Final

Kwasi Mensah

2022-12-21

Author: Kwasi Mensah Final Output: pdf_document Attribution statement:

I did the homework by myself, with help from the book and the professor and the following sources: https://www.r-tutor.com/elementary-statistics/multiple-linear-regression/estimated-multiple-regression-equation https://www.displayr.com/variance-inflation-factors-vifs/

#R Markdown #Run these three functions to get a clean test of homework code

```
dev.off() #Clear the graph window

## null device
##           1

cat('\014') #Clear the console

rm(list = ls()) #Clear user objects from the environment
```

#R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this: ##

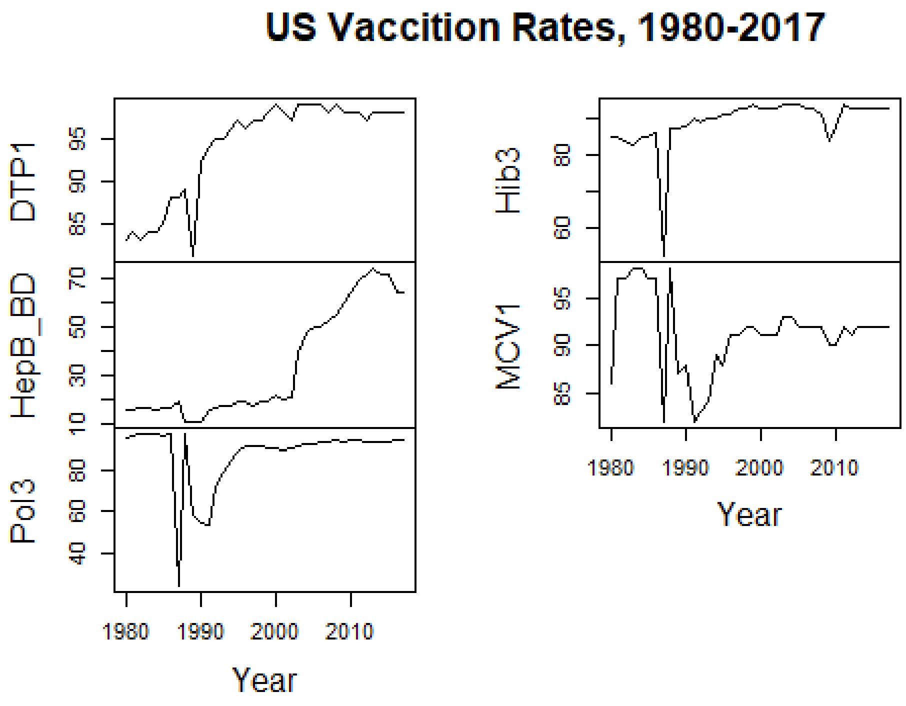
1. How have U.S. vaccination rates varied over time? Are vaccination rates increasing or decreasing? Which vaccination has the highest rate at the conclusion of the time series? Which vaccination has the lowest rate at the conclusion of the time series? Which vaccine has the greatest volatility?

```
load("C:/Users/lmori/Downloads/allSchoolsReportStatus.RData")
load("C:/Users/lmori/Downloads/districts10.RData")
load("C:/Users/lmori/Downloads/usVaccines.RData")
par(mar = c(1, 1, 1, 1))
summary(usVaccines)

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

```
tail(usVaccines)
```

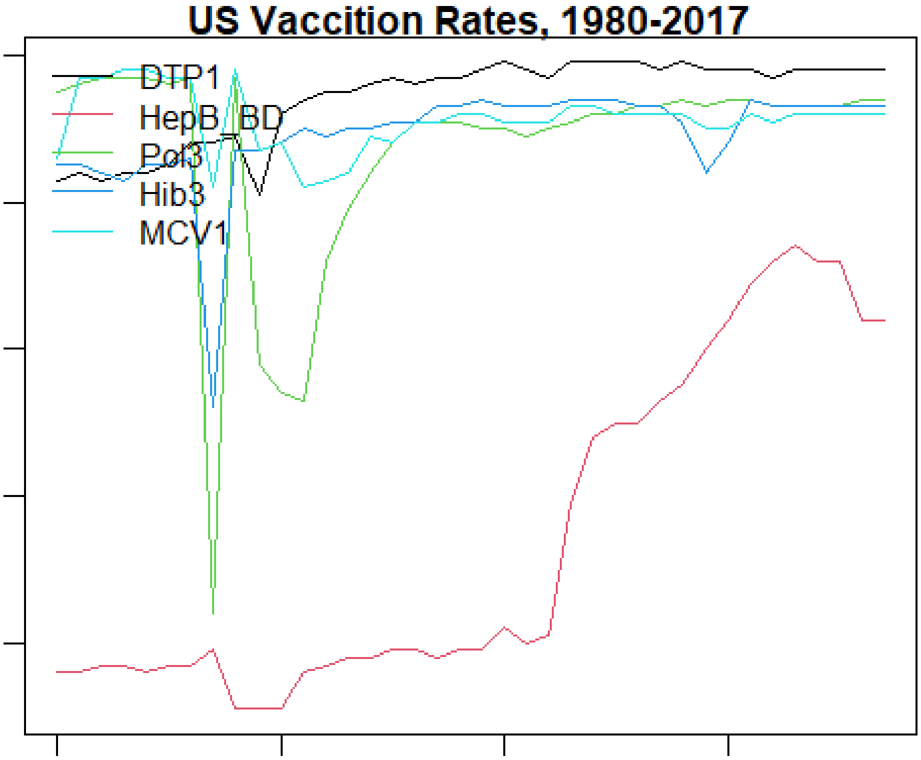
```
##       DTP1 HepB_BD Pol3 Hib3 MCV1
## [33,]   97      72   93   93   91
## [34,]   98      74   93   93   92
## [35,]   98      72   93   93   92
## [36,]   98      72   93   93   92
## [37,]   98      64   94   93   92
## [38,]   98      64   94   93   92
```

```
plot(usVaccines, xlab="Year", main="US Vaccition Rates, 1980-2017")
```

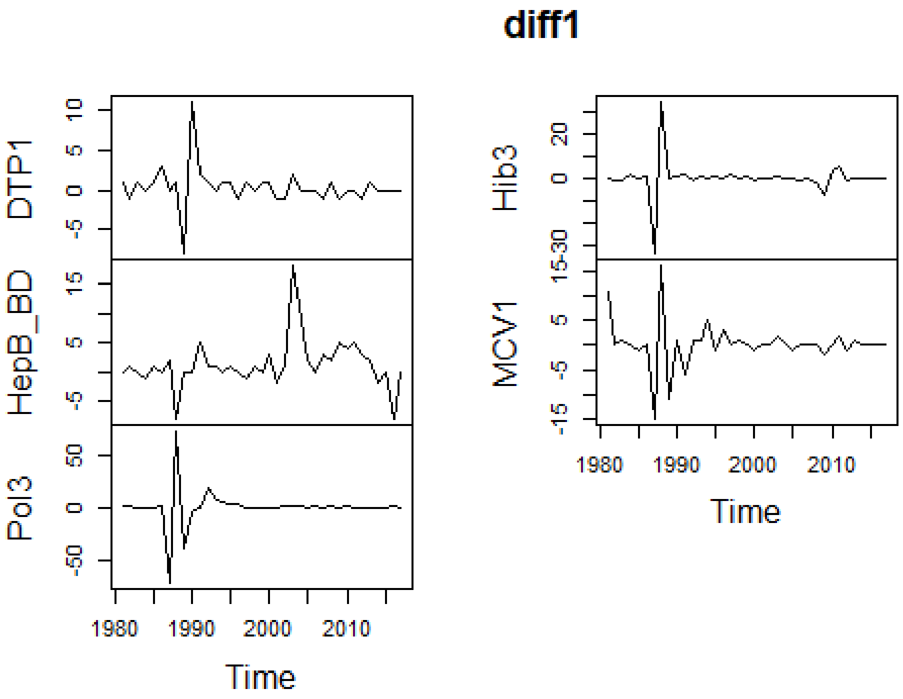


```
#U.S vaccination rates did not vary much over time, the rate tended to remain
more or less stagnant after 1990 other than the rate of the HepB_BD vaccine
which increased over time before dropping toward the end of the time series.
```

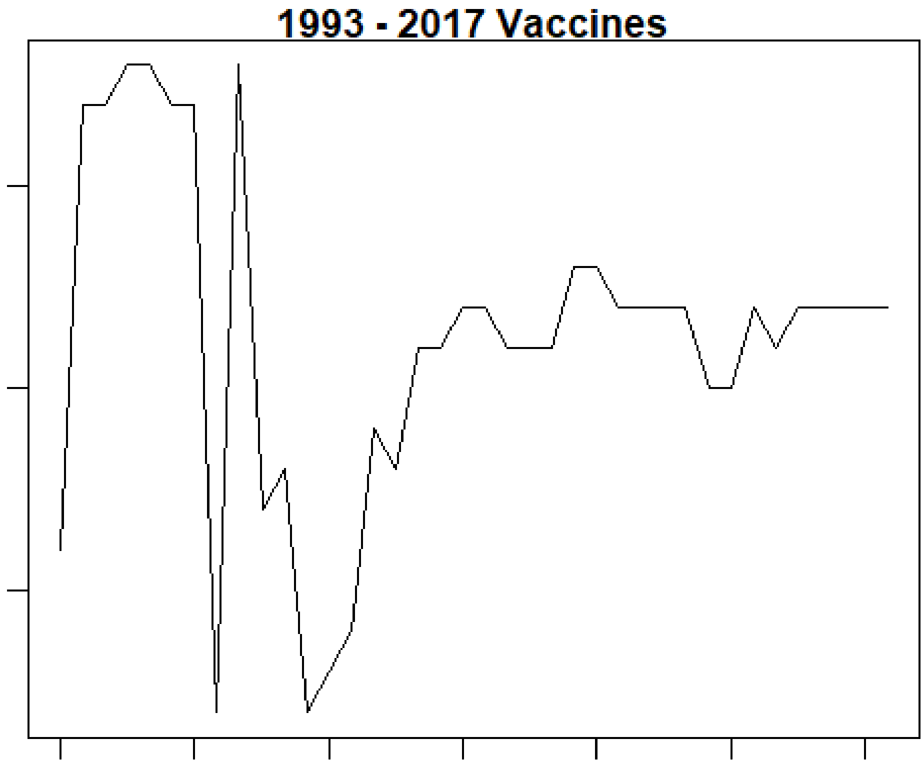
```
#U.S vaccination rates increased over time then became more or less stagnant.
#The DTP1 vaccine had the highest rate at the conclusion of the time series.
#The HepB_BD vaccine had the lowest vaccination rate at the end of the time
series.


