General Linear Model - "Theoretical Concepts"

Cathy Maugis-Rabusseau

INSA Toulouse / IMT
GMM 116
cathy.maugis@insa-toulouse.fr

2023 - 2024

Outline

Outline

1. **Chapter 2: General definitions**
   - Regular linear model
   - Examples of Gaussian linear model

Linear model

Definition

Let $Y = (Y_1, \ldots, Y_i, \ldots, Y_n)'$ be a response variable. A **linear model** is defined by

$$\underset{n\times 1}{Y} = \underset{n\times k}{X}\underset{k\times 1}{\theta} + \underset{n\times 1}{\varepsilon},$$

where

- X is a n rows $\times$ k columns matrix with $\boldsymbol{k} < \boldsymbol{n}$, $X \in \mathcal{M}_{n,k}(\mathbb{R})$,
- θ is an unknown vector of size k,
- $\varepsilon \in \mathbb{R}^n$ is the vector of errors

Remark: k is linked to the number p of explanatory variables. For instance, $k = 1 + p$ for a linear regression model with an intercept.

Regular linear model

Definition

A linear model $Y = X\theta + \varepsilon$ is **regular** if the matrix X is regular, i.e the rank of X is equal to k.
Otherwise ($\text{rg}(X) = r < k$), the model is **singular**.

Proposition

Let $X \in \mathcal{M}_{n,k}(\mathbb{R})$. The following propositions are equivalent:

- X is a matrix of rank k.
- The application $X : \mathbb{R}^k \to \mathbb{R}^n$ is injective.
- The matrix $X'X$ is invertible.

Summary

$X'X$ is invertible if and only if the model is regular.

Regular linear model

$0_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \updownarrow n$

Remark

If X is regular then, by injectivity of X :

$$X\theta = 0_n \Rightarrow \theta = 0_k \text{ for all } \theta \in \mathbb{R}^k.$$

This property ensures that the columns of X are linearly independent in $\mathbb{R}^n$ and guarantees the uniqueness of θ.

$X\theta = X\tilde{\theta} \Rightarrow \theta = \tilde{\theta}$

In some cases, the matrix X cannot be regular. However, we will see that it is sometimes possible to overcome this problem by adding identifiability constraints on the parameters to be estimated.

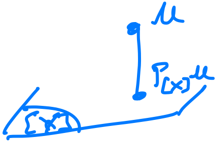
Unless explicitly mentioned, the matrix X will be assumed to be regular in the sequel.

Proposition

Let $X \in \mathcal{M}_{n,k}(\mathbb{R})$ be a regular matrix. Then the projection matrix on $[X] := Im(X)$ is $P_{[X]} = X(X'X)^{-1}X'$.

This matrix $P_{[X]}$, often denoted H, is called the **Hat Matrix**.

Proof:



- $\forall u \in \mathbb{R}^n, u = \underbrace{P_{[X]}u}_{\in[X]} + \underbrace{u - P_{[X]}u}_{\in [X]^{\perp}\ ?}$
- We show that $u - P_{[X]}u \in [X]^{\perp}$: $\forall v \in \mathbb{R}^k$,

$$\langle Xv, u - P_{[X]}u \rangle$$

$$\begin{aligned} = \underbrace{(Xv)}_{\in[X]}'(u - P_{[X]}u) &= v'X'(u - X(X'X)^{-1}X'u) \\ &= v'X'u - v'\underbrace{(X'X)(X'X)^{-1}}_{=I}X'u = 0 \end{aligned}$$

Hypotheses on ε

déterministe

Aléatoire.

$$Y = X\theta + \varepsilon$$

- **Hypothesis H1** : The errors are centered $\mathbb{E}[\varepsilon] = 0_n$.

  This hypothesis ensures that

$$\mathbb{E}[Y] = X\theta = \sum_{j=1}^{k} \theta_j X^{(j)}$$

  where $X^{(j)}$ is the j-th column of X.

  Y is on average a **linear** combination of the $X^{(j)}$ $\Rightarrow$ **linear** model.

Hypotheses on ε

- **Hypothesis H2** : The variance of errors is constant:

$$\mathbb{E}[\varepsilon_i^2] = \mathrm{Var}(\varepsilon_i) = \sigma^2, \forall i = 1, \ldots, n$$

where σ^2 is an unknown parameter to be estimated.

This hypothesis implies that

$$Var(Y_i) = Var((X\theta)_i + \varepsilon_i)$$
$$= Var(\varepsilon_i)$$

$$\forall i = 1, \cdots, n, \; Var(Y_i) = \sigma^2$$

It is often reasonable to assume that **H2** is true.
If **H2** is not satisfied, it is possible to set up a statistical treatment of the linear model ... this however requires much more work.

Hypotheses on ε

- **Hypothesis H3** : The variables ε_i are independent.
- **Hypothesis H4** : The errors follow a Gaussian law:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \forall i \in \{1, \cdots, n\}$$

The hypotheses **H1-H4** imply that:

$$Y = X\theta + \varepsilon$$

$$Y \sim \mathcal{N}_n\left(X\theta, \sigma^2 I_n\right)$$

$$\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n).$$

Hypotheses on ε

The assumption of normality of errors can be justified:

- A theoretical argument: the ε_i can be characterized as measurement errors. According to the Central Limit Theorem, if all these effects are independent with the same zero mean and the same "small" variance, their sum tends towards a Gaussian variable.
  The Gaussian distribution models fairly well all the situations where a fluctuation is the result of several independent causes.

- A practical argument: it is easy to control if a random variable follows a Gaussian law. By studying a posteriori the distribution of the calculated errors (residuals) and comparing it to the theoretical (Gaussian) distribution, it is often observed that it can be considered as approaching Gaussian law.

Summary

Summary

- Linear model:

$$Y = X\theta + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$$

with $Y \in \mathbb{R}^n$, $X \in \mathcal{M}_{n,k}(\mathbb{R})$, $\theta \in \mathbb{R}^k$, $\varepsilon \in \mathbb{R}^n$

- Regular model if $\mathrm{rg}(X) = k$, otherwise it is singular.
- Regular model $\Leftrightarrow$ X injective $\Leftrightarrow$ $X'X$ invertible
- Orthogonal projection matrix on $[X] = Im(X)$: $P_{[X]} = X(X'X)^{-1}X'$

Outline

1. **Chapter 2: General definitions**
    - Regular linear model
    - Examples of Gaussian linear model

Linear regression

- Goal: Explain a quantitative response variable Y by several quantitative explanatory variables $x^{(1)}, \cdots, x^{(p)}$.
- Linear regression model:

$$Y_i = \underbrace{\boxed{\theta_0 + \theta_1 x_i^{(1)} + \cdots + \theta_p x_i^{(p)}}}_{\text{déterministe .}} + \varepsilon_i,$$

with

  - $\theta_0, \theta_1, \cdots, \theta_p$ unknown parameters
  - $\varepsilon_1, \cdots, \varepsilon_n$ i.i.d $\mathcal{N}(0, \sigma^2)$, σ^2 to be estimated.

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(1)} & \dots & x_1^{(j)} & \dots & x_1^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_i^{(1)} & \dots & x_i^{(j)} & \dots & x_i^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & \dots & x_n^{(j)} & \dots & x_n^{(p)} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

One-way ANOVA

- Goal: Explain a quantitative response variable Y by one qualitative (categorical) explanatory variable (called factor) with I modalities.

