University of East Anglia

BIO-5023YB
2020
Spring term – week 2
Introduction to statistics

Dr Philip Leftwich – p.leftwich@uea.ac.uk

Introduction to statistics

Learning outcomes

- Understand model fitting

- Most statistical models we require can be fitted using linear models and the least squares approach

- Understand significance and the null hypothesis

Week 2: Intro to Statistical Modelling

Everything you ever wanted to know about statistical models

But were afraid to ask!

- As scientists we want to be able to explain phenomena

  - The links between diet and ageing
  - The role of DNA methylation in cancer
  - Foraging behaviour in passerine birds
  - Why don't undergrads start their coursework on time?

To do this, we collect data, build statistical models and test hypotheses

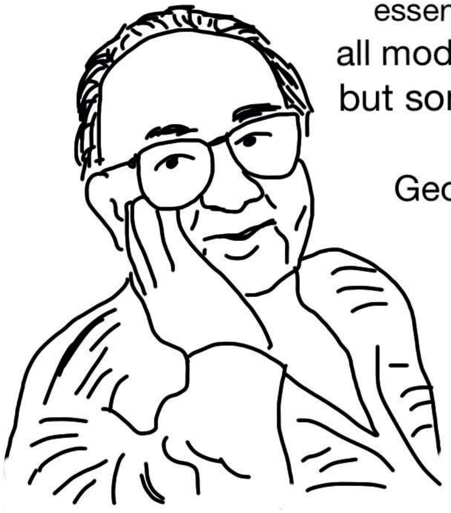
Models don’t have to be complicated

Representing a range of values by a mean and standard deviation is a simple model

Models are useful e.g. if we want to know what the average potato weighs we could try to weigh every potato in existence

BUT if we weigh a sample of potatoes (our mean and s.d. will act as a model for the average potato weight).

Models are always wrong



essentially,
all models are wrong,
but some are useful

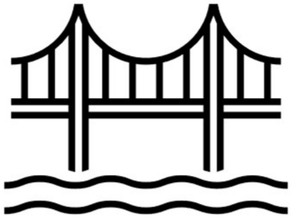
George E. P. Box

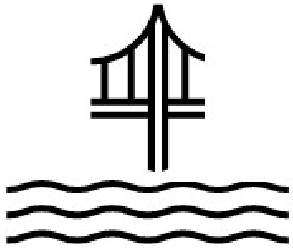
A model cannot be perfect – it is by definition an approximation.

But it doesn't matter as long as it is *good enough* to be useful.

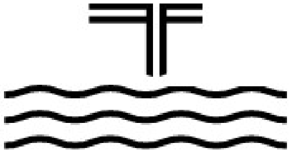
Real world

Model 1
Good fit





Model 2
Moderate fit



Model 3
Poor fit

Models are always wrong

- Fit

What is explained by the model

- Residuals

Variance which is *not* explained by the model

Model fitting

Model 1 was a *good fit* there are differences between it and the real world example – but it is basically a good replica

Model 2 – has some similarities, contains the basic structures, but there are also some big differences (missing one support tower) – this is a *moderate fit.* A model with moderate fit might get the broad trends right but fail to make accurate predictions

Model 3 – is a *poor fit* it is obviously a massive over-simplification and would be likely to completely inaccurate

Models are a way of testing hypotheses

A Model that is a good fit for data allows us to test our hypotheses

- Make an observation
- Form a hypothesis
- Collect data
- Test our hypothesis with a statistical model

Making a hypothesis

What are the important parts of a good hypothesis?

Making a hypothesis

Turn this research question into a hypothesis

“I think plant inbreeding is deleterious to their fitness, inbred plants have a stunted growth”

What would the Null hypothesis be?

Our statistical models test what the likelihood of observing a result is “*if the Null Hypothesis were true”*

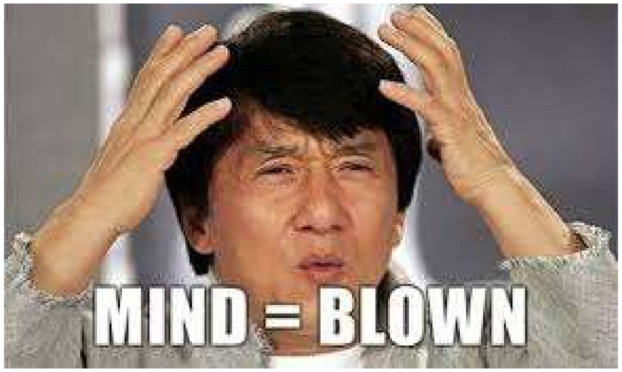
(General) Linear models

Almost all of the models that we will require to test & describe the data we encounter are going to be variations of *linear models*

You will here a lot about various types of statistical tests/models (t-test, ANOVA, ANCOVA, regression etc.) These are in fact **all just identical systems based on linear models**

A linear model is simply a model that is based upon a straight line

A *general* linear model includes the error as well as the fit for the model



(General) Linear models

So… If you can fit a straight line to data in order to explain differences or associations – then you can use a general linear model.

