Parametric models and logistic regression

Ciaran Evans

Warmup activity

Work on the activity (handout) with a neighbor, then we will discuss as a class.

Warmup

$$\text{odds} = \frac{\pi}{1 - \pi}$$

If $\pi = 0.2$, calculate the odds.

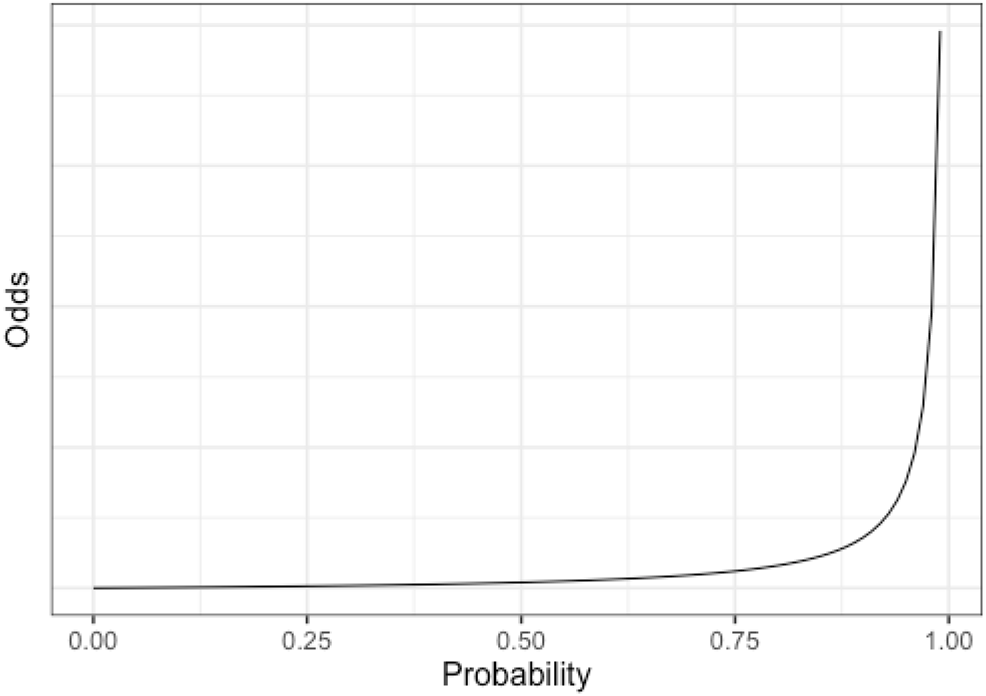
Warmup

$$\text{odds} = \frac{\pi}{1-\pi}$$

What happens to odds as $\pi \to 0$? As $\pi \to 1$?

