L3: Hypotheses Testing with Linear Regression

Aneta Havlínová, Michal Hakala

25th of April, 2023

TODAY'S LECTURE

1. Hypotheses testing intuition
2. Ordinary Least Squares (OLS) model
3. OLS model validation
4. Multiple OLS - practical example
5. OLS model assumptions

Anet

Michal

Hypotheses

1. Investment in sustainability has a positive impact on company revenues.

1. Healthy diet decreases the time needed for a recovery from a viral disease.

1. The higher the unemployment of a city, the higher the crime rate.

1. Women are less likely to get a promotion.

Hypotheses Terminology

EXAMPLE: Unemployment rate has a significant impact on the crime rate.

Null hypothesis: There is no relationship between the unemployment rate and the crime rate.

Alternative hypothesis:. A higher unemployment rate leads to a higher crime rate.

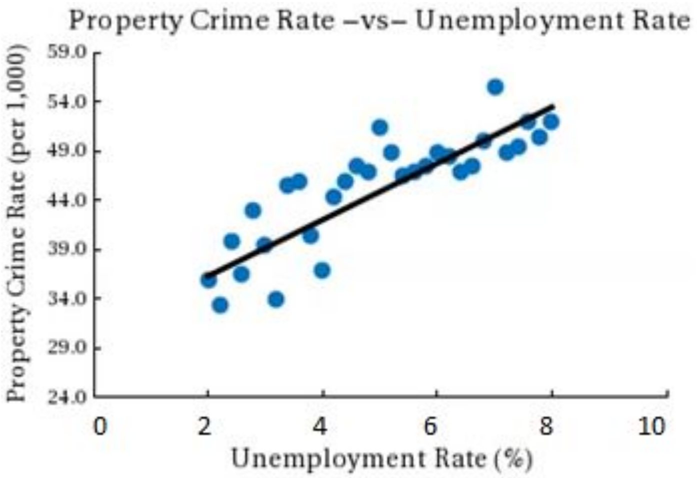
Goal: to **(not) reject the null hypothesis**

We never say that we accept a hypothesis!

czechitas

Hypotheses Testing Intuition

Null hypothesis: There is <u>no relationship</u> between unemployment rate and crime rate.



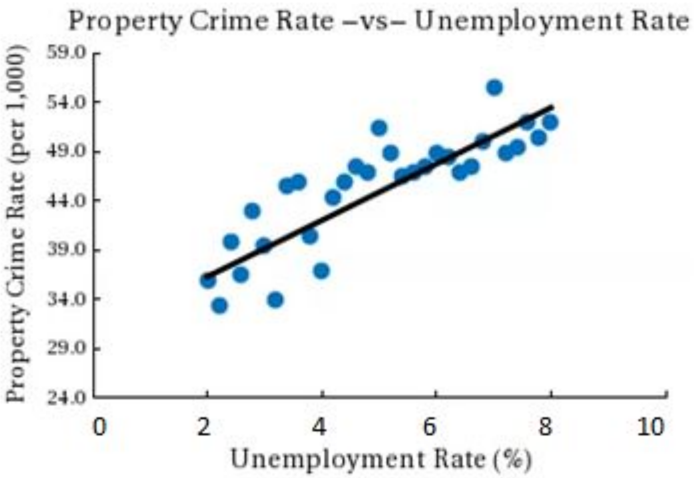
1) We want to **generalize the relationship** between unemployment and crime we fit a straight line through data.

This way, we make an assumption that the relationship is linear.

czechitas

Hypotheses Testing Intuition

Null hypothesis: There is <u>no relationship</u> between unemployment rate and crime rate.



1) We want to **generalize the relationship** between unemployment and crime □ we fit a straight line through data

2) We want to **test if the relationship is statistically significant**

Linear Regression – Ordinary Least Squares

$$crime\ rate = \beta_0 + \beta_1\, unemployment\ rate + \varepsilon$$

Null hypothesis H_0: There is no relationship between unemployment rate and crime rate.

$$H_0: \beta_1 = 0$$

→ This is a convention.
→ We expect we will reject the null hypothesis.

czechitas

Linear Regression – Ordinary Least Squares

$$crime\ rate = \beta_0 + \beta_1\, unemployment\ rate + \varepsilon$$

Null hypothesis H_0: There is <u>no relationship</u> between unemployment rate and crime rate.

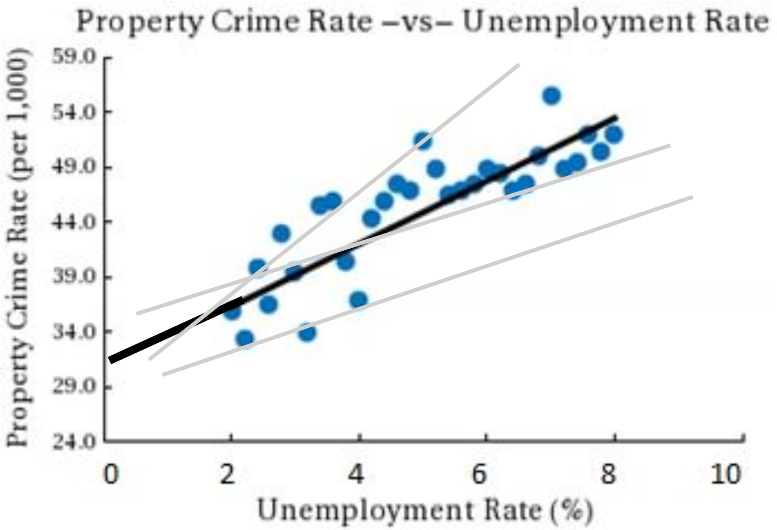
$$H_0: \beta_1 = 0$$

We reject the null hypothesis, if our calculated β_1 *is "far enough from zero".*