- One-way ANOVA model:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \text{ for } i = 1, \cdots, I\,;\ j = 1, \cdots, n_i,$$

  with

  - $\mu_1, \cdots, \mu_I$ unknown parameters
  - $\varepsilon_{11}, \cdots, \varepsilon_{In_I}$ i.i.d $\mathcal{N}(0, \sigma^2)$ with σ^2 to be estimated.

  In order to write this model matricially, the observations are arranged by factor modality (level):

$$Y = (\underbrace{Y_{1\,1}, \cdots, Y_{1\,n_1}}_{[Y]_1}, \underbrace{Y_{2\,1}, \cdots, Y_{2n_2}}_{[Y]_2}, \cdots, \underbrace{Y_{I\,1}, \cdots, Y_{I\,n_I}}_{[Y]_I})'$$

One-way ANOVA

$$Y_{ij} = \mu_i + \varepsilon_{ij} \text{ for } i = 1, \cdots, I\,;\, j = 1, \cdots, n_i,$$

$$\Leftrightarrow$$

$$Y = \begin{pmatrix} [Y]_1 \\ \vdots \\ [Y]_I \end{pmatrix} = \begin{pmatrix} \mathbb{1}_{n_1} & 0_{n_1} & \cdots & \cdots & 0_{n_1} \\ 0_{n_2} & \boxed{\mathbb{1}_{n_2}} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{n_I} & 0_{n_I} & \cdots & 0_{n_I} & \mathbb{1}_{n_I} \end{pmatrix} \overbrace{\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix}}^{\theta}_{k=I} + \begin{pmatrix} [\varepsilon]_1 \\ \vdots \\ [\varepsilon]_I \end{pmatrix}$$

Remarks :

- $k = I$
- X is full rank so the model is regular.

One-way ANOVA

If we consider the following model:

$$Y_{ij} = \alpha + \mu_i + \varepsilon_{ij} \text{ for } i = 1, \cdots, I \,;\, j = 1, \cdots, n_i,$$

$$\Leftrightarrow$$

$$Y = \begin{pmatrix} [Y]_1 \\ \vdots \\ [Y]_I \end{pmatrix} = \begin{pmatrix} \mathbb{1}_{n_1} & \mathbb{1}_{n_1} & 0_{n_1} & \cdots & \cdots & 0_{n_1} \\ \mathbb{1}_{n_2} & 0_{n_2} & \mathbb{1}_{n_2} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbb{1}_{n_I} & 0_{n_I} & 0_{n_I} & \cdots & \cdots & \mathbb{1}_{n_I} \end{pmatrix} \begin{pmatrix} \alpha \\ \mu_1 \\ \vdots \\ \mu_I \end{pmatrix} + \begin{pmatrix} [\varepsilon]_1 \\ \vdots \\ [\varepsilon]_I \end{pmatrix}$$

somme

Remarks :

- $k = I + 1$
- X is no longer full rank ($r = I < k$) so the model is singular.

Outline

Context and goals

- Model :

$$\begin{cases} Y = X\theta + \varepsilon \\ \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n) \end{cases}$$

  The model is assumed to be regular.

- Goals:
  - Estimate the unknown parameters $\theta \in \mathbb{R}^k$ and $\sigma^2 > 0$
  - Study the law of the estimators
  - Deduce confidence intervals and predictions

Outline

2 Chapter 3: Parameter estimation

- Estimation of θ
- Adjusted values and residuals
- Estimator of σ^2
- Standard errors of $\widehat{\theta}_j$, $\widehat{Y_i}$, $\widehat{\varepsilon}_i$
- Confidence Intervals
- Prediction Interval
- Measure for goodness-of-fit

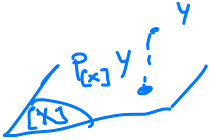
Least squares estimator of θ

Least squares estimator (*Estimateur des moindres carrés* EMC) of θ:

$$\begin{aligned}\widehat{\theta} &= \arg\min_{\vartheta} \underbrace{\|Y - X\vartheta\|^2}_{\sum_{i=1}^{n} (Y_i - (X\theta)_i)^2} \\ &= \arg\min_{\vartheta} SSR(\vartheta) \\ &= \arg\min_{\vartheta} (Y - X\vartheta)'(Y - X\vartheta)\end{aligned}$$

$$\min_{\theta} \|Y - X\theta\|^2 = \min_{u \in [X]} \|Y - u\|^2$$

$$= \|Y - P_{[X]}Y\|^2$$



$$X\hat{\theta} = P_{[X]}Y = X(X'X)^{-1}X'Y$$

X régulière donc injectivité

$$\Rightarrow \hat{\theta} = (X'X)^{-1}X'Y.$$

Least squares estimator of θ

Theorem

- Regular linear model: $Y = X\theta + \varepsilon$.
- The least squares estimator $\widehat{\theta}$ is defined by

$$\widehat{\theta} = (X'X)^{-1}X'Y.$$

The least squares estimator $\widehat{\theta}$ satisfies the following property:

$$X\widehat{\theta} = P_{[X]}Y.$$

Remark: When the errors are Gaussian, the least squares estimator $\widehat{\theta}$ exactly corresponds to the maximum likelihood estimator.

Theorem

In the framework of a regular Gaussian linear model,

$$\widehat{\theta} \sim \mathcal{N}_k\left(\theta, \sigma^2(X'X)^{-1}\right).$$

Proof:

- $Y = X\theta + \varepsilon$ and $\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$ thus $Y \sim \mathcal{N}_n(X\theta, \sigma^2 I_n)$
- $\widehat{\theta} = \underbrace{(X'X)^{-1}X'}_{=A\in\mathcal{M}_{k,n}(\mathbb{R})} Y \sim \mathcal{N}_k(AX\theta, \sigma^2 AA')$
- $\mathbb{E}[\widehat{\theta}] = AX\theta = \cancel{(X'X)^{-1}}\cancel{X'X}\theta = \theta$ $\longrightarrow$ $\hat{\theta}$ est un estimateur sans biais de θ
- $\text{Var}(\widehat{\theta}) = \sigma^2 AA' = \sigma^2\cancel{(X'X)^{-1}}\cancel{X'X}(X'X)^{-1} = \sigma^2(X'X)^{-1}$

Outline

2. **Chapter 3: Parameter estimation**
   - Estimation of θ
   - Adjusted values and residuals
   - Estimator of σ^2
   - Standard errors of $\widehat{\theta}_j$, $\widehat{Y_i}$, $\widehat{\varepsilon}_i$
   - Confidence Intervals
   - Prediction Interval
   - Measure for goodness-of-fit

Definitions

- For each Y_i, we obtain an **adjusted (predicted) value** $\widehat{Y}_i$ by the adjusted model:

$$\widehat{Y} = (\widehat{Y}_1, \cdots, \widehat{Y}_n)' = X\widehat{\theta} = X(X'X)^{-1}X'Y = P_{[X]}Y.$$

(projection of Y on $Im(X)$)

- The errors ε_i are estimated by the **residuals**:

$$\hat{\varepsilon} = Y - \widehat{Y} = (I_n - P_{[X]})Y = P_{[X]^\perp}Y$$

Properties

Proposition

- $\widehat{Y} \sim \mathcal{N}_n\left(X\theta, \sigma^2 P_{[X]}\right)$ where $P_{[X]} = X(X'X)^{-1}X'$
- $\widehat{\varepsilon} \sim \mathcal{N}_n\left(0_n, \sigma^2(I_n - P_{[X]})\right)$
- The random variables $\widehat{Y}$ and $\widehat{\varepsilon}$ are independent.
- The random variables $\widehat{\theta}$ and $\widehat{\varepsilon}$ are independent.

