Lecture 4: Univariate Linear Regression

Paul Röttger

Applied Analytical Statistics

10th of February 2026

Correction | One-sided hypothesis testing

For a two-sided test: $H_0\colon \mu = \mu_0$ and $H_A\colon \mu \neq \mu_0$ → all possible outcomes

For a one-sided test: $H_A\colon \mu > \mu_0$ … **but what is H_0?**

In principle, we would want H_0 to be complementary: $H_0\colon \mu \leq \mu_0$

However, hypothesis testing requires us to **specify a single null distribution**.
This is why it is convenient to **use a simple** null $H_0\colon \mu = \mu_0$ even in the one-sided case.

In practice, $H_0\colon \mu \leq \mu_0$ and $H_0\colon \mu = \mu_0$ usually lead to equivalent results because the latter describes the "worst case". H_0 is easier to reject for values $\mu < \mu_0$ than for $\mu = \mu_0$.

→ only $\mu = \mu_0$ is relevant for controlling α and calculating p-values.

Details | Holm-Bonferroni correction for multiple comparisons

Family-wise error rate (FWER) is the probability of making one or more false discoveries, i.e. Type I errors, when performing multiple hypothesis tests.

Without correction, FWER increases as we perform more tests that each have fixed α.

Simple Bonferroni correction controls FWER by dividing per-test α by number of tests $\mathbf{m}$.

Holm-Bonferroni correction is uniformly more powerful → less increase in Type II error rate

Sort p-values from smallest to largest: $p_{(1)}, \ldots, p_{(m)}$.

Compare $p_{(k)} \leq \frac{\alpha}{m-k+1}$ **sequentially**, starting from $p_{(1)}$.

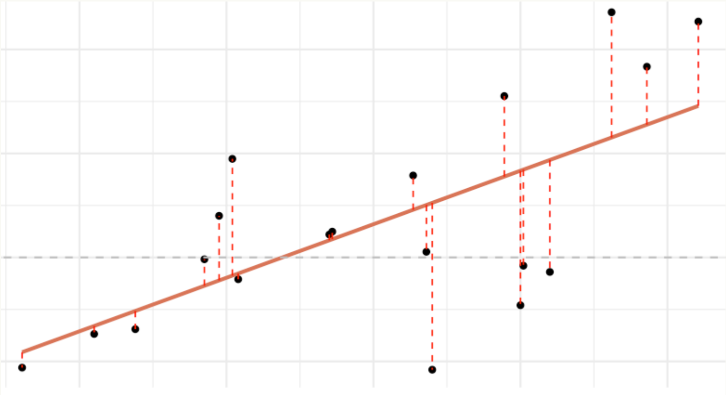
Stop once a test fails to reject null, do not reject null for any larger p-values.

$$p_{(1)} \leq \frac{\alpha}{m}$$

$$p_{(2)} \leq \frac{\alpha}{m-1}$$

$$\cdots$$

Plan for today | Univariate linear regression



Regression is one of the most powerful tools in the statistical analysis toolkit.

Today we **start with its most simplest form**: one outcome, one predictor, linear model

In the next weeks, we keep expanding: multiple predictors, non-linear models, ...

In the second half of class, we will **discuss summative expectations**.
Then, we will run a **peer feedback session** for your draft summative proposals.

Regression | Motivation

Regression provides a **unified framework** for describing the relationship between variables: One **outcome variable** vs. one or multiple **predictor variables**.

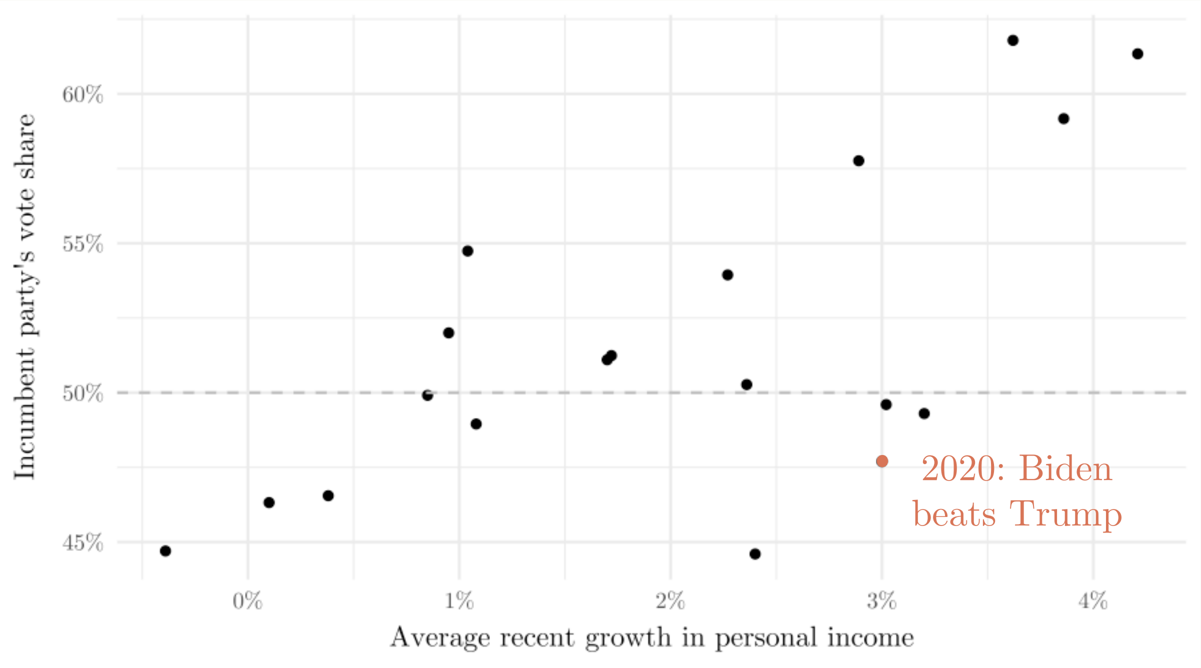
1. Quantifying **strength and direction** of associations.
2. Expressing **uncertainty** about associations.
3. Drawing **conclusions** about statistical hypotheses.

