State of Vaccination Rates in California School Districts

Matt A. Beck

9/10/2020

Introduction

This report includes analysis of three datasets, which includes:

- Time series data from the World Health Organization reporting vaccination rates in the U.S. for five common vaccines
- A list of California kindergartens and whether they reported vaccination data to the state in 2013
- A sample of California public school districts from the 2013 data collection, along with specific numbers and percentages for each district.

Author's Note

Auxillary research of CDC vaccination history (https://www.cdc.gov/mmwr/preview/mmwrhtml/su6004a9.htm) and actuals from the World Health Organization (https://apps.who.int/immunization_monitoring/globalsummary/coverages?c=USA) lends credence to the notion that the datasets used for this analysis should be considered purely as directional - data quality for the 1980's in particular is difficult, given the lack of reliable coverage data from the period, and that more robust survey results on vaccinations did not become available until around 1995. Most glaring in the WHO data are the low coverage rates for Hepatitis B, for which the CDC reports "By 2000, at least 90% of infants were being vaccinated annually."

Executive Summary

Looking across the last several decades, vaccination rates at a national level have largely remained above 90%, with Hepatitis B as the lone exception. Changes in vaccination rates are examined, revealing the Flu vaccine to be the most actively fluctuating. Vaccination reporting rates between public and private schools at the national level are also examined, indicating private schools report vaccination rates at a significantly lower proportion than public schools. California falls below the national average in vaccination rates across all of the vaccines in the standard array, with the exception of Hepatitis B. Vaccination rates from one vaccine type to another are highly correlated, indicating that if a person receives one vaccine, it is almost certain they will receive or have received all of the vaccines present in the dataset.

Varying methods of statistical analysis are used to examine the relationship between reporting completion, reported vaccination completion rates, and belief-based exemption rates across California school districts using an array of available demographic variables. These variables include the percentage of reported children in poverty, percentage of students receiving free meals, percentage of family poverty, and a ratio of enrolled students to schools in a given district. These analyses indicate potential relationships between:

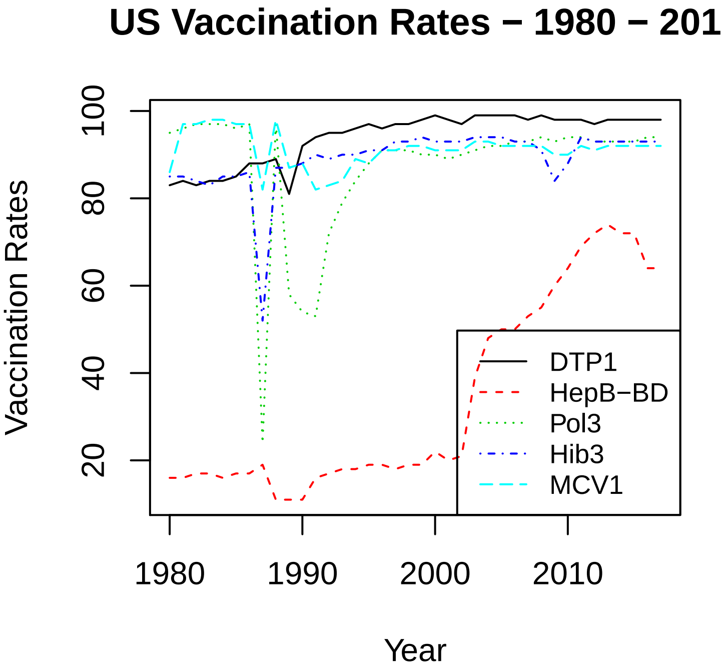
- Vaccination reporting completion and the ratio of students per school in the district
- Reported Vaccination rates and both in-district free meal eligibility and student to school ratio (positive relationship)

- Belief-based Exemptions and children under the poverty level in the district (positive relationship), in-district free meal eligibility and student to school ratio (negative relationship)

