IST-772 - Final Exam

Jean Paul Uwimana

9/11/2020

Final Exam by Jean Paul Uwimana: I produced the material below with no assistance

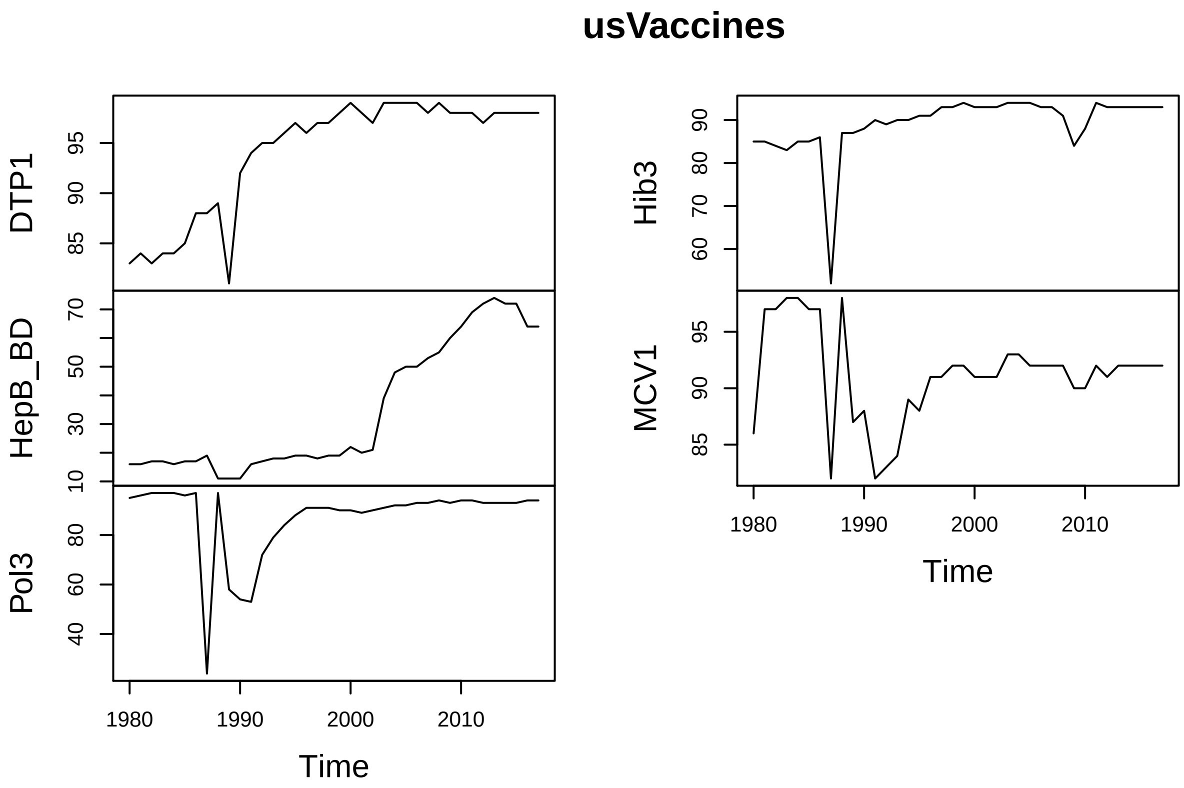
Loading the datasets

```
load("~/Syracuse_University/Summer2020/IST-772/11_Week/Final_Exam_Docs/districts21.RData")
load("~/Syracuse_University/Summer2020/IST-772/11_Week/Final_Exam_Docs/allSchoolsReportStatus 1.RData")
load("~/Syracuse_University/Summer2020/IST-772/11_Week/Final_Exam_Docs/usVaccines 2.RData")
```

Descriptive Reports

- Plotting US vaccines time series

```
plot(usVaccines)
```



- Defining a function to compute variability (SD) in each time series

```
computeVar <- function(vaccine)
{
  # differencing time series
  diffSeries <- diff(usVaccines)
  # computing the SD for a given vaccine name
  SD <- sd(diffSeries[, deparse(substitute(vaccine))]) #No quotes needed on f(x) call
  return(SD)
}

# Computing volatility for each vaccine
computeVar(DTP1)
```

```
## [1] 2.443352
```

```
computeVar(HepB_BD)
```

```
## [1] 4.155751
```

```
computeVar(Pol3)
```

```
## [1] 18.7609
```

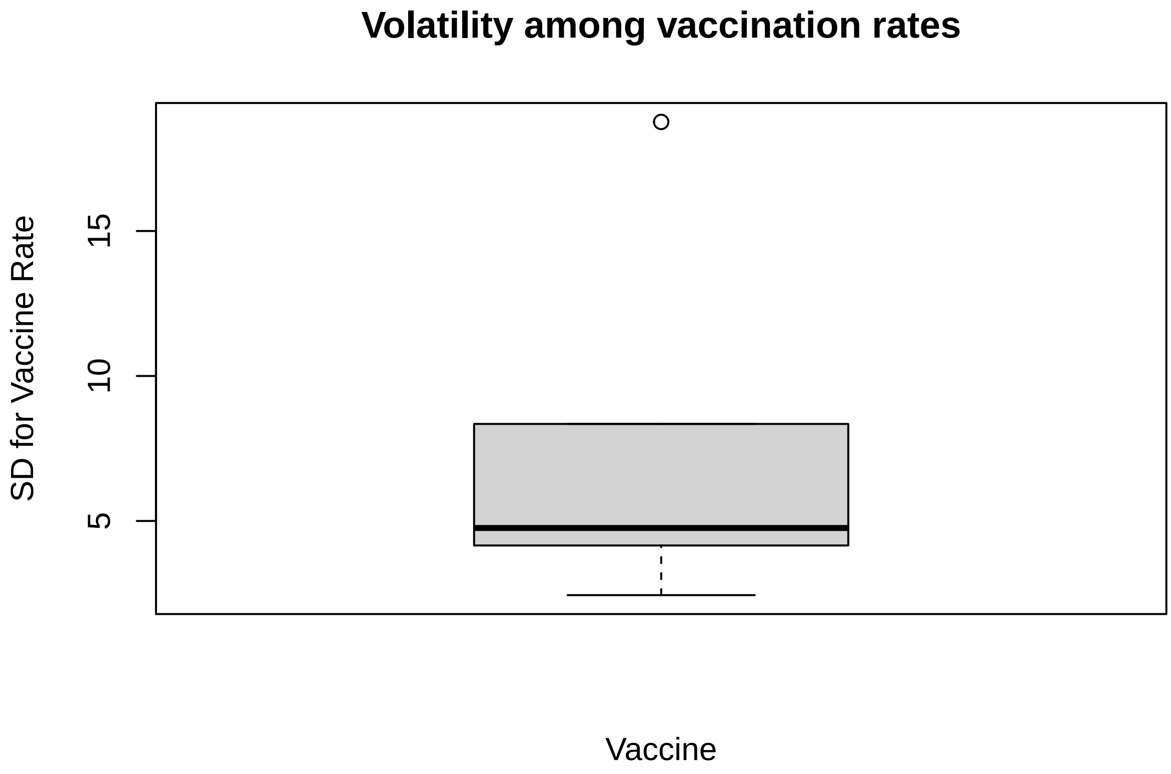
```
computeVar(Hib3)
```

```
## [1] 8.347106
```

```
computeVar(MCV1)
```

```
## [1] 4.758113
```

```
# Plotting volatility level among vaccines
boxplot(c(computeVar(DTP1), computeVar(HepB_BD), computeVar(Pol3),
          computeVar(Hib3), computeVar(MCV1)), xlab = 'Vaccine',
        ylab = 'SD for Vaccine Rate',
        main = 'Volatility among vaccination rates')
```



