BIEM/BIEF

A.Y. 2024/2025

BLAB

HANDOUTS

STATISTICS

-SECOND PARTIAL-

WRITTEN BY

JOYCE COLING

TEACHING DIVISION

This handout is written by students with no intention of replacing university materials.

It is a useful tool for studying the subject, but does not guarantee preparation as exhaustive and complete as the material recommended by the University.

Confidence Interval

GENERAL FRAMEWORK

REMINDER: - point estimation: a single value assigned to the parameter
- interval estimation: a range of values within the parameter is very "likely" to be found

Consider examples on "Commuters" → let's derive the interval estimate

X = "Distance" ~ $N(\mu, \sigma^2 = 100)$

sample of 25 commuters → n=25

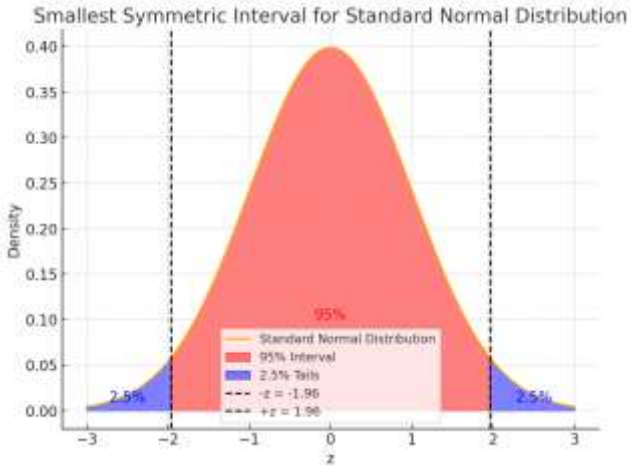
The sample mean is the natural estimator of μ, so:

$\underline{X} \sim N(\mu, \frac{\sigma^2}{n} = \frac{100}{25} = 4)$

$$SE(\underline{X}) = \sigma_{\underline{X}} = \sqrt{4} = 2$$

$Z = \frac{\underline{X} - \mu}{2} \sim N(0,1)$

→ Let's find the smallest symmetric interval where the standard normal r.v. has a 95% probability of being within the interval



P(Z < +z) = 0.975

↳ $P(-1.96 < Z < +1.96) = 0.95$

→ Knowing that $Z = \frac{\underline{X} - \mu}{2}$

$$P(-1.96 < \frac{\underline{X} - \mu}{2} < +1.96) = 0.95$$

$$P(\mu - 3.92 < \underline{X} < \mu + 3.92) = 0.95$$

→ Remember that the population mean is our goal, we rewrite:

$$P(\underline{X} - 3.92 < \mu < \underline{X} + 3.92) = 0.95$$

→ Interpretation: it's equal to GSY. The probability that the random interval [$\underline{X} - 3.92, \underline{X} + 3.92$]

contains the constant value μ

↳ this random interval is therefore an interval estimator of μ

➜ If now we draw a sample, we obtain a numeric interval $(\underline{X} - 3.92, \underline{X} + 3.92)$ ← that can be interpreted as a interval estimate of μ

CONFIDENCE INTERVAL: FORMAL DEFINITION

Let ϑ be a generic unknown parameter of the population. Consider two random variable A_ϑ and B_ϑ, that are both function of the random sample $(X_1, X_2, \dots, X_n)$, such that:
$P(A_\vartheta < \vartheta < B_\vartheta) = 1 - \alpha$, with $0 < \alpha < 1$
If a and b are the observed values of A_ϑ and B_ϑ, we define *"confidence interval A2 level* $100(1-\alpha)\%$ *for* ϑ*"*
The interval $(a; b)$, formally we can write: $ci_{1-\alpha}(\vartheta) = (a; b)$
Note that:

$(A_\vartheta; B_\vartheta)$ ⇒ interval estimator of ϑ
$(a; b)$ ⇒ interval estimate of ϑ, that we generally call "confidence int."
$100(1-\alpha)\%$ ⇒ confidence level of the interval

What is the interpretation of the Confidence Level?
Frequentist interpretation of the "confidence level":
The confidence level $100(1-\alpha)\%$ can be interpreted in this way:

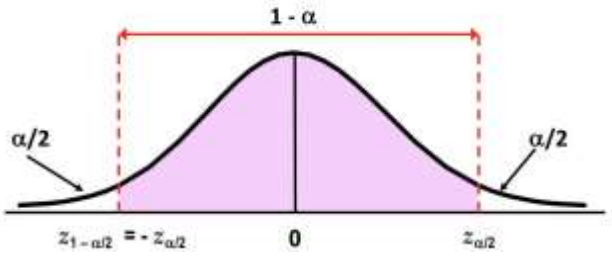
⇒ Selecting a large number of samples, from the same population and all of the same size n, $100(1-\alpha)\%$ of the intervals created starting from those samples will contain the real value of the parameters [$100(\alpha)\%$ will not]
⇒ Repeating sampling procedure

⇒ Confidence intervals cases ↗ GROUP A: single population
↘ GROUP B: differences between two population

A) CONFIDENCE INTERVALS: SINGLE POPULATION

Case A1: C.I. for the mean of a normal pop. (variance known)

Let X_1, X_2, …, X_n be a random sample of size *n* selected from a normal population, with mean μ and a known variance σ^2. In this case, $\underline{X} \sim N(\mu, \frac{\sigma^2}{n}) \rightarrow \frac{\underline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$



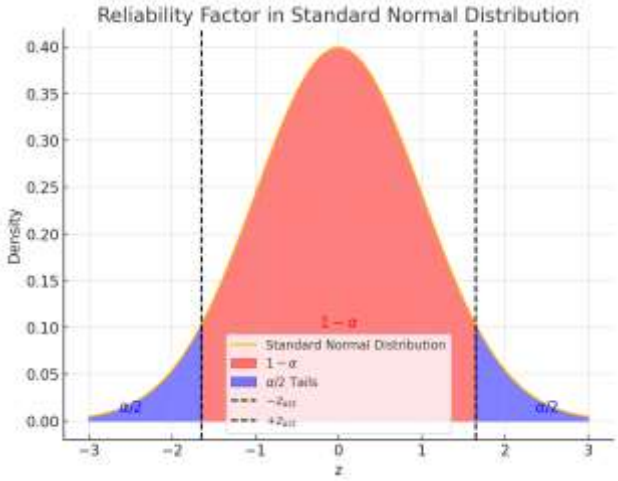
➜ $Prob(-z_{\frac{\alpha}{2}} \leq \frac{\underline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$

➜ $Prob(\underline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \underline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$

Given the observed sample mean $\underline{x}$, we define the $100(1-\alpha)\%$ confidence interval for μ, the following:

$$ci_{1-\alpha}(\mu) = \underline{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = \underline{x} \pm z_{\frac{\alpha}{2}} \cdot SE(\underline{X}) \qquad \text{NOTE: } [ci(\mu) = \underline{x} \pm ME]$$

- L.C.L. ⇒ Lower Confidence Limit: $\underline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
- U.C.L. ⇒ Upper Confidence Limit: $\underline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
- M.E. ⇒ Margin of Error: $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = z_{\frac{\alpha}{2}} \cdot SE(\underline{X})$
- W ⇒ Width of the interval: $2 \cdot ME$
- RELIABILITY FACTOR: $z_{\frac{\alpha}{2}}$ ⇒ it's the quantile of order $\left(1 - \frac{\alpha}{2}\right)$ of the standard normal distribution



$$p(z < z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$$

So the reliability factor $z_{\frac{\alpha}{2}}$ it's determined entirely by the confidence level $(1-\alpha)$.

The most common situations are: $1-\alpha = 90\% \quad \Rightarrow \quad z_{\frac{\alpha}{2}} = 1.645$

$$1-\alpha = 95\% \quad \Rightarrow \quad z_{\frac{\alpha}{2}} = 1.96$$

$$1-\alpha = 99\% \quad \Rightarrow \quad z_{\frac{\alpha}{2}} = 2.575$$

How to reduce the Margin of Error?

