

Linear regression

Hicham Janati · *INSEA* · hjanati@insea.ac.ma

The Gaussian linear model

The house sales table

house	price (\$)	living area (m ²)	quality (1–10)	year built	lot area (m ²)	dist. to center (km)
1	98 800	86	4	1936	855	12.1
2	148 400	111	4	1940	564	3.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮
400	188 300	120	7	2003	917	8.2

A real-estate agency asks: "**What** makes a house **expensive** ?"

To answer this question we can try to fit two different models:

Model the joint distribution
 $\mathbb{P}(\text{price, area, quality, ...})$



Model the conditional
 $\mathbb{P}(\text{price} \mid \text{area, quality, ...})$

Which model is *richer*, i.e. has more information ?

The goal is to *explain* the price:
Conditional modeling is enough

Response and covariates

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \varepsilon$$



response · target
dependent variable
FR: variable à expliquer
dépendante · cible



covariates · features
explanatory variables · predictors
FR: variables explicatives
covariables · prédicteurs

Stack the n houses, one row each: the **design matrix**

$$\mathbb{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{pmatrix} \in \mathbb{R}^{n \times (d+1)}$$

The Gaussian linear model

Definition: the Gaussian linear model for a response Y given covariates $\mathbf{x}$:

$$Y \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}, \sigma^2)$$

(A1) correct mean: $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$

equivalently $Y = \beta_0 + \boldsymbol{\beta}^\top \mathbf{X} + \varepsilon$,

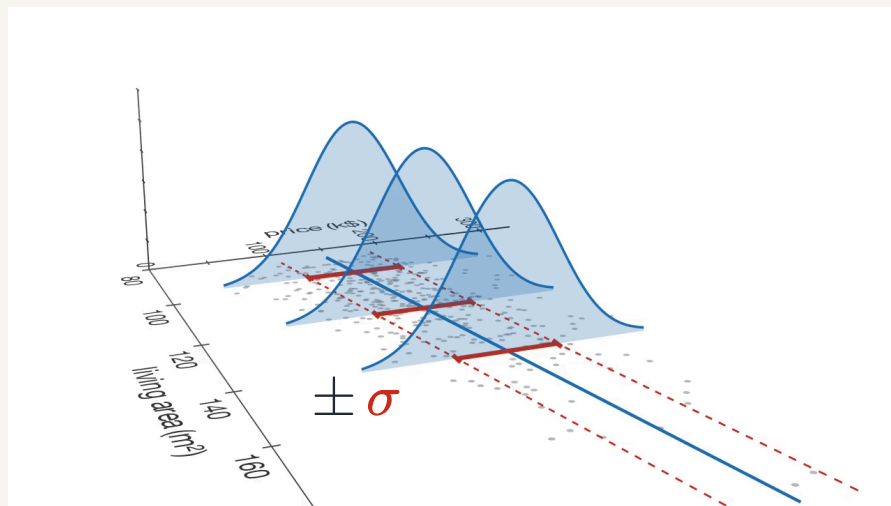
with $\mathbb{E}(\varepsilon \mid \mathbf{X}) = 0$

(A2) i.i.d observations

(A3) homoscedasticity: $\text{Var}(\varepsilon \mid \mathbf{X} = \mathbf{x}) = \sigma^2$

i.e. the **same variance** at every $\mathbf{x}$

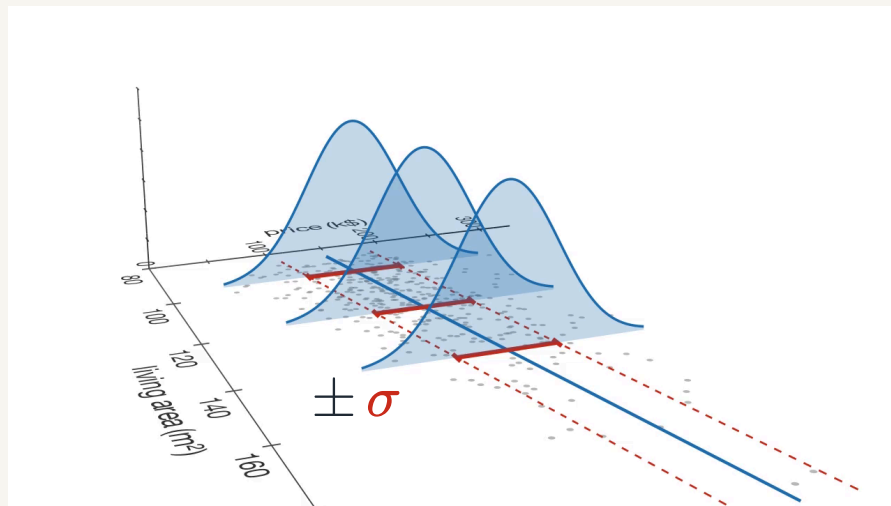
(A4) conditional Gaussianity of the deviations



The Gaussian linear model

Definition: the Gaussian linear model for a response Y given covariates $\mathbf{x}$:

$$Y \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}(\beta_0 + \beta^\top \mathbf{x}, \sigma^2)$$



Theorem (OLS): under (A1)–(A4), with $\mathbf{X}$ the design matrix:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \quad \hat{\beta} \mid \mathbf{X} \sim \mathcal{N}(\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$$

We have a **distribution**: confidence intervals, tests...

