Regression Modelling in R

Sonja N. Tang and Fernando Guntoro

Imperial College London

Contents

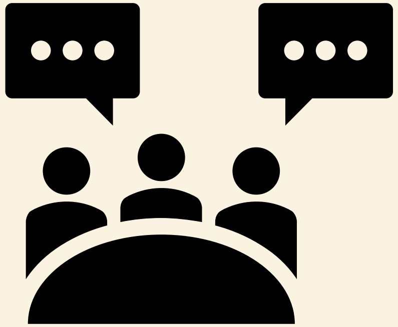
Introductions

- **Name:** Sonja
- **Origin:** Vienna, Austria
- **Department:** Epidemiology and Biostatistics and MRC Centre for Environment and Health
- **PhD topic:** Causal networks between metabolites
- **Favourite movie:** Everything Everywhere All At Once

- **Name:** Fernando
- **Origin:** Jakarta, Indonesia
- **Department:** Epidemiology and Biostatistics and Infectious Disease Epidemiology
- **PhD topic:** Multi-omics analysis of COVID-19 severity and long COVID
- **Favourite movie:** Hacksaw Ridge

Introductions

- Name
- Department
- PhD topic
- Favourite movie?



Head to menti.com

Code: **2575 5157**

Learning outcomes

Imperial College London

1. **Identify** the correlation coefficient as a single measure of linear association.
2. **Apply** general linear models to model a response variable in terms of a single or multiple variables.
3. **Evaluate** model fitness by comparing the results produced by the model with your data.
4. **Present** model fitness using data visualisation techniques.
5. **Interpret** regression model results from scientific papers.

Table of contents

1. Theory (~45 min)

- Background
- Linear regression (Least squares method, R^2, p-values)
- Logistic regression

Break (10 min)

2. Practical (~60 min)

3. Interpreting a study (~30 min)

Main idea of regression modelling

The problem: We have loads of data and we want to **describe the relationship**.

A solution: We build a **regression model**. There are many regression models. Today we're focussing on:

A. Linear regression

B. Logistic regression

Weight	Height
0.4	1.1
1.2	1.9
1.9	1.7
2.0	2.8
2.8	2.3

What is regression modelling?

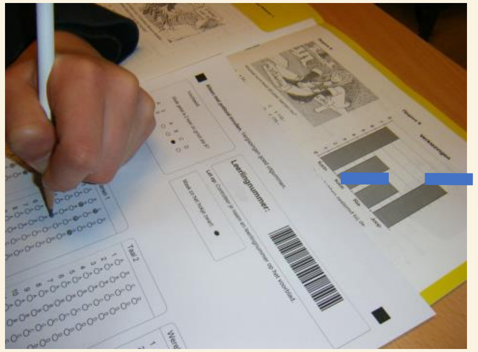
In statistics, regression modelling is a process for **estimating** a **line** or **curve** that **best represents the general trend** between one **dependent variable (Y)** and one or more **independent variables (X).**

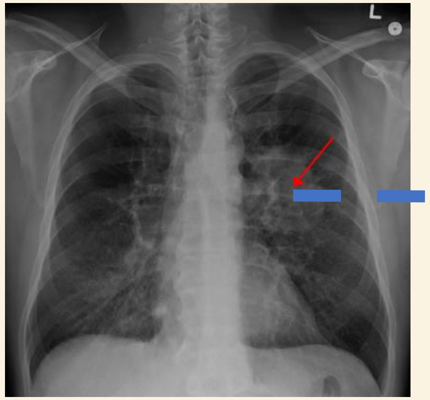
'outcome', 'response', 'label'

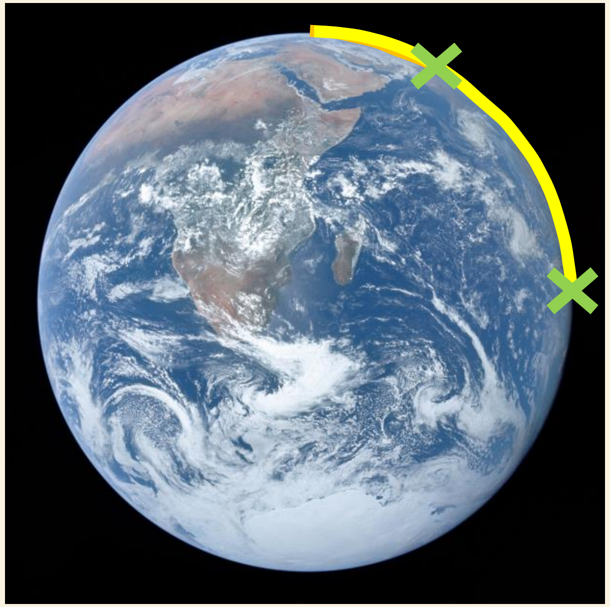
'predictors', 'covariates', 'explanatory variables', 'features'

Imperial College London

When do we need regression modelling?



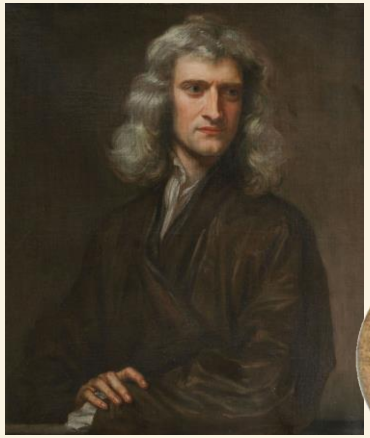




When do we need regression modelling?

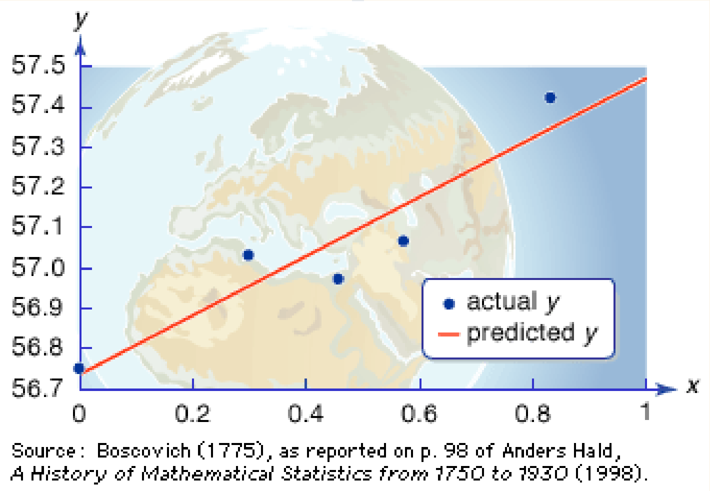
Imperial College London

Jacques Cassini
(1677-1756)



