Data Exploration: Gender and World View

Owen Bernstein

October 14, 2021

In this Data Exploration assignment, you will work with data that has been modified from the Barnhart et al. (2020) article. You will investigate whether certain types of countries are more likely to initiate conflicts with other countries. Note that you are only working with the data used to generate the *Monadic Findings* of the paper - that is, you will examine whether democracies initiate fewer conflicts than autocracies.

If you have a question about any part of this assignment, please ask! Note that the actionable part of each question is **bolded**.

The Suffragist Peace

Data Details:

- File Name: `suffrage_data.csv`
- Source: These data are from Barnhart et al. (2020).

Variable Name	Variable Description
`ccode1`	Unique country code
`country_name`	Country name
`year`	Year
`init`	The number of overall conflicts initiated by the country specified by `ccode1` during the year specified by `year`
`init_autoc`	The number of overall conflicts initiated by the country specified by `ccode1` **with autocracies** during the year specified by `year`
`init_democ`	The number of overall conflicts initiated by the country specified by `ccode1` **with democracies** during the year specified by `year`
`democracynosuff`	Indicator variable for a democracy without women's suffrage. 1 if the country is a democracy without women's suffrage, 0 otherwise.
`suffrage`	Indicator variable for a country with women's suffrage
`autocracy`	Indicator variable for a country with an autocratic government
`nuclear`	Indicator variable for whether the country is a nuclear power
`wcivillibs`	Measure of the degree of civil liberty women enjoy, ranging from 0-1, where higher values mean women have more civil liberties
`polity`	Polity score for the country specified by `ccode1` during the year specified by `year`

Question 1

Part a

Before getting started, it is a good idea to take a look at the structure of the data. This data set is different from what we've seen so far. Until now, all the data we've looked at has had the individual as the unit of observation. This means that each row of the data corresponds to a single individual, and the columns correspond to some characteristics of that individual, like their responses to a survey. When working with data, it is important to understand the unit of observation, along with other characteristics of the data. The unit of observation is the object about which data is collected. That could be, say, an individual, a country, a football game, or an episode of TV. **Take a look at the data to determine the unit of observation. Note that the structure isn't exactly the same as the data used in Barnhart et al. (2020).**

Part b

Is war rare or common? **Make a histogram of the main dependent variable, `init`. Comment on what you see, being sure to keep the unit of observation and the definition of the `init` variable in mind. Is what you see surprising? What does it say about the frequency of initiating conflict?**

Question 2

How were the `autocracy` and `suffrage` variables defined? Can autocracies also have women's suffrage (at least in this coding scheme)? What reasons do the authors of the paper give for these coding decisions and how do you think it might affect their findings? (Hint: take a close look at pages 651 and 652 of the original article.)

Question 3

The democratic peace - i.e. the propensity for democracies to avoid conflict with each other, and to avoid conflict more generally - is an empirical regularity. The theory, as originally posed, is not gendered. Do the data support the democratic peace theory? **Ignoring suffrage status for now, do the data suggest that modern democracies initiate fewer conflicts than autocracies? Do democracies tend to initiate conflict more with autocracies or other democracies?**

Question 4

Now that we've taken a look at the classic democratic peace theory, let's take an initial look at how women's suffrage is related to initiating conflict. **Conduct a bivariate regression, modeling the number of conflicts initiated with women's suffrage (i.e. `init ~ suffrage`). This will help inform you about how the number of conflict initiated in a year depends on women's suffrage. Report the coefficient on `suffrage`. Interpret your results. If you like, extend the problem by reporting the 95% confidence interval for the `suffrage` coefficient. Is the relationship statistically significant?**

The `lm()` function is used to calculate regressions in R. Here is a guide to linear regression in R that may be helpful.

Question 5

The model in the previous question was very simple; we modeled initiation only as a function of suffrage. In reality, the relationship is probably more complicated - conflict initiation probably depends on more than just women's suffrage. **Look at the other variables available in the data and find one or more that you think may also be related to conflict initiation. Explain why you think so, then add the variable(s) to the right side of the regression (as explanatory variables) in question 4. Interpret what you find.**

Question 6: Data Science Question

Estimate a regression of the following form: `init ~ suffrage + polity + polity*suffrage`, where `polity*suffrage` is the interaction between polity score and women's suffrage. Compare this to the same model but without the interaction term. Interpret your results.

