Problem Set 3

Santiago Vidal Calvo

2025-12-07

```
# Question 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.

#

# Answer 1:

# The main goal of Fearon and Laitin's study is to explain why civil wars

# > occur where and when they do, rather than just describing patterns or

# > predicting future conflicts.

# They compare different explanations for civil war, especially the idea

# > that ethnic and religious diversity causes conflict versus the idea

# > that conditions favoring insurgency (like poverty and weak states)

# > matter more.

# So their goal is essentially causal inference about the determinants of

# > civil war onset, even though they do not use modern causal language.

# The authors clearly say they are challenging the "ethnic conflict"

# > conventional wisdom, which makes their objective easy to understand,

# > but they do not formally define a causal estimand in the way newer

# > papers sometimes do.

# Question 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.
#
# Answer 2:
# The paper does not use the word "estimand," but we can infer what they
# > care about.
# The theoretical estimand is basically the causal effect of changing a
# > country's characteristics (income, state capacity, ethnic diversity,
# > terrain, population, etc.) on the probability of a civil war beginning.
# The empirical estimands are the coefficients in their logit regressions,
# > which they treat as estimates of how each variable changes the log
# > odds of civil war onset, holding the others constant.
# They could be clearer by stating explicitly which causal effect each
# > coefficient is supposed to represent, and by separating descriptive
# > patterns from the causal parameters they hope to learn about.
# Question 3: Identification Strategy
# 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.
#
# Answer 3:
# Since the data are observational, Fearon and Laitin rely on regression
# > with control variables as their identification strategy.
# They include many potential confounders, like income, regime type,
```

```
# > mountains, population, political instability, oil exports, and
# > whether a state is new.
# The idea is that once these controls are in the model, the coefficient
# > on a variable like income can be interpreted as the causal effect of
# > income on civil war risk.
# In other words, they assume that controlling for this long list of
# > variables removes important sources of bias.
# They do not use tools like natural experiments or instrumental
# > variables, so the identification argument depends heavily on the
# > assumption that there are no major omitted confounders and that the
# > functional form of the model is appropriate.
# Question 4: Assessment of Findings
# 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?
#
# Answer 4:
# The main conclusion is that conditions favoring insurgency-especially
# > low income and weak states-predict civil war much better than ethnic
# > or religious diversity does.
# This conclusion is supported by their regressions, which show strong
# > and robust effects for income, political instability, terrain, and
# > population, while ethnic fractionalization becomes small and
# > insignificant once those factors are controlled for.
```

```
# Given their approach, I think the identification strategy supports
# > their claims in a broad, qualitative sense.
# It is still hard to treat the coefficients as precise causal effects,
# > because we cannot rule out omitted variables or measurement error in
# > the civil war data.
# Overall, the model seems like a reasonable approximation to the data
# > generating process, and the data are good by the standards of
# > cross-national conflict research, but the results remain correlational
# > rather than definitively causal.
# Question 5: Broader Contribution
# Despite any weaknesses, can this research still inform our
# > understanding of the world? If so, how?
#
# Answer 5:
# Even with its limitations, the paper makes an important contribution.
# It shifts attention away from "ancient ethnic hatreds" and toward state
# > weakness, poverty, and the logistics of insurgency as key drivers of
# > civil war.
# That change in focus has influenced a lot of later work and remains
# > part of how scholars and policymakers think about internal conflict.
# The paper also shows how careful large-N analysis can challenge
# > popular stories that seem intuitive but do not hold up in the data.
# So even if the identification is not perfect, the study still improves
# > our understanding of civil wars by reframing what we should be
# > looking at when we ask why some countries experience these conflicts.
```

```
# Question 1
# Load the thermometers.csv data from the data folder on the github
```

```r
# > repo. Use the birth_year variable to create a new age variable

# > (Note: This survey was taken in 2017).

#

# We load the data and create an age variable using birth_year.

thermometers <- read.csv("/Users/santividal5/Desktop/R/thermometers.csv")

thermometers$party_id <- factor(thermometers$party_id)
thermometers$sex <- factor(thermometers$sex)
thermometers$race <- factor(thermometers$race)
thermometers$educ <- factor(thermometers$educ)

thermometers$age <- 2017 - thermometers$birth_year

head(thermometers[, c("birth_year", "age")])
```

```
##   birth_year age
## 1       1931  86
## 2       1952  65
## 3       1931  86
## 4       1952  65
## 5       1939  78
## 6       1959  58
```

```r
# This head() output lets me check that someone born in 1950 appears as

# > age 67 in 2017, and so on, which confirms that age was computed