There are several assumptions we make when using linear models, and over the next few weeks we will learn what these assumptions *are* and how to test whether your data and model meet these assumptions, but today we learn the most important assumption

Assumption 1:
There is a linear relationship between your dependent ~ independent variable

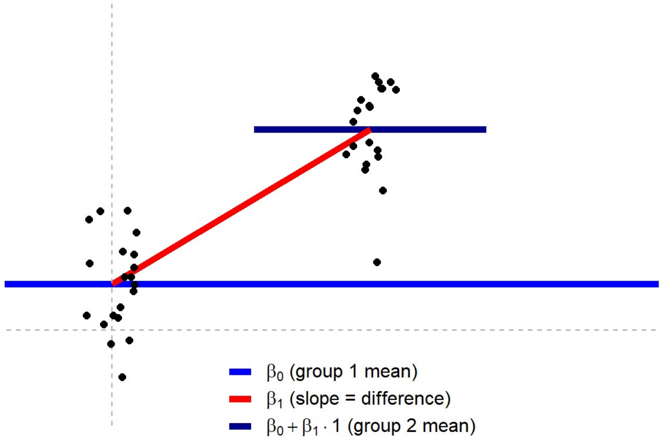
What is the equation for a straight line?

In the equation of the straight line $y = mx + c$
what do m and c stand for?

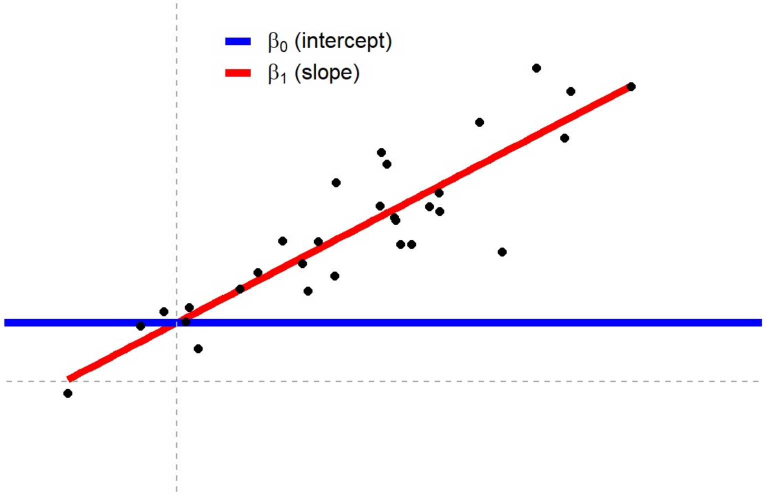
a) c is the gradient(slope) and m is the y-intercept
b) m is the x-intercept and c is the y-intercept
c) m is the gradient(slope) and c is the y-intercept
d) m is the gradient(slope) and c is the x-intercept

Least squares

The general strategy of the linear model is to quantify the overall variability in the data set and to **produce a straight line equation** which has the smallest amount of overall difference *between* the data points and the mapped line. **The intercept and slope are the fit of the model – which we use to explain and predict**



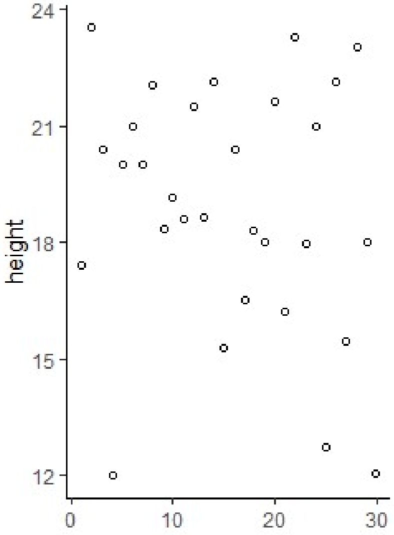
Linear model for the difference in mean between two or more categories - ANOVA

