Analysis of COVID-19 Testing Data within Institutions of Higher Education in the United States

An In-depth Bayesian Hierarchical Beta-Binomial and Regression Model Application

Yicheng Shen, Yucheng Yang, Mitchell Wang

11/15/2021

Abstract

In this project, we are interested in modelling the positivity rate of COVID in higher education institutions in the U.S. Based on our Bayesian hierarchical Beta-Binomial and regression models, we found that Carleton in 2021 has done a better job controlling the spread of COVID on campus compared with 2020, when the majority of students first began to return. The level of COVID threats in the surrounding area, particularly Rice county, is a significant predictor associated with changes in the number of COVID cases within Carleton, whereas temperature is less relevant in explaining any rise or fall of positive cases on campus. Our findings further suggest that for all higher education institutions, the institution's total enrollment, test rate (number of tests administrated divided by the enrollment) and type of the institution (either being a liberal arts college or a university) are significant predictors affecting the positive rates of COVID testing. A larger student body, less frequent testing and being a university can all be associated with higher COVID positive rates on campus.

Keywords: Bayesian Inference; Markov chain Monte Carlo Methods; COVID-19; Public Health among Higher Education Institutions; Just Another Gibbs Sampler (JAGS)

Introduction

Much of the year 2021 has been featured with mourning grief and crucial challenges as governments and scientists around the world continue to study how to battle the COVID-19 pandemic and restore citizens' hope from the enduring health threat. However, the insufficiency of public health resources and concerns of community infections persist as we walk towards the third year since the initial outbreak of the pandemic. As of November 12, 2021, the CDC reports that the total positive coronavirus cases in the United States have exceeded 46.8 million with nearly 760,000 related deaths. These alarming statistics constantly remind us that this pandemic is far from being over given the current situation.

Meanwhile, tens of thousands of new cases have emerged on college campuses according to the New York Times, with most of the on-campus cases announced since students returned to campus for the fall term in 2020. Despite usually having high vaccination rates, mask mandates and social distancing requirements, these densely populated campus communities are still vulnerable to this highly contagious disease. As more and more schools began to actively test their population and publish relevant data, it is not only feasible but also imperative that we conduct proper analysis on the spread of COVID within campus communities.

In this project, we are interested in exploring whether Carleton College has effectively contained the infectious disease and protected its campus population from the pandemic, using the weekly testing numbers and positive cases published by Carleton's COVID-19 working groups and committees since the beginning of fall term in 2020. Based on our Bayesian hierarchical Beta-Binomial and regression models, we found that Carleton in 2021 has done a better job controlling the potential spread of COVID on campus compared with 2020, when the majority of students first began to return to campus. The level of COVID threats in the surrounding area, particularly Rice county, is a significant predictor associated with changes of COVID cases within Carleton, whereas temperature is less relevant in explaining any rise or fall of positive cases on campus. We further applied our methodologies to analyze a larger sample of 30 higher education institutions in the U.S. Our findings suggest that the institution's total enrollment, test rate (number of tests administrated over total enrolled students) and type of the institution (either being a liberal arts college or university) are significant predictors affecting the positive rates of COVID testing. A larger student body, less frequent testing and being a university can all be associated with higher COVID infection rates on campus.

Data

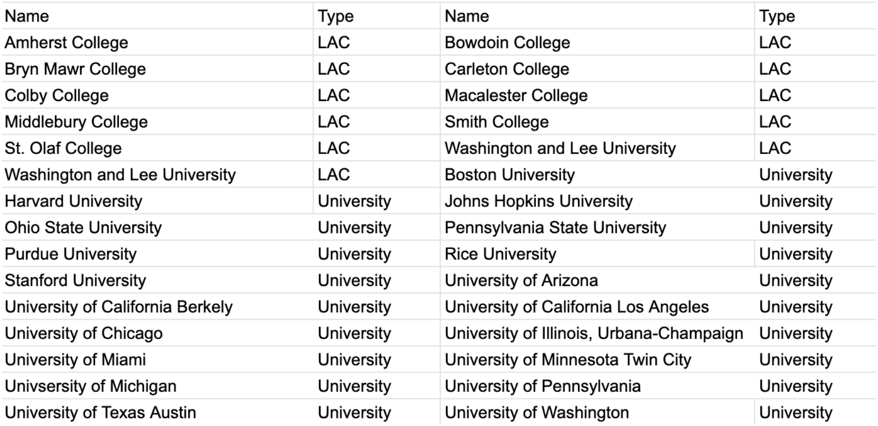
As the public health impacts of colleges reopening has become apparent, there has been a substantive move towards greater transparency of COVID testing and reports. The most notable efforts of ensuring information transparency are presented by the COVID dashboards, now maintained by a significant number of institutions to publish their COVID policies, statistics and future plans. The best dashboards, according to the We Rate COVID Dashboards rating scheme, are updated at least once every weekday and include information not only about the positive number of cases but also about the total number of COVID tests conducted and the frequency of testing. These online resources provide us with detailed and easily accessible data that can be used for our model fitting and analysis.

To obtain a full picture of the COVID positive rates among the U.S. higher education institutions, we extracted and complied data from the COVID dashboard of 30 institutions including universities and liberal arts colleges, including Carleton College, our very own campus. For Carleton College, the dashboard publishes 57 weeks of COVID testing data across four terms. The earliest record started on August 9th, 2020 and the latest one was on November 12th, 2021. The data set contains the total number of tests conducted each week, the number of positive and negative tests, and the positive rates within that 7-day period. Among the 57 weeks of records, we have 39 weeks when classes were in session and the college had at least 80% of students on campus, and the rest 18 weeks were summer, winter or spring breaks when most of the student and faculty population were absent.

As full-time college students, our primary interests are to better understand our college's COVID responses and conditions during academic terms, when the majority of the students spend most of the time on campus. Therefore, our analysis would focus on the data from four terms: 10 weeks each for 2020 fall, 2021 winter and 2021 spring, and 9 weeks for 2021 fall term (when the latest record ends).

When looking into a broader population of U.S. colleges and universities, our sampled institutions are largely determined by the availability of clear, published and comprehensive records. Since it is the graduate school application season for college seniors, our group complied a list of potential institutions that we are interested in applying or learning more about so that the project could provide us with more relevant insights. Based on whether their COVID dashboards are clear and usable, we selected 30 institutions in total, consisting of 19 universities and 11 liberal arts colleges (LAC). These universities and colleges generally have similar length of records as Carleton, starting from summer of 2020 until November of 2021. The full list of our sampled institutions are shown in Table 1. Since schools have various reporting standards and different lengths of semesters or trimesters, it is hard to cross compare term-specific statistics between these institutions. For each institution, we recorded the total number of COVID tests conducted and the total positive cases reported since the summer of 2020. We also searched for their total enrollment this fall term as collected by the U.S. News, representing their population size.

