

Modèle Linéaire Gaussien

Bruno Pelletier

Outline

Introduction

Modèle

Estimation des paramètres

Propriétés des estimateurs

Intervalles de confiance

Prévision

Encore des tests

Introduction

Introduction

Rappel : Régression linéaire multiple

$$\mathbf{Y} = \mathbb{X}\beta + \epsilon, \quad \text{avec}$$

$$\mathbb{E}[\epsilon] = 0 \quad \text{et} \quad \mathbb{V}(\epsilon) = \sigma^2 \mathbf{I}_n.$$

- Comment effectuer une inférence sur β , autre que ponctuelle, e.g., intervalles/domaines de confiance ?
- La loi de $\hat{\beta}$ est inconnue : utilisation d'inégalités (éventuellement peu fines).
- Si la loi de $\hat{\beta}$ était connue :
 - accès direct à des régions de confiances,
 - test sur les composantes de β .

~~ Formuler une hypothèse supplémentaire sur le vecteur des bruits ϵ : **vecteur gaussien**

Modèle

Modèle

$$\mathbf{Y} = \mathbb{X}\beta + \epsilon,$$

avec $\mathbb{X} \in \mathcal{M}_{n,p}(\mathbb{R})$, $\beta \in \mathbb{R}^p$, et ϵ vecteur aléatoire de \mathbb{R}^n .

Hypothèses sur ϵ

- $\mathbb{E}[\epsilon] = 0$ (bruits centrés).
- $\mathbb{V}(\epsilon) = \sigma^2 \mathbf{I}_n$ (homoscédasticité et décorrélation).
- ϵ est un vecteur gaussien de \mathbb{R}^n .

La loi de ϵ est donc **connue** (loi Normale sur \mathbb{R}^n) :

$$\epsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n).$$

Loi des observations

Loi de $\mathbf{Y} = \mathbb{X}\beta + \epsilon$?

Rappel

- Toute transformation affine d'un vecteur gaussien est un vecteur gaussien, i.e., si U est un vecteur gaussien de \mathbb{R}^n , $A \in \mathcal{M}_{m,n}(\mathbb{R})$ et $b \in \mathbb{R}^m$, alors $AU + b$ est un vecteur gaussien (de \mathbb{R}^m).
- En outre, si $U \sim \mathcal{N}_n(\mu, \Sigma)$, alors $AU + b \sim \mathcal{N}_m(A\mu + b, A\Sigma A')$.

↔ : il suffit d'identifier les moments.

Loi de \mathbf{Y}

On a $\mathbb{E}[\mathbf{Y}] = \mathbb{X}\beta$ et $\mathbb{V}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$, d'où :

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbb{X}\beta, \sigma^2 \mathbf{I}_n).$$

Loi des observations

- $\epsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$:
 - les ϵ_i sont décorrélatés,
 - ϵ est un vecteur gaussien,
 - donc les ϵ_i sont indépendants.
- $\mathbf{Y} \sim \mathcal{N}_n(\mathbb{X}\beta, \sigma^2 \mathbf{I}_n)$:
 - les Y_i sont décorrélatés,
 - \mathbf{Y} est un vecteur gaussien,
 - donc les Y_i sont indépendants.

Remarques

- $\epsilon_i \sim \mathcal{N}(0, \sigma^2) \rightsquigarrow$ les bruits sont IID.
- $Y_i \sim \mathcal{N}(x_i\beta, \sigma^2) \rightsquigarrow$ les Y_i sont indépendants mais ne sont pas identiquement distribués.

Estimation des paramètres

Estimation des paramètres

Régression linéaire multiple

Utilisation du principe des Moindres Carrés Ordinaires.

Modèle Linéaire Gaussien

- La loi des observations est connue : $\mathbf{Y} \sim \mathcal{N}_n (\mathbb{X}\beta, \sigma^2 \mathbf{I}_n)$.
- Modèle statistique paramétré dominé par la mesure de Lebesgue sur \mathbb{R}^n .
- Estimation des paramètres par Maximum de Vraisemblance.