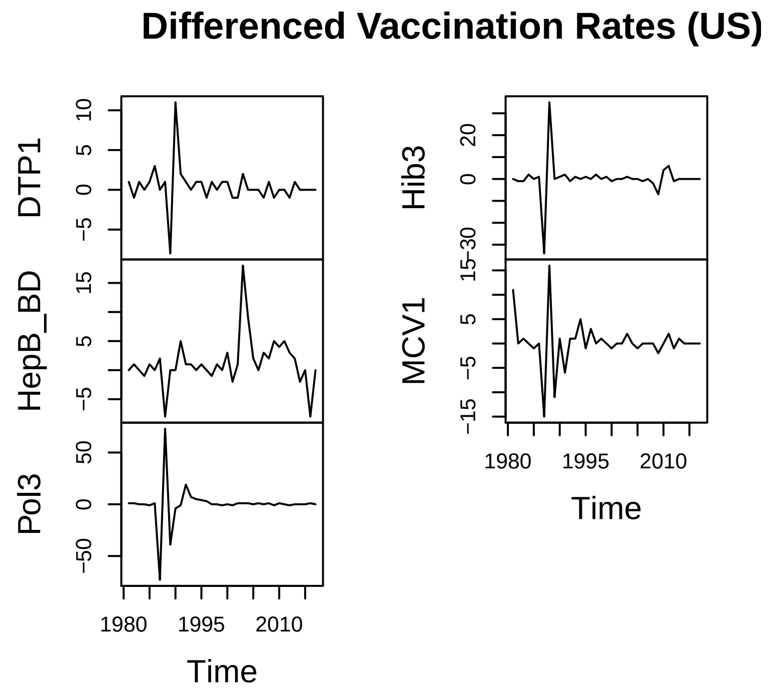
The evidence taken together, it is recommended our state legislator allocate financial assistance to school districts in rural areas with high proportions of individuals (especially children) who are under the poverty level.

US Vaccination Rates over Time

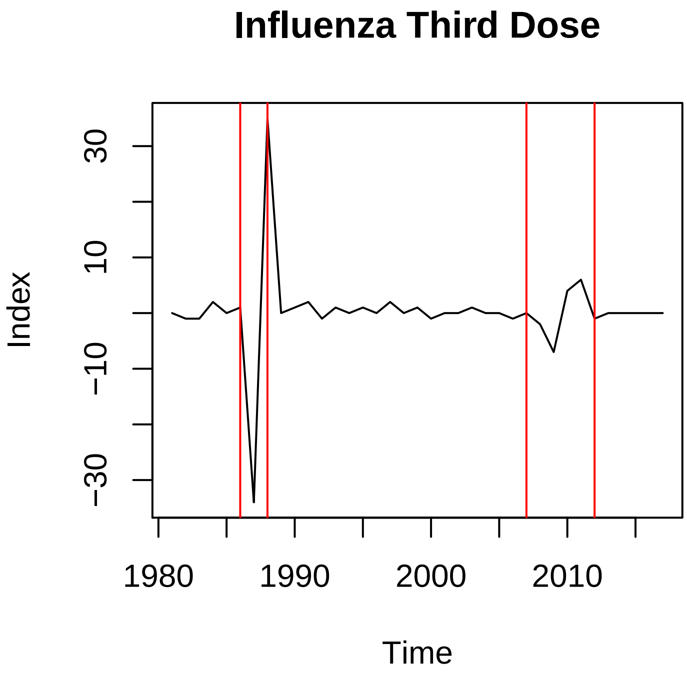
Vaccination efforts in the United States from the 80's through 2017 can generally be characterized as a success, with many of the measured vaccination rates remaining in the low 90 percent range through the majority of the period.

