Spring 2021 - Final Examination

Zijun Yi

```
## Loading required package: lattice

## Loading required package: ggplot2

##
## Attaching package: 'psych'

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

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

## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

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

## Loading required package: viridisLite

##
## Attaching package: 'dlookr'

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

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

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

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Successfully loaded changepoint package version 2.2.2
##  NOTE: Predefined penalty values changed in version 2.2.  Previous penalty values with a postfix 1 i
## Loading required package: carData
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
## The following object is masked from 'package:purrr':
##
##     some
## The following object is masked from 'package:psych':
##
##     logit
##
## Attaching package: 'nlme'
## The following object is masked from 'package:dplyr':
##
##     collapse
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
## This is DHARMa 0.4.1. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa')
## Loading required package: coda
## 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)
## ##
```

Instructions

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

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

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

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

Data

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

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

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

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

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

```
'data.frame':   700 obs. of  14 variables:
 $ DistrictName    : Name of the district
 $ WithDTP         : Percentage of students in the district with the DTP vaccine
```

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

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

Data Cleaning

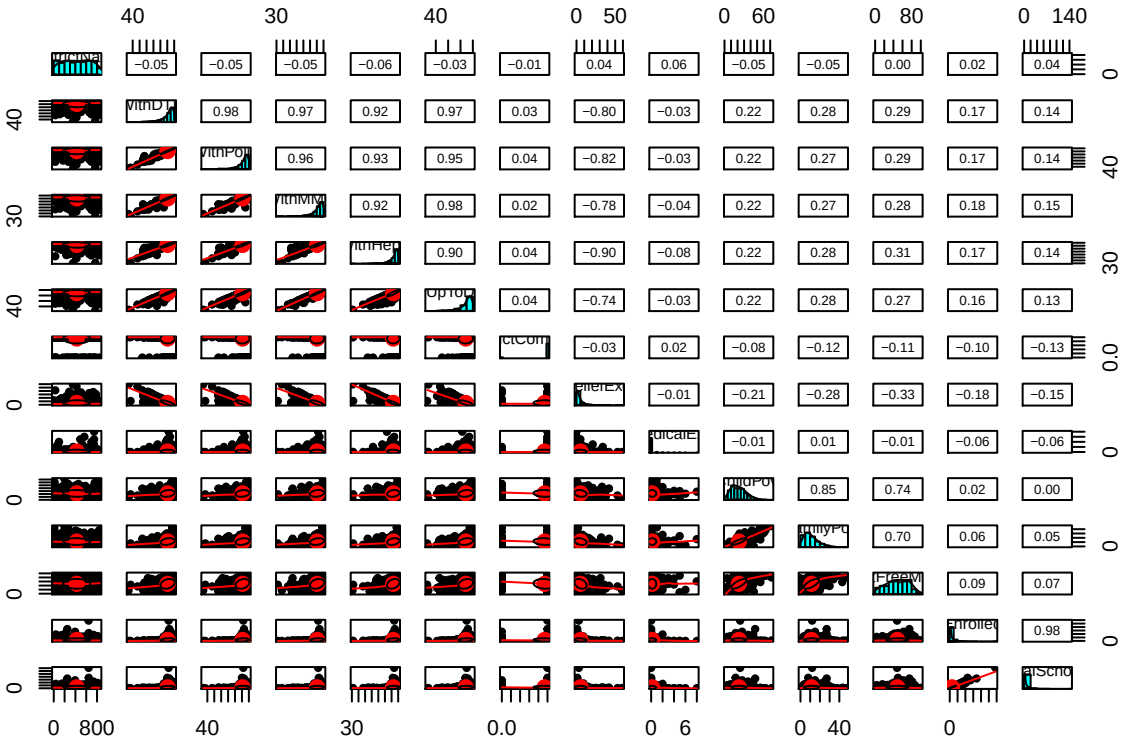
```
districts.filtered <- districts %>%
  filter(PctUpToDate <= 100) %>%
  filter(PctBeliefExempt <= 100) %>%
  filter(PctMedicalExempt <= 100) %>%
  filter(PctChildPoverty <= 100) %>%
  filter(PctFamilyPoverty <= 100) %>%
  filter(PctFreeMeal <= 100) # %>%
  # filter(DistrictComplete != FALSE)
```

Run Data Cleaning diagnose

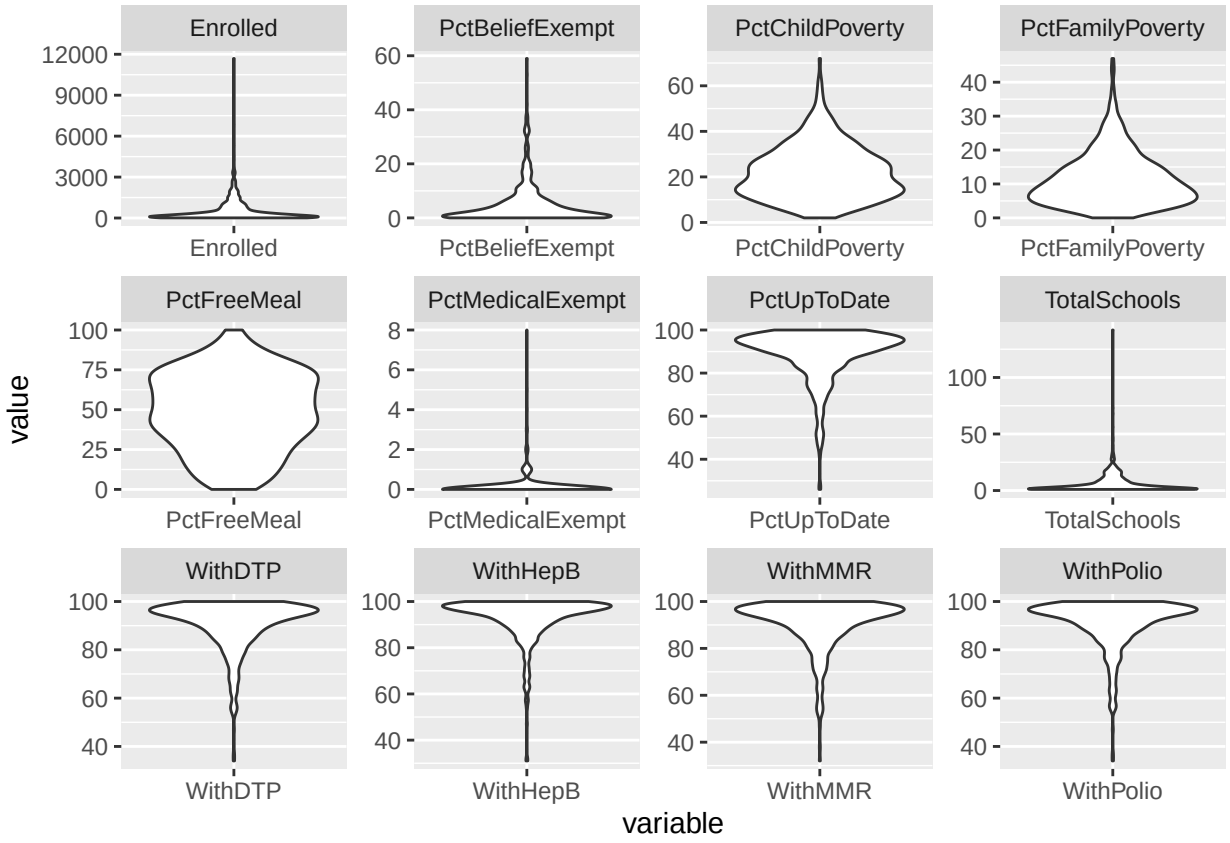
```
sum(is.na(districts.filtered))
```

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

```
# pairs panel on correlation
pairs.panels(districts.filtered)
```



```
# Violin plot to show distribution
districts.filtered %>% 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")
```

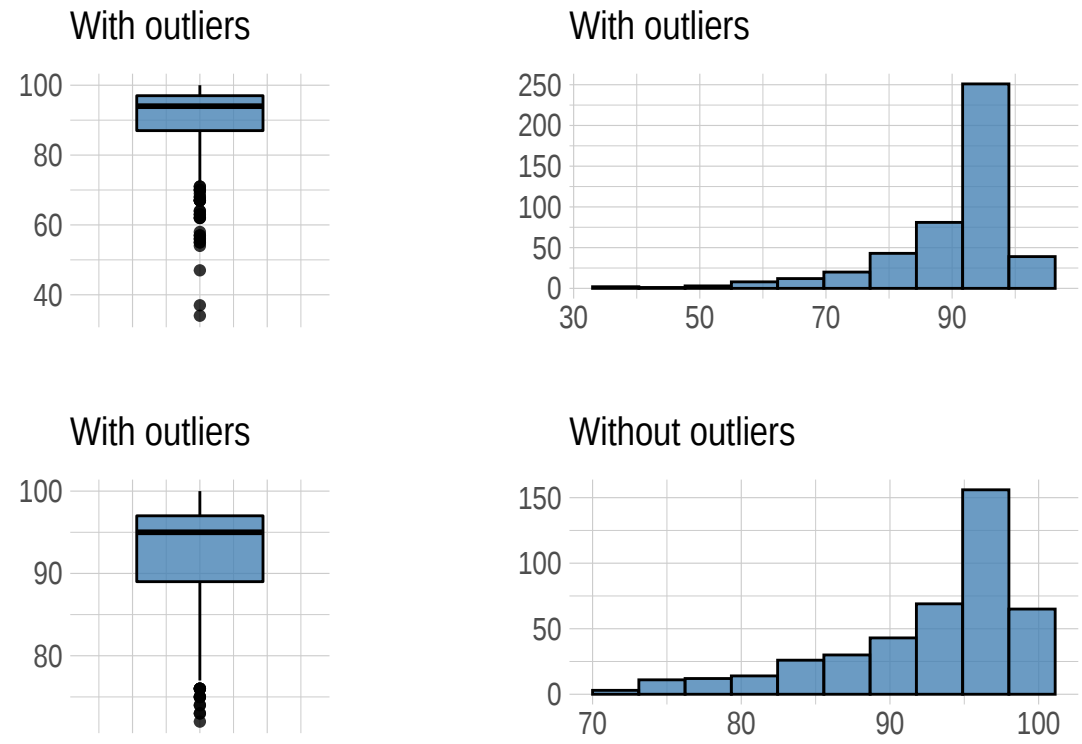



```
# looking for outlier
diagnose_outlier(districts.filtered)
```

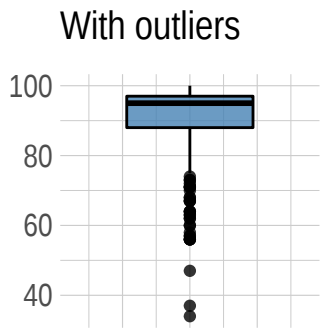
```
##           variables outliers_cnt outliers_ratio outliers_mean   with_mean
## 1           WithDTP           31       6.739130     61.225806  90.4782609
## 2         WithPolio           33       7.173913     62.393939  90.9217391
## 3           WithMMR           33       7.173913     60.878788  90.4500000
## 4          WithHepB           33       7.173913     65.181818  92.6413043
## 5       PctUpToDate           29       6.304348     54.344828  88.5760870
## 6   PctBeliefExempt           41       8.913043     26.121951   5.1869565
## 7  PctMedicalExempt           39       8.478261      1.974359   0.1673913
## 8   PctChildPoverty            7       1.521739     60.000000  22.0239130
## 9  PctFamilyPoverty           16       3.478261     34.500000  11.3347826
## 10      PctFreeMeal            0       0.000000           NaN  48.6260870
## 11         Enrolled           41       8.913043   2843.756098 563.2760870
## 12     TotalSchools           34       7.391304     33.029412   6.5434783
##    without_mean
## 1     92.592075
## 2     93.126464
## 3     92.735363
## 4     94.763466
## 5     90.879350
## 6      3.138425
## 7      0.000000
## 8     21.437086
## 9     10.500000
```

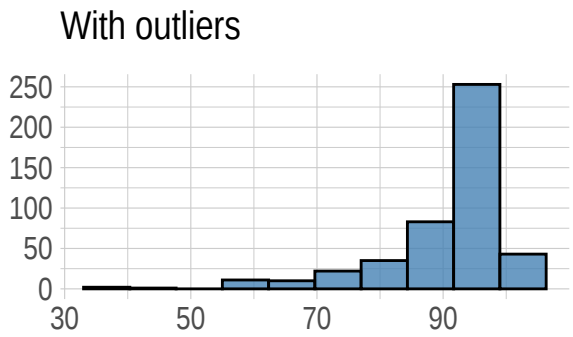
```
## 10    48.626087
## 11   340.126492
## 12     4.429577
```

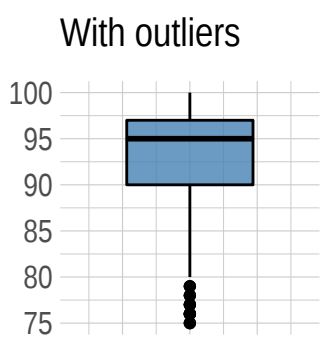
```
# plotting outlier
plot_outlier(districts.filtered)
```

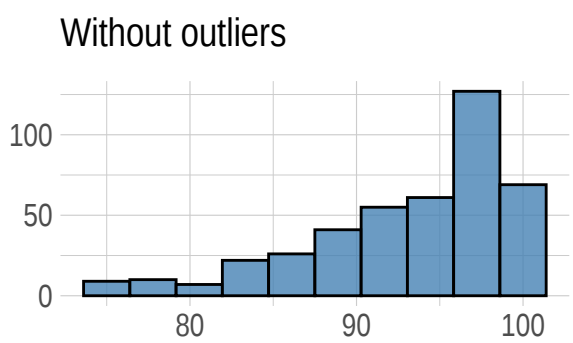


Outlier Diagnosis Plot (WithPolio)



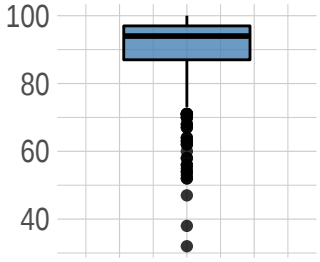




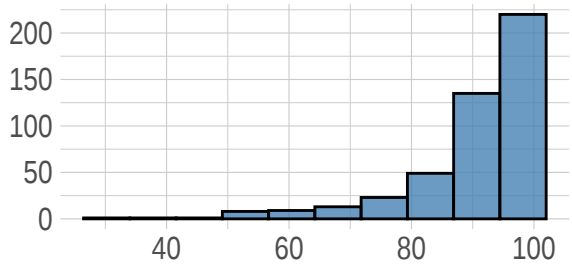


Outlier Diagnosis Plot (WithMMR)

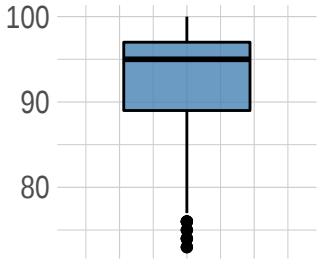
With outliers



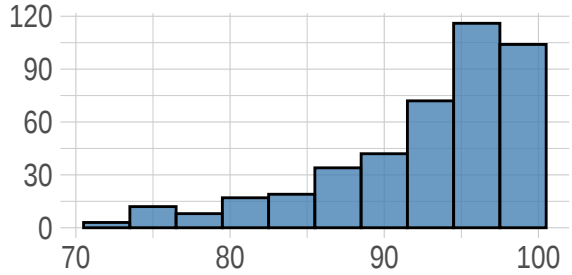
With outliers



With outliers

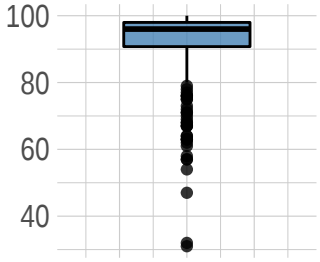


Without outliers

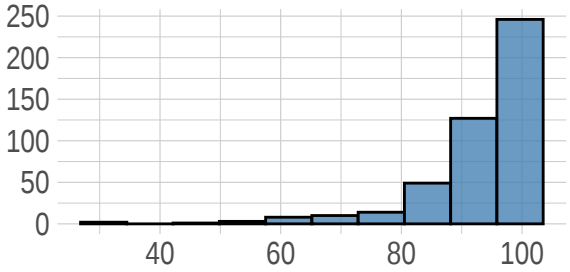


Outlier Diagnosis Plot (WithHepB)

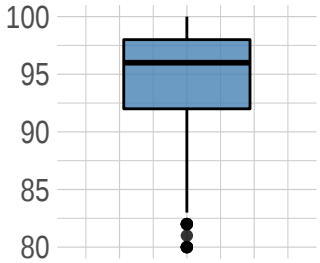
With outliers



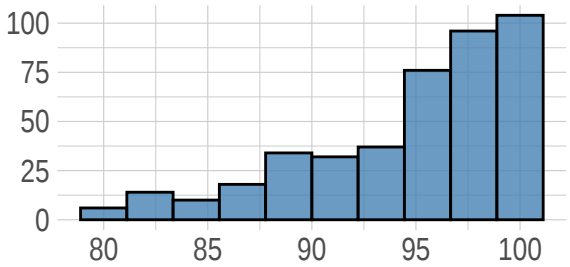
With outliers



With outliers

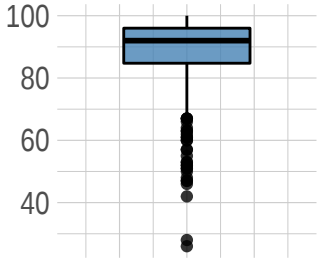


Without outliers

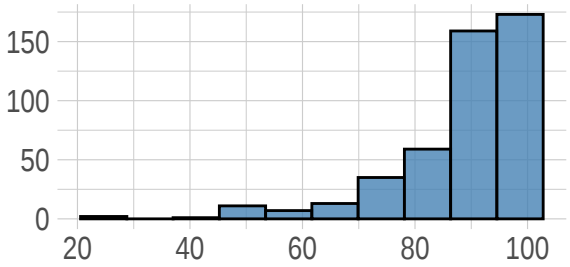


Outlier Diagnosis Plot (PctUpToDate)

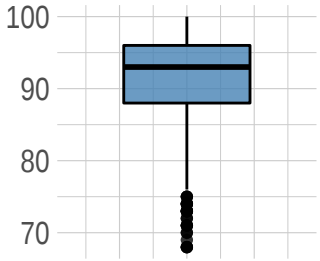
With outliers



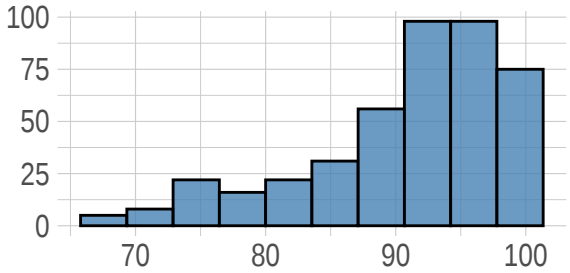
With outliers



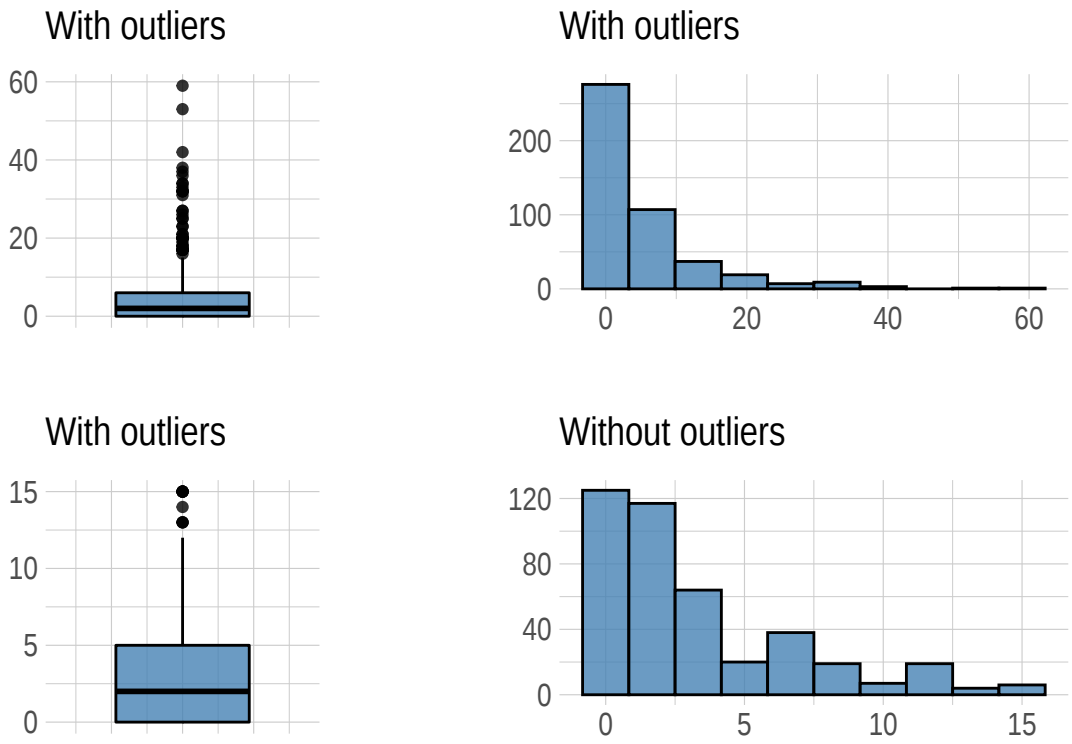
With outliers



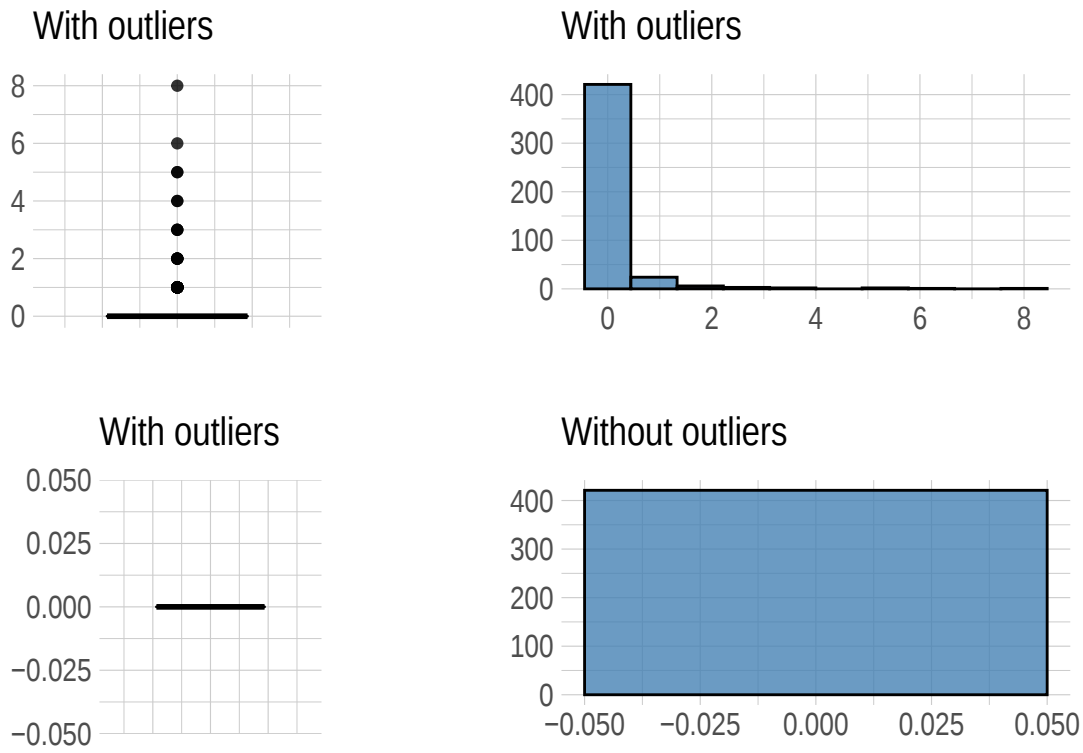
Without outliers



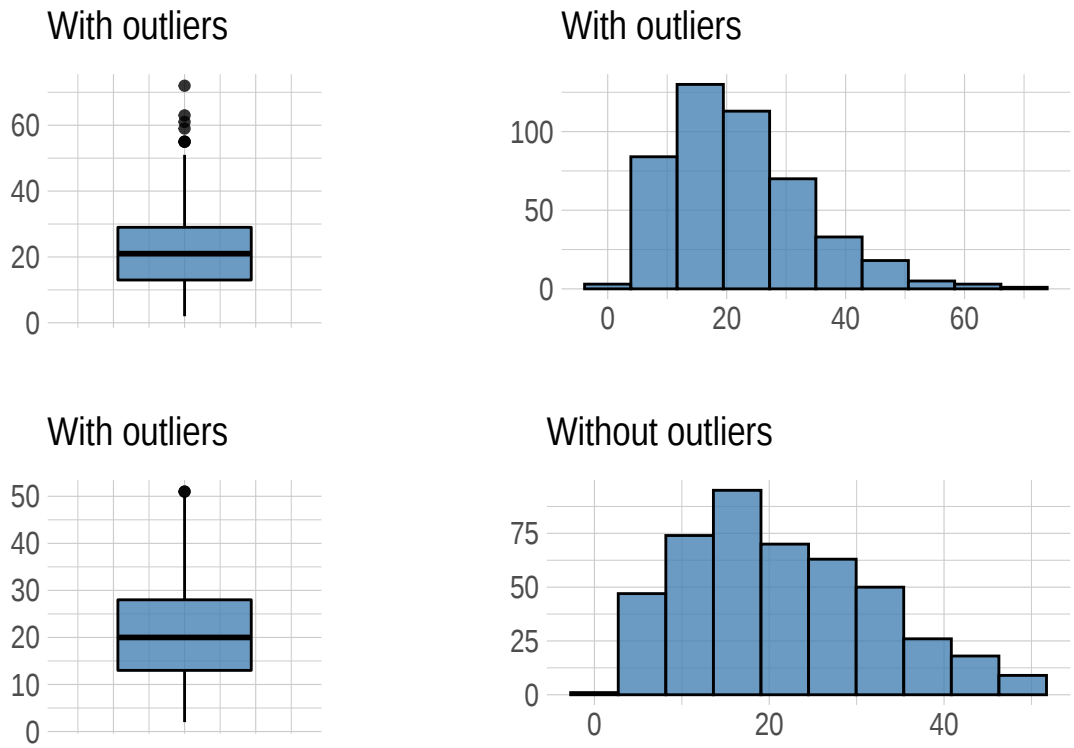
Outlier Diagnosis Plot (PctBeliefExempt)



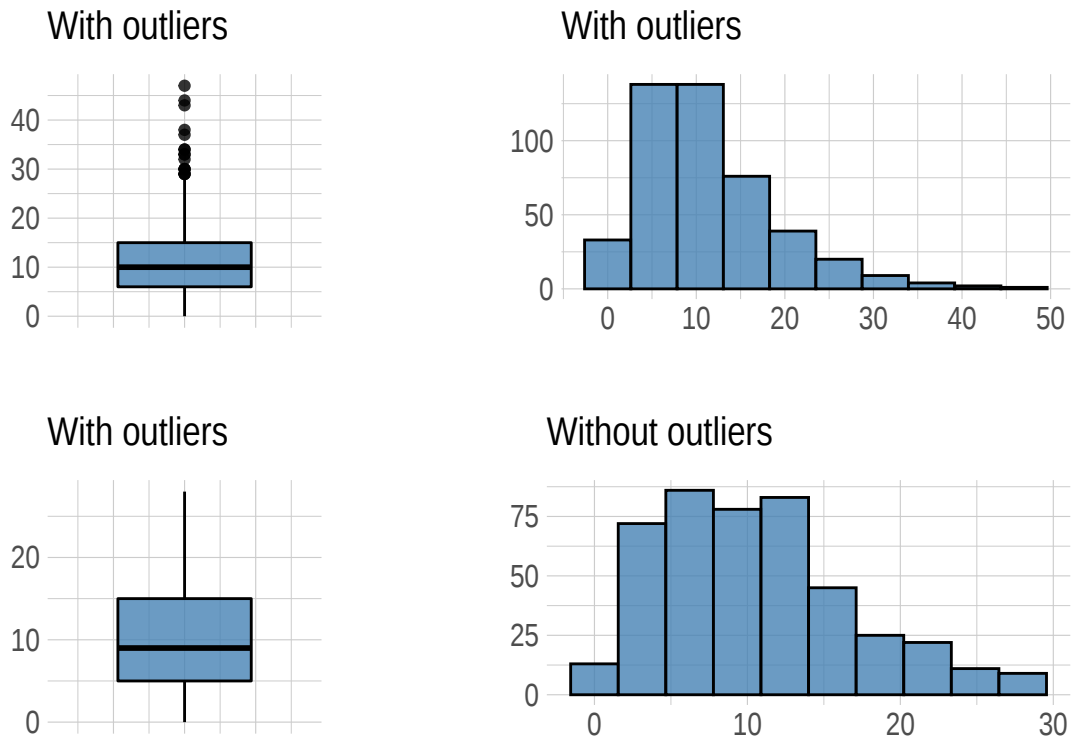
Outlier Diagnosis Plot (PctMedicalExempt)



Outlier Diagnosis Plot (PctChildPoverty)

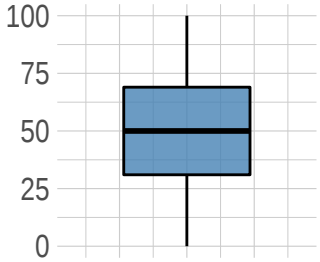


Outlier Diagnosis Plot (PctFamilyPoverty)

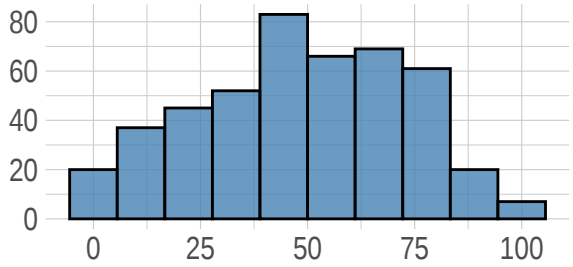


Outlier Diagnosis Plot (PctFreeMeal)

