Hypothesis Testing in Linear Regression

Renato Assunção
ESRI and DCC/UFMG

Hypothesis testing

- This is a broad topic and there are many controversies around it.
- There is a general theory of hypothesis testing. We will see it in the second half of the course.
- As in the case of confidence intervals, here we will focus only on the main hypothesis tests associated with the linear regression model.
- We start again with our basic example of the compressive strength of concrete blocks.

Back to the compressive strength of concrete example

- Aim: To predict the compressive strength of concrete based on material composition.
- 1,030 observations with 8 predictors.
- The last two highlighted columns show the extremes of the 95% confidence intervals for each coefficient.
- In these slides, we will connect those confidence intervals with the main hypothesis tests used in linear regression.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.3312	26.586	-0.878	0.380	-75.500	28.837
cement	0.1198	0.008	14.113	0.000	0.103	0.136
slag	0.1039	0.010	10.247	0.000	0.084	0.124
flyash	0.0879	0.013	6.988	0.000	0.063	0.113
water	-0.1499	0.040	-3.731	0.000	-0.229	-0.071
superplasticizer	0.2922	0.093	3.128	0.002	0.109	0.476
coarseaggregate	0.0181	0.009	1.926	0.054	-0.000	0.037
fineaggregate	0.0202	0.011	1.887	0.059	-0.001	0.041
age	0.1142	0.005	21.046	0.000	0.104	0.125

Figure: statsmodels output with the confidence interval columns highlighted

Confidence interval for cement

- The variable cement has a 95% confidence interval equal to (0.103, 0.136).
- When cement changes from cement to cement+1, the compressive strength increases by an amount between 0.103 and 0.136.
- The uncertainty about the value of the coefficient is reflected in the width of the confidence interval.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.3312	26.586	-0.878	0.380	[illegible]	[illegible]
cement	0.1198	0.008	14.113	0.000	0.103	0.136
slag	0.1039	0.010	10.247	0.000	[illegible]	[illegible]
flyash	0.0879	0.013	6.988	0.000	0.063	0.113
water	-0.1499	0.040	-3.731	0.000	-0.229	-0.071
superplasticizer	0.2922	0.093	3.128	0.002	0.109	0.476
coarseaggregate	0.0181	0.009	1.926	0.054	-0.000	0.037
fineaggregate	0.0202	0.011	1.887	0.059	-0.001	0.041
age	0.1142	0.005	21.046	0.000	0.104	0.125

Figure: output highlighting the confidence interval for cement

Confidence interval for water

- The variable water has a negative effect.
- Its confidence interval is $(-0.229, -0.071)$.
- When water changes from water to water+1, the compressive strength decreases by an amount between 0.229 and 0.071.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.3312	26.586	-0.878	0.380	-75.500	28.837
cement	0.1198	0.008	14.113	0.000	0.103	0.136
slag	0.1039	0.010	10.247	0.000	0.084	0.124
flyash	0.0879	0.013	6.988	0.000	[illegible]	[illegible]
water	-0.1499	0.040	-3.731	0.000	-0.229	-0.071
superplasticizer	0.2922	0.093	3.128	0.002	0.109	0.476
coarseaggregate	0.0181	0.009	1.926	0.054	-0.000	0.037
fineaggregate	0.0202	0.011	1.887	0.059	-0.001	0.041
age	0.1142	0.005	21.046	0.000	0.104	0.125

Figure: output highlighting the confidence interval for water

Prediction with uncertainty

- In the first case, although there is uncertainty, we still predict a positive effect for increasing `cement`.
- In the second case, again with uncertainty, we predict a negative effect for increasing `water`.
- A third case is when the confidence interval contains the value zero.

Confidence interval for `fineaggregate`

- The confidence interval for `fineaggregate` is $(-0.001, 0.041)$.
- When `fineaggregate` changes from `fineaggregate` to `fineaggregate+1`, the compressive strength may decrease or increase.
- The uncertainty about the coefficient includes uncertainty even about the direction of the effect.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.3312	26.586	-0.878	0.380	-75.500	28.837
cement	0.1198	0.008	14.113	0.000	0.103	0.136
slag	0.1039	0.010	10.247	0.000	0.084	0.124
flyash	0.0879	0.013	6.988	0.000	0.063	0.113
water	-0.1499	0.040	-3.731	0.000	-0.229	-0.071
superplasticizer	0.2922	0.093	3.128	0.002	0.109	0.476
coarseaggregate	0.0181	0.009	1.926	0.054	-0.000	0.037
fineaggregate	0.0202	0.011	1.887	0.059	-0.001	0.041
age	0.1142	0.005	21.046	0.000	0.104	0.125

Figure: output highlighting the confidence interval for fineaggregate

Confidence intervals containing zero

- Confidence intervals containing the value zero form a special class.
- These are coefficients for which we are not confident about the existence or even the sign of the effect.
- The effect may be positive, negative, or exactly zero.
- What happens if the true coefficient is actually zero?
- In that case, the variable may be removed from the model.

