Spring 2024 - IST 686 - Final Examination

Ximeng Deng

Instructions

Your goal for this final exam is to conduct the necessary analyses of vaccination rates in California schools and school districts (due today) and then write up a technical report for a scientifically knowledgeable staff member in a California state legislator's office (due Friday).

Use this notebook during the in-class session to run the statistical analyses needed to answer the questions given. Note that not all questions for the final report are included, just the ones that require an analysis. Include both frequentist and Bayesian inferential evidence (i.e., it is not sufficient to just examine the data and say what you see). Submit the knitted notebook by the end of the exam period. Be sure to proofread your final knitted submission to ensure that everything is included and readable (e.g., that the code does not run off the edge of the page).

You may not *seek nor receive assistance, help, coaching, guidance, or support from any system or any human except your instructor at any point during this exam. Seeking or obtaining improper assistance will result in a 0 for this exam.*

Data

You have a personalized RData file available on Blackboard area that contains two data sets that pertain to vaccinations for the United States as a whole and for Californian school districts. In addition, you will find on Blackboard a CSV file, All Schools.csv, with data about 7,381 individual schools. A description of the data was provided earlier.

Paste the R code to read your data and for your data exploration and cleaning here. Be sure to explain what you were looking for and your reasons for any changes.

read all schools csv file and load datasets

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(psych)
```

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

```
schools <- read_csv("/Users/ximengdeng/Desktop/2024spring/ist686/final/All Schools(1).csv")
```

```
## Rows: 7381 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): PUBLIC/ PRIVATE, Public School District ID, PUBLIC SCHOOL DISTRICT...
## dbl (11): SCHOOL CODE, ENROLLMENT, UP_TO_DATE, CONDITIONAL, PME, PBE_BETA, D...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
load("~/Desktop/2024spring/ist686/final/datasets9(2)(1).RData")
```

get descriptive statistics

```
summary(districts)
```

```
##                    DistrictName    WithDTP        WithPolio
##  ABC Unified              :  1   Min.   : 23.0   Min.   : 23.0
##  Ackerman Charter         :  1   1st Qu.: 86.0   1st Qu.: 87.0
##  Acton-Agua Dulce Unified:  1   Median : 93.0   Median : 94.0
##  Adelanto Elementary      :  1   Mean   : 89.8   Mean   : 90.2
##  Alameda Unified          :  1   3rd Qu.: 97.0   3rd Qu.: 97.0
##  Albany City Unified      :  1   Max.   :100.0   Max.   :100.0
##  (Other)                  :694
##     WithMMR          WithHepB       PctUpToDate    DistrictComplete
##  Min.   : 23.00   Min.   : 23.00   Min.   : 23.0   Mode :logical
##  1st Qu.: 86.00   1st Qu.: 90.00   1st Qu.: 84.0   FALSE:43
##  Median : 94.00   Median : 96.00   Median : 92.0   TRUE :657
##  Mean   : 89.79   Mean   : 92.26   Mean   : 88.4
##  3rd Qu.: 97.00   3rd Qu.: 98.00   3rd Qu.: 96.0
##  Max.   :100.00   Max.   :100.00   Max.   :200.0
##
##  PctBeliefExempt  PctMedicalExempt PctChildPoverty PctFamilyPoverty
##  Min.   : 0.000   Min.   :0.0000   Min.   : 2.00   Min.   : 0.00
##  1st Qu.: 0.750   1st Qu.:0.0000   1st Qu.:13.00   1st Qu.: 5.00
##  Median : 2.000   Median :0.0000   Median :21.00   Median : 9.00
##  Mean   : 5.623   Mean   :0.1471   Mean   :22.33   Mean   :11.48
##  3rd Qu.: 7.000   3rd Qu.:0.0000   3rd Qu.:29.00   3rd Qu.:16.00
##  Max.   :77.000   Max.   :8.0000   Max.   :72.00   Max.   :47.00
```

```
##
##   PctFreeMeal        Enrolled        TotalSchools
##  Min.   :  0.00   Min.   :   10.0   Min.   :  1.000
##  1st Qu.: 28.75   1st Qu.:   50.5   1st Qu.:  1.000
##  Median : 48.00   Median :  201.5   Median :  3.000
##  Mean   : 47.65   Mean   :  630.8   Mean   :  7.253
##  3rd Qu.: 69.00   3rd Qu.:  684.2   3rd Qu.:  8.000
##  Max.   :100.00   Max.   :54238.0   Max.   :582.000
##  NA's   :244
```

```
str(districts)
```

```
## 'data.frame':    700 obs. of  14 variables:
##  $ DistrictName    : Factor w/ 846 levels "ABC Unified",..: 316 582 663 35 785 291 392 374 39 55 ...
##  $ WithDTP         : num  100 100 95 96 89 93 96 97 93 98 ...
##  $ WithPolio       : num  100 100 96 97 89 94 96 97 95 98 ...
##  $ WithMMR         : num  100 100 97 98 89 98 97 97 90 98 ...
##  $ WithHepB        : num  100 100 97 98 89 95 97 97 95 99 ...
##  $ PctUpToDate     : num  100 94 95 96 89 91 95 97 90 98 ...
##  $ DistrictComplete: logi  TRUE TRUE TRUE FALSE TRUE FALSE ...
##  $ PctBeliefExempt : num  0 0 2 1 11 0 2 3 1 1 ...
##  $ PctMedicalExempt: num  0 0 0 0 0 0 0 0 2 0 ...
##  $ PctChildPoverty : num  29 23 4 41 27 20 24 32 44 27 ...
##  $ PctFamilyPoverty: num  10 11 3 22 14 10 13 18 27 14 ...
##  $ PctFreeMeal     : num  32 72 3 NA 69 NA 56 75 79 78 ...
##  $ Enrolled        : num  42 17 2321 613 64 ...
##  $ TotalSchools    : num  1 1 21 8 1 23 10 1 31 4 ...
##  - attr(*, "na.action")= 'omit' Named int [1:67] 14 40 63 107 124 158 182 196 199 207 ...
##   ..- attr(*, "names")= chr [1:67] "14" "40" "63" "107" ...
```

```
describe(districts)
```

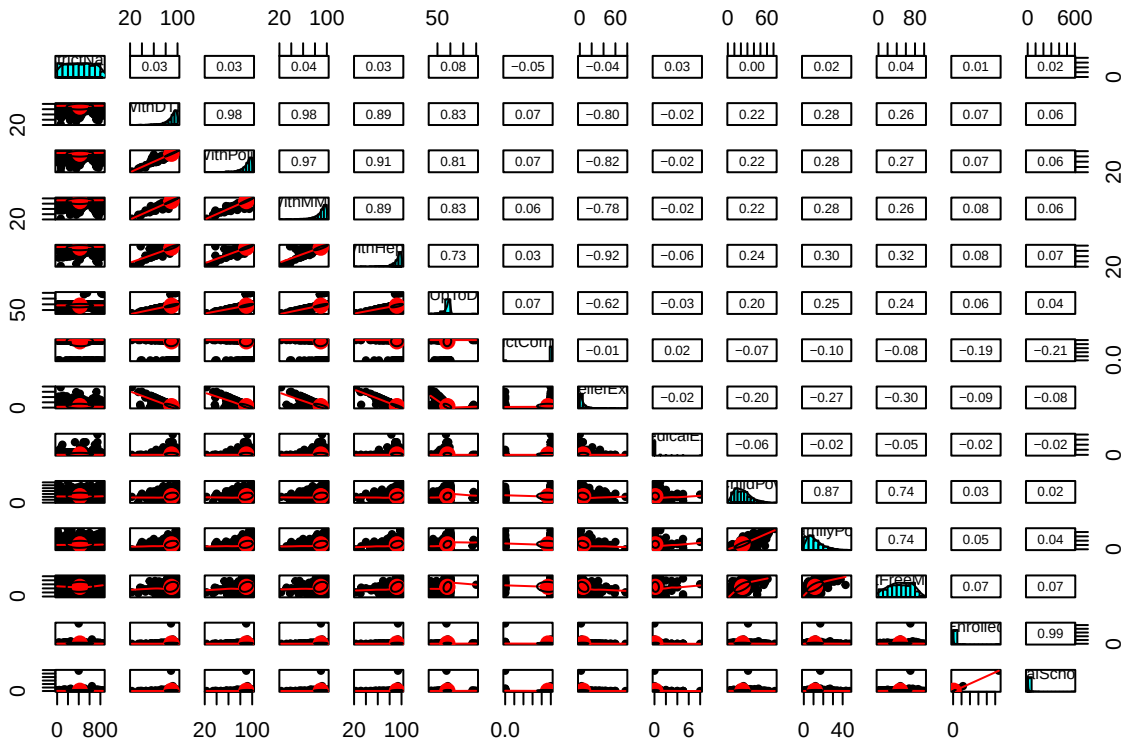
```
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
```

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

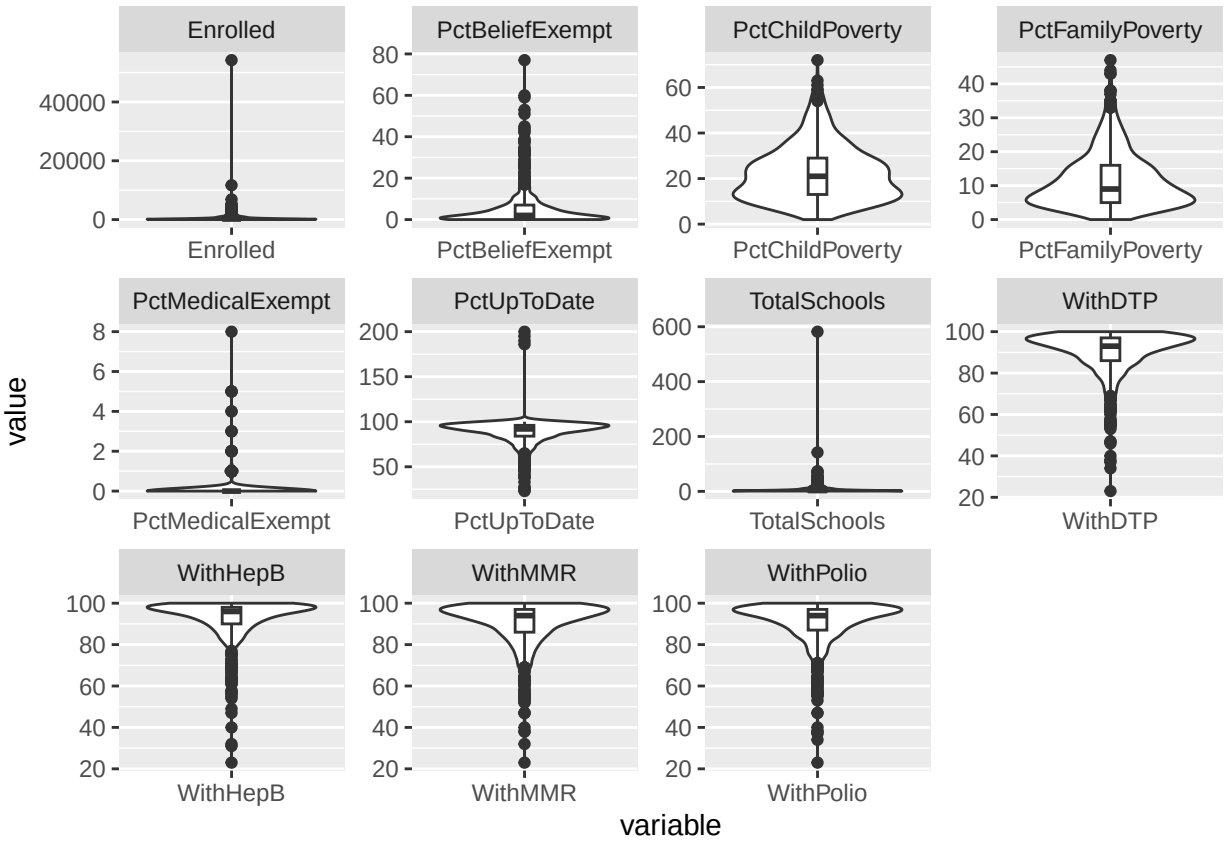
```
##                  vars   n   mean      sd median trimmed    mad min   max range
## DistrictName*       1 700 420.49  246.21  417.5  420.02 309.86   1   846   845
## WithDTP             2 700  89.80   11.03   93.0   91.85   7.41  23   100    77
## WithPolio           3 700  90.20   11.00   94.0   92.29   5.93  23   100    77
## WithMMR             4 700  89.79   11.34   94.0   91.92   5.93  23   100    77
## WithHepB            5 700  92.26    9.90   96.0   94.26   4.45  23   100    77
## PctUpToDate         6 700  88.40   14.99   92.0   90.19   7.41  23   200   177
## DistrictComplete    7 700    NaN      NA     NA     NaN     NA Inf  -Inf  -Inf
## PctBeliefExempt     8 700   5.62    8.86    2.0    3.63   2.97   0    77    77
## PctMedicalExempt    9 700   0.15    0.63    0.0    0.00   0.00   0     8     8
## PctChildPoverty    10 700  22.33   12.18   21.0   21.19  11.86   2    72    70
## PctFamilyPoverty   11 700  11.48    8.18    9.0   10.43   7.41   0    47    47
## PctFreeMeal        12 456  47.65   24.66   48.0   48.16  29.65   0   100   100
## Enrolled           13 700 630.75 2227.23  201.5  360.07 263.16  10 54238 54228
## TotalSchools       14 700   7.25   24.07    3.0    4.34   2.97   1   582   581
##                   skew kurtosis    se
```

```
## DistrictName*     0.01    -1.20  9.31
## WithDTP          -2.16     5.93  0.42
## WithPolio        -2.26     6.50  0.42
## WithMMR          -2.11     5.44  0.43
## WithHepB         -2.82    10.68  0.37
## PctUpToDate       0.57    15.26  0.57
## DistrictComplete    NA       NA    NA
## PctBeliefExempt   3.26    14.18  0.33
## PctMedicalExempt  6.74    57.31  0.02
## PctChildPoverty   0.83     0.46  0.46
## PctFamilyPoverty  1.21     1.45  0.31
## PctFreeMeal      -0.13    -0.95  1.15
## Enrolled         20.22   477.56 84.18
## TotalSchools     19.86   463.05  0.91
```

```
pairs.panels(districts)
```




```
districts %>%
  pivot_longer(cols = c("WithDTP", "WithPolio", "WithMMR", "WithHepB", "PctUpToDate",
                        "PctBeliefExempt", "PctMedicalExempt", "PctChildPoverty", "PctFamilyPoverty", #
                        "Enrolled", "TotalSchools"),
               names_to = "variable",
               values_to = "value") %>%
  ggplot(aes(x = variable, y = value, values_drop_na = TRUE)) +
  geom_violin() +
  geom_boxplot(width=0.1) +
  facet_wrap(~ variable, scales = "free")
```




```
describe(schools)
```

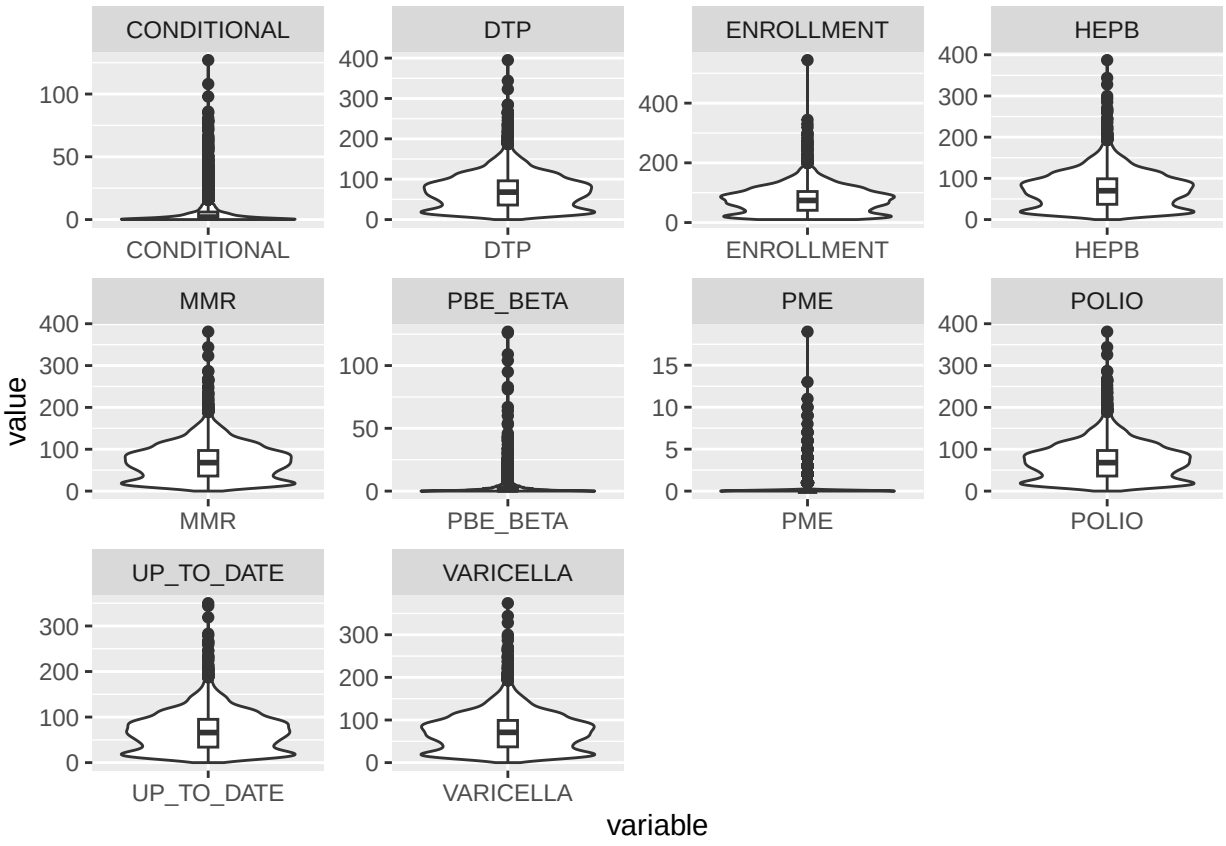
```
##                            vars    n       mean         sd  median    trimmed
## SCHOOL CODE                   1 7381 5546463.18 1978353.31 6043806 6041254.70
## PUBLIC/ PRIVATE*              2 7381       1.78       0.42       2       1.85
## Public School District ID*    3 7381     523.66     257.75     531     539.27
## PUBLIC SCHOOL DISTRICT*       4 7381     468.26     209.51     538     479.27
## CITY*                         5 7381     490.49     248.02     511     498.91
## COUNTY*                       6 7381      28.44      13.75      30      28.43
## SCHOOL NAME*                  7 7381    3153.71    1795.41    3138    3155.02
## ENROLLMENT                    8 6982      75.99      43.80      74      73.20
## UP_TO_DATE                    9 6982      68.56      41.63      66      65.73
## CONDITIONAL                  10 6982       4.93       8.54       2       3.06
## PME                          11 6982       0.14       0.66       0       0.00
## PBE_BETA                     12 6982       2.35       5.27       1       1.39
## DTP                          13 6982      70.10      42.03      68      67.28
## POLIO                        14 6982      70.44      42.22      68      67.63
## MMR                          15 6982      70.21      42.18      68      67.42
## HEPB                         16 6982      72.06      42.69      70      69.27
## VARICELLA                    17 6982      72.44      42.87      71      69.65
## REPORTED*                    18 7381       1.95       0.23       2       2.00
##                               mad    min     max   range  skew kurtosis
## SCHOOL CODE                62811.83 100016 7105125 7005109 -2.25     3.44
## PUBLIC/ PRIVATE*               0.00      1       2       1 -1.33    -0.24
## Public School District ID*   346.93      1     847     846 -0.26    -1.12
## PUBLIC SCHOOL DISTRICT*      198.67      1     848     847 -0.51    -0.63
## CITY*                        286.14      1     913     912 -0.28    -1.02
```

```
## COUNTY*                       16.31      1      58      57 -0.01    -0.65
## SCHOOL NAME*                2263.93      1    6266    6265  0.00    -1.17
## ENROLLMENT                    45.96     10     544     534  0.80     2.52
## UP_TO_DATE                    44.48      0     350     350  0.71     1.04
## CONDITIONAL                    2.97      0     127     127  4.21    28.17
## PME                            0.00      0      19      19 10.50   178.06
## PBE_BETA                       1.48      0     127     127 10.22   176.52
## DTP                           44.48      0     395     395  0.73     1.24
## POLIO                         44.48      0     381     381  0.71     1.13
## MMR                           44.48      0     381     381  0.71     1.12
## HEPB                          45.96      0     387     387  0.70     1.15
## VARICELLA                     45.96      0     374     374  0.69     1.07
## REPORTED*                      0.00      1       2       1 -3.94    13.51
##                                  se
## SCHOOL CODE                23027.47
## PUBLIC/ PRIVATE*               0.00
## Public School District ID*     3.00
## PUBLIC SCHOOL DISTRICT*        2.44
## CITY*                          2.89
## COUNTY*                        0.16
## SCHOOL NAME*                  20.90
## ENROLLMENT                     0.52
## UP_TO_DATE                     0.50
## CONDITIONAL                    0.10
## PME                            0.01
## PBE_BETA                       0.06
## DTP                            0.50
## POLIO                          0.51
## MMR                            0.50
## HEPB                           0.51
## VARICELLA                      0.51
## REPORTED*                      0.00
```

```
schools %>%
  pivot_longer(cols = c("ENROLLMENT", "UP_TO_DATE", "CONDITIONAL", "PME", "PBE_BETA",
                        "DTP", "POLIO", "MMR", "HEPB", "VARICELLA", "CONDITIONAL","PME", "PBE_BETA"),
               names_to = "variable",
               values_to = "value") %>%
  ggplot(aes(x = variable, y = value, values_drop_na = TRUE)) +
  geom_violin() +
  geom_boxplot(width=0.1) +
  facet_wrap(~ variable, scales = "free")
```

```
## Warning: Removed 3990 rows containing non-finite values (`stat_ydensity()`).
```

```
## Warning: Removed 3990 rows containing non-finite values (`stat_boxplot()`).
```



Check for missing values

```
sapply(districts, function(x) sum(is.na(x)))
```

```
##     DistrictName          WithDTP        WithPolio          WithMMR
##                0                0                0                0
##         WithHepB      PctUpToDate DistrictComplete  PctBeliefExempt
##                0                0                0                0
## PctMedicalExempt  PctChildPoverty PctFamilyPoverty      PctFreeMeal
##                0                0                0              244
##         Enrolled     TotalSchools
##                0                0
```

There are 244 NAs out of 700 observations in the PctFreeMeal column. I may consider dropping this column for future analysis if necessary.

```
library(dplyr)
districts <- districts %>%
  dplyr::select(-PctFreeMeal)
```

Correct the data for PctUpToDate in the districts dataset

We observed that 4 districts have a percentage of students with completely up-to-date vaccines exceeding 100%, which is implausible. Therefore, I checked all schools in those 4 districts within the all schools dataset,

and calculated their percentages manually by grouping by `PUBLIC SCHOOL DISTRICT`, calculating the sum of `UP_TO_DATE` divided by the sum of `ENROLLMENT`. Upon rounding, we found that the results were consistently 100 less than the erroneous data. Thus, I corrected the `PctUpToDate` in the districts dataset by subtracting 100 from those 4 rows.

```
library(stringr)

district_Pct_over_100 <- districts %>%
  filter(PctUpToDate > 100) %>%
  dplyr::select(DistrictName) %>%
  mutate(DistrictName = tolower(DistrictName))

matchingSchools <- schools %>%
  mutate(`PUBLIC SCHOOL DISTRICT` = tolower(`PUBLIC SCHOOL DISTRICT`)) %>%
  semi_join(district_Pct_over_100, by = c(`PUBLIC SCHOOL DISTRICT` = "DistrictName"))

print(matchingSchools)
```

```
## # A tibble: 37 x 18
##    `SCHOOL CODE` `PUBLIC/ PRIVATE` Public School Distri~1 PUBLIC SCHOOL DISTRI~2
##            <dbl> <chr>             <chr>                  <chr>
##  1       6029219 PUBLIC            627240                 newport-mesa unified
##  2       6029227 PUBLIC            627240                 newport-mesa unified
##  3       6029268 PUBLIC            627240                 newport-mesa unified
##  4       6029300 PUBLIC            627240                 newport-mesa unified
##  5       6029326 PUBLIC            627240                 newport-mesa unified
##  6       6029334 PUBLIC            627240                 newport-mesa unified
##  7       6029375 PUBLIC            627240                 newport-mesa unified
##  8       6029391 PUBLIC            627240                 newport-mesa unified
##  9       6029409 PUBLIC            627240                 newport-mesa unified
## 10       6029433 PUBLIC            627240                 newport-mesa unified
## # i 27 more rows
## # i abbreviated names: 1: `Public School District ID`,
## #   2: `PUBLIC SCHOOL DISTRICT`
## # i 14 more variables: CITY <chr>, COUNTY <chr>, `SCHOOL NAME` <chr>,
## #   ENROLLMENT <dbl>, UP_TO_DATE <dbl>, CONDITIONAL <dbl>, PME <dbl>,
## #   PBE_BETA <dbl>, DTP <dbl>, POLIO <dbl>, MMR <dbl>, HEPB <dbl>,
## #   VARICELLA <dbl>, REPORTED <chr>
```

```
matchingSchools %>%
  group_by(`PUBLIC SCHOOL DISTRICT`) %>%
  summarise(
    Total_UpToDate = sum(UP_TO_DATE, na.rm = TRUE),
    Total_Enrollment = sum(ENROLLMENT, na.rm = TRUE),
    Percentage_UpToDate = (sum(UP_TO_DATE, na.rm = TRUE) / sum(ENROLLMENT, na.rm = TRUE)) * 100 )
```

```
## # A tibble: 4 x 4
##   `PUBLIC SCHOOL DISTRICT` Total_UpToDate Total_Enrollment Percentage_UpToDate
##   <chr>                              <dbl>            <dbl>               <dbl>
## 1 newport-mesa unified                1458             1706                85.5
## 2 pierce joint unified                 114              115                99.1
## 3 westminster                         1222             1298                94.1
## 4 yreka union elementary               110              123                89.4
```

```r
districts <- districts %>%
  mutate(PctUpToDate = ifelse(PctUpToDate > 100, PctUpToDate - 100, PctUpToDate))
```

In the districts dataset, "Enrolled", "TotalSchools" are highly skewed, so I applied log transformation.

In the all schools dataset, the variables CONDITIONAL, PME, and PBE_BETA are highly skewed; And in the districts dataset, 'PctBeliefExempt' and 'PctMedicalExempt' are also highly skewed, primarily due to many zeros, which precludes the use of log transformation. Although research offers several solutions, I will maintain their current values for now.

Boulton, A. J., & Williford, A. (2018). Analyzing skewed continuous outcomes with many zeros: A tutorial for social work and youth prevention science researchers. Journal of the Society for Social Work and Research, 9(4), 721-740. https://doi.org/10.1086/701235

log transformation on Enrolled and TotalSchools

```r
districts$EnrolledLog <- log(districts$Enrolled)
districts$TotalSchoolsLog <- log(districts$TotalSchools)
districts <- districts %>%
  dplyr::select(-Enrolled, -TotalSchools)
```