- $\uparrow n \Rightarrow \downarrow ME$
- $\downarrow (1-\alpha) \Rightarrow \downarrow ME$
- $\downarrow \sigma \Rightarrow \downarrow ME$ (not directly actionable)

Case A2: C.I. for the mean of a normal population (variance unknown)

Normal population = $X \sim N \Rightarrow \frac{\underline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

If σ is unknown, we use S to estimate it ⇒ However, the random variable $T = \frac{\underline{X}-\mu}{\frac{S}{\sqrt{n}}}$ is no longer normal → it is distributed as ***Student's t***, with $(n-1)$ degrees of freedom: $\frac{\underline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim t_{n-1}$

Student's t R.V.:
- similar to the Standard Normal (bell-shaped, symmetric, centered in zero)
- as *n* increases, the Student's t can be approximated by the Standard Normal

Interval definition

Let $X_1, X_2, \ldots, X_n$ be a random sample of size *n* selected from a normal population, with mean μ and a unknown variance σ^2. Given the observed sample mean $\underline{x}$, we define the $100(1-\alpha)\%$ confidence interval for μ, the following:

$$ci_{1-\alpha}(\mu) = \underline{x} \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} = \underline{x} \pm t_{n-1,\frac{\alpha}{2}} \cdot se(\underline{X})$$

NOTE:

- M.E. ⇒ Margin of Error: $t_{n-1,\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} = t_{n-1,\frac{\alpha}{2}} \cdot se(\underline{X})$ **(estimate of the Standard Error!!)**
- W ⇒ Width of the interval: $2 \cdot ME$
- RELIABILITY FACTOR: $z_{\frac{\alpha}{2}}$ ⇒ it's the quantile of order $\left(1-\frac{\alpha}{2}\right)$ of the t_{n-1}

LARGE SAMPLES CONFIDENCE INTERVALS

Case A3: Large sample C.I. for a pop. mean

It's typically used when:
- Pop. distribution is unknown (or not normal)
- Pop. variance is unknown
- The sample size is big ($n \geq 30$)

Under this conditions we can write: $\frac{\underline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim N(0;1)$

So the $100(1-\alpha)\%$ confidence interval for μ, is:

$$ci_{1-\alpha}(\mu) = \underline{x} \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \simeq \underline{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$$

NOTE:
- All the C.I. are in the form: $\underline{x} \pm ME$
- The width of the C.I. when the variance of the Pop. is unknown depends on the observed sample variance

Case A4: Large sample C.I. for a pop. proportion

When the sample is large ($n \cdot p(1-p) > 5$), we know that:

$$\hat{P} \approx N\left(p, \frac{p(1-p)}{n}\right) \Rightarrow \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0;1) \qquad E(\hat{P}) = p \qquad Var(\hat{P}) = \frac{p(1-p)}{n}$$

However p is unknown, so we use $\hat{p}$ as an approximation.

It is still true that, if $(n \cdot \hat{p}(1-\hat{p}) > 5)$, we have: $\frac{\hat{P}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0;1)$

So the $100(1-\alpha)\%$ confidence interval for the pop. proportion is:

$$ci_{1-\alpha}(p) = \hat{p} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} = \hat{p} \pm z_{\frac{\alpha}{2}} \cdot se(\hat{p})$$

SAMPLE SIZE DETERMINATION

We set a target value for the Margin of Error (ME) and determine the sample size (n) accordingly

- **C.I. FOR THE POP. MEAN** (variance known)

$$ME = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow \quad n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 \cdot \sigma^2}{ME^2}$$

- **C.I. FOR A POP. PROPORTION**

$$ME = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \Rightarrow \quad n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 \cdot \hat{p}\,(1-\hat{p})}{ME^2}$$

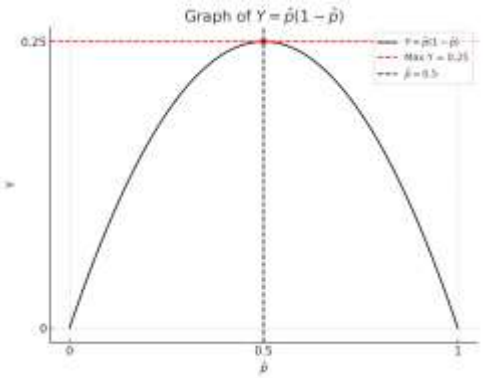
$\hat{p}$ is not known before the sample selection, so we substitute $\hat{p}\,(1-\hat{p})$ with 0.25, so we have:

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 \cdot 0.25}{ME^2}$$

Why 0.25?

$Y = f(\hat{p}) = \hat{p}\,(1-\hat{p}) = \hat{p} - (\hat{p})^2$

We choose $\hat{p} = 0.25$ that maximizes the expression $\hat{p}\,(1-\hat{p})$, so that the sample size will be in any case sufficient to guarantee the requires ME (that means we make sure that n will be "safely high")



Confidence Interval pt.2

CONFIDENCE INTERVAL: FORMAL DEFINITION

B) CONFIDENCE INTERVALS: DIFFERENCE BETWEEN TWO POPULATIONS

Case B1: C.I. for the difference between two pop. means (dependent samples)

Population 1 - described by a v.a. X	Population 2 - described by a v.a. Y
$X_1, \dots, X_{n_X}$ sample of size n_X, iid X with mean μ_X and variance σ_X^2	$Y_1, \dots, Y_{n_Y}$ sample of size n_Y iid Y with mean μ_Y and variance σ_Y^2
$\bar{X} = (X_1 + \dots + X_{n_X})/n_X$	$\bar{Y} = (Y_1 + \dots + Y_{n_Y})/n_Y$
$S_X^2 = \sum_{i=1}^{n_X} \frac{(X_i - \bar{X})^2}{n_X - 1}$	$S_Y^2 = \sum_{i=1}^{n_Y} \frac{(Y_i - \bar{Y})^2}{n_Y - 1}$

Let's consider two dependent / paired random samples of *n* pairs, selected from two normal pop. X and Y, with means μ_X and μ_Y (and typically unknown variances).
After the sample selection, we have a set of *n* pairs:

$$x_i = x_1, x_2, x_3, \dots, x_n$$
$$y_i = y_1, y_2, y_3, \dots, y_n$$
$$d_i = d_1, d_2, d_3, \dots, d_n$$

↳ Calling $d_i = x_i - y_i$ the difference between the pairs of observations, $\underline{d} = \underline{x} - \underline{y}$ the sample mean of the realizations of the differences and S_D the standard deviation of the differences. We define the $100(1-\alpha)\%$ confidence interval for $(\mu_X - \mu_Y)$, the following:

$$ci_{1-\alpha}(\mu_X - \mu_Y) = \underline{d} \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{S_D}{\sqrt{n}} = \underline{d} \pm t_{n-1,\frac{\alpha}{2}} \cdot se(\underline{D})$$

where $S_D = \sqrt{S_X^2 + S_Y^2 - 2S_{XY}}$ is the ***corrected sample variance***

In case of known variances (rare), we have:

$$ci_{1-\alpha}(\mu_X - \mu_Y) = \underline{d} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma_D}{\sqrt{n}} = \underline{d} \pm z_{\frac{\alpha}{2}} \cdot SE(\underline{D})$$

where $\sigma_D = \sqrt{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$

Case B2: C.I. for the difference between two pop. means (independent samples)

Consider two random independent samples with n_X and n_Y observations, selected from two normal populations X and Y, with mean μ_X and μ_Y , and unknown variances.
Let $\underline{X}$ and $\underline{Y}$ be the sample means and S_X^2 and S_Y^2 the sample variances.

1. In the rare case of known variances, the $100(1-\alpha)\%$ confidence interval is:

$$ci_{1-\alpha}(\mu_X - \mu_Y) = (\underline{X} - \underline{Y}) \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} = (\underline{X} - \underline{Y}) \pm z_{\frac{\alpha}{2}} \cdot SE(\underline{X} - \underline{Y})$$

2. In case the unknown pop. variances are assumed to be equal ($\sigma_X^2 = \sigma_Y^2 = \sigma^2$), the $100(1-\alpha)\%$ confidence interval for $(\mu_X - \mu_Y)$ is:

$$ci_{1-\alpha}(\mu_X - \mu_Y) = (\underline{x} - \underline{y}) \pm t_{n_X+n_Y-2,\frac{\alpha}{2}} \cdot \sqrt{\frac{S_{POOL}^2}{n_X} + \frac{S_{POOL}^2}{n_Y}}$$

where $S_{POOL}^2 = \frac{(n_X-1)\,S_X^2 + (n_Y-1)\,S_Y^2}{n_X+n_Y-2}$ is the ***pooled corrected sample variance***

substituted to σ_X and σ_Y to estimate the $se(\underline{X} - \underline{Y})$

3. In case of unknown variances assumed to be different the C.I. is built using the estimates s_x^2 and s_y^2; the reliability factor is still based on Student's t, but with a different number of degrees of freedom (can only be obtained with RStudio).

⚠ **NOTE:** The confidence intervals built in case 2 & 3 have ≠ ME, due both to ≠ Standard Error Estimates (se) and to ≠ reliability factors.

4. When the distribution of the two pop. is not normal (or unknown) but the sample size of both the samples is large ($n_x \geq 30$ and $n_y \geq 30$), we can apply the CLT and base the C.I. on the Standard Normal Distribution. In this case the reliability factor will be $Z_{\frac{\alpha}{2}}$

Case B3: C.I. for the difference between two proportions

Population 1	Population 2
$X_1, \dots, X_{n_X}$ sample of size n_X, iid X distributed according to a Bernoulli of parameter p_X (i.e. X takes on a value of 1 or 0 depending on whether a success is observed or not, and p_X = proportion of successes in the population), with $E(X) = p_X$ and $Var(X) = p_X(1-p_X)$	$Y_1, \dots, Y_{n_Y}$ sample of size n_Y iid Y distributed according to a Bernoulli parameter p_Y (i.e. Y takes on a value of 1 or 0 depending on whether a success is observed or not, and p_Y = proportion of successes in the population), with $E(Y) = p_Y$ and $Var(Y) = p_Y(1-p_Y)$
$\hat{P}_X = (X_1 + \dots + X_{n_X})/n_X$ = sample proportion of successes	$\hat{P}_Y = (Y_1 + \dots + Y_{n_Y})/n_Y$ = sample proportion of successes

Consider two random independent samples with n_X and n_Y observations, selected from two Bernoulli pop. X and Y, of parameters P_X and P_Y .
Let $\hat{P}_X$ and $\hat{P}_Y$ be the corresponding sample proportions. In the case of independent samples, we have seen that:

$$Var(\hat{P}_X - \hat{P}_Y) = \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}$$

