Lecture 2: Fitting and interpreting logistic regression models

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)

Logistic regression model

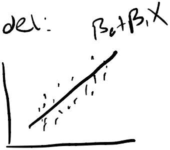
$p_i = \mathbb{E}[Y_i \mid X_i]$

$$Y_i \sim Bernoulli(p_i) \qquad \text{(random component)}$$

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

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.

Two options for writing linear regression model:



Option 1: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

captures individual variability

Option 2: If we assume $\varepsilon_i \sim N(0, \sigma^2)$, then

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

$\mu_i = \beta_0 + \beta_1 X_i$ (systematic)

no ε_i here b/c random component captures randomness

Fitting the logistic regression model

$$Y_i \sim Bernoulli(p_i)$$

"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)
```

formula (in R)

family specifies distribution of response variable

logistic regression: family = binomial

linear regression: family = gaussian

poisson " " : " = poisson

Fitting the logistic regression model

$$Y_i \sim Bernoulli(p_i)$$

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

```
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 ***
##
```

Making predictions

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

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

- What is the predicted odds of dengue for a patient with a WBC of 10?
- For a patient with a WBC of 10, is the predicted probability of dengue > 0.5, < 0.5, or $= 0.5$?
- What is the predicted *probability* of dengue for a patient with a WBC of 10?

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

$WBC = 10$

$$\text{log odds} = 1.737 - 0.361(10)$$
$$= -1.873$$

$$\text{odds} = e^{-1.873} = \underbrace{0.154}_{<1}$$

if $p = 0.5 \iff \text{odds} = 1 \iff \text{log odds} = 0$

$p < 0.5 \iff \text{odds} < 1$ $\Rightarrow$ probability < 0.5

$p > 0.5 \iff \text{odds} > 1$

$$\hat{p} = \frac{e^{-1.873}}{1 + e^{-1.873}} \approx 0.133$$

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

$$p = \frac{\text{odds}}{1 + \text{odds}}$$

$$e^{\beta_0+\beta_1 x} = \frac{e^{\text{log odds}}}{1 + e^{\text{log odds}}}$$

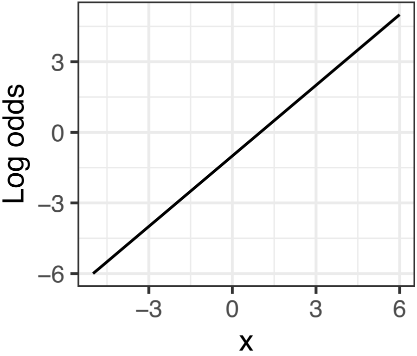
$$p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

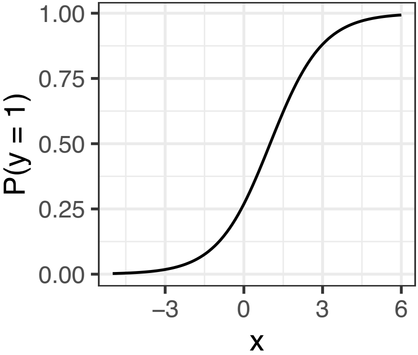
Shape of the regression curve

$Y_i \sim Bernoulli(p_i)$

$= P(Y_i = 1)$

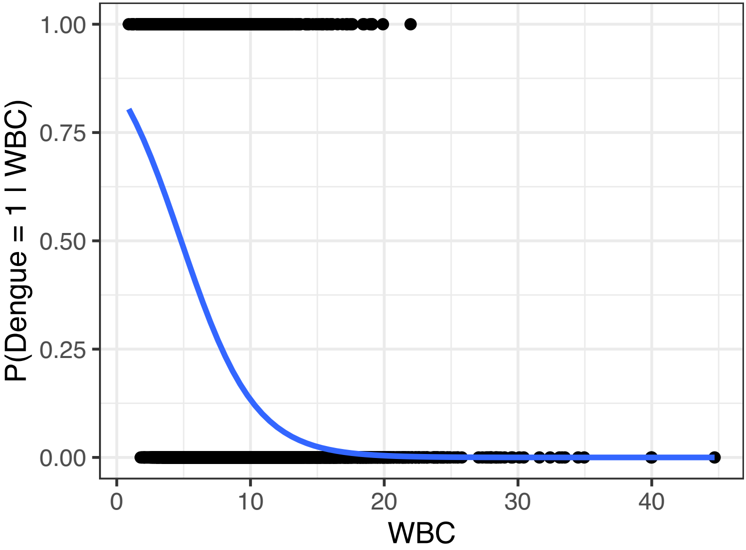
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \ X_i \qquad p_i = \frac{e^{\beta_0+\beta_1 \ X_i}}{1+e^{\beta_0+\beta_1 \ X_i}}$$





$(\beta_1 > 0)$

Plotting the fitted model for dengue data



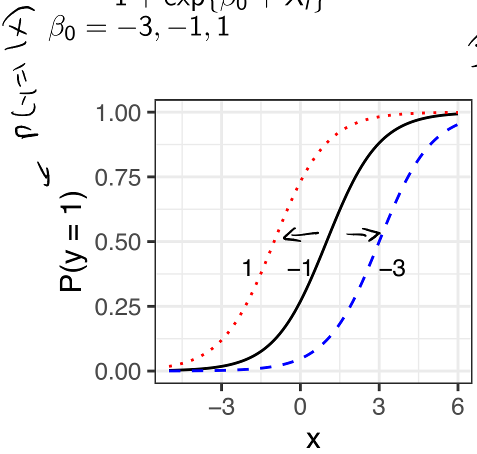
Shape of the regression curve

$\beta_1 = -1$

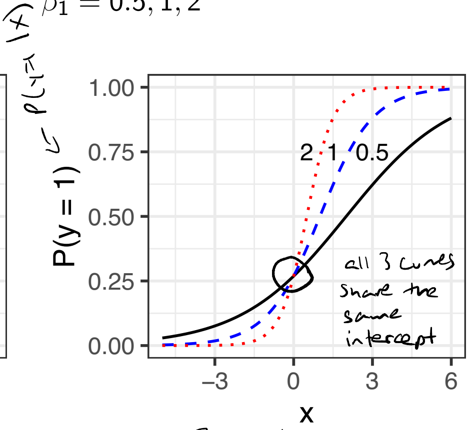
How does the shape of the fitted logistic regression depend on β_0 and β_1?

$$p_i = \frac{\exp\{\beta_0 + X_i\}}{1 + \exp\{\beta_0 + X_i\}} \text{ for } \beta_0 = -3, -1, 1$$

$$p_i = \frac{\exp\{-1 + \beta_1 \ X_i\}}{1 + \exp\{-1 + \beta_1 \ X_i\}} \text{ for } \beta_1 = 0.5, 1, 2$$



$\beta_1 = 1$



$\beta_0 = -1$

Interpretation

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

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

- Are patients with a higher WBC more or less likely to have dengue?
- What is the change in *log odds* associated with a unit increase in WBC?
- What is the change in *odds* asociated with a unit increase in WBC?

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

$$WBC \to WBC + 1$$

$$\log \text{odds}\,(WBC+1) = \cancel{1.737} - 0.361\,(\cancel{WBC} + 1)$$

$$-\ \log \text{odds}\,(WBC) = \cancel{1.737} - 0.361\,(\cancel{WBC})$$

$$= -0.361 \qquad \text{(log odds decreases by } 0.361)$$

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

a one-unit increase in WBC is associated with a change in odds of dengue by a <u>factor</u> of $e^{-0.361} \approx 0.7$