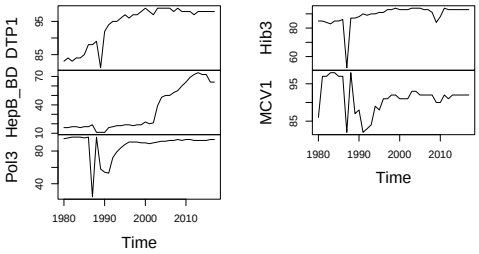
U.S.A. Vaccination Report

Introduction

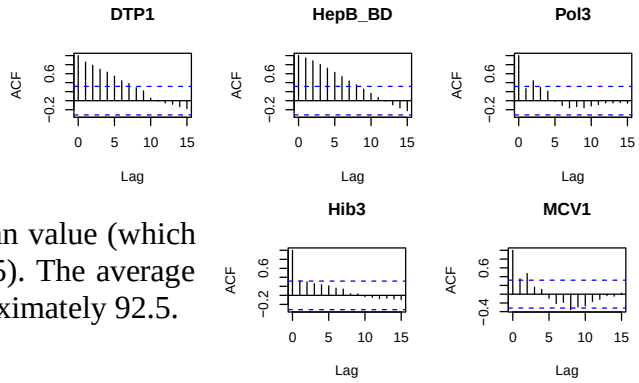
As the most cost-effective way to prevent people from contracting diseases, the California state legislator is seeking measures to improve the vaccination coverage rates as well as the reporting rates. Therefore, this report aims to analyze the two datasets with descriptive details, to find factors that affect the overall vaccination rates and reporting rates, and to provide insightful opinions after comprehensive data analysis.

Descriptive Overview of US Vaccination



According to the time series plot which shows the percentage of vaccine rates from 1980 to 2017, the number of HepB_BD and DTP1 increased significantly and maintained at a high level by the end of 2017. The vaccination rate of Hib3 has a relatively small increase while that of Pol3, which started at a fairly high number, decrease sharply, and bounced back to the original level. On the contrary, the vaccine rate of MCV1 was volatile, fluctuating around 90, and has stayed stable since 1995.

A notable trend could be detected in DTP1's and HepB_BD since a gradual decrease could be observed in their responding of ACF plot. The majority of lines for the rest ACF plots are within the significant range, indicating that the original time series plot is stationary. As for the mean level of vaccine rates for recent years (2010 to 2017), DTP1 has the highest mean value (which is 97.875) while HepB_BD has the lowest one (68.875). The average value of the rest is similar to each other, which is approximately 92.5.



Descriptive Overview of California Vaccination

The mean levels of four vaccine rates across districts in California are similar. DTP, Polio, and MMR have a similar number (which is approx. 90). HepB has the highest average, which is 92. Additionally, there was a significant positive correlation among the four variables, implying students who did not take one specific vaccine tended to reject the other three. Comparing the vaccination levels in the California and the U.S overall, California government successfully sustained a high vaccine rate for all those four types of vaccines in 2017. Most of the vaccine rates in California are at the same level as those of vaccines in the U.S. The average value of HepB is higher than the national HepB_BD immunization rate while that of DTP in California, however, is slightly lower than that of DTP1.

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

Regression Model Evaluation

The correlation result shows that PctUpToDate and PctBeliefExempt has a relatively stronger correlation with PctFamilyPoverty (which is 0.25 and -0.24 respectively) while DistrictComplete has a relatively stronger correlation with TotalSchools (which is -0.2). However, one needs to use the log-transformed predictors to improve normality due to the skewness issues.

Firstly, four linear regression models with individual predictors have been run respectively to predict PctBeliefExempt. Each predictor is statistically significant and has a negative relationship with PctBeliefExempt. Having the highest R-square value, logEnrolled could explain 7.672% of the variance of PctBeliefExempt, which is unsatisfactory. However, to find out the collective relationship between predictors and PctBeliefExempt, it is necessary to evaluate the four predictors as a whole after data scaling.

