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

```
RMSE <- function(y, yhat) {
  SSE = sum((y-yhat)^2)
  return(sqrt(SSE/length(y)))
}
```

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 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 (Daniel)

Explain motivation for group testing

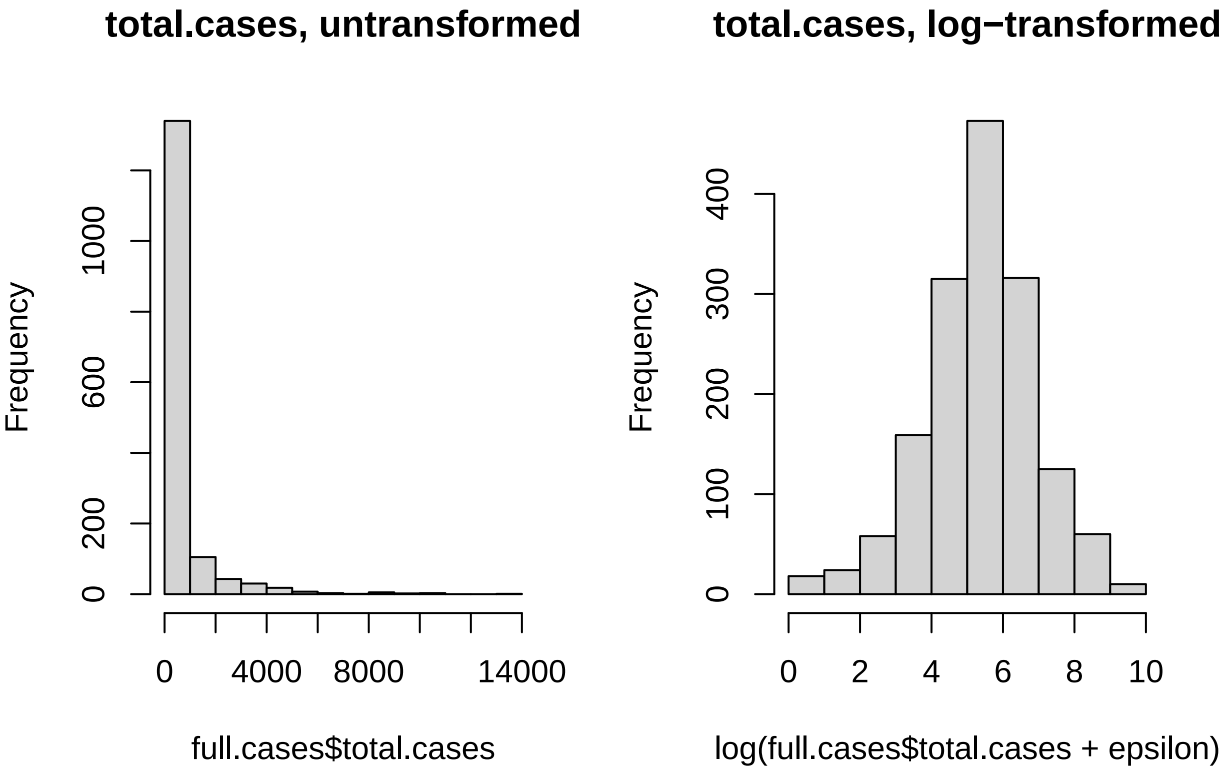
Checking the assumptions for t-based methods

