STAT 108: Lab 8

Tim Lanthier

3/10/2022

Github Repository: https://github.com/talanthier/lab-08

Lab 8: Multinomial Logistic Regression

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

In this lab we will be using data from the 2016 General Social Survey. Here we will use multinomial regression to understand the relationship between political views and a person's attitudes towards government spending on mass transportation.

```
gss <- read_csv("raw data/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.
```

```
glimpse(gss)
```

```
## Rows: 2,590
## Columns: 6
## $ natmass  <chr> "Too little", "Too little", "Too much", "Too little", "About ~
## $ age      <chr> "47", "61", "43", "55", "53", "50", "23", "71", "86", "32", "~
## $ sex      <chr> "Male", "Male", "Female", "Female", "Female", "Male", "Female~
## $ sei10    <dbl> 87.9, 38.3, 21.8, 39.7, 44.6, 80.7, 20.1, 32.0, 13.2, 20.8, 2~
## $ region   <chr> "New england", "New england", "New england", "New england", "~
## $ polviews <chr> "Moderate", "Liberal", "Moderate", "Slightly liberal", "Sligh~
```

Here we will be trying to predict `natmass`, which is each person's response to the following prompt:

> "We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount... are we spending too much, too little, or about the right amount on mass transportation?"

Note the `polviews` is the response to the following prompt:

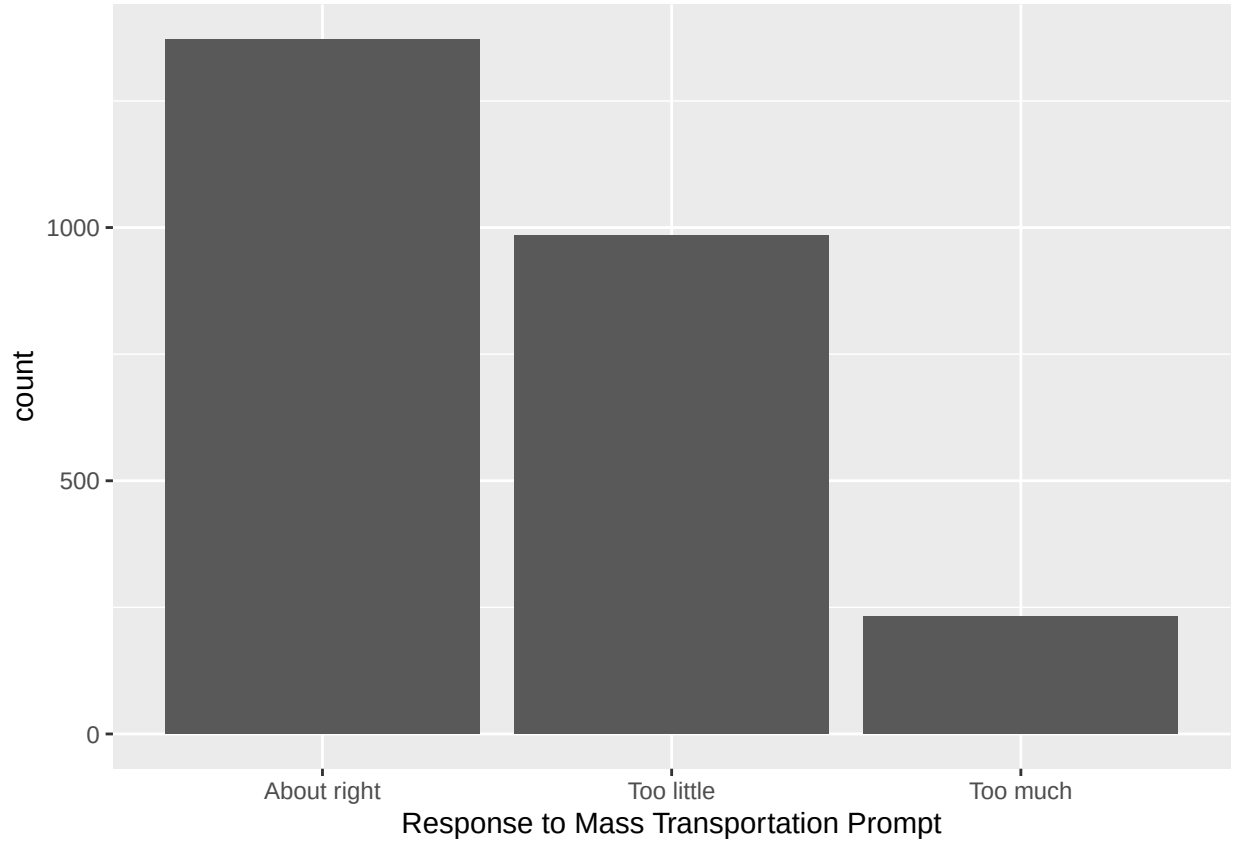
> "We hear a lot of talk these days about liberals and conservatives. I'm going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal - point 1 - to extremely conservative - point 7. Where would you place yourself on this scale?"