```r
cor(districts[sapply(districts, is.numeric)], use = "complete.obs")
```

```
##                     WithDTP   WithPolio     WithMMR    WithHepB PctUpToDate
## WithDTP          1.00000000  0.98164459  0.97685462  0.89072650  0.95889735
## WithPolio        0.98164459  1.00000000  0.96676214  0.90551032  0.94027555
## WithMMR          0.97685462  0.96676214  1.00000000  0.88978894  0.96715492
## WithHepB         0.89072650  0.90551032  0.88978894  1.00000000  0.84332389
## PctUpToDate      0.95889735  0.94027555  0.96715492  0.84332389  1.00000000
## PctBeliefExempt -0.79928929 -0.81945744 -0.78455517 -0.91863033 -0.72075073
## PctMedicalExempt -0.01802398 -0.01727579 -0.02255113 -0.05551026 -0.02015874
## PctChildPoverty   0.22392698  0.22337341  0.22040402  0.23713731  0.22398612
## PctFamilyPoverty  0.28168386  0.28008137  0.27673923  0.29782603  0.27457760
## EnrolledLog       0.29125085  0.29424889  0.29799141  0.30826147  0.28370107
## TotalSchoolsLog   0.20967375  0.21140500  0.21486128  0.21629904  0.20434909
##                  PctBeliefExempt PctMedicalExempt PctChildPoverty
## WithDTP               -0.79928929      -0.01802398      0.22392698
## WithPolio             -0.81945744      -0.01727579      0.22337341
## WithMMR               -0.78455517      -0.02255113      0.22040402
## WithHepB              -0.91863033      -0.05551026      0.23713731
## PctUpToDate           -0.72075073      -0.02015874      0.22398612
## PctBeliefExempt        1.00000000      -0.02148321     -0.20130543
## PctMedicalExempt      -0.02148321       1.00000000     -0.05692281
## PctChildPoverty       -0.20130543      -0.05692281      1.00000000
## PctFamilyPoverty      -0.26554357      -0.02163590      0.86777683
## EnrolledLog           -0.27277972      -0.05431268     -0.05301706
## TotalSchoolsLog       -0.19931812      -0.04411751     -0.08728632
##                  PctFamilyPoverty EnrolledLog TotalSchoolsLog
## WithDTP               0.281683859  0.29125085     0.209673754
## WithPolio             0.280081365  0.29424889     0.211405001
## WithMMR               0.276739230  0.29799141     0.214861275
## WithHepB              0.297826029  0.30826147     0.216299038
```

```
## PctUpToDate           0.274577602  0.28370107     0.204349092
## PctBeliefExempt      -0.265543568 -0.27277972    -0.199318123
## PctMedicalExempt     -0.021635900 -0.05431268    -0.044117509
## PctChildPoverty       0.867776828 -0.05301706    -0.087286318
## PctFamilyPoverty      1.000000000  0.05546011    -0.005408707
## EnrolledLog           0.055460112  1.00000000     0.916319391
## TotalSchoolsLog      -0.005408707  0.91631939     1.000000000
```

create a new column in the schools data frame

```
schools$PctUpToDate <- schools$UP_TO_DATE / schools$ENROLLMENT
cor(schools[sapply(schools, is.numeric)], use = "complete.obs")
```

```
##             SCHOOL CODE  ENROLLMENT  UP_TO_DATE CONDITIONAL         PME
## SCHOOL CODE  1.00000000 -0.18926660 -0.17197653 -0.04248789 -0.01096738
## ENROLLMENT  -0.18926660  1.00000000  0.97260752  0.27647566  0.09890612
## UP_TO_DATE  -0.17197653  0.97260752  1.00000000  0.08125671  0.07770968
## CONDITIONAL -0.04248789  0.27647566  0.08125671  1.00000000  0.02498296
## PME         -0.01096738  0.09890612  0.07770968  0.02498296  1.00000000
## PBE_BETA    -0.14429294  0.16745663  0.04232908  0.03241211  0.04236220
## DTP         -0.17225359  0.98144116  0.99699843  0.13787225  0.07981121
## POLIO       -0.17222073  0.98265878  0.99637880  0.14875381  0.08044297
## MMR         -0.17331383  0.98195069  0.99578978  0.14854329  0.07813969
## HEPB        -0.17161476  0.98869798  0.98927448  0.21475018  0.07928582
## VARICELLA   -0.17170350  0.98910828  0.98788688  0.22371486  0.07922669
## PctUpToDate  0.04134012  0.18434404  0.35854053 -0.56629368 -0.05035894
##                PBE_BETA         DTP       POLIO         MMR        HEPB
## SCHOOL CODE -0.14429294 -0.17225359 -0.17222073 -0.17331383 -0.17161476
## ENROLLMENT   0.16745663  0.98144116  0.98265878  0.98195069  0.98869798
## UP_TO_DATE   0.04232908  0.99699843  0.99637880  0.99578978  0.98927448
## CONDITIONAL  0.03241211  0.13787225  0.14875381  0.14854329  0.21475018
## PME          0.04236220  0.07981121  0.08044297  0.07813969  0.07928582
## PBE_BETA     1.00000000  0.04745810  0.04476004  0.04415761  0.04427995
## DTP          0.04745810  1.00000000  0.99925490  0.99815403  0.99478415
## POLIO        0.04476004  0.99925490  1.00000000  0.99835413  0.99584641
## MMR          0.04415761  0.99815403  0.99835413  1.00000000  0.99483096
## HEPB         0.04427995  0.99478415  0.99584641  0.99483096  1.00000000
## VARICELLA    0.04413276  0.99400203  0.99521351  0.99442589  0.99917553
## PctUpToDate -0.37652214  0.32185930  0.31685832  0.31926245  0.27432427
##               VARICELLA PctUpToDate
## SCHOOL CODE -0.17170350  0.04134012
## ENROLLMENT   0.98910828  0.18434404
## UP_TO_DATE   0.98788688  0.35854053
## CONDITIONAL  0.22371486 -0.56629368
## PME          0.07922669 -0.05035894
## PBE_BETA     0.04413276 -0.37652214
## DTP          0.99400203  0.32185930
## POLIO        0.99521351  0.31685832
## MMR          0.99442589  0.31926245
## HEPB         0.99917553  0.27432427
## VARICELLA    1.00000000  0.26845363
## PctUpToDate  0.26845363  1.00000000
```

Descriptive Reporting

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

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

```
usVaccines_ts <- ts(usVaccines, start = 1980, frequency = 1)
```

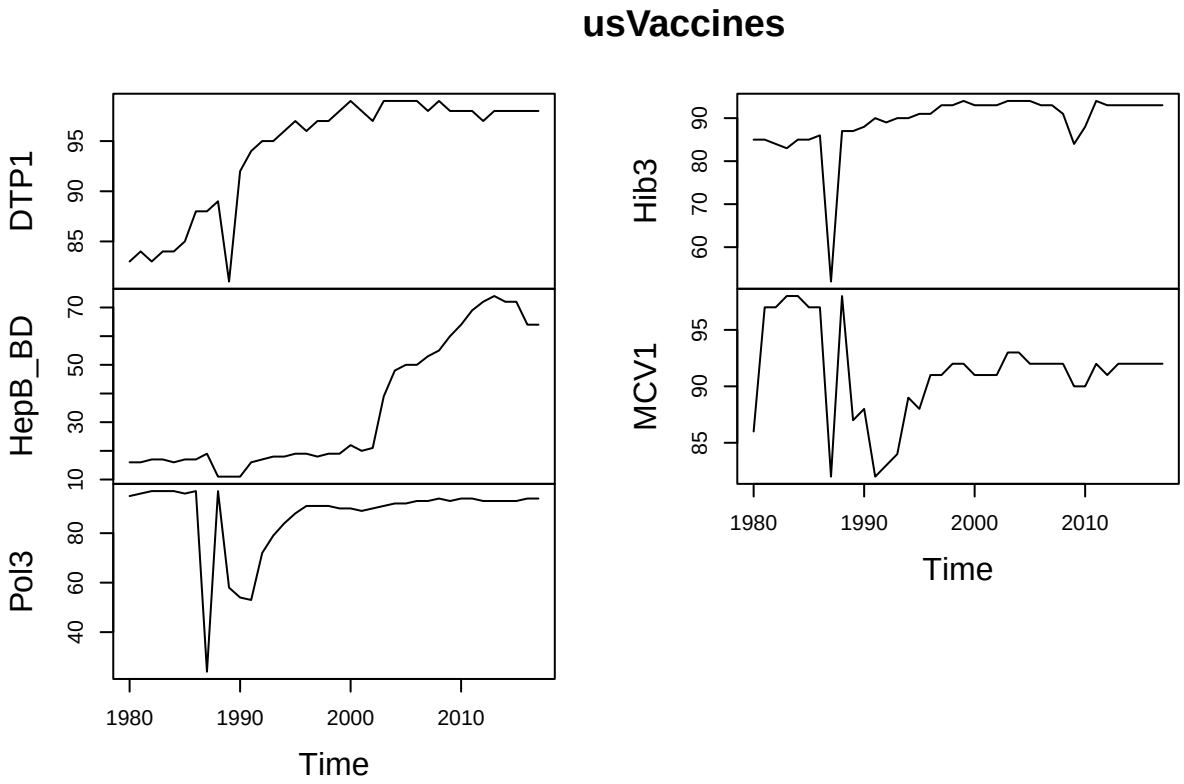
```
str(usVaccines)
```

```
##  Time-Series [1:38, 1:5] from 1980 to 2017: 83 84 83 84 84 85 88 88 89 81 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:5] "DTP1" "HepB_BD" "Pol3" "Hib3" ...
```

```
summary(usVaccines)
```

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

```
plot(usVaccines)
```



in the usVaccines – Time series data from the World Health Organization reporting vaccination rates in the U.S. for five common vaccines from 1980 to 2017."DTP1" "HepB_BD" "Pol3" "Hib3" "MCV1". we can see a huge drop during 1987 to 1991.

A pure polysaccharide vaccine was licensed for use in the United States in 1985 and remained in use until 1988. The first Hib conjugate vaccine was licensed in 1987. It is somewhat confusing to have data from 1980 to 1985 (and 1987) given these licensing dates.

https://www.cdc.gov/vaccines/pubs/pinkbook/hib.html

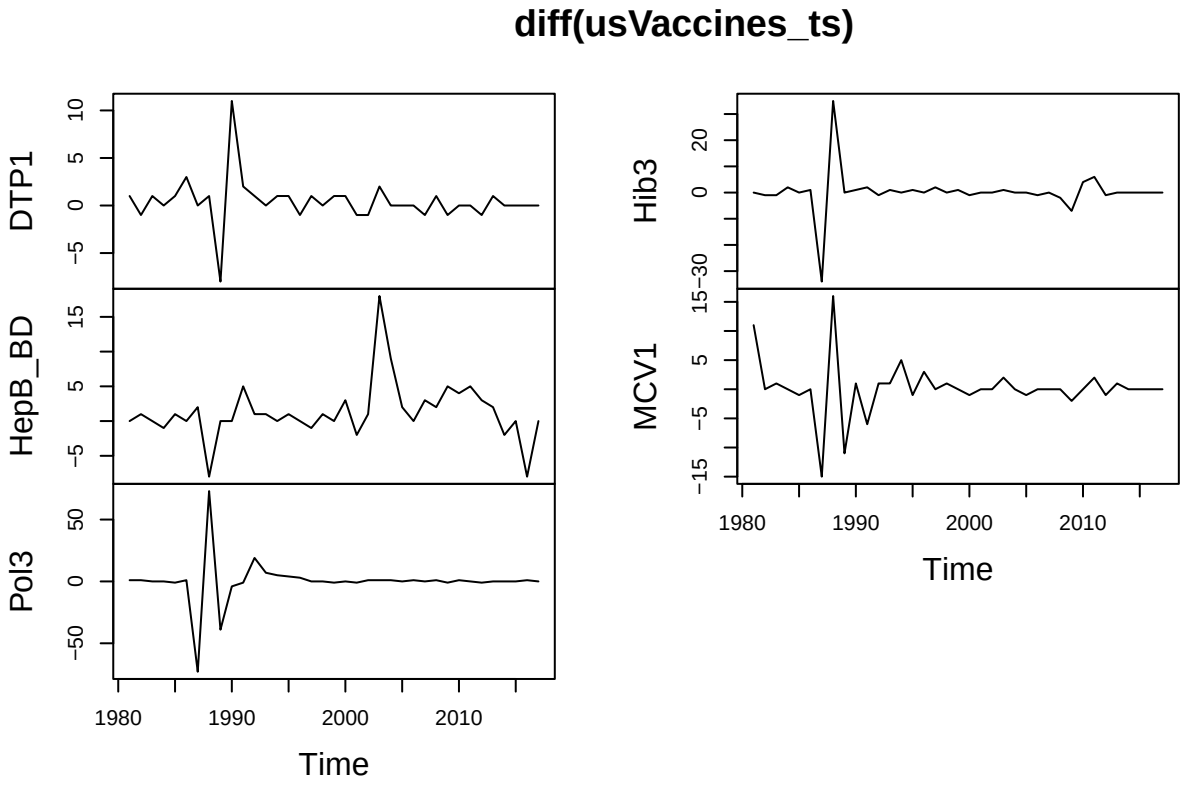
Regarding the data trends: DTP1 had a significant drop in 1989. HepB_BD had a decline from 1988 to 1990. Pol3 showed a huge decrease in 1987, a rebound in 1988, followed by a moderate decline from 1989 to 1991, and then a gradual increase. Hib3 had a significant drop in 1987. MCV1 dropped in 1987, recovered in 1988, dropped again from 1990 to 1993, and then slowly rose.

I have not found a comprehensive historical reference that explains these phenomena, an article from The Washington Post in 1991 partly attributes the reason to a cessation in data collection.

https://www.washingtonpost.com/archive/politics/1991/10/09/us-immunization-rates-uncertain/554e1354-781e-45ed-9df9-307e7a12575d/

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

```
plot(diff(usVaccines_ts))
```




```
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
adf.test(diff(usVaccines_ts[, "DTP1"]))
```

```
## Warning in adf.test(diff(usVaccines_ts[, "DTP1"])): p-value smaller than
## printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  diff(usVaccines_ts[, "DTP1"])
## Dickey-Fuller = -4.5333, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

```
adf.test(diff(usVaccines_ts[, "HepB_BD"]))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(usVaccines_ts[, "HepB_BD"])
## Dickey-Fuller = -1.958, Lag order = 3, p-value = 0.5896
## alternative hypothesis: stationary
```

```r
adf.test(diff(usVaccines_ts[, "Pol3"]))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(usVaccines_ts[, "Pol3"])
## Dickey-Fuller = -2.9361, Lag order = 3, p-value = 0.2082
## alternative hypothesis: stationary
```

```r
adf.test(diff(usVaccines_ts[, "Hib3"]))
```

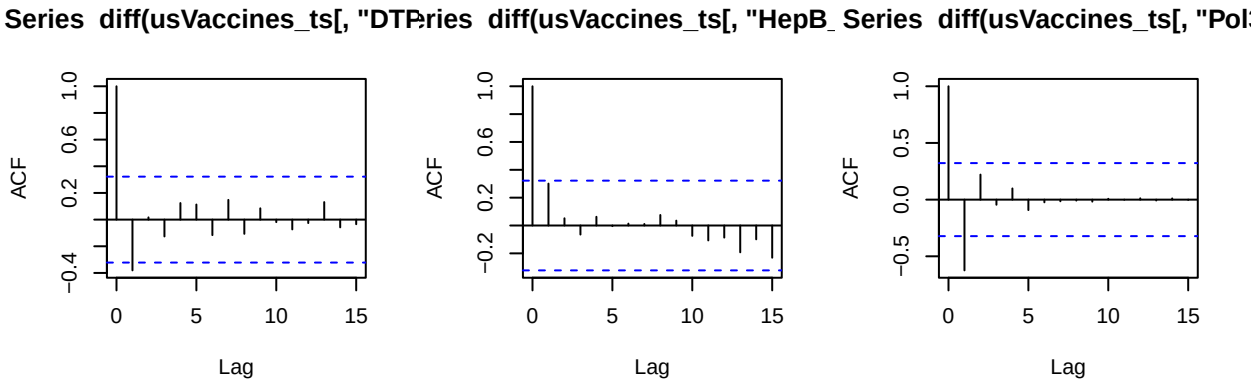
```
## Warning in adf.test(diff(usVaccines_ts[, "Hib3"])): p-value smaller than
## printed p-value
```

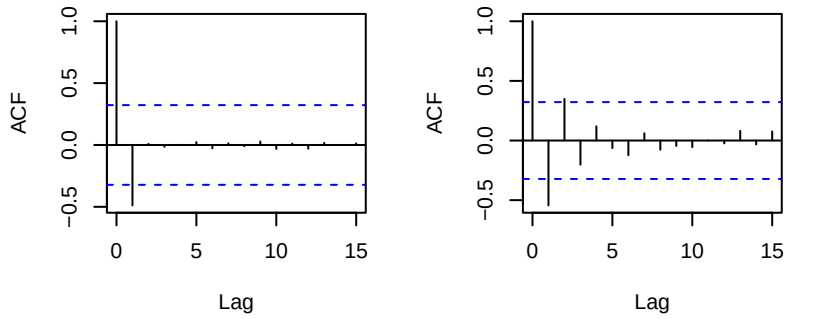
```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(usVaccines_ts[, "Hib3"])
## Dickey-Fuller = -4.3152, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

```r
adf.test(diff(usVaccines_ts[, "MCV1"]))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(usVaccines_ts[, "MCV1"])
## Dickey-Fuller = -3.3982, Lag order = 3, p-value = 0.07305
## alternative hypothesis: stationary
```

```r
par(mfrow=c(2,3))
acf(diff(usVaccines_ts[, "DTP1"]))
acf(diff(usVaccines_ts[, "HepB_BD"]))
acf(diff(usVaccines_ts[, "Pol3"]))
acf(diff(usVaccines_ts[, "Hib3"]))
acf(diff(usVaccines_ts[, "MCV1"]))
```

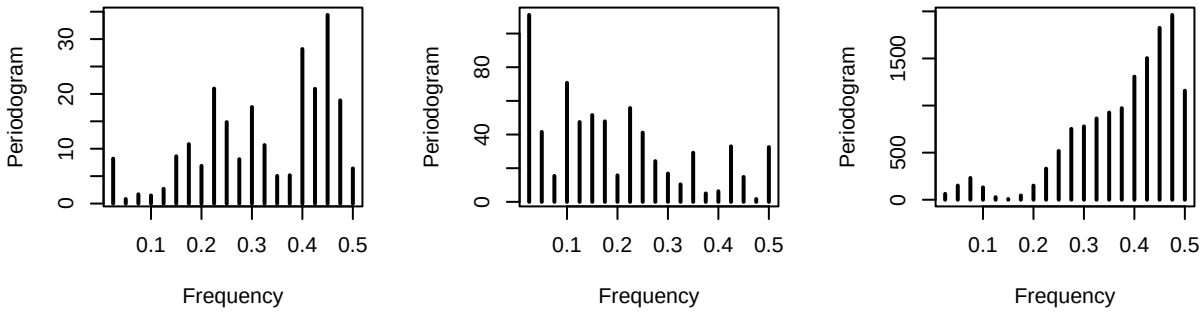


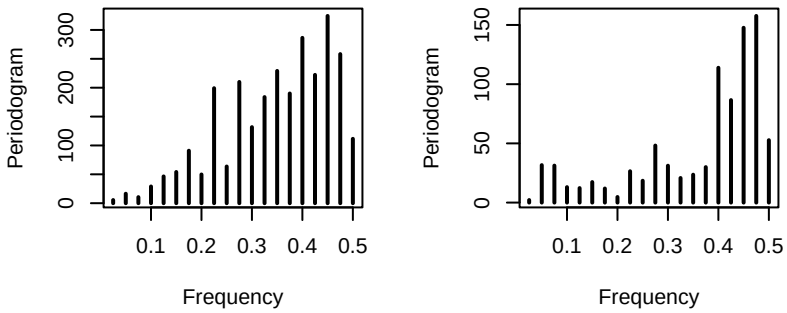



```
library(TSA)
```

```
##
## Attaching package: 'TSA'

## The following object is masked from 'package:readr':
##
##     spec

## The following objects are masked from 'package:stats':
##
##     acf, arima

## The following object is masked from 'package:utils':
##
##     tar
```

```
par(mfrow=c(2,3))
p1 <- periodogram(diff(usVaccines_ts[,1]))
p2 <- periodogram(diff(usVaccines_ts[,2]))
p3 <- periodogram(diff(usVaccines_ts[,3]))
p4 <- periodogram(diff(usVaccines_ts[,4]))
p5 <- periodogram(diff(usVaccines_ts[,5]))
```






```
1/p1$freq[which.max(p1$spec)]
```

```
## [1] 2.222222
```

```
1/p2$freq[which.max(p2$spec)]
```

```
## [1] 40
```

```
1/p3$freq[which.max(p3$spec)]
```

```
## [1] 2.105263
```

```
1/p4$freq[which.max(p4$spec)]
```

```
## [1] 2.222222
```

```
1/p5$freq[which.max(p5$spec)]
```

```
## [1] 2.105263
```

c. *What are the mean U.S. vaccination rates when including only recent years in the calculation of the mean?*

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. Examine your analyses for the previous questions or run another analysis to decide what a reasonable recent period is, i.e., a period during which the rates are relatively constant.

```r
library(changepoint)
```

```
## 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 2.2.4
##  See NEWS for details of changes.
```

```r
par(mfrow=c(2,3))
cpt.mean(usVaccines_ts[, "DTP1"])
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## 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
```

```r
plot(cpt.mean(usVaccines_ts[, "DTP1"]))

cpt.mean(usVaccines_ts[, "HepB_BD"])
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## 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
```

```
plot(cpt.mean(usVaccines_ts[, "HepB_BD"]))

cpt.mean(usVaccines_ts[, "Pol3"])
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## 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
```

```
plot(cpt.mean(usVaccines_ts[, "Pol3"]))

cpt.mean(usVaccines_ts[, "Hib3"])
```

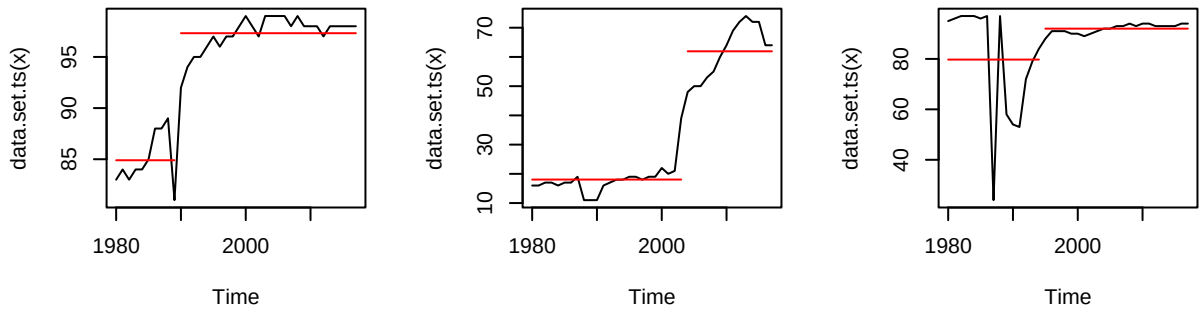
```
## Class 'cpt' : Changepoint Object
##        ~~   : S4 class containing 12 slots with names
##               cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## 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
```

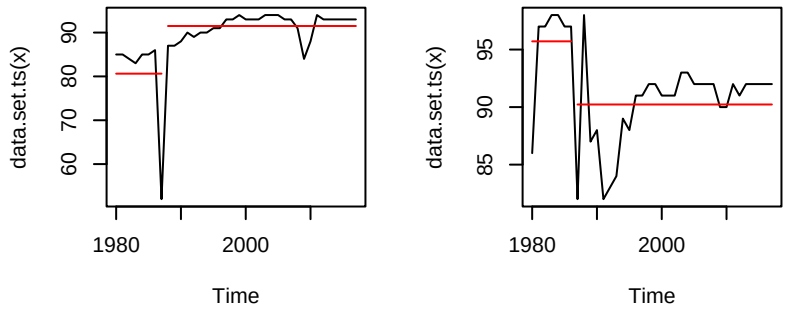
```
plot(cpt.mean(usVaccines_ts[, "Hib3"]))

cpt.mean(usVaccines_ts[, "MCV1"])
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## 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
```

```
plot(cpt.mean(usVaccines_ts[, "MCV1"]))
```






```
par(mfrow=c(2,3))
cpt.var(diff(usVaccines_ts[, "DTP1"]))
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
```

```
## 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
```

```
plot(cpt.var(diff(usVaccines_ts[, "DTP1"])))
cpt.var(diff(usVaccines_ts[, "HepB_BD"]),method = "PELT")
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.83275
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations :
```

```
plot(cpt.var(diff(usVaccines_ts[, "HepB_BD"]),method = "PELT"))
cpt.var(diff(usVaccines_ts[, "Pol3"]))
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## 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
```

```
plot(cpt.var(diff(usVaccines_ts[, "Pol3"])))
cpt.var(diff(usVaccines_ts[, "Hib3"]))
```

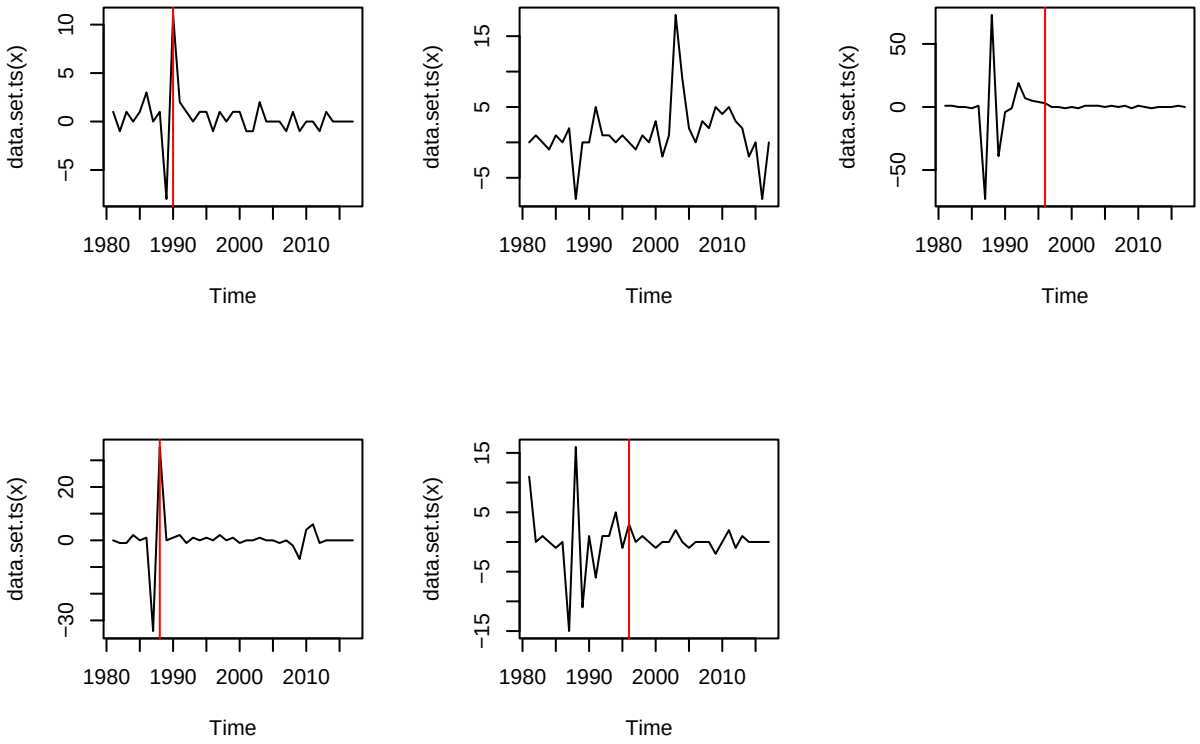
```
## Class 'cpt' : Changepoint Object
```

```
##        ~~   : S4 class containing 12 slots with names
##               cpttype date version data.set method test.stat pen.type pen.value minseglen cpts ncpts
##
## Created on  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## 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
```

```
plot(cpt.var(diff(usVaccines_ts[, "Hib3"])))
cpt.var(diff(usVaccines_ts[, "MCV1"]))
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## 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
```

```
plot(cpt.var(diff(usVaccines_ts[, "MCV1"])))
```




```
par(mfrow=c(2,3))
cpt.var(usVaccines_ts[, "DTP1"], method = "PELT")
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations : 10
```

```
plot(cpt.var(usVaccines_ts[, "DTP1"], method = "PELT"))

cpt.var(usVaccines_ts[, "HepB_BD"], method = "PELT")
```

```
## 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  : Thu Apr  6 10:58:50 2023
```

