Matthew L. Pergolski

IST 772

Dr. Block

12/21/2021

**Final Exam**

**Beginning Statement**

“I produced the material below with no assistance [direct quote from IST 772 class syllabus].” Note: Homework questions from the book may have been copied/pasted into the document for both the student and viewer’s convenience. Quotes may also have been used from the class textbook.

**Final Exam Question 1**

**Question:**

1. How have U.S. vaccination rates varied over time? Are vaccination rates increasing or decreasing? Which vaccination has the highest rate at the conclusion of the time series? Which vaccination has the lowest rate at the conclusion of the time series? Which vaccine has the greatest volatility?

**Answer/Student Response:**

In general, without differencing the data (i.e., applying the diff() command), we can see a general increase trend in all attributes that associate with the 'usVaccines' data. We can confirm this through the plot() command as well as the boxplot() command as well. Although the beginning of the time period observes volatility, all vaccines appear to end at higher values compared to the beginning of the time series. At the conclusion of the time series, we can see which vaccine(s) have the highest and lowest value by using the tail() commands. The highest at the end of the interval appears to be 'DTP1' with a rate value of 98 while 'HepB\_BD' is by far the lowest with a rate value of just 64. With visually looking at the box plots for these vaccine rates, we see the 'HepB\_BD' also has the 'longer' spread of values, which suggests a more substantial volatility compared to the other variables.

Diagram, engineering drawing

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Graphical user interface

Description automatically generated

A screen shot of a computer

Description automatically generated with low confidence

**Homework Question 2**

**Question:**

2. What proportion of public schools reported vaccination data? What proportion of private schools reported vaccination data? Was there any credible difference in overall reporting proportions between public and private schools?

**Answer/Student Response:**

The following was generated and observed:

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, histogram

Description automatically generated

0.97418 (~97%) of public schools reported vaccinations. 0.8471801 (~85%) of private schools reported vaccinations. In order to generate this statistic, the ‘allSchoolsReportStatus’ data was split into two groups: public and private. From here, they were split into two more groups – those who reported vaccinations and those who did not. The number of schools (for both public and private, respectively) that reported vaccinations were divided by the total number of schools (public or private, respectively). Please reference the R-code script for additional information.

The observed p-value from the frequentist t-test provides a value of 2.2e-16, which suggests

statistical significance. The confidence interval also ranges from 0.1091376 to 0.1448623.

Although this frequentist confidence interval is a 'long-term' solution (which may or may not apply

to this specific time the test was ran), we do not see that this interval spans across zero, which

also gives us confidence that we can reject the null hypothesis. There does seem to be a credible

difference in overall reporting between public and private schools. In order to run the BESTmcmc() command the dataset was trimmed and sampled. From the results of the Bayesian test, we see a mean difference that does unfortunately span across zero, with the 95% HDI showing ~-6 on the lower bound and ~6 on the upper bound, as shown above.

I believe the trimmed dataset unfortunately may have skewed the results, so the frequentist perspective will be viewed as more legitimate, with statistical significance over the smaller mean difference. Regardless, the difference does not seem to be very large.

**Homework Question 3**

**Question:**

3. What are 2013 vaccination rates for individual vaccines (i.e., DOT, Polio, MMR, and HepB) in California public schools? How do these rates for individual vaccines in California districts compare with overall US vaccination rates (make an informal comparison to the final observations in the time series)?

**Answer/Student Response:**

The following code was generated and observed:

Text

Description automatically generated

Text

Description automatically generated

A merge took place of the allSchoolsReportStatus and the districts dataset. This merge took place in order to identify vaccination rates for 2013 public school districts. From here, since the rates are categorized as a percentage of those who did not receive the vaccines, the rates were ultimately subtracted from 100 to provide a rate of those who did get vaccinated. We see that the vaccination rate, on average, was 91.56842% for DTP in 2013 public schools, 91.37544% for MMR in 2013 public schools, 91.37544% for MMR in 2013 public schools, and 93.13684 for HepB in 2013 public schools. These rates are comparableto the ending of the dataset except for HepB\_BD -- where the average of HepB is significantly greater than the last remaining value in the usVaccines dataset for HepB.

