AE 9: Odds

Notes

Packages

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

heart_disease <- read_csv(here::here("data/framingham.csv")) %>%
  select(totChol, TenYearCHD) %>%
  drop_na() %>%
  mutate(high_risk = as.factor(TenYearCHD)) %>%
  select(totChol, high_risk)
```

Linear regression vs. logistic regression

State whether a linear regression model or logistic regression model is more appropriate for each scenario:

1. Use age and education to predict if a randomly selected person will vote in the next election.
2. Use budget and run time (in minutes) to predict a movie's total revenue.
3. Use age and sex to calculate the probability a randomly selected adult will visit Duke Health in the next year.

Heart disease

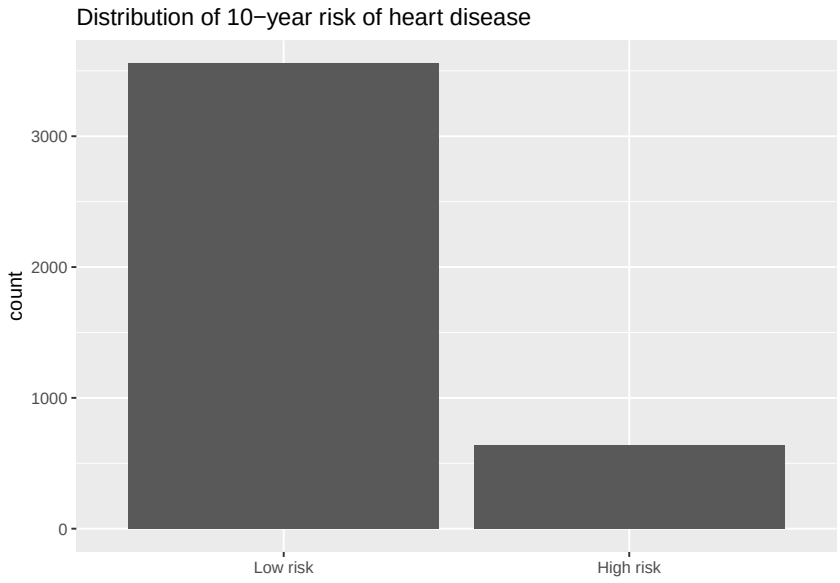
Data: Framingham study

This data set is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. We want to use the total cholesterol to predict if a randomly selected adult is high risk for heart disease in the next 10 years.

- `high_risk`:
  - 1: High risk of having heart disease in next 10 years
  - 0: Not high risk of having heart disease in next 10 years
- `totChol`: total cholesterol (mg/dL)

Outcome: `high_risk`

```
ggplot(data = heart_disease, aes(x = high_risk)) +
  geom_bar() +
  scale_x_discrete(labels = c("1" = "High risk", "0" = "Low risk")) +
  labs(
    title = "Distribution of 10-year risk of heart disease",
    x = NULL)
```



```
heart_disease %>%
  count(high_risk)
```

```
# A tibble: 2 x 2
  high_risk     n
  <fct>     <int>
1 0          3555
2 1           635
```

Calculating probability and odds

1. What is the probability a randomly selected person in the study is **not** high risk for heart disease?

```
heart_disease %>%
  count(high_risk) %>%
  mutate(p = n / sum(n))
```

```
# A tibble: 2 x 3
  high_risk     n     p
  <fct>     <int> <dbl>
1 0          3555 0.848
2 1           635 0.152
```

```
p_not <- heart_disease %>%
  count(high_risk) %>%
  mutate(p = n / sum(n)) %>%
  filter(high_risk == 0) %>%
  pull(p)
```

The probability that a randomly selected person in the study is not high risk for heart disease is 0.848.

2. What are the **odds** a randomly selected person in the study is **not** high risk for heart disease?

```
odds_not <- p_not / (1 - p_not)
odds_not
```

```
[1] 5.598425
```

The odds that a randomly selected person in the study is not high risk for heart disease is 5.598.

Logistic regression model

Fit a logistic regression model to understand the relationship between total cholesterol and risk for heart disease.

Let π be the probability an adult is high risk. The statistical model is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \times TotChol_i$$

```
heart_disease_fit <- logistic_reg() %>%
  set_engine("glm") %>%
  fit(high_risk ~ totChol, data = heart_disease, family = "binomial")

tidy(heart_disease_fit) %>% kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-2.894	0.230	-12.607	0
totChol	0.005	0.001	5.268	0

3. Write the regression equation. Round to 3 digits.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -2.894 + 0.005 \times TotChol_i$$

Calculating log-odds, odds and probabilities

Based on the model, if a randomly selected person has a total cholesterol of 250 mg/dL,

4. What are the log-odds they are high risk for heart disease?

```
new_person <- tibble(totChol = 250)
log_odds <- predict(heart_disease_fit, new_data = new_person, type = "raw")
log_odds
```

```
        1
-1.674053
```

5. What are the odds they are high risk for heart disease?

```
odds <- exp(log_odds)
odds
```

```
        1
0.1874856
```

6. What is the probability they are high risk for heart disease? *Use the odds to calculate your answer.*

```
# using odds
odds / (1 + odds)
```

```
        1
0.1578845
```

```
# using predict
predict(heart_disease_fit, new_data = new_person, type = "prob")
```

```
# A tibble: 1 x 2
  .pred_0 .pred_1
    <dbl>   <dbl>
1   0.842   0.158
```

Comparing observations

Suppose a person's cholesterol changes from 250 mg/dL to 200 mg/dL.

7. How do you expect the log-odds that this person is high risk for heart disease to change?

```
new_people <- tibble(totChol = c(250, 200))
log_odds <- predict(heart_disease_fit, new_data = new_people, type = "raw")
log_odds
```

```
        1         2
-1.674053 -1.918089
```

8. How do you expect the odds that this person is high risk for heart disease to change?

```
# odds
exp(log_odds)
```

```
        1         2
0.1874856 0.1468874
```

```
# probabilities
## using odds
odds / (1 + odds)
```

```
        1
0.1578845
```

```r
## using predict
predict(heart_disease_fit, new_data = new_people, type = "prob")
```

```
# A tibble: 2 x 2
  .pred_0 .pred_1
    <dbl>   <dbl>
1   0.842   0.158
2   0.872   0.128
```