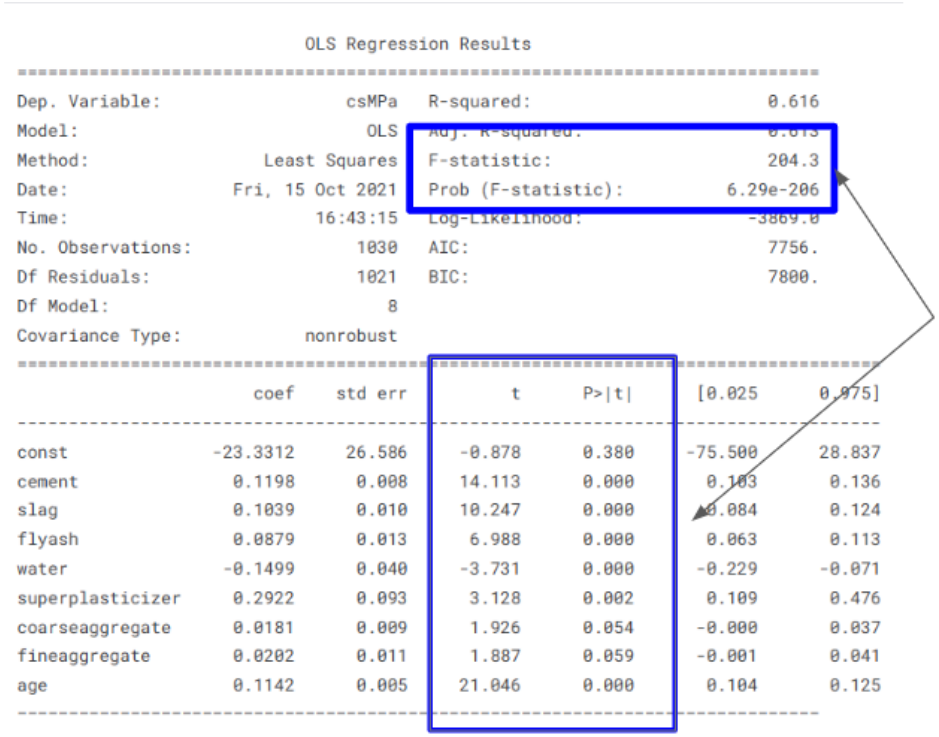
$$(Y \mid \boldsymbol{x}) \sim N(\beta_0 + \beta_1 x_1 + 0 x_2 + \beta_3 x_3, \sigma^2).$$

Testing whether the true coefficient is zero

- A very simple way to discard irrelevant variables in linear regression is to inspect the confidence intervals.
- If the interval contains zero, the variable may be discarded.
- This rule is simple, but it has some disadvantages that we will discuss later when we discuss the multicollinearity problem.
- This simple method is based on a hypothesis test that asks whether the true coefficient is zero.
- In addition to confidence intervals, we also look at the p-values of the tests.

Where are the hypothesis tests in the output?

- The main hypothesis tests associated with linear regression appear directly in the standard regression output.
- We now connect confidence intervals with the theory of hypothesis testing.



OLS Regression Results

Dep. Variable:	csMPa	R-squared:	0.616
Model:	OLS	Adj. R-squared:	0.613
Method:	Least Squares	F-statistic:	204.3
Date:	Fri, 15 Oct 2021	Prob (F-statistic):	6.29e-206
Time:	16:43:15	Log-Likelihood:	-3869.0
No. Observations:	1030	AIC:	7756.
Df Residuals:	1021	BIC:	7800.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.3312	26.586	-0.878	0.380	-75.500	28.837
cement	0.1198	0.008	14.113	0.000	0.103	0.136
slag	0.1039	0.010	10.247	0.000	0.084	0.124
flyash	0.0879	0.013	6.988	0.000	0.063	0.113
water	-0.1499	0.040	-3.731	0.000	-0.229	-0.071
superplasticizer	0.2922	0.093	3.128	0.002	0.109	0.476
coarseaggregate	0.0181	0.009	1.926	0.054	-0.000	0.037
fineaggregate	0.0202	0.011	1.887	0.059	-0.001	0.041
age	0.1142	0.005	21.046	0.000	0.104	0.125

Figure: statsmodels output with t, p-value, F statistic, and confidence interval columns highlighted

The stochastic model again

- Assume that the linear regression model is the data-generating mechanism.
- There exists an unknown vector of coefficients that generates the data and that we want to learn:

$$\boldsymbol{\beta}^* = \left(\beta_0^*, \quad \beta_1^*, \quad \ldots, \quad \beta_p^*\right)^\top.$$