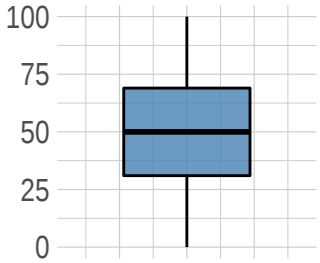
With outliers



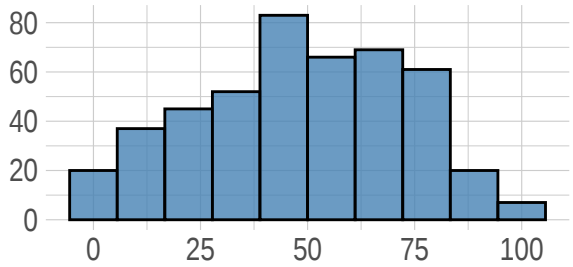
With outliers



With outliers

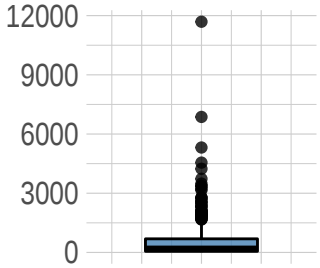


Without outliers

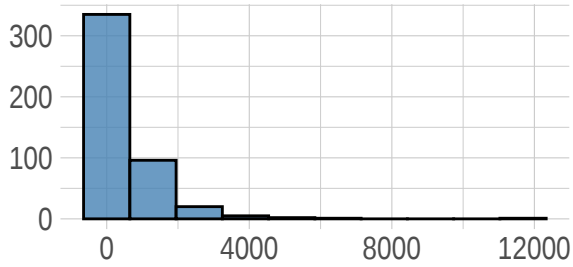


Outlier Diagnosis Plot (Enrolled)

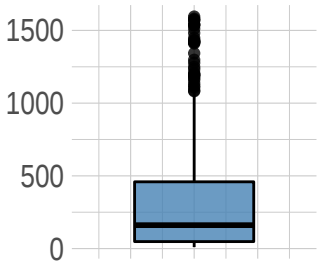
With outliers



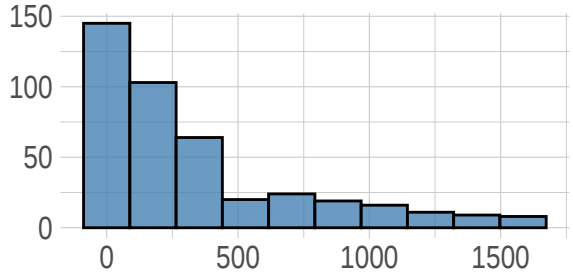
With outliers

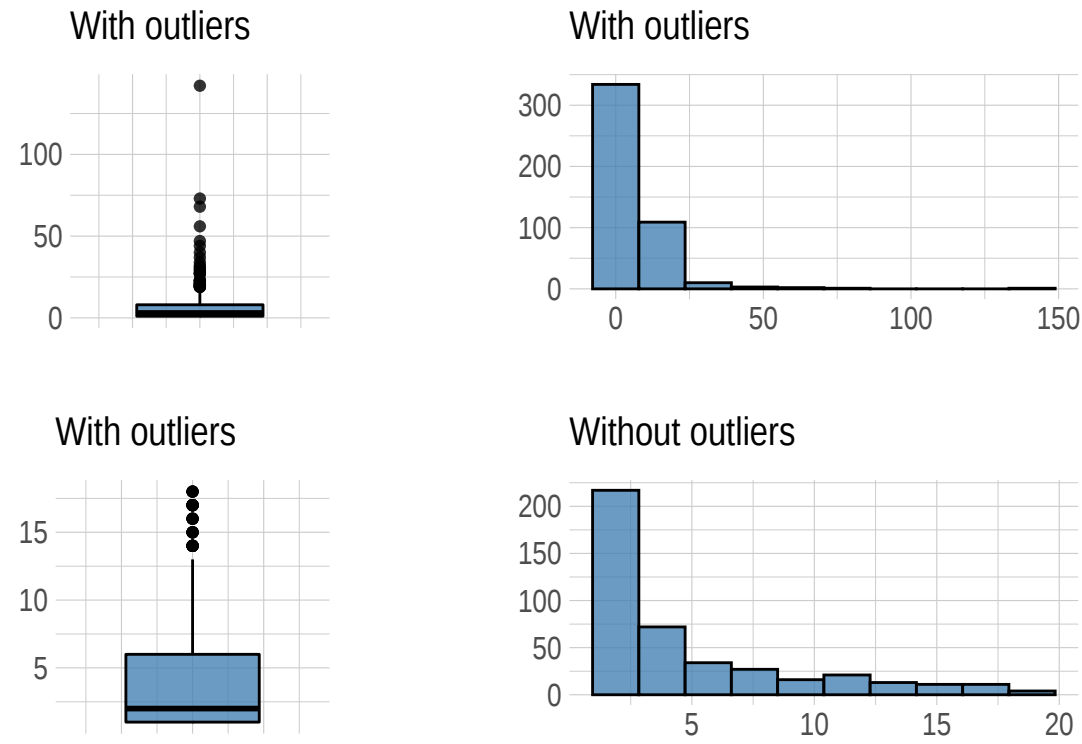


With outliers



Without outliers






```
districts.filtered %>% filter(PctUpToDate < 40)
```

```
##                DistrictName WithDTP WithPolio WithMMR WithHepB PctUpToDate
## 1 Trinidad Union Elementary      37        37      32       32          26
## 2      Lagunitas Elementary      34        34      38       31          28
##   DistrictComplete PctBeliefExempt PctMedicalExempt PctChildPoverty
## 1             TRUE              11                0              19
## 2             TRUE              59                0               2
##   PctFamilyPoverty PctFreeMeal Enrolled TotalSchools
## 1                7          44       19            1
## 2                0          16       29            2
```

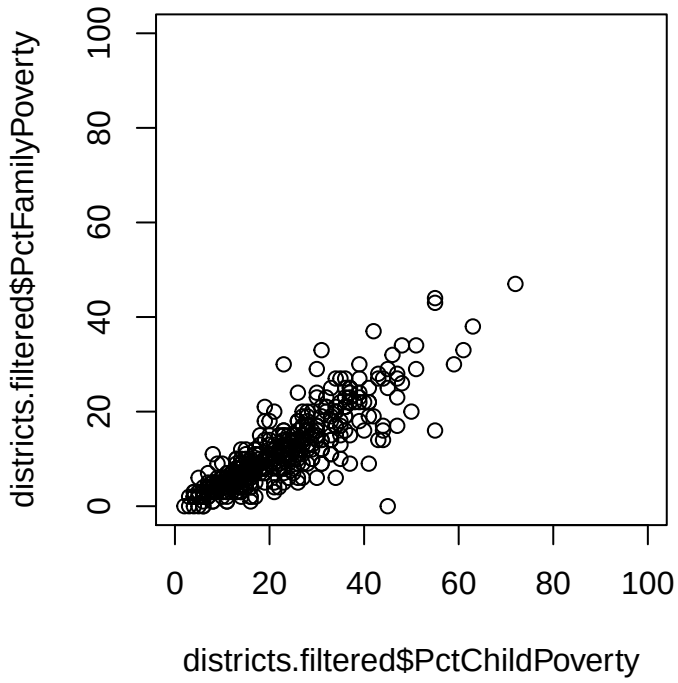
```
districts.diagnose <- districts.filtered %>%
  mutate(WithDTP_PctBelieveExempt   = WithDTP + PctBeliefExempt) %>%
  mutate(WithPolio_PctBelieveExempt = WithPolio + PctBeliefExempt) %>%
  mutate(WithMMR_PctBelieveExempt   = WithMMR + PctBeliefExempt) %>%
  mutate(WithHepB_PctBelieveExempt  = WithHepB + PctBeliefExempt) %>%
  mutate(WithDTP_PctMedicalExempt   = WithDTP + PctMedicalExempt) %>%
  mutate(WithPolio_PctMedicalExempt = WithPolio + PctMedicalExempt) %>%
  mutate(WithMMR_PctMedicalExempt   = WithMMR + PctMedicalExempt) %>%
  mutate(WithHepB_PctMedicalExempt  = WithHepB + PctMedicalExempt)
```

```
summary(districts.diagnose)
```

```
##                              DistrictName    WithDTP         WithPolio
##  Ackerman Charter                   :  1   Min.   : 34.00   Min.   : 34.00
##  Acton-Agua Dulce Unified           :  1   1st Qu.: 87.00   1st Qu.: 88.00
```

```
##  Alameda Unified                  :  1   Median : 94.00   Median : 95.00
##  Albany City Unified              :  1   Mean   : 90.48   Mean   : 90.92
##  Alexander Valley Union Elementary:  1   3rd Qu.: 97.00   3rd Qu.: 97.00
##  Alhambra Unified                 :  1   Max.   :100.00   Max.   :100.00
##  (Other)                          :454
##     WithMMR          WithHepB       PctUpToDate     DistrictComplete
##  Min.   : 32.00   Min.   : 31.00   Min.   : 26.00   Mode :logical
##  1st Qu.: 87.00   1st Qu.: 90.75   1st Qu.: 84.75   FALSE:29
##  Median : 94.00   Median : 96.00   Median : 92.00   TRUE :431
##  Mean   : 90.45   Mean   : 92.64   Mean   : 88.58
##  3rd Qu.: 97.00   3rd Qu.: 98.00   3rd Qu.: 96.00
##  Max.   :100.00   Max.   :100.00   Max.   :100.00
##
##  PctBeliefExempt  PctMedicalExempt PctChildPoverty PctFamilyPoverty
##  Min.   : 0.000   Min.   :0.0000   Min.   : 2.00   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:13.00   1st Qu.: 6.00
##  Median : 2.000   Median :0.0000   Median :21.00   Median :10.00
##  Mean   : 5.187   Mean   :0.1674   Mean   :22.02   Mean   :11.33
##  3rd Qu.: 6.000   3rd Qu.:0.0000   3rd Qu.:29.00   3rd Qu.:15.00
##  Max.   :59.000   Max.   :8.0000   Max.   :72.00   Max.   :47.00
##
##   PctFreeMeal        Enrolled         TotalSchools     WithDTP_PctBelieveExempt
##  Min.   :  0.00   Min.   :   10.00   Min.   :  1.000   Min.   : 48.00
##  1st Qu.: 31.00   1st Qu.:   55.75   1st Qu.:  1.000   1st Qu.: 94.00
##  Median : 50.00   Median :  202.00   Median :  3.000   Median : 98.00
##  Mean   : 48.63   Mean   :  563.28   Mean   :  6.543   Mean   : 95.67
##  3rd Qu.: 69.00   3rd Qu.:  688.50   3rd Qu.:  8.000   3rd Qu.:100.00
##  Max.   :100.00   Max.   :11691.00   Max.   :142.000   Max.   :100.00
##
##  WithPolio_PctBelieveExempt WithMMR_PctBelieveExempt WithHepB_PctBelieveExempt
##  Min.   : 48.00             Min.   : 43.00           Min.   : 43.00
##  1st Qu.: 95.00             1st Qu.: 94.75           1st Qu.: 98.00
##  Median : 98.00             Median : 98.00           Median : 99.00
##  Mean   : 96.11             Mean   : 95.64           Mean   : 97.83
##  3rd Qu.:100.00             3rd Qu.:100.00           3rd Qu.:100.00
##  Max.   :100.00             Max.   :100.00           Max.   :100.00
##
##  WithDTP_PctMedicalExempt WithPolio_PctMedicalExempt WithMMR_PctMedicalExempt
##  Min.   : 34.00           Min.   : 34.00             Min.   : 32.00
##  1st Qu.: 87.00           1st Qu.: 89.00             1st Qu.: 87.00
##  Median : 94.00           Median : 95.00             Median : 94.00
##  Mean   : 90.65           Mean   : 91.09             Mean   : 90.62
##  3rd Qu.: 97.00           3rd Qu.: 97.00             3rd Qu.: 97.00
##  Max.   :100.00           Max.   :100.00             Max.   :100.00
##
##  WithHepB_PctMedicalExempt
##  Min.   : 31.00
##  1st Qu.: 91.00
##  Median : 96.00
##  Mean   : 92.81
##  3rd Qu.: 98.00
##  Max.   :100.00
##
```

```
par(pty="s")
plot(districts.filtered$PctChildPoverty, districts.filtered$PctFamilyPoverty,
ylim = c(0,100),
xlim = c(0,100))
```




```
lmout <- lm(PctChildPoverty ~ PctFamilyPoverty, districts.filtered)
summary(lmout)
```

```
##
## Call:
## lm(formula = PctChildPoverty ~ PctFamilyPoverty, data = districts.filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.924  -3.726  -0.656   2.405  37.490
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.50992    0.49944   15.04   <2e-16 ***
## PctFamilyPoverty  1.28048    0.03631   35.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.068 on 458 degrees of freedom
## Multiple R-squared:  0.7308, Adjusted R-squared:  0.7303
## F-statistic:  1244 on 1 and 458 DF,  p-value: < 2.2e-16
```

Descriptive Reporting

1. *Basic Introductory Paragraph*

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

It is a fact that low vaccination rate will bring back dangerous disease such as Diphtheria, Polio, Measles and Hepatitis. This report will analysis the vaccination rate in California public school, how poverty, religious and medical factor effects them.

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?*

```
usVaccines.DTP1 <- as.data.frame(usVaccines)$DTP1 %>% ts( start= c(1980), frequency=1)
usVaccines.HepB_BD <- as.data.frame(usVaccines)$HepB_BD %>% ts( start= c(1980), frequency=1)
usVaccines.Pol3 <- as.data.frame(usVaccines)$Pol3 %>% ts( start= c(1980), frequency=1)
usVaccines.Hib3 <- as.data.frame(usVaccines)$Hib3 %>% ts( start= c(1980), frequency=1)
usVaccines.MCV1 <- as.data.frame(usVaccines)$MCV1 %>% ts( start= c(1980), frequency=1)

par(mfrow=c(5,2),
    oma=c(2,2,0,0)+0.1,
    mar=c(1,3,1,0)+0.1,
    mgp=c(2,1,0), font.lab=2, cex.lab=1.1)

usVaccines.DTP1.diff <-  diff(usVaccines.DTP1)
usVaccines.DTP1.cp.mean <- cpt.mean(usVaccines.DTP1)
usVaccines.DTP1.cp.mean %>% plot(main="Mean Changepoint Plot", ylab = "DTP")
usVaccines.DTP1.cp.mean
```

```
## Class 'cpt' : Changepoint Object
##         ~~   : S4 class containing 12 slots with names
##                cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sat May 15 18:39:34 2021
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 10
```

```
usVaccines.DTP1.cp.var <- cpt.var(usVaccines.DTP1.diff)
usVaccines.DTP1.cp.var %>% plot(main="Variance Difference Plot",ylab = "")
usVaccines.DTP1.cp.var
```

```
## Class 'cpt' : Changepoint Object
```

```
##        ~~   : S4 class containing 12 slots with names
##               cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sat May 15 18:39:34 2021
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.83275
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 10
```

```
usVaccines.HepB_BD.diff <-  diff(usVaccines.HepB_BD)
usVaccines.HepB_BD.cp.mean <- cpt.mean(usVaccines.HepB_BD)
usVaccines.HepB_BD.cp.mean %>% plot(ylab = "HepB_BD")
usVaccines.HepB_BD.cp.mean
```

```
## Class 'cpt' : Changepoint Object
##        ~~   : S4 class containing 12 slots with names
##               cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sat May 15 18:39:34 2021
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 24
```

```
usVaccines.HepB_BD.cp.var <- cpt.var(usVaccines.HepB_BD.diff)
usVaccines.HepB_BD.cp.var %>% plot(ylab = "")
usVaccines.HepB_BD.cp.var
```

```
## Class 'cpt' : Changepoint Object
##        ~~   : S4 class containing 12 slots with names
##               cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sat May 15 18:39:34 2021
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.83275
```

```
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations :
```

```
usVaccines.Pol3.diff <-  diff(usVaccines.Pol3)
usVaccines.Pol3.cp.mean <- cpt.mean(usVaccines.Pol3)
usVaccines.Pol3.cp.mean %>% plot(ylab = "Pol3")
usVaccines.Pol3.cp.mean
```

```
## Class 'cpt' : Changepoint Object
##         ~~   : S4 class containing 12 slots with names
##                cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sat May 15 18:39:34 2021
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 15
```

```
usVaccines.Pol3.cp.var <- cpt.var(usVaccines.Pol3.diff)
usVaccines.Pol3.cp.var %>% plot(ylab = "")
usVaccines.Pol3.cp.var
```

```
## Class 'cpt' : Changepoint Object
##         ~~   : S4 class containing 12 slots with names
##                cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sat May 15 18:39:34 2021
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.83275
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 16
```

```
usVaccines.Hib3.diff <-  diff(usVaccines.Hib3)
usVaccines.Hib3.cp.mean <- cpt.mean(usVaccines.Hib3)
usVaccines.Hib3.cp.mean %>% plot(ylab = "Hib3")
usVaccines.Hib3.cp.mean
```

```
## Class 'cpt' : Changepoint Object
##         ~~   : S4 class containing 12 slots with names
##                cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
```

```
## Created on  : Sat May 15 18:39:34 2021
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 8
```

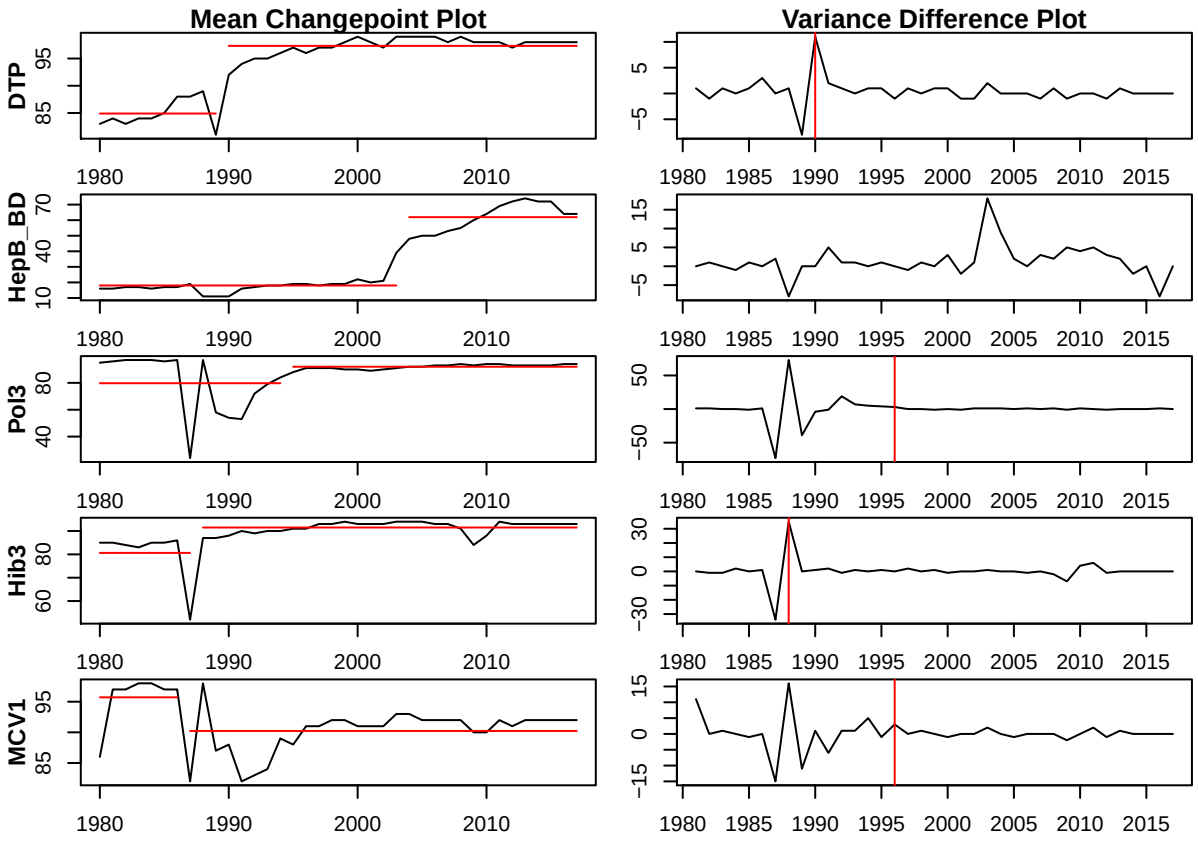
```
usVaccines.Hib3.cp.var <- cpt.var(usVaccines.Hib3.diff)
usVaccines.Hib3.cp.var %>% plot(ylab = "")
usVaccines.Hib3.cp.var
```

```
## Class 'cpt' : Changepoint Object
##         ~~   : S4 class containing 12 slots with names
##                cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sat May 15 18:39:34 2021
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.83275
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 8
```

```
usVaccines.MCV1.diff <-  diff(usVaccines.MCV1)
usVaccines.MCV1.cp.mean <- cpt.mean(usVaccines.MCV1)
usVaccines.MCV1.cp.mean %>% plot(ylab = "MCV1")
usVaccines.MCV1.cp.mean
```

```
## Class 'cpt' : Changepoint Object
##         ~~   : S4 class containing 12 slots with names
##                cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sat May 15 18:39:34 2021
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 7
```

```
usVaccines.MCV1.cp.var <- cpt.var(usVaccines.MCV1.diff)
usVaccines.MCV1.cp.var %>% plot(ylab = "")
```




```
usVaccines.MCV1.cp.var
```

```
## Class 'cpt' : Changepoint Object
##         ~~   : S4 class containing 12 slots with names
##                cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Sat May 15 18:39:34 2021
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.83275
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 16
```

Fig 1 - US average Vaccination rate(in percentages) on Diphtheria/Pertussis/Tetanus(DTP), Hepatitis B(HepB_BD), Polio(Pol3), Influenza(Hib3) and Measles(MCV1)

Polio, Hib, MMR all have a great vaccination rate above 80 percent, and are relatively consistent, except at around 1985 - 1990 there is an drop in data. HepB_BD,and DTP had seen a increase from 85 to 95 and

from 20 to 60.

```
summary(usVaccines.Hib3.cp.mean)
```

```
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 8
```

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

```
par(mfrow=c(2,3),mar=c(2,2,4,1)+0.1)

acf(usVaccines.DTP1)
acf(usVaccines.HepB_BD)
acf(usVaccines.Pol3)
acf(usVaccines.Hib3)
acf(usVaccines.MCV1)
```

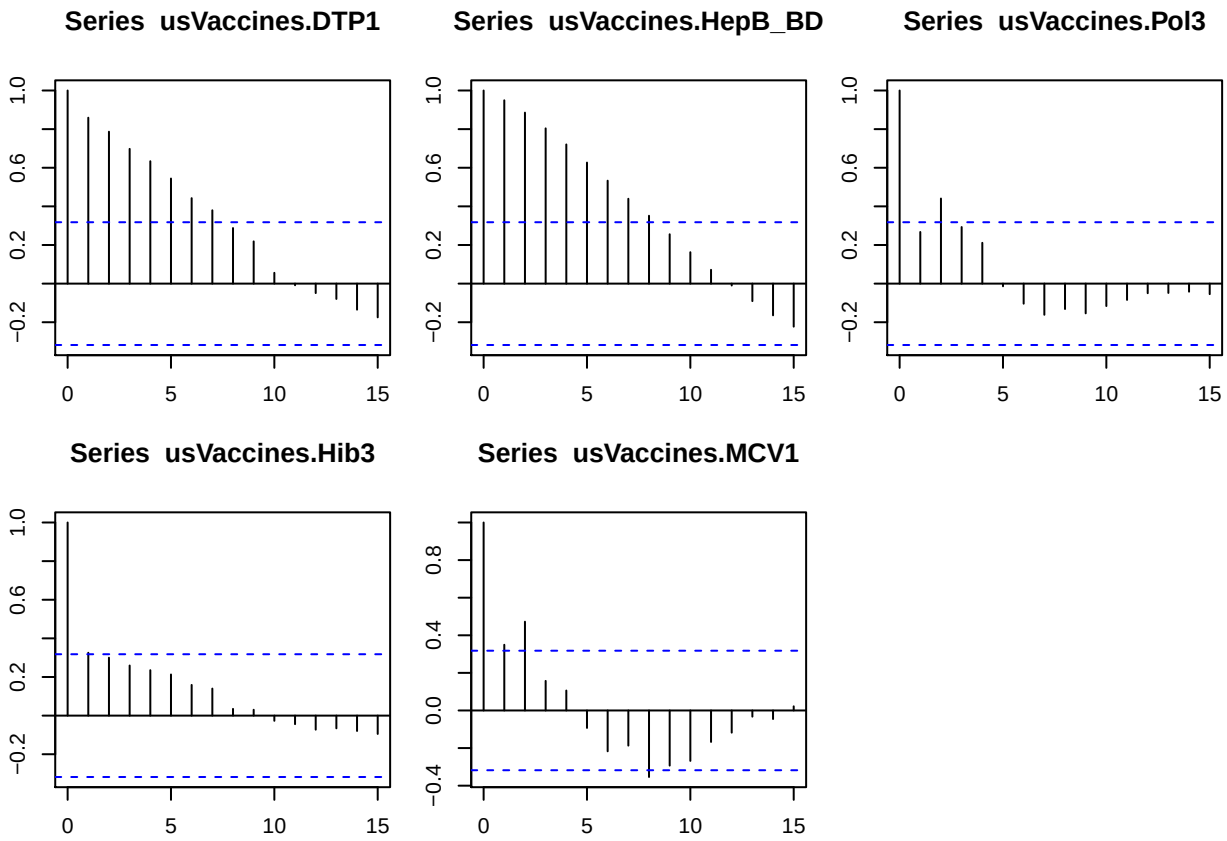


Fig 2.1 - Trend analysis On 5 type of vaccination rate in the US

```
par(mfrow=c(2,3),mar=c(2,2,4,1)+0.1)

acf(usVaccines.DTP1.diff)
```

```
acf(usVaccines.HepB_BD.diff)
acf(usVaccines.Pol3.diff)
acf(usVaccines.Hib3.diff)
acf(usVaccines.MCV1.diff)
```

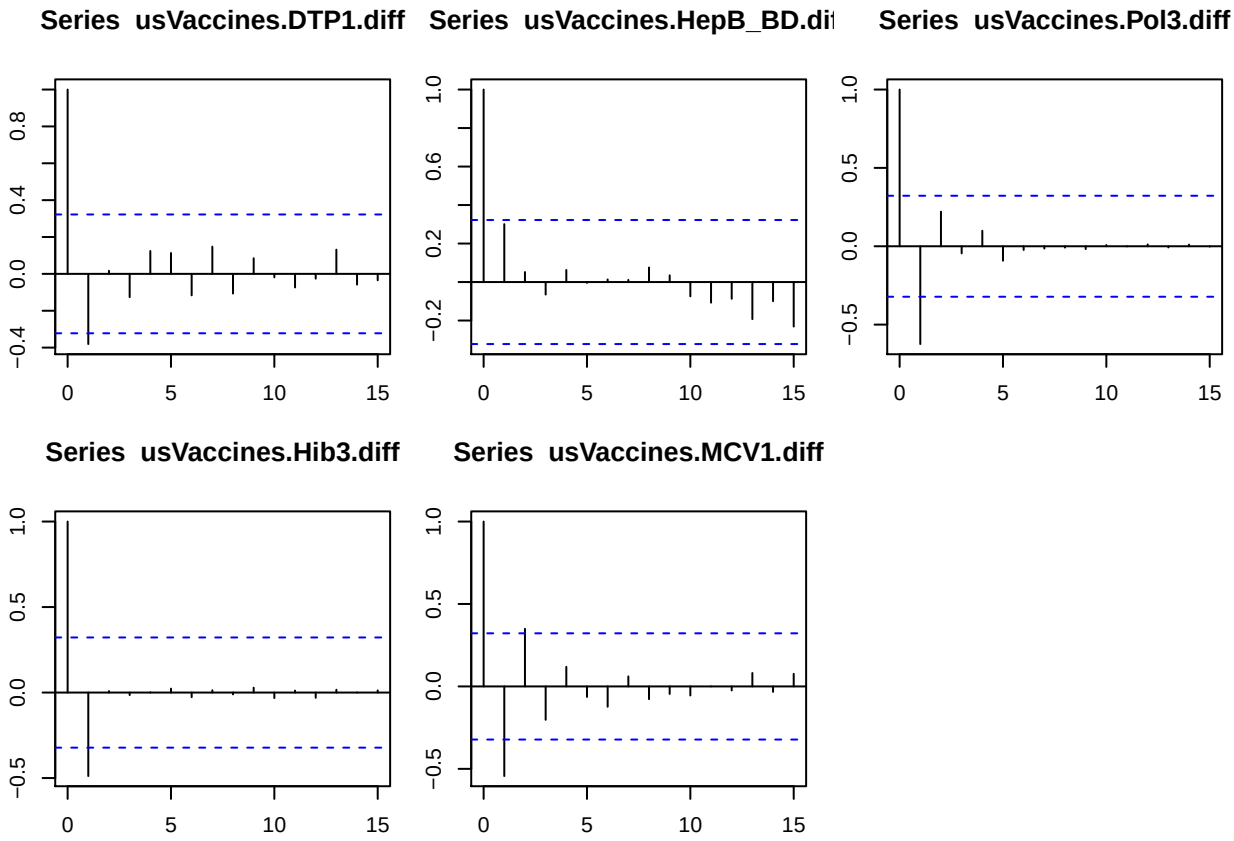


Fig 2.2 - Seasonality analysis On 5 type of vaccination rate in the US

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)?***

```
usVaccines.recent <- window(usVaccines, start = c(1980,24))
as.data.frame(usVaccines.recent) %>% colMeans()
```

```
##     DTP1  HepB_BD     Pol3     Hib3     MCV1
## 98.26667 60.40000 93.06667 92.20000 91.80000
```

