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

Daniel de Castro and Laura Appleby

December 13, 2022

Introduction

Paragraph expressing our motivations for pursuing this project, a brief analysis plan, and our hypotheses.

Description of data and source

Our data for this project comes from three sources:

1. NYTimes Covid-19 Data. This is publicly available on GitHub and was the source for some of the NYTimes maps and data visuals during the 2020-2021 era of the pandemic. It includes cases from 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 publicly available data on colleges across the globe. It has many possible variables including demographics, admission rates, University affiliation, etc. The IPEDS data center allowed us to select certain Universities and variables. The smallest subset of Universities that included all from the NYTimes database (by IPEDS ID) was 6125 rows, with all US 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.

The data has $n = 1,855$ rows after removing Universities without stats or without matching IPEDS ids.

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`, `catholic`, 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 `md`. 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 `md` 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 ADJUST NUMBER! columns:

- **ipeds_id**
- **institution.name**
- **state**
- **private**
- **religious.affiliation**
- **religious**
- **catholic**
- **tuition**
- **total.headcount**
- **percent.american.native**
- **percent.asian**
- **percent.black**
- **percent.hispanic.latino**
- **percent.pacific.islander**
- **percent.white**
- **percent.two.more.races**
- **percent.NA.race**
- **percent.nonres.alien**
- **percent.women**
- **avg.grant.money**
- **grad.rate**
- **percent.fin.aid**
- **percent.student.loan**
- **occupational.degree**
- **hs.equivalent.degree**

- **NCAA.football**
- **percent.disability**
- **on.campus.housing**
- **dorm.capacity**
- **city**
- **college**
- **cases**
- **cases_2021**
- **religious**
- **catholic**
- **private**
- **mask.mandated.days**
- **total.cases**
- **gap20repub**

