Fitting logistic regression models

where did the ε_i go?

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

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

random   random

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma_\varepsilon^2)$$

Y_i is <u>not</u> iid b/c μ_i (or p_i) changes, so not identically distributed

$$\Downarrow$$

$$\boxed{\begin{aligned} Y_i | X_i &\sim N(\mu_i, \sigma_\varepsilon^2) \\ \mu_i &= \beta_0 + \beta_1 X_i \end{aligned}}$$

(random component)

(systematic componet)

not random

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

$$\underbrace{\log\left(\frac{p_i}{1-p_i}\right)}_{\in(-\infty,\infty)} = \underbrace{\beta_0 + \beta_1 X_i}_{\in(-\infty,\infty)}$$ (systematic component)

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

$$Y_i = p_i + \varepsilon_i$$

$$\varepsilon_i = \begin{cases} 1 - p_i & \text{w/ prob } p_i \\ -p_i & \text{w/ prob } 1 - p_i \end{cases}$$

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

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

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

link function

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

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

1 unit increase in WBC is associated w/ a change in odds of dengue by a factor of 0.697

Last time: Logistic regression model

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

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

How should we interpret the slope -0.361?

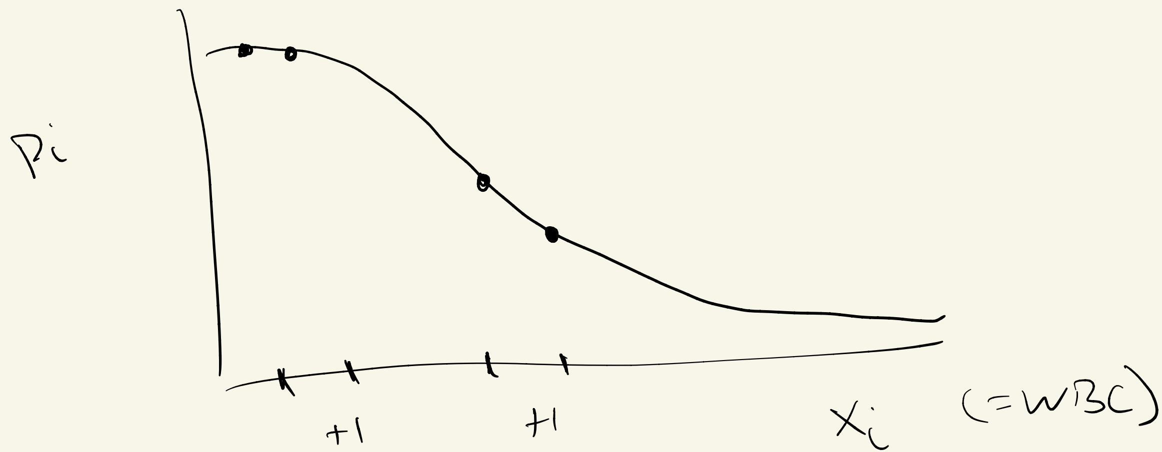
log odds: 1 unit increase in WBC is associated w/ a decrease of 0.361 in log odds of dengue

odds: $\text{odds} = \exp\{1.737 - 0.361\, WBC_i\}$

$WBC \to WBC + 1$

$\exp\{1.737 - 0.361\, WBC\} \to \exp\{1.737 - 0.361\, WBC - 0.361\}$

$$\text{odds ratio} = \frac{\cancel{e^{1.737-0.361 WBC}}\; e^{-0.361}}{\cancel{e^{1.737-0.361 WBC}}} = e^{-0.361} \approx 0.697$$



p_i
+1
+1
x_i
(=WBC)

$\hat{p}_i \in [0, 1]$

Getting probabilities

$Y_i =$ 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 do I calculate estimated probabilities $\hat{p}_i$?