However, $Var(\hat{P}_X - \hat{P}_Y)$ is never known (it depends on p_X and p_Y), but it can be estimated by replacing them with the observed sample proportions $\hat{p}_X$ and $\hat{p}_Y$
the $100(1-\alpha)\%$ confidence interval for the difference between two pop. proportions $(p_X - p_Y)$ is:

$$ci_{1-\alpha}(p_X - p_Y) = (\hat{p}_X - \hat{p}_Y) \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} = (\hat{p}_X - \hat{p}_Y) \pm z_{\frac{\alpha}{2}} \cdot se(\hat{P}_X - \hat{P}_Y)$$

NOTE: This is true when both the samples are big enough, because in that case we have that:

$$\frac{(\hat{P}_X - \hat{P}_Y) - (p_X - p_Y)}{\sqrt{\frac{\hat{P}_X(1-\hat{P}_X)}{n_X} + \frac{\hat{P}_Y(1-\hat{P}_Y)}{n_Y}}} \approx N(0;1)$$

Hypothesis testing

DEFINITIONS

STATISTICAL HYPOTHESIS

= a statement about a parameter θ of the population

HYPOTHESIS TESTING

= a procedure that consists in defining statistical hypotheses and in deciding on the basis of the sample whether they are acceptable or not

CHARACTERISTIC OF STATISTICAL HYPOTHESES

In the "Test Theory", hypothesis testing is a "conflict" between two opposing hypotheses:

$H_0 \Rightarrow$ NULL HYPOTHESIS

↘

not overlapped: only one can be true!!

↗

$H_1 \Rightarrow$ NULL HYPOTHESIS

The two hypotheses (H_0 and H_1) can be:

- SIMPLE: if they specify a single value for the parameter (i.e. when you have the "=")
- COMPOSITE: if they specify a range of values (<, >, ≥, ≤, ≠)

MEANING	H_0 (null)	H_1 (alternative)
	Status quo (true until proven otherwise)	Challenges the status quo (it's what the researchers is trying to prove)
CONTENT	Parameter of the population (never sample)	
ADMITTED SYMBOLS	=, ≥, ≤	≠, <, >
RESULTS	Reject or Fail to reject	Support or Do not support

IMPORTANT CONSIDERATION: the test is always conducted starting from the assumption that the **null hypothesis is true** (under H_0)

HYP. TESTING PROCEDURE

1. Specify the hypotheses
2. Identify the test statistic (and its distribution under H_0)
3. Formulate the decision rule
4. Insert the sample data and take a decision

TEST STATISTIC

= a sample statistic (reminder: a R.V. function of the random sample) used to perform a test of hyp., whose distribution must be known under the null hyp.

DECISION RULE

= it consists in dividing the sample space into two complementary regions: (R) rejection region and (A) acceptance region

CRITICAL VALUE

= value that defines the two regions (R) and (A)

➜ *How to determine the critical value?*
Determined a priori (in advance) by setting the **significance level** (α) of the test
↳ it's a small probability value that defines the critical value of the test (typically α = 0.05, 0.01, 0.1)

CONSEQUENCES OF THE DECISION

DECISION	GROUND TRUTH	
	H_0 IS TRUE	H_0 IS FALSE
REJECT H_0	**TYPE I ERROR** α	CORRECT DECISION $1-\beta$
FAIL TO REJECT H_0	CORRECT DECISION $1-\alpha$	**TYPE II ERROR** β

- Probability of TYPE I ERROR:
  α = P(reject H_0 | H_0 is true)
  ↳ α is the significance level of the test
  ↳ α is set in advance

- Probability of TYPE II ERROR:
  β = P(fail to reject H_0 | H_0 is false)
  ↳ givem α, we can compute β

- POWER OF THE TEST
  $1-\beta$ = P(reject H_0 | H_0 is false)

What are the factors affecting β (and so $1-\beta$)?

- $\alpha \downarrow \Rightarrow \beta \uparrow$ (keeping constant everything else)
- $n \uparrow \Rightarrow \beta \downarrow$ (but also $\alpha \downarrow$)
- $\sigma^2 \downarrow \Rightarrow \beta \downarrow$ (but also $\alpha \downarrow$)

Hypothesis testing cases

CASES

- GROUP A: Single Population
- GROUP B: Difference between two population
- GROUP C: Non parametric tests

GROUP A

CASE A1: TEST FOR THE MEAN OF A NORMAL POPULATION (variance known)

- STEP 1: Definition of the hypotheses
  - UTT: Upper-Tail Test (one-sided alternative hyp., right tail)

    $H_0: \mu = \mu_0$ (or $\mu \leq \mu_0$)

    $H_1: \mu > \mu_0$
  - LTT: Lower-Tail Test (one-sided alternative hyp., left tail)

    $H_0: \mu = \mu_0$ (or $\mu \geq \mu_0$)

    $H_1: \mu < \mu_0$
  - TTT: Two-Tail Test (two-sided alt. hyp.)

    $H_0: \mu = \mu_0$

    $H_1: \mu \neq \mu_0$

- STEP 2: Identify the test statistic and its distribution under H_0

  ⇒ If the null hyp. is true (under H_0)

  $X \sim N(\mu, \sigma^2) \Rightarrow X \sim N(\mu_0, \frac{\sigma^2}{n})$

  $$z = \frac{\underline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

- STEP 3: Decision rule
  - UTT: under H_0

    Take the sample realization of the test stat.

    $\underline{x}$ (or equiv. $z = \frac{\underline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$) ⇒ observed value of the sample stat.

    Decision rule: we reject H_0 if:

    1. $\underline{x} > \underline{x}_* = \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$ or equivalently 2. $z = \frac{\underline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha$

    z_α is the quantile of order $(1-\alpha)$ of the $N \sim (0,1)$

  - LTT: under H_0

    Decision rule: we reject H_0 if:

    1. $\underline{x} < \underline{x}_* = \mu_0 - z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$ or equivalently 2. $z = \frac{\underline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} < -z_\alpha$

- TTT: H_1: under H_0

Decision rule: we reject H_0 if:

1. $\underline{x} > \underline{x}_* = \mu_0 + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ or $\underline{x} < \underline{x'}_* = \mu_0 - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

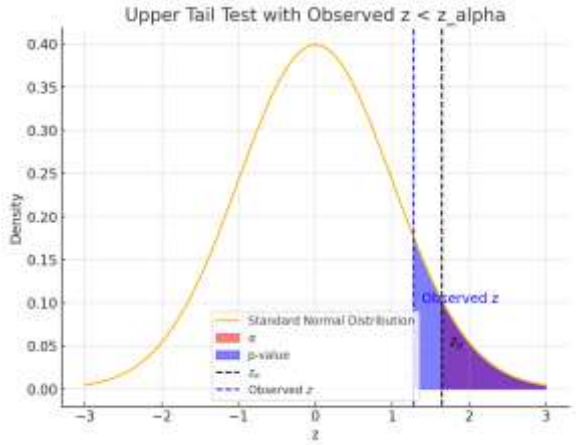
   or equivalently

2. $|z| = \left|\frac{\underline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right| > z_{\frac{\alpha}{2}}$

P-VALUE

= Probability of obtaining (under H_0) a value of the test statistic that is equal or more extreme than the observed value

Take an Upper Tail Test $H_1: \mu > \mu_0$



$\alpha = P(\underline{X} > \underline{x}_* \mid H_0: \mu = \mu_0)$ [defined in advance]

$P = P(\underline{X} > \underline{x} \mid H_0: \mu = \mu_0)$ [calculated based on sample results]

CASE A2: TEST FOR THE MEAN OF A NORMAL POPULATION (variance unknown)

➢ STEP 1: Hypotheses definition: UTT, LTT, TTT (identical to Case A1)

➢ STEP 2: Identify the test statistic and its distribution under H_0

$$t = \frac{\underline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

➢ STEP 3: Decision rule

- UTT: We reject H_0 if $t = \frac{\underline{X}-\mu_0}{\frac{S}{\sqrt{n}}} > t_{n-1,\alpha}$
- LTT: We reject H_0 if $t = \frac{\underline{X}-\mu_0}{\frac{S}{\sqrt{n}}} < -t_{n-1,\alpha}$
- TTT: We reject H_0 if $|t| = \left|\frac{\underline{X}-\mu_0}{\frac{S}{\sqrt{n}}}\right| > t_{n-1,\frac{\alpha}{2}}$

  or for <u>all the cases</u> we reject H_0 if $P - value < \alpha$

CASE A3: LARGE SAMPLE TEST FOR A POP. MEAN

If the distribution of the pop. is unknown (or not normal),we can apply the CLT when the sample is large enough, so the distribution of the test statistic under H_0 will be approx. normal

➢ STEP 1: Hypotheses definition: UTT, LTT, TTT (identical to Case A1)

➢ STEP 2: Identify the test statistic and its distribution under H_0 (for $n \geq 30$)

$$\frac{\underline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \approx N(0,1)$$

➢ STEP 3: Decision rule

↳ Same as Case A2, but the critical values will be based on the normal distribution (the mean, for instance, using z_α instead of $t_{n-1,\alpha}$)

NOTE ON THE HYPOTHESES:

Why in the Upper Tail Test, we use the same decision rule for H_0: $\mu = \mu_0$ and H_0: $\mu \leq \mu_0$?

