Duke Datathon Writeup

Linda Cao, Georgie Stammer, Enzo Moraes Mescall, Erik Mendes Novak

2021-11-07

Introduction

Immediately after going over the data set our group was interested in section P of the questionnaire where individuals were asked a series of intense political questions and responded with their opinions. This provides a unique subjective insight of a population's perspective on their civic space. The data set provides information about the material conditions, and their perceived material conditions, of the surveyed and thus allows us to attempt to estimate the impact of an individual's material conditions on their political leanings. Notably, all the questions in section P are on a scale of 1 - 4 where answering 1 indicates support of a larger authoritarian government while answering 4 indicates support for a more democratic form of government.

Research Question: *To what extent do an individual's material conditions affect their leanings toward authoritarian form of government?*

Methodology

Our statistical intent was to construct a general linear model and have our response variable be some encapsulation of the answers in section P of the questionnaire. We settled on the simplest possibility: to sum each interviewee's responses to the 29 relevant questions and divide the result by 29, thus finding their mean political leaning. This method, however, works best when assuming all questions are weighed equally and provoke a similar spread of responses, which we could not assume. For one, in 29 questions it is very possible that the wording in some of them leads to biased results. Similarly, questions could be phrased similarly or cover similar roles such that their responses would be highly correlated, even if all questions are ostensibly independent. Thus, a two-step approach was taken to increase confidence in our necessary assumptions.

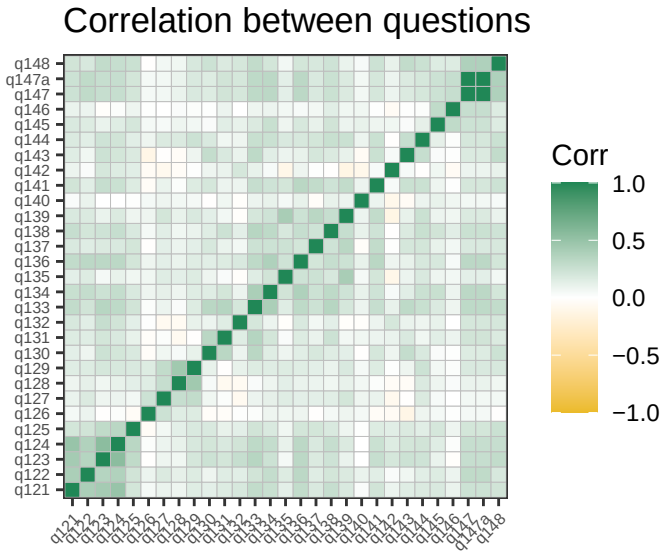
Firstly, we calculated the median and interquartile range (IQR) of the responses for each question. These choices were made because each question uses a Likert scale, meaning answers are discrete and ordinal. In this scenario, the spacing between answer options cannot be effectively calculated, so using the median for a measure of centrality sidesteps any spacing inconsistencies; a similar rationale applies to the IQR.

From those calculations, all but three questions had an IQR of 1 and a median at 2 or 3. An IQR of 1 indicates that the question had reasonable but not polarizing spread, and thus a more all-filling distribution. A median of 2 or 3 indicates that the population, generally, did not hold a lopsided opinion on the issue. The three questions that did not fit this context were questions 123, 125 and 147a. 123 had a median of 4, indicating that at least half the respondents felt strongly about it, and thus the question becomes incapable of capturing nuance. 125 had an IQR of 2, indicating that proportionately fewer respondents had a mild response to it, with many more answering a 1 or a 4. Finally, question 147a had an IQR of 0, indicating that over half the respondents answered the same, also indicating very little spread of opinions.

