STAT115: Introduction to Biostatistics

University of Otago
Ōtākou Whakaihu Waka

Lecture 19: Introduction to Linear Regression

Outline

- Continue to explore relationships between two variables
- Go beyond summary statistics
  - Look into a statistical model for the relationship
    - What the model looks like
    - Fitted model
    - Residuals

Recall: motivating examples

- The size of brushtail possums
  - Compare total length (mm) to head length (cm)
- Height of STAT110 students
  - Compare father's height (cm) to son's height (cm)
- Squat weight of international power lifters
  - Comparing body weight (kg) to max squat weight (kg)

Recall: correlation

- The correlation r measures the strength of linear relationship between two variables x and y
- The correlation is limited
- What might we want to know?
  1. Possum data: predict head length from a measurement of total length
  2. Height data: understanding and quantifying heritability of height as a trait
  3. Powerlifting: compare the squat weight of an athlete to their peers of a similar weight
- Correlation does not help us for 1 and 3
  - Limited for 2: quantifies the linear relationship, but does not describe it
    - What is the expected difference in height between a son with father who is 170 cm tall, and a son with father who is 180 cm tall?

Statistical model

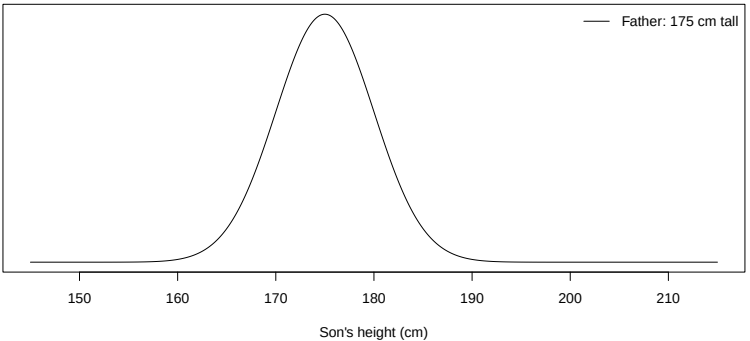
- To overcome these problems we will look to a statistical model
  - Extension of our previous models
- Explore relationship between continuous variables x and y
  - e.g. x is father's height, y is son's height
- The variable y is referred to as the outcome variable
  - Can also be called the response variable, or dependent variable
- The variable x is referred to as the predictor variable
  - Can also be called the explanatory variable, or independent variable
- The idea: the predictor variable helps us 'predict' the outcome variable

Statistical model

- Our description will make use of the father/son height example
  - Interest is in understanding the relationship the height of NZ male university students and their fathers
  - Sample is from (former) students in STAT110
- Using probability to describe data
- Recall concept of conditional probability: $Pr(A \mid B)$
  - Here we are looking at a probability density for $y|x$
    - We have the height of a father (x) and son (y)
    - Given a father's height (x), we specify a model for son's height (y)
    - We will specify a normal model
- Look at it graphically

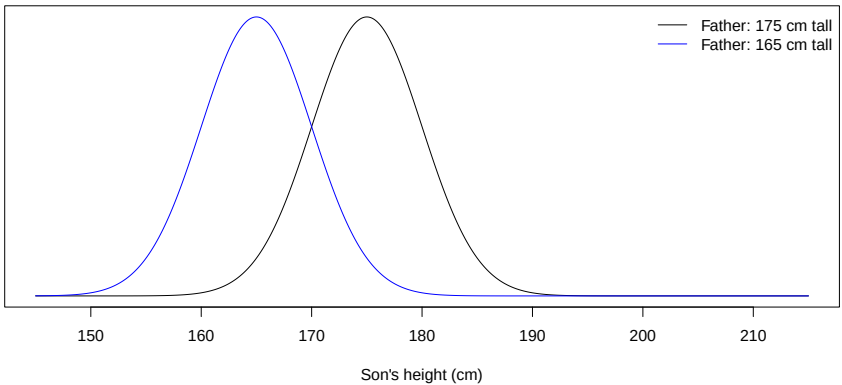
Statistical model

- Consider the subpopulation at particular value of x
  - e.g. sons with fathers who are 175 cm tall ($x = 175$)
  - Assume that son's height is normally distribution
    - For the sake of explanation: sons are expected to be the same height as their fathers