Isaac Newton
(1643-1727)

Arc length in Paris toise
(roughly 1.9m)



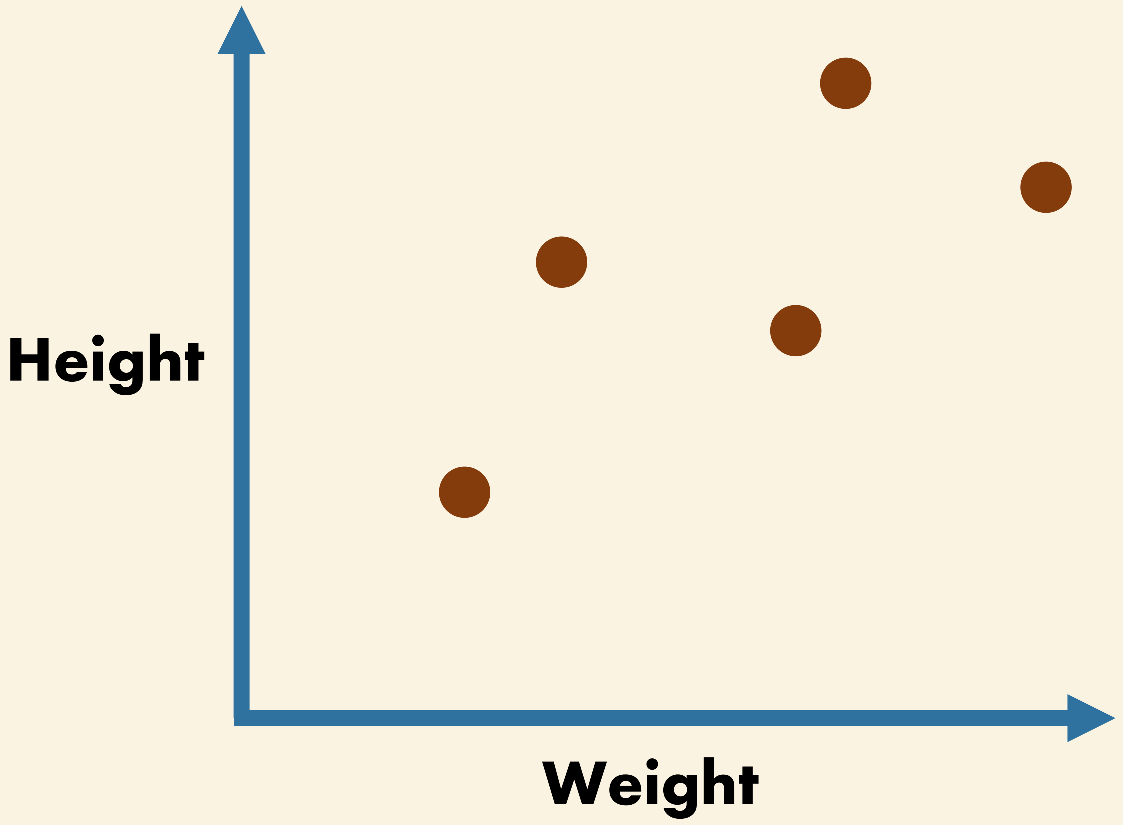
Lattitude in degrees

Linear regression using the least squares method

Imperial College London

Dependent variable (Y) and independent variable (X)

Height	Weight
1.1	0.4
1.9	1.2
1.7	1.9
2.8	2.0
2.3	2.8



Linear regression

1. Use **least-squares** to fit a line to the data.
2. Calculate **R^2.**
3. Calculate a ***p*-value** for R^2.

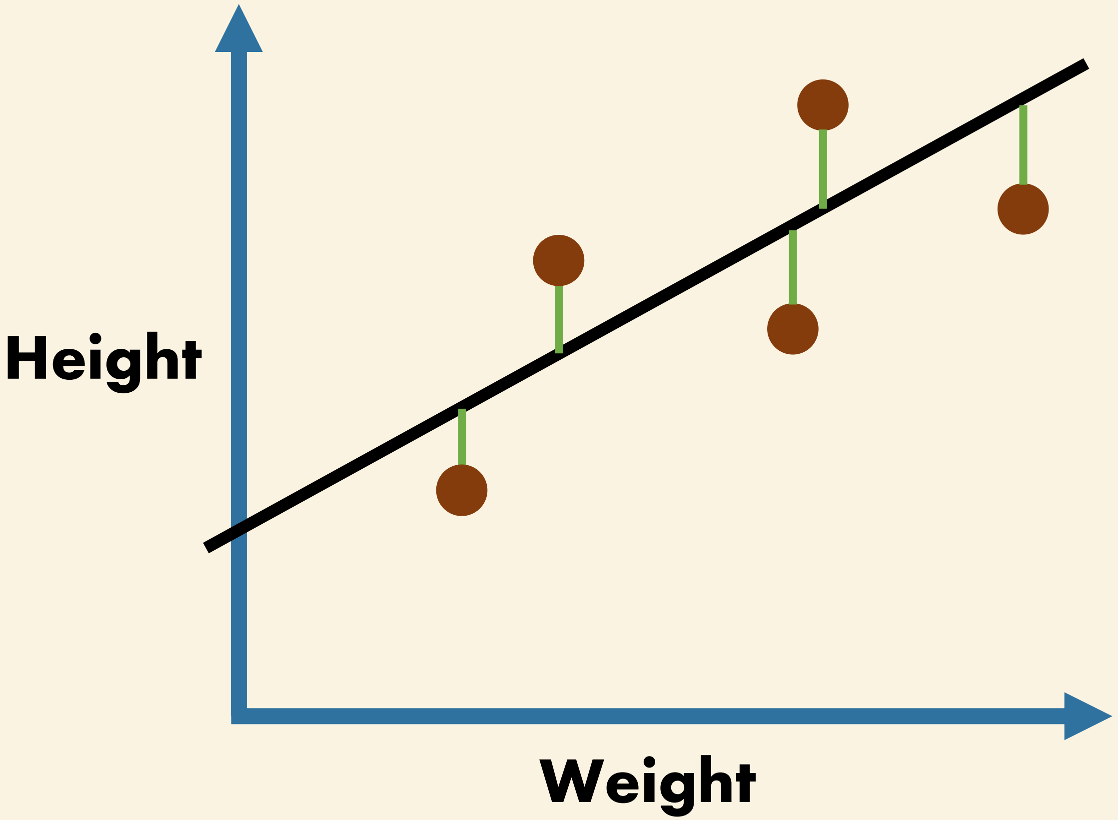
Linear regression

1. Use **least-squares** to fit a line to the data.
2. Calculate the **R^2**.
3. Calculate a ***p*-value** for R^2.

Least-squares minimises the
Sum of the Squared Residuals (SSR)