Exploratory Data Analysis

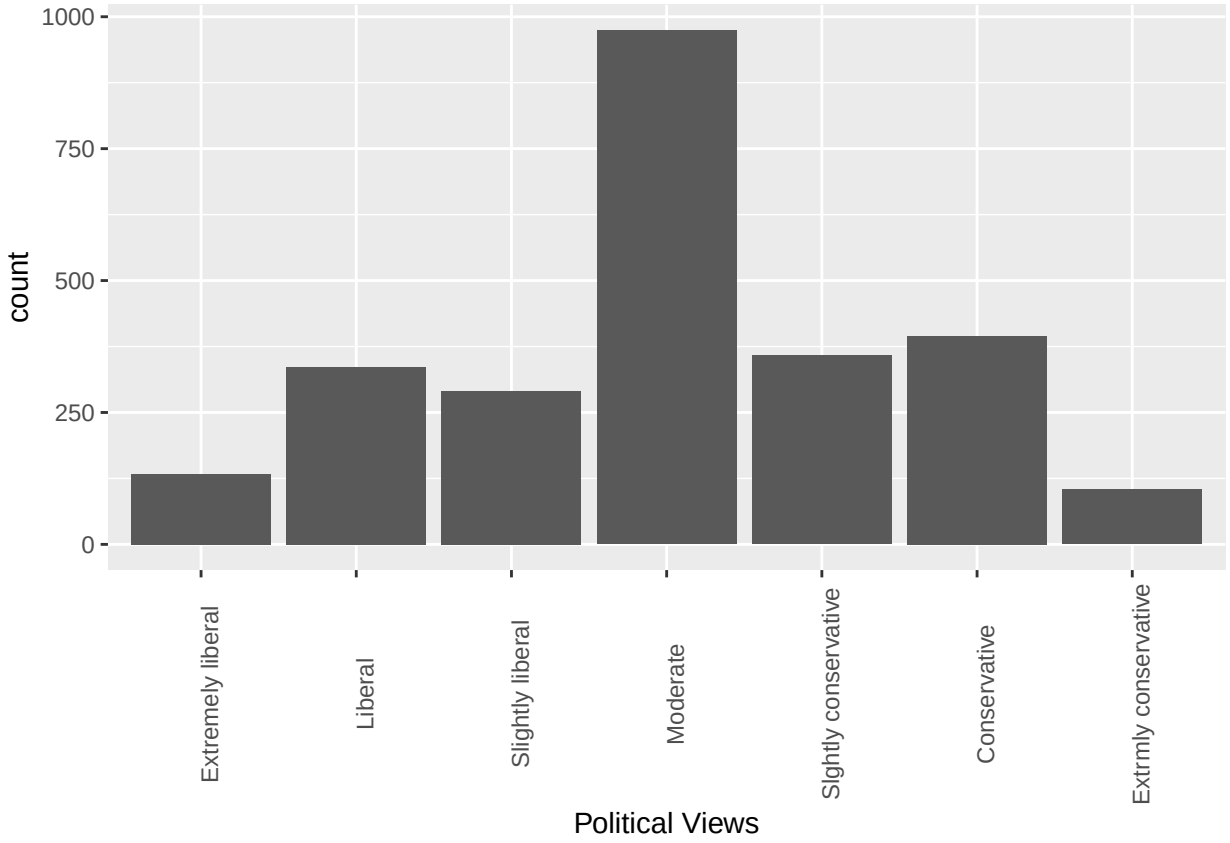
We will start by converting `natmass` to a factor with "About right" as the baseline. We're also going to reorder the levels for `polviews` so the ordering makes more sense.

```
gss <- gss %>%
  mutate(natmass = relevel(as.factor(natmass), "About right"),
         polviews = fct_relevel(polviews, 'Extremely liberal', 'Liberal', 'Slightly liberal',
                                                  'Moderate', 'Slghtly conservative', 'Conservative',
```

```
ggplot(data = gss, aes(natmass)) +
  geom_bar() +
  labs(x = 'Response to Mass Transportation Prompt')
```

