Final Exam

Maya Mileva

due date: Dec 21st, 2019

Task

Your goal for this final exam is to conduct the necessary analyses and then write up a technical report for a scientifically knowledgeable staff member in a state legislator's office. Thus, you should provide sufficient numeric and graphical detail that the staff member can create a comprehensive briefing for a legislator. You can assume that the staff member understands the concept of statistical significance. Your report should include a few graphics created by R, keeping in mind that you must provide some accompanying text to explain each graphic that you include in your report.

These three data sets all pertain to vaccinations. The first and second datasets are the same for everyone and are mainly included to provide context for interpretation of the results. Most of the substantive analyses occur in reference to the third dataset. This third dataset is different for every student and results will vary depending upon the sample the student received.

Datasets

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

DTP1 = First dose of Diphtheria/Pertussis/Tetanus vaccine;

HepB_BD = Hepatitis B, Birth Dose;

Pol3 = Polio third dose;

Hib3 – Influenza third dose;

MCV1 = Measles first dose

```
## Load the data
load.Rdata(filename="usVaccines.RData", "usVaccines" )

## Explore the usVaccines.Rdata dataset
head(usVaccines)
```

DTP1	HepB_BD	Pol3	Hib3	MCV1
83	16	95	85	86
84	16	96	85	97
83	17	97	84	97
84	17	97	83	98
84	16	97	85	98
85	17	96	85	97

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

```
any(is.na(usVaccines))# no missing values
```

```
## [1] FALSE
```

2. allSchoolsReportStatus.RData – A list of California kindergartens and whether they reported vaccination data to the state in 2013

```
## Load the data
load.Rdata("allSchoolsReportStatus.RData", "allSchoolsReportStatus")

## Explore the allSchoolsReportStatus dataset
head(allSchoolsReportStatus, n=3)
```

name	pubpriv	reported
AGUA DULCE ELEMENTARY	PUBLIC	Y
MEADOWLARK ELEMENTARY	PUBLIC	Y
CALIFORNIA SCHOOL FOR THE DEAF-FREMONT	PUBLIC	Y

```
str(allSchoolsReportStatus)
```

```
## 'data.frame':    7381 obs. of  3 variables:
##  $ name    : chr  "AGUA DULCE ELEMENTARY" "MEADOWLARK ELEMENTARY" "CALIFORNIA SCHOOL FOR THE DEAF-FR
##  $ pubpriv : chr  "PUBLIC" "PUBLIC" "PUBLIC" "PUBLIC" ...
##  $ reported: chr  "Y" "Y" "Y" "Y" ...
```

```
summary(allSchoolsReportStatus)
```

name	pubpriv	reported
Length:7381	Length:7381	Length:7381
Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character

```r
any(is.na(allSchoolsReportStatus))# no missing values
```

```
## [1] FALSE
```

3. districtsX.RData – (Where X is the number of your particular dataset) A sample of California public school districts from the 2013 data collection, along with specific numbers and percentages for each district.

$ DistrictName : Name of the district

$ WithoutDTP : Percentage of students without the DTP vaccine

$ WithoutPolio : Percentage of students without the Polio vaccine

$ WithoutMMR : Percentage of students without the MMR vaccine

$ WithoutHepB : Percentage of students without the Hepatitis B vaccine

$ PctUpToDate : Percentage of all enrolled students with completely up-to-date vaccines

$ DistrictComplete: Boolean indicating whether or not the district's reporting was complete

$ PctBeliefExempt : Percentage of all enrolled students with belief exceptions

$ PctChildPoverty : Percentage of children in the district living below the poverty line

$ PctFreeMeal : Percentage of children in the district eligible for free student meals

$ PctFamilyPoverty: num Percentage of families in the district living below the poverty line

$ Enrolled : Total number of enrolled students in the district

$ TotalSchools : Total number of different schools in the district

```r
## Load the data
load.Rdata("districts12.RData", "districts12")

## Explore the allSchoolsReportStatus dataset
head(districts12, n=3)
```

	DistrictName	WithoutDTP	WithoutPolio	WithoutMMR	WithoutHepB	PctUpToDate	Dist
118	Lowell Joint	6	5	6	5	94	TRU
413	Chowchilla Elementary	2	3	3	2	96	TRU
610	Harmony Union Elementary	38	38	38	39	61	TRU

```r
str(districts12)
```

```
## 'data.frame':    700 obs. of  13 variables:
##  $ DistrictName    : Factor w/ 846 levels "ABC Unified",..: 406 137 288 306 81 455 373 210 133 677 .
##  $ WithoutDTP      : num  6 2 38 4 6 0 33 24 25 4 ...
##  $ WithoutPolio    : num  5 3 38 3 5 0 40 22 24 4 ...
##  $ WithoutMMR      : num  6 3 38 4 6 0 30 24 24 4 ...
##  $ WithoutHepB     : num  5 2 39 1 5 0 30 15 20 4 ...
##  $ PctUpToDate     : num  94 96 61 95 93 100 53 72 75 96 ...
##  $ DistrictComplete: logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ PctBeliefExempt : num  4 1 38 0 4 0 23 11 17 3 ...
##  $ PctChildPoverty : num  13 25 11 25 9 55 22 22 37 11 ...
##  $ PctFreeMeal     : num  28 69 26 66 22 77 68 69 50 33 ...
```

```
##  $ PctFamilyPoverty: num  6 15 2 12 5 16 8 4 9 7 ...
##  $ Enrolled        : num  341 271 64 1020 871 33 30 46 59 696 ...
##  $ TotalSchools    : num  5 1 2 9 8 1 1 1 2 9 ...
```

```
summary(districts12)
```

DistrictName	WithoutDTP	WithoutPolio	WithoutMMR	WithoutHepB	PctUpToDate	Dis
ABC Unified : 1	Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.000	Min. : 24.00	Mo
Ackerman Charter : 1	1st Qu.: 3.00	1st Qu.: 3.00	1st Qu.: 3.00	1st Qu.: 2.000	1st Qu.: 84.00	FA
Acton-Agua Dulce Unified: 1	Median : 7.00	Median : 6.00	Median : 6.00	Median : 5.000	Median : 92.00	TR
Adelanto Elementary : 1	Mean :10.22	Mean : 9.82	Mean :10.27	Mean : 7.851	Mean : 87.76	NA
Alameda Unified : 1	3rd Qu.:14.00	3rd Qu.:13.00	3rd Qu.:14.00	3rd Qu.:10.000	3rd Qu.: 96.00	NA
Albany City Unified : 1	Max. :72.00	Max. :70.00	Max. :72.00	Max. :69.000	Max. :100.00	NA
(Other) :694	NA	NA	NA	NA	NA	NA

```
any(is.na(districts12))# no missing values
```

```
## [1] FALSE
```

Introductory/Descriptive Reports:

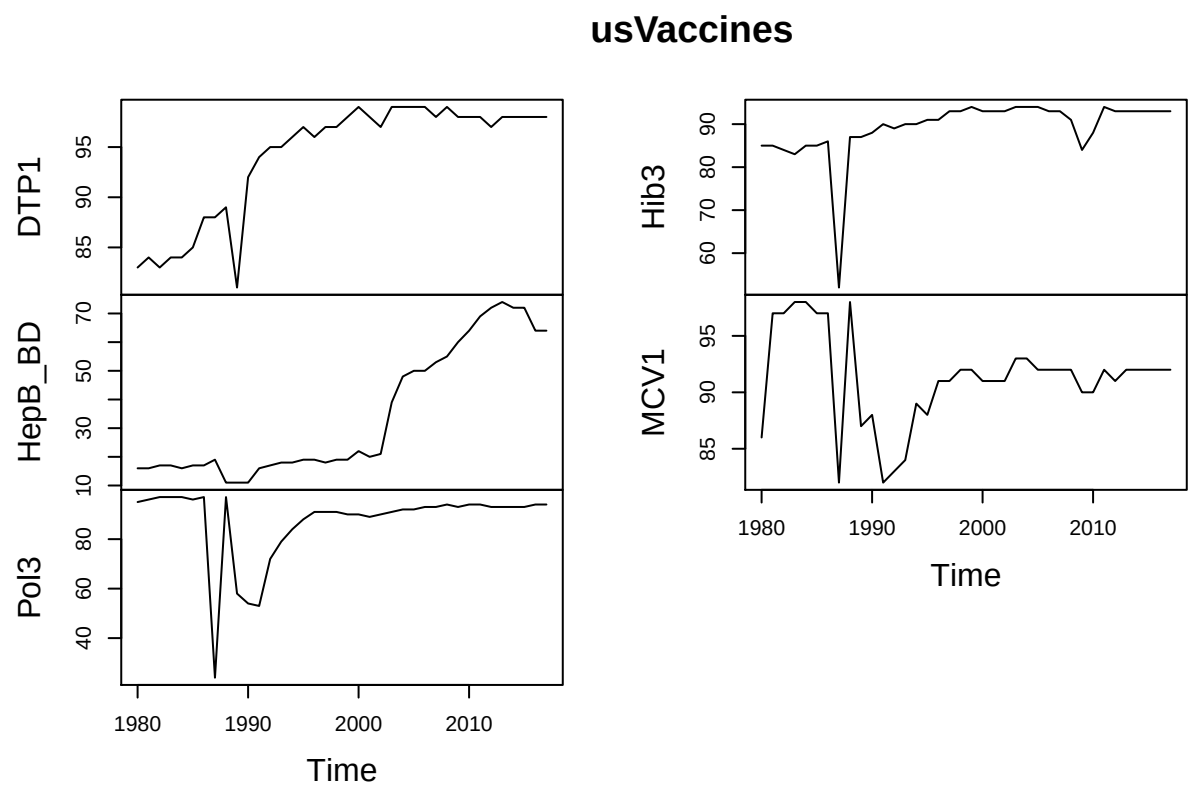
1. How have U.S. vaccination rates varied over time? Are vaccination rates increasing or decreasing? Which vaccination has the highest rate at the conclusion of the time series? Which vaccination has the lowest rate at the conclusion of the time series? Which vaccine has the greatest volatility?

- use usVaccines data

```
## Correlate the raw data
cor(usVaccines)
```

	DTP1	HepB_BD	Pol3	Hib3	MCV1
DTP1	1.0000000	0.5905157	0.1730162	0.5593254	-0.2095028
HepB_BD	0.5905157	1.0000000	0.3255994	0.3282156	0.0819240
Pol3	0.1730162	0.3255994	1.0000000	0.6025809	0.7308557
Hib3	0.5593254	0.3282156	0.6025809	1.0000000	0.2108128
MCV1	-0.2095028	0.0819240	0.7308557	0.2108128	1.0000000

```
## Plot the data
plot(usVaccines)
```




```
str(usVaccines) # Structure confirms time series
```

```
##  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" ...
```

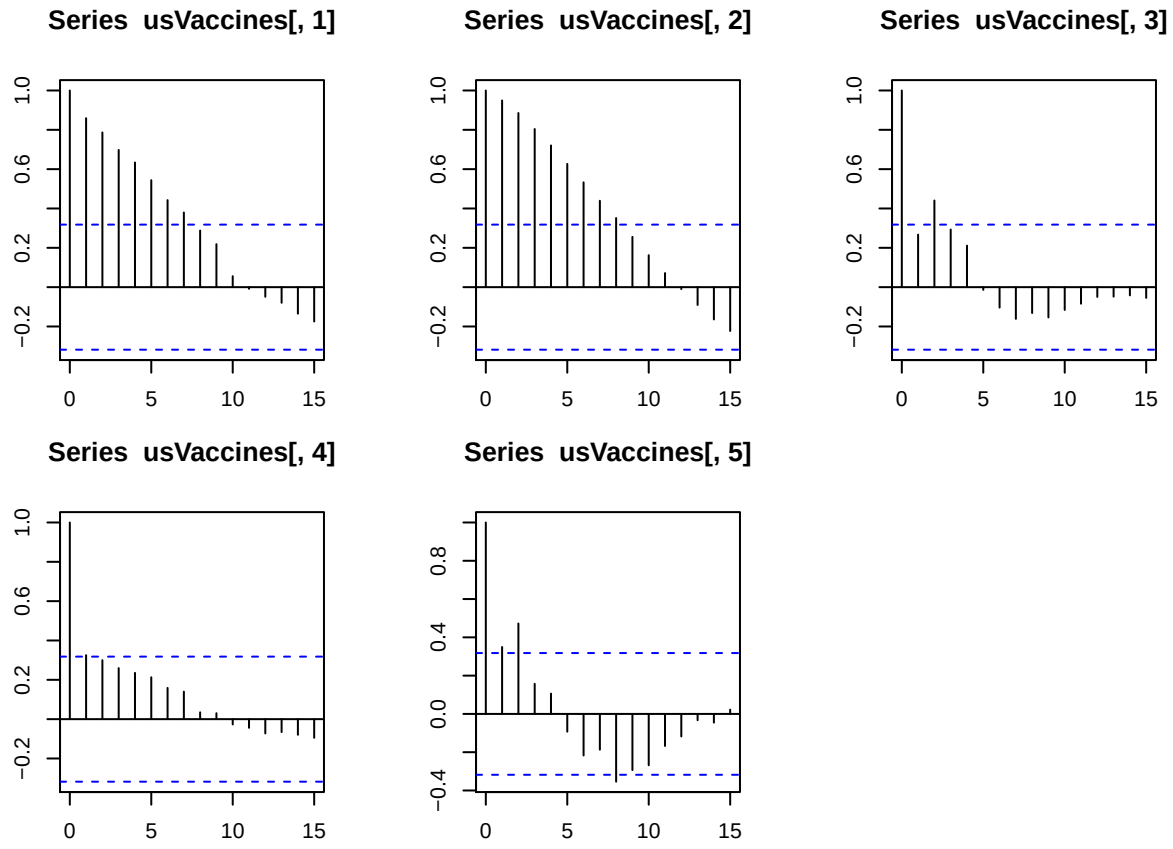
```
# plot(usVaccines[,1])
# plot(usVaccines[,2])
# plot(usVaccines[,3])
# plot(usVaccines[,4])
# plot(usVaccines[,5])

class(usVaccines)
```

```
## [1] "mts"    "ts"     "matrix"
```

```
cycle(usVaccines)
```

```
## Time Series:
## Start = 1980
## End = 2017
## Frequency = 1
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```r
#boxplot(usVaccines~cycle(usVaccines))

par(mar = c(3,3,3,3))
par(mfrow =c(2,3))
## Examine the auto-correlation function (ACF)
# acf(usVaccines)
acf(usVaccines[,1])
acf(usVaccines[,2])
acf(usVaccines[,3])
acf(usVaccines[,4])
acf(usVaccines[,5])
```




```r
adf.test(usVaccines[,1]) # not stationary
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, 1]
## Dickey-Fuller = -0.87963, Lag order = 3, p-value = 0.943
## alternative hypothesis: stationary
```

```r
adf.test(usVaccines[,2]) # not stationary
```

```
##
```

```
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, 2]
## Dickey-Fuller = -1.9729, Lag order = 3, p-value = 0.5839
## alternative hypothesis: stationary
```

```
adf.test(usVaccines[,3]) # not stationary
```

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

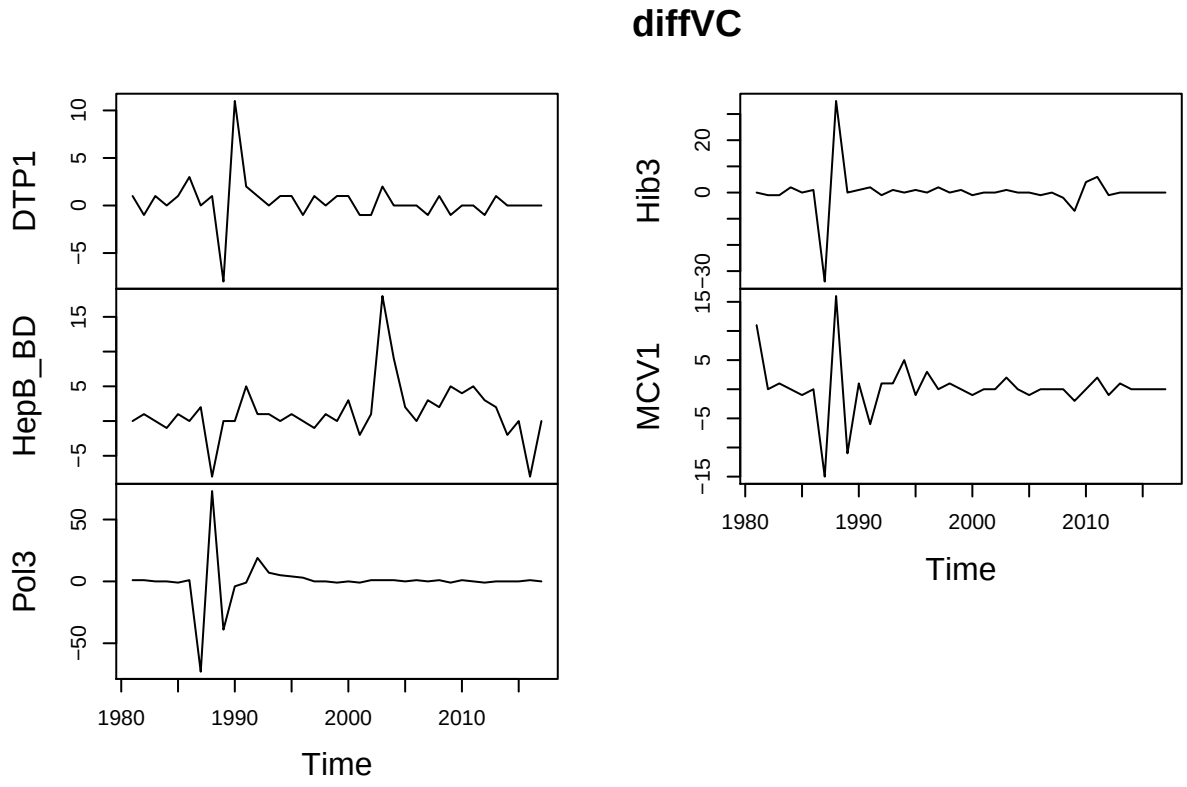
```
adf.test(usVaccines[,4]) # not stationary
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, 4]
## Dickey-Fuller = -2.3377, Lag order = 3, p-value = 0.4414
## alternative hypothesis: stationary
```

```
adf.test(usVaccines[,5]) # not stationary
```

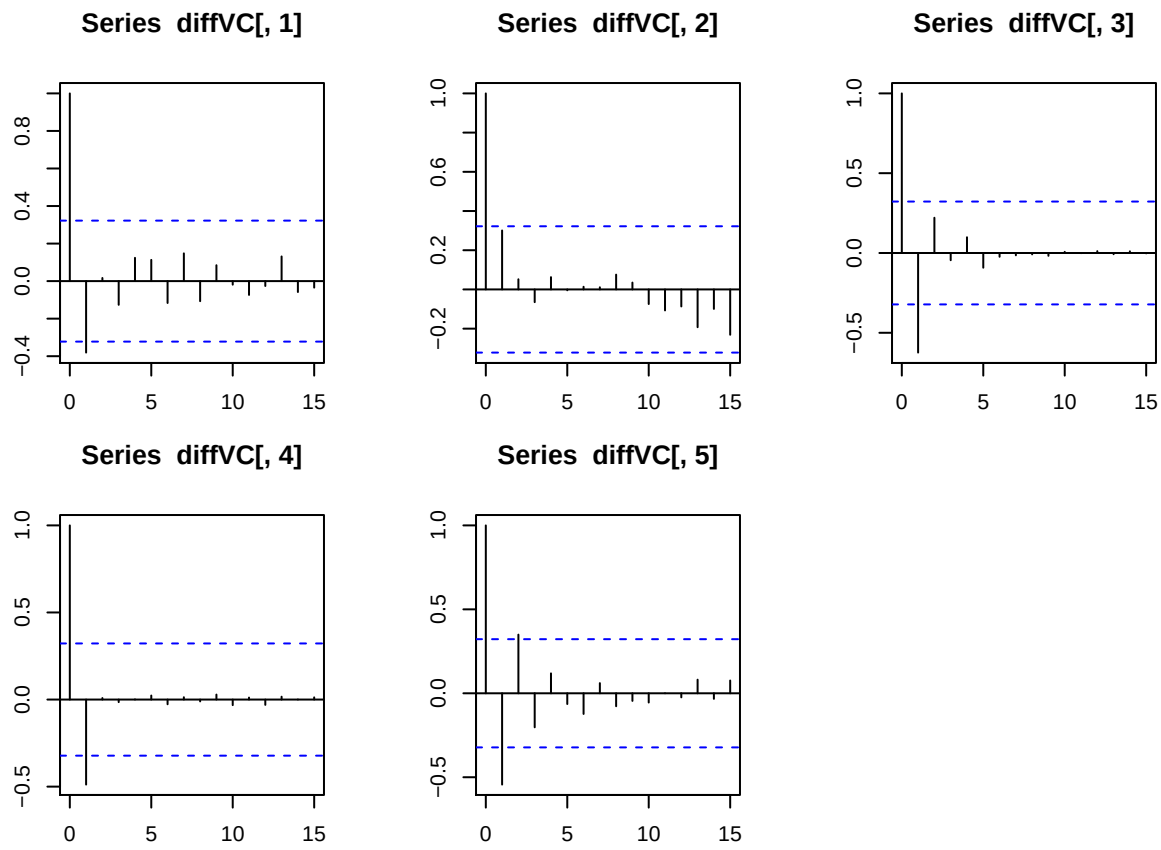
```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, 5]
## Dickey-Fuller = -2.5324, Lag order = 3, p-value = 0.3652
## alternative hypothesis: stationary
```

```
## Create a differenced dataset
diffVC <- diff(usVaccines)
plot(diffVC, type = "l")
```



After differencing, the trend should be removed. In this case is not. Differencing time series means, to subtract each data point in the series from its successor. For most time series patterns, 1 or 2 differencing is necessary to make it stationary series.

```
par(mar = c(3,3,3,3))
par(mfrow = c(2,3))
## Examine the auto-correlation function (ACF)
acf(diffVC[,1]) # seasonalal
acf(diffVC[,2])
acf(diffVC[,3])
acf(diffVC[,4])
acf(diffVC[,5])
```



We can see that the acf of the residuals is centered around 0.

Dickey Fuller test for stationarity allows us to more definitively say, does this particular time series have any trend left in it. A significant DickeyFuller test means that a time series is stationary. We use the test after differencing(in our case) or decomposition to confirm that trend, seasonality, and other time artifacts have been removed.

```
# Run the Augmented Dickey-Fuller test on each time series to remove stationary
adf.test(diffVC[,1])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diffVC[, 1]
## Dickey-Fuller = -4.5333, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

```
adf.test(diffVC[,2])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diffVC[, 2]
## Dickey-Fuller = -1.958, Lag order = 3, p-value = 0.5896
## alternative hypothesis: stationary
```

```r
adf.test(diffVC[,3])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diffVC[, 3]
## Dickey-Fuller = -2.9361, Lag order = 3, p-value = 0.2082
## alternative hypothesis: stationary
```

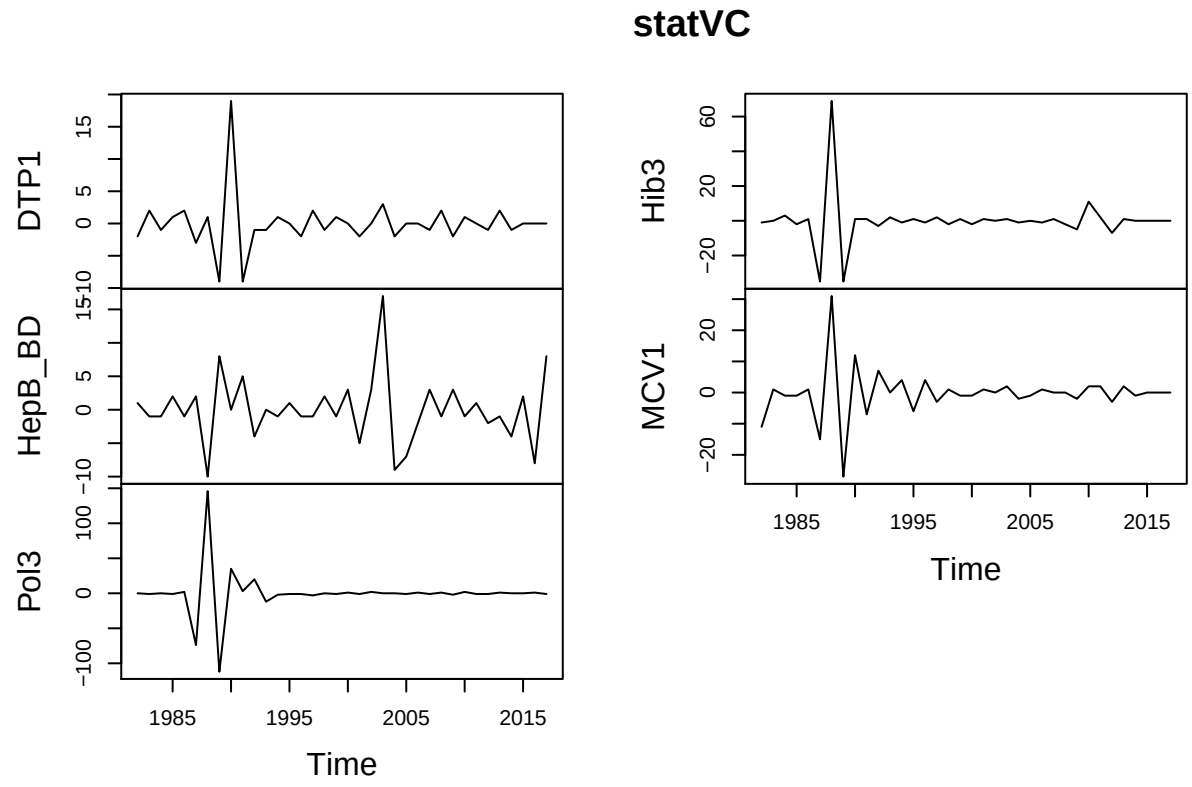
```r
adf.test(diffVC[,4])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diffVC[, 4]
## Dickey-Fuller = -4.3152, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

```r
adf.test(diffVC[,5])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diffVC[, 5]
## Dickey-Fuller = -3.3982, Lag order = 3, p-value = 0.07305
## alternative hypothesis: stationary
```

```r
# ndiffs(diffVC)
statVC <- diff(diffVC)
plot(statVC)
```




```
# Run the Augmented Dickey-Fuller test on each time series to remove stationary
adf.test(statVC[,1])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  statVC[, 1]
## Dickey-Fuller = -6.2933, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

```
adf.test(statVC[,2])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  statVC[, 2]
## Dickey-Fuller = -4.1017, Lag order = 3, p-value = 0.01702
## alternative hypothesis: stationary
```

```
adf.test(statVC[,3])
```

```
##
##  Augmented Dickey-Fuller Test
##
```

```
## data:  statVC[, 3]
## Dickey-Fuller = -5.8893, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

```
adf.test(statVC[,4])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  statVC[, 4]
## Dickey-Fuller = -5.7323, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

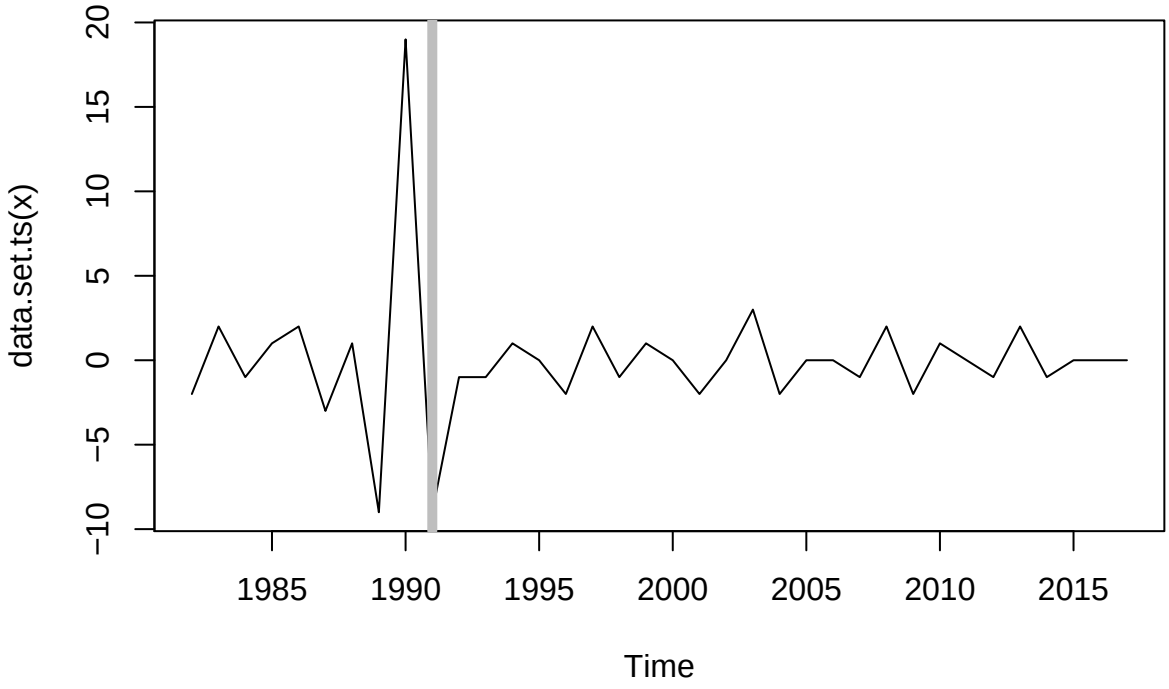
```
adf.test(statVC[,5])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  statVC[, 5]
## Dickey-Fuller = -5.1152, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

The output has a Dickey Fuller value in it for each one of the five. It also shows a lag order that was the limit of the tests that were done and then a p-value(0.01), that is significant, and that means that the time series is stationary. If it's non-significant, then we have concerns that there may be trend in there or possibly other time series artifacts.

We can continue with further analysis of our time series.

