R Problem set 3 at 2

Nathaniel Williams

2025-12-02

#Reading

1. Research Goals Goal of the study is casual inference, The causal effect of different variables and the occurrence of Civil War. The authors state their desire to parse through a few different common reasons in the literature, and propose a new framework of understanding A strength of how the set their hypothesis up is empirically testing the hypothesis of the previous standings in the literature (ethnicity and religious diversity) and then their own so their is a direct comparison built from the same observations. A weakness of this paper overall is that it is a prime example of "Kitchen Sink Regression," and it is hard to parse out and make sense of the sheer number of hypothesis and test.

2. Estimates

The theoretical estimands for the hypothesis are the causal effect of some ethnic/religious diversity variable or some structural factor favoring insurgency (terrain, state weakness, geography) on the outbreak of civil war.

The empirical estimands are the coefficient of the different independent variables in a logit regression model on the binary dv (No civil war or civil war)

Certain variables are more unclear than others.

For hypothesis 7, which is "Among countries with an ethnic minority comprising at least 5% of the population, greater ethnic diversity should be associated with a higher risk of ethnic civil war." The goal is to measure for ethnic civil wars, ones where fighters were majority mobilized on ethnic lines. For an ethnic civil war, you must have a ethnic minority present,with the capability of plausibly rising up and engaging in conflict. hence the floor condition of 5%. Presence of ethnic minorities -> higher likelihood of ethnic conflict The issue here is that the authors are claiming a rate increase in comparison to nations with no minorities above 5%, but if there are no ethnic minorities (or they are so small as to be insignificant) how could an ethnic conflict ever occur in the comparison group? You are virtually guaranteed to see an increase when moving from an essentially impossible environment for ethnic conflict to a possible one.

Another example of a possible problem is in hypothesis 8 which says "The presence of rough terrain, poorly served roads, at a distance from the centers of state power, should favor insurgency and civil war. So should the availability of foreign cross-border sanctuaries and a local population that can be induced not to denounce insurgents to government agents."

This is all focused on describing the observation that having an environment for insurgency leads to civil war. What that environment is is here classified through measures such as terrain or distance from state-power centers. The issue is that terrain and size of the country (distance from centers of power) are impacted from how borders are drawn. Many nations of interest for this study in Africa had borders imposed on them by European power following the world wars. This also means that borders impact ethnicity within the country as many were cut up by the partition.

Therefore, this matter of geography is not independent from issues such as whether the country was a former colony, whether it created or had its borders imposed, and the type of ethnic groups that end up being majority or minority.

3. Id The authors use several observations in a multivariate logistic regression over a cross-national time survey. Assumptions are: All confounders have been accounted and controlled for Explanatory variables are lagged by one year which prevents simultenaity bias The proxy variables used are accurate measures of the real world event they are interested in

4. It is unlikely that the goal of causal inference is met, and that the all else equal assumptions of regression has been met. The paper is too broad in its goals and it is unable to satisfy the idea that it has controlled for all confounders. The empirical strategy for some hypothesis that are more simple appear okay (for example ethnic diversity in a nation is generally easy to accurately measure), but many have holes that can be poked (such as H7 and H8). Other proxies of question are items such as terrain being categorized by mountainous regions, not accounting for more diverse biomes that may still create issues, such as jungles that are still prone to guerrilla warfare.

5. The paper can still be important, it shifts the debate away from base ethnic or religious grievances toward material and structural explanations such as state capacity and infrastructure. It also suggest a need for more critical thinking in policies aimed at stability beyond new borders or blind democracy building. Instead emphasizing the need for strong institutional and infrastructure building, in addition to policies of security aid.

#Sim

```
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
```

```
#Load in data set
Thermo<- read.csv('https://raw.githubusercontent.com/MLBurnham/pols_602/refs/heads/main/data/thermomete
head(Thermo)
```

```
##   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
```

```
#New age variable, year of study (2017) minus birth year
Thermo$Age<- 2017 - Thermo$birth_year
#Median of union thermoter,remove not observed
median(Thermo$ft_unions, na.rm = TRUE)
```

```
## [1] 51
```

```
#51

#standard deviation to access the spread of the data
sd(Thermo$ft_unions, na.rm = TRUE)
```

```
## [1] 31.78824
```

```
#31.7

#Examples for each category in the demographic
median(Thermo$ft_unions[Thermo$sex=='Male'], na.rm = TRUE)
```

```
## [1] 50
```

```
#50
sd(Thermo$ft_unions[Thermo$sex=='Male'],na.rm = TRUE)
```

```
## [1] 32.82687
```

```
#32.8
median(Thermo$ft_unions[Thermo$sex=='Female'], na.rm = TRUE)
```

```
## [1] 56
```