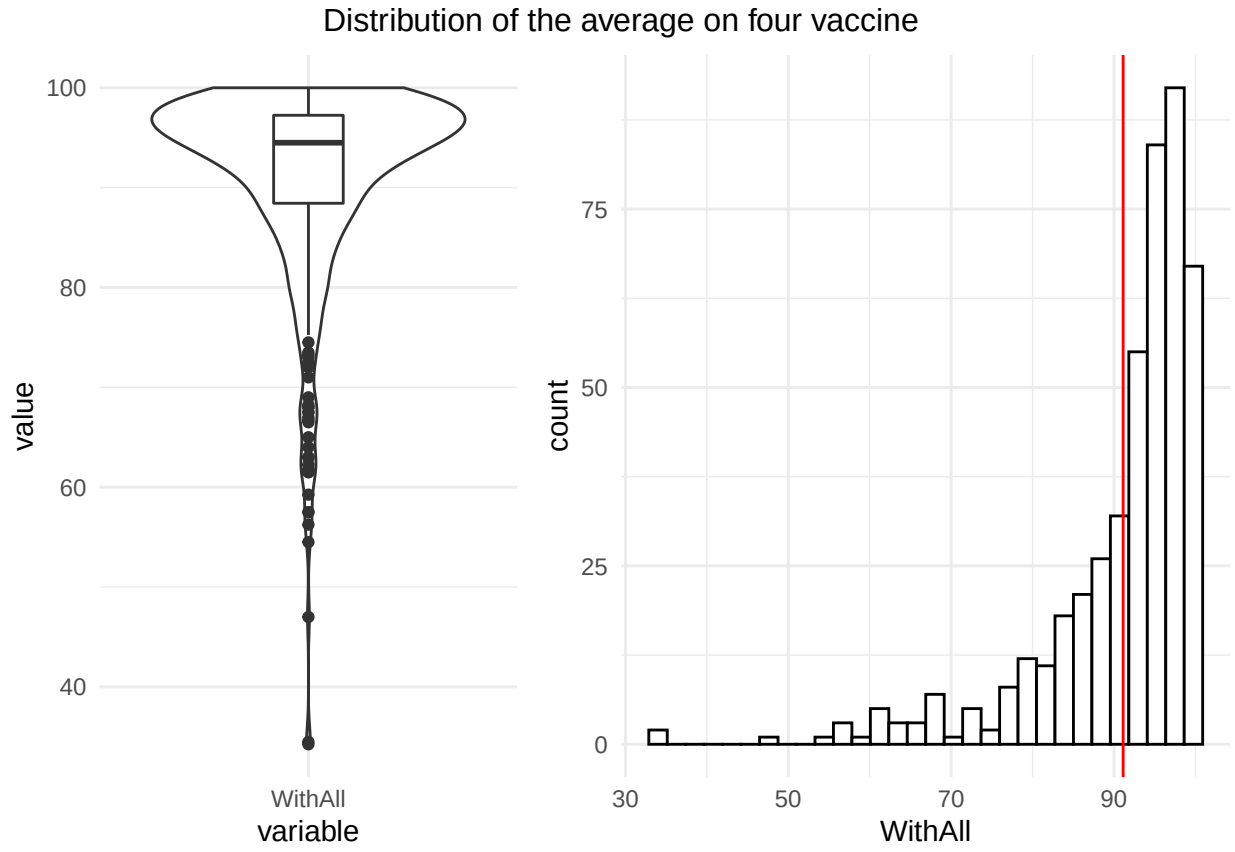
3. *Descriptive Overview of California Vaccinations*

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?***

```
districts.filtered.withall <- districts.filtered %>% mutate(WithAll = (WithDTP + WithPolio + WithMMR + W

p1 <- districts.filtered.withall %>%
```

```
  pivot_longer(cols=c(WithAll) , names_to = "variable",
                          values_to = "value", values_drop_na = TRUE) %>%
  ggplot(aes(x = variable, y = value)) + geom_violin() + geom_boxplot(width = 0.2) +
    theme_minimal()

p2 <- districts.filtered.withall %>%
  ggplot(aes(x = WithAll)) + geom_histogram(color="black", fill="white", bins=30) +
    geom_vline(aes(xintercept = mean(WithAll)),color="red") +
    theme_minimal()

grid.arrange(p1,p2,nrow = 1,widths = c(1.5,2),
  top = "Distribution of the average on four vaccine")
```




```
describe(districts.filtered)
```

```
## # A tibble: 12 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      460     0  90.5    10.3    0.480   10      -2.08     5.32     34
##  2 WithPolio    460     0  90.9    10.1    0.473    9      -2.19     5.90     34
##  3 WithMMR      460     0  90.4    10.6    0.495   10      -2.12     5.34     32
##  4 WithHepB     460     0  92.6     9.49   0.442    7.25   -2.79    10.4      31
##  5 PctUpToDa~   460     0  88.6    11.8    0.551   11.2    -2.04     5.00     26
##  6 PctBelief~   460     0   5.19    7.97   0.372    6       2.89    10.7       0
##  7 PctMedica~   460     0   0.167   0.729  0.0340   0       6.40    49.4       0
##  8 PctChildP~   460     0  22.0    11.7    0.545   16       0.811    0.742     2
```

```
##  9 PctFamily~   460     0  11.3     7.80   0.364    9       1.24     2.02      0
## 10 PctFreeMe~   460     0  48.6    24.1    1.12    38      -0.138   -0.855     0
## 11 Enrolled     460     0 563.    964.    45.0    633.      5.14    44.2      10
## 12 TotalScho~   460     0   6.54   10.8    0.502    7       6.10    60.6       1
## # ... with 16 more variables: p01 <dbl>, p05 <dbl>, p10 <dbl>, p20 <dbl>,
## #   p25 <dbl>, p30 <dbl>, p40 <dbl>, p50 <dbl>, p60 <dbl>, p70 <dbl>,
## #   p75 <dbl>, p80 <dbl>, p90 <dbl>, p95 <dbl>, p99 <dbl>, p100 <dbl>
```

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?***

```
districts.filtered %>% dplyr::select(WithDTP, WithPolio, WithMMR, WithHepB) %>% pairs.panels()
```

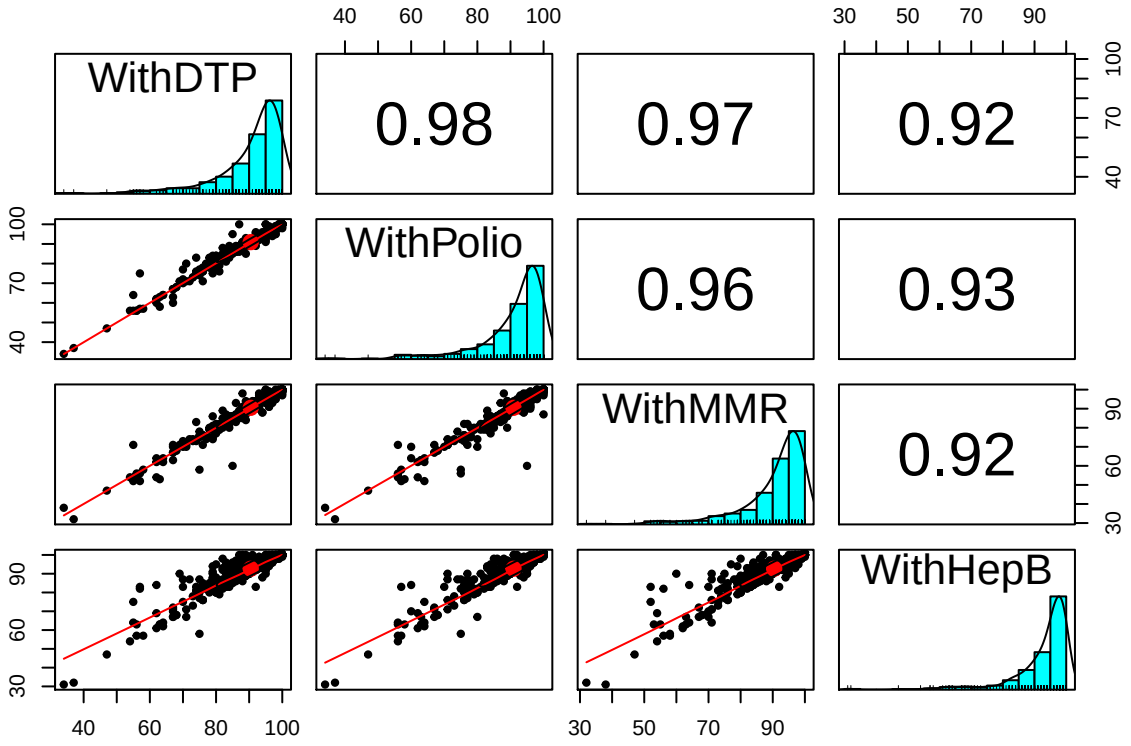


Fig 3.1 - Four Type of Vaccination Correlation in California

```
districts.filtered %>% dplyr::select(WithDTP, WithPolio, WithMMR, WithHepB) %>% .^6 %>% pairs.panels()
```

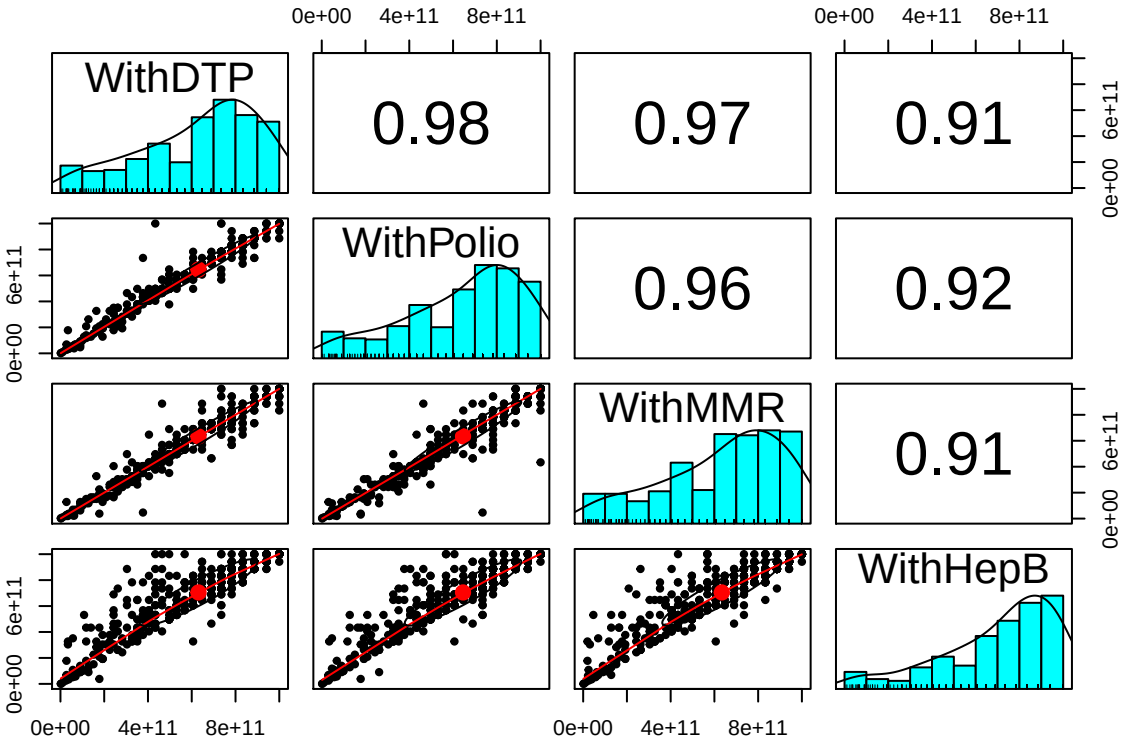


Fig 3.2 - Four Type of Vaccination Correlation in California after power transformation

```
districts.filtered.flattened <- districts.filtered %>%
  pivot_longer(cols=c(WithDTP, WithPolio, WithMMR, WithHepB),
               names_to = "vaccineType",
               values_to = "value")

ggplot(districts.filtered.flattened,aes(x=vaccineType,y=value)) + geom_boxplot() +
  theme_minimal()
```

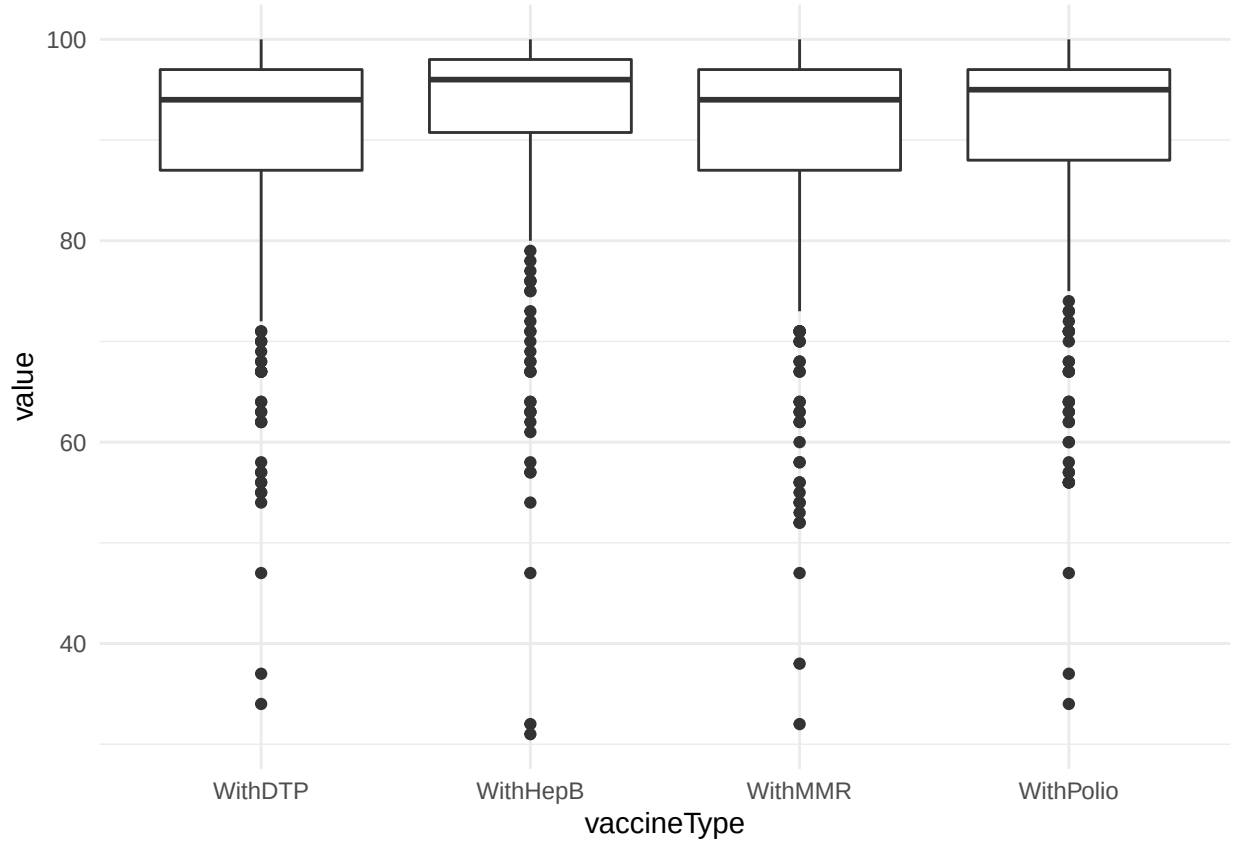


Fig 3.3 - Box plots of Vaccination rate in California

```
aovOut <- aov(value^6 ~ vaccineType, districts.filtered.flattened)
summary(aovOut)
```

```
##               Df    Sum Sq   Mean Sq F value  Pr(>F)
## vaccineType    3 1.654e+24 5.514e+23   7.774 3.7e-05 ***
## Residuals   1836 1.302e+26 7.092e+22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The summary of the model have a p-value of 0.000037. We can confidently reject the null hypothesis and say there is a different of mean between these 4 groups Although we do not know which group are different, or even it's possible that all 4 groups are different. Cobain with the correlation graph(Fig 3.2) and the boxplot(Fig 3.3), my inference is the Hepatitis B vaccination rate are different with the other 3 groups. The Hepatitis B vaccine have a correlation from .92~.93, where the other groups are all above .95.

```
plot(aovOut,which=2)
```

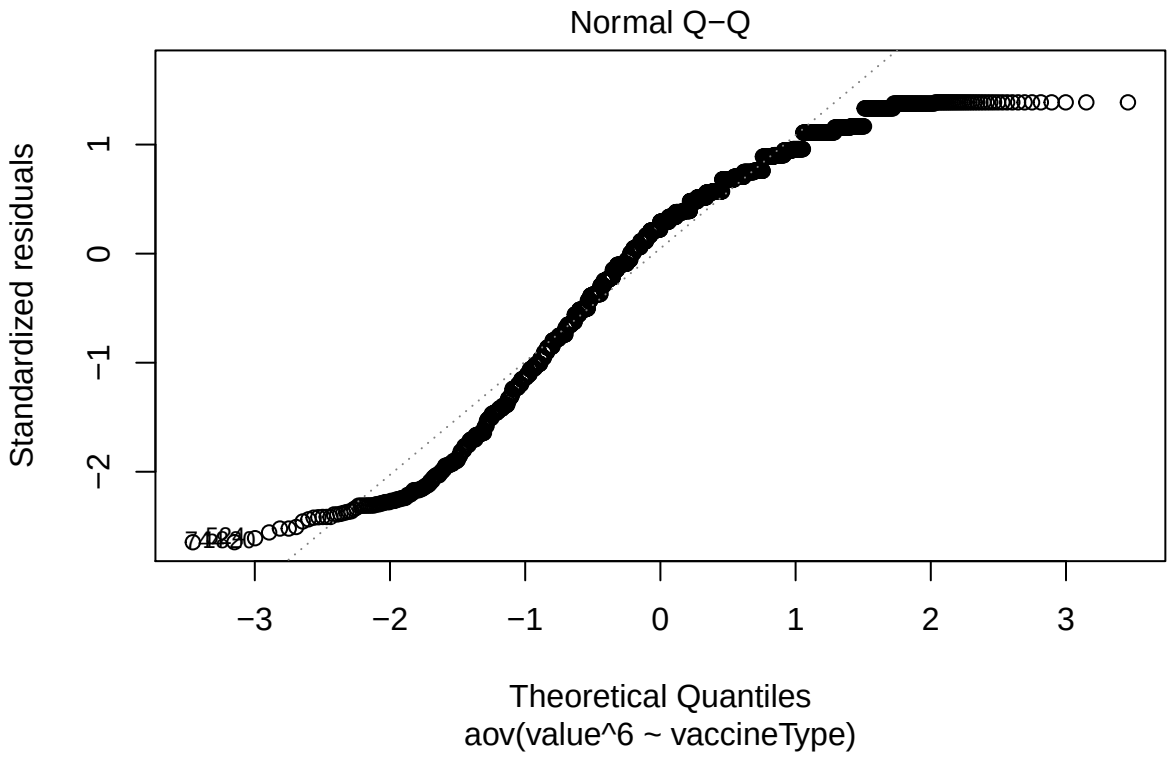


Fig 3.4 - anova model residual after power 6 transformation

```
simulationOutput1 <- simulateResiduals(fittedModel = aovOut, n = 250)
```

```
## Warning in checkModel(fittedModel): DHARMa: fittedModel not in class of
## supported models. Absolutely no guarantee that this will work!
```

```
plot(simulationOutput1)
```

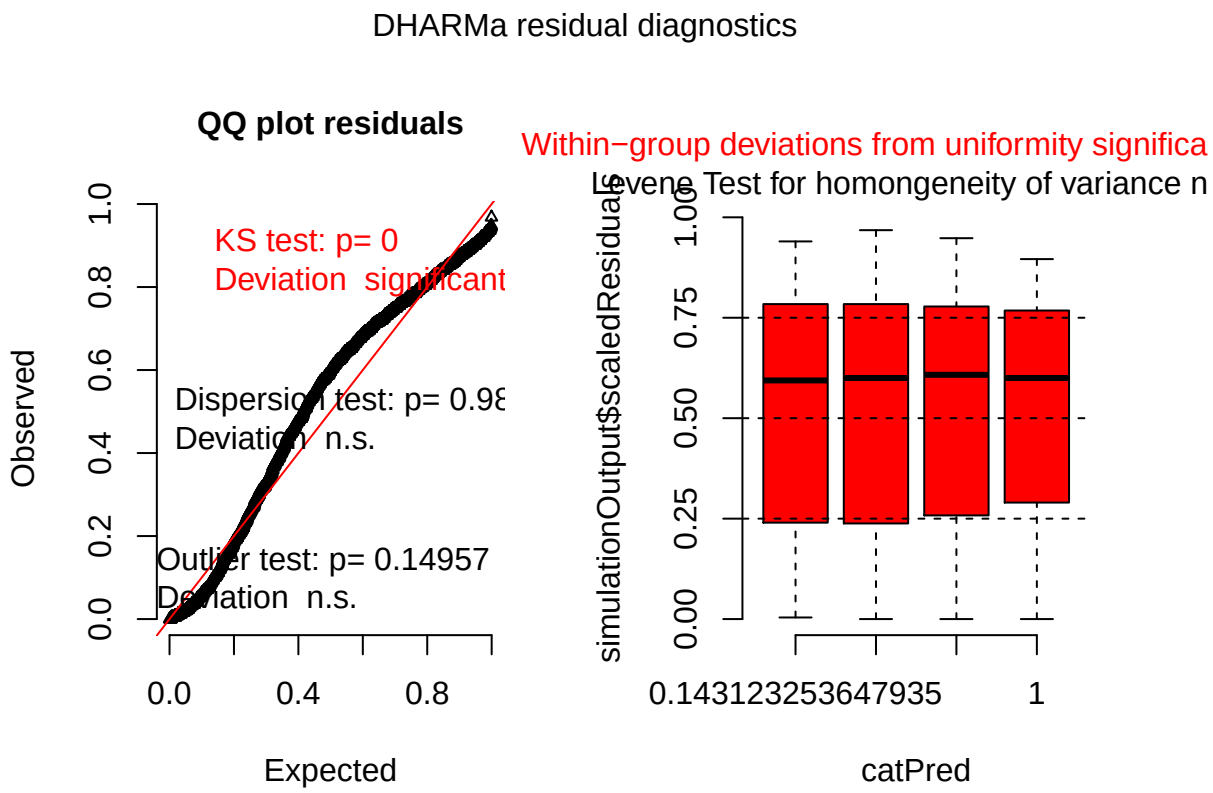


Fig 3.5 - DHARMa simulation of the anova model

The simulation showed mostly normal distribution of the residuals

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

```
usVaccines.last.DTP1 <- tail(usVaccines, n=1)[,1]
usVaccines.last.HepB_BD <- tail(usVaccines, n=1)[,2]
usVaccines.last.Pol3 <- tail(usVaccines, n=1)[,3]
usVaccines.last.Hib3 <- tail(usVaccines, n=1)[,4]
usVaccines.last.MCV1 <- tail(usVaccines, n=1)[,5]

usVaccines.recent.DTP1 <- as.data.frame(usVaccines.recent)$DTP1 %>% mean()
usVaccines.recent.HepB_BD <- as.data.frame(usVaccines.recent)$HepB_BD %>% mean()
usVaccines.recent.Pol3 <- as.data.frame(usVaccines.recent)$Pol3 %>% mean()
usVaccines.recent.Hib3 <- as.data.frame(usVaccines.recent)$Hib3 %>% mean()
usVaccines.recent.MCV1 <- as.data.frame(usVaccines.recent)$MCV1 %>% mean()
```

```
districts.filtered.averages <- districts.filtered %>%
  pivot_longer(cols=c(WithDTP, WithPolio, WithHepB, WithMMR) , names_to = "variable",
                          values_to = "value", values_drop_na = TRUE) %>%
  group_by(variable) %>%
  summarise(mean = mean(value), median = median(value)) %>%
  add_column(usRecentMean = c(usVaccines.recent.DTP1,
                              usVaccines.recent.HepB_BD,
                              usVaccines.recent.MCV1,
```

```
                                usVaccines.recent.Pol3))
```

```
districts.filtered.averages
```

```
## # A tibble: 4 x 4
##   variable   mean median usRecentMean
##   <chr>     <dbl>  <dbl>        <dbl>
## 1 WithDTP    90.5     94         98.3
## 2 WithHepB   92.6     96         60.4
## 3 WithMMR    90.4     94         91.8
## 4 WithPolio  90.9     95         93.1
```

```
districts.filtered %>%
  pivot_longer(cols=c(WithDTP, WithPolio, WithHepB, WithMMR) , names_to = "variable",
                          values_to = "value", values_drop_na = TRUE) %>%
  ggplot(aes(x = variable, y = value)) + geom_violin() + facet_wrap(~variable, scales="free") +
    geom_hline(data = districts.filtered.averages, aes(yintercept = median, color = 'CA Median')) +
    geom_hline(data = districts.filtered.averages, aes(yintercept = mean, color = 'CA Mean')) +
    geom_hline(data = districts.filtered.averages, aes(yintercept = usRecentMean, color = 'US Mean', li
    scale_linetype_identity() +
    theme_minimal() +
    theme(legend.title = element_blank()) +
    scale_color_manual(values=c('red','orange','blue'))
```

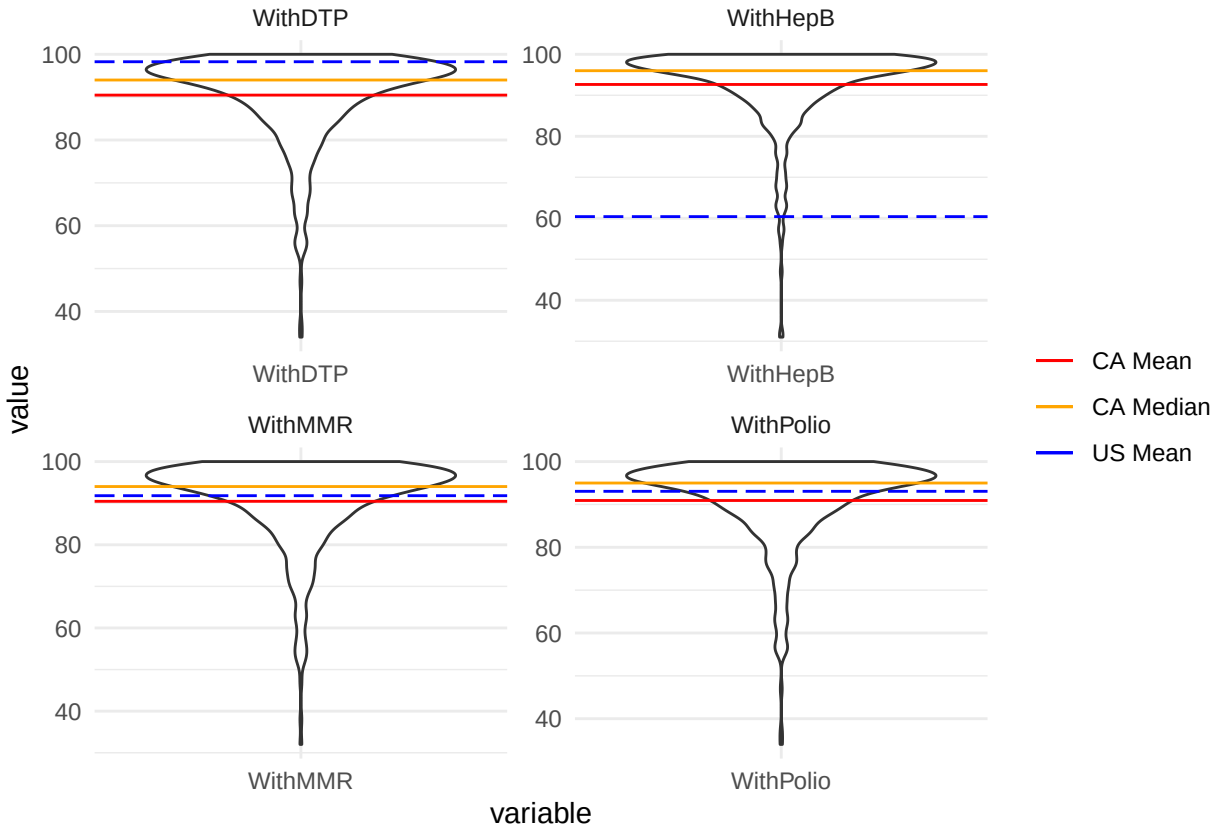


Fig 4.1 - Violin plot on 4 types of vaccination rate in the CA distracts with California average(Red), California Median(Orange) and US average(Blue)