Familiar ideas:
correlations
confidence intervals
hypothesis tests

Regression **generalises more specialised tools for analysis**: more than just two variables, different variable types, ...

In this course, we mainly use regression for **analytical inference** but regression can also be a powerful tool for prediction.

Simple linear regression | Working example



Data: US election results vs. economic performance

Source: RegOS + updates

🤔 What kind of relationship can we **plausibly assume** here?

Simple linear regression | Population model

The **simple linear regression model** specifies the relationship between an outcome Y and a single predictor X **at the population level** based on coefficients β_0 and β_1:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where ε_i is the **error term**, capturing unobserved factors affecting Y_i.

We **assume** that the conditional mean of Y given X is linear in X :

$$E[Y \mid X] = \beta_0 + \beta_1 X$$

so that

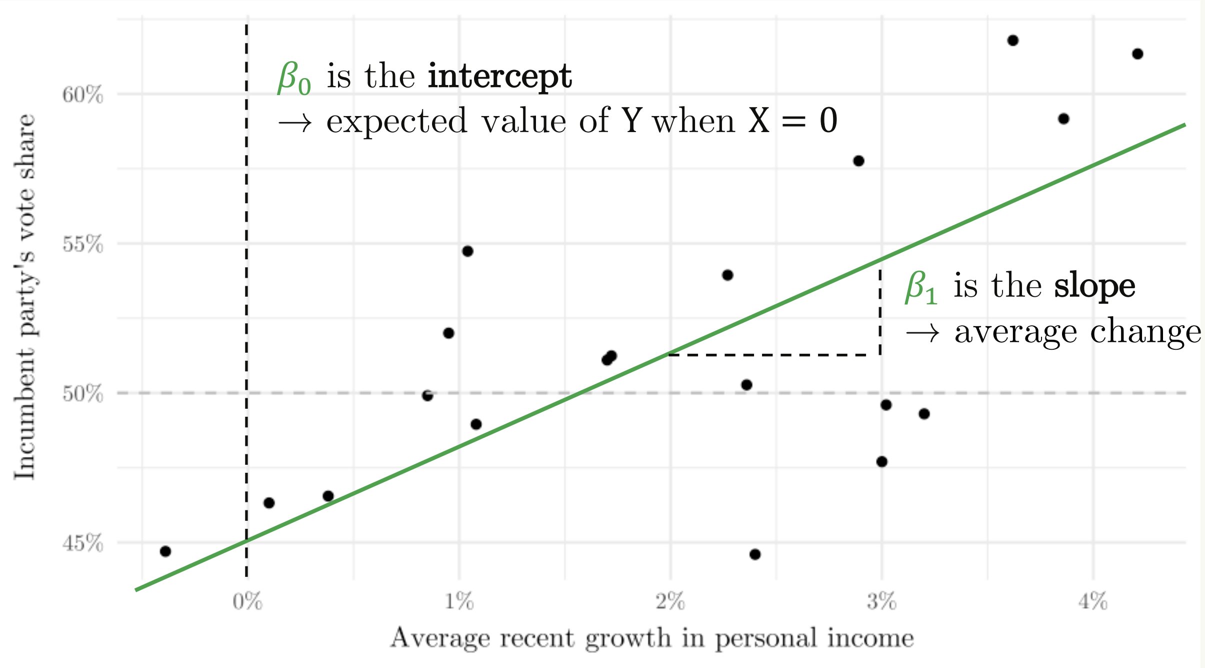
$$Y_i \ = E[Y \mid X_i] + \varepsilon_i$$

This is a **modelling assumption** about the average relationship between X and Y.

All models are wrong, but some are useful!

