Characterizing the relationship between the religious affiliation and incidences of Covid-19 at U.S. universities.

Daniel de Castro and Laura Appleby

December 14, 2022

Introduction

Although Covid-19 pandemic may feel like a thing of the past, there is still much research to do in order to understand its nascence, spread, and lasting health effects. Indeed, we, as a society, must be prepared to handle the next highly-transmissible pathogen that may arise. As university students whose first two years in school were far from normal on account of the Covid-19 pandemic — whose lives were dominated by case counts and triweekly testing — we are interested in exploring the development of the pandemic at colleges and universities across the United States. In particular, we are interested in the differences in how different academic institutions were affected by the pandemic.

It has been well established that, at least in the United States, responses to the Covid-19 pandemic were tightly bound to political and ideological questions, the responses to which differ across the administrations of the wide range of universities in this country. In this paper, we will explore the differences in the impact of the pandemic between institutions with religious affiliations and those without religious affiliations. We will attempt to characterize the relationship between case counts during the fall and spring semesters of the 2020-21 academic year — the academic year in which responses to the pandemic were perhaps most varied across universities — and whether a university has a religious affiliation. We not only hope to answer the question of whether religious affiliation has a statistically significant relationship with case counts during the 2020-21 academic year — we also hope to discover the other characteristics of universities may have an effect on this relationship. To do so, we will employ a range of statistical inference procedures, from group testing to hierarchical multilevel models.

Description of data and source

Our data for this project comes from three sources:

1. NYTimes Covid-19 Data. This data set is publicly available on GitHub and was the source for various New York Times maps and visuals in its coverage of the pandemic. It includes case counts from July 2020 - May 2021, and we have specifically selected cases at Universities. This dataset has 1948 entries and includes 2020 cases, 2021 cases, University IPEDS ID, University Name, State, etc.

2. IPEDS Data Center. This is a publicly available data set on colleges and universities across the globe. It has many possible variables including demographics, admission rates, religious affiliation, etc. The IPEDS data center allowed us to select which variables to include in our data set. The smallest subset of universities that included all from the NYTimes database (by IPEDS ID) was 6125 rows, with all U.S. universities.

3. Centers for Disease Control. This publicly available data set tracks mask mandates in each state from April 8, 2020, to August 15, 2021.

4. The presidential elections data from STAT 139 Problem Set 4. Of this data set, we will be particularly focused on the `gap20repub` predictor, which we will use to help characterize the political climate of the

state in which each institution is located.

After removing universities that were missing lots of data in the pull from the IPEDS data center, we were left with $n = 1,705$ observations.

Data Cleaning Procedures

Our first step in cleaning the data is to read our data from CSV files into R data frames. The `colleges` data frame stores the NYT data on Covid cases at universities, while the `ipeds` data frame stores the data will most of our predictor variables (university characteristics) taken from IPEDS. We then rename most of the columns in `ipeds` to make them shorter and easier to work with.

Next, we merge the `colleges` and `ipeds` data frames on the `ipeds_id` column and remove institutions with no IPEDS data. We then create the `religious` and `private` columns, which are simply indicators for whether an institution has any religious affiliation, whether it has a catholic affiliation, and whether it is a private university. Finally, we drop any unnecessary or redundant columns from the data frame.

We then look to add a column to the data frame that addresses the extent to which mask mandates were present in the state in which each institution is located. We read out the mask mandates data from the CDC into a data frame from the CSV file, treat the appropriate columns as factors, and convert `date` into R's `Date` type.

The next step is to create a new simpler data frame to merge with the product of the last manipulations. This data frame contains only two columns: One with the name of each state, and the other with the number of days between July 1, 2020, and May 26, 2021, during which face masks were required in public in that state. We then merge this data frame with our previous data frame to create `full.cases`, and create a column `total.cases` in `full.cases` that sums the `cases` and `cases_2021` columns.

Finally, we read the presidential election data from Problem Set 4 into a data frame, create a two-column data frame with the columns `state` and `gap20repub`, and merge this data frame with our data frame of observations on the `state` variable.

Description of variables

After performing the data cleaning procedures outlined above, we are left with the following predictors:

- **ipeds_id** This is the IPEDS id, specific to each institution and a commonly used identifier for Universities, created by IPEDS.

- **institution.name** This is the institution name, as filed by each University with IPEDS.

- **state** The state the university is in. This variable has state titles, not abbreviations or codes.

- **private** "A classification of whether an institution is operated by publicly elected or appointed officials or by privately elected or appointed officials and derives its major source of funds from private sources" [IPEDS]. This is simplified to an indicator: "Yes" for private, "No" for a public school.

- **religious** Indicator with "Yes" if the institution is religious, "No" if not.

- **tuition** "Published tuition and fees, 2020-21 for academic year reporters only" [IPEDS].

- **total.headcount** "12-month unduplicated headcount: 2020-21" [IPEDS].

We elected to include the following demographic information to account for established inequalities in wealth and access to healthcare in the United States.

- **percent.american.native** "Percent of 12-month unduplicated headcount that are American Indian or Alaska Native" [IPEDS].

- **percent.asian** "Percent of 12-month unduplicated headcount that are Asian" [IPEDS].

- **percent.black** "Percent of 12-month unduplicated headcount that are Black" [IPEDS].

- **percent.hispanic.latino** "Percent of 12-month unduplicated headcount that are Hispanic or Latino" [IPEDS].

- **percent.pacific.islander** "Percent of 12-month unduplicated headcount that are Native Hawaiian or Other Pacific Islander" [IPEDS].

- **percent.white** "Percent of 12-month unduplicated headcount that are White" [IPEDS].

- **percent.two.more.races** "Percent of 12-month unduplicated headcount that are two or more races" [IPEDS].

- **percent.NA.race** "Percent of 12-month unduplicated headcount that are race/ethnicity unknown" [IPEDS].

- **percent.women** "Percent of 12-month unduplicated headcount that are women" [IPEDS].

The following predictors were included to better characterize each institution, including the opportunities available to and the general level of affluence of the student body, as these factors may be related to access to healthcare or attitudes towards the Covid-19 pandemic.

- **grad.rate** "Graduation rate of first-time, full-time degree or certificate-seeking students" [IPEDS]. Included in our analysis to better characterize each institution, its student body, and its

- **percent.fin.aid** "Percent of full-time first-time undergraduates awarded any financial aid" [IPEDS].

- **percent.student.loan** "Percent of full-time first-time undergraduates awarded student loans" [IPEDS].

- **occupational.degree** Indicator for if an occupational certificate, degree, or other formal award are offered at the institution.

- **hs.equivalent.degree** Indicator for if a high0-school-equivalent certificate, degree, or other formal award are offered at the institution.

Other included predictors:

- **on.campus.housing** Indicator for if on-campus housing is offered at the institution.

- **dorm.capacity** "The maximum number of students that the institution can provide residential facilities for, whether on or off campus. (off-campus dormitory space that is reserved by the institution)" [IPEDS].

- **college** See `institution.name`.

- **cases**

- **cases_2021**

- **mask.mandated.days**

- **total.cases**

- **gap20repub**

Group Testing

To begin our analyses, we will perform a range of group tests to determine whether the institutions with a religious affiliation had a higher true average number of Covid cases than religious institutions without a religious affiliation. In the following analyses, we will test whether or not the true average number of cases in the fall semester — i.e., the case counts for just 2020 in the NYT data set, `cases` — were different for these two groups of institutions, as well as whether the true average number of cases for the entire academic year — see the description of `total.cases` — different for the two groups of institutions. We will perform both sets of analyses, because they are so fast and easy to perform and interpret. Later, as we fit linear models to perform more complicated statistical inference procedures with respect to our data set, we will demonstrate

why a focus will be placed on case counts from the fall semester only (see "Determining whether data from 2020 or 2020 and 2021 should be used").

Checking the assumptions for t-based methods

We begin with the most commonly used test for a difference in means: the Student-t test. Of course, before we can perform this test, we must make sure that its underlying assumptions are reasonable. Because we have no reason to assume that the variances in the observations between both groups — religiously affiliated and not religiously affiliated universities — are the same, we will use the unpooled t-test. For unpooled t-based test for a difference in sample means, there are three assumptions:

1) **Observations are independent.** We claim that this observation is reasonable. Even though the geographic proximity of universities might have caused there to be some correlation between case counts at these different schools, we claim that since universities are somewhat insular communities — students tend to stay within their own social sphere of their university — this is not a major concern. Overall, with such a rich data set, in which a wide and diverse range of characteristics are expressed across all 1,705 observations, this assumption of independence of observations seems reasonable.

2) **Groups are independent of one another.** There is no reason to suggest that the two groups in question — religious institutions and non-religious institutions — are not independent from one another. Above, we justified that all of the observations in our data set are sufficiently independent from one another — if this is true, then it is reasonable to assume that these two groups of institutions are independent of one another.

3) **Observations are normally distributed.** As is shown in the plot below (left), both of the response variables in question are *not* normally distributed; after log-transforming these responses, however, we see that their distributions are symmetrical enough to satisfy this assumption. Of course, we do not need to seek perfect normality with respect to these distributions, since t-based statistical inference procedures are robust to this assumption (as was demonstrated in Question 5 in Problem Set 3).
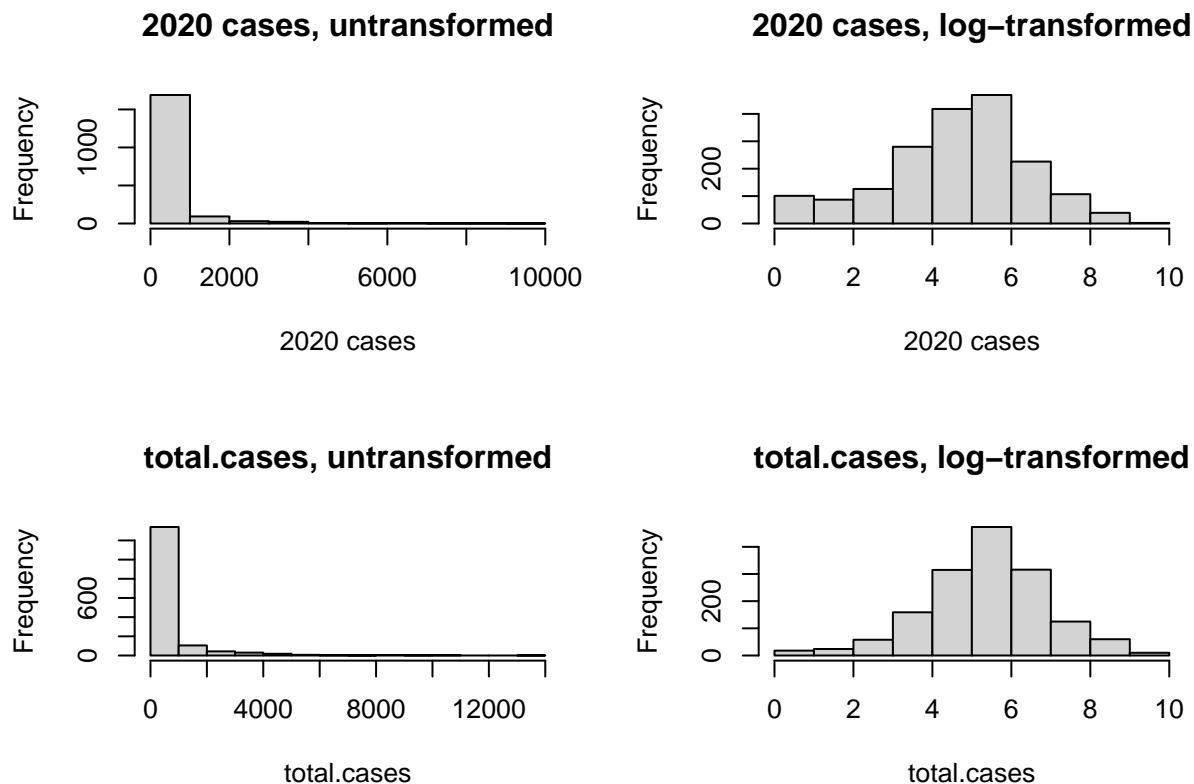
The Normal Q-Q Plots below corroborate the conclusion that this assumption is reasonable. In each plot,

