Welcome to STA 214

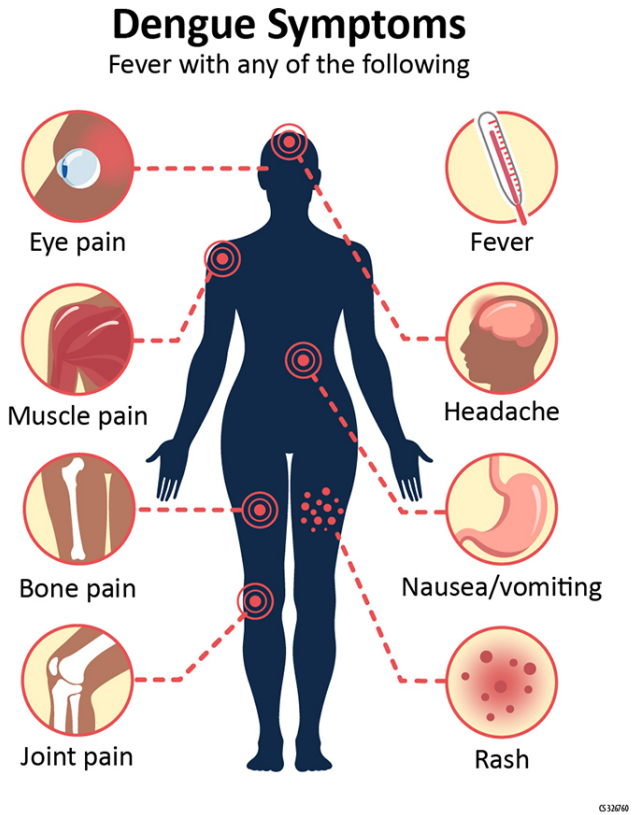
Dr. Ciaran Evans

Agenda

- Introductions
- Beginning logistic regression
- Plan for week 1 and the semester
- Syllabus highlights

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)

Research questions

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

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

- Looking @ relationships between dengue and explanatory variables
  - fit regression model!
- Performance assessments (accuracy, sensitivity, specificity, etc.)
- Compare nested models & hypothesis tests

Exploratory data analysis (EDA)

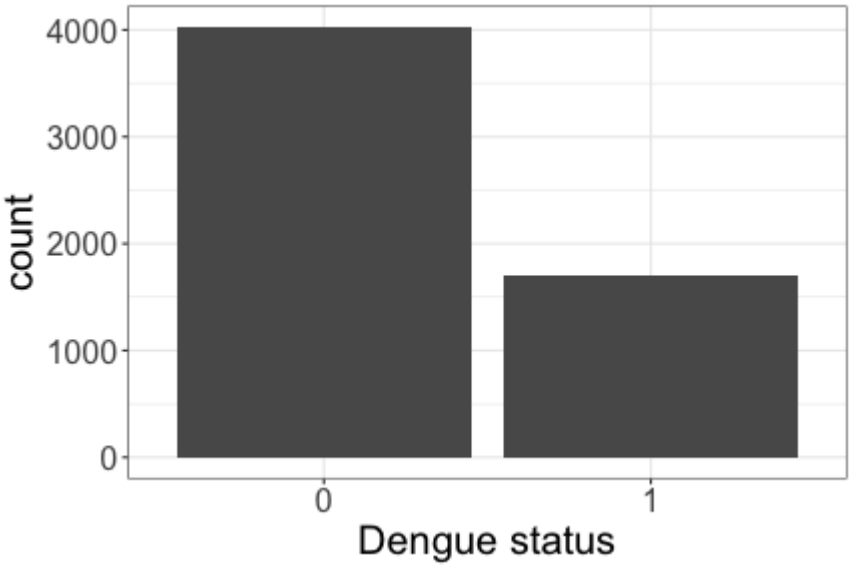
> What plot can we use to visualize the response (dengue status)?

Bar chart.

