Lesson 1: Simple Linear Regression

Zacharie Naulet
zacharie.naulet@universite-paris-saclay.fr

Contents

Roadmap and informations

- 7 séances avec cours + TP
- Programme du cours (4 chapitres de cours + 8 séances de TP):
  1. Régression linéaire simple (Séances 1, 2).
  2. Régression linéaire multiple (Séances 2 à 5).
  3. Sélection de variables (Séances 5 et 6).
  4. Grande dimension et tests multiples (Séance 7).
- Modalités (théoriques) d'évaluation des connaissances:
  - Contrôle Continu (Séance 3), coeff 0.2.
  - Compte rendu TP (séance 5 ou 6), coeff 0.3.
  - Examen de 1ère session, coeff 0.5.
- Références:
  - Régression avec R, Cornillon et Matzner-Lober (2011).
- Acknowledgments: Most of the material here originates from Sylvain Arlot's course. I am only responsible for typesetting the notes.

1 Introduction and motivations

Goal: Explain the variations of a quantitative variable Y from the measurement of another quantitative variable X.

Example 1 (CML Section 1.1)**.** *Historical motivation, Sir Francis Galton (1885). While working on heredity, his goal was to explain the heights of the sons (Y) in term of the heights of the fathers (X). He noticed that when the father was taller than the average "taller than mediocrity", then his son was likely to be smaller than him, and conversely, if the father was smaller than the average "shorter than mediocrity", then his son was likely to be shorter than him. This led him to consider his theory of "regression toward mediocrity".*

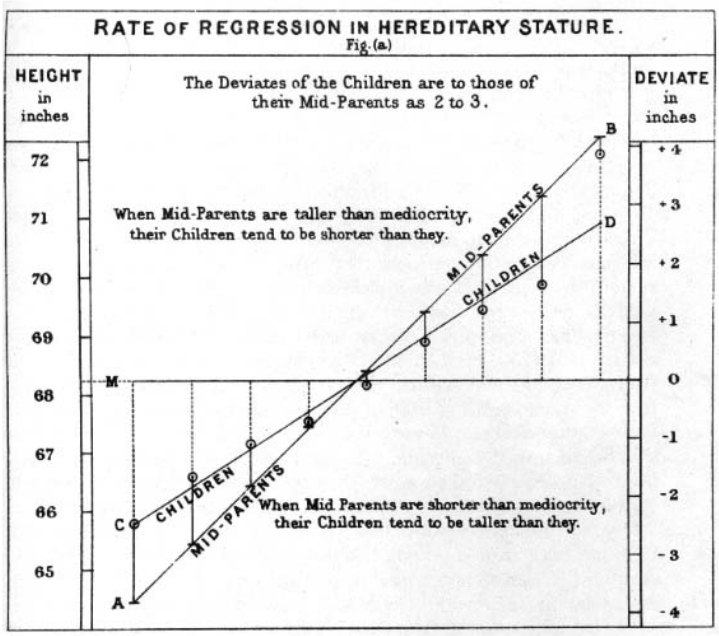


Figure 8.8. Galton's graphical illustration of regression; the circles give the average heights for groups of children whose midparental heights can be read from the line AB. The difference between the line CD (drawn by eye to approximate the circles) and AB represents regression toward mediocrity. (From Galton, 1886a.)

Example 2 (Running example for this course)**.** *Air pollution, see exercices. The goal is to understand the ozone (O_3) concentration in term of the temperature. This is important to predict, in particular for fragile persons.*

- *In a first time we record the maximum concentration in O_3 each days and the temperature T, say at noon. Remark that we know (elementary chemistry) that the O_3 concentration increases as T increases.*
- *Then, we want to predict in advance the O_3 concentration in the next days from the temperature (which can be predicted, see Météo-France).*

We refer to CML for other examples and details. In particular:

- Blood pressure as a function of age;
- Wheat crop yield (*i.e.* agricultural output) as a function of fertilizer quantity;
- Treatment effect as a function of dose;
- etc..

2 Mathematical formalism

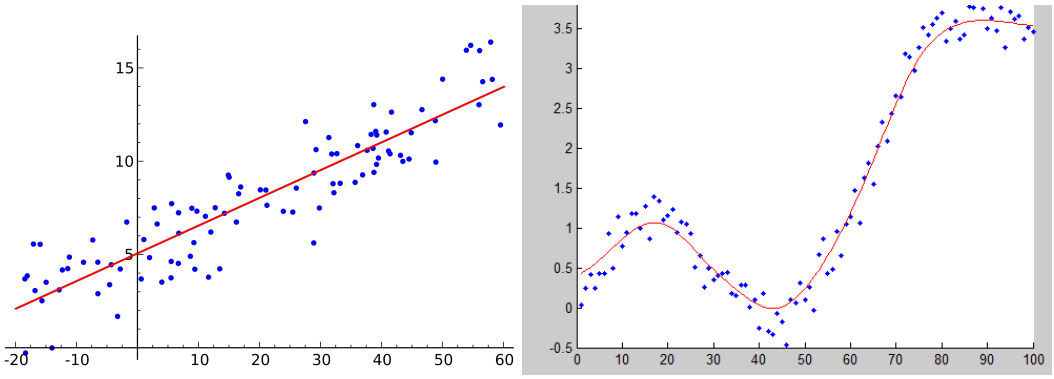
2.1 Principle

- We have observations $(x_1, y_1), \ldots, (x_n, y_n)$ with $(x_i, y_i) \in \mathbb{R}^2$.

- y_i is the **response variable** (what we want to explain); This is assumed random.
- x_i is the **explanatory variable** (aka covariate, predictor); May be random or not random.

The role of x_i and y_i is not symmetric!!

- We are seeking for a function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x_i) \approx y_i$.
- We assume a linear relation between x_i and y_i (indeed affine), *i.e.*
  - $f(x) = \beta_1 + \beta_2 x$, $\beta_1, \beta_2 \in \mathbb{R}$.
  - β_1 and β_2 to be estimated from the data.
  - β_1 is called **intercept**.
  - β_2 is called **slope**.
- **First things first:** Always plot the data to see if a linear model is relevant (remark that later on we will find diagnostics to see if it is relevant to consider linear regression).