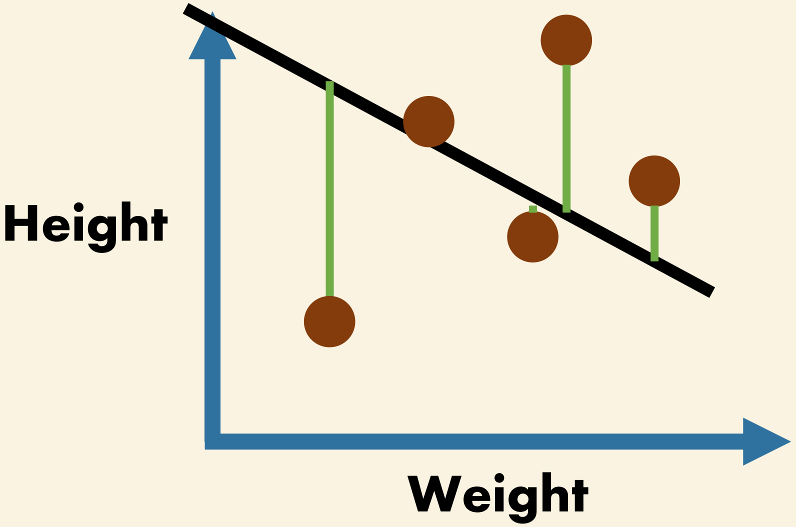
Residual = Observed - Fitted

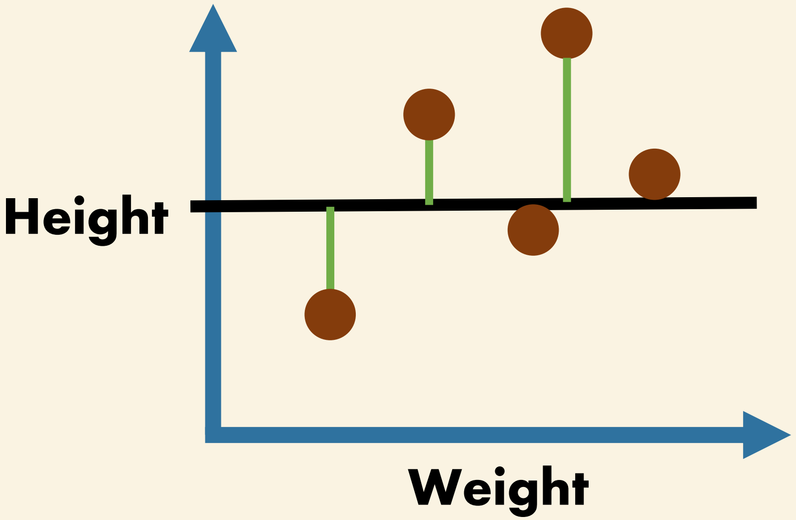
$$SSR = \sum_{i=1}^{n} (Observed_i - Fitted_i)^2$$

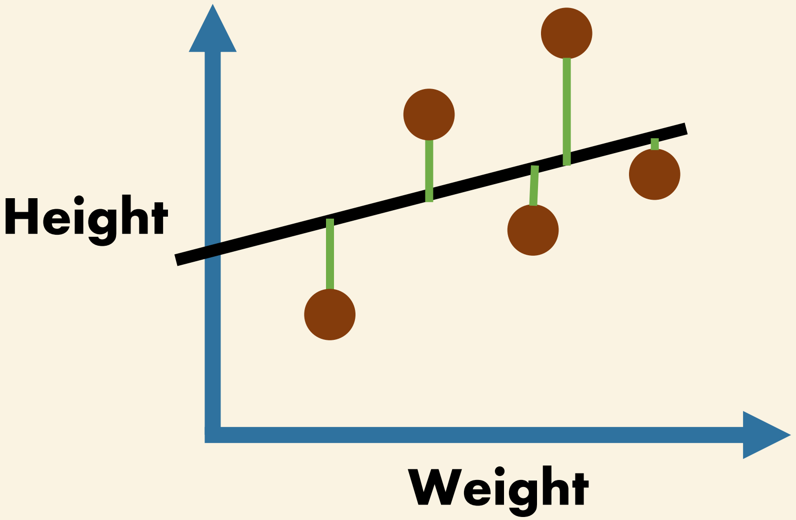


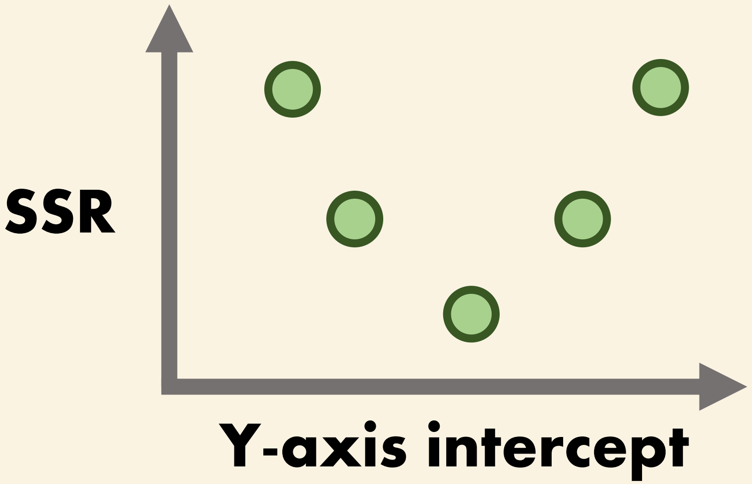
Linear regression

1. Use **least-squares** to fit a line to the data.
2. Calculate **R^2**.
3. Calculate a ***p*-value** for R^2.









y = mx + b, where
y = how far up
x = how far along
m = slope or gradient
b = the y-intercept