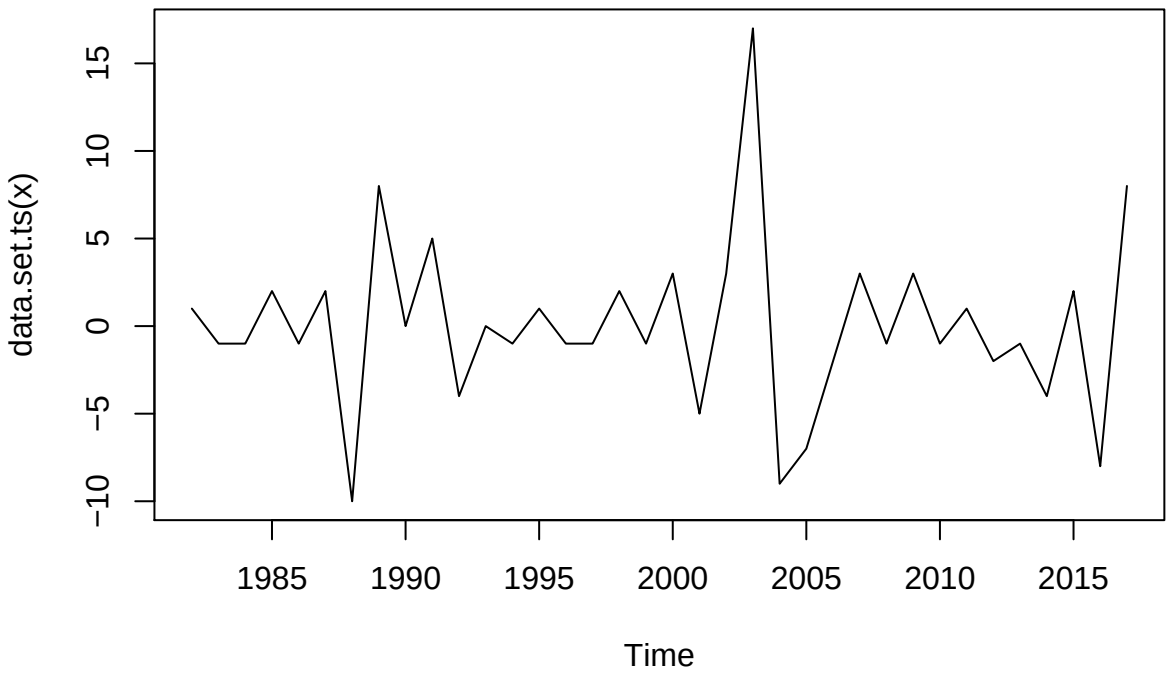
```
diffVCcp1 <- cpt.var(statVC[,1])
plot(diffVCcp1,cpt.col="grey",cpt.width=5)
```




```
diffVCcp1 # change in volatility occurred late in 1991 (DTP1 = First dose of Diphtheria/Pertussis/Tetan
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.75056
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 10
```

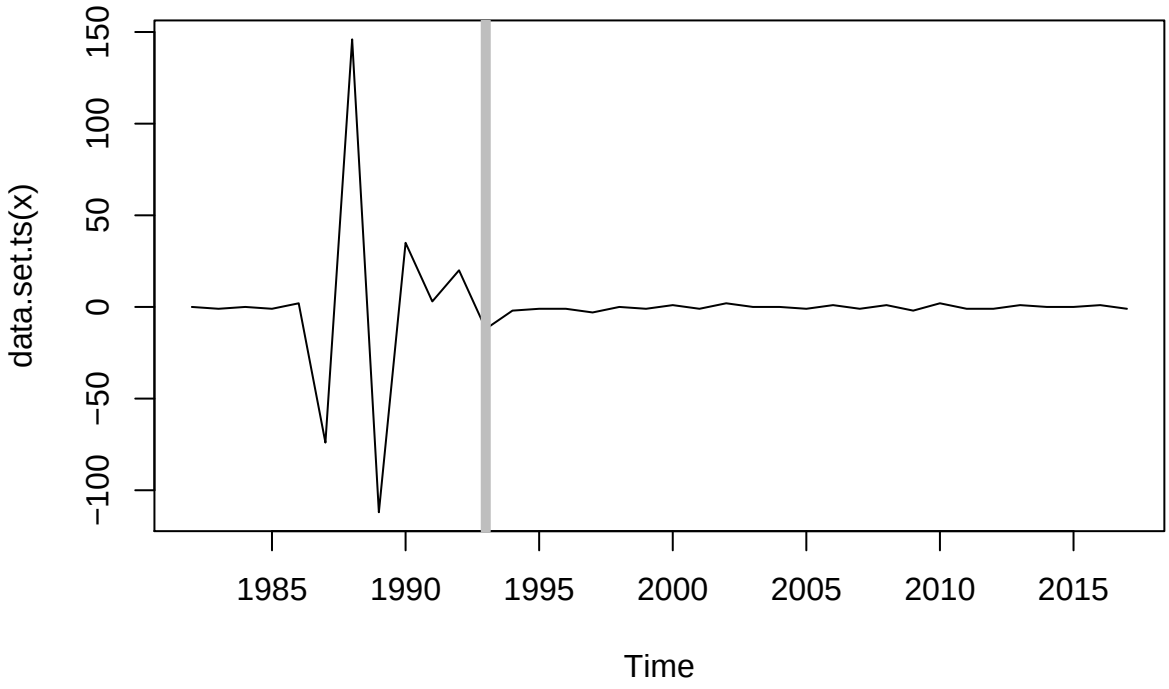
```
diffVCcp2 <- cpt.var(statVC[,2])
plot(diffVCcp2,cpt.col="grey",cpt.width=5)
```




```
diffVCcp2 # no change point detected (HepB_BD = Hepatitis B, Birth Dose)
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.75056
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations :
```

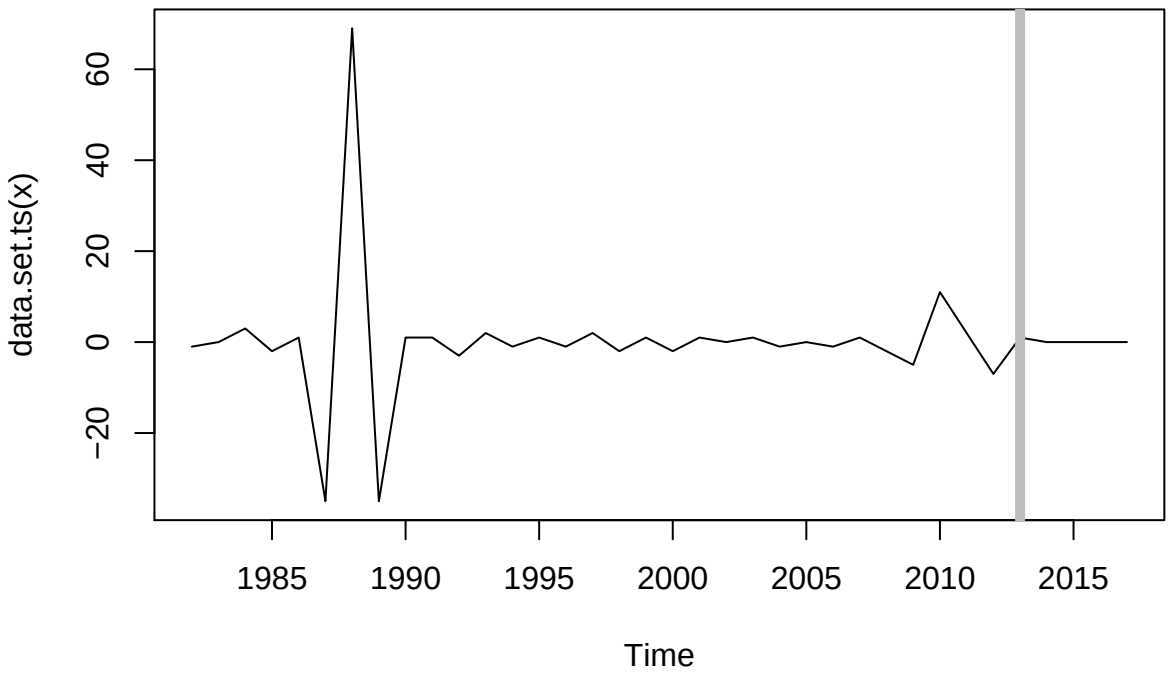
```
diffVCcp3 <- cpt.var(statVC[,3])
plot(diffVCcp3,cpt.col="grey",cpt.width=5)
```




```
diffVCcp3 # change in volatility occurred late in 1993 (Pol3 = Polio third dose)
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.75056
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 12
```

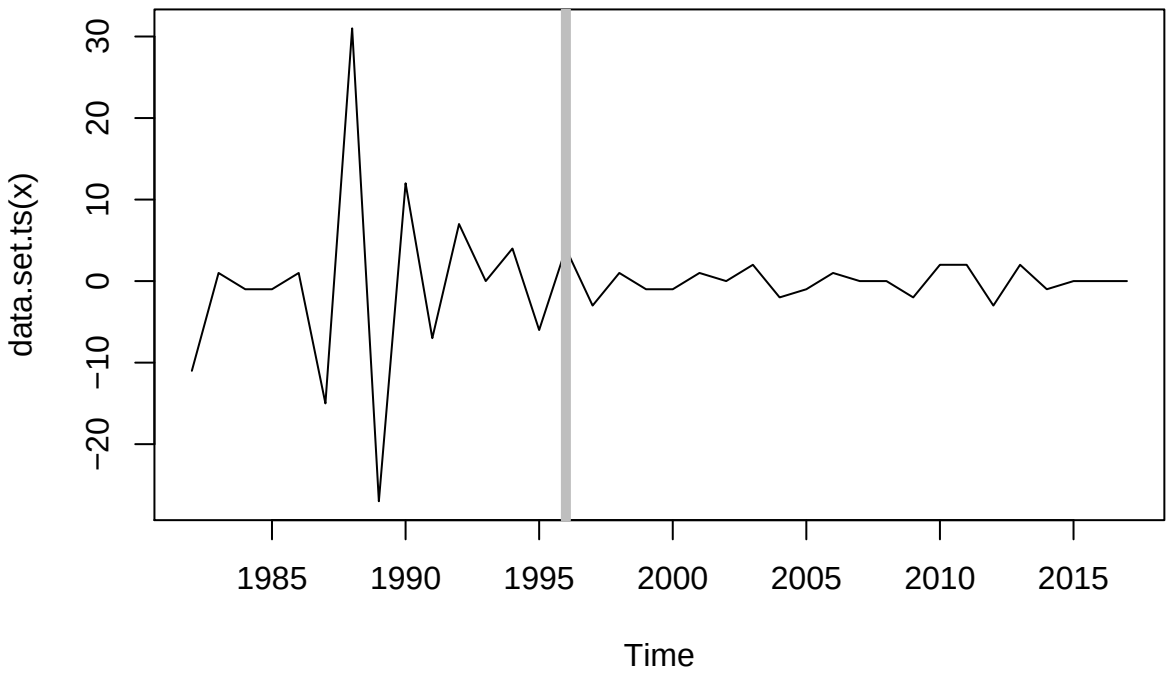
```
diffVCcp4 <- cpt.var(statVC[,4])
plot(diffVCcp4,cpt.col="grey",cpt.width=5)
```




```
diffVCcp4 # change in volatility occurred  in 2013 (Hib3 - Influenza third dose)
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.75056
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 32
```

```
diffVCcp5 <- cpt.var(statVC[,5])
plot(diffVCcp5,cpt.col="grey",cpt.width=5)
```




```
diffVCcp5 # change in volatility occurred late in 1996 (MCV1 = Measles first dose)
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.75056
## Minimum Segment Length : 2
## Maximum no. of cpts   : 1
## Changepoint Locations : 15
```

In a change point analysis, an algorithm searches through the time series data to detect and document major transitions; provides one way of getting an insight into something important that may be happening in a time series. So change point analysis is useful to see if you can detect when some kind of event or intervention has occurred in a time series. This is a very useful way to do some research. We start measuring something at a time when we know not much is going on, then we keep measuring it through a period where there might be important changes happening. When we subject the data to analysis, we can notice whether or

not a meaningful change has occurred and that can be accomplished through a process called change point analysis.

cpt.var() finds the change point in the variability of the differenced time series. Change in volatility occurred at different points for each vaccine.

The major change point for **DTP1** occurs at 10. This indicates that the vaccine rate started to increase in 1991, and also entered a period of more intense volatility (grater variance). The "Type of Penalty" in the output refers to a mathematical formulation that determines how sensitive the algorithm is to detecting changes. MBIC = 10.75056 is like a statistical line, everything that crosses that like need to be examined.

There is no any change point for **HepB_BD**.

The major change point for **Pol3** occurs at 12. This indicates that the vaccine rate started to increase in 1993.

The major change point for **Hib3** occurs at 32. This indicates that the vaccine rate started to increase in 2013.

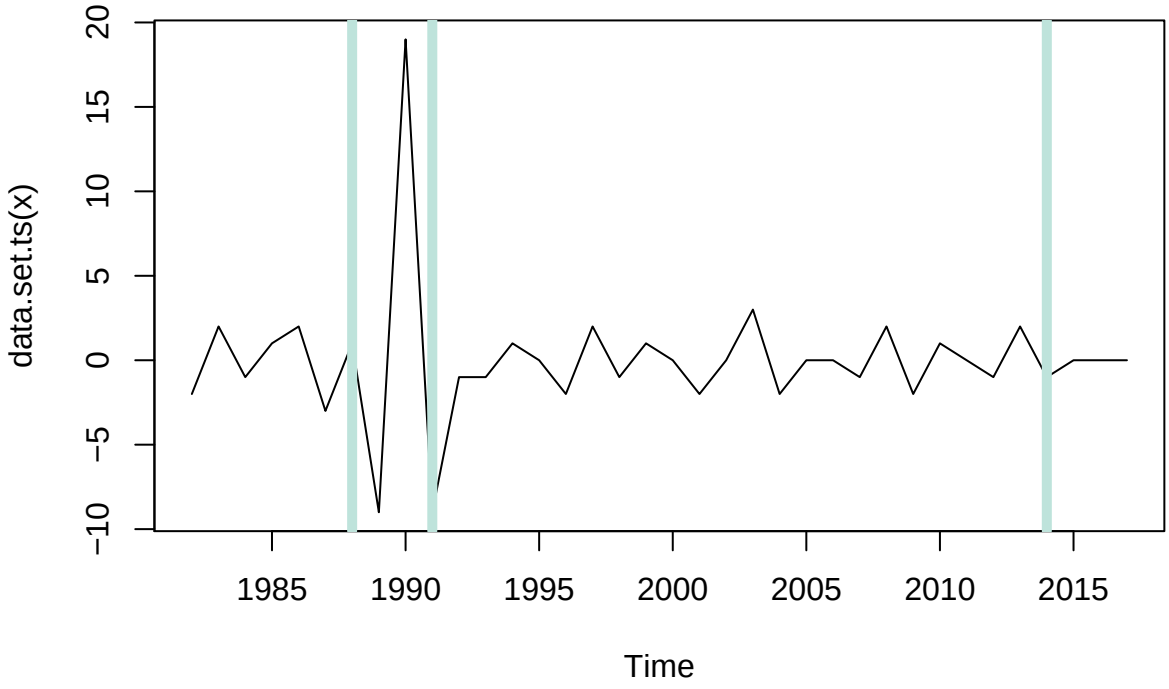
The major change point for **MCV1** occurs at 15. This indicates that the vaccine rate started to increase in 1996.

In addition, for each of the 4 vaccines(won't include HepB_BD) we can find the number of change points in variance throughout the whole time period.

```
cptVarOut1 <- cpt.var(statVC[,1], method = "PELT")
cptVarOut1
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.75056
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations : 7 10 33
```

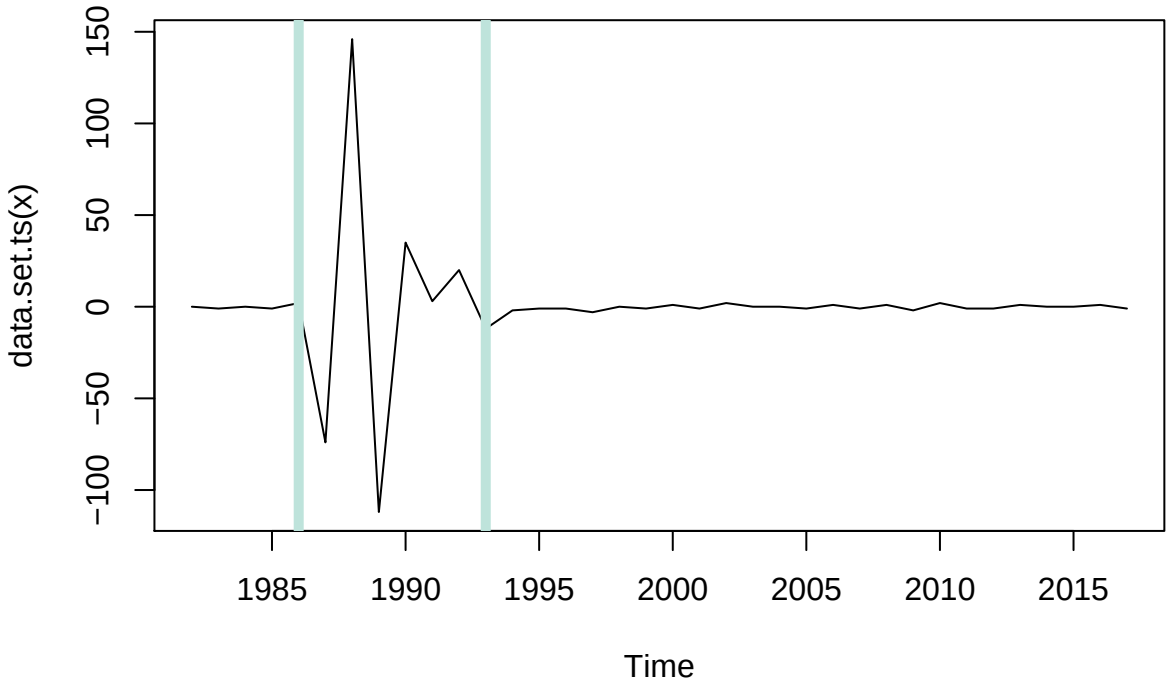
```
plot(cptVarOut1, cpt.col="#BEE3DB",cpt.width=5)
```




```
cptVarOut3 <- cpt.var(statVC[,3], method = "PELT")
cptVarOut3
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.75056
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations : 5 12
```

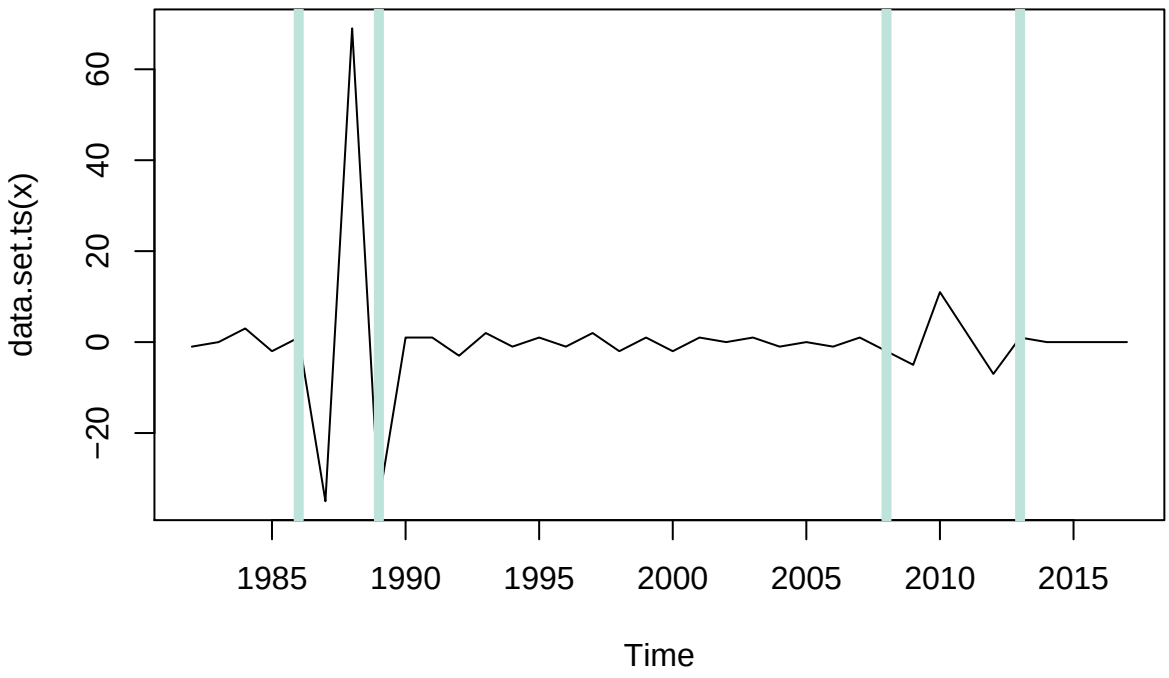
```
plot(cptVarOut3, cpt.col="#BEE3DB",cpt.width=5)
```




```
cptVarOut4 <- cpt.var(statVC[,4], method = "PELT")
cptVarOut4
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.75056
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations : 5 8 27 32
```

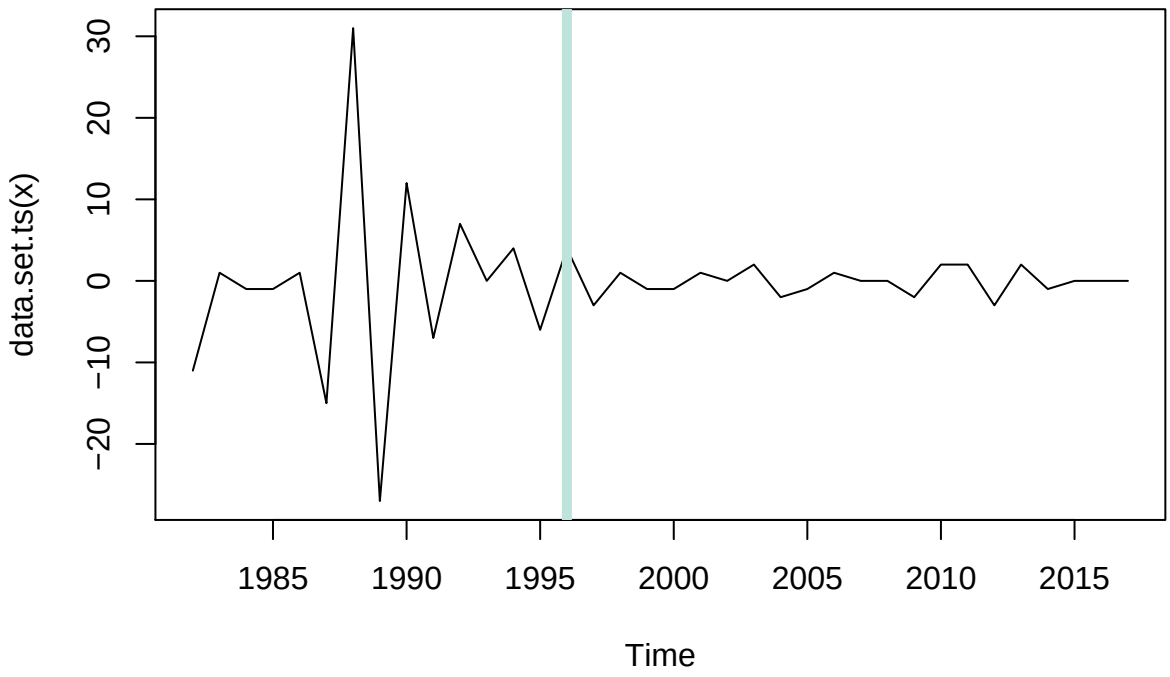
```
plot(cptVarOut4, cpt.col="#BEE3DB",cpt.width=5)
```




```
cptVarOut5 <- cpt.var(statVC[,5], method = "PELT")
cptVarOut5
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.75056
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations : 15
```

```
plot(cptVarOut5, cpt.col="#BEE3DB",cpt.width=5)
```



DTP1: 3

HepB_BD: 0

Pol3: 2

Hib3: 4

MCV1: 1

Hib3 had the most change points and the greatest variance.

The cpt.mean() function allows us to detect transition points where the mean of a time series changes substantively.

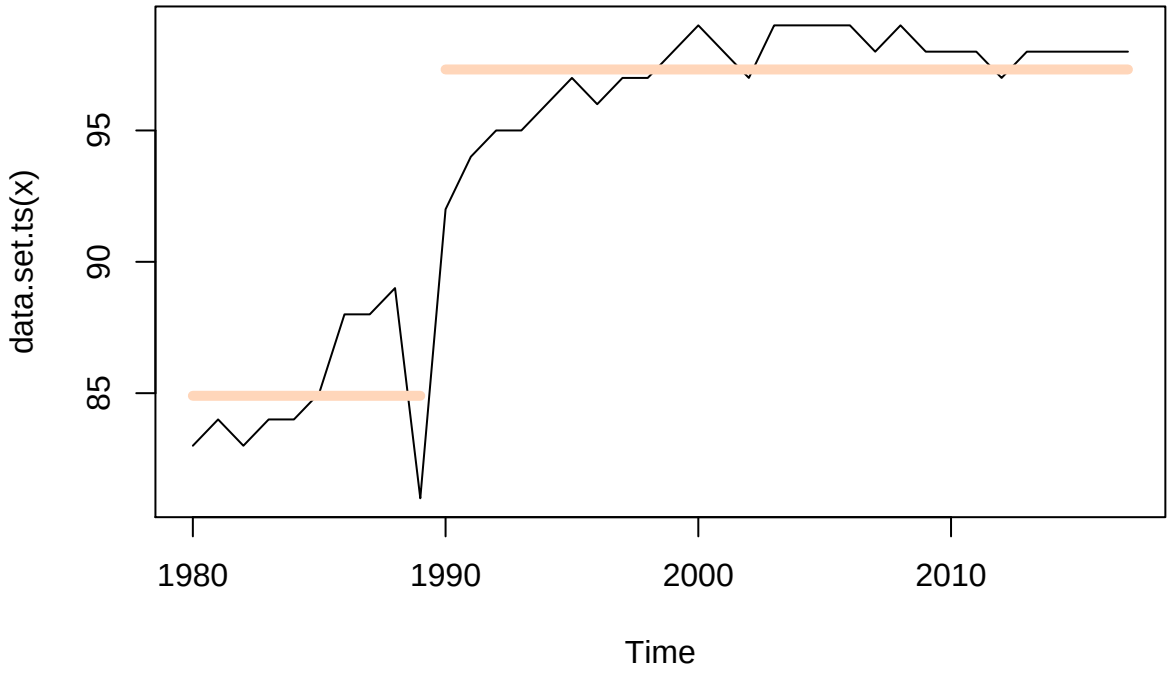
We have left the trend in the data. That is important because the change point analysis of means is usually concerned with places where a time series may have increased or decreased very suddenly.

```
# DTP1 rates variation over time
diffVCcpM1 <- cpt.mean(usVaccines[,1])
diffVCcpM1
```

```
## 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  : Sun Dec 08 06:47:09 2019
```

```
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 10
```

```
plot(diffVCcpM1, cpt.col="#FFD6BA",cpt.width=5)
```



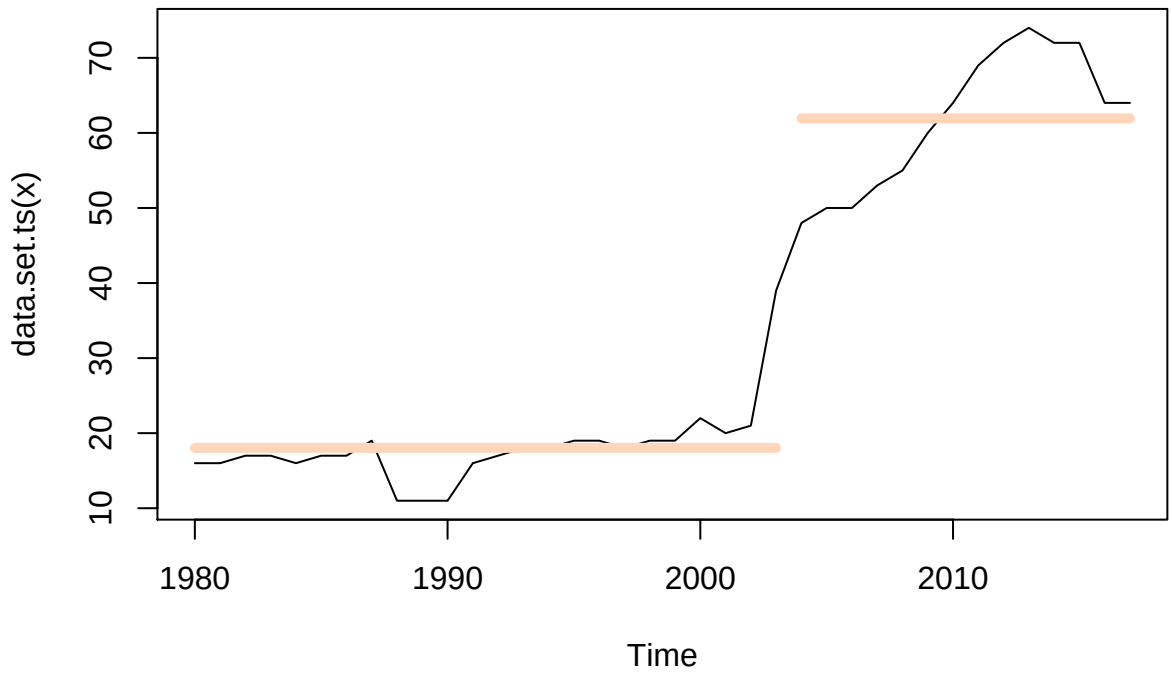
cpt.mean() detected a change in the mean of the time series for DTP1 at point 2. The change in means is documented by the horizontal lines on the plot. The first line shows sort of the typical level, ranging from about 1980 to the beginning of 1989. There are some fluctuations in there, we do have a little bit of a growth pattern between 1985 and 1989. Then in 1990 the index jumps up to a higher level, somewhere around 97. Each horizontal line represents the mean level of an index across the whole time period covered by the line. The change point is 2 (it was 10 with the var).

