

# 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 with 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 finally create a column `total.cases` in `full.cases` that sums the `cases` and `cases_2021` columns.

## 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

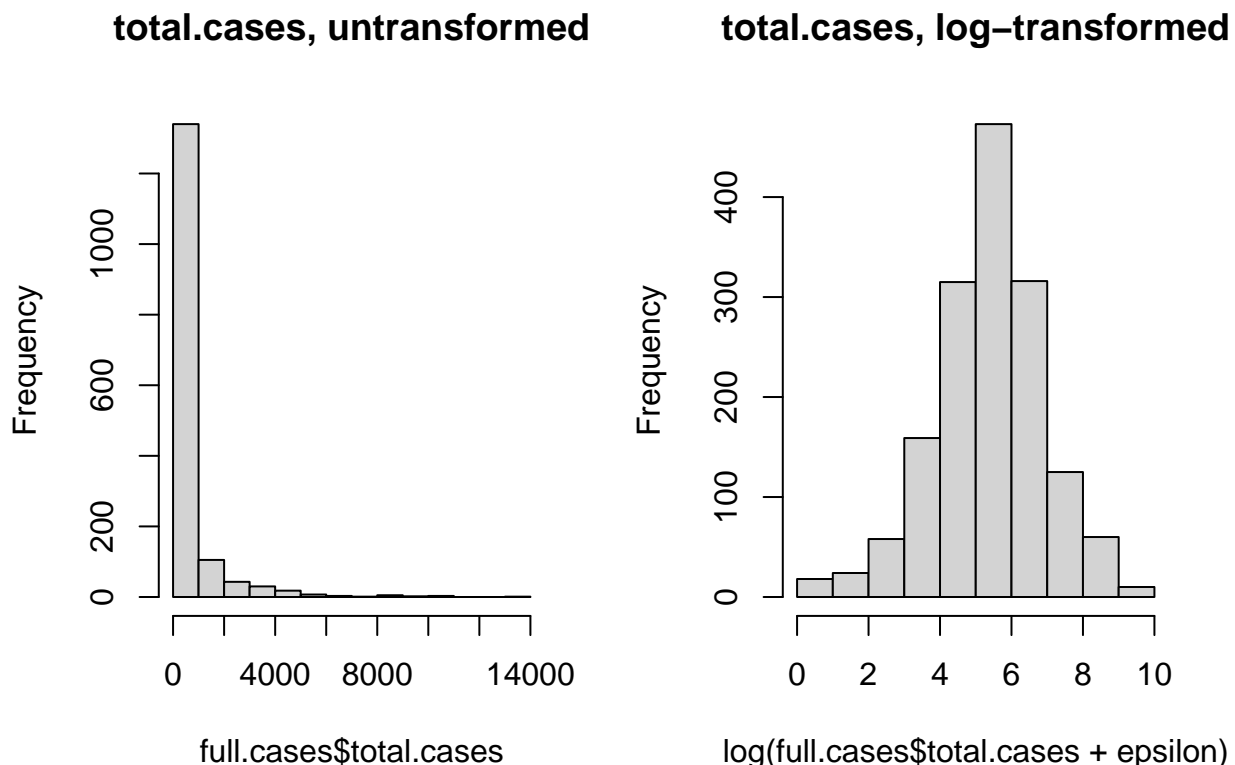
## Checking the assumptions for *t*-based methods