Table 1. List of American Higher Education Institutions analyzed in this Project



Name	Type	Name	Type
Amherst College	LAC	Bowdoin College	LAC
Bryn Mawr College	LAC	Carleton College	LAC
Colby College	LAC	Macalester College	LAC
Middlebury College	LAC	Smith College	LAC
St. Olaf College	LAC	Washington and Lee University	LAC
Washington and Lee University	LAC	Boston University	University
Harvard University	University	Johns Hopkins University	University
Ohio State University	University	Pennsylvania State University	University
Purdue University	University	Rice University	University
Stanford University	University	University of Arizona	University
University of California Berkely	University	University of California Los Angeles	University
University of Chicago	University	University of Illinois, Urbana-Champaign	University
University of Miami	University	University of Minnesota Twin City	University
Univsersity of Michigan	University	University of Pennsylvania	University
University of Texas Austin	University	University of Washington	University

In addition to COVID testing numbers and positive cases, our project takes other factors into considerations when building models. In the case of Carleton College, we think our adjacent community could be quite influential, since there have been repeated incidents in which students were infected outside of campus, either when they visited downtown Northfield or other nearby areas. So we found the weekly records of COVID cases reported by Rice Country (where Carleton is situated in) Department of Public Health (2021). Dr. Prince Allotey's recent research also shed some lights on our model by pointing out that temperature can be a significant factor affecting the infection and mortality rates of COVID. The National Oceanic and Atmospheric Administration under the U.S. Department of Commerce (USDOC, 2021) provides daily weather service at the county level. We therefore selected the weekly average temperature available in the data set.

Methods

Our project relies primarily on Bayesian statistical methods and Just Another Gibbs Sampler (JAGS), which is a algorithm for simulating from Bayesian hierarchical models using Markov chain Monte Carlo method, developed by Martyn Plummer (Plummer, 2003).

As college seniors who have been used to online and hybrid learning mode since the pandemic, we are relatively uninformed of the level COVID spread over the past year, specifically the details and exact statistics among and beyond our communities are not always fully understood. Therefore, we consider ourselves possessing very weakly informative or diffused prior belief on this subject.

In both within Carleton and between institutions scenarios, we recognize that there could be nested group structures in our data. Observations within each group may be correlated, and knowing the case of one period or one school can tell us more information about the others. We also want to be able to generalize to future scenarios and broader populations based on our sampled data. Consequently, we decide to try a hierarchical Beta-Binomial model structure to analyze the rate of positive cases among total tests conducted.

Model Specification for Carleton COVID Dashboard Analysis

When using Beta-Binomial models to model and evaluate the positive COVID test rates within Carleton, we assume that the number of people tested as positive follows a binomial distribution with each test being an independent Bernoulli trial. There also exist other potential factors that could affect the positive test rate of COVID test in Carleton and the relationship between the positive test rate and these factors are linear. The factors considered include the weekly temperature of Rice County and the weekly positive cases of COVID in Rice County. These variables are included in our model as regressors in the link function of positive test rate. In the model equations shown in the following sections, an asteroid on the side of a variable denotes that this variable has been standardized when fitting the model.

For the first model (Hierarchical), we assume that the number of positive COVID cases $Y_{i,k}$ follows a Binomial distribution with probability θ_k and trials m_i, where θ_k represents the chance of a COVID test administrated in Carleton came back positive and m_i represents the number of tests administered in the corresponding week. The subscript k corresponds with one specific term of Carleton (Fall 2020, Winter 2021, Spring 2021, Fall 2021) and the subscript i corresponds with one specific week (39 weeks in total). We assume that the four θ_k follow the same Beta distribution with parameters α, β, which are constructed using the estimated mean μ and sample size η. We gave a weakly informative prior of Beta(1,1) to μ and a weekly informative prior of the Logistic(log(100),1) to η which indicates that a priori about the shrinkage $\lambda = \eta/(\eta + 100)$ is uniformly distributed on (0, 1).

- Sampling (likelihood): for i in 1,. . . ,39 and k in 1, 2, 3, 4:

$$y_{i,k}|\theta_k, m_i \overset{i.i.d.}{\sim} \binom{m_i}{\theta_k}$$

- Prior for θ_k, for k= 1, 2, 3, 4:

$$\theta_k|\alpha, \beta \sim Beta(\alpha, \beta)$$

- Hyperprior

$$\alpha = \mu\eta;\ \ \beta = (1-\mu)\eta$$

$$\mu \sim Beta(1,1);\ \ log(\eta) = Logistic(log(100), 1)$$

With our (weakly informative) priors, we update the belief with the sampled data to obtain our posterior:

- Posterior

$$
\begin{aligned}
\pi(\theta_k|Y_{i,k},\mu,\eta) &\propto Prior \times Likelihood \\
&\propto \pi(\mu,\eta) \times L(\theta_k|Y_{i,k}) \\
&\propto \pi(\mu,\eta) \times \prod f(Y_{i,k}|\theta_k, m_i) \\
&\propto \pi(\mu) \times \pi(\eta) \times \prod f(Y_{i,k}|\theta_k, m_i)
\end{aligned}
$$

For the second model (Regression), we used the logistic regression as a link function on the probability, and the regressors included are the weekly temperature of Rice County and the weekly positive COVID cases in Rice County. We denote the intercept as β_0, the coefficient for standardized weekly positive COVID cases in Rice County as β_1, the coefficient for standardized weekly temperature of Rice County as β_2. We gave weakly informative prior to the coefficients.

- Sampling: for i in 1,. . . ,39:

$$
y_i|\theta_i, m_i \overset{i.i.d.}{\sim} \binom{m_i}{\theta_i}
$$

$$
logit(\theta_i) = log\frac{\theta_i}{1-\theta_i} = \beta_0 + \beta_1 \times rice_i^* + \beta_2 \times temp_i^*
$$

$$
\theta_i = \frac{exp(\beta_0 + \beta_1 \times rice_i^* + \beta_2 \times temp_i^*)}{1 + exp(\beta_0 + \beta_1 \times rice_i^* + \beta_2 \times temp_i^*)}
$$

- Prior for θ_i, for i = 1,. . . ,39:

$$
\beta_0 \sim \mathcal{N}(0,\, 100)
$$

$$
\beta_1 \sim \mathcal{N}(0,\, 100)
$$

$$
\beta_2 \sim \mathcal{N}(0,\, 100)
$$

- Posterior

$$
\begin{aligned}
\pi(\beta_0,\beta_1,\beta_2|y_1,...,y_i,m_1,...,m_i) &\propto \pi(\beta_0,\beta_1,\beta_2) \times L(\beta_0,\beta_1,\beta_2|y_i,m_i) \\
&\propto \pi(\beta_0)\pi(\beta_1)\pi(\beta_2) \times \prod f(y_1,...,y_n|\beta_0,\beta_1,\beta_2,m_i)
\end{aligned}
$$

Model Specification for American Institutions of Higher Education

We decide to use Beta-Binomial models to study the positive COVID test rates across these institutions. In our setting, we assume that the number of people tested as positive follows a binomial distribution with the total number of tests administrated denoted as n_j; each test will be one independent Bernoulli trial with probability of p_j as the chance of one COVID test administrated in these institutions came back positive. The subscript j specifies that these parameters are institution-specific and j itself correspond to institution index number in our sample. We think that there exist potential factors that affect p_j and the relationship between p_j and these factors are linear as shown in Figure 4 in the Results section. The factors we considered include the institution type (university versus LAC), institution enrollment, and the test rate (calculated as the total number of tests administrated from 2020 to 2021 divided by enrollment for each institution). We also assume that there exists some baseline positive test rate for all educational institutions in the US. In this section, we are interested in studying how well the educational institutions in the US are dealing with the COVID pandemic by investigating p_j across these different schools.

