Lecture 5: Multivariate Linear Regression

Paul Röttger

Applied Analytical Statistics

17th of February 2026

Summative proposals | Feedback

Overall: great job! Very exciting project ideas, cool datasets, plausible methods.

Common challenges:

Choosing regression models: Linear regression may not be appropriate. Consider type of outcome and distribution. Model choice should follow from data-generating process.

Working with hierarchical data: Many datasets are clustered. Observations within clusters are not independent. This invalidates standard inference → **mixed-effects models**.

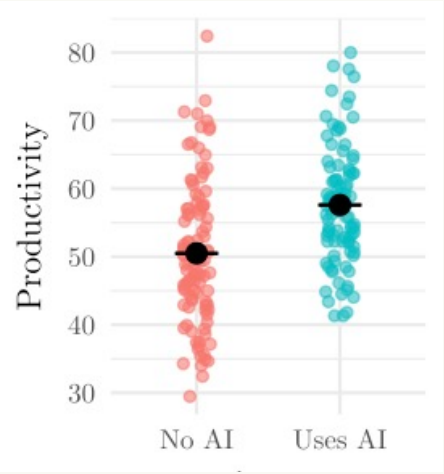
Working with time series data: Observations across time are rarely independent. This invalidates standard inference → **time series models**.

Plan for today | Multivariate linear regression

[TO BE ADDED]

Motivation | Limitations of univariate regression

Univariate regression allows us to quantify relationships between two variables.



$$\hat{Y} = 50.5 + 7.1\,X_1$$

Observation: Researchers who use AI tools are more productive.

Possible explanation:
AI use increases productivity.

🤔 ... but what else could explain this observation?

Alternative explanation:
Skilled workers adopt AI and are more productive.

Motivation | Omitted variable bias

Let's assume that the **true population model** includes two regressors:

$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ where in our example Y is productivity, X_1 is AI use, X_2 is skill.

However, so far we would have estimated coefficients for a univariate population model:

$Y = \alpha_1 X_1 + u$ where X_2 is not accounted for

Then by $\alpha_1 = \mathrm{Cov}(X_1, Y)/\mathrm{Var}(X_1)$ we can show that α_1 is **biased** relative to β_1:

$$\alpha_1 = \beta_1 + \beta_2 \cdot \frac{\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)}$$ $\rightarrow \alpha_1$ mixes direct effect of X_1 + indirect pathway through X_2

Motivation | Omitted variable bias (cont'd)

OLS is still unbiased for the parameter of the model we estimate:

$E[\hat{\alpha}_1] = \alpha_1$ $\rightarrow \hat{\alpha}_1$ is an unbiased estimator of α_1

BUT **bias in population parameters** carries through to estimated coefficients:

$$E[\hat{\alpha}_1] = \alpha_1 = \beta_1 + \beta_2 \cdot \frac{Cov(X_1, X_2)}{Var(X_1)}$$

$\rightarrow \hat{\alpha}_1$ is a biased estimator of β_1

Under which conditions can we ignore X_2?

α_1 is unbiased relative to β_1 IF:

- $\beta_2 = 0 \rightarrow$ the omitted variable X_2 does not affect Y
- $Cov(X_1, X_2) = 0 \rightarrow$ the omitted variable X_2 is uncorrelated with the included regressor X_1.

Motivation | Multivariate regression as adjustment

Multivariate regression addresses OVB by modelling the **conditional mean**:

$$E[Y \mid X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

population model of multivariate linear regression

Each coefficient measures the association between one regressor and the outcome **holding other observed variables constant** = **"controlling" for observed variables**

RQ: Is AI use associated with increased worker productivity?

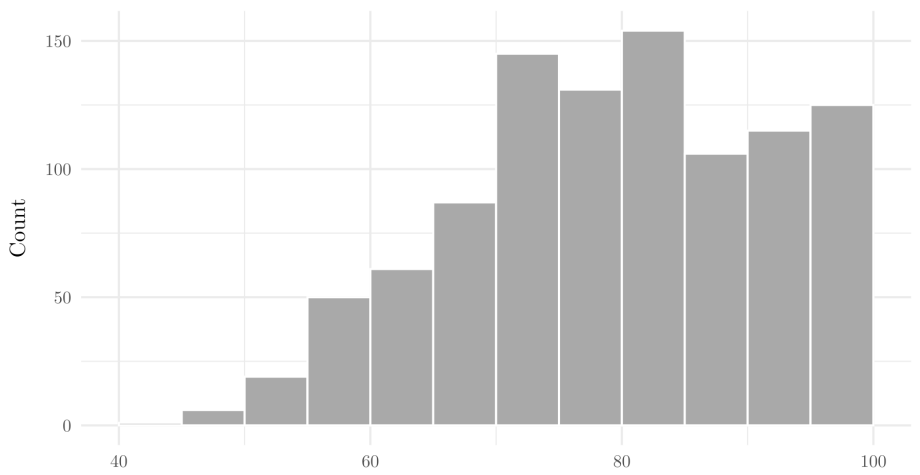
→ we want to compare like with like (i.e. similar in relevant characteristics) to estimate direct association

