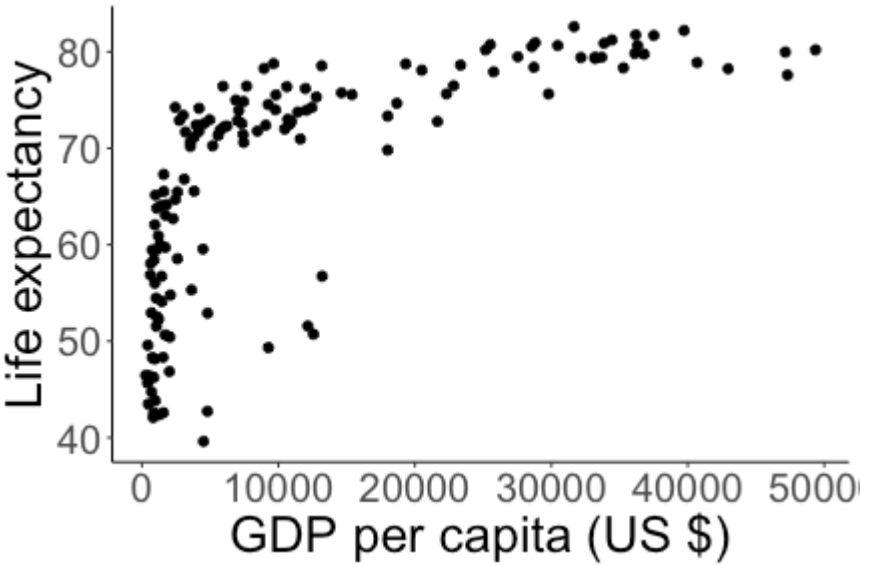
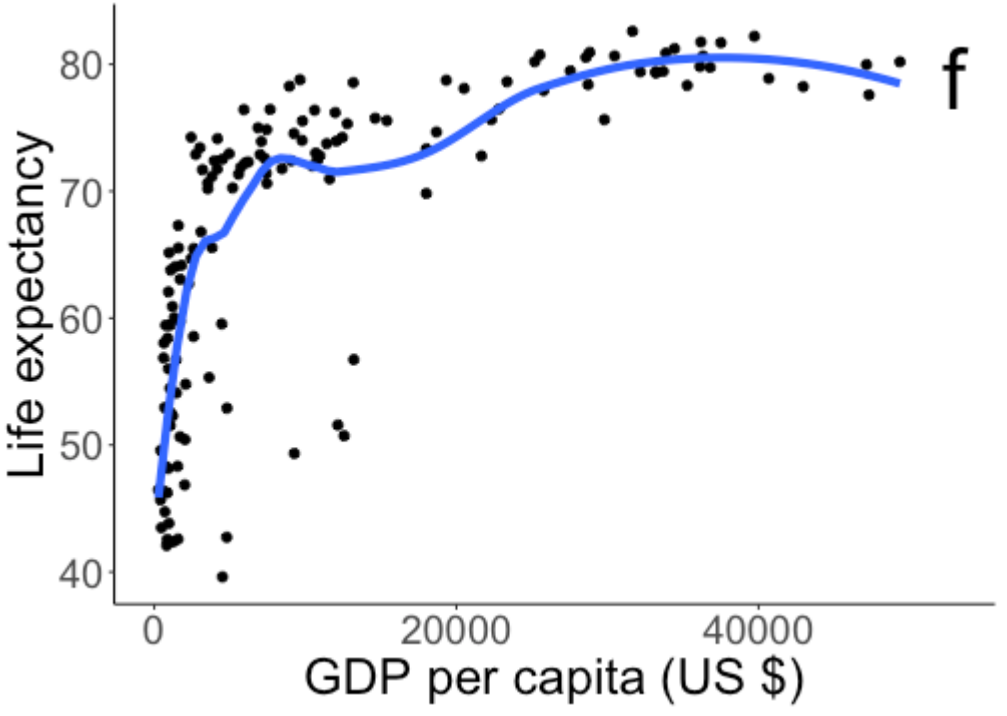
Intro to Regression

Modeling



- Given a value of GDP per capita, what would I predict for life expectancy?
- What is the relationship between GDP per capita and life expectancy?

