Inferential statistics

SingStats

What's ahead:

We will go through the fundamentals of statistical modelling:

- Hypothesis testing - theoretical overview
- t-tests (statistically compare two groups)
- OLS regression

Inferential statistics

Statistics is a branch of mathematics that deals with collecting, organising, analysing, reading and presenting data. It can also be further divided into:

- **Data collection and handling**
- **Descriptive statistics:** How the data can be summarized.
- **Inferential statistics:** Using the data to make predictions.

Example: Return to Office in 2023

Assume that there is an imaginary organization where the admin is trying to estimate how many days each employee aims to spend on-site in offices A and B. The idea is to check whether the same pattern is observed in all offices.

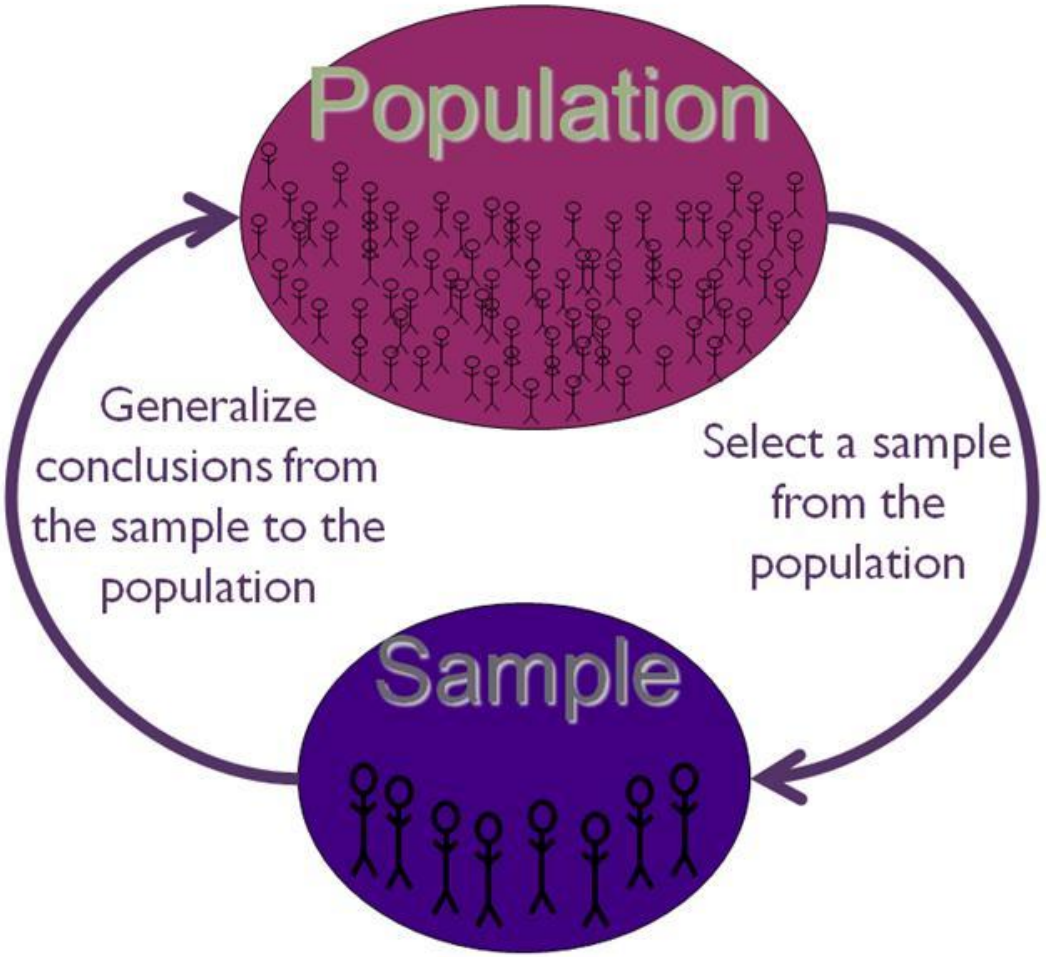
Admin is working very hard to gather accurate information. They contact some employees in offices A and B to ask them about their intentions. One key thing to address is that different people of the company are equally represented. (**Data Collection**).

Once the data have been collected, they have initial ideas. They know the plans made by employees of different departments or of age groups. (**Descriptive statistics**).

Based on all the information collected (departments, worker composition, ect), they can try to predict what number of days of the employees will work in either office - or another company. (**Inferential statistics**)

From sample to population

Statistics is the science of using the sample to make inference about the entire population.



In our example...

We could only rely on the collected samples and make a verdict like:

In the sample collected, we got that in office A, employees aim to spend **3 days/week** in the office, while on site B it's **2.3 days/week**!

This statement ignores the following problem:

What if the difference is just by chance?

In order to make inference about the whole population from our sample, we need to make some assumptions about the whole population.

Hypotheses

Every problem in inferential statistics starts with a hypothesis.

In all the problems of this nature that we encounter here, there are two kinds of hypotheses:

Null hypothesis (H_0): the "no effect" or "status quo" statement.

Alternative hypothesis (H_1 or H_a): the claim you'll conclude if data contradicts H_0.

The alternative hypothesis can fully or partially contradict the null hypothesis.

Coming back to our example:

Option 1:

- H0: The mean number of hours that employees want to spend on-site on site A and B are (statistically) the same.
- H1: The mean number of hours that employees want to spend on-site on site A is (statistically) **greater** than the mean number of hours that employees want to spend on-site on site B.

Option 2:

- H1: The mean number of hours that employees want to spend on-site on site A is (statistically) **lower**.

Option 3:

- H1: The mean number of hours that employees want to spend on-site on site A is (statistically) **different**.

NULL HYPOTHESIS

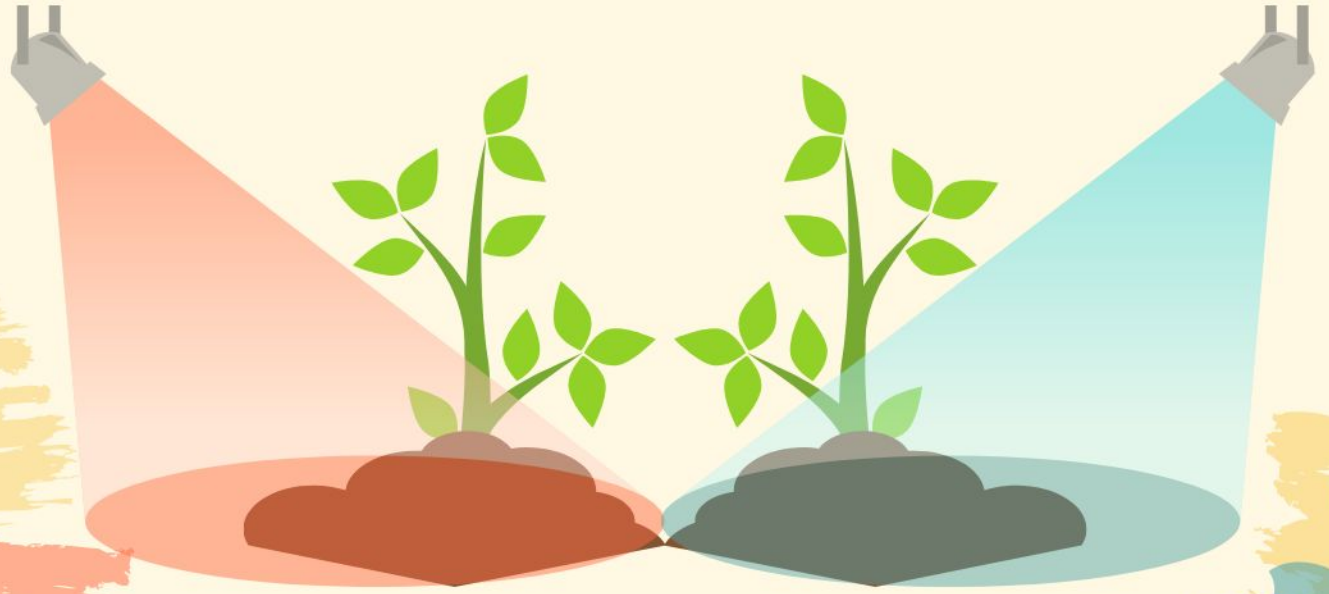
THE NULL HYPOTHESIS ASSUMES THERE IS NO DIFFERENCE BETWEEN TWO GROUPS.

