Linear Model 1: A New Equation

Lecture 7

Dr Jennifer Mankin
7 March 2022

UNIVERSITY OF SUSSEX

Overview

- Reminder: the TAP!
- The Linear Model
  - What is modeling?
  - Model with continuous predictor
  - Model with categorical predictor

Reminder: The TAP

The **take-away paper** is currently live!

- See Take-Away Paper Information:
  - Download the Rmd document to complete
  - All information on preparing and submitting the assessment
  - All necessary background information, tips, and FAQs

Objectives

After this lecture you will understand:

- What a statistical model is and why they are useful
- The equation for a linear model with one predictor
  - b_0 (the intercept)
  - b_1 (the slope)
- Using the equation to predict an outcome
- How to read scatterplots and lines of best fit

The Linear Model

- Extremely common and fundamental testing paradigm
  - Predict the outcome y from one or more predictors (xs)
  - Our first (explicit) contact with statistical modeling
- A **statistical model** is a mathematical expression that captures the relationship between variables
  - All of our test statistics are actually models!

Maps as Models

- A map is a simplified depiction of the world
  - Captures the important elements (roads, cities, oceans, mountains)
  - *Doesn't* capture individual detail (where your gran lives)
- Depicts **relationships** between locations and geographical features
  - Helps you **predict** what you will encounter in the world
  - E.g. if you keep walking south eventually you'll fall in the sea!

Statistical Models

- A model is a simplified depiction of some relationship
  - We want to **predict** what will happen in the world
  - But the world is complex and full of noise (randomness)
- We can build a model to try to capture the important elements
  - Gather a sample that (we assume) is representative of the population
  - Investigate and quantify the relationships in that sample (ie construct a model)
  - Change/adjust the model to see what might happen with different parameters

Statistical Models

- **Why** might it be useful to create a model like this?
- Can you think of any recent examples of such models?
- One example of modelling you might all be familiar with!

Predictors and Outcomes

- Now we start assigning our variables roles to play
- The **outcome** is the variable we want to explain
  - Also called the dependent variable, or DV
- The **predictors** are variables that may have a relationship with the outcome
  - Also called the independent variable(s), or IV(s)
- We measure or manipulate the predictors, then quantify the systematic change in the outcome
  - NB: **YOU** (the researcher) assign these roles!

General Model Equation

$$outcome = model + error$$

- We can use models to **predict** the outcome for a particular case
- This is always subject to some degree of **error**

Linear Model Equation

$$y_i = b_0 + b_1 x_{1i} + e_i$$

- y_i: the predicted value of the outcome
- b_0: the intercept
- b_1: the slope
- x_{1i}: the predictor
- e_i: the error in prediction

You may know her as \`$y = ax + b$\`!

Linear Model Equation

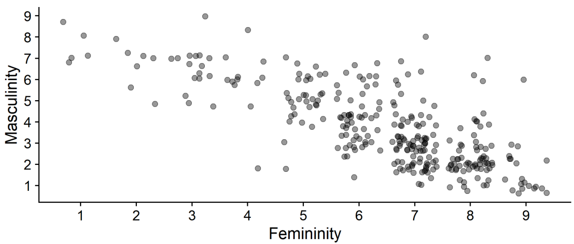
$$y_i = b_0 + b_1 x_{1i} + e_i$$

- We will next see:
  - How we can create a line that captures the relationship between those two variables
  - How we can adapt this general LM equation to describe that line

Visualising the Line

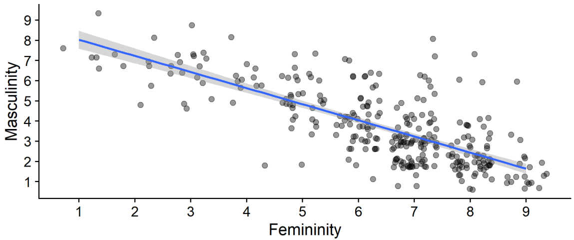
- Where would you draw a line through these dots that best captures where they tend to fall?

Plot | R code



Visualising the Line

Plot | R code



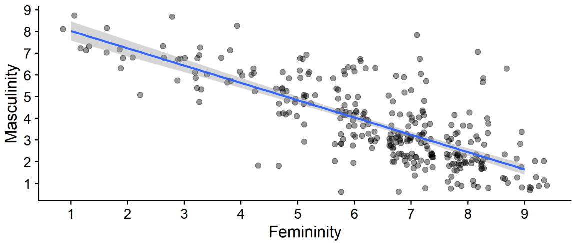
Visualising the Line

- The data points tend to be higher up on the right and lower down on the left
  - So as the variable on *x* (here, ratings of femininity) increases...
  - The variable on *y* (here, ratings of masculinity) tends to decrease
  - This represents a **negative relationship** between *x* and *y*: as one goes up, the other goes down
- Our line captures this by going downwards from left to right

