# Final Project: Data Ethics and Policy

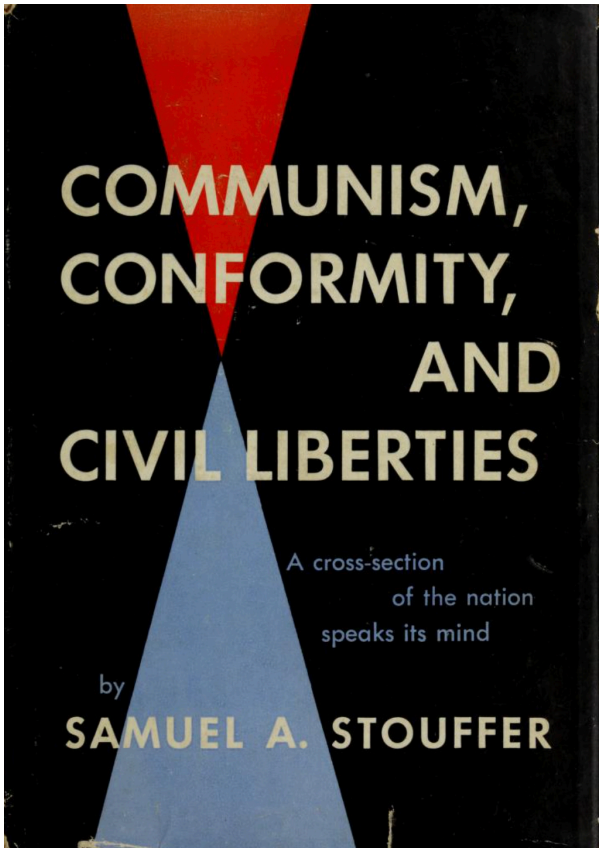
Christy Hsu

## Abstract

This project looked into an attempt in the fifties that set out to measure individual tolerance and use that measure to evaluate the impact the anti-Communism Cold War agenda—both abroad and at home—had on U.S. citizens. I saw the picking up of this 70-year-old survey data, and the analysis that accompanied it, as a chance to stroll through the data ethics concerns that can arise when holding the hope to learn about the world from data

## Introduction

The online poll data library host by Roper Center was where I first came across the Stouffer Study of 1954 *Communism, Conformity, and Civil Liberties: A Cross-Section of the Nation Speaks Its Mind* by Samuel Andrew Stouffer(**?**)And, I gain access to the entire dataset on ICPSR(**?**)




I found Stouffer's attempt in the fifties to design public opion polls and, construct an innovative way of measuring the latent properties tolerance and fear in at the individual level very interesting. The tolerance scale and the perception of internal communist danger scale are not included in the data, thus a major part of this project involved returning these two target variables in order to complete the picture and reproduce that basis on which Stouffer built his arguments. I learned much from this practice of reading and researching that past effort into conceptualizing and operationalizing tolerance and fear, and got my hands dirty to really apply those framings and methods to the data. It gave me a chance to reflect on the many artistic and arbitrary decisions that the researcher made throughout this data analysis process.

Source: Article Notebook

## Harvesting from Historical Data Collection Efforts: A More Friendly Format

Complying ICPSR's redistribution policy, the converted data files are not provided here. Instead, the author provides STATA .do and .dct files, which were constructed based on the reading of the codebook. Please down the the dataset in ASCII format from ICPSR and should be able to apply to decode the .txt files from both samples.(**?**)

```
read-ascii-files
    gp-decode.do
    lead-decode.do
    sample1.dct
    sample2.dct
```

### Returning target variables to the data

The tolerance that Stouffer argued upon.(**?**)

#### Preparing the `code_df` data frame

1. Cleaning column names and binding the two samples

Source: Article Notebook

Source: Article Notebook

#### Adding Binary and Ternary Variables `leader`,`interested` and `categorizer`: these are the variables that Stouffer specified in his book as ways to divide the respondents and make comparisons.

| leader | n |
|---|---|
| 0 | 4933 |
| 1 | 1500 |

Source: Article Notebook

| leader | interested | n | pct |
|---|---|---|---|
| 0 | less | 2155 | 43.68538 |
| 0 | more | 2778 | 56.31462 |
| 1 | less | 210 | 14.00000 |
| 1 | more | 1290 | 86.00000 |

