Fall 2021 - Final Examination

Dixitha Kasturi

Instructions

Your goal for this final exam is to conduct the necessary analyses of vaccination rates in California schools and 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 7 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, so you do not need to define those.

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 and say what you see); 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. Please place the answers in the text (not R comments) after the relevant analysis. 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 for the question you are answering. Please keep your answers concise and focused on the question asked. 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 (e.g., that the code does not run off the edge of the page).

You may not *receive assistance, help, coaching, guidance, or support from any human except your instructor at any point during this exam. Obtaining improper assistance will result in a 0 for 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 a personalized 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:

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 (NB. your sample may be slightly different). 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. Note that the data are about districts, not individual students, so be careful that you do not commit an ecological fallacy by stating conclusions about individuals.

In addition, you will find on Blackboard a CSV file, All Schools.csv, with data about 7,381 individual schools.

```
'data.frame' 7,381 obs. of 18 variables:
 $ SCHOOL CODE              : School ID number
 $ PUBLIC/ PRIVATE          : School status, "PUBLIC" or "PRIVATE" (note the space in the variable name
 $ Public School District ID: School district ID (only if public)
 $ PUBLIC SCHOOL DISTRICT   : School district name (only if public)
 $ CITY                     : City name
 $ COUNTY                   : Country name
 $ SCHOOL NAME              : School name
 $ ENROLLMENT               : Total number of enrolled students in the school
 $ UP_TO_DATE               : Number of students with completely up-to-date vaccines
 $ CONDITIONAL              : Number of students missing some vaccine without an exemption
 $ PME                      : Number of students with a medical exemption
 $ PBE_BETA                 : Number of students with a personal belief exemption
 $ DTP                      : Number of students in the district with the DTP vaccine
 $ POLIO                    : Number of students in the district with the Polio vaccine
 $ MMR                      : Number of students in the district with the MMR vaccine
 $ HEPB                     : Number of students in the district with Hepatitis B vaccine
 $ VARICELLA                : Number of students in the district with Varicella vaccine
 $ REPORTED                 : Whether the school reported vaccination data (Y or N)
```

Exploratory Data Analysis:

1. Loading Data:

```
setwd('C:/Users/kastu/Desktop/Syracuse/Fall21/IST_772/final exam')

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
schools <- read_csv('All Schools.csv')
```

```
## Rows: 7381 Columns: 18

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): PUBLIC/ PRIVATE, Public School District ID, PUBLIC SCHOOL DISTRICT...
## dbl (11): SCHOOL CODE, ENROLLMENT, UP_TO_DATE, CONDITIONAL, PME, PBE_BETA, D...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Data <- load('datasets14(2).RData')
districts <- districts
usVaccines <- usVaccines
#View(districts)
#View(schools)
```

2. Data understanding :

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
str(districts)
```

```
## 'data.frame':    700 obs. of  14 variables:
##  $ DistrictName    : Factor w/ 846 levels "ABC Unified",..: 808 276 731 515 611 752 255 61 496 190 .
##  $ WithDTP         : num  70 100 63 86 92 96 82 98 88 96 ...
##  $ WithPolio       : num  70 100 63 87 92 96 82 98 89 97 ...
##  $ WithMMR         : num  70 100 63 86 92 97 82 99 88 96 ...
##  $ WithHepB        : num  70 100 63 91 92 97 82 99 92 98 ...
##  $ PctUpToDate     : num  70 100 63 79 92 94 82 97 85 94 ...
##  $ DistrictComplete: logi  TRUE TRUE TRUE FALSE TRUE TRUE ...
##  $ PctBeliefExempt : num  29 0 37 4 8 0 18 0 6 0 ...
##  $ PctMedicalExempt: num  0 0 0 1 0 2 0 0 0 0 ...
##  $ PctChildPoverty : num  22 36 23 14 13 33 20 12 16 21 ...
##  $ PctFamilyPoverty: num  14 16 5 7 7 17 14 7 8 10 ...
##  $ PctFreeMeal     : num  40 76 39 44 NA 73 17 NA NA NA ...
##  $ Enrolled        : num  337 394 171 192 265 ...
##  $ TotalSchools    : num  2 3 1 5 6 4 1 10 21 11 ...
##  - attr(*, "na.action")= 'omit' Named int [1:67] 14 40 63 107 124 158 182 196 199 207 ...
##   ..- attr(*, "names")= chr [1:67] "14" "40" "63" "107" ...
```

```
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*       1 700 413.10  241.68  414.5  411.01 300.23   1   846   845
## WithDTP             2 700  89.58   11.17   93.0   91.64   7.41  23   100    77
## WithPolio           3 700  90.00   11.10   94.0   92.09   5.93  23   100    77
## WithMMR             4 700  89.58   11.45   94.0   91.71   5.93  23   100    77
## WithHepB            5 700  92.12    9.98   96.0   94.12   4.45  23   100    77
## PctUpToDate         6 700  88.20   14.89   92.0   89.99   7.41  23   191   168
## DistrictComplete    7 700    NaN      NA     NA     NaN     NA Inf  -Inf  -Inf
## PctBeliefExempt     8 700   5.72    8.80    3.0    3.76   4.45   0    77    77
## PctMedicalExempt    9 700   0.15    0.66    0.0    0.00   0.00   0     8     8
## PctChildPoverty    10 700  22.25   12.02   20.5   21.16  11.12   2    72    70
## PctFamilyPoverty   11 700  11.44    8.16    9.0   10.35   7.41   0    47    47
## PctFreeMeal        12 454  48.89   23.88   50.0   49.62  26.69   0   100   100
## Enrolled           13 700 604.98 2185.57  219.5  347.11 282.44  10 54238 54228
## TotalSchools       14 700   6.96   23.53    3.0    4.18   2.97   1   582   581
##                   skew kurtosis    se
## DistrictName*     0.05    -1.16  9.13
## WithDTP          -2.17     6.12  0.42
## WithPolio        -2.26     6.62  0.42
## WithMMR          -2.11     5.66  0.43
## WithHepB         -2.80    10.68  0.38
## PctUpToDate       0.32    13.71  0.56
## DistrictComplete    NA       NA    NA
## PctBeliefExempt   3.15    13.35  0.33
## PctMedicalExempt  6.79    56.72  0.02
## PctChildPoverty   0.83     0.54  0.45
```

```
## PctFamilyPoverty  1.29     1.80  0.31
## PctFreeMeal      -0.22    -0.86  1.12
## Enrolled         21.25   515.32 82.61
## TotalSchools     21.02   506.58  0.89
```

- For districts data, as seen by the statistics, most of the attributes/columns are heavily skewed as reported by the skewness and also are tailed as shown by the "kurtosis" values.
- Visualizing the skewness using violin plots and histograms(to see direction of outliers) :

```
library(ggplot2)
library(Hmisc)
```

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

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

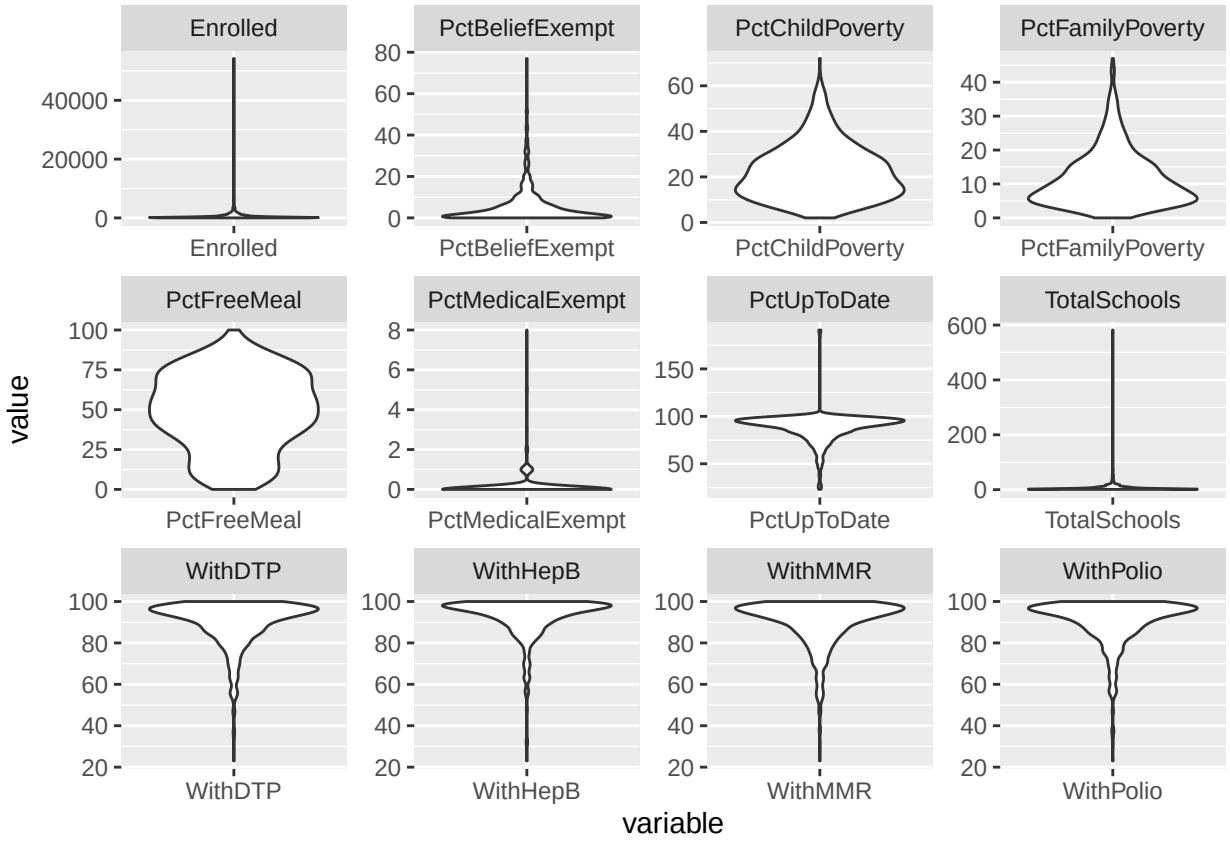
```
##
## Attaching package: 'Hmisc'
```

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

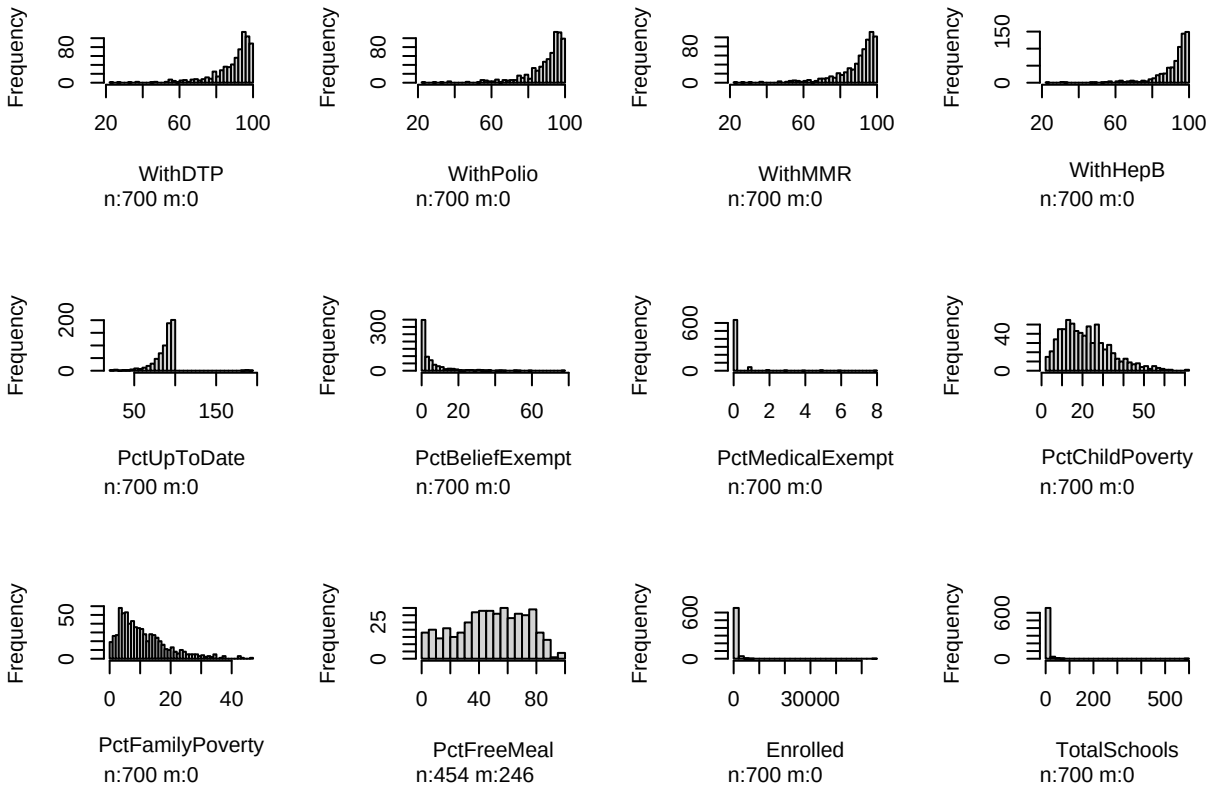
```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
#Violin Plots
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")
```




```
#Histograms
hist.data.frame(subset(districts,select = -c(DistrictName, DistrictComplete)))
```



- The violin plots show skewness of variables which can also be seen by the histograms. withDTP,withPolio, wihtMMR, withHepB, PctUpToDate are all heavily left skewed, meaning the have longer left tails. Percentage of children and Families in district living below the poverty line are right skewed.
- The two main observations from the violin plots and histograms are :
    1. Heavy skewness is observed in the Total number of enrolled students , Total number of different schools in the district, Percentage of all enrolled students with medical exceptions.
    2. The y-axis from violin plot of percentage of students with completely up-to-date vaccines goes beyond a value of 150. In percentages usually we take from a scale of 0 to 100, so there is an issue with this column. This can be observed from the outliers plot. The skewness(as seen in the violin plot) is due to the presence of outliers. The histogram for this column shows that the scale runs beyond 100 on x-axis. PctUpToDate column has to be looked into further.
- Now that we know that our data has skewness, we should look for outliers. We do so by plotting the data:

```
library(dlookr)
```

```
## Imported Arial Narrow fonts.
```

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

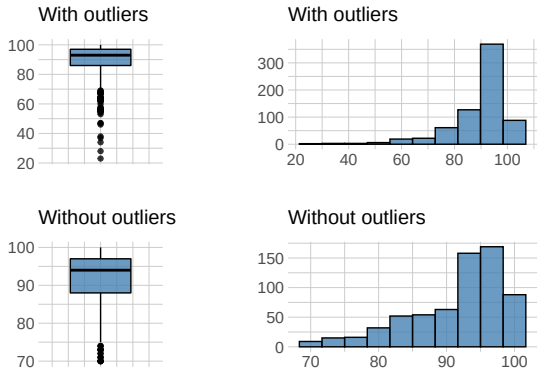
```
## The following object is masked from 'package:Hmisc':
##
##     describe

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

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

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

```
diagnose_outlier(districts)
```

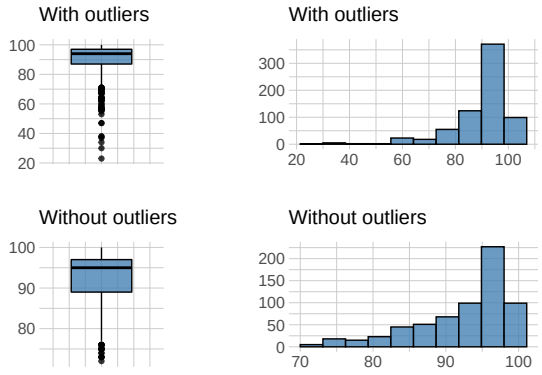
```
##           variables outliers_cnt outliers_ratio outliers_mean   with_mean
## 1           WithDTP           44       6.285714     57.181818  89.5771429
## 2         WithPolio           50       7.142857     59.060000  90.0014286
## 3           WithMMR           44       6.285714     56.590909  89.5800000
## 4          WithHepB           48       6.857143     62.854167  92.1242857
## 5       PctUpToDate           48       6.857143     62.041667  88.1985714
## 6   PctBeliefExempt           62       8.857143     28.612903   5.7242857
## 7  PctMedicalExempt           62       8.857143      1.709677   0.1514286
## 8   PctChildPoverty           14       2.000000     58.214286  22.2485714
## 9  PctFamilyPoverty           25       3.571429     36.040000  11.4371429
## 10      PctFreeMeal            0       0.000000           NaN  48.8898678
## 11         Enrolled           62       8.857143   3483.758065 604.9828571
## 12     TotalSchools           48       6.857143     43.479167   6.9628571
##    without_mean
## 1      91.75000
## 2      92.38154
## 3      91.79268
## 4      94.27914
## 5      90.12423
## 6       3.50000
## 7       0.00000
## 8      21.51458
## 9      10.52593
## 10     48.88987
## 11    325.22727
## 12      4.27454
```

```
plot_outlier(districts)
```

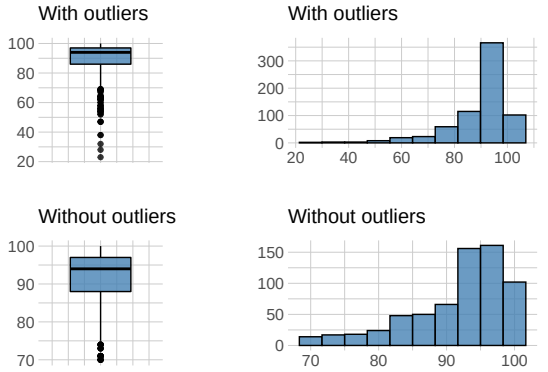
Outlier Diagnosis Plot (WithDTP)



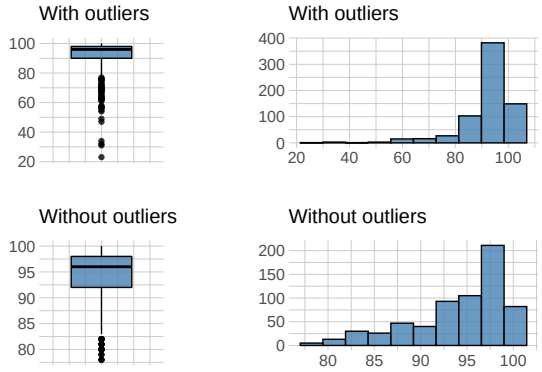
Outlier Diagnosis Plot (WithPolio)



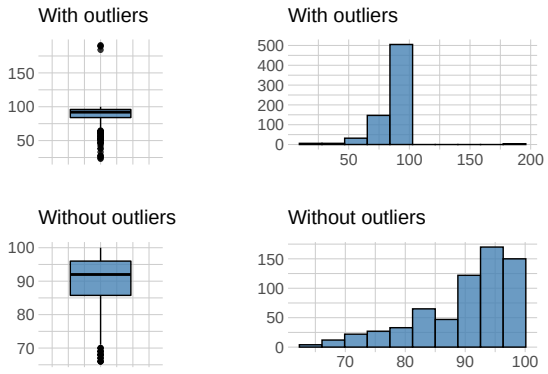
Outlier Diagnosis Plot (WithMMR)



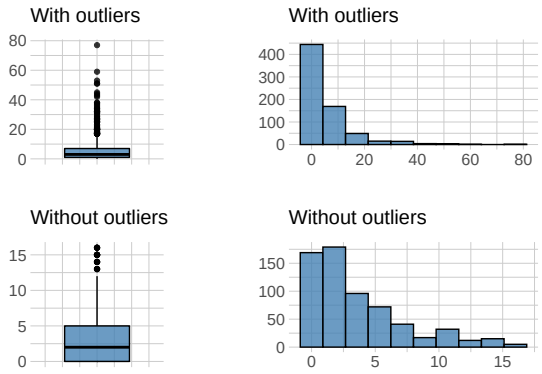
Outlier Diagnosis Plot (WithHepB)

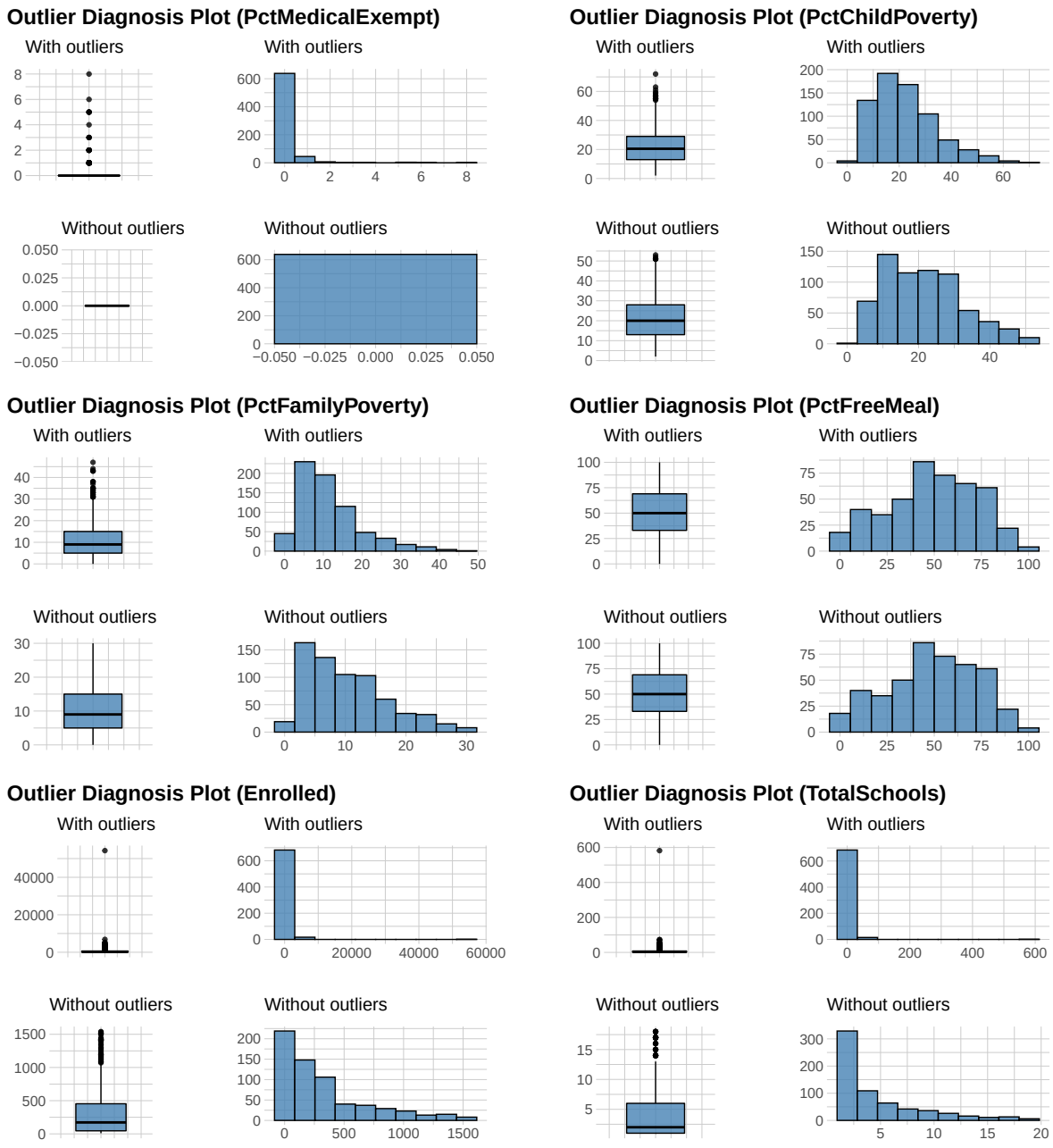


Outlier Diagnosis Plot (PctUpToDate)