Visualising the Line

- Two key *parameters*: where the line starts, and its slope

Plot | R code



Modeling Gender Ratings

We can make some estimates:

- The line would cross the *y*-axis somewhere between 8 and 9 (close to 9)
  - $b_0 \approx 8.5$
- Every time we go up one point on the femininity scale, masculinity goes down by a little less than one point
  - $b_1 \approx -0.8$

Modeling Gender Ratings

$$y_i = b_0 + b_1 x_{1i} + e_i$$

- y_i (outcome): Masculinity
- x_{1i} (predictor): Femininity
- b_0 (intercept): the predicted value of masculinity when femininity is 0
- b_1 (slope): **change** in masculinity associated with a **unit change** in femininity

$$Masculinity_i = b_0 + b_1 Femininity_{1i} + e_i$$

Modeling Gender Ratings

How do we get the real numbers?

```
##
## Call:
## lm(formula = gender_masc ~ gender_fem, data = gensex)
##
## Coefficients:
## (Intercept)   gender_fem
##      8.8246      -0.7976
```

Adapt our equation to include the real *b* values:

$$Masculinity_i = 8.82 - 0.8 \times Femininity_{1i} + e_i$$

Predicting Gender

- We can now use this model to **predict** someone's rating of masculinity, if we know their rating of femininity
    - someone who doesn't identify strongly with femininity: `gender_fem` = 3
    - What would the model **predict** for this person's masculinity rating?

$$Masculinity_i = 8.82 - 0.8 \times Femininity$$

Predicting Gender

$Masculinity_i = 8.82 - 0.8 \times Femininity_{1i}$

- $Masculinity_i = 8.82 - 0.8 \times 3$
- $Masculinity_i = 6.42$

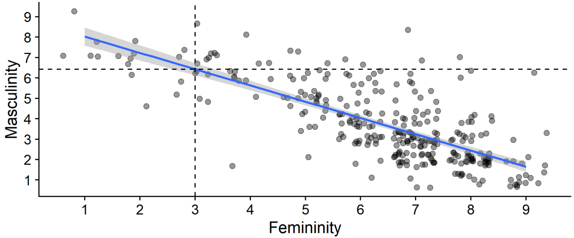
So, someone with femininity = 3 is **predicted** to have a masculinity rating of 6.42

- This is subject to some (unknowable!) degree of error

Predicting Gender

Someone with a femininity rating of 3 is **predicted** to have a masculinity rating of 6.42

Plot | R code



Interim Summary

- The linear model predicts the outcome *y* based on a predictor *x*
  - General form: $y_i = b_0 + b_1 x_{1i} + e_i$
  - b_0, the intercept, is the value of *y* when *x* is 0
  - b_1, the slope, is the change in *y* for every unit change in *x*
- The slope, b_1, is the key piece of information, because it represents the relationship between the predictor and the outcome
- Up next: categorical predictors

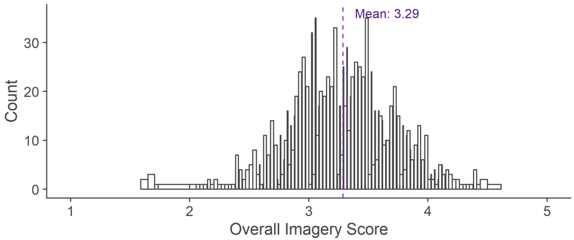
Words and Colours

In Tutorial 5, we looked at synaesthesia and imagery

- Let's revisit those ideas using the linear model!
- If I wanted to **predict** the next random person's overall imagery score...
  - What would be the most sensible *estimate*?

Making Predictions

Plot | R code



Making Predictions

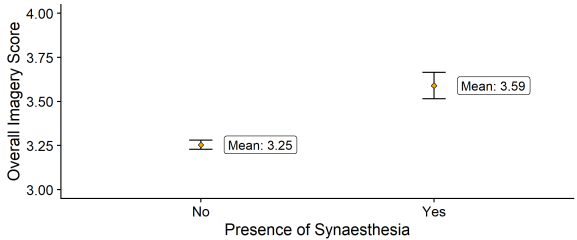
- Without any other information, the best estimate is the mean of the outcome
  - But we *do* have more information!
- Grapheme-colour synaesthetes score higher than non-synaesthetes on overall imagery on average
  - We could make a better **prediction** if we knew whether that person was a synaesthete
  - Use the mean score in the synaesthete vs non-synaesthete groups

