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.

The data is 1,855 rows after removing Universities without stats or without matching IPEDS ids. It has 40 columns, including IPEDS id, university name, cases, and predictor variables based on college attributes.

Data Cleaning Procedures

For this exploratory data analysis, the first step 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 the `control` column from the data frame, since it now contains redundant information.

Finally, we 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. Below, 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.

We then 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 2020 presidential election voter gap data into a data frame, and merge this data frame — with two columns: state and 2020 voter gap — with our data frame of observations on the `state` variable.

```
elections.data <- read.csv("data/pres_elections.csv")
voter.gap <- elections.data[,c("state", "gap20repub")]
voter.gap$state[voter.gap$state == "DC"] <- "Washington, D.C."

full.cases <- merge(full.cases, voter.gap, by="state")
write.csv(full.cases, "full_cases.csv")
```

Description of variables

Add description of all the variables considered in our analyses. COULD PROBABLY SAVE THIS FOR THE END

Group Testing

Explain motivation for group testing

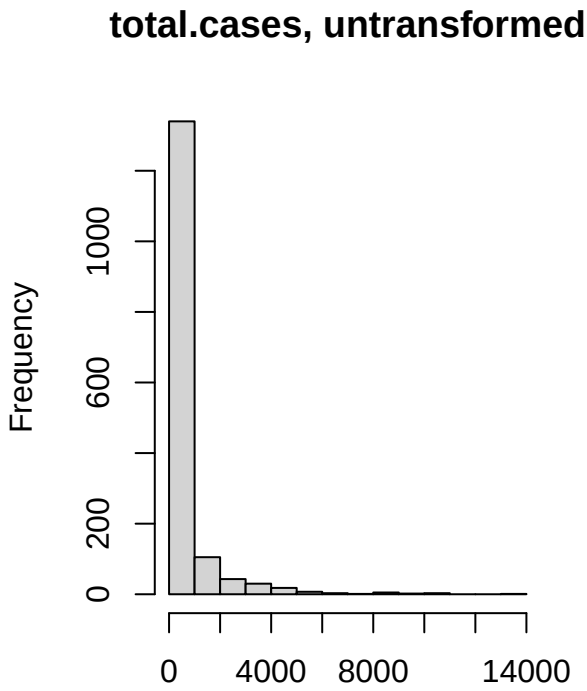
Checking the assumptions for t-based methods

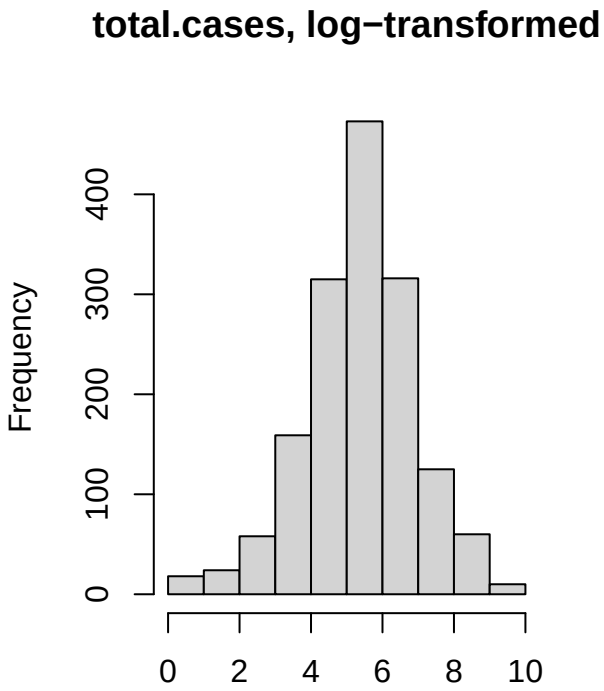
For unpooled t-based test for a difference in sample means, there are three assumptions:

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

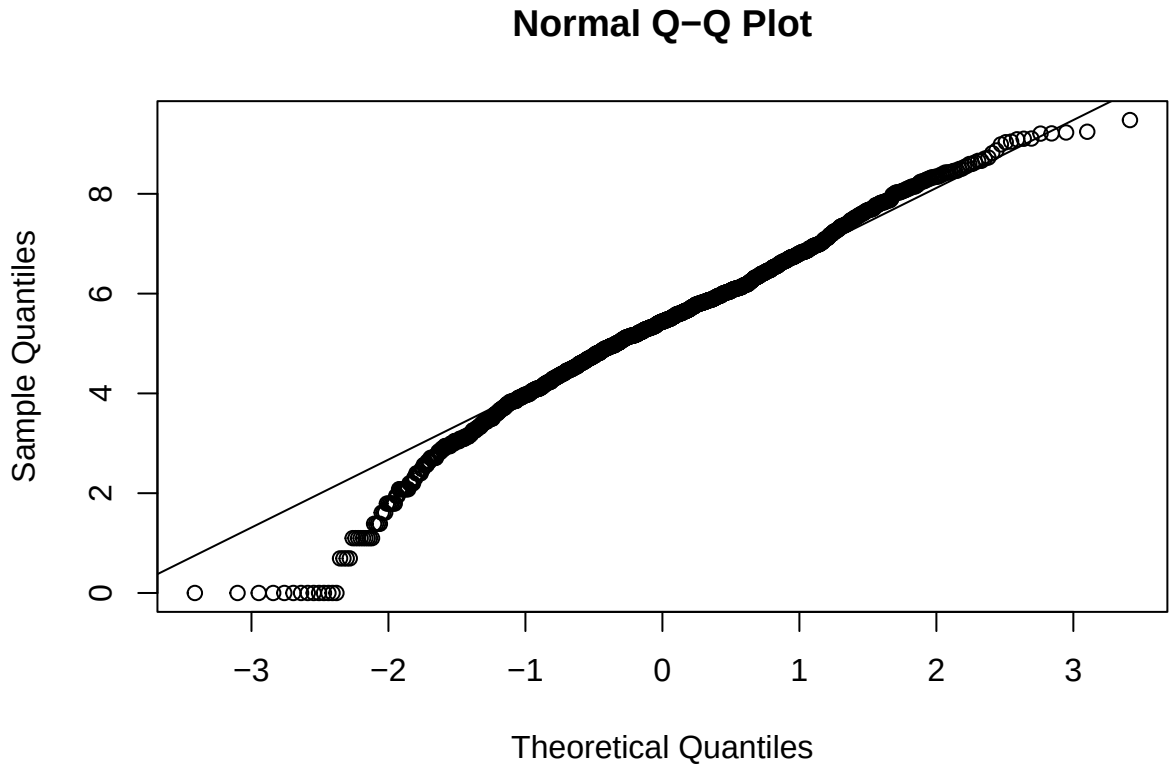
```
epsilon <- 1

par(mfrow=c(1,2))
hist(full.cases$total.cases, main="total.cases, untransformed")
hist(log(full.cases$total.cases + epsilon), main="total.cases, log-transformed")
```






```
qqnorm(log(full.cases$total.cases + epsilon))
qqline(log(full.cases$total.cases + epsilon))
```



Student-t Tests

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

```
t.test(log(total.cases + epsilon) ~ religious, data=full.cases)
```

```
##
##  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

```
for.rel.affil <- full.cases[,c("total.cases", "religious.affiliation")]
for.rel.affil <- for.rel.affil[complete.cases(for.rel.affil),]
summary(aov(total.cases ~ religious.affiliation, data=for.rel.affil))
```

