefExempt+Enrolled_log+PctFamilyPoverty+DistrictCo

summary(uptodate_mcmc_out3)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                                      Mean       SD  Naive SE Time-series SE
## mu                                87.562954  0.25526 0.0025526      0.0025443
## WithMMR-WithMMR                    1.174150  0.01680 0.0001680      0.0001680
## PctBeliefExempt-PctBeliefExempt    0.166191  0.02140 0.0002140      0.0002140
## Enrolled_log-Enrolled_log          0.007582  0.07848 0.0007848      0.0007848
## PctFamilyPoverty-PctFamilyPoverty  0.016996  0.01515 0.0001515      0.0001515
## DistrictComplete_new-FALSE        -0.469965  0.25456 0.0025456      0.0025797
## DistrictComplete_new-TRUE          0.469965  0.25456 0.0025456      0.0025797
## sig2                               9.363279  0.50266 0.0050266      0.0050040
## g_DistrictComplete_new             1.520294 38.37840 0.3837840      0.3837840
## g_continuous                       5.172977  5.80709 0.0580709      0.0583937
##
## 2. Quantiles for each variable:
##
##                                      2.5%       25%       50%      75%
## mu                                87.05388 87.392736 87.566888 87.73774
## WithMMR-WithMMR                    1.14085  1.162733  1.174029  1.18572
## PctBeliefExempt-PctBeliefExempt    0.12369  0.151825  0.165986  0.18053
## Enrolled_log-Enrolled_log         -0.14328 -0.047015  0.007271  0.06182
## PctFamilyPoverty-PctFamilyPoverty -0.01220  0.006808  0.016978  0.02706
## DistrictComplete_new-FALSE        -0.97369 -0.638048 -0.469689 -0.29607
## DistrictComplete_new-TRUE         -0.01834  0.296071  0.469689  0.63805
## sig2                               8.43999  9.012929  9.342755  9.69213
## g_DistrictComplete_new             0.03998  0.108572  0.225186  0.55394
## g_continuous                       1.20835  2.378973  3.664079  5.87361
##                                     97.5%
## mu                                88.05943
## WithMMR-WithMMR                    1.20726
## PctBeliefExempt-PctBeliefExempt    0.20815
## Enrolled_log-Enrolled_log          0.16043
## PctFamilyPoverty-PctFamilyPoverty  0.04717
```

```
## DistrictComplete_new-FALSE          0.01834
## DistrictComplete_new-TRUE           0.97369
## sig2                               10.40214
## g_DistrictComplete_new              5.57335
## g_continuous                       18.55305
```

Result 7th

A linear regression was performed twice to estimate the percentage of all enrolled students with completely up-to-date vaccines with use of WithMMR, PctBeliefExempt, Enrolled_log, PctFamilyPoverty and DistrictComplete_new as the predictors in the first model and WithMMR, PctBeliefExempt, Enrolled_log, PctFreeMeal and DistrictComplete_new as the predictors in the second model

Bi-variate exploratory data analysis for both linear models noted that the variables were somewhat skewed with a hint of a non-linear relationship. As the distributions were highly skewed for Enrolled and TotalSchools, so the data were log transformed for analysis, which generally improved the skew and the linearity of the relationship. The variables with percentage values were not transformed to not to hamper the data meaning and its overall efect on the dependent variable.

First linear regression found strong support for the relationship (F(5,690)=2198, p-value<0.001, adjusted R2 = 0.9405). Among predictors, WithMMR (b=1.174561, t=69.421, p<0.001) and PctBeliefExempt (b=0.166178, t=7.709, p<0.001) were significant. Enrolled_log (b=0.008933, t=0.114, p>0.05), PctFamilyPoverty(b=0.017453, t=1.152, p>0,05) and DistrictComplete_newTRUE i.e., district reporting completed(b=1.017080, t=1.944, p>0.05) were not significant as the p value is greater than 0.05 and we failed to reject the null hypothesis.

Second linear regression also found strong support for the relationship (F(5,690)=2199, p-value<0.001, adjusted R2 = 0.9405). Among predictors, WithMMR (b=1.175350, t=69.804, p<0.001) and PctBeliefExempt (b=0.168964, t=7.761, p<0.001) were significant. Enrolled_log (b=0.008159, t=0.104, p>0.05), PctFreeMeal(b=0.006792, t=1.360, p>0,05) and DistrictComplete_newTRUE i.e., district reporting completed(b=1.020383, t=1.956, p>0.05) were not significant as the p value is greater than 0.05 and we failed to reject the null hypothesis.

A Bayesian regression also found overwhelming evidence in support of both models with significant predictors WithMMR and PctBeliefExempt. Whereas Enrolled_log, PctFamilyPoverty, PctFreeMeal and DistrictComplete_newTRUE and DistrictComplete_newFALSE are not significant as they include 0 which tells us that they are not good predictors because there is chance that mean value is 0.

The BayesFactor analysis for the 1st model shows that Bayes Factor of 2.696566e+417:1 are very strong odds in the favor of alternative hypothesis. Similary, the BayesFactor analysis for the 2nd model shows that Bayes Factor of 2.711148e+417:1 are very strong odds in the favor of alternative hypothesis.So we reject the null hypothesis which suggest that Intercept only model is better.The sampled coefficients had similar values in both models, a mean of about 1.174 for WithMMR with an 95% HDI ranging about of 1.14 to 1.20 and a mean of about 0.166 for PctBeliefExempt with an 95% HDI ranging about of 0.12 to 0.20. This result is perfectly aligning with the traditional linear model analysis.

Overall, we can say that WithMMR and PctBeliefExempt provide an excellent estimate of the percentage of all enrolled students with with completely up-to-date vaccines.

8. *In predicting the percentage of all enrolled students with completely up-to-date vaccines, is there an interaction between PctChildPoverty and Enrolled?*

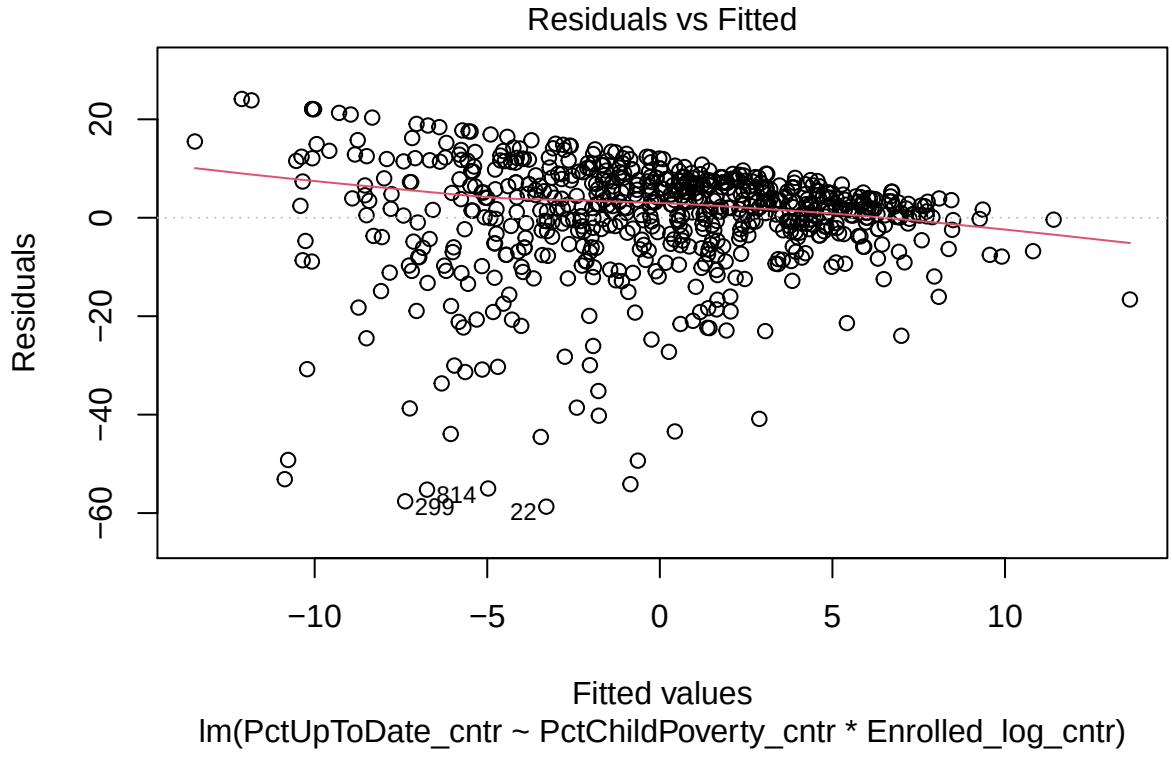
```
# we need to center the predictor variables first to proceed with regression analysis on intercation te

