US Vaccine Analysis

By Leland Ball

2020-09-14

Contents

Introduction

Vaccinations remain an important part of a healthy modern society. The United States has been vaccinating its population for as long as there have been vaccines. One of the most effective ways of administering important vaccines is within the school system, where younger citizens can receive needed treatment at the appropriate age. Funding at the local, state, and federal levels can be tied to this activity, so it is important to get it right.

Belief exemptions for vaccinations exist in the US, which plays into the religious liberties that the country was founded on. Nevertheless it is important to vaccinate as many people as possible to prevent the spread of preventable diseases to others, especially vulnerable populations. To this end, data has been collected on US and California district schools in hopes of ascertaining the best method to apply funding that will increase vaccination percentages.

In this analysis, three datasets of vaccine information will be analyzed for the purposes of gaining insight into how the current Californian public school system's vaccine rates can be improved.

Analysis

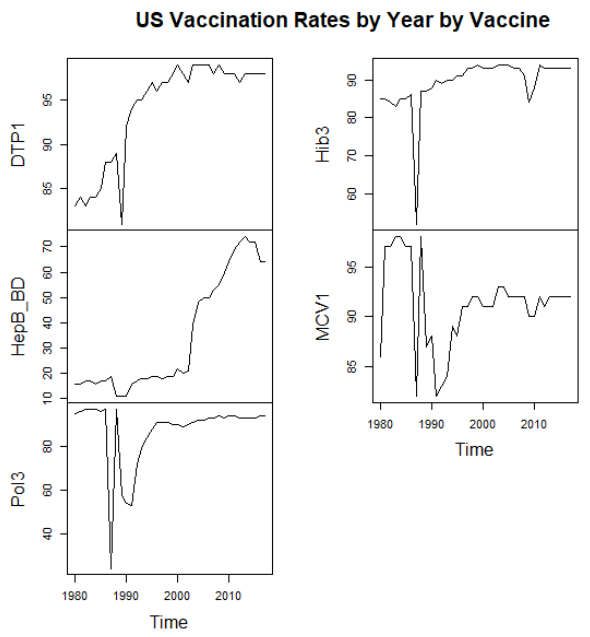
About the Data

- Three datasets of vaccine information for the USA, California Kindergartens, and California public schools
- All rows in the data are "complete" and not missing any data in any column. No data balancing is necessary (for time-series analysis)
- Data normalization carried out in preparation for GLM and LM models (see "Predictive Analysis" section)

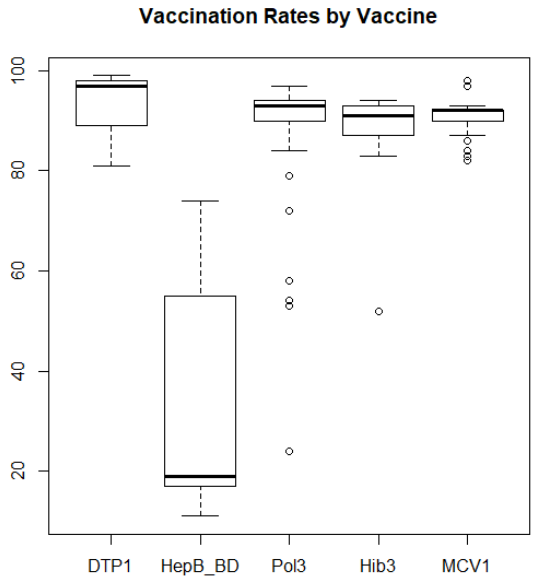
Descriptive Report

How Have U.S. Vaccination Rates Varied over Time?

Over the years, the vaccination rate for most vaccines has generally trended upwards, and leveled off at a high percentage (low-mid 90%). The two exceptions to this is the HepB_BD vaccine which trended upward until settling around 65%, and the MCV1 vaccine which has been north of 80% for the duration of the data in this analysis. The following diagrams illustrate these changes over time.



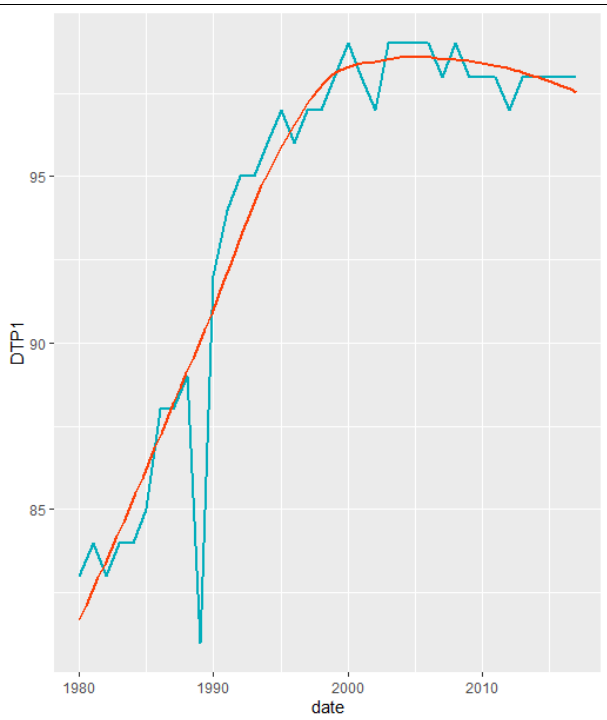
Plot 1. Time-series plots of vaccine percentages over time. This raw data shows some volatility and trends.



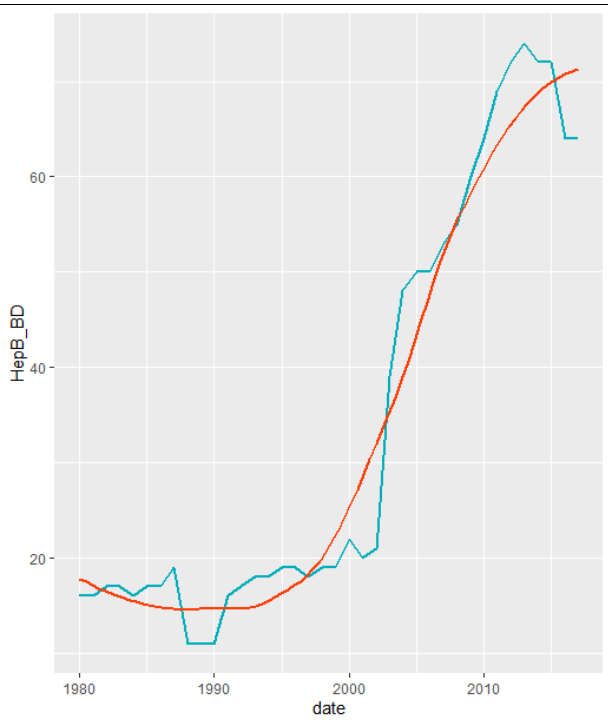
Plot 2. Boxplot of vaccination rates illustrates the variation and long-term trends of the vaccination rate for each vaccine.

Are vaccination rates increasing or decreasing?

