Lab 4 - Multinomial Regression - Questions

Yifei Chen

2025-02-28

Lab Goal: Predict voting frequency using demographic variables Data source: FiveThirtyEight "Why Many Americans Don't Vote" survey Method: Multinomial logistic regression

0.1 Data

The data for this assignment comes from an online Ipsos survey that was conducted for the FiveThirtyEight article "Why Many Americans Don't Vote". You can read more about the survey design and respondents in the README of the GitHub repo for the data.

Respondents were asked a variety of questions about their political beliefs, thoughts on multiple issues, and voting behavior. We will focus on using the demographic variables and someone's party identification to understand whether a person is a probable voter.

The variables we'll focus on were (definitions from the codebook in data set GitHub repo):

- `ppage`: Age of respondent
- `educ`: Highest educational attainment category.
- `race`: Race of respondent, census categories. Note: all categories except Hispanic were non-Hispanic.
- `gender`: Gender of respondent
- `income_cat`: Household income category of respondent
- `Q30`: Response to the question "Generally speaking, do you think of yourself as a…"
  - 1: Republican
  - 2: Democrat
  - 3: Independent
  - 4: Another party, please specify
  - 5: No preference

    - -1: No response

- `voter_category`: past voting behavior:

    - **always**: respondent voted in all or all-but-one of the elections they were eligible in
    - **sporadic**: respondent voted in at least two, but fewer than all-but-one of the elections they were eligible in
    - **rarely/never**: respondent voted in 0 or 1 of the elections they were eligible in

You can read in the data directly from the GitHub repo:

```r
library(nnet)
library(car)
library(tidyverse)
library(emmeans)
library(ggeffects)
library(knitr)
library(patchwork)
library(broom)
library(parameters)
library(easystats)
```

```r
voter_data <- read_csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/non-vot
```

1 Lab

- The variable `Q30` contains the respondent's political party identification. Make a new variable that simplifies `Q30` into four categories: "Democrat", "Republican", "Independent", "Other" ("Other" also includes respondents who did not answer the question).

```r
voter_data <- voter_data %>%
  mutate(pol_ident_new = case_when(
    Q30==1 ~ "Rep",
    Q30==2 ~ "Dem",
    Q30==3 ~ "Indep",
    TRUE ~ "Other"
  ))
```

- The variable `voter_category` identifies the respondent's past voter behavior. Relevel the variable to make rarely/never the baseline level, followed by sporadic, then always

```r
#Enter your code
voter_data$voter_category <- factor(voter_data$voter_category,
                                    levels = c("rarely/never", "sporadic", "always"))
levels(voter_data$voter_category)
```

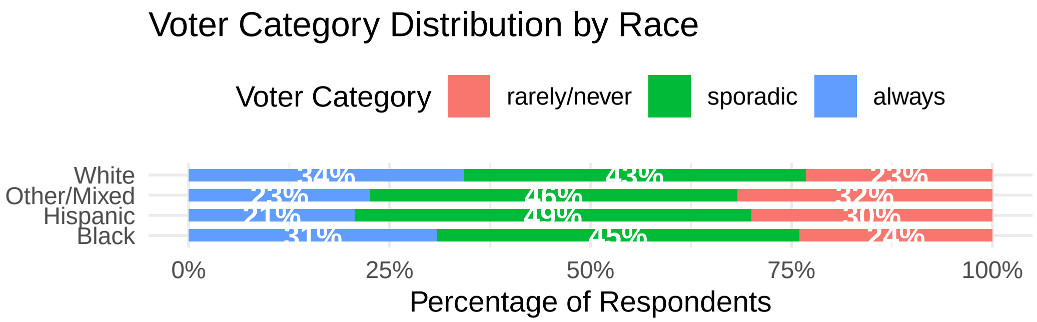
```
[1] "rarely/never" "sporadic"     "always"
```

