Data Exploration: Symbolic Politics

Yanxi Fang

October 21, 2021

```
# load the data containing the tibble protest_df
load('RN_2001_data.RData')
```

Question 1

The two outcome variables of interest are listed in Figure 2, and are the respondent's favorability toward the police, and the level of discrimination that the recipient perceives the Black population in the US as facing. The former is on a 1-to-4 scale, with lower values being more favorable and higher values being less favorable. The latter is on a 1-to-5 scale, with lower values representing less discrimination and higher values representing more discrimination. (Note: in Figure 2, the authors use a measure of "unfavorability", which is presumably on the same 4-point scale, but in reverse.)

Question 8: Data Science Question

Part a

Below, I ran a regression with favorability toward the police as the outcome variable, and two predictor variables: party and the onset of the protests. The results are printed below.

Both predictor variables have statistically significant coefficients. The -0.1274 coefficient on the party variable means that an increase of 1 in the `pid7` variable (i.e. one step to the right on the left-right spectrum, with "Strong Democrat" at the furthest left and "Strong Republican" at the furthest right) is predicted to decrease a person's unfavorability of the police (i.e. increase a person's favorability of the police) by 0.1274 on average, when holding the protest-onset binary variable constant. In contrast, the 0.1671 coefficient on the protest-onset binary variable means that when holding party affiliation constant, the post-protest unfavorability of the police is predicted to increase by 0.1671 (i.e. a favorability decrease of 0.1671) on average. The coefficient, 2.4786, is the starting value of the police unfavorability that serves as the foundation for changes in the predicted unfavorability based on the values of the predictor variables and the estimated coefficients.

```r
# create binary variable for whether protests have started
protest_df <- protest_df %>%
  mutate(protest_binary = ifelse(day_running >0, 1, 0))

# run linear model with two explanatory variables
model8a <- lm(group_favorability_the_police ~ pid7 + protest_binary, data = protest_df)

# print results
summary(model8a)
```

```
##
## Call:
## lm(formula = group_favorability_the_police ~ pid7 + protest_binary,
##     data = protest_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5182 -0.7139 -0.2237  0.6489  2.4135
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.4785683  0.0034454  719.38   <2e-16 ***
## pid7           -0.1274360  0.0007361 -173.13   <2e-16 ***
## protest_binary  0.1670863  0.0039022   42.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9659 on 326794 degrees of freedom
##   (51710 observations deleted due to missingness)
## Multiple R-squared:  0.08807,    Adjusted R-squared:  0.08807
## F-statistic: 1.578e+04 on 2 and 326794 DF,  p-value: < 2.2e-16
```

Part b

The linear functional form does not accurately model the relationship because linear regressions assume a continuous, unbounded outcome variable; in this case, the outcome variable (police favorability/unfavorability) is neither unbounded nor continuous, since it is only able to take on four integer values (1 through 4). In this sense, a linear regression will likely produce predicted outcomes that do not fall within the domain of the outcome variable, requiring additional assumptions (e.g. rounding up/down) to actually get predicted values that are realistic.

In addition, the linear functional form assumes that the relationship between the predictor and outcome variables will always be constant regardless of the starting/initial values of the predictor variables. This is, again, not necessarily the most accurate model. For instance, someone who is a "Strong Democrat" (labeled as `1` for `pid7`) and moves to a `2` on the `pid7` scale will probably experience only a small change in their view of the police, while someone who is an "Independent" (labeled as `4` for `pid7`) and moves to either a `3` or a `4` will probably experience a relatively larger change in their view of the police.

Thus, for this question, I chose to use an entirely different functional form: a logistic (logit) model. The logit model requires a binary (0 or 1) outcome, which is something that is actually done in the paper (the right-hand side of Figure 3). Specifically, rather than focusing on respondents' 1-to-4 evaluation fo the police, we could choose to focus specifically on whether respondents viewed police "very unfavorably" (i.e. setting responses of `4` as `1`s, with all other responses as `0`s). In addition, the logit model addresses the "constant-effect" assumption of the linear model, since predicted changes to the outcome variable are larger at middle values of the predictor variables than at the extreme values.

Below, I run a logit model with a large number of variables, such as race and ethnicity (including the binary variable for `hispanic`), age, whether someone went to college, and the respondent's household income – all of which are factors that influence political outlooks and thus attitudes toward the police. The results are shown below.