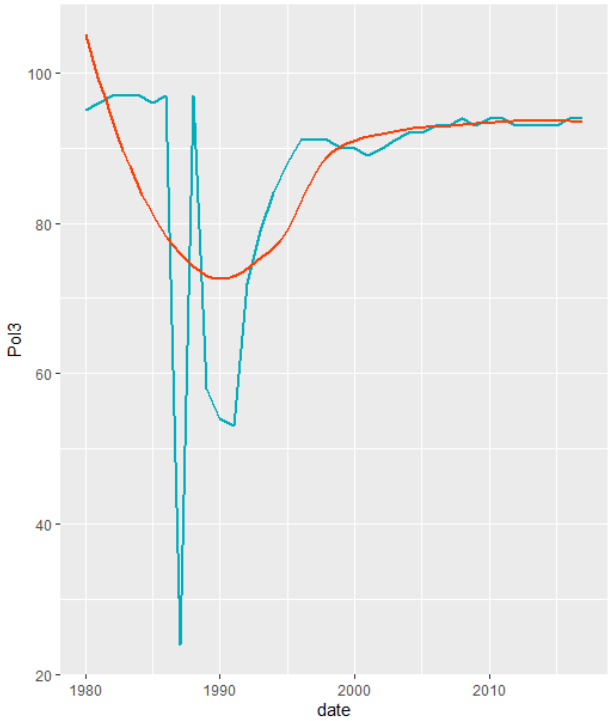
While vaccination rates appear to be increasing on the whole, this is not necessarily the case across all vaccines. The plots below show the rates for each vaccine, and a running average trend line plotted in orange. These graphs show that while vaccines such as DTP1 have increased since 1980, there has recently been a minor decline. Other vaccines can be said to be holding steady.



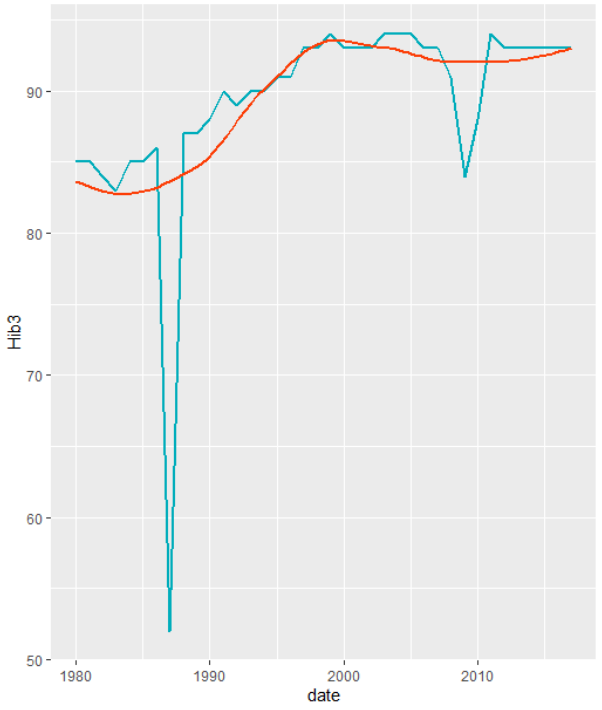
Plot 3. DTP1 Vaccination Rates by Date with Trend Line



Plot 4. HepB_BD Vaccination Rates by Date with Trend Line



Plot 5. Pol3 Vaccination Rates by Date with Trend Line



Plot 6. Hib3 Vaccination Rates by Date with Trend Line

Plot 7. MCV1 Vaccination Rates by Date with Trend Line

Not content with merely a visual inspection, a dependent samples t-test (one sample test on differences) was conducted on the difference in vaccine rates from the beginning of the data in 1980 to the end of the data in 2017. The Null Hypothesis for this test states that there is no difference in the means of these two groups. The Alternative Hypothesis states that there is in fact a difference in the means of these two groups.

The results of this test yielded a p-value of $p=0.1514>a=0.05$. The p-value being greater than the alpha value implies that we cannot reject the Null Hypothesis. The difference in these five vaccines from 1980 to 2017 is not in itself statistically significant from a Frequentist point of view.

A Bayesian test was also performed upon the same differenced vaccination rates. The Null and Alternative Hypotheses are identical in this case. The 95% Highest Density Interval (HDI) for this analysis was -14.16 to 46.31. This range includes zero, so the Null Hypothesis cannot be rejected. This leaves the Alternative Hypothesis: There is no credible statistical difference in vaccination rates since 1980, when all five vaccines are considered.

Which vaccination has the highest rate at the conclusion of the time series?

At the conclusion of the time series in 2017, the **DTP1** vaccine had the highest vaccination rate in the U.S.

DTP1	HepB_BD	Pol3	Hib3	MCV1
98	64	94	93	92

Which vaccination has the lowest rate at the conclusion of the time series?

At the conclusion of the time series in 2017, the **HepB_BD** vaccine had the lowest vaccination rate in the U.S.

DTP1	**HepB_BD**	Pol3	Hib3	MCV1
98	**64**	94	93	92

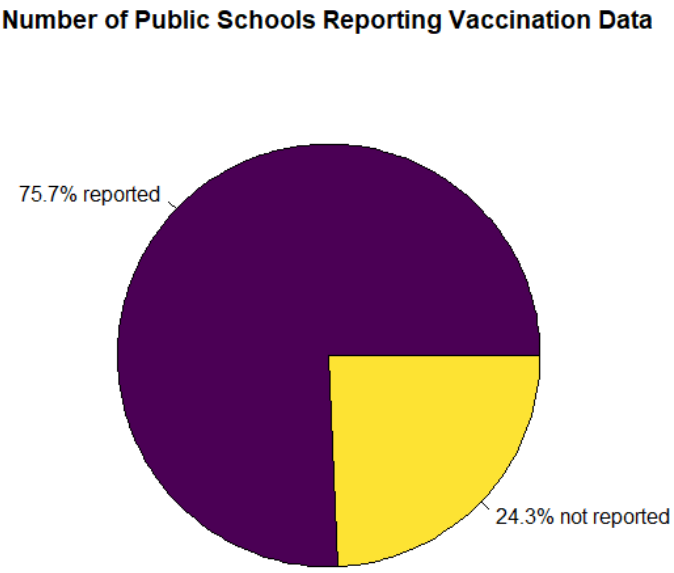
Which vaccine has the greatest volatility?

The standard deviation of each vaccine's data was calculated as a measurement of variance. While **HepB_BD** was found to have the most variation in the raw data, this is likely because it also has one of the largest ranges, with significantly many values. Pol3 shares this range, but only due to a few brief dips over the years, not large trends as is seen in HepB_BD's data. To identify which vaccine had the largest variance regardless of trends, a differencing technique was applied to the data and the standard deviation was calculated again. At this point, **Pol3** showed the largest variation in the data, which highlights the steepness and large magnitude of the changes in its data. See Plot 1. The standard deviations are as follows:

vaccine	sd	de-trended sd
DTP1	5.87	2.44
HepB_BD	**22.54**	4.16
Pol3	15.35	**18.76**
Hib3	7.14	8.35
MCV1	4.19	4.76

What proportion of public schools reported vaccination data?

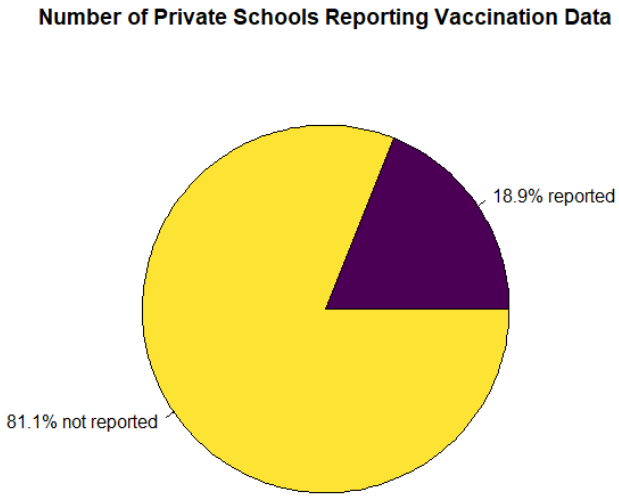
Seventy-Six percent of public schools reported vaccination data for the year 2013



Plot 8. Public School Vaccination Reporting Percent

What proportion of private schools reported vaccination data?