Vraisemblance du modèle

$$\mathcal{L}(\beta, \sigma^2; Y_1, \dots, Y_n) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{\|\mathbf{Y} - \mathbb{X}\beta\|^2}{2\sigma^2}\right).$$

Maximisation de la vraisemblance

Log-Vraisemblance

$$\ln \mathcal{L}(\beta, \sigma^2; Y_1, \dots, Y_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbb{X}\beta\|^2.$$

Point critique

$$\begin{cases} \nabla_{\beta} \ln \mathcal{L}(\beta, \sigma; Y_1, \dots, Y_n) = 0 \\ \frac{\partial}{\partial(\sigma^2)} \ln \mathcal{L}(\beta, \sigma; Y_1, \dots, Y_n) = 0 \end{cases} \iff \begin{cases} \mathbb{X}' \mathbb{X} \beta - \mathbb{X}' \mathbf{Y} = 0 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \|\mathbf{Y} - \mathbb{X}\beta\|^2 = 0 \end{cases}$$

Estimateurs du maximum de vraisemblance

$$\hat{\beta} = (\mathbb{X}' \mathbb{X})^{-1} \mathbb{X}' \mathbf{Y} \quad \text{et} \quad \tilde{\sigma}^2 = \frac{1}{n} \|\mathbf{Y} - \mathbb{X}\hat{\beta}\|^2.$$

Estimation sans biais de la variance

- Le calcul donne

$$\mathbb{E} [\tilde{\sigma}^2] = \frac{n-p}{n} \sigma^2.$$

↔ l'estimateur du maximum de vraisemblance de σ^2 est **biaisé**.

- On préfèrera utiliser l'estimateur **sans biais** suivant :

$$\hat{\sigma}^2 = \frac{1}{n-p} \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 = \frac{n}{n-p} \tilde{\sigma}^2.$$

- Remarque : Le modèle contient **p paramètres** $\beta_0, \dots, \beta_{p-1}$.

Propriétés des estimateurs

Propriétés des estimateurs

$\hat{\beta}$ est un estimateur sans biais de β

- Calcul analogue à celui de l'estimateur des MCO :

$$\mathbb{E} [\hat{\beta}] = (\mathbb{X}' \mathbb{X})^{-1} \mathbb{X}' \mathbb{E} [\mathbf{Y}] = (\mathbb{X}' \mathbb{X})^{-1} \mathbb{X}' \mathbb{X} \beta = \beta.$$

$\hat{\sigma}^2$ est un estimateur sans biais de σ^2

- par construction.

En outre, la loi de \mathbf{Y} étant connue, on peut déterminer la loi exacte des EMV $\hat{\beta}$ et $\hat{\sigma}^2$.

Loi de $\hat{\beta}$

Modèle

$$\mathbf{Y} = \mathbb{X}\beta + \epsilon \quad \rightsquigarrow \mathbf{Y} \sim \mathcal{N}_n (\mathbb{X}\beta, \sigma^2 \mathbf{I}_n).$$