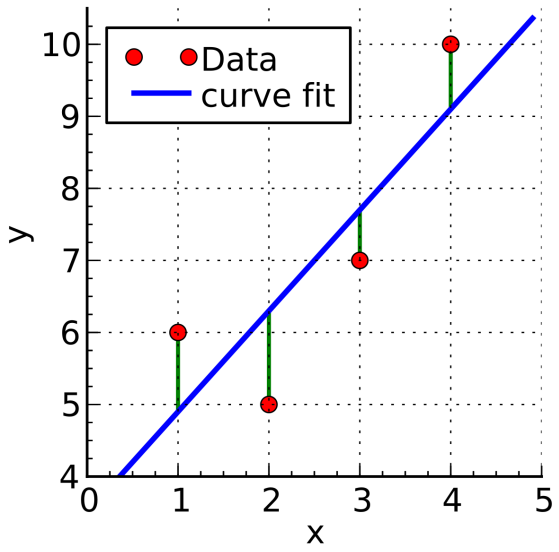


2.2 How to find the regression function

- General principle: Find the function that minimizes the error (with respect to some cost function) on the data [Legendre (1805), Gauss (1795-1809)].
- Here, we are interested in **Least Squares estimation**, where the cost function is $x \mapsto x^2$ (quadratic risk).

Definition 1. *Let $\mathcal{F}$ be the set of all affine functions, i.e. $\mathcal{F} := \{f : f(x) = \beta_1 + \beta_2 x,\ \beta_1, \beta_2 \in \mathbb{R}\}$. Then, a **Least Squares** estimator of the regression function f based on the observation of $(x_1, y_1), \ldots, (x_n, y_n)$ is any*

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (y_i - f(x_i))^2; \tag{1}$$



- Note that we are not restricted to affine relations.
  - We can model quadratic relations: $x_i \mapsto x_i^2$, *i.e.* the model is $y_i = \beta_1 + \beta_2 x_i^2$;
  - logarithmic relations: $\log(y_i) = \beta_1 + \beta_2 \log(x_i)$;
  - etc...

  But in every case, we need a linear relation between the parameters (β_1 and β_2), and we need to know the relation ahead of time.

- The quadratic risk is not the only option.
  - ℓ_1 risk: find the $f \in \mathcal{F}$ that minimizes $\sum_{i=1}^n |y_i - f(x_i)|$;
  - ℓ_∞ risk: find the $f \in \mathcal{F}$ that minimizes $\max_{i=1,\dots,n} |y_i - f(x_i)|$;
  - etc...

  In this course we will always and only consider the quadratic risk and the least squares estimator of equation (1).

2.3 What do we do with our estimator

- For simplicity, we can rewrite equation (1) as

$$(\hat{\beta}_1, \hat{\beta}_2) \in \arg\min_{(\beta_1,\beta_2)\in\mathbb{R}^2} \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2. \tag{2}$$

- Given our estimator $(\hat{\beta}_1, \hat{\beta}_2)$, we can
  - Model a relation between x and y, *i.e.* estimate the "true" β_1 by $\hat{\beta}_1$ and the "true" β_2 by $\hat{\beta}_2$ + confidence intervals (to understand some phenomenon).

- Quantify the relevance of the model (is there a relationship between x and y ? Is the linear model enough ?)
- Do prediction: given that we observed $(x_1, y_1), \dots, (x_n, y_n)$ and we know x_{n+1}, what can we predict about y_{n+1} ? (see for instance air pollution)
  * Propose $\hat{y}^p_{n+1} = \hat{\beta}_1 + x_{n+1}\hat{\beta}_2$ (we want to have $\mathbb{E}[(\hat{y}^p_{n+1} - y_{n+1})^2]$ as small as possible);
  * Propose $IP(x_{n+1})$ a prediction interval (we want $y_{n+1} \in IP(x_{n+1})$ with probability $\geq 1 - \alpha$).

2.4 Statistical model

- Formally, a statistical model is a set of probability distributions that are good "candidates" to explain our observations $(x_1, y_1), \dots, (x_n, y_n)$. Each probability distribution corresponds to a parameter, here (β_1, β_2), and we aim at finding the parameter(s) that explain the best our data.

- The **standard model** for simple linear regression is

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad i = 1, \dots, n. \tag{3}$$

  Here:

  - $(\beta_1, \beta_2) \in \mathbb{R}^2$ is the parameter of the model (what we want to estimate from the observations). This is **unknown**;
  - $x_1, \dots, x_n \in \mathbb{R}$ are the **explanatory variables**. In this course, they are assumed to be non random. This is **known**;
  - $Y_1, \dots, Y_n \in \mathbb{R}$ are the **response variables**. They are random and **known** (this is what we observe).
  - $\varepsilon_1, \dots, \varepsilon_n$ is the **noise**, see below. Here, it is assumed that $\varepsilon_1 \perp \dots \perp \varepsilon_n$ and that the all have the same distribution (*i.i.d.*). These are **unknown**.

- Note that in this course we assume here that $x_1, \dots, x_n$ are **deterministic**, *i.e.* not random. It is sometimes the case that x_i's are random, that $x_1, \dots, x_n$ are the realizations of some random variables $X_1, \dots, X_n$. If it is the case, we can always reduce the problem to the situation where $x_1, \dots, x_n$ are fixed by working conditional on $X_1, \dots, X_n$.

- One of the goal is to find estimates of (β_1, β_2) based upon the observation of $Y_1 = y_1, \dots, Y_n = y_n$ and the knowledge of $x_1, \dots, x_n$.

- Regarding the **noise** $\varepsilon_1, \dots, \varepsilon_n$, we will do the following classical assumptions:

  - They are not depending on $x_1, \dots, x_n$ (homoskedasticity);
  - $\varepsilon_1, \dots, \varepsilon_n$ are *i.i.d.*
  - The (common) distribution of the ε_i's is not necessarily known, but:
    * They are centered, *i.e.* $\mathbb{E}[\varepsilon_i] = 0$;
    * They have a finite variance (known or unknown), *i.e.* $\mathbb{E}[\varepsilon_i^2] := \sigma^2 < \infty$;

- Interpretation of the noise:

- Contain the "missing" information in x_i (and the intercept) to explain y_i (*e.g.* the O_3 concentration does not only depend on the temperature, but also on other less important factors, such that the weather, other pollutants in the atmosphere, etc., for which we don't necessarily have measurements and we can thus model by noise)
- Eventually, measurement errors (imperfect material, imperfect user).