This result indicates that just as the DTP1 rates started to increase rapidly in 1991, it also entered a period of more intense volatility, that is, substantially greater variance. The change points at 2 and 10 are part of the way through 1991.

```r
# HepB_BD variation over time
diffVCcpM2 <- cpt.mean(usVaccines[,2])
diffVCcpM2
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 24
```

```r
plot(diffVCcpM2, cpt.col="#FFD6BA",cpt.width=5)
```



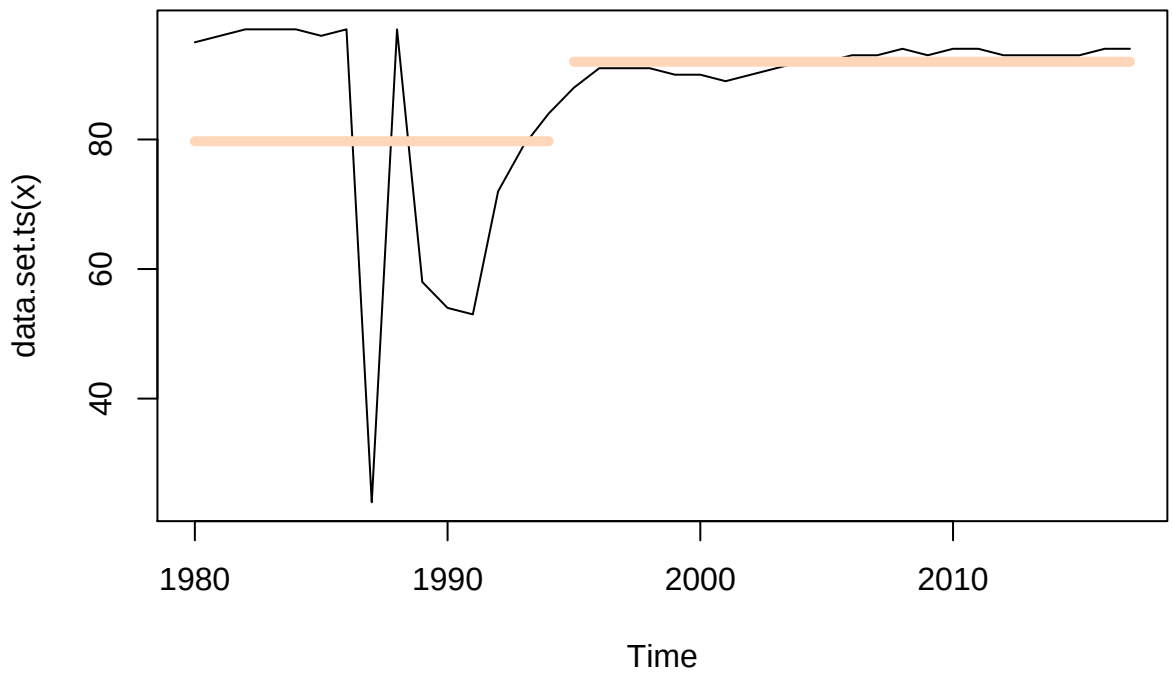
As we didn‘t see any change point in variance, there are some in the mean here. cpt.mean() detected a change point in the mean of this time series at point 24. On the plots we can see both points in time where

the change occur and how big the shift was. It is consistent from 1980 to 2003 and there is a jump up to a higher level. Interestingly from 2009 to 2013, the incidence of acute hepatitis B infection increased by 114% in Kentucky, Tennessee, and West Virginia, while remaining stable in the U.S. overall.

```
# Pol3 variation over time
diffVCcpM3 <- cpt.mean(usVaccines[,3])
diffVCcpM3
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 15
```

```
plot(diffVCcpM3, cpt.col="#FFD6BA",cpt.width=5)
```



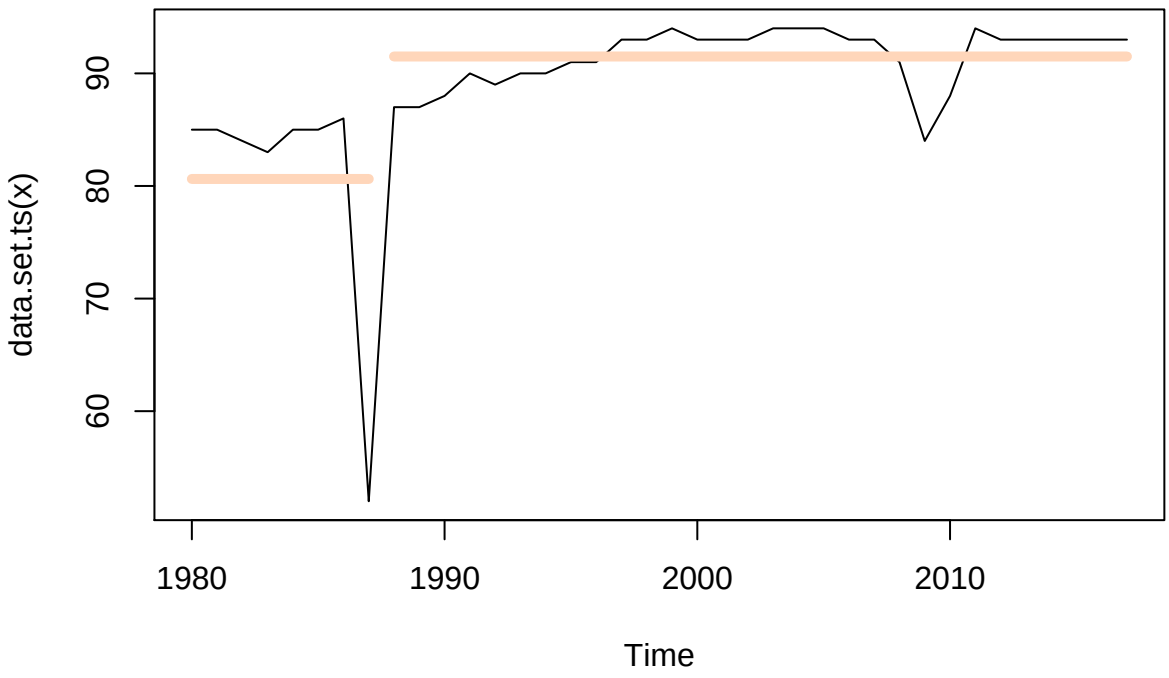
The change point here is at 15. The rise in the rates is from 80 to 90. The first line shows sort of the typical level, ranging from about 1980 to the beginning of 1993. There are some fluctuations in there - going up and all the way down at 1987. Then in 1993 the index jumps up to a higher level, somewhere around 90 and stay around that level.

This result indicates that just as the Pol3 rates started to increase rapidly in 1993, it also entered a period of more intense volatility, that is, substantially greater variance. The change points at 12 and 15 are part of the way through 1993.

```
# Hib3 variation over time
diffVCcpM4 <- cpt.mean(usVaccines[,4])
diffVCcpM4
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 8
```

```
plot(diffVCcpM4, cpt.col="#FFD6BA",cpt.width=5)
```

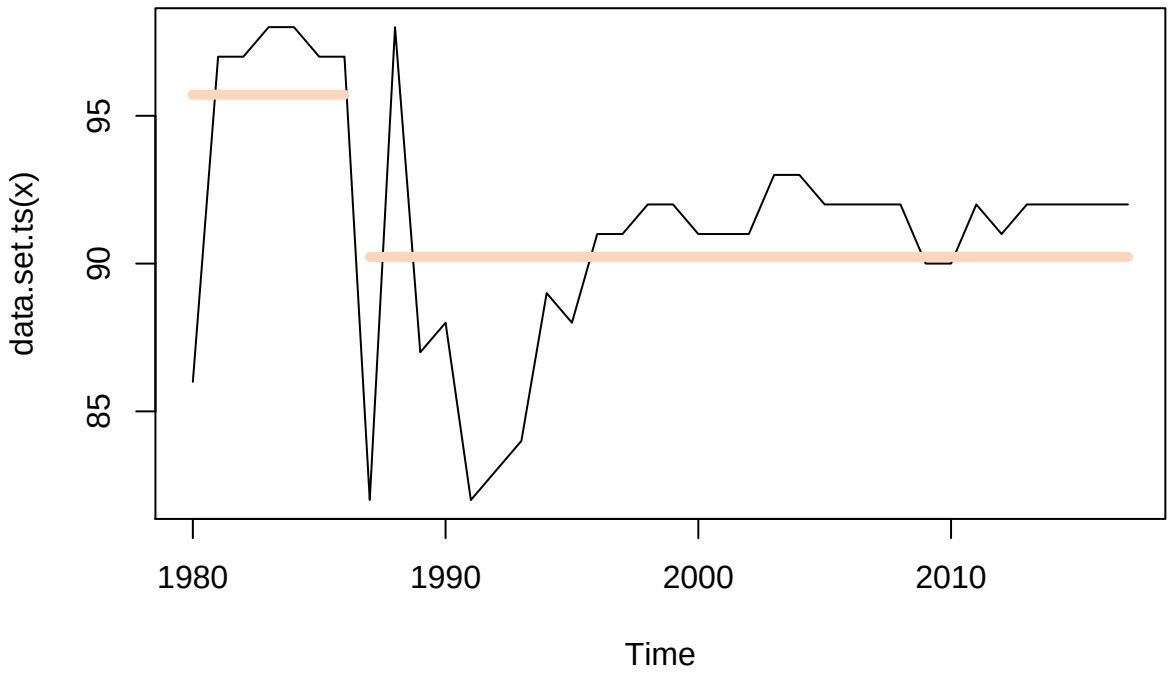


Change point is at 8. From 1980 to 1987 is at 80. There is a sudden drop down in 1987(similar to what we observed with Hib3). After that the rate jump up to a higher level. That drop down might be the cause for variance.

```
# MCV1 variation over time
diffVCcpM5 <- cpt.mean(usVaccines[,5])
diffVCcpM5
```

```
## 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  : Sun Dec 08 06:47:09 2019
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.2
## Changepoint type      : Change in mean
## Method of analysis    : AMOC
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.91276
## Minimum Segment Length : 1
## Maximum no. of cpts   : 1
## Changepoint Locations : 7
```

```
plot(diffVCcpM5, cpt.col="#FFD6BA",cpt.width=5)
```



Each horizontal line represents the mean level of rate across the whole time period covered by the line. There is decrease in the rate here. There is that drop down around 1987 too. In 1996 the final growing trend starts. The change point here is 7(it was 15 with the var). The first line shows sort of the typical level, ranging from about 1980 to the beginning of 1987. There are some fluctuations in there, we do have a little bit of a growth pattern between 1981 and 1987. Then in 1988 the index jumps up to a higher level, somewhere around 98.

The DTP1 were at higher levels around 1987, in comparison to the rest. We observed growth in the trends (only MCV1 has decreased mean level of rates). HepB_BD has the biggest jump (from 20 to 60).

```
## Find the maximum values across each series
lapply(usVaccines[,2:5],max)
```

```
## $HepB_BD
## [1] 74
##
## $Pol3
## [1] 97
##
## $Hib3
## [1] 94
##
## $MCV1
## [1] 98
```

MCV1 has the highest rate of 98 at the conclusion of the time series.

```
## Find the minimum values across each series
lapply(usVaccines[,2:5],min)
```

```
## $HepB_BD
## [1] 11
##
## $Pol3
## [1] 24
##
## $Hib3
## [1] 52
##
## $MCV1
## [1] 82
```

HepB_BD has the lowest rate of 11 at the conclusion of the time series.

```
## Find the first and the last value
usVaccines[1,]
```

```
##    DTP1 HepB_BD    Pol3    Hib3    MCV1
##      83      16      95      85      86
```

```
usVaccines[38,]
```

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

If we look at the first and the last value of each one we can see that all the rates increased from 1980 to 2017, only Pol3 decreased from 95 to 94.

```
## Find the variance for each series
lapply(usVaccines,var)
```

```
## $DTP1
## [1] 34.42959
##
## $HepB_BD
## [1] 508.0085
##
## $Pol3
## [1] 235.7041
##
## $Hib3
## [1] 50.92745
##
## $MCV1
## [1] 17.53698
```

```r
lapply(usVaccines,sd)
```

```
## $DTP1
## [1] 5.867673
##
## $HepB_BD
## [1] 22.53904
##
## $Pol3
## [1] 15.35266
##
## $Hib3
## [1] 7.136347
##
## $MCV1
## [1] 4.187718
```

HepB_BD has the greatest overall SD of differenced series: 22.53904.

2. What proportion of public schools reported vaccination data? What proportion of private schools reported vaccination data? Was there any credible difference in overall reporting proportions between public and private schools?

- use allSchoolsReportStatus data

```r
attach(allSchoolsReportStatus)
## all schools reported vaccination data
anyY <- allSchoolsReportStatus[which(reported == "Y"),]
detach(allSchoolsReportStatus)
count(anyY$pubpriv[anyY$pubpriv == "PUBLIC"]) # 5584 public school reported vaccination data
```

x	freq
PUBLIC	5584

```r
count(anyY$pubpriv[anyY$pubpriv == "PRIVATE"]) # 1397   private schools reported vaccination data
```

x	freq
PRIVATE	1397

```r
totalPublic <- count(allSchoolsReportStatus$pubpriv[allSchoolsReportStatus$pubpriv== "PUBLIC"])
totalPublic # 5732
```

x	freq
PUBLIC	5732

```
totalPrivate <- count(allSchoolsReportStatus$pubpriv[allSchoolsReportStatus$pubpriv== "PRIVATE"])
totalPrivate # 1649
```

x	freq
PRIVATE	1649

```
count(anyY$pubpriv[anyY$pubpriv == "PUBLIC"])/totalPublic
```

x	freq
NA	0.97418

97.4% of public schools reported vaccination data.

```
count(anyY$pubpriv[anyY$pubpriv == "PRIVATE"])/totalPrivate
```

x	freq
NA	0.8471801

84.7% of private schools reported vaccination data.

We have two categorical factors (pubpriv and reported), and we want to see if they are related to one another. The following call to ftable() extracts a split of reported vaccines by school, which we then test for independence from the reported status with the chi-square null hypothesis test.

```
schoolsYN <- ftable(allSchoolsReportStatus, row.vars = 2, col.vars = "reported")
schoolsYN
```

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

```
chiOut <- chisq.test(schoolsYN, correct = FALSE)
chiOut
```

```
##
##  Pearson's Chi-squared test
##
## data:  schoolsYN
## X-squared = 402.97, df = 1, p-value < 2.2e-16
```

What proportion of public schools reported vaccination data?

```
5584/148 #(repored/not reported)
```

```
## [1] 37.72973
```

What proportion of private schools reported vaccination data?

```
1397/252 #(repored/not reported)
```

```
## [1] 5.543651
```

The reported value of chi-square, 402.97 on one degree of freedom, has an extremely low corresponding p-value, well below the standard alpha level of p< .05. Thus we reject the null hypothesis of independence between schools and reported status. These two factors are not independent and by inspecting the 2x2 contingency table we can see that the proportion of reported vaccines in the private schools was considerably lower.

```
chiOut$residuals # how far is the obsereud from expected
```

```
##         reported         N         Y
## pubpriv
## PRIVATE          17.204118 -4.118163
## PUBLIC           -9.227619  2.208822
```

Residuals represent how far an observed value was from the expected value. A large positive residual means that the observation for that cell was too much lower than expected. Large residuals (negative or positive) indicate the cells that made the most powerful contribution to the value of chi-square. In our example that would be PRIVATE/N; PUBLIC/N. This cells showed where the "action" is with respect to non-independance.

The BayesFactor packge produce posterior distributions for the fequencies (or proportions)in the cell of the contingency table.

```
ctBFout <- contingencyTableBF(schoolsYN, sampleType = "poisson", posterior = F)
ctBFout
```

```
## Bayes factor analysis
## --------------
## [1] Non-indep. (a=1) : 1.150548e+69 ±0%
##
## Against denominator:
##   Null, independence, a = 1
## ---
## Bayes factor type: BFcontingencyTable, poisson
```

The Bayes factor of 1.150548e+69 :1 in favor of the alternative hypothesis that the two factors are not independent from one another (they are associated). Because the reported Bayes factor excess 150:1, we can treat it as a strong evidence in favor of the alternative hypothesis(nonindependence). Therefore, in the situation the Bayes factor and the null hypothesis concur with each other.

```
set.seed(1234)
ctMCMCOut <- contingencyTableBF(schoolsYN, sampleType="poisson", posterior=TRUE, iterations=10000)

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]  252.9 15.55   0.1555         0.1555
## lambda[2,1]  148.8 12.25   0.1225         0.1225
## lambda[1,2] 1398.0 37.79   0.3779         0.3779
## lambda[2,2] 5581.5 74.02   0.7402         0.7402
##
## 2. Quantiles for each variable:
##
##               2.5%    25%    50%    75%  97.5%
## lambda[1,1]  223.4  242.2  252.7  263.3  284.1
## lambda[2,1]  125.8  140.4  148.4  156.7  174.0
## lambda[1,2] 1324.8 1372.2 1397.5 1423.2 1472.9
## lambda[2,2] 5436.1 5532.2 5581.6 5631.0 5729.7
```

The resulting object, ctMCMCOut, contains the result of the 10000 samples in the form of means and HDIs for each of the cell counts.

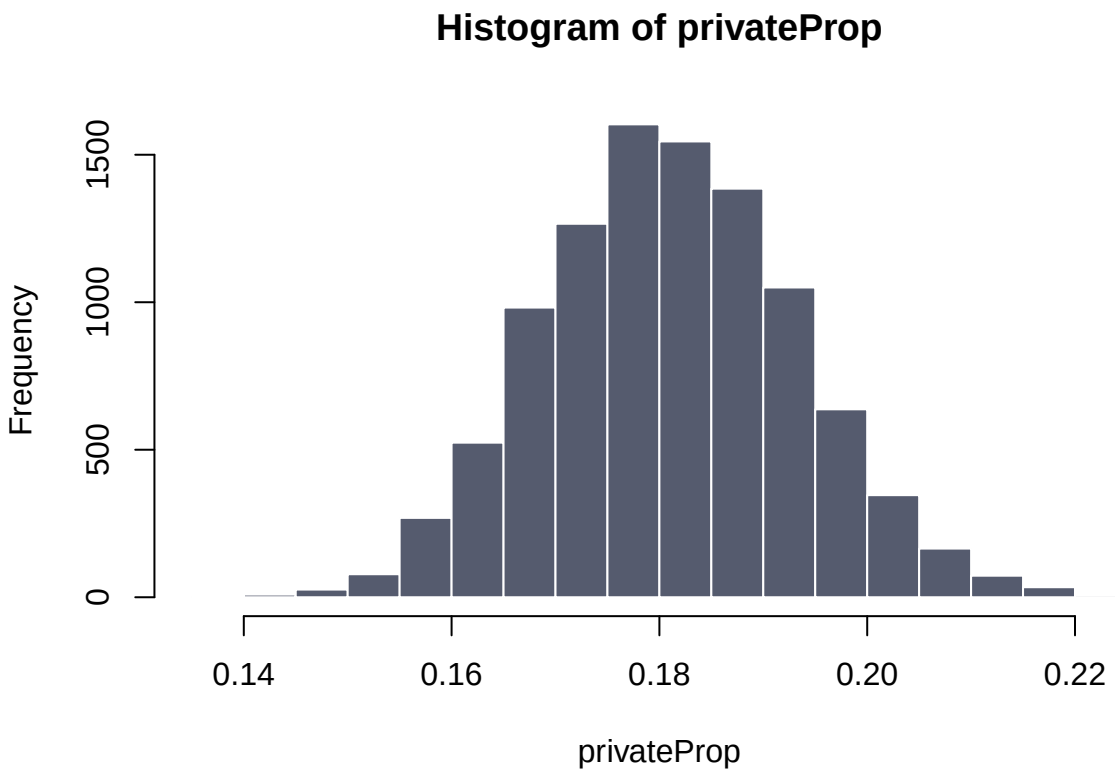
The means at the first section closely match the content of the cells in the original data.

```
schoolsYN
```

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

We don't learn anything new from those means. The second table shows quantiles for each variable, including the boundaries of the 95% highest density interval (HDI)in the fist and the last column.

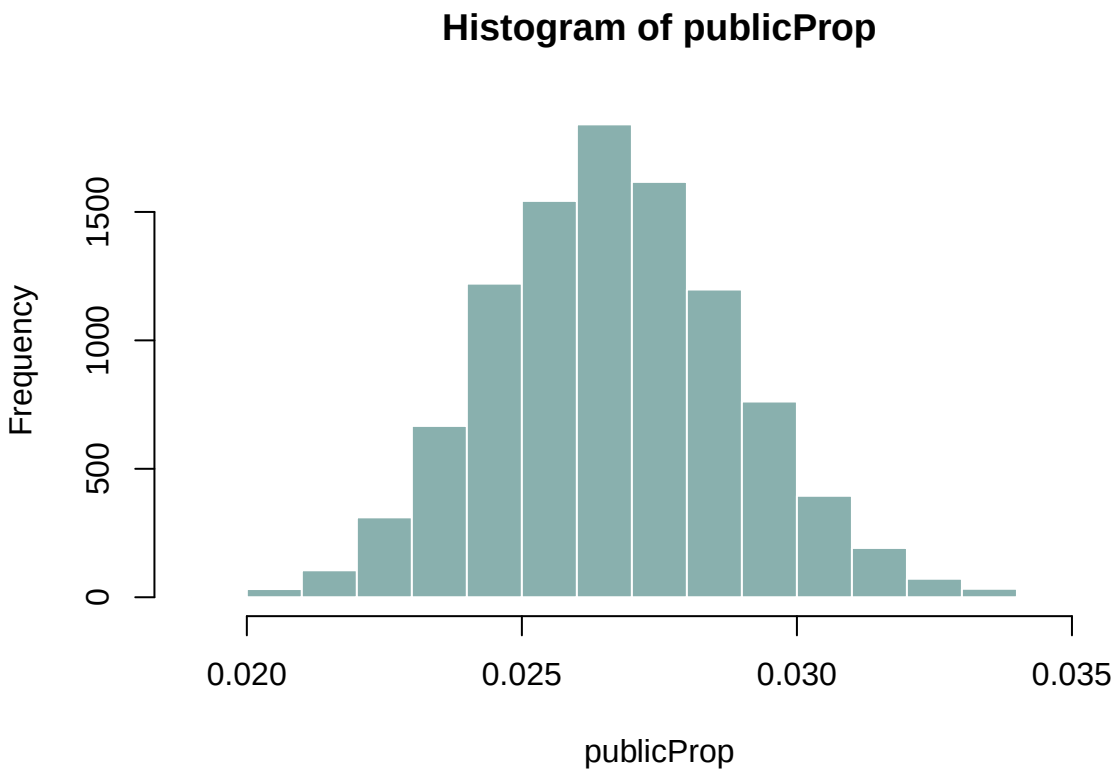
```
privateProp <- ctMCMCOut[,"lambda[1,1]"] / ctMCMCOut[,"lambda[1,2]"]
hist(privateProp
     , border = "white"
     , col = "#555B6E")
```




```
# [1,1] private/N; [1,2] privare/Y - (252/1397)
```

PRIVATE/N have lower proportion than PRIVATE/Y. Most common value is around 0.18.

```
publicProp <- ctMCMCOut[,"lambda[2,1]"] / ctMCMCOut[,"lambda[2,2]"]
hist(publicProp
     , border = "white"
     , col = "#89B0AE")
```

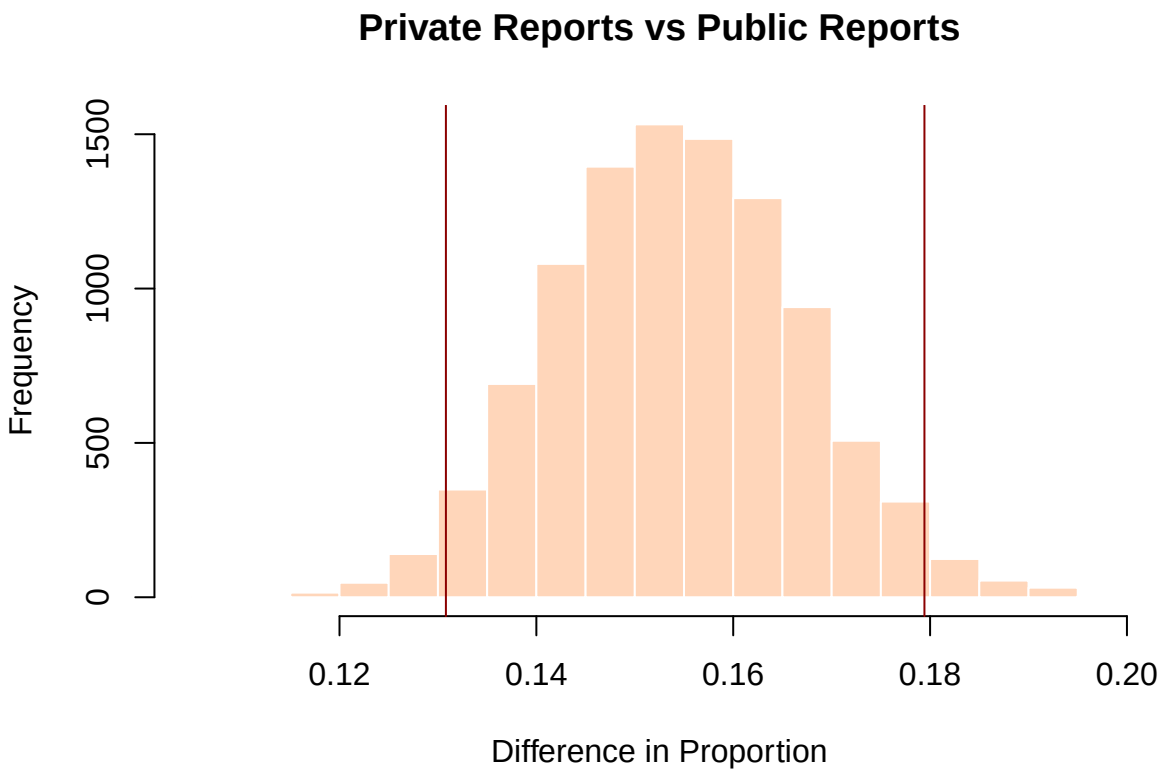



```
# [2,1] PUBLIC/N; [2,2] PUBLIC/Y - (148/5584)
```

Public/Y have higher proportion than PUBLIC/N. The center of the distribution a little bit higher than 0.026.

This two posterior distributions represent the two columns of our original table. We will try find the difference between them by substracting them.

```
## Substract PRIVATE and PUBLIC
subs <- privateProp - publicProp
#range(subs)
hist(subs
     , xlab =  c("Difference in Proportion")
     , main = "Private Reports vs Public Reports"
     , border = "white"
     , col = "#FFD6BA"
     )
## Show HDI
abline(v=quantile(subs, c(0.025)), col = "darkred") # low end
abline(v=quantile(subs, c(0.975)), col = "darkred") # high end
```




```
mean(subs)
```

```
## [1] 0.1543894
```

The center of this distribution is a difference in proportions of 0.15. HDI is marked by ablines. In the population, the proportion shifts about 0.05, although there is small likelihood that the difference in proportions could be as little as about 0.13 or as much as about 0.18.

Was there any credible difference in overall reporting proportions between public and private schools?

HDI does not overlap with 0(none of this posterior estimates of the difference in proportions differed by 0). There is definitely relationship between those two categorical variables and we modeled it with our posterior estimates. In proportion only 26/1000 public school didn't reported vaccination(37.7times as many public schools reported as they did not). 1/8 of the private school did not reported(5.5 times as many private schools reported as they did not). Public schools are did better job reporting in the subset we explored.

3. What are 2013 vaccination rates for individual vaccines (i.e., DOT, Polio, MMR, and HepB) in California public schools? How do these rates for individual vaccines in California districts compare with overall US vaccination rates (make an informal comparison to the final observations in the time series)?

- uses districts12 dataset

```
## 2013 vaccination rates for individual vaccines in overall US vaccination rates
usVaccines[time(usVaccines) == "2013"]
```

```
## [1] 98 74 93 93 92
```

```
## Get the finlal value of each series
usVaccines[38,]
```

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

```
head(districts12, n=3) # A sample of California public school districts from the 2013 data collection,
```

	DistrictName	WithoutDTP	WithoutPolio	WithoutMMR	WithoutHepB	PctUpToDate	Dist
118	Lowell Joint	6	5	6	5	94	TRU
413	Chowchilla Elementary	2	3	3	2	96	TRU
610	Harmony Union Elementary	38	38	38	39	61	TRU

What are 2013 vaccination rates for individual vaccines (i.e., DOT, Polio, MMR, and HepB) in California public schools?

```
# DTP
DTP <- round(100 - mean(districts12$WithoutDTP),2) # 89.78

# Polio
Polio <- round(100 - mean(districts12$WithoutPolio),2) # 90.18

# MMR
MMR <- round(100 - mean(districts12$WithoutMMR),2) # 89.73

# HepB
HepB <- round(100 - mean(districts12$WithoutHepB),2) # 92.15
```

How do these rates for individual vaccines in California districts compare with overall US vaccination rates (make an informal comparison to the final observations in the time series)?

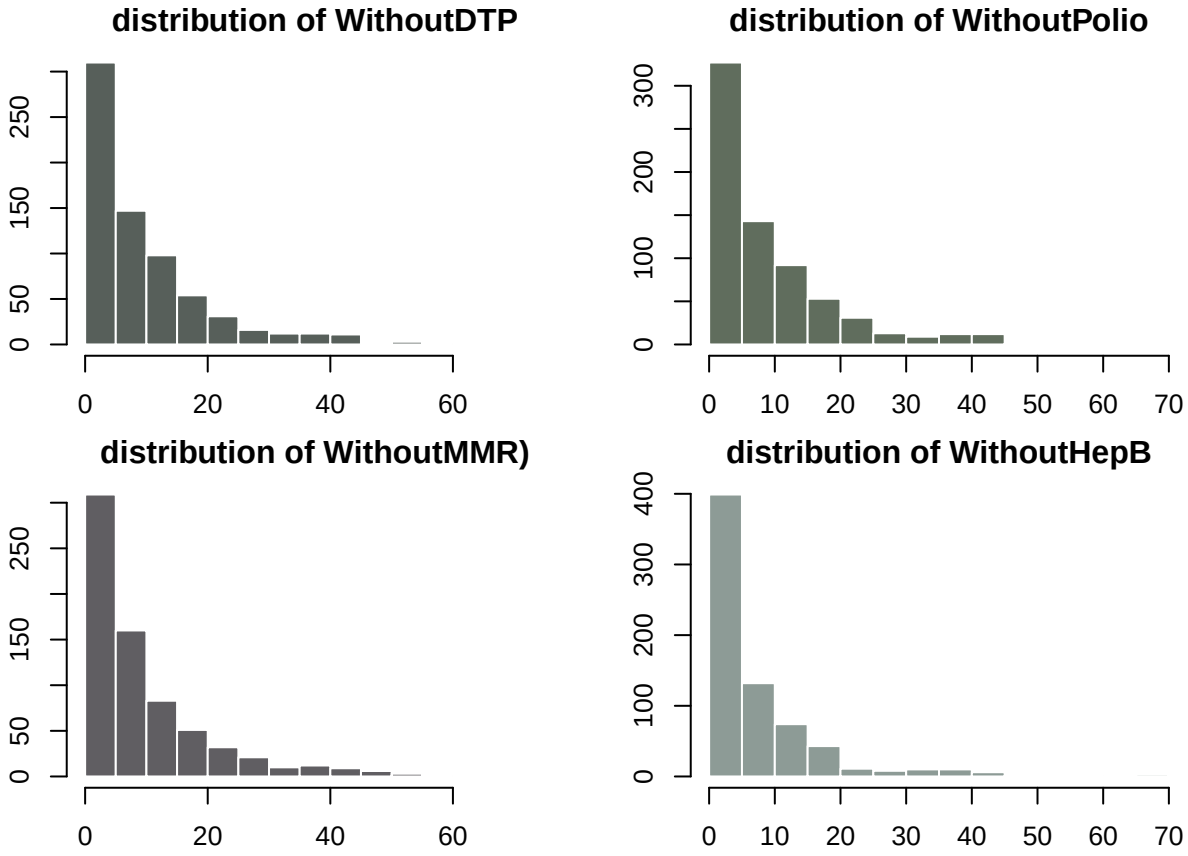
All the retes are lower compared to overall US vaccination rates for 2013 and the last values of the time series, only HepB has an increase compared to 2013 and of the final observation of the time series.

4. Among districts, how are the vaccination rates for individual vaccines related? In other words, if students are missing one vaccine are, they missing all of the others?

```
df <- districts12[,2:5]
cor(df)
```

	WithoutDTP	WithoutPolio	WithoutMMR	WithoutHepB
WithoutDTP	1.0000000	0.9814768	0.9764722	0.9073381
WithoutPolio	0.9814768	1.0000000	0.9638816	0.9228270
WithoutMMR	0.9764722	0.9638816	1.0000000	0.9051035
WithoutHepB	0.9073381	0.9228270	0.9051035	1.0000000

```
par(mar = c(2,2,2,2))
par(mfrow = c(2,2))
hist(df[,1], main = "distribution of WithoutDTP"
     , border = "white", col = "#575F5A")
hist(df[,2], main = "distribution of WithoutPolio"
     , border = "white", col = "#606D5D")
hist(df[,3], main = "distribution of WithoutMMR)"
     , border = "white", col = "#605E62")
hist(df[,4], main = "distribution of WithoutHepB"
     , border = "white", col = "#8D9B96")
```



The plots show similar dsitributions.