George Box (1919–2013)

Simple linear regression | Interpreting regression coefficients



β_0 is the **intercept**
→ expected value of Y when X = 0

β_1 is the **slope**
→ average change in Y for a one-unit increase in X

Example population model:
$\beta_0 = 45, \quad \beta_1 = 2.5$
$\rightarrow E[Y \mid X_i] = 45 + 2.5X_i$

Simple linear regression | Estimated model

As always, we do not observe the population but one finite, noisy sample.
The parameters in our assumed population model are fixed but unknown: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

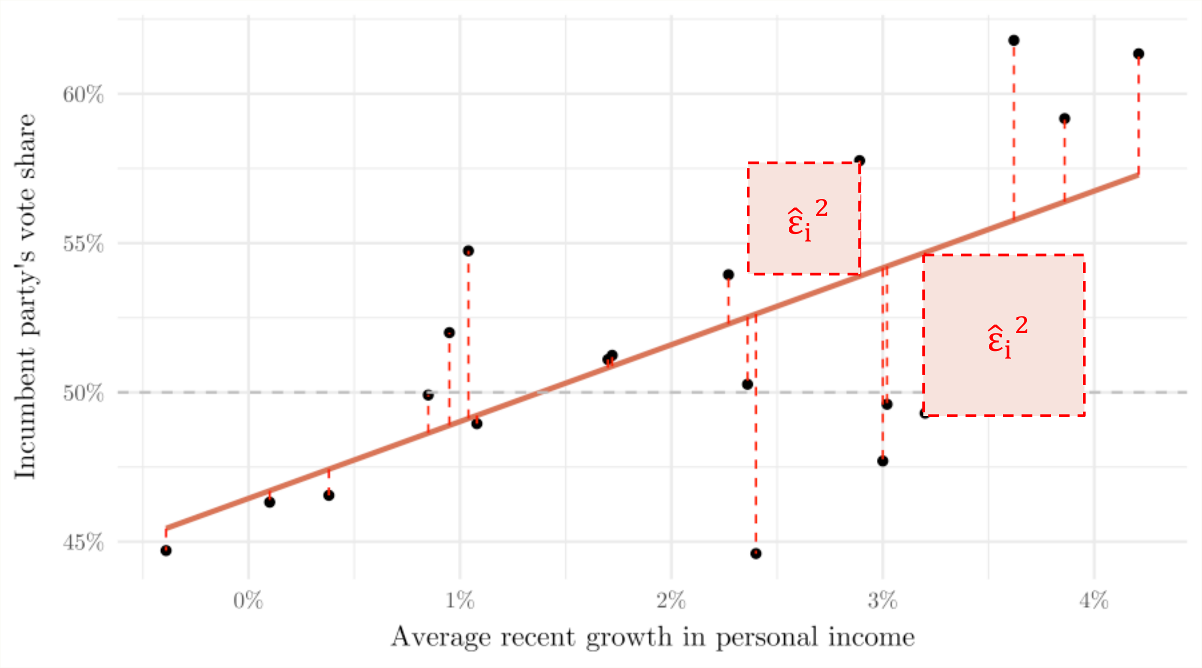
As always, we want to obtain sample estimates of population parameters: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
Here, our regression coefficients are statistics that vary across samples, i.e. random variables.

🤔 How do we choose $\hat{\beta}_0$ and $\hat{\beta}_1$? → find a "line of best fit"

The residuals $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ are the observed errors from our estimated model.

The most common method for estimating $\hat{\beta}_0$ and $\hat{\beta}_1$ is by **ordinary least squares (OLS)**:
We choose $\hat{\beta}_0$, $\hat{\beta}_1$ to **minimise the sum of the squared residuals** $\sum_i \hat{\varepsilon}_i^2$

Ordinary Least Squares | Residuals and OLS



$\hat{\varepsilon}_i$ is the vertical distance between the estimated ("fitted") regression line and observation i.

We select intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ to minimise the sum of squared residuals $\hat{\varepsilon}_i$

🤔 Why do we **square** residuals before minimising their sum?

Ordinary Least Squares | Deriving OLS estimators

We choose $\hat{\beta}_0$, $\hat{\beta}_1$ to minimise the sum of squared residuals / residual sum of squares (RSS):

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \mathrm{RSS} = \min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

We solve this **minimisation problem** by taking partial derivatives and setting to zero:

$$\frac{\partial \mathrm{RSS}}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \quad \Rightarrow \sum_{i=1}^{n} \hat{\varepsilon}_i = 0.$$

→ **residuals sum to zero**

geometrically: $\hat{\varepsilon} \perp \mathbf{1}$ (from intercept)

$$\frac{\partial \mathrm{RSS}}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} X_i \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0 \quad \Rightarrow \sum_{i=1}^{n} X_i \hat{\varepsilon}_i = 0.$$

→ **residuals are orthogonal to the predictor**

geometrically: $\hat{\varepsilon} \perp \mathbf{X}$ (from slope)

Ordinary Least Squares | Deriving OLS estimators (cont'd)

By solving the system of equations on the previous slide we get to:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad \text{and} \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

→ exact closed-form solutions

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the **OLS estimators** of population slope β_1 and population intercept β_0.