3 Point estimation using least squares

3.1 Definition

- Refer to CML Section 1.4.
- Until now, we implicitly defined the least squares estimator, but we never said if it is unique, or even if it exists, and even less how to compute it.
- Define the **Least Squares criterion** $S : \mathbb{R}^2 \to \mathbb{R}_+$

$$S(\beta_1, \beta_2) \coloneqq \sum_{i=1}^{n} \left(y_i - \beta_1 - \beta_2 x_i\right)^2 \tag{4}$$

- Then, a *Least Squares Estimator* (of β_1 and β_2, based upon $(x_1, y_1), \dots (x_n, y_n)$), is any

$$(\hat{\beta}_1, \hat{\beta}_2) \in \arg\min_{(\beta_1,\beta_2)\in\mathbb{R}^2} S(\beta_1, \beta_2). \tag{5}$$

- Questions we want to answer in this course:
  - Is there a minimizer of S ?
  - If there is a minimizer, is it unique ?
  - If there is a minimizer (eventually unique) how to compute it ?
  - Do the minimizer has good statistical properties (in particular, is it close to β_1 and β_2, and does it allow to make good predictions ?)

3.2 Closed-form solution

- **Generic method**: we want to minimize S, which is convex and differentiable $\Rightarrow$ it suffices to differentiate.
- We define the **normal equations**, *i.e.* $(\hat{\beta}_1, \hat{\beta}_2)$ is solution to

$$\frac{\partial S}{\partial \beta_1}(\hat{\beta}_1, \hat{\beta}_2) = 0, \text{ and } \frac{\partial S}{\partial \beta_2}(\hat{\beta}_1, \hat{\beta}_2) = 0. \tag{6}$$

- Another (perhaps more direct) approach. Let $\mathbf{1} \coloneqq (1, \dots, 1)^T \in \mathbb{R}^n$, $\boldsymbol{y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ and $\boldsymbol{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^n$. Then,

$$S(\beta_1, \beta_2) = \|\boldsymbol{y} - \beta_1 \mathbf{1} - \beta_2 \boldsymbol{x}\|^2. \tag{7}$$

Letting,

$$\bar{x} \coloneqq \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ and } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{8}$$

we can rewrite (details left as an exercise!)

$$\begin{aligned} S(\beta_1, \beta_2) &= \|(\boldsymbol{y} - \bar{y}\mathbf{1}) - \beta_2(\boldsymbol{x} - \bar{x}\mathbf{1}) + (\bar{y} - \beta_1 - \beta_2\bar{x})\mathbf{1}\|^2 && (9)\\ &= \|(\boldsymbol{y} - \bar{y}\mathbf{1}) - \beta_2(\boldsymbol{x} - \bar{x}\mathbf{1})\|^2 + (\bar{y} - \beta_1 - \beta_2\bar{x})^2\|\mathbf{1}\|^2 && (10)\\ &= \|(\boldsymbol{y} - \bar{y}\mathbf{1}) - \beta_2(\boldsymbol{x} - \bar{x}\mathbf{1})\|^2 + n(\bar{y} - \beta_1 - \beta_2\bar{x})^2. && (11) \end{aligned}$$

Then we notice the following:

- The first term in equation (11) does not depend on β_1 !
- The second term is always non-negative, and minimal when $\beta_1 = \bar{y} + \beta_2\bar{x}$. Since the first term does not depend on $\beta_1 \implies \hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}$.
- Remains to find $\hat{\beta}_2$, which is the minimizer of $\beta_2 \mapsto \|(\boldsymbol{y} - \bar{y}\mathbf{1}) - \beta_2(\boldsymbol{x} - \bar{x}\mathbf{1})\|^2$. We rewrite,

$$\|(\boldsymbol{y} - \bar{y}\mathbf{1}) - \beta_2(\boldsymbol{x} - \bar{x}\mathbf{1})\|^2 = \|\boldsymbol{x} - \bar{x}\mathbf{1}\|^2\beta_2^2 - 2(\boldsymbol{x} - \bar{x}\mathbf{1})^T(\boldsymbol{y} - \bar{y}\mathbf{1})\beta_2 + \|\boldsymbol{y} - \bar{y}\mathbf{1}\|^2 \tag{12}$$

- This a second order polynomial with dominating coefficient > 0 (under mild assumptions), it has a unique minimizer given by

$$\hat{\beta}_2 = \frac{(\boldsymbol{x} - \bar{x}\mathbf{1})^T(\boldsymbol{y} - \bar{y}\mathbf{1})}{\|\boldsymbol{x} - \bar{x}\mathbf{1}\|^2}. \tag{13}$$

- Remark that if $\boldsymbol{x} - \bar{x}\mathbf{1} = (0, \dots, 0)^T$, then the minimizer of S is not unique and there is not so much we can do. To avoid this scenario, we will always make the following assumption, which guarantees that $\boldsymbol{x} - \bar{x}\mathbf{1} \neq 0$.

  Assumption 1 (H_A)**.** *There exists $(i, j) \in \{1, \dots, n\}^2$ such that $x_i \neq x_j$, i.e. there are at least two covariates that are different.*

- **Conclusion**: Under the assumption H_A, there exists a unique minimizer of S, which is given by

$$\begin{cases} \hat{\beta}_2 = \frac{(\boldsymbol{x} - \bar{x}\mathbf{1})^T(\boldsymbol{y} - \bar{y}\mathbf{1})}{\|\boldsymbol{x} - \bar{x}\mathbf{1}\|^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}. \end{cases} \tag{14}$$

  That is, under H_A, there exists a unique least sqaures estimator, and it is given by the expression above.

- What do we do with $(\hat{\beta}_1, \hat{\beta}_2)$?
  - **Trend lines** (droite de régression), $y = \hat{\beta}_1 + \hat{\beta}_2 x$;
  - **Adjusted values** (valeurs ajustées), $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$, $i = 1, \dots, n$. Those are estimators of $\mathbb{E}[y_i] = \beta_1 + \beta_2 x_i$ (which is unknown as we only observe $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$).
  - **Predict values** : Given a new x_{n+1}, we can predict $\hat{y}^p_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}$ (different from y_{n+1}!)

- **Important**! $\hat{\beta}_1$ and $\hat{\beta}_2$ are **random variables**, *i.e.* if we observe a new set of observations $y_1, \dots, y_n$, then the values for $(\hat{\beta}_1, \hat{\beta}_2)$ will be different (whereas the values of (β_1, β_2) are the same). That is, the trend line we obtain depends on our observations of $y_1, \dots, y_n$.

3.3 Statistical properties

Eventually skip this section depending on schedule.