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

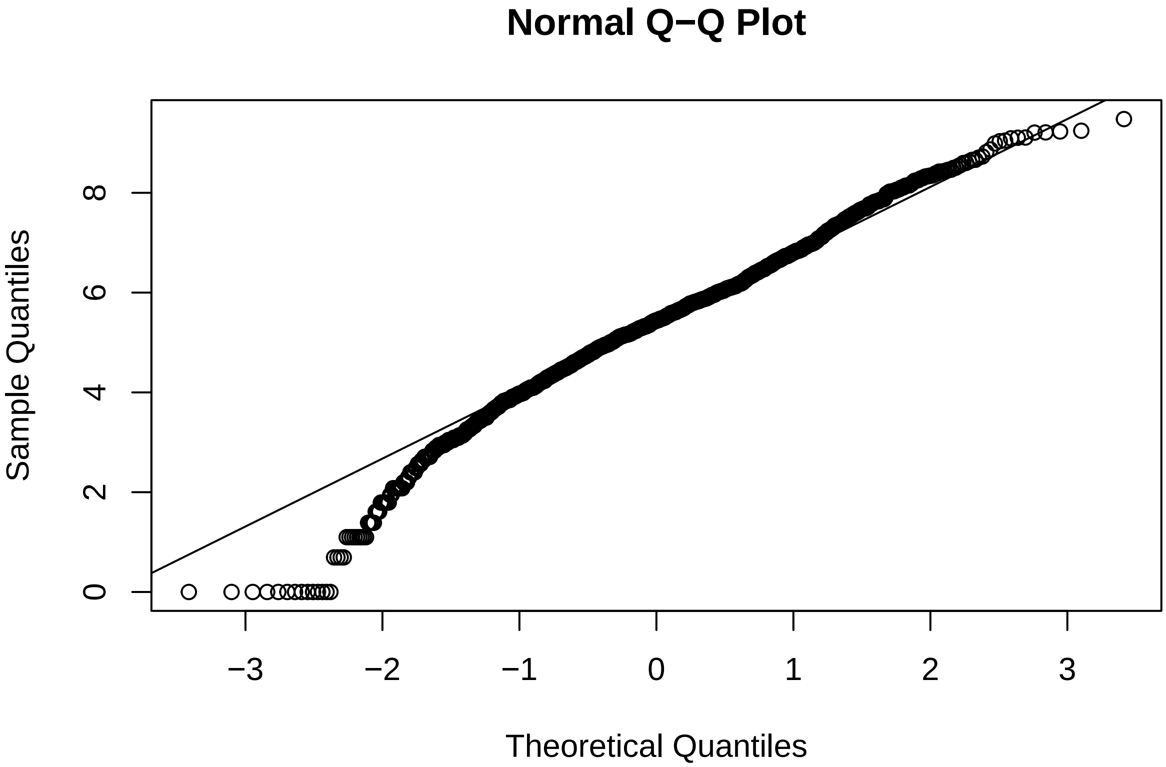
1) **Observations are independent.** Explain why reasonable.
2) **Groups are independent of one another.** Explain why reasonable.
3) **Observations are normally distributed.**

```r
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

State motivations and hypotheses for the next test

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

```
##
##  Welch Two Sample t-test
##
## data:  log(total.cases + epsilon) by catholic
## t = -1.2478, df = 179.01, p-value = 0.2137
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.37880584  0.08532016
## sample estimates:
##  mean in group No mean in group Yes
##          5.348917          5.495659
```

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$total.cases[full.cases$re
## 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

```
# 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
lm1 <- lm(log(total.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)

# No idea why the below does not work
# lm1 <- lm(log(total.cases + epsilon) ~ . - (avg.grant.money + dorm.capacity + dorm.room.price + FIPS.
# bad <- c("avg.grant.money", "dorm.capacity", "dorm.room.price", )
summary(lm1)
```

