Logistic regression

STA 210 - Summer 2022

Yunran Chen

Welcome

Announcements

- Any questions on project proposals?
- Exam 2 is due on 11:59 pm today.

Topics

- Logistic regression for binary response variable
- Relationship between odds and probabilities
- Use logistic regression model to calculate predicted odds and probabilities

Computational setup

```r
# load packages
library(tidyverse)
library(tidymodels)
library(knitr)
library(Stat2Data)

# set default theme and larger font size for ggplot2
ggplot2::theme_set(ggplot2::theme_minimal(base_size = 20))
```

Predicting categorical outcomes

Types of outcome variables

Quantitative outcome variable:

- Sales price of a house in Levittown, NY
- **Model**: Expected sales price given the number of bedrooms, lot size, etc.

Types of outcome variables

Quantitative outcome variable:

- Sales price of a house in Levittown, NY
- **Model**: Expected sales price given the number of bedrooms, lot size, etc.

Categorical outcome variable:

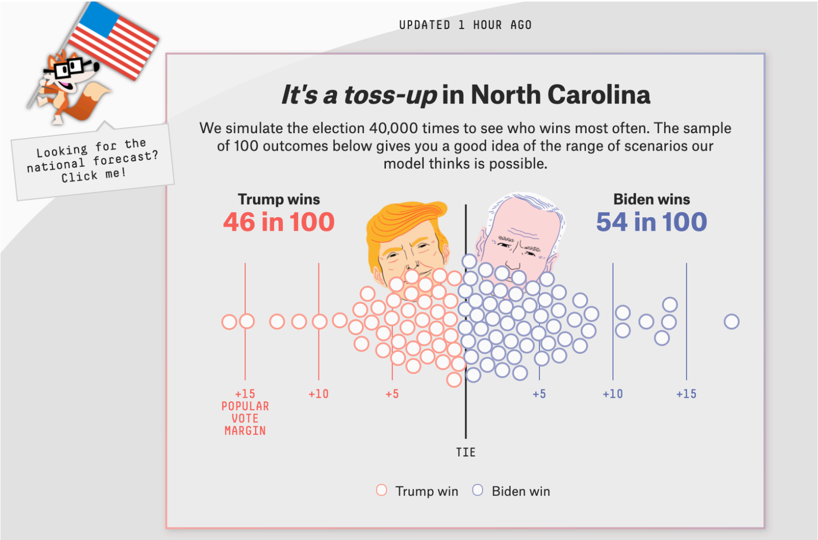
- High risk of coronary heart disease
- **Model**: Probability an adult is high risk of heart disease given their age, total cholesterol, etc.

Models for categorical outcomes

Logistic regression
2 Outcomes
1: Yes, 0: No

Multinomial logistic regression
3+ Outcomes
1: Democrat, 2: Republican, 3: Independent

2020 election forecasts



Source: FiveThirtyEight Election Forcasts

NBA finals predictions

Sunday, June 5 FINALS

Game 2 • 8 p.m. Eastern	RAPTOR SPREAD	WIN PROB.	SCORE
Celtics 1-0	-2	57%	
Warriors		43%	

Source: FiveThirtyEight 2021-22 NBA Predictions

Do teenagers get 7+ hours of sleep?

Students in grades 9 - 12 surveyed about health risk behaviors including whether they usually get 7 or more hours of sleep.