```
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations :
```

```
plot(cpt.var(usVaccines_ts[, "HepB_BD"], method = "PELT"))

cpt.var(usVaccines_ts[, "Pol3"])
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 13
```

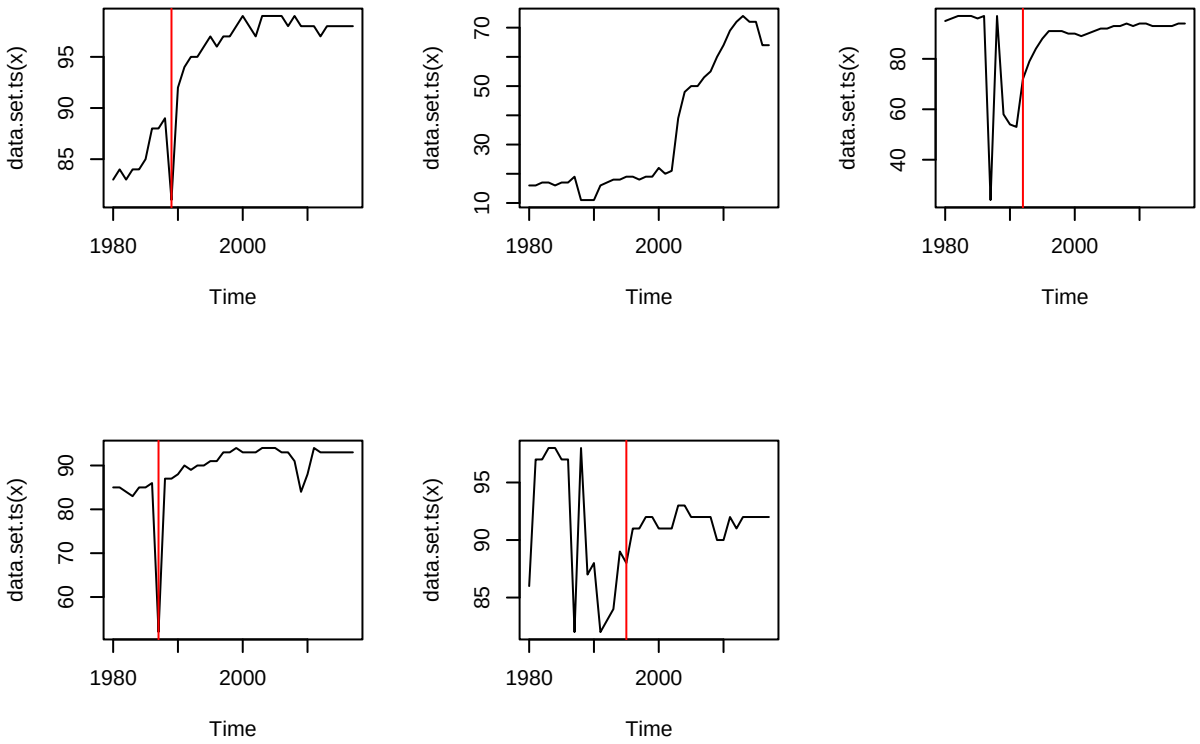
```
plot(cpt.var(usVaccines_ts[, "Pol3"]))

cpt.var(usVaccines_ts[, "Hib3"])
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 8
```

```r
plot(cpt.var(usVaccines_ts[, "Hib3"]))

cpt.var(usVaccines_ts[, "MCV1"])
```

```
## 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  : Thu Apr  6 10:58:50 2023
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 16
```

```r
plot(cpt.var(usVaccines_ts[, "MCV1"]))
```




```r
recent_vaccines <- usVaccines[25:38, ]

colMeans(recent_vaccines)
```

```
##     DTP1   HepB_BD     Pol3     Hib3     MCV1
## 98.21429 61.92857 93.21429 92.07143 91.71429
```

```r
mean(recent_vaccines)
```

```
## [1] 87.42857
```

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?***

```r
mean(districts$WithDTP)
```

```
## [1] 89.79571
```

```r
mean(districts$WithPolio)
```

```
## [1] 90.20429
```

```r
mean(districts$WithMMR)
```

```
## [1] 89.78714
```

```r
mean(districts$WithHepB)
```

```
## [1] 92.26286
```

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?***

```r
cor(districts %>% dplyr::select(c(WithDTP,WithPolio,WithMMR,WithHepB,PctUpToDate)))
```

```
##                WithDTP WithPolio   WithMMR  WithHepB PctUpToDate
## WithDTP     1.0000000 0.9816446 0.9768546 0.8907265   0.9588973
## WithPolio   0.9816446 1.0000000 0.9667621 0.9055103   0.9402755
## WithMMR     0.9768546 0.9667621 1.0000000 0.8897889   0.9671549
## WithHepB    0.8907265 0.9055103 0.8897889 1.0000000   0.8433239
## PctUpToDate 0.9588973 0.9402755 0.9671549 0.8433239   1.0000000
```

```r
cor.test(districts[,"WithDTP"], districts[,"PctUpToDate"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  districts[, "WithDTP"] and districts[, "PctUpToDate"]
## t = 89.281, df = 698, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9524745 0.9644680
## sample estimates:
##       cor
## 0.9588973
```

```
library("BayesFactor")
```

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

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

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

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

```
## ************
## Welcome to BayesFactor 0.9.12-4.7. If you have questions, please contact Richard Morey (richarddmorey
##
## Type BFManual() to open the manual.
## ************
```

```
bfCorTest <- function (x,y)
{
zx <- scale(x)
zy <- scale(y)
zData <- data.frame(x=zx,rhoNot0=zy)
bfOut <- generalTestBF(x ~ rhoNot0, data=zData)
mcmcOut <- posterior(bfOut,iterations=10000)
print(summary(mcmcOut[,"rhoNot0"]))
return(bfOut)
}

bfCorTest(districts[,"WithDTP"], districts[,"PctUpToDate"])
```

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

```
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##      0.9585294      0.0107152      0.0001072      0.0001072
##
## 2. Quantiles for each variable:
##
##   2.5%    25%    50%    75%  97.5%
## 0.9375 0.9512 0.9587 0.9658 0.9794
```

```
## Bayes factor analysis
## --------------
## [1] rhoNot0 : 2.300632e+379 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
cor.test(districts[,"WithHepB"], districts[,"PctUpToDate"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  districts[, "WithHepB"] and districts[, "PctUpToDate"]
## t = 41.459, df = 698, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8204961 0.8634664
## sample estimates:
##       cor
## 0.8433239
```

```
bfCorTest(districts[,"WithHepB"], districts[,"PctUpToDate"])
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##      0.8424194      0.0205381      0.0002054      0.0001997
##
## 2. Quantiles for each variable:
##
##   2.5%    25%    50%    75%  97.5%
## 0.8022 0.8284 0.8423 0.8563 0.8828
```

```
## Bayes factor analysis
## --------------
## [1] rhoNot0 : 1.368053e+186 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
cor.test(districts[,"WithPolio"], districts[,"PctUpToDate"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  districts[, "WithPolio"] and districts[, "PctUpToDate"]
## t = 72.975, df = 698, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9310452 0.9483034
## sample estimates:
##       cor
## 0.9402755
```

```
bfCorTest(districts[,"WithPolio"], districts[,"PctUpToDate"])
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##      0.9398156      0.0128829      0.0001288      0.0001288
##
## 2. Quantiles for each variable:
##
##   2.5%    25%    50%    75%  97.5%
## 0.9142 0.9312 0.9398 0.9485 0.9649

## Bayes factor analysis
## --------------
## [1] rhoNot0 : 1.848858e+324 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
cor.test(districts[,"WithMMR"], districts[,"PctUpToDate"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  districts[, "WithMMR"] and districts[, "PctUpToDate"]
## t = 100.52, df = 698, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9619974 0.9716226
## sample estimates:
##       cor
## 0.9671549
```

```
bfCorTest(districts[,"WithMMR"], districts[,"PctUpToDate"])
```

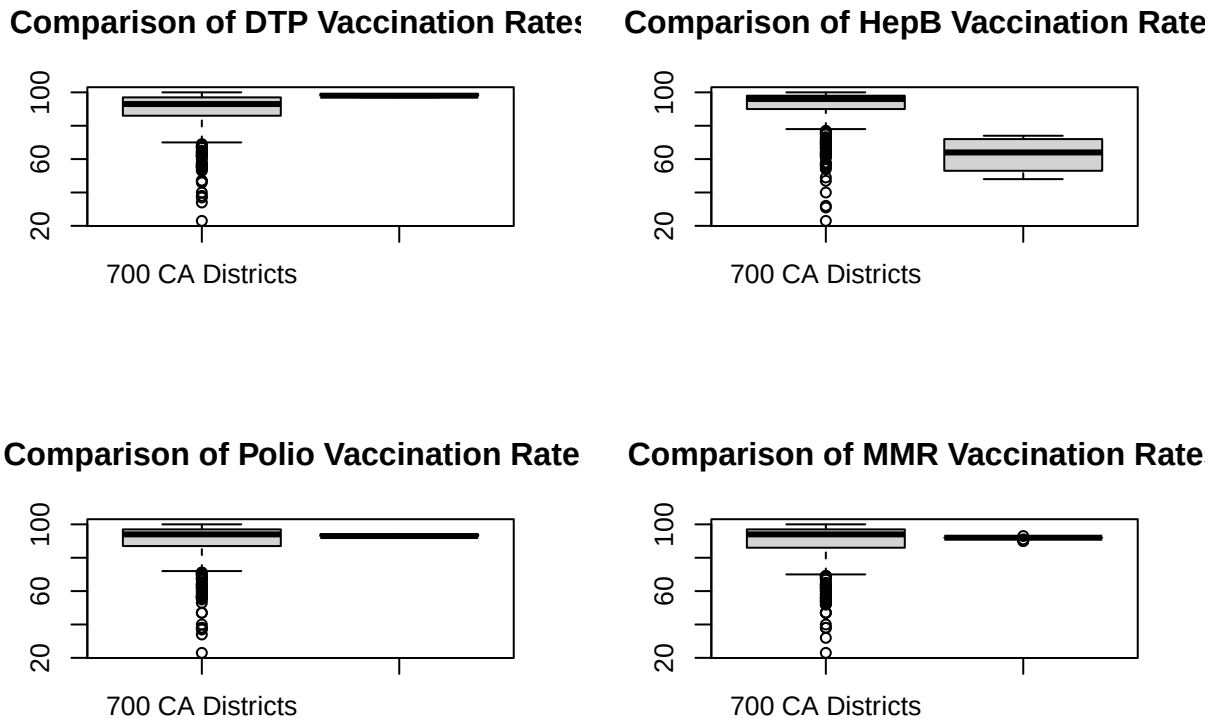
```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##      9.669e-01      9.708e-03      9.708e-05      9.910e-05
##
## 2. Quantiles for each variable:
##
##   2.5%    25%    50%    75%  97.5%
## 0.9482 0.9604 0.9668 0.9733 0.9859
```

```
## Bayes factor analysis
## --------------
## [1] rhoNot0 : 4.596466e+412 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

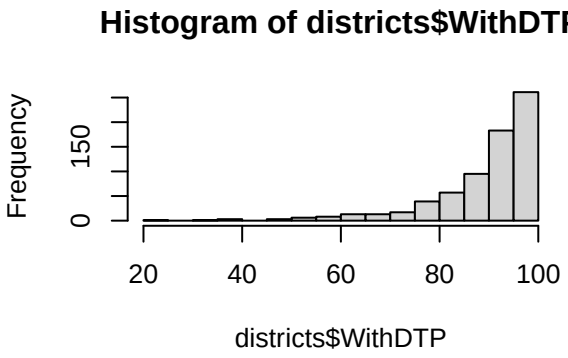
c. ***How do these Californian vaccination levels compare to U.S. vaccination levels (recent years only)? Note any patterns you notice and run any appropriate statistical tests.***

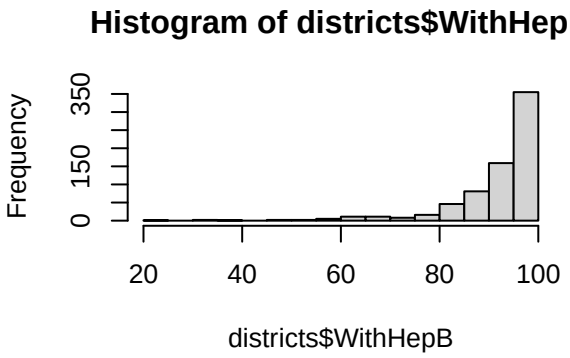
```
recent_vaccines <- usVaccines[25:38, ]


par(mfrow=c(2,2))
boxplot(districts$WithDTP, recent_vaccines[,"DTP1"],
```

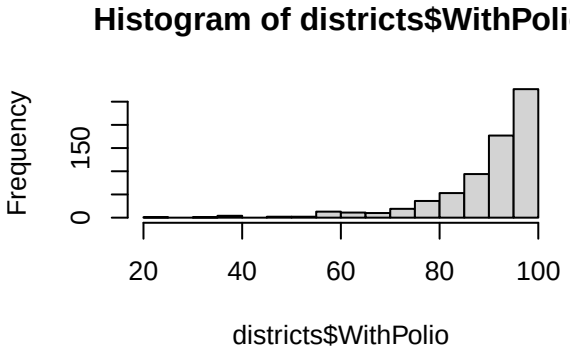
```
        names = c("700 CA Districts", "U.S. Recent Years"),
        main = "Comparison of DTP Vaccination Rates")

boxplot(districts$WithHepB, recent_vaccines[,"HepB_BD"],
        names = c("700 CA Districts", "U.S. Recent Years"),
        main = "Comparison of HepB Vaccination Rates")

boxplot(districts$WithPolio, recent_vaccines[,"Pol3"],
        names = c("700 CA Districts", "U.S. Recent Years"),
        main = "Comparison of Polio Vaccination Rates")

boxplot(districts$WithMMR, recent_vaccines[,"MCV1"],
        names = c("700 CA Districts", "U.S. Recent Years"),
        main = "Comparison of MMR Vaccination Rates")
```

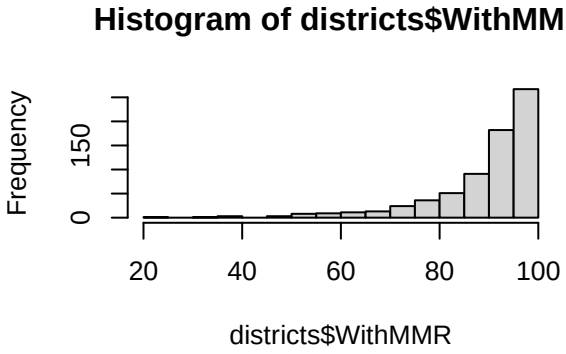



```
par(mfrow=c(2,2))
hist(districts$WithDTP)
hist(districts$WithHepB)
hist(districts$WithPolio)
hist(districts$WithMMR)
```










```
t.test(districts$WithDTP, recent_vaccines[,"DTP1"])
```

```
##
##  Welch Two Sample t-test
##
## data:  districts$WithDTP and recent_vaccines[, "DTP1"]
## t = -18.926, df = 448.14, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.292766 -7.544377
## sample estimates:
## mean of x mean of y
##  89.79571  98.21429
```

```
t.test(districts$WithPolio, recent_vaccines[,"Pol3"])
```

```
##
##  Welch Two Sample t-test
##
## data:  districts$WithPolio and recent_vaccines[, "Pol3"]
## t = -6.6037, df = 316.03, p-value = 1.698e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.906799 -2.113201
## sample estimates:
## mean of x mean of y
##  90.20429  93.21429
```

```r
t.test(districts$WithMMR, recent_vaccines[,"MCV1"])
```

```
##
##  Welch Two Sample t-test
##
## data:  districts$WithMMR and recent_vaccines[, "MCV1"]
## t = -3.9986, df = 234.17, p-value = 8.545e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.876667 -0.977619
## sample estimates:
## mean of x mean of y
##  89.78714  91.71429
```

```r
t.test(districts$WithHepB, recent_vaccines[,"HepB_BD"])
```

```
##
##  Welch Two Sample t-test
##
## data:  districts$WithHepB and recent_vaccines[, "HepB_BD"]
## t = 12.076, df = 13.597, p-value = 1.191e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  24.93162 35.73695
## sample estimates:
## mean of x mean of y
##  92.26286  61.92857
```

```r
library(BEST)
```

```
## Loading required package: HDInterval
```

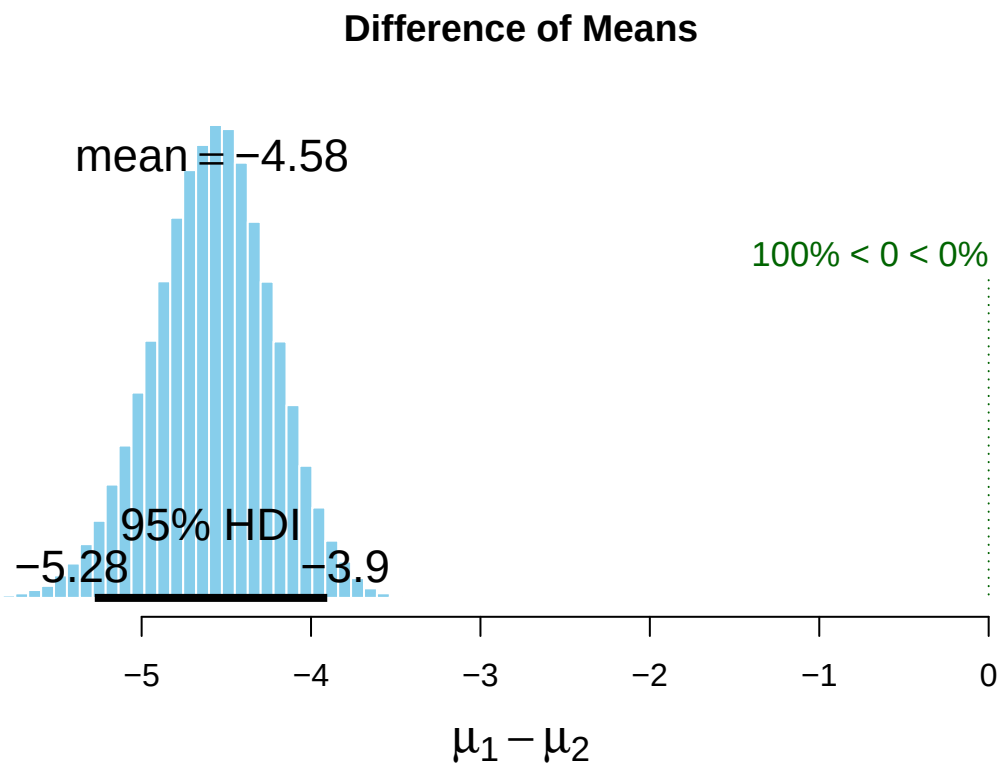
```r
vaccineDTPBest <- BESTmcmc(districts$WithDTP, recent_vaccines[,1],rnd.seed = 1234)
```

```
## Waiting for parallel processing to complete...

## done.
```

```r
vaccineDTPBest
```

```
## MCMC fit results for BEST analysis:
## 100002 simulations saved.
##           mean     sd  median    HDIlo  HDIup Rhat n.eff
## mu1    93.4984 0.3138 93.5037 92.87884 94.112    1 23972
## mu2    98.0820 0.1298 98.0555 97.85833 98.373    1 28227
## nu      1.9741 0.2178  1.9604  1.57126  2.417    1 15618
## sigma1  5.2869 0.3155  5.2806  4.68156  5.917    1 16857
## sigma2  0.3482 0.1754  0.3314  0.01521  0.664    1 18292
##
## 'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.
## 'Rhat' is the potential scale reduction factor (at convergence, Rhat=1).
## 'n.eff' is a crude measure of effective sample size.
```

```
plot(vaccineDTPBest)
```



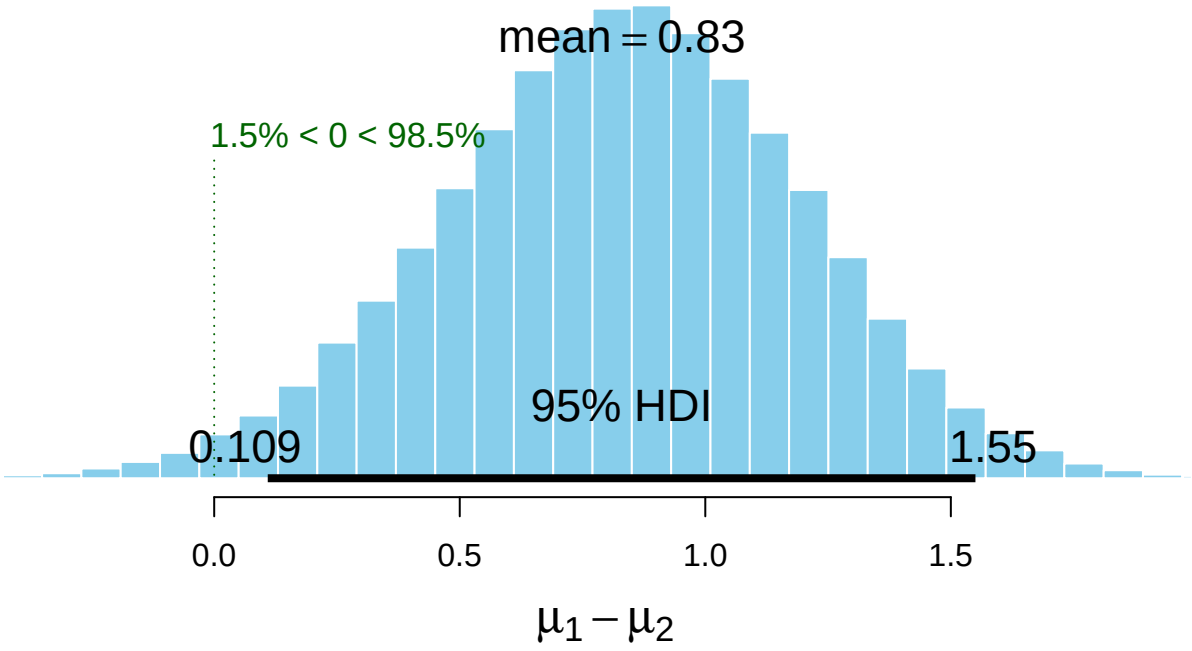

```
vaccinePolioBest <- BESTmcmc(districts$WithPolio, recent_vaccines[,"Pol3"],rnd.seed = 1234)
```

```
## Waiting for parallel processing to complete...done.
```

```
vaccinePolioBest
```

```
## MCMC fit results for BEST analysis:
## 100002 simulations saved.
##           mean     sd  median   HDIlo  HDIup Rhat n.eff
## mu1    94.0285 0.2940 94.0330 93.4571 94.605    1 26363
## mu2    93.1981 0.2171 93.1834 92.7911 93.652    1 52737
## nu      1.8981 0.1925  1.8851  1.5381  2.285    1 19337
## sigma1  5.0019 0.2958  4.9941  4.4307  5.584    1 18674
## sigma2  0.6067 0.1960  0.5801  0.2676  1.006    1 36584
##
## 'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.
## 'Rhat' is the potential scale reduction factor (at convergence, Rhat=1).
## 'n.eff' is a crude measure of effective sample size.
```

```
plot(vaccinePolioBest)
```



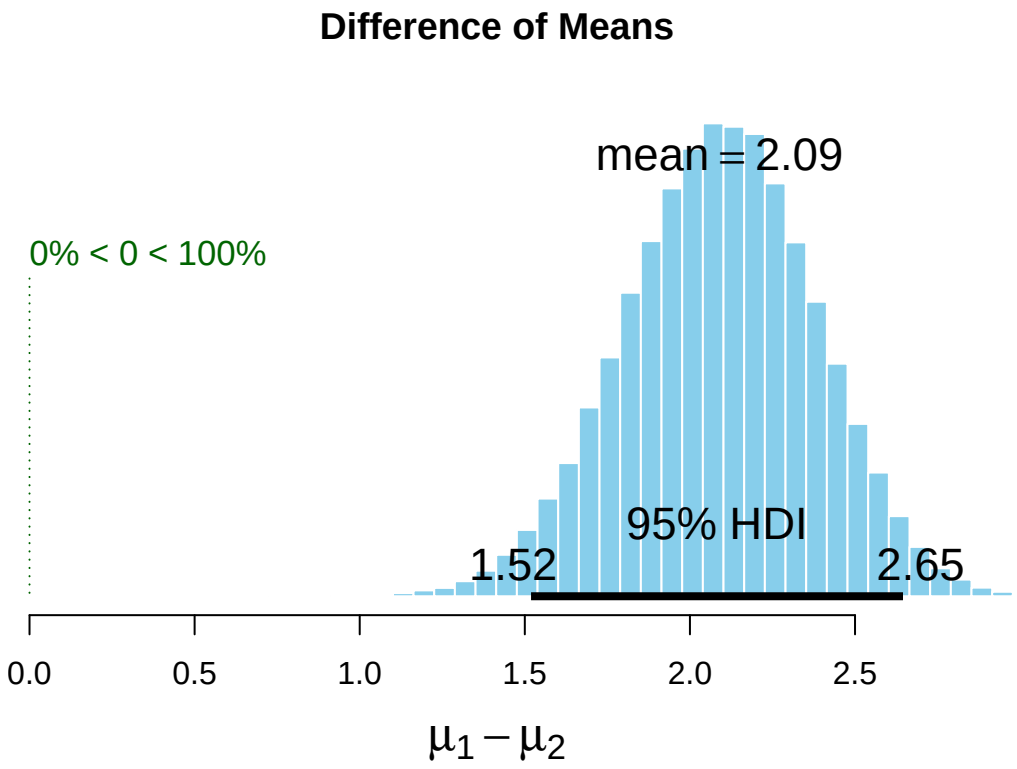

```
vaccineMMRBest <- BESTmcmc(districts$WithMMR, recent_vaccines[,"MCV1"],rnd.seed = 1234)
```

```
## Waiting for parallel processing to complete...done.
```

```
vaccineMMRBest
```

```
## MCMC fit results for BEST analysis:
## 100002 simulations saved.
##            mean      sd   median    HDIlo   HDIup  Rhat n.eff
## mu1    94.08957 0.28936 94.09489 93.52246 94.6498 1.000 29298
## mu2    91.99967 0.01072 91.99992 91.98310 92.0162 1.022  5077
## nu      1.56220 0.14225  1.55237  1.29835  1.8491 1.001 15315
## sigma1  4.86045 0.29192  4.84990  4.29473  5.4339 1.000 20461
## sigma2  0.02557 0.02791  0.01672  0.01123  0.0691 1.020  1579
##
## 'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.
## 'Rhat' is the potential scale reduction factor (at convergence, Rhat=1).
## 'n.eff' is a crude measure of effective sample size.
```

```
plot(vaccineMMRBest)
```