In the social sciences, we use interaction terms in regressions to capture heterogeneous effects. As an example of how to implement and interpret this type of model, suppose we wanted to understand the relationship between education on the one hand (as the outcome variable), and age and gender on the other hand (as explanatory variables). We might think that the effect of age on education depends on whether you're talking about men or women. Maybe for men, age has no effect on education, but for women, there is a negative effect, as older women were discouraged or barred from seeking higher education. To assess whether this is true, we can use an interaction between gender and age. You can model this in R using this formula in the `lm()` function: `education ~ age + female + age*female` (supposing gender is coded into a binary variable `female`). Here, `age*female` is what creates the interaction.

Lets say that we ran this regression in R and found that the model looks like this: $education = 1.5 + .005 * age + .01 * female + -.4 * age * female$. Here, the coefficient on `age` is .005, .01 on `female`, and -.4 on the interaction between the two. Without an interaction, to interpret the coefficient on `age`, we would say the effect of `age` on `education` is .005. However, the interaction term modifies that relationship - the effect of `age` on `education` now depends on gender.

To see this, we must plug in values for `female` and `age`. When `female` = 0, then the interaction term vanishes, and then the effect of `age` on `education` is .005. In other words, for non-women, there is a very small relationship between age and education. Now plugging in `female` = 1, the effect of `age` on `education` becomes .005 (the coefficient on age) + -.4 (the coefficient on the interaction) = -.395. In other words, the effect of age on education among women is negative.

Interpreting the effect of gender is a bit more complicated, and in this case is nonsensical. To do so, we set `age` = 0 (which doesn't make a ton of sense) to find that the effect of `female` on `education` is .01 when `age` = 0. Always pay attention to whether the coefficients you're focusing on are even substantively meaningful.

```r
mod_1 <- lm(init ~ suffrage + polity*suffrage, data = s_data)
mod_2 <- lm(init ~ suffrage + polity, data = s_data)

stargazer(mod_1, mod_2, type = "text")
```

```
##
## ===================================================================
##                                    Dependent variable:
##                      -----------------------------------------------
##                                           init
##                               (1)                     (2)
## -------------------------------------------------------------------
## suffrage                    0.064*                  0.0001
##                            (0.038)                 (0.026)
##
## polity                      -0.002                 -0.004**
##                            (0.002)                 (0.002)
##
```