Then, we considered the possibility that some questions might have been measuring equivalent components in respondents' political inclinations. It is undeniable that one's political inclination is a multifaceted and complex viewpoint, such that even seemingly different questions may in fact elucidate preferences that are in fact fundamentally similar. That is, though each question was independent, some could be highly correlated with others. If so, we would effectively be measuring a particular facet of one's political inclination twice, thereby assigning it excess weight in the calculation of an average. To combat this, we calculated the correlation coefficient for all pair choices given the remaining 26 questions. We used the Spearman rho

correlation coefficient given the ordinal monotonic nature of the data, as this choice is more natural for discrete data with intangible option spacing. We set the threshold after which we would consider questions too correlated at 0.38, because we wanted to ensure all remaining questions were at most weakly correlated. Literature often claims anything above 0.4 is at least moderately correlated, so we left some buffer and set it at 0.38.

```
ggcorrplot(corr_ideology, ggtheme = theme_bw(), tl.cex = 6,
           colors = c("#ebbb2a", "white", "#208756")) +
  labs(title = "Correlation between questions")
```



There were 4 pairs flagged: 121 and 124, 128 and 129, 131 and 133, and 133 and 134. The remaining task was to decide which of these questions to remove and which to keep. Because 134 and 131 did not meet the threshold in their correlation, it was sensible to remove 133, thereby removing two problematic pairs already. For the remaining two pairs, we calculated the mean correlation coefficient between each question and all other questions (so, from the 26 questions, we only did not sum its correlation with itself, which is always 1, or the pair's correlation, flagged as too high). The mean correlations were as follows: 121 – 0.119662228, 124 – 0.12142794, 128 – 0.103254203, 129 – 0.103928434. Clearly, keeping 121 was better; finally, though very close, 128 edged out 129.

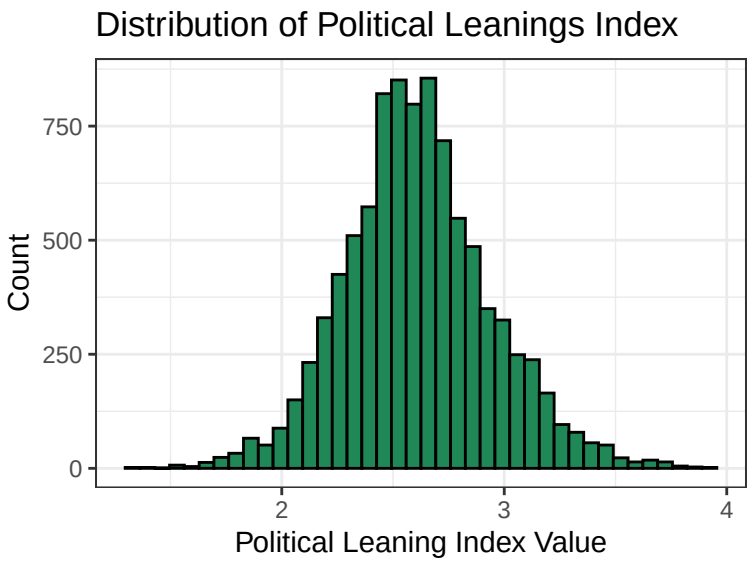
It is important to note we were not interested in multicollinearity in general, hence why we did not use VIF. This is because even if one set of answers could be written as a long linear combination of other answers, we thought this relationship would be sufficiently distinct such that these questions would still be measuring fundamentally different facets about a respondent's political inclinations. Meanwhile, as an example, questions 121 and 124 had very similar wordings and concepts, as follows: [INSERT WORDING FOR THESE TWO QUESTIONS HERE].

Having removed these 6 questions, and having 23 remaining, we felt substantially more confident in our assumption that for the mean political inclination of each respondent we could simply take the unweighted sum of all their responses

Exploratory Data Analysis

```
ggplot(data = final_data, aes(x = avg)) +
  geom_histogram(fill = "#208756", color = "black", bins = 40) +
  labs(title = "Distribution of Political Leanings Index",
       x = "Political Leaning Index Value",
```

```
       y = "Count")  +
  theme_bw()
```




```
kable(learning_index, digits = 3,
      caption = "Summary Statistics for Learning Index",
      col.names = c("Mean", "Max", "Median", "Min", "IQR", "SD"))
```

Table 1: Summary Statistics for Learning Index

Mean	Max	Median	Min	IQR	SD
2.613	3.926	2.593	1.333	0.415	0.33

The distribution is unimodal, looks vaguely normal and doesn't have any apparent outliers.