NULL HYPOTHESIS
H_0

LIGHT COLOR HAS NO EFFECT ON PLANT GROWTH

ALTERNATIVE HYPOTHESIS
H_A

LIGHT COLOR AFFECTS PLANT GROWTH



SCIENCENOTES.ORG

Another example

- H0: Light color has no effect on plant growth

- H1: Light color affects plant growth

But, where is the zero in this null hypothesis? The zero is hidden in the more mathematical formulation of the above hypotheses.

- H0: The difference in the growth of plants under different lights is 0.

- H1: The difference in the growth of plants under blue light and the plants under red light is different than 0.

In summary…

Null hypothesis (H_0): the "no effect" or "status quo" statement.

Alternative hypothesis (H_1 or H_a): the claim you'll conclude if data contradicts H_0.

You pick a critical region (the set of values that lead you to reject H_0). That critical region is chosen so that, if H_0 is actually true, you'll only land in it with some small probability α.

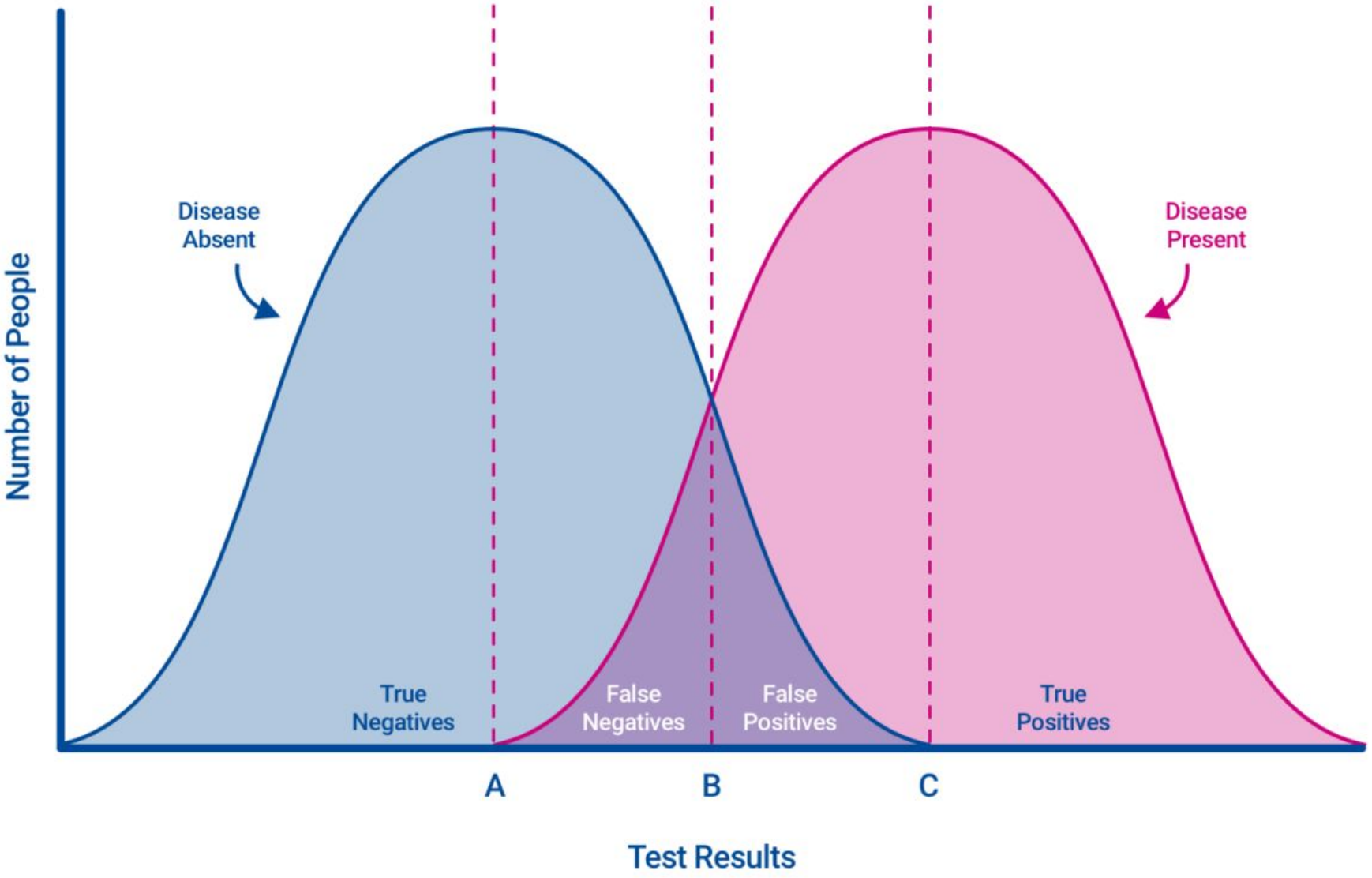
From Definitions to Decisions

We've defined

- Null hypothesis (H_0)
- Alternative hypothesis (H_1)

Next: How we actually perform a hypothesis test

Key idea: reduce our data to one number → compare to a known distribution and decide whether to reject the null


Number of People
Disease Absent
Disease Present
True Negatives
False Negatives
False Positives
True Positives
A
B
C
Test Results

Type I vs Type II error

You decide to get tested for COVID-19 based on mild symptoms. There are two errors that could potentially occur:

Type I error (false positive): the test result says you have coronavirus, but you actually don't.

Type II error (false negative): the test result says you don't have coronavirus, but you actually do.

Type I Error (Alpha)

Definition
The probability of rejecting H_0 when H_0 is true.

Interpretation
It's your tolerance for "false alarms."
By convention, alpha is usually 5%.

Choosing α
Lowering α → smaller chance of a false positive → critical region shrinks.
But shrinking the region → making it harder to reject H_0 → increases β

Errors and its types

		Reality	
		Cancer	No cancer
Mammogram	Cancer	9 (True positive)	79 (False positive)
	No cancer	1 (False negative)	911 (True negative)

Errors and its types

Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

Key Takeaways

1. **α (Type I rate):** risk of a false positive. You set it before seeing the data. Usually 5%.