1. Change of US vaccination rate

- The vaccination rate seems to have increased over time, except for Pol3 which declined by 1% (1980 vs 2017)

```
# vaccination rates in 1980
head(usVaccines, 1)
```

```
##      DTP1 HepB_BD Pol3 Hib3 MCV1
## [1,]   83      16   95   85   86
```

```
# vaccination rates in 2017
tail(usVaccines, 1)
```

```
##       DTP1 HepB_BD Pol3 Hib3 MCV1
## [38,]   98      64   94   93   92
```

```
# average vaccination rates between 1980 - 2017
round(apply(usVaccines, 2, mean), 2)
```

```
##    DTP1 HepB_BD    Pol3    Hib3    MCV1
##   94.05   34.21   87.16   89.21   91.24
```

- The vaccination with the highest rate at the conclusion of the time series is `DTP1` which stands at `98%`
- The vaccination with the lowest rate at the conclusion of the time series is `HepB_BD`, with a rate of `64%`.
- The vaccine with the highest volatility is `Pol3`: 19 as seen in the boxplot above. It is way up in the outlier section of the plot compared to other vaccine standard deviations in the plot.

2. The proportion of public schools reported vaccination data

```r
public <- dplyr::filter(allSchoolsReportStatus, pubpriv == 'PUBLIC')
pubReported <- dplyr::filter(allSchoolsReportStatus,  reported == 'Y' & pubpriv == 'PUBLIC')
round(nrow(pubReported) / nrow(public), 2)
```

```
## [1] 0.97
```

- The proportion of private schools reported vaccination data

```r
private <- dplyr::filter(allSchoolsReportStatus, pubpriv == 'PRIVATE')
privReported <- dplyr::filter(allSchoolsReportStatus,  reported == 'Y' & pubpriv == 'PRIVATE')
round(nrow(privReported) / nrow(private), 2)
```

```
## [1] 0.85
```

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

```r
propDiff <- prop.test(x = c(nrow(pubReported),  nrow(privReported)),
          n = c(nrow(public),  nrow(private)))
# Printing proportion difference test results
propDiff
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(nrow(pubReported), nrow(privReported)) out of c(nrow(public), nrow(private))
## X-squared = 400.49, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1087641 0.1452357
## sample estimates:
##    prop 1    prop 2
## 0.9741800 0.8471801
```

There appears to be a significant difference in proportions between public and private schools based on the results produced by `R` `prop.test()`. The p-value is statistically significant ($p < 0.001$). And the 95% confidence interval show that if we run the process of sampling the proportions multiple times, over the long run, ninety five percent of the time, the difference of proportions between public and private schools that reported vaccination data would be between 0.11 and 0.15. Thus, we reject the null hypothesis that the proportions of public schools that reported vaccine data is equal to that of private schools.

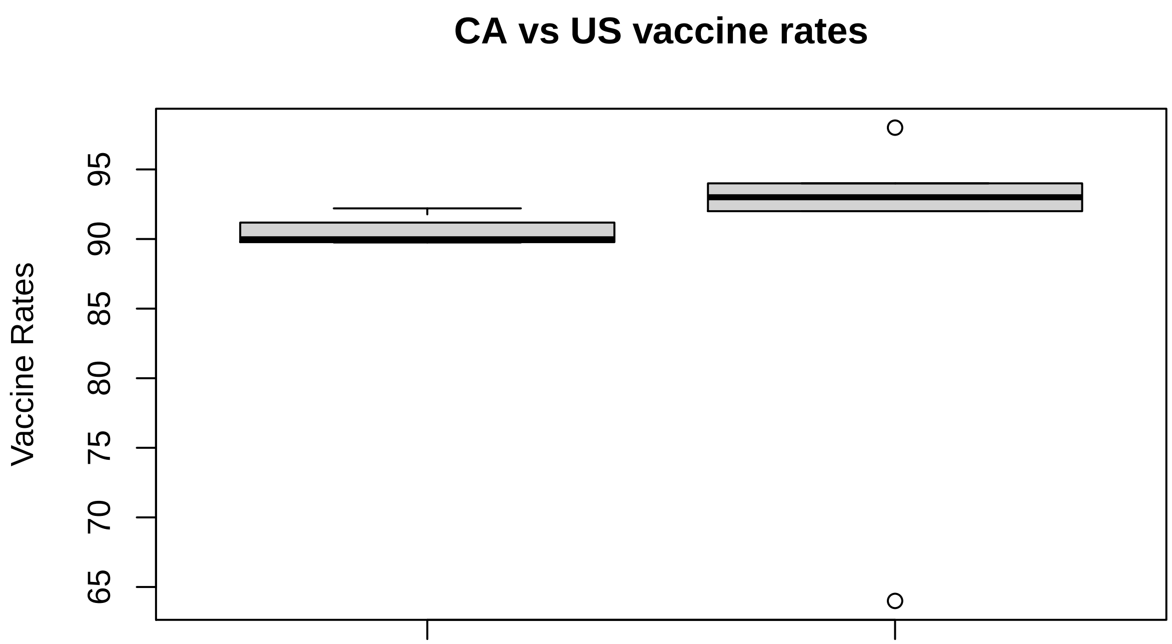
3. 2013 vaccination rates for individual vaccines

```r
# Function to compute vaccination rate in CA for each vaccine
computeRate <- function(vaccineName)
{
  withoutVaccine <- round(mean(districts[[deparse(substitute(vaccineName))]]), 2)
  vaccineRate <- 100 - withoutVaccine # vaccine rate = 100 - those with no vaccine
  return(vaccineRate)
}
```

```r
DTPRate <- computeRate(WithoutDTP)
PolioRate <- computeRate(WithoutPolio)
MMRRate <- computeRate(WithoutMMR)
HepBRate <- computeRate(WithoutHepB)
CARate <- c('DTP1' = DTPRate, 'Pol3' = PolioRate, 'MMR' = MMRRate, 'HepB' = HepBRate)
```

