Problem Set 3

Fernanda Valdez

2025-11-19

Part 1: Paper Analysis 1. Research Goals: 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 causal inference because it aims to understand what causes civil wars. Specifically, the authors want to know if the end of the Cold War could have led to civil wars around the world or if these civil wars were caused by other circumstances that began long before the Cold War. Additionally, they want to examine two theories on civil war which are grievance and insurgency. Also, they want to know why civil wars occurred in some countries and not others as well as why these civil wars occurred when they did. Though the authors clearly stated their goals, it is a monumental task because they are looking at many variables. One could argue that their scope is too broad in that they want to compare civil wars in many countries (161 countries to be exact) rather than a specific group of countries in a region, for example.

2. Estimands: Have the authors sufficiently defined their theoretical and empirical estimands? Discuss what these estimands are and explain how the authors could clarify them if necessary.

Though authors include many hypotheses and theoretical estimands, their study could benefit from more clarity on what exactly they are are trying to measure and how they are doing so. For example, their hypothesis 7 on ethnic civil war is unclear because a country cannot have an ethnic civil war if they do not have an ethnic minority. Essentially, their study could benefit from clarification on their theoretical estimands.

Theoretical estimands:

H1: Average causal effect of ethnic/religious diversity on the risk of civil war

(empirical estimand: regression coefficient between ethnic/religious diversity and the risk of civil war)

H2: Average causal effect of higher levels of per capita income on the effect of ethnic diversity on the probability of civil war

(empirical estimand: regression coefficient between higher levels of per capita income and the effect of ethnic diversity)

H3: Average causal effect of ethnic majority and an ethnic minority on the risk of civil war

(empirical estimand: regression coefficient between ethnic majority and an ethnic minority and the risk of civil war)

H4: Average causal effect of political democracy and civil liberties on the risk of civil war

(empirical estimand: regression coefficient between political democracy and civil liberties and the risk of civil war)

H5: Average causal effect of policies favoring a specific language or religion on the risk of civil war

(empirical estimand: regression coefficient between policies favoring a specific language or religion and the risk of civil war)

H6: Average causal effect of greater income inequality on the risk of civil war

(empirical estimand: regression coefficient between greater income inequality and the risk of civil war)

H7: Average causal effect of ethnic minorities on the risk of ethnic civil war

(empirical estimand: regression coefficient between ethnic minorities and the risk of civil war)

H8: Average causal effect of rough terrain on the risk of insurgency

(empirical estimand: regression coefficient between rough terrain and the risk of civil war)

H9: Average causal effect of proxies for insurgency on civil war

(empirical estimand: regression coefficient between proxies for insurgency and the risk of civil war)

H10: Average causal effect of political and military technology of insurgencies on the risk of civil war

(empirical estimand: regression coefficient between political and military technology of insurgencies and the risk of civil war)

H11: Average causal effect of small number of rebels on the risk of civil war

(empirical estimand: regression coefficient between small number of rebels and the risk of civil war)

3. Identification Strategy: What does the research do to ensure that the empirical estimand is a good measure of the theoretical estimand? Describe the authors' identification strategy. (what do they control for?) The authors control for prior war, that is, if a country already had a civil war underway in the year before. They also control for per capita income, recent political instability, and ethnic diversity.
4. Assessment of Findings: Does the identification strategy support the authors' claims? For example, could the regression coefficients be credibly interpreted as causal effects if causal inference is the goal? Does the model adequately represent the real-world data-generating process? Does the data credibly measure the phenomena being studied?

No, they control for many variables. However, it is impossible to consider for every possible variable. There are many factors that can contribute to what causes a civil war. So, for example, because this is an observational study (no randomization) then there are many unobservable variables that they need to consider. Since their study is so broad, it is difficult to say that they considered every relevant confounders in their study. Additionally, what they did use, we also cannot confidently say that, for example, using GDP is a great variable to use in order to measure a country's vulnerability to going into civil war.

5. Broader Contribution: Despite any weaknesses, can this research still inform our understanding of the world? If so, how?

Despite this research's weaknesses, this research can still inform us about the variables that might affect civil wars. It also demonstrates how trying to measure too many things could undermine your claims.