mydistricts_new$PctUpToDate_cntr <- scale(mydistricts_new$PctUpToDate, center=T, scale= F)
```

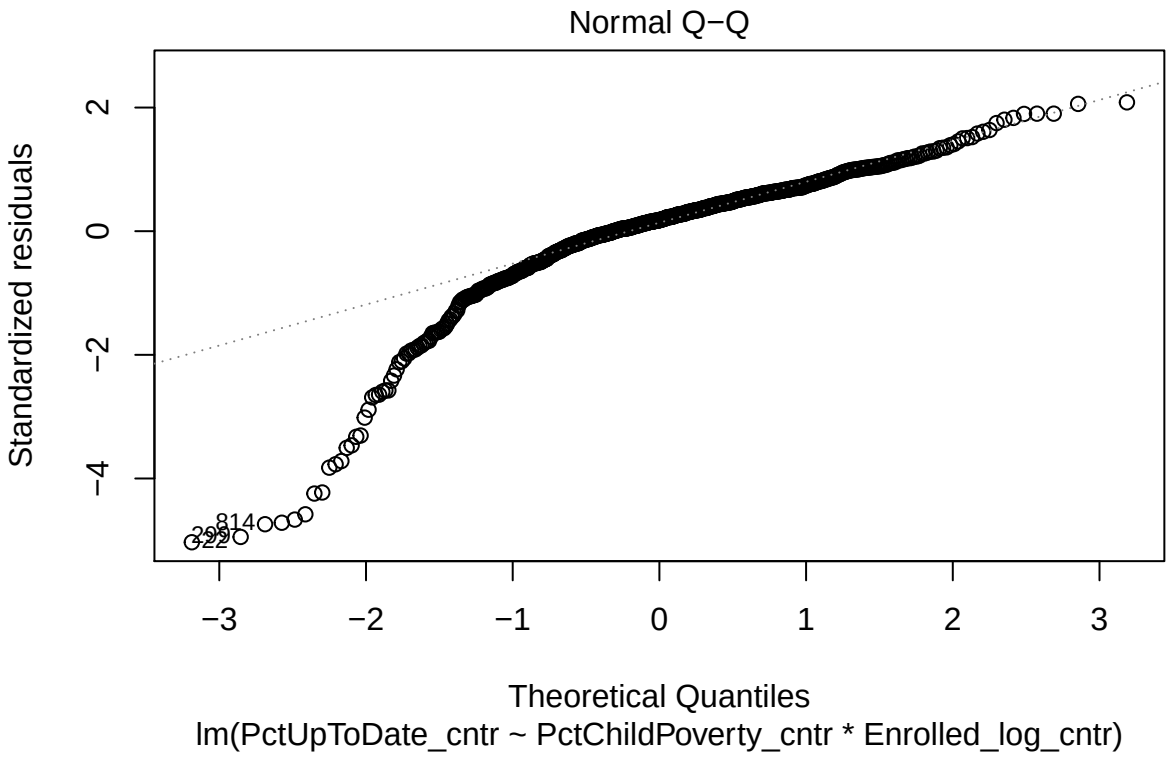
```
mydistricts_new$PctChildPoverty_cntr <- scale(mydistricts_new$PctChildPoverty, center=T, scale= F)

mydistricts_new$Enrolled_log_cntr <- scale(mydistricts_new$Enrolled_log, center=T, scale= F)


# conducting the linear regression on centered data
lm_updt_intr <- lm(PctUpToDate_cntr~PctChildPoverty_cntr*Enrolled_log_cntr, data =mydistricts_new)

# checking Multicolinearity

vif(lm_updt_intr)
```

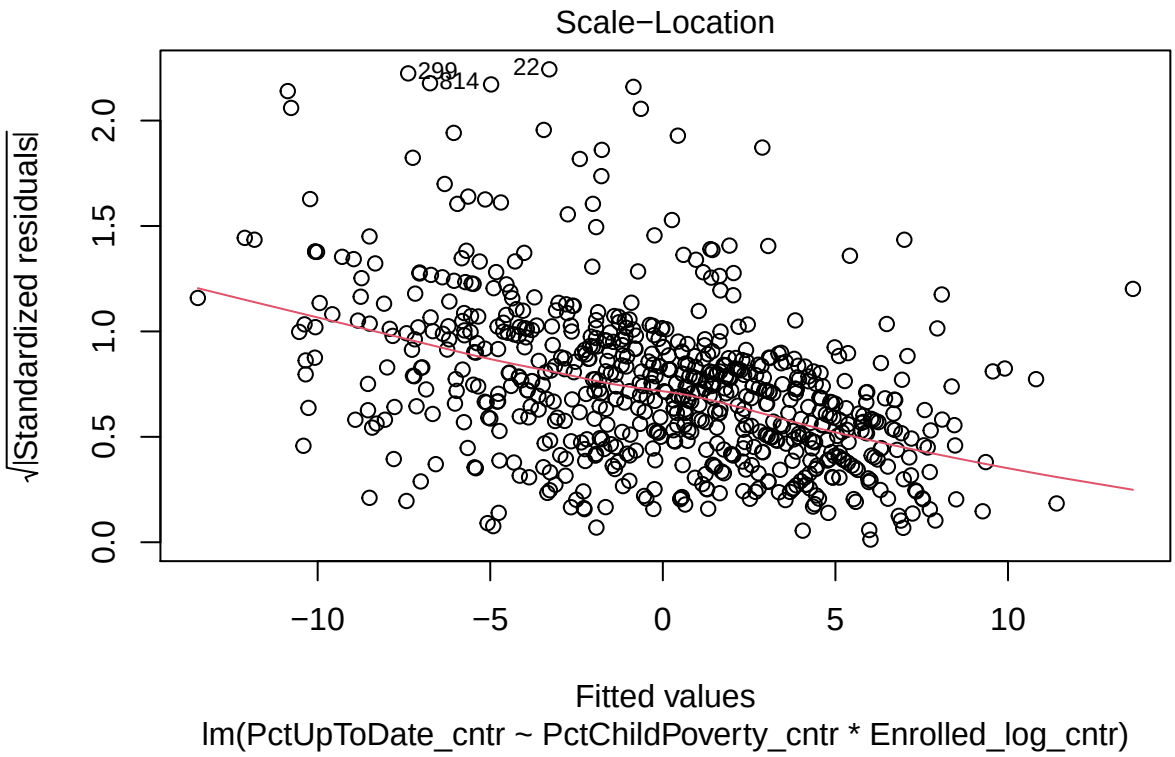
```
##                   PctChildPoverty_cntr                         Enrolled_log_cntr
##                                1.025568                                  1.011772
## PctChildPoverty_cntr:Enrolled_log_cntr
##                                1.025989
```

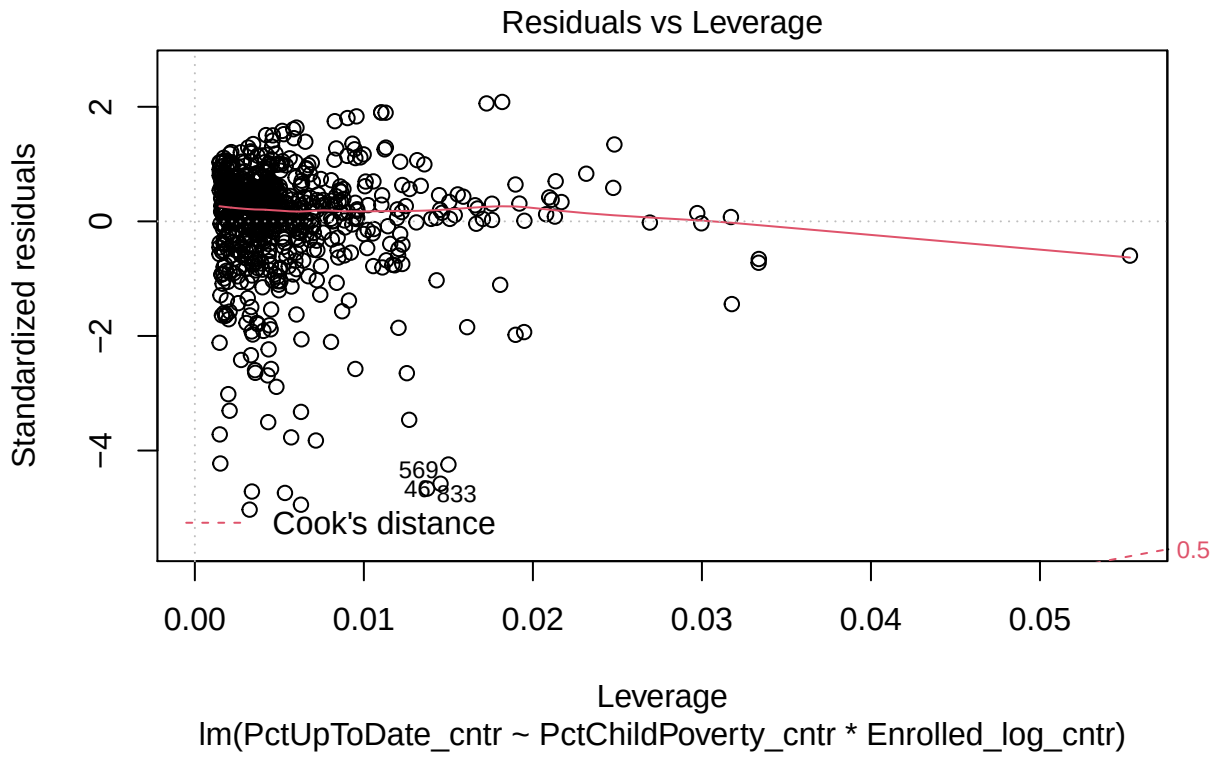
In the above results of multicolinearity, we can see that the variance inflation factor is low i.e. < 5 for centered interaction data PctChildPoverty_cntr:Enrolled_log_cntr [1.025989]. let us check the residual plot and histographic plot for residual's distribution.

