# prob set 3

2025-11-13

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     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
```

```
library(dplyr)
library(ggplot2)
library(ggeffects)
```

```
## Warning: package 'ggeffects' was built under R version 4.5.2
```

## Part 1

### 1.

The goal of this study is causal inference. Specifically, this study tries to falsify the conventional opinion that civil wars are caused by ethnic and religious diversity, and to propose an alternative explanation based on conditions that favor insurgency. However, although the authors clearly stated the puzzle, target of falsification and alternative theory, it seems that they did not clearly state whether they are seeking prediction or causal relation. (And their method also did not reach the standard of causality)

### 2.

Here are the theoretical and empirical estimands:

| | |
|---|---|
| The effect of ethnic/religious diversity | The coefficient on **Ethnolinguistic Fractionalization (ELF)** and **Religious Fractionalization**. |
| The effect of ethnic polarization | The coefficient on a **dummy variable** for countries with a dominant majority and a significant minority. |
| The effect of political grievances | The coefficient on the **Polity IV democracy score** or the **Freedom House civil liberties score**. |
| The effect of state discrimination | The coefficient on **dummies for discriminatory language/religious policies** . |
| The effect of economic grievances | The coefficient on the **Gini coefficient** (income inequality). |

| | |
|---|---|
| The effect of State Capacity | The coefficient on (lagged) **Per Capita Income**. |
| The effect of favorable terrain | The coefficient on **Percentage of Mountainous Terrain**. |
| The effect of a large, hard-to-control territory/population | The coefficient on **log(Population)**. |
| The effect of peripheral or separated territory | The coefficient on the **"Noncontiguous State"** dummy. |
| The effect of state weakness/instability | The coefficient on the **"New State"** dummy and the **"Instability"** dummy. |
| The effect of a state apparatus weakened by resource reliance | The coefficient on the **"Oil Exporter"** dummy |

Overall, the authors clearly stated their empirical and theoretical estimand, but there is still some ambiguity in the definition of state capacity effect. In the article, the state capacity is measured by logged income per capita, however, it represents 3 mechanisms (state capacity, rebel opportunity cost and infrastructure) at the same time. Further clarification is needed by finding suitable measurement for the 3 mechanisms respectively.

### 3.

The identification strategy multivariate logit analysis. Basically the authors try to add as much control variables as they can. However, this strategy can not tackle the unobserved variables, as well as the problem of endogeneity. Specifically, they cannot argue that their independent variable–income per capita is not influenced by the dependent variable and omitted variables. For example, civil war could lead to the economic breakdown and lower the income, which is the reversed causality.

### 4.

This article has theoretical contribution in related field, but its empirical strategy is not satisfactory under modern standards of political science.

Firstly, the identification strategy does not fully support the authors' causal claims, and the coefficients cannot be credibly interpreted as causal effects, as there is significant endogeneity problem.

Secondly, the multivariate logit model cannot adequately represent the DGP in real world. It ignores some critical problems like the feedback loop and the spacial spillover effect.

Lastly, the data mostly credibly measures the phenomena being studied, despite the bundled measurement and mechanisms of state capacity and income.

### 5.

Despite the methodological weakness, this paper contribute to the knowledge of political science by providing a shift of paradigm in the research of civil war, namely from ethinity and religion state capacity. This contribution allows researchers to consider civil war as a solvable, structural problem of political economy, state capacity, and governance.

## Part 2

### 1.

```r
thermo <- read.csv("https://raw.githubusercontent.com/MLBurnham/pols_602/refs/heads/main/data/thermomet