Controlling for variables other than AI use isolates the association between AI use and productivity that is not explained by the included covariates of AI use.

Controlling for confounders is NOT ENOUGH to establish causality (→ Week 7).

Multivariate linear regression | Working example

Data: AI assessments for 1,000 job candidates, based on CV and coding test results.



AI hiring score (0-100)

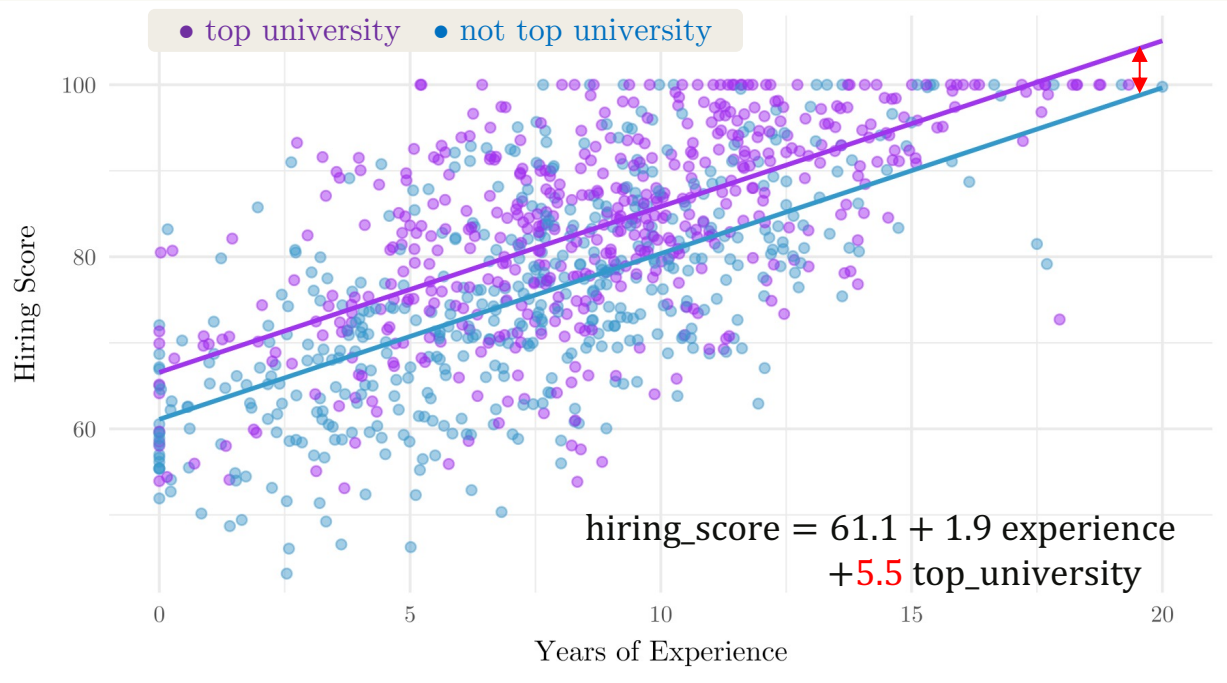
Potential regressors:

- experience (years)
- top_university (yes/no)
- coding_score (0-100)

Interpretation | Single regressor

$\hat{\beta}_1 = 2.1$: Each one-year increase in experience is associated with a 2.1-point increase in AI hiring score.

Interpretation | Two regressors



$\hat{\beta}_1 = 1.9$: Each one-year increase in experience is associated with a 1.9-point increase in AI hiring score, **ceteris paribus**

$\hat{\beta}_2 = 5.5$: Holding a degree from a top university is associated with a 5.5-point increase in AI hiring score, **ceteris paribus**

Interpretation | Three regressors

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     30.52895    1.65612  18.434   <2e-16 ***
experience       1.02505    0.07766  13.199   <2e-16 ***
top_university1 -0.66819    0.58530  -1.142    0.254
coding_score     0.61121    0.03105  19.685   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$\hat{\beta}_2$ for top_university is no longer significant, after controlling for coding_score.

coding_score correlates with hiring_score and top_university → **omitted variable**.