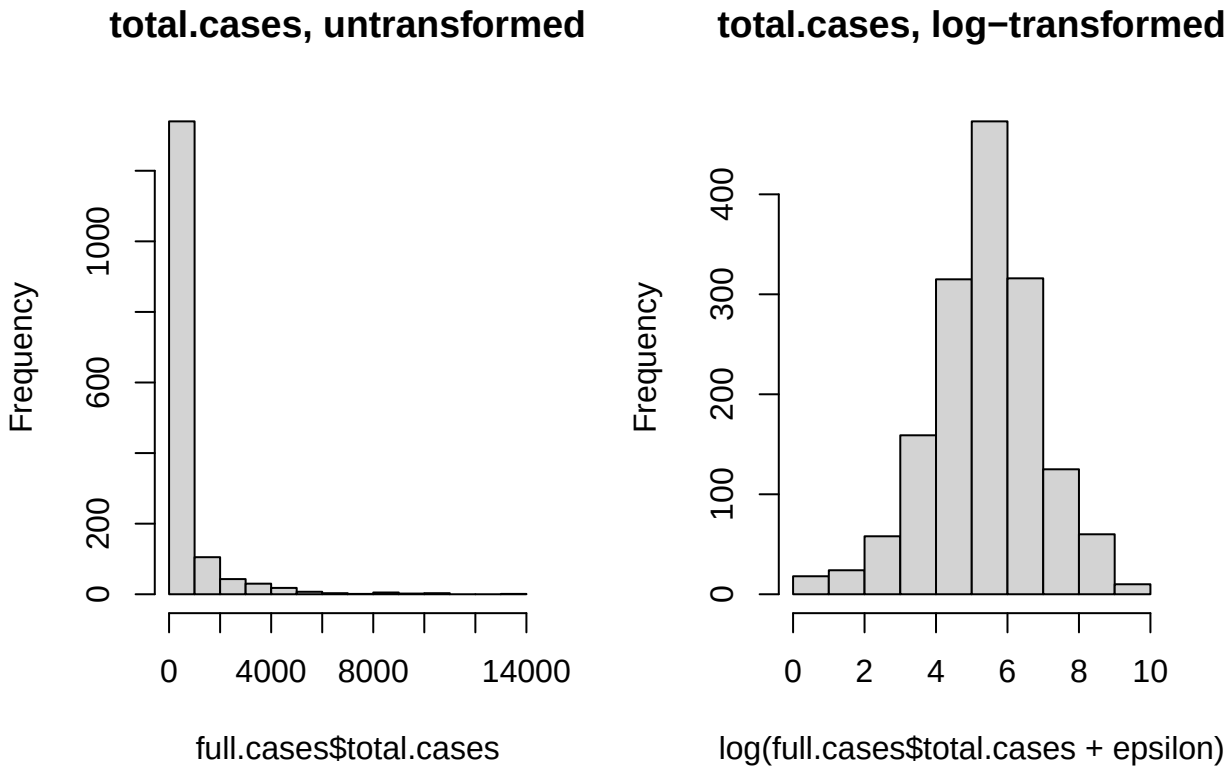
Group Testing

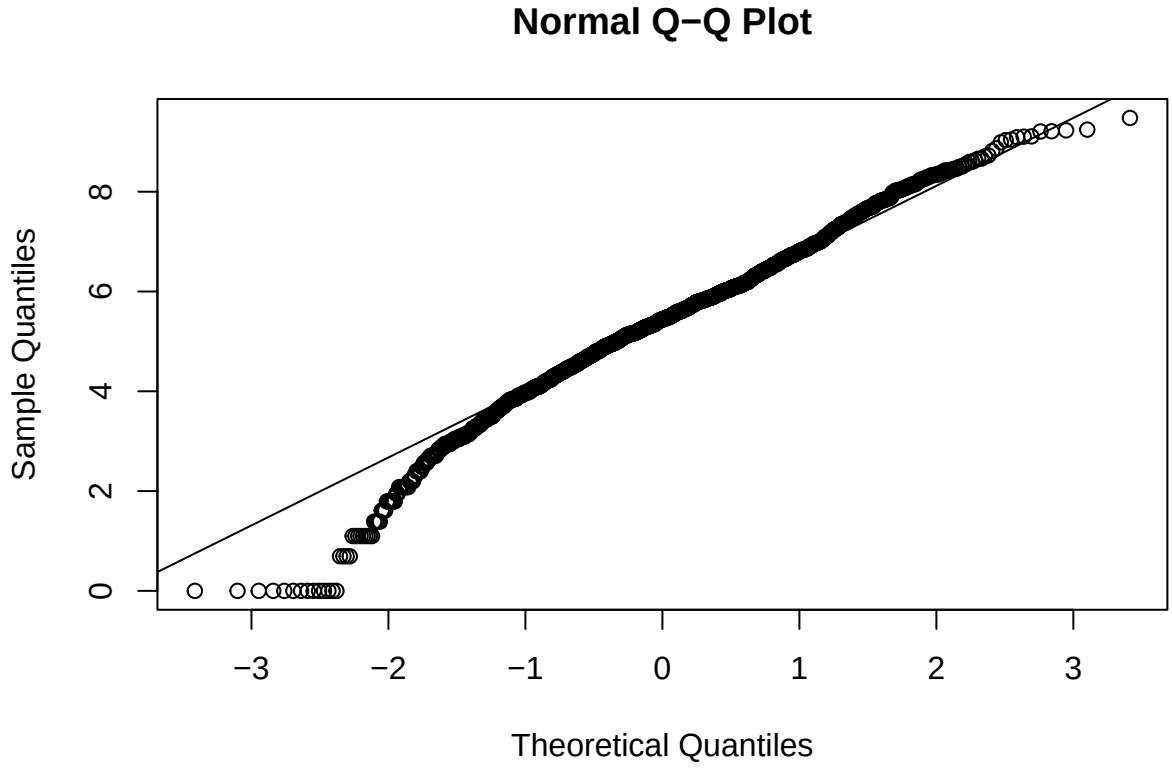
Explain motivation for group testing

Checking the assumptions for t-based methods

For unpooled t-based test for a difference in sample means, there are three assumptions:

1) **Observations are independent.** Explain why reasonable.
2) **Groups are independent of one another.** Explain why reasonable.
3) **Observations are normally distributed.**





Student-t Tests

State hypotheses, add note confirming the use of $\alpha = 0.05$ as our confidence level throughout the paper.

##

```
##  Welch Two Sample t-test
##
## data:  log(total.cases + epsilon) by religious
## t = -0.56432, df = 885.72, p-value = 0.5727
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.1961100  0.1085196
## sample estimates:
##  mean in group No mean in group Yes
##          5.351261          5.395056
```

Interpret statistical significance

ANOVA

```
##                           Df    Sum Sq Mean Sq F value Pr(>F)
## religious.affiliation   47 5.560e+07 1183010   0.844  0.766
## Residuals             1510 2.118e+09 1402478
```

Shows that just breaking observations into religious vs. not religious (or catholic vs. not catholic) is much more informative than considering the specific religious affiliation of each school.

