Lab 07: General Social Survey

Logistic regression

Stats is 'Fun' - Dav King, Luke Thomas, Thomas Barker, Harry Liu

11/11/22

Setup

Load packages and data

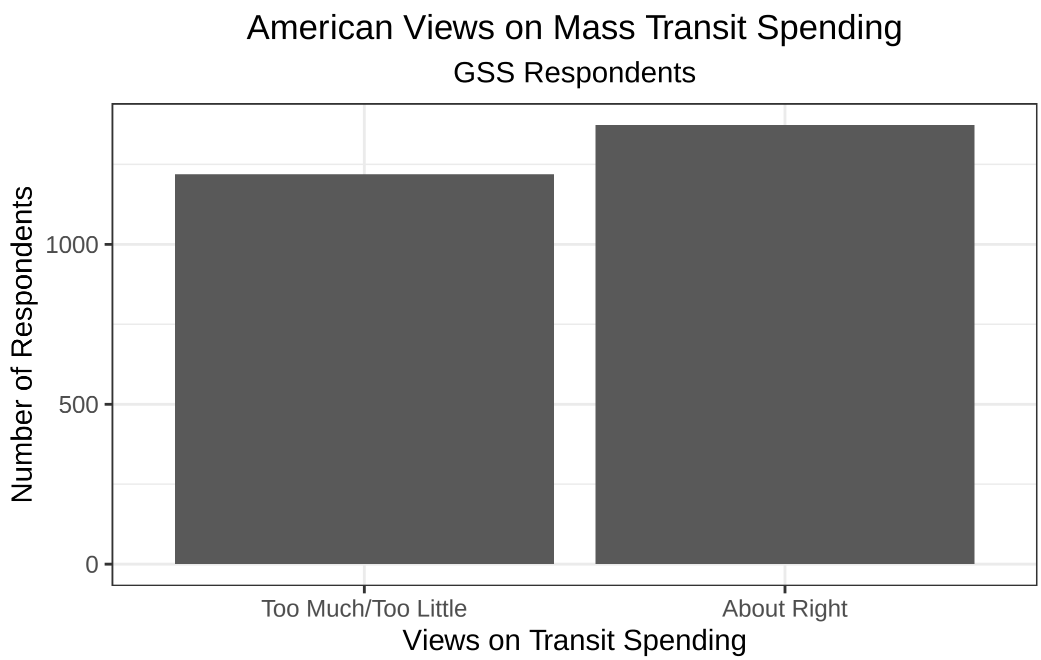
```
# add packages + data

library(tidyverse)
library(tidymodels)
library(knitr)

gss <- read_csv("data/gss2016.csv") |>
  mutate(polviews = case_when(polviews == 'Extrmly conservative' ~
   ↪  'Extremely conservative',
                              polviews == 'Slghtly conservative' ~
 ↪  'Slightly conservative',
                              TRUE ~ polviews))
```

[Select this page for the "Workflow & formatting" in Gradescope.]