2. **β (Type II rate):** risk of missing a real effect. It depends on α, sample size, effect size, and variability.

3. **Trade-off:** reducing α (being stricter) makes it harder to achieve high power unless you boost sample size or target a larger effect.

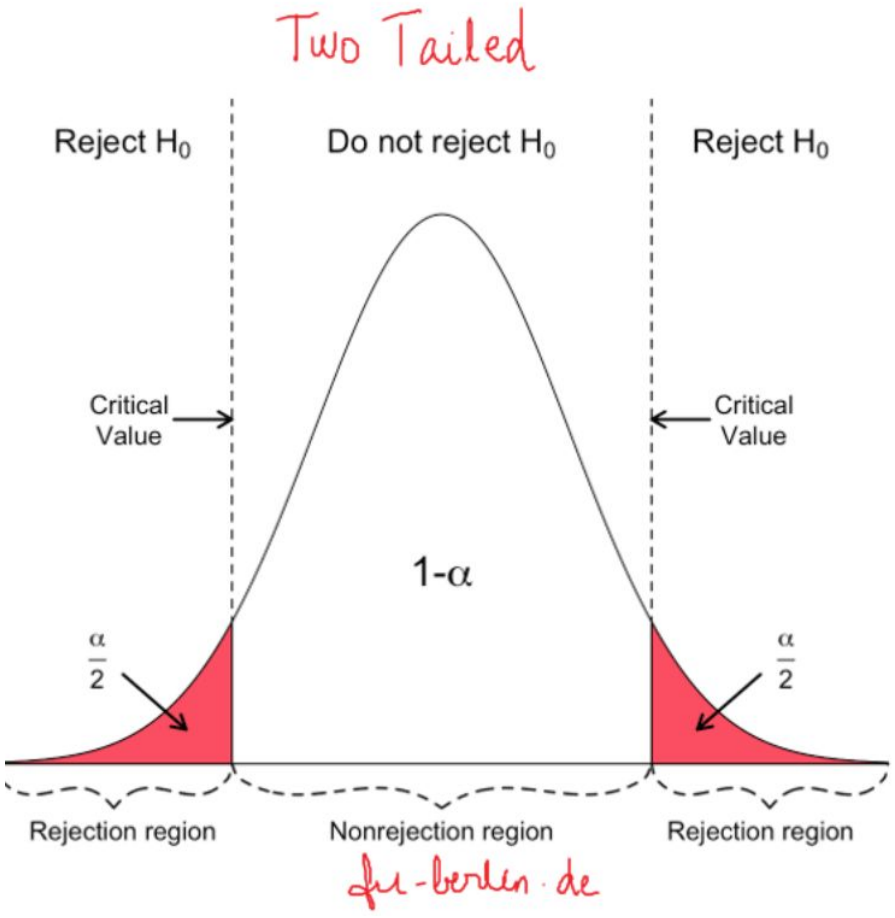
Test Statistic & Decision Rule

Goal: Reduce your sample data into a single summary number

Makes it easy to decide if the data are consistent with the null hypothesis

Key Steps:

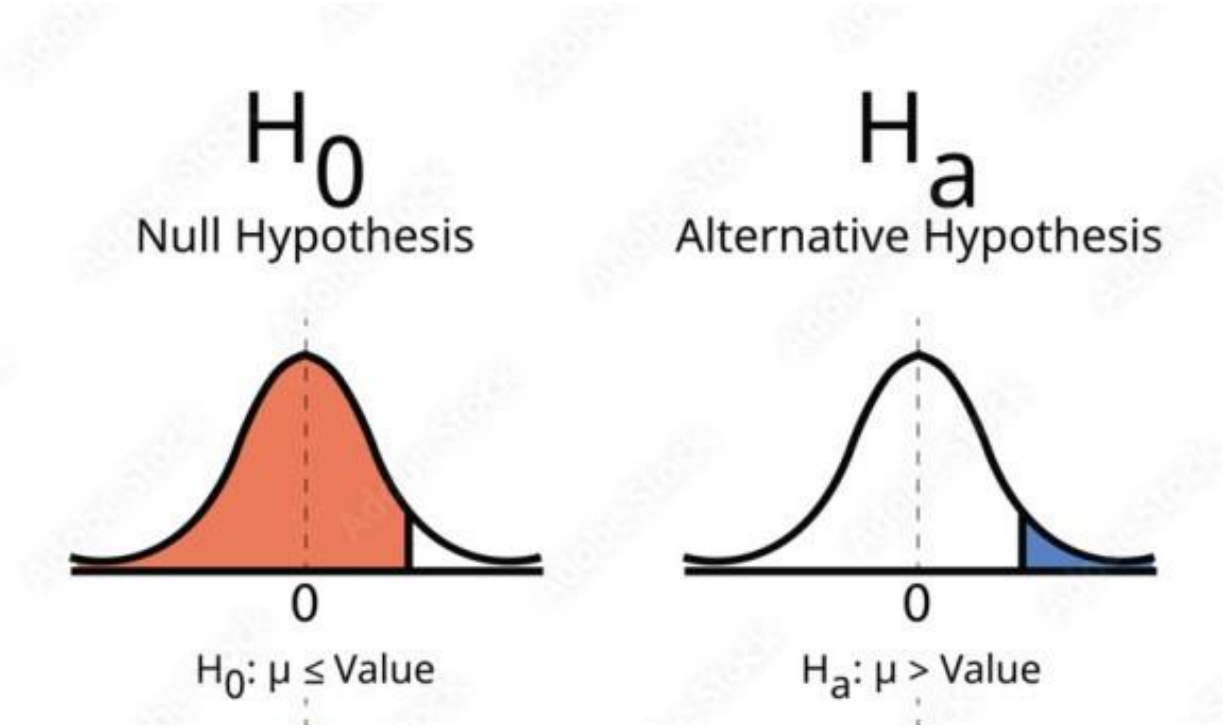
- Compute the test statistic
- Choose a significance level, α (e.g. 0.05) and find the critical threshold, tα/2, df
- Decision rule: if t is within the critical threshold, do not reject



P-value (MOST important!)

Probability of randomness - how sure are we that the result is random / not meaningful?

We want this probability to be less than 5% or 0.05.



P-value in a t-test

Probability that the two groups of data are a part of the same population.

We usually want this probability to be less than 0.05.

When & Why Use the t-Test

T-tests assess *differences between groups*.

