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**

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

Group Testing

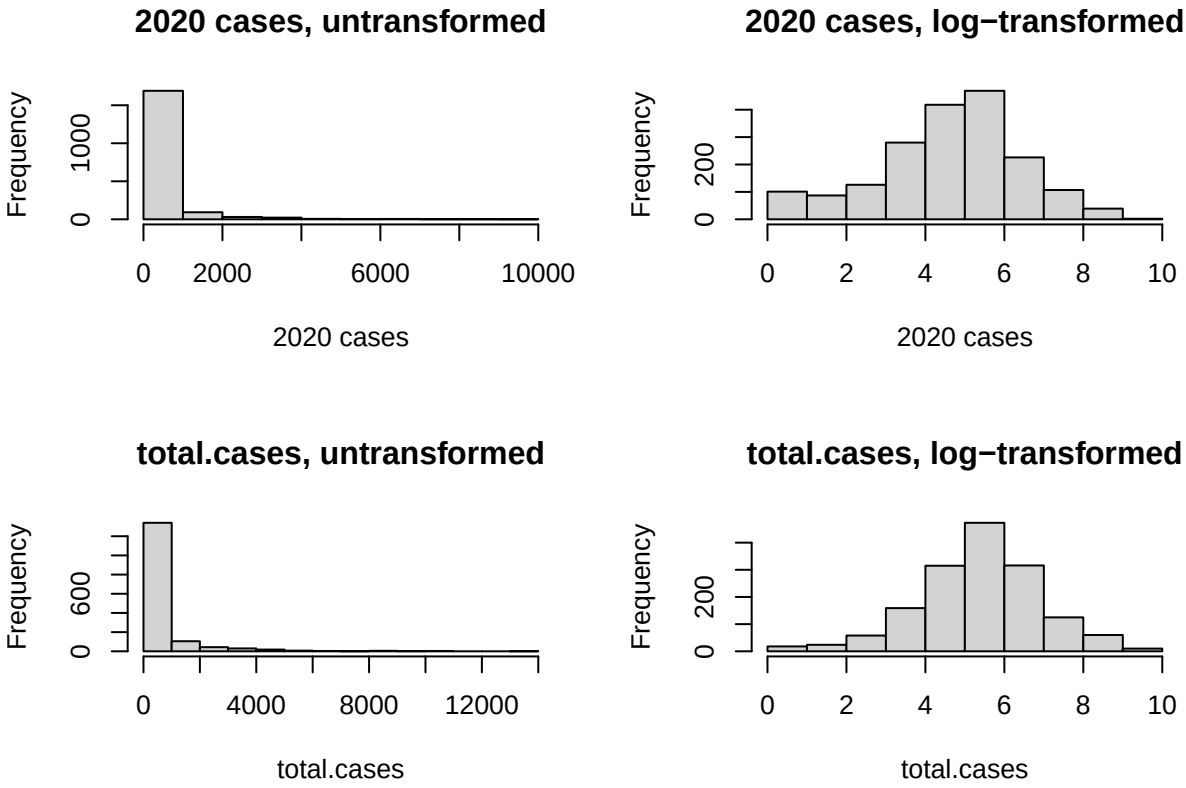
To begin our analyses, we will perform a range of group tests to determine whether the institutions with a religious affiliation had a higher true average number of Covid cases than religious institutions without a religious affiliation. In the following analyses, we will test whether or not the true 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 bfit 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

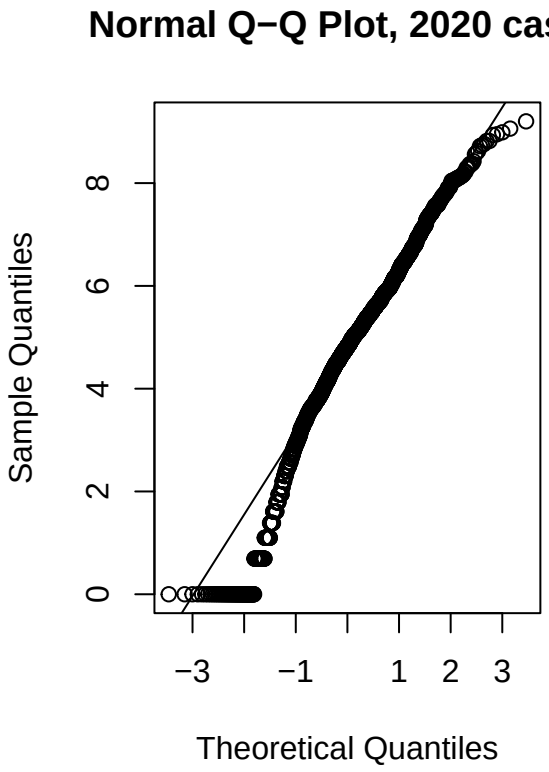
We begin with the most commonly used test for a difference in means: the Student-t test. Of course, before we can perform this test, we must make sure that its underlying assumptions are reasonable. Because we have no reason to assume that the variances in the observations between both groups — religiously 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:

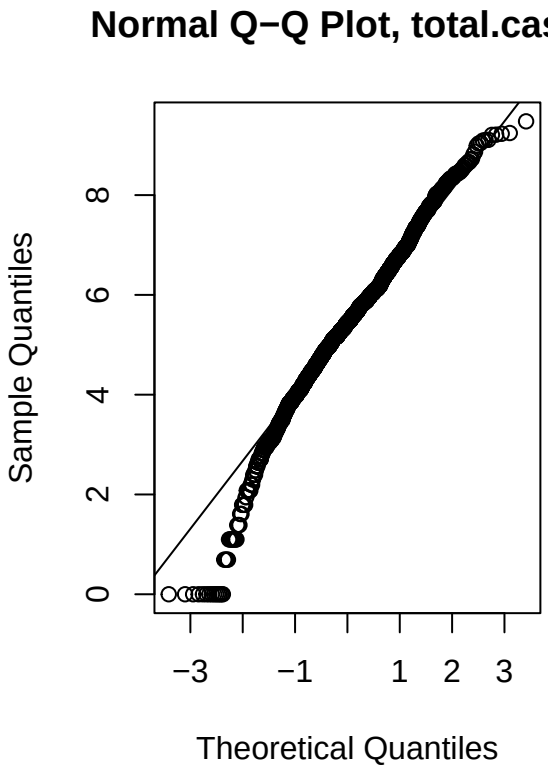
1) **Observations are independent.** We claim that this observation is reasonable. Even though the geographic proximity of universities — or just a relationship between universities in which students from one frequently visit the other — might have caused there to be some correlation between case counts 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 1,855 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, there is no reason to suggest that these two groups of religious institutions would not be 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 distirbution is 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 test itself. 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. 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 a assumption on the distribution of the data-generating process of the sample follows. 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

