Data Exploration: Symbolic Politics

Your name here

October 21, 2021

In this Data Exploration assignment we will explore Reny and Newman's (2021) finding that opinions towards the police and about the level of discrimination faced by Black Americans were impacted by the spread of protests in the wake of the killing of George Floyd. You will recreate, present, and assess those claims as well as creating your own regression models to test which attitudes change and when.

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

Opinion Mobilization: The George Floyd Protests

Data Details:

- File Name: `RN_2001_data.RData`
- Source: These data are from Reny and Newman (2021).

Variable Name	Variable Description
`race_ethnicity`	Race or ethnicity. Levels labelled in data: 1-White, 2-Black or AfAm, 3-American Indian or Alaskan Native, 4 through 14-Asian or Pacific Islander (details in labels), and 15-Some other race
`hispanic`	Of Hispanic, Latino, or Spanish origin. Levels labelled in data: 1-Not Hispanic, 2-15 Hispanic of various origins
`day_running`	Day relative to onset of George Floyd protests (day 0)
`age`	Respondent's age
`female`	Binary indicator variable: 1 if respondent female, 0 otherwise
`college`	Binary indicator variable: 1 if respondent attended college, 0 otherwise
`household_income`	Household pre-tax income ranging from 1 (less than $15,000) to 24 (more than $250,000). Details for other levels in labels.
`pid7`	Party identification on a seven point scale with strong, weak, lean: 1-Strong Democrat to 7-Strong Republican with 4-Independent.
`ideo5`	Ideological self placement: 1-Very liberal, 2-Liberal, 3-Moderate, 4-Conservative, 5-Very Conservative
`vote_clinton`	Indicator variable for whether the respondent said they voted for Clinton in the 2016 presidential election
`group_favorability_the_police`	Favorability towards the police: 1-Very favorable, 2-Somewhat favorable, 3-Somewhat unfavorable, 4-Very unfavorable

Variable Name	Variable Description
discrimination_blacks	Perceptions of the level of discrimination in US faced by Blacks: 1-None at all, 2-A little, 3-A moderate amount, 4-A lot, 5-A great deal
day	The date the respondent took the survey
group_fav_white_black	The difference in respondents favorability towards Blacks subtracted from their favorability towards whites (each on four point scale). Ranges from -3 to 3.
racial_attitudes_generations	Agreement with the statement that generations of slavery and discrimination have made it difficult for Blacks to work their way out of the lower class: 1-Strongly Agree to 5-Strongly Disagree
interest	Degree to which respondent claims to follow politics: 1-Most of the time, 2-Some of the time, 3-Only now and then, 4-Hardly at all
group_favorability_jews	Favorability towards Jews: 1-Very favorable, 2-Somewhat favorable, 3-Somewhat unfavorable, 4-Very unfavorable
group_favorability_whites	Favorability towards whites: 1-Very favorable, 2-Somewhat favorable, 3-Somewhat unfavorable, 4-Very unfavorable
group_favorability_evangelicals	Favorability towards evangelicals: 1-Very favorable, 2-Somewhat favorable, 3-Somewhat unfavorable, 4-Very unfavorable
group_favorability_socialists	Favorability towards socialists: 1-Very favorable, 2-Somewhat favorable, 3-Somewhat unfavorable, 4-Very unfavorable
protest	Indicator variable if survey respondent lived in area that would at any point have a BLM protest in the wake of the killing of George Floyd
n_protests	Number of eventual BLM protests in area where resident lived

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

#Note that the data is saved in the form of a tibble, a special table using the dplyr package that has

head(protest_df$race_ethnicity)
```

```
## <labelled<double>[6]>: What is your race? Provided by LUCID.
## [1] 6 1 1 2 1 1
##
## Labels:
##  value                              label
##      1                              White
##      2         Black, or African American
##      3   American Indian or Alaska Native
##      4               Asian (Asian Indian)
##      5                    Asian (Chinese)
##      6                   Asian (Filipino)
##      7                   Asian (Japanese)
##      8                     Asian (Korean)
##      9                  Asian (Vietnamese)
##     10                       Asian (Other)
##     11 Pacific Islander (Native Hawaiian)
##     12       Pacific Islander (Guamanian)
##     13          Pacific Islander (Samoan)
```