For the first model (Model 1), we used the logit function as a link function on p_j, which is a simple linear regression including an intercept, the institution type, the standardized enrollment, and the standardized test rate. We denote the intercept as β_0, the coefficient for institution type variable as β_1, the coefficient for the standardized enrollment as β_2, and the coefficient for the standardized test rate as β_3. For all the coefficients mentioned above, we assume they all follow a weakly informative prior $N(0, 100)$.

- Below are the equations for Model 1:

- Sampling: for j in 1,. . . ,30:

$$Y_j | p_j, n_j \overset{i.i.d.}{\sim} \binom{n_j}{p_j}$$

$$logit(p_j) = \beta_0 + \beta_1 \times Type_j + \beta_2 \times Enrollment^*_j + \beta_3 \times TestRate^*_j$$

- Prior for $\mathbf{p}_j$, for j= 1,. . . ,30:

$$\beta_0 \sim \mathcal{N}(0,\ 100); \beta_1 \sim \mathcal{N}(0,\ 100)$$

$$\beta_2 \sim \mathcal{N}(0,\ 100);\ \beta_3 \sim \mathcal{N}(0,\ 100)$$

- Posterior

$$\pi(\beta_0, \beta_1, \beta_2, \beta_3 | y_1, ..., y_j, n_1, ..., n_j) \propto \pi(\beta_0, \beta_1, \beta_2, \beta_3) \times L(\beta_0, \beta_1, \beta_2, \beta_3 | y_1, ..., y_j, n_1, ..., n_j)$$

$$\pi(\beta_0, \beta_1, \beta_2, \beta_3) = \pi(\beta_0)\pi(\beta_1)\pi(\beta_2)\pi(\beta_3)$$

In the second model (Model 2), we assume that p_j follows a Beta distribution with parameters a_j and b_j. We will use reverse elicitation to make posterior inferences on a_j and b_j. We define the mean of the Beta distribution as $\mu_j = \frac{a_j}{a_j + b_j}$ and the sample size as $\eta_j = a_j + b_j$. Then we use the logit function as a link function on μ_j with regressors as the intercept, the institution type, and the test rate. Similarly, we denotes the intercept as β_0, the coefficient for institution type as β_1, and the coefficient for test rate as β_3. Again, the coefficients mentioned here all follow a weakly informative prior $N(0, 100)$.

- The equations for Model 2 are shown below:

- Sampling: for j in 1,. . . ,30:

$$Y_j|p_j, n_j \overset{i.i.d.}{\sim} \binom{n_j}{p_j}$$

$$p_j|a_j, b_j \sim Beta(a_j, b_j)$$

- Prior for p_j, for j= 1,. . . ,30:

$$a_j = \eta_j \mu_j;\ b_j = \eta_j(1 - \mu_j)$$

$$logit(\mu_j) = \beta_0 + \beta_1 \times Type_j + \beta_3 \times TestRate_j$$

$$\eta_j = exp(log\ \eta_j)$$

$$log\ \eta_j \sim Logistic(logn, 1)$$

- Hyperprior

$$\beta_0 \sim \mathcal{N}(0,\ 100)$$

$$\beta_1 \sim \mathcal{N}(0,\ 100)$$

$$\beta_3 \sim \mathcal{N}(0,\ 100)$$

$$logn = log(100)$$

- Posterior

$$\begin{aligned}
\pi(p_j|y_j, \mu, \eta) &\propto Prior \times Likelihood \\
&\propto \pi(\mu_j, \eta_j) \times L(p_j|y_1, ..., y_j) \\
&\propto \pi(\mu_j, \eta_j) \times L(\beta_0, \beta_1, \beta_3|y_1, ..., y_j, n_1, ..., n_j) \\
&\propto \pi(\mu_j, \eta_j) \times \prod f(y_j|\beta_0, \beta_1, \beta_3, n_j) \\
&\propto \pi(\mu_j) \times \pi(\eta_j) \times \prod f(y_j|\beta_0, \beta_1, \beta_3, n_j)
\end{aligned}$$

For all models, we made an additional 10000 draws for every chain after an adaptation period of 1000 draws and a burn-in period of 5000 draws, and we kept every 5th draws to reduce the effect of temporal correlation between consecutive MCMC draws. For convergence consistency, we ran 3 chains for both models. We checked convergence and efficiency using trace plots and autocorrelation plots discussed in sections below.

Results

Carleton Covid Dashboard Analysis

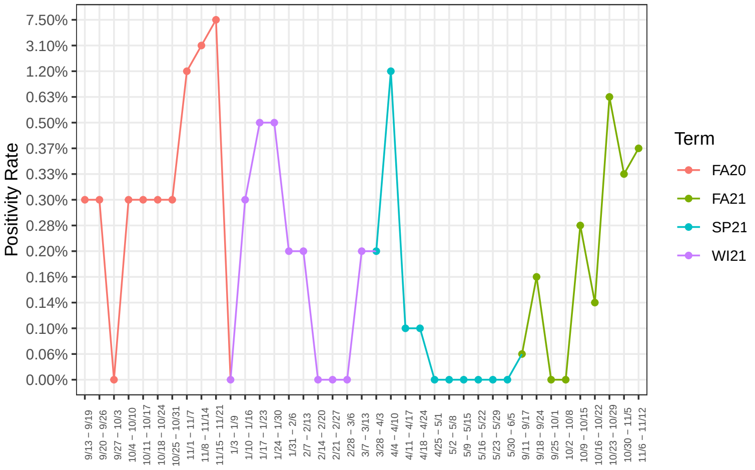


Figure 1. Carleton COVID Positivity Rate by Week

From Figure 1, we can see the that the positive rate remains around 0.5% for all the weeks except for a spike at around November 2020 where the positivity rate reached more than 6.5% and a spike at around April 2021 where the positivity rate reached 1%. As we can see in Figure 2, there is an adequate relationship between Carleton's COVID cases with Rice County cases at a weekly basis. However, the relationship between Carleton's cases and the temperature does not seem to be obvious.

In order to adjust variables' scales and make the coefficients for the regression model easier to interpret, we chose to standardize both the weekly number of COVID cases in Rice County and the weekly average temperature of Rice County. Specifically, we performed the standard procedure of taking the z-score of each observation.

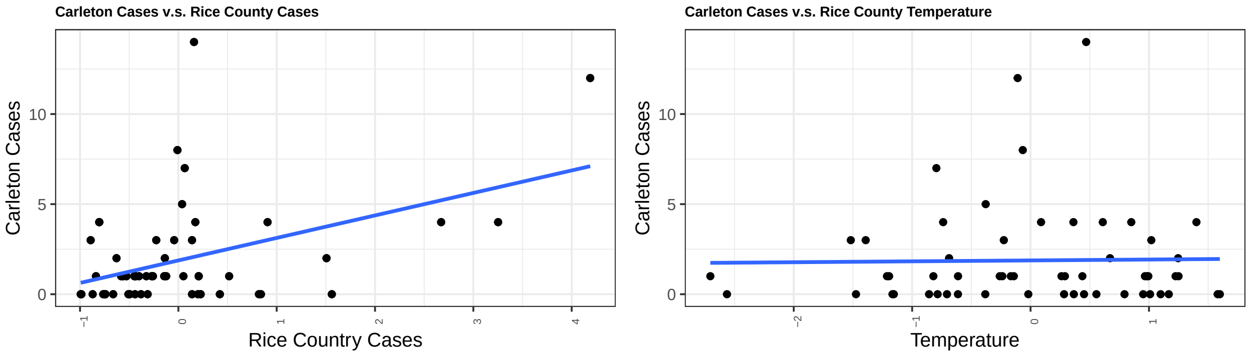


Figure 2. The correlation of Carleton COVID cases between temperature and Rice county cases. Both temperature and Rice county cases are standardized.

