Spring 2021 - Final Examination

Zijun Yi

github: https://github.com/zyi103/ist772-vaccination-rate-analysis

Instructions

Your goal for this final exam is to conduct the necessary analyses of vaccination rates in California school districts and then write up a technical report for a scientifically knowledgeable staff member in a California state legislator's office. You should provide sufficient numeric and graphical detail that the staff member can create a comprehensive briefing for a legislator (see question 10 for specific points of interest). You can assume that the staff member understands the concept of statistical significance and other basic concepts like mean, standard deviation, and correlation.

For this exam, the report writing is very important: Your responses will be graded on the basis of clarity; conciseness; inclusion and explanation of specific and appropriate statistical values; inclusion of both frequentist and Bayesian inferential evidence (i.e., it is not sufficient to just examine the data); explanation of any included tabular material and the appropriate use of graphical displays when/if necessary. It is also important to conduct a thorough analysis, including both data exploration and cleaning and appropriate diagnostics. Bonus points will be awarded for work that goes above expectations.

In your answer for each question, make sure you write a narrative with complete sentences that answers the substantive question. You can choose to put important statistical values into a table for readability, or you can include the statistics within your narrative. Be sure that you not only report what a test result was, but also what that result means substantively. Make sure to include enough statistical information so that another analytics professional could review your work. 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. Finally, be sure to proofread your final knitted submission to ensure that everything is included and readable.

You may not *receive assistance, help, coaching, guidance, or support from any human except your instructor at any point during 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!*

Data

You have an RData file available on Blackboard area that contains two data sets that pertain to vaccinations for the U.S. as a whole and for Californian school districts. The U.S. vaccine data is a time series and the California data is a sample of end-of-year vaccination reports from n=700 school districts. Here is a description of the datasets:

usVaccines – 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 (i.e., DTP); HepB_BD = Hepatitis B, Birth Dose (HepB); Pol3 = Polio third dose (Polio); Hib3 – Influenza third dose; MCV1 = Measles first dose (included in MMR))

districts – A sample of California public school districts from the 2017 data collection, along with specific numbers and percentages for each district:

```
'data.frame':   700 obs. of  14 variables:
 $ DistrictName    : Name of the district
 $ WithDTP         : Percentage of students in the district with the DTP vaccine
 $ WithPolio       : Percentage of students in the district with the Polio vaccine
 $ WithMMR         : Percentage of students in the district with the MMR vaccine
 $ WithHepB        : Percentage of students in the district with Hepatitis B vaccine
 $ PctUpToDate     : Percentage of students with completely up-to-date vaccines
 $ DistrictComplete: Boolean showing whether or not district's reporting was complete
 $ PctBeliefExempt : Percentage of all enrolled students with belief exceptions
 $ PctMedicalExempt: Percentage of all enrolled students with medical exceptions
 $ PctChildPoverty : Percentage of children in district living below the poverty line
 $ PctFamilyPoverty: Percentage of families in district living below the poverty line
 $ PctFreeMeal     : Percentage of students in the district receiving free or reduced cost meals
 $ Enrolled        : Total number of enrolled students in the district
 $ TotalSchools    : Total number of different schools in the district
```

As might be expected, the data are quite skewed: districts range from 1 to 582 schools enrolling from 10 to more than 50,000 students. Further, while most districts have low rates of missing vaccinations, a handful are quite high. Be sure to note problems the data cause for the analysis and address any problems you can.

Data Cleaning

Data points that not possiable to be ture.

Run Data Cleaning diagnose

creating data points to test for the credibility

Ratio between Child and Family poverty prompts question on data collection.

Descriptive Reporting

1. *Basic Introductory Paragraph*

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

It is a fact that low vaccination rate will bring back dangerous disease such as Diphtheria, Polio, Measles and Hepatitis. This report will analysis the vaccination rate in California public school, how poverty, religious and medical factor effects them.

2. *Descriptive Overview of U.S. Vaccinations*

You have U.S. vaccination data going back 38 years, but the staff member is only interested in recent vaccination rates as a basis of comparison with California schools.

a. *How have U.S. vaccination rates varied over time?*

Fig 1 - US average Vaccination rate(in percentages) on Diphtheria/Pertussis/Tetanus(DTP), Hepatitis B(HepB_BD), Polio(Pol3), Influenza(Hib3) and Measles(MCV1)

Polio, Hib, MMR all have a great vaccination rate above 80 percent, and are relatively consistent, except at around 1985 - 1990 there is an drop in data. HepB_BD,and DTP had seen a increase from 85 to 95 and from 20 to 60.

b. *Are there notable trends or cyclical variation in U.S. vaccination rates?*

Fig 2.1 - Trend analysis On 5 type of vaccination rate in the US

Fig 2.2 - Seasonality analysis On 5 type of vaccination rate in the US

c. *What are the mean U.S. vaccination rates when including only recent years in the calculation of the mean (examine your answers to the previous question to decide what a reasonable recent period is, i.e., a period during which the rates are relatively constant)?*

3. *Descriptive Overview of California Vaccinations*

Your districts dataset contains four variables that capture the individual vaccination rates by district: WithDTP, WithPolio, WithMMR, and WithHepB.

a. *What are the mean levels of these variables across districts?*

Fig 3.1 - Distribution of the average on four vaccine

We have a mean at 91.1, standard deviation at 9.94 and a negative skewness.