Warmup

$$\text{odds} = \frac{\pi}{1-\pi}$$



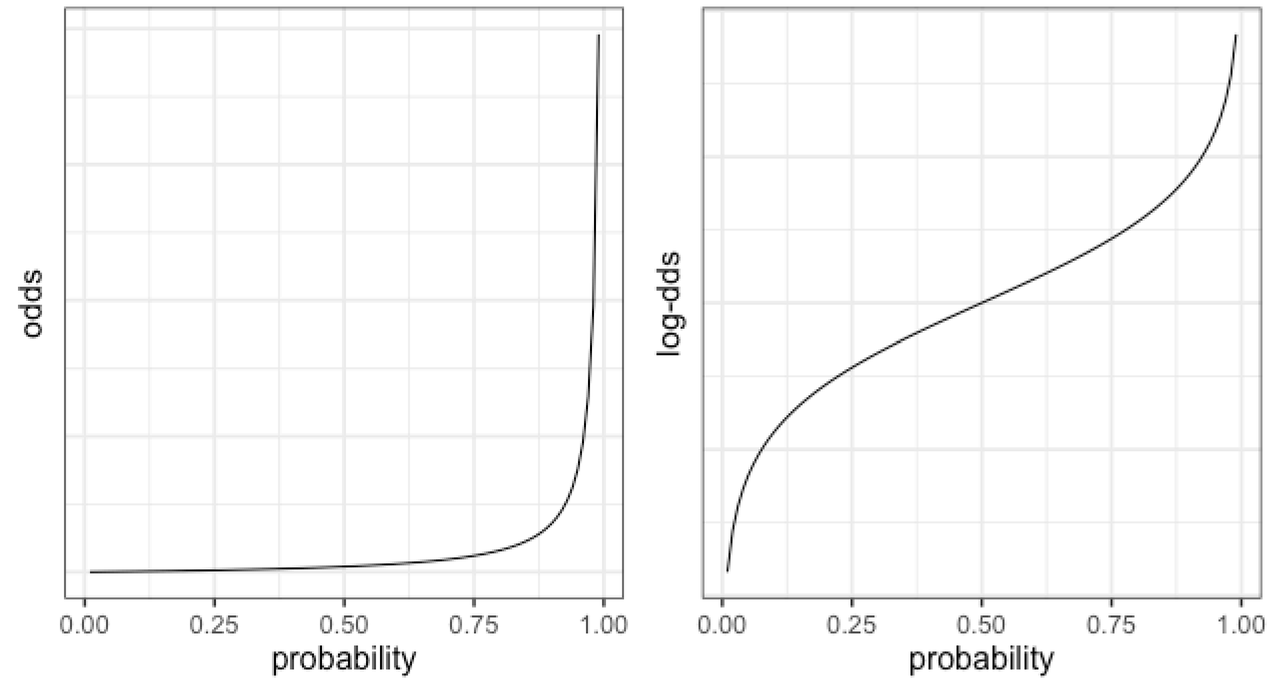
Warmup

$$\text{log-odds} = \log(\text{odds}) = \log\left(\frac{\pi}{1-\pi}\right)$$

What happens to log-odds as $\pi \to 0$? As $\pi \to 1$?