#Time-Series in One Chart
is.ts(usVaccines)

## [1] TRUE

ts.plot(usVaccines, col = 1:5,  xlab="Year", ylab="Vaccination Rates",
main="US Vaccition Rates, 1980-2017")
legend("topleft", colnames(usVaccines), lty=1, col = 1:5, bty ="n")
```



```
diff1<-diff(usVaccines)
plot(diff1)
```



```
#   When we remove the trend component and see the graphs after differencing,
we see that over time vaccination rates remained stagnant for the most part,
other than HepB_BD vaccine  over the selected time period(1980-2017). We can
see that all the vaccines rates dropped in the early years between 1985 and
1990. We can see that there is an increase in vaccination rate between 1985-
1990 for the DTP1, Pol3,MCV1, and Hib3 vaccines. We see that there was a jump
in the vaccination rates of the HepB_BD vaccine between 2000 and 2005.



runTStests1 <- function (testSeries, seriesName) {
  # The basics
  print(seriesName)
  ts.plot(testSeries, main="1993 - 2017 Vaccines")
  print("Amount of decline Over the years")
  print(max(testSeries) - min(testSeries))
  print("Total change in trend:")
  print(testSeries[length(testSeries)] - testSeries[1])
  print("Max value:")
  print(max(testSeries))
  print("Min value:")
  print(min(testSeries))
}
######### Function for Stationarity and Volatility#########
runTStests2 <- function (testSeries, seriesName) {
  # Examine stationarity
  diffTestSeries <- diff(testSeries[,seriesName])
  diffTestSeries
  acf(diffTestSeries, main="ACF: Differenced Series")
  require(tseries)
  a1<-adf.test(diffTestSeries)
  a1

  # Look at volatility
  plot(diffTestSeries, main="Differenced Series")
  print("Overall SD of differenced series:")
  print(sd(diffTestSeries))
  library(changepoint)
  cptOut <- cpt.var(diffTestSeries, method = "PELT")
  print(cptOut)
  plot(cptOut, main="Variance Changepoint Plot")

}

######################################################################
# MCV1
usVaccines1<- data.frame(usVaccines)
testSeries1 <- ts(usVaccines1$MCV1 , frequency = 12)
runTStests1(testSeries1,"MCV1" )

## [1] "MCV1"
```

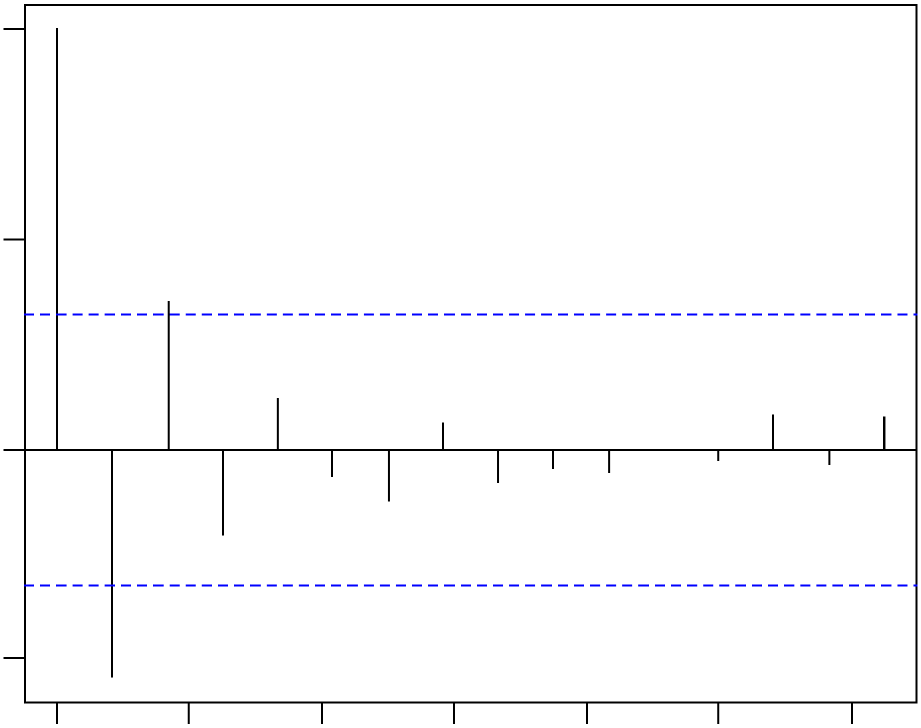


```
## [1] "Amount of decline Over the years"
## [1] 16
## [1] "Total change in trend:"
## [1] 6
## [1] "Max value:"
## [1] 98
## [1] "Min value:"
## [1] 82
```

```
testSeries1_1 <- ts(usVaccines1, frequency = 12)
runTStests2(testSeries1_1,"MCV1" )
```

```
## Loading required package: tseries
```

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



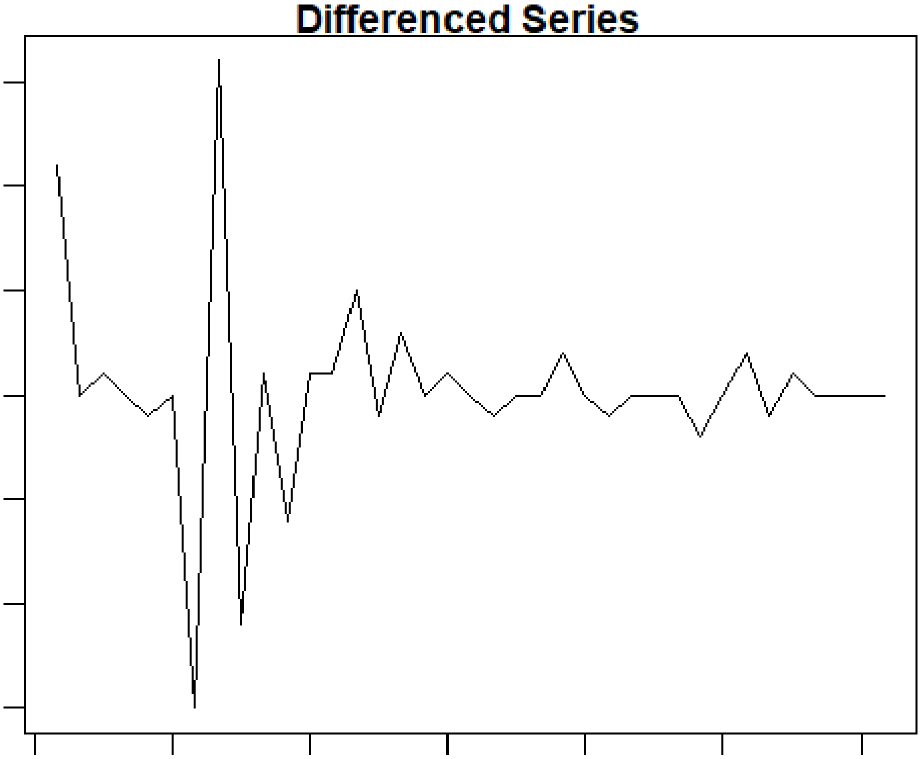
```
## [1] "Overall SD of differenced series:"
## [1] 4.758113
```

```
## Warning: package 'changepoint' was built under R version 4.2.2
```

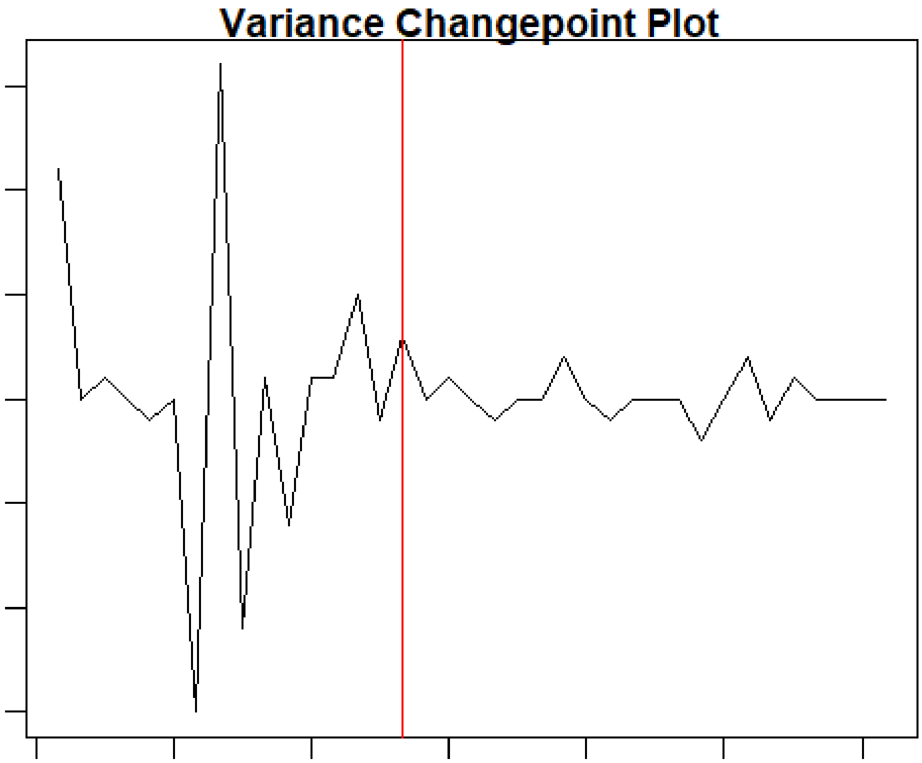
```
## Loading required package: zoo
```

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

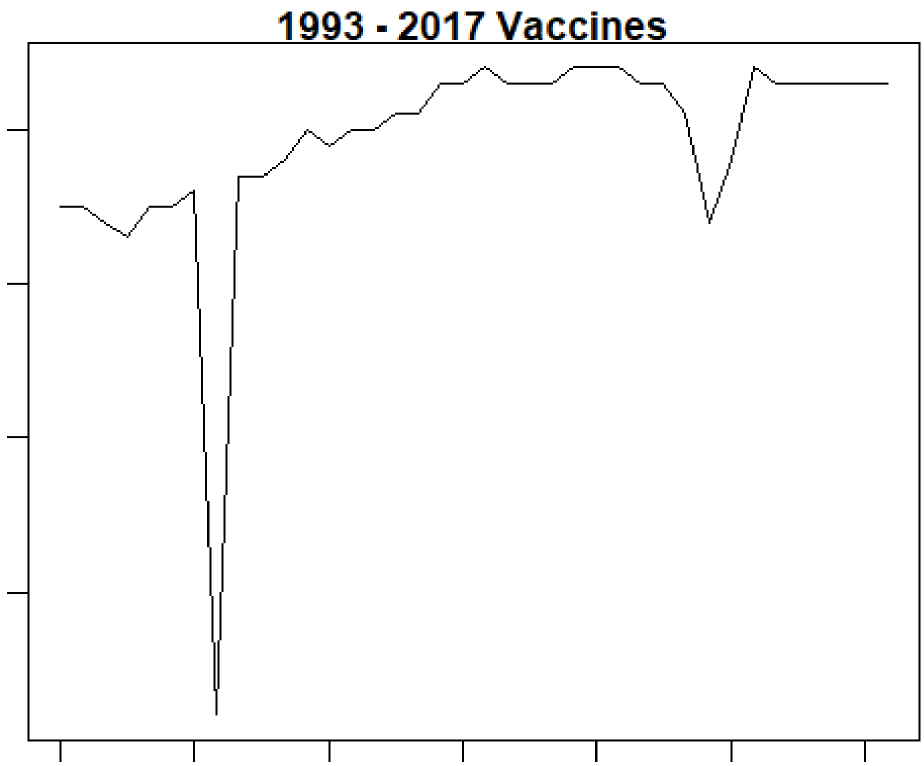
```
## Successfully loaded changepoint package version 2.2.4
##  See NEWS for details of changes.
```