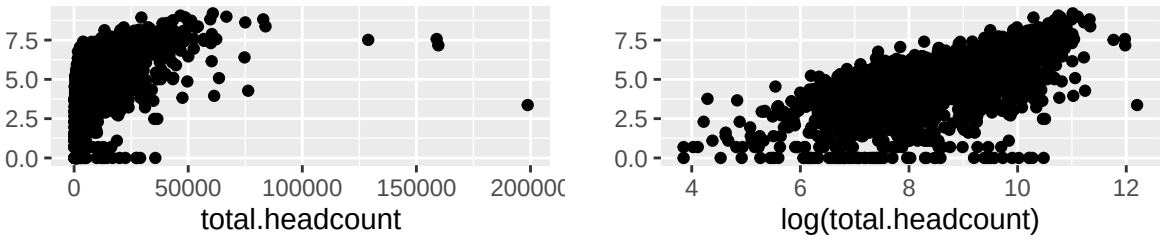
As the results above show, neither test was statistically significant at the 0.05 confidence level; we thus again fail to reject the null hypothesis, concluding that there is no statistically significant evidence to suggest that the there is an association between religious affiliation and average quantile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

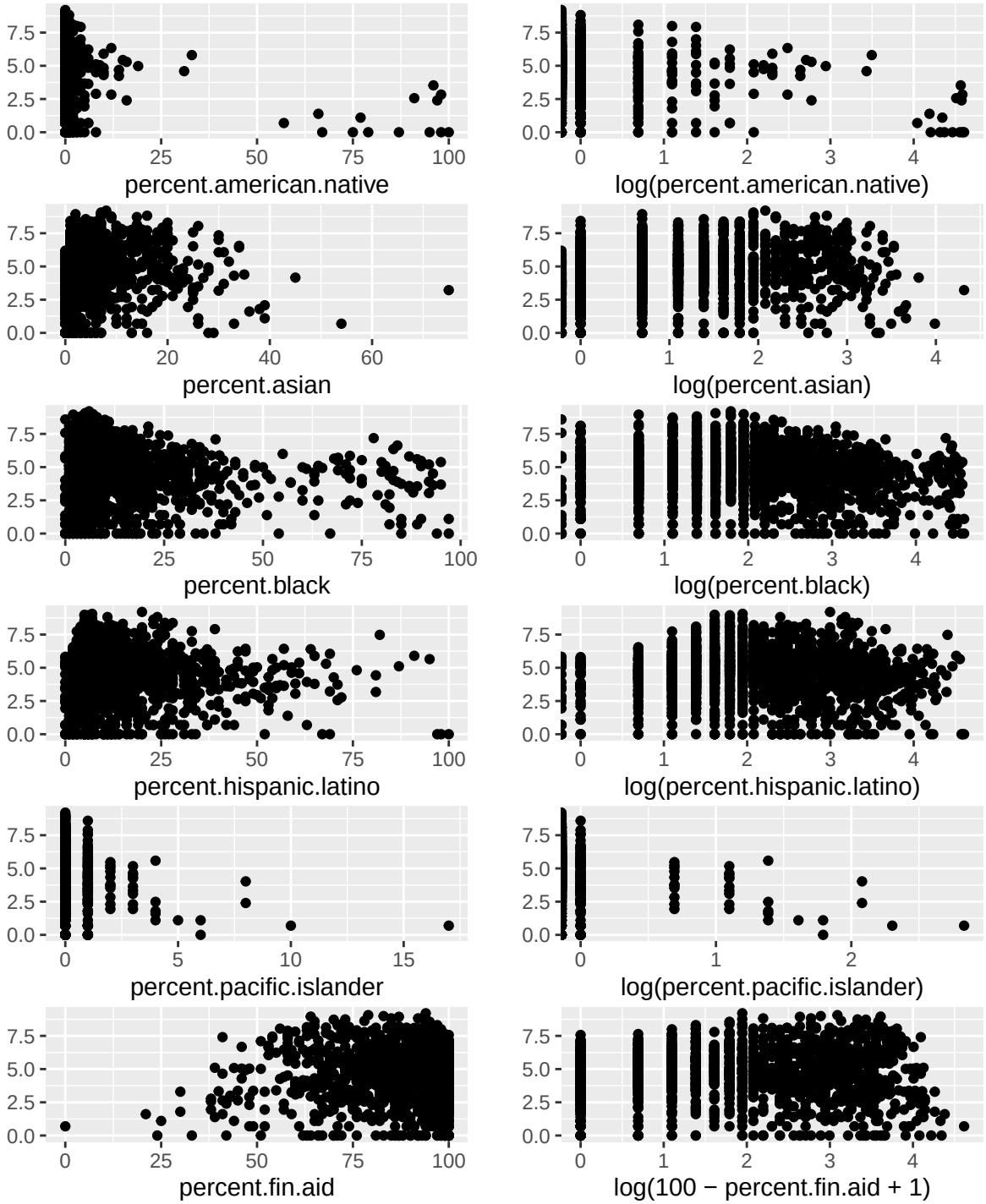
Basic Linear Regression Models

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

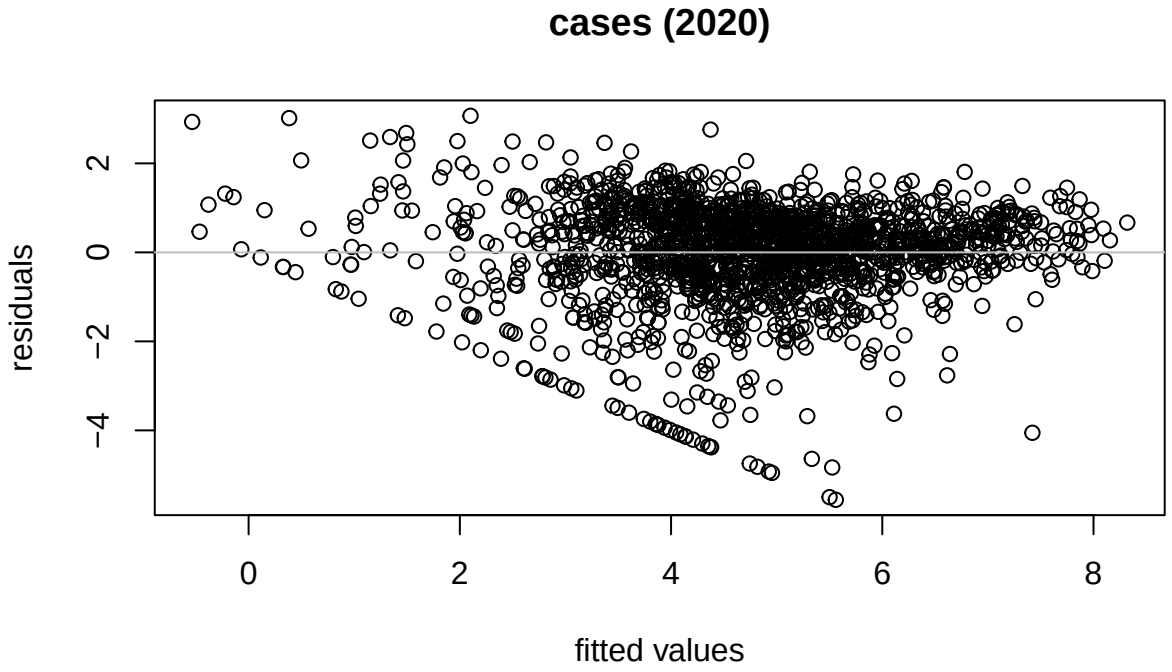
Before we can fit more complicated models, we must first check the assumptions of linear regression. We begin by checking the assumption of linearity. To do so, we plotted `cases` versus all of the quantitative predictors that we would like to include in our analyses (the complete list of these plots can be found in Appendix A). 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.

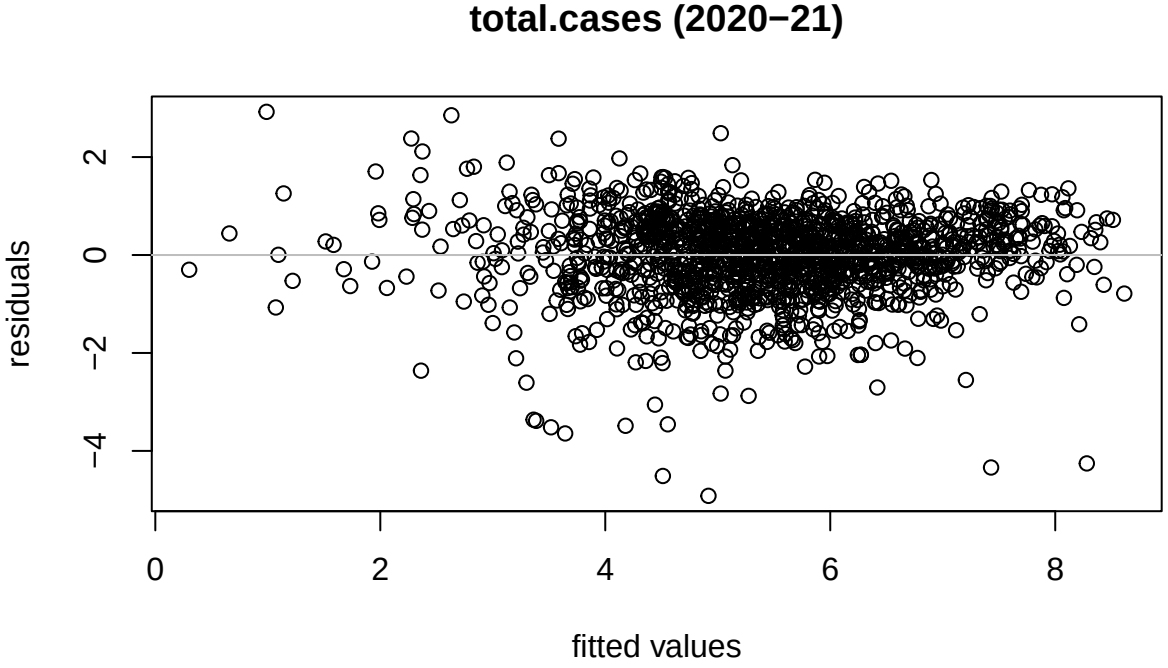


As is shown in the plot above, the spread of the residuals is not constant across the entire range of fitted values — thus, it is called into question whether the assumption of homoskedasticity is reasonable in this case. The behavior of this graph is interesting: The lower bound for the residuals appears to follow a negatively sloped line, and one that contains a solid number of points. ADDRESS ASK TF!

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

	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	

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 represents the change in the number of `cases` that we would expect given that a school were to be religiously affiliated, with all other predictors held constant. The coefficient estimate for `religiousYes` is quite different in the two models fit to data from only one of the two subgroups, which 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

Comments: - Assumed use of total.cases - can change - should we do a polynomial model? not possible with religous but something else? - What do you think of the explanations? Is there too much here?