The 1.0 values of the diagonal are the correlations between each variable and itself. There are two triangles of correlation data, one above diagonal and one below. The two triangles are transposed versions of each other: they contain the same information, so we really need to look at the lower triangle.

Percentage of student without vaccine are highly correlated (strong positive correlation, al over .9), suggesting the possibility that these variables might in some senses be redundant with each other.

Null hypothesis testing on the correlation – the procedure for testing the significance of the correlation coefficient assumes a null hypothesis of rho = 0.

In many cases researchers report the results of the null hypothesis significant test on each correlation coefficient. We can compare them in pair too.

```r
cor.test(df$WithoutDTP, df$WithoutPolio)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$WithoutDTP and df$WithoutPolio
## t = 135.35, df = 698, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9785437 0.9840122
## sample estimates:
##       cor
## 0.9814768
```

```r
cor.test(df$WithoutPolio, df$WithoutMMR)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$WithoutPolio and df$WithoutMMR
## t = 95.616, df = 698, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9582210 0.9687875
## sample estimates:
##       cor
## 0.9638816
```

```r
cor.test(df$WithoutMMR, df$WithoutHepB)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$WithoutMMR and df$WithoutHepB
## t = 56.24, df = 698, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8907435 0.9176583
## sample estimates:
##       cor
## 0.9051035
```

```r
cor.test(df$WithoutHepB, df$WithoutDTP)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$WithoutHepB and df$WithoutDTP
## t = 57.021, df = 698, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.8932973 0.9196098
## sample estimates:
##       cor
## 0.9073381
```

```
cor.test(df$WithoutDTP, df$WithoutMMR)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$WithoutDTP and df$WithoutMMR
## t = 119.63, df = 698, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9727574 0.9796856
## sample estimates:
##       cor
## 0.9764722
```

```
cor.test(df$WithoutPolio, df$WithoutHepB)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$WithoutPolio and df$WithoutHepB
## t = 63.291, df = 698, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9110237 0.9331193
## sample estimates:
##      cor
## 0.922827
```

The outputs above has three sections: the first three lines are the conventional null hypothesis test with an assumption of rho = 0. For our example the null hypothesis will be that rho, the population correlation coefficient between WithoutPolio and WithoutHepB, is zero. The alternative hypothesis will be simply the logical opposite and incorporates the possibility of nonzero correlation that is either positive or negative. We have detected statistical significance (p-value < 2.2e-16) in all relatioships. The test statistic is a t-test on a transformed version of the correlation coefficient. The test yields a t-value of each one. Df are the degrees of freedom – how many elements are free to vary in a statistical system. One way of thinking about p-value is to say that there is 2.2e-16 chance of observing an absolute value of t this high or higher under the assumption that the population value of rho = 0. Using the conventional p < .05 threshold for alpha to evaluate this result, we can reject the null hypothesis of **rho=0**. The cor.test() also provides 95% confidence interval around the point estimate of r. We can define CI as follows: if we repeated this sampling process many times and each time constructed a confidence interval around the calculated value of r, about 95% of those constructed confidence intervals would contain the true population value of rho. In conclusion we cam say that 95% CI for rho has tight range in all of the observations. Importantly CI does not straddle with 0, result that concur with the result from the significance test and we have a sense of certainty that the correlation is positive. In a student is misiisng one vaccine, there is a big chance he is missing all of them.

We can try to see all the correlations with corr.test() from psych package.

```r
class(df)
```

```
## [1] "data.frame"
```

```r
df1<- as.matrix(df)

corr.test(df1, adjust = "none")
```

```
## Call:corr.test(x = df1, adjust = "none")
## Correlation matrix
##              WithoutDTP WithoutPolio WithoutMMR WithoutHepB
## WithoutDTP         1.00         0.98       0.98        0.91
## WithoutPolio       0.98         1.00       0.96        0.92
## WithoutMMR         0.98         0.96       1.00        0.91
## WithoutHepB        0.91         0.92       0.91        1.00
## Sample Size
## [1] 700
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##              WithoutDTP WithoutPolio WithoutMMR WithoutHepB
## WithoutDTP            0            0          0           0
## WithoutPolio          0            0          0           0
## WithoutMMR            0            0          0           0
## WithoutHepB           0            0          0           0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

```r
cor.test.p <- function(x){
    FUN <- function(x, y) cor.test(x, y)[["p.value"]]
    z <- outer(
      colnames(x),
      colnames(x),
      Vectorize(function(i,j) FUN(x[,i], x[,j]))
    )
    dimnames(z) <- list(colnames(x), colnames(x))
    z
}

cor.test.p(df1)
```

	WithoutDTP	WithoutPolio	WithoutMMR	WithoutHepB
WithoutDTP	0	0	0	0
WithoutPolio	0	0	0	0
WithoutMMR	0	0	0	0
WithoutHepB	0	0	0	0

Predictive Analyses:

(For all of these analyses, use **PctChildPoverty**, **PctFreeMeal**, **PctFamilyPoverty**, **Enrolled**, and **TotalSchools** as predictors. Transform variables as necessary to improve prediction and/or interpretability.

In general, if there is a Bayesian version of an analysis available, you are expected to run that analysis in addition to the frequentist version of the analysis.)

5. What variables predict whether or not a district's reporting was complete?

```
unique(districts12$DistrictComplete)
```

```
## [1]  TRUE FALSE
```

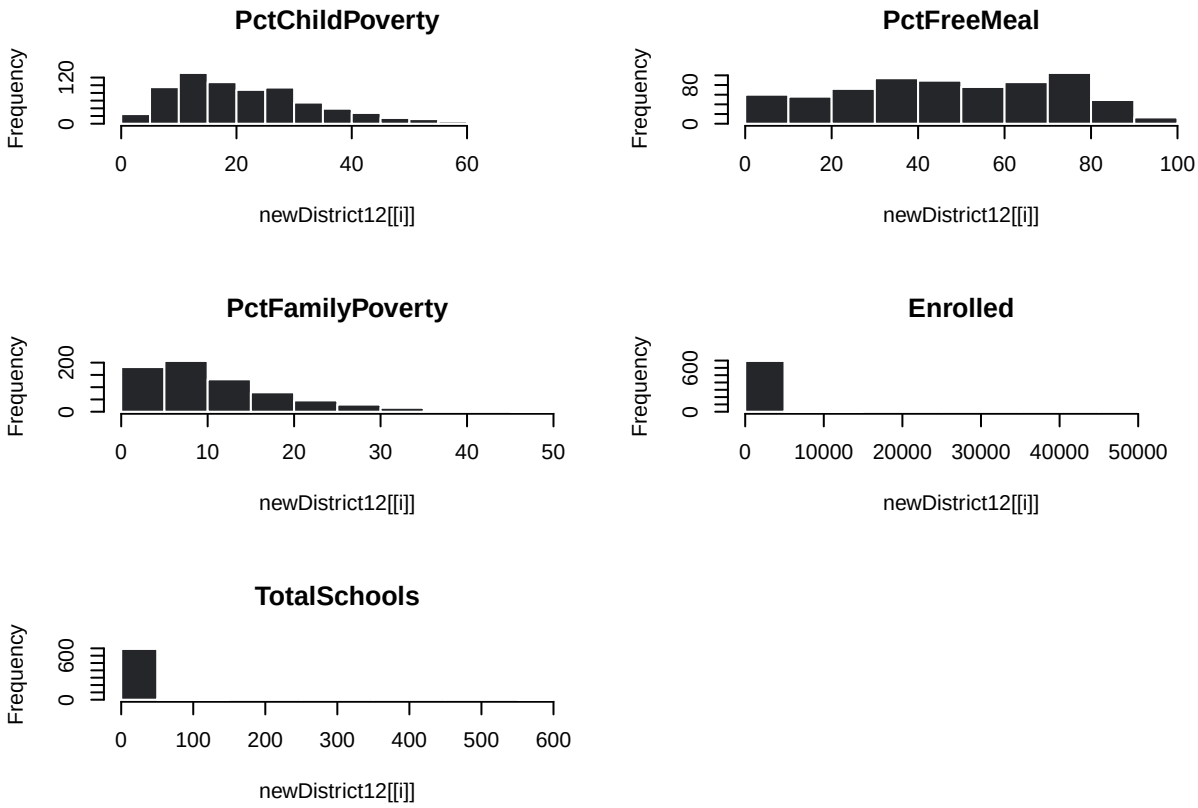
```
districts12$DistrictComplete <- factor(districts12$DistrictComplete, levels = c("TRUE", "FALSE"))
```

The dependent variable is categorical (true/false) so we have to use logistic regression fpt the prediction.

```
newDistrict <- districts12[7:13]
newDistrict12 <- newDistrict[,-2]
head(newDistrict12)
```

	DistrictComplete	PctChildPoverty	PctFreeMeal	PctFamilyPoverty	Enrolled	TotalSchools
118	TRUE	13	28	6	341	5
413	TRUE	25	69	15	271	1
610	TRUE	11	26	2	64	2
163	TRUE	25	66	12	1020	9
550	TRUE	9	22	5	871	8
361	TRUE	55	77	16	33	1

```
par(mfrow=c(3,2))
for (i in c(2:6)) hist(newDistrict12[[i]],main=colnames(newDistrict12)[i], col = "#252629", border ="wh
```



We have to use logistic regression to predict DistrictComplete, using variables PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools as predictors. The predictors are not normaly distributed.

```
round(cor(newDistrict12[,c(2:6)]),2)
```

	PctChildPoverty	PctFreeMeal	PctFamilyPoverty	Enrolled	TotalSchools
PctChildPoverty	1.00	0.76	0.86	0.03	0.02
PctFreeMeal	0.76	1.00	0.73	0.06	0.06
PctFamilyPoverty	0.86	0.73	1.00	0.04	0.03
Enrolled	0.03	0.06	0.04	1.00	0.99
TotalSchools	0.02	0.06	0.03	0.99	1.00
The poverty and fre	e meal are highly	correlated whi	ch make sense.		

Next, we will create and interpret a basic logistic regression model using glm.

- GLM Output

```
glmOut <- glm(formula = DistrictComplete ~ Enrolled +TotalSchools+PctChildPoverty + PctFreeMeal+PctFamil
              family = binomial(link="logit"),
              data = newDistrict12)
summary(glmOut)
```

```
##
```

```
## Call:
## glm(formula = DistrictComplete ~ Enrolled + TotalSchools + PctChildPoverty +
##     PctFreeMeal + PctFamilyPoverty, family = binomial(link = "logit"),
##     data = newDistrict12)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2082  -0.3467  -0.2908  -0.2515   2.6828
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.6245740  0.4499206  -8.056 7.88e-16 ***
## Enrolled         -0.0020908  0.0006375  -3.280 0.001038 **
## TotalSchools      0.2059085  0.0559274   3.682 0.000232 ***
## PctChildPoverty  -0.0311363  0.0310881  -1.002 0.316561
## PctFreeMeal       0.0120245  0.0109768   1.095 0.273323
## PctFamilyPoverty  0.0481995  0.0398236   1.210 0.226154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 312.23  on 699  degrees of freedom
## Residual deviance: 276.12  on 694  degrees of freedom
## AIC: 288.12
##
## Number of Fisher Scoring iterations: 6
```

In the equation we can see the "link function" - in this case indicating binomial(). By specifying "binomial()" we invoke the inverse logit as the basis of fitting the X variablesto the Y variable.

The "Deviance Residuals" show diagnostic information about the distribution of the residuals after the model is fit. The mean of the residuals should be always 0, in our case slightly under 0.

```
mean(residuals(glmOut))
```

```
## [1] -0.1735764
```

The fact that the median residual is slightly negative suggest that the distribution of the residuals is slightly positive skewed.

These residuals represent error of prediction. If there is residual that is strongly positive or strongly negative, it might suggest problem, such as present of an outlier like in this case.

The output shows that the **intercept** is *statistically significant*. The value of the intercept is not very meaningful to us, but we must keep it in the model to make sure that other coefficient are calibrated correctly.

The coefficient of **PctChildPoverty** is *not significantly different from 0*, based on a Wald's z-test value of -1 and associated p-value of 0.316561. Thus we *fail to reject* the null hypothesis that the log-odds of PctChildPoverty is equal to 0 in the population.

The coefficient of **PctFreeMeal** is *not significantly different from 0*, based on a Wald's z-test value of 1.09 and associated p-value of 0.273323. Thus we *fail to reject* the null hypothesis that the log-odds of PctFreeMeal is equal to 0 in the population.

The coefficient of **PctFamilyPoverty** is *not significantly different from 0*, based on a Wald's z-test value of 1.21 and associated p-value of 0.226154. Thus we *fail to reject* the null hypothesis that the log-odds of PctFamilyPoverty is equal to 0 in the population.

The coefficient on the **Enrolled** predictor is *statistically significant*, based on the Wald's z-test value of -3.28 and the associated p-value. Because p-value (0.001038) <.05 we can reject the null hypothesis that the log-odds of Enrolled is 0 in the population. The Wald's z-test is calculated by dividing the coefficient value by the standard error.

The coefficient on the **TotalSchools** predictor is *statistically significant*, based on the Wald's z-test value of 3.68 and the associated p-value. Because p-value (0.000232) <.05 we can reject the null hypothesis that the log-odds of TotalSchools is 0 in the population. The Wald's z-test is calculated by dividing the coefficient value by the standard error.

Enrolled and TotalSchools are the only good predictors. All these coefficients are log-odds values, we need to convert them to regular odds for easier interpretation.

```
confint(glmOut)
```

	2.5 %	97.5 %
(Intercept)	-4.5702585	-2.7977791
Enrolled	-0.0033833	-0.0008021
TotalSchools	0.1006258	0.3235060
PctChildPoverty	-0.0941436	0.0271771
PctFreeMeal	-0.0093088	0.0338042
PctFamilyPoverty	-0.0308557	0.1254190

```
exp(cbind(OR = coef(glmOut), confint(glmOut)))
```

	OR	2.5 %	97.5 %
(Intercept)	0.0266605	0.0103553	0.0609453
Enrolled	0.9979113	0.9966224	0.9991983
TotalSchools	1.2286407	1.1058628	1.3819645
PctChildPoverty	0.9693434	0.9101520	1.0275497
PctFreeMeal	1.0120971	0.9907344	1.0343821
PctFamilyPoverty	1.0493800	0.9696155	1.1336233

We usually ignore odds ratios for intercept terms.

The odds of having complete reporting increase by 0.99 for every enrolled student. In this case the odds are almost 1:1 The 95% confidence interval for Enrolled ranged from 0.996:1 up to 0.999:1 (really close to 1, but doesn't straddle), expressed in plain odds; if the study was repeated 100 times, 95% of similarly constructed intervals would contain the true population value.

The odds of 1.22:1for TotalSchools shows that for every additional school in the district, complete reporting is about 1.22% more likely to be TRUE. The 95% confidence interval for TotalSchools ranged from 1.10:1 up to 1.38:1.

The confidence intervals for PctChildPoverty, PctFreeMeal and PctFamilyPoverty stradles with 1:1, confirming the nonsignificant results for that coefficient. The definition of CI is that if you constructed a very large number of similar experiments based on new samples, 95% of the CI you would calculate would contain the population value.

We have to take a look at the deviance too.

Null deviance: 312.23 on 699 degrees of freedom

Residual deviance: 276.12 on 694 degrees of freedom

AIC: 288.12

The last line shows how many iterations of the model fitting it took to get the final model. AIC is a measure of stress in the model. The "Null Deviance" shows amount of error in the model, if we pretend there is no connection between X variables and Y variable. It shows what would happen if the predictors had no predictive value. The null model shows 699 degrees of freedom for calculating the proportion of TRUE and FALSE. The null model in some ways represents the null hypothesis. The next line shows how much error is reduced by introducing the X variables. We lose 5 degrees by introducing 5 variables. By introducing 5 predictors we reduced error from 312.23 to 276.12 (which cost 3 degree of freedom).

The difference between the null model and the residual model is distributed as chi-square and can be used as an omnibus test.

```
# Compare null model to two predictor model
anova(glmOut, test = "Chisq")
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	699	312.2259	NA
Enrolled	1	14.677648	698	297.5483	0.0001275
TotalSchools	1	16.803608	697	280.7447	0.0000415
PctChildPoverty	1	1.356336	696	279.3884	0.2441736
PctFreeMeal	1	1.829040	695	277.5593	0.1762410
PctFamilyPoverty	1	1.441292	694	276.1180	0.2299304

These results make sense in the light of the significance test on the coefficients and confirms the utility of a model that contains only Enrolled and TotalSchools.

```
table(round(predict(glmOut, type= "response")), newDistrict12$DistrictComplete)
```

/	TRUE	FALSE
0	657	37
1	2	4

The off diagonal items, 37 and 2 are all the errorneous predictions.

```
## Overall accuracy
(657+4)/(657+4+2+37)
```

```
## [1] 0.9442857
```

```
(37+2)/(657+4+2+37) # error rate
```

```
## [1] 0.05571429
```

The overall accuracy was 94%.

Conclusion

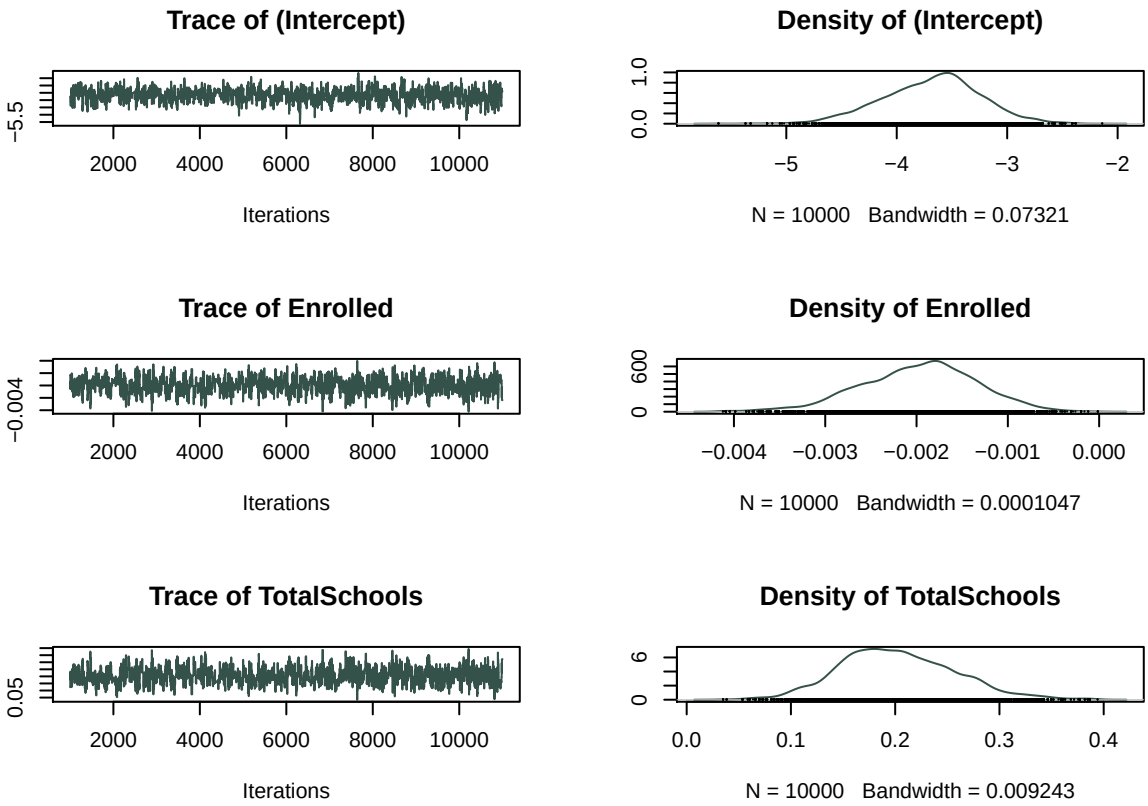
We tested a measures of PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled and TotalSchools to see if they could predict wheater or not a district's reporting was complete. A chi-square omnibus test on the result of logistic regression was significant for model with the 2 predictors, chisq(2) = 16.8036, p<.0001. Only the Wald's z-test on the Enrolled and TotalSchools coefficient was significant, z= -8.05, p<.05 and z=3.28, p<.05. When converted to odds, the coefficient for Enrolled was 0.99 suggesting that for each unit increase in enrolled, the odds of the district's reporting to be complete is .99:1. Coefficient for TotalSchools was 1.22(odds 1.22:1).

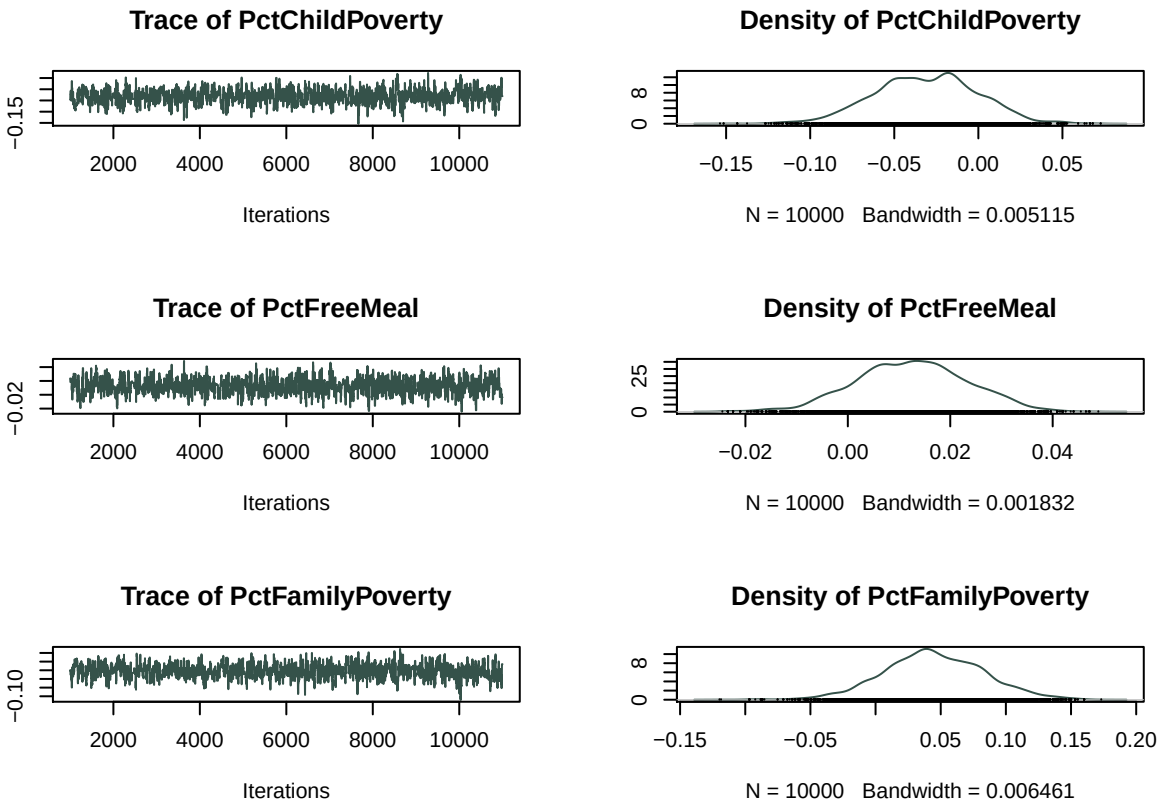
- Bayesian Analysis

```
newDistrict12$DistrictComplete <- as.numeric(newDistrict12$DistrictComplete)- 1
bayesLogitOut <- MCMClogit(formula = newDistrict12$DistrictComplete ~  Enrolled +
                           TotalSchools + PctChildPoverty + PctFreeMeal+PctFamilyPoverty, data = newDi
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)      -3.683791 0.4412208 4.412e-03      1.842e-02
## Enrolled         -0.001973 0.0006323 6.323e-06      2.801e-05
## TotalSchools      0.202389 0.0550204 5.502e-04      2.430e-03
## PctChildPoverty  -0.031246 0.0304494 3.045e-04      1.329e-03
## PctFreeMeal       0.012477 0.0109052 1.091e-04      4.703e-04
## PctFamilyPoverty  0.044700 0.0384598 3.846e-04      1.689e-03
##
## 2. Quantiles for each variable:
##
##                      2.5%       25%       50%      75%      97.5%
## (Intercept)      -4.584664 -3.972266 -3.649187 -3.38833 -2.8271090
## Enrolled         -0.003308 -0.002376 -0.001922 -0.00154 -0.0007942
## TotalSchools      0.103821  0.162600  0.198570  0.23861  0.3208837
## PctChildPoverty  -0.090270 -0.052017 -0.031006 -0.01067  0.0245840
## PctFreeMeal      -0.008577  0.005118  0.012564  0.01982  0.0329034
## PctFamilyPoverty -0.032748  0.019193  0.043205  0.07107  0.1209419
```

```
plot(bayesLogitOut, col = "#35524A", border = "white")
```


Trace of (Intercept)
−5.5
2000
4000
6000
8000
10000
Iterations
Density of (Intercept)
0.0
1.0
−5
−4
−3
−2
N = 10000 Bandwidth = 0.07321
Trace of Enrolled
−0.004
2000
4000
6000
8000
10000
Iterations
Density of Enrolled
0
600
−0.004
−0.003
−0.002
−0.001
0.000
N = 10000 Bandwidth = 0.0001047
Trace of TotalSchools
0.05
2000
4000
6000
8000
10000
Iterations
Density of TotalSchools
0
6
0.0
0.1
0.2
0.3
0.4
N = 10000 Bandwidth = 0.009243



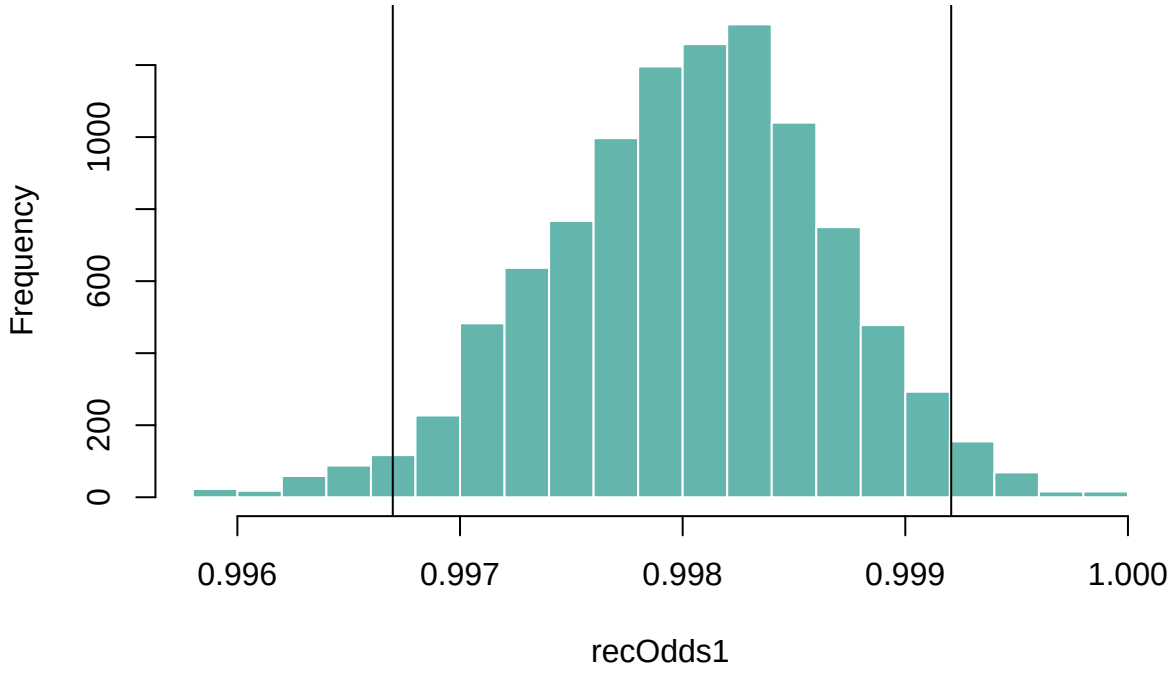
Trace plots show the progress of the MCMC estimation process. Density plots show the posterior distribution of each coefficient. The last three are centered near 0 which confirm that there isn't much going on with those variables and might not be a good predictors. They are all quite normally shaped and the central region of 95% under the curve is where in all likelihood the parameter of interest lies.

The output of MCMC focuses on describing the distribution of the parameters representing both the intercept and the coefficients calibrated in log-odds.

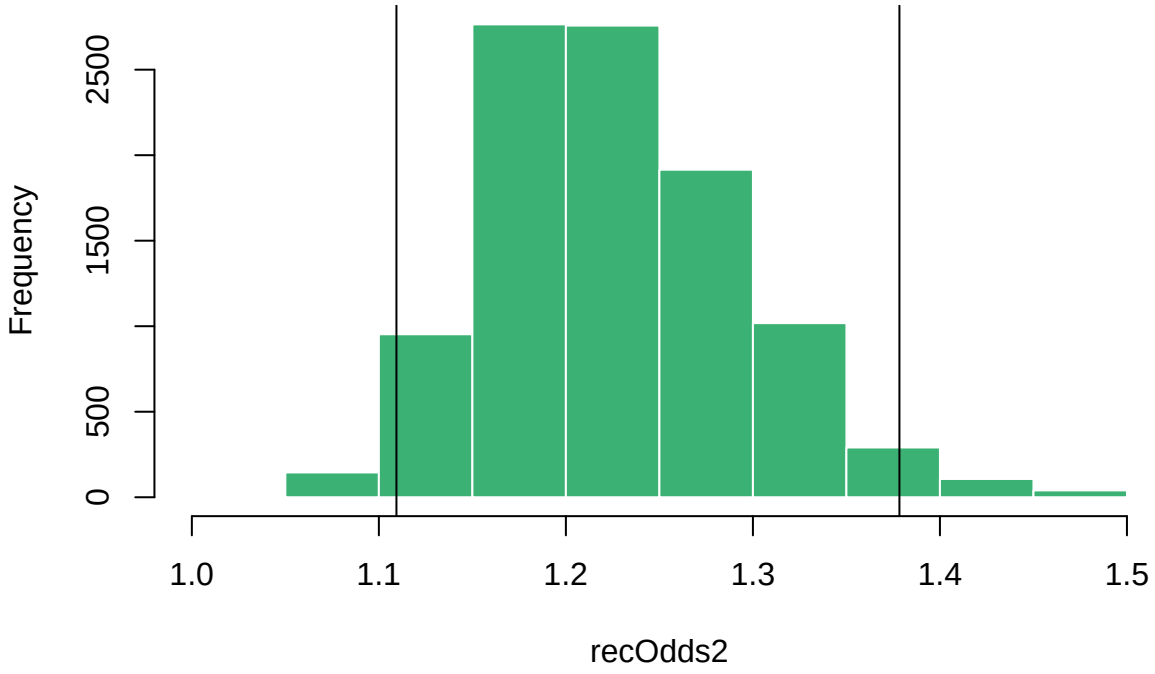
The point estimates in the current output are listed under the "Mean" column and are similar to the output from the traditional logistic regression. The next column "SD" corresponds to the standard error in the output. The most common points of interest will be the log-odd coefficients of the predictors. In the second output we can clearly see that the HDIs for PctChildPoverty, PctFreeMeal and PctFamilyPoverty overlap with 0, so the population parameter for each of them lies somewhere near 0. We need to convert Enrolled and TotalSchools to plain odds in order to interpret it, because the iintervals does not overlap with 0 and we can use that predictor.