```
## 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.max param.est
##
## Created on  : Thu Dec 15 03:47:50 2022
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.83275
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations : 16
```



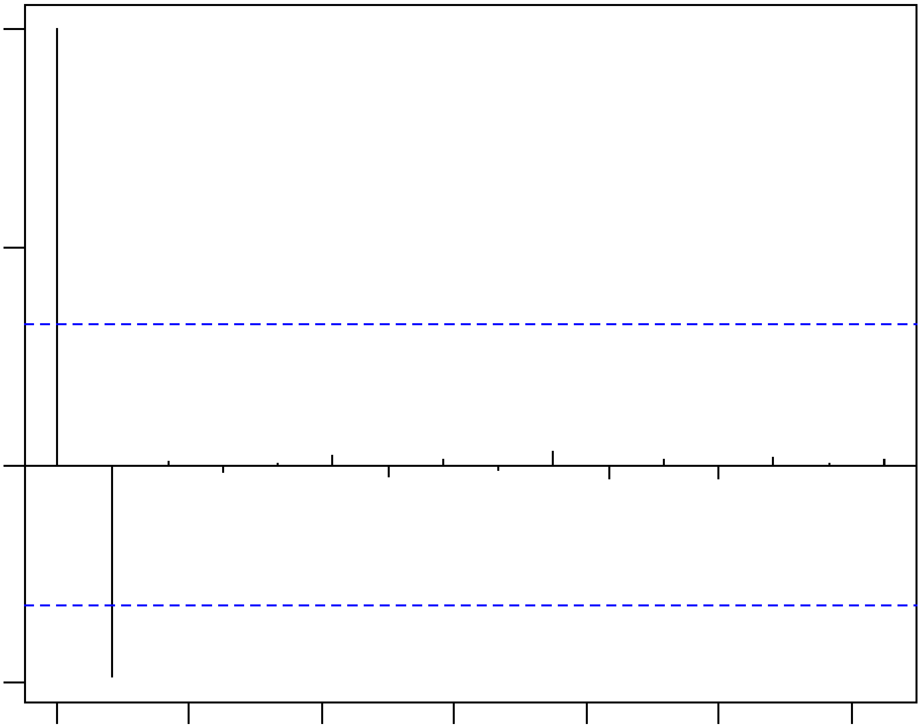
```
# Hib3
testSeries2 <- ts(usVaccines1$Hib3, frequency = 12)
runTStests1(testSeries2,"Hib3" )

## [1] "Hib3"
```

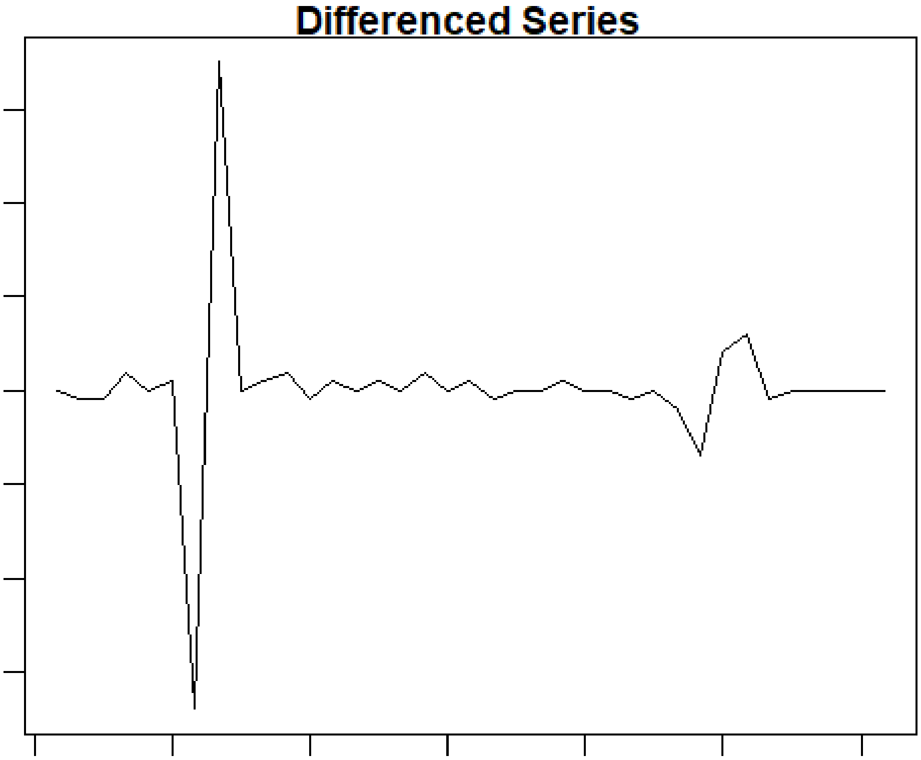


```
## [1] "Amount of decline Over the years"
## [1] 42
## [1] "Total change in trend:"
## [1] 8
## [1] "Max value:"
## [1] 94
## [1] "Min value:"
## [1] 52
```

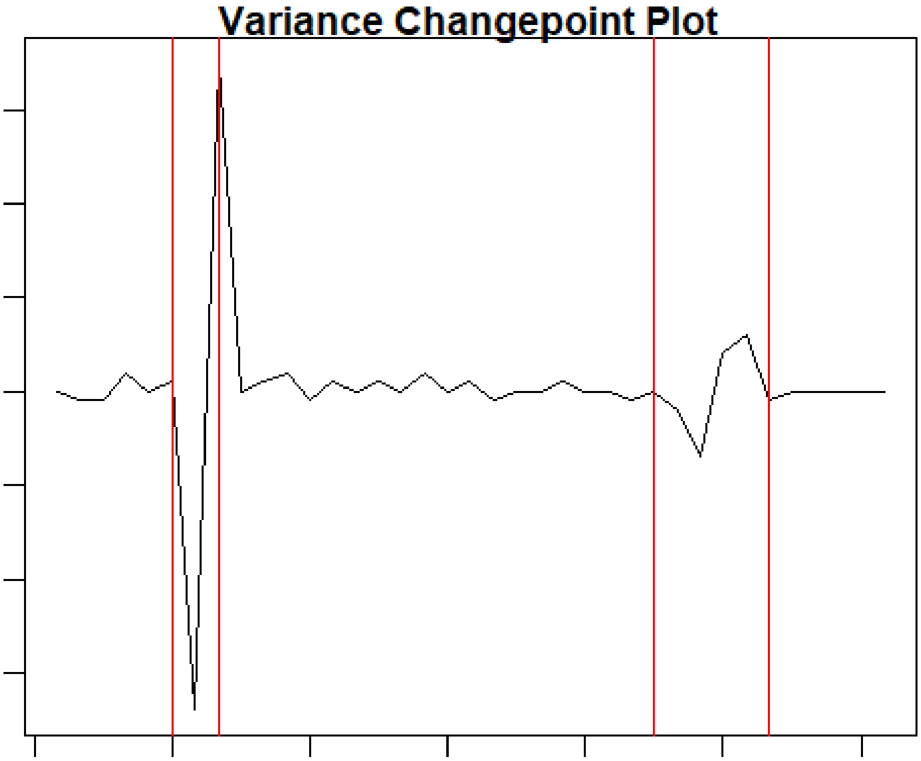
```
testSeries2_1 <- ts(usVaccines1, frequency = 12)
runTStests2(testSeries2_1,"Hib3" )
```

```
## Warning in adf.test(diffTestSeries): p-value smaller than printed p-value
```

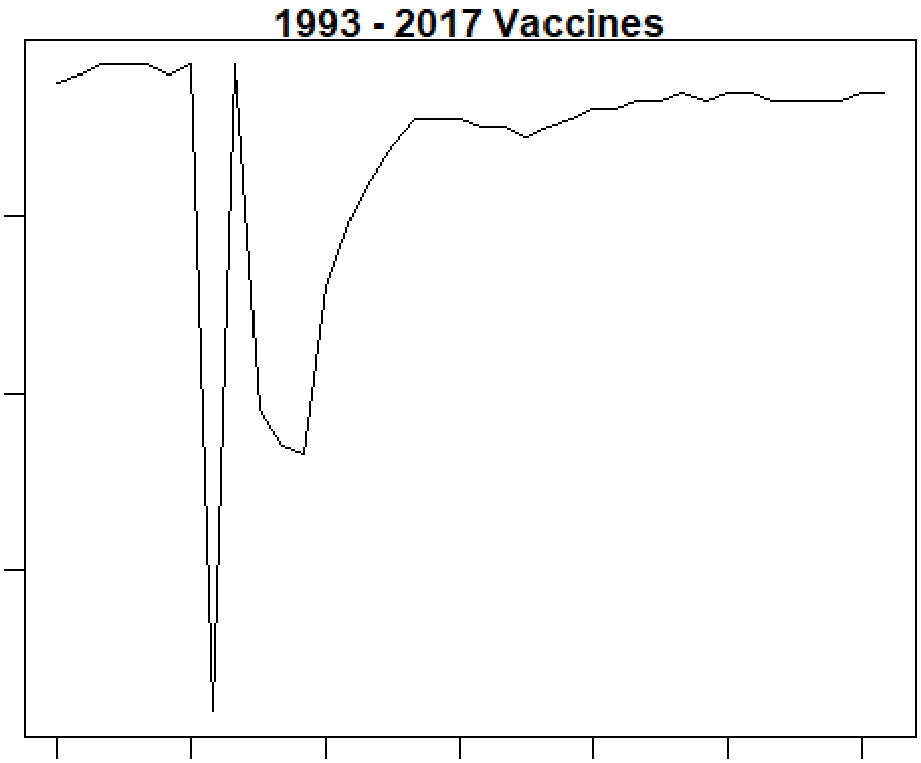