Comparison between CA vaccine rates and US final observations in the time series

```
boxplot(CARate, tail(usVaccines, 1), ylab = 'Vaccine Rates',
        main = 'CA vs US vaccine rates')
```



Comparing standard deviations

```
# sd for CA vaccination rates
sd(CARate)
```

```
## [1] 1.162766
```

```
# sd for US vaccination rates (last observations in time series)
sd(tail(usVaccines, 1))
```

```
## [1] 13.7186
```

Comparing CA vs US vaccine rates using t-test

```
# comparison of individual vaccines in CA districts vs overall US vaccination rates
t.test(CARate, tail(usVaccines, 1))
```

```
##
##  Welch Two Sample t-test
##
## data:  CARate and tail(usVaccines, 1)
## t = 0.36957, df = 4.0717, p-value = 0.7301
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -14.71449  19.26949
## sample estimates:
```

```
## mean of x mean of y
##   90.4775   88.2000
```

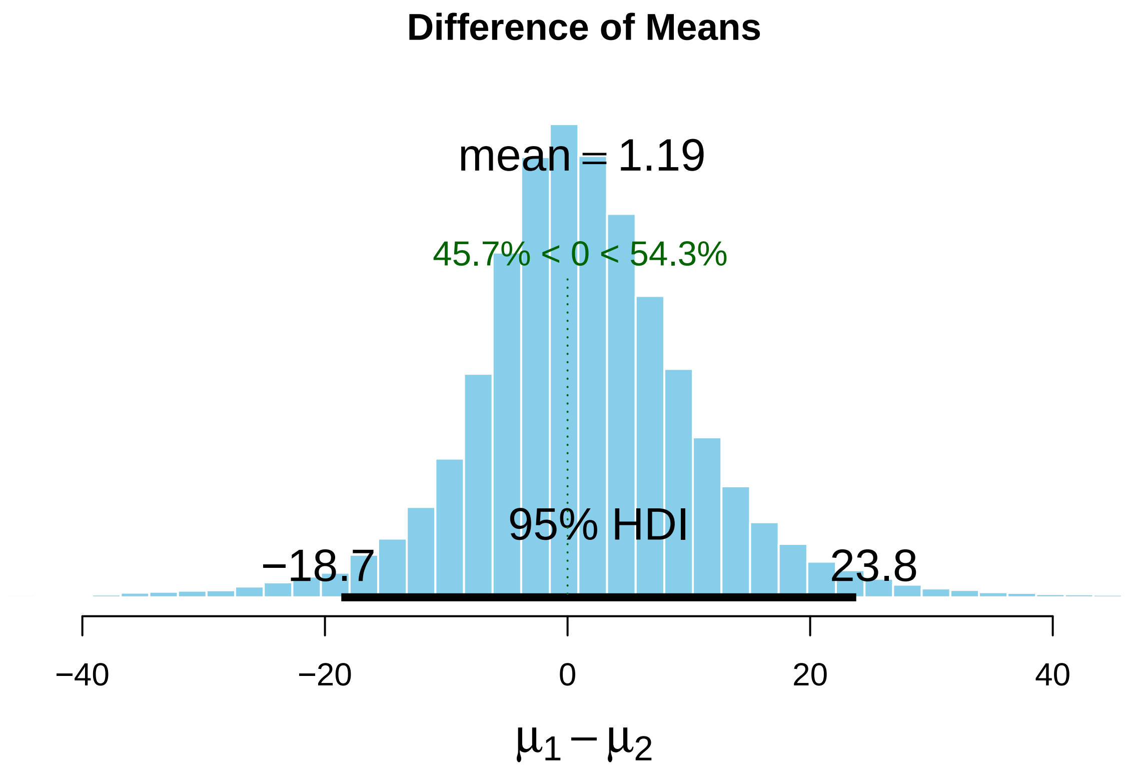
Based on the above boxplot, the median for the US vaccination rates for the last observations of the time series seems is slightly higher than that of CA districts. It stands at 91%, while the median for California school districts is 90%. However, looking at how the data is distributed one might argue that CA districts' vaccination rates are better. The CA rates first quartile is 89.75%, the 3rd quartile is 92.047%. On the other hand, the US median stands at 91%, while the first and third quartile are 66.8 and 97.6, respectively. The US vaccination rates is much more spread out with the standard deviation of 13.72, while the standard deviation for CA districts is just 1.16. An official t-test statistic was conducted and it showed that there does not appear to be a significant difference in the vaccine mean rates between CA and the final observations of the US vaccine rates based on a very high p-value (0.73), a weak t-value (0.37) and the 95% CI (-14.71 to 19.27) which clearly overlap with zero. Thus, I fail to reject the null hypothesis that the CA school districts vaccine rate is the same as that of the final observations of the US.

