# Problem Set 3, POLS 602

Neeraj Tom Savio

2025-12-11

**Part 1:Reading**

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.
   A: The goal of the study is causal inference. The authors try to determine what causes civil wars: whether it is ethnic and religious grievances or whether it is conditions that allow for insurgencies such as weak state capacity, mountainous terrain, etc. The authors articulate that they are trying to reject the conventional understanding that religious and ethnically diverse states are more likely to descend into civil war, which helps in better understanding the goals of the study.
2. 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.

A: The theoretical estimand for hypothesis 1 is the average effect of diversity on higher risk of civil war. The empirical estimand is the regresssion coefficent of the authors' measure for the probabilistic chance of civil war.

The theoretical estimand for hypothesis 2 is the average causal effect of higher income per capita on the probability of civil war onset. The empirical estimand is the regression coefficient of the income per capita for the probabilistic chance of civil war.

The theoretical estimand for hypothesis 3 is the average causal effect of having an ethnic majority and a signficant ethnic minority on the probability for civil war. The empirical estimand is the regression coefficient of measures of ethnic groups for the probabilistic chance of civil war.

The theoretical estimand for hypothesis 4 is the average causal effect of political liberties on higher risk of civil war. The empirical estimand is the regression coefficient between measures of political liberties of countries and the probabilitic chance of civil war.

The theoretical estimand for hypothesis 5 is the average causal effect of discriminatory policies in states with minorities on the risk for civil war. The empirical estimand is the regression coefficient of measures of discriminatory policies in states with minorities on the probabilistic chance of civil war.

The theoretical estimand for hypothesis 6 is the average causal effect of greater income inequality on the chances for civil war. The empirical estimand is the regression coefficient of measure of income inequality on the probabilistic chance of civil war.

The theoretical estimand for hypothesis 7 is the average causal effect of greater ethnic diversity in countries with ethnic diversity greater than 5% on the risk of civil war. The empirical estimand is the regression coefficient of measures of greater ethnic diversity on the probabilistic chance of civil war.

The theoretical estimand for hypothesis 8 is the average causal effect of rough terrain poorly served by road at a distance from the centers of state power, availability of foreign sancturaries and the willingness of local population to work against the insurgents on the risk of civil war. The empirical estimand is the regression coefficient measures on the probabilistic chance of civil war.

The theoretical estimand for hypothesis 9 is the average causal effect of relative strength of the insurgents on the risk of civil war. The empirical estimand is the reggresion coefficient on the relative strength of the insurgency on the probabilistic chance of civil war.

The theoretical estimand for hypothesis 10 is the average causal effect of a newly independent state, political instability, government that mixes democratic with non-democratic features, large population, territories separated by water, a supportive diasopora, productive land, and oil producing state on the risk of civil war. The empirical estimand is the regression coefficient of these measures on the probabilitic chance of civil war.

The theoretical estimand for hypothesis 11 is the average causal effect of political democracy, the presence of civil liberties, higher income inequality, nondiscriminatory linguistic religious policies on the risk of civil war. The empirical estimand is the regression coefficient of these measures on the probabilistic chance of civil war.

3. The way you connect your theoretical estimand to your empirical estimand is known as identification—in other words, what does the research do to ensure that the empirical estimand is a good measure of the theoretical estimand? Describe the authors' identification strategy.

A: The authors attempt to connect their theoretical and empirical estimands by controlling for confounders. However, they don't convincingly control for every confounder.Their identification strategy could have been better

4. Provide an overall assessment of the paper and its conclusions. 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?

A: Overall the paper has some flaws. Its identification strategy is not robust as they do not convincingly control for all the confounders that could bias observational data. Their measurement strategy is also pretty weak as they do not convince that empirical estimands match the theoretical estimands. For example, their argument that per capita income is a proxy for the overall capacities of a state does not make sense.

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

A: The research performed is still relevant as it helps drive scholarly understanding of the factors that drive civil war. For example, they show that religious or ethnically diverse states do not have an an increased intrinsic propensity toward civil war.

### Part 2: Data Analysis

```
## Introductory Code

rm(list=ls()) #Clearing the environment
setwd("~/Desktop/POLS 602/R Files") #Setting up the working directory
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

1. Load the thermometers.csv data from the data folder on the github repo. Use the birth_year variable to create a new age variable (Note: This survey was taken in 2017).

