Lecture 2: 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

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 \, \text{WBC}_i$$

Logistic regression model

$$Y_i \sim \text{Bernoulli}(p_i)$$

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

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

$$Y_i \sim \text{Bernoulli}(p_i)$$

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

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

Fitting the logistic regression model

$$Y_i \sim \text{Bernoulli}(p_i)$$

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

```
Call:
glm(formula = Dengue ~ WBC, family = binomial, data = dengue)

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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6955.8  on 5719  degrees of freedom
```

Making predictions

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

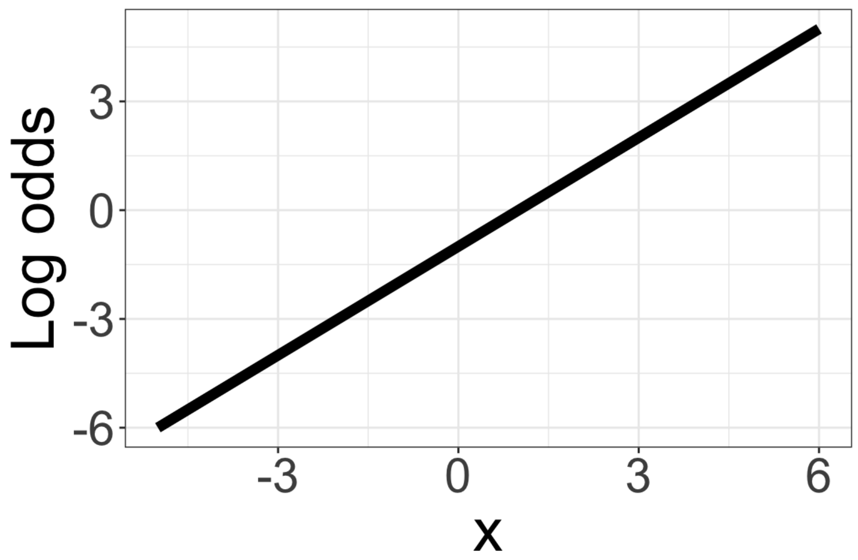
Work in groups of 2-3 on the following questions:

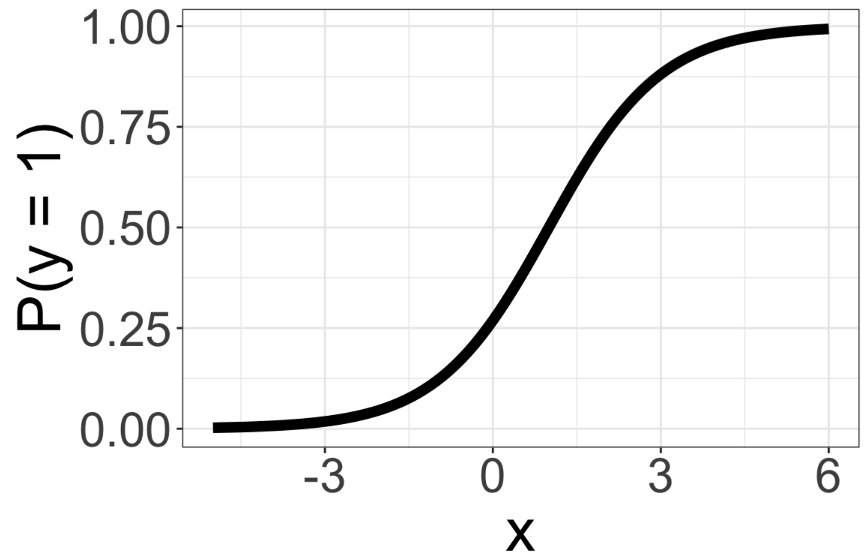
- What is the predicted odds of dengue for a patient with a WBC of 10?
- For a patient with a WBC of 10, is the predicted probability of dengue > 0.5, < 0.5, or $= 0.5$?
- What is the predicted *probability* of dengue for a patient with a WBC of 10?