Outlier Diagnosis Plot (PctBeliefExempt)

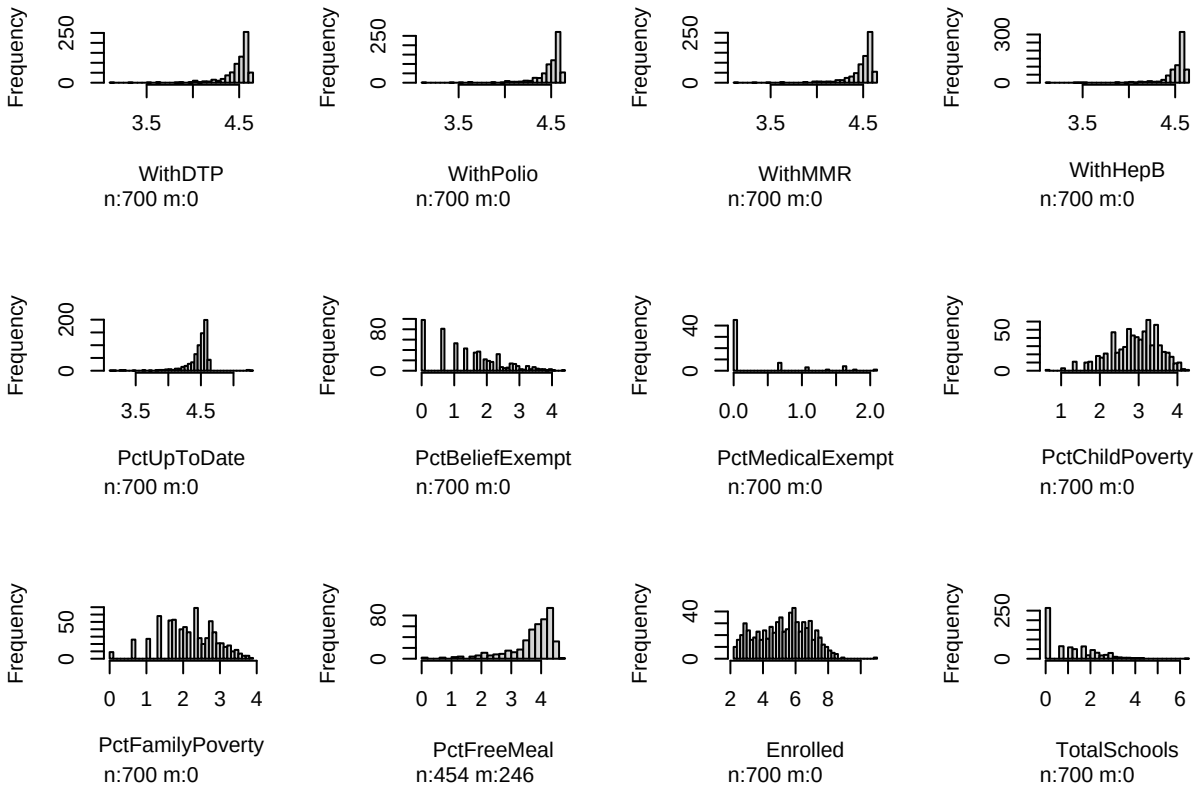




3. Data Cleaning:

- 1. As mentioned above, we look for the heavily skewed columns :

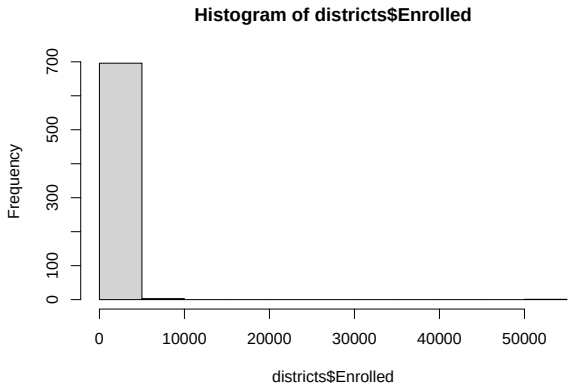
```
df <- subset(districts,select = -c(DistrictName, DistrictComplete))
df <- lapply(df[,1:12 ],log)
hist.data.frame(df)
```

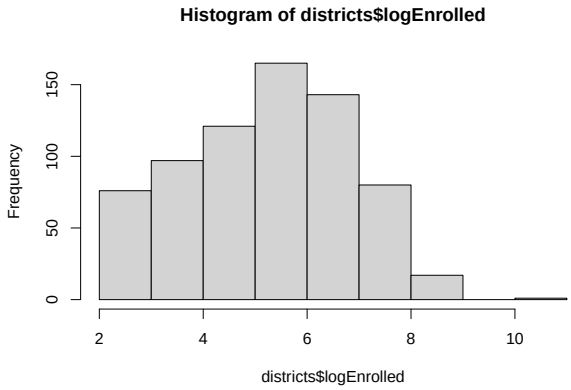


We apply log transformations on all columns except DistrictNames and DistrictComplete. We observe that the two columns that actually had some skewness removed are Enrolled, TotalSchools, MedicalExempt. So instead of taking log transformation on all columns we only take these.

a. Enrolled : Mean =604.98, number of outliers = 62. These outliers and skewness is understandable as the number of enrolled students varies per district, as some districts might have more or less schools which leads to the differences in number. We can apply log transformation to reduce the skewness and have the data be more normally distributed. The results of the transformation on the skewness can be seen through the histograms

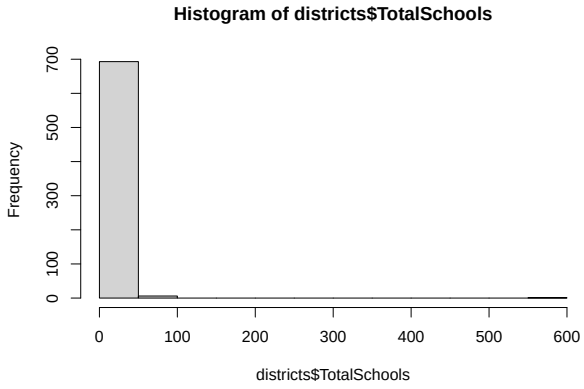
```
districts$logEnrolled <- log(districts$Enrolled)
hist(districts$Enrolled)
hist(districts$logEnrolled)
```

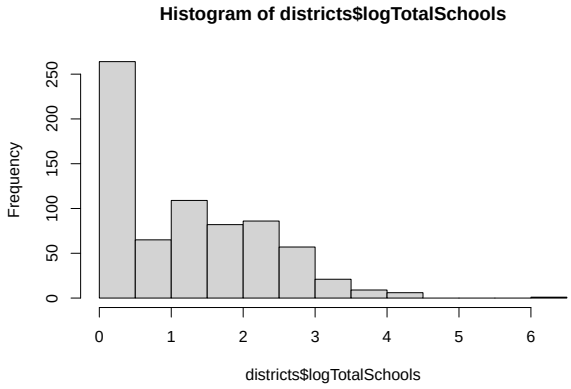




b. TotalSchools : Mean = 6.97, number of outliers = 48. Some districts which are bigger in size might have more schools but on an average there were 7 schools(as shown in the descriptive stats). So the outliers can be justified in away and donot have to be removed. We can apply log trasnformation to get rid of this skewnwss and view it through the hsitograms.

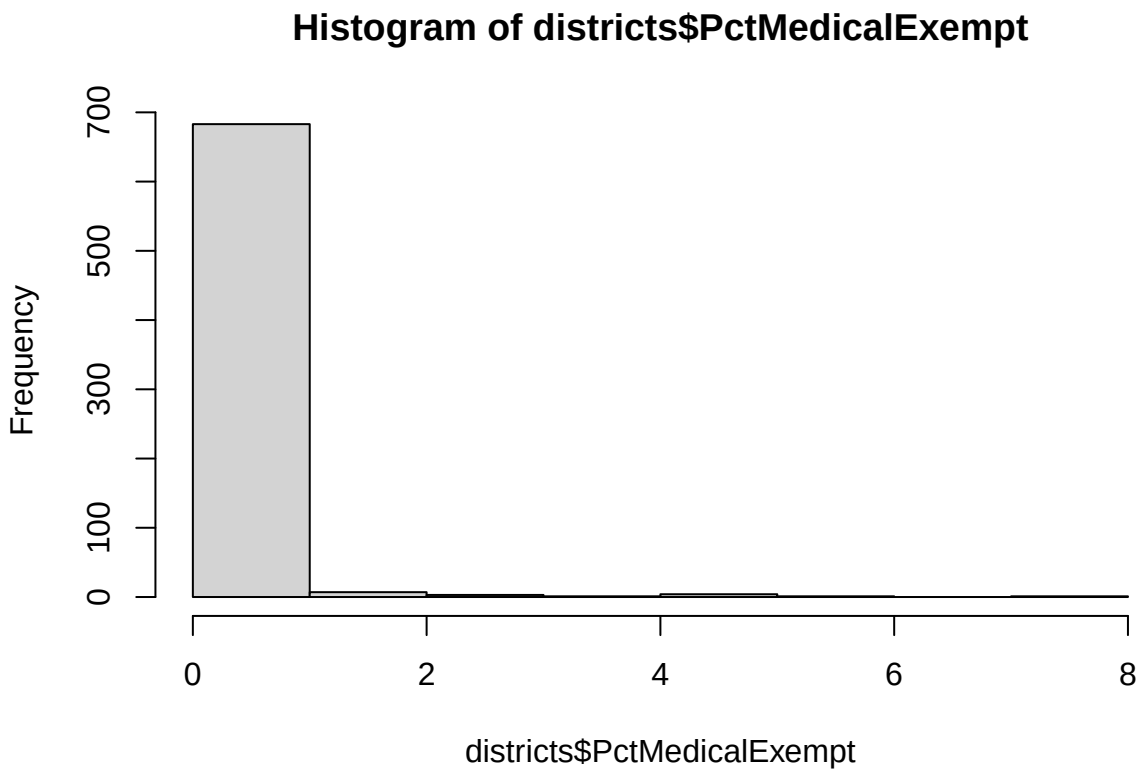
```
districts$logTotalSchools <- log(districts$TotalSchools)
hist(districts$TotalSchools)
hist(districts$logTotalSchools)
```





c. PctMedicalExempt: Mean = 0.1514286, number of outliers = 61. Some districts might have different conditions for medical exemptions/more number of schools leading to a higher number of medical exempts. The range of number of exempts is 0 to 8. The difference is not high, so we can let it be without using transformation.

```
hist(districts$PctMedicalExempt)
```



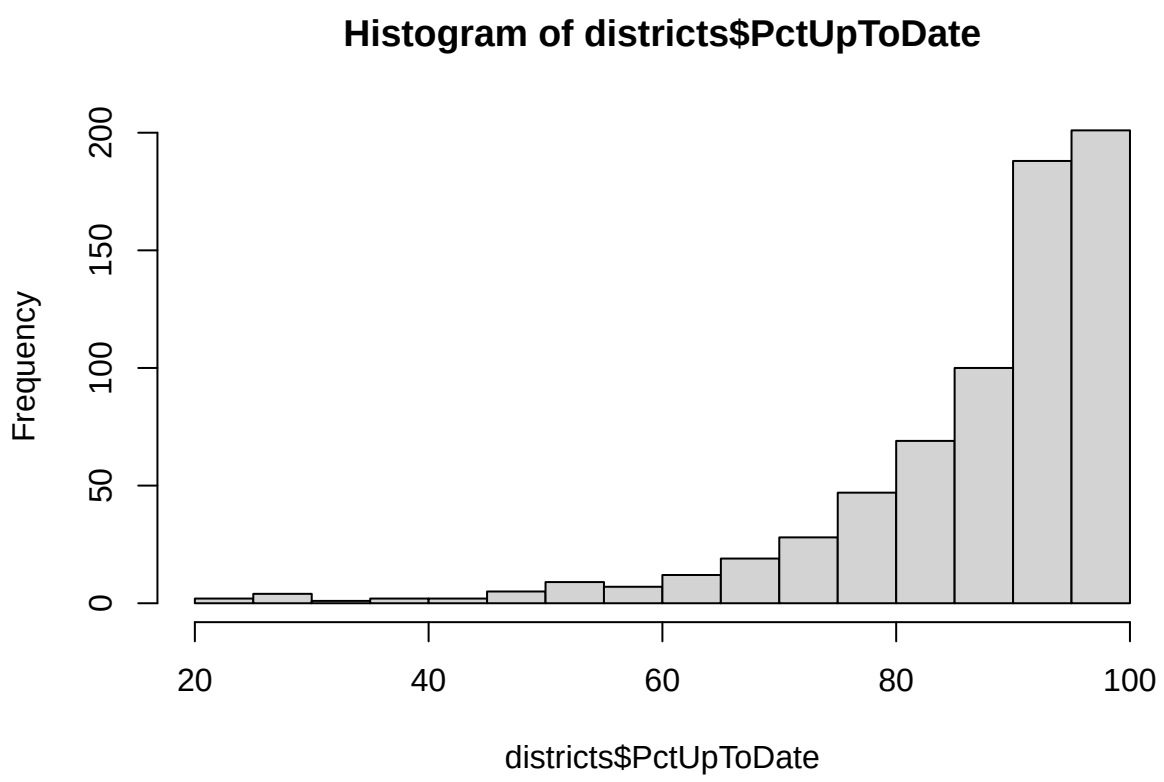

```
range(districts$PctMedicalExempt) # 0 to 8
```

```
## [1] 0 8
```

2. From the diagnostics we see that the number of outliers are atleast greater than 10 and in somecases even as high as 62. As our data only has 700 observations, getting rid of these outliers completely, will reduce out number of observations and a lot of data can be lost in this process. The only exception of outliers that we don't let go, is for the percentage of upto date vaccinations. The scale for this is >100 and the plot shows heavy skewness.

- To handle the outliers in the PctUpToDate column, we first check for values that are greater than 100. Essentially, these are considered as outliers in our case as percentage > 100 does not make sense.

```
# Counting number of uptodate values that are >100
count <- sum(districts$PctUpToDate > 100) # 4 rows.

# Removing those rows with percentage value >100
districts <- districts[-which(districts$ PctUpToDate >100),]
hist(districts$PctUpToDate)
```



- We only removed outliers for the PctUpToDate column and the plot now shows that there are no values which are >100, as the count was only and removing 4 observations would not alter our analysis much but the same does not hold true for other variable with higher number of outliers.

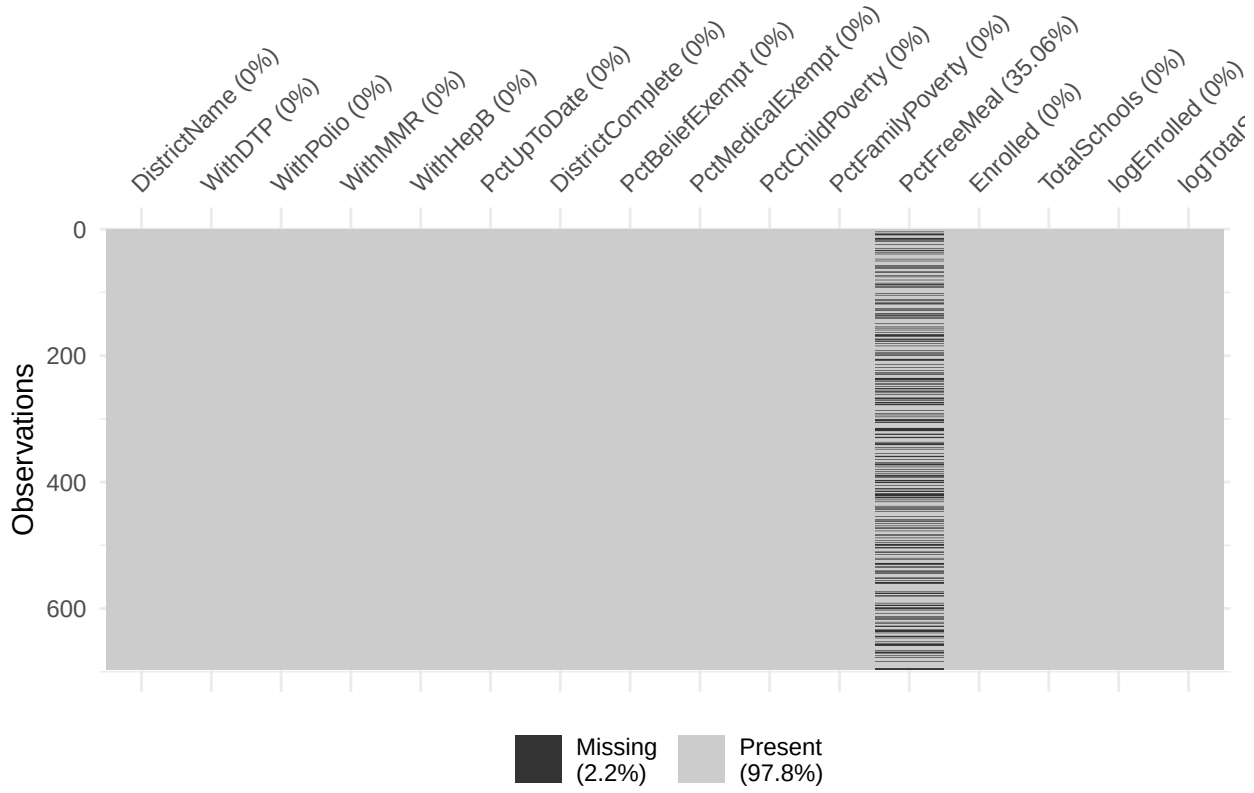
```
describe(districts)
```

```
## # A tibble: 14 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  89.6   1.12e+1  0.425   11     -2.17      6.15  23
##  2 WithPolio    696     0  90.0   1.11e+1  0.422   10     -2.26      6.66  23
##  3 WithMMR      696     0  89.6   1.15e+1  0.435   11     -2.12      5.68  23
##  4 WithHepB     696     0  92.1   1.00e+1  0.379    8     -2.80     10.7   23
##  5 PctUpToDa~   696     0  87.6   1.28e+1  0.486   12.2   -2.12      5.59  23
##  6 PctBelief~   696     0   5.74  8.81e+0  0.334    6      3.15     13.4    0
##  7 PctMedica~   696     0   0.151 6.56e-1  0.0249   0      6.82     57.3    0
##  8 PctChildP~   696     0  22.3   1.20e+1  0.456   16      0.833     0.554  2
##  9 PctFamily~   696     0  11.5   8.18e+0  0.310   10.2    1.30      1.82   0
## 10 PctFreeMe~   452   244  48.8   2.39e+1  1.13    36.2   -0.217    -0.852  0
## 11 Enrolled     696     0 604.    2.19e+3 83.1    597     21.3     518.    10
## 12 TotalScho~   696     0   6.96  2.36e+1  0.894    7     21.1     509.     1
## 13 logEnroll~   696     0   5.25  1.57e+0  0.0595   2.47  -0.0292   -0.757  2.30
## 14 logTotalS~   696     0   1.15  1.13e+0  0.0427   2.08   0.638    -0.257  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(visdat)
cbind(lapply(lapply(districts, is.na), sum))
```

```
##                  [,1]
## DistrictName     0
## WithDTP          0
## WithPolio        0
## WithMMR          0
## WithHepB         0
## PctUpToDate      0
## DistrictComplete 0
## PctBeliefExempt  0
## PctMedicalExempt 0
## PctChildPoverty  0
## PctFamilyPoverty 0
## PctFreeMeal      244
## Enrolled         0
## TotalSchools     0
## logEnrolled      0
## logTotalSchools  0
```

```
vis_miss(districts)
```



- From the visualization and counting, it is observed that the Percentage of students in the district

receiving free or reduced cost meals have 246 missing values, which account to 35.06% of the data in that column. So this column is problematic and should be set aside.

```
# Dataset after cleaning
districts_new <- subset(districts,select = -c(PctFreeMeal))
```

Descriptive Reporting

1. *Basic Introductory Paragraph*

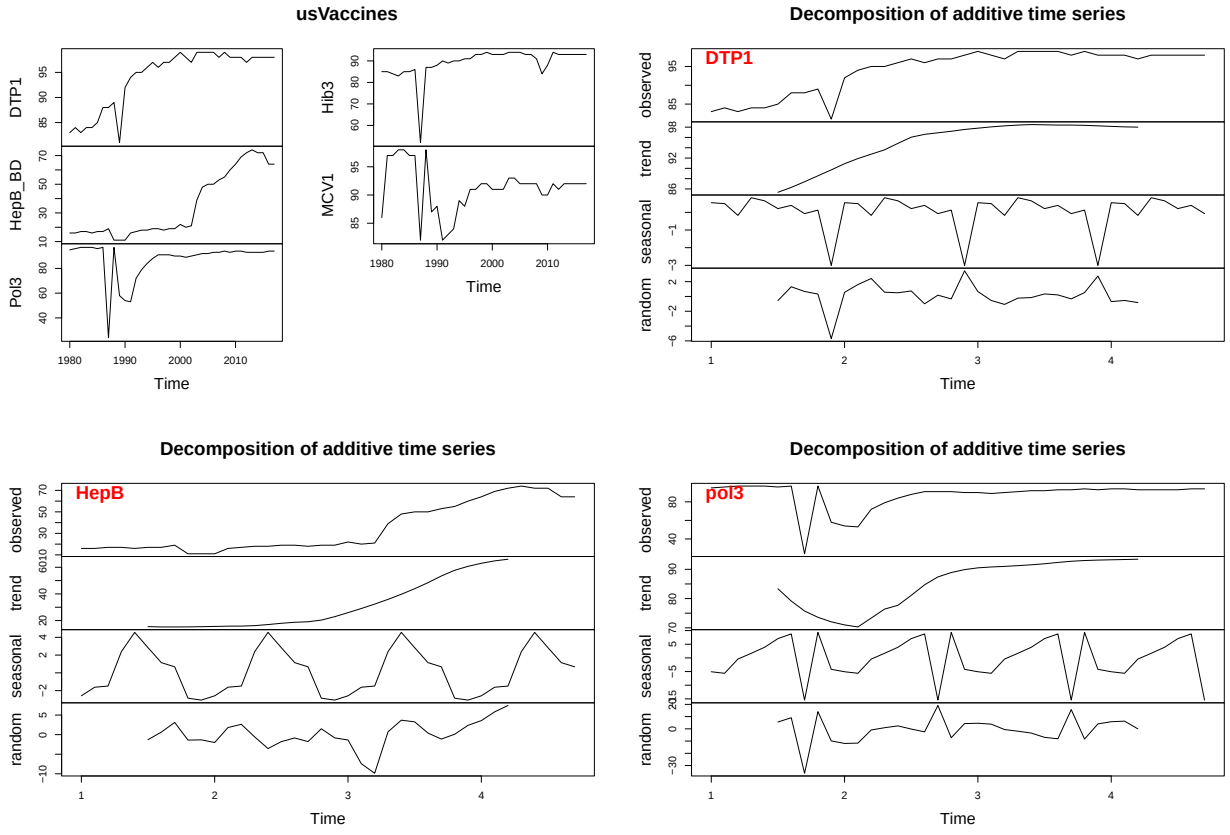
In your own words, write about three sentences of introduction addressing the staff member in the state legislator's office. Frame the problem/topic that your report addresses.

