lab 8

Aisha Lakshman

3/11/2022

Packages

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(nnet)
library(knitr)
library(broom)
library(patchwork)
```

Data

```r
gss <- read_csv("gss2016.csv",
  na = c("", "Don't know", "No answer",
         "Not applicable"),
         guess_max = 2867) %>%
  select(natmass, age, sex, sei10, region, polviews) %>%
  drop_na()
```

```
## Rows: 2867 Columns: 935
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (810): wrkstat, marital, martype, childs, age, degree, sex, race, born, ...
## dbl (106): year, id_, hrs2, sphrs2, sibs, agekdbrn, educ, emailmin, emailhr,...
## lgl  (19): bigbang1, spwrkgvt, where6, away8, where8, away9, where9, mar10, ...
##
## 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.
```

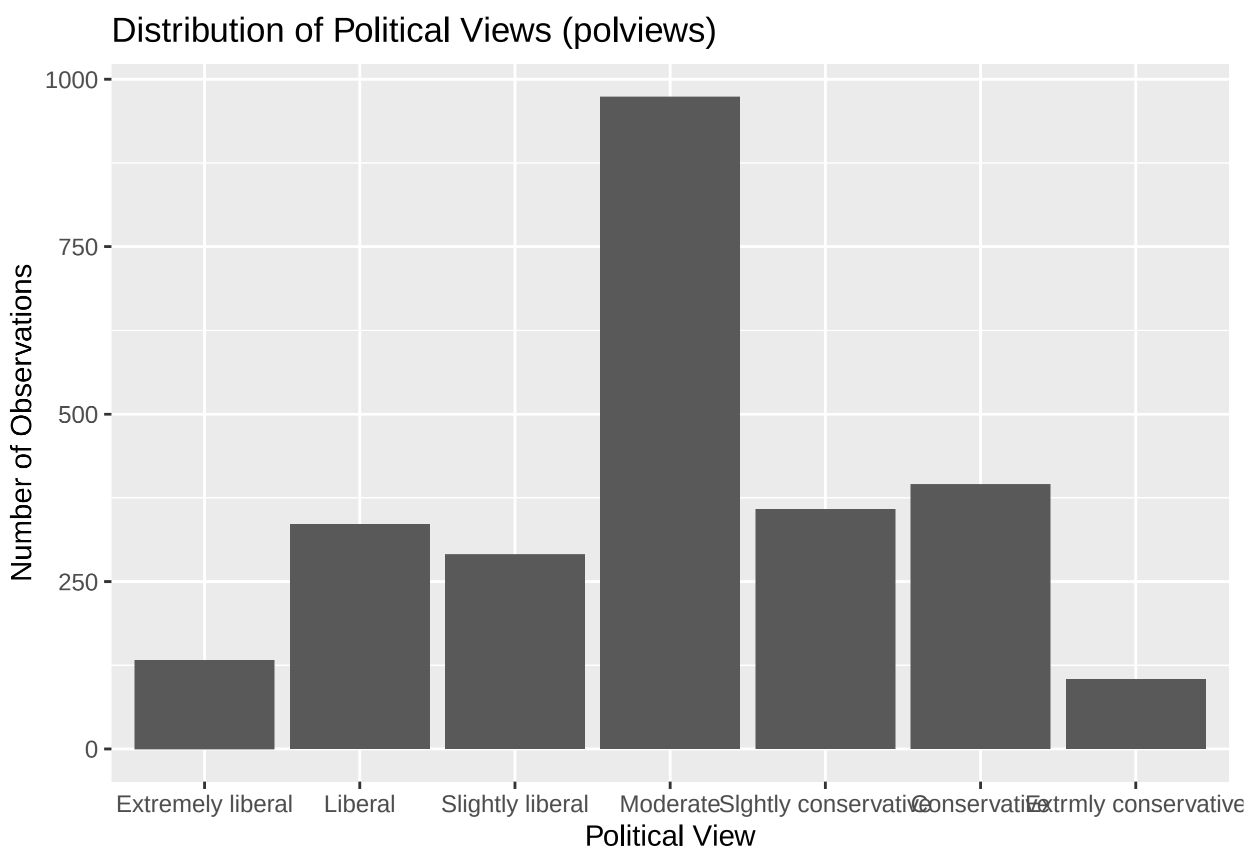
Part I: Exploratory Data Analysis

Exercise 1