the data for the most part follows the line that we would expect if it were perfectly normally distributed; although, since the data sits below the line at both the positive and negative extremes of the graph, the left tail is a bit fatter and the right tail a bit skinnier than the perfectly normal distribution in each of these sample distributions.
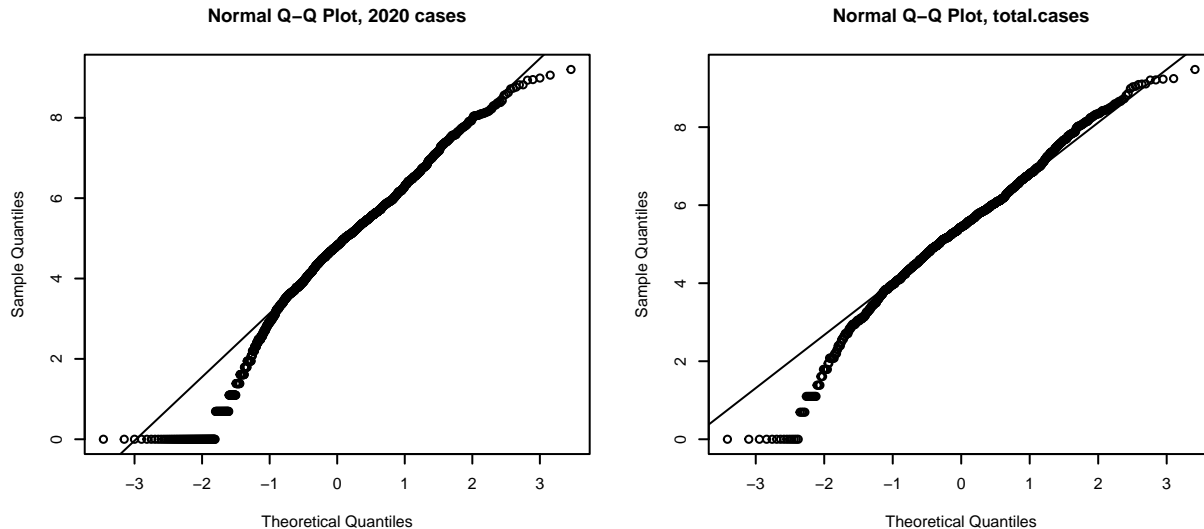
Student-t Tests

Now that we have shown the assumptions of Student-t tests to be reasonable, we can proceed to perform the tests themselves. We perform two two-sided t-tests for a difference in means. In each case, the null hypothesis is that the true mean case counts in the two groups — religiously affiliated and non-religiously affiliated institutions — are equal over the given time period, and the alternative is that they are not equal. As in the rest of the paper, we will use the standard confidence level of $\alpha = 0.05$, for the sake of convenience and conformity with the wider statistics community. Below are the results from these tests.

sample	test.statistic	df	p.value
2020 only	1.379947	969.4587	0.1679212
entire year	-0.564321	885.7181	0.5726786

In both tests, we see that our p-value is greater than the significance level of 0.05; we thus fail to reject the null hypothesis in both cases — we do not have statistically significant evidence to suggest that the true mean case counts, both for the fall semester and for the entire year, are different between the religiously affiliated and non-religiously affiliated institutions. Note that the degrees of freedom — which are not integers owing to R's use of the Welch approximation — are different for the two tests. This is because the test using the `total.cases` data has fewer observations, since some schools in our sample did not report case counts for the spring semester. This issue is tackled in greater detail below (see "Determining whether data from 2020 or 2020 and 2021 should be used").

Non-Parametric Testing — Wilcox Rank Sum Test

We will also perform the non-parametric Wilcox Rank Sum Test to determine whether the case counts are different between religiously affiliated and non-religiously affiliated schools. The strength of this non-parametric test lies in the fact that it does not rely on an assumption on the distribution of the data-generating process of the sample. Thus, though we are confident in having shown above that the distributions of the log-transformed responses are sufficiently approximately normal, for the sake of completeness, we will perform this non-parametric test, and see if it leads us to the same conclusions.

Again, we will perform the Wilcox Rank Sum Test on data from just the fall semester and on data from the entire academic year. For the following two tests, the null hypothesis is that the true average quantiles within the two groups — when the data from both groups are ranked together — are the same, while the alternative hypothesis is that there is an association between group status and the average quantile of the observations in the entire population.

| sample	test.statistic	p.value
2020 only	320824.0	0.2990076
entire year	225800.5	0.7547593

As the results above show, neither test was statistically significant at the 0.05 confidence level; we thus again fail to reject the null hypothesis, concluding that there is no statistically significant evidence to suggest that the there is an association between religious affiliation and average quantile of case counts in either sample.
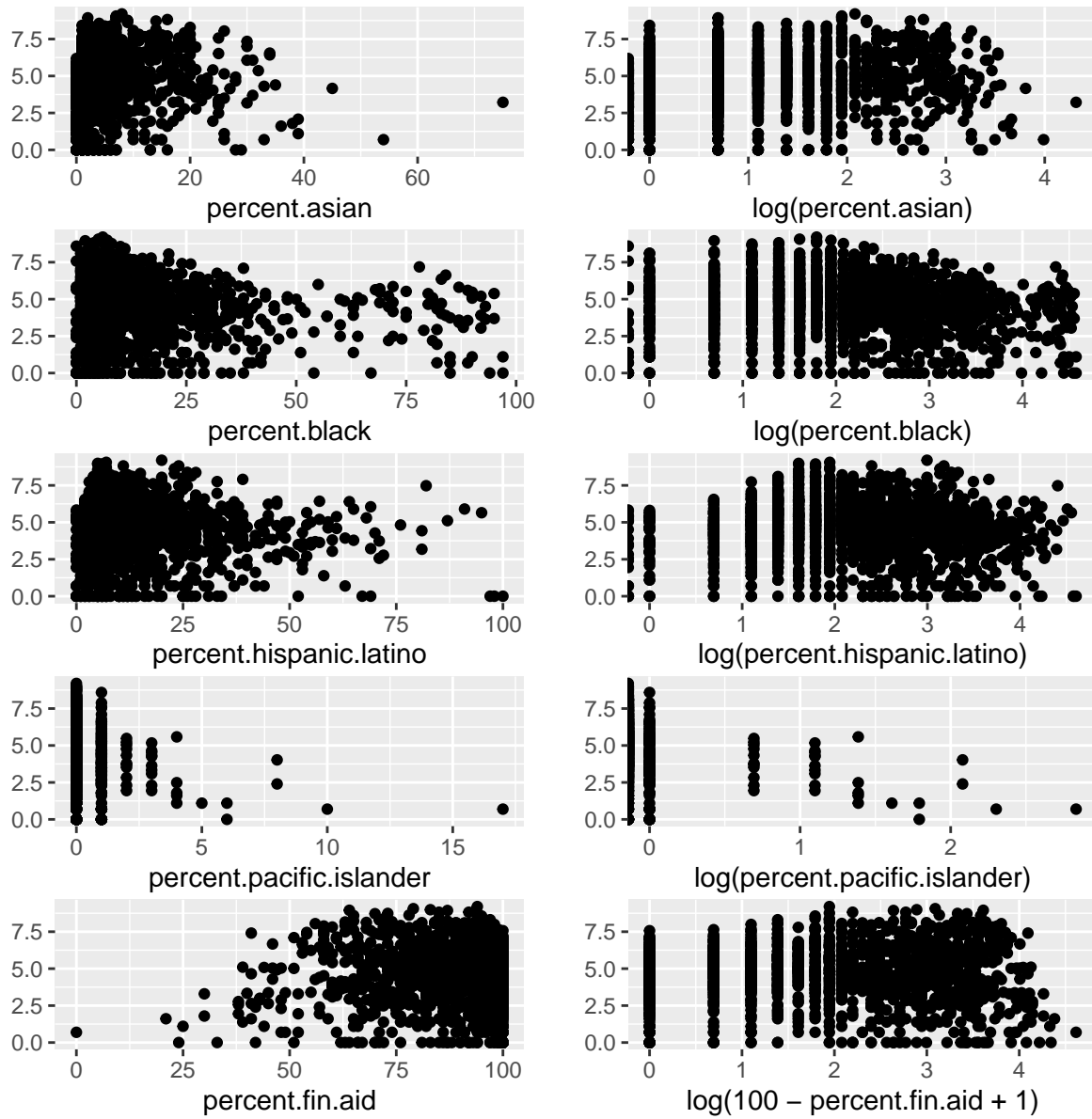
Basic Linear Regression Models

Of course, group tests are limited in that they do not take into account potential confounding variables that might reveal the significance of an institution's having a religious affiliation. To take into account the effects of these potential confounding variables — which we have already identified at length, as evidenced by the extensive list of predictors we compiled before beginning our analyses — we will fit linear models. We will then use these linear models to perform statistical inference on the coefficient of the `religious` predictor, determining whether there is a statistically significant relationship between an institution's religious affiliation and the number of cases that it recorded.

### Checking the Assumptions of Linear Regression Models

Before we can fit more complicated models, we must first check the assumptions of linear regression. We begin by checking the assumption of linearity. To do so, we plotted `cases` versus all of the quantitative predictors that we would like to include in our analyses. We the identified that `total.headcount`, `percent.american.native`, `percent.asian`, `percent.black`, `percent.hispanic.latino`, and `percent.pacific.islander` would benefit from being log-transformed; given the left-skewness of its distribution, we also determined that `percent.fin.aid` would be best transformed using the following transformation `log(100 - percent.fin.aid + 1)`. The following plots show that the distribution of these predictors before and after being transformed. *In these plots, the y-axis is always log-transformed cases* (2020); *the axis label is not printed in order to save space.*
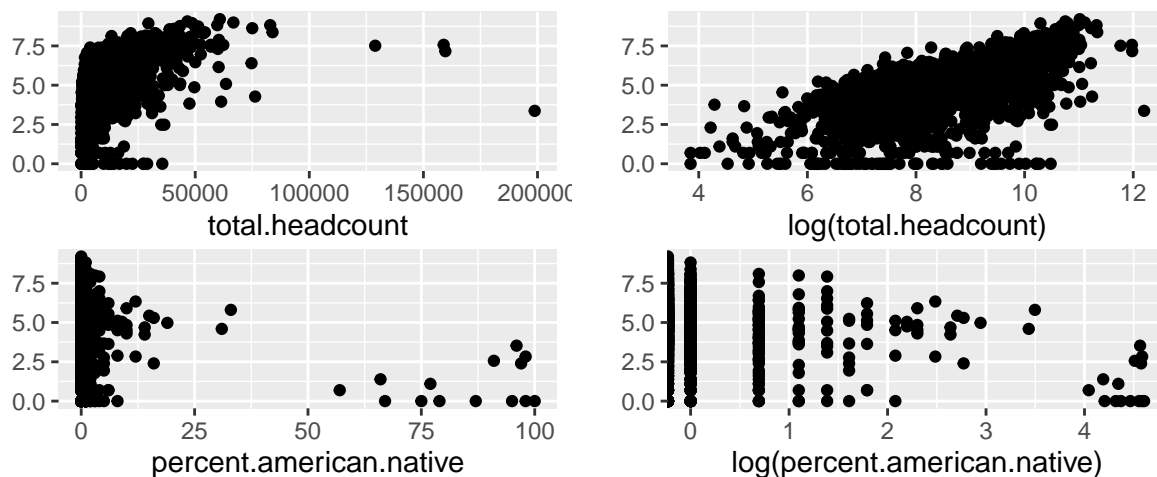
We claim that with these transformations, the assumption of linearity is reasonable.