Because if we reject the null hyp. H_0: $\mu = \mu_0$, we will consequently reject all the values smaller than μ_0

Same considerations for Lower Tail Test

CASE A4: TEST FOR THE PROPORTION OF A POPULATION (LARGE SAMPLE)

➢ STEP 1: Definition of the hypotheses

- UTT: Upper-Tail Test (one-sided alternative hyp., right tail)

  H_0: $P = P_0$ (or $P \leq P_0$)

  H_1: $P > P_0$
- LTT: Lower-Tail Test (one-sided alternative hyp., left tail)

  H_0: $P = P_0$ (or $P \geq P_0$)

  H_1: $P < P_0$
- TTT: Two-Tail Test (two-sided alt. hyp.)

  H_0: $P = P_0$

  H_1: $P \neq P_0$

➢ STEP 2: Identify the test statistic and its distribution under H_0 (large sample: $n\,p_0(1-p_0) > 5$)

$$z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0,1)$$

➢ STEP 3: Decision rule

We reject H_0 if the observed value of the test stat ($z = \frac{\hat{P}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$) satisfies one of the following conditions:

- UTT: $z > z_\alpha$
- LTT: $z < -z_\alpha$
- TTT: $|z| > z_{\frac{\alpha}{2}}$

or for <u>all the cases</u> we reject H_0 if $P - value < \alpha$

GROUP B

TEST FOR THE DIFFERENCE BETWEEN TWO POP. MEANS

System of hypotheses:

- UTT: $H_0: \mu_X - \mu_Y = d_0$ (or $\leq d_0$)
  $H_1: \mu_X - \mu_Y > d_0$
- LTT: $H_0: \mu_X - \mu_Y = d_0$ (or d_0)
  $H_1: \mu_X - \mu_Y < d_0$
- TTT: $H_0: \mu_X - \mu_Y = d_0$
  $H_1: \mu_X - \mu_Y \neq d_0$

  Typically $d_0 = 0 \Rightarrow \mu_X - \mu_Y = 0 \Rightarrow \mu_X = \mu_Y$

CASE B1: DEPENDENT SAMPLES

- X and Y are two dependent populations (normally distributed)
- Having dependent samples (pairs of obs.), the test is based on the r.v. $D = X - Y$

➢ STEP 1: Hypotheses definition

➢ STEP 2: Identify the test statistic and its distribution under H_0

$$t = \frac{\underline{D} - d_0}{\frac{S_D}{\sqrt{n}}} \sim t_{n-1}$$

➢ STEP 3: Decision rule

We reject H_0 if the observed value of the test stat ($\frac{\underline{D} - d_0}{\frac{S_D}{\sqrt{n}}}$) is:

- UTT: $t > t_{n-1,\alpha}$
- LTT: $t < -t_{n-1,\alpha}$
- TTT: $|t| > t_{n-1,\frac{\alpha}{2}}$

  or for all the cases we reject H_0 if $P - value < \alpha$

CASE B2: INDEPENDENT SAMPLES

- X and Y are two normal populations
- Let's start assuming unknown variances but assumed to be equal

➢ STEP 1: Hypotheses definition

➢ STEP 2: Identify the test statistic and its distribution under H_0

$$t = \frac{(\underline{X} - \underline{Y}) - d_0}{\sqrt{\frac{S^2_{POOL}}{n_X} + \frac{S^2_{POOL}}{n_Y}}} \sim t_{n_X + n_Y - 2}$$

➢ STEP 3: Decision rule
We reject H_0 if the observed value of the test stat is:

- UTT: $t > t_{n_X+n_Y-2,\alpha}$
- LTT: $t < -t_{n_X+n_Y-2,\alpha}$
- TTT: $|t| > t_{n_X+n_Y-2,\frac{\alpha}{2}}$

or for <u>all the cases</u> we reject H_0 if $P-value < \alpha$

HYPOTHESIS TESTING AND CONFIDENCE INTERVALS CONNECTION

Consider a generic <u>TwoTail Test</u> :

$H_0: \vartheta = \vartheta_0$
$H_1: \vartheta \neq \vartheta_0$

If we <u>fail to reject</u> H_0 at α level of significance ⇔ The value ϑ_0 is included in the $1-\alpha$ confidence interval

If we <u>reject</u> H_0 at α level of significance ⇔ The value ϑ_0 is NOT included in the $1-\alpha$ confidence interval

Most typical case is:

$H_0: \mu_X - \mu_Y = 0$ $\quad$ $H_0: p = 0.5$
$H_1: \mu_X - \mu_Y \neq 0$ $\quad$ $H_1: p \neq 0.5$

<u>CASE B3</u>: TEST FOR THE DIFFERENCE BETWEEN TWO POP. PROPORTIONS

➢ STEP 1: Definition of the hypotheses

$H_0: P_X - P_Y = d_0$ → UTT $H_1: P_X - P_Y > d_0$
→ LTT $H_1: P_X - P_Y < d_0$
→ TTT $H_1: P_X - P_Y \neq d_0$

➢ STEP 2: Identify the test statistic and its distribution under H_0 (large sample: $n\,p_0(1-p_0) > 5$)

$$z = \frac{\hat{P} - d_0}{\sqrt{\frac{\hat{P}_X(1-\hat{P}_X)}{n_X} + \frac{\hat{P}_Y(1-\hat{P}_Y)}{n_Y}}} \approx N(0,1)$$

➢ STEP 3: Decision rule
We reject H_0 if:

- obs. value of the test stat is " more extreme" than the critical value
- $P-value < \alpha$

GROUP C

With non parametric tests there are no assumptions about normality or population parameters (like variance)

CASE C1: GOODNESS OF FIT TEST

We check, using these tests, if the observed data "fits" a specified distribution of the pop.
Consider a population where the statistical units can be classified according to K categories

CATEGORY	1	2	…	K	TOT	
Probability (under H_0)	P_1	P_2	…	P_K	1	⇐
Observed absolute frequencies (O_i)	O_1	O_2	…	O_K	n	⇒ SAMPLE DATA
Expected absolute frequencies (E_i)	$E_1 = n \cdot P_1$	$E_2 = n \cdot P_2$	…	$E_K = n \cdot P_K$	n	⇐

How to perform the test ?

➢ Measure the distance between the obs.freq. and the exp.freq.

➢ So define the test stat and the distribution under H_0

$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$ is distributed as CHI-SQUARED with K-1 degrees of freedom

CHI-SQUARED DISTRIB. PROPERTIES:

- can only assume non-negative values
- it's asymmetrical
- right skewed
- it's a family of distribution dependent on a single parameter ν = K-1 degrees of freedom

➢ Decision rule → rejection region only on the right tail

↳ We reject H_0 if:

observed value of $\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{K-1,\alpha}$

or we reject if $P - value < \alpha$

CASE C2: TEST OF STATISTICAL INDEPENDENCE

Consider a population whose units can be classified according two categorical variables A and B

A\B	B_1 , B_2 , B_3 , … , B_c	P_i	CROSSTAB A x B
A_1 , A_2 , A_3 , … , A_r	P_{11} P_{12} P_{13} … P_{1c} P_{21} P_{31} … P_{r1} P_{r2} P_{r3} … P_{rc}	. . .	P_{ij} : joint frequencies P_i : marginal freq. by row P_j : marginal freq. by column
P_J	. . .		

Consider now these two hypotheses:

H_0: A and B are statistically independent

H_1: A and B are statistically dependent

In case of statistical independence (under H_0): $P_{ij} = P_i \cdot P_j$

↳ it is known that in case of independence each joint frequency equals the product of the corresponding marginal frequency

➢ Collect a sample and calculate the Observed Absolute joint frequency (O_{ij})

A\B	B_1 , B_2 , B_3 , … , B_c	R_i	(r x c) CROSSTAB Legend: - O_{ij} : **observed absolute** frequencies - R_i : marginal abs. freq. by row - C_j : margina abs. freq. by column
A_1 , A_2 , A_3 , … , A_r	O_{11} O_{12} O_{13} … O_{1c} O_{21} O_{31} … O_{r1} O_{r2} O_{r3} … OP_{rc}	R_1 . . . R_r	
C_J	C_1 . . . C_c	n	

If A and B are independent, we expect that the Observed freq. (O_{ij}) to be very similar to the Expected freq. in case of independence (E_{ij})

$$E_{ij} = \frac{R_i \cdot C_j}{n} \quad , \qquad \forall i,j$$

➢ Idea: Measure the distance between O_{ij} and E_{ij} ⇒ if it's big enough we reject the null hyp. of independence

➢ Identify the test statistic and its distribution under H_0 (if the sample is large enough)

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

Larger sample? ⇒ if $E_{ij} \geq 5$ for at least 80% of the cells in the cross tab

➢ Decision rule:

↳ We reject H_0 if

observed value of $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} > \chi^2_{(r-1)(c-1),\alpha}$

or we reject if $P - value < \alpha$

Linear regression model

It's an extension of the linear correlation index

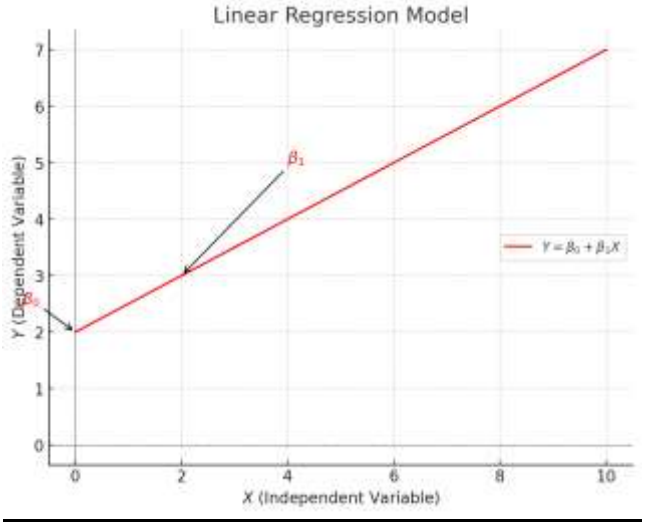
↳ The linear regression estimates the asymmetric linear relationship between two variables (X and Y):

$$Y = f(X)$$

↓ ↳ independent variable

dependent variable

Linear Equation: $Y = \beta_0 + \beta_1 \cdot X$



Goals of the linear regression

↗ Interpretation of the relationship between X and Y

↘ Prediction: Given X, what is the expected Y?