The association between top_university and hiring_score in the two-regressor model was **confounded by coding ability**.

In the two-regressor model, $\hat{\beta}_2$ included indirect association via coding score.

Key equations | From scalar to matrix notation

	Univariate → scalar	**Multivariate** → matrix
Regression model:	$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$	$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$
OLS estimator:	$\hat{\beta}_1 = \mathrm{Cov}(X, Y)/\mathrm{Var}(X)$	$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
Sampling distribution:	$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum (X_i - \bar{X})^2}\right)$	$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$,
Standard error:	$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2 / \sum (X_i - \bar{X})^2}$	$SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}}$

Key equations | Multivariate regression model

$n \times 1$

outcome for n observations

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$n \times 1$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

error terms for n observations

design matrix: n observations, $k+1$ regressors

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n} & \dots & X_{kn} \end{pmatrix}$$

$\mathrm{n} \times (\mathrm{k}+1)$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

regression coefficients, for intercept + k regressors

$(\mathrm{k}+1) \times 1$

🤔 Why $k+1$ regressors? → includes intercept

Key equations | Covariance matrix

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

design matrix: $n \times (k+1)$

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n} & \dots & X_{kn} \end{pmatrix}$$

Variance of coefficient estimates depends on variance and covariance of regressors!

covariance matrix: $(k+1) \times (k+1)$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} c_{00} & c_{01} & \cdots & c_{0k} \\ c_{10} & c_{11} & \cdots & c_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k0} & c_{k1} & \cdots & c_{kk} \end{pmatrix}$$

diagonal elements are **variances**:

$$\text{Var}(\hat{\beta}_j) = \sigma^2 c_{jj} = \text{SE}(\hat{\beta}_j)^2$$

off-diagonal elements are **covariances**:

$$\text{Cov}(\hat{\beta}_j, \hat{\beta}_m) = \sigma^2 c_{jm}$$

where $c_{jm} = \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{jm}$

Inference | Individual t-tests

For individual coefficients, we test for **evidence of a conditional association** using the t-test:

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}$$

The test statistic measures the distance between our sample coefficient and the coefficient value under the null in SE units (see Week 3).

Under standard OLS assumptions and H_0, the t-statistic is t-distributed: $T \sim t_{n-k-1}$

→ adjust df for number of coefficients

The p-value is the probability of observing a t-statistic at least this extreme under the null.

We reject H_0 if $p < \alpha$ and do not reject H_0 if $p \geq \alpha$, where $p = Pr(|T_{n-k-1}| \geq |t| \mid H_0)$

Inference | Overall F-test

We may also want to test **whether all regressors jointly matter**: $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$
Under the null, none of the regressors are linearly associated with outcome Y.

$$F = \frac{ESS/k}{RSS/(n-k-1)}$$

ESS: explained sum of squares = variation explained by our fitted model

RSS: residual sum of squares = variation left unexplained

Under standard OLS assumptions and H_0, the F-statistic is F-distributed: $F \sim F_{k,\, n-k-1}$

🤔 What's the intuition?

→ numerator = explained variation per regressor
→ denominator = unexplained variation per degree of freedom

We may reject the overall F-test even if we fail to reject all individual coefficient t-tests.

Goodness of fit | Adjusted R^2

Standard R^2 always increases when we add more regressors to our model:

$$R^2 = 1 - \frac{RSS}{TSS}$$

OLS minimises RSS, so adding regressors can only reduce (or leave unchanged) RSS

→ standard $\mathbf{R^2}$ rewards overfitting

Adjusted R^2 penalises unnecessary complexity by accounting for loss of degrees of freedom:

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

we lose k+1 degrees of freedom from estimating coefficients

we lose 1 degree of freedom from estimating sample mean of Y

Keep in mind: **our goal in statistical inference is NOT to optimise (adjusted) $\mathbf{R^2}$.**
We design models to answer substantive research questions, not chase fit statistics.

Linear regression | Assumptions

$$E(\hat{\beta}) = \beta$$

Two main assumptions to obtain **unbiased estimates**:

Exogeneity: Regressors X are uncorrelated with unobserved determinants of Y.