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.**

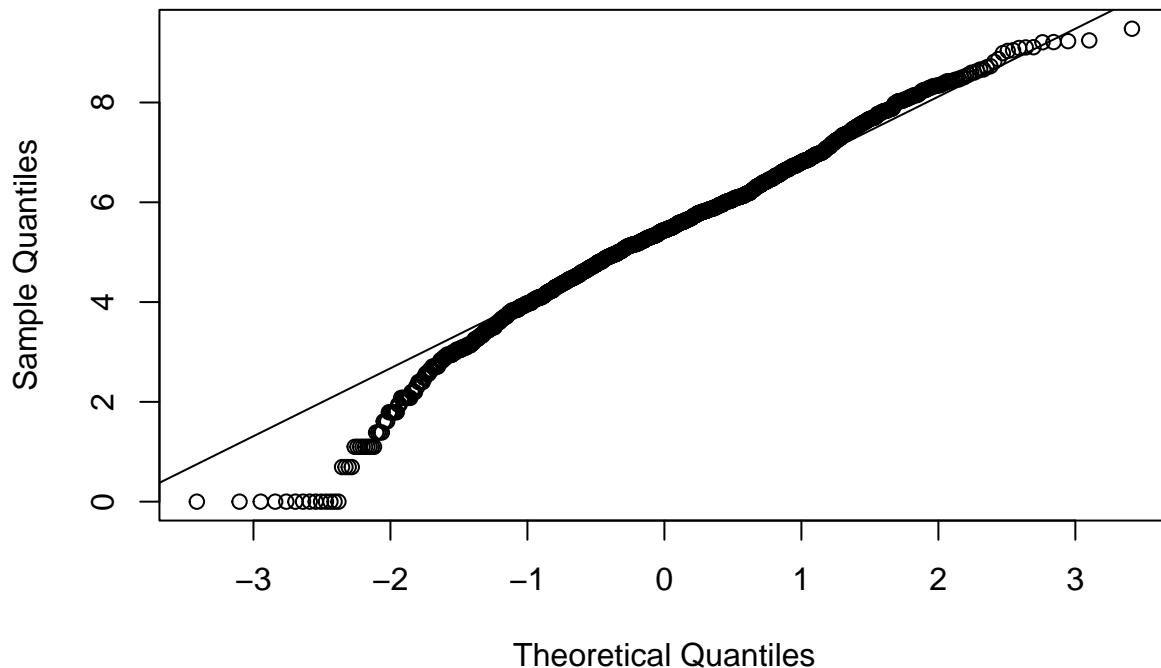
```
epsilon <- 1

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))
```

## Normal Q-Q Plot



## 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 — Wilcoxon 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$total.cases[full.cases$religious == "No"]
## 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