We know that $\mathrm{Var}_n(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ and $\mathrm{Cov}_n(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\,(Y_i - \overline{Y})^2$

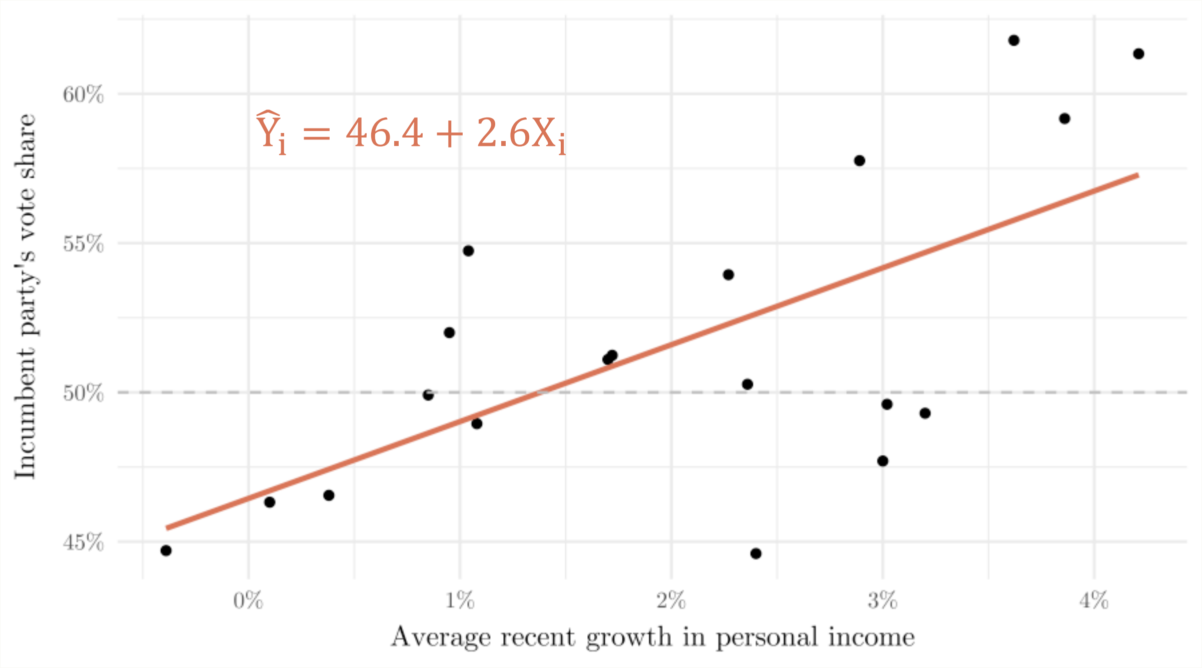
Therefore: $\hat{\beta}_1 = \frac{\mathrm{Cov}_n(X,Y)}{\mathrm{Var}_n(X)}$.

how strongly do X and Y move together?`

→ enables per-unit interpretation of slope coefficient

how much does X itself vary?