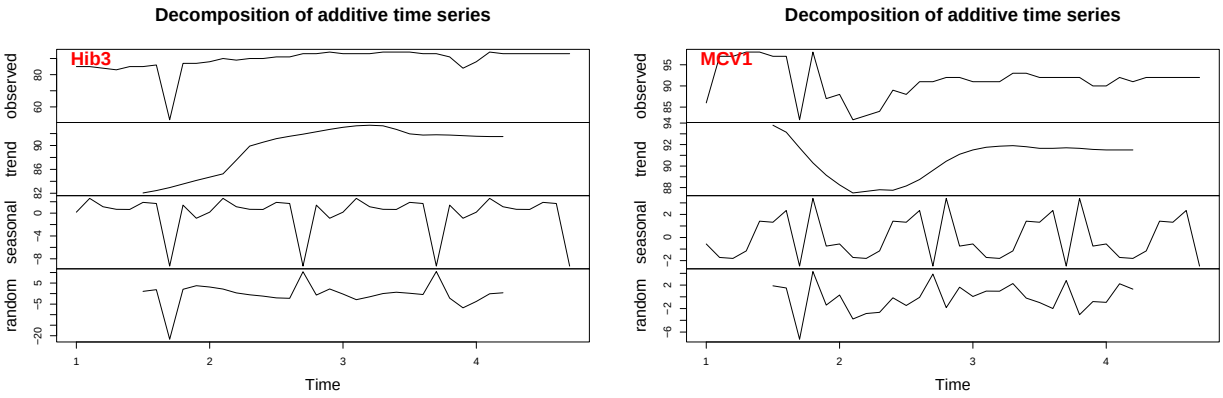
We are analyzing the vaccination rates of 4 vaccines : DPT,Polio,MMR and HepatitisB at schools in california respective to their districts and comparing their rates to the overall vaccination rate in the United States. The goal of our analysis is to find factors which are affecting the vaccination rates. We want to improve the rates in vaccinations, to do so understanding the factors that increase/decrease the rates, is important.

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.

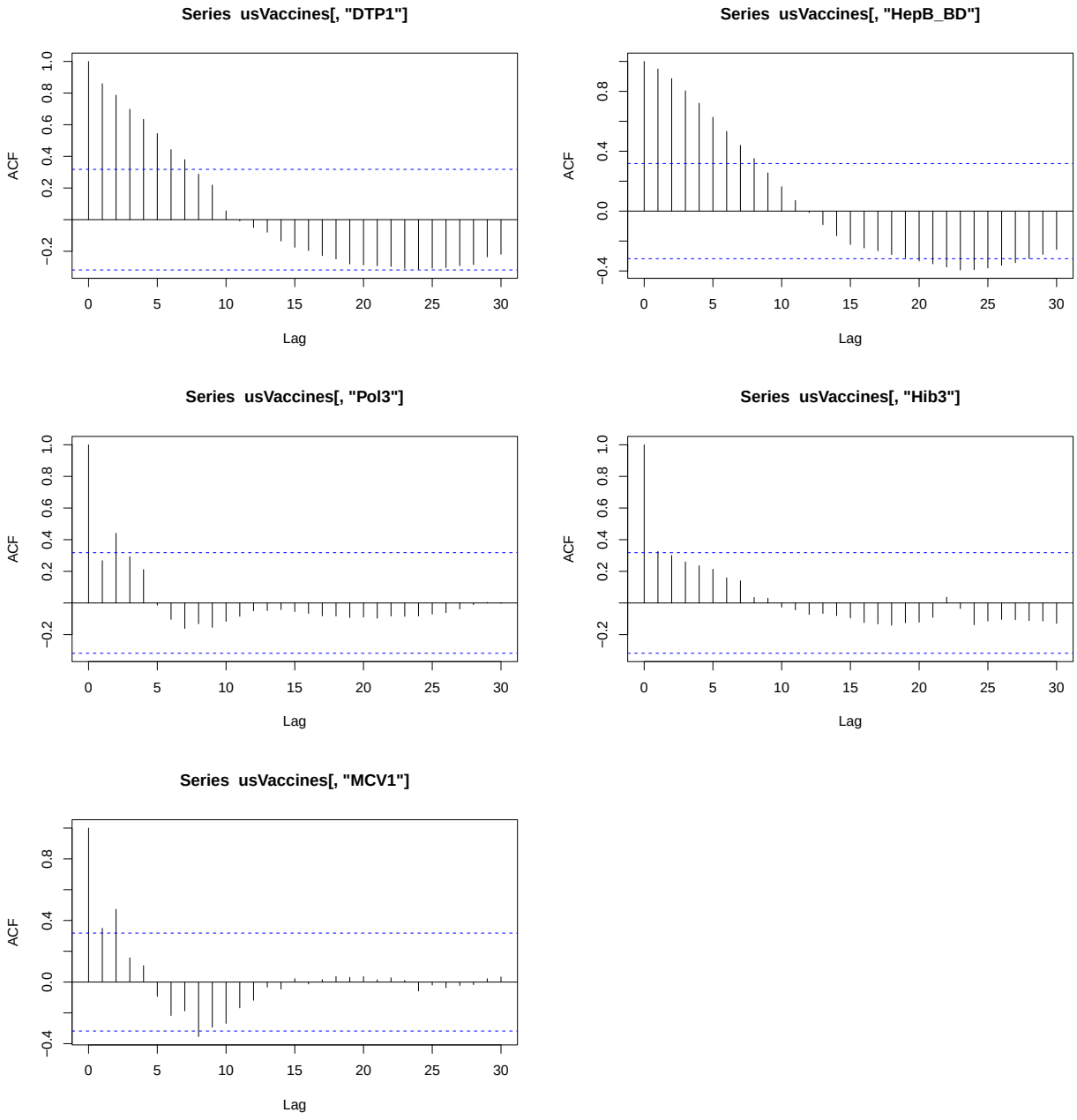
a. *How have U.S. vaccination rates varied over time?*





- For the first dose of vaccines, DTP1, HepB_BD and Hib3 had an upward trend overall which is justified by their decomposed individual components. It is noticed that in the 2nd half of 1980's or early 1990's, the rate for DTP1 and Hib3 increased.
- For HepB_BD, the rate of vaccinations took a steep rise in early 2000's. In september 1999 FDA approved a 2-dose schedule of hepatitis B vaccination for adolescents 11-15 years of age using Recombivax HB (Merck) with the 10 µg (adult) dose at 0 and 4-6 months later
- In may 2001,A combined hepatitis A inactivated and hepatitis B (recombinant) vaccine (Twinrix by SmithKline Beecham) was licensed.
- When we look at the dips/decreases,pol3,hib3,mcv1 had decreases in the 2nd half of 1980's. In 1989, recommendations for 2nd doses of measles-containing vaccine were issued by both ACIP and the AAP. During the mid- to late-1980s, a high proportion of reported measles cases were in school-aged children (5-19 years) who had been appropriately vaccinated before. This could mean that the previously given dose was wearing out or stopped working.
- For Polio, we see that there is a sharp dip from around 1986 and then there is a sharp rise. In 1988, a bill was passed to get rid of polio. In Dec 1990, a better poliovirus vaccine (Ipol by Pasteur Méérieux Vaccins et Serums) was licensed.

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



- When our timeseries is decomposed, we are nota 100% sure if that decomposition captured the true trends and seasonality. A timeseries that does not have trend and cyclic components is said to be stationary. We use the auto correlation function to check if these components were correctly separated.The ACF correlates a variable with itself at a later time period.
- In an acf plot, the horizontal lines indicate the significance levels for the series to be stationary. The correlations should be insignificant(below the line/threshold) for the process to be stationary. No patter should repeat(seasonality should not be there). The first correlation is ignored as for all it will be 1. But nothing concrete can be said.We will have to perform an inferential test about whether or not this is a stationary process by using the augmented Dickey–Fuller test, adf.test().

- DPT1, HepB_BD are significant as there are correlations that are above the threshold.The height of the bar pokes out above or below the horizontal dotted lines. there could be a pattern
- For Hib3,MCV1 and Pol3 the autocorrelations are insignificant(the values lie below the threshold/significance level)
- To further check for stationarity, we perform the adf test:

```
library(tseries)
```

```
## 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 test we see that none of the time-series are significant, the p-values for all of them are insignificant at a 0.05 level.So we reject the alternate hypothesis that they are stationary. But also, we fail to reject the null hypothesis that they are not stationary. In a way we see trends and cyclicality in the time-series through the plots. This test stands as substantial evidence for it.

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)?***

- Our time-series data doesn't explicitly state which row corresponds to which year. We add another column to identify. We have 38 rows. The year from which our data starts is 1980(inclusive). So our data runs from 1980 to 2017. We have to change the usVaccines to a dataframe to add another column.

```
usVaccines_df <- cbind(data.frame(usVaccines),year = seq(1980,2017, by =1))
summary(usVaccines_df[1:5])
```

```
##       DTP1          HepB_BD          Pol3            Hib3
##  Min.   :81.00   Min.   :11.00   Min.   :24.00   Min.   :52.00
##  1st Qu.:89.75   1st Qu.:17.00   1st Qu.:90.00   1st Qu.:87.00
##  Median :97.00   Median :19.00   Median :93.00   Median :91.00
##  Mean   :94.05   Mean   :34.21   Mean   :87.16   Mean   :89.21
##  3rd Qu.:98.00   3rd Qu.:54.50   3rd Qu.:94.00   3rd Qu.:93.00
##  Max.   :99.00   Max.   :74.00   Max.   :97.00   Max.   :94.00
##       MCV1
##  Min.   :82.00
##  1st Qu.:90.00
##  Median :92.00
##  Mean   :91.24
##  3rd Qu.:92.00
##  Max.   :98.00
```

- The mean values for overall 38 year period are, DTP1 :94.05, HepB_BD:34.21, Pol3:87.16, Hib3:89.21, MCV1:91.24
- We now, consider the recent years, we take a 10 year period for convenience. So recent years would be 2008 to 2017. Taking a 5 year period would be less, so i went with 10 years.

```
tail(usVaccines_df, 10)# 10 yearsof data
```

```
##    DTP1 HepB_BD Pol3 Hib3 MCV1 year
## 29   99      55   94   91   92 2008
## 30   98      60   93   84   90 2009
```

```
## 31   98      64   94   88   90 2010
## 32   98      69   94   94   92 2011
## 33   97      72   93   93   91 2012
## 34   98      74   93   93   92 2013
## 35   98      72   93   93   92 2014
## 36   98      72   93   93   92 2015
## 37   98      64   94   93   92 2016
## 38   98      64   94   93   92 2017
```

```
summary(usVaccines_df[32:36,1:5]) # for years 2011 to 2015
```

```
##       DTP1          HepB_BD          Pol3           Hib3           MCV1
##  Min.   :97.0   Min.   :69.0   Min.   :93.0   Min.   :93.0   Min.   :91.0
##  1st Qu.:98.0   1st Qu.:72.0   1st Qu.:93.0   1st Qu.:93.0   1st Qu.:92.0
##  Median :98.0   Median :72.0   Median :93.0   Median :93.0   Median :92.0
##  Mean   :97.8   Mean   :71.8   Mean   :93.2   Mean   :93.2   Mean   :91.8
##  3rd Qu.:98.0   3rd Qu.:72.0   3rd Qu.:93.0   3rd Qu.:93.0   3rd Qu.:92.0
##  Max.   :98.0   Max.   :74.0   Max.   :94.0   Max.   :94.0   Max.   :92.0
```

- For DTP1, Pol3,Hib3,MCV1 the values/rates have been preety much constant from 2011 to 2017. The values for HepB_BD are more or less constant from 2011 to 2015. So we go through this timeperiod and check the means for these timeperiods using the summary function.
- Mean values when only recent years are considered. DTP1:97.8, HepB_BD:71.8, Pol3:93.2,Hib3 :93.2, MCV1:91.8. The rates for Hepatitis B vaccinations were less for that timeframe.

```
usVaccines_recent <- usVaccines_df[32:36,]
```

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?*

```
mean(districts_new$WithDTP)
```

```
## [1] 89.56609
```

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

```
## [1] 92.10201
```

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

```
## [1] 89.99282
```

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

```
## [1] 89.56897
```

- By the way i understood the question, it is to calculate mean for the 4 vaccines overall.
- For DTP it is 89.56609, HepB is 92.10201,Polio is 89.99282 and MMR is 89.56897. Overall DTP,Polio and MMR had more or less equal rates whereas HepB had slightly higher rate.

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?***

```
cor(districts_new[,c("WithDTP","WithHepB","WithPolio","WithMMR")])
```

```
##             WithDTP  WithHepB WithPolio   WithMMR
## WithDTP   1.0000000 0.8896178 0.9815751 0.9775077
## WithHepB  0.8896178 1.0000000 0.9050617 0.8901386
## WithPolio 0.9815751 0.9050617 1.0000000 0.9663793
## WithMMR   0.9775077 0.8901386 0.9663793 1.0000000
```

- The values of correlations are high(close to 1 ,strongly correlated) and positive. This means that if a student has takes one vaccine, it is highly likely that he/she has also taken the other vaccines.

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

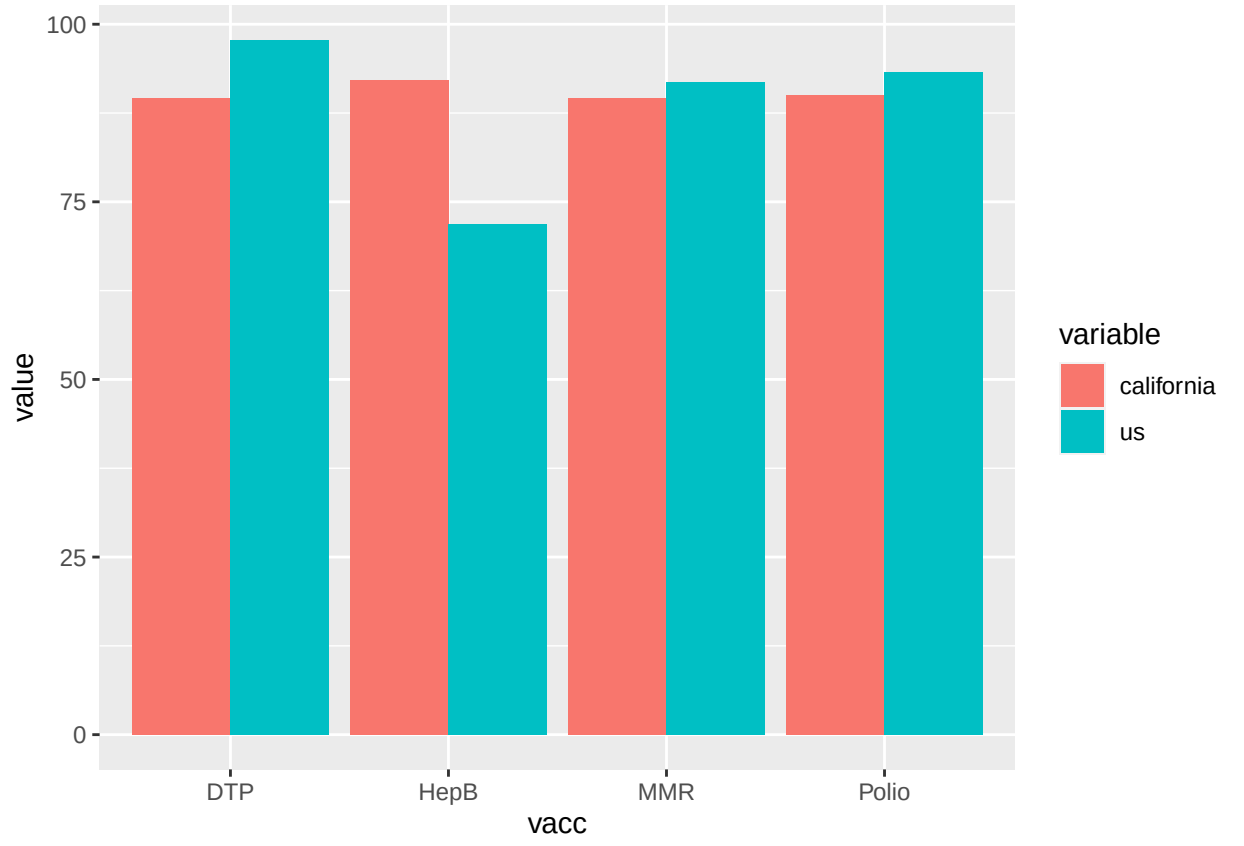
```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

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

```
california <- c(mean(districts_new$WithDTP),mean(districts_new$WithHepB),
                mean(districts_new$WithPolio),mean(districts_new$WithMMR))

us <- c(mean(usVaccines_recent$DTP1),mean(usVaccines_recent$HepB_BD),
        mean(usVaccines_recent$Pol3),mean(usVaccines_recent$MCV1))

vacc <- c("DTP","HepB","Polio","MMR")
Us_Cali <- us-california

df <- data.frame(california,us,vacc,Us_Cali)
df
```

```
##   california   us  vacc    Us_Cali
## 1   89.56609 97.8   DTP   8.233908
## 2   92.10201 71.8  HepB -20.302011
## 3   89.99282 93.2 Polio   3.207184
## 4   89.56897 91.8   MMR   2.231034
```

```
ggplot(melt(data.frame(california,us,vacc),id.vars = "vacc"),
       aes(x=vacc, y=value, fill=variable)) +
    geom_bar(stat='identity', position='dodge')
```



- We see that for DTP,MMR and Polio rates for USA for the recent years timeframe that was considered is higher than Californian rates. Where as for HepB the vaccination rates are higher compared to USA

4. *Comparison of public and private schools (i.e., from the All Schools data)*

a. *What proportion of public schools reported vaccination data?*

- I understood this question in 2 ways, first is taking proportion of actually reported data(excluding na's) and second is, overall proportion. So i did both

```
public <- schools[schools$`PUBLIC/ PRIVATE`== 'PUBLIC',]
describe(public)
```

```
## # 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 SCHOOL C~  5732     0 5.17e+6 2.09e+6 2.76e+4 44098.   -1.99      1.98 100016
##  2 ENROLLME~  5584   148 8.79e+1 4.02e+1 5.38e-1    51     0.906     4.34     10
##  3 UP_TO_DA~  5584   148 7.96e+1 3.87e+1 5.17e-1    50     0.713     1.83      2
##  4 CONDITIO~  5584   148 5.56e+0 9.11e+0 1.22e-1     7     3.99     25.3       0
##  5 PME        5584   148 1.55e-1 7.10e-1 9.50e-3     0    10.1     161.        0
##  6 PBE_BETA   5584   148 2.54e+0 5.65e+0 7.56e-2     3    10.1     165.        0
##  7 DTP        5584   148 8.13e+1 3.90e+1 5.21e-1    50     0.741     2.15      2
##  8 POLIO      5584   148 8.18e+1 3.91e+1 5.23e-1    50     0.722     2.01      2
##  9 MMR        5584   148 8.15e+1 3.90e+1 5.22e-1    50     0.719     2.00      2
## 10 HEPB       5584   148 8.35e+1 3.94e+1 5.27e-1    50     0.721     2.07      2
## 11 VARICELLA  5584   148 8.40e+1 3.95e+1 5.29e-1    51     0.710     1.96      2
## # ... 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>
```

```
proportion_no_na_public <- (nrow(public)-148)/nrow(schools)
proportion_no_na_public
```

```
## [1] 0.7565371
```

```
# With Na's
proportion_na_public <- nrow(public)/nrow(schools)
proportion_na_public
```

```
## [1] 0.7765885
```

If it is with Na's then 77.6% public schools reported vaccination rates If we get rid of NA values in vaccinations, then 75.65% public schools rpeorted vaccination rates.

b. ***What proportion of private schools reported vaccination data?***

Using the same methodology as above:

```
private <- schools[schools$`PUBLIC/ PRIVATE`== 'PRIVATE',]
describe(private)
```

```
## # 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 SCHOOL CODE  1649     0 6.84e+6 4.19e+5 1.03e+4 63790    -8.17   119.   113001
##  2 ENROLLMENT   1398   251 2.84e+1 1.70e+1 4.55e-1    18     2.08     7.13     10
##  3 UP_TO_DATE   1398   251 2.43e+1 1.54e+1 4.13e-1    17     2.03     7.54      0
##  4 CONDITIONAL  1398   251 2.42e+0 5.00e+0 1.34e-1     3     4.93    38.5       0
##  5 PME          1398   251 8.37e-2 4.00e-1 1.07e-2     0    10.5    188.        0
##  6 PBE_BETA     1398   251 1.61e+0 3.24e+0 8.67e-2     2     4.07    22.8       0
##  7 DTP          1398   251 2.52e+1 1.57e+1 4.20e-1    17     2.13     8.32      0
##  8 POLIO        1398   251 2.52e+1 1.57e+1 4.21e-1    17     2.14     8.42      0
##  9 MMR          1398   251 2.50e+1 1.57e+1 4.19e-1    17     2.10     8.03      0
```

```
## 10 HEPB         1398   251 2.62e+1 1.63e+1 4.36e-1    18     2.14     7.86      0
## 11 VARICELLA    1398   251 2.63e+1 1.63e+1 4.36e-1    18     2.13     7.72      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>
```

```
#without Na's
proportion_no_na_private <- (nrow(private)-251)/nrow(schools)
proportion_no_na_private
```

```
## [1] 0.1894052
```

```
# With Na's
proportion_na_private <- nrow(private)/nrow(schools)
proportion_na_private
```

```
## [1] 0.2234115
```

- If it is with Na's then 22.34% private schools reported vaccination rates
- If we get rid of NA values in vaccinations, then 18.94% private schools reported vaccination rates.

c. ***Was there any credible difference in reporting between public and private schools?***

```
tab <- table(schools$REPORTED,schools$`PUBLIC/ PRIVATE`)
tab
```

```
##
##     PRIVATE PUBLIC
##   N     252    148
##   Y    1397   5584
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 400.49, df = 1, p-value < 2.2e-16
```

- A significant p-value means that there is difference in reporting between public and private schools. It can also be justified from the ratio of reporting that was done.

d. ***Does the proportion of students with up-to-data vaccinations vary from county to county?***

```
anov <- aov(UP_TO_DATE ~ COUNTY, data= schools)
summary(anov)
```

```
##               Df   Sum Sq Mean Sq F value Pr(>F)
## COUNTY        57   960203   16846   10.47 <2e-16 ***
## Residuals   6924 11139890    1609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 399 observations deleted due to missingness
```

- The p-value is <0.05, so it is significant and tells that there might be a difference in vaccinations rate from county to county. Though thereis missing data we can still rely on these outputs.

5. *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.

- From put above analysis we saw that the Vaccination rates fir HepB in california are Higher than overall USA rates, but for other 3, they are more or less on par with the US. The mean level of vaccination rates for DTP,Polio and MMR are all above 95%. For private schools, reporting and vaccinations should be made better as their values/proportion is very less compared to the public schools.Vaccination rates differ from county to county and this could be because of the different number of schools in each county. There are mising values in the schools data, but we choose to go by what is available. Some districts reported 100% vaccinations. The overall rates had an upward trend.