Proof in course

Loi de $\hat{Y} = X\hat{\theta} = P_{[X]} Y$

$$\hat{\theta} \sim \mathcal{N}_k\left(\theta, \sigma^2 (X'X)^{-1}\right)$$

donc $X\hat{\theta} \sim \mathcal{N}_n\left(X\theta, \sigma^2 \underbrace{X(X'X)^{-1}X'}_{P_{[X]}}\right)$

Loi de $\hat{\varepsilon} = Y - \hat{Y} = P_{[X]^\perp} Y$

$$Y \sim \mathcal{N}_n(X\theta, \sigma^2 I_n)$$

$$\hat{\varepsilon} = P_{[X]^\perp} Y \sim \mathcal{N}_n(\blacksquare, \blacktriangle)$$

$$\begin{aligned}
\blacksquare &= P_{[X]^\perp} X\theta \\
&= \left(I_n - X(X'X)^{-1}X'\right) X\theta \\
&= X\theta - X\cancel{(X'X)^{-1}}\,\cancel{X'X}\,\theta = 0_n
\end{aligned}$$

$$\begin{aligned}
\blacktriangle &= \sigma^2 P_{[X]^\perp} I_n P_{[X]^\perp}' \\
&= \sigma^2 P_{[X]^\perp} = \sigma^2\left(I_n - X(X'X)^{-1}X'\right)
\end{aligned}$$

$\hat{Y} \perp\!\!\!\perp \hat{\varepsilon}$? $\qquad \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$

$$\begin{aligned}
\hat{Y} &= P_{[X]} Y = P_{[X]}(X\theta + \varepsilon) \\
&= X\theta + P_{[X]}\varepsilon
\end{aligned}$$

$$\hat{\varepsilon} = P_{[X]^\perp} Y = P_{[X]^\perp}(X\theta + \varepsilon) = P_{[X]^\perp}\varepsilon$$

$\Rightarrow$ Cochran $\Rightarrow \hat{\varepsilon} \perp\!\!\!\perp \hat{Y}$

$\hat{\theta} \perp\!\!\!\perp \hat{\varepsilon}$?

$$\hat{\theta} = (X'X)^{-1}X'X\hat{\theta} = g(X\hat{\theta}) = g(\hat{Y})$$

$\hat{Y} \perp\!\!\!\perp \hat{\varepsilon}$

$\Rightarrow \quad \hat{\theta} \perp\!\!\!\perp \hat{\varepsilon}$.

Outline

2 Chapter 3: Parameter estimation

- Estimation of θ
- Adjusted values and residuals
- Estimator of σ^2
- Standard errors of $\widehat{\theta}_j$, $\widehat{Y_i}$, $\widehat{\varepsilon}_i$
- Confidence Intervals
- Prediction Interval
- Measure for goodness-of-fit

Estimator of σ^2

Theorem

Let $\widehat{\theta}$ be the least squares estimator of θ.
Under the hypotheses H1-H4, and if $X \in \mathcal{M}_{nk}(\mathbb{R})$, then

$$= \sum_{i=1}^{n} (\widehat{\varepsilon}_i)^2$$

$$\widehat{\sigma^2} = \frac{\|\widehat{\varepsilon}\|^2}{n-k} = \frac{\|Y - \widehat{Y}\|^2}{n-k} = \frac{\|Y - X\widehat{\theta}\|^2}{n-k} = \frac{SSR(\widehat{\theta})}{n-k}$$

is an unbiased estimator of σ^2, independent of $\widehat{\theta}$.
Moreover,

$$\frac{(n-k)\,\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-k).$$

Proof in course

Estimateur de σ^2

$\varepsilon_1 - - - \varepsilon_n \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

$$\frac{1}{(*)} \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2 = \frac{1}{(*)} \| Y - X\hat{\theta} \|^2 = \frac{SSR(\hat{\theta})}{(*)}$$

$$SSR(\hat{\theta}) = \| Y - X\hat{\theta} \|^2$$

$$= \| Y - P_{[X]} Y \|^2$$

$P_{[X]^\perp} Y = P_{[X]^\perp}(X\theta + \varepsilon)$

$$= \| P_{[X]^\perp} Y \|^2 = \| P_{[X]^\perp} \varepsilon \|^2$$

$\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$

Th de Cochran $\quad \| P_{[X]^\perp} \varepsilon \|^2 \sim \sigma^2 \chi^2(\dim([X]^\perp))$

$$\dim([X]^\perp) = \dim(\mathbb{R}^n) - \dim([X])$$
$$= n - k \quad (rg(X) = k \text{ régulier})$$

$$\mathbb{E}[SSR(\hat{\theta})] = \sigma^2 (n-k)$$

$\iff \mathbb{E}\left[\frac{SSR(\hat{\theta})}{\underbrace{n-k}_{=(*)}} \right] = \sigma^2 \longrightarrow$ estim. sans biais

$\hat{\sigma}^2 \perp\!\!\!\perp \hat{\theta}$?

$$\hat{\sigma}^2 = \frac{\| P_{[X]^\perp} \varepsilon \|^2}{n-k} \qquad X\hat{\theta} = X\theta + P_{[X]} \varepsilon$$

$$\hat{\sigma}^2 = g\left(P_{[X]^\perp}\varepsilon\right) \qquad X\hat{\theta} = f\left(P_{[X]}\varepsilon\right)$$

$\Rightarrow$ Cochran $\Rightarrow$ $\hat{\sigma}^2 \perp\!\!\!\perp X\hat{\theta}$

Puis $\hat{\theta} = (X'X)^{-1}X' \underbrace{X\hat{\theta}}$

donc $\hat{\theta} \perp\!\!\!\perp \hat{\sigma}^2$.

Outline

2. **Chapter 3: Parameter estimation**
   - Estimation of θ
   - Adjusted values and residuals
   - Estimator of σ^2
   - Standard errors of $\widehat{\theta}_j$, $\widehat{Y_i}$, $\widehat{\varepsilon}_i$
   - Confidence Intervals
   - Prediction Interval
   - Measure for goodness-of-fit

Standard errors

- $\Gamma_{\widehat{\theta}} = \sigma^2(X'X)^{-1}$ is estimated by $\widehat{\Gamma}_{\widehat{\theta}} = \widehat{\sigma}^2(X'X)^{-1}$
$\Longrightarrow$ standard error of $\widehat{\theta}_j$ is $se_j = \sqrt{\widehat{\sigma}^2[(X'X)^{-1}]_{jj}}$

- $Var(\widehat{Y}) = \sigma^2 \ P_{[X]} = \sigma^2 X(X'X)^{-1}X'$ is estimated by $\widehat{\sigma}^2 P_{[X]}$
$\Longrightarrow$ standard error of $\widehat{Y_i}$ is $\sqrt{\widehat{\sigma}^2(P_{[X]})_{ii}}$

- Standard error of $\widehat{\varepsilon}_i$ is $\sqrt{\widehat{\sigma}^2[1-(P_{[X]})_{ii}]}$

- Standardized residual $= \dfrac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2}}$

- Studentized residual $= \dfrac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2[1-(P_{[X]})_{ii}]}}$

Outline

2 Chapter 3: Parameter estimation

$IC_{1-\alpha}(\theta_j)$

- $\widehat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2(X'X)^{-1})$ thus $\widehat{\theta}_j \sim \mathcal{N}(\theta_j, \underbrace{\sigma^2[(X'X)^{-1}]_{jj}}_{\sigma^2 v_j})$