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

```
recLogOdds1 <- as.matrix(bayesLogitOut[,"Enrolled"])
recOdds1 <- apply(recLogOdds1,1,exp)
hist(recOdds1, main=NULL, col = "#64B6AC", border = "white")
abline(v=quantile(recOdds1,probs=c(0.025,0.975)))
```




```
#mean(recOdds1) # 0.9980289

recLogOdds2 <- as.matrix(bayesLogitOut["TotalSchools"])
recOdds2 <- apply(recLogOdds2,1,exp)
hist(recOdds2, main=NULL, col = "#3BB273", border = "white")
abline(v=quantile(recOdds2,probs=c(0.025,0.975)))
```




```
#mean(recOdds2) # 1.226189
```

The first histogram shows a distribution centered around .99 consistent with the results we obtained from glm(). The HDI bounds here are similar to but not identical to the confidence interval from glm() too.

The second histogram shows a distribution centered around 1,22 consistent with the results we obtained from glm() too. Odds of 0.99:1 for the coefficient on the Enrolled and 1.22:1 on TotalSchools.

Conclusion

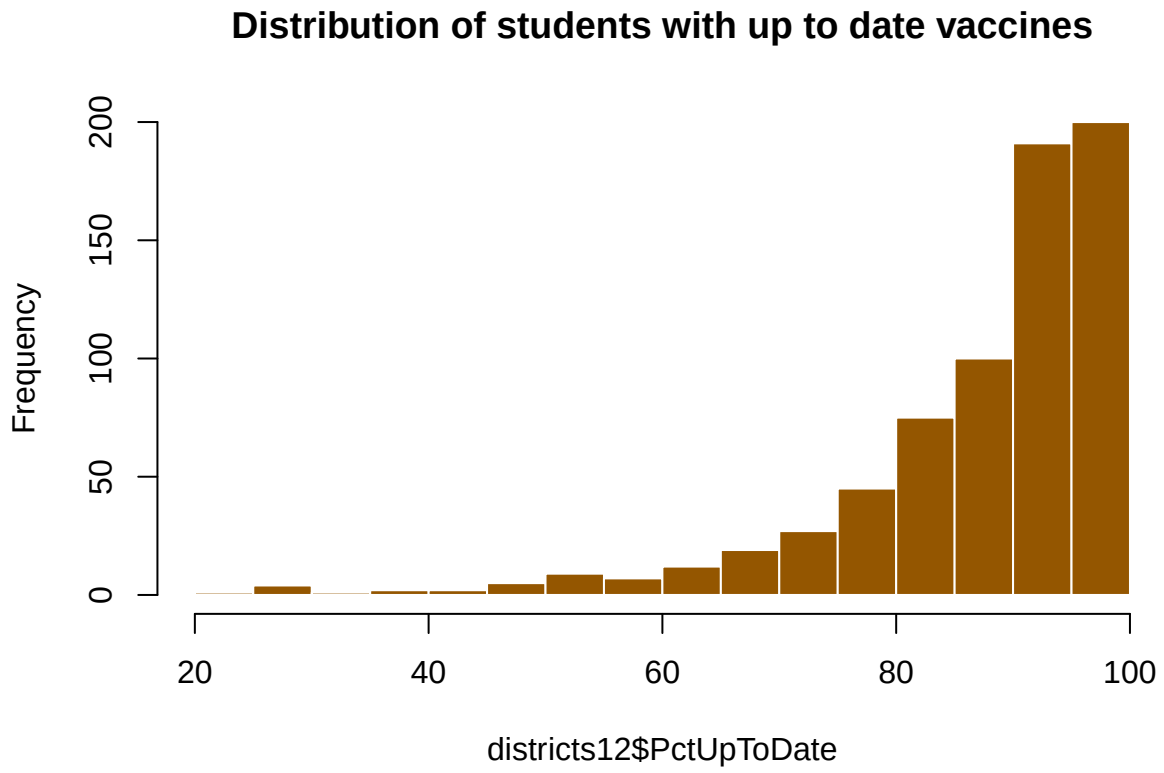
We examined the data from sample of California public school district from the 2013 data collection and tried to see if PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled and TotalSchools could predict wheatear or not a district's reporting was complete. We conducted Bayesian logistic analysis, using all of 5 of them to predict DistrictComplete. The posterior distributions of the coefficients for PctChildPoverty, PctFreeMeal, PctFamilyPoverty (calibrated as log odds) overlapped squarely with 0, suggesting that they were not meaningful predictor of engine. In contrast the HDI for Enrolled and TotalSchools did not overlap with zero. When converted to regular odd, the mean value of the posterior distribution for Enrolled was .99:1 suggesting that for every additional enrolled individual, the district reporting is about 1% more likely to be complete. Converted to regular odd, the mean value of the posterior distribution for TotalSchools was 1.22:1 suggesting that for every additional enrolled individual, the district reporting is about 1.22% more likely to be complete. The confusion matrix showed overall error rate of 0.06% indicating that the logistic model was good.

6.What variables predict the percentage of all enrolled students with completely up-to-date vaccines?

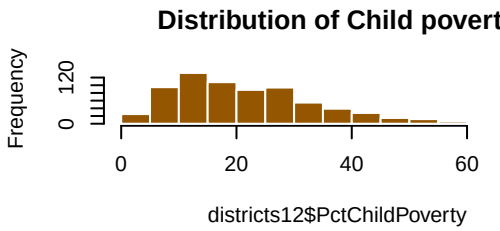
```
class(districts12$PctUpToDate)
```

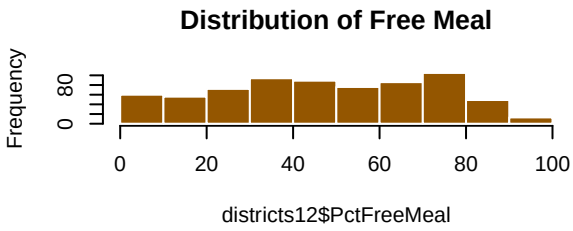
```
## [1] "numeric"
```

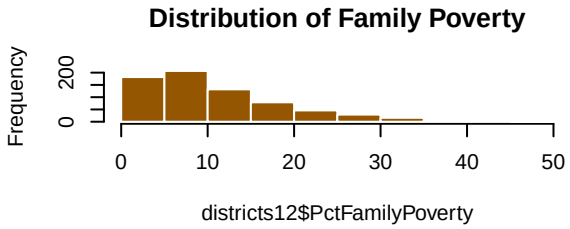
```
hist(districts12$PctUpToDate, main = "Distribution of students with up to date vaccines"
     , border = "white",
     col = "#945600")
```

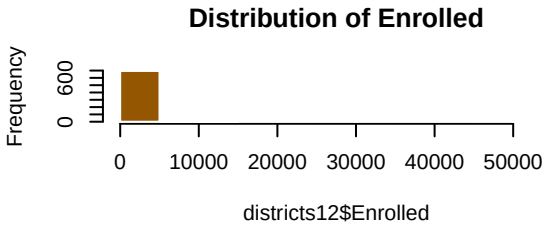


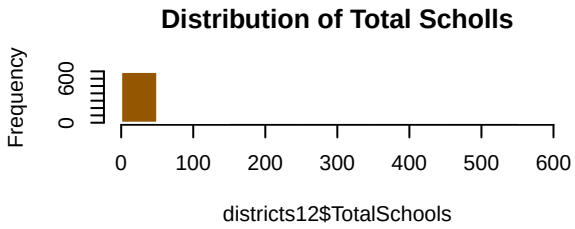

```
par(mfrow= c(3,2))
hist(districts12$PctChildPoverty, main = "Distribution of Child poverty"
     , border = "white",
     col = "#945600")
hist(districts12$PctFreeMeal, main = "Distribution of Free Meal"
     , border = "white",
     col = "#945600")
hist(districts12$PctFamilyPoverty, main = "Distribution of Family Poverty"
     , border = "white",
     col = "#945600")
hist(districts12$Enrolled, main = "Distribution of Enrolled"
     , border = "white",
     col = "#945600")
hist(districts12$TotalSchools, main = "Distribution of Total Scholls"
     , border = "white",
     col = "#945600")
#districts12$Enrolled <- districts12$Enrolled/100
```












```
## Correlate the variables to find the connection
newdis <- districts12[,6:13]
newdist <- newdis[-(2:3)]
head(newdist)
```

	PctUpToDate	PctChildPoverty	PctFreeMeal	PctFamilyPoverty	Enrolled	TotalSchools
118	94	13	28	6	341	5
413	96	25	69	15	271	1
610	61	11	26	2	64	2
163	95	25	66	12	1020	9
550	93	9	22	5	871	8
361	100	55	77	16	33	1

```
cor(newdist)
```

	PctUpToDate	PctChildPoverty	PctFreeMeal	PctFamilyPoverty	Enrolled	TotalSchools
PctUpToDate	1.0000000	0.2251438	0.2606492	0.2534934	0.0601608	0.0463027
PctChildPoverty	0.2251438	1.0000000	0.7562022	0.8641509	0.0265675	0.0212344
PctFreeMeal	0.2606492	0.7562022	1.0000000	0.7286677	0.0649288	0.0593207
PctFamilyPoverty	0.2534934	0.8641509	0.7286677	1.0000000	0.0388149	0.0318555
Enrolled	0.0601608	0.0265675	0.0649288	0.0388149	1.0000000	0.9940612
TotalSchools	0.0463027	0.0212344	0.0593207	0.0318555	0.9940612	1.0000000

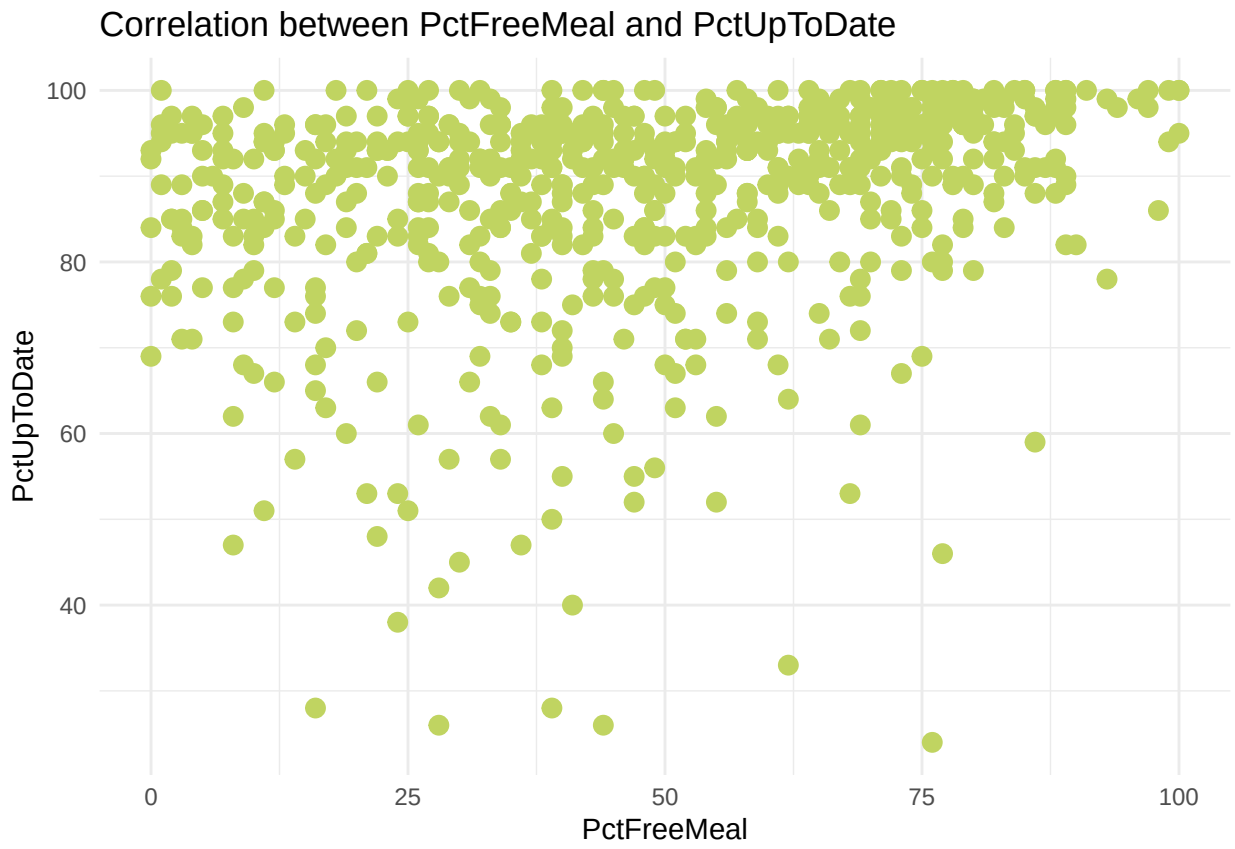
The Pearson Product-Moment Correlation, aka "r", expresses the association between two metric variables on a scale of -1 to +1. Values of r near -1 or +1 are strong; values near 0 are weak.

In the example above we are calculating a cross product between scores and taking the average of the sum of a cross product in order to find out a measure of association between two metric variables. It starts off as a quantity called the covariance. And then we standardize it into a scale that goes from minus 1 to plus 1. And that is r, the Pearson product-moment correlation.

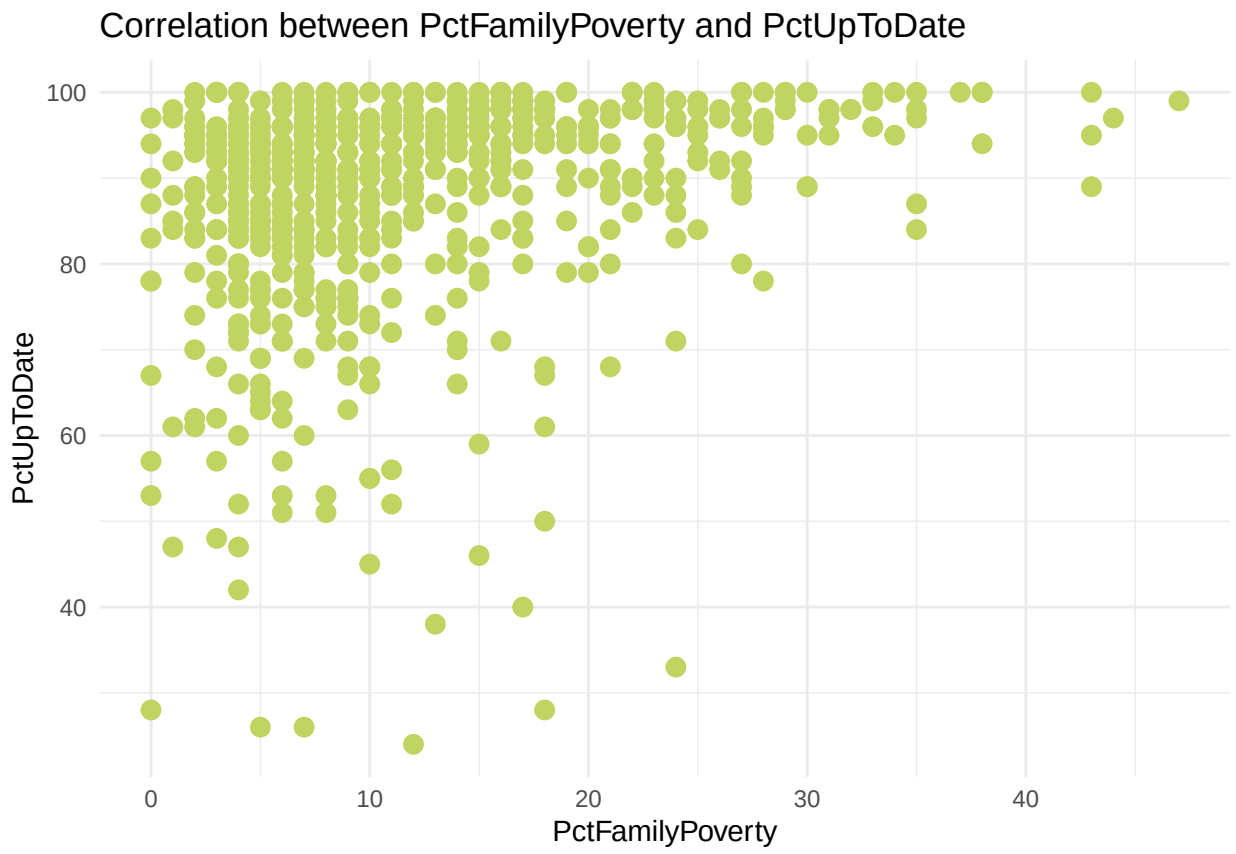
We are going to run a multiple regression analysis on the district12 data with lm(), using PctUpToDate as the dependent variable and PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools as predictors. The outcome variable is skewed.

If we take a look at the correlation matrix, we will see 1's along the diagonal representing a correlation between a variable and itself. And then it has two halves, the lower triangle and the upper triangle, which are mirror images of one another. They contain the same information. We will be trying to predict the PctUpToDate variable. PctFreeMeal and PctFamilyPoverty has the highest correlation with PctUpToDate, which might be the two best predictors. Looks the poverty might affects up to date vaccination status. Lets examine some correlation plots.

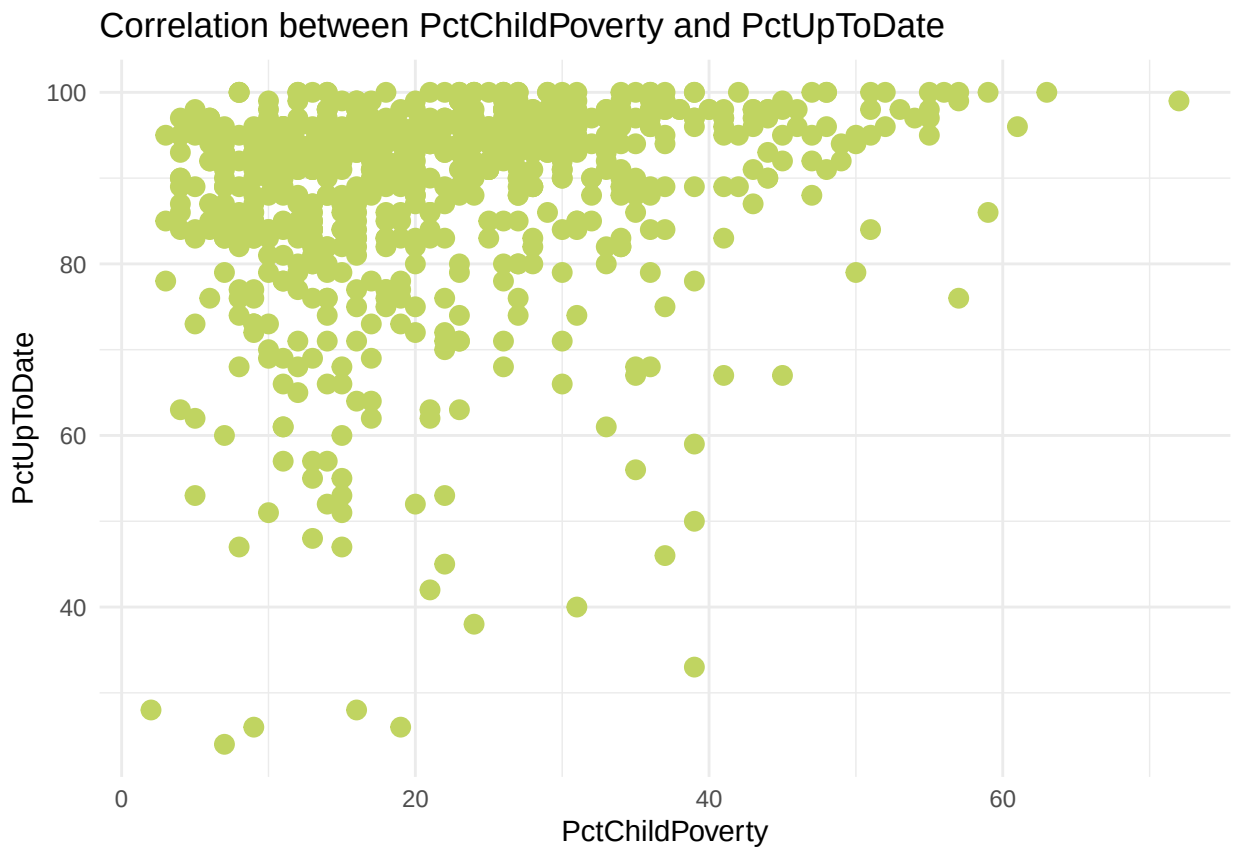
```
ggplot(newdist, aes(x=PctFreeMeal, y=PctUpToDate)) + geom_point(col = "#C0D461", size = 3) + theme_mini
```




```
ggplot(newdist, aes(x=PctFamilyPoverty, y=PctUpToDate)) + geom_point(col = "#C0D461", size = 3) + theme_
```



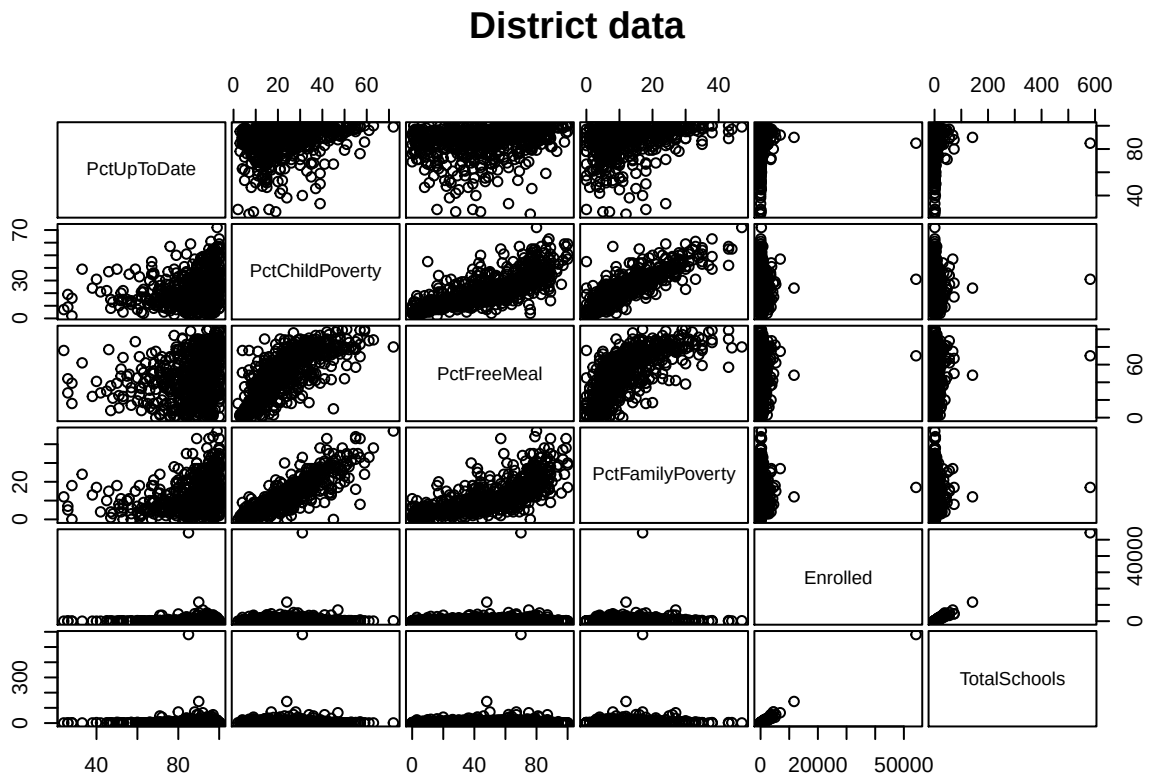

```
ggplot(newdist, aes(x=PctChildPoverty, y=PctUpToDate)) + geom_point(col = "#C0D461", size = 3) + theme_
```



There isn't any big positive or negative correlation,like we saw from the cor matrix.

Lets look at all the variables.

```
pairs(newdist, main = "District data", gap = 1/4)
```



Pairs plot shows us the pattern of correlation using a scatterplot for each pair of variables. If we have a cigar-shaped scatterplot that points from the lower left the upper right, that's an indication of a strong positive variable. That same cigar shape pointing from the upper left down to the lower right indicates an inverted relationship, but again, a strong one. Any weak relationship between two variables and the scatterplot of points is more circular or more disordered. And that would indicate a relatively small value of r that's close to 0.

```
skewness(newdist$PctUpToDate)
```

```
## [1] -2.069381
```

Transformations to reduce skewness:

```
PctUpToDateNew <- sqrt(districts12$PctUpToDate)
PctUpToDateNew1 <- log(districts12$PctUpToDate)
PctUpToDateNew2 <- log10(districts12$PctUpToDate)
PctUpToDateNew3 <- asin(districts12$PctUpToDate)
PctUpToDateNew4 <- atan(districts12$PctUpToDate)

skewness(PctUpToDateNew)
```

```
## [1] -2.580835
```

```
skewness(PctUpToDateNew1)
```

```
## [1] -3.28394
```

```
skewness(PctUpToDateNew2)
```

```
## [1] -3.28394
```

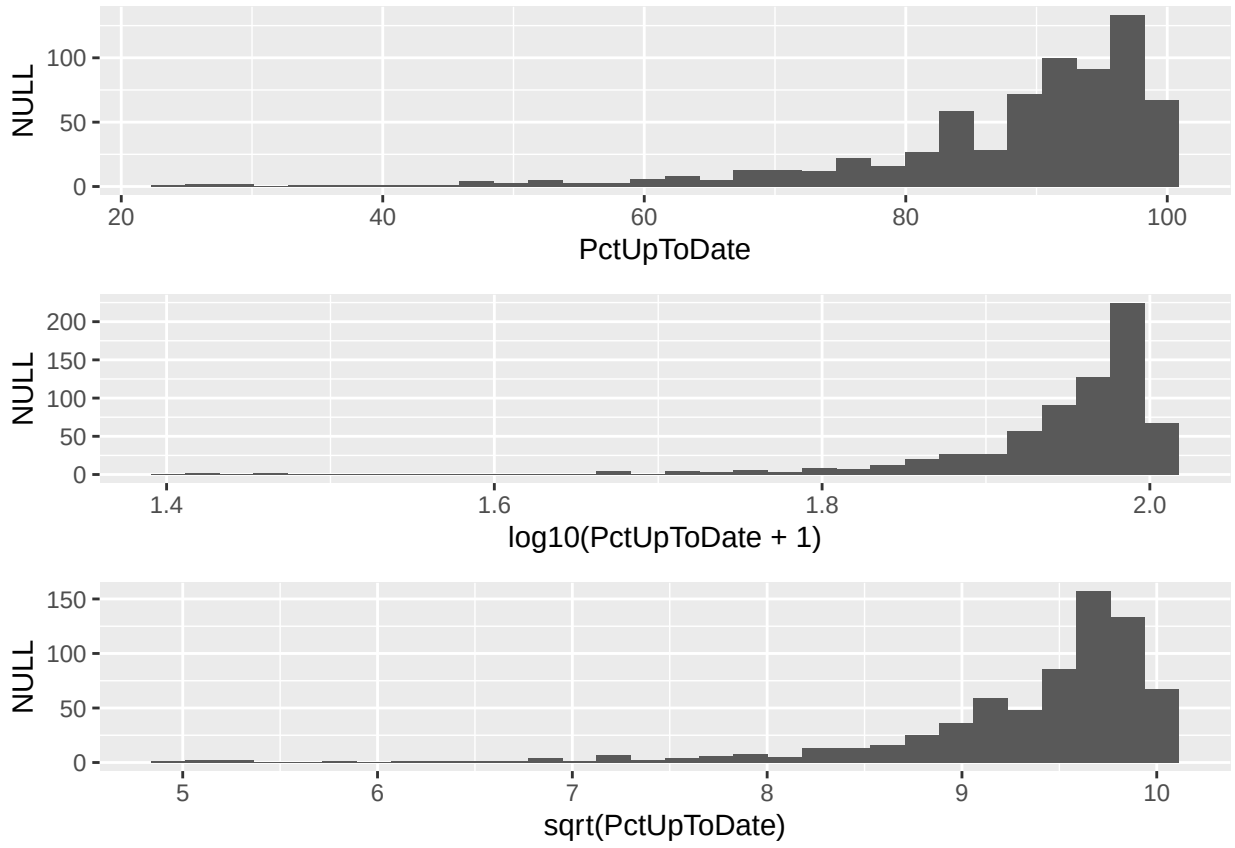
```
skewness(PctUpToDateNew3)
```

```
## [1] NA
```

```
skewness(PctUpToDateNew4)
```

```
## [1] -5.407817
```

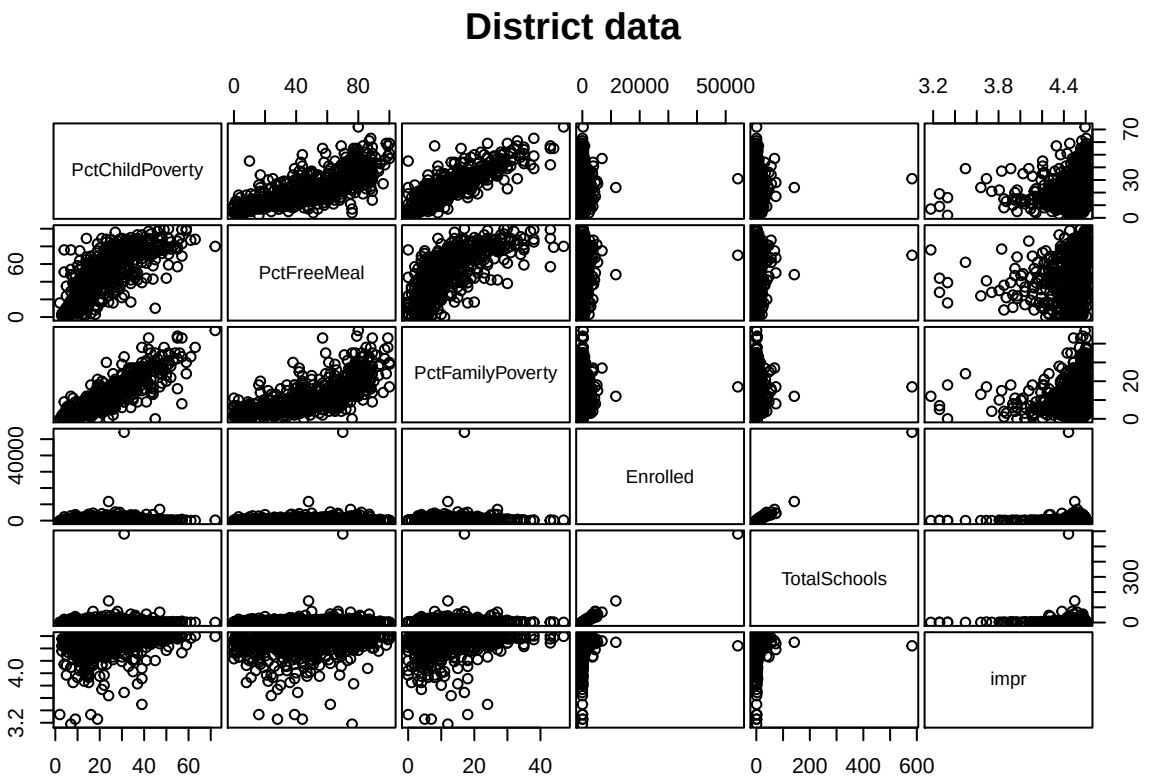
```
p1 <- qplot(x=PctUpToDate, data = districts12)
p2 <-qplot(x = log10(PctUpToDate+1), data = districts12)
p3 <- qplot(x=sqrt(PctUpToDate), data = districts12)
grid.arrange(p1,p2,p3, ncol = 1)
```



```
## Check the correlation with some of the adjusted outcome variables
newdistIM <- newdist[-1]
newdistIM$impr <- PctUpToDateNew1
cor(newdistIM)
```