Modeling Imagery

For non-synaesthetes, mean overall imagery = 3.25

- We will treat them as the **baseline** and give them a group code of 0

Plot | R code

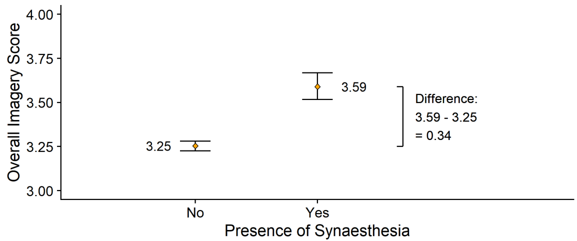


Modeling Imagery

For synaesthetes, mean overall imagery = 3.59

- We will treat them as the **comparison** group and give them a group code of 1

Plot | R code



Modeling Imagery

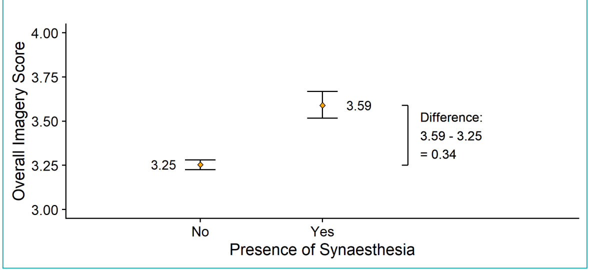
We want to write an equation that will give a different prediction depending on whether someone is a synaesthete or not

- $y_i = b_0 + b_1 x_{1i} + e_i$
  - y = Overall imagery score
  - x_1 = Synaesthesia (0 = No, 1 = Yes)
- $OverallImagery_i = b_0 + b_1 Syn_{1i}$
- How do we find out b_0 and b_1?

Estimating the Line

- Where would you draw a line through these dots that best captures where they tend to fall?

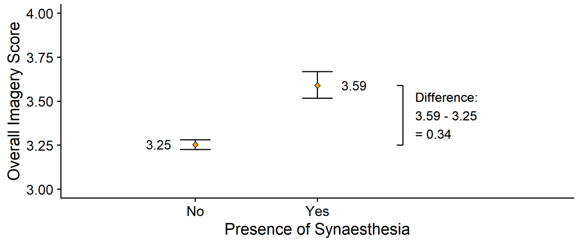
Plot | R code



Estimating the Line

This line is our **linear model**, with the same properties as the last one!

Plot | R code



Modeling Imagery

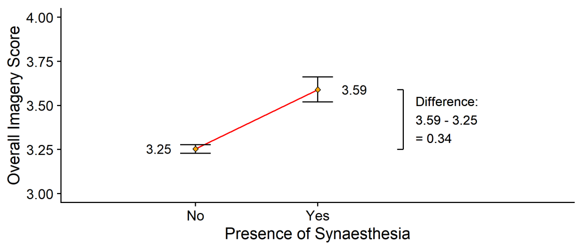
- The line starts from the **mean** of the non-synaesthete group = 3.25
  - This is the **intercept**, b_0
  - The predicted value of the outcome when the predictor is 0
  - Our predictor is `syn` group, where no synaesthesia = 0
- When we switch from looking at non-synaesthetes to synaesthetes, predicted overall imagery changes by 0.34
  - This is the **slope** of the line, b_1
  - The change in the outcome for every **unit change** in the predictor
  - Here, a "unit change" means switching groups, from 0 (non-syn) to 1 (syn)

$$OverallImagery_i = 3.25 + 0.34 \times Syn_{1i}$$

Using `lm()`

```
##
## Call:
## lm(formula = overall_img ~ syn, data = syn_data)
##
## Coefficients:
## (Intercept)       synYes
##      3.2539       0.3361
```

Plot | R code



Checking Predictions

If I wanted to **predict** the next random person's overall imagery score...

- First, ask them if they're a synaesthete or not!
- "Yes" = 1, "No" = 0

$$OverallImagery_i = 3.25 + 0.34 \times Syn_{1i}$$

If yes, then $Syn_{1i} = 1$:

- $OverallImagery_i = 3.25 + 0.34 \times 1$
- $OverallImagery_i = 3.59$

If no, then $Syn_{1i} = 0$:

- $OverallImagery_i = 3.25 + 0.34 \times 0$
- $OverallImagery_i = 3.25$

So, we can predict imagery score based on group membership, just as we predicted masculinity score based on femininity score earlier!

Welcome to the World of `lm()`

- The **l**inear **m**odel (`lm()`) will be our focus from here on out
  - If this is unfamiliar to you, it's **highly recommended** that you revise linear equations!
  - Visualisation on the Analysing Data website
  - Khan Academy intro to linear equations
  - Learning Statistics with R - see Chapter V, Linear Regression
- Linear models will be crucial for **the rest of your degree**

Summary

- The linear model expressed the relationship between at least one predictor, *x*, and an outcome, *y*
    - Linear model equation: $y_i = b_0 + b_1 x_{1i} + e_i$
    - Key for statistical testing is the parameter b_1, with expresses the relationship between *x* and *y*

  Used to **predict** the outcome for a given value of the predictor
-
- Next week: LM2 - significance and model fit
- Don't forget to do the TAP!