```
##                          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

```
wilcox.test(x = full.cases$total.cases[full.cases$religious == "Yes"],
            y = full.cases$total.cases[full.cases$religious == "No"],
            alternative='two.sided', exact = FALSE, correct = FALSE,
            conf.int = TRUE)
```

```
##
##  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

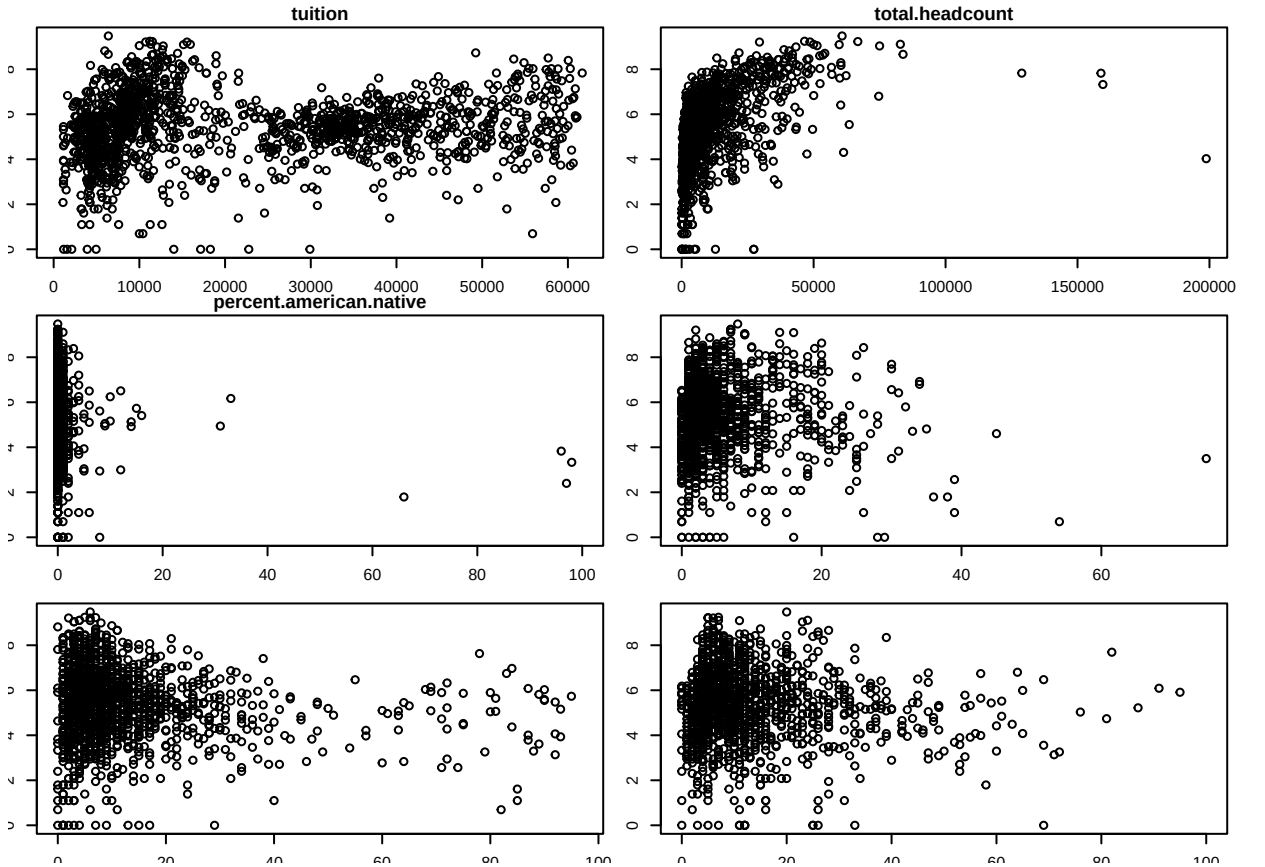
Basic Linear Regression Models

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.

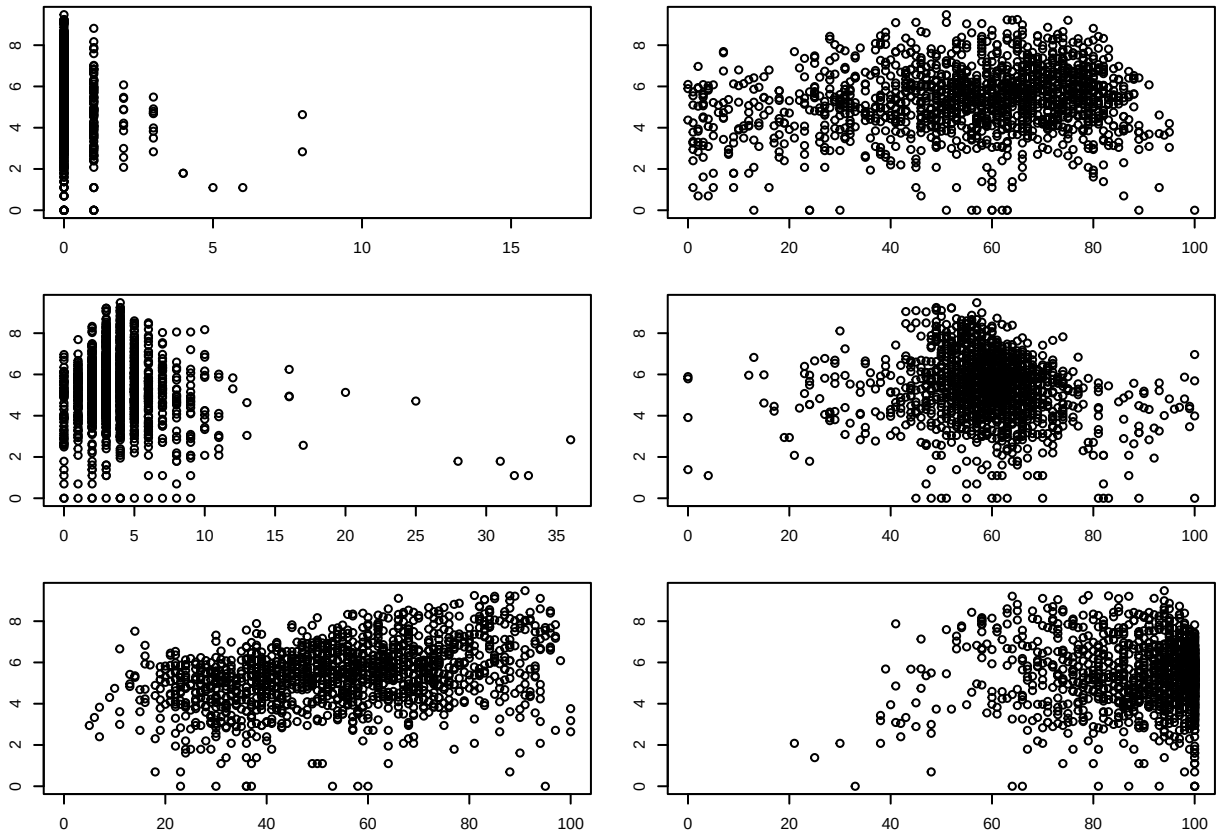
Checking the Assumptions of Linear Regression Models