## 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.

```
# 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
```

```
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))
```

```
##
## Call:
## lm(formula = 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)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13488 -0.02893 -0.00546  0.02275  0.36797
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.141e-02  2.134e-01   0.428  0.668525
## religiousYes      2.942e-02  5.824e-03   5.051  5.35e-07 ***
## catholicYes     -1.239e-02  6.410e-03  -1.933  0.053504 .
## tuition           2.768e-07  2.763e-07   1.002  0.316582
## total.headcount  -4.080e-07  6.790e-07  -0.601  0.548088
## undergrad.headcount  1.069e-06  8.594e-07   1.244  0.213790
## percent.american.native  1.901e-03  2.392e-03   0.795  0.426946
## percent.asian     -2.032e-03  2.185e-03  -0.930  0.352734
## percent.black     -3.107e-04  2.131e-03  -0.146  0.884134
## percent.hispanic.latino -8.024e-05  2.139e-03  -0.038  0.970086
## percent.pacific.islander  2.732e-03  6.652e-03   0.411  0.681437
## percent.white      3.322e-04  2.128e-03   0.156  0.875987
## percent.two.more.races -2.791e-03  2.326e-03  -1.200  0.230550
## percent.NA.race   -1.043e-03  2.117e-03  -0.493  0.622383
## percent.nonres.alien -7.343e-04  2.153e-03  -0.341  0.733119
## percent.women     -8.447e-04  1.569e-04  -5.385  9.34e-08 ***
## grad.rate         9.716e-04  1.711e-04   5.677  1.87e-08 ***
## percent.fin.aid   -2.610e-04  1.754e-04  -1.488  0.137131
## percent.disability  3.668e-04  3.310e-04   1.108  0.268124
## on.campus.housingYes  2.514e-03  7.129e-03   0.353  0.724434
## stateAlaska      -4.846e-02  5.472e-02  -0.886  0.376059
## stateArizona     -2.007e-02  2.989e-02  -0.671  0.502101
## stateArkansas      1.238e-02  2.108e-02   0.587  0.557030
## stateCalifornia   -4.858e-02  1.890e-02  -2.570  0.010322 *
## stateColorado     -2.696e-02  2.070e-02  -1.302  0.193115
## stateConnecticut  -1.472e-02  1.960e-02  -0.751  0.452729
## stateDelaware      4.084e-02  3.825e-02   1.068  0.285880
## stateFlorida     -3.149e-02  1.799e-02  -1.751  0.080332 .
## stateGeorgia     -5.516e-03  1.895e-02  -0.291  0.771010
## stateHawaii        3.931e-02  5.740e-02   0.685  0.493604
## stateIdaho       -5.228e-02  2.952e-02  -1.771  0.076961 .
## stateIllinois     -3.425e-02  1.752e-02  -1.956  0.050828 .
## stateIndiana     -7.493e-03  1.885e-02  -0.397  0.691108
## stateIowa        -1.640e-02  2.001e-02  -0.820  0.412508
## stateKansas       -3.584e-02  2.664e-02  -1.345  0.178844
## stateKentucky     -1.922e-02  1.827e-02  -1.052  0.293061
## stateLouisiana    -3.460e-03  2.108e-02  -0.164  0.869652
```

```

## stateMaine -5.557e-02 2.330e-02 -2.385 0.017280 *
## stateMaryland -3.666e-02 2.025e-02 -1.810 0.070567 .
## stateMassachusetts -4.335e-02 1.709e-02 -2.537 0.011357 *
## stateMichigan -6.824e-03 1.757e-02 -0.388 0.697858
## stateMinnesota -1.831e-02 1.747e-02 -1.048 0.294910
## stateMississippi -3.571e-02 2.659e-02 -1.343 0.179516
## stateMissouri -2.356e-02 1.892e-02 -1.245 0.213403
## stateMontana -1.578e-02 3.244e-02 -0.486 0.626891
## stateNebraska -1.593e-02 2.384e-02 -0.668 0.504096
## stateNevada -1.987e-02 2.973e-02 -0.668 0.504201
## stateNew Hampshire 2.752e-02 2.259e-02 1.218 0.223439
## stateNew Jersey -3.730e-02 1.851e-02 -2.016 0.044146 *
## stateNew Mexico -3.348e-02 3.752e-02 -0.892 0.372402
## stateNew York -2.126e-02 1.589e-02 -1.338 0.181274
## stateNorth Carolina 1.572e-03 1.765e-02 0.089 0.929046
## stateNorth Dakota 1.088e-02 3.219e-02 0.338 0.735535
## stateOhio -3.192e-02 1.680e-02 -1.901 0.057687 .
## stateOklahoma -3.602e-02 3.299e-02 -1.092 0.275240
## stateOregon -6.024e-02 2.112e-02 -2.852 0.004449 **
## statePennsylvania -2.903e-02 1.616e-02 -1.797 0.072737 .
## stateRhode Island 7.803e-03 2.322e-02 0.336 0.736864
## stateSouth Carolina 4.037e-02 1.906e-02 2.118 0.034443 *
## stateSouth Dakota 1.399e-02 5.118e-02 0.273 0.784602
## stateTennessee -3.705e-02 1.960e-02 -1.890 0.059060 .
## stateTexas -2.730e-02 1.831e-02 -1.491 0.136231
## stateUtah -3.513e-03 2.885e-02 -0.122 0.903117
## stateVermont -8.213e-02 2.458e-02 -3.342 0.000868 ***
## stateVirginia -7.066e-03 1.799e-02 -0.393 0.694665
## stateWashington -3.475e-02 1.907e-02 -1.822 0.068797 .
## stateWashington, D.C. -6.769e-02 2.683e-02 -2.523 0.011820 *
## stateWest Virginia 8.546e-03 2.304e-02 0.371 0.710768
## stateWisconsin -2.606e-02 1.827e-02 -1.427 0.153964
## stateWyoming 6.018e-03 3.793e-02 0.159 0.873990
## privateYes 1.229e-02 9.430e-03 1.304 0.192682
## percent.student.loan 2.056e-04 1.231e-04 1.670 0.095179 .
## occupational.degreeYes -5.572e-03 4.773e-03 -1.167 0.243390
## hs.equivalent.degreeYes -1.958e-03 6.682e-03 -0.293 0.769634
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04871 on 873 degrees of freedom
## (908 observations deleted due to missingness)
## Multiple R-squared: 0.4706, Adjusted R-squared: 0.4263
## F-statistic: 10.63 on 73 and 873 DF, p-value: < 2.2e-16

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

##
## Call:
## lm(formula = log(percent.cases + epsilon) ~ religious + private +
## tuition, data = full.cases)
##
## Residuals:
## Min 1Q Median 3Q Max
## -0.12584 -0.03696 -0.01342 0.02519 0.36211