```
## [1] "Overall SD of differenced series:"
## [1] 8.347106
## Class 'cpt' : Changepoint Object
##        ~~   : S4 class containing 12 slots with names
##               cpttype date version data.set method test.stat pen.type
pen.value minseglen cpts ncpts.max param.est
##
## Created on  : Thu Dec 15 03:47:50 2022
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.83275
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations : 6 8 27 32
```

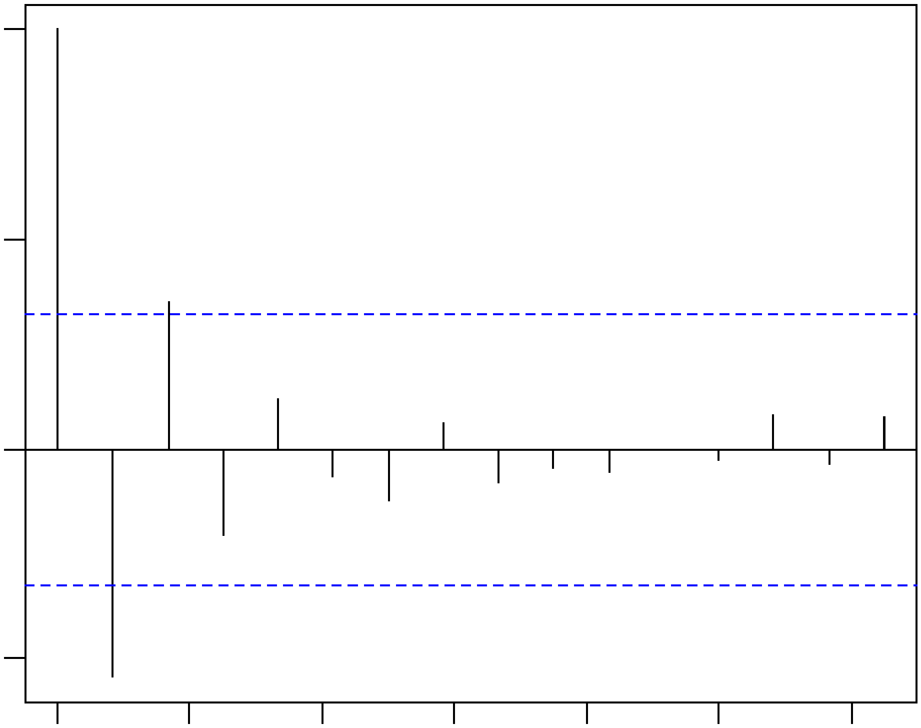


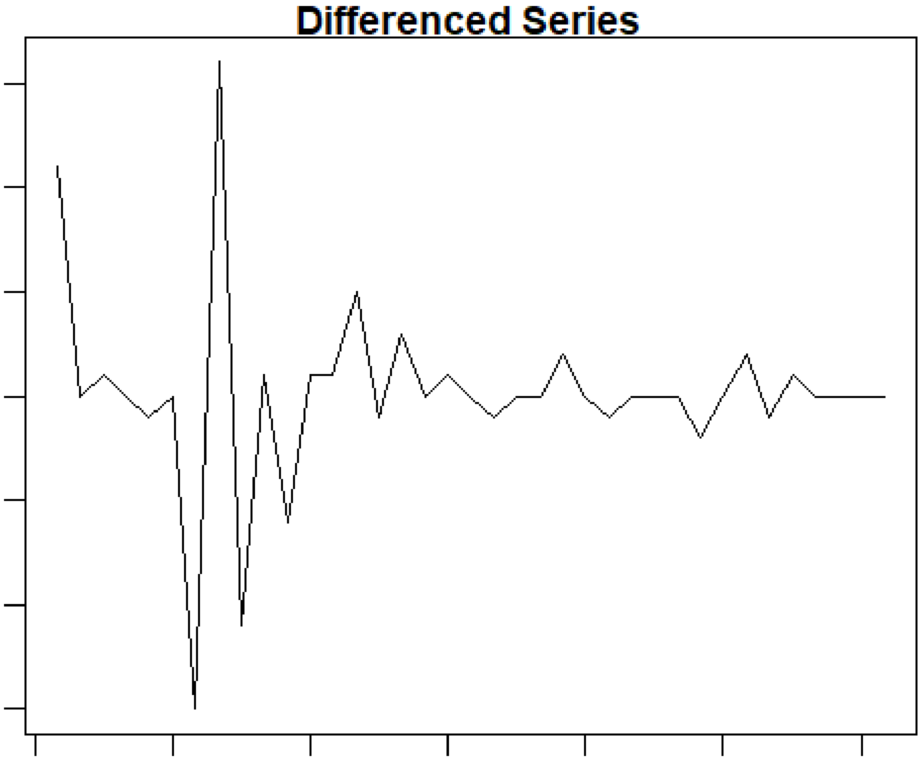
```
# Pol3
testSeries3 <- ts(usVaccines1$Pol3, frequency = 12)
runTStests1(testSeries3,"Pol3" )

## [1] "Pol3"
```



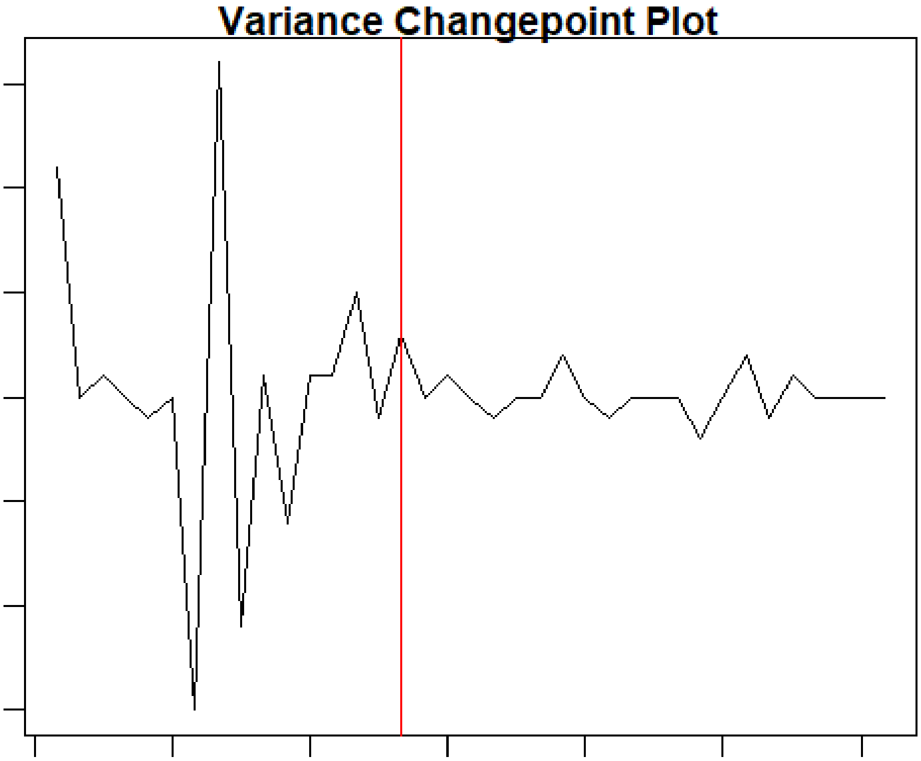
```
## [1] "Amount of decline Over the years"
## [1] 73
## [1] "Total change in trend:"
## [1] -1
## [1] "Max value:"
## [1] 97
## [1] "Min value:"
## [1] 24
```

```
testSeries3_1 <- ts(usVaccines1, frequency = 12)
runTStests2(testSeries3_1,"MCV1" )
```

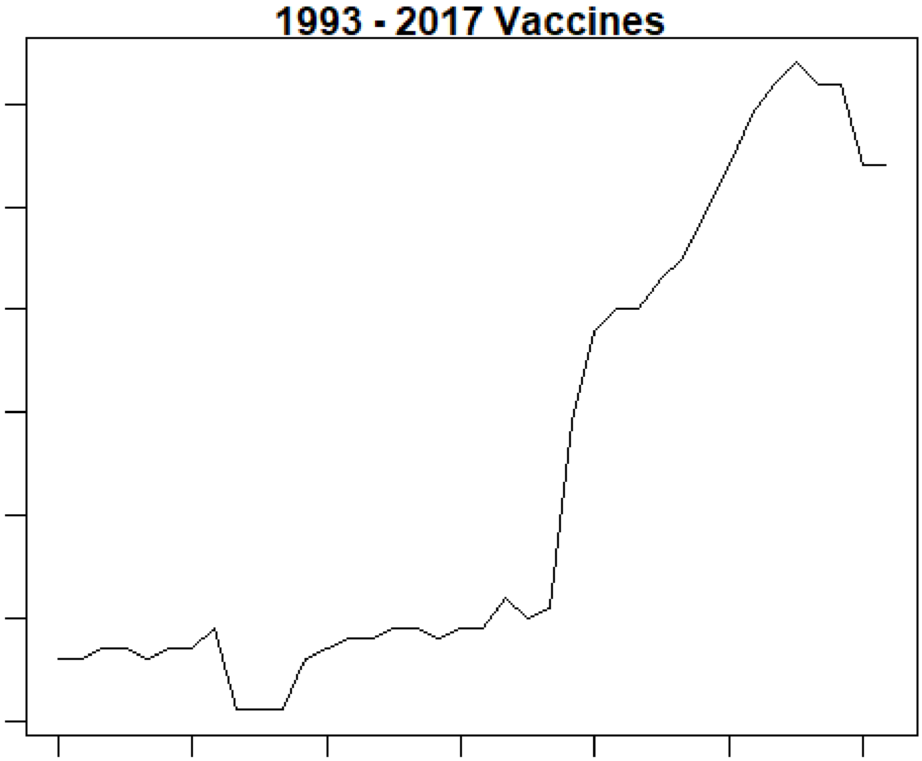