```
# plotting model
plot(lm_updt_intr)
```

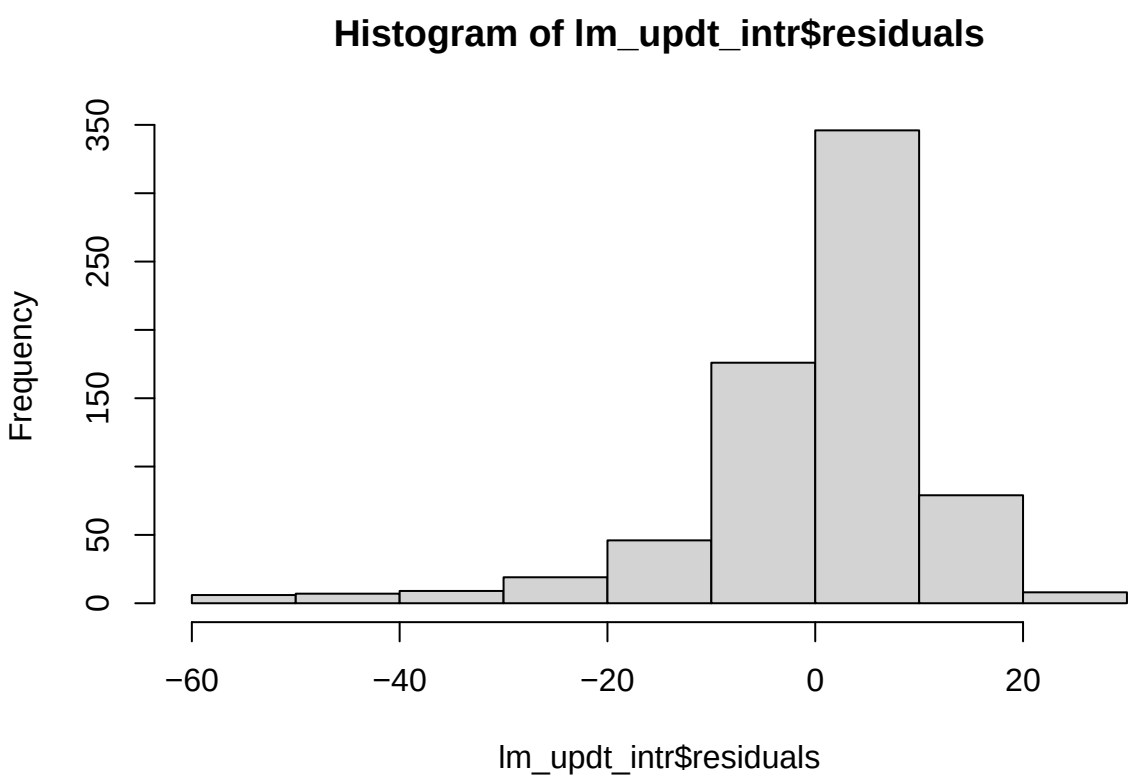



Normal Q−Q
Standardized residuals
2
0
−2
−4
814
299
622
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(PctUpToDate_cntr ~ PctChildPoverty_cntr * Enrolled_log_cntr)


Scale−Location
299
814
22
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
−10
−5
0
5
10
Fitted values
lm(PctUpToDate_cntr ~ PctChildPoverty_cntr * Enrolled_log_cntr)




```
# checking the distribution for residuals
hist(lm_updt_intr$residuals)
```



In above plots, we can see that the distribution for residual is highly left skewed. This skewness can also be seen in the residual Q-Q plot, the lower side is not aligning properly with red line. But Residual vs Leverage plot shows that there is no point out of contour lines shown by Cook's distance.

Checking the significance in the results

```
summary(lm_updt_intr)
```

```
##
## Call:
## lm(formula = PctUpToDate_cntr ~ PctChildPoverty_cntr * Enrolled_log_cntr,
##     data = mydistricts_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.692  -3.611   2.088   6.797  24.131
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -0.05712    0.44458  -0.128    0.898
## PctChildPoverty_cntr                  0.22740    0.03755   6.055  2.3e-09 ***
## Enrolled_log_cntr                     2.41176    0.28184   8.557  < 2e-16 ***
## PctChildPoverty_cntr:Enrolled_log_cntr -0.03770    0.02538  -1.485    0.138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 11.68 on 692 degrees of freedom
## Multiple R-squared:  0.1328, Adjusted R-squared:  0.129
## F-statistic: 35.32 on 3 and 692 DF,  p-value: < 2.2e-16
```

In the above interaction model experiment, model updt_lm_out2 has , Multiple R-squared = 0.1328 and Adjusted R-squared = 0.129 represents the proportion of about 13% variation in PctUpToDate_cntr (about its mean) explained by the multiple linear regression model with predictors in the model.

Calculating beta weights below to verify how the standardized variations have been changed for all predictors

```
# checking beta weights to see standardized deviation
#install.packages("lm.beta")
library(lm.beta)
summary(lm.beta(lm_updt_intr))
```

```
##
## Call:
## lm(formula = PctUpToDate_cntr ~ PctChildPoverty_cntr * Enrolled_log_cntr,
##     data = mydistricts_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.692  -3.611   2.088   6.797  24.131
##
## Coefficients:
##                                       Estimate Standardized Std. Error t value
## (Intercept)                           -0.05712      0.00000    0.44458  -0.128
## PctChildPoverty_cntr                   0.22740      0.21708    0.03755   6.055
## Enrolled_log_cntr                      2.41176      0.30470    0.28184   8.557
## PctChildPoverty_cntr:Enrolled_log_cntr -0.03770     -0.05326    0.02538  -1.485
##                                       Pr(>|t|)
## (Intercept)                              0.898
## PctChildPoverty_cntr                    2.3e-09 ***
## Enrolled_log_cntr                       < 2e-16 ***
## PctChildPoverty_cntr:Enrolled_log_cntr    0.138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.68 on 692 degrees of freedom
## Multiple R-squared:  0.1328, Adjusted R-squared:  0.129
## F-statistic: 35.32 on 3 and 692 DF,  p-value: < 2.2e-16
```

Now conducting a Bayesian linear regression analysis, using the facilities in the BayesFactor package.

```
library(BayesFactor)

# Calculating Bayes Factor
lmbf_updt_intr <- lmBF(PctUpToDate_cntr~PctChildPoverty_cntr*Enrolled_log_cntr, data =mydistricts_new)

lmbf_updt_intr
```

```
## Bayes factor analysis
## --------------
```

```
## [1] PctChildPoverty_cntr * Enrolled_log_cntr : 1.270309e+18 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
# running the MCMC test on the lmbf_updt_intr using posterior distrbution and 10000 iterations to verif
lmbf_updt_intr1 <- lmBF(PctUpToDate_cntr~PctChildPoverty_cntr*Enrolled_log_cntr, data =mydistricts_new,