What remains to be shown is that the assumption of homoskedasticity is reasonable. In order to check this assumption, we fit basic regression models for `cases` and `total.cases` using all of the predictors in our predictor set.

cases (2020)
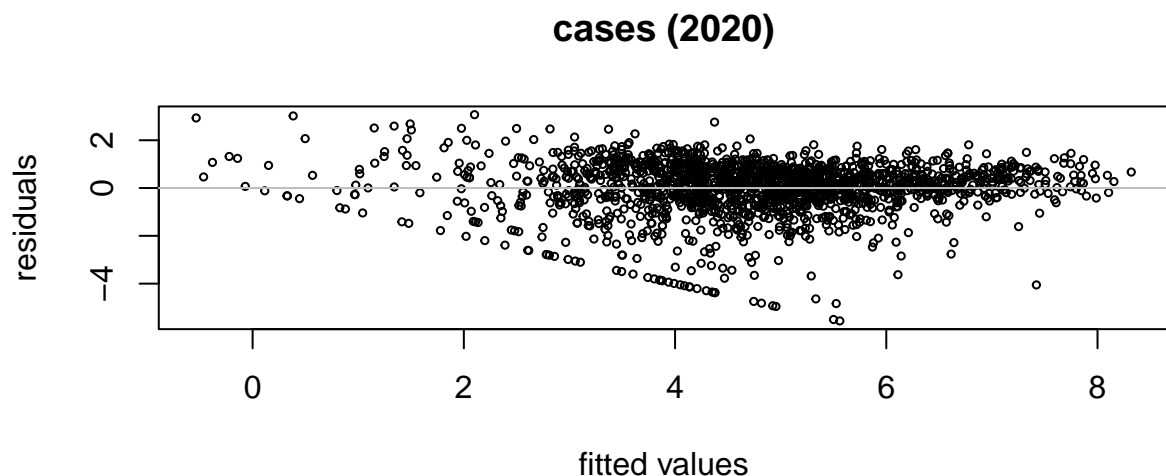
As is shown in the plot above, the spread of the residuals is not constant across the entire range of fitted values — thus, it is called into question whether the assumption of homoskedasticity is reasonable in this case. The behavior of this graph is interesting: Towards the left extreme of the graph, there appear to be bands of observations whose positions on this plot form negatively sloped lines. The left-most line corresponds to all those observations where the case counts are zero; the second-left-most line corresponds to the observations where the case counts are one; the next corresponds to the observations where the case count is two, etc. As one moves closer to the center of the plot, this pattern disappears, and the plot assumes a form that is much better suited to linear regression. Thus, this problem arises from the fact that our response variable is discrete, and from the fact that we have so many observations with very low case counts. As it stands, we must be wary of this assumption of homoskedasticity, which might affect the standard errors that we will use to calculate the statistical significance of coefficient estimates. Later in the paper, we will perform a sensitivity analysis using Poisson regression to corroborate the results we obtain using Gaussian regression.

Below, we recreate the same plot, but for the case counts for the entire 2020-21 academic year.

total.cases (2020–21)
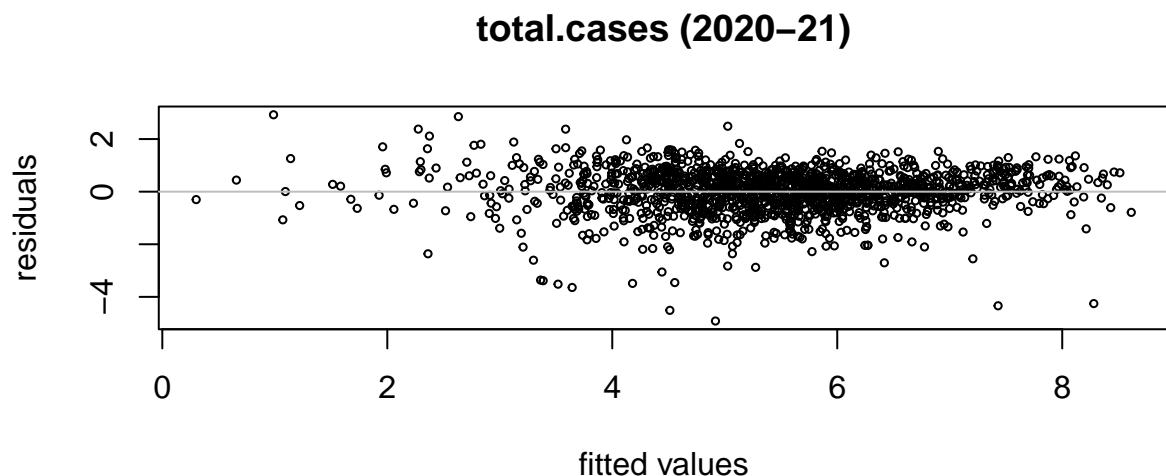
In the above plot for the model for `total.cases`, the spread of the residuals appears much more consistent across the entire range of fitted values. If one looks carefully enough, however, they will see that the same behavior observed in the preceding plot is present in this one, the difference being that there are fewer schools that reported very low case counts (0, 1, or 2). This is likely due to the fact that, with an extra five months of data, our response took on enough different values to begin to behave more like a continuous variable; in addition, a not-insignificant number of schools that simply did not report case data for the spring semester (297 observations).

	all schools		schools with data for both years		schools with only 2020 data	
	Est.	p	Est.	p	Est.	p
(Intercept)	−5.719	<0.001	−4.975	<0.001	−2.338	0.109
religiousYes	0.102	0.252	0.318	<0.001	−0.701	0.049
tuition	0.000 02	<0.001	0.000 01	<0.001	−0.000 003	0.878
log(total.headcount + epsilon)	0.920	<0.001	0.915	<0.001	0.491	<0.001
log(percent.american.native + epsilon)	0.013	0.811	0.170	<0.001	−0.300	0.072
log(percent.asian + epsilon)	−0.125	0.026	−0.225	<0.001	−0.0006	0.998
log(percent.black + epsilon)	0.195	<0.001	0.205	<0.001	−0.057	0.716
log(percent.hispanic.latino + epsilon)	0.060	0.172	0.075	0.046	0.223	0.110
log(percent.pacific.islander + epsilon)	−0.289	0.015	−0.366	<0.001	0.118	0.719
percent.white	0.018	<0.001	0.015	<0.001	0.014	0.017
percent.two.more.races	−0.025	0.026	−0.015	0.115	−0.066	0.097
percent.women	−0.011	<0.001	−0.014	<0.001	0.002	0.784
grad.rate	0.021	<0.001	0.020	<0.001	0.017	0.037
log(100 - percent.fin.aid + epsilon)	−0.043	0.162	0.024	0.349	−0.275	0.019
on.campus.housingYes	0.823	<0.001	0.670	<0.001	1.096	<0.001
gap20repub	0.013	<0.001	0.012	<0.001	0.028	<0.001
privateYes	−0.677	<0.001	−0.377	0.003	−0.475	0.331
percent.student.loan	0.005	0.002	0.007	<0.001	−0.0006	0.923
mask.mandated.days	0.0002	0.545	−0.0007	0.001	−0.000 07	0.941
occupational.degreeYes	−0.184	0.017	−0.011	0.865	−0.938	<0.001
hs.equivalent.degreeYes	0.136	0.154	−0.083	0.287	0.314	0.381
Num.Obs.	1705		1469		236	
R2	0.604		0.694		0.408	
R2 Adj.	0.599		0.690		0.353	
Log.Lik.	−2590.545		−1808.194		−424.879	
RMSE	1.11		0.83		1.46	

Determining whether data from 2020 or 2020 and 2021 should be used

Naturally, there is a temptation to focus on case data for the entire academic year (the entire time period during which case data was collected by the *New York Times*). There is one problem, however: While the *New York Times* was able to compile case data for all of the schools in our data set for the fall semester of that year, it was not able to find data for 297 schools for the spring semester. This is a nontrivial number of observations given that our data set only consists of 1,705 different institutions; thus, to determine whether it would be sound to remove these 297 observations and fit models and perform statistical inferences on data from the entire academic year, we must determine whether the two groups of schools in question — those that reported data for the entire academic year, and those that did not — are sufficiently similar to one another.

The first step is to compare the coefficients of our baseline model (`lm6` above) when it is fit to all 1705 observations, as well as individually to the two different groups of schools in question. The coefficient estimates, along with the p-values (not adjusted to be heteroskedastically consistent) are given below.

As we can see, when the sample of the institutions used to fit the model changes, some of the coefficient estimates change vastly. Take, for example, the coefficient estimate for `religionYes`, which can be interpreted as the natural log of the change in cases that we would expect if a given school were to have a religious affiliation rather than not have one. This coefficient estimate is not statistically significant when all institutions are considered together, is positive and statistically significant when the institutions that reported data for the entire academic year are considered, and negative and statistically significant when the institutions that only reported data for the fall are considered. In fact, if one were to inspect the table above, they would notice that the majority of predictors included in this baseline model experienced changes in their coefficient

estimates and statistical significance as the sample of institutions was changed.

Clearly, then, there are some underlying differences between the schools that did and the schools that did not report case data for the spring semester of the 2020-21 academic year. Consider the coefficient estimate for `religiousYes`, which is quite different in the two models fit to data from only one of the two subgroups; this is especially interesting for our research purposes. Indeed, searching for the differences between these subgroups in terms of `religious` and other predictors might help us to answer the questions about the association between religious affiliation and reported Covid cases that are motivating this paper.

First, we investigate the difference in the political leanings of the states in which these schools were located, as given by `gap20repub`. To do so, we perform a t-test (the assumptions that allow us to do so have been explored above), with the null hypothesis being that the true means `gap20repub` are equal between the two groups of institutions, and the alternative being that the true means are different. The result of this test is shown below.

	predictor	test.statistic	df	p.value
t	gap20repub	-6.289246	445.719	0

With a p-value that is much smaller than our $\alpha = 0.05$ (again, because we do not assume equal variances between the groups, the Welch approximation for degrees of freedom is used), we reject the null hypothesis, concluding that the true mean `gap20repub` is different for the schools that did report case data for 2021 versus those that did not. As it turns out, the 95% t-based confidence interval for the mean `gap20repub` for the schools that did report 2021 case data minus the mean `gap20repub` for the schools that did not is (-9.468120, -4.959647) — thus, we conclude that the schools that did report 2021 case data are located in states that are considerably more left-leaning than the schools that did not.