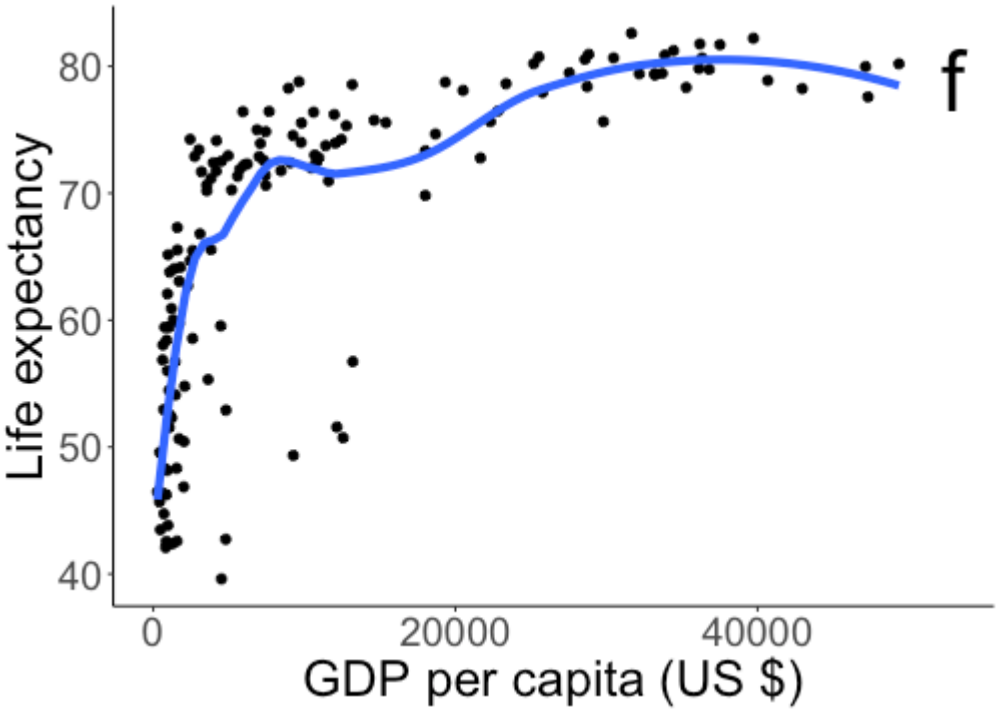
General model



$$\text{life expectancy} \stackrel{?}{=} f(\text{GDP per capita})$$

Do all of the observations fall exactly on the curve?