Height = slope x Weight + intercept

Height = 0.5 x Weight + 1.1

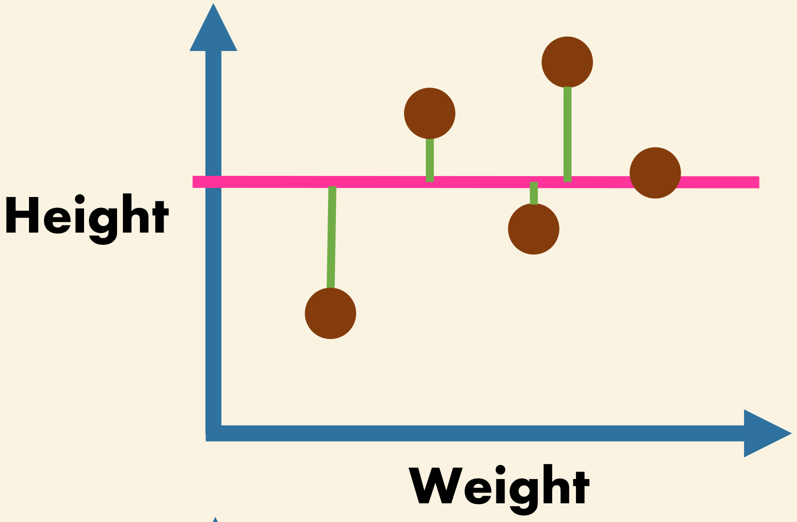
Linear regression

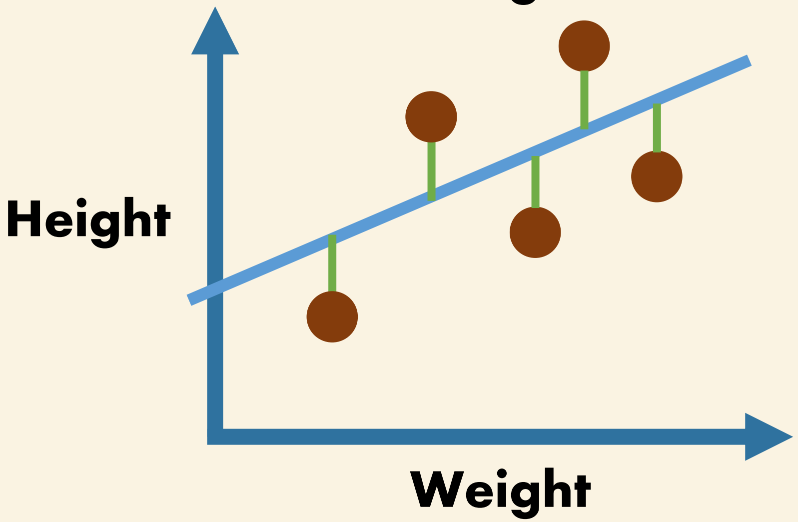
1. Use **least-squares** to fit a line to the data.
2. Calculate the **R^2.**
3. Calculate a ***p*-value** for R^2.

R^2 is the proportion of the variation in the dependent variable that is explained by the independent variable.

$$R^2 = \frac{SSR(mean) - SSR(fitted\ line)}{SSR(mean)}$$

$$R^2 = \frac{1.61 - 0.55}{1.61} = 0.66$$



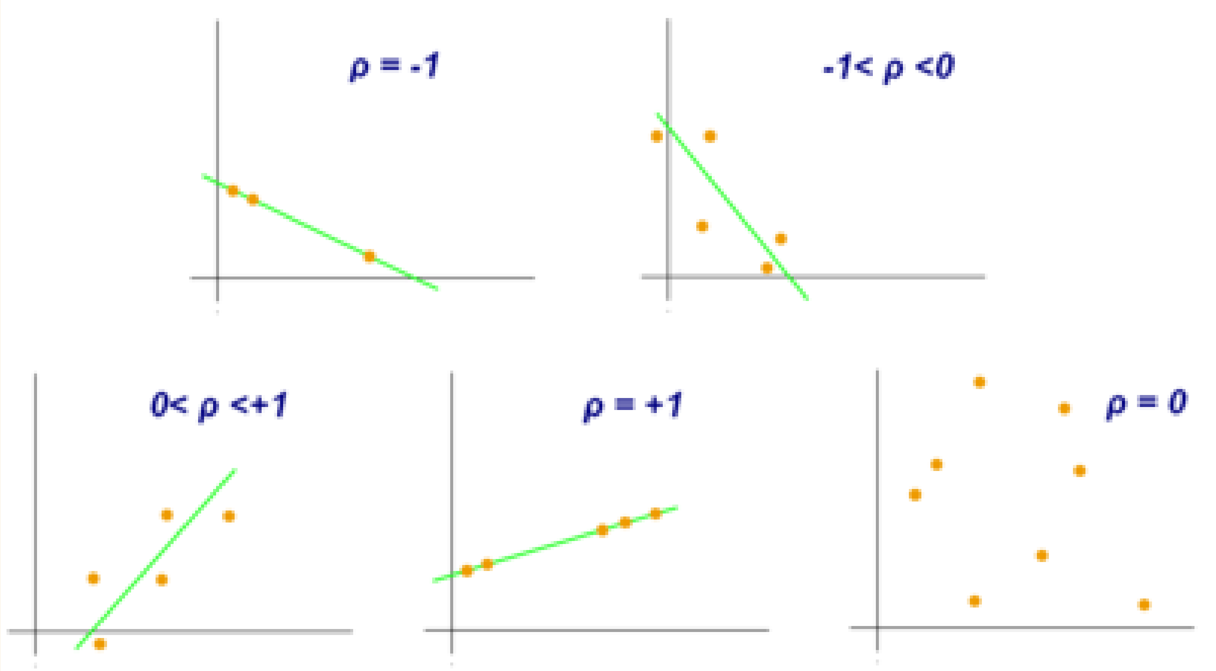


Pearson correlation coefficient (ρ)

The Pearson correlation coefficient (ρ, or rho) is the measure of **linear correlation** between two sets of data.

$\rho = r$

$\rho^2 = r^2 = R^2$

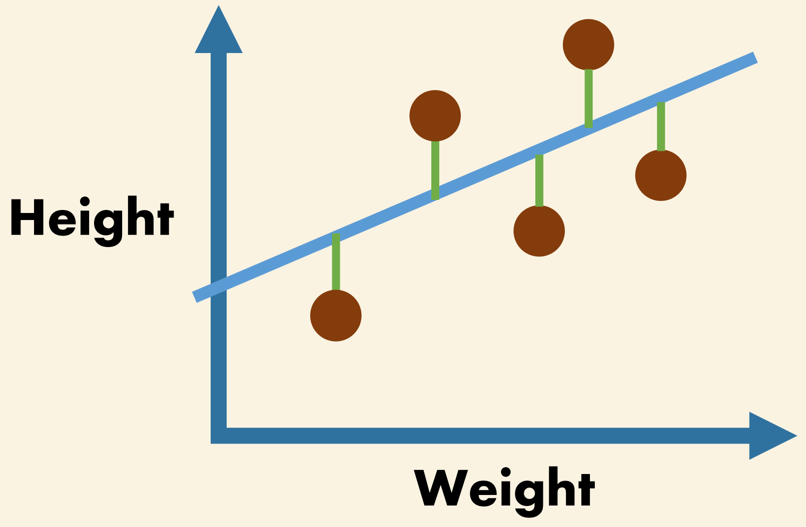


Linear regression

1. Use **least-squares** to fit a line to the data.
2. Calculate the **R^2.**
3. Calculate a ***p*-value** for R^2.

The *p*-value for our R^2 tells us the probability that random data could result in a similar or better R^2.