```
##
## Call:
## lm(formula = log(total.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
## -4.6176 -0.4187  0.0529  0.5416  2.4679
##
## Coefficients:
```

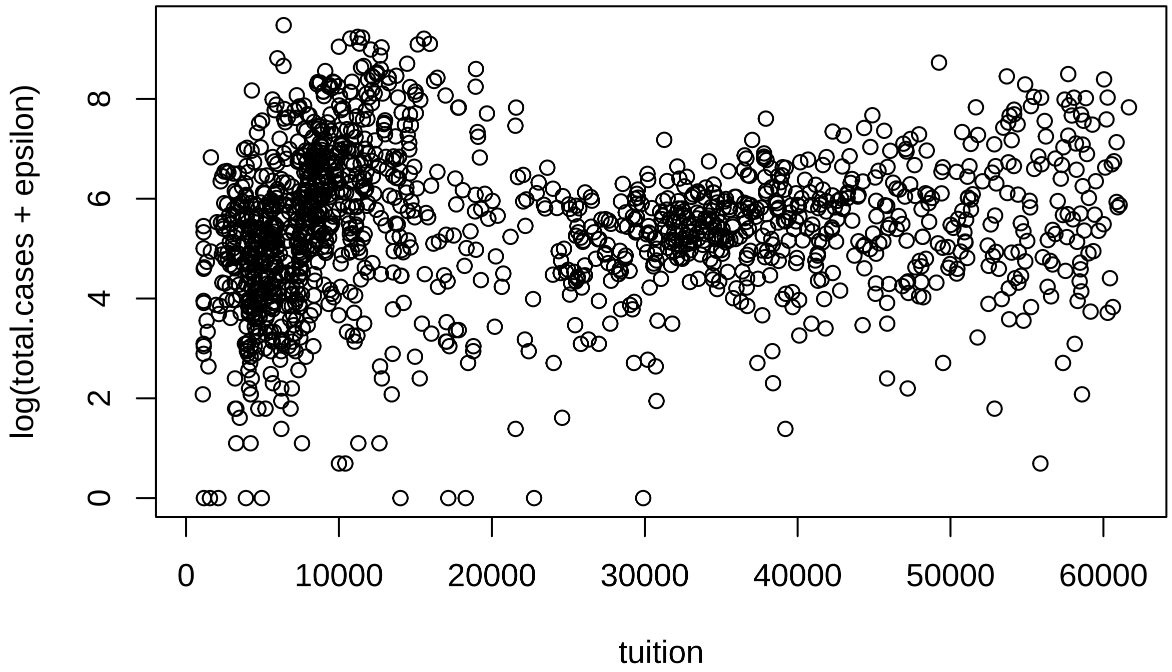
```
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               3.133e+00  3.962e+00   0.791 0.429309    
## religiousYes              1.002e-01  1.081e-01   0.927 0.354151    
## catholicYes               2.178e-01  1.190e-01   1.831 0.067498 .  
## tuition                   1.391e-05  5.128e-06   2.713 0.006799 ** 
## total.headcount           8.060e-05  1.260e-05   6.395 2.61e-10 ***
## undergrad.headcount      -1.640e-05  1.595e-05  -1.028 0.304336    
## percent.american.native   8.186e-03  4.440e-02   0.184 0.853778    
## percent.asian            -2.867e-02  4.056e-02  -0.707 0.479788    
## percent.black            -3.197e-03  3.956e-02  -0.081 0.935608    
## percent.hispanic.latino  -1.106e-02  3.971e-02  -0.279 0.780677    
## percent.pacific.islander -5.576e-02  1.235e-01  -0.452 0.651714    
## percent.white             9.726e-03  3.950e-02   0.246 0.805571    
## percent.two.more.races   -5.312e-02  4.318e-02  -1.230 0.218910    
## percent.NA.race          -1.346e-02  3.929e-02  -0.342 0.732064    
## percent.nonres.alien      2.927e-03  3.996e-02   0.073 0.941623    
## percent.women            -6.029e-03  2.912e-03  -2.070 0.038722 *  
## grad.rate                 1.983e-02  3.177e-03   6.241 6.76e-10 ***
## percent.fin.aid           6.433e-03  3.256e-03   1.976 0.048520 *  
## percent.disability       -1.801e-02  6.145e-03  -2.930 0.003476 ** 
## on.campus.housingYes      7.649e-01  1.323e-01   5.780 1.04e-08 ***
## stateAlaska              -1.898e+00  1.016e+00  -1.869 0.061957 .  
## stateArizona             -5.295e-01  5.548e-01  -0.954 0.340148    
## stateArkansas             2.640e-01  3.913e-01   0.675 0.500063    
## stateCalifornia          -7.421e-01  3.509e-01  -2.115 0.034704 *  
## stateColorado             3.229e-01  3.843e-01   0.840 0.400975    
## stateConnecticut          2.738e-01  3.638e-01   0.752 0.451988    
## stateDelaware             3.939e-01  