Welcome to STA 214

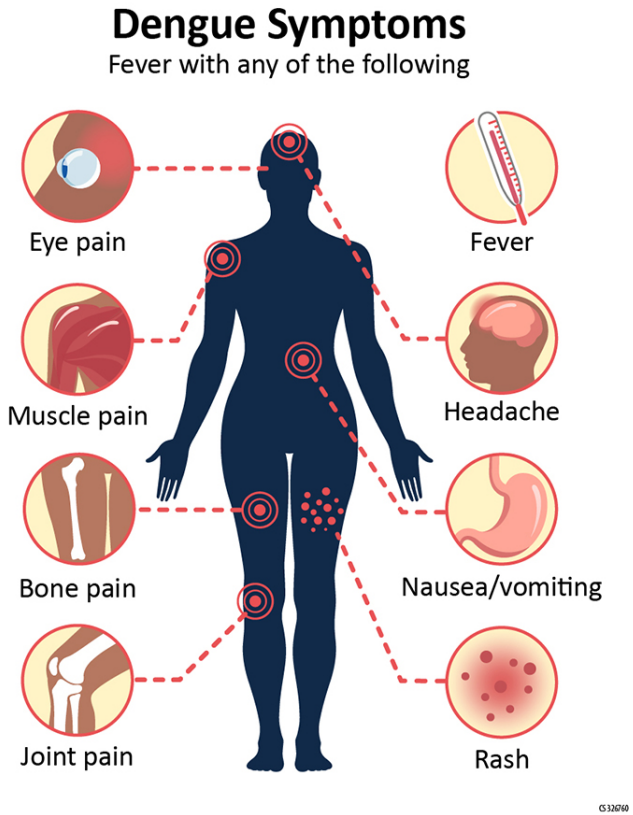
Ciaran Evans

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)

Warmup activity

Work on the activity (handout) with a neighbor, then we will discuss as a class.

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 might you try to address each of these questions?

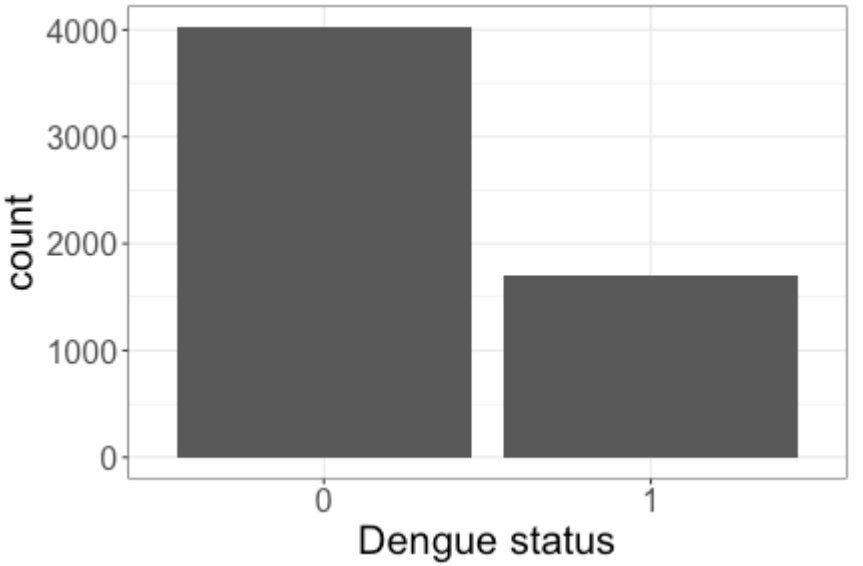
Exploratory data analysis (EDA)

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

Exploratory data analysis (EDA)

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

Answer: Bar chart



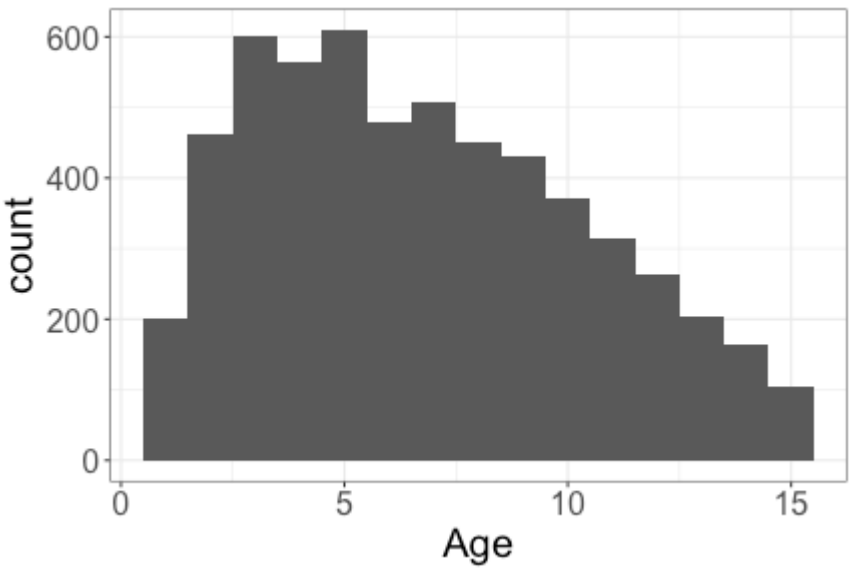
Exploratory data analysis (EDA)

> 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

In the class activity (handout), 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
- I will collect the handout at the end of class

Class activity

> What do you think are some of the potential risks and benefits of the dengue study?