For both the Hierarchical Model and the Regression Model, the trace plots and ACF plots have all shown that our draws are well-mixed and our parameters have converged to their posterior regions. The overlaid density plots have shown that the draws from all three chains all converged to the same distribution with similar density curves, for both of the models. The posterior predictive draws also suggest the models are adequate, as the observed mean lies in the center of the distribution of the simulated data mean.

However, for the Regression Model, the residual plot indicates that there is one potential outlier that affected the constant variance assumption. More specifically, it's the 10th week of Fall term in 2020 that had an unusually high positive test rate. Considering the limited data we have, we chose to proceed with our analysis and acknowledging the fact that our posterior inferences would be more conservative accordingly.

For the Hierarchical Model, given our priors and data, there is a 95% chance that the ratio of Fall 2021's positivity rate over Fall 2020's positivity rate is within the range from 0.218 to 0.652, indicating that the Carleton is doing better in terms of controlling COVID in Fall 2021 comparing to Fall 2020. Additionally, for Fall 2020, the positivity rate is significantly higher than the rest of the terms as shown in Figure 3.

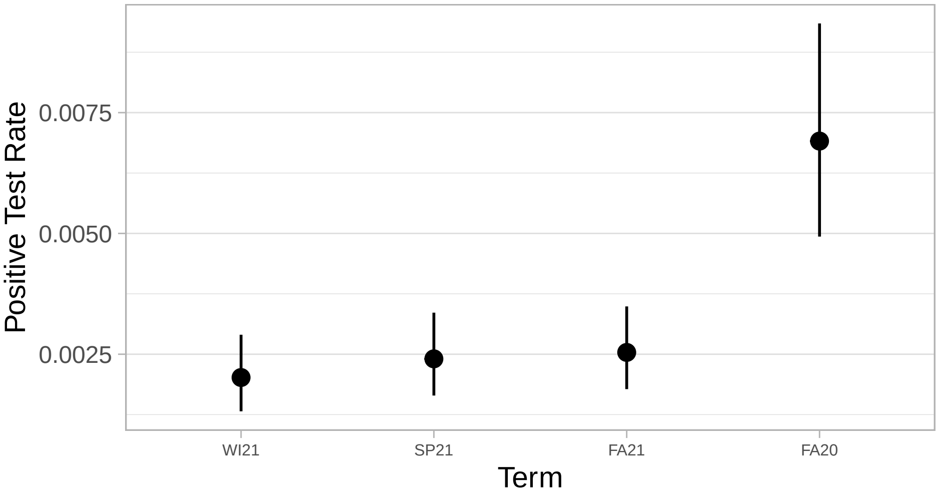


Figure 3. The Posterior Predictive Distribution of COVID Positivity Rate during different terms at Carleton.

Looking at the Regression Model, we know that for one standard deviation increase of standardized Rice County Cases, the expected percentage increase in the odds of a Carleton COVID test is 104% and there is a 95% probability that the increase of odds will range from 81.87% to 128.51%, given the prior and data. For one standard deviation increase of standardized Rice County Temperature, the expected percentage change in the odds of a Carleton COVID test is -2% and this percentage change of odds will be from 21% lower to 20% higher with a 95% probability.

Covid Trend in American Institutions of Higher Education

Figure 4 below shows that there exist a medium negative correlation between the test rate and positivity rate and a medium positive correlation between the positivity rate and enrollment. Positivity rate is defined as total number of positive tests divided by the total number of tests administrated.

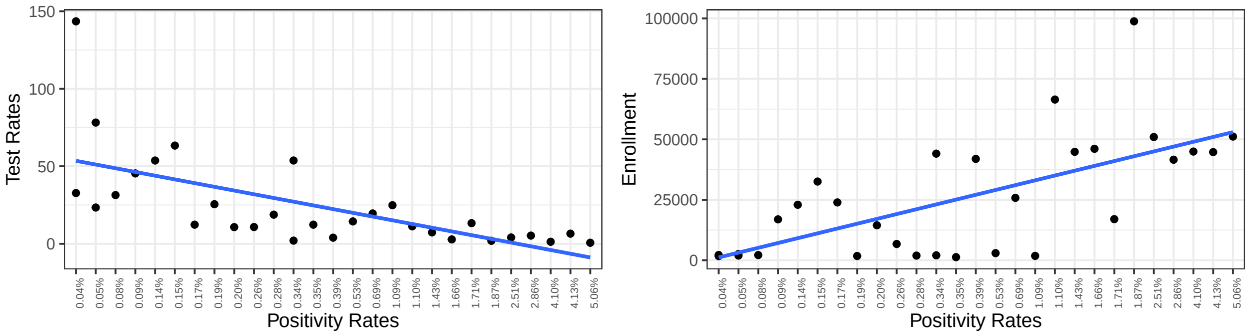


Figure 4. The correlation of observed average probability of tested positive for COVID between test rates and enrollment.

Looking at Model 1, we know that if the enrollment number and test rate are at their average level among these institutions, the positive test rate of a university is 0.0142 (β_0); for a LAC, this baseline level could decrease by an average of 50.20% (β_1). Looking at the credible intervals for these coefficients, we also know there is a 95% chance that the baseline positive test rate β_0 for university is between 0.0140 to 0.0143; for β_1, we know that the baseline positive test rate of LAC is from 47.27% to 53.02% lower than the baseline positive test rate of a university.

Holding the institution type and standardized test rate constant, with every one standard deviation increase in the standardized enrollment, there will be a 16.03% increase in the odds of one COVID testing being positive (β_2), and we know that this increase will range from 14.97% to 17.10% with a 95% probability. Holding the institution type and standardized enrollment constant, with every one standard deviation increase in the standardized test rate, there will be a 72.29% decrease in the odds of one COVID testing being positive (β_3), and we know that this increase will be from 71.93% to 72.65% with a 95% probability.

Using Model 2, we know that if the test rate is at the average level among these institutions, the baseline probability of one test being positive for university is 0.0451 (β_0), this baseline level could reasonably range from 0.0301 to 0.0700 with a 95% probability. For LAC we will see an average of 66.32% decrease in the odds of one COVID testing being positive (β_1), this decrease could be from 23.37% to 84.00% with a 95% probability. Holding the institution type still, with every increase of 1 in the test rate variable, there will be a 2.89% decrease in the odds of one COVID testing being positive (β_3). This decrease could range from 1.21% to 4.60%, with a 95% possibility.

For Model 1, the trace plots and ACF plots show that our draws are well-mixed and our parameters have converged to their posterior regions. The overlaid density plots show that the draws from three separate chains all converged to the same distribution with roughly the same density curves. However, the posterior predictive check indicates that Model 1 fails to generate data that are similar to the observed data. The residual plot of Model1 shows that its residuals could potentially be non-normally distributed and we should be concerned with whether the way we specified our model satisfied the assumptions of the simple linear regression model.