Part 2: 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(tidyr)
library(ggplot2)
```

```r
require(ggplot2)
set.seed(7965)
```

1. Load the thermometer

```r
# Load thermometer data
setwd("~/Desktop/POLS 602")
thermometer <- read.csv("thermometers.csv")
# A few observations
head(thermometer)
```

```
##   birth_year    sex  race    party_id                 educ ft_black ft_white
## 1       1931 Female White    Democrat               4-year       51       50
## 2       1952 Female White  Republican               2-year       98       90
## 3       1931   Male White Independent High school graduate       87       90
## 4       1952   Male White  Republican               4-year       90       85
## 5       1939 Female White    Democrat               2-year      100       50
## 6       1959 Female Black    Democrat            Post-grad       98       70
##   ft_hisp ft_asian ft_muslim ft_jew ft_christ ft_fem ft_immig ft_gays ft_unions
## 1      79       50        50     50        50     99       95      50        80
## 2      95      100        61    100        98     65       96      82        62
## 3      91       88        49     25        50     74       77      77       100
## 4      90       96        80     91        94     25       91      71        20
## 5     100      100       100    100        28    100      100     100       100
## 6      99      100       100    100       100     73      100      54        80
##   ft_police ft_altright ft_evang ft_dem ft_rep
## 1        76           1       50     88     21
## 2        95          50       96     86     96
## 3        78           0        2     91     20
## 4        94          50       70     22     83
## 5        28          NA       NA     99     NA
## 6        24           4       53     53      4
```

```r
# New age variable (survey done in 2017)
thermometer$age <- 2017 - thermometer$birth_year
thermometer$birth_year<-as.numeric(as.character(thermometer$birth_year))
thermometer$age<-2017 - thermometer$birth_year
```

2. Spread, central tendency, and visualization.

```r
# Spread and Central Tendency: All observations for Hispanics
summary(thermometer$ft_hisp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   51.00   75.00   70.18   90.00  100.00     146
```

```r
mean(thermometer$ft_hisp, na.rm=TRUE)
```

```
## [1] 70.18005
```

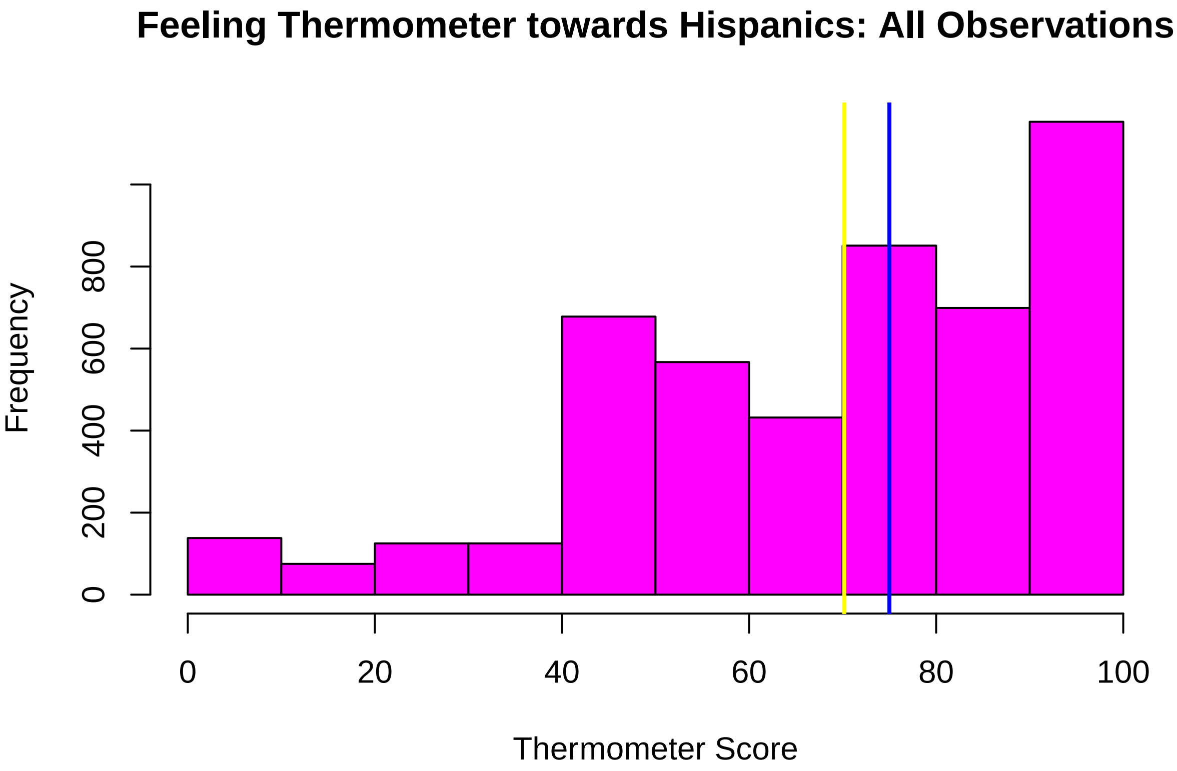
```r
median(thermometer$ft_hisp, na.rm=TRUE)
```