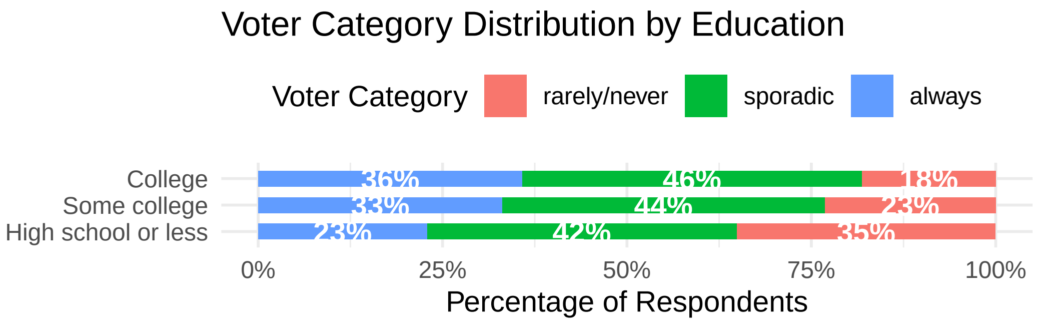
- Center the age variable to make the intercept more interepretable. That is, so that it reflects the log-odds for an average-aged person rather than a 0-year old person

```r
# enter code
voter_data$age_centered <- voter_data$ppage - mean(voter_data$ppage, na.rm = TRUE)
```

- In the FiveThirtyEight article, the authors include visualizations of the relationship between the voter category and demographic variables such as race, age, education, etc. Select two demographic variables. For each variable, try to replicate the visualizations and interpret the plot to describe its relationship with voter category. Have fun with it: https://www.mikelee.co/posts/2020-02-08-recreate-fivethirtyeight-chicklet-stacked-bar-chart-in-ggplot2.

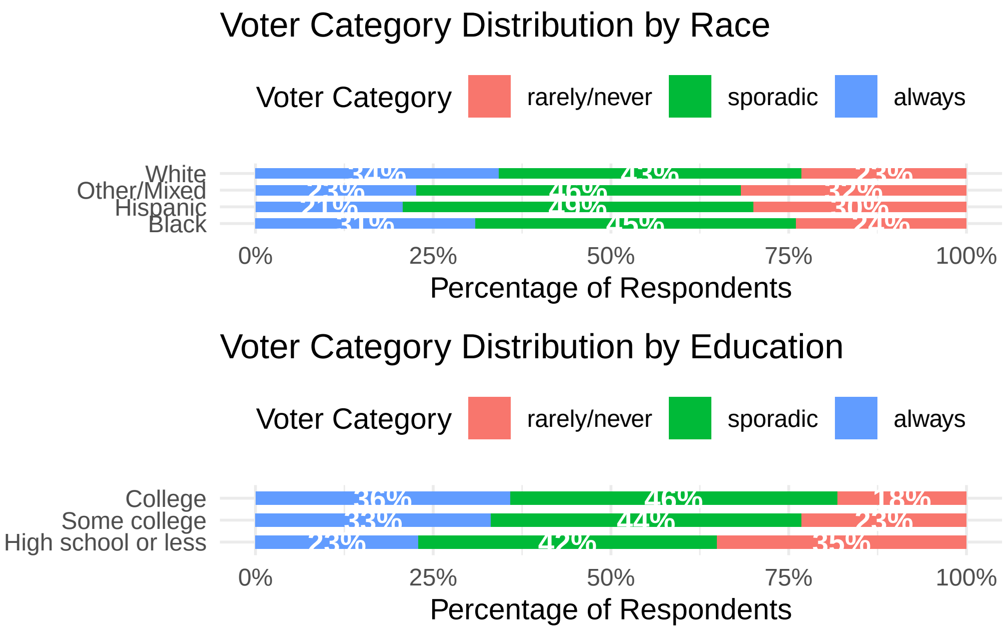
```r
# library
library(ggplot2)
library(viridis)
library(cowplot)
library(forcats)
library(scales)

# Enter code
race_data <- voter_data %>%
  group_by(race, voter_category) %>%
  summarise(count = n(), .groups = 'drop') %>%
  group_by(race) %>%
  mutate(percentage = count / sum(count))

education_data <- voter_data %>%
  group_by(educ, voter_category) %>%
  summarise(count = n(), .groups = 'drop') %>%
  group_by(educ) %>%
  mutate(percentage = count / sum(count))

education_data$educ <- factor(education_data$educ, levels=c("High school or less", "Some col
plot_race <- ggplot(race_data, aes(x = percentage, y = race, fill = voter_category)) +
  geom_col(position = "stack", width = 0.6) +  # Stacked bars
  geom_text(aes(label = scales::percent(percentage, accuracy = 1)),
            position = position_stack(vjust = 0.5), size = 4, fontface = "bold", color = "whi
  scale_x_continuous(labels = scales::percent_format(accuracy = 1)) +
  labs(
    title = "Voter Category Distribution by Race",
    x = "Percentage of Respondents", y = NULL,
    fill = "Voter Category"
  ) +
  theme_minimal() +
  theme(legend.position = "top")

plot_education <- ggplot(education_data, aes(x = percentage, y = educ, fill = voter_category)
  geom_col(position = "stack", width = 0.6) +  # Stacked bars
  geom_text(aes(label = scales::percent(percentage, accuracy = 1)),
            position = position_stack(vjust = 0.5), size = 4, fontface = "bold", color = "whi
  scale_x_continuous(labels = scales::percent_format(accuracy = 1)) +
  labs(
    title = "Voter Category Distribution by Education",
    x = "Percentage of Respondents", y = NULL,
    fill = "Voter Category"
  ) +
  theme_minimal() +
  theme(legend.position = "top")

plot_grid(plot_race, plot_education, ncol = 1)
```





