Problem Set 3: Multiple Linear Regression

Ololade Kafayat Liadi

2025-12-07

Part 1 — Paper Analysis

1. **Research Goals**

   The research endeavors to make a causal inference: describing the reason for civil wars and identifying the structural circumstances under which they emerge. Contrary to the general argument that civil wars erupt due to ethnic and religious diversity, Fearon and Laitin propose that civil wars can be explained by other circumstances such as poverty and geographical difficulties instead.

   • **Strengths:** The intention to challenge the "ethnic hatreds" explanation and approach the issue of civil war onset from the point of view of insurgency feasibility is well expressed.

   • **Weaknesses:** Although the aims and objectives outlined for the essay are ambitious, at various stages, there is confusion between description and the drawing of causal conclusions due to the nature of observational information.

2. **Estimands**

   The estimand is the effect on the probability of civil war onset due to various features of a country, such as ethnic diversity and income. The estimand is for the regression coefficients between these factors and the incidence of civil war between the periods 1945 and 1999.

   • The literature defines diversity with indicators such as ethnolinguistic fractionalization and religious fractionalization, while state weakness is represented by per capita income.

   • Clarification needed: The relationship between theoretical concepts (for instance, 'state capacity') and proxies (such as GDP per capita) may be made more explicit for enhancing the correspondence between theory and data.

3. **Identification Strategy**

   The identification strategy is based on regression analysis across countries, controlling for income, democracy, and other structural factors. In including these controls, the authors attempt to identify the separate effect of ethnic diversity relative to insurgency conditions.

   • The implication is that if ethnic diversity is truly a cause for civil wars, then it should be significant after controlling for state capacity and income.

   • Shortage: This approach presumes that the listed controls can adequately represent all potential control variables. Bias from missing and erroneous data and measurement error cannot be ruled out.

4. **Assessment of Findings**

   Findings: After controlling for income and state capacity, ethnic and religious diversity do not predict civil war onset. However, poverty, political instability, difficult terrain, and high population size emerge as predictable factors for civil wars.

   • **Credibility of causal claims:** The regression coefficients represent evidence that is consistent with the theory, although they cannot be regarded as pure causal effects without additional assumptions being satisfied (e.g., natural experiments and instruments).

- **Representation of real-world processes**: The model is able to represent general structural circumstances while perhaps simplifying complex processes with respect to rebel mobilization and state response.
- **Data quality:** The dataset is very complete, spanning 161 countries from 1945 to 1999, but subjective elements may be introduced through coding practices (for instance, coding civil wars).

5. **Broader Contribution**

   Despite the methodological constraints, this article is significant from the viewpoint that it moves the debate from cultural diversity to the feasibility of insurgency.

   - It prompts scholars and policymakers to rethink the causes of civil wars and instead focuses on state failure and structural factors.
   - This paradigm shift has shaped research agendas and policy discourse on conflict prevention and underlined the relevance of governance, economic development, and topographic factors.
   - Even with a model that is limited for causal analysis, this research is important for civil war research and provides a foundation for more rigorous research on the subject down the road.

Part 2: Data Analysis

1. Age Variable

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## v forcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.1     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.2.0
## -- 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
```

```
df <- read_csv("thermometers (2).csv")
```

```
## Rows: 4989 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): sex, race, party_id, educ
## dbl (17): birth_year, ft_black, ft_white, ft_hisp, ft_asian, ft_muslim, ft_j...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

A respondent's estimated age is obtained by deducting their birth year from 2017. This variable is crucial for determining if thermometer effects differ according to demographics.