6. *Inferential reporting about districts*

For every item below except question c, 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.

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

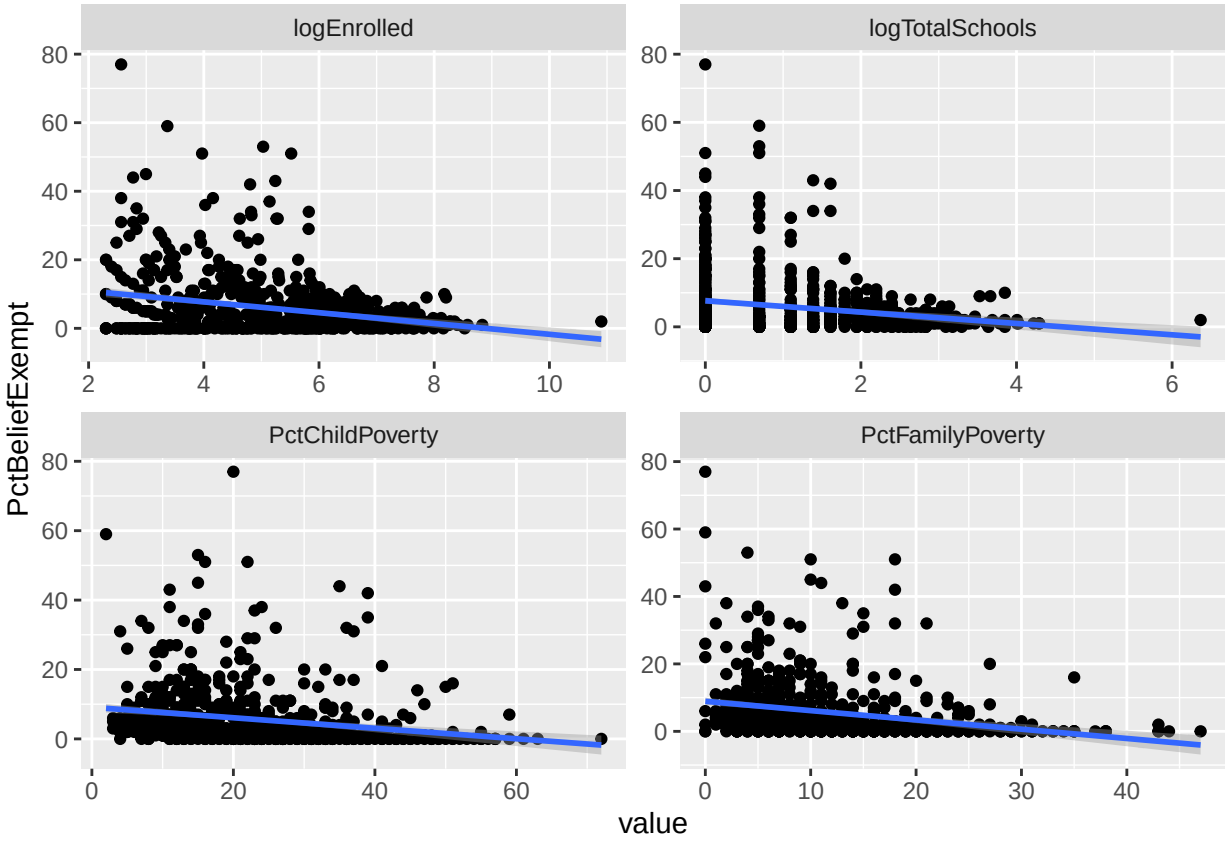
- Here our independent variables are PctChildPoverty, PctFamilyPoverty,logenrolled,logtotalschools(log columns because we tranformed our data in EDA step). The dependent variable is PctBeliefExempt.

```
districts_inference <- subset(districts_new,select = c(logEnrolled,
                                                       logTotalSchools,
                                                       PctChildPoverty,
                                                       PctFamilyPoverty,
                                                       PctBeliefExempt))
```

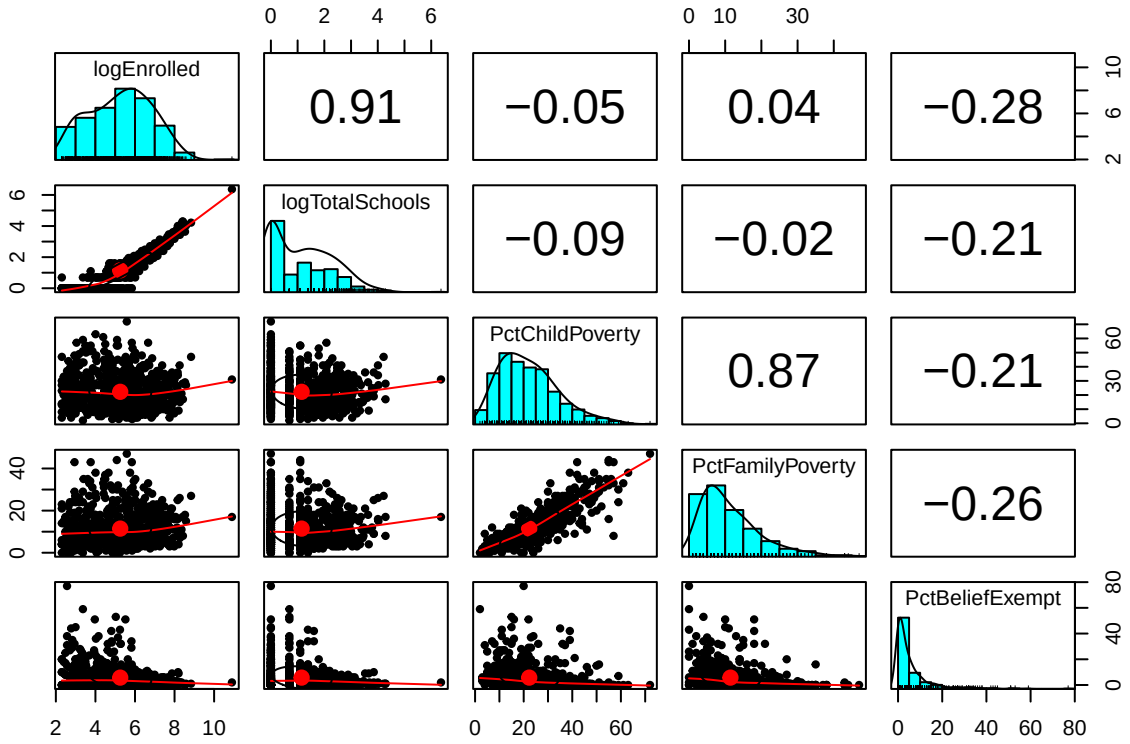
- We do bivariate exploration of data to understand te independent and dependent variable relationships better.
- a) In this analysis, we're looking for bivariate outliers and non-linear relationships. Plotting scatter plots to check for any patterns/problems

```
districts_inference %>% pivot_longer(-PctBeliefExempt,
                                     names_to="variable",
                                     values_to="value",
                                     values_drop_na = TRUE) %>%
             ggplot(aes(x=value, y=PctBeliefExempt)) +
  geom_point() + geom_smooth(method = "lm") + facet_wrap( ~ variable, scales="free")
```

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




```
pairs.panels(districts_inference)
```



- The above plots show that the data is more or less normally distributed with no issues.
- b) We now check for correlations between the variables considered.

```
districts_corr <- cor(districts_inference, use="pairwise.complete.obs")
signif(districts_corr)
```

```
##                 logEnrolled logTotalSchools PctChildPoverty PctFamilyPoverty
## logEnrolled       1.0000000       0.9139180      -0.0545198        0.0373644
## logTotalSchools   0.9139180       1.0000000      -0.0875150       -0.0231195
## PctChildPoverty  -0.0545198      -0.0875150       1.0000000        0.8688450
## PctFamilyPoverty  0.0373644      -0.0231195       0.8688450        1.0000000
## PctBeliefExempt  -0.2809900      -0.2128760      -0.2064430       -0.2554740
##                 PctBeliefExempt
## logEnrolled            -0.280990
## logTotalSchools        -0.212876
## PctChildPoverty        -0.206443
## PctFamilyPoverty       -0.255474
## PctBeliefExempt         1.000000
```

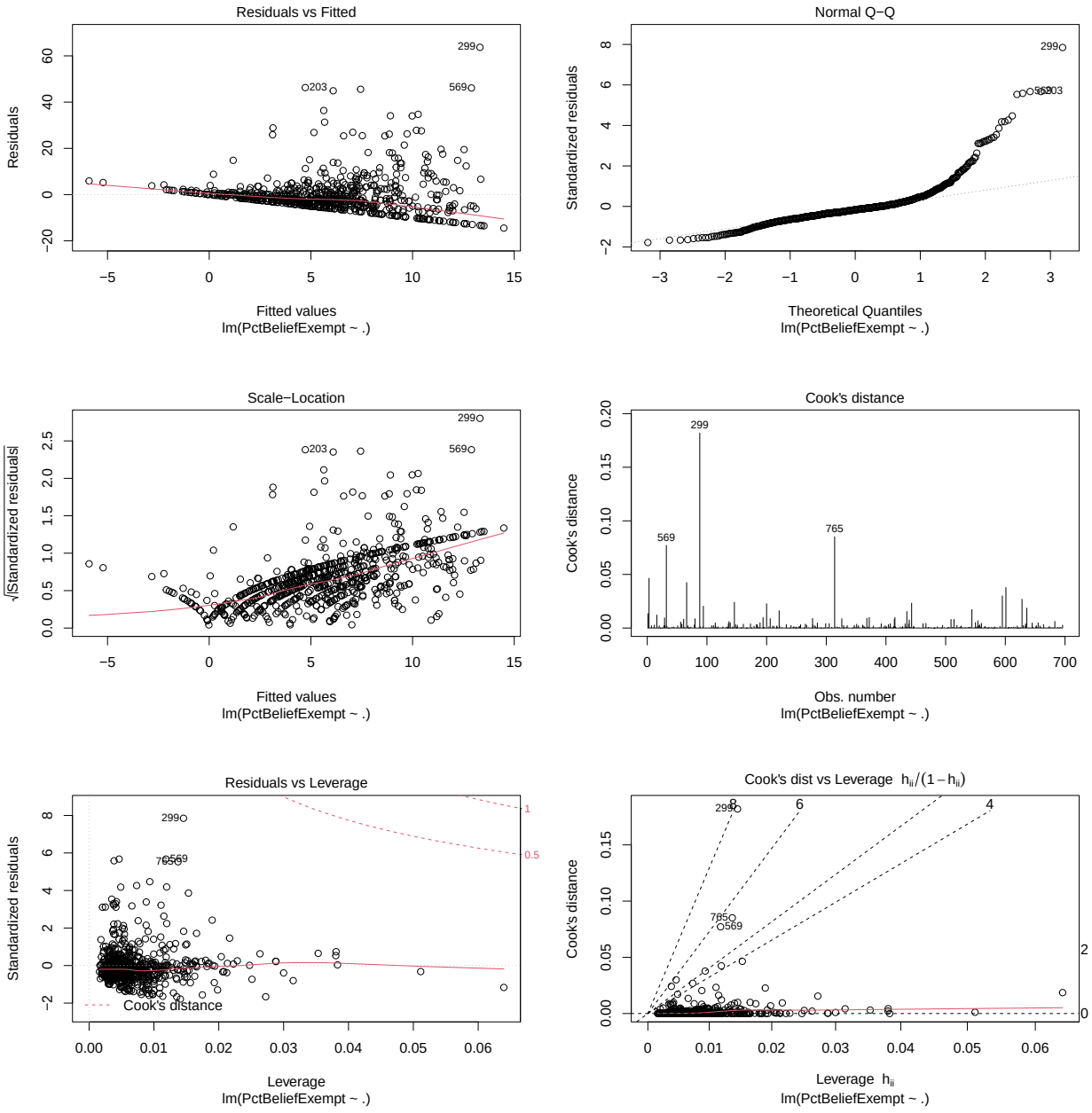
```
sort(districts_corr[,5])
```

```
##      logEnrolled PctFamilyPoverty  logTotalSchools  PctChildPoverty
##       -0.2809896       -0.2554742       -0.2128765       -0.2064428
##  PctBeliefExempt
##        1.0000000
```

- We see that the correlations with independent variables for the dependent variable are not that high, they are more closer to 0 than -1(because we are considering sign here). They are negatively correlated but not strong

```
#Linear Model

lm_belief_all <- lm(PctBeliefExempt ~.,data =districts_inference)
```

- a) First we check the residuals:



- Ideally, for the predictors to makeup to a good model, the residuals should not deviate a lot from the red line in resiudals vs fitted plot. The normal QQ plot

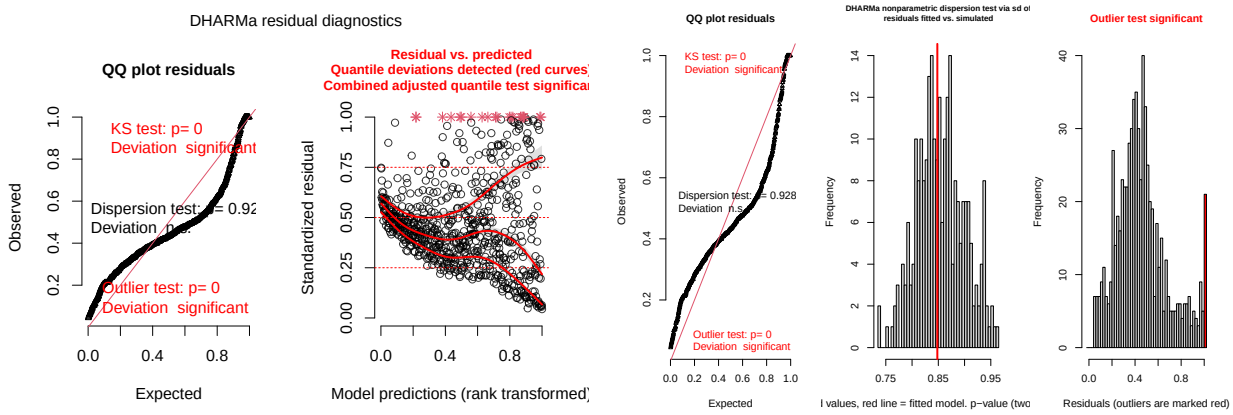
- 203,299,569 as they have been marked as outliers in the plots.We look into this data:

```
districts_new[c(203,299,569),]
```

```
##                  DistrictName WithDTP WithPolio WithMMR WithHepB PctUpToDate
## 244                Arvin Union      99        99      99      100          98
## 585 Petaluma Joint Union High      73        75      75       73          73
## 799         Cutten Elementary      88        88      98       98          88
##     DistrictComplete PctBeliefExempt PctMedicalExempt PctChildPoverty
## 244             TRUE               0                0              45
## 585             TRUE              25                0               9
## 799             TRUE               2                0              24
##     PctFamilyPoverty Enrolled TotalSchools logEnrolled logTotalSchools
## 244               35      411            3    6.018593        1.098612
## 585                5       52            1    3.951244        0.000000
## 799               11       95            1    4.553877        0.000000
```

- Arvin Union,Petaluma Joint Union High,Cutten Elementary are marked as outliers, when we observe the data, it they more or less are inline with the mean values in(WithDTP,WithPolio,WithMMR,WithHepB).???

```
## This is DHARMa 0.4.4. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa')
```



- DharMa simulations show that there is some deviation from the ideal red line in the qq plot. Ideally i would have transformed the data to remove skewness and outliers, but because we only have 700 obsrvations, i will let it be.
- b) Checking for multicollinearity which was hinted in the correlation values:

```
library(car)
```

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

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

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

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

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

```
vif(lm_belief_all)
```

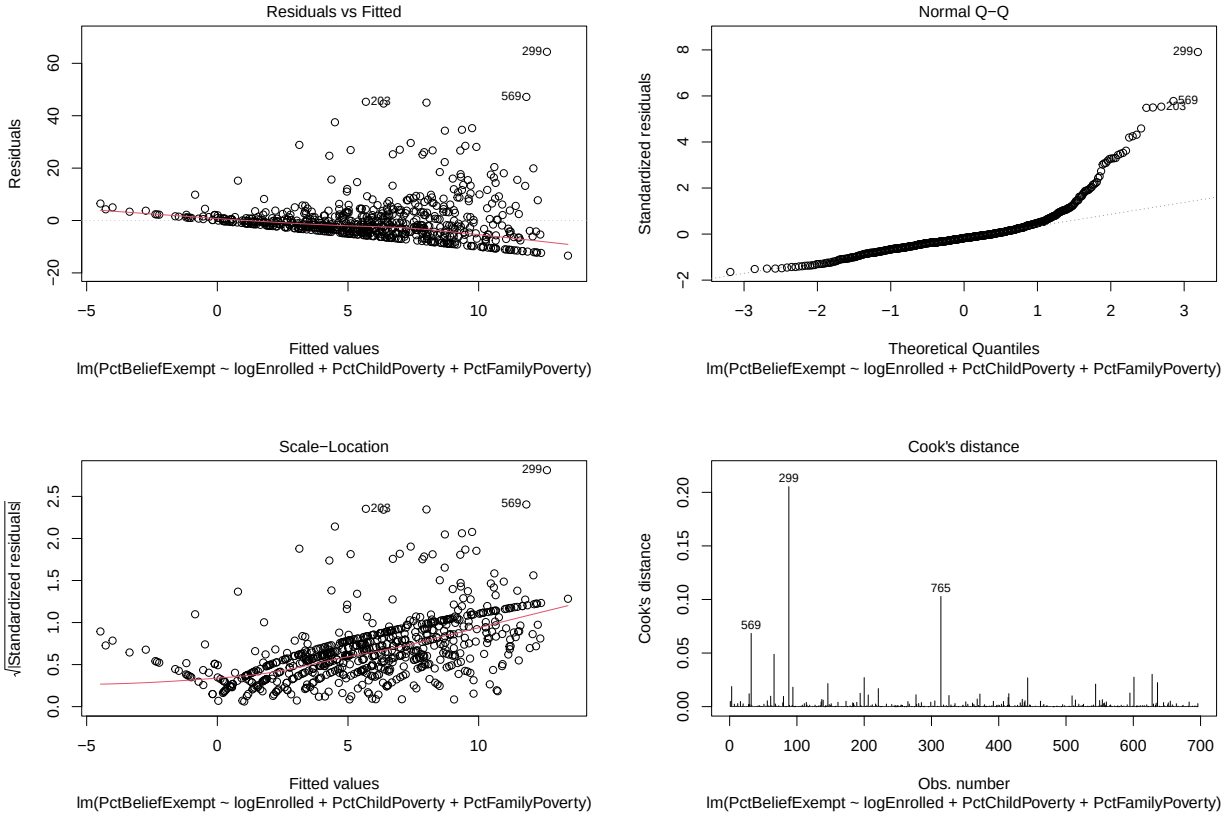
```
##      logEnrolled  logTotalSchools  PctChildPoverty PctFamilyPoverty
##         6.302123         6.217437         4.226525         4.268450
```

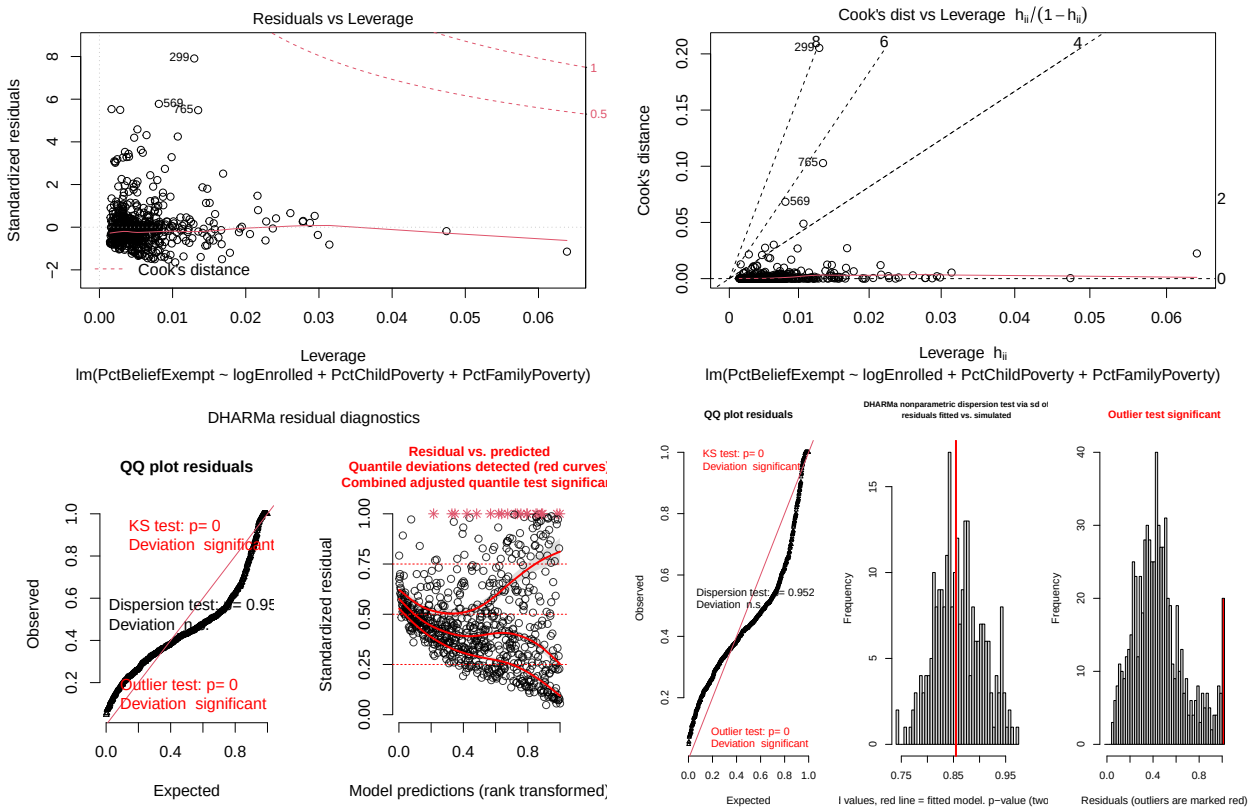
- Enrolled and Total Number of schools are highly correlated.We can get rid of either of them to check our values again:

```
lm_belief_2 <- lm(PctBeliefExempt ~ logEnrolled+PctChildPoverty+PctFamilyPoverty ,
                  data =districts_inference)
vif(lm_belief_2)
```

```
##      logEnrolled  PctChildPoverty PctFamilyPoverty
##         1.033341         4.209976         4.203331
```

- Removing Total Schools reduced the effect of multicollinearity.





- Not perfect, but its is okay. We move to the next steps

```
summary(lm_belief_2)
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ logEnrolled + PctChildPoverty +
##     PctFamilyPoverty, data = districts_inference)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.418  -4.184  -1.414   1.474  64.388
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      17.07023    1.29681  13.163  < 2e-16 ***
## logEnrolled      -1.54275    0.20126  -7.665 6.05e-14 ***
## PctChildPoverty  -0.02504    0.05301  -0.472  0.63680
## PctFamilyPoverty -0.23234    0.07795  -2.981  0.00298 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.195 on 692 degrees of freedom
## Multiple R-squared:  0.1393, Adjusted R-squared:  0.1356
## F-statistic: 37.34 on 3 and 692 DF,  p-value: < 2.2e-16
```

- A linear model was generated to predict the Pct of belief exempt using percentage of child poverty, percentage of family poverty and enrolled number.
- The null hypothesis is that R-squared value for population is 0. F(692,3) = 37.34, in favor of alternate hypothesis and p-value(2.2e-16) is <0.05. So our test is significant and we reject the null hypothesis. Had the null not been rejected then likelihood of observing a F-value value > 37.44 is less)
- The overall R-squared value is 0.1393. The adjusted R-squared is significant with a value is 0.1356, the 3 independent variables account to 13.56% of the data variability. The median of residuals is not around 0, this is because of the outliers that we didnot get rid of.