```
##     14           Pacific Islander (Other)
##     15                    Some other race
##    777             Not asked in this wave
```

```
head(protest_df$household_income)
```

```
## <labelled<double>[6]>: What is your current annual household income before taxes? Provided by L...
## [1] 21  8  7  1 NA  1
##
## Labels:
##  value                   label
##      1      Less than $14,999
##      2     $15,000 to $19,999
##      3     $20,000 to $24,999
##      4     $25,000 to $29,999
##      5     $30,000 to $34,999
##      6     $35,000 to $39,999
##      7     $40,000 to $44,999
##      8     $45,000 to $49,999
##      9     $50,000 to $54,999
##     10     $55,000 to $59,999
##     11     $60,000 to $64,999
##     12     $65,000 to $69,999
##     13     $70,000 to $74,999
##     14     $75,000 to $79,999
##     15     $80,000 to $84,999
##     16     $85,000 to $89,999
##     17     $90,000 to $94,999
##     18     $95,000 to $99,999
##     19   $100,000 to $124,999
##     20   $125,000 to $149,999
##     21   $150,000 to $174,999
##     22   $175,000 to $199,999
##     23   $200,000 to $249,999
##     24     $250,000 and above
##    777 Not asked in this wave
```

Question 1

As usual it is important to first examine the structure of the data. What are the two main outcome variables of interest to Reny and Newman? How were they measured and how are they coded in the data? What was the treatment? **Take a look at the data and determine which are the two outcome variables of interest. Observe the scale of each.**

##Question 2

###Part a R has a special 'date' class for storing and manipulating dates as seen below. Date variables can conveniently be logically compared and arithmetically manipulated. Using the day variable find out how many days the dataset spans. **First check using the code below that the day variable is of the class 'date'. Next subtract the latest day in the sample from the first day to calculate the timespan covered by the dataset. Hint: functions like max() and min() work for date variables too!**

```r
class(protest_df$day)
```

```
## [1] "Date"
```

###Part b On what date is the treatment said to have occurred? **Find the date for which the day_running variable is 0. Then modify the code below to add a variable to each row for whether or not the observation was before or after treatment.**

```r
#Change the object to be the date of the protest spread, remember to put it in quotes if you copy/paste

protest_df_bydate <- protest_df %>%  mutate(before = ifelse(day<as.Date("INSERT_TREATMENT_DAY_HERE"), 1
```