summary(lmbf_updt_intr1)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                                             Mean      SD  Naive SE
## mu                                      -0.001998 0.44529 0.0044529
## PctChildPoverty_cntr                     0.223492 0.03715 0.0003715
## Enrolled_log_cntr                        2.364209 0.28283 0.0028283
## PctChildPoverty_cntr.&.Enrolled_log_cntr -0.036933 0.02498 0.0002498
## sig2                                   136.777290 7.37099 0.0737099
## g                                        0.140243 0.27336 0.0027336
##                                         Time-series SE
## mu                                          0.0044529
## PctChildPoverty_cntr                        0.0003715
## Enrolled_log_cntr                           0.0028283
## PctChildPoverty_cntr.&.Enrolled_log_cntr    0.0002498
## sig2                                        0.0737099
## g                                           0.0027336
##
## 2. Quantiles for each variable:
##
##                                              2.5%       25%        50%
## mu                                       -0.88024  -0.29803  2.446e-04
## PctChildPoverty_cntr                      0.15078   0.19857  2.233e-01
## Enrolled_log_cntr                         1.80461   2.17233  2.366e+00
## PctChildPoverty_cntr.&.Enrolled_log_cntr -0.08505  -0.05397 -3.679e-02
## sig2                                   123.13481 131.81545  1.364e+02
## g                                        0.02379   0.05082  8.286e-02
##                                             75%     97.5%
## mu                                       0.3011   0.86337
## PctChildPoverty_cntr                     0.2487   0.29635
## Enrolled_log_cntr                        2.5536   2.92682
## PctChildPoverty_cntr.&.Enrolled_log_cntr -0.0201   0.01187
## sig2                                   141.5085 152.12247
## g                                        0.1461   0.60181
```

8Th result

A linear regression was performed to estimate the percentage of all enrolled students with completely up-to-date vaccines with use of interaction between PctChildPoverty and Enrolled_log. We decided to centered the data to get the better interaction results between two predictors.

_We decided to interpret the interaction term first, in case it influences how we make sense out of the linear main effects. In this case the interaction PctChildPoverty_cntr*Enrolled_log_cntr coefficient is statistically not significantly different from 0 with a t-value of -1.485 and a p-value greater than 0.05 and we failed to reject the null hypothesis._

Bi-variate exploratory data analysis done earlier made us use the data which was log transformed for analysis, which generally improved the skew and the linearity of the relationship. The main linear effect found strong support for the relationship (F(3,692)=35.32, p-value<0.001, adjusted R2 = 0.129). Among predictors, PctChildPoverty_cntr (b=0.22740, t=6.055, p<0.001) and Enrolled_log_cntr (b=2.41176, t=8.557, p<0.001) were significant.

A Bayesian regression also found overwhelming evidence in support of a model with significant predictors PctChildPoverty and Enrolled_log. The BayesFactor analysis shows that Bayes Factor of 1.270309e+18:1 are very strong odds in the favor of alternative hypothesis that a interaction model is better than intercept model. So we reject the null hypothesis which suggest that Intercept only model is better. The sampled coefficients had similar values, a mean of 0.222966 for PctChildPoverty_cntr with an 95% HDI of 0.15019 to 0.29506, and a mean of 2.361761 for Enrolled_log_cntr with an 95% HDI of 1.81599 to 2.92110. Apart from this, we can see that, a mean of -0.037283 for an interaction PctChildPoverty_cntr.&.Enrolled_log_cntr with 95% HDI of -0.08596 to 0.01264 shows that HDI has 0, which tells us that interaction is not a good predictor because there is chance that mean value is 0. This result is perfectly aligning with the traditional linear model analysis.

Overall, we can say that interaction between PctChildPoverty and Enrolled_log is not significant to estimate of the percentage of all enrolled students with with completely up-to-date vaccines , but still interaction model's main linear effect will be able to predict the percentage of all enrolled students with with completely up-to-date vaccines correctly.

9. *Which, if any, of the four predictor variables predict whether or not a district's reporting was complete?*

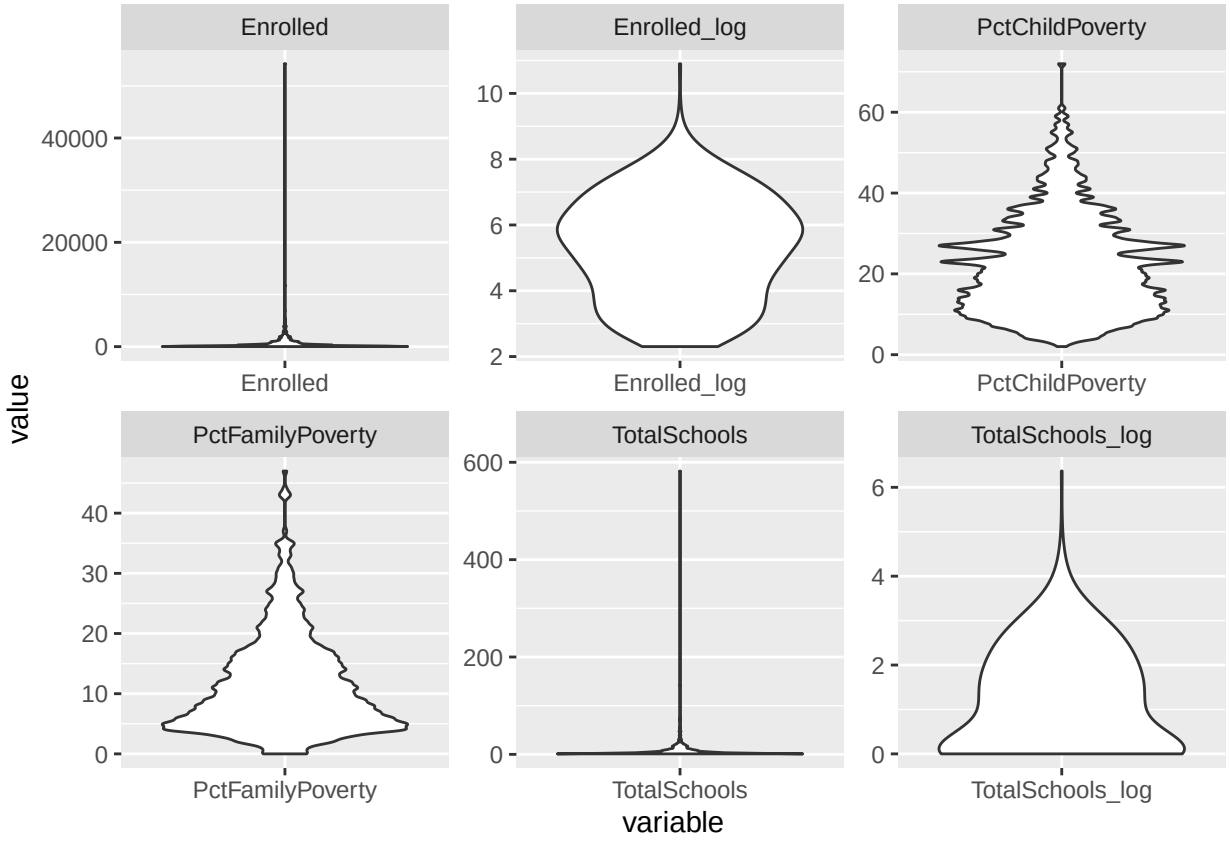
```
# Creating a new dataset

district_reporting <- subset(mydistricts_new, select = c("DistrictComplete_new","PctChildPoverty",
                                                         "PctFamilyPoverty","Enrolled","TotalSchools","Enrolle

diagnose(district_reporting)
```

```
## # A tibble: 7 x 6
##   variables        types  missing_count missing_percent unique_count unique_rate
##   <chr>            <chr>          <int>           <dbl>        <int>       <dbl>
## 1 DistrictComplet~ factor             0               0            2     0.00287
## 2 PctChildPoverty  numer~             0               0           59     0.0848
## 3 PctFamilyPoverty numer~             0               0           40     0.0575
## 4 Enrolled         numer~             0               0          455     0.654
## 5 TotalSchools     numer~             0               0           44     0.0632
## 6 Enrolled_log     numer~             0               0          455     0.654
## 7 TotalSchools_log numer~             0               0           44     0.0632
```

```
describe(district_reporting)
```

```
## # A tibble: 6 x 26
##   variable          n    na   mean     sd se_mean    IQR skewness kurtosis   p00
##   <chr>         <int> <int>  <dbl>  <dbl>   <dbl>  <dbl>    <dbl>    <dbl> <dbl>
## 1 PctChildPove~   696     0  22.3  1.20e1  0.453   16      0.769     0.368  2
## 2 PctFamilyPov~   696     0  11.5  8.03e0  0.304   11      1.21      1.64   0
## 3 Enrolled        696     0 634.   2.23e3 84.6    631.    20.3     480.    10
## 4 TotalSchools    696     0   7.28 2.41e1  0.914    7     20.0     467.     1
## 5 Enrolled_log    696     0   5.29 1.58e0  0.0600   2.56  -0.0201   -0.763  2.30
## 6 TotalSchools~   696     0   1.17 1.15e0  0.0437   2.08   0.632    -0.328  0
## # ... with 16 more variables: p01 <dbl>, p05 <dbl>, p10 <dbl>, p20 <dbl>,
## #   p25 <dbl>, p30 <dbl>, p40 <dbl>, p50 <dbl>, p60 <dbl>, p70 <dbl>,
## #   p75 <dbl>, p80 <dbl>, p90 <dbl>, p95 <dbl>, p99 <dbl>, p100 <dbl>
```

```
library(tidyverse)
district_reporting %>% pivot_longer(cols=-c(DistrictComplete_new), names_to="variable",
                        values_to="value", values_drop_na = TRUE) %>%
ggplot(aes(x=variable, y=value)) + geom_violin(bw=.5) + facet_wrap( ~ variable, scales="free")
```



As we can see from above plots and the results of the describe function, the skewness has been reduced because of log transformed data in the Enrolled and TotalSchools data. We will proceed with the analysis of whether the reporting for district is complete or not using log transformed variables.

Also as there are no missing values we can proceed to make the model using four predictors as required PctChildPoverty, PctFamilyPoverty, Enrolled_log, TotalSchools_log.

```
# checking the correlation Matrix

round(cor(district_reporting[2:dim(district_reporting)[2]]),3)
```

```
##                   PctChildPoverty PctFamilyPoverty Enrolled TotalSchools
## PctChildPoverty             1.000            0.864    0.026        0.021
## PctFamilyPoverty            0.864            1.000    0.044        0.038
## Enrolled                    0.026            0.044    1.000        0.994
## TotalSchools                0.021            0.038    0.994        1.000
## Enrolled_log               -0.080            0.018    0.416        0.406
## TotalSchools_log           -0.099           -0.020    0.474        0.480
##                   Enrolled_log TotalSchools_log
## PctChildPoverty         -0.080           -0.099
## PctFamilyPoverty         0.018           -0.020
## Enrolled                 0.416            0.474
## TotalSchools             0.406            0.480
## Enrolled_log             1.000            0.917
## TotalSchools_log         0.917            1.000
```

From the above correlation matrix, we can see that PctFamilyPoverty and PctChildPoverty are highly positively correlated [0.864] with each other. Similarly, Enrolled_log and TotalSchools_log are highly positively correlated [0.917] with each other. This may cause some issues in the logistic regression ahead.

```
# creating Generalized linear model(logistic regression)

district_glm_out <- glm(DistrictComplete_new~PctChildPoverty+PctFamilyPoverty+Enrolled_log+TotalSchools

# checking Multicolinearity
library(car)
vif(district_glm_out)
```

```
##  PctChildPoverty PctFamilyPoverty     Enrolled_log TotalSchools_log
##         4.680916         4.695987        14.459060        14.479050
```

In the above results of Multicolinearity, the vif value of Enrolled_log and TotalSchools_log is greater than 10 and needs to be fixed.