Nineteen percent of public schools reported vaccination data for the year 2013

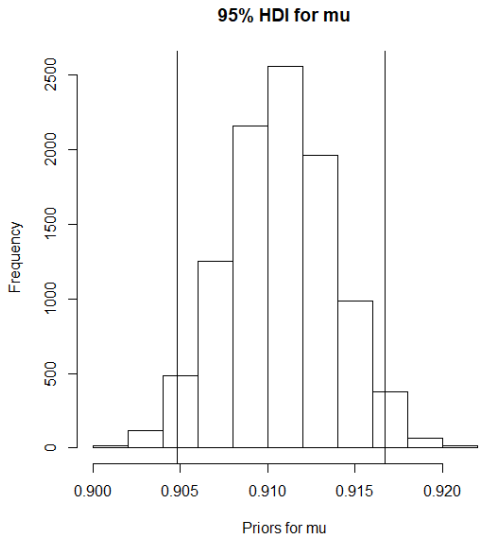


Plot 9. Private School Vaccination Reporting Percent

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

Credible difference implies more than a feeling in overall reporting proportions. To this end, a frequentist ANOVA test was conducted with the Null Hypothesis being that both groups were sampled from the same population, while the Alternative Hypothesis stated that The groups come from different populations. The results showed a large F-score of 426, and significant p=<2e-16 < a=0.05 leading to the rejection of the Null Hypothesis. These groups come from different populations, implying that public and private schools have credible differences between them when it comes to overall reporting proportions

An additional Bayesian Analysis had similar results. In a Bayesian Factor Analysis, the Bayesian ANOVA found very large odds (1.770577e+88 ±0% to 1) that the public and private school groups had different reporting proportions . Inspection of the 95% Highest Density Interval of 0.90 to 0.92 shows no overlap with 0, and coefficient point-estimate is 0.91 corroborating the Frequentist findings that there is a statistically significant difference between these two groups.



Plot 10. 95% HDI for Mean Difference in Public/Private School Reporting Means

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

The vaccination rates (as a percent) are included below. This 2013 California Public Schools data does not include the Hib3 (influenza, third dose) vaccine included in the U.S. statistics as a whole.

DTP	HepB	Polio	MMR
89.7	92.2	90.1	89.8

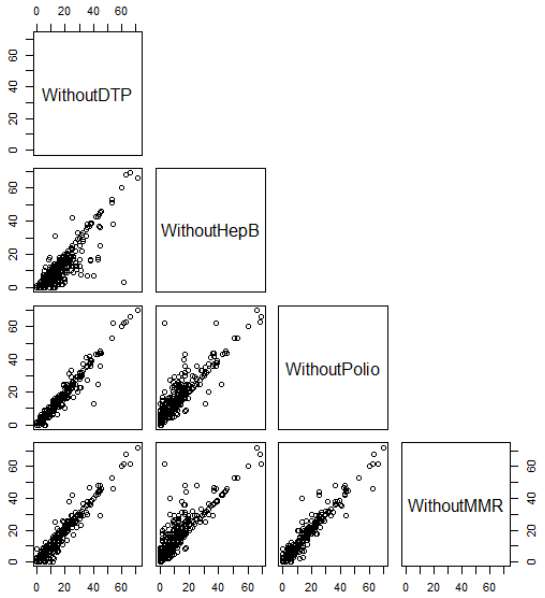
How do these rates for individual vaccines in California districts compare with overall US vaccination rates?

The 2013 mean California Public School vaccination rates were under the US average rates for the same year for DTP1/DTP, Polio/Pol3, MMR/MCV1 and over average for HepB/HepB_BD This remains true for the US average vaccination rates in 2017. Informally, CA has a much higher rate of HepB/HepB_BD vaccination than the US average.

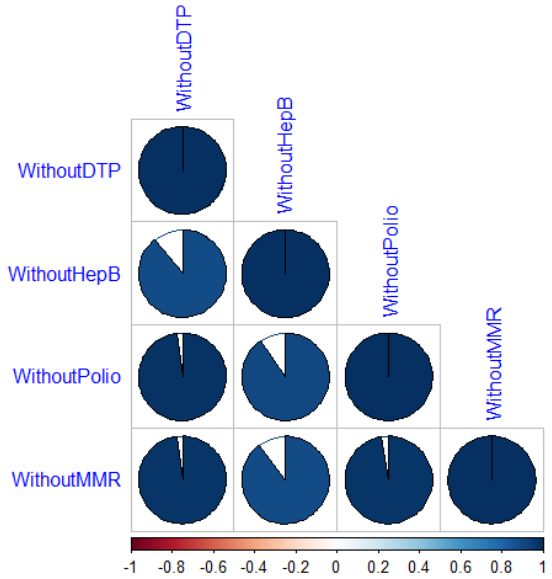
dataset	year	DTP1	HepB_BD	Pol3	Hib3	MCV1
USA	2013	98	74	93	93	92
USA	2017	98	64	94	93	92
California	2013	89.7	92.2	90.1	NA	89.8

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? To measure this effect, the correlation among each vaccine's adoption rate over the years was analyzed. This was conducted on the Californian vaccine rates from 2013, across both public and private schools. The results showed statistically significant correlation among all vaccines at an alpha level of 0.05. As a formality, the Null Hypothesis for the significance test carried out states that there is no correlation between a given two vaccines. The Alternative Hypothesis states that there is correlation between a given two vaccines. This test was carried out automatically using the corrplot() function, seen on the right. Only values exceeding the given significance are plotted. All values are plotted.



Plot 11. Correlation Plots by Vaccine Type



Plot 12. Correlation Percent by Vaccine Type

The raw correlation values have been included below.

	WithoutDTP	WithoutHepB	WithoutPolio	WithoutMMR
WithoutDTP	1	0.8905016	0.981788	0.978404
WithoutHepB	0.890502	1	0.905733	0.896808
WithoutPolio	0.981788	0.9057329	1	0.971818
WithoutMMR	0.978404	0.8968082	0.971818	1

A Bayesian correlation test was also run between all pairs of vaccines. Results from this test show a high degree of correlation, which was able to be measured quite accurately. The "Inf" results in some of the results are artifacts of rounding very large numbers and should be taken to be extremely likely odds. The point estimate and Factor Analysis odds (showing odds of significance) are included below.