	PctChildPoverty	PctFreeMeal	PctFamilyPoverty	Enrolled	TotalSchools	impr
PctChildPoverty	1.0000000	0.7562022	0.8641509	0.0265675	0.0212344	0.1945220
PctFreeMeal	0.7562022	1.0000000	0.7286677	0.0649288	0.0593207	0.2210301
PctFamilyPoverty	0.8641509	0.7286677	1.0000000	0.0388149	0.0318555	0.2124643
Enrolled	0.0265675	0.0649288	0.0388149	1.0000000	0.9940612	0.0623193
TotalSchools	0.0212344	0.0593207	0.0318555	0.9940612	1.0000000	0.0508403
impr	0.1945220	0.2210301	0.2124643	0.0623193	0.0508403	1.0000000

```
pairs(newdistIM, main = "District data", gap = 1/4)
```

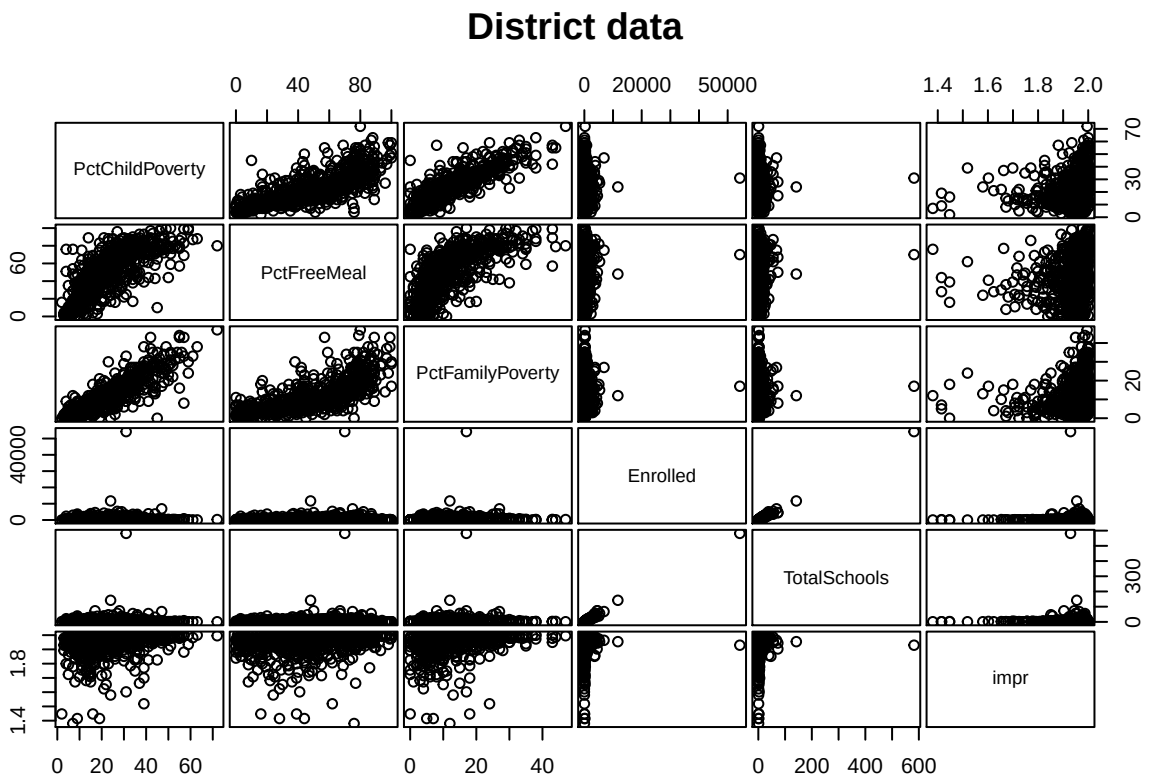



```
newdistIM$impr <- PctUpToDateNew2
cor(newdistIM)
```

	PctChildPoverty	PctFreeMeal	PctFamilyPoverty	Enrolled	TotalSchools	impr
PctChildPoverty	1.0000000	0.7562022	0.8641509	0.0265675	0.0212344	0.1945220
PctFreeMeal	0.7562022	1.0000000	0.7286677	0.0649288	0.0593207	0.2210301
PctFamilyPoverty	0.8641509	0.7286677	1.0000000	0.0388149	0.0318555	0.2124643

	PctChildPoverty	PctFreeMeal	PctFamilyPoverty	Enrolled	TotalSchools	impr
Enrolled	0.0265675	0.0649288	0.0388149	1.0000000	0.9940612	0.0623193
TotalSchools	0.0212344	0.0593207	0.0318555	0.9940612	1.0000000	0.0508403
impr	0.1945220	0.2210301	0.2124643	0.0623193	0.0508403	1.0000000

```
pairs(newdistIM, main = "District data", gap = 1/4)
```



The correlation with the dependent variable is even smaller.

```
regOut <- lm(PctUpToDate~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty+
               Enrolled + TotalSchools, data = districts12)
summary(regOut)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
##     Enrolled + TotalSchools, data = districts12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.917  -3.375   3.115   7.197  18.298
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      82.061208   1.057030  77.634  < 2e-16 ***
## PctChildPoverty  -0.060310   0.079039  -0.763  0.44569
## PctFreeMeal       0.087179   0.028568   3.052  0.00236 **
## PctFamilyPoverty  0.255886   0.111582   2.293  0.02213 *
## Enrolled          0.005814   0.001880   3.092  0.00207 **
## TotalSchools     -0.517295   0.173618  -2.980  0.00299 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12 on 694 degrees of freedom
## Multiple R-squared:  0.09101,    Adjusted R-squared:  0.08446
## F-statistic:  13.9 on 5 and 694 DF,  p-value: 5.839e-13
```

The first section is labeled residuals. Residuals are just the errors of prediction, and the lm() model generates a whole list of them in order to do its somewhat squared prediction errors calculation. In the output we see just a description of the distribution of them. It shows the minimum prediction error, the maximum prediction error, and then the first quartile, third quartile, and the median. They should be normally distributed with a median near zero. The fact that the median is 3.15 suggests that there is some skewness in the residuals, they are not that symmetrically distributed.

In comparison:

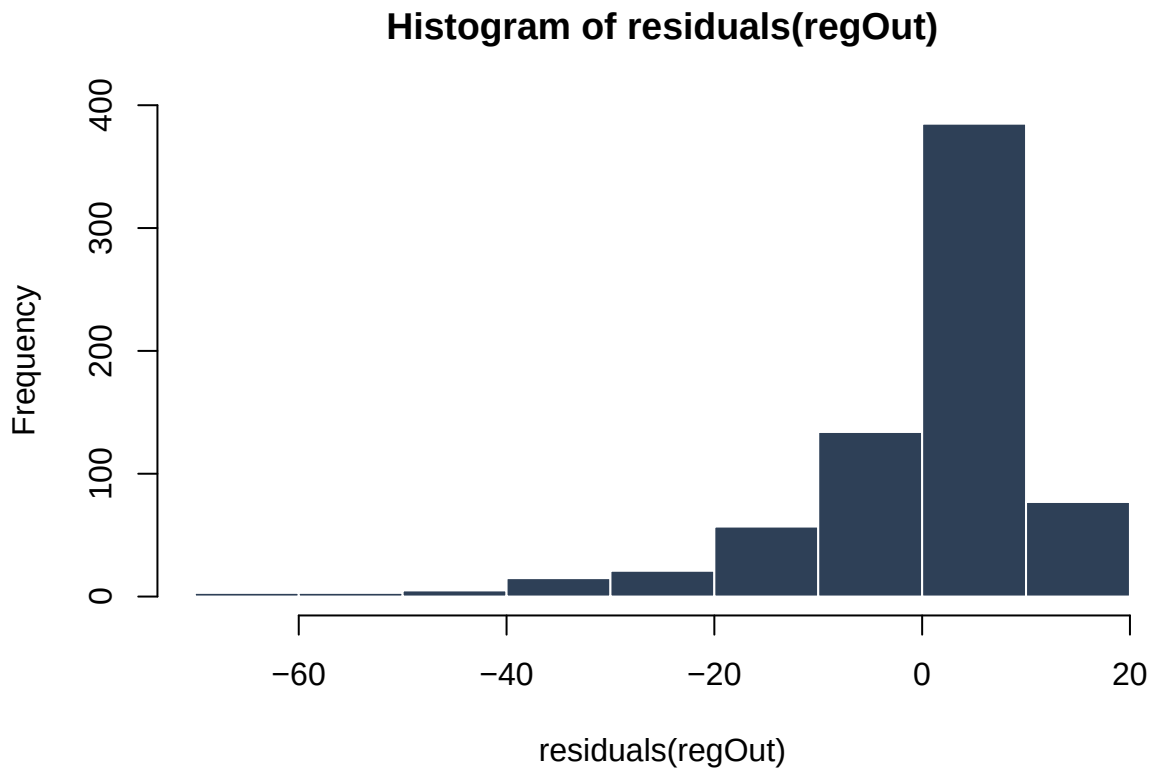
```
regOut1 <- lm(impr~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty+
               Enrolled + TotalSchools, data = newdistIM)
summary(regOut1)
```

```
##
## Call:
## lm(formula = impr ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
##     Enrolled + TotalSchools, data = newdistIM)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57089 -0.01092  0.01947  0.04043  0.09470
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.906e+00  6.793e-03 280.548   <2e-16 ***
## PctChildPoverty  -1.632e-04  5.079e-04  -0.321   0.7481
## PctFreeMeal       4.566e-04  1.836e-04   2.487   0.0131 *
## PctFamilyPoverty  1.169e-03  7.171e-04   1.630   0.1036
## Enrolled          3.074e-05  1.208e-05   2.545   0.0112 *
## TotalSchools     -2.695e-03  1.116e-03  -2.415   0.0160 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07714 on 694 degrees of freedom
## Multiple R-squared:  0.06492,    Adjusted R-squared:  0.05818
## F-statistic: 9.637 on 5 and 694 DF,  p-value: 6.591e-09
```

```
summary(residuals(regOut))
```

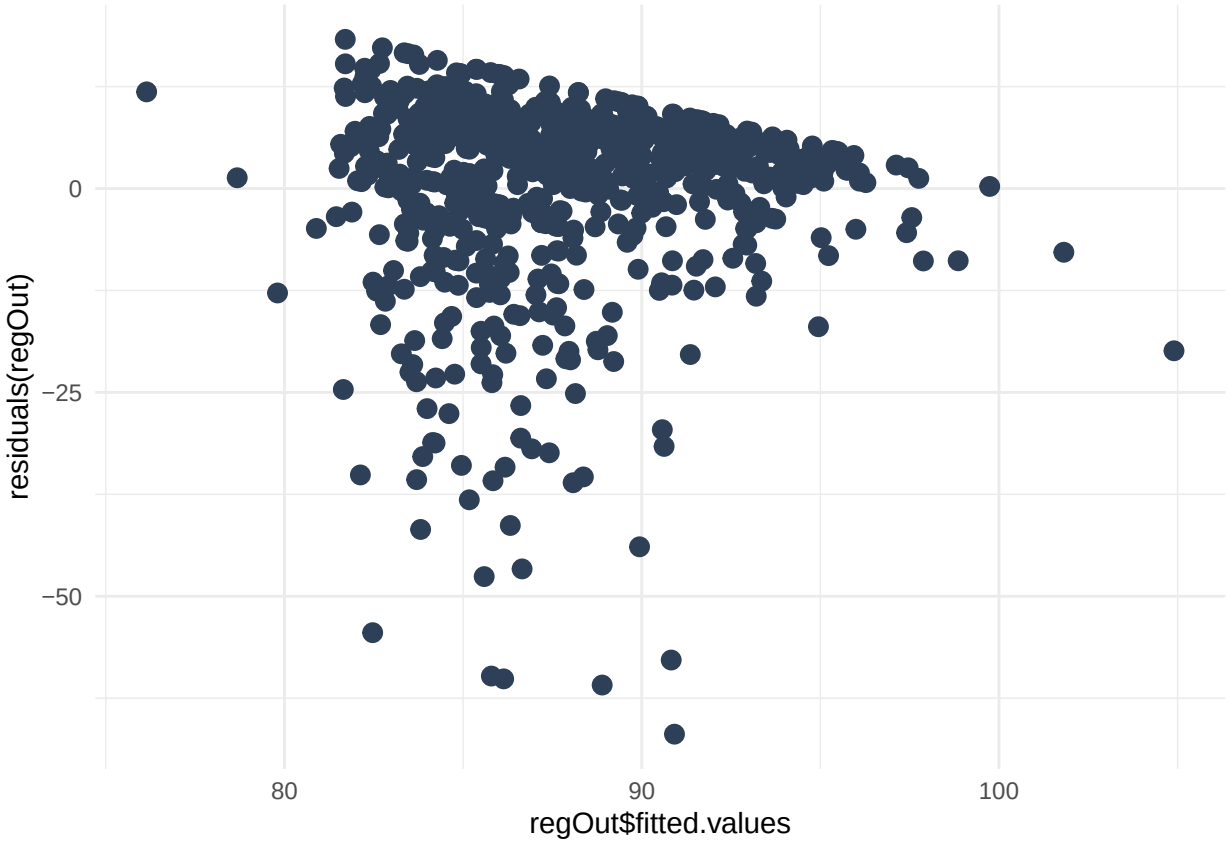
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-66.91679	-3.37482	3.115005	0	7.19665	18.29808

```
## Histogram of the residuals
hist(residuals(regOut)
     , col = "#2E4057"
     , border = "white"
     , main = "Histogram of residuals(regOut)")
```



We want to check and make sure that they are normally distributed and centered on 0 (that undicates that the regression model is doing good). Most of the cases are around 0, and the range that we saw earlier is between -66.9 and 18.29. Some improvement of the model can be done or there might be outliers in the data.

```
ggplot(regOut, aes(x=regOut$fitted.values, y=residuals(regOut))) +
  geom_point(colour = "#2E4057", size = 3) + theme_minimal()
```



The residuals a mostly normaly distributed, some clutered together forming heteroscedastic (cone-shaped) pattern.

We should begin interpreting a regression equation by making sure that the R squared is significant. If we look at the bottom, we see Multiple R squared of 0.09101 and an Adjusted R squared of 0.08446. Adjusted R-squared is slightly penalized relative to the multiple R-squared because we have five predictors. In social science, R squared could be anywhere from about 0.1 to about 0.5. In our case we have really low values, which means that predictors accounted for about .85% variability in the percentage of all enrolled students with completely up-to-date vaccines. That's backed up by the F-test. F-value is 13.9. That's a big F value, the expected F value is 1, so anything that's wildly in excess of 1 is likely to be significant. The degrees of freedom is 5 and 694. So the 5 refers to the number of the predictors, and 694 = 700-5-1, starting with 700 observation, 5 df are lost for the predictors and one for calculating the Y-intercept (F of 5 and 694 is equal to 13,9). The associated p-value is vanishingly small. It certainly is below that conventional threshold of 0.05 so we are rejecting the null hypothesis that R squared in the population is equal to 0. We've passed the omnibus test and we can go on and interpret the coefficients.

The coefficient section begins to show us the key results we need to know. The fist column is "Estimate" and show us the intercept in the first line and the slopes/B-weights on the predictors in the lines bellow. So the upper number, 82.06, is the intercept, and the lower numbers, -0.06, 0,08, 0.25. 0.005 and -0.52 are the B-weights. For each, the estimate is the statistical value of interest. The std.error estimates variability of the underlying sampling distribution. Together, these two values to calculate "t" and an associated null hypothesis test that each estimate coefficient is equal to 0. In all the cases but PctChildPoverty, the value of "t" and the associated probability clearly indicates that we should reject the null hypothesis ($p < .05$).

- Multiple regression analysis on the district12 data with lmBF()

```
regOutBF <- lmBF(PctUpToDate~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty+
               Enrolled + TotalSchools, data = districts12)
regOutBF
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools : 4537547065 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

That confirm that PctChildPoverty is not a good predictor.

```
regOutBF1 <- lmBF(PctUpToDate~  PctFreeMeal + PctFamilyPoverty +
               Enrolled + TotalSchools, data = districts12)
regOutBF1
```

```
## Bayes factor analysis
## --------------
## [1] PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools : 20424863319 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

The Bayes factor is an odd ratio showing the likelihood of the stated alternative hypothesis (model with nonzero weights for the predictors) divided by the likelihood of the null model (intercept only model).

The results shows that the odds are in favor of the alternative hypotheses(very strong posite evidence that PctFreeMeal, PctFamilyPoverty, Enrolled and TotalSchools are nonzero), in the sense that the model containing those variables as predictors is hugely favored over a model that only contains the Y-intercept.

Taken together, the result of the conventional analysis and the Bayesian analysis indicate that we we do have a solid predictive model with those 4 variables.

- lmBF() with the options posterior=TRUE and iterations=10000

```
regOutMCMC <- lmBF(PctUpToDate~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
               Enrolled + TotalSchools, data = districts12, 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                87.758175 0.45346 0.0045346      0.0045346
## PctChildPoverty   -0.059944 0.07782 0.0007782      0.0007782
## PctFreeMeal        0.083898 0.02776 0.0002776      0.0002757
## PctFamilyPoverty   0.248700 0.11097 0.0011097      0.0011097
## Enrolled           0.005576 0.00184 0.0000184      0.0000184
## TotalSchools      -0.496032 0.16988 0.0016988      0.0016988
## sig2             143.972803 7.69152 0.0769152      0.0769152
## g                  0.055604 0.05243 0.0005243      0.0005243
##
## 2. Quantiles for each variable:
##
##                       2.5%        25%        50%        75%     97.5%
## mu                86.877409  87.448089  87.750993  88.070183  88.65776
## PctChildPoverty   -0.211502  -0.111887  -0.060053  -0.006962   0.08975
## PctFreeMeal        0.029493   0.065273   0.083947   0.102438   0.13824
## PctFamilyPoverty   0.036393   0.172970   0.249286   0.323087   0.46286
## Enrolled           0.001901   0.004346   0.005567   0.006828   0.00918
## TotalSchools      -0.827438  -0.611497  -0.494805  -0.383120  -0.15829
## sig2             129.525315 138.647188 143.717084 149.032266 159.32581
## g                  0.014815   0.027995   0.041593   0.064596   0.17763
```

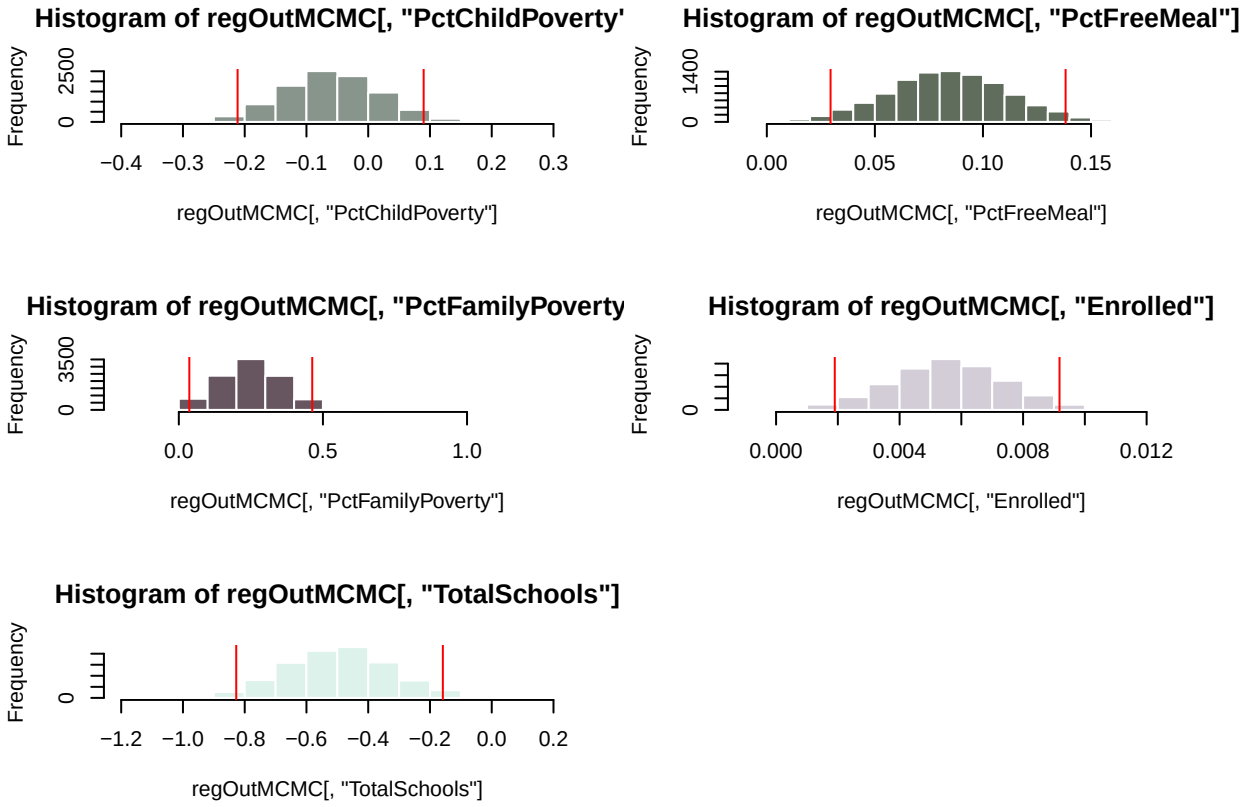
In the command above, we run lmBF() with posterior = TRUE and iterations= 10000 to sample from the posterior distribution using MCMC technique. Looking at the MCMC output first, we see both the means of the respective distribution and the 95% HDI.

If we compare the Bayesian output and the LM procedure, we need to look at the mean column. The values are really close (-0.059 compared to -0.060; 0.083 compared to 0.087; 0.248 compared to 0.255; 0.005 compared to 0.005; -0.496 compared to -0.517). The top line of the tables is labeled mu - the grand mean for the data set. There are two specific columns: Naive SE and Time-series SE. SE stands for standard error (indication of dispersion in a sampling distribution). Naive SE calculates the standard error in the normal way. The Time-series SE accounts for the fact that there might be some, autocorrelation among the different estimates of the standard error.

The second table is overview of the highest density interval. We have the 2.5% and 97.5% boundaries of the HDI for each of the B-weights. These boundaries mark the edges of the central region of the posterior distribution for each B-weight. sig2 is the error variance in the regression equation. And we can use that error variance to calculate a distribution of r-squared values.

```
par(mfrow = c(3,2))
hist(regOutMCMC[,"PctChildPoverty"], col = "#88958D", border = "white")
abline(v=quantile(regOutMCMC[,"PctChildPoverty"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC[,"PctChildPoverty"], c(0.975)), col = "red")

hist(regOutMCMC[,"PctFreeMeal"], col = "#606D5D", border = "white")
abline(v=quantile(regOutMCMC[,"PctFreeMeal"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC[,"PctFreeMeal"], c(0.975)), col = "red")

hist(regOutMCMC[,"PctFamilyPoverty"], col = "#675660", border = "white")
abline(v=quantile(regOutMCMC[,"PctFamilyPoverty"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC[,"PctFamilyPoverty"], c(0.975)), col = "red")

hist(regOutMCMC[,"Enrolled"], col = "#D3CDD7", border = "white")
```

```
abline(v=quantile(regOutMCMC[,"Enrolled"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC[,"Enrolled"], c(0.975)), col = "red")

hist(regOutMCMC[,"TotalSchools"], col = "#DDF2EB", border = "white")
abline(v=quantile(regOutMCMC[,"TotalSchools"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC[,"TotalSchools"], c(0.975)), col = "red")
```



We can see pretty symmetric distribution and the lower and upper bounds of the HDI. The intervalPctChildPoverty does not overlap with 0, confirming the insignificance of that variable. The rest of the intervals does not overlap with 0 providing evidence that the population value of those B-weights credibly differ from 0.

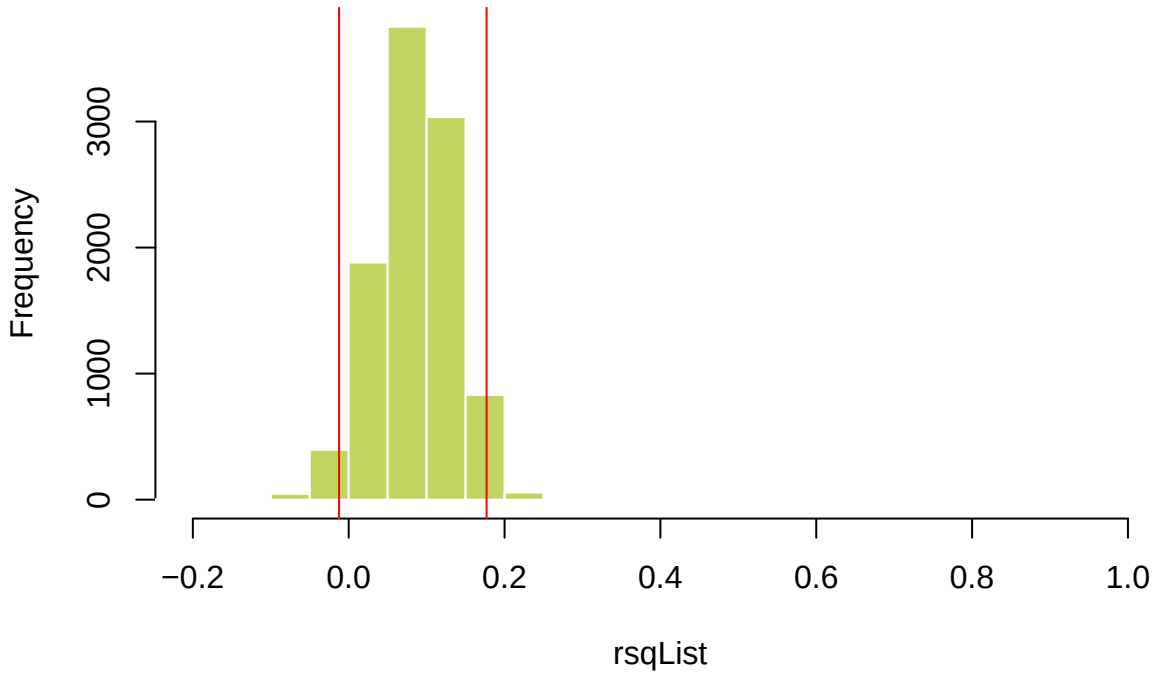
```
var(districts12$PctUpToDate)
```

```
## [1] 157.3808
```

```
## Calculate a distribution of r-squared values
rsqList <- 1 - (regOutMCMC[,"sig2"] / 157.3808)
hist(rsqList
     , xlim = c(-0.2,1)
     , main = ""
     , col = "#C0D461", border = "white")
mean(rsqList) # overall mean R-square is 0.085
```

```
## [1] 0.08519462
```

```
## Draw boundaries of the 95% HDI
abline(v=quantile(rsqList,c(0.025)), col="red")
abline(v=quantile(rsqList,c(0.975)), col="red")
```



The mean value of this distribution came out to 0.085, which is slightly higher than the adjusted R-square of 0.08446. The mean B-weight on the variables are smaller in magnitude for the Bayesian analysis, than for the conventional analysis. Bayesian analysis finds that variables are not as strong predictors as the conventional analysis indicated. Bayesian model can give us a clear-eyed view of the likely range of the possibilities for the predictive strength of our model. In the underlying population, it is credible for us to expect R-square as low as about -0.01 or as high as about 0.177, with the most likely value pf R-square in that central region surrounding 0.085.

These results strengthen the conclusions we made. Using the "frequentist" (traditional) null hypothesis test, we rejected the null hypothesis for all four predictors as well as the overall R-squared. The Highest Density Intervals (HDIs) from the MCMC output showed estimates for the coefficients and R squared that concur with the frequentist model. The Bayes Factor overwhelmingly favored a model that includes the four predictors.

```
gvlma::gvlma(x = lm(impr~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty+
               Enrolled + TotalSchools, data = newdistIM))
```

```
##
## Call:
## lm(formula = impr ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
##     Enrolled + TotalSchools, data = newdistIM)
##
```

```
## Coefficients:
##      (Intercept)   PctChildPoverty       PctFreeMeal  PctFamilyPoverty
##        1.906e+00        -1.631e-04         4.566e-04         1.169e-03
##         Enrolled      TotalSchools
##        3.074e-05        -2.695e-03
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma::gvlma(x = lm(impr ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +      Enrolled + Total
##
##                      Value p-value                   Decision
## Global Stat        9252.6883  0.0000 Assumptions NOT satisfied!
## Skewness           1366.3680  0.0000 Assumptions NOT satisfied!
## Kurtosis           7883.2895  0.0000 Assumptions NOT satisfied!
## Link Function         0.4988  0.4800    Assumptions acceptable.
## Heteroscedasticity    2.5319  0.1116    Assumptions acceptable.
```

Summary of the results

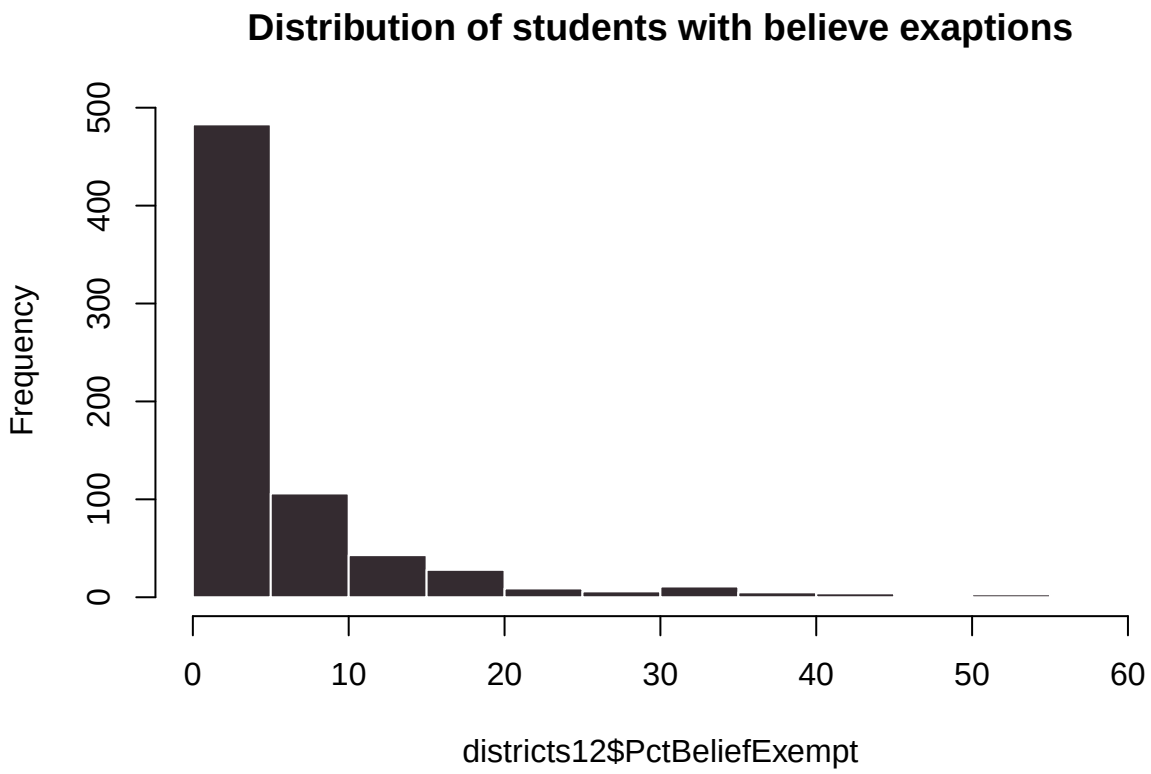
We tested a model that used five variables(PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools) to predict completely up-to-dates vaccines. A Bayesian analysis of this model showed a mean posterior estimate for R-squared of 0.085, with the highest density interval ranging from roughly -0.01 to 0.177. The traditional analysis confirmed this result with a slightly more pesimistic R-squared of 0.084. The F-test on this value was $F(5, 694)=13.9$, $p<.05$, so we reject the null hypothesis that R-squared was equal to zero. Only four predictors were significant with B-weights of 0.087(PctFreeMeal), 0.255 (PctFamilyPoverty), 0.005(Enrolled), -0.517(TotalSchools). The Bayes factor of 20424863319 was strongly in favor of the four-predictor model (in comparison with an intercept-only model). We have to mention that R-square is really close to 0, and its HDI with 0 which makes the model not that good.