- We know that

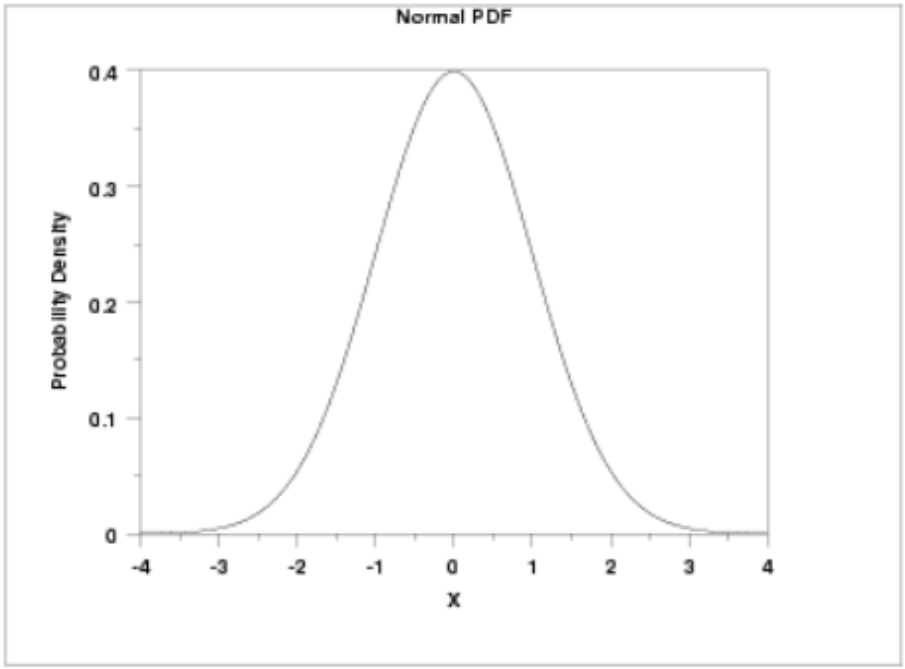
$$\hat{\boldsymbol{\beta}} \sim N_{p+1}\left(\boldsymbol{\beta}^*, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right).$$

- Therefore, each coordinate $\hat{\beta}_j$ of the estimated vector oscillates like a Gaussian around its true and unknown value β_j^*.

$$\hat{\beta}_j \sim N\left(\beta_j^*, \sigma^2[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{jj}\right).$$

Standard normal Gaussian

- We know a lot about the behavior of Gaussian random variables.
- Consider the standard Gaussian $\mathcal{N}(0, 1)$.
  - It rarely falls outside the interval $(-2, 2)$.
  - The probability of being in $(-2, 2)$ is approximately 0.95.
- Also, $X \sim \mathcal{N}(\mu, \sigma^2)$ can be transformed into a standard Gaussian: $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$.



Back to linear regression

- We apply this idea to one coefficient of the least-squares estimator.
- As

$$\hat{\beta}_j \sim N(\beta_j^*, \sigma^2(\mathrm{X}'\mathrm{X})^{-1}[jj])$$

we obtain, by standardization, that

$$Z = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\sigma^2[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{jj}}} \sim \mathcal{N}(0,1).$$

- This the valid whenever we use the **true** and **unknown** value of the β_j^* coefficient in the numerator.
- With confidence intervals, we used the fact that Z lies between -2 and 2 with high probability and then reversed the inequality to obtain a reasonable interval of values for the true unknown coefficient.

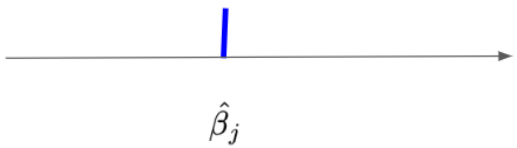


Figure: The estimated coefficient $\hat{\beta}_j$ is a point in the real line.

A fundamental point

- If we subtract the wrong value in the numerator, instead of subtracting the true coefficient β_j^*, then the resulting quantity will **not** follow a standard normal distribution.
- This is the key idea behind hypothesis testing.
- For example, suppose that $\hat{\beta}_j \sim \mathcal{N}(5.5, 4.0)$. That is, $\beta_j^* = 5.5$.
- Then,

$$Z = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{4.0}} = \frac{\hat{\beta}_j - 5.5}{\sqrt{4.0}} \sim \mathcal{N}(0, 1).$$

- What happens if we subtract a value different from the true value $\beta_j^* = 5.5$?
- For example, what happens if we use 0 instead of the true $\beta_j^* = 5.5$?

