# IST772, Standard Homework Heading

# Student name: Kelsey Johnson

# Homework number: Final Examination

# Date due: 6/27/21

# Attribution statement: (choose only one)

# I did this homework by myself, with help from the book

# 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  
# Set working directory   
setwd("C:/Users/HP/OneDrive/Desktop/Grad School/IST772 Quantitative Reasoning/Week 10/Final") # Change to the folder containing your homework data files

The datasets are:

• usVaccines.Rdata • allSchoolsReportStatus.RData • districtsX.RData

Here is a description of each dataset:

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

Time-Series [1:38, 1:5] from 1980 to 2017: - attr(\*, “dimnames”)=List of 2 ..$ : NULL ..$ : chr [1:5] “DTP1” “HepB\_BD” “Pol3” “Hib3” “MCV1”…

(Note: DTP1 = First dose of Diphtheria/Pertussis/Tetanus vaccine; HepB\_BD = Hepatitis B, Birth Dose; Pol3 = Polio third dose; Hib3 – Influenza third dose; MCV1 = Measles first dose)

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

‘data.frame’: 7381 obs. of 3 variables: $ name : Name of the school $ pubpriv : “PUBLIC” or “PRIVATE” $ reported: “Y” or “N”

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:

‘data.frame’: 700 obs. of 13 variables: $ 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

## Research Questions:

## Question 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("usVaccines.RData")  
load("allSchoolsReportStatus.RData")  
load("districts13.RData")  
library(changepoint)

## Warning: package 'changepoint' was built under R version 4.0.5

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.4

##   
## Attaching package: 'zoo'

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

## Successfully loaded changepoint package version 2.2.2  
## NOTE: Predefined penalty values changed in version 2.2. Previous penalty values with a postfix 1 i.e. SIC1 are now without i.e. SIC and previous penalties without a postfix i.e. SIC are now with a postfix 0 i.e. SIC0. See NEWS and help files for further details.

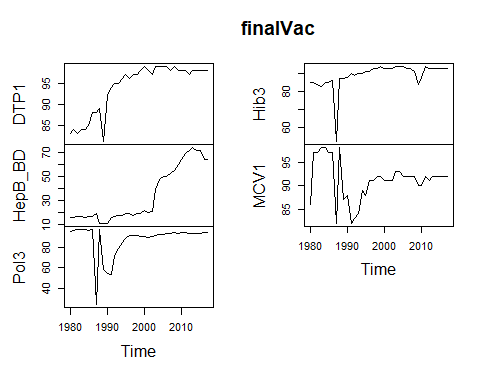
finalVac<-usVaccines  
str(finalVac)

## 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(finalVac)

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

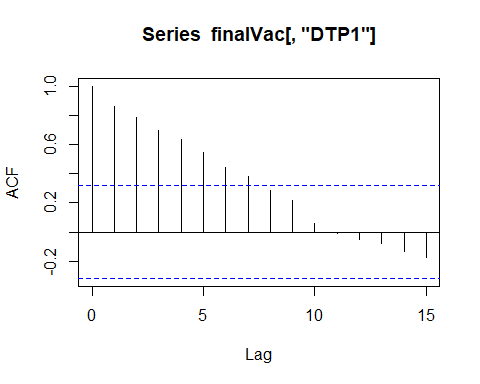
plot(finalVac)



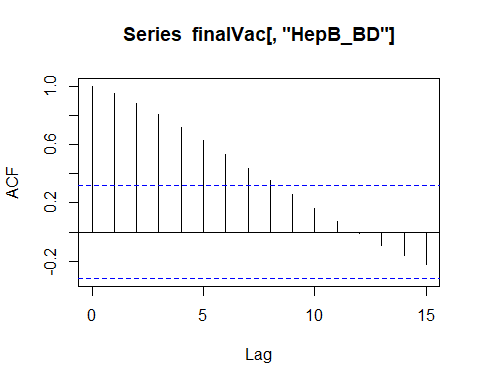
cor(finalVac)

## 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

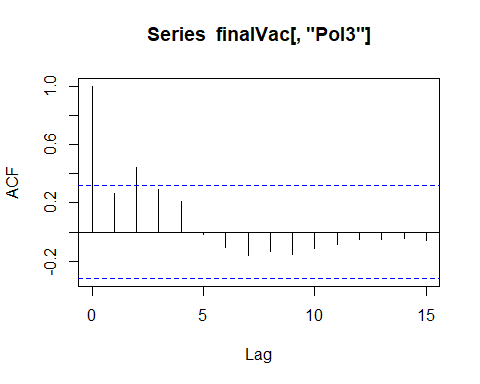
acf(finalVac[,"DTP1"],)



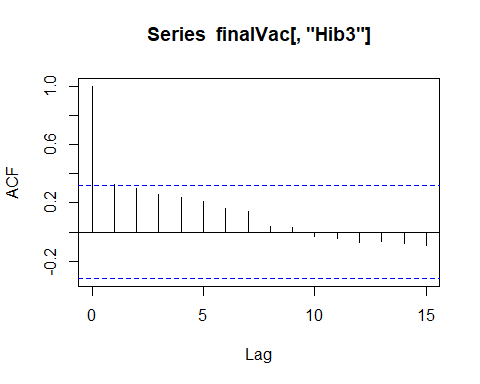
acf(finalVac[,"HepB\_BD"])