- We formalize in the next assumption what we already said about the noise.

  Assumption 2 (H_B)**.** *The variables $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d, $\mathbb{E}[\varepsilon_1] = 0$, $\mathbb{E}[\varepsilon_1^2] = \sigma^2 < \infty$ and $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ for all $i \neq j$ (no correlation).*

- Then, we have the following (Exercise: prove it)

  Proposition 1 (CML, Proposition 1.1)**.** *Under H_A and H_B, $\hat{\beta}_1$ is an unbiased estimator of β_1 and $\hat{\beta}_2$ is an unbiased estimator of β_2; that is $\mathbb{E}[\hat{\beta}_1] = \beta_1$ and $\mathbb{E}[\hat{\beta}_2] = \beta_2$.*

  Proposition 2. *Under H_A and H_B,*

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{n} \frac{\frac{1}{n}\sum_{i=1}^n x_i^2}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}, \text{ and, } \mathrm{var}(\hat{\beta}_2) = \frac{\sigma^2}{n} \frac{1}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} \tag{15}$$

- Some remarks about the variance:
  - It is proportional to the variance of the noise σ^2, *i.e.* the more noise, the more it is difficult to estimate β_1 and β_2.
  - It is proportional to $\frac{1}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$. The more the x_i's are spaced, the better the estimate is.
  - The variance decreases as n increases (the more data we have, the better it is).

- We may wonder if we can find better estimators than the least squares estimator. Indeed the **Gauss-Markov** theorem partially answers this question: if the estimator is unbiased and depends linearly on $(y_1, \ldots, y_n)$, then it cannot be better than the least squares estimator. Note that this do not prove that the least squares is the best estimator. Indeed, there exist biased linear estimators that are always better (details next year ?).

  Theorem 1 (Gauss-Markov)**.** *Under H_A and H_B, the least squares estimator has the smallest variance among all the unbiased and linear estimators of (β_1, β_2). Formally, if $(\tilde{\beta}_1, \tilde{\beta}_2)$ is another unbiased linear estimator of (β_1, β_2), then $\mathrm{var}(\tilde{\beta}_1) \geq \mathrm{var}(\hat{\beta}_1)$ and $\mathrm{var}(\tilde{\beta_2}) \geq \mathrm{var}(\hat{\beta}_2)$.*

- Exercise: prove Gauss-Markov's theorem.

3.4 Residual variance

- It may be the case that we don't know the noise's variance σ^2.
- σ^2 is called **residual variance**. The term residual is to distinguish with the variance of Y_i, which may also contains some variability due to X_i in case X_i's are random.
- If we don't know σ^2, we may want to estimate it. How ?
- We define the *residuals*

$$\hat{\varepsilon}_i := y_i - \hat{y}_i = y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i). \tag{16}$$

- The residual $\hat{\varepsilon}_i$ is the difference between the **observed value** y_i and the **adjusted value** $\hat{y}_i$. It estimates ε_i.
- We deduce an estimate of σ^2:

$$\hat{\sigma}^2 := \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2. \tag{17}$$

Proposition 3 (CML Proposition 1.4)**.** *$\hat{\sigma}^2$ is an unbiased estimator of σ^2, i.e. $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.*

- Why $n-2$? Intuitively, we lost two degrees of freedom by estimating $\hat{\beta}_1$ and $\hat{\beta}_2$.

3.5 Prediction

- By definition: $Y_{n+1} = \beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}$.
- Prediction: $\hat{y}^p_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}$

Proposition 4. *Under H_A and H_B,*

$$\mathbb{E}[\hat{y}^p_{n+1}] = \mathbb{E}[Y_{n+1}], \text{ and, } \operatorname{var}(\hat{y}^p_{n+1}) = \frac{\sigma^2}{n}\Big(1 + \frac{(x_{n+1}-\bar{x})^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\Big). \tag{18}$$

- Exercise: prove Proposition 4.
- The variance indicates how well $\hat{y}^p_{n+1}$ estimates $\mathbb{E}[Y_{n+1}]$ (notice that this is $\neq$ from Y_{n+1}).
  - $\hat{y}^p_{n+1}$ is unbiased!
  - $\operatorname{var}(\hat{y}^p_{n+1})$ small if:
    - σ^2 is small;
    - $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ is large;
    - x_{n+1} is close to $\bar{x}$ (easier to predict close to the mean!).
- Optimality of the prediction ?
  - We saw that $\hat{y}^p_{n+1}$ is (eventually) a good estimate of $\mathbb{E}[Y_{n+1}]$.
  - What about Y_{n+1} ?
  - Define the prediction error: $\hat{\varepsilon}^p_{n+1} = Y_{n+1} - \hat{y}^p_{n+1}$.

    Proposition 5. *Under H_A and H_B,*

$$\mathbb{E}[\hat{\varepsilon}^p_{n+1}] = 0, \text{ and, } \operatorname{var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2 + \operatorname{var}(\hat{y}^p_{n+1}). \tag{19}$$

  - Exercise: prove Proposition 5.
  - Intuitively, the error decomposes in two parts:
    - How accurate we are in estimating $\mathbb{E}[Y_{n+1}] \Rightarrow \operatorname{var}(\hat{y}^p_{n+1})$;
    - How close to its expectation is Y_{n+1} in average ? $\Rightarrow \sigma^2$. This is due to the fact that $\hat{y}^p_{n+1}$ is a good estimate of $\mathbb{E}[Y_{n+1}]$, but it is impossible to predict the noise ε_{n+1}.

3.6 Analysis of variance

- Pythagoras theorem (see also Fig. 1): $\|\boldsymbol{y} - \bar{y}\mathbf{1}\|^2 = \|\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}\|^2 + \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$. Equivalently,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{20}$$

$$= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 \tag{21}$$

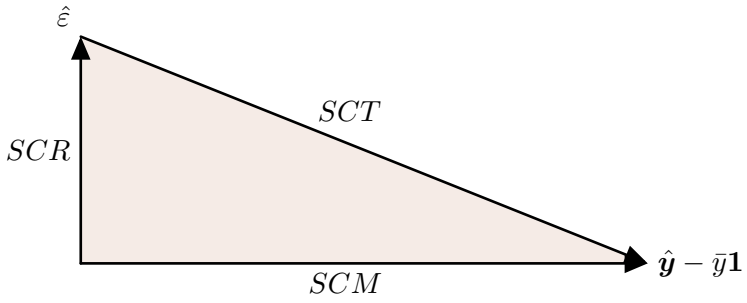


Figure 1: Representation of Pythagoras' theorem: the vectors $\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}$ and $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ are orthogonal, *i.e.* the residuals and the (centered) adjusted values are orthogonal.