Reading the output

The regression output

Read the output to decide **what influences the price**: each line tests $H_0 : \beta_j = 0$.

covariate	estimate (\$)	SE	95% CI	p (t-test)	p (asymptotic)
living area	870	39	[793, 947]	$7.2 \cdot 10^{-72}$	$3.1 \cdot 10^{-110}$
quality	24 000	1 150	[21 740, 26 260]	$1.1 \cdot 10^{-65}$	$1.0 \cdot 10^{-96}$
year built	-2	38	[-76.7, 72.7]	0.96	0.96
lot area	6.7	2.1	[2.6, 10.8]	0.0015	0.0014
dist. to center	-2 400	330	[-3 049, -1 751]	$1.9 \cdot 10^{-12}$	$3.5 \cdot 10^{-13}$

- year built: ?

The regression output

Read the output to decide **what influences the price**: each line tests $H_0 : \beta_j = 0$.

covariate	estimate (\$)	SE	95% CI	p (t-test)	p (asymptotic)
living area	870	39	[793, 947]	$7.2 \cdot 10^{-72}$	$3.1 \cdot 10^{-110}$
quality	24 000	1 150	[21 740, 26 260]	$1.1 \cdot 10^{-65}$	$1.0 \cdot 10^{-96}$
year built	-2	38	[-76.7, 72.7]	0.96	0.96
lot area	6.7	2.1	[2.6, 10.8]	0.0015	0.0014
dist. to center	-2 400	330	[-3 049, -1 751]	$1.9 \cdot 10^{-12}$	$3.5 \cdot 10^{-13}$

- **year built**: *not significant* ($p = 0.96$), the *real* coefficient could be 0.
- **living area**: ?

The regression output

Read the output to decide **what influences the price**: each line tests $H_0 : \beta_j = 0$.

covariate	estimate (\$)	SE	95% CI	p (t-test)	p (asymptotic)
living area	870	39	[793, 947]	$7.2 \cdot 10^{-72}$	$3.1 \cdot 10^{-110}$
quality	24 000	1 150	[21 740, 26 260]	$1.1 \cdot 10^{-65}$	$1.0 \cdot 10^{-96}$
year built	-2	38	[-76.7, 72.7]	0.96	0.96
lot area	6.7	2.1	[2.6, 10.8]	0.0015	0.0014
dist. to center	-2 400	330	[-3 049, -1 751]	$1.9 \cdot 10^{-12}$	$3.5 \cdot 10^{-13}$

- **year built**: *not significant* ($p = 0.96$), the *real* coefficient could be 0.
- **living area**: *significant*, +870 \$ per m².
- **quality**: ?

The regression output

Read the output to decide **what influences the price**: each line tests $H_0 : \beta_j = 0$.

covariate	estimate (\$)	SE	95% CI	p (t-test)	p (asymptotic)
living area	870	39	[793, 947]	$7.2 \cdot 10^{-72}$	$3.1 \cdot 10^{-110}$
quality	24 000	1 150	[21 740, 26 260]	$1.1 \cdot 10^{-65}$	$1.0 \cdot 10^{-96}$
year built	-2	38	[-76.7, 72.7]	0.96	0.96
lot area	6.7	2.1	[2.6, 10.8]	0.0015	0.0014
dist. to center	-2 400	330	[-3 049, -1 751]	$1.9 \cdot 10^{-12}$	$3.5 \cdot 10^{-13}$

- **year built**: *not significant* ($p = 0.96$), the *real* coefficient could be 0.
- **living area**: *significant*, +870 \$ per m².
- **quality**: *significant*, +24 000 \$ per quality point.

All else being equal ?

$$\mathbb{E}(\textcolor{teal}{Y} \mid \mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$$

"+870 \$ per m², all else being equal" is a **partial derivative**:

$$\frac{\partial \mathbb{E}(\textcolor{teal}{Y} \mid \mathbf{x})}{\partial x_j} = \beta_j : \quad \text{move } x_j, \text{ keep every other covariate } x_i \ (i \neq j) \text{ **fixed**}$$

Can we always assume “everything else stays fixed” ?

When two covariates $\mathbf{X}_i, \mathbf{X}_j$ are strongly **correlated**: if one moves, the other moves too.

What happens to the OLS formula $\hat{\beta} = (\textcolor{violet}{X}^\top \textcolor{violet}{X})^{-1} \textcolor{violet}{X}^\top \textcolor{teal}{y}$ when columns of $\textcolor{violet}{X}$ are **highly correlated** ?

