Problem Set 3

Ellie Choe

2025-11-18

Paper Analysis

1. Research Goals

The primary goal of this study is not purely causal inference but a combination of description, theory testing, and prediction. The authors aim to determine whether structural conditions favoring insurgency cause civil war onset, explicitly challenging conventional wisdom regarding post-Cold War instability and ethnic grievances. They successfully articulate their research objectives by contrasting their “insurgency” theory with existing “ethnic” models. However, while their theoretical goal is causal, their methodological framework lacks a formal definition of a causal estimand. They rely on observing predictive associations through logistic regression rather than strictly identifying causal effects using a potential outcomes framework.

2. Estimands

The theoretical estimand is the causal effect of insurgency-favoring conditions on civil war onset. However, the connection between their theoretical concepts and empirical estimands is tenuous. First, the hypothesis structure is diffuse; by testing numerous hypotheses (H1–H11) largely to debunk conventional wisdom, the specific causal quantity of interest becomes obscured. Second, measurement validity is a concern. “Rough terrain” is proxied solely by mountainous terrain, omitting other favorable terrains like swamps or jungles—an omission the authors themselves acknowledge. More critically, “Poverty” is measured by GDP per capita. While the authors argue this proxies for “state capacity” (police/military strength), it also conflates other mechanisms, such as the opportunity cost for rebels. Thus, the coefficient on GDP captures a composite effect rather than the specific effect of state capacity. To clarify their estimands, the authors could have explicitly defined the causal quantity of interest, such as the average treatment effect of weak state capacity on civil war probability, and justified their measurement strategies more rigorously.

3. Identification Strategy

Fearon and Laitin employ a large-N cross-country design covering 161 countries from 1945 to 1999. Their identification strategy relies primarily on the assumption of selection on observables, using logistic regression with statistical controls rather than an explicit causal identification mechanism. To mitigate endogeneity, specifically simultaneity bias, they use lagged independent variables. They also conduct extensive robustness checks using alternative datasets and regional controls. Despite these efforts, the strategy does not fully address reverse causality or omitted variable bias because no instrumental variable or natural experiment is used to isolate exogenous variation in insurgency conditions.

4. Assessment of Findings

The empirical findings largely support the authors' claims that structural factors like low income and rough terrain increase civil war risk, while ethnic diversity does not. However, the credibility of these findings is limited by validity concerns. First, interpreting the regression coefficients as strictly causal is problematic due to potential endogeneity. Second, the logistic regression model assumes a linear additive effect, which may oversimplify the complex, path-dependent data-generating process of civil war. Finally, regarding measurement, the proxy measures create mechanism indeterminacy. Since the data cannot distinguish whether the income effect is due to state weakness or rebel opportunity costs, the specific causal conclusion that "state weakness" is the primary driver is not fully supported by the data.

5. Broader Contribution

Despite these methodological limitations, this paper makes a seminal theoretical contribution. It shifted the analytical focus from "ethnic grievances" to "structural opportunities" for insurgency. By demonstrating that civil wars occur not because of diversity but because states are weak, it effectively debunked the view that ethnic partitions are a viable solution. This framework opened the door for future research on civil conflict, encouraging scholars and policymakers to focus on strengthening state administrative and police capabilities rather than solely on democratic reforms or ethnic rights.