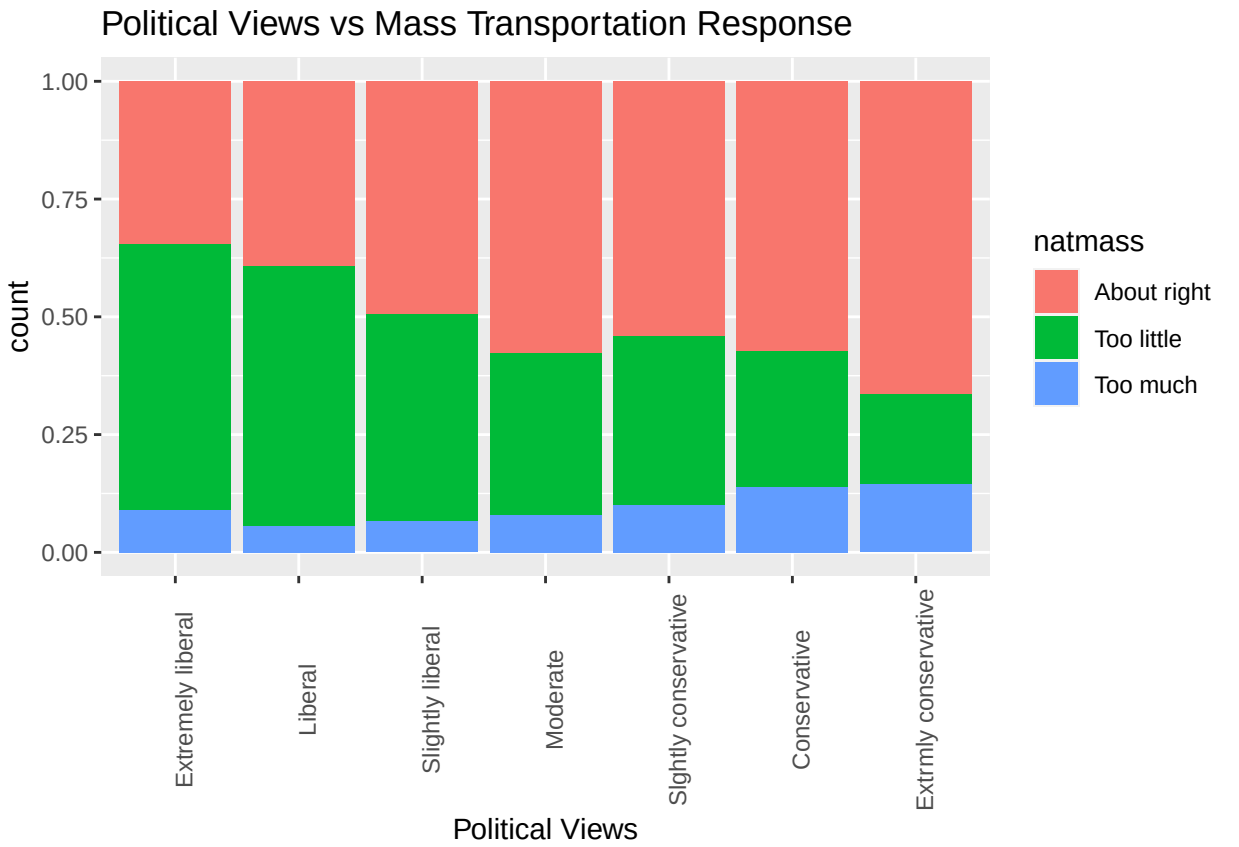


```
ggplot(data = gss, aes(polviews)) +
  geom_bar() +
  labs(x = 'Political Views')+
  theme(axis.text.x = element_text(angle = 90))
```



So here it looks like most people in the dataset identify themselves as moderate. Also most participants appear to think that the government is either spending "about much" or "too little" on mass transportation projects. Now we will look at the interaction between these 2 variables.

```
ggplot(data = gss, aes(fill = natmass, x = polviews)) +
  geom_bar(position = 'fill') +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = 'Political Views vs Mass Transportation Response', x = 'Political Views')
```



Unsurprisingly, there appears to be a pretty clear relationship between political views and their thoughts on government spending on mass transportation. It appears that the proportion of those who's response to the mass transportation prompt is "about right" or "too much" is larger for those who identify themselves as more conservative. Meanwhile there is a larger proportion of those who had the response of "too little" for those who identify themselves as more liberal.

Now we will investigate `age`. First we will convert `age` to a numeric variable. Note that we currently have an age for "89 or older". So we will treat all those who are "89 or older" as just 89. The distribution for `age` is as follows:

```
gss['age'][gss['age'] == '89 or older'] = '89'
gss <- gss %>%
  mutate(age = as.numeric(age))

ggplot(data = gss, aes(age)) +
  geom_histogram(binwidth = 2) +
  labs(x = 'Age', title = 'Distribution of Age')
```

Multinomial Logistic Regression Model

Now we will fit a model using `age`, `sex`, `sei10`, and `region` to understand the difference in opinions on mass transportation government spending. Seeing as `natmass` contains 3 different responses. Since there are more than 2 possible responses, it would be inappropriate to us Logistic Regression. Hence a multinomial logistic regression model would be more appropriate.

