Parametric models and logistic regression

This week

- Office hour schedule posted
- HW 1 released on course website
- Download R and RStudio
  - Instructions on course website
  - Please come to office hours or contact me if you have problems!

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)

Goal: Build a model to predict dengue status

Parametric modeling

A regression model is an example of a more general process called **parametric modeling**

- **Step 1:** Choose a reasonable distribution for Y_i ← linear regression: $Y_i \sim N(\mu_i, \sigma_\varepsilon^2)$
- **Step 2:** Build a model for the parameters of interest
- **Step 3:** Fit the model ↑ linear regression: $\mu_i = \beta_0 + \beta_1 x_i$

Step 1: Bernoulli distribution

Definition: Let Y_i be a binary random variable, and $\pi_i = P(Y_i = 1)$. Then $Y_i \sim Bernoulli(\pi_i)$.

$1 - \pi_i = P(Y_i = 0)$

A **random variable** is an event that has a set of possible outcomes, but we don't know which one will occur

- Here $Y_i = 0$ or 1
- Our goal is to use the observed data to estimate $\pi_i = P(Y_i = 1)$

Linear:

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

$\mu_i = \beta_0 + \beta_1 x_i$

Binary response:

$Y_i \sim Bernoulli(\pi_i)$

Second attempt at a model

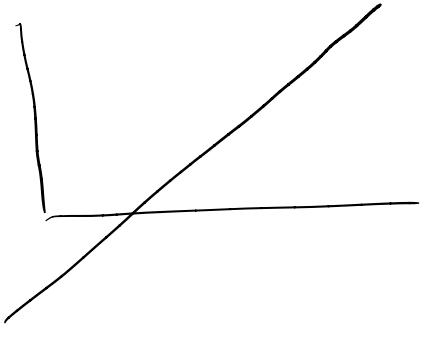
$$Y_i \sim Bernoulli(\pi_i) \quad \pi_i = P(Y_i = 1 | Age_i)$$

$$\pi_i = \beta_0 + \beta_1 Age_i$$

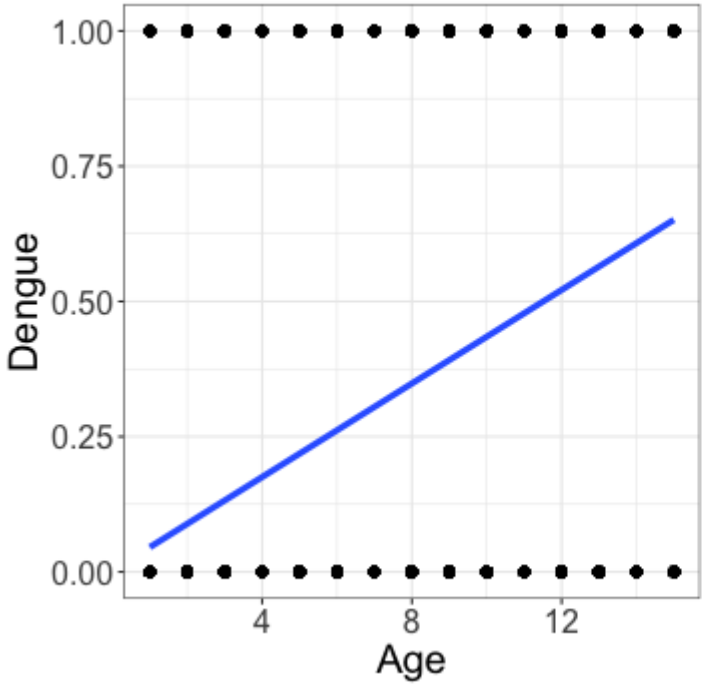
> Are there still any potential issues with this approach?