For Model 2, we also see that our draws are well-mixed and our parameters have converged to their posterior regions from checking the trace plots and ACF plots. The overlaid density plots also show that the draws from three separate chains all converged to the same distribution with roughly the same density curves. Similarly, we also conducted the posterior predictive check on Model 2. The 100 simulated data have a very similar density curve to the density curve of the observed data. Checking the posterior estimates of the mean number of positive tests for each of those institutions confirmed the fact that our model is able to generate data that mimic the observed data, and therefore we will be able to conclude that Model 2 is adequate and is preferable to Model 1. Looking at the residual plot of Model 2, we should be able to say that the residuals are normally distributed and the model assumptions are met.

Comparing the DIC stats of Model 1 and Model 2, we also get to conclude that Model 2 is a better fit to our data and should be able to give us more reasonable posterior inferences than Model1.

The posterior inference made on p_j using Model 2 is shown in Figure 5 below:

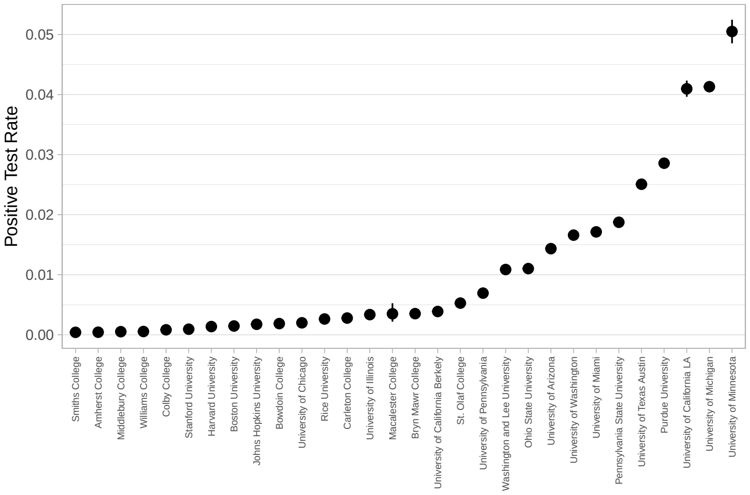


Figure 5. Posterior Inference of the chance of one test being positive across institutions from Model 2

Discussion & Conclusion

The Hierarchical model and the Regression model provide valuable insights into how Carleton has handled COVID from 2020 to 2021. The Hierarchical model reflects a significant decrease in the possibility of one COVID being positive in Fall Term 2021 compared with the previous year, which is mainly contributed by the spike of positive cases a the end of 2020. In the Regression model, the standardized Rice County COVID cases are shown to be significant in terms of predicting Carleton's COVID cases. Specifically, one standard deviation increase in the Rice County COVID cases is correlated with a 104% increase of the odds of a Carleton COVID test being positive. This is congruent to the fact that as Carleton is located in Rice County, a lot of transmission came from the residents and workers in Rice County. Contrary to our expectation, the temperature is not influential to the COVID situation in Carleton community. The residual plot from the Regression model indicates our model's soundness is affected by one potential outlier. Potential ways to further improve our models include getting a larger number of weekly records and quantifying Carleton's COVID policy as a variable.

The latter two models regarding COVID cases on a wider range of American institutions reveal more information about school-to-school difference in coping with COVID. In specific, we could see that more populated universities with lower test rates usually have a higher possibility of getting positive result from one COVID test than those smaller liberal art colleges that conduct more frequent tests over their populations. The interesting finding is that although from looking at the posterior predictive distribution, Carleton College has a relatively low COVID positive rate among 30 institutions, this performance is in fact below average among the 11 liberal arts colleges in our sample, suggesting that Carleton still has much to work on when comparing with its peers.

Our models have limitations that should be well acknowledged. For example, certain assumptions for linear regression models are not well satisfied with limited data and the existence of extreme outliers. Despite having adequate effective sample sizes, the amount of available data we have access to is still limited due to various reporting standards, thus our conclusion is not strongly generalizable when applying to more institutions. Some variables we selected, for example, the number of enrollment and test rates, are correlated with each other, raising the issue of collinearity. We could also choose more informative priors when setting up the models, since the simple linear regression models are not robust to the choice of hyperpriors.

In summary, our project provides a positive answer to the research question about whether Carleton handled COVID well, with its positive rates decreasing over the terms. We also identified that the situation in the surrounding area is a significant predictor for COVID cases within Carleton. When put into a bigger context, Carleton and other liberal art peers have done a better job overall than larger universities at testing and controlling the spread of COVID, although we should acknowledge that some liberal arts colleges have outperformed Carleton along this process.

Reference

Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd international workshop on distributed statistical computing (Vol. 124, No. 125.10, pp. 1-10).

Rice County Department of Public Health. (2021, November 18). Current situation. Current Situation | Rice County, MN. Retrieved November 19, 2021, from http://www.co.rice.mn.us/492/Coronavirus-disease-COVID-19-Current-Sit.

The New York Times. (2020, August 26). Tracking the coronavirus at U.S. colleges and Universities. The New York Times. Retrieved November 13, 2021, from https://www.nytimes.com/interactive/2020/us/covid-college-cases-tracker.html.

U.S. Department of Commerce, T. N. O. and A. A. (2021, October 6). Rice County Climate Records. Climate. Retrieved November 18, 2021, from https://www.weather.gov/wrh/climate?wfo=lox.

U.S. News & World Report 2021. (2021). U.S. news rankings: National Liberal Arts College and University Rankings. Retrieved November 19, 2021, from https://www.usnews.com/best-colleges/rankings/national-liberal-arts-colleges.