The plots can be combined into a single plot using the patchwork package.

```r
library(patchwork)
combined_plot <- plot_race + plot_education + plot_layout(ncol = 1)
combined_plot
```



- Fit a model using mean-centered age, race, gender, income, and education to predict voter category. Show the code used to fit the model, but do **not** display the model output.

```
library(nnet)
model <- multinom(voter_category ~ age_centered + race + gender + income_cat + educ, data = v
```

```
# weights:  36 (22 variable)
initial  value 6411.501317
iter  10 value 5869.948482
iter  20 value 5728.474131
final  value 5693.312867
converged
```

- *Should party identification be added to the model?*
- #Hint: Use an anova test to make the determination

```
model2 <- multinom(voter_category ~ age_centered + race + gender + income_cat + educ + pol_id
                   data = voter_data)
```

```
# weights:  45 (28 variable)
initial  value 6411.501317
iter  10 value 5818.012349
iter  20 value 5709.034111
iter  30 value 5621.228937
final  value 5616.390878
converged
```

```
# Perform an ANOVA test to compare the two models
anova(model, model2)
```

```
                                                             Model Resid. df
1                  age_centered + race + gender + income_cat + educ     11650
2 age_centered + race + gender + income_cat + educ + pol_ident_new     11644
  Resid. Dev   Test    Df LR stat. Pr(Chi)
1   11386.63           NA       NA      NA
2   11232.78 1 vs 2     6  153.844       0
```

```
> #Enter answer based on your code: Adding party identification significantly improves the mo
```

Use the model you select for the remainder of the assignment.

1.1 LRT

- Run the full model and report overall significance of each of the terms

```r
car::Anova(model2, type="II") %>%
  kable(format = "markdown", digits = 3)
```

	LR Chisq	Df	Pr(>Chisq)
age_centered	638.297	2	0.000
race	52.652	6	0.000
gender	6.028	2	0.049
income_cat	67.721	6	0.000
educ	154.137	4	0.000
pol_ident_new	153.844	6	0.000

1.2 Marginal Effects Political Group - Emmeans

```r
#Get estimated marginal means from the model

#using
multinomial_analysis <- emmeans(model2, ~ pol_ident_new|voter_category)


coefs = contrast(regrid(multinomial_analysis, "log"),"trt.vs.ctrl1",  by="pol_ident_new")
# you can add a parameter to the above command, ref = newbaseline, if you want to change bas

update(coefs, by = "contrast") %>%
  kable(format = "markdown", digits = 3)
```

contrast	pol_ident_new	estimate	SE	df	t.ratio	p.value
sporadic - (rarely/never)	Dem	0.961	0.070	28	13.722	0.000
always - (rarely/never)	Dem	0.480	0.074	28	6.498	0.000
sporadic - (rarely/never)	Indep	0.591	0.077	28	7.643	0.000
always - (rarely/never)	Indep	-0.049	0.084	28	-0.590	0.900
sporadic - (rarely/never)	Other	0.078	0.087	28	0.902	0.747
always - (rarely/never)	Other	-0.835	0.110	28	-7.577	0.000
sporadic - (rarely/never)	Rep	0.883	0.084	28	10.469	0.000
always - (rarely/never)	Rep	0.327	0.089	28	3.672	0.004