Vaccine Pair Compared	Bayes Factor Odds	Point Estimate
WithoutHepB & WithoutDTP	3.13E+236	0.8899
WithoutPolio & WithoutDTP	Inf	0.9818
WithoutMMR & WithoutDTP	Inf	0.9781
WithoutPolio & WithoutHepB	8.67E+257	0.9051
WithoutMMR & WithoutHepB	9.23E+244	0.8962
WithoutMMR & WithoutPolio	Inf	0.9715

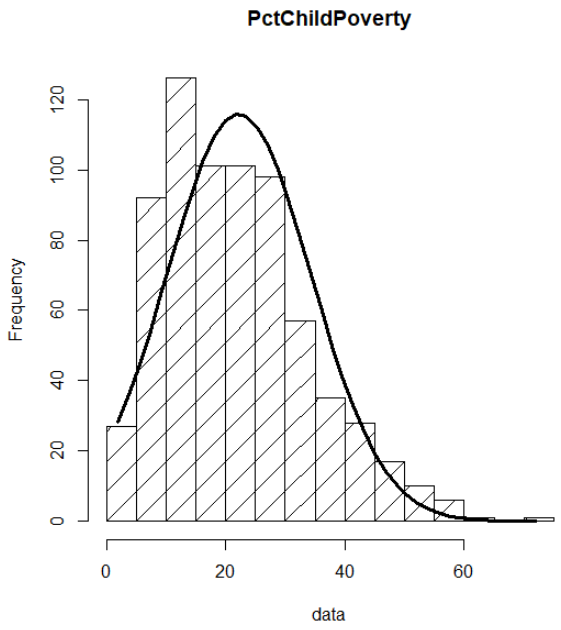
In summary, both the Frequentist and Bayesian methods agree that if a school has students who are missing one vaccine, they are very likely to be missing other vaccines as well. This holds true across all vaccine types.

Predictive Analysis

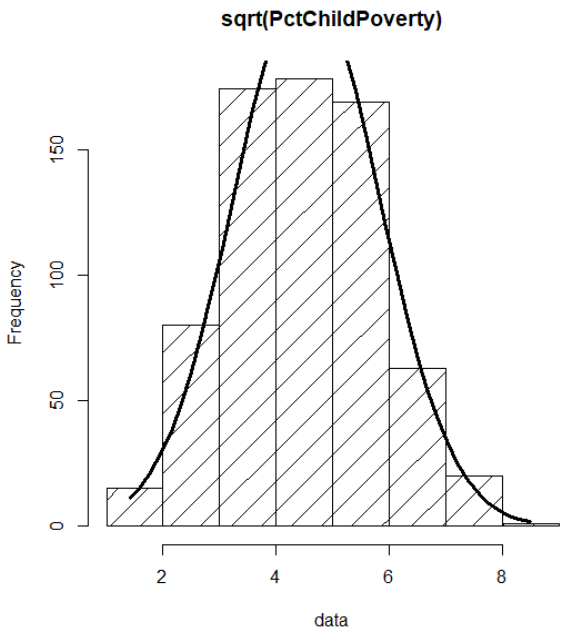
Variable Transformations - Exploring how well each variable fits a visual normal curve

Many of these models expect normalized data. To this end, the relevant variables in the data were given visual inspection for conformity with the normal curve, and transformed where appropriate. What follows is a before/after view of each variable's distribution alongside a normal curve matching the appropriate standard deviation and mean of the variable.

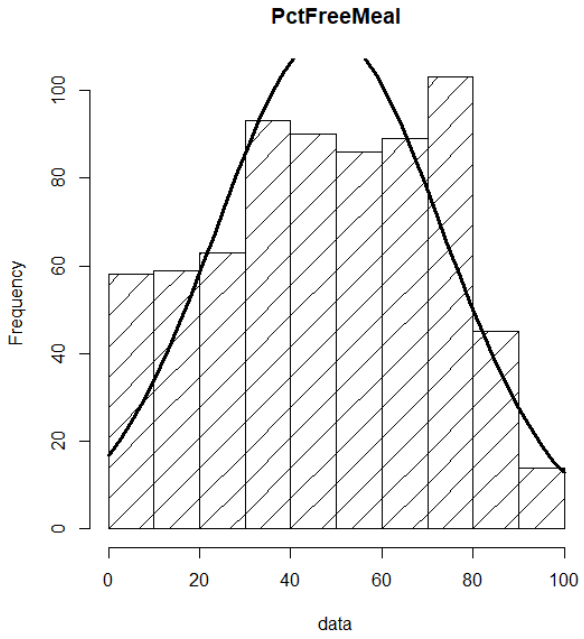
Unmodified Variable Distribution	Transformed Variable Distribution



Plot 13. PctChildPoverty Variable, Pre-Normalization

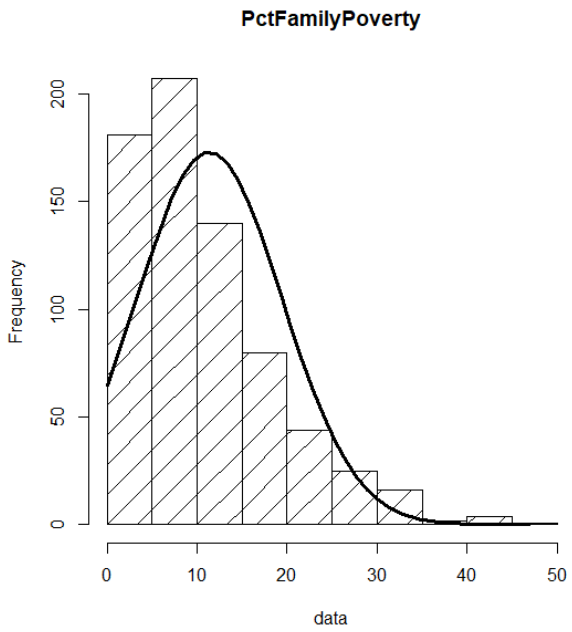


Plot 14. PctChildPoverty Variable, Post-Normalization via Square Root

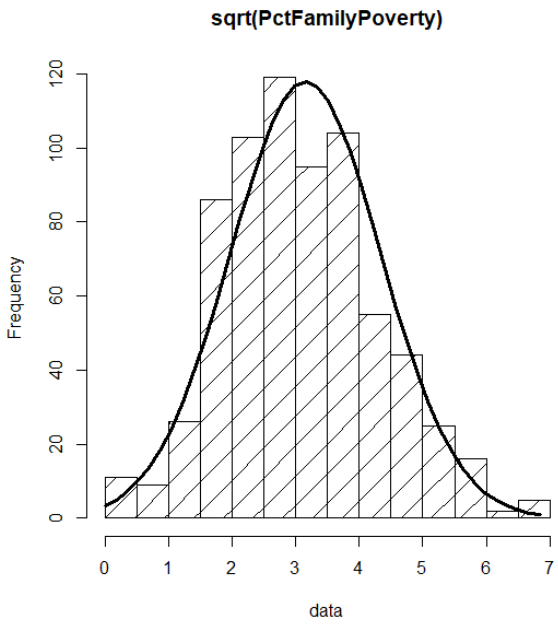


Plot 15. PctFreeMeal Variable

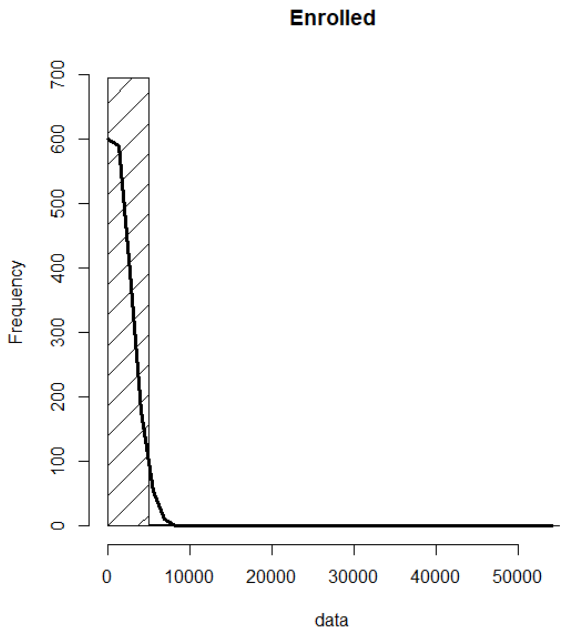
No variable transformations were found to significantly normalize the distribution of this variable