Non-Parametric Testing — Wilcox Rank Sum Test

Motivations for non-parametric testing, and hypotheses

```
##
##  Wilcoxon rank sum test
##
## data:  full.cases$total.cases[full.cases$religious == "Yes"] and full.cases$total.cases[full.cases$r
## W = 225800, p-value = 0.7548
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -22.99995  29.00002
## sample estimates:
## difference in location
##               4.000088
```

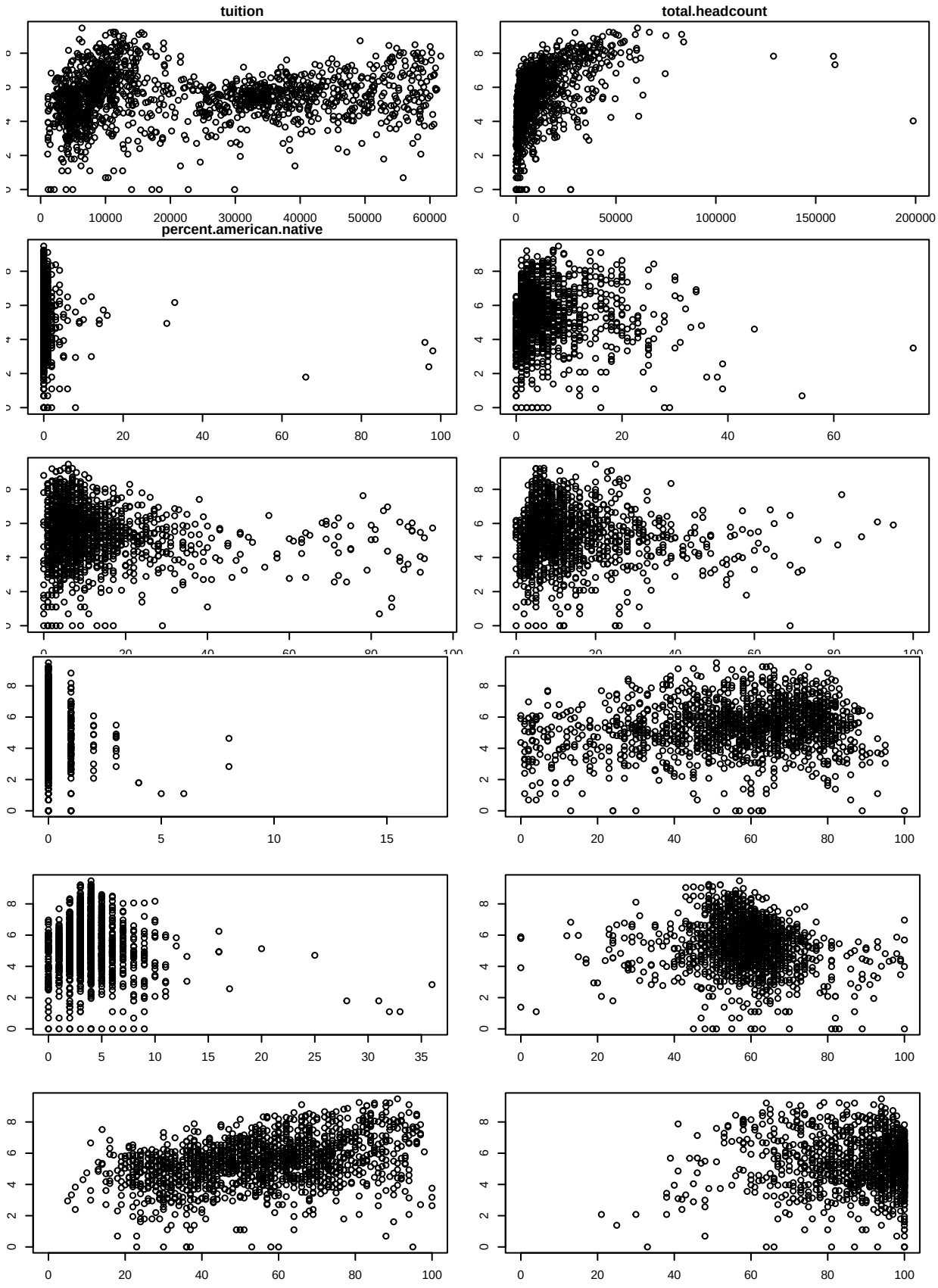
Interpret significance

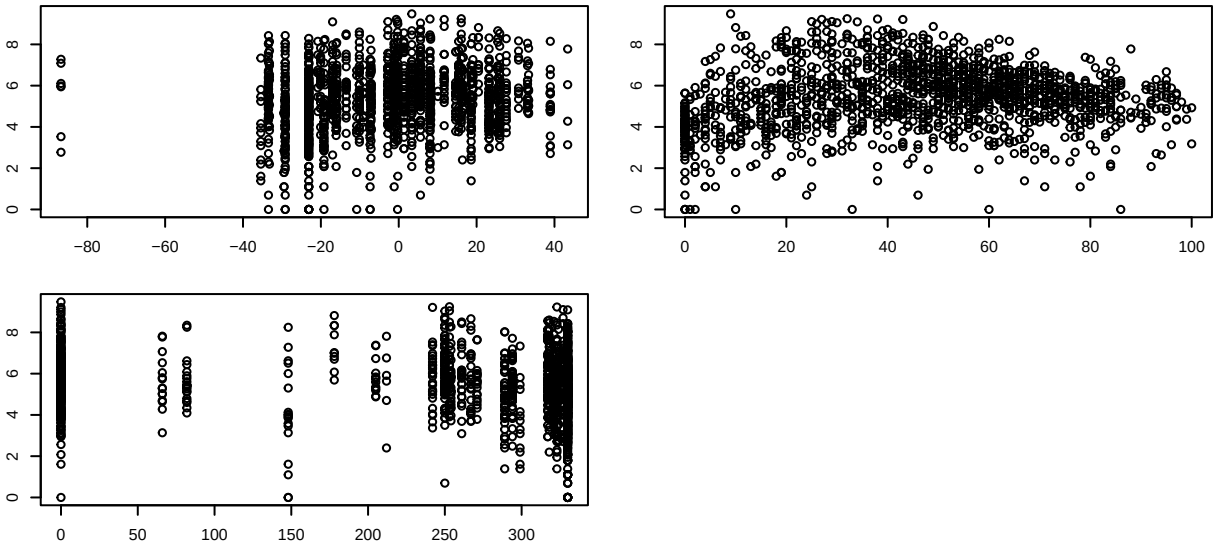
Basic Linear Regression Models

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.

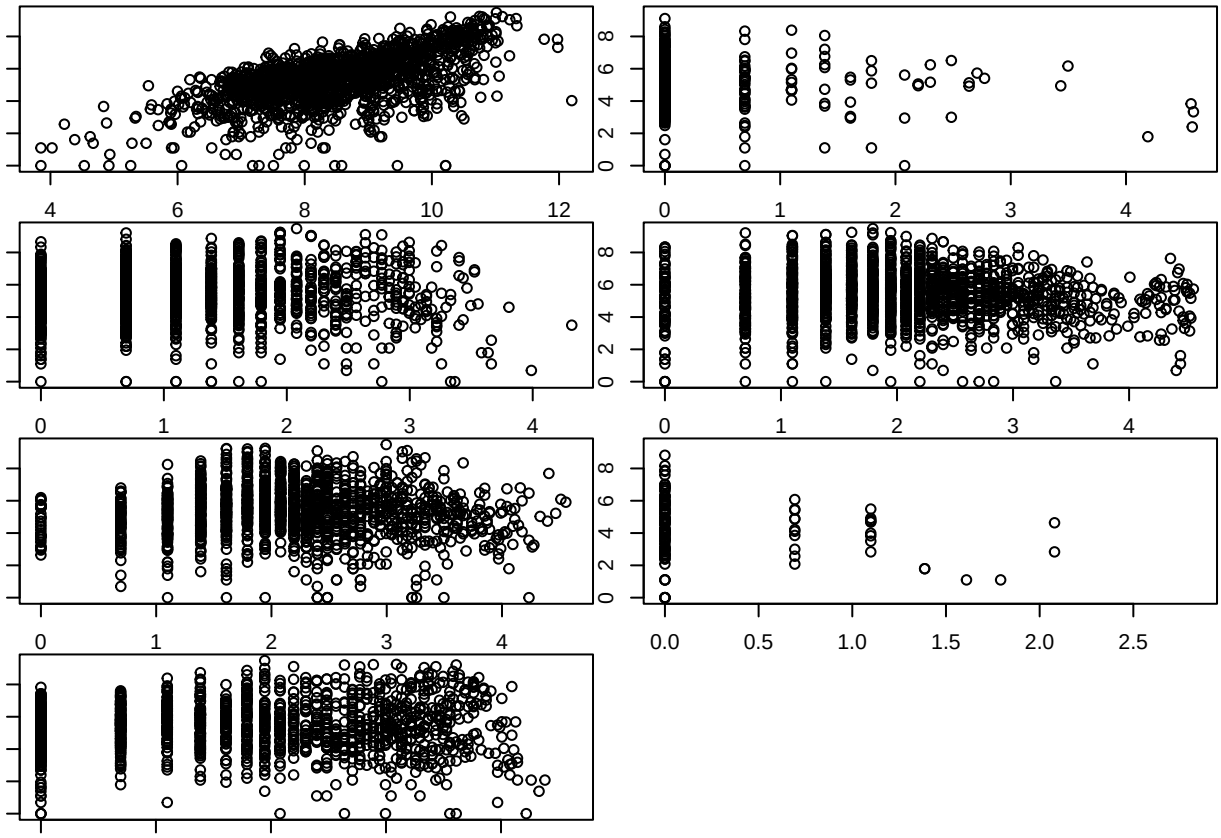
Checking the Assumptions of Linear Regression Models