Applicability:
Sample size small (n < 30) or population variance unknown
Data roughly normally distributed

Variants:
One-sample t-test: compare sample mean to a known value
Two-sample t-test: compare means of two independent groups
Paired t-test: compare means of matched or repeated measurements

Design	"Compare"	Example
One-sample	mean vs. benchmark	Is the average score ≠ 75?
Independent groups	group A mean vs. group B mean	Drug vs. placebo on blood pressure
Paired observations	before vs. after on same subject	Weight before & after diet program

Example: Paired t-Test

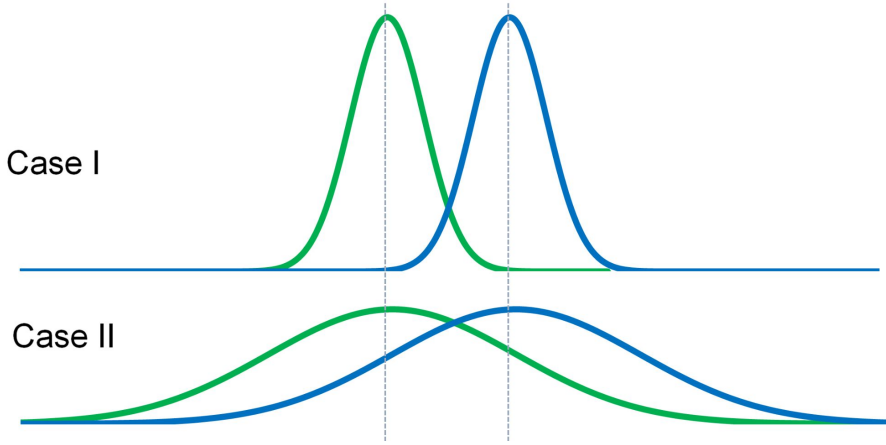
Context: Measure reaction time of 12 students

- **Condition A:** no caffeine
- **Condition B:** after two cups of coffee

Results: $\bar{d}=-35$ ms, $t=-3.2$, $p=0.008$

Interpretation:

- **p-value (0.008):** only a 0.8% chance of observing this if caffeine had no effect → **reject** Ho
- **Confidence interval:** we're 95% confident
- **Practical take-away:** coffee appears to speed up reaction by ~35 ms on average



What is regression analysis?

Regression analysis also assesses the relationship between two or more (typically continuous) variables. Specifically:

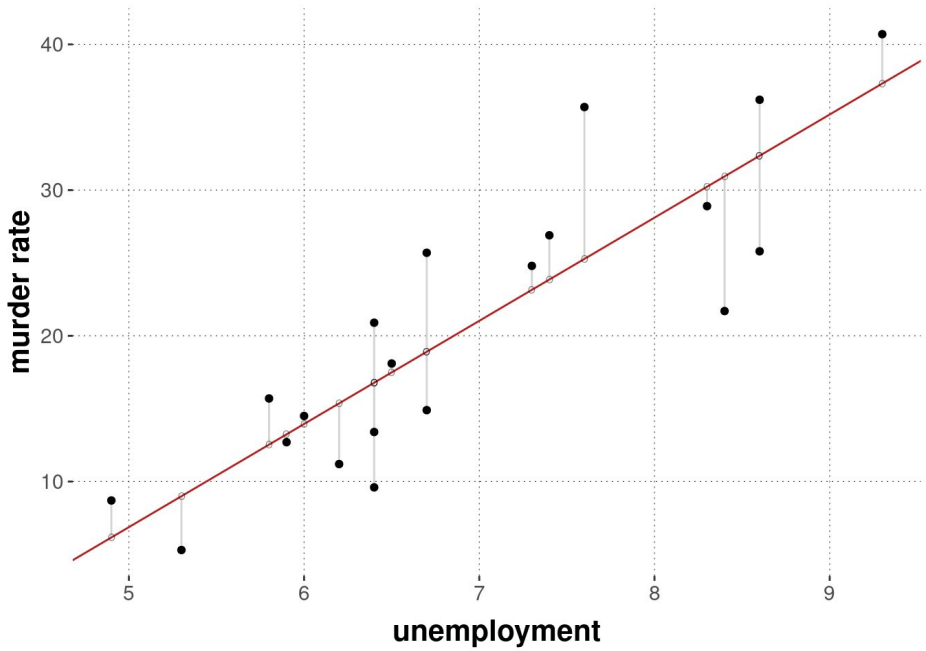
- **Whether** there an association between two (or more) variables
- The **strength** of the association between two (or more) variables

What OLS Does: Summary

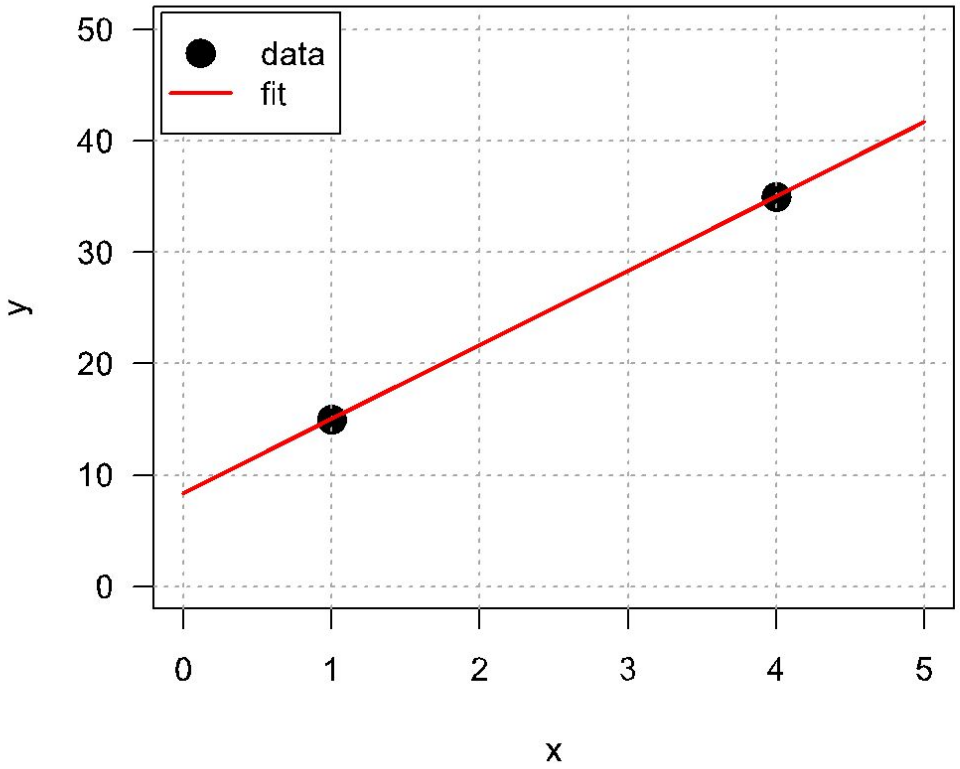
- **Goal:** Model the relationship between a predictor x and outcome y

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- **Line-Fitting:**
  - Fits a straight line through data
  - Chooses parameters to **minimize** the sum of squared errors
- **Error Term (ε)**
  - Captures all variation **not** explained by the line