Comparing and Visualizing the t-test using Bayesian technique

```
t_test <- BEST::BESTmcmc(CARate, tail(usVaccines, 1))
```

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

```
plot(t_test)
```



Based on the above graph, there's a 95% probability that there's no evidence of difference in means between CA vaccination rate and that of the final observations of the US. The test also shows that 46% of the time the difference in means is below zero, and 54% of the time it is above zero. Additionallly, the 95% HDI overlap with zero (-18.9 to 23.1) which further confirms the absence of any avidence to suggest that there's a difference in means between CA vaccination rate and that of the final observations of the US.

4. Among districts, checking whether there's a correlation between a student missing one vaccine vs missing all others

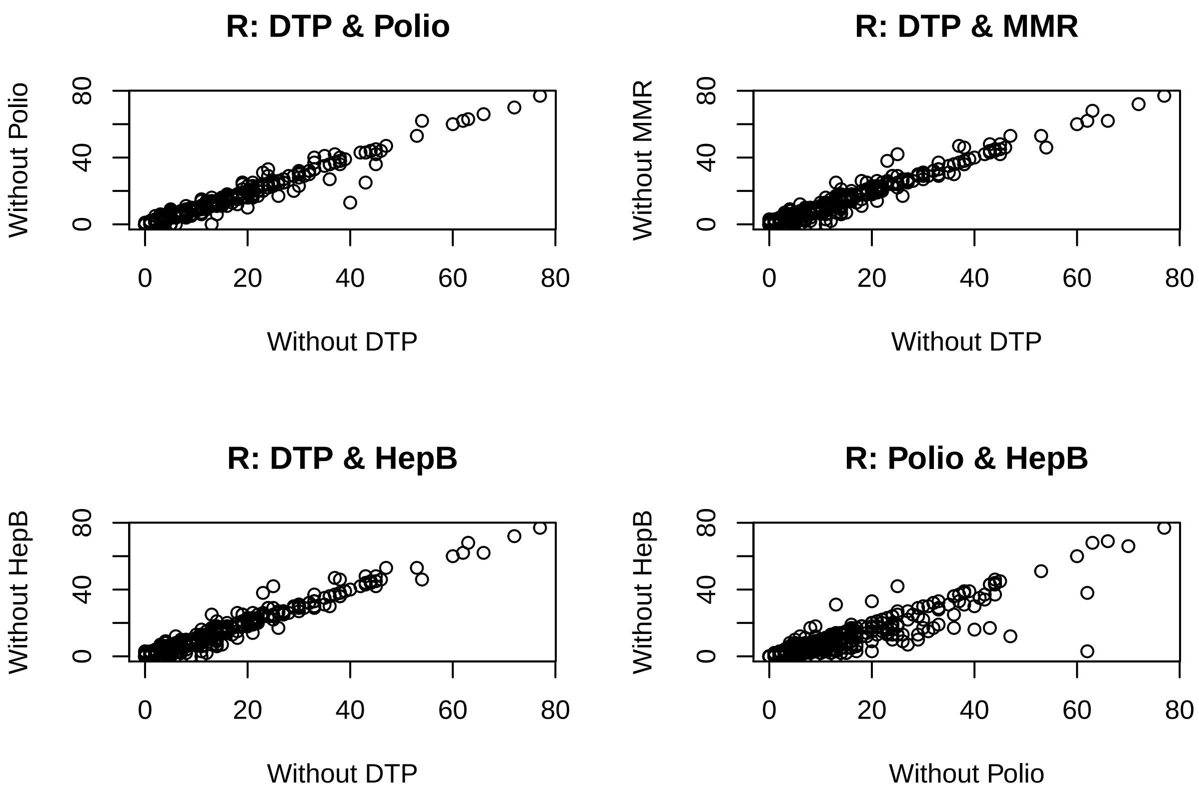
```
knitr::kable(round(cor(districts[, 2:5]), 2))
```

	WithoutDTP	WithoutPolio	WithoutMMR	WithoutHepB
WithoutDTP	1.00	0.98	0.98	0.89
WithoutPolio	0.98	1.00	0.97	0.91
WithoutMMR	0.98	0.97	1.00	0.90
WithoutHepB	0.89	0.91	0.90	1.00

Based on the above correlation results, it's clear that the rate of missing a vaccine is highly correlated among individual vaccines. The lowest correlation coefficient is between `DTP` and `HepB` which is `0.89`, all others are `0.90` or above. Thus, the results are telling us that if a student misses one vaccine, he or she will most likely miss all other vaccines.

Below is the graph of correlations between missing vaccines accross all the school districts. The correlation shows nice cigar-shaped graphs between individual vaccines which is evidence of a strong positive correlation.

```
par(mfrow = c(2, 2))
plot(districts$WithoutDTP, districts$WithoutPolio,
     main = "R: DTP & Polio ",
     xlab = 'Without DTP', ylab = 'Without Polio')
plot(districts$WithoutDTP, districts$WithoutMMR,
     main = "R: DTP & MMR",
     xlab = 'Without DTP', ylab = 'Without MMR')
plot(districts$WithoutDTP, districts$WithoutMMR,
     main = "R: DTP & HepB",
     xlab = 'Without DTP', ylab = 'Without HepB')
plot(districts$WithoutPolio, districts$WithoutHepB,
     main = "R: Polio & HepB",
     xlab = 'Without Polio', ylab = 'Without HepB')
```



Predictive Analysis

5. The variables that predict whether or not a district's reporting was complete

- Data Transformation: Transforming `DistrictComplete` column from `logical` to `factor` and adding a factor column replacing True = 1 and False = 0 for the Bayesian model

```
districts$DistrictComplete <- as.factor(districts$DistrictComplete)
districts$DistrictCompleteTransformed <- ifelse(districts$DistrictComplete == T, 1, 0)
```

- Fitting the Generalized Linear Model

```
modelFit <- glm(DistrictComplete ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
                  Enrolled + TotalSchools, family = binomial(), data = districts)
```

- Bayesian model using the transformed variable (DistrictCompleteTransformed)

```
modelFitBayes <- MCMCpack::MCMClogit(DistrictCompleteTransformed ~ PctChildPoverty +
                                       PctFreeMeal + PctFamilyPoverty + Enrolled +
                                       TotalSchools, data = districts)
```