OLS MODEL:
CALCULATING BETA COEFFICIENT

Linear Regression – Fitted Line

$$crime\ rate = \beta_0 + \beta_1\, unemployment\ rate + \varepsilon$$



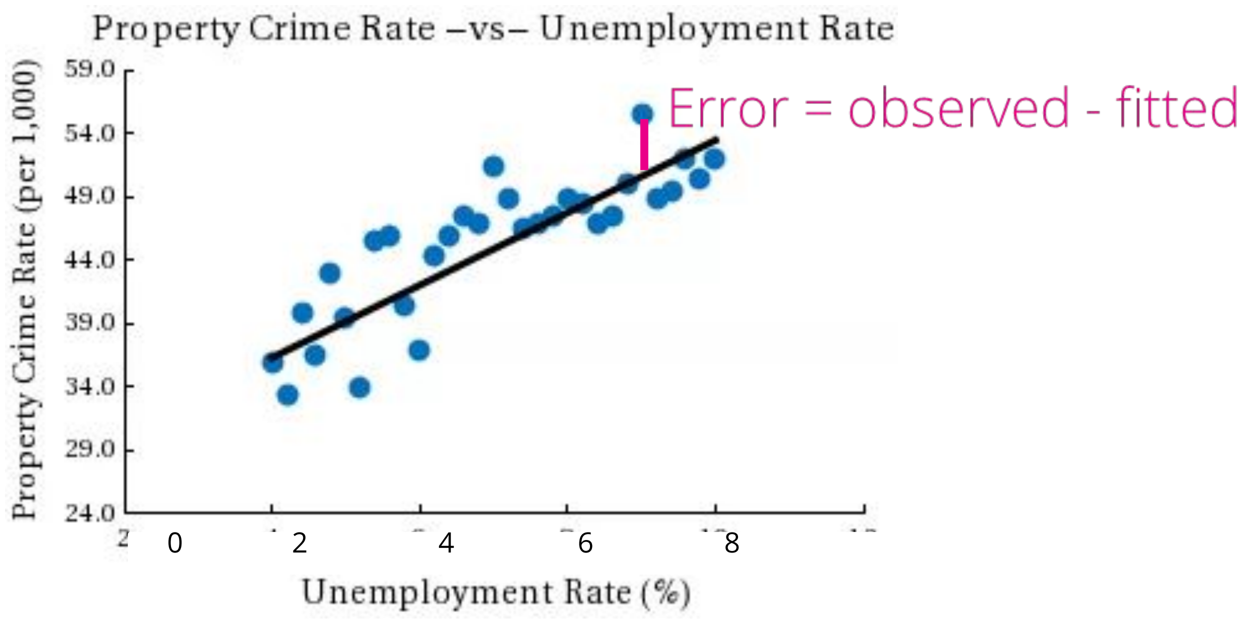
β_0 – intersection with y-axis

β_1 – slope of the fitted line

How do we know which line should be fitted?

Linear Regression – Fitted Line

$$crime\ rate = \beta_0 + \beta_1\, unemployment\ rate + \boxed{\varepsilon}$$



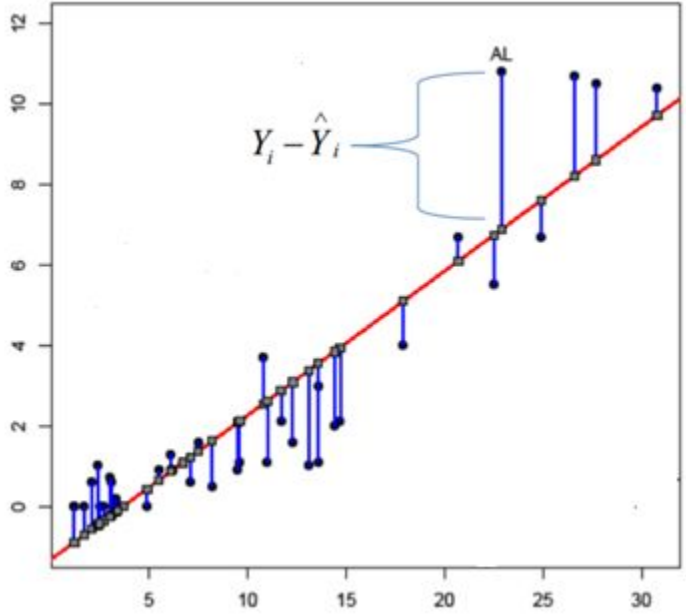
Linear Regression – Fitted Line

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\boldsymbol{\varepsilon = y - (\beta_0 + \beta_1 x)}$$

czechitas

Minimization of Sum of Squared Residuals



$$\min \sum \varepsilon_i^2$$

$$\min \sum (y_i - (b_0 + b_1 x_i))^2$$

Minimization of Sum of Squared Residuals

$$\min \sum (y_i - (b_0 + b_1 x_i))^2$$

Take derivative with respect to b_1 and set it equal to 0.

$$\Rightarrow \quad b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Minimization of Sum of Squared Residuals

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Unemployment rate (%)	Crime rate per 1000 habitants
4	38
4.2	39
6.1	45
7.3	48
5.2	42
3.9	36
5.5	44
…	…

AVG = 4.9 AVG = 39

czechitas

Minimization of Sum of Squared Residuals

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{(4 - \mathbf{4.9})(38 - \mathbf{39}) + (4.2 - \mathbf{4.9})(39 - \mathbf{39}) + \ldots}{(4 - \mathbf{4.9})^2 + (4.2 - \mathbf{4.9})^2 + \ldots}$$

Unemployment rate (%)	Crime rate per 1000 habitants
4	38
4.2	39
6.1	45
7.3	48
5.2	42
3.9	36
5.5	44
…	…

AVG = 4.9

AVG = 39

Minimization of Sum of Squared Residuals

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= 2.86$$

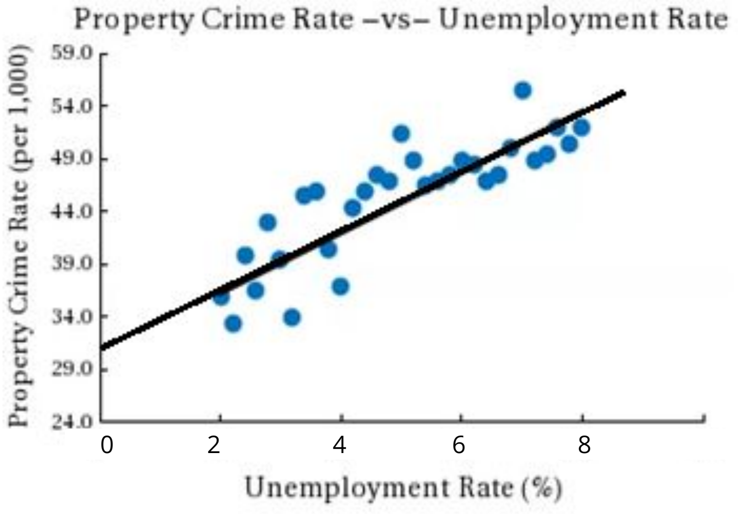
Unemployment rate (%)	Crime rate per 1000 habitants
4	38
4.2	39
6.1	45
7.3	48
5.2	42
3.9	36
5.5	44
…	…

$$crime\,rate = \beta_0 + 2.86\ unemployment\,rate + \varepsilon$$

czechitas

Fitted Model

$$crime\ rate = \mathbf{29.4} + \mathbf{2.86}\ unemployment\ rate + \varepsilon$$