```
## [1] "Overall SD of differenced series:"
## [1] 4.758113
## Class 'cpt' : Changepoint Object
##        ~~   : S4 class containing 12 slots with names
##                cpttype date version data.set method test.stat pen.type
pen.value minseglen cpts ncpts.max param.est
##
## Created on  : Thu Dec 15 03:47:50 2022
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.83275
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations : 16
```

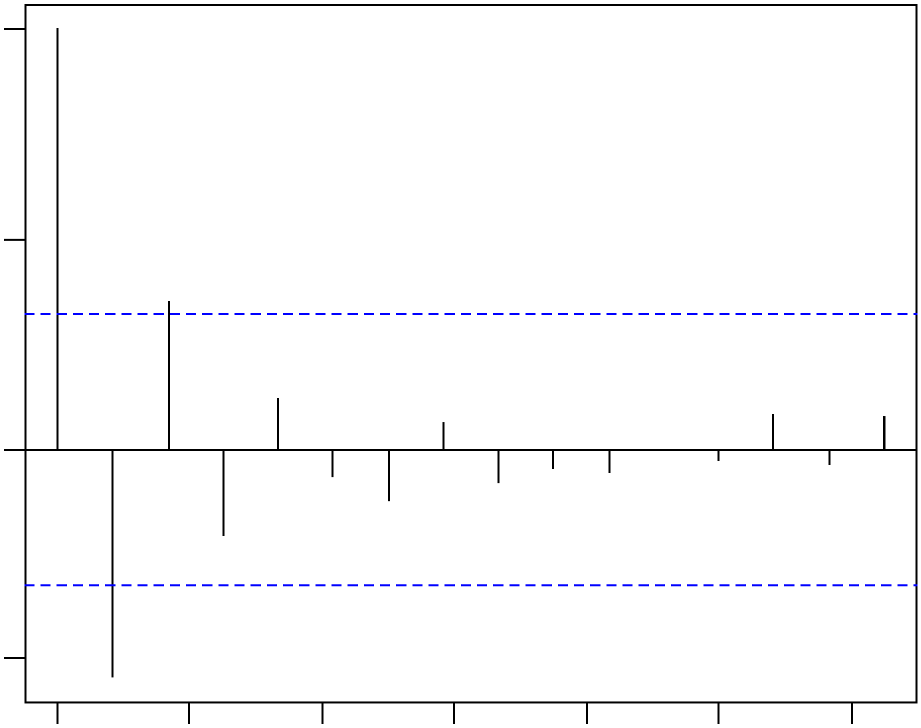


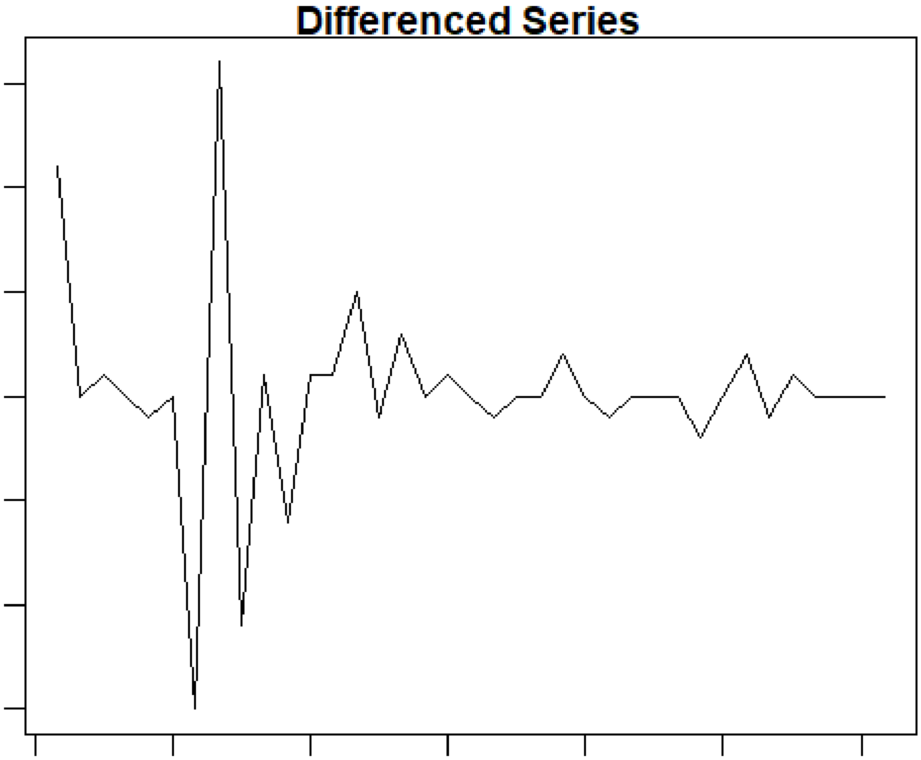
```
#HepB_BD
testSeries4 <- ts(usVaccines1$HepB_BD, frequency = 12)
runTStests1(testSeries4,"HepB_BD" )

## [1] "HepB_BD"
```



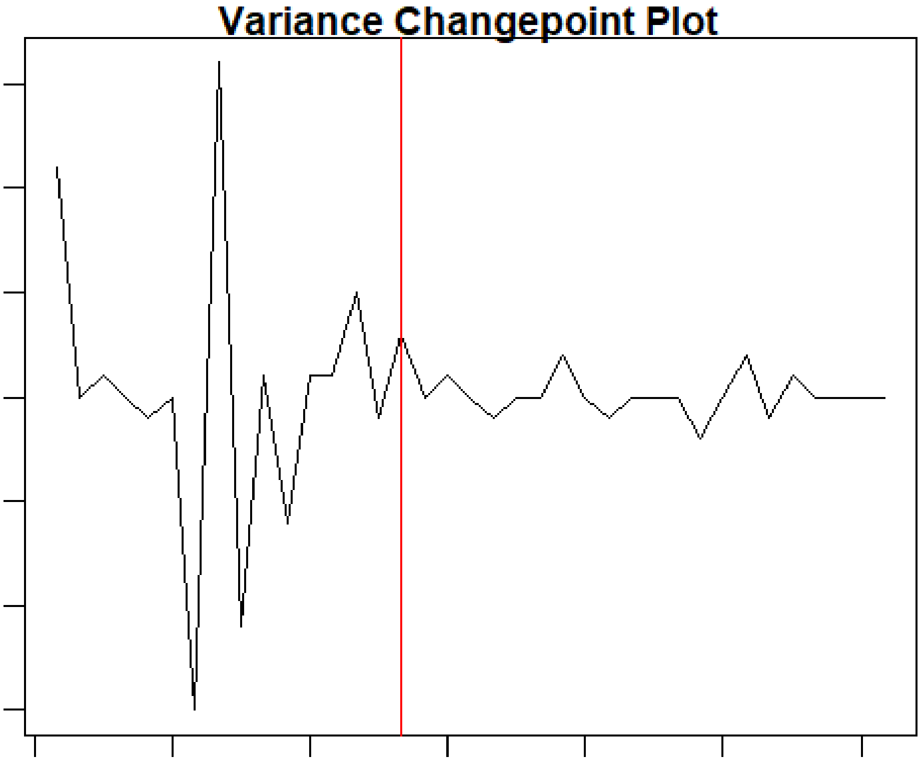
```
## [1] "Amount of decline Over the years"
## [1] 63
## [1] "Total change in trend:"
## [1] 48
## [1] "Max value:"
## [1] 74
## [1] "Min value:"
## [1] 11
```

```
testSeries4_1 <- ts(usVaccines1, frequency = 12)
runTStests2(testSeries4_1,"MCV1" )
```

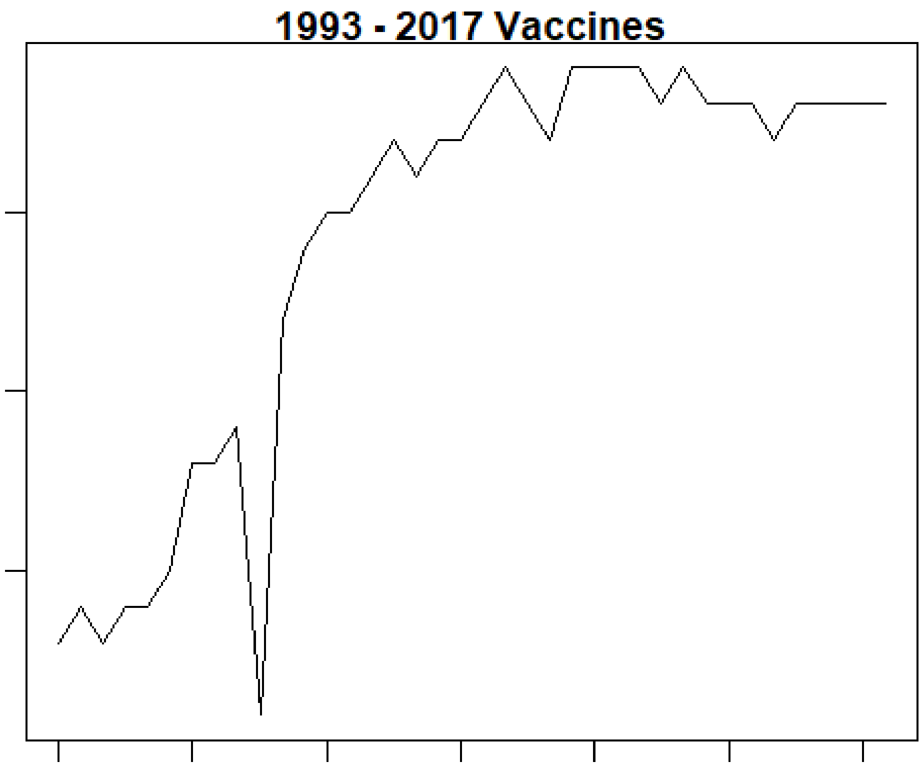




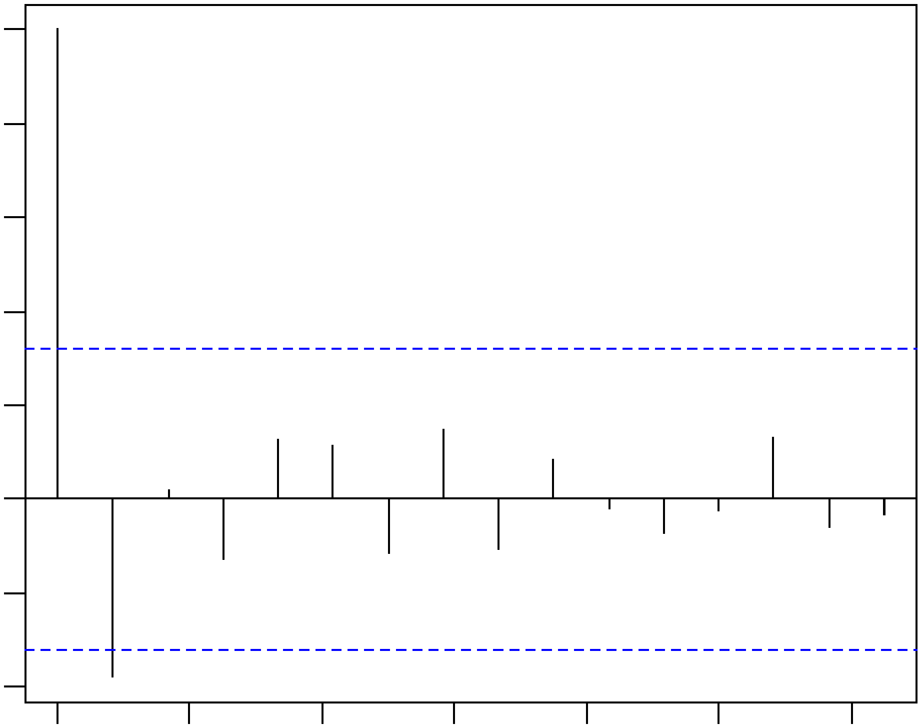
```
## [1] "Overall SD of differenced series:"
## [1] 4.758113
## Class 'cpt' : Changepoint Object
##        ~~   : S4 class containing 12 slots with names
##               cpttype date version data.set method test.stat pen.type
pen.value minseglen cpts ncpts.max param.est
##
## Created on  : Thu Dec 15 03:47:50 2022
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.83275
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations : 16
```

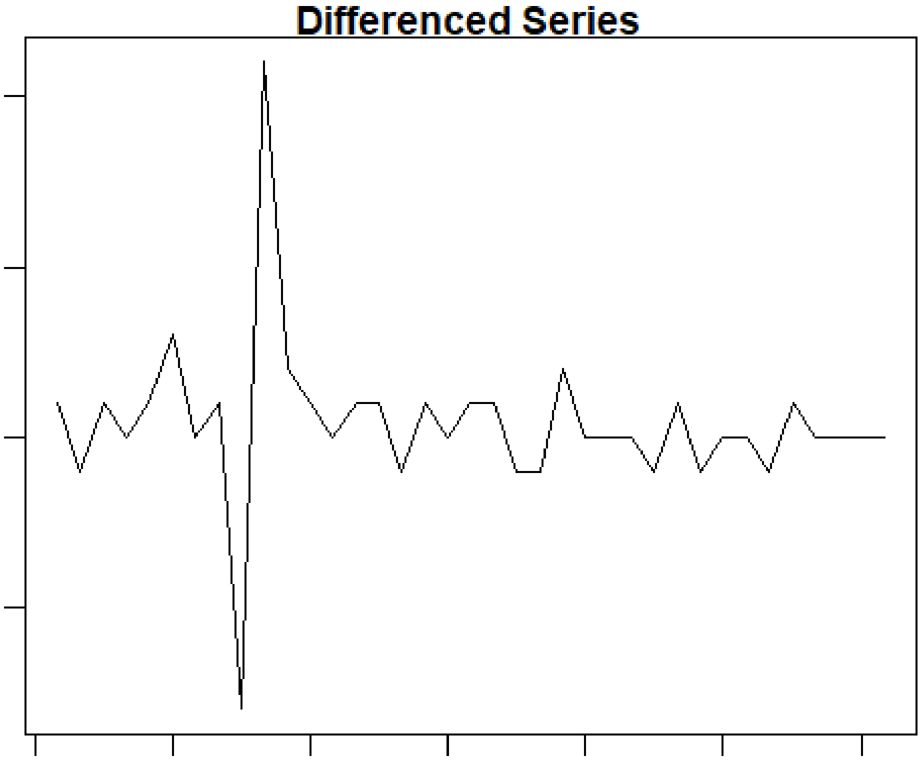