```r
#Importing the data

thermometer <- read.csv("https://raw.githubusercontent.com/MLBurnham/pols_602/refs/heads/main/data/therm

thermometer$age <- 2025 - thermometer$birth_year

thm_cln <- thermometer %>%
  drop_na(ft_black, ft_white, ft_hisp, ft_asian, ft_muslim, ft_jew, ft_christ, ft_fem, ft_immig, ft_gays

head(thm_cln)
```

```
##   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       1959 Female Black    Democrat           Post-grad       98       70
## 6       1944   Male White Independent High school graduate       10       50
##   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      99      100       100    100       100     73      100      54        80
## 6      26       50         1     50        95     50        1       1         1
##   ft_police ft_altright ft_evang ft_dem ft_rep age
## 1        76           1       50     88     21  94
## 2        95          50       96     86     96  73
## 3        78           0        2     91     20  94
## 4        94          50       70     22     83  73
## 5        24           4       53     53      4  66
## 6        95          50       50      1     50  81
```

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
#Picking sex as the categorical demographic of interest
#Picking feeling thermometer ft_altright

#Mean of the selected feeling thermometer
avg_alt <- mean(thm_cln$ft_altright)
print(avg_alt)
```

```
## [1] 29.23568
```

```r
#Standard Deviation of the selected feeling thermometer
sd_alt <- sd(thm_cln$ft_altright)
print(sd_alt)
```

```
## [1] 29.44393
```

```r
#Disaggregating average feeling thermometer score by sex
avg_alt_m <- mean(thm_cln$ft_altright[thm_cln$sex == "Male"])
```

```r
print(avg_alt_m)
```

```
## [1] 30.58773
```

```r
avg_alt_f <- mean(thm_cln$ft_altright[thm_cln$sex == "Female"])
print(avg_alt_f)
```

```
## [1] 27.77046
```

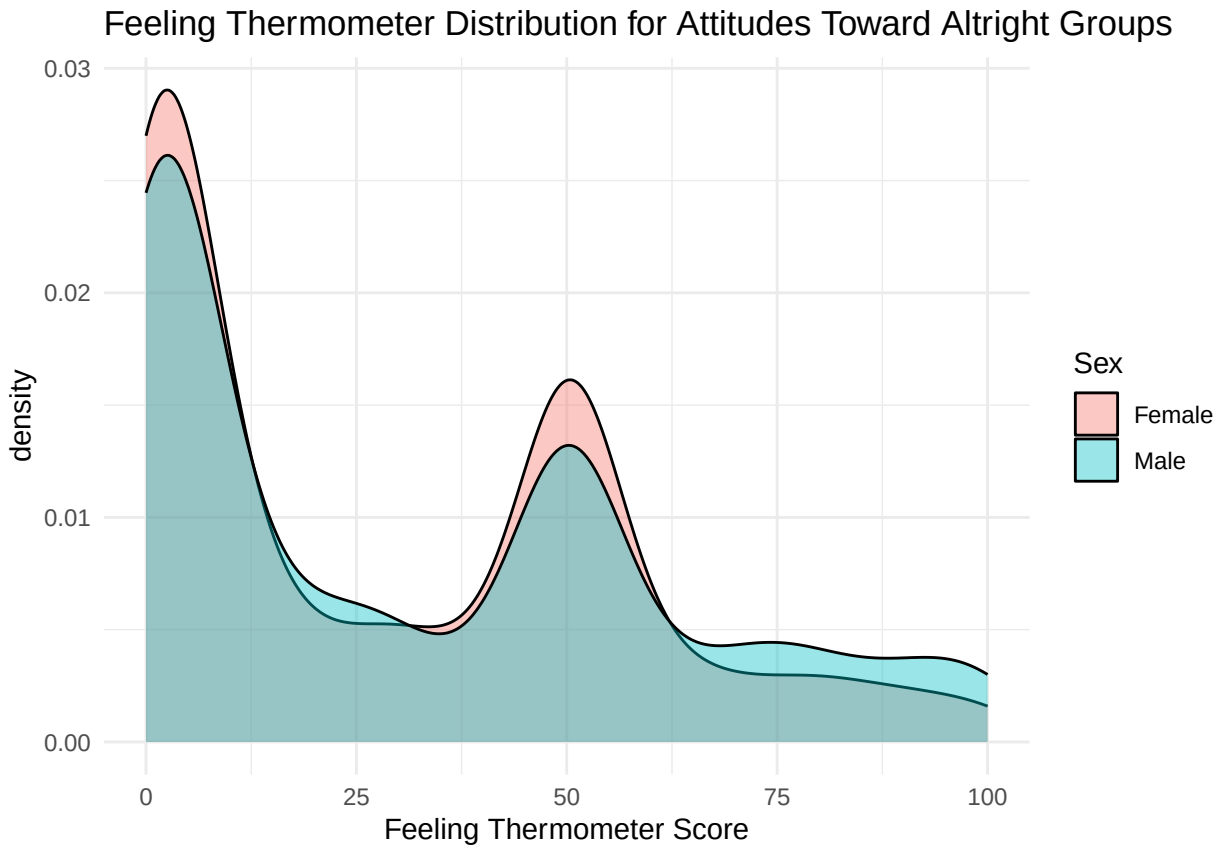
```r
#Obtaining standard deviation of the feeling thermometer for each demographic
#..category
sd_alt_m <- sd(thm_cln$ft_altright[thm_cln$sex == "Male"])
print(sd_alt_m)
```

```
## [1] 30.64753
```

```r
sd_alt_f <- sd(thm_cln$ft_altright[thm_cln$sex == "Female"])
print(sd_alt_f)
```

```
## [1] 28.01723
```

```r
#Using GGPLOT to create density plots of the distribution of feeling thermometer
#..scores disaggregated by sex
ggplot(thm_cln, aes(x = ft_altright, fill = sex)) +
  geom_density(alpha = 0.4) +
  theme_minimal() +
  labs (
    title = "Feeling Thermometer Distribution for Attitudes Toward Altright Groups",
    x = "Feeling Thermometer Score",
    fill = "Sex"
  )
```