Source: Article Notebook

Source: Article Notebook

| leader | categorizer | n | pct |
|---|---|---|---|
| 0 | agree | 3127 | 63.389418 |
| 0 | disagree | 1495 | 30.306102 |
| 0 | dont know | 311 | 6.304480 |
| 1 | agree | 806 | 53.733333 |
| 1 | disagree | 669 | 44.600000 |
| 1 | dont know | 25 | 1.666667 |

Source: Article Notebook

## Scale1: Willingness to Tolerate Nonconformist

### Conceptual Tolerance and Operational Tolerance

The questionnaires used to rank respondents into six tolerance groups focused on four types of nonconformists:

- A person who is against all churches and religion (atheist)
- A person who favors government ownership (socialist)
- An alleged communist (someone whose loyalty has been questioned by a Congressional Committee but swears under oath they have never been a communist)
- An admitted communist

respondent were asked about their approval of 3 types of disposition against the above nonconfromist, and whether they agree or disagree the limitation or deprivation of the nonconformist's civil liberties, for example:

1. Freedom speech:
    - "If _____ wants to make a speech in your community, should he be allowed to speak or not?"
2. Book censor:
    - "Suppose he wrote a book that is in your public library. Somebody in your community suggests the book should be removed. Would you favor removing it, or not?"
3. Employment:
    - Should a radio singer who is a nonconformist be fired or not?
    - Should a college or university teacher be fired or not?
    - Should a high school teacher be fired or not?

- Should someone working in a defense plant be fired or not?
- Should a store clerk be fired or not?

4. Boycott:

    - "Suppose the radio program he is on advertises a brand of soap. Somebody in your community suggests you stop buying that soap. Would you stop or not?"

### 0 to 5: Scaling Individual Tolerance

I ran into many challenges replicating Stouffer's results. Both the overall proportions across tolerance rankings, was unable to reproduce the group counts applying further breakdowns, such as by age, region, education, thus for comparison.

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

| leader | tolerance_group | n | pct |
|---|---|---|---|
| 0 | tolerance0 | 651 | 13.196838 |
| 0 | tolerance1 | 871 | 17.656598 |
| 0 | tolerance2 | 1263 | 25.603081 |
| 0 | tolerance3 | 951 | 19.278330 |
| 0 | tolerance4 | 395 | 8.007298 |
| 0 | tolerance5 | 802 | 16.257855 |
| 1 | tolerance0 | 79 | 5.266667 |
| 1 | tolerance1 | 182 | 12.133333 |
| 1 | tolerance2 | 180 | 12.000000 |
| 1 | tolerance3 | 246 | 16.400000 |
| 1 | tolerance4 | 205 | 13.666667 |
| 1 | tolerance5 | 608 | 40.533333 |

Source: Article Notebook

| tolerance_group | n |
|---|---|
| tolerance0 | 730 |
| tolerance1 | 1053 |
| tolerance2 | 1443 |

| tolerance_group | n |
|---|---|
| tolerance3 | 1197 |
| tolerance4 | 600 |
| tolerance5 | 1410 |

Source: Article Notebook

### Broader Tolerance Rank Groups: `less tolerant`, `in-between` and `more tolerant`

Source: Article Notebook

| tolerance_broader0 | n | pct |
|---|---|---|
| more tolerant | 813 | 54.2 |
| in between | 426 | 28.4 |
| less tolerant | 261 | 17.4 |

Source: Article Notebook

### Attempt2

Allowing some inconsistency?