```
# DTP1
testSeries5 <- ts(usVaccines1$DTP1, frequency = 12)
runTStests1(testSeries5,"DTP1" )

## [1] "DTP1"
```



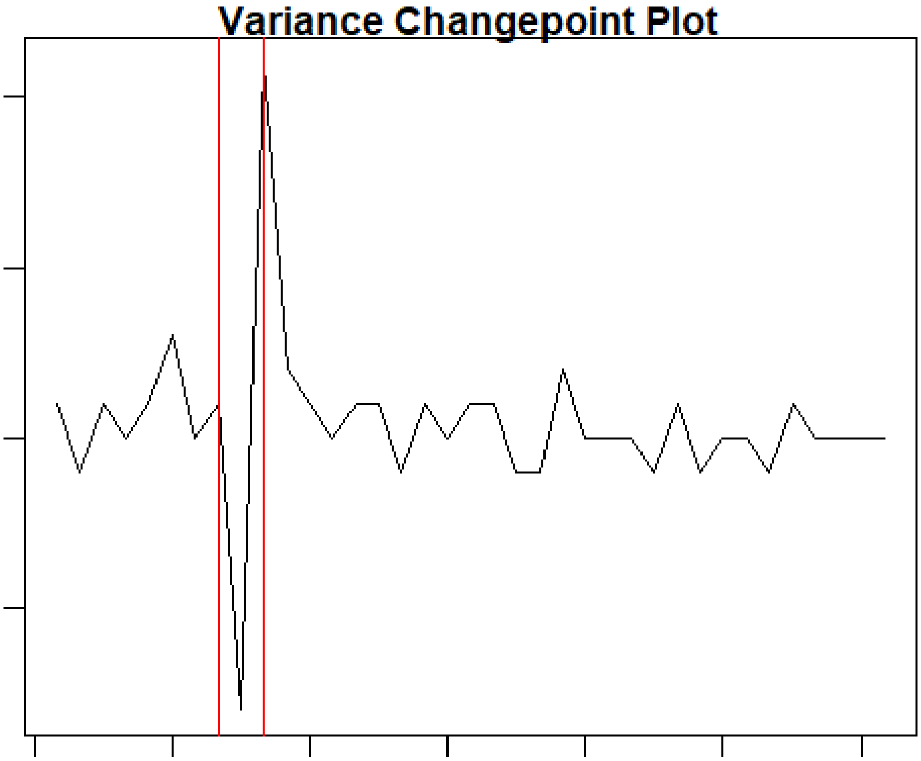
```
## [1] "Amount of decline Over the years"
## [1] 18
## [1] "Total change in trend:"
## [1] 15
## [1] "Max value:"
## [1] 99
## [1] "Min value:"
## [1] 81

testSeries5_1 <- ts(usVaccines1, frequency = 12)
runTStests2(testSeries5_1,"DTP1" )

## Warning in adf.test(diffTestSeries): p-value smaller than printed p-value
```





Differenced Series

```
## [1] "Overall SD of differenced series:"
## [1] 2.443352
## 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.max param.est
##
## Created on  : Thu Dec 15 03:47:50 2022
##
## summary(.)  :
## ----------
## Created Using changepoint version 2.2.4
## Changepoint type      : Change in variance
## Method of analysis    : PELT
## Test Statistic  : Normal
## Type of penalty       : MBIC with value, 10.83275
## Minimum Segment Length : 2
## Maximum no. of cpts   : Inf
## Changepoint Locations : 8 10
```



```
#The Hib3 vaccine had the highest overall SD of differenced series being
8.347106, thus having the greatest volatility.

# When we plot the differencing trends we see these difference over the
subjected time period.  The red line  in the graph signifies the change-point
location. The change-point shows us where there is a significant shift occurs
in the mean level during a certain period of time.
```

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?

```r
#Contingency Table
ft1 <- ftable(allSchoolsReportStatus, row.vars = 2, col.vars = "reported")
ft1

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

total_reported<-1397+5584+148+252
#What proportion of public schools reported vaccination data?
(5584/total_reported)*100

## [1] 75.65371

#About 76% of public schools reported vaccination data.



#What proportion of private schools reported vaccination data?
(1397/total_reported)*100

## [1] 18.92697

#About 19% of private schools reported vaccination data.


#Was there any credible difference in overall reporting proportions between
public and private schools?
chisq.test(ft1,correct=F)

##
##  Pearson's Chi-squared test
##
## data:  ft1
## X-squared = 402.97, df = 1, p-value < 2.2e-16

#This is a standard null hypothesis test shows us the chi-squared value of
402.97 with a p-value of 2.2e-16 which is well below our alpha threshold of
p<.05. As a result we will reject the null hypothesis in support of the
alternative hypothesis of credible differences in overall reporting
proportions between public and private schools.
library(BayesFactor)

## Warning: package 'BayesFactor' was built under R version 4.2.2

## Loading required package: coda

## Loading required package: Matrix

## ************
## Welcome to BayesFactor 0.9.12-4.4. If you have questions, please contact
```

```
Richard Morey (richarddmorey@gmail.com).
##
## Type BFManual() to open the manual.
## ************

library(BEST)

## Warning: package 'BEST' was built under R version 4.2.2

## Loading required package: HDInterval

## Warning: package 'HDInterval' was built under R version 4.2.2

ctBFout <-contingencyTableBF(ft1,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

#When we take the Bayesian approach to the Chi-Square Test we get the Bayes
Factor of 1.150548e+69:1  in favor of the alternative hypothesis of credible
differences in overall reporting proportions between public and private
schools due the factor being in excess of our threshold of 3:1.
```

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)?

```
summary(districts)

##  DistrictName         WithoutDTP      WithoutPolio      WithoutMMR
##  Length:700          Min.   : 0.00   Min.   : 0.000   Min.   : 0.00
##  Class :character    1st Qu.: 3.00   1st Qu.: 3.000   1st Qu.: 3.00
##  Mode  :character    Median : 7.00   Median : 6.000   Median : 6.00
##                      Mean   :10.13   Mean   : 9.747   Mean   :10.18
##                      3rd Qu.:14.00   3rd Qu.:13.000   3rd Qu.:14.00
##                      Max.   :77.00   Max.   :77.000   Max.   :77.00
##   WithoutHepB      PctUpToDate     DistrictComplete PctBeliefExempt
##  Min.   : 0.000   Min.   : 23.00   Mode :logical    Min.   : 0.000
##  1st Qu.: 2.000   1st Qu.: 84.00   FALSE:43         1st Qu.: 1.000
##  Median : 4.000   Median : 92.00   TRUE :657        Median : 2.000
##  Mean   : 7.739   Mean   : 87.88                    Mean   : 5.683
##  3rd Qu.:10.000   3rd Qu.: 96.00                    3rd Qu.: 7.000
##  Max.   :77.000   Max.   :100.00                    Max.   :77.000
##  PctChildPoverty  PctFreeMeal     PctFamilyPoverty    Enrolled
##  Min.   : 2.00   Min.   :  0.00   Min.   : 0.00    Min.   :   10.00
##  1st Qu.:13.00   1st Qu.: 30.00   1st Qu.: 5.00    1st Qu.:   52.75
##  Median :20.00   Median : 49.50   Median : 9.00    Median :  202.00
##  Mean   :22.11   Mean   : 48.43   Mean   :11.27    Mean   :  610.63
##  3rd Qu.:29.00   3rd Qu.: 69.00   3rd Qu.:15.00    3rd Qu.:  673.50
##  Max.   :72.00   Max.   :100.00   Max.   :47.00    Max.   :54238.00
##   TotalSchools
##  Min.   :  1.000
##  1st Qu.:  1.000
##  Median :  3.000
##  Mean   :  7.081
##  3rd Qu.:  8.000
##  Max.   :582.000
```

```
#Find individual rates for vaccines in California public schools 2013
#DTP vaccine rate 89.87%
100-(mean(districts$WithoutDTP))
```