Printing frequentist model results

```
summary(modelFit)
```

```
##
## Call:
## glm(formula = DistrictComplete ~ PctChildPoverty + PctFreeMeal +
##     PctFamilyPoverty + Enrolled + TotalSchools, family = binomial(),
##     data = districts)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7645   0.2300   0.2716   0.3241   1.8247
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.9636754  0.5004647   7.920 2.38e-15 ***
## PctChildPoverty   0.0247009  0.0317553   0.778  0.43666
## PctFreeMeal      -0.0109869  0.0118925  -0.924  0.35556
## PctFamilyPoverty -0.0597363  0.0398015  -1.501  0.13339
## Enrolled          0.0017366  0.0006798   2.554  0.01064 *
## TotalSchools     -0.1741938  0.0578641  -3.010  0.00261 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 283.78  on 699  degrees of freedom
## Residual deviance: 252.23  on 694  degrees of freedom
## AIC: 264.23
##
## Number of Fisher Scoring iterations: 6
```

```r
# converting coefficient to regular odds for interpretation
round(exp(coef(modelFit)), 2)
```

```
##      (Intercept)  PctChildPoverty      PctFreeMeal PctFamilyPoverty
##            52.65             1.03             0.99             0.94
##         Enrolled     TotalSchools
##             1.00             0.84
```

```r
# getting confidence intervals from the model
round(exp(confint(modelFit)), 2)
```

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

```
##                  2.5 % 97.5 %
## (Intercept)      21.15 152.10
## PctChildPoverty   0.97   1.09
## PctFreeMeal       0.97   1.01
## PctFamilyPoverty  0.87   1.02
## Enrolled          1.00   1.00
## TotalSchools      0.74   0.94
```

```r
# getting R-squared
round(BaylorEdPsych::PseudoR2(modelFit), 2)['Nagelkerke']
```

```
## Nagelkerke
##       0.13
```

- Frequentist results
    - According to the frequentist analysis above, the coefficients (regular odds) for `Enrolled` and `TotalSchools`, `1.00` and `0.84`, respectively, are significantly different from `0` based on the z-test values of `2.554` and `-3.01` and their associated p-values of `0.01064` and `0.00261`. Thus, I reject the null hypothesis that the odds for the total number of enrolled students and total number of different schools in the districts are zero in the population.

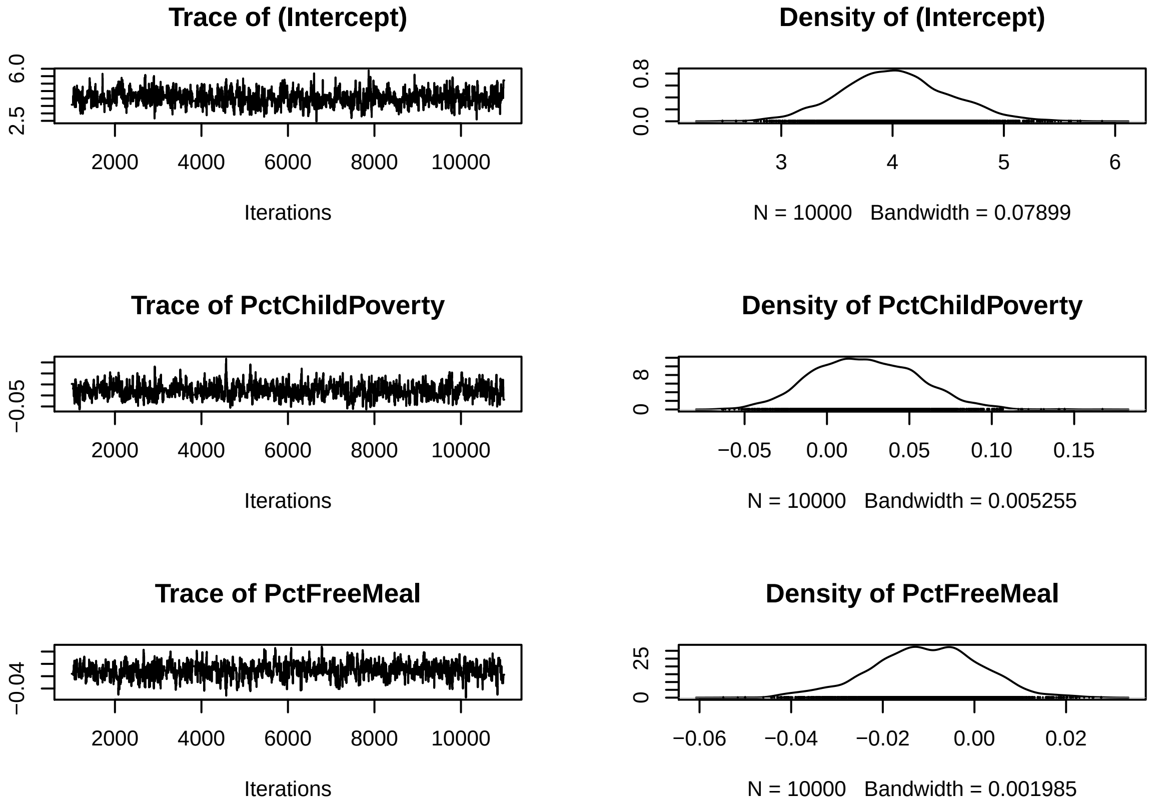
  - The 95% confidence interval (CI) of also confirm the results of the hypothesis test. The `Enrolled` ranges between `1.0003825` and `1.0030840`, while the `TotalSchools` ranges between `0.7442703` and `0.9366998`. On the other hand, the 95% CI for all other variables included in this model i.e.: PctChildPoverty (0.9657056 - 1.0931498) , PctFreeMeal (0.9657810 - 1.0120267) and PctFamilyPoverty (0.8717782 - 1.0192992) straddle `1` as seen in their confidence interval ranges. The fact that they straddle 1, effectively makes them not containing any predictive value to contribute to the model. And therefore, I fail to reject the null on these three variables. This leaves us with just `Enrolled` and `TotalSchools` as the best predictors as to whether a district reporting was complete
  - Although, there are two variables that are deemed predictors of whether a district reporting was complete, the results of R-squared show that the proportion of `DistrictComplete` accounted by predictor variables is only 0.13. This proportion is very small and as such it is a strong indication that this is not a good model that can accurately predict the response variable, `DistrictComplete`.