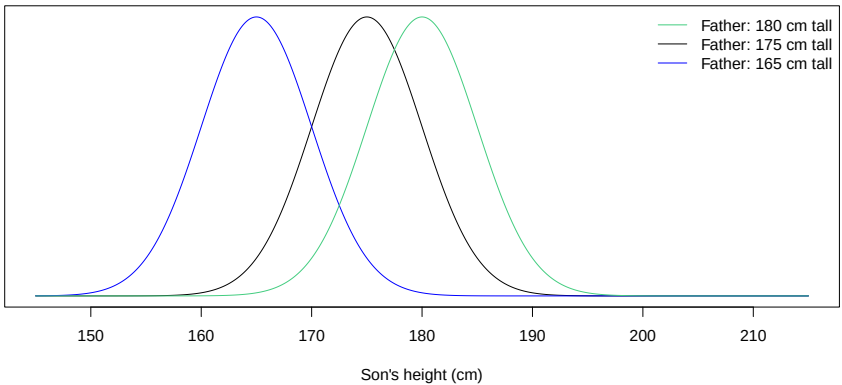
Statistical model

- Subpopulation at a given value of x: outcome variable is normally distributed
- For fathers who are 165 cm tall (blue)



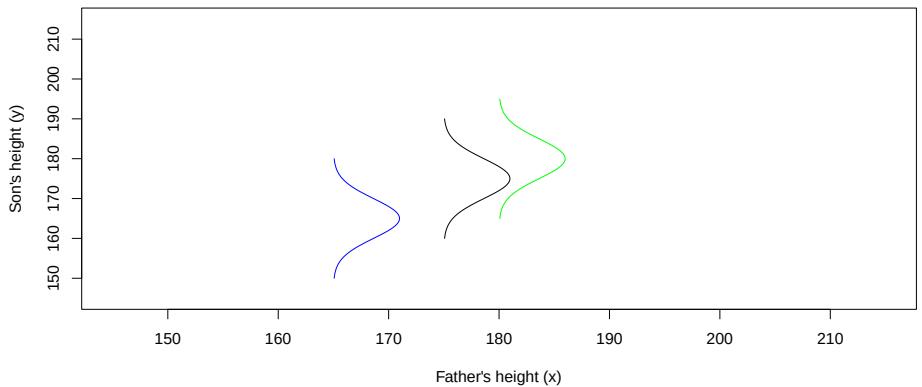
Statistical model

- Subpopulation at a given value of x: outcome variable is normally distributed
- For fathers who are 180 cm tall (green)



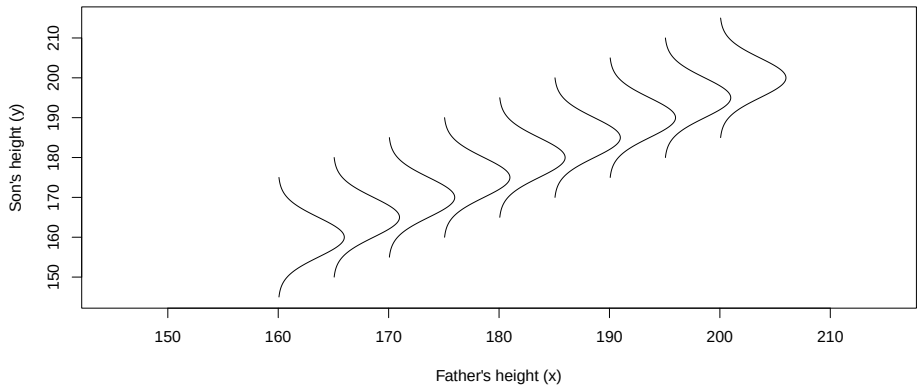
Turning it sideways

- Visualise it with outcome variable on y-axis, and predictor variable on x-axis
  - The same distributions are given below



