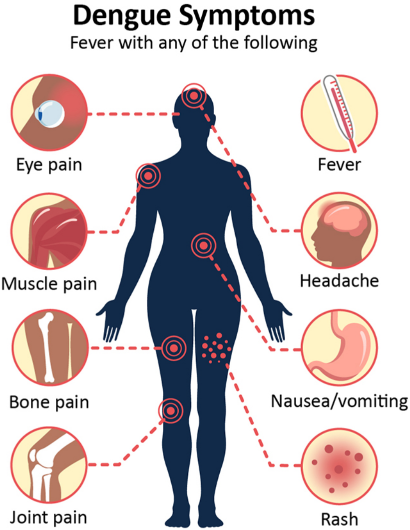
Lecture 1: Intro to logistic regression

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

Predicted \ Truth	Dengue=0	Dengue=1
Dengue = 0		
Dengue = 1		

- 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.

- regression model (e.g., logistic regression)
- prediction metrics (accuracy, confusion matrices, ROC curves, etc..)
- VIFs
- model selection (nested tests, e.g. LRTs, AIC, BIC)
  similar to nested F tests
- tests for individual coefficients in a model
  (e.g. $H_0: \beta_1 = 0$)

Fitting a model: initial attempt

What if we try a linear regression model?

$$Y_i = \text{dengue status of } i\text{th patient}$$

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

What are some potential issues with this linear regression model?

$\beta_0 + \beta_1 WBC_i + \varepsilon_i \quad \in (-\infty, \infty)$

(assuming $\beta_1 \neq 0$)

$Y_i \in \{0, 1\}$

Second attempt

random component: describes distribution of Y_i, in terms of some param.

systematic component: relates params to explanatory variables

Let's rewrite the linear regression model:

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

$$\begin{aligned}
\mathbb{E}[Y_i \mid WBC_i] &= \mathbb{E}[\beta_0 + \beta_1 WBC_i + \varepsilon_i \mid WBC_i] \\
&= \beta_0 + \beta_1 WBC_i + \underbrace{\mathbb{E}[\varepsilon_i]}_{0} \\
&= \beta_0 + \beta_1 WBC_i
\end{aligned}$$

$$Y_i \mid WBC_i \sim N(\beta_0 + \beta_1 WBC_i, \sigma_\varepsilon^2)$$

$$\Rightarrow \quad Y_i \mid WBC_i \sim N(\mu_i, \sigma_\varepsilon^2) \quad \text{(random component)}$$

$$\mu_i = \beta_0 + \beta_1 WBC_i \quad \text{(systematic component)}$$

Problem: $Y_i = 0$ or 1 $\Rightarrow$ Y_i is not normal! Let's use Bernoulli instead

Second attempt

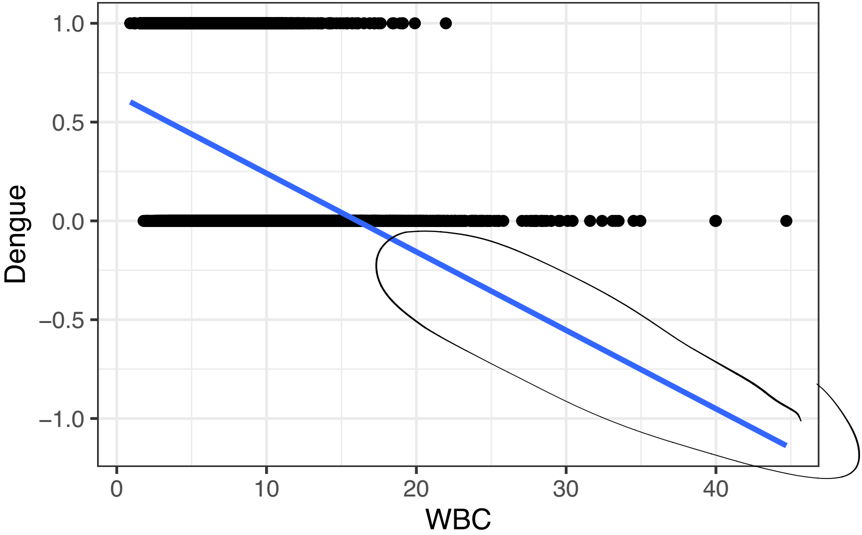
$$Y_i \sim Bernoulli(p_i) \qquad p_i = \mathbb{P}(Y_i = 1 | WBC_i)$$

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

Are there still any potential issues with this approach?