While the numerical results are harder to interpret for the logit model, their magnitude and sign (positive/negative) are helpful. Below, it is clear that all estimated coefficients are statistically significant, so everything can be interpreted. The party ID, Hispanic, age, college, and household income variables all had negative coefficients, meaning that a respondent with higher values of any one of those variables (being more "right" on the left-right spectrum, being Hispanic, being older, being college educated, having higher household income) is predicted to be less likely to see the police "very unfavorably" on average, and when holding other predictor variables constant. In contrast, the positive coefficients on `race_ethnicity` and `protest_binary` suggest that non-white respondents, as well as respondents being asked after the protests had occurred, are predicted to be more likely to see the police "very unfavorably" on average, when holding other predictor variables constant.

```r
# run a logit model

# create outcome variable: similar to actual paper
protest_df <- protest_df %>%
  mutate(police_unfavorable_binary = ifelse(group_favorability_the_police == 4, 1, 0))

# add race, education, etc.
model8b <- glm(police_unfavorable_binary ~ pid7 + race_ethnicity + hispanic +
               age + college + household_income + protest_binary,
               data = protest_df, family = binomial(link = "logit"))

summary(model8b)
```

```
##
## Call:
## glm(formula = police_unfavorable_binary ~ pid7 + race_ethnicity +
##     hispanic + age + college + household_income + protest_binary,
##     family = binomial(link = "logit"), data = protest_df)
##
## Deviance Residuals:
```

```
##      Min       1Q   Median       3Q      Max  
## -1.2075  -0.5446  -0.4000  -0.2844   3.0838  
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)       0.1629428  0.0206130   7.905 2.68e-15 ***
## pid7             -0.1589875  0.0027180 -58.494  < 2e-16 ***
## race_ethnicity    0.0110989  0.0014472   7.669 1.73e-14 ***
## hispanic         -0.0091088  0.0022149  -4.113 3.91e-05 ***
## age              -0.0335244  0.0003920 -85.524  < 2e-16 ***
## college          -0.2659996  0.0136379 -19.504  < 2e-16 ***
## household_income -0.0326325  0.0009147 -35.675  < 2e-16 ***
## protest_binary    0.5453103  0.0125157  43.570  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 225733  on 311659  degrees of freedom
## Residual deviance: 206914  on 311652  degrees of freedom
##   (66847 observations deleted due to missingness)
## AIC: 206930
## 
## Number of Fisher Scoring iterations: 5
```

Part c

Below, I used the MASS package to run an ordinal probit model using the same model as part b.

It is difficult to compare coefficients because the models and even the outcome variables aren't the same (this is an ordered probit model, whereas I had fitted a binary logit model in part b). Although the coefficients have the same signs (positive/negative) as before, their magnitudes/absolute values are substantially different for some variables, while other variables have relatively similar coefficients. For instance, the race/ethnicity variable has similar coefficients of 0.011 and 0.012, while the protest variable changed from `0.545` to `0.210`.

```
library(MASS)
select <- dplyr::select

# modify outcome variable to become a factor
protest_df$police_favor_factor <- as.factor(protest_df$group_favorability_the_police)

# run ordinal probit model
model8c1 <- polr(data = protest_df,
                 formula = police_favor_factor ~ pid7 + race_ethnicity + hispanic +
                  age + college + household_income + protest_binary, method = "probit")

# print results
summary(model8c1)
```

```
## Call:
## polr(formula = police_favor_factor ~ pid7 + race_ethnicity + 
##     hispanic + age + college + household_income + protest_binary, 
##     data = protest_df, method = "probit")
```

```
##
## Coefficients:
##                     Value Std. Error  t value
## pid7             -0.13826  0.0008943 -154.592
## race_ethnicity    0.01204  0.0005753   20.921
## hispanic         -0.00410  0.0008099   -5.062
## age              -0.01438  0.0001263 -113.828
## college          -0.01444  0.0045003   -3.209
## household_income -0.01522  0.0003068  -49.599
## protest_binary    0.20986  0.0045922   45.700
##
## Intercepts:
##     Value     Std. Error t value
## 1|2   -1.6007    0.0080  -199.6846
## 2|3   -0.6560    0.0077   -84.8192
## 3|4    0.0360    0.0077     4.6619
##
## Residual Deviance: 753773.30
## AIC: 753793.30
## (66847 observations deleted due to missingness)
```

Section

```r
# group by day
protest_byday_df <- protest_df %>%
  group_by(day_running) %>%
  summarise(mean_police = mean(group_favorability_the_police, na.rm = T))

protest_df <- protest_df %>%
  left_join(protest_byday_df, by = "day_running")

# pre and post treatment
protest_pre <- protest_byday_df %>%
  filter(day_running < 0)
protest_post <- protest_byday_df %>%
  filter(day_running >= 0)

# linear regression for pre-treatment
pre_lm <- lm(mean_police ~ day_running, data = protest_pre)
summary(pre_lm)
```

```
##
## Call:
## lm(formula = mean_police ~ day_running, data = protest_pre)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18314 -0.04492 -0.00198  0.04391  0.36927
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.954e+00  8.373e-03 233.440  < 2e-16 ***
## day_running -2.953e-04  4.605e-05  -6.413 5.28e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07405 on 312 degrees of freedom
## Multiple R-squared:  0.1165, Adjusted R-squared:  0.1136
## F-statistic: 41.13 on 1 and 312 DF,  p-value: 5.281e-10
```

