Regression Modelling in R

Sonja N. Tang and Fernando Guntoro

Contents

- Introductions
- Regression
- Correlation
- Residuals and Least squares
- Model fitting
- Example – Pima Indians Diabetes Database
- Visualization
- Interpretation and Application

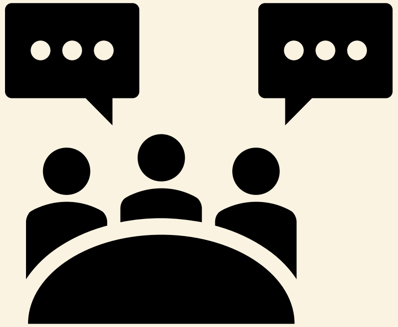
Introductions

- **Name:** Sonja
- **Origin:** Vienna, Austria
- **Department:** Epidemiology and Biostatistics and MRC Centre for Environment and Health
- **PhD topic:** Causal network analysis for health data
- **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?



Scan for Menti quiz:

or

Head to menti.com, code:

Learning outcomes

1. **Define and explain** fundamental concepts of regression modelling.
2. **Formulate, apply, and compare** regression models based on a research question.
3. **Estimate** regression coefficients using R and **interpret** them in the context of the question.
4. **Interpret** regression model results from scientific papers.

Table of contents

1. Theory (~45 min)

- Background
- Linear regression
- Logistic regression

Break (5 min)

2. Practical I: linear regression (~45 min)

Break (5 min)

3. Practical II: logistic regression (~30 min)

4. Interpreting a study (~15 min)

Imperial College London

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

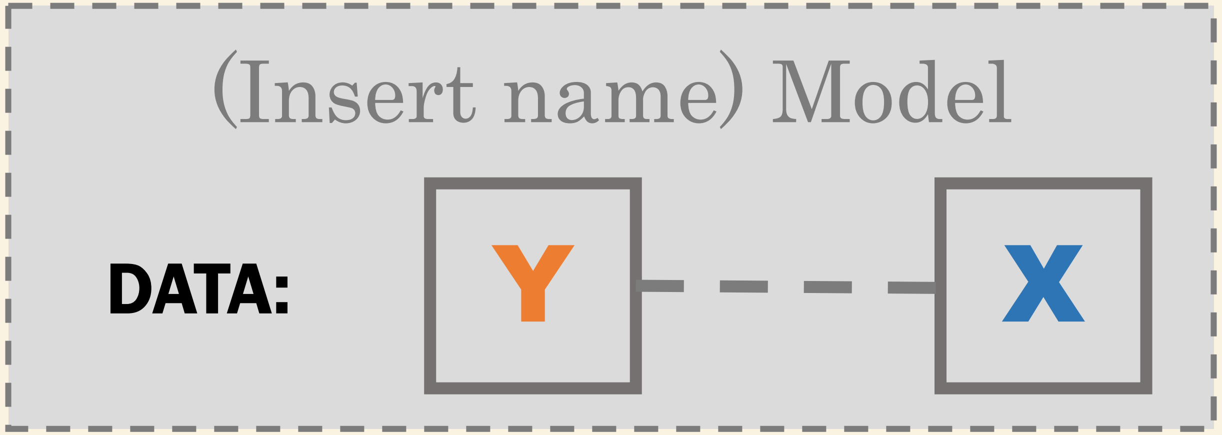
Height	Weight
1.1	0.4
1.9	1.2
1.7	1.9
2.8	2.0
2.3	2.8

What is regression modelling?

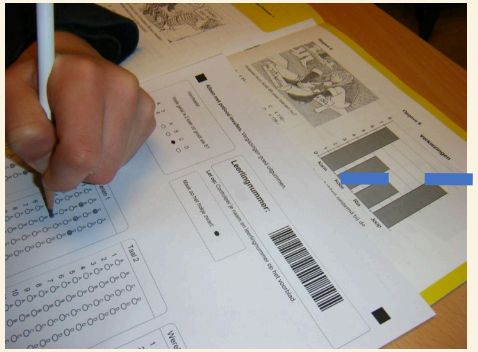
In statistics, regression modelling is a process for **estimating** a **line** or **curve** that **best represents the general trend** between one **outcome variable (Y)** and one or more **predictor variables (X)**.

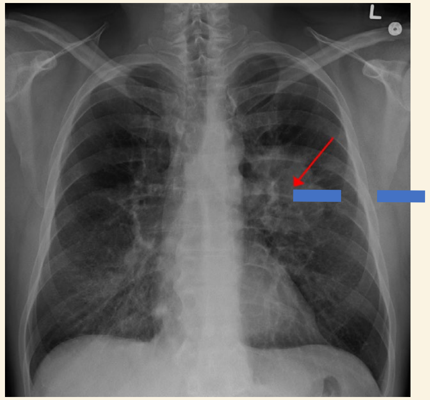
'dependent', 'response', 'label'

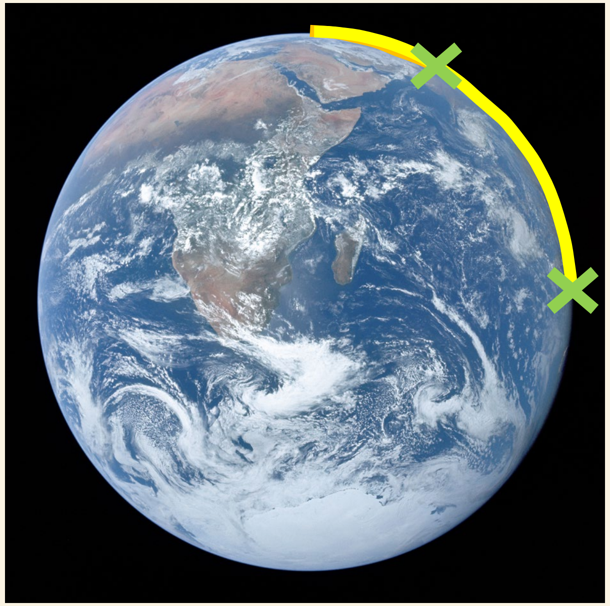
'independent', 'covariates', 'explanatory variables', 'features'



When do we need regression modelling?