$$\frac{\hat{p}_i}{1-\hat{p}_i} = \exp\{1.737 - 0.361\, WBC_i\}$$

$$\hat{p}_i = e^{1.737 - 0.361\, WBC_i}(1-\hat{p}_i)$$

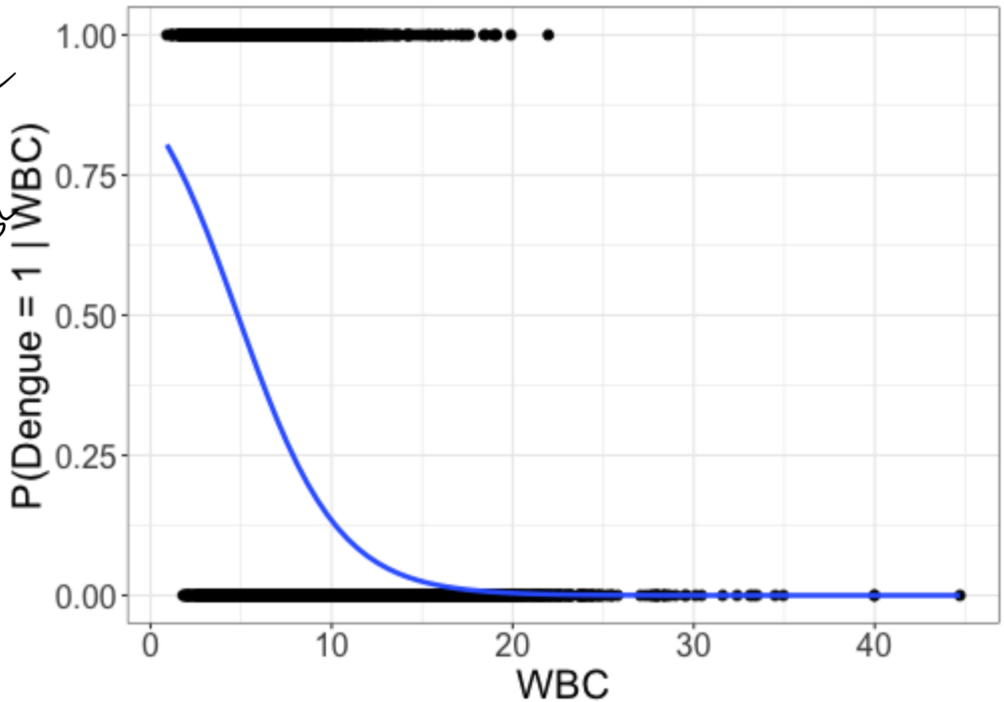
$$\hat{p}_i(1 + e^{1.737 - 0.361\, WBC_i}) = e^{1.737 - 0.361\, WBC_i}$$

$$\Rightarrow \quad \hat{p}_i = \frac{e^{1.737 - 0.361\, WBC_i}}{1 + e^{1.737 - 0.361\, WBC_i}} = \frac{\text{odds}}{1 + \text{odds}}$$

curve!

Plotting the fitted model for dengue data

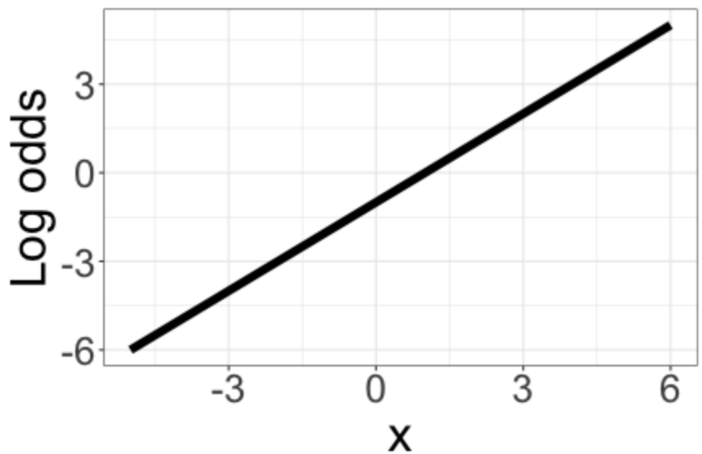
$$\frac{e^{1.737-0.361WBC}}{1+e^{1.737-0.361WBC}}$$

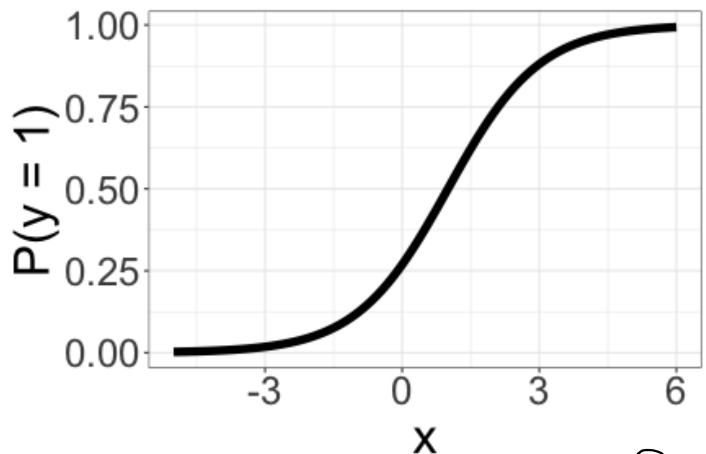


Shape of the regression curve

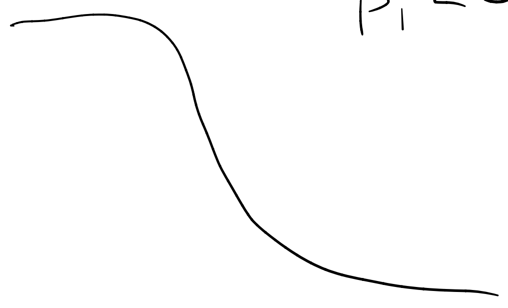
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \; X_i \qquad p_i = \frac{e^{\beta_0+\beta_1 \; X_i}}{1+e^{\beta_0+\beta_1 \; X_i}}$$