When assumptions fall

covariate	estimate (\$)	SE	95% CI	p (t-test)	p (asymptotic)
living area	870	39	[793, 947]	$7.2 \cdot 10^{-72}$	$3.1 \cdot 10^{-110}$
quality	24 000	1 150	[21 740, 26 260]	$1.1 \cdot 10^{-65}$	$1.0 \cdot 10^{-96}$
year built	-2	38	[-76.7, 72.7]	0.96	0.96
lot area	6.7	2.1	[2.6, 10.8]	0.0015	0.0014
dist. to center	-2 400	330	[-3 049, -1 751]	$1.9 \cdot 10^{-12}$	$3.5 \cdot 10^{-13}$

Drop (A4) Gaussianity: what happens ?

$$T_j = \hat{\beta}_j / \widehat{\text{SE}}(\hat{\beta}_j) \not\sim t(n - d - 1)$$

When assumptions fall

covariate	estimate (\$)	SE	≈ 95% CI	≈ p (t-test)	p (asymptotic)
living area	870	39	[793, 947]	$7.2 \cdot 10^{-72}$	$3.1 \cdot 10^{-110}$
quality	24 000	1 150	[21 740, 26 260]	$1.1 \cdot 10^{-65}$	$1.0 \cdot 10^{-96}$
year built	-2	38	[-76.7, 72.7]	0.96	0.96
lot area	6.7	2.1	[2.6, 10.8]	0.0015	0.0014
dist. to center	-2 400	330	[-3 049, -1 751]	$1.9 \cdot 10^{-12}$	$3.5 \cdot 10^{-13}$

Drop (A4) Gaussianity: what happens ?

$$T_j = \hat{\beta}_j / \widehat{\text{SE}}(\hat{\beta}_j) \not\sim t(n - d - 1) \quad T_j \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 1)$$

When assumptions fall

covariate	estimate (\$)	SE	≈ 95% CI	≈ p (t-test)	p (asymptotic)
living area	870	39	[793, 947]	$7.2 \cdot 10^{-72}$	$3.1 \cdot 10^{-110}$
quality	24 000	1 150	[21 740, 26 260]	$1.1 \cdot 10^{-65}$	$1.0 \cdot 10^{-96}$
year built	-2	38	[-76.7, 72.7]	0.96	0.96
lot area	6.7	2.1	[2.6, 10.8]	0.0015	0.0014
dist. to center	-2 400	330	[-3 049, -1 751]	$1.9 \cdot 10^{-12}$	$3.5 \cdot 10^{-13}$

Drop (A4) Gaussianity: what happens ?

$$T_j = \hat{\beta}_j / \widehat{\text{SE}}(\hat{\beta}_j) \not\sim t(n - d - 1) \quad T_j \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 1)$$

The asymptotic p-values and confidence intervals stay valid for large n : the **central limit theorem**.

When assumptions fall

covariate	estimate (\$)	SE	≈ 95% CI	≈ p (t-test)	p (asymptotic)
living area	870	39	[793, 947]	$7.2 \cdot 10^{-72}$	$3.1 \cdot 10^{-110}$
quality	24 000	1 150	[21 740, 26 260]	$1.1 \cdot 10^{-65}$	$1.0 \cdot 10^{-96}$
year built	-2	38	[-76.7, 72.7]	0.96	0.96
lot area	6.7	2.1	[2.6, 10.8]	0.0015	0.0014
dist. to center	-2 400	330	[-3 049, -1 751]	$1.9 \cdot 10^{-12}$	$3.5 \cdot 10^{-13}$

Drop (A3), homoscedasticity: what happens ?

$$\widehat{\text{Var}}(\hat{\beta} \mid \mathbf{X}) \neq \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

When assumptions fall

covariate	estimate (\$)	SE	95% CI	p (t-test)	p (asymptotic)
living area	870	39	{793, 947}	7.2 · 10⁻⁷²	3.1 · 10⁻¹¹⁰
quality	24 000	1150	{21 740, 26 260}	1.1 · 10⁻⁶⁵	1.0 · 10⁻⁹⁶
year built	-2	38	{-76.7, 72.7}	0.96	0.96
lot area	6.7	2.1	{2.6, 10.8}	0.0015	0.0014
dist. to center	-2 400	330	{-3 049, -1 751}	1.9 · 10⁻¹²	3.5 · 10⁻¹³

Drop (A3), homoscedasticity: what happens ?

$$\widehat{\text{Var}}(\hat{\beta} \mid \mathbf{X}) \neq \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

Fix 1: a new variance estimator, robust to heteroscedasticity (HC, see textbook).

Fix 2: the empirical **bootstrap** estimator (later in these slides).

When assumptions fall

covariate	estimate (\$)	SE	95% CI	p (t-test)	p (asymptotic)
living area	870	39	{793, 947}	$7.2 \cdot 10^{-72}$	$3.1 \cdot 10^{-110}$
quality	24 000	1150	{21 740, 26 260}	$1.1 \cdot 10^{-65}$	$1.0 \cdot 10^{-96}$
year built	-2	38	{-76.7, 72.7}	0.96	0.96
lot area	6.7	2.1	{2.6, 10.8}	0.0015	0.0014
dist. to center	-2 400	330	{-3 049, -1 751}	$1.9 \cdot 10^{-12}$	$3.5 \cdot 10^{-13}$

Drop (A2), independence: what happens ?

