IST 686

Final Examination Report – Ximeng(Simone) Deng

Spring 2024

Prepare your final exam report by answering the questions in this document. In your answer for each question, make sure you write a narrative with complete sentences. Keep your answers concise and focused on the question asked, but provide sufficient numeric and graphical detail that the staff member can create a comprehensive briefing for a legislator (see question 7 for specific points of interest). Make sure to include enough statistical information so that another analytics professional could review your work. You can assume that the staff member understands the concept of statistical significance and other basic concepts like mean, standard deviation, and correlation, so you do not need to define those. Be sure that you not only report what a test result was, but also what that result means substantively for the question you are answering.

You can choose to put important statistical values into a table for readability, or you can include the statistics within your narrative. Your report can include graphics created by R, keeping in mind that if you do include a graphic, you will have to provide some accompanying narrative text to explain what it is doing in your report. You can delete this instruction page and the prompts in italics, but keep the section headings. Finally, be sure to proofread your final submission to eliminate typos, check grammar and to ensure that everything is included and readable.

I will assume that the statistical results in your answers refer to the corresponding section of your analysis notebook. If you decide you need to rerun an analysis, you can do so, but you will only receive credit for interpreting the new results. If you do rerun an analysis, please be sure to reference it and upload a knitted PDF with the analysis.

You may use an LLM such as ChatGPT to help prepare your answers. If you do so, you must disclose specifically how you used it in the final section of the report. You **may not** seek nor receive assistance, help, coaching, guidance, or support from any human except your instructor at any point during this exam. Seeking or obtaining improper assistance will result in a 0 for this exam. Your instructor will be available by email throughout the report writing period if you have questions, but don't wait until the last minute!

Your responses will be graded on clarity; conciseness; inclusion and explanation of specific and appropriate statistical values; explanation of any included tabular material and the appropriate use of graphical displays when/if necessary. Bonus points will be awarded for work that goes above expectations.

1. Introductory paragraph

In your own words, write about three sentences of introduction addressing the staff member in the state legislator's office. Frame the problem/topic that your report addresses.

This report addresses the issue of vaccination rates and reporting compliance across California' elementary and unified schools. Based on data from the WHO on U.S. vaccination rates for five common vaccines, along with data on 7,381 California public and private schools, and a sample of 700 California public school districts from the 2017 data collection, we aim to compare Californian vaccination levels to U.S. averages, explore variations among counties, and provide actionable insights that can guide the allocation of resources and interventions to improve public health outcomes in schools throughout the state.

Descriptive Reporting

2. Descriptive Overview of U.S. Vaccinations

a. How have U.S. vaccination rates varied over time? [See notebook 2.a.]

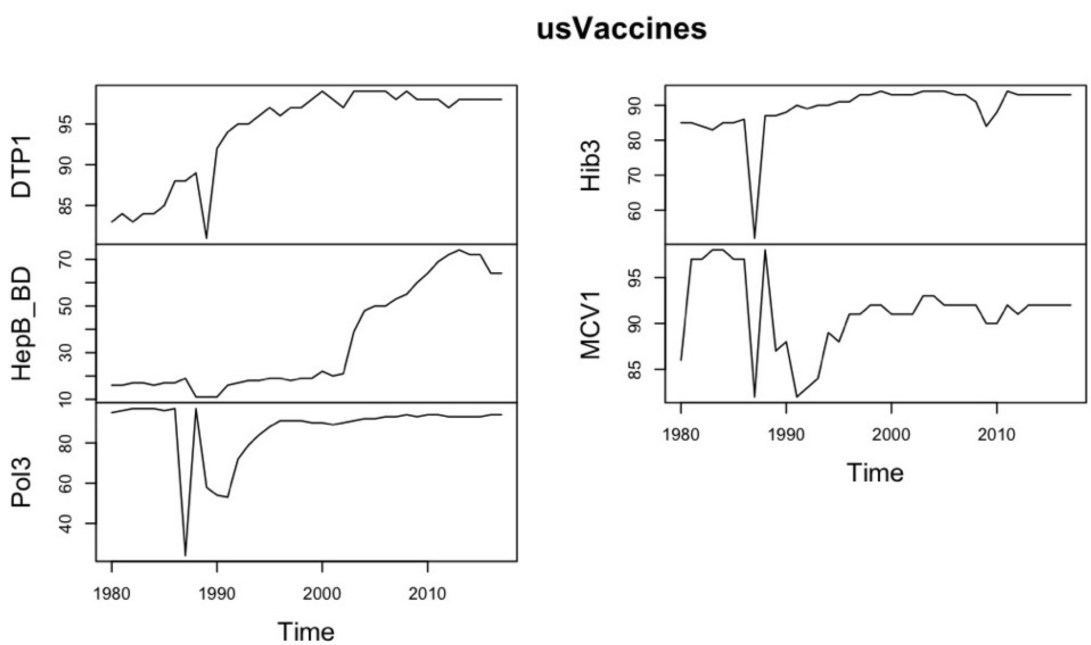

Figure 1

DTP1	HepB_BD	Pol3	Hib3	MCV1
Min. :81.00	Min. :11.00	Min. :24.00	Min. :52.00	Min. :82.00
1st Qu.:89.75	1st Qu.:17.00	1st Qu.:90.00	1st Qu.:87.00	1st Qu.:90.00
Median :97.00	Median :19.00	Median :93.00	Median :91.00	Median :92.00
Mean :94.05	Mean :34.21	Mean :87.16	Mean :89.21	Mean :91.24
3rd Qu.:98.00	3rd Qu.:54.50	3rd Qu.:94.00	3rd Qu.:93.00	3rd Qu.:92.00
Max. :99.00	Max. :74.00	Max. :97.00	Max. :94.00	Max. :98.00

Figure 2

Based on the time series data from the World Health Organization reporting vaccination rates in the U.S. for five common vaccines from 1980 to 2017, I plotted the data (Figure 1) and examined their structure with a summary of descriptive statistics (Figure 2).

The U.S. vaccination rates for 5 common vaccines generally show an increasing trend or stabilization at high coverage levels from 1980 to 2017. However, we can also see a huge drop during 1987 to 1991 for all the 5 vaccines. DTP1 had a significant drop in 1989. HepB_BD had a decline from 1988 to 1990. Pol3 showed a huge decrease in 1987, a rebound in 1988, followed by a moderate decline from 1989 to 1991, and then a gradual increase. Hib3 had a significant drop in 1987. MCV1 dropped in 1987, recovered in 1988, dropped again from 1990 to 1993, and then slowly rose.

A pure polysaccharide vaccine was licensed for use in the United States in 1985 and remained in use until 1988. The first Hib conjugate vaccine was licensed in 1987. https://www.cdc.gov/vaccines/pubs/pinkbook/hib.html It is somewhat confusing to have data from 1980 to 1985 (and 1987) given these licensing dates. For HepB_BD, Pol3, and MCV1, their coverage before 1991 (except for the sudden drop) was higher than the coverage after the sudden drop and subsequent rebound. I have not found a comprehensive historical reference that explains these phenomena, an article from The Washington Post in 1991 partly attributes the reason to a cessation in data collection.

https://www.washingtonpost.com/archive/politics/1991/10/09/us-immunization-rates-uncertain/554e1354-781e-45ed-9df9-307e7a12575d/

HepB_BD was introduced later as compared to other vaccines, with initial data around the 15% in the early years and a significant increase over time. There is a sharp rise in the early 2000s, reaching a peak near 74%, and then stabilizing around 60% range towards the end of the data period.

b. Are there significant trends or cyclical variation in U.S. vaccination rates? [See notebook 2.b.]

