HW 06: Logistic regression

Binomial responses and overdispersion

```r
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 becom
```

```r
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

```r
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.
```

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

Part 1

```r
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

```r
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.

For each mile increase in the distance the community is from the proposed railroad, the odds that the community voted Yes for the railroad changes by a multiplicative factor of 0.7467685 i.e., it declines by 25.3% after adjusting for racial composition.

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.

drop in deviance

$$H_0 : distance, percentblack$$

$$H_1 : distance, percentblack, dist \times percBlack$$

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

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

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	7.551	0.638	11.828	0	6.300	8.809
distance	-0.614	0.057	-10.701	0	-0.727	-0.502
pctBlack	-0.065	0.009	-7.057	0	-0.083	-0.046
distance:pctBlack	0.005	0.001	5.974	0	0.004	0.007

```r
anova(rr_model_alt, rr_model_int,test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: pctYes ~ distance + pctBlack
Model 2: pctYes ~ distance + pctBlack + distance * pctBlack
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         8     307.22
2         7     274.23  1   32.984 9.294e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We perform a drop in deviance, and our p-value is very small so we reject the null in favor of the model with the interaction term. We find that there is significant evidence supporting the full model.

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)

For a community that is right on the railroad, the odds that the community voted Yes for the railroad changes by a multiplicative factor of 0.9370675 i.e. it decreases by about 6.3%.

For a community 15 miles away from the proposed railroad,

Exercise 4

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

```r
1-pchisq(274.23, 7)
```

```
[1] 0
```

Part 2

Exercise 5

Fit the quasibinomial model. How did the coefficients change from the original model? How did the standard errors change?

```r
rr_model_qb <- glm(pctYes ~ distance + pctBlack + distance*pctBlack, data = rr,
                    family = quasibinomial, weight = NumVotes)

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

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	7.551	4.585	1.647	0.144	-2.308	17.913
distance	-0.614	0.412	-1.490	0.180	-1.529	0.250
pctBlack	-0.065	0.066	-0.982	0.359	-0.198	0.090
distance:pctBlack	0.005	0.006	0.832	0.433	-0.009	0.019

Exercise 6

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

```r
# code for next step
```