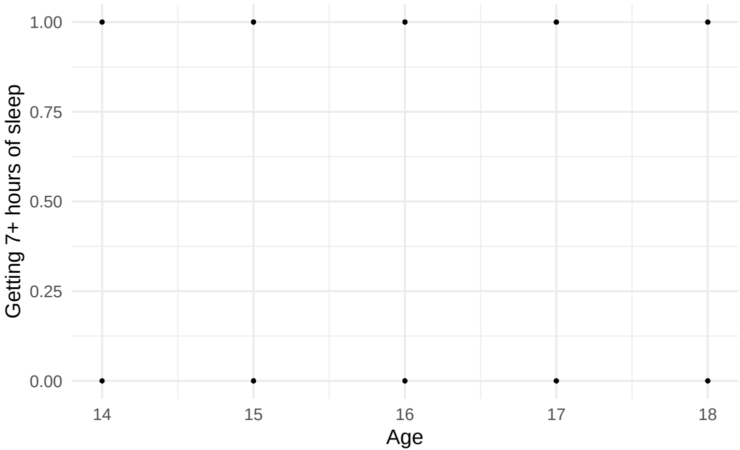
`Sleep7`

1: yes

0: no

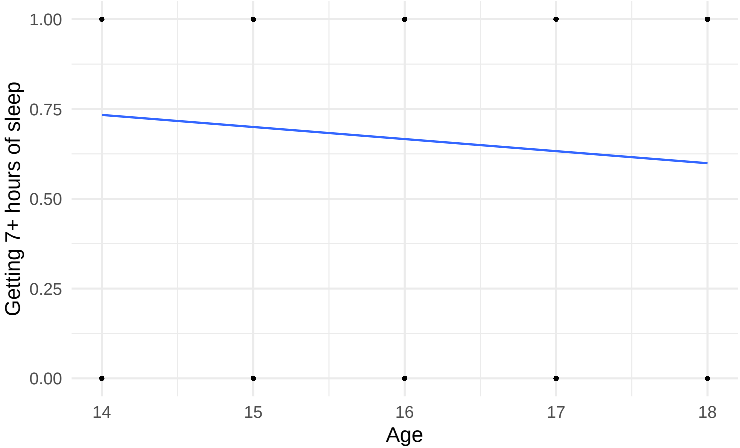
```
# A tibble: 446 x 6
      Age Sleep7 Sleep                Smok
    <int>  <int> <fct>                <fct
 1     16      1 8 hours              Yes
 2     17      0 5 hours              Yes
 3     18      0 5 hours              Yes
 4     17      1 7 hours              Yes
 5     15      0 4 or less hours No
 6     17      0 6 hours              No
 7     17      1 7 hours              No
 8     16      1 8 hours              Yes
 9     16      1 8 hours              No
10     18      0 4 or less hours Yes
# ... with 436 more rows
```