Exploratory data analysis (EDA)

> What plot can we use to visualize the response (dengue status)?

Answer: Bar chart



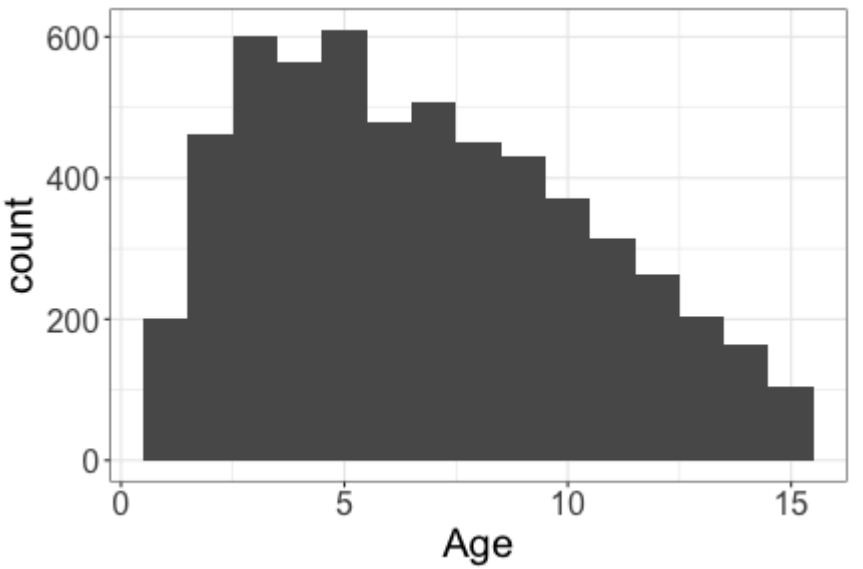
Exploratory data analysis (EDA)

> What plot can we use to visualize an explanatory variable like Age?

Exploratory data analysis (EDA)

> What plot can we use to visualize an explanatory variable like Age?

Answer: Histogram (or boxplot)



Class activity

https://sta214-s23.github.io/class_activities/ca_lecture_1.html

In the class activity, you'll start to explore the relationship between age and dengue.

Spend a few minutes to do the following:

- Say hi to the people around you, introduce yourself
- Work in groups on the class activity
- You don't need to submit your work

Class activity

> What is the (empirical) probability that a patient in the study has dengue?

$P(\text{Dengue} = 1) =$ $\frac{1697}{5720} \approx 0,3$

Class activity

> What is the (empirical) probability that a 5 year old patient has dengue? What about a 10 year old patient?

$$P(\text{Dengue} = 1 | \text{Age} = 5) = \frac{122}{122 + 487} = 0.2$$

$$P(\text{Dengue} = 1 | \text{Age} = 10) = \frac{185}{185 + 185} = 0.5$$

Odds

> What are the (empirical) odds that a 5 year old patient has dengue?

$$\text{odds} = \frac{\pi}{1-\pi}$$

5 year old patient: $\text{odds} = \frac{0.2}{1-0.2} = 0.25$

$$= \frac{P(\text{Dengue} \mid \text{Age}=5)}{P(\text{No dengue} \mid \text{Age}=5)}$$

$P(\text{patient has dengue} \mid \text{Age}=5) = 0.25 \times P(\text{don't have dengue} \mid \text{Age}=5)$

patients $1, \ldots, 5720$

Fitting a model: initial attempt

What if we try a linear regression model?

