Logistic regression

Ciaran Evans

Last time: 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)

Research goal: Predict dengue status using diagnostic measurements

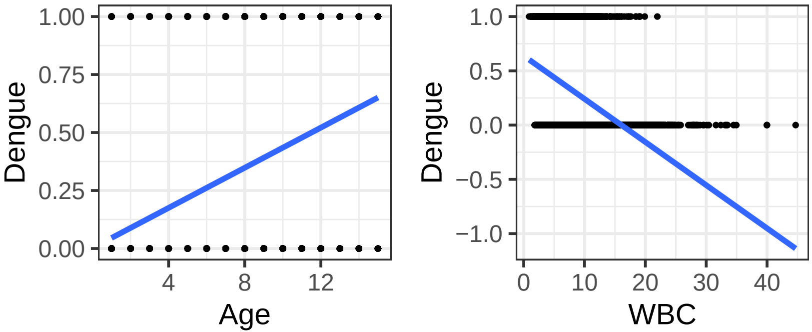
Last time: initial attempt

What if we try a linear regression model?

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

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

Don't fit linear regression with a binary response



Last time: rewriting the linear regression model

$$
\begin{aligned}
Y_i | WBC_i &\sim N(\mu_i, \sigma^2) && \text{(random component)} \\
\mu_i &= \beta_0 + \beta_1 WBC_i && \text{(systematic component)}
\end{aligned}
$$

Second attempt

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

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

Are there still any potential issues with this approach?

Fixing the issue: logistic regression

$$Y_i | WBC_i \sim Bernoulli(p_i)$$

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

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

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

Logistic regression model

$$Y_i | WBC_i \sim Bernoulli(p_i)$$

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

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

$$Y_i | WBC_i \ \sim \ Bernoulli(p_i)$$

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

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

Fitting the logistic regression model

$$Y_i | WBC_i \sim Bernoulli(p_i)$$

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

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

```
##
## Call:
## glm(formula = Dengue ~ WBC, family = binomial, data = de
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.73743    0.08499   20.44   <2e-16 ***
## WBC         -0.36085    0.01243  -29.03   <2e-16 ***
## ---
```

Activity

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

Activity

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

What is the predicted *odds* of dengue for a patient with a white blood cell count of 15?

Activity

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

What is the predicted *probability* of dengue for a patient with a WBC of 15?

Interpretation: Activity 2

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

Interpretation

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

Are patients with a higher WBC more or less likely to have dengue?

Interpretation

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

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

Interpretation

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

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

Coefficient interpretation

$$\log\left(\frac{\widehat{p}_i}{1-\widehat{p}_i}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

Fitting a *logistic* regression model?

Linear regression: minimize $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2$

Question: Should we minimize a similar sum of squares for a *logistic* regression model?

Motivation: likelihoods and estimation

Let $Y \sim \textit{Bernoulli}(p)$ be a Bernoulli random variable, with $p \in [0, 1]$. We observe 5 independent samples from this distribution:

$$Y_1 = 1, \; Y_2 = 1, \; Y_3 = 0, \; Y_4 = 0, \; Y_5 = 1$$

The true value of p is unknown, so two friends propose different guesses for the value of p: 0.3 and 0.7. Which do you think is a "better" guess?

Likelihood

Definition: Let $\mathbf{y} = (Y_1, ..., Y_n)$ be a sample of n observations, and let $f(\mathbf{y}|\theta)$ denote the joint pdf or pmf of $\mathbf{y}$, with parameter(s) θ. The *likelihood function* is

$$L(\theta|\mathbf{y}) = f(\mathbf{y}|\theta)$$

Example: Bernoulli data