- Introducing SCT (*somme des carrés totale*), SCM (*somme des carrés modèle*), and SCR (*somme des carrés résiduels*),

$$SCT = SCM + SCR. \tag{22}$$

- Intuitively:
  - SCT is n times the empirical variance of $(y_1, \ldots, y_n)$;
  - SCM is the fraction of the variance explained by the model;
- Coefficient of determination:

$$R^2 := \cos^2(\theta) = \frac{SCM}{SCT} = \frac{\|\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}\|^2}{\|\boldsymbol{y} - \bar{y}\mathbf{1}\|^2} = 1 - \frac{SCR}{SCT}. \tag{23}$$

- Interpretation:
  - $R^2 = 1$: the model explains completely the observations (no residual).
  - $R^2 = 0$: the model explains nothing.
  - $R^2 \approx 0$: evidence that the model is not adequate.
  - R^2 not close to 0 but also not close to 1: model is incomplete.
  - $R^2 \approx 1$: complete model.
- We will give a more precise meaning to this in Section 5.

4 Confidence intervals in the Normal mean regression model

4.1 The Normal mean regression model

- Until now, we only assumed a few on $\varepsilon_1, \ldots, \varepsilon_n$ (iid, finite variance).
- We do a bit more now:

  Assumption 3 (H_C)**.** *$\varepsilon_1, \ldots, \varepsilon_n$ are iid, and $\varepsilon_i \sim N(0, \sigma^2)$.*

- Obviously: $H_C \Rightarrow H_B$ (*i.e.*, H_C is stronger).
- Under H_C, we can determine the law of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\sigma}$; and use this to construct CI or tests. Note that this was not needed for point estimation, for which the results hold in more generality.
- Remark: If H_C is valid, then we can improve on Gauss-Markov's theorem, $(\hat{\beta}_1, \hat{\beta}_2)$ is of minimal variance among all the unbiased estimators of (β_1, β_2) (*i.e.* and not only among unbiased linear).
- Remark: $(\hat{\beta}_1, \hat{\beta}_2)$ is the Maximum Likelihood Estimator of (β_1, β_2) within the normal mean regression model.

4.2 Useful distributions

- χ-square distribution: If $Z_i, \ldots, Z_n$ are i.i.d $N(0,1)$, then $S \coloneqq \sum_{i=1}^n Z_i^2$ has a $\chi^2(n-2)$ distribution (χ-square with n degrees of freedom).
- Student t-distribution: If $Z \sim N(0,1)$ and $U \sim \chi^2(n)$, then the random variable $T \coloneqq \frac{Z}{\sqrt{U/n}}$ has $\mathcal{T}(n)$ distribution (Student t-distribution with n degrees of freedom).
- Fisher F-distribution: If $U_1 \sim \chi^2(n_1)$ and $U_2 \sim \chi^2(n_2)$ are independent, then $F \coloneqq \frac{U_1/n_1}{U_2/n_2}$ has a $\mathcal{F}(n_1, n_2)$ distribution (Fisher F-distribution with n_1, n_2 degrees of freedom).

4.3 Law of least squares estimator

- To construct CIs, we need to know the law of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\sigma}^2$.

  Proposition 6. *Under H_A and H_C, the following are true.*

  1. *$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ has a Normal distribution with expectation $\beta = (\beta_1, \beta_2)$ and variance $\frac{\sigma^2}{n} V$, with*
  $$V = \frac{1}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \tag{24}$$
  2. *Marginally, $\hat{\beta}_j$ has a $N(\beta_j, \sigma^2_{\hat{\beta}_j})$ distribution, with*
  $$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{n} \frac{\frac{1}{n}\sum_{i=1}^n x_i^2}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}, \text{ and } \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{n} \frac{1}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{25}$$
  3. *$\frac{n-2}{\sigma^2}\hat{\sigma}^2$ has $\chi^2(n-2)$ distribution (i.e. a χ-square distribution with $n-2$ degrees of freedom).*

4. $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ *and* $\hat{\sigma}^2$ *are independent.*

- We admit the proof of Proposition 6 (Exercise: consequence of Cochran's theorem).
- Proposition 6 is in general not enough to construct CIs: this is because in general we don't know σ^2.

Proposition 7. *Under* H_A *and* H_C*:*

1. $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{T}(n-2)$*, with*

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{n} \frac{\frac{1}{n}\sum_{i=1}^n x_i^2}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}, \text{ and } \hat{\sigma}^2_{\hat{\beta}_2} = \frac{\hat{\sigma}^2}{n} \frac{1}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{26}$$

2. *(Bonus).* $\frac{n}{2\hat{\sigma}^2}(\hat{\beta} - \beta)^T V^{-1} (\hat{\beta} - \beta) \sim \mathcal{F}(2, n-2)$ *(with* $\hat{\beta} = (\beta_1, \beta_2)$*).*

- Proof of item (1) is immediate consequence of Proposition 6 and the definition of the Student t-distribution, and item (2) of Proposition 6 and the definition of Fisher's F-distribution.

4.4 Confidence intervals for $\beta_1, \beta_2, \sigma^2$

- Recall the definition of a CI:
  - $I(\beta_j) \subset \mathbb{R}$ is a confidence interval of confidence level $1 - \alpha \in [0, 1]$ for β_j if

$$\mathbb{P}_{(\beta_1, \beta_2, \sigma^2)}(I(\beta_j) \ni \beta_j) \geq 1 - \alpha, \quad \forall (\beta_1, \beta_2, \sigma^2). \tag{27}$$

  - Note that $I(\beta_j)$ is random, not β_j.