We repeat this sort of analysis for `total.headcount`, `tuition`, `percent.fin.aid`, `mask.mandated.days`, `on.campus.housing`, and `religious`, predictors that we have chosen based on large differences in the coefficient estimates produced by the models above when fit to only one of the two subgroups of institutions. We use a two-sided t-test for the quantitative variables and two-sided z-tests for proportions (ADDRESS) for the `categorical` predictors. The hypotheses are similar to above, the null being that the true means/proportions are equal between the two groups, and the alternative being that they are not equal. As per the results shown below, the tests for all six of these predictors showed statistically significant differences between the two groups of universities.

predictor	test.statistic	df	p.value
total.headcount	10.4880993	872.6989	0.0000000
tuition	3.7184096	397.6510	0.0002293
percent.fin.aid	0.5400659	349.3054	0.5894957
mask.mandated.days	8.9492453	376.7757	0.0000000
on.campus.housing	26.5612204	1.0000	0.0000003
religious	14.2542757	1.0000	0.0001597

The existence of these statistically significant differences between these two groups of institutions leads us to two distinct conclusions: Firstly, that we should perform all further analyses on data only from the fall 2020 semester, and secondly, that we should consider interaction effects in our model to help better isolate and take into account these demonstrated differences between these two different groups of institutions.

Linear Models with Interaction Effects

Above we performed several linear regressions with no interactions and multiple predictors. Our "base model" includes all of the predictors we include in this paper, which are: `religious`, `tuition`, `total.headcount`, `percent.american.native`, `percent.asian`, `percent.black`, `percent.hispanic.latino`, `percent.pacific.islander`, `percent.white`, `percent.two.more.races`, `percent.women`, `grad.rate`, `percent.fin.aid`, `on.campus.housing`, `gap20repub`, `private`, `percent.student.loan`, `mask.mandated.days`, `occupational.degree` and `hs.equivelant.degree`. Above we discuss using 2020 cases, 2021 cases, or the total number of cases, concluding that we will use only 2020 cases for the remaining models.

We will first look at three linear models with all of the predictors in this paper: our "base model," an interaction model between the `religious` predictor and the others (`religiousInteraction`), and a full interaction model (`fullInteraction`).

```
##                                           coefficients        p.value
## tuition                                   1.739771e-05   1.611698e-05
## log(total.headcount + epsilon)            9.199545e-01  6.421515e-151
## log(percent.asian + epsilon)             -1.251276e-01   2.570484e-02
## log(percent.black + epsilon)              1.946336e-01   8.798378e-06
## log(percent.pacific.islander + epsilon)  -2.890057e-01   1.462735e-02
## percent.white                             1.788797e-02   1.144046e-22
## percent.two.more.races                   -2.547028e-02   2.626922e-02
## percent.women                            -1.076282e-02   2.452712e-05
## grad.rate                                 2.101062e-02   2.383138e-19
## on.campus.housingYes                      8.225306e-01   4.661025e-18
## gap20repub                                1.274761e-02   2.534125e-11
## privateYes                               -6.765733e-01   5.084869e-06
## percent.student.loan                      5.314923e-03   1.791965e-03
## occupational.degreeYes                   -1.840886e-01   1.735707e-02
```

Above is the base model. We see that there are a lot of statistically significant coefficients on the predictors, holding all else constant. These are noted with a star next to them in the above table. Notably, the coefficient for the `religious` variable does not carry statistical significance, with a p-value of .25190, significantly above $\alpha = 0.05$. This could be for a number of reasons, including that 1) Religious affiliation (either Yes: Religious or No: Not Religious) simply has no correlation with 2020 Covid cases or 2) The `religious` variable shares predictive power (has some degree of collinearity) with other predictor variables.

We will determine this relationship by fitting the `religiousInteraction` model; a linear model that includes the interactions between the religious variable and the remaining confounding variables.

```
##                                                      coefficients        p.value
## religiousYes                                         3.665377e+00   2.895900e-04
## tuition                                              1.749468e-05   7.125827e-04
## log(total.headcount + epsilon)                       9.491104e-01  6.266898e-129
## log(percent.black + epsilon)                         2.514204e-01   3.332000e-07
## log(percent.pacific.islander + epsilon)             -3.163746e-01   2.291030e-02
## percent.white                                        1.903071e-02   7.737699e-20
## percent.women                                       -1.105465e-02   2.047427e-04
## grad.rate                                            2.223123e-02   2.615399e-18
## on.campus.housingYes                                 8.374462e-01   4.790326e-18
## gap20repub                                           1.318101e-02   6.180288e-09
## privateYes                                          -6.565577e-01   2.534542e-04
## percent.student.loan                                 5.515932e-03   4.141881e-03
## religiousYes:log(percent.american.native + epsilon) -4.244315e-01   1.001051e-02
## religiousYes:log(percent.black + epsilon)           -3.386581e-01   2.683179e-03
## religiousYes:grad.rate                              -1.332992e-02   4.141586e-02
## religiousYes:on.campus.housingYes                   -1.664619e+00   1.366724e-03
## religiousYes:mask.mandated.days                      2.275318e-03   6.433301e-05
```

The `religiousInteraction` model clarifies the significance of the `religious` predictor: When accounting for the interactions between `religious` and other predictors, as seen in this model, `religious` becomes a statistically significant predictor, with a p-value of 0.000290, significantly below the p-value of $\alpha = 0.05$. This indicates that the previous "base model" had `religious` as an insignificant predictor because it interacted with other variables which were collinear and therefore shared a lot of predictive power. Given the number of confounding variables, this result makes sense. We see in this model that the coefficient on the `religiousYes` variable is 3.665, indicating that when looking at a religiously-affiliated school versus one that is not, holding

all other variables constant, the religious affiliation alone implies an increase of $e^{3.665}$ Covid cases, which is a significant increase.

`religious` had a statistically significant interactive term with 5 other variables, which were: `percent.black`, `percent.american.native`, `grad.rate`, `on.campus.housing`, and `mask.mandated.days`. For example, the term `religiousYes:mask.mandated.days` in the summary results of the interaction model had a coefficient of 2.275e-03 and a p-value of 6.43e-05. This implies that it is statistically significant in this model, and for religiously affiliated schools, a one-day increase in mask mandates correlates with an increase in log Covid cases compared to its non-religious counterpart (specifically $e^{0.002275}$. This result is interesting, and could prompt further investigation in a different paper.

Given the significance of the interaction terms in our partial interaction model, we will now look at a full interaction model with all predictive variables in this paper to predict the 2020 Covid cases. Significant results relating to the `religious` predictor are printed below.

```
##                                        coefficients    p.value
## religiousYes:log(percent.asian + epsilon)  -0.39360179 0.04228886
## religiousYes:log(percent.black + epsilon)  -0.43880409 0.02889200
## religiousYes:on.campus.housingYes          -1.49643791 0.03841221
## religiousYes:mask.mandated.days             0.00177513 0.03507104
```

The `religious` variable in the above model with all interactive terms is not statistically significant in predicting 2020 Covid cases at Universities, with a p-value of 0.290212. But, very few terms alone are significant in the model; only `percent.black`, `percent.two.more.races`, and `on.campus.housing`. Even the predictor `total.headcount` is not statistically significant in this model, whereas it is consistently very significant in other models given the obvious relationship between the total number of people on a given campus and the number of Covid cases amongst them. This result suggests that a lot of the variables are correlated with one another to varying degrees. Thus, there are many interactions, and therefore predictive terms in this model. These have taken the predictive power from the `religious` variable by itself; though some interactions with the `religious` term are statistically significant, such as `percent.asian`, `percent.black`, `mask.mandated.days`, and `on.campus.housing`. This indicates that in relation to 2020 Covid cases, the religious affiliation has an affect on the predictive power of those variables. Ultimately a high number of predictive terms in a model dealing with highly collinear data will inherently diminish the predictive power from a term that shares a lot of correlation with others (referencing `religious`, amongst other terms). With so many predictors, the t-tests used to determine the statistical significance of each coefficient estimate are also not very powerful, as the number of degrees of freedom is quite high.

```
##                     Model R.Squared
## 1            base model 0.6035471
## 2 religiousInteraction 0.6208401
## 3      fullInteraction 0.7215086
```

We see in the above table that the R^2 value for the three models increases as the number of interactive terms increases. This indicates the significance of the interactive terms in the prediction model. It is unsurprising that the greatest increase is between the `religiousInteraction` model and the `fullInteraction` model. This can be explained by the sheer number of interactive terms in the model, as well as the collinearity between a lot of variables in the data.

We now perform ESS F-tests to compare the predictive power the base model to `religiousInteraction` and `religiousInteraction` to `fullInteraction`. In each test, the null hypothesis is that the true coefficients of the additional predictors included in the larger model — `religiousInteraction` in the first test, and `fullInteraction` in the second — are all zero; the alternative hypothesis is that the true coefficient of one or more of these predictors is not zero, i.e., that the additional predictors provide explanatory power. The result of these ESS F-tests are shown below.

```
##                                  ESS.F.test F.statistic  df      p.value
## 1      base model vs. religiousInteraction    4.221356   18 7.246993e-09
## 2 religiousInteraction vs. fullInteraction    3.160287  171 4.805415e-32
```

In the first test, because the p-values of each test are below our $\alpha = 0.05$, we reject the null hypothesis, and conclude that there is evidence that the `religious` interaction terms contribute to the model. Furthermore, in the second ESS F-test between the `religiousInteraction` model and the `fullInteraction` model, we again reject the null hypothesis, concluding that there is evidence that the additional interaction terms in `fullInteraction` have some explanatory power. This indicates that both `religious` interaction with the confounding variables (as discussed above) and the predictors interacting with one another are significant and improve (from the R^2 results) their respective models.

We used a variance-covariance matrix result to view the covariance values between `religious` and all other variables to create a simplified linear regression model with the most correlated confounding variables removed. The refined full interaction model is below, with the terms `private`, `percent.pacific.islander`, `percent.hispanic.latino`, `on.campus.housing`, `percent.native.american`, `gap20repub`, `percent.women`, `grad.rate`, `percent.white`, `percent.student.load`, and `mask.mandated.days` removed.

```
##                                            coefficients      p.value
## religiousYes                               2.223538e+00 0.0046053175
## religiousYes:tuition                       2.387962e-05 0.0014909237
## religiousYes:log(total.headcount + epsilon) -3.492881e-01 0.0001207003
```

We see above that we have artificially made the `religious` predictor significant in the full interactive model by removing terms that it is collinear with. But, this comes at the expense of the R^2 value, now 0.5368185 which has decreased even from the "base model" with all predictors but no interactions. This indicates that the variables we removed are important in the prediction of 2020 Covid cases.

```
## [1] 898.0029
```

```
## [1] 897.9843
```

```
## [1] 898.079
```

In Search of a Parsimonious Model: Sequential Variable Selection Models

Of course, both `religiousInteraction` and `fullInteraction` are large, complicated models with many predictors — this makes them hard to interpret, and also leaves our significance tests diminished in statistical power. In order to find more parsimonious models — i.e., remove some predictors from these models — we performed Sequential Variable Selection using AIC as our model comparison criteria, stepping backward, forward, and in both directions. Brief descriptions and results from each resulting model are below.