**Homework Question 4**

**Question:**

4. Among districts, how are the vaccination rates for individual vaccines related? In other

words, if students are missing one vaccine are they missing all of the others?

**Answer/Student Response:**

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Among the districts, when looking at a correlation matrix, we see high degrees of correlation between the different attributes variables. Since these variables are categorized as those who did not receive the vaccine, this suggests that, if one student is to miss the DTP vaccine, for example, it is highly likely they will also miss the Polio vaccine (correlation value of 0.9803893). If we dive deeper into this example of DTP and Polio, we find that both the frequentist approach as well as Bayesian point to similar results (that the variables are highly correlated). We see a p-value for the frequentist approach come out to be 2.2e-16, which suggests statistical significance when comparing it to our conventional threshold of 0.05. When performing the Bayesian test, we see a rhoNot0 value of 4.937621e+489 to 1, which is extremely high. If we remember our standard for determining strong odds for this test, we would indicate anything greater than 150 to 1 would be strong odds. To reiterate, although we did just two vaccines in this example, we can see other correlation measures in the standard cor() command.

**Homework Question 5**

**Question:**

5. What variables predict whether or not a district's reporting was complete?

**Answer/Student Response:**

The following as generated and observed:

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated with medium confidence

When comparing all variables to the 'DistrictComplete' attribute, we see that 2 variables have statistical significance and may be sufficient predictors: (1) 'Enrolled' (p-value = 8.24e-11) and (2) 'TotalSchools (p-value = 2.20e-12).' The intercept is also significant with a p-value at 0.006. Although the overall p-value shown is 2.2e-16, we see that the R and R-squared values are around 0.1, which does not suggest a strong model.

**Homework Question 6**

**Question:**

6. What variables predict the percentage of all enrolled students with completely up-to-

date vaccines?

**Answer/Student Response:**

The following was generated and observed:

Graphical user interface, application, website

Description automatically generated

Text

Description automatically generated with medium confidence

Table

Description automatically generated with low confidence

When comparing all variables to the 'PctUpToDate' attribute, we see that 3 variables have statistical significance and may be sufficient predictors: (1) 'WithoutDTP' (p-value = 6.41e-14) and (2) 'WithoutMMR (p-value = WithoutMMR), (3) 'PctBeliefExempt' (p = 2.86e-11). The intercept is also significant with a p-value at 2e-16. For the the R and R-squared values, we see values of are around 0.94, which suggests a relatively strong model. When running the Bayesian version of this test on this predicting variables, we see similar results and do not see the HDI span across zero for any of the chosen attributes in the summary(bayesian.model.6) function.

**Homework Question 7**

**Question:**

7. What variables predict the percentage of all enrolled students with belief exceptions?

**Answer/Student Response:**

The following was generated and observed:

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated with low confidence

A picture containing table

Description automatically generated

When comparing all variables to the 'PctBeliefExempt' attribute, we see that 4 variables have statistical significance and may be sufficient predictors: (1) 'WithoutDTP' (p-value = 0.00116), (2) 'WithoutHepB (p-value = 2e-16), (3) 'PctUpToDate' (p = 2.86e-11), and (4) 'PctFreeMeal' (p = 0.00402). The intercept is also significant with a p-value at 1.91e-11. For the R and R-squared figures, we see values of 0.84/0.85, which suggests a relatively strong model. When running the Bayesian version of this test on this predicting variables, we see similar results and most of the variables do not have an HDI that span across zero in the summary(bayesian.model.7) function. It is important to note that we do see the 'PctFreeMeal' variable have an HDI of -0.01943 on the lower bound and 0.001079 on the upper bound, which unfortunately spans across zero – this gives us less confidence that this is a true predictor for the 'PctBeliefExempt' attribute.