Before we can fit more complicated models, we must first check the assumptions of linear regression. We begin by checking the assumption of linearity, plotting `cases` versus all of the quantitative predictors that we would like to include in our analyses:


tuition
total.headcount
percent.american.native

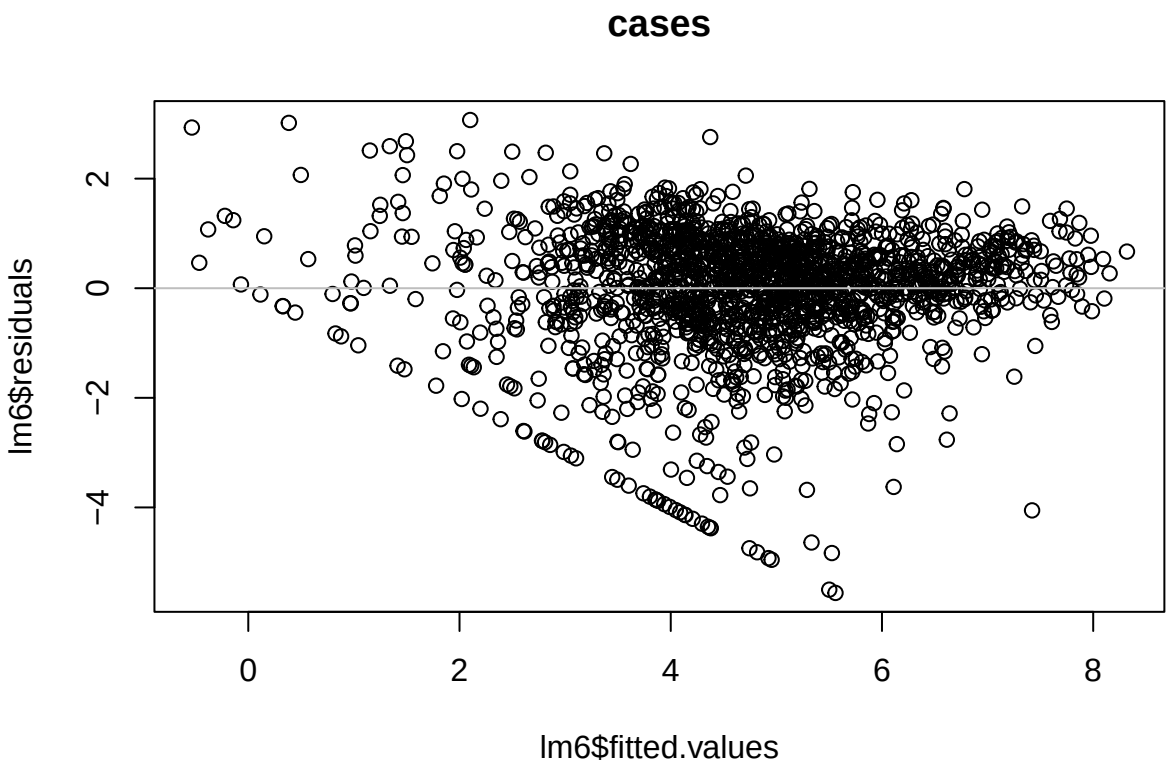


We identify that `total.headcount`, `percent.american.native`, `percent.asian`, `percent.black`, `percent.hispanic.latino`, and `percent.pacific.islander` would benefit from being log-transformed; given the left-skewness of its distribution, we also determined that `percent.fin.aid` would be best transformed using the following transformation `log(100 - percent.fin.aid + 1)`. The following plots show that the distribution of these predictors are much better after being transformed.



After these transformations, the linearity assumptions seems much more reasonable for each of the quantitative predictors.

In order to check the assumption of homoskedasticity, we fit basic regression models for `cases` and `total.cases` using all of the predictors in our predictor set.

