HW 04: Logistic regression

Karen Dong

2023-11-15

Set up

```r
library(tidyverse)
library(tidymodels)
library(knitr)
```

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

Exercises

Exercise 1

The authors chose to only include data from people who were eligible to vote for at least four election cycles because they would be able to analyze their voting behavior across multiple election cycles, rather than just one cycle. We would be able to analyze people's patterns of voting behavior across longer periods of time, instead of attributing voting behavior from a single voting cycle.

Exercise 2

```
voter_data$frequent_voter <- as.integer(voter_data$voter_category == "always")

table(voter_data$frequent_voter)
```

```
   0    1 
4025 1811 
```

```
mean(voter_data$frequent_voter) * 100
```

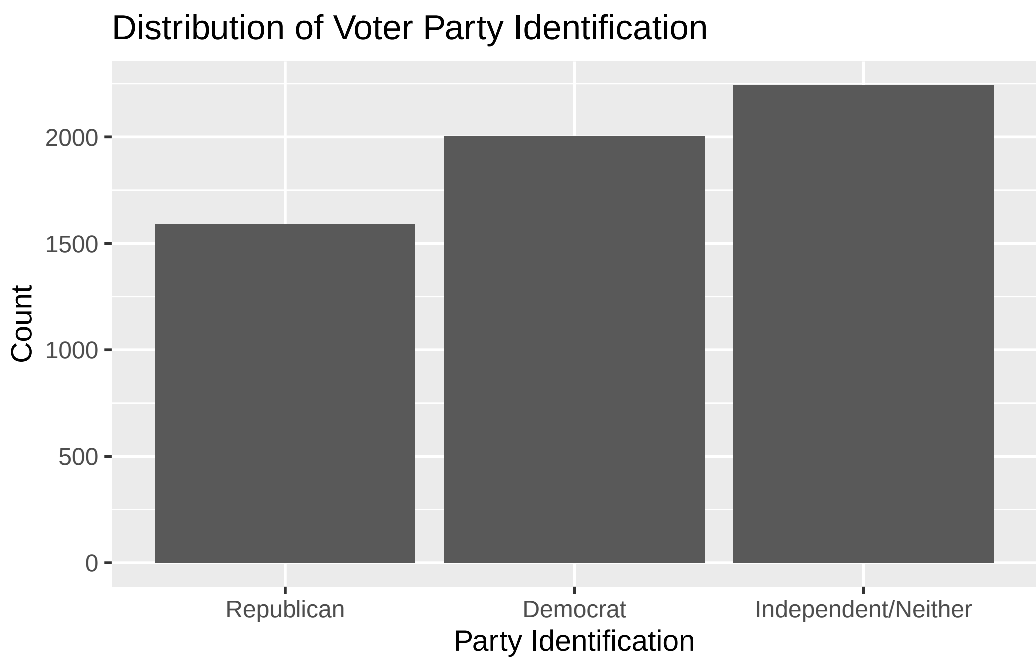
```
[1] 31.03153
```

Approximately 31.032% of respondents in the data said they voted "in all or all-but-one of the elections they were eligible in."

Exercise 3

```
voter_data$party_id <- factor(voter_data$Q30, levels = c(1, 2, 3, 4, 5, -1),
                              labels = c("Republican", "Democrat",
                                         "Independent/Neither",
                                         "Independent/Neither",
                                         "Independent/Neither",
                                         "Independent/Neither"))
```

```
voter_data |>
  ggplot(aes(x = party_id)) +
  geom_bar() +
  labs(title = "Distribution of Voter Party Identification",
       x = "Party Identification", y = "Count")
```