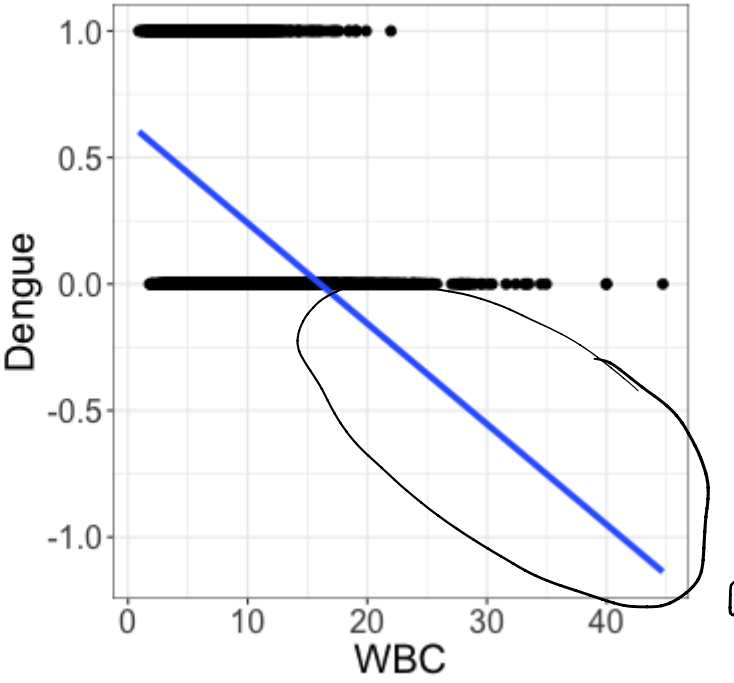
$\pi_i \in [0,1]$

but $\beta_0 + \beta_1 Age_i \in (-\infty, \infty)$ (unless $\beta_1 = 0$)



Don't fit linear regression with a binary response





$\hat{\pi}_i$ not in $[0,1]$

Fixing the issues

$$Y_i \sim \text{Bernalli}(\pi_i)$$

$$?? = \underbrace{\beta_0 + \beta_1 \text{Age}_i}_{\in (-\infty, \infty)}$$

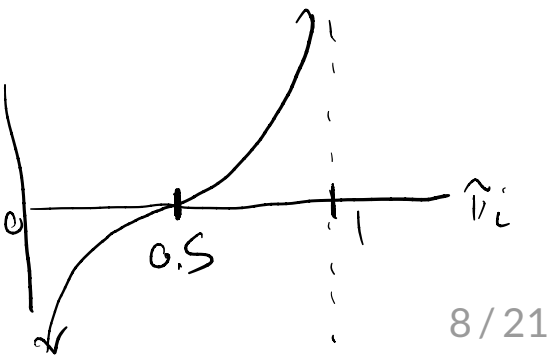
- So we need $?? \in (-\infty, \infty)$ too!
- Also, $??$ needs to be a function of π_i

$$\pi_i \in [0,1] \qquad \pi_i = 0$$

$$\frac{\pi_i}{1-\pi_i} \in [0, \infty) \qquad \pi_i = 1$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) \in (-\infty, \infty)$$

$\pi_i = 0.5 \Rightarrow \log\left(\frac{\pi_i}{1-\pi_i}\right) = 0$



Logistic regression model

$Y_i =$ dengue status (0 = negative, 1 = positive)

$Age_i =$ age (in years)

Random component: $Y_i \sim Bernoulli(\pi_i)$

Specifies distribution of Y_i

Systematic component: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \, Age_i$

relates distribution to explanatory variables

log = natural log (Ln)

log odds aka logit

not $\log_{10}$

Logistic regression model

$Y_i =$ dengue status (0 = negative, 1 = positive)

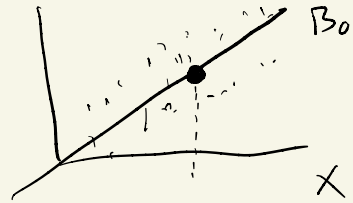
$Age_i =$ age (in years)

Random component: $Y_i \sim Bernoulli(\pi_i)$

Systematic component: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \, Age_i$

> 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.

Linear regression: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ $\quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$



ε_i captures randomness in Y_i

Alternative form:

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

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

} No ε_i! The distribution of Y_i captures randomness

Logistic regression:

(random) $Y_i \sim \text{Bernalli}(\pi_i)$

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

} No ε_i! The distribution of Y_i captures randomness

Fitting the logistic regression model

$$Y_i \sim Bernoulli(\pi_i)$$

"generalized linear model"

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

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

tells us distribution of response

for logistic regression, use binomial family

Fitting the logistic regression model

$$Y_i \sim Bernoulli(\pi_i)$$

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

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

```
...
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.454345   0.075068  -32.70   <2e-16 ***
## Age          0.217312   0.008826   24.62   <2e-16 ***
## ---
...
```

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -2.45 + 0.22\,Age_i$$

Class activity

https://sta214-s23.github.io/class_activities/ca_lecture_2.html

- Spend 5--7 minutes working in pairs on the questions
- We will then discuss as a class

Class activity

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -2.45 + 0.22\ Age_i$$

> What is the predicted odds of dengue for a 5 year old patient?

