Lesson 2: Multiple Linear Regression

Zacharie Naulet
zacharie.naulet@math.u-psud.fr

Contents

Motivations

- Air pollution (O_3)
  - The model with one explanatory variable (*i.e.* the temperature at noon) is too simplistic
  - Need more complex models ? $O_3 = \beta_1 + \beta_2 T + \beta_3 \sqrt{T} + \beta_4 T^4 + \varepsilon$?
  - And/Or need more explanatory variables ? (Nebulosity, wind, ...)
- Other examples mentioned in Lecture 1:
  - Diabetes as function of : age, blood pressure, body mass index (BMI), concentration in blood of certain proteins, ...
  - Wheat crop yield *i.e.* agricultural output) as function of : fertilizer quantity, gene expression, ...
  - ...

1 Mathematical formalism

- **Goal**: Explain the variations of a quantitative variable $Y \in \mathbb{R}$ from the measurement of multiple other quantitative variables $X_1, \ldots, X_p \in \mathbb{R}$.
  - Observation is $(y_i, x_{i,1}, \ldots, x_{i,p})$, $i = 1, \ldots, n$.
  - We want to find f linear (affine) such that $y_i \approx f(x_{i,1}, \ldots, x_{i,p})$ for all i.
- Convenient convention/usage (which we will always use), is to always choose $x_{i,1} = 1$:
  - f linear $\Leftrightarrow f(x_{i,1}, \ldots, x_{i,p}) = \sum_{j=1}^{p} \beta_j x_{i,j} = \beta_1 + \sum_{j=2}^{p} \beta_j x_{i,j}$.
  - Using matrix/vector notations (more compact):
    - $x_i = (x_{i,1}, \ldots, x_{i,p})$;
    - $\beta = (\beta_1, \ldots, \beta_j)$;
    - $\sum_{j=1}^{p} \beta_j x_{i,j} = \beta_1 + \sum_{j=2}^{p} \beta_j x_{i,j} = \beta^T x_i$.

    * Note that β_1 is the intercept!
    * Given observations $(y_1, x_1), \dots, (y_n, x_n)$ we want to find $\beta \in \mathbb{R}^p$ such that $y_i \approx \beta^T x_i$ for all $i = 1, \dots, n$.

- Statistical model:
  - We assume for $i = 1, \dots, n$:

$$Y_i = \beta^T x_i + \varepsilon_i = \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i$$

  - The covariates $x_1, \dots, x_n$ are dterministic (not random);
  - Only $\varepsilon_1, \dots, \varepsilon_n$ (and thus $Y_1, \dots, Y_n$) are random.

- Statistical model in vector/matrix form:
  - Define

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \in \mathbb{R}^n, \quad \boldsymbol{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^p.$$

  - Be careful with the dimensions, in general $n \neq p$!
  - Then we can rewrite the model as:

$$Y = \boldsymbol{X}\beta + \varepsilon.$$

2 Point estimation using least squares

2.1 Theoretical definition

- How do we estimate the parameter β given $\boldsymbol{X}$ and $Y = (y_1, \dots, y_n)$?
  - Least squares estimator;
  - *i.e.* find $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ minimizing $\beta \mapsto \|y - \boldsymbol{X}\beta\|^2$;
  - Equivalently, $\hat{\beta}$ minimizes:

$$(\beta_1, \dots, \beta_p) \mapsto \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \Big)^2.$$

  - Equivalently,

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \|y - \boldsymbol{X}\beta\|^2.$$

2.2 Geometrical interpretation

2.2.1 Reminders of linear algebra : orthogonal projections

- Two sets $V, W \subseteq \mathbb{R}^n$ are orthogonal if $\forall v \in V$ and $\forall w \in W$ we have $v^T w = 0$.
- For any $V \subset \mathbb{R}^n$ subspace of $\mathbb{R}^n$ we let

$$\begin{aligned} V^\perp &:= \{w \in \mathbb{R}^n \,:\, v^T w = 0\} \\ &\equiv \text{Linear subspace orthogonal to } V. \end{aligned}$$

- Any element of $\mathbb{R}^n$ can be written uniquely as the sum of one element from V and one element from $V^\perp$:

$$\mathbb{R}^n \ni x = \underbrace{P_V(x)}_{\in V} + \underbrace{P_{V^\perp}(x)}_{\in V^\perp};$$

  - P_V: Orthogonal projection on V;
  - $P_{V^\perp}$: Orthogonal projection on $V^\perp$.

- **Pythagoras theorem**:

$$\|x\|^2 = \|P_V(x)\|^2 + \|P_{V^\perp}(x)\|^2.$$

- $P_V : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map:
  - There is a $n \times n$ matrix P_V such that $P_V(x) = P_V(x)$;
  - Properties: $P_V^2 = P_V$ (projection), $P_V^T = P_V$ (orthogonality), $\mathrm{Tr}(P_V) = \dim(V)$.
  - If V is generated by $v_1, \ldots, v_p \in \mathbb{R}^n$ (not necessarily a basis for V). Then, letting $X_V = (v_1 \cdots v_p)$ the $n \times p$ matrix whose columns are the vectors $v_1, \ldots, v_p$:

$$P_V = X_V (X_V^T X_V)^{-1} X_V^T.$$

2.2.2 Geometrical interpretation of $\hat{\beta}$

- Define the image of $\boldsymbol{X}$:

$$\begin{aligned} \mathrm{Im}(\boldsymbol{X}) &= \{\boldsymbol{X} v \,:\, v \in \mathbb{R}^p\} \subseteq \mathbb{R}^n \\ &= \text{Linear space generated by the columns of } \boldsymbol{X} \end{aligned}$$

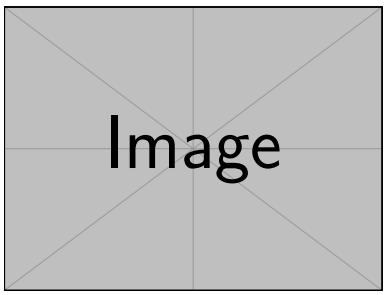
- In other words, letting

$$v_j = \begin{pmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{pmatrix} \in \mathbb{R}^n, \quad j = 1, \ldots, p$$

  $u \in \mathrm{Im}(\boldsymbol{X}) \Leftrightarrow \exists (a_1, \ldots, a_p) \in \mathbb{R}^p$ such that $u = a_1 v_1 + \cdots + a_p v_p$.