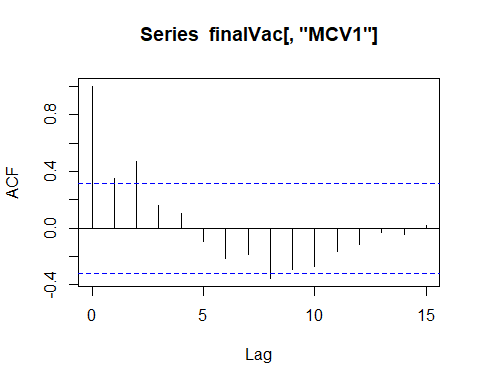
acf(finalVac[,"Pol3"])



acf(finalVac[,"Hib3"])



acf(finalVac[,"MCV1"])



plot(cpt.var(finalVac[,"Pol3"]))

Chart, histogram

Description automatically generated

plot(cpt.mean(finalVac[,"Pol3"]))

Chart

Description automatically generated

Vaccine rates in general have trended higher since 1980 year over year with some variability. However, DTP1, POL3 and Hib3 all continued growing until around 2000 and have remained fairly steady except for a small dip in Hib3 around 2010. HepB\_BD had continued to increase through 2010 and began to decrease slightly afterwards. By running an autocorrelation analysis on the individual vaccines, the strength and whether a trend is present or not is visible. Three of the vaccines (DTP1, HepB\_BD, and Hib3) all show a trend. They also have the strongest correlation with DTP1 and HepB\_BD having the highest at 0.5905157.

MCV1 had a large amount of vulnerability between 1987 when it dropped to a count of 82 and rose to a count of 98 by 1988. Then it had another dip beginning in 1989 through 1991 before it finally began to increase and remain steady since 2000 with a count of 90-93 until the end of our data. Overall, vaccine rates are increasing. DTP1 had the highest rate in 2017 at the end of the data at 98. HepB\_BD had the lowest rate at conclusion with 64. POL3 had the greatest volatility with the lowest count being 24 and highest count being 97 (range of 73). The change point analysis visuals show where the difference in variance and mean values shifted.

## Question 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?

#allSchoolsReportStatus  
schoolstatus<-allSchoolsReportStatus  
#schoolstatus  
str(schoolstatus)

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

schoolvaccines<-matrix(c(252,1397,148,5584), ncol=2, byrow=TRUE)  
colnames(schoolvaccines)<-c("No", "Yes")  
rownames(schoolvaccines)<-c("PRIVATE", "Public")  
schoolvaccines<-as.table(schoolvaccines)  
#schoolvaccines  
#vaccine grand total  
margin.table(schoolvaccines)

## [1] 7381

#marginal row totals  
margin.table(schoolvaccines,1)

## PRIVATE Public   
## 1649 5732

#marginal column totals  
margin.table(schoolvaccines,2)

## No Yes   
## 400 6981

#Find vaccine probabilities  
schoolprobs<-schoolvaccines/margin.table(schoolvaccines)  
schoolprobs

## No Yes  
## PRIVATE 0.03414172 0.18926975  
## Public 0.02005148 0.75653705

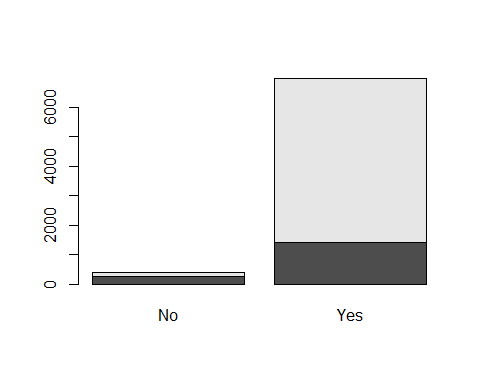
#Total Probability of No vs Yes  
schoolmargins<-margin.table(schoolvaccines,2)/margin.table(schoolvaccines)  
schoolmargins

## No Yes   
## 0.0541932 0.9458068

#Total Probability of Public vs Private  
schoolmargins2<-margin.table(schoolvaccines,1)/margin.table(schoolvaccines)  
schoolmargins2

## PRIVATE Public   
## 0.2234115 0.7765885

barplot(schoolvaccines)



Public schools reported a proportion of 77.66% (0.7765885) while private schools reported a proportion of 22.34% (0.2234115) from the total of 7381 schools. There is a large difference in reporting between public and private schools. For private schools, 15.28% of the total of private schools reported No for vaccines while 84.7% reported yes. For public schools, only 2.58% reported No for vaccines with 97.42% reporting yes. This is a gap of 12.72% when comparing private and public schools who reported vaccination data. The total reporting of both private and public is 94.6% while total of 5.4% of both schools did not report vaccinations.

## Question 3

What are 2013 vaccination rates for individual vaccines 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(finalVac)

## 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

#districts  
californiaschools<-districts  
str(districts)

## 'data.frame': 700 obs. of 13 variables:  
## $ DistrictName : chr "Happy Valley Elementary" "Raymond-Knowles Union Elementary" "South Pasadena Unified" "Saugus Union" ...  
## $ WithoutDTP : num 37 10 12 10 9 1 7 6 5 5 ...  
## $ WithoutPolio : num 42 10 10 10 10 1 5 5 5 4 ...  
## $ WithoutMMR : num 47 10 12 9 10 1 7 4 4 5 ...  
## $ WithoutHepB : num 37 10 10 6 11 1 3 4 4 4 ...  
## $ PctUpToDate : num 47 90 85 89 85 99 90 92 95 91 ...  
## $ DistrictComplete: logi TRUE FALSE TRUE FALSE TRUE TRUE ...  
## $ PctBeliefExempt : num 32 0 8 4 6 0 2 2 3 0 ...  
## $ PctChildPoverty : num 8 30 9 7 13 37 9 29 8 24 ...  
## $ PctFreeMeal : num 8 63 9 17 2 80 36 58 13 69 ...  
## $ PctFamilyPoverty: num 1 6 5 3 4 25 4 15 4 13 ...  
## $ Enrolled : num 19 10 365 1120 175 ...  
## $ TotalSchools : num 1 2 3 15 1 4 3 20 4 17 ...