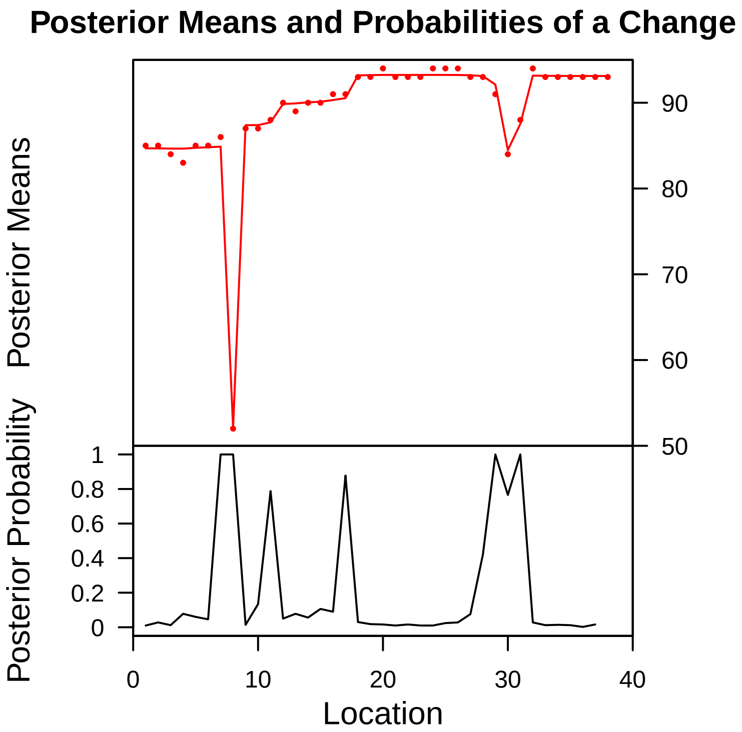


In terms of success, the DTP1 vaccine (shown in the chart above in black), has steadily risen from a low of ~80% in 1989 and remained above 90 percent through at least 2017, placing the highest among the vaccines tracked in the data. Polio, Influenza, and Measles all fall in the high 80 to low 90 percent range, and barring a dip in Flu around 2009, have either risen or remained fairly steady from 1995 onwards. Less successful is Hepatitis B, underwent a significant increase in adoption in the early 2000's, but recently peaked near 70% adoption.



Plotting the year-over-year differences of vaccination coverage helps to better capture the variation in coverage rates, where above it's clear that some vaccines experience more notable changes in coverage than others. Analysis of changepoints in the provided data produced the following:





Of all the vaccines measured, the Influenza vaccine produced the most changepoints, indicating that this vaccine has the greatest volatility in coverage. Bayesian analysis of this volatility indicated that there is a high probability (virtually at 100%) of changepoints in year-over-year differences occuring 4 times in the recorded data, illustrated by the provided chart of posterior means and associated probabilities.

Proportions of Vaccination Data Reporting

Table 1: Schools by Reporting Status - Public and Private ### Proportion of Public Schools reporting vaccination data: 97%

	reported	N	Y
pubpriv			
PRIVATE		252	1397
PUBLIC		148	5584

Proportion of Private Schools reporting vaccination data: 85%

Reporting rates for both public and private schools in California were analyzed to determine whether a credible difference existed in reporting rates. To make this determination, two primary methods of statistical analyis were used (frequentist and Bayesian). This is true of all proceeding analysis.

Significance Testing on Difference in Proportions

```
##
##  Pearson's Chi-squared test
##
## data:  SchoolReportMF
## X-squared = 402.97, df = 1, p-value < 0.00000000000000022
```

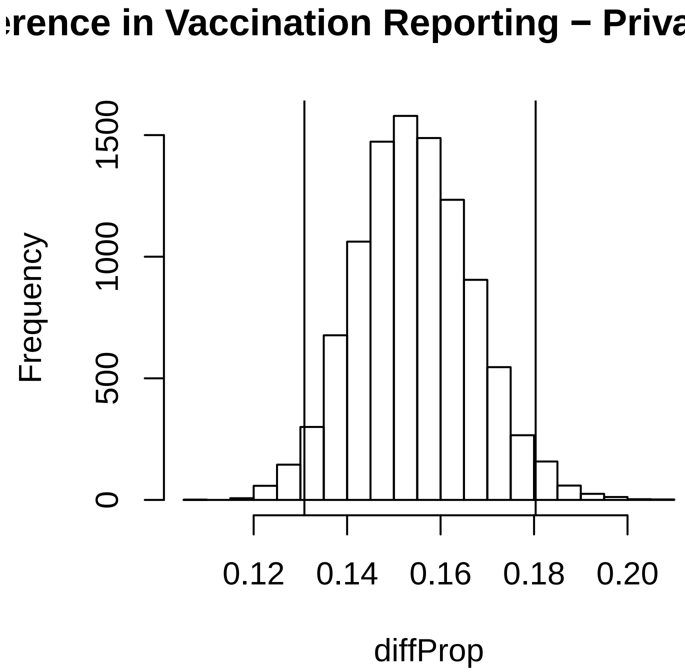
```
## Bayes factor analysis
## --------------
## [1] Non-indep. (a=1) : 1150547528972173601926488008460602286480242866844006468248684682882428 ±0%
##
## Against denominator:
##   Null, independence, a = 1
## ---
## Bayes factor type: BFcontingencyTable, poisson
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##               Mean    SD Naive SE Time-series SE
## lambda[1,1]  252.9 15.86   0.1586         0.1586
## lambda[2,1]  148.8 12.09   0.1209         0.1232
## lambda[1,2] 1397.0 37.07   0.3707         0.3707
## lambda[2,2] 5581.1 74.98   0.7498         0.7246
##
## 2. Quantiles for each variable:
##
##               2.5%    25%    50%    75%  97.5%
## lambda[1,1]  223.1  242.2  252.4  263.2  285.2
## lambda[2,1]  125.8  140.6  148.6  156.9  173.4
## lambda[1,2] 1325.2 1371.8 1396.9 1421.6 1470.1
## lambda[2,2] 5434.7 5530.7 5581.1 5631.7 5726.6
```

```
## [1] 0.1811527
```

```
## [1] 0.02667489
```

```
## [1] 0.1544778
```