7. What variables predict the percentage of all enrolled students with belief exceptions?

```
class(districts12$PctBeliefExempt)
```

```
## [1] "numeric"
```

```
hist(districts12$PctBeliefExempt, main = "Distribution of students with believe exaptions"
     , border = "white",
     col = "#342B30")
```



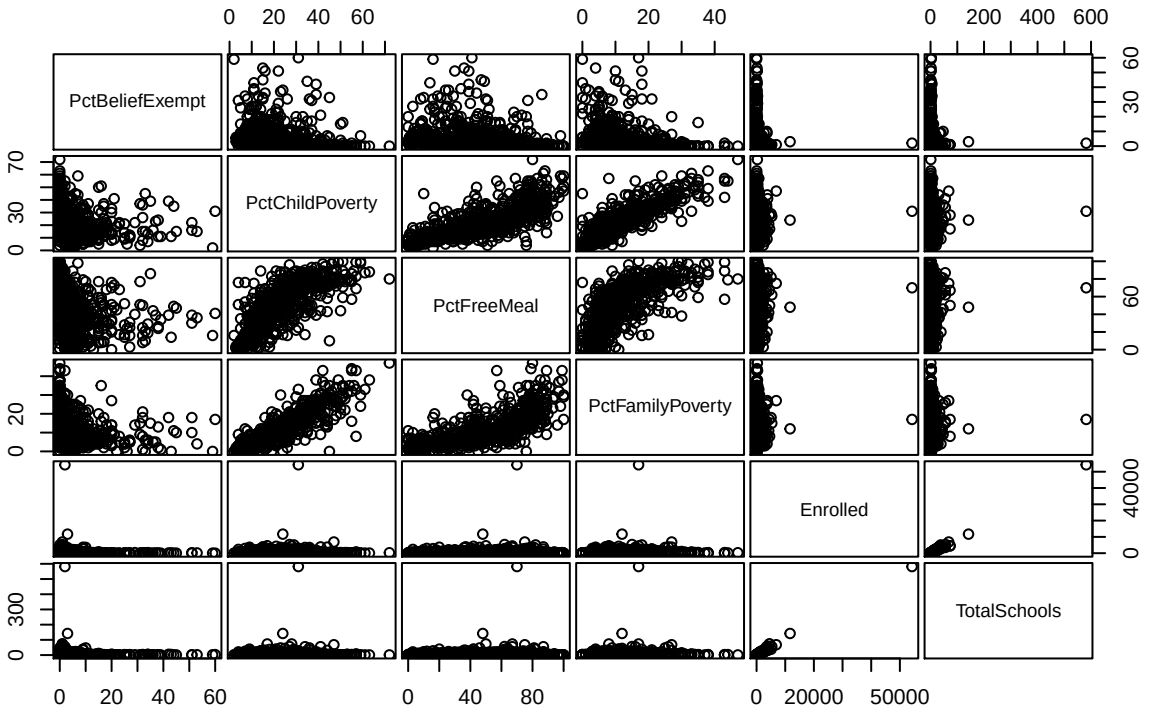

```
dis <- districts12[,8:13]
cor(dis)
```

	PctBeliefExempt	PctChildPoverty	PctFreeMeal	PctFamilyPoverty	Enrolled	TotalSchools
PctBeliefExempt	1.0000000	-0.1980644	-0.3078808	-0.2411090	-0.0902692	-0.0809891
PctChildPoverty	-0.1980644	1.0000000	0.7562022	0.8641509	0.0265675	0.0212344
PctFreeMeal	-0.3078808	0.7562022	1.0000000	0.7286677	0.0649288	0.0593207
PctFamilyPoverty	-0.2411090	0.8641509	0.7286677	1.0000000	0.0388149	0.0318555
Enrolled	-0.0902692	0.0265675	0.0649288	0.0388149	1.0000000	0.9940612
TotalSchools	-0.0809891	0.0212344	0.0593207	0.0318555	0.9940612	1.0000000

There is moderate negative correlation between PctBeliefExempt and PctFreeMeal, PctFamilyPoverty - they might be good predictors.

Lets look at all the variables.

```
pairs(dis, main = "District data", gap = 1/4)
```



We see similar picture to the one before - wealth might be of a factor for change in the outcome variable.

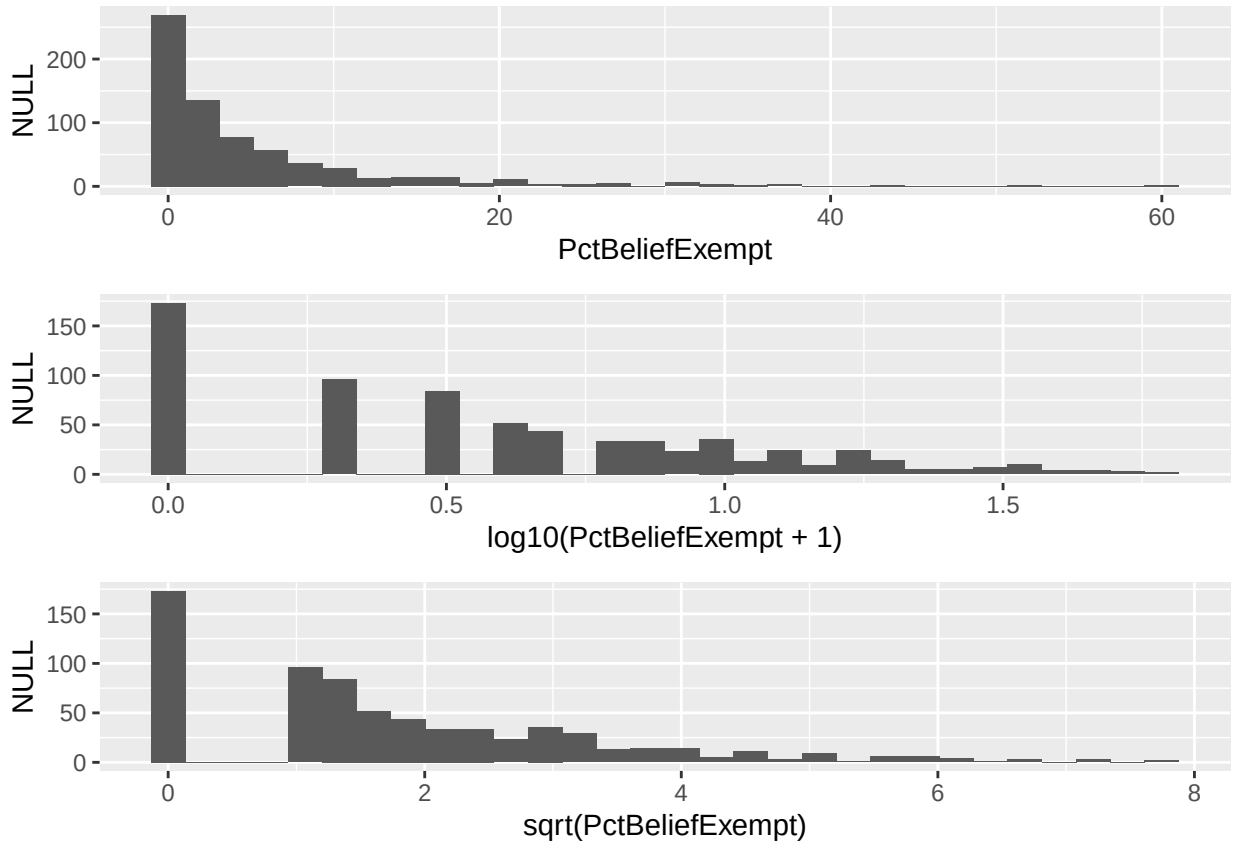
```
skewness(districts12$PctBeliefExempt)
```

```
## [1] 2.894354
```

```
## Check if transforming help reduce skewness and improve relationship between the outcome variable and
PctBeliefExemptNew <- sqrt(districts12$PctBeliefExempt)
skewness(PctBeliefExemptNew)
```

```
## [1] 0.9623478
```

```
p1 <- qplot(x=PctBeliefExempt, data = districts12)
p2 <-qplot(x = log10(PctBeliefExempt+1), data = districts12)
p3 <- qplot(x=sqrt(PctBeliefExempt), data = districts12)
grid.arrange(p1,p2,p3, ncol = 1)
```




```
distr <- dis[-1]
distr$improvedVar <-PctBeliefExemptNew
cor(distr)
```

	PctChildPoverty	PctFreeMeal	PctFamilyPoverty	Enrolled	TotalSchools	improvedVa
PctChildPoverty	1.0000000	0.7562022	0.8641509	0.0265675	0.0212344	-0.333264
PctFreeMeal	0.7562022	1.0000000	0.7286677	0.0649288	0.0593207	-0.442867
PctFamilyPoverty	0.8641509	0.7286677	1.0000000	0.0388149	0.0318555	-0.368763
Enrolled	0.0265675	0.0649288	0.0388149	1.0000000	0.9940612	-0.067429
TotalSchools	0.0212344	0.0593207	0.0318555	0.9940612	1.0000000	-0.056091
improvedVar	-0.3332641	-0.4428672	-0.3687630	-0.0674291	-0.0560910	1.000000
We have higher cor	relation here from	before, the t	ransformation of th	e outcome var	iable helped.	

```
regOutS <- lm(PctBeliefExempt~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty+
                Enrolled + TotalSchools, data = dis)
summary(regOutS)
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFreeMeal +
##     PctFamilyPoverty + Enrolled + TotalSchools, data = dis)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -11.802  -4.008  -2.054   0.611  52.534
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.142821   0.724429  14.001  < 2e-16 ***
## PctChildPoverty   0.143606   0.054169   2.651  0.00821 **
## PctFreeMeal      -0.115209   0.019579  -5.884 6.21e-09 ***
## PctFamilyPoverty -0.176548   0.076472  -2.309  0.02126 *
## Enrolled         -0.002542   0.001288  -1.973  0.04893 *
## TotalSchools      0.211611   0.118988   1.778  0.07577 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.227 on 694 degrees of freedom
## Multiple R-squared:  0.1136, Adjusted R-squared:  0.1072
## F-statistic: 17.79 on 5 and 694 DF,  p-value: < 2.2e-16
```

The first section is labeled residuals. Residuals are just the errors of prediction, and the lm() model generates a whole list of them in order to do its somewhat squared prediction errors calculation. In the output we see just a description of the distribution of them. It shows the minimum prediction error, the maximum prediction error, and then the first quartile, third quartile, and the median. They should be normally distributed with a median near zero. The fact that the median is -2.053 suggests that there is some skewness in the residuals, they are not that symmetrically distributed.

```
regOutTransf <- lm(improvedVar~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty+
               Enrolled + TotalSchools, data = distr)
summary(regOutTransf)
```

```
##
## Call:
## lm(formula = improvedVar ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
##     Enrolled + TotalSchools, data = distr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1690 -0.8873 -0.2033  0.5782  5.6751
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.0278777  0.1225404  24.709  < 2e-16 ***
## PctChildPoverty   0.0181428  0.0091629   1.980  0.04809 *
## PctFreeMeal      -0.0254012  0.0033118  -7.670 5.84e-14 ***
## PctFamilyPoverty -0.0356641  0.0129356  -2.757  0.00599 **
## Enrolled         -0.0005143  0.0002179  -2.360  0.01855 *
## TotalSchools      0.0453546  0.0201274   2.253  0.02455 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.392 on 694 degrees of freedom
## Multiple R-squared:  0.2126, Adjusted R-squared:  0.2069
## F-statistic: 37.47 on 5 and 694 DF,  p-value: < 2.2e-16
```

After transformation of the outcome variable all the predictors are statistically significant with p-value $<$ .05.

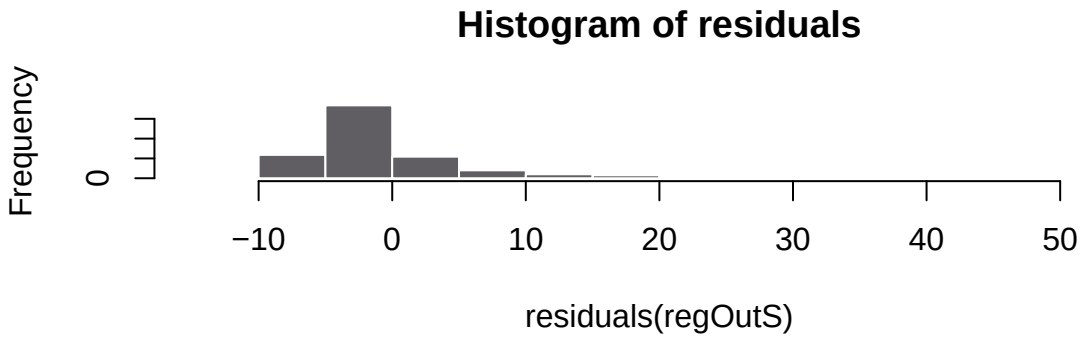
```
summary(residuals(regOutS))
```

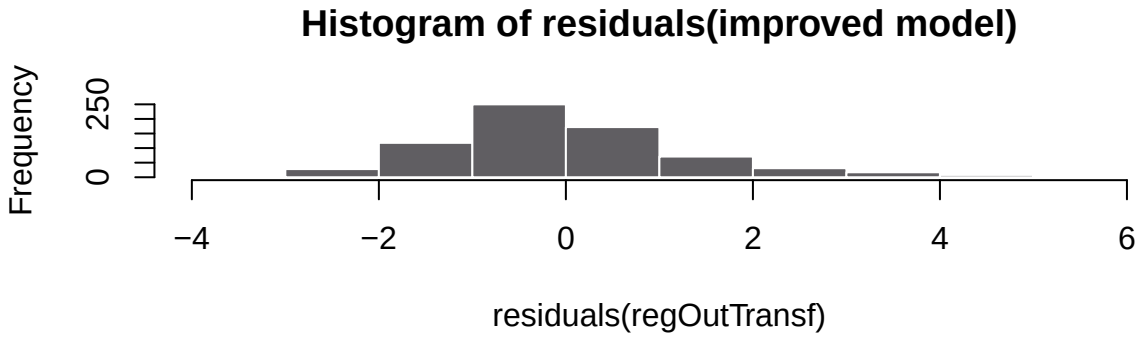
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-11.80174	-4.007793	-2.054492	0	0.6113103	52.5336

```
summary(residuals(regOutTransf))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.168981	-0.8873348	-0.203283	0	0.5781556	5.675053

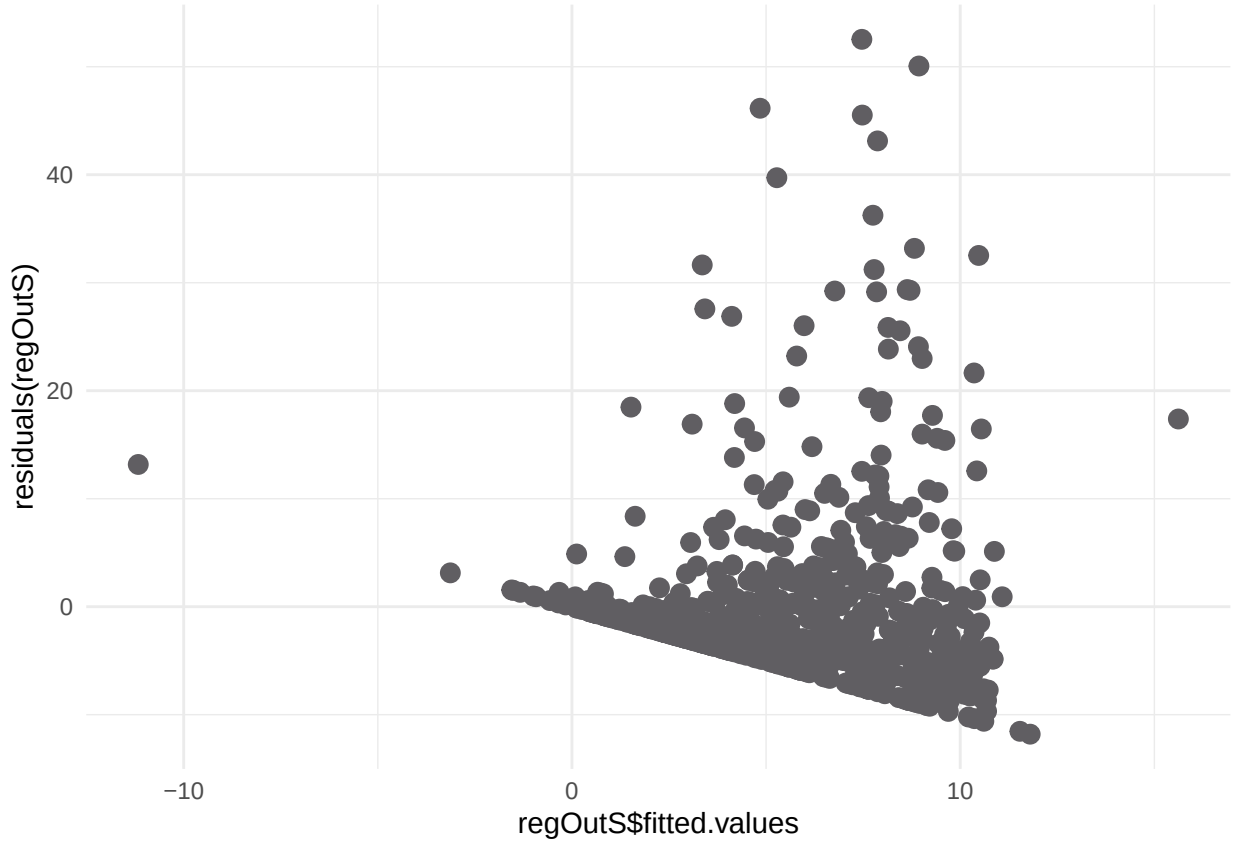
```
par(mfrow = c(2,1))
## Histogram of the residuals
hist(residuals(regOutS)
     , col = "#605E62"
     , border = "white"
     , main = "Histogram of residuals")

## Histogram of the residuals
hist(residuals(regOutTransf)
     , col = "#605E62"
     , border = "white"
     , main = "Histogram of residuals(improved model)")
```



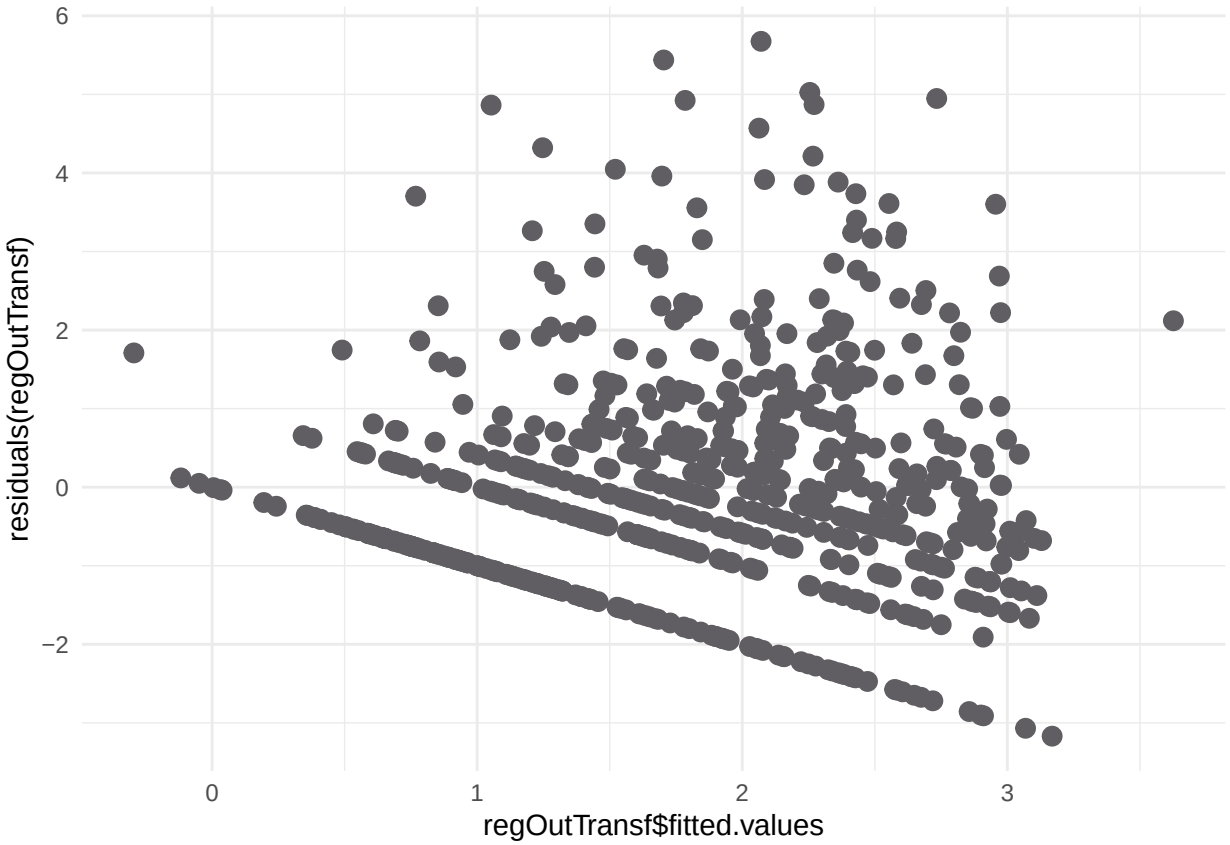


The improved model shows almost normaly distributed residuals.

```
ggplot(regOutS, aes(x=regOutS$fitted.values, y=residuals(regOutS))) +
  geom_point(colour = "#605E62", size = 3) + theme_minimal()
```




```
ggplot(regOutTransf, aes(x=regOutTransf$fitted.values, y=residuals(regOutTransf))) +
  geom_point(colour = "#605E62", size = 3) + theme_minimal()
```



In the first plot the residuals are clutered together forming heteroscedastic (cone-shaped) pattern. We can see that they are more spread out in the second plot.

We should begin interpreting a regression equation by making sure that the R squared is significant. If we look at the bottom, we see Multiple R squared of 0.1136 and an Adjusted R squared of 0.1072(0.2126 and 0.2069 for the improved model). Adjusted R-squared is slightly penalized relative to the multiple R-squared because we have five predictors. In social science, R squared could be anywhere from about 0.1 to about 0.5. In our case we have really normal values, which means that predictors accounted for about 10.7% variability in the percentage of all enrolled students with belief exceptions. That's backed up by the F-test. F-value is 17.79(37.47 for the improved model). That's a big F value, the expected F value is 1, so anything that's wildly in excess of 1 is likely to be significant. The degrees of freedom is 5 and 694. So the 5 refers to the number of the predictors, and 694 = 700-5-1, starting with 700 observation, 5 df are lost for the predictors and one for calculating the Y-intercept (F of 5 and 694 is equal to 17.79). The associated p-value is vanishingly small(2.2e-16. It certainly is below that conventional threshold of 0.05 so we are rejecting the null hypothesis that R squared in the population is equal to 0. We've passed the omnibus test and we can go on and interpret the coefficients.

The coefficient section begins to show us the key results we need to know. The fist column is "Estimate" and show us the intercept in the first line and the slopes/B-weights on the predictors in the lines bellow. So the upper number, 10.14, is the intercept, and the lower numbers are the B-weights. For each, the estimate is the statistical value of interest. The std.error estimates variability of the underlying sampling distribution. Together, these two values to calculate "t" and an associated null hypothesis test that each estimate coefficient is equal to 0. In all the cases but TotalSchools in the first model, the value of "t" and the associated probability clearly indicates that we should reject the null hypothesis ($p < .05$). All values of "t" in the second model are significant.

- Multiple regression analysis on the district12 data with lmBF()

```
regOutBF1 <- lmBF(PctBeliefExempt~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty+
               Enrolled + TotalSchools, data = dis)
regOutBF1
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools : 2.122793e+13 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
regOutBF2 <- lmBF(improvedVar~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty+
               Enrolled + TotalSchools, data = distr)
regOutBF2
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools : 5.413776e+30 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

The Bayes factor is an odd ratio showing the likelihood of the stated alternative hypothesis (model with nonzero weights for the predictors) divided by the likelihood of the null model (intercept only model).

The results shows that the odds are in favor of the alternative hypotheses(very strong posite evidence that PctFreeMeal, PctFamilyPoverty, Enrolled and PctChildPoverty are nonzero), in the sense that the model containing those variables as predictors is hugely favored over a model that only contains the Y-intercept.

Taken together, the result of the conventional analysis and the Bayesian analysis indicate that we we do have a solid predictive model with those 4 variables. The second model all the variables are good predictors.

- lmBF() with the options posterior=TRUE and iterations=10000

```
regOutMCMC1 <- lmBF(PctBeliefExempt~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
               Enrolled + TotalSchools, data = dis, posterior = TRUE,
                   iterations = 10000)
summary(regOutMCMC1)
```

```
##
## 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                 5.690333 0.315826 3.158e-03      3.158e-03
## PctChildPoverty    0.138990 0.053247 5.325e-04      5.325e-04
## PctFreeMeal       -0.111601 0.019147 1.915e-04      1.955e-04
## PctFamilyPoverty  -0.170300 0.075590 7.559e-04      7.559e-04
## Enrolled          -0.002439 0.001279 1.279e-05      1.279e-05
## TotalSchools       0.202887 0.118094 1.181e-03      1.181e-03
## sig2              67.732754 3.637740 3.638e-02      3.638e-02
## g                  0.062580 0.064521 6.452e-04      6.579e-04
##
## 2. Quantiles for each variable:
##
##                        2.5%       25%       50%       75%      97.5%
## mu                 5.075626  5.475445  5.682951  5.904196  6.319e+00
## PctChildPoverty    0.034371  0.103043  0.139467  0.174900  2.430e-01
## PctFreeMeal       -0.148944 -0.124650 -0.111679 -0.098590 -7.410e-02
## PctFamilyPoverty  -0.316660 -0.221649 -0.171161 -0.118398 -2.009e-02
## Enrolled          -0.004935 -0.003305 -0.002424 -0.001586  4.418e-05
## TotalSchools      -0.026698  0.123777  0.201563  0.282506  4.339e-01
## sig2              60.984659 65.204494 67.640058 70.112482  7.527e+01
## g                  0.016955  0.031890  0.046551  0.071870  2.070e-01
```

```
## Improved model
regOutMCMC2 <- lmBF(improvedVar~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
               Enrolled + TotalSchools, data = distr, posterior = TRUE,
                   iterations = 10000)
summary(regOutMCMC2)
```

```
##
## 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                   1.80186 0.0527833 5.278e-04      5.373e-04
## PctChildPoverty      0.01764 0.0091542 9.154e-05      9.154e-05
## PctFreeMeal         -0.02482 0.0032860 3.286e-05      3.362e-05
## PctFamilyPoverty    -0.03480 0.0129159 1.292e-04      1.317e-04
## Enrolled            -0.00050 0.0002174 2.174e-06      2.263e-06
## TotalSchools         0.04406 0.0200701 2.007e-04      2.088e-04
## sig2                 1.94015 0.1052956 1.053e-03      1.053e-03
## g                    0.09727 0.0921542 9.215e-04      9.808e-04
##
## 2. Quantiles for each variable:
##
##                         2.5%        25%        50%        75%      97.5%
## mu                   1.6981343  1.7660385  1.8017781  1.8374526  1.905e+00
## PctChildPoverty     -0.0007617  0.0114888  0.0176864  0.0237921  3.560e-02
## PctFreeMeal         -0.0312602 -0.0270449 -0.0248131 -0.0225881 -1.844e-02
## PctFamilyPoverty    -0.0604151 -0.0434349 -0.0348306 -0.0259042 -9.555e-03
```

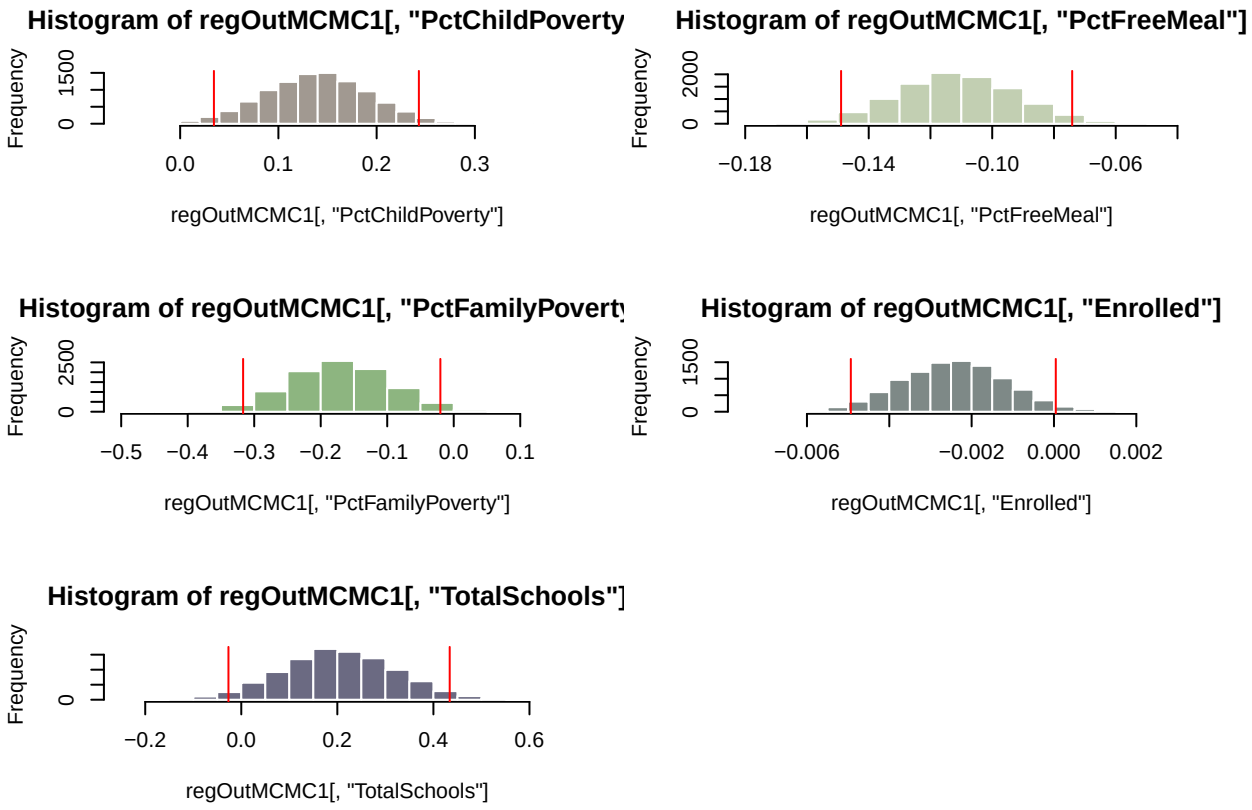
```
## Enrolled         -0.0009332 -0.0006425 -0.0005004 -0.0003535 -7.366e-05
## TotalSchools      0.0045927  0.0306043  0.0441043  0.0572093  8.375e-02
## sig2              1.7471227  1.8663649  1.9352195  2.0091380  2.156e+00
## g                 0.0263375  0.0492447  0.0726414  0.1133258  3.123e-01
```

In the command above, we run lmBF() with posterior = TRUE and iterations= 10000 to sample from the posterior distribution using MCMC technique. Looking at the MCMC output first, we see both the means of the respective distribution and the 95% HDI.