In general, *p*-values below 0.05 give us a large confidence in the results of our analysis.



Height = 0.5 x Weight + 1.1
R^2 = 0.66

p-value = 0.1

One dependent variable and multiple independent variables

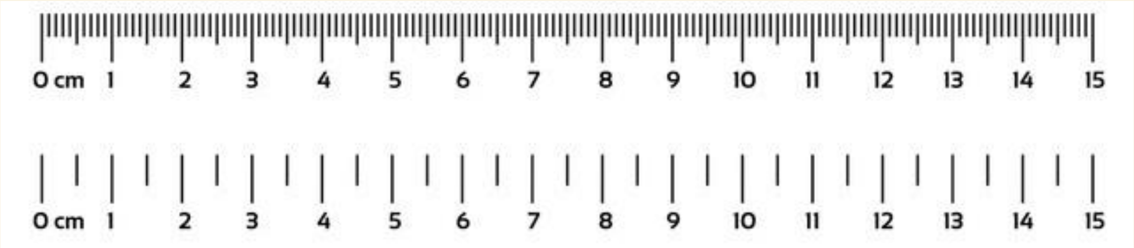
Height	Weight	Shoe size	Favourite colour
1.1	0.4	36	Green
1.9	1.2	41	Blue
1.7	1.9	39	Blue
2.8	2.0	43	Orange
2.3	2.8	44	Yellow

Discrete and continuous data

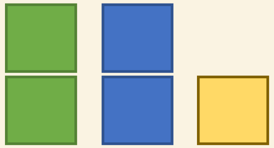
Height	Weight	Shoe size	Favourite colour
1.1	0.4	36	Green
1.9	1.2	41	Blue
1.7	1.9	39	Blue
2.8	2.0	43	Green
2.3	2.8	44	Yellow

Continuous data is measurable and can take any numeric value within a range.

The precision of the measurements is only limited by the tools we use, e.g. height in cm or mm:



Discrete data is countable and only takes specific values. We count the number of people who sit in the categories.

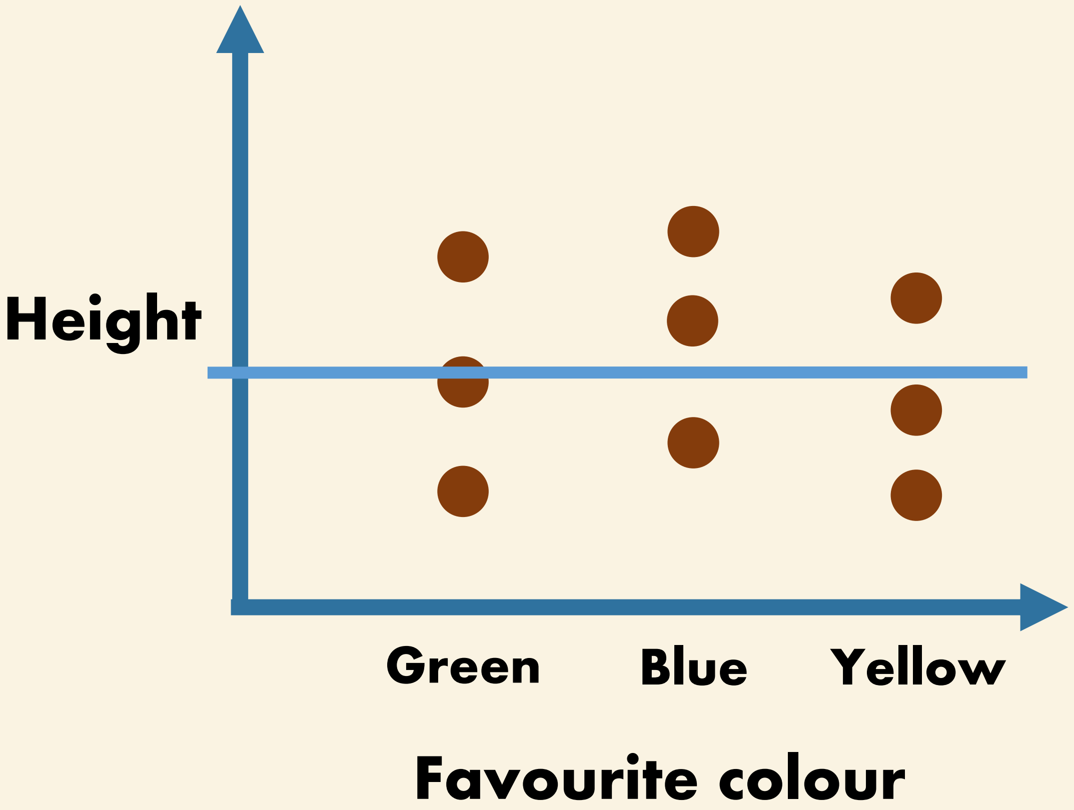


Two people love the colour green, two blue, and one yellow.