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

```
model1 <- lm(thm_cln$ft_altright ~ thm_cln$sex)
summary(model1)
```

```
##
## Call:
## lm(formula = thm_cln$ft_altright ~ thm_cln$sex)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30.59 -26.77 -10.59  22.23  72.23
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      27.7705     0.7297  38.058  < 2e-16 ***
## thm_cln$sexMale   2.8173     1.0118   2.784  0.00539 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.41 on 3384 degrees of freedom
## Multiple R-squared:  0.002286,   Adjusted R-squared:  0.001991
## F-statistic: 7.753 on 1 and 3384 DF,  p-value: 0.005393
```

4. Create a new dataframe that only contains rows for Democrats and Republicans. Create a new binary

variable for party_id

```
thm_new <- thm_cln[!(thm_cln$party_id == "Independent"), ]
thm_new$dem <- ifelse(thm_new$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.

```
model2 <- lm(thm_new$dem ~ thm_new$ft_black + thm_new$age + thm_new$age*thm_new$sex)
summary(model2)
```

```
##
## Call:
## lm(formula = thm_new$dem ~ thm_new$ft_black + thm_new$age + thm_new$age *
##     thm_new$sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7775 -0.4863  0.2646  0.4222  0.9037
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.3355577  0.0771995   4.347 1.44e-05 ***
## thm_new$ft_black              0.0051218  0.0004292  11.933  < 2e-16 ***
## thm_new$age                  -0.0017553  0.0010485  -1.674   0.0942 .
## thm_new$sexMale              -0.0238265  0.1047197  -0.228   0.8200
## thm_new$age:thm_new$sexMale -0.0013367  0.0015365  -0.870   0.3844
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4782 on 2214 degrees of freedom
## Multiple R-squared:  0.08302,    Adjusted R-squared:  0.08136
## F-statistic: 50.11 on 4 and 2214 DF,  p-value: < 2.2e-16
```

Including an interaction term of age with sex in order to examine whether there is a differential impact of age on democratic party membership disaggregated by sex.