contrast	pol_ident_new	estimate	SE	df	t.ratio	p.value

```
# Pairwise comparisons
contrast(coefs, "revpairwise", by = "contrast") %>%
  kable(format = "markdown", digits = 3)
```

contrast1	contrast	estimate	SE	df	t.ratio	p.value
Indep - Dem	sporadic - (rarely/never)	-0.370	0.094	28	-3.933	0.003
Other - Dem	sporadic - (rarely/never)	-0.883	0.103	28	-8.578	0.000
Other - Indep	sporadic - (rarely/never)	-0.513	0.107	28	-4.807	0.000
Rep - Dem	sporadic - (rarely/never)	-0.078	0.099	28	-0.787	0.860
Rep - Indep	sporadic - (rarely/never)	0.292	0.099	28	2.965	0.029
Rep - Other	sporadic - (rarely/never)	0.805	0.109	28	7.404	0.000
Indep - Dem	always - (rarely/never)	-0.529	0.101	28	-5.255	0.000
Other - Dem	always - (rarely/never)	-1.315	0.125	28	-10.508	0.000
Other - Indep	always - (rarely/never)	-0.786	0.129	28	-6.072	0.000
Rep - Dem	always - (rarely/never)	-0.153	0.104	28	-1.470	0.468
Rep - Indep	always - (rarely/never)	0.376	0.104	28	3.605	0.006
Rep - Other	always - (rarely/never)	1.162	0.130	28	8.969	0.000

1.3 Marginal Effects of Education - Emmeans

```
#Get estimated marginal means from the model

#using
multinomial_analysis <- emmeans(model2, ~ educ|voter_category)


coefs = contrast(regrid(multinomial_analysis, "log"),"trt.vs.ctrl1",  by="educ")
# you can add a parameter to the above command, ref = newbaseline, if you want to change bas

update(coefs, by = "contrast") %>%
  kable(format = "markdown", digits = 3)
```

contrast	educ	estimate	SE	df	t.ratio	p.value
sporadic - (rarely/never)	College	0.986	0.076	28	12.904	0.000

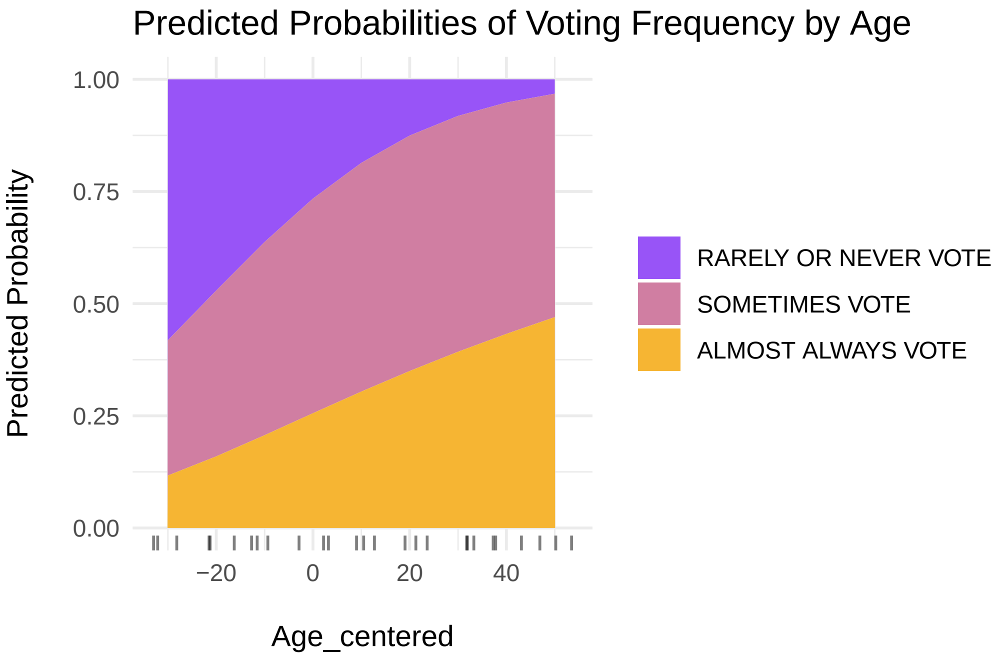
contrast	educ	estimate	SE	df	t.ratio	p.value
always - (rarely/never)	College	0.477	0.080	28	5.960	0.000
sporadic - (rarely/never)	High school or less	0.187	0.069	28	2.705	0.031
always - (rarely/never)	High school or less	-0.711	0.080	28	-8.883	0.000
sporadic - (rarely/never)	Some college	0.707	0.074	28	9.512	0.000
always - (rarely/never)	Some college	0.167	0.079	28	2.114	0.112

```r
# Pairwise comparisons
contrast(coefs, "revpairwise", by = "contrast") %>%
  kable(format = "markdown", digits = 3)
```