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

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

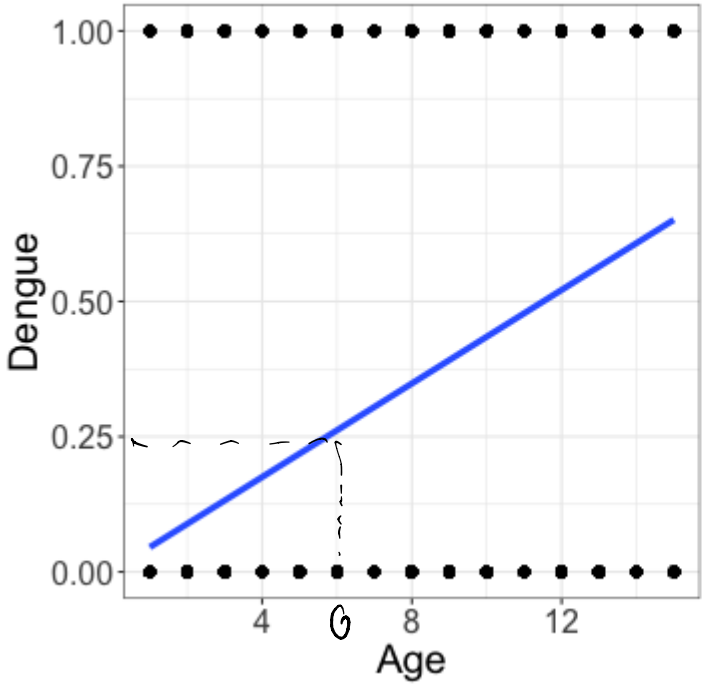
> What are some potential issues with this linear regression model? Discuss with your neighbors for 1--2 minutes, then we will discuss as a class.

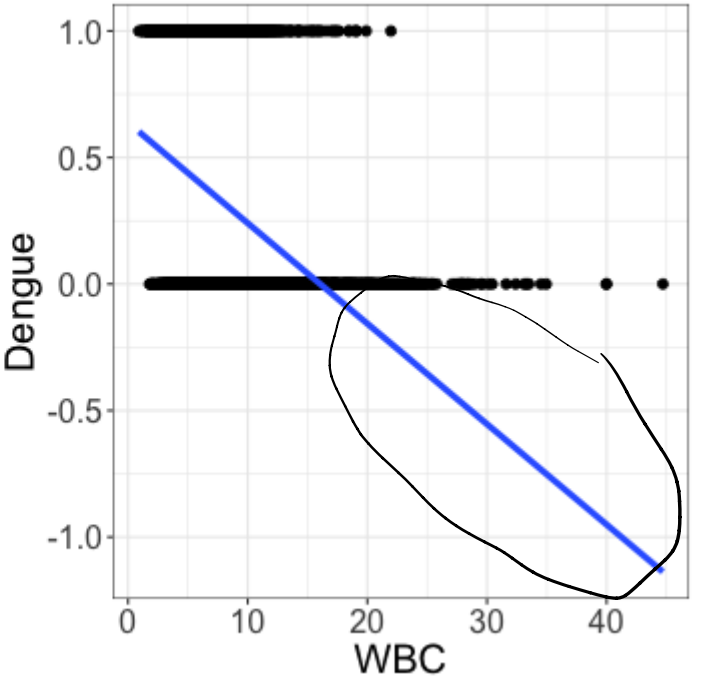
Y_i is binary (either 0 or 1)

$\beta_0 + \beta_1 Age_i + \varepsilon_i$ is continuous

It is impossible to have $Y_i = \beta_0 + \beta_1 Age_i + \varepsilon_i$

Don't fit linear regression with a binary response





$Age = 6 \rightarrow \hat{Dengue} = 0.25$

?

$WBC > 15 \rightarrow \hat{Dengue} < 0$

Revisiting the linear regression model

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

$\Rightarrow$ $Y_i \mid X_i$ is also Normal

$\Rightarrow$ $Y_i \sim N(\underbrace{\beta_0 + \beta_1 X_i}_{\mu_i}, \sigma_\varepsilon^2)$

Two parts to my regression model:

$$Y_i \sim N(\mu_i, \sigma_\varepsilon^2)$$ ← specifies distribution of Y

$$\mu_i = \beta_0 + \beta_1 X_i$$ ← relates distribution to X_i

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 e.g., $Y_i \sim N(\mu_i, \sigma_\varepsilon^2)$
- **Step 2:** Build a model for the parameters of interest e.g., $\mu_i = \beta_0 + \beta_1 X_i$
- **Step 3:** Fit the model

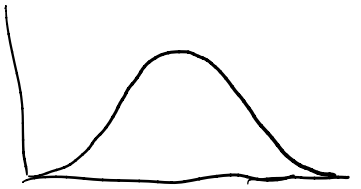
e.g. In R: lm(y ~ x , ...)

