Problem Set 3

Luke Haner

2025-11-04

#Paper Analysis

#1. Is the goal of the study causal inference, description, prediction, or something else? Have the authors clearly stated their goals? Describe any strengths or weaknesses in how the authors articulate their research objectives.

The goal of the study is primarily causal inference as the authors try to decide what is the causal mechanism in the rise of civil wars post WW2 through the fall of the USSR. Specifically the authors assert that civil war onset is due to factors that favor insurgency rather than primarily ethnic fractionaliztion.

A strength in how the authors articulate their research objectives is that they clearly highlight how argument departs and contrasts from previous literature. Additionally, they quickly explain reasons why previous explanations were insufficient, and then empirically showed so, alongside outlining and giving support for their alternative theories.

A weakness I think the authors have in articulating research objective is that they address too many alternative explanations at once. At times the causal relationships they are looking for in each theory and how it ties together becomes a little tough to distinguish which makes the connections between theories to the overall arguments a little unclear.

Estimands: Have the authors sufficiently definied their theorticial and emprical estimands. Discuss what these are and explain and how the authors could clarify if necessary.

Hypothesis 1: Ethnic or Religious diversity increases Civil War Risk

Theoretical Estimand: The causal effect of a country's ethnic and religious diversity on the probability that a civil war begins

Empirical Estimand: The estimated change in the probability of civil war onset from a one unit change in ethnic and religious diversity measures, holding other factors constant.

Hypothesis 2: The effect or ethic/religious diversity on probability of civil war onset is increases at higher levels of per capital income

Theoretical Estimand: The causal effect of diversity interacting with modernization (per capital income) on the probability of civil war onset

Empirical Estimand: The estimated change in probability of civil war onset from the interaction between high per capital income and religious/ethnic diversity, holding other factors constant.

Hypothesis 3: Countries with ethnic majority and significant ethnic minority are at greater risk for civil war

Theoretical Estimand: the causal effect of a country's minority-majority ethnic group structure on the probability of civil war onset

Empirical Estimand: The estimated change in the probability of civil war onset associated with presence of ethnic majority and significant minority group, holding other factors constant.

Clarification: The authors bring up this hypothesis as more of an implication of previous work, rather than developing out it out and testing it.

Hypothesis 4: Measures of political democracy and civil liberty should be associated with lower risks of civil war onset

Theoretical Estimand: The causal effect of a country's level of political democracy and civil liberties on the probability that a civil war begins

Empirical Estimand: The estimated change in probability of civil war onset associated with higher democracy or civil liberty scores, holding other factors constant.

Hypothesis 5: Policies that discriminate in favor of a particular language or religion should raise th risk of civil war onset in states with religious or linguistic minorities

Theoretical Estimand: The causal effect of state policies that favor a particular language or religion on the probability of civil war onset

Empirical Estimand: The estimated change in the probability of civil war onset associated with discriminatory polices, holding other favors constant

Hypothesis 6: Greater income inequality should be associated with higher risks of civil war onset

Theoretical Estimand: The causal effect of income inequality within a country on the probability of civil war onset

Empirical Estimand: The estimated change in the probability of civil war onset from higher measures of income inequality (using Gini coefficients), holding other factors constant.

Hypothesis 7: Among countries with an ethic minority of at least 5% of the total population, greater ethnic diversity should associate with a higher risk of ethnic civil war

Theoretical Estimand: The causal effect of ethnic diversity (within countries that have a minority group of at east 5% of the total population) on the probability of civil war onset

Empirical Estimand: The estimated change in the probability of ethnic civil war onset from the increases of ethnic diversity within countries that already have at least one ethnic group that compromises 5% of the total population, holding other factors constant.

Clarification: This hypothesis extends from the discussion on Horowitz;s 1985 theory of ethnic conflict rather than from the authors theoretical framework. It is not developed/considered much, especially due to measurement issues in defining what counts as a "ethnic civil war" is.

Hypothesis 8(a,b,c): Terrain and Local Support and civil war onset

Theoretical Estimand: The causal effect of insurgency favoring conditions (such as rough terrains, access to sanctuaries in foreign countries, and local population cooperation), on the probability of civil war onset.

