IST772 – Quantitative Reasoning for Data Science - Final Report

Mahesh Kumar Badam Venkata

Introduction:

Using the vaccine data of U.S. as a whole and Californian school districts, a descriptive and inferential analysis was performed and the results of which are documented and explained in a statistical reasoning approach. Below points are being highlighted as key points as part of the analysis.

Descriptive Overview of U.S. Vaccinations:

1. In the recent years, from 2000 to 2010, except for Hepatitis B, Birth Dose which had a spike in around 2003, the vaccination rates are almost stable in the range of above 90.
2. Again, in the recent years, in attempt to decompose the time series plot of vaccination rates, it is observed that a slowly increasing and stabilizing trend along with a triangular seasonality in the plot.
3. The mean U.S. vaccination rates are 97.76%, 43.96%, 91.16%, 92.16%, 91.08% for First dose of Diphtheria/Pertussis/Tetanus vaccine, Hepatitis B (Birth Dose), Polio third dose, Influenza third dose, Measles first dose respectively for the recent years.

Descriptive Overview of California Vaccinations:

4. The mean levels of the variables WithoutDTP, WithoutPolio, WithoutMMR and WithoutHepB that show the percentage of students without DTP, Polio, MMR, and Hepatitis B are 10, 9.53, 9.98, and 7.54 respectively.
5. From the correlation table below of vaccination rates in Californian districts, it is observed that row represents the missing vaccine likelihood of a student when the student missed another one already shown in the column. The value represents the degree out of 1 which is at the diagonal missing vaccine likelihood with itself.

	WithoutDTP	WithoutPolio	WithoutMMR	WithoutHepB
WithoutDTP	1.00	0.98	0.97	0.89
WithoutPolio	0.98	1.00	0.96	0.90
WithoutMMR	0.97	0.96	1.00	0.89
WithoutHepB	0.89	0.90	0.89	1.00

6. Comparing Californian vaccine rates to the US as a whole, the districts are also the same range as above 90% except for Hepatitis B which is significantly low for the country than California state.

Overall, it can be stated that California is at a very good standing having near the country level vaccination rates except for Hepatitis B. Nonetheless, the Hepatitis B vaccination rate has only been increasing in the recent years, California is on top contributing to an increase in the vaccination rates.

Inferential Reporting:

7. Out of the four predictors, 'PctChildPoverty', 'PctFamilyPoverty', 'Enrolled', and 'TotalSchools' to predict percentage of enrolled students with belief exceptions, firstly, 'Enrolled' and 'TotalSchools' variables are log transformed and the logarithm of those variables are used to predict the percentage. Of which, log of Enrolled, Log of

TotalSchools, PctFamilyPoverty variables have the statistical significance which is deduced based on the Pr value less than 0.05 and the variable PctChildPoverty stated above have Pr value greater than 0.05 and it doesn't hold true for the Hypothesis test in favor of alternative hypothesis: predictors can be used to predict dependent variable. Also, it is backed by Bayes Factor of very high value and the results of Bayesian analysis, the High-Density Interval has the values of linear model. Pr value means that the probability of the coefficient of the variables out of the range of the obtained coefficient which needs to be less than 0.05 for having statistical significance.

8. Similarly, to predict percentage of all enrolled students with completely up-to-date vaccines, logEnrolled, PctFamilyPoverty, and Log of TotalSchools have statistical significance with Pr value less than 0.05. Again, Bayesian analysis and Bayesian Factor support the results of hypothesis test.
9. The results of attempts to use any set of variables to predict percentage of enrolled students with completely up-to-date vaccines, WithoutDTP, WithoutMMR, WithoutHepB, WithoutPolio, PctBeliefExempt provide the top statistical significance or in our case we term it as R-squared which is highest, 0.944 and all the independent values have the Pr value far less than 0.05, important for modelling.
10. There is an interaction between PctChildPoverty and Enrolled, it is understood when using just the two variables and comparing it with the model with includes the interaction term of both variables. The latter model outperformed the prior model, and it is again supported by very high Bayesian Factor, which means there is a greater evidence for the alternative hypothesis that the variables with interaction term has higher predictability than the ones without. Below is the figure for reference.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	81.0315700	1.1001842	73.653	< 2e-16	***
PctChildPoverty	0.2792997	0.0409679	6.818	2.01e-11	***
Enrolled	0.0049923	0.0012641	3.949	8.63e-05	***
PctChildPoverty:Enrolled	-0.0001570	0.0000416	-3.773	0.000175	***

11. Out of the four variables mentioned above, the Enrolled and the TotalSchools are to be log transformed to be used for prediction. So, the PctFamilyPoverty, logEnrolled, logTotSchools, are three variables that can used to predict whether or not a district's reporting was complete. This statement is through using Pr value less than 0.05 for confirmation of predictability and using Bayesian Factor and High-Density Interval in Bayesian analysis to support our alternate hypothesis and the results of Hypothesis Test.

Conclusion:

At a higher level, the best predictor variables for prediction of percentages of enrolled students with good statistical significance are obtained. They can be used to predict the enrollments. To improve vaccination reporting rates, it is important to identify the districts that are not completely reported, and it can be predicted again using the variables above. To improve vaccination rates, it is important to improve the individual vaccination rates and percent of belief exceptions, financial assistance would be needed where there is low vaccination rate and low percent of belied exceptions.