Linearity: The conditional mean E[Y | X] is a linear function of parameters **β**.

Four additional assumptions for **correct standard errors**: → valid CIs, valid tests

Homoskedasticity: Variation in the error term ε is constant across regressors X.

Independence: Error terms ε are uncorrelated across observations.

No perfect **multicollinearity**: Regressors X are not perfectly correlated with each other.

Normality: The error term ε is normally distributed.

→ not important for large n because of CLT

Assumptions | Exogeneity

Exogeneity requires that $E[\varepsilon \mid X] = 0$, which is equivalent to $Cov(X, \varepsilon) = 0$.
= There are no systematic unobserved factors correlated with our regressors.

Exogeneity is a **design assumption** that can fail for many reasons:

Omitted variable bias: unobserved variable Z affects Y and is correlated with X.

Reverse causality: Y affects X but we estimate X predicting Y.

Measurement error: true regressor is X* but we observe $X = X^* + w$

Selection bias: inclusion in sample depends on unobserved determinants of Y.

In all cases:
$Cov(X, \varepsilon) \neq 0$.

Assumptions | What to do about exogeneity

No statistical test can verify exogeneity in observational data.
Exogeneity is a design assumption, not a statistical property.

Common strategies for mitigating exogeneity concerns:

→ today

Control for observed confounders by including them in the regression model.

→ Week 8?

For panel data, include **fixed effects** to remove time-invariant unobserved hetereogeneity.

In **research design**, try to guarantee exogeneity, e.g. through randomised treatments.

→ Week 7

When writing a research paper, **consider what is in the error term ε:**
Is ε plausibly uncorrelated with X? Why would this assumption be violated?

Assumptions | Linearity

We assume that **the true conditional mean is linear in parameters β:**

$$E[Y \mid X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

population model of
multivariate linear regression

This is an assumption about the **functional form** of the data-generating process.

Suppose the true model is $E[Y \mid X] = \beta_0 + \beta_1 X + \beta_2 X^2$

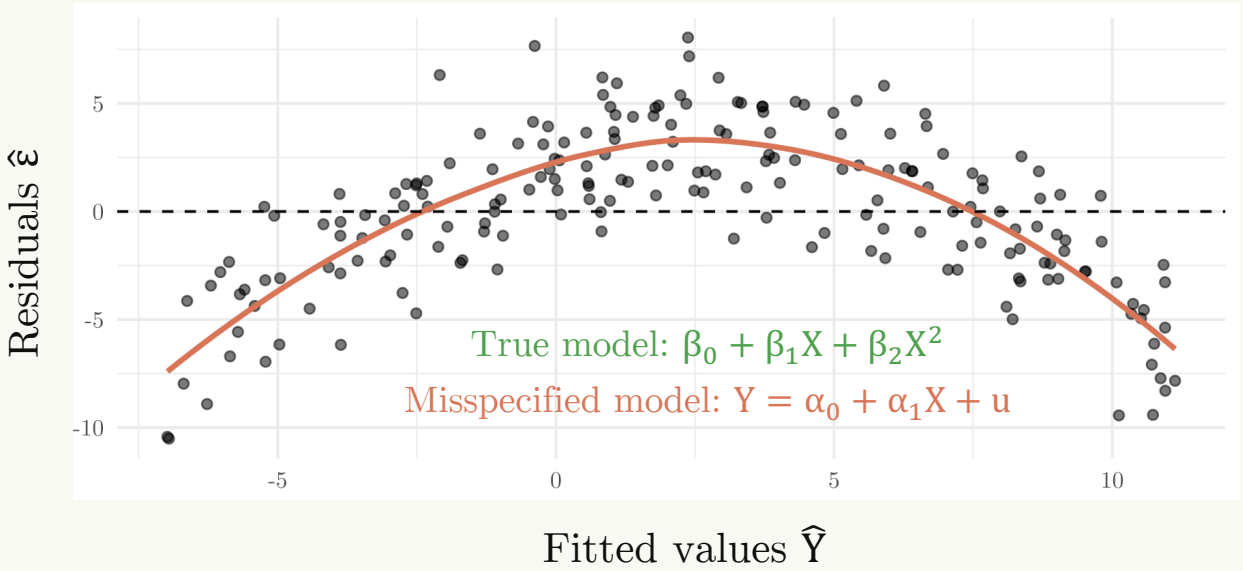
But we estimate $Y = \alpha_0 + \alpha_1 X + u$

Then $E(\widehat{\alpha_1}) = \beta_1$.