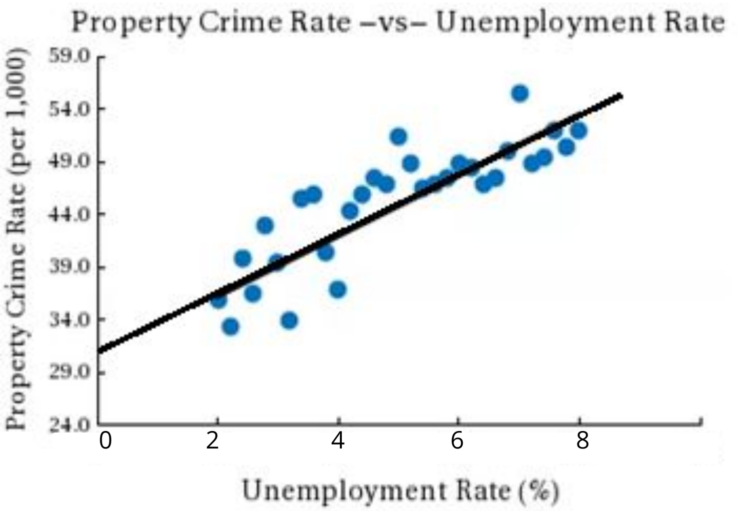


Unemployment rate (%)	Crime rate per 1000 habitants
4	38
4.2	39
6.1	45
7.3	48
5.2	42
3.9	36
5.5	44
…	…

czechitas

Interpretation

$$crime\ rate = \mathbf{29.4} + \mathbf{2.86}\ unemployment\ rate + \varepsilon$$

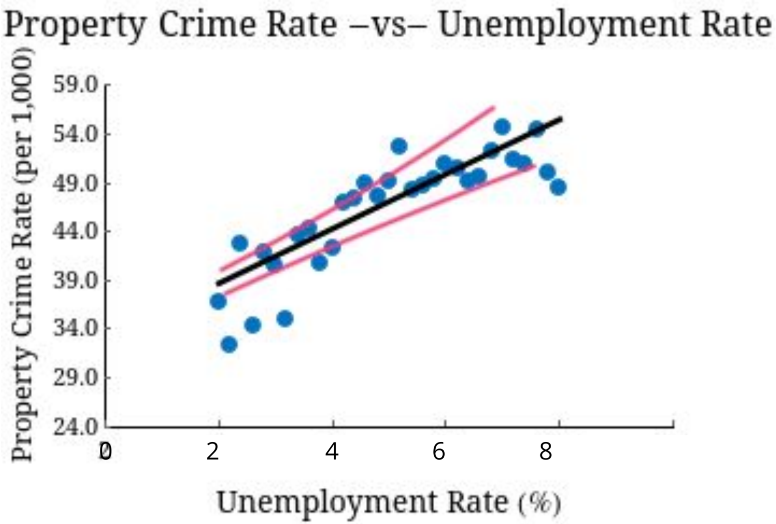


One percentage point increase in the unemployment rate

→ Increase of 2.86 in property crime rate on average.

Standard errors and confidence interval of beta estimate

$$crime\ rate = \mathbf{29.4} + \mathbf{2.86}\ unemployment\ rate + \varepsilon$$



Standard errors: represent the average distance that the observed values have from the regression line

→ base for confidence interval.

Standard Errors of Beta Coefficient

$$s_{\hat{\beta}_1} = \sqrt{\frac{\sum_i \hat{\epsilon}_i^2}{(n-2)\sum_i (x_i - \bar{x})^2}}$$