The first statistical test produced a Chi-Squared value of 402.97, which has an associated p-value of well well below .001. Were there no difference in reporting proportion between public and private schools, one would expect a Chi-Squared value close to 1 (the associated degrees of freedom). The probability of getting a result of ~403 (or p-value of <.001) is so low that it is well outside any reasonable threshold to believe that is true (otherwise said that the null hypothesis should be kept). Thus it does appear that these proportions are significantly different from one another.

The second statistical test generated a value (known as a Bayes Factor) greater than 1 trillion, which can be interpreted to mean the odds that this difference in proportions would be found in a population are even above 1,000,000,000,000 to 1 - virtually irrefutable evidence that public school vaccination reporting rates are not the same as those in private schools in the wider population of public and private schools in California. The difference in proportions is 95% likely to be between 13 and 18, with the most likely value near 15.

Vaccination Rates in CA vs. US

Comparison the mean (average) values from California data in 2013

Table 2: Measures of Centrality and Dispersion - CA Vaccination Coverage 2013 and the WHO data on US Vaccination Coverage in 2013

WithDTP	WithPolio	WithMMR	WithHepB
Min. :23	Min. :23	Min. :23	Min. :23
1st Qu.:86	1st Qu.:87	1st Qu.:86	1st Qu.:90
Median :93	Median :94	Median :94	Median :96
Mean :90	Mean :90	Mean :90	Mean :92
3rd Qu.:97	3rd Qu.:97	3rd Qu.:97	3rd Qu.:98
Max. :100	Max. :100	Max. :100	Max. :100

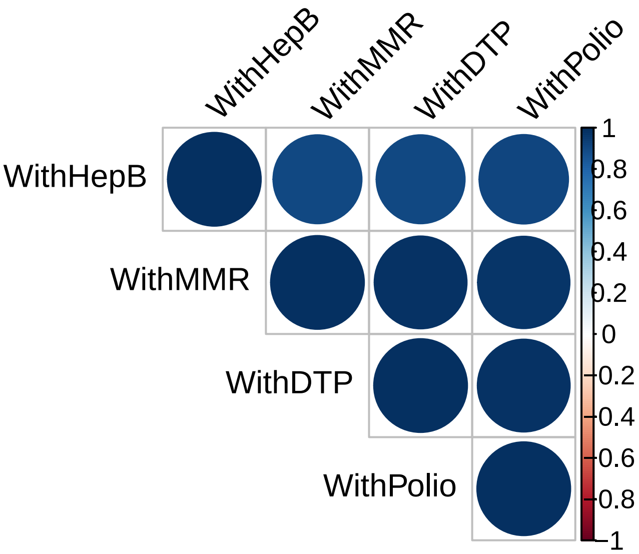
DTP1	HepB_BD	Pol3	Hib3	MCV1
98	74	93	93	92

indicates that California Vaccination Rates (in 2013):

- Rank below the national average in DTP (8 points below)
- Rank below the national average in Polio (3 points below)
- Rank below the national average in Measles (2 points below)
- Ranks well above the national average in Hepatitis B (28 points above)

How are Vaccination Rates related?

	WithDTP	WithPolio	WithMMR	WithHepB
WithDTP	1	0.98	0.98	0.9
WithPolio	0.98	1	0.97	0.91
WithMMR	0.98	0.97	1	0.9
WithHepB	0.9	0.91	0.9	1



There is a very strong relationship (correlation) between receiving one vaccine and receiving others. The plot and table above show very high correlation values for vaccination rates (closer to 1 indicates higher strength of relationship), indicating that it's virtually a given that a child with one vaccine has already received the others. This reflects the fact that infant vaccinations are often distributed in short succession of one another, if not simultaneously. The opposite is then also true - if a student is missing one vaccine, it is very likely they are missing the remainder.

Which predict district's reporting was complete or not?