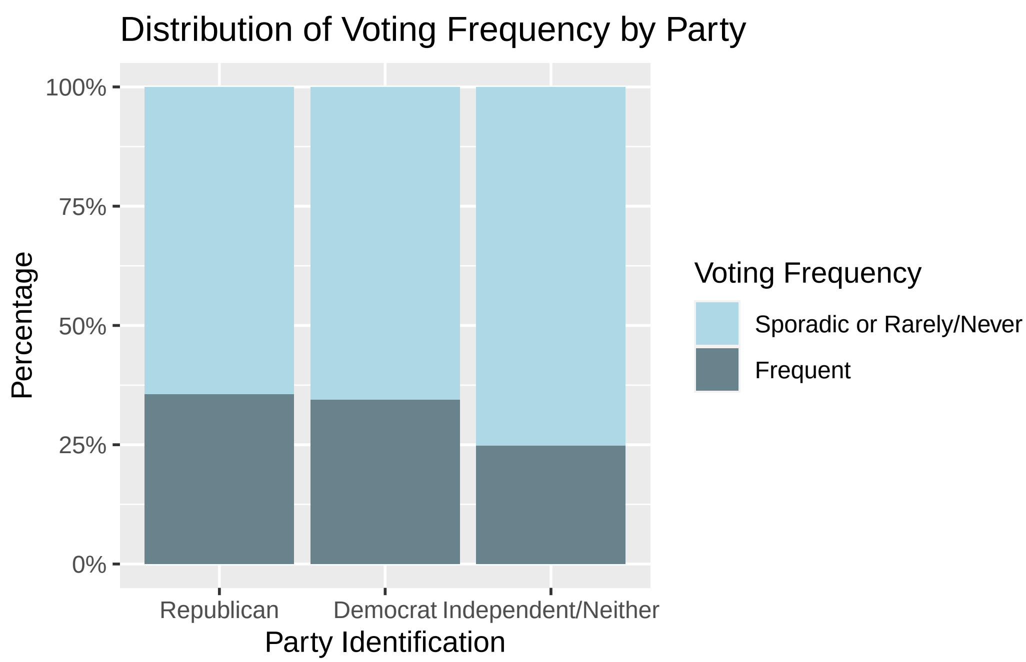
In this data set, the most frequent category of `party_id` is independent/neither.

\pagebreak

Exercise 4

```
voter_data$frequent_voter <- as.factor(voter_data$frequent_voter)

voter_data |>
  ggplot(aes(x = party_id, fill = frequent_voter)) +
  geom_bar(position = "fill", stat = "count") +
  labs(title = "Distribution of Voting Frequency by Party",
       x = "Party Identification",
       y = "Percentage") +
  scale_y_continuous(labels = scales::percent_format(scale = 100)) +
  scale_fill_manual(values = c("0" = "lightblue", "1" = "lightblue4"),
                    name = "Voting Frequency",
                    labels = c("Sporadic or Rarely/Never", "Frequent"))
```



Voters who identify with the Republican party have the highest proportion of frequent voters, voters who identify with Democrat have slightly lower proportion, but the second highest proportion of frequent voters, and those who are Independent/Neither have the lowest proportion of frequent voters.

Exercise 5

```r
set.seed(29)
voter_split <- initial_split(voter_data, prop = 0.75)
voter_train <- training(voter_split)
voter_test <- testing(voter_split)

voter_fit <- logistic_reg() |>
  set_engine("glm") |>
  fit(frequent_voter ~ ppage + educ + race + gender + income_cat,
      data = voter_train, family = "binomial")

voter_fit |>
  tidy() |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-2.090	0.164	-12.761	0.000
ppage	0.028	0.002	13.500	0.000
educHigh school or less	-0.655	0.094	-6.956	0.000
educSome college	-0.113	0.084	-1.341	0.180
raceHispanic	-0.409	0.134	-3.060	0.002
raceOther/Mixed	-0.465	0.171	-2.716	0.007
raceWhite	0.118	0.094	1.253	0.210
genderMale	-0.089	0.069	-1.304	0.192
income_cat$40-75k	0.096	0.102	0.938	0.348
income_cat$75-125k	0.257	0.094	2.733	0.006
income_catLess than $40k	-0.258	0.112	-2.296	0.022

The coefficient of `ppage` is 0.028, which means the predicted odds of a person being a frequent voter increases by approximately 1.028 $\exp\{0.028\}$ times for each year their age increases, holding all other variables constant.

Exercise 6

$H_0 : \beta_{Democrat} = \beta_{Independent} = 0$

H_a: at least one $\beta_{party\_id} \neq 0$

```
voter_fit1 <- logistic_reg() |>
  set_engine("glm") |>
  fit(frequent_voter ~ ppage + educ + race + gender
      + income_cat + party_id,
      data = voter_train, family = "binomial")
```

```
anova(voter_fit$fit, voter_fit1$fit, test = "Chisq") |>
  tidy() |> kable(digits = 3)
```

term	df.residual	residual.deviance	df	deviance	p.value
frequent_voter ~ ppage + educ + race + gender + income_cat	4366	5072.595	NA	NA	NA
frequent_voter ~ ppage + educ + race + gender + income_cat + party_id	4364	5052.151	2	20.444	0

The p-value is 0, which is very small, so we reject the null hypothesis. The data provides sufficient evidence that the coefficient of `party_id` is not equal to 0. Therefore, we should add it to the model.

Exercise 7