Example: subtracting the wrong expected value

If we use zero instead of the true value $\beta_j^* = 5.5$, then

$$\frac{\hat{\beta}_j - 0}{\sqrt{4.0}} = \frac{\hat{\beta}_j \pm \beta_j^* - 0}{\sqrt{4.0}} == \frac{\hat{\beta}_j \pm 5.5 - 0}{\sqrt{4.0}}$$

$$= \underbrace{\frac{\hat{\beta}_j - 5.5}{\sqrt{4.0}}}_{\mathcal{N}(0,1)} + \underbrace{\frac{5.5 - 0}{\sqrt{4.0}}}_{\text{a constant}}$$

$$= \mathcal{N}(0,1) + \frac{\beta_j^* - 0}{\sqrt{4.0}} = \mathcal{N}(\frac{\beta_j^*}{2}, 1)$$

Therefore, instead of being centered at zero, the distribution is shifted by the constant $\beta_j^*/2$:

$$\frac{\hat{\beta}_j - 0}{\sqrt{4.0}} \sim \mathcal{N}\left(\frac{\beta_j^*}{2}, 1\right).$$

Summary of the idea

Suppose that

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_j^*, \sigma^2[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{jj}\right) = \mathcal{N}\left(\beta_j^*, v^2\right)$$

When we use the true value β_j^* in the numerator, we obtain

$$\frac{\hat{\beta}_j - \beta_j^*}{v} \sim \mathcal{N}(0,1),$$

If we use zero instead, then

$$\frac{\hat{\beta}_j - 0}{v} \sim \mathcal{N}\left(\frac{\beta_j^*}{v}, 1\right).$$

This last quantity is standard normal if and only if the true coefficient is zero:

$$\beta_j^* = 0.$$

Summary, part 2

- If we use zero in the numerator **and** the hypothesis

$$H_0 : \beta_j^* = 0 \quad \text{is true,}$$

then the standardized quantity

$$\frac{\hat{\beta}_j - 0}{v} \sim \mathcal{N}(0, 1)$$

- Also, it falls between -2 and 2 with high probability (≈ 0.95).
- However, if the hypothesis is false, the Gaussian $(\hat{\beta}_j - 0)/v$ will be centered at a value different from zero
- and it may very easily fall outside the interval $(-2, 2)$.
- How do we use this result in practice?

Testing the null hypothesis $H_0 : \beta_j^* = 0$

- There are only two possibilities:
  - either $H_0 : \beta_j^* = 0$ is true
  - or $H_a : \beta_j^* \neq 0$ is true.
- If we believe $H_0 : \beta_j^* = 0$ is true, we may drop the corresponding feature from the linear regression.
- If we believe that $H_a : \beta_j^* \neq 0$ is true, we should keep the corresponding feature in the linear regression model.
- What we should believe?
- We look at the data to find evidence for one of these hypothesis.

Testing the null hypothesis $H_0 : \beta_j^* = 0$

- Let $t = (\hat{\beta}_j - 0)/v$.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.3312	26.586	-0.878	0.380	-75.500	28.837
cement	0.1198	0.008	14.113	0.000	0.103	0.136
slag	0.1039	0.010	10.247	0.000	0.084	0.124
flyash	0.0879	0.013	6.988	0.000	0.063	0.113
water	-0.1499	0.040	-3.731	0.000	-0.229	-0.071
superplasticizer	0.2922	0.093	3.128	0.002	0.109	0.476
coarseaggregate	0.0181	0.009	1.926	0.054	-0.000	0.037
fineaggregate	0.0202	0.011	1.887	0.059	-0.001	0.041
age	0.1142	0.005	21.046	0.000	0.104	0.125

Figure: Values of $t = (\hat{\beta}_j - 0)/v$ for each variable.

- Note that v has a different value for each variable but we avoid overburden the notation by not writing v_j.

Testing the null hypothesis $H_0 : \beta_j^* = 0$

- Let $t = (\hat{\beta}_j - 0)/v$.
  - Check if $t \in (-2, 2)$. If so, this event is compatible with $H_0 : \beta_j^* = 0$ and we may drop the corresponding feature from the model.
  - However, if $t \notin (-2, 2)$ and we insist on believing that $H_0 : \beta_j^* = 0$ is true, we are holding a belief that a rare event happened.
  - Rare events may happen, of course.
  - However, we have a more simple explanation for the event that $t \notin (-2, 2)$: it seems more plausible that simply $H_a : \beta_j^* \neq 0$ is true, because then the event $t \notin (-2, 2)$ has a larger probability and we are not seeing a rare event.
- In summary, if the observed t falls outside this interval, that is evidence that the true and unknown coefficient is different from zero, and the variable should **not** be removed. Otherwise, drop the variable.