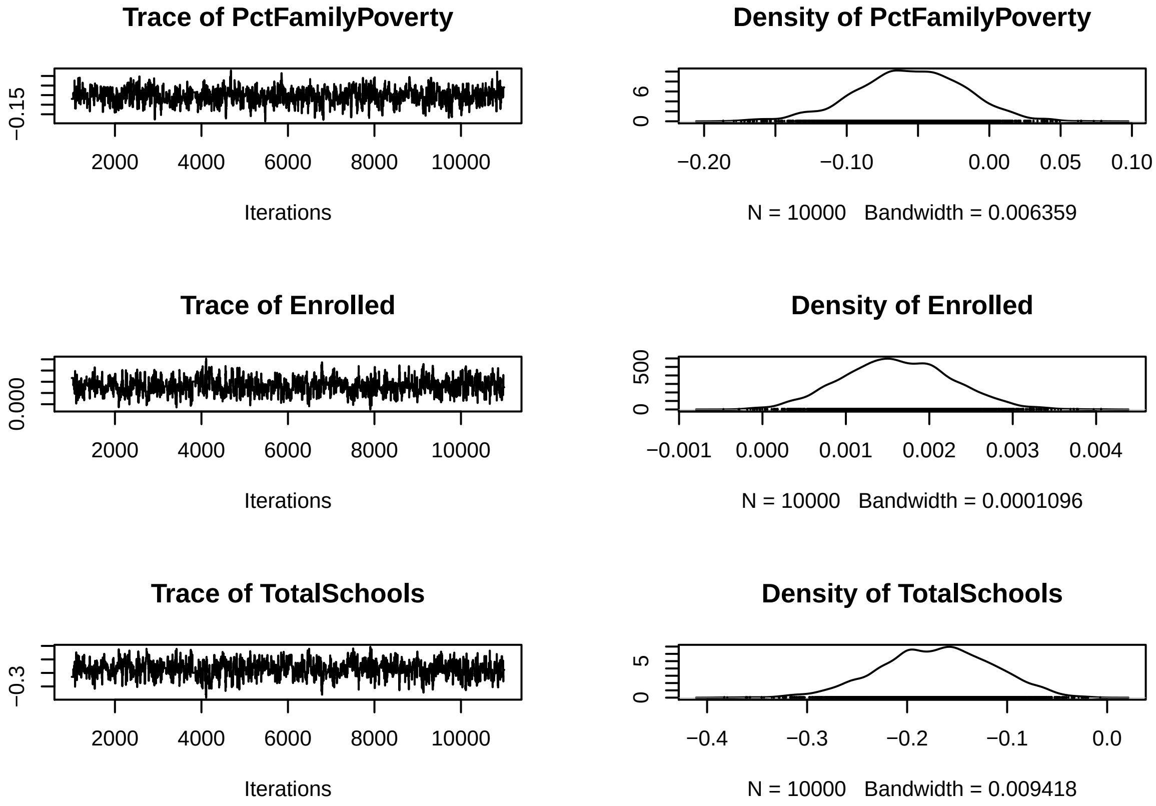
- Printing Bayesian results

```
# model summary
summary(modelFitBayes)
```

```
##
## 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)       4.02667 0.4815493 4.815e-03      2.104e-02
## PctChildPoverty   0.02332 0.0312810 3.128e-04      1.441e-03
## PctFreeMeal      -0.01060 0.0120252 1.203e-04      5.745e-04
## PctFamilyPoverty -0.05536 0.0378511 3.785e-04      1.669e-03
## Enrolled          0.00162 0.0006524 6.524e-06      2.937e-05
## TotalSchools     -0.17066 0.0560571 5.606e-04      2.539e-03
##
## 2. Quantiles for each variable:
##
##                        2.5%        25%       50%       75%     97.5%
## (Intercept)       3.1238440  3.7035000  4.010546  4.333543  5.010265
## PctChildPoverty  -0.0340221  0.0004172  0.022645  0.046099  0.086479
## PctFreeMeal      -0.0356899 -0.0184492 -0.010447 -0.002616  0.011872
## PctFamilyPoverty -0.1336779 -0.0798783 -0.054777 -0.028446  0.015098
## Enrolled          0.0003704  0.0011751  0.001598  0.002055  0.002907
## TotalSchools     -0.2832254 -0.2066605 -0.168284 -0.130939 -0.065668
```

```
# plotting the density plots
plot(modelFitBayes)
```


Trace of (Intercept)
6.0
2.5
2000
4000
6000
8000
10000
Iterations
Density of (Intercept)
0.8
0.0
3
4
5
6
N = 10000   Bandwidth = 0.07899
Trace of PctChildPoverty
−0.05
2000
4000
6000
8000
10000
Iterations
Density of PctChildPoverty
8
0
−0.05
0.00
0.05
0.10
0.15
N = 10000   Bandwidth = 0.005255
Trace of PctFreeMeal
−0.04
2000
4000
6000
8000
10000
Iterations
Density of PctFreeMeal
25
0
−0.06
−0.04
−0.02
0.00
0.02
N = 10000   Bandwidth = 0.001985



+ As seen above, the density plots for `PctChildPoverty`, `PctFreeMeal` and `PctFamilyPoverty` clearly overlap with zero. Further solidifying the evidence from the significance test that their coefficients were not significant.