Caractérisation

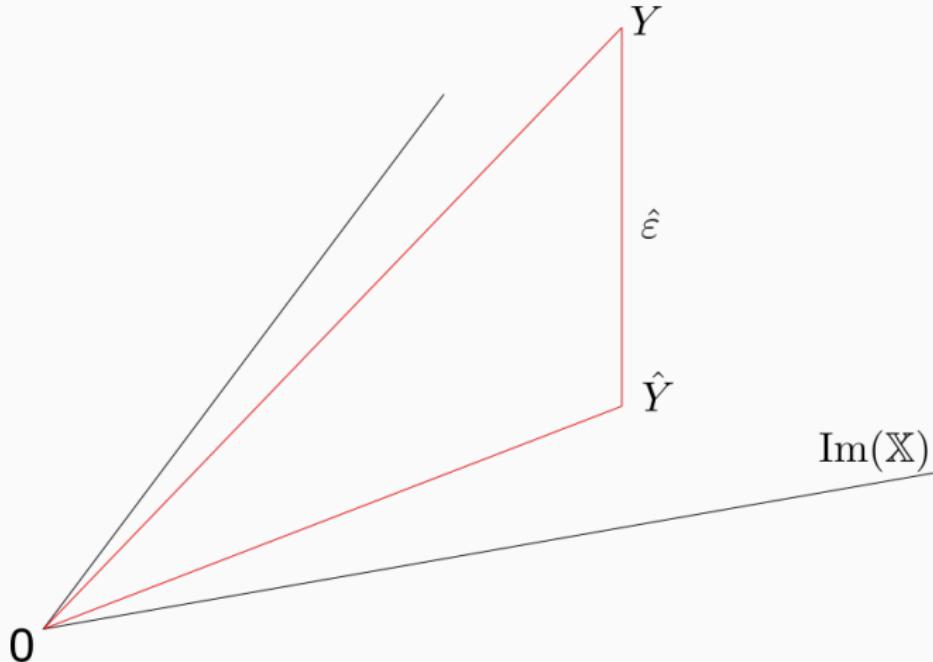
- $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y}$: transformation linéaire d'un vecteur gaussien. Donc $\hat{\beta}$ est un vecteur gaussien (dans \mathbb{R}^p).

Moments

- $\mathbb{E} [\hat{\beta}] = \beta,$
- $\mathbb{V} (\hat{\beta}) = \mathbb{V} ((\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y}) = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}' \mathbb{V} (\mathbf{Y}) ((\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}')'$
 $= \sigma^2 (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} = \sigma^2 (\mathbb{X}'\mathbb{X})^{-1}$
 $\rightsquigarrow \hat{\beta} \sim \mathcal{N}_p (\beta, \sigma^2 (\mathbb{X}'\mathbb{X})^{-1}).$

Loi de $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{1}{n-p} \|\mathbf{Y} - \mathbb{X}\hat{\beta}\|^2 = \frac{1}{n-p} \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \frac{1}{n-p} \|\hat{\epsilon}\|^2.$$



Rappel : Théorème de Cochran

Théorème (Cochran)

Soit $U \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ et soit E un sous-espace vectoriel de \mathbb{R}^n de dimension p . On note Π_E la projection orthogonale sur E .

Alors :

- $\Pi_E U$ et $U - \Pi_E U$ sont des vecteurs gaussiens avec

$$\Pi_E U \sim \mathcal{N}_n(0, \sigma^2 \Pi_E) \quad \text{et} \quad U - \Pi_E U \sim \mathcal{N}_n(0, \sigma^2(\mathbf{I} - \Pi_E)).$$

- $\Pi_E U$ et $(U - \Pi_E U)$ sont indépendants.
- Lois des normes au carré des projetés :

$$\frac{\|\Pi_E U\|^2}{\sigma^2} \sim \chi^2(p) \quad \text{et} \quad \frac{\|U - \Pi_E U\|^2}{\sigma^2} \sim \chi^2(n-p).$$

Loi de $\hat{\sigma}^2$

- Le théorème de Cochran est appliqué avec

$$U = Y - \mathbb{X}\beta \quad \text{et} \quad E = \text{Im}(\mathbb{X}).$$

- Projection orthogonale sur $\text{Im}(\mathbb{X})$:

$$\Pi_{\text{Im}(\mathbb{X})} = \mathbb{X} (\mathbb{X}' \mathbb{X})^{-1} \mathbb{X}'$$

- Projections :

$$\begin{aligned}\Pi_{\text{Im}(\mathbb{X})} U &= \hat{\mathbf{Y}} - \mathbb{X}\beta \\ U - \Pi_{\text{Im}(\mathbb{X})} U &= \mathbf{Y} - \hat{\mathbf{Y}}.\end{aligned}$$

