Fitting and interpreting logistic regression models

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

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

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

linear model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

captures individual variability

Logistic regression model

$$Y_i \sim Bernoulli(p_i)$$

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

> 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.

$Y_i \sim N(\mu_i, \sigma_\varepsilon^2)$

$\mu_i = \beta_0 + \beta_1 X_i$

No ε_i!

Random component captures the 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)
```

specifies distribution of response

linear regression: family = gaussian

Fitting the logistic regression model

GLMS

Systematic component:

$g(E[Y]) = \ldots\ldots$

$\beta_0 + \beta_1 x_i$

$E[Y_i] = p_i$

$$Y_i \sim Bernoulli(p_i)$$

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

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

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

(not t) (not t)

```
...
## 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 ***
## ---
...
```

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

Making predictions

= natural log (ln) not $\log_{10}$ not $\log_2$

$$Y_i \sim Bernoulli(p_i)$$

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

Work in groups of 2-3 for 5 minutes 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?

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

$WBC = 10$

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

$$\text{odds} = e^{-1.873} = 0.154$$

$\Rightarrow$ probability < 0.5

$p < 0.5 \Rightarrow$ odds < 1, log odds < 0

$p > 0.5 \Rightarrow$ odds > 1, log odds > 0

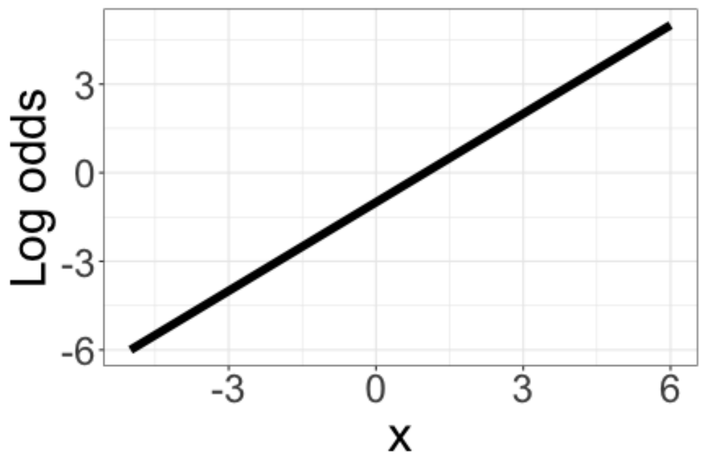
$p = 0.5 \Rightarrow$ odds $= 1$, log odds $= 0$

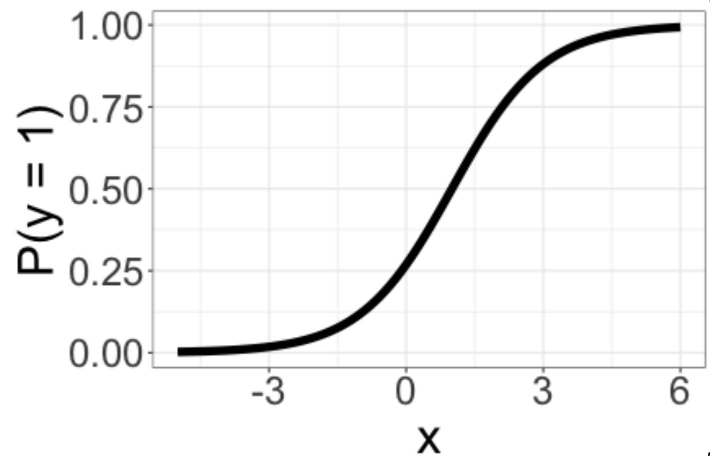
$$\hat{p}_i = e^{-1.873}(1-\hat{p}_i) \Rightarrow \hat{p}_i = \frac{e^{-1.873}}{1+e^{-1.873}} = 0.133$$

Shape of the regression curve

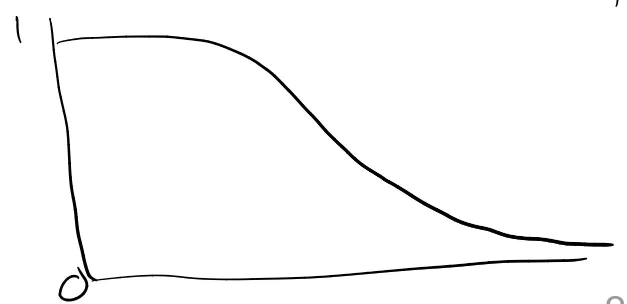
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1\, X_i \qquad p_i = \frac{e^{\beta_0+\beta_1\, X_i}}{1+e^{\beta_0+\beta_1\, X_i}}$$

$\beta_1 > 0$

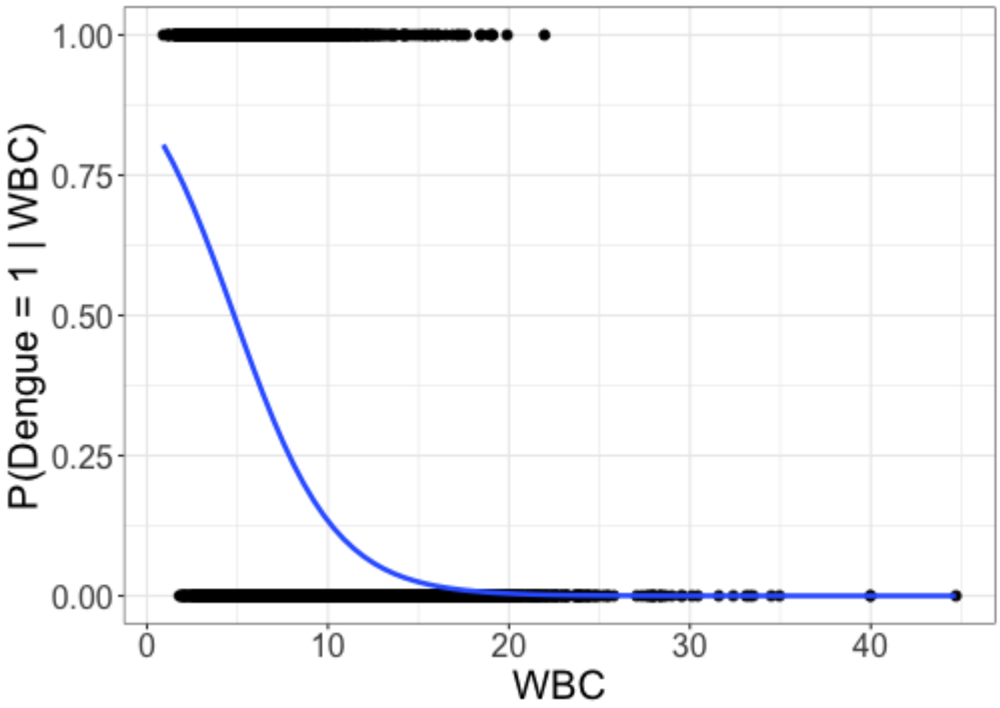




$\beta_1 < 0$



Plotting the fitted model for dengue data



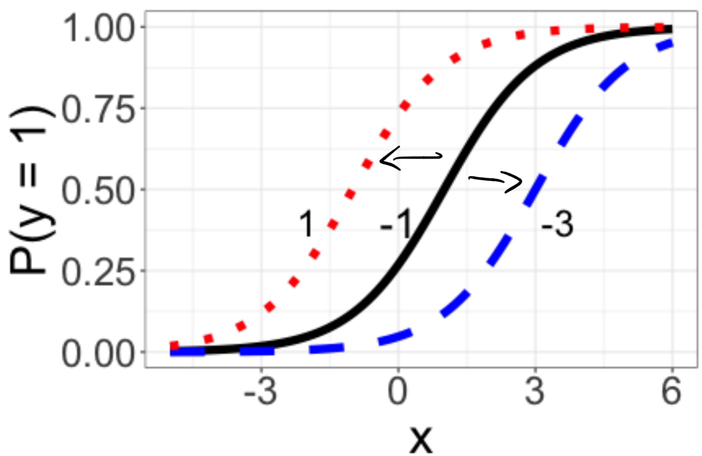
Shape of the regression curve

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

$\beta_1 = 1$

$$p_i = \frac{\exp\{\beta_0 + X_i\}}{1 + \exp\{\beta_0 + X_i\}}$$

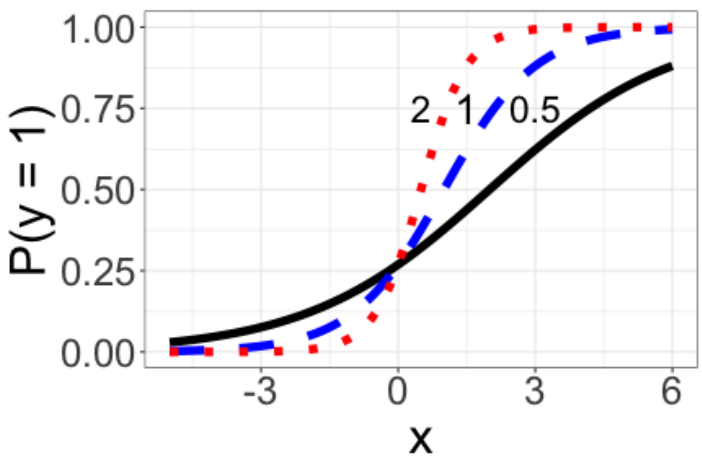
for $\beta_0 = -3, -1, 1$



$\beta_0 = -1$

$$p_i = \frac{\exp\{-1 + \beta_1 X_i\}}{1 + \exp\{-1 + \beta_1 X_i\}}$$

for $\beta_1 = 0.5, 1, 2$



Interpretation

$$Y_i \sim Bernoulli(p_i)$$

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

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

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

- log odds: a one unit increase in WBC is associated w/ a decrease of 0.361 in log odds of dengue

- odds: $e^{-0.361} = 0.7$

  a one unit increase in WBC is associated w/ a decrease in the odds of dengue by a factor of 0.7

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

$$= 0.7$$

Recap: ways of fitting a *linear* regression model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_k X_{i,k} + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