Ordinary Least Squares | Fitted regression model

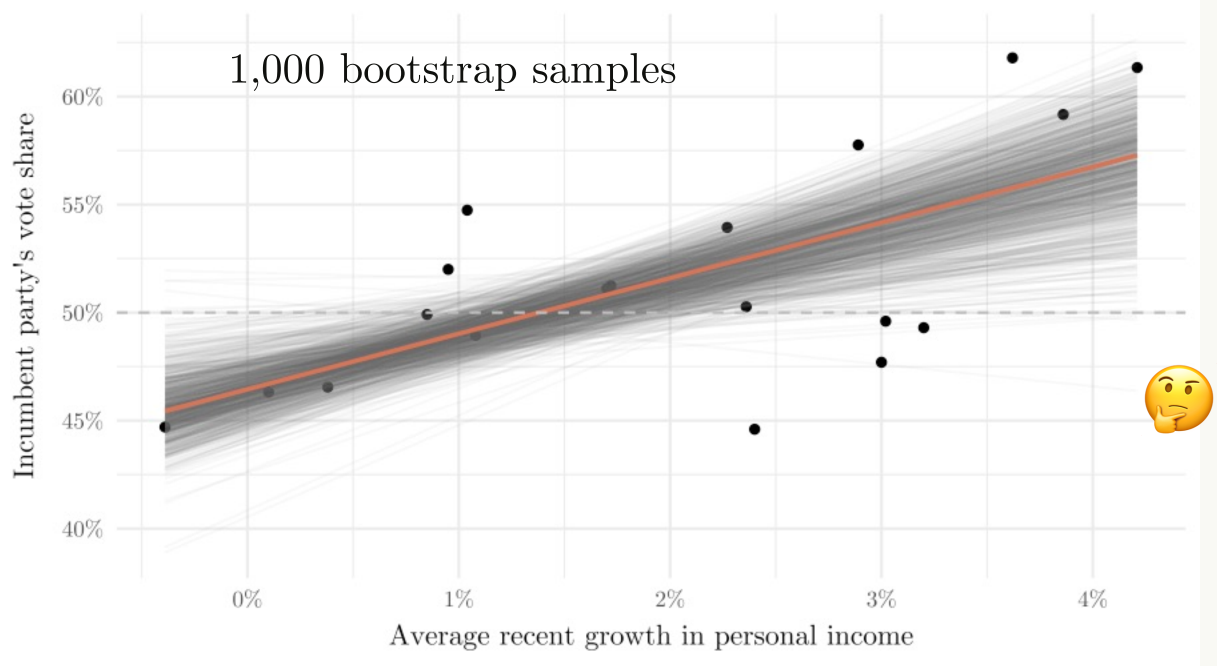


→ expected change

$\widehat{\beta}_1 = 2.6$: a 1pp increase in “average recent growth in personal income” is associated with a 2.6pp increase in “incumbent party’s vote share”

$\widehat{\beta}_0 = 46.4$: for 0% “average recent growth in personal income”, the expected “incumbent party’s vote share” is 46.4%.

Uncertainty in regression | Coefficients as random variables



We derived $\hat{\beta}_0$ and $\hat{\beta}_1$ as functions of our sample.

$\hat{\beta}_0$ and $\hat{\beta}_1$ will vary across repeated samples.

🤔 How do we quantify uncertainty?

As in previous weeks, we want to **approximate the sampling distribution** in order to perform inference.

Uncertainty in regression | Sampling distribution of OLS

Under **standard OLS assumptions**, the OLS slope has a well-defined sampling distribution:

→ linearity, independence, homoskedasticity: will cover next week

$$\hat{\beta}_1 \sim N(\underbrace{\beta_1}_{\text{population level}}, \mathrm{Var}(\hat{\beta}_1)) \quad \text{where} \quad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{and} \quad \mathrm{Var}(\varepsilon|X) = \sigma^2$$

population level

→ homoskedasticity

🤔 What is the intuition behind the different terms in $\mathrm{Var}(\hat{\beta}_1)$?

σ^2 is the population variance of the error term, given X.
→ irreducible noise in outcome after controlling for predictor

$\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the spread of our predictor.
→ horizontal information in predictor, growing with n

Familiar problem: σ^2 is a fixed but unknown population parameter.

Uncertainty in regression | Coefficient standard error

Our goal is to **characterise the sampling distribution of our regression coefficients**, so that we can quantify uncertainty around our estimated coefficients.

The **standard error of coefficient $\boldsymbol{\widehat{\beta}_1}$** is the standard deviation of its sampling distribution:

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{X}_\mathrm{i} - \overline{\mathrm{X}})^2}} \qquad \text{where} \qquad \hat{\sigma}^2 = \frac{1}{n-2} \sum \hat{\varepsilon}_i^2$$

🤔 Why 1/(n-2) in $\hat{\sigma}^2$?

residual df = 2, one for each estimated parameter

Larger samples n → less uncertainty.
Larger error terms $\hat{\varepsilon}_i^2$ → more uncertainty.
More variation in X → less uncertainty.

Uncertainty in regression | Coefficient confidence intervals

Definition: A 95% confidence interval (**CI**) is a **procedure** that, in repeated sampling, produces intervals that contain the true population parameter 95% of the time.

Now that we know $\mathrm{SE}(\hat{\beta}_1)$, we can apply the same logic as for sample statistics (Week 2):

$$\mathrm{CI}_{1-\alpha} = \left[\hat{\beta}_1 \pm t_{\alpha/2,\, n-2} \times \mathrm{SE}(\hat{\beta}_1)\right]$$

🤔 Why the t distribution instead of the normal?

We use $\hat{\sigma}^2$ to estimate σ^2 when calculating $\mathrm{Var}(\hat{\beta}_1) = \mathrm{SE}(\hat{\beta}_1)^2$.
This **uncertainty propagates** and we need to adjust for in our sampling distribution.

→ same intuition as for z-test vs. t-test (Week 3)

We adjust by using the t distribution, which has heaver tails than the normal distribution.
As before, each estimated coefficient uses up one degree of freedom $df = n - 2$.

Uncertainty in regression | Hypothesis tests about coefficients

We test claims about population parameters: H_0: $\beta_1 = 0$ vs. H_A: $\beta_1 \neq 0$

Under the null, there is no linear association between X and Y in the population.
Knowing predictor X, on average, does not provide information about outcome Y.