- Conclusion : la loi de $\hat{\sigma}^2$ est caractérisée par

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

Lien entre $\hat{\beta}$ et $\hat{\sigma}^2$

- On déduit également du théorème de Cochran que $\Pi_{\text{Im}(\mathbb{X})}U$ et $U - \Pi_{\text{Im}(\mathbb{X})}U$ sont indépendants, d'où :

$$\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{Y} - \hat{\mathbf{Y}}.$$

- Or, comme $\mathbf{Y} - \hat{\mathbf{Y}} \perp\!\!\!\perp \text{Im}(\mathbb{X})$:

$$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\hat{\mathbf{Y}} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\hat{\mathbf{Y}}.$$

- Conclusion :

$$\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2.$$

Intervalles de confiance

Loi d'une composante $\hat{\beta}_j$ de $\hat{\beta}$

- On a montré que :

$$\hat{\beta} \sim \mathcal{N}_p \left(\beta, \sigma^2 (\mathbb{X}' \mathbb{X})^{-1} \right).$$

- Donc

$$\hat{\beta}_j \sim \mathcal{N} \left(\beta_j, \sigma^2 (\mathbb{X}' \mathbb{X})_{j+1,j+1}^{-1} \right), \quad j = 0, \dots, p-1.$$

- Construction d'une statistique pivotale :

$$\frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{(\mathbb{X}' \mathbb{X})_{j+1,j+1}^{-1}}} \sim \mathcal{N}(0, 1).$$

↔ Remplacer σ (inconnu) par $\hat{\sigma}$.

Rappel : Loi de Student

Soit Z et V deux variables aléatoires indépendantes telles que :

$$Z \sim \mathcal{N}(0, 1) \quad \text{et} \quad V \sim \chi^2(\nu).$$

Alors

$$\frac{Z}{\sqrt{\frac{V}{\nu}}} \sim \mathcal{T}(\nu),$$

Loi de Student à ν degrés de liberté.

Statistique pivotale sur β_j

- $\hat{\beta}_j$ et $\hat{\sigma}^2$ sont indépendants.
- $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$. D'où :

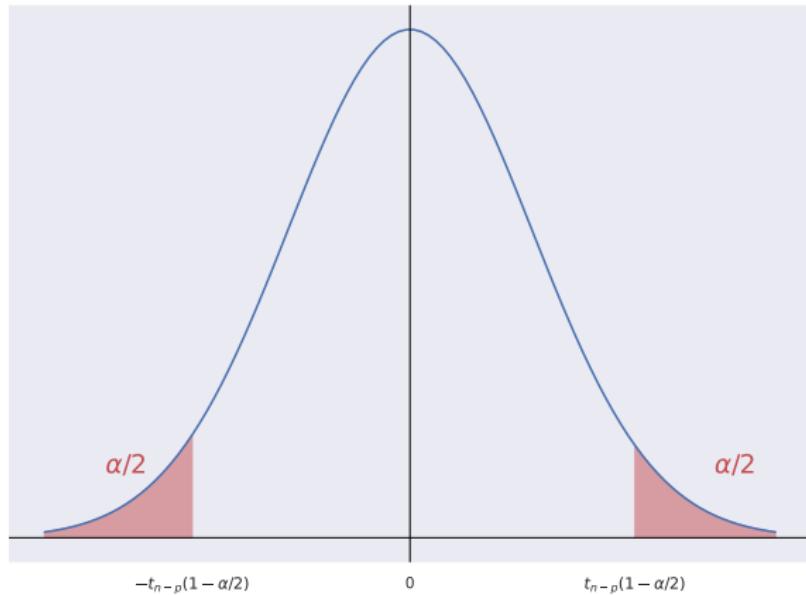
$$\frac{\frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{(\mathbb{X}'\mathbb{X})_{j+1,j+1}^{-1}}}}{\sqrt{\frac{(n-p)\hat{\sigma}^2}{\sigma^2}}} \sim \mathcal{T}(n-p).$$

- Conclusion :

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{(\mathbb{X}'\mathbb{X})_{j+1,j+1}^{-1}}} \sim \mathcal{T}(n-p).$$

- Remarque : $\mathcal{T}(\nu) \xrightarrow{\nu \rightarrow \infty} \mathcal{N}(0, 1)$.