Data Analysis

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
# load the data
thermometers <- read.csv("https://raw.githubusercontent.com/MLBurnham/pols_602/refs/heads/main/data/the
# create a new variable age
thermometers$age <- 2017-(thermometers$birth_year)
```

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.

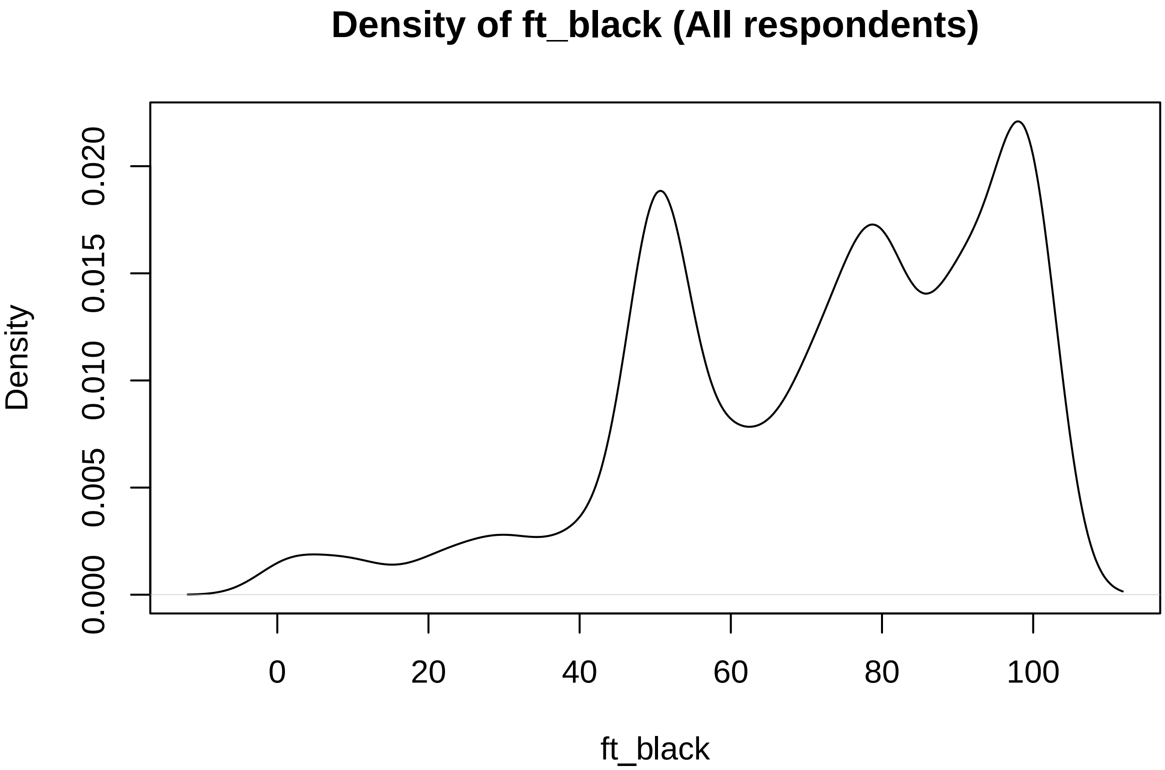
```r
# The ft_black thermometers have a mean of 71.33, a median of 76, a standard deviation of 23.98, and ra
summary(thermometers$ft_black)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   51.00   76.00   71.33   91.00  100.00     131
```

```r
sd(thermometers$ft_black, na.rm = TRUE)
```

```
## [1] 23.98103
```

```
# Visualize the distribution of all observations with a density plot
plot(density(thermometers$ft_black, na.rm = TRUE),
     main = "Density of ft_black (All respondents)",
     xlab = "ft_black", ylab = "Density")
```