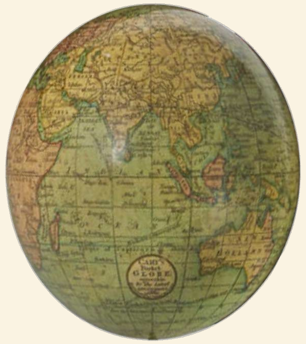


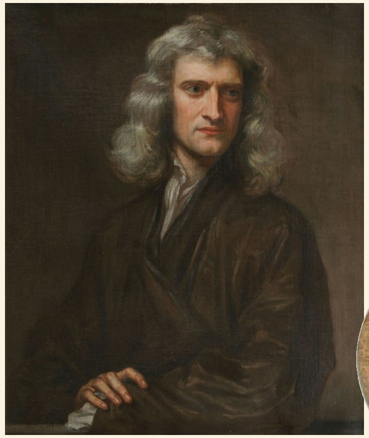




When do we need regression modelling?

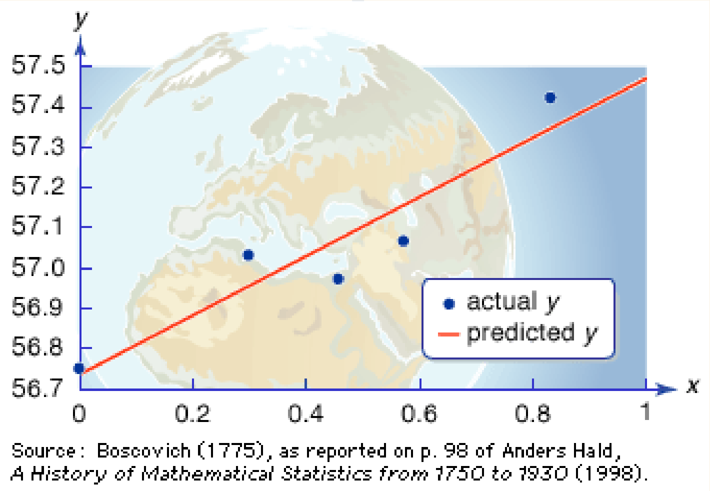
Jacques Cassini
(1677-1756)





Isaac Newton
(1643-1727)

Arc length in Paris toise
(roughly 1.9m)

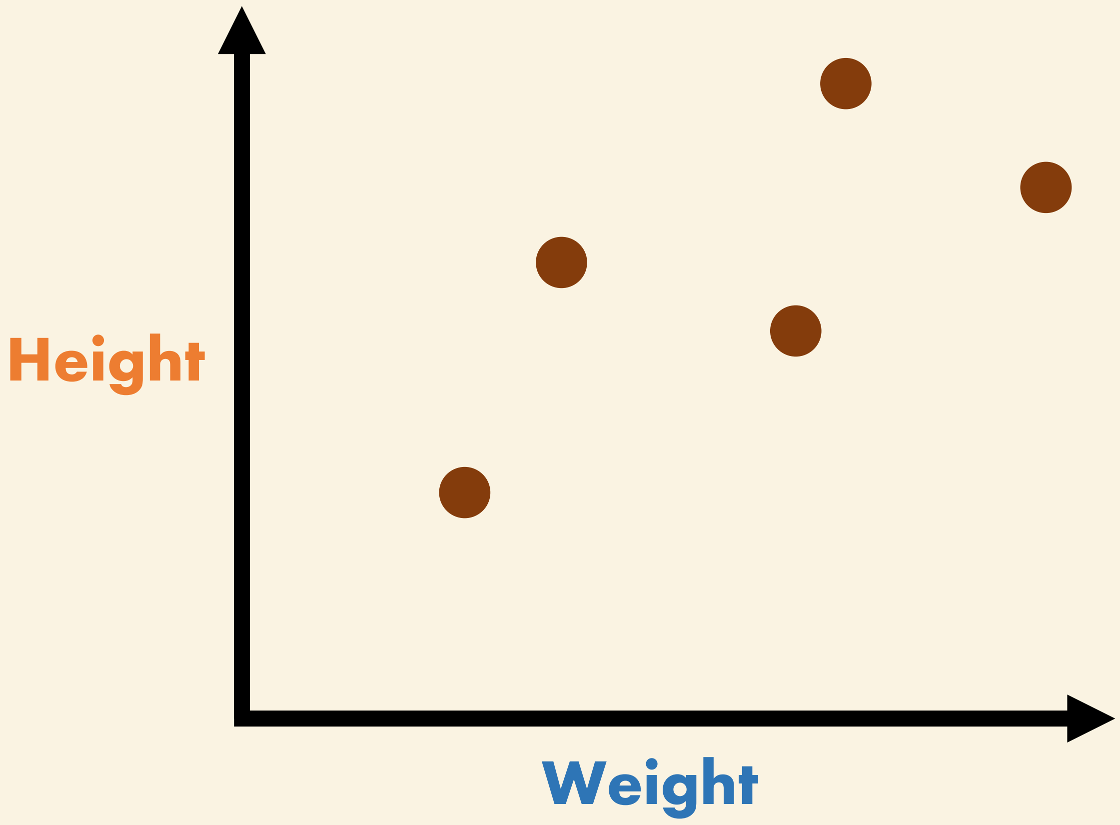


Latitude in degrees

Linear regression using the least squares method

One outcome (Y) and one predictor (X)