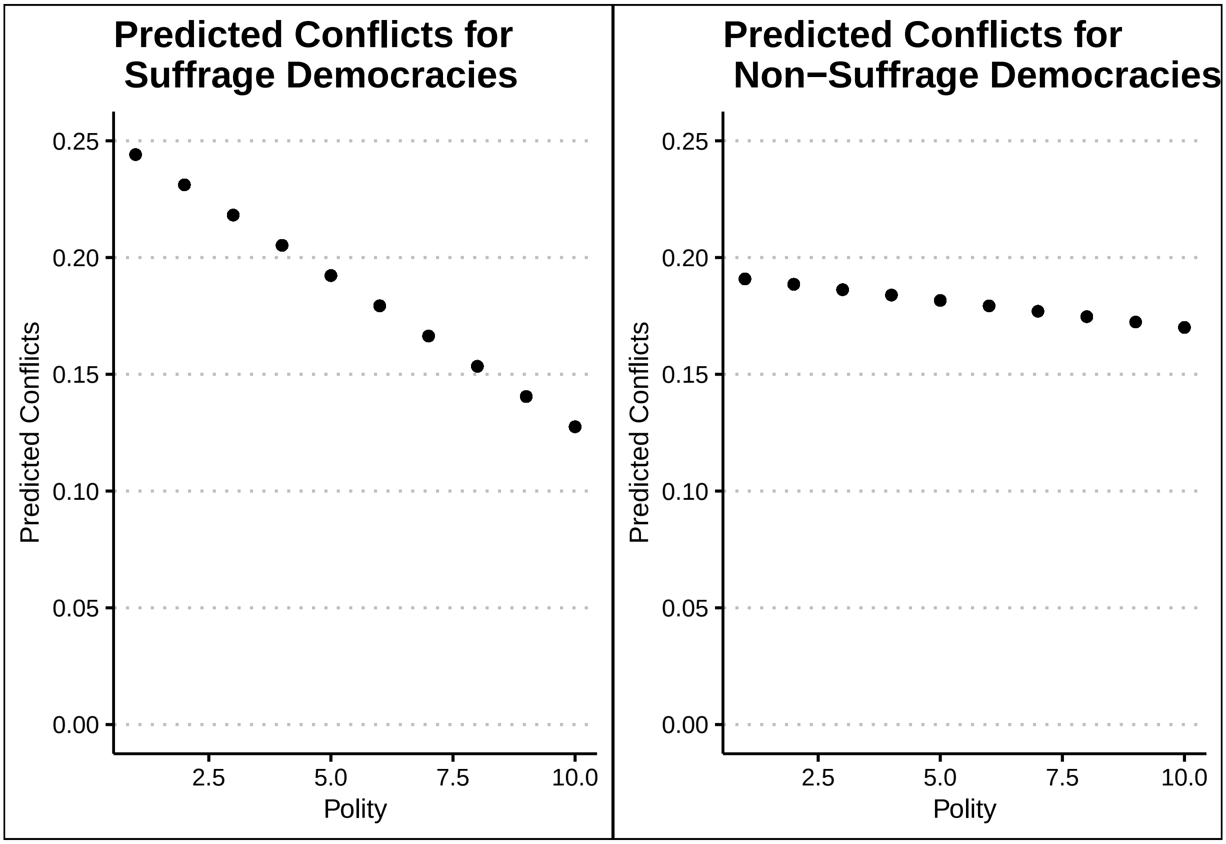
```
## suffrage:polity            -0.011**
##                             (0.005)
##
## Constant                   0.193***                0.185***
##                             (0.013)                 (0.012)
##
## -------------------------------------------------------------------
## Observations                 9,865                   9,865
## R2                           0.002                   0.002
## Adjusted R2                  0.002                   0.002
## Residual Std. Error    0.659 (df = 9861)       0.659 (df = 9862)
## F Statistic         8.132*** (df = 3; 9861) 9.682*** (df = 2; 9862)
## ===================================================================
## Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

Question 7: Data Science Question

When using regression, especially with interactions, sometimes it is useful to visualize the results. **Create two plots of the predicted number of conflicts per year on the y-axis and Polity score on the x-axis (among countries with a Polity score greater than or equal to one only), split by suffrage. That is, one plot should plot the predicted number of conflict per year among suffrage democracies, and the other among non-suffrage democracies. This way you will be able to visualize the interaction between suffrage and Polity score that we saw in the previous question. This guide may be helpful in doing so - it uses a different type of regression model (binary logit), but the principle of prediction is the same. Make sure to hold the `suffrage` variable at 0 or 1. Comment on what you find.**

```
suffrage_data <- s_data %>%
  filter(polity > 0) %>%
  filter(suffrage == 1) %>%
  mutate(predicted_val = predict(mod_1, newdata = .))


no_suffrage_data <- s_data %>%
  filter(polity > 0) %>%
  filter(suffrage == 0) %>%
  mutate(predicted_val = predict(mod_1, newdata = .))

plot_1 <- suffrage_data %>%
  ggplot(aes(x = polity, y = predicted_val)) +
  geom_point() +
  theme_clean() +
  labs(x = "Polity", y = "Predicted Conflicts", title = "Predicted Conflicts for \n Suffrage Democracie
  ylim(0, 0.25)

plot_2 <- no_suffrage_data %>%
  ggplot(aes(x = polity, y = predicted_val)) +
  geom_point() +
  theme_clean() +
  labs(x = "Polity", y = "Predicted Conflicts", title = "Predicted Conflicts for \n Non-Suffrage Democra
  ylim(0, 0.25)

end_plot <- plot_1 + plot_2
```

```
end_plot
```



Question 8

One of the advantages of the data we have is that we can plot trends over time. **Group countries into those with and without suffrage and plot the average number of disputes initiated by those countries in each year covered by the data. Comment on what you find.**