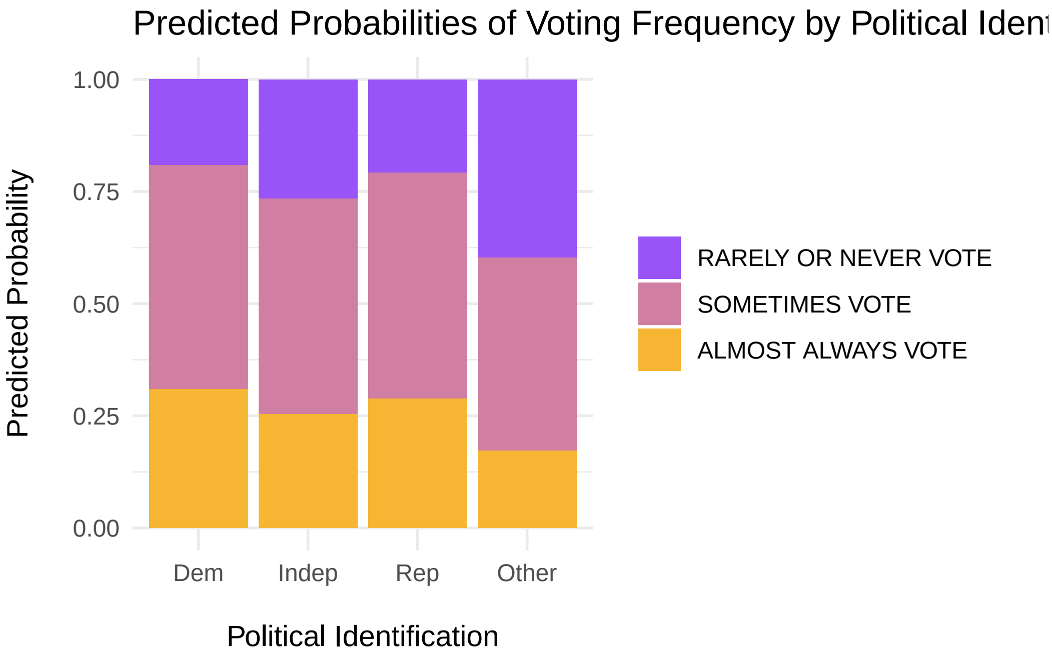
contrast1	contrast	estimate	SE	df	t.ratio	p.value
High school or less - College	sporadic - (rarely/never)	-0.799	0.095	28	-8.416	0.000
Some college - College	sporadic - (rarely/never)	-0.278	0.092	28	-3.030	0.014
Some college - High school or less	sporadic - (rarely/never)	0.520	0.088	28	5.920	0.000
High school or less - College	always - (rarely/never)	-1.188	0.104	28	-11.394	0.000
Some college - College	always - (rarely/never)	-0.310	0.097	28	-3.207	0.009
Some college - High school or less	always - (rarely/never)	0.878	0.098	28	8.995	0.000

- Next, plot the predicted probabilities of voter category as a function of Age and Party ID

```r
ggemmeans(model2, terms = c("age_centered")) %>%
  ggplot(aes(x = x, y = predicted, fill = response.level)) +
  geom_area() +
  geom_rug(sides = "b", position = "jitter", alpha = .5) +
  labs(x = "\nAge_centered", y = "Predicted Probability\n",
       title = "Predicted Probabilities of Voting Frequency by Age") +
  scale_fill_manual(
    name = NULL,
    values = c("always" = "#F6B533", "sporadic" = "#D07EA2", "rarely/never" = "#9854F7"),
    labels = c("RARELY OR NEVER VOTE    ", "SOMETIMES VOTE    ", "ALMOST ALWAYS VOTE    "),
    breaks = c("rarely/never", "sporadic", "always")
  ) +
  theme_minimal()
```

