HW 06: Logistic regression

Binomial responses and overdispersion

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(broom)
library(knitr)

# add other packages as needed
```

Data: Supporting railroads in the 1870s

The data set RR_Data_Hale.csv contains information on support for referendums related to railroad subsidies for 11 communities in Hale County, Alabama in the 1870s. The data were originally collected from the US Census by historian Michael Fitzgerald and analyzed as part of a thesis project by a student at St. Olaf College. The variables in the data are

- `pctBlack`: percentage of Black residents in the county
- `distance`: distance the proposed railroad is from the community (in miles)
- `YesVotes`: number of "yes" votes in favor of the proposed railroad line
- `NumVotes`: number of votes cast in the election

```
rr <- read_csv("../data/RR_Data_Hale.csv")
```

```
Rows: 12 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): County
dbl (7): popBlack, popWhite, popTotal, pctBlack, distance, YesVotes, NumVotes

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.
```

```
rr <- rr |>
  mutate(pctYes = YesVotes/NumVotes,
         emp_logit = log(pctYes / (1 - pctYes)),
         inFavor = if_else(pctYes > 0.5, "Yes", "No"))
```

Part 1

```
rr_model <- glm(cbind(YesVotes, NumVotes - YesVotes) ~ distance + pctBlack,
                data = rr, family = binomial)

tidy(rr_model, conf.int = TRUE) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.222	0.297	14.217	0.000	3.644	4.809
distance	-0.292	0.013	-22.270	0.000	-0.318	-0.267
pctBlack	-0.013	0.004	-3.394	0.001	-0.021	-0.006

Alternate model syntax

```
rr_model_alt <- glm(pctYes ~ distance + pctBlack, data = rr,
                    family = binomial, weight = NumVotes)

tidy(rr_model_alt, conf.int = TRUE) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.222	0.297	14.217	0.000	3.644	4.809
distance	-0.292	0.013	-22.270	0.000	-0.318	-0.267
pctBlack	-0.013	0.004	-3.394	0.001	-0.021	-0.006

Exercise 1

Interpret the coefficient of distance in the context of the data.

Exercise 2

Use a likelihood ratio test or drop-in-deviance test to determine if the interaction between `distance` and `pctBlack` should be added to the model.

```
# code to test interaction
```

Exercise 3

Use the model selected in the previous exercise. Interpret the effect of the demographics for a community that is…

- Right on the proposed railroad (distance = 0)
- 15 miles away from the proposed railroad (distance = 15)

Exercise 4

Conduct the appropriate test to assess if the model selected in Exercise 2 is good fit for the data.

```
# code for goodness-of-fit test
```

Part 2

> **Exercise 5**
>
> Fit the quasibinomial model. How did the coefficients change from the original model? How did the standard errors change?

```
# code to fit quasibinomial model
```

> **Exercise 6**
>
> Based on the results from Exercise 5, what might be your next step in the analysis? If possible, conduct that step below.

```
# code for next step
```