MODEL SPECIFICATION AND COEFFICIENT ESTIMATION

In the population: $Y = \beta_0 + \beta_1 \cdot X \Rightarrow$ *is this reasonable?*

↳ Deterministic relationship (like $T_F = 32 + 1.8T_C$)

More reasonable equation considering the real world:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

↓ deterministic component

↳ random component

Can we use the sample data to estimate β_0 and β_1 ?

↳ In the sample we collect *n* pairs (x_i, y_i)

➢ How can we estimate the "best" regression line coefficient starting from our sample data?

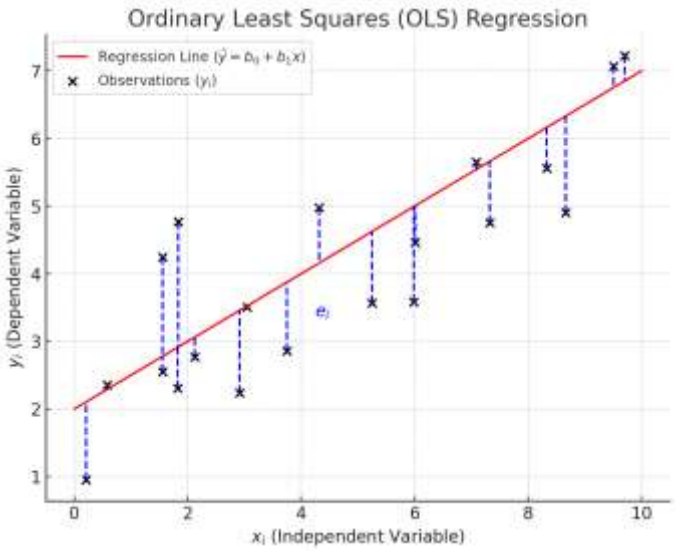
$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

↳ use the sample to estimate β_0 and β_1

- (β_0, β_1) : parameters of the linear regression model (population)
- $(\hat{\beta}_0, \hat{\beta}_1)$: estimators of the parameters (random variable)
- (b_0, b_1) : estimates of the parameters (2 numbers)
  - ↳ ESTIMATED MODEL: $\hat{y}_i = b_0 + b_1 \cdot x_i$
    - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$ ⇒ **estimator** of $\beta_0 + \beta_1 \cdot x_i$
    - $\hat{y}_i = b_0 + b_1 \cdot x_i$ ⇒ **estimate** of $\beta_0 + \beta_1 \cdot x_i$

ORDINARY LEAST SQUARES (OLS)

= Method to find the "best" estimates



The red line represents the regression line $\hat{y} = b_0 + b_1 x$

The black points are the observed values y_i

The blue dashed lines represent the residuals $e_i = y_i - \hat{y}_i$

Goal: Find the line that minimizes the squared errors

$$min \sum_{i=1}^{n} e_i^2 \Rightarrow min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The OLS method will find the best b_0 *and* b_1 (those which minimizes $\sum e_i^2$)

LEAST SQUARES ESTIMATES:

$$b_1 = \frac{S_{XY}}{S_X^2} = r_{XY} \cdot \frac{S_Y}{S_X}$$

$$b_0 = \underline{y} - b_1 \cdot \underline{x}$$

ESTIMATED REGRESSION LINE

$$\hat{y}_i = b_0 + b_1 \cdot x_i$$

b_1 (slope): it represents the estimated **average change** of the Y for one unit increase of the X

b_0 (constant): it is the expected value of the Y when $X = 0$

CHARACTERISTICS OF THE LINEAR REGRESSION MODEL

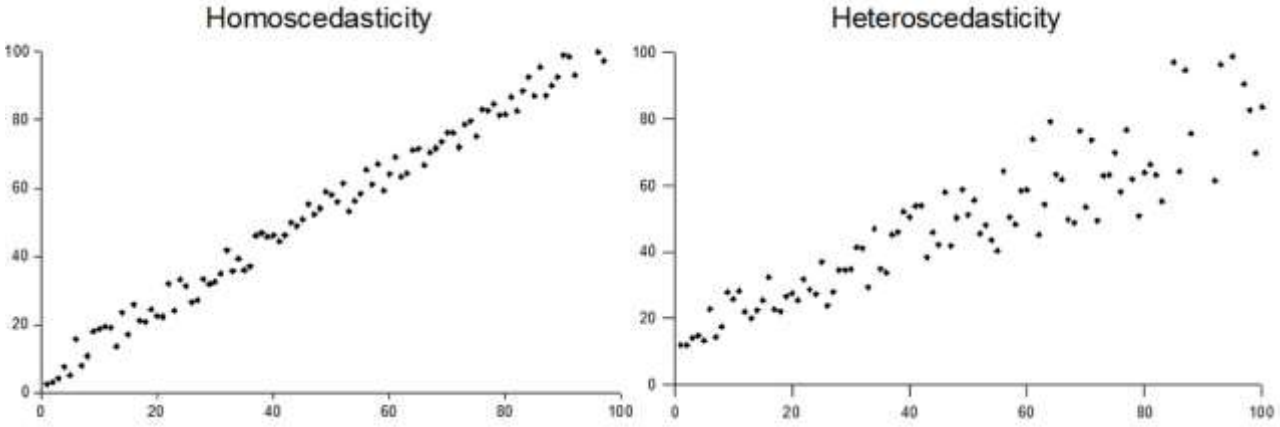
ASSUMPTIONS

The model relies on some important assumptions in order to guarantee the goodness of the estimators (and to make inference on the parameters)

↳ **Assumptions' List**:

N	Assumption	Description	Considerations / Implications
1	$E(\varepsilon_i) = 0$, $\forall i$	*The mean of the errors is zero*	If $E(\varepsilon_i) = 0$ then $E(Y_i \mid X = x_i) = \beta_0 + \beta_1 x_i$
2	$VAR(\varepsilon_i) = \sigma^2$, $\forall i$	*The variance of the errors is constant (HOMOSKEDASTICITY)*	$Var(\varepsilon_i) = Var(Y_i) = \sigma^2$ as $Var(\beta_0 + \beta_1 x_i) = 0$
3	$Cor(\varepsilon_i, \varepsilon_j) = 0$, $\forall i, j$	*Absence of autocorrelation of the errors*	If $Cor(\varepsilon_i, \varepsilon_j) = 0$ then $Cor(Y_i, Y_j) = 0$

4	$\varepsilon_i \sim N(0;\sigma^2)$, $\forall i$	*The errors are normally distributed*	The first three assumptions are the "core" conditions to obtain optimal estimates from the model. This last assumption is optional, needed only if we want to make inference on the parameters. If $\varepsilon_i \sim N(0;\sigma^2)$ then $Y_i \sim N(\beta_0 + \beta_1 x_i; \sigma^2)$



PROPERTIES OF OLS ESTIMATORS

It can be easily demonstrated that:

$$\hat{\beta}_1 = \frac{(x_i - \underline{x})}{(n-1)\,S_X^2}\,Y_i = \sum_{i=1}^{n} w_i Y_i$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \underline{x}\, w_i\right) Y_i = \sum_{i=1}^{n} k_i Y_i$$

} Least squares estimators are **linear combinations** of $Y_1, Y_2, \ldots, Y_n$

- UNBIASEDNESS: $E(\hat{\beta}_1) = E(\sum_{i=1}^{n} w_i Y_i) = \sum_{i=1}^{n} w_i\, E(Y_i) = \beta_1$
  $E(\hat{\beta}_0) = E(\sum_{i=1}^{n} k_i Y_i) = \sum_{i=1}^{n} k_i\, E(Y_i) = \beta_0$

- The OLS estimators of β_0 *and* β_1 are BLUE
  ↳ Best Linear Unbiased Estimators ⇒ with the lowest variance (most efficient)

$$Var(\hat{\beta}_0) = E[\hat{\beta}_0 - E(\hat{\beta}_0)]^2 = E(\hat{\beta}_0 - \beta_0)^2$$
$$Var(\hat{\beta}_1) = E[\hat{\beta}_0 - E(\hat{\beta}_0)]^2 = E(\hat{\beta}_0 - \beta_0)^2$$

It can be proved that under weak assumptions:

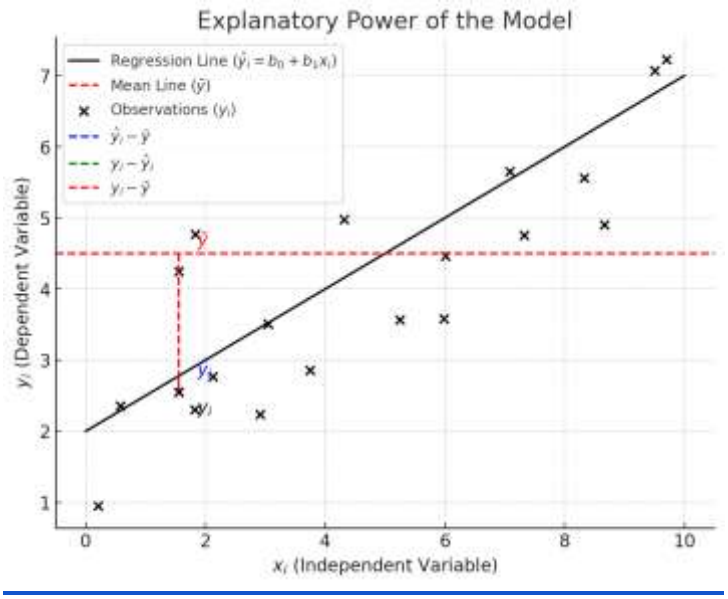
$$Var(\hat{\beta}_1) = Var(\sum_{i=1}^{n} w_i Y_i) = \sum_{i=1}^{n} w_i^2 Var(Y_i) = Var(Y) \sum_{i=1}^{n} w_i^2 = \frac{\sigma_\varepsilon^2}{(n-1)\,S_X^2}$$

$$Var(\hat{\beta}_0) = Var(\sum_{i=1}^{n} k_i Y_i) = \sum_{i=1}^{n} k_i^2 \, Var(Y_i) =$$

$$Var(Y) \sum_{i=1}^{n} k_i^2 = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\, S_X^2}\right)$$

➢ The variances of the estimators tend to zero as n ↑ ⇒ The distribution of the estimators is increasingly concentrated around the respite parameter as n ↑

EXPLANATORY POWER OF THE MODEL



One obs: $(y_i - \underline{y}) = (\hat{y}_i - \underline{y}) + (y_i - \hat{y}_i)$

All the obs: $\sum_i (y_i - \underline{y}) = \sum_i (\hat{y}_i - \underline{y})^2 + \sum_i (y_i - \hat{y}_i)^2$ ⇒ DECOMPOSITION OF THE DEVIANCE

SST SSR SSE

- SST: Deviance of Y
- SSR: Sum of squares of the regression

$$\sum_i (\hat{y}_i - \underline{y})^2 = b_1^2 \sum_i (\hat{y}_i - \underline{y})^2 = b_1^2 \, (n-1) \, s_X^2$$

- SSE: Sum of squares of the errors

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

COEFFICIENT OF DETERMINATION

$$R^2 = \frac{SSR}{SST} = \left(1 - \frac{SSE}{SST}\right)$$

➢ PROPERTIES:

- $0 \leq R^2 \leq 1$
- Interpretation of the R^2: percentage of the variance of the Y explained by the regression
- In a simple linear regression (one X) we have that: $R^2 = \rho_{XY}{}^2$

ESTIMATE OF THE VARIANCE OF THE MODEL (= VARIANCE OF THE ERROR)

From the assumptions we have: $Var(\varepsilon_i) = \sigma_\varepsilon^2$

↳ Estimator of σ_ε^2 : $\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i)]^2}{(n-2)}$

↳ Estimate of σ_ε^2 : $s_\varepsilon^2 = \frac{\sum_{i=1}^{n} [y_i - (b_0 + b_1 \cdot x_i)]^2}{(n-2)} = \frac{\sum_i e_i^2}{n-2} = MSE = \frac{SSE}{n-2}$

⇩

$$s_\varepsilon = \sqrt{\frac{\sum e_n^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$ ⇒ Standard Error of the Model

Sampling distribution of $\hat{\beta}_1$

Under the "strong" assumptions of the Model:

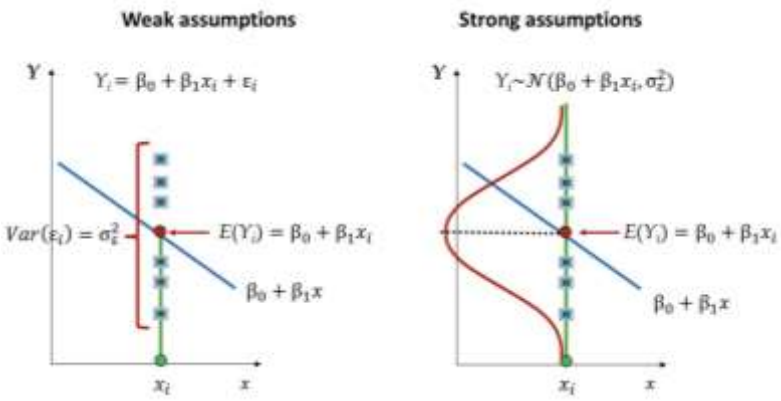
$\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \Rightarrow Y_i \sim N(\beta_0 + \beta_1 \cdot x_i, \sigma_\varepsilon^2)$

$E(\hat{\beta}_1) = \beta_1$ (unbiased)

$Var(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 = \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \underline{x})^2} = \frac{\sigma_\varepsilon^2}{(n-1) \cdot S_X^2}$

⇩

$$SE(\hat{\beta}_1) = \sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma_\varepsilon^2}{(n-1) \cdot S_X^2}}$$



Given the strong assumption of the Model:

1) $\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0,1)$ ⇒ This is true

because $\hat{\beta}_1$ is a linear transformation of Y_i

Since $Var(\varepsilon_i) = \sigma_\varepsilon^2$ is unknown, also $SE(\hat{\beta}_1)$ is not known, so we must estimate it.

Estimator: $SE(\hat{\beta}_1) = S_{\hat{\beta}_1} = \sqrt{\frac{\sigma_\varepsilon^2}{(n-1) \cdot S_X^2}}$ (unbiased)

↳ Estimate: $se(\hat{\beta}_1) = s_{\hat{\beta}_1} = \sqrt{\frac{S_\varepsilon^2}{(n-1) \cdot S_X^2}}$ Reminder: $S_\varepsilon^2 = \frac{SSE}{n-2}$

Eventually we have:

2) $\frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t_{n-2}$

Inference of $\hat{\beta}_1$

CONFIDENCE INTERVAL FOR β_1

$$ci_{1-\alpha}(\beta_1) = b_1 \pm t_{n-2, \frac{\alpha}{2}} \cdot SE(\hat{\beta}_1)$$

TEST ON β_1

➢ Hypotheses definition:

UTT	LTT	TTT
H_0: $\beta_1 = \beta_1^*$ (or $\beta_1 \leq \beta_1^*$) H_1: $\beta_1 > \beta_1^*$	H_0: $\beta_1 = \beta_1^*$ (or $\beta_1 \geq \beta_1^*$) H_1: $\beta_1 < \beta_1^*$	H_0: $\beta_1 = \beta_1^*$ H_1: $\beta_1 \neq \beta_1^*$

➢ Identify test statistic and its distribution (under H_0)

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

➢ Decision rules:

↳ $t = \frac{b_1 - \beta_1^*}{s_\varepsilon \sqrt{\frac{1}{(n-1)\cdot s_X^2}}} = \frac{b_1 - \beta_1^*}{se(\hat{\beta}_1)}$

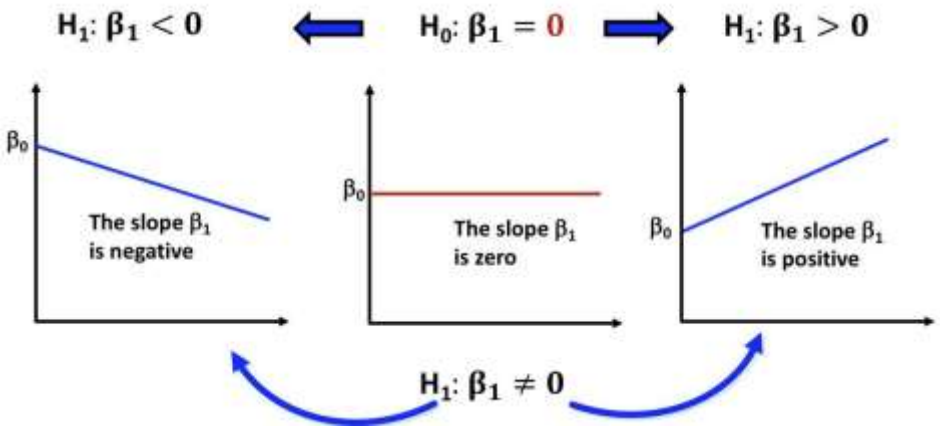
↳ We reject H_0 if

- UTT: $t > t_{n-2,\alpha}$
- LTT: $t < -t_{n-2,\alpha}$
- TTT: $|t| > t_{n-1,\frac{\alpha}{2}}$

or for <u>all the cases</u> we reject H_0 if $P\ value < \alpha$

In R you have the quantity $\frac{b_1}{s_{\hat{\beta}_1}}$ in the output together with the p-value

Typically we set $\beta_1^* = 0$ that means testing if X and Y are linearly independent



Multiple Linear Regression model

We extend the simple linear model by considering the case where the dependent variable is assumed to be associated with several explanatory variables:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon$$

Y = dependent variable (*r.v.*)

$x_1, x_2, \ldots, x_K$ = independent variable (*deterministic, known*)

β_0= intercept of the linear model

$\beta_1, \beta_2, \ldots, \beta_K$= model parameters

ε = error (*r.v.*)

We expect the model to fit the data better, as we are considering **more information to explain the dependent variable**.