Intervalle de confiance sur une composante β_j



$$\mathbb{P} \left(-t_{n-p}(1 - \alpha/2) \leq \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{(\mathbb{X}'\mathbb{X})_{j+1,j+1}^{-1}}} \leq t_{n-p}(1 - \alpha/2) \right) = 1 - \alpha.$$

Intervalle de confiance sur une composante β_j

D'où un intervalle de confiance de niveau $1 - \alpha$ sur β_j :

$$\left[\hat{\beta}_j - t_{n-p}(1 - \alpha/2)\hat{\sigma}\sqrt{(\mathbb{X}'\mathbb{X})_{j+1,j+1}^{-1}} ; \hat{\beta}_j + t_{n-p}(1 - \alpha/2)\hat{\sigma}\sqrt{(\mathbb{X}'\mathbb{X})_{j+1,j+1}^{-1}} \right].$$

Remarques :

- A partir de la statistique pivotale, il est possible de construire des intervalles de confiance unilatères.
- Cet intervalle de confiance est exact, i.e. non asymptotique.

Test sur une composante de β

Modèle

$$Y_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i.$$

- La $j^{\text{ème}}$ variable est-elle significative ? En d'autres termes, peut-on considérer que β_j est nul ?

~~~ test de nullité sur  $\beta_j$

- Plus généralement :

$$\begin{array}{l|ll} & H_0 & : \beta_j = \beta^* \\ & H_1 & : \beta_j \neq \beta^*. \end{array}$$

# Test sur une composante $\beta_j$ de $\beta$

## Test

$$\begin{array}{l} \left| \begin{array}{ll} H_0 & : \beta_j = \beta^* \\ H_1 & : \beta_j \neq \beta^*. \end{array} \right. \end{array}$$

## Statistique de test

$$T(\mathbf{Y}) = \frac{\hat{\beta}_j - \beta^*}{\hat{\sigma} \sqrt{(\mathbb{X}'\mathbb{X})_{j+1,j+1}^{-1}}} \sim \mathcal{T}(n-p) \quad \text{sous } H_0.$$

## Règle de décision

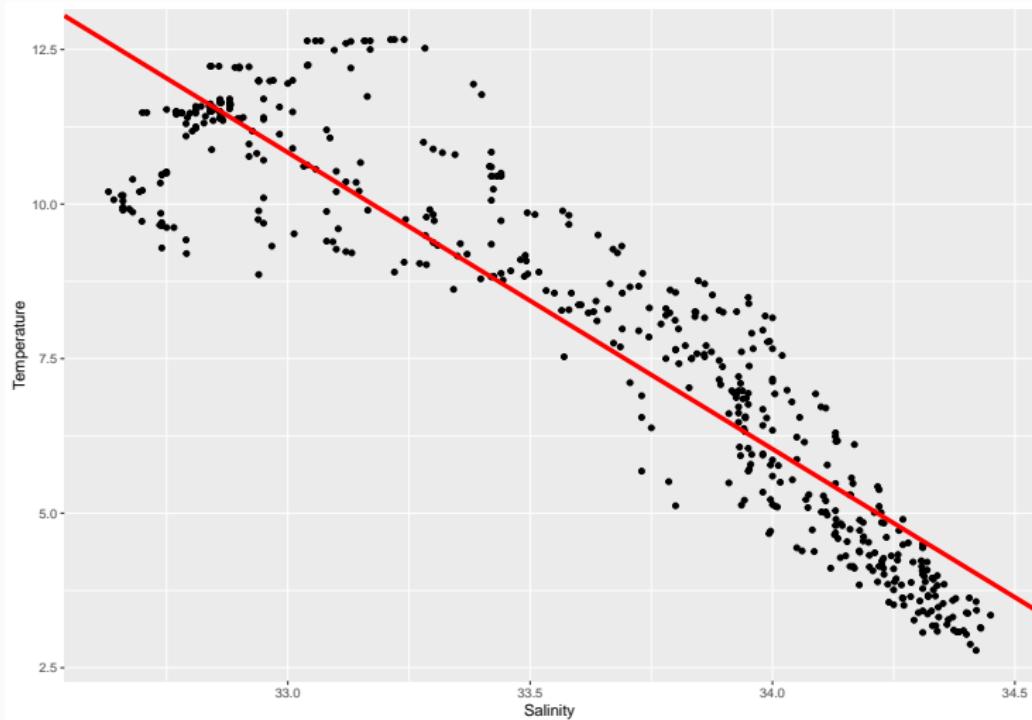
Rejeter  $H_0$  si :  $|T(\mathbf{Y})| > t_{n-p}(1 - \alpha/2)$ .

~ Produit un test de niveau  $\alpha$ .

## Remarques

- Test de nullité : prendre  $\beta^* = 0$ .
- p-value =  $\mathbb{P}(|\mathcal{T}(n-p)| > |T^{\text{obs}}|)$ .

## Exemple : Salinité et Température (CalCOFI) i



Droite de régression :  $y = 169.1178 - 4.79646x$ .

## Exemple : Salinité et Température (CalCOFI) ii