To see which predictors have the biggest impact, we can look at standardized coefficients, which are based on standardized variables, meaning that each gives the impact of 1 standarded deviation change in the predictor on the outcome variable

```
library(lm.beta)
summary(lm.beta(lm_belief_2))
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ logEnrolled + PctChildPoverty +
##     PctFamilyPoverty, data = districts_inference)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.418  -4.184  -1.414   1.474  64.388
##
## Coefficients:
##                  Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)      17.07023      0.00000    1.29681  13.163  < 2e-16 ***
## logEnrolled      -1.54275     -0.27480    0.20126  -7.665 6.05e-14 ***
## PctChildPoverty  -0.02504     -0.03418    0.05301  -0.472  0.63680
## PctFamilyPoverty -0.23234     -0.21551    0.07795  -2.981  0.00298 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.195 on 692 degrees of freedom
## Multiple R-squared:  0.1393, Adjusted R-squared:  0.1356
## F-statistic: 37.34 on 3 and 692 DF,  p-value: < 2.2e-16
```

- According to the coefficients : we cannot interpret PctChildPoverty as it is not significant.Whereas we the other two are significant. We reject the null hypothesis that the B-weights for PctFamilyPoverty and LogEnrolled are 0.
- To interpret the values of coefficients : Every unit increase in logenrolled, decreases the pctbeliefexempt by 1.54, whereas every unit increase in family poverty(if family poverty percentage rises by 1%), the percentage of be;ief exempt goes down by -0.23
- Performing Bayesian Analysis:

```
library(BayesFactor)
```

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

```
## 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.
## ************
```

```
belief_mcmc <- lmBF(PctBeliefExempt ~ logEnrolled + PctChildPoverty +
    PctFamilyPoverty,
    data=districts_inference,
    posterior=TRUE, iterations=10000)
summary(belief_mcmc)
```

```
##
## 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                5.74453 0.31453 0.0031453       0.003145
## logEnrolled      -1.50956 0.19966 0.0019966       0.001925
## PctChildPoverty  -0.02391 0.05240 0.0005240       0.000524
## PctFamilyPoverty -0.22837 0.07729 0.0007729       0.000784
## sig2             67.29292 3.64872 0.0364872       0.036417
## g                 0.13974 0.23774 0.0023774       0.002377
##
## 2. Quantiles for each variable:
##
##                     2.5%      25%      50%     75%    97.5%
## mu                5.13257  5.53285  5.74180  5.9573  6.37125
## logEnrolled      -1.89143 -1.64474 -1.51136 -1.3742 -1.11967
## PctChildPoverty  -0.12947 -0.05900 -0.02399  0.0112  0.07871
## PctFamilyPoverty -0.38229 -0.28091 -0.22797 -0.1762 -0.07689
## sig2             60.54504 64.73791 67.18913 69.6744 74.73721
## g                 0.02449  0.05171  0.08323  0.1476  0.58781
```

- We ran the Bayesian Linear regression using lmBF() function with posterior as true and 10000 iterations using the MCMC technique for sampling.
- In the first part, Mean column are the parameter estimates values for the coefficients of our independentvariables(PctChildPoverty,PctFamilyPoverty,LogEnrolled). For LogEnrolled it is -1.5118, for

PctFamilyPoverty it is -0.228 and for PctChildPoverty it is -0.023 which are very close to the values that we generated using the lm() function.

- In the 2nd part we see the 95% HDI interval values(2.5% and 97.5%) for each of the B-weights. The HDI interval values are the edges of the central region of the posterior distribution for each of the variable considered.For logEnrolled there is a 95% chance that the coefficient value/B-weight will lie between -1.906 and -1.11348. For PctFamilyPoverty the range is from -0.37879 to -0.0748, for PctChild-Poverty the range is from -0.12610 to 0.07890(this interval contains 0,tells us that PctChildPoverty is not a good predictor because mean value can be 0). As intervals for PctFamilyPoverty and logEnrolled donot contain 0, we can say that a model with these two variables variables as independent variables/predictors will be better than just the y-intercept. All of these findings run parallel with our findings from the frequentist method. Having PctChildPoverty insignificant.

- sig2 here gives the model precision for 10000 iterations.It gives the summary of the error in the model.R squared is (1-sig2)/variance of dependent variable. So to get bigger value of RSquared, the sig2value should be less.

```
library(BayesFactor)
belief_mcmc_bf <- lmBF(PctBeliefExempt ~ logEnrolled + PctChildPoverty +
    PctFamilyPoverty,
    data=districts_inference)
belief_mcmc_bf
```

```
## Bayes factor analysis
## --------------
## [1] logEnrolled + PctChildPoverty + PctFamilyPoverty : 1.658167e+19 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

- According to the model being used, the bayes factor is 1.658167e+19 ±0.01% which is a very high value. It shows very high and strong odds in favor of the alternative hypothesis that the model using PctFamilyPoverty, logEnrolled and PctChildPoverty as predictors/independent variables is highly favored over the model that only has the y-intercept. Though this stands as strong evidence, it doesnot give us information about which variable effects the outcome variable more.

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

- Here our independent variables are PctChildPoverty, PctFamilyPoverty,logenrolled,logtotalschools(log columns because we tranformed our data in EDA step). The dependent variable is PctUpToDate.

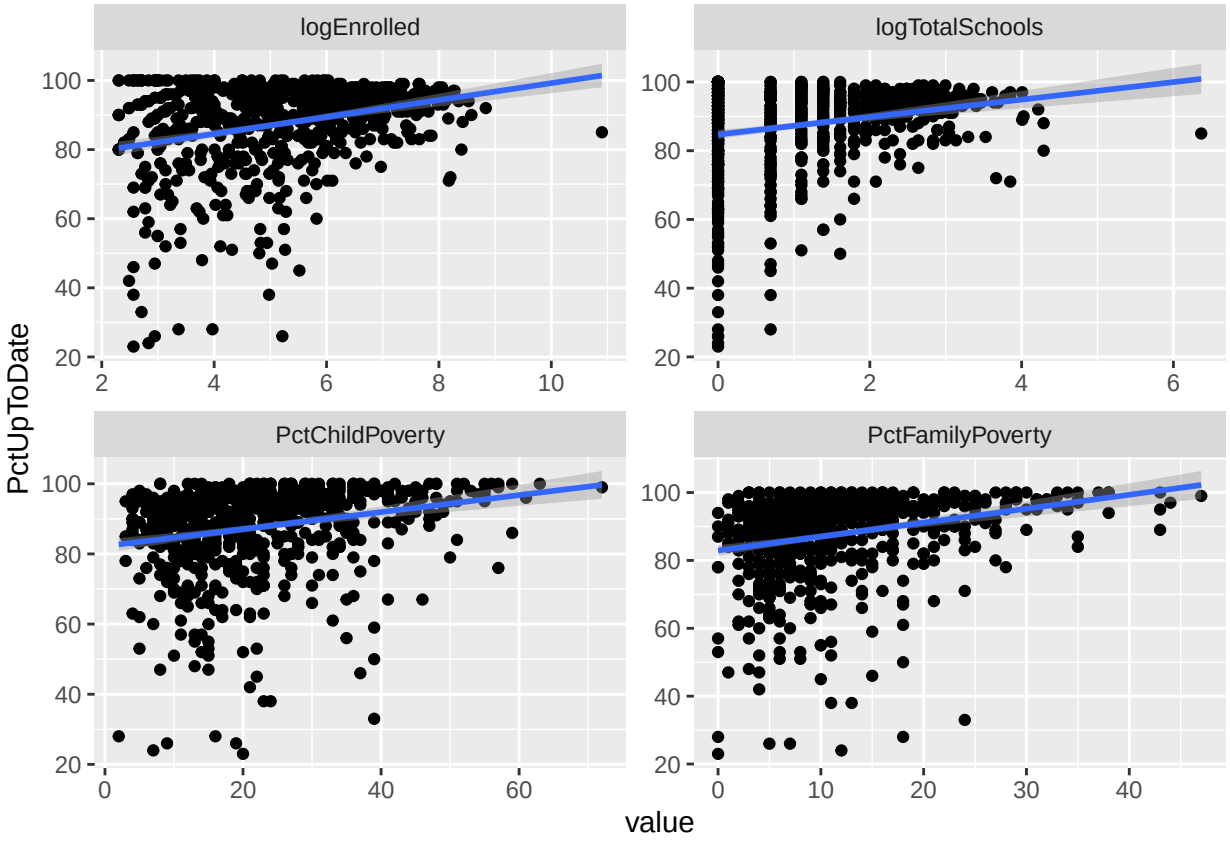
```
districts_inference_2 <- subset(districts_new,select = c(logEnrolled,
                                                         logTotalSchools,
                                                         PctChildPoverty,
                                                         PctFamilyPoverty,
                                                         PctUpToDate))
```

- We di bivariate exploration of data to understand te independent and dependent variable relationships better.

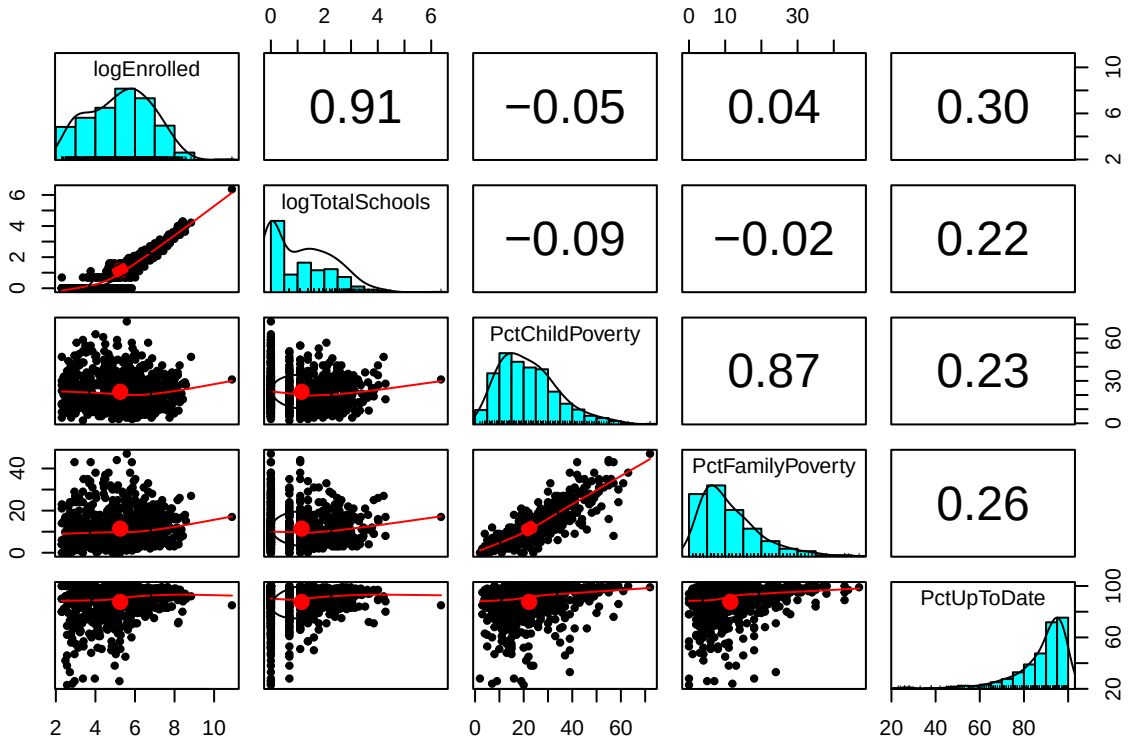
- a) In this analysis, we're looking for bivariate outliers and non-linear relationships. Plotting scatter plots to check for any patterns/problems

```
districts_inference_2 %>% 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'
```




```
pairs.panels(districts_inference_2)
```



- The above plots show that the data is more or less normally distributed with no issues.
- b) We now check for correlations between the variables considered.

```
districts_corr_2 <- cor(districts_inference_2, use="pairwise.complete.obs")
signif(districts_corr_2)
```

```
##                 logEnrolled logTotalSchools PctChildPoverty PctFamilyPoverty
## logEnrolled       1.0000000       0.9139180      -0.0545198        0.0373644
## logTotalSchools   0.9139180       1.0000000      -0.0875150       -0.0231195
## PctChildPoverty  -0.0545198      -0.0875150       1.0000000        0.8688450
## PctFamilyPoverty  0.0373644      -0.0231195       0.8688450        1.0000000
## PctUpToDate       0.2992440       0.2231860       0.2273960        0.2617610
##                 PctUpToDate
## logEnrolled        0.299244
## logTotalSchools    0.223186
## PctChildPoverty    0.227396
## PctFamilyPoverty   0.261761
## PctUpToDate        1.000000
```

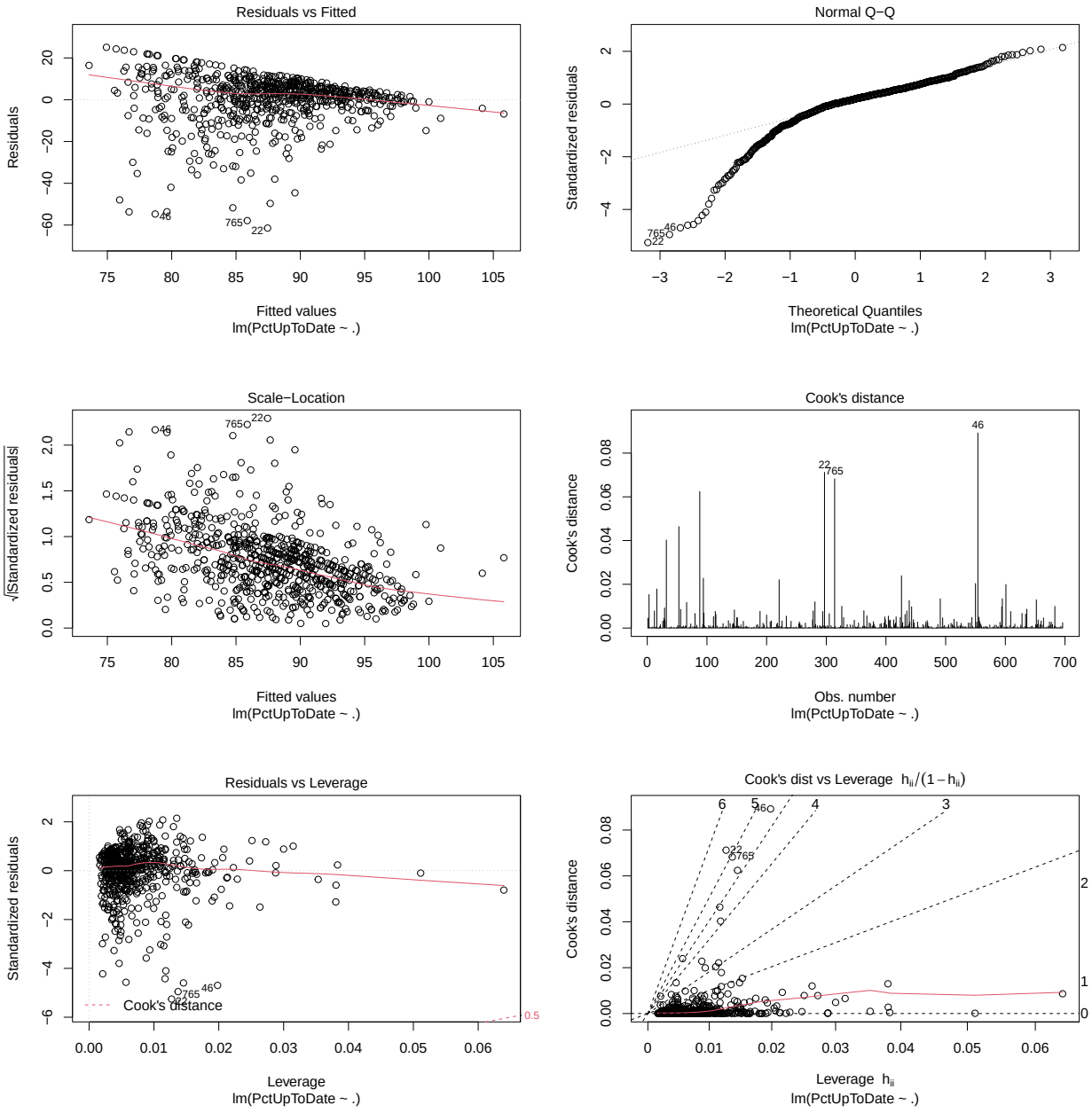
```
sort(districts_corr_2[,5])
```

```
##  logTotalSchools  PctChildPoverty PctFamilyPoverty      logEnrolled
##        0.2231861        0.2273957        0.2617606        0.2992443
##      PctUpToDate
##        1.0000000
```

- We see that the correlations with independent variables for the dependent variable are not that high, they are more closer to 0 than 1(because we are considering sign here). Strong correaltions between logEnrolled an LogToTalSchools and PctFamilyPoverty and PctUpToDate

```
#Linear Model

lm_uptodate_all <- lm(PctUpToDate ~.,data =districts_inference_2)
```

- a) First we check the residuals:



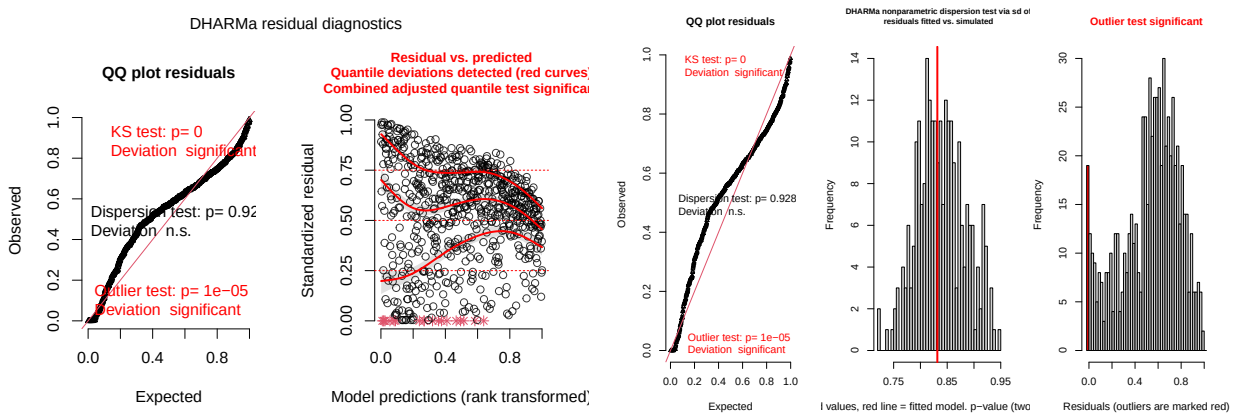
- Ideally, for the predictors to makeup to a good model, the residuals should not deviate a lot from the red line in residuals vs fitted plot.

- 46,22 as they have been marked as outliers in the plots.We look into this data:

```
districts_new[c(46,22),]
```

```
##             DistrictName WithDTP WithPolio WithMMR WithHepB PctUpToDate
## 143 Cucamonga Elementary      92        94      98       94          88
## 331       Carmel Unified      82        84      74       81          66
##     DistrictComplete PctBeliefExempt PctMedicalExempt PctChildPoverty
## 143             TRUE               1                1              23
## 331             TRUE              11                0              15
##     PctFamilyPoverty Enrolled TotalSchools logEnrolled logTotalSchools
## 143                8      358            3    5.880533        1.098612
## 331                4      175            3    5.164786        1.098612
```

- Cucamonga Elementary,Carmel Unified are marked as outliers, when we observe the data, it they more or less are inline with the mean values in(WithDTP,WithPolio,WithMMR,WithHepB). The PctChild-Poverty seems a little high



- DharMa simulations show that there is some deviation from the ideal red line in the qq plot. Ideally i would have transformed the data to remove skewness and outliers, but because we only have 700 obsrvations, i will let it be.
- b) Checking for multicollinearity which was hinted in the correlation values:

```
library(car)
vif(lm_uptodate_all)
```

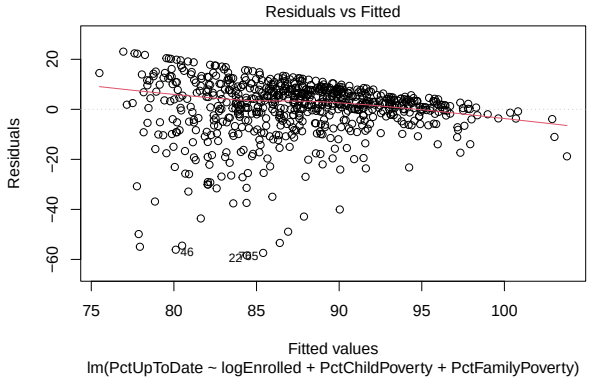
```
##      logEnrolled  logTotalSchools  PctChildPoverty PctFamilyPoverty
##         6.302123         6.217437         4.226525         4.268450
```

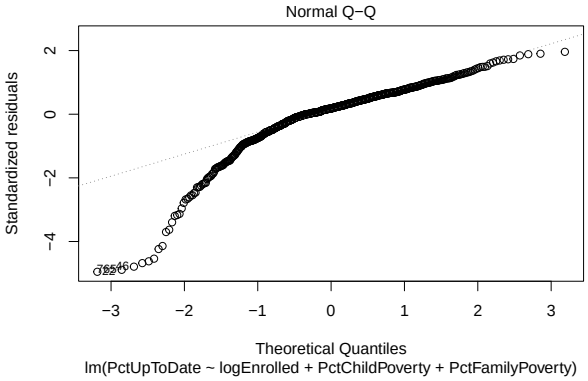
- Enrolled and Total Number of schools are highly correlated.We can get rid of either of them to check our values again:

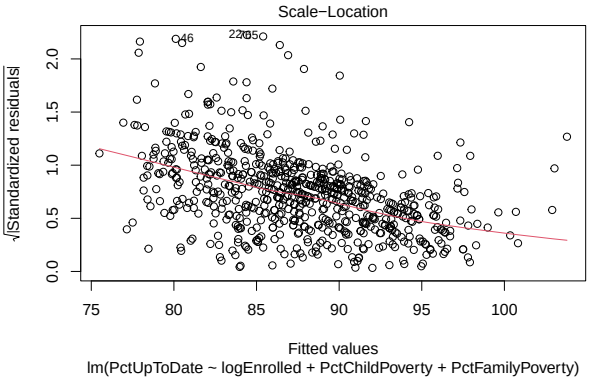
```
lm_uptodate_2 <- lm(PctUpToDate ~ logEnrolled+PctChildPoverty+PctFamilyPoverty ,
                    data =districts_inference_2)
vif(lm_uptodate_2)
```

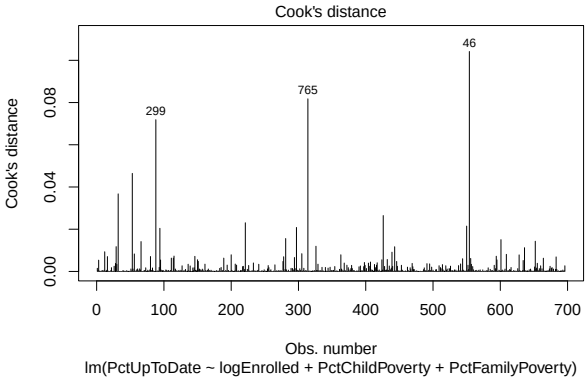
```
##      logEnrolled  PctChildPoverty PctFamilyPoverty
##         1.033341         4.209976         4.203331
```

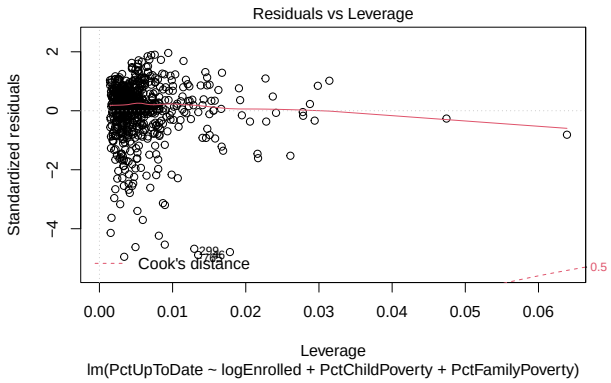
- Removing Total Schools reduced the effect of multicollinearity.

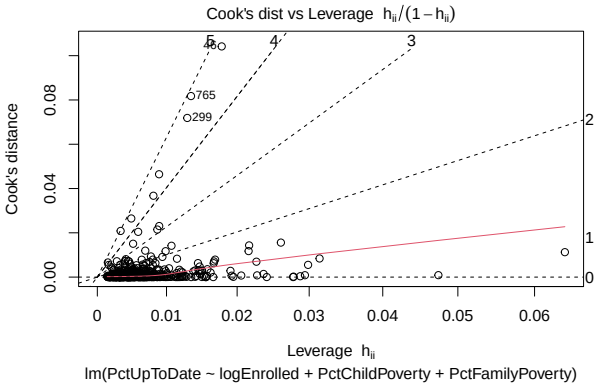


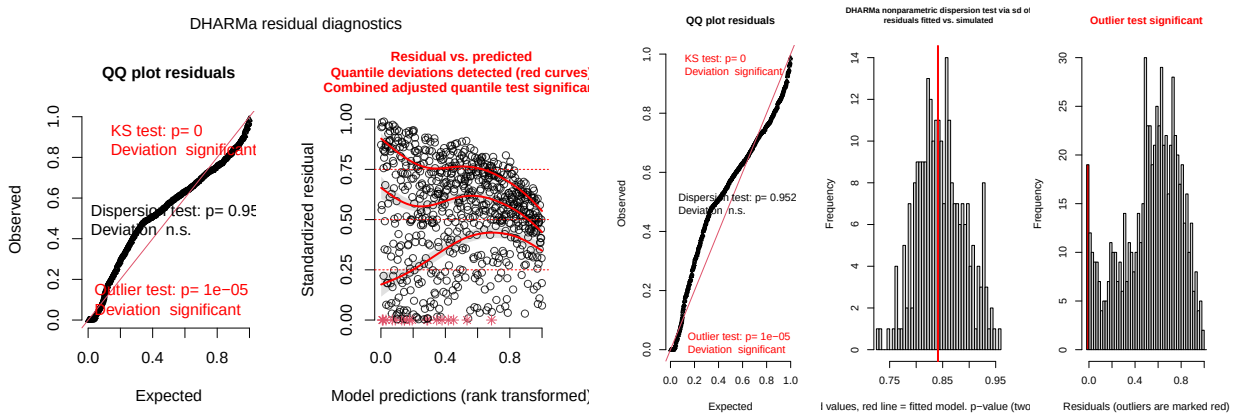












- Not perfect, but its sure improved. We move to the next steps

```
summary(lm_uptodate_2)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ logEnrolled + PctChildPoverty + PctFamilyPoverty,
##     data = districts_inference_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.419  -3.903   2.125   7.101  23.061
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      69.41298    1.87010  37.117   <2e-16 ***
## logEnrolled       2.44399    0.29024   8.421   <2e-16 ***
## PctChildPoverty   0.11303    0.07645   1.479   0.1397
## PctFamilyPoverty  0.24868    0.11241   2.212   0.0273 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.82 on 692 degrees of freedom
## Multiple R-squared:  0.1551, Adjusted R-squared:  0.1514
## F-statistic: 42.34 on 3 and 692 DF,  p-value: < 2.2e-16
```

- A linear model was generated to predict the Pct of uptodate using percentage of child poverty, percentage of family poverty and enrolled number.
- The null hypothesis is that R-squared value for population is 0. $F(692,3) = 42.34$, in favor of alternate hypothesis and p-value(2.2e-16) is <0.05. So our test is significant and we reject the null hypothesis. Had the null not been rejected then likelihood of observing a F-value value > 42.34 is less)
- The overall R-squared value is 0.1551. The adjusted R-squared is significant with a value is 0.1514, the 3 independent variables account to 15.14% of the data variability. The median of residuals is not around 0, this is because of the outliers that we didnot get rid of.

To see which predictors have the biggest impact, we can look at standardized coefficients, which are based on standardized variables, meaning that each gives the impact of 1 standarded deviation change in the predictor on the outcome variable

```
library(lm.beta)
summary(lm.beta(lm_uptodate_2))
```

```
##
## Call:
## lm(formula = PctUpToDate ~ logEnrolled + PctChildPoverty + PctFamilyPoverty,
##     data = districts_inference_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.419  -3.903   2.125   7.101  23.061
##
## Coefficients:
##                  Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)      69.41298      0.00000    1.87010  37.117   <2e-16 ***
## logEnrolled       2.44399      0.29910    0.29024   8.421   <2e-16 ***
## PctChildPoverty   0.11303      0.10601    0.07645   1.479   0.1397
## PctFamilyPoverty  0.24868      0.15848    0.11241   2.212   0.0273 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.82 on 692 degrees of freedom
## Multiple R-squared:  0.1551, Adjusted R-squared:  0.1514
## F-statistic: 42.34 on 3 and 692 DF,  p-value: < 2.2e-16
```

- According to the coefficients : we cannot interpret PctChildPoverty as it is not significant.Whereas we the other two are significant. We reject the null hypothesis that the B-weights for PctFamilyPoverty and LogEnrolled are 0 as they are significant
- To interpret the values of coefficients : Every unit increase in logenrolled, increases the pctuptodate by 2.44, whereas every unit increase in family poverty(if family poverty percentage rises by 1%), the percentage of uptodate goes up by 0.2486
- Performing Bayesian Analysis:

```
library(BayesFactor)
uptodate_mcmc <- lmBF(PctUpToDate ~ logEnrolled + PctChildPoverty +
    PctFamilyPoverty,
    data=districts_inference_2,
    posterior=TRUE, iterations=10000)
summary(uptodate_mcmc)
```

```
##
## 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                87.6171 0.44056 0.0044056      0.0044056
```

```
## logEnrolled        2.3940 0.29112 0.0029112      0.0029611
## PctChildPoverty    0.1102 0.07609 0.0007609      0.0007481
## PctFamilyPoverty   0.2447 0.11197 0.0011197      0.0011000
## sig2             139.9133 7.50228 0.0750228      0.0749412
## g                  0.1527 0.31599 0.0031599      0.0032439
##
## 2. Quantiles for each variable:
##
##                       2.5%       25%       50%      75%    97.5%
## mu                 86.75542  87.31931  87.61202  87.9130  88.4884
## logEnrolled         1.82801   2.19524   2.39371   2.5909   2.9710
## PctChildPoverty    -0.03872   0.05897   0.11127   0.1620   0.2583
## PctFamilyPoverty    0.02474   0.16950   0.24563   0.3199   0.4625
## sig2              126.09933 134.75625 139.60739 144.9050 155.2153
## g                   0.02649   0.05568   0.08889   0.1567   0.6694
```

- We ran the Bayesian Linear regression using lmBF() function with posterior as true and 10000 iterations using the MCMC technique for sampling.
- In the first part, Mean column are the parameter estimates values for the coefficients of our independentvariables(PctChildPoverty,PctFamilyPoverty,LogEnrolled). For LogEnrolled it is 2.3983, for PctFamilyPoverty it is 0.2418 and for PctChildPoverty it is 0.1125 which are very close to the values that we generated using the lm() function.
- In the 2nd part we see the 95% HDI interval values(2.5% and 97.5%) for each of the B-weights. The HDI interval values are the edges of the central region of the posterior distribution for each of the variable considered.For logEnrolled there is a 95% chance that the coefficient value/B-weight will lie between 1.83384 and 2.9714. For PctFamilyPoverty the range is from 0.02204 to 0.4628, for PctChildPoverty the range is from -0.03444 to 0.2628(this interval contains 0,tells us that PctChildPoverty is not a good predictor because mean value can be 0). As intervals for PctFamilyPoverty and logEnrolled donot contain 0, we can say that a model with these two variables variables as independent variables/predictors will be better than just the y-intercept. All of these findings run parallel with our findings from the frequentist method. Having PctChildPoverty insignificant.
- sig2 here gives the model precision for 10000 iterations.It gives the summary of the error in the model.R squared is (1-sig2)/variance of dependent variable. So to get bigger value of RSquared, the sig2value should be less.

```
library(BayesFactor)
uptodate_mcmc_bf <- lmBF(PctUpToDate ~ logEnrolled + PctChildPoverty +
    PctFamilyPoverty,
    data=districts_inference_2)
uptodate_mcmc_bf
```

```
## Bayes factor analysis
## --------------
## [1] logEnrolled + PctChildPoverty + PctFamilyPoverty : 8.854202e+21 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

- According to the model being used, the bayes factor is 8.854202e+21 ±0% which is a very high value. It shows very high and strong odds in favor of the alternative hypothesis that the model using PctFamilyPoverty, logEnrolled and PctChildPoverty as predictors/independent variables is highly favored over the model that only has the y-intercept. Though this stands as strong evidence, it doesnot give us information about which variable effects the outcome variable more.

c. ***Using any set or combination 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?***

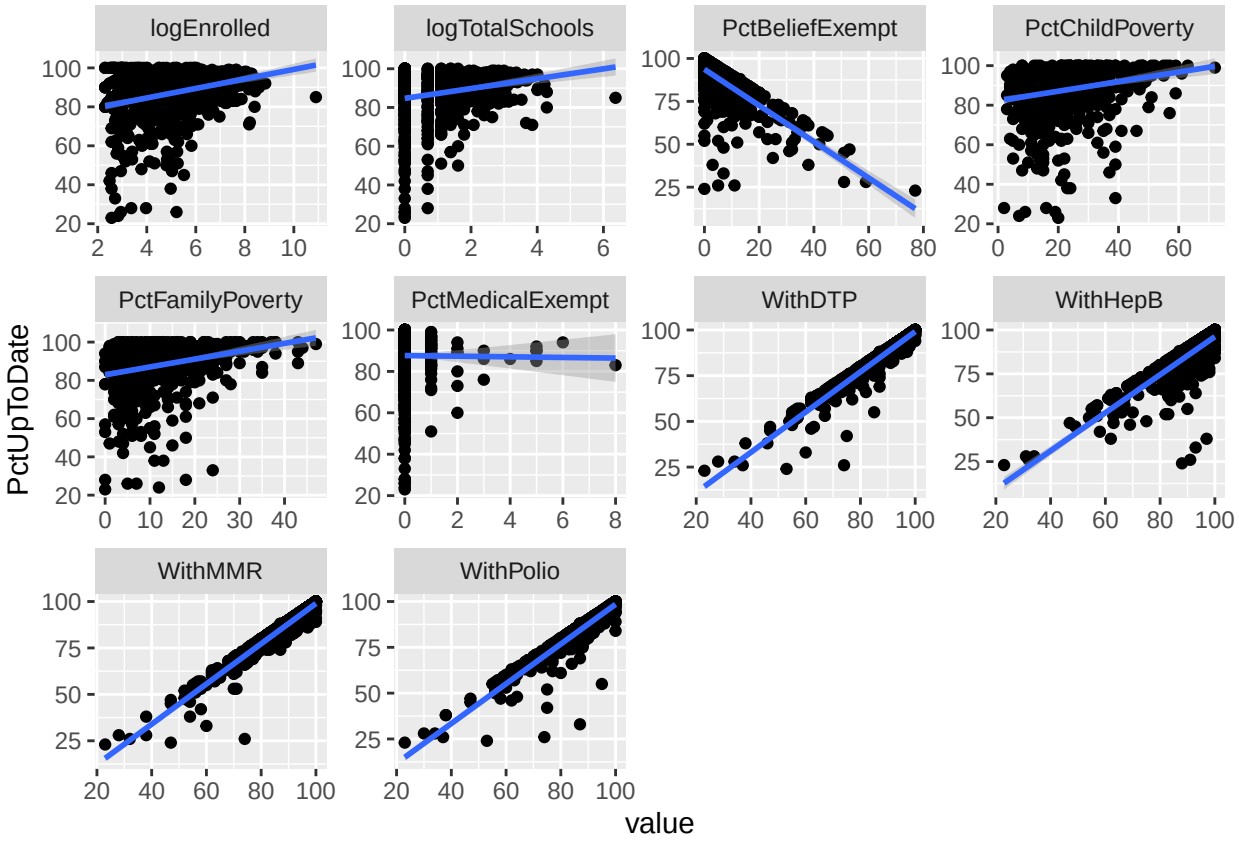
- Here our independent variables are all except The dependent variable is PctUpToDate.

```
districts_inference_3 <- subset(districts_new,select = -c(DistrictName,
                                                          DistrictComplete,
                                                          Enrolled,TotalSchools))
```

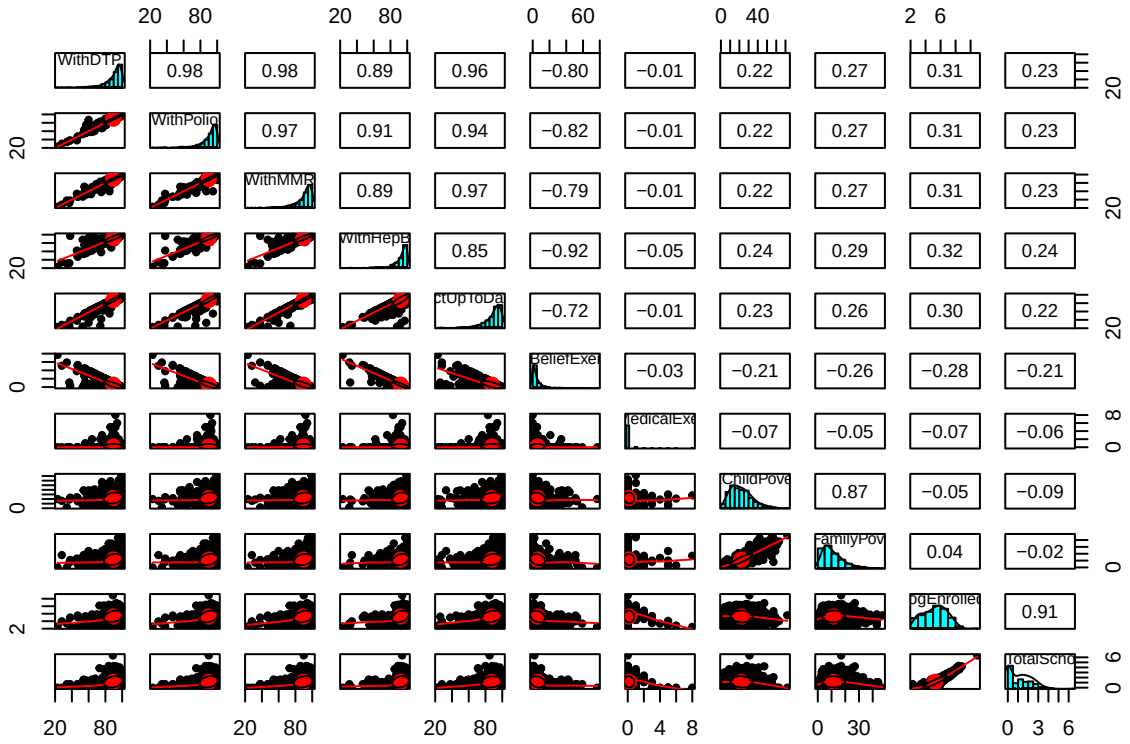
- We di bivariate exploration of data to understand te independent and dependent variable relationships better.
- a) In this analysis, we're looking for bivariate outliers and non-linear relationships. Plotting scatter plots to check for any patterns/problems

```
districts_inference_3 %>% 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'
```




```
pairs.panels(districts_inference_3)
```



- The above plots show that the data is more or less normally distributed around the line with no issues. Medical exempt seems a little biased but we will proceed further
- b) We now check for correlations between the variables considered.

```
districts_corr_3 <- cor(districts_inference_3, use="pairwise.complete.obs")
signif(districts_corr_3)
```

```
##                      WithDTP   WithPolio     WithMMR   WithHepB PctUpToDate
## WithDTP           1.00000000  0.98157500  0.97750800  0.8896180   0.9603770
## WithPolio         0.98157500  1.00000000  0.96637900  0.9050620   0.9407690
## WithMMR           0.97750800  0.96637900  1.00000000  0.8901390   0.9687760
## WithHepB          0.88961800  0.90506200  0.89013900  1.0000000   0.8450570
## PctUpToDate       0.96037700  0.94076900  0.96877600  0.8450570   1.0000000
## PctBeliefExempt  -0.79926500 -0.81998200 -0.78576700 -0.9189000  -0.7246970
## PctMedicalExempt -0.00576495 -0.00733702 -0.00816547 -0.0488398  -0.0077475
## PctChildPoverty   0.22284700  0.22367300  0.22448600  0.2405210   0.2273960
## PctFamilyPoverty  0.26582700  0.26761300  0.26501300  0.2868560   0.2617610
## logEnrolled       0.30736300  0.30715100  0.31025700  0.3207870   0.2992440
## logTotalSchools   0.23138500  0.23114700  0.23037000  0.2350280   0.2231860
##                  PctBeliefExempt PctMedicalExempt PctChildPoverty
## WithDTP               -0.7992650      -0.00576495       0.2228470
## WithPolio             -0.8199820      -0.00733702       0.2236730
## WithMMR               -0.7857670      -0.00816547       0.2244860
## WithHepB              -0.9189000      -0.04883980       0.2405210
```

```
## PctUpToDate           -0.7246970      -0.00774750       0.2273960
## PctBeliefExempt        1.0000000      -0.02566240      -0.2064430
## PctMedicalExempt      -0.0256624       1.00000000      -0.0727744
## PctChildPoverty       -0.2064430      -0.07277440       1.0000000
## PctFamilyPoverty      -0.2554740      -0.04715990       0.8688450
## logEnrolled           -0.2809900      -0.06699010      -0.0545198
## logTotalSchools       -0.2128760      -0.05814530      -0.0875150
##                   PctFamilyPoverty logEnrolled logTotalSchools
## WithDTP                  0.2658270   0.3073630       0.2313850
## WithPolio                0.2676130   0.3071510       0.2311470
## WithMMR                  0.2650130   0.3102570       0.2303700
## WithHepB                 0.2868560   0.3207870       0.2350280
## PctUpToDate              0.2617610   0.2992440       0.2231860
## PctBeliefExempt         -0.2554740  -0.2809900      -0.2128760
## PctMedicalExempt        -0.0471599  -0.0669901      -0.0581453
## PctChildPoverty          0.8688450  -0.0545198      -0.0875150
## PctFamilyPoverty         1.0000000   0.0373644      -0.0231195
## logEnrolled              0.0373644   1.0000000       0.9139180
## logTotalSchools         -0.0231195   0.9139180       1.0000000
```

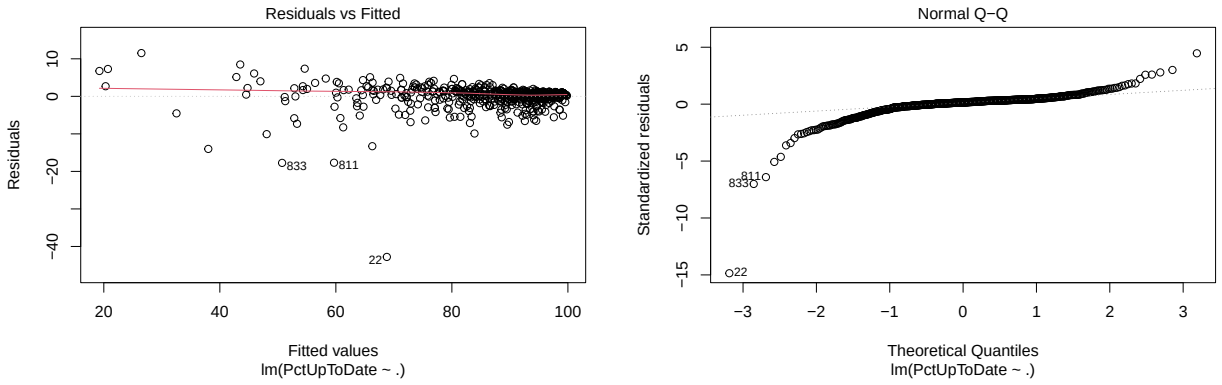
- WithDTP, WithPolio, WithMMR, WithHepB, PctUpToDate have strong correlations between them.