```
gss <- gss %>%
  mutate(natmass = fct_relevel(natmass, "About right", "Too little", "Too much"))
```

Exercise 2

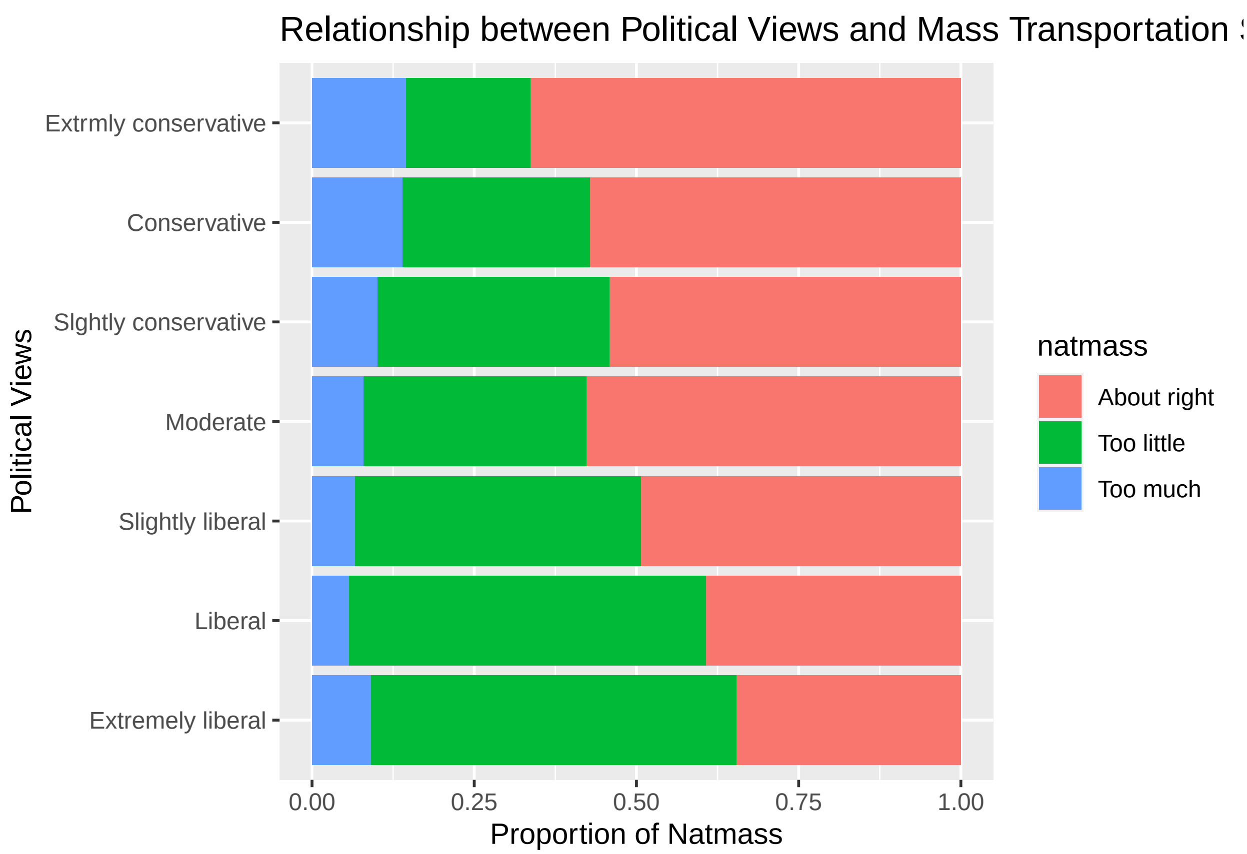
```
gss <- gss %>%
  mutate(polviews = fct_relevel(polviews, "Extremely liberal", "Liberal", "Slightly liberal", "Moderate

ggplot(data = gss, aes(x = polviews)) +
  geom_bar() +
  labs(title = "Distribution of Political Views (polviews)",
       x="Political View",
       y="Number of Observations")
```



The political view that occurs most frequently in this data set is "Moderate".

Exercise 3

```
ggplot(data = gss, aes(fill=natmass, x=polviews)) +
  geom_bar(position="fill") +
  labs(title = "Relationship between Political Views and Mass Transportation Spending Views",
       x="Political Views",
       y="Proportion of Natmass") +
  coord_flip()
```



This plot demonstrates that liberals believe that government spending on mass transportation is insufficient. The more conservative a person is, the more likely they believe government spending on mass transportation is adequate or excessive.

Exercise 4

