

Lecture 1: Intro to logistic regression

Ciaran Evans

Motivating example: Dengue fever

Dengue fever: a mosquito-borne viral disease affecting 400 million people a year

Motivating example: Dengue data

Data: Data on 5720 Vietnamese children, admitted to the hospital with possible dengue fever. Variables include:

- ▶ *Sex*: patient's sex (female or male)
- ▶ *Age*: patient's age (in years)
- ▶ *WBC*: white blood cell count
- ▶ *PLT*: platelet count
- ▶ other diagnostic variables. . .
- ▶ *Dengue*: whether the patient has dengue (0 = no, 1 = yes)

Motivating example: Dengue data

Research questions:

- ▶ How well can we predict whether a patient has dengue?
- ▶ Which diagnostic measurements are most useful?
- ▶ Is there a significant relationship between WBC and dengue?

Research questions

- ▶ How well can we predict whether a patient has dengue?
- ▶ Which diagnostic measurements are most useful?
- ▶ Is there a significant relationship between WBC and dengue?

How can I answer each of these questions? Discuss with a neighbor for 2 minutes, then we will discuss as a class.

Fitting a model: initial attempt

What if we try a linear regression model?

Y_i = dengue status of i th patient

$$Y_i = \beta_0 + \beta_1 WBC_i + \varepsilon_i \quad \varepsilon_i \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

What are some potential issues with this linear regression model?

Second attempt

Let's rewrite the linear regression model:

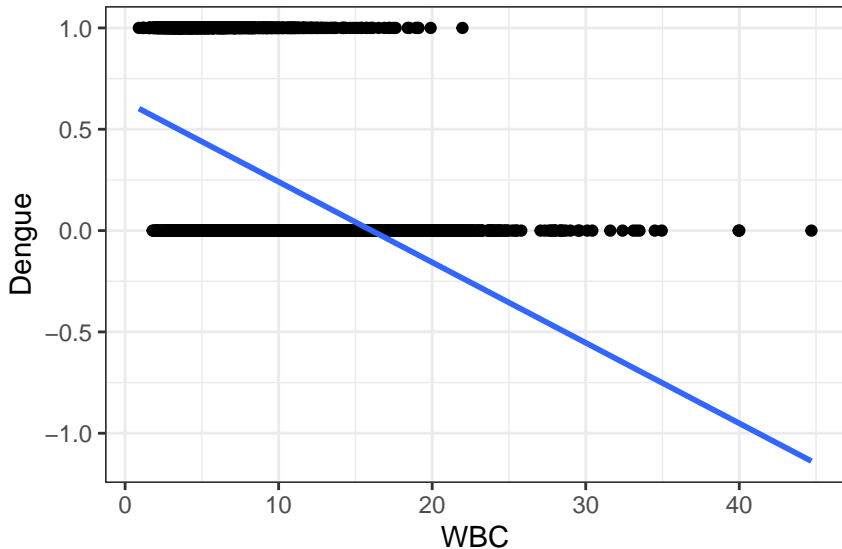
Second attempt

$$Y_i \sim \text{Bernoulli}(p_i) \quad p_i = \mathbb{P}(Y_i = 1 | WBC_i)$$

$$p_i = \beta_0 + \beta_1 WBC_i$$

Are there still any potential issues with this approach?

Don't fit linear regression with a binary response



Fixing the issue: logistic regression

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$g(p_i) = \beta_0 + \beta_1 \text{WBC}_i$$

where $g : (0, 1) \rightarrow \mathbb{R}$ is unbounded.

Usual choice: $g(p_i) = \log \left(\frac{p_i}{1 - p_i} \right)$

Odds

Definition: If $p_i = \mathbb{P}(Y_i = 1|WBC_i)$, the **odds** are $\frac{p_i}{1 - p_i}$

Example: Suppose that $\mathbb{P}(Y_i = 1|WBC_i) = 0.8$. What are the *odds* that the patient has dengue?

Odds

Definition: If $p_i = \mathbb{P}(Y_i = 1 | WBC_i)$, the **odds** are $\frac{p_i}{1 - p_i}$

The probabilities $p_i \in [0, 1]$. The linear function

$\beta_0 + \beta_1 WBC_i \in (-\infty, \infty)$. What range of values can $\frac{p_i}{1 - p_i}$ take?

Log odds

$$g(p_i) = \log \left(\frac{p_i}{1 - p_i} \right)$$

Binary logistic regression

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{WBC}_i$$

Note: Can generalize to $Y_i \sim \text{Binomial}(m_i, p_i)$, but we won't do that yet.

Example: simple logistic regression

$Y_i = \text{dengue status (0 = no, 1 = yes)}$ $Y_i \sim \text{Bernoulli}(p_i)$

$$\log \left(\frac{\hat{p}_i}{1 - \hat{p}_i} \right) = 1.737 - 0.361 \text{ WBC}_i$$

Work in groups of 2-3 for 5 minutes on the following questions:

- ▶ Are patients with a higher WBC more or less likely to have dengue?
- ▶ Interpret the estimated slope in context of a unit change in the log odds.
- ▶ What is the change in *odds* associated with a unit increase in WBC?