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



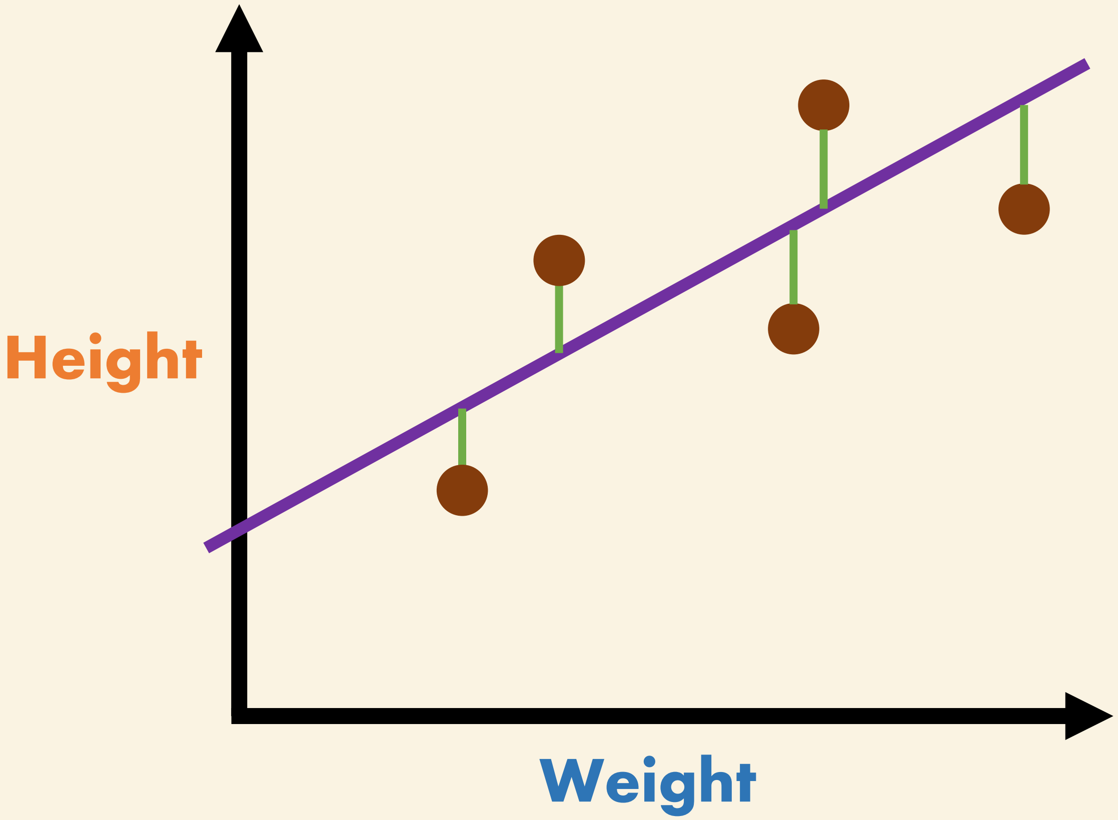
Linear regression

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

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

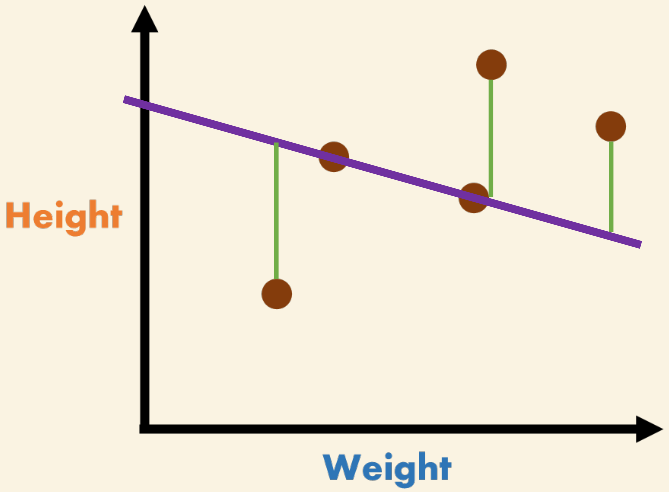
Residual = Observed - Fitted

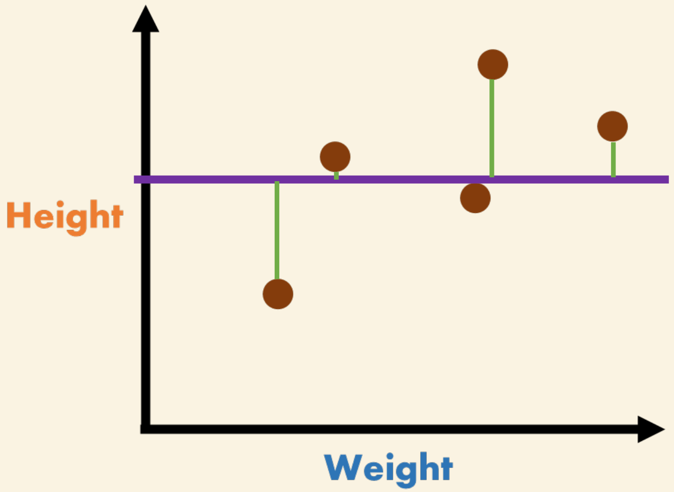
$$SSR = \sum_{i=1}^{n} (Observed_i - Fitted_i)^2$$