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 now look at interaction effects and their significance in linear regression models for this dataset. First, we will look at a model with fewer predictors; including only religious, total.headcount, grad.rate, percent.fin.aid, dorm.capacity, and gap20repub to predict 2020 cases. We chose these variables because we anticipate there to not be a high degree of multicollinearity between them, and because they were statistically significant in predicting 2020 cases in a simple linear regression above. See lm8 for the model with no interaction effects, lm9 for interaction effects with the religious variables, and lm10 for the all interaction effects between the variables in lm8:

```
##
## Call:
## lm(formula = log(cases + epsilon) ~ religious + log(total.headcount +
##     epsilon) + +percent.fin.aid + log(dorm.capacity + epsilon) +
##     gap20repub, data = full.cases)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9882 -0.5520  0.2316  0.7838  2.7815
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -3.823349   0.329768 -11.594  < 2e-16 ***
## religiousYes                  -0.004540   0.073225  -0.062 0.950572
## log(total.headcount + epsilon)  0.759048   0.026616  28.519  < 2e-16 ***
## percent.fin.aid                 0.008806   0.002432   3.621 0.000302 ***
## log(dorm.capacity + epsilon)    0.240114   0.010579  22.697  < 2e-16 ***
## gap20repub                      0.015364   0.001517  10.126  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.21 on 1774 degrees of freedom
##   (75 observations deleted due to missingness)
## Multiple R-squared:  0.5451, Adjusted R-squared:  0.5438
## F-statistic: 425.1 on 5 and 1774 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = log(cases + epsilon) ~ (log(total.headcount + epsilon) +
##     +percent.fin.aid + log(dorm.capacity + epsilon) + gap20repub) *
##     religious, data = full.cases)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9220 -0.5596  0.1982  0.7829  2.8863
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             -3.845958   0.364250 -10.559
## log(total.headcount + epsilon)           0.760943   0.029062  26.183
## percent.fin.aid                          0.008824   0.002661   3.316
## log(dorm.capacity + epsilon)             0.243142   0.010744  22.630
## gap20repub                               0.019504   0.001749  11.150
```

```
## religiousYes                                     1.197673   0.850026   1.409
## log(total.headcount + epsilon):religiousYes -0.062032   0.079376  -0.781
## percent.fin.aid:religiousYes                -0.007917   0.006619  -1.196
## log(dorm.capacity + epsilon):religiousYes    0.005719   0.062374   0.092
## gap20repub:religiousYes                     -0.017325   0.003512  -4.933
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## log(total.headcount + epsilon)               < 2e-16 ***
## percent.fin.aid                             0.000931 ***
## log(dorm.capacity + epsilon)                 < 2e-16 ***
## gap20repub                                   < 2e-16 ***
## religiousYes                                0.159015
## log(total.headcount + epsilon):religiousYes 0.434616
## percent.fin.aid:religiousYes                0.231786
## log(dorm.capacity + epsilon):religiousYes   0.926961
## gap20repub:religiousYes                     8.84e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.203 on 1770 degrees of freedom
##   (75 observations deleted due to missingness)
## Multiple R-squared:  0.5519, Adjusted R-squared:  0.5496
## F-statistic: 242.2 on 9 and 1770 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = log(cases + epsilon) ~ (religious + log(total.headcount +
##     epsilon) + +percent.fin.aid + log(dorm.capacity + epsilon) +
##     gap20repub)^2, data = full.cases)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7449 -0.5329  0.2165  0.7580  2.7159
##
## Coefficients:
##                                                              Estimate
## (Intercept)                                                -6.271e+00
## religiousYes                                                1.833e+00
## log(total.headcount + epsilon)                              1.027e+00
## percent.fin.aid                                             5.050e-02
## log(dorm.capacity + epsilon)                               -3.097e-02
## gap20repub                                                 -1.135e-02
## religiousYes:log(total.headcount + epsilon)                -1.432e-01
## religiousYes:percent.fin.aid                               -1.023e-02
## religiousYes:log(dorm.capacity + epsilon)                   5.839e-02
## religiousYes:gap20repub                                    -1.006e-02
## log(total.headcount + epsilon):percent.fin.aid             -4.644e-03
## log(total.headcount + epsilon):log(dorm.capacity + epsilon)  3.535e-02
## log(total.headcount + epsilon):gap20repub                   5.328e-03
## percent.fin.aid:log(dorm.capacity + epsilon)               -5.064e-04
## percent.fin.aid:gap20repub                                 -4.846e-05
## log(dorm.capacity + epsilon):gap20repub                    -2.066e-03
##                                                            Std. Error t value
## (Intercept)                                                 1.265e+00  -4.959
```

```
## religiousYes                                                 8.667e-01   2.115
## log(total.headcount + epsilon)                               1.544e-01   6.651
## percent.fin.aid                                              1.436e-02   3.516
## log(dorm.capacity + epsilon)                                 1.033e-01  -0.300
## gap20repub                                                   1.933e-02  -0.587
## religiousYes:log(total.headcount + epsilon)                  8.133e-02  -1.761
## religiousYes:percent.fin.aid                                 6.920e-03  -1.478
## religiousYes:log(dorm.capacity + epsilon)                    6.323e-02   0.923
## religiousYes:gap20repub                                      3.682e-03  -2.733
## log(total.headcount + epsilon):percent.fin.aid               1.787e-03  -2.598
## log(total.headcount + epsilon):log(dorm.capacity + epsilon)  7.594e-03   4.655
## log(total.headcount + epsilon):gap20repub                    1.425e-03   3.740
## percent.fin.aid:log(dorm.capacity + epsilon)                 7.576e-04  -0.668
## percent.fin.aid:gap20repub                                   1.382e-04  -0.351
## log(dorm.capacity + epsilon):gap20repub                      5.634e-04  -3.667
##                                                             Pr(>|t|)
## (Intercept)                                                 7.77e-07 ***
## religiousYes                                                0.034605 *
## log(total.headcount + epsilon)                              3.87e-11 ***
## percent.fin.aid                                             0.000449 ***
## log(dorm.capacity + epsilon)                                0.764365
## gap20repub                                                  0.557176
## religiousYes:log(total.headcount + epsilon)                 0.078370 .
## religiousYes:percent.fin.aid                                0.139661
## religiousYes:log(dorm.capacity + epsilon)                   0.355896
## religiousYes:gap20repub                                     0.006344 **
## log(total.headcount + epsilon):percent.fin.aid              0.009447 **
## log(total.headcount + epsilon):log(dorm.capacity + epsilon) 3.49e-06 ***
## log(total.headcount + epsilon):gap20repub                   0.000190 ***
## percent.fin.aid:log(dorm.capacity + epsilon)                0.504015
## percent.fin.aid:gap20repub                                  0.725919
## log(dorm.capacity + epsilon):gap20repub                     0.000252 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.186 on 1764 degrees of freedom
##   (75 observations deleted due to missingness)
## Multiple R-squared:  0.5657, Adjusted R-squared:  0.562
## F-statistic: 153.2 on 15 and 1764 DF,  p-value: < 2.2e-16
```

We see from the model without interaction effects that all chosen variables are statistically significant in predicting total cases in this specific model. Notably, the coefficient on `religiousYes` is positive and statistically significant, indicating that a school being religious predicts a higher number of cases than a non-religious institution. The interaction model with the variables from the previous models and their interactions with the religious variable has a slightly higher r-squared value than the model without interactions; 0.551881, versus 0.5450971.