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

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

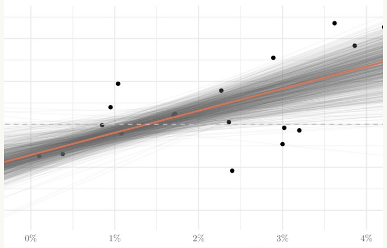
Under standard OLS assumptions and H_0: $\beta_1 = 0$, the t statistic is t-distributed: $\mathrm{T} \sim t_{n-2}$

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 $p = \Pr(|\mathrm{T}_{n-2}| \geq |\mathrm{t}| \mid \mathrm{H}_0)$

Uncertainty in regression | Bootstrap for coefficients

Analytical SEs and CIs rely on theoretical assumptions about the sampling distribution.
Instead, we can estimate the sampling distribution of $\hat{\beta}_0$, $\hat{\beta}_1$ directly from the data.



- We treat our sample as a **stand-in for the population**.
- We repeatedly draw samples **with replacement**.
- We **recompute the** coefficient for each sample.
- We use the resulting "**bootstrap distribution" of the** coefficient as a stand-in for the unobserved sampling distribution.

We can **estimate coefficient SEs** based on the bootstrap distribution.
We can **construct coefficient CIs** from percentiles of the bootstrap distribution.
We can **reject $\mathbf{H_0}$** at α significance level if the null value lies outside the bootstrapped $\mathtt{CI}_{1-\alpha}$.

Goodness of fit | Decomposition of outcome variation

Goodness of fit describes how well our fitted regression model fits our data.

To quantify goodness of fit, we first **decompose the variation in our outcome Y**:

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

Residual Sum of Squares (RSS):
Unexplained variation, in residuals

Total Sum of Squares (TSS):
Total variation in outcome Y

Explained Sum of Squares (ESS):
Variation explained by fitted regression

Total variation in the outcome is the sum of explained and unexplained variation.

$$\text{TSS} = \text{ESS} + \text{RSS}$$

These are all sample statistics that we can easily calculate.

Goodness of fit | R^2 coefficient of determination

In linear regression, the most common goodness-of-fit measure is R^2

$$R^2 = \frac{\text{Explained Sum of Squares (ESS)}}{\text{Total Sum of Squares (TSS)}} = \%\ \text{of total variation explained by fitted model}$$

R^2 is a proportion, so measured on a 0-1 scale. In simple linear regression, $R^2 = \mathtt{Corr}(\mathrm{X}, \mathrm{Y})^2$.

Small R^2 = limited practical significance, even if coefficients are statistically significant.

🤔 When is a large R^2 achievable? When is it not?

→ simple mechanical process vs human behaviour

🤔 When is a small R^2 acceptable?

→ okay for explanation, not prediction

Simple linear regression | Categorical predictors

We focused on scalar predictors, but regression can also handle categorical predictors:

Data: Human ratings (0-100) for answers from LLM A vs. LLM B for 50 questions.

Assumed population model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where $X \in \{0,1\}$ indicates LLM $\in \{A, B\}$.

Let's assume $\hat{\beta}_0 = 50$ and $\hat{\beta}_1 = 20$. 🤔 What is the interpretation of these coefficients?

The expected (mean) human rating for LLM A is 50.

LLM B is rated, on average, 20 points higher than LLM A.

In this setting, testing $H_0\colon \beta_1 = 0$ is the same as testing for equality of group means.
→ **Regression generalises the t-test framework!**

Recap | Key takeaways from Week 4

Linear regression assumes a linear population model.
We model the conditional mean of an outcome as a linear function of a predictor.

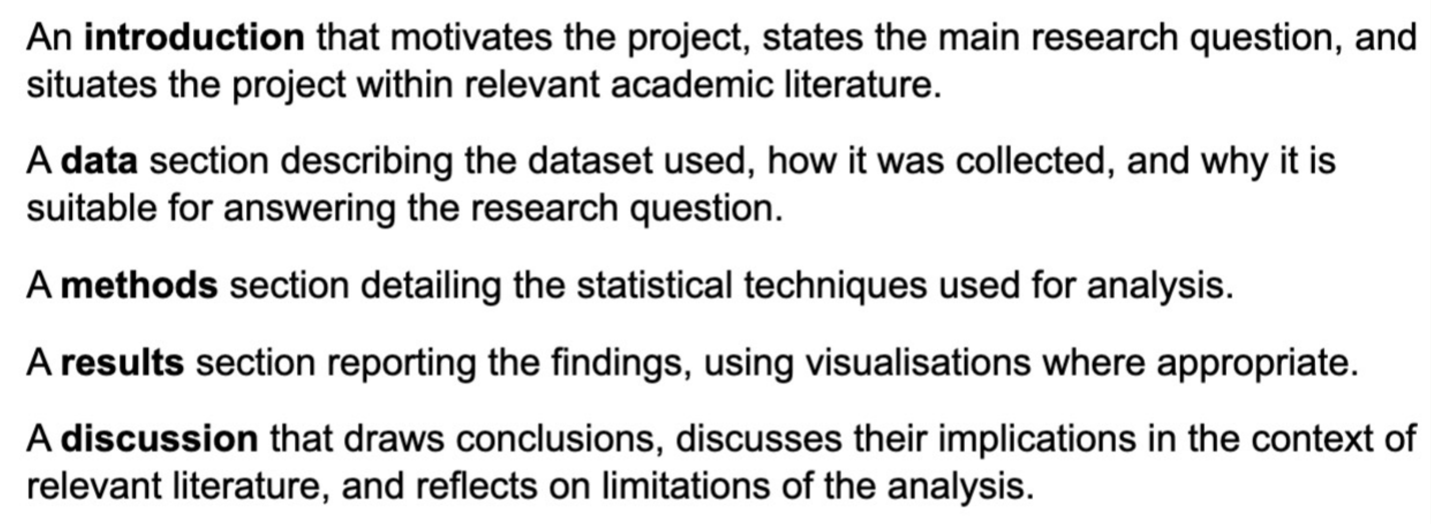
OLS estimates regression coefficients by minimising squared residuals.
Residuals sum to zero and are orthogonal to the predictor, yielding closed-form estimators.