Empirical Estimand: The estimated change in the probability of civil war onset from the presence of insurgency favoring conditions, holding other factors constant.

Clarification: Some of the conditions may be hard to measure which makes the empirical proxy not fully representative of the true mechanism at play.

Hypothesis 9: State Capacity (Per Capita Income) and Civil War Onset

Theoretical Estimand: The causal effect of a state overall financial, administrative, and enforcement capacity (proxies by per capital income) on the probability of civil war onset

Empirical Estimand: The estimated change in the probability fo civil war onset from higher per capita income, a proxy for state capacity, holding other factors constant.

Clarification: The approach and measurement links a lot fo capacity measurements to per capital income which may not be main mechanism in place driving the various avenues of state capacity mentioned.

Hypothesis 10: Insurgency Favoring Factors and onset of Civil War

Theoretical Estimand: The causal effect of insurgency favoring conditions (new statehood, political instability, mixed regime type, large country population, territory base separation from state center, foreign government/diaspora willing to provide support, high value low weight natural resources, and oil revenues) on the probability that a civil war begins in a given year.

Empirical Estimand: The estimated change in the probability of civil war onset from each of the previously mentioned insurgency favoring conditions, holding other factors constant.

Clarification: Several of the conditions are hard to measure directly. The authors try to overcome this by using various proxies for measurement, but they may not be fully interpretative of the various mechanisms at play.

#Identification Strategy

The identification strategy used throughout the paper is multivariate logit regression with a large amount of control variables to try and isolate the causal effect of insurgency conditions on civil war onset.

#Assessment of Findings

The paper provides a lot of new understanding of the factors that influence civil war onset, and furthers the literature from considering ethnic/religious fractionalization as the primary cause. I think the identification strategy is a little limiting for the amount of hypotheses presented, and as such it is difficult to ascertain the results as strong causal claims/relationships. I think this topic has a complex data-generating process which is not easy to model, and isn't fully captured by the work done in the paper. Additionally, a lot of key measures tested in various hypotheses are hard to measure adequately, requiring the usage of various proxies. These proxies work and are supplemented by theory, but I think they may be leaving out some of the true mechanisms at play. Most, if not all, of the data is credibly measured except for a handful of cases where the measure itself is hard to nail down. Overall, the findings show statistically convincing correlation and weak causation.

#Broader Considerations

I think this research can still inform our understanding of the world. It provides a first look into a wide array of factors for researches to delve into that may not have been previously covered. Some are very location specific such as how rugged geography increases the chance of civil war onset, and other are more broad in looking in terms of state economic capacity. In essence, I think the findings expand the areas of research on the topic and iterate that perhaps no one reason is the standalone cause for civil war onset.

#Data Analysis

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
therm_data <- read.csv("thermometers.csv")
```

#1.Use the birth_year variable to create a new age variable (Note this survey was taken in 2017)

```r
#age of the person when they took the survey
therm_data$survey_age <- 2017 - therm_data$birth_year

head(therm_data$survey_age)
```

```
## [1] 86 65 86 65 78 58
```

#2.Pick one of the feeling thermometers and one of the categorical demographic. variables (sex, race, party_id, or educ). Describe the spread and central tendency of the feeling thermometer both for all observations, and for each category in the demographic variable you chose. Use histograms or density plots to visualize the distribution

```r
#choosing feelings toward immigration/immigrants and demographic variable of education
summary(therm_data$ft_immig)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   50.00   65.00   61.92   82.00  100.00     197
```

```r
summary(therm_data$educ)
```

```
##    Length     Class      Mode
##      4989 character character
```

```r
#view different category
unique(therm_data$educ)
```

```
## [1] "4-year"                 "2-year"                 "High school graduate"
## [4] "Post-grad"              "Some college"           "No HS"
```

```r
#Spread and Central Tendency of the ft_immig (use na.rm = true to take out missing values)
median(therm_data$ft_immig, na.rm = TRUE)
```

```
## [1] 65
```

```r
sd(therm_data$ft_immig, na.rm = TRUE)
```

```
## [1] 27.19318
```

```r
#The median of the feeling thermometer, which is a measure of the central tendency of the data is 65 wh