- The **adjusted values vector** $\hat{Y} = \boldsymbol{X}\hat{\beta}$ is the orthogonal projection of the observation vector Y onto the space $\mathrm{Im}(\boldsymbol{X})$.

$$\hat{Y} = \begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^p x_{1,j}\hat{\beta}_j \\ \vdots \\ \sum_{j=1}^p x_{n,j}\hat{\beta}_j \end{pmatrix} = \boldsymbol{X}\hat{\beta}.$$



- Using the mathematical results from the previous section we deduce the result (provided that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, see after)

$$\hat{Y} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y. \tag{1}$$

2.3 Closed form solution

Assumption 1 (H_A)**.** $\operatorname{rank}(\boldsymbol{X}) = p \leq n$. *Equivalently,* $\dim(\boldsymbol{X}) = p \leq n$.

- Remarks:
  - If $p = 2$, this is completely equivalent to H_A in Lecture 1.
  - If $\boldsymbol{X}$ has rank $p \leq n$, then $\boldsymbol{X}^T\boldsymbol{X}$ is a $p \times p$ matrix with full rank, and hence invertible.
- Under H_A we can obtain the expression for $\hat{\beta}$ from equation (1):

  Proposition 1. *If H_A is valid, then $\hat{\beta} \in \arg\min_{\beta\in\mathbb{R}^d} \|y - \boldsymbol{X}\beta\|^2$ exists and is unique. It is given by*

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y.$$

- Remark (Identifiability): If H_A is not valid, then the formula $\hat{Y} = \boldsymbol{X}\hat{\beta}$ remains true, but the estimator $\hat{\beta}$ is not unique (*i.e.* there can be many minimizers of $\hat{\beta} \mapsto \|y - \boldsymbol{X}\beta\|^2$, but for any two minimizers $\hat{\beta}_1$ and $\hat{\beta}_2$ we have $\boldsymbol{X}\hat{\beta}_1 = \boldsymbol{X}\hat{\beta}_2$).
- Remark: $\hat{\beta}$ can also be obtained from solving the normal equations, as in the Lecture 1.

2.4 Statistical properties

- As in the Lesson 1, we need some assumptions on $\varepsilon_1, \ldots, \varepsilon_n$

  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).*

- Note that this is exactly the same assumption as in Lecture 1.

  Proposition 2. *If H_A and H_B are valid, then*

  1. *$\hat{\beta}$ is an unbiased estimator of β: $\mathbb{E}[\hat{\beta}] = \beta$;*
  2. *$\operatorname{var}(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ (covariance matrix).*

- Remarks

- If $p = 2$ and $\boldsymbol{X} = (\mathbb{1}\ \boldsymbol{x}_n)$, this is exactly the result of Lecture 1.
- We can obtain the quadratic risk of $\hat{\beta}$ (exercice: prove it)

$$\begin{aligned}\mathbb{E}[\|\hat{\beta} - \beta\|^2] &= \sigma^2 \mathrm{Tr}((\boldsymbol{X}^T\boldsymbol{X})^{-1}) \\ &= \sigma^2 \sum_{j=1}^{p} \mathrm{var}(\hat{\beta}_j).\end{aligned}$$

- Variance if proportional to the noise level σ^2. The smaller is the noise variance, the easier the estimation is.
- Variance is proportional to $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The larger are the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$ are, the easier the estimation is.

- Can we do better than least squares estimation ?

  Theorem 1 (Gauss-Markov)**.** *Under H_A and H_B the least squares estimator $\hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE). That is, for any other linear unbiased estimator $\tilde{\beta}$ the matrix $\mathrm{var}(\hat{\beta}) - \mathrm{var}(\tilde{\beta})$ is positive semi-definite.*

2.5 Residuals and residual variance

2.5.1 Definition of residuals and basic properties

- We let $P_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ denote the orthogonal projection matrix onto $\mathrm{Im}(\boldsymbol{X})$.
- Define the vector of residuals:

$$\hat{\varepsilon} := Y - \hat{Y}$$

- We have:

$$\begin{aligned}\hat{\varepsilon} &= Y - \boldsymbol{X}\hat{\beta} \\ &= Y - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y \\ &= (I - P_{\boldsymbol{X}})Y.\end{aligned}$$

- We remark that $\hat{Y}$ is the orthogonal projection of Y onto $\mathrm{Im}(\boldsymbol{X})$, and thus the residuals are the projection of Y onto $\mathrm{Im}(\boldsymbol{X})^{\perp}$.
- $\mathrm{Im}(\boldsymbol{X})^{\perp}$ is called **residual space**.

  Proposition 3. *Under H_A and H_B:*

  1. $\mathbb{E}[\hat{\varepsilon}] = 0$;
  2. $\mathrm{var}(\hat{\varepsilon}) = \mathbb{E}[\hat{\varepsilon}\hat{\varepsilon}^T] = \sigma^2(I - P_{\boldsymbol{X}})$;
  3. $\mathrm{cov}(\hat{\varepsilon}, \hat{Y}) = 0$ *(covariance matrix of dimension $n \times n$ is null).*

- Remarks:
  - Residuals $\hat{\varepsilon}_i$ are in general correlated (while the ε_i are not by H_B) $\Rightarrow$ this is because $I - P_{\boldsymbol{X}}$ is in general not diagonal.
  - BUT: the variance-covariance of the residuals is known, so that we can "decorrelate" the residuals (see later!)