Linear regression with discrete measurements

- Old linear regression: **Height** = 0.5 x **Weight** + 1.1
- New linear regression: **Height** = 0.1 x **Favourite colour** + 1.1

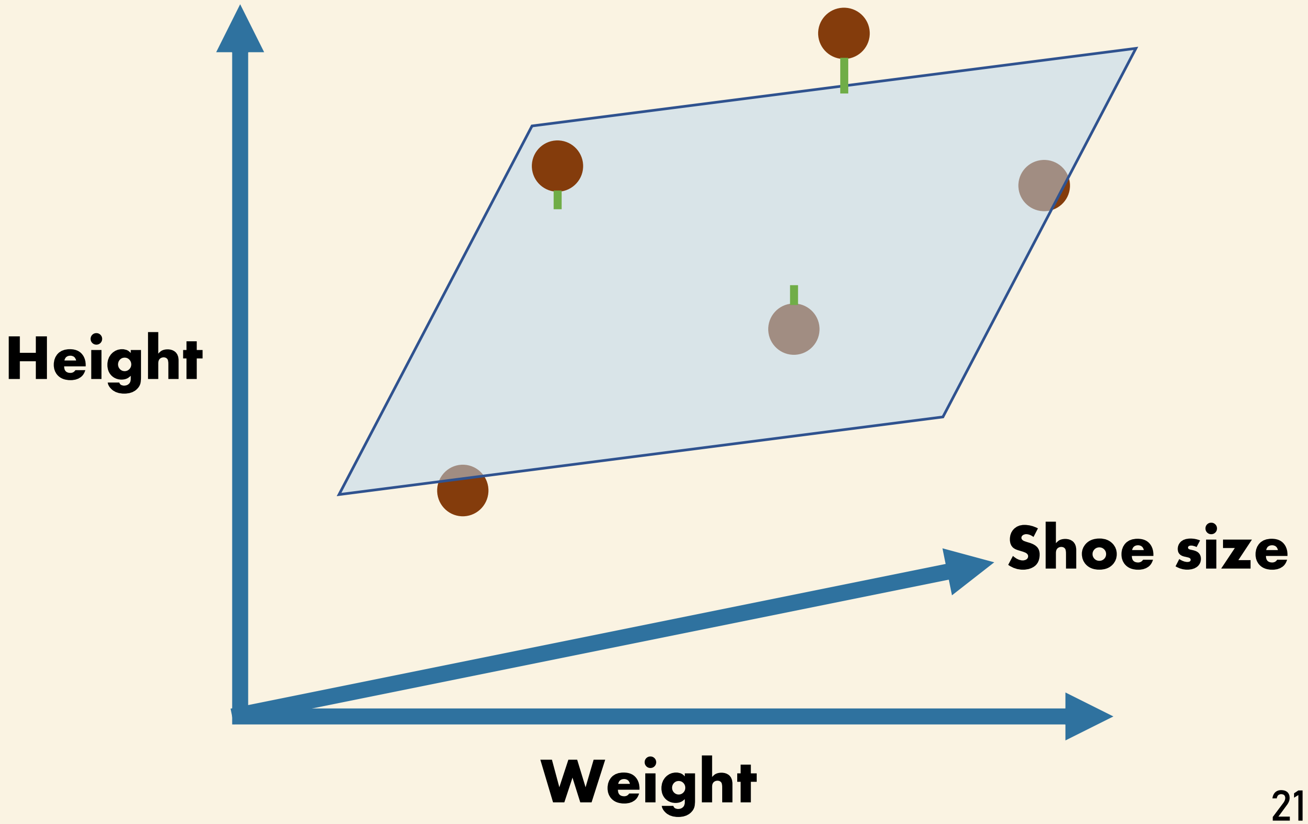
Height	Favourite colour
1.1	Green
1.9	Blue
1.7	Blue
2.8	Green
2.3	Yellow



Multiple linear regression

- Linear regression: **Height** = 0.5 x **Weight** + 1.1
- Multiple linear regression: **Height** = 0.5 x **Weight** + 0.3 x **Shoe size** + 1.1

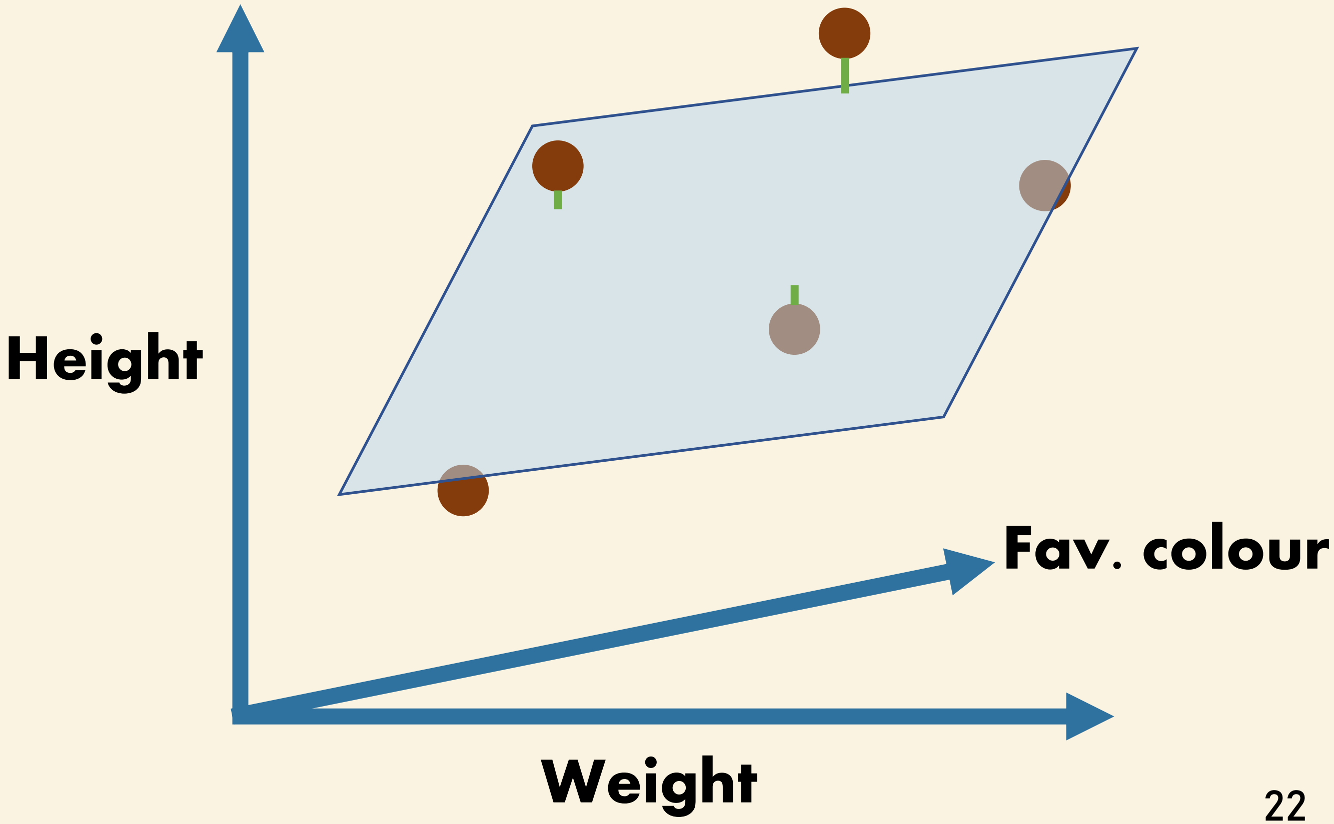
Height	Weight	Shoe size
1.1	0.4	36
1.9	1.2	41
1.7	1.9	39
2.8	2.0	43
2.3	2.8	44



