Logistic Regression

Odds + probabilities

Prof. Maria Tackett

[Click here for PDF of slides]

Topics

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

Types of response variables

Quantitative response variable:

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

Types of response variables

Quantitative response variable:

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

Categorical response 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 response variables

Logistic Regression

2 Outcomes

1: Yes, 0: No

Models for categorical response variables

Logistic Regression

2 Outcomes

1: Yes, 0: No

Multinomial Logistic Regression

3+ Outcomes

1: Democrat, 2: Republican, 3: Independent

Models for categorical response variables

Logistic Regression

2 Outcomes

1: Yes, 0: No

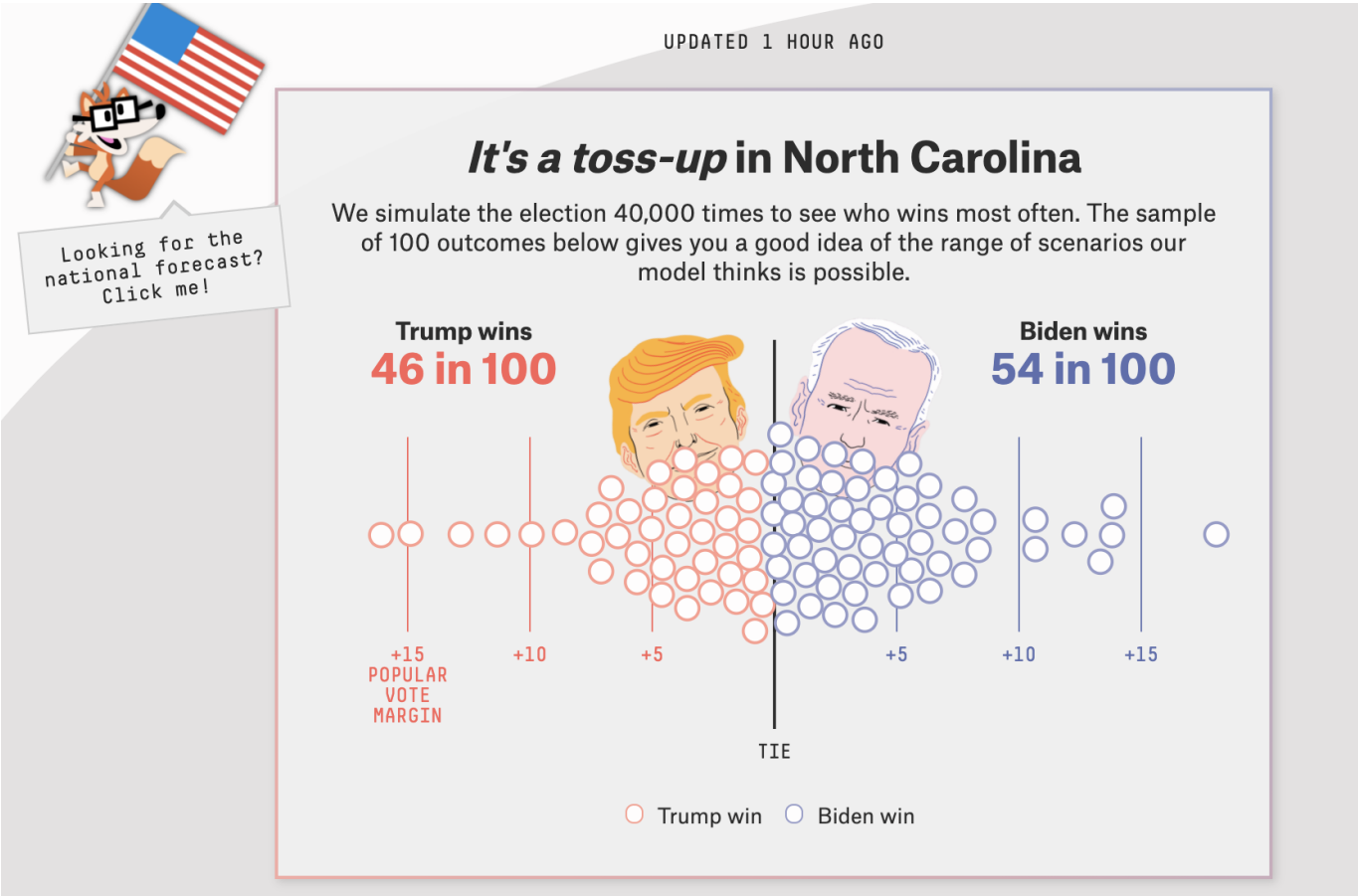
Multinomial Logistic Regression

3+ Outcomes

1: Democrat, 2: Republican, 3: Independent

Let's focus on logistic regression models for now.

FiveThirtyEight 2020 election forcasts



[FiveThirtyEight Election Forcasts]

FiveThirtyEight NBA finals predictions

Friday, Oct. 2 FINALS

Game 2 • FINAL	RAPTOR SPREAD	WIN PROB.	SCORE
Heat		43%	114
Lakers 2-0	-2	57%	✓ 124

Wednesday, Sept. 30 FINALS

Game 1 • FINAL	RAPTOR SPREAD	WIN PROB.	SCORE
Heat	-5	68%	98
Lakers 1-0		32%	✓ 116