2.5.2 Estimation of the residual variance

- Recall that the residual variance is defined by $\sigma^2 \coloneqq \mathbb{E}[\varepsilon_1^2]$ (under H_B, the ε_i's are iid with mean zero, so that the residual variance is indeed the common variance of the noise).
- We remark that

$$\mathrm{Tr}(\mathrm{var}(\hat{\varepsilon})) = \sum_{i=1}^{n} \mathrm{var}(\hat{\varepsilon}_i) = \sigma^2 \mathrm{Tr}(I - P_{\boldsymbol{X}}) = \sigma^2 \cdot (n-p).$$

- Consequently:

$$\hat{\sigma}^2 \coloneqq \frac{\|\varepsilon_i\|^2}{n-p} = \frac{SCR}{n-p}$$

satisfies (exercise: prove it):

Proposition 4. *Under H_A and H_B $\hat{\sigma}$ as defined above sis unbiased for σ^2, i.e. $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.*

- Remarks:
  - The number of degree of freedoms is $n - p$.
  - We can estimate the variance of $\hat{\beta}$ (indeed, variance matrix) $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ by

$$\hat{\sigma}^2_{\hat{\beta}} = \hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

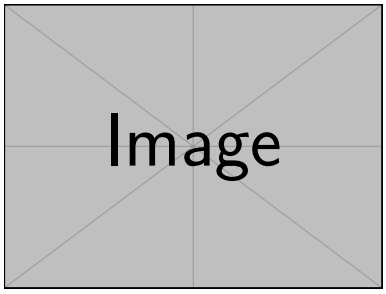
  - And therefore the variance of a single coefficient $\hat{\beta}_j$ $(j = 1, \ldots, p)$, that is $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{j,j}$ can be estimated by

$$\hat{\sigma}^2_{\hat{\beta}_j}(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{j,j}$$

2.6 Coefficient of determination

- Recall the geometric interpretation $\hat{Y} = P_{\boldsymbol{X}}Y$, *i.e.* $\hat{Y}$ is the orthogonal projection of Y onto $\mathrm{Im}(\boldsymbol{X})$, and $\hat{\varepsilon}$ is the orthogonal projection of Y onto $\mathrm{Im}(\boldsymbol{X})^{\perp}$. Then we have Pythagoras' theorem:

$$\|Y\|^2 = \|\hat{Y}\|^2 + \|\hat{\varepsilon}\|^2$$



- If $\mathbb{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$ is in $\mathrm{Im}(\boldsymbol{X})$. Which is **always assumed** anyway in this course because we set $x_{i,1} = 1$ for all $i = 1, \ldots, n$. Then, we can rewrite Pythagoras' theorem as (using $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$):

$$\|Y - \bar{Y} \cdot \mathbb{1}\|^2 = \|\hat{Y} - \bar{Y} \cdot \mathbb{1}\|^2 + \|\hat{\varepsilon}\|^2$$
$$SCT = SCM + SCR$$

- If $\mathbb{1} \in \mathrm{Im}(\boldsymbol{X})$, we define the coefficient of determination R^2 by

$$R^2 \coloneqq \frac{SCM}{SCT} = \frac{\|\hat{Y} - \mathbb{1}\|^2}{\|Y - \bar{Y} \cdot \mathbb{1}\|^2} = \cos^2(\theta) = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{Y} \cdot \mathbb{1}\|^2}.$$

- Remark: If $\mathbb{1} \in \mathrm{Im}(\boldsymbol{X})$, then R^2 can be interpreted as the fraction of the variance of the observations explained by the model.

- Problem: The larger is p and the better becomes R^2. In other words, if we add a lot of covariates to our model, we might think that our model is better because R^2 increases. This is not correct!

- Indeed, if we add to $X_1, \ldots, X_p$ a new variable X_{p+1} which is not informative (say the color of the shoe of the professor), then $\|\hat{\varepsilon}\|^2$ decreases by $\approx \sigma^2$, while $\|Y - \bar{Y} \cdot \mathbb{1}\|^2$ remains the same; that is $R^2 = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{Y} \cdot \mathbb{1}\|^2}$ increases.

- Solution: **Adjusted coefficient of determination**. If $\mathbb{1} \in \mathrm{Im}(\boldsymbol{X})$:

$$R^2_a = 1 - \frac{\|\hat{\varepsilon}\|^2/(n-p)}{\|Y - \bar{Y} \cdot \mathbb{1}\|^2/(n-1)}.$$

2.7 Prediction

- Given a new predictor $x_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})^T$ (beware the transposition!) and we wish to predict the value of the next outcome $Y_{n+1} = x_{n+1}^T \beta$. We can use

$$\hat{y}^p_{n+1} \coloneqq x_{n+1}^T \hat{\beta}.$$

- Under H_A and H_B:

$$\mathbb{E}[\hat{y}^p_{n+1}] = x_{n+1}^T \beta = \mathbb{E}[Y_{n+1}],$$

and,

$$\mathbb{E}[(\hat{y}^p_{n+1} - y_{n+1})^2] = \mathrm{var}(\hat{y}^p_{n+1} - y_{n+1}) = \sigma^2 \Big(1 + x_{n+1}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} x_{n+1}\Big),$$

where

  - 1 comes from the uncertainty due to the noise ε_{n+1},
  - $x_{n+1}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} x_{n+1}$ comes from the uncertainty due to using the estimator $\hat{\beta}$ in place of the true β, *i.e.* uncertainty in estimating $\mathbb{E}[Y_{n+1}]$.

3 Confidence intervals in the Normal mean regression model

3.1 The Normal mean regression model

- Recall Lesson 1: Confidence intervals require the knowledge of the distribution (law) of the estimator $\hat{\beta} \Rightarrow$ we need to do (more) assumptions on the ε_i's.

- We do the exact same assumption as in the Lesson 1 (*aka* Normal mean regression model):

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

- Remark (Bonus): Under H_C we can show that $\hat{\beta}$ is the *Maximum Likelihood Estimator* (see the course on likelihood methods).