Data Analysis

After cleaning up the data we created a linear model using the responses to questions SE001 through SE017 and then selected the most relevant predictor variables to keep in the model based off of their p-values with the threshold being $p > 0.05$. We iterated through the predictors removing the entries with the highest p-value and re-running regression until arriving at the following model:

```
linear_model <- lm(avg ~ country + gender + education, data = final_data) %>%
  tidy(conf.int = TRUE) %>%
  kable(digits = 3, caption = "Linear Regression Model for Political Index")
linear_model
```

Table 2: Linear Regression Model for Political Index

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.767	0.014	202.361	0.000	2.740	2.794
countryHong Kong	-0.296	0.014	-21.311	0.000	-0.324	-0.269
countryKorea	-0.204	0.011	-18.378	0.000	-0.225	-0.182
countryChina	-0.473	0.012	-38.357	0.000	-0.497	-0.449
countryMongolia	-0.430	0.012	-35.948	0.000	-0.454	-0.407

term	estimate	std.error	statistic	p.value	conf.low	conf.high
countryPhilippines	-0.327	0.012	-27.324	0.000	-0.350	-0.304
countryTaiwan	-0.268	0.012	-22.565	0.000	-0.291	-0.245
countryThailand	-0.436	0.012	-36.208	0.000	-0.460	-0.412
gender	-0.014	0.006	-2.396	0.017	-0.025	-0.003
education	0.016	0.001	19.557	0.000	0.014	0.017

Which we may represent in the equation:

$$\hat{PolIndex} = 2.767 - 0.296\ countryHK_i\ - 0.204\ countryKO_i - 0.473\ countryCN_i\ - 0.430\ countryMN_i$$

$$-0.327\ countryPH_i\ - 0.268\ countryTW_i\ - 0.014_i\ gender_i + 0.016\ education_i\ + \epsilon_i, \qquad \epsilon \sim N(, \sigma_\epsilon^2)$$

The y-intercept represents a male individual from Japan with 0 years of formal education. All the countries are indicator variables and these carry the most weight on the response variable. Given that zero is not included in any of the countries' confidence intervals, we are 95% confident that the average political index for each country is not the same. Gender is also categorical but having a small impact of only a -0.014 difference in female political indices and we are 95% confident that the political index for males differs from -0.025 to -0.003 compared to females, holding all other predictor variables constant. The only numerical is years of education where we are 95% confident that for every additional year in education, the respondent's political index increases by 0.014 to 0.017, holding all other predictor variables constant.

Conclusion

These findings would match our intuition about education where increasing years of schooling dissuades authoritarian ideology. Another point to consider is that belonging to countries generally associated to authoritarian regimes, like the PRC, Mongolia and Thailand, generally have a much greater negative effect on an individual's political ideology than belonging to Japan or Korea.

```
anova(lm(avg ~ country + gender + education, data = final_data)) %>%
  tidy() %>%
  kable(digits = 3, caption = "ANOVA of Linear Regression Model")
```

Table 3: ANOVA of Linear Regression Model

term	df	sumsq	meansq	statistic	p.value
country	7	251.046	35.864	467.623	0
gender	1	1.215	1.215	15.844	0
education	1	29.333	29.333	382.474	0
Residuals	9062	694.997	0.077	NA	NA

According to the ANOVA table, all predictor values have a very low p-value, indicating that they are statistically significant to the model. However, the model has an adjusted R^2 value of 0.2876, even with a large F-statistic and low p-value. Therefore, we can conclude that this model is statistically significant but may not be best represented by linear regression.