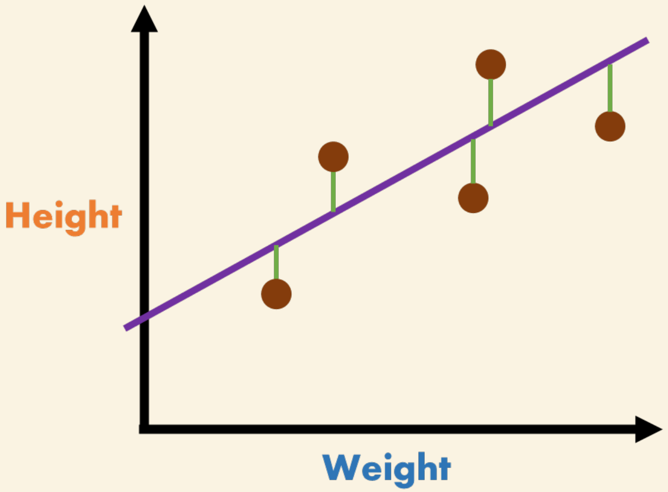


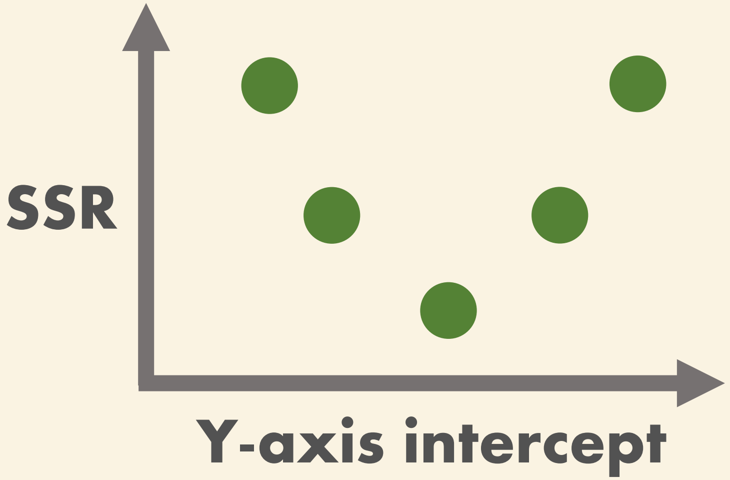
Linear regression

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









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

Height = slope x Weight + intercept
Height = 0.5 x Weight + 1.1

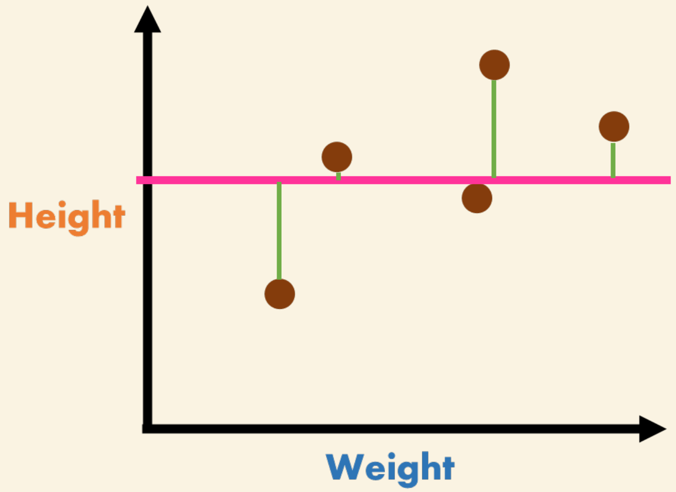
Linear regression

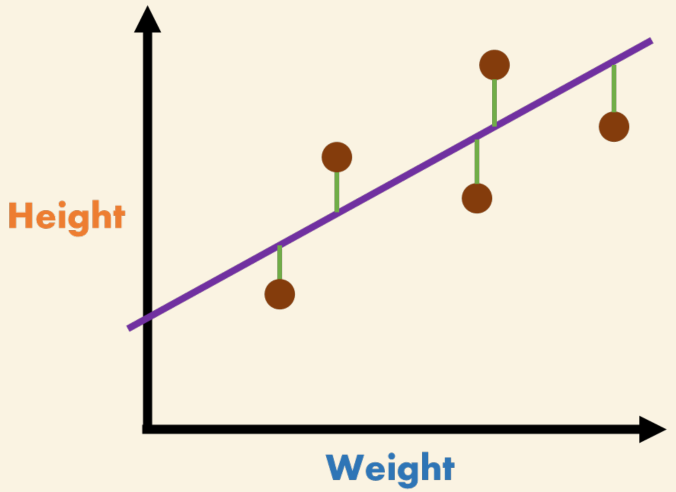
1. Use **least-squares** to fit a line to the data.
2. Calculate the **R^2.**
3. Calculate a ***p*-value** for R^2.

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

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

$$R^2 = \frac{1.6 - 0.5}{1.6} = 0.7$$





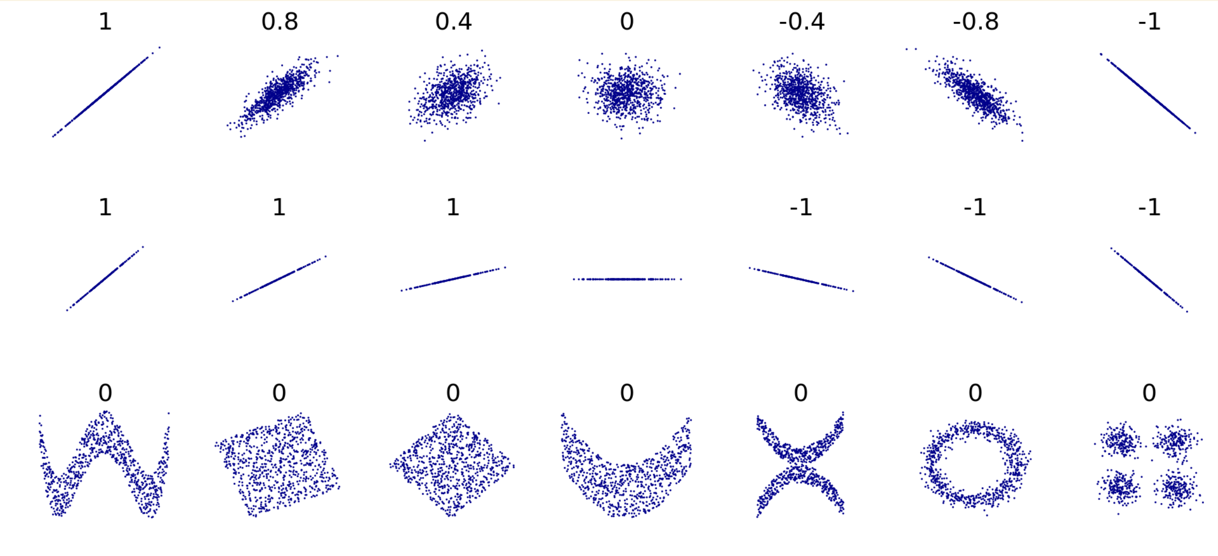
Pearson correlation coefficient (ρ)

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

The word Correlation is made of **Co-** (meaning "together"), and **Relation**.

$\rho = r$

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



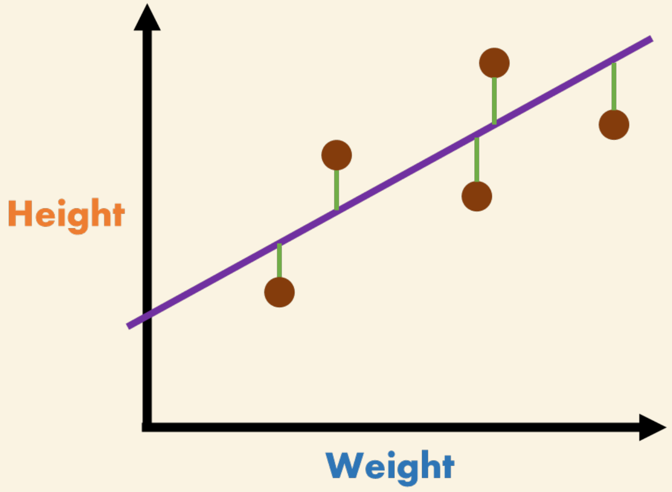
- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases
- The value shows how good the correlation is (not how steep the line is), and if it is positive or negative.