So the interpretations of the intercept make more sense, we will also be using `age_cent` (`age` centered about its median) instead of `age`. We will use the median since the mean likely will not be an integer and we want to keep `age_cent` as integer values. Seeing as `age` doesn't appear to be heavily skewed, using the median instead of the mean will have little impact on our model.

```
gss <- gss %>%
  mutate(age_cent = age - median(age),
         sex = as.factor(sex),
         region = as.factor(region))

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

```
## # weights:  39 (24 variable)
## initial  value 2845.405828
## iter  10 value 2380.898304
## iter  20 value 2331.976986
## iter  30 value 2327.224253
## final  value 2327.223281
## converged
```

```
tidy(model, conf.int = TRUE, exponentiate = FALSE) %>%
  kable(digits = 3, format = 'markdown')
```

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
Too little	(Intercept)	-1.001	0.139	-7.208	0.000	-1.273	-0.729
Too little	age_cent	0.004	0.002	1.607	0.108	-0.001	0.009
Too little	sexMale	0.196	0.085	2.300	0.021	0.029	0.364
Too little	sei10	0.009	0.002	5.282	0.000	0.006	0.013
Too little	regionE. sou. central	0.272	0.189	1.434	0.151	-0.099	0.643
Too little	regionMiddle atlantic	-0.030	0.164	-0.184	0.854	-0.352	0.292
Too little	regionMountain	0.183	0.177	1.034	0.301	-0.164	0.529
Too little	regionNew england	0.595	0.201	2.959	0.003	0.201	0.989
Too little	regionPacific	0.409	0.151	2.704	0.007	0.112	0.705
Too little	regionSouth atlantic	0.123	0.139	0.883	0.377	-0.150	0.396
Too little	regionW. nor. central	0.030	0.196	0.151	0.880	-0.355	0.414
Too little	regionW. sou. central	-0.086	0.169	-0.508	0.611	-0.417	0.245
Too much	(Intercept)	-1.630	0.221	-7.378	0.000	-2.063	-1.197
Too much	age_cent	0.016	0.004	3.945	0.000	0.008	0.024
Too much	sexMale	0.553	0.145	3.809	0.000	0.269	0.838
Too much	sei10	-0.010	0.003	-3.018	0.003	-0.016	-0.003
Too much	regionE. sou. central	-0.285	0.350	-0.816	0.414	-0.970	0.400
Too much	regionMiddle atlantic	-0.162	0.278	-0.585	0.559	-0.707	0.382
Too much	regionMountain	-0.021	0.303	-0.070	0.944	-0.616	0.574
Too much	regionNew england	0.853	0.289	2.946	0.003	0.285	1.420
Too much	regionPacific	0.296	0.242	1.221	0.222	-0.179	0.771
Too much	regionSouth atlantic	-0.263	0.242	-1.086	0.278	-0.737	0.212
Too much	regionW. nor. central	0.138	0.302	0.458	0.647	-0.454	0.730
Too much	regionW. sou. central	-0.583	0.310	-1.878	0.060	-1.191	0.026

The model for the log odds is shown above. We have an intercept for "too much" of -1.630. This means that a person who is of median age (49 in our dataset), female, from the East North Central region, and with a Social Economic Index of 0 will have odds of having the opinion "Too much" over "About right" of $e^{-1.63} = 0.196$.

Now we also have a coefficient for `age_cent` in the "too little" part of our model of 0.004. This means that holding all else constant, if the same person were to be one year older, the odds of having the opinion "too little" over "about right" would be multiplied by a factor of $e^{0.004} = 1.004$ from the original level.

Now we will examine the effect of political views. To do this, we will using a Chi-squared test to check whether it would be appropriate to add `polviews` to our model. If we let $\beta_{polviews}$ be the coefficient for `polviews` in our regression model we have the following hypotheses.

$$H_0 : \beta_{polviews} = 0$$
$$H_a : \beta_{polviews} \neq 0$$

So we have the null hypothesis that the true coefficient for `polviews` would be 0. That is `polviews` is an insignificant predictor of the response to the survey. Meanwhile the alternate hypothesis is that the true coefficient is nonzero. That is `polviews` is a significant predictor of the response to the survey.