Proposition 8. *Under* H_A *and* H_C*.*

1. $CI(\beta_j) \coloneqq [\hat{\beta}_j - t_{n-2}(1-\alpha/2)\hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t_{n-2}(1-\alpha/2)\hat{\sigma}_{\hat{\beta}_j}]$ *is a* CI *for* β_j *with confidence level* $1 - \alpha$*, where* $t_{n-2}(1-\alpha/2)$ *is the quantile of order* $1 - \alpha/2$ *of the Student* t*-distribution with* $n - 2$ *degrees of freedom.*
2. $[\frac{(n-1)\hat{\sigma}^2}{c_{n-2}(1-\alpha/2)}, \frac{(n-2)\hat{\sigma}^2}{c_{n-2}(\alpha/2)}]$ *is an* CI *for* σ^2 *with confidence level* $1 - \alpha$*, where* $c_{n-2}(\gamma)$ *is the quantile of order* γ *of the* $\chi^2(n-2)$ *distribution.*

- Proof is immediate from Proposition 7 and Proposition 6.
- Remarks:
  - The Proposition 8 gives two-sided (*bilatère*) and symmetric intervals. This is not the only possibility (one-sided, or non-symmetric).
  - The Proposition 8 gives interval marginally for β_1 and β_2, but not for $\beta = (\beta_1, \beta_2)$.
    - A confidence region for $\beta = (\beta_1, \beta_2)$ with level $1 - \alpha$ is any (random) set $E_\alpha \subseteq \mathbb{R}^2$ such that $\mathbb{P}_{(\beta_1, \beta_2, \sigma^2)}(E_\alpha \ni (\beta_1, \beta_2)) \geq 1 - \alpha$ for all $(\beta_1, \beta_2, \sigma^2)$.
    - The set $CI(\beta_1) \times CI(\beta_2)$ is always a confidence region of level $\geq 1 - 2\alpha$ for (β_1, β_2), but this does note take into accounts the correlations between $\hat{\beta}_1$ and $\hat{\beta}_2$!
    - We can do better using item (2) of Proposition 7 $\Rightarrow$ confidence ellipses (see the Fig. 2).

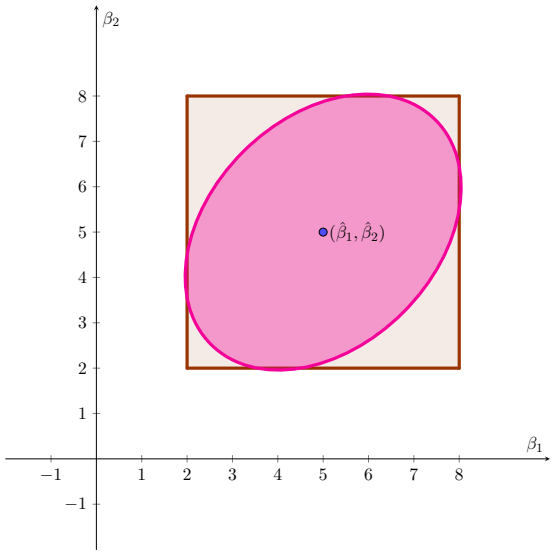


Figure 2: Confidence ellipse versus products of individual confidence intervals for (β_1, β_2) : both the ellipse (in pink) and the products of individuals CI (in brown) are centered on $(\hat{\beta}_1, \hat{\beta}_2)$ but the ellipse takes into accounts the correlations between $\hat{\beta}_1$ and $\hat{\beta}_2$ and is thus slightly better.

4.5 Prediction intervals

- Often called : "prediction intervals for y_{n+1}".
- We are looking for a random interval $I \subseteq \mathbb{R}$ such that

$$\mathbb{P}_{(\beta_1,\beta_2,\sigma^2)}(I \ni Y_{n+1}) \geq 1-\alpha, \quad \forall(\beta_1, \beta_2, \sigma^2). \tag{28}$$

Proposition 9. *Under H_A and H_C,*

$$IP(Y_{n+1}) \coloneqq \left[\hat{y}^p_{n+1} \pm t_{n-2}(1-\alpha/2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{1}{n}\frac{(x_{n+1}-\bar{x})^2}{\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2}}\right], \tag{29}$$

is a prediction interval of y_{n+1} with level $1-\alpha$.

- Remarks:
  - We got an extra $+1$ compared with confidence interval: intuitively, we have to predict the *noise*, which cannot be estimated, thus we are always less confident in the predicting Y_{n+1} than we are in estimating $\mathbb{E}[Y_{n+1}]$, see the Fig. 3.
  - The diameter of $IP(Y_{n+1})$ does not $\to 0$ as $n \to \infty$. This is still because the noise cannot be estimated. As $n \to \infty$, we are more and more confident about what is $\mathbb{E}[Y_{n+1}]$, but it is still impossible to predict the value of ε_{n+1}... (hence the extra σ term, corresponding to $\text{var}(\epsilon_{n+1})^{1/2}$).

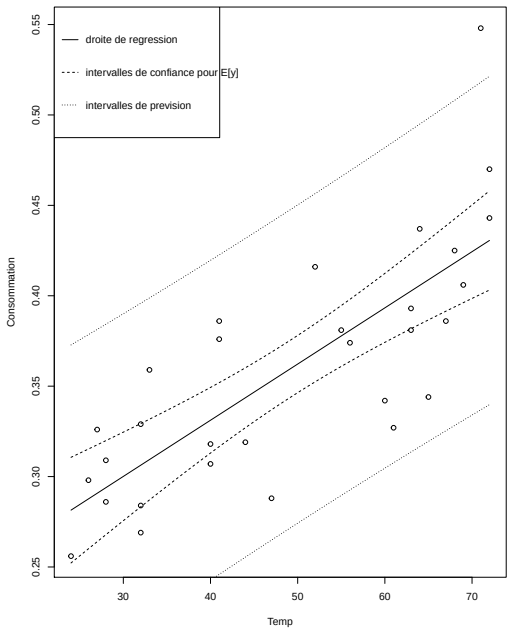


Figure 3: Prediction intervals for Y_{n+1} versus confidence interval for $\mathbb{E}[Y_{n+1}]$: We plotted the regression function in straight line, and for all values of temperature $x \in [25, 80]$ we plotted the associated confidence interval for $\mathbb{E}[Y] = \beta_1 + \beta_2 x$ and the prediction interval. This gives a collection of confidence and prediction intervals whose endpoints are plotted in dashed lines (confidence intervals) and with points (prediction intervals). We can see that the prediction intervals are always larger as we have to take into account the noise which can't be estimated. We can also notice that the intervals are smaller around the average temperature $x \approx 50$ than for the extreme values of temperature. This reflects the fact that prediction is easier for values of x close to $\bar{x}$.