```

```
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.830e-02  2.451e-03  15.627 < 2e-16 ***
## religiousYes  3.744e-02  4.741e-03   7.896 5.58e-15 ***
## privateYes   -1.361e-02  7.616e-03  -1.788  0.0741 .
## tuition      1.292e-06  1.821e-07   7.095 2.00e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05657 on 1471 degrees of freedom
## (380 observations deleted due to missingness)
## Multiple R-squared:  0.1898, Adjusted R-squared:  0.1881
## F-statistic: 114.9 on 3 and 1471 DF, p-value: < 2.2e-16

summary(lm3 <-lm(log(percent.cases+epsilon)~(religious + tuition + percent.white+percent.black+mask.mandated.days)^2, data = full.cases))

##
## Call:
## lm(formula = log(percent.cases + epsilon) ~ (religious + tuition +
##      percent.white + percent.black + mask.mandated.days)^2, data = full.cases)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13558 -0.03115 -0.00840  0.02062  0.34810
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.243e-02  1.959e-02   1.655  0.0980 .
## religiousYes      3.625e-02  2.328e-02   1.557  0.1197
## tuition          -9.967e-07  4.677e-07  -2.131  0.0332 *
## percent.white     3.709e-04  2.610e-04   1.421  0.1555
## percent.black    -4.648e-05  3.019e-04  -0.154  0.8777
## mask.mandated.days -6.492e-05  6.355e-05  -1.022  0.3072
## religiousYes:tuition -6.292e-07  2.989e-07  -2.105  0.0355 *
## religiousYes:percent.white -1.001e-04  2.416e-04  -0.414  0.6786
## religiousYes:percent.black -3.698e-04  3.477e-04  -1.064  0.2877
## religiousYes:mask.mandated.days 5.043e-05  3.420e-05   1.474  0.1406
## tuition:percent.white 4.019e-08  5.532e-09   7.265 6.07e-13 ***
## tuition:percent.black 2.367e-08  1.186e-08   1.995  0.0462 *
## tuition:mask.mandated.days -5.434e-10  9.251e-10  -0.587  0.5570
## percent.white:percent.black -6.428e-06  4.201e-06  -1.530  0.1262
## percent.white:mask.mandated.days -4.134e-08  8.388e-07  -0.049  0.9607
## percent.black:mask.mandated.days 1.015e-06  1.006e-06   1.008  0.3134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05206 on 1459 degrees of freedom
## (380 observations deleted due to missingness)
## Multiple R-squared:  0.3194, Adjusted R-squared:  0.3124
## F-statistic: 45.64 on 15 and 1459 DF, p-value: < 2.2e-16

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

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

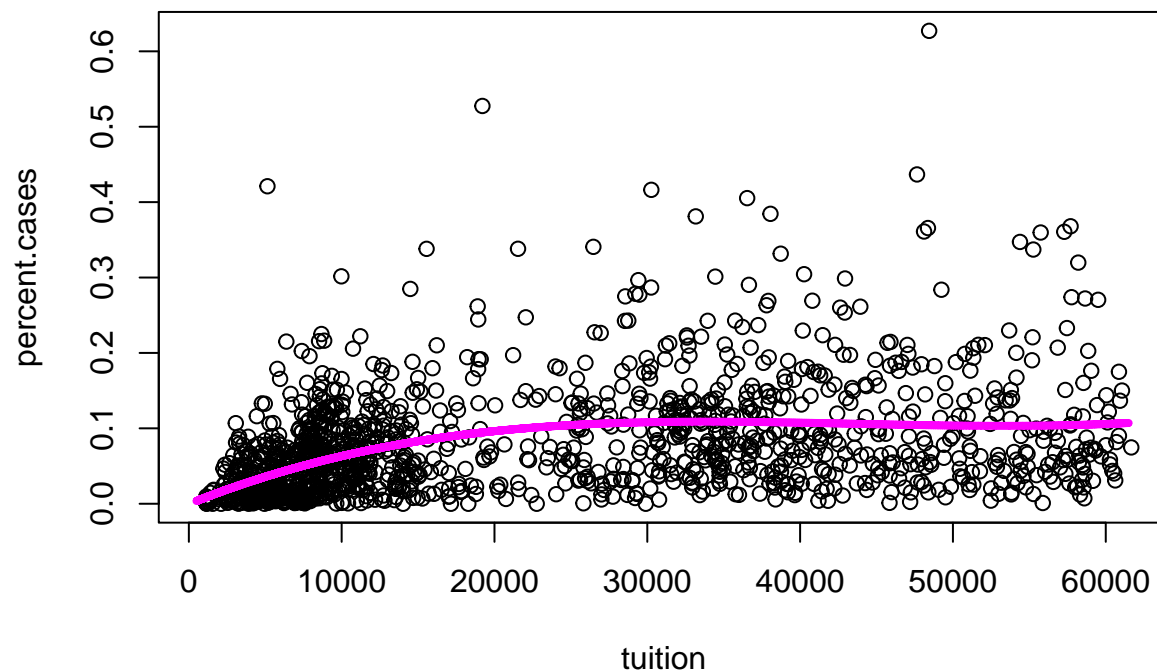
```
summary(lm5)
```