# > correctly.
```

```r
# Question 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.

#

# I use ft_immig (feelings toward immigrants) and party_id.

summary(thermometers$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
sd(thermometers$ft_immig, na.rm = TRUE)
```

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

```r
# The summary shows the minimum, quartiles, median, mean, and maximum,

# > and the standard deviation tells me that ratings are quite spread

# > out, with many people giving very low or very high scores.

by(
thermometers$ft_immig,
thermometers$party_id,
summary
)
```

```
## thermometers$party_id: Democrat
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   53.00   77.00   71.66   90.00  100.00      63
## ------------------------------------------------------------
## thermometers$party_id: Independent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   50.00   62.00   61.54   81.00  100.00      60
## ------------------------------------------------------------
## thermometers$party_id: Not sure
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   42.25   54.00   56.88   77.50  100.00       7
## ------------------------------------------------------------
## thermometers$party_id: Other
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   50.00   69.00   65.39   88.00  100.00       7
## ------------------------------------------------------------
## thermometers$party_id: Republican
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    30.0    50.0    50.2    71.0   100.0      60
```

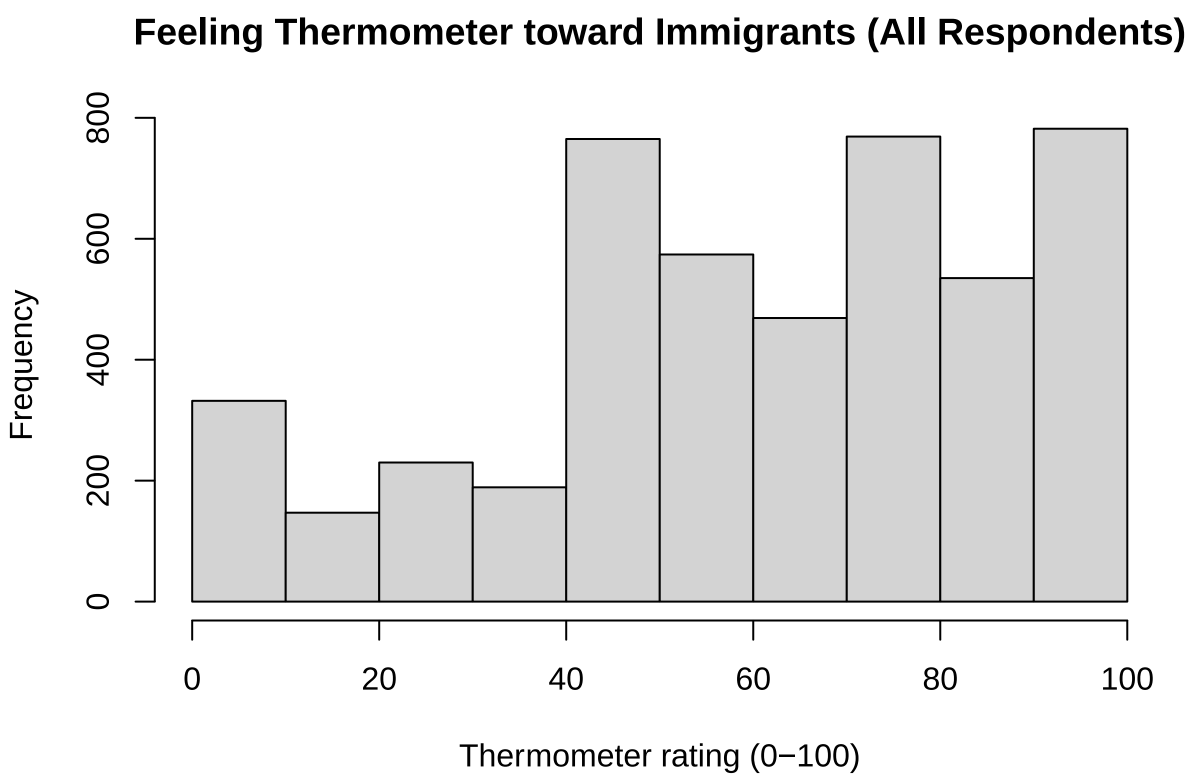
```r
# This by() output shows the distribution for each party group.

# Democrats tend to have higher average ratings toward immigrants,

# > Republicans tend to have lower ratings, and Independents and other

# > groups fall in between.

party_means <- tapply(
thermometers$ft_immig,
thermometers$party_id,
mean,
na.rm = TRUE
)

party_sds <- tapply(
thermometers$ft_immig,
thermometers$party_id,
sd,
```

```r
na.rm = TRUE
)

cbind(mean = party_means, sd = party_sds)
```

```
##                 mean       sd
## Democrat    71.65829 23.75100
## Independent 61.53880 26.46124
## Not sure    56.87500 26.28860
## Other       65.39024 26.22183
## Republican  50.20192 27.46507
```

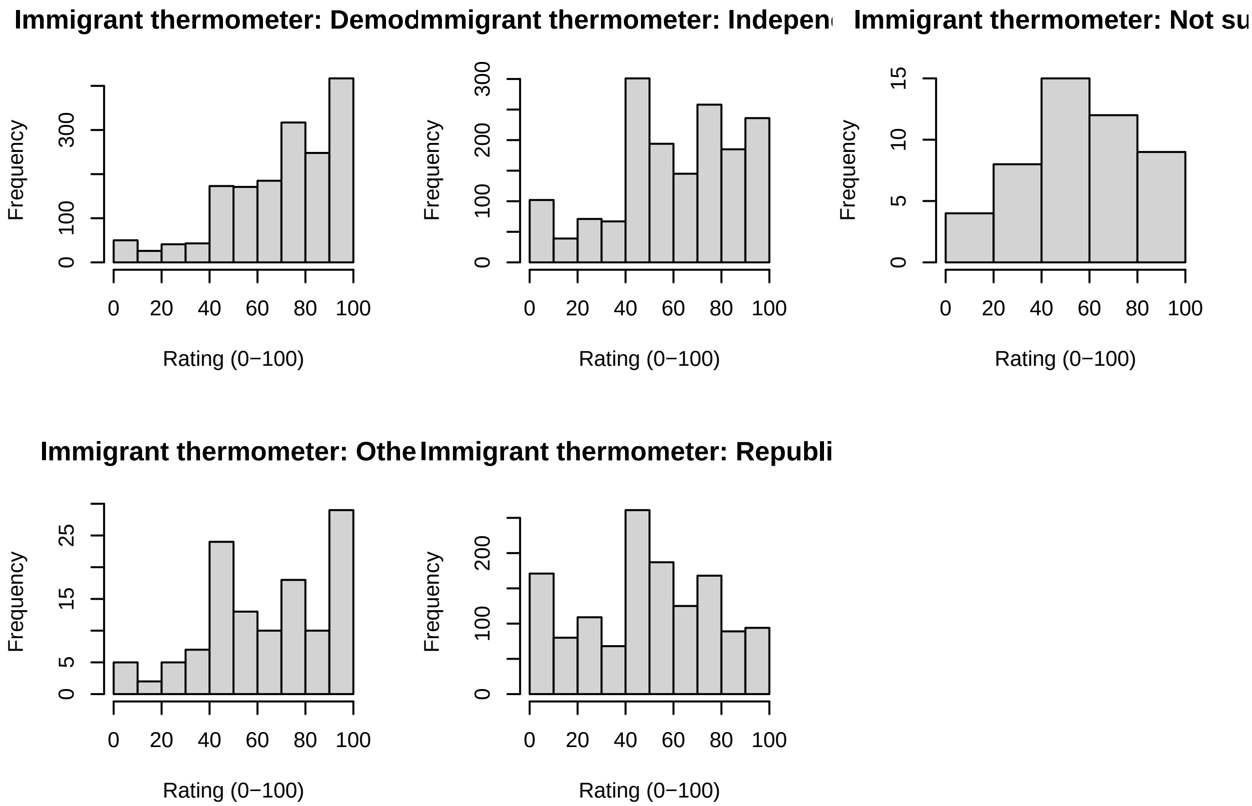
```r
# This small table makes it easy to compare average warmth and

# > variability toward immigrants across the different party_id

# > categories.
```

```r
# Now we plot histograms for the feeling thermometer.

hist(
thermometers$ft_immig,
main = "Feeling Thermometer toward Immigrants (All Respondents)",
xlab = "Thermometer rating (0-100)"
)
```




```r
par(mfrow = c(2, 3))

for (p in levels(thermometers$party_id)) {
hist(
thermometers$ft_immig[thermometers$party_id == p],
```

```
main = paste("Immigrant thermometer:", p),
xlab = "Rating (0-100)",
xlim = c(0, 100)
)
}

par(mfrow = c(1, 1))
```