Turning attention now to what demographic data can help inform whether a district will provide complete vaccination reporting:

```
##
## Call:
## glm(formula = DistrictComplete ~ PctChildPoverty + PctFreeMeal +
##     PctFamilyPoverty + school_student_ratio, family = binomial(),
##     data = districts)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7665   0.2514   0.3094   0.3679   0.8640
##
## Coefficients:
##                       Estimate Std. Error z value     Pr(>|z|)
## (Intercept)           2.912146   0.515216   5.652 0.0000000158 ***
## PctChildPoverty       0.028111   0.026935   1.044       0.2966
## PctFreeMeal          -0.016571   0.010018  -1.654       0.0981 .
## PctFamilyPoverty     -0.046873   0.034800  -1.347       0.1780
## school_student_ratio   0.010036   0.004301   2.333       0.0196 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 306.65  on 699  degrees of freedom
## Residual deviance: 293.88  on 695  degrees of freedom
## AIC: 303.88
##
## Number of Fisher Scoring iterations: 6
```

```
##          (Intercept)      PctChildPoverty          PctFreeMeal
##           18.3962256            1.0285101            0.9835654
##     PctFamilyPoverty school_student_ratio
##            0.9542089            1.0100861
```

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

```
##                          2.5 %    97.5 %
## (Intercept)          6.9454952 52.721466
## PctChildPoverty      0.9777703  1.086721
## PctFreeMeal          0.9642155  1.002954
## PctFamilyPoverty     0.8914077  1.022452
## school_student_ratio 1.0020193  1.019019
```

```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##      0.041639315      0.002506183      0.018075373      0.050957283
## McKelvey.Zavoina           Effron            Count        Adj.Count
##      0.092441385      0.013412853               NA               NA
##              AIC    Corrected.AIC
##    303.877020572    303.963475904
```

```
##
## 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)           2.96214 0.53212 0.0053212      0.0228509
## PctChildPoverty       0.03047 0.02667 0.0002667      0.0010679
## PctFreeMeal          -0.01750 0.01041 0.0001041      0.0004360
## PctFamilyPoverty     -0.04699 0.03409 0.0003409      0.0014156
## school_student_ratio  0.01023 0.00433 0.0000433      0.0001756
##
## 2. Quantiles for each variable:
##
##                           2.5%       25%      50%      75%    97.5%
## (Intercept)           1.962655  2.595961  2.94369  3.29966 4.075569
## PctChildPoverty      -0.019844  0.012288  0.02968  0.04782 0.085631
## PctFreeMeal          -0.038315 -0.024356 -0.01717 -0.01036 0.002667
## PctFamilyPoverty     -0.113641 -0.069780 -0.04683 -0.02371 0.018808
## school_student_ratio  0.002199  0.007246  0.01005  0.01292 0.019237
```

```
## [1] 1.010282
```

```
##     2.5%
## 1.002201
```

```
##    97.5%
## 1.019423
```

Through analysis of the summarized outputs of a model predicting reporting completion by the district and a summary of district demographics, we conclude the following:

- The p-values for percentage of children in poverty, children receiving free meals, and families in poverty in the district are all above *.05*, and therefore do not allow us to reject the null hypothesis for this model at an alpha threshold of .95, indicating that they do not add predictive power to the model. Because omitting them from the model reduces measures for predictive power (Nagelkirk Pseudo R-Squared decrease from .05 to .01), they will remain in the model.
- However, a calculated ratio of students per school (meant to approximate the values of enrolled student count and total schools) has a p-value below *.001*, which is evidence in favor of rejecting the null hypothesis, suggesting the ratio of students to schools in the district adds predictive power to this model.

In terms of log odds, there appears to be a very slight *1.01:1* change in odds for each increase in student school ratio, indicating that there is a virtually negligible effect of student to school ratio on the likelihood of a district completing their reporting.

Bayesian analysis further allows us to conclude with 95% confidence that the effect of student-to-school ratio on the log odds of a district's reporting completion is an increase between *.002 - .019%*, with the most likely value calculated around *.01*.

The status quo model generated a Nagelkerke pseudo-R-squared value of *0.05*. Used as a loose replacement for an adjusted R-squared value in a typical linear regression, this roughly speaking says that the given model can explain about 6% of the variance predicting district completion.

Overall is clear that the provided data does not offer conclusive evidence that any of these district demographics substantively impact vaccination reporting rates.