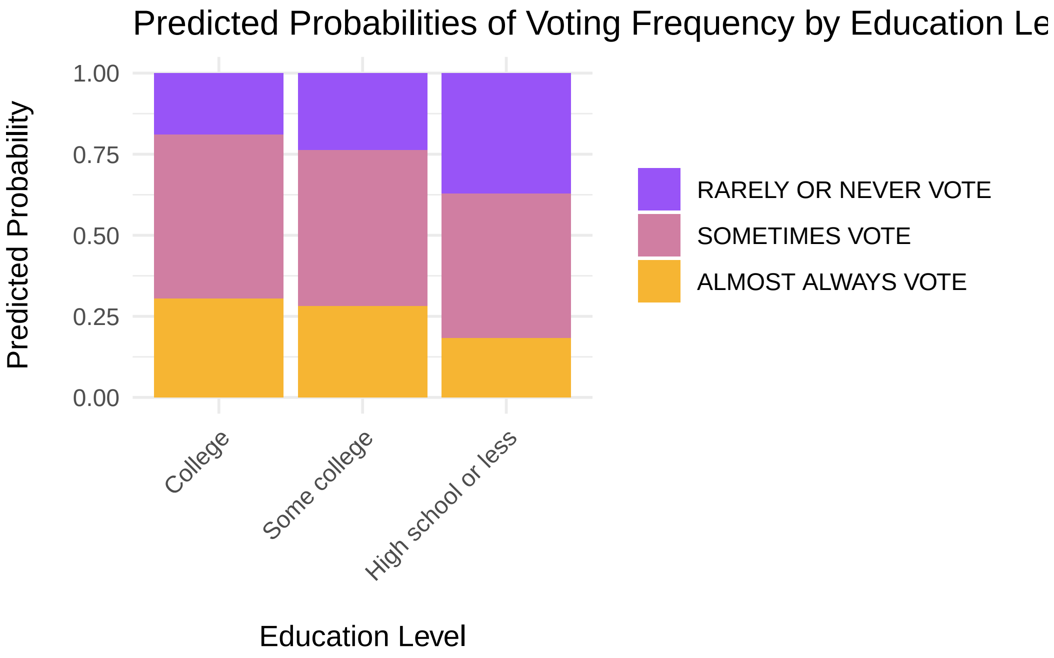


```
ggemmeans(model2, terms = c("pol_ident_new")) %>%
  ggplot(aes(x = x, y = predicted, fill = response.level)) +
  geom_col(position = "stack") +
  labs(x = "\nPolitical Identification", y = "Predicted Probability\n",
       title = "Predicted Probabilities of Voting Frequency by Political Identification") +
  scale_fill_manual(
    name = NULL,
    values = c("always" = "#F6B533", "sporadic" = "#D07EA2", "rarely/never" = "#9854F7"),
    labels = c("RARELY OR NEVER VOTE    ", "SOMETIMES VOTE    ", "ALMOST ALWAYS VOTE    "),
    breaks = c("rarely/never", "sporadic", "always")
  ) +
  theme_minimal()
```



Plot predicted probabilities as a function of education and voting frequency.

```
ggemmeans(model2, terms = c("educ")) %>%
  ggplot(aes(x = x, y = predicted, fill = response.level)) +
  geom_col(position = "stack") +
  labs(x = "\nEducation Level", y = "Predicted Probability\n",
       title = "Predicted Probabilities of Voting Frequency by Education Level") +
  scale_fill_manual(
    name = NULL,
    values = c("always" = "#F6B533", "sporadic" = "#D07EA2", "rarely/never" = "#9854F7"),
    labels = c("RARELY OR NEVER VOTE    ", "SOMETIMES VOTE    ", "ALMOST ALWAYS VOTE    "),
    breaks = c("rarely/never", "sporadic", "always")
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
## Write-up
```