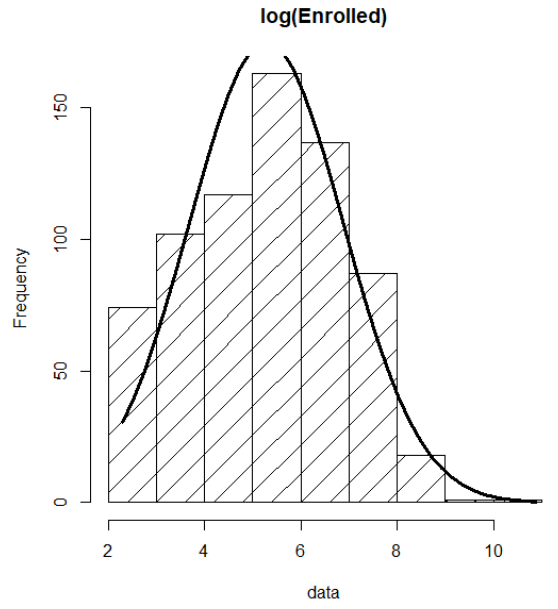
Plot 16. PctFamilyPoverty Variable, Pre-Normalization



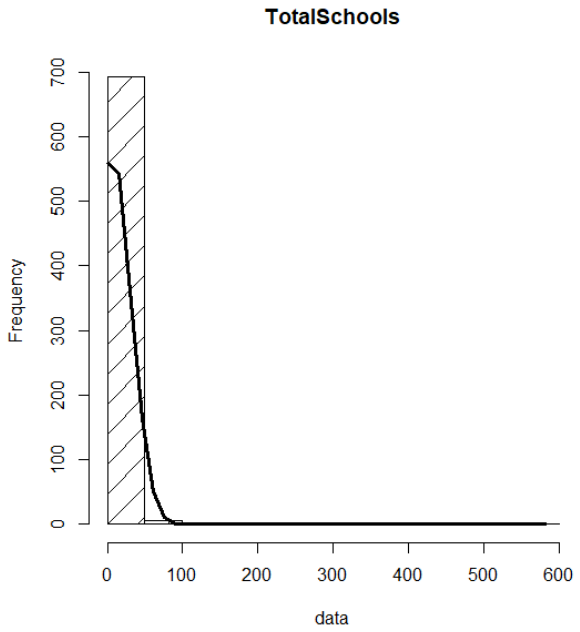
Plot 17. PctFamilyPoverty Variable, Post-Normalization via Square Root



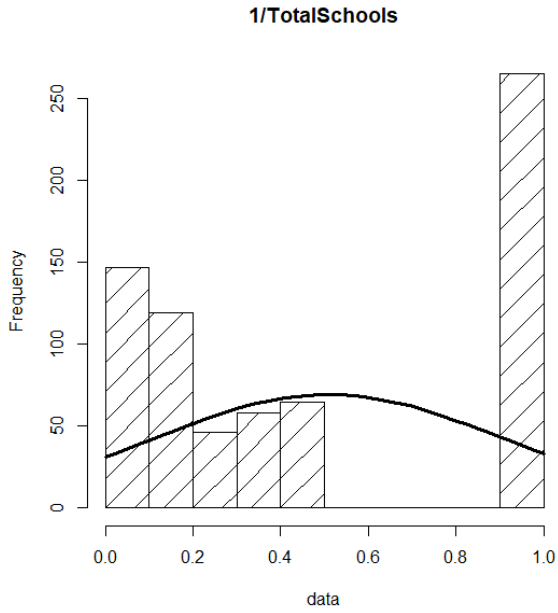
Plot 18. Enrolled Variable, Pre-Normalization



Plot 19. Enrolled Variable, Post-Normalization via Logarithm



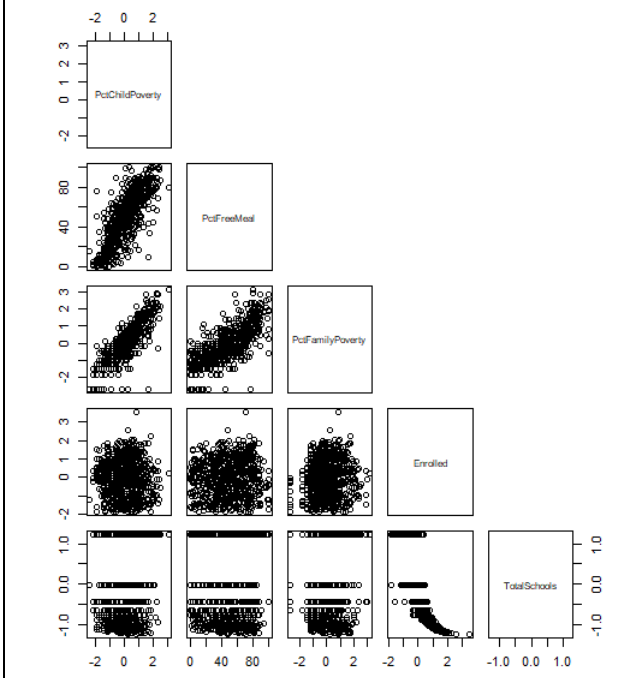
Plot 20. TotalSchools Variable, Pre-Normalization



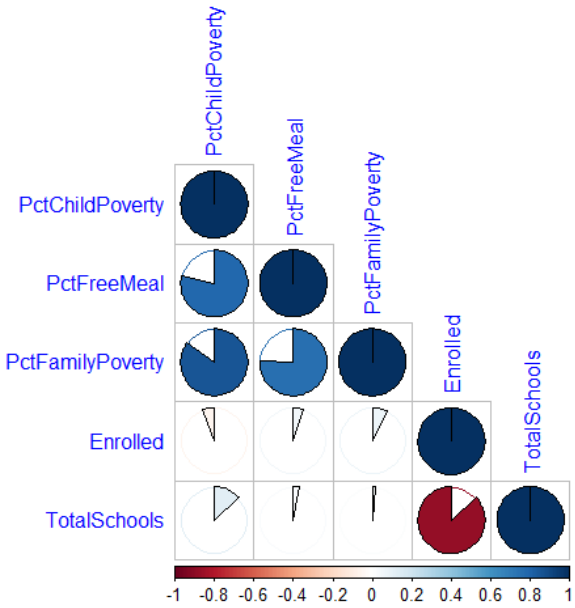
Plot 21. TotalSchools Variable, Post-Normalization via Ratio

It is important to note that after this variable transformation, the output of various coefficients and odds are scaled differently. Even real-odds likelihoods that are normally in terms of the units of the coefficients will instead need to be interpreted in units of the transformed variables, so a real-odds for a coefficient's effect of 2:1 would actually be 2^2:1, in the case of a square root.

After the data transformation, a brief look was made into correlations between these variables. There is a significant, strong, positive correlation between PctChildPoverty, PctFreeMeal and PctFamilyPoverty. There is also a significant, strong, negative correlation between Enrolled and TotalSchools.



Plot 21b. Correlations of Transformed Variables



Plot 21c. Correlations of Transformed Variables

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

PctFreeMeal, **Enrolled**, and **TotalSchools**

For this analysis, the DistrictComplete variable was converted into a binary numeric value for use with the following model. A Generalized Linear Model was used to perform logistic regression in order to explore how well each of the following variables can predict DistrictComplete (whether or not a district had complete reporting): PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools. The variable descriptions are located in the Data Dictionary section of the Appendix.