P-value

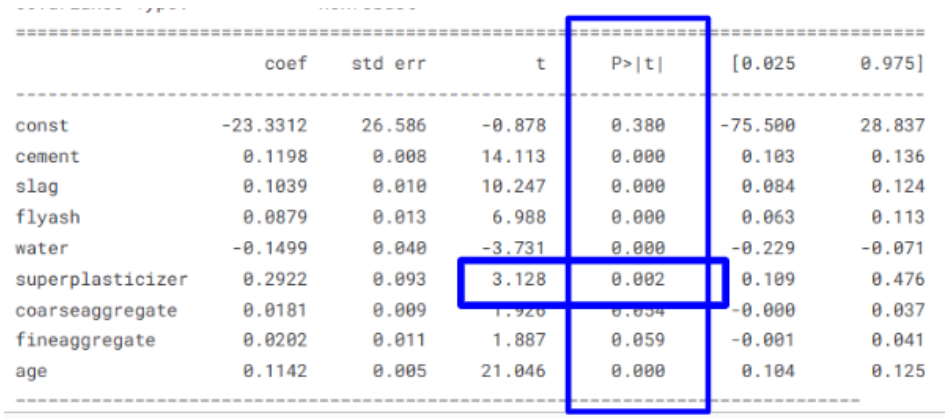
- A p-value is a probability.
- It assumes that the null hypothesis $H_0 : \beta_j^* = 0$ is true.
- Under that assumption, $t = \frac{\hat{\beta}_j - 0}{v} \sim \mathcal{N}(0, 1)$.
- We observe $t = 3.128$ for `superplasticizer`, which is outside the interval $(-2, 2)$ and therefore, it does not seem that t is following the $\mathcal{N}(0, 1)$ distribution.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.3312	26.586	-0.878	0.380	-75.500	28.837
cement	0.1198	0.008	14.113	0.000	0.103	0.136
slag	0.1039	0.010	10.247	0.000	0.084	0.124
flyash	0.0879	0.013	6.988	0.000	0.063	0.113
water	-0.1499	0.040	-3.731	0.000	-0.229	-0.071
superplasticizer	0.2922	0.093	3.128	0.002	0.109	0.476
coarseaggregate	0.0181	0.009	1.926	0.054	-0.000	0.037
fineaggregate	0.0202	0.011	1.887	0.059	-0.001	0.041
age	0.1142	0.005	21.046	0.000	0.104	0.125

Figure: Output highlighting the p-value column

What does the p-value measure?

- If $H_0 : \beta_j^* = 0$ is true, then t should be in $(-2, 2)$ with high probability.
- For `superplasticizer`, $t = 3.128$, outside the interval $(-2, 2)$.
- The p-value measures how extreme this observed value of t is **assuming that the null hypothesis is true**.
- For `superplasticizer`, if $H_0 : \beta_j^* = 0$ is true, the event that t is extreme as 3.128 is only 0.002.



	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.3312	26.586	-0.878	0.380	-75.500	28.837
cement	0.1198	0.008	14.113	0.000	0.103	0.136
slag	0.1039	0.010	10.247	0.000	0.084	0.124
flyash	0.0879	0.013	6.988	0.000	0.063	0.113
water	-0.1499	0.040	-3.731	0.000	-0.229	-0.071
superplasticizer	0.2922	0.093	3.128	0.002	0.109	0.476
coarseaggregate	0.0181	0.009	1.926	0.054	-0.000	0.037
fineaggregate	0.0202	0.011	1.887	0.059	-0.001	0.041
age	0.1142	0.005	21.046	0.000	0.104	0.125

Figure: Output highlighting the t and p-value for superplasticizer

Observed test statistic

We observed

$$|t_{obs}| = 3.128.$$

The p-value is the probability, under the null hypothesis, of obtaining a value at least this extreme.

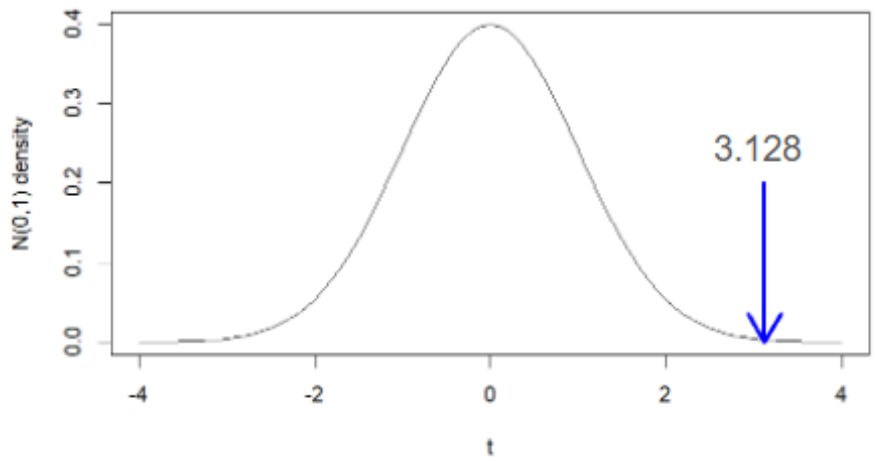


Figure: normal density with the observed t marked in the tail

That is, the p-=value is the area under the standard normal density beyond the observed t value.

