IST 772 Final Exam

Vaccinations and schools

Emilio Ramos Monzalvo

2021

**Part 1: Descriptive Reports**

**1. Vaccination Rates Over Time**

In order to combat infectious diseases, the US has kept track of multiple vaccines’ rates over time. The list of vaccination rates includes Tetanus (DTP1), Hepatitis B (HepB), Polio (Pol3), Influenza (Hib3), and Measles (MCV1) where each vaccine is showing unique a trend and volatility. The DPT1 vaccines showed a steady growth with a small drop in the early 1990’s, but it eventually plateaus in the 2000’s where it remained with a steady rate of more than 96% which makes it the vaccine with the highest rate. The HepB vaccine shows the biggest growth of all from about 18% to 60% which a significant state change in the early 2000’s, but even at 60% it is the vaccine with the smallest vaccination rate in the most recent records. The Pol3 vaccine had the highest volatility since the rate changed more than 30% four different times from 1980 to the present where it eventually converges to about 94% making it the second most used vaccine in the present. Hib3 is one of the most consistent vaccines with only a significant drop in the early 1990’s while it recovered after a few years and stayed constantly at about 90%. Finally, the MCV1 vaccine showed a relatively high volatility since it kept changing from more than 95% to less than 90% and vice versa about 5 times until it converged to just over 90% in the early 2000’s.

**Figure 1: US Vaccination Rates Over Time**

Shape

Description automatically generated with low confidence **Shape

Description automatically generated with low confidence**

**Figure 2: Most Recent Vaccination Rates**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **DTP1** | **HepB** | **Pol3** | **Hib3** | **MCV1** |
| 98% | 64% | 94% | 93% | 92% |

**2. School Vaccination Records**

In 2013, the state of California gathered vaccination records from kindergartens and reported whether the school complied with the request and whether the school is private or public. With that information we were able to summarize the results into a table as shown in Figure 3. Public Schools had a reporting rate of 5584 out of 5732 which equals to about 97.4% while Private Schools had a reporting rate of 1397 out of 1649 which equals to about 84.7%. This shows a higher reporting rate from Public Schools, but we performed a Chi-Squared test and its Bayesian equivalent to see if there is a statistically significant difference among the two school types.

**Figure 3: School Type vs Vaccination Records Provided**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Not Reported** | **Reported** | **Total** |
| **Private** | 252 | 1397 | 1649 |
| **Public** | 148 | 5584 | 5732 |
| **Total** | 400 | 6981 | 7381 |

**Figure 4: Chi-Squared Test Results**

Text

Description automatically generated

**Figure 5: Bayesian Difference in Proportions Results**

Graphical user interface, text, application

Description automatically generated Chart, histogram

Description automatically generated

In both tests, we found significant evidence to say the proportions of reporting rates are different with the Public Schools reporting rate being higher than Private Schools. The Bayesian Posterior analysis showed the difference in proportions to always be positive in the resulting 95% Highest Density Interval (HDI) which means Public Schools are more likely to report vaccination records. Similarly, the Bayesian Factor resulted in a value of 1.15e69 which means there is supports the posterior distribution analysis. Finally, the chi-squared test resulted in a p-value less than 0.05 which means we reject the null of the school types having the same proportions. Overall, private schools are showing a significant difference in vaccine reporting relative to public schools even if both types of schools have a reporting rate higher than 80%.

**3. Vaccination Rates in California Schools vs US Rates**

In order to check where California stands in vaccination rates, we can compare the school vaccination rates against the US rates. The most current rates are highlighted as red lines in Figure 7. For DTP1, the national rate is 98% while the California student rate is 92%.6 when we sum up the total number of students with the vaccine and divide by the total number of students. Similarly, we conducted the same calculations for the rest of the vaccines where the results are shown in Figure 6. The California student vaccination rate is more than 90% for all vaccines, but it is less than the National rate for DPT1 and Pol3. For Pol3, only 50% of school have a rate higher than the National rate while it is only 25% for DPT1. On the bright side, the vaccination rate for MCV1 and HepB are well above the national rate.