#create age var.
```

```r
thermo <- thermo %>%
  mutate(age = 2017-birth_year)
```

## 2.

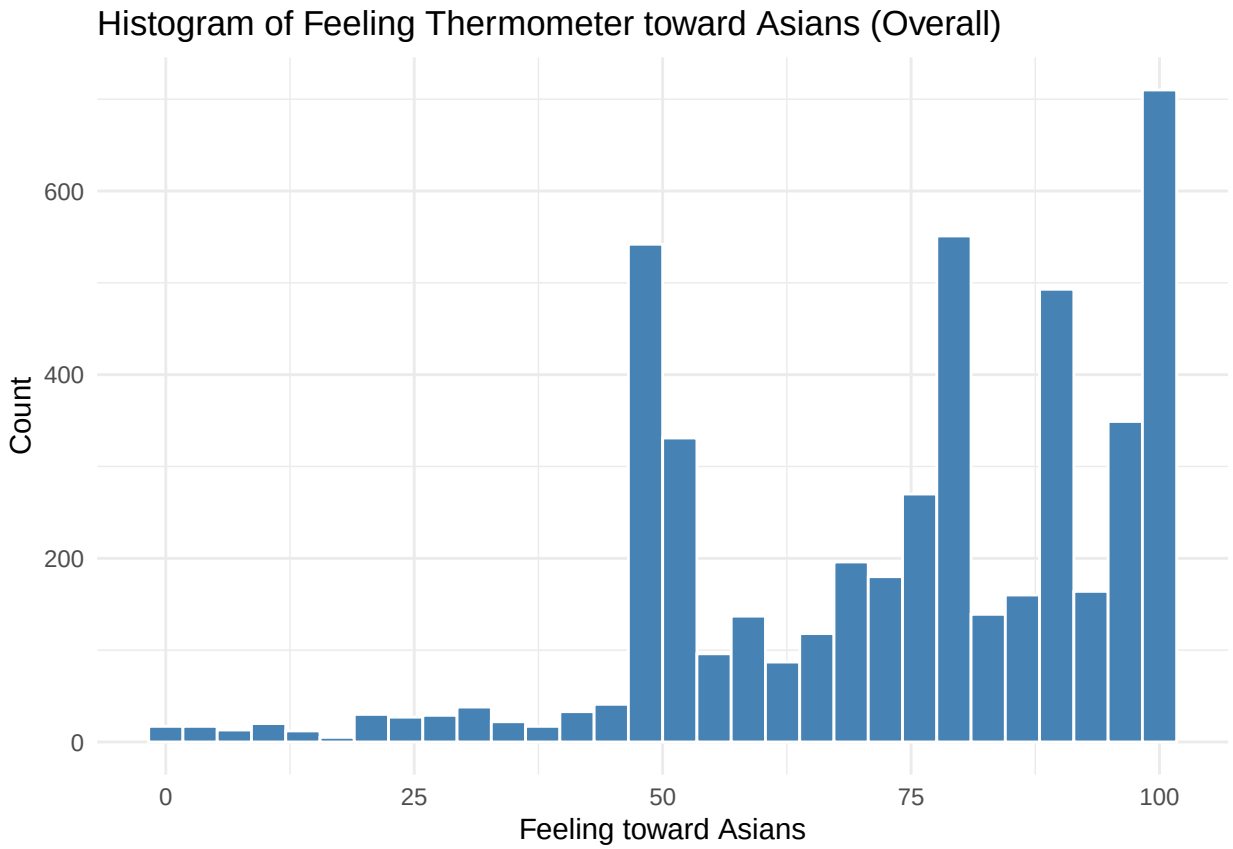
```r
#summarize the mean, median and sd of feeling on asian, for whole population
sum1 <- thermo %>%
  summarise(
    mean_ft = mean(ft_asian, na.rm = TRUE),
    median_ft = median(ft_asian, na.rm = TRUE),
    sd_ft = sd(ft_asian, na.rm = TRUE),
    n = n()
  )
print(sum1)
```