Two-sided p-value

$$p\text{-value} = P\big(|t| > |t_{obs}| \mid H_0 \text{ is true}\big).$$

In this example,

$$P\big(|t| > 3.128 \mid H_0 \text{ is true}\big) = 0.002.$$

This is a very small probability, so the observed value is not compatible with the null hypothesis.

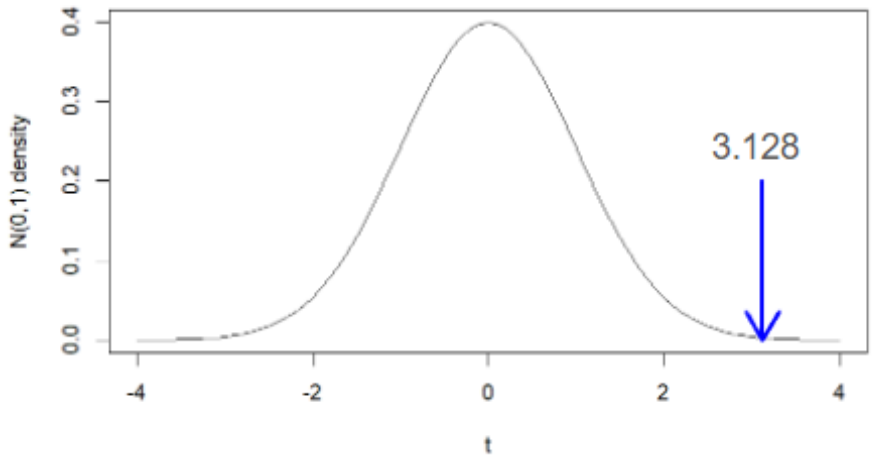


Figure: normal density with the observed t marked in the tail

Using p-values in tests

- Very small p-value (smaller than 0.05): the hypothesis $\beta_j^* = 0$ is not compatible with the data.

$$\Rightarrow \text{Reject the null hypothesis } H_0 : \beta_j^* = 0.$$

- Large p-value (greater than 0.05): the null hypothesis is compatible with the data.

$$\Rightarrow \text{Do not reject the null hypothesis.}$$

A practical detail

- The denominator of the t ratio must be estimated from the data, because σ^2 is unknown.
- Therefore, the exact calculation of the p-value uses the Student-t distribution with

$$n - (p + 1)$$

  degrees of freedom, rather than the standard normal distribution.
- However, if we have a large amount of data compared to the number of features (i.e., if $n - (p + 1) > 30$), using the Student-t is practically identical to using the standard normal distribution.

Equivalences between confidence intervals and tests

- There are direct equivalences between confidence intervals and hypothesis tests.
- p-value < 0.05 if and only if 0 does not belong to the 95% confidence interval.
- If 0 does not belong to the 95% confidence interval, we reject the null hypothesis.

One more test

We can also test whether **all** features may be set to zero at the same time.

$$H_0 : \beta_1^* = \beta_2^* = \cdots = \beta_p^* = 0.$$

Equivalently,

$$H_0 : \begin{pmatrix} \beta_0^* \\ \beta_1^* \\ \vdots \\ \beta_p^* \end{pmatrix} = \begin{pmatrix} \beta_0^* \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

OLS Regression Results

Dep. Variable:	csMPa	R-squared:	0.616
Model:	OLS	Adj. R-squared:	0.613
Method:	Least Squares	F-statistic:	204.3
Date:	Fri, 15 Oct 2021	Prob (F-statistic):	6.29e-206
Time:	16:43:15	Log-Likelihood:	-3869.0
No. Observations:	1030	AIC:	7756.
Df Residuals:	1021	BIC:	7800.
Df Model:	8		
Covariance Type:	nonrobust		

Figure: Output highlighting the F statistic and its p-value

How can we test this?

- If all feature coefficients are zero, the residuals with or without the features should be similar.
- Regression without the features leaves only the column of ones.
- This means projecting $\boldsymbol{Y}$ onto the space of linear combinations of the vector $(1, 1, \ldots, 1)^T$.
- The fitted value under the null model is simply

$$\bar{y}\,(1, 1, \ldots, 1)^T.$$

Two residual sums of squares

If the null hypothesis is true,

$$H_0 : \beta_1^* = \cdots = \beta_p^* = 0,$$

then

$$y_i - \hat{y}_i \approx y_i - \bar{y}.$$

Why? Because in that case

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \approx \hat{\beta}_0.$$

So we compare the two residual sums of squares:

$$\sum_i (y_i - \hat{y}_i)^2 \qquad \text{and} \qquad \sum_i (y_i - \bar{y})^2.$$

The F statistic

- We must also consider the corresponding degrees of freedom.
- The ratio of two **independent** chi-squared distributions, each divided by its degrees of freedom, has a known distribution: the F distribution.
- In this context, we use the statistic

$$F = \frac{\sum_i (y_i - \bar{y})^2/(n-1)}{\sum_i (y_i - \hat{y}_i)^2/(n-(p+1))}.$$

- The letter F comes from (Ronald) Fisher.