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 = 0.7

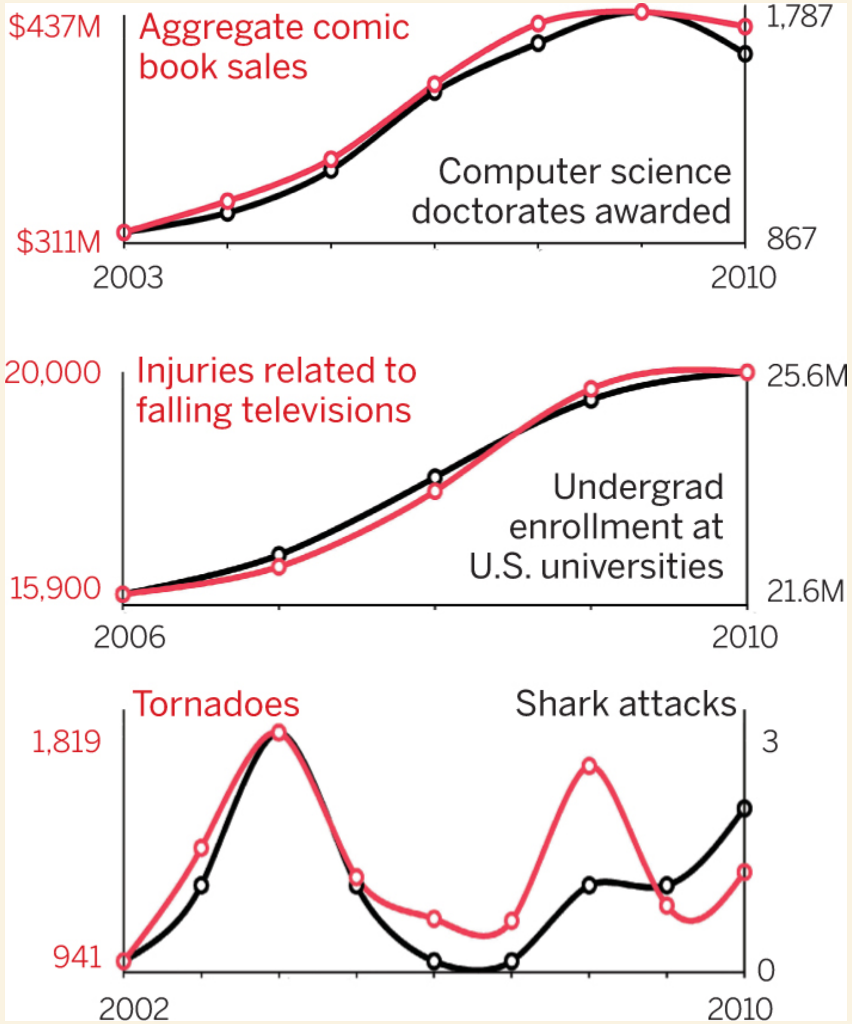
***p*-value = 0.1**

Correlation is not always causation

Height ~ Weight

Height <- Weight

Height -> Weight



Source: Tyler Vigen for Science Magazine

One outcome and multiple predictors

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

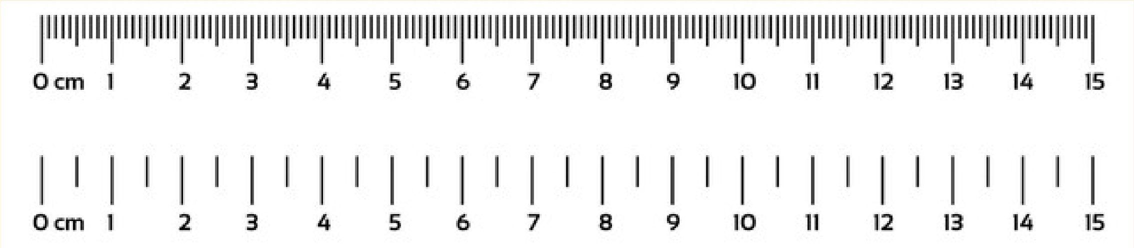
Imperial College London

Continuous and discrete data

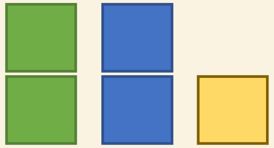
Height	Weight	Shoe size	Favourite colour
1.1	0.4	36	Green
1.9	1.2	41	Blue
1.7	1.9	39	Blue
2.8	2.0	43	Orange
2.3	2.8	44	Yellow

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.

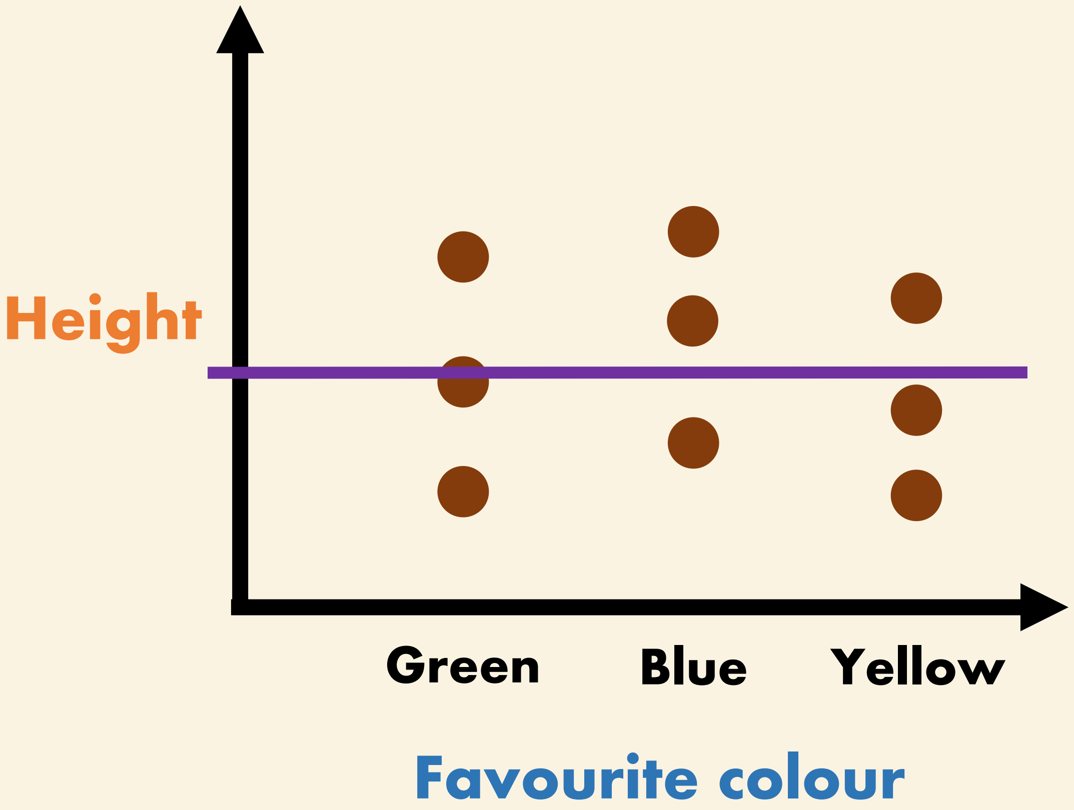


Two people love the colour green, two blue, and one yellow.