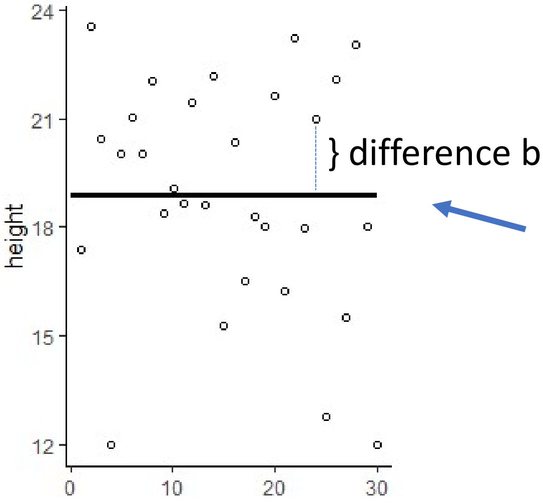


Linear model for the relationship between two continuous variables - regression

Calculating least squares

To start calculating least squares we start by measuring the differences from the individual data points to some reference point (this might be a regression line or the mean of a categorical group).





} difference between datapoint and mean

Intercept =
Overall mean

Calculating least squares

So the line representing the mean is our model and the distances between each data point and the model are the residual differences.

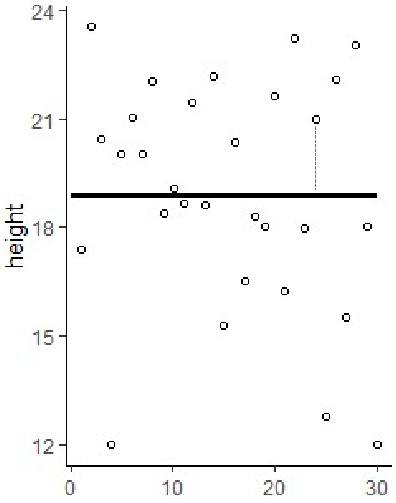
If we had a normally distributed data set then:

total sum of deviance =
(0.4) + (1.4) + (0.4) + (-1.6) + (-0.6) = 0

So there is no residual difference in our model???

How do we avoid this? **SUM OF SQUARES**

$(0.4)^2 + (1.4)^2 + (0.4)^2 + (-1.6)^2 + (-0.6)^2 =$ **5.2**



Calculating least squares

Sum of square is a good measure of the *accuracy* of our model.

BUT there are two issues

- it is squared from the original data (so not on the same scale)

- Sum of squares will *keep* going up as the sample size increases

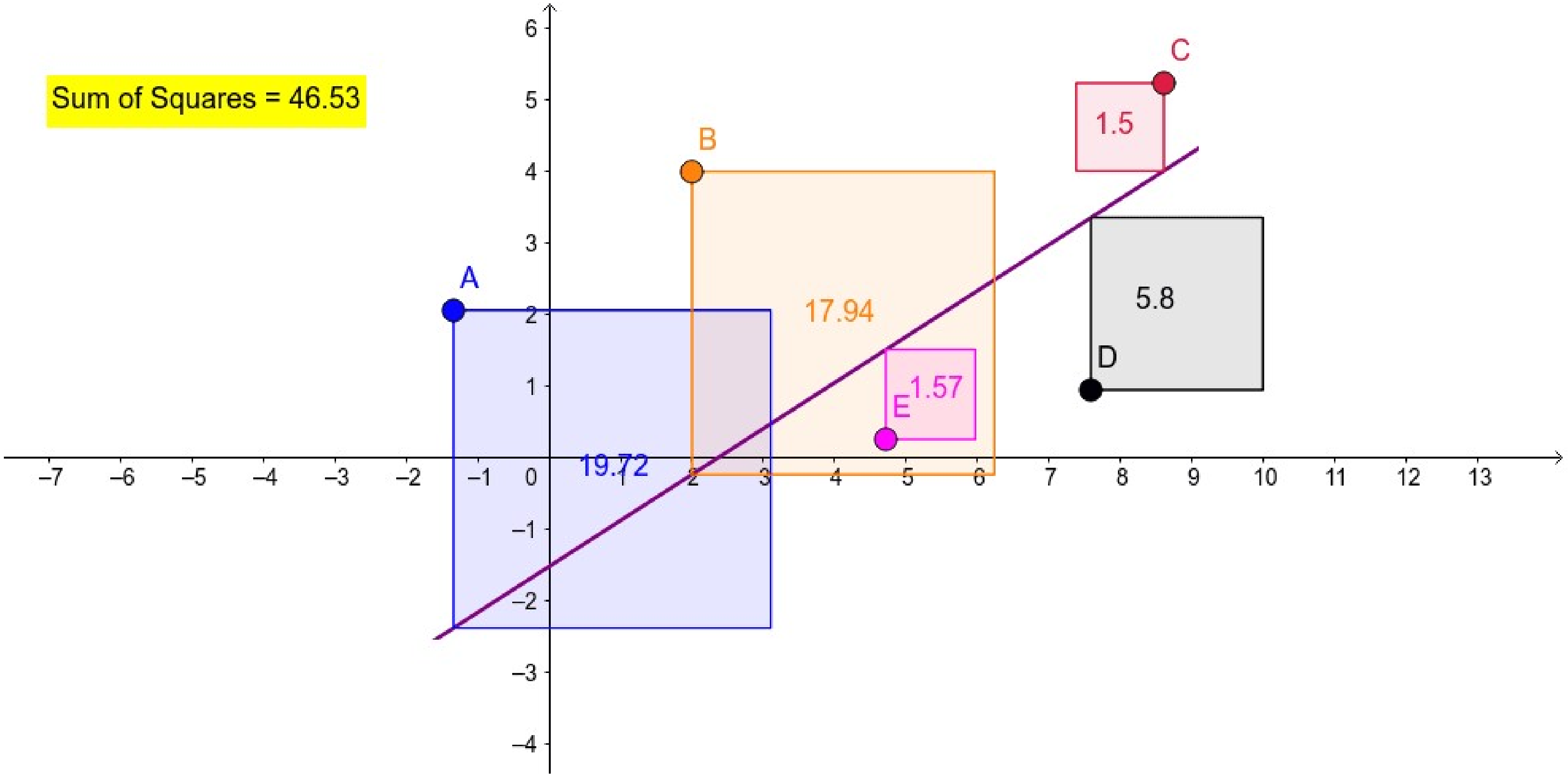
So if we divide it by the degrees of freedom & square root the result $= \sqrt{\frac{SS}{N-1}}$