#Spread and Central Tendency for each category of education level
education_summary <- therm_data %>%
  group_by(educ) %>%
  summarise(
    median_ft = median(ft_immig, na.rm = TRUE),
    sd_ft = sd(ft_immig, na.rm = TRUE)
  )

education_summary
```

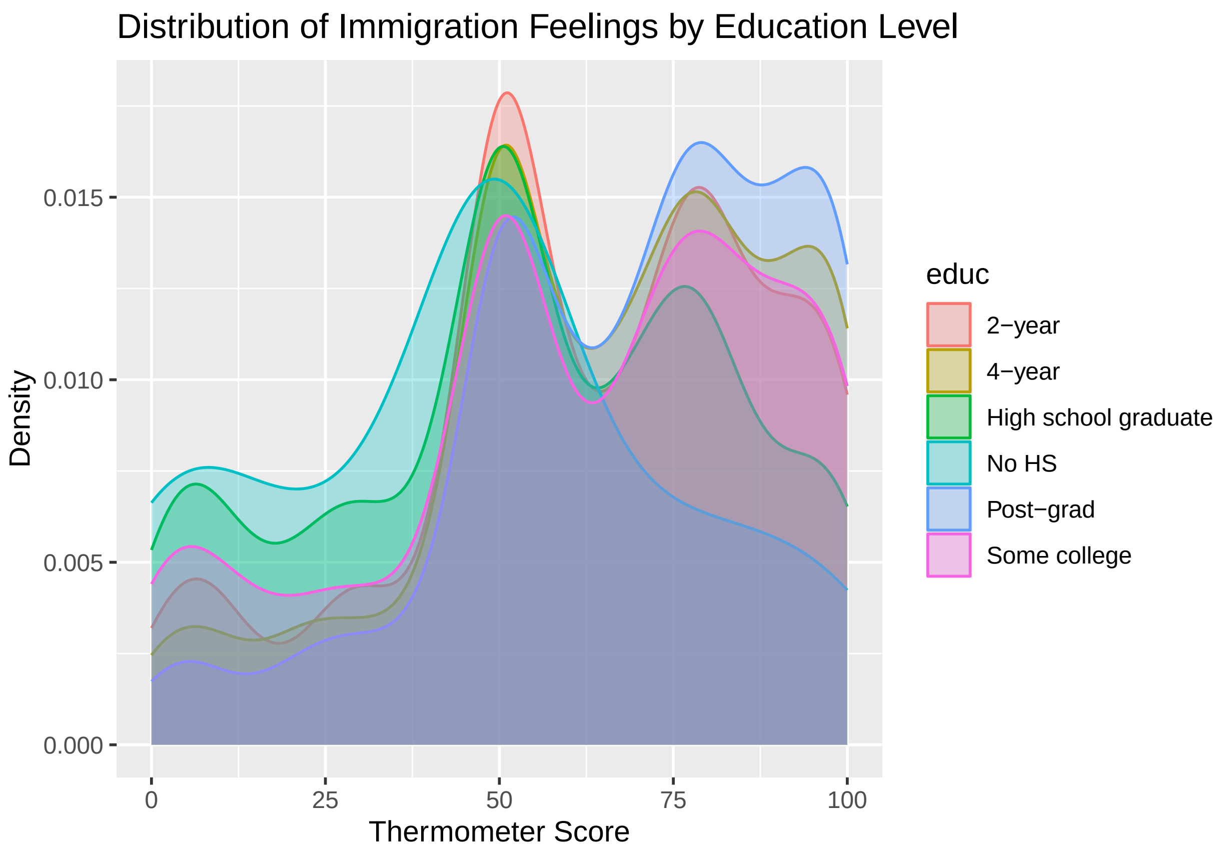
```
## # A tibble: 6 x 3
##   educ                  median_ft sd_ft
##   <chr>                     <dbl> <dbl>
## 1 2-year                       65  26.4
## 2 4-year                       69  25.6
## 3 High school graduate         52  28.2
## 4 No HS                        50  28.9
## 5 Post-grad                    74  24.3
## 6 Some college                 65  28.6
```

```
#The central tendency and spread of the data broken down by education level suggests that respondents w
```

#Histogram to visulize dstirbution

```
ggplot(therm_data, aes(x = ft_immig, color = educ, fill = educ)) +
  geom_density(alpha = 0.3) +
  labs(
    title = "Distribution of Immigration Feelings by Education Level",
    x = "Thermometer Score",
    y = "Density"
  )
```

```
## Warning: Removed 197 rows containing non-finite outside the scale range
## (‘stat_density()‘).
```



#3. Fit a regression model to estimate the conditional mean of the feeling thermometer for each category in the demogrpahic variable you choose.