$$crime\ rate = \beta_0 + \beta_1\, unemployment\ rate + \boxed{\varepsilon}$$

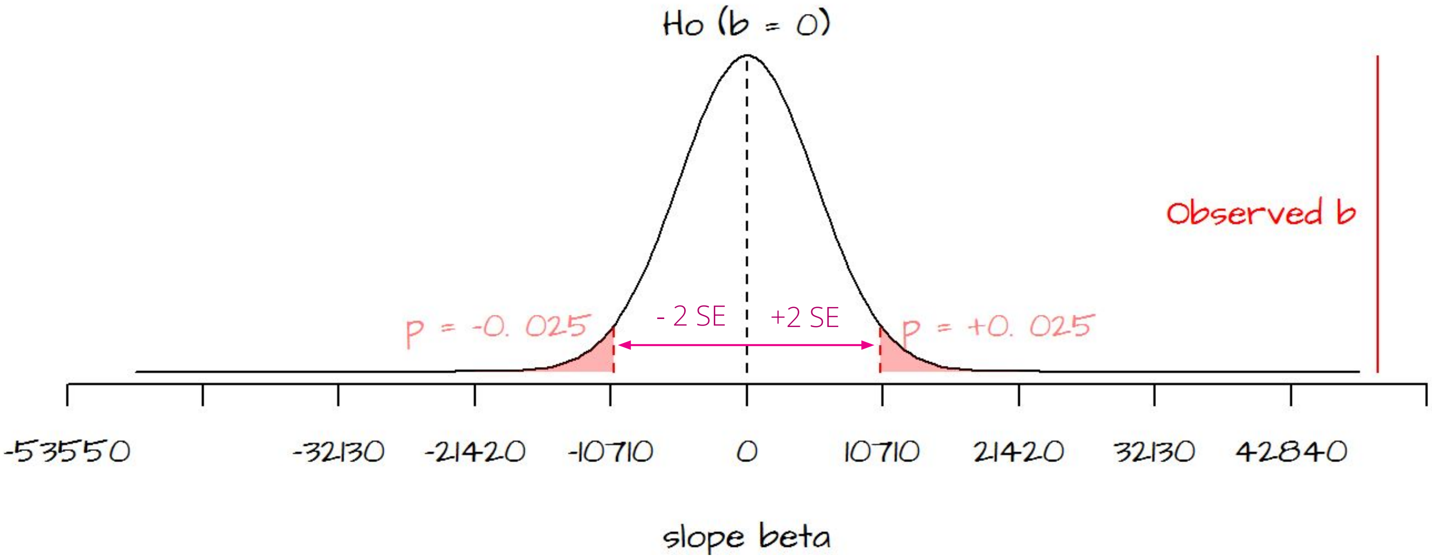
residuals

n
= sample size
= number of rows in the data

Unemployment rate (%)	Crime rate per 1000 habitants
4	38
4.2	39
6.1	45
7.3	48
…	…

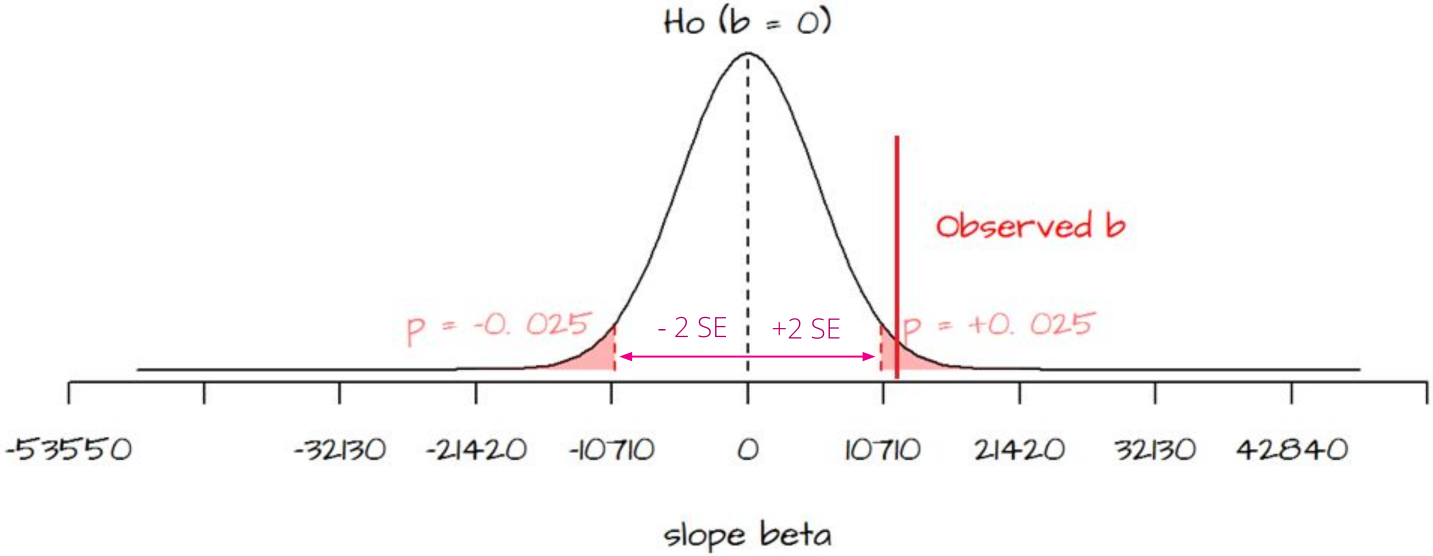
AVG X = 4.9

Hypotheses Testing with OLS: Is Beta “far enough from zero”?



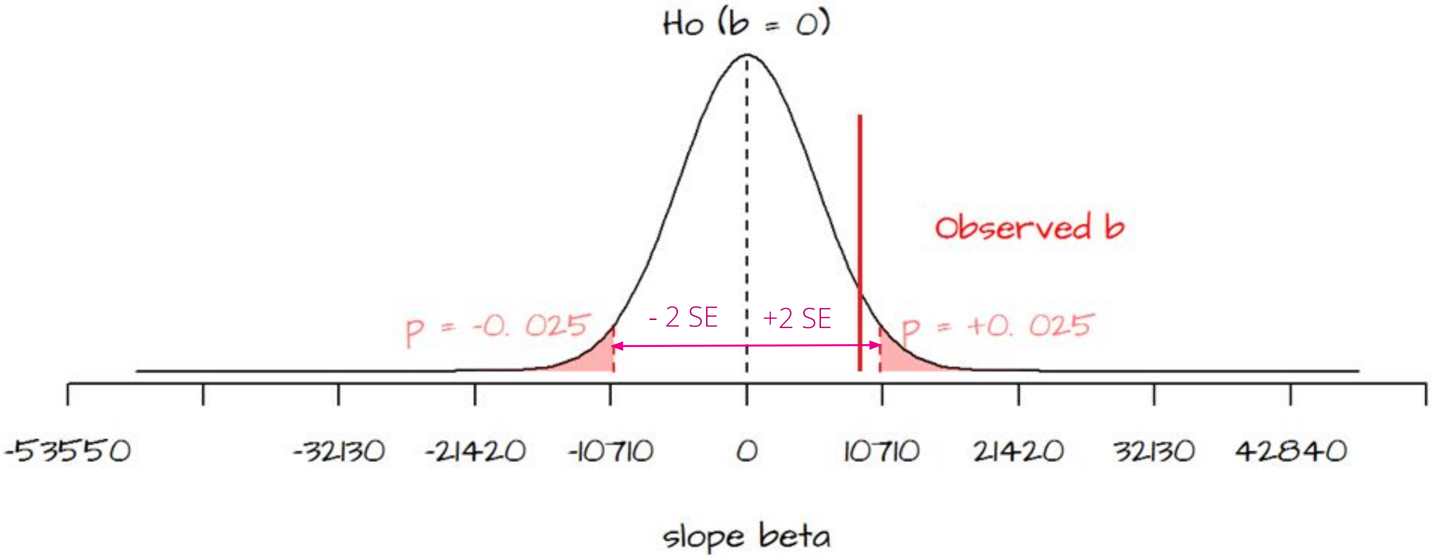
Here, we reject H_0 with 99,99% confidence

Hypotheses Testing with OLS: Is Beta “far enough from zero”?



