Introduction 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
  - A subset of questions will be graded
  - 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 a **C** in the course:

- Receive credit for at least 4 homework assignments
- Master at least 80% of the questions on one exam

To get a **B** in the course:

- Receive credit for at least 5 homework assignments
- Master at least 80% of the questions on two exams

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

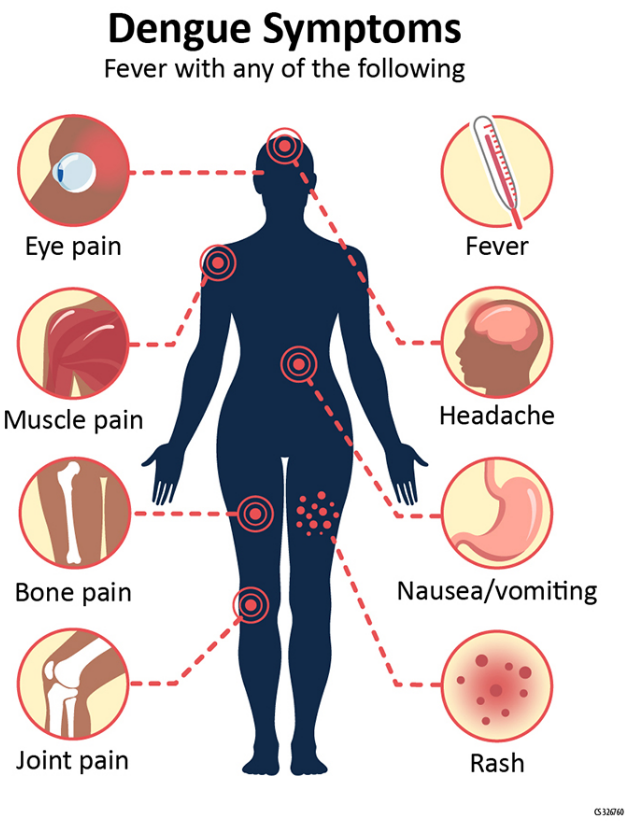
- Receive credit for at least 5 homework assignments
- Master at least 80% of the questions on all three exams

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

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)

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.

- EDA (plots of dengue vs. WBC, dengue vs. Age, etc...)
  - ask experts / clients for context
- Fit regression model (e.g., logistic regression)
- CIs, hyp. tests, effect sizes for coefficients
- Model selection (stepwise selection, penalized, etc.)
- prediction metrics (confusion matrix, accuracy, etc.)
  - cross validation

Fitting a model: initial attempt

What if we try a linear regression model?

$$Y_i = \text{dengue status of } i\text{th \ patient}$$

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

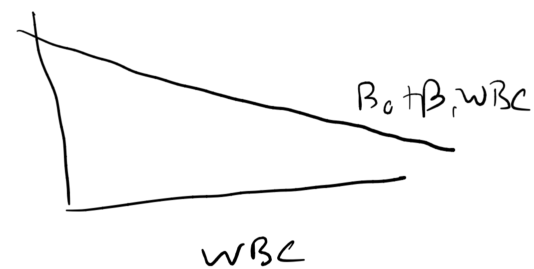
> What are some potential issues with this linear regression model?

$\beta_0 + \beta_1 WBC_i + \varepsilon_i$

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

and is continuous

but Y_i is binary!



Second attempt

Let's rewrite the linear regression model:

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

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

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

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

$\mu_i = \beta_0 + \beta_1 WBC_i$ (systematic component)

<u>Problem</u>: $Y_i = 0$ or 1 $\Rightarrow$ $Y_i \mid WBC_i$ is <u>not</u> normal

Let's use Bernoulli instead!

Second attempt

random component
$$Y_i \sim Bernoulli(p_i) \qquad p_i = \mathbb{P}(Y_i = 1|WBC_i)$$

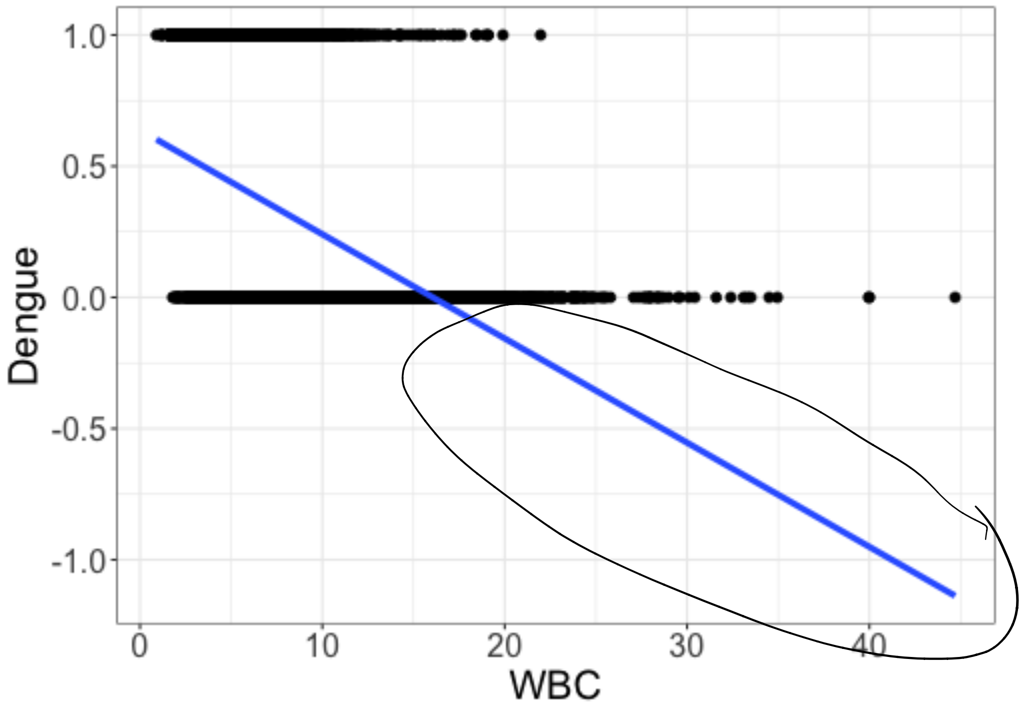
systematic component (wrong)
$$p_i = \beta_0 + \beta_1 WBC_i$$