3.2 Some reminders about Normal random vectors

- Let $Z \in \mathbb{R}^n$, $Z = (Z_1, \dots, Z_n)^T$ be random vector.
- The expectation of Z is $\mathbb{E}[Z] := (\mathbb{E}[Z_1], \dots, \mathbb{E}[Z_n])^T$ (if it exists).
- The variance of Z (or covariance matrix, or variance-covariance matrix), when it exists, is given by

Matrice de variance covariance

$$\mathrm{var}(Z) := \mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^T] = \begin{pmatrix} \mathrm{var}(Z_1) & \mathrm{cov}(Z_1, Z_2) & \cdots & \mathrm{cov}(Z_1, Z_n) \\ \mathrm{cov}(Z_2, Z_1) & \mathrm{var}(Z_2) & \ddots & \vdots \\ \vdots & & \ddots & \vdots \\ \mathrm{cov}(Z_n, Z_1) & \cdots & \cdots & \mathrm{var}(Z_n). \end{pmatrix}.$$

  It is a $n \times n$ matrix such that the (i, j)-th coefficient if equal to $\mathrm{cov}(Z_i, Z_j) = \mathrm{cov}(Z_j, Z_i)$ (symmetric matrix).

- If $\boldsymbol{A}$ is a non random $m \times n$ matrix , then
  - $\mathbb{E}[\boldsymbol{A}Z] = \boldsymbol{A}\mathbb{E}[Z] \in \mathbb{R}^m$;
  - $\mathrm{var}(\boldsymbol{A}Z) = \boldsymbol{A}\mathrm{var}(Z)\boldsymbol{A}^T$ is a $m \times m$ symmetric matrix.
- Let $I_m = \mathrm{diag}(1, \dots, 1)$ be the $m \times m$ identity matrix. We define the multivariate Normal distribution $\mathcal{N}(0, I_m)$ such that $U = (U_1, \dots, U_m)^T$ and each entries $U_i \in \mathbb{R}$ is independent and has a standard $\mathcal{N}(0, 1)$ distribution on $\mathbb{R}$.
- For any $n \times m$ matrix $\boldsymbol{A}$ and $\mu \in \mathbb{R}^n$, we say that $Z \in \mathbb{R}^n$ has a $\mathcal{N}(\mu, \boldsymbol{A}\boldsymbol{A}^T)$ distribution, if $Z \overset{d}{=} \mu + \boldsymbol{A}U$ for $U \sim \mathcal{N}(0, 1)$.
- Consequences:
  - $\mathbb{E}[Z] = \mu$ and $\mathrm{var}(Z) = \boldsymbol{A}\boldsymbol{A}^T =: \Sigma$.
  - If $Z \sim \mathcal{N}(\mu, \Sigma)$ with Σ a $n \times n$ positive-definite and symmetric matrix[1] then for any $m \times n$ matrix $\boldsymbol{B}$ and any $\eta \in \mathbb{R}^m$ we have

$$\eta + \boldsymbol{B}Z \sim \mathcal{N}(\eta + \boldsymbol{B}\mu, \boldsymbol{B}\Sigma\boldsymbol{B}^T).$$

- The following is well-known:

  Theorem 2 (Cochran's theorem)**.** *Let $Z \sim \mathcal{N}(0, \sigma^2 I_n)$, $V \subset \mathbb{R}^n$ and P_V the orthogonal projection matrix onto V. Then, the random vectors $P_V Z$ and $P_{V^\perp} Z = (I_n - P_V)Z$ are independent, $P_V Z \sim \mathcal{N}(0, \sigma^2 P_V)$, $P_{V^\perp} Z \sim \mathcal{N}(0, \sigma^2 P_{V^\perp})$, and*

$$\frac{\|P_V Z\|^2}{\sigma^2} \sim \chi^2(\dim(V)).$$

[1] This entails that the law is well-defined, and indeed that $\Sigma = \boldsymbol{A}\boldsymbol{A}^T$ for some matrix A.

3.3 Statistical properties and law of estimators

- Under H_C we can improve a bit Gauss-Markov's theorem:

  Theorem 3 (Gauss-Markov under H_C)**.** *Assume H_A and H_C. Then $(\hat{\beta}, \hat{\sigma}^2)$ has minimal variance among all the unbiased estimators of (β, σ^2). More formally, if $\hat{C}$ is the variance matrix of $(\hat{\beta}, \hat{\sigma}^2)$ and $\tilde{C}$ is the variance matrix of another unbiased estimator $(\tilde{\beta}, \tilde{\sigma}^2)$ of (β, σ^2), then $\tilde{C} - \hat{C}$ is a positive definite matrix.*

- We can also determine the distribution (law) of the estimators (this follows from Cochran's theorem, *i.e.* Theorem 2, exercise: prove it)

  Proposition 5 (Distribution of the estimator)**.** *Assume H_A and H_C. Then,*

  1. $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$;
  2. $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$;
  3. $\hat{\beta}$ *and* $\hat{\sigma}^2$ *are independent (!).*

- When σ is not known, the following is also useful to construct confidence intervals (and tests).

  Proposition 6. *Assume H_A and H_C. For all $j = 1, \dots, p$ let*

  $$T_j := \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{j,j}}}.$$

  Then $T_j \sim \mathcal{T}(n-p)$.

  Remark: the proof is rather immediate from the definition of the Student distribution and from Proposition 5 (exercise!).

3.4 Confidence intervals for β_j and for σ^2

- Recall that in this Lecture we always assume that we don't know σ^2. In this case we can immediately obtain:
  - a confidence interval for β_j $(j = 1, \dots, p)$ from Proposition 6; and
  - a confidence interval for σ^2 using Proposition 5.