Regression coefficients are random variables with sampling distributions.
Uncertainty depends on sample characteristics, summarised by standard errors.

Inference in regression extends the logic of earlier weeks.
We approximate sampling distributions to obtain CIs and run hypothesis tests.

Summative | Structure

You are asked to write a **4,000-word research paper**.



An **introduction** that motivates the project, states the main research question, and situates the project within relevant academic literature.

A **data** section describing the dataset used, how it was collected, and why it is suitable for answering the research question.

A **methods** section detailing the statistical techniques used for analysis.

A **results** section reporting the findings, using visualisations where appropriate.

A **discussion** that draws conclusions, discusses their implications in the context of relevant literature, and reflects on limitations of the analysis.

Summative | Choosing a research question

Your goal is to **demonstrate your mastery** of the statistical methods taught in this course.

Pick a question that concerns **the relationship between two or more variables**

RQ: To what extent is worker productivity associated with worker AI use?

... or a **difference across groups**.

RQ: Are workers who use AI more productive than those who do not?

Avoid causal language unless your data is explicitly set up to enable causal analysis.

~~**RQ**: Does using AI make workers more productive?~~

Summative | Choosing a research question (cont'd)

This is **not a replication exercise**, so do not just take an existing RQ + dataset.

You will have to **explain why your RQ is interesting / relevant**.
Clear motivation will be rewarded, but you are not expected to do groundbreaking research.

Try to find a **RQ that relates to your research interests**.
This will make the summative more worthwhile for you (and usually lead to better results).

You can work on your MSc thesis topic, but keep in mind that the emphasis is different:

- Here, you need to focus on applying statistical methods from this course.
- Here, you will not get credit for collecting novel data.

You are not graded on finding strong effects. You are graded on strong design & reasoning.

Summative | Choosing a research question (cont'd)

The best projects will do more than just run a single regression or hypothesis test.

Set yourself up to run more than just one analysis, by nesting research questions

RQ1: To what extent is worker productivity associated with worker AI use?

RQ2: How does this association vary across different levels of worker experience?

... or considering different ways of operationalising your constructs of interest

number of tasks completed vs. lines of code written

... and incorporating different methods we covered in class.

analytical CIs and bootstrap CIs

Explaining the choices you make is a great way of demonstrating your understanding!

Summative | Choosing a dataset

I strongly suggest not collecting new data specifically for this course.
If you have new data from another project or plan on using the data elsewhere – fine.

You will have to **describe your chosen dataset and its collection** in the summative but you will not get any credit for having collected the data yourself.

If you are unsure where to start, I recommend looking for:

- A reputable source (e.g. a trusted organisation, a well-published paper)
- A large-ish sample size (i.e. 100s or 1,000s of observations rather than a handful)
- A good number variables (e.g. dozens of questions in a survey)

Never assume that you can use a dataset. Request access, download, explore asap!

Summative | Choosing and describing your methods

Your **primary method of analysis** should be a method from this course.

In scope methods include hypothesis testing, regression, bootstrapping...

Out of scope methods include predictive ML methods like XGBoost and SVMs...

You can use more sophisticated/complex methods <u>that build on the methods in this course</u>, but you have to justify why and clearly demonstrate your understanding.

If you are unsure about whether your method is in scope, please ask me.
Mikhail will not be involved in grading your summative, so he cannot speak on this.

Compared to a regular research paper, **this summative places extra emphasis on methods**.
This means: explain modelling choices, check assumptions, show diagnostics, test robustness...