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:

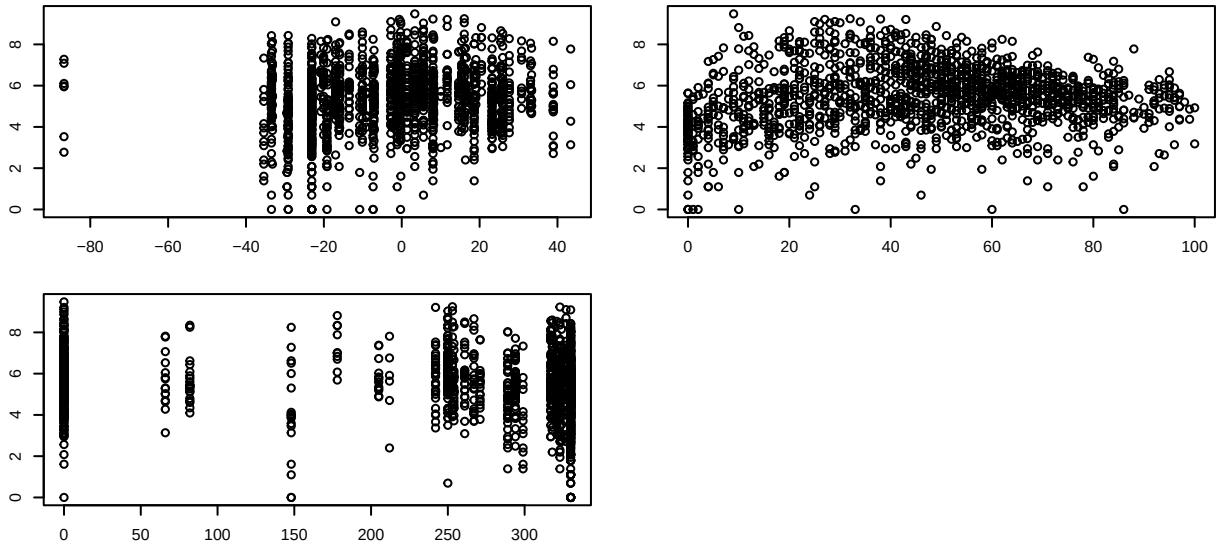
```
par(mfrow=c(3,2), mar=1.5 * c(1,1,1,1), cex=0.5)
plot(log(total.cases + epsilon) ~ tuition, data=full.cases,
     main="tuition")
plot(log(total.cases + epsilon) ~ total.headcount, data=full.cases,
     main="total.headcount")
plot(log(total.cases + epsilon) ~ percent.american.native, data=full.cases,
     main="percent.american.native")
plot(log(total.cases + epsilon) ~ percent.asian, data=full.cases)
plot(log(total.cases + epsilon) ~ percent.black, data=full.cases)
plot(log(total.cases + epsilon) ~ percent.hispanic.latino, data=full.cases)
```