```
districts.filtered %>%
  pivot_longer(cols=c(WithDTP, WithPolio, WithHepB, WithMMR) , names_to = "variable",
                           values_to = "value", values_drop_na = TRUE) %>%
  ggplot(aes(x = variable, y = value)) + geom_boxplot() + facet_wrap(~variable, scales="free") +
    geom_hline(data = districts.filtered.averages, aes(yintercept = mean, color = 'CA Mean')) +
    geom_hline(data = districts.filtered.averages, aes(yintercept = usRecentMean, color = 'US Mean', li
    scale_linetype_identity() +
    theme_minimal() +
    theme(legend.title = element_blank()) +
    scale_color_manual(values=c('red','blue'))
```

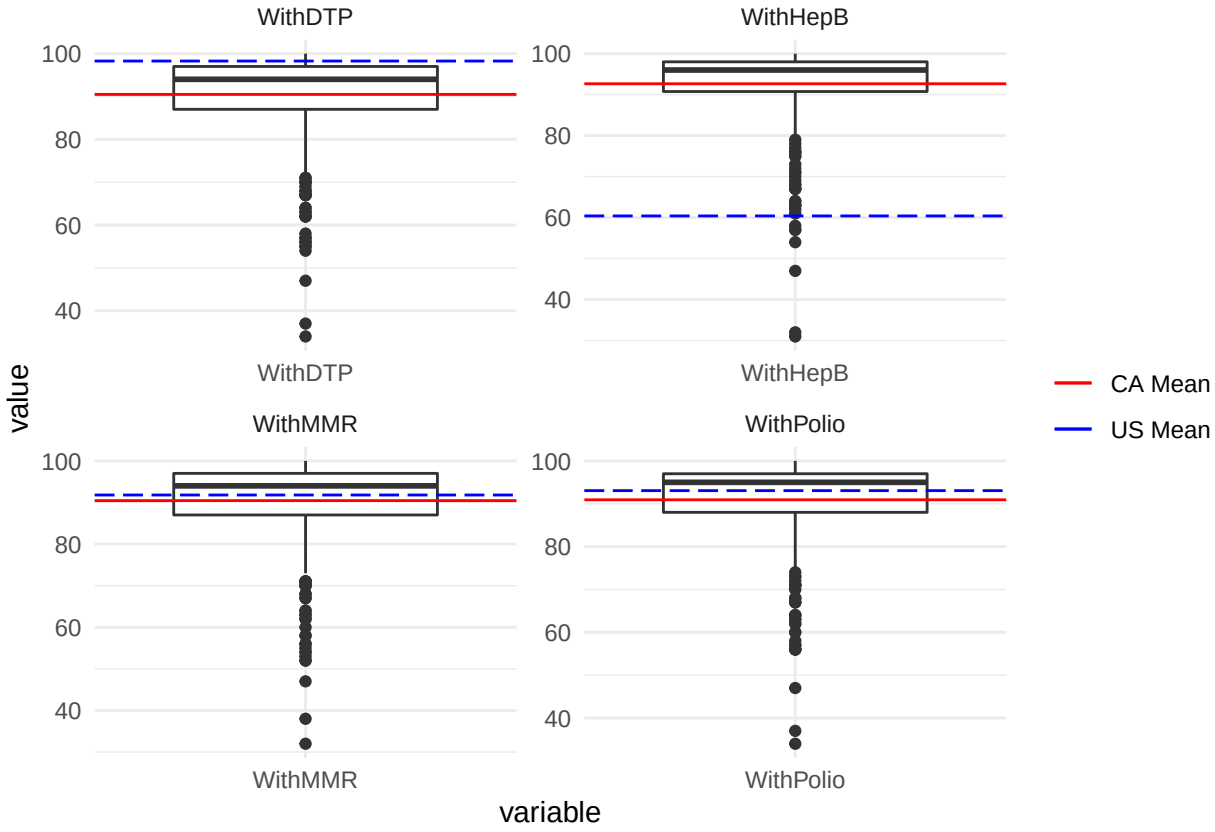


Fig 4.2 - Box plot on 4 types of vaccination rate in the CA distracts with California average(Red) and US average(Blue)

4. *Conclusion Paragraph for Vaccination Rates*

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

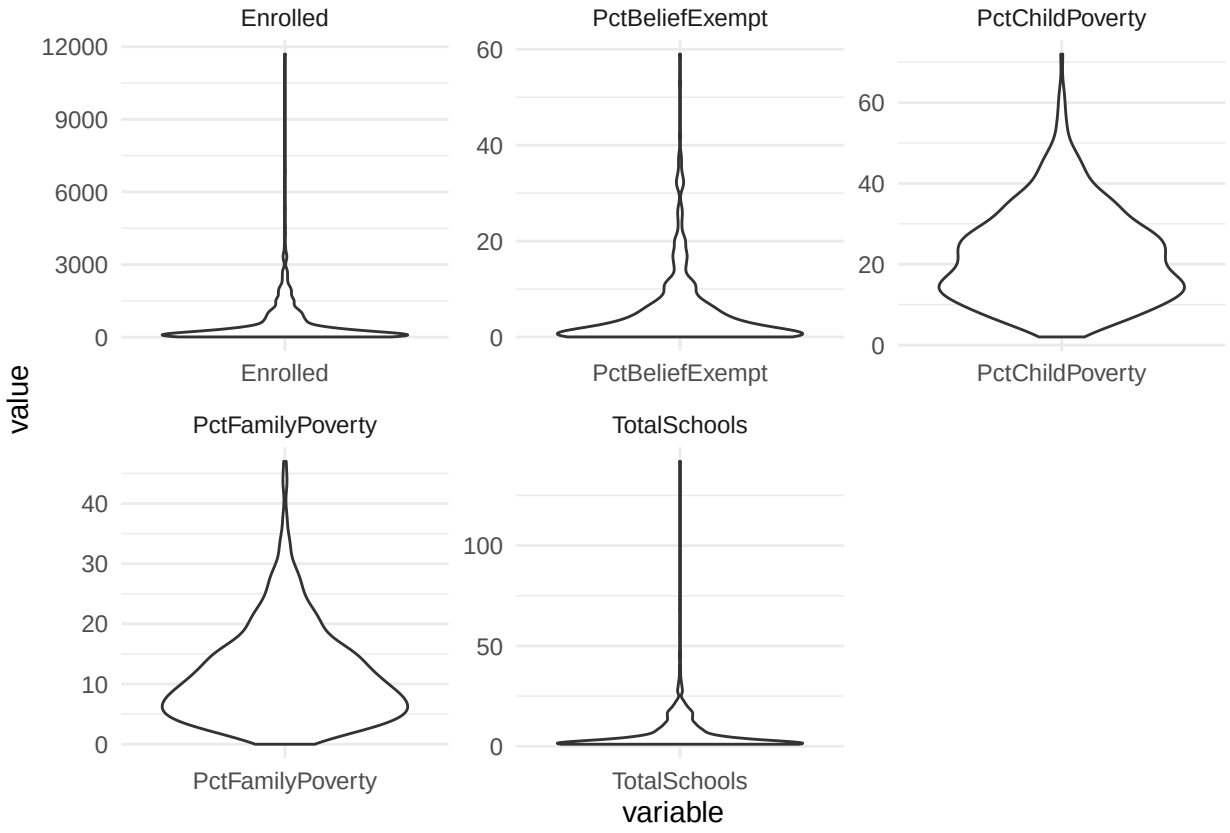
California have great vaccination rate in Hepatitis b, above average in MMR and Polio, and fall behind in DTP. However, for MMR and Polio, both have a mean below the US average and a median above the average. It means there are districts with lower vaccination rate that's dragging the average down. Therefore we should focus on the districts with low vaccination rates on MMR and Polio vaccine, and increase DTP vaccination rate in all districts.

Inferential Reporting

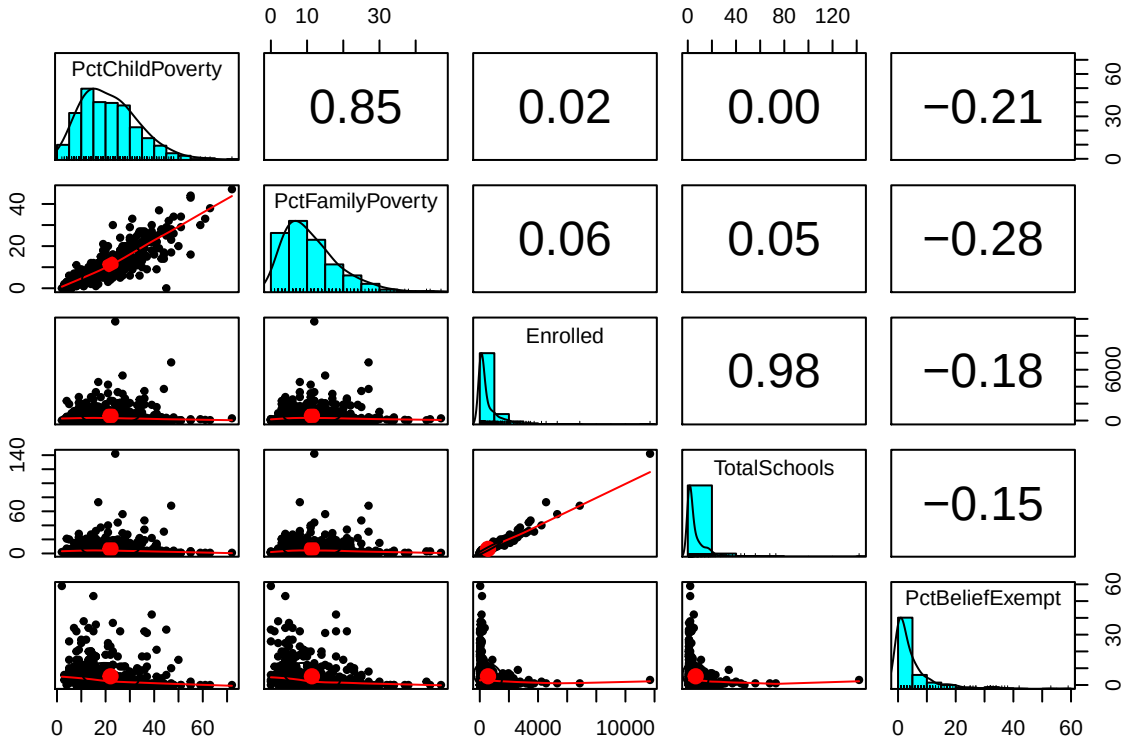
For every item below except 7, use PctChildPoverty, PctFamilyPoverty, Enrolled, and TotalSchools as the four predictors. Explore the data and transform variables as necessary to improve prediction and/or interpretability. Be sure to include appropriate diagnostics and modify your analyses as appropriate.

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

```
districts.filtered %>%
  pivot_longer(cols=c(PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools, PctBeliefExempt) , name
                           values_to = "value", values_drop_na = TRUE) %>%
  ggplot(aes(x = variable, y = value)) + geom_violin() + facet_wrap(~variable, scales="free") +
    theme_minimal() +
    theme(legend.title = element_blank())
```



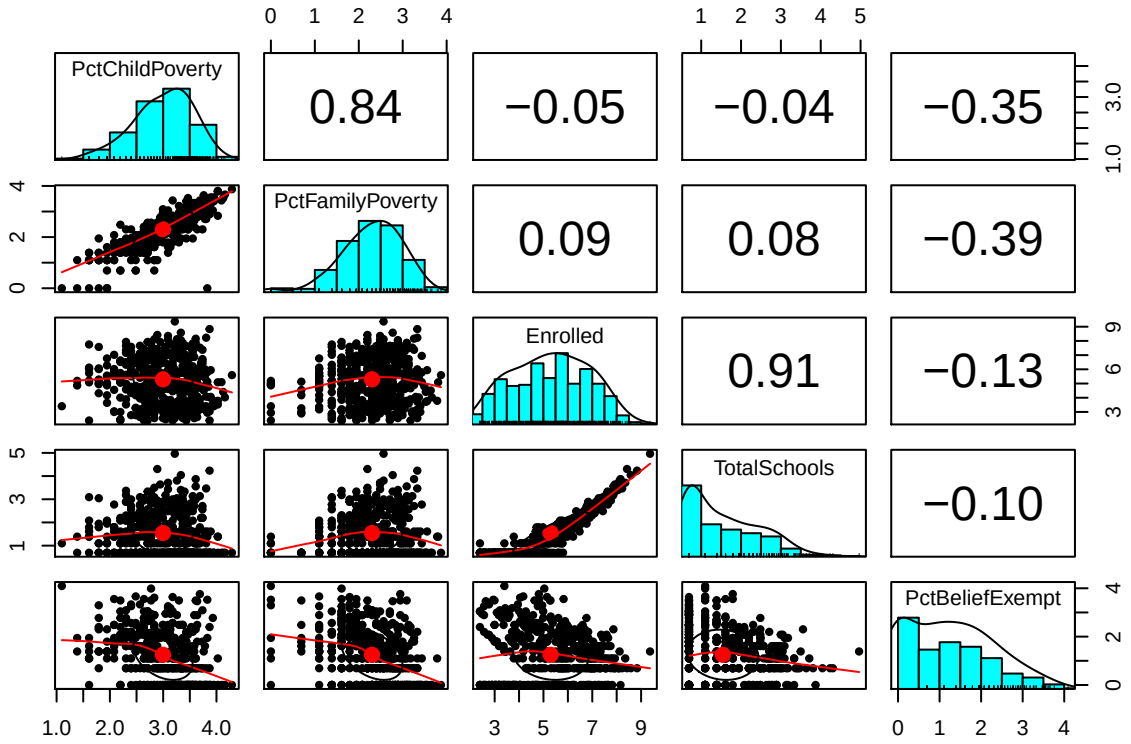

```
districts.filtered %>% dplyr::select(PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools, PctBeli
```




```
districts.filtered.belief.log <- districts.filtered %>% dplyr::select(PctChildPoverty, PctFamilyPoverty
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

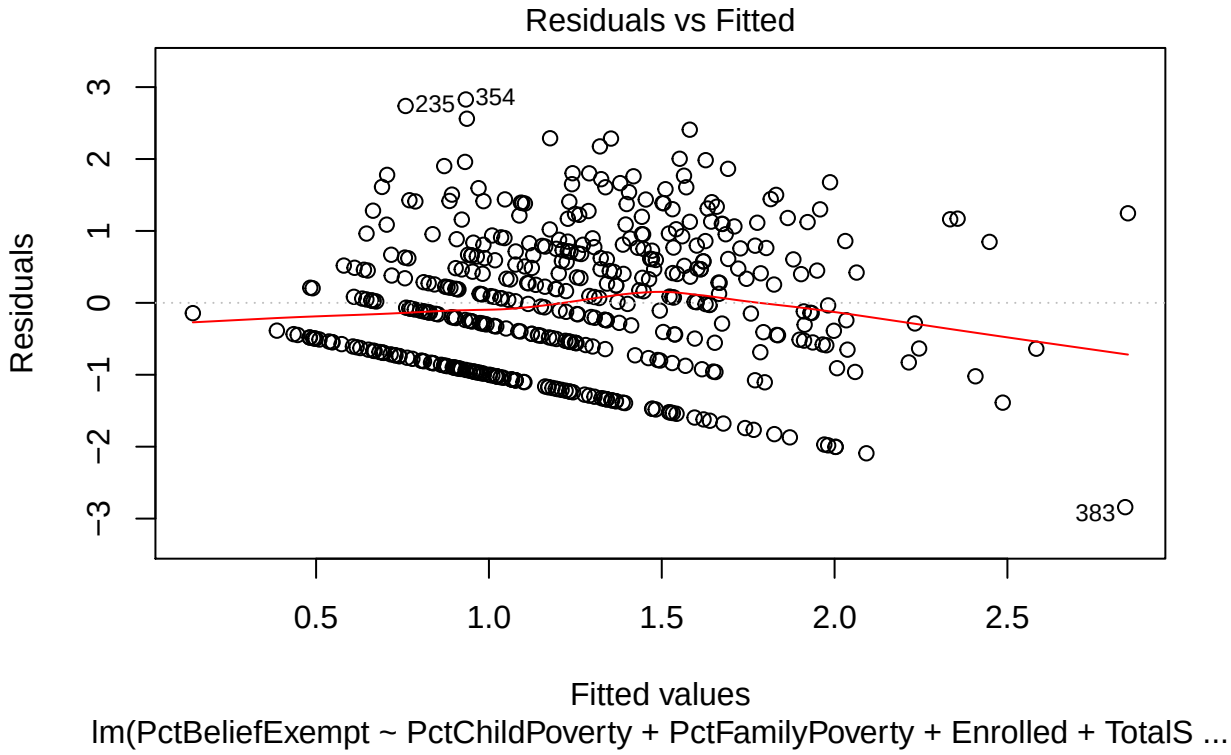
```
pairs.panels(districts.filtered.belief.log)
```

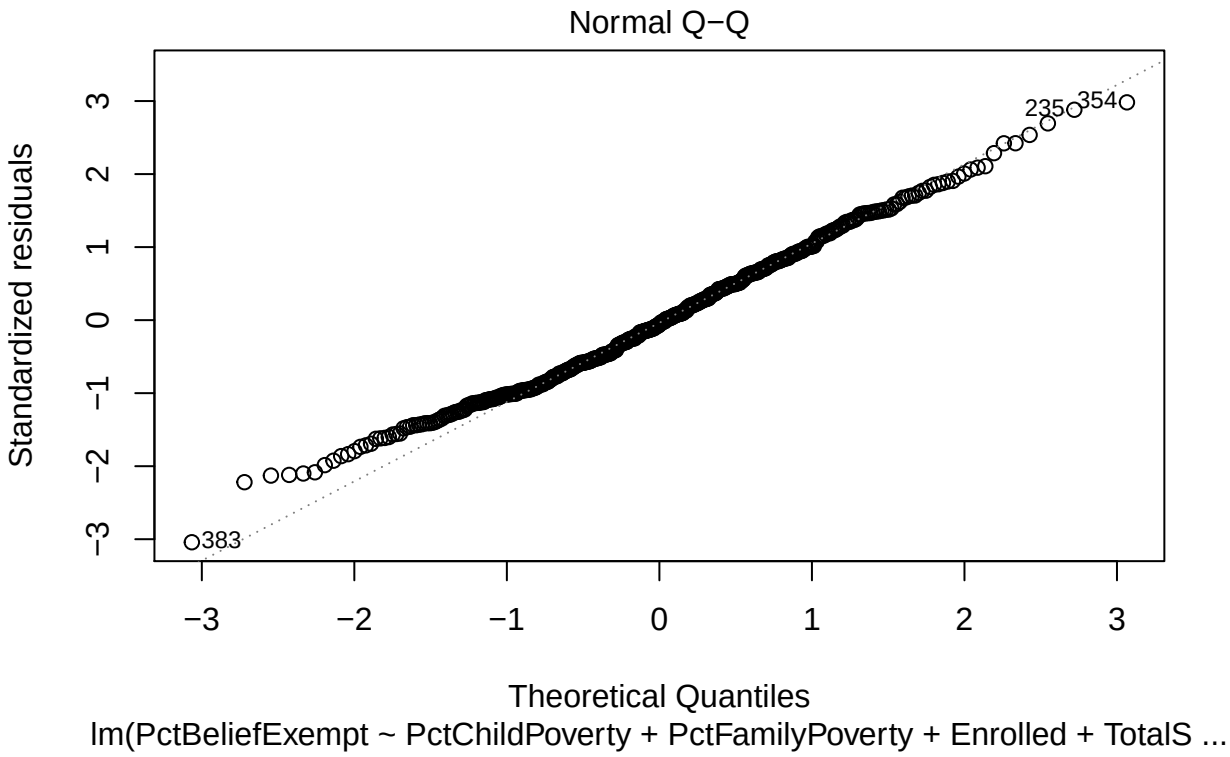



```
lmOut1 <- lm(PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty + Enrolled + TotalSchools, districts.
summary(lmOut1)
```

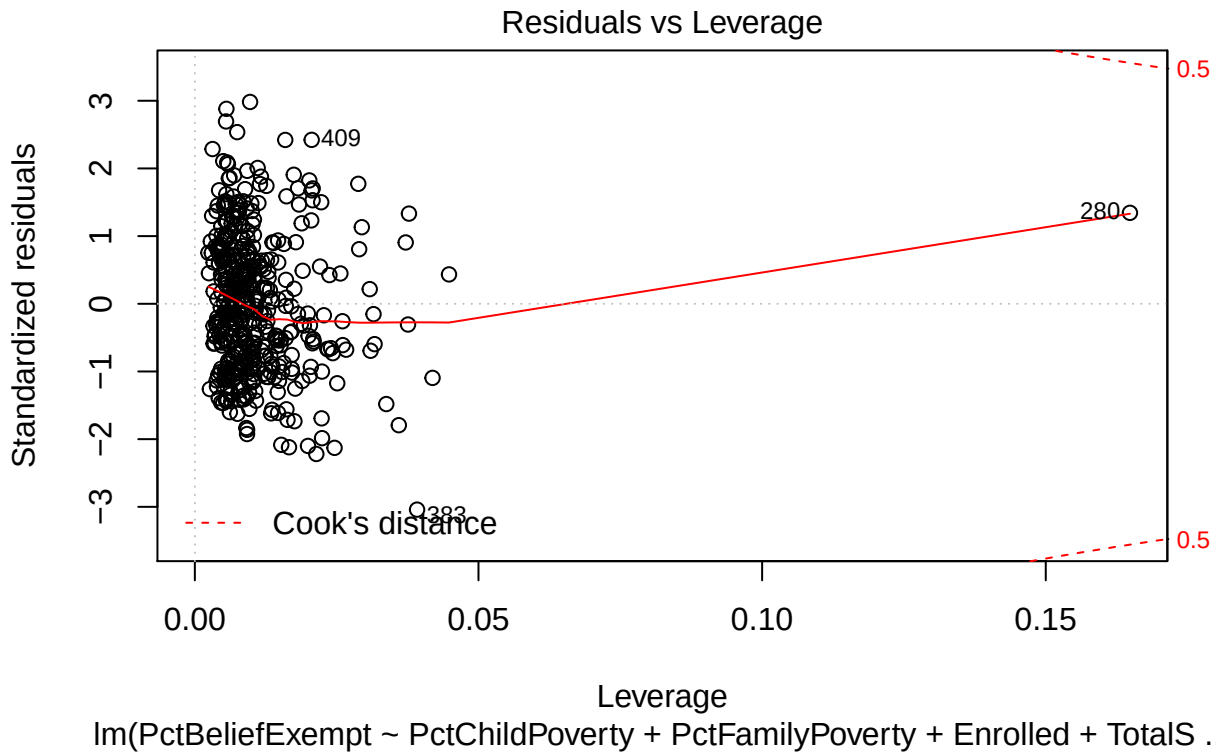
```
##
## Call:
## lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty +
##     Enrolled + TotalSchools, data = districts.filtered.belief.log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.84088 -0.72771 -0.05711  0.66090  2.82798
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.38008    0.36477   9.266  < 2e-16 ***
## PctChildPoverty   -0.18389    0.15186  -1.211 0.226550
## PctFamilyPoverty  -0.45335    0.12235  -3.705 0.000237 ***
## Enrolled          -0.14075    0.07089  -1.985 0.047694 *
## TotalSchools       0.13600    0.12163   1.118 0.264082
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9529 on 455 degrees of freedom
## Multiple R-squared:  0.1676, Adjusted R-squared:  0.1603
## F-statistic: 22.91 on 4 and 455 DF,  p-value: < 2.2e-16
```

```
plot(lmOut1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty + Enrolled + TotalS ...
235
354
383
0.5
1.0
1.5
2.0
2.5
−3
−2
−1
0
1
2
3


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty + Enrolled + TotalS ...
383
235
354
−3
−2
−1
0
1
2
3


Scale−Location
235
354
383
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
2.5
Fitted values
lm(PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty + Enrolled + TotalS ...



Model result isn't great, the r^2 value is too low, and we have a point in Residuals vs Leverage graphs that's way out.

```
vif(lmOut1)
```

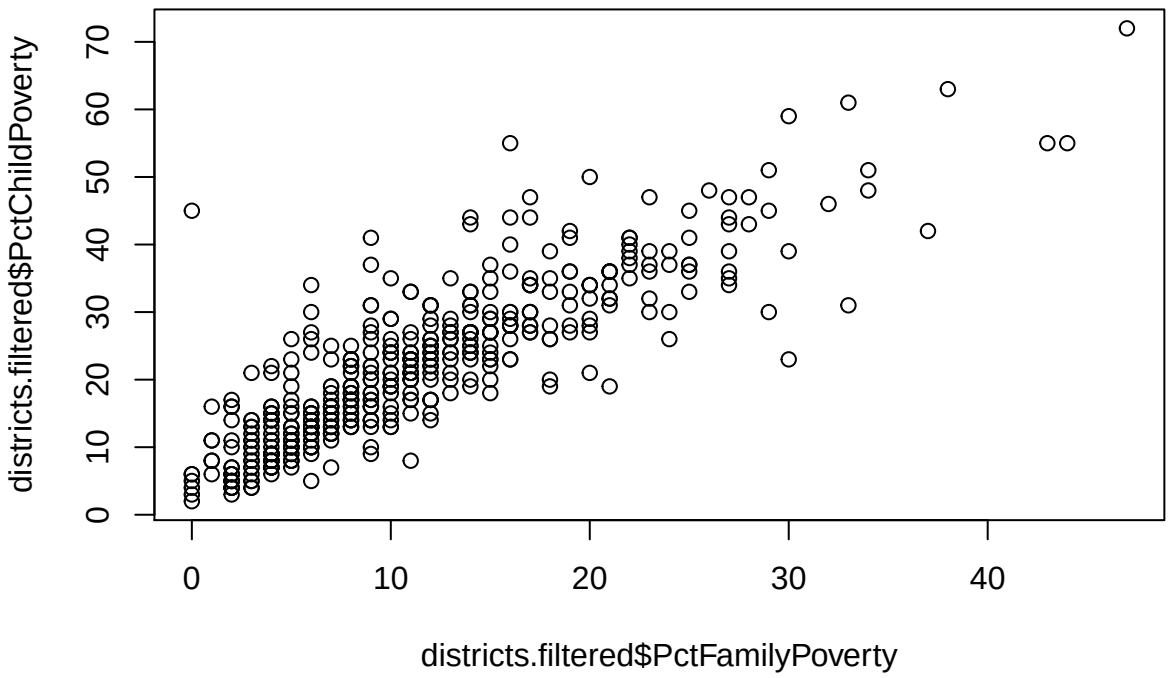
```
##  PctChildPoverty PctFamilyPoverty         Enrolled     TotalSchools
##         3.596307         3.617163         6.055710         5.950180
```

VIF looks ok, but we have two variable that's over 5, but it's not larger than 10 we will leave it for now.

```
districts.filtered %>% dplyr::select(DistrictName, PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSc
```

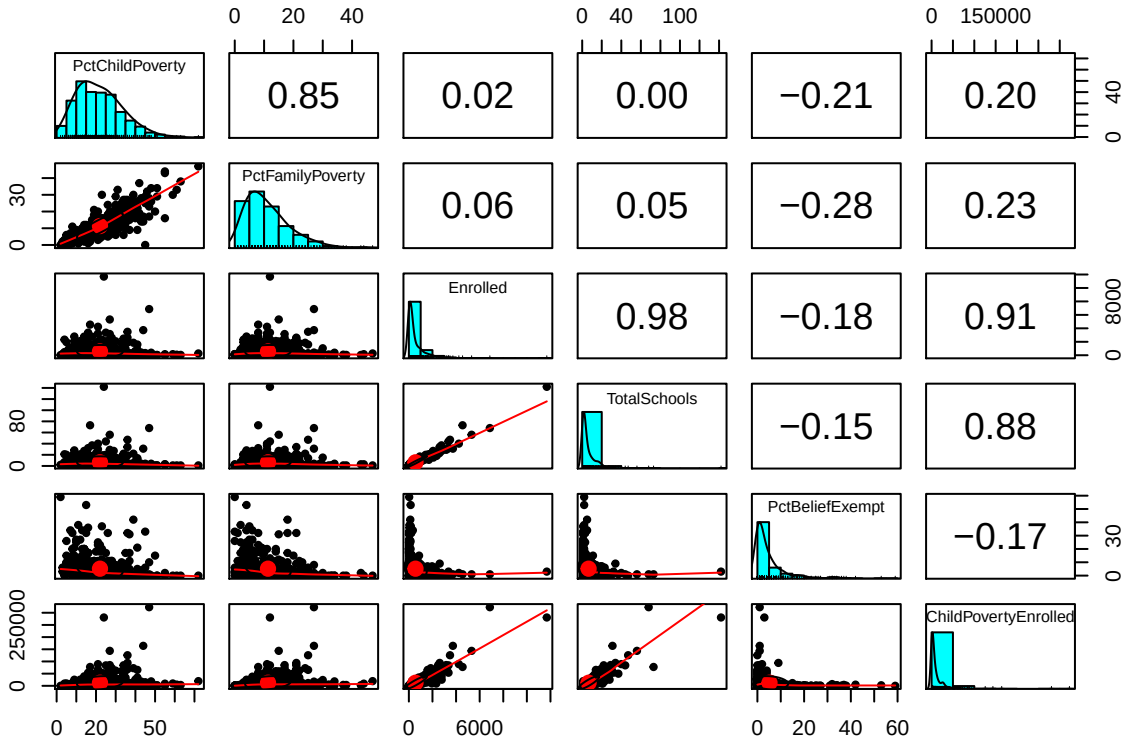
```
##                    DistrictName PctChildPoverty PctFamilyPoverty Enrolled
## 280 Mt. Baldy Joint Elementary              45                0       18
##     TotalSchools PctBeliefExempt
## 280            1              33
```

```
plot(districts.filtered$PctFamilyPoverty,districts.filtered$PctChildPoverty)
```

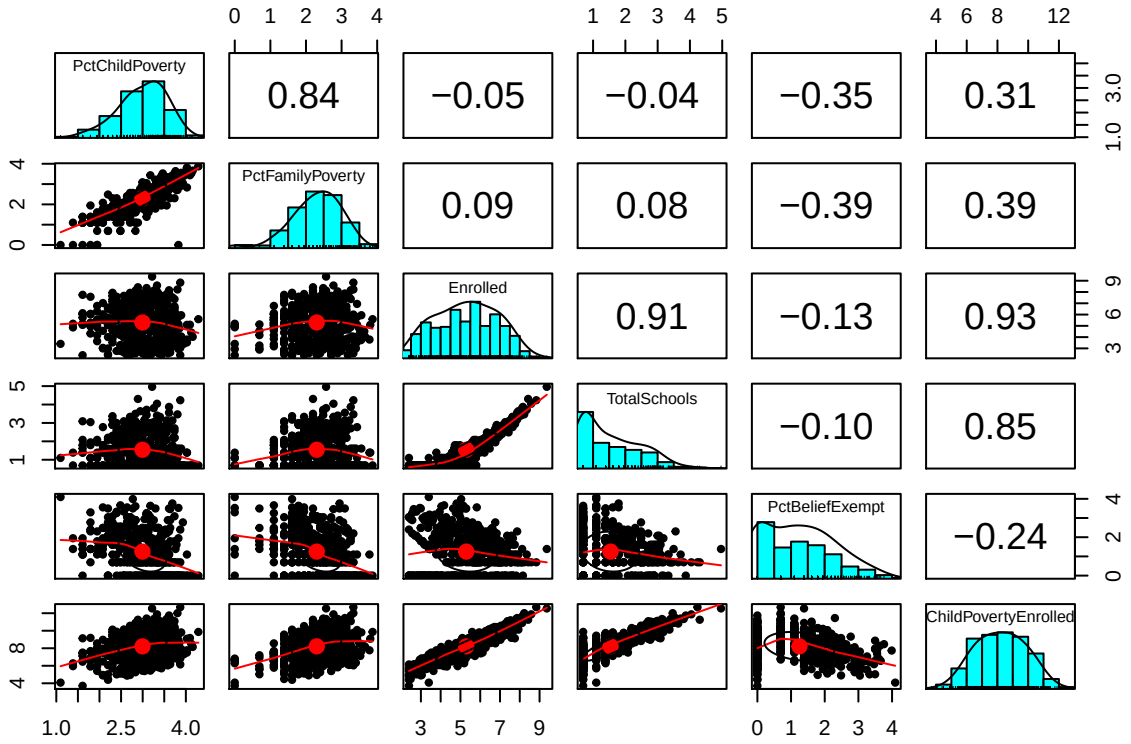


We looked more in to the data point on index 280, the Mt. Baldy Joint Elementary school. There is a extreme ratio of percentage of child poverty verse percentage family poverty. At 45 percent, we will have 8 student out of 18 to be in child poverty while no family there are, I think this might be an outlier due to data collecting, but we cant take it out yet, more research needed. Because of the unsatisfied result, I'm going to add more columns. According to PPIC(Public Policy Institute of California), the age to be consider as a child is till 17. After some research, I will assume the data set that's provided to me is the K-12 education, therefore I'm going to mutate a new columns called `ChildPovertyEnrolled` by multiple them.