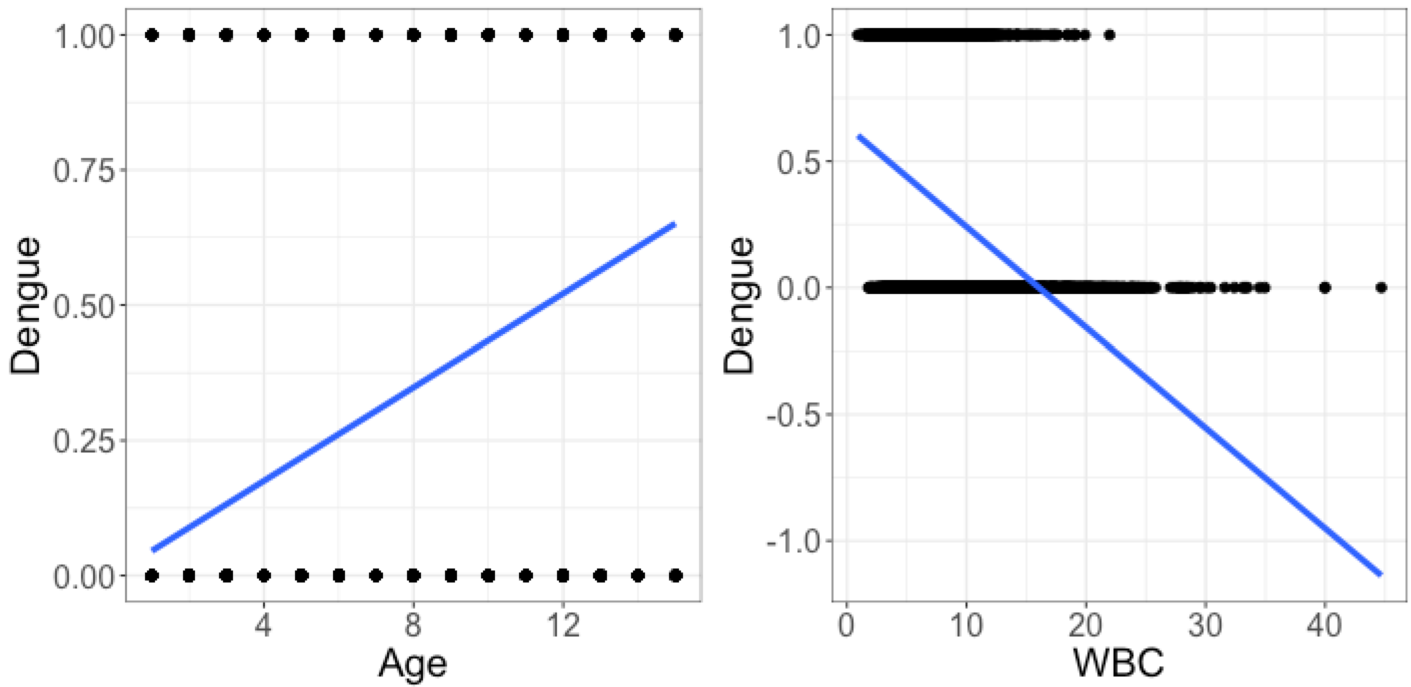
Warmup

$$\text{log-odds} = \log(\text{odds}) = \log\left(\frac{\pi}{1-\pi}\right)$$



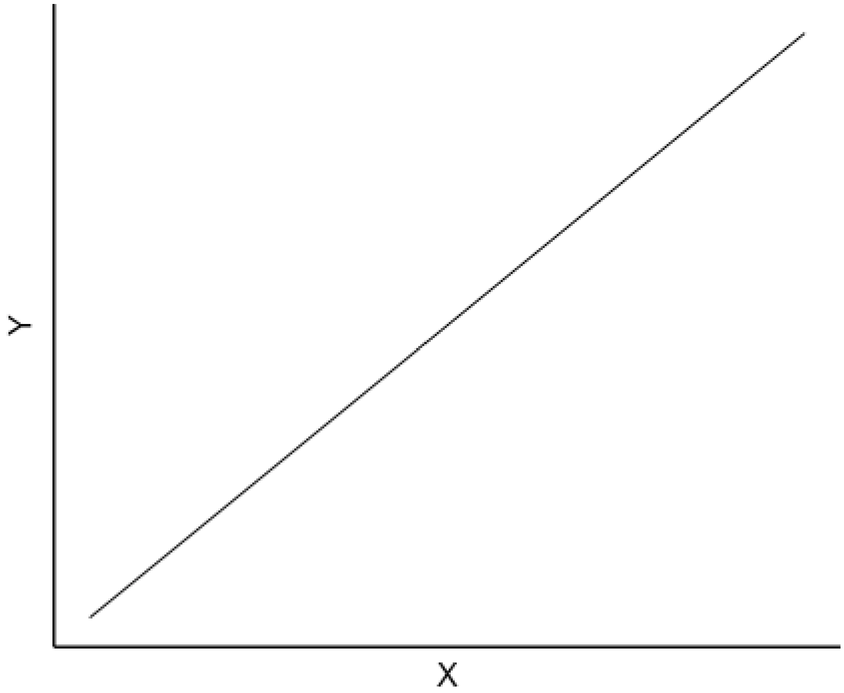
Last time

Don't fit linear regression with a binary response



Revisiting the linear regression model

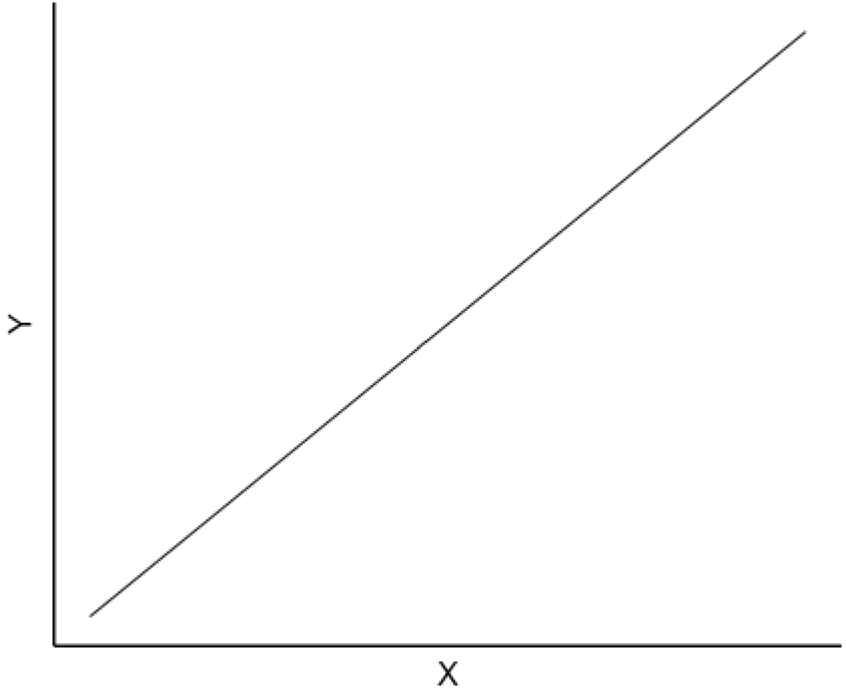
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



> Will all of the observations fall exactly on the line?