Backward step model: For this model, we started with `fullInteraction`, stepping backwards until a reduction in AIC was no longer possible. This process removed 116 predictors, leaving the resulting model, `back.step`, with 94 predictors. This list includes `religiousYes`, as well as interaction between `religious` and eight other predictors: `tuition`, `log(total.headcount + epsilon)`, `log(percent.american.native + epsilon)`, `log(percent.asian + epsilon)`, `log(percent.black + epsilon)`, `percent.two.more.races`, `on.campus.housingYes`, and `mask.mandated.days`. Of the nine predictors related to `religious` included in this model, only 5 were significant at the 0.05 level: `religiousYes:tuition`, `religiousYes:percent.two.more.races`, `religiousYes:log(percent.asian + epsilon)`, `religiousYes:on.campus.housingYes`, and `religiousYes:mask.mandated.days`, in the table below. Thus, according to this model, the statistically significant relationship between `religious` and `cases` changes in nature based on the predictors `log(percent.asian + epsilon)`, `on.campus.housingYes`, `mask.mandated.days`, `percent.two.more.races`, and `tuition`.

```
##                                            coefficients      p.value
## religiousYes:log(percent.asian + epsilon)  -0.380279872 0.0035694845
## religiousYes:on.campus.housingYes          -1.312094381 0.0051891062
## religiousYes:mask.mandated.days             0.002078556 0.0001400012
```

Forward step: For this model, we started with our base model, and stepped forward until a reduction in AIC was no longer possible. This process added 55 predictors, leaving `forward.step` with 75 predictors. In terms of the predictors related to `religious`, this list included `religiousYes`, `religiousYes:mask.mandated.days`, `religiousYes:on.campus.housingYes`, `religiousYes:tuition`, `religiousYes:log(percent.asian + epsilon)`, and `religiousYes:percent.two.more.races`. Interestingly, all of the `religious` interaction terms in this model were statistically significant at the 0.05 level, but the predictor `religiousYes` itself was not. We can interpret this model as saying that whether a school has a religious affiliation changes the relationship between `cases` and these other predictors, but that `religious` does not have a significant association with `cases` in of itself.

```
##                                            coefficients      p.value
## religiousYes:mask.mandated.days             2.086638e-03 1.023453e-05
## religiousYes:on.campus.housingYes          -1.280268e+00 4.999928e-03
## religiousYes:tuition                        2.037133e-05 6.101862e-03
## religiousYes:log(percent.asian + epsilon)  -3.071756e-01 1.071404e-02
## religiousYes:percent.two.more.races         6.124955e-02 4.001867e-02
```

Forward step (from back): For this model, we started with the back step model (described above, with 94 predictors), and stepped forward until a reduction in AIC was no longer possible. This process did not actually add any predictors to the back step model, leaving `forward.step` with 94 predictors. But, the significance of each predictor changed, as only three `religious` interaction terms are statistically significant in this model. These terms are `religiousYes:mask.mandated.days`, `religiousYes:on.campus.housingYes` and `religiousYes:log(percent.asian + epsilon)`.

```
##                                            coefficients      p.value
## religiousYes:log(percent.asian + epsilon)  -0.380279872 0.0035694845
## religiousYes:on.campus.housingYes          -1.312094381 0.0051891062
## religiousYes:mask.mandated.days             0.002078556 0.0001400012
```

Forward step (from intercept): For this model, we stepped forward from an intercept model. As a result, `forward.step1` incorporated 67 predictors. This list of predictors did not include `religiousYes`, or any interaction term with `religious`. This could be taken as evidence that `religious` really is not a relevant predictor at all, at least compared to the many other available predictors and their interactions.

Bidirectional step: For this model, we used our base model as our starting point, and stepped in both directions (alternating forwards and backwards), until an improvement in AIC was no longer possible. As a result, 55 predictors were added to the base model to produce `both.step`, a model with 75 total predictors. This list of predictors for this model was extremely similar to the list of predictors included in the `forward.step` model (from the "base model"); only four predictors were different between the two models, none of which were related to `religious`. Predictably, the same interaction terms with `religious` that were significant in `forward.step` were statistically significant in this model as well (noted below), and `religiousYes` was again not statistically significant. This model corroborates the idea that whether an institution is religious changes the relationship between `cases` and these other predictors, but that there is no relationship between the predictor `religious` and the response `cases`.

```
##                                            coefficients      p.value
## religiousYes:mask.mandated.days             2.008876e-03 0.0000221325
## religiousYes:on.campus.housingYes          -1.415322e+00 0.0022346083
## religiousYes:tuition                        1.789561e-05 0.0203496761
## religiousYes:log(percent.asian + epsilon)  -2.891088e-01 0.0170956127
## religiousYes:percent.two.more.races         6.074036e-02 0.0413932297
```

The results of the five step models explored come to very similar conclusions: though `religious` alone was sometimes included as a predictor and sometimes not, it was never statistically significant in the models. Conversely, five interactive terms stood out, namely, `religiousYes:mask.mandated.days`, `religiousYes:on.campus.housingYes`, `religiousYes:tuition`, `religiousYes:log(percent.asian + epsilon)`, and `religiousYes:percent.two.more.races`. Three of these (`mask.mandated.days`,

14

`on.campus.housing`, and `percent.asian`) were statistically significant in four of the models, and the remaining two were significant in three of those four. The fifth model, the forward interaction model starting from the intercept model, had no statistically significant interactive terms with `religious`. Of these interactions, `mask.mandated.days` was always the most statistically significant with `religious`. Next was `on.campus.housing`, then typically `percent.asian`. The significance of these interaction terms can be interpreted as an indication that the relationship between case counts and the religious affiliation of a school changes based on the values of these other predictors.

```
##                              Model      AIC
## 1                   Backward Step 4816.395
## 2      Forward (from base model) 4833.366
## 3  Forward (from backward model) 4816.395
## 4 Forward (from intercept model) 4910.775
## 5               Both Directions 4828.786
```

We see in the above table that the models with the lowest AIC (this indicates a "better" model) is the Backward Step model and the Forward (from the backward model). These are the same models, as discussed above, as the forward procedure did not add any variables to the starting backward model. Notably, these included all five discussed interactive terms with `religious`. Conversely, the model with the greatest AIC was the Forward direction model starting from the intercept model. This was the model without the `religious` interaction terms.

We thought that the backward step model may be the worst for our data due to the high collinearity, as the procedure would force the model to consider and keep collinear variables because it starts with them. Conversely, a forward stepwise model could entirely omit them from the procedure. While the run time for the backward step model was the longest and is indicative of considering so many interactive terms, it has a higher AIC and does not seem to differ much with variable significance from other step models we considered.

LASSO for Variable Section

Because sequential variable selection can be sensitive to outliers and high-leverage observations, and in order to get a better idea of which of the three sequential variable selection models might be the most reliable, we will employ an alternative to this method: LASSO. A penalized form of regression, LASSO is able to function as a method of variable selection, since, unlike ridge regression, it actually does shrink coefficient estimates to exactly zero.

We will apply LASSO to both the `religiousInteraction` and `fullInteraction` models, and interpret the results of each model. In each case, we will use cross-validation to select the best regularizing constant λ with which to fit each model.

First, we do so for the `religiousInteraction` model. We plot the λ values that we cross validated to make sure that the minimum value given by R is not the lower or upper bound of those considered (which would suggest that we widen our range of λ values considered); we then fit a LASSO model using the formula from `religiousInteraction`; call this model LASSO Model 1.
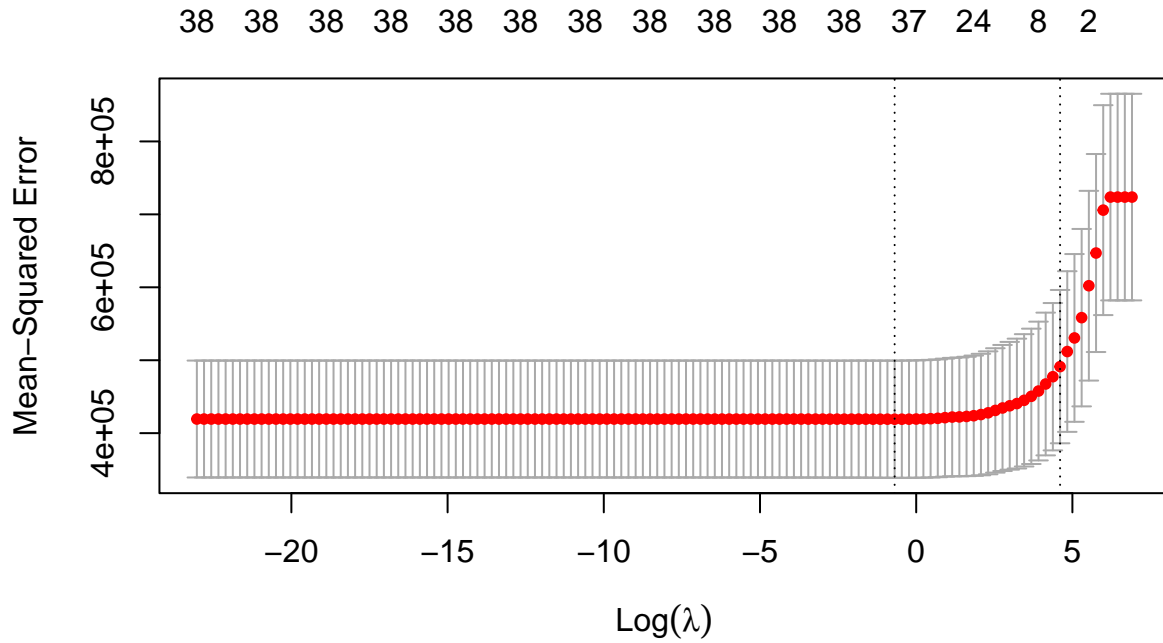
In LASSO Model 1, only the intercept estimate and the coefficient estimate for `religiousYes:percent.women` were shrunk all the way to zero. We now perform the same procedure with the `fullInteraction` model, creating LASSO Model 2:
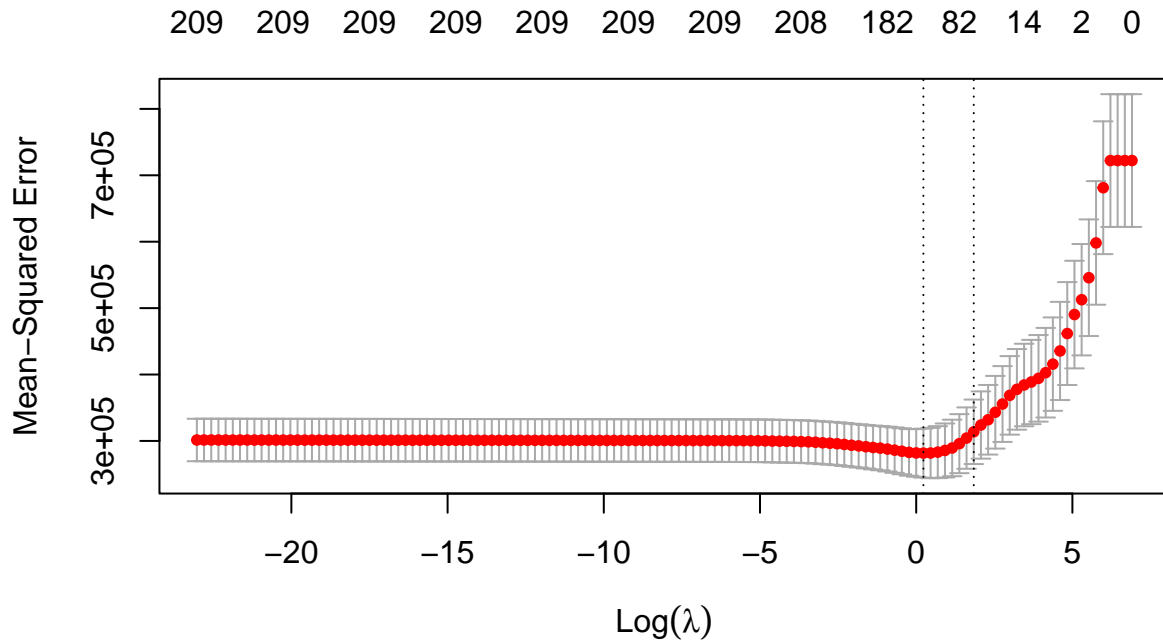
This time around, there are 11 coefficient estimates that were shrunk all the way to zero; of those related to `religious`, only `religiousYes:percent.women` was shrunk to zero.