We get some sort of <u>standard deviation!!!</u> = 1.14

To understand how least squares could calculate a mean – the intercept will be placed wherever the **least squares** are required e.g. the point which produces the smallest s.d.

Least squares for regression

This least squares approach is also how we calculate a regression fit – where can we put the line through our data to produce the least squares value and minimize residuals while maximizing fit.



Least squares for a difference model

This least squares approach is also how we calculate a regression fit – where can we put the line through our data to produce the least squares value.

But what about for differences e.g ANOVA?

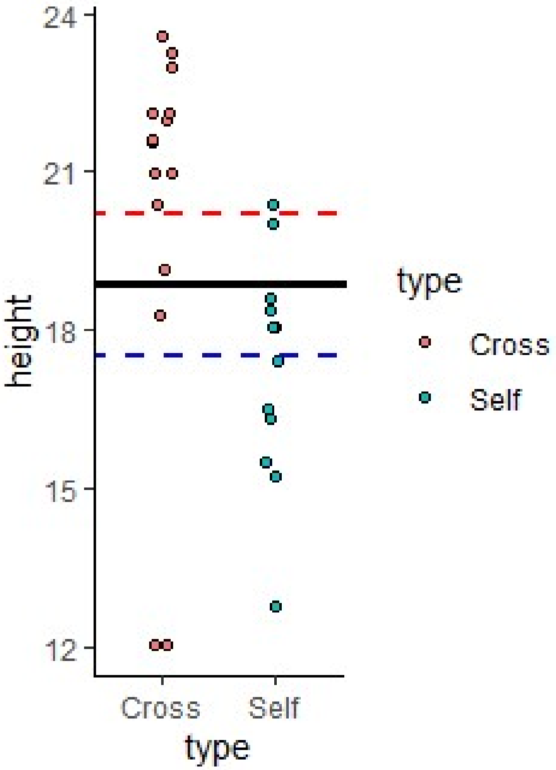
Here we calculate 3 different squares

SST = Sum of squares total e.g. what we calculated before

SSA = sum of squares between treatment means (e.g. red and blue line)

SSE = sum of squares of errors e.g. sum of squares between points within a group to the group mean

This allows us to compare squares within and between groups.



Least squares for a difference model

This least squares approach is also how we calculate a regression fit – where can we put the line through our data to produce the least squares value.