[2019-20 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`

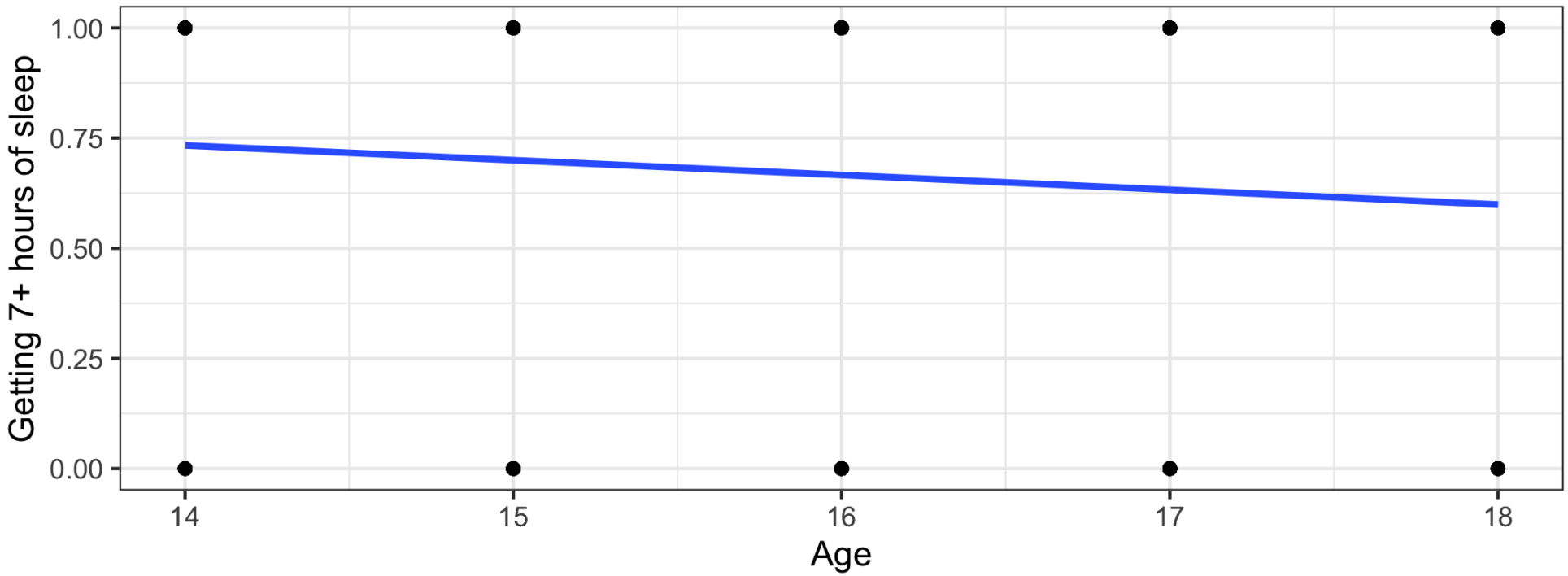
1: yes

0: no

Age	Sleep7
16	1
17	0
18	0
17	1
15	0
17	0
17	1
16	1
16	1
18	0

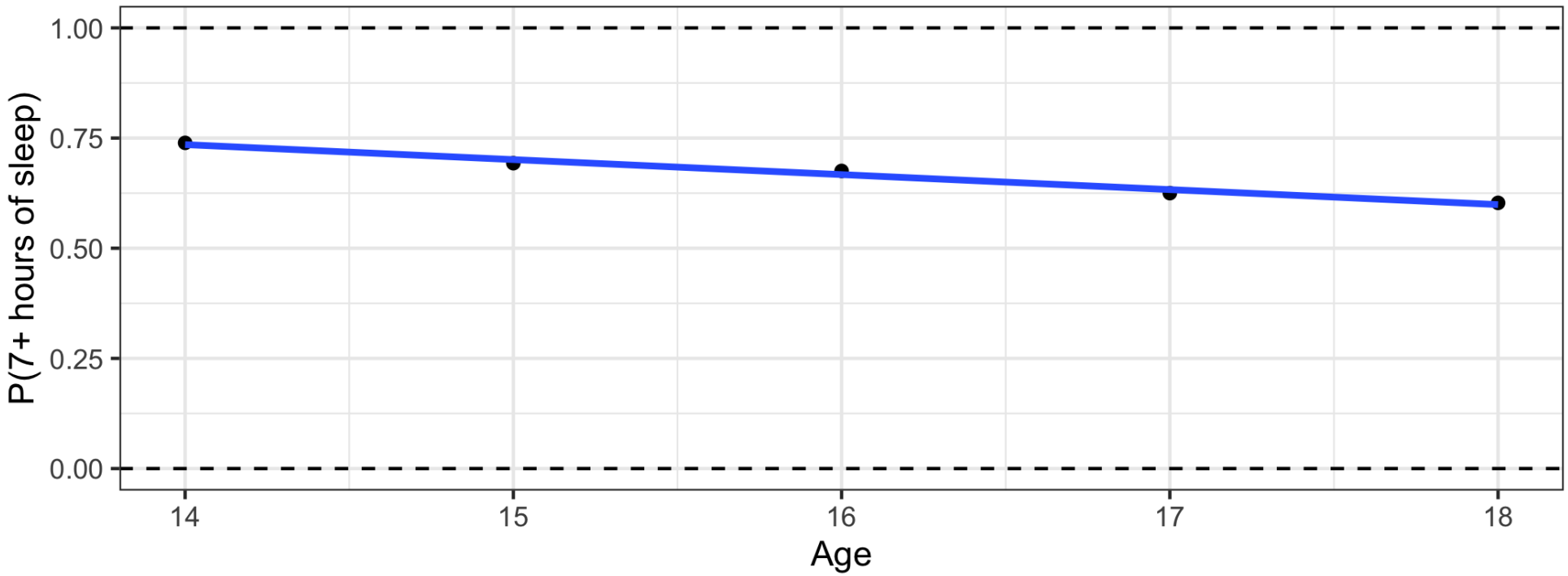
Let's fit a linear regression model

Response: Y = 1: yes, 0: no



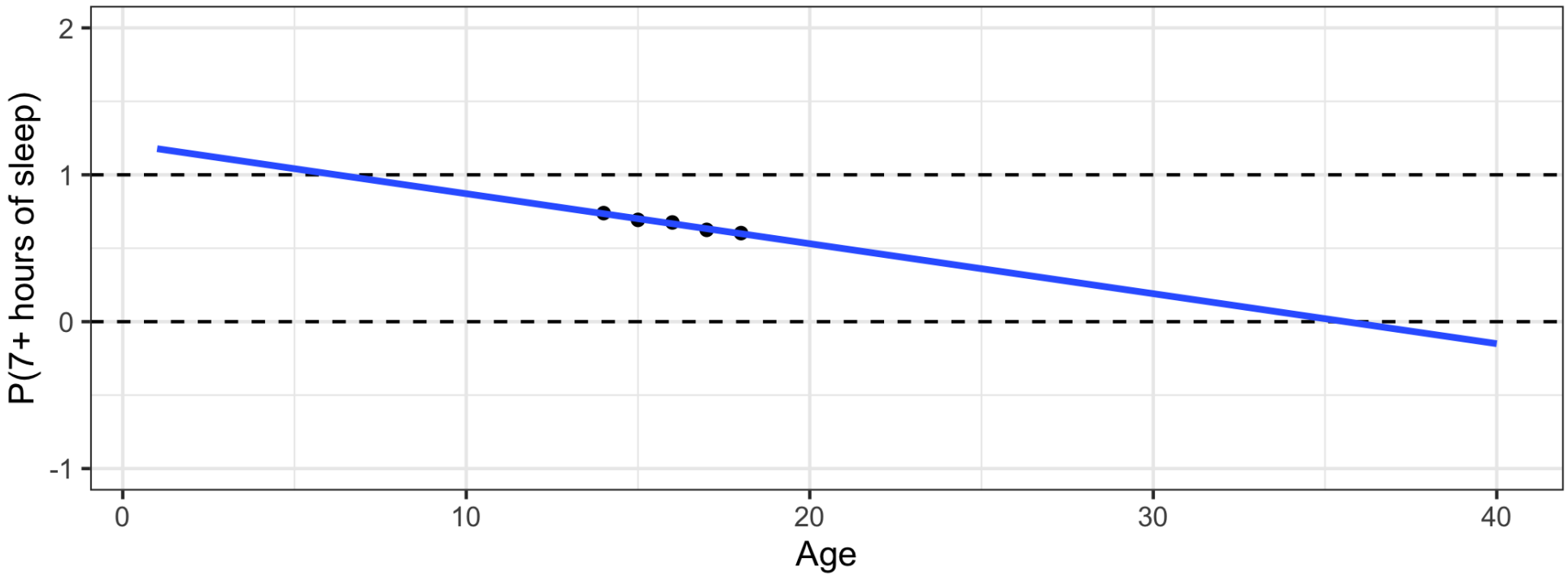
Let's use proportions

Response: Probability of getting 7+ hours of sleep



