Lecture 1: Intro to logistic regression

Agenda

- Introductions
- Overview of course details
- Begin logistic regression
- HW1 released on course website

Class overview

- STA 711 focuses on *statistical inference*: estimation, confidence intervals, and hypothesis testing
- Throughout the semester, topics will be initially motivated by logistic regression
- We will continue with inference and GLMs in STA 712 (Generalized Linear Models)

Grading philosophy

- Focusing on grades can detract from the learning process
- Homework should be an opportunity to *practice* the material. It is ok to make mistakes when practicing, as long as you make an honest effort
- Errors are a good opportunity to learn and revise your work
- Partial credit and weighted averages of scores make the meaning of a grade confusing. Does an 85 in the course mean you know 85% of everything, or everything about 85% of the material?

Grading in this course

- I will give you feedback on every assignment
- All assignments are graded as Mastered / Not yet mastered
- If you haven't yet mastered something, you get to try again!

Course components

- Regular homework assignments
  - Practice material from class
  - You may resubmit “Not yet mastered” questions once
- 3 take-home exams
  - Opportunity to demonstrate mastery of course material
  - Optional make-up exams for “Not yet mastered” questions
- Optional final exam
  - Final opportunity to demonstrate mastery

Assigning grades

To get an **A-** in the course:

- Receive credit for at least *N-2* homework assignments
- Master at least 80% of the questions on all three exams

To get an **A** in the course:

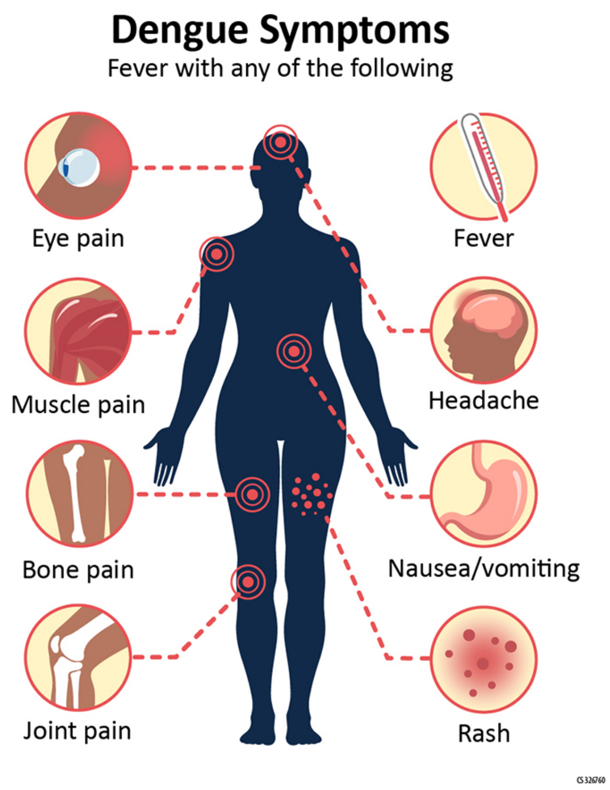
- Receive credit for at least *N-1* homework assignments
- Master at least 80% of the questions on all three exams

Late work and resubmissions

- You get a bank of **5** extension days. You can use 1–2 days on any assignment, exam, or project.
- No other late work will be accepted (except in extenuating circumstances!)

Motivating example: Dengue fever

Dengue fever: a mosquito-borne viral disease affecting 400 million people a year



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)

Motivating example: Dengue data

Research questions:

- How well can we predict whether a patient has dengue?
- Which diagnostic measurements are most useful?
- Is there a significant relationship between WBC and dengue?

Research questions

- How well can we predict whether a patient has dengue?
- Which diagnostic measurements are most useful?
- Is there a significant relationship between WBC and dengue?

How can I answer each of these questions? Discuss with a neighbor for 2 minutes, then we will discuss as a class.

- t-tests (difference in distribution in feature between dengue & non-dengue patients)
- Fit a regression model (e.g. logistic regression)
- Model comparison (model selection (nested tests, AIC, BIC)
- Prediction metrics (confusion matrices, accuracy, etc.)

Fitting a model: initial attempt

What if we try a linear regression model?

$$Y_i = \text{dengue status of ith patient}$$

$$Y_i = \beta_0 + \beta_1 \text{WBC}_i + \varepsilon_i \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$$

What are some potential issues with this linear regression model?

non-dengue

$Y_i \in \{0, 1\}$

dengue

$\beta_0 + \beta_1 WBC_i + \varepsilon_i \in (-\infty, \infty)$

but Y_i is binary!