Here, we reject H_0 with 95% confidence

Hypotheses Testing with OLS: Is Beta “far enough from zero”?



And so on..

czechitas

Hypotheses Testing with OLS: Is Beta “far enough from zero”?

Statistical significance: Is the p-value small enough?

If **p-value = 0.10** → we have **90% confidence**, that our variable is significant

If **p-value = 0.05** → we have **95% confidence**, that our variable is significant

If **p-value = 0.01** → we have **99% confidence**, that our variable is significant

czechitas

OLS MODEL: VALIDATION

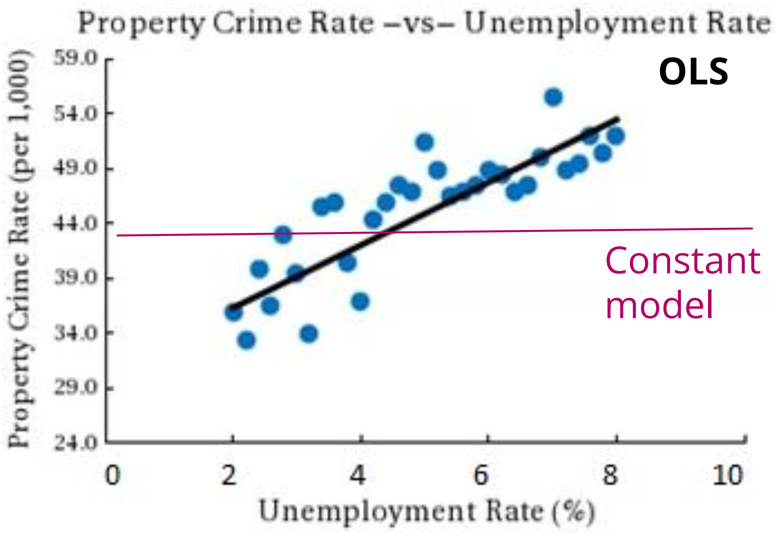
OLS Model Validation

F-test:

H0: Model with no independent variables fits the data as well as your model

→ we want to reject H0

$$crime\ rate = \beta_0 + \beta_1\, unemployment\ rate + \varepsilon$$



OLS Model Validation

R-squared:

- Indicates the percentage of the variance in the dependent variable that the independent variables explain collectively

- 0-100% scale (the higher the better)

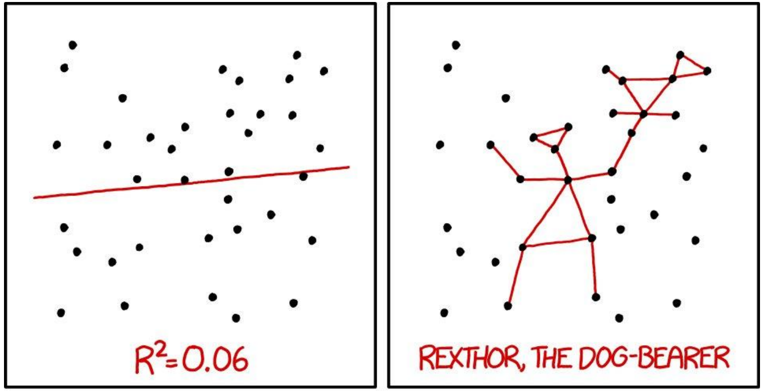
$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

OLS Model Validation

R-squared:

- Indicates the percentage of the variance in the dependent variable that the independent variables explain collectively
- 0-100% scale (the higher the better)



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

czechitas

OLS MODEL:
MULTIPLE LINEAR REGRESSION

Multiple Linear Regression (OLS)

$$y= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Multiple Linear Regression (OLS)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Example:

Impact of various marketing investments on product sales

$$Sales = \beta_0 + \beta_1 Youtube\ Ads + \beta_2\ Facebook\ Ads + \beta_3\ Newspaper\ Ads + \varepsilon$$

Hypotheses Testing with OLS

Assumption: Investment in Facebook advertising has a positive impact on sales.

$$Sales = \beta_0 + \beta_1 Youtube\ Ads + \boxed{\beta_2}\ Facebook\ Ads + \beta_3\ Newspaper\ Ads + \varepsilon$$

??

Hypotheses Testing with OLS

Null hypothesis: Investment in Facebook advertising has NO impact on sales.

$$Sales = \beta_0 + \beta_1 Youtube\ Ads + \boxed{\beta_2} Facebook\ Ads + \beta_3\ Newspaper\ Ads + \varepsilon$$

??

→ Economic significance: is β_2 large enough for our business?

→ Statistical significance: Is p-value small enough?

OLS MODEL:

MULTIPLE LINEAR REGRESSION WITH EXAMPLE IN PYTHON

Hypotheses Testing with OLS

EXAMPLE: Impact of marketing investments (youtube, facebook, newspaper) on sales

Data in thousands USD

youtube	facebook	newspaper	sales
276.12	45.36	83.04	26.52
53.40	47.16	54.12	12.48
20.64	55.08	83.16	11.16
181.80	49.56	70.20	22.20
216.96	12.96	70.08	15.48
10.44	58.68	90.00	8.64

Hypotheses Testing with OLS in Python

EXAMPLE: Impact of marketing investments (youtube, facebook, newspaper) on sales

```
# Multiple linear model
model_full = smf.ols(formula='sales ~ facebook + newspaper + youtube',
                     data=marketing).fit()

print(model_full.summary())
```

Hypotheses Testing with OLS – Results in Python

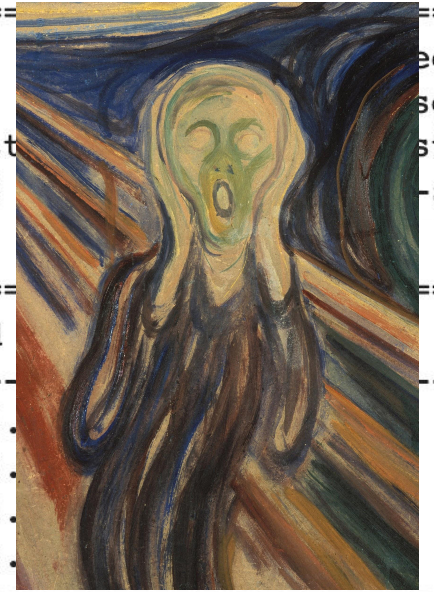
EXAMPLE: Impact of marketing investments (youtube, facebook, newspaper) on sales