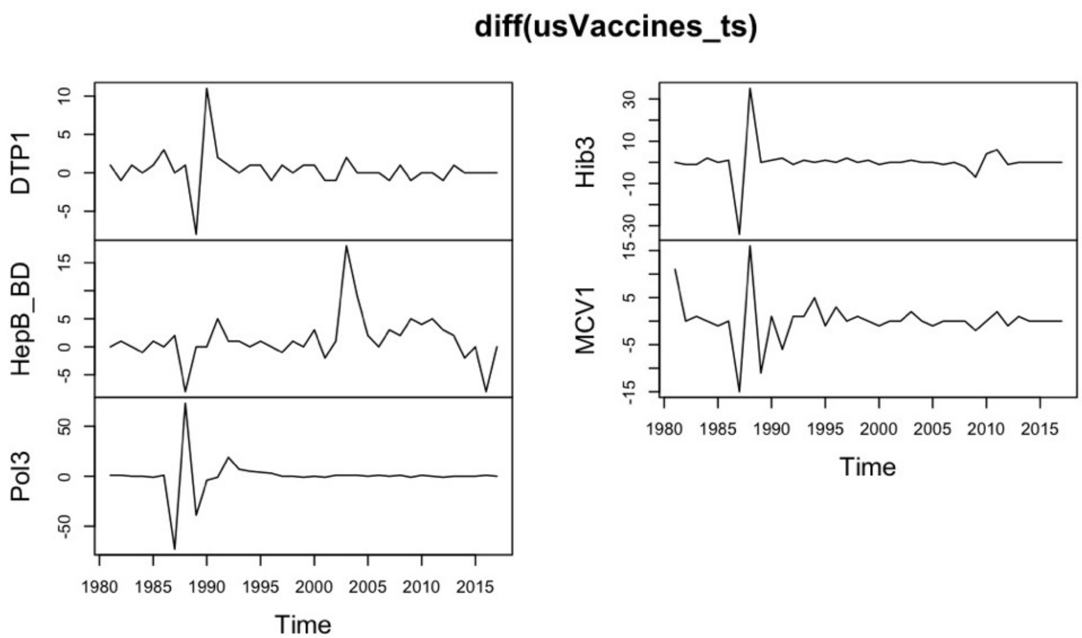

Figure 3

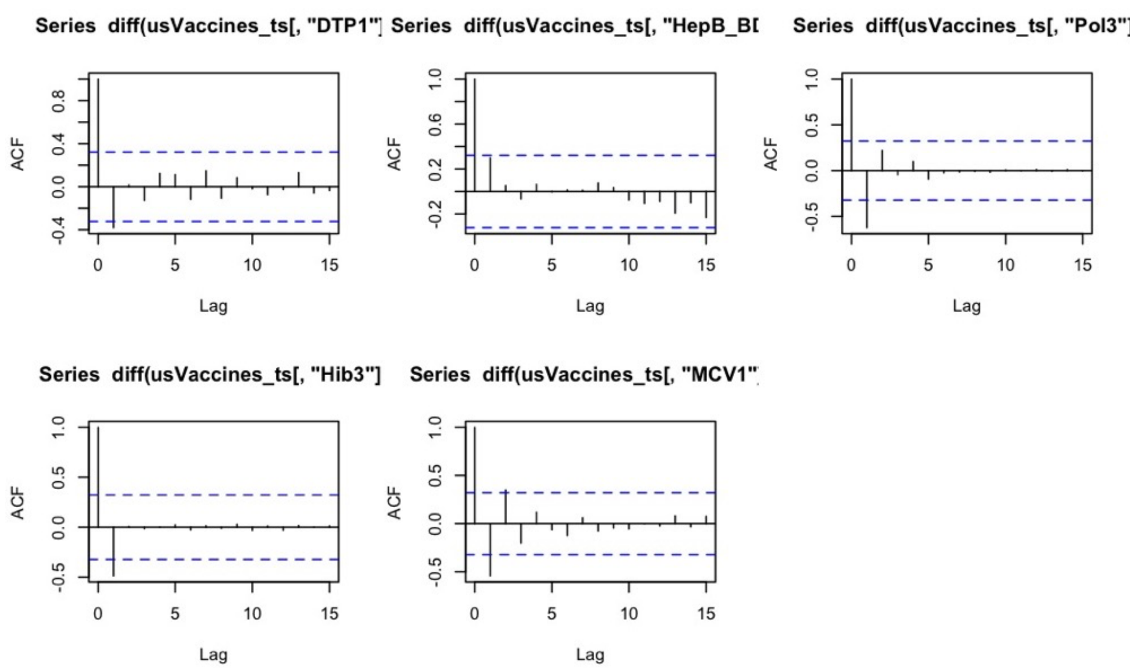

Figure 4

I applied differencing to the 5 vaccines, creating first-lag differences to remove any trends occurring over time and to demonstrate how the result is a stationary series (Figure 3). Then, I tested the stationarity of each series using the Augmented Dickey-Fuller (ADF) test, adf.test(). The alternative hypothesis for this test is that the process is stationary.

For DTP1, Dickey-Fuller = -4.5333, $p < .01$. We reject the null hypothesis, suggesting that the differenced series is stationary. For HepB_BD, Dickey-Fuller = -1.958, p-value = 0.5896. We fail to reject the null hypothesis, indicating that the differenced series is not stationary. Pol3, Dickey-Fuller = -2.9361, p-value = 0.2082, also suggesting non-stationarity. For Hib3, Dickey-Fuller = -4.3152, $p < .01$. We reject the null hypothesis, indicating stationarity. MCV1, Dickey-Fuller = -3.3982, p-value = 0.07305. We fail to reject the null hypothesis, suggesting that the series is not stationary.

However, I examined the auto-correlation function, acf() using the differenced data (Figure 4), which showed no evidence of cyclical variations.

In addition, I analyzed the periodograms for the differenced data. The results showed that there is some cyclical variation in DTP1, Pol3, Hib3, MCV1 roughly every 2.1 to 2.2 years, while no shorter cyclical variation for HepB_BD.

In short, though observing fluctuations in vaccination rates, we cannot conclude that there are significant trends or cyclical variation in U.S. vaccination rates.

c. What are the mean U.S. vaccination rates when including only recent years in the calculation of the mean? [See notebook 2.c.]