Source: Article Notebook

| tolerance | n |
|---|---|
| tolerance0 | 674 |
| tolerance1 | 530 |
| tolerance2 | 1334 |
| tolerance3 | 1229 |
| tolerance4 | 807 |
| tolerance5 | 1859 |

Source: Article Notebook

| leader | tolerance | n | pct |
|---|---|---|---|
| 0 | tolerance0 | 604 | 12.244070 |
| 0 | tolerance1 | 491 | 9.953375 |
| 0 | tolerance2 | 1179 | 23.900264 |

| leader | tolerance | n | pct |
|---|---|---|---|
| 0 | tolerance3 | 993 | 20.129738 |
| 0 | tolerance4 | 554 | 11.230488 |
| 0 | tolerance5 | 1112 | 22.542064 |
| 1 | tolerance0 | 70 | 4.666667 |
| 1 | tolerance1 | 39 | 2.600000 |
| 1 | tolerance2 | 155 | 10.333333 |
| 1 | tolerance3 | 236 | 15.733333 |
| 1 | tolerance4 | 253 | 16.866667 |
| 1 | tolerance5 | 747 | 49.800000 |

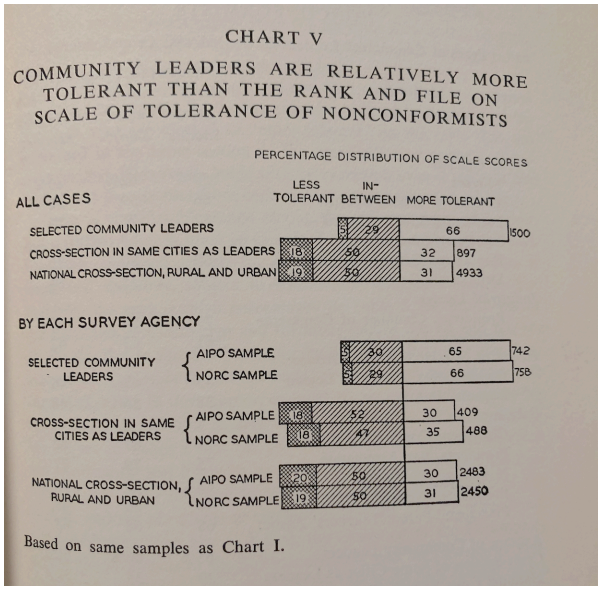
Source: Article Notebook

Source: Article Notebook

| leader | tolerance_broader | n | pct |
|---|---|---|---|
| 0 | more tolerant | 1666 | 33.772552 |
| 0 | in between | 2172 | 44.030002 |
| 0 | less tolerant | 1095 | 22.197446 |
| 1 | more tolerant | 1000 | 66.666667 |
| 1 | in between | 391 | 26.066667 |
| 1 | less tolerant | 109 | 7.266667 |

Source: Article Notebook

To answer this question(**?**)




### us region

Source: Article Notebook

| us_region | tolerance_broader | n | pct |
|---|---|---|---|
| east | more tolerant | 276 | 68.148148 |
| east | in between | 107 | 26.419753 |
| east | less tolerant | 22 | 5.432099 |
| midwest | more tolerant | 348 | 69.322709 |
| midwest | in between | 123 | 24.501992 |
| midwest | less tolerant | 31 | 6.175299 |
| south | more tolerant | 225 | 57.251908 |
| south | in between | 123 | 31.297710 |
| south | less tolerant | 45 | 11.450382 |
| west | more tolerant | 151 | 75.500000 |
| west | in between | 38 | 19.000000 |
| west | less tolerant | 11 | 5.500000 |

Source: Article Notebook

## Scale2: Scale of the Perception on the Internal Communist Danger

Source: Article Notebook

Source: Article Notebook

| danger | n |
|---|---|
| danger0 | 507 |
| danger1 | 682 |
| danger2 | 2116 |
| danger3 | 1170 |
| danger4 | 1093 |
| danger5 | 865 |

Source: Article Notebook

### Broader rank groups

Source: Article Notebook

| danger_broader | n |
|---|---|
| great threat | 1958 |
| in between | 3286 |
| little threat | 1189 |

Source: Article Notebook

Source: Article Notebook

## Evaluating Operationallizations: Reliability and Validity, insights from classfication algorithms

### Data and Measures: Validity

From the conceptual variable tolerance to the operationalized definition of tolerance, Stouffer proposed *h-technique* to map answers of the respondent to a tolerance score corresponding to their degree of tolerance.(**?**) But does this tolerance scale really measuring people's tolerance or is it measuring something else?(**?**)