- That is, we have the following:

  Proposition 7. *Assume H_A and H_C. Then,*

  1. $[\hat{\beta}_j \pm t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{j,j}}]$ *is a two-sided confidence interval for β_j with level $1-\alpha$ (recall that $t_{n-p}(1-\alpha/2)$ is the quantile of order $1-\alpha/2$ of the $\mathcal{T}(n-p)$ distribution).*
  2. $[\frac{(n-p)\hat{\sigma}^2}{c_{n-p}(\alpha/2)}, \frac{(n-p)\hat{\sigma}^2}{c_{n-p}(1-\alpha/2)}]$ *is a confidence interval for σ^2 with level $1-\alpha$ (recall that $c_{n-p}(\gamma)$ is the quantile of order γ of the $\chi^2(n-p)$ distribution).*

- Bonus:
  - It is also possible to construct a confidence region for the whole parameter β (not only the individual coordinates). As in Lesson 1, such confidence region is in general different from the product of individual CIs for each coordinates because it takes into account the correlation between the $\hat{\beta}_j$'s.
  - Exercise: It is also possible to construct a CI for $\mathbb{E}[Y_{n+1}] = x_{n+1}^T\beta$.

3.5 Bonus: prediction intervals

- Given a new x_{n+1} we want to predict the value of Y_{n+1} (see Section 2.7) and also to derive a prediction interval (which contains Y_{n+1} with probability $1-\alpha$).

  Proposition 8. *Assume H_A and H_C.* $[x_{n+1}^T\hat{\beta} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1+x_{n+1}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}x_{n+1}}]$ *is a prediction interval for Y_{n+1} with confidence level $1-\alpha$.*

- Exercise: prove the previous proposition.

4 Hypothesis testing

4.1 Statement of the problem, motivating example

- Problem: we want to answer a closed question (yes/no).
- Example (ozone O_3):
  1. Is the O_3 concentration affected by the wind speed W ?
  2. is there a nebulosity N effect ?
  3. Is the O_3 concentration affected by W or T ?
- Recall the model:
$$O_3 = \beta_1 + \beta_T \cdot T + \beta_W \cdot W + \beta_N \cdot N + \varepsilon, \qquad \text{(MC)}$$
  where
  - O_3 is the vector of ozone concentrations,
  - T is the vector of temperatures at $12:00$,
  - W is the vector of recorded wind speeds,
  - N is the vector of recorded nebulosities.
- Then we can formulate the previous questions in **mathematical terms**:
  1. Test $H_0 : \beta_W = 0$ (no effect) against $H_1 : \beta_W \neq 0$ (effect).
  2. Test $H_0 : \beta_N = 0$ (no effect) against $H_1 : \beta_N \neq 0$ (effect).
  3. Test $H_0 : \beta_W = 0$ and $\beta_T = 0$ (no effect) against $H_1 : \beta_W \neq 0$ or $\beta_T \neq 0$.

  la question 3 est un test de fischer donc n'apparait pas dans la sortie de summary(lm)
- We remark that in every case, this is equivalent to test the complete model (MC) against a **submodel** $M_0 \subseteq$ MC. Example:
  1. $M_0 : O_3 = \beta_1 + \beta_T \cdot T + \beta_N \cdot N + \varepsilon$ against MC;
  2. $M_0 : O_3 = \beta_1 + \beta_T \cdot T + \beta_W \cdot W + \varepsilon$ against MC;
  3. $M_0 : O_3 = \beta_1 + \beta_T \cdot T + \varepsilon$ against MC.

4.2 F-test between nested models (simple version)

- We want to test **nested models**. model emboité
- That is for a subset $J_0 \subset \{1, \dots, p\}$:
  - $H_0 : \beta_j = 0$ for all $j \notin J_0$
    $\Leftrightarrow$ Model $M_0 : Y = \boldsymbol{X}_{J_0}\beta_{J_0} + \varepsilon$
    $\Leftrightarrow$ Model $M_0 : Y_i = \sum_{j \in J_0} \beta_j x_{i,j} + \varepsilon_i$ for $i = 1, \dots, n$.
  - H_1 : the complete model $Y = \boldsymbol{X}\beta + \varepsilon$ is correct.
- Notations:
  - J_0 set of indices of covariables of the model M_0 (*i.e.* all the covariables except the ones we want to test the effect)
  - $\beta_{J_0} = (\beta_j)_{j \in J_0} \in \mathbb{R}^{p_0}$ the vector constructed from β by keeping only the coordinates with index in J_0 (and $p_0 = |J_0|$).
  - $\boldsymbol{X}_{J_0}$ is the $n \times p_0$ matrix constructed from $\boldsymbol{X}$ by keeping only the columns with indices in J_0.
- Define the adjusted values:
  - In the model M_0: $\hat{Y}_{J_0} = \boldsymbol{X}_{J_0}\hat{\beta}_{J_0}$ where $\hat{\beta}_{J_0} = (\boldsymbol{X}_{J_0}^T \boldsymbol{X}_{J_0})^{-1} \boldsymbol{X}_{J_0}^T Y$
  - The the complete model: $\hat{Y} = \boldsymbol{X}\hat{\beta}$.
- Define also Fisher's F-statistics for testing M_0 against the complete model

$$F := \frac{\|\hat{Y}_{J_0} - \hat{Y}\|^2/(p - p_0)}{\|Y - \hat{Y}\|^2/(n - p)}.$$

Proposition 9. *Assume H_A and H_C. Then*

1. *F follows a $\mathcal{F}(p - p_0, n - p_0)$ distribution.*
2. *The test which rejects H_0 if $F \geq f_{p-p_0,n-p_0}(1 - \alpha)$ has level α, where $f_{p-p_0,n-p_0}(1 - \alpha)$ is the quantile of order $1 - \alpha$ of the $\mathcal{F}(p - p_0, n - p_0)$ distribution (this test is known as the F-test).*

4.3 Example 1: Testing one parameter

- We want to test: $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$;
- That is $J_0 = \{1, \dots, p\} \backslash \{j\}$.
- Exercise: Show that in this case

$$F = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2_{\hat{\beta}_j}},$$

and then this test is equivalent to Student's test.