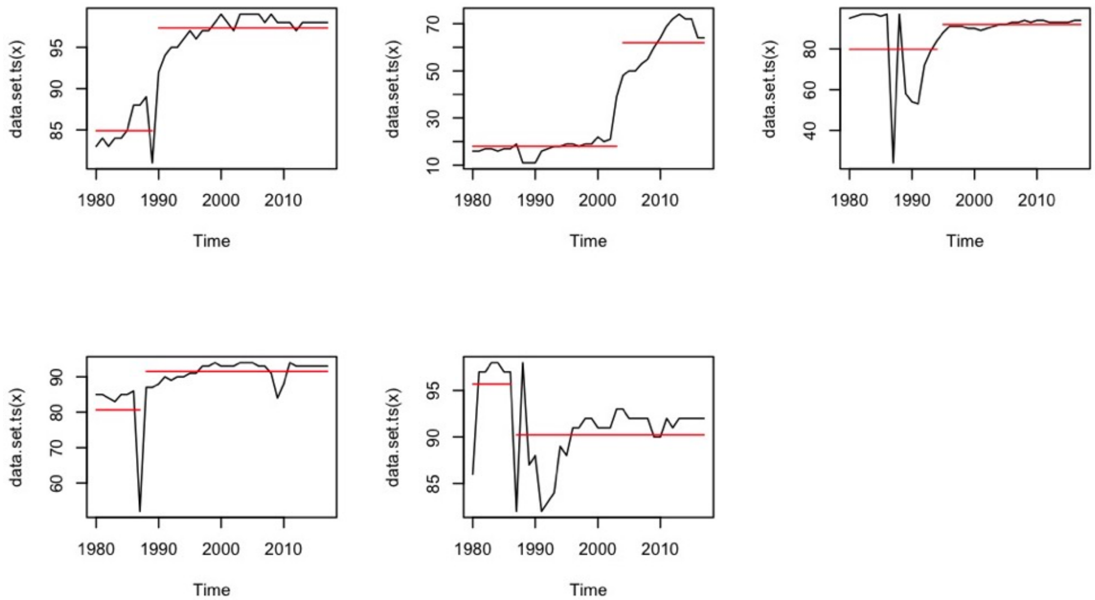

Figure 5

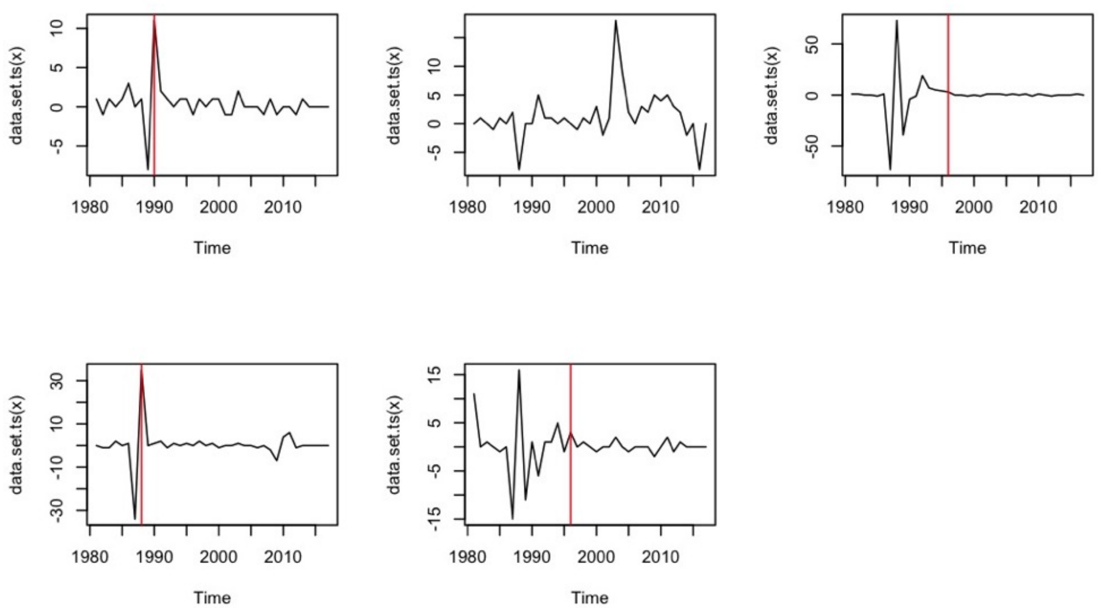

Figure 6

I used the cpt.mean and cpt.var functions from the changepoint package to identify changes in the mean and variance(on the differenced data) in the time series data for each vaccine. This method helped me to decide the most recent period during which vaccination rates have been stable, in terms of both mean and variance (Figure 5 and Figure 6).

DTP1 has a mean change point in 1900 and a variance change point in 1990. HepB_BD has a mean change point in 2004; however, no variance change points are showing even with the PELT method. Pol3 has a mean change point in 1995 and a variance change point in 1996. Hib3 has both a mean and a variance change point in 1988. MCV1 has a mean change point in 1987 and a variance change point in 1996.

I chose to use the period from 2004 to 2017 as the recent years based on the latest mean change point for HepB_BD in 2004. I also checked that the post-2004 data for all vaccines appear stable in both mean and variance.

The mean U.S. vaccination rates for the recent years are as follows: DTP1 at 98.21429, HepB_BD at 61.92857, Pol3 at 93.21429, Hib3 at 92.07143, and MCV1 at 91.71429. The overall mean is 87.42857.

3. Descriptive Overview of California Vaccinations

a. What are the mean levels of the four vaccination rate variables across districts? [See notebook 3.a.]

The mean vaccination rates across districts in California are as follows: DTP at 89.79571, Polio at 90.20429, MMR at 89.78714, and HepB at 92.26286.

b. Among districts, how are the vaccination rates for individual vaccines related? In other words, if there are students with one vaccine, are students likely to have all the others? [See notebook 3.b.]

The correlation analysis of the vaccination data across California school districts reveals strong positive relationships between individual vaccine coverage and the overall percentage of students with completely up-to-date vaccines (figure 7). The correlations between HepB and other vaccines, though slightly lower, were still strong at above 0.84.

	WithDTP	WithPolio	WithMMR	WithHepB	PctUpToDate
WithDTP	1.0000000	0.9816446	0.9768546	0.8907265	0.9588973
WithPolio	0.9816446	1.0000000	0.9667621	0.9055103	0.9402755
WithMMR	0.9768546	0.9667621	1.0000000	0.8897889	0.9671549
WithHepB	0.8907265	0.9055103	0.8897889	1.0000000	0.8433239
PctUpToDate	0.9588973	0.9402755	0.9671549	0.8433239	1.0000000

Figure 7

The All Schools dataset indicates that there are 847 unique public school districts, while our districts dataset contains 700 observations, representing a sample of California public school districts. Therefore, we require inferential reasoning about correlation.

I conducted the cor.test on each vaccine with PctUpToDate. For withDTP and the PctUpToDate, $t(698) = 89.281$, $p < .01$. We reject the null hypothesis of rho = 0. The 95% confidence interval for rho, around the point estimate of $r = 0.9589$, ranged from 0.9525 to 0.9645. If we repeated this sampling process many times and each time constructed a confidence interval around the calculated value of r, about 95% of those constructed intervals would contain the true population value, rho, though we don't know for sure if any particular one does. However, the confidence interval does not straddle 0. This evidence suggests that the correlation between "withDTP" and "PctUpToDate" is more than could be expected by chance, we have a sense of certainty that the correlation is positive. There is a < 2.2e-16 chance of observing an absolute value of t this high or higher under the assumption that the population value of rho = 0.

I also performed Bayesian tests on the correlation coefficient. The mean correlation in the posterior distribution of rho was 0.9587. The 95% HDI ranges from 0.9375 to 0.9797, which doesn't span 0. The Bayes factor of 2.300632e+379:1 strongly in favor of the alternative hypothesis, providing overwhelmingly evidence that the population correlation, rho, between "withDTP" and "PctUpToDate", is not 0. We have the credibility that the population correlation is a positive value lying somewhere in the range of 0.9375 to 0.9797, and close to a central value of 0.9587.

Similarly, For withHepB and the PctUpToDate, $t\ (698) = 41.459$, $p < .01$. The point estimate for the correlation was 0.8433. The 95% confidence interval for rho ranged from 0.8205 to 0.8635, excluding 0. The mean correlation in the Bayesian posterior distribution of rho was 0.8422. The 95% HDI ranges from 0.8026 to

0.8823. The Bayes factor of 1.368053e+186:1 is in favor of the alternative hypothesis.

For withPolio and the PctUpToDate, t (698) = 72.975, $p < .01$. The point estimate for the correlation was 0.9403. The 95% confidence interval for rho ranged from 0.931 to 0.948. The mean correlation in the posterior distribution of rho was 0.94. The 95% HDI ranges from 0.9148 to 0.9654. The Bayes factor of 1.848858e+324:1 was in favor of the alternative hypothesis.

For withMMR and the PctUpToDate, t (698) = 100.52, $p < .01$. The point estimate for the correlation was 0.9672. The 95% confidence interval for rho ranged from 0.962 to 0.9716. The mean correlation in the posterior distribution of rho was 0.9668. The 95% HDI ranges from 0.948 to 0.9859. The Bayes factor of 4.596466e+412:1 was in favor of the alternative hypothesis.

For all vaccines, we reject the null hypothesis of rho = 0. All confidence intervals and HDIs consistently exclude 0, suggesting strong positive relationships and are more than could be expected by chance. All have overwhelming Bayes Factors favor the alternative hypothesis.

In conclusion, the evidence suggests that if there are students with one vaccine, students are very likely to be up-to-date with all others.

c. How do these Californian vaccination levels compare to U.S. vaccination levels (recent years only)? [See notebook 3.c.]