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     570.3
Date:                Sun, 23 Apr 2023   Prob (F-statistic):           1.58e-96

==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3526.6672    374.290      9.422      0.000    2788.515    4264.820
facebook       0.1885      0.009     21.893      0.000       0.172       0.206
newspaper     -0.0010      0.006     -0.177      0.860      -0.013       0.011
youtube        0.0458      0.001     32.809      0.000       0.043       0.049
==============================================================================
```

Hypotheses Testing with OLS – Results in Python

EXAMPLE: Impact of marketing investments (youtube, facebook, newspaper) on sales

```
                                 OLS Regression Results
=========================                    ============================
Dep. Variable:                               ed:                    0.897
Model:                                       squared:               0.896
Method:                      Least           stic:                  570.3
Date:                     Sun, 23            -statistic):        1.58e-96


=========================                    ============================
                  coef       std             P>|t|      [0.025     0.975]
-------------------------                    ----------------------------
Intercept    3526.6672       374.            0.000    2788.515   4264.820
facebook        0.1885         0.            0.000       0.172      0.206
newspaper      -0.0010         0.            0.860      -0.013      0.011
youtube         0.0458         0.            0.000       0.043      0.049
=========================                    ============================
```



OLS Results Interpretation

EXAMPLE: Impact of marketing investments (youtube, facebook, newspaper) on sales

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     570.3
Date:                Sun, 23 Apr 2023   Prob (F-statistic):           1.58e-96
```

F-stats is statistically significant (p-value < 0.05), so the model makes sense overall.

czechitas

OLS Results Interpretation

EXAMPLE: Impact of marketing investments (youtube, facebook, newspaper) on sales

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     570.3
Date:                Sun, 23 Apr 2023   Prob (F-statistic):           1.58e-96
```

R^2 is high (we want it as close to 1 as possible), so our variables explain sales well.

OLS Results Interpretation

EXAMPLE: Impact of marketing investments (youtube, facebook, newspaper) on sales

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3526.6672	374.290	9.422	0.000	2788.515	4264.820
facebook	0.1885	0.009	21.893	0.000	0.172	0.206
newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011
youtube	0.0458	0.001	32.809	0.000	0.043	0.049

Youtube and Facebook investments are statistically significant because their p-values are nearly zero.

Newspaper investment is not significant.

OLS Results Interpretation

$$Sales = \beta_0 + \beta_1 Youtube\ Ads + \beta_2\ Facebook\ Ads + \beta_3\ Newspaper\ Ads + \varepsilon$$

Run model in Python

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3526.6672	374.290	9.422	0.000	2788.515	4264.820
facebook	0.1885	0.009	21.893	0.000	0.172	0.206
newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011
youtube	0.0458	0.001	32.809	0.000	0.043	0.049

Resulting equation

$$Sales = \mathbf{3.527} + \mathbf{0.046} * Youtube\ Ads + \mathbf{0.189} * Facebook\ Ads - \mathbf{0.001} * Newspaper\ Ads$$

czechitas

OLS Results Interpretation

EXAMPLE: Impact of marketing investments (youtube, facebook, newspaper) on sales

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3526.6672	374.290	9.422	0.000	2788.515	4264.820
facebook	$\boxed{0.1885}$	0.009	21.893	0.000	0.172	0.206
newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011
youtube	0.0458	0.001	32.809	0.000	0.043	0.049

If FB investment increases by 1000 USD

→ sales increase by 189 USD <u>on average</u>, keeping <u>other variables fixed.</u>

$$Sales = \mathbf{3.527} + \mathbf{0.046} * Youtube\ Ads + \mathbf{0.189} * Facebook\ Ads - \mathbf{0.001} * Newspaper\ Ads$$

czechitas

Quiz

Variables that are significant at **5% or lower** level:

A. x1, x2, x4, x5

A. only x2

A. x1, x2, x4

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1675     0.1384  -1.210  0.23281
x1            0.5306     0.1754   3.025  0.00414
x2           -0.4115     0.1769  -2.326  0.02470
x3            0.1289     0.1673   0.771  0.44510
x4           -0.5884     0.1818  -3.237  0.00230
x5           -0.2476     0.1432  -1.728  0.09094
```