Exercises

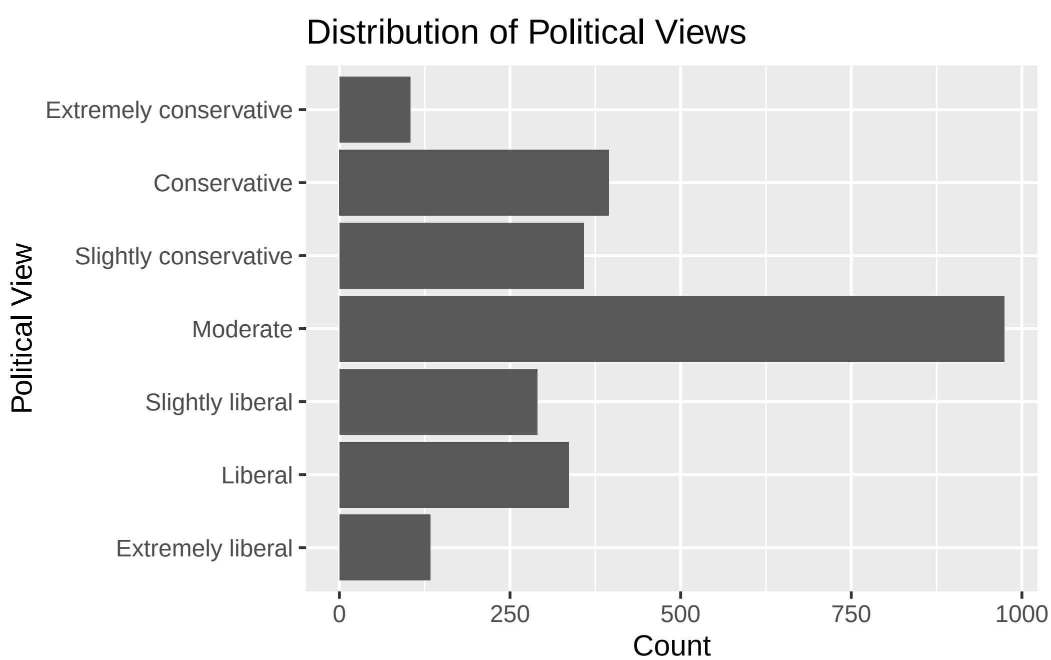
Exercise 1

```
gss <- gss %>%
  mutate(transit = factor(if_else(natmass == "About right", 1, 0)))
ggplot(gss, aes(x = transit)) +
  geom_bar() +
  scale_x_discrete(labels = c("Too Much/Too Little", "About Right")) +
  theme_bw() +
  labs(x = "Views on Transit Spending", y = "Number of Respondents",
       title = "American Views on Mass Transit Spending", subtitle =
        ↪  "GSS Respondents") +
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
   ↪  element_text(hjust = 0.5))
```



Exercise 2

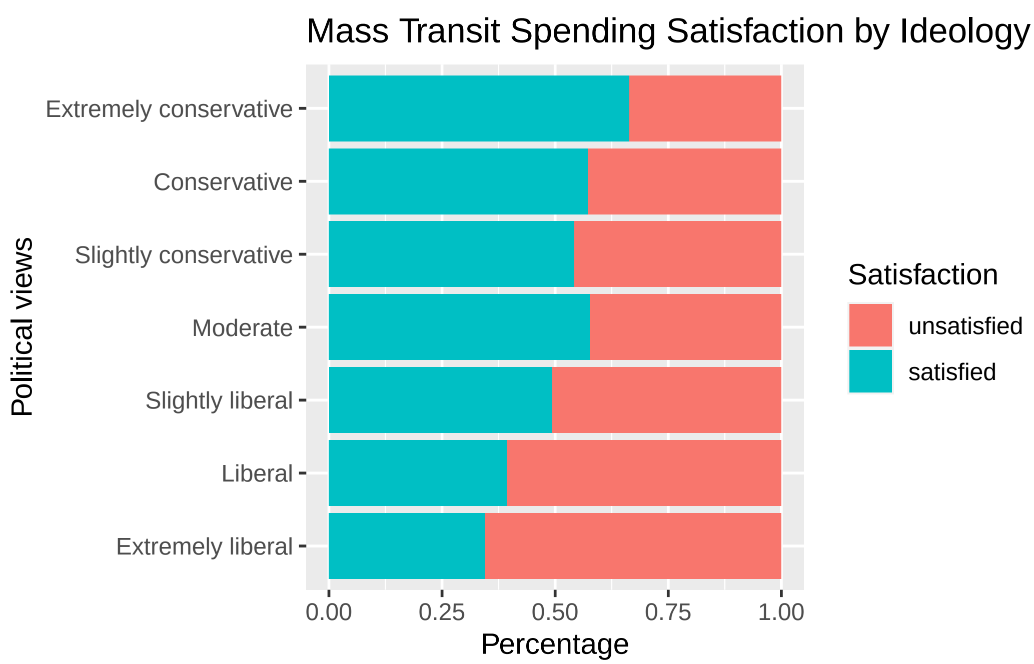
```r
gss <- gss |>
  mutate(polviews = factor(polviews, levels = c("Extremely liberal",
                                                 "Liberal",
                                                 "Slightly liberal",
                                                 'Moderate',
                                                 'Slightly
                                                  ↪  conservative',
                                                 'Conservative',
                                                 'Extremely
                                                  ↪  conservative')))
ggplot(data = gss, aes(y=polviews)) +
  geom_bar() +
  labs(x = 'Count',
       y = 'Political View',
       title = 'Distribution of Political Views')
```



Moderate political view occurs most frequently in this data set.