$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -2.45 + 0.22(5) = -1.35$

$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = e^{-1.35} = 0.26$

Class activity

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -2.45 + 0.22\ Age_i$$

For a 5 year old patient, is the predicted probability of dengue $> 0.5, < 0.5$, or $= 0.5$?

$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$ < 0 $\quad \hat{\pi} < 0.5$

> 0 $\quad \hat{\pi} > 0.5$

$= 0$ $\quad \hat{\pi} = 0$

$-1.35 < 0 \quad \Rightarrow \quad \hat{\pi} < 0.5$

Class activity

$$\hat{\pi}_i = \frac{e^{-2.45 + 0.22\,Age_i}}{1 + e^{-2.45 + 0.22\,Age_i}}$$

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -2.45 + 0.22\ Age_i$$

> What is the predicted *probability* of dengue for a 5 year old patient?

$\hat{\pi}_i = 0.21$

$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -1.35$

$\Rightarrow \frac{\hat{\pi}_i}{1-\hat{\pi}_i} = e^{-1.35}$ $\Rightarrow \hat{\pi}_i = (1-\hat{\pi}_i)e^{-1.35}$

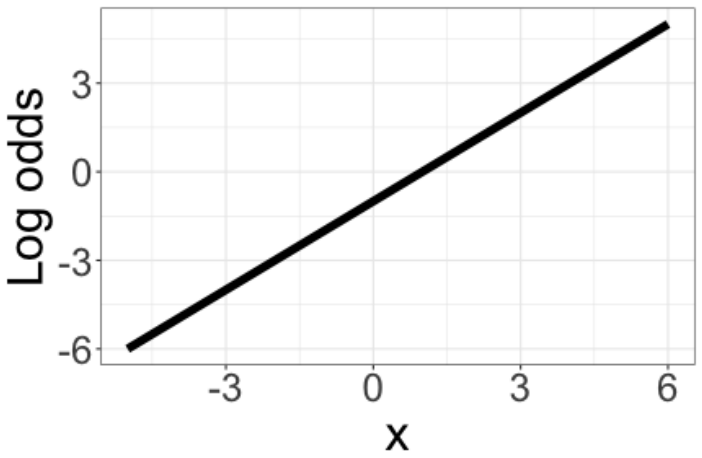
$\Rightarrow \hat{\pi}_i = e^{-1.35} - e^{-1.35}\hat{\pi}_i$ $\Rightarrow \hat{\pi}_i + e^{-1.35}\hat{\pi}_i = e^{-1.35}$

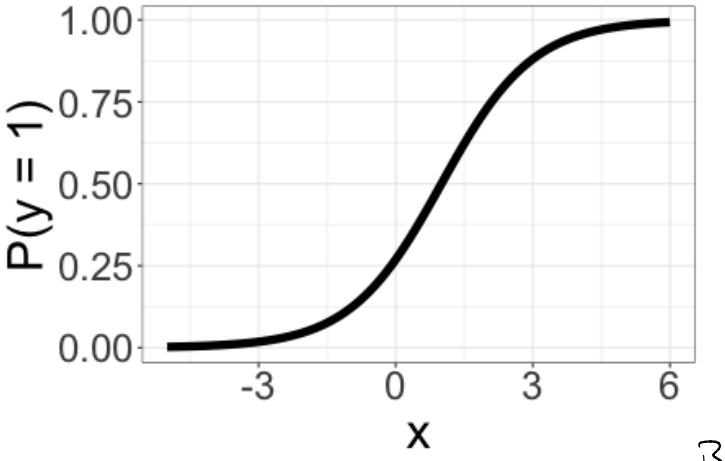
$\Rightarrow \hat{\pi}_i = \frac{e^{-1.35}}{1+e^{-1.35}} = 0.21$

