Factors driving the reopening of colleges amid the Covid-19 crisis

Pedro Vallocci[1], Toshiya Yoshida[1]
University of California, Santa Cruz[1]

Abstract

U.S. College administrators have been struggling with managing the campus operation, dealing with the effects of the COVID-19 pandemic throughout the country. Some universities are conducting classes fully online and placing restrictions on entry to the campus facilities. In contrast, other universities reopened at the beginning of the academic year 2020 or decided to adopt a so-called hybrid mode that mixes online and in-person instruction. This study aims to know what factors can predict the college instruction mode in the Fall by performing regression analysis and evaluating the predictive performance. Driven by our supposition and past work, we use institutional data, the severity of the spread of COVID-19, and regional political orientation in the model. The analyses suggest that public universities with a large endowment situated in areas with higher population density and a higher percentage of Biden votes in the 2020 presidential election tend to provide online instruction in the Fall.

KEY WORDS: COVID-19, College Reopening Plan, Logistic regression.

1. Introduction

1.1 Description

During the Fall of 2020, colleges were faced with significant challenges during the COVID-19 pandemic, where contradictory forces led colleges to reopen or not. We first guessed that variables pertinent to financial information affect the instruction plan. For example, financially vulnerable colleges are more likely to resume in-person classes to attract students who do not like remote instruction and bring their students to the campus to secure the income from their facilities such as dormitory. Other factors such as institutional characteristics (e.g., private or public), the reported cases of COVID-19, or sporting activity are considered in this project.

On the other hand, wealthy universities with larger endowments are more likely to prioritize preventative measures by offering fully or primarily online instruction.

1.2 Prior analysis

Studies on influential factors on the college instruction plans are being discussed by researchers. For example, Felson and Adamczyk (2020) conducted hierarchical logistic regression models to predict instruction modes and included random effects for higher education states and systems. This study suggested that the state's political orientation and some college data (e.g., facilities, financial vulnerability) can predict the college's instruction mode.

2. Dataset

We gained and merged several data sources to collect the variables mentioned above: college reopening data, COVID-19 data, college's institutional and financial data, and political data.

The primary dataset is the college reopening data from Davidson's College Initiative (CCI).[1] The dataset consists of instruction methods in the United States as of October 30, 2020, with some basic institutional variables such as enrollment and the cases of COVID-19 reported in the corresponding county. To incorporate variables related to financial stability, we use Galloway's (2020) worksheet to evaluate the U.S. colleges' value and vulnerability. Specifically, this worksheet includes financial and enrollment information imported from the Integrated Postsecondary Education Data System (IPEDS), U.S. News & World Report, Niche.com, and the Center on Education and the Workforce. Galloway's worksheet also calculated indicators such as student life score or vulnerability score. We imported the universities' total sports revenue from the National Collegiate Athletic Association (NCAA). County-level population data was imported from the United States Department of Agriculture and used to normalize COVID-19 related variables. Furthermore, we also imported the presidential election results data following Felson and Adamczyk's (2020) study. We imported the county-level share for Joe Biden in the 2020 presidential election, while Felson and Adamczyk (2020) used the state-level share for Hillary Clinton in the 2016 presidential election.[2]

2.1 Data manipulation

We first merged the above datasets by carefully comparing the university names or the county names. However, the merged dataset is not ready for analysis for reasons. First of all, the original data has five categories for the instruction mode: fully online, primarily online, hybrid, primarily in-person, and fully in-person. The frequency

[1] We scraped the dataset from the Chronicle of Higher Education, which relied on CCI.

[2] We imported the results of the presidential election in 2020 from Tony McGovern's website (https://github.com/tonmcg/US_County_Level_Election_Results_08-20).

for these categories are plotted in Figure 1. It is obvious that the number of fully in-person is small and the distribution is seriously unbalanced. Since this makes the regression analysis and evaluation of predictive performance difficult, we aggregated them into a variable with two levels by relabelling the latter three categories as "in-person," and the remaining two categories are relabelled as "online."