```
par(mfrow=c(3,2), mar=2 * c(1,1,1,1), cex=0.5)
plot(log(total.cases + epsilon) ~ percent.pacific.islander, data=full.cases)
plot(log(total.cases + epsilon) ~ percent.white, data=full.cases)
plot(log(total.cases + epsilon) ~ percent.two.more.races, data=full.cases)
plot(log(total.cases + epsilon) ~ percent.women, data=full.cases)
plot(log(total.cases + epsilon) ~ grad.rate, data=full.cases)
plot(log(total.cases + epsilon) ~ percent.fin.aid, data=full.cases)
```

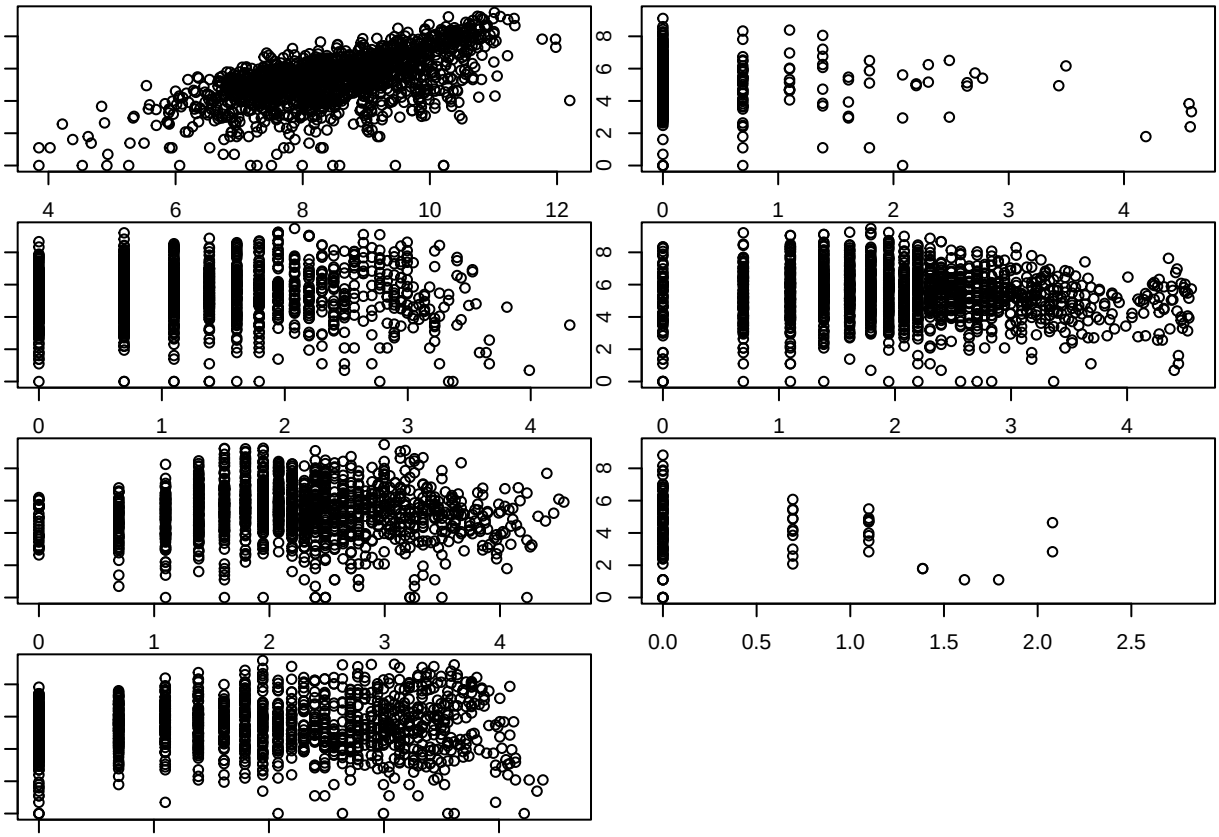