**Figure 6: California Overall Student Vaccination Rates vs US Rates**

|  |  |  |
| --- | --- | --- |
|  | **California Vaccination Rate** | **US Vaccination Rate** |
| **DPT1** | 92.6% | 98% |
| **Pol3** | 93.1% | 94% |
| **MCV1** | 92.9% | 92% |
| **HepB** | 95.2% | 64% |

**Figure 7: California Individal School Vaccination Rates vs US Rates**

**A picture containing text, receipt

Description automatically generatedText

Description automatically generated**

To check the overall school average against the US rates we ran a t-test and a Bayesian t-test to check for a difference in mean. In the frequentist t-test shown in Figure 8, we see that the school average rate was concluded to be below the national average for DTP1, Pol3, and MCV1 while it was much higher for HepB since the confidence intervals are below the national rates. On the other hand, the Bayesian T-Test showed a different result for Pol3 and MCV1 where Pol3 showed no difference from the National rate while MCV1 HDI was above the national rate. These two differences in results come from the frequentist version of the test being driven by outliers in the data which can be seen in Figure 7. Overall, the vaccination rate among schools is better for MCV1, and HepB while California still needs to work on the DTP1 vaccine to get it up to the national rate.

**Figure 8: Frequentist T-Test Results for School Average Vaccination Records**

**Table

Description automatically generated**

**Figure 9: Bayesian T-Test Results for School Average Vaccination Records**

**Table

Description automatically generated**

**4. Relationship Among Missing Vaccines**

In order to check if a vaccine is less popular than the rest, or if there is any commonality among students with at least one missing vaccine, we perform a correlation analysis among the percentage of missing vaccines per school. It is important to note that we do not have individual student records, so we are unable to tell whether an individual student without at least one vaccine does not have all vaccines or just a one or two depending on the number of students without a vaccine. Consequently, we have to perform a Pearson's product-moment correlation test and its Bayesian counterpart on schools without 100% up to date vaccination records for all of its students and schools that did report . In Figure 10 we can see how only 43 schools have a perfect record out of 700. In addition, 122 schools that do not have a perfect record and each of the students without at least one vaccine do not have any of the four vaccines.

**Figure 10: Districts Vaccination Records Summary**

|  |  |  |
| --- | --- | --- |
|  | **Not all Vaccinated** | **All Students Up To Date** |
| **All Individual Vaccine Missing do not equal Total Missing** | 535 | 0 |
| **All Individual Vaccine Missing equals Total Missing** | 122 | 43 |

**Figure 11: Person’s Correlation Test for Different Vaccine Rates**

Table, Excel

Description automatically generated

**Figure 12: Bayesian Correlation Test for Different Vaccine Rates**

**Table

Description automatically generated**

Based on the Frequentist test for Correlation and the Baysian version, we see that all of the missing vaccine rates are strongly correlated to one and other. The correlation between a school’s percentage of missing Polio, MCV1, and DTP1 show a strong positive correlation higher than .9. On the other hand, the HepB vaccine has a lower correlation to the other vaccines with a value between .8 and .9. This means that it is still strongly correlated to the other vaccines, but it is more likely to be missing just Polio, MCV1 and DTP1 than for a student to be missing all of the vaccines.

**Part 2: Predictive Analysis**

In order to find deeper understandings among districts and their vaccine reports and rates. The districts dataset includes five descriptive columns that will help us understand what kind of districts have high reporting rates, high vaccination rates, and belief exceptions. The five columns are Percent of Children in Poverty, Percent of Children receiving a Free Meal, Percent of Family Poverty, Total Number of Students Enrolled and the Total Number of Schools. The columns given can have strong correlations with each other, so we plotted them and computed the correlation matrix in order to understand the relationships prior to the modeling. Right off the bat, we can see how Child Poverty, Free Meal Percentage and Family Poverty are strongly correlated to each other while the Number of Enrolled Students and the number of schools in the district are almost 1:1. The dataset also includes one big outlier shown in Figure 13 in both the Total Schools and the Enrolled boxplots that will be taken out of the analysis since they can affect the true results.

**Figure 13: Boxplot of Predictive Variables**

**Shape

Description automatically generated with low confidence**

**Figure 14: Pair Plot of Predictive Variables**

**Shape

Description automatically generated with low confidence**

**Figure 15: Correlation Among Predictive Variables**

**Table

Description automatically generated**

**5. Predicting Whether a District Reported Vaccination Records**

In order to predict whether a district will provide a report to the state about its vaccination records, we identified the prediction column to be binary, so we used a generalized linear model (GLM) with a logistic link function, and its bayesian equivalent to find the important variables. It is also important to note that most of the districts did complete the reports as it is shown in Figure 16.