```
## [1] 75
```

```r
sd(thermometer$ft_hisp, na.rm=TRUE)
```

```
## [1] 23.95509
```

```r
# Visualize: All observations for Hispanics
hist(thermometer$ft_hisp,
     main="Feeling Thermometer towards Hispanics: All Observations",
     xlab="Thermometer Score",
     col="magenta",
     border="black")
# With Mean
abline(v = mean(thermometer$ft_hisp, na.rm = TRUE), col = "yellow", lwd = 2)
# With Median
abline(v = median(thermometer$ft_hisp, na.rm = TRUE), col = "blue", lwd = 2)
```

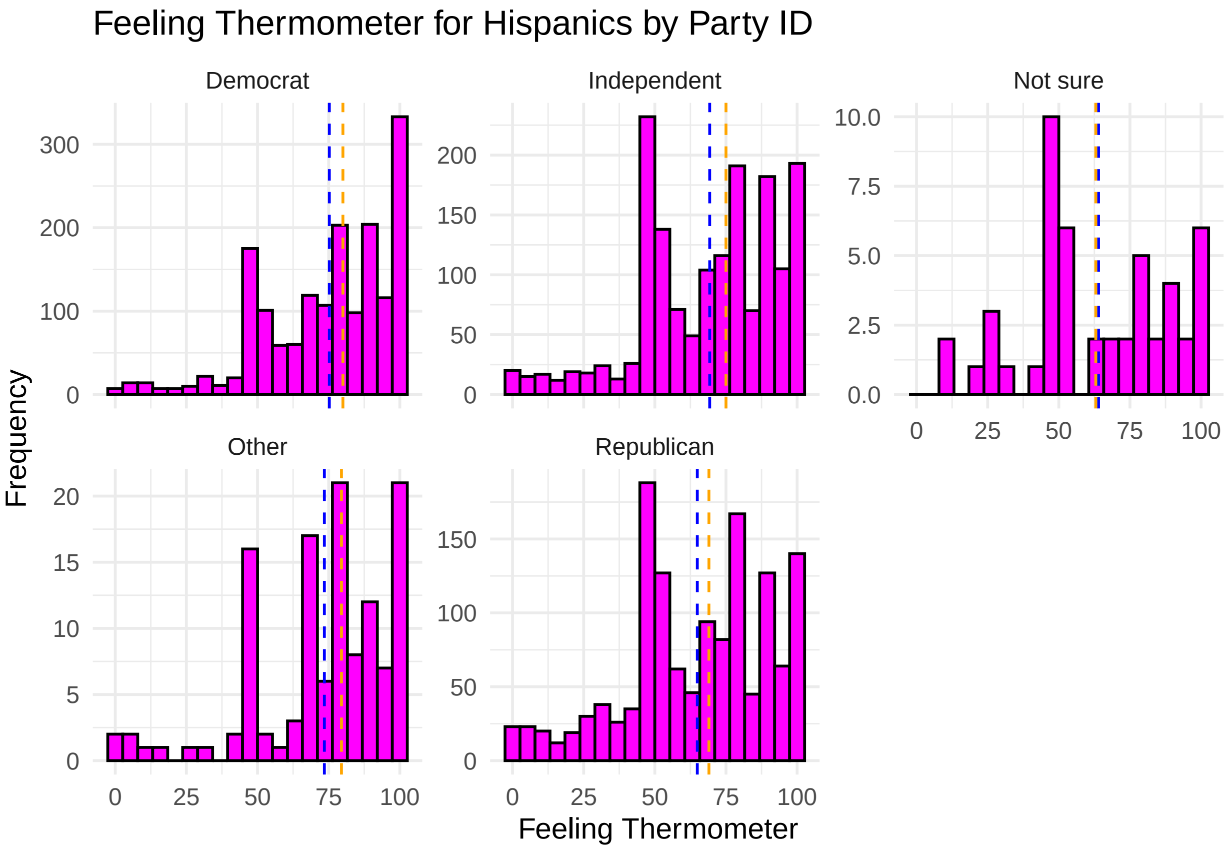


```r
# Spread and Central Tendency: Party ID for Hispanics
# Hispanics: Party ID
party <- thermometer %>%
  group_by(party_id) %>%
  summarise(
    mean_ft = mean(ft_hisp, na.rm = TRUE),
    median_ft = median(ft_hisp, na.rm = TRUE),
    sd_ft = sd(ft_hisp, na.rm = TRUE),
    n = n()
  )
print(party)
```