The strength of this logistic model was compared with the null model using AIC scores. The Null deviance received a score of 328, df=699 while the Residual deviance for the logistic model received a score of 255, df=694. The lower AIC scores shows that the logistic model performed better at explaining the dependent variable (DistrictComplete) than the null model.

Further exploration of each coefficient in the model was undertaken. To begin with, a Null Hypothesis for each coefficient was created stating that the given coefficient had no effect upon the dependent variable. The Alternative Hypothesis being that there was an effect. Analyzing the remaining coefficients, only Enrolled and TotalSchools was found to have a significant effect at p<a=0.05. The remaining coefficients had no such significant effect at this level. This led to the rejection of the Null Hypothesis for the Enrolled and TotalSchools variables, implying that they do have an effect on DistrictComplete, while we were unable to reject the Null Hypothesis for any of the other terms in the

model. This leads us to conclude that PctChildPoverty, PctFreeMeal, and PctFamilyPoverty have no statistically significant effect upon DistrictComplete at the significance level of a=0.05.

A Chi-Squared ANOVA test on this logistic GLM found identical results (using the same hypothesis above). Again the Enrolled and TotalSchools variables were found to have statistically significant effect upon DistrictComplete at the level of a=0.05.

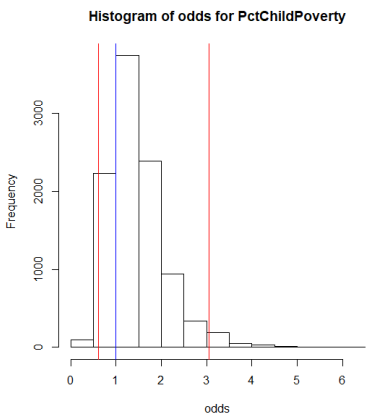
A pseudo-R-Squared value was calculated for this model and found to have a Nagelkerke value of 0.26 and a McFadden value of 0.22. This indicates that about 22% to 26% of the variance in the DistrictComplete variable can be modeled by the other five terms. While this model is statistically significant, it does not explain a majority of the reason why a district would have incomplete reporting.

A Variance Index Factor (VIF) analysis was done on all terms in this GLM with the following results:

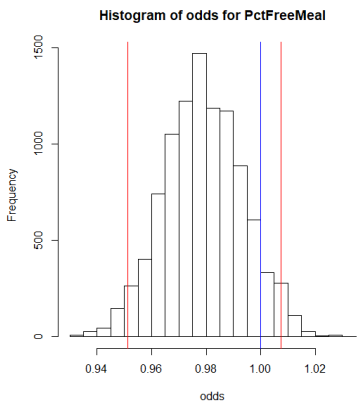
PctChildPoverty	PctFreeMeal	PctFamilyPoverty	Enrolled	TotalSchools
8.979757	3.188573	7.267617	4.189732	4.023604

Values of over 5 or 10 show multicollinearity that adversely affect the model. It is likely that PctFamilyPoverty and PctChildPoverty are collinear. This makes sense as they are both related forms of poverty. Thus, one or both of these terms should be removed from the model. However, upon removing either one from the model, no additional statistical significance was seen in the coefficients. When building a model with just TotalSchools, Enrolled, and one of (PctFreeMeal, or PctFamilyPoverty), the two terms PctFreeMeal or PctFamilyPoverty were found to have a statistically significant effect upon the model, however these models all received lower pseudo-R-squared values and higher AIC values. This implies that the best Frequentist model incorporates PctFreeMeal, Enrolled, and TotalSchools, and leaves out PctChildPoverty and PctFamilyPoverty for reasons of multicollinearity and lack of statistical significance as terms in the model.

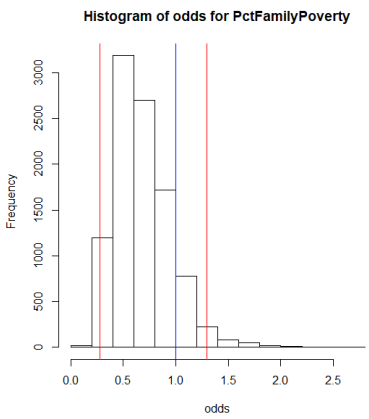
A Bayesian logistic model was then employed to analyze the same data, using the same five terms. The 95% confidence intervals for each real-odds likelihood that the coefficient has an effect on the model are included below. It is important to note that if this interval overlaps with 1 (implying 1:1 odds) then it is most likely that this coefficient has no effect upon the model. For clarity, 1 is marked with a blue line, while the HDI endpoints are marked with red lines.



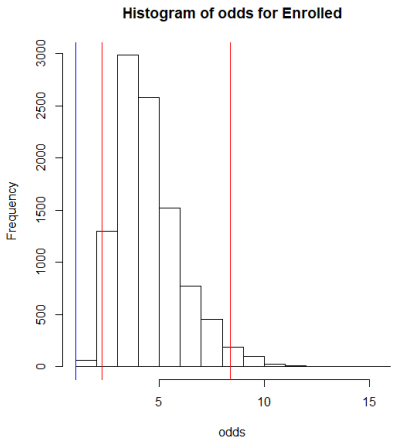
Plot 22. HDI Odds for PctChildPoverty Showing Insignificance



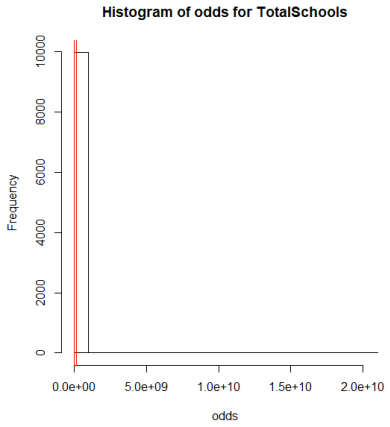
Plot 23. HDI Odds for PctFreeMeal Showing Insignificance



Plot 24. HDI Odds for PctFamilyPoverty Showing Insignificance



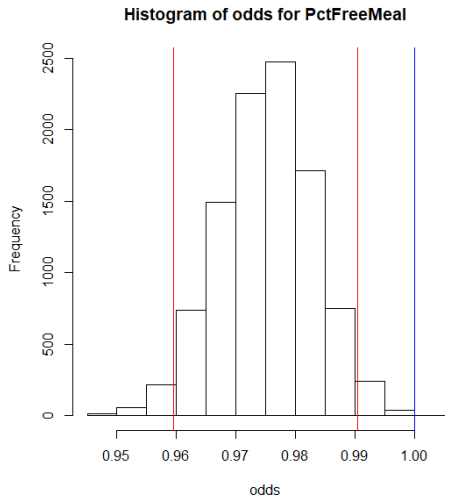
Plot 25. HDI Odds for Enrolled Showing Significance



Plot 26. HDI Odds for TotalSchools Showing Significance

The only difficult-to-interpret histogram is the odds for TotalSchools, which has a very narrow range, but does not include 1. These plots show that the only two coefficients that have a significant effect upon the model are Enrolled and TotalSchools. This agrees with the Frequentist analysis of the same data.

Going back to further exploration of the removal of multicollinear terms which have no effect upon the model, like with the Frequentist method, a Bayesian GLM with just PctFreeMeal, Enrolled, and TotalSchools was constructed, which then found significance to the PctFreeMeal term which was not there before.