```
## [1] 89.86571
```

```
#Polio vaccine rate 90.25%
100-(mean(districts$WithoutPolio))
```

```
## [1] 90.25286
```

```
#MMR vaccine rate 89.82%
100-(mean(districts$WithoutMMR))
```

```
## [1] 89.82
```

```
#Hepatitis B vaccine rate 92.26%
100-(mean(districts$WithoutHepB))
```

```
## [1] 92.26143
```

```
summary(usVaccines)
```

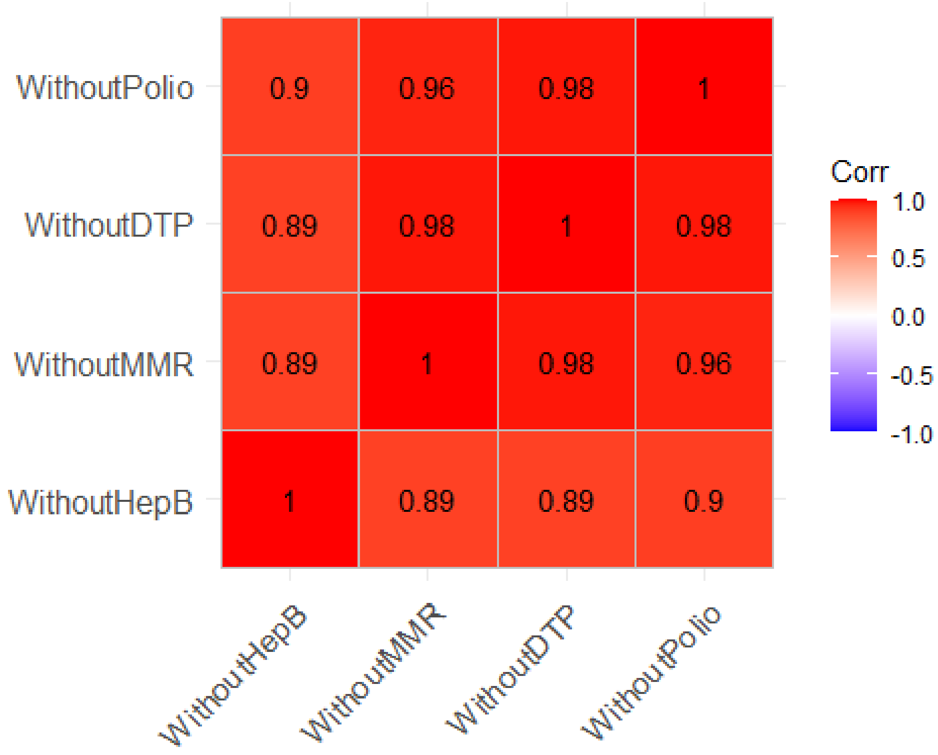
```
##       DTP1          HepB_BD           Pol3            Hib3
##  Min.   :81.00   Min.   :11.00   Min.   :24.00   Min.   :52.00
##  1st Qu.:89.75   1st Qu.:17.00   1st Qu.:90.00   1st Qu.:87.00
##  Median :97.00   Median :19.00   Median :93.00   Median :91.00
##  Mean   :94.05   Mean   :34.21   Mean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

```
tail(usVaccines)

##       DTP1 HepB_BD Pol3 Hib3 MCV1
## [33,]   97      72   93   93   91
## [34,]   98      74   93   93   92
## [35,]   98      72   93   93   92
## [36,]   98      72   93   93   92
## [37,]   98      64   94   93   92
## [38,]   98      64   94   93   92

print(usVaccines[34,])

##    DTP1 HepB_BD    Pol3    Hib3    MCV1
##      98      74      93      93      92

#   When we compare the individual vaccines in California districts with the
overall US vaccination rates we could take a look at their averages in
addition to taking a closer look at the rates in the US during 2013
specifically. We see that for the Measles vaccine the average rate of about
92% is which is close to the mean of the Measles vaccine of the overall US
vaccination rate of 89%%. We also see that the DTP vaccine average rate in
California was around 90% compared to that of the overall Us vaccination time
series average DTP vaccination rate of about 94%. When we look at Polio
vaccine for California we see an average rate of about 90% when compared to
the that of the usVaccines data set with an average mean of about 87%. In
addition when we look at the average rates of the Hepatitis B vaccination
rates we find interesting results.
#   We see that average rate for the Hepatitis B vaccine in California public
schools was about 92% when compare to the vaccination rates for the
USvaccines data set has an average of about 34%, this will cause us to look
at our time series to compare actual estimated rates during the 2013 year. We
do notice in out time series that there is a very large spike in the
vaccination rates of the Hepatitis B vaccine over the years. If we take a
further look at our time series we see the rates for the year 2013 for DTP1
being 98%, HepB 74%, Pol3, 93%, and MCv1 92%. These rates for the overall US
are much closer when comparing the time series to the vaccination rates of
the California public school districts. Thus when we look at the plotted time
series for HepB we can assume that California's high vaccination rate in
Hepatitis B had an effect of some kind on the overall US Vaccination rate for
Hepatitis B increasing around that period.
```

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?

```
is.ts(districts)

## [1] FALSE

districts10N<- districts[,-1]
districts10N<- data.frame(districts10N)
cor1<-cor(districts10N[1:4])
cor1
```

```
##              WithoutDTP WithoutPolio WithoutMMR WithoutHepB
## WithoutDTP    1.0000000    0.9811088  0.9759526   0.8902883
## WithoutPolio  0.9811088    1.0000000  0.9644609   0.9048356
## WithoutMMR    0.9759526    0.9644609  1.0000000   0.8907565
## WithoutHepB   0.8902883    0.9048356  0.8907565   1.0000000

library(ggcorrplot)

## Loading required package: ggplot2

ggcorrplot(cor1,hc.order = TRUE,lab = TRUE)
```



```
#If there is a percentage of students who are missing one vaccine are they
missing all of the others we can see this in the correlation matrix that this
idea may be true due to the high positive correlation between each of the
variables pertaining not having a specific vaccine. We can assume that this
may be because of percentage of students enrolled with belief exceptions, but
we would have to do more research to confirm this.
```

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

```
#Change Districtcomplete column into numeric values.
districts10<-districts
districts10$DistrictComplete<-as.numeric(districts10$DistrictComplete)
#view(districts10) #True = 1 False = 0
districts10N<-data.frame(scale(districts10[,-1],center=T,scale = F))

#Linear Regression
```

```r
regg1<- lm(DistrictComplete ~ PctChildPoverty + PctFreeMeal+
PctFamilyPoverty+ Enrolled+ TotalSchools, data=districts10)
summary(regg1)
```

```
##
## Call:
## lm(formula = DistrictComplete ~ PctChildPoverty + PctFreeMeal +
##     PctFamilyPoverty + Enrolled + TotalSchools, data = districts10)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99516  0.02074  0.04580  0.07585  0.65176
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.007e+00  2.074e-02  48.580  < 2e-16 ***
## PctChildPoverty   1.530e-03  1.521e-03   1.006    0.315
## PctFreeMeal      -8.203e-04  5.442e-04  -1.507    0.132
## PctFamilyPoverty -3.556e-03  2.217e-03  -1.604    0.109
## Enrolled          2.183e-04  4.042e-05   5.400 9.14e-08 ***
## TotalSchools     -2.204e-02  3.740e-03  -5.894 5.88e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2301 on 694 degrees of freedom
## Multiple R-squared:  0.0899, Adjusted R-squared:  0.08335
## F-statistic: 13.71 on 5 and 694 DF,  p-value: 8.749e-13
```

```r
# The results show the the Enrolled P-value 9.14e-08 and  TotalSchools P-
value 5.88e-09variables predict predict whether or not a district’s reporting
was complete.  These variables P-values are under our threshold of .05, in
addition our overall regression analysis has a p-value of 8.749e-13 meaning
that our overall analysis is statically significant, and we can reject the
null hypothesis in support of the alternative hypothesis.



#library(BEST)
#library(BayesFactor)

#Bayes Approach to Linear Regression
reggbf1 <- lmBF(DistrictComplete ~ PctChildPoverty + PctFreeMeal+
PctFamilyPoverty+ Enrolled+ TotalSchools, data=districts10N)
summary(reggbf1)
```

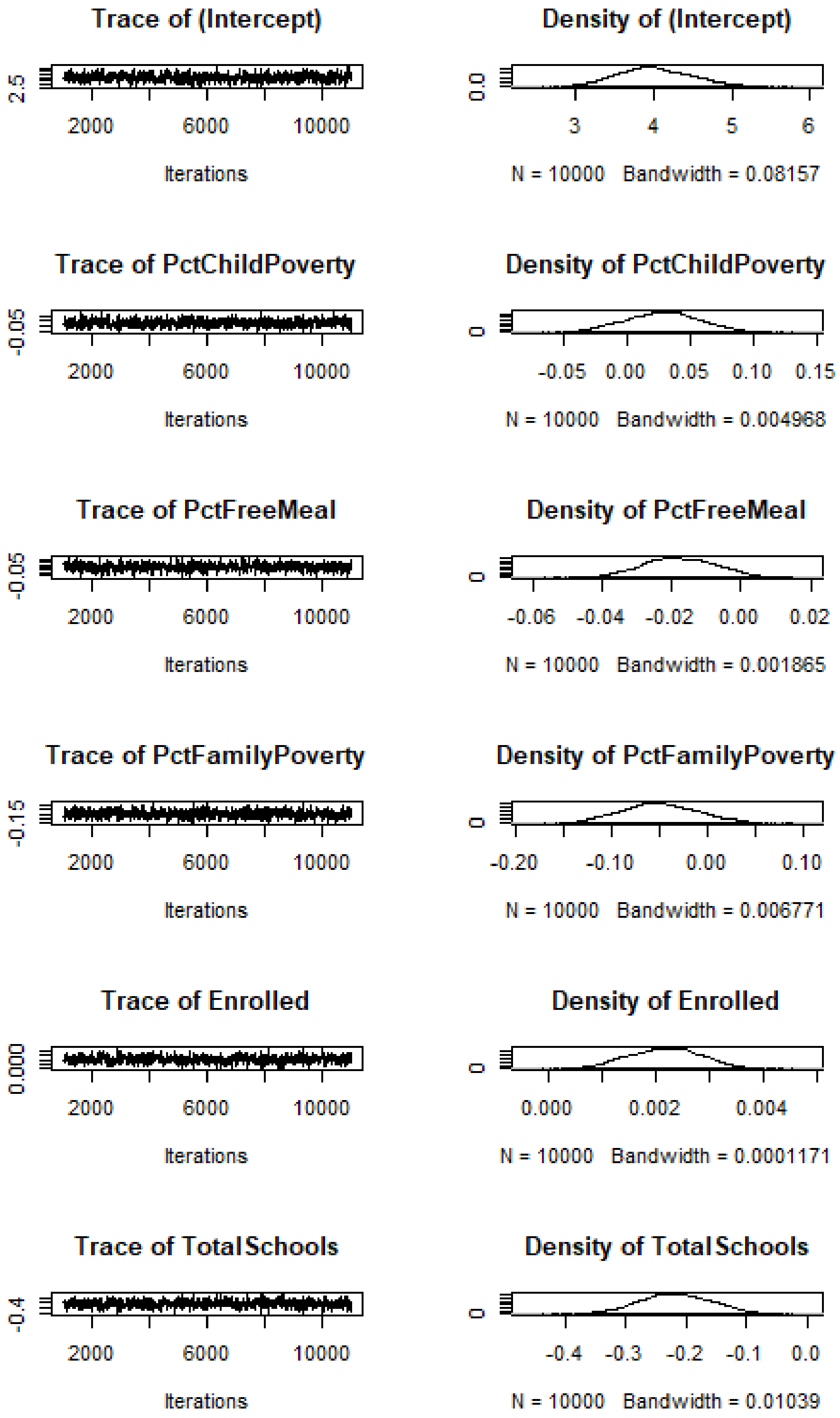
```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled +
TotalSchools : 3017311149 ±0.01%
##
```