$\dfrac{\hat\theta_j - \theta_j}{\sqrt{\sigma^2 v_j}} \sim \mathcal{N}(0,1)$

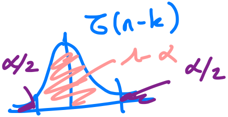
$\cancel{((X'X)^{-1})_{jj}}$

donc

- $(n-k)\ \widehat{\sigma^2} \sim \sigma^2\chi^2(n-k)$

$\hat\theta \perp\!\!\!\perp \hat\sigma^2$ donc

- According to Cochran's theorem, $\widehat{\theta}_j$ and $\widehat{\sigma^2}$ are independent



$$\Longrightarrow \frac{\widehat{\theta}_j - \theta_j}{\sqrt{\widehat{\sigma^2}[(X'X)^{-1}]_{jj}}} \sim \mathcal{T}(n-k).$$

- Let $t_{1-\frac{\alpha}{2}}$ be the $(1-\alpha/2)$-quantile of the Student law with $(n-k)$ df.

Then $\mathbb{P}\left(\left| \dfrac{\hat\theta_j - \theta_j}{\sqrt{\hat\sigma^2 v_j}} \right| \leq t_{1-\alpha/2} \right) = 1-\alpha \iff \mathbb{P}(-t \leq \dots \leq t) = 1-\alpha$

$$IC_{1-\alpha}(\theta_j) = \left[\ \widehat{\theta}_j \pm t_{1-\frac{\alpha}{2}} \sqrt{\widehat{\sigma}^2[(X'X)^{-1}]_{jj}}\ \right] = \left[\widehat{\theta}_j \pm t_{1-\frac{\alpha}{2}} se_j\right].$$

$$IC_{1-\alpha}(\theta_j) = ?$$

- $\hat{\theta} = (X'X)^{-1}X'Y$ estimateur de θ

donc $\hat{\theta}_j$ estimateur de θ

- $\hat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2 (X'X)^{-1})$

donc $\hat{\theta}_j \sim \mathcal{N}(\theta_j, \sigma^2 \underbrace{[(X'X)^{-1}]_{jj}}_{:= s_j^2})$

donc $$\frac{\hat{\theta}_j - \theta_j}{\sqrt{\sigma^2 s_j^2}} \sim \mathcal{N}(0,1).$$

- σ^2 étant un paramètre inconnu, on l'estime par $\hat{\sigma}^2 = \frac{1}{n-k} \| Y - X\hat{\theta} \|^2$.

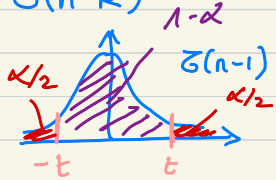
- D'après le théorème de Cochran

$$\frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k)$$

et $\hat{\theta} \perp\!\!\!\perp \hat{\sigma}^2$ donc $\hat{\theta}_j \perp\!\!\!\perp \hat{\sigma}^2$

Ainsi $$\frac{\hat{\theta}_j - \theta_j}{\sqrt{\hat{\sigma}^2 s_j^2}} \sim \mathcal{T}(n-k)$$

- Soit t le $1-\frac{\alpha}{2}$ quantile d'une $\mathcal{T}(n-k)$



car si $T \sim \mathcal{T}(n-k)$

$$\mathbb{P}(|T| \leq t) = 1-\alpha$$

$$\iff \mathbb{P}(T \leq t) = \frac{\alpha}{2} + 1 - \alpha = 1 - \frac{\alpha}{2}$$

- Finalement

$$\mathbb{P}\left(\left| \frac{\hat{\theta}_j - \theta_j}{\sqrt{\hat{\sigma}^2 s_j^2}} \right| \leq t \right) = 1-\alpha$$

$$\iff \mathbb{P}\left(-t \leq \frac{\hat{\theta}_j - \theta_j}{\sqrt{\hat{\sigma}^2 s_j^2}} \leq t \right) = 1-\alpha$$

$$\iff \mathbb{P}\left(\hat{\theta}_j - t\sqrt{\hat{\sigma}^2 s_j^2} \leq \theta_j \leq \hat{\theta}_j + t\sqrt{\hat{\sigma}^2 s_j^2} \right) = 1-\alpha$$

$$\iff \mathbb{P}\left(\theta_j \in \left[\hat{\theta}_j \pm t \sqrt{\hat{\sigma}^2 \left[(X'X)^{-1}\right]_{jj}} \right] \right) = 1-\alpha$$

$IC_{1-\alpha}((X\theta)_i)$

- $\mathbb{E}[Y_i] = (X\theta)_i =$ the average response of Y_i
- $(X\theta)_i$ is estimated by $\widehat{Y_i} = (X\widehat{\theta})_i \sim \mathcal{N}\left((X\theta)_i, \sigma^2[X(X'X)^{-1}X']_{ii}\right)$
- $(n-k)\widehat{\sigma^2} \sim \sigma^2\chi^2(n-k)$
- $\hat{\theta}$ and $\widehat{\sigma^2}$ are independent thus $\widehat{Y_i} \perp\!\!\!\perp \widehat{\sigma^2}$.

$$\Longrightarrow \frac{\widehat{Y}_i - (X\theta)_i}{\sqrt{\widehat{\sigma}^2[X(X'X)^{-1}X']_{ii}}} \sim \mathcal{T}(n-k)$$

$$IC_{1-\alpha}((X\theta)_i) = \left[\widehat{Y_i} \pm t_{n-k,1-\alpha/2} \times \sqrt{\widehat{\sigma}^2[X(X'X)^{-1}X']_{ii}}\right]$$

on estime θ par $\hat{\theta} = (X'X)^{-1}X'Y$

$\Rightarrow$ on estime $(X\theta)_i$ par $(X\hat{\theta})_i$

$\hat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2 (X'X)^{-1})$

donc $X\hat{\theta} \sim \mathcal{N}_n(X\theta, \sigma^2 X(X'X)^{-1}X')$

donc $(X\hat{\theta})_i \sim \mathcal{N}((X\theta)_i, \sigma^2 [P_{[X]}]_{ii})$

donc $\dfrac{(X\hat{\theta})_i - (X\theta)_i}{\sqrt{\sigma^2 (P_{[X]})_{ii}}} \sim \mathcal{N}(0,1)$

Or σ^2 inconnue donc on l'estime

par $\hat{\sigma}^2 = \dfrac{1}{n-k} \| Y - X\hat{\theta} \|^2$

$\dfrac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k)$

$\hat{\sigma}^2 \perp\!\!\!\perp \hat{\theta}$ donc $\hat{\sigma}^2 \perp\!\!\!\perp (X\hat{\theta})_i$

$\Rightarrow \dfrac{(X\hat{\theta})_i - (X\theta)_i}{\sqrt{\hat{\sigma}^2 (P_{[X]})_{ii}}} \sim \mathcal{T}(n-k)$

Soit t le $1-\frac{\alpha}{2}$ quantile d'une $\mathcal{T}(n-k)$

$$\mathbb{P}\left(\left| \quad \right| \leq t \right) = 1 - \alpha$$

$$\iff \mathbb{P}\left(-t \leq \quad \leq t \right) = 1 - \alpha$$

$$\iff \mathbb{P}\left((X\theta)_i \in \left[(X\hat{\theta})_i \pm t \sqrt{\hat{\sigma}^2 (P_{[X]})_{ii}} \right] \right) = 1 - \alpha.$$