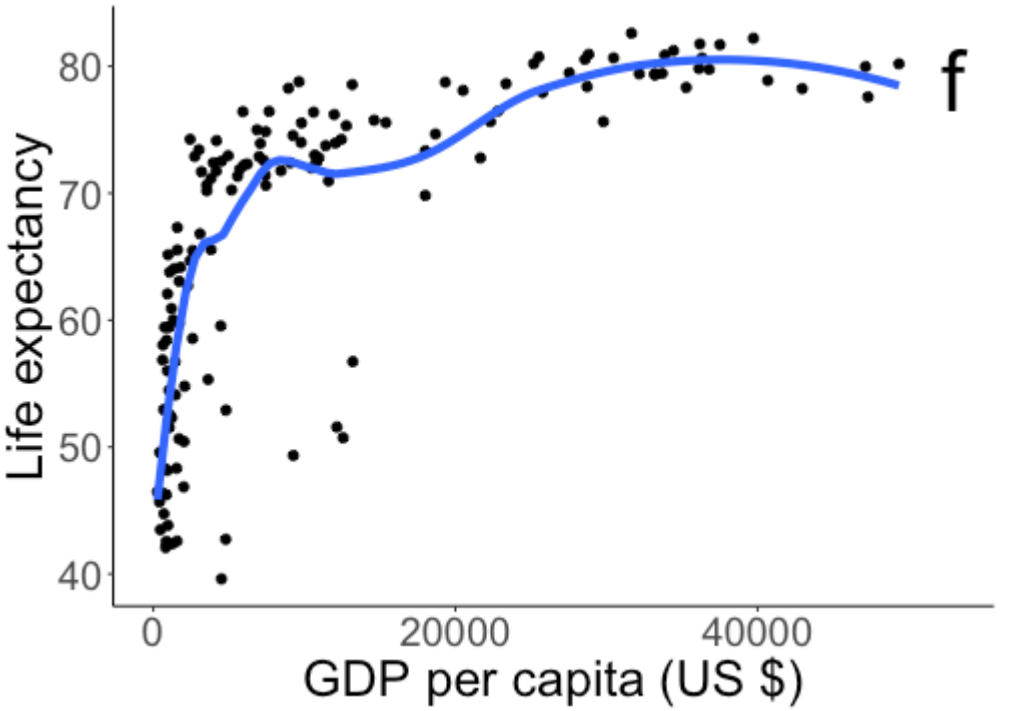
General model



$$\text{life expectancy} = f(\text{GDP per capita}) + \varepsilon$$

- f = average life expectancy, given GDP per capita
- ε = random error term (how observations vary around f)

Regression



$$\text{life expectancy} = f(\text{GDP per capita}) + \varepsilon$$

- But, we don't actually know what f is!
- Regression: estimate f

Goals of regression

In general, write

$$y = f(x) + \varepsilon$$

- x = predictor, aka independent or explanatory variable
- y = response, aka dependent variable

Goals of regression

$$y = f(x) + \varepsilon$$

How do we estimate f? Depends on our goal.

- **Prediction:** given a value of x, what is our "best guess" for the value of y?
  - don't care about the form of f, just want to get good predictions
- **Interpretation/association:** What is the relationship between x and y?
  - want to get form of f right
- **Causal inference:** If I change x, how does y change?
  - need study design that allows for causal conclusions (e.g., randomized experiment)

Prediction, interpretation/association, or causal inference?

Scenario: A beer company is conducting a social media marketing campaign, and wants to identify individuals who are likely to buy their beer based on Facebook and Instagram activity. The company doesn't care about understanding the relationship, they just want to accurately target likely customers.

(A) Prediction

(B) Interpretation/association

(C) Causal inference

Prediction, interpretation/association, or causal inference?

Scenario: A beer company is conducting a social media marketing campaign, and wants to identify individuals who are likely to buy their beer based on Facebook and Instagram activity. The company doesn't care about understanding the relationship, they just want to accurately target likely customers.

(A) Prediction

(B) Interpretation/association

(C) Causal inference

Answer: Prediction

Prediction, interpretation/association, or causal inference?

Scenario: The beer company advertises on several different platforms: Facebook, Instagram, YouTube, and on several popular podcasts. They want to know whether the amount they spend on each platform is associated with an increase in sales.

(A) Prediction

(B) Interpretation/association

(C) Causal inference

Prediction, interpretation/association, or causal inference?

Scenario: The beer company advertises on several different platforms: Facebook, Instagram, YouTube, and on several popular podcasts. They want to know whether the amount they spend on each platform is associated with an increase in sales.

(A) Prediction

(B) Interpretation/association

(C) Causal inference

Answer: Interpretation/association

Prediction, interpretation/association, or causal inference?

Scenario: The beer company wants to target social media influencers who can help sell their beer. Unfortunately, no one in the company knows any influencers. So they decide to identify influencers by finding Instagram posts with pictures of beer. To do so, the company trains a neural network which takes an Instagram image as input, and outputs either "contains beer" or "does not contain beer". They don't care how the neural network works, they just want to identify images of beer.

(A) Prediction

(B) Interpretation/association

(C) Causal inference

Prediction, interpretation/association, or causal inference?