Appendix A

```
library(tidyverse)
library(knitr)
library(broom)
library(patchwork)
library(haven)
library(yardstick)
library(ggcorrplot)
```

```
Wave1_data <- read_sav("Wave1_20170906.sav")
Wave1_data$index <- 1:nrow(Wave1_data)

predictors <- select(Wave1_data, c(country, level3,
                                   se002:se012a, se014:pgnum, index))
```

```
ideology_responses <- Wave1_data %>%
  select(index, q121:q148)

# This is our janky method of removing responses which failed to answer more
# than 4 questions by turning NAs into huge numbers and removing entries with
# sums greater than 4.9 million, indicating that they had 5 or more NAs.
# We chose 5 NAs because otherwise most of the Chinese and Thai data gets
# removed.
ideology_responses[is.na(ideology_responses)] <- 100000000

na_sum <- ideology_responses %>%
  mutate(sum = rowSums(ideology_responses[])) %>%
  filter(sum <499999999)

na_sum[na_sum > 100000] <- NA

cleaned_ideology_0 <- na_sum[, -31]
cleaned_ideology <- cleaned_ideology_0[, -1]

# removing rows using the methodology explained
added_avg <- na_sum %>%
  select(index, q121, q122, q126:q128, q130:q132, q134:q147, q148)

added_avg$avg <- rowMeans(subset(na_sum, select = c(q121:q148)), na.rm = TRUE)

# calculating the political index
average_ideology <- select(added_avg, c(index, avg))

clean_data <- inner_join(average_ideology, predictors, by = "index")
```

```
final_data <- clean_data %>%
  mutate(country = as.factor(country)) %>%
  mutate(urban_rural = level3 - 1) %>%
  mutate(gender = se002 - 1) %>%
  mutate(age = se003a) %>%
  mutate(marital_status = as.factor(se004a - 1)) %>%
  mutate(education = se005a) %>%
  mutate(household_members = se008a) %>%
```

```r
  mutate(household_generations = se008b) %>%
  mutate(subjective_status = se017) %>%
  mutate(religion = case_when(se006 == 11 ~ 0, se006 < 11 ~ 1))

levels(final_data$country) <- c("Japan", "Hong Kong", "Korea", "China",
                          "Mongolia", "Philippines", "Taiwan", "Thailand")

linear_model <- lm(avg ~ country + gender + education, data = final_data) %>%
  tidy(conf.int = TRUE) %>%
  kable(digits = 3, title = "Linear Regression Model for Political Index")
```

```r
summary(lm(avg ~ country + gender + education, data = final_data))
```

```
##
## Call:
## lm(formula = avg ~ country + gender + education, data = final_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17289 -0.17044 -0.00676  0.16198  1.28771
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.7668171  0.0136727 202.361   <2e-16 ***
## countryHong Kong   -0.2963615  0.0139064 -21.311   <2e-16 ***
## countryKorea       -0.2036964  0.0110840 -18.378   <2e-16 ***
## countryChina       -0.4728282  0.0123270 -38.357   <2e-16 ***
## countryMongolia    -0.4303382  0.0119711 -35.948   <2e-16 ***
## countryPhilippines -0.3270026  0.0119677 -27.324   <2e-16 ***
## countryTaiwan      -0.2677718  0.0118669 -22.565   <2e-16 ***
## countryThailand    -0.4360120  0.0120418 -36.208   <2e-16 ***
## gender             -0.0140057  0.0058466  -2.396   0.0166 *
## education           0.0155385  0.0007945  19.557   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2769 on 9062 degrees of freedom
##   (204 observations deleted due to missingness)
## Multiple R-squared:  0.2883, Adjusted R-squared:  0.2876
## F-statistic:   408 on 9 and 9062 DF,  p-value: < 2.2e-16
```