```
# Spread and central tendency of ft_black by race (one of the categorical demographic variables)
library(dplyr)
```

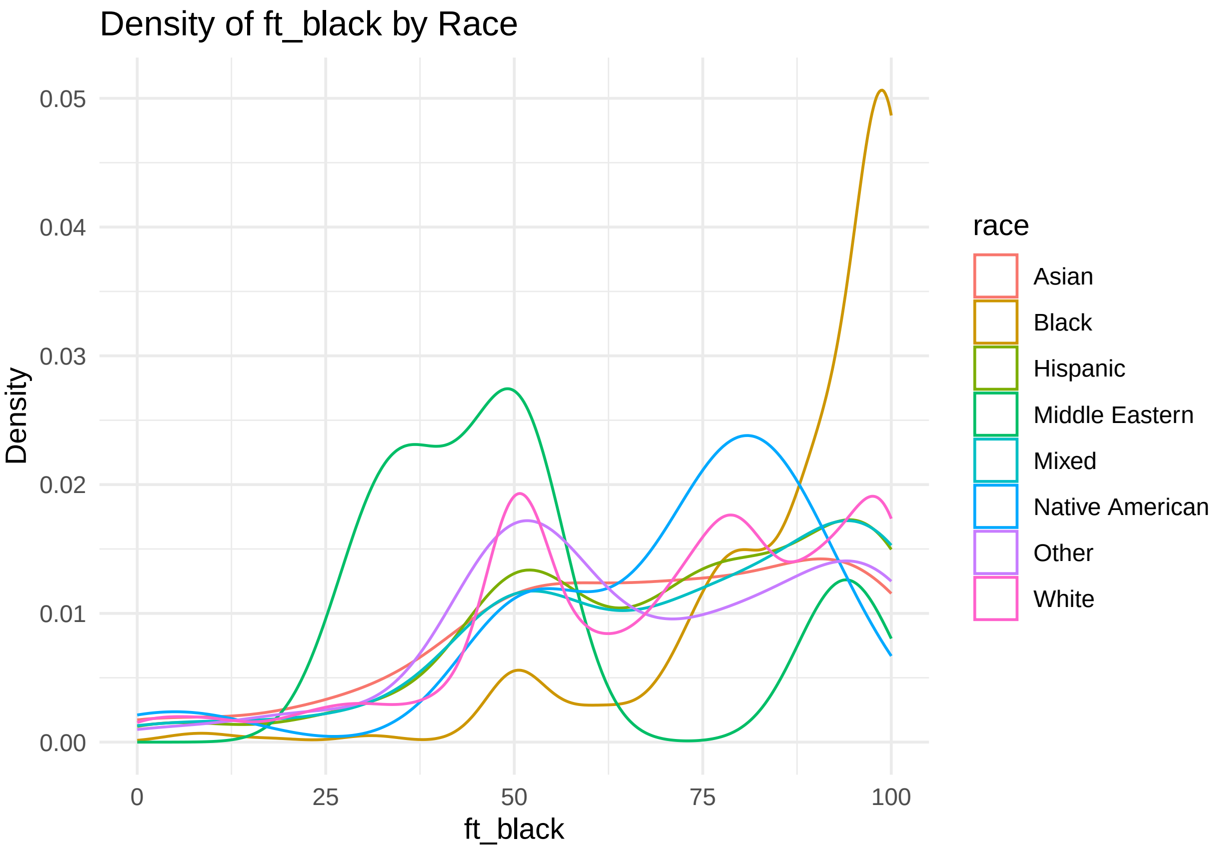
```
##
## Attaching package: 'dplyr'

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

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

```
race_summary <- thermometers %>%
  group_by(race) %>%
  summarise(
    mean = mean(ft_black, na.rm = TRUE),
    median = median(ft_black, na.rm = TRUE),
    sd = sd(ft_black, na.rm = TRUE),
    min = min(ft_black, na.rm = TRUE),
    max = max(ft_black, na.rm = TRUE)
  )
print(race_summary)
```

```
## # A tibble: 8 x 6
##   race              mean median    sd   min   max
##   <chr>            <dbl>  <dbl> <dbl> <int> <int>
## 1 Asian             68.0   71.5  25.7     0   100
## 2 Black             87.9   94    16.4     5   100
## 3 Hispanic          71.2   75    24.0     1   100
## 4 Middle Eastern    52.8   50    24.5    31    94
## 5 Mixed             72.2   78    25.1     0   100
## 6 Native American   69.5   78    22.0     0    97
## 7 Other             68.1   69    24.4     1   100
## 8 White             69.8   75    23.9     0   100
```

```
# Density plot showing the distribution of ft_black for each race
library(ggplot2)
ggplot(thermometers, aes(x = ft_black, color = race)) +
  geom_density() +
  labs(title = "Density of ft_black by Race",
       x = "ft_black", y = "Density") +
  theme_minimal() # Density plot showing the distribution of ft_black for each race
```

```
## Warning: Removed 131 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 demographic variable you chose.