Scenario: The beer company wants to target social media influencers who can help sell their beer. Unfortunately, no one in the company knows any influencers. So they decide to identify influencers by finding Instagram posts with pictures of beer. To do so, the company trains a neural network which takes an Instagram image as input, and outputs either "contains beer" or "does not contain beer". They don't care how the neural network works, they just want to identify images of beer.

(A) Prediction

(B) Interpretation/association

(C) Causal inference

Answer: Prediction

Prediction, interpretation/association, or causal inference?

Scenario: The beer company made their Facebook ads with Comic Sans, but it turns out that Comic Sans isn't cool anymore. The company considers switching their advertising font to Papyrus, but they want to know whether changing the font will lead to more sales.

(A) Prediction

(B) Interpretation/association

(C) Causal inference

Prediction, interpretation/association, or causal inference?

Scenario: The beer company made their Facebook ads with Comic Sans, but it turns out that Comic Sans isn't cool anymore. The company considers switching their advertising font to Papyrus, but they want to know whether changing the font will lead to more sales.

(A) Prediction

(B) Interpretation/association

(C) Causal inference

Answer: Causal inference

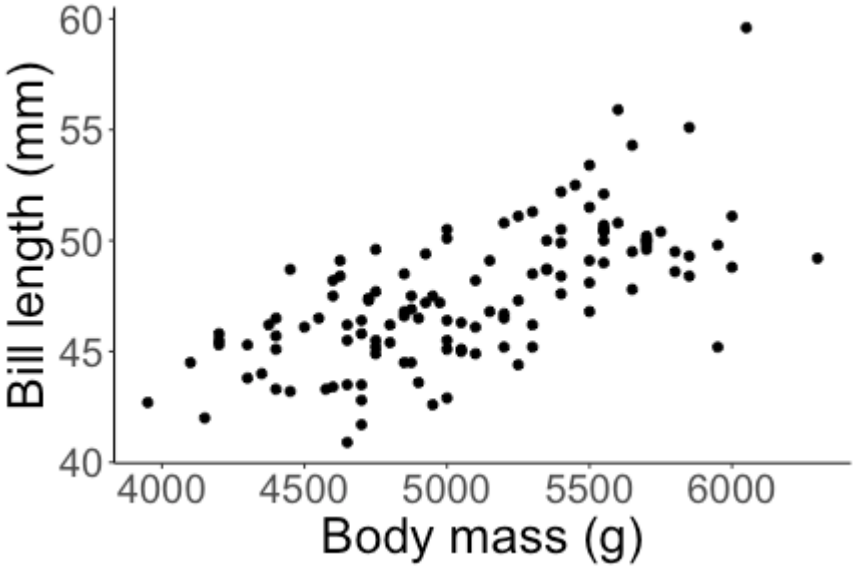
Some regression methods

$$y = f(x) + \varepsilon$$

- Models often used for prediction (fewer assumptions about f):
  - tree-based methods
  - local regression
  - neural networks
  - $+$ others
- Models often used for interpretation/association (more assumptions about f):
  - linear regression
  - generalized linear models
  - penalized regression
  - $+$ others

Linear regression

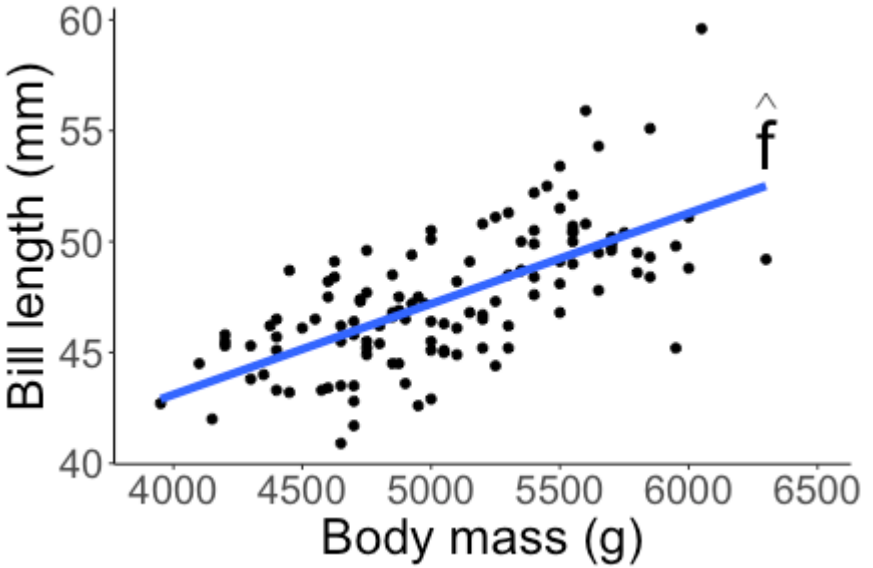
What is the relationship between body mass and bill length for Gentoo penguins?