```
## Error: Problem with `mutate()` column `before`.
## i `before = ifelse(day < as.Date("INSERT_TREATMENT_DAY_HERE"), 1, 0)`.
## x character string is not in a standard unambiguous format
```

Question 3

###Part a Compare the average for each outcome variable before and after the onset of the protests. Are the differences statistically significant? **Calculate the outcome variable means for before and after treatment. Conduct a test as to whether the differences in means are statistically significant. Hint: you can use either the t.test() function or difference_in_means() from the estimatr package**

###Part b It might be that the period before and after the treatment was different in ways in addition to the onset of the protests. Use the same procedure as above to check for differences between two means of a survey response measuring favorability towards a group besides the police. **Calculate the means from before and after the treatment and conduct a test of statistical significance of the difference for another measure of group favorability that was recorded in the survey (e.g. evangelicals, Jews, socialists, or whites). Is there also a substantive or statistically significant difference on that variable? Should that change our confidence in attributing the opinion changes found in part a to the George Floyd protests?**

Question 4

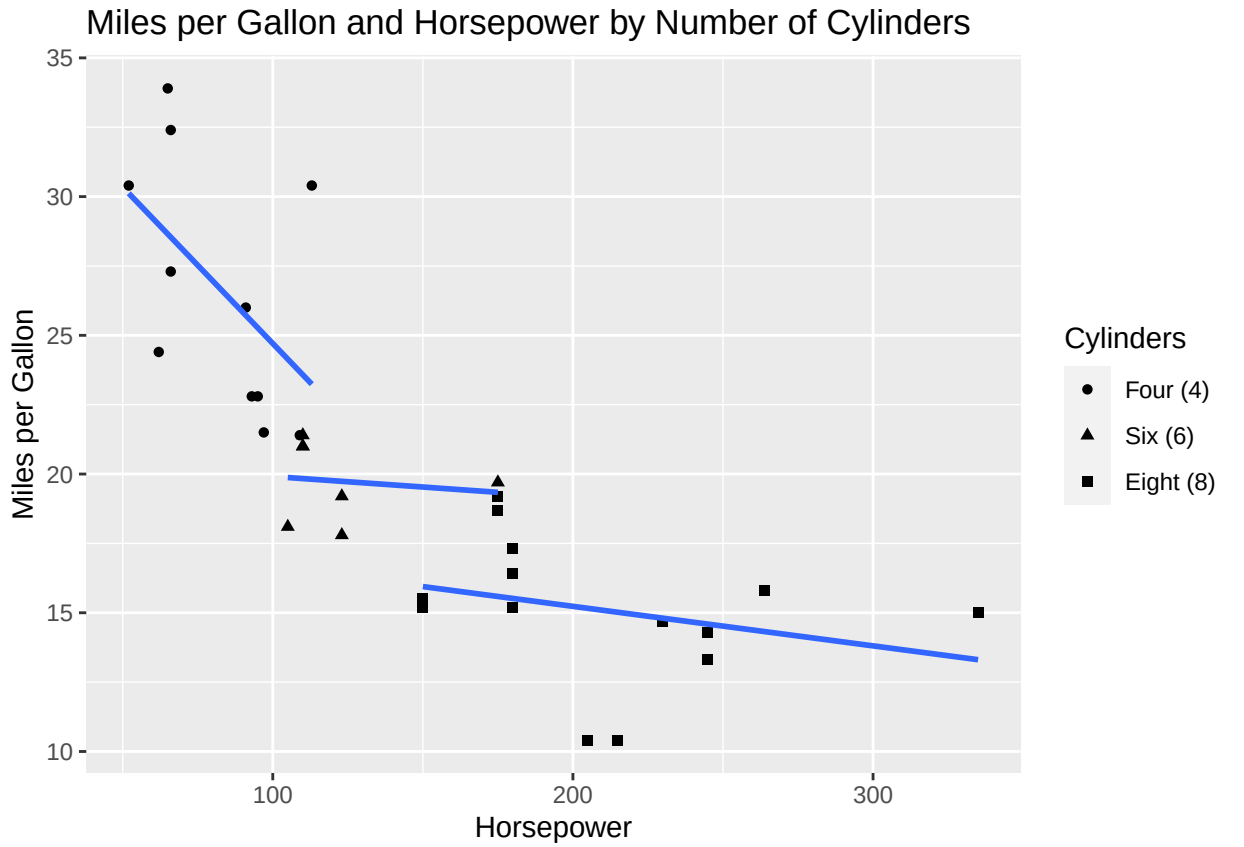
###Part a In order to create figures similar to the panels in Figure 2 in Reny and Newman (2021) we must first manipulate the data to be more usable. If we intend to graph the average of each outcome variable for each day, on what variable should we group the data using group_by? **Create a new object that is the data split out by the appropriate group and producing the average for each of the two outcome variables for each day. Also be sure to preserve an indicator for whether the observations are from before or after the spread of the protests.**

###Part b Graph the results for the entire sample. **Graph the results for the entire sample for both outcome variables by day. Include a vertical line demarcating when the protests started to spread. Does there appear to be a shift in the outcome variables from before to after the protests began to spread?**

###Part c It might be useful to more clearly illustrate the differences in the trend lines before and after the protests began. **Modify the code below to include a separate line of best fit for before and after the protests began. Does the trend line align with your previous reading of the graph? Remember to add a vertical line demarcating for the onset of treatment.**

```r
#An example of how to do multiple lies of best fit using example data from mtcars (mtcars is a dataset

ggplot(data=mtcars, aes(x=hp, y = mpg, shape=as.factor(cyl))) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  scale_shape_discrete("Cylinders", labels=c("Four (4)", "Six (6)", "Eight (8)")) +
  ggtitle("Miles per Gallon and Horsepower by Number of Cylinders") +
  xlab("Horsepower") +
  ylab("Miles per Gallon")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Question 5

###Part a The attitudes in question are no doubt highly influenced by the respondent's race and ethnicity. How do the graphs from question 4 differ for white and Black respondents. **Subset the data to include only white respondents and recreate the graphs from part c of question 4. Do the same with the data from only Black respondents. How do these differ from each other? Hint: Be careful when subsetting white responses to not also include Hispanic responses.**