A more philosophical flight: Karl Popper

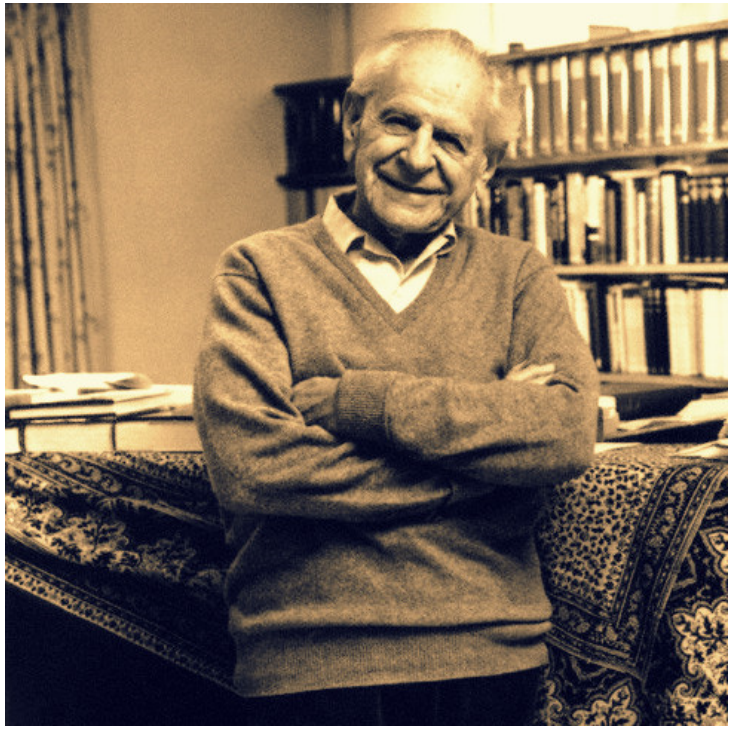


Figure: Sir Karl Popper

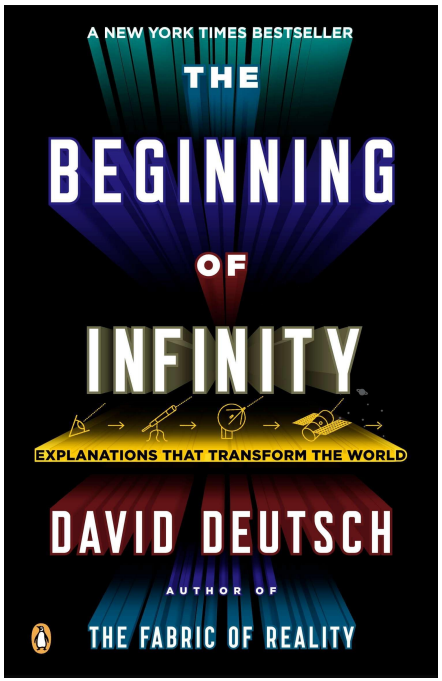
- 1902–1994.
- Vienna $\rightarrow$ New Zealand $\rightarrow$ England (from 1946 onward).
- Philosopher of science.
- What makes a theory scientific?
- Are psychoanalysis and Marxism scientific? Is Einstein's theory scientific?

Books in Portuguese

Figure: Three Portuguese editions of Popper's books

A more modern spokesperson: David Deutsch

A contemporary defender of the Popperian view is David Deutsch.



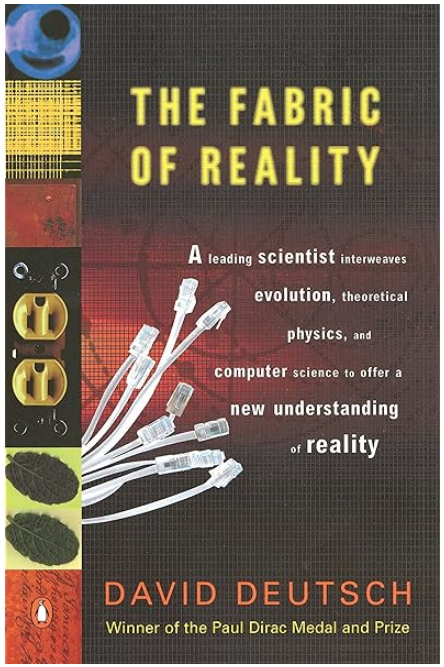


Figure: Books by David Deutsch and TED talk screenshot at `https://www.ted.com/talks/david_deutsch_a_new_way_to_explain_explanation`

Context

- Theories that seemed perfect are later shown to be wrong.
- Newtonian theory lasted for centuries and is still used today.
- But it was a shock to discover, around 1920 through Einstein's work, that it was not the perfect explanation of the physical world.
- Psychoanalysis was a great success in Vienna around 1920–1930. Was it science?
- Is Marxism a scientific theory of history?
- How do we know whether a theory is scientific?
- How do we know whether a scientific theory is correct?