5 Tests

5.1 Introduction and reminders

- Problem: We want to answer a precise question
  - Does the temperature affects air pollution ? In a negative way ? In a positive way ?
  - Is the linear model explaining the variability of Y ?
- We need to do **hypothesis testing**.
- We briefly recall the basics of the Neyman-Pearson's approach to hypothesis testing:
  - The goal is to decide (from the data) between two complementary hypotheses: the null hypothesis H_0 and the alternative hypothesis H_1 (one of which has to be true, and they cannot be both true).
  - The data provides incomplete information about the problem, which means that most of time we cannot decide with certitude which is the correct hypothesis: we can make errors.

- We can be mistaken in two ways:
  - H_0 is true and we choose H_1 (Type I error);
  - H_1 is true and we choose H_0 (Type II error).
- We can control one of the two errors, but not the two error simultaneously. The convention is to control the Type I error $\Rightarrow$ level of the test.
- The level α of the test is define as the (maximum) probability of choosing H_1 while H_0 is true, *i.e.* $\mathbb{P}_{H_0}(\text{reject } H_0) \leq \alpha$.
- This implies that the role of H_0 and H_1 are very different (non symmetric), and we have to pay attention to the way we choose the competing hypotheses. That is, once we have chosen α, we have no control on the Type II error (then the only way to reduce the Type II error is to collect more data!).
- The p-value of the test is the smallest value of α for which the test rejects H_0, *i.e.* p-value $\leq \alpha \Leftrightarrow$ reject H_0.

- How do we do the test in practice ?
  - Step 1: Choose the model (*i.e.* What do I assume about the data I collected)
  - Step 2: Select the competing hypotheses H_0 and H_1
  - Step 3: Choose a **test statistics** T (important: we need to know the distribution of T under H_0)
  - Step 4: Determine the shape of the rejection zone $\mathcal{R}$, *i.e.* for which values of T we will reject/accept H_0 ?
  - Step 5: Choose α and
  - Step 6: Decision
    - Alternative 1: Calibrate the test, *i.e.* choose the "size" of the rejection zone such that $\mathbb{P}_{H_0}(T \in \mathcal{R}) \leq \alpha$, and compute the observed value of the test statistic t^{obs}; Choose between H_0 and H_1 according to whether or not $t^{obs} \notin \mathcal{R}$.
    - Alternative 2: Compute the p-value and choose between H_0 and H_1 according to whether or not p-value $> \alpha$.
  - Step 7: Conclude, *i.e.* answer the question asked.

5.2 Testing the parameters of the model

- Testing if the slope is null (Student's test) ?
  - Model: $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $i = 1, \ldots, n$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d;
  - $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$
  - Test statistic:
$$T = \sqrt{n}\hat{\beta}_2 \Big/ \sqrt{\frac{\hat{\sigma}^2}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}} \tag{30}$$
    If H_0 is true, then $T \sim \mathcal{T}(n-2)$.
  - We reject if $|t^{obs}| \geq t_{n-2}(1-\alpha/2)$ (two-sided test)

- p-value: $\mathbb{P}_{\mathcal{T}(n-2)}(|T| > |t^{obs}|)$.

- Testing if the intercept is null (Student's test) ?
  - Model: $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $i = 1, \ldots, n$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d;
  - $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$
  - Test statistic:
$$T = \sqrt{n}\hat{\beta}_1 \Big/ \sqrt{\hat{\sigma}^2 \Big(1 + \frac{\bar{x}^2}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}\Big)}. \tag{31}$$
    If H_0 is true, then $T \sim \mathcal{T}(n-2)$.
  - We reject if $|t^{obs}| \geq t_{n-2}(1 - \alpha/2)$ (two-sided test)
  - p-value: $\mathbb{P}_{\mathcal{T}(n-2)}(|T| > |t^{obs}|)$.

- Many other possibilities:
  - One-sided tests: $H_0 : \beta_2 \leq 0$ versus $H_1 : \beta_2 > 0$, etc.
  - $H_0 : \beta_2 = 10$ versus $H_1 : \beta_2 \neq 10$, etc.
  - etc.

5.3 Analysis of variance

- See Lesson 2 for more details.

- We want to test the following hypotheses:
  - H_0 : The simpler model M_0 is correct : $Y_i = \beta_1 + \varepsilon_i$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$; versus
  - H_1 : The linear model M_1 is correct: $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

- Note that this is equivalent to: $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$, but here we propose a different approach compared with the previous section. This approach will be justified in the next Lesson (multiple linear regression).

- Recall R^2, SCT, SCR, SCE.

- Basic intuition: if the model M_1 is correct, then it explains better the data, *i.e.* R^2 should be higher. But $M_0 \subset M_1$, so indeed R^2 is always better if we choose M_1. Intuitively, we will choose to use M_1 only if R^2 is significantly better (otherwise it is preferable to keep the number of parameters as low as possible). The question is then, what is significantly better ?

- We have (see Cochran's theorem for details)
  - $SCR = (n-2)\sigma^2 \sim \sigma^2 \chi^2(n-2)$ (which is true both under H_0 and H_1);
  - If H_0 is true: $SCT \sim \sigma^2 \chi^2(n-1)$
  - If H_0 is true: $SCM \sim \sigma^2 \chi^2(1)$.

- Test statistic (Fisher's statistic):
$$F = \frac{SCM/1}{SCR/(n-2)} \tag{32}$$
If H_0 is true $F \sim \mathcal{F}(1, n-2)$ (this is because SCM and SCR are independent, consequence of Cochran's theorem).

- We reject H_0 is $F \geq f_{1,n-2}(1-\alpha)$ (quantile of order $1-\alpha$ of the $\mathcal{F}(1, n-2)$ distribution).

- Intuitively, F is large when M_1 explains better the variability of Y than pure randomness).

- Remark:
$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT} = \frac{1}{1 + \frac{SCR}{SCM}} = \frac{1}{1 + \frac{n-2}{F}}. \tag{33}$$
Hence, F large $\Leftrightarrow R \approx 1$.

- p-value of the test : $\mathbb{P}_{\mathcal{F}(1,n-2)}(F \geq f^{obs})$.

- Remark: This test is completely equivalent to Student's test for $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$. This is because $F = T^2$.

6 Diagnostics

- Goal: Checking if the hypotheses we did are valid:
  - H_A: this is immediate and easy.
  - H_B of H_C: less obvious.
    - Linearity / Goodness of fit $\Leftrightarrow \varepsilon_i$ centered, non correlated to x_i
    - Homoscedasticity (*i.e.* all ε_i have the same variance)
    - The ε_i are non-correlated, *i.e.* $i \neq j \Rightarrow \mathbb{E}[\varepsilon_i \varepsilon_j] = 0$.
    - If we assumed H_C: ε_i have normal distribution.
  - In general we want to check is the previous are true globally, except perhaps for a few number of observations (outliers).