czechitas

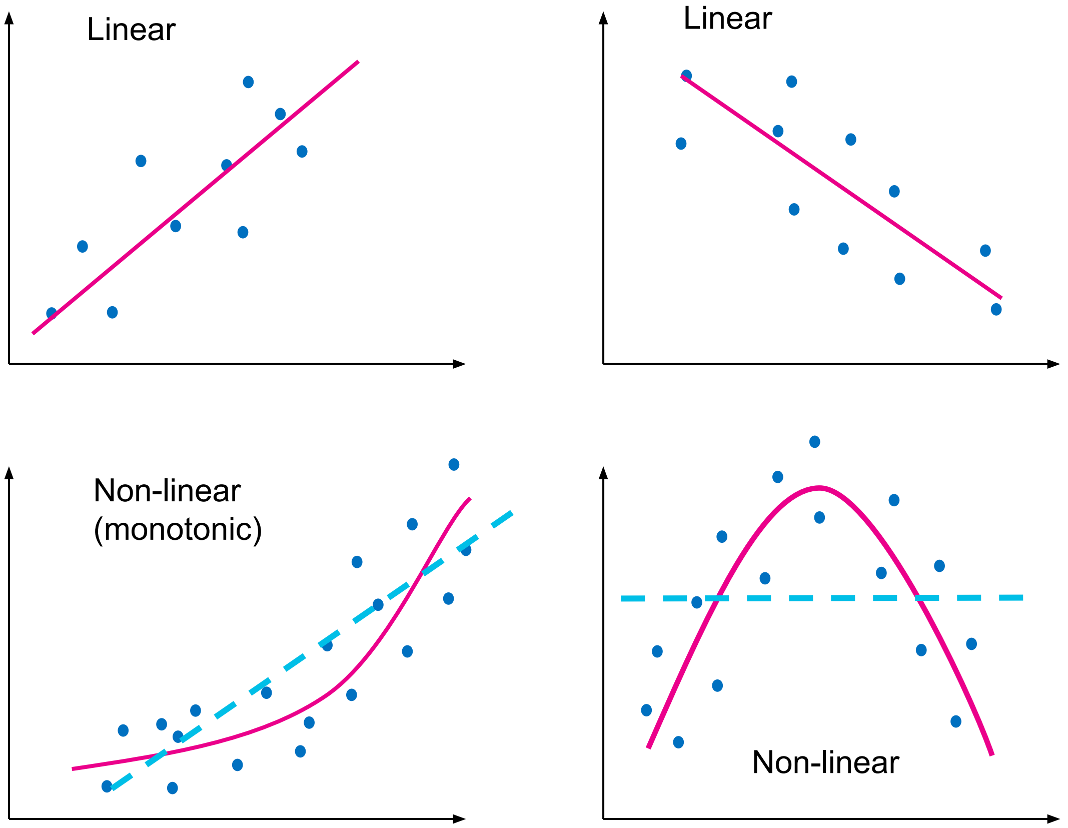
OLS MODEL: ASSUMPTIONS

OLS Assumptions

1. Linear relationship
2. No multicollinearity
3. Random sample
4. No omitted variable
5. Homoskedasticity
6. Normality

OLS Assumptions

1. Linear Relationship



Main Takeaway

- Relationship should be linear
- Non-linear relationship may or may not jeopardize our conclusion

Formal Assumption

$$Y_i = X_i^T \boldsymbol{\beta} + e_i, \qquad \mathbb{E}[e_i | X_i] = 0$$

$$X_i = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_k \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

2. No Multicollinearity

Multicollinearity - occurrence of high correlations among two or more independent variables in a multiple regression model.

youtube	facebook	newspaper	sales
276.12	45.36	83.04	26.52
53.40	47.16	54.12	12.48
20.64	55.08	83.16	11.16
181.80	49.56	70.20	22.20
216.96	12.96	70.08	15.48
10.44	58.68	90.00	8.64

2. No Multicollinearity

WHY?

- An isolated relationship between each independent variable and the dependent variable is needed.
- Stronger multicollinearity ⇒ higher standard errors (explodes to infinity for correlation approaching 1).

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3526.6672	374.290	9.422	0.000	2788.515	4264.820
facebook	$\boxed{0.1885}$	0.009	21.893	0.000	0.172	0.206
newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011
youtube	0.0458	0.001	32.809	0.000	0.043	0.049

If FB investment increases by 1000 USD

→ sales increase by 189 USD on average, keeping other variables fixed

2. No Multicollinearity

HOW TO TEST?

- **Correlation matrix:** correlation above, say, 70% may be problematic

	youtube	facebook	newspaper
youtube	1.000		
facebook	0.055	1.000	
newspaper	0.057	0.354	1.000

- **Variance Inflation Factor (VIF)**
  - Above 5: multicollinearity might be present
  - Above 10: multicollinearity certainly present

2. No Multicollinearity

SOLUTION?

1. Remove variable
   - remove one of the two highly correlated variables
   - hypotheses or theory should guide your decision
2. Specialized methods
   - ridge regression, LASSO, elastic net, principal component analysis
   - better for large datasets with many variables

3. Random Sample & Sample Bias

Random Sample:

- Individual observations are independent from each other
- All individuals have the same probability of sampling

Examples of violations:

- Analyzing impact of education on income using one individual over her/his lifetime
- Analyzing the impact of education on income when high-income individuals are less willing to share information about their income
- MSD project example: analyzing the productivity of farms in France only for farms that have good data about productivity

4. No Omitted Variable

Causal Impact

"Impact of X on Y while everything else remains the same."

Wage Example

Consider two models

$$wage = \alpha_0 + \alpha_1 education + e$$

$$wage = \beta_0 + \beta_1 education + \beta_2 ability + u$$

- Problem: education and ability is correlated $\Rightarrow$
  α_1 captures impact of education and partially impact of ability
- "Solution": we have to add ability to the model to control it ("keep it the same")
- Omitted Variable Bias: $\alpha_1 - \beta_1$

czechitas

4. No Omitted Variable

Possible solutions:

1. Include all relevant variables
2. Panel models (covered next time)
3. Randomized experiment
4. Regression Discontinuity Design
5. ...

4. No Omitted Variable - Correlation Is Not Causation

Examples of omitted variable bias (funny?)

- Regression of children injuries on ice cream consumption within one month has positive beta
  ⇒ ice cream is causing injuries
- Regression of health on recent visit of hospital has negative beta
  ⇒ hospitals have negative impact on health

Violation of OLS Assumptions

1. Linear relationship → biased betas (underestimates or overestimated betas)
2. No multicollinearity → high variance of beta estimates
3. Random sample → biased betas
4. No omitted variable → biased betas

The first four assumptions are crucial to obtain correct betas.

5. Homoskedasticity
6. Normality

Unbiasedness of OLS

1. Linear relationship
2. No multicollinearity
3. Random sample
4. No omitted variable

5. Homoskedasticity
6. Normality

Violation of the other assumptions does not make beta estimates invalid.

It makes **statistical inference invalid** (standard errors, p-values, ...).

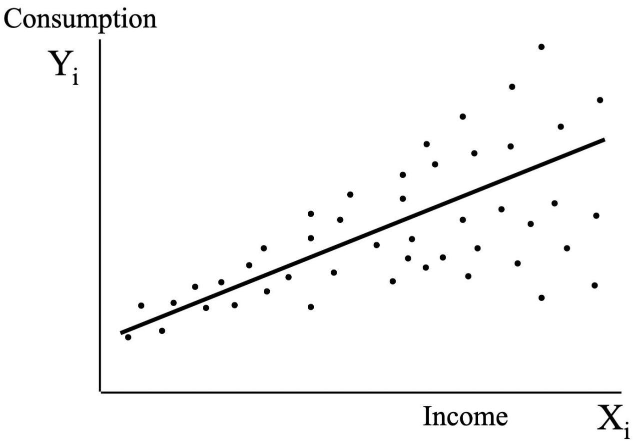
5. Homoskedasticity

MEANING

Variance of residuals is the same across all values of the independent variables.