```
# checking model assumptions
library(performance)
```

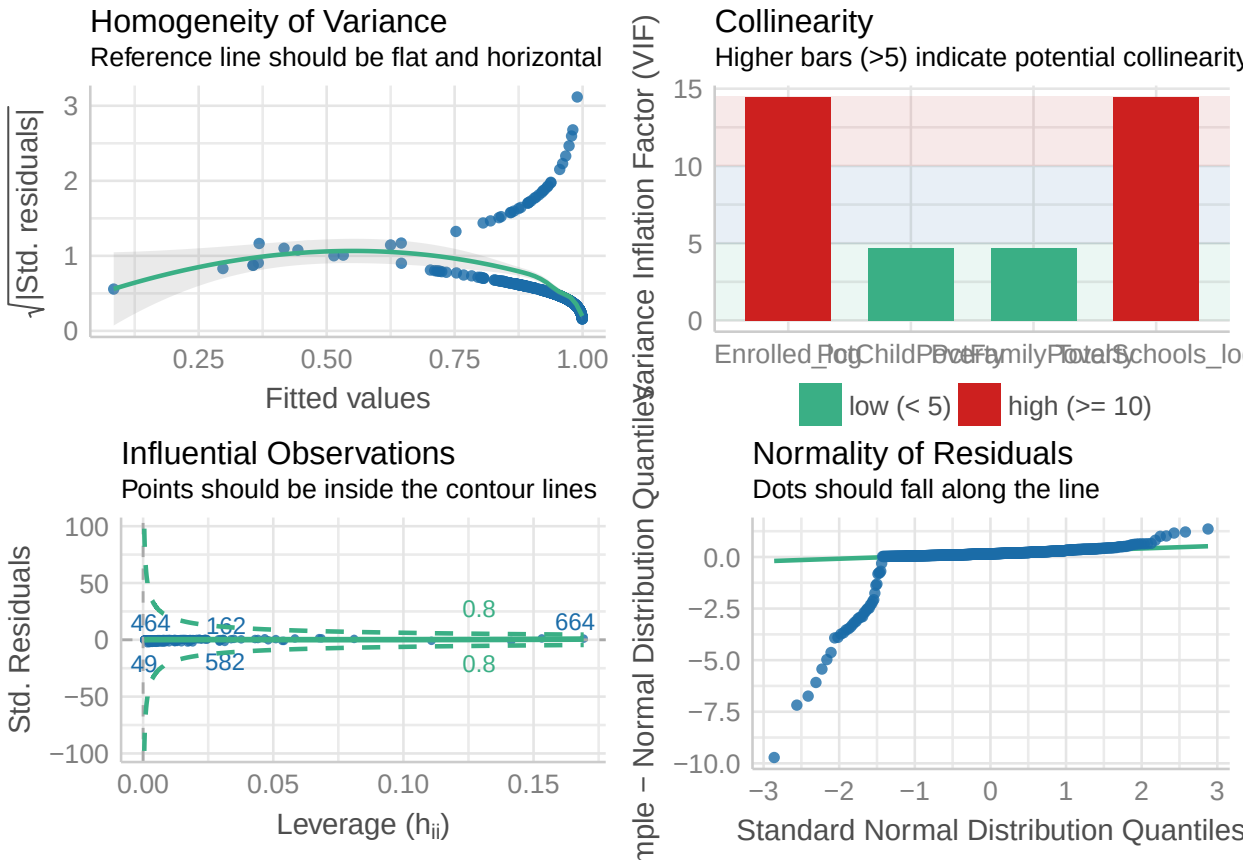
```
## Warning: package 'performance' was built under R version 4.0.5
```

```
library(see)
```

```
## Warning: package 'see' was built under R version 4.0.5
```

```
check_model(district_glm_out)
```

```
## Loading required namespace: qqplotr
```



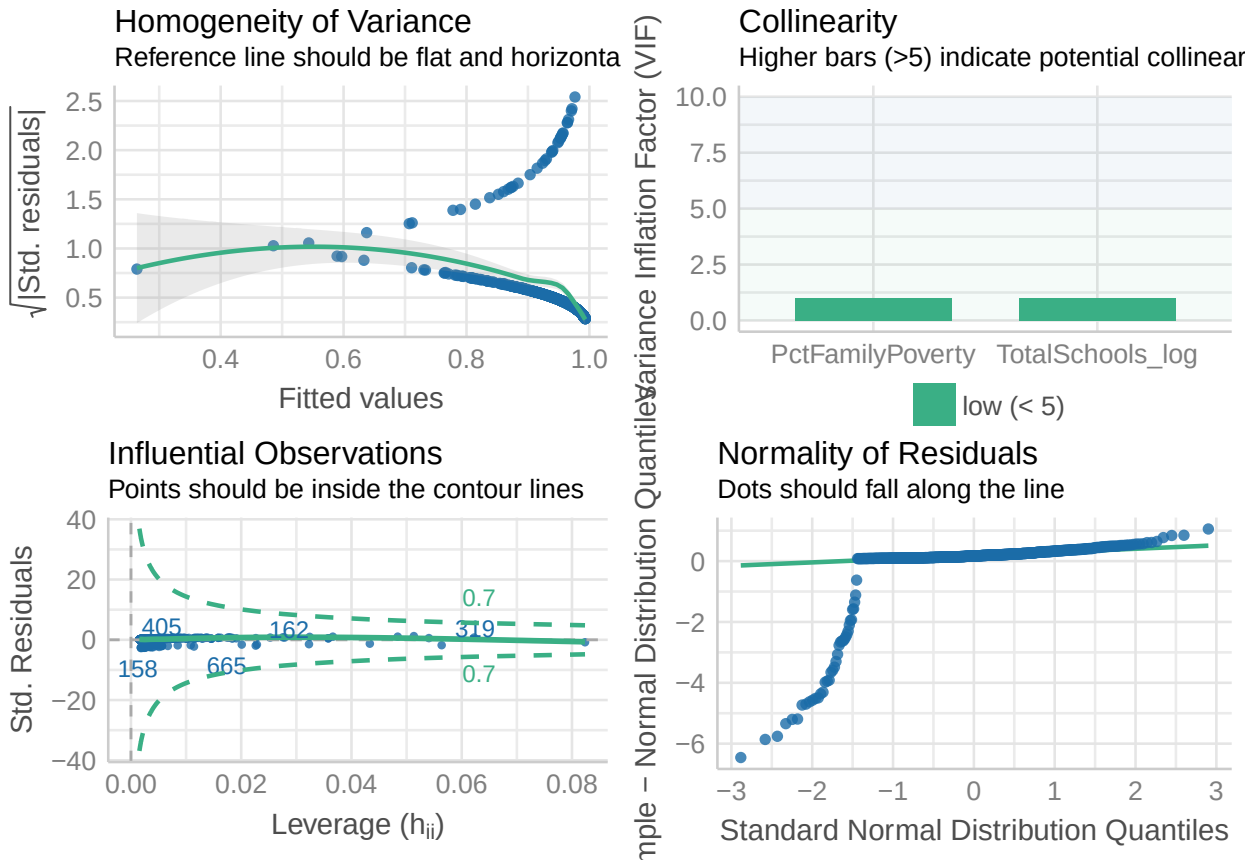
In the above plots as we can see there are colinearity issues and Influential observations plot shows some points on or out of countour lines, we have to remove highly self correlated variable with i.e. one with high variance inflation factor.

```
# changing the model predictors to remove the collinearity issues which was mentioned in the correlatio

district_glm_out2 <- glm(DistrictComplete_new~PctFamilyPoverty+TotalSchools_log, data = district_reporti

# checking Multicolinearity
library(car)
vif(district_glm_out2)
```

```
## PctFamilyPoverty TotalSchools_log
##         1.024313         1.024313
```

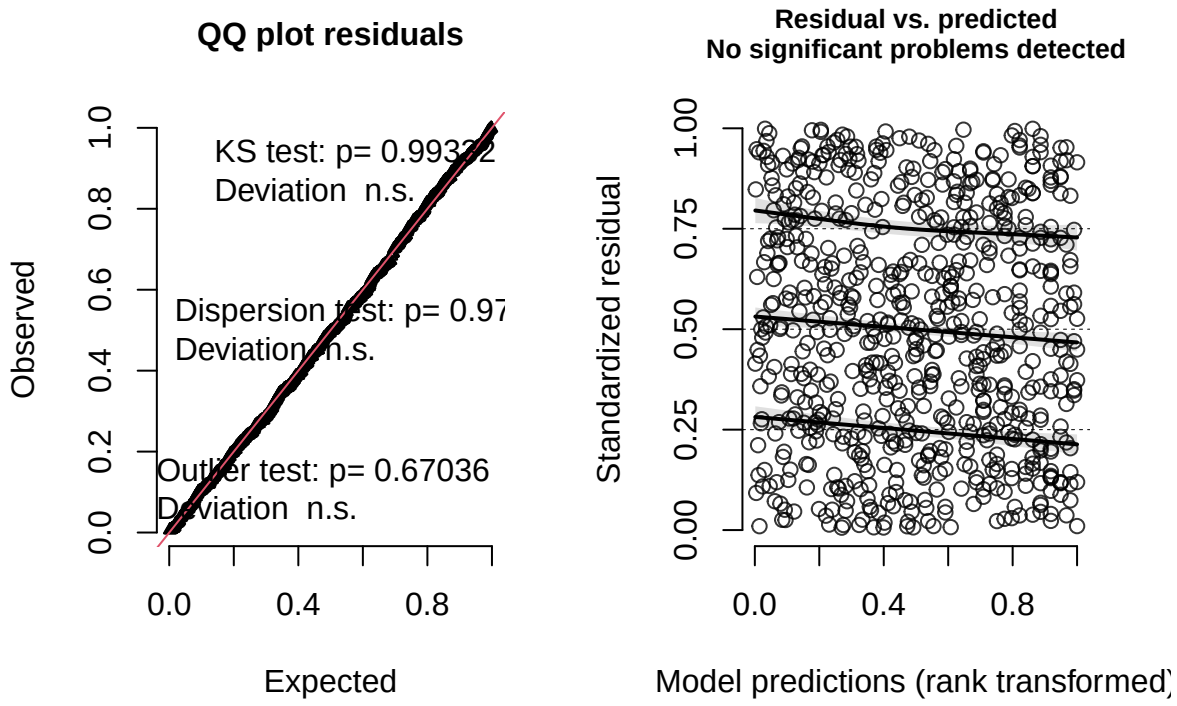
In the above results of multicolinearity, we can see that the variance inflation factor is less than 5 and we are good to go ahead with further analysis.