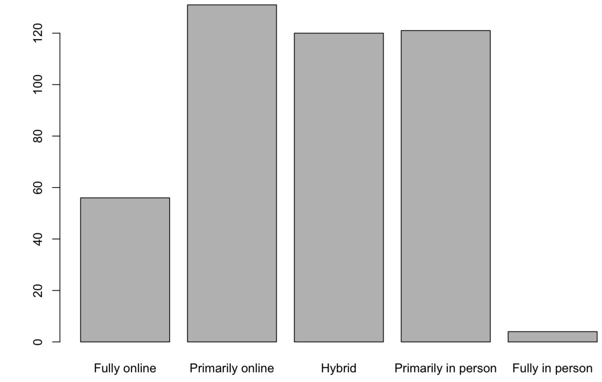

Figure 1: Distribution of instruction mode of the colleges

Another manipulation we did is a transformation of some variables. To be more specific, we added the log-transformation of enrollment, full-time enrollment, sports revenue, population, endowment per student, and instructional wages into the merged dataset. Figure 2 and 3 are an example of how log-transformation works. The distribution is symmetric and looks Gaussian in Figure 3.

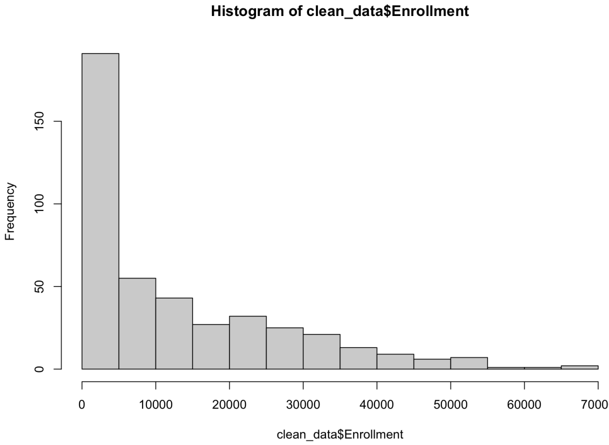

Figure 2: Histogram of Enrollment

The merged dataset has other seriously unbalanced categorical variables. For example, Tables 1 and 2 show the distribution of the categorical variables: Category of the college, and the dummy variable for Historically black colleges and universities (HBCU) or Tribal colleges. From the tables, it is evident that they are extremely unbalanced.

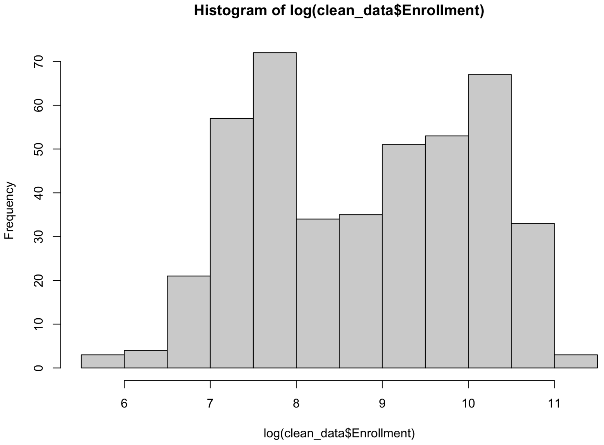

Figure 3: Histogram of log-transformed Enrollment

Table 1: Distribution of Categories of the colleges

Profit, 4-year	Public, 2-year	Public 4-year
288	1	144

Table 2: Distribution of HBCU or Tribal colleges

Yes	No
4	429

We omitted these variables from our estimation process since they are not suited for regression analysis and prediction.

2.2 EDA

After the data processing, we performed an exploratory analysis on the prepared dataset. We output boxplots of categorical variables against the instruction mode. For instance, Figure 4 represents the boxplots of four variables: share of Joe Biden vote in 2020, population, log-transformed population, and endowment per full-time students.

As for these variables, the county level share of Biden's votes, the county population, and the log-transformed county population are different for different instruction modes. We can also find a difference in endowment per full-time student for different instruction modes. More specifically, the shape of the distribution around the mean is almost the same.