Plot the data



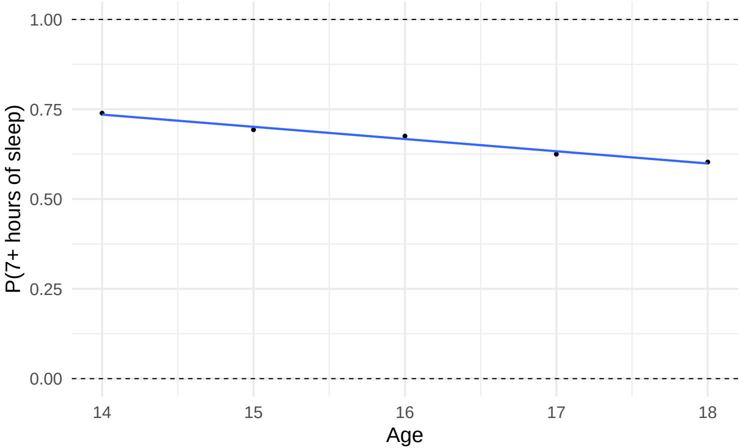
Let's fit a linear regression model

Outcome: $Y = 1$: yes, 0: no



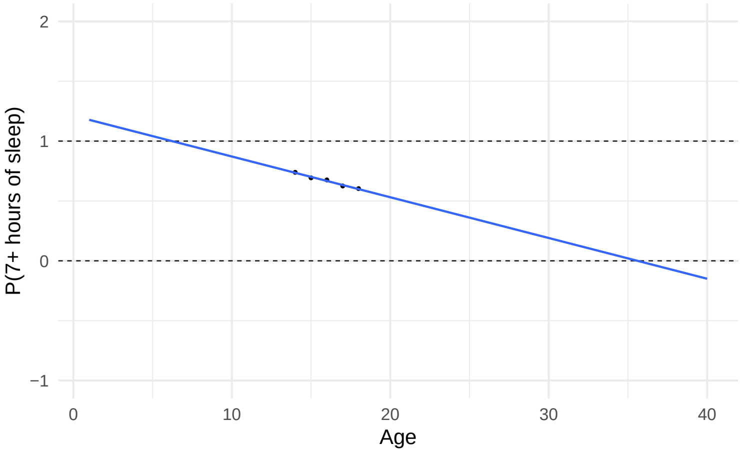
Let's use proportions

Outcome: Probability of getting 7+ hours of sleep



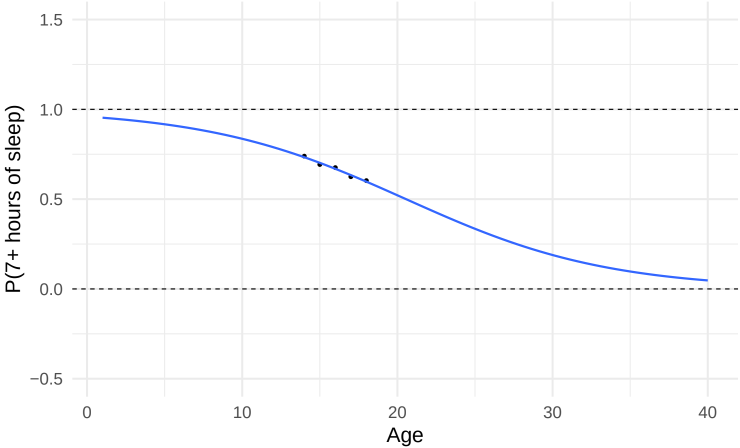
What happens if we zoom out?

Outcome: Probability of getting 7+ hours of sleep



This model produces predictions outside of 0 and 1.

Let's try another model



This model (called a ***logistic regression model****) only produces predictions between 0 and 1.*

The code

```
ggplot(sleep_age, aes(x = Age, y = prop)) +
  geom_point() +
  geom_hline(yintercept = c(0,1), lty = 2) +
  stat_smooth(method ="glm", method.args = list(family = "b
              fullrange = TRUE, se = FALSE) +
  labs(y = "P(7+ hours of sleep)") +
  xlim(1, 40) +
  ylim(-0.5, 1.5)
```

Different types of models

Method	Outcome	Model
Linear regression	Quantitative	$Y = \beta_0 + \beta_1 \ X$
Logistic regression	Binary	$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \ X$

Odds and probabilities

Binary response variable

- $Y = 1:$ yes, $0:$ no
- π: **probability** that $Y = 1$, i.e., $P(Y = 1)$
- $\frac{\pi}{1-\pi}$: **odds** that $Y = 1$
- $\log\left(\frac{\pi}{1-\pi}\right)$: **log odds**
- Go from π to $\log\left(\frac{\pi}{1-\pi}\right)$ using the **logit transformation**

Odds

Suppose there is a **70% chance** it will rain tomorrow

- Probability it will rain is $\mathbf{p = 0.7}$
- Probability it won't rain is $\mathbf{1 - p = 0.3}$
- Odds it will rain are **7 to 3**, **7:3**, $\mathbf{\frac{0.7}{0.3} \approx 2.33}$

Are teenagers getting enough sleep?

```
# A tibble: 2 x 3
  Sleep7     n     p
   <int> <int> <dbl>
1      0   150 0.336
2      1   296 0.664
```

Are teenagers getting enough sleep?

```
# A tibble: 2 x 3
  Sleep7     n     p
   <int> <int> <dbl>
1      0   150 0.336
2      1   296 0.664
```

$P(7+ \text{ hours of sleep}) = P(Y = 1) = p = 0.664$

Are teenagers getting enough sleep?

```
# A tibble: 2 x 3
  Sleep7     n     p
   <int> <int> <dbl>
1      0   150 0.336
2      1   296 0.664
```