Class activity

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

$P(\text{Dengue} = 1) =$

Class activity

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

$$P(\text{Dengue} = 1 | \text{Age} = 5) =$$

$$P(\text{Dengue} = 1 | \text{Age} = 10) =$$

Odds

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

Fitting a model: linear regression?

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

Explanatory variable: $Age_i = \text{dengue status of } i\text{th patient}$

> What would a *linear* regression model look like for these variables?

Fitting a model: initial attempt

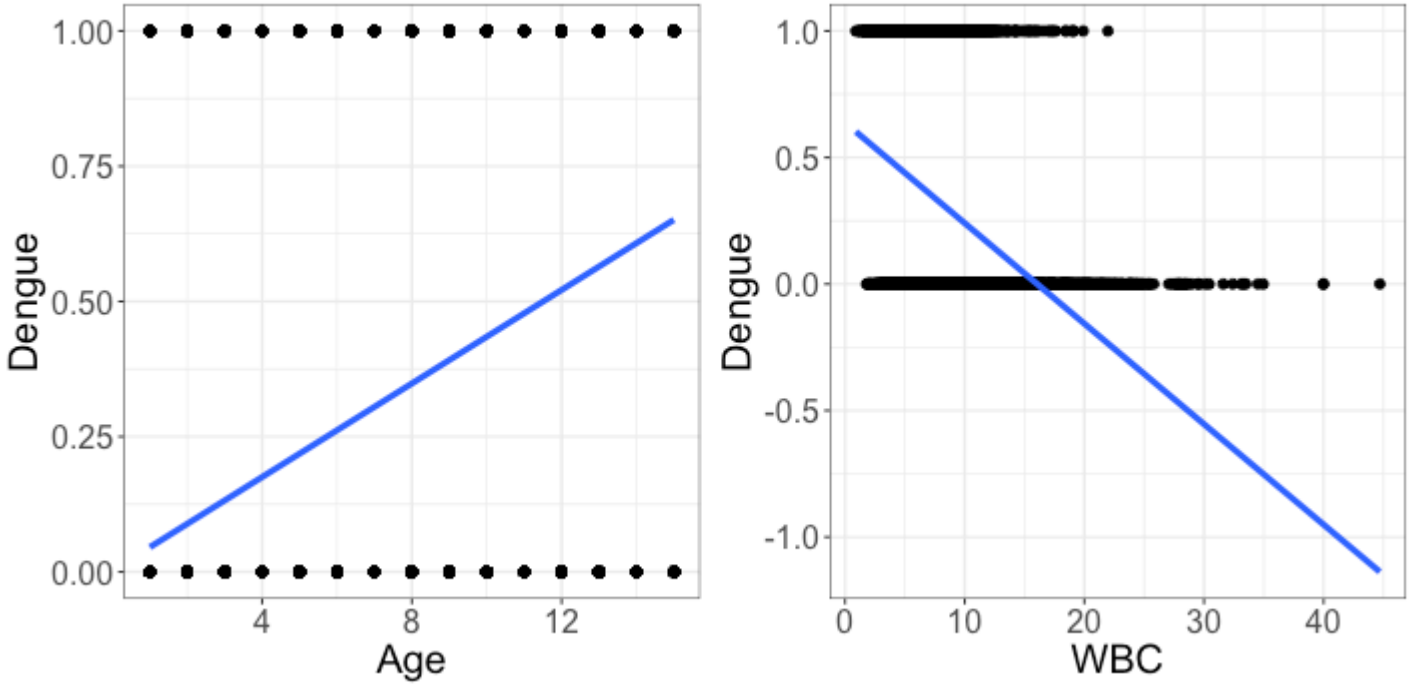
What if we try a linear regression model?

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

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

> 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.

Don't fit linear regression with a binary response



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 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 (tentative)

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

Course goals:

- Be able to model data with different types of response variables
- Extend tools from 112 to address violations of the usual linear regression assumptions
- Emphasize sound statistical and data analysis practices

Course prerequisites

Prerequisites:

- STA 112 and MTH 111 (Calculus I)

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

Expectations

- Complete any assigned reading ahead of class
- Bring laptop each day
- Submit class activities (graded for effort, not completeness or correctness)
- Attend department seminars (more info to follow)

Course components

- Class participation (graded for effort)
- HW assignments (roughly one per week)
- Exams (2 midterms, 1 final)
- Project

AI policy

- I will *never* use AI to grade your work; all feedback you receive will be directly from me and the TA
- Collaboration with other students, and AI assistance, is permitted on homework
  - Assistance does *not* mean uploading the assignment to ChatGPT and copying the answers
  - You must cite collaborators and external resources
- See syllabus for further details

For next time

- Make sure R and RStudio are installed
- Instructions are provided on the course website
- **Reading:** sections 3.3.1 and 6.2 in the textbook