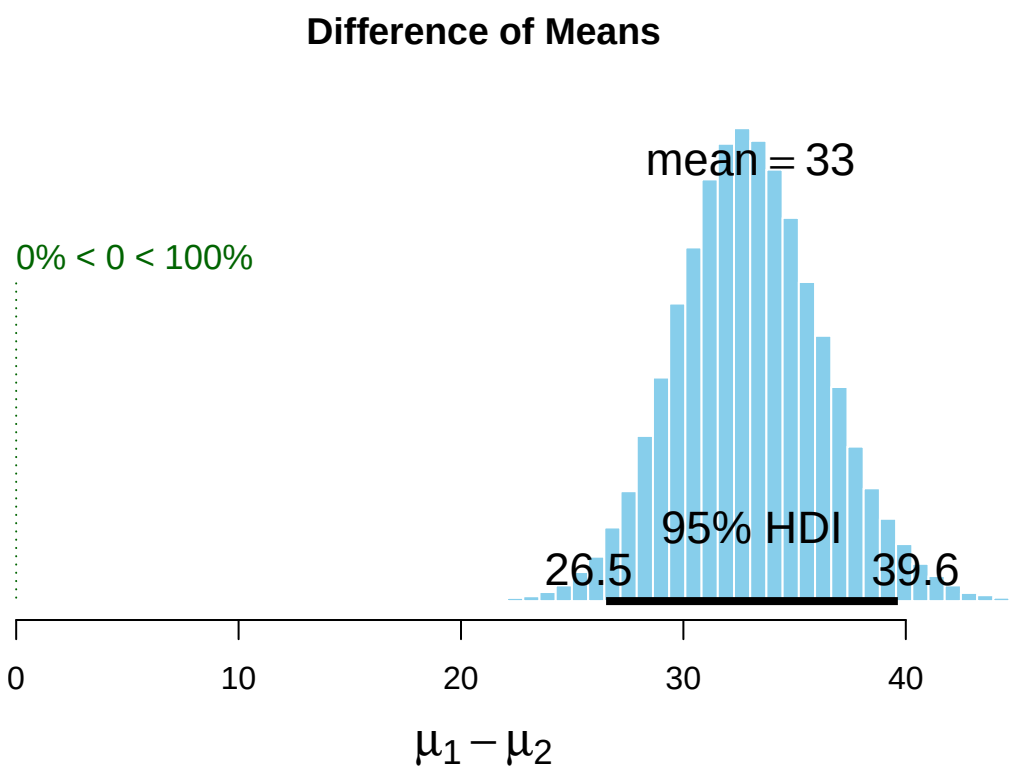

```
vaccineHepBBest <- BESTmcmc(districts$WithHepB, recent_vaccines[,2],rnd.seed = 1234)
```

```
## Waiting for parallel processing to complete...done.
```

```
vaccineHepBBest
```

```
## MCMC fit results for BEST analysis:
## 100002 simulations saved.
##          mean     sd median  HDIlo  HDIup Rhat n.eff
## mu1    95.846 0.2251 95.849 95.411 96.291    1 27265
## mu2    62.823 3.3241 62.914 56.065 69.172    1 56868
## nu      1.677 0.1559  1.668  1.385  1.992    1 21019
## sigma1  3.751 0.2275  3.747  3.307  4.198    1 21208
## sigma2  8.791 2.6556  8.404  4.323 14.158    1 39380
##
## 'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.
## 'Rhat' is the potential scale reduction factor (at convergence, Rhat=1).
## 'n.eff' is a crude measure of effective sample size.
```

```
plot(vaccineHepBBest)
```




```
kruskalTestPolio <- kruskal.test(list(districts$WithPolio, recent_vaccines[,"Pol3"]))
print(kruskalTestPolio)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(districts$WithPolio, recent_vaccines[, "Pol3"])
## Kruskal-Wallis chi-squared = 0.19229, df = 1, p-value = 0.661
```

```
kruskalTestDTP <- kruskal.test(list(districts$WithDTP, recent_vaccines[,"DTP1"]))
print(kruskalTestDTP)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(districts$WithDTP, recent_vaccines[, "DTP1"])
## Kruskal-Wallis chi-squared = 19.053, df = 1, p-value = 1.271e-05
```

```
kruskalTestMMR <- kruskal.test(list(districts$WithMMR, recent_vaccines[,"MCV1"]))
print(kruskalTestMMR)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(districts$WithMMR, recent_vaccines[, "MCV1"])
## Kruskal-Wallis chi-squared = 1.3126, df = 1, p-value = 0.2519
```

```
kruskalTestHepB <- kruskal.test(list(districts$WithHepB, recent_vaccines[,"HepB_BD"]))
print(kruskalTestHepB)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(districts$WithHepB, recent_vaccines[, "HepB_BD"])
## Kruskal-Wallis chi-squared = 36.577, df = 1, p-value = 1.467e-09
```

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

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

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

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

for the all schools dataset, there are 7,381 obs. of 18 variables. There are 5732 public schools, and 1649 private schools. For public schools, 148 not reported, and for private schools, 252 not reported. One priviate school has reported as N but actually has values in the vaccines column. so, therefore, 399 schools have NAs in 10 columns related to vaccines.

```
schools$REPORTED[schools$`SCHOOL CODE` == 6143788] <- "Y"
```

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

```
##
##     PRIVATE PUBLIC
##   N     251    148
##   Y    1398   5584
```

```
PrivateProp <- 1398 / (251 + 1398)
PublicProp <- 5584 / (148 + 5584)
PrivateProp
```

```
## [1] 0.8477865
```

```
PublicProp
```

```
## [1] 0.97418
```

b. ***Was there any credible difference in the proportion reporting between public and private schools?***

```
chisq.test(contingency_table, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 400.07, df = 1, p-value < 2.2e-16
```

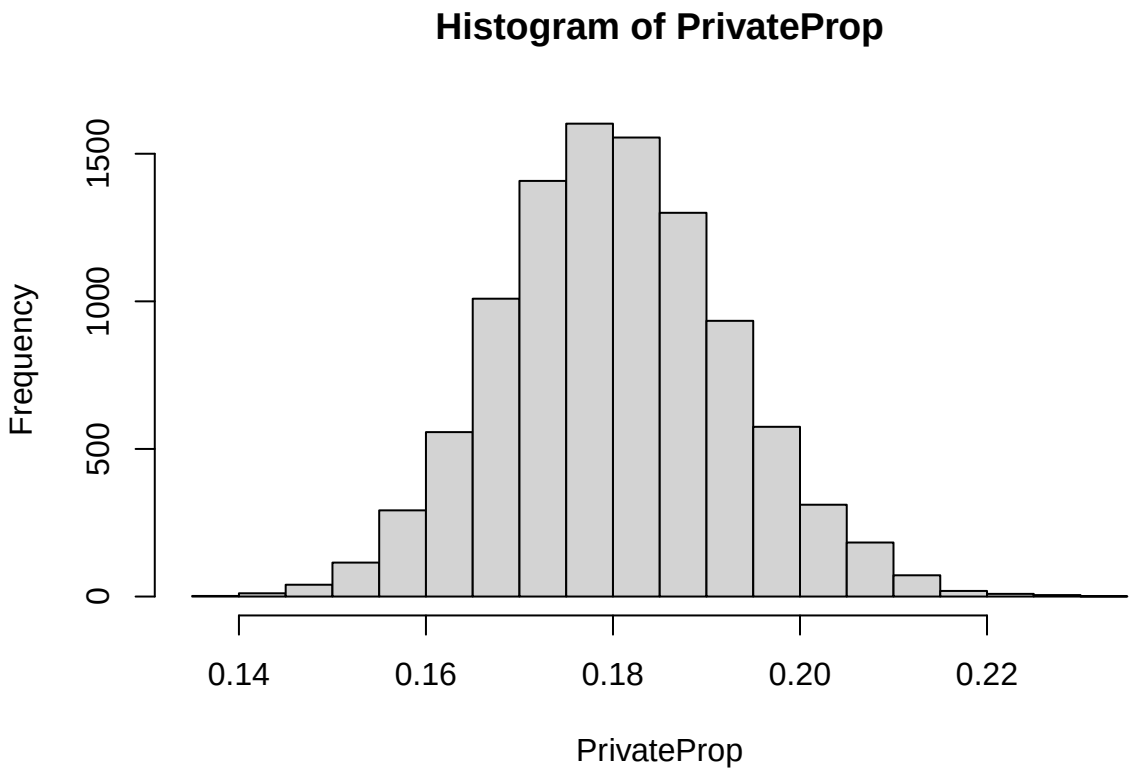
```
library("BayesFactor")
ctBFout <- contingencyTableBF(contingency_table, sampleType = "poisson", posterior=FALSE)
ctBFout
```

```
## Bayes factor analysis
## --------------
## [1] Non-indep. (a=1) : 3.656717e+68 ±0%
##
## Against denominator:
##   Null, independence, a = 1
## ---
## Bayes factor type: BFcontingencyTable, poisson
```

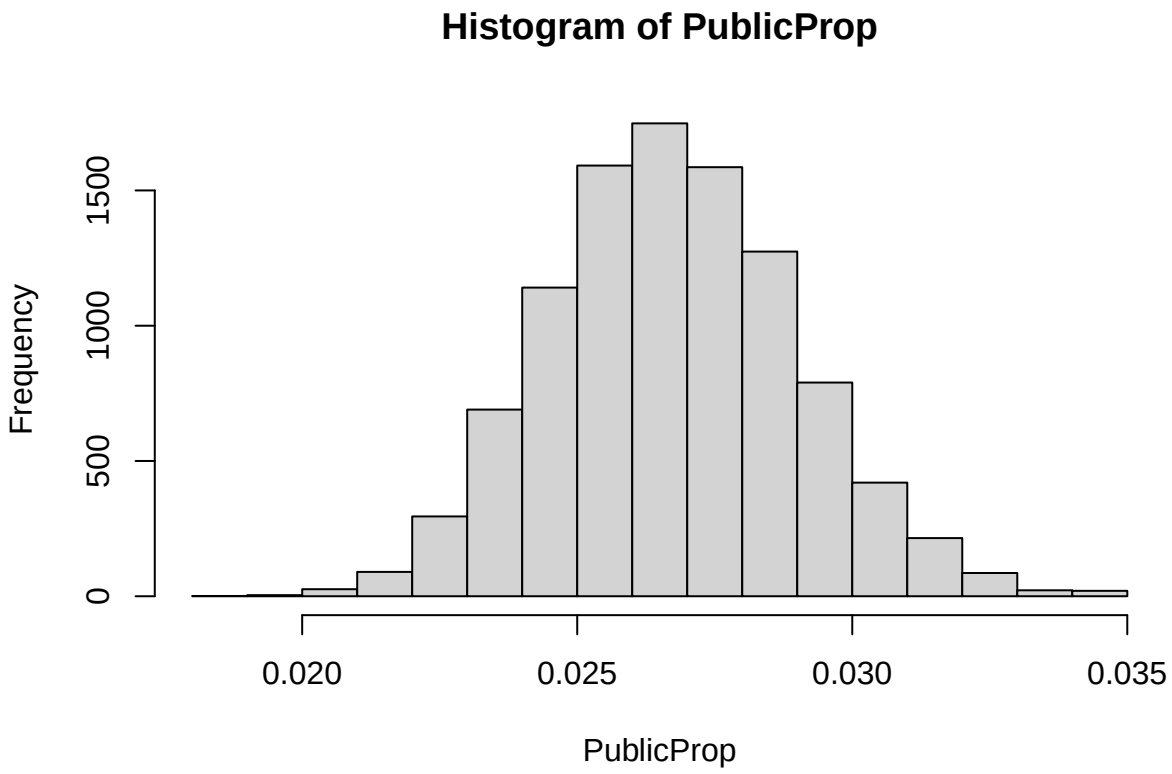
```
ctMCMCout <- contingencyTableBF(contingency_table, sampleType = "poisson", posterior=TRUE, iterations =
summary(ctMCMCout)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##               Mean    SD Naive SE Time-series SE
## lambda[1,1]  251.9 15.93   0.1593         0.1593
## lambda[2,1] 1398.3 37.14   0.3714         0.3819
## lambda[1,2]  149.2 12.34   0.1234         0.1234
## lambda[2,2] 5582.2 74.27   0.7427         0.7427
##
## 2. Quantiles for each variable:
##
##               2.5%    25%    50%    75%  97.5%
## lambda[1,1]  222.0  240.8  251.4  262.4  284.4
## lambda[2,1] 1327.7 1372.7 1397.8 1422.7 1472.2
## lambda[1,2]  126.2  140.7  148.8  157.3  174.0
## lambda[2,2] 5436.1 5532.1 5581.6 5632.3 5727.5
```

```
PrivateProp <- ctMCMCout[,"lambda[1,1]"] / ctMCMCout[,"lambda[2,1]"]
PublicProp <- ctMCMCout[,"lambda[1,2]"] / ctMCMCout[,"lambda[2,2]"]
diffProp <- PrivateProp - PublicProp

hist(PrivateProp)
```



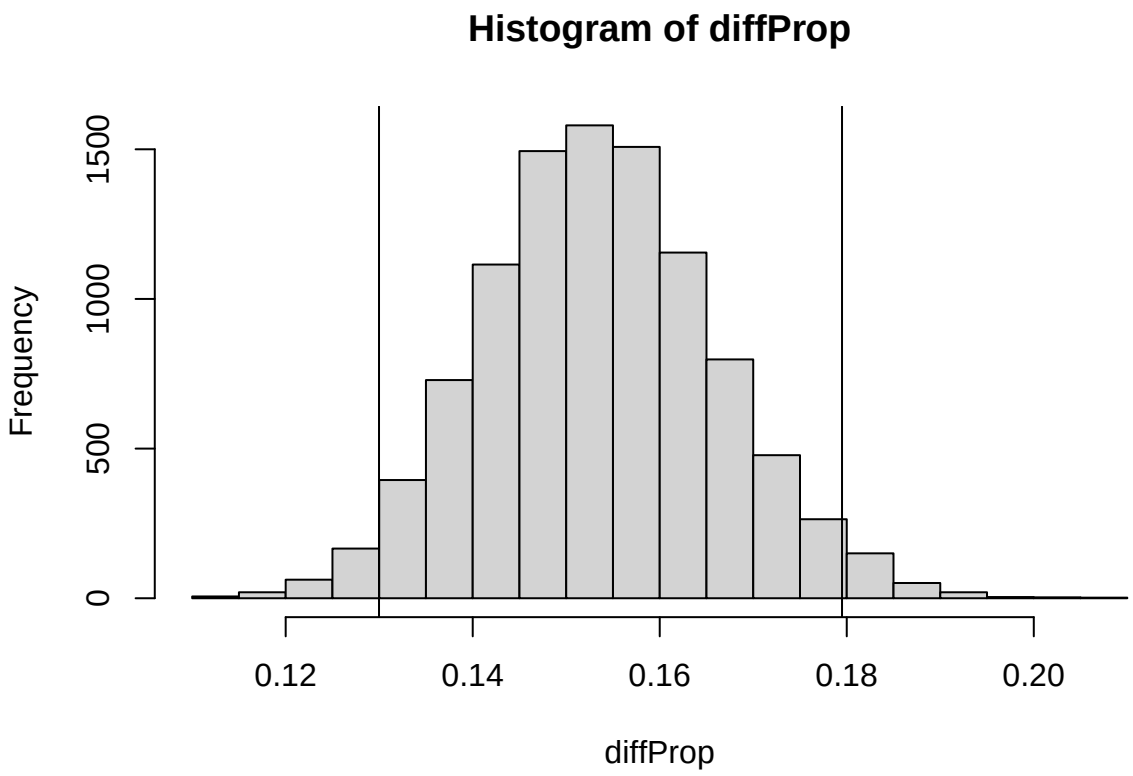

```
hist(PublicProp)
```




```
hist(diffProp)
mean(diffProp)
```

```
## [1] 0.1535379
```

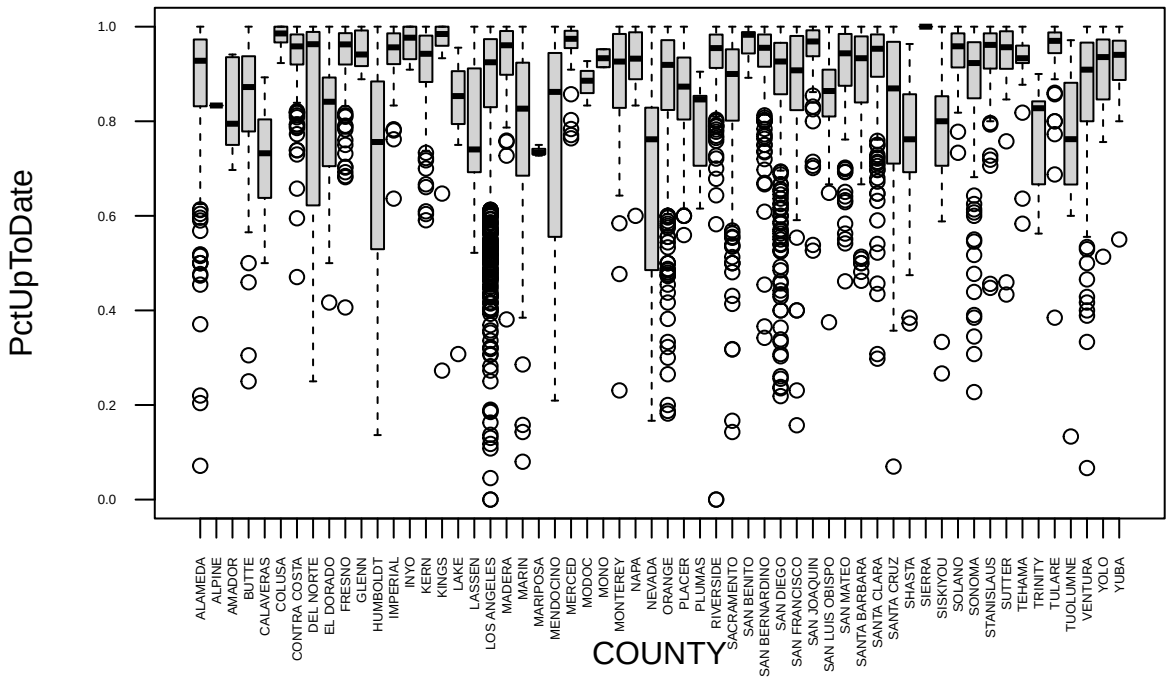
```
abline(v=quantile(diffProp,c(0.025)))
abline(v=quantile(diffProp,c(0.975)))
```



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

```
schools$PctUpToDate <- schools$UP_TO_DATE / schools$ENROLLMENT
```

```
boxplot(PctUpToDate ~ COUNTY, data = schools, las = 2, cex.axis = 0.4)
```



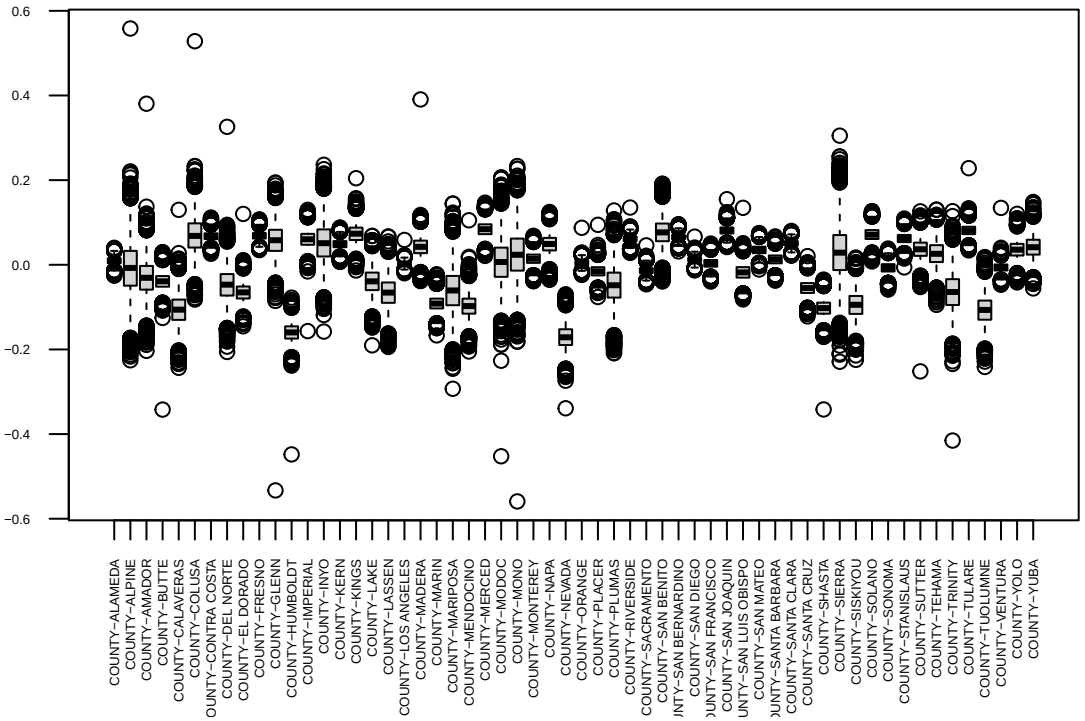
```
vaccineAnova <- aov(PctUpToDate ~ COUNTY, data = schools)
summary(vaccineAnova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## COUNTY        57  12.92  0.2267    13.1 <2e-16 ***
## Residuals   6924 119.81  0.0173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 399 observations deleted due to missingness
```

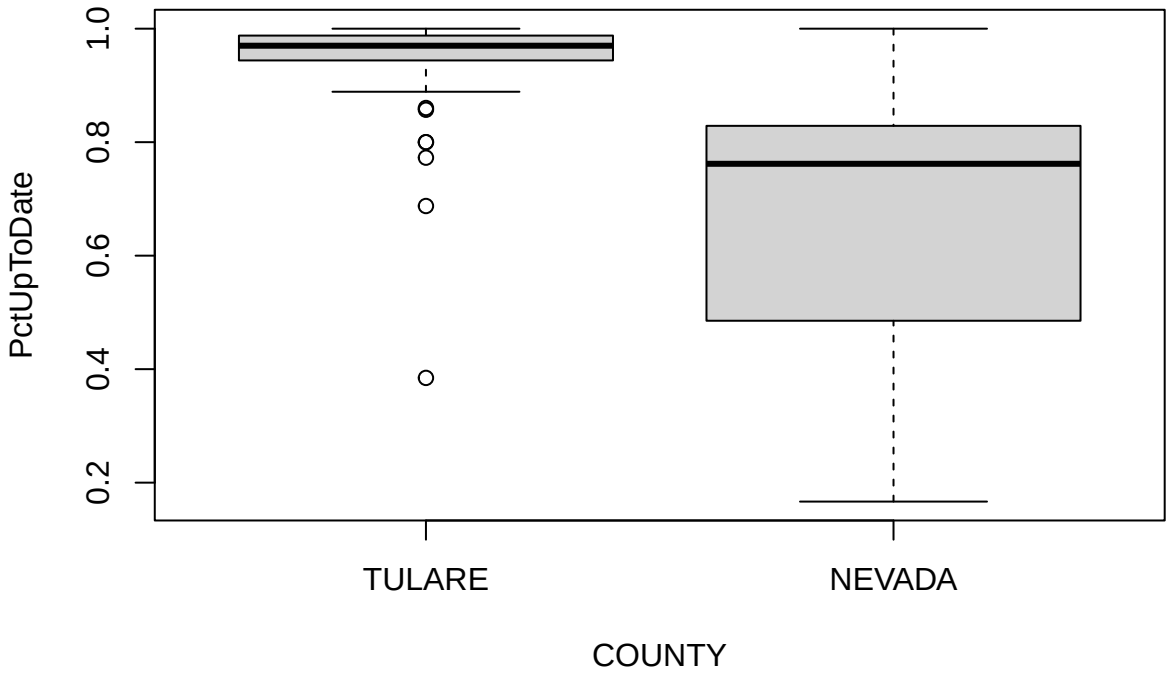
```
library("BayesFactor")
library(BEST)

clean_schools <- schools[!is.na(schools$PctUpToDate), ]
clean_schools$COUNTY <- as.factor(clean_schools$COUNTY)

vaccineOut <- anovaBF(PctUpToDate ~ COUNTY, data = clean_schools)
```

```
## Warning: data coerced from tibble to data frame
```

```
mcmcOut <- posterior(vaccineOut, iterations = 10000)

boxplot(as.matrix(mcmcOut[,2:59]), las = 2, cex.axis = 0.4)
```




```
filtered_data <- clean_schools[clean_schools$COUNTY %in% c("TULARE", "NEVADA"), ]
filtered_data$COUNTY <- factor(filtered_data$COUNTY, levels = c("TULARE", "NEVADA"))
boxplot(PctUpToDate ~ COUNTY, data = filtered_data)
```



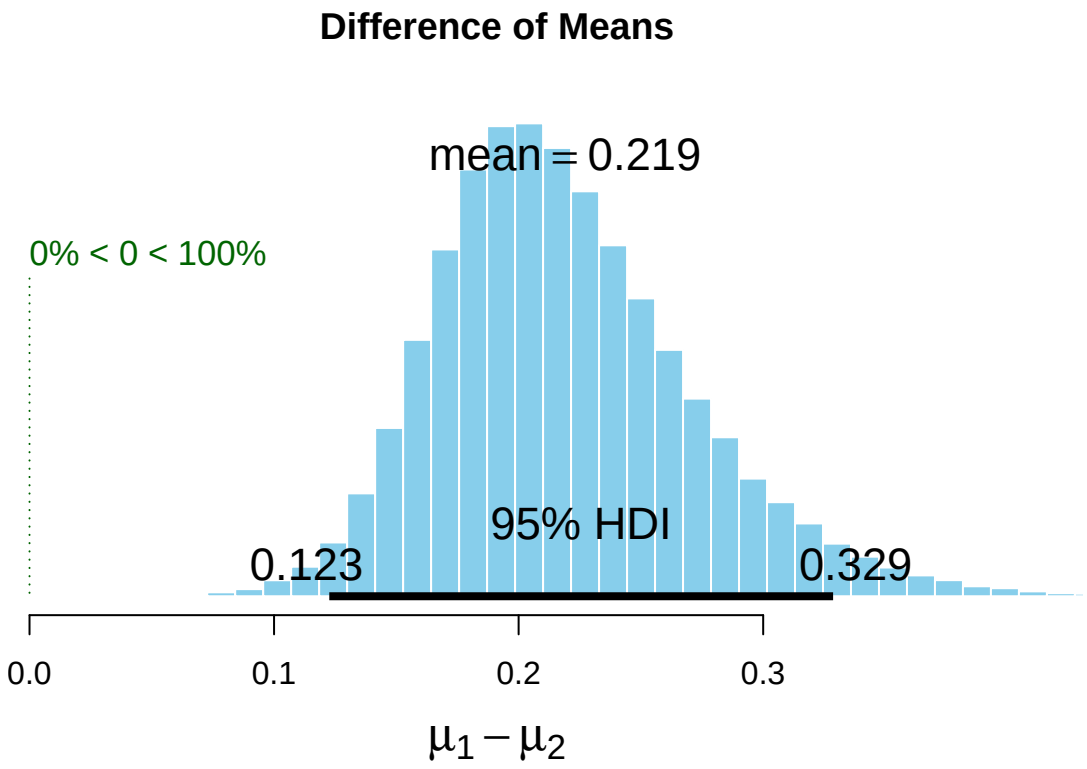

```
library(BEST)
CountyPctUpToDateBest <- BESTmcmc(clean_schools$PctUpToDate[clean_schools$COUNTY == "TULARE"],
                                  clean_schools$PctUpToDate[clean_schools$COUNTY == "NEVADA"], rnd.seed
```

```
## Waiting for parallel processing to complete...done.
```

```
CountyPctUpToDateBest
```

```
## MCMC fit results for BEST analysis:
## 100002 simulations saved.
##           mean       sd median   HDIlo   HDIup Rhat n.eff
## mu1    0.97102 0.003147 0.9711 0.96483 0.97718    1 47802
## mu2    0.75205 0.053087 0.7583 0.64010 0.84634    1 35471
## nu     2.02627 0.460776 1.9657 1.22115 2.95016    1 23377
## sigma1 0.02375 0.003012 0.0236 0.01803 0.02979    1 28353
## sigma2 0.16767 0.055668 0.1610 0.07004 0.27794    1 25520
##
## 'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.
## 'Rhat' is the potential scale reduction factor (at convergence, Rhat=1).
## 'n.eff' is a crude measure of effective sample size.
```

```
plot(CountyPctUpToDateBest)
```