Exercise 3

```
ggplot(data = gss, aes(fill=transit, x=polviews)) +
    geom_bar(position="fill") +
    labs(title = "Mass Transit Spending Satisfaction by Ideology",
         x = "Political views", y = "Percentage") +
    scale_fill_discrete(name = "Satisfaction", labels = c("unsatisfied",
     ↪  "satisfied")) +
    coord_flip()
```

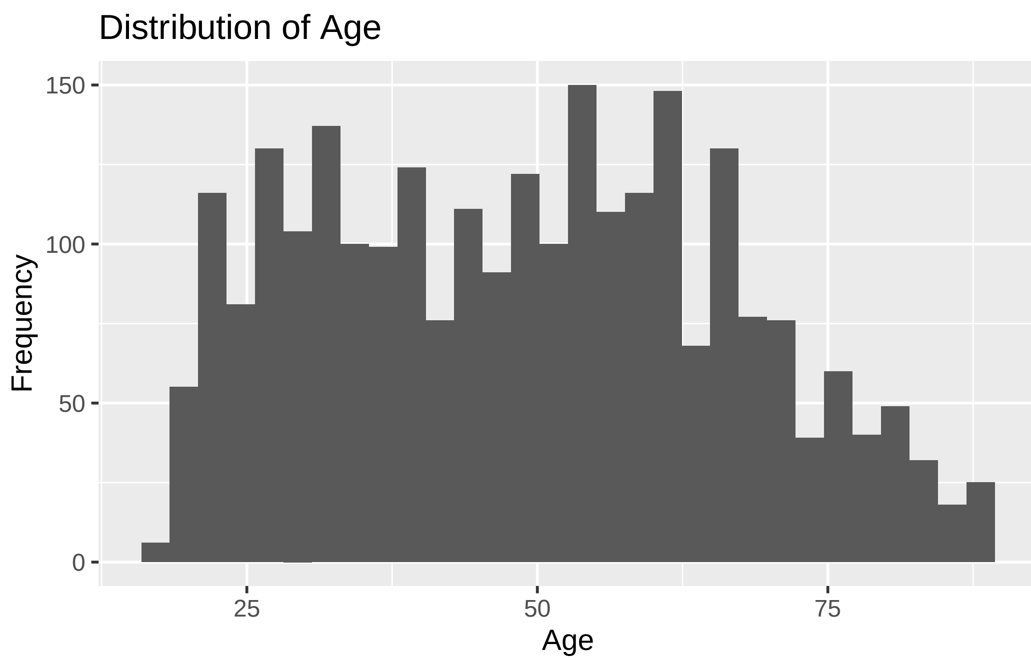


As a person's political view becomes more liberal, the percentage of people that are satisfied with mass transportation spending generally decreases, while the percentage of people that are unsatisfied (either think its too much or too little) with mass transportation spending generally increases.

Exercise 4

```
gss <- gss |>
  mutate(age = if_else(age == "89 or older", "89", age))

gss <- gss |>
  mutate_at('age', as.numeric)
```

```
gss |>
  ggplot(aes(x = age)) +
  geom_histogram(bins = 30) +
  labs(x = "Age",
       y = "Frequency",
       title = "Distribution of Age")
```



Exercise 5

Satisfaction with spending on mass transportation is a binary response variable. Thus, we want a model that will run between satisfaction and dissatisfaction, without trying to make linear regression predictions that are much less meaningful. Using a logistic regression model, we can predict the odds that a randomly selected person is satisfied with spending on mass transit - giving us much more meaningful and nuanced conclusions than we would get by simply classifying people into either “satisfied” or “unsatisfied” without these odds.