Suppose we observe data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, where $X_i = (1, X_{i,1}, \ldots, X_{i,k})^T$.

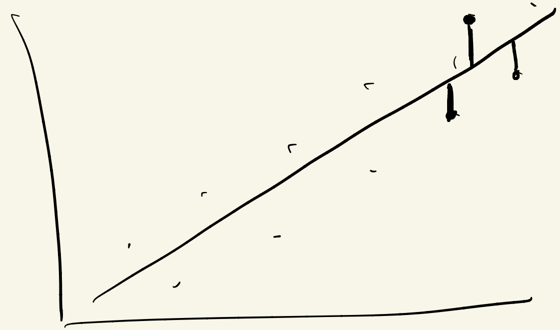
How do we fit this linear regression model? That is, how do we estimate

- Minimize SSE
- Projection
- Maximize likelihood

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

Minimize SSE

$$SSE = \hat{\sum_{i=1}} \underbrace{(Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k X_{ik})^2}_{\text{squared errors}}$$



Minimize:

$$\frac{\partial SSE}{\partial \beta_0} \overset{set}{=} 0$$

$$\frac{\partial SSE}{\partial \beta_1} \overset{set}{=} 0$$

$$\vdots$$

$$\frac{\partial SSE}{\partial \beta_k} \overset{set}{=} 0$$

k+1 equations

k+1 unknowns

choose $\hat{\beta}_0, \ldots, \hat{\beta}_k$ to

solve system

Projection

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots \\ 1 & x_{21} & \cdots \\ \vdots & \vdots & \\ 1 & x_{n1} & \cdots \end{pmatrix} \in \mathbb{R}^{n \times (k+1)}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$\hat{Y} = X\hat{\beta}$$

want $\hat{Y}$ "close" to Y

$$\|Y - \hat{Y}\| = \sqrt{SSE}$$

$\Leftrightarrow$ minimize SSE

Summary: three ways of fitting linear regression models

- Minimize SSE, via derivatives of $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i,1} - \cdots - \beta_k X_{i,k})^2$
- Minimize $||Y - \widehat{Y}||$ (equivalent to minimizing SSE)
- Maximize likelihood (for *normal* data, equivalent to minimizing SSE)

> Which of these three methods, if any, is appropriate for fitting a logistic regression model? Do any changes need to be made for the logistic regression setting?