The principle of falsifiability

- One way to demarcate science from non-science.
- A theory is scientific if it can, in principle, be shown to be false.

A historical test

- In 1919, Sir Arthur Stanley Eddington and Frank Watson Dyson led two expeditions to observe a total solar eclipse: one in Africa and one in Sobral, Ceará, Brazil.
- The goal was to measure how much starlight bends as it passes close to the Sun.

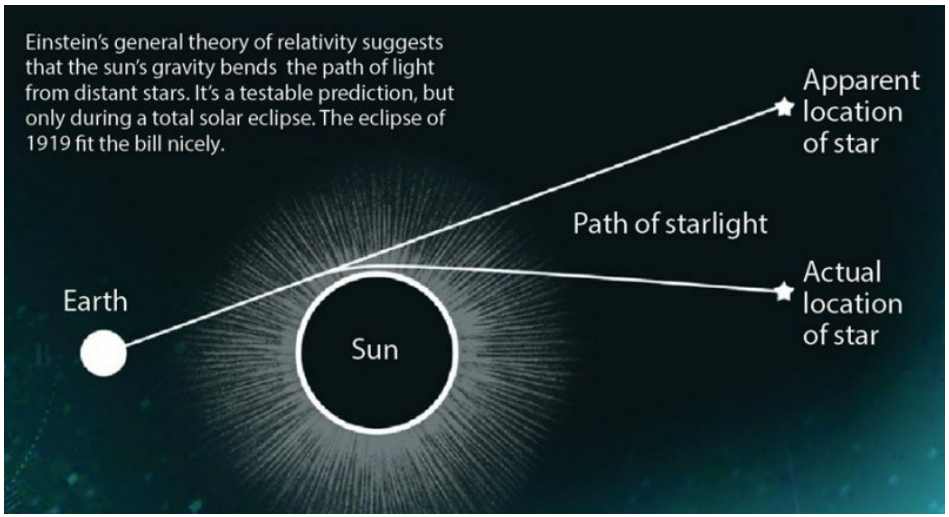


Figure: Sun bending the light and showing an apparent star position.

A historical test

- The results confirmed Einstein's theory of general relativity: light is bent by gravity.
- This made Einstein a worldwide celebrity.
- The key point is that, if the results had not confirmed the prediction, the theory would have been shown to be false.

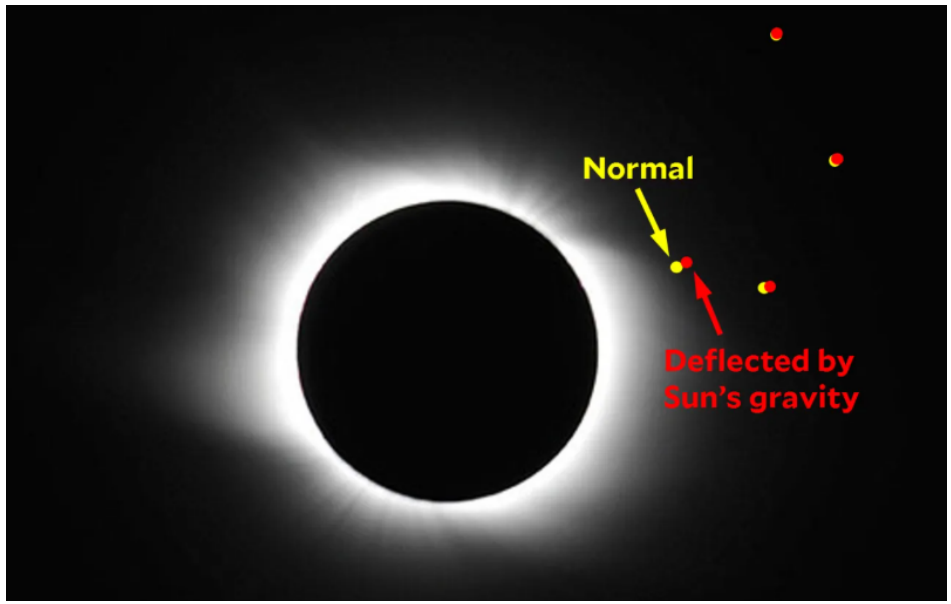


Figure: Eclipse photograph with real and apparent star positions

The principle of falsifiability revisited

- A theory is scientific if it can be tested and potentially refuted.
- If it is refuted, the theory is wrong: it is false.
- If it is not refuted, this does **not** mean it has been proved true.

Conjectures and refutations

- Scientific knowledge is not what we know to be true.
- Rather, it is the collection of theories that we have not yet managed to refute.
- Theories that cannot be tested in this way are not necessarily absurd, but they are not scientific.
- And where do scientific ideas come from?
- There is no scientific method for generating ideas and theories in the first place.