```
par(mfrow=c(3,2), mar=2 * c(1,1,1,1), cex=0.5)
plot(log(total.cases + epsilon) ~ gap20repub, data=full.cases)
plot(log(total.cases + epsilon) ~ percent.student.loan, data=full.cases)
plot(log(total.cases + epsilon) ~ mask.mandated.days, data=full.cases)
```



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-skenwess 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.

```
par(mfrow=c(4,2), mar=c(1,1,1,1))
plot(log(total.cases + epsilon) ~ log(total.headcount), data=full.cases)
plot(log(total.cases + epsilon) ~ log(percent.american.native), data=full.cases)
plot(log(total.cases + epsilon) ~ log(percent.asian), data=full.cases)
plot(log(total.cases + epsilon) ~ log(percent.black), data=full.cases)
plot(log(total.cases + epsilon) ~ log(percent.hispanic.latino), data=full.cases)
plot(log(total.cases + epsilon) ~ log(percent.pacific.islander), data=full.cases)
plot(log(total.cases + epsilon) ~ log(100-percent.fin.aid+1), data=full.cases)
```

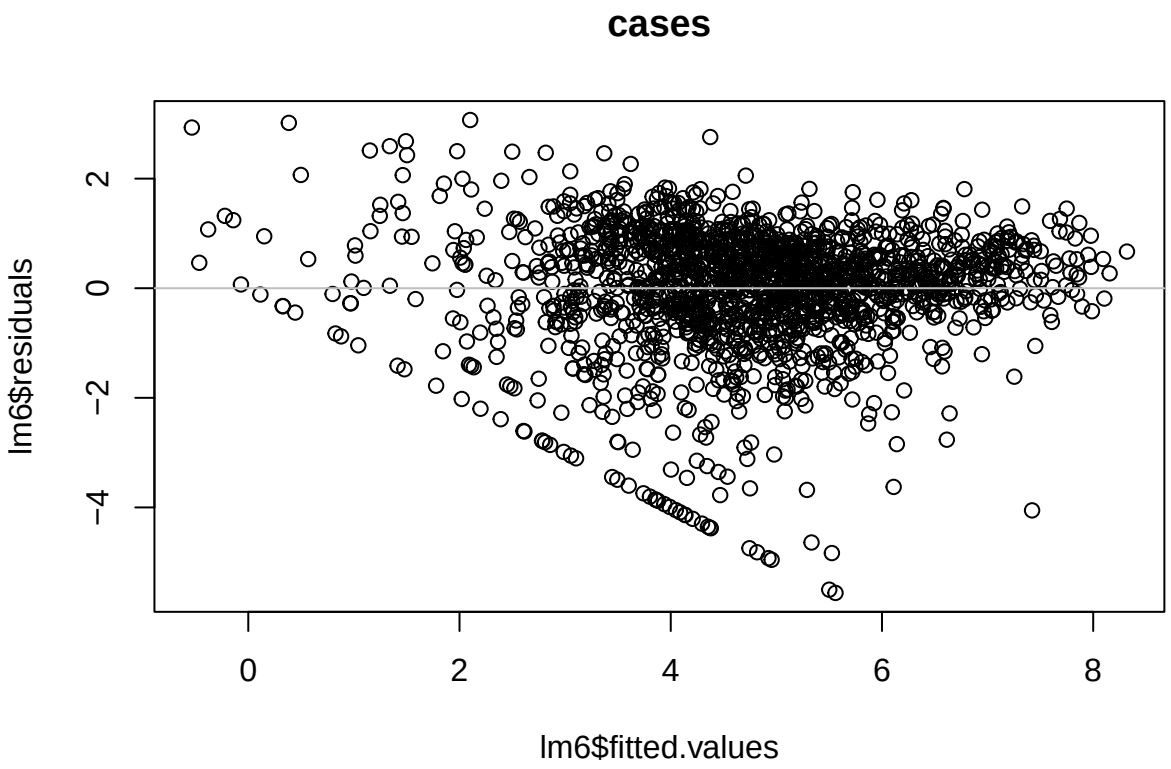


After these transformations, the linearity assumptions seems much more reasonable for each of the quantitative predictors.

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.