Second attempt

Let's rewrite the linear regression model:

$$Y_i = \beta_0 + \beta_1 WBC_i + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

$$E[Y_i | WBC_i] = E[\beta_0 + \beta_1 WBC_i + \varepsilon_i \ | WBC_i]$$

$$= \beta_0 + \beta_1 WBC_i + \underbrace{E[\varepsilon_i]}_{0}$$

$$= \beta_0 + \beta_1 WBC_i$$

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

$$\Rightarrow \quad Y_i | WBC_i \sim N(\mu_i, \sigma_\varepsilon^2) \qquad \text{(random component)}$$

$$\mu_i = \beta_0 + \beta_1 WBC_i \qquad \text{(systematic component)}$$

Problem: $Y_i = 0$ or 1 $\Rightarrow$ $Y_i | WBC_i$ is not normal

Let's use the Bernoulli instead!

Second attempt

$$Y_i \sim \text{Bernoulli}(p_i) \quad p_i = \mathbb{P}(Y_i = 1 | \text{WBC}_i)$$

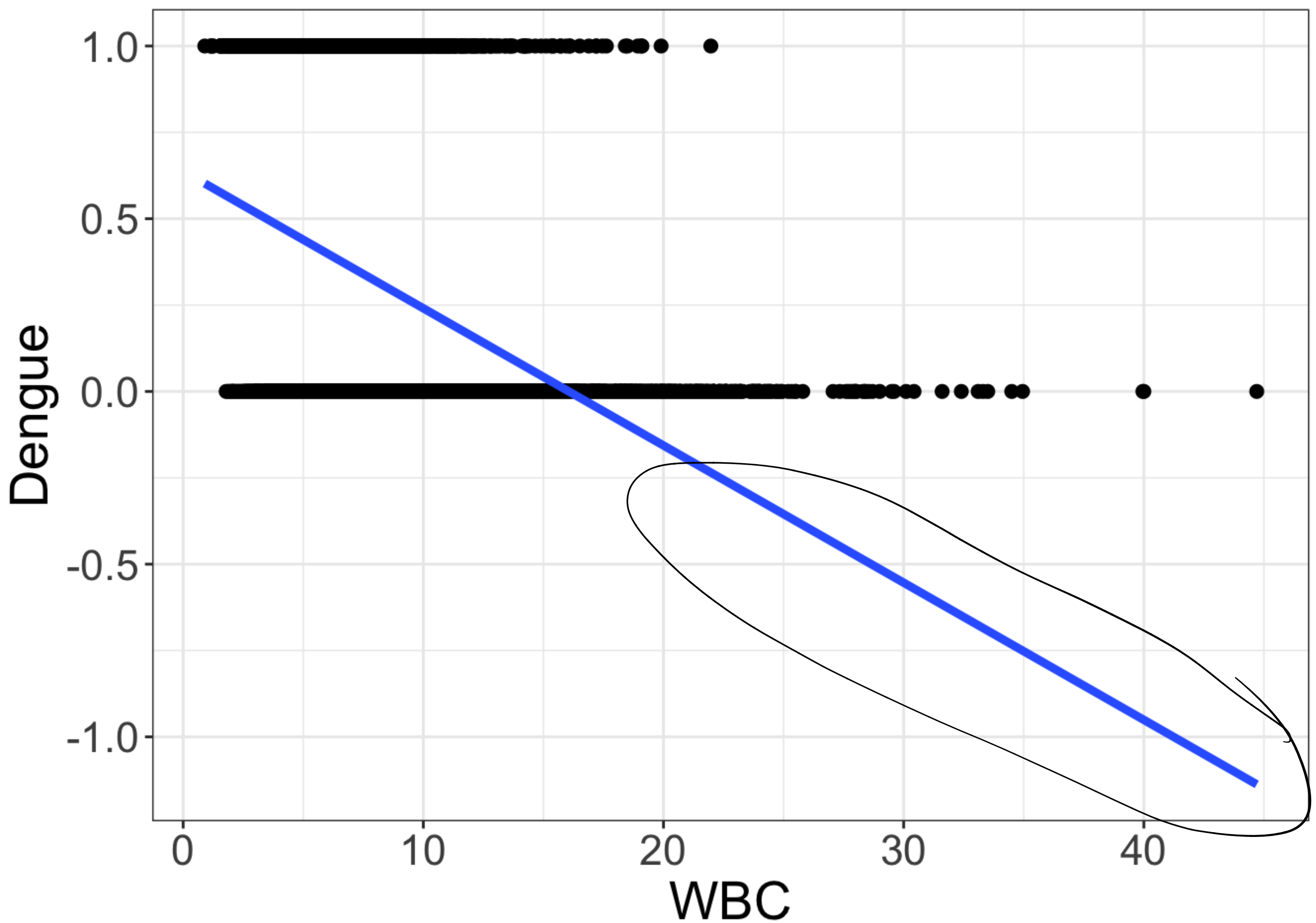
$$p_i = \beta_0 + \beta_1 \text{WBC}_i$$