```
districts.filtered.belief.count <- districts.filtered %>%
                                        mutate(ChildPovertyEnrolled = PctChildPoverty * Enrolled) %>%
                                        dplyr::select(PctChildPoverty, PctFamilyPoverty, Enrolled, To
pairs.panels(districts.filtered.belief.count)
```



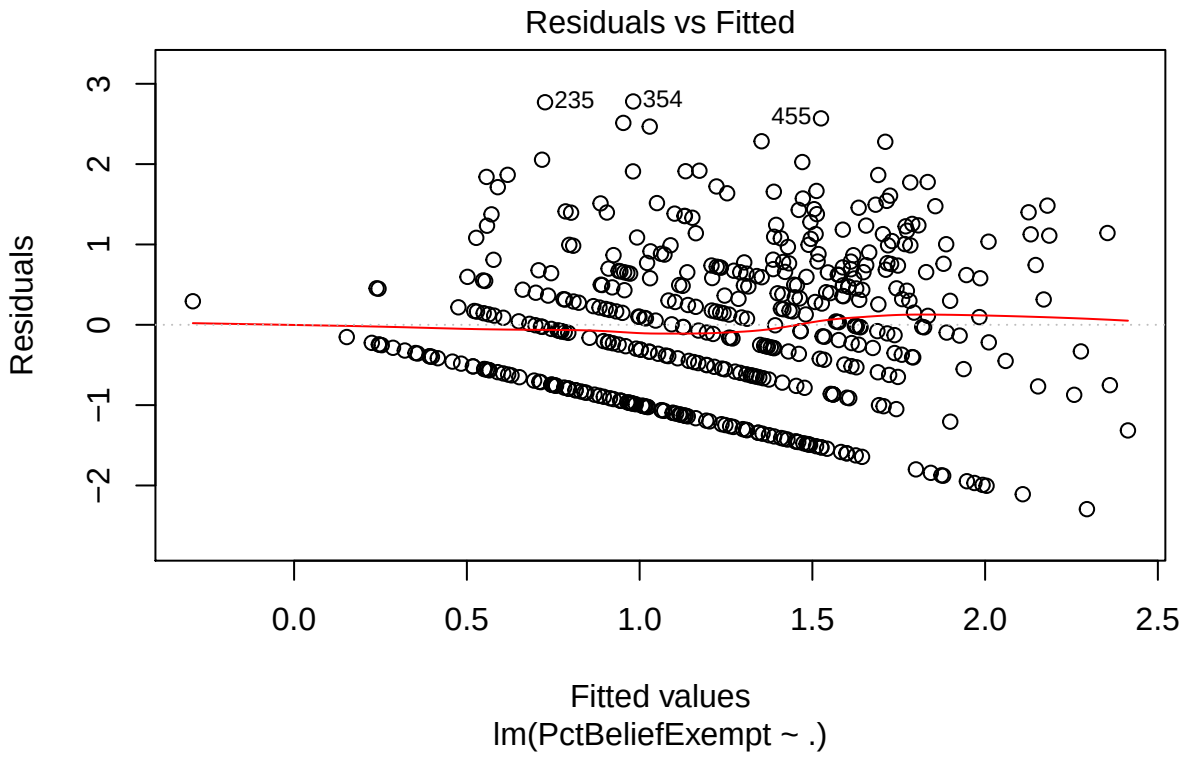

```
districts.filtered.belief.count.log <- log(districts.filtered.belief.count + 1)
pairs.panels(districts.filtered.belief.count.log)
```

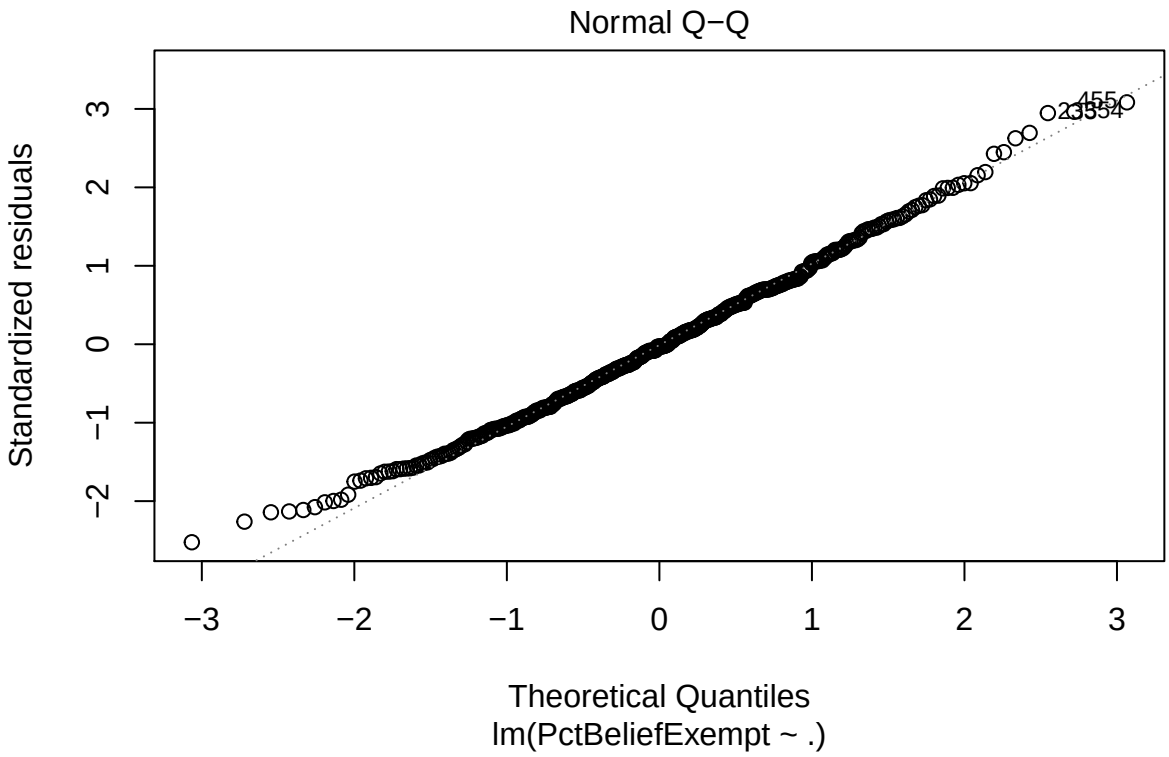


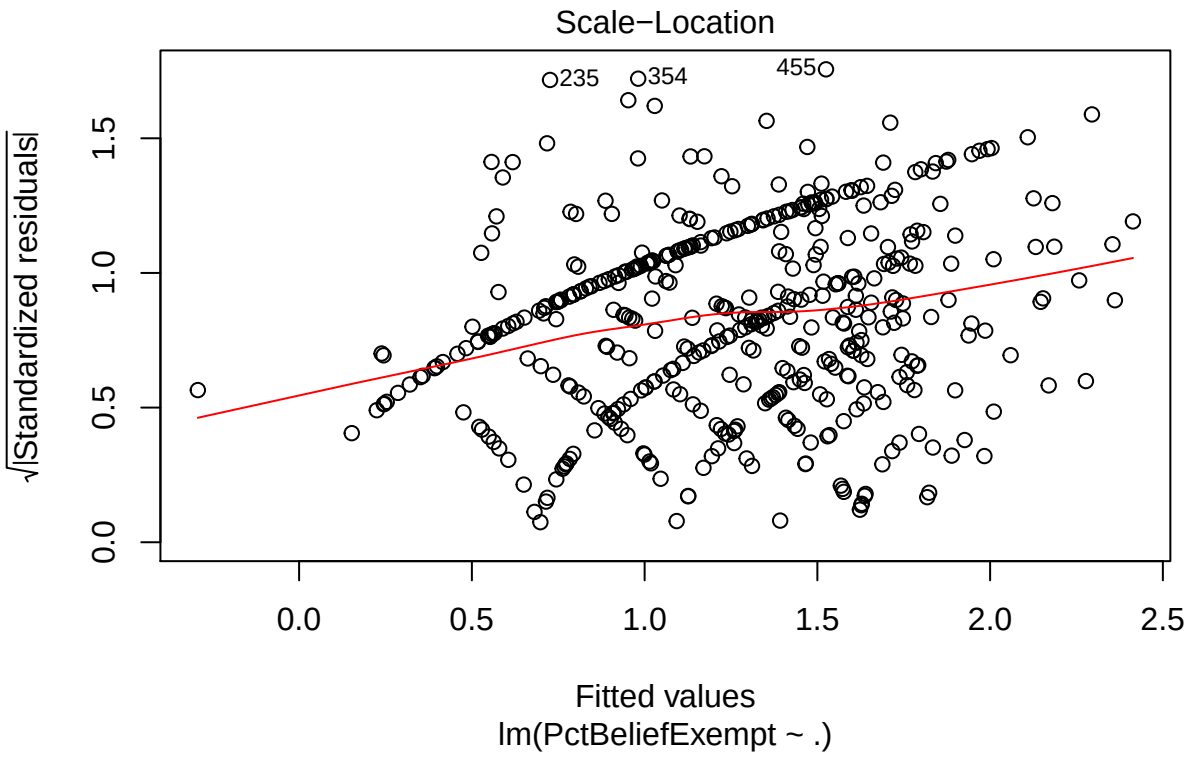

```
lmOut2 <- lm(PctBeliefExempt ~ ., districts.filtered.belief.count.log)
summary(lmOut2)
```

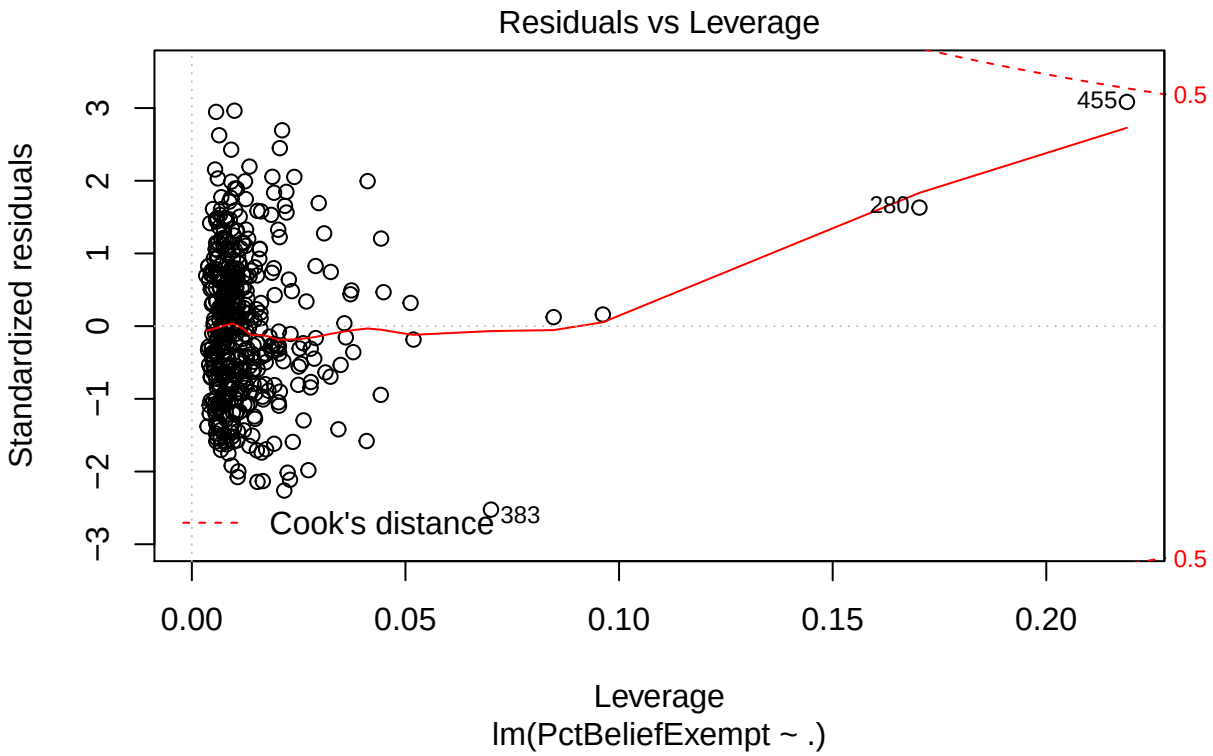
```
##
## Call:
## lm(formula = PctBeliefExempt ~ ., data = districts.filtered.belief.count.log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.29463 -0.66896 -0.02777  0.64679  2.77976
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.2014     0.9282   6.681 6.95e-11 ***
## PctChildPoverty        -8.2022     2.4350  -3.368 0.000820 ***
## PctFamilyPoverty       -0.4427     0.1211  -3.656 0.000286 ***
## Enrolled               -7.7628     2.3113  -3.359 0.000849 ***
## TotalSchools            0.2773     0.1277   2.171 0.030421 *
## ChildPovertyEnrolled    7.4635     2.2622   3.299 0.001046 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9427 on 454 degrees of freedom
## Multiple R-squared:  0.1871, Adjusted R-squared:  0.1782
## F-statistic:  20.9 on 5 and 454 DF,  p-value: < 2.2e-16
```

```
plot(lmOut2)
```


Residuals vs Fitted
235
354
455
Residuals
3
2
1
0
−1
−2
0.0
0.5
1.0
1.5
2.0
2.5
Fitted values
lm(PctBeliefExempt ~ .)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(PctBeliefExempt ~ .)
455
3354
−3
−2
−1
0
1
2
3


Scale−Location
235
354
455
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
2.5
Fitted values
lm(PctBeliefExempt ~ .)



So the model is not better and we have 1 points on the l leverage graph that's almost out of the 0.5 value. I took a deeper look and the percentage of the belief exempt is at 59, which is the highest in the whole dataset.

```
districts.filtered %>% dplyr::select(DistrictName, PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSc
```

```
##                 DistrictName PctChildPoverty PctFamilyPoverty Enrolled
## 409 Spencer Valley Elementary              23                5      171
## 455      Lagunitas Elementary               2                0       29
##     TotalSchools PctBeliefExempt
## 409            1              37
## 455            2              59
```

```
vif(lmOut2)
```

```
##      PctChildPoverty     PctFamilyPoverty             Enrolled
##           944.726787             3.619744          6576.843662
##         TotalSchools ChildPovertyEnrolled
##             6.704382          7131.535246
```

bad vif score we will need to drop some variables.

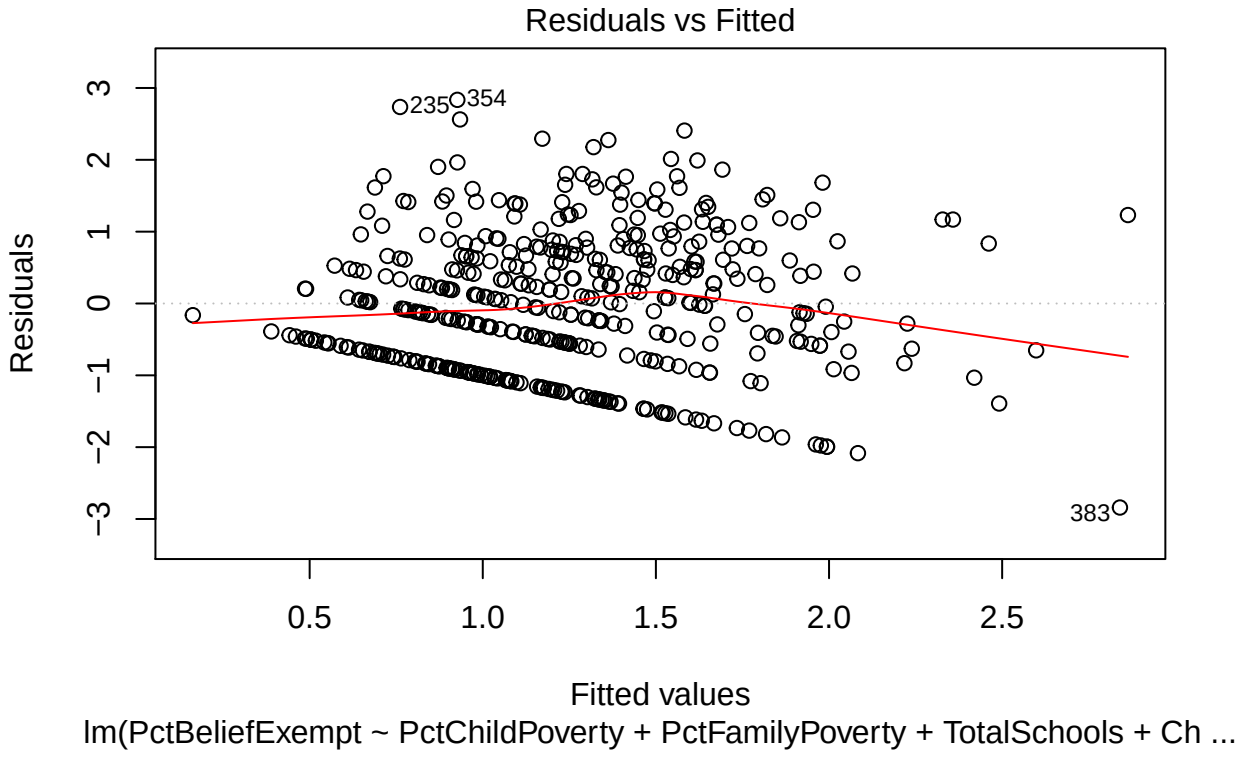
```
colnames(districts.filtered.belief.count.log)
```

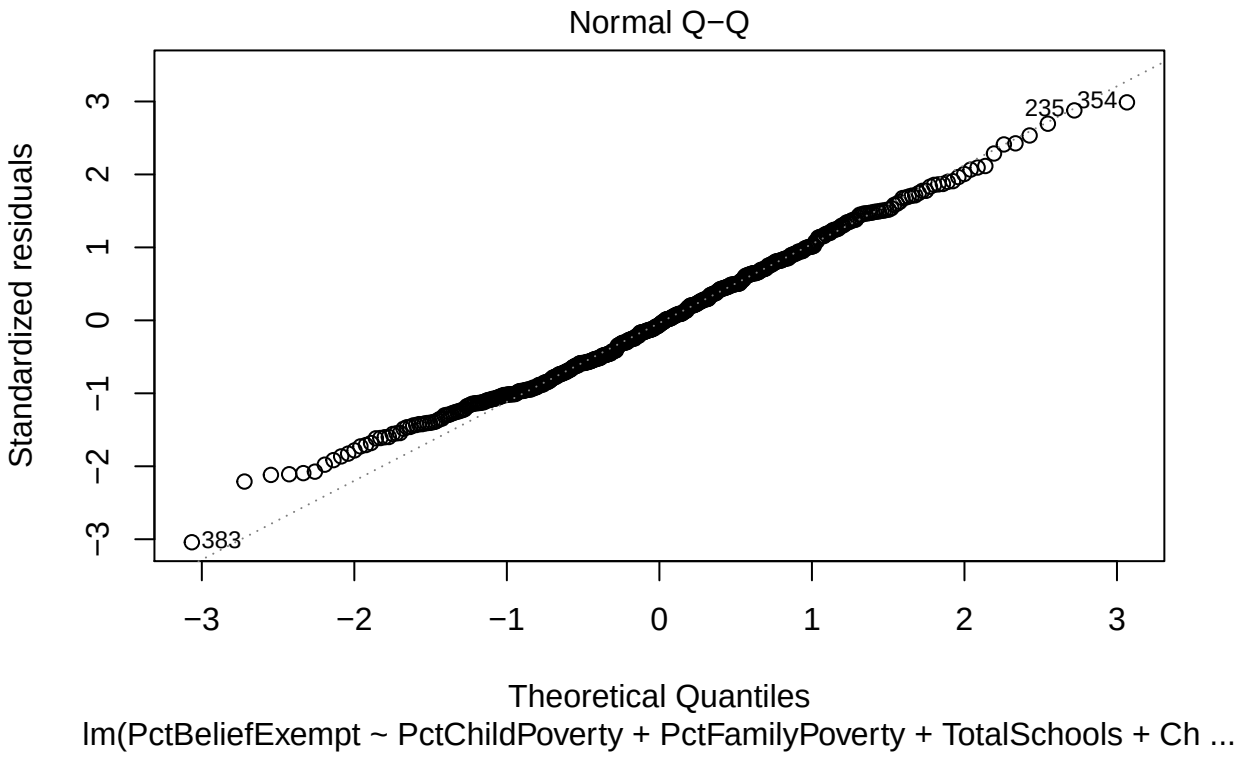
```
## [1] "PctChildPoverty"      "PctFamilyPoverty"     "Enrolled"
## [4] "TotalSchools"         "PctBeliefExempt"      "ChildPovertyEnrolled"
```

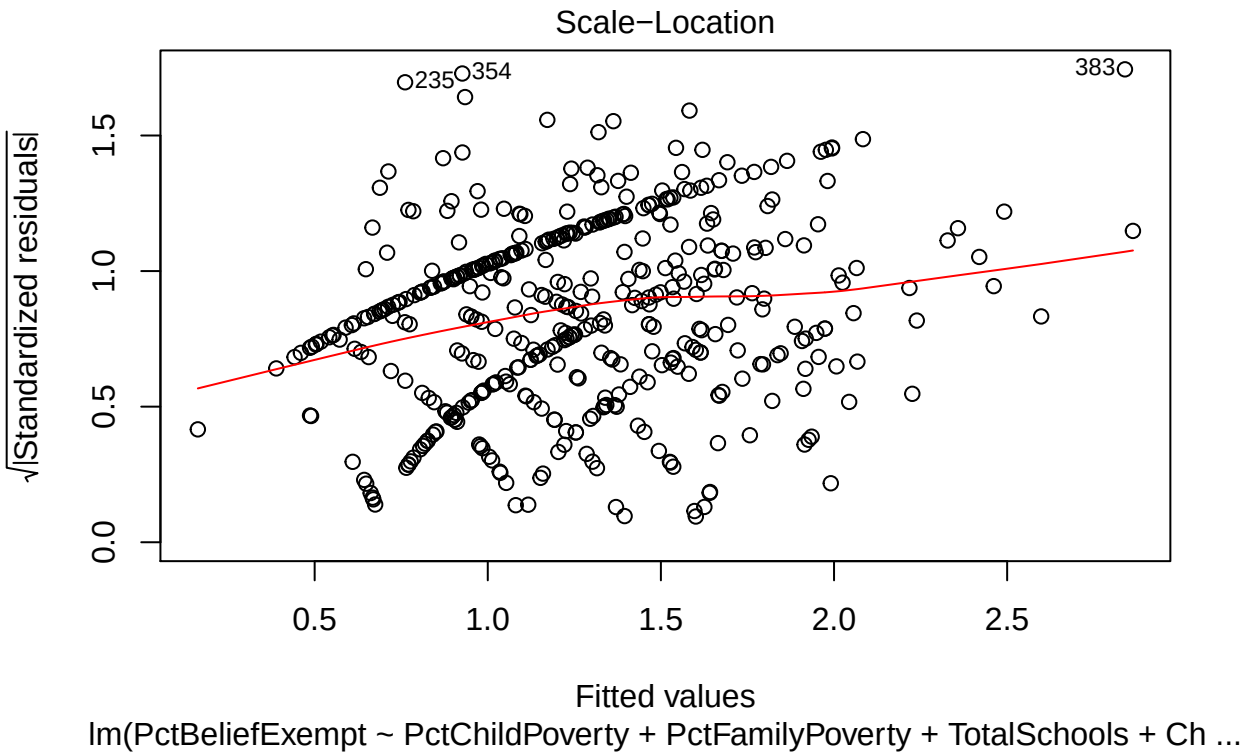
```
lmOut3 <- lm(PctBeliefExempt ~ PctChildPoverty+PctFamilyPoverty+TotalSchools+ChildPovertyEnrolled, dist
summary(lmOut3)
```

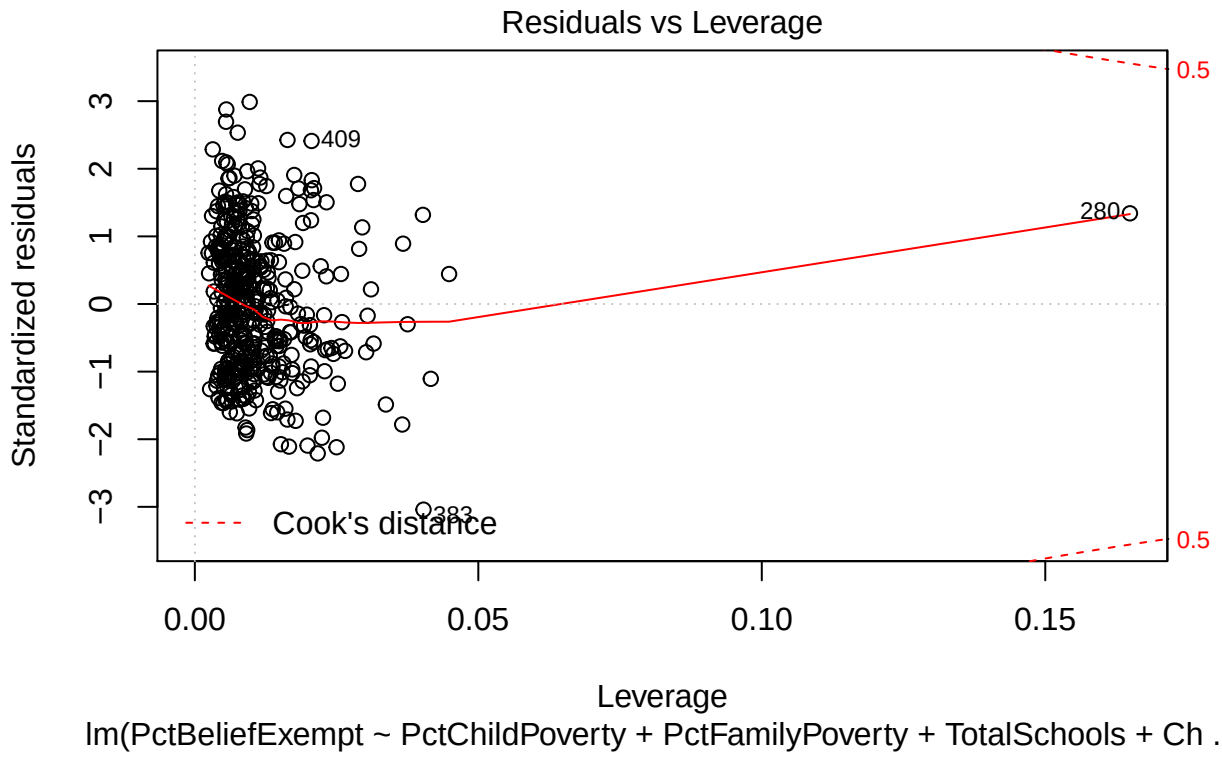
```
##
## Call:
## lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty +
##     TotalSchools + ChildPovertyEnrolled, data = districts.filtered.belief.count.log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.84017 -0.72447 -0.05725  0.65848  2.83471
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.30771    0.34912   9.475  < 2e-16 ***
## PctChildPoverty       -0.04135    0.16030  -0.258 0.796569
## PctFamilyPoverty      -0.45516    0.12239  -3.719 0.000225 ***
## TotalSchools           0.12239    0.12044   1.016 0.310064
## ChildPovertyEnrolled -0.13083    0.06942  -1.885 0.060108 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9533 on 455 degrees of freedom
## Multiple R-squared:  0.1669, Adjusted R-squared:  0.1596
## F-statistic: 22.79 on 4 and 455 DF,  p-value: < 2.2e-16
```

```
plot(lmOut3)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty + TotalSchools + Ch ...
383
235
354
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
235
354
383
0.5
1.0
1.5
2.0
2.5
Fitted values
lm(PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty + TotalSchools + Ch ...

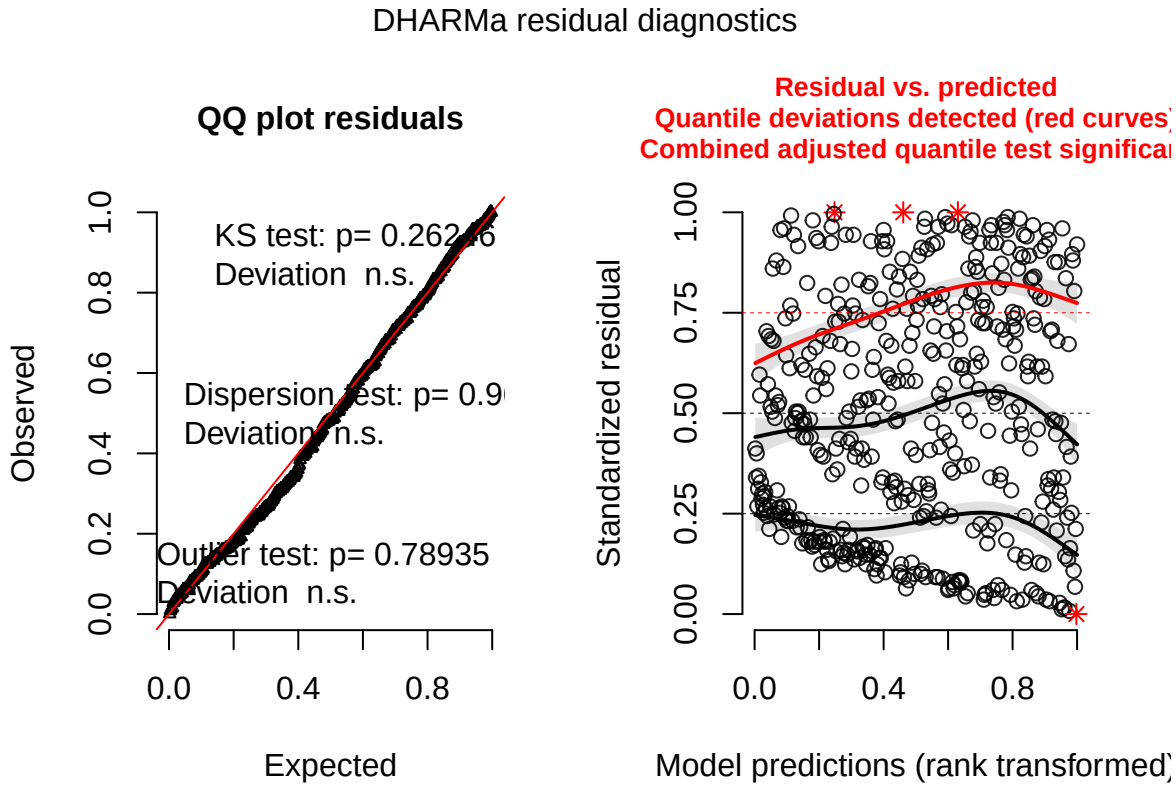



```
vif(lmOut3)
```

```
##      PctChildPoverty     PctFamilyPoverty         TotalSchools
##             4.003621             3.616336             5.829752
## ChildPovertyEnrolled
##             6.566449
```

Although both model1 and model3 all have significant p-value from the test, we are going to use the first model as the R^2 is better and it's easier to interpret.

```
simulationOutput2 <- simulateResiduals(fittedModel = lmOut1, n = 250)
plot(simulationOutput2)
```



One more test before the interpretation, The DHARMA simulation residual looks normal.