```
gss <- gss %>%
  mutate(age = if_else(age == "89 or older", 89, as.numeric(age)))
```

```
## Warning in replace_with(out, !condition, false, fmt_args(~false), glue("length
## of {fmt_args(~condition)}")): NAs introduced by coercion
```

Part II: Multinomial Logistic Regression Model

Exercise 5

Because our response variable, Natmass, is a categorical variable with more than two categories, a multinomial logistic regression model is the best choice for this problem. Logistic regression is used to solve classification problems, and because our classifier has three categories, we cannot use a binomial model.

Exercise 6

```r
model <- multinom(natmass ~ ., data = gss)
```

```
## # weights:  57 (36 variable)
## initial  value 2845.405828
## iter  10 value 2308.054489
## iter  20 value 2277.361046
## iter  30 value 2276.038249
## iter  40 value 2275.922824
## final  value 2275.922640
## converged
```

```r
tidy(model) %>%
  kable(format = "markdown", digits = 4)
```

y.level	term	estimate	std.error	statistic	p.value
Too little	(Intercept)	-0.4149	0.2584	-1.6058	0.1083
Too little	age	0.0062	0.0025	2.4478	0.0144
Too little	sexMale	0.2174	0.0870	2.4996	0.0124
Too little	sei10	0.0081	0.0018	4.4463	0.0000
Too little	regionE. sou. central	0.3339	0.1923	1.7359	0.0826
Too little	regionMiddle atlantic	-0.0815	0.1674	-0.4865	0.6266
Too little	regionMountain	0.1377	0.1798	0.7658	0.4438
Too little	regionNew england	0.4660	0.2053	2.2701	0.0232
Too little	regionPacific	0.3637	0.1539	2.3636	0.0181
Too little	regionSouth atlantic	0.1319	0.1418	0.9296	0.3526
Too little	regionW. nor. central	0.0306	0.1993	0.1535	0.8780
Too little	regionW. sou. central	-0.0275	0.1715	-0.1606	0.8724
Too little	polviewsLiberal	-0.2016	0.2226	-0.9057	0.3651
Too little	polviewsSlightly liberal	-0.5969	0.2267	-2.6330	0.0085
Too little	polviewsModerate	-0.9695	0.2026	-4.7847	0.0000
Too little	polviewsSlghtly conservative	-0.9400	0.2224	-4.2264	0.0000
Too little	polviewsConservative	-1.2207	0.2237	-5.4558	0.0000
Too little	polviewsExtrmly conservative	-1.6962	0.3199	-5.3021	0.0000
Too much	(Intercept)	-1.8496	0.4356	-4.2463	0.0000
Too much	age	0.0143	0.0041	3.4804	0.0005
Too much	sexMale	0.5349	0.1462	3.6596	0.0003
Too much	sei10	-0.0099	0.0032	-3.0785	0.0021
Too much	regionE. sou. central	-0.3234	0.3508	-0.9217	0.3567
Too much	regionMiddle atlantic	-0.1435	0.2791	-0.5143	0.6070

y.level	term	estimate	std.error	statistic	p.value
Too much	regionMountain	-0.0255	0.3048	-0.0835	0.9334
Too much	regionNew england	0.8785	0.2922	3.0065	0.0026
Too much	regionPacific	0.3403	0.2438	1.3956	0.1628
Too much	regionSouth atlantic	-0.2740	0.2428	-1.1283	0.2592
Too much	regionW. nor. central	0.1593	0.3038	0.5243	0.6001
Too much	regionW. sou. central	-0.6018	0.3114	-1.9328	0.0533
Too much	polviewsLiberal	-0.6307	0.4113	-1.5333	0.1252
Too much	polviewsSlightly liberal	-0.6699	0.4110	-1.6298	0.1031
Too much	polviewsModerate	-0.6797	0.3510	-1.9362	0.0528
Too much	polviewsSlghtly conservative	-0.4011	0.3768	-1.0645	0.2871
Too much	polviewsConservative	-0.0798	0.3640	-0.2193	0.8264
Too much	polviewsExtrmly conservative	-0.3064	0.4429	-0.6918	0.4891

Exercise 7

The fact that the coeficients of the intercepts for "Too Little" and "Too Much" are both negative indicates that the model will favor the more neutral baseline in its predictions.

Exercise 8

The age coefficient of "Too little" versus the baseline is slightly positive. This indicates that as people get older, the likelihood that they believe mass transportation spending is insufficient rises.

Exercise 9

According to the null hypothesis, political views have no effect on attitudes toward spending on mass transportation. According to the alternative hypothesis, political beliefs influence people's attitudes toward spending on mass transportation. In terms of statistics, I will contrast the above model with one that does not include the polviews variable. The null hypothesis is true if the reduced model has a lower AIC.