### Reliability

Source: Article Notebook

### 3-class classification using the strict measure of tolerance score `tolerance_broader0`

Source: Article Notebook

Source: Article Notebook

| tolerance_broader0 | n |
|---|---|
| in between | 426 |
| less tolerant | 261 |
| more tolerant | 813 |

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

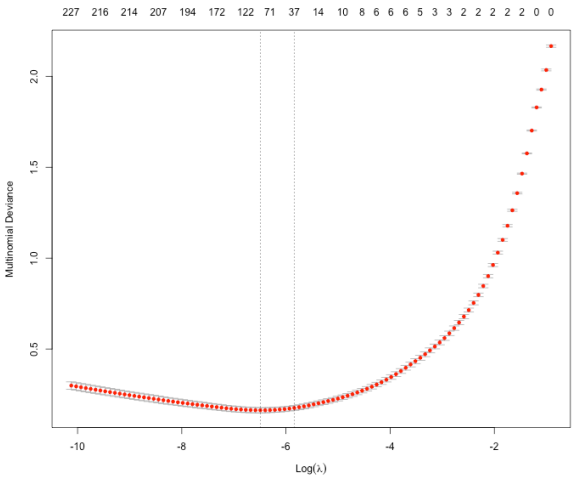
Source: Article Notebook

Source: Article Notebook

```
clf0 <- cv.glmnet(
  x_train, y_train, family = "multinomial",
  alpha = 1, type.multinomial = "ungrouped"
)

lambda_1se0 <- clf0$lambda.1se
lambda_1se0
# 0.002905948

png("clf0-plot.png", width = 700, height = 600)
plot(clf0)
dev.off()
```




Source: Article Notebook

```
[1] 0.9902837
```

Source: Article Notebook

```
[1] 0.989899
```

Source: Article Notebook

```
Confusion Matrix and Statistics

               Reference
Prediction      more tolerant in between less tolerant
  more tolerant           400          4             3
  in between                0        511             3
  less tolerant             1          2           363

Overall Statistics

               Accuracy : 0.9899
                 95% CI : (0.9828, 0.9946)
    No Information Rate : 0.4017
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9847
```

```
 Mcnemar's Test P-Value : 0.1577

Statistics by Class:

                     Class: more tolerant Class: in between
Sensitivity                        0.9975            0.9884
Specificity                        0.9921            0.9961
Pos Pred Value                     0.9828            0.9942
Neg Pred Value                     0.9989            0.9922
Prevalence                         0.3116            0.4017
Detection Rate                     0.3108            0.3970
Detection Prevalence               0.3162            0.3994
Balanced Accuracy                  0.9948            0.9922
                     Class: less tolerant
Sensitivity                        0.9837
Specificity                        0.9967
Pos Pred Value                     0.9918
Neg Pred Value                     0.9935
Prevalence                         0.2867
Detection Rate                     0.2821
Detection Prevalence               0.2844
Balanced Accuracy                  0.9902
```

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

| predictor | n |
|---|---|
| (Intercept) | 3 |
| admitted_communist_allow_speak5 | 2 |
| admitted_communist_buy_soap5 | 2 |
| admitted_communist_put_jail5 | 2 |
| alledged_communist_fire_teacher5 | 2 |
| censor_antireligion_book5 | 2 |
| censor_antireligion_speaker5 | 2 |
| censor_antireligion_speaker8 | 2 |
| censor_socialist_book5 | 2 |
| censor_socialist_speaker8 | 2 |
| many_communists_defense_plants8 | 2 |
| many_communists_us5 | 2 |