Table 1: Appendix A1. COVID Testing and Positive Cases within Carleton College

Period	Rice_County	Total_Tests	Negative_Results	Positive_Results	Positivity_Rate	Term	Week	Ave_temp
8/9 - 8/15	44	161	157	4	2.50%	SB20	1	70.07
8/16 - 8/22	70	235	233	2	0.90%	SB20	2	67.00
8/23 - 8/29	53	185	185	0	0.00%	SB20	3	74.00
8/30 - 9/5	31	465	462	3	0.60%	SB20	4	62.50
9/6 - 9/12	34	1519	1519	0	0.00%	SB20	5	51.21
9/13 - 9/19	44	1277	1273	4	0.30%	FA20	1	54.36
9/20 - 9/26	39	324	323	1	0.30%	FA20	2	61.43
9/27 - 10/3	53	340	340	0	0.00%	FA20	3	53.29
10/4 - 10/10	97	319	318	1	0.30%	FA20	4	47.43
10/11 - 10/17	80	340	339	1	0.30%	FA20	5	50.93
10/18 - 10/24	98	332	331	1	0.30%	FA20	6	30.07
10/25 - 10/31	194	334	333	1	0.30%	FA20	7	25.93
11/1 - 11/7	560	340	336	4	1.20%	FA20	8	44.00
11/8 - 11/14	785	385	373	12	3.10%	FA20	9	40.07
11/15 - 11/21	646	53	49	4	7.50%	FA20	10	27.57
11/22 - 11/28	387	18	16	2	11.10%	WB20	1	28.57
11/29 - 12/5	395	16	16	0	0.00%	WB20	2	26.64
12/6 - 12/12	288	1	1	0	0.00%	WB20	3	30.05
12/13 - 12/19	285	17	17	0	0.00%	WB20	4	19.07
12/20 - 12/26	226	3	3	0	0.00%	WB20	5	19.29
12/27 - 1/2	184	0	0	0	0.00%	WB20	6	12.94
1/3 - 1/9	240	1026	1025	1	0.00%	WI21	1	18.14
1/10 - 1/16	173	2350	2343	7	0.30%	WI21	2	26.43
1/17 - 1/23	157	664	661	3	0.50%	WI21	3	14.57
1/24 - 1/30	130	666	663	3	0.50%	WI21	4	12.07
1/31 - 2/6	104	659	658	1	0.20%	WI21	5	18.50
2/7 - 2/13	77	656	655	1	0.20%	WI21	6	-11.50
2/14 - 2/20	50	661	661	0	0.00%	WI21	7	-8.71
2/21 - 2/27	53	656	656	0	0.00%	WI21	8	25.21
2/28 - 3/6	90	672	672	0	0.00%	WI21	9	28.21
3/7 - 3/13	85	641	640	1	0.20%	WI21	10	37.50
3/14 - 3/20	98	2	2	0	0.00%	SB21	1	34.64
3/21 - 3/27	123	1042	1041	1	0.10%	SB21	2	37.07
3/28 - 4/3	169	2716	2711	5	0.20%	SP21	1	34.71
4/4 - 4/10	187	899	885	14	1.20%	SP21	2	51.57
4/11 - 4/17	194	780	779	1	0.10%	SP21	3	39.43
4/18 - 4/24	125	935	934	1	0.10%	SP21	4	38.93
4/25 - 5/1	107	1043	1043	0	0.00%	SP21	5	41.86
5/2 - 5/8	88	731	731	0	0.00%	SP21	6	49.50
5/9 - 5/15	65	694	694	0	0.00%	SP21	7	47.86
5/16 - 5/22	53	528	528	0	0.00%	SP21	8	65.43
5/23 - 5/29	16	329	329	0	0.00%	SP21	9	58.00
5/30 - 6/5	17	207	207	0	0.00%	SP21	10	62.29
8/14 - 8/20	117	150	150	0	0.00%	SB21	1	73.64
8/21 - 8/27	99	268	267	1	0.37%	SB21	2	67.07
8/28 - 9/3	142	384	383	1	0.26%	SB21	3	66.57
9/4 - 9/10	145	1792	1791	1	0.06%	SB21	4	61.78
9/11 - 9/17	171	1649	1648	1	0.06%	FA21	1	62.00
9/18 - 9/24	189	2501	2497	4	0.16%	FA21	2	59.14
9/25 - 10/1	193	420	420	0	0.00%	FA21	3	64.07
10/2 - 10/8	197	420	420	0	0.00%	FA21	4	61.14
10/9 - 10/15	143	713	711	2	0.28%	FA21	5	55.57
10/16 - 10/22	115	722	721	1	0.14%	FA21	6	48.00
10/23 - 10/29	162	1263	1255	8	0.63%	FA21	7	40.93
10/30 - 11/5	184	915	912	3	0.33%	FA21	8	37.75
11/6 - 11/12	298	542	538	4	0.37%	FA21	9	49.43

Table 2: Appendix A2. COVID Testing and Positive Cases within 38 Universities and Colleges

Institution	Total_Tests	Positive_Results	Positive_Rate	Type	Enrollment
Carleton College	36418	101	0.28%	LAC	1940
St. Olaf College	42573	225	0.53%	LAC	2953
University of Minnesota Twin City	33456	1692	5.06%	University	51147
Macalester College	4163	14	0.34%	LAC	2049
Smiths College	71434	29	0.04%	LAC	2183
Harvard University	1232467	1687	0.14%	University	22947
University of Chicago	155482	310	0.20%	University	14467
Ohio State University	747182	8232	1.10%	University	66444
Johns Hopkins University	294505	515	0.17%	University	23917
University of California Los Angeles	57955	2376	4.10%	University	44947
University of California Berkely	164714	638	0.39%	University	41910
Univsersity of Michigan	291887	12060	4.13%	University	44718
Boston University	2062000	3013	0.15%	University	32551
University of Washington	129257	2144	1.66%	University	46081
Rice University	72783	191	0.26%	University	6740
University of Texas Austin	206942	5189	2.51%	University	50950
University of Miami	226038	3874	1.71%	University	17003
Pennsylvania State University	190381	3567	1.87%	University	98783
Amherst College	250419	108	0.04%	LAC	1745
Williams College	153423	84	0.05%	LAC	1962
Middlebury College	60320	31	0.05%	LAC	2580
Bryn Mawr College	16008	56	0.35%	LAC	1300
Washington and Lee University	45308	494	1.09%	LAC	1822
Bowdoin College	45356	84	0.19%	LAC	1777
University of Illinois, Urbana-Champaign	2368302	7951	0.34%	University	44087
Purdue University	218650	6247	2.86%	University	41574
Colby College	67709	56	0.08%	LAC	2155
University of Arizona	328668	4713	1.43%	University	44831
Stanford University	767676	714	0.09%	University	16914
University of Pennsylvania	504221	3496	0.69%	University	25806

title: “Final Project”
output: pdf_document

Data Wrangling

```
#load data
library(readxl)
Covid_Carleton<- read.csv("Covid Data - Carleton.csv")
```

```
#data wrangling -- separate terms
Covid_Carleton <- Covid_Carleton %>%
  filter(Term == "FA20"| Term == "WI21" | Term == "SP21" | Term == "FA21")  %>%
  mutate(term=recode(Term, "FA20" = 1,"WI21" = 2,"SP21" = 3,"FA21" = 4))
```

EDA

```
#mutate variable, rename variable
Covid_Summary <- Covid_Carleton %>%
  mutate(Term = recode(term, "1"="FA20","2"="WI21","3"="SP21","4"="FA21")) %>%
  group_by(term) %>%
  summarise(Mean_Positive_Case = mean(Positive_Results),Mean_Positive_Rate_ = mean(Positivity_Rate))

# at most 4 decimal places
knitr::kable(Covid_Summary, digits = 3)
```

```
#standardize variable rice_standard
Covid_Carleton$Rice_Standard <- as.vector(scale(Covid_Carleton$Rice_County))
#EDA
ggplot(Covid_Carleton, aes(Rice_Standard, Positive_Results, group = 1)) +
  geom_point() + theme_bw() +
  geom_point() + geom_smooth(method = "lm", se = F) +
  xaxis_text(angle = 90, size = 6) +
  ggtitle("Relationship between Carleton Cases and Rice County Cases")

Covid_Carleton$Temp_Standard <- as.vector(scale(Covid_Carleton$Ave_temp))
ggplot(Covid_Carleton, aes(Temp_Standard, Positive_Results, group = 1)) +
  geom_point() + geom_smooth(method = "lm", se = F) +
  theme_bw() +
  xaxis_text(angle = 90, size = 6)+
  ggtitle("Relationship between Carleton Cases and Temperature")
```

```
#time series plot of positivity rate in different period
ggplot(Covid_Carleton, aes(Period, Positive_Results, group = 1)) +
  geom_point() +  geom_line() +
  theme_light() +
  xaxis_text(angle = 90, size = 6)+ggtitle("Covid Cases by Week")
```

```
#boxplot of positive rate by term
Covid_Carleton  %>%
  group_by(Term) %>%
  ggplot(aes(Term, Positive_Results)) + geom_boxplot() +
  theme_light() +
  coord_flip()+
  ggtitle("Positive Case by Term")
```