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 :10.34 Mean : 9.923 Mean :10.32   
## 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. : 23.0 Mode :logical Min. : 0.000   
## 1st Qu.: 2.000 1st Qu.: 84.0 FALSE:43 1st Qu.: 1.000   
## Median : 4.000 Median : 92.0 TRUE :657 Median : 2.000   
## Mean : 7.863 Mean : 87.7 Mean : 5.746   
## 3rd Qu.:10.000 3rd Qu.: 96.0 3rd Qu.: 7.000   
## Max. :77.000 Max. :100.0 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.: 29.00 1st Qu.: 5.00 1st Qu.: 51.75   
## Median :21.00 Median : 50.00 Median : 9.00 Median : 196.50   
## Mean :22.28 Mean : 48.29 Mean :11.39 Mean : 629.71   
## 3rd Qu.:30.00 3rd Qu.: 69.00 3rd Qu.:15.00 3rd Qu.: 675.75   
## Max. :72.00 Max. :100.00 Max. :47.00 Max. :54238.00   
## TotalSchools   
## Min. : 1.000   
## 1st Qu.: 1.000   
## Median : 3.000   
## Mean : 7.253   
## 3rd Qu.: 8.000   
## Max. :582.000

#California Totals  
DTP<-1-(sum(californiaschools$WithoutDTP)/sum(californiaschools$Enrolled))  
DTP

## [1] 0.9835799

Polio<-1-(sum(californiaschools$WithoutPolio)/sum(californiaschools$Enrolled))  
Polio

## [1] 0.9842423

MMR<-1-(sum(californiaschools$WithoutMMR)/sum(californiaschools$Enrolled))  
MMR

## [1] 0.9836139

HepB<-1-(sum(californiaschools$WithoutHepB)/sum(californiaschools$Enrolled))  
HepB

## [1] 0.9875136

#total up to date  
sum(californiaschools$PctUpToDate)

## [1] 61390

#Total Enrolled  
sum(californiaschools$Enrolled)

## [1] 440800

#Total Without DTP  
sum(californiaschools$WithoutDTP)

## [1] 7238

#Total Without Polio  
sum(californiaschools$WithoutPolio)

## [1] 6946

#Total Without MMR  
sum(californiaschools$WithoutPolio)

## [1] 6946

#Total Without HepB  
sum(californiaschools$WithoutPolio)

## [1] 6946

#Comparison of US to CA  
california<-c(98.4, 98.4, 98.4, 98.8)  
us<-c(94.05, 87.16, 91.24, 34.21)  
comparison<-data.frame(california, us)  
comparison

## california us  
## 1 98.4 94.05  
## 2 98.4 87.16  
## 3 98.4 91.24  
## 4 98.8 34.21

The 2013 rates for individual vaccines in public schools are 98.4% for DTP, Polio, and MMR while HepB is 98.8%. These rates are higher than the average US vaccination rates for all 4 vaccines. The largest difference coming from HepB which California has a rate of 98.8% while the US is only at 34.21%. There are 5.7% of students in California who have not been vaccinated due to a belief exemption. One reason why California is higher than the rest of the US appears to be because they require all children who attend private or public schools to be vaccinated. There are several states who offer medical, religious and philosophical exemptions which can cause the overall US rate to decline.

## Question 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?

str(californiaschools)

## 'data.frame': 700 obs. of 13 variables:  
## $ DistrictName : chr "Happy Valley Elementary" "Raymond-Knowles Union Elementary" "South Pasadena Unified" "Saugus Union" ...  
## $ WithoutDTP : num 37 10 12 10 9 1 7 6 5 5 ...  
## $ WithoutPolio : num 42 10 10 10 10 1 5 5 5 4 ...  
## $ WithoutMMR : num 47 10 12 9 10 1 7 4 4 5 ...  
## $ WithoutHepB : num 37 10 10 6 11 1 3 4 4 4 ...  
## $ PctUpToDate : num 47 90 85 89 85 99 90 92 95 91 ...  
## $ DistrictComplete: logi TRUE FALSE TRUE FALSE TRUE TRUE ...  
## $ PctBeliefExempt : num 32 0 8 4 6 0 2 2 3 0 ...  
## $ PctChildPoverty : num 8 30 9 7 13 37 9 29 8 24 ...  
## $ PctFreeMeal : num 8 63 9 17 2 80 36 58 13 69 ...  
## $ PctFamilyPoverty: num 1 6 5 3 4 25 4 15 4 13 ...  
## $ Enrolled : num 19 10 365 1120 175 ...  
## $ TotalSchools : num 1 2 3 15 1 4 3 20 4 17 ...

summary(californiaschools)