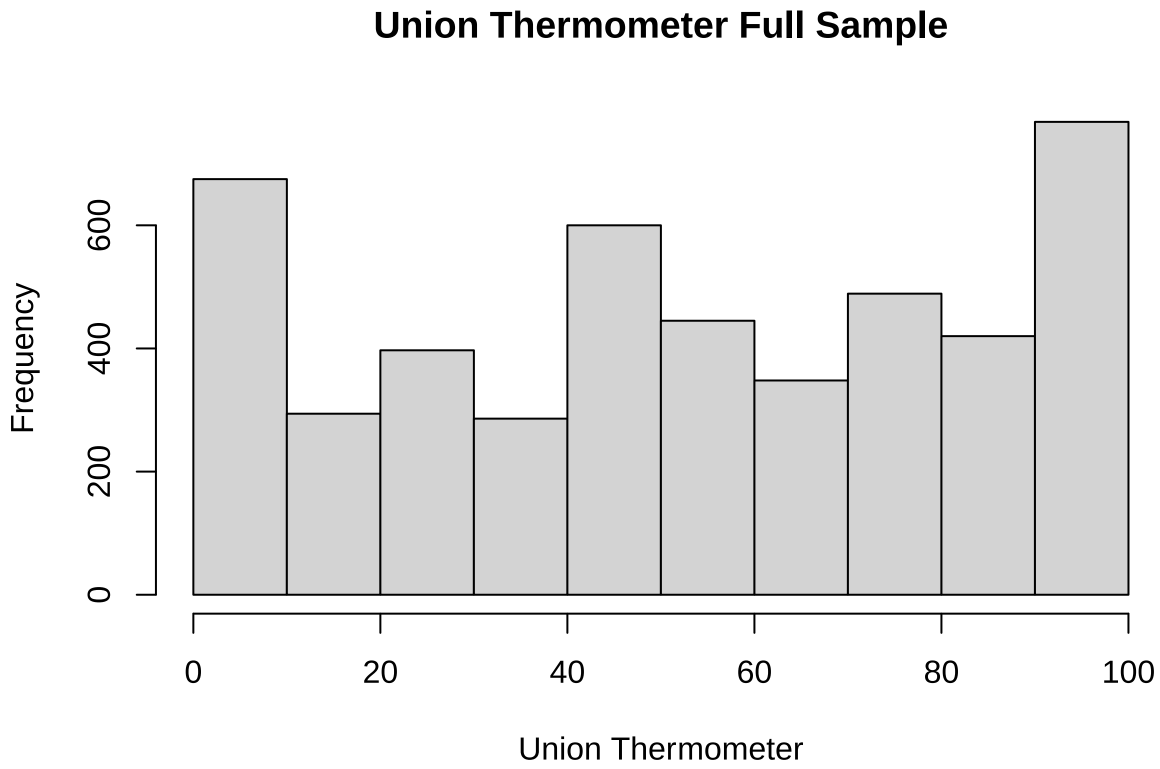
```
#56
sd(Thermo$ft_unions[Thermo$sex=='Female'], na.rm = TRUE)
```

```
## [1] 30.22849
```