```
##
## Call:
## lm(formula = percent.cases ~ poly(tuition, 3, raw = TRUE), data = full.cases.tuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10801 -0.03985 -0.01046  0.02602  0.52299
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.167e-04  6.369e-03  -0.065  0.94784
## poly(tuition, 3, raw = TRUE)1  8.181e-06  1.087e-06   7.528 8.91e-14 ***
## poly(tuition, 3, raw = TRUE)2 -1.975e-10  4.249e-11  -4.649 3.64e-06 ***
## poly(tuition, 3, raw = TRUE)3  1.511e-15  4.616e-16   3.273 0.00109 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06421 on 1471 degrees of freedom
## (240 observations deleted due to missingness)
## Multiple R-squared:  0.1894, Adjusted R-squared:  0.1878
## F-statistic: 114.6 on 3 and 1471 DF,  p-value: < 2.2e-16
```

```
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
```



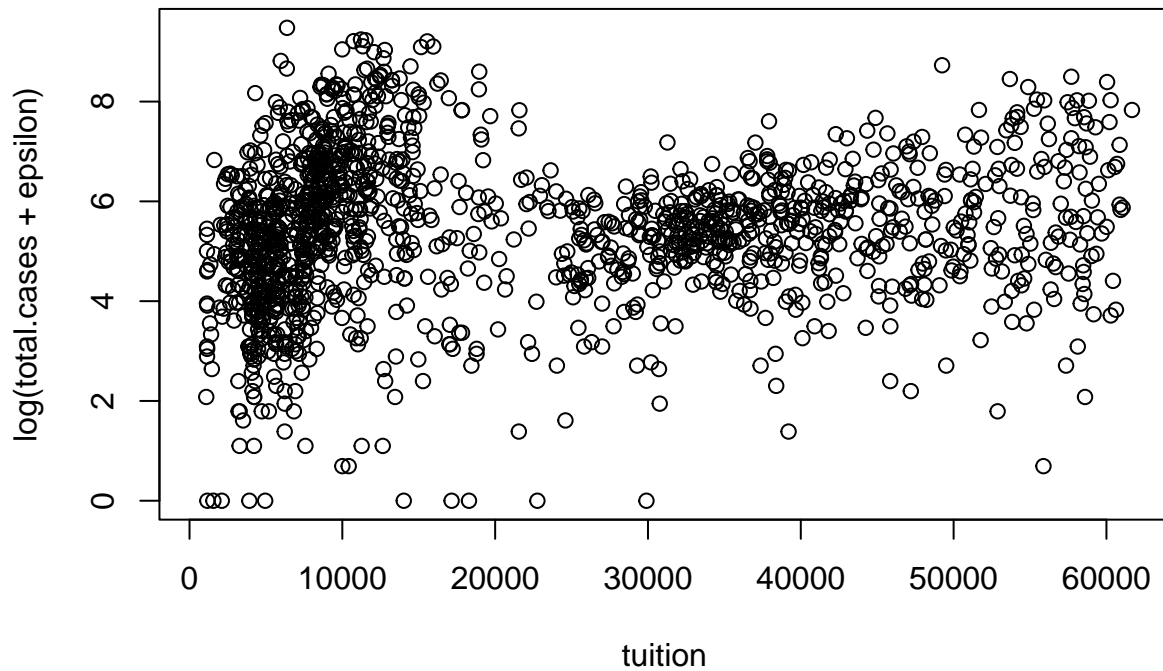