Differences between political groups and voting behavior - Emmeans

```r
multi_an <- emmeans(model2, ~ pol_ident_new|voter_category)

coefs = contrast(regrid(multi_an, "log"),"trt.vs.ctrl1",  by="pol_ident_new")

update(coefs, by = "contrast") %>%
  kable(format = "markdown", digits = 3)
```

contrast	pol_ident_new	estimate	SE	df	t.ratio	p.value
sporadic - (rarely/never)	Dem	0.961	0.070	28	13.722	0.000
always - (rarely/never)	Dem	0.480	0.074	28	6.498	0.000
sporadic - (rarely/never)	Indep	0.591	0.077	28	7.643	0.000
always - (rarely/never)	Indep	-0.049	0.084	28	-0.590	0.900
sporadic - (rarely/never)	Other	0.078	0.087	28	0.902	0.747
always - (rarely/never)	Other	-0.835	0.110	28	-7.577	0.000
sporadic - (rarely/never)	Rep	0.883	0.084	28	10.469	0.000
always - (rarely/never)	Rep	0.327	0.089	28	3.672	0.004

```
# get difference between yes-no and fair-excellent
contrast(coefs, "revpairwise", by = "contrast") %>%
  kable(format = "markdown", digits = 3)
```

contrast1	contrast	estimate	SE	df	t.ratio	p.value
Indep - Dem	sporadic - (rarely/never)	-0.370	0.094	28	-3.933	0.003
Other - Dem	sporadic - (rarely/never)	-0.883	0.103	28	-8.578	0.000
Other - Indep	sporadic - (rarely/never)	-0.513	0.107	28	-4.807	0.000
Rep - Dem	sporadic - (rarely/never)	-0.078	0.099	28	-0.787	0.860
Rep - Indep	sporadic - (rarely/never)	0.292	0.099	28	2.965	0.029
Rep - Other	sporadic - (rarely/never)	0.805	0.109	28	7.404	0.000
Indep - Dem	always - (rarely/never)	-0.529	0.101	28	-5.255	0.000
Other - Dem	always - (rarely/never)	-1.315	0.125	28	-10.508	0.000
Other - Indep	always - (rarely/never)	-0.786	0.129	28	-6.072	0.000
Rep - Dem	always - (rarely/never)	-0.153	0.104	28	-1.470	0.468
Rep - Indep	always - (rarely/never)	0.376	0.104	28	3.605	0.006
Rep - Other	always - (rarely/never)	1.162	0.130	28	8.969	0.000

Differences between education level and voting behavior - Emmeans

Last part of the assignment: Interpret the results from running the following code for your model

```
multi_an <- emmeans(model2, ~ educ|voter_category)

coefs = contrast(regrid(multi_an, "log"),"trt.vs.ctrl1",  by="educ")

update(coefs, by = "contrast") %>%
  kable(format = "markdown", digits = 3)
```

contrast	educ	estimate	SE	df	t.ratio	p.value
sporadic - (rarely/never)	College	0.986	0.076	28	12.904	0.000
always - (rarely/never)	College	0.477	0.080	28	5.960	0.000
sporadic - (rarely/never)	High school or less	0.187	0.069	28	2.705	0.031
always - (rarely/never)	High school or less	-0.711	0.080	28	-8.883	0.000
sporadic - (rarely/never)	Some college	0.707	0.074	28	9.512	0.000
always - (rarely/never)	Some college	0.167	0.079	28	2.114	0.112

```
# get difference between yes-no and fair-excellent
contrast(coefs, "revpairwise", by = "contrast") %>%
  kable(format = "markdown", digits = 3)
```

contrast1	contrast	estimate	SE	df	t.ratio	p.value
High school or less - College	sporadic - (rarely/never)	-0.799	0.095	28	-8.416	0.000
Some college - College	sporadic - (rarely/never)	-0.278	0.092	28	-3.030	0.014
Some college - High school or less	sporadic - (rarely/never)	0.520	0.088	28	5.920	0.000
High school or less - College	always - (rarely/never)	-1.188	0.104	28	-11.394	0.000
Some college - College	always - (rarely/never)	-0.310	0.097	28	-3.207	0.009
Some college - High school or less	always - (rarely/never)	0.878	0.098	28	8.995	0.000

Enter your interpretation here: For political party, dem and rep are more likely to vote more frequently than indp and other. Other seems to vote the least frequently overall. Higher education levels are associated with increased odds of being a more frequent voters.