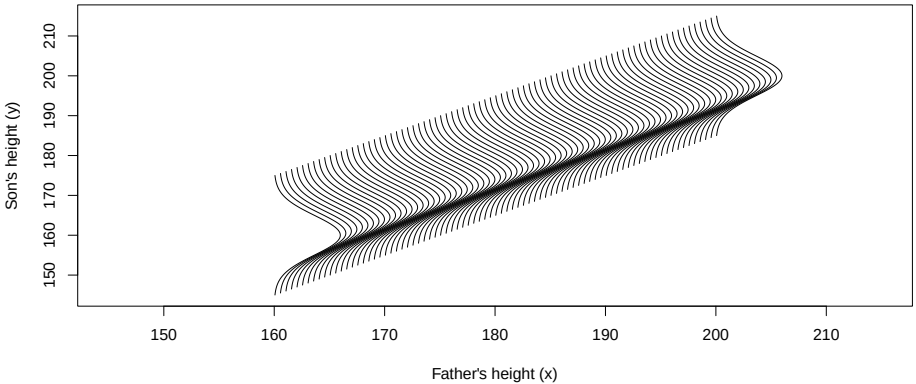
Turning it sideways

- Including some other values of x (father's height)



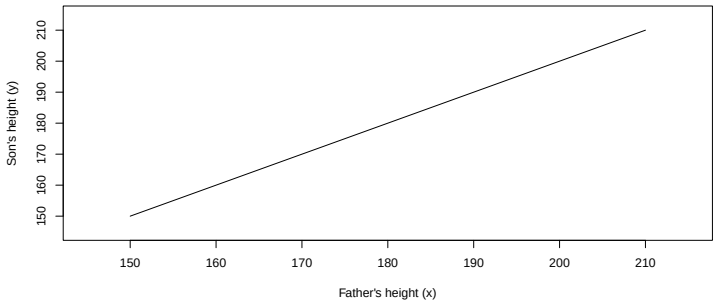
Turning it sideways

- Including even more values of x (father's height)



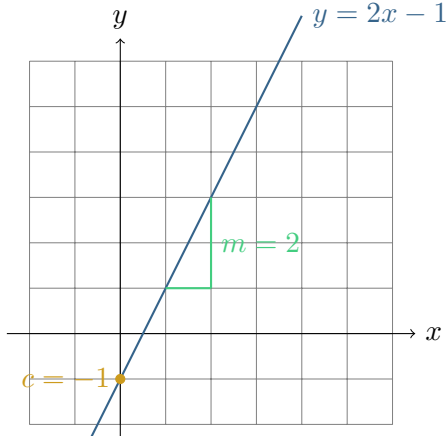
Linear regression

- The outcome variable, y, can be written in terms of two pieces:
  - outcome $=$ mean response $+$ error
- The mean response (what we expect) is assumed to vary with the predictor x
  - Expected height of a son is different if father is 165 cm vs father who is 180 cm
- We assume the mean response is a straight line
  - e.g. continuing the father and son height example, the mean response is



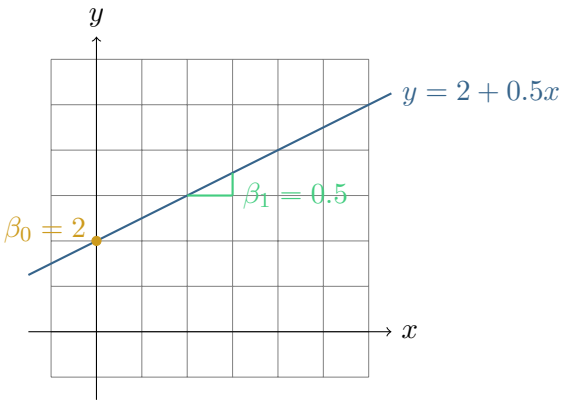
Revision: equation for a straight line

- Mathematical equation: $y = mx + c$
  - Intercept c: where it crosses the y-axis ($x = 0$)
  - Slope m



Revision: equation for a straight line

- We will use the equation: $\beta_0 + \beta_1 x$
  - Convention: use β_0 and β_1 in place of c and m
    - Intercept β_0: where it crosses the y-axis ($x = 0$)
    - Slope β_1



Understanding the model: population level

- Putting this together we have:

$$\underbrace{y}_{\text{outcome}} = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \underbrace{\varepsilon}_{\text{error}}$$

- The mean response is given by the straight line: $\mu_y = \beta_0 + \beta_1 x$
  - Gives us the expected value of y in the population for a given value of x
- The mean will be different for two different values of x
- For $x = 165$ cm:
  - Mean is: $\mu_y = \beta_0 + \beta_1 \times 165$
- For $x = 180$ cm:
  - Mean is: $\mu_y = \beta_0 + \beta_1 \times 180$