```
# The overall histogram shows many scores around 50 and many near 100,
# > with some very low values, which suggests polarization.
# The separate histograms show that Democrats cluster more at high scores
# > and Republicans have more low to mid-range scores, which matches the
# > differences we saw in the summary statistics.
```

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

```r
# We regress ft_immig on party_id so each coefficient is a difference

# > in group means relative to the baseline party.

model1 <- lm(ft_immig ~ party_id, data = thermometers)

summary(model1)
```

```
##
## Call:
## lm(formula = ft_immig ~ party_id, data = thermometers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.658 -16.202   2.342  19.461  49.798
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          71.6583     0.6321 113.371  < 2e-16 ***
## party_idIndependent -10.1195     0.9040 -11.194  < 2e-16 ***
## party_idNot sure    -14.7833     3.7825  -3.908 9.42e-05 ***
## party_idOther        -6.2680     2.4139  -2.597  0.00944 **
## party_idRepublican  -21.4564     0.9451 -22.702  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.84 on 4787 degrees of freedom
##   (197 observations deleted due to missingness)
## Multiple R-squared:  0.09796,    Adjusted R-squared:  0.09721
## F-statistic:   130 on 4 and 4787 DF,  p-value: < 2.2e-16
```

```r
# In this model the intercept is the average immigrant thermometer

# > score for the baseline party (usually Democrats).

# Each party_id coefficient tells us how much higher or lower that

# > party's mean is compared to Democrats.

# For example, the coefficient for Republicans is large and negative,

# > which means Republicans score immigrants much lower on average than

# > Democrats do.
```

```r
# Question 4

# Create a new dataframe that only contains rows for Democrats and

# > Republicans. Create a new binary variable for party_id.

#

# We keep only Democrats and Republicans and then code Republican = 1
```

```r
# > and Democrat = 0.

dr_data <- subset(
thermometers,
party_id %in% c("Democrat", "Republican")
)

dr_data$party_bin <- ifelse(dr_data$party_id == "Republican", 1, 0)

table(dr_data$party_id, dr_data$party_bin)
```

```
##
##                 0    1
##   Democrat    1734    0
##   Independent    0    0
##   Not sure       0    0
##   Other          0    0
##   Republican     0 1412
```

```r
# This table lets me confirm that all Democrats are coded as 0 and all

# > Republicans are coded as 1 in the new party_bin variable.
```

```r
# Question 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.

#

# I include ft_immig (immigrants) and ft_police (police) plus their

# > interaction. Immigration and law-and-order attitudes are both

# > plausibly related to partisanship.

model2 <- lm(party_bin ~ ft_immig * ft_police, data = dr_data)

summary(model2)
```

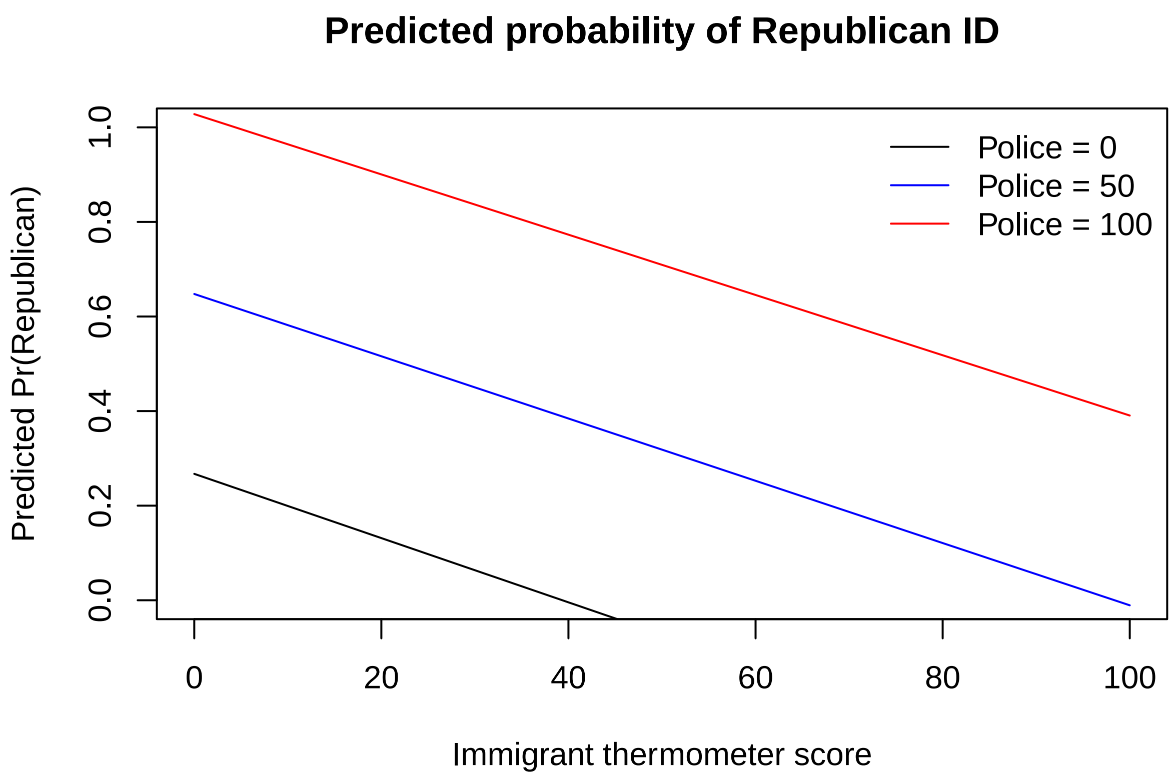
```