In addition, the multiple regression model allows to study the relationship between the dependent variable and more explanatory variables, and to take into account/**control for factors** that possibly influence the dependent variable (Simpson's paradox)

Main aspect to consider when using/interpreting a Multiple Linear Regression model

1) Interpretation of the coefficients

In the simple regression, a change of 1 unit of X corresponds to an **average change in** Y (proportional to the slope of the line).

In the multiple regression, the interpretation is basically the same, but additionally we say that the marginal change of the Y due to a 1 unit increase in **one** of the independent variables must only consider the case where **all the other** independent variables are **FIXED** (i.e. *keeping them constant* or *ceteris paribus*)

2) Explanatory power of the model (goodness of fit)

As in the case of simple regression:

$$R^2 = \frac{SSR}{SST} = \left(1 - \frac{SSE}{SST}\right)$$

In the case of multiple regression, a modification of the R^2 is also considered, the so-called **adjusted R^2** which also takes into account the sample size but above all the **number of explanatory variables in the model**.

$$Adjusted\ R^2 = 1 - \frac{\frac{SSE}{(n-K-1)}}{\frac{SST}{(n-1)}}$$

COMPONENT	D.F.
SSE	$(n-K-1)$
SSR	K
SST	$(n-1)$

Why? If the number of explanatory variables included in the model is very high (compared to the number of cases), the R^2 may be excessively high and not provide a reliable measure

of the model's fit (at the limit, if as many explanatory variables are used as there are cases, the R^2 would be equal to 1 regardless of the quality of the model)

The **adjusted R^2** is used to **compare the fit of models with a different number of explanatory variables**

3) Estimator/estimate of the variance of the model (= of the error)

The variances of the least-squares estimators depend on the **variance of the errors**, σ_ε^2, **which is unknown and must be estimated.**

The **unbiased estimator** of σ_ε^2 in case of a model with *K* explanatory variables (in addition to the sum of the squared errors divided by $(n-K-1)$ (degrees of freedom), and the corresponding estimate is:

$$s_\varepsilon^2 = \frac{SSE}{(n-K-1)} \rightarrow \text{simple regression: } s_\varepsilon^2 = \frac{SSE}{(n-2)} \; K = 1$$

Which is an estimate of the variance of the errors taking into account the number of observations and on the **number of estimated parameters on which the model is based** (which are the *K* coefficients for the explanatory variables and the intercept, hence *K+ 1*)

The square root of s_ε^2, s_ε , is called **standard error of the model** **or** **standard error of the residuals**

4) Confidence intervals and Hypothesis testing on INDIVIDUAL model coefficients

All inferences about the individual parameters of the model (intercepts and slopes) are based on estimators/statistics with $(n-K-1)$ degrees of freedom

For instance, the $1-\alpha$ confidence interval for any individual parameter of the model β_k is:

$$ci_{1-\alpha}(\beta_k) = \left[b_k \pm t_{n-K-1,\frac{\alpha}{2}} \cdot se(\hat{\beta}_k)\right]$$

While the hypothesis testing procedure for the following null hypothesis:

H_0: $\beta_k = \beta_k^*$ (commonly $\beta_k = 0$)

Against alternative hypotheses such as:

H_1: $\beta_k < \beta_k^*$ or H_1: $\beta_k \neq \beta_k^*$ or H_1: $\beta_k > \beta_k^*$

Is based on the following test statistic (and its distribution under H_0):

$$\frac{\hat{\beta}_k - \beta_k^*}{SE(\hat{\beta}_k)} \sim t_{n-K-1}$$

Decision rules ⇒ We reject H_0 if

- UTT: $t > t_{n-K-1,\alpha}$
- LTT: $t < -t_{n-K-1,,\alpha}$
- TTT: $|t| > t_{n-K,\frac{\alpha}{2}}$

So the rejection region (and the critical values) of the test will be based on the Student's T distribution with $(n-K-1)$ degrees of freedom.

5) Hypothesis testing on all model coefficients (F test)

The so-called ***F*** test is used to check the ***overall significance of a model,*** i.e. whether there is a significant relationship between the dependent variable and the set of **all** explanatory variables.

Specifically, one is interested in assessing whether the model ***includes at least one explanatory variable that is significant for the explanation of the dependent variable***

This problem can be addressed by testing the following hypotheses:

H₀: $\beta_1 = \ldots = \beta_k = 0$

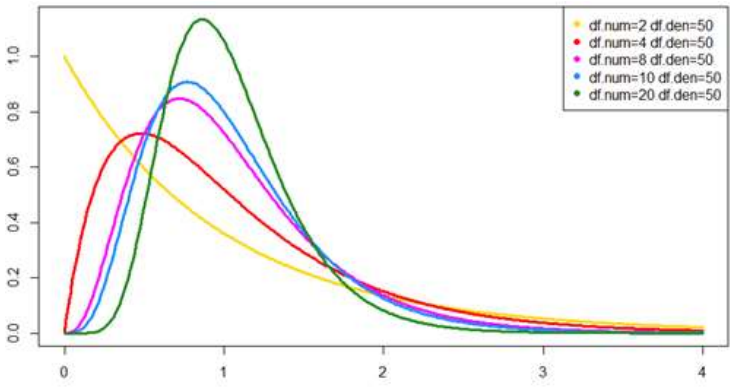
H₁: At least one parameter β_k is $\neq 0$

If the null hypothesis is rejected, it can be concluded that ***at least*** **one** of the coefficients in the model is significantly different from zero and that the model therefore has some validity (subject to checking which coefficient is significant by means of the specific *t-tests* for each explanatory variable).

To test the hypothesis, it is necessary to find a test statistic whose distribution under the null hypothesis is known. In this case, the test statistic is:

$$F = \frac{\frac{SSR}{K}}{\frac{SSE}{(n-K-1)}}$$

which, under H₀: $\beta_1 = \ldots = \beta_k = 0$ has a distribution F with K degrees of freedom at the numerator and $(n-K-1)$ degrees of freedom at the denominator, $F_{K,(n-K-1)}$



A random variable with an F distribution can only assume **non-negative** values. The ***F*-distribution is right skewed,** and depends on two parameters, called degrees of freedom (of the numerator and denominator), which influence its shape.

The null hypothesis will be rejected for very high values of the test statistic:

$$F = \frac{\frac{SSR}{K}}{\frac{SSE}{(n-K-1)}} = \frac{MSR}{MSE}$$

indicating that SSR is much larger than SSE (since SST = SSR + SSE) and that the model therefore includes at least one variable that can explain the dependent variable.

These values are found on the **right** tail of the distribution of ***F*** under H₀

- The rejection region α of a test with significance level is therefore:
  Rejection region: $F > F_{\alpha,K,(n-K-1)}$ **the percentile of order** $(1-\alpha)$ **of a** $F_{K,(n-K-1)}$ **r.v.**

6) Multicollinearity

DEFINITION: the presence of a strong linear correlation between two or more explanatory variables

➢ typically $Cor(X_i, X_j) \geq |0.7|$

The information content of one variable is shared with the information content of one or more other variables: redundant information.

Including in a model two or more highly correlated variables can lead to:

1. **Increase complexity** of the model without an increase of the explanatory power
2. **Difficulty in identifying the marginal effect of a single independent variable** on the dependent variable (because the effect is shared with one or more other independent variables).
3. **"Unstable" estimation of the coefficients**, with high standard errors and low values of t statistics → loss of efficiency of the estimators, higher variability of the estimates!

How can we detect the presence of multicollinearity in a regression?

POSSIBLE EVIDENCES of the presence of a multicollinearity problem in the model:

1. Regression coefficients are very different (e.g. inverse signs), from those that could be expected according to economic hypotheses, logic or experience.
2. The coefficients of the variables that are considered economically/logically relevant are not statistically significant.
3. The model is globally significant (*F TEST*), but all the coefficients are not individually significant (*T TESTS*).
4. A significant variable is no longer significant after another independent variable has been added to the model.
5. Coefficients that drastically change their estimates (even sign inversions) after the inclusion of a new explanatory variable
6. The estimate of the variance of the error increases with the inclusion of a new variable to the model (and also the estimated standard errors of the coefficients increases)

Model Residuals' analysis

GOAL: Before utilizing a linear model for predictions, it is crucial to assess whether the necessary assumptions for making inferences are met. This has to be done for BOTH the simple linear regression model and the multiple linear regression model.

Various procedures and tests can be employed for this purpose, but **we adopt a graphical/qualitative approach**. Here below we describe how to use **regression residuals** ($e_i = (y_i - \hat{y}_i)$) to evaluate key aspects of the model assumptions.