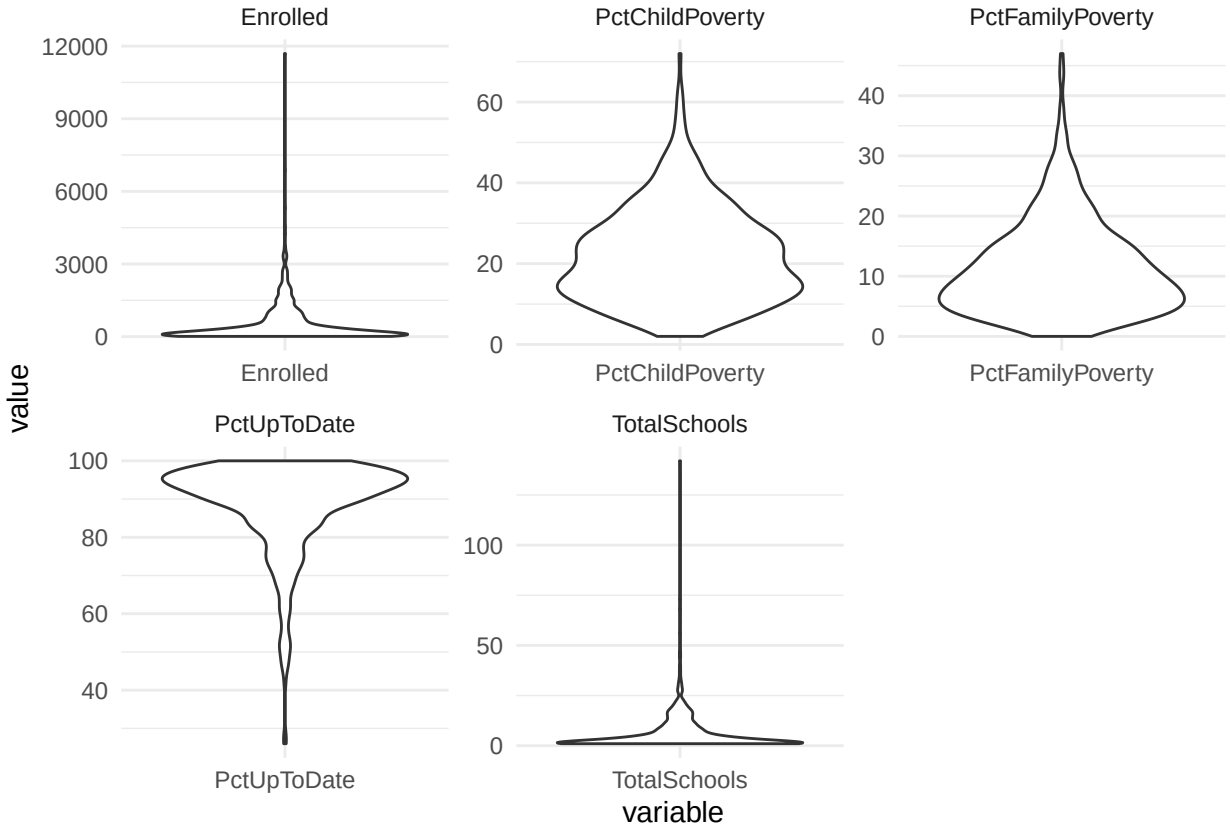
```
beta1 <- lm.beta(lm(PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty + Enrolled + TotalSchools, dis
summary(beta1)
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty +
##     Enrolled + TotalSchools, data = districts.filtered.belief.log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.84088 -0.72771 -0.05711  0.66090  2.82798
##
## Coefficients:
##                  Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)       3.38008      0.00000    0.36477   9.266  < 2e-16 ***
## PctChildPoverty  -0.18389     -0.09822    0.15186  -1.211 0.226550
## PctFamilyPoverty -0.45335     -0.30141    0.12235  -3.705 0.000237 ***
## Enrolled         -0.14075     -0.20897    0.07089  -1.985 0.047694 *
## TotalSchools      0.13600      0.11666    0.12163   1.118 0.264082
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9529 on 455 degrees of freedom
## Multiple R-squared:  0.1676, Adjusted R-squared:  0.1603
## F-statistic: 22.91 on 4 and 455 DF,  p-value: < 2.2e-16
```

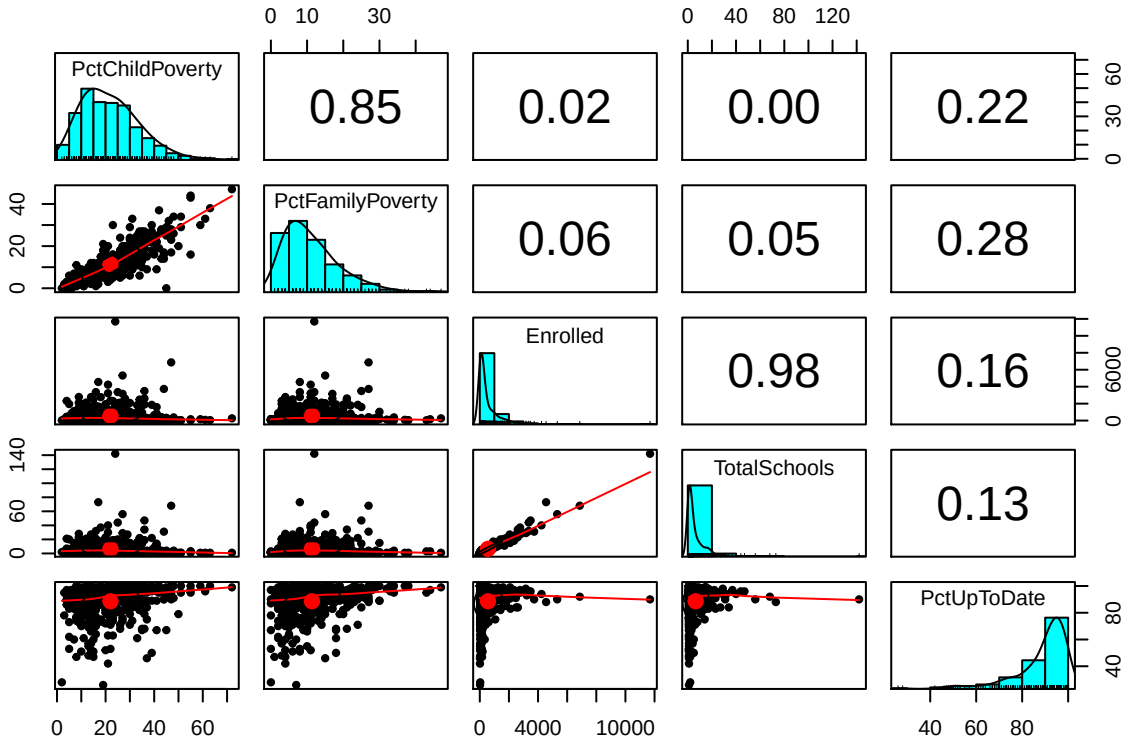
The two significant variables are the percentage of family poverty and the number of enrolled students. They are both negatively correlated to the percent of believe exceptions. It means the more poverty in the area, less believe exception and more enrolled students less believe exception. One thing to note is that all variables are log transformed, the real changes in percentage of the believe exception will be in log. So every 1 standard deviation increase in percentage of family poverty will have 0.453 standard deviation decrease in the log(percentage of belief exception).

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

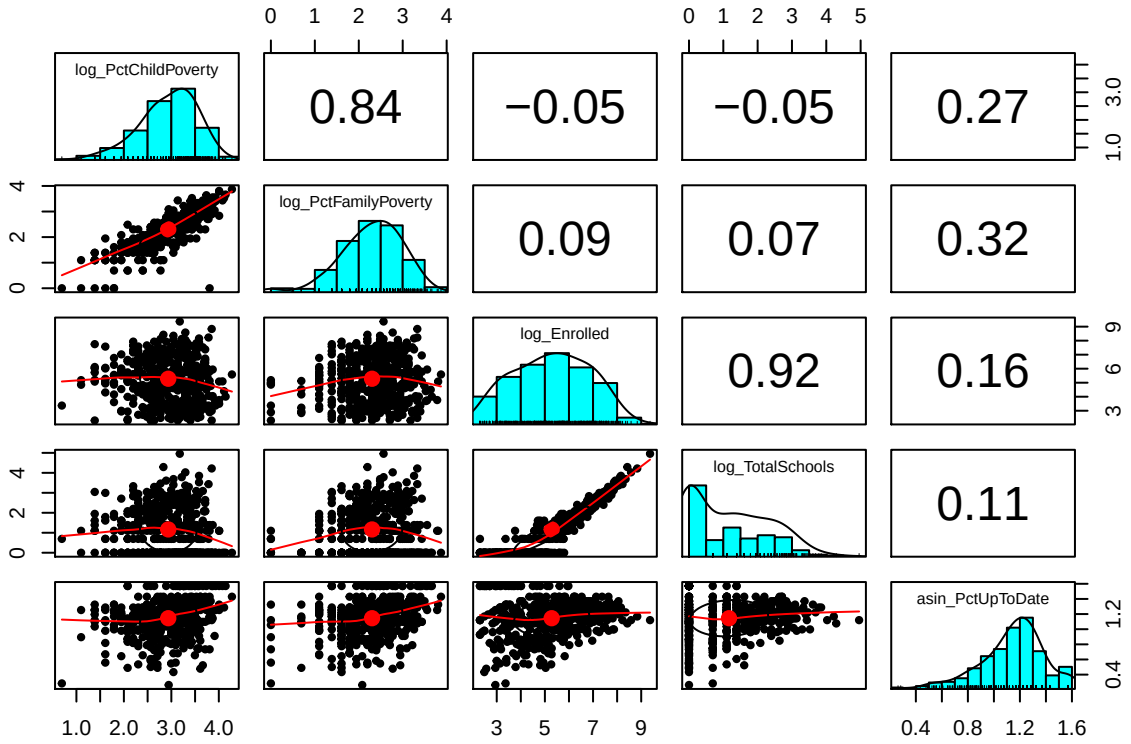
```
districts.filtered %>%
  pivot_longer(cols=c(PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools, PctUpToDate) , names_t
                          values_to = "value", values_drop_na = TRUE) %>%
  ggplot(aes(x = variable, y = value)) + geom_violin() + facet_wrap(~variable, scales="free") +
    theme_minimal() +
    theme(legend.title = element_blank())
```




```
districts.filtered.uptodate <- districts.filtered %>% dplyr::select(PctChildPoverty, PctFamilyPoverty, 
pairs.panels(districts.filtered.uptodate)
```



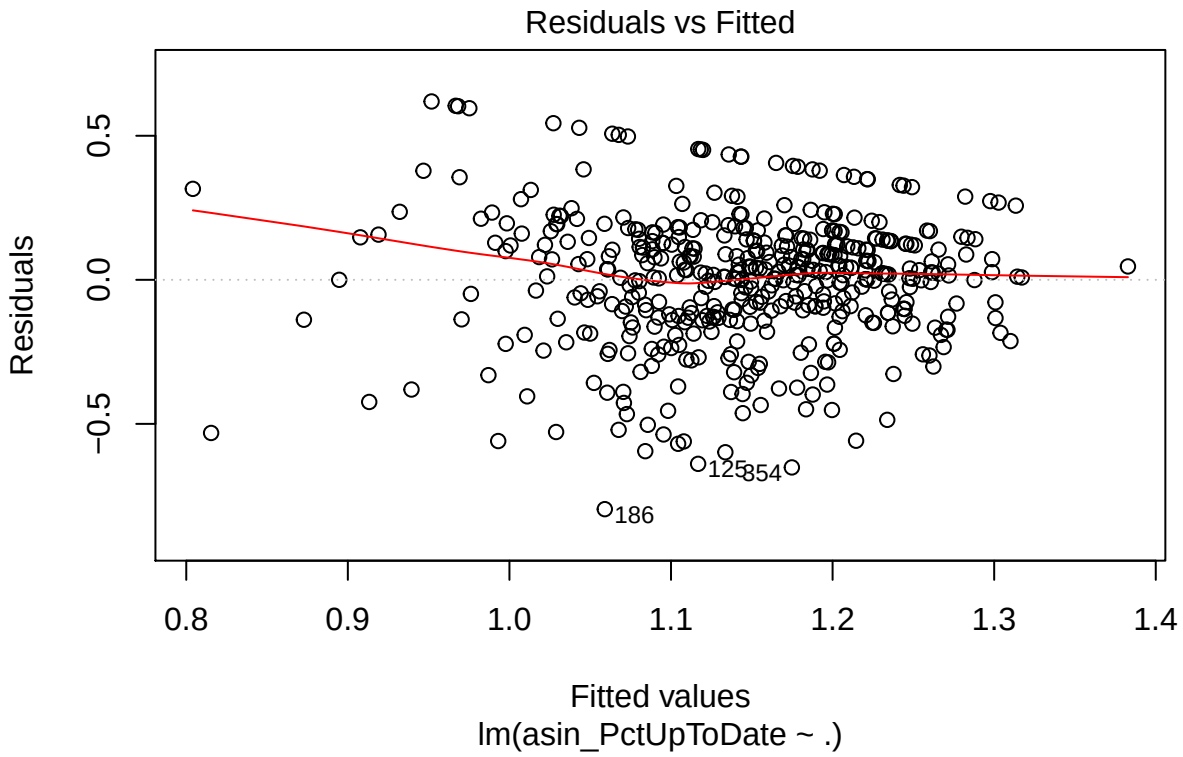

```
districts.filtered.uptodate.transformed <- districts.filtered.uptodate %>% mutate(log_PctChildPoverty =
                                                                                  log_PctFamilyPoverty
                                                                                  log_Enrolled = log(
                                                                                  log_TotalSchools =
                                                                                  asin_PctUpToDate =
                                                                                dplyr::select(log_PctChild

pairs.panels(districts.filtered.uptodate.transformed)
```

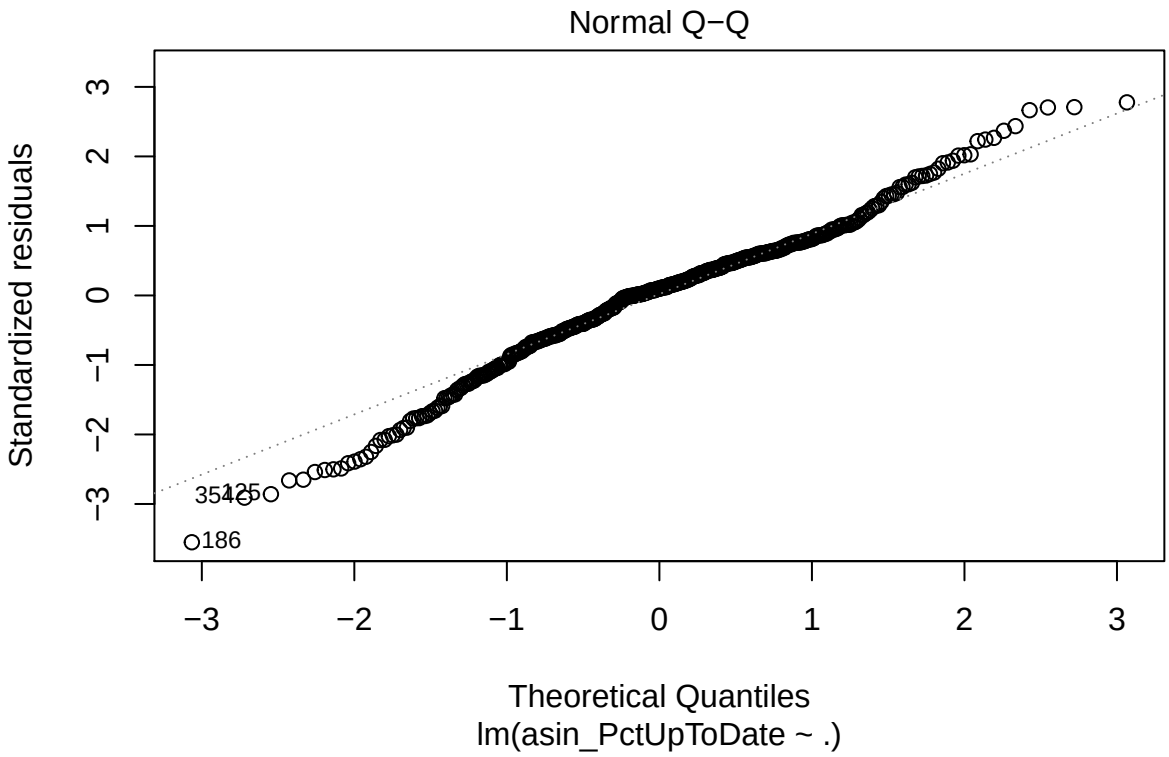


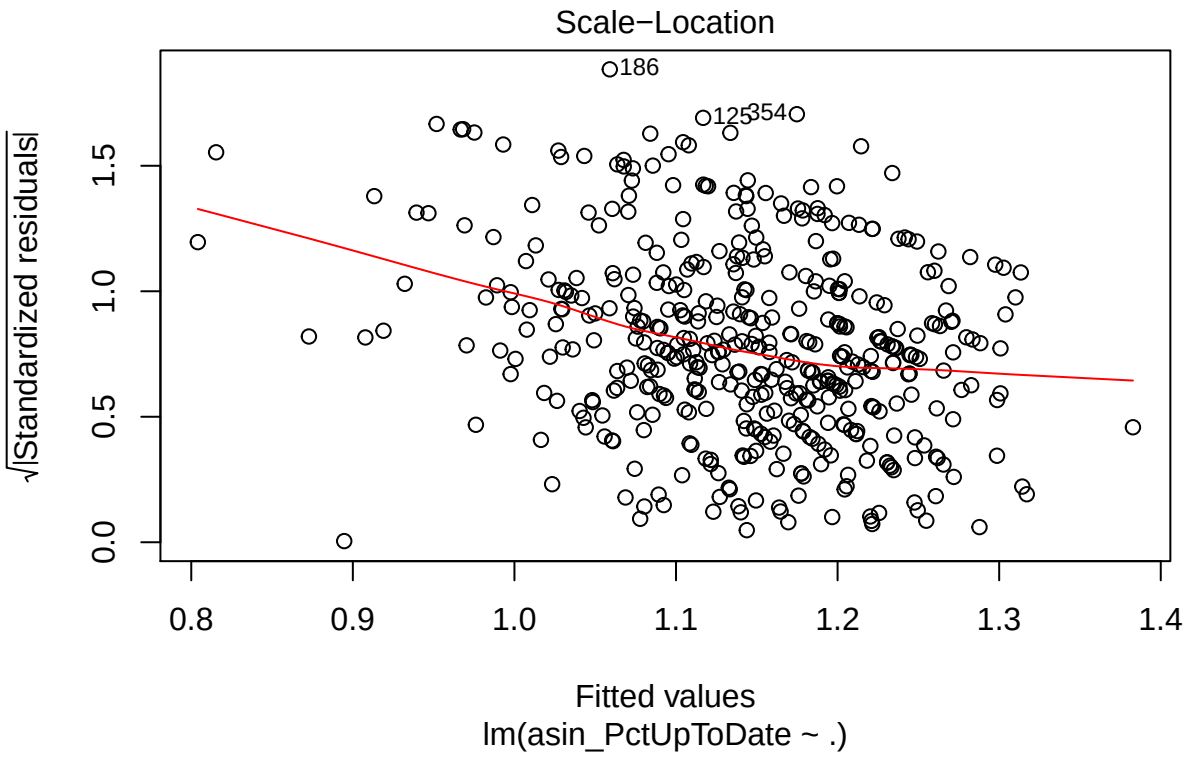

```
lmOut4 <- lm(asin_PctUpToDate ~ ., districts.filtered.uptodate.transformed)
summary(lmOut4)
```

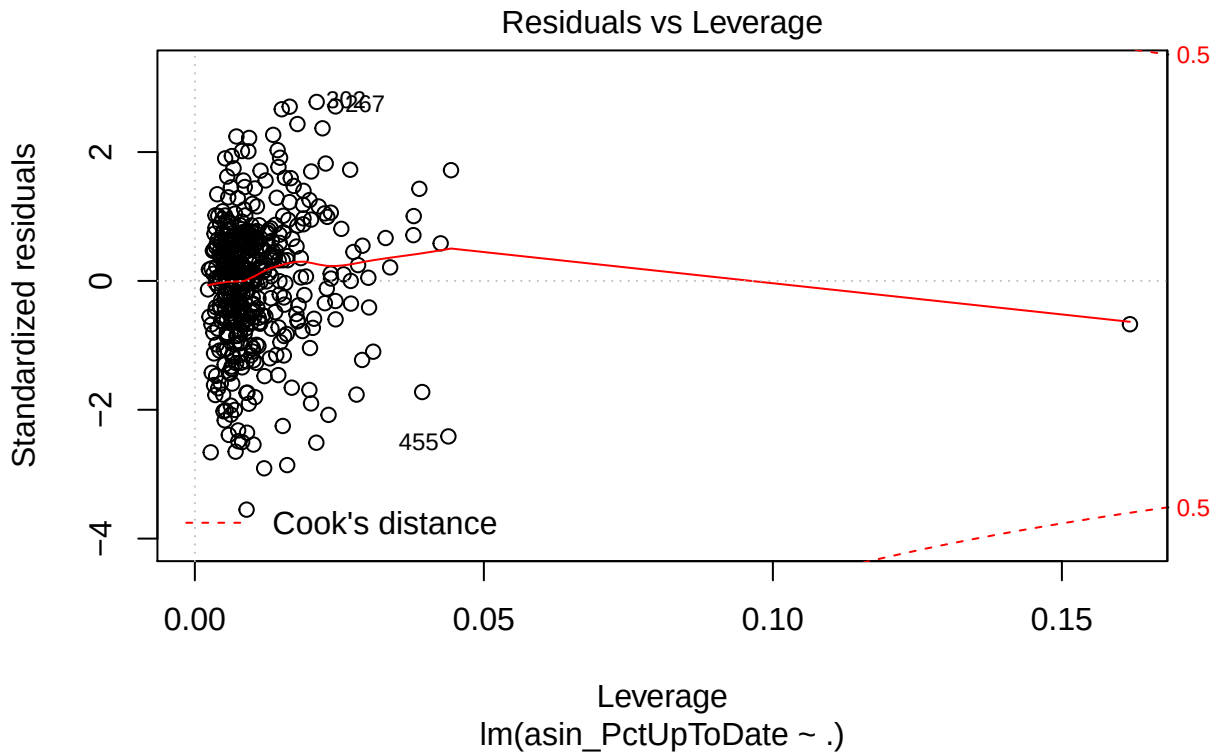
```
##
## Call:
## lm(formula = asin_PctUpToDate ~ ., data = districts.filtered.uptodate.transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79599 -0.12624  0.02219  0.13564  0.61899
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.66660    0.08912   7.480 3.86e-13 ***
## log_PctChildPoverty    0.01628    0.03328   0.489  0.62492
## log_PctFamilyPoverty   0.09500    0.02882   3.296  0.00106 **
## log_Enrolled           0.04990    0.01719   2.903  0.00387 **
## log_TotalSchools      -0.04405    0.02333  -1.888  0.05967 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2252 on 455 degrees of freedom
## Multiple R-squared:  0.1269, Adjusted R-squared:  0.1192
## F-statistic: 16.53 on 4 and 455 DF,  p-value: 1.179e-12
```

```
plot(lmOut4)
```


Residuals vs Fitted
Residuals
0.5
0.0
−0.5
125
354
186
0.8
0.9
1.0
1.1
1.2
1.3
1.4
Fitted values
lm(asin_PctUpToDate ~ .)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(asin_PctUpToDate ~ .)
186


Scale−Location
186
125
354
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.8
0.9
1.0
1.1
1.2
1.3
1.4
Fitted values
lm(asin_PctUpToDate ~ .)



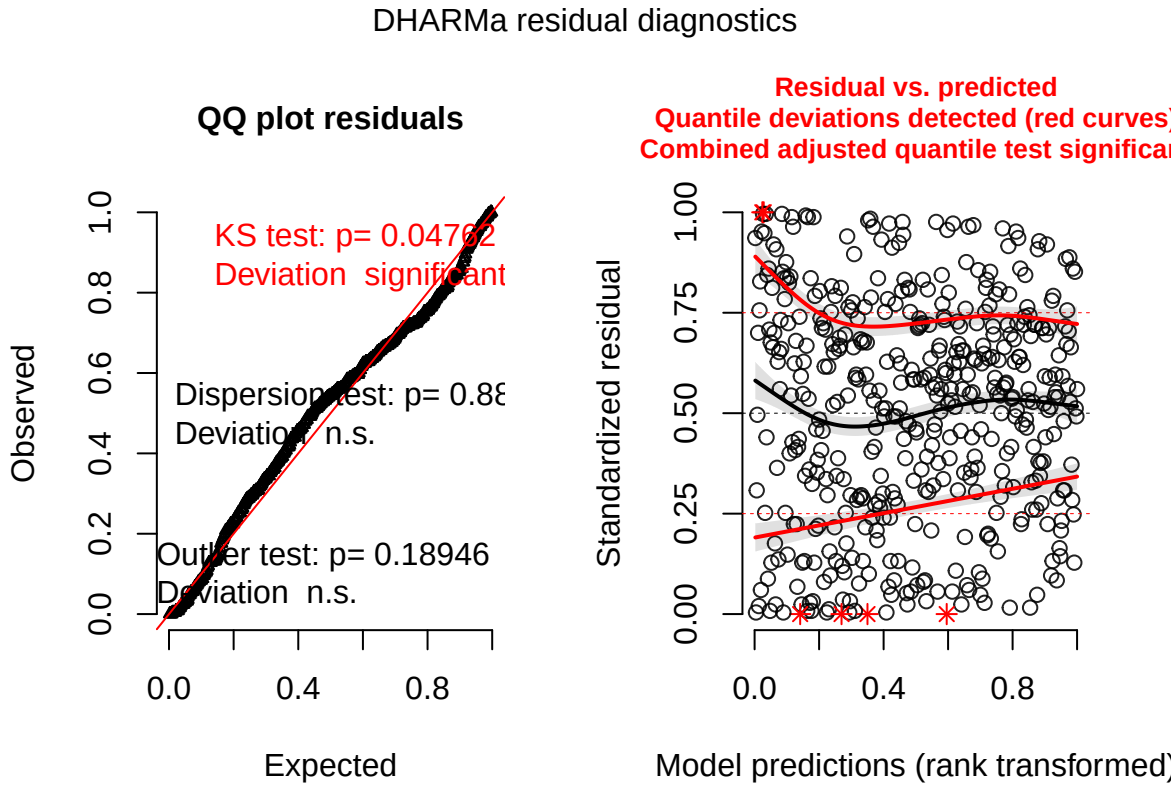
Model p-value is good, although the r^2 is not great, the significant variables are percentage of family poverty, enrolled students and total number of school in the districts. residuals and leverage also looks normal.

```
vif(lmOut4)
```

```
##  log_PctChildPoverty log_PctFamilyPoverty         log_Enrolled
##             3.570321             3.594154             6.511105
##     log_TotalSchools
##             6.423644
```

vif looks ok.

```
simulationOutput3 <- simulateResiduals(fittedModel = lmOut4, n = 250)
plot(simulationOutput3)
```



DHARMa simulation residual looks normal.

```
beta2 <- lm.beta(lm(asin_PctUpToDate ~ ., districts.filtered.uptodate.transformed))
summary(beta2)
```

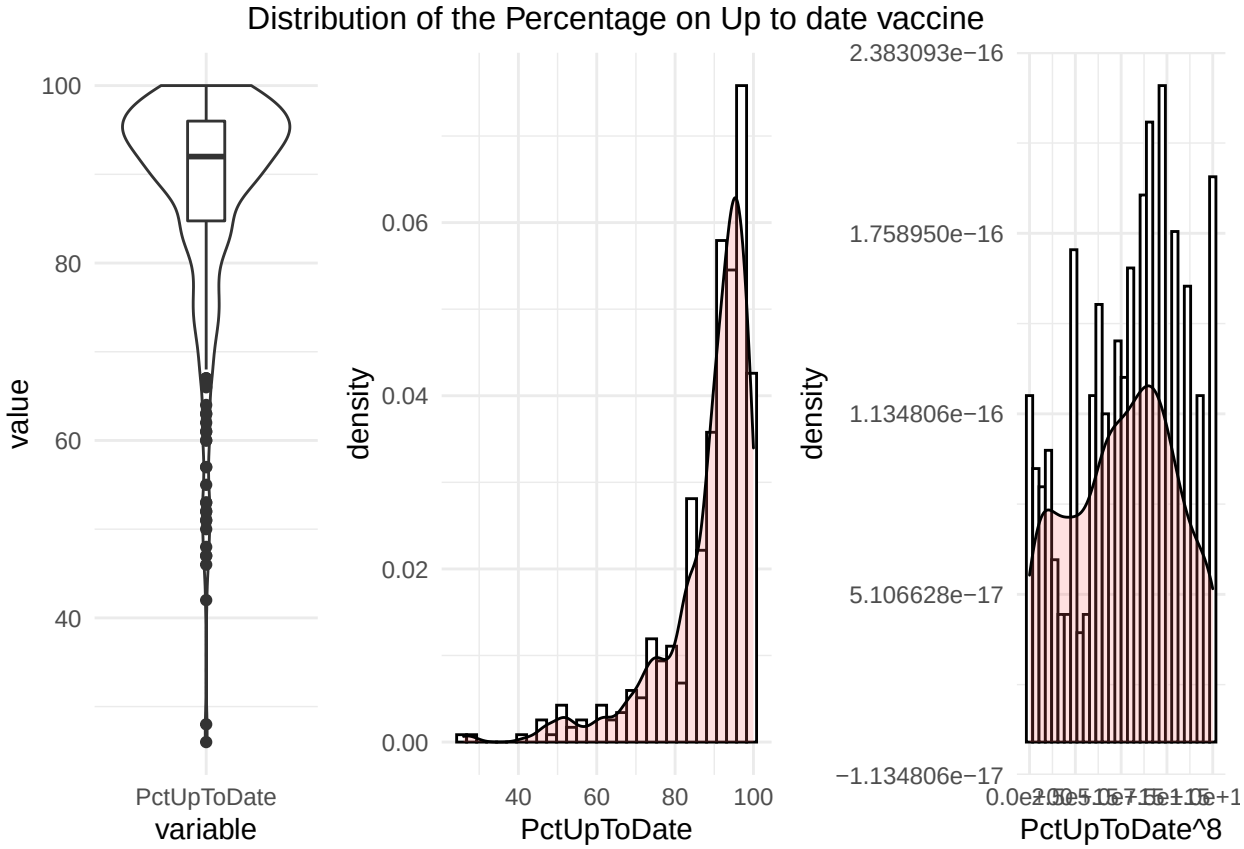
```
##
## Call:
## lm(formula = asin_PctUpToDate ~ ., data = districts.filtered.uptodate.transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79599 -0.12624  0.02219  0.13564  0.61899
##
## Coefficients:
##                       Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)            0.66660      0.00000    0.08912   7.480 3.86e-13 ***
## log_PctChildPoverty    0.01628      0.04050    0.03328   0.489  0.62492
## log_PctFamilyPoverty   0.09500      0.27371    0.02882   3.296  0.00106 **
## log_Enrolled           0.04990      0.32454    0.01719   2.903  0.00387 **
## log_TotalSchools      -0.04405     -0.20961    0.02333  -1.888  0.05967 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2252 on 455 degrees of freedom
## Multiple R-squared:  0.1269, Adjusted R-squared:  0.1192
## F-statistic: 16.53 on 4 and 455 DF,  p-value: 1.179e-12
```

There are three significant variables that predicts the percentage of up-to-date vaccines. The variables with

the best predictive power is the percentage of family poverty, then the enrolled students, and the number of schools. Increasing in the percentage of family poverty and the number of enrolled students will increase the percentage of up-to-date vaccine, while increasing in total number of school will result in decrease of the percentage of up-to-date vaccine.