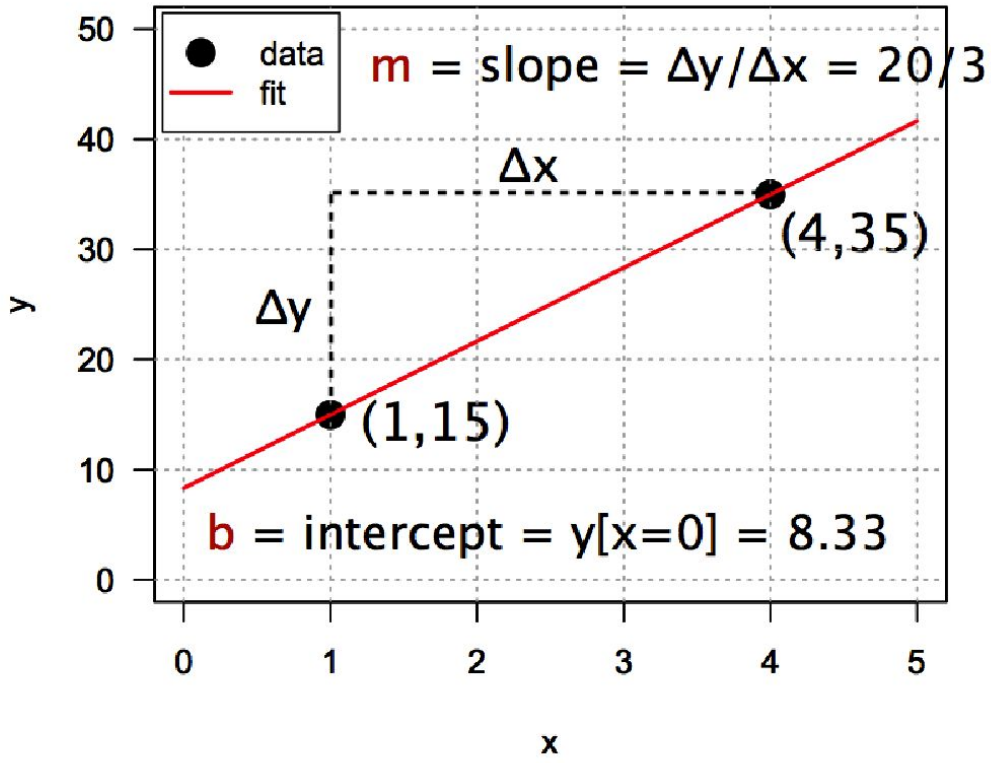


Example



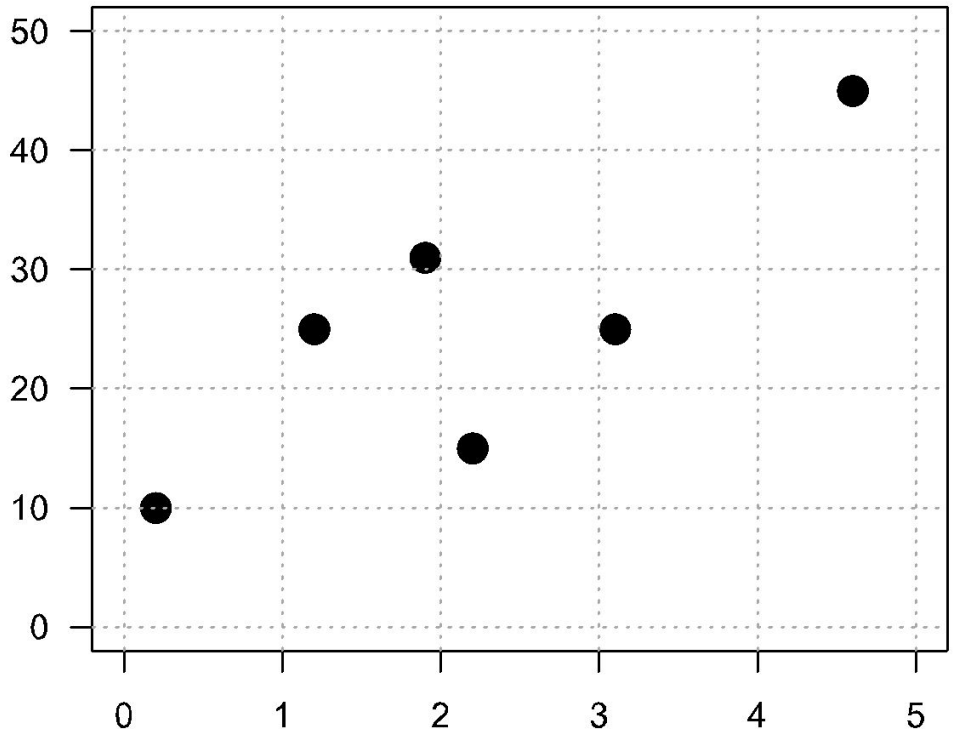
Example

Slope:



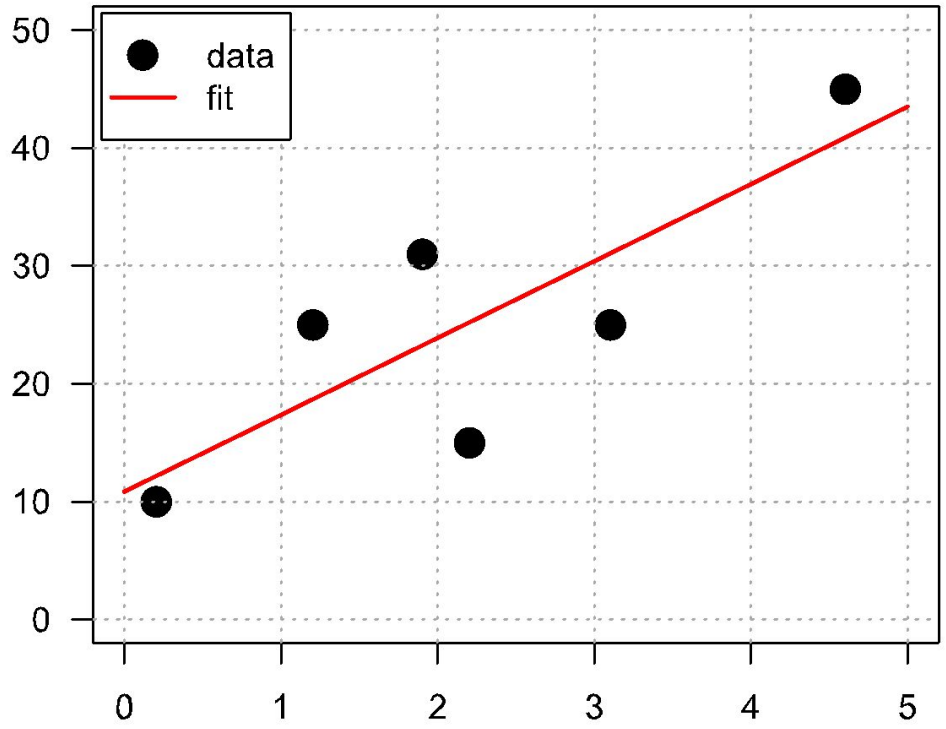
The real world

What does the real world give us? Noise!



The real world

What does the real world give us? Noise!