```
## # A tibble: 5 x 5
##   party_id    mean_ft median_ft sd_ft     n
##   <chr>         <dbl>     <dbl> <dbl> <int>
## 1 Democrat       75.2      80    22.0  1734
## 2 Independent    69.3      75    23.7  1658
## 3 Not sure       64.0      63    25.5    55
## 4 Other          73.5      79.5  23.4   130
## 5 Republican     64.9       69    25.2   1412
```

```
# Hispanics: Visualize by Party ID
ggplot(thermometer, aes(x = ft_hisp)) +
  geom_histogram(bins = 20, fill = "magenta", color = "black") +
  geom_vline(data = party, aes(xintercept = mean_ft), color = "blue", linetype = "dashed") +
  geom_vline(data = party, aes(xintercept = median_ft), color = "orange", linetype = "dashed") +
  facet_wrap(~party_id, scales = "free_y") +
  labs(
    title = "Feeling Thermometer for Hispanics by Party ID",
    x = "Feeling Thermometer",
    y = "Frequency"
  ) +
  theme_minimal()
```

```
## Warning: Removed 146 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



EXPLAIN: All observations: The thermometer rating for Hispanics indicates that the mean is 70, the median is 75, and the standard deviation of 23.96. This shows that generally, people feel warmly towards Hispanics. For Party ID for Hispanics: The Democrats have a mean of 75, median of 80, standard deviation of 22. This shows that Democrats feel warmly towards Hispanics, even more than the population as a whole. The Republicans have a mean of 65, median of 69, and standard deviation of 22. This shows that Republicans feel less warmly towards Hispanics and even more so than the Democrats and population as a whole. The Independents have a mean of 69, median of 75, and standard deviation of 23.69. They still feel warm towards Hispanics, but land between Democrats and republicans. Those that are not sure have a mean of 73, median of 63, and standard deviation of 25.5 and those that identify as other have a mean of 73, median of 79.5, and standard deviation fo 23.

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

```
model = lm(ft_hisp ~ party_id, data=thermometer)
summary(model)
```

```
## 
## Call:
## lm(formula = ft_hisp ~ party_id, data = thermometer)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -75.237 -16.277   4.763  19.763  36.041 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          75.2371     0.5744 130.976  < 2e-16 ***
## party_idIndependent  -5.9597     0.8214  -7.256 4.63e-13 ***
## party_idNot sure    -11.2779     3.4192  -3.298 0.000979 ***
## party_idOther        -1.7452     2.1953  -0.795 0.426673    
## party_idRepublican  -10.3051     0.8584 -12.005  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 23.59 on 4838 degrees of freedom
##   (146 observations deleted due to missingness)
## Multiple R-squared:  0.03073,    Adjusted R-squared:  0.02993 
## F-statistic: 38.35 on 4 and 4838 DF,  p-value: < 2.2e-16
```

4. Create a new data frame and new binary variable.

```
# New data frame only for Democrat and Republican
D_R <- subset(thermometer, party_id %in% c("Democrat", "Republican"))
# New binary variable for party_id where Democrat = 1, Republican = 0
D_R$D_or_R <- ifelse(D_R$party_id == "Democrat", 1, 0)
```

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

Here, I'm using the feeling thermometer and race as an interaction term to predict how this interaction or rather, how race interacts with the feeling thermometer for the police to predict if you are a democrat or republican.

```
model <- lm(D_or_R~ft_police*race, data=D_R)
summary(model)
```

```
## 
## Call:
## lm(formula = D_or_R ~ ft_police * race, data = D_R)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.18808 -0.35845  0.03898  0.42323  0.85224 
## 
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                        0.938630   0.181449   5.173 2.45e-07 ***
## ft_police                         -0.003370   0.002655  -1.269  0.20446
## raceBlack                          0.057665   0.190631   0.302  0.76229
## raceHispanic                       0.227262   0.223694   1.016  0.30973
## raceMixed                          0.351011   0.246540   1.424  0.15462
## raceNative American                0.944802   0.457219   2.066  0.03887 *
## raceOther                          0.325508   0.349673   0.931  0.35198
## raceWhite                          0.249449   0.184309   1.353  0.17602
## ft_police:raceBlack                0.002200   0.002797   0.786  0.43166
## ft_police:raceHispanic            -0.004263   0.003103  -1.374  0.16961
## ft_police:raceMixed               -0.005717   0.003526  -1.621  0.10507
## ft_police:raceNative American -0.015471   0.005711  -2.709  0.00679 **
## ft_police:raceOther               -0.007794   0.004458  -1.748  0.08049 .
## ft_police:raceWhite               -0.005363   0.002684  -1.998  0.04577 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4426 on 3069 degrees of freedom
##   (63 observations deleted due to missingness)
## Multiple R-squared:  0.2129, Adjusted R-squared:  0.2096
## F-statistic: 63.85 on 13 and 3069 DF,  p-value: < 2.2e-16
```