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

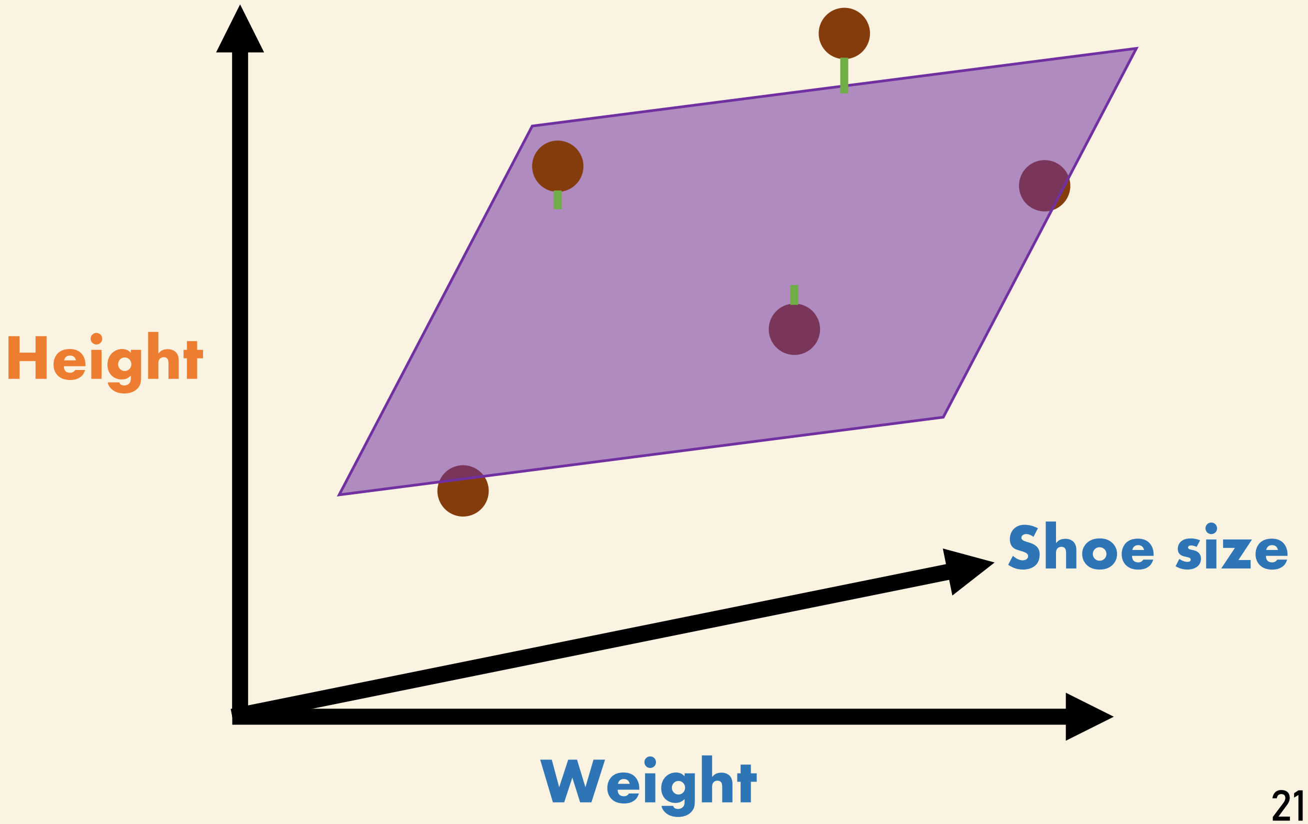
Height	Fav. colour
1.1	Green
1.9	Blue
1.7	Blue
2.8	Green
2.3	Yellow