Exercise 6

```r
set.seed(6)
gss_split <- initial_split(gss)
gss_train <- training(gss_split)
gss_test <- testing(gss_split)
```

Exercise 7

```
gss_spec <- logistic_reg() |>
  set_engine('glm')

gss_rec1 <- recipe(transit ~ age + sex + sei10 + region, data =
↪   gss_train) |>
  step_center(all_numeric_predictors())

gss_wflow1 <- workflow() |>
  add_model(gss_spec) |>
  add_recipe(gss_rec1)

gss_fit <- gss_wflow1 |>
  fit(gss_train)

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

term	estimate	std.error	statistic	p.value
(Intercept)	0.293	0.118	2.484	0.013
age	-0.006	0.003	-2.362	0.018
sexMale	-0.269	0.093	-2.886	0.004
sei10	-0.008	0.002	-4.149	0.000
regionE. sou. central	-0.127	0.209	-0.605	0.545
regionMiddle atlantic	0.203	0.178	1.139	0.255
regionMountain	-0.042	0.188	-0.221	0.825
regionNew england	-0.615	0.214	-2.879	0.004
regionPacific	-0.362	0.164	-2.202	0.028
regionSouth atlantic	0.122	0.151	0.807	0.420
regionW. nor. central	0.088	0.213	0.414	0.679
regionW. sou. central	0.190	0.185	1.026	0.305

Exercise 8

For a person who is female, lives in the east north central region of the country, is the mean age of ~48.6 years, and has the mean SEI10 score of ~46.02, the odds of being satisfied with spending on mass transportation are expected to be 1.340 (exp(0.293)).

For each additional year in age, the odds of being satisfied with spending on mass transportation are expected to multiply by a factor of 0.994 (exp(-0.006)), holding sex, region, and SEI10 constant.

Exercise 9

```r
views_spec <- logistic_reg() |>
  set_engine('glm')

views_rec <- recipe(transit ~ age + sex + sei10 + region + polviews,
↪   data = gss_train) |>
  step_center(all_numeric_predictors())

views_wflow <- workflow() |>
  add_model(views_spec) |>
  add_recipe(views_rec)

views_fit <- views_wflow |>
  fit(gss_train)

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

term	estimate	std.error	statistic	p.value
(Intercept)	-0.382	0.237	-1.615	0.106
age	-0.008	0.003	-3.091	0.002
sexMale	-0.298	0.095	-3.143	0.002
sei10	-0.007	0.002	-3.417	0.001
regionE. sou. central	-0.149	0.212	-0.705	0.481
regionMiddle atlantic	0.233	0.181	1.285	0.199
regionMountain	0.011	0.191	0.057	0.954
regionNew england	-0.510	0.218	-2.341	0.019
regionPacific	-0.340	0.167	-2.039	0.041
regionSouth atlantic	0.125	0.153	0.817	0.414
regionW. nor. central	0.064	0.216	0.295	0.768
regionW. sou. central	0.138	0.187	0.737	0.461
polviewsLiberal	0.081	0.246	0.330	0.741
polviewsSlightly liberal	0.464	0.251	1.854	0.064
polviewsModerate	0.841	0.222	3.785	0.000
polviewsSlightly conservative	0.817	0.243	3.369	0.001
polviewsConservative	0.927	0.243	3.813	0.000
polviewsExtremely conservative	1.271	0.326	3.900	0.000