```
lm6 <- lm(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, data=full.cases)

plot(lm6$residuals ~ lm6$fitted.values, main="cases")
abline(h=0, col="gray")
```



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:

```
vcov.robust = vcovHC(lm6, type="HC")
data.frame(ols=round(summary(lm6)$coef[,'Std. Error'],4),
           robust=round(sqrt(diag(vcov.robust)),4),
           diff=round(abs(sqrt(diag(vcov.robust)) -
                          summary(lm6)$coef[,'Std. Error']),4))
```

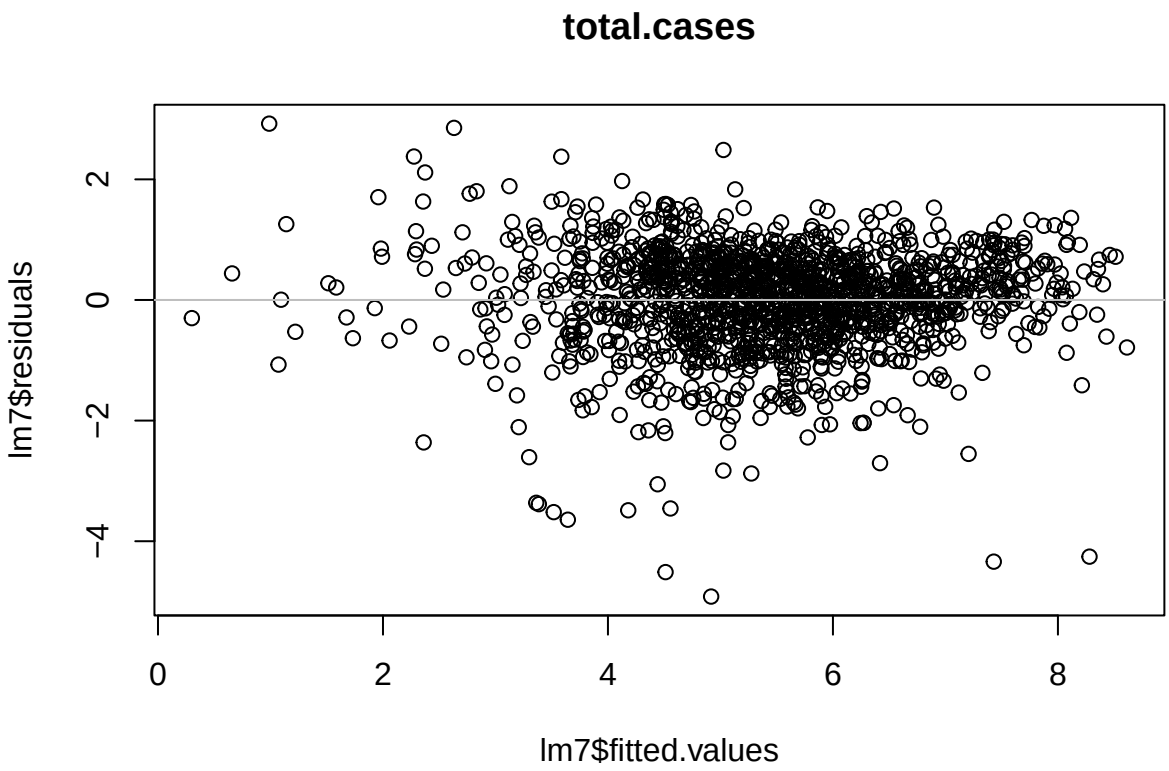
```
##                                        ols robust   diff
## (Intercept)                         0.4099 0.4703 0.0604
## religiousYes                        0.0892 0.0870 0.0022
## tuition                             0.0000 0.0000 0.0000
## log(total.headcount + epsilon)      0.0316 0.0369 0.0052
## 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

```
lm7 <- lm(log(total.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, data=full.cases)

plot(lm7$residuals ~ lm7$fitted.values, main="total.cases")
abline(h=0, col="gray")
```