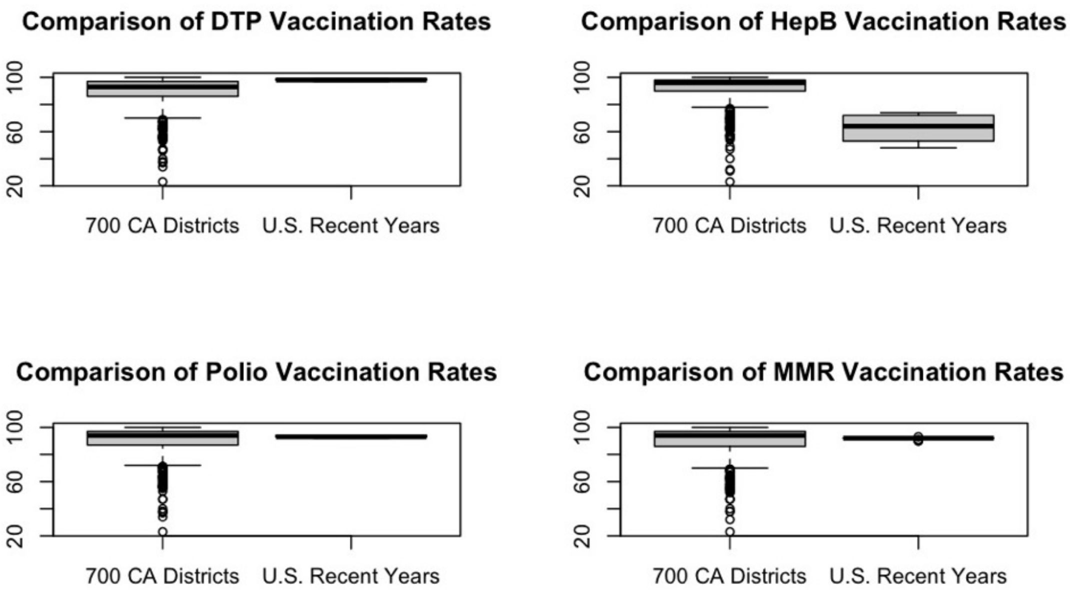

Figure 8

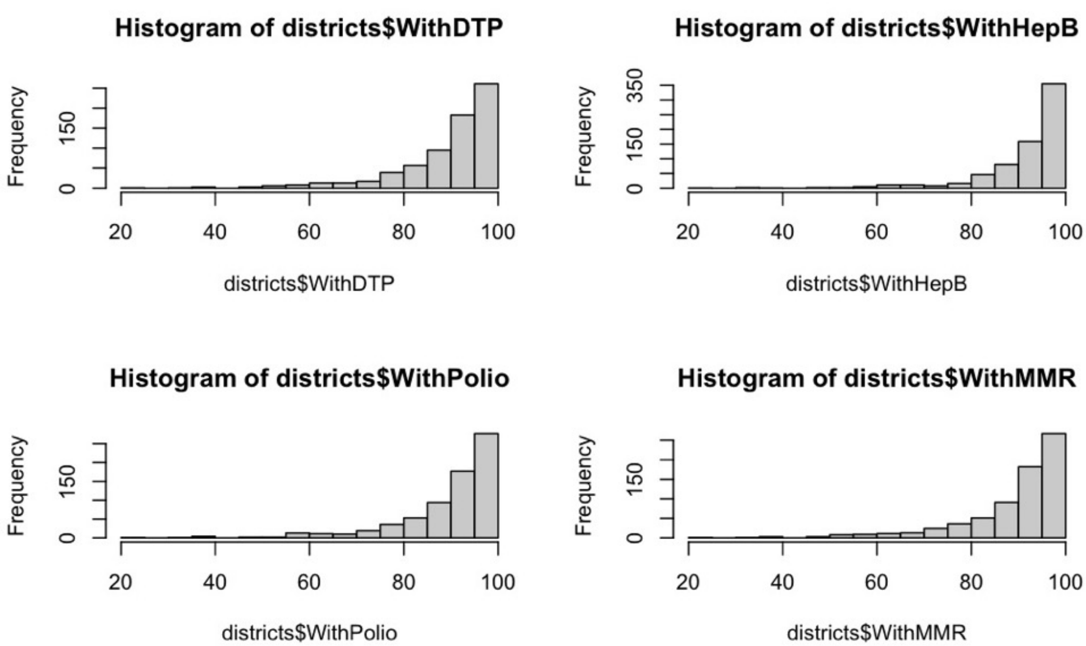

Figure 9

Figure 8 shows box plots illustrating the differences in vaccination rates for 4 vaccines—DTP, HepB, Polio, and MMR—between California districts and the U.S. in recent years. I conducted both frequentist and Bayesian t-tests to compare the DTP, HepB, Polio, and MMR vaccination rates between California school districts and the overall U.S. rates for recent years. Since the percentage of students in California districts with these vaccines are not normally distributed (Figure 9), I also conducted non-parametric Kruskal-Wallis tests using kruskal.test().

For **DTP**, t (448.14) = -18.926, $p < .01$. Besides, with df = 1, the Kruskal-Wallis Chi-squared = 19.053, $p < .01$. We reject the null hypothesis that the true differences in means and medians are equal to 0. The 95% confidence interval for the difference in mean DTP vaccination rates between California districts and the overall U.S. data ranged from -9.2928 to -7.5444. If we replicate the whole study many times, on average, we would expect that approximately 95 out of 100 times, our interval contains the true mean difference, though we don't know for sure if any particular one does. However, since the CI doesn't span 0, the population mean difference is more than could be expected by chance, meaning the mean DTP vaccination rate in California districts do seem to be lower than the recent overall U.S. rate.

The HDI from the BESTmcmc() procedure shows that there's a 95% chance that the population mean difference falls in the range of -5.28 to -3.9. The population value of the mean difference is in the region near -4.59. The likelihood of a population mean difference of 0 or smaller is 100%. There's a very consistence evidence that the DTP vaccination rates in California are lower than the recent overall U.S. rates.

For **HepB**, t (13.597) = 12.076, $p < .01$. With df = 1, the Kruskal-Wallis Chi-squared = 36.577, $p < .01$. We reject the null hypothesis that the true differences in means and medians are equal to 0. The 95% confidence interval for the difference in mean HepB vaccination rates between California districts and the overall U.S. data ranged from 24.9316 to 35.737. If we replicate the whole study many times, on average, we would expect that approximately 95 out of 100 times, our interval contains the true mean difference, though we don't know for sure if any particular one does. However, since the CI doesn't span 0, the population mean difference is more than could be expected by chance, meaning the HepB vaccination rates in California districts do seem to be higher than the recent overall U.S. rates.

The HDI from the BESTmcmc() procedure shows that there's a 95% chance that the population mean difference falls in the range of 26.5 to 39.6. The population value of the mean difference is in the region near 33. The likelihood of a population mean difference of 0 or larger is 100%. There's a very consistence evidence that the HepB vaccination rates in California districts are much higher than the recent overall U.S. rates.

For **Polio**, $t\ (316.03) = -6.6037$, $p < .01$, 95% CI ranged from -3.9068 to -2.113, suggesting lower rates in California. However, with df = 1, the Kruskal-Wallis Chi-squared = 0.19229, $p > .05$. We fail to reject the null hypothesis that the true difference in medians is equal to 0. Observed difference could be due to the sampling.

The HDI from the BESTmcmc() procedure shows that there's a 95% chance that the population mean difference in Polio vaccination rates between California districts and the overall U.S. data falls in the range of 0.109 to 1.55. The population value of the mean difference is in the region near 0.83. The likelihood of a population mean difference of 0 or larger is 98.5%, 0 or smaller is 1.5%. There's a slight tendency towards the Polio vaccination rates in California districts higher than in the U.S. But it's not definite.

For **MMR**, $t\ (234.17) = -3.9986$, $p < .01$, 95% CI ranged from -2.8767 to -0.9777, suggesting lower rates in California. However, with df = 1, the Kruskal-Wallis Chi-squared = 1.3126, p > .05. We fail to reject the null hypothesis that the true difference in medians is equal to 0. Observed difference could be due to the sampling.

The HDI from the BESTmcmc() procedure shows that there's a 95% chance that the population mean difference in MMR vaccination rates between California districts and the overall U.S. data falls in the range of 1.52 to 2.65. The population value of the mean difference is in the region near 2.09. The likelihood of a population mean difference of 0 or larger is 100%. There's evidence that the MMR vaccination rates in California districts are higher than the recent overall U.S. rates.

Here, the results from Kruskal-Wallis test and BESTmcmc() analysis aren't consistent. It's possible that there are differences in means while median differences are not significant, potentially due to the influence of outliers or skewed data distributions.

d. Provide one or two sentences of your professional judgment about where California school districts stand with respect to vaccination rates in the larger context of the U.S.

California school districts generally have similar vaccination rates for MMR and Polio compared to recent overall U.S. rates, and have notably higher rates for

HepB, though all four vaccines, including DTP—which shows slightly lower mean rates—feature some districts with lower vaccination outliers.

4. Comparison of public and private schools

a. What proportion of public schools and what proportion of private schools reported vaccination data? [See notebook 4.a.]