$P(\text{7+ hours of sleep}) = P(Y = 1) = p = 0.664$

$P(< \text{7 hours of sleep}) = P(Y = 0) = 1 - p = 0.336$

Are teenagers getting enough sleep?

```
# A tibble: 2 x 3
  Sleep7     n     p
   <int> <int> <dbl>
1      0   150 0.336
2      1   296 0.664
```

$P(\text{7+ hours of sleep}) = P(Y = 1) = p = 0.664$

$P(< \text{7 hours of sleep}) = P(Y = 0) = 1 - p = 0.336$

$P(\text{odds of 7+ hours of sleep}) = \frac{0.664}{0.336} = 1.976$

From odds to probabilities

odds

$$\omega = \frac{\pi}{1 - \pi}$$

probability

$$\pi = \frac{\omega}{1 + \omega}$$

Logistic regression

From odds to probabilities

(1) **Logistic model**: log odds $= \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1\ X$
(2) **Odds =** $\exp\left\{\log\left(\frac{\pi}{1-\pi}\right)\right\} = \frac{\pi}{1-\pi}$
(3) Combining (1) and (2) with what we saw earlier

$$\text{probability} = \pi = \frac{\exp\{\beta_0 + \beta_1\ X\}}{1 + \exp\{\beta_0 + \beta_1\ X\}}$$

Logistic regression model

Logit form:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \ X$$

Logistic regression model

Logit form:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \ X$$

Probability form:

$$\pi = \frac{\exp\{\beta_0 + \beta_1 \ X\}}{1 + \exp\{\beta_0 + \beta_1 \ X\}}$$

Risk of coronary heart disease

This dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. We want to use age to predict if a randomly selected adult is high risk of having coronary 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

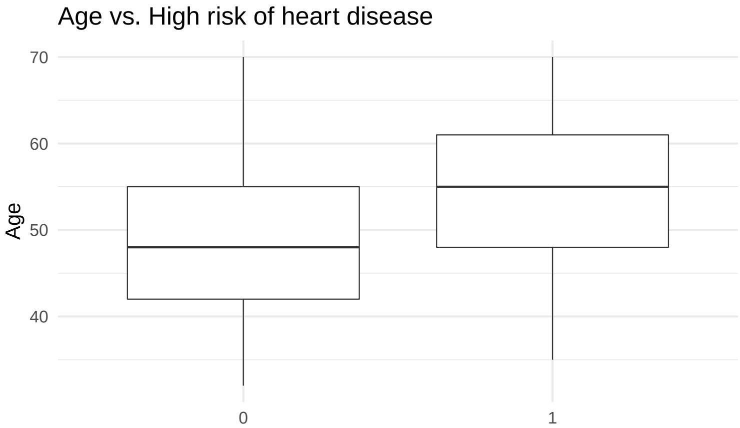
`age`: Age at exam time (in years)

Data: heart

```
# A tibble: 4,240 x 2
     age high_risk
   <dbl> <fct>
 1    39 0
 2    46 0
 3    48 0
 4    61 1
 5    46 0
 6    43 0
 7    63 1
 8    45 0
 9    52 0
10    43 0
# ... with 4,230 more rows
```

High risk vs. age

```
ggplot(heart_disease, aes(x = high_risk, y = age)) +
  geom_boxplot() +
  labs(x = "High risk - 1: yes, 0: no",
       y = "Age",
       title = "Age vs. High risk of heart disease")
```



Let's fit the model

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

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

term	estimate	std.error	statistic	p.value
(Intercept)	-5.561	0.284	-19.599	0
age	0.075	0.005	14.178	0

The model

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

term	estimate	std.error	statistic	p.value
(Intercept)	-5.561	0.284	-19.599	0
age	0.075	0.005	14.178	0