```
reduced_model <- multinom(natmass ~ age + sex + sei10 + region, data = gss)
```

```
## # weights:  39 (24 variable)
## initial  value 2845.405828
## iter  10 value 2345.298055
## iter  20 value 2328.421434
## iter  30 value 2327.225660
## final  value 2327.223281
## converged
```

```
reduced_model$AIC
```

```
## [1] 4702.447
```

```
model$AIC
```

```
## [1] 4623.845
```

The model with the polviews variable has a lower AIC. As a result, the alternate hypothesis is correct. For the remainder of the lab, we will use the full model.

Part III: Model Fit

Exercise 11

```
fitted <- model$fitted.values
resid <- model$residuals
head(fitted)
```

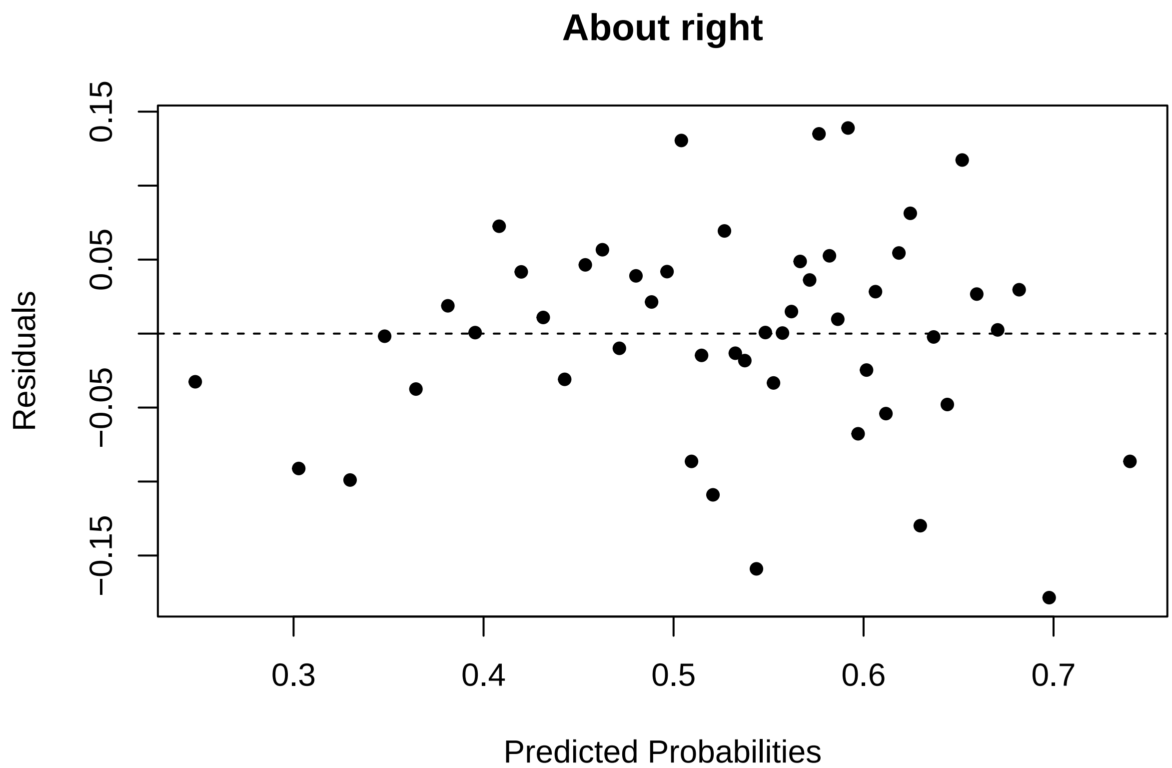
```
##   About right Too little   Too much
## 1   0.3824439  0.5151564 0.10239965
## 2   0.2715367  0.5756776 0.15278570
## 3   0.5246593  0.3253687 0.14997198
## 4   0.4155186  0.4653015 0.11917992
## 5   0.4138702  0.4762595 0.10987027
## 6   0.3142385  0.5904887 0.09527281
```