In the all schools dataset, there are 7,381 observations of 18 variables, including 5,732 public schools and 1,649 private schools. Of these, 148 public schools and 252 private schools did not report their vaccination data (REPORTED == N). Figure 10 shows the original contingency table. However, one private school (SCHOOL CODE: 6143788), was marked as 'N', but actually has values in the vaccine columns. I have corrected its REPORTED status from 'N' to 'Y'. Consequently, this aligns with the fact that there are 399 schools with NAs in 10 columns related to vaccines.

Figure 11 shows the updated contingency table. Out of 5,732 public schools, 97.418% reported their vaccination data, whereas 84.78% of the 1,649 private schools reported their vaccination data.

	PRIVATE	PUBLIC
N	252	148
Y	1397	5584

Figure 10

	PRIVATE	PUBLIC
N	251	148
Y	1398	5584

Figure 11

b. Was there any credible difference in reporting between public and private schools? [See notebook 4.b.]

The reported value of Chi-squared, 400.07 on one degree of freedom, $p < .01$. Thus, we reject the null hypothesis of independence in reporting between public and private schools.

The Bayes factor of 3.656717e+68 suggesting that the odds are strongly in favor of the alternative hypothesis of non-independence.

The contingency table shows that in the Private school's column, the proportion of Not Reported versus Reported is 251:1398 = 0.1795. In the Public school's column, the proportion is 148:5548 = 0.0267.

Running contingencyTableBF() with posterior sampling, the difference in proportions across 10,000 posterior estimates of the Not Reported/Reported ratio for private and public schools shows that the 95% HDI for the difference in proportions spans from approximately 0.1292 to 0.1791.

There is a 95% probability that the true difference in proportions falls within this range with the mean difference being 0.1535. This evidence suggests a credible difference in reporting between public and private schools, with public schools more likely to report their vaccination data.

c. Does the proportion of students with up-to-data vaccinations vary from county to county? Report significant details. [See notebook 4.c.]

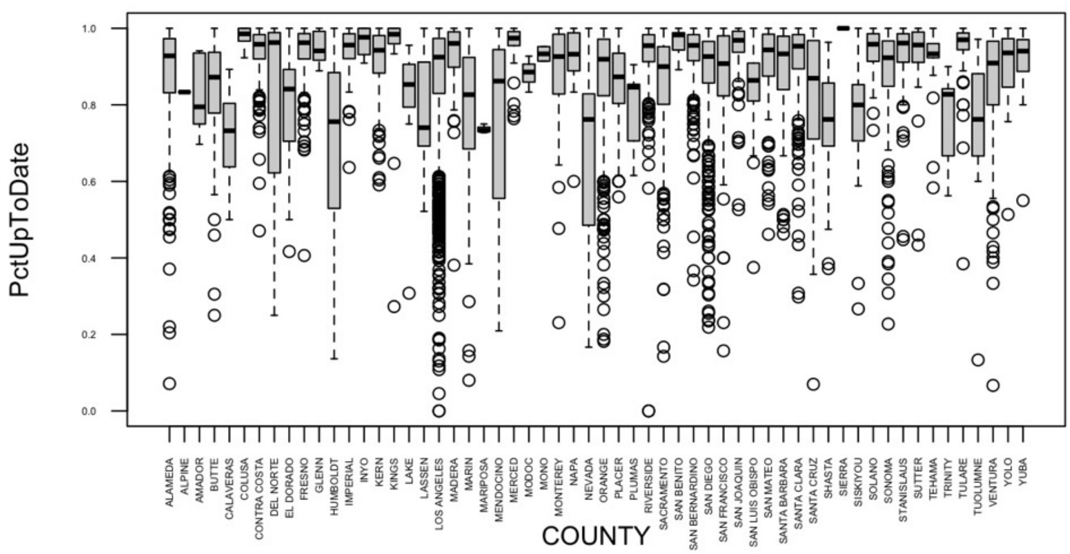

Figure 12

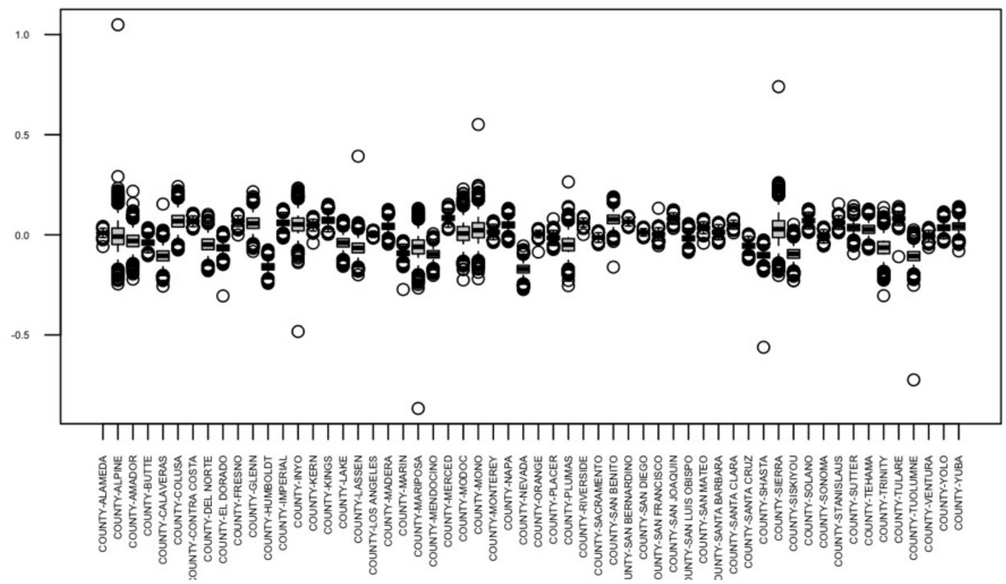

Figure 13

Figure 14

Figure 12 shows the boxplot for the distribution of students with up-to-data vaccinations across counties in California.

Running the ANOVA test, $F\ (57,6924) = 13.1$, $p < 0.01$. We reject the null hypothesis, where the null hypothesis is that all counties have the same proportion of students with up-to-data vaccinations and there's no difference between group means.

A Bayesian analysis was performed to compare the proportions of students with up-to-date vaccinations across 58 counties in California, N=7,381. The analysis was performed with the anovaBF command from the R package BayesFactor. The Bayes factor comparing the model with groups to a model without groups was 1.046271e+110, showing extremely strong evidence that differences among counties impact vaccination rates. Figure 13 shows the Boxplot of the posterior distributions for the groups.

Examination of the HDIs for the groups suggests that counties Butte, Humboldt, Calaveras, El Dorado, Marin, Mendocino, Nevada, Santa Cruz, Shasta, Siskiyou, and Tuolumne had the lowest proportion of students with up-to-date vaccinations (a mean of -0.04, -0.16, -0.10, -0.06, -0.09, -0.1, -0.17, -0.05, -0.1, -0.1, and -0.1 below the average, HDIs -0.07 to 0, -0.2 to -0.12, -0.17 to -0.04, -0.11 to -0.02, -0.12 to -0.06, -0.15 to -0.04, -0.23 to -0.12, -0.09 to -0.02, -0.14 to -0.06, -0.16 to -0.04, and -0.17 to -0.04 respectively). And counties Contra Costa, Fresno, Imperial, Kern, Kings, Merced, Napa, Riverside, San Benito, San Bernardino, Santa Clara, Solano, Stanislaus, and Tulare showed higher proportions of students with up-to-date vaccinations, with means of 0.07, 0.07, 0.06, 0.05, 0.07, 0.08, 0.05, 0.06, 0.08, 0.07, 0.05, 0.07, 0.06, and 0.08 above the average. The corresponding HDIs were 0.05 to 0.09, 0.05 to 0.09, 0.02 to 0.1, 0.03 to 0.07, 0.03 to 0.11, 0.05 to 0.12, 0.01 to 0.09, 0.05 to 0.08, 0.02 to 0.14, 0.05 to 0.08, 0.03 to 0.07, 0.04 to 0.1, 0.04 to 0.09, and 0.06 to 0.1 respectively.

I conducted a Bayesian t-test to compare the groups NEVADA and TULARE. The 95% HDI boundary values calculated by BESTmcmc() ranged from 0.123 to 0.329. There is a 95% chance that the population mean difference between the two groups falls within this range. The population value of the mean difference is approximately 0.219. The likelihood that the population mean difference is 0 or larger is 100%. There is consistent evidence suggesting that TULARE has a higher proportion of students with up-to-date vaccinations than NEVADA. Figure 14 displays the box plots for these two counties.