## 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 :10.34 Mean : 9.923 Mean :10.32   
## 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. : 23.0 Mode :logical Min. : 0.000   
## 1st Qu.: 2.000 1st Qu.: 84.0 FALSE:43 1st Qu.: 1.000   
## Median : 4.000 Median : 92.0 TRUE :657 Median : 2.000   
## Mean : 7.863 Mean : 87.7 Mean : 5.746   
## 3rd Qu.:10.000 3rd Qu.: 96.0 3rd Qu.: 7.000   
## Max. :77.000 Max. :100.0 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.: 29.00 1st Qu.: 5.00 1st Qu.: 51.75   
## Median :21.00 Median : 50.00 Median : 9.00 Median : 196.50   
## Mean :22.28 Mean : 48.29 Mean :11.39 Mean : 629.71   
## 3rd Qu.:30.00 3rd Qu.: 69.00 3rd Qu.:15.00 3rd Qu.: 675.75   
## Max. :72.00 Max. :100.00 Max. :47.00 Max. :54238.00   
## TotalSchools   
## Min. : 1.000   
## 1st Qu.: 1.000   
## Median : 3.000   
## Mean : 7.253   
## 3rd Qu.: 8.000   
## Max. :582.000

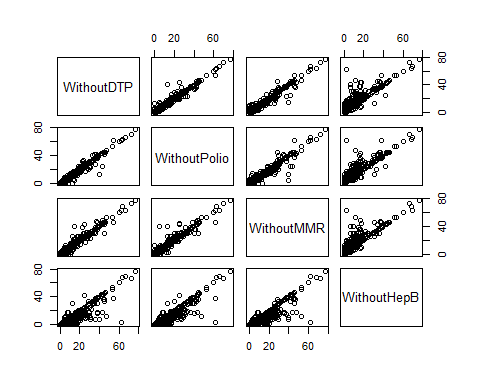
#transform District to Num and drop DistrictName  
californiaschools$DistrictComplete <- ifelse(californiaschools$DistrictComplete == "F",1,ifelse(californiaschools$DistrictComplete == "T",2,0))  
newcaliforniaschools<-californiaschools[c(-1, -6:-13)]  
str(newcaliforniaschools)

## 'data.frame': 700 obs. of 4 variables:  
## $ WithoutDTP : num 37 10 12 10 9 1 7 6 5 5 ...  
## $ WithoutPolio: num 42 10 10 10 10 1 5 5 5 4 ...  
## $ WithoutMMR : num 47 10 12 9 10 1 7 4 4 5 ...  
## $ WithoutHepB : num 37 10 10 6 11 1 3 4 4 4 ...

#Correlation  
cor(newcaliforniaschools)

## WithoutDTP WithoutPolio WithoutMMR WithoutHepB  
## WithoutDTP 1.0000000 0.9831203 0.9775468 0.8947948  
## WithoutPolio 0.9831203 1.0000000 0.9682234 0.9079517  
## WithoutMMR 0.9775468 0.9682234 1.0000000 0.8955538  
## WithoutHepB 0.8947948 0.9079517 0.8955538 1.0000000

plot(newcaliforniaschools)



#ANOVA  
anovaCAschools<-aov(WithoutDTP ~ WithoutPolio + WithoutMMR + WithoutHepB, data = newcaliforniaschools)  
summary(anovaCAschools)

## Df Sum Sq Mean Sq F value Pr(>F)   
## WithoutPolio 1 85055 85055 29480.36 <2e-16 \*\*\*  
## WithoutMMR 1 927 927 321.28 <2e-16 \*\*\*  
## WithoutHepB 1 11 11 3.74 0.0535 .   
## Residuals 696 2008 3   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The highest correlation is between without DTP and without Polio at 98.31%. This means if you don’t have the DTP vaccine, you are 98.31% likely to also not have had a Polio vaccine.The next highest is Without DTP and Without MMR at 97.7%. This means if you don’t have the DTP vaccine, you are 97.7% likely to also not have had an MMR vaccine. The third highest is if you don’t have polio you won’t have MMR at 96.82%. Either way, the unvaccinated rates in school remains at 5.7%. The ANOVA shows that with the Sum of Squares for variance that there is more variance between groups and a smaller amount of variance within groups. There are 696 remaining degrees of freedom. The only variable that is not significant is WithoutHepB. This is not so surprising as WithoutHepB had the lowest amount of correlation between variables.

## 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.)

## Question 5

What variables predict whether or not a district’s reporting was complete?

#install.packages("caret")  
library(caret)

caschools<-districts  
str(caschools)

## 'data.frame': 700 obs. of 13 variables:  
## $ DistrictName : chr "Happy Valley Elementary" "Raymond-Knowles Union Elementary" "South Pasadena Unified" "Saugus Union" ...  
## $ WithoutDTP : num 37 10 12 10 9 1 7 6 5 5 ...  
## $ WithoutPolio : num 42 10 10 10 10 1 5 5 5 4 ...  
## $ WithoutMMR : num 47 10 12 9 10 1 7 4 4 5 ...  
## $ WithoutHepB : num 37 10 10 6 11 1 3 4 4 4 ...  
## $ PctUpToDate : num 47 90 85 89 85 99 90 92 95 91 ...  
## $ DistrictComplete: logi TRUE FALSE TRUE FALSE TRUE TRUE ...  
## $ PctBeliefExempt : num 32 0 8 4 6 0 2 2 3 0 ...  
## $ PctChildPoverty : num 8 30 9 7 13 37 9 29 8 24 ...  
## $ PctFreeMeal : num 8 63 9 17 2 80 36 58 13 69 ...  
## $ PctFamilyPoverty: num 1 6 5 3 4 25 4 15 4 13 ...  
## $ Enrolled : num 19 10 365 1120 175 ...  
## $ TotalSchools : num 1 2 3 15 1 4 3 20 4 17 ...