**Homework Question 8**

**Question:**

8. What's the big picture, based on all of 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. What have you learned from the data and analyses that might inform this

question?

**Answer/Student Response:**

All in all, we see that vaccine rates are generally increasing over time, which is positive news for the public as a whole in addition to the legislator’s office. Although most vaccines have rates in the 90s, there is one, ‘HepB’ that unfortunately is around the rate of ~ 60. The legislator should strive to continue these upward trends and achieve a 100% vaccination rate for all vaccines.

A place to focus on in terms of overall vaccine rate is increasing private schools who report their aggregate vaccination status. Although ~97% of public schools reported vaccinations, it was discovered that only ~85% of private schools reported their rates. If we gain additional information in these reporting figures, we may find that vaccination rates are higher overall. An overall statistical significance was found in terms of the difference in reporting data for public versus private schools, with a p-value equating to 2.2e-16. To reiterate, we should attempt to collect and report as much data as we can – the more data, the better and more concrete our overall results will be.

Moreover, when taking into consideration the analysis that was done on the ‘districts’ dataset, we generally found high correlation between the various vaccines that were represented in the data, and found that some may also be good predictors. To share a few examples from the above analysis, we observed statistical significance when running our linear models and identifying ‘PctBeliefExempt’ versus the ‘WithoutHepB’ variable – as well as the ‘PctBeliefExempt’ ~ ‘PctFreeMeal variables. Generally when running Bayesian techniques on these tests, we also found most had a 95% HDI that did not span zero, which only strengthened our beliefs about the validity of our results. When conducting a Bayesian (bfCorTest) test, the correlated results were found to be 4.937621e+489:1, which is exceptional odds by our standard (i.e., results greater than 150:1 is considered ‘strong’ odds). As a reminder, this customer function was introduced in the IST 772 class textbook that allowed us to find an Bayesian developed odds figure that would help place results in perspective in terms of statistical significance. The full function is defined as follows:

bfCorTest <- function (x,y) #Get r from BayesFactor

{

zx <- scale(x) #Standardize X

zy <- scale(y) #Standardize Y

zData <- data.frame(x=zx,rhoNot0=zy) #Put in a data frame

bfOut <- generalTestBF(x ~ rhoNot0, data=zData) #linear coefficient

mcmcOut <- posterior(bfOut,iterations=10000) #posterior samples

print(summary(mcmcOut[,"rhoNot0"])) #Show the HDI for r

return(bfOut) #Return Bayes factor object

I believe enough information was discovered that allows us to believe that, generally, if a student misses one vaccine, they are more likely to miss other vaccines as well (even if only a few others). Since there appeared to be statistical evidence that suggested those children who qualified for free meals are likely to miss (some) vaccines, we may be able to infer that this socio-economic group may be struggling to prioritize health vaccinations due to financial limitations.

Because of the above, financial assistance should be provided to these families so they can begin to prioritize vaccinations within their children. To be brief, educational programs and other forms of information (with an emphasis on financial assistance) may help increase vaccination rates for those belonging to this specific group. Similar to COVID-19 vaccines now, more should be allocated to ensure vaccinations in children increase (while taking the above information into account).

In terms of which variables predict whether or not a district's reporting was complete, the following were found to be statistically significant: ‘Enrolled’ and ‘TotalSchools’. Subsequently, ‘WithoutDTP,’ ‘WithoutMMR,’ and ‘PctBeliefExempt’ were found to predict ‘PctUpToDate.’ Lastly, ‘WithoutDTP,’ ‘WithoutHepB,’ ‘PctUpToDate,’ and ‘PctFreeMeal’ were all found to be statistically significant in predicting ‘PctBeliefExempt.’

As more data becomes available, we hope to learn more insights on how vaccination data within the ‘districts’ dataset is related.