> Are there still any potential issues with this approach?

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

(unless $\beta_1 \equiv 0$)

Don't fit linear regression with a binary response



If WBC > 15, predictions < 0

instead: fit a curve!

Fixing the issue: logistic regression

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

$$g(p_i) = \beta_0 + \beta_1 WBC_i$$ Systematic component

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

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

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

link function: links parameter p_i to predictor WBC_i

log odds aka logit

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?

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

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

Odds

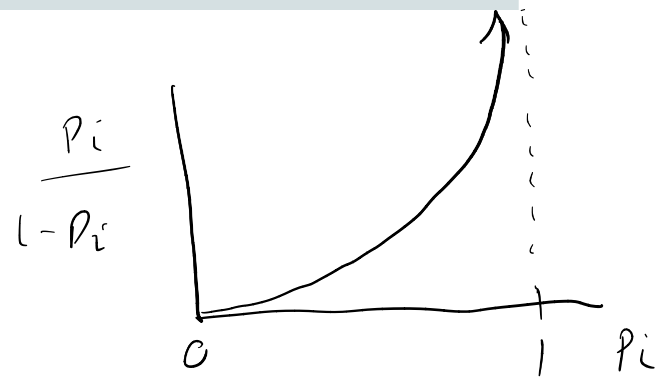
Definition: If $p_i = \mathbb{P}(Y_i = 1 | WBC_i)$, the **odds** are $\dfrac{p_i}{1 - p_i}$

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

if $p = 0$ odds $= 0$

if $p = 1$ odds $= \infty$

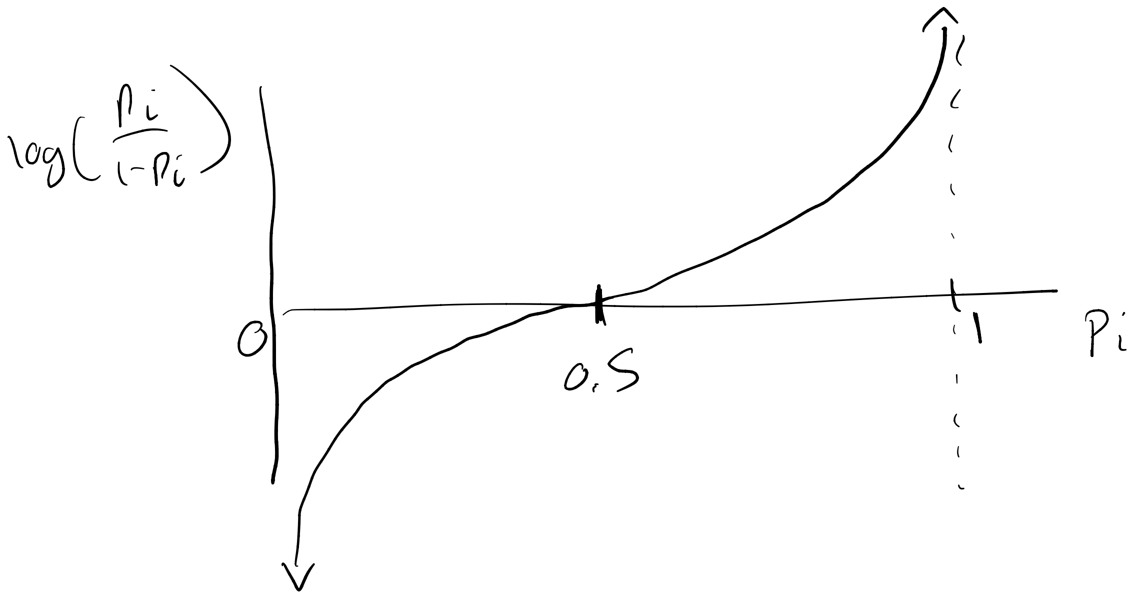
odds $\in [0, \infty)$



Log odds

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

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



Binary logistic regression

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

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

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

random component: specifies distribution of Y_i

systematic component: relates distribution to explanatory variable(s)

Example: simple logistic regression with dengue

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

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

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?