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

```
p1 <- districts.filtered.uptodate %>%
  pivot_longer(cols=c(PctUpToDate) , names_to = "variable",
                           values_to = "value", values_drop_na = TRUE) %>%
  ggplot(aes(x = variable, y = value)) + geom_violin() + geom_boxplot(width = 0.2) +
    theme_minimal()

p2 <- districts.filtered.withall %>%
  ggplot(aes(x = PctUpToDate, y=..density..)) +
    geom_histogram(color="black", fill="white", bins=30) +
    geom_density(alpha=.2, fill="#FF6666") +
    theme_minimal()

p3 <- districts.filtered.withall %>%
  ggplot(aes(x = PctUpToDate^8, y=..density..)) +
    geom_histogram(color="black", fill="white", bins=30) +
    geom_density(alpha=.2, fill="#FF6666") +
    theme_minimal()

grid.arrange(p1,p2,p3,nrow = 1,widths = c(1.5,2,2),
  top = "Distribution of the Percentage on Up to date vaccine")
```



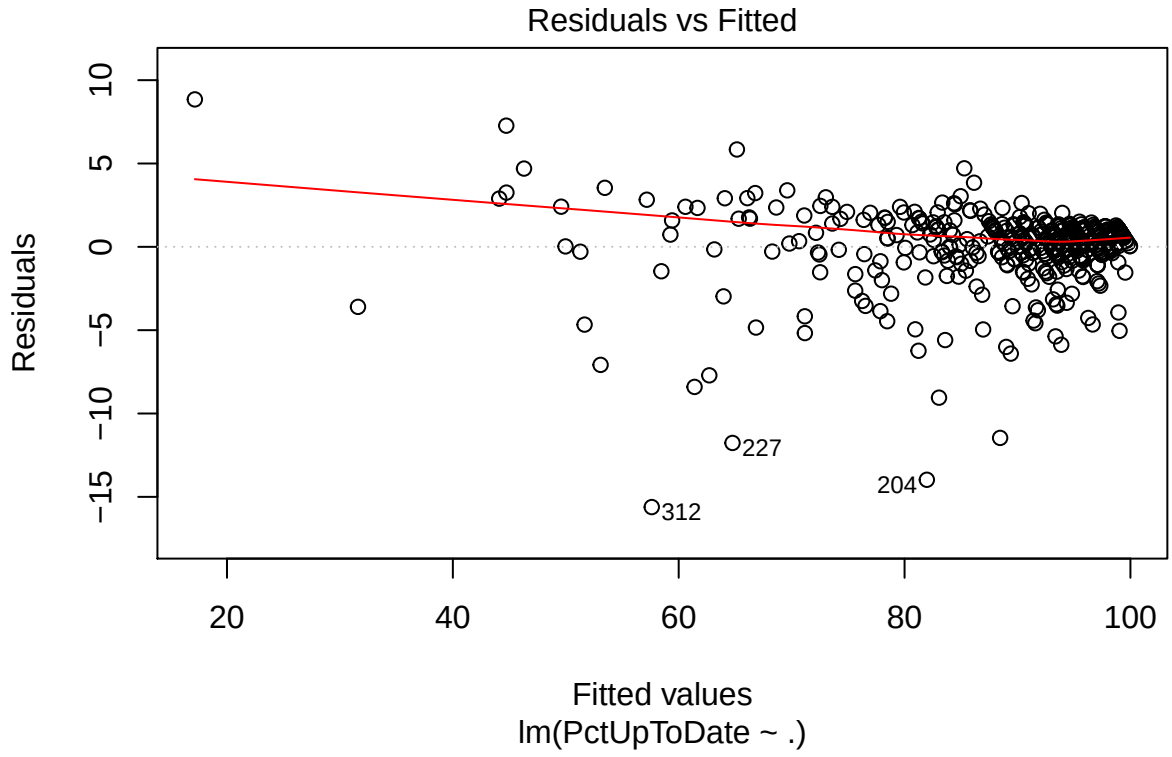
We will look at the distribution of the dependent variables first, it's very left skewed.

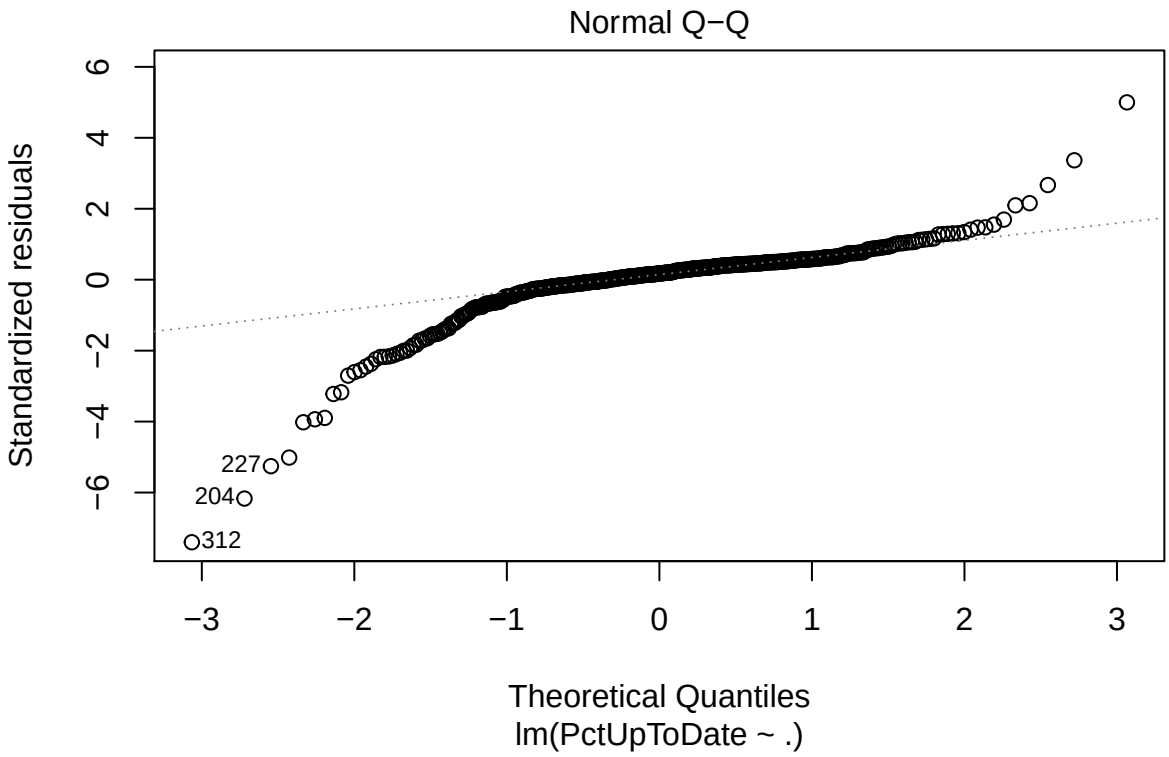
```
districts.filtered.all.features <- subset(districts.filtered, select = -c(DistrictName))

lmOut5 <- lm(PctUpToDate~. , districts.filtered.all.features)
summary(lmOut5)
```

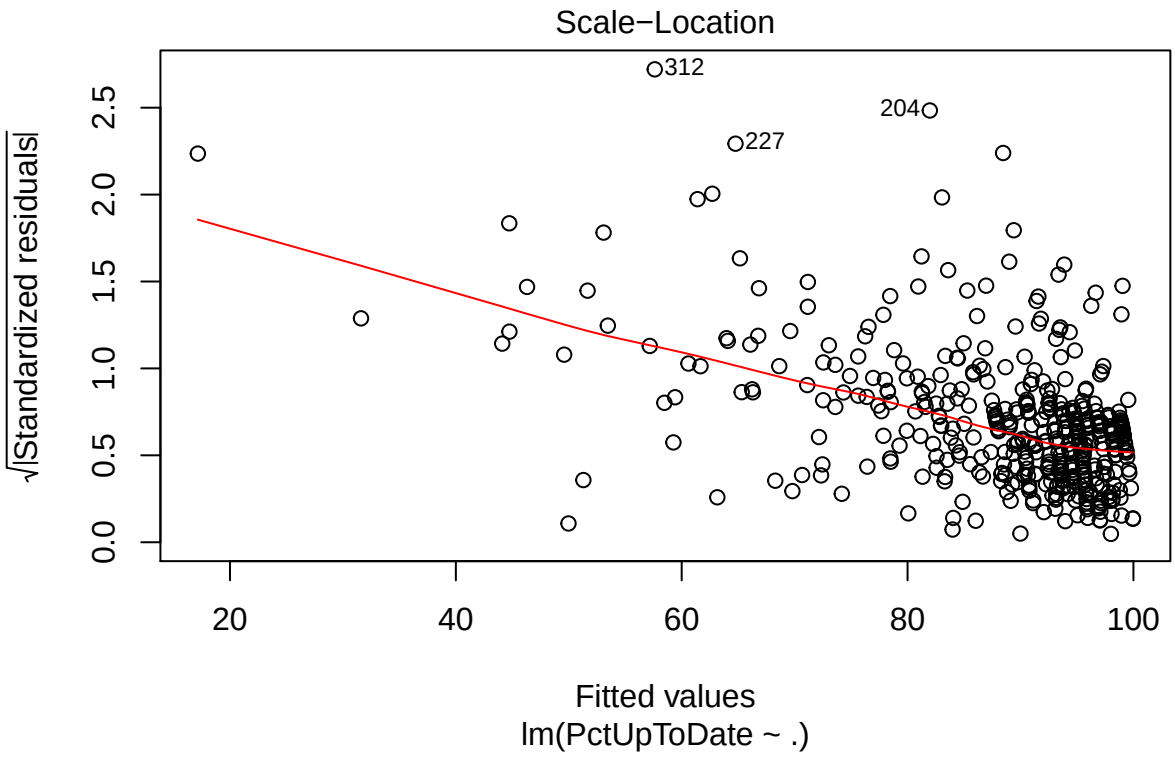
```
##
## Call:
## lm(formula = PctUpToDate ~ ., data = districts.filtered.all.features)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6130  -0.4225   0.4053   1.0844   8.8460
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.841e+01  2.739e+00 -10.373  < 2e-16 ***
## WithDTP               3.082e-01  7.220e-02   4.269 2.40e-05 ***
## WithPolio             5.773e-02  6.463e-02   0.893 0.372217
## WithMMR               7.523e-01  4.751e-02  15.836  < 2e-16 ***
## WithHepB              1.494e-01  4.395e-02   3.399 0.000736 ***
## DistrictCompleteTRUE  6.816e-01  4.598e-01   1.483 0.138883
## PctBeliefExempt       2.337e-01  3.277e-02   7.132 4.02e-12 ***
## PctMedicalExempt      2.048e-01  1.542e-01   1.328 0.184831
## PctChildPoverty      -1.746e-02  1.984e-02  -0.880 0.379232
## PctFamilyPoverty      4.833e-02  2.792e-02   1.731 0.084215 .
```

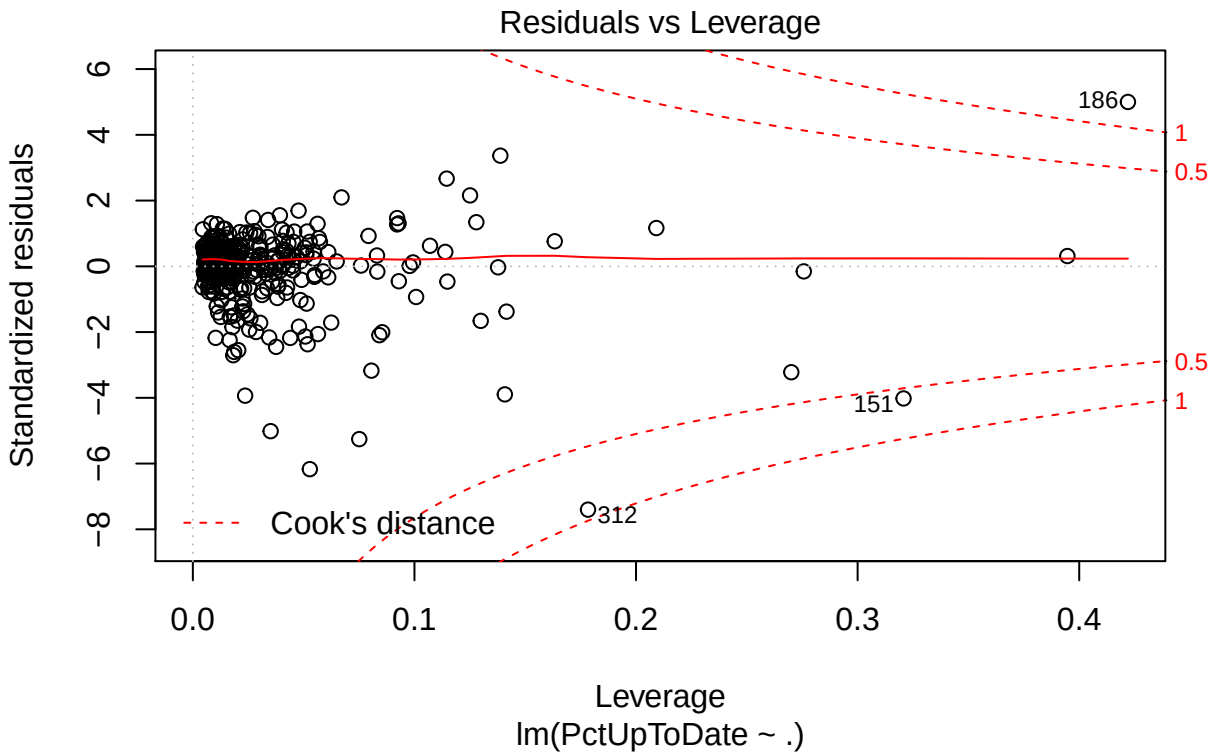
```
## PctFreeMeal          -1.295e-03  7.150e-03  -0.181 0.856316
## Enrolled              3.958e-04  5.441e-04   0.727 0.467342
## TotalSchools         -3.650e-02  4.863e-02  -0.750 0.453351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.327 on 447 degrees of freedom
## Multiple R-squared:  0.9622, Adjusted R-squared:  0.9612
## F-statistic: 949.2 on 12 and 447 DF,  p-value: < 2.2e-16
```

```
plot(lmOut5)
```

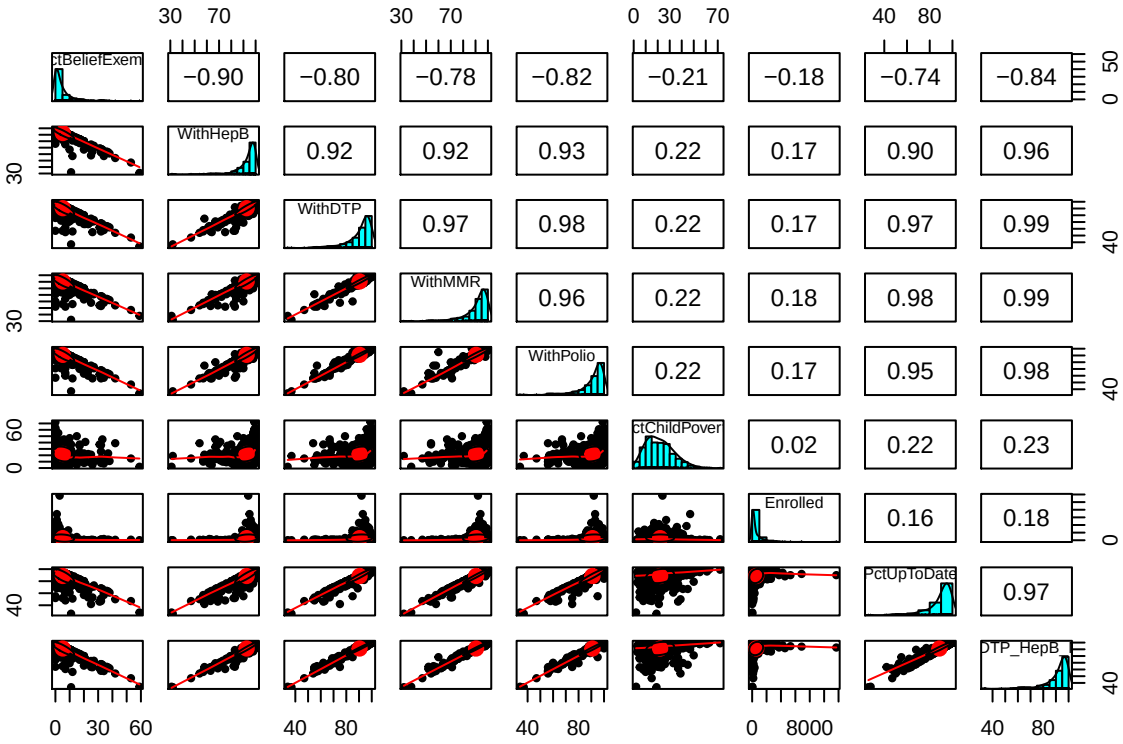



Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
−6
227
204
312
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(PctUpToDate ~ .)


Scale−Location
312
204
227
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
20
40
60
80
100
Fitted values
lm(PctUpToDate ~ .)




```
districts.filtered.all.features <- districts.filtered %>% dplyr::select(PctBeliefExempt, WithHepB, WithI
                                                WithDTP_HepB_MMR = ((WithDTP + WithMMR + WithHepB)/3))
pairs.panels(districts.filtered.all.features)
```

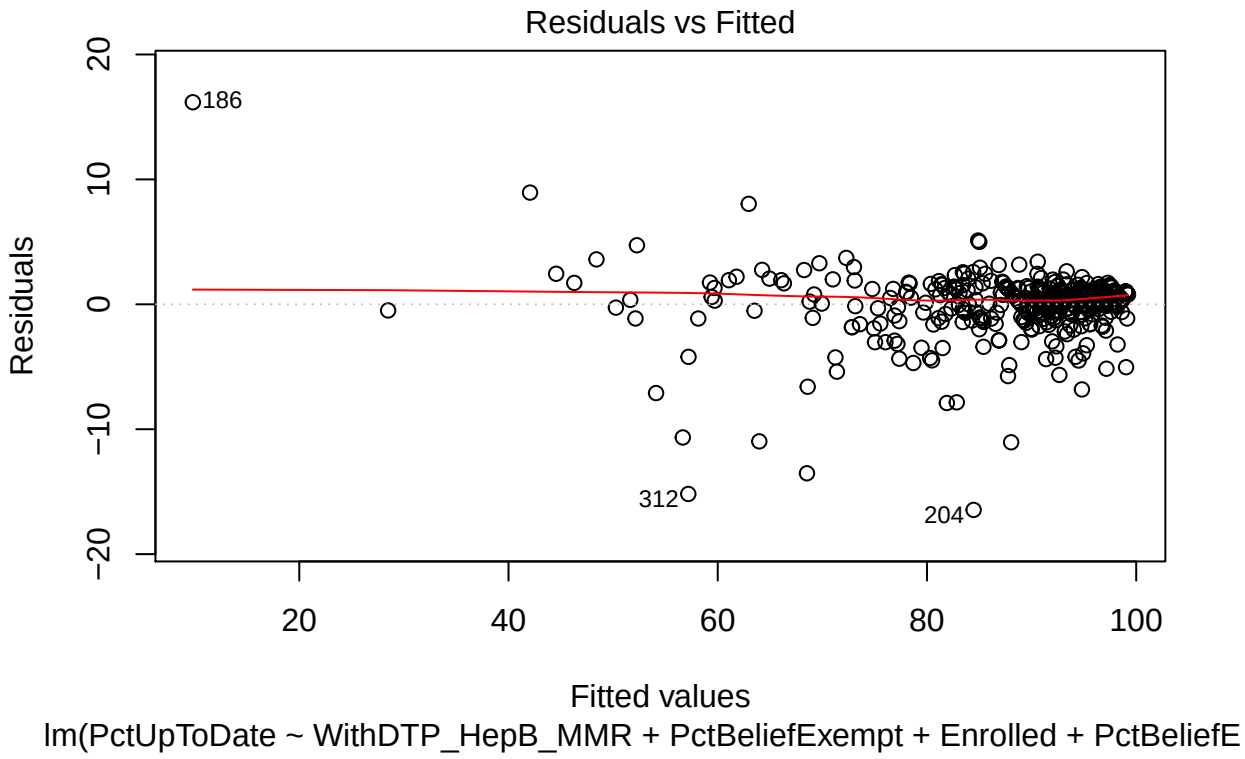


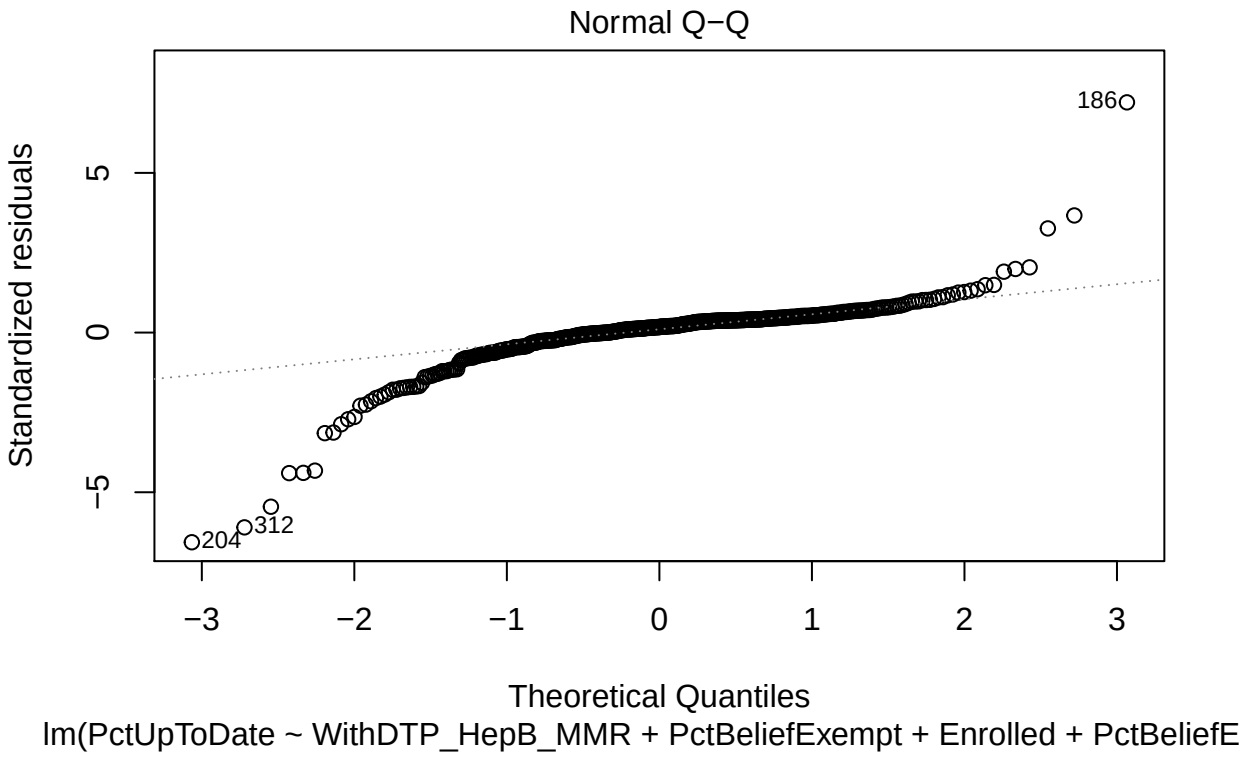

```
lmOut6 <- lm(PctUpToDate ~ WithDTP_HepB_MMR + PctBeliefExempt + Enrolled + PctBeliefExempt * Enrolled +
summary(lmOut6)
```

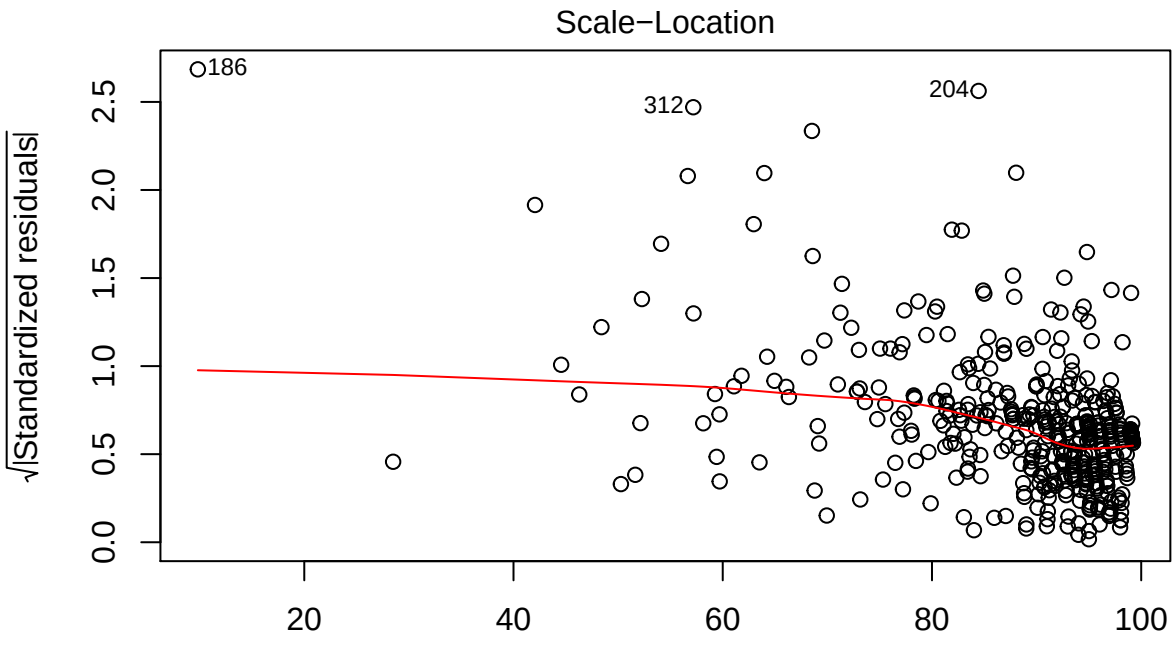
```
##
## Call:
## lm(formula = PctUpToDate ~ WithDTP_HepB_MMR + PctBeliefExempt +
##     Enrolled + PctBeliefExempt * Enrolled + PctChildPoverty +
##     Enrolled * PctChildPoverty, data = districts.filtered.all.features)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.4589  -0.5193   0.4351   1.0548  16.1753
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -4.166e+01  2.127e+00 -19.584   <2e-16 ***
## WithDTP_HepB_MMR             1.406e+00  2.218e-02  63.399   <2e-16 ***
## PctBeliefExempt              3.688e-01  2.867e-02  12.867   <2e-16 ***
## Enrolled                    -3.382e-04  4.019e-04  -0.842    0.400
## PctChildPoverty              4.153e-03  1.174e-02   0.354    0.724
## PctBeliefExempt:Enrolled     6.809e-05  6.321e-05   1.077    0.282
## Enrolled:PctChildPoverty     8.077e-06  1.243e-05   0.650    0.516
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.519 on 453 degrees of freedom
```

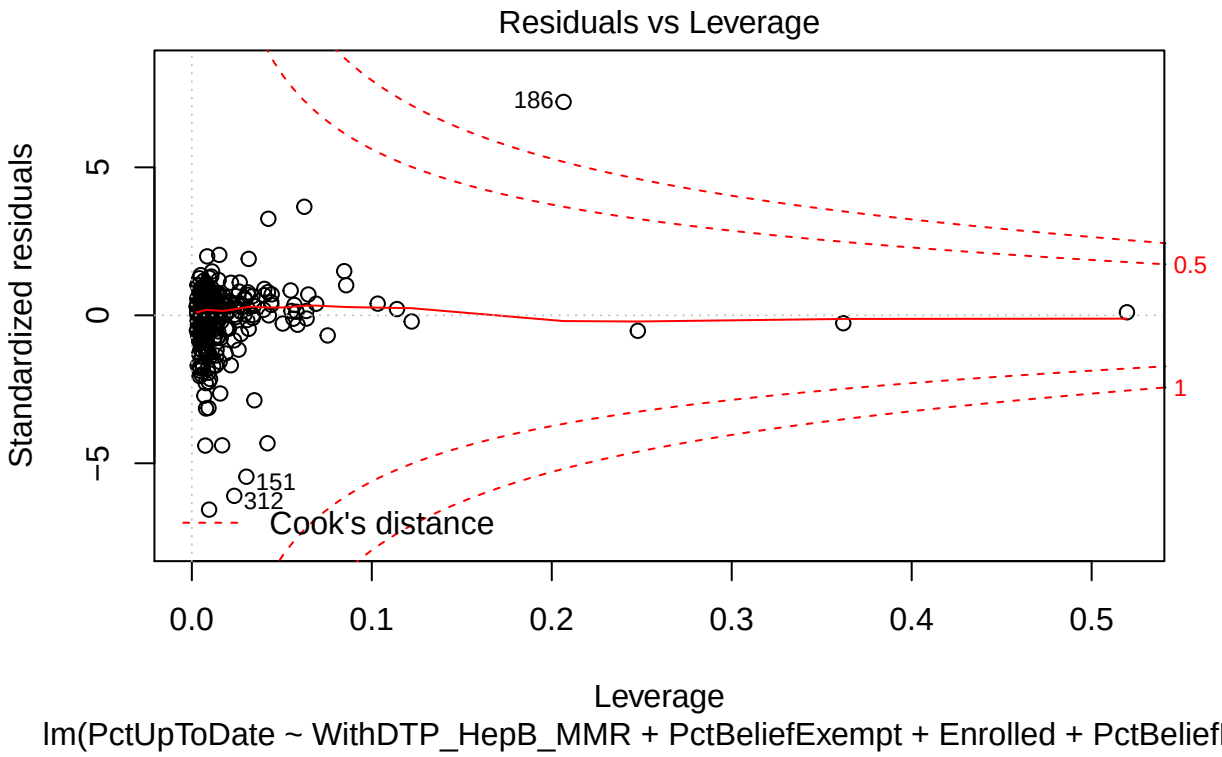
```
## Multiple R-squared:  0.9552, Adjusted R-squared:  0.9546
## F-statistic:  1608 on 6 and 453 DF,  p-value: < 2.2e-16
```

```
plot(lmOut6)
```