```
full_model <- multinom(natmass ~ age_cent + sex + sei10 + region + polviews, data = gss)
```

```
## # weights:  57 (36 variable)
## initial  value 2845.405828
## iter  10 value 2328.597380
```

```
## iter  20 value 2280.584841
## iter  30 value 2276.240392
## iter  40 value 2275.930072
## final  value 2275.922613
## converged
```

```
anova(model, full_model, test = "Chisq") %>%
  kable(format = "markdown", digits = 3)
```

Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
age_cent + sex + sei10 + region	5156	4654.447		NA	NA	NA
age_cent + sex + sei10 + region + polviews	5144	4551.845	1 vs 2	12	102.601	0

As we see, we have a p-value of very close to 0 (at least to 3 significant digits). Since this is the case, we have sufficient evidence to reject the null hypothesis that $\beta_{polviews} = 0$. Hence `polviews` is a significant predictor of `natmass`. Thus for the remainder of the lab we will be using `full_model` which includes `polviews`.

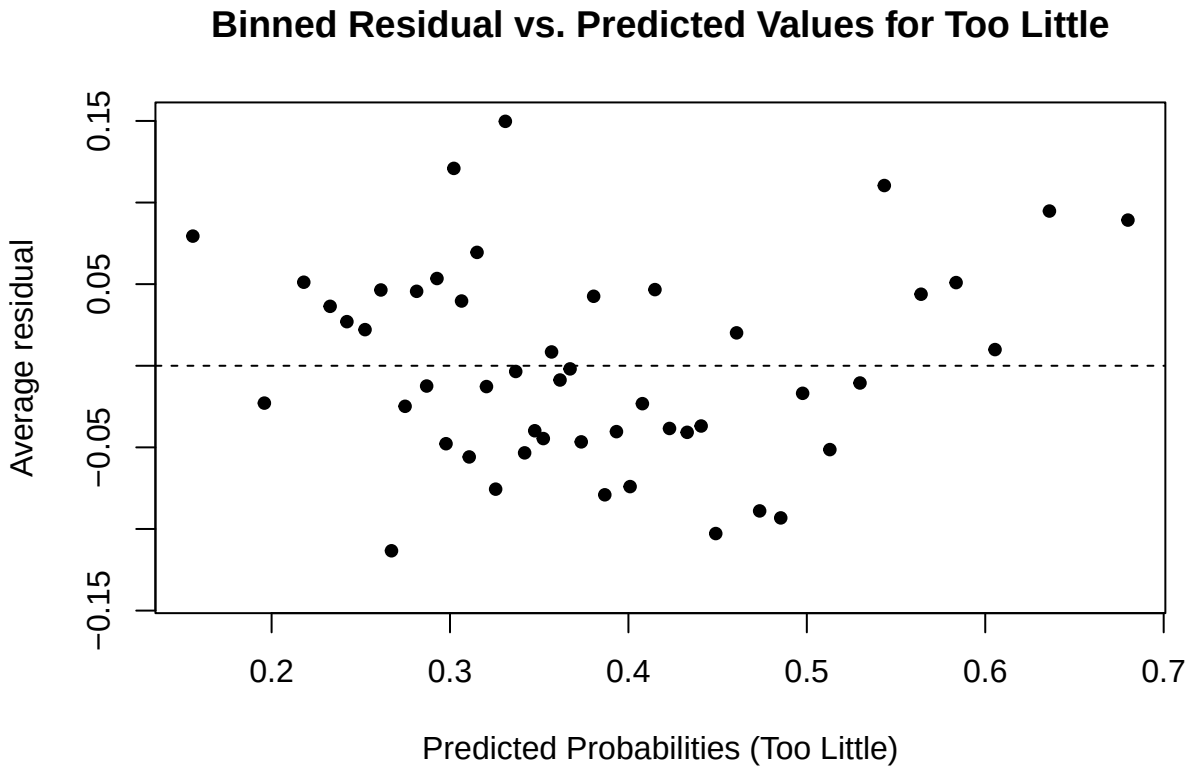
Model Fit

Now we will assess the fit of our full model.

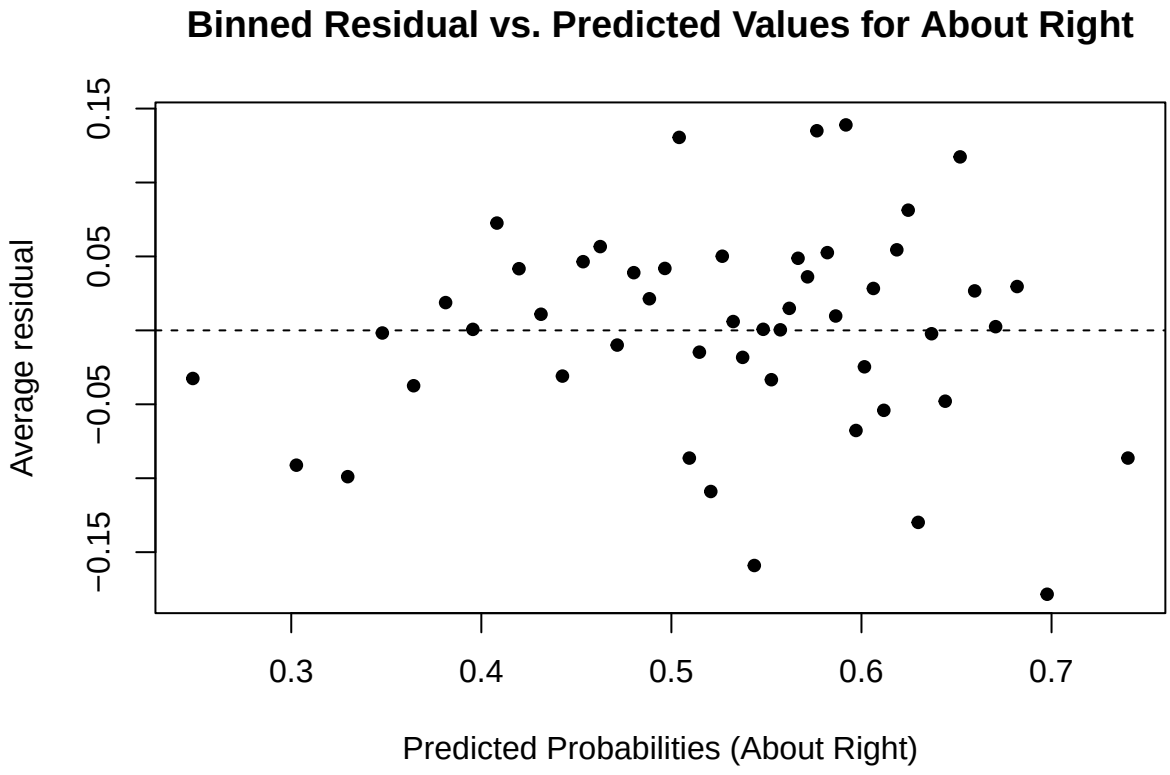
```
gss <- gss %>%
  mutate(obs_num = 1:n())