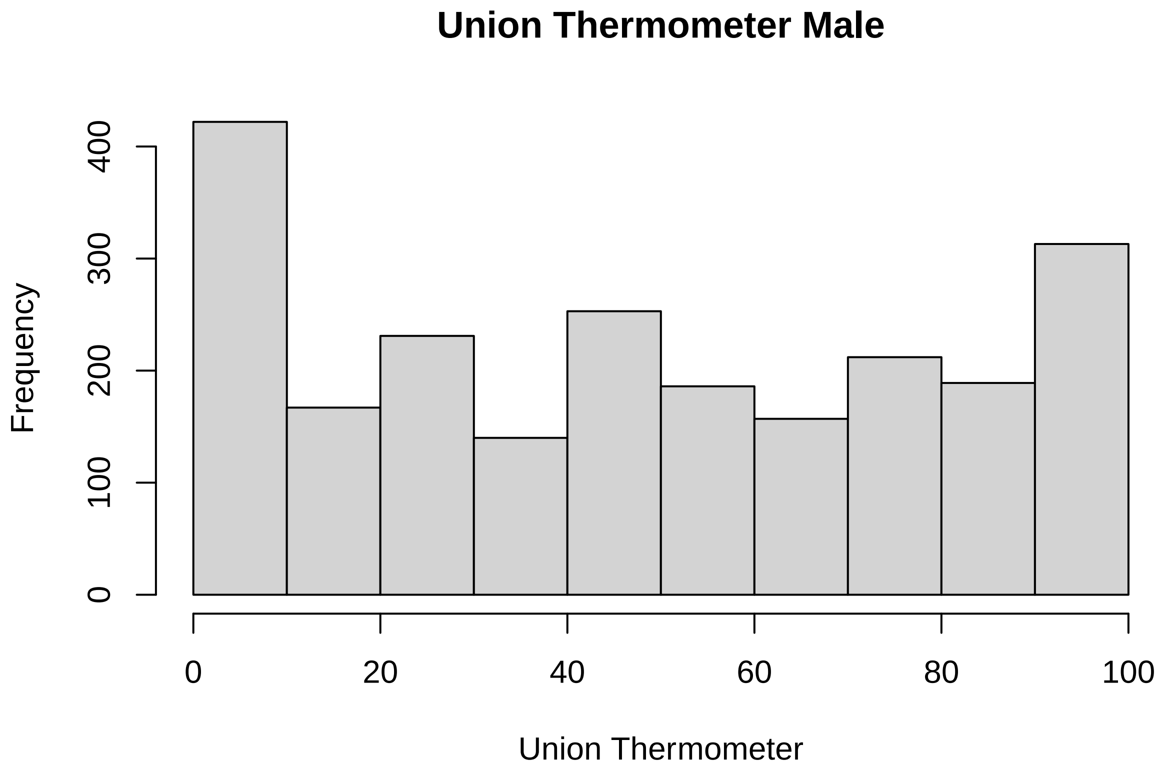
```
#30.2
#Checking that there are no other categories (other, prefer not to say)
unique(Thermo$sex)
```

```
## [1] "Female" "Male"
```

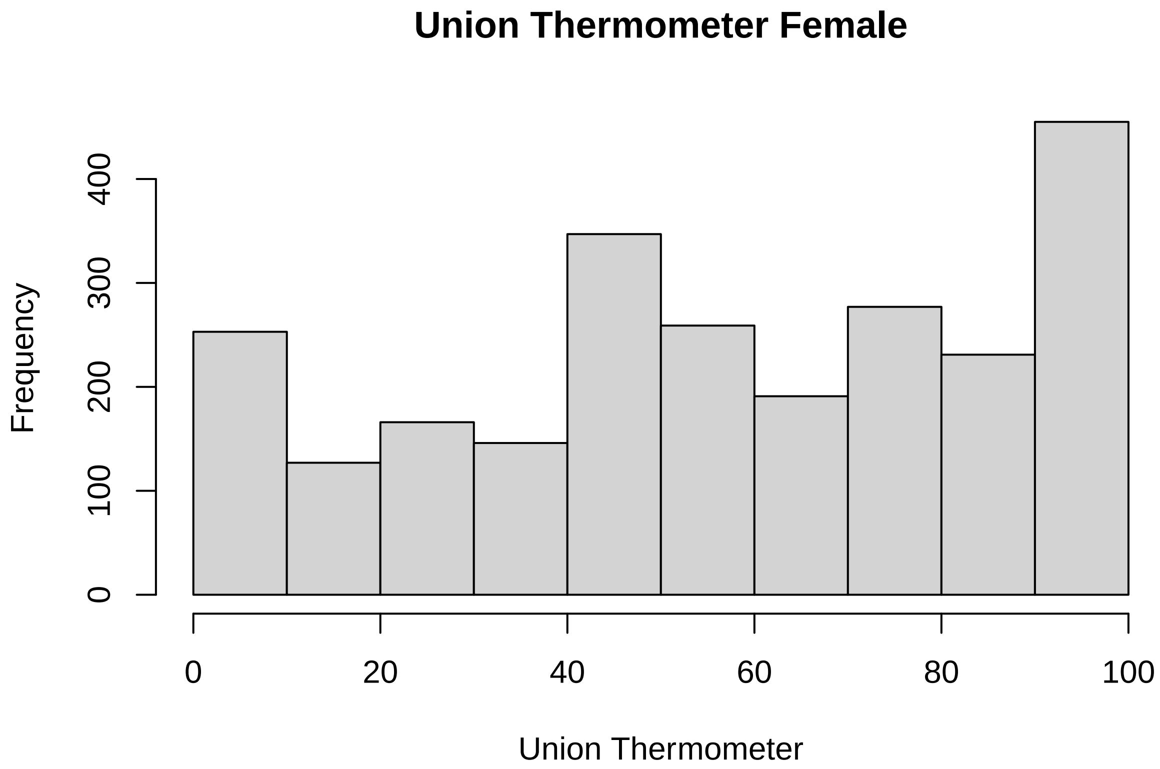
```
#Histograms
#Full Sample
hist((Thermo$ft_unions),
main = 'Union Thermometer Full Sample', xlab = 'Union Thermometer')
```



```
#Just male
hist(Thermo$ft_unions[Thermo$sex=='Male'], main = 'Union Thermometer Male', xlab = 'Union Thermometer')
```



```
#Just female
hist(Thermo$ft_unions[Thermo$sex=='Female'], main = 'Union Thermometer Female', xlab = 'Union Thermomet
```