```
head(resid)
```

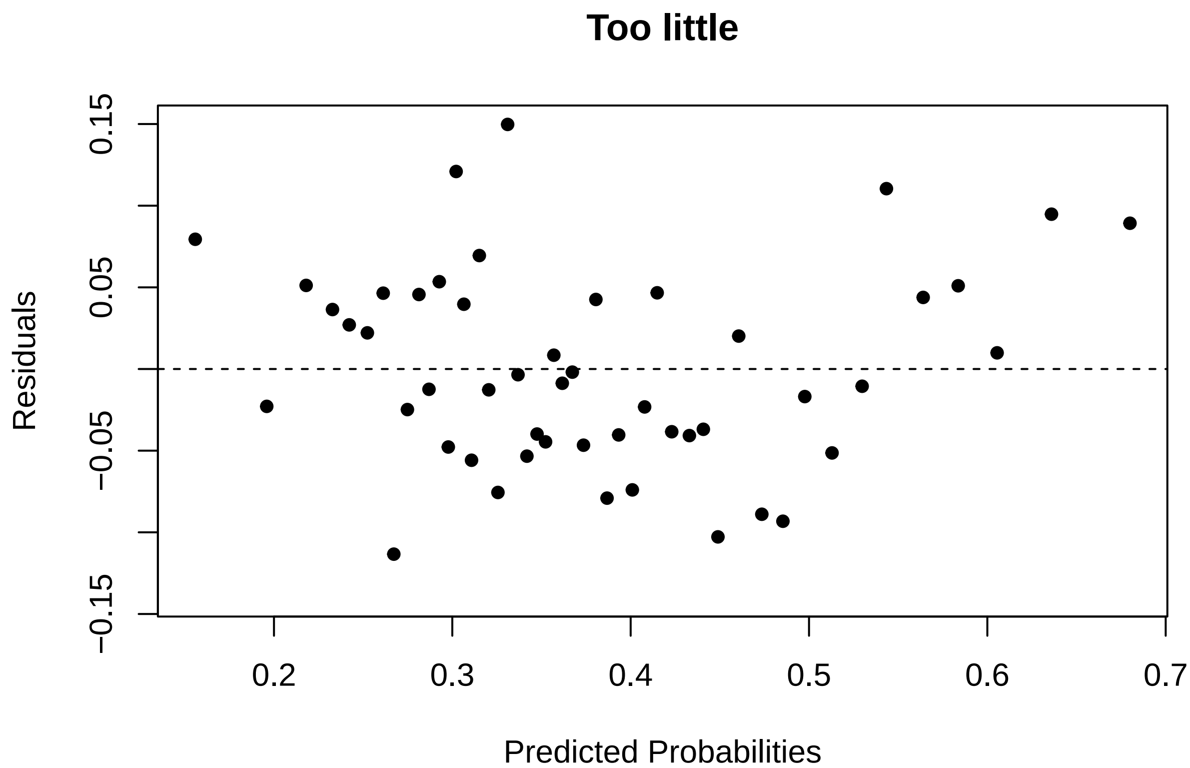
```
##   About right Too little    Too much
## 1  -0.3824439  0.4848436 -0.10239965
## 2  -0.2715367  0.4243224 -0.15278570
## 3  -0.5246593 -0.3253687  0.85002802
## 4  -0.4155186  0.5346985 -0.11917992
## 5   0.5861298 -0.4762595 -0.10987027
## 6  -0.3142385  0.4095113 -0.09527281
```

Exercise 12

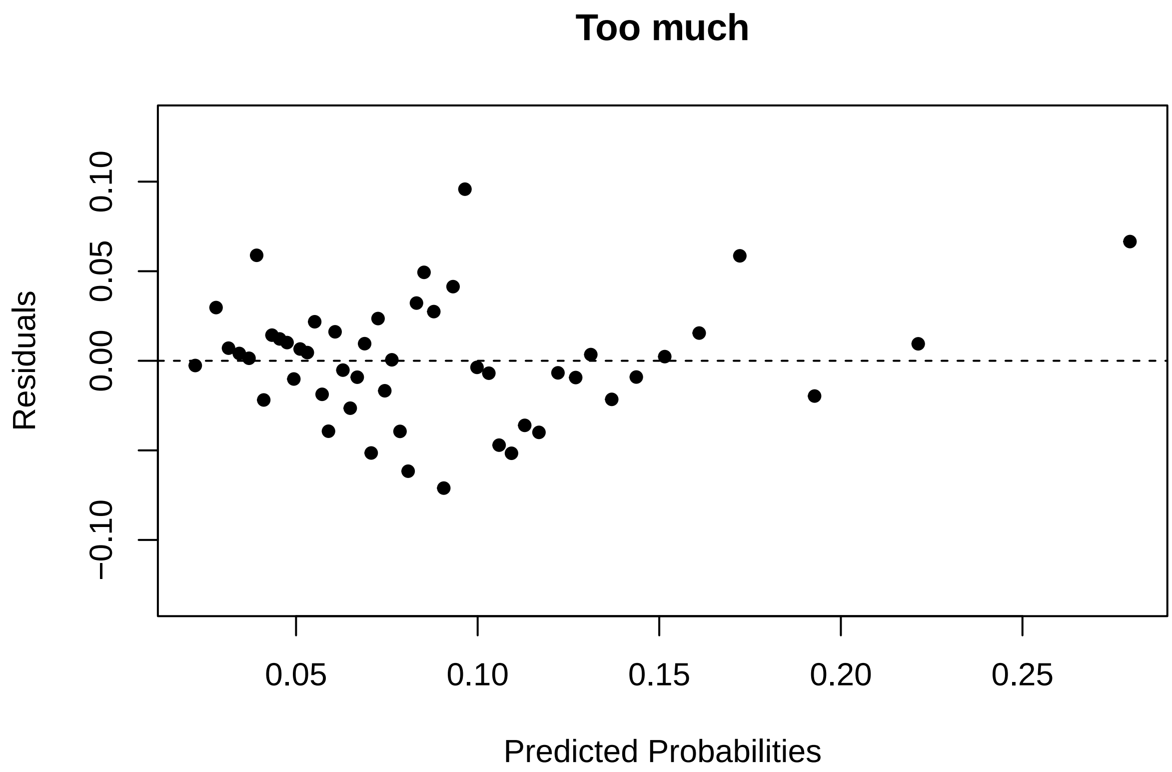
```
p1 <- arm::binnedplot(x = fitted[,1], y = resid[,1],
                xlab = "Predicted Probabilities",
                ylab = "Residuals",
                main = "About right",
                col.int = FALSE)
```