View(caschools)  
summary(caschools)

## 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 :10.34 Mean : 9.923 Mean :10.32   
## 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. : 23.0 Mode :logical Min. : 0.000   
## 1st Qu.: 2.000 1st Qu.: 84.0 FALSE:43 1st Qu.: 1.000   
## Median : 4.000 Median : 92.0 TRUE :657 Median : 2.000   
## Mean : 7.863 Mean : 87.7 Mean : 5.746   
## 3rd Qu.:10.000 3rd Qu.: 96.0 3rd Qu.: 7.000   
## Max. :77.000 Max. :100.0 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.: 29.00 1st Qu.: 5.00 1st Qu.: 51.75   
## Median :21.00 Median : 50.00 Median : 9.00 Median : 196.50   
## Mean :22.28 Mean : 48.29 Mean :11.39 Mean : 629.71   
## 3rd Qu.:30.00 3rd Qu.: 69.00 3rd Qu.:15.00 3rd Qu.: 675.75   
## Max. :72.00 Max. :100.00 Max. :47.00 Max. :54238.00   
## TotalSchools   
## Min. : 1.000   
## 1st Qu.: 1.000   
## Median : 3.000   
## Mean : 7.253   
## 3rd Qu.: 8.000   
## Max. :582.000

caschools$DistrictComplete<-as.numeric(caschools$DistrictComplete)  
caschools$PctChildPoverty<-caschools$PctChildPoverty/100  
caschools$PctFreeMeal<-caschools$PctFreeMeal/100  
caschools$PctFamilyPoverty<-caschools$PctFamilyPoverty/100  
caschools$PctUpToDate<-caschools$PctUpToDate/100  
caschools$PctBeliefExempt<-caschools$PctBeliefExempt/100  
  
#Linear Regression  
reportinglm<-lm(DistrictComplete ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools, family="binomial", data=caschools)

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'family' will be disregarded

summary(reportinglm)

##   
## Call:  
## lm(formula = DistrictComplete ~ PctChildPoverty + PctFreeMeal +   
## PctFamilyPoverty + Enrolled + TotalSchools, data = caschools,   
## family = "binomial")  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.99235 0.02140 0.04575 0.07563 0.65911   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.0025451 0.0202358 49.543 < 2e-16 \*\*\*  
## PctChildPoverty 0.1698168 0.1473804 1.152 0.250   
## PctFreeMeal -0.0844435 0.0541422 -1.560 0.119   
## PctFamilyPoverty -0.3333398 0.2115495 -1.576 0.116   
## Enrolled 0.0002181 0.0000360 6.059 2.25e-09 \*\*\*  
## TotalSchools -0.0221164 0.0033291 -6.643 6.20e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2283 on 694 degrees of freedom  
## Multiple R-squared: 0.1037, Adjusted R-squared: 0.09719   
## F-statistic: 16.05 on 5 and 694 DF, p-value: 5.471e-15

#Logistic Regression - All variables  
set.seed(1299)  
sampled <- sample(c(TRUE, FALSE), nrow(caschools), replace = TRUE, prob=c(0.7, 0.3))  
train <- caschools[sampled, ]  
test <- caschools[!sampled, ]  
  
reporting<-glm(DistrictComplete ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools, family="binomial", data=train)  
summary(reporting)

##   
## Call:  
## glm(formula = DistrictComplete ~ PctChildPoverty + PctFreeMeal +   
## PctFamilyPoverty + Enrolled + TotalSchools, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0251 0.1939 0.2521 0.3145 1.8905   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.255331 0.671596 6.336 2.36e-10 \*\*\*  
## PctChildPoverty 4.242982 4.056158 1.046 0.29553   
## PctFreeMeal -2.765220 1.473529 -1.877 0.06057 .   
## PctFamilyPoverty -3.821294 5.528759 -0.691 0.48946   
## Enrolled 0.002316 0.001030 2.249 0.02452 \*   
## TotalSchools -0.211456 0.081295 -2.601 0.00929 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 179.27 on 486 degrees of freedom  
## Residual deviance: 157.95 on 481 degrees of freedom  
## AIC: 169.95  
##   
## Number of Fisher Scoring iterations: 6

#plot(reporting)

pred<-round(predict(reporting, test, type="response"))  
str(pred)

## Named num [1:213] 1 1 0 1 1 1 1 1 1 1 ...  
## - attr(\*, "names")= chr [1:213] "418" "216" "630" "45" ...

real<- as.numeric(test$DistrictComplete)  
  
confusionMatrix(table(real, pred))

## Confusion Matrix and Statistics  
##   
## pred  
## real 0 1  
## 0 1 20  
## 1 0 192  
##   
## Accuracy : 0.9061   
## 95% CI : (0.8587, 0.9417)  
## No Information Rate : 0.9953   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.0827   
##   
## Mcnemar's Test P-Value : 2.152e-05   
##   
## Sensitivity : 1.000000   
## Specificity : 0.905660   
## Pos Pred Value : 0.047619   
## Neg Pred Value : 1.000000   
## Prevalence : 0.004695   
## Detection Rate : 0.004695   
## Detection Prevalence : 0.098592   
## Balanced Accuracy : 0.952830   
##   
## 'Positive' Class : 0   
##