Dependent rows lead to incorrect standard errors.

Fix 1: a new variance estimator, robust to heteroscedasticity (HC, see textbook).

Fix 2: the empirical **bootstrap** estimator (later in these slides).

When assumptions fall

covariate	estimate (\$)	SE	95% CI	p (t-test)	p (asymptotic)
living area	870	39	{793, 947}	$7.2 \cdot 10^{-72}$	$3.1 \cdot 10^{-110}$
quality	24 000	1150	{21740, 26260}	$1.1 \cdot 10^{-65}$	$1.0 \cdot 10^{-96}$
year built	-2	38	{-76.7, 72.7}	0.96	0.96
lot area	6.7	2.1	{2.6, 10.8}	0.0015	0.0014
dist. to center	-2 400	330	{-3 049, -1751}	$1.9 \cdot 10^{-12}$	$3.5 \cdot 10^{-13}$

Drop (A1), conditional linear mean: what happens ?

When assumptions fall

covariate	estimate (\$)	SE	95% CI	p (t-test)	p (asymptotic)
living area	870	39	{793, 947}	$7.2 \cdot 10^{-72}$	$3.1 \cdot 10^{-110}$
quality	24 000	1150	{21 740, 26 260}	$1.1 \cdot 10^{-65}$	$1.0 \cdot 10^{-96}$
year built	-2	38	{-76.7, 72.7}	0.96	0.96
lot area	6.7	2.1	{2.6, 10.8}	0.0015	0.0014
dist. to center	-2 400	330	{-3 049, -1 751}	$1.9 \cdot 10^{-12}$	$3.5 \cdot 10^{-13}$

Drop (A1), conditional linear mean: what happens ?

Biased coefficients: nothing in the table is reliable anymore.

Assumptions are not verifiable

How can we make sure that our data verify (A1)-(A4) ?



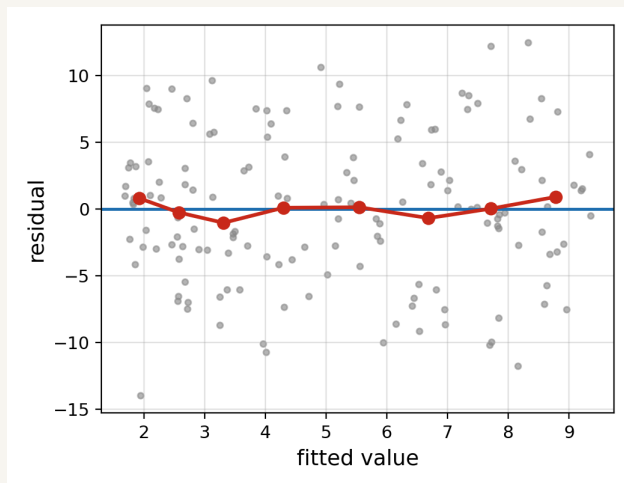
We can't ! All we can do is gather **red flags**: diagnostics that can *detect* problems but never *certify* that A1-A4 are true.

Diagnostics: evidence, not proof

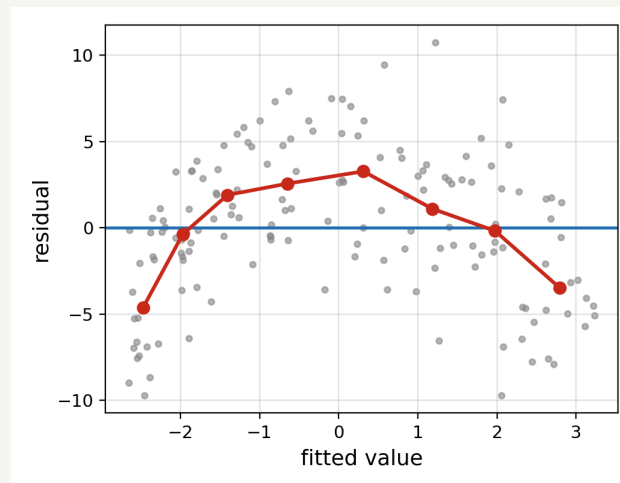
Reading the residuals: the mean

(A1) implies $\mathbb{E}(\varepsilon \mid \mathbf{X}\boldsymbol{\beta}) = 0$: the noise averages to 0 at every value of the mean.

$$\mathbf{x}_i^\top \boldsymbol{\beta} \approx \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} \quad (\text{fitted value } \hat{y}_i), \quad \varepsilon_i \approx r_i = y_i - \hat{y}_i \quad (\text{residual})$$