Step 1: Choose a reasonable distribution for Y_i

> What do I mean by a *distribution*?

- distribution describes spread of Y_i

e.g. Normal :



Step 1: Choose a reasonable distribution for Y_i

> What do I mean by a *distribution*?

- A **distribution** tells us what outcomes are possible for Y_i, and how often these outcomes occur.

Here the possible values of Y_i are 0 (no dengue) and 1 (dengue).

> How often do these values occur in the population?

Step 1: Choose a reasonable distribution for Y_i

> What do I mean by a *distribution*?

- A **distribution** tells us what outcomes are possible for Y_i, and how often these outcomes occur.

Here the possible values of Y_i are 0 (no dengue) and 1 (dengue).

> How often do these values occur in the population?

- We don't know, so we will estimate from the sample
- We assume the probability $Y_i = 1$ depends on Age_i

Step 1: Choose a reasonable distribution for Y_i

> How should I describe the distribution of Y_i?

$Y_i = 0$ or 1

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

That's it!

Bernoulli distribution

"is distributed as"

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

> What do I mean by a *random variable*?

Bernoulli distribution

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

> What do I mean by a *random variable*?

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 $\pi_i = P(Y_i = 1)$

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?

Fixing the issues

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

Next steps

- We will spend the next few weeks talking in depth about logistic regression
- Download R and RStudio today or tomorrow
  - Instructions on course website
  - Please come to office hours or contact me if you have problems!
- Bring laptop to class
- HW 1 released on the course website
- Course codebook also on the course website

Semester plan

- Logistic regression
- Multinomial regression
- Poisson regression
- Mixed effects
- Parametric bootstrapping (time permitting)

Course prerequisites

Prerequisites:

- STA 112 (previously numbered 212) and MTH 111 (Calculus I)
- You may *not* take this class if you previously took STA 279 with me or Dr. Dalzell!

I expect you to be familiar with:

- EDA (Exploratory data analysis)
- Linear regression with multiple predictors
- Comparing and interpreting models
- Confidence intervals and hypothesis tests
- Basic R computing

Getting help

- Office hours
  - sign up for 15-minute time slots
  - see course web page
- Email (evansc@wfu.edu)

Diversity and Inclusion

In this class, we will embrace diversity of age, background, beliefs, ethnicity, gender, gender identity, gender expression, national origin, neurotype, race, religious affiliation, sexual orientation, and other visible and non-visible categories. The university and I do not tolerate discrimination.

- Let me know ways to improve the effectiveness of the course for you personally, or for other students or student groups.
- *Names/Pronouns:* You deserve to be addressed in the manner you prefer. To guarantee that I address you properly, you are welcome to tell me your pronoun(s) and/or preferred name at any time, either in person or via email.

Course components

Component	Weight
Homework and Labs	20%
Exam 1	15%
Exam 2	15%
Final Exam	20%
Project 1	15%
Project 2	15%

Extensions and late work

Extensions: You have a bank of **5** extension days, which you may use over the course of the semester. You may use either one or two days for a given assignment. *Additional extensions may be given, on an individual basis, in extenuating circumstances.*

Late work: An assignment will be marked off 20% for every 24 hours it is late (past the original due date).

If you know you cannot turn in an assignment (for instance, if you are ill or there is a family emergency), let me know before the assignment is due, and we will work something out. There will be no grade changes after our last day of class.