###Part b As we have learned partisanship heavily influences how people take in and process new information. **Split the sample into Democrats, Republicans and independents and use them to produce the same graphs as part a (either all in the same figure or separate). Compare both the level and the trends for each party affiliation. What could this imply about how partisanship affects processing?**

##Question 6:

###Part a The graphs in questions 4 and 5 indicate that the effects dissipate as time progresses past the onset of the protests. **Explain why that might be the case? What does this indicate about whether or not attitudes towards the police are symbolic or not?**

###Part b One way to look at the effect decay is to bin the post-protest data and compare averages. **Split the post-protest data into however many groups you choose and compare the period directly after the protest with the latest period in the data. What are the differences in means for the outcomes?**

Question 7

###Part a What are some reasons we might be unconvinced by the comparison of aggregate survey results from a time before and after an event? Do you think they apply here?

###Part b There is often a problem in conducting surveys of non-response bias. That is, the people who answer surveys may differ from the people who do not answer surveys and the differences may vary over time. This is especially damaging to inference when non-response is correlated with the outcomes being measured. For example after a series of damaging headlines supporters of a politician may be less willing to answer phone surveys about that politician. As a result we would potentially observe an exaggeration of the negative effects of the scandal on a politician's polled approval rating. Test whether this is the case in the Reny and Newman data. **Test whether there is balance between the respondents before and after the onset of the protests along two demographic traits that you would expect to correlate with the measured responses to the outcome variables.**

###Part c Racial resentment is often considered a symbolic attitude in strength and consistency. Examine the before and after levels of racial resentment as measured by the question from the racial resentment scale about the impact of generations of slavery and discrimination (racial_attitudes_generations). **Graph the average racial_attitudes_generations (remember the direction of how it is coded!) by day like other outcome variables. Does it behave like the other outcome variables? Does the data support that racial attitudes are symbolic attitudes?**

Question 8: Data Science Question

###Part a Run an initial regression examining the relationship between favorability towards the police, party, and treatment. **Run a regression examining party and the onset of the protests' effect on favorability towards the police. Interpret the results**