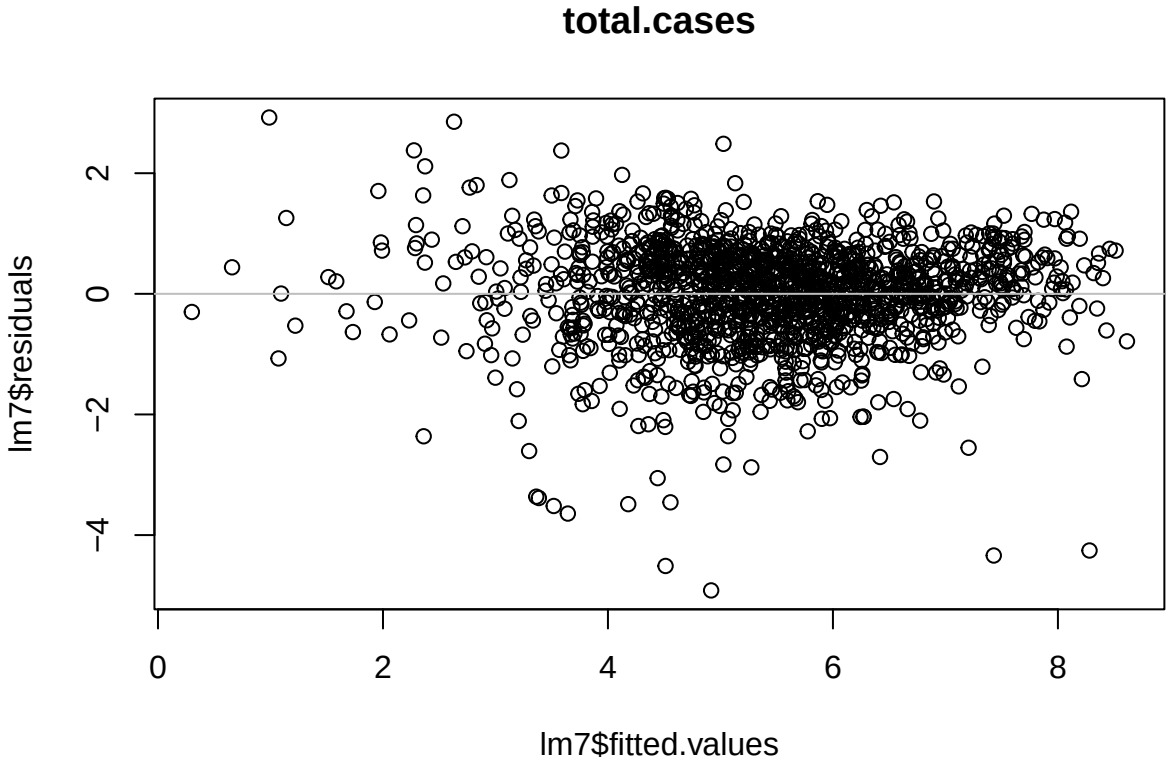


As is shown in the plot above, the spread of the residuals is not constant across the entire range of fitted values — thus, it is called into question whether the assumption of homoskedasticity is reasonable in this case. Given that the vast majority of residuals are clustered around a region with consistent spread, we believe that the assumption of homoskedasticity will be reasonable enough to accept in this case. As a check, we will compare the standard errors generated by heteroskedasticity-consistent method with those generated by the standard OLS approach:

```
##                                       ols robust   diff
## (Intercept)                        0.4099 0.4703 0.0604
## religiousYes                       0.0892 0.0870 0.0022
## tuition                            0.0000 0.0000 0.0000
## log(total.headcount + epsilon)     0.0316 0.0369 0.0052
## log(percent.american.native + epsilon)  0.0556 0.0681 0.0125
## log(percent.asian + epsilon)       0.0560 0.0649 0.0089
## log(percent.black + epsilon)       0.0437 0.0492 0.0055
## log(percent.hispanic.latino + epsilon)  0.0435 0.0504 0.0068
## log(percent.pacific.islander + epsilon) 0.1183 0.1363 0.0180
## percent.white                      0.0018 0.0022 0.0004
## percent.two.more.races             0.0115 0.0141 0.0027
## percent.women                      0.0025 0.0028 0.0003
## grad.rate                          0.0023 0.0026 0.0003
## log(100 - percent.fin.aid + epsilon)    0.0308 0.0325 0.0018
## on.campus.housingYes               0.0939 0.1122 0.0183
## gap20repub                         0.0019 0.0019 0.0000
## privateYes                         0.1479 0.1907 0.0429
## percent.student.loan               0.0017 0.0019 0.0002
## mask.mandated.days                 0.0003 0.0003 0.0000
## occupational.degreeYes             0.0773 0.0809 0.0036
## hs.equivalent.degreeYes            0.0954 0.0980 0.0026
```

As the table above shows, the standard errors are the most part consistent between the two models. ADDRESS

ASK TF!



In the above plot for the model for `total.cases`, the spread of the residuals is much more consistent across the entire range of fitted values. Though it is not entirely uniform, it would appear reasonable enough to operate under the assumption of homoskedasticity, especially given that, as was demonstrated in Problem Set 3, t-based statistical inference procedures done using linear models are fairly robust to slight violations in the assumption of homoskedasticity.