Shape of the regression curve

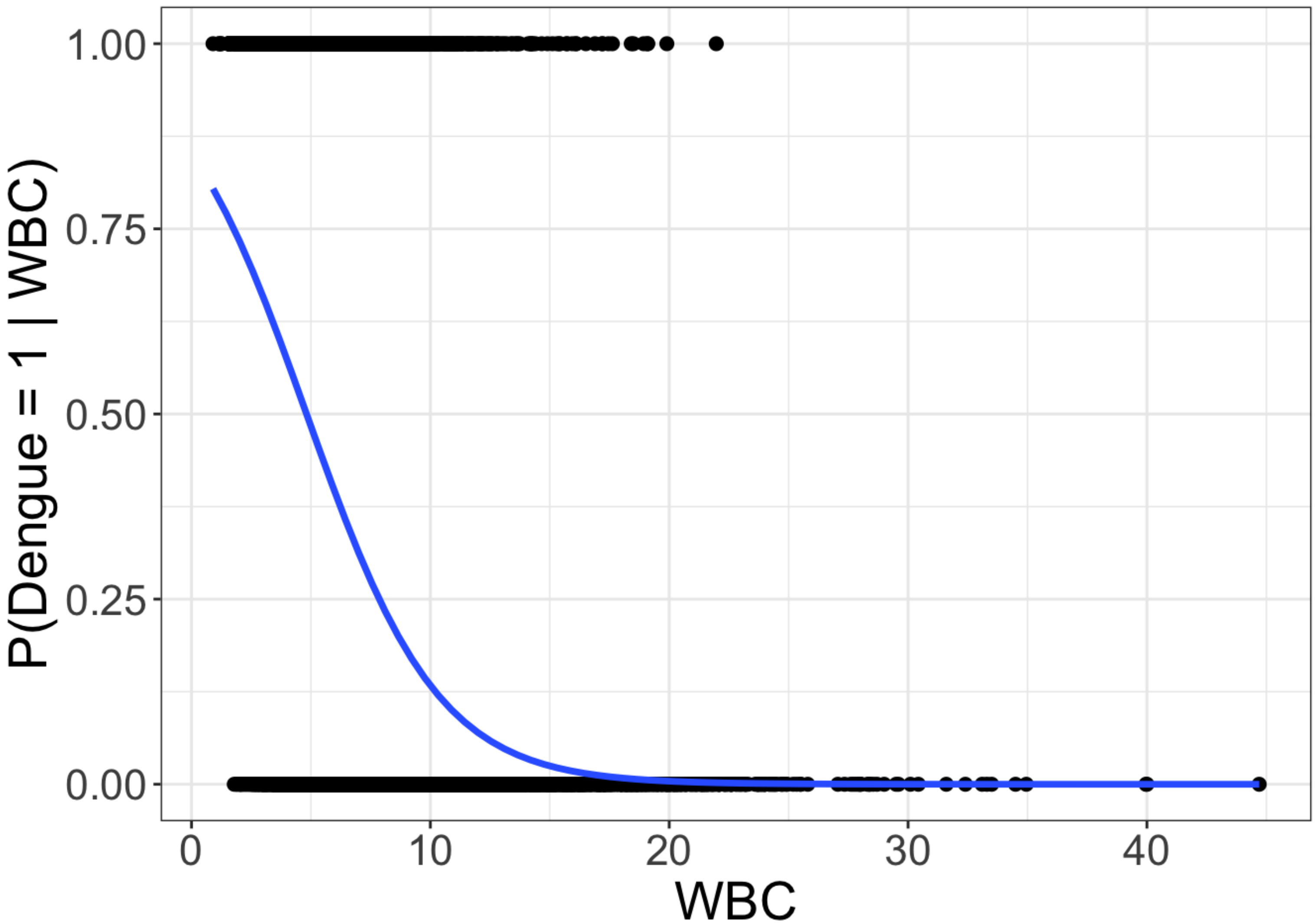
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$$

$$p_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$$





Plotting the fitted model for dengue data

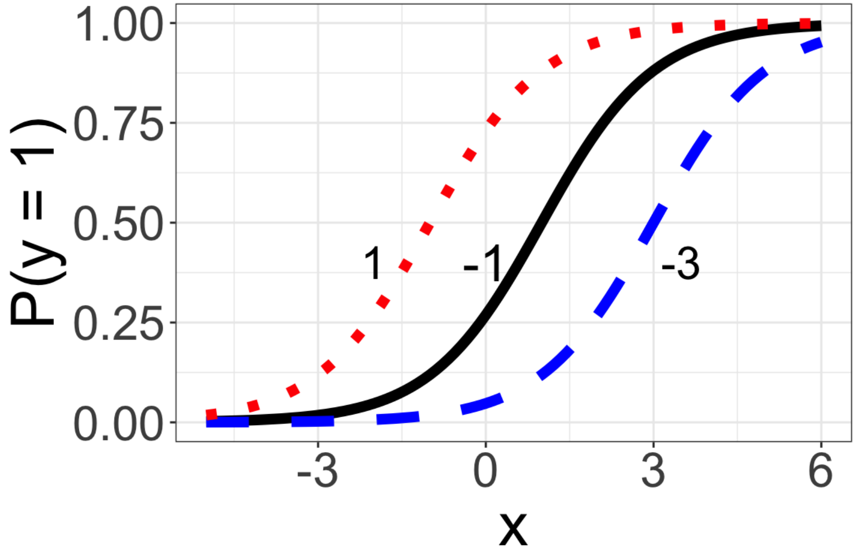


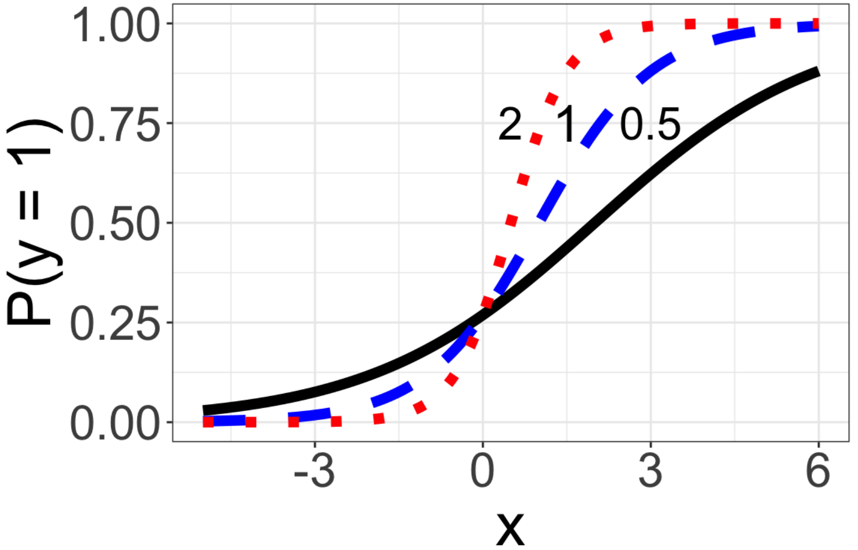
Shape of the regression curve

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

$$p_i = \frac{\exp\{\beta_0 + X_i\}}{1 + \exp\{\beta_0 + X_i\}} \quad \text{for} \quad p_i = \frac{\exp\{-1 + \beta_1 X_i\}}{1 + \exp\{-1 + \beta_1 X_i\}}$$

$\beta_0 = -3, -1, 1$ for $\beta_1 = 0.5, 1, 2$





Interpretation

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

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?

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

$$\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$$

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?