Revisiting the linear regression model

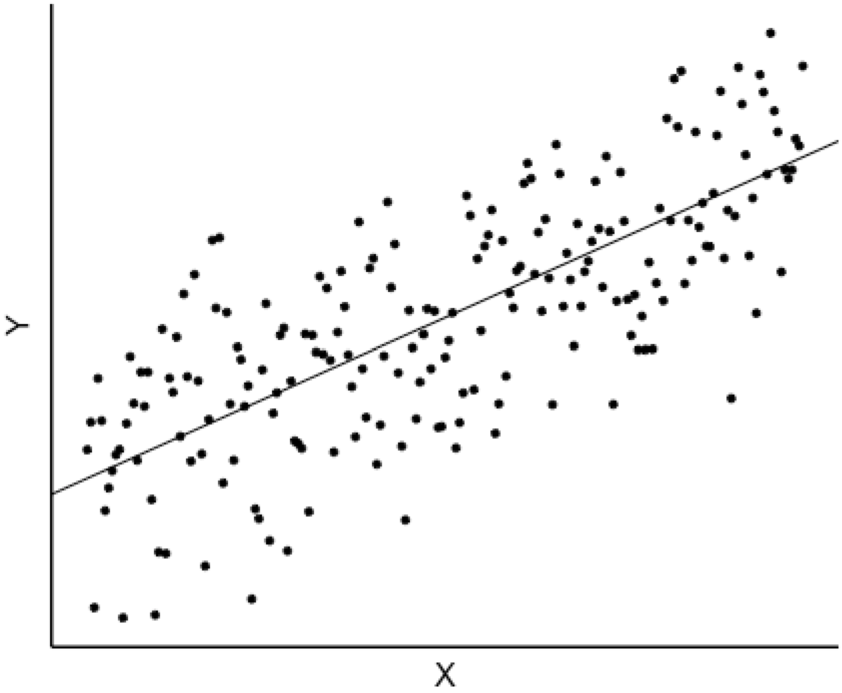
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Given a value of X, how do I know where the values of Y are likely to be?

Revisiting the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



What do we often assume about the distribution of ε?

Revisiting the linear regression model

Parametric modeling

A regression model is an example of a more general process called **parametric modeling**

- **Step 1:** Choose a reasonable distribution for Y_i
- **Step 2:** Build a model for the parameters of interest
- **Step 3:** Fit the model

Step 1: Choose a reasonable distribution for Y_i

What do I mean by a *distribution*?

Step 1: Choose a reasonable distribution for Y_i

> What do I mean by a *distribution*?

- A **distribution** tells us what outcomes are possible for Y_i, and how often these outcomes occur.

Here the possible values of Y_i are 0 (no dengue) and 1 (dengue).

> How often do these values occur in the population?

Step 1: Choose a reasonable distribution for Y_i

> What do I mean by a *distribution*?

- A **distribution** tells us what outcomes are possible for Y_i, and how often these outcomes occur.

Here the possible values of Y_i are 0 (no dengue) and 1 (dengue).

> How often do these values occur in the population?

- We don't know, so we will estimate from the sample
- We assume the probability $Y_i = 1$ depends on Age_i

Step 1: Choose a reasonable distribution for Y_i

> How should I describe the distribution of Y_i?

Bernoulli distribution

Definition: Let Y_i be a binary random variable, and $\pi_i = P(Y_i = 1)$. Then $Y_i \sim Bernoulli(\pi_i)$.

> What do I mean by a *random variable*?

Bernoulli distribution

Definition: Let Y_i be a binary random variable, and $\pi_i = P(Y_i = 1)$. Then $Y_i \sim Bernoulli(\pi_i)$.

> What do I mean by a *random variable*?

A **random variable** is an event that has a set of possible outcomes, but we don't know which one will occur

- Here $Y_i = 0$ or 1
- Our goal is to use the observed data to estimate $\pi_i = P(Y_i = 1)$

Second attempt at a model

$$Y_i \sim Bernoulli(\pi_i) \qquad \pi_i = P(Y_i = 1 | Age_i)$$

$$\pi_i = \beta_0 + \beta_1 Age_i$$

Are there still any potential issues with this approach?

Fixing the issues

Logistic regression model

$Y_i =$ dengue status (0 = negative, 1 = positive)

$Age_i =$ age (in years)

Random component: $Y_i \sim Bernoulli(\pi_i)$

Systematic component: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\ Age_i$

Logistic regression model

$Y_i =$ dengue status (0 = negative, 1 = positive)

$Age_i =$ age (in years)

Random component: $Y_i \sim Bernoulli(\pi_i)$

Systematic component: $$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \; Age_i$$

> Why is there no noise term ε_i in the logistic regression model? Discuss for 1--2 minutes with your neighbor, then we will discuss as a class.

Fitting the logistic regression model

$$Y_i \sim Bernoulli(\pi_i)$$

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

```
m1 <- glm(Dengue ~ Age, data = dengue,
          family = binomial)
summary(m1)
```

Fitting the logistic regression model

$$Y_i \sim Bernoulli(\pi_i)$$

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

```
m1 <- glm(Dengue ~ Age, data = dengue,
          family = binomial)
summary(m1)
```

```
...
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.454345   0.075068  -32.70   <2e-16 ***
## Age          0.217312   0.008826   24.62   <2e-16 ***
...
```

Class activity

- Work with a neighbor on the class activity (handout)
- I will collect your work at the end of class

For next time, read sections 6.4 and 6.6 in the textbook