| predictor | n |
|---|---|

Source: Article Notebook

Source: Article Notebook

### 3-class classification: `tolerance_broader` based on the measure with wiggle room

Source: Article Notebook

Source: Article Notebook

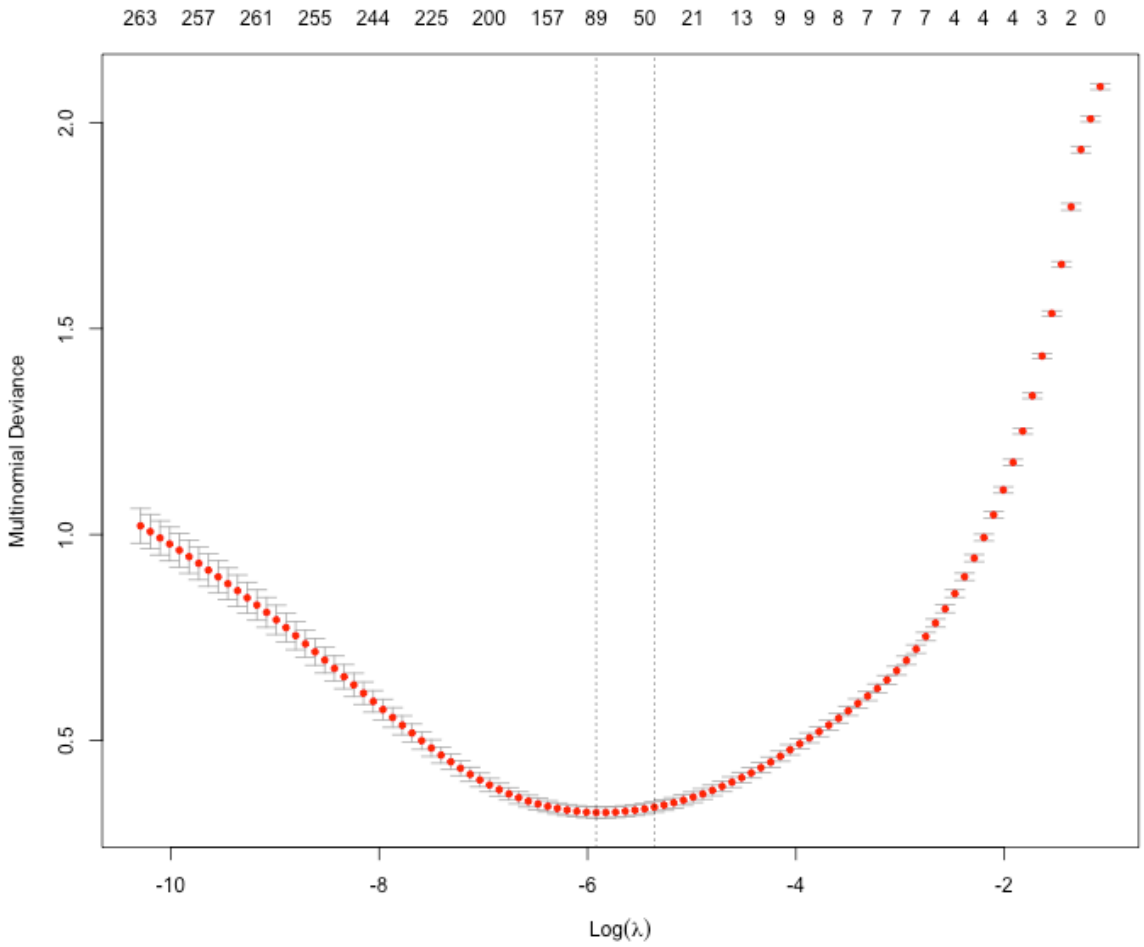
Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

```
clf <- cv.glmnet(
  x_train, y_train,family = "multinomial",
  alpha = 1, type.multinomial = "ungrouped"
)
# png("clf-plot.png", width = 700, height = 600)
# plot(clf)
# dev.off()
lambda_best <- clf$lambda.1se
# 0.00391
lambda_seq <- clf$lambda
# saveRDS(lambda_seq, file = "data/clf-lambda-seq.rds")
```