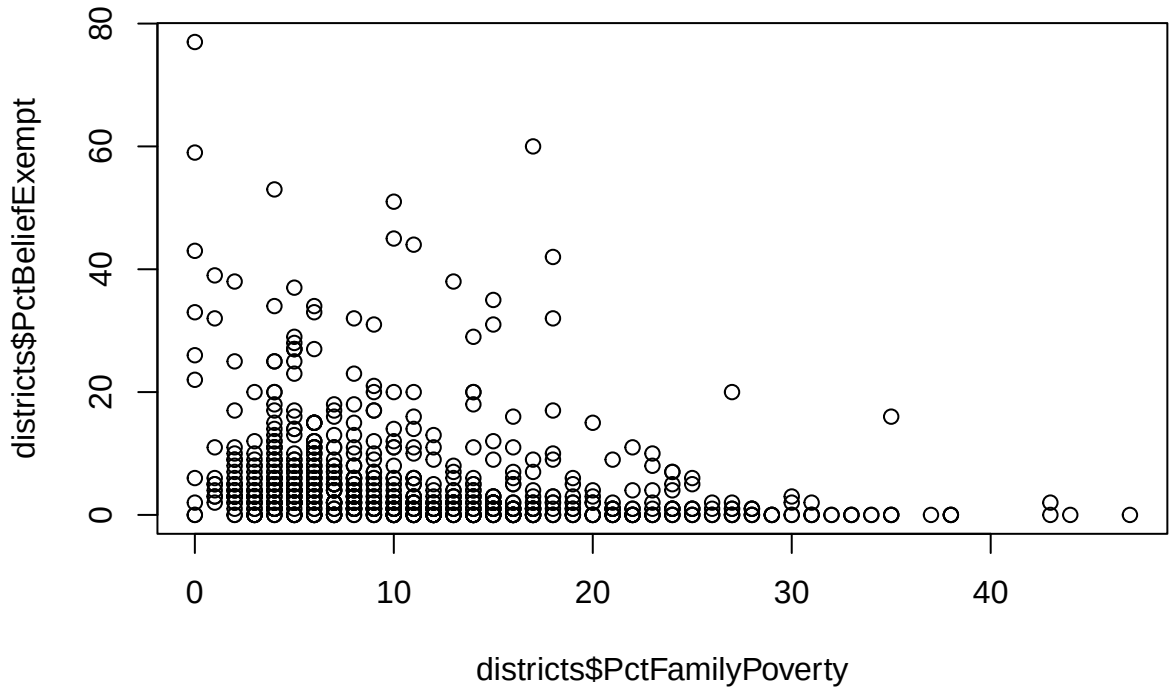

```
vaccineAnova <- aov(PctUpToDate ~ COUNTY, data = schools)
tukey_result <- TukeyHSD(vaccineAnova)
#tukey_result

# comment out the print command, as the output is too long
```

5. *Inferential reporting about districts*

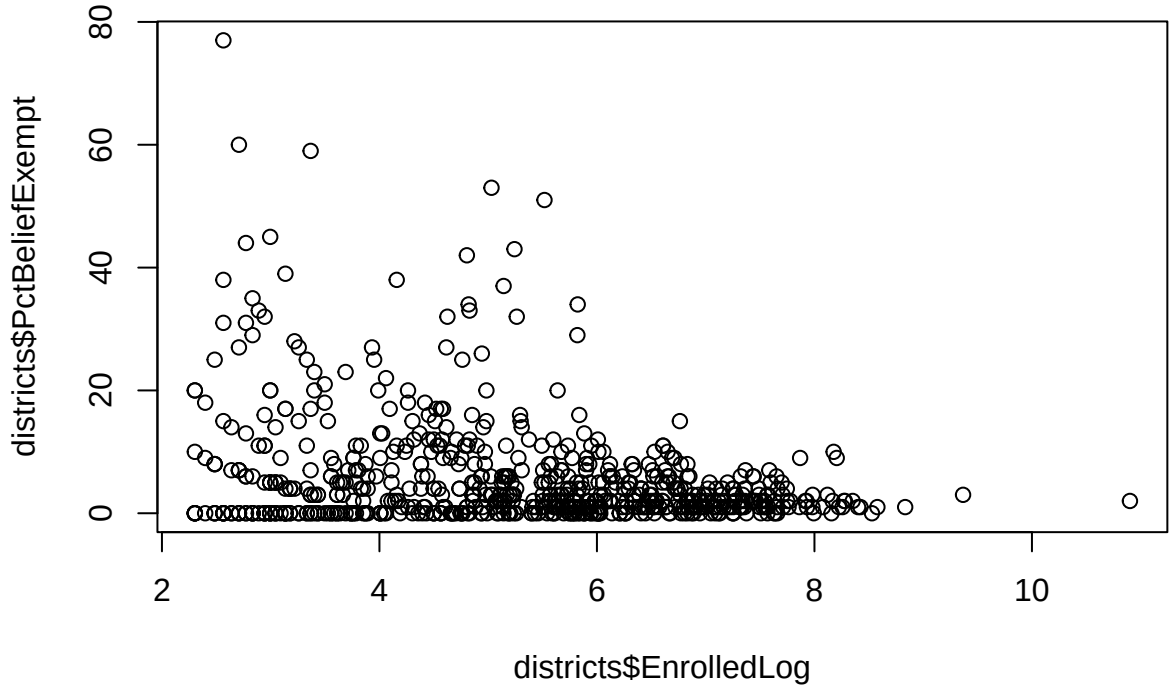
For every item below except question b, use PctChildPoverty, PctFamilyPoverty, Enrolled, and TotalSchools as the four predictors, transformed as necessary to improve prediction and/or interpretability. Be sure to include appropriate diagnostics and modify your analyses as appropriate.

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

```
plot(districts$PctFamilyPoverty, districts$PctBeliefExempt)
```




```
plot(districts$EnrolledLog, districts$PctBeliefExempt)
```




```
districtsQ5 <- districts %>% dplyr::select(c(PctChildPoverty, PctFamilyPoverty, EnrolledLog, TotalSchool
cor(districtsQ5)
```

```
##                  PctChildPoverty PctFamilyPoverty EnrolledLog TotalSchoolsLog
## PctChildPoverty       1.00000000      0.867776828 -0.05301706    -0.087286318
## PctFamilyPoverty      0.86777683      1.000000000  0.05546011    -0.005408707
```

```
## EnrolledLog          -0.05301706      0.055460112  1.00000000     0.916319391
## TotalSchoolsLog      -0.08728632     -0.005408707  0.91631939     1.000000000
## PctBeliefExempt      -0.20130543     -0.265543568 -0.27277972    -0.199318123
##                   PctBeliefExempt
## PctChildPoverty       -0.2013054
## PctFamilyPoverty      -0.2655436
## EnrolledLog           -0.2727797
## TotalSchoolsLog       -0.1993181
## PctBeliefExempt        1.0000000
```

```
model <- lm(PctBeliefExempt ~ PctChildPoverty + PctFamilyPoverty + EnrolledLog + TotalSchoolsLog, data =

library(car)
```

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

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

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

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

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

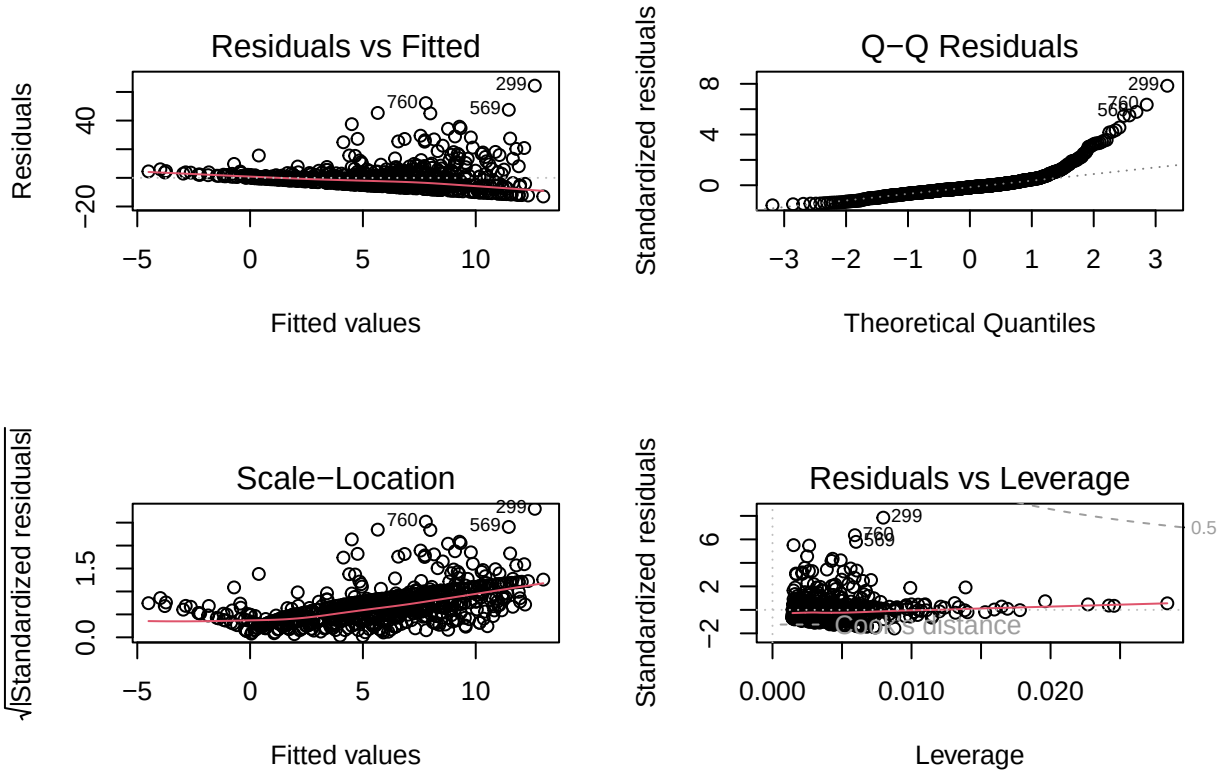
```
vif(model)
```

```
##  PctChildPoverty PctFamilyPoverty      EnrolledLog  TotalSchoolsLog
##         4.237053         4.283179         6.492899         6.380393
```

```
model <- lm(PctBeliefExempt ~ PctFamilyPoverty + EnrolledLog , data = districts)

library(car)
vif(model)
```

```
## PctFamilyPoverty      EnrolledLog
##         1.003085         1.003085
```

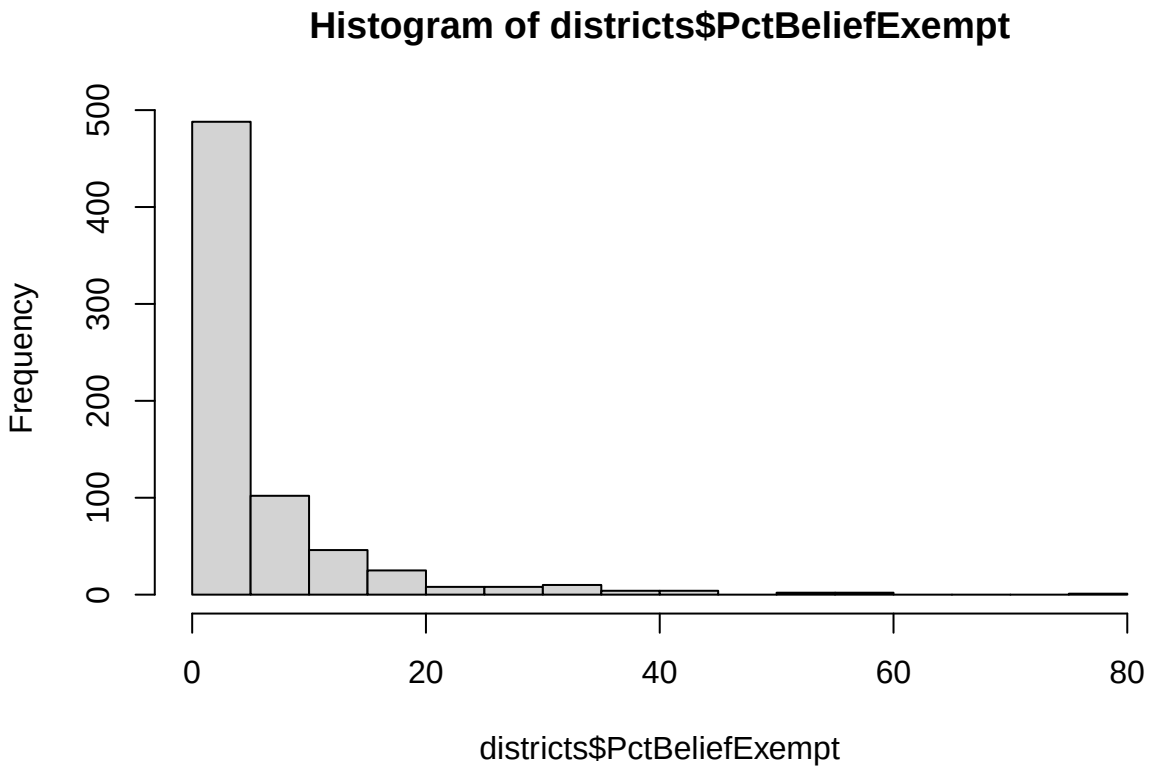
```
par(mfrow=c(2,2))
plot(model)
```



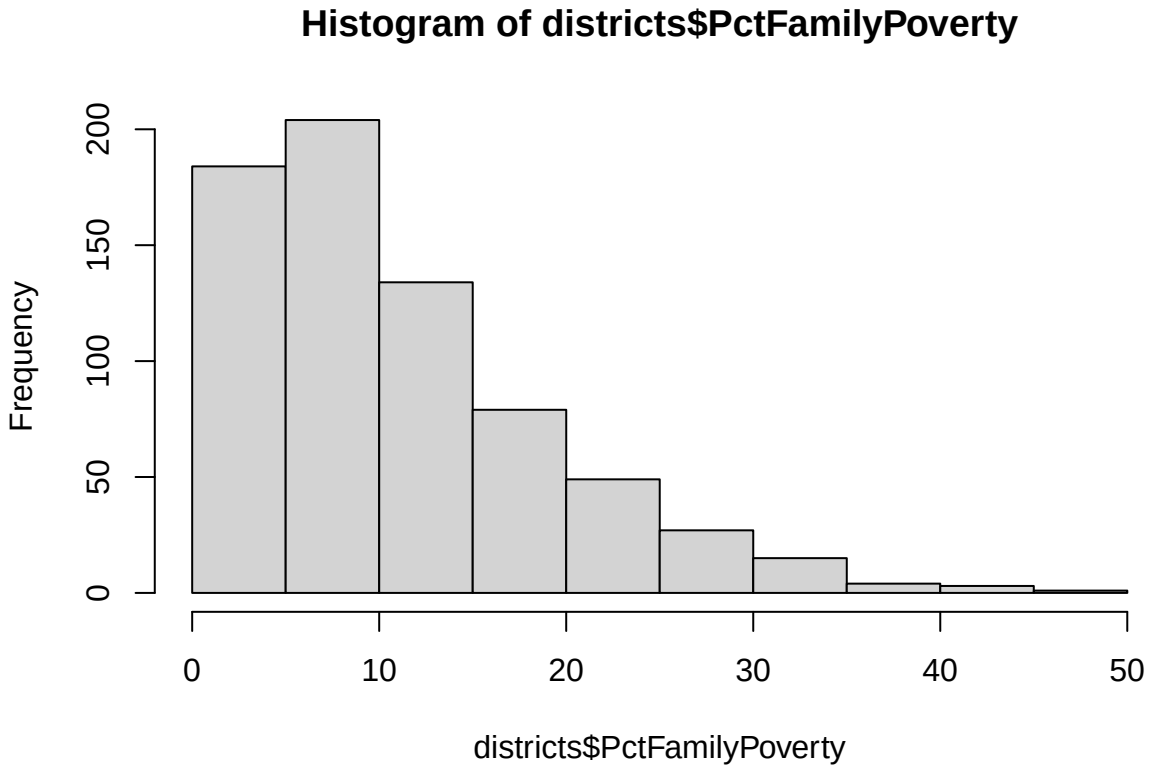

```
library(gvlma)
gvlma(model)
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ PctFamilyPoverty + EnrolledLog,
##     data = districts)
##
## Coefficients:
##      (Intercept)  PctFamilyPoverty       EnrolledLog
##          16.2999           -0.2719           -1.4361
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = model)
##
##                      Value p-value                   Decision
## Global Stat        6522.061 0.00000 Assumptions NOT satisfied!
## Skewness           1038.483 0.00000 Assumptions NOT satisfied!
## Kurtosis           5473.220 0.00000 Assumptions NOT satisfied!
## Link Function         5.378 0.02039 Assumptions NOT satisfied!
## Heteroscedasticity    4.980 0.02565 Assumptions NOT satisfied!
```

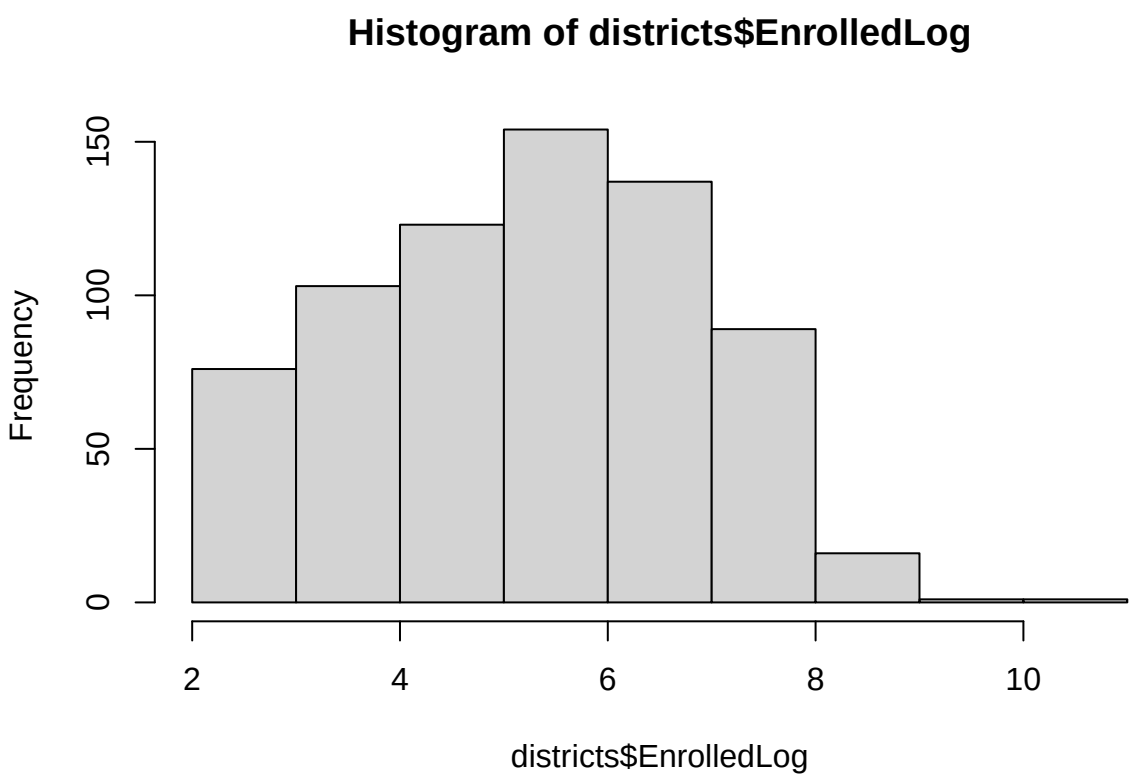
```
hist(districts$PctBeliefExempt)
```




```
hist(districts$PctFamilyPoverty)
```



```r
hist(districts$EnrolledLog)
```



There's evidence of a curvilinear relationship

While the residuals show outliers and a heavy-tailed distribution, and the gvlma indicates problems with skewness, kurtosis, and heteroscedasticity, the model still provides a potentially useful approximation of the underlying relationship. Given the current analysis context, we will proceed with the model while acknowledging these limitations.

```r
summary(model)
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ PctFamilyPoverty + EnrolledLog,
##     data = districts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.993  -4.207  -1.472   1.578  64.384
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      16.29988    1.13867  14.315  < 2e-16 ***
## PctFamilyPoverty -0.27191    0.03814  -7.129 2.54e-12 ***
## EnrolledLog      -1.43609    0.19549  -7.346 5.71e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.238 on 697 degrees of freedom
```

```
## Multiple R-squared:  0.1373, Adjusted R-squared:  0.1348
## F-statistic: 55.47 on 2 and 697 DF,  p-value: < 2.2e-16
```

```
library(lm.beta)
lm.beta(model)
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ PctFamilyPoverty + EnrolledLog,
##     data = districts)
##
## Standardized Coefficients::
##      (Intercept) PctFamilyPoverty      EnrolledLog
##               NA       -0.2511878       -0.2588488
```

```
regOutBF <- lmBF(PctBeliefExempt ~ PctFamilyPoverty + EnrolledLog, data = districts)
summary(regOutBF)
```

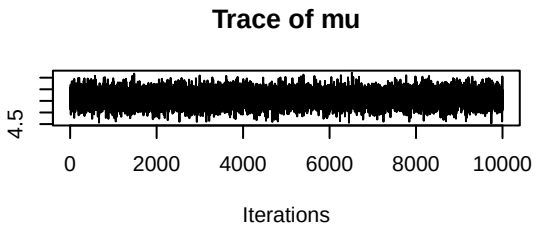
```
## Bayes factor analysis
## --------------
## [1] PctFamilyPoverty + EnrolledLog : 8.821259e+19 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

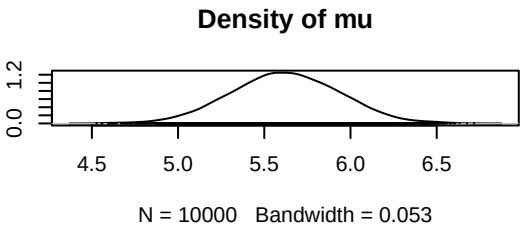
```
regOutMCMC <- lmBF(PctBeliefExempt ~ PctFamilyPoverty + EnrolledLog,
                   data = districts,
                   posterior = TRUE,
                   iterations = 10000)
summary(regOutMCMC)
```

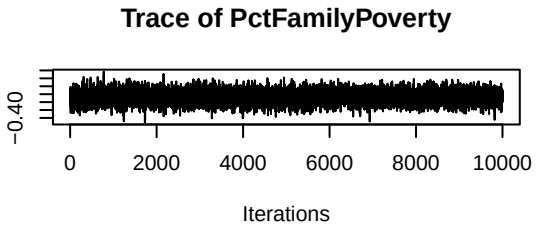
```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                      Mean      SD  Naive SE Time-series SE
## mu                 5.6220 0.31545 0.0031545      0.0031042
## PctFamilyPoverty -0.2682 0.03874 0.0003874      0.0003949
## EnrolledLog      -1.4140 0.19754 0.0019754      0.0019754
## sig2             68.0691 3.68867 0.0368867      0.0368867
## g                 0.2583 0.73096 0.0073096      0.0073096
##
## 2. Quantiles for each variable:
##
##                      2.5%      25%      50%      75%    97.5%
```

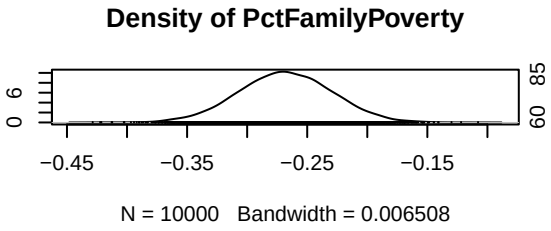
```
## mu                5.00780  5.41136  5.6210  5.8360  6.2347
## PctFamilyPoverty -0.34458 -0.29432 -0.2683 -0.2421 -0.1925
## EnrolledLog      -1.79756 -1.54733 -1.4148 -1.2793 -1.0222
## sig2             61.26655 65.49977 67.9116 70.4848 75.6630
## g                 0.02873  0.06747  0.1185  0.2369  1.3309
```

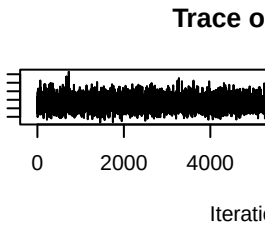
```
plot(regOutMCMC)
```

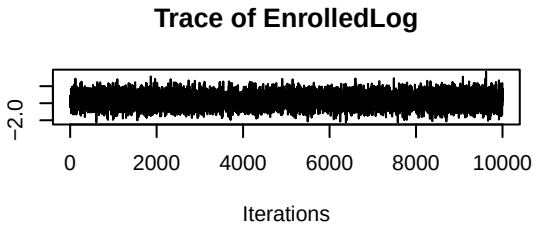


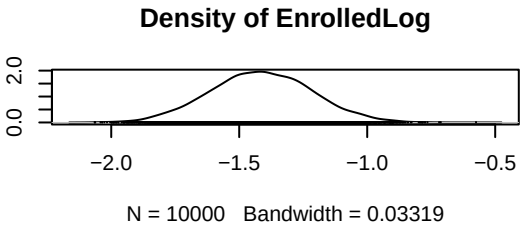


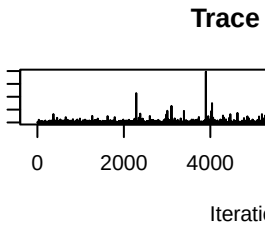













```
rsqList <- 1 - (regOutMCMC[,"sig2"] / var(districts$PctBeliefExempt))
mean(rsqList)
```

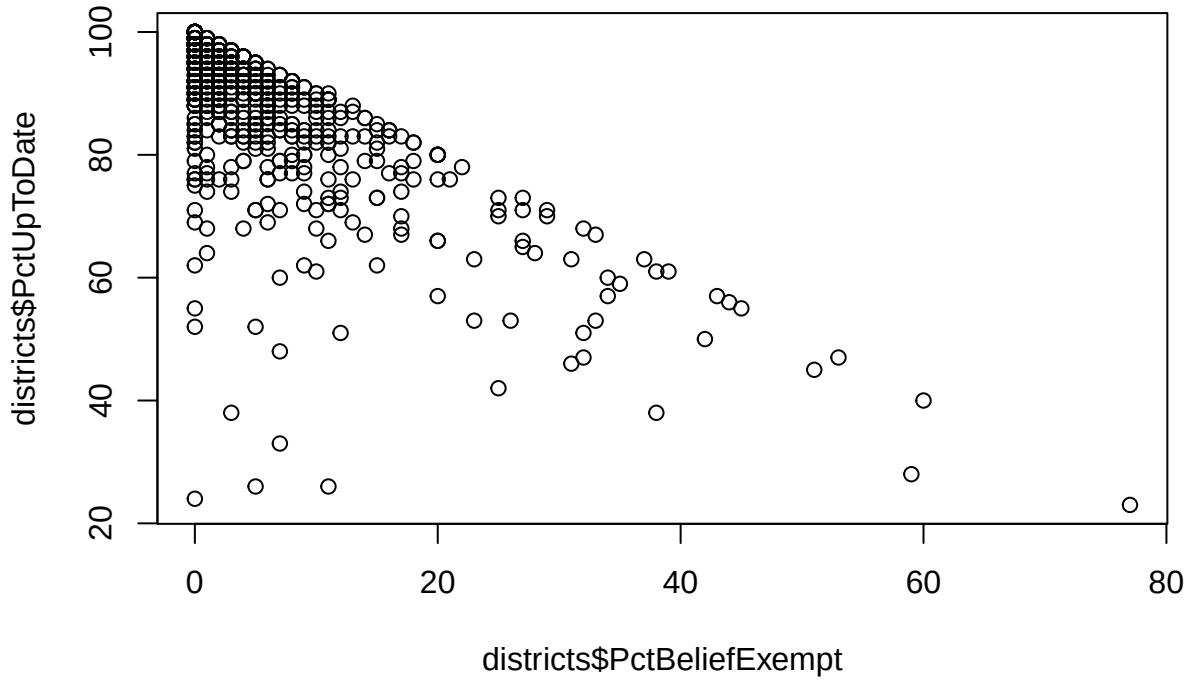
```
## [1] 0.1322593
```