4.4 Example 2: Global F-test

- We want to test:
  - H_0: the model $M_0 : Y = \beta_1 + \varepsilon$ (*i.e.* $\beta_2 = \cdots = \beta_p = 0$;
  - H_1: the complete model $M_1 : Y = \boldsymbol{X}\beta + \varepsilon$.
    Equivalently $Y_i = \beta_1 + \sum_{j=2}^{p} \beta_j x_{i,j} + \varepsilon_i$ for $i = 1, \ldots, n$.
- This corresponds to $J_0 = \{1\}$, and thus $\hat{Y}_{J_0} = \bar{y} \cdot \mathbb{1}$. Then,
  This formula is wrong, this must be $(n-p)/(p-1)$ instead of $(p-1)/(n-p)$

$$\begin{aligned} F &= \frac{\|\hat{Y}_{J_0} - \hat{Y}\|^2/(p-p_0)}{\|Y - \hat{Y}\|^2/(n-p)} \\ &= \frac{\|\bar{y} \cdot \mathbb{1} - \hat{Y}\|^2/(p-1)}{\|Y - \hat{Y}\|^2/(n-p)} \\ &= \frac{p-1}{n-p}\frac{SCM}{SCR} \\ &= \frac{p-1}{n-p}\frac{R^2}{1-R^2}. \end{aligned}$$

- Remarks:
  - The test is sometimes called the R^2 test.
  - This is the test performed by R when using $lm()$ (see line F-statistics).

4.5 Bonus: F-test between nested models (general version)

- Remark that we can rewrite the complete model as $Y = m + \varepsilon$, where $m \in \mathrm{Im}(\boldsymbol{X})$.
- So in the general version, we define the complete model as $Y = m + \varepsilon$ with $m \in \Omega$, a subspace of $\mathbb{R}^n$ with $\dim(\Omega) < n$.
- We want to test:
  - H_0 : submodel of Ω, *i.e.* $Y = m + \varepsilon$ with $m \in \omega \subset \Omega$; against
  - H_1 : the model $m \in \Omega \backslash \omega$ is correct.
- Let P_Ω the orthogonal projection onto Ω, P_ω the orthogonal projection onto ω, $p = \dim(\Omega)$ and $p_0 = \dim(\omega$. The F-statistics is

$$F = \frac{\|P_\omega Y - P_\Omega Y\|^2/(p-p_0)}{\|Y - P_\Omega Y\|^2/(n-p)}.$$

5 Diagnostics

5.1 Goals

- Verify the assumptions we made.
  - Checking H_A is immediate, no need for complex diagnostics (just linear algebra).
  - Checking H_B or H_C is less direct (recall that depending on whether or not we want to do CI/Tests we use H_B or H_C).
    - Goodness of fit, linearity ($\Leftrightarrow \varepsilon_i$ are centered).
    - Homoscedasticity of the ε_i's (*i.e.* same variance).
    - Independence of the ε_i's.
    - If H_C: Normality of the ε_i ?
- Checking the influence of a single observations $\Leftrightarrow$ atypical observations.
- We essentially do analysis of residuals and of the projection matrix $P_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

5.2 Various types of residuals

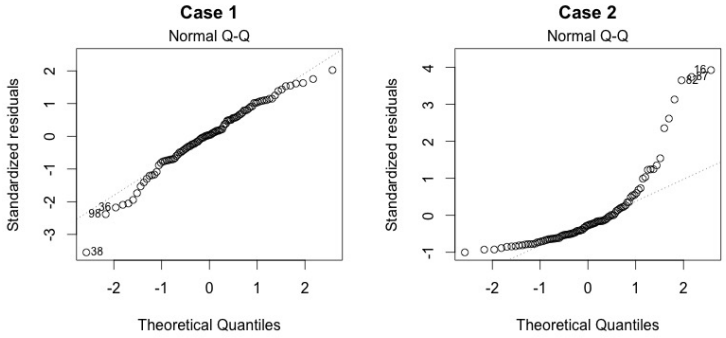
- Recall the definition of the **residuals**. For $i = 1, \dots, n$:
$$\hat{\varepsilon}_i := Y_i - \hat{Y}_i = Y_i - x_i^T\hat{\beta}.$$
- Since $\mathbb{E}[\hat{\varepsilon}_i] = 0$ and $\text{var}(\hat{\varepsilon}_i) = \sigma^2(1 - (P_{\boldsymbol{X}})_{i,i}) =: \sigma^2(1 - h_{i,i})$, the "right scale" to analysis the residuals is $\sigma\sqrt{1-h_{i,i}}$; *i.e.* we expect the residuals to lie within $[-c\sigma\sqrt{1-h_{i,i}}, c\sigma\sqrt{1-h_{i,i}}]$ for $c > 0$ no more than 2 or 3 (except for a few of them). This leads to the definion of **normalized residuals**:
$$r_i := \frac{\hat{\varepsilon}_i}{\sigma\sqrt{1-h_{i,i}}}$$
$\Rightarrow$ Problem: most of time we don't know σ.
- Solution: Define the **standardized residuals**:
$$t_i := \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{i,i}}}$$
$\Rightarrow$ Problem the distribution (law) of r_i is complex because the residuals $\hat{\varepsilon}_i$ are not independent of $\hat{\sigma}$, which makes not convenient the use of the standardized residuals.
- Solution: **Studentized residuals** (aka Jackknifed residuals),
$$t_i^* := \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{i,i}}},$$
suit une loi de student
where $\hat{\sigma}_{(-i)}$ is the estimator of the residual variance computed using only the observations (y_j, x_j) for $j \neq i$ (*i.e.* we remove the observation (y_i, x_i)).
- Summary in Table 1.
- Advice:
  - Favor the studentized residuals (homoscedastic, distribution known).
  - Avoid the $\hat{\varepsilon}_i$ which are naturally heteroscedastic.