Source: Article Notebook

```
[1] 0.958803
```

Source: Article Notebook

```
[1] 0.950272
```

Source: Article Notebook

```
Confusion Matrix and Statistics
```

```
              Reference
Prediction      more tolerant in between less tolerant
  more tolerant           505          4             5
  in between                9        489            12
  less tolerant             7         27           229

Overall Statistics

               Accuracy : 0.9503
                 95% CI : (0.9369, 0.9615)
    No Information Rate : 0.4048
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9222

 Mcnemar's Test P-Value : 0.04548

Statistics by Class:

                     Class: more tolerant Class: in between
Sensitivity                        0.9693            0.9404
Specificity                        0.9883            0.9726
Pos Pred Value                     0.9825            0.9588
Neg Pred Value                     0.9793            0.9601
Prevalence                         0.4048            0.4040
Detection Rate                     0.3924            0.3800
Detection Prevalence               0.3994            0.3963
Balanced Accuracy                  0.9788            0.9565
                     Class: less tolerant
Sensitivity                        0.9309
Specificity                        0.9673
Pos Pred Value                     0.8707
Neg Pred Value                     0.9834
Prevalence                         0.1911
Detection Rate                     0.1779
Detection Prevalence               0.2044
Balanced Accuracy                  0.9491
```

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

| predictor | n |
|---|---|
| (Intercept) | 3 |
| admitted_communist_allow_speak5 | 2 |
| admitted_communist_buy_soap5 | 2 |
| admitted_communist_put_jail5 | 2 |
| admitted_communist_remove_book_library5 | 2 |
| alledged_communist_fire_clerk5 | 2 |
| alledged_communist_fire_clerk8 | 2 |
| alledged_communist_fire_defense_plant5 | 2 |
| alledged_communist_fire_high_school_teacher5 | 2 |
| alledged_communist_fire_singer5 | 2 |
| alledged_communist_fire_teacher5 | 2 |
| alledged_communist_make_community_speech5 | 2 |

Source: Article Notebook

### Are these selected Items capturing most of the variances?

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

### Learning: Predicting Tolerance Score without the original 15 items

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

| tolerance_broader0 | n |
|---|---|
| in between | 426 |
| less tolerant | 261 |
| more tolerant | 813 |

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

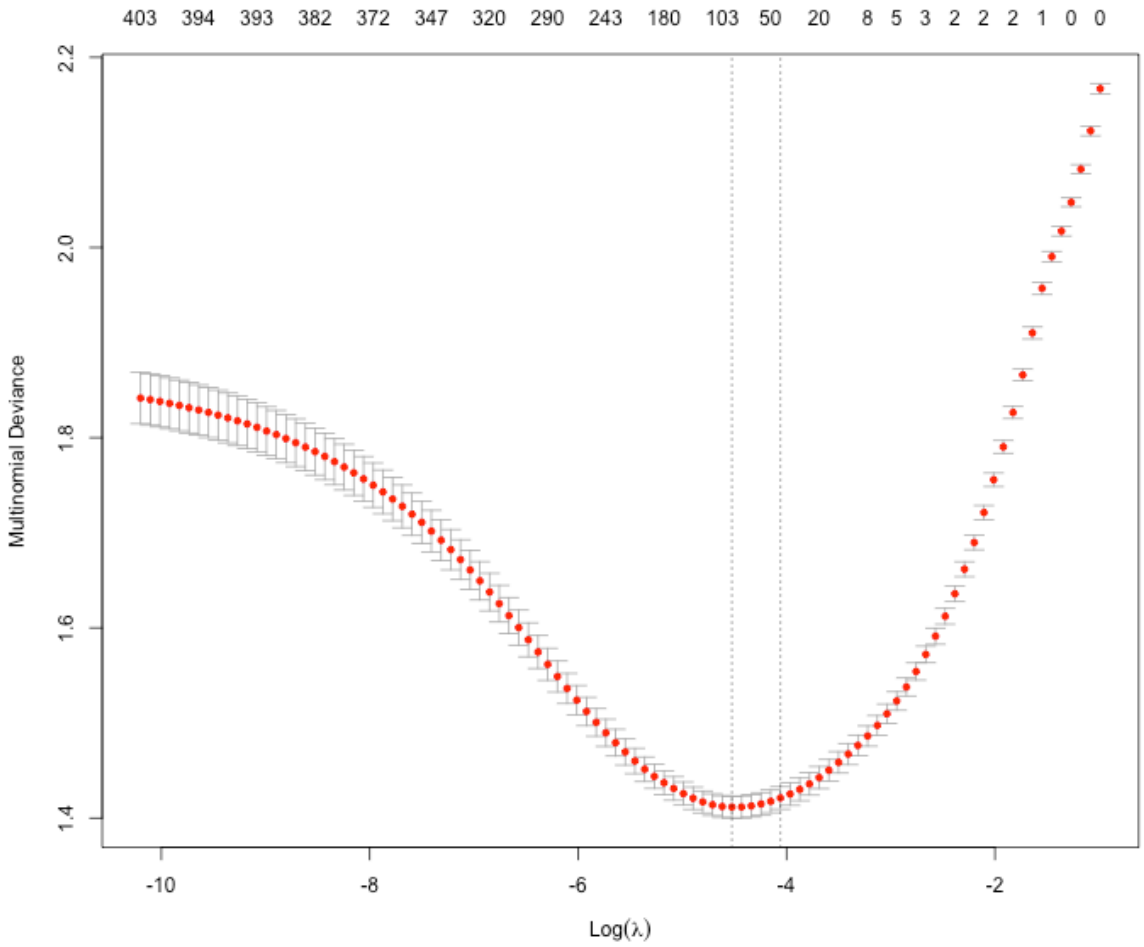
Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

```r
clf3 <- cv.glmnet(
  x_train, y_train, family = "multinomial",
  alpha = 0.75, type.multinomial = "ungrouped"
)
lambda_1se3 <- clf3$lambda.1se
lambda_1se3
# 0.01726083
lambda_seq3 <- clf3$lambda
# saveRDS(lambda_seq3, file = "data/clf3-lambda-seq.rds")

# png("image/clf3-plot.png", width = 700, height = 600)
# plot(clf3)
# dev.off()
```