Table 1 Summary of the percentage of all the enrolled students with belief expectations

	Estimate	Pr(>|t|)	Multiple R-square	p-value	95% HDI
logTotSchools	-1.5872	**5.33e-08**	0.04154	5.334e-08	-2.1223 to -0.994

logEnrolled	-1.5561	**8.53e-14**	0.07672	8.535e-14	-1.9432 to -1.144
logPctChildPoverty	-2.841	**4.15e-07**	0.03607	4.154e-07	-3.8622 to -1.739
logPctFamilyPoverty	-2.3115	**3.07e-08**	0.04369	3.066e-08	-3.8904 to -1.691

Next, a multiple regression model with four predictors has been run to predict PctUpToDate. Two out of the four are statistically significant. The R-square value of 0.1249 suggests that 12.49% of the variance in PctUpToDate was accounted for by those four predictors collectively, which is still unsatisfactory.

Table 2 Summary of the percentage of all enrolled students with completely up to date vaccines

	Estimate	Pr(>|t|)	Multiple R-square	p-value
logTotSchools	-2.9305	**0.00197**	0.1249	2.2e-16
logEnrolled	4.0118	**9.46e-09**		
logPctChildPoverty	2.1221	0.13499		
logPctFamilyPoverty	1.5064	0.18246		

Additionally, having the combination in table 3 could yield the highest significant multiple R-square (0.1355). All those predictors are statistically significant. logEnrolled and logPctChildPoverty are positively associated with PctUpToDate while logTotSchools is negatively associated with PctUpToDate.

Table 3 Best set of predictors for predicting the percentage of all enrolled students with completely up-to-date vaccines

	Estimate	Pr(>|t|)	Multiple R-square	p-value
logTotSchools	-3.3010	**1.90e-10**	0.1355	< 2.2e-16
logEnrolled	4.4765	**0.000619**		
logPctChildPoverty	4.1777	**2.75e-08**		

To validate the interaction between PctChildPoverty and Enrolled, the p-value for Delta-R-square provides an evidence. The result of Bayes Factor shows the odd of 98.58:1 in favor of the model that includes the interaction, which is consistent with the results of model compare analysis.

Table 4 Interaction between Enrolled and Percentage of children in district living below the poverty line

	Estimate	Pr(>|t|)	R-square	p-value	p-value for Delta-R-square	Bayes Factor	95% HDI
PctChildPoverty	0.178262	**8.82e-06**	0.05842	9.623e-10	0.00024825	98.57982 ±0.01%	1.469 to 3.057
Enrolled	0.001632	**8.13e-05**					1.654 to 2.784
PctChildPoverty:Enrolled	-0.000164	**0.000248**					2.42812 to 5.471

Lastly, logistic regression is conducted to predict the completeness of a districts reporting. The p-values associated with logEnrolled and logTotSchools indicates that they are statistically significant. Besides, logEnrolled and logTotSchools may be correlated. Although a district with a high logTotSchools tends to have a lower probability of reporting rate, the truth that district with a high logTotSchools tends to have a high logEnrolled, which in turn associated with a higher probability of complete reporting rate. However, without further verification, one cannot assert that assumption is valid.

Table 5 Logistic regression of predicting whether a district's reporting is complete

	Estimate	exp(Estimate)	Pr(>|t|)	exp(95% C.I.)
logTotSchools	-3.1084	0.04467319	**2.80e-09**	0.01505 to 0.11816
logEnrolled	1.8229	6.19007942	**2.38e-07**	3.17372 to 12.76800
logPctChildPoverty	0.3180	1.37436513	0.633	0.36343 to 4.92042
logPctFamilyPoverty	-0.7373	0.47841109	0.182	0.15850 to 1.37487

Conclusion

In terms of vaccination rates, compared with the U.S. overall, California state government had an outstanding average value for each of the four vaccines in 2017. Especially, the percentage of HepB was notably higher than the nationwide HepB_BD vaccine rate for recent years. Nonetheless, there is a space for improvement. To increase the vaccination rates, one might wish to minimize the PctBeliefExempt and maximum the PctUpToDate by having a larger number of enrolled students and a lower number of schools in a district. Similarly, to improve the reporting rates, the legislator might need to decrease the number of schools or increase the number of enrolled students in a district.

This report failed to use variance inflation factor (vif) analysis for multicollinearity detection. If any predictor has a vif value that is above 5, that predictor might have the issue of multicollinearity.