$IC_{1-\alpha}(X_0\theta)$

- Let $X_0 \in \mathcal{M}_{1k}(\mathbb{R})$ be a new point and $X_0\theta$ the average response
- $X_0\theta$ is estimated by $\widehat{Y_0} = X_0\widehat{\theta} \sim \mathcal{N}(X_0\theta, \sigma^2 X_0(X'X)^{-1}X_0')$
- $(n-k)\widehat{\sigma^2} \sim \sigma^2\chi(n-k)$
- $\hat{\theta}$ and $\widehat{\sigma^2}$ are independent

$$\Longrightarrow \frac{\widehat{Y_0} - (X_0\theta)}{\sqrt{\widehat{\sigma}^2[X_0(X'X)^{-1}X_0']}} \sim \mathcal{T}(n-k)$$

$$IC_{1-\alpha}(X_0\theta) = \left[\widehat{Y_0} \pm t_{n-k,1-\alpha/2} \times \sqrt{\widehat{\sigma^2} X_0(X'X)^{-1}X_0'}\ \right].$$

Reprise en cours le 18/09

Outline

2. **Chapter 3: Parameter estimation**
   - Estimation of θ
   - Adjusted values and residuals
   - Estimator of σ^2
   - Standard errors of $\widehat{\theta}_j$, $\widehat{Y_i}$, $\widehat{\varepsilon}_i$
   - Confidence Intervals
   - Prediction Interval
   - Measure for goodness-of-fit

Prediction interval

- Important: understand the difference between a confidence interval of $X_0\theta$ and a prediction interval
  - In the first case, we want to predict an average response corresponding to these explanatory variables $X_0\theta$
  - In the second case, we try to predict a new "individual" response value.
- If we want to predict in which interval the result of a new point $X_0 \in \mathcal{M}_{1k}(\mathbb{R})$ will belong, two types of uncertainty must be taken into account:
  - the uncertainty in the estimation of the average response $X_0\theta$
  - the uncertainty of the error term ε_0

Prediction interval

- The response Y_0 associated to a new point X_0:

$$Y_0 = X_0\theta + \varepsilon_0,$$

where $\varepsilon_0 \perp\!\!\!\perp \varepsilon_i$ $(1 \leq i \leq n)$ and $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2) \Rightarrow Y_0 \sim \mathcal{N}(X_0\theta, \sigma^2)$

- The predicted value is

$$\widehat{Y_0} = X_0\widehat{\theta} \sim \mathcal{N}(X_0\theta, \sigma^2 X_0(X'X)^{-1}X_0').$$

- $Y_0 \perp\!\!\!\perp \hat{Y}_0$ (since $\varepsilon_0 \perp\!\!\!\perp \varepsilon_i$) thus

$$Y_0 - \widehat{Y_0} \sim \mathcal{N}\left(0, \sigma^2(1 + X_0(X'X)^{-1}X_0')\right).$$

Prediction interval

- Otherwise,

$$\widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - X\widehat{\theta})^2 \sim \frac{\sigma^2}{n-k}\chi^2(n-k)$$

and since $\widehat{\sigma}^2$ is independent to $\widehat{\theta}$ and ε_0 (since $\varepsilon_0 \perp\!\!\!\perp \varepsilon_i$),

$$\Longrightarrow \frac{Y_0 - \widehat{Y_0}}{\widehat{\sigma}\sqrt{1 + X_0(X'X)^{-1}X_0'}} \sim \mathcal{T}(n-k)$$

- Finally, the prediction interval of the response Y_0 is defined by

$$IC_{1-\alpha}(Y_0) = \left[\widehat{Y_0} \pm t_{n-k,1-\alpha/2} \times \widehat{\sigma}\sqrt{1 + X_0(X'X)^{-1}X_0'}\right].$$

Outline

2. **Chapter 3: Parameter estimation**
   - Estimation of θ
   - Adjusted values and residuals
   - Estimator of σ^2
   - Standard errors of $\widehat{\theta}_j$, $\widehat{Y_i}$, $\widehat{\varepsilon}_i$
   - Confidence Intervals
   - Prediction Interval
   - Measure for goodness-of-fit

Decomposition of the variability

$$SST = SSE + SSR$$

with

- **Total sum of squares**:

$$SST = \|Y - \bar{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- **Explained sum of squares**:

$$SSE = \|\widehat{Y} - \bar{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(\widehat{Y_i} - \overline{Y})^2$$

- **Residual sum of squares**:

$$SSR = \|Y - \widehat{Y}\|^2 = \sum_{i=1}^{n}(\widehat{\varepsilon}_i)^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$$

Coefficient of determination

- According to the least squares criterion used to estimate the parameters, we seek to minimize the Sum of Squares of Residuals (SSR) and therefore to maximize the Explained Sum of Squares (SSE).

- Coefficient of determination = measure for goodness-of-fit

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = \frac{var(\widehat{Y})}{var(Y)} \in [0, 1]$$

- The closer R^2 is to 1, the better the model fits the data

Summary

In the framework of a regular linear model

- $\widehat{\theta} = (X'X)^{-1}X'Y \sim \mathcal{N}_k(\theta, \sigma^2(X'X)^{-1})$
- $\widehat{\sigma^2} = \frac{\|Y - X\widehat{\theta}\|^2}{n-k} \sim \frac{\sigma^2}{n-k}\chi^2(n-k)$
- $\widehat{\theta}$ and $\widehat{\sigma^2}$ are independent
- Know the definitions of the adjusted (predicted) values $\widehat{Y} = X\widehat{\theta} = P_{[X]}Y$ and the residuals $\widehat{\varepsilon} = Y - \widehat{Y}$
- Know how to build
  - a confidence interval of a parameter
  - a confidence interval of an average response
  - a prediction interval
- Variance decomposition

$$\underbrace{\|Y - \bar{Y}\mathbb{1}_n\|^2}_{SST} = \underbrace{\|Y - \widehat{Y}\|^2}_{SSR} + \underbrace{\|\widehat{Y} - \bar{Y}\mathbb{1}_n\|^2}_{SSE}$$

and $R^2 = \frac{SSE}{SST}$.

Outline

Outline

3 **Chapter 4: Fisher's test**

- Submodel - Hypothesis
- Fisher's test
- Confidence Interval (region) of $C\theta$

Goal

- A Gaussian linear model

$$Y = X\theta + \varepsilon, \text{ with } \varepsilon \sim \mathcal{N}_n\left(0_n, \sigma^2 I_n\right)$$

- We want to test the nullity of some components of θ or of some linear combinations of the components of θ: e.g $\theta_j = 0$; $\theta_j = \theta_k = 0$ or $\theta_j = \theta_k$.

- This amounts to comparing a reference model with a reduced or constrained model (called **submodel**). This approach therefore aims to determine whether the model used can be simplified or not.

- Goal : Build a suitable testing procedure

Examples

- Example 1:
  - Simple linear model: $Y_i = a + bx_i + \varepsilon_i$
  - Submodel with a null slope: $Y_i = a + \varepsilon_i$
- Example 2:
  - One-way ANOVA model: $Y_{ij} = \mu_i + \varepsilon_{ij}$
  - Submodel (no factor effect): $Y_{ij} = \mu + \varepsilon_{ij}$

The null hypothesis $\mathcal{H}_0$

To define the null hypothesis, we introduce a matrix $C \in \mathcal{M}_{qk}(\mathbb{R})$ where

- k denotes the number of parameters of the reference model
- q the number of constraints tested ($1 \leq q \leq k$) such that:

$$\mathcal{H}_0 : C\theta = 0_q \text{ with } C \in \mathcal{M}_{qk}(\mathbb{R})$$

The matrix C is assumed to be full rank ($rg(C) = q$).