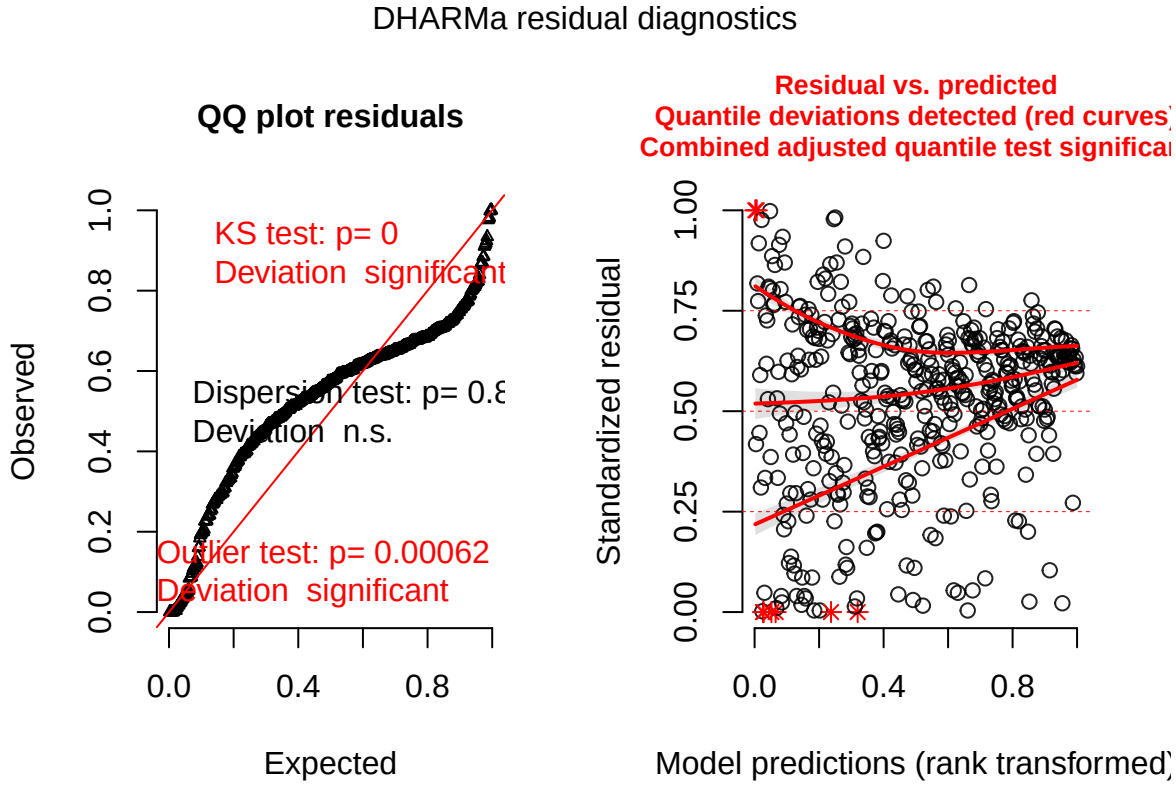

Normal Q−Q
Standardized residuals
5
0
−5
186
312
204
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(PctUpToDate ~ WithDTP_HepB_MMR + PctBeliefExempt + Enrolled + PctBeliefE


Scale−Location
186
312
204
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
20
40
60
80
100
Fitted values
lm(PctUpToDate ~ WithDTP_HepB_MMR + PctBeliefExempt + Enrolled + PctBeliefE



In this model we took the average of three vaccines, WithDTP, WithMMR, and WithHepB as the polio is not that significant in predicting the total up to date result. I also put in Belief Exempt, Enrolled, Child Poverty and interation between enrolled and child poverty, enrolled with belief Exempt. They are not significant for this model. I put them in because it gives 0.01 increase in the Adjusted R^2. Overall we got an Adjusted R^2 value of 0.9546. The residuals vs Fitted graphs looks normal. The Q-Q plot showed there are some point that's not on the line, same as the Leverage plot where we have 186 way outside of the 1.0 standard deviation, but the line is still normal. We will need look into these point more to see if its an outlier below.

```
simulationOutput4 <- simulateResiduals(fittedModel = lmOut6, n = 500)
plot(simulationOutput4)
```



The simulated residual does not look great. The test think there might be outliers. Also our dependent variables are not a normal distribution when we fit the model. This can also cause problems in the simulations.

```
vif(lmOut6)
```

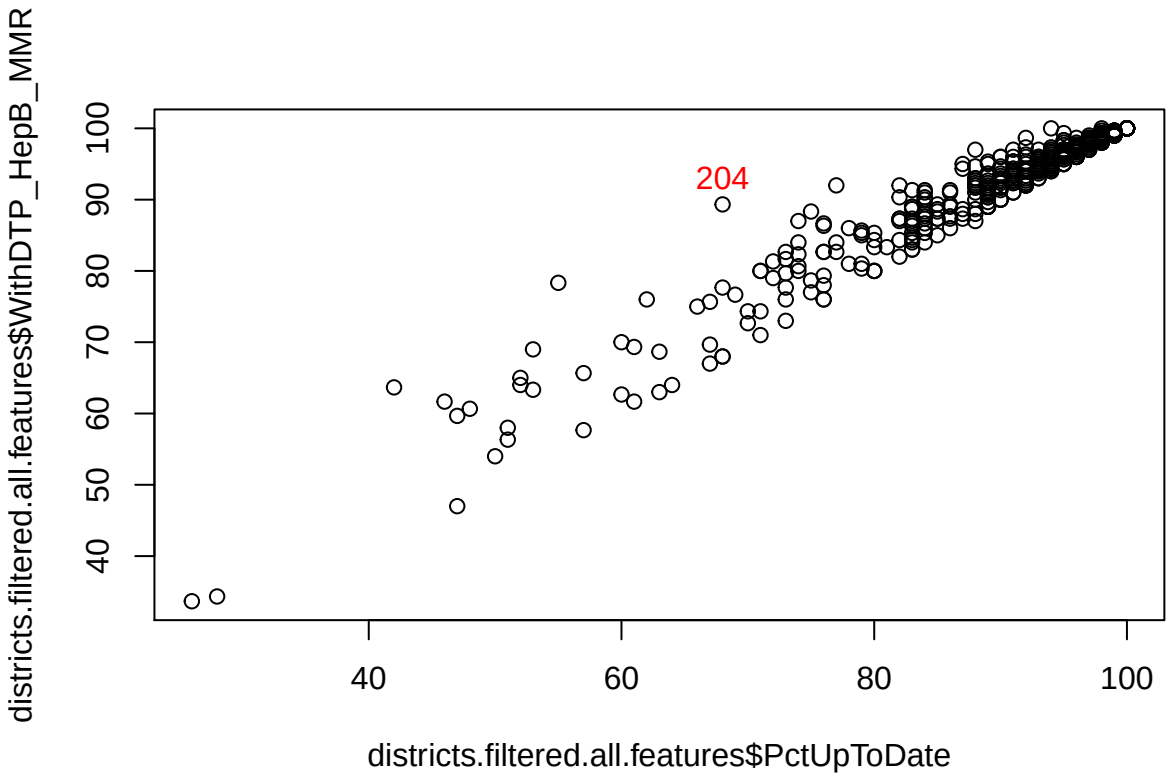
```
##          WithDTP_HepB_MMR          PctBeliefExempt                 Enrolled
##                  3.504382                 3.778776                10.861856
##           PctChildPoverty PctBeliefExempt:Enrolled Enrolled:PctChildPoverty
##                  1.361551                 2.436747                 8.330555
```

All the variable are have a vif score under 10, we can keep them.

```
districts.filtered.all.features %>% .[c(204,186),]
```

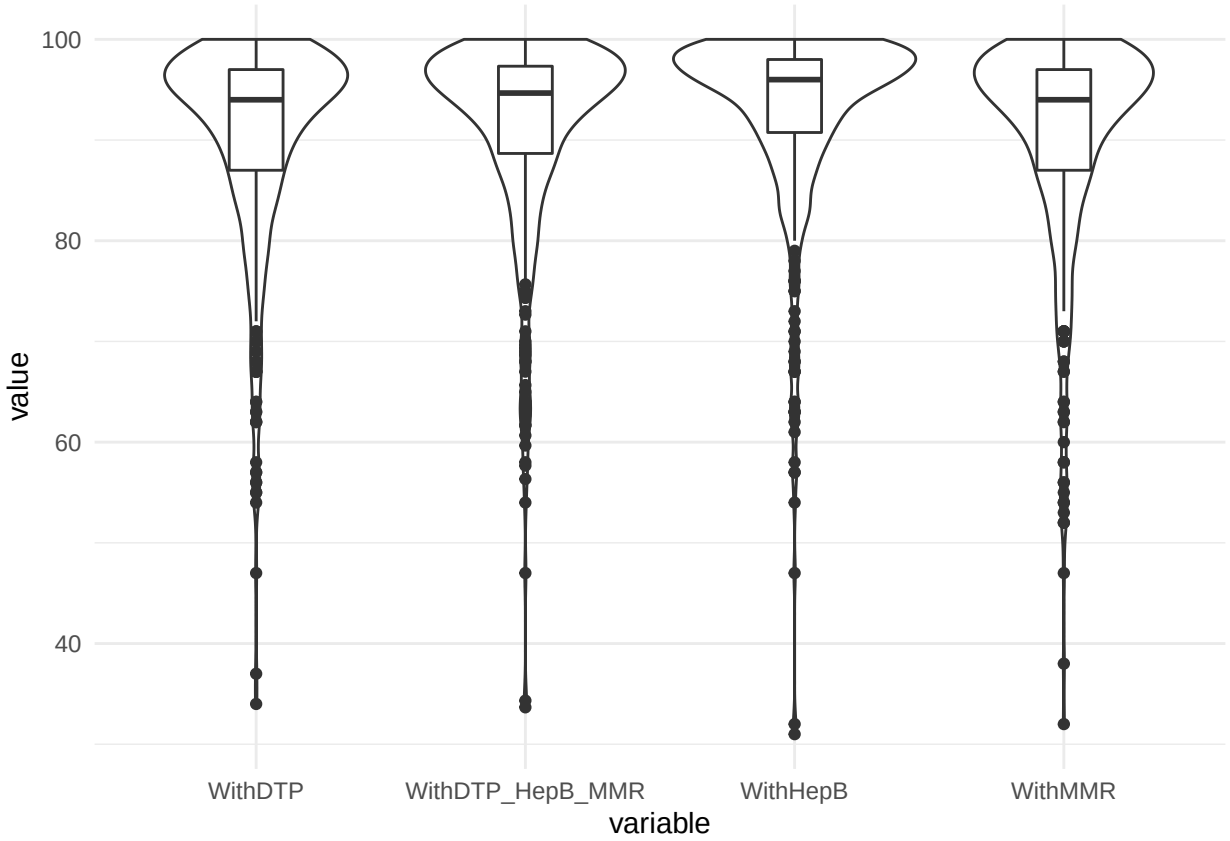
```
##     PctBeliefExempt WithHepB WithDTP WithMMR WithPolio PctChildPoverty Enrolled
## 204               1       95      89      84        85              35       75
## 186              11       32      37      32        37              19       19
##     PctUpToDate WithDTP_HepB_MMR
## 204          68         89.33333
## 186          26         33.66667
```

```
plot(districts.filtered.all.features$PctUpToDate,districts.filtered.all.features$WithDTP_HepB_MMR)
text(68,93,"204",col='red')
```



The point 204, Sausalito Marin City, is out side of the normal trend. From Google this place shows up as a tourist place, nothing special other then that. We need more information to find out if it's an outlier.

```
districts.filtered.all.features %>%
  pivot_longer(cols=c(WithHepB, WithDTP, WithMMR, WithDTP_HepB_MMR) , names_to = "variable",
                          values_to = "value", values_drop_na = TRUE) %>%
  ggplot(aes(x = variable, y = value)) + geom_violin() + geom_boxplot(width = 0.2) +
    theme_minimal()
```



Point 186, Trinidad Union Elementary, showed up way off the leverage plot, therefore are classified as outlier in diagnose analysis. No special information from google, need more investigation.

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

From the previous model. We can confidently say there is no significant evidence that PctChildPoverty, Enrolled, nor an interation variable between PctChildPoverty and Enrolled predicts the percentage of all enrolled students with completely up-to-date vaccines.

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

```
districts.filtered.completed <- subset(districts.filtered, select = -c(DistrictName))

glmOut1 <- glm(DistrictComplete~. , family = binomial(), districts.filtered.completed)
summary(glmOut1)
```

```
##
## Call:
## glm(formula = DistrictComplete ~ ., family = binomial(), data = districts.filtered.completed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8403   0.2188   0.2885   0.3794   1.4748
##
```

```
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.4968857  4.6051869   0.759   0.4477
## WithDTP          -0.0062813  0.1241459  -0.051   0.9596
## WithPolio         0.0209275  0.1141036   0.183   0.8545
## WithMMR          -0.1352420  0.0921814  -1.467   0.1423
## WithHepB          0.0303219  0.0723325   0.419   0.6751
## PctUpToDate       0.1021002  0.0737015   1.385   0.1660
## PctBeliefExempt  -0.0273908  0.0526952  -0.520   0.6032
## PctMedicalExempt  0.2120620  0.3609449   0.588   0.5569
## PctChildPoverty   0.0531613  0.0377847   1.407   0.1594
## PctFamilyPoverty -0.0925046  0.0479030  -1.931   0.0535 .
## PctFreeMeal      -0.0266979  0.0135316  -1.973   0.0485 *
## Enrolled          0.0010889  0.0008216   1.325   0.1851
## TotalSchools     -0.1157068  0.0679472  -1.703   0.0886 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 216.44  on 459  degrees of freedom
## Residual deviance: 193.69  on 447  degrees of freedom
## AIC: 219.69
##
## Number of Fisher Scoring iterations: 6
```

```
model_performance(glmOut1)
```

```
## # Indices of model performance
##
## AIC     |     BIC | Tjur's R2 |  RMSE | Sigma | Log_loss | Score_log | Score_spherical |   PCP
## ------------------------------------------------------------------------------------------------
## 219.693 | 273.399 |     0.058 | 0.238 | 0.658 |    0.211 |      -Inf |           0.002 | 0.889
```

We ran a model that used all the variables, the performance indicates model is only fitting one side of the label as the score_log is -inf. So we gonna try with less variables and see if there is a better result.

```
glmOut2 <- glm(DistrictComplete~ TotalSchools + PctFreeMeal + PctFamilyPoverty, family = binomial(), dis
summary(glmOut2)
```

```
##
## Call:
## glm(formula = DistrictComplete ~ TotalSchools + PctFreeMeal +
##     PctFamilyPoverty, family = binomial(), data = districts.filtered.completed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6560   0.2578   0.3194   0.3918   1.3337
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.98716    0.56113   7.106  1.2e-12 ***
## TotalSchools     -0.02365    0.01103  -2.143   0.0321 *
## PctFreeMeal      -0.01400    0.01195  -1.172   0.2414
## PctFamilyPoverty -0.02642    0.03044  -0.868   0.3855
## ---
```

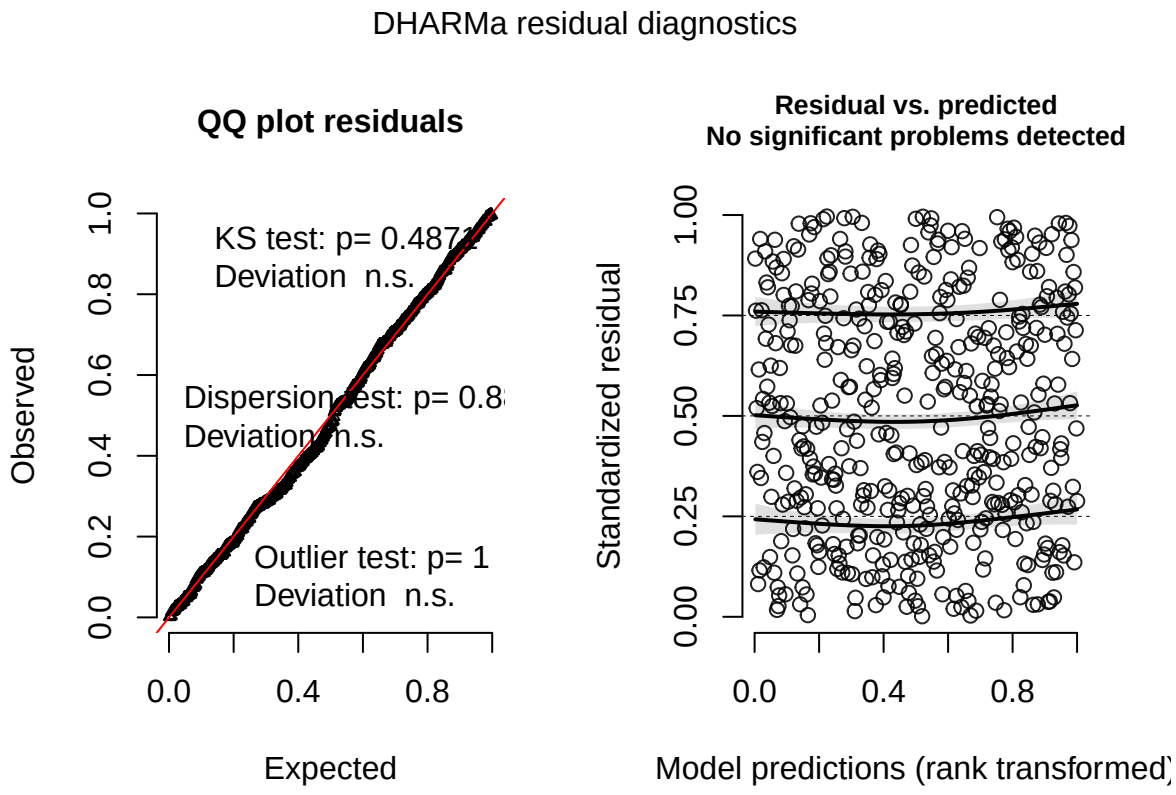
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 216.44  on 459  degrees of freedom
## Residual deviance: 205.47  on 456  degrees of freedom
## AIC: 213.47
##
## Number of Fisher Scoring iterations: 6
```

```
model_performance(glmOut2)
```

```
## # Indices of model performance
##
## AIC     |     BIC | Tjur's R2 |  RMSE | Sigma | Log_loss | Score_log | Score_spherical |   PCP
## ------------------------------------------------------------------------------------------------
## 213.472 | 229.997 |     0.025 | 0.241 | 0.671 |    0.223 |      -Inf |           0.002 | 0.885
```

Still the same results with less variables.

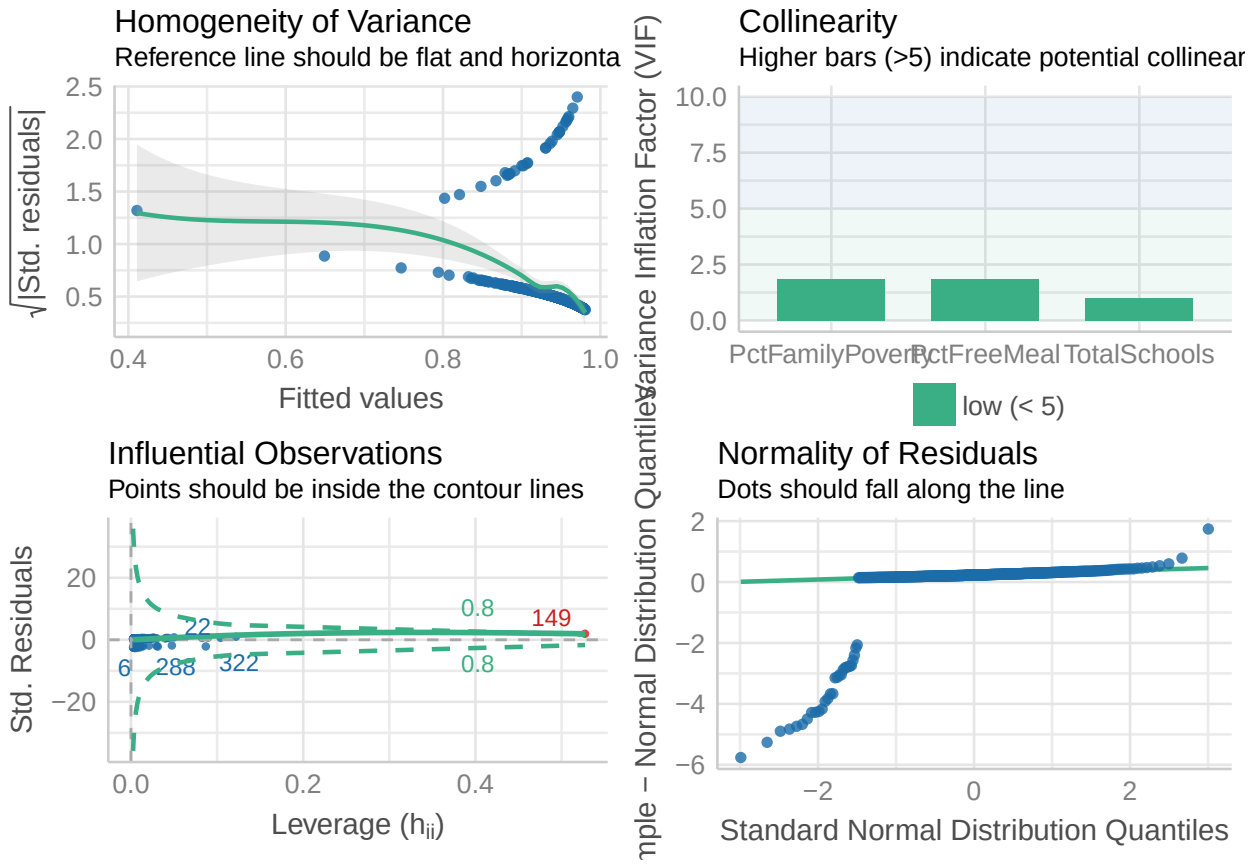
```
simulationOutput6 <- simulateResiduals(fittedModel = glmOut2, n = 250)
plot(simulationOutput6)
```



The simulation residual look normal.

```
check_model(glmOut2)
```

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



The model has bad residuals, it only fitted one side of the data. we are going to try with MCMC simulations.

```
bayesLogitOut1 <- MCMClogit(DistrictComplete~., districts.filtered.completed)
summary(bayesLogitOut1)
```

```
##
## 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)       6.068685 5.8865587 5.887e-02      5.154e-01
## WithDTP          -0.010343 0.1337110 1.337e-03      1.011e-02
## WithPolio         0.036054 0.1201432 1.201e-03      8.560e-03
## WithMMR          -0.149827 0.0989982 9.900e-04      7.354e-03
## WithHepB          0.020501 0.0841101 8.411e-04      6.761e-03
## PctUpToDate       0.090306 0.0801715 8.017e-04      5.880e-03
## PctBeliefExempt  -0.052953 0.0660855 6.609e-04      5.721e-03
## PctMedicalExempt  0.461535 0.4475284 4.475e-03      3.802e-02
## PctChildPoverty   0.058422 0.0433526 4.335e-04      3.605e-03
## PctFamilyPoverty -0.096707 0.0501464 5.015e-04      3.794e-03
## PctFreeMeal      -0.028041 0.0146934 1.469e-04      1.116e-03
## Enrolled          0.001261 0.0008913 8.913e-06      6.837e-05
```

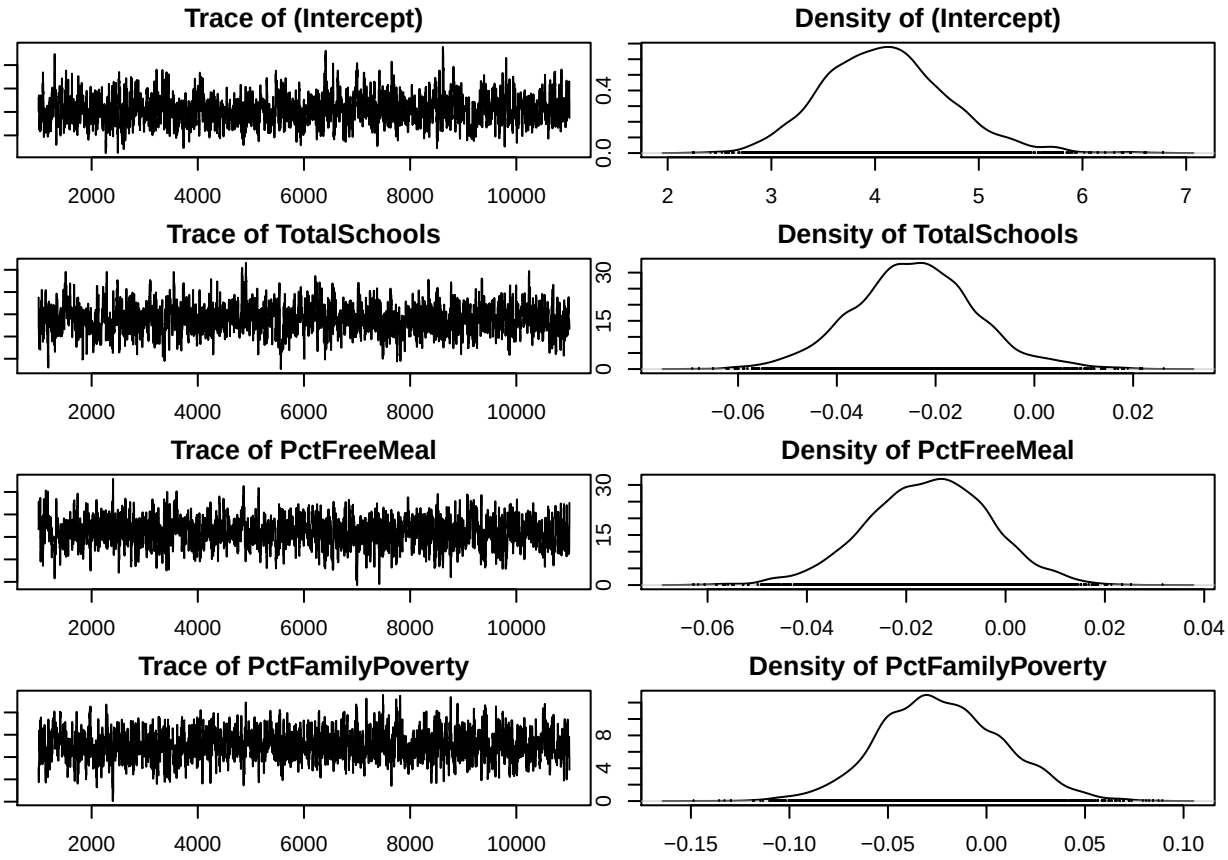
```
## TotalSchools     -0.133017 0.0747887 7.479e-04      5.789e-03
##
## 2. Quantiles for each variable:
##
##                       2.5%        25%       50%       75%      97.5%
## (Intercept)      -3.1069209  1.9335801  5.436915  9.363350 19.2816970
## WithDTP          -0.2646831 -0.0986789 -0.004475  0.072483  0.2505975
## WithPolio        -0.1752554 -0.0489880  0.020597  0.110171  0.3004939
## WithMMR          -0.3547609 -0.2138824 -0.144955 -0.079044  0.0399031
## WithHepB         -0.1388193 -0.0331040  0.022576  0.075451  0.2032810
## PctUpToDate      -0.0746209  0.0396403  0.092556  0.148118  0.2364789
## PctBeliefExempt  -0.2040750 -0.0925358 -0.044265 -0.007480  0.0597262
## PctMedicalExempt -0.2668645  0.1495132  0.402706  0.725432  1.4742866
## PctChildPoverty  -0.0186771  0.0254479  0.059506  0.090956  0.1449335
## PctFamilyPoverty -0.1932863 -0.1325017 -0.097341 -0.059048 -0.0013103
## PctFreeMeal      -0.0574259 -0.0380751 -0.026498 -0.017184 -0.0003741
## Enrolled         -0.0003547  0.0006357  0.001167  0.001785  0.0033054
## TotalSchools     -0.3049966 -0.1787281 -0.126510 -0.080749 -0.0019027
```

There are three variables that did not cross 0, but it's really really close to overlay on 0, therefore we need to test with these variables alone.

```
bayesLogitOut2 <- MCMClogit(DistrictComplete~ TotalSchools+PctFreeMeal+PctFamilyPoverty, districts.filt
summary(bayesLogitOut2)
```

```
##
## 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)       4.10103 0.59687 0.0059687      0.0231088
## TotalSchools     -0.02428 0.01246 0.0001246      0.0004878
## PctFreeMeal      -0.01558 0.01235 0.0001235      0.0004713
## PctFamilyPoverty -0.02279 0.03121 0.0003121      0.0011557
##
## 2. Quantiles for each variable:
##
##                      2.5%      25%      50%       75%    97.5%
## (Intercept)       3.03706  3.68031  4.07754  4.462656 5.375866
## TotalSchools     -0.04891 -0.03222 -0.02430 -0.016230 0.001928
## PctFreeMeal      -0.04082 -0.02373 -0.01529 -0.007206 0.008670
## PctFamilyPoverty -0.08210 -0.04471 -0.02443 -0.001911 0.039931
```

```
par(mar=c(2,1,2,1))
plot(bayesLogitOut2)
```



The simulations showed no good results here, and we can't not make a confusion matrix because the model did not fit both side of the data. With these data and model we can conclude there is no predictor that indicates if the district had completed reporting.

10. *Concluding Paragraph*

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

reference

https://www.cdc.gov/vaccines/vac-gen/whatifstop.htm