Interpretation

- What do β_0 and β_1 represent?
- The mean will be different for two different values of x
  - Mean is: $\mu_y = \beta_0 + \beta_1 x$
- For someone with a father one cm taller $(x + 1)$, the mean response is
  - Mean is: $\mu_y = \beta_0 + \beta_1(x + 1) = \beta_0 + \beta_1 x + \beta_1$
- β_1 is the difference between these
  - β_1 is the change in mean response when x increases by one unit
    - Change in the expected height of two male NZ university students whose fathers differ in height by 1 cm
- β_0 is the mean response when $x = 0$
  - May make no sense in many examples
    - Mean response for a son with a father of height 0 cm: physically impossible

From mean response to individual response

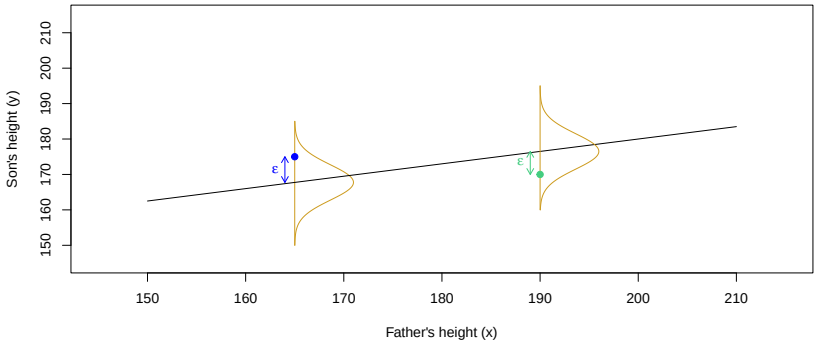
- The linear regression model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Error term ε (greek letter epsilon) describes how an individual response differs from the mean of their subpopulation
  - Subpopulation: all individuals in the population with the same value of x
- We assume that variation within a given subpopulation is normally distributed
  - ε is normally distributed with mean 0 and variance σ_ε^2
    - σ_ε tells us how variable individual observations are within their subpopulation

Visualising subpopulation

- Suppose that the true regression model for height is $y = 110 + 0.35x + \varepsilon$
  - Mean response (black line)
  - Normal model for the errors (gold)
  - Individual with $y = 175$ and $x = 165$ (blue point)
  - Individual with $y = 170$ and $x = 190$ (green point)

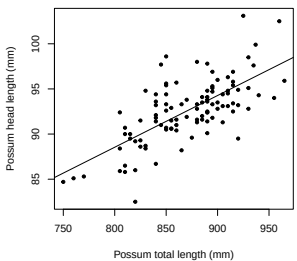


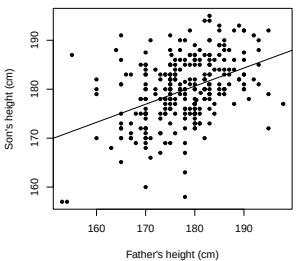
Statistical model: data

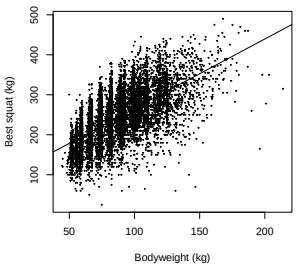
- The linear regression model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- The errors mean that data will not fall exactly on the line
  - Like the data we have!







It's quiz time!

- Suppose that the true regression model for height is

$$y = 110 + 0.35x + \varepsilon$$

- Decide whether the following statements are true or false:
  1. Consider the subpopulation of all students with fathers of height $x = 200$ cm. The mean height of those students is 180 cm.
  2. On average, students with fathers of height $x = 201$ cm are 0.35 cm taller than students with fathers of height $x = 200$cm.
  3. All students with fathers of height $x = 190$ cm are taller than all students with fathers of height $x = 170$ cm.
  4. Students with fathers of height $x = 0$ cm are 110 cm tall on average

Summary

- Introduced a statistical model for the relationship between x and y
  - Outcome variable, y
  - Predictor variable, x
  - For a given value of x, y is assumed to be normally distributed
- Understand the linear regression model
  - Mean response
  - Error
  - Interpretation
- Looking forward: how do we fit a linear regression to data?