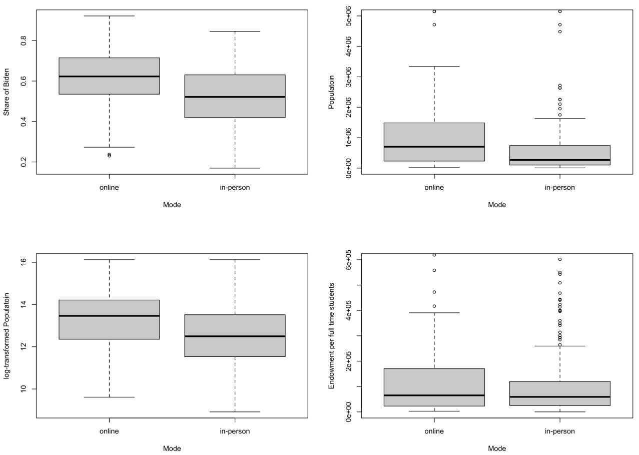

Figure 4: Boxplots of four variables against instruction mode

3. Model estimation

We propose a logistic relationship between *Mode* and some of the explanatory variables in our dataset. The resulting dataset, however, has a large number of explanatory variables. Also, it has a significant amount of correlation among them. This correlation occurs partly by design, because several variables in Galloway's spreadsheet are functions of others; some of them are even exact linear combinations (e.g., *Education.Score*, *Credential.Score*, and *Vulnerability.Score*). Therefore, to find the most significant subset of regressors, we perform LASSO (least absolute shrinkage and selection operator).

Our dataset, however, has the categorical variables *Category*, *State*, and *Hospital*. Since each of the categories' underlying values have merely an ordinal meaning, we must perform one-hot encoding before using LASSO. One-hot encoding expands each categorical value with n levels into n different binary variables. After one-hot encoding, the binary variable *Cat_i*, for example, is true if, and only if, the categorical variable *Cat* assumes category i in the sampled unit.

Moreover, to suit our dataset for this shrinkage method, we normalize our dataset to demean it and scale the variables' variances to 1. We then use the package *gamlr* to run LASSO. We choose binomial as the response type.

Figures 5 and 6 show the result of the 5-fold cross-validation for the LASSO penalty selection. Figures 5 and 6 shows us that the optimal penalty factor λ lies, in log, somewhere around -3.3.

After the process of cross-validation, LASSO chooses an optimal penalty value λ and returns a model where the only variables with a nonzero coefficient are:

- Intercept

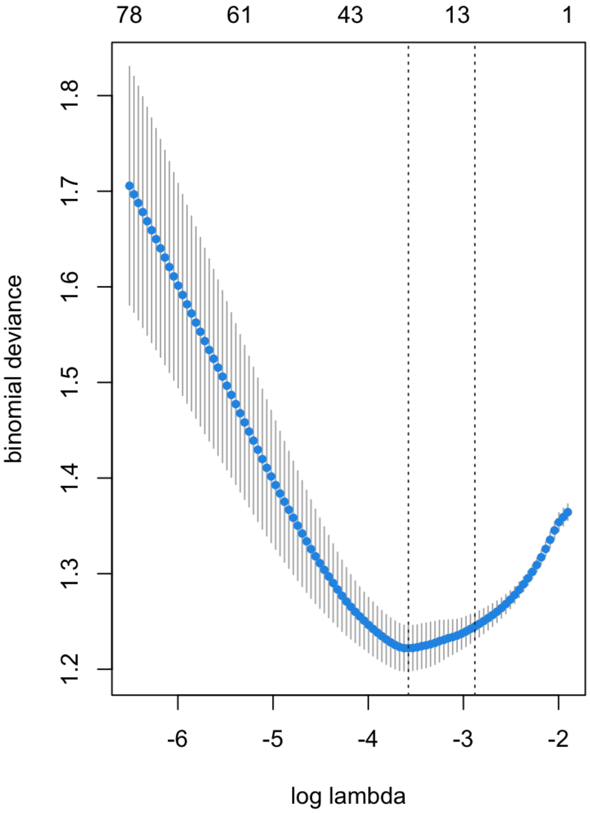

Figure 5: 5-fold cross-validation results for LASSO and varying λ

- Category
- State dummies for California, District of Columbia, Maryland, Texas, and Vermont
- Credential score
- Endowment per full-time student
- Percentage of democrat vote in the 2020 election
- log(Population)