Plot 27. HDI Odds for PctFreeMeal Now Showing Significance

Again this shows that in the absence of the poverty terms, the PctFreeMeal term does indeed have significance and should be included in the model going forward.

To summarize the effects of these coefficients:

The likelihood that a district's reporting was complete increases as the number of students eligible to receive a free meal decreases.

The likelihood that a district's reporting was complete increases as the number of students enrolled increases.

The likelihood that a district's reporting was complete increases as the total number of schools in a district decreases.*

*This coefficient was inverted by the 1/x data normalization, but that inversion has been taken into account for this statement.

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

PctFreeMeal and **Enrolled**

As this question is specifically asking for “completely” up-to-date vaccine percentages, the data for vaccine percentages (PctUpToDate) was first transformed into a binary numeric variable for use with the following models. With this new binarized dependent variable, a logistical model was employed to evaluate the effect that PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools have upon PctUpToDate (binarized).

The model with all five terms had a Pseudo-R-Squared Nagelkerke value of 0.27 and a McFadden value of 0.23, implying that this model accounts for an estimated 23% to 27% of the variance in PctUpToDate.

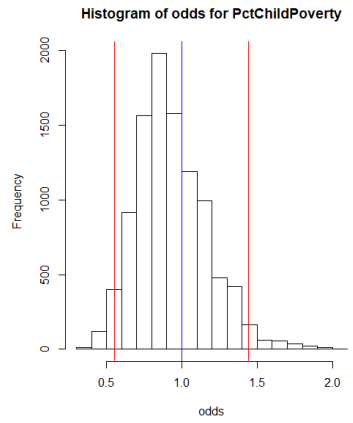
A Null Hypothesis stating that each given term had no effect upon PctUpToDate was used along with an Alternative Hypothesis stating that a given term does have an effect upon PctUpToDate. Low p-values for PctFreeMeal and Enrolled were found to be significant at the a=0.05 level, causing us to reject the Null Hypothesis in favor of the Alternative Hypothesis for these cases. In all other cases (PctChildPoverty, PctFamilyPoverty, and TotalSchools) the Null Hypothesis could not be rejected.

A further Chi-Squared ANOVA test conducted on this model found that PctChildPoverty and Enrolled were significant (per the above hypothesis, with a significance of a=0.05). This somewhat contradictory result led to the construction of the next model.

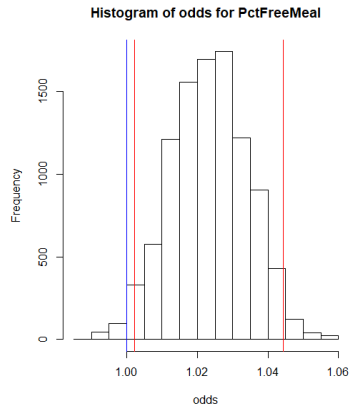
The next model compared just the significant terms: PctChildPoverty, PctFreeMeal, and Enrolled. The logistical regression model again reported **no** significance to PctChildPoverty, and in this case, the Chi-Squared ANOVA test results were different, and in agreement with the GLM that PctChildPoverty did not have significance at this alpha level. It is worth noting that a model containing just the PctFreeMeal and Enrolled terms received Psuedo-R-Squared values of Nagelkerke: 0.26 and McFadden: 0.22, which was around the same values obtained in the earlier models.

VIF analysis found no significant multicollinearity among any variables in this instance.

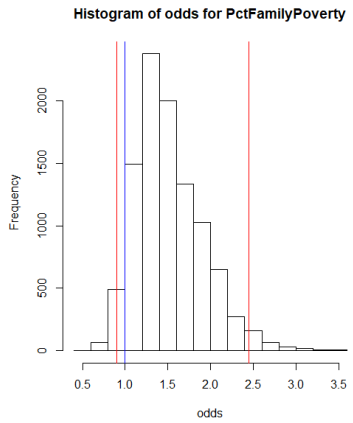
A Bayesian logistic model was then created to predict PctUpToDate with the same terms as in the Frequentist model. Upon analyzing the coefficients it was found that PctFreeMeal and Enrolled were the only two variables that had a likelihood of affecting PctUpToDate. Below is a plot of each coefficient's distribution with its 95% HDI. The blue line represents 1, implying a 1:1 real-odds likelihood of affecting the coefficient. A the 1:1 odds should lie outside of the 95% confidence interval to be considered to have a significant effect upon the model.



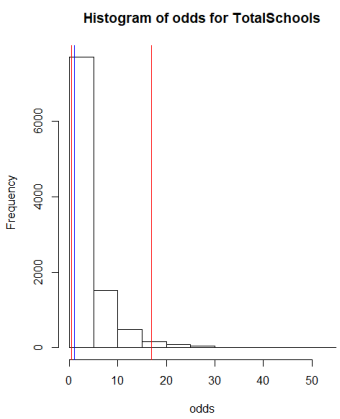
Plot 28. HDI Odds for PctChildPoverty Showing Insignificance



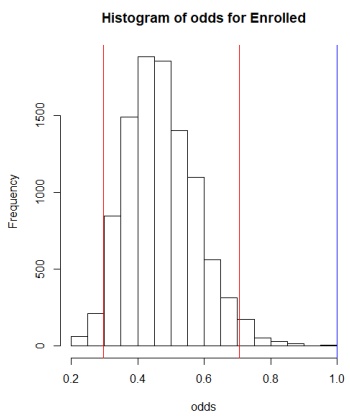
Plot 29. HDI Odds for PctFreeMeal Showing Significance



Plot 30. HDI Odds for PctFamilyPoverty Showing Insignificance



Plot 31. HDI Odds for TotalSchools Showing Insignificance



Plot 32. HDI Odds for Enrolled Showing Significance

These results agree with the Frequentist model initially calculated, and show that PctFreeMeal and Enrolled are the only two significant predictors of PctUpToDate (100%).

To summarize the effects of these coefficients:

The percentage of enrolled students with completely up-to-date vaccines increases as the number of students eligible to receive a free meal increases.

The percentage of enrolled students with completely up-to-date vaccines increases as the number of enrolled students decreases.

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

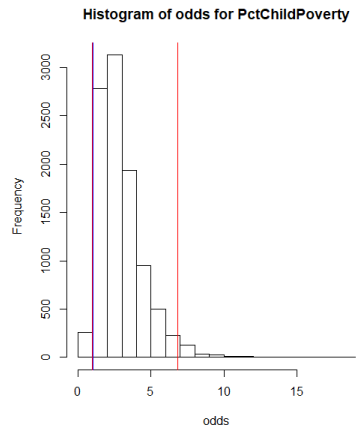
TotalSchools, **Enrolled**, and **PctFreeMeal**

Due to the non-binary nature of the dependent variable (PctBeliefExempt), and all independent variables, a linear model was chosen to analyze this topic. A linear model was constructed with the following Null Hypothesis: These terms have no effect upon PctBeliefExempt. The Alternative Hypothesis stated that: these terms do have an effect upon PctBeliefExempt. The constructed model was significant at p=2.2e-16<a=0.05. The F-Statistic was also large at 31.21 on 5 and 694 degrees of freedom. This results in the rejection of the Null Hypothesis, leaving the Alternative Hypothesis that these five terms do have a statistically significant effect upon PctBeliefExempt. The Adjusted R-Squared value was low, at 0.18, implying that about 18% of the variation in PctBeliefExempt could be explained by these five variables.