Example for $k = 3$

- $\mathcal{H}_0 : \theta_2 = 0 \quad C\theta = 0$ with $C = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$ and $q = 1$.
- $\mathcal{H}_0 : \theta_3 = \theta_2 \quad C\theta = 0$ with $C = \begin{pmatrix} 0 & -1 & 1 \end{pmatrix}$ or $C = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}$ and $q = 1$.
- $\mathcal{H}_0 : \theta_3 = \theta_2 = 0 \quad C\theta = 0_2$ with $C = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $q = 2$.

The null hypothesis $\mathcal{H}_0$

- Let Z be a matrix such that

$$Im(Z) \subset Im(X) \text{ and } k_0 = dim(Im(Z)) < k = dim(Im(X))$$

- The model defined by $Y = Z\beta + \varepsilon$ is called a **submodel** of the linear model $Y = X\theta + \varepsilon$.
- Most often, Z is a matrix made up of k_0 columns of X with $k_0 < k$ and β is a vector of length k_0.
- If we consider a general model $Y = R + \varepsilon$, the problem consists of testing

$$\mathcal{H}_0 : R \in Im(Z) \text{ against } \mathcal{H}_1 : R \in Im(X) \backslash Im(Z).$$

Outline

3. **Chapter 4: Fisher's test**
   - Submodel - Hypothesis
   - Fisher's test
   - Confidence Interval (region) of $C\theta$

Test of Fisher-Snedecor

$$(M_0) : Y = Z\beta + \varepsilon \qquad\qquad (M_1) : Y = X\theta + \varepsilon$$

Theorem

- $\mathcal{H}_0 : C\theta = 0_q \,(R \in [Z])$ against $\mathcal{H}_1 : C\theta \neq 0_q \,(R \in [X]\backslash[Z])$
- Test statistics:

$$F = \frac{(SSR_0 - SSR)/(k - k_0)}{SSR/(n-k)} = \frac{\|X\widehat{\theta} - Z\widehat{\beta}\|^2/(k-k_0)}{\|Y - X\widehat{\theta}\|^2/(n-k)}$$

  with $SSR_0 = \|Y - Z\widehat{\beta}\|^2$ and $SSR = \|Y - X\widehat{\theta}\|^2$.

- Under $\mathcal{H}_0$, $F \sim \mathcal{F}(k - k_0, n - k)$
- Rejection zone: $\mathcal{R}_\alpha = \{F > f_{1-\alpha}\}$
  where $f_{1-\alpha}$ is the $(1-\alpha)$-quantile of $F(k-k_0, n-k)$ which ensures that $\mathbb{P}_{H_0}(F > f_{1-\alpha}) = \alpha$

Proof

- $SSR = \|Y - X\widehat{\theta}\|^2 = \|Y - P_{[X]}Y\|^2 = \|P_{[X]^\perp}Y\|^2 = \|P_{[X]^\perp}\varepsilon\|^2$
  $SSR_0 = \|Y - Z\widehat{\beta}\|^2 = \|P_{[Z]^\perp}\varepsilon\|^2$
- Under $\mathcal{H}_0$, $[X]^\perp \subset [Z]^\perp$ thus $A \oplus^\perp [Z] = [X]$ and $dim(A) = k - k_0$
- Using Pythagore, $\|P_{[Z]^\perp}\varepsilon\|^2 = \|P_{[X]^\perp}\varepsilon\|^2 + \|P_A\varepsilon\|^2$ thus
  $SSR_0 - SSR = \|P_A\varepsilon\|^2$
- By Cochran's theorem, since $\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$,

$$SSR =\sim \sigma^2\chi^2(n-k) \text{ and } SSR_0 - SSR \sim \sigma^2\chi^2(k-k_0)$$

- Finally, under $\mathcal{H}_0$,

$$F = \frac{(SSR_0 - SSR)/(k-k_0)}{SSR/(n-k)} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(k-k_0, n-k)$$

- $SSR_0 - SSR = \|P_A Y\|^2 = \|P_{[X]}Y - P_{[Z]}Y\|^2 = \|X\widehat{\theta} - Z\widehat{\beta}\|^2$

Test of Fisher-Snedecor

We can also write the Fisher test statistics under the following form:

$$F = \frac{[C\widehat{\theta}\,]'\left[C(X'X)^{-1}C'\right]^{-1}[C\widehat{\theta}\,]}{q\,\widehat{\sigma^2}} \text{ with } q = k - k_0.$$

This last expression has the advantage of not requiring the estimation of the constrained model to test $\mathcal{H}_0 : C\theta = 0_q$ against $\mathcal{H}_1 : C\theta \neq 0_q$.

Particular case with $q = 1$: Student's Test

- For $q = 1$, the F-test is equivalent to the Student's test
- Construction:
  - $\widehat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2(X'X)^{-1})$ thus $C\widehat{\theta} \sim \mathcal{N}_1(C\theta, \sigma^2 C(X'X)^{-1}C')$
  - $C(X'X)^{-1}C' \in \mathbb{R}$ thus we may divide by this scalar
  - Under $\mathcal{H}_0$,

$$\frac{C\widehat{\theta}}{\sqrt{\sigma^2 C(X'X)^{-1}C'}} \sim \mathcal{N}(0,1) \text{ and } \frac{(n-k)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-k) \text{ and } \widehat{\theta} \perp\!\!\!\perp \widehat{\sigma^2}$$

  - The test statistics is deduced:

$$T := \frac{C\widehat{\theta}}{\sqrt{\widehat{\sigma^2} C(X'X)^{-1}C'}} \underset{\mathcal{H}_0}{\sim} \mathcal{T}(n-k)$$

  - Rejection zone: $\mathcal{R}_\alpha = \{|T| > t_{n-k,1-\alpha/2}\}$ where $t_{n-k,1-\alpha/2}$ is the $(1-\frac{\alpha}{2})$-quantile of $\mathcal{T}(n-k)$

Outline

3 **Chapter 4: Fisher's test**

- Submodel - Hypothesis
- Fisher's test
- Confidence Interval (region) of $C\theta$

$IC_{1-\alpha}(C\theta)$ when $q = 1$

Since $q = 1$, we make the usual construction with a Student law:

- $\widehat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2(X'X)^{-1})$
- $C\widehat{\theta} \sim \mathcal{N}(C\theta, \sigma^2\Delta)$ with $\Delta = C(X'X)^{-1}C' \in \mathbb{R}$
- $(n-k)\widehat{\sigma^2}/\sigma^2 \sim \chi^2(n-k)$ and $\widehat{\theta} \perp\!\!\!\perp \widehat{\sigma^2}$
- Thus

$$\frac{C\widehat{\theta} - C\theta}{\widehat{\sigma}\sqrt{\Delta}} \sim \mathcal{T}(n-k).$$

$$\Longrightarrow IC_{1-\alpha}(C\theta) = \left[C\widehat{\theta} \pm t_{n-k,1-\alpha/2}\sqrt{\widehat{\sigma^2}C(X'X)^{-1}C'}\right].$$