7.099e-01   0.555 0.579176    
## stateFlorida             -1.320e-01  3.339e-01  -0.395 0.692667    
## stateGeorgia              2.658e-01  3.517e-01   0.756 0.450069    
## stateHawaii               4.020e-01  1.066e+00   0.377 0.706064    
## stateIdaho                1.538e-02  5.480e-01   0.028 0.977623    
## stateIllinois            -3.830e-02  3.252e-01  -0.118 0.906263    
## stateIndiana             -3.350e-01  3.500e-01  -0.957 0.338677    
## stateIowa                -2.848e-01  3.714e-01  -0.767 0.443391    
## stateKansas              -3.070e-01  4.945e-01  -0.621 0.534805    
## stateKentucky            -3.847e-01  3.392e-01  -1.134 0.257100    
## stateLouisiana           -2.735e-02  3.913e-01  -0.070 0.944291    
## stateMaine               -6.332e-01  4.325e-01  -1.464 0.143496    
## stateMaryland            -4.984e-01  3.759e-01  -1.326 0.185194    
## stateMassachusetts       -6.213e-01  3.172e-01  -1.959 0.050471 .  
## stateMichigan            -1.098e-01  3.262e-01  -0.337 0.736467    
## stateMinnesota            1.692e-01  3.244e-01   0.522 0.602009    
## stateMississippi         -1.326e-01  4.935e-01  -0.269 0.788159    
## stateMissouri            -1.317e-01  3.512e-01  -0.375 0.707794    
## stateMontana              2.727e-01  6.022e-01   0.453 0.650827    
## stateNebraska             1.595e-01  4.426e-01   0.360 0.718699    
## stateNevada               1.754e-01  5.519e-01   0.318 0.750730    
## stateNew Hampshire       -3.253e-02  4.194e-01  -0.078 0.938195    
## stateNew Jersey          -5.277e-02  3.435e-01  -0.154 0.877955    
## stateNew Mexico          -8.613e-01  6.965e-01  -1.237 0.216542    
## stateNew York            -3.734e-01  2.949e-01  -1.266 0.205824    
## stateNorth Carolina       9.092e-02  3.276e-01   0.278 0.781412    
## stateNorth Dakota        -4.675e-01  5.976e-01  -0.782 0.434319    
## stateOhio                 -6.914e-01  3.118e-01  -2.218 0.026830 *
## stateOklahoma              1.206e-01  6.124e-01   0.197 0.843902
## stateOregon               -9.179e-01  3.921e-01  -2.341 0.019464 *
## statePennsylvania         -4.055e-01  2.999e-01  -1.352 0.176772
## stateRhode Island          3.280e-01  4.310e-01   0.761 0.446800
## stateSouth Carolina        4.074e-01  3.538e-01   1.151 0.249864
## stateSouth Dakota         -6.221e-02  9.500e-01  -0.065 0.947802
## stateTennessee             3.044e-02  3.639e-01   0.084 0.933361
## stateTexas                 1.388e-01  3.398e-01   0.408 0.683146
## stateUtah                  1.493e-01  5.356e-01   0.279 0.780518
## stateVermont              -1.146e+00  4.562e-01  -2.512 0.012168 *
## stateVirginia             -6.681e-02  3.340e-01  -0.200 0.841517
## stateWashington           -8.494e-01  3.541e-01  -2.399 0.016658 *
## stateWashington, D.C.     -2.518e-01  4.981e-01  -0.506 0.613307
## stateWest Virginia        -1.073e-01  4.277e-01  -0.251 0.801919
## stateWisconsin            -3.005e-01  3.391e-01  -0.886 0.375686
## stateWyoming              -2.637e-01  7.042e-01  -0.375 0.708093
## privateYes                -6.426e-01  1.750e-01  -3.671 0.000256 ***
## percent.student.loan       8.729e-04  2.285e-03   0.382 0.702564
## occupational.degreeYes     9.142e-02  8.860e-02   1.032 0.302428
## hs.equivalent.degreeYes    1.197e-01  1.240e-01   0.965 0.334832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9042 on 873 degrees of freedom
##   (908 observations deleted due to missingness)
## Multiple R-squared:  0.6492, Adjusted R-squared:  0.6199
## F-statistic: 22.13 on 73 and 873 DF,  p-value: < 2.2e-16
```

Checking the Assumptions of Linear Regression

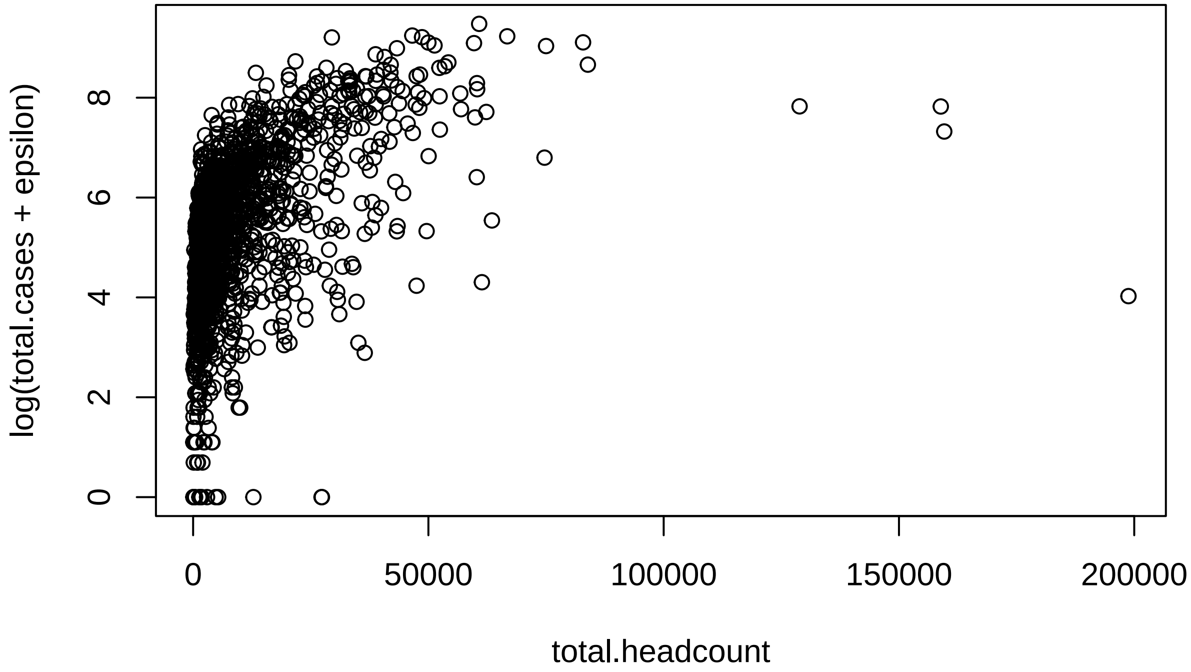
As discussed above in the group testing section, we believe that it is reasonable to assume that the observations in this data set are independent of one another, and that the log-transformed response variable `total.cases` is distributed normally. Thus, what remains to shown is that the following three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

1) **Linearity.**

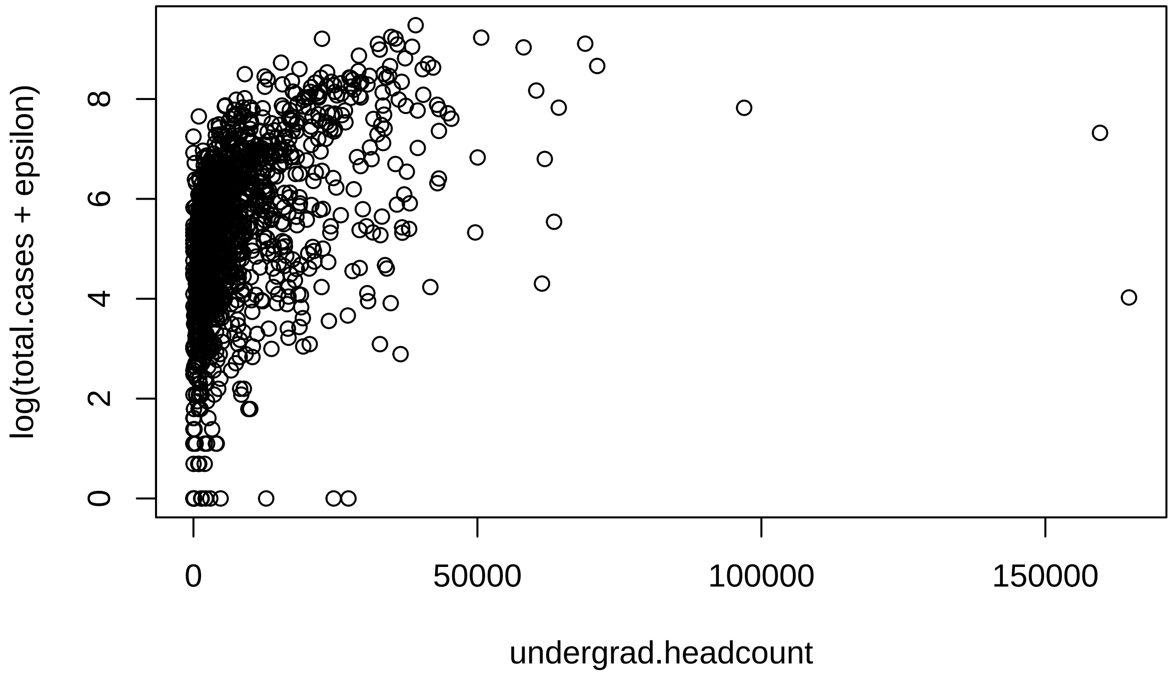
```
# par(mfrow=c(2,2))
plot(log(total.cases + epsilon) ~ tuition, data=full.cases)
```



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

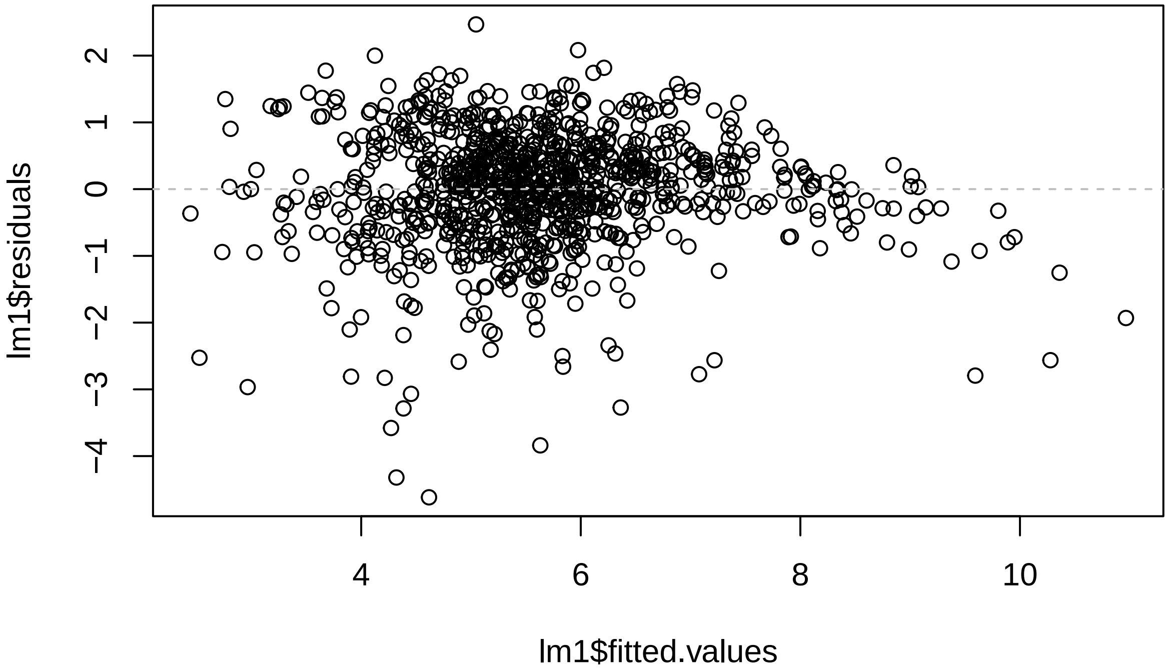