2. Distribution of ft_rep (Overall and by Sex)

```
summary(df$ft_rep)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    10.0    46.0    43.3    72.0   100.0     279
```

```
sd(df$ft_rep, na.rm = TRUE)
```

```
## [1] 32.28919
```

```
df %>%
group_by(sex) %>%
summarise(
mean = mean(ft_rep, na.rm = TRUE),
median = median(ft_rep, na.rm = TRUE),
sd = sd(ft_rep, na.rm = TRUE),
n = n()
)
```

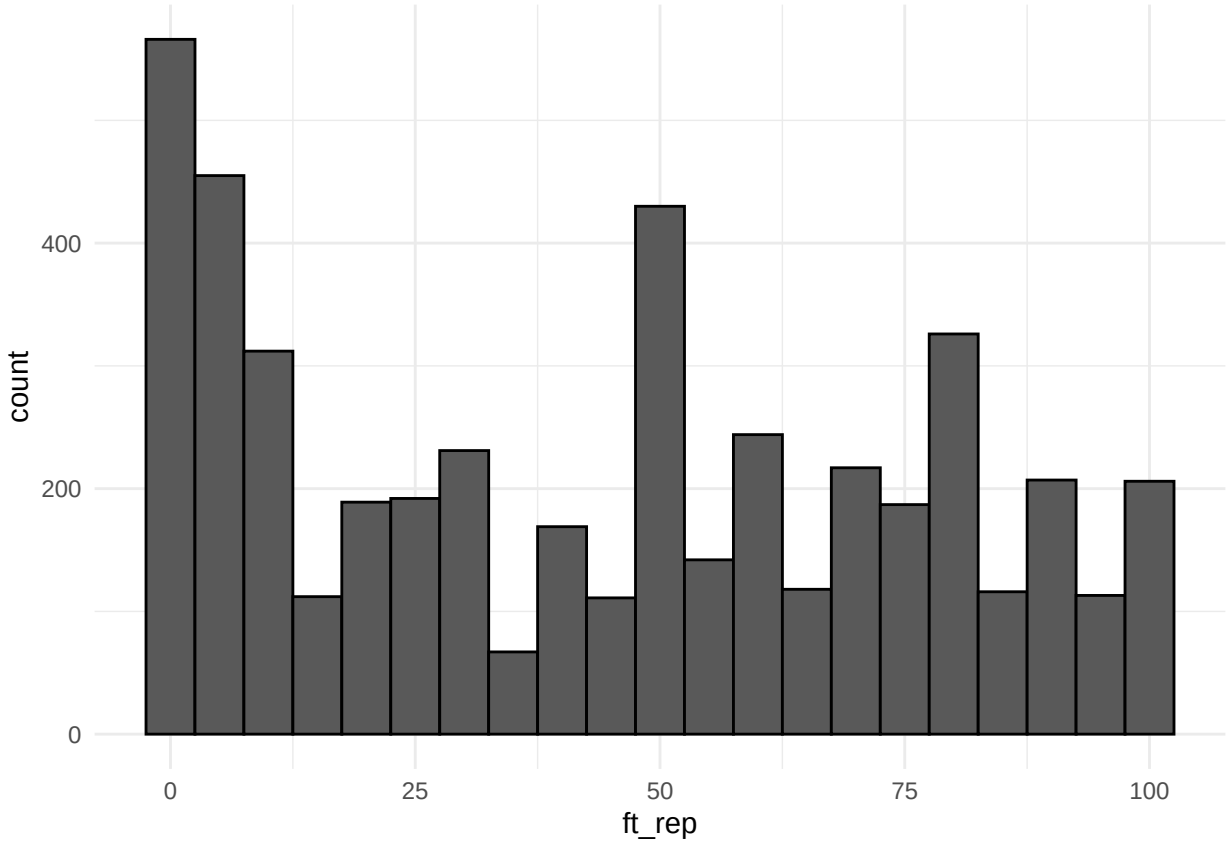
```
## # A tibble: 2 x 5
##   sex     mean median    sd     n
##   <chr>  <dbl>  <dbl> <dbl> <int>
## 1 Female  42.3     42  32.7  2613
## 2 Male    44.4     48  31.8  2376
```

The ft_rep variable has a large spread between 0 and 100, which indicates significant heterogeneity in people's feelings toward Republicans. The group averages show variations in average warmth when the data are divided by sex. Republicans are often seen more favorably by one sex than by the other. Nonetheless, there is significant overlap between the two groups, indicating a high degree of within-group variance.