```r
# Fit lm(ft_black ~ race) to estimate the conditional mean of ft_black for each racial category
lm_ft_race <- lm(ft_black ~ race, data = thermometers)
summary(lm_ft_race)
```

```
##
## Call:
## lm(formula = ft_black ~ race, data = thermometers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.904 -18.817   5.183  20.183  41.200
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           67.9595     2.7286  24.906  < 2e-16 ***
## raceBlack             19.9443     2.9732   6.708  2.2e-11 ***
## raceHispanic           3.2884     3.1304   1.050    0.294
## raceMiddle Eastern  -15.1595    10.8459  -1.398    0.162
## raceMixed              4.2250     3.5769   1.181    0.238
## raceNative American    1.4951     4.9133   0.304    0.761
## raceOther              0.1866     3.6926   0.051    0.960
## raceWhite              1.8579     2.7542   0.675    0.500
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 23.47 on 4850 degrees of freedom
##   (131 observations deleted due to missingness)
## Multiple R-squared:  0.04337,    Adjusted R-squared:  0.04199
## F-statistic: 31.41 on 7 and 4850 DF,  p-value: < 2.2e-16
```

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

```r
# Creat a new dataframe that only contains rows for Democrats and Republicans
df_demrep <- thermometers %>% filter(party_id %in% c("Democrat", "Republican"))
# Create a new binary variable: Republican = 1, Democrat = 0
df_demrep <- df_demrep %>% mutate(party_binary = ifelse(party_id == "Republican", 1, 0))

head(df_demrep)
```

```
##   birth_year    sex  race   party_id      educ ft_black ft_white ft_hisp
## 1       1931 Female White   Democrat    4-year       51       50      79
## 2       1952 Female White Republican    2-year       98       90      95
## 3       1952   Male White Republican    4-year       90       85      90
## 4       1939 Female White   Democrat    2-year      100       50     100
## 5       1959 Female Black   Democrat Post-grad       98       70      99
## 6       1969   Male White   Democrat    4-year       56       41      56
##   ft_asian ft_muslim ft_jew ft_christ ft_fem ft_immig ft_gays ft_unions
## 1       50        50     50        50     99       95      50        80
## 2      100        61    100        98     65       96      82        62
## 3       96        80     91        94     25       91      71        20
## 4      100       100    100        28    100      100     100       100
## 5      100       100    100       100     73      100      54        80
## 6       71        69     71        70    100       90     100        90
##   ft_police ft_altright ft_evang ft_dem ft_rep age party_binary
## 1        76           1       50     88     21  86            0
## 2        95          50       96     86     96  65            1
## 3        94          50       70     22     83  65            1
## 4        28          NA       NA     99     NA  78            0
## 5        24           4       53     53      4  58            0
## 6        60           0       25     90     10  48            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.

```
lm_interaction <- lm(party_binary ~ ft_fem * age, data = df_demrep)
summary(lm_interaction)
```

```
## 
## Call:
## lm(formula = party_binary ~ ft_fem * age, data = df_demrep)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.99849 -0.23292 -0.00322  0.20977  1.04191 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  8.276e-01  6.346e-02  13.041   <2e-16 ***
## ft_fem      -9.218e-03  9.573e-04  -9.628   <2e-16 ***
## age          2.408e-03  1.046e-03   2.302   0.0214 *  
## ft_fem:age  -7.731e-06  1.594e-05  -0.485   0.6278    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3767 on 2958 degrees of freedom
##   (184 observations deleted due to missingness)
## Multiple R-squared:  0.4258, Adjusted R-squared:  0.4252 
## F-statistic: 731.1 on 3 and 2958 DF,  p-value: < 2.2e-16
```