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

Naturally, there is a temptation to forget about the case data for only 2020 and 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 1855 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 1855 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 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

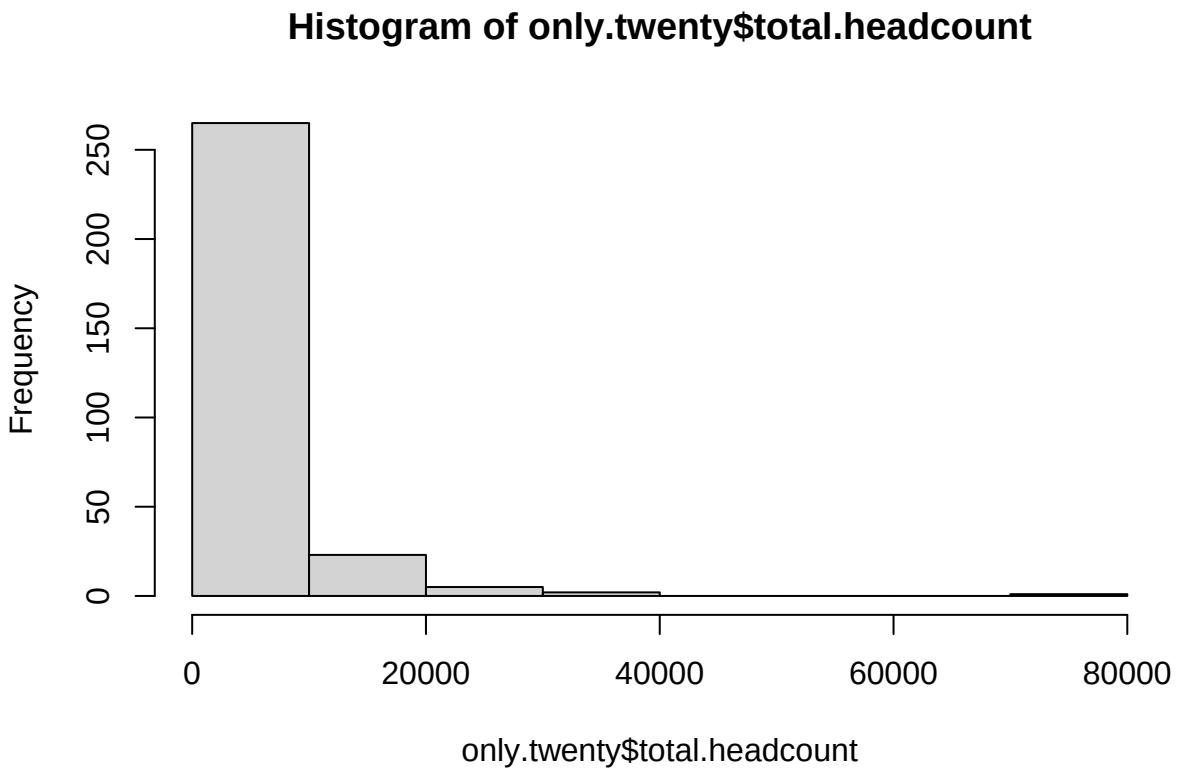
	all schools		schools with data for both years		schools with only 2020 data	
	Est.	p	Est.	p	Est.	p
(Intercept)	−5.719	<0.001	−4.975	<0.001	−2.338	0.109
religiousYes	0.102	0.252	0.318	<0.001	−0.701	0.049
tuition	0.000 02	<0.001	0.000 01	<0.001	−0.000 003	0.878
log(total.headcount + epsilon)	0.920	<0.001	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	
AIC						
BIC						
Log.Lik.	−2590.545		−1808.194		−424.879	
RMSE	1.11		0.83		1.46	