Confidence region of $C\theta \in \mathbb{R}^q$ with $q \geq 2$

- $\widehat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2(X'X)^{-1})$
- $C\widehat{\theta} \sim \mathcal{N}_q(C\theta, \sigma^2\Delta)$ with $\Delta = C(X'X)^{-1}C' \in \mathcal{M}_q(\mathbb{R})$ thus

$$\frac{[C\widehat{\theta} - C\theta]'\Delta^{-1}[C\widehat{\theta} - C\theta]}{\sigma^2} \sim \chi^2(q).$$

- $(n-k)\widehat{\sigma^2}/\sigma^2 \sim \chi^2(n-k)$ and $\widehat{\sigma^2} \perp\!\!\!\perp C\widehat{\theta}$.
- We deduce that

$$A := \frac{[C\widehat{\theta} - C\theta]'\Delta^{-1}[C\widehat{\theta} - C\theta]}{q\ \widehat{\sigma^2}} \sim \mathcal{F}(q, n-k).$$

- Finally, $\mathbb{P}(A \leq f_{q,n-k,1-\alpha}) = \mathbb{P}(C\theta \in RC) = 1 - \alpha$ where RC is the confidence ellipsoid defined by:

$$RC = \left\{u \in \mathbb{R}^q;\ (C\widehat{\theta} - u)'[C(X'X)^{-1}C']^{-1}(C\widehat{\theta} - u) \leq q\widehat{\sigma^2} f_{q,n-k,1-\alpha}\right\}.$$

Summary

Summary

- Know how to write the hypotheses of a Fisher's test
- Know how to justify that one model is a sub-model of another
- Know the form of the Fisher test statistics, its law under $\mathcal{H}_0$ and know how to define the quantities that compose it according to the context
- Know how to carry out the construction of a Fisher's test
- Know how to construct a Student's test when $q = 1$
- Know how to construct a confidence interval for $C\theta$.
  Don't learn the formula!

Outline

Outline

4. **Chap 5: Singular models, orthogonality and hypotheses on errors**
   - Back to the hypotheses H1-H4
   - Singular models
   - Orthogonality

Hypotheses H1-H4

- H_1 : $\mathbb{E}[\varepsilon] = 0_n$ i.e $\mathbb{E}[Y] = X\theta$
- H_2 : $\mathbb{E}[\varepsilon_i] = \sigma^2, \forall i = 1, \ldots, n$
- H_3 : The random variables ε_i are independent.
- H_4 : The errors ε_i follow a Gaussian law

Gaussian law (H4)

- The assumption of normality of errors is the most difficult to verify in practice.
- The usual normality testing procedures (Kolmogorov-Smirnov, Cramer-Von Mises, Anderson-Darling or Shapiro-Wilks)
    - require the observation of the ε_i errors themselves
    - significant loss of power when they are applied to the residuals $\widehat{\varepsilon_i} = Y_i - \widehat{Y}_i$
- Plot Henry lines or QQ-plots to highlight obvious differences.

Properties of $\widehat{\theta}$ without H4

$$\widehat{\theta} = (X'X)^{-1}X'Y.$$

- $\widehat{\theta}$ remains unbiased, $\mathbb{E}[\widehat{\theta}] = \theta$, under H1.
- The variance-covariance matrix of $\widehat{\theta}$ remains equal to $\sigma^2(X'X)^{-1}$ under H2 and H3. This property is of little interest if H1 is not true.
- $\widehat{\theta}$ is no longer an optimal estimator among the unbiased estimators, but it remains so among the linear unbiased estimators under H1-H3.
- $\widehat{\theta}$ follows a Gaussian law under H3 and H4. If H4 is not satisfied, $\widehat{\theta}$ is **asymptotically Gaussian**.

Properties of $\widehat{\sigma^2}$

We assume that σ^2 is well defined (H2 true) and

$$\widehat{\sigma^2} = \frac{1}{n-k}\|Y - X\widehat{\theta}\|^2 \text{ with } \widehat{\theta} = (X'X)^{-1}X'Y.$$

Then we have the following properties:

- Under H1-H3, $\mathbb{E}[\widehat{\sigma^2}] = \sigma^2$ (even if H4 is not satisfied)
- $(n-k)\widehat{\sigma^2} \not\sim \sigma^2\chi^2(n-k)$ as soon as H4 is not satisfied.
- Under H1-H3, $\widehat{\sigma^2} \underset{n\to+\infty}{\overset{\mathbb{P}}{\to}} \sigma^2$ even if H4 is not satisfied.

Models with correlations

- It is possible to model correlations between errors, for example by supposing that these errors come from an ARMA (Autoregressive–moving-average model) process, which makes it possible to no longer need the hypothesis H3

- It is also possible to model the links by random effects models and thus study a mixed model

Outline

4. **Chap 5: Singular models, orthogonality and hypotheses on errors**
   - Back to the hypotheses H1-H4
   - Singular models
   - Orthogonality

Example of over-parametrized model

Two-way ANOVA model without interaction:
we assume that the both factors have 2 levels resp. and the 4 combinations are only observed once:

$$\begin{aligned} Y_{11} &= \mu + a_1 + b_1 + \varepsilon_{11} \\ Y_{12} &= \mu + a_1 + b_2 + \varepsilon_{12} \\ Y_{21} &= \mu + a_2 + b_1 + \varepsilon_{21} \\ Y_{22} &= \mu + a_2 + b_2 + \varepsilon_{22} \end{aligned}$$

The vector $\theta = (\mu, a_1, a_2, b_1, b_2)'$ and the matrix X of this model is :

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

For all vectors $v = (\alpha + \beta, -\alpha, -\alpha, -\beta, -\beta)$, $Xv = 0_4$.
The values μ, a_i, b_i for $i = 1$ or 2 are thus not uniquely identifiable.
The model is **over-parametrized**: 5 unknown parameters and only 4 df.

Singular model

Definition

The model is called **singular** or **no regular** when the matrix X is no injective ($\exists \theta \neq 0_k$ such that $X\theta = 0_n$).

Two remarks :

- $X\widehat{\theta} = P_{[X]}Y$ remains unique
- $\widehat{\theta}$ no unique : if $u \in Ker(X)$ then $\widehat{\theta} + u$ is also solution.

Singular model

X is not regular $\Longrightarrow X'X$ is not invertible.

Definition

Let M be a matrix. Then the matrix M^- is **a generalized inverse matrix** of M if

$$MM^-M = M.$$

The generalized inverse matrix always exists:
$(X'X)$ defines a bijective application from $Ker(X)^\perp$ to itself. It is thus sufficient to neglect the part contained in the kernel: we take the inverse on $Ker(X)^\perp$, completed arbitrary on $Ker(X)$.

The definition of $(X'X)^-$ is not unique!

It is then possible to generalize the results of the regular case.

Singular model

Proposition

If $(X'X)^-$ is a generalized inverse matrix of $X'X$, then $\widehat{\theta} = (X'X)^- X'Y$ is one solution of

$$(X'X)\widehat{\theta} = X'Y.$$

We start by noticing that

$$\forall \omega \in \mathbb{R}^k, < X\omega, P_{[X]^\perp} Y >=< \omega, X' P_{[X]^\perp} Y >= 0$$

thus

$$X'Y = X'P_{[X]}Y + X'P_{[X]^\perp}Y = X'P_{[X]}Y.$$

Thus, $\exists u \in \mathbb{R}^k,\ X'Y = X'Xu$. Finally,

$$(X'X)\widehat{\theta} = (X'X)(X'X)^- X'Y = (X'X)(X'X)^- X'Xu = X'Xu = X'Y.$$

Singular model - Identifiability constraints

In general, we prefer to remove the indeterminacy of $\hat{\theta}$ by setting constraints, often in order to give a more intuitive interpretation to θ.