Residual	Observed ?	Under H_B	Under H_C
ε_i	no	$\mathbb{E}[\varepsilon_i] = 0$, $\text{var}(\varepsilon_i) = \sigma^2$	$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$
$\hat{\varepsilon}_i$	yes	$\mathbb{E}[\hat{\varepsilon}] = 0$, $\text{var}(\hat{\varepsilon}) = \sigma^2(I_n - P_{\boldsymbol{X}})$ (correlated!)	$\hat{\varepsilon} \sim \mathcal{N}(0, \sigma^2(I_n - P_{\boldsymbol{X}})$
r_i	no	$\mathbb{E}[r_i] = 0$, $\text{var}(r_i) = 1$ but correlated!	$r \sim \mathcal{N}(0, D(I - P_X)D)$
t_i	yes	$\mathbb{E}[t_i] \approx 0$, $\text{var}(t_i) \approx 1$	Exists but complex
t_i^*	yes	$\mathbb{E}[t_i^*] = 0$, $\text{var}(t_i^*) \approx 1$ (but correlated)	$t_i^* \sim \mathcal{T}(n - p - 1)$

Table 1: Summary of the different residuals. Here $D = \text{diag}(1/(1 - h_{i,i})$.

5.3 Normality (if H_C)

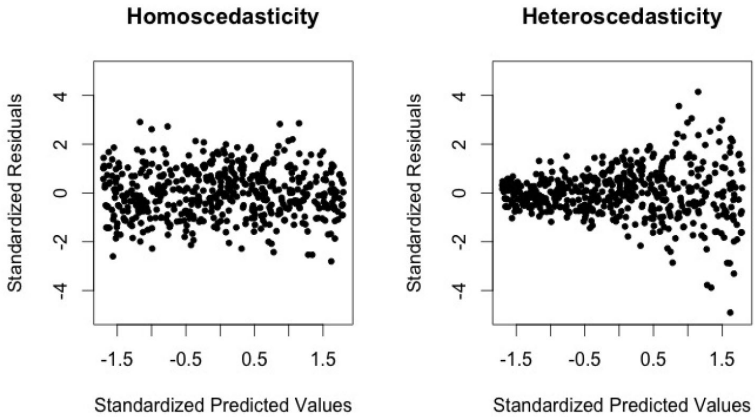
- First diagnostic, QQPlot: Plot the empirical quantiles of the residuals versus the quantile of the $\mathcal{N}(0, 1)$ distribution



- Second diagnostic: Normality test (Shapiro-Wilks, Kolmogorov, etc...)
- What if the diagnostic fails ? (*i.e.* we suspect non normality of the ε_i)
  - The linear model is quite robust to mispecification as long as the law of ε_i is symmetric.
  - Normality necessary only for confidence intervals and tests (+ there exist methods for confidence intervals without normality assumptions, see for instance bootstrap)
  - Sometimes transforming Y can solve the issue (e.g. $\log(Y)$).

5.4 Homoscedasticity

- No precise procedure.
- Various plots possible:
  - t_i^* or $|t_i^*|$ as a function of i
  - t_i^* or $|t_i^*|$ as a function of date of observation i
  - t_i^* or $|t_i^*|$ as a function of $\hat{y}_i$
  - t_i^* or $|t_i^*|$ as a function of $x_{i,j}$ for each $j = 1, \ldots, p$
- Issue if we observe a trend in t_i^*:



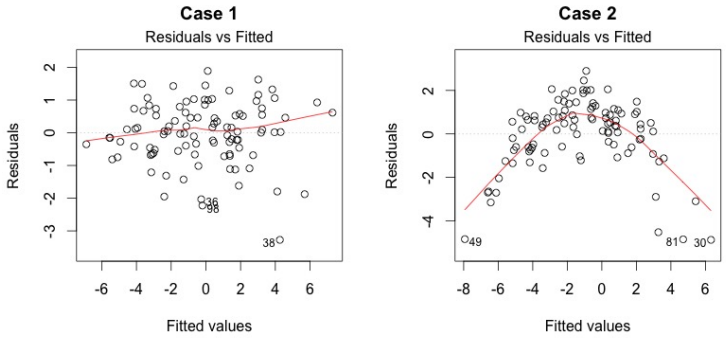
- Some tests are possible, but this requires to have a model for the cause of the heteroscedasticity.
- What to do if we diagnose heteroscedasticity ?
  - $\hat{\beta}$ remains unbiased (but the formula for the variance are wrong)
  - We cannot trust confidence intervals and tests.
  - Transformations of Y or $\boldsymbol{X}$ might solve the issue.
  - Not easy to diagnose heteroscedasticity in general.

5.5 Trend in residuals

We shall observe no trend/structure in the residuals. If we do, this can be caused by various reasons, which we can diagnose using different diagnostics.

5.5.1 Structure due to incomplete or bad modeling

- Missing variable ? Or relation is not exactly linear (*i.e.* need to transform some of the covariables) ?
- Can eventually be seen on the residual plots:



- We might obtain a finer analysis by plotting the **partial residuals**

$$\hat{\varepsilon}^{(j)}_{partial} = \hat{\varepsilon} + \hat{\beta}_j \begin{pmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{pmatrix}$$

- Some intuition about partial residuals:
  - We want to know if the j-th covariate has a **linear** effect on the response Y, or is the effect is non linear.
  - Assuming we don't know yet if the effect is linear, we can assume the following model for Y:

$$Y_i = \sum_{k \neq j} x_{i,k}\beta_k + \psi(x_{i,j}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

    where ψ is some function. Then our goal is to check if ψ is linear or not.
  - One possibility would be to plot the scatter plot $(x_{i,j}, \psi(x_{i,j}))$ to get an idea of what ψ looks like. However, we know what are the $x_{i,j}$'s, but we don't know what is ψ. Nevertheless, ψ can be estimated since:

$$\psi(x_{i,j}) = Y_i - \sum_{k \neq j} x_{i,k}\beta_k - \varepsilon_i.$$

  - From the previous that $\hat{\varepsilon}^{(j)}_{partial}$ is a an estimator of the vector $(\psi(x_{1,j}), \ldots, \psi(x_{n,j}))^T$.
- Exists other methods (partial regression...)