#Logistic Regression Best Variables  
reporting1<-glm(DistrictComplete ~ Enrolled + TotalSchools, family="binomial", data=train)  
summary(reporting1)

##   
## Call:  
## glm(formula = DistrictComplete ~ Enrolled + TotalSchools, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7586 0.2569 0.2832 0.2930 1.9485   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.312536 0.254250 13.029 < 2e-16 \*\*\*  
## Enrolled 0.002323 0.001045 2.223 0.02622 \*   
## TotalSchools -0.213993 0.082665 -2.589 0.00963 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 179.27 on 486 degrees of freedom  
## Residual deviance: 164.04 on 484 degrees of freedom  
## AIC: 170.04  
##   
## Number of Fisher Scoring iterations: 6

#plot(reporting1)

pred<-round(predict(reporting1, test, type="response"))  
str(pred)

## Named num [1:213] 1 1 0 1 1 1 1 1 1 1 ...  
## - attr(\*, "names")= chr [1:213] "418" "216" "630" "45" ...

real<- as.numeric(test$DistrictComplete)  
  
confusionMatrix(table(real, pred))

## Confusion Matrix and Statistics  
##   
## pred  
## real 0 1  
## 0 1 20  
## 1 0 192  
##   
## Accuracy : 0.9061   
## 95% CI : (0.8587, 0.9417)  
## No Information Rate : 0.9953   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.0827   
##   
## Mcnemar's Test P-Value : 2.152e-05   
##   
## Sensitivity : 1.000000   
## Specificity : 0.905660   
## Pos Pred Value : 0.047619   
## Neg Pred Value : 1.000000   
## Prevalence : 0.004695   
## Detection Rate : 0.004695   
## Detection Prevalence : 0.098592   
## Balanced Accuracy : 0.952830   
##   
## 'Positive' Class : 0   
##

Before running any models the Percentage Variables that were integers were changed into percentages (PctChildPoverty, PctFreeMeal, PctFamilyPoverty). This will make interpreting the results easier.

The first model created was a linear model, there were only two variables which were significant and that was Enrolled and TotalSchools. The R-squared was 0.1037 with an adjusted R-squared of 0.09719. The F-statistic was 16.05 on 5 with DF of 694. The overall p-value was less than our alpha of 0.05.

Then the model was ran using logistic regression. This resulted in two variables being significant (Enrolled, and TotalSchools) along with the Intercept. The null hypothesis is that there is no relationship between DistrictsComplete and the variables. The z-test value of 6.336 and the associated p-value of 2.36e-10 means we reject the null hypothesis as the p-value is lower than the alpha. The null hypothesis for PctFreeMeal is that there is no relationship between the intercept and PctFreeMeal. Because the p-value is 0.06057, which is greater than the alpha we fail to reject the hypothesis. The null hypothesis for PctChildPoverty is that there is no relationship between PctChildPoverty and the intercept. Because the p-value is 0.29553, which is greater than our alpha we fail to reject the hypothesis. The null hypothesis for PctFamilyPoverty is that there is no relationship between it and the intercept. Because the p-value is 0.48946, which is greater than the alpha we fail to reject the hypothesis. For our variable Enrolled, we reject the null hypothesis that there is no relationship between Enrolled and the intercept because the p-value is 0.02452. Lastly, our variable of TotalSchools has a p-value of 0.00929 which means we reject the null hypothesis that there is no relationship between it and the intercept.

Because the DistrictComplete variable is binomial, the logistic regression provides a better model. It was set to binomial and split the data into training and test data. After using the training data to create the model, the prediction was created using the test data. Once the confusion matrix was built the accuracy was calculated to be 90.61% which is better than the linear model which had an R-squared of 0.1037 and an Adjusted R-squared of 0.09719.

Based on the results of significant and insignificant variables - the best predictors were Enrolled and TotalSchools.

## Question 6

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

#Linear Regression  
uptodate1<-lm(PctUpToDate ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools, data=caschools)  
summary(uptodate1)

##   
## Call:  
## lm(formula = PctUpToDate ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +   
## Enrolled + TotalSchools, data = caschools)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.67616 -0.03122 0.03209 0.07191 0.18618   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.218e-01 1.084e-02 75.791 < 2e-16 \*\*\*  
## PctChildPoverty -1.004e-01 7.897e-02 -1.271 0.204192   
## PctFreeMeal 9.245e-02 2.901e-02 3.187 0.001502 \*\*   
## PctFamilyPoverty 2.983e-01 1.134e-01 2.632 0.008679 \*\*   
## Enrolled 6.466e-05 1.929e-05 3.352 0.000847 \*\*\*  
## TotalSchools -5.758e-03 1.784e-03 -3.228 0.001304 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1223 on 694 degrees of freedom  
## Multiple R-squared: 0.09142, Adjusted R-squared: 0.08487   
## F-statistic: 13.97 on 5 and 694 DF, p-value: 5.028e-13

#Bayes Factor  
bayesuptodate <- lmBF(PctUpToDate ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools, data=caschools,posterior=F)  
summary(bayesuptodate)

## Bayes factor analysis  
## --------------  
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools : 5276913702 ±0.01%  
##   
## Against denominator:  
## Intercept only   
## ---  
## Bayes factor type: BFlinearModel, JZS