```
quantile(rsqList, c(0.025,0.975))
```

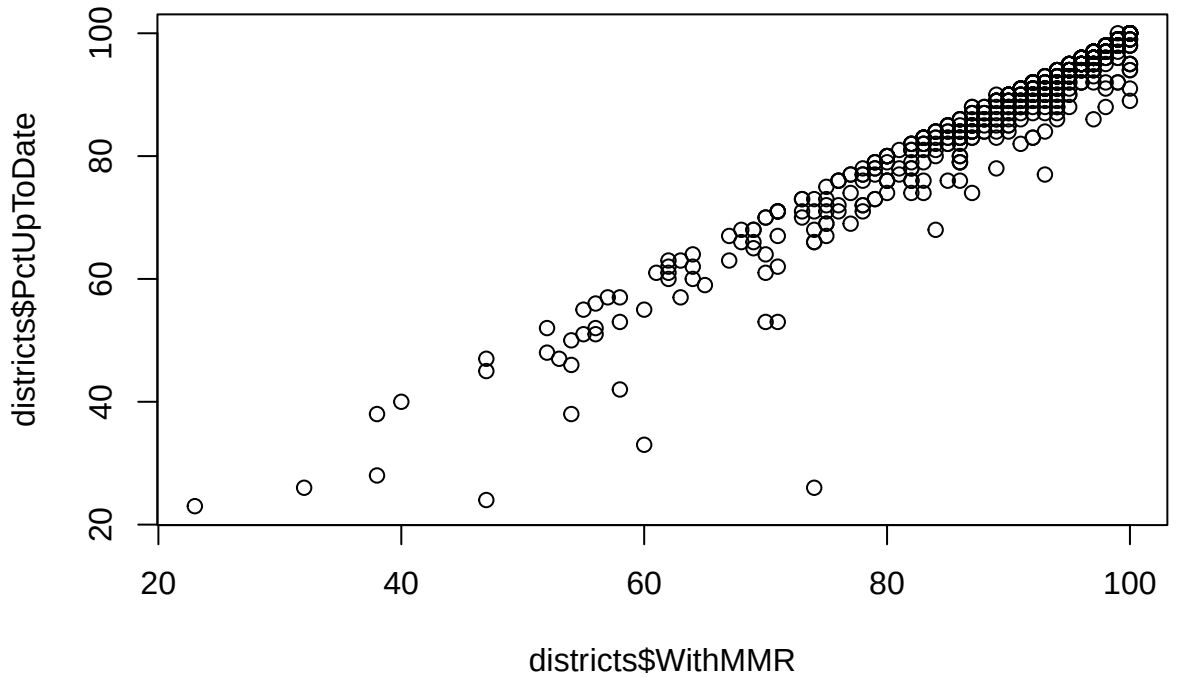
```
##       2.5%      97.5%
## 0.03545321 0.21897832
```

b. *Using any set or combination of predictors that you want to use, what combination gives the best R-squared in predicting the percentage of all enrolled students with completely up-to-date vaccines while still being an acceptable regression?*

```
plot(districts$PctBeliefExempt, districts$PctUpToDate)
```



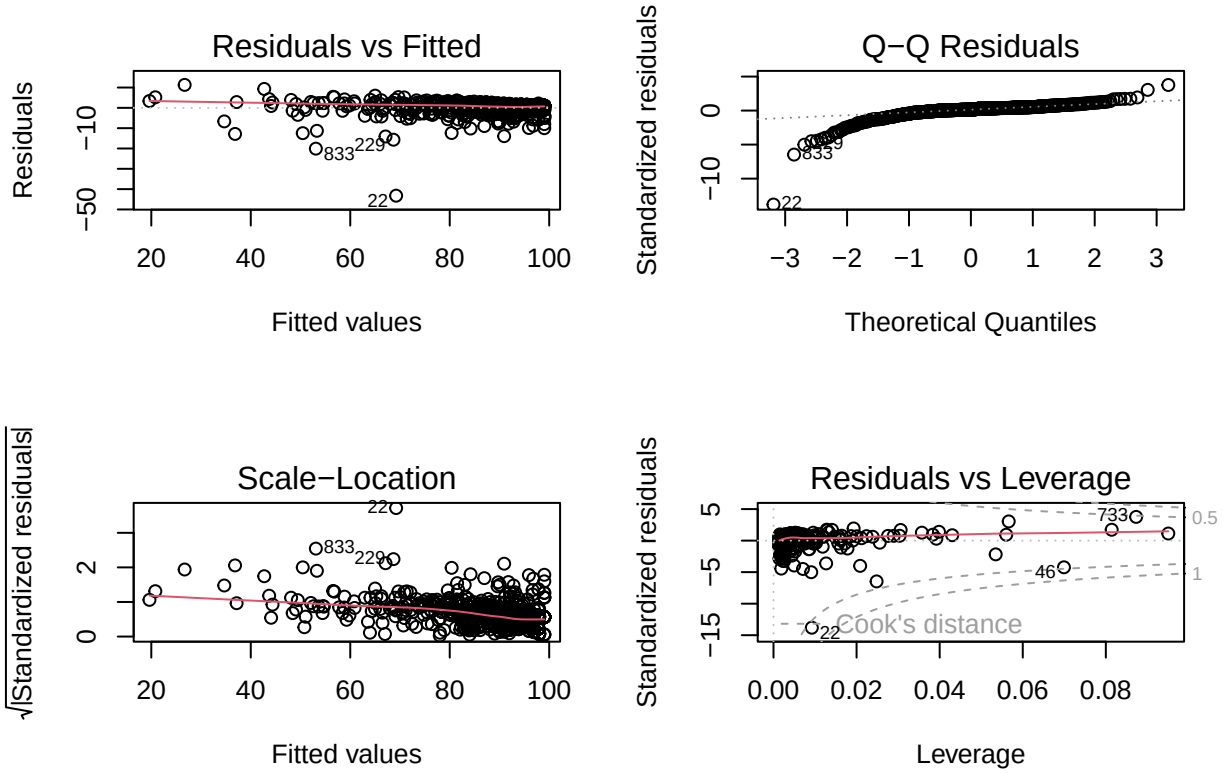

```
plot(districts$WithMMR, districts$PctUpToDate)
```




```
lm_modelNew <- lm(PctUpToDate ~ WithMMR + PctBeliefExempt, data = districts)

library(car)
vif(lm_modelNew)
```

```
##          WithMMR PctBeliefExempt
##         2.600962        2.600962
```

```
par(mfrow=c(2,2))
plot(lm_modelNew)
```




```
districts[rownames(districts) == "22", ]
```

```
##               DistrictName WithDTP WithPolio WithMMR WithHepB PctUpToDate
## 22 Alpine Union Elementary      74        74      74       91          26
##    DistrictComplete PctBeliefExempt PctMedicalExempt PctChildPoverty
## 22             TRUE               5                0               9
##    PctFamilyPoverty EnrolledLog TotalSchoolsLog
## 22                5    5.214936               0
```

```
districts[rownames(districts) == "46", ]
```

```
##      DistrictName WithDTP WithPolio WithMMR WithHepB PctUpToDate
## 46 Warner Unified      53        53      47       88          24
##    DistrictComplete PctBeliefExempt PctMedicalExempt PctChildPoverty
## 46             TRUE               0                0               7
##    PctFamilyPoverty EnrolledLog TotalSchoolsLog
## 46               12    2.833213               0
```

```
districts[rownames(districts) == "733", ]
```

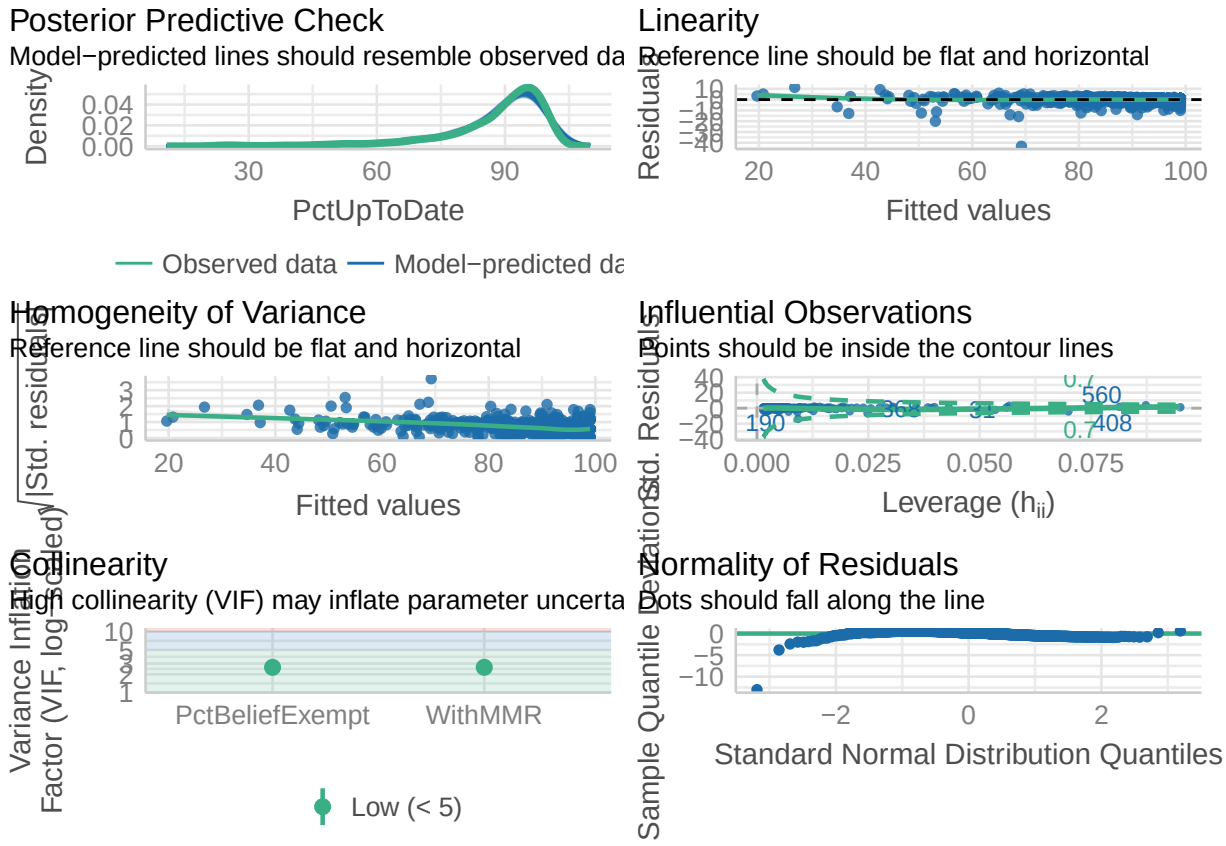
```
##           DistrictName WithDTP WithPolio WithMMR WithHepB PctUpToDate
## 733 Fort Bragg Unified      38        38      38       97          38
##     DistrictComplete PctBeliefExempt PctMedicalExempt PctChildPoverty
```

```
## 733            FALSE               3               0              23
##     PctFamilyPoverty EnrolledLog TotalSchoolsLog
## 733               11    4.976734       0.6931472
```

```
library(gvlma)
gvlma(lm_modelNew)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ WithMMR + PctBeliefExempt, data = districts)
##
## Coefficients:
##     (Intercept)          WithMMR  PctBeliefExempt
##        -18.2937           1.1730           0.1422
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = lm_modelNew)
##
##                       Value   p-value                   Decision
## Global Stat        94636.373 0.000e+00 Assumptions NOT satisfied!
## Skewness            3138.475 0.000e+00 Assumptions NOT satisfied!
## Kurtosis           91470.766 0.000e+00 Assumptions NOT satisfied!
## Link Function          3.137 7.654e-02    Assumptions acceptable.
## Heteroscedasticity    23.994 9.663e-07 Assumptions NOT satisfied!
```

```
library(performance)
check_model(lm_modelNew)
```




```
summary(lm_modelNew)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ WithMMR + PctBeliefExempt, data = districts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.222  -0.433   0.515   1.303  11.291
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -18.29365    1.62075 -11.287  < 2e-16 ***
## WithMMR           1.17304    0.01692  69.334  < 2e-16 ***
## PctBeliefExempt   0.14218    0.02166   6.565 1.02e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.144 on 697 degrees of freedom
## Multiple R-squared:  0.9392, Adjusted R-squared:  0.939
## F-statistic:  5379 on 2 and 697 DF,  p-value: < 2.2e-16
```

```
library(lm.beta)
lm.beta(lm_modelNew)
```

```
##
```

```
## Call:
## lm(formula = PctUpToDate ~ WithMMR + PctBeliefExempt, data = districts)
##
## Standardized Coefficients::
##     (Intercept)         WithMMR PctBeliefExempt
##              NA      1.04477042      0.09892931
```

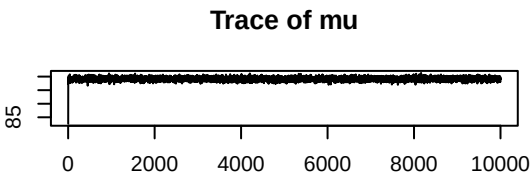
```
regOutBF <- lmBF(PctUpToDate ~ WithMMR + PctBeliefExempt, data = districts)
summary(regOutBF)
```

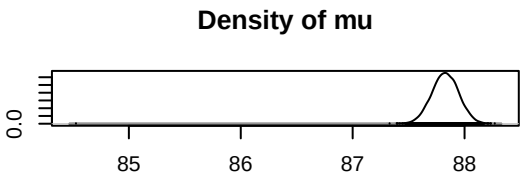
```
## Bayes factor analysis
## --------------
## [1] WithMMR + PctBeliefExempt : 6.658546e+419 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

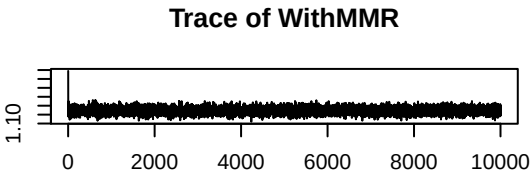
```
regOutMCMC <- lmBF(PctUpToDate ~ WithMMR + PctBeliefExempt,
                   data = districts,
                   posterior = TRUE,
                   iterations = 10000)
summary(regOutMCMC)
```

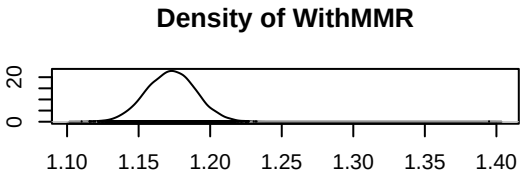
```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                     Mean       SD  Naive SE Time-series SE
## mu               87.8273  0.12334 0.0012334      0.0012334
## WithMMR           1.1727  0.01724 0.0001724      0.0001755
## PctBeliefExempt   0.1422  0.02182 0.0002182      0.0002182
## sig2              9.9315  0.54866 0.0054866      0.0054866
## g                13.8717 43.90360 0.4390360      0.4582247
##
## 2. Quantiles for each variable:
##
##                     2.5%     25%     50%     75%   97.5%
## mu               87.59439 87.7479 87.8277 87.9081 88.0593
## WithMMR           1.13944  1.1609  1.1729  1.1842  1.2062
## PctBeliefExempt   0.09908  0.1275  0.1426  0.1569  0.1848
## sig2              8.93086  9.5676  9.9166 10.2745 11.0154
## g                 1.63557  3.7631  6.4554 12.6010 68.2555
```

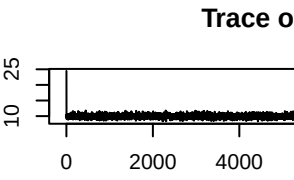
```
plot(regOutMCMC)
```

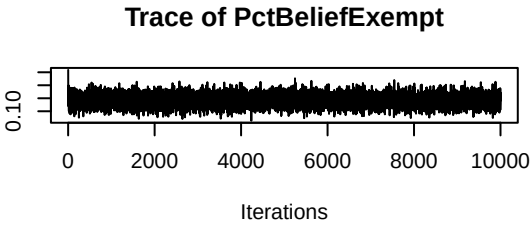


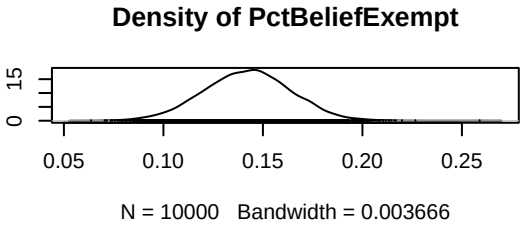


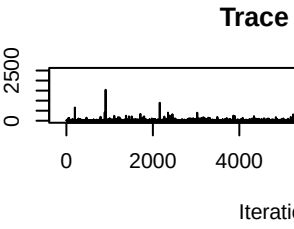













```
rsqList <- 1 - (regOutMCMC[,"sig2"] / var(districts$PctBeliefExempt))
mean(rsqList)
```

```
## [1] 0.8733939
```

```
quantile(rsqList, c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.8595762 0.8861500
```

try PCA

```
library(paran)
```

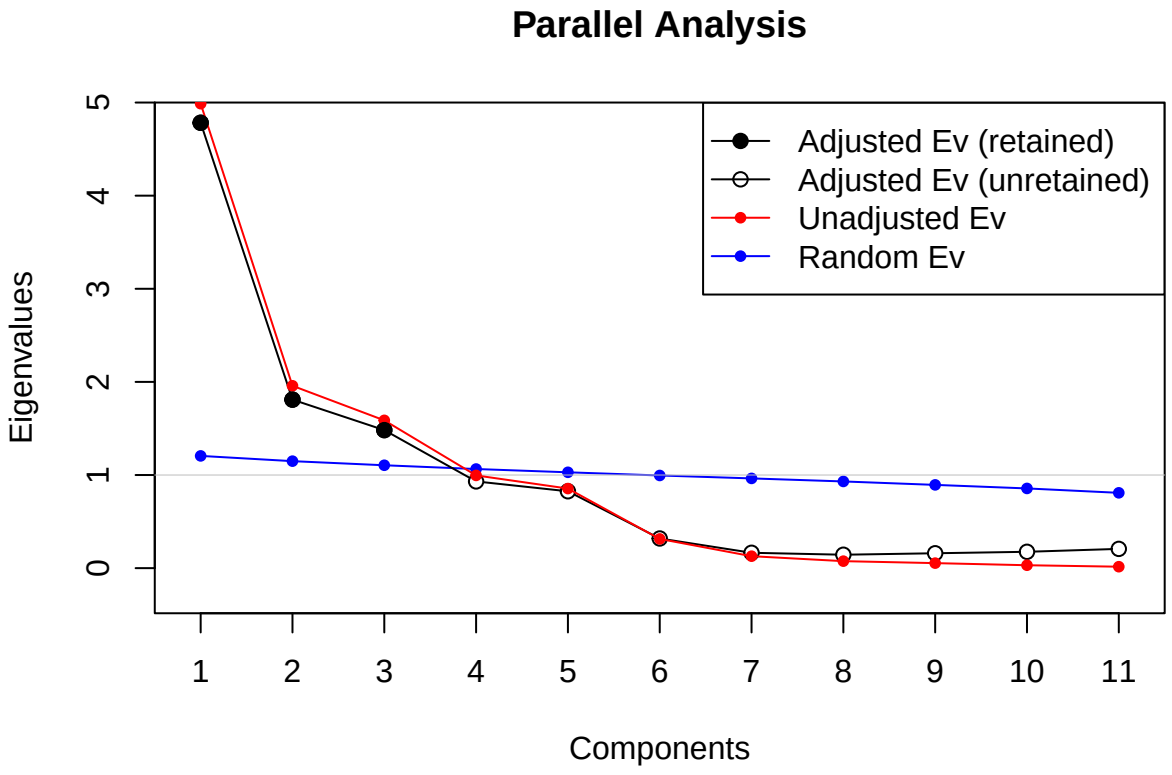
```
## Loading required package: MASS
```

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

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

```
paranOut <- districts %>% dplyr::select(-c(DistrictName,PctUpToDate)) %>%
paran::paran(iterations=330, graph=TRUE)
```

```
##
## Using eigendecomposition of correlation matrix.
## Computing: 10%  20%  30%  40%  50%  60%  70%  80%  90%  100%
##
##
## Results of Horn's Parallel Analysis for component retention
## 330 iterations, using the mean estimate
##
## ----------------------------------------------------
## Component   Adjusted    Unadjusted    Estimated
##             Eigenvalue  Eigenvalue    Bias
## ----------------------------------------------------
## 1           4.782578    4.987712      0.205133
## 2           1.809523    1.958274      0.148751
## 3           1.482180    1.587007      0.104826
## ----------------------------------------------------
##
## Adjusted eigenvalues > 1 indicate dimensions to retain.
## (3 components retained)
```




```
library(psych)
prinOut <- districts %>% dplyr::select(-c(DistrictName,PctUpToDate)) %>%
  psych::principal(nfactors = paranOut$Retained,
residuals = FALSE, rotate="varimax", scores=TRUE)
```

```r
print(prinOut, cut=0.4, sort=TRUE)
```

```
## Principal Components Analysis
## Call: psych::principal(r = ., nfactors = paranOut$Retained, residuals = FALSE,
##     rotate = "varimax", scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                   item   RC1   RC2   RC3    h2    u2 com
## WithPolio            2  0.97             0.955 0.045 1.0
## WithDTP              1  0.97             0.946 0.054 1.0
## WithMMR              3  0.96             0.933 0.067 1.0
## WithHepB             4  0.95             0.920 0.080 1.0
## PctBeliefExempt      6 -0.89             0.804 0.196 1.0
## TotalSchoolsLog     11        0.93       0.925 0.075 1.1
## EnrolledLog         10        0.89       0.891 0.109 1.2
## DistrictComplete     5       -0.48       0.315 0.685 1.7
## PctMedicalExempt     7                   0.024 0.976 1.9
## PctChildPoverty      8              0.94 0.914 0.086 1.1
## PctFamilyPoverty     9              0.93 0.906 0.094 1.1
##
##                        RC1  RC2  RC3
## SS loadings           4.70 1.94 1.90
## Proportion Var        0.43 0.18 0.17
## Cumulative Var        0.43 0.60 0.78
## Proportion Explained  0.55 0.23 0.22
## Cumulative Proportion 0.55 0.78 1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.06
##  with the empirical chi square  234.83  with prob <  5.2e-36
##
## Fit based upon off diagonal values = 0.98
```

```r
scoresData <- data.frame(prinOut$scores) %>% mutate(PctUpToDate = districts$PctUpToDate)
```

```r
cor(scoresData)
```

```
##                        RC1           RC2           RC3 PctUpToDate
## RC1          1.000000e+00 -1.768574e-16 -3.078129e-16  0.92208609
## RC2         -1.768574e-16  1.000000e+00 -1.287160e-15  0.05278925
## RC3         -3.078129e-16 -1.287160e-15  1.000000e+00  0.09224878
## PctUpToDate  9.220861e-01  5.278925e-02  9.224878e-02  1.00000000
```

```r
summary(scoresData)
```

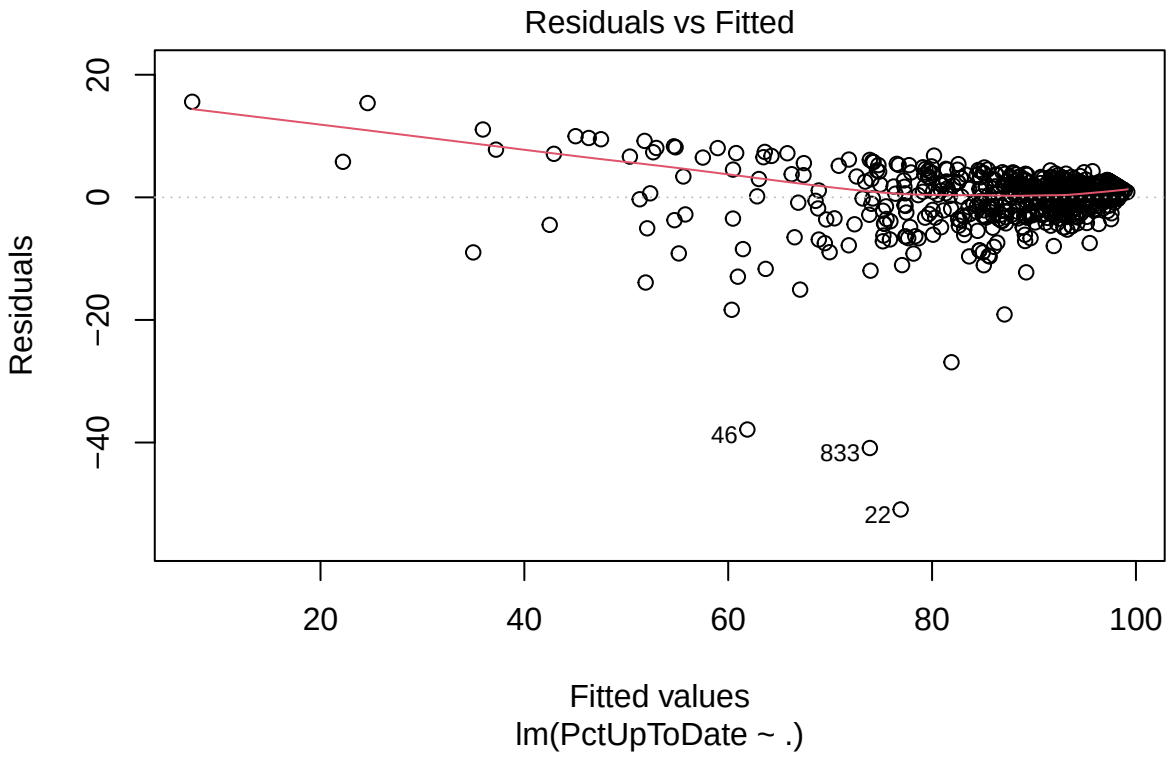
```
##       RC1                RC2                RC3            PctUpToDate
##  Min.   :-6.8649   Min.   :-2.21725   Min.   :-1.6108   Min.   : 23.00
##  1st Qu.:-0.2400   1st Qu.:-0.83023   1st Qu.:-0.7649   1st Qu.: 84.00
##  Median : 0.3508   Median :-0.03576   Median :-0.1916   Median : 92.00
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 87.83
##  3rd Qu.: 0.6233   3rd Qu.: 0.70859   3rd Qu.: 0.4977   3rd Qu.: 96.00
##  Max.   : 0.9998   Max.   : 4.98130   Max.   : 4.0975   Max.   :100.00
```

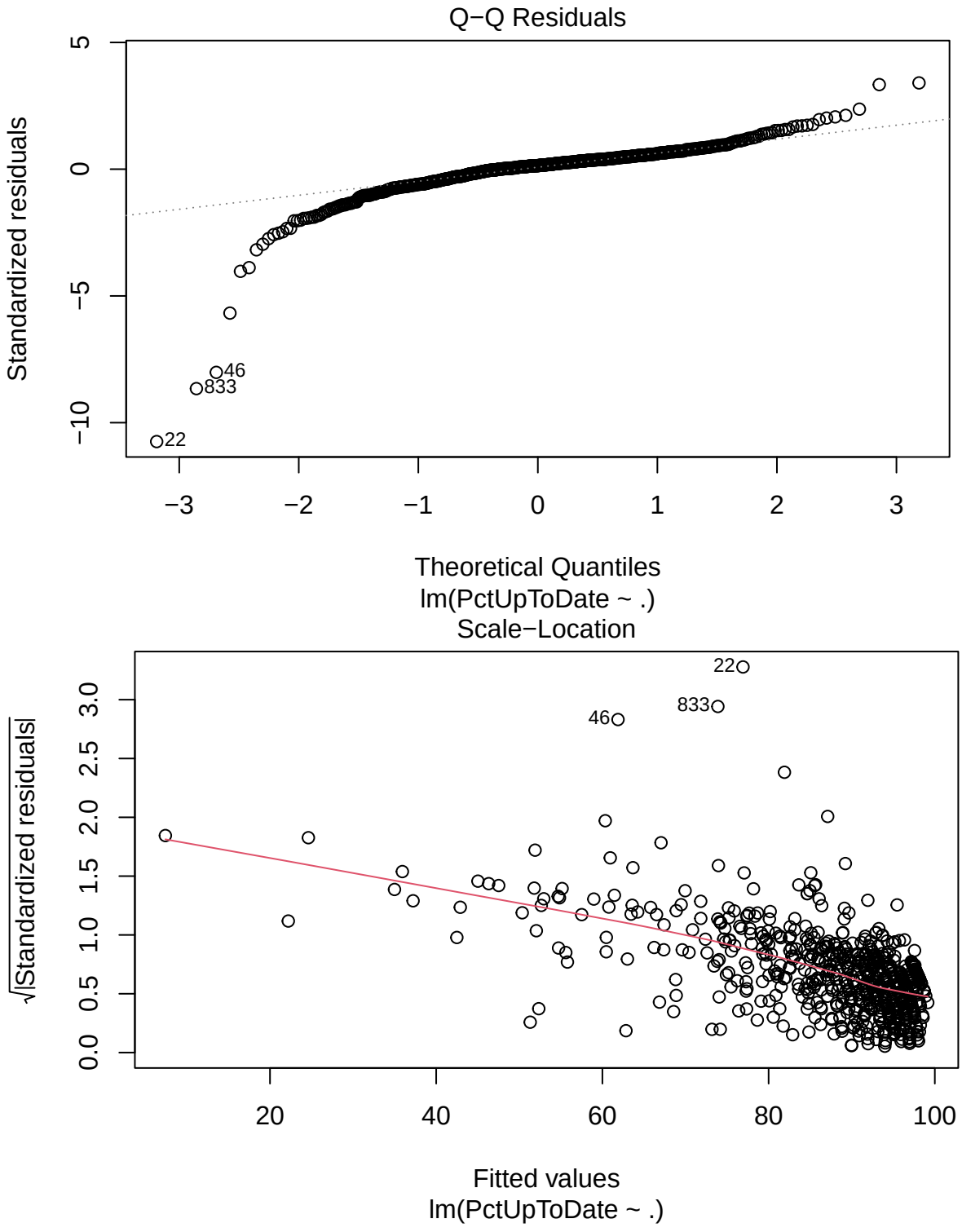
```r
scoresData.lm <- lm(PctUpToDate ~ ., data=scoresData)
```

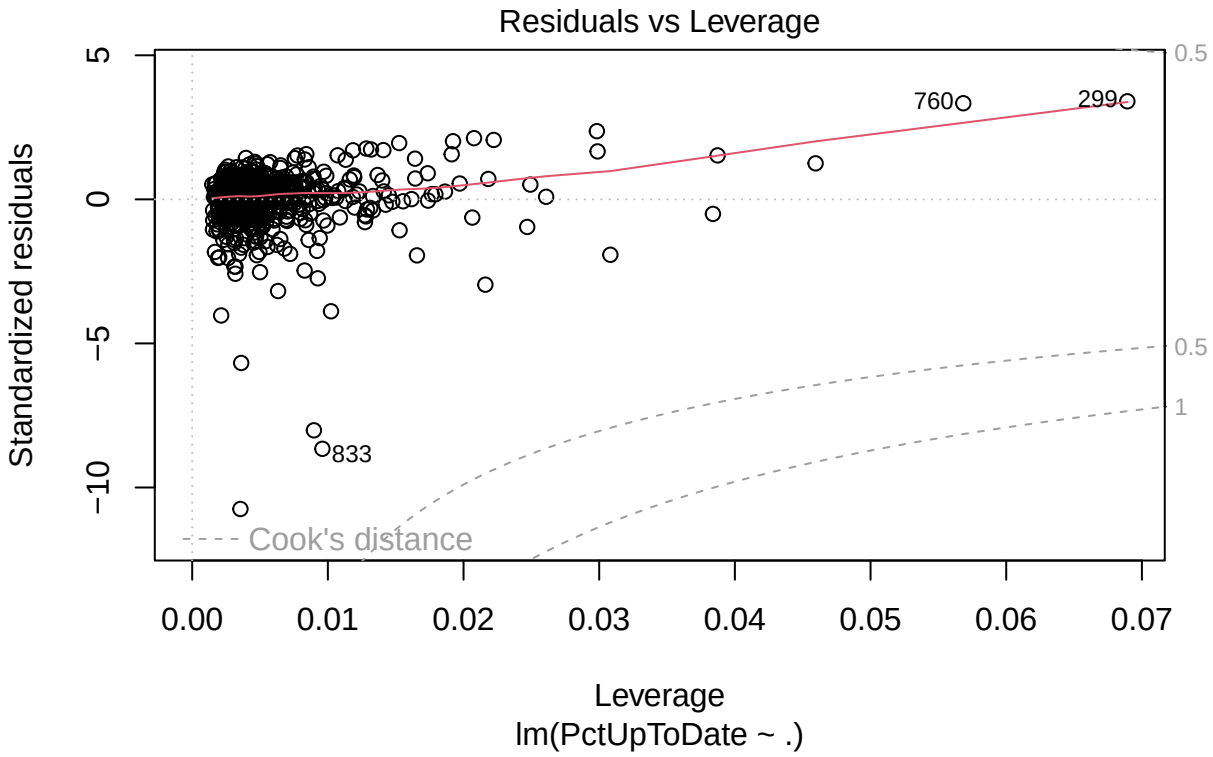
```r
vif(scoresData.lm)
```

```
## RC1 RC2 RC3
##   1   1   1
```

```r
plot(scoresData.lm)
```




Q−Q Residuals
Standardized residuals
5
0
−5
−10
46
833
22
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(PctUpToDate ~ .)
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
22
833
46
20
40
60
80
100
Fitted values
lm(PctUpToDate ~ .)