pred_probs <- as_tibble(predict(full_model, type = 'probs')) %>%
  mutate(obs_num = 1:n())


resid <- as_tibble(residuals(full_model)) %>%
  mutate(obs_num = 1:n())

full_model_aug <- inner_join(gss, resid, by = 'obs_num') %>%
  mutate(preds = predict(full_model, type = 'class'))

full_model_aug <- inner_join(full_model_aug, pred_probs, by = 'obs_num')
```

Note that `About right.x` are the residuals and `About right.y` are the predicted probabilities. Now we will check the linearity assumption for our multinomial logistic regression model.

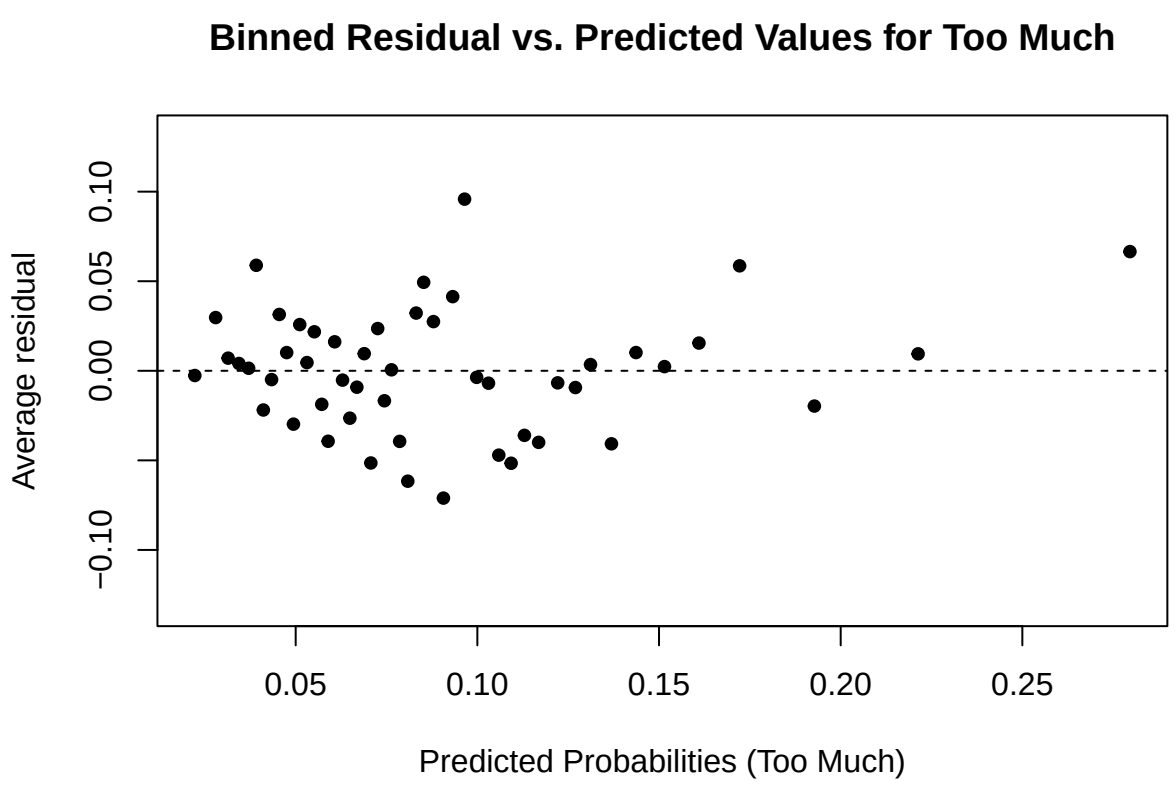
```
nbins <- sqrt(nrow(full_model_aug))