The central tendency of the unions thermometer (the median) is 51 The union thermoter has a high spread, one change in standard deviation is plus or minus 31.7 points on the 100 scale. The central tendency for males is 50 and 56 for females The spread is higher than the sample mean at 32.8 for males, and slightly lower for females at 30.2

```
#Fit A regression model
Model1<-lm(ft_unions ~ sex, data=Thermo)

summary(Model1)
```

```
##
## Call:
## lm(formula = ft_unions ~ sex, data = Thermo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.454 -26.454   1.078  27.546  51.078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.4543     0.6362  90.305   <2e-16 ***
## sexMale      -8.5327     0.9176  -9.299   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.5 on 4720 degrees of freedom
##   (267 observations deleted due to missingness)
## Multiple R-squared:  0.01799,    Adjusted R-squared:  0.01778
## F-statistic: 86.47 on 1 and 4720 DF,  p-value: < 2.2e-16
```

$\hat{ftunions} = 57.45 - 8.53Male$

```r
Predictions <- predict(Model1)
summary(Predictions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.92   48.92   57.45   53.35   57.45   57.45
```

```r
Predmeans<-predict(Model1, newdata=data.frame(sex=unique(Thermo$sex)), interval = 'confidence')
Predmeans
```

```
##        fit      lwr      upr
## 1 57.45432 56.20703 58.70162
## 2 48.92159 47.62525 50.21792
```

```r
# The conditional means within 95% confidence
```

```r
#Democrat and Republican Data frame
#Make a new data frame that is just a subset only pulling democrat and republicans
ThermDR<- subset.data.frame(Thermo, party_id %in% c('Democrat','Republican'))
#Use if else to override the party_id variable with 1 for democrats and 0 for republicans
ThermDR$party_id <- ifelse(ThermDR$party_id == 'Democrat',1,0)

#Build a predictive model
Prmodel<-lm(party_id ~ ft_unions * ft_immig + sex, data=ThermDR)
# Better feelings toward unions and immigrants would suggest liberal ideology and more alignment with d
summary(Prmodel)
```

```
##
## Call:
## lm(formula = party_id ~ ft_unions * ft_immig + sex, data = ThermDR)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1058 -0.2601 -0.0240  0.2648  0.9973
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.071e-02  3.084e-02   1.644   0.1002
## ft_unions           5.987e-03  5.660e-04  10.578  < 2e-16 ***
## ft_immig            1.056e-03  5.042e-04   2.095   0.0363 *
## sexMale            -5.840e-02  1.458e-02  -4.005 6.36e-05 ***
## ft_unions:ft_immig  3.508e-05  8.060e-06   4.352 1.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3865 on 2905 degrees of freedom
##   (236 observations deleted due to missingness)
## Multiple R-squared:  0.3949, Adjusted R-squared:  0.3941
## F-statistic:   474 on 4 and 2905 DF,  p-value: < 2.2e-16
```