```
voter_fit1 |>
  tidy() |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-2.050	0.187	-10.952	0.000
ppage	0.028	0.002	13.094	0.000
educHigh school or less	-0.645	0.095	-6.798	0.000
educSome college	-0.106	0.085	-1.244	0.213
raceHispanic	-0.380	0.135	-2.815	0.005
raceOther/Mixed	-0.411	0.173	-2.380	0.017
raceWhite	0.169	0.100	1.687	0.092
genderMale	-0.061	0.069	-0.881	0.378
income_cat$40-75k	0.106	0.102	1.034	0.301
income_cat$75-125k	0.262	0.094	2.782	0.005
income_catLess than $40k	-0.240	0.113	-2.128	0.033
party_idDemocrat	0.080	0.091	0.883	0.377
party_idIndependent/Neither	-0.277	0.087	-3.165	0.002

Political party does have an effect on the odds of whether a person is a frequent voter. The statistically significant level is when `party_id` is Independent/Neither, since it has a p-value of 0, whereas when `party_id` is Democrat, there is a high p-value of 0.377. When a person is Independent/Neither, the odds of being a frequent voter is 0.758 exp{-0.277} times the odds of a Republican being a frequent voter, holding all else constant.

Exercise 8

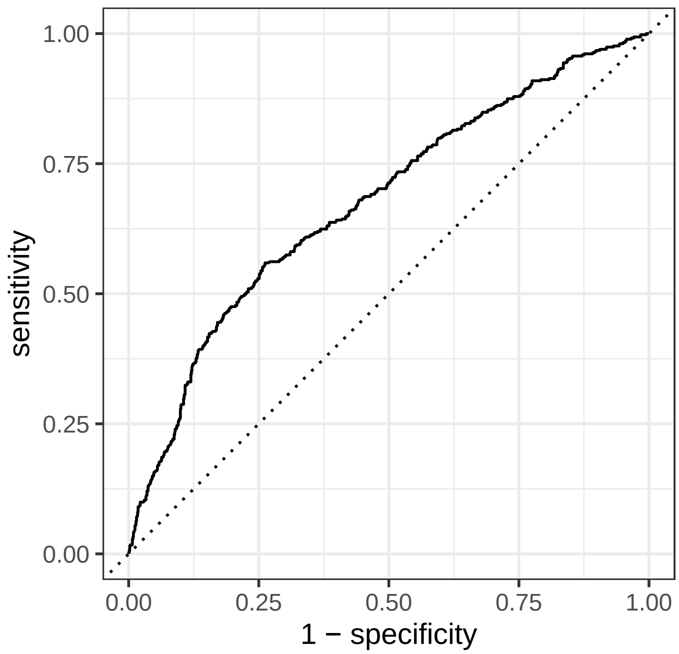
The model I selected is consistent with the statement since the coefficients tell us about how each condition impacts the odds of someone being a frequent voter.

- The coefficient for `income_cat` less than $40k is negative while other `income_cat` categories have a positive coefficient, so lower income has lower odds of a person being a frequent voter.
- The coefficient for `ppage` is positive, which means lower age has lower odds of a person being a frequent voter.
- Additionally, the coefficient for `educ` high school or less education is 0.539 more negative than some college, so lower levels of education has lower odds of a person being a frequent voter.

Exercise 9

```
voter_pred <- predict(voter_fit1, voter_test,
                      type = "prob") |>
  bind_cols(voter_test)
```

```
voter_pred |>
  roc_curve(
    truth = frequent_voter,
    .pred_1,
    event_level = "second") |>
  autoplot()
```




```
voter_pred |>
  roc_auc(
    truth = frequent_voter,
    .pred_1,
    event_level = "second")
```

```
# A tibble: 1 x 3
```

```
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.676
```

The AUC is approximately 0.676, which means this model is a moderate fit since it is above 0.5, but it is not as close to 1 as it is to 0.5.

Exercise 10

```r
cutoff_prob <- 0.25

voter_pred |>
  mutate(voter_predicted =
           as_factor(if_else(.pred_1 >= cutoff_prob, 1, 0))) |>
  conf_mat(truth = frequent_voter, estimate = voter_predicted)
```

```
          Truth
Prediction   0   1
         0 426 103
         1 570 360
```

- **Sensitivity:** 360 / (360 + 103) = 0.778
- **Specificity:** 426 / (570 + 426) = 0.428
- **False negative rate:** 0.222
- **False positive rate:** 0.572