##
## Call:
## lm(formula = party_bin ~ ft_immig * ft_police, data = dr_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02797 -0.35270 -0.00885  0.36106  1.33806
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)         2.673e-01  6.226e-02   4.293 1.82e-05 ***
## ft_immig           -6.793e-03  8.966e-04  -7.577 4.69e-14 ***
## ft_police           7.607e-03  7.479e-04  10.172  < 2e-16 ***
## ft_immig:ft_police  4.198e-06  1.091e-05   0.385      0.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4184 on 2985 degrees of freedom
##   (157 observations deleted due to missingness)
## Multiple R-squared:  0.2937, Adjusted R-squared:  0.293
## F-statistic: 413.7 on 3 and 2985 DF,  p-value: < 2.2e-16
```

```r
# The coefficient on ft_immig is negative, so higher warmth toward

# > immigrants is associated with a lower probability of being

# > Republican.

# The coefficient on ft_police is positive, so higher warmth toward the

# > police is associated with a higher probability of being Republican.

# The interaction term is small and not statistically significant, so

# > the effect of immigrant attitudes does not seem to change much as

# > police attitudes change.

# This simple linear probability model gives a reasonable summary of how

# > these two issue attitudes are related to party identification in the

# > data.
```

```r
# Question 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?

#

# We vary ft_immig from 0 to 100 and look at predicted probabilities

# > when ft_police is 0, 50, or 100.

immig_seq <- seq(0, 100, by = 1)

new_low <- data.frame(
ft_immig = immig_seq,
ft_police = 0
```

```
)

new_mid <- data.frame(
ft_immig = immig_seq,
ft_police = 50
)

new_high <- data.frame(
ft_immig = immig_seq,
ft_police = 100
)

pred_low <- predict(model2, newdata = new_low)
pred_mid <- predict(model2, newdata = new_mid)
pred_high <- predict(model2, newdata = new_high)

plot(
immig_seq,
pred_low,
type = "l",
ylim = c(0, 1),
xlab = "Immigrant thermometer score",
ylab = "Predicted Pr(Republican)",
main = "Predicted probability of Republican ID"
)

lines(immig_seq, pred_mid, col = "blue")
lines(immig_seq, pred_high, col = "red")

legend(
"topright",
legend = c("Police = 0", "Police = 50", "Police = 100"),
col = c("black", "blue", "red"),
lty = 1,
bty = "n"
)
```




```
# The three lines all slope downward, meaning that as warmth toward
# > immigrants increases, the predicted probability of being Republican
# > falls, regardless of how someone feels about the police.
# The red line (police = 100) is highest at every point, which shows
# > that people who feel very warmly toward the police are more likely
# > to be Republican for any given level of immigrant warmth.
# These patterns are associations from survey data, not guaranteed
# > causal effects.
# Party identification and these attitudes likely influence each other,
# > and all of them are driven by deeper ideological and social factors,
# > so we should be careful not to claim that changing the thermometer
# > score would by itself cause someone to change parties.
```