In addition, the Tukey HSD post-hoc procedure revealed differences between each pair of counties.

5. Inferential reporting about districts

a. Which of the four predictor variables predicts the percentage of all enrolled students with belief exceptions? [See notebook 5.a.]

In the districts dataset, Enrolled and TotalSchools are highly skewed, so I applied a log transformation.

PctBeliefExempt and PctMedicalExempt are also skewed, primarily due to many zeros, which precludes the use of log transformation. Although research offers several solutions, I will maintain their current values for now.

Boulton, A. J., & Williford, A. (2018). Analyzing skewed continuous outcomes with many zeros: A tutorial for social work and youth prevention science researchers. Journal of the Society for Social Work and Research, 9(4), 721-740. https://doi.org/10.1086/701235

	PctChildPoverty	PctFamilyPoverty	EnrolledLog	TotalSchoolsLog	PctBeliefExempt
PctChildPoverty	1.00000000	0.867776828	-0.05301706	-0.087286318	-0.2013054
PctFamilyPoverty	0.86777683	1.000000000	0.05546011	-0.005408707	-0.2655436
EnrolledLog	-0.05301706	0.055460112	1.00000000	0.916319391	-0.2727797
TotalSchoolsLog	-0.08728632	-0.005408707	0.91631939	1.000000000	-0.1993181
PctBeliefExempt	-0.20130543	-0.265543568	-0.27277972	-0.199318123	1.0000000

Figure 15

I checked the correlation among variables (Figure 15) and checked multicollinearity when building the linear regression model with all 4 predictor variables. The VIF values indicated multicollinearity, suggesting the need to remove either PctChildPoverty or PctFamilyPoverty, and either EnrolledLog or TotalSchoolsLog. I chose PctFamilyPoverty over PctChildPoverty and EnrolledLog over TotalSchoolsLog based on their stronger correlations with PctBeliefExempt.

Subsequently, **I rebuilt the model using only PctFamilyPoverty and EnrolledLog as predictors for PctBeliefExempt. This adjustment resolved the problem of multicollinearity.**

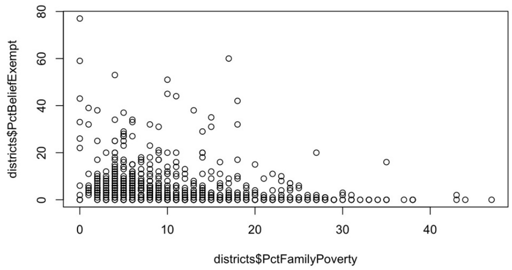

Figure 16

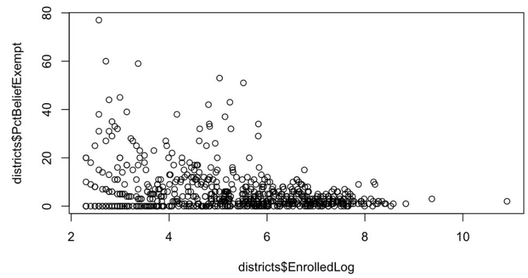

Figure 17

Figures 16 and 17 show the scatter plot of PctFamilyPoverty vs. PctBeliefExempt and EnrolledLog vs. PctBeliefExempt.

When checking the residuals, there's evidence of a curvilinear relationship. While the residuals showed some outliers and a heavy-tailed distribution, and the gvlma indicates problems with skewness, kurtosis, and heteroscedasticity, the model still provides a potentially useful approximation of the underlying relationship. Given the current analysis context, we will proceed with the model while acknowledging these limitations.

The results of the regression were significant, F (2, 697) = 55.47, $p < 0.01$. The null hypothesis test on this R-squared—which asserts that R-squared is actually 0 in the population—has meaning that we reject the null hypothesis. The adjusted R-squared was 0.1348, meaning that the percentage of families in district living below the poverty line and the log of total number of enrolled students in the district variables explained about 13.48% of the variance in percentage of all enrolled students with belief exceptions on vaccination.

The coefficient for PctFamilyPoverty is -0.27(significant, $p < .01$), suggesting a negative relationship with the dependent variable. The coefficient for the EnrolledLog is -1.44 (significant, $p < .01$), also indicating a negative relationship. For each one percentage point increase in PctFamilyPoverty, there is an associated decrease of 0.27 percentage points in the PctBeliefExempt. And for each one unit increase in the log of enrollment, there is an associated decrease of 1.44 percentage points in the PctBeliefExempt.

The standardized coefficients show that the negative relationship between EnrolledLog and PctBeliefExempt is very slightly stronger than that between PctFamilyPoverty and PctBeliefExempt.

A Bayes factor of 8.821259e+19 shows that the odds are overwhelmingly in favor of the alternative hypothesis, suggesting that a model containing the EnrolledLog and PctFamilyPoverty as predictors is hugely favored over a model that only contains the Y-intercept, forcing all the B-weights on the predictors to be 0.

The 95% HDI for posterior distribution of R-squared runs from 3.88% to 21.87% with a mean of 13.31%. The 95% HDI for posterior distribution of the coefficient for EnrolledLog shows that there is 95% chance that the population coefficient is from -1.8 to -1. The 95% HDI for posterior distribution of the coefficient for PctFamilyPoverty shows that there is 95% chance that the population coefficient is from -0.34 to -0.19. The results from the Bayesian analysis are very close to those from the traditional analysis.

b. What combination of predictors gives the best prediction of the percentage of all enrolled students with completely up-to-date vaccines? [See notebook 5.b.]

First, I checked for missing values in the districts dataset. There are 244 NAs out of 700 observations in the PctFreeMeal column. I dropped this column.

Among checking correlation and multicollinearity among the variables, I selected **WithMMR and PctBeliefExempt as predictor variables to predict the percentage of all enrolled students with completely up-to-date vaccines**. They gave the best prediction. Figures 18 and 19 show the scatter plot of PctBeliefExempt vs. PctUpToDate and WithMMR vs. PctUpToDate.

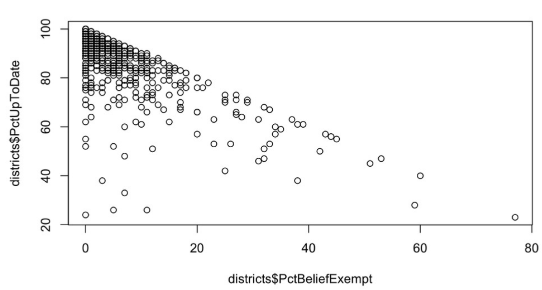

Figure 18

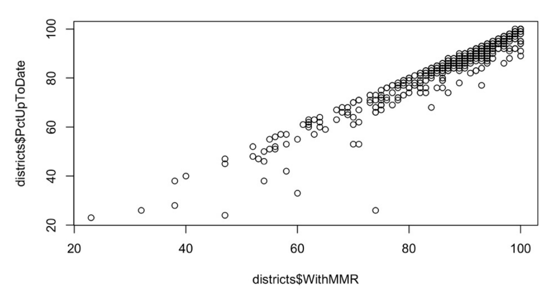

Figure 19

The relationship is not strictly linear, yet it remains acceptable. There are some outliers, but the residuals look okay. We proceed with the model while acknowledging these limitations.

The results of the regression were significant, F (2, 697) = 5379, $p < 0.01$. The null hypothesis test on this R-squared—which asserts that R-squared is 0 in the population—has meaning that we reject the null hypothesis. The adjusted R-squared was 0.939, meaning that the percentage of students in the district with the MMR vaccine and the percentage of all enrolled students with belief exceptions variables explained about 93.9% of the variance in percentage of students with completely up-to-date vaccines.

The coefficient for WithMMR is 1.173 (significant, $p < .01$), suggesting a positive relationship with the dependent variable. The coefficient for the PctBeliefExempt is 0.14218 (significant, $p < .01$). Though PctBeliefExempt itself is negatively correlate with the PctUpToDate, after controlling for WithMMR, there is a positive relationship between the PctBeliefExempt and the PctUpToDate.

The standardized coefficients showed that there's a stronger positive relationship between WithMMR and PctUpToDate than the percentage of belief exempt and the PctUpToDate.