arm::binnedplot(x = full_model_aug$'Too little.y', y = full_model_aug$'Too little.x',
                xlab = "Predicted Probabilities (Too Little)",
                main = "Binned Residual vs. Predicted Values for Too Little",
                col.int = FALSE)
```




```
arm::binnedplot(x = full_model_aug$'About right.y', y = full_model_aug$'About right.x',
                xlab = "Predicted Probabilities (About Right)",
                main = "Binned Residual vs. Predicted Values for About Right",
                col.int = FALSE)
```




```
arm::binnedplot(x = full_model_aug$'Too much.y', y = full_model_aug$'Too much.x',
                xlab = "Predicted Probabilities (Too Much)",
                main = "Binned Residual vs. Predicted Values for Too Much",
                col.int = FALSE)
```




```
full_model_aug %>%
  group_by(sex) %>%
  summarise(mean_resid_about_right = mean(`Too little.x`),
            mean_resid_about_right = mean(`About right.x`),
            mean_resid_about_right = mean(`Too much.x`))
```

```
## # A tibble: 2 x 2
##   sex    mean_resid_about_right
##   <fct>                   <dbl>
## 1 Female            0.000000214
## 2 Male             -0.000000106
```

```
full_model_aug %>%
  group_by(region) %>%
  summarise(mean_resid_about_right = mean(`Too little.x`),
            mean_resid_about_right = mean(`About right.x`),
            mean_resid_about_right = mean(`Too much.x`))
```

```
## # A tibble: 9 x 2
##   region          mean_resid_about_right
##   <fct>                            <dbl>
## 1 E. nor. central           0.000000163
## 2 E. sou. central          -0.0000000769
## 3 Middle atlantic          -0.0000000516
## 4 Mountain                 -0.0000000707
## 5 New england               0.000000286
## 6 Pacific                   0.000000110
```

```
## 7 South atlantic            0.0000000982
## 8 W. nor. central           0.0000000204
## 9 W. sou. central           0.0000000244
```

```
full_model_aug %>%
  group_by(polviews) %>%
  summarise(mean_resid_about_right = mean(`Too little.x`),
            mean_resid_about_right = mean(`About right.x`),
            mean_resid_about_right = mean(`Too much.x`))
```

```
## # A tibble: 7 x 2
##   polviews             mean_resid_about_right
##   <fct>                                 <dbl>
## 1 Extremely liberal                0.000000368
## 2 Liberal                         -0.000000384
## 3 Slightly liberal                -0.000000115
## 4 Moderate                         0.000000110
## 5 Slghtly conservative             0.0000000579
## 6 Conservative                     0.000000362
## 7 Extrmly conservative             0.000000182
```

Looking at the binned residual plots, there does not appear to be a clear relationship between the binned residuals and predicted probabilities. Looking at the average residuals across the groups for the categorical variables, it looks like the mean residuals are consistent across groups (all around order of 10^{-7} to 10^{-8}). Based off of these plots, there doesn't appear to be any violations to the linearity assumption.

Using the Model

Now we will take a look at our model.