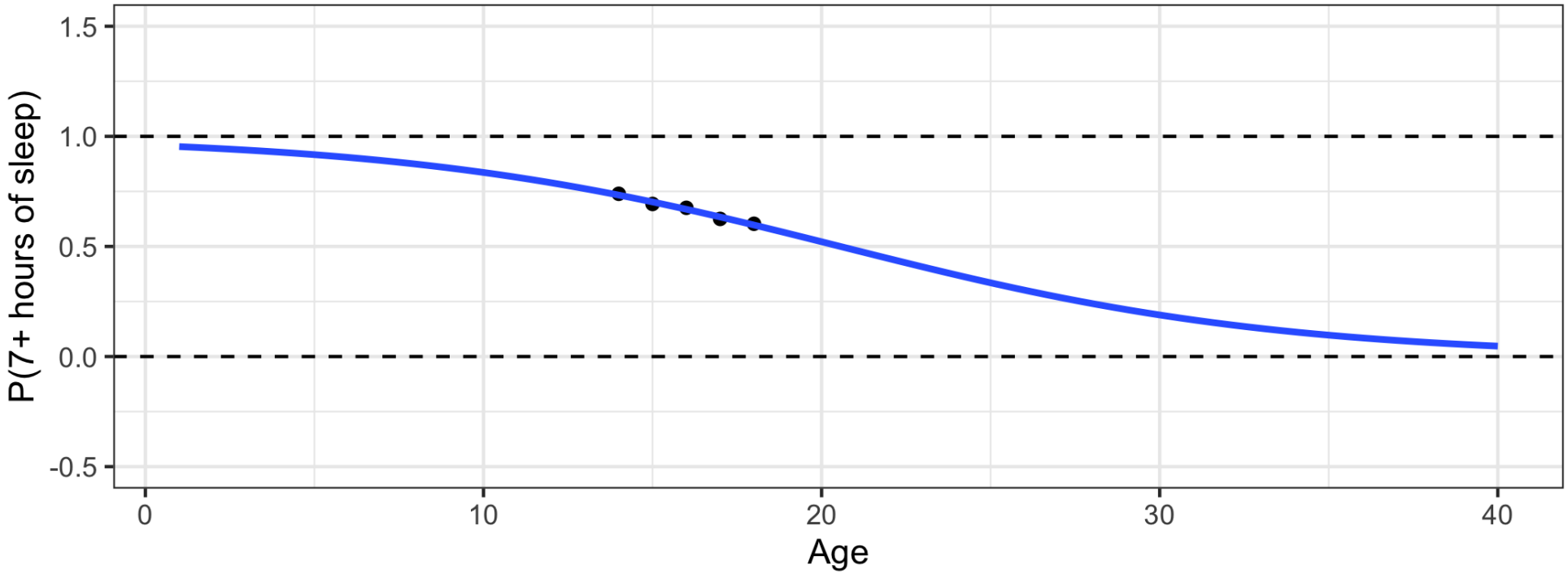
What happens if we zoom out?

Response: 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.

Different types of models

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

Binary response variable

- $Y = 1: \text{ yes}, 0: \text{ no}$

Binary response variable

- $Y = 1: \text{ yes}, 0: \text{ no}$
- π: **probability** that $Y = 1$, i.e., $P(Y = 1)$

Binary response variable

- $Y = 1: \text{ yes}, 0: \text{ no}$
- π: **probability** that $Y = 1$, i.e., $P(Y = 1)$
- $\frac{\pi}{1-\pi}$: **odds** that $Y = 1$

Binary response variable

- $Y = 1: \text{ yes}, 0: \text{ no}$
- 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**

Binary response variable

- $Y = 1: \text{ yes}, 0: \text{ no}$
- 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

Odds

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

- Probability it will rain is $\mathbf{p = 0.7}$

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

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**, $\frac{\mathbf{0.7}}{\mathbf{0.3}} \approx \mathbf{2.33}$

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
```

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

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
```