If we compare the Bayesian output and the LM procedure, we need to look at the mean column. The values are really close to the lm model. The top line of the tables is labeled mu - the grand mean for the data set. There are two specific columns: Naive SE and Time-series SE. SE stands for standard error (indication of dispersion in a sampling distribution). Naive SE calculates the standard error in the normal way. The Time-series SE accounts for the fact that there might be some, autocorrelation among the different estimates of the standard error.

The second table is overview of the highest density interval. We have the 2.5% and 97.5% boundaries of the HDI for each of the B-weights. These boundaries mark the edges of the central region of the posterior distribution for each B-weight. sig2 is the error variance in the regression equation. And we can use that error variance to calculate a distribution of r-squared values.

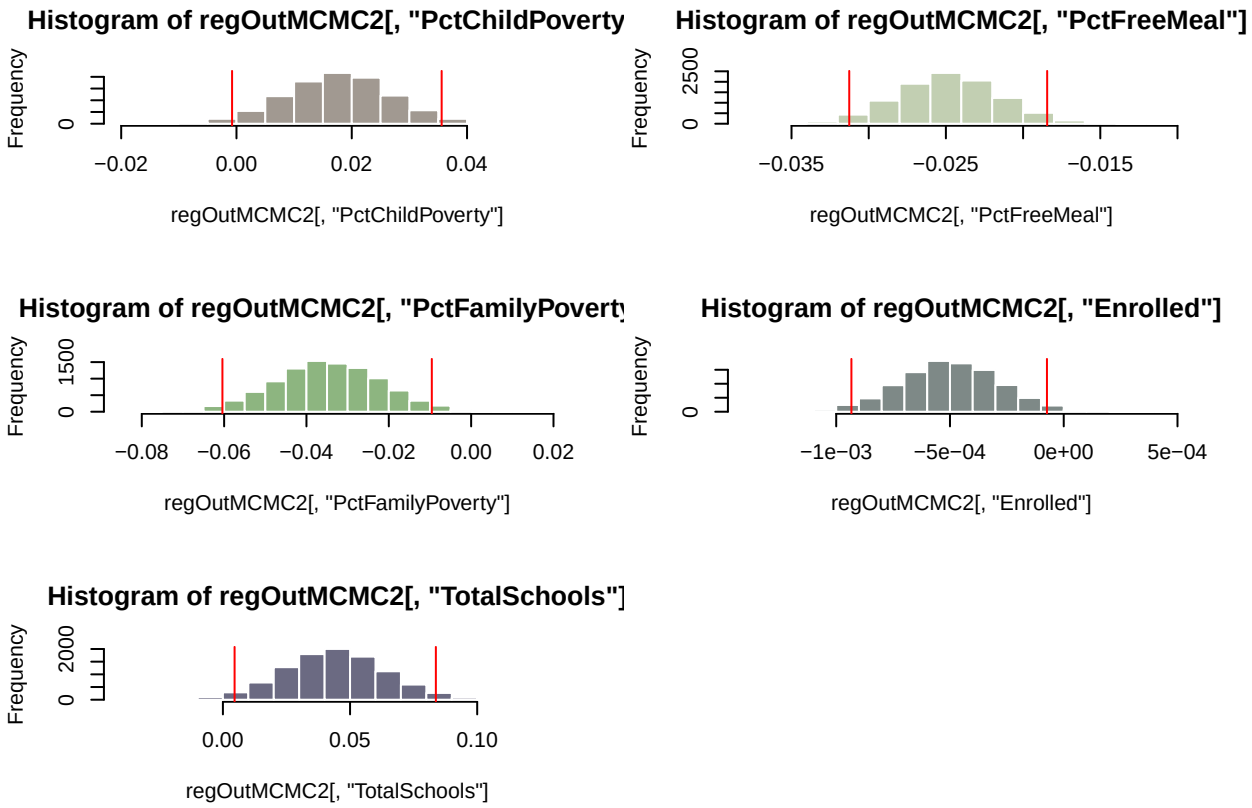
```
par(mfrow = c(3,2))
hist(regOutMCMC1[,"PctChildPoverty"], col = "#A19991", border = "white")
abline(v=quantile(regOutMCMC1[,"PctChildPoverty"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC1[,"PctChildPoverty"], c(0.975)), col = "red")

hist(regOutMCMC1[,"PctFreeMeal"], col = "#C2CFB2", border = "white")
abline(v=quantile(regOutMCMC1[,"PctFreeMeal"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC1[,"PctFreeMeal"], c(0.975)), col = "red")

hist(regOutMCMC1[,"PctFamilyPoverty"], col = "#8DB580", border = "white")
abline(v=quantile(regOutMCMC1[,"PctFamilyPoverty"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC1[,"PctFamilyPoverty"], c(0.975)), col = "red")

hist(regOutMCMC1[,"Enrolled"], col = "#7E8987", border = "white")
abline(v=quantile(regOutMCMC1[,"Enrolled"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC1[,"Enrolled"], c(0.975)), col = "red")

hist(regOutMCMC1[,"TotalSchools"], col = "#6B6A82", border = "white")
abline(v=quantile(regOutMCMC1[,"TotalSchools"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC1[,"TotalSchools"], c(0.975)), col = "red")
```



TotalSchool straddles with 0. Enrolled upper bound is right on 0 for the first model.

Improved model

```
par(mfrow = c(3,2))
hist(regOutMCMC2[,"PctChildPoverty"], col = "#A19991", border = "white")
abline(v=quantile(regOutMCMC2[,"PctChildPoverty"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC2[,"PctChildPoverty"], c(0.975)), col = "red")

hist(regOutMCMC2[,"PctFreeMeal"], col = "#C2CFB2", border = "white")
abline(v=quantile(regOutMCMC2[,"PctFreeMeal"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC2[,"PctFreeMeal"], c(0.975)), col = "red")

hist(regOutMCMC2[,"PctFamilyPoverty"], col = "#8DB580", border = "white")
abline(v=quantile(regOutMCMC2[,"PctFamilyPoverty"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC2[,"PctFamilyPoverty"], c(0.975)), col = "red")

hist(regOutMCMC2[,"Enrolled"], col = "#7E8987", border = "white")
abline(v=quantile(regOutMCMC2[,"Enrolled"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC2[,"Enrolled"], c(0.975)), col = "red")

hist(regOutMCMC2[,"TotalSchools"], col = "#6B6A82", border = "white")
abline(v=quantile(regOutMCMC2[,"TotalSchools"], c(0.025)), col = "red")
abline(v=quantile(regOutMCMC2[,"TotalSchools"], c(0.975)), col = "red")
```

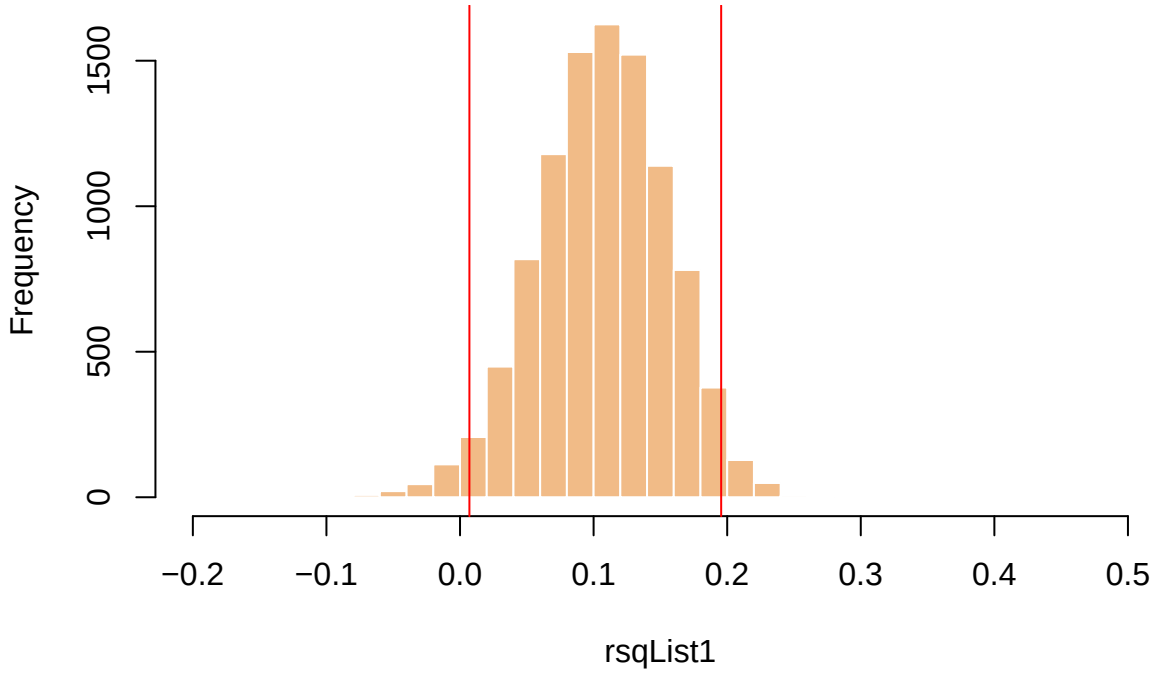


None of the intervals straddles with 0 now.

```
# var(dis$PctBeliefExempt)
## Calculate a distribution of r-squared values
rsqList1 <- 1 - (regOutMCMC1[,"sig2"] / var(dis$PctBeliefExempt))


hist(rsqList1
     , xlim = c(-0.2,0.5)
     , main = ""
     , col = "#F1BB87", border = "white")
mean(rsqList1) # overall mean R-square is 0.107
```

```
## [1] 0.1064729
```

```
## Draw boundaries of the 95% HDI
abline(v=quantile(rsqList1,c(0.025)), col="red")
abline(v=quantile(rsqList1,c(0.975)), col="red")
```

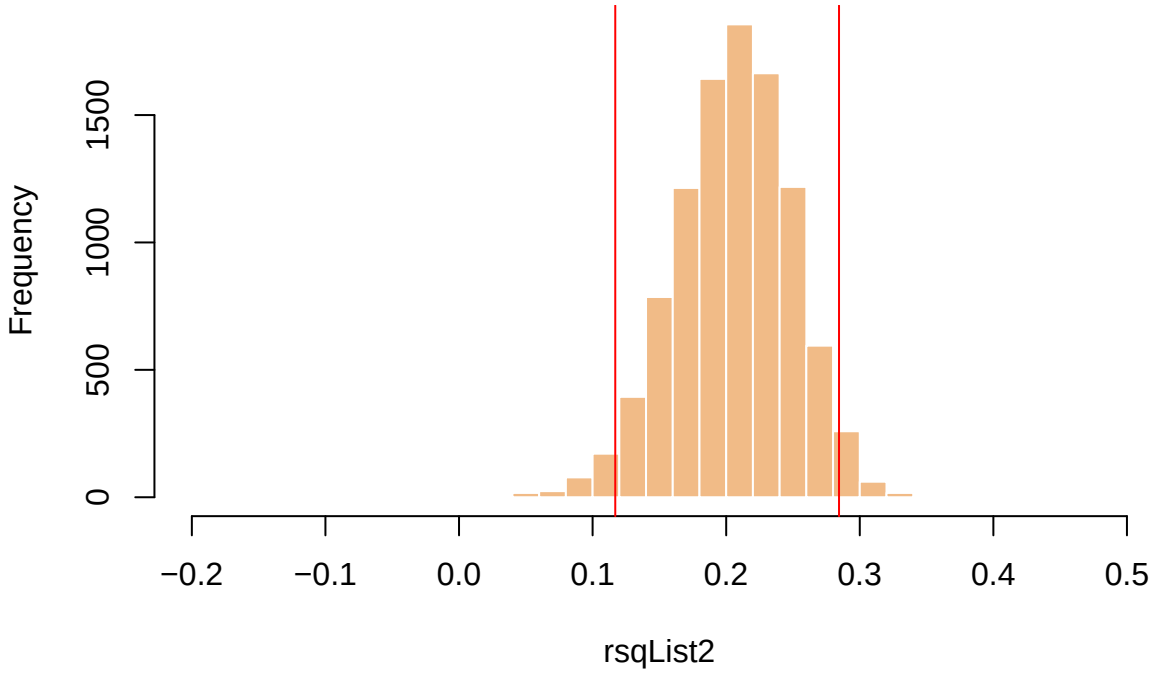



```
# quantile(rsqList1,c(0.025)) # 0.007025534
# quantile(rsqList1,c(0.975)) # 0.1954934
```

```
## Calculate a distribution of r-squared values for the improved model
rsqList2 <- 1 - (regOutMCMC2[,"sig2"] / var(distr$improvedVar))
hist(rsqList2
     , xlim = c(-.2,.5)
     , main = ""
     , col = "#F1BB87", border = "white")
mean(rsqList2) # overall mean R-square is 0.2054043
```

```
## [1] 0.2054043
```

```
## Draw boundaries of the 95% HDI
abline(v=quantile(rsqList2,c(0.025)), col="red")
abline(v=quantile(rsqList2,c(0.975)), col="red")
```




```
#quantile(rsqList2,c(0.025))# 0.1164406
#quantile(rsqList2,c(0.975))# 0.284142
```

The mean value of this distribution came out to 0.107(0.205 for the second model), which is quite a bit lower than the adjusted R-square of 0.1072(0.206 for the adjusted outcome variable model). The mean B-weight on the variables are smaller in magnitude for the Bayesian analysis, than for the conventional analysis which might be the reason for that difference in the R-square. Bayesian analysis finds that variables are not as strong predictors as the conventional analysis indicated. Bayesian model can give us a clear-eyed view of the likely range of the possibilities for the predictive strength of our model. In the underlying population, it is credible for us to expect R-square as low as about 0.007 or as high as about 0.195, with the most likely value pf R-square in that central region surrounding 0.107. The second model give us a little bit better HDI, still really close to 0 with a low bound at 0.116, and high at 0.284, with the most likely value of R-square at 0.205.

These results strengthen the conclusions we made. Using the "frequentist" (traditional) null hypothesis test, we rejected the null hypothesis for four predictors in the raw model, and for five predictors in the adjusted one. The Highest Density Intervals (HDIs) from the MCMC output showed estimates for the coefficients and R squared that concur with the frequentist models. The Bayes Factor for the raw model overwhelmingly favored a model that includes the four predictors. The adjusted model keept the five predictors.

```
gvlma::gvlma(x = lm(PctBeliefExempt~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty+
               Enrolled + TotalSchools, data = dis))
```

```
##
## Call:
```

```
## lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFreeMeal +
##     PctFamilyPoverty + Enrolled + TotalSchools, data = dis)
##
## Coefficients:
##      (Intercept)   PctChildPoverty       PctFreeMeal  PctFamilyPoverty
##        10.142821          0.143606         -0.115209         -0.176548
##         Enrolled      TotalSchools
##        -0.002542          0.211611
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma::gvlma(x = lm(PctBeliefExempt ~ PctChildPoverty + PctFreeMeal +      PctFamilyPoverty + Enroll
##
##                      Value p-value                   Decision
## Global Stat        4461.2331  0.0000 Assumptions NOT satisfied!
## Skewness            978.5170  0.0000 Assumptions NOT satisfied!
## Kurtosis           3481.1506  0.0000 Assumptions NOT satisfied!
## Link Function         1.3037  0.2535    Assumptions acceptable.
## Heteroscedasticity    0.2618  0.6089    Assumptions acceptable.
```

```
gvlma::gvlma(x = lm(improvedVar~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty+
               Enrolled + TotalSchools, data = distr))
```

```
##
## Call:
## lm(formula = improvedVar ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
##     Enrolled + TotalSchools, data = distr)
##
## Coefficients:
##      (Intercept)   PctChildPoverty       PctFreeMeal  PctFamilyPoverty
##        3.0278777          0.0181428         -0.0254012         -0.0356641
##         Enrolled      TotalSchools
##       -0.0005143          0.0453546
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma::gvlma(x = lm(improvedVar ~ PctChildPoverty + PctFreeMeal +      PctFamilyPoverty + Enrolled
##
##                       Value   p-value                   Decision
## Global Stat        216.9609 0.0000000 Assumptions NOT satisfied!
## Skewness           117.3577 0.0000000 Assumptions NOT satisfied!
## Kurtosis            88.5521 0.0000000 Assumptions NOT satisfied!
## Link Function       10.8347 0.0009962 Assumptions NOT satisfied!
## Heteroscedasticity   0.2163 0.6418499    Assumptions acceptable.
```

Summary of the results

We tested a model that used five variables(PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools) to predict the percentage of all enrolled students with belief exceptions. A Bayesian analysis of the raw model showed a mean posterior estimate for R-squared of 0.205, with the highest density interval ranging from roughly 0.007 to 0.195. Then a Bayesian analysis of the adjusted model showed a mean posterior estimate for R-squared of 0.107, with the highest density interval ranging from roughly 0.116 to 0.284. The traditional analysis confirmed this result with a slightly more optimistic R-squared of 0.107(0.206 for the second model). The F-test on those values were $F(5, 694)=17.9$, $p<.05$ and $F(5, 694)=37.47$, $p<.05$, so we reject the null hypothesis that R-squared was equal to zero. Only four predictors were significant in the raw model with B-weights of 0.143 (PctChildPoverty), -0.115 (PctFreeMeal), -0.176 (PctFamilyPoverty), -0.002(Enrolled). The adjusted model bring back the variable TotalSchools as significant, which is confirmed by the Bayes analysis for that model too. The Bayes factor of was strongly in favor of the four predictor model (in comparison with an intercept-only model). The adjusted model Bayes factor of 5.413776e+30 was strongly in favor of the five predictors. We have to mention that R-square is really close to 0, and its HDI with 0 which makes the model not that good and take into consideration the transformation on the outcome variable.

8. What's the big picture, based on all of the foregoing analyses? The staff member in the state legislator's office is interested to know how to allocate financial assistance to school districts to improve both their vaccination rates and their reporting compliance. What have you learned from the data and analyses that might inform this question?

```
mean(districts12$PctUpToDate) # 87,7% up to date
```

```
## [1] 87.76143
```

```
mean(districts12$PctBeliefExempt) # %5.6% exemp
```

```
## [1] 5.685714
```

```
mean(districts12$PctChildPoverty) # 22% child poverty
```

```
## [1] 22.19857
```

```
mean(districts12$PctFamilyPoverty) # 11% family poverty
```

```
## [1] 11.54571
```

```
mean(districts12$PctFreeMeal) # 48.1 free meals
```

```
## [1] 48.19
```

```
sum(districts12$TotalSchools) # 5080 schools
```

```
## [1] 5080
```

```
mean(districts12$DistrictComplete) # 94%
```

```
## [1] NA
```

There are 5080 schools, with 5.6% examption rate, which is high. Child poverty id 22% and the family poverty is 11%. Probably financial assistance have to be allocated in those problem groups.

Vaccination involves a virus or bacteria being purposely introduced to an individual, usually through injection, in order for the person's immune system to develop resistance to a specific disease. Vaccination is one of the most effective means of preventing infectious diseases and is responsible for the reduction of a number of diseases including measles and polio, and the complete eradication of small pox. As of 2018, an estimated 86 percent of one-year-olds worldwide had been vaccinated against measles, while Merck & Co., one of the largest pharmaceutical companies in the world, generated over 6.5 billion dollars in vaccine revenues in 2017. Some diseases, such as measles, have made a return in the United States because parents in some areas have failed to vaccinate their children. Last year, Minnesota suffered the state's worst measles outbreak in decades. It was sparked by anti-vaccine activists who targeted an immigrant community, spreading misinformation about the measles vaccine. Most of the 75 confirmed cases were young, unvaccinated Somali American children.

Vaccination rates has varied over time due to many reasons. The major change point for **DTP1** vaccine rate occurs in 1991 (started to increase). **Pol3** rate started to increase in 1993, **Hib3** increase rate in 2013 and **MCV1** occurs at increase in 1996. Those are major changing point of variance for each of the time series. The DTP1 were at higher levels around 1987, in comparison to the rest. We observed growth in the trends (only MCV1 has decreased mean level of rates).

HepB_BD has the biggest jump (from 20 to 60). MCV1 has the highest rate of 98 at the conclusion of the time series. HepB_BD has the lowest rate of 11 at the conclusion of the time series. If we look at the first and the last value of each one, we can see that all the rates increased from 1980 to 2017, only Pol3 decreased from 95 to 94. HepB_BD has the greatest overall SD of differenced series: 22.53904. Another problem with the vaccines is that the school don't report the data. Some kids got accepted without a vaccine and probably they never got it, or schools has less than 95% vaccination rate, which is a problem according to the laws. There is definitely relationship between those two categorical variables, and we modeled it with our posterior estimates. In proportion only 26/1000 public school didnt reported vaccination (37.7times as many public schools reported as they did not). 1/8 of the private school did not reported (5.5 times as many private schools reported as they did not). Public schools are did better job reporting in the subset we explored. Vaccination rates for California Public Schools are not that good: DTP 89.7%, Polio 90.1%, MMR 89.7% and HepB 92.1%. All the rates are lower compared to overall US vaccination rates for 2013 and the last values of the time series, only HepB has an increase compared to 2013 and of the final observation of the time series. The correlation between students without vaccines is high suggesting that if one student misses one vaccine, he probably misses all of them. The null hypothesis testing on the correlation - the procedure for testing the significance of the correlation coefficient confirmed that theory. Student exemption from the vaccines is based on parents believes or medical conditions. Most parents dont want their kind vaccinated because of the risk of complications (autism for example).

We tested a measures of PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled and TotalSchools to see if they could predict wheater or not a district's reporting was complete. A chi-square omnibus test on the result of logistic regression was significant for model with the 2 predictors, chisq(2) = 16.8036, p<.0001. Only the Wald's z-test on the Enrolled and TotalSchools coefficient was significant, z= -8.05, p<.05 and z=3.28, p<.05. When converted to odds, the coefficient for Enrolled was 0.99 suggesting that for each unit increase in enrolled, the odds of the district's reporting to be complete is .99:1. Coefficient for TotalSchools was 1.22(odds 1.22:1). We conducted Bayesian logistic analysis, using all of 5 of them to predict DistrictComplete. The posterior distributions of the coefficients for PctChildPoverty, PctFreeMeal, PctFamilyPoverty (calibrated as log odds) overlapped squarely with 0, suggesting that they were not meaningful predictor of engine. In contrast the HDI for Enrolled and TotalSchools did not overlap with zero. When converted to regular odd, the mean value of the posterior distribution for Enrolled was .99:1 suggesting that for every additional enrolled individual, the district reporting is about 1% more likely to be complete. Converted to regular odd, the mean value of the posterior distribution for TotalSchools was 1.22:1 suggesting that for every additional enrolled individual, the district reporting is about 1.22% more likely to be complete. The confusion matrix

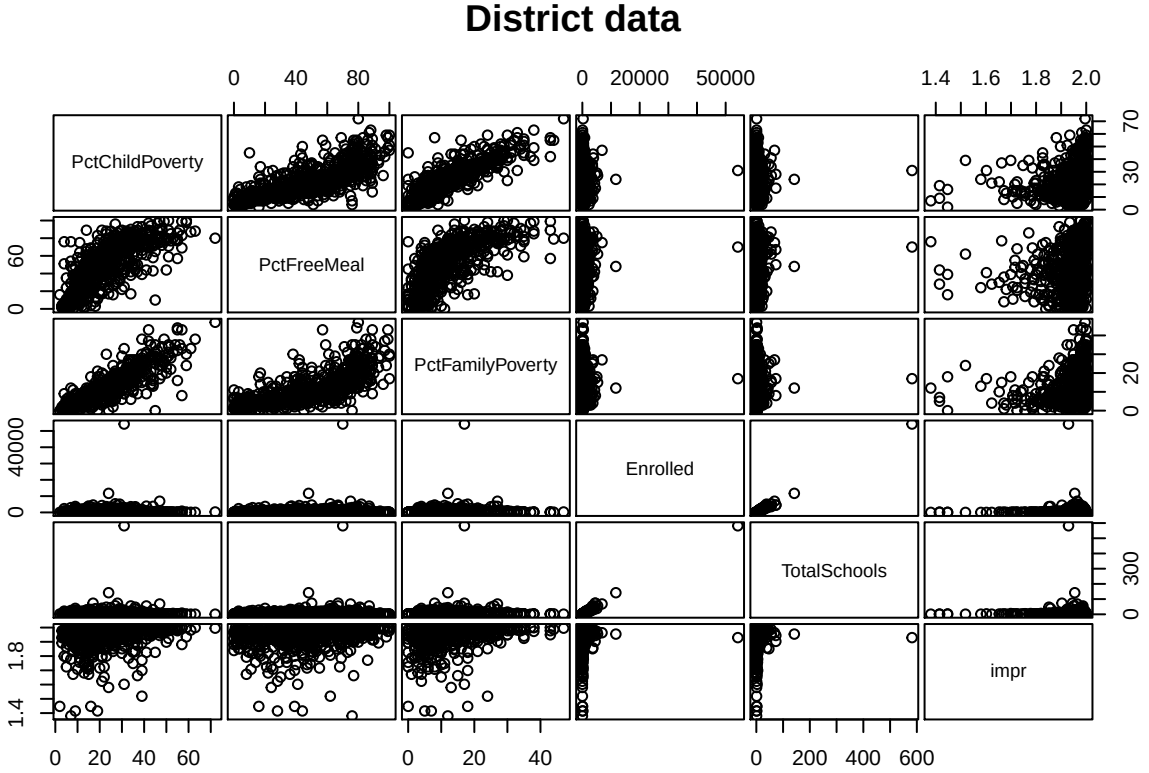
showed overall error rate of 0.06% indicating that the logistic model was good.

We tested also a model that used five variables (PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools) to predict completely up-to-dates vaccines. A Bayesian analysis of this model showed a mean posterior estimate for R-squared of 0.085, with the highest density interval ranging from roughly -0.01 to 0.177. The traditional analysis confirmed this result with a slightly more pesimistic R-squared of 0.084. The F-test on this value was F(5, 694)=13.9, p<.05, so we reject the null hypothesis that R-squared was equal to zero. Only four predictors were significant with B-weights of 0.087(PctFreeMeal), 0.255 (PctFamilyPoverty), 0.005(Enrolled), -0.517(TotalSchools). The Bayes factor of 20424863319 was strongly in favor of the four-predictor model (in comparison with an intercept-only model). We have to mention that R-square is really close to 0, and its HDI with 0 wich makes the model not that good. We tested a model that used five variables (PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools) to predict the percentage of all enrolled students with belief exceptions. A Bayesian analysis of the raw model showed a mean posterior estimate for R-squared of 0.205, with the highest density interval ranging from roughly 0.007 to 0.195. Then a Bayesian analysis of the adjusted model showed a mean posterior estimate for R-squared of 0.107, with the highest density interval ranging from roughly 0.116 to 0.284. The traditional analysis confirmed this result with a slightly more optimistic R-squared of 0.107(0.206 for the second model). The F-test on those values were F (5, 694) =17.9, p<.05 and F(5, 694)=37.47, p<.05, so we reject the null hypothesis that R-squared was equal to zero. Only four predictors were significant in the raw model with B-weights of 0.143 (PctChildPoverty), -0.115 (PctFreeMeal), -0.176 (PctFamilyPoverty), -0.002(Enrolled). The adjusted model bring back the variable TotalSchools as significant, which is confirmed by the Bayes analysis for that model too. The Bayes factor of was strongly in favor of the four-predictor model (in comparison with an intercept-only model). The adjusted model Bayes factor of 5.413776e+30 was strongly in favor of the five predictors. We have to mention that R-square is really close to 0, and its HDI with 0 which makes the model not that good and take into consideration the transformation on the outcome variable.

```
cor(newdistIM)
```

	PctChildPoverty	PctFreeMeal	PctFamilyPoverty	Enrolled	TotalSchools	impr
PctChildPoverty	1.0000000	0.7562022	0.8641509	0.0265675	0.0212344	0.1945220
PctFreeMeal	0.7562022	1.0000000	0.7286677	0.0649288	0.0593207	0.2210301
PctFamilyPoverty	0.8641509	0.7286677	1.0000000	0.0388149	0.0318555	0.2124643
Enrolled	0.0265675	0.0649288	0.0388149	1.0000000	0.9940612	0.0623193
TotalSchools	0.0212344	0.0593207	0.0318555	0.9940612	1.0000000	0.0508403
impr	0.1945220	0.2210301	0.2124643	0.0623193	0.0508403	1.0000000

```
pairs(newdistIM, main = "District data", gap = 1/4)
```



Pairs plot shows high positive correlation between PctChildPoverty, PctFreeMeal and PctFamilyPoverty. That means if the percentage of one goes up, the otherwill go up too. As we saw in the logistic regression model PctChildPoverty, PctFreeMeal and PctFamilyPoverty were not a good predictor for reporting compliance, but they had significant values in the linear regression model. School districts can assist those people in order to improve vaccination rates. Lower income class not always can afford health insurance to cover all vaccination taxes.

Number of enrolled students and the total schools were factors that predict the reporting compliance in CA public schools. California schools have to submit vaccine rates but face no consequences if they don't so maybe some additional fines from not reporting can be created and the money collected from those fines can go towards law income families or towards more free meals. Private schoold reporting rates have to be improved.