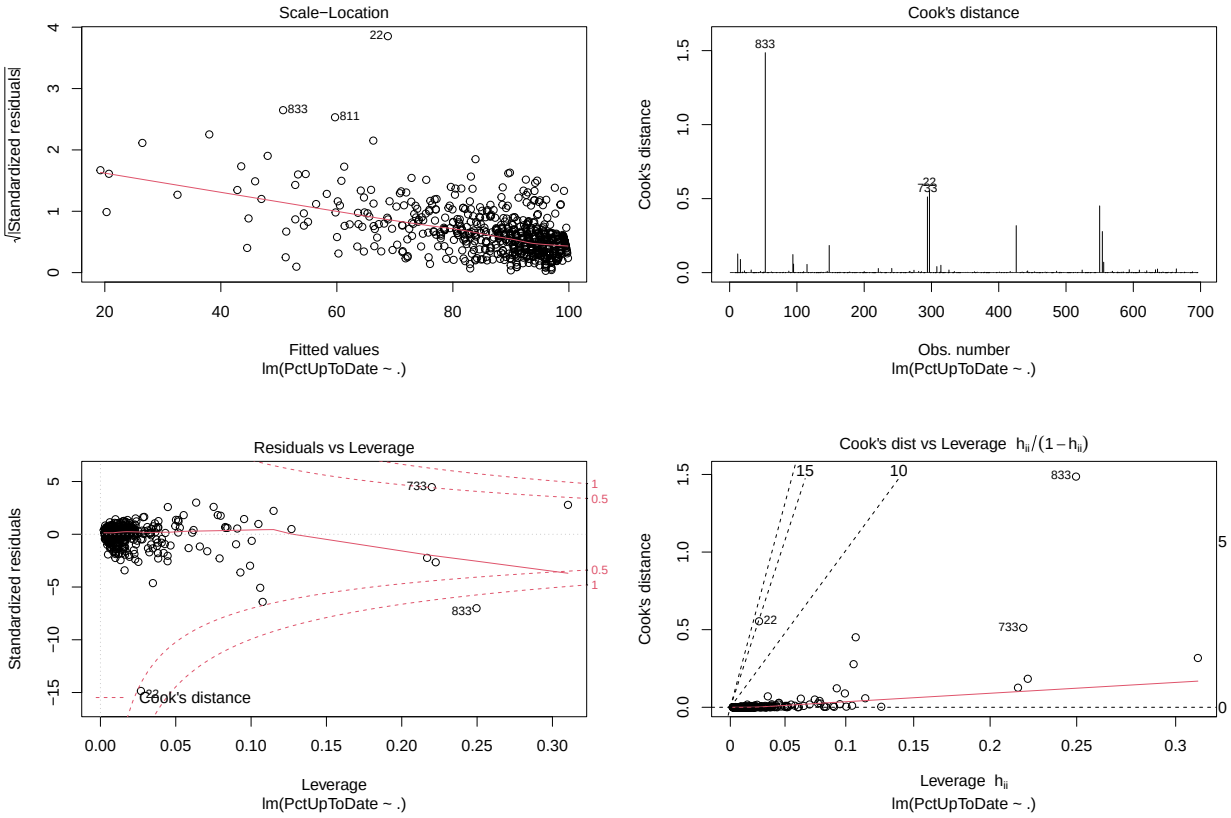
```
sort(districts_corr_3["PctUpToDate",])
```

```
##  PctBeliefExempt PctMedicalExempt  logTotalSchools  PctChildPoverty
##     -0.724697296     -0.007747497      0.223186073      0.227395697
## PctFamilyPoverty      logEnrolled         WithHepB        WithPolio
##      0.261760552      0.299244265      0.845056911      0.940768749
##          WithDTP          WithMMR      PctUpToDate
##      0.960377067      0.968776058      1.000000000
```

```
#Linear Model

lm_uptodate_all3 <- lm(PctUpToDate ~.,data =districts_inference_3)
```

- a) First we check the residuals:





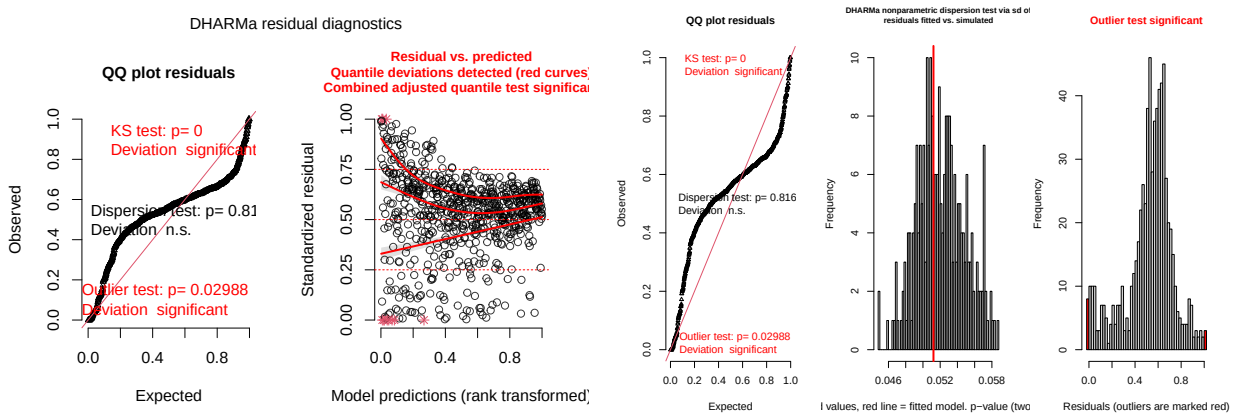
- Ideally, for the predictors to makeup to a good model, the residuals should not deviate a lot from the red line in residuals vs fitted plot.
- 22,11 as they have been marked as outliers in the plots.We look into this data:

```
districts_new[c(22,11),]
```

```
##                DistrictName WithDTP WithPolio WithMMR WithHepB PctUpToDate
## 331          Carmel Unified      82        84      74       81          66
## 686 Anderson Valley Unified      79        84      86       91          79
##     DistrictComplete PctBeliefExempt PctMedicalExempt PctChildPoverty
## 331             TRUE              11                0              15
## 686             TRUE               0                0              36
##     PctFamilyPoverty Enrolled TotalSchools logEnrolled logTotalSchools
## 331                4      175            3    5.164786        1.098612
## 686               19       43            1    3.761200        0.000000
```

- Anderson Valley Unified,Carmel Unified are marked as outliers, when we observe the data, it they more or less are inline with the mean values in(WithDTP,WithPolio,WithMMR,WithHepB). The PctChild-Poverty seems a little high

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

## Warning in asinh(z): NaNs produced

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



- DharMa simulations show that there is some deviation from the ideal red line in the qq plot. Ideally i would have transformed the data to remove skewness and outliers, but because we only have 700 obsrvations, i will let it be.
- b) Checking for multicollinearity which was hinted in the correlation values:

```
library(car)
vif(lm_uptodate_all3)
```

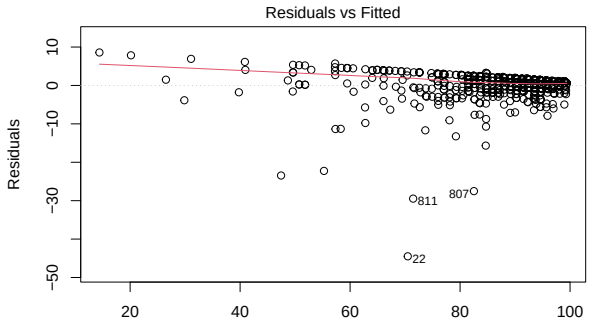
```
##          WithDTP        WithPolio          WithMMR         WithHepB
##        42.367479        32.132727        24.948992        13.453354
##  PctBeliefExempt PctMedicalExempt  PctChildPoverty PctFamilyPoverty
##         7.087355         1.059260         4.271250         4.324119
##      logEnrolled  logTotalSchools
##         6.753576         6.343505
```

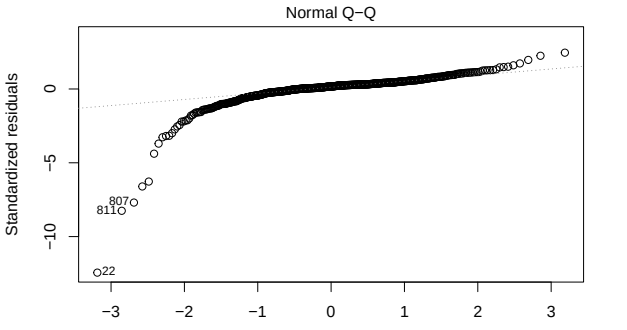
- We get rid of WithPolio, WithMMR, WithHepB bcause their values are very high. Logenrolled and logtotalschools are correlated so we removed one of them. PctBeliefExempt is removed as well.

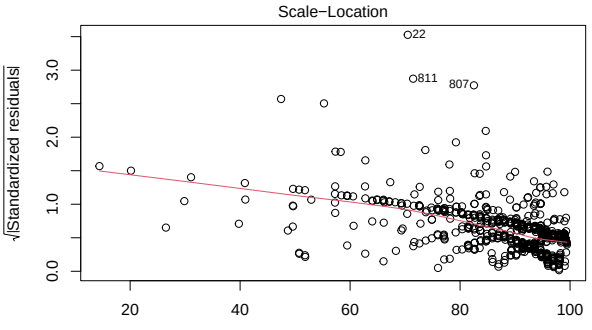
```
lm_uptodate_3 <- lm(PctUpToDate ~ WithDTP + PctMedicalExempt+ PctFamilyPoverty + logEnrolled,
                    data =districts_inference_3)
vif(lm_uptodate_3)
```

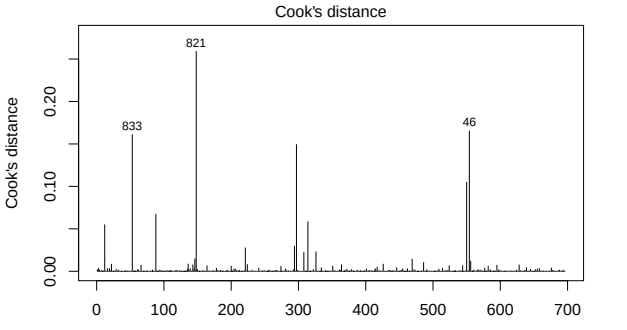
```
##          WithDTP PctMedicalExempt PctFamilyPoverty      logEnrolled
##         1.190396         1.007355         1.081350         1.112520
```

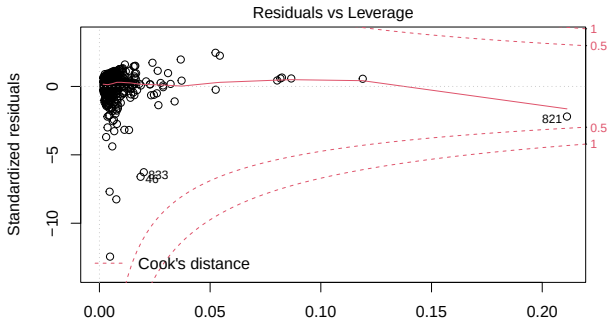
- Removing those variables reduced the effect of multicollinearity.

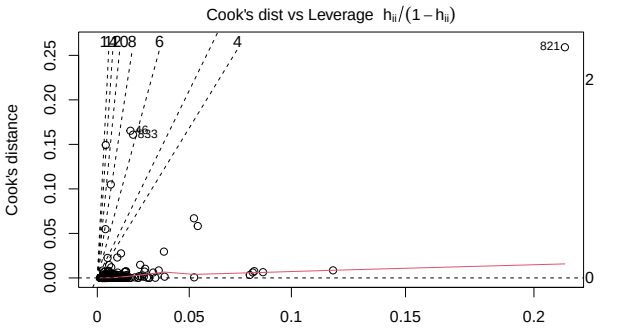

Residuals vs Fitted
Residuals
Fitted values
lm(PctUpToDate ~ WithDTP + PctMedicalExempt + PctFamilyPoverty + logEnrolle ..

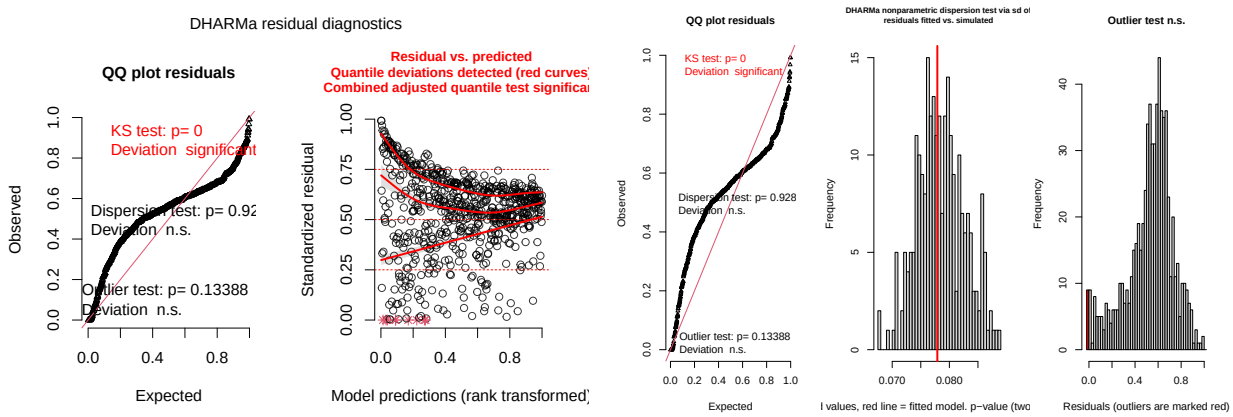

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(PctUpToDate ~ WithDTP + PctMedicalExempt + PctFamilyPoverty + logEnrolle ..


Scale−Location
√|Standardized residuals|
Fitted values
lm(PctUpToDate ~ WithDTP + PctMedicalExempt + PctFamilyPoverty + logEnrolle ..


Cook's distance
Cook's distance
Obs. number
lm(PctUpToDate ~ WithDTP + PctMedicalExempt + PctFamilyPoverty + logEnrolle ..


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(PctUpToDate ~ WithDTP + PctMedicalExempt + PctFamilyPoverty + logEnrolle ..


Cook's dist vs Leverage hii/(1 − hii)
Cook's distance
Leverage hii
lm(PctUpToDate ~ WithDTP + PctMedicalExempt + PctFamilyPoverty + logEnrolle ..



- Not perfect, but its sure improved. We move to the next steps

```
summary(lm_uptodate_3)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ WithDTP + PctMedicalExempt + PctFamilyPoverty +
##     logEnrolled, data = districts_inference_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.487  -0.582   0.611   1.412   8.584
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -10.89894    1.10812  -9.836   <2e-16 ***
## WithDTP            1.09635    0.01325  82.769   <2e-16 ***
## PctMedicalExempt  -0.03088    0.20800  -0.148    0.882
## PctFamilyPoverty   0.01115    0.01729   0.645    0.519
## logEnrolled        0.03855    0.09134   0.422    0.673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.584 on 691 degrees of freedom
## Multiple R-squared:  0.9224, Adjusted R-squared:  0.9219
## F-statistic:  2053 on 4 and 691 DF,  p-value: < 2.2e-16
```

- A linear model was generated to predict the Pct of uptodate using withDTP, percentage of family poverty, percentage of medical exempts and enrolled number.
- The null hypothesis is that R-squared value for population is 0. F(691,4) = 2053, in favor of alternate hypothesis and p-value(2.2e-16) is <0.05. So our test is significant and we reject the null hypothesis. Had the null not been rejected then likelihood of observing a F-value value > 2053 is less)
- The overall R-squared value is 0.9224. The adjusted R-squared is significant with a value is 0.9219, the 4 independent variables account to 92.19% of the data variability. The median of residuals is around 0, showing normal distribution.

To see which predictors have the biggest impact, we can look at standardized coefficients, which are based on standardized variables, meaning that each gives the impact of 1 standarded deviation change in the predictor on the outcome variable

```
library(lm.beta)
summary(lm.beta(lm_uptodate_3))
```

```
##
## Call:
## lm(formula = PctUpToDate ~ WithDTP + PctMedicalExempt + PctFamilyPoverty +
##     logEnrolled, data = districts_inference_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.487  -0.582   0.611   1.412   8.584
##
## Coefficients:
##                    Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)      -10.898942     0.000000   1.108123  -9.836   <2e-16 ***
## WithDTP            1.096346     0.957029   0.013246  82.769   <2e-16 ***
## PctMedicalExempt  -0.030880    -0.001579   0.207999  -0.148    0.882
## PctFamilyPoverty   0.011149     0.007105   0.017292   0.645    0.519
## logEnrolled        0.038550     0.004718   0.091337   0.422    0.673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.584 on 691 degrees of freedom
## Multiple R-squared:  0.9224, Adjusted R-squared:  0.9219
## F-statistic:  2053 on 4 and 691 DF,  p-value: < 2.2e-16
```

- According to the coefficients : we cannot interpret PctMedicalExempt,PctFamilyPOverty,logEnrolled as they are not significant.Whereas WithDTP is significant. We reject the null hypothesis that the B-weight for WithDTP is 0. So we know that PctMedicalExempt,PctFamilyPOverty,logEnrolled have no effect in contribyting to the percetage of uptodate value
- To interpret the values of coefficients : Every unit increase in WithDTP, the PctUpToDate goes up by 1.09.
- Performing Bayesian Analysis:

```
library(BayesFactor)
uptodate2_mcmc <- lmBF(PctUpToDate ~ logEnrolled +WithDTP + PctMedicalExempt+
                         PctFamilyPoverty ,
                       data=districts_inference_3,
                       posterior=TRUE, iterations=10000)
summary(uptodate2_mcmc)
```

```
##
## 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                87.62123 0.13701 0.0013701      0.0013701
## logEnrolled        0.03890 0.09158 0.0009158      0.0009158
## WithDTP            1.09547 0.01329 0.0001329      0.0001329
## PctMedicalExempt -0.02878 0.21629 0.0021629      0.0021629
## PctFamilyPoverty  0.01143 0.01731 0.0001731      0.0001639
## sig2             12.90787 0.75575 0.0075575      0.0076846
## g                 3.88891 4.44652 0.0444652      0.0444652
##
## 2. Quantiles for each variable:
##
##                       2.5%        25%      50%      75%    97.5%
## mu                87.35786 87.5294389 87.62156 87.71167 87.88635
## logEnrolled      -0.13768 -0.0229023  0.03983  0.10041  0.21364
## WithDTP           1.06902  1.0865789  1.09549  1.10458  1.12130
## PctMedicalExempt -0.43686 -0.1689224 -0.02662  0.11492  0.37829
## PctFamilyPoverty -0.02216 -0.0001564  0.01136  0.02305  0.04423
## sig2             11.58940 12.4276284 12.86950 13.37070 14.32918
## g                 0.91774  1.7662216  2.70543  4.38132 14.38373
```

- We ran the Bayesian Linear regression using lmBF() function with posterior as true and 10000 iterations using the MCMC technique for sampling.
- In the first part, Mean column are the parameter estimates values for the coefficients of our independent variables(PctMedicalExempt,PctFamilyPoverty,LogEnrolled,WithDTP). The values are smiliar to the ones geenrated by the lm() fucntion.
- In the 2nd part we see the 95% HDI interval values(2.5% and 97.5%) for each of the B-weights. The HDI interval values are the edges of the central region of the posterior distribution for each of the variable considered.For logEnrolled there is a 95% chance that the coefficient value/B-weight will lie between -0.13892 and 0.22441. For PctFamilyPoverty the range is from -0.02278 to 0.04463, for PctMedicalExempt the range is from -0.44308 to 0.37136. All of these intervals contains 0,tells us that they are not a good predictors because mean value can be 0).
- Interval range for WithDTP is from 1.06957 to 1.12152. As interval for WithDTP does not contain 0, we can say that a model with these this variable as independent variable/predictor will be better than just the y-intercept. All of these findings run parallel with our findings from the frequentist method.
- sig2 here gives the model precision for 10000 iterations.It gives the summary of the error in the model.R squared is (1-sig2)/variance of dependent variable. So to get bigger value of RSquared, the sig2value should be less.

```
library(BayesFactor)
uptodate2_mcmc_bf <- lmBF(PctUpToDate ~ logEnrolled +WithDTP + PctMedicalExempt+ PctFamilyPoverty ,
                          data=districts_inference_3)
uptodate2_mcmc_bf
```

```
## Bayes factor analysis
## --------------
## [1] logEnrolled + WithDTP + PctMedicalExempt + PctFamilyPoverty : 2.556332e+377 ±0%
##
```

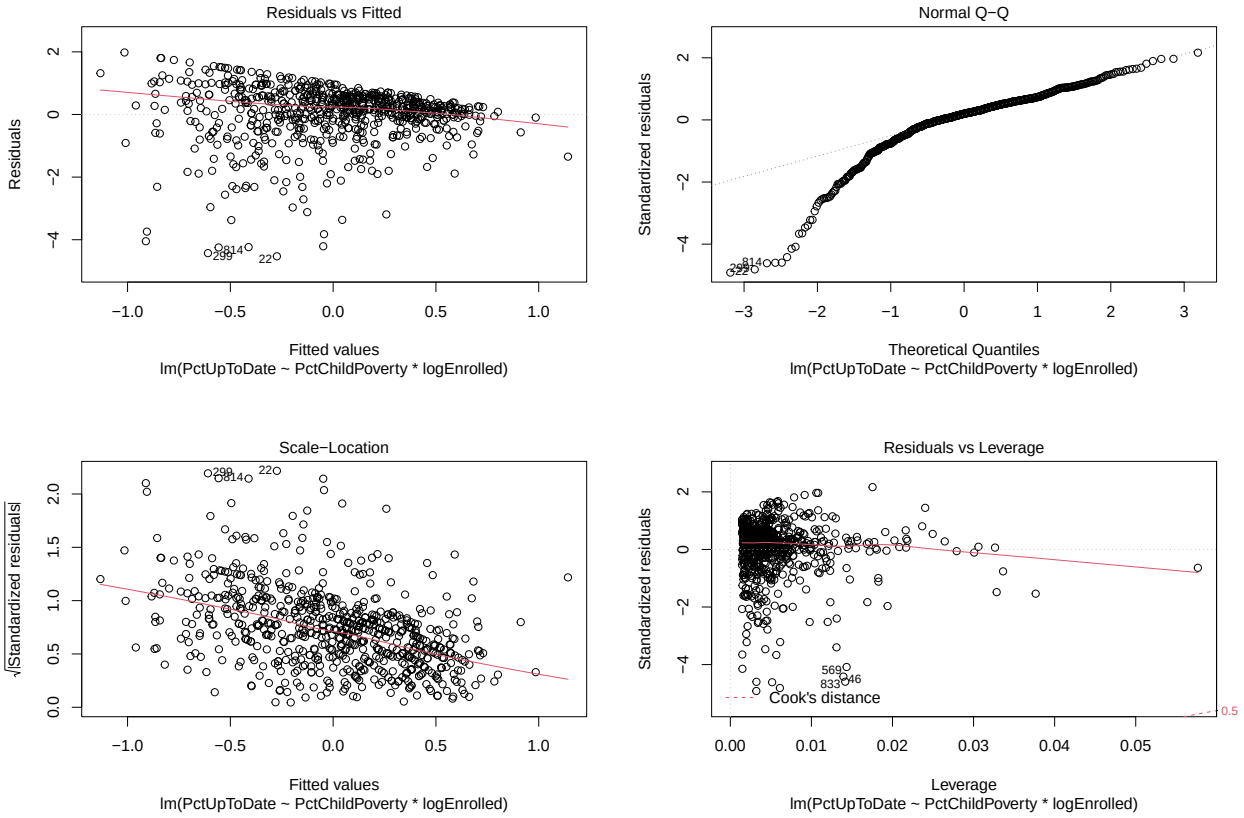
```
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

- According to the model being used, the bayes factor is 2.556332e+377 ±0% which is a very high value. It shows very high and strong odds in favor of the alternative hypothesis that the model using PctFamilyPoverty, logEnrolled ,WithDTP and PctMedicalExempt as predictors/independent variables is highly favored over the model that only has the y-intercept. Though this stands as strong evidence, it doesnot give us information about which variable effects the outcome variable more.

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

```
df_new <- subset(districts_new, select = c(logEnrolled,
                                           PctChildPoverty,
                                           PctUpToDate))
df_new <- data.frame(scale(df_new),center = T,scale = F)
lm <- lm(PctUpToDate ~ PctChildPoverty * logEnrolled, data  = df_new)
summary(lm)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ PctChildPoverty * logEnrolled, data = df_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5294 -0.2707  0.1809  0.5442  1.9797
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.003361   0.035039  -0.096    0.924
## PctChildPoverty              0.236154   0.035433   6.665 5.41e-11 ***
## logEnrolled                  0.315713   0.035108   8.993  < 2e-16 ***
## PctChildPoverty:logEnrolled -0.061729   0.038557  -1.601    0.110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9227 on 692 degrees of freedom
## Multiple R-squared:  0.1523, Adjusted R-squared:  0.1486
## F-statistic: 41.43 on 3 and 692 DF,  p-value: < 2.2e-16
```



- The resiudals look fine ,we go with bayesian analysis

```
lmbf <- lmBF(PctUpToDate ~ logEnrolled+ PctChildPoverty, data = df_new)
lmbf
```

```
## Bayes factor analysis
## --------------
## [1] logEnrolled + PctChildPoverty : 7.316619e+21 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

- Bayes factor for the model with no interaction is 7.316619e+21 ±0%. Strong odds in favor of the alternate hypothesis that percentage of students with uptodate vaccines can be predicted using the log enrolled percentage and percentage of childern under the considered poverty line.

```
lmbf2 <- lmBF(PctUpToDate ~ logEnrolled* PctChildPoverty, data = df_new)
lmbf2
```

```
## Bayes factor analysis
## --------------
## [1] logEnrolled * PctChildPoverty : 2.835457e+21 ±0%
```

```
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

- Strong odds in favor of alternate hypothesis 2.835457e+21 ±0%. Interaction term was used here.