**Figure 16: Distribution of Completed Vaccination Reports by District**

**Shape

Description automatically generated with low confidence**

The Omnious Test results shown in Figure 17 show cleary how Percent Child Poverty, Percent Free Mean and Percent Family Poverty do not have an impact in whether a school reported their vaccination records or not since they did not pass the Ombnious Chi-Squared test nor the Wald z-test with p-values more than 0.05. On the other hand, the Number of Students Enrolled pass the ombnious test, but failed the Wald z-test. This could be an effect of its correlation to the number of schools feature. Its ratio is extremely close to one which means that its impact is minimal per student added, so we reran the model with a new feature called which gives the number of students enrolled divided by 100. When we looked at the results, we see a much bigger impact of 20% likeliness to report vaccination records for every 100 students added in the district. Finally, the number of schools in the district is the only variable that passed the Omnibus test, the Wald z-test, and has a significant impact on the likeliness a district will report their vaccination records with a value of 0.84. This means that for every school added to the district, the likeliness the district will report ther vaccination records down by 16%.

In the Bayesian equivalent, we got similar results where the 95% HDI shows the Percent Child Poverty, Percent Free Mean and Percent Family Poverty having no impact to the model since their interval overlaps with 1 while the Number of Schools and the Number Enrolled came out as significant. The only difference from the GLM is that the significant features had a bigger impact. The Number Enrolled had an impact closer to 23% per 100 students added while the Number of Schools in the district had an impact of 18%.

|  |  |
| --- | --- |
| **Constant** | **Value** |
| Intercept | 52.28 |
| Percent Child Poverty | 1.03 |
| Percent Free Meal | 0.99 |
| Percent Family Poverty | 0.94 |
| Number Enrolled  Number Enrolled/100 | 1  1.2 |
| Total Schools | 0.84 |

**Figure 17: Generalized Linear Model Output**

**Table

Description automatically generated**

**Figure 18: Generalized Linear Model Omnibus Test Output**

Table

Description automatically generated

**Figure 19: Bayesian Generalized Linear Model Output**

|  |  |
| --- | --- |
| **Constant** | **Value** |
| Intercept | 55.32 |
| Percent Child Poverty | 1.03 |
| Percent Free Meal | 0.99 |
| Percent Family Poverty | 0.94 |
| Number Enrolled  Number Enrolled/100 | 1  1.23 |
| Total Schools | 0.82 |

A picture containing clock, dark

Description automatically generated

**6. Predicting Up to Date Vaccination Record Percentage**

In regards to the percentage of students with Up to Date Vaccination Records per district, we will continue using the same variables, but we will be training a regular linear model to predict the continuous variable. This means we will be doing the same for the Bayesian equivalent of the linear model. It is also important to note that the percentage is most likely high even if there are some outliers below the 60% mark as shown in Figure 21.

**Figure 21: Distribution of Percentage of Fully Vaccinated Students Per District**

Shape

Description automatically generated with low confidence

The linear model resulted in an F-statistic of 18.52 and a p-value less than 0.05 which means we can reject the null hypothesis which says the R-squared value is equal to 0. I.e, the model is not randomly predicting the percentage of fully vaccinated students. On the other hand, the reported adjusted R-squared value is less than 0.01 which means the model is not predicting it accurately. When we review the t-values of the individual predictors, we find that almost all predictors except the Percentage of Child Poverty comes out as significant. This can be due to its correlation to the Percent of Free Meals column which the linear model ignores.

The Bayesian Inference on Coeffificients and R-squared gave back similar results where the R-squared is close to 0.11 and the only insignificant predictor is the Percent of Child Poverty. The Bayesian Factor analysis did return a significant value, but this only furthers the assumption that this model is better than a intercept only model.

Overall, the predictability of the Percent of Fully Vaccinated students per district exists, but it is very weak to make any assumptions on the predictors other than the positive or negative controbutions by each variable. Therefore, we see that the Percent of Students with Free Meals, the Number Enrolled and the Percent of Family Poverty predictors have a positive contribution while the Number of Schools has a negative contribution.

