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 ($0 = \text{no}$, $1 = \text{yes}$)

Research goal: Predict dengue status using diagnostic measurements

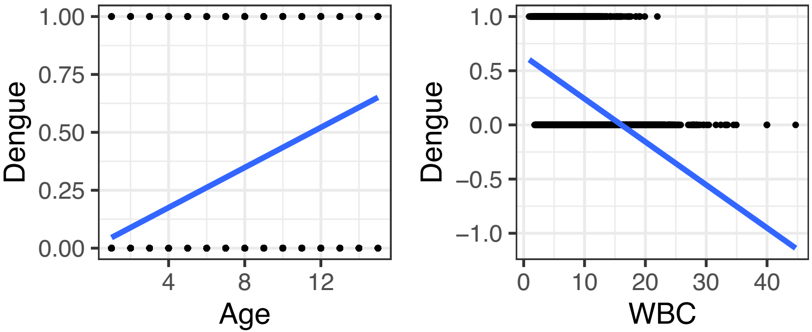
Last time: initial attempt

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

describe distribution of $Y_i | WBC_i$

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

specifies how parameters of the distribution depend on WBC

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?

<u>Problem</u>: $p_i \in [0, 1]$

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

(potentially)

Fixing the issue: logistic regression

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

link function $\rightarrow$

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

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

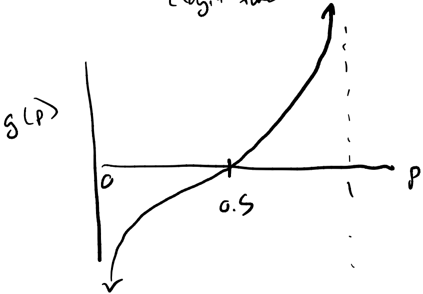
$p_i \in (0,1)$

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

(logit function)

$\frac{p_i}{1-p_i} = \underline{\text{odds}}$ $\in (0, \infty)$

$\log\left(\frac{p_i}{1-p_i}\right) =$ log odds $\in (-\infty, \infty)$



Logistic regression model

$$Y_i | WBC_i \sim Bernoulli(p_i)$$ ← capturing randomness about Y_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.

Linear model:

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

$\Longleftrightarrow$

$Y_i | X_i \sim N(\mu_i, \sigma^2)$

$\mu_i = \beta_0 + \beta_1 X_i$

Fitting the logistic regression model

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

"generalized linear model"

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

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

family specifies distribution of the response

Fitting the logistic regression model

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

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

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

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

z-values instead of t-values

```
##
## 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?

$$\widehat{\log\ odds} = 1.737 - 0.361(15)$$

$$= -3.678$$

$$\widehat{odds} = e^{-3.678} = 0.0253$$

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?

$$\text{odds} = \frac{p}{1-p} \quad \Leftrightarrow \quad p = \frac{\text{odds}}{1+\text{odds}}$$

$$= \frac{e^{\log \text{odds}}}{1+e^{\log \text{odds}}}$$

$$\hat{p}_i = \frac{e^{-3.678}}{1+e^{-3.678}} \approx 0.025$$

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?

less likely (negative slope)

$\hat{\beta}_1 < 0$

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?

-0.361

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?

$$\frac{\text{odds when } WBC = x+1}{\text{odds when } WBC = x} = \frac{e^{1.737 - 0.361(x+1)}}{e^{1.737 - 0.361x}}$$

$$= e^{-0.361}$$

$$= 0.697$$

A one unit increase in WBC is associated with a change in the odds of dengue by a <u>factor</u> of 0.697 (30.3% decrease)

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?