In order to determine which of the remaining predictors in each model are statistically significant, we refit a linear model using all the coefficients that were not shrunk to zero in each of LASSO Models 1 and 2. Note that even though the intercept was shrunk to zero by LASSO in both cases, we still include an intercept in this new model to make interpretation of the coefficient estimates in the resulting models easier. Call the new linear models LASSO-Selected Linear Model 1 and LASSO-Selected Linear Model 2. Tables 3 and 4 show the coefficients that we found to be statistically significant in each of these two models.

Table 3: LASSO-Selected Linear Model 1

	coefficients	p.value
religiousYes	3.7873072	0.0000828
tuition	0.0000175	0.0006961
log(total.headcount + epsilon)	0.9491069	0.0000000
log(percent.black + epsilon)	0.2503230	0.0000004
log(percent.pacific.islander + epsilon)	-0.3143841	0.0236531
percent.white	0.0190369	0.0000000
percent.women	-0.0104931	0.0000577
grad.rate	0.0222903	0.0000000
on.campus.housingYes	0.8381158	0.0000000
gap20repub	0.0131994	0.0000000
privateYes	-0.6582761	0.0002423
percent.student.loan	0.0055166	0.0041280
religiousYes:log(percent.american.native + epsilon)	-0.4278807	0.0093028
religiousYes:log(percent.black + epsilon)	-0.3328287	0.0029103
religiousYes:grad.rate	-0.0130525	0.0445093
religiousYes:on.campus.housingYes	-1.7000670	0.0008971
religiousYes:mask.mandated.days	0.0022565	0.0000693

In both LASSO-Selected Linear Models 1 and 2, we find that `religiousYes` is statistically significant. Recall that this is the categorical indicator variable for `religious`, which can be interpreted as the natural logarithm of the multiplicative change in cases that we would expect for an institution with a given predictor set if it were to have a religious affiliation, rather than not have one, and all the other predictors were to remain constant.

Interestingly, while five different `religious` interaction coefficients are statistically significant in LASSO-Selected Linear Model 1, only one `religious` interaction coefficient is statistically significant in LASSO-Selected Linear Model 2: `religiousYes:mask.mandated.days`. Thus, while the first model suggests that the relationship between `religious` and `cases` is affected by the value of five other predictors in the model, the second suggests that the association between `religious` and `cases` changes based on the value of `mask.mandated.days` of a given observation. It appears that the interaction effects deemed to be statistically significant in LASSO-Selected Linear Model 1 shared some level of collinearity with some of the additional interaction effects that were included in LASSO-Selected Linear Model 2. In this way, when these other interaction effects were included in the model, the `religious` interaction effects were no longer statistically significant, as the relationships that they helped to describe were better captured by other predictors.

The plot below demonstrates that there does indeed exist collinearity between the predictors of LASSO Model 2 (and thus LASSO-Selected Model 2). We have chosen to present this plot, because it is more readable than a printout of a massive variance-covariance matrix. This plot shows the trajectories of the coefficient estimates of LASSO Model 2 as the regularizing constant λ increases on a logarithmic scale. As is clear, not all of the coefficient estimates are shrunk uniformly towards zero; instead, some display sharp increases in magnitude as λ increases. This demonstrates that there is collinearity between some of the predictors: As one predictor in a collinear pair is shrunk to zero, the magnitude of the coefficient estimate for the other increases in order to make up for the lost predictive power.

Table 4: LASSO-Selected Linear Model 2

	coefficients	p.value
religiousYes	3.1755840	0.0407812
log(percent.black + epsilon)	1.0835193	0.0452619
on.campus.housingYes	4.4174840	0.0035187
religiousYes:mask.mandated.days	0.0016696	0.0474806
tuition:percent.two.more.races	-0.0000049	0.0130804
tuition:grad.rate	-0.0000007	0.0313698
tuition:percent.student.loan	-0.0000006	0.0088660
log(total.headcount + epsilon):log(percent.black + epsilon)	-0.1062718	0.0266786
log(total.headcount + epsilon):grad.rate	0.0089707	0.0000360
log(percent.american.native + epsilon):gap20repub	-0.0118856	0.0093360
log(percent.american.native + epsilon):privateYes	-0.8983133	0.0398615
log(percent.asian + epsilon):log(percent.pacific.islander + epsilon)	-0.5202528	0.0482986
log(percent.asian + epsilon):gap20repub	0.0129322	0.0019242
log(percent.black + epsilon):gap20repub	-0.0096110	0.0001176
log(percent.hispanic.latino + epsilon):percent.white	0.0060773	0.0054232
log(percent.hispanic.latino + epsilon):grad.rate	-0.0091300	0.0122518
percent.white:on.campus.housingYes	-0.0143575	0.0338577
percent.two.more.races:mask.mandated.days	-0.0002355	0.0407143
percent.two.more.races:occupational.degreeYes	-0.0736443	0.0427468
percent.women:log(100 - percent.fin.aid + epsilon)	0.0054505	0.0291722
percent.women:on.campus.housingYes	-0.0285868	0.0046361
percent.women:gap20repub	0.0006105	0.0003450
percent.women:percent.student.loan	0.0003121	0.0334358
grad.rate:log(100 - percent.fin.aid + epsilon)	0.0060796	0.0205208
grad.rate:gap20repub	-0.0004421	0.0081894
grad.rate:hs.equivalent.degreeYes	-0.0144991	0.0411769
on.campus.housingYes:percent.student.loan	0.0149356	0.0082776
on.campus.housingYes:mask.mandated.days	-0.0021618	0.0119621
on.campus.housingYes:occupational.degreeYes	-0.8963147	0.0024896
on.campus.housingYes:hs.equivalent.degreeYes	-0.5515519	0.0392638
percent.student.loan:hs.equivalent.degreeYes	0.0198189	0.0004598

LASSO Model 2: Coefficient Estimate Trajectories
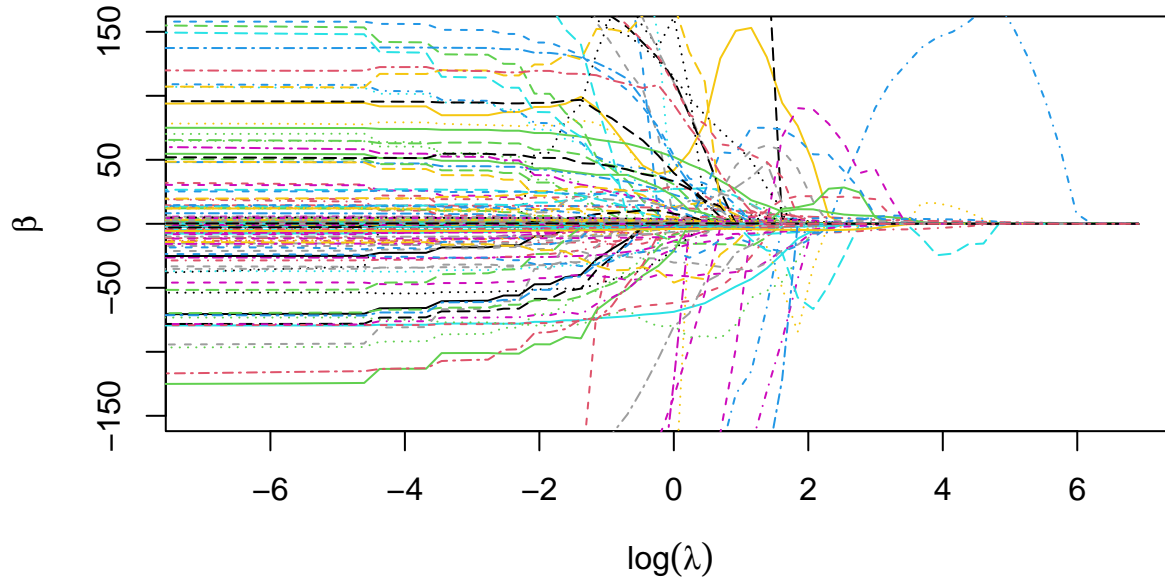
In the end, we lean towards accepting the conclusions presented by the second model: With a wider range of interaction effects to consider, it is less likely that the statistically significant predictors were only marked as such because they help explain a relationship between `cases` and another predictor with which they are collinear. The conclusion of the LASSO models alone appears to be that there is a statistically significant relationship between `religious` and `cases`, and that the nature of this relationship changes based on the value of `mask.mandated.days`. Interestingly, in LASSO-Selected Model 2, the interaction coefficient `religiousYes:mask.mandated.days` has a positive coefficient: An increase in the number of mask mandated days during the 2020-21 academic year is associated, in this model, with an increase in the multiplicative change that we would expect for a given school if that school were to have a religious affiliation, compared to if that same school were to not have a religious affiliation. This phenomenon also seems to be present in LASSO-Selected Model 1, in which the coefficient estimate for `religious:mask.mandated.days` is also positive.

Hierarchical Multi-level Models

We next chose to explore the relationship between 2020 Covid cases and our list of predictors using hierarchical multi-level models, namely the mixed effects models from the `lme4` package in R. We thought this model type would work well with our data because the data can be organized hierarchically as follows: The first level is at the university level; each university has predictors, such as `grad.rate` and `religious`, which are specific to each university. The second level in the hierarchy is the state level, where Universities in each specific state will share cultural, legislative, etc. characteristics that can be vaguely accounted for by sorting the "groups" by states. For example, mask mandates, which were decided by state, were in effect at that stage of the pandemic. Finally, these models have another advantageous capability of being able to account for "weight," meaning we can account for the affect `total.headcount` has on the number of Covid cases.