A Bayes factor of 6.658546e+419 shows that the odds are overwhelmingly in favor of the alternative hypothesis, suggesting that a model containing the WithMMR and PctBeliefExempt as predictors is hugely favored over a model that only contains the Y-intercept, forcing all B-weights on the predictors to be 0.

The Bayesian regression results are similar to but not identical to the traditional regression results. The 95% HDI for posterior distribution of R-squared runs from 85.88% to 88.61% with a mean of 87.34%. This is somewhat lower than the traditional analysis result.

c. In predicting the percentage of all enrolled students with completely up-to-date vaccines, is there an interaction between PctChildPoverty and Enrolled? If so, interpret the interaction term. [See notebook 5.c.]

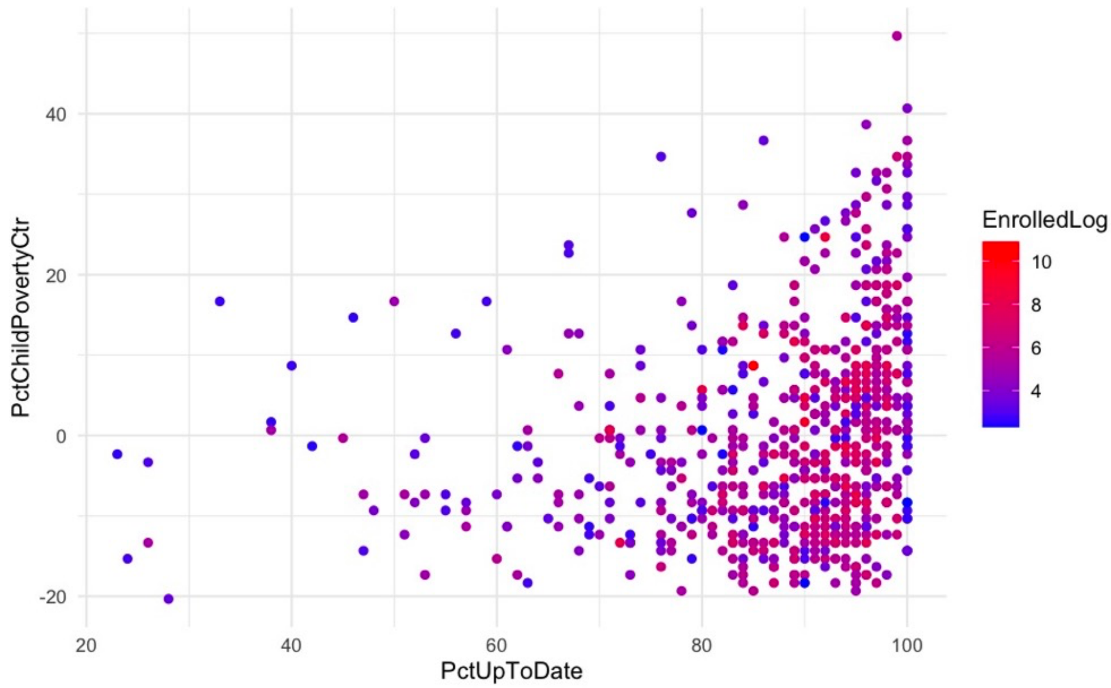

Figure 20

A regression was conducted to test whether there is an interaction between PctChildPoverty and Enrolled in predicting the percentage of all enrolled students with completely up-to-date vaccines (N = 700).

Figure 20 shows the scatter plot on interaction effect of PctChildPoverty and EnrolledLog on PctUpToDate. There is no obvious pattern to suggest a strong interaction effect.

Though there are some outliers, non-linearity, skewness, and kurtosis present, the residuals appear okay.

The regression was significant F (3,696) = 37.66, p < .01. We reject the null hypothesis that R-squared is equal to 0. The adjusted R-squared is 0.1359, suggesting that PctChildPoverty, EnrolledLog and their interaction explained about 13.59% of the variance in PctUpToDate.

Both PctChildPoverty (b = 0.24, p < .01), and EnrolledLog (b = 2.38, p < .01) were significant predictors; however, **the interaction was not significant.**

The Bayes Factor was 0.2362, which is approximately 4.2337 when inverted the fractional odds ratio. **This provides evidence in favor of the model without the interaction.**

d. Which, if any, of the four predictor variables predict whether or not a district's reporting was complete? [See notebook 5.d.]

	PctChildPoverty	PctFamilyPoverty	EnrolledLog	TotalSchoolsLog	DistrictComplete
PctChildPoverty	1.00000000	0.867776828	-0.05301706	-0.087286318	-0.07383085
PctFamilyPoverty	0.86777683	1.000000000	0.05546011	-0.005408707	-0.09854408
EnrolledLog	-0.05301706	0.055460112	1.00000000	0.916319391	-0.14050086
TotalSchoolsLog	-0.08728632	-0.005408707	0.91631939	1.000000000	-0.22967828
DistrictComplete	-0.07383085	-0.098544081	-0.14050086	-0.229678279	1.00000000

Figure 21

After checking the correlation (Figure 21) and multicollinearity among the variables, **I selected PctFamilyPoverty and TotalSchoolsLog as predictor variables to predict whether a district's reporting was complete.**

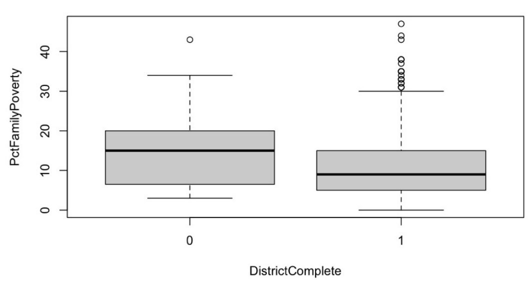

Figure 22

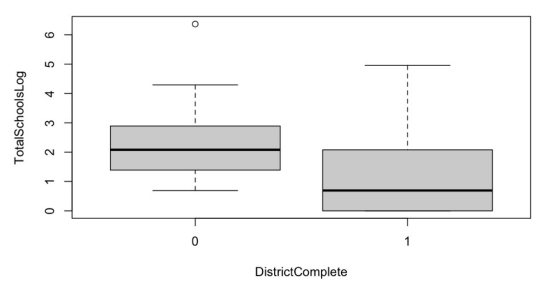

Figure 23

Figures 22 and 23 show the distribution of these variables relative to whether district reporting was complete or not.

DistrictComplete is a binary variable, and a logistic regression was conducted. There's evidence of a curvilinear relationship, outliers, and a heavy-tailed distribution, while DHARMa results show that the residuals look okay, with no significant problems detected. We will proceed with the model while acknowledging these limitations.

The logistic regression results indicate that both PctFamilyPoverty ($z = -2.732$, $p < .01$) and TotalSchoolsLog ($z = -5.571$, $p < .01$) were statistically significant predictors of whether a district's reporting was complete. We reject the null hypothesis that introducing these 2 predictors into the model caused zero reduction of the model error.

For each one-unit increase in the PctFamilyPoverty, the odds of a district's reporting being completed decrease by a factor of approximately 0.95 (95% CI: 0.92 to 0.99), holding TotalSchoolsLog constant. Similarly, for each one-unit increase in the TotalSchoolsLog, the odds of a district's reporting being complete decrease by a factor of approximately 0.47 (95% CI: 0.35 to 0.61), holding PctFamilyPoverty constant.

The difference between the null deviance and the model deviance, in this case $323.23 - 282.74 = 40.49$, is distributed as chi-square. A chi-square omnibus test on the results of logistic regression was significant,

for PctFamilyPoverty, $chisq(1) = 6.024$, $p < .05$, for TotalSchoolsLog, $chisq(1) = 34.465$, $p < .01$. These results suggest that a model containing these 2 variables significantly predicts reporting completion compared to a model without any predictors.

Tjur's R2 is reported at 0.078. We might loosely interpret that the model explains a small proportion of the variance in reporting completion.

The model's fit was also evaluated using a confusion matrix. It indicates an accuracy of 94%, with a sensitivity of 100% and a specificity of 2.33%. The model is not very effective, since the no-information rate is 93.86%. This is a result of the imbalance in the dataset. For future analyses, applying different weights to the model could help address this imbalance and improve the model's effectiveness across both categories.