$\beta_1 > 0$





$\beta_1 < 0$



Shape of the regression curve

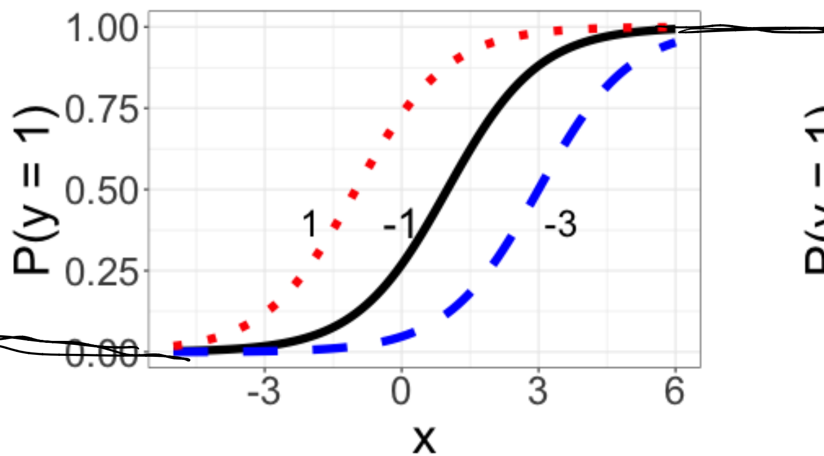
How does the shape of the fitted logistic regression depend on β_0 and β_1?

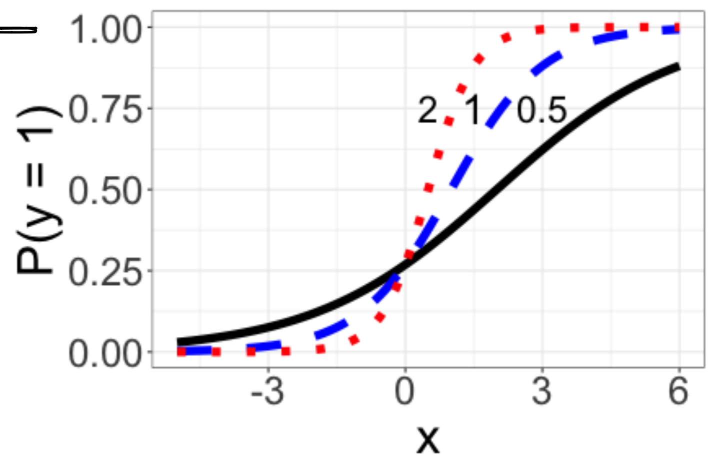
$$p_i = \frac{\exp\{\beta_0 + X_i\}}{1 + \exp\{\beta_0 + X_i\}}$$

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

vs. lm

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

distribution of response

not t!

```
...
##             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
```

(ignore for now)

(instead of R^2)

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.

1) Minimize SSE
2) Projection $\left(\Rightarrow (X^T X)^{-1} X^T Y \right)$
3) MLE

Method 1: Minimize SSE

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

$$\left.\begin{aligned} \frac{\partial SSE}{\partial \beta_0} &\overset{\text{set}}{=} 0 \\ \frac{\partial SSE}{\partial \beta_1} &\overset{\text{set}}{=} 0 \\ &\vdots \\ \frac{\partial SSE}{\partial \beta_k} &\overset{\text{set}}{=} 0 \end{aligned}\right\}$$

$k+1$ equations

$k+1$ unknowns

$\Rightarrow$ solve

Method 2: Projection argument

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R} \qquad X = \begin{pmatrix} 1 & X_{11} & \cdots \\ 1 & X_{21} & \cdots \\ \vdots & \vdots & \\ 1 & X_{n1} & \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \hat{Y} = X\hat{\beta}$$

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

$\Rightarrow \|Y - \hat{Y}\|$ to be small

$$\|Y - \hat{Y}\| = \left(\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_k X_{ik})^2 \right)^{1/2}$$

$\Longleftrightarrow$ minimizing SSE!

Method 3: Maximizing likelihood

$$Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}, \sigma_\varepsilon^2)$$

(independence)

$$L(\beta_0, \ldots, \beta_k, \sigma_\varepsilon^2) = \prod_{i=1}^{n} f(Y_i; \beta_0, \beta_1, \ldots, \beta_k, \sigma_\varepsilon^2)$$

$$= (2\pi)^{-\frac{n}{2}} \sigma_\varepsilon^{-n} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \underbrace{\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2}_{\text{SSE!}} \right\}$$

maximizing likelihood $\Leftrightarrow$ minimizing SSE (for normal data)

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 $(X_1, Y_1), \dots, (X_n, Y_n)$. 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