Can see across the board that religious status of school is a statistically significant predictor in the above models. Could also analyze specifics?

## Checking the Assumptions of Linear Regression

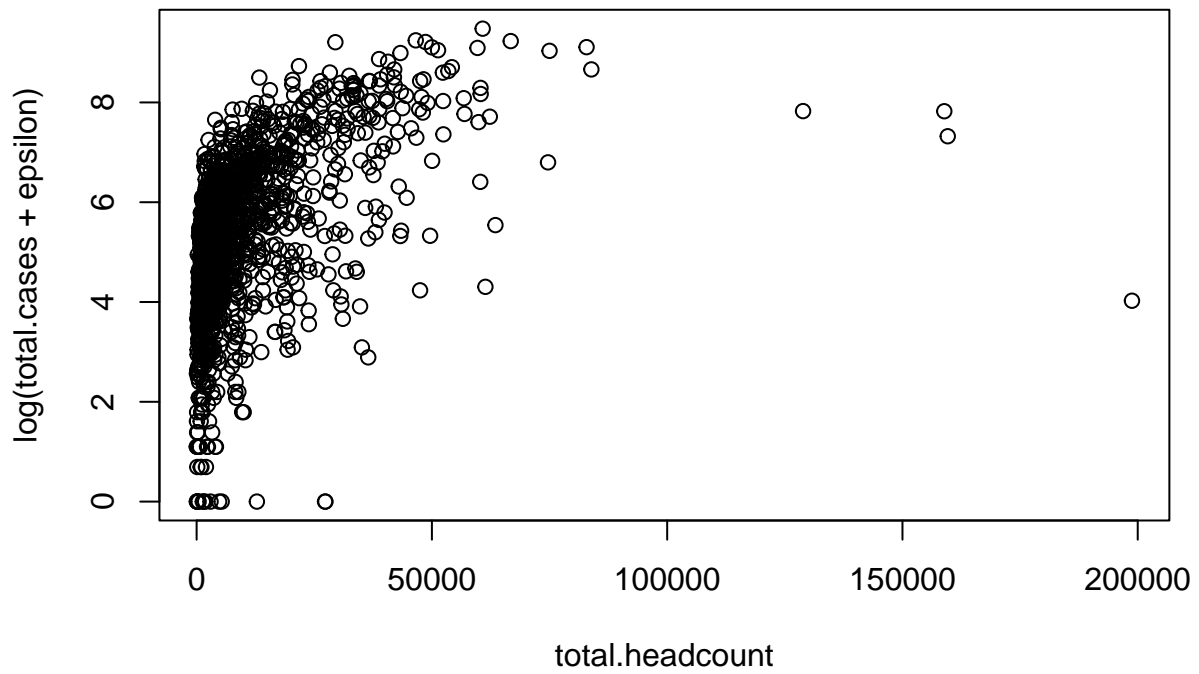
As discussed above in the group testing section, we believe that it is reasonable to assume that the observations in this data set are independent of one another, and that the log-transformed response variable `total.cases` is distributed normally. Thus, what remains to shown is that the following three assumptions are reasonable:

### 1) Linearity.

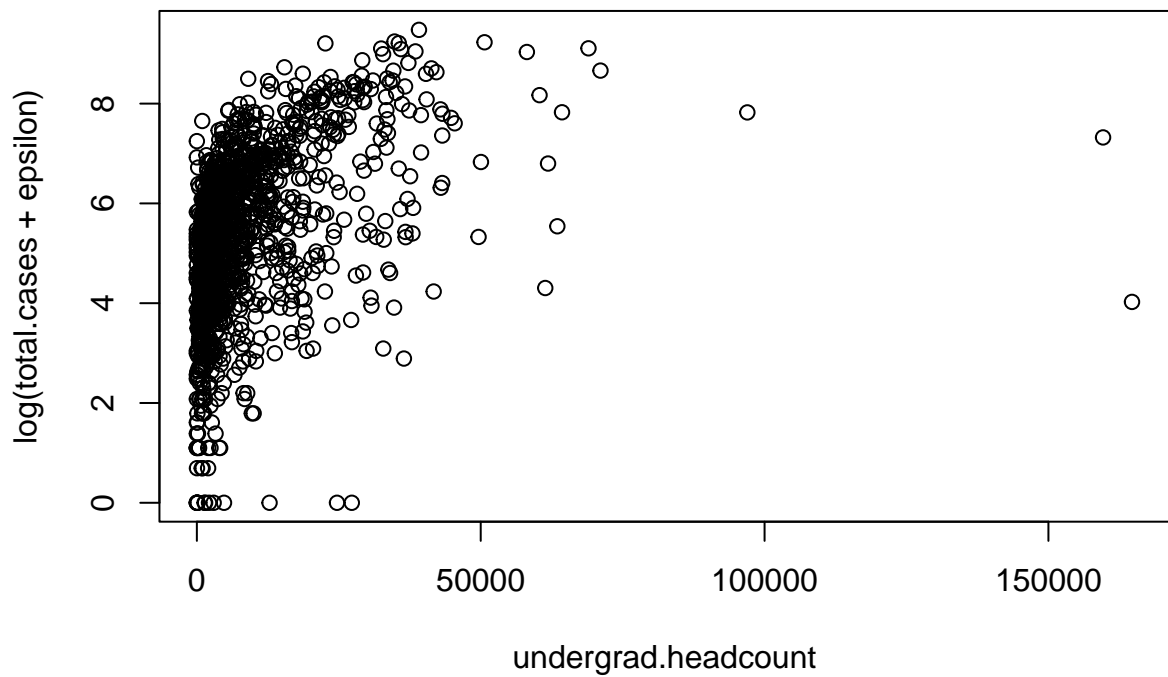
```
# par(mfrow=c(2,2))
plot(log(total.cases + epsilon) ~ tuition, data=full.cases)
```



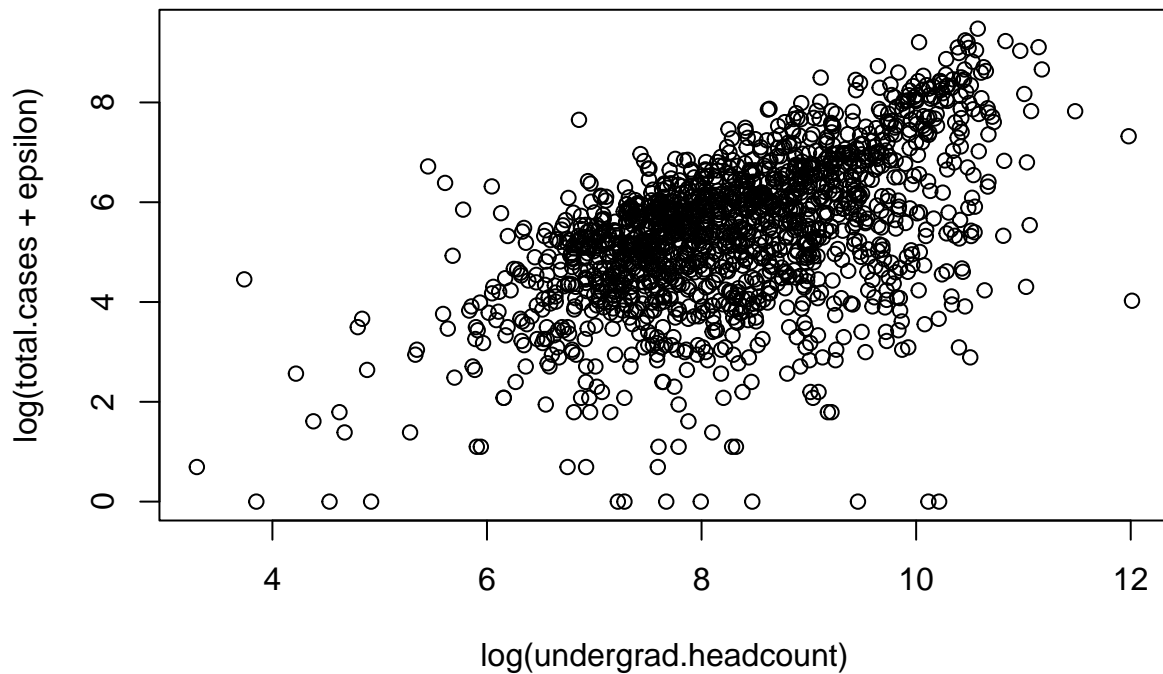
```
plot(log(total.cases + epsilon) ~ total.headcount, data=full.cases)
```