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 of the others?*

Fig 3.2 - Four Type of Vaccination Correlation in California

Fig 3.3 - Four Type of Vaccination Correlation in California after power transformation

Fig 3.4 - Box plots of Vaccination rate in California

The summary of the model have a p-value of 0.000037. We can confidently reject the null hypothesis and say there is a different of mean between these 4 groups Although we do not know which group are different, or even it's possible that all 4 groups are different. Cobain with the correlation graph(Fig 3.2) and the boxplot(Fig 3.3), my inference is the Hepatitis B vaccination rate are different with the other 3 groups. The Hepatitis B vaccine have a correlation from .92~.93, where the other groups are all above .95.

Fig 3.5 - anova model residual after power 6 transformation

Fig 3.6 - DHARMa simulation of the anova model

The simulation showed mostly normal distribution of the residuals

c. *How do these Californian vaccination levels compare to U.S. vaccination levels (recent years only)? Note any patterns you notice.*

Fig 4.1 - Violin plot on 4 types of vaccination rate in the CA distracts with California average(Red), California Median(Orange) and US average(Blue)

Fig 4.2 - Box plot on 4 types of vaccination rate in the CA distracts with California average(Red) and US average(Blue)

4. *Conclusion Paragraph for Vaccination Rates*

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

California have great vaccination rate in Hepatitis b, above average in MMR and Polio, and fall behind in DTP. However, for MMR and Polio, both have a mean below the US average and a median above the average. It means there are districts with lower vaccination rate that's dragging the average down. Therefore we should focus on the districts with low vaccination rates on MMR and Polio vaccine, and increase DTP vaccination rate in all districts.

Inferential Reporting

For every item below except 7, use PctChildPoverty, PctFamilyPoverty, Enrolled, and TotalSchools as the four predictors. Explore the data and transform variables as necessary to improve prediction and/or interpretability. Be sure to include appropriate diagnostics and modify your analyses as appropriate.

5. *Which of the four predictor variables predicts the percentage of all enrolled students with belief exceptions?*

Fig 5.1 - Distribution of PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools, PctBeliefExempt

Fig 5.2 - Correlation plot of PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools, PctBeliefExempt

Fig 5.3 - Correlation plot of PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools, PctBeliefExempt after log transformation

Model result isn't great, the r^2 value is too low, and we have a point in Residuals vs Leverage graphs that's way out.

VIF looks ok, but we have two variable that's over 5, but it's not larger than 10 we will leave it for now.

Fig 5.4 - Scatterplot of Family poverty and Child poverty, with indicator of potential outlier

We looked more in to the data point on index 280, the Mt. Baldy Joint Elementary school. There is a extreme ratio of percentage of child poverty verse percentage family poverty. At 45 percent, we will have 8 student out of 18 to be in child poverty while no family there are, I think this might be an outlier due to data collecting, but we cant take it out yet, more research needed. Because of the unsatisfied result, I'm going to add more columns. According to PPIC(Public Policy Institute of California), the age to be consider as a child is till 17. After some research, I will assume the data set that's provided to me is the K-12 education, therefore I'm going to mutate a new columns called `ChildPovertyEnrolled` by multiple them.

Fig 5.5 - Correlation plot of PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools, PctBeliefExempt, ChildPovertyEnrolled

Fig 5.6 - Correlation plot of PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools, PctBeliefExempt, ChildPovertyEnrolled after log transformation

So the model is not better and we have 1 points on the l leverage graph that's almost out of the 0.5 value. I took a deeper look and the percentage of the belief exempt is at 59, which is the highest in the whole dataset.

bad vif score we will need to drop some variables.

Although both model1 and model3 all have significant p-value from the test, we are going to use the first model as the R^2 is better and it's easier to interpret.

Fig 5.7 - MCMCsimulation of linear regression model 1

One more test before the interpretation, The DHARMA simulation residual looks normal.

The two significant variables are the percentage of family poverty and the number of enrolled students. They are both negatively correlated to the percent of believe exceptions. It means the more poverty in the area, less believe exception and more enrolled students less believe exception. One thing to note is that all variables are log transformed, the real changes in percentage of the believe exception will be in log. So every 1 standard deviation increase in percentage of family poverty will have 0.453 standard deviation decrease in the log(percentage of belief exception).

6. *Which of the four predictor variables predicts the percentage of all enrolled students with completely up-to-date vaccines?*

Fig 6.1 - Distribution of PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools, PctUpToDate

Fig 6.2 - Correlation plot of PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools, PctUpToDate

Fig 6.3 - Correlation plot of PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools, PctUpToDate after transformation

Model p-value is good, although the r^2 is not great, the significant variables are percentage of family poverty, enrolled students and total number of school in the districts. residuals and leverage also looks normal.

vif looks ok.