This is linear **in parameters**.
Linearity assumption is not about variables.

Assumptions | Diagnostics for linearity

$E[Y \mid X] = X\beta$ implies $E[\hat{\varepsilon}_i \mid X_i] \approx 0$ if our model is correctly specified.



Residual vs. fitted plot allows for visual test of assumption $E[\hat{\varepsilon}_i \mid X_i] \approx 0$

Curvature indicates violation of linearity assumption.

Assumptions | What to do about non-linearity

Linearity is an assumption about parameters, not variables.
We can **transform variables** to create linearity in the conditional mean.

We can **transform the outcome**:

$$\log(Y) = \beta_0 + \beta_1 X + \varepsilon$$

Useful for strictly positive or right-skewed Y, or multiplicative effects.
$1 - \exp(\beta_1)$ % increase in Y for 1-unit increase in X.

We can **transform predictors**:

$$Y = \beta_0 + \beta_1 \log(X) + \varepsilon$$

Useful for diminishing returns to X.
$\beta_1/100$ increase in Y for 1% increase in X.

We can **add polynomial terms**:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Useful for inverted-U relationships between X and Y.
Direction of association depends on value of X.

Assumptions | Homoskedasticity

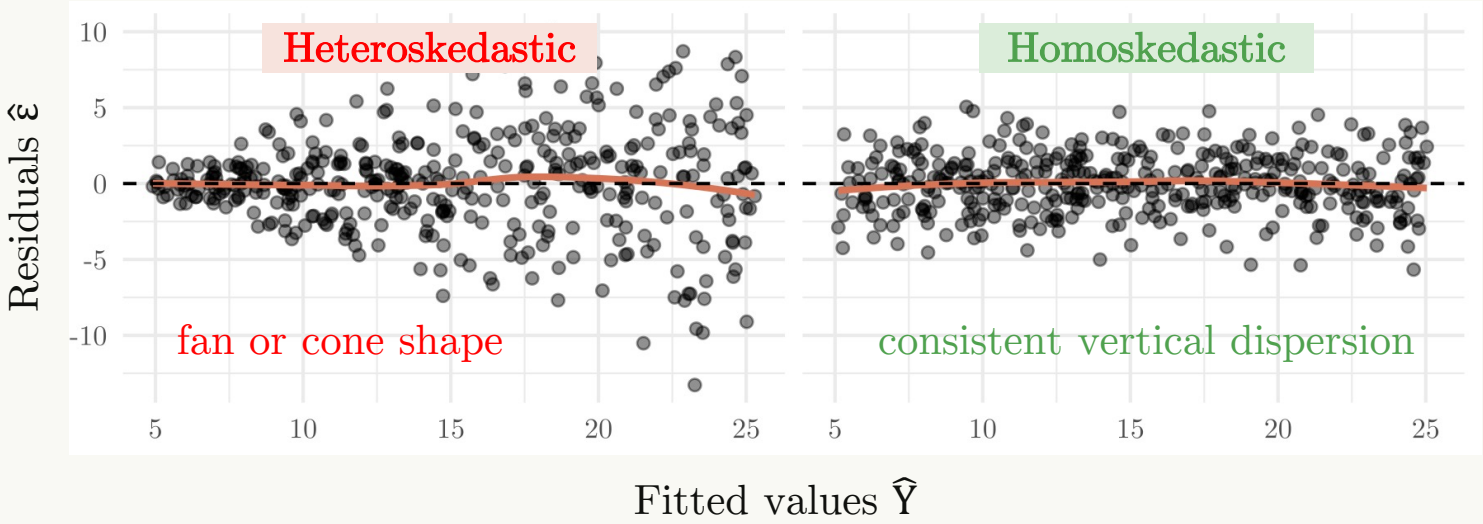
We assume that **variation in the error term ε is constant across regressors X**:

$\text{Var}(\varepsilon_i \mid X_i) = \sigma^2$ → the **spread of unexplained variation** does not depend on X

If errors are heteroskedastic, standard errors are incorrect.
OLS coefficients are still unbiased, but CIs, hypothesis tests, and p-values are misleading.

Assumptions | Diagnostics for homoskedasticity

As a visual test for homoskedasticity, we look at **dispersion in the residuals vs. fitted plot:**



To address heteroskedasticity, use **robust (HC) standard errors**. Almost always a good idea!