```
plot(log(total.cases + epsilon) ~ undergrad.headcount, data=full.cases)
```



What about these individual cases?

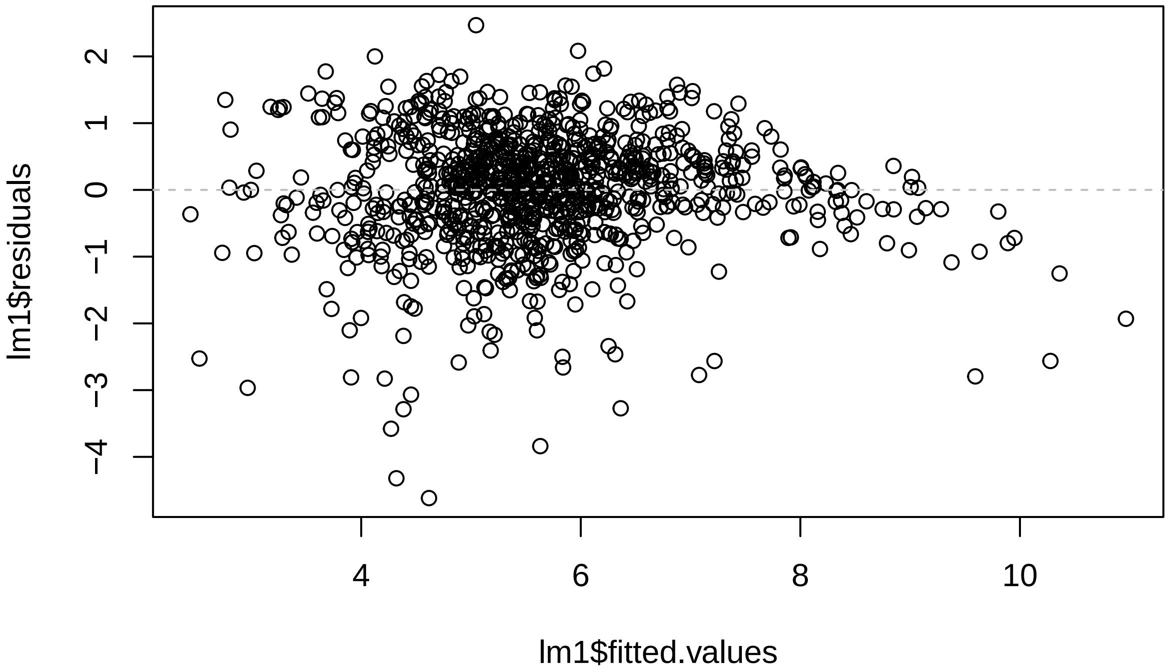
```
plot(lm1$residuals ~ lm1$fitted.values)
abline(h=0, col="gray", lty=2)
```



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

Penalized Regression Models (Ridge and LASSO)

Non-Parametric Models (Decision Trees, Random Forests)

Decision Tree

justify control parameters