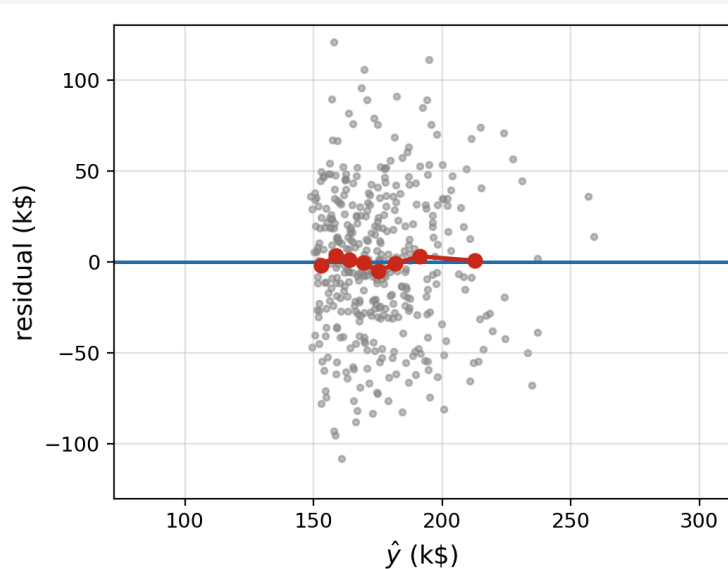
Symmetric around 0: no red flag



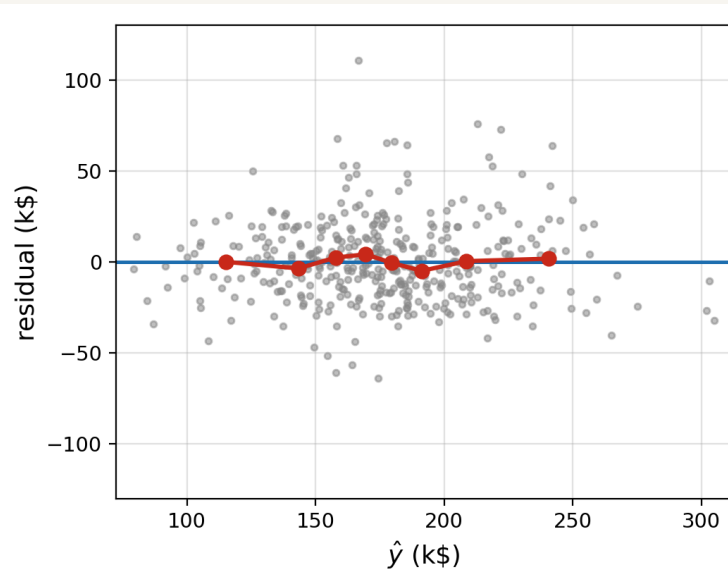
Non-zero mean: **(A1)** contradicted

A clean plot proves nothing

price ~ area (residual sd 39.7 k\$)



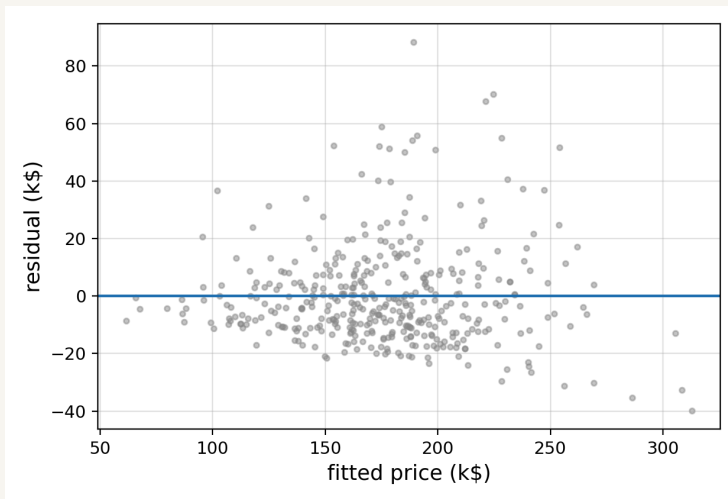
price ~ area + **quality** (residual sd 22.5 k\$)



A clean residual plot is **not sufficient** to conclude (A1): an omitted variable can hide perfectly.

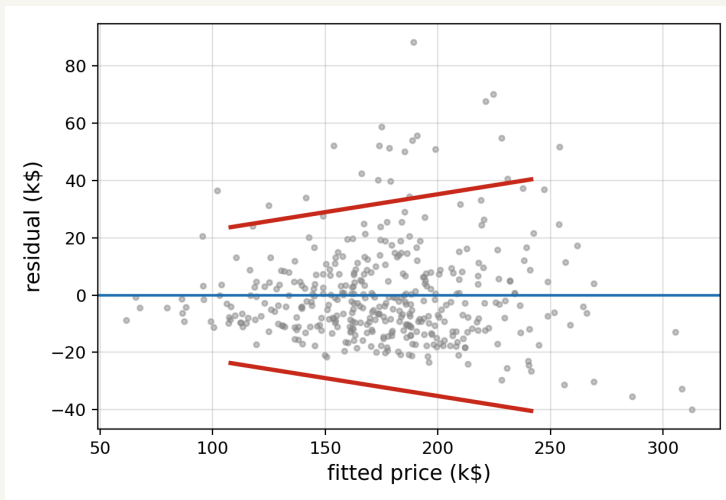
Reading the residuals: the spread

(A1 + A3) imply $\text{Var}(\varepsilon \mid \mathbf{X}\beta) = \sigma^2$, constant everywhere



Reading the residuals: the spread

(A1 + A3) imply $\text{Var}(\varepsilon \mid \mathbf{X}\beta) = \sigma^2$, constant everywhere



a funnel
(entonnoir) →



A *funnel*: the spread grows with the fitted price, **(A3) contradicted**.