The model with a reduced number of coefficients fitted by LASSO, however, is biased. Thus, we refit a logistic model using the regressors above and the *glm* package, assessing the significance of all the included regressors. We perform a backward elimination of variables, removing at each iteration the variable with the highest p-value until the largest p-value is below a threshold of 0.05. Thus, we aim to have a significance level of at least 95% for all our regressors.

We also tested interaction terms between the regressors. In no case, the interaction terms turned out to be significant at a 95% significance level.

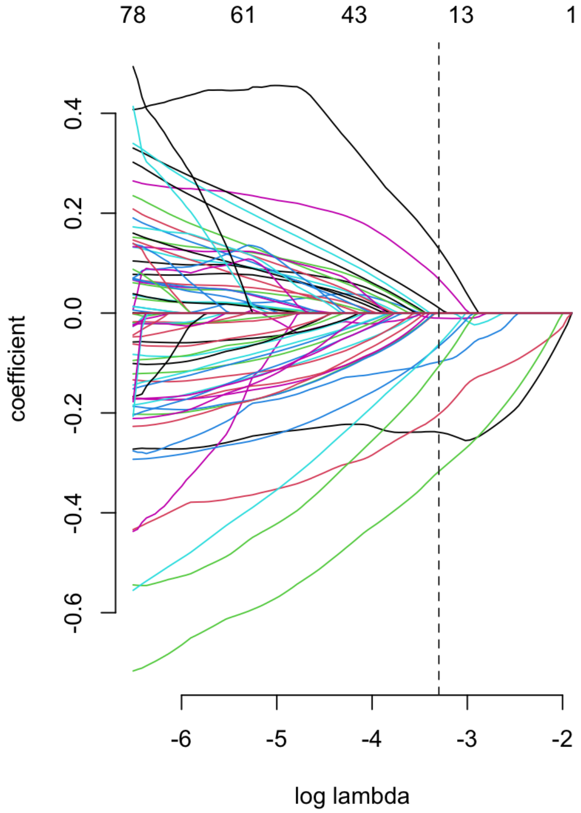

Figure 6: Profiles of the LASSO coefficients as the tuning parameter λ is varied

After doing backwards elimination, we find the following model, described as Model 1 in Table 3:

$$\begin{aligned} logit(p_i) = \mu + \alpha_{1,i}Private + \alpha_{2,i}StateCA + \\ \alpha_{3,i}StateTX + \beta_{1,i}Endowment + \beta_{2,i}log(Pop) + \\ \beta_{3_i}PctDem \end{aligned} \quad (1)$$

In the following analysis:

- μ is the intercept.
- p_i corresponds to the probability of teaching being in-person.
- *Private* is a variable that denotes if a college is privately owned.
- *StateCA* is 1 if the college is in California; and is 0 otherwise.
- *StateMD* is 1 if the college is in Maryland; and is 0 otherwise.
- *StateTX* is 1 if the college is in Texas; and is 0 otherwise.
- *Endowment* represents the college's endowment per full time student.
- *log(Pop)* denotes the logarithm of the county population
- *PctDem* represents the percentage of the population who voted democrat in the 2020 elections

We also fit two more models, Model 2 and Model 3, in Table 3. In Model 2, we maintain $State_MD$ and remove *PctDem* instead; in Model 3, we refit the model with all variables chosen by the LASSO but substituting *Endowment* by $log(Endowment)$.

Using Akaike's Information Criterion, we notice that Model 1 and Model 2 are the preferred ones. However, Model 2 uses a less balanced variable ($StateMD$) while discarding the more balanced *PctDem*; therefore, due to the similar AIC, we would prefer Model 1. The predictive power analysis also provides further evidence in favor of Model 1 (see Table 4).