Problem : $p_i \in [0,1]$ but $\beta_0 + \beta_1 WBC_i \in (-\infty, \infty)$

Don't fit linear regression with a binary response



IF WBC > 15, $\hat{p} < 0$ (bad)

Fixing the issue: logistic regression

$$Y_i \sim Bernoulli(p_i)$$

random component

link function

$$g(p_i) = \beta_0 + \beta_1 WBC_i$$

systematic component

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

i.e. $p_i = g^{-1}(\beta_0 + \beta_1 WBC_i)$

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

log odds

(aka logit)

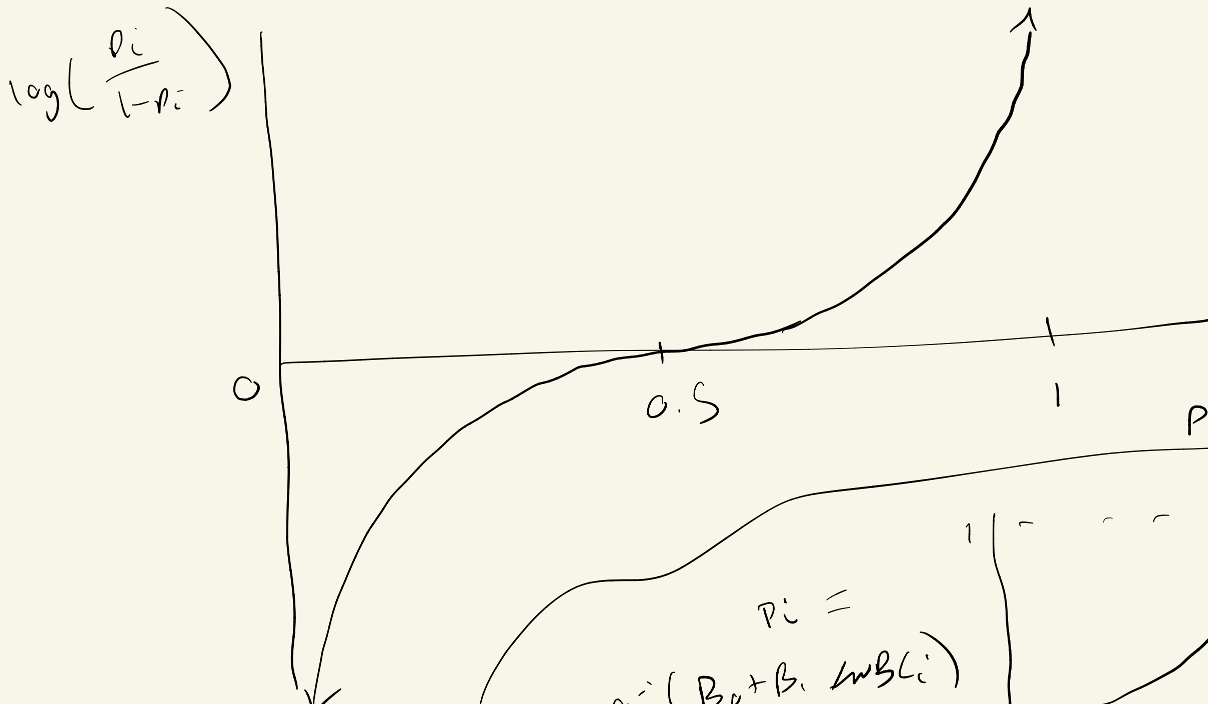
$\in (-\infty, \infty)$

$\frac{p_i}{1-p_i}$ = odds

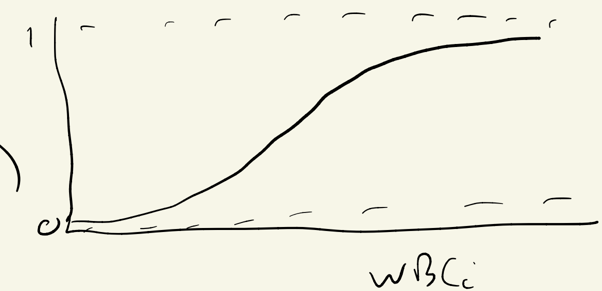
$(0, \infty)$

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

$$\log\left(\frac{0.5}{1-0.5}\right) = 0$$



$$p_i = g^{-1}(\beta_0 + \beta_1 \, WBC_i)$$



Odds

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

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

$$Y_i \sim Bernoulli(p_i)$$

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

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

Example: simple logistic regression

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

$$\log\left(\frac{\widehat{p}_i}{1-\widehat{p}_i}\right) = 1.737 - 0.361\ WBC_i$$

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?
- What is the change in the log odds associated with a unit increase in WBC?
- What is the change in *odds* asociated with a unit increase in WBC?