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

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

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
```

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

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

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

From odds to probabilities

odds

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

From odds to probabilities

odds

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

probability

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

Logistic model: from odds to probabilities

1️⃣ **Logistic model**: log odds = $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$

Logistic model: from odds to probabilities

1️⃣ **Logistic model**: log odds = $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$

2️⃣ **odds** = $\exp\left\{\log\left(\frac{\pi}{1-\pi}\right)\right\} = \frac{\pi}{1-\pi}$

Logistic model: from odds to probabilities

1️⃣ **Logistic model**: log odds = $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$

2️⃣ **odds** = $\exp\left\{\log\left(\frac{\pi}{1-\pi}\right)\right\} = \frac{\pi}{1-\pi}$

Combining 1️⃣ and 2️⃣ 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

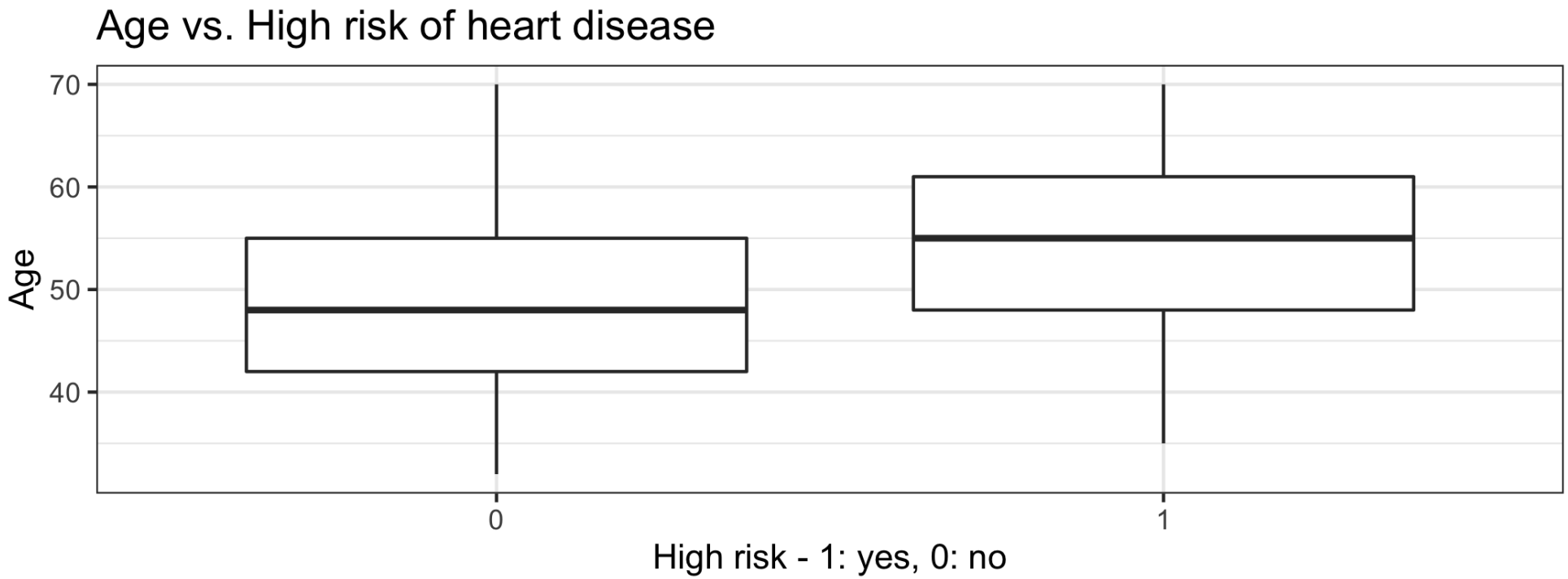
This dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. We want to use **`age`** to predict if a randomly selected adult is high risk of having coronary heart disease in the next 10 years.

`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)

High risk vs. age



Let's fit the model

```
high_risk_model <- glm(high_risk ~ age,
                       data = heart_data,
                       family = "binomial")
tidy(high_risk_model) %>%
  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

Let's fit the model

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

```
predict(high_risk_model)
```

```
##       1       2       3       4       5       6       7       8
## -2.650 -2.127 -1.978 -1.007 -2.127 -2.351 -0.858 -2.202 -1.
```

Predicted log odds

```
predict(high_risk_model)
```

```
##      1      2      3      4      5      6      7      8
## -2.650 -2.127 -1.978 -1.007 -2.127 -2.351 -0.858 -2.202 -1.
```

For observation 1

$$\text{predicted odds} = \hat{\omega} = \frac{\hat{\pi}}{1-\hat{\pi}} = \exp\{-2.650\} = 0.071$$

Predcited probabilities

```
predict(high_risk_model,
        type = "response")
```

```
##     1     2     3     4     5     6     7     8     9    10
## 0.066 0.106 0.122 0.267 0.106 0.087 0.298 0.100 0.157 0.087
```

Predcited probabilities

```
predict(high_risk_model,
        type = "response")
```

```
##     1     2     3     4     5     6     7     8     9    10
## 0.066 0.106 0.122 0.267 0.106 0.087 0.298 0.100 0.157 0.087
```

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

Recap

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