```
# checking model assumptions
library(performance)
library(see)
check_model(district_glm_out2)
```

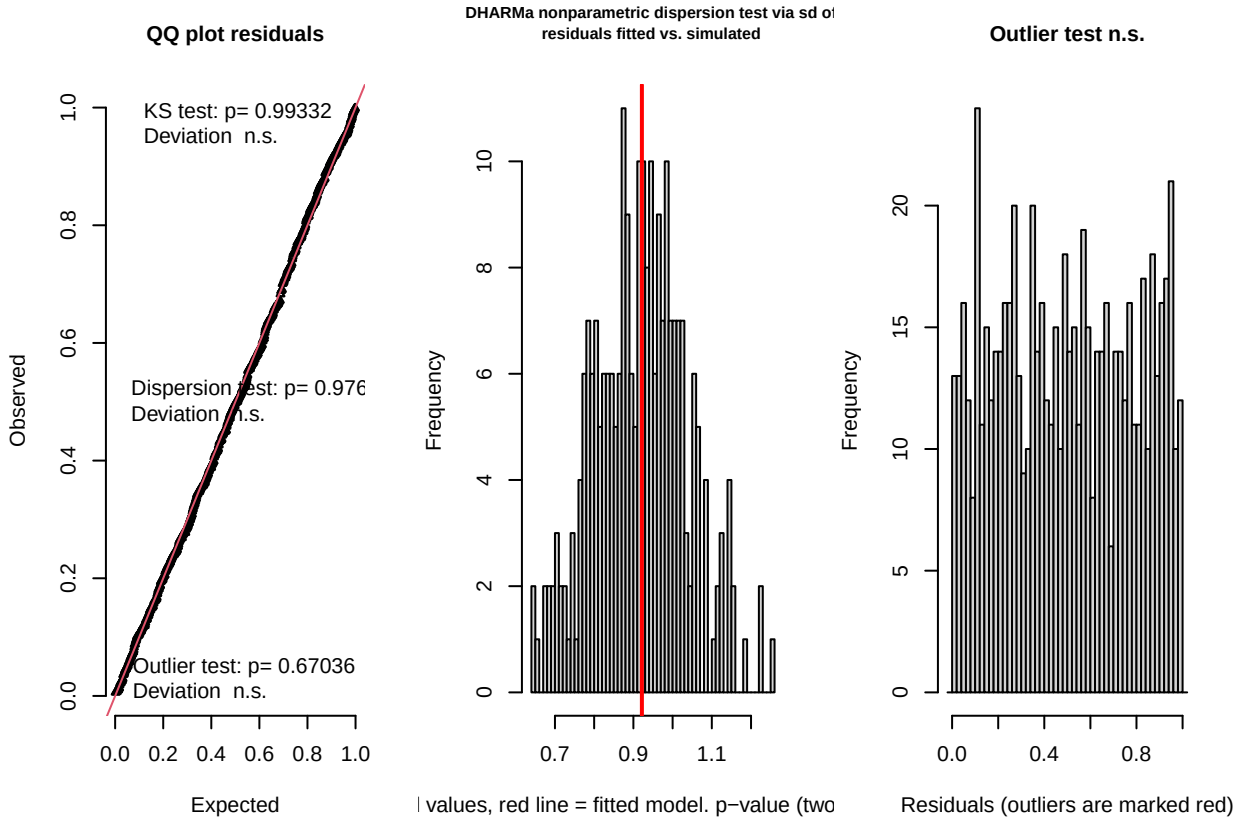


In the above plots, we can see that we have reduced potential collinearity issues from the model. The Homogeneity of Variance plot has less curve. Also the Influential Observations are well inside the contour lines.

```
#install.package("DHARMa")
library(DHARMa)
simulationOutput <- simulateResiduals(fittedModel = district_glm_out2, n = 250)
plot(simulationOutput)
```




```
# The test are run as follows:

testResiduals(simulationOutput)
```




```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.016169, p-value = 0.9933
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.007, p-value = 0.976
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 4, observations = 696, p-value = 0.6704
```

```
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.001568052 0.014649058
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                           0.005747126

## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.016169, p-value = 0.9933
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.007, p-value = 0.976
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 4, observations = 696, p-value = 0.6704
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.001568052 0.014649058
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                           0.005747126
```

```
summary(district_glm_out2)
```

```
##
## Call:
## glm(formula = DistrictComplete_new ~ PctFamilyPoverty + TotalSchools_log,
##     family = binomial(link = "logit"), data = district_reporting)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7393   0.1612   0.2387   0.3495   1.2016
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)       5.00324    0.48006  10.422  < 2e-16 ***
## PctFamilyPoverty -0.06760    0.01963  -3.445 0.000572 ***
## TotalSchools_log -0.76661    0.14348  -5.343 9.14e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 294.88  on 695  degrees of freedom
## Residual deviance: 254.34  on 693  degrees of freedom
## AIC: 260.34
##
## Number of Fisher Scoring iterations: 6
```

Running Omnibus test on district_glm_out2 model to verify whether the model is significant

```
anova(district_glm_out2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: DistrictComplete_new
##
## Terms added sequentially (first to last)
##
##
##                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                695     294.88
## PctFamilyPoverty  1   8.9978       694     285.88  0.002703 **
## TotalSchools_log  1  31.5356       693     254.34 1.958e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the above experiment, we can see that the predictor coefficients have reduced the Null deviance from 294.88 to Residual deviance of 254.34, which suggests that the model is predicting the canceled reservations by reducing the residual errors. The normal odds are predicting whether the reservation is likely to get canceled i.e, 1. The difference between the null deviance and the model deviance, in this case (294.88–254.34) = 40.54, is distributed as chi-square and can be directly tested from the output model. This is the omnibus test for this analysis, equivalent to the significance test of R-squared on a linear regression model. We can see the probability of observing a chi-square value of 8.9978 for PctFamilyPoverty on one degree of freedom and 31.5356 for TotalSchools_log on one degree of freedom is extremely low and well below our conventional thresholds of alpha, so we can reject the null hypothesis that introducing two predictor variables into the model caused zero reduction of model error. We can consider this rejection of the null hypothesis as evidence that the two-predictor model is preferred over the null model.

Converting the coefficient and 95% CI values to normal odds as below:

```
exp(coef(district_glm_out2))
```

```
##      (Intercept) PctFamilyPoverty TotalSchools_log
##      148.8949754        0.9346332        0.4645842
```

```
exp(confint(district_glm_out2))
```

```
## Waiting for profiling to be done...
```

```
##                       2.5 %      97.5 %
## (Intercept)      61.6107997 407.8250108
## PctFamilyPoverty  0.8997610   0.9722755
## TotalSchools_log  0.3465215   0.6102860
```

```
# converting normal odds of TotalSchools_log to TotalSchools by converting it back to normal value
exp(0.4645842)
```

```
## [1] 1.591352
```

```
# converting 95% CI of normal odds of TotalSchools_log to TotalSchools by converting it back to normal
exp(0.3465215) #2.5% quantile of CI
```

```
## [1] 1.41414
```

```
exp(0.6102860) #97.5% quantile of CI
```

```
## [1] 1.840958
```

checking the model performance

```
model_performance(district_glm_out2)
```

```
## Warning in Ops.factor(n, resp): '*' not meaningful for factors
```

```
## Warning in Ops.factor(n, resp): '*' not meaningful for factors
```

```
## # Indices of model performance
##
## AIC     |     BIC | Tjur's R2 |  RMSE | Sigma | Log_loss | Score_log | Score_spherical |   PCP
## ------------------------------------------------------------------------------------------------
## 260.343 | 273.979 |     0.081 | 0.218 | 0.606 |    0.183 |      -Inf |           0.001 | 0.905
```

In the above model performance results, we can see that Tjur's R2 = 0.081 represents the proportion of about 8.10% variation in ADR (about its mean) explained by the multiple logistic regression model with predictors in the model.

Creating the Confusion matrix to verify the model accuracy

```
#install.packages("caret")
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.5
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
predicted_district<-round(predict(district_glm_out2,type="response"))

sum(predicted_district) # number we predict to be hired
```

```
## [1] 694
```

```
sum(as.numeric(district_reporting$DistrictComplete_new)) # number actually hired
```

```
## [1] 1354
```

```
confusion<-table(predicted_district, ifelse(district_reporting$DistrictComplete_new == "TRUE",1,0))

confusion
```

```
##
## predicted_district   0   1
##                  0   1   1
##                  1  37 657
```

```
addmargins(confusion)
```

```
##
## predicted_district   0   1 Sum
##                0     1   1   2
##                1    37 657 694
##                Sum  38 658 696
```

```
confusionMatrix(confusion, positive="1")
```

```
## Confusion Matrix and Statistics
##
##
## predicted_district   0   1
##                  0   1   1
##                  1  37 657
##
##                Accuracy : 0.9454
##                  95% CI : (0.9258, 0.9611)
##     No Information Rate : 0.9454
##     P-Value [Acc > NIR] : 0.543
##
##                   Kappa : 0.0448
##
```

```
##  Mcnemar's Test P-Value : 1.365e-08
##
##             Sensitivity : 0.99848
##             Specificity : 0.02632
##          Pos Pred Value : 0.94669
##          Neg Pred Value : 0.50000
##              Prevalence : 0.94540
##          Detection Rate : 0.94397
##    Detection Prevalence : 0.99713
##       Balanced Accuracy : 0.51240
##
##        'Positive' Class : 1
##
```

As we can see in the confusion matrix, the model is able to predict whether the District Complete reporting is complete:TRUE or whether the District Complete reporting is bot complete: FALSE with 0.9454 or 94.54% accuracy.