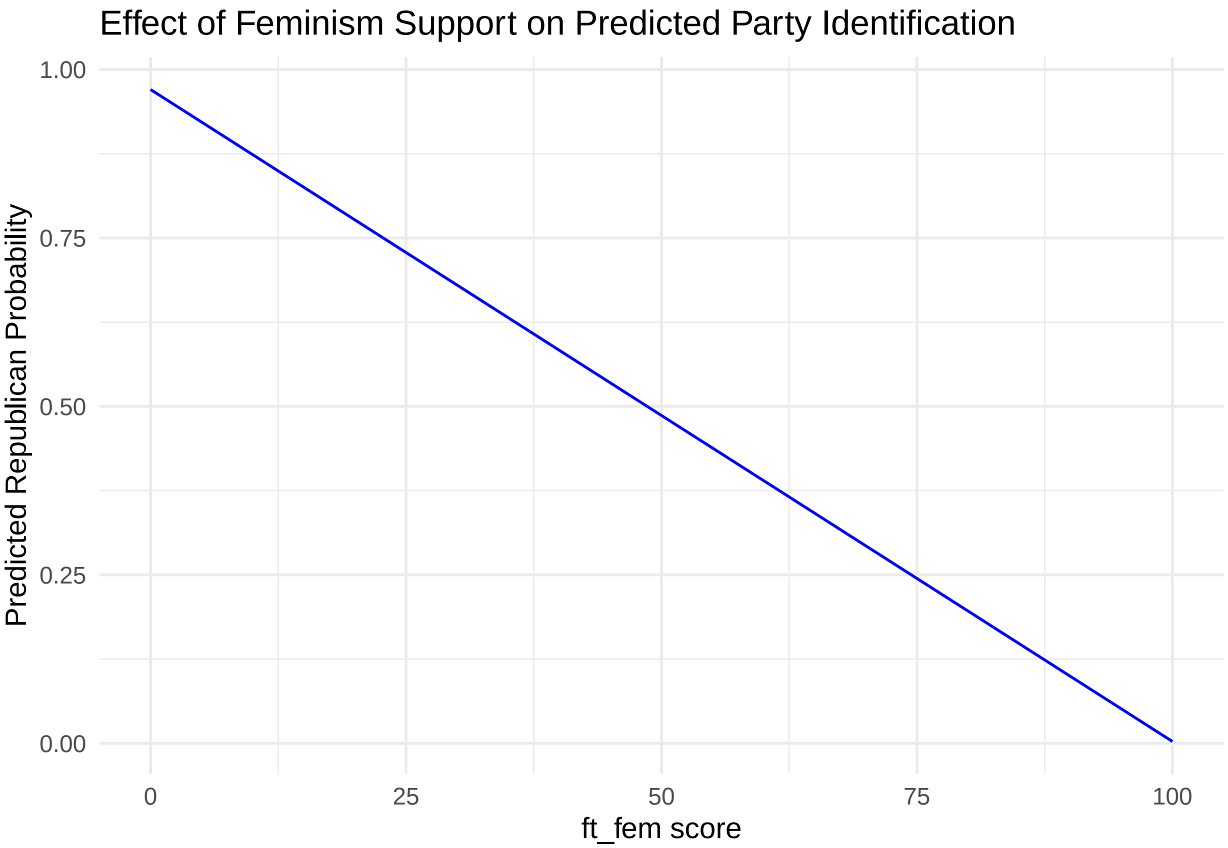
I include ft_fem and age because younger respondents and those with higher feminism support are more likely to identify as Democrats.The interaction term allows the effect of feminism support on party identification to vary by age, reflecting that the influence of feminist attitudes may be stronger among younger individuals.

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

The ft_fem coefficient (-0.009218) indicates that, when age is 0, a one-point increase in ft_fem decreases the probability of being Republican by about 0.92%, or equivalently, increases the chance of being a Democrat. The age coefficient (0.002408) shows that, when ft_fem is 0, each additional year of age increases the probability of being Republican by about 0.24%. The interaction term (ft_fem:age = -0.000007731) indicates that as age increases, the negative effect of ft_fem on Republican probability becomes slightly stronger, though the effect is very small and not statistically significant.

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?

```r
library(ggplot2)
mean_age <- mean(df_demrep$age, na.rm = TRUE)
ft_fem_seq <- seq(min(df_demrep$ft_fem, na.rm = TRUE),
                  max(df_demrep$ft_fem, na.rm = TRUE), length.out = 100)
data_predicted <- data.frame(ft_fem = ft_fem_seq,
                      age = mean_age)
data_predicted$predicted <- predict(lm_interaction, newdata = data_predicted)
ggplot(data_predicted, aes(x = ft_fem, y = predicted)) +
  geom_line(color = "blue") +
  labs(title = "Effect of Feminism Support on Predicted Party Identification",
       x = "ft_fem score",
       y = "Predicted Republican Probability") +
  theme_minimal()
```



The plot illustrates the predicted probability of being Republican as a function of feminism support (ft_fem), with age held constant at its mean value. As ft_fem increases, the predicted probability of being Republican decreases, indicating that respondents with higher feminism support are more likely to identify as Democrats. This plot depicts an association observed in the data, and should not be interpreted as evidence of a causal relationship between feminism support and party affiliation.