```
tree1 <- rpart(log(total.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, control=list(minsplit=1, cp=0, maxdepth=20))
tree1$variable.importance
```

```
##          total.headcount                    state      undergrad.headcount
##              1255.613373              1042.732197              1030.869546
##        on.campus.housing                  tuition                grad.rate
##               653.633274               505.107583               466.418024
##     percent.student.loan            percent.white            percent.asian
##               327.292357               286.261336               266.836745
##  percent.hispanic.latino     percent.nonres.alien            percent.women
##               195.490742               181.034061               177.014518
##          percent.NA.race     hs.equivalent.degree            percent.black
##               106.017737               104.905026               104.668977
##          percent.fin.aid percent.pacific.islander                  private
##                98.304111                76.618318                65.980777
##    percent.two.more.races      occupational.degree  percent.american.native
##                54.413318                44.214190                35.146096
##                religious                 catholic       percent.disability
##                20.085109                11.946021                 3.857854
```

fit a well-pruned decision tree

```
tree2 <- prune(tree1, tree1$cptable[,"CP"][which.min(tree1$cptable[,"xerror"])])
```

Bagged Model

```
mtry <- 39
maxnodes.bag <- c(50,100,200,500)
ntree <- 200
train.RMSE <- matrix(nrow=length(maxnodes.bag), ncol=1)

selected.columns <- c("")

for (i in 1:length(maxnodes.bag)) {
  rf <- randomForest(log(total.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,
                     maxnodes=maxnodes.bag,
                     ntree=ntree, mtry=length(colnames(full.cases)) - 1)
  train.RMSE[i,1] <- RMSE(log(full.cases$total.cases + epsilon), rf$predicted)
}

best.index.bag <- which.min(train.RMSE[,1])
best.index.bag
```

Mixed-Effect Models

Conclusions