Multiple linear regression

- Linear regression: **Height** = 0.5 x **Weight** + 1.1
- Multiple linear regression: **Height** = 0.5 x **Weight +** 0.3 x **Fav. colour** + 1.1

Height	Weight	Favourite colour
1.1	0.4	Green
1.9	1.2	Blue
1.7	1.9	Blue
2.8	2.0	Green
2.3	2.8	Yellow



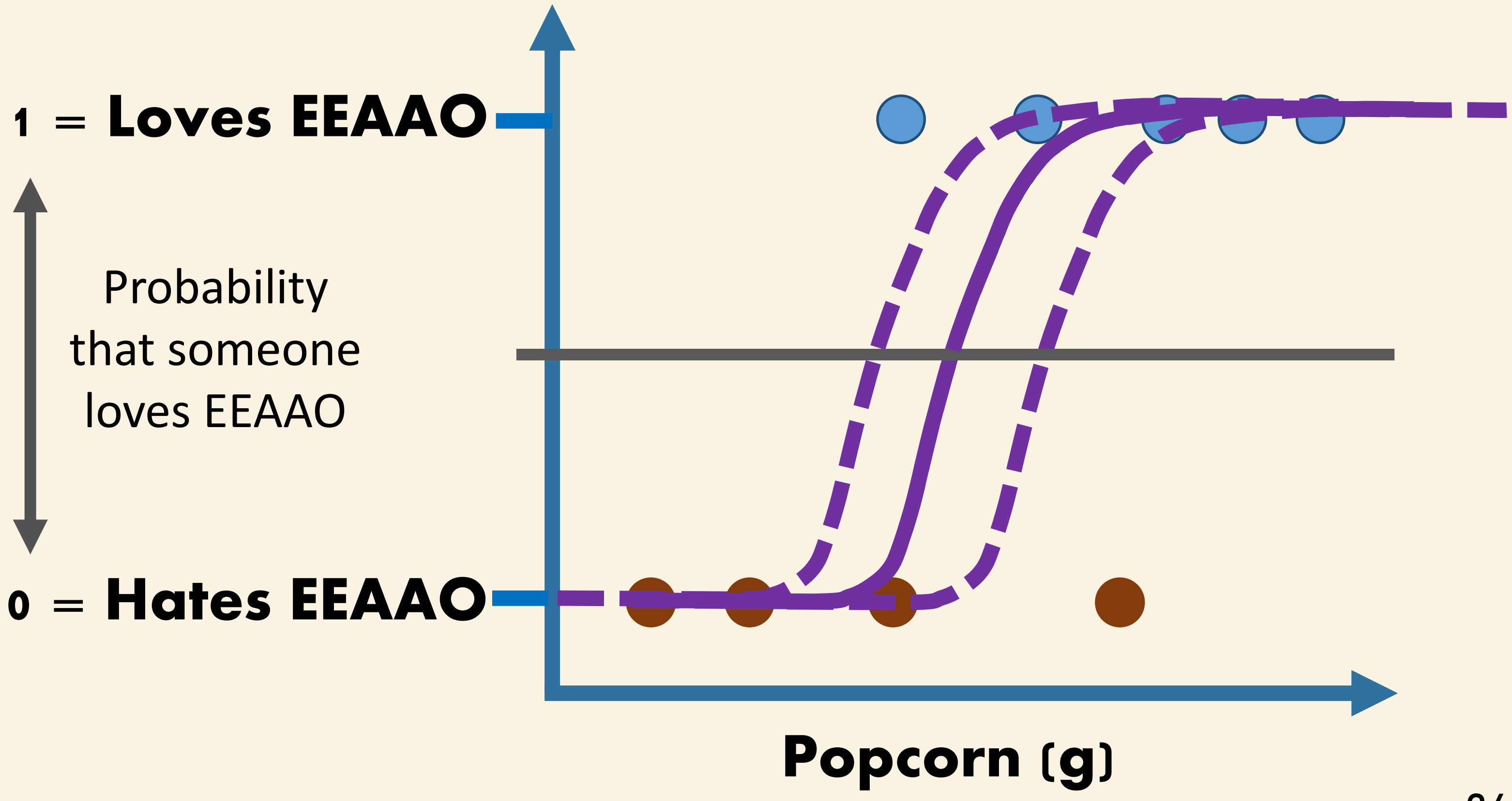
Logistic regression

1. Use **maximum likelihood** to fit an S-shaped logistic function to the data.
2. Calculate the R^2.
3. Calculate the p-value.

Logistic regression

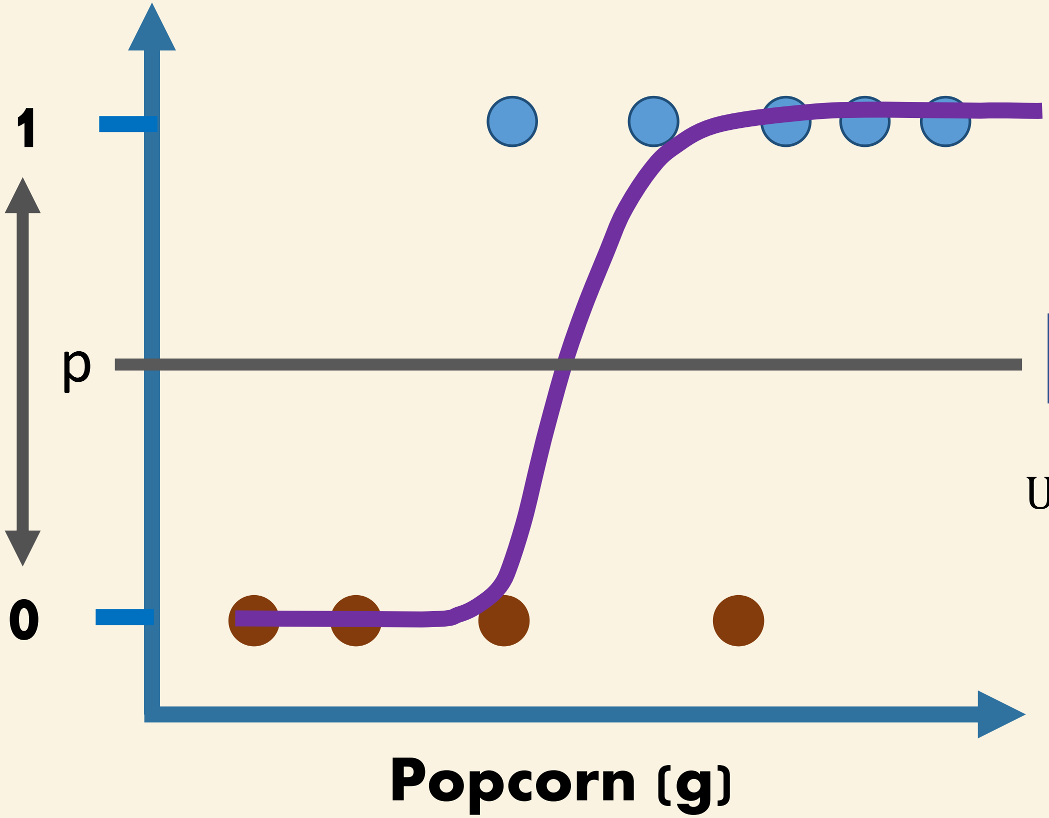
1. Use **maximum likelihood** to fit an S-shaped logistic function to the data.
2. Calculate the R^2.
3. Calculate the p-value.