The F-statistic is 13.97 with a p-value of 5.028e-13 which is below the alpha value of 0.05. This shows there is a relationship between PctUpToDate and the variables. The intercept has a t-value of 75.791 and a near 0 p-value. Use of the intercept in the equation is supported. The PctChildPoverty coefficient is -1.004e-01, has a t-value of -1.271 and a p-value of greater than the alpha of 0.05. Use of this coefficient in the equation is not supported. The PctFreeMeal coefficient has t-value of 3.187 and a p-value of 0.001502. This is below the 0.05 alpha. Use of the coefficient in the equation is supported. The PctFamilyPoverty coefficient has t-value of 2.632 and a p-value of 0.008679. This is below the 0.05 alpha and use of the coefficient in the equation is supported. The Enrolled coefficient has t-value of 3.352 and a p-value of 0.000847. This is below the 0.05 alpha and use of the coefficient in the equation is supported. The TotalSchools coefficient has t-value of -3.228 and a p-value of 0.001304. This is below the 0.05 alpha and use of the coefficient in the equation is supported. Overall, 4 out of the 5 variables are significant and we would reject the null hypothesis that there is no relationship between PctUpToDate and the variables. For the one insignificant variable (PctChildPoverty), we would fail to reject the null hypothesis.

Because the linear model has 1 variable that is not statistically significant, the model should be re-ran without it (PctChildPoverty) and see if it performs better.

After running the linear regression, a Bayes Factor was ran. The result was 5276913702 ±0.01%. This provides odds in favor of the alternative hypothesis. We reject the null hypothesis that there is no relationship between PctBeliefExempt and the variables. This supports the alternative hypothesis that there is a relationship between PctBeliefExempt and the variables. This finding agrees with the frequentist finding which showed high correlations between these same variables and a p-value of 9.109e-12 which is below the alpha.

## Question 7

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

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

##   
## Call:  
## lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFreeMeal +   
## PctFamilyPoverty + Enrolled + TotalSchools, data = caschools)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.12785 -0.04100 -0.02002 0.00656 0.65439   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.011e-01 7.615e-03 13.281 < 2e-16 \*\*\*  
## PctChildPoverty 1.884e-01 5.546e-02 3.397 0.000721 \*\*\*  
## PctFreeMeal -1.197e-01 2.037e-02 -5.873 6.62e-09 \*\*\*  
## PctFamilyPoverty -2.393e-01 7.961e-02 -3.005 0.002747 \*\*   
## Enrolled -2.736e-05 1.355e-05 -2.020 0.043811 \*   
## TotalSchools 2.293e-03 1.253e-03 1.830 0.067667 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08591 on 694 degrees of freedom  
## Multiple R-squared: 0.1163, Adjusted R-squared: 0.11   
## F-statistic: 18.27 on 5 and 694 DF, p-value: < 2.2e-16

#Bayes Factor  
bayesbelief <- lmBF(PctBeliefExempt ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools, data=caschools,posterior=F)  
summary(bayesbelief)

## Bayes factor analysis  
## --------------  
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools : 6.025845e+13 ±0.01%  
##   
## Against denominator:  
## Intercept only   
## ---  
## Bayes factor type: BFlinearModel, JZS

The F-statistic is 18.27 with a p-value of < 2.2e-16 which is below our alpha value of 0.05. This shows there is a relationship between PctBeliefExempt and the variables. The intercept has a t-value of 13.281 and a near 0 p-value. Use of the intercept in the equation is supported. The PctChildPoverty coefficient has a t-value of 3.397 and a p-value of 0.000721 which is less than the alpha of 0.05. Use of this coefficient in the equation is supported. The PctFreeMeal coefficient has t-value of -5.873 and a p-value of 6.62e-09. This is below the 0.05 alpha. Use of the coefficient in the equation is supported. The PctFamilyPoverty coefficient has t-value of -3.005 and a p-value of 0.002747. This is below the 0.05 alpha and use of the coefficient in the equation is supported. The Enrolled coefficient has t-value of -2.020 and a p-value of 0.043811. This is below the 0.05 alpha and use of the coefficient in the equation is supported. The TotalSchools coefficient has t-value of 1.830 and a p-value of 0.067667. This is above the 0.05 alpha and use of the coefficient in the equation is not supported. Overall, 4 out of the 5 variables are significant and thus we would reject the null hypothesis that there is no relationship between PctBeliefExempt and the variables. For the one insignificant variable (TotalSchools), we would fail to reject the null hypothesis.

After running the linear regression, a Bayes Factor was ran. The result was 6.025845e+13 ±0.01%. This provides odds in favor of the alternative hypothesis. We reject the null hypothesis that there is no relationship between PctBeliefExempt and the variables. This supports the alternative hypothesis that there is a relationship between PctBeliefExempt and the variables. This finding agrees with the frequentist finding which showed high correlations between these same variables and a p-value of 9.109e-12 which is below the alpha.

## Question 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?

First off, vaccine rates in general have trended higher since 1980 year over year with some variability. Below you can see that most of these have remained steady year over year as well except for Polio and MCV1.

Diagram, engineering drawing

Description automatically generated

Overall, both public and private schools have a completion rate of 94.6% providing vaccine reports. Private schools have the lowest overall completion rate of 84.7%.

Chart

Description automatically generated

The 2013 rates for individual vaccines in public schools are 98.4% for DTP, Polio, and MMR while HepB is 98.8%. These rates are higher than the average US vaccination rates for all 4 vaccines. The largest difference coming from HepB which California has a rate of 98.8% while the US is only at 34.21%. There are 5.7% of students in California who have not been vaccinated due to a belief exemption.