Are there still any potential issues with this approach?

Problem: $p_i \in [0,1]$ but $\beta_0 + \beta_1 WBC_i \in (-\infty, \infty)$

(unless $\beta_0 = 0$)

Don’t fit linear regression with a binary response



if WBC > 15,
$\hat{p} < 0$
(bad)

Fixing the issue: logistic regression

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

random component

$$g(p_i) = \beta_0 + \beta_1 \text{WBC}_i$$

systematic component

where $g : (0, 1) \to \mathbb{R}$ is unbounded.

Usual choice: $g(p_i) = \underbrace{\log\left(\frac{p_i}{1 - p_i}\right)}_{\text{log odds}}$

$\frac{p_i}{1-p_i}$ = odds

$\in (0, \infty)$

link function

(links parameter p_i to predictor WBC_i)

log odds

aka logit

$\in (-\infty, \infty)$

link function: strictly increasing & bijective

Odds

Definition: If $p_i = \mathbb{P}(Y_i = 1|WBC_i)$, the **odds** are $\dfrac{p_i}{1 - p_i}$

Example: Suppose that $\mathbb{P}(Y_i = 1|WBC_i) = 0.8$. What are the *odds* that the patient has dengue?

$$\text{odds} = \frac{0.8}{1-0.8} = \frac{0.8}{0.2} = 4$$

prob. patient has dengue = 4 × prob. patient does not have dengue

Odds

Definition: If $p_i = \mathbb{P}(Y_i = 1|WBC_i)$, the **odds** are $\frac{p_i}{1-p_i}$

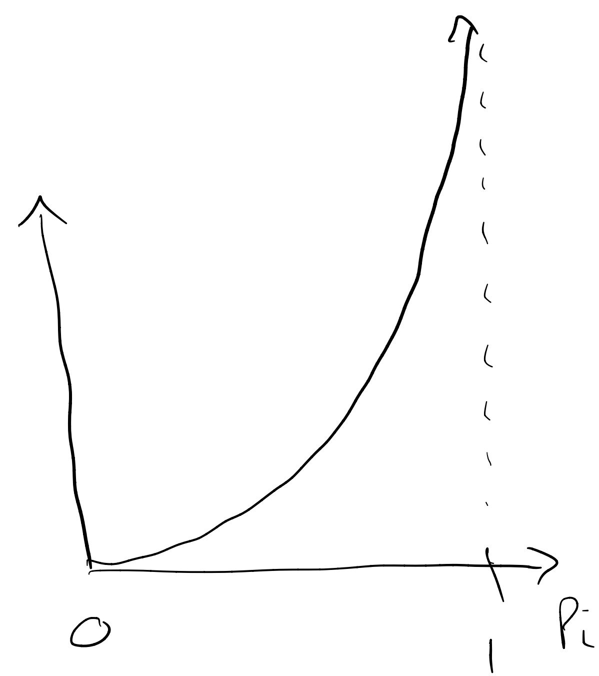
The probabilities $p_i \in [0, 1]$. The linear function $\beta_0 + \beta_1 WBC_i \in (-\infty, \infty)$. What range of values can $\frac{p_i}{1-p_i}$ take?