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

```
#   With the Bayes approach to linear regression we get a Bayes Factor of
3017311149 , which is which is well over the odds cut off of 3:1. This
analysis results support our alternative hypothesis, and we will as a result
reject the null hypothesis. This lines up with our liner regression as our
analysis had a p-value of 8.749e-13, which is lower than .05, thus we will
reject the null hypothesis as well, in support with our alternative
hypothesis


#General Linear Model
reg1_1 <- glm(DistrictComplete ~ PctChildPoverty + PctFreeMeal+
PctFamilyPoverty+ Enrolled+ TotalSchools, data=districts10, family =
binomial())
summary(reg1_1)
```

```
##
## Call:
## glm(formula = DistrictComplete ~ PctChildPoverty + PctFreeMeal +
##     PctFamilyPoverty + Enrolled + TotalSchools, family = binomial(),
##     data = districts10)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7740   0.2359   0.2962   0.3617   2.2593
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.932783   0.474947   8.280  < 2e-16 ***
## PctChildPoverty   0.029277   0.029934   0.978 0.328043
## PctFreeMeal      -0.017344   0.010593  -1.637 0.101570
## PctFamilyPoverty -0.052975   0.039897  -1.328 0.184246
## Enrolled          0.002184   0.000650   3.360 0.000781 ***
## TotalSchools     -0.213030   0.058580  -3.637 0.000276 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 323.23  on 699  degrees of freedom
## Residual deviance: 287.43  on 694  degrees of freedom
## AIC: 299.43
##
## Number of Fisher Scoring iterations: 6
```

```
# The results in our general linear model show the the  Enrolled P-value
0.012984  , TotalSchools P-value 0.017414  variables predict whether or not a
district’s reporting was complete.  These variables P-values are under our
threshold of .05, which are the same variables from our regression model that
we found to be the predictors in support of our alternative hypothesis.

#Bayseian Estimation of Logistic Regression
library(MCMCpack)

## Warning: package 'MCMCpack' was built under R version 4.2.2

## Loading required package: MASS

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

## ## Copyright (C) 2003-2022 Andrew D. Martin, Kevin M. Quinn, and Jong Hee
Park

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

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

regbf1_1<- MCMClogit(DistrictComplete ~ PctChildPoverty + PctFreeMeal+
PctFamilyPoverty+ Enrolled+ TotalSchools, data=districts10 )
summary(regbf1_1)

##
## Iterations = 1001:11000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                      Mean        SD  Naive SE Time-series SE
## (Intercept)       4.02397 0.4855533 4.856e-03      2.177e-02
## PctChildPoverty   0.02814 0.0295744 2.957e-04      1.211e-03
## PctFreeMeal      -0.01786 0.0111017 1.110e-04      5.053e-04
## PctFamilyPoverty -0.04880 0.0403062 4.031e-04      1.745e-03
## Enrolled          0.00217 0.0006972 6.972e-06      3.193e-05
## TotalSchools     -0.21778 0.0618717 6.187e-04      2.793e-03
##
## 2. Quantiles for each variable:
##
##                        2.5%       25%       50%       75%     97.5%
## (Intercept)       3.1236516  3.690819  3.991884  4.350473  5.045506
## PctChildPoverty  -0.0289039  0.007345  0.028361  0.048042  0.086296
## PctFreeMeal      -0.0403494 -0.025268 -0.017988 -0.010224  0.003079
## PctFamilyPoverty -0.1237541 -0.076047 -0.050505 -0.021629  0.033207
## Enrolled          0.0007954  0.001687  0.002183  0.002646  0.003491
## TotalSchools     -0.3386296 -0.258565 -0.217799 -0.175504 -0.096965

# With the Bayes approach to linear regression the 95% HDI ranges for the
Enrolled predictor is 0.0007954 to 0.003491  with a mean of 0.00217  which is
very close to the coefficient results from the general linear regression of
0.002184. With the Bayes approach to linear regression the 95% HDI ranges for
the hp(Gross horsepower) predictor is -0.3386296 to -0.096965, with a mean
of-0.213030 which is very close to the coefficient results from the general
linear regression of -0.21778. Since none of the interval pass 0 our results
are statistically significant, and we can reject the null hypothesis, in
support of the alternative hypothesis.

# We can get a more detailed view of these HDIs in the graphs belows
plot(regbf1_1)
```



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

```
reg3 <- lm(PctUpToDate ~ PctChildPoverty + PctFreeMeal+
PctFamilyPoverty+Enrolled+ TotalSchools, data=districts10)
summary(reg3)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ PctChildPoverty + PctFreeMeal +
PctFamilyPoverty +
##     Enrolled + TotalSchools, data = districts10)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.580  -3.259   3.093   7.064  17.502
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      82.365450   1.078196  76.392  < 2e-16 ***
## PctChildPoverty  -0.137897   0.079098  -1.743 0.081714 .
## PctFreeMeal       0.096061   0.028298   3.395 0.000726 ***
## PctFamilyPoverty  0.354470   0.115272   3.075 0.002187 **
## Enrolled          0.005234   0.002101   2.491 0.012984 *
## TotalSchools     -0.463518   0.194465  -2.384 0.017414 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.96 on 694 degrees of freedom
## Multiple R-squared:  0.09362,    Adjusted R-squared:  0.08709
## F-statistic: 14.34 on 5 and 694 DF,  p-value: 2.247e-13
```

```
# The results show the the PctFamilyPoverty P-value 0.002187 , PctFreeMeal P-
value 0.000726, Enrolled P-value 0.012984  , TotalSchools P-value 0.017414
variables predict the percentage of all enrolled students with completely up-
to-date vaccines.  These variables P-values are under our threshold of .05,
in addition our overall regression analysis has a p-value of 2.247e-13
meaning that our overall analysis is statically significant, and we can
reject the null hypothesis, supporting the alternative hypothesis of
PctFamilyPoverty, PctFreeMeal, TotalSchools, and Enrolled variables predict
the percentage of all enrolled students with completely up-to-date vaccines.

regbf3<-lmBF(PctUpToDate ~ PctChildPoverty +
PctFreeMeal+PctFamilyPoverty+Enrolled+ TotalSchools, data=districts10N)
summary(regbf3)
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled +
TotalSchools : 11886710731 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS

#With the Bayes approach to linear regression we get a Bayes Factor of
11886710731  which is which is well over the odds cut off of 3:1. This
analysis results support our alternative hypothesis, and we will as a result
reject the null hypothesis. This lines up with our liner regression as our
analysis had a p-value of 2.247e-13, which is lower than .05, thus we will
reject the null hypothesis as well, in support with our alternative
hypothesis that PctFamilyPoverty, PctFreeMeal, TotalSchools, and Enrolled
variables predict the percentage of all enrolled students with completely up-
to-date vaccines.
```

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

```
# Linear Regression
reg4<-lm(PctBeliefExempt ~ PctChildPoverty + PctFreeMeal+
PctFamilyPoverty+Enrolled+ TotalSchools, data=districts10N)
summary(reg4)

##
## Call:
## lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFreeMeal +
##     PctFamilyPoverty + Enrolled + TotalSchools, data = districts10N)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.739  -3.956  -1.952   0.764  65.480
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -3.529e-15  3.127e-01   0.000 1.000000
## PctChildPoverty   1.865e-01  5.470e-02   3.410 0.000688 ***
## PctFreeMeal      -1.203e-01  1.957e-02  -6.145 1.35e-09 ***
## PctFamilyPoverty -2.403e-01  7.972e-02  -3.014 0.002671 **
## Enrolled         -2.613e-03  1.453e-03  -1.798 0.072623 .
## TotalSchools      2.200e-01  1.345e-01   1.636 0.102346
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.273 on 694 degrees of freedom
## Multiple R-squared:  0.1232, Adjusted R-squared:  0.1169
## F-statistic:  19.5 on 5 and 694 DF,  p-value: < 2.2e-16

# The results show the the PctFamilyPoverty P-value 0.002671, PctFreeMeal P-
value 1.35e-09, and PctChildPoverty P-value 0.000688 , variables predict the
percentage of all enrolled students with belief exceptions.  These variables
P-values are under our threshold of .05, in addition our overall regression
analysis has a p-value of 2.2e-16 meaning that our overall analysis is
statically significant, and we can reject the null hypothesis in support of
```

```
the alternative hypothesis.

# Bayes Approach to Linear Regression
regbf4 <- lmBF(PctBeliefExempt ~ PctChildPoverty + PctFreeMeal+
PctFamilyPoverty+Enrolled+ TotalSchools, data=districts10N)
summary(regbf4)

## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled +
TotalSchools : 8.276841e+14 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS

#With the Bayes approach to linear regression we get a Bayes Factor of
8.276841e+14, which is which is well over the odds cut off of 3:1. This
analysis results support our alternative hypothesis, and we will as a result
reject the null hypothesis. This lines up with our liner regression as our
analysis had a p-value of 2.2e-16, which is lower than .05, thus we will
reject the null hypothesis as well, in support with our alternative
hypothesis that PctFamilyPoverty, PctFreeMeal, PctChildPoverty, variables
predict the percentage of all enrolled students with belief exceptions.   .
library(car)

## Loading required package: carData

vif(reg4)

##  PctChildPoverty      PctFreeMeal PctFamilyPoverty         Enrolled
##         4.410932         2.380507         4.147192       105.713128
##     TotalSchools
##       105.572378

#As previously stated, a "Rule of Thumb" we could use when interpreting vif()
is that a value of 1 means that the predictor variable is not correlated with
other variables. The higher the value, the greater the correlation of the
variable with other variables. We see that the PctFamilyPoverty, PctFreeMeal,
PctChildPoverty variables, that predict the percentage of all enrolled
students with belief exceptions all have low valued VIF scores while the two
variable that were not correlated to the outcome variable have similar
extremely high VIF scores.
```

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?

```
#   When we look at the districts data set as a sample population we see that
there is an average of about 88% of student who have their vaccinations up to
date. We see that the variable PctFamilyPoverty comes up as a statistically
significant predictor to both the PctUpToDate variable which is the
percentage of enrolled students with up to date vaccines, and PctBeliefExempt
which is the percentage of all enrolled students with belief exceptions.
#   We believe that a good idea for state legislators to allocate financial
assistance to school districts in a few areas as suggestions. For starters
there is a high correlation that if a student is missing one vaccination they
is a high chance they are missing all, and we see the number of enrolled
students is a predictor for this variable. We believe putting some financial
assistance in enrollment effort would help increase the percentage of
completed vaccine reporting. In addition we see the area of poverty plays a
role whether or not students are vaccinated, we believe investing in free
vaccination sites and programs for student would help legislators reach their
goals of improving both their vaccination rates and their reporting
compliance. We also believe that it may help in decreasing the average of
students who have beliefs exceptions, who may in actuality have issues
affording the fees for the vaccine for their families.
#   We also see that the there is a large average of students who are
eligible for free meals, this variable came up a predictor for the percentage
of student who are up to date with their vaccinations. We can conclude that
further assistance in free student meals will both increase enrollment rates
thus increasing the compliance reporting and student vaccination rates. In
addition a good measure may be to host more educational programs of the
benefits and myths of vaccines to decrease the number of student with belief
exceptions.
#   We feel that the big picture is that a large number of student have their
vaccines, which is almost as high as the Us vaccination rates between 1980-
2017. We can also see that there is a correlation when a student is missing a
vaccine chances are they are missing all of their vaccine. We see that there
are reasons such as poverty, schools actual reporting was complete, and
belief exceptions, but that only looks to be the reason why a small
```

```
percentage of students do not have their vaccine and it is something that we
should do further research in.
```