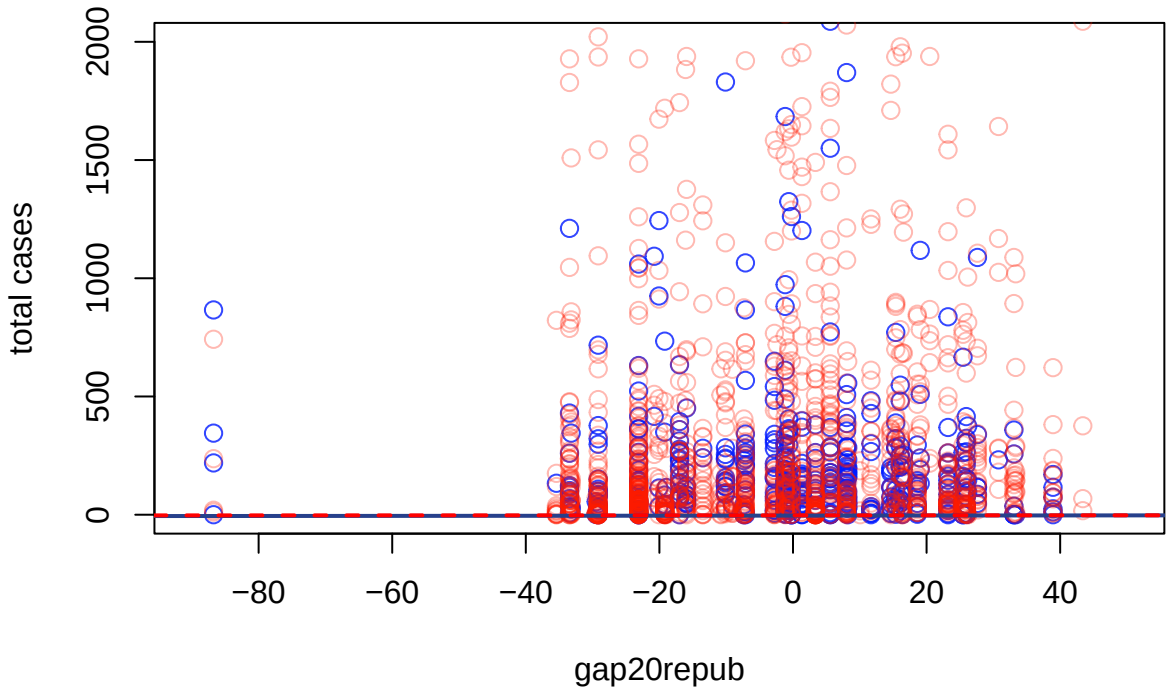
There was only one interaction effect that was statistically significant, that between `gap20repub` and `religious`. Even though `gap20repub` has a statistically significant positive effect when predicting cases across models when keeping `religious` constant (as one might expect; institutions in more republican states likely had fewer COVID-19 precautions, resulting in disproportionately higher cases), the relationship between `gap20repub` and `total.cases` is not quite as strong for institutions that are religious (though the coefficient is close to 0). This is statistically significant, with the negative interaction effect coefficient having a p-value of 0.00155, which is significantly below the level $\alpha = 0.05$.

Only one statistically significant interaction effect explains why the R^2 value for the `religious` interaction model was not a lot greater than that for the no interaction model. Conversely, the R^2 for the full interaction effect model (with these few selected variables) is a significant amount greater, with value 0.5657176.

```
## Analysis of Variance Table
##
## Model 1: log(cases + epsilon) ~ religious + log(total.headcount + epsilon) +
##     +percent.fin.aid + log(dorm.capacity + epsilon) + gap20repub
## Model 2: log(cases + epsilon) ~ (log(total.headcount + epsilon) + +percent.fin.aid +
##     log(dorm.capacity + epsilon) + gap20repub) * religious
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1774 2598.4
## 2   1770 2559.6  4    38.749 6.6988 2.384e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: log(cases + epsilon) ~ religious + log(total.headcount + epsilon) +
##     +percent.fin.aid + log(dorm.capacity + epsilon) + gap20repub
## Model 2: log(cases + epsilon) ~ (religious + log(total.headcount + epsilon) +
##     +percent.fin.aid + log(dorm.capacity + epsilon) + gap20repub)^2
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   1774 2598.4
## 2   1764 2480.6 10    117.78 8.3758 2.02e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ESS F-test, we see that there is evidence that the `religious` interaction terms contribute to the model. More significantly, there is evidence that the other interaction terms in the third model contribute even more significantly. This indicates that though `religious` has some interaction with the confounding variables, these predictors interact with one another more significantly.



The above plot confirms that adding the interaction between `religious` and the other selected variables,

though statistically significant for the variable `gap20repub`, does not change the model significantly. This is clear because the two predictive lines are nearly identical. They are low compared visually to the data due to the high number of 0 cases.

We will next look at a full interaction model with all variables as used in the base linear regression model, as well as their interaction with only religious. We will simply compare R^2, p-values, and RMSE to determine the importance of interaction in the models.