```
plot(log(total.cases + epsilon) ~ undergrad.headcount, data=full.cases)
```

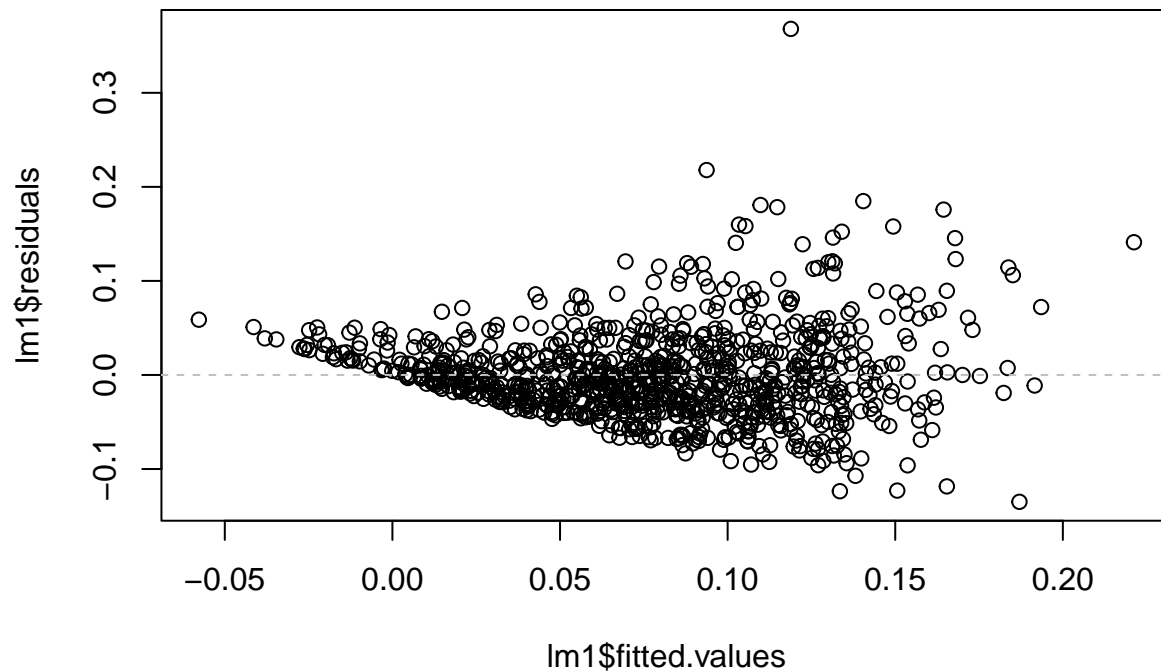