A capability of the `lme4` mixed effects package is that it can produce both random intercept and random slope models. We decided that with the data we have, we can only look at random intercept mixed effects models. This is due to the number of observations we have; there are a total of 1855 Universities, which means that in each group (by state + Washington DC), there is an average of just over 36 data points per group. Especially given the differing densities of Universities (there are so many in Massachusetts), there are simply not enough observations per group to justify using and analyzing a random slope model. Fortunately, the nature of these mixed effects models means they will not overfit to the fewer observations in a given group.

Massive model - this is what I am unsure on, if this is a valid use of all of the predictors.

First we fit a model with all of the predictors from our "base model", except `total.headcount`, which we accounted for using the "weights =" functionality of the `lmer` modelling. We used a random intercept model and grouped by state, meaning that the results include the coefficients for each predictor, which are fixed by state, as well as the intercept, which should differ by state. Below is a model of the coefficient values, which we will use to best visualize a line graph.

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```


log(cases + 1)
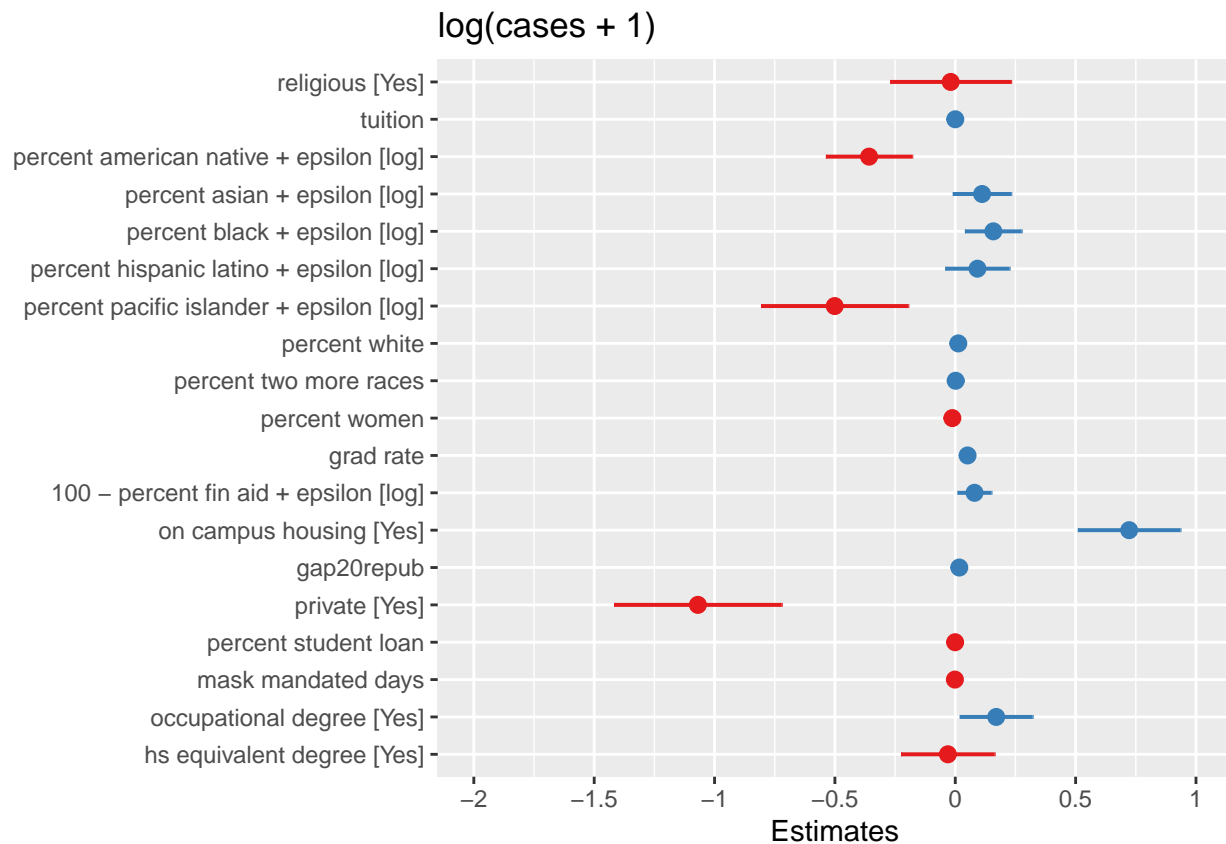
We see in the above model that certain coefficent estimates are significantly larger than the others. It is significant to note that `religiousYes` actualy has a negative affect on the 2020 Covid cases in this model, which is contrary to all previous models we have explored.

We chose `percent.black` as for the x-axis below to visualize the process for this random intercept model because unlike `private` and `on.campus.housing`, it is a continuous predictor, and unlike `percent.pacific.islander` and `percent.american.native`, it does not have a lot of Universities with 0%. Otherwise, it has the largest coefficient, and will therefore create a more dramatic graph.
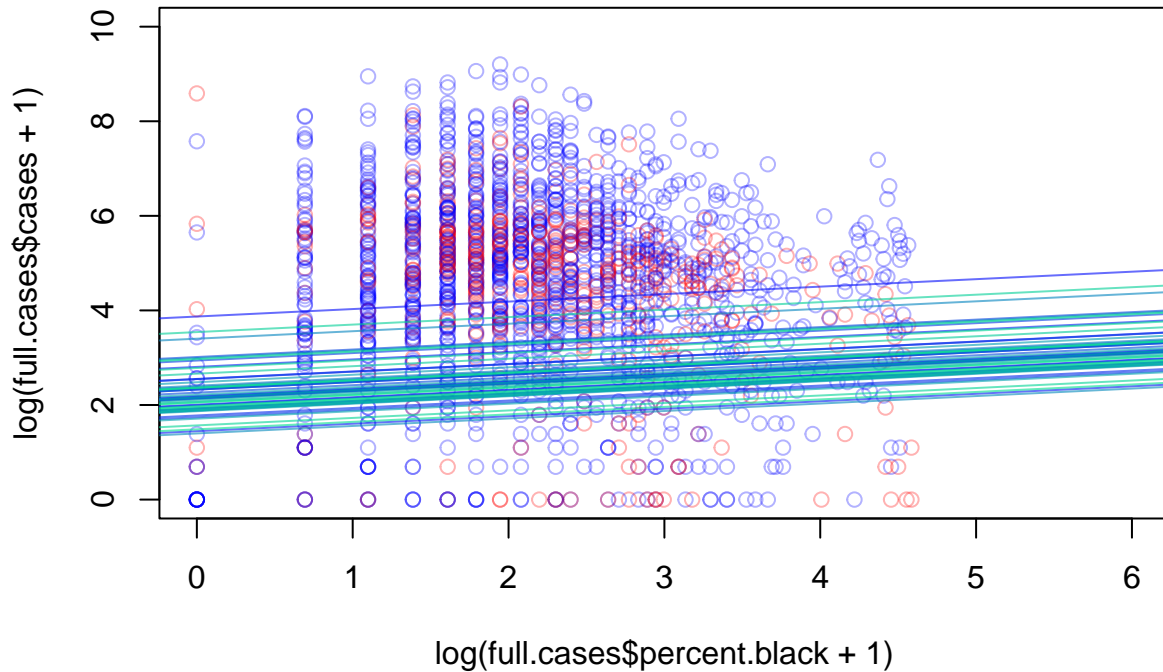
The above plot explores the relationship between `percent.black` and 2020 Covid cases in institutions, as well as plots the lines for each state (and DC) generated from our mixed effects (random intercept) model. We can see that the states have some spread in terms of differing intercepts. We can also see that, similar to the interaction models we explored, there is no significant visual relationship between religious institutions (red dots) and non-religious (blue dots) when comparing `percent.black` and `full.cases`.