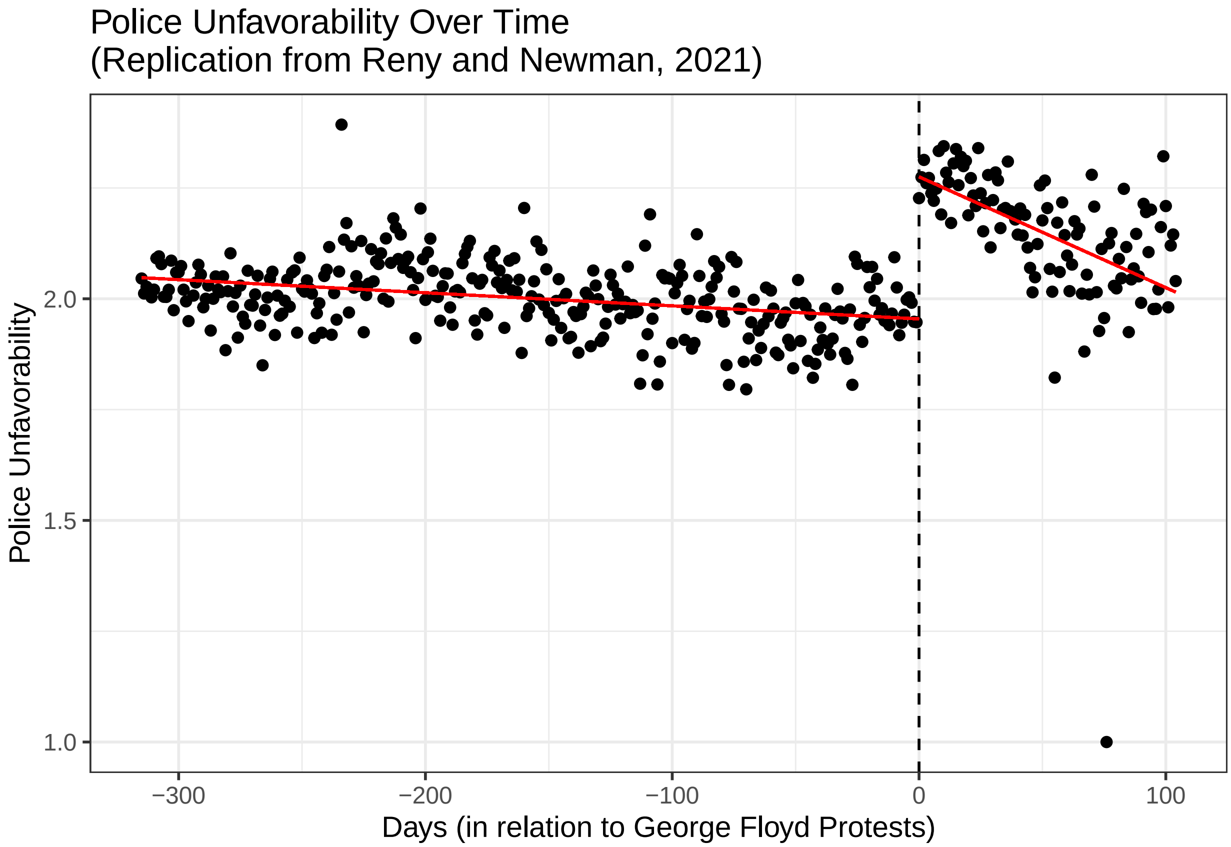
```r
section_intercept_p <- summary(pre_lm)$coefficients[1,1]
section_coefficient_p <- summary(pre_lm)$coefficients[2,1]

# linear regression for post-treatment
post_lm <- lm(mean_police ~ day_running, data = protest_post)
summary(post_lm)
```

```
##
## Call:
## lm(formula = mean_police ~ day_running, data = protest_post)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08506 -0.04843  0.01289  0.06820  0.29389
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.2751164  0.0273013  83.334  < 2e-16 ***
## day_running -0.0025008  0.0004536  -5.513 2.62e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1409 on 103 degrees of freedom
## Multiple R-squared:  0.2279, Adjusted R-squared:  0.2204
## F-statistic:  30.4 on 1 and 103 DF,  p-value: 2.624e-07
```

```
section_intercept <- summary(post_lm)$coefficients[1,1]
section_coefficient <- summary(post_lm)$coefficients[2,1]

# plot
protest_byday_df %>%
  ggplot(aes(x = day_running, y = mean_police)) +
  geom_point() +
  geom_vline(xintercept = 0, linetype = "dashed") +
  geom_segment(x = 0, xend = 104, y = section_intercept + 0*section_coefficient,
             yend = section_intercept + 104*section_coefficient, color = "red") +
  geom_segment(x = -315, xend = 0, y = section_intercept_p + (-315)*section_coefficient_p,
             yend = section_intercept_p + 0*section_coefficient_p, color = "red") +
  ggtitle("Police Unfavorability Over Time \n(Replication from Reny and Newman, 2021)") +
  xlab("Days (in relation to George Floyd Protests)") + ylab("Police Unfavorability") +
  theme_bw()
```



```
ggsave("police_unfavorability_time.png", height = 4, width = 8)
```