A picture containing text

Description automatically generated

One reason why California is higher than the rest of the US appears to be because they require all children who attend private or public schools to be vaccinated. There are several states who offer medical, religious and philosophical exemptions which can cause the overall US rate to decline.

All three of the models that were run, show that there is a relationship between the intercept and some of the variables for each one. Beginning with predicting DistrictComplete we know that the most significant relationship is between the intercept and variables Enrolled and TotalSchools. For predicting PctUpToDate, both the frequentist and Bayes Factor show that we reject the null hypothesis that there is no relationship between PctUpToDate and our significant variables (PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools). For predicting PctBeliefExempt, both the frequentist and Bayes Factor show that we reject the null hypothesis that there is no relationship between PctUpToDate and our significant variables (PctChildPoverty, PctFreeMeal, PctFamilyPoverty, and Enrolled).

Based on this information it is clear that the priority should be to improve the number of schools who meet vaccination completions. Until this rate can be improved, there will be incomplete vaccination information which means that children attending schools are vulnerable to be infected. The analysis shows private schools have the lowest overall completion rate at 84.7%. Public schools have a higher rate at 98.4%. In looking at the districts, the analysis shows that higher number of enrolled students and total schools in the district have problems completing reporting mandates.

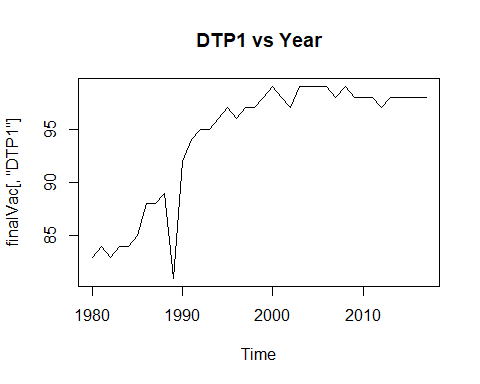
It is recommended that the state allocate resources to collect more data and more school information (recent data would most likely reflect an increase in the number of schools). This will help to understand what is causing the gap between public and private school reporting and help increase the overall total school reporting of vaccinations. At this point the sample is small and we need more information about the underlying factors of both types of schools as to why they are or are not completing reporting mandates. One other variable that is not accounted for and should be is children who are home-schooled. While they do not need to worry about infecting other students, they do need to worry about infecting other family members, some of whom could be students who are not home-schooled. This can lead to outbreaks of infections happening at schools.

The second recommendation is to shift resources to the districts with the highest enrollment counts. Districts with the most schools and most students should be used specifically for reporting. Any other resources such as free meals that take away from efforts to complete reporting should be re-evaluated.

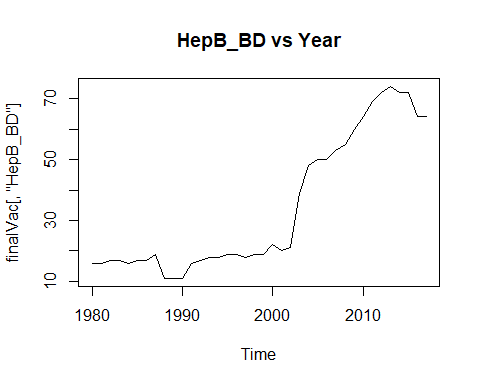
The third recommendation is to increase vaccination rates. Schools with higher belief exemption rates lead to having less students vaccinated. Although this exemption does help with discrimination of students it does increase the risk of other students contracting the disease (even if vaccinated). There is a clear correlation that if you do not have a DTP vaccine you are unlikely to have the others. Schools should re-evaluate the exemption policies and the state could allocate money towards changing these policies. If they can create stricter exemptions or have a no exemption policy, the rates will increase.

## Appendix

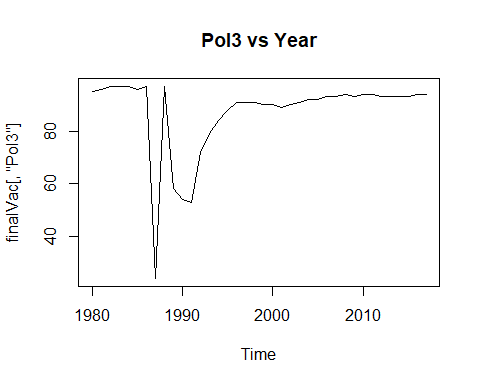
plot.ts(finalVac[,'DTP1'], main = "DTP1 vs Year")



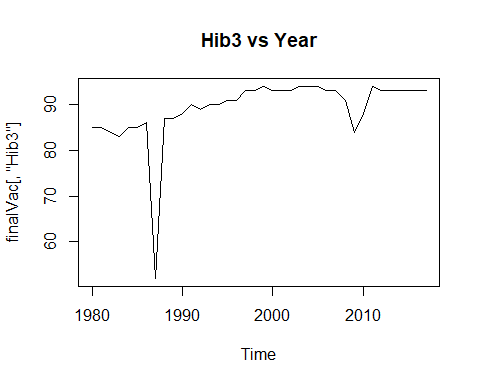
plot.ts(finalVac[,'HepB\_BD'], main = "HepB\_BD vs Year")



plot.ts(finalVac[,'Pol3'], main = "Pol3 vs Year")



plot.ts(finalVac[,'Hib3'], main = "Hib3 vs Year")



plot.ts(finalVac[,'MCV1'], main = "MCV1 vs Year")