Describing relationships: OLS

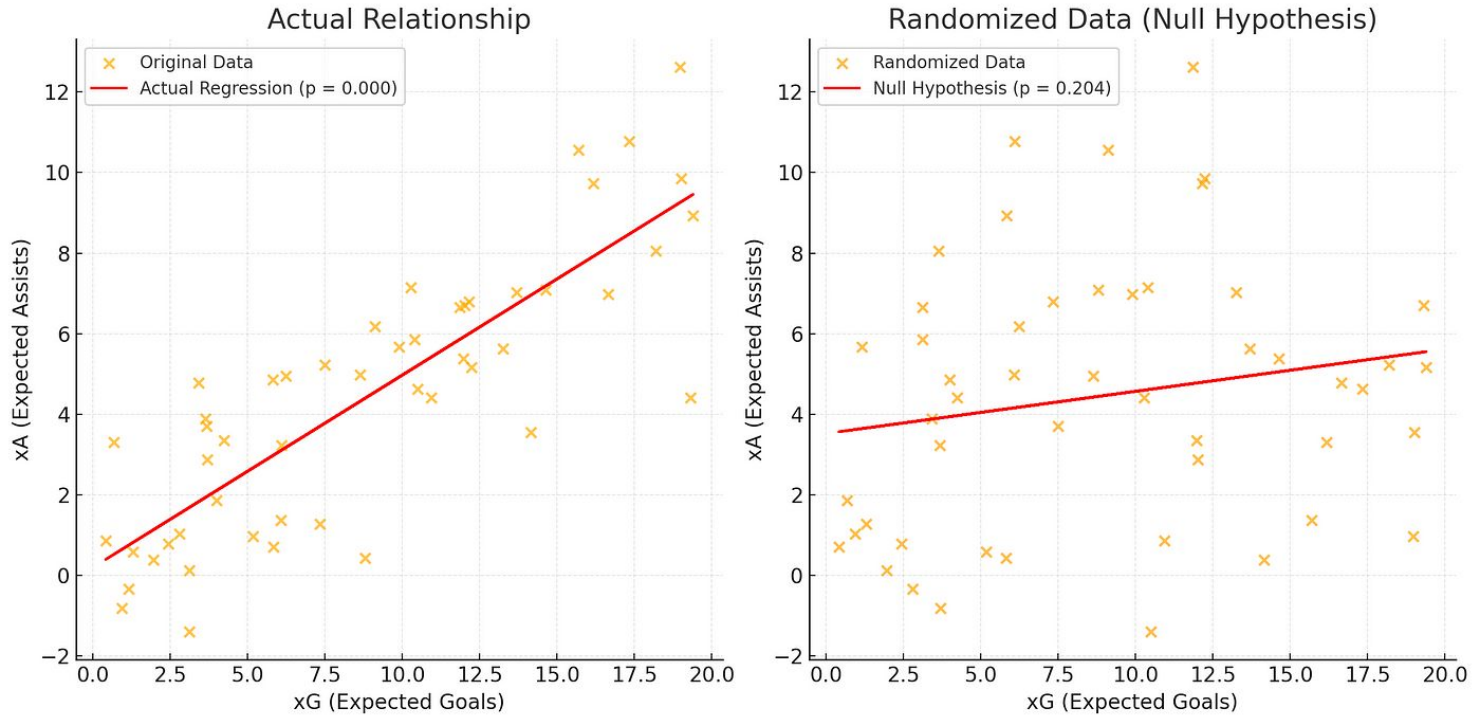
We need some system to consistently estimate the relationship between variables when there is 'noise' in the data (which there almost always is!)

Using Ordinary Least Squares (OLS regression) is the most common way of doing this.

P-value in a Regression

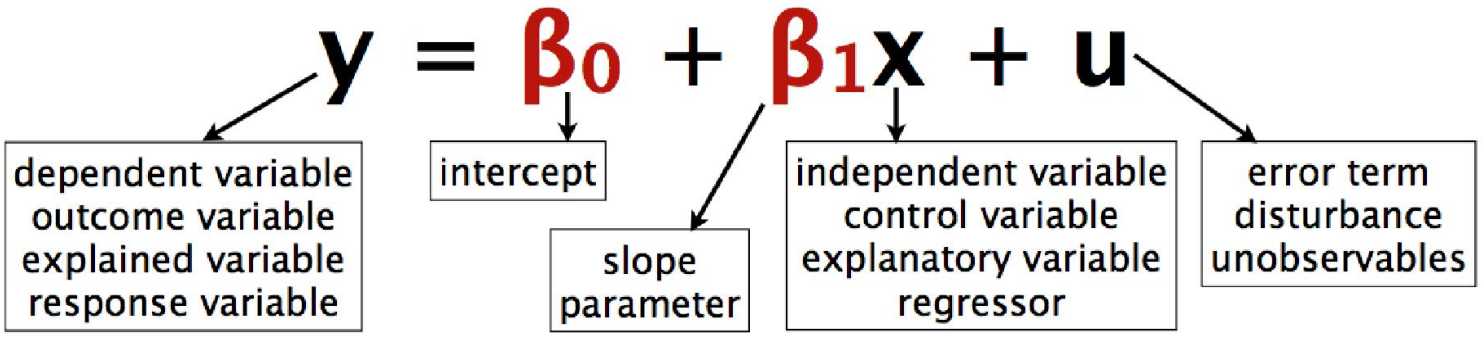
Probability that the two variables do not have a meaningful relationship.

We usually want this probability to be less than 5%.



Describing relationships

When we use OLS regression, we slightly alter the equation:

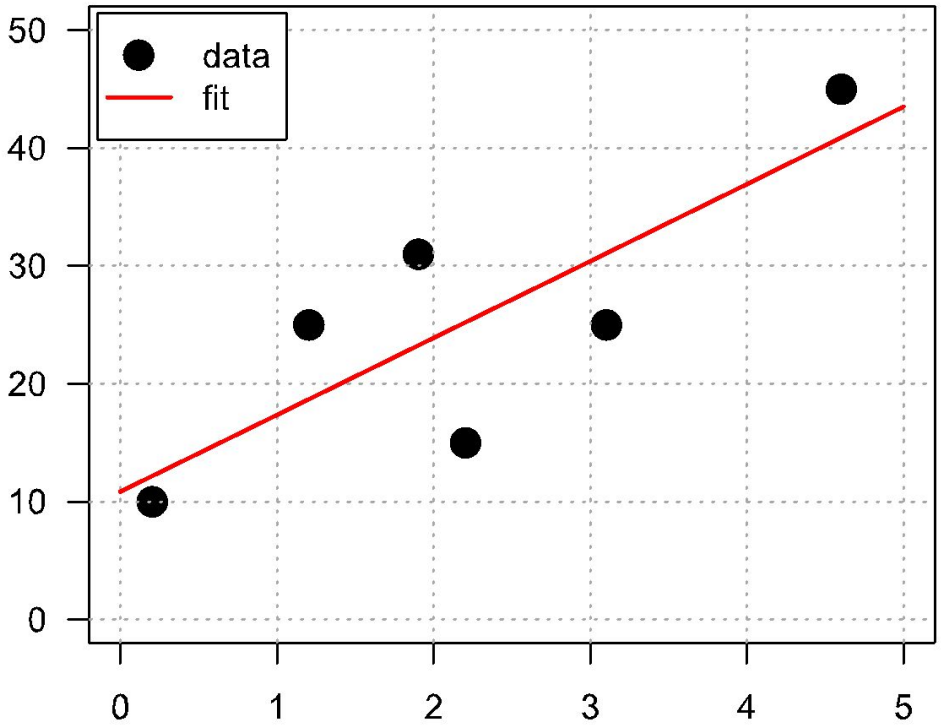


Refining the ideas: *e*

The error term (e) contains unobserved factors.