6. The variables that predict the percentage of enrolled students with completely up-to-date vaccines

- Fitting the linear model

```
modelFitLinear <- lm(PctUpToDate ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
                     Enrolled + TotalSchools, data = districts)
```

- Fitting the Bayesian linear model

```
modelFitLinearBayes <- BayesFactor::lmBF(PctUpToDate ~ PctChildPoverty + PctFreeMeal +
                                             PctFamilyPoverty + Enrolled + TotalSchools,
                                           data = districts, posterior = T, iterations = 10000)
```

- Printing frequentist results for linear model

```
summary(modelFitLinear)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
##     Enrolled + TotalSchools, data = districts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.914  -3.269   3.374   7.121  18.748
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      82.636291   1.086527  76.055  < 2e-16 ***
## PctChildPoverty  -0.152169   0.079865  -1.905 0.057151 .  
## PctFreeMeal       0.102734   0.029459   3.487 0.000519 ***
## PctFamilyPoverty  0.336652   0.112871   2.983 0.002958 ** 
## Enrolled          0.006970   0.002042   3.413 0.000680 ***
## TotalSchools     -0.622906   0.188534  -3.304 0.001002 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.14 on 694 degrees of freedom
## Multiple R-squared:  0.09148,    Adjusted R-squared:  0.08494 
## F-statistic: 13.98 on 5 and 694 DF,  p-value: 4.91e-13
```

- Printing Bayesian results for linear model

```r
summary(modelFitLinearBayes)
```

```
## 
## 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                87.852318 0.457279 4.573e-03      4.573e-03
## PctChildPoverty   -0.147723 0.078501 7.850e-04      7.850e-04
## PctFreeMeal        0.099225 0.029083 2.908e-04      3.010e-04
## PctFamilyPoverty   0.325987 0.110003 1.100e-03      1.100e-03
## Enrolled           0.006714 0.002021 2.021e-05      2.021e-05
## TotalSchools      -0.600125 0.186279 1.863e-03      1.863e-03
## sig2             147.492917 7.908139 7.908e-02      8.114e-02
## g                  0.057406 0.075043 7.504e-04      7.504e-04
## 
## 2. Quantiles for each variable:
## 
##                        2.5%        25%        50%        75%      97.5%
## mu                86.961626  87.545247  87.849975  88.162443  88.752461
## PctChildPoverty   -0.302272  -0.200206  -0.148332  -0.095136   0.006462
## PctFreeMeal        0.042534   0.079795   0.098930   0.118630   0.156583
## PctFamilyPoverty   0.104964   0.253195   0.326295   0.400312   0.537533
## Enrolled           0.002742   0.005362   0.006751   0.008054   0.010657
## TotalSchools      -0.968222  -0.723042  -0.603346  -0.475985  -0.233853
## sig2             132.722327 142.065317 147.210073 152.601586 163.549702
## g                  0.015210   0.028559   0.042279   0.066060   0.190346
```

```r
# Bayes Factor
BayesFactor::lmBF(PctUpToDate ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
                  Enrolled + TotalSchools, data = districts)
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools : 5404801648 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

- Frequentist model interpretation
  - The coefficients for PctFreeMeal, PctFamilyPoverty, Enrolled and TotalSchools, `0.102734`, `0.336652`, `0.006970` and `-0.622906`, respectively, are significantly different from zero based on solid t-values and their associated p-values. Thus, I reject the null hypothesis that the model composed by just the y-intercept is no different than the one with the variables above. However, the coefficient for PctChildPoverty, `-0.152169`, is not significantly different from zero based on its t-value of `-1.905` and its associated p-value (0.057) which is greater than the conventional threshold of `0.05`. Thus, I fail to reject the null hypothesis that `PctChildPoverty` has no effect on `PctUpToDate`
  - The R-squared value is very tiny, just `0.09148` and Adjusted R-squared is `0.08494`. This R-squared value suggests that the proportion of `PctUpToDate` that is accounted by the predictor variables, namely, PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled and TotalSchools is only `9.15%`. Additionally, the F-test on the null hypothesis that R-squared is equal to 0, is a fairly decent number F(5,694) = 13.98, again I can reject the null hypothesis, as the associated p-value is less than `0.05`.
- Bayesian model interpretation
  - The 95% HDI for B-weights of all other variables except `PctChildPoverty`, do NOT overlap with 0, providing evidence that the population value of B-weights for PctFreeMeal, PctFamilyPoverty, Enrolled, TotalSchools differ from 0.
  - Additionaly, the Bayes Factor produced a substantially large number (5404801648) which is a very strong positive evidence that B-weights for the aforementioned variables are nonzero.

7. The variables that predict the percentage of all enrolled students with belief exceptions

- Fitting linear model 2

```r
modelFitLinear2 <- lm(PctBeliefExempt ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
                  Enrolled + TotalSchools, data = districts)
```

- Fitting the Bayesian linear model 2

```r
modelFitLinearBayes2 <- BayesFactor::lmBF(PctBeliefExempt ~ PctChildPoverty + PctFreeMeal +
                                            PctFamilyPoverty + Enrolled + TotalSchools,
                                          data = districts, posterior = T, iterations = 10000)
```

- Printing frequentist results for linear model 2

```r
summary(modelFitLinear2)
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFreeMeal +
##     PctFamilyPoverty + Enrolled + TotalSchools, data = districts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.618  -4.071  -2.104   0.767  65.612
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.722944   0.752897  12.914  < 2e-16 ***
## PctChildPoverty   0.193841   0.055342   3.503  0.00049 ***
## PctFreeMeal      -0.116223   0.020413  -5.694 1.84e-08 ***
## PctFamilyPoverty -0.239158   0.078213  -3.058  0.00232 **
## Enrolled         -0.003183   0.001415  -2.249  0.02481 *
## TotalSchools      0.270093   0.130643   2.067  0.03907 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.415 on 694 degrees of freedom
## Multiple R-squared:  0.1093, Adjusted R-squared:  0.1028
## F-statistic: 17.02 on 5 and 694 DF,  p-value: 6.702e-16
```

- Printing Bayesian results for linear model 2