```
##    mean_ft median_ft    sd_ft    n
## 1 74.18704        79 21.56114 4989
```

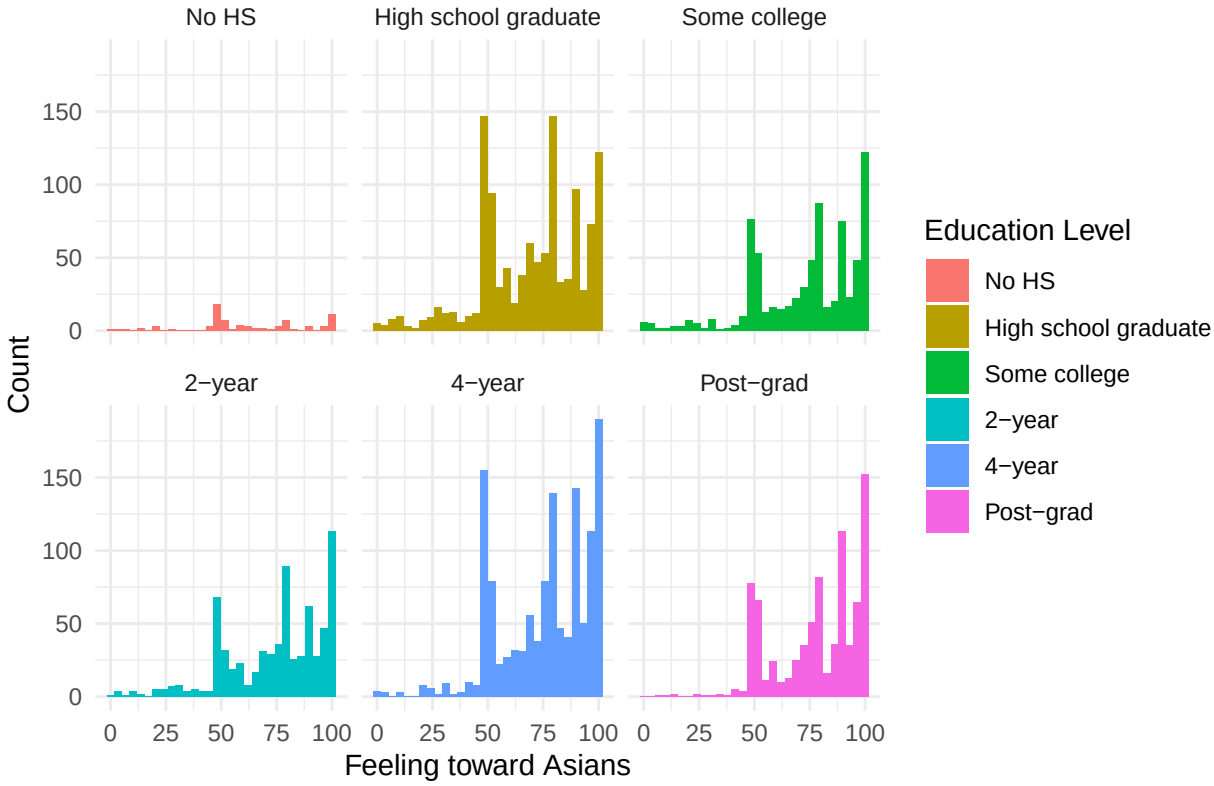
```r
#Relevel the educ variable so that it's ranked correctly
thermo$educ <- factor(thermo$educ,
                      levels = c("No HS", "High school graduate",
                                 "Some college", "2-year", "4-year","Post-grad"))

#summarize the mean, median and sd of feeling on asian, by different educ lvl
sum2 <- thermo %>%
  group_by(educ) %>%
  summarise(
    mean_ft = mean(ft_asian, na.rm = TRUE),
    median_ft = median(ft_asian, na.rm = TRUE),
    sd_ft = sd(ft_asian, na.rm = TRUE),
    n = n()
  )
print(sum2)
```

```
## # A tibble: 6 x 5
##   educ                  mean_ft median_ft sd_ft     n
##   <fct>                   <dbl>     <dbl> <dbl> <int>
## 1 No HS                    62.4        59  25.7    84
## 2 High school graduate     69.9        74  22.4  1223
## 3 Some college             73.7        79  22.8   763
## 4 2-year                   74.8        80  21.8   728
## 5 4-year                   76.0        80  20.3  1342
## 6 Post-grad                78.5        82  18.9   849
```

```r
#histogram for whole population
ggplot(thermo %>% filter(!is.na(ft_asian)), #remove na
       aes(x = ft_asian)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 30) +
  labs(
    title = "Histogram of Feeling Thermometer toward Asians (Overall)",
    x = "Feeling toward Asians",
    y = "Count"
  ) +
  theme_minimal()
```




```
#histogram by different educ lvl
ggplot(thermo %>% filter(!is.na(ft_asian)),
    aes(x = ft_asian, fill = educ)) +
    geom_histogram(bins = 30) +
    facet_wrap(~educ)+
    labs(
    title = "Histogram of Feeling Thermometer toward Asians by Education Level",
    x = "Feeling toward Asians",
    y = "Count",
    fill = "Education Level"
  ) +
    theme_minimal()
```




**3.**

```
#fit the regression
model1 <- lm(ft_asian ~ educ, data = thermo)
summary(model1)
```

```
##
## Call:
## lm(formula = ft_asian ~ educ, data = thermo)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75.98 -18.54   4.18  18.02  37.56
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 62.436      2.413  25.869  < 2e-16 ***
## educHigh school graduate     7.428      2.492   2.981  0.00289 **
## educSome college            11.247      2.537   4.433 9.52e-06 ***
## educ2-year                  12.384      2.543   4.871 1.15e-06 ***
## educ4-year                  13.549      2.485   5.453 5.21e-08 ***
## educPost-grad               16.100      2.524   6.379 1.95e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.32 on 4838 degrees of freedom
```