**Figure 22: Linear Model Output**

|  |  |
| --- | --- |
| **Constant** | **Value** |
| Intercept | 81.6 |
| Percent Child Poverty | -0.11 |
| Percent Free Meal | 0.09 |
| Percent Family Poverty | 0.3 |
| Number Enrolled  Number Enrolled/100 | 0.01  1.018 |
| Total Schools | -0.72 |

**Table

Description automatically generated**

**Figure 23: Bayesian Linear Model Bayesian Factor Output**

|  |  |
| --- | --- |
| **Constant** | **Value** |
| Intercept | 87.9 |
| Percent Child Poverty | -0.10 |
| Percent Free Meal | 0.08 |
| Percent Family Poverty | 0.29 |
| Number Enrolled  Number Enrolled/100 | 0.01  0.99 |
| Total Schools | -0.70 |

A picture containing clock, dark

Description automatically generated

A picture containing graphical user interface

Description automatically generated

**7. Predicting the Percentage of Students with Belief Exceptions**

Similar to the prediction of the percent of students with up to date vaccination records, we do the same analysis but with the percent of students with belief exceptions as the prediction variable. Now, the percentage of students with belief exceptions is most likely below 20% even if there are some outliers past the 19% mark. This means that most belief exceptions are not related to the type of district it is since only a few show a high percentage.

**Figure 24: Distribution of Percentage of Students with Belief Exceptions Per District**

Shape

Description automatically generated with low confidence

The linear model resulted in an F-statistic of 21.08 and a p-value less than 0.05 which means we can reject the null hypothesis which says the R-squared value is equal to 0. I.e, the model is not randomly predicting the percentage of fully vaccinated students similar to the model from part 6. The reported adjusted R-squared value is about 0.13 which means the linear model is not following well the percentage of students with belief exceptions. When we review the t-values of the individual predictors, we find that almost all predictors come out as significant with a corresponding p-value less than 0.05. On the other hand, the estimates coefficient values are close to 0 which means each variable has a small impact on the percentage of students with belief exceptions.

The Bayesian Inference on Coeffificients and R-squared gave back similar results where the R-squared is close to 0.13 and all the 95% HDI’s of the predictors do not overlap with 0. The Bayesian Factor analysis did return a significant value, but this only furthers the assumption that this model is better than a intercept only model.

Overall, the predictability of the Percent of Students with Belief Exceptions per district exists, but it is very weak with the current predictors to make any assumptions on the predictors other than the positive or negative controbutions by each variable. Therefore, we see that the Percent of Students with Free Meals, the Number Enrolled and the Percent of Family Poverty, and the Number of Students Enrolled predictors have a negative contribution while the Number of Schools and the Percent of Child Poverty has a positive contribution to the percent of Students with Belief Exceptions.

**Figure 25: Linear Model Output**

|  |  |
| --- | --- |
| **Constant** | **Value** |
| Intercept | 10.4 |
| Percent Child Poverty | 0.16 |
| Percent Free Meal | -0.12 |
| Percent Family Poverty | -0.22 |
| Number Enrolled  Number Enrolled/100 | -0.01  -0.53 |
| Total Schools | 0.33 |

**Table

Description automatically generated**

**Figure 26: Bayesian Linear Model Bayesian Factor Output**

|  |  |
| --- | --- |
| **Constant** | **Value** |
| Intercept | 5.7 |
| Percent Child Poverty | 0.16 |
| Percent Free Meal | -0.1 |
| Percent Family Poverty | -0.21 |
| Number Enrolled  Number Enrolled/100 | -0.01  -0.53 |
| Total Schools | 0.32 |

A picture containing clock, dark, cloudy

Description automatically generated

**Text

Description automatically generated with medium confidence**

**8. Big Picture**

In order to increase vaccination rates and their reporting compliance in school districts, we analyzed multiple historical records to try and correlate vaccination rates to multiple descriptive variables of the districts. Overall, we have found that the descriptive features of Child Poverty Percentage, Family Poverty Percentage, Percentage of Students Elegible for Free Meals, Number of Students Enrolled in School and the Number of Schools do not accurately explain vaccination rates and the belief-based vaccine exceptions. This means that the vaccination rates cannot be explained as a matter of poverty in the district nor student count in a district. Other descriptive variables outside of this study might be a bigger driving factor in the vaccination rate of a district.

In terms of reporting compliance by the district, the team was able to find a strong predictability based on the Number of Schools and the Number of Students Enrolled in the District. Ironically, the two have an opposing effect on the reporting compliance by the district. If the number of schools increases by 1, then the district is 16% less likely to comply. On the other hand, if the number of students increases by 100, then a district is 20% more likely to comply. This means that adding schools to a district makes it more difficult to comply with the vaccination reports while having more students makes the schools more likely to comply. This could be an effect of the education budget of a district increasing as the number of students increase. Consequently, adding more resources to districts with small education budgets or small head count per the number of schools will help increases the chances that the district will comply with the vaccination reports.