Multiple linear regression

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

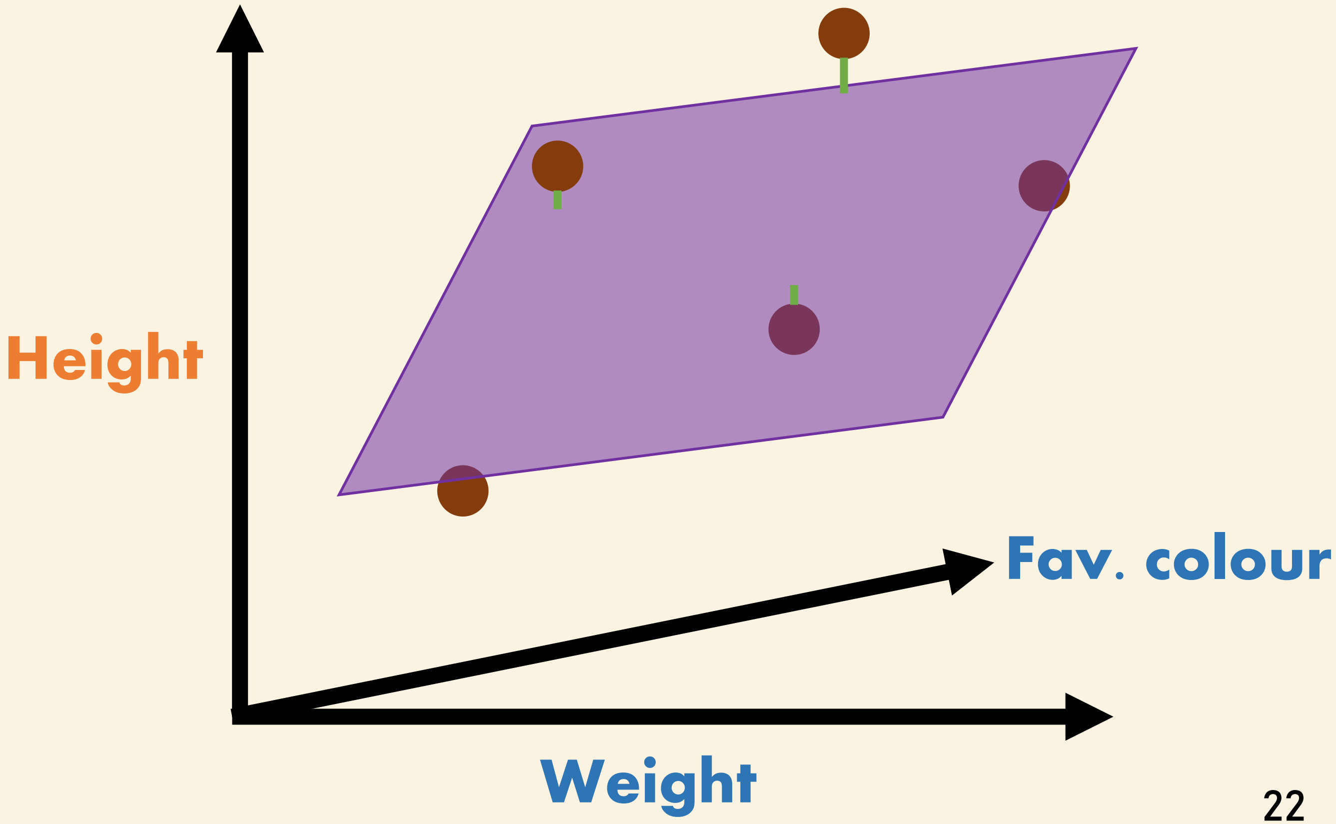
Height	Weight	Shoe size
1.1	0.4	36
1.9	1.2	41
1.7	1.9	39
2.8	2.0	43
2.3	2.8	44



Multiple linear regression

- Simple 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	Fav. 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