```
##   (145 observations deleted due to missingness)
## Multiple R-squared:  0.02367,    Adjusted R-squared:  0.02266
## F-statistic: 23.46 on 5 and 4838 DF,  p-value: < 2.2e-16
```

**4.**

```r
#filter for new dataset
thermo1 <- thermo %>%
  filter(party_id %in% c("Democrat","Republican"))

#create binary variable based on party id
thermo1 <- thermo1 %>%
  mutate(party_bi = ifelse(party_id == "Democrat" ,1,0))
```

**5.**

```r
#adjust the reference group of race
thermo1$race <- factor(thermo1$race,
                  levels  = c("White","Black","Hispanic",
                              "Asian","Mixed","Native American","Other"))

#fit model
model2 <- glm(party_bi ~  ft_gays * age + race + sex + educ,
              data = thermo1,
              family = binomial(link = "logit"))
summary(model2)
```

```
##
## Call:
## glm(formula = party_bi ~ ft_gays * age + race + sex + educ, family = binomial(link = "logit"),
##     data = thermo1)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.6937013  0.6507282  -4.140 3.48e-05 ***
## ft_gays                     0.0549229  0.0077015   7.131 9.93e-13 ***
## age                         0.0133216  0.0086696   1.537 0.124394
## raceBlack                   3.1334190  0.2526595  12.402  < 2e-16 ***
## raceHispanic                0.3810148  0.2000504   1.905 0.056833 .
## raceAsian                   1.1185949  0.4177507   2.678 0.007414 **
## raceMixed                   0.8323733  0.3401426   2.447 0.014400 *
## raceNative American        -0.1900952  0.5618192  -0.338 0.735094
## raceOther                  -0.3821948  0.4144245  -0.922 0.356408
## sexMale                    -0.3154972  0.0902520  -3.496 0.000473 ***
## educHigh school graduate -0.3532663  0.3351923  -1.054 0.291919
## educSome college           -0.3965066  0.3473058  -1.142 0.253594
## educ2-year                 -0.4197997  0.3450576  -1.217 0.223754
## educ4-year                 -0.3110635  0.3384387  -0.919 0.358036
## educPost-grad              -0.3657367  0.3442791  -1.062 0.288088
## ft_gays:age                -0.0002674  0.0001241  -2.154 0.031211 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4145.5  on 3019  degrees of freedom
## Residual deviance: 3078.5  on 3004  degrees of freedom
##   (126 observations deleted due to missingness)
## AIC: 3110.5
##
## Number of Fisher Scoring iterations: 5
```

This model tests the influence of feeling towards homosexual on predicting party identity. Theoretically, the attitude on LGBTQ+ people is a critical distinction between Democrats and Republicans. Thus I suppose the feeling towards homosexual can be a good predictor on party identity.