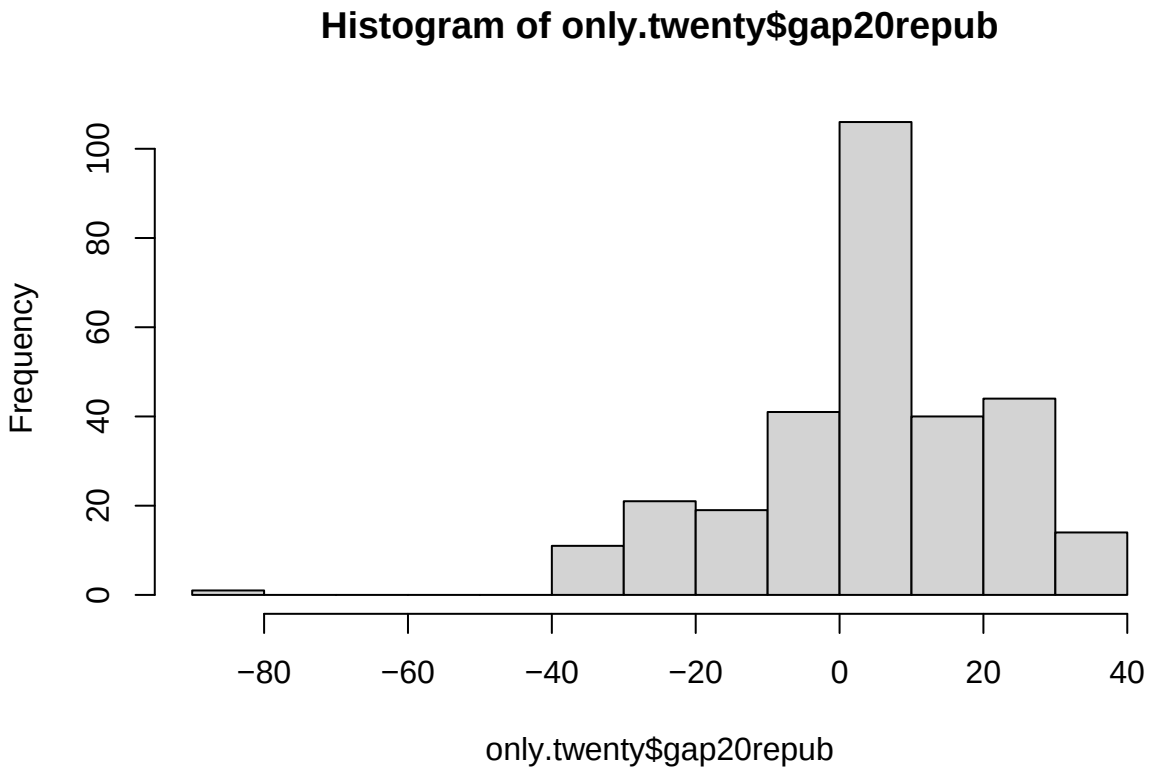
of predictors included in this baseline models experienced changes in coefficient estimates and statistical significance as the sample of institutions considered was changed.

Let's find out more about the schools that only reported data for 2020:

```
##
##  No Yes
## 193 104
```

```
##
##  No Yes
## 163 134
```





t-test for difference in political leanings

```
##
##  Welch Two Sample t-test
##
## data:  both$gap20repub and only.twenty$gap20repub
## t = -6.2892, df = 445.72, p-value = 7.621e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.468120 -4.959647
## sample estimates:
## mean of x mean of y
## -3.264288  3.949596
```

t-test for difference in size

```
##
##  Welch Two Sample t-test
##
## data:  both$total.headcount and only.twenty$total.headcount
## t = 10.488, df = 872.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4533.923 6621.486
## sample estimates:
## mean of x mean of y
##  9739.015  4161.311
```

z-test for proportion private

```
##
##  2-sample test for equality of proportions with continuity correction
```

```
##
## data:  matrix(c(sum(both$private == "Yes"), sum(both$private == "No"), sum(only.twenty$private == "Y
## X-squared = 0.19867, df = 1, p-value = 0.6558
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.0797248  0.0477145
## sample estimates:
##    prop 1    prop 2
## 0.4351733 0.4511785
```

z-test for `religious`

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  matrix(c(sum(both$religious == "Yes"), sum(both$religious == "No"), sum(only.twenty$religious
## X-squared = 14.254, df = 1, p-value = 0.0001597
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.16719658 -0.04661894
## sample estimates:
##    prop 1    prop 2
## 0.2432606 0.3501684
```

z-test for `on.campus.housing`

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  matrix(c(sum(both$on.campus.housing == "Yes"), sum(both$on.campus.housing == "No"), sum(only.
## X-squared = 26.561, df = 1, p-value = 2.553e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.07907841 0.19829091
## sample estimates:
##    prop 1    prop 2
## 0.7952503 0.6565657
```

THE TWO SAMPLES OF SCHOOLS ARE CLEARLY DIFFERNET WITH RESPECT TO CERTAIN PREDICTORS — INTERACTION EFFECTS MIGHT BE AT PLAY. WE SHOULD RESTRICT SAMPLE TO CASES IN 2020, AND EXPLORE INTERACTION EFFECTS

Linear Models with Interaction Effects

In Search of a Parsimonious Model: Sequential Variable Selection Models

Honestly not sure how to interpret and forward step model and the both-directions step model take forever to run.

LASSO for Variable Section

Hierarchical Multi-level Models

Conclusions

Laura's work from before