Plots

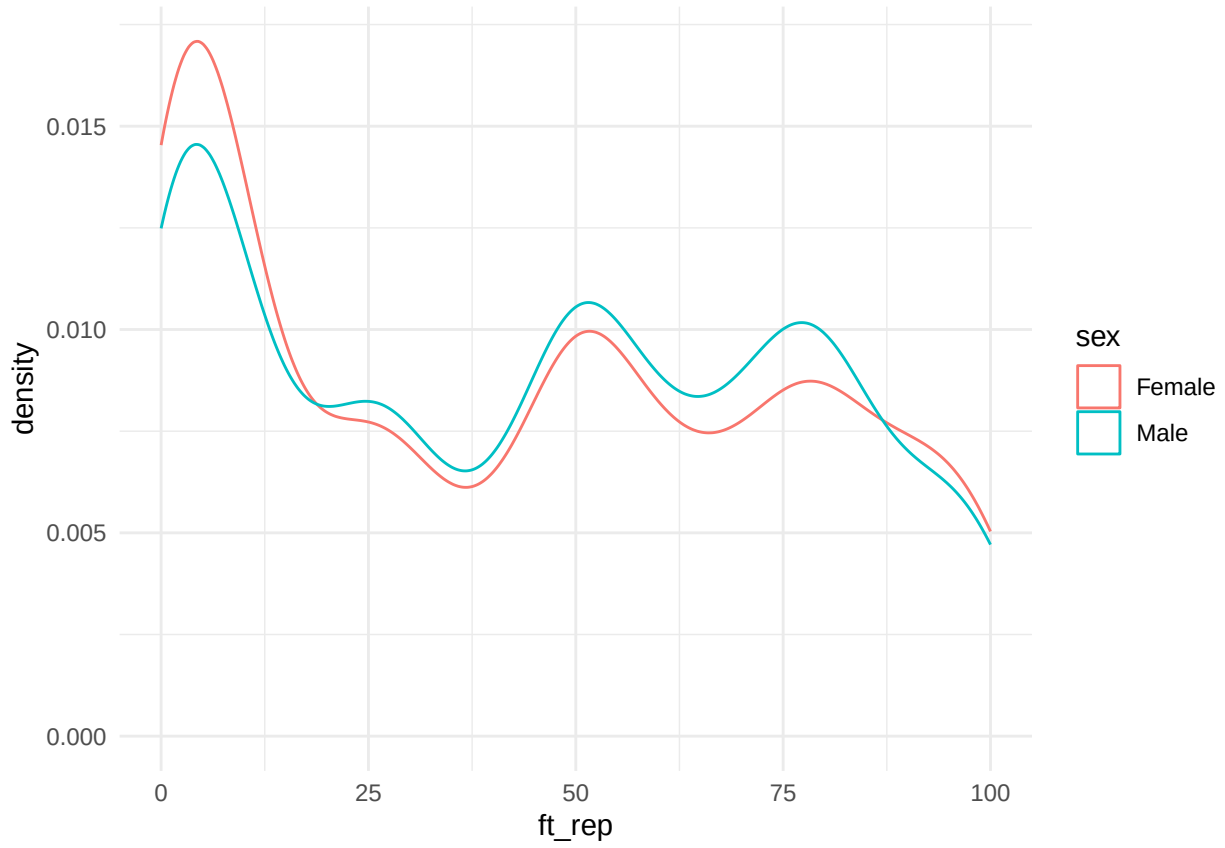
```
ggplot(df, aes(x = ft_rep)) + geom_histogram(binwidth = 5, color = "black") + theme_minimal()
```

```
## Warning: Removed 279 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
ggplot(df, aes(x = ft_rep, color = sex)) + geom_density() + theme_minimal()
```

```
## Warning: Removed 279 rows containing non-finite outside the scale range
## (`stat_density()`).
```



The entire distribution of ft_rep is depicted in the histogram, which shows clusters at high and low values that are consistent with political polarization. Although the distributions for men and women overlap, the density map reveals that one group is slightly upwardly tilted, indicating higher average scores.