```
## [1] 0.6856778
```

```
## [1] 0.6208401
```

```
## [1] 0.7215086
```

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

```
## [1] 898.1031
```

```
## [1] 898.0947
```

```
## Analysis of Variance Table
##
## Model 1: log(cases + epsilon) ~ religious + tuition + log(total.headcount +
##     epsilon) + log(percent.american.native + epsilon) + log(percent.asian +
##     epsilon) + log(percent.black + epsilon) + log(percent.hispanic.latino +
##     epsilon) + log(percent.pacific.islander + epsilon) + percent.white +
##     percent.two.more.races + percent.women + grad.rate + log(100 -
##     percent.fin.aid + epsilon) + on.campus.housing + gap20repub +
##     private + percent.student.loan + mask.mandated.days + occupational.degree +
##     hs.equivalent.degree
## Model 2: log(cases + epsilon) ~ religious * (tuition + log(total.headcount +
##     epsilon) + log(percent.american.native + epsilon) + log(percent.asian +
##     epsilon) + log(percent.black + epsilon) + log(percent.hispanic.latino +
##     epsilon) + log(percent.pacific.islander + epsilon) + percent.white +
##     percent.two.more.races + percent.women + grad.rate + log(100 -
##     percent.fin.aid + epsilon) + on.campus.housing + gap20repub +
##     private + percent.student.loan + mask.mandated.days + occupational.degree +
##     hs.equivalent.degree)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1684 2084.3
## 2   1666 1993.4 18    90.918 4.2214 7.247e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: log(cases + epsilon) ~ religious * (tuition + log(total.headcount +
##     epsilon) + log(percent.american.native + epsilon) + log(percent.asian +
##     epsilon) + log(percent.black + epsilon) + log(percent.hispanic.latino +
##     epsilon) + log(percent.pacific.islander + epsilon) + percent.white +
##     percent.two.more.races + percent.women + grad.rate + log(100 -
##     percent.fin.aid + epsilon) + on.campus.housing + gap20repub +
##     private + percent.student.loan + mask.mandated.days + occupational.degree +
##     hs.equivalent.degree)
## Model 2: log(cases + epsilon) ~ (religious + tuition + log(total.headcount +
##     epsilon) + log(percent.american.native + epsilon) + log(percent.asian +
##     epsilon) + log(percent.black + epsilon) + log(percent.hispanic.latino +
##     epsilon) + log(percent.pacific.islander + epsilon) + percent.white +
```

```
##     percent.two.more.races + percent.women + grad.rate + log(100 -
##     percent.fin.aid + epsilon) + on.campus.housing + gap20repub +
##     private + percent.student.loan + mask.mandated.days + occupational.degree +
##     hs.equivalent.degree)^2
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1   1666 1993.4
## 2   1495 1464.2 171    529.26 3.1603 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: log(cases + epsilon) ~ religious + log(total.headcount + epsilon) +
##     +percent.fin.aid + log(dorm.capacity + epsilon) + gap20repub
## Model 2: log(cases + epsilon) ~ (religious + log(total.headcount + epsilon) +
##     +percent.fin.aid + log(dorm.capacity + epsilon) + gap20repub)^2
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   1774 2598.4
## 2   1764 2480.6 10    117.78 8.3758 2.02e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Talk a couple p-values
- From the ESS F-test, we can see that the interaction terms with just religious contribute to the model. Furthermore, the full interaction terms contribute in addition to the just religious terms.
- conclude this section? there is a huge weight with interaction terms, but religious does not interact with other variables as much as they interact with one another.

In Search of a Parsimonious Model: Sequential Variable Selection Models

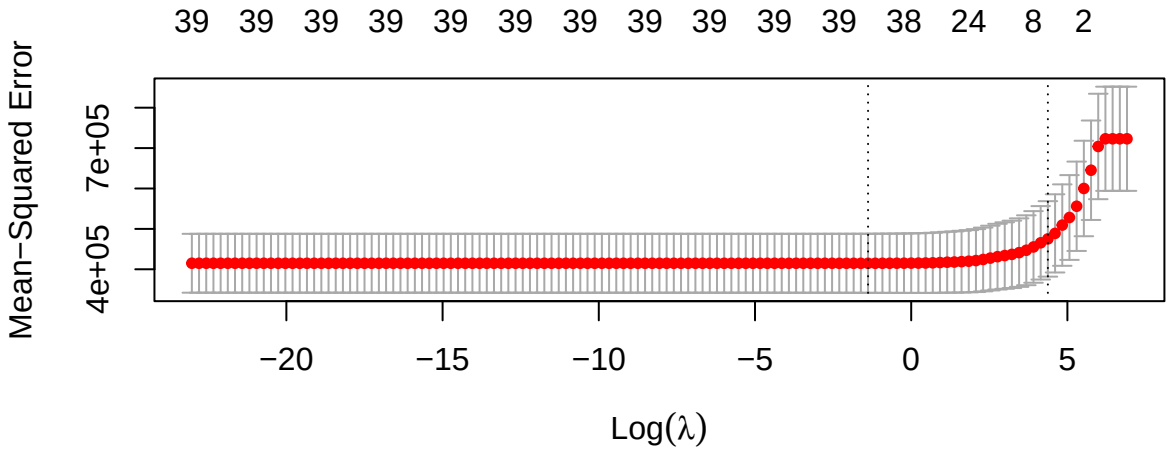
Honestly not sure how to interpret and forward step model and the both-directions step model take forever to run.

LASSO for Variable Section

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.