```
plot(log(total.cases + epsilon) ~ log(undergrad.headcount), data=full.cases)
```



What about these individual cases?

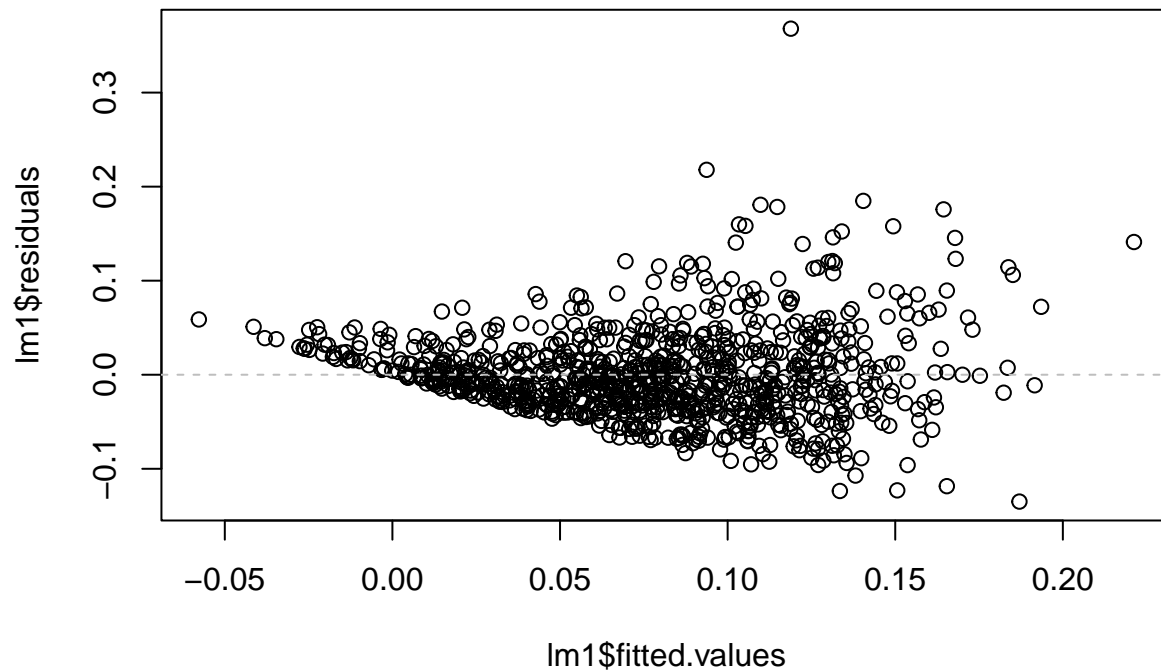
```
plot(lm1$residuals ~ lm1$fitted.values)
abline(h=0, col="gray", lty=2)
```



Based on above plot, no other clear pattern revealed — nothing obviously flawed with this assumption, so we continue forward.

## 2) Constant Variance.

```
plot(lm1$residuals ~ lm1$fitted.values)
abline(h=0, col="gray", lty=2)
```



## Sequential Variable Selection Models

```
df = full.cases[,c("percent.cases", "state", "mask.mandated.days", "private", "tuition", "total.headcount")]
df = na.omit(df)

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

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

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.campus.housing

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 Selection**

**Mixed-Effect Models**

**Conclusions**