```
regression_model1 <- lm(ft_immig ~ educ, data = therm_data)

summary(regression_model1)
```

```
##
## Call:
## lm(formula = ft_immig ~ educ, data = therm_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.906 -15.146   2.661  21.169  53.416
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                62.604      1.009  62.045  < 2e-16 ***
## educ4-year                  2.542      1.252   2.030   0.0424 *
## educHigh school graduate   -8.104      1.275  -6.354 2.30e-10 ***
## educNo HS                 -16.019      3.203  -5.001 5.90e-07 ***
## educPost-grad               6.303      1.372   4.593 4.47e-06 ***
## educSome college           -1.426      1.412  -1.010   0.3126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 26.68 on 4786 degrees of freedom
##   (197 observations deleted due to missingness)
## Multiple R-squared:  0.03863,    Adjusted R-squared:  0.03762
## F-statistic: 38.46 on 5 and 4786 DF,  p-value: < 2.2e-16
```

#4.Create a new dataframe that only contains rows for Democracts and Republicans, create a new binary variable for party_id

```
#creating binary variable for party id; Democrats equal 1 and republicans equal 0
partyid_binary <- ifelse(therm_data$party_id == 'Democrat', 1, 0)

#adding to original dataframe
therm_data$partid_binary <- partyid_binary

print(head(therm_data$partid_binary))
```

```
## [1] 1 0 0 0 1 1
```

```
#create new dataframe only containing rows for republicans and democrats
partisan_data <- subset(therm_data, party_id %in% c('Democrat', 'Republican')) #new dataframe should in
```

#5. Use multiple linear regression to build a model that predicts your binary party id variable. Use any combination of varibales you like, but you should include at least one feeling thermometer and one interaction term. Justify your model.

```
#creating multiple linear regression model predicting party id from a binary variable with the feeling
multiple_regression1 <- lm(partid_binary ~ ft_immig * educ, data = partisan_data)

#Results show the expected value of being a Democrat (coded 1) based on education, feeling thermometer,
summary(multiple_regression1)
```

```
##
## Call:
## lm(formula = partid_binary ~ ft_immig * educ, data = partisan_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8813 -0.4587  0.1729  0.3684  0.9454
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       0.1315713  0.0549964   2.392   0.0168 *
## ft_immig                          0.0069798  0.0008025   8.698   <2e-16 ***
## educ4-year                       -0.0488155  0.0706051  -0.691   0.4894
## educHigh school graduate          0.0374929  0.0651404   0.576   0.5649
## educNo HS                        -0.0404332  0.1375209  -0.294   0.7688
## educPost-grad                    -0.1688502  0.0817666  -2.065   0.0390 *
## educSome college                  0.0732075  0.0761683   0.961   0.3366
## ft_immig:educ4-year               0.0005391  0.0010177   0.530   0.5963
## ft_immig:educHigh school graduate -0.0011846  0.0009828  -1.205   0.2281
## ft_immig:educNo HS                0.0002502  0.0024410   0.103   0.9184
## ft_immig:educPost-grad            0.0022059  0.0011538   1.912   0.0560 .
## ft_immig:educSome college         -0.0009678  0.0011126  -0.870   0.3845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4583 on 3011 degrees of freedom
##   (123 observations deleted due to missingness)
## Multiple R-squared:  0.1538, Adjusted R-squared:  0.1508
## F-statistic: 49.77 on 11 and 3011 DF,  p-value: < 2.2e-16
```

```
#Justification

#I include education, immigration attitudes and the interaction between the two to best see how feeling
```

#6.The coefficients in your model represent the change in what?