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. EXPAND UPON IF NECESSARY

Inferences

Now that we have concluded that the assumptions of linear regression are reasonable, we can begin to perform inferences using these models.

```
modelsummary(lm6)
```

Linear Models with Interaction Effects

In Search of a Parsimonious Model: Sequential Variable Selection Models

```
df = full.cases[,c("percent.cases","gap20repub", "mask.mandated.days", "private","tuition","total.headc

df = na.omit(df)

summary(model1 <- lm(log(percent.cases+epsilon) ~ mask.mandated.days + private+tuition+total.headcount+t


interactionModel <- lm(log(percent.cases+epsilon) ~ (mask.mandated.days+ private+tuition+total.headcoun


back.step <- step(model1, direction = "backward", k = 2)

## including states, excluding days mandated:
# removed columns are control, undergrad.headcount, percent.pacific.islander,on.campus.housing

## including days mandated, excluding states:
# removed columns are control, percent.american.native + percent.two.more.races+percent.NA.race,on.camp

forward.step = step(model1, scope = list(upper = formula(interactionModel)), direction = "forward")

model0 = lm(log(percent.cases+1) ~ 1, full.cases)

step = step(model1, scope = list(lower = formula(model0), upper = formula(interactionModel)),
direction = "both")
```

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

```
# Not included in the below model: NCAA.football, religious.affiliation,
# mask.mandated days, as well as those subtracted from the formula of
# the commented out lm1 below
```

	Model 1
(Intercept)	−5.719
	(0.410)
religiousYes	0.102
	(0.089)
tuition	0.000 02
	(0.000 004)
log(total.headcount + epsilon)	0.920
	(0.032)
log(percent.american.native + epsilon)	0.013
	(0.056)
log(percent.asian + epsilon)	−0.125
	(0.056)
log(percent.black + epsilon)	0.195
	(0.044)
log(percent.hispanic.latino + epsilon)	0.060
	(0.044)
log(percent.pacific.islander + epsilon)	−0.289
	(0.118)
percent.white	0.018
	(0.002)
percent.two.more.races	−0.025
	(0.011)
percent.women	−0.011
	(0.003)
grad.rate	0.021
	(0.002)
log(100 - percent.fin.aid + epsilon)	−0.043
	(0.031)
on.campus.housingYes	0.823
	(0.094)
gap20repub	0.013
	(0.002)
privateYes	−0.677
	(0.148)
percent.student.loan	0.005
	(0.002)
mask.mandated.days	0.0002
	(0.0003)
occupational.degreeYes	−0.184
	(0.077)
hs.equivalent.degreeYes	0.136
	(0.095)
Num.Obs.	1705
R2	0.604
R2 Adj.	0.599
AIC	
BIC	
Log.Lik.	−2590.545
RMSE	1.11

```r
full.cases$percent.cases <- full.cases$total.cases/full.cases$total.headcount

summary(lm1 <- lm(log(percent.cases + epsilon) ~ religious + catholic +
            tuition + total.headcount +
            undergrad.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 + grad.rate + percent.fin.aid +
            percent.disability + on.campus.housing + state +
            private + percent.student.loan +
            occupational.degree + hs.equivalent.degree,
          data=full.cases))

summary(lm2 <- lm(log(percent.cases + epsilon) ~ religious + private + tuition, full.cases))

summary(lm3 <-lm(log(percent.cases+epsilon)~(religious + tuition + percent.white+percent.black+mask.mand

full.cases.tuit = full.cases[complete.cases(full.cases[,c("tuition")]),]

lm5 = lm(percent.cases ~ poly(tuition, 3, raw=TRUE), data = full.cases.tuit)
summary(lm5)

x=500:61500
yhat = predict(lm5,new=data.frame(tuition=x))

plot(percent.cases~tuition,data=full.cases.tuit)
lines(yhat~x,col="magenta",lwd=4)

df = full.cases

# avg.grant.money, dorm.capacity, dorm.room.price, academic.degree
```