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

```
## 
## Call:
## lm(formula = PctUpToDate ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty + 
##     school_student_ratio, data = districts)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -66.528  -3.492   2.499   6.675  22.438
```

```
##
## Coefficients:
##                      Estimate Std. Error t value             Pr(>|t|)
## (Intercept)          77.22283    1.26278  61.153 < 0.0000000000000002 ***
## PctChildPoverty      -0.10498    0.07822  -1.342               0.1800
## PctFreeMeal           0.11848    0.02780   4.262       0.000023017553 ***
## PctFamilyPoverty      0.20477    0.11148   1.837               0.0667 .
## school_student_ratio  0.06472    0.01034   6.262       0.000000000666 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.71 on 695 degrees of freedom
## Multiple R-squared:  0.1388, Adjusted R-squared:  0.1339
## F-statistic: 28.01 on 4 and 695 DF,  p-value: < 0.00000000000000022
```

PctChildPoverty	PctFreeMeal	PctFamilyPoverty	school_student_ratio
4.429	2.398	4.064	1.058

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + school_student_ratio : 2395196966343094272 ±0
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

This model produced a value for R-squared of *0.1339*, with an F-test *F(4, 695)=28.01, p< 0.00000000000000022*, providing reasonably strong evidence to reject the null hypothesis that R-squared was equal to zero - all this to say, this model can be used to estimate the effect of the provided demographic factors on percentage of students with full vaccinations The low R-squared value (closer to 0 than 1) does indicate this model's predicted effects should be further explored in subsequent analysis, as it is unable to explain more than ~13 of the variation in the data.

Significant Variables:

- Percentage of Students who use the free meals program (positive effect)
- School-to-Student ratio (positive effect)

The resulting Bayes Factor - Well above *1 Trillion* - strongly favors a model that includes these variables to predict the percentage of students with full vaccinations over a model without these factors.

These results certainly add import to the conclusion that districts with higher numbers of students eligible for free meals, and higher volume of students, possess a higher percentage of students with up-to-date vaccines, based on the positive coefficients on these variables and the strong indicators that these factors possess some predictive power.

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

```
##
```

```
## Call:
## lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFreeMeal +
##     PctFamilyPoverty + school_student_ratio, data = districts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.741  -3.997  -1.676   1.263  64.399
##
## Coefficients:
##                       Estimate Std. Error t value             Pr(>|t|)
## (Intercept)          13.233408   0.890899  14.854 < 0.0000000000000002 ***
## PctChildPoverty       0.122928   0.055182   2.228               0.0262 *
## PctFreeMeal          -0.121368   0.019610  -6.189        0.00000000103 ***
## PctFamilyPoverty     -0.114505   0.078652  -1.456               0.1459
## school_student_ratio -0.041673   0.007292  -5.715        0.00000001629 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.262 on 695 degrees of freedom
## Multiple R-squared:  0.1496, Adjusted R-squared:  0.1447
## F-statistic: 30.57 on 4 and 695 DF,  p-value: < 0.00000000000000022
```

PctChildPoverty	PctFreeMeal	PctFamilyPoverty	school_student_ratio
4.429	2.398	4.064	1.058

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + school_student_ratio : 175953998308925964298
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

This model produced a considerably high value for R-squared of *0.8148*, with an F-test *F(4, 695)=30.57, p<0.00000000000000022*, once again providing reasonably strong evidence to reject the null hypothesis that R-squared was equal to zero, and that this model can be used to estimate the effect of the provided demographic factors on the percentage of students with religious belief exemptions. Significant Variables:

- Percentage of children in the district living below the poverty line (positive effect)
- Percentage of students who are eligible for free meals (negative effect)
- School-to-Student ratio (negative effect)

The resulting Bayes Factor - Well above *1 Trillion* - again strongly favors a model that includes these variables to predict the percentage of students with religious belief exemptions over a model without these factors.

Summary of Findings

When considered in concert, there are definite indicators of a tale of two districts - one in an urban center, where access and proximity to healthcare is abundant, and that of a more remote rural location, where the nearest hospital is many miles away. Initial modeling showed that only a proxy measure for population had any significance in predicting district reporting completion (school/student ratio), and measures of poverty and food scarcity combined with higher enrolled student counts (think: city) show a positive relationship with up-to-date vaccination rates. In contrast, higher rates of belief exemptions to vaccination appeared to be related to areas with consistent levels of poverty, but fewer students (think: rural).

Above all, it is recommended additional data be gathered on the rural and urban splits in these districts to confirm the indicators in the provided data. Assuming these findings can be corroborated, it is then recommended that the legislator allocate financial assistance to rural areas with high concentrations of students below the poverty level.