Assumptions | Independence

We assume that **error terms ε are uncorrelated across observations.**
After accounting for X, what remains must not be systematically related across observations.

$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j.$ → regression analogue of **i.** in **i**.i.d. assumption (Week 2)

Independence is a **design assumption** that fails for specific types of data:

Clustered data: Errors share a group-level component. e.g. users across platforms

Longitudinal data: Errors are auto-correlated. e.g. cohort health data over time

Network data: Observations influence each other. e.g. citation networks

If independence is violated, standard errors are incorrect because standard formulas overestimate the amount of useful signal in our data. Effective sample size is $<$ n!

Assumptions | What to do about independence

There is no general statistical test to verify independence in observational data.
Independence (like exogeneity) is a design assumption, not a statistical property.

For clustered data with known structure, we can use **clustered standard errors**:

$$\text{Var}(\hat{\beta}) = (X'X)^{-1}\left(\sum_{j=1}^{G} X_j' \hat{\varepsilon}_j \hat{\varepsilon}_j' X_j\right)(X'X)^{-1}$$

Depending on data type, we can model the dependence directly:

Mixed-effects models that eliminate within-group correlation in the error term.

Time-series models that include dependence across subsequent observations.

We will probably cover these in Week 8.

Assumptions | Multicollinearity

We assume **no perfect multicollinearity** for mathematical reasons:

The OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ requires $\mathbf{X}'\mathbf{X}$ to be invertible.

→ fails when one regressor is an exact linear combination of others.

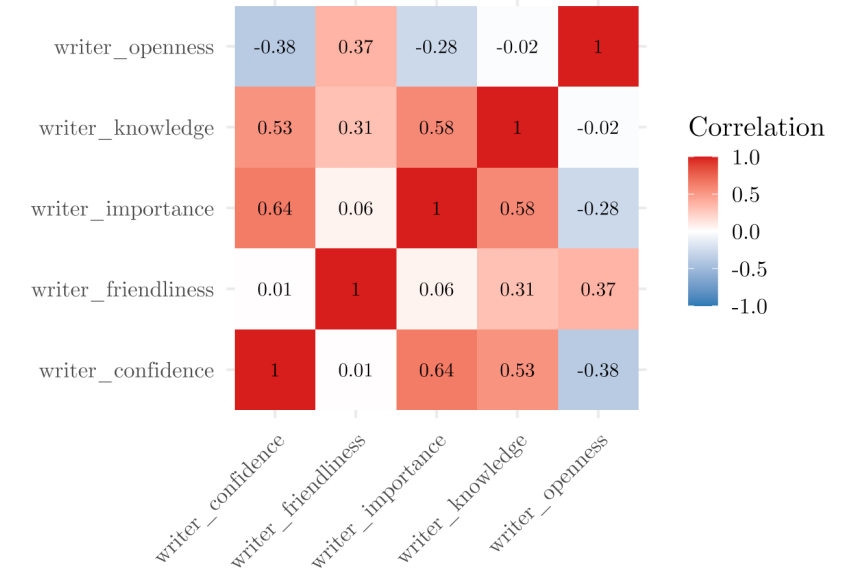
Under **strong multicollinearity**, unbiased estimation is possible, but SEs are inflated because the variance of coefficient estimates depends on the covariance structure of the regressors:

$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ where $\text{Var}(\hat{\beta}_j) = \sigma^2 c_{jj}$ which grows with multicollinearity.

OLS estimates coefficients using **variation in X that is independent of other regressors**.
If X_1 and X_2 move together almost perfectly, this makes it hard to identify separate effects.

Assumptions | Diagnosing multicollinearity

We can check for strong collinearity between **pairs of regressors** in a **correlation matrix**:



(from tutorial)

Data: human ratings of political opinion paragraphs (0-100 scale)

High pairwise correlation is a warning sign.

Rule of thumb: $|r|>0.8$ is a problem.

Assumptions | Diagnosing multicollinearity (cont'd)

However, a **single regressor may be explained by a combination of others**.
We can check for multivariate multicollinearity using the **Variance Inflation Factor**:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$ where R_j^2 comes from regressing X_j on all other regressors.

If R_j^2 is high, most variation in X_j is explained by other regressors.

$\sqrt{\text{VIF}_j}$ = how much larger the SE is compared to if X_j had 0 correlation to other regressors.

Rule of thumb: $\text{VIF} > 10$ indicates high multicollinearity that should be addressed.

To **address multicollinearity**, combine collinear variables or drop redundant ones.
(if necessary!)