✅ Homoskedasticity:

❌ Heteroskedasticity:



5. Homoskedasticity

Solution of Heteroskedasticity

- Robust standard errors (White standard errors)

Recommendation

- Use robust standard errors unless you have strong reason to believe the errors are homoskedastic, do not perform test of heteroskedasticity (sequential testing)

- Non-robust errors under heteroskedasticity
  $\Rightarrow$ errors are inconsistent (potentially completely wrong for any sample size)

- Robust errors under homoskedasticity
  $\Rightarrow$ errors are consistent, but inefficient (= estimation is less precise, but increasing sample size gives “true” value)

czechitas

5. Homoskedasticity

HOW TO TEST?

Two tests are commonly used:

- **Breusch-Pagan Test** – tests simple form of heteroskedasticity
- **White Test** - tests various forms of heteroskedasticity

- Null hypothesis: homoskedasticity
- Available in statistical software

6. Normality

Hypothesis test $\mathbb{H}_0: \beta_j = 0$

- How unlikely it is to obtain estimate $\widehat{\beta_j}$ or something more distant from 0 when $\mathbb{H}_0$ is true? (= p-value)
- We need to know distribution of $\widehat{\beta_j}$ to answer the question

Need normality of betas (under $\mathbb{H}_0$)

$$\widehat{\beta}_j \sim N\left(0, \sigma^2_{\widehat{\beta_j}}\right)$$

Normality holds when:

1. Residuals are normally distributed

OR

2. We have large sample

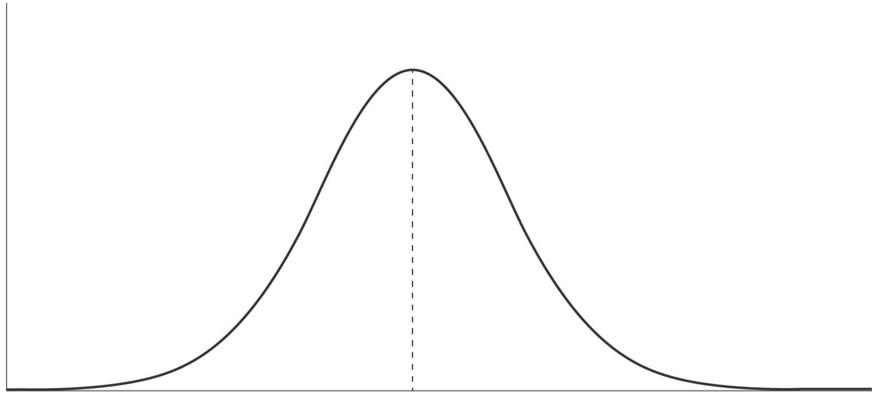
6. Normality of residuals

$$Sales = \mathbf{3.527} + \mathbf{0.046} * Youtube\ Ads + \mathbf{0.189} * Facebook\ Ads - \mathbf{0.001} * Newspaper\ Ads$$

sales	Estimated Sales
26.52	24.63
12.48	14.81
11.16	14.77
22.20	21.12
15.48	15.83
8.64	14.97

→

Residuals
1.89
-2.33
-3.61
1.08
-0.35
-6.33



- Not tested in practice

6. Asymptotic normality

- When number of observations is high the estimates of betas are approximately normal
- Follows from central limit theorem (and few other theorems)
- In practice you almost always rely on asymptotic normality

Quiz

Which statement is **false** about OLS assumptions?

A. Heteroskedasticity implies we have invalid p-values

A. We have to use a random sample

A. Homoskedasticity cannot be statistically tested

A. Omitted variable causes bias in OLS estimators

SUMMARY

Summary

Hypotheses testing: (not) rejecting our assumption with help of historical data

→ We did not predict anything today :)

Example: *"Women earn lower salaries than men."*

czechitas

Summary

Hypotheses testing: (not) rejecting our assumption with help of historical data

"Women earn lower salaries than men."

$$Income = \beta_0 + \beta_1 FemaleGender + \beta_2 Education + \beta_3 Age + \varepsilon$$

Personal ID	FemaleGender	Income	Education (years)	Age
2343	1	50 000	17	35
1213	0	35 000	15	32
4533	0	40 000	15	53
4563	0	100 000	19	51
…	…	…	…	…

Summary

When testing with **linear model (OLS)**, we are interested in:

- Model performance

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     570.3
Date:                Sun, 23 Apr 2023   Prob (F-statistic):           1.58e-96
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3526.6672    374.290      9.422      0.000    2788.515    4264.820
facebook       0.1885      0.009     21.893      0.000       0.172       0.206
newspaper     -0.0010      0.006     -0.177      0.860      -0.013       0.011
youtube        0.0458      0.001     32.809      0.000       0.043       0.049
==============================================================================
```

Summary

When testing with **linear model (OLS),** we are interested in:

- Model performance
- Beta coefficients

INTERPRETATION:

If FB investment increases by 1000 USD

→ sales increase by 189 USD on average, keeping other variables fixed

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Sun, 23 Apr 2023	Prob (F-statistic):	1.58e-96

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3526.6672	374.290	9.422	0.000	2788.515	4264.820
facebook	0.1885	0.009	21.893	0.000	0.172	0.206
newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011
youtube	0.0458	0.001	32.809	0.000	0.043	0.049

Summary

When testing with **linear model (OLS)**, we are interested in:

- Model performance
- Beta coefficients
- Statistical significance

We want p-value < 0.1,
ideally even p-value < 0.05

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Sun, 23 Apr 2023	Prob (F-statistic):	1.58e-96

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3526.6672	374.290	9.422	0.000	2788.515	4264.820
facebook	0.1885	0.009	21.893	0.000	0.172	0.206
newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011
youtube	0.0458	0.001	32.809	0.000	0.043	0.049

Summary

We need to check that **model assumptions** hold.

1. Linear relationship → biased betas
2. No multicollinearity → high variance of beta estimates
3. Random sample → biased betas
4. No omitted variable → biased betas

Values of betas

5. Homoskedasticity → invalid inference (use robust errors)
6. Normality → invalid inference (large sample desired)

Validity of inference

czechitas

Thank you for your attention!