Next we explore the relationship between `gap20repub` and `religious` and their interactive effect on 2020 Covid cases grouped by state (still using mixed effects, random intercept). We are exploring this given the feature of the mixed effects models in exploring predictors at the second level of the heirarchical model (ie. the state-level). `gap20repub` is unique by state, but not by institutions in each state. We formally test below whether the effect of university religious affiliation (on 2020 Covid cases) depends on `gap20repub`:

```
## Data: full.cases
## Models:
## lmer3b: log(cases + 1) ~ gap20repub + religious + (1 | state)
## lmer3: log(cases + 1) ~ gap20repub * religious + (1 | state)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer3b    5 8706.2 8733.9 -4348.1   8696.2
## lmer3     6 8700.5 8733.6 -4344.2   8688.5 7.7511  1   0.005368 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
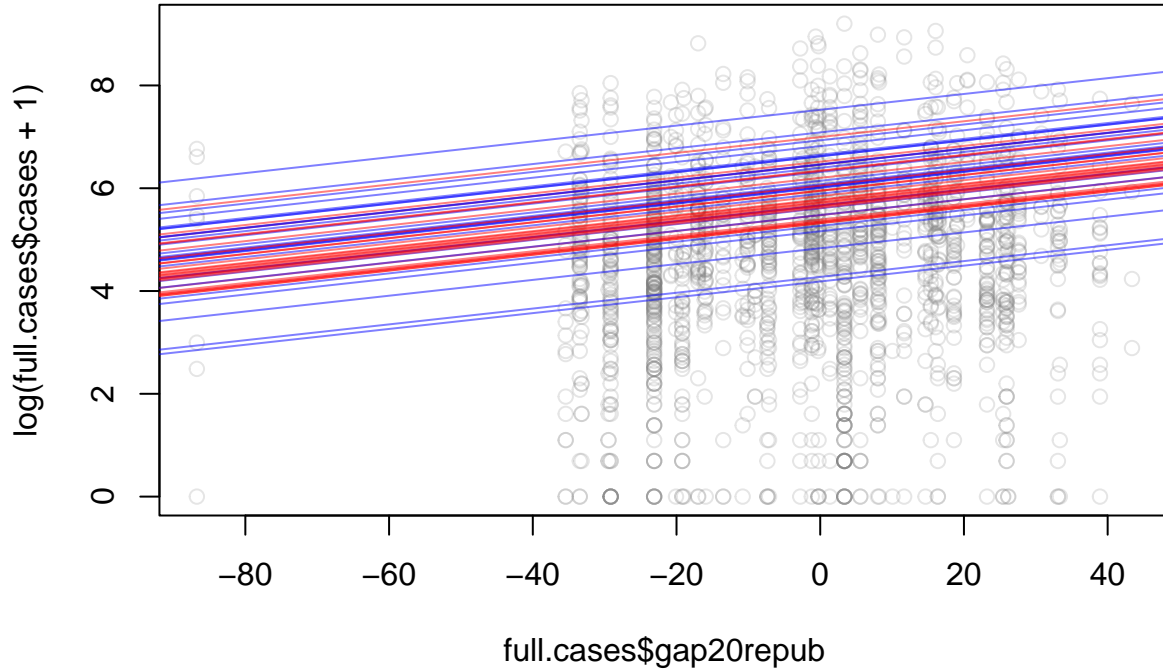
The result of the anova test between nested models mixed effects models (one with interactive effect, one without) indicates that there is evidence that the interaction between `gap20repub` and `religious` is significant in the predictive model, and the effect of religious affiliation upon 2020 Covid cases does depend on `gap2020repub`. Though this was not flagged as a particularly significant interaction in the above models, it is not surprising, per se: It is clear that the `religious` significantly interacts with many other predictors.

Sensitivity Analysis using Poisson Regression

In "Checking the Assumptions of Linear Regression Models", we found behavior in our data set that calls into question the assumption of homoskedasticity that is used in all of the above methods, which are based on Gaussian regression. Because our response is count data, Poisson regression is an obvious candidate for improving the violation of this assumption. The first thing we do to verify this fact is to fit a Poisson regression model using the same formula as our base model from above; as the residuals versus fitted values plot below shows, Poisson regression does, for the most part, fix the intractable violations of homoskedasticity that we encountered with Gaussian regression.

Residuals vs Fitted
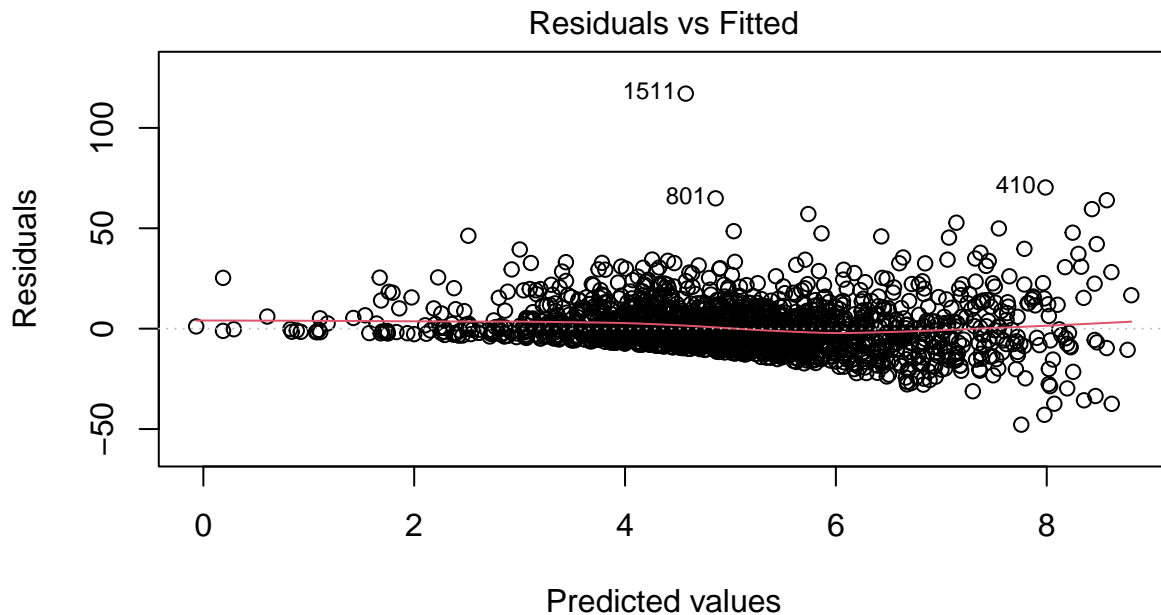
Predicted values
glm(cases ~ religious + tuition + log(total.headcount + epsilon) + log(perc ...

With this model, however, the problem is not completely solved: Because the spread in the responses seems to increase as fitted values increase, it seems that weighted regression would be the best approach in the Poisson setting. The most obvious candidate for weights is the predictor `total.headcount` — it is reasonable to think that as school size increases, the differences approaches to handling the outbreak of Covid-19 might be manifested in increasingly different case count numbers. The below plot shows that employing weighted-least-squares Poisson regression using `1/total.headcount` as weights for the most part fixes the problem of heteroskedasticity.

Residuals vs Fitted

Predicted values
glm(formula(p1))
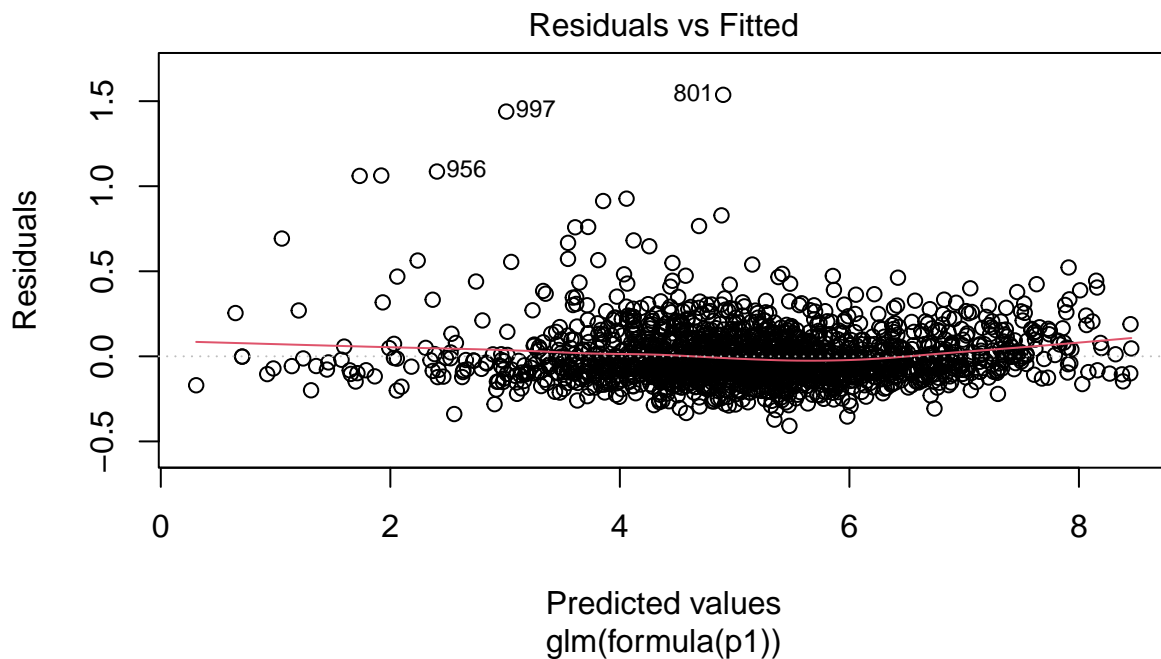
Clearly, there are still some outliers, which might be due to a number of factors, including exceptionally frequent testing, comprehensive Covid contact tracing systems, superspreader events on campus, or simply exceptionally large outbreaks on campus; however, the assumption of homoskedasticity now seems to be reasonable enough to trust our standard errors.

Table 5: Base Model: Gaussian vs. Poisson

coefficient	Gaussian.Estimate	Gaussian.p.value	Poisson.Estimate	Poisson.p.value
gap20repub	0.0127476	0.0000000	0.0083467	0.0000000
grad.rate	0.0210106	0.0000000	0.0170122	0.0000000
log(percent.asian + epsilon)	-0.1251276	0.0257048	-0.2202536	0.0000565
log(percent.black + epsilon)	0.1946336	0.0000088	0.1259964	0.0080049
log(percent.pacific.islander + epsilon)	-0.2890057	0.0146274	-0.0738136	0.6778458
log(total.headcount + epsilon)	0.9199545	0.0000000	0.9386192	0.0000000
occupational.degreeYes	-0.1840886	0.0173571	-0.1626333	0.0050808
on.campus.housingYes	0.8225306	0.0000000	0.7586564	0.0000000
percent.student.loan	0.0053149	0.0017920	0.0030923	0.0721520
percent.two.more.races	-0.0254703	0.0262692	-0.0287106	0.0491504
percent.white	0.0178880	0.0000000	0.0145168	0.0000000
percent.women	-0.0107628	0.0000245	-0.0104932	0.0001242
privateYes	-0.6765733	0.0000051	-0.2976105	0.0214963
tuition	0.0000174	0.0000161	0.0000081	0.0270510

Table 6: Full Interaction Model: Gaussian vs. Poisson (Only predictors related to 'religious')

coefficient	Gaussian.Estimate	Gaussian.p.value	Poisson.Estimate	Poisson.p.value
religiousYes:log(percent.asian + epsilon)	-0.3936018	0.0422889	-0.2979514	0.0457438
religiousYes:log(percent.black + epsilon)	-0.4388041	0.0288920	-0.1931332	0.2352499
religiousYes:mask.mandated.days	0.0017751	0.0350710	0.0004273	0.4971038
religiousYes:on.campus.housingYes	-1.4964379	0.0384122	-1.1822020	0.0390593

We now fit Poisson regression models using the same formulas as the most notable Gaussian models, and compare the lists of coefficient estimates related to religious affiliation marked as statistically significant by the two approaches. Note that when determinging the statistical significance of a coefficient estimate in the Poisson models, we use heteroskedastic-consistent standard errors to account for possible violations of distribution assumption that the mean of the response is equal to its variance. The tables below illustrates the results of these analyses.

In the base model, only one coefficient that was statistically significant in the Gaussian model was no longer statistically significant in the Poisson model: `percent.student.loan`. Otherwise, though the p-values may have changed, the predictors that were statistically significant in the Gaussian model were statistically significant in the Poisson model as well.

With respect to the `fullInteraction` model, of the linear or interaction effects related to the predictor `religious`, two of the four coefficient estimates remained statistically significant in the Poisson model — `religiousYes:log(percent.asian + epsilon)` and `religiousYes:on.campus.housingYes` — while the other two did not.

In the Poisson regression version of LASSO-Selected Model 2, only the coefficient estimate for `religiousYes` remained statistically significant; `religiousYes:mask.mandated.days` was no longer statistically significant after we assumed that the errors took on the Poisson distribution.

Table 7: LASSO-Selected Model 2: Gaussian vs. Poisson (Only predictors related to 'religious')

coefficient	Gaussian.Estimate	Gaussian.p.value	Poisson.Estimate	Poisson.p.value
religiousYes	3.1755840	0.0407812	2.6427005	0.0102538
religiousYes:mask.mandated.days	0.0016696	0.0474806	0.0001859	0.7572110

Table 8: Backwards Sequential Variable Selection Model: Gaussian vs. Poisson (Only predictors related to 'religious')

coefficient	Gaussian.Estimate	Gaussian.p.value	Poisson.Estimate	Poisson.p.value
religiousYes:log(percent.asian + epsilon)	-0.3802799	0.0035695	-0.1859713	0.0945386
religiousYes:mask.mandated.days	0.0020786	0.0001400	0.0008273	0.0473759
religiousYes:on.campus.housingYes	-1.3120944	0.0051891	-1.1023241	0.0048741

Of the three `religious` interaction effects that were determined to be statistically significant in the Gaussian `back.step` model, only two remained statistically significant in the Poisson regression recreation of this model: `religiousYes:mask.mandated.days` and `religiousYes:on.campus.housingYes`. `religiousYes:log(percent.asian + epsilon)` was not statistically significant in the Poisson model.

Overall, we have observed that most of the coefficients that were marked as statistically significant in the Gaussian models fit earlier in the paper remained statistically significant under the assumptions of Poisson regression. This corroborates our analyses and findings from earlier in the paper, and allows us to now draw the overarching conclusions of our analysis.

Conclusions

General points to discuss:

- Brief summary of first three sections: Group Testing, Assumptions, 2020/21/total.cases decision
- Linear model recap, determining the significance of `religious` relies a lot on its interactions
- AIC step comes to similar conclusions?

Discussing inferences: - Which coefficients related to `religious` (linear and interaction effects) did we find to be significant at each stage of our analyses? - Where the findings differ, which models do we choose to give more credit to? Which findings are we more confident in (based on number of models that support the significance of a predictor)? - Giving the interpretation of every predictor that is discussed?