Loves EEAAO	Popcorn (g)
1	95
0	50
1	100
1	85
0	60



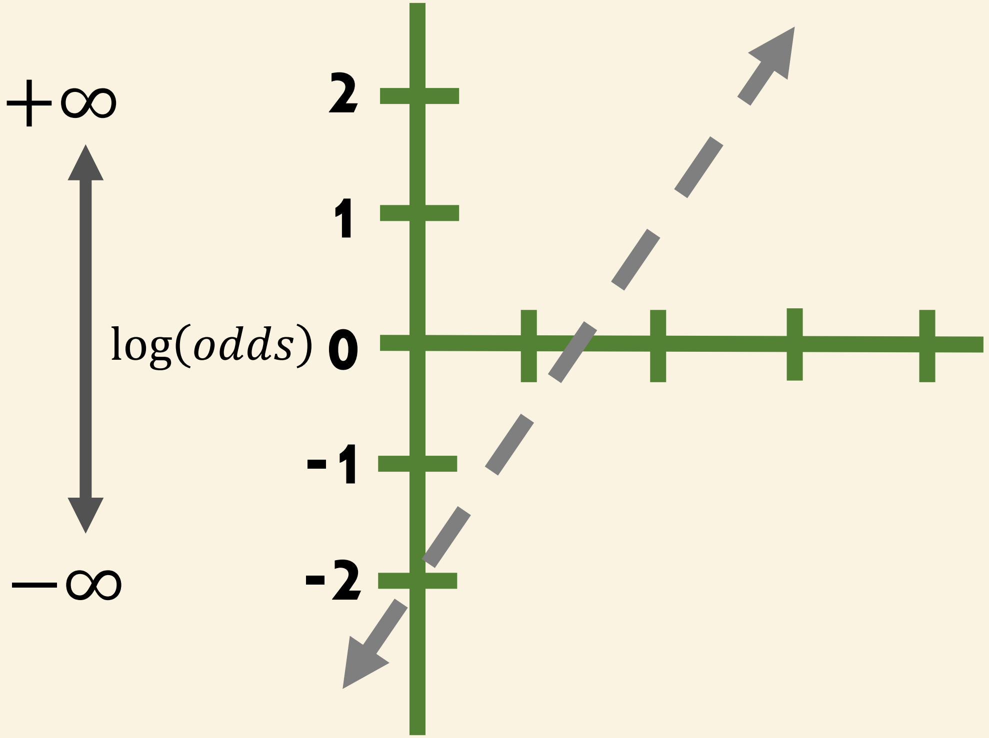
Logistic regression

1. Use **maximum likelihood** to fit an S-shaped logistic function to the data.
2. Calculate the R^2.
3. Calculate the p-value.



Use logit function:

$$\log\left(\frac{p}{1-p}\right)$$



y = mx + b

log(odds) = slope x popcorn (g) + intercept

Multiple logistic regression

- As with linear regression, we can use multiple discrete and continuous independent variables.

Loves EEAAO	Popcorn (g)	Loves Hacksaw Ridge	Astrological sign
1	95	0	Aquarius
0	50	1	Virgo
1	100	0	Taurus
1	85	1	Gemini
0	60	1	Leo

Practical session

Literature workshop

Mental health and caregiving experiences of family carers supporting people with psychosis (Sin *et al*, 2021)

tinyurl.com/2as79xtv

Workshop questions

Spend 10 minutes to skim through the Abstract and Table 1-3.

1. What was the aim of the study?
2. What were the dependent and independent variables?
3. Interpret the regression coefficients in Table 3.

Imperial College London

Workshop answers

1. What was the aim of the study?

To explore the associations between demographic, carer characteristics, and mental health outcomes of family carers supporting an individual with psychosis.

Workshop answers

2. What were the dependent and independent variables?

Dependent variable: Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS); range 14-70, higher score better wellbeing

Independent variable: (9) age, gender, ethnicity, employment status, highest education level achieved, marital status, relationship with CfP, living arrangement, duration of care.

Workshop answers

3. Interpret the regression coefficients in Table 3.

e.g. *Age of CfP*

For every unit increase in age of CfP (1 year):

- **(Coefficient + CI)** WEMWBS on average slightly increases by 0.29 with a 95% CI 0.1 to 0.5, after adjusting for other variables in the model
- **(p-value)** there is a strong evidence ($p<0.01$) that this association is not caused by random chance

Next steps

1. **Resources:** StatQuest, STHDA, RPubs, Imperial Graduate School
2. **Statistics fundamentals** (histograms, probability distributions, etc.)
3. **Machine learning** (classification and prediction)

Learning outcomes

- **Identify** the correlation coefficient as a single measure of linear association.
  - ρ, or rho, has values between -1 and 1 and reflects linear correlation.
- **Apply** general linear models to model a response variable in terms of a single or multiple variables.
  - lm(y ~ x) and glm(y ~ x, family = binomial)
- **Evaluate** model fitness by comparing the results produced by the model with your data.
  - R-squared, p-value
- **Present** model fitness using data visualisation techniques.
  - plot(y ~ x)
- **Interpret** regression model results from scientific papers.

Head to menti.com

Code: **7452 3861**

Thank you for listening!