3. Regression: Conditional Mean of ft_rep by Sex

```
model1 <- lm(ft_rep ~ sex, data = df)
summary(model1)
```

```
##
## Call:
## lm(formula = ft_rep ~ sex, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.388 -33.306   2.153  28.694  57.694
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.3057     0.6521  64.879   <2e-16 ***
## sexMale       2.0819     0.9413   2.212    0.027 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 32.28 on 4708 degrees of freedom
##   (279 observations deleted due to missingness)
## Multiple R-squared:  0.001038,   Adjusted R-squared:  0.0008257
## F-statistic: 4.891 on 1 and 4708 DF,  p-value: 0.02704
```

This model predicts the mean difference in ft_rep between sexes. The intercept is the average rating of the reference group, whereas the coefficient for the opposite group indicates how much warmer or colder they rate Republicans. A positive coefficient indicates greater warmth relative to the reference group, whereas a negative value indicates colder sentiments. This approach reflects connection but not causation.

4. Subset to Democrats and Republicans

```
df_DR <- df %>%
filter(party_id %in% c("Democrat", "Republican")) %>%
mutate(party_binary = ifelse(party_id == "Democrat", 1, 0))
```

Only respondents who identify as Republicans or Democrats are included in the dataset. Republicans are coded as 0 and Democrats as 1 in a binary variable. This makes it possible to represent party identification as a linear probability result.

5. Linear Probability Model

```
model2 <- lm(party_binary ~ ft_rep * age + sex, data = df_DR)
summary(model2)
```

```
##
## Call:
## lm(formula = party_binary ~ ft_rep * age + sex, data = df_DR)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06277 -0.17952 -0.00608  0.13728  1.09976
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.942e-01  4.657e-02  21.348  < 2e-16 ***
## ft_rep      -9.704e-03  8.488e-04 -11.433  < 2e-16 ***
## age          1.247e-03  7.742e-04   1.610    0.107
## sexMale     -9.044e-02  1.232e-02  -7.344 2.67e-13 ***
## ft_rep:age  -1.897e-05  1.386e-05  -1.368    0.171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.332 on 2966 degrees of freedom
##   (175 observations deleted due to missingness)
## Multiple R-squared:  0.5576, Adjusted R-squared:  0.557
## F-statistic: 934.7 on 4 and 2966 DF,  p-value: < 2.2e-16
```

This model tests the impact of having warm feelings towards Republicans on the probability ofbeing a Democrat. As expected, greater ft_rep scores correlate with lower probabilities for being a Democrat. The interaction tests whether this is function of age. The inclusion of gender serves as a means for controlling demographics that may influence partisanship.