- Method: Analysis of residuals, plots, tests (see practicals).

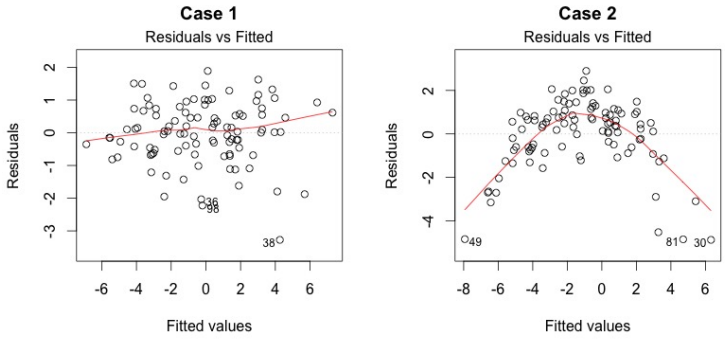
- More formal diagnostics will be made in Lesson 2. These are only preliminary diagnostics.

6.1 Some useful diagnostic plots

- Depending on what we want to analyze: plot $\hat{\varepsilon}_i$ as a function of i, $\hat{y}_i$, x_i, ...
  - Can detect outliers: if $|\hat{\varepsilon}_i|/\hat{\sigma}$ is too large. If it is the case, need to understand why. This is visible if we plot $\hat{\varepsilon}_i$ as a function of i
  - Rescaling the residuals $\hat{\varepsilon}_i$ by the residual standard deviation $\hat{\sigma}$ is a good idea (Intuitively, the residuals must be in average within $[-2\hat{\sigma}, 2\hat{\sigma}]$ with high probability.
  - $\hat{t}_i \coloneqq \hat{\varepsilon}_i/\hat{\sigma}$ is the standardized residuals $\Rightarrow$ Law is rather complicated, so indeed we prefer to work with the Studentized/Jacknife residuals $\hat{t}_i^*$ (see practicals Q6 and/or Lesson 2).

– With high probability : $\hat{t}_i^* \in [-2, 2]$, if there is a point for which $|\hat{t}_i^*| \gg 2$, then it is likely to be an outlier, we have to wonder why.

– $\hat{\varepsilon}_i$ as a function of $\hat{y}_i$: can be used to detect if the residuals have non-linear patterns (and hence the model does not explain completely the data, *i.e.* if the model is correct, the residuals are random and centered, see bonus section).



– $\hat{\varepsilon}_i$ as a function of $\hat{y}_i$: Can also be used to detect heteroscedasticity.

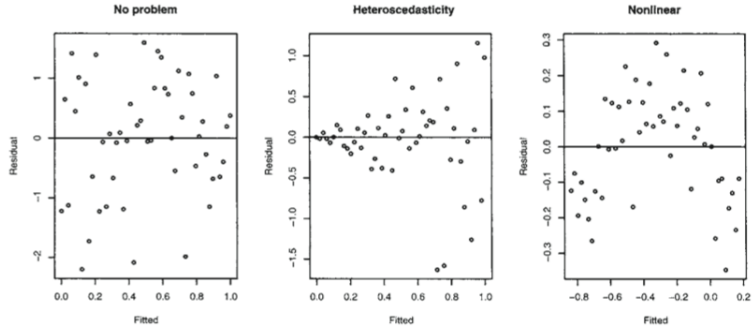
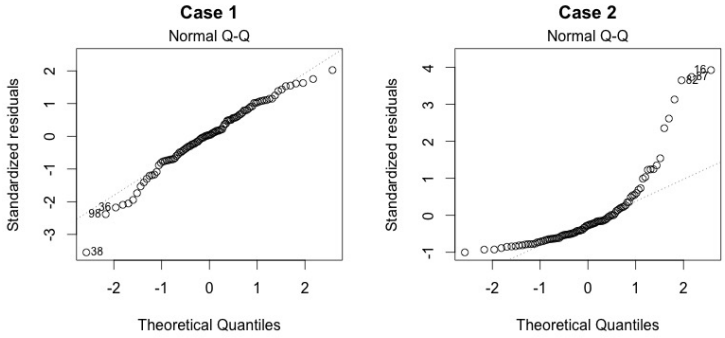


Figure 4.1 *Residuals vs. fitted plots—the first suggests no change to the current model while the second shows nonconstant variance and the third indicates some nonlinearity, which should prompt some change in the structural form of the model.*

– QQ-Plot: to determine if the residuals are normal. Plot the empirical quantiles of $\hat{\varepsilon}_i$ as a function of the theoretical quantile of the Normal distribution.



6.2 Some useful tests

- Testing for the independence of ε_i's: Durbin-Watson test.
  - Test the null hypothesis $\hat{\varepsilon}_i = U_i$ versus H_1 : there is $\rho \neq 0$ such that $\hat{\varepsilon}_i = \rho\varepsilon_{i-1} + U_i$, where $U_1, \ldots, U_n$ are i.i.d normal random variables, *i.e.* test if ε_i and ε_{i-1} are linearly correlated or not.
  - Test statistic is:
  $$d = \frac{\sum_{i=2}^{n}(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}. \tag{34}$$
  Law under H_0 is known, but depends on $x_1, \ldots, x_n$.
  - R: command durbinWatsonTest.
  - What to do if we reject H_0 ?
    - The point estimation still ok in average ($\hat{\beta}_1$ and $\hat{\beta}_2$ are still unbiased), but the variance might be large (and most importantly, not controlled).
    - This is problematic for confidence intervals (the model is wrong, we don't know the laws of $\hat{\beta}_1$ and $\hat{\beta}_2$ anymore).
- Testing for normality of ε_i's:
  - Various tests: Shapiro-Wilk, Kolmogorov, ...
  - Details in Lesson 2.
  - What to do if the ε_i's are not Normal ?
    - Not an issue for point estimation (as long as H_B remains true!)
    - If the law of the ε_i's is symmetric, impact is in general mild on the CI's and tests (though we have to be careful)
    - If the law is non symmetric: we cannot really trust CI's and tests anymore.
    - Transformation of the data can sometimes solve the issue.

6.3 Bonus: Law of the residuals, different types of residuals

- Goal is to make more formal the visual diagnostics above, by determining the law of $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)$ and related quantities.
- See lesson 2.