Proposition

Let X be a singular matrix with $rg(X) = r < k$ so that there are $k - r$ redundant parameters. Let M be a matrix with $k - r$ rows and k columns, $rg(M) = k - r$ and such that:

$$Ker(M) \cap Ker(X) = \{0_k\}.$$

Then,

- the matrix $(X'X + M'M)$ is invertible and its inverse matrix is a generalized inverse matrix of $X'X$
- the vector $\widehat{\theta} = (X'X + M'M)^{-1}X'Y$ is the unique solution of $\left\{ \begin{array}{l} X'X\alpha = X'Y \\ M\alpha = 0_{k-r}. \end{array} \right.$

Proof

1. Show that $X'X + M'M$ is invertible: show that
$$A = \begin{pmatrix} X \\ M \end{pmatrix} \in \mathcal{M}_{n+k-r,k}(\mathbb{R})$$
is injective and thus $A'A$ is invertible.

2. Consider the following minimization problem:
$$g : \alpha \mapsto \|Y - X\alpha\|^2 + \|M\alpha\|^2.$$
Write $g(\alpha)$ under the form $g(\alpha) = \|\tilde{Y} - A\alpha\|^2$ with $\tilde{Y}$ to precise.
Deduce that $\widehat{\theta}$ is solution of $\left\{ \begin{array}{c} X'X\alpha = X'Y \\ M\alpha = 0_{k-r}. \end{array} \right.$

3. Show that this solution is unique.

Singular model - Identifiability constraints

Example : One-way ANOVA model

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{ij} \text{ for } i = 1, \cdots, 4 \text{ and } j = 1.$$

The associated matrix X is:

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

- If we consider the constraint $M = (0\ 1\ 1\ 1\ 1)$:

$$M\theta = 0 \Leftrightarrow \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0.$$

  thus we impose that the sum of the differential effects is null.

- If we consider the constraint $M = (0\ 1\ 0\ 0\ 0)$: $M\theta = \alpha_1 = 0$ thus we impose that the first category is the reference.

Estimable functions and contrasts

When X is singular, it is possible to define an estimator with a generalized inverse matrix. What about the testing procedures? In particular, are these constraints systematically necessary?

Definition

A linear function $C\theta$ is called **estimable function** if it doesn't depend on the particular solution of the over-parameterized model (doesn't depend on the type of constraint chosen). It can be checked that it satisfies $C\theta = DX\theta$ where D is a full ranked matrix.

Definition

A linear combination $C\theta$ is a **contrast** if $C\mathbb{1} = 0$.

In analysis of variance, most of the linear combinations which are tested are contrasts.

Outline

4. **Chap 5: Singular models, orthogonality and hypotheses on errors**
   - Back to the hypotheses H1-H4
   - Singular models
   - Orthogonality

Orthogonality for regular models

- Orthogonality permits to simplify the computation but also the interpretation in a linear model
- A linear model most often admits a natural decomposition of the parameters θ and thus a decomposition of the associated design matrix X
- We will be interested here in the possible orthogonality of the various spaces associated with this decomposition

Example

Consider the multiple regression model on three variables $x^{(1)}$, $x^{(2)}$ et $x^{(3)}$:

$$Y_i = \mu + \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + \theta_3 x_i^{(3)} + \varepsilon_i, i = 1, \cdots, n > 4.$$

The vector θ contains 4 terms: μ, θ_1, θ_2, θ_3 and the matrix X four columns. Quite naturally here, we can consider the decomposition into four elements. Then the matrix X is the concatenation of 4 column vectors. The orthogonality of the partition will then strictly correspond to the orthogonality of the 4 one-dimensional spaces: $[\mathbb{1}]$, $[x^{(1)}]$, $[x^{(2)}]$ and $[x^{(3)}]$.

Example

Consider the quadratic regression model depending on two variables $x^{(1)}$ and $x^{(2)}$: $\forall i = 1, \cdots, n > 6$,

$$Y_i = \mu + \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + \gamma_1 \left(x_i^{(1)}\right)^2 + \gamma_2 \left(x_i^{(2)}\right)^2 + \delta x_i^{(1)} x_i^{(2)} + \varepsilon_i.$$

Here we can consider the partition:

- the constant μ
- linear effects θ_1, θ_2
- squares γ_1, γ_2
- cross product δ

The orthogonality here is the orthogonality of the vector sub-spaces $[\mathbb{1}]$, $[(x^{(1)}, x^{(2)})]$, $\left[\left((x^{(1)})^2, (x^{(2)})^2\right)\right]$ and $[x^{(1)}x^{(2)}]$.

Orthogonality for regular models

Definition

Let $Y = X\theta + \varepsilon$ be a regular linear model.
Consider a partition in m terms of X and θ:

$$Y = X_1\theta_1 + \cdots + X_m\theta_m + \varepsilon,$$

where X_j is a matrix of size (n, k_j) and $\theta_j \in \mathbb{R}^{k_j}$ with $k_j \in \{1, \cdots, k\}$ for $j = 1, \cdots, m$ and with $\sum_{j=1}^{m} k_j = k$.
This partition is said **orthogonal** if the following sub-spaces of $\mathbb{R}^n$ are orthogonal:

$$[X_1], \cdots, [X_m]$$

A consequence of the orthogonality of a linear model is that the information matrix $X'X$ has a block diagonal structure, each block being associated with each element of the partition.

Orthogonality for regular models

Proposition

Consider a regular linear model with an orthogonal partition:

$$Y = X_1\theta_1 + \cdots + X_m\theta_m + \varepsilon.$$

Then

- The estimators of the different effects $\widehat{\theta}_1, \ldots, \widehat{\theta}_m$ are independent (non-correlated under non Gaussian model)
- For $l = 1, \cdots, m$, the expression of $\widehat{\theta}_l$ does not depend on the presence or absence of the other terms θ_j in the model.

Orthogonality for singular models

Definition

Consider a partition for a singular linear model

$$Y = X_1\theta_1 + \cdots + X_m\theta_m + \varepsilon.$$

Consider a system of constraints $C_1\theta_1 = 0, \cdots, C_m\theta_m = 0$ that make the model identifiable. We say that these constraints make the partition orthogonal if the sub-spaces

$$V_j = \{X_j\theta_j;\ \theta_j \in Ker(C_j)\},\ j = 1, \cdots, m$$

are orthogonal.

The idea is to choose constraints that make the model orthogonal.
We will see that this definition takes on its full meaning with the example of the two-way ANOVA model.

Summary

Summary

In this chapter, you are expected to understand

- the problem of parameter estimation for a singular linear model
- the interest of having orthogonality for a linear model

You are not expected to know these results but you will apply them in the framework of ANOVA and ANCOVA.

References I

[1] Jean-Marc Azais and Jean-Marc Bardet. *Le modèle linéaire par l'exemple-2e éd.: Régression, analyse de la variance et plans d'expérience illustrés avec R et SAS*. Dunod, 2012.

[2] Jean-Jacques Daudin. *Le modèle linéaire et ses extensions-Modèle linéaire général, modèle linéaire généralisé, modèle mixte, plans d'expériences (Niveau C)*. 2015.

[3] Nalini Ravishanker, Zhiyi Chi, and Dipak K Dey. *A first course in linear model theory*. CRC Press, 2021.

[4] Alvin C Rencher and G Bruce Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008.