Checking Gaussianity (A4)

If **(A4)** holds, the standardized residuals $r_1, \dots, r_n$ behave like draws from $\mathcal{N}(0, 1)$.

How can we **detect** non-Gaussianity ?

Exercise 1

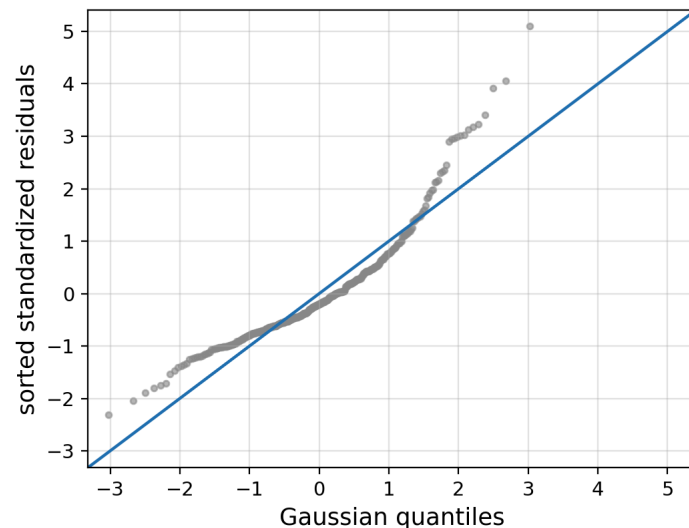
Sorted residuals

Assume $r_1, \dots, r_n \sim \mathcal{N}(0, 1)$ i.i.d and denote F the Gaussian CDF (*Cumulative distribution function*).

1. Define the empirical CDF (fonction de répartition empirique) F_n of $r_1, \dots, r_n$.
2. Sort them: $r_{(1)} < \dots < r_{(n)}$. What is $F_n(r_{(i)})$?
3. Admit the Glivenko–Cantelli theorem: $\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \rightarrow 0$ almost surely. What is the limit of $F_n(r_{(i)})$ when $n \rightarrow +\infty$? Deduce what $r_{(i)}$ should be close to with large n .

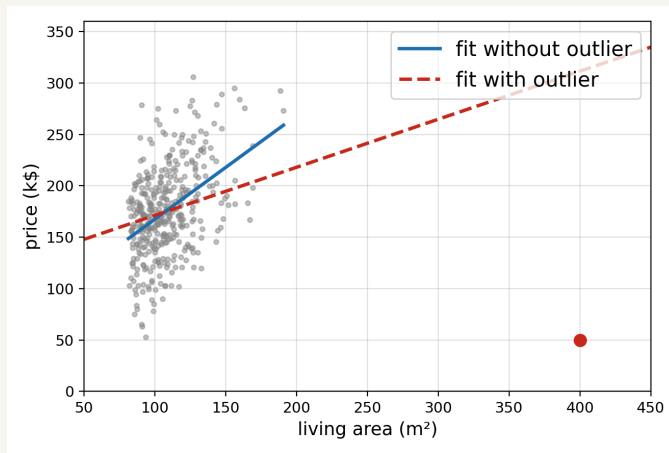
Diagnosing Gaussian errors (A4)

Gaussian residuals: the points $(\Phi^{-1}(\frac{i}{n}), r_{(i)})$ lie close to the line $y = x$. This is the **QQ-plot**.



How do you *read* this QQ plot ?

Effect of an extreme outlier



One sale was entered as **400 m²** of living area for **50 k\$**. How much do you expect the fitted line to change ?



A single outlier drags the slope from 1 007 \$/m² (blue) to 90 \$/m² (red).

Here we had one covariate, one scatter plot. With **many covariates**, how can we detect it ?

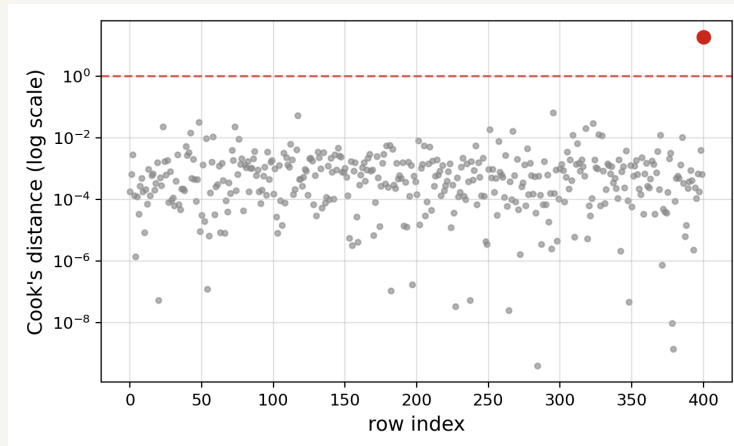
Cook's distance

Cook's distance is a score for each observation: how far do the predictions $\hat{y}$ move when row i is deleted ?