```
# The coefficients in my model represent how a one unit change in immigration attitudes, having a certa
```

#7. Select one of the feeling thermometers in your model and plot how your predicted values change as the feeling thermometer changes. Interpret your results. Can this reasonably be interpreted as a causal effect?

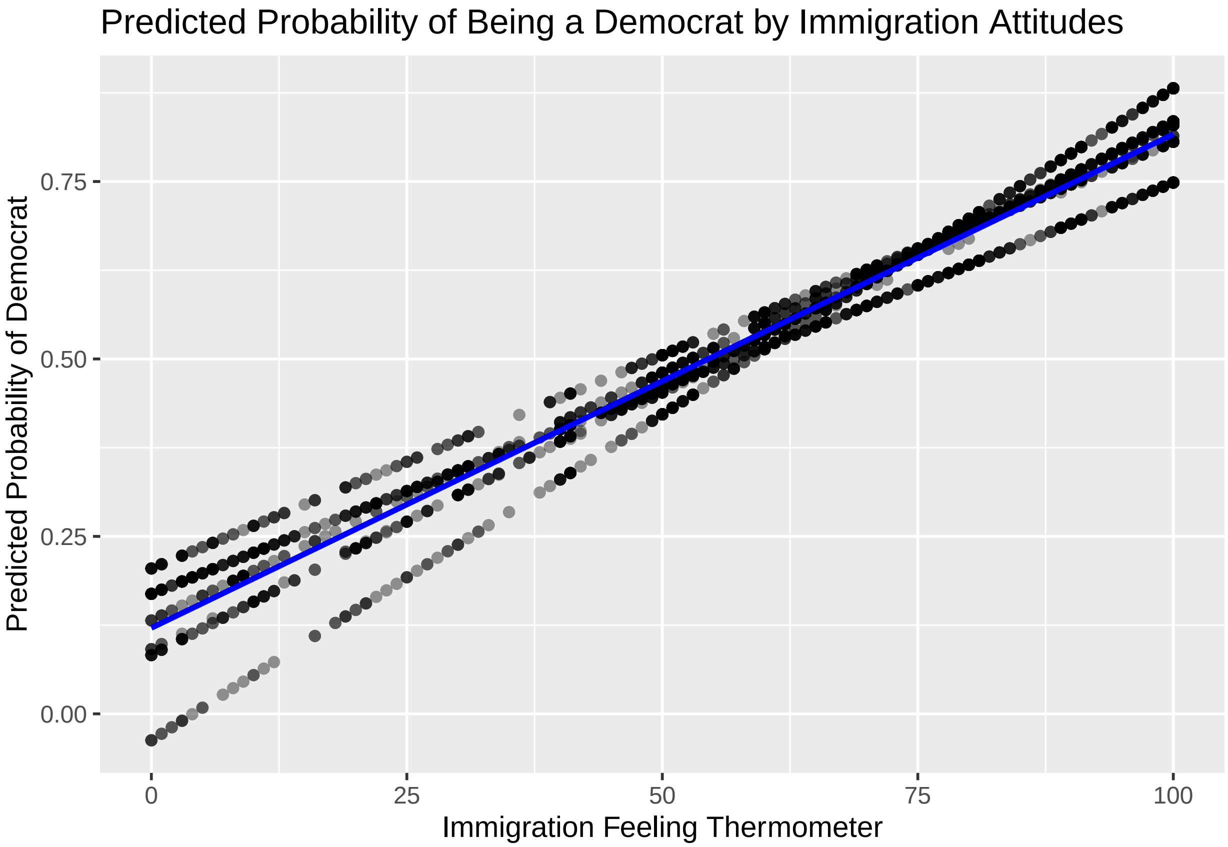
```
#adding predicted values from multiple linear regression model to exisitng data
partisan_data$predicted_val <- predict(multiple_regression1, newdata = partisan_data)

#Plot of predicted values as immigration feeling thermometer changes
ggplot(partisan_data, aes(x = ft_immig, y = predicted_val)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", color = "blue") +
  xlab("Immigration Feeling Thermometer") +
  ylab("Predicted Probability of Democrat") +
  ggtitle("Predicted Probability of Being a Democrat by Immigration Attitudes")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 123 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 123 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



```
#Interpretation

# From the plot, it can be shown there is a positive relationship between immigration attitudes and pro
```