Class activity | Group assignment based on your RQs

G1: **Language, Communication, and Bias in AI & Media** – Caleb Agoha, Noha Mahgoub, Yunjia Qi

G2: **AI, Generative Models, and Evaluation** – Max Davy, Howard Leong, Audrey Yip

G3: **Media, Platforms, and Audience Response** – Sophie Bair, Charlotte Peart, Michi Wong

G4: **Po** **Institutions** – Celikhan Baylan, Grahm Gaydos, Caleb Tan

G5: **So** **Adoption** – Min Jung, Mia Kussman, Isaac Backer

G6: **Health, Medicine, and Neuroscience** – Amelia Mercado, Laura Wegner, Ines Trichard

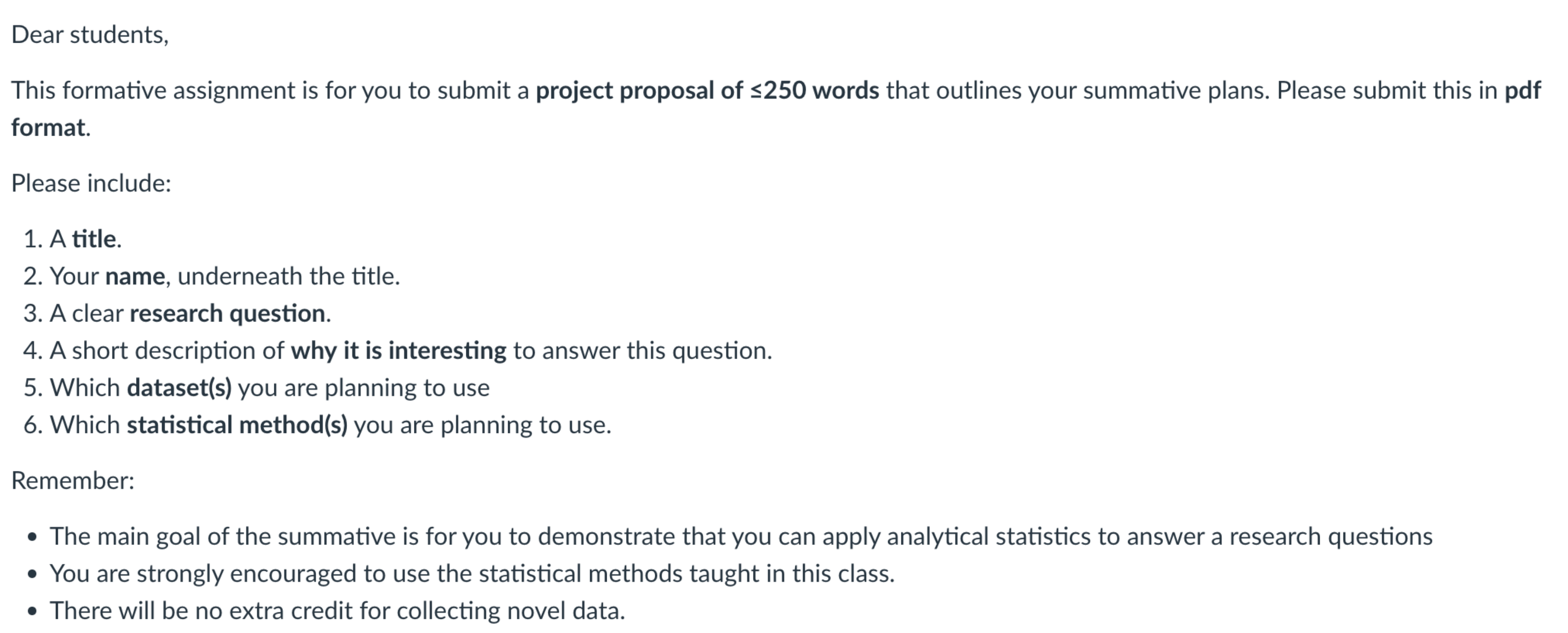
G7: **Education, Labor, and Socioeconomic Outcomes** – Rehmat Arora, Yilin Qian, Yue Zhang

G8: **Culture, Mobility, Lifestyle** – Teo Canmetin, Alena Tsvetkova, Fucheng Wang, Nesma Hammouda

Same groups as in Week 2!

Everyone not named: please get together in groups of 3.

Class activity | Proposal instructions



Dear students,

This formative assignment is for you to submit a **project proposal of ≤250 words** that outlines your summative plans. Please submit this in **pdf format**.

Please include:

1. A **title**.
2. Your **name**, underneath the title.
3. A clear **research question**.
4. A short description of **why it is interesting** to answer this question.
5. Which **dataset(s)** you are planning to use
6. Which **statistical method(s)** you are planning to use.

Remember:

- The main goal of the summative is for you to demonstrate that you can apply analytical statistics to answer a research questions
- You are strongly encouraged to use the statistical methods taught in this class.
- There will be no extra credit for collecting novel data.

Class activity | Peer review questions

Is there a clear **motivation** for this project?

Does this project seem **feasible** within the constraints of this course?

Does the **dataset** seem appropriate for answering the RQ?

Does the **statistical method** seem appropriate for answering the RQ?
  Does it seem **in scope** for this course?

Does the project/RQ **avoid causal claims** that the project's design cannot support?

When discussing with your groupmates, try to be a **constructive reviewer**:
How could the project be improved? How could limitations be addressed and overcome?