Now conducting a Bayesian logistic regression analysis, using the facilities in the MCMCpack package.

```
#install.packages("MCMCpack")    # Download MCMCpack package
library(MCMCpack) # Load the package
```

```
## Warning: package 'MCMCpack' was built under R version 4.0.5
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)
```

```
## ## Copyright (C) 2003-2021 Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park
```

```
## ##
## ## Support provided by the U.S. National Science Foundation
```

```
## ## (Grants SES-0350646 and SES-0350613)
## ##
```

Running the MCMClogit() function using the model district_glm_out2

```
district_reporting$DistrictComplete_new1 <- ifelse(district_reporting$DistrictComplete_new == "TRUE",1,
```

```
district_MCMC_out <- MCMClogit(DistrictComplete_new1~PctFamilyPoverty+TotalSchools_log, data = district_
```

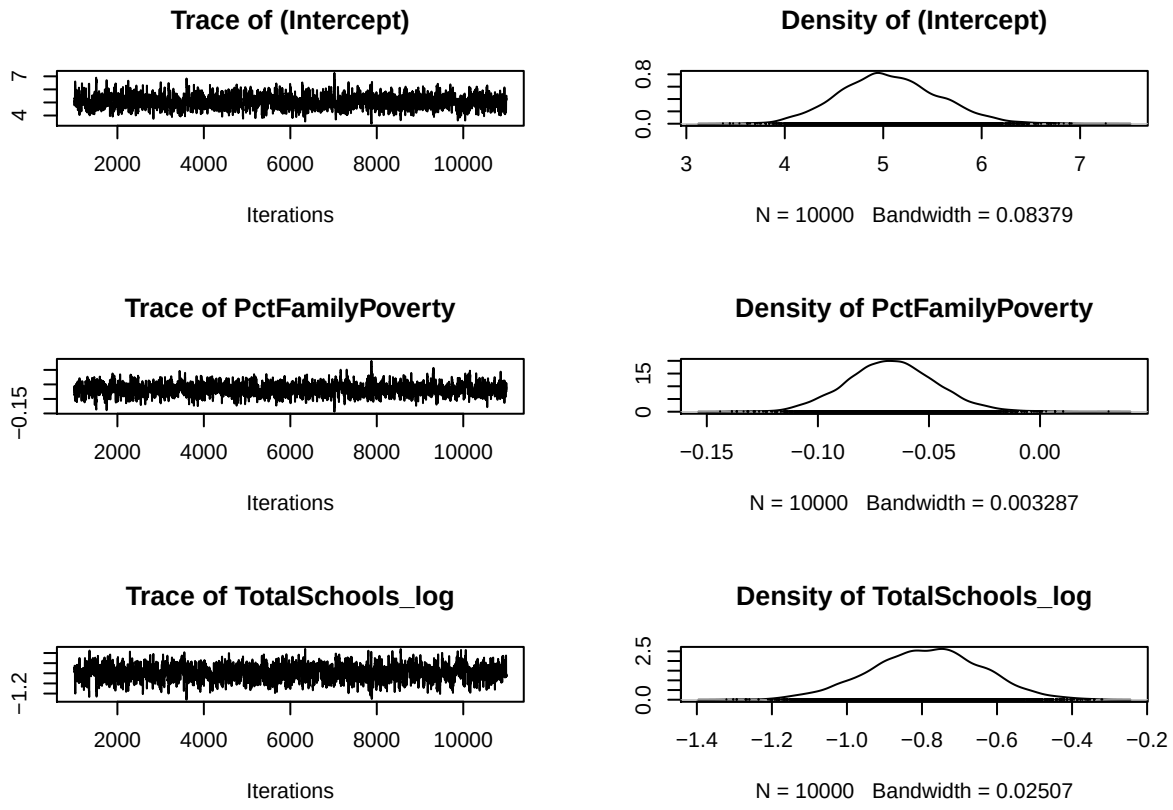
Running summary() on the output object.

```
summary(district_MCMC_out)
```

```
##
## Iterations = 1001:11000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                       Mean      SD  Naive SE Time-series SE
## (Intercept)       5.06477 0.49877 0.0049877      0.0165397
## PctFamilyPoverty -0.06687 0.02016 0.0002016      0.0006774
## TotalSchools_log -0.78787 0.14921 0.0014921      0.0048251
##
## 2. Quantiles for each variable:
##
##                      2.5%      25%      50%      75%    97.5%
## (Intercept)       4.1259  4.71602  5.03764  5.38822  6.07790
## PctFamilyPoverty -0.1059 -0.07998 -0.06731 -0.05376 -0.02581
## TotalSchools_log -1.0887 -0.88977 -0.78433 -0.68777 -0.50268
```

Creating a plot of the MCMC output

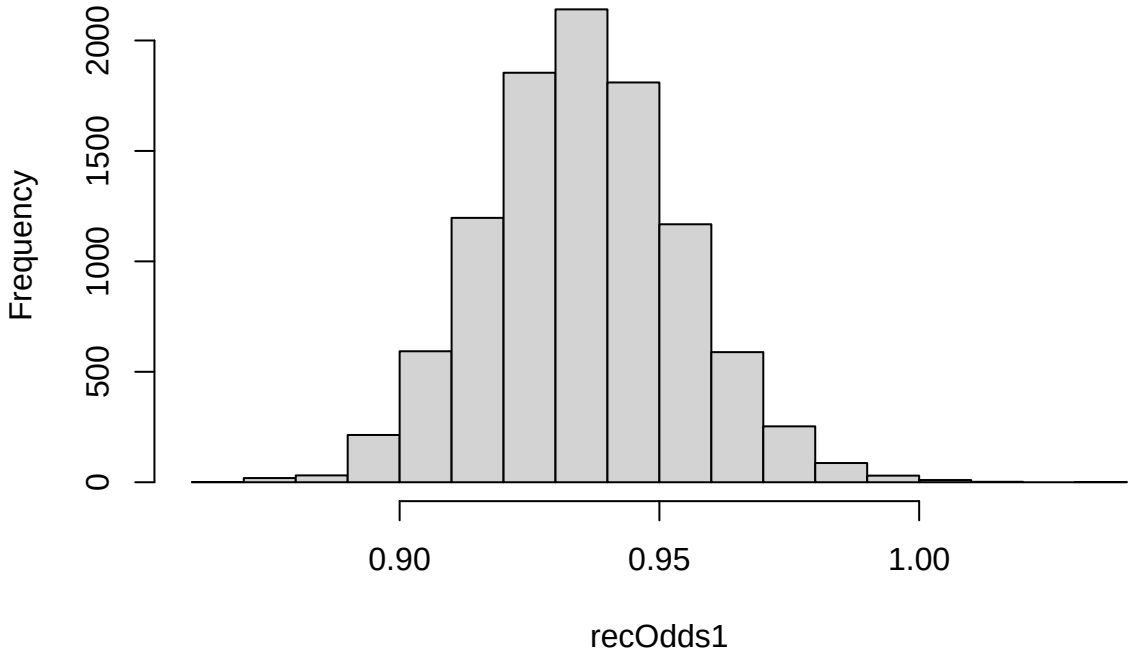
```
plot(district_MCMC_out)
```



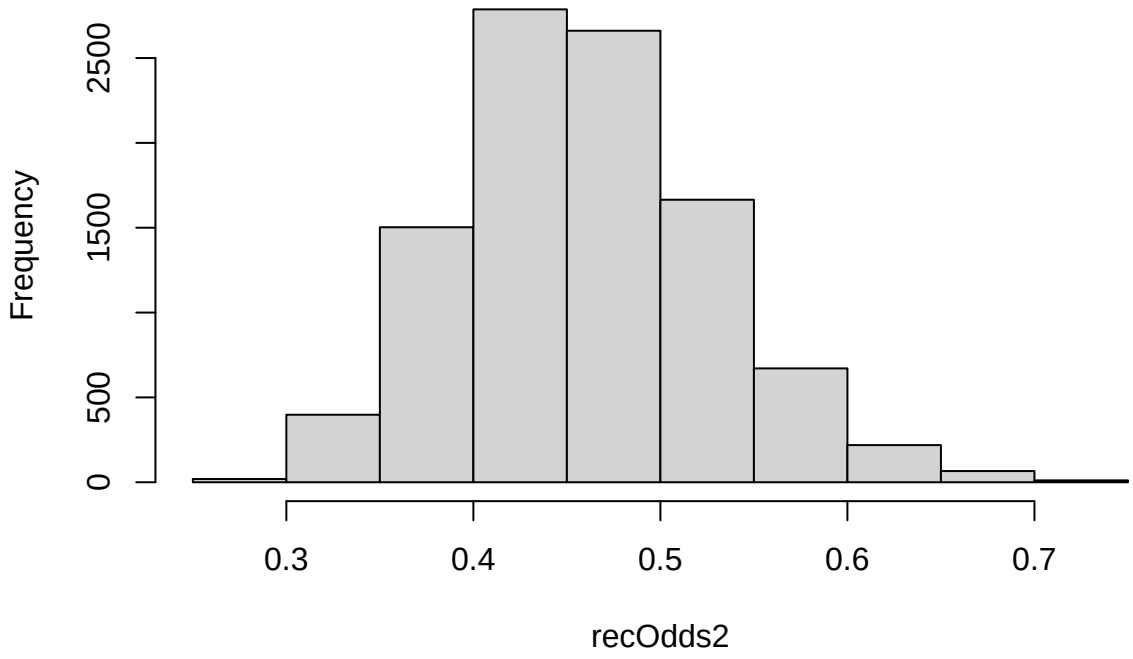
The above trace plot shows every iteration it has been taken to get the HDI region.

