Fitting logistic regression models

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)

Last time: Logistic regression model

$$Y_i = \text{dengue status } (0 = \text{negative}, 1 = \text{positive})$$

$$Y_i \sim Bernoulli(p_i)$$

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

We get n observations $(WBC_1, Y_1), \ldots, (WBC_n, Y_n)$. Want estimates $\widehat{\beta}_0, \widehat{\beta}_1$

Last time: Logistic regression model

$Y_i = \text{dengue status } (0 = \text{no}, 1 = \text{yes}) \quad Y_i \sim Bernoulli(p_i)$

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

How should we interpret the slope -0.361?

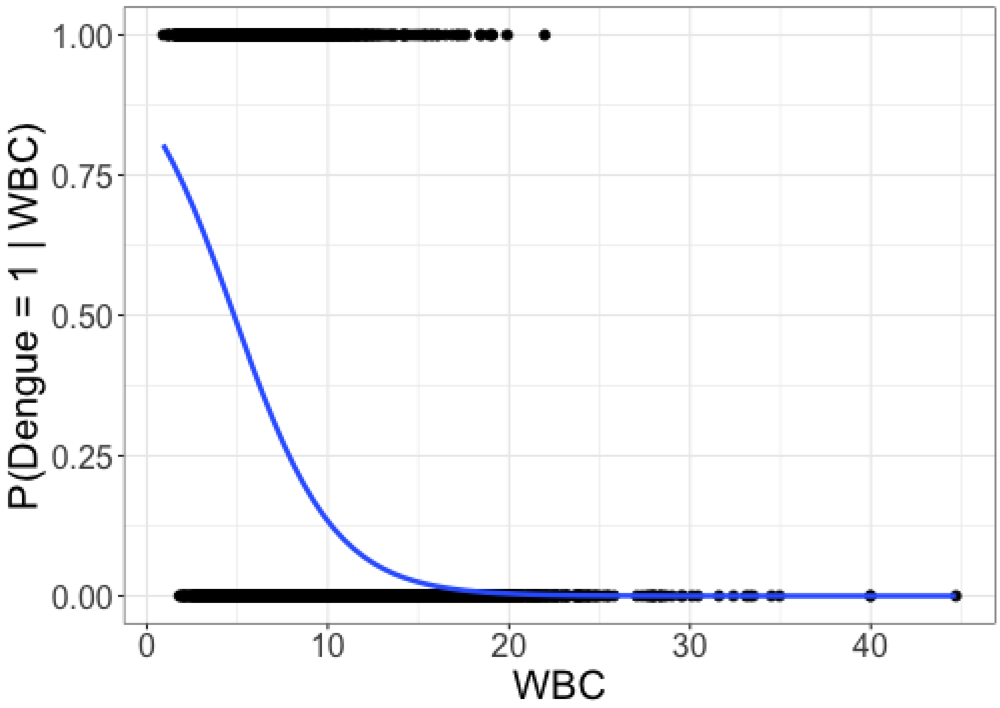
Getting probabilities

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

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

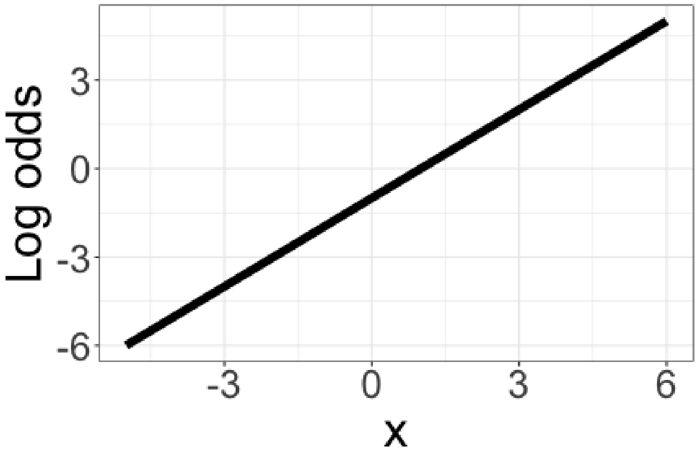
> How do I calculate estimated probabilities $\hat{p}_i$?

