Science, Scientific Discovery, and Statistics
Seminar 3: Correlation and Causation

What we are doing today

- Correlation as a pattern in data (association)
- How correlation is measured for different variable types:
  - categorical vs categorical (contingency tables & conditional proportions)
  - binary categorical vs continuous (difference in means)
  - continuous vs continuous (scatterplots, Pearson correlation)
- Simple linear regression: the "best line" idea (least squares)
- How to read a regression table (like in academic papers)

Covariance

Idea

Two variables covary if they tend to change together with a pattern.

For two numerical variables X and Y:

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mu_X)(Y - \mu_Y)\big]$$

- $\mathrm{Cov}(X, Y) > 0$: high X tends to come with high Y (positive association)
- $\mathrm{Cov}(X, Y) < 0$: high X tends to come with low Y (negative association)

Correlation

What correlation measures

Correlation is a single number between -1 and 1 that summarizes how strongly two numerical variables move together using a straight line.

- $r \approx +1$: strong positive linear relationship (dots tilt up tightly)
- $r \approx -1$: strong negative linear relationship (dots tilt down tightly)
- $r \approx 0$: little/no linear relationship (cloud with no clear tilt)

Key intuitions

- **Unit-free:** correlation is unchanged if you switch units (e.g., pounds to kilograms).

$$r_{XY} = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

- Correlation is just covariance standardised by the variables' standard deviations.

Two categorical variables: contingency table intuition

Think of a contingency table as a neat way to compare **rates** across groups.

	Did not vote	Vote
Not unemployed	O_{11}	O_{12}
Unemployed	O_{21}	O_{22}

The key comparison

Compute the **voting rate** in each group ("within-row proportion"):

$$\textbf{Among unemployed} \quad \frac{O_{22}}{O_{21} + O_{22}}$$

$$\textbf{Among not unemployed} \quad \frac{O_{12}}{O_{11} + O_{12}}$$

- If the two rates are similar $\Rightarrow$ weak/no association.
- If the two rates are very different $\Rightarrow$ strong association.
- Main habit: always ask **"Compared to which baseline?"**.

Which statements describe a correlation?

Statements:

1. People who live to be 100 years old typically take vitamins.
2. Cities with more crime tend to hire more police officers.
3. Successful people have spent at least ten thousand hours honing their craft.
4. Most politicians facing a scandal win reelection.
5. Older people vote more than younger people.

Rule of thumb

A statement describes a **correlation** only if it compares **variation in both variables**:

How does Y change as X changes?

Not correlation: 1, 3, 4

They describe a **single group** or a **single level** of X.

Correlation statements: 2, 5

They make an explicit **comparison across levels of** X.

Simple linear regression: the model

We now choose a direction: X helps explain Y.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- α (intercept): predicted Y when $X = 0$
- β (slope): average change in Y for a $+1$ change in X
- ε_i (error): everything about Y not captured by X

Fitted values and residuals

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i, \qquad \hat{\varepsilon}_i = Y_i - \hat{Y}_i.$$

What is the "best line"?

Regression chooses the line that makes residuals "small" overall.

Ordinary Least Squares (OLS)

Pick $(\hat{\alpha}, \hat{\beta})$ to minimize:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \big(Y_i - (\alpha + \beta X_i)\big)^2.$$

Closed-form solutions:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}, \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}.$$

- If X and Y covary strongly $\Rightarrow$ bigger $|\hat{\beta}|$
- If X barely varies $\Rightarrow$ slope is hard to estimate

Using regression: prediction + fit

Prediction: plug in a value of X.

$$\widehat{Y}(x^\star) = \hat{\alpha} + \hat{\beta} x^\star.$$

Correlation vs causation (intuition)

A strong $\hat{\beta}$ can arise because:

$$X \leftarrow Z \rightarrow Y \quad \text{(a confounder drives both)}$$

If Z is omitted, $\hat{\beta}$ may mix association with confounding.

How to read a regression table

Table: OLS regression of Trust in Politicians on EU Referendum Preference

	(1)	(2)	(3)
Stay (1=Stay, 0=Leave)	0.45^{***}	0.32^{***}	0.28^{**}
	(0.08)	(0.09)	(0.11)
Age (years)		0.01^{***}	0.01^{***}
		(0.00)	(0.00)
Education (years)			0.02^{**}
			(0.01)
Constant	2.10^{***}	1.20^{***}	1.05^{***}
	(0.05)	(0.18)	(0.22)
Controls	No	Yes	Yes
N	2,000	2,000	2,000
R^2	0.06	0.12	0.13

Notes: Standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Column (1) is bivariate; columns (2)–(3) add controls. Coefficients are differences in the dependent variable units (here: trust scale points).