$\hat{\beta}$: OLS on all n rows

$\hat{\beta}_{-i}$: OLS without row i

$$D_i = \frac{\|\mathbb{X}\hat{\beta} - \mathbb{X}\hat{\beta}_{-i}\|^2}{(d+1)\hat{\sigma}^2}$$



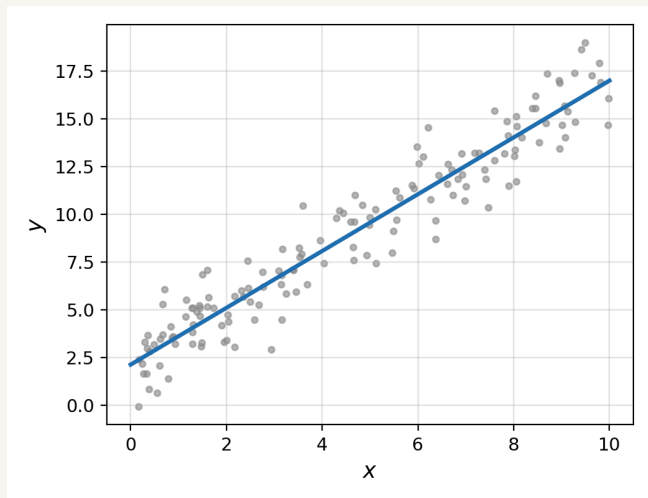
Practical rule: $D_i > 1$ flags observation i as having **too much influence** on the fit: inspect it individually.

Evaluating the model

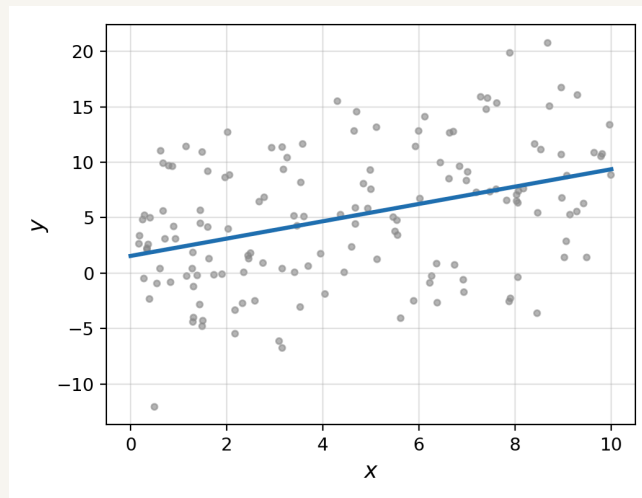
R^2 , recapped

$$R^2 = 1 - \frac{\text{RSS}}{\sum_i (y_i - \bar{y})^2}$$

the share of the variance of Y the fitted mean absorbs



$R^2 = 0.92$: most of the variance absorbed

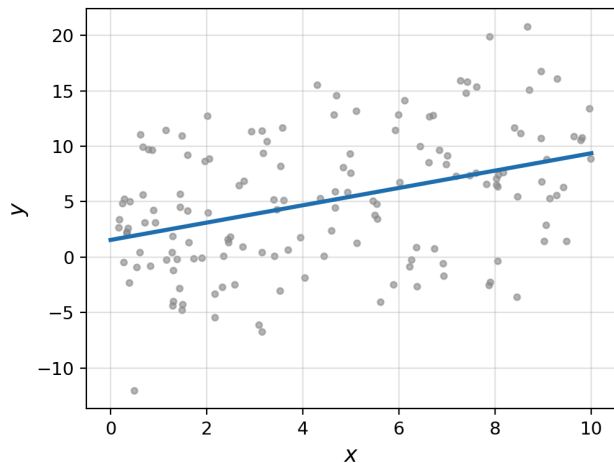


$R^2 = 0.15$: same line, noisier response

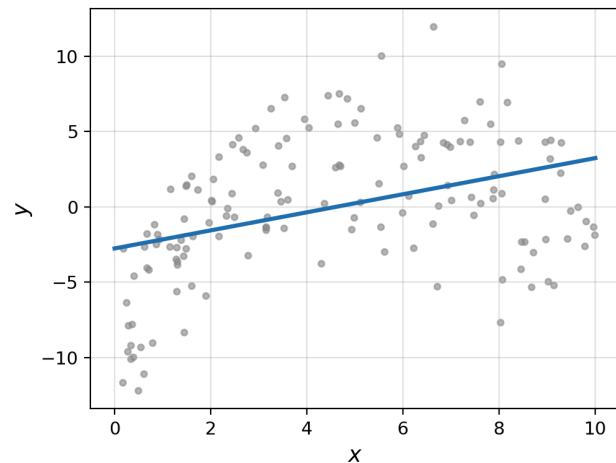
A **valid** model can have a low R^2 : inference stays exact, the intervals are just wider.

Why is R^2 low ?

Two clouds, the **same** low $R^2 = 0.15$. Same diagnosis ?