> How would you describe this relationship?

Linear regression

> When the relationship appears linear, let's use a line!



$$\text{bill length} = f(\text{body mass}) + \varepsilon$$

$\widehat{f}$ (blue line) is our *estimate* of f

Simple linear regression

One predictor (body mass), one response (bill length):

$$\text{bill length} = f(\text{body mass}) + \varepsilon$$

Linear regression: $f(\text{body mass}) = \beta_0 + \beta_1 \text{ body mass}$

- β_0: intercept of line
- β_1: slope of line

> Do we know β_0 and β_1?

Simple linear regression

One predictor (body mass), one response (bill length):

$$\text{bill length} = f(\text{body mass}) + \varepsilon$$

Linear regression: $f(\text{body mass}) = \beta_0 + \beta_1 \text{ body mass}$

- β_0: intercept of line
- β_1: slope of line

> Do we know β_0 and β_1?

No! We only have a sample of data, so we *estimate* the relationship:

$$\widehat{\text{bill length}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{ body mass}$$

Notation

Assumed truth: $y = \beta_0 + \beta_1 x + \varepsilon$

Estimate from sample: $\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

- y = quantitative response variable
- x = quantitative predictor
- β_0 = intercept of true regression line
- β_1 = slope of true regression line
- $\widehat{\beta}_0$ = intercept of estimated regression line
- $\widehat{\beta}_1$ = slope of estimated regression line
- $\hat{y}$ = estimated response

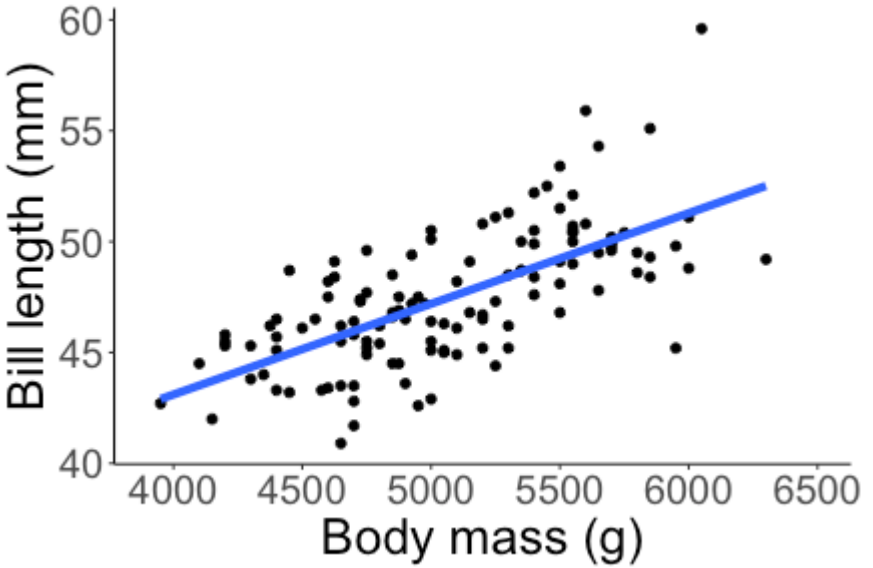
Notation

Assumed truth: $y = \beta_0 + \beta_1 x + \varepsilon$

Estimate from sample: $\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

- y = quantitative response variable
- x = quantitative predictor
- β_0 = intercept of true regression line **(parameter)**
- β_1 = slope of true regression line **(parameter)**
- $\widehat{\beta}_0$ = intercept of estimated regression line **(parameter estimate)**
- $\widehat{\beta}_1$ = slope of estimated regression line **(parameter estimate)**
- $\hat{y}$ = estimated response