Notice that, for all models in Table 3, the p-values of the Z-statistics of all coefficients are less or equal than 0.05. This result means that, in all cases, we can reject the null hypothesis H_0 for each coefficient β, where $H_0 : \beta = 0$, at a 95% significance level.

4. Model evaluation

We evaluate our model's predictive performance in-sample using the hold-out method, dividing our sample in a train set and in a test set, which correspond to 90% and 10% of the sample, respectively. We perform 5000 iterations, reshuffling the model at each time. We reestimate the model for each iteration on the train set and evaluate its predictive performance in the test set. The predictive performance of Model 1 can be seen in Figure 7.

We also evaluate our models in terms of their predictive power. The comparison between the models can be seen in Table 4.

Since logistic regression does not rely on the assumption of Gaussian-distributed errors, plots of residual vs. fitted values, or normal Q.Q. plots are not useful. Instead, we use the binned plot, shown in Figure 8. In the binned plot, we divide the data into bins based on their fitted values and then plot the average residual versus the average fitted value for each bin. We expect that the 95% of the residuals fall into the error bounds. Therefore, we conclude that the residuals in our analysis behave as expected.

Interpreting the coefficients in Table 3, we recalling Equation 1 and that p_i corresponds to the probability of in-person teaching, the models give us evidence that private schools are more likely to provide in-person instruction.

	logit(p_i)		
	Model 1	Model 2	Model3
(Intercept)	5.038^{***}	5.662^{***}	8.697^{***}
	(1.172)	(1.162)	(1.7295)
Private	0.877^{***}	0.9223^{***}	1.1108^{***}
	(0.2437)	(0.2436)	(0.2829)
StateCA	-2.235^{**}	-2.208^{**}	-2.2657^{**}
	(0.7686)	(0.7681)	(0.7662)
StateMD		-2.323^{*}	
		(1.142)	
StateTX	1.965^{*}	2.236^{**}	2.0059^{*}
	(0.8238)	(0.8217)	(0.8183)
log(Endowment)			-0.3444^{***}
			(0.1036)
Endowment	$-1.88\times10^{-6**}$	$-2.08\times10^{-6***}$	
	(5.83×10^{-7})	(5.82×10^{-7})	
log(PoP)	-0.2946^{**}	-0.4378^{***}	-0.3198^{**}
	(0.1061)	(0.0910)	(0.1073)
PctDem	-2.126^{*}		-2.0593^{*}
	(0.8747)		(0.8744)
Observations	433	433	433
AIC	508.98	508.92	511.88

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 3: Fitted models

	Predictive power (mean)
Model 1	0.7045
Model 2	0.6821
Model 3	0.6845

Table 4: Comparison of the predictive power of the three models

Partly funded by tax revenue, public universities may be more resilient to temporary income loss due to their closing for in-person activities.

The models also give us evidence that universities in California (and Maryland, in Model 2), even controlling for the states' democratic leanings in most counties, are more likely to provide online teaching than expected.

Conversely, universities in Texas are more prone to stay open than expected. In Texas, it is possible that the state's republican leaning, besides each county's leaning, also weighs on the university's decision. We must consider that *PctDem* is a county-level variable, and some Texan universities are located in democratic-leaning counties (e.g., Universities in Austin, Dallas, Houston, El Paso).

Universities that switched to online teaching are probably wealthier, according to Table 3. They can afford the loss of income that comes from the closing. In contrast, less affluent universities may be more dependent on revenues that diminish or disappear in the case of online teaching: sports revenues, meal plan revenues, and tuition fees. This relationship was not evident initially since wealthier universities can also afford to provide widespread testing. It is generally questionable if the possibility of testing influenced the colleges' reopening decision since the regressor *Hospital* was not found to be significant in any of the models we regressed.

Interestingly, the variables containing COVID-19 cases per county inhabitant (total accumulated cases, or cases in the previous 60 days) were not significant in the LASSO estimation. The "county vulnerability to a COVID-19 outbreak" effect might have been captured by the *log(Pop)* variable. More populated counties were arguably where the first COVID-19 outbreaks happened, are probably more densely populated and, therefore, more prone to agglomerations. It is comprehensible that more universities in more populated counties were more cautious, which can be seen in Table 3.