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -5.561 + 0.075 \times \text{age}$$

where $\hat{\pi}$ is the predicted probability of being high risk

Predicted log odds

```
# A tibble: 4,240 x 8
   high_risk   age .fitted .resid .std.resid      .hat .sig
   <fct>     <dbl>   <dbl>  <dbl>      <dbl>     <dbl>  <db
 1 0            39  -2.65  -0.370     -0.370 0.000466   0.89
 2 0            46  -2.13  -0.475     -0.475 0.000322   0.89
 3 0            48  -1.98  -0.509     -0.509 0.000288   0.89
 4 1            61  -1.01   1.62       1.62  0.000706   0.89
 5 0            46  -2.13  -0.475     -0.475 0.000322   0.89
 6 0            43  -2.35  -0.427     -0.427 0.000384   0.89
 7 1            63  -0.858  1.56       1.56  0.000956   0.89
 8 0            45  -2.20  -0.458     -0.458 0.000342   0.89
 9 0            52  -1.68  -0.585     -0.585 0.000262   0.89
10 0            43  -2.35  -0.427     -0.427 0.000384   0.89
# ... with 4,230 more rows
```

Predicted log odds

```
# A tibble: 4,240 x 8
   high_risk   age .fitted .resid .std.resid     .hat .sigm
   <fct>     <dbl>   <dbl>  <dbl>      <dbl>    <dbl> <dbl
 1 0            39  -2.65  -0.370     -0.370 0.000466  0.89
 2 0            46  -2.13  -0.475     -0.475 0.000322  0.89
 3 0            48  -1.98  -0.509     -0.509 0.000288  0.89
 4 1            61  -1.01   1.62       1.62  0.000706  0.89
 5 0            46  -2.13  -0.475     -0.475 0.000322  0.89
 6 0            43  -2.35  -0.427     -0.427 0.000384  0.89
 7 1            63  -0.858  1.56       1.56  0.000956  0.89
 8 0            45  -2.20  -0.458     -0.458 0.000342  0.89
 9 0            52  -1.68  -0.585     -0.585 0.000262  0.89
10 0            43  -2.35  -0.427     -0.427 0.000384  0.89
# ... with 4,230 more rows
```

For observation 1

$$\text{predicted odds} = \hat{\omega} = \frac{\hat{\pi}}{\quad} = \exp\{-2.650\} = 0.071$$

Predicted probabilities

```
# A tibble: 4,240 x 2
   .pred_0 .pred_1
     <dbl>   <dbl>
 1   0.934  0.0660
 2   0.894  0.106
 3   0.878  0.122
 4   0.733  0.267
 5   0.894  0.106
 6   0.913  0.0870
 7   0.702  0.298
 8   0.900  0.0996
 9   0.843  0.157
10   0.913  0.0870
# ... with 4,230 more rows
```

Predicted probabilities

```
# A tibble: 4,240 x 2
   .pred_0 .pred_1
     <dbl>   <dbl>
 1   0.934  0.0660
 2   0.894  0.106
 3   0.878  0.122
 4   0.733  0.267
 5   0.894  0.106
 6   0.913  0.0870
 7   0.702  0.298
 8   0.900  0.0996
 9   0.843  0.157
10   0.913  0.0870
# ... with 4,230 more rows
```

$$\text{predicted probabilities} = \hat{\pi} = \frac{\exp\{-2.650\}}{1 + \exp\{-2.650\}} = 0.066$$

Predicted classes

```
# A tibble: 4,240 x 1
   .pred_class
   <fct>
 1 0
 2 0
 3 0
 4 0
 5 0
 6 0
 7 0
 8 0
 9 0
10 0
# ... with 4,230 more rows
```

Default prediction

For a logistic regression, the default prediction is the `class`.

```
# A tibble: 4,240 x 1
   .pred_class
   <fct>
 1 0
 2 0
 3 0
 4 0
 5 0
 6 0
 7 0
 8 0
 9 0
10 0
# ... with 4,230 more rows
```

Observed vs. predicted

What does the following table show?

```
# A tibble: 2 x 3
  high_risk .pred_class     n
  <fct>     <fct>       <int>
1 0         0            3596
2 1         0             644
```

Recap

- Logistic regression for binary response variable
- Relationship between odds and probabilities
- Used logistic regression model to calculate predicted odds and probabilities

Application exercise

ae-9-odds