A Bayesian logistic regression was also performed on the same data. Examination of the traces showed that the sampling had converged. The mean posterior odds for PctFamilyPoverty were 0.95 (95% HDI: 0.92 to 0.99), meaning that a unit increase in PctFamilyPoverty is associated with a 0.95x decrease in the likelihood of reporting complete. For TotalSchoolsLog, the mean posterior odds were 0.46 (95% HDI: 0.34 to 0.6), meaning that a unit increase in TotalSchoolsLog is associated with a 0.46x decrease in the likelihood of reporting complete. The results are similar to but not identical to the logistic regression results.

6. Concluding Paragraph

Describe your conclusions, based on the foregoing analyses. The staff member in the state legislator's office is interested to know how to allocate financial assistance to school districts to improve both their vaccination rates and their reporting compliance. Make sure you have at least one sentence that makes a recommendation about improving vaccination rates and at least one that makes a recommendation about improving reporting rates. Finally, say what further analyses might be helpful to answer these questions and any additional data you would like to have.

Financial assistance should be prioritized for schools in counties such as Butte, Humboldt, Calaveras, El Dorado, Marin, Mendocino, Nevada, Santa Cruz, Shasta, Siskiyou, and Tuolumne. These areas were identified as having the lowest proportions of students with up-to-date vaccinations. Allocating funds to these counties could focus on enhancing vaccination awareness programs and improving access to vaccines. Resources should also focus on public school districts where low outliers in vaccine rates are located, as shown in Figure 8. A list of these district names is provided in the new code notebook part 6.

PctFamilyPoverty and PctChildPoverty are positively correlated with PctUpToDate. In regions experiencing the highest poverty, there may already be more targeted health interventions. Therefore, resources should be allocated not only to regions with the highest levels of poverty but also to those with moderate levels of poverty.

Given that PctBeliefExempt is negatively correlated with PctUpToDate, targeted educational campaigns could be developed to address vaccine hesitancy. These programs should respect community beliefs while providing clear, evidence-based information about the benefits of vaccinations. This dual approach may help reduce exemptions based on personal beliefs without alienating the community.

Given that individuals who receive one vaccine are likely to receive others, the reluctance to vaccinate may stem not from ignorance about specific diseases but from broader concerns about vaccines. Therefore, it is crucial to implement comprehensive educational campaigns that emphasize the importance of vaccinations for children, addressing common fears and misconceptions to improve overall vaccination rates.

Regarding the reporting compliance, private schools could increase their reporting compliance. Initiatives could include training programs for school administrators on the importance of vaccination data accuracy, streamlined reporting processes, and potential incentives for timely and accurate reporting.

Further analyses and Additional data: It would be beneficial to have qualitative data on why public schools are not reporting complete vaccinations, as our current data analysis does not explain much.

LLM disclosure

If you used an LLM to prepare any answers to this exam, disclose so here. For each question, paste in the prompt you gave and explain how you used the response you received.

Question 2.c

My prompt: I've chosen to use the period from 2004 to 2017 as the recent years based on the latest mean changepoint for HepB_BD in 2004. Is my analysis correct?

ChatGPT suggested that while this makes sense for HepB_BD, other vaccines show earlier changepoints for both mean and variance, which suggests that their data characteristics might have stabilized earlier. It might be beneficial to check if post-2004 data for all vaccines indeed show stability in both mean and variance. This would justify using this period universally across all vaccines. If not, consider tailoring the recent period for each vaccine based on its last significant changepoint, or analyze the implications of using a uniform period despite these differences.

Question 3.c

My prompt: Does the BESTmcmc() in R require normal distribution? How do we conclude if results from the Kruskal-Wallis test and BESTmcmc() analysis are not consistent?

ChatGPT suggested to consider the measures of central tendency: The discrepancy might arise from the differences in what each test measures (mean vs. median). It's possible for the mean differences to be significant while median differences are not, especially if the data distribution is skewed or contains outliers.

Question 3.d

My prompt: "California school districts generally have similar vaccination rates for MMR and Polio compared to recent overall U.S. rates. The HepB vaccination rate in California school districts is much higher, though it does include lower outliers. For DTP, California has a little bit lower mean vaccination rate compared to the recent overall U.S. averages. For all 4 vaccines, California has some district outliers with lower rates.” Can you concise it into 1 or 2 sentences?

ChatGPT: California school districts generally have similar vaccination rates for MMR and Polio compared to recent overall U.S. rates, and have notably higher rates for HepB, though all four vaccines, including DTP—which shows slightly lower mean rates—feature some districts with lower vaccination outliers.

Question 5.d

My prompt: “The model’s fit was also evaluated using a confusion matrix. It indicates an accuracy of 94%, with a sensitivity of 100% and a specificity of 2.33%. The model is not very effective, since the no-information rate is 93.86%. It is much better at predicting completed districts than non-completed ones.” Please revise this part, saying that this is due to the imbalance of Completed and Not completed in the dataset. For future analysis, we might apply different weights when building model.

ChatGPT: The model’s fit was evaluated using a confusion matrix, which indicates an accuracy of 94%, with a sensitivity of 100% and a specificity of 2.33%. However, the model is not very effective, largely due to the no-information rate being 93.86%. This high no-information rate is a result of the imbalance in the dataset, where the number of completed districts far outweighs the number of non-completed ones. Consequently, the model is significantly better at predicting completed districts than non-completed ones. For future analyses, applying different weights to the model when building it could help address this imbalance and improve the model's effectiveness across both categories.

Question 6

My prompt: Here is the question: “Describe your conclusions, based on the foregoing analyses. The staff member in the state legislator’s office is interested to know how to allocate financial assistance to school districts to improve both their vaccination rates and their reporting compliance. Make sure you have at least one sentence that makes a recommendation about improving vaccination rates and at least one that makes a recommendation about improving reporting rates. Finally, say what further analyses might be helpful to answer these questions and any additional data you would like to have.”

Please revise my answers: “Financial assistance should allocate to schools in counties in Butte, Humboldt, Calaveras, El Dorado, Marin, Mendocino, Nevada, Santa Cruz, Shasta, Siskiyou, and Tuolumne, since they had the lowest proportion of students with up-to-date vaccinations below the average. PctBeliefExempt is negatively correlated with the PctUpToDate. We probably could educate people on keeping their beliefs but at the same time receiving necessary vaccinations. Since people who have received one vaccine usually also receive other vaccines or all of them, the reason people do not vaccinate may not be ignorance of a particular disease, but rather out of worries of vaccines in a broader sense. There should be more educational campaigns to publicize the need for children to be vaccinated. Regarding the reporting compliance, private schools should increase their reporting compliance."

ChatGPT: Improving Vaccination Rates:

Financial assistance should be prioritized for schools in counties such as Butte, Humboldt, Calaveras, El Dorado, Marin, Mendocino, Nevada, Santa Cruz, Shasta, Siskiyou, and Tuolumne. These areas were identified as having the lowest proportions of students with up-to-date vaccinations. Allocating funds to these counties could focus on enhancing vaccination awareness programs and improving access to vaccines, particularly in schools where logistical and financial barriers may impede vaccine uptake.

Addressing Vaccine Hesitancy:

Given that PctBeliefExempt is negatively correlated with PctUpToDate, targeted educational campaigns could be developed to address vaccine hesitancy. These programs should respect community beliefs while providing clear, evidence-based information about the benefits of vaccinations. This dual approach may help reduce exemptions based on personal beliefs without alienating the community.

Given that individuals who receive one vaccine are likely to receive others, the reluctance to vaccinate may stem not from ignorance about specific diseases but from broader concerns about vaccines. Therefore, it is crucial to implement comprehensive educational campaigns that emphasize the importance of vaccinations for children, addressing common fears and misconceptions to improve overall vaccination rates.

Improving Reporting Compliance:

Special attention should be given to improving reporting compliance, particularly among private schools, where compliance rates are notably lower. Initiatives could include training programs for school administrators on the importance of vaccination data accuracy, streamlined reporting processes, and potential incentives for timely and accurate reporting.

My Prompt: when getting boxplot(districts$PctUpToDate), boxplot(districts$WithMMR), boxplot(districts$WithPolio), boxplot(districts$WithHepB), boxplot(districts$WithDTP), print a list of those outliers' DistrictName

ChatGPT: code in notebook part 6.