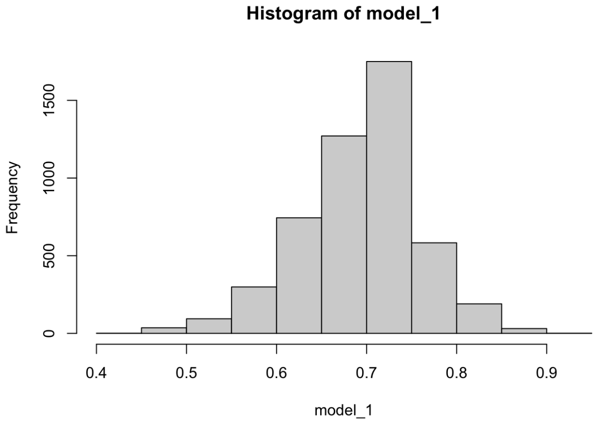

Figure 7: Histogram of the predictive performance of Model 1, 5000 draws

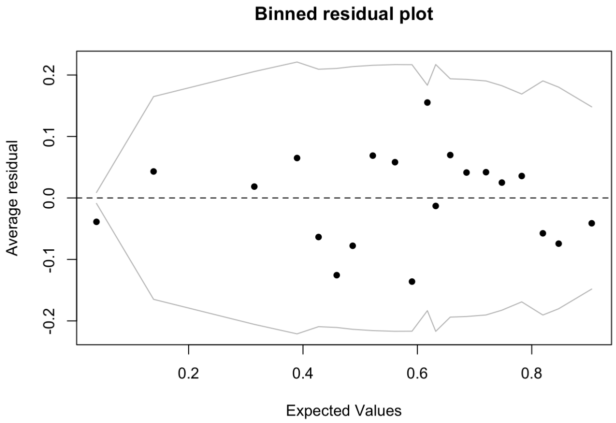

Figure 8: Binned plot of residuals, Model 1

Model 1 also shows that universities in more Democratic counties were more likely to switch to online instruction. As the press widely reported, President Donald Trump supported the universities' reopening in Fall 2020. He pushed for the refusal of the U.S. permanence to international students registered for online classes only. Democratic leaders, such as the State governors of New York and California, were opposed to reopening. We aimed to capture the increasing political polarization in the U.S. with this variable, and it showed to be significant in Model 1.

5. Conclusion

In this model, we concluded that public, wealthier universities in more populous and more Democratic-leaning counties were more likely to switch to online teaching in Fall 2020.

The predictive power found in Table 4 reached 70% in our most accurate model. This power could undoubtedly be improved. One serious drawback we had in our study is the lack of clear boundaries in Davidson's spreadsheet between the intermediary states "primarily online", "hybrid", and "primarily in-person". This ambiguity probably had an impact on the predictive power of our model, and could be targeted in a future expansion of this work.

However, our model shows that universities need economic resilience to overcome a pandemic. This entails saving for tail-risk events.

6. References

- Felson, J., and Adamczyk, A. (2020), "Examining the decision to offer in-person college instruction during the COVID-19 era A multilevel analysis of the factors that affected intentions to open," *medRxiv*, doi:10.1101/2020.10.15.20213363.
- Galloway, S. (2020), "USS University," *No Mercy / No Malice*, Retrieved November 9, 2020 (https://www.profgalloway.com/uss-university).
- McGovern, Tony., "US County Level Election Results 08-20," Retrieved December 8, 2020 (https://github.com/tonmcg/US_County_Level_Election_Results_08-20).
- National Collegiate Athletic Association. (2020), "NCAA Financial Database [Data visualization dashboard]," Retrieved December 7, 2020 (http://www.ncaa.org/about/resources/research/finances-intercollegiateathletics-database).
- United States Department of Agiculture., "Population estimates for the U.S., States, and counties, 2010-2019," Retrieved November 6, 2020 (https://www.ers.usda.gov/data-products/county-level- data-sets/download-data.aspx).