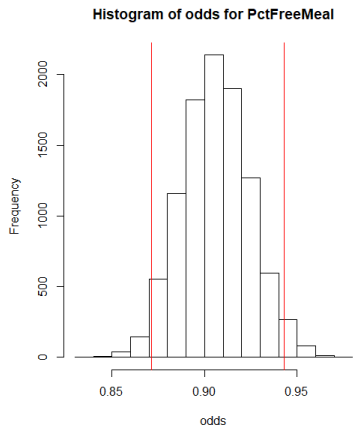
Drilling down into each coefficient (with similar Null and Alternative Hypotheses) shows that PctFreeMeal, Enrolled, and TotalSchools all have a significant effect at a=0.05. An additional VIF analysis found no concerning amount of multicollinearity in the model's coefficients.

A linear Bayesian Factor Analysis built with the same dependent and independent variables as the Frequentist model was found to have a high likelihood of being statistically significant at 2.52162e+25 to 1 odds vs. an intercept-only model. Each coefficient's 95% HDI was plotted to show the significance of

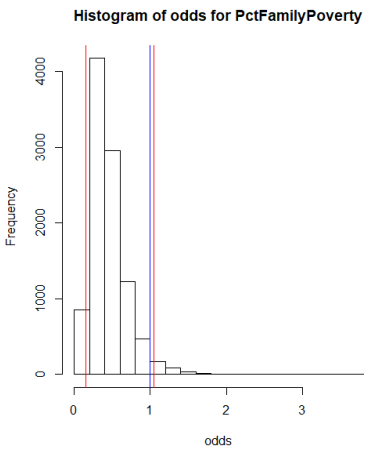
effect it had on the model. The blue lines denote 1:1 odds that imply no effect. If this blue bar is within the 95% HDI then the term is considered to have no appreciable effect at the 95% significance level.



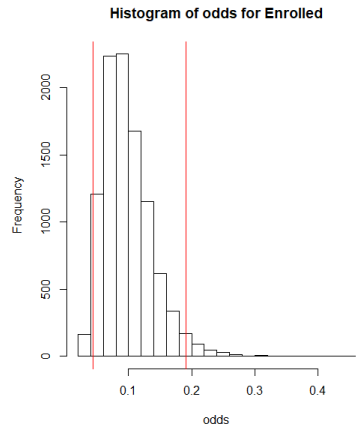
Plot 33. HDI Odds for PctChildPoverty Showing Insignificance



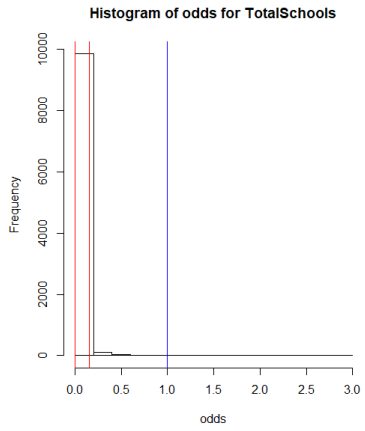
Plot 34. HDI Odds for PctFreeMeal Showing Significance



Plot 35. HDI Odds for PctFamilyPoverty Showing Insignificance



Plot 36. HDI Odds for Enrolled Showing Significance



Plot 37. HDI Odds for TotalSchools Showing Significance

From the analysis of the linear Bayesian model's coefficients it can be seen that TotalSchools, Enrolled, and PctFreeMeal all have an appreciable effect upon PctBeliefExempt. This agrees with the Frequentist findings, above.

To summarize the effects of these coefficients:

Vaccine belief exceptions increase as the number of students who are eligible to receive free meals decreases.

Vaccine belief exceptions increase as school size decreases.

Vaccine belief exceptions increase as district size increases.*

*This coefficient was inverted by the 1/x data normalization, but that inversion has been taken into account for this statement.

Conclusions

What's the big picture, based on all of the foregoing analyses?

To begin, much of the initial descriptive analysis focused on the state of the US as a whole, and reporting in public vs. private schools. To summarize, vaccine rates in the US are not fluctuating much, or trending in any particular direction by much. It cannot be said for certain that California is statistically behind in this arena, but with three out of four vaccines seeing lower-than-US-average adoption rates, California is not ahead of the game either. Private schools report much less of their vaccine information than public schools do, to a statistically significant amount.

Some other findings were present, and it is worth mentioning that reporting metrics may be tweaked to remove some of the less useful ones. For instance, it was found that as the student population of a school decreases, that school is more likely to have a completely up-to-date record. This is not surprising, as smaller schools are less likely to have fewer unvaccinated students simply due to the smaller sample sizes.

As seen from the district reporting model, the likelihood that a district's reporting was complete increases as the number of students eligible to receive a free meal decreases. This is not to say that students in need of a meal are themselves the issue, but it seems that conversely, schools with many students in need of assistance are less likely to have completed their reporting. Taken together with this model's other significant terms which state that a district is more likely to complete its reporting if it has more students, and if there are fewer schools in its district, imply a certain picture.

The picture painted by the results of this model show that schools with more students are better able to complete reporting. Perhaps this is because smaller schools do not have the resources to complete such a task. Supply the smaller schools with the resources to accomplish this reporting.

Furthermore, while none of the models scrutinized in this analysis were able to find significance with the direct metrics for poverty, there is one metric that indirectly implies poverty: students eligible to receive a free meal. Free meals, as it was found, being a highly correlated proxy for the other measurements of poverty. These students are not wealthy. But for the models in this analysis, the presence of more students in need of a meal corresponds to incomplete vaccine reporting. It could be that schools with many such students are themselves underfunded, or that it takes more money to account for these students who may have other issues reporting their vaccine history. These schools are likely in need of more resources to tackle the problem.

Finally, this model found a link between district sizes and reporting completeness. Smaller districts were better at completing their reporting. There may not be an easy way to tie this finding back to funding.

Additional research would be needed to see why this link exists. Perhaps smaller districts are more able to address vaccine reporting. It is not feasible to shrink district size simply to accommodate vaccine reporting, however.

Moving on from reporting to improving rates of vaccination, there are a few interesting observations. First among them are the findings in the question predicting the percentage of all enrolled students with completely up-to-date vaccines. The percentage of enrolled students with completely up-to-date vaccines increases as the number of students eligible to receive a free meal increases. This finding makes little sense, but could be interpreted in light of another finding: It was found that vaccine belief exceptions and students eligible to receive free meals are inversely related. It could be that the belief exceptions being less present in free-meal-eligible students, that by nature of their less-likely belief exceptions to vaccines, this is causing an increase in vaccine rates at schools where many such otherwise low-income students are.

In light of this, it is recommended that for the end-goal of increasing vaccination rates, some funding should be considered for outreach to schools and districts where fewer students are eligible for free meals. These may be more affluent neighborhoods. Perhaps education and outreach efforts from within schools and neighborhoods, in partnership with local health officials, faith leaders, and trend influencers could open a dialogue with parents and students who are currently objecting to vaccines on the grounds of a belief exception.

In a way that may highlight a geographic trend, it was found that vaccine belief exceptions increase as school size decreases, and increase as district sizes increases. It would be odd to suspect that the school size itself causes a belief exception in parents and students, but perhaps more rural schools in larger districts tend to a smaller size, and the populations of students with belief exceptions are more common in these areas. Additional research would be required in this area, but could also be a starting point to fund educational outreach related to vaccines in a geographic direction.

Appendix

Data Dictionary

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 familes 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

Complete R Code

```
# see external file data-and-code/Leland_Ball_Vaccine_Analysis.R
```