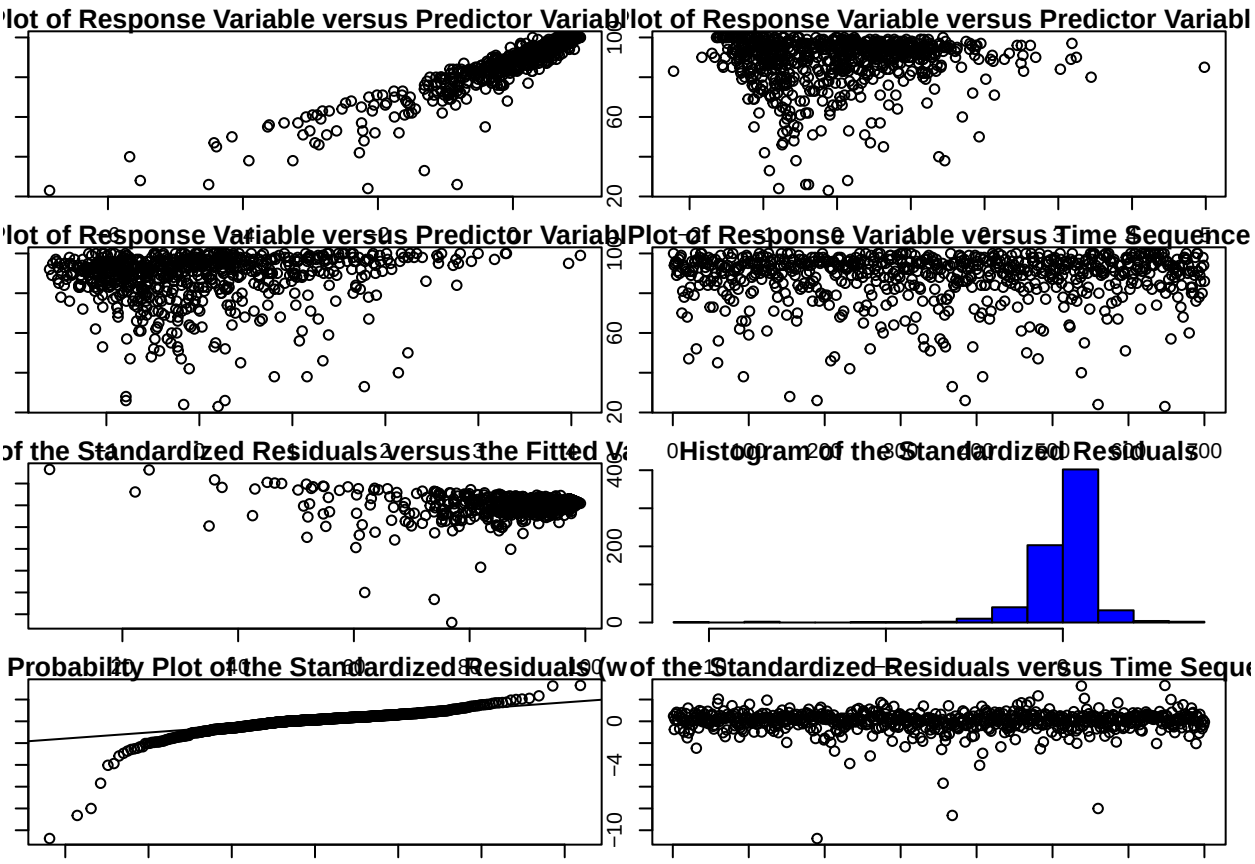


These look okay.

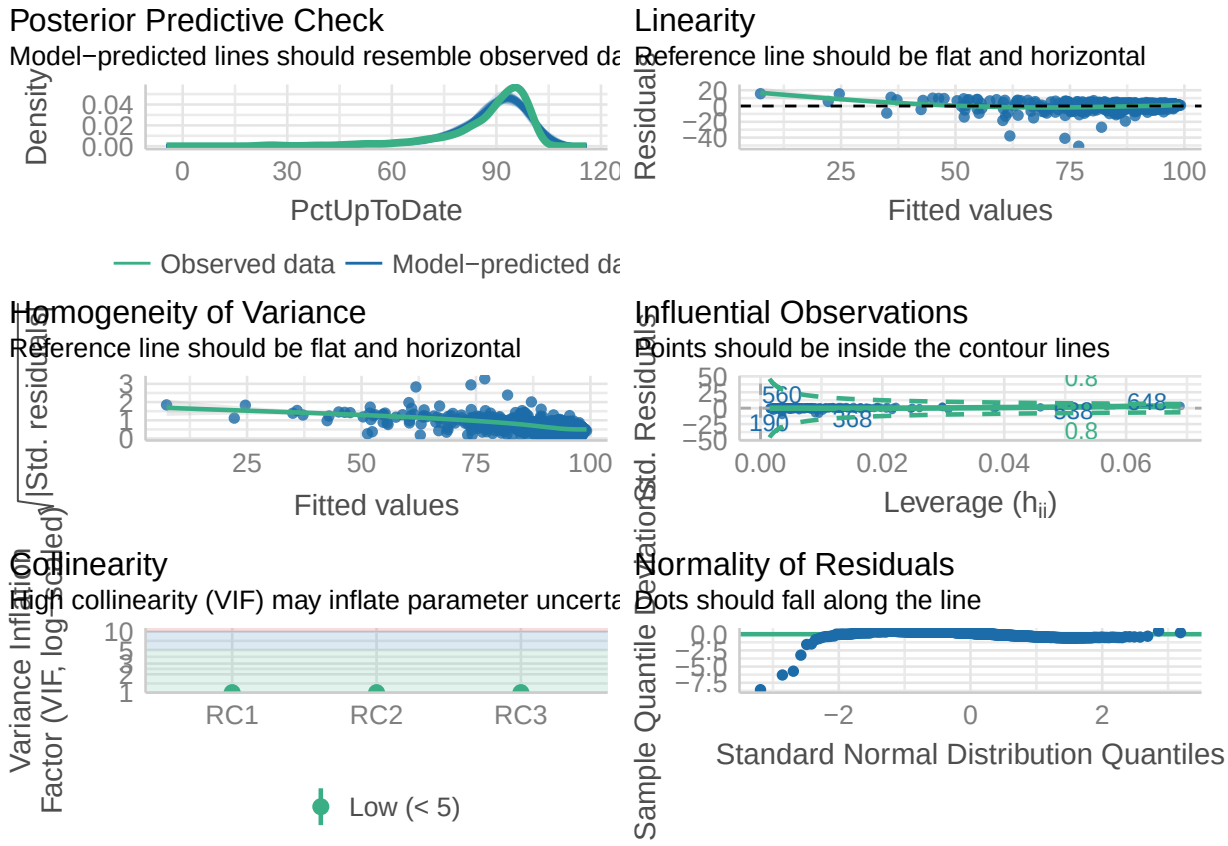
```
library(gvlma)
scoresData.lm.gvlma<-gvlma(scoresData.lm)
scoresData.lm.gvlma
```

```
##
## Call:
## lm(formula = PctUpToDate ~ ., data = scoresData)
##
## Coefficients:
## (Intercept)          RC1          RC2          RC3
##      87.830       11.737        0.672        1.174
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = scoresData.lm)
##
##                       Value   p-value                   Decision
## Global Stat        3.532e+04 0.000e+00 Assumptions NOT satisfied!
## Skewness           1.871e+03 0.000e+00 Assumptions NOT satisfied!
## Kurtosis           3.340e+04 0.000e+00 Assumptions NOT satisfied!
## Link Function      5.251e+01 4.281e-13 Assumptions NOT satisfied!
## Heteroscedasticity 3.075e-04 9.860e-01    Assumptions acceptable.
```

```
par(mar = c(1, 1, 1, 1))
plot(scoresData.lm.gvlma)
```




```
library(performance)
check_model(scoresData.lm)
```

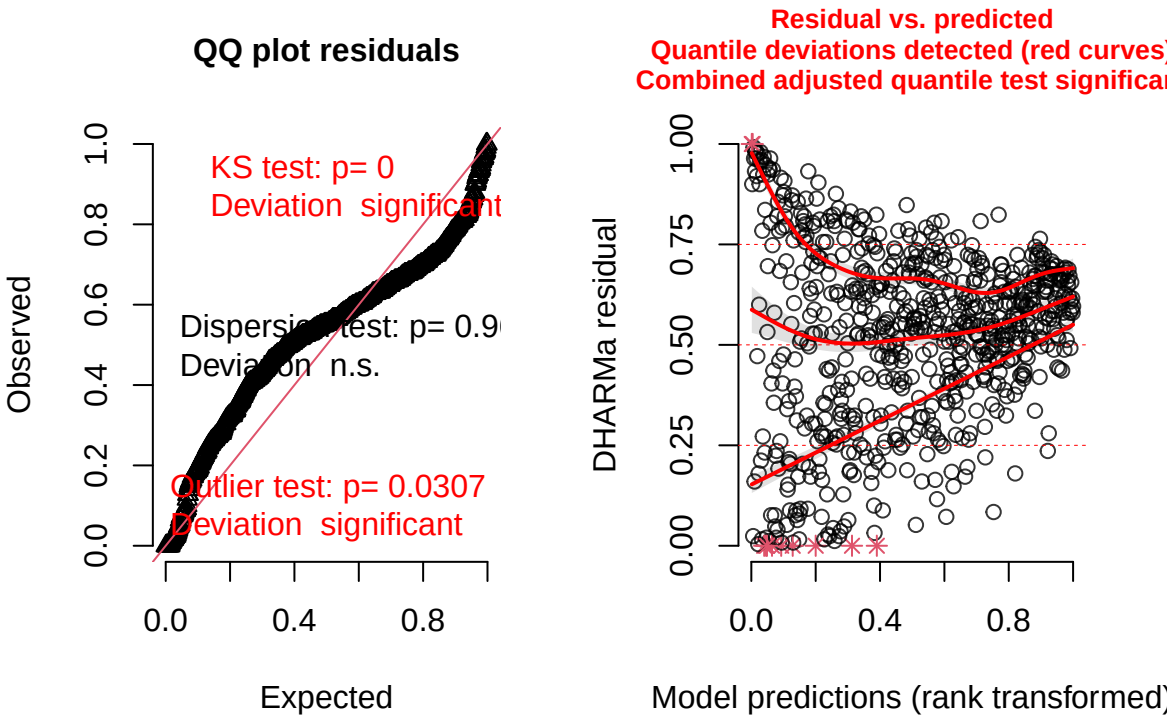



```
library(DHARMa)
```

```
## This is DHARMa 0.4.6. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa')
```

```
simulationOutput <- simulateResiduals(fittedModel = scoresData.lm, n = 250)
plot(simulationOutput)
```

```
## qu = 0.5, log(sigma) = -2.718267 : outer Newton did not converge fully.
```



There's evidence of a curvilinear relationship

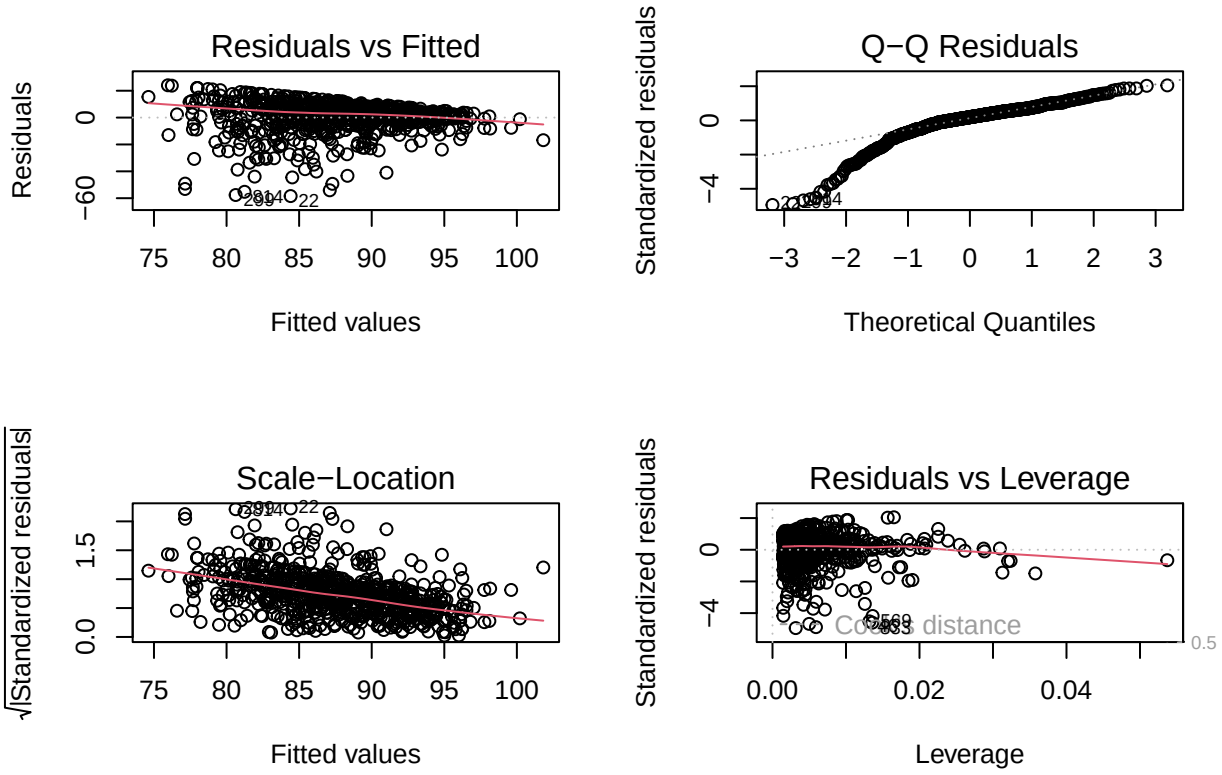
While the residuals show outliers and a heavy-tailed distribution, and the gvlma indicates problems with skewness, kurtosis, and heteroscedasticity, the model still provides a potentially useful approximation of the underlying relationship. Given the current analysis context, we will proceed with the model while acknowledging these limitations.

```
summary(scoresData.lm)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ ., data = scoresData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.918  -1.403   0.667   2.127  15.589
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  87.8300     0.1794 489.550  < 2e-16 ***
## RC1          11.7373     0.1795  65.375  < 2e-16 ***
## RC2           0.6720     0.1795   3.743 0.000197 ***
## RC3           1.1742     0.1795   6.540 1.19e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.747 on 696 degrees of freedom
## Multiple R-squared:  0.8615, Adjusted R-squared:  0.8609
## F-statistic:  1444 on 3 and 696 DF,  p-value: < 2.2e-16
```

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

```
districts$PctChildPovertyCtr <- as.numeric(scale(districts$PctChildPoverty,
                                                 center = TRUE,
                                                 scale = FALSE))
districts$EnrolledLogCtr <- as.numeric(scale(districts$EnrolledLog,
                                             center = TRUE,
                                             scale = FALSE))

model2 <- lm(PctUpToDate ~ PctChildPovertyCtr * EnrolledLogCtr, data = districts)


par(mfrow=c(2,2))
plot(model2)
```




```
summary(gvlma(model2))
```

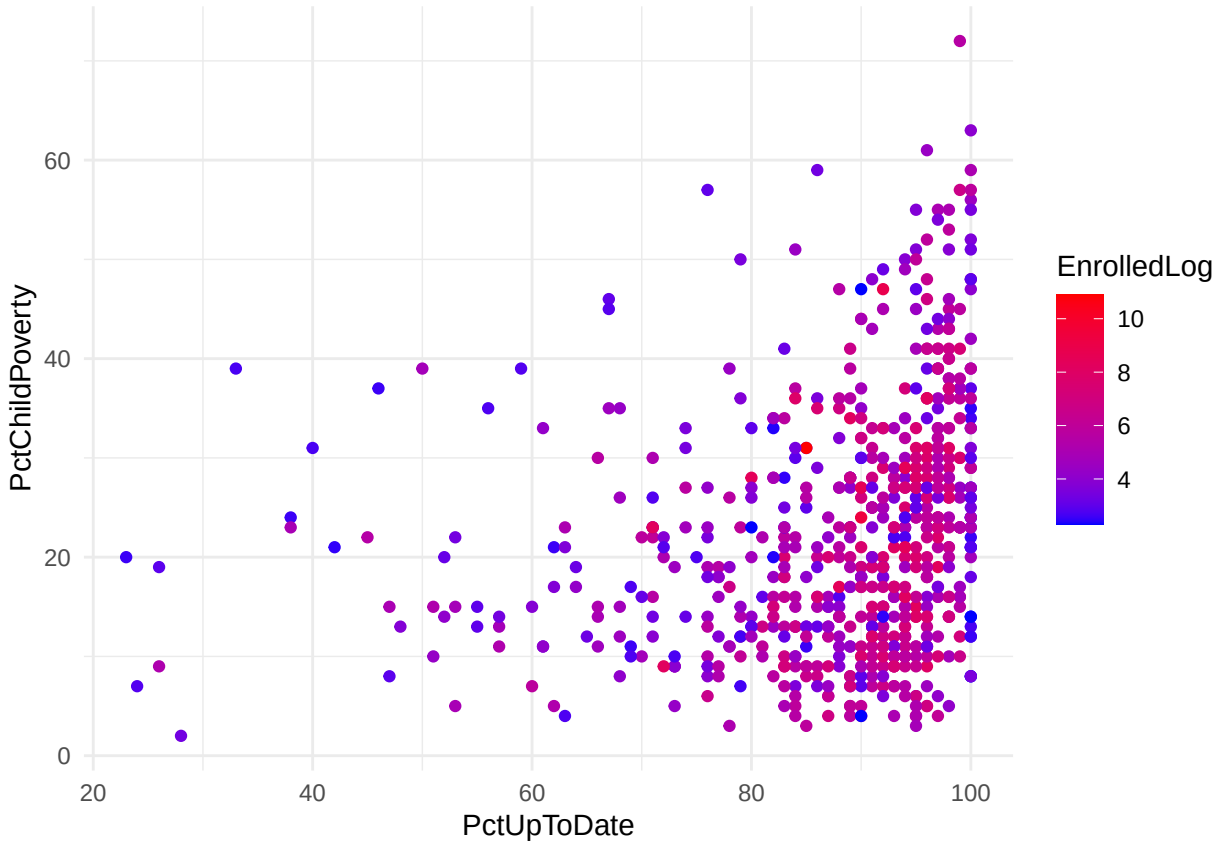
```
##
## Call:
## lm(formula = PctUpToDate ~ PctChildPovertyCtr * EnrolledLogCtr,
##     data = districts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.414  -3.667   2.232   6.820  24.039
```

```
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      87.79834    0.44796 195.994  < 2e-16 ***
## PctChildPovertyCtr                0.24422    0.03717   6.570 9.84e-11 ***
## EnrolledLogCtr                    2.37754    0.28097   8.462  < 2e-16 ***
## PctChildPovertyCtr:EnrolledLogCtr -0.03076    0.02506  -1.227     0.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.83 on 696 degrees of freedom
## Multiple R-squared:  0.1396, Adjusted R-squared:  0.1359
## F-statistic: 37.66 on 3 and 696 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = model2)
##
##                      Value p-value                   Decision
## Global Stat        1055.5096  0.0000 Assumptions NOT satisfied!
## Skewness            364.8020  0.0000 Assumptions NOT satisfied!
## Kurtosis            690.1140  0.0000 Assumptions NOT satisfied!
## Link Function         0.3006  0.5835    Assumptions acceptable.
## Heteroscedasticity    0.2930  0.5883    Assumptions acceptable.
```

```
lmOutBayes1 <- lmBF(PctUpToDate ~ PctChildPovertyCtr + EnrolledLogCtr, data = districts)
lmOutBayes2 <- lmBF(PctUpToDate ~ PctChildPovertyCtr * EnrolledLogCtr, data = districts)

lmOutBayes2 / lmOutBayes1
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPovertyCtr * EnrolledLogCtr : 0.2362283 ±0.01%
##
## Against denominator:
##   PctUpToDate ~ PctChildPovertyCtr + EnrolledLogCtr
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
library(ggplot2)

ggplot(districts, aes(x=PctUpToDate, y=PctChildPoverty, color=EnrolledLog)) +
  geom_point(alpha=1) +
  scale_color_gradient(low="blue", high="red") +
  theme_minimal()
```



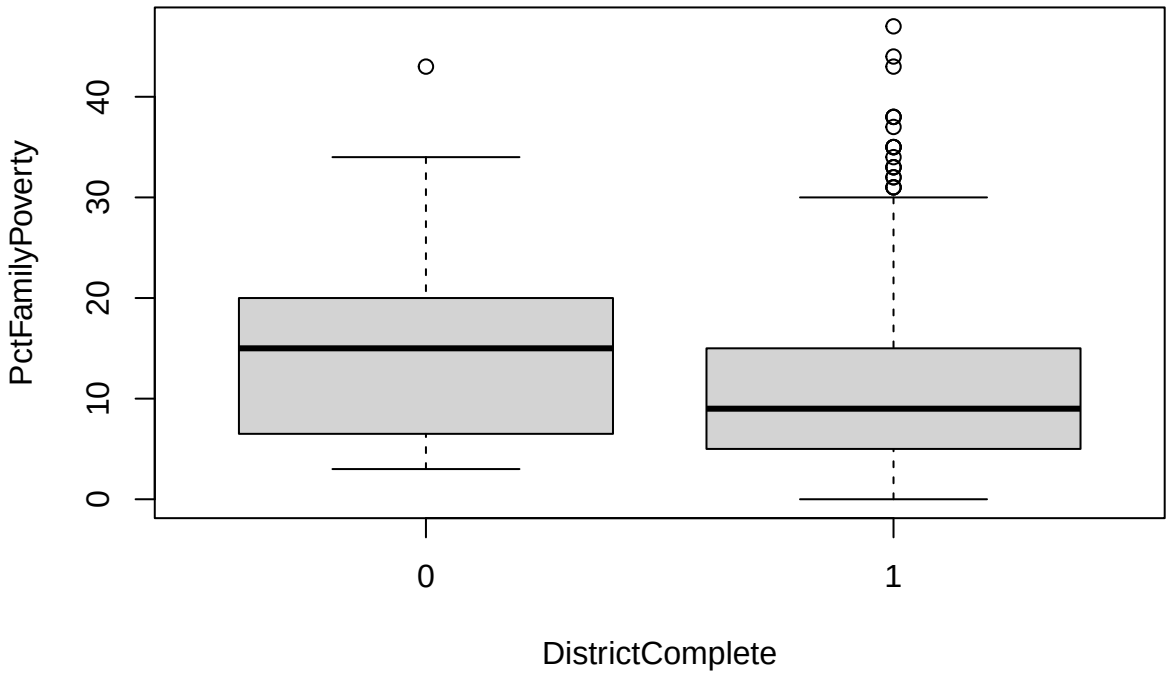
d. ***Which, if any, of the four predictor variables predict whether or not a district's reporting was complete?***

```
districts$DistrictComplete <- as.numeric(districts$DistrictComplete)

districtsQ5D <- districts %>% dplyr::select(c(PctChildPoverty, PctFamilyPoverty, EnrolledLog, TotalSchoo
cor(districtsQ5D)
```