5.5.2 Temporal or spatial structure

- Visually: Can be seen by plotting t_i^* as a function of i or date or location
- Example: see Figure 4.5 in CML for an example of spatial structure.
- Autocorrelation of the residuals:
  - Durbin-Watson test:
  - H_0: the ε_i's are independent; against
  - H_1: the ε_i's follow an $AR(1)$ process.
- In general difficult to distinguish between temporal/spatial structure and modeling issues (see for instance Figure 4.4 in CML: autocorrelated noise versus missing variable give the same residual plot).

5.6 Influential observations

Questions:

- is the individual observation (y_i, x_i) well explained by the model ?

- Does this observation play an important role in the fit ? (in particular on its own adjusted value $\hat{y}_i$?)
- Does this observation influence (too much) the value of $\hat{\beta}$?[2]

5.6.1 Outliers

- An outlier is an observation (y_i, x_i) whose residual $\hat{\varepsilon}_i$ is abnormally large (compared to σ^2), and thus badly explained by the model.
- Formally, (y_i, x_i) is an **outlier** if calcul des résidus studentisés

$$|t_i^*| > t_{n-p-1}(1-\alpha/2)$$

for α well chosen (say $\alpha = 0.05$).

- Remark: If the model is correct, we expect to observe $\alpha \times n$ outliers in a n-sample observation, so there is an issue only of we have more than $\alpha \times n$ outliers.
- What to do if we detect too many outliers ?
  - It is important to understand why (measurement errors? ...)
  - If we understand the cause, we may eventually remove them.
  - Or we can keep them but check that those observations do not influence too much the estimators $\hat{\beta}$ and $\hat{\sigma}^2$.

5.6.2 Leverage les points leviers

- We analyze the projection matrix $P_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.
  For simplicity we write $h_{i,j} = (P_{\boldsymbol{X}})_{i,j}$.
- $h_{i,i}$ is called **leverage** of the observation (y_i, x_i).
- The observation (y_i, x_i) is called a **leverage point** if $h_i > \frac{2p}{n}$ (or $\frac{3p}{n}$ depending on the authors). The Fig. 1 illustrates what is a leverage point in the context of simple linear regression ($p = 2$).
- The intuition is that $\hat{y}_i = h_{i,i}y_i + \sum_{j\neq i} h_{i,j}y_j$. Then, when $h_{i,i}$ is large, the observation (y_i, x_i) plays an important role in its own adjusted value $\hat{y}_i$ (*i.e* $\hat{y}_i$ is really influenced by y_i).
- $\text{var}(\hat{y}_i) = \sigma^2 h_{i,i}$ so $h_{i,i}$ also measures the variability of $\hat{y}_i$.
- Why $\frac{2p}{n}$? In average $\frac{1}{n}\sum_{i=1}^n h_{i,i} = \frac{\text{Tr}(P_{\boldsymbol{X}})}{n} = \frac{p}{n}$, so that $\frac{2p}{n}$ is considered as a large value.
- As for the outliers, it is important to identify the leverage points and to understand why they are leverage points.

[2] Intuitively, if the model is good, a single observation should not influence too much the estimation.

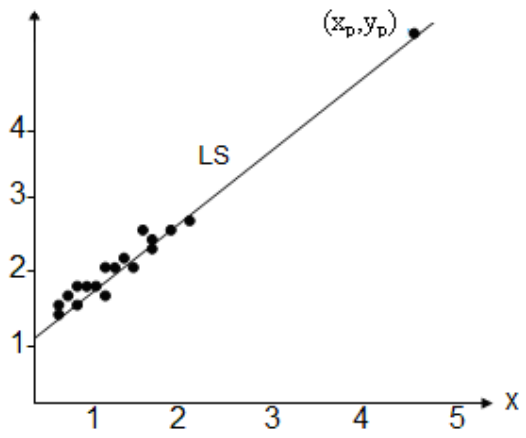


Figure 1: Example of a leverage point. We can see that the point (x_p, y_p) is highly influential in estimating $\hat{\beta}$: a small change in the observed value of y_p will induce a large change in the value of $\hat{\beta}$. Remark that in this case, the leverage point is not an outlier (outlier $\neq$ leverage point, but there can be points that are both outliers and leverage points).

5.6.3 Cook's distance

- Cook's distance naturally combine the two concepts of outliers and leverage points and measures the influence of a single observation (y_i, x_i) on the global fit.
- **Definition**:
$$C_i = \frac{\|\hat{Y} - \boldsymbol{X}\hat{\beta}_{(-i)}\|^2}{p\hat{\sigma}^2} = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})^T \boldsymbol{X}^T \boldsymbol{X})(\hat{\beta}_{(-i)} - \hat{\beta})}{p\hat{\sigma}^2},$$
where $\hat{\beta}_{(-i)}$ is the estimator obtained by removing the observation (y_i, x_i) from the data.
- We can show that
$$C_i = \frac{h_{i,i}}{p(1-h_{i,i})^2} \cdot \frac{\hat{\varepsilon}_i^2}{\hat{\sigma}^2}.$$
Thus,
  - it is not necessary to compute $\hat{\beta}_{(-i)}$ to compute C_i (useful);
  - We see immediately that C_i gets large if $h_{i,i}$ is large and/or $|t_i|$ is large (compromise between outlier and leverage).
- What should we consider as a large value for C_i ? Usage: $C_i > 1$ problematic.

5.7 Conclusion

- Important to know how to perform the diagnostics quickly (because we need to do them every time we remove an outlier and/or change the model, *i.e.* transform some variables, add more variables, etc...)
- Critical before any interpretation of the results.

- We note that there are plenty of other diagnostics not discussed here (we refer to CML for details).