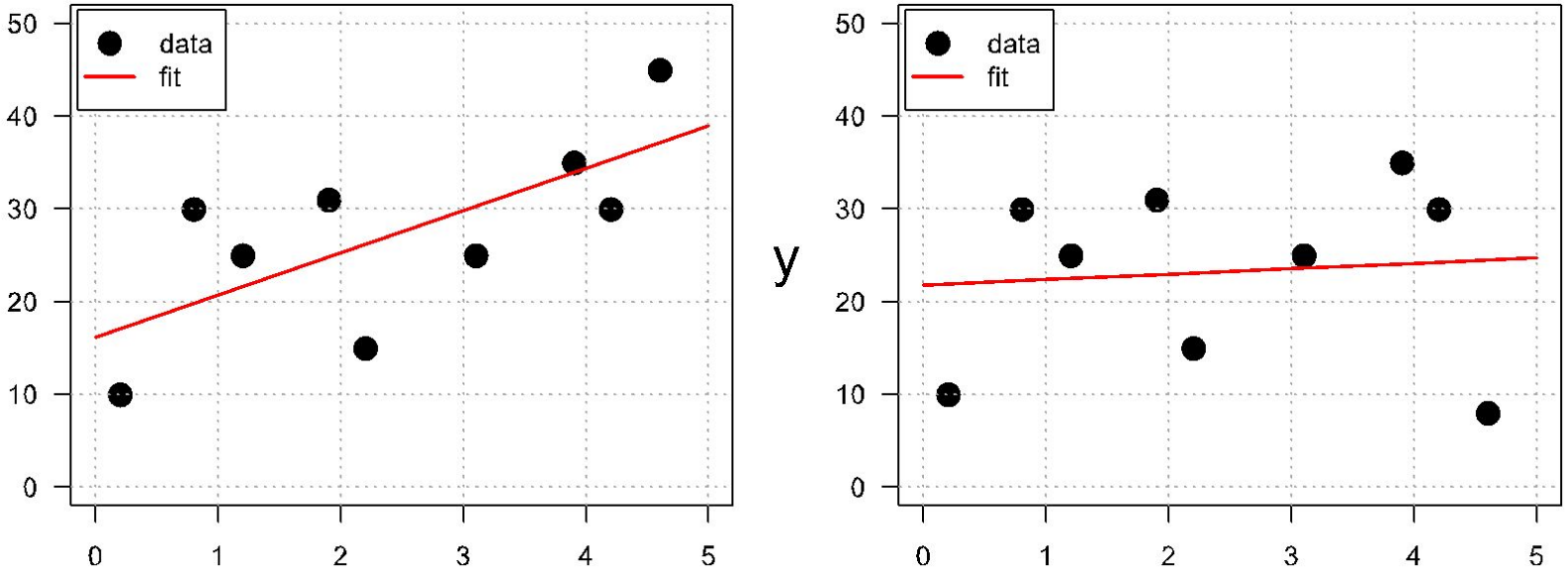
$\hat{y}_i$ (y hat) = predicted value of y for a given x

The residual (u) is the difference between (predicted) $\hat{y}_i$ and (actual) y_i



Why OLS?

- It's computationally simple
- Avoids big misses (minimizing sum u^2)
- But, it is sensitive to outliers



An Example

We might ask whether there is a relationship between spending and salary.

What would this model look like?

$$Salary\text{A} = \beta_0 + \beta_1 spending\text{A} + e$$

How would we interpret β_0?
How would we interpret β_1?
How would we interpret e?
How would we describe this regression?

Interpreting β0, β1, R-squared, and p-values

Slope β^1:

- Estimated change in y for a one-unit increase in x
- E.g. β^1=2.5 → on average, y rises by 2.5 units per 1x

Intercept β^0:

- Predicted y when x=0 (context-dependent interpretation)

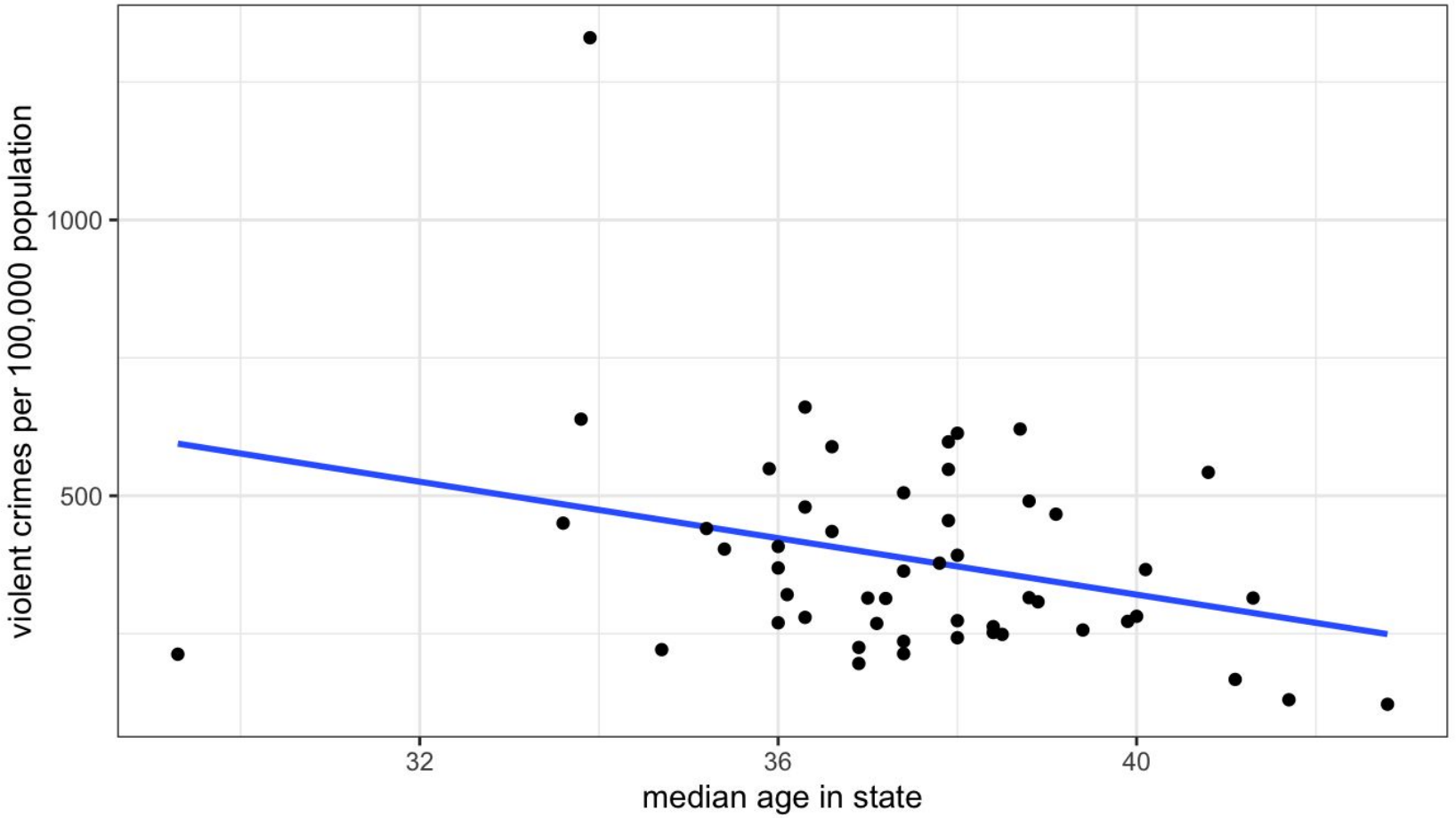
R-squared:

- Proportion of total variance in y explained by the model
- Ranges from 0 (no fit) to 1 (perfect fit)

p-Values (for β1):

- Small p (e.g. < 0.05) → evidence that x is a significant predictor

Example


violent crimes per 100,000 population
1000
500
32
36
40
median age in state

Regression Output in Python

OLS Regression Results

Dep. Variable:	Calorie_Burnage	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.006
Method:	Least Squares	F-statistic:	0.04975
Date:	Tue, 29 Sep 2020	Prob (F-statistic):	0.824
Time:	17:48:00	Log-Likelihood:	-1145.8
No. Observations:	163	AIC:	2296.
Df Residuals:	161	BIC:	2302.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	346.8662	160.615	2.160	0.032	29.682	664.050
Average_Pulse	0.3296	1.478	0.223	0.824	-2.588	3.247

Omnibus:	124.542	Durbin-Watson:	1.620
Prob(Omnibus):	0.000	Jarque-Bera (JB):	938.541
Skew:	2.927	Prob(JB):	1.58e-204
Kurtosis:	13.195	Cond. No.	811.

Regression Output in Python

OLS Regression Results

Dep. Variable:	y	R-squared:	0.669
Model:	OLS	Adj. R-squared:	0.667
Method:	Least Squares	F-statistic:	299.2
Date:	Mon, 01 Mar 2021	Prob (F-statistic):	2.33e-37
Time:	16:19:34	Log-Likelihood:	-88.686
No. Observations:	150	AIC:	181.4
Df Residuals:	148	BIC:	187.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.2002	0.257	-12.458	0.000	-3.708	-2.693
x1	0.7529	0.044	17.296	0.000	0.667	0.839

Omnibus:	3.538	Durbin-Watson:	1.279
Prob(Omnibus):	0.171	Jarque-Bera (JB):	3.589
Skew:	0.357	Prob(JB):	0.166
Kurtosis:	2.744	Cond. No.	43.4

Key Takeaways

Recap:

- Hypotheses → define H0 and H1
- Errors → understand Type I/II trade-off
- Test Statistic → collapse data to one number
- Decision → compare to threshold, draw conclusion
- Regression → fit line, interpret coefficients

Why EDA before Modeling?

Understand your variables

- Are any wildly skewed or heavy-tailed? Non-normal predictors can pull your estimates around.

  Outliers in X or Y can unduly influence an OLS fit.

Check linearity: Scatterplots reveal whether a straight-line makes sense, or whether you might need log/√ transforms.

Spot collinearity

- Highly correlated Xs inflate variances (unstable $\hat{\beta}$). A quick correlation matrix or VIF table flags trouble.

Plan transformations

- If you see skew (e.g. right-tail), consider log transforms on the fly.

Catch data errors early

- Typos or impossible values (e.g. negative ages) often stand out in a histogram or boxplot.

Tips for EDA

Histograms + boxplots side by side give both global shape and outlier detail.

Scatter plots can reveal non-linearity.

Correlation heatmap → VIF: Once you have corr, plug into a small VIF loop to decide if you need to drop or combine variables.

Log/√ transforms: If skew > 1 (in absolute), np.log1p() or np.sqrt() dramatically "straighten" relationships.

Automate EDA: Company usually use customized one-click reports, but when working with new data, we need to be reading the charts by hand first.

Beware of bias!

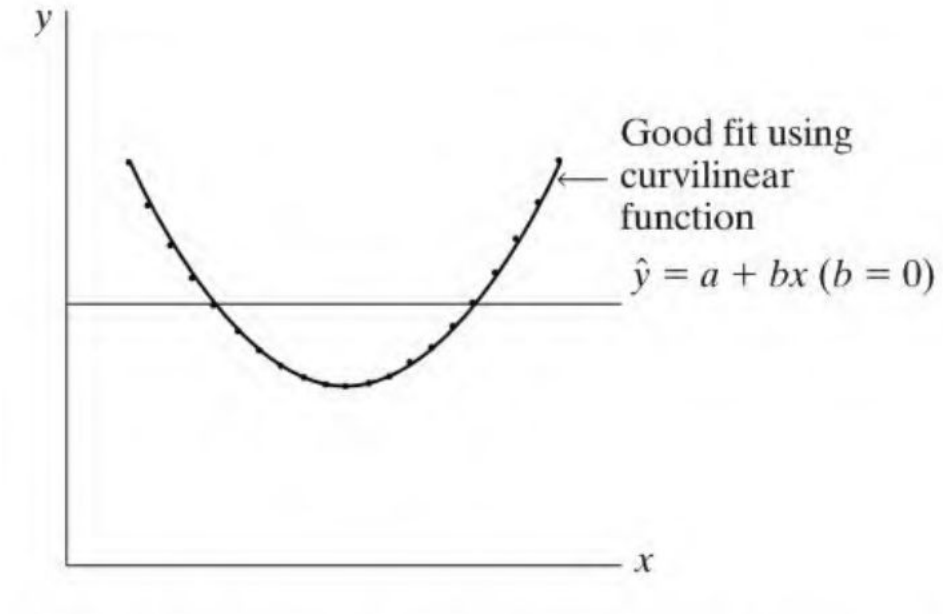
OLS can be dangerous, since numbers can give a false sense of confidence. For β_0 and β_1 to be unbiased and the line to be BLUE (best linear unbiased estimator), several conditions must hold (**Gauss-Markov assumptions**):

- *Linear relationship* between y and x
- *Random sample* of data (to make inferences)
- *No perfect collinearity* in explanatory variables x
- *Exogeneity: zero conditional mean* (expected value of u is 0 for all values of x)
- *Homoscadasticity: u* has the same variance for all values of x

We'll never meet these perfectly, but we can try to test…

A1: Linear relationship

Is the relationship between x and y linear? If it is not, we need to transform the data or use a curvilinear function. More on that in coming weeks…



The relationship depicted on the on the left is NOT linear!

A2: Random sample

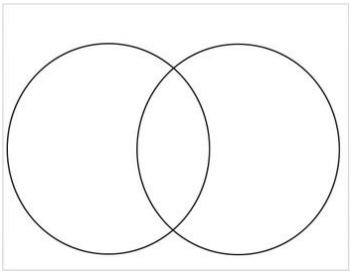
This is obvious. As with other inference tests, inferences based on non-random samples are (more) subject to bias.

A3: No perfect collinearity

Independent variables must not be perfectly correlated with one another.

Example: do you prefer English or French as a primary language?

Eng	1	1	0	0	1
Fr	0	0	1	1	0



A4: Exogeneity (zero conditional mean)

Errors terms should be independent of the IVs. This means they should not be correlated with the IV… i.e., they should be *random* noise.

In practice, there should be no **omitted variable bias** and no **autocorrelation**.

A5: Constant variance for *e*

We assume **homoscadasticity**. This means that the error term is constant across values of *x* (ie, the model uncertainty is identical across observations). When we don't have this, we have the problem of **heteroscadasticity**.