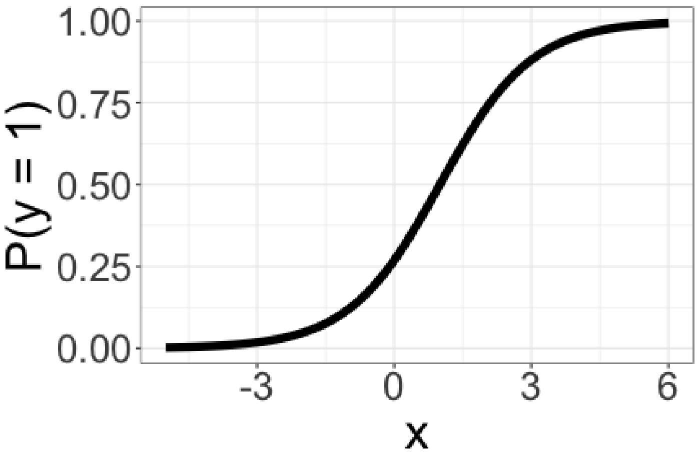
Plotting the fitted model for dengue data



Shape of the regression curve

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



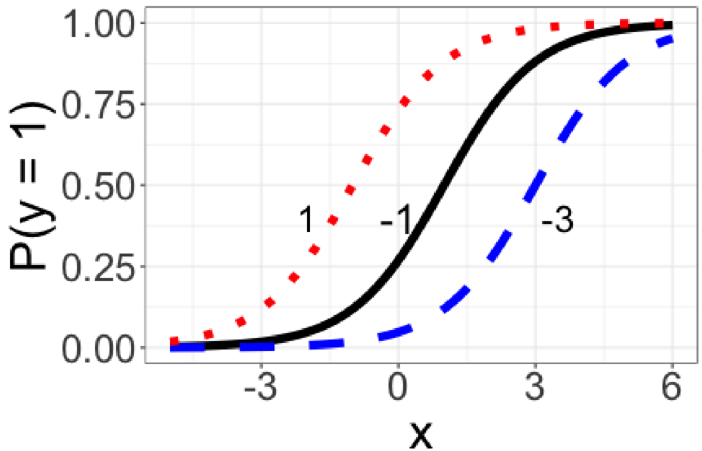


Shape of the regression curve

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

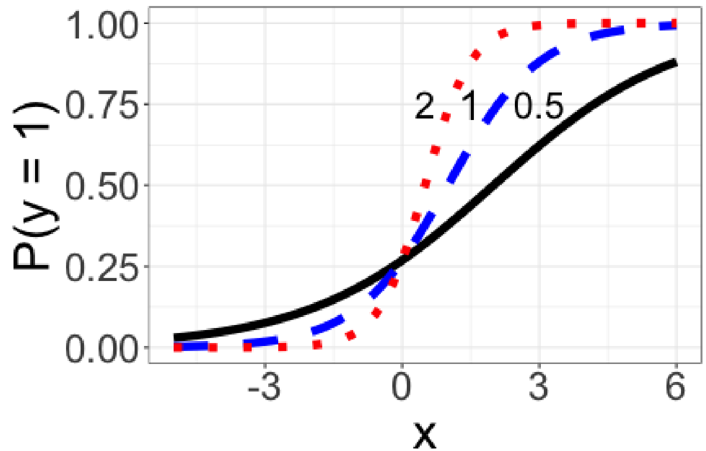
$$p_i = \frac{\exp\{\beta_0 + X_i\}}{1 + \exp\{\beta_0 + X_i\}}$$

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



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

for $\beta_1 = 0.5, 1, 2$



Fitting logistic regression in R

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

```
...
##             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 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6955.8  on 5719  degrees of freedom
## Residual deviance: 5529.8  on 5718  degrees of freedom
## AIC: 5533.8
##
## Number of Fisher Scoring iterations: 5
```

Recap: ways of fitting a *linear* regression model

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

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

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

Discuss with your neighbor for 2--3 minutes.

Method 1: Minimize SSE

Method 2: Projection argument

Method 3: Maximizing likelihood

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 $||\widehat{Y}||$ (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?
>
> Discuss with your neighbor for 2--3 minutes.

Maximum likelihood for logistic regression

$$Y_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}$$

Suppose we observe independent samples $(X_1, Y_1), \ldots, (X_n, Y_n)$. Write down the likelihood function

$$L(\beta) = \prod_{i=1}^{n} f(Y_i; \beta)$$

for the logistic regression problem. Take 2--3 minutes, then we will discuss as a class.

Maximum likelihood for logistic regression

$L(\beta) =$

> I want to choose β to maximize $L(\beta)$. What are the usual steps to take?

Initial attempt at maximizing likelihood

$$L(\beta) = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i}$$

$$\ell(\beta) =$$

Iterative methods for maximizing likelihood

Fisher scoring

Fisher scoring for logistic regression