JAGS

```
# JAGS code
modelString <-"
model{
## sampling
for (i in 1:N){
 y[i] ~ dbin(theta[term[i]], n[i])}

## priors
for (j in 1:M){
 theta[j] ~ dbeta(alpha, beta)}

alpha <- mu * eta
beta <- (1-mu) *eta
mu ~ dbeta(1,1)
eta <- exp(logeta)
logeta ~ dlogis(log(100), 1)
}"
```

```
#data details
y <- Covid_Carleton$Positive_Results
n <- Covid_Carleton$Total_Tests
N <- length(y)
M <- length(unique(Covid_Carleton$Term))
term <- Covid_Carleton$term
the_data <- list(y = y, n=n, N=N,M=M, term = term)

# seed for JAGS
init = list(
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987654),
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987653),
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987652)
            )

# MCMC draws
carleton_post <- run.jags(
  model = modelString,
  n.chains = 3,
  data = the_data,
  monitor = c("theta"),
  adapt = 1000,
  burnin = 5000,
  sample = 10000,
  silent.jags = TRUE  # Eliminates progress bar
)
```

Diagnostics

```
#convert to mcmc object
post_mcmc <- as.data.frame(as.mcmc(carleton_post))
#trace plot
mcmc_trace(post_mcmc)
#acf plot
mcmc_acf(post_mcmc)
#overlayed density
mcmc_dens_overlay(carleton_post$mcmc)
#summary mcmc
summary(post_mcmc)
```

Posterior Predictive

```
# JAGS code

modelString <-"
model{
## sampling
for (i in 1:N){
 y[i] ~ dbin(theta[term[i]], n[i])}

## priors
for (j in 1:M){
 theta[j] ~ dbeta(alpha, beta)}

alpha <- mu *eta
beta <- (1-mu)*eta
mu ~ dbeta(1,1)
eta <- exp(logeta)
logeta ~ dlogis(log(100), 1)
for (i in 1:N) {
   y_pred[i] ~ dbin(theta[term[i]], n[i])
}

}
"
```

```
# data details
y <- Covid_Carleton$Positive_Results
n <- Covid_Carleton$Total_Tests
N <- length(y)
M <- length(unique(Covid_Carleton$Term))
term <- Covid_Carleton$term
the_data <- list(y = y, n=n, N=N,M=M, term = term)

#seed
init = list(
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987654),
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987653),
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987652)
            )
#MCMC draws
```

```
carleton_pred <- run.jags(
  model = modelString,
  n.chains = 3,
  data = the_data,
  monitor = c("y_pred"),
  adapt = 1000,
  burnin = 5000,
  sample = 10000,
  thin=5,
  silent.jags = TRUE  # Eliminates progress bar
)
```

```
#posterior predictive check
ppc_dens_overlay(y = Covid_Carleton$Positive_Results, yrep = carleton_pred$mcmc[[1]][1:100,])
```

Inference

```
#make posterior inference plot
post_intervals <- mcmc_intervals_data(carleton_post$mcmc) %>%
  mutate(para = c("FA20","WI21","SP21","FA21"))
head(post_intervals)

slice(post_intervals, 1:4) %>%
  ggplot(
    aes(x = reorder(para, (m)), y = (m), ymin = (ll), ymax = (hh))) +
  geom_pointrange() +
  theme_light() +
  xaxis_text(angle = 0, size = 6) +
  theme(
    panel.grid.major.x = element_blank(),
    panel.grid.minor.x = element_blank()
)+
xlab("Term") + ylab("Positive Test Rate")
```

Logistical Regression

```
#JAGS
modelString <-"
model {
## likelihood
for (i in 1:n){
   y[i] ~ dbin(p[i], N[i])
}
## priors and regression
for (i in 1:n){
   logit(p[i]) <- beta0 +beta1*rice[i]+beta2*temp[i]
   }
## hyperpriors
beta0 ~ dnorm(0, 0.0001)
beta1 ~ dnorm(0, 0.0001)
beta2 ~ dnorm(0, 0.0001)

}
```

```
"
# data details
Covid_Carleton <- Covid_Carleton %>%
  mutate(Rice_Standard = as.vector(scale(Rice_County)),
         temp_standard = as.vector(scale(Ave_temp)))


y <- Covid_Carleton$Positive_Results
n <- length(y)
N <- Covid_Carleton$Total_Tests
rice <- Covid_Carleton$Rice_Standard
temp <- Covid_Carleton$temp_standard

the_data <- list(y=y,n=n,N=N,rice=rice,temp=temp)

# seed
init = list(
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987654),
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987653),
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987652)
            )

# MCMC draws
carleton_post <- run.jags(
  model = modelString,
  n.chains = 3,
  data = the_data,
  monitor = c("beta0","beta1","beta2","y_pred"),
  adapt = 1000,
  burnin = 5000,
  sample = 10000,
  thin=5,
  silent.jags = TRUE  # Eliminates progress bar
)
```

```
# convert to mcmc
post_mcmc <- as.data.frame(as.mcmc(carleton_post))
quantile(post_mcmc$beta0,c(0.05,0.95))
quantile(post_mcmc$beta1,c(0.05,0.95))
quantile(post_mcmc$beta2,c(0.05,0.95))
```

```
# diagnostics
#trace plot
mcmc_trace(post_mcmc)
#acf
mcmc_acf(post_mcmc)
#density plot
mcmc_dens_overlay(carleton_post$mcmc)
#summary
summary(post_mcmc)
```

Posterior Predictive

```
#JAGS
modelString <-"
model {
## likelihood
for (i in 1:n){
   y[i] ~ dbin(p[i], N[i])
}
## priors and regression
for (i in 1:n){
   logit(p[i]) <- beta0 +beta1*rice[i]+beta2*temp[i]
   }
## hyperpriors
beta0 ~ dnorm(0, 0.0001)
beta1 ~ dnorm(0, 0.0001)
beta2 ~ dnorm(0, 0.0001)

for (i in 1:n) {
   y_pred[i] ~ dbin(p[i], N[i])
}

}
"
# MCMC draws
carleton_pred <- run.jags(
  model = modelString,
  n.chains = 3,
  data = the_data,
  monitor = c("y_pred"),
  adapt = 1000,
  burnin = 5000,
  sample = 10000,
  thin=5,
  silent.jags = TRUE  # Eliminates progress bar
)
```

```
#posterior preditive check
ppc_dens_overlay(y = Covid_Carleton$Positive_Results, yrep = carleton_pred$mcmc[[1]][1:100,])
```