Exercise 10

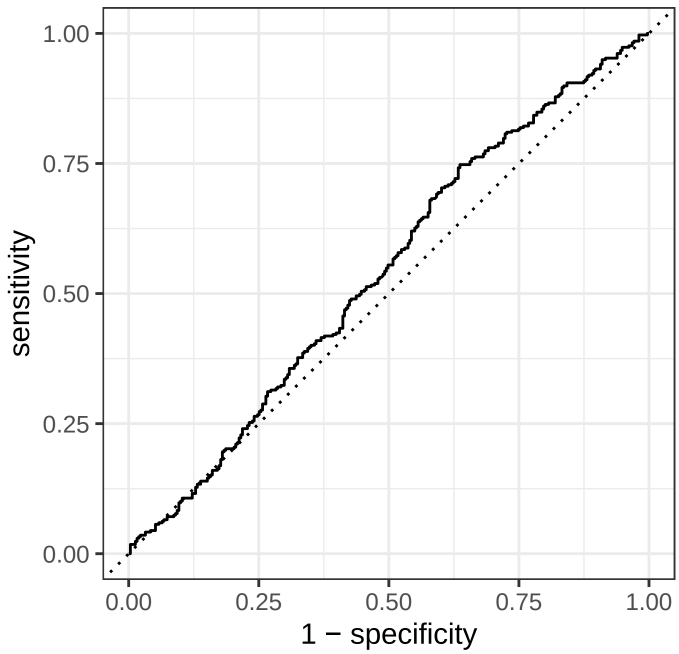
```r
gss_pred <- predict(gss_fit, gss_test, type = "prob") |>
  bind_cols(gss_test)
gss_pred
```

```
# A tibble: 648 x 9
   .pred_0 .pred_1 natmass       age sex    sei10 region         polviews transit
     <dbl>   <dbl> <chr>       <dbl> <chr>  <dbl> <chr>          <fct>    <fct>
 1   0.577   0.423 Too little     55 Female  39.7 New england    Slightl~ 0
 2   0.707   0.293 Too little     50 Male    80.7 New england    Slightl~ 0
 3   0.296   0.704 Too much       23 Female  20.1 Middle atlan~ Slightl~ 0
 4   0.371   0.629 About right    86 Female  13.2 Middle atlan~ Slightl~ 1
 5   0.622   0.378 About right    43 Male    39.2 New england    Liberal  1
 6   0.298   0.702 About right    23 Female  21.6 Middle atlan~ Slightl~ 1
 7   0.289   0.711 About right    25 Female  14.8 Middle atlan~ Liberal  1
 8   0.552   0.448 Too little     71 Male    82.5 Middle atlan~ Liberal  0
 9   0.391   0.609 About right    22 Male    39.9 Middle atlan~ Moderate 1
10   0.369   0.631 About right    32 Male    20.7 Middle atlan~ Liberal  1
# ... with 638 more rows
```

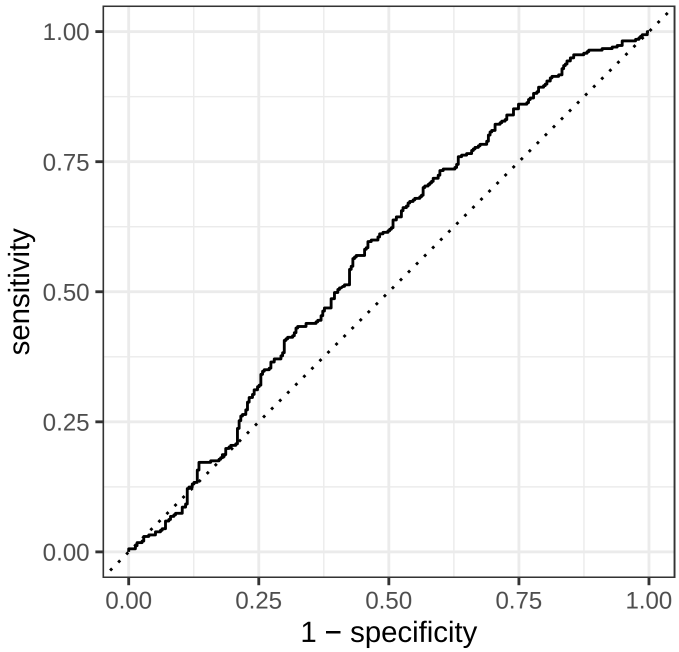
```r
views_pred <- predict(views_fit, gss_test, type = "prob") |>
  bind_cols(gss_test)
views_pred
```

```
# A tibble: 648 x 9
   .pred_0 .pred_1 natmass       age sex    sei10 region         polviews transit
     <dbl>   <dbl> <chr>       <dbl> <chr>  <dbl> <chr>          <fct>    <fct>
 1   0.608   0.392 Too little     55 Female  39.7 New england    Slightl~ 0
 2   0.726   0.274 Too little     50 Male    80.7 New england    Slightl~ 0
 3   0.257   0.743 Too much       23 Female  20.1 Middle atlan~ Slightl~ 0
 4   0.360   0.640 About right    86 Female  13.2 Middle atlan~ Slightl~ 1
 5   0.734   0.266 About right    43 Male    39.2 New england    Liberal  1
 6   0.332   0.668 About right    23 Female  21.6 Middle atlan~ Slightl~ 1
 7   0.414   0.586 About right    25 Female  14.8 Middle atlan~ Liberal  1
 8   0.691   0.309 Too little     71 Male    82.5 Middle atlan~ Liberal  0
 9   0.340   0.660 About right    22 Male    39.9 Middle atlan~ Moderate 1
10   0.513   0.487 About right    32 Male    20.7 Middle atlan~ Liberal  1
# ... with 638 more rows
```

```
gss_pred |>
  roc_curve(
    truth = transit,
    .pred_1,
    event_level = "second"
  ) |>
  autoplot()
```




```
views_pred |>
  roc_curve(
    truth = transit,
    .pred_1,
    event_level = "second"
  ) |>
  autoplot()
```




```
gss_pred |>
  roc_auc(
    truth = transit,
    .pred_1,
    event_level = "second"
  )
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.541
```

```
views_pred |>
  roc_auc(
    truth = transit,
    .pred_1,
    event_level = "second"
  )
```