Besides, the attitude difference between different age groups of voters are often observed. Thus I choose age as a interaction term as I could examine whether the influence of feeling towards homosexual is different among age groups.

I added race, sex and education as control variable to avoid omitted variable bias. These are standard controls for demographics.

I use Logistic Regression as the dependent variable is a binary variable. We can make sure the predicted probability is valid compared with linear regression.

## 6.

The coefficient of $ft\_gays$ is about 0.0549, and is significant at p<0.001 level. It represents that having one point higher of feeling thermometers on homosexual people can predict an increase of 0.0549 log odds in probability of being a Democrat.

The coefficient of interaction term $ft\_gays * age$ is about -0.0002, and is significant at p<0.05 level. It represents that the positive effect of feeling thermometers on homosexual people on probability of being a Democrat, decreases about -0.0002 log odds when age increases one year.

The coefficient of $race : Black$ is about 3.1334, and is significant at p<0.001 level. It represents that being Black people can predict an increase of 3.1334 log odds in probability of being a Democrat, compared to White people.

The coefficient of $race : Asian$ is about 1.1185, and is significant at p<0.01 level. It represents that being White people can predict an increase of 1.1185 log odds in probability of being a Democrat,compared to White people.

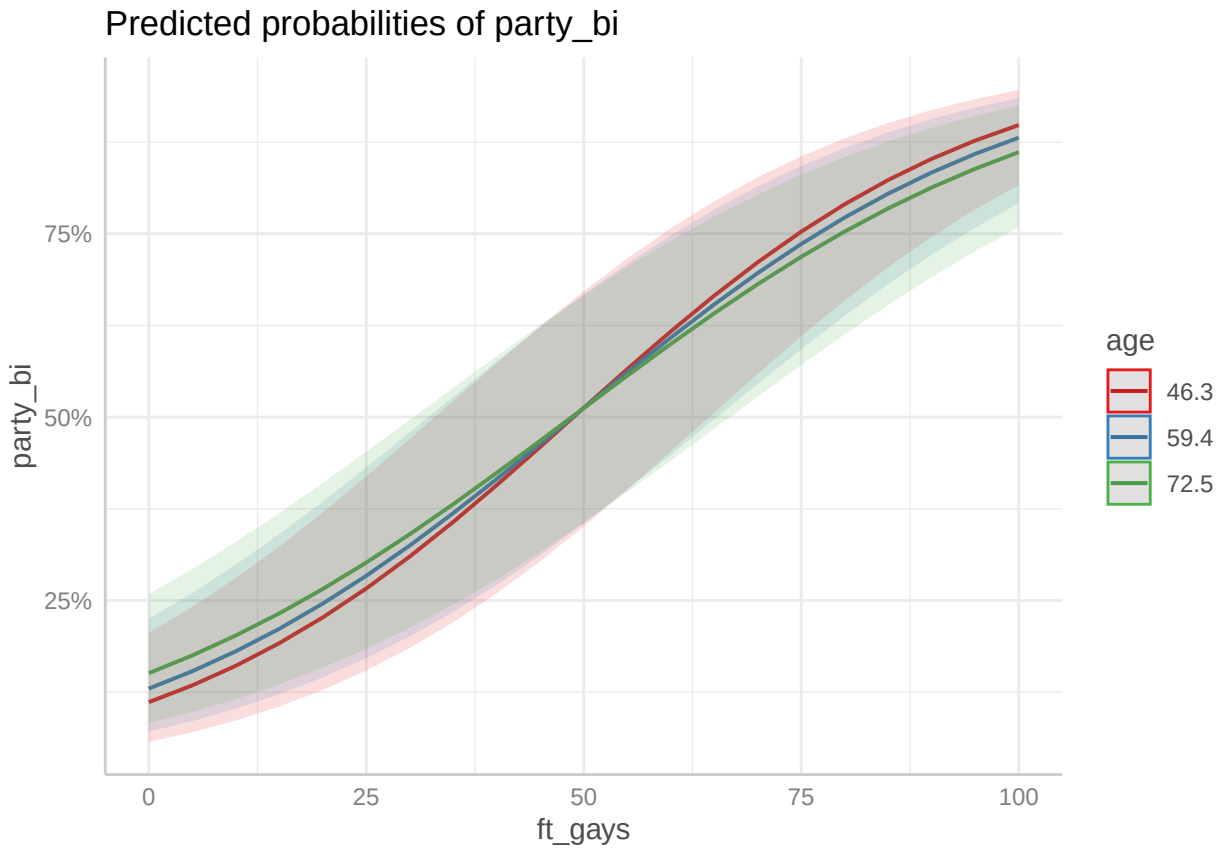
The coefficient of $sex : Male$ is about -0.3154, and is significant at p<0.001 level. It represents that being male can predict an decrease of 0.3154 log odds in probability of being a Democrat,compared to female people.

## 7.

```
#plot the predicted value
pred <- ggpredict(model2, terms = c("ft_gays", "age"))
```

```
## Data were 'prettified'. Consider using `terms="ft_gays [all]"` to get
##   smooth plots.
```

```
plot(pred)
```




The figure shows that as $ft\_gays$, the feeling thermometer towards homosexual increases, the probability of being a Democrats also increases. Further, we can find the line for the youngest group (Age 46.3, red line) has the steepest slope, while the line for the oldest group (Age 72.5, green line) is the flattest. The figure shows that the influence of feeling towards homosexual on party identity is conditioned on age. The feeling thermometer has a stronger effect on the probability of being a Democrat for younger individuals than elderly people.

Beside, this cannot be interpreted as a causal effect. Because we only examined the correlation, without any causal identification in this simple logit model. Specifically, it is based on observational data and cannot exclude the omitted variable bias and endogeneity.