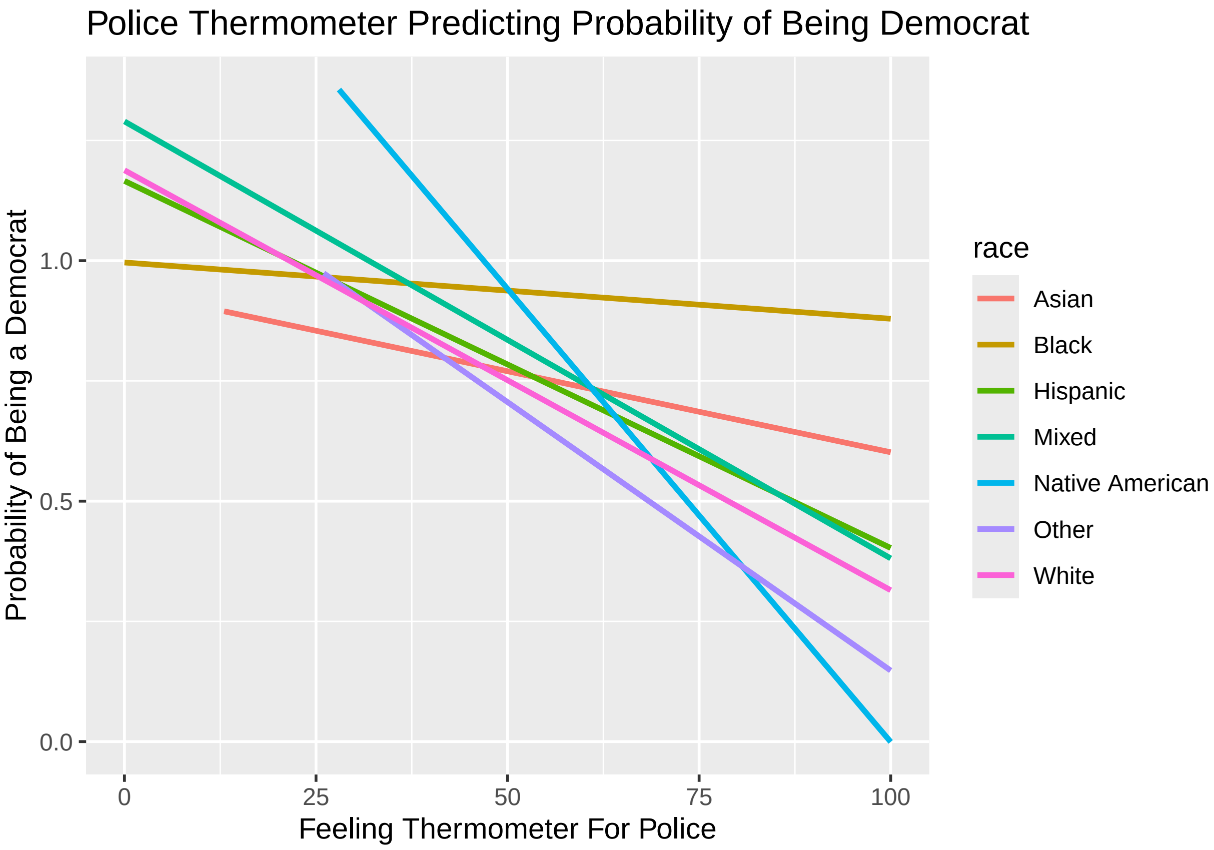
6. The coefficients in model represent the change in what? The changes in coefficients show that racial group is less likely to be a Democrat.
7. Select one of the feeling thermometers in your model and plot how your predicted values change as the feeling thermometer changes. Interpret your results. Can this reasonably be interpreted as a causal effect?

Yes, as the feeling thermometer increases to 100, the probability of being a democrat decreases regardless of race.

```
ggplot(data= D_R, mapping = aes(x=ft_police, y = D_or_R, color = race)) +
  geom_smooth(method = "lm", se=FALSE) + labs( x= "Feeling Thermometer For Police",
                                   y = "Probability of Being a Democrat",
                                   title = "Police Thermometer Predicting Probability of Being Democra
```

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

```
## Warning: Removed 63 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```


Police Thermometer Predicting Probability of Being Democrat
Probability of Being a Democrat
1.0
0.5
0.0
0
25
50
75
100
Feeling Thermometer For Police
race
Asian
Black
Hispanic
Mixed
Native American
Other
White