↳ OLS: MIN $\sum e_i^2$

PROCEDURES

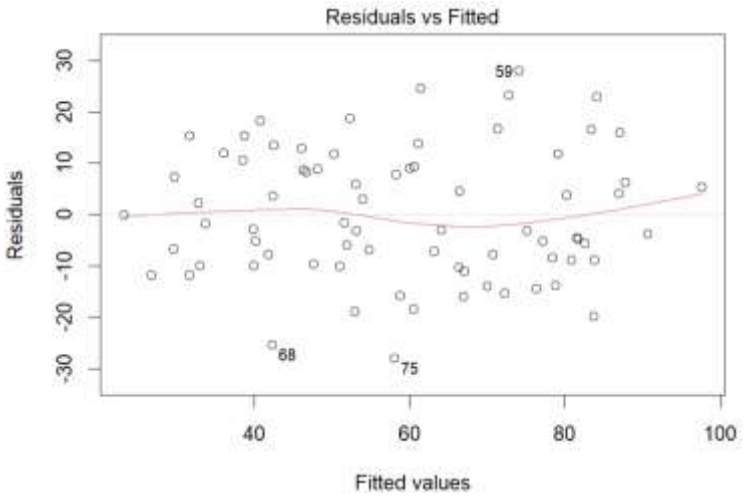
1) Linearity assumption and absence of systematic structural patterns

Assess whether the residuals of the model supports the assumption of linearity and the absence of any systematic structural patterns.

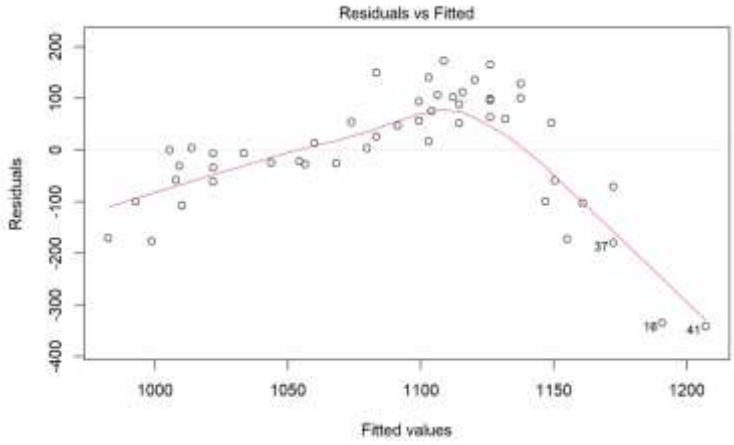
- Limited utility in simple linear regression because we can simply do the scatterplot of Y vs. X to assess non-linearities
- Very insightful for multiple linear regression model

TOOL TO USE: Scatterplot of residuals e_i vs. Predicted (fitted) values $\hat{y}_i$

R CODE: `plot(model,which=1)`



Linearity seems ok here:
At different values of the fitted values, the residuals show an average value that is fairly constant.
The red line in the plot is quite linear, with no big deviation from the horizontal line starting from the origin. It seems we don't have any residual structural pattern here.



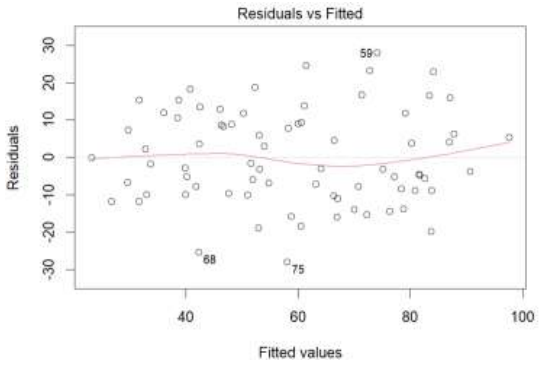
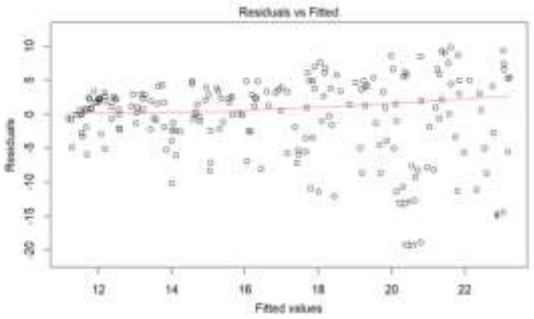
The residuals are characterized by an **obvious non-linear structure**: the average value of the residuals is changing for different values of the fitted Y (that is dependent on the values of the X variables used for the prediction)

2) Constant Variance of Errors (Homoscedasticity):

Verify if the variance of errors remains constant, a critical assumption for reliable model performance. [Assumption! $Var(\varepsilon_i) = \sigma_\varepsilon^2$]

TOOL TO USE: Scatterplot of residuals e_i vs. Predicted (fitted) values $\hat{y}_i$

R CODE: `plot(model,which=1)`

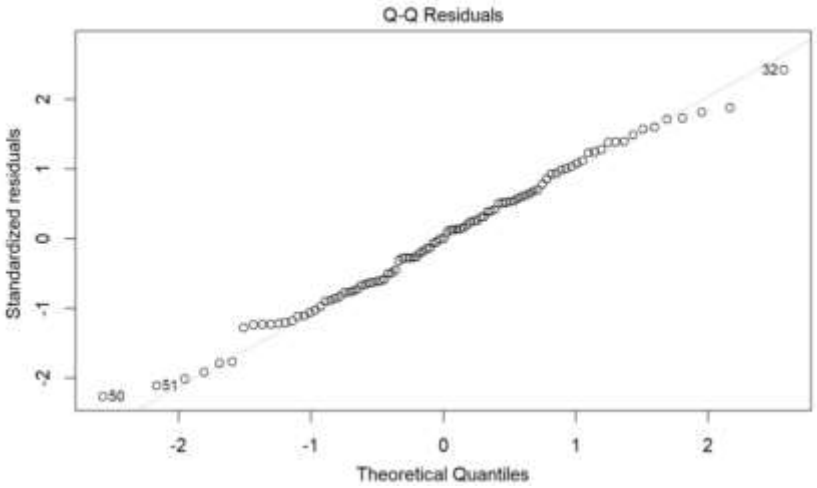
Constant variance evidences → OK!! (Homoskedasticity)	**Non-Constant** variance evidence → NOT OK, possible violation of assumptions. (Heteroskedasticity)
	

3) Normal Distribution of Errors:

Check if errors follow a normal distribution, contributing to the robustness of statistical inference tools (CI and Tests) [$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$]

TOOL TO USE: Q-Q Plot of residuals e_i

R CODE: `plot(model,which=2)`



Although not strictly necessary for inferential purposes as the sample size is sufficiently large, we check whether the residuals have an approximately normal distribution, particularly to assess skewness and possible tails.

The distribution is approximately normal, although some deviation is observed on the tails.

4) Absence of Outliers or Influential observations:

Identify specific cases that disproportionately influence the estimated regression model

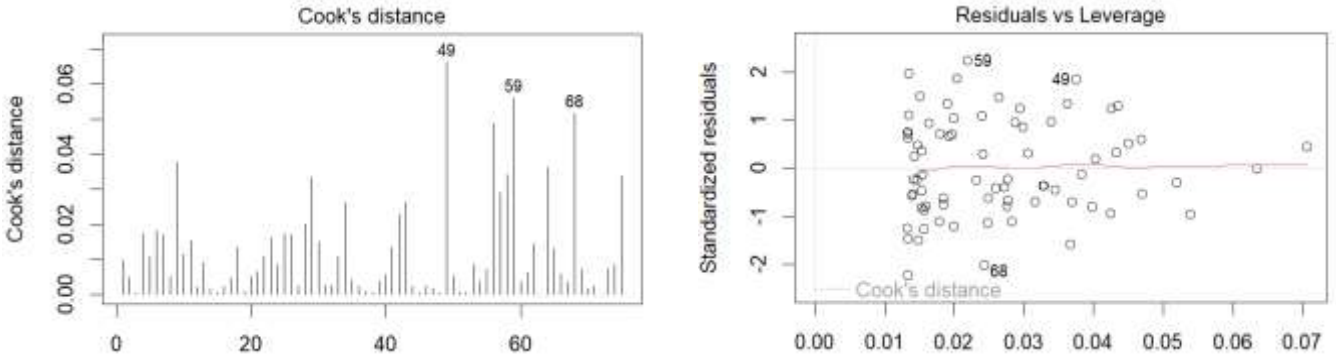
TOOL TO USE: Cook's distance and Leverage plot → just high level qualitative interpretation!

R CODE: `plot(model,which=3)`

R CODE: `plot(model,which=4)`

The model should be valid for all observations, and each observation should have equal weight in determining the model. The characteristics of the reported points should be assessed.

In the following two plots the observations with **high residuals** and also **high leverage** are highlighted (those with a reported number: 49, 59, 68). These observations may be influential on the regression estimates. However, they are few (3 out of a hundred of data points) and not so far away from the other observations.



5) Correlation Between Errors

Investigate potential correlations between errors, particularly relevant in certain applications like the analysis of time series or panel data. $[\ Cor(\varepsilon_i\ ,\varepsilon_j) = 0\ ]$

TOOL TO USE: ACF plot and specific hypothesis testing (not in the syllabus)

FOR DOUBTS OR SUGGESTIONS ON THE HANDOUTS

JOYCE COLING

rina.coling@studbocconi.it

@joyce.coling

+39 3911092533

FOR INFO ON THE TEACHING DIVISION

VITTORIA NASONTE

vittoria.nasonte@studbocconi.it

@_vittorian_

+39 3274441476

ELENA CACIOLI

elena.cacioli@studbocconi.it

@elenacaciolii_

+39 3928931605

TEACHING DIVISION

OUR PARTNERS

700+ CLUB

ETHAN SUSTAINABILITY

DELIVERY VALLEY
NO GENDER KITCHEN

LA PIADINERIA