$$\frac{p_i}{1-p_i} \in [0, \infty)$$

if $p_i = 0$ odds $= 0$

if $p_i = 1$ odds $= \infty$

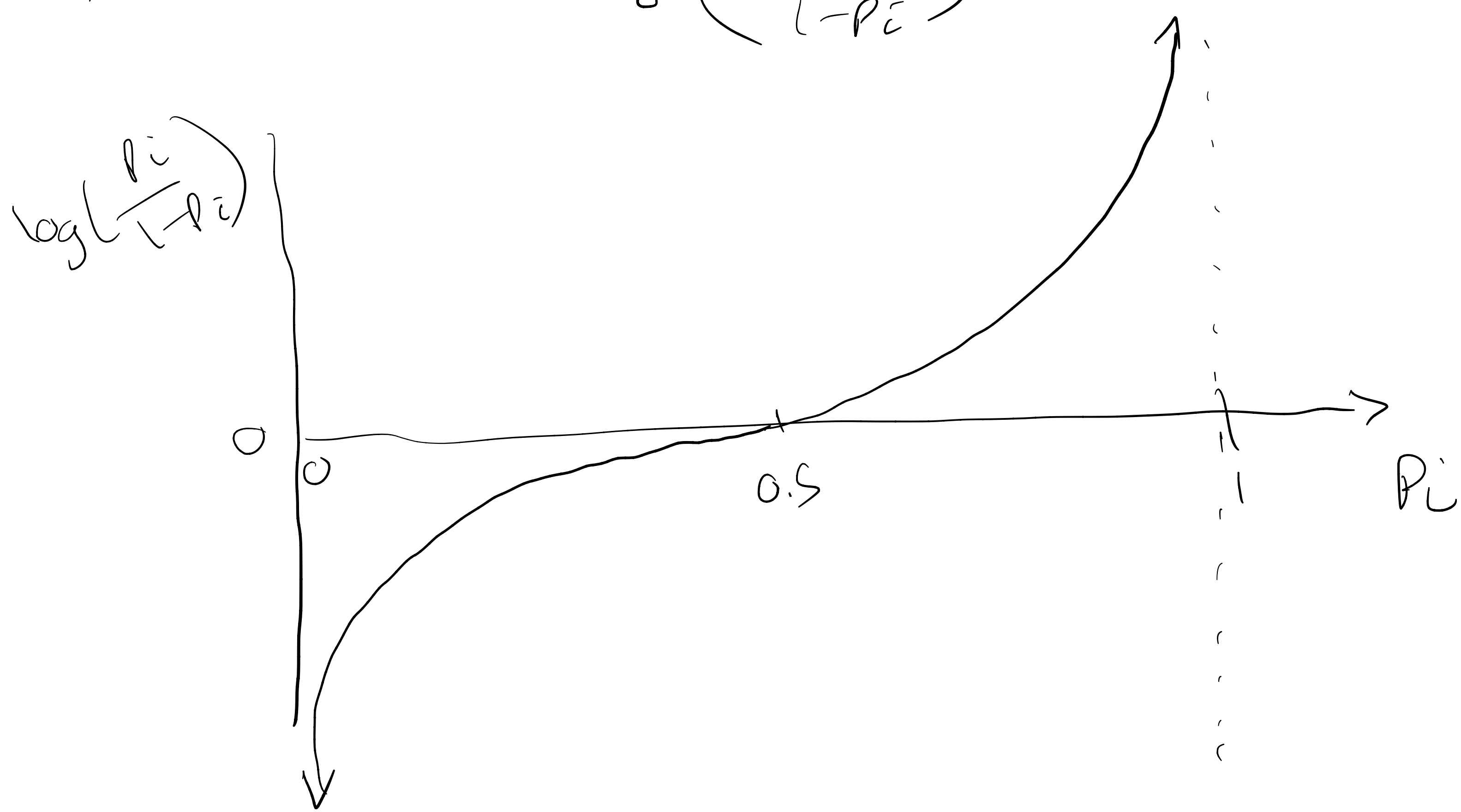
$\frac{p_i}{1-p_i}$



Log odds

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

$\frac{p_i}{1-p_i} \in [0,\infty) \Rightarrow \log\left(\frac{p_i}{1-p_i}\right) \in (-\infty,\infty)$



Binary logistic regression

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

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

Note: Can generalize to $Y_i \sim \text{Binomial}(m_i, p_i)$, but we won't do that yet.

Example: simple logistic regression

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

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

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?
- Interpret the estimated slope in context of a unit change in the log odds.
- What is the change in *odds* asociated with a unit increase in WBC?