Residual Diagnostics

```
#residual of SLR
resids <- Covid_Carleton %>%
mutate(
beta0 = mean(post_mcmc$beta0),
beta1 = mean(post_mcmc$beta1),
beta2 = mean(post_mcmc$beta2),
phat_logit = beta0 + beta1 * Rice_Standard + beta2 * temp_standard,
phat = exp(phat_logit)/(1+exp(phat_logit)),
yhat = rbinom(39,Total_Tests, phat),
resid = Positive_Results - yhat
)
```

```
# residuals against fitted values
ggplot(data = resids, aes(x = yhat, y = resid)) + # fitted values
  geom_hline(yintercept = 0, linetype = 2, color = "blue") +
  geom_point() + theme_bw()
qqnorm(resids$resid)
qqline(resids$resid)

#against explanatory variable
ggplot(data = resids, aes(x = Rice_Standard, y = resid)) + # fitted values
  geom_hline(yintercept = 0, linetype = 2, color = "blue") +
  geom_point() + theme_bw()
```

```
#quantile of parameter
quantile(post_mcmc$beta2,c(0.05,0.95))
```

Between instituiton comparison

EDA

model mu (Model2)

```
#JAGS
modelString_mu <-"
model {
## likelihood
for (i in 1:N){
   y[i] ~ dbin(p[i], n[i])
}

## priors and regression
for (i in 1:N){
   p[i] ~ dbeta(a[i], b[i])
   a[i] <- phi[i] * mu[i]
   b[i] <- phi[i] * (1 - mu[i])
   logit(mu[i]) <- beta0 +beta1*type[i] +beta3*test_rate[i]
   phi[i] <- exp(logeta[i])
   logeta[i] ~ dlogis(logn, 1)

}
## hyperpriors

beta0 ~ dnorm(0, 0.0001)
beta1 ~ dnorm(0, 0.0001)
beta3 ~ dnorm(0, 0.0001)

## predictive
for (i in 1:N) {
   y_pred[i] ~ dbin(p[i], n[i])
}

}

"
```

```r
# seed
init = list(
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987654),
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987653),
            list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 987652)
            )



# data details
y <- cross$Positive_Results
n <- cross$Total_Tests
type <- cross$type_num
test_rate<- cross$Test_Rate
N <- length(y)
the_data2 <- list("y" = y, "n" = n, "N" = N, "type" = type, "test_rate" = test_rate,
                 "logn" = log(100))


#mcmc draws
posterior_mu <- run.jags(modelString_mu,
                 data = the_data2,
                 n.chains = 3,
                 monitor = c("beta0","beta1","beta3", "p"),
                 adapt = 1000,
                 burnin = 5000,
                 sample = 10000,
                 thin = 5,
                 inits = init, silent.jags = TRUE)

posterior_mu_pred <- run.jags(modelString_mu,
                 data = the_data2,
                 n.chains = 3,
                 monitor = c("y_pred"),
                 adapt = 1000,
                 burnin = 5000,
                 sample = 10000,
                 thin = 5,
                 inits = init, silent.jags = TRUE)
```

model p standardized (Model1)

model diagnostics

```r
# diagnostics for model p standradized
mcmc_trace(posterior_p_stand$mcmc)
mcmc_acf(posterior_p_stand$mcmc)
mcmc_dens_overlay(posterior_p_stand$mcmc)
```

```r
# diagnostics for model mu
mcmc_trace(posterior_mu$mcmc)
mcmc_acf(posterior_mu$mcmc)
mcmc_dens_overlay(posterior_mu$mcmc)
```

```
# gelman -- good
gelman.diag(posterior_mu$mcmc)

# geweke -- good
geweke.diag(posterior_mu$mcmc)

# effective sample size -- good
effectiveSize(posterior_mu$mcmc)

# pretty good
summary(posterior_mu$mcmc)
```

Posterior predictive check

```
#not good
ppc_dens_overlay(cross$Positive_Results, posterior_p_stand_pred$mcmc[[1]][1:100, ])
#kinda bad
ppc_stat_grouped(y = cross$Positive_Results,
                 yrep= as.matrix(posterior_p_stand_pred$mcmc[[1]]),
                 group = cross$Institution)
```

```
ppc_dens_overlay(cross$Positive_Results, posterior_mu_pred$mcmc[[1]][1:100, ])
#actually pretty good
ppc_stat_grouped(y = cross$Positive_Results,
                 yrep= as.matrix(posterior_mu_pred$mcmc[[1]]),
                 group = cross$Institution)
```

inference plot

observed rate vs posterior intervals

```
# intervals
post_intervals_mu <- mcmc_intervals_data(posterior_mu$mcmc, regex_pars = "p", prob_outer = 0.9)

# plot observed rate vs poterior intervals
ggplot(cross, mapping =aes(x = (cross$Positive_Results)/(cross$Total_Tests))) +
  geom_point(data.frame(post_intervals_mu[1:30, ]),
             mapping = aes(y = post_intervals_mu[1:30,  7] %>% unlist()), alpha = 0.5) +
  geom_linerange(data.frame(post_intervals_mu[1:30, ]),
                 mapping = aes(ymin = post_intervals_mu[1:30,  5] %>% unlist(),
                               ymax =post_intervals_mu[1:30,  9] %>% unlist()), alpha = 0.5) +
  geom_abline(slope = 1, intercept = 0, color = "dodgerblue")
```

residual diagnostics for Model1 and Model2

```
# residual diagnostics for model p standardized
post_p_stand<-as.mcmc(posterior_p_stand)
# post_p_stand[, 5:34]
resids_p_stand <- cross %>%
mutate(
   phat = colMeans(post_p_stand[, 5:34]),
   yhat = rbinom(30,cross$Total_Tests, phat),
   resid = Positive_Results - yhat
)

# vs fitted values
ggplot(data = resids_p_stand, aes(x = yhat, y = resid)) + # fitted values
  geom_hline(yintercept = 0, linetype = 2, color = "blue") +
  geom_point() + theme_bw()
qqnorm(resids_p_stand$resid)
qqline(resids_p_stand$resid)

# vs z_test rate
ggplot(data = resids_p_stand, aes(x = z_test_rate, y = resid)) + # fitted values
  geom_hline(yintercept = 0, linetype = 2, color = "blue") +
  geom_point() + theme_bw()


# vs z_enrol
ggplot(data = resids_p_stand, aes(x = z_enrol, y = resid)) + # fitted values
  geom_hline(yintercept = 0, linetype = 2, color = "blue") +
  geom_point() + theme_bw()

## qqplot of model p stand
qqnorm(resids_p_stand$resid)
qqline(resids_p_stand$resid)
```

```
# resiudal diagnostics for model mu
post_mu<-as.mcmc(posterior_mu)
# post_mu[, 4:33]
resids_mu <- cross %>%
mutate(
   phat = colMeans(post_mu[, 4:33]),
   yhat = rbinom(30,cross$Total_Tests, phat),
   resid = Positive_Results - yhat
)

# vs fitted values
ggplot(data = resids_mu, aes(x = yhat, y = resid)) + # fitted values
  geom_hline(yintercept = 0, linetype = 2, color = "blue") +
  geom_point() + theme_bw()
qqnorm(resids_mu$resid)
qqline(resids_mu$resid)

# vs test rate
ggplot(data = resids_mu, aes(x = cross$Test_Rate, y = resid)) + # fitted values
  geom_hline(yintercept = 0, linetype = 2, color = "blue") +
  geom_point() + theme_bw()
```