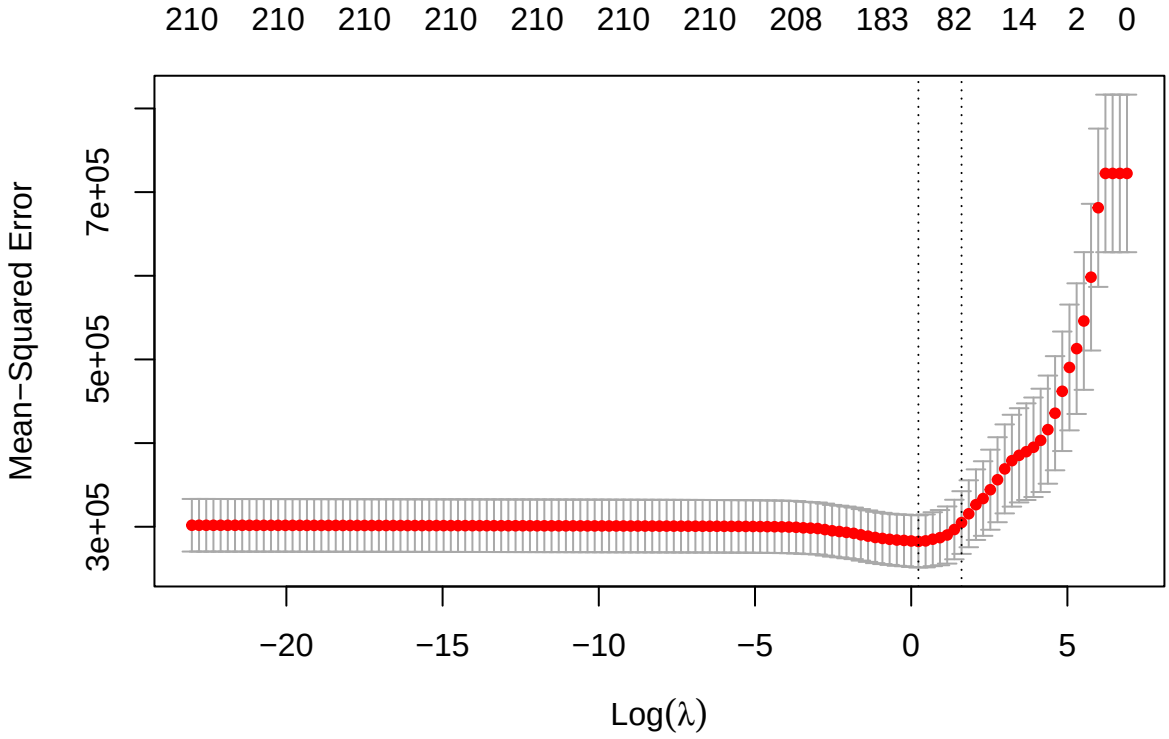
We will apply LASSO to both the `religiousInteraction` and `fullInteraction` models, and interpret the results of each model. In each case, we will use cross-validation to select the best regularizing constant λ with which to fit each model.

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 — and print the list of coefficients whose estimates were shrunk all the way to zero.




```
## [1] "(Intercept)"
```

Clearly, the list is short — most of the predictors were not shrunk all the way to zero. We now perform the same procedure with the `fullInteraction` model, creating LASSO Model 2:




```
## [1] "(Intercept)"
## [2] "tuition:log(total.headcount + epsilon)"
## [3] "log(percent.american.native + epsilon):percent.two.more.races"
## [4] "log(percent.hispanic.latino + epsilon):log(percent.pacific.islander + epsilon)"
## [5] "percent.two.more.races:gap20repub"
## [6] "gap20repub:hs.equivalent.degreeYes"
```

This time around, there are 11 coefficient estimates that were shrunk all the way to zero.

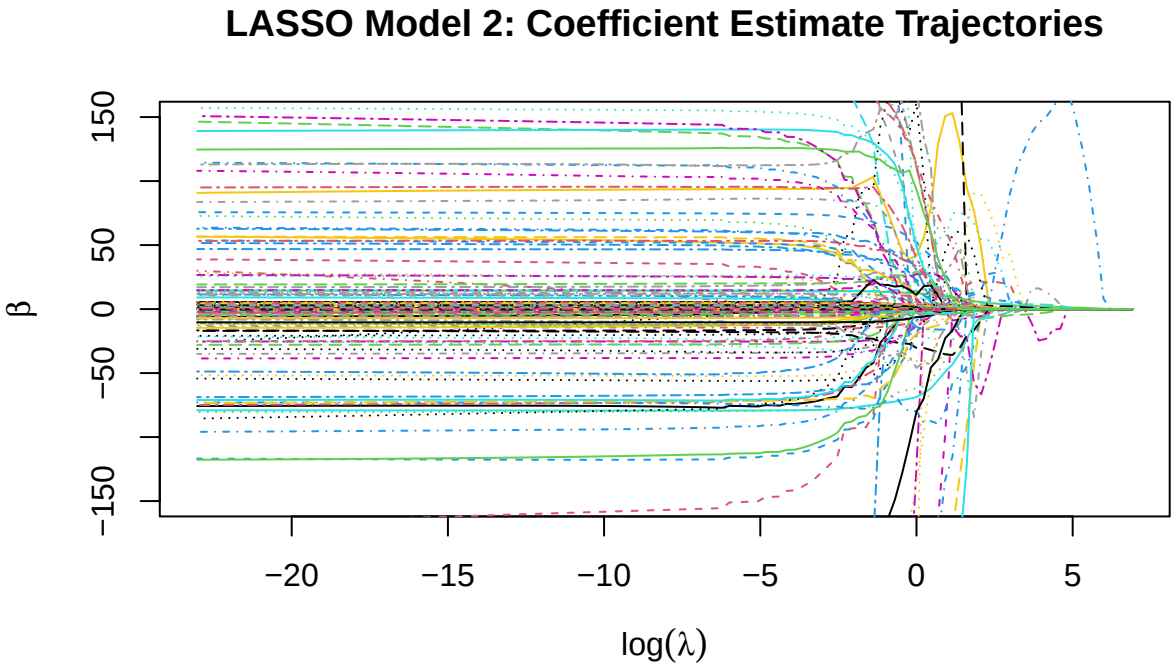
In order to determine which of the remaining predictors in each model are statistically significant, we refit a linear model using all the coefficients that were not shrunk to zero in each of LASSO Models 1 and 2. Note that even though the intercept was shrunk to zero by LASSO in both cases, we still 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. We then list the coefficients that we find to be statistically significant.

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 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 whether a school has a religious affiliation affects the relationship between five other predictors and the case counts for the fall 2020 semester, the second suggests that only the association between `mask.mandated.days` and `cases` changes based on whether or not an institution has a religious affiliation. 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 interaction effects 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.