Estimated regression line



$$\widehat{\text{bill length}} = 26.74 + 0.004 \text{ body mass}$$

- $\widehat{\beta}_0 = 26.74$
- $\widehat{\beta}_1 = 0.004$

Estimated regression line

$$\widehat{\text{bill length}} = 26.74 + 0.004 \text{ body mass}$$

Suppose a Gentoo penguin has mass 5000g. What is the predicted bill length?

Estimated regression line

$$\widehat{\text{bill length}} = 26.74 + 0.004 \text{ body mass}$$

Suppose a Gentoo penguin has mass 5000g. What is the predicted bill length?

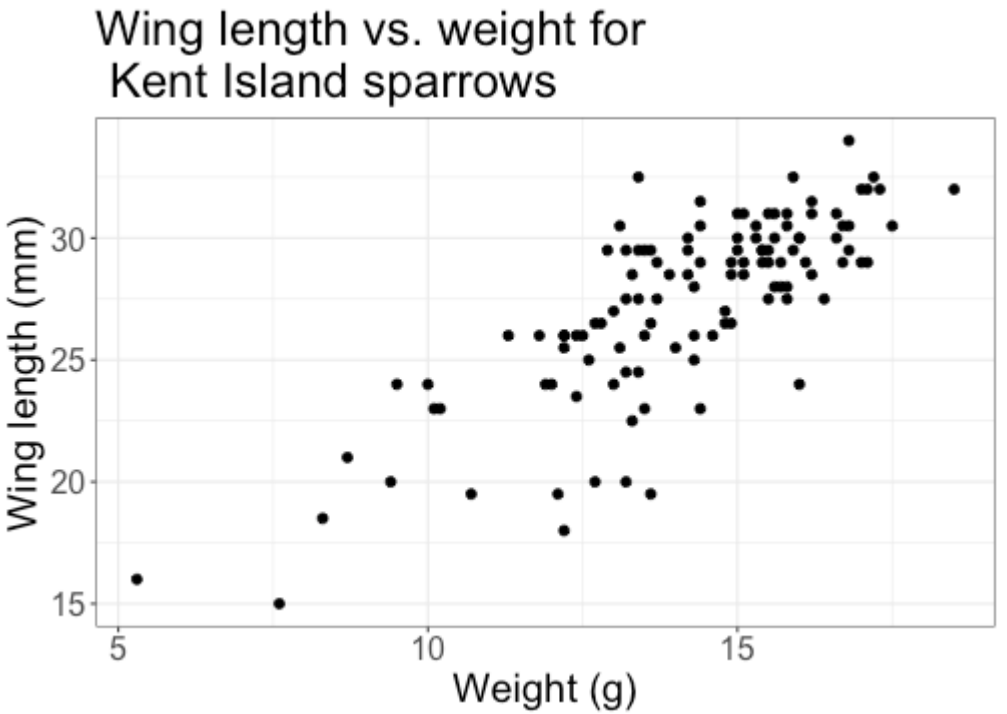
Answer:

$$26.74 + 0.004 \cdot 5000 = 46.74$$

Class activity

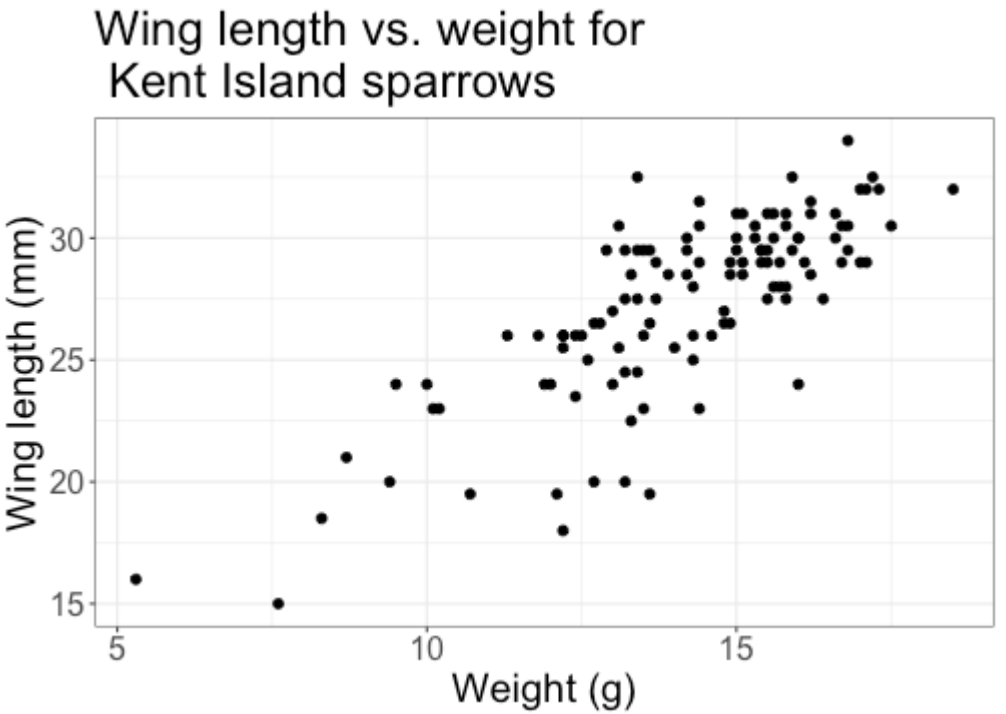
Work on the activity (handout) with a neighbor, then we will discuss as a class.

Class activity



Is linear regression an appropriate choice?

Class activity



> Is linear regression an appropriate choice?

Yes! The relationship looks approximately linear.

Class activity

$$\widehat{\text{wing length}} = 8.755 + 1.313 \text{ weight}$$

$\widehat{\beta}_0 =$

$\widehat{\beta}_1 =$

Class activity

$$\widehat{\text{wing length}} = 8.755 + 1.313 \text{ weight}$$

$\widehat{\beta}_0 = 8.755$

$\widehat{\beta}_1 =$

Class activity

$$\widehat{\text{wing length}} = 8.755 + 1.313 \text{ weight}$$

$\widehat{\beta}_0 = 8.755$

$\widehat{\beta}_1 = 1.313$

Class activity

$$\widehat{\text{wing length}} = 8.755 + 1.313 \text{ weight}$$

- Estimated wing length when weight = 15g:
- Estimated wing length when weight = 16g:

Class activity

$$\widehat{\text{wing length}} = 8.755 + 1.313 \text{ weight}$$

- Estimated wing length when weight = 15g: 28.45 mm
- Estimated wing length when weight = 16g:

Class activity

$$\widehat{\text{wing length}} = 8.755 + 1.313 \text{ weight}$$

- Estimated wing length when weight = 15g: 28.45 mm
- Estimated wing length when weight = 16g: 29.763 mm

Class activity

$$\widehat{\text{wing length}} = 8.755 + 1.313 \text{ weight}$$

- Estimated wing length when weight = 15g: 28.45 mm
- Estimated wing length when weight = 16g: 29.763 mm

Change in estimated wing length when weight increases by 1g:

Class activity

$$\widehat{\text{wing length}} = 8.755 + 1.313 \text{ weight}$$

- Estimated wing length when weight = 15g: 28.45 mm
- Estimated wing length when weight = 16g: 29.763 mm

Change in estimated wing length when weight increases by 1g: 1.313 mm

- = $\widehat{\beta}_1$

Next time

- How we fit a regression line
- Reading: Section 1.1 in textbook