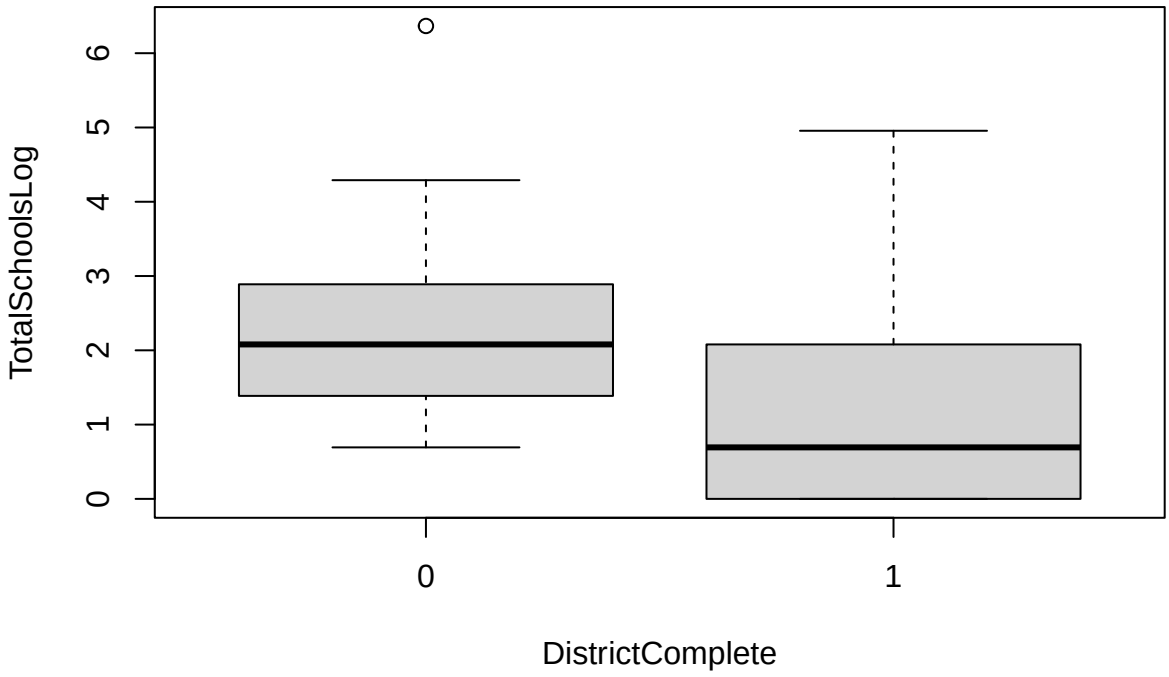
```
##                   PctChildPoverty PctFamilyPoverty EnrolledLog TotalSchoolsLog
## PctChildPoverty        1.00000000      0.867776828 -0.05301706    -0.087286318
## PctFamilyPoverty       0.86777683      1.000000000  0.05546011    -0.005408707
## EnrolledLog           -0.05301706      0.055460112  1.00000000     0.916319391
## TotalSchoolsLog       -0.08728632     -0.005408707  0.91631939     1.000000000
## DistrictComplete      -0.07383085     -0.098544081 -0.14050086    -0.229678279
##                   DistrictComplete
## PctChildPoverty       -0.07383085
## PctFamilyPoverty      -0.09854408
## EnrolledLog           -0.14050086
## TotalSchoolsLog       -0.22967828
## DistrictComplete       1.00000000
```

```
modelGlm <- glm(DistrictComplete ~ PctChildPoverty +PctFamilyPoverty +
                  EnrolledLog +TotalSchoolsLog, data = districts,
                family = binomial())
```

```r
boxplot(PctFamilyPoverty ~ DistrictComplete, data = districts)
```




```r
boxplot(TotalSchoolsLog ~ DistrictComplete, data = districts)
```



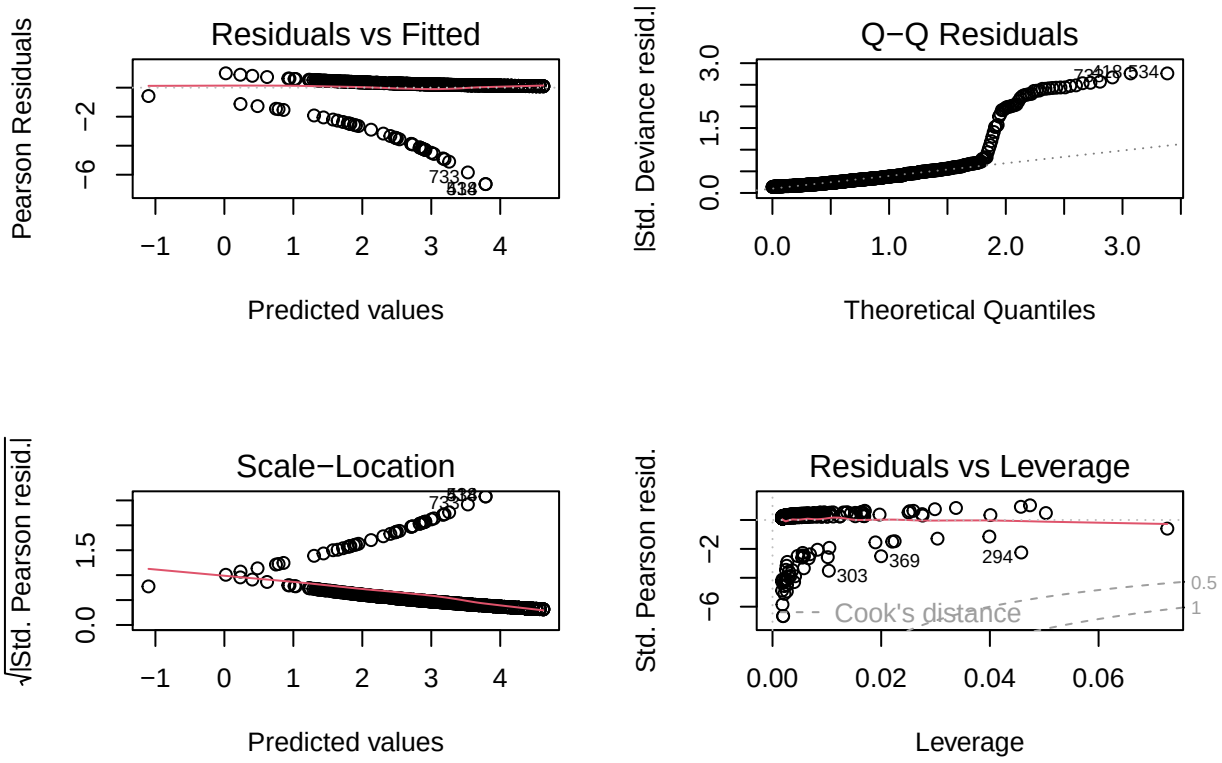

```r
library(performance)
library(see)
library(car)
library(DHARMa)
vif(modelGlm)
```

```
##  PctChildPoverty PctFamilyPoverty      EnrolledLog  TotalSchoolsLog
##         4.548039         4.579030        15.578199        15.511391
```

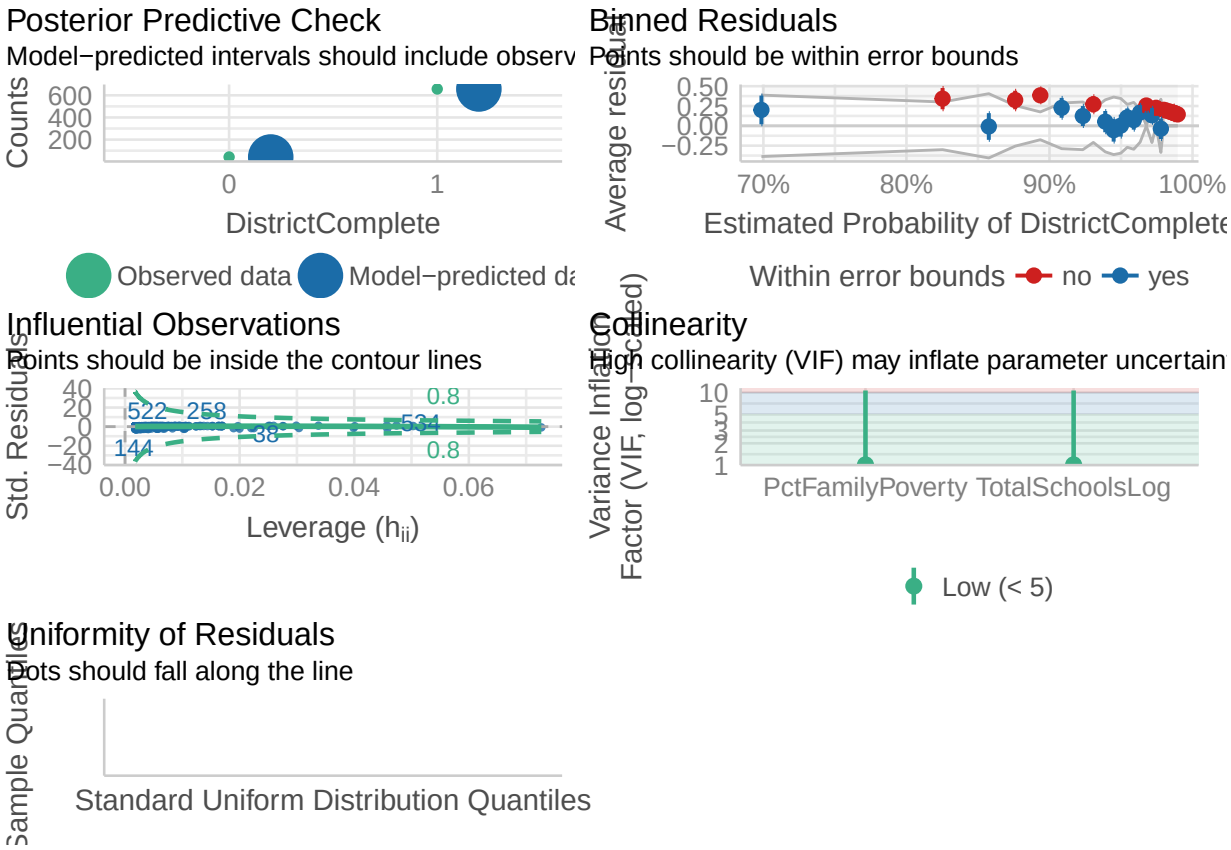
```
modelGlm <- glm(DistrictComplete ~ PctFamilyPoverty +TotalSchoolsLog, data = districts,
                family = binomial())
vif(modelGlm)
```

```
## PctFamilyPoverty  TotalSchoolsLog
##         1.011096         1.011096
```

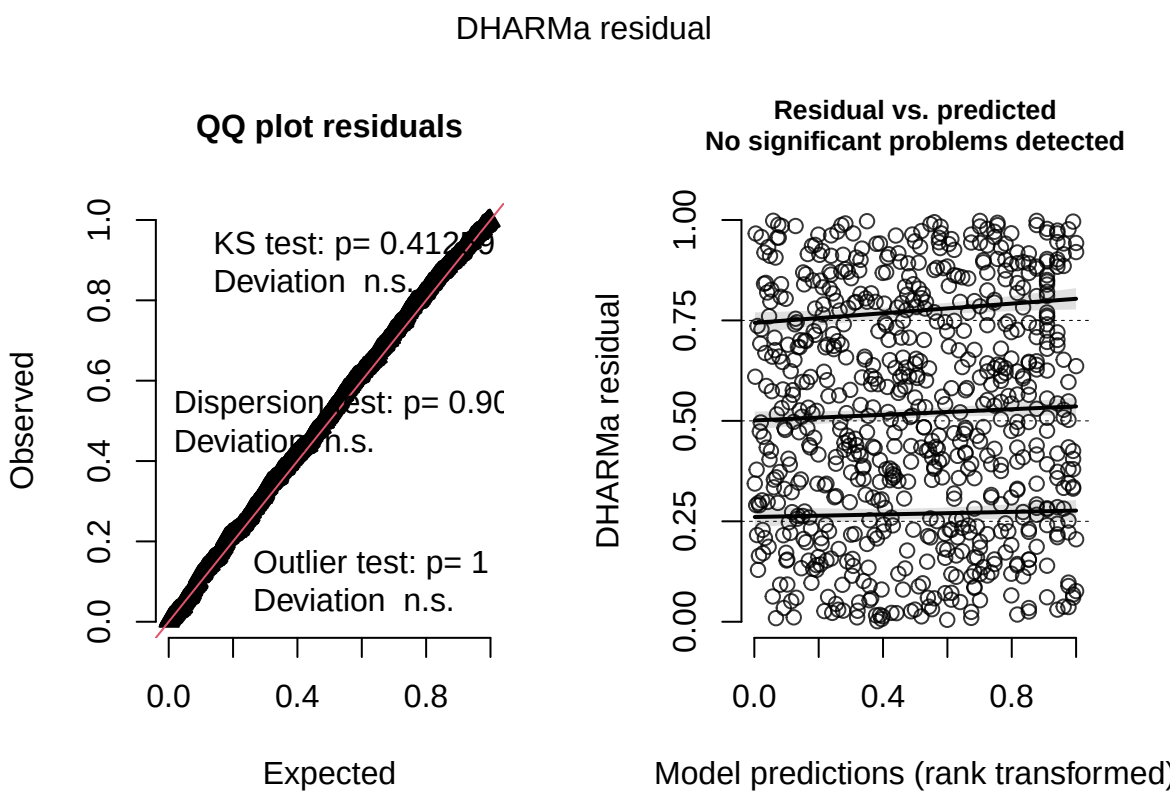
```
par(mfrow=c(2,2))
plot(modelGlm)
```



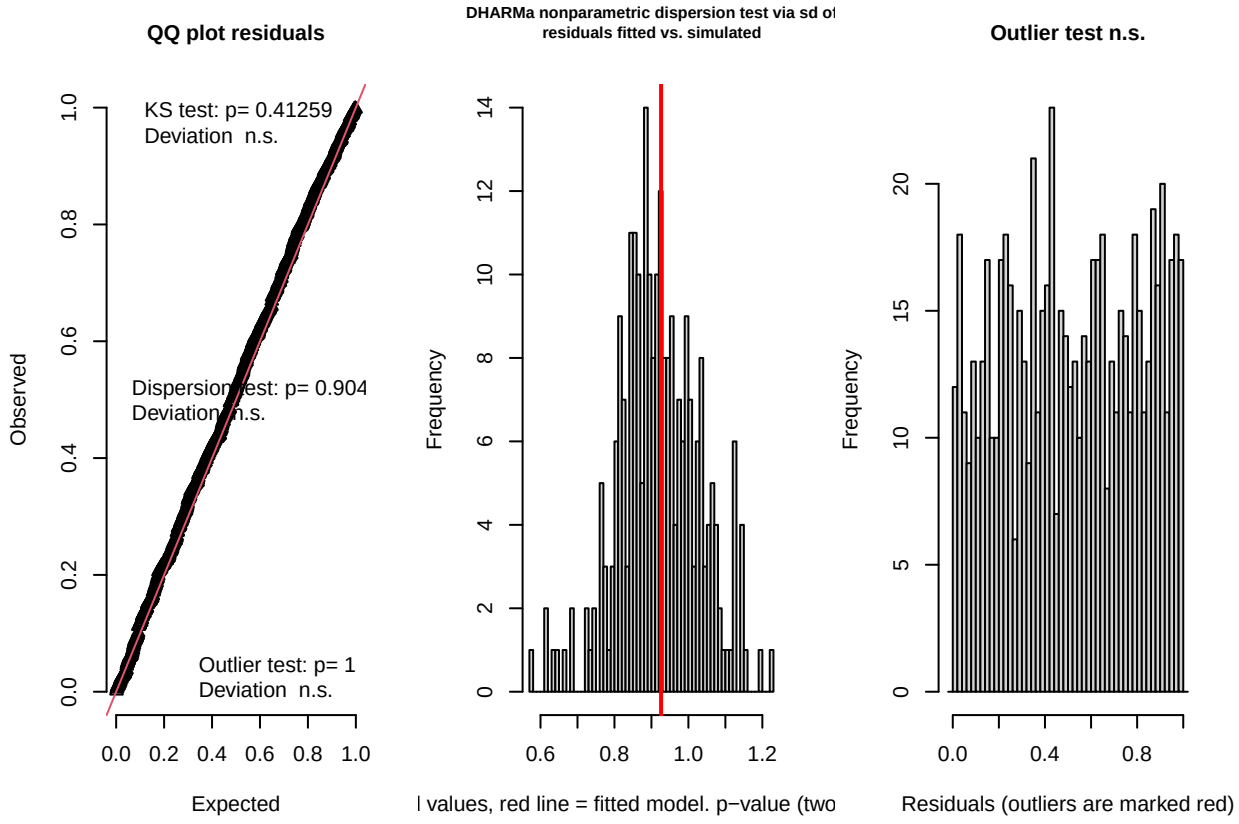

```
check_model(modelGlm)
```




```
simulationOutput <- simulateResiduals(fittedModel = modelGlm, n=250)
plot(simulationOutput)
```




```
testResiduals(simulationOutput)
```



```
## $uniformity
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.033481, p-value = 0.4126
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.0061, p-value = 0.904
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 5, observations = 700, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.002323215 0.016589774
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                           0.007142857
```

```
## $uniformity
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.033481, p-value = 0.4126
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.0061, p-value = 0.904
## alternative hypothesis: two.sided
##
##
## $outliers
##
```

```
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 5, observations = 700, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.002323215 0.016589774
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                           0.007142857
```

There's evidence of a curvilinear relationship.

While the residuals show outliers and a heavy-tailed distribution, the model still provides a potentially useful approximation of the underlying relationship. Given the current analysis context, we will proceed with the model while acknowledging these limitations.

```
summary(modelGlm)
```

```
##
## Call:
## glm(formula = DistrictComplete ~ PctFamilyPoverty + TotalSchoolsLog,
##     family = binomial(), data = districts)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       4.62448    0.43739  10.573  < 2e-16 ***
## PctFamilyPoverty -0.05144    0.01883  -2.732  0.00629 **
## TotalSchoolsLog  -0.76190    0.13675  -5.571 2.53e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 323.23  on 699  degrees of freedom
## Residual deviance: 282.74  on 697  degrees of freedom
## AIC: 288.74
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coef(modelGlm))
```

```
##      (Intercept) PctFamilyPoverty  TotalSchoolsLog
##      101.9494513        0.9498618        0.4667796
```

```
exp(confint(modelGlm))
```

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

```
##                      2.5 %      97.5 %
## (Intercept)      45.4890013 254.3432935
## PctFamilyPoverty  0.9161027   0.9867404
## TotalSchoolsLog   0.3531812   0.6054821
```

```r
anova(modelGlm,test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: DistrictComplete
##
## Terms added sequentially (first to last)
##
##
##                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                699     323.23
## PctFamilyPoverty  1    6.024       698     317.21   0.01411 *
## TotalSchoolsLog   1   34.465       697     282.74 4.341e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model_performance(modelGlm)
```

```
## # Indices of model performance
##
## AIC     |    AICc |     BIC | Tjur's R2 |  RMSE | Sigma | Log_loss | Score_log | Score_spherical |
## ------------------------------------------------------------------------------------------------------
## 288.744 | 288.778 | 302.397 |     0.078 | 0.231 | 1.000 |    0.202 |      -Inf |           0.001 | 0
```

```r
library(caret)
```

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

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

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

```r
predicted <- round(predict(modelGlm,type="response"))
confusion <- table(predicted,districts$DistrictComplete)
confusionMatrix(confusion,positive="1")
```

```
## Confusion Matrix and Statistics
##
##
## predicted   0   1
##         0   1   0
##         1  42 657
##
##                Accuracy : 0.94
##                  95% CI : (0.9198, 0.9564)
```

```
##     No Information Rate : 0.9386
##     P-Value [Acc > NIR] : 0.4778
##
##                   Kappa : 0.0428
##
##  Mcnemar's Test P-Value : 2.509e-10
##
##             Sensitivity : 1.00000
##             Specificity : 0.02326
##          Pos Pred Value : 0.93991
##          Neg Pred Value : 1.00000
##              Prevalence : 0.93857
##          Detection Rate : 0.93857
##    Detection Prevalence : 0.99857
##       Balanced Accuracy : 0.51163
##
##        'Positive' Class : 1
##
```

Even though the accuracy was 0.94, the specificity was very low as 0.02325. This is due to the imbalance in the dataset, where there are far more observations with complete district, while very few are not complete. The no-information rate is 0.9386. The model tend to predict all the Not complete districts as complete to achive a seemly high accuracy.

```
library(MCMCpack)
```

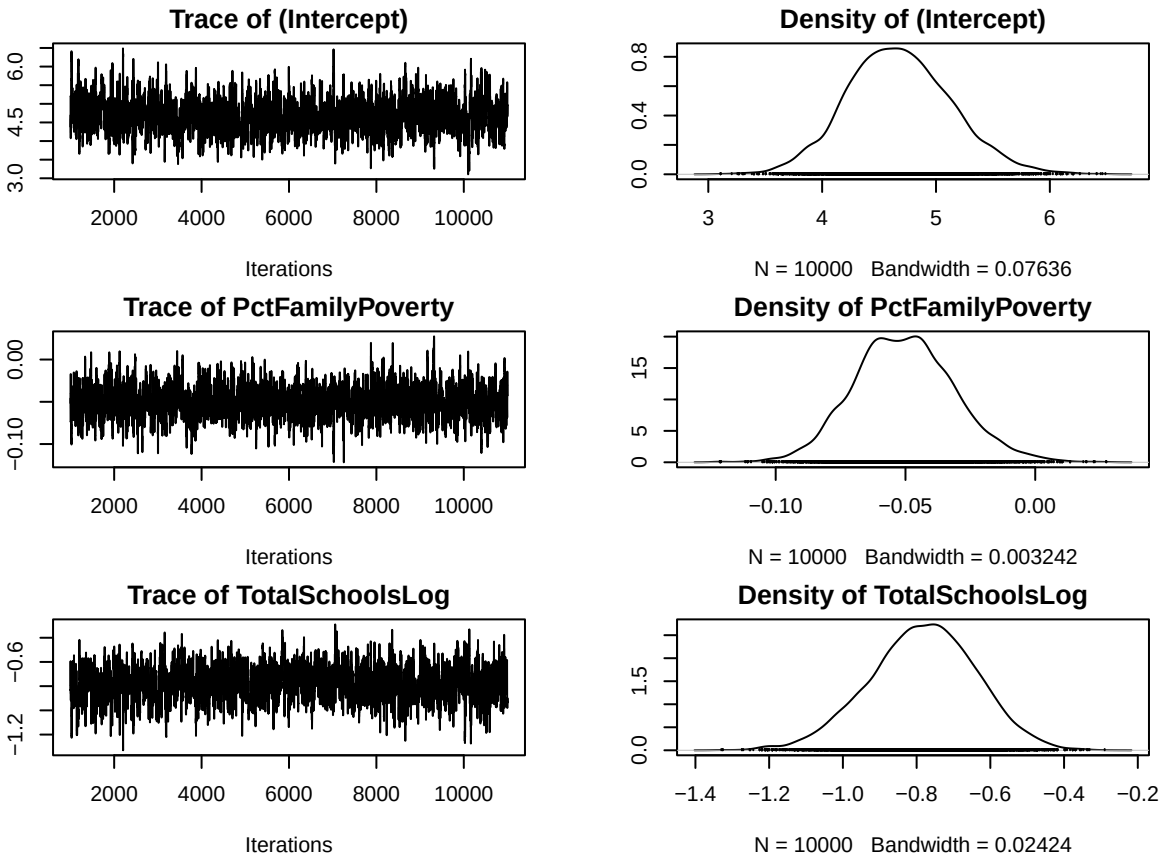
```
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)

## ## Copyright (C) 2003-2024 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)
## ##
```

```
bayesLogitOut <- MCMClogit(formula = DistrictComplete ~ PctFamilyPoverty +TotalSchoolsLog, data = distri
summary(bayesLogitOut)
```

```
##
## 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)       4.67889 0.45776 0.0045776      0.0155149
```

```
## PctFamilyPoverty -0.05047 0.01938 0.0001938      0.0006682
## TotalSchoolsLog  -0.78384 0.14430 0.0014430      0.0050834
##
## 2. Quantiles for each variable:
##
##                      2.5%      25%      50%      75%    97.5%
## (Intercept)       3.82702  4.36440  4.66244  4.97346  5.62477
## PctFamilyPoverty -0.08666 -0.06363 -0.05074 -0.03777 -0.01087
## TotalSchoolsLog  -1.07598 -0.87722 -0.77811 -0.68305 -0.51209
```

```
par(mar = c(4, 4, 2, 2))

plot(bayesLogitOut)
```




```
exp(mean(bayesLogitOut[,"PctFamilyPoverty"]))
```

```
## [1] 0.9507794
```

```
exp(quantile(bayesLogitOut[,"PctFamilyPoverty"],c(0.025)))
```

```
##      2.5%
## 0.9169885
```

```r
exp(quantile(bayesLogitOut[,"PctFamilyPoverty"],c(0.975)))
```

```
##     97.5%
## 0.9891853
```

```r
exp(mean(bayesLogitOut[,"TotalSchoolsLog"]))
```

```
## [1] 0.4566508
```

```r
exp(quantile(bayesLogitOut[,"TotalSchoolsLog"],c(0.025)))
```

```
##      2.5%
## 0.3409621
```

```r
exp(quantile(bayesLogitOut[,"TotalSchoolsLog"],c(0.975)))
```

```
##     97.5%
## 0.5992415
```

6. Concluding Paragraph

```r
identify_outliers <- function(data) {
  quartiles <- quantile(data, probs=c(.25, .75))
  iqr <- IQR(data)
  lower_bound <- quartiles[1] - 1.5 * iqr
  upper_bound <- quartiles[2] + 1.5 * iqr
  return(data < lower_bound | data > upper_bound)
}

outliers_pctUpToDate <- districts$DistrictName[identify_outliers(districts$PctUpToDate)]
outliers_withMMR <- districts$DistrictName[identify_outliers(districts$WithMMR)]
outliers_withPolio <- districts$DistrictName[identify_outliers(districts$WithPolio)]
outliers_withHepB <- districts$DistrictName[identify_outliers(districts$WithHepB)]
outliers_withDTP <- districts$DistrictName[identify_outliers(districts$WithDTP)]

all_outliers <- unique(c(outliers_pctUpToDate, outliers_withMMR, outliers_withPolio, outliers_withHepB,
print(all_outliers)
```

```
##  [1] Happy Valley Elementary
##  [2] Shaffer Union Elementary
##  [3] Lucerne Valley Unified
##  [4] Burnt Ranch Elementary
##  [5] Cinnabar Elementary
##  [6] Bolinas-Stinson Union
##  [7] Dunsmuir Elementary
##  [8] Shasta Union Elementary
##  [9] Lagunitas Elementary
## [10] Alpine Union Elementary
```

```
## [11] Twin Ridges Elementary
## [12] Wilmar Union Elementary
## [13] Fieldbrook Elementary
## [14] Bonny Doon Union Elementary
## [15] Camino Union Elementary
## [16] Monte Rio Union Elementary
## [17] San Lorenzo Valley Unified
## [18] Hughes-Elizabeth Lakes Union Elementary
## [19] Alta-Dutch Flat Union Elementary
## [20] Sebastopol Union Elementary
## [21] Mesa Union Elementary
## [22] William S. Hart Union High
## [23] Mendocino Unified
## [24] Blue Lake Union Elementary
## [25] Laytonville Unified
## [26] Happy Camp Union Elementary
## [27] Trinidad Union Elementary
## [28] Gorman Elementary
## [29] Fort Bragg Unified
## [30] Harmony Union Elementary
## [31] Arcata Elementary
## [32] Spencer Valley Elementary
## [33] Marcum-Illinois Union Elementary
## [34] Larkspur-Corte Madera
## [35] Needles Unified
## [36] Emery Unified
## [37] Freshwater Elementary
## [38] Southern Humboldt Joint Unified
## [39] Camptonville Elementary
## [40] Warner Unified
## [41] Julian Union Elementary
## [42] Big Sur Unified
## [43] Dehesa Elementary
## [44] Curtis Creek Elementary
## [45] Mountain Empire Unified
## [46] Blochman Union Elementary
## [47] Gold Oak Union Elementary
## [48] Gravenstein Union Elementary
## [49] Mt. Baldy Joint Elementary
## [50] Hickman Community Charter
## [51] Fall River Joint Unified
## [52] Big Springs Union Elementary
## [53] North Cow Creek Elementary
## [54] Westwood Unified
## [55] Montecito Union Elementary
## [56] Waterford Unified
## [57] Cayucos Elementary
## [58] Twin Hills Union Elementary
## [59] Summerville Union High
## [60] Petaluma Joint Union High
## [61] Big Oak Flat-Groveland Unified
## [62] Cold Spring Elementary
## [63] Pacific Elementary
## [64] South Bay Union Elementary
```

```
## [65] Fort Sage Unified
## 846 Levels: ABC Unified Ackerman Charter ... Yucaipa-Calimesa Joint Unified
```