```
p2 <- arm::binnedplot(x = fitted[,2], y = resid[,2],
                xlab = "Predicted Probabilities",
                ylab = "Residuals",
                main = "Too little",
                col.int = FALSE)
```




```
p3 <- arm::binnedplot(x = fitted[,3], y = resid[,3],
                xlab = "Predicted Probabilities",
                ylab = "Residuals",
                main = "Too much",
                col.int = FALSE)
```



Exercise 13

```
aboutright_avg_resid <- mean(resid[,1])
toolittle_avg_resid <- mean(resid[,2])
toomuch_avg_resid <- mean(resid[,3])
aboutright_avg_resid
```

```
## [1] -2.238446e-06
```

```
toolittle_avg_resid
```

```
## [1] 1.712403e-06
```

```
toomuch_avg_resid
```

```
## [1] 5.260433e-07
```

Part IV: Using The Model

Exercise 16

According to the model, the more liberal an individual is, the more liberal their attitude toward spending on mass transportation is "too little". In contrast, the more conservative a person is, the more they believe that spending on mass transportation is "too much".

Exercise 17

```r
gss <- gss %>%
  mutate(pred_probs = predict(model, type = "class"))

gss %>%
  count(natmass, pred_probs)
```

```
## # A tibble: 8 x 3
##   natmass     pred_probs      n
##   <fct>       <fct>       <int>
## 1 About right About right  1151
## 2 About right Too little    219
## 3 About right Too much        2
## 4 Too little  About right   646
## 5 Too little  Too little    339
## 6 Too much    About right   196
## 7 Too much    Too little     36
## 8 Too much    Too much        1
```

misclassification rate = (219 + 2 + 646 + 196 + 36) / 2590 = 0.424