6. The coefficients in your model represent the change in what?
   A: The intercept denotes the likelihood that a female member with a feeling thermometer score of 0 toward the black community is associated with the democratic party. The thm_new$ft_black shows the change in likelihood of democratic association for every 1 score increase in the feeling thermometer toward the black community. The thm_new$age shows the change in democratic identification for a 1-unit increase in age for females. The thm_new$sexMale is the change in likelihood for males that they will identify with the democratic party. The thm_new$age:thm_new$sexMale shows the change in democratic identification for a 1-unit increase in age for males.

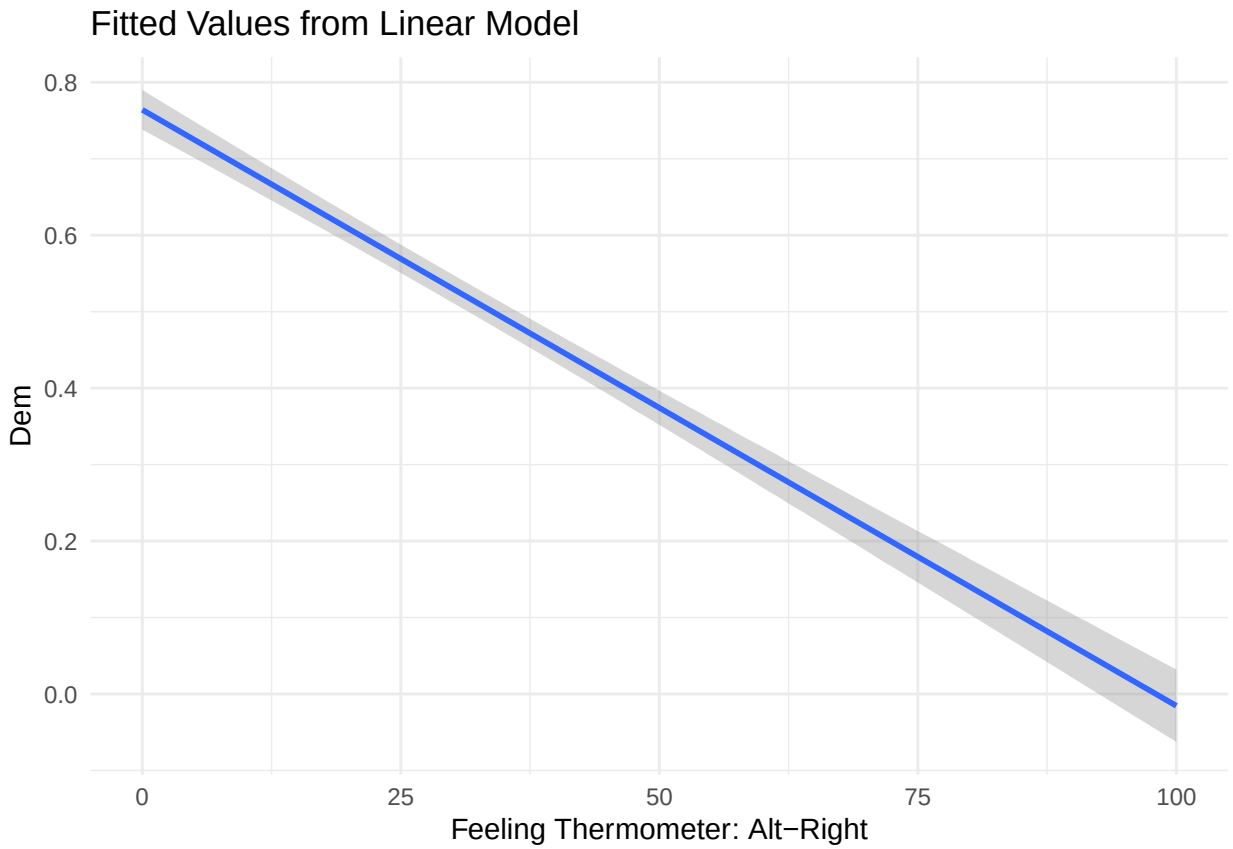
7. Select one of the feeling thermometers in your model and plot how your predicted values change as the feeling thermometer changes. Interpret your results. Can this reasonably be interpreted as a causal effect?

```
model3 <- lm(formula = thm_new$dem ~ thm_new$ft_altright)
summary(model3)
```

```
##
## Call:
## lm(formula = thm_new$dem ~ thm_new$ft_altright)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7641 -0.3743  0.2359  0.2671  1.0154
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.7640565  0.0131936   57.91   <2e-16 ***
## thm_new$ft_altright -0.0077949  0.0003151  -24.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4418 on 2217 degrees of freedom
## Multiple R-squared:  0.2163, Adjusted R-squared:  0.216
## F-statistic:   612 on 1 and 2217 DF,  p-value: < 2.2e-16
```

```
ggplot(thm_new, aes(x = ft_altright, y = dem)) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    x = "Feeling Thermometer: Alt-Right",
    y = "Dem",
    title = "Fitted Values from Linear Model"
  ) +
  theme_minimal()
```

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




We can't reasonably assume a causal effect as we can't be sure that all confounders have been controlled for and that there is no bias.