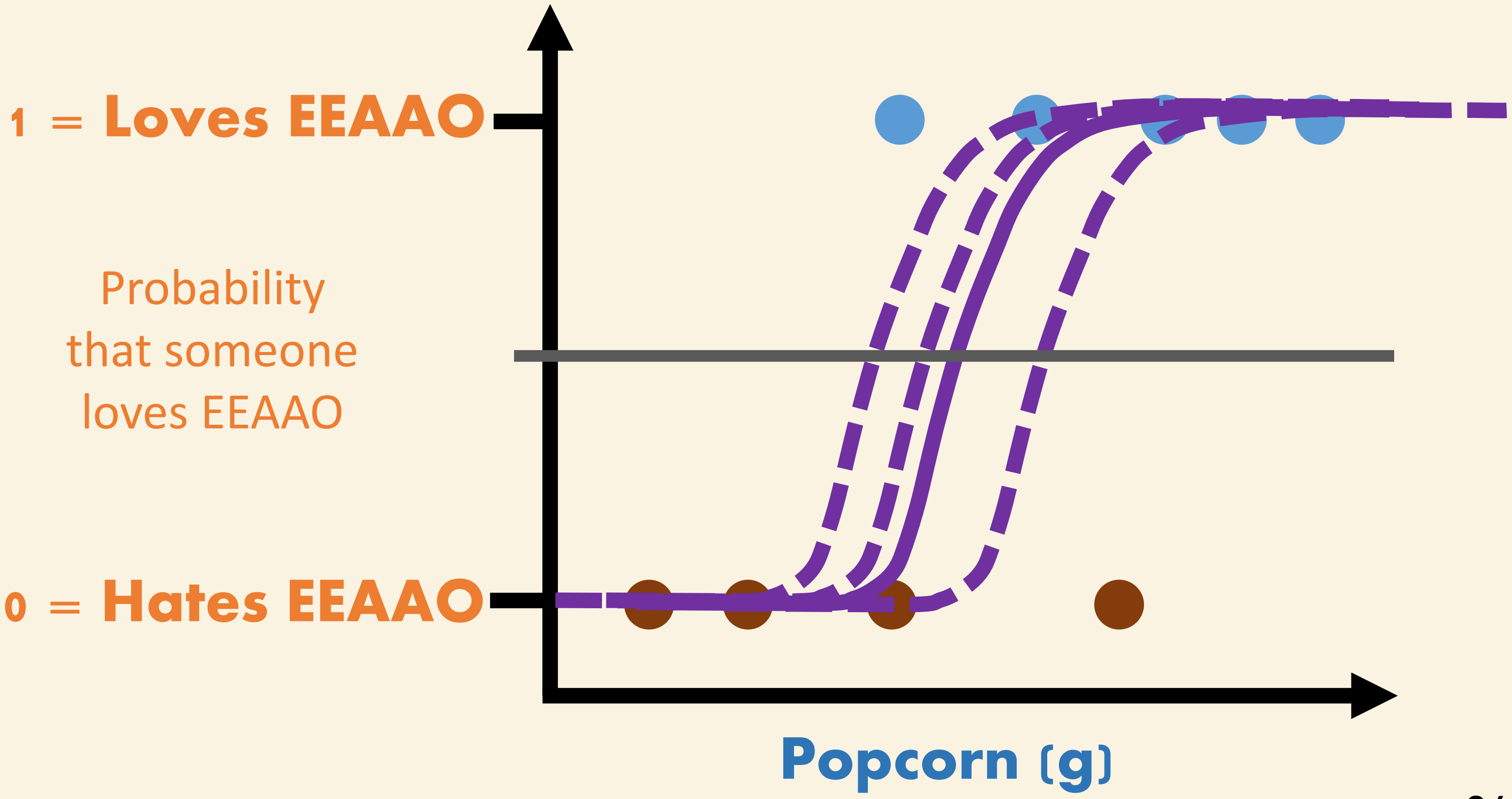
Logistic regression

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

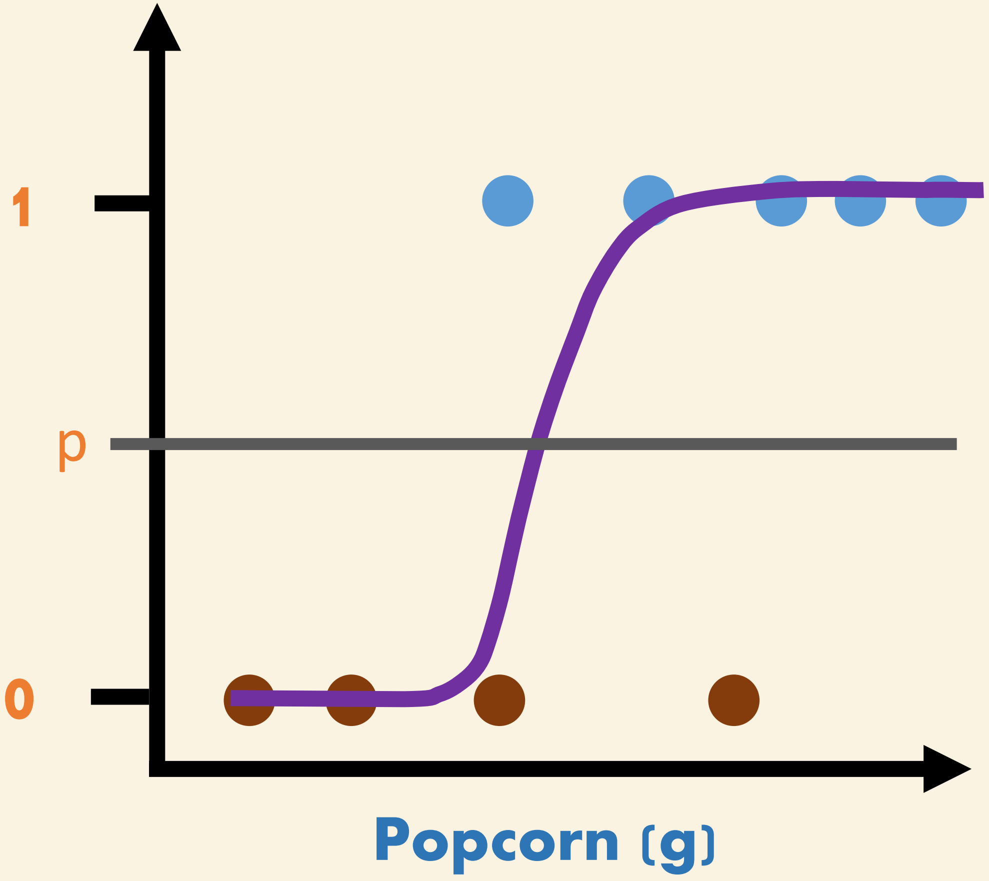
1. Use **maximum likelihood** to fit an S-shaped logistic function to the data.
2. Calculate the R^2.
3. Calculate the p-value.

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



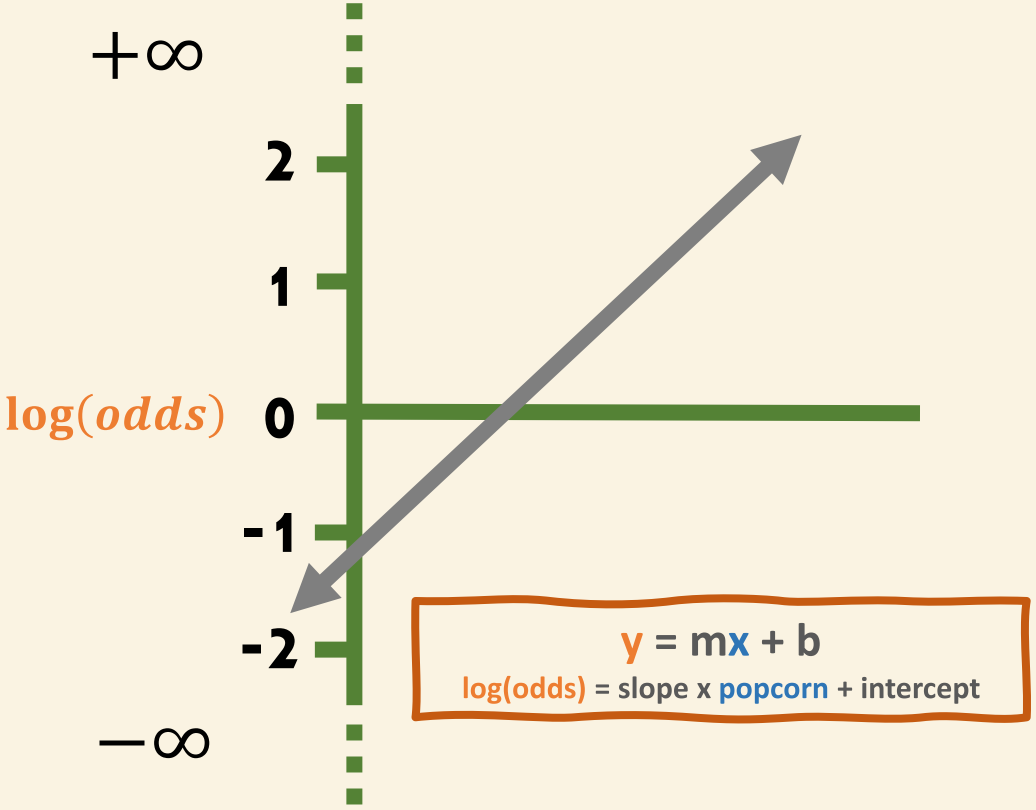
Logistic regression

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:

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



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 – but why use R?

Practical session – but why use R?

Imperial College London

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.**

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.

Imperial College London

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, Coursera
2. **Statistics fundamentals:** histograms, probability distributions, hypothesis testing
3. **Machine learning:** regression, classification, clustering, dimensionality reduction

Learning outcomes

1. **Define and explain** fundamental concepts of regression modelling.
    - Regression models contain one outcome and one or multiple predictors.
    - Regression modelling consists of fitting a line or curve to the data and calculating the R^2 and p-value.

2. **Formulate, apply, and compare** regression models based on a research question.
    - Formulate and apply bespoke lm(y ~ x) and glm(y ~ x, family = binomial) models.
    - Identify potential covariates or confounding variables that should be considered in a regression model.

3. **Estimate** regression coefficients using R and **interpret** them in the context of the question.
    - Assess the fit of a regression model using measures such as R-squared and adjusted R-squared.

4. **Interpret** regression model results from scientific papers.

Graduate School feedback form

Attendance link