```
lmbf2/lmbf
```

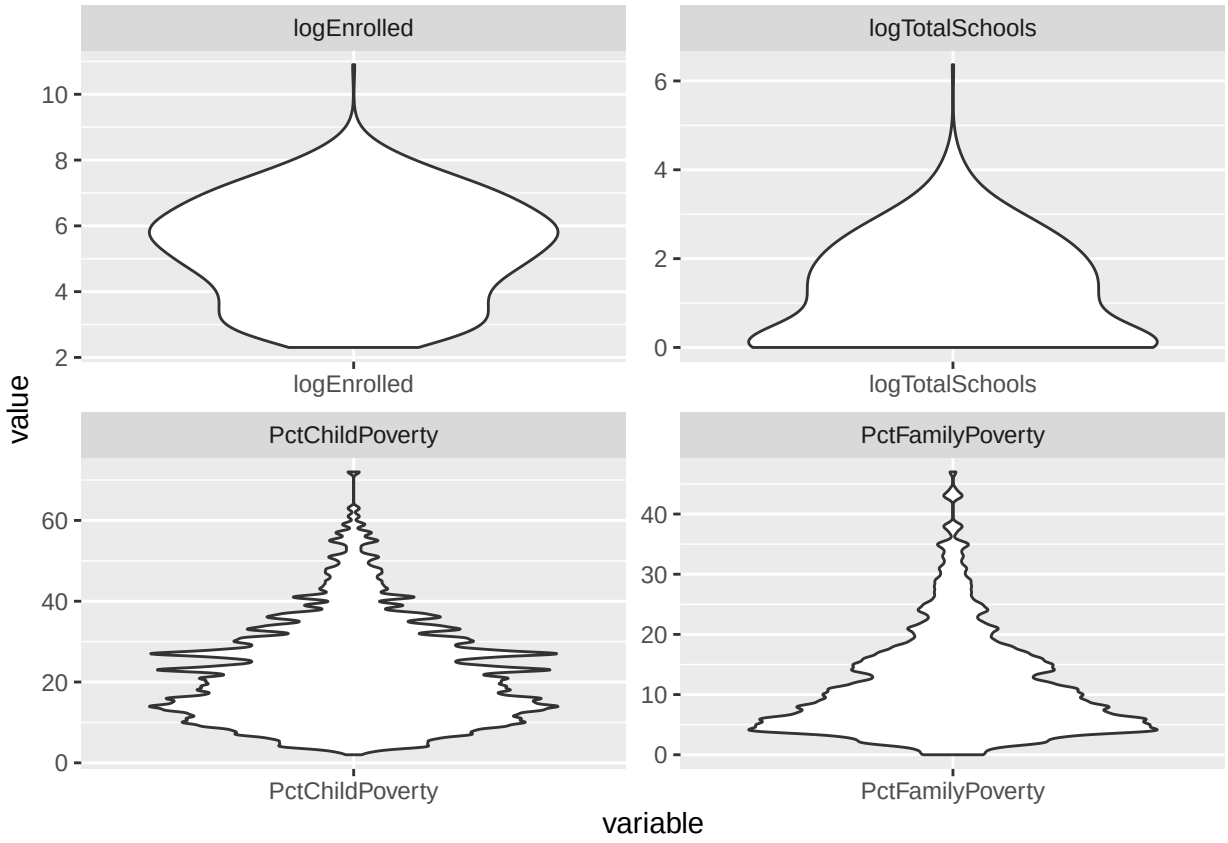
```
## Bayes factor analysis
## --------------
## [1] logEnrolled * PctChildPoverty : 0.3875365 ±0.01%
##
## Against denominator:
##   PctUpToDate ~ logEnrolled + PctChildPoverty
## ---
## Bayes factor type: BFlinearModel, JZS
```

- The ratio shows that the odds 0.3875 are in favor of the model which considered interaction. But the value is so small that it is not worth mentioning. Though it is understood that the moel with intercatin is better than the one without it.

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

```
districts_inference_5 <- subset(districts_new,
                                select = c(logEnrolled,
                                           logTotalSchools,
                                           PctChildPoverty,
                                           PctFamilyPoverty,
                                           DistrictComplete))
```

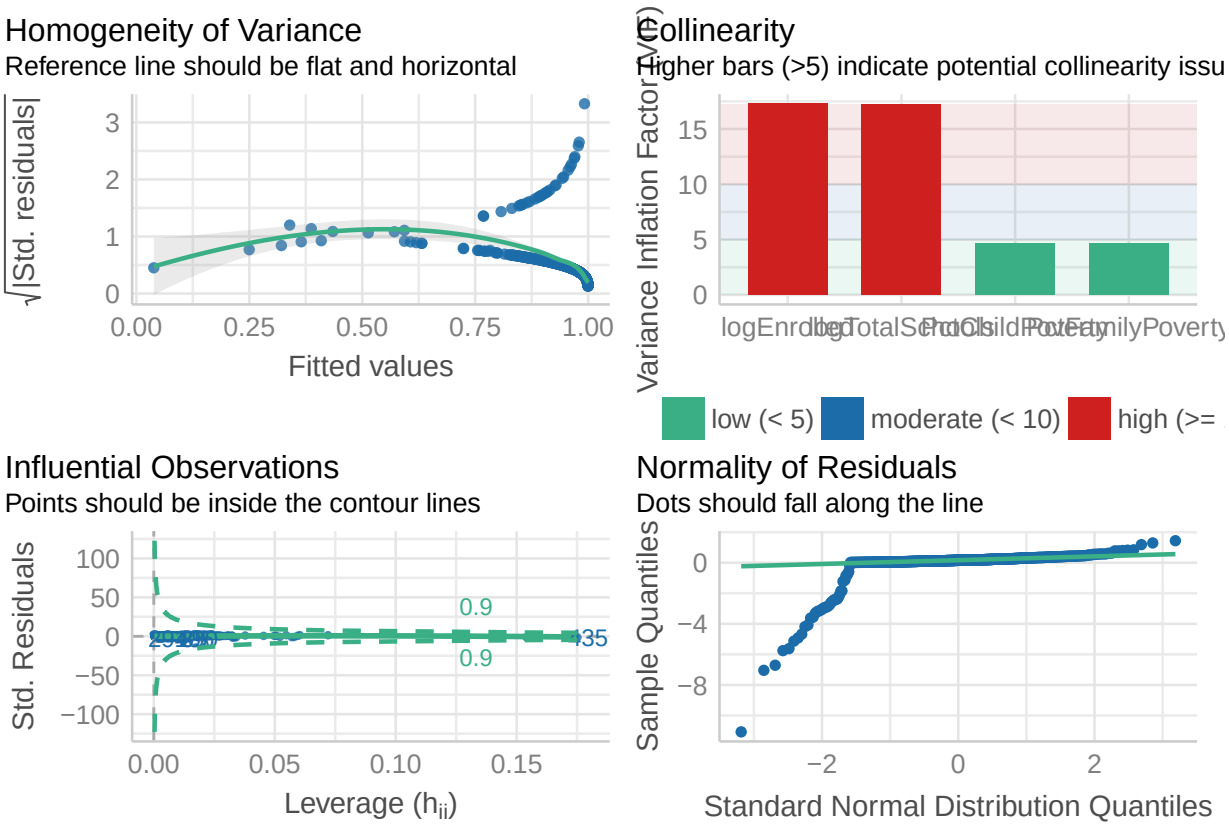
```
library(tidyverse)
districts_inference_5 %>% pivot_longer(cols=-c(DistrictComplete),
                                       names_to="variable",
                        values_to="value", values_drop_na = TRUE) %>%
ggplot(aes(x=variable, y=value)) +
  geom_violin(bw=.5) + facet_wrap( ~ variable, scales="free")
```



- We know thatlogEnrolled and logTotalSchools are correlated.
- We create a logistic regression model

```
log_mod <- glm(formula = DistrictComplete ~.,
               data = districts_inference_5,
               family = binomial(link = "logit"))
```

```
library(performance)
library(see)
check_model(log_mod)
```



- The refrence line is curved, there is collinearity(plot 2), the normality residulas plot shows that there are deviations.
- checking multicollinearity and we remove logtotalschools

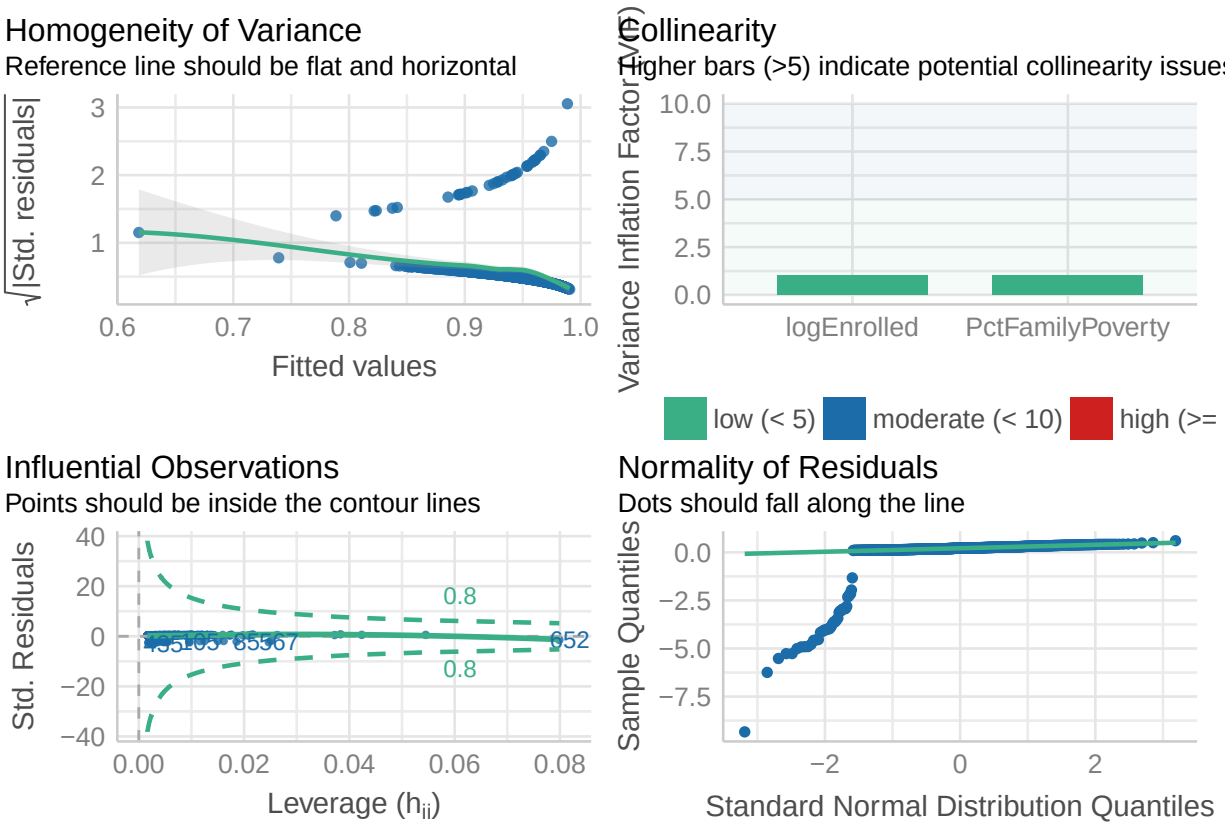
```
vif(log_mod)
```

```
##       logEnrolled  logTotalSchools  PctChildPoverty PctFamilyPoverty
##         17.284585        17.268081         4.667681         4.664407
```

- We remove PctChild poverty as the value is high and it can be correlated to family poverty. For the same reason we remove LogTotalSchools

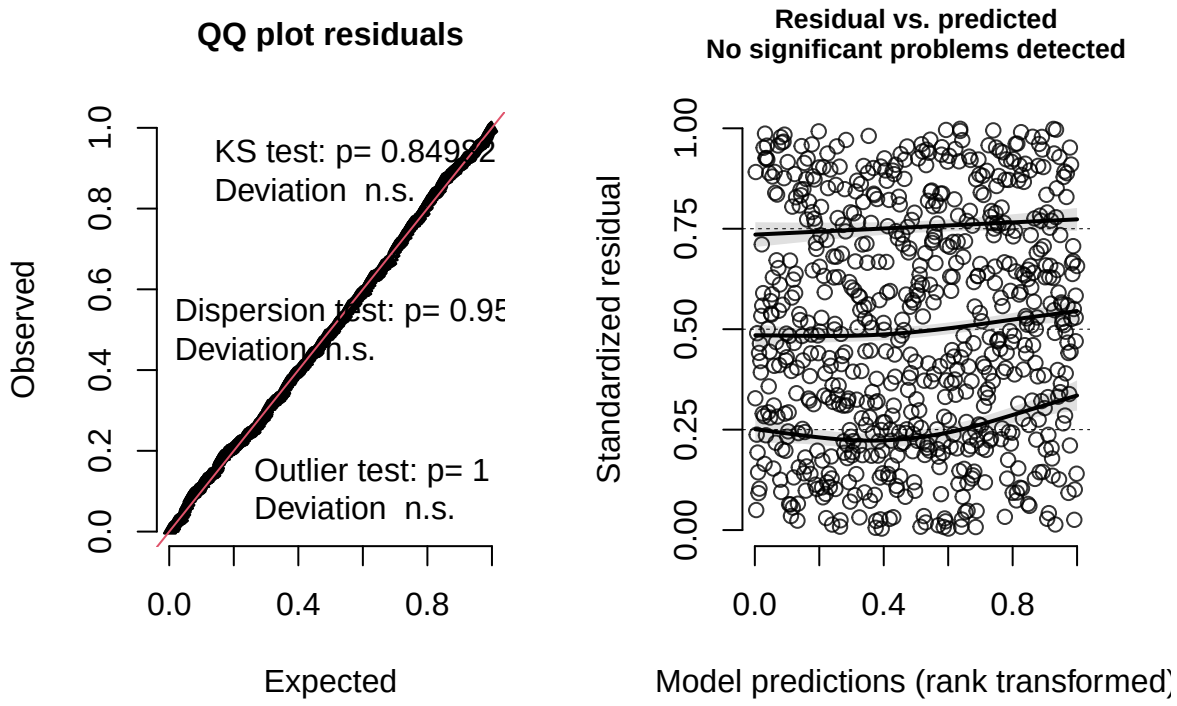
```
glm2 <- glm(formula = DistrictComplete ~logEnrolled + PctFamilyPoverty,
            data = districts_inference_5,
            family = binomial(link = "logit"))
```

```
check_model(glm2)
```

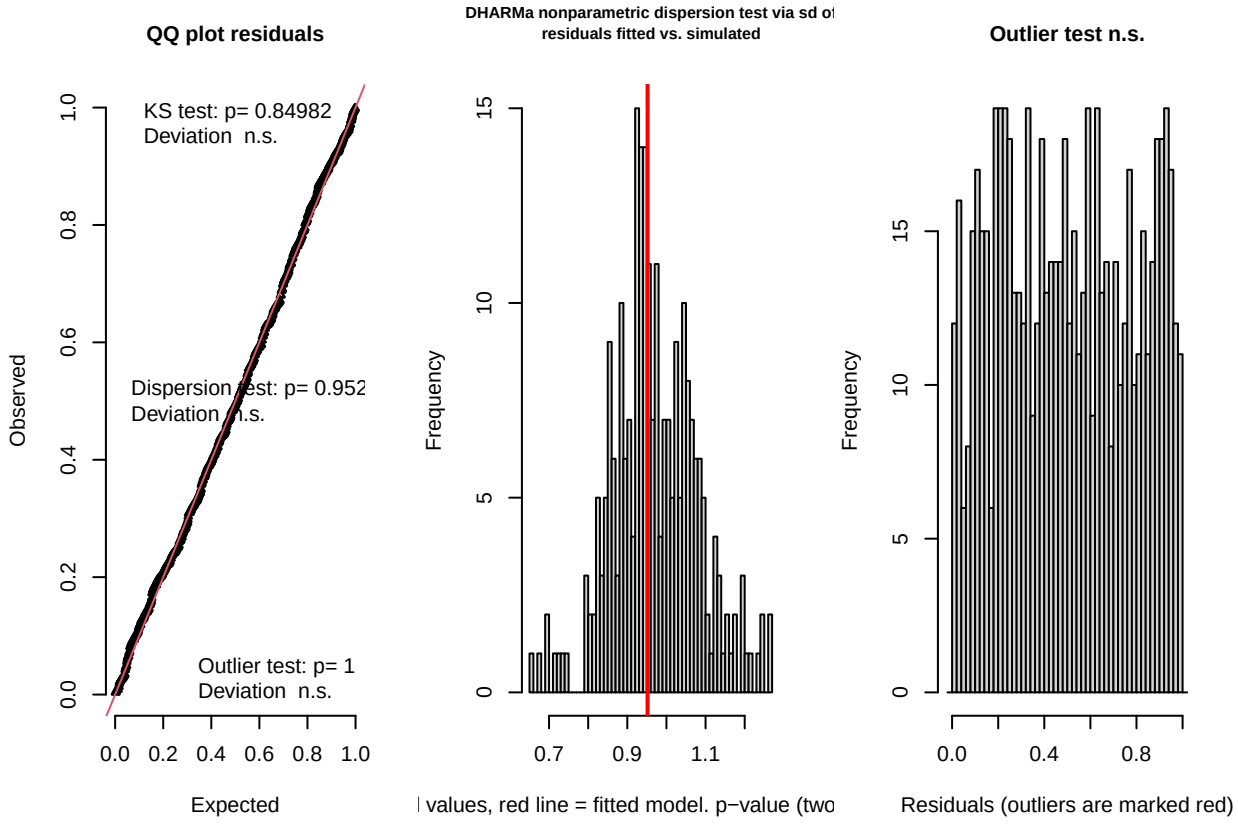


- collinearity is removed and reference line is not as curved as it was before. Though resiudals still deviate.

```
simulationOutput5 <- simulateResiduals(fittedModel = glm2, n = 250)
plot(simulationOutput5)
```




```
testResiduals(glm2)
```




```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.023148, p-value = 0.8498
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.9807, 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) = 5, observations = 696, p-value = 1
```

```
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.00233659 0.01668466
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.007183908

## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.023148, p-value = 0.8498
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.9807, 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) = 5, observations = 696, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.00233659 0.01668466
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.007183908
```

- The residuals lie perfectly on the line. so we can move ahead of our analysis and check the

```
summary(glm2)
```

```
##
## Call:
## glm(formula = DistrictComplete ~ logEnrolled + PctFamilyPoverty,
##     family = binomial(link = "logit"), data = districts_inference_5)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9922   0.2138   0.3036   0.3803   0.7774
##
```

```
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        5.62797    0.77433   7.268 3.64e-13 ***
## logEnrolled       -0.42290    0.11501  -3.677 0.000236 ***
## PctFamilyPoverty -0.03149    0.01917  -1.643 0.100392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 300.55  on 695  degrees of freedom
## Residual deviance: 283.18  on 693  degrees of freedom
## AIC: 289.18
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coef(glm2)) # Convert log odds to odds
```

```
##      (Intercept)      logEnrolled PctFamilyPoverty
##      278.0968861        0.6551474        0.9690016
```

```
exp(confint(glm2)) # Look at confidence intervals
```

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

```
##                       2.5 %       97.5 %
## (Intercept)      66.0128376 1388.8693950
## logEnrolled       0.5188140    0.8160816
## PctFamilyPoverty  0.9345626    1.0080234
```

- Coefficients : The intercept is not significant(p-value > 0.05) so we donot interpet it. PctFamilyPoverty is also not significant at 0.05 alpha level. According to the wald's Z test. So we fail to reject the null hypothesis that the coefficient/log odds can be 0 in the population. The coefficient of logEnrolled is statistically significant(pvalue < 0.05), the Wald's z-test value is -3.677 . We reject the null hypothesis that the coefficient/log-odds of logEnrolled is 0 in the population.
- To make sense out of the coefficients we convert them to normal values instead of logs.
- For logEnrolled, the odds are 1:0.65, meaning that in getting a TRUE(Completed reporting of vaccination) there is a 6.5% increase with increase in logEnrolled.
- For PctFamilyPoverty, for every unit change/increase in PctFamilyPoverty, there is a 9.6% more likely chance that the District completed their vaccination reporting.

```
library(MCMCpack)
```

```
## 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)
## ##
```

```
bayes_log <- MCMClogit(formula = DistrictComplete ~ logEnrolled + PctFamilyPoverty,
                       data = districts_inference_5)
summary(bayes_log)
```

```
##
## 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)       5.69382 0.77154 0.0077154      0.0251838
## logEnrolled      -0.42921 0.11397 0.0011397      0.0037315
## PctFamilyPoverty -0.03148 0.01907 0.0001907      0.0006254
##
## 2. Quantiles for each variable:
##
##                      2.5%      25%      50%      75%     97.5%
## (Intercept)       4.21497  5.16580  5.68192  6.20969  7.265056
## logEnrolled      -0.64778 -0.50597 -0.42907 -0.35216 -0.204948
## PctFamilyPoverty -0.06666 -0.04455 -0.03193 -0.01894  0.008276
```

- For log-odds, the above output describes the posteriro distributions for logEnrolled and PctFamily-Poverty as the independent variables.
- The point estimates are quite similar to what was generated in the previous model.
- In the second part, quantiles are displayed which are the 95% HDI intervals. Log enrolled can have value between -0.6477 and -0.204, PctFamilyPoverty can have -0.066 and 0.008. This includes 0, so we cannot be sure that our value is not 0.So there is a possibility that out coefficient cen be 0 for PctFamilyPoverty
- So we can say that using logEnrolled to predict if the district completed reportiong or not is bettwe than additonally using PctFamilyPoverty.

7. *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.

- Throughout our analysis we tried ot find the factors which contribute the best in predicting the vaccination rates according to belief exemptions, percentage of uptodate vaccinatios and so on.We also saw that private schools reported less compared to public schools. According to the regression analysis we did, we verified our approaches using frequentist and bayesian methods. Contradictory results werenot observed, so we can be confidence with our interpretations. As californian rates fall behind in DTP,POLIO and MMR compared to the US rates,these vaccines should be given out more. Family Povertyline infact had influence in reducing the vaccination rates, if free vaccinations are provided, that would definitely help the vaccniation rates increase.
- There is an intercation between percentage of students that enrolled and percentage of children living below the poverty line, lookimg into this futher would help understand better what alternatives can be provided to increase the vaccination rates. So this tells us in a way to focus on areas where the overall standard of living is below the poverty line.