Shape of the regression curve

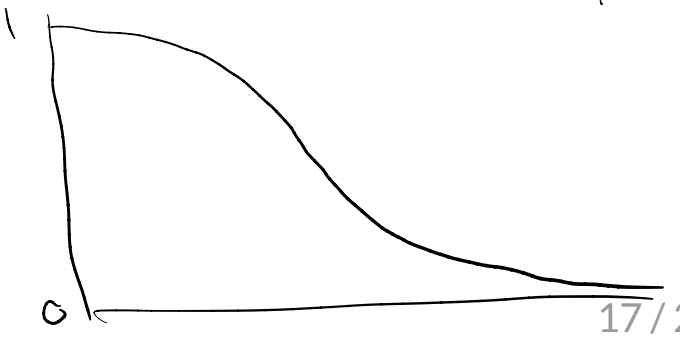
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i \qquad \pi_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$$

(B, 70)





B, 40

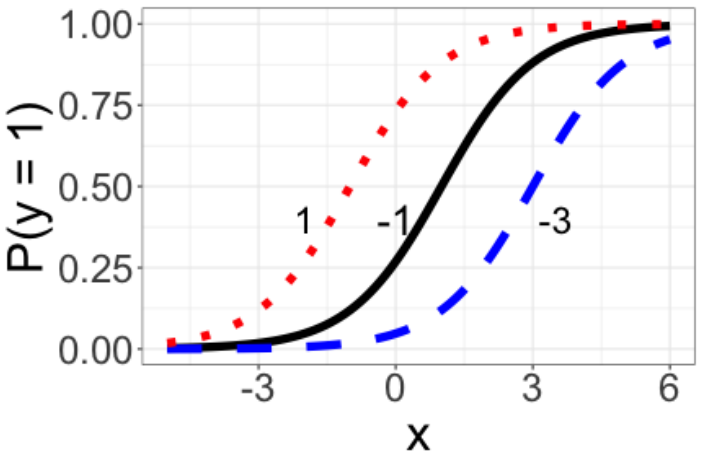


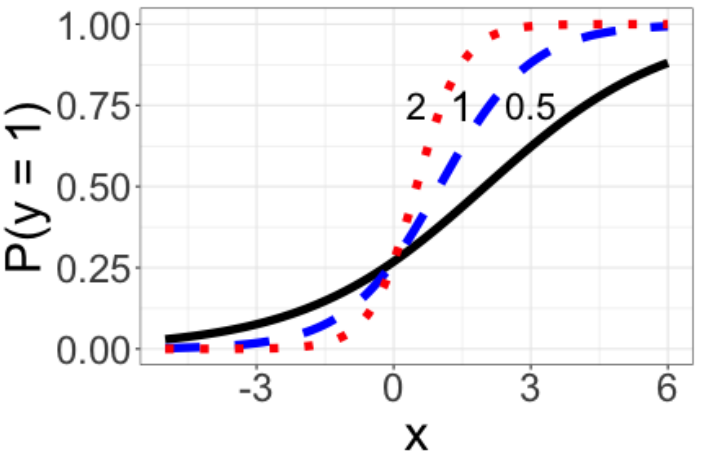
Shape of the regression curve

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

$$\pi_i = \frac{\exp\{\beta_0 + x_i\}}{1 + \exp\{\beta_0 + x_i\}} \quad \text{for} \quad \beta_0 = -3, -1, 1$$

$$\pi_i = \frac{\exp\{-1 + \beta_1\, x_i\}}{1 + \exp\{-1 + \beta_1\, x_i\}} \quad \text{for } \beta_1 = 0.5, 1, 2$$





Interpretation

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -2.45 + 0.22\ Age_i$$

> How would I interpret the slope and intercept of this fitted model in terms of the *log odds* that a patient has dengue?

Interpretation

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -2.45 + 0.22\ Age_i$$

> How do you think I would interpret the slope and intercept of this fitted model in terms of the *odds* that a patient has dengue?

Adding more variables

Now let's add WBC as a variable to the model:

```
m2 <- glm(Dengue ~ Age + WBC, data = dengue,
          family = binomial)
summary(m2)
```

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = 0.34 + 0.15\ Age_i - 0.31 WBC_i$$

> How should I interpret each coefficient in the fitted model?