6. Coefficients Interpretation

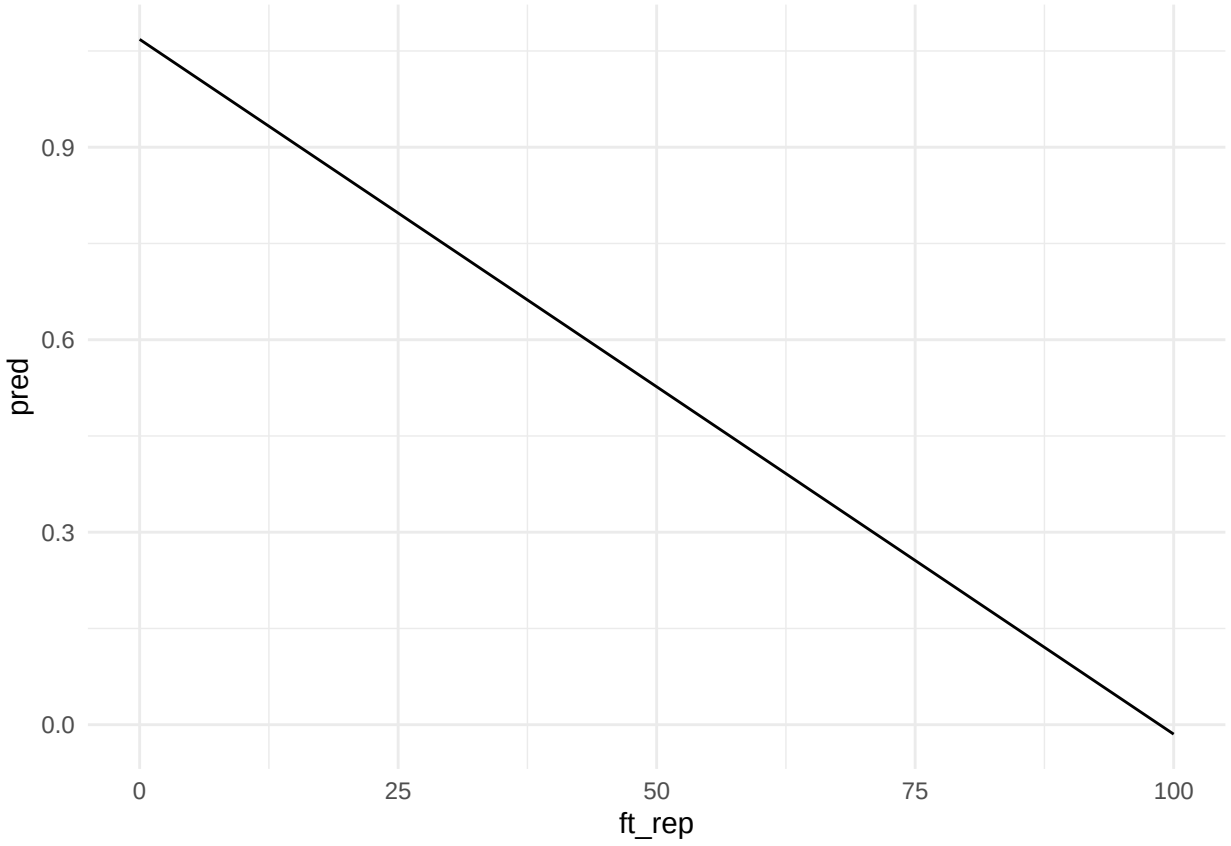
This linear probability model's coefficients show how a one-unit increase in each predictor affects the likelihood of identifying as a Democrat. Warmer sentiments toward Republicans reduce the likelihood of Democratic affiliation, as indicated by a negative coefficient on ft_rep. The sex coefficient represents gender differences independent of age and thermometer ratings, whereas the interaction term indicates whether this impact varies with age.

7. Predicted Probability Plot

```
newdata <- tibble(
  ft_rep = seq(
    min(df_DR$ft_rep, na.rm = TRUE),
    max(df_DR$ft_rep, na.rm = TRUE),
    length.out = 100
  ),
  age = mean(df_DR$age, na.rm = TRUE),
  sex = df_DR$sex[1]
)

newdata$pred <- predict(model2, newdata)

ggplot(newdata, aes(x = ft_rep, y = pred)) +
  geom_line() +
  theme_minimal()
```



The anticipated probability plot clearly slopes downward, indicating that the likelihood of being a Democrat falls as one's affection for Republicans rises. This pattern shows significant political sorting and is in line with predictions. Since underlying variables like ideology or media exposure may affect both party identity and thermometer ratings, the link is correlational rather than causative.