We can improve our view of the parameter estimates of the coefficient by converting the distribution from log odds to plain odds.

```
recLogOdds1 <- as.matrix(district_MCMC_out[,"PctFamilyPoverty"])
recOdds1 <- exp(recLogOdds1)
hist(recOdds1, main=NULL)
```

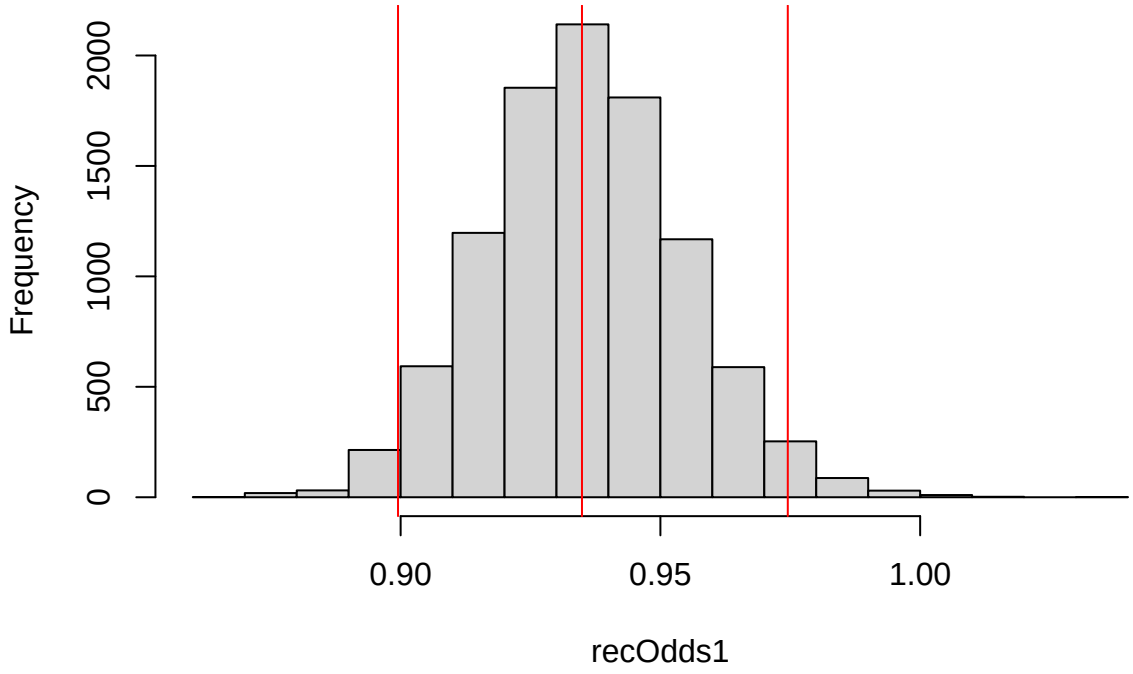



```
recLogOdds2 <- as.matrix(district_MCMC_out[,"TotalSchools_log"])
recOdds2 <- exp(recLogOdds2)
hist(recOdds2, main=NULL)
```

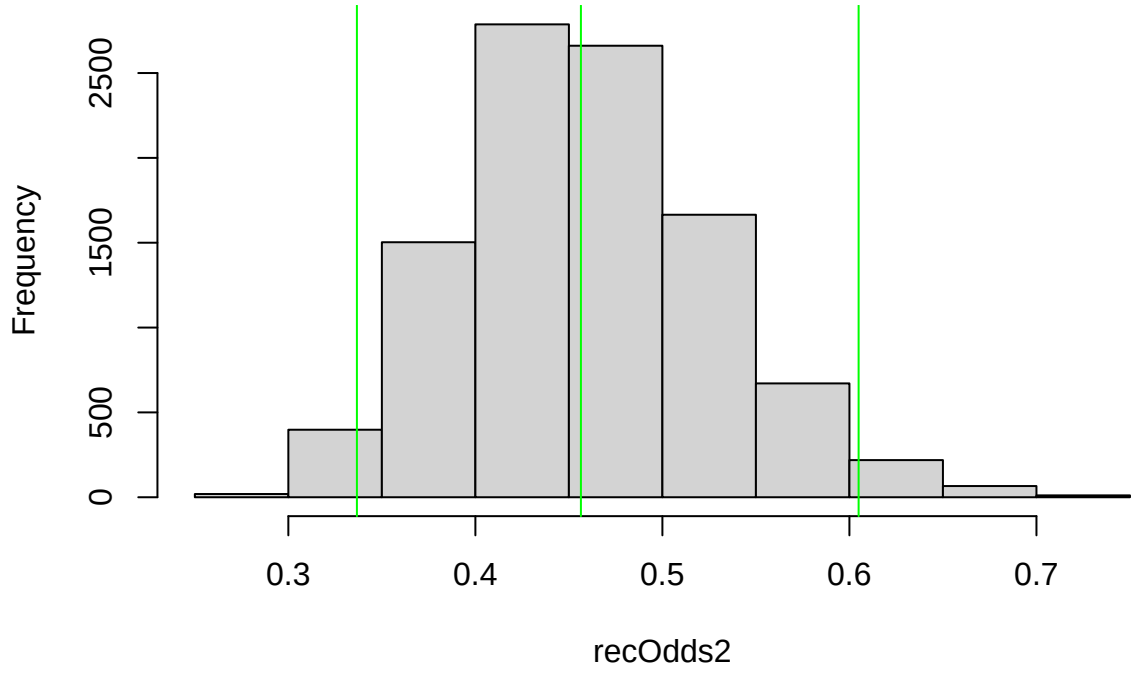


Marking HDI boundaries on above histograms: distributions of predictors obtained from Bayesian logistic regression analysis

```
hist(recOdds1, main=NULL)
abline(v=quantile(recOdds1,c(0.025, 0.5, 0.975)),col="red")
```




```
hist(recOdds2, main=NULL)
abline(v=quantile(recOdds2,c(0.025, 0.5, 0.975)),col="green")
```



_In the 1st plot for recOdds1, the extreme red lines represents the 95% HDI boundaries and the middle red line is for the median. Similarly, in the 2nd plot for recOdds2, the extreme green lines represents the 95% HDI boundaries and the middle green line is for the median.___

9Th Result

A logistic regression was performed on data to test how Percentage of children in district living below the poverty line(PctChildPoverty), Percentage of families in district living below the poverty line(PctFamilyPoverty), Total number of enrolled students in the district(Enrolled) and Total number of different schools in the district(TotalSchools) predict whether or not district's reporting was complete(DistrictComplete), dichotomized.

As explained earlier, the distributions were highly skewed for Enrolled and TotalSchools, so the data were log transformed for analysis, which generally improved the skew and the linearity of the relationship. The variables with percentage values were not transformed to not to hamper the data meaning and its overall effect on the dependent variable.

In addition to this, after getting normal odds of TotalSchools_log, we converted the log transformation back to TotalSchools to gather the exact normal odds and 95% CI of Total number of different schools in the district.

The results showed a significant association of both PctFamilyPoverty (b=-0.06760, Z(693)=-3.445, p<0.001) and TotalSchools_log (b=-0.76661, Z(693)=-5.343, p<0.01). In regular odds, each unit increase in Percentage of families in district living below the poverty line changes the odds of the district's reporting to be complete by 0.9346332; each unit increase in the TotalSchools increases the odds 1.591 times. The 95% confidence intervals were 0.899 to 0.972 for PctFamilyPoverty and 1.41414 to 1.840958 for TotalSchools.

The model showed poor performance with a Tjur's pseudo-R2 of 0.081 and but shows an accuracy of 94.54%.A Bayesian analysis reached similar conclusions, though the log odds of sampled coefficients differed slightly: a

mean of -0.06687 for PctFamilyPoverty with an HDI of -0.1059 to -0.02581 and -0.78787 for TotalSchool_log with an HDI of -1.0887 to -0.50268. Neither HDI contains 0, suggesting that these variables are predictive. In summary, the higher the Percentage of families in district living below the poverty line and Total number of different schools in the district, the greater the chance of the district's reporting being completed.

10. *Concluding Paragraph*

Describe your conclusions, based on all of the foregoing analyses. As well, the staff member in the state legislator's office is interested to know how to allocate financial assistance to school districts to improve both their vaccination rates and their reporting compliance. Make sure you have at least one sentence that makes a recommendation about improving vaccination rates. Make sure you have at least one sentence that makes a recommendation about improving reporting rates. Finally, say what further analyses might be helpful to answer these questions and any additional data you would like to have.

Conclusion and Recommendations:

In the above analysis we are trying to understand which factors are affecting to predict Percentage of all enrolled students with belief exceptions, percentage of students with completely up-to-date vaccines and whether or not district's reporting was complete.

Among all the regression analysis, Frequentist approach and Baysian approches are coinciding with each other. So we can conclude and suggest some recommendations with strong confidence.

The above analysis shows us that if there are students with one vaccine, there is high chance that students likely to have all of the others vaccines because of the high correlation within Tetanus, Polio, MMR and Hepatitis B. With this result in mind we can setup a system to get all the 4 vaccinations reporting done at all vaccination places.

Furthermore, we need to allocate financial resources by investing into charity drives or aid drives to the reduce the percentage of families in district living below the poverty line. Also we can direct some financial help to the schools in the districts by introducing free meals drives or free education drives about importance of having vaccines. This can be linked to a new system that after done with vaccinations incentives such as coupons, will be given to the helpers assisting government official in reporting. These strategies would help in getting reporting done where we have large number of students, greater area to cover and large number of schools.

Also we need to have some kind of interaction between total number of enrolled students in the district and percentage of children in district living below the poverty line. The interaction would help for more in depth understanding to allocate financial resources accurately to the districts and which would essentially help us to get more vaccinations and reporting done.

The drives for enrolled students mentioned above will also help to improve percentage of students with completely up-to-date vaccines. There is very less information available on percentage of all enrolled students with belief exceptions. But we can say from the correlation matrix that this is highly negatively correlated with percentage of students with completely up-to-date vaccines. We can study this thoroughly and provide some more understanding on improving the overall vaccination rates if more information is available to us. The availability of Hib3 and MCV1 vaccines rates for all Californian districts and the availability of MMR vaccine rates all over the US could have also shed some insights in understandning the into understanding and improving vaccination and reporting rates.