Fig 6.4 - MCMC simulation on linear model 4

DHARMa simulation residual looks normal.

There are three significant variables that predicts the percentage of up-to-date vaccines. The variables with the best predictive power is the percentage of family poverty, then the enrolled students, and the number of schools. Increasing in the percentage of family poverty and the number of enrolled students will increase the percentage of up-to-date vaccine, while increasing in total number of school will result in decrease of the percentage of up-to-date vaccine. But because the model only have a r2 of 0.11. it only explain 11% of the changes in up to date percentages.

7. *Using any set of predictors that you want to use, what's the best R-squared you can achieve in predicting the percentage of all enrolled students with completely up-to-date vaccines while still having an acceptable regression?*

Fig 7.1 - Distribution of the Percentage on Up to date vaccine

We will look at the distribution of the dependent variables first, it's very left skewed.

Fig 7.2 - Correlation plot on PctBeliefExempt, WithHepB, WithDTP, WithMMR, WithPolio, PctChildPoverty, Enrolled, PctUpToDate

In this model we took the average of three vaccines, WithDTP, WithMMR, and WithHepB as the polio is not that significant in predicting the total up to date result. I also put in Belief Exempt, Enrolled, Child Poverty and interation between enrolled and child poverty, enrolled with belief Exempt. They are not significant for this model. I put them in because it gives 0.01 increase in the Adjusted R^2. Overall we got an Adjusted R^2 value of 0.9546. The residuals vs Fitted graphs looks normal. The Q-Q plot showed there are some point that's not on the line, same as the Leverage plot where we have 186 way outside of the 1.0 standard deviation, but the line is still normal. We will need look into these point more to see if its an outlier below.

Fig 7.2 - MCMC simulation on linear model 6

The simulated residual does not look great. The test think there might be outliers. Also our dependent variables are not a normal distribution when we fit the model. This can also cause problems in the simulations.

All the variable are have a vif score under 10, we can keep them.

Fig 7.3 - Scatterplot of percentage up-to-date and average of DTP, HepB, MMR vaccination rate to show potential outlier

The point 204, Sausalito Marin City, is out side of the normal trend. From Google this place shows up as a tourist place, nothing special other then that. We need more information to find out if it's an outlier.

Fig 7.4 - Violin plot on DTP, HepB, MMR vaccination rate to show potential outlier

Point 186, Trinidad Union Elementary, showed up way off the leverage plot, therefore are classified as outlier in diagnose analysis. The point have the lowest vaccnation rate with DTP, HepB, MMR. No special information from google, need more investigation.

Model have a close to 0 p-value, it's statistically significant. The variable that has the most predictive power on up-to-date vaccination rate is the DTP HepB MMR, then we have Belief Exception. They both have a positive coefficient, so increase in DTP, HepB, MMR and Belief Exception with lead to increase in up-to-date vaccination rates.

8. *In predicting the percentage of all enrolled students with completely up-to-date vaccines, is there an interaction between PctChildPoverty and Enrolled?*

From the previous model. We can confidently say there is no significant evidence that PctChildPoverty, Enrolled, nor an interation variable between PctChildPoverty and Enrolled predicts the percentage of all enrolled students with completely up-to-date vaccines.

9. *Which, if any, of the four predictor variables predict whether or not a district's reporting was complete?*

We ran a model that used all the variables, the performance indicates model is only fitting one side of the label as the score_log is -inf. So we gonna try with less variables and see if there is a better result.

Still the same results with less variables.

The simulation residual look normal.

The model has bad residuals, it only fitted one side of the data. we are going to try with MCMC simulations.

There are three variables that did not cross 0, but it's really really close to overlay on 0, therefore we need to test with these variables alone.

The simulations showed no good results here, and we can't not make a confusion matrix because the model did not fit both side of the data. With these data and model we can conclude there is no predictor that indicates if the district had completed reporting.

10. *Concluding Paragraph*

Describe your conclusions, based on all of the foregoing analyses. As well, 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. Make sure you have at least one sentence 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.

Poverty does not have a significant correlation with vaccination, on the percentage of students in school. To improve the vaccination rate, California state should focus on DTP vaccines out of the four types. As well as on districts such as Trinidad Union Elementary, who have the lowest vaccnation rates that drags down the average. To improve on up-to-date vaccination rate, I suggest to improve on DTP, HepB, and MMR vaccination rate. These three vaccines are a significant factor that relates to the over all up to date percentages. Father investigation we can do is to work on the getting better reporting rates, or data on districts that didn't finish reporting. As well as investigate on data points that likely to be outlier. Next step in the research should also include more general student in public, not just students in school. Since this study can exlude people who didn't went to school, and it's relationship with poverty in child and family.

reference

https://www.cdc.gov/vaccines/vac-gen/whatifstop.htm