Source: Article Notebook

```
[1] 0.7188107
```

Source: Article Notebook

```
[1] 0.6985237
```

Source: Article Notebook

```
Confusion Matrix and Statistics
```

```
              Reference
Prediction      more tolerant in between less tolerant
  more tolerant           264         65            39
  in between              125        435           130
  less tolerant            12         17           200

Overall Statistics

               Accuracy : 0.6985
                 95% CI : (0.6726, 0.7235)
    No Information Rate : 0.4017
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5322

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: more tolerant Class: in between
Sensitivity                        0.6584            0.8414
Specificity                        0.8826            0.6688
Pos Pred Value                     0.7174            0.6304
Neg Pred Value                     0.8509            0.8626
Prevalence                         0.3116            0.4017
Detection Rate                     0.2051            0.3380
Detection Prevalence               0.2859            0.5361
Balanced Accuracy                  0.7705            0.7551
                     Class: less tolerant
Sensitivity                        0.5420
Specificity                        0.9684
Pos Pred Value                     0.8734
Neg Pred Value                     0.8403
Prevalence                         0.2867
Detection Rate                     0.1554
Detection Prevalence               0.1779
Balanced Accuracy                  0.7552
```

Source: Article Notebook

### Acquired the class specific variables and their coefficients

Source: Article Notebook

Source: Article Notebook

| predictor | n |
|---|---|
| (Intercept) | 3 |
| admitted_communist_fire_clerk5 | 2 |
| admitted_communist_lose_citizenship5 | 2 |
| break_friendship_former_communist5 | 2 |
| censor_antireligion_teacher5 | 2 |
| child_rearing_respect3 | 2 |
| last_grade_finished_school5 | 2 |
| remember_hiss_caught8 | 2 |
| admitted_communist_fire_college_professor5 | 1 |
| admitted_communist_lose_citizenship8 | 1 |
| alledged_communist_buy_soap5 | 1 |
| alledged_communist_fire_singer5 | 1 |

Source: Article Notebook

### Worrying about Internal Communist Threats

```
    Pearson's Chi-squared test

data:  toler_rigid_tb
X-squared = 491.09, df = 10, p-value < 2.2e-16
```

Source: Article Notebook

```
                agree   disagree  dont know
tolerance0  3.6776281 -5.3964437  1.1128421
tolerance1 -2.3955483 -1.0211719 10.7874512
tolerance2  4.5040579 -6.1915000  0.3030713
tolerance3  2.7420911 -2.5744267 -2.8481376
tolerance4 -0.6175252  2.2641002 -3.6331099
tolerance5 -7.2561266 11.9240239 -5.4354457
```