$\hat{partyid} = .05 + .006ftunions + .001ftimmig - .05Male + 0.000035ftunion * ftimmig$

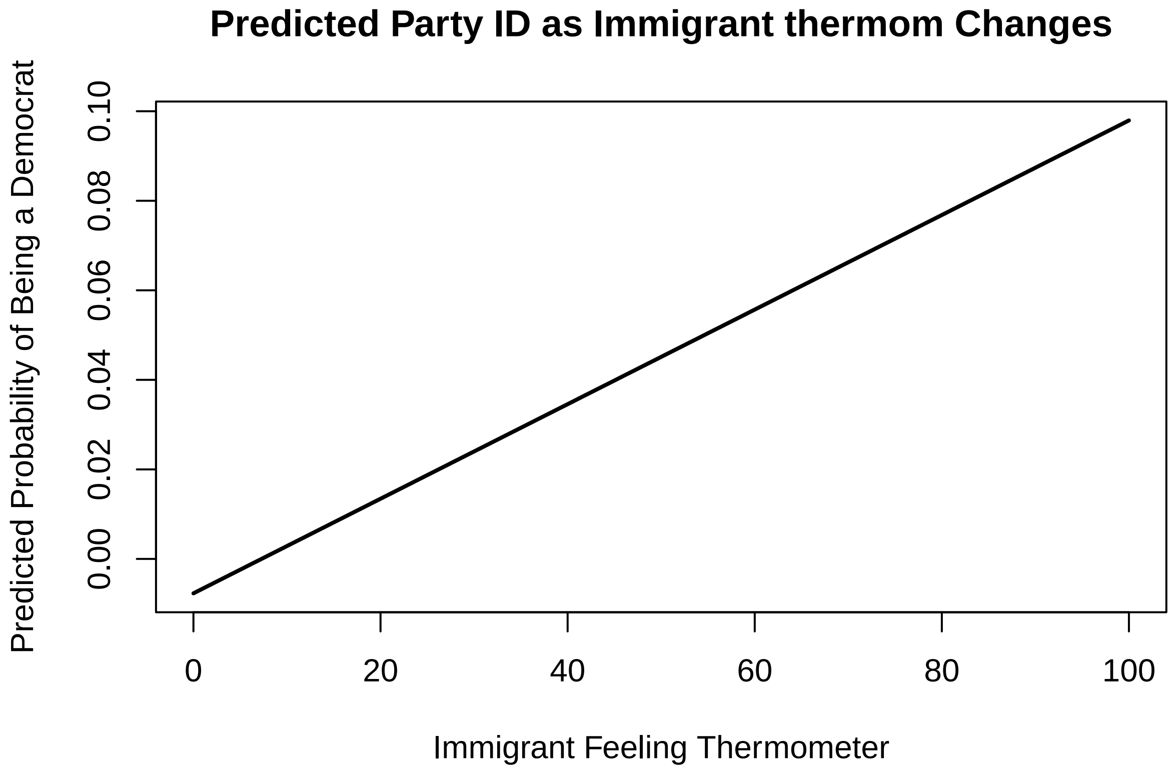
Party id is binary so the coefficients represent percentage point increases in the odds of being a democrat (1).

```r
# Plot as feeling thermometer changes
# Generate sequence of thermometer values to substitute into model
ft_seq <- seq(0, 100,length.out = 200)

# Create a simulated data set that holds other variables constant to isolate effect of just increasing
pred_data <- data.frame(
  sex = "Male",
  ft_unions = 0 ,
  ft_immig = ft_seq
)

# Making a variable with the predictions from the model and simmed data
pred_vals <- predict(Prmodel, newdata = pred_data)

# Plot
plot(ft_seq, pred_vals,
     type = "l",
     lwd = 2,
     xlab = "Immigrant Feeling Thermometer",
     ylab = "Predicted Probability of Being a Democrat",
     main = "Predicted Party ID as Immigrant thermom Changes")
```



Having a positive opinion of immigrants leads to up to a 10 percentage point increase in the odds of being a democrat, interestingly only turns negative when extremely close to zero. I would not see this as causal as I do not think the assumption of all confounders being controlled for has been met.