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

```
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.573
```

The model that includes political views as a predictor is a better fit, because its AUC is closer to 1 than the model that doesn't include political views as a predictor ($.573 > .541$).

Exercise 11

```
views_pred |>
  roc_curve(
    truth = transit,
    .pred_1,
    event_level = "second"
  )
```

```
# A tibble: 648 x 3
   .threshold specificity sensitivity
        <dbl>       <dbl>       <dbl>
 1   -Inf          0             1
 2      0.180      0             1
 3      0.182      0.00322       1
 4      0.219      0.00322       0.997
 5      0.230      0.00322       0.994
 6      0.233      0.00643       0.994
 7      0.235      0.00965       0.994
 8      0.238      0.0129        0.994
 9      0.243      0.0129        0.991
10      0.266      0.0161        0.991
# ... with 638 more rows
```

We would use a cutoff probability of 55.70% to classify observations in "satisfied with mass transportation spending" versus "not satisfied".

Reasoning: Under such a circumstance when the political organization wants to send political mailings only to the adults who are currently satisfied with current spending on mass transportation while avoiding to send the mailing to those who are not satisfied, the best scenario is to try to seek a balanced solution where both sensitivity and specificity is taken into consideration - that been said, we want to find a cutoff where both sensitivity and specificity could be considerably high (we don't want high false negative as we don't want to miss sending emails to people that are in reality satisfied, but we also don't want high false positive as we don't want to send political mailing to the wrong people). Placing sensitivity and specificity at the same importance, we decide to choose a cutoff probability that have the same sensitively and specificity. And by checking the table, we found that at a cutoff probability of 55.70%, the sensitivity and specificity are the closest (sensitivity ~56.59%, specificity ~56.68).

Exercise 12

```
cutoff_prob <- 0.5570
views_pred |>
  mutate(views_predicted = as_factor(if_else(.pred_1 >= cutoff_prob, 1,
   ↪  0))) |>
  conf_mat(truth = transit, estimate = views_predicted)
```

```
          Truth
Prediction   0   1
         0 176 147
         1 135 190
```

```
sensitivity <- 190 / (147 + 190)
sensitivity
```

```
[1] 0.5637982
```

```
specificity <- 176 / (135 + 176)
specificity
```

```
[1] 0.5659164
```

```
false_nagative <- 147 / (147 + 190)
false_nagative
```

```
[1] 0.4362018
```

```
false_positive <- 135 / (135 + 176)
false_positive
```

```
[1] 0.4340836
```

- Sensitivity = 190 / (147 + 190) = 56.38%
- Specificity = 176 / (135 + 176) = 56.59%
- False negative rate = 147 / (147 + 190) = 43.62%
- False positive rate = 135 / (135 + 176) = 43.41%