$R^2 = 0.15$: noisy data, linear model is *fine*

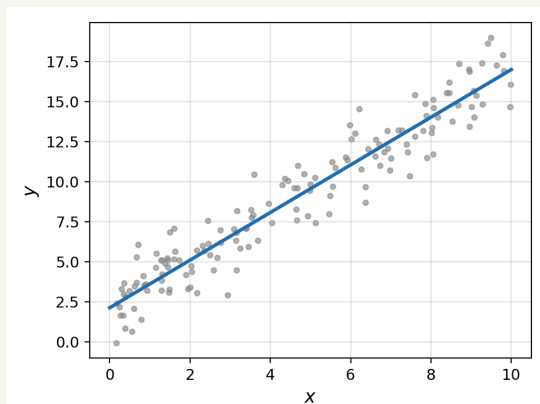


$R^2 = 0.15$: non-linear, model wrong

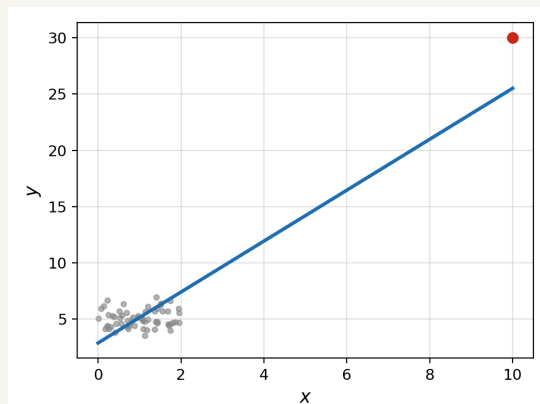
R^2 alone cannot tell which: the **residual plots** of the diagnostics section can (maybe).

Is a high R^2 always good ?

Two fits, both with a high R^2 . Equally trustworthy ?



$R^2 = 0.92$: a real linear relationship



$R^2 = 0.78$: one far point manufactures the fit

High R^2 can be good, or **misleading**: check influence (Cook's distance) and the residuals.

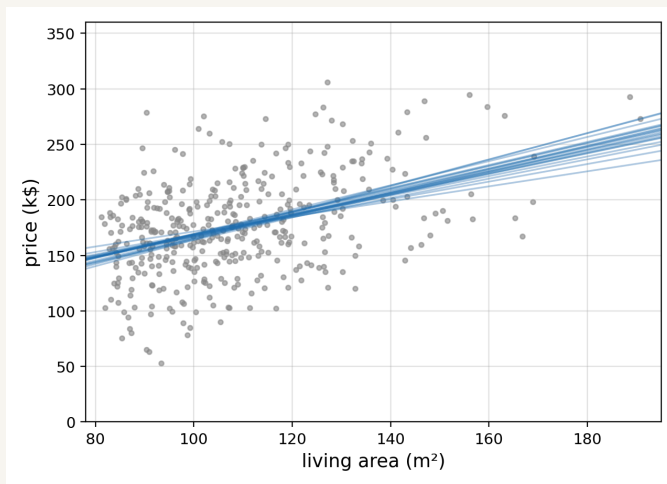
The bootstrap

Seeing the sampling distribution

A3-A4 do not hold: can we still estimate uncertainty around $\hat{\beta}$?



Idea: compute many estimators $\hat{\beta}^{*1}, \dots, \hat{\beta}^{*B}$ on **bootstrap** samples:



the data: n houses

sample b : n rows drawn **with replacement** (bigger dot: drawn more than once)

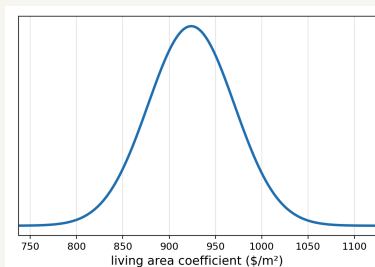
its fit: $\hat{\beta}^{*b}$

$b = 1, 2, \dots, B$

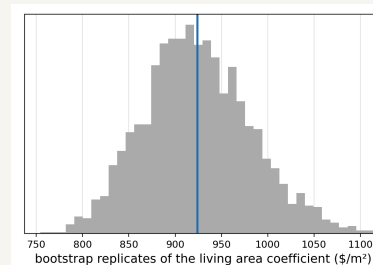
Analytical law, empirical law

Under (A1)–(A4): one analytical law:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$$



Bootstrap: B refitted "models" $\hat{\beta}^{*1}, \dots, \hat{\beta}^{*B}$:
an *empirical* law, no formula.



$$\bar{\beta}_j^* = \frac{1}{B} \sum_b \hat{\beta}_j^{*b}$$

$$\widehat{\text{SE}}_{\text{boot}}(\hat{\beta}_j) = \sqrt{\frac{1}{B-1} \sum_b (\hat{\beta}_j^{*b} - \bar{\beta}_j^*)^2}$$

Empirical mean and std over the replicates;
percentile CI: the empirical 2.5% / 97.5%
quantiles.

Contents

1. The Gaussian linear model
2. Reading the output
3. Diagnostics: evidence, not proof
4. Evaluating the model
5. The bootstrap
6. What to remember