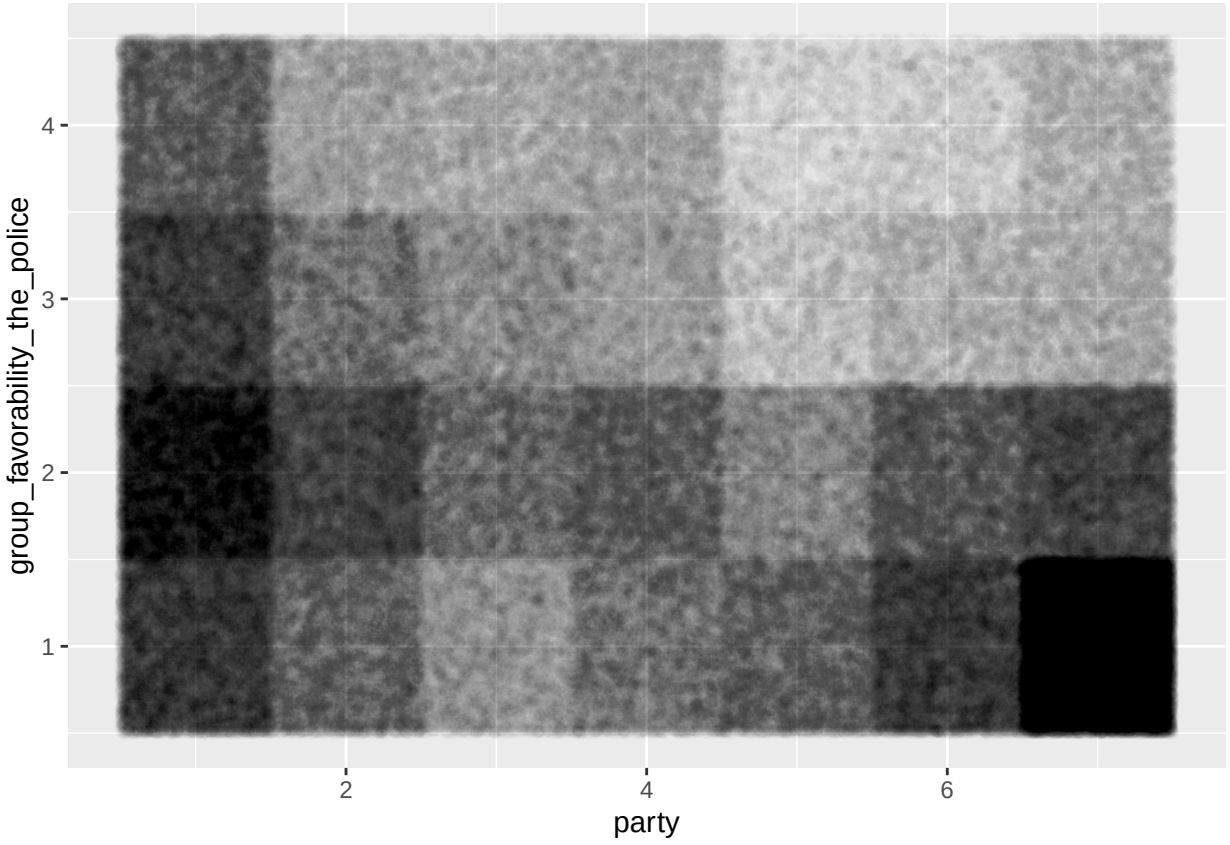
```
protest_df <- protest_df %>%
  mutate(party = as.double(pid7))

protest_df_stand <- protest_df %>%
  mutate(party.standardised = (party-mean(party, na.rm = TRUE))/sd(party, na.rm = TRUE)) %>%
  mutate(day_running.standardised = (day_running-mean(day_running))/sd(day_running)) %>%
  mutate(n_protests.standardised = (n_protests-mean(n_protests))/sd(n_protests)) %>%
  select(group_favorability_the_police, party.standardised, day_running.standardised, n_protests.standa

model <- lm(group_favorability_the_police ~ party + day_running, protest_df)
model_stand <- lm(group_favorability_the_police ~ party.standardised + day_running.standardised, protes
```

###Part b The above functional form probably does not accurately model the relationship of all the relevant covariates in the dataset. What functional form would you recommend using and why? What covariates would you add? Is there need for an interaction term? **Run a regression of your specification and interpret the results. Justify your choices in modeling.**

```r
# exploring what each variable looks like plotted
ggplot(protest_df, aes(x = party, y = group_favorability_the_police)) +
  geom_point(alpha = 0.01, position=position_jitter(height=.5, width=.5))
```

```
## Don't know how to automatically pick scale for object of type haven_labelled/vctrs_vctr/double. Defau
```