But what about for differences

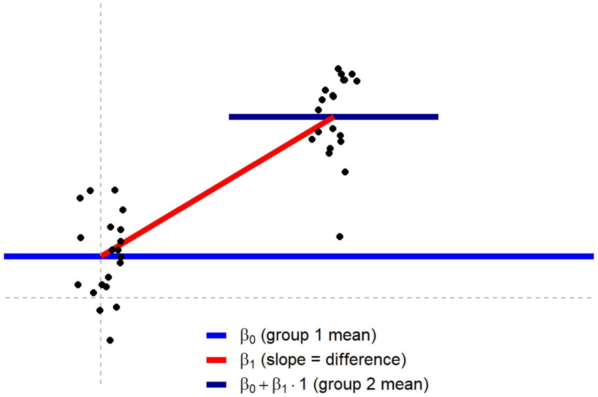
Here we calculate 3 different sum of squares

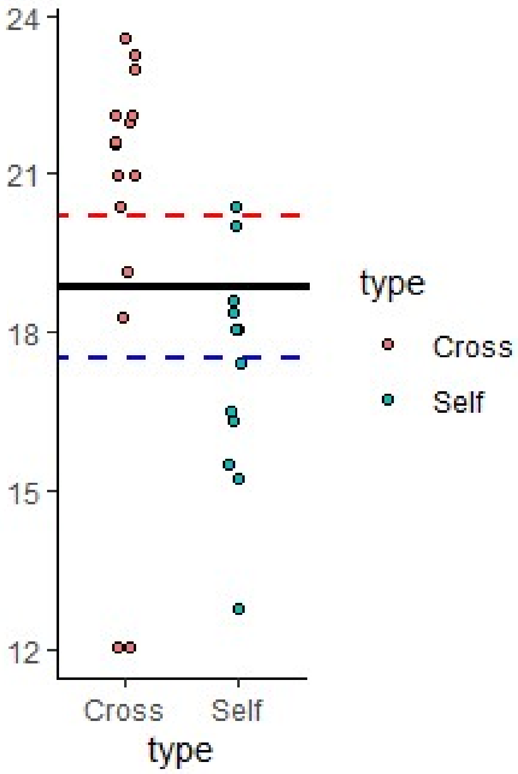
One of our groups is *arbitrarily* chosen to be the Intercept (usually alphabetical).

SST – value of total squares to the intercept

SS – treatment squares (SSA, SSB etc.) squared difference between treatment mean and intercept

SSE – error squares – squared difference between Individual data points and hypothetical treatment mean





F-value

Divide the treatment variance by the error variance

e.g. SS / SSE

The more *signal* there is compare to *noise* the higher the F-value will be.

In our example it is *F =* 5.94

In an ANOVA table we can input F and the degrees of freedom (29)

Note in an ANOVA table you get 1 df used for *each* non-intercept treatment (here *1*) and “the rest” so $F_{1,28}$

if we looked at four groups it would be $F_{3,26}$

With F and df we can look up our P-value

What is P?

What is P?

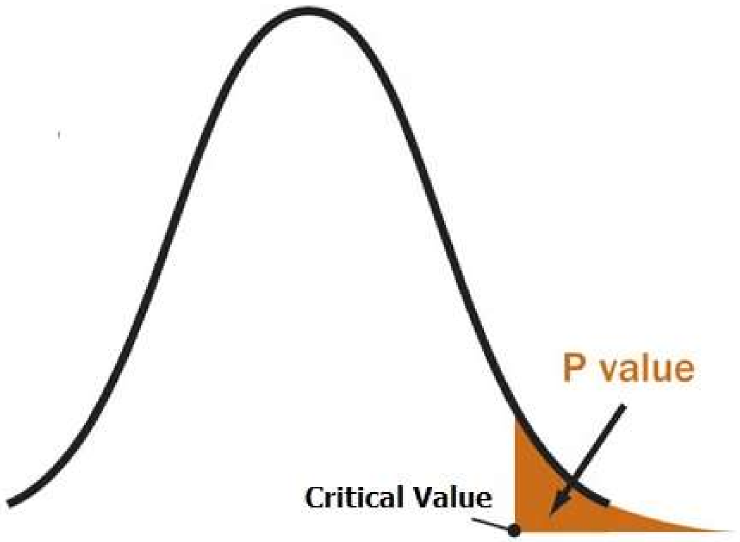
P-value

Simply put – if our *Null Hypothesis* is true, what is the probability of observing this F-value at this sample size?

>5% is our critical threshold for *rejecting* the Null Hypothesis.

Type I error – rejecting our Null Hypothesis incorrectly
(false positive)

Type II error – accepting the Null Hypothesis when it is false
(false negative)



What next?

Your next workshop will be to work through the plant crossing data and apply a linear model for yourself – calculate F, understand the slope and write up your results.

We will see how our linear model provides not just a test of significance, but important biological data on the relationships between these groups.