```r
summary(modelFitLinearBayes2)
```

```
##
## 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                5.704166 0.315611 3.156e-03      3.139e-03
## PctChildPoverty   0.187977 0.054278 5.428e-04      5.204e-04
## PctFreeMeal      -0.112568 0.020321 2.032e-04      2.064e-04
## PctFamilyPoverty -0.231222 0.077094 7.709e-04      7.709e-04
## Enrolled         -0.003076 0.001388 1.388e-05      1.244e-05
## TotalSchools      0.261129 0.128192 1.282e-03      1.138e-03
## sig2             70.818120 3.806588 3.807e-02      3.807e-02
## g                 0.062094 0.073102 7.310e-04      8.054e-04
##
## 2. Quantiles for each variable:
##
##                       2.5%       25%       50%       75%      97.5%
## mu                5.087618  5.495197  5.705587  5.908888  6.3362069
## PctChildPoverty   0.082011  0.151960  0.187893  0.223974  0.2962442
## PctFreeMeal      -0.152413 -0.126430 -0.112475 -0.099015 -0.0724668
## PctFamilyPoverty -0.385121 -0.283498 -0.231229 -0.178437 -0.0822854
## Enrolled         -0.005764 -0.004019 -0.003071 -0.002133 -0.0003891
## TotalSchools      0.008421  0.173908  0.261343  0.348161  0.5093714
## sig2             63.752777 68.190190 70.690387 73.365093 78.5483503
## g                 0.016609  0.031116  0.045897  0.071817  0.1986032
```

```r
# Bayes Factor
BayesFactor::lmBF(PctBeliefExempt ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
                  Enrolled + TotalSchools, data = districts)
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled + TotalSchools : 4.120335e+12 ±0%
##
## Against denominator:
```

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

- Frequentist model interpretation
  - The coefficients for variables PctChildPoverty (0.193841), PctFreeMeal (-0.116223), PctFamilyPoverty (-0.239158), Enrolled (-0.003183), TotalSchools (0.270093) are all significantly different from zero based on their respective t-values and associated p-values (also statiscally significant). Also, the F-test value F(5,694) = 17.02 and its associated p-value (p < 0.001) is large enough to allow me to reject the null that the coefficients for the above variables are zero in the population.
  - The results of the R-squared, although, rather small, indicates that the variation of the outcome variable, PctBeliefExempt, accounted by the predictor variables is just 11% (R-squared = 0.11)
- Bayesian model interpretation
  - The 95% HDI for B-weights of all the variables, do NOT overlap with 0. Which provides further evidence that the B-weights of the predictor variables differ from zero. And finally, the Bayes Factor produced a substantially large number (4120335000000) which is a very strong positive evidence that B-weights for the all the variables are nonzero in a model predicting `PctBeliefExempt` (Percentage of belief exception).

8. The big picture

- In this report, I conducted various statistical analyses to compare how the US vaccination rates for the most popular vaccines (DTP1, HepB_DB, Pol3, Hib3, and MCV1) progressed between 1980 and 2017. I have also looked at how the vaccination rates in the state of California fare in comparison with the United States vaccination rates. Additionally, I analyzed what the vaccination reporting rate is for California public schools versus private schools, and which variables if any, are credible in predicting whether or not a school district's reporting was complete, which variables could predict the percentage of enrolled students who are up-to-date with vaccines, and finally which variables may predict the percentage of enrolled students with belief exception.
- A descriptive analysis of the data, concluded that the US vaccination rates have increased over time as the rates for DTP1, HepB_BD, Pol3, Hib3, MCV1 have increased by 15%, 48%, -1%, 8%, 6%, respectively, between 1980 and 2017. Although, most vaccines showed a strong growth rate, Pol3 declined by one point percentage (1980: 95% vs 2017: 94%). And HepB_BD, despite having a strong growth (1980: 16% vs 2017: 64%), its rate still remains relatively lower at 64% in 2017 vs all others which are 90+%. Moreover, during the last year's observations (2017), the median vaccination rate for the US was 91%, while California stood at 90%. California vaccination rate was less volatile, however. When it came to reporting compliance, California public schools had a higher vaccination rate than private schools, 97% vs 85%, respectively. On average, California districts had very strong vaccination rates: DTP1 (89.75), Pol3 (90.16), MMR (89.80), and HepB (92.20). In regards to the reporting compliance, in California, 94.86% school districts had completed their reporting vs 5.14% who had not, while 87.85 students were up-to-date on their vaccines. Another remarkable aspect of the analysis is that there's a strong correlation between missing vaccines. In other words, if a student misses one vaccine, they are very likely going to miss all the vaccines as demonstrated in the correlation table of the analysis.
- Predictive analyses
  - Results of the frequentist method showed that the number of enrolled students and the total number of schools in a district provide a 13% predictive value as to whether or not a district reporting compliance was complete. The Bayesian analysis of the same model also provided a strong evidence in favor of the significance test that the number of enrolled students and total number of schools in a district provide predictive value in determining the reporting compliance. The following is a summary of how (Enrolled + TotalSchools) variables are likely to predict whether or not a district compliance was complete:
    - ∗ As the number of enrolled students increases by a percentage point, the odds of reporting completion increase by 0.2%
    - ∗ As the number of total schools increases by a percentage point, the odds of reporting completion

decrease by 16%

- Once again, both the frequentist and Bayesian techniques were in agreement that PctFreeMeal, PctFamilyPoverty, Enrolled and TotalSchools are predictors for the percentage of students with up-to-date vaccines. These results were backed by very strong p-values < 0.05 overall and an enormous Bayes factor value of 5404801648, plus the 95% HDI from the Bayesian model do not overlap with zero, providing a strong evidence of the predictive value found in these variables. The following is a summary of how these variables are likely to impact the percentage of being up-to-date on vaccines:
  * Every percent increase in child poverty would likely result in 0.15 decrease in percentage of up-to-date
  * Every perent increase in free meals would result in 0.10 increase in percentage of up-to-date Every percent increase in family poverty would likely result in 0.34 increase in percentage of up-to-date
  * Every percent increase in enrolled students would likely result in 0.006 increase in percentage of up-to-date
  * Every percent increase in total number of schools would likely result in 0.62 decrease in percentage of up-to-date
- Finally, it was demonstrated through the frequentist and Bayesian analyses that the percentage of students with belief exception (PctBeliefExempt) can be predicted by the following variables PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, TotalSchools in the district in the following ways:
  * Every percent increase in child poverty would result in 0.19 increase in percentage of number of enrolled students with belief exceptions
  * Every percent increase in free meals would result in 0.12 decrease in percentage of belief exceptions.
  * Every percent increase in family poverty would result in 0.24 decrease in percentage of belief exceptions
  * Every percent increase in number of enrolled students would result in 0.003 decrease in percentage of belief exceptions
  * Every percent increase in number of total schools would result in 0.27 increase in percentage of belief exceptions.