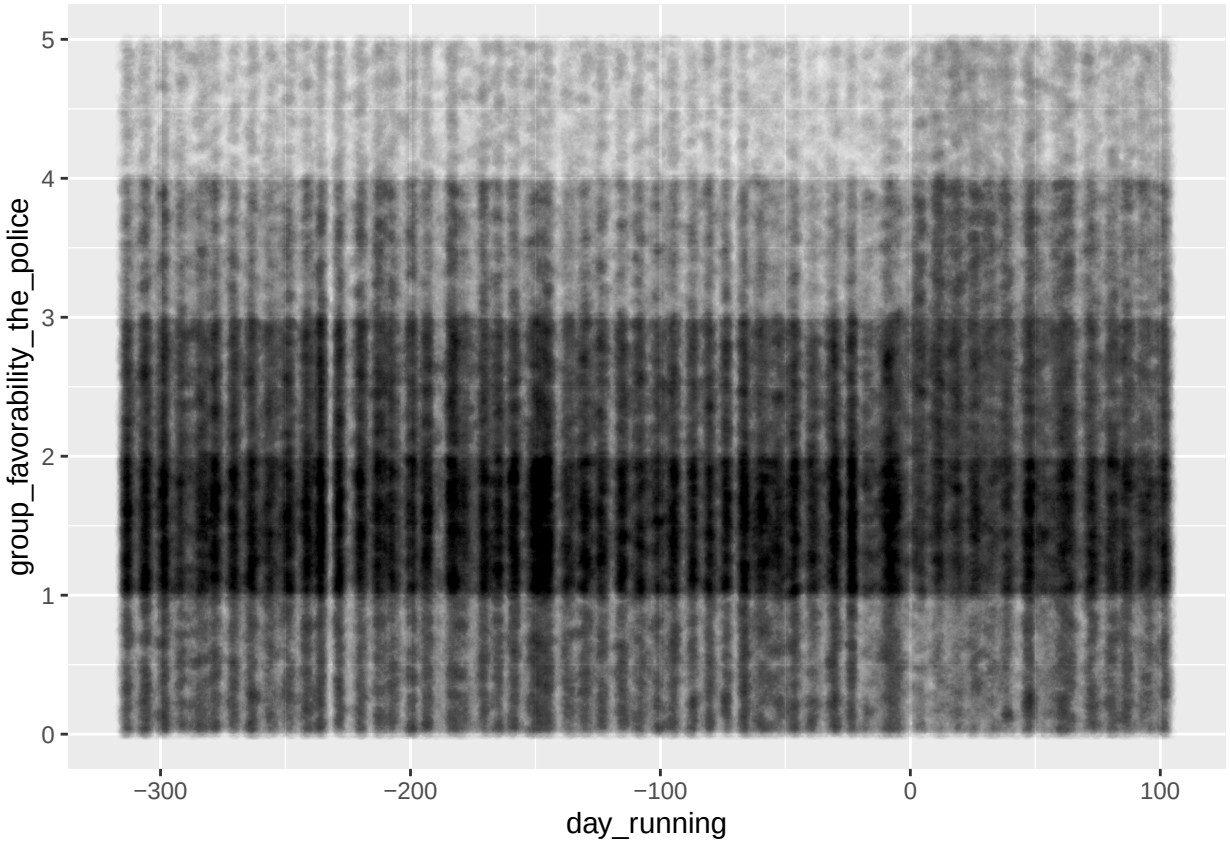
```
## Warning: Removed 51710 rows containing missing values (geom_point).
```




```r
ggplot(protest_df, aes(x = day_running, y = group_favorability_the_police)) +
  geom_point(alpha = 0.01, position=position_jitter(height=1, width=1))
```

```
## Don't know how to automatically pick scale for object of type haven_labelled/vctrs_vctr/double. Defau
```

```
## Warning: Removed 33807 rows containing missing values (geom_point).
```




```
# Honestly not altogether sure how functional form applies here and cannot find any resources as to how

# But I would recommend adding n_protests as a covariate, as that would pose an interesting question of

# Thus we have our new model:

new_model <- lm(group_favorability_the_police ~ party + day_running + n_protests + party*day_running + p

# To more easily compare the terms, let's standardize:

new_model_stand <- lm(group_favorability_the_police ~ party.standardised + day_running.standardised + n_
```

###Part c Linear models are not well suited for bounded ordinal responses. Instead ordinal logit or probit models are frequently employed in order to capture a) that the outcomes are restricted to a scale (in the case of police unfavorability 1-4) and b) that the differences between different rungs on the scale are not necessarily equivalent (going from very unfavorable to somewhat unfavorable is not necessarily the same difference as going from somewhat unfavorable to somewhat favorable). **Using the code below from the MASS package run an ordinal probit model using the same model as part b. How do the coefficients differ from part b?**

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.6.3
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
select <- dplyr::select

polr(data = protest_df_stand, formula = as.factor(group_favorability_the_police) ~ party.standardised +
```

```
## Call:
## polr(formula = as.factor(group_favorability_the_police) ~ party.standardised +
##     day_running.standardised + n_protests.standardised + party.standardised *
##     day_running.standardised + party.standardised * n_protests.standardised,
##     data = protest_df_stand, method = "probit")
##
## Coefficients:
##                            party.standardised
##                                    -0.33729850
##                      day_running.standardised
##                                     0.03862780
##                       n_protests.standardised
##                                     0.02178769
## party.standardised:day_running.standardised
##                                    -0.01821830
##  party.standardised:n_protests.standardised
##                                     0.01752583
##
## Intercepts:
##        1|2        2|3        3|4
## -0.3179435  0.5946187  1.2536540
##
## Residual Deviance: 810820.33
## AIC: 810836.33
## (51710 observations deleted due to missingness)
```