```
tidy(full_model, conf.int = TRUE) %>%
  kable(digits = 3, format = 'markdown')
```

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
Too little	(Intercept)	-0.113	0.233	-0.486	0.627	-0.570	0.344
Too little	age_cent	0.006	0.003	2.448	0.014	0.001	0.011
Too little	sexMale	0.217	0.087	2.500	0.012	0.047	0.388
Too little	sei10	0.008	0.002	4.446	0.000	0.005	0.012
Too little	regionE. sou. central	0.334	0.192	1.736	0.083	-0.043	0.711
Too little	regionMiddle atlantic	-0.082	0.167	-0.488	0.626	-0.410	0.246
Too little	regionMountain	0.138	0.180	0.766	0.444	-0.215	0.490
Too little	regionNew england	0.466	0.205	2.271	0.023	0.064	0.868
Too little	regionPacific	0.364	0.154	2.363	0.018	0.062	0.665
Too little	regionSouth atlantic	0.132	0.142	0.929	0.353	-0.146	0.410
Too little	regionW. nor. central	0.030	0.199	0.152	0.879	-0.360	0.421
Too little	regionW. sou. central	-0.028	0.171	-0.161	0.872	-0.364	0.309
Too little	polviewsLiberal	-0.202	0.223	-0.906	0.365	-0.638	0.235
Too little	polviewsSlightly liberal	-0.597	0.227	-2.633	0.008	-1.041	-0.153
Too little	polviewsModerate	-0.969	0.203	-4.785	0.000	-1.367	-0.572
Too little	polviewsSlghtly conservative	-0.940	0.222	-4.226	0.000	-1.376	-0.504
Too little	polviewsConservative	-1.221	0.224	-5.456	0.000	-1.659	-0.782
Too little	polviewsExtrmly conservative	-1.696	0.320	-5.302	0.000	-2.323	-1.069
Too much	(Intercept)	-1.148	0.392	-2.931	0.003	-1.916	-0.380
Too much	age_cent	0.014	0.004	3.480	0.001	0.006	0.022
Too much	sexMale	0.535	0.146	3.660	0.000	0.248	0.821

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
Too much	sei10	-0.010	0.003	-3.079	0.002	-0.016	-0.004
Too much	regionE. sou. central	-0.324	0.351	-0.922	0.356	-1.011	0.364
Too much	regionMiddle atlantic	-0.144	0.279	-0.516	0.606	-0.691	0.403
Too much	regionMountain	-0.025	0.305	-0.081	0.935	-0.622	0.573
Too much	regionNew england	0.879	0.292	3.009	0.003	0.306	1.452
Too much	regionPacific	0.340	0.244	1.395	0.163	-0.138	0.818
Too much	regionSouth atlantic	-0.274	0.243	-1.129	0.259	-0.750	0.202
Too much	regionW. nor. central	0.158	0.304	0.521	0.603	-0.437	0.754
Too much	regionW. sou. central	-0.601	0.311	-1.931	0.053	-1.211	0.009
Too much	polviewsLiberal	-0.630	0.411	-1.533	0.125	-1.437	0.176
Too much	polviewsSlightly liberal	-0.671	0.411	-1.631	0.103	-1.476	0.135
Too much	polviewsModerate	-0.680	0.351	-1.936	0.053	-1.368	0.008
Too much	polviewsSlghtly conservative	-0.401	0.377	-1.064	0.287	-1.140	0.337
Too much	polviewsConservative	-0.080	0.364	-0.219	0.827	-0.793	0.634
Too much	polviewsExtrmly conservative	-0.306	0.443	-0.692	0.489	-1.174	0.562

Looking at our model, we see the coefficient for the political views for Too little appear to all be negative with those more liberal closer to 0. This suggests that more conservative political views are associated with a lower odds of the response to the prompt being `Too little`. Meanwhile the coefficients for `Too much` political views are also negative, but are farther from 0 for more liberal views. So the log odds of their response being "too much" over "about right" decreases across all political views. The same is true for "too little". But the log odds of responding "too little" and additionally the odds themselves decreases more for those with more conservative political views compared to the baseline of "about right". Meanwhile the odds of responding "too much" decreases more with more liberal people than those with more conservative views. So it looks like those with more conservative views tend to respond with "too much" more than those with more liberal views. Also those with more conservative views tend to respond with "too little" less than those with more liberal views. That being said, overall, both more left and right leaning participants seem to respond with "about right."

Now we will look at the predictions we made.

```
full_model_aug %>%
  group_by(natmass, preds) %>%
  summarise(n = n()) %>%
  kable(format="markdown")
```

```
## `summarise()` has grouped output by 'natmass'. You can override using the `.groups` argument.
```

natmass	preds	n
About right	About right	1151
About right	Too little	219
About right	Too much	2
Too little	About right	645
Too little	Too little	340
Too much	About right	196
Too much	Too little	36
Too much	Too much	1

So it looks like we misclassified $36 + 196 + 645 + 2 + 219 = 1098$ of our observations. With 2590 observations in our dataset, we have a misclassification rate of $\frac{1098}{2590} = 0.424$.