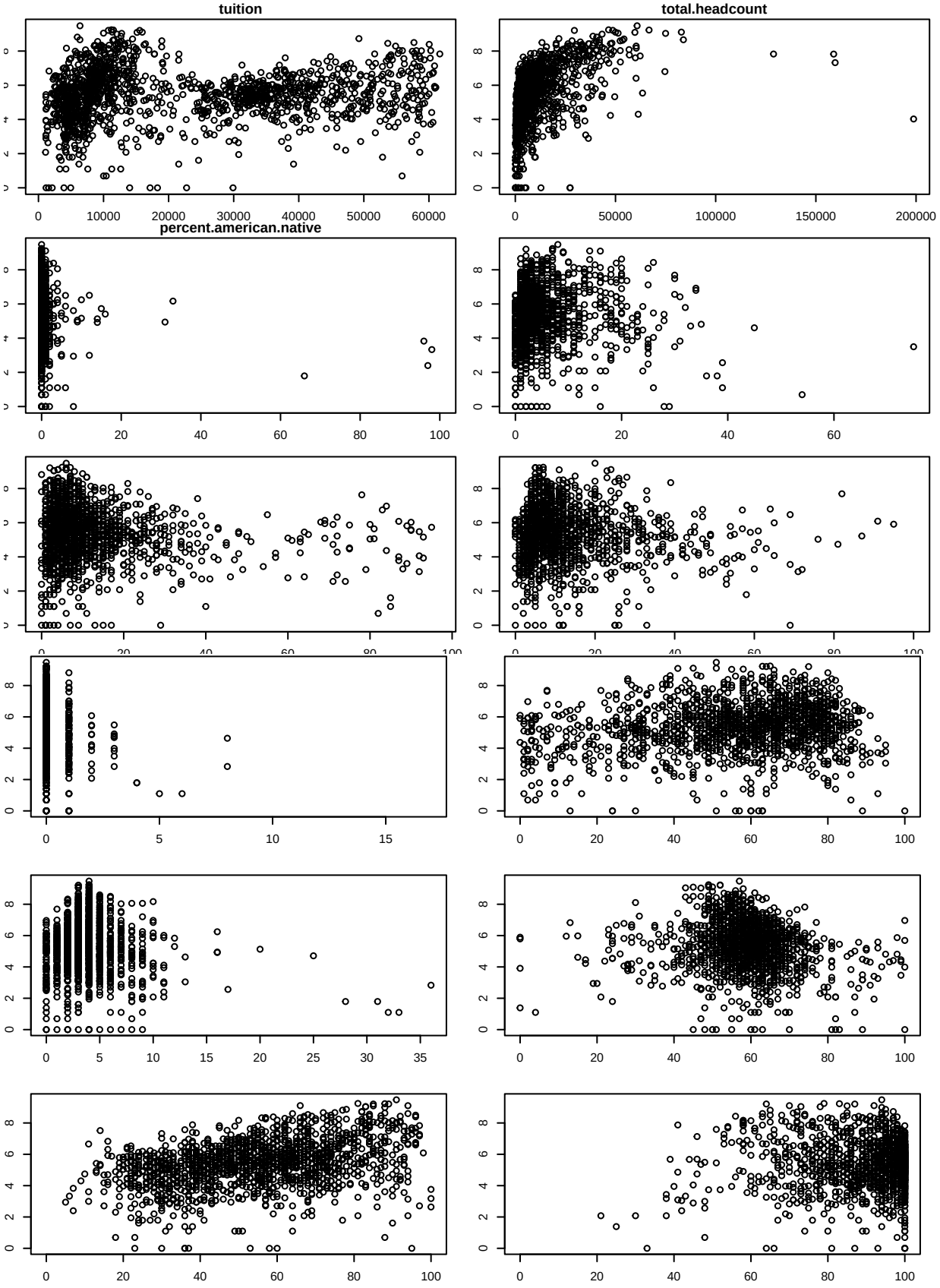
In the end, we lean towards accepting the conclusions presented by the second model: With a wider range of interaction effects to consider, it is less likely that the statistically significant predictors were only marked as such because they help explain a relationship between `cases` and another predictor with which they are collinear.

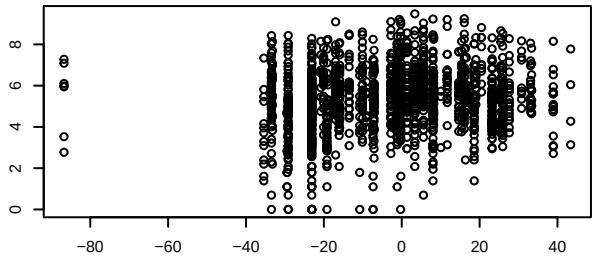
LASTLY, COMPARE LASSO RESULTS TO SEQUENTIAL VARIABLE SELECTION RESULTS

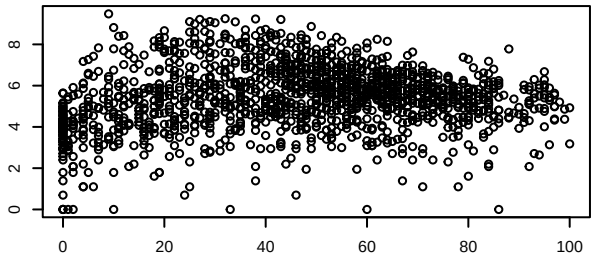
Hierarchical Multi-level Models

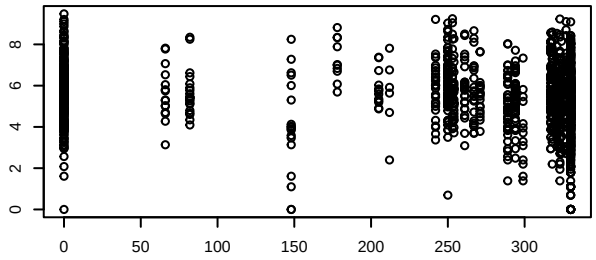
Conclusions

Appendix A: Plots to Check the Assumption of Linearity




0
2
4
6
8
−80
−60
−40
−20
0
20
40


0
2
4
6
8
0
20
40
60
80
100


0
2
4
6
8
0
50
100
150
200
250
300