```
> m <- lm(Temperature~Salinity,data=bottle)
> summary(m)
```

Call:

```
lm(formula = Temperature ~ Salinity, data = bottle)
```

Residuals:

| Min      | 1Q       | Median   | 3Q      | Max     |
|----------|----------|----------|---------|---------|
| -2.79153 | -0.75022 | -0.06611 | 0.66100 | 3.04295 |

Coefficients:

|                | Estimate  | Std. Error | t value  | Pr(> t )   |         |   |
|----------------|-----------|------------|----------|------------|---------|---|
| (Intercept)    | 169.11780 | 3.03735    | 55.68    | <2e-16 *** |         |   |
| Salinity       | -4.79646  | 0.09031    | -53.11   | <2e-16 *** |         |   |
| ---            |           |            |          |            |         |   |
| Signif. codes: | 0 ‘***’   | 0.001 ‘**’ | 0.01 ‘*’ | 0.05 ‘.’   | 0.1 ‘ ’ | 1 |

Residual standard error: 1.123 on 491 degrees of freedom

Multiple R-squared: 0.8517, Adjusted R-squared: 0.8514

F-statistic: 2821 on 1 and 491 DF, p-value: < 2.2e-16

## Exemple : Salinité et Température (CalCOFI) iii

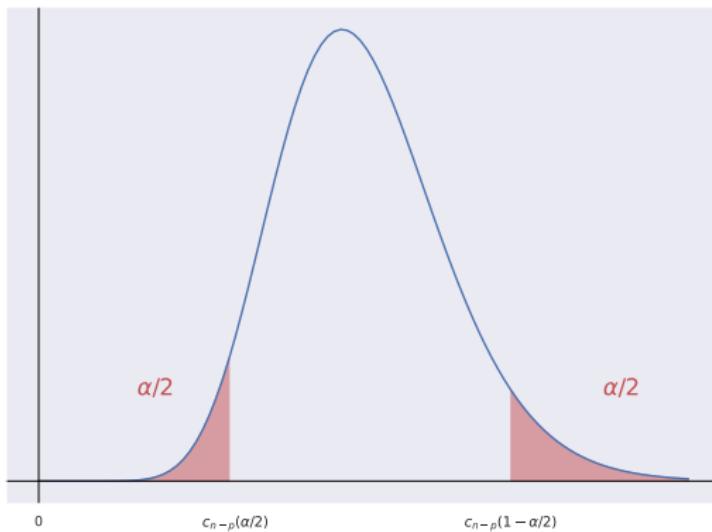
Coefficients:

|                | Estimate  | Std. Error | t value | Pr(> t ) |      |     |      |      |     |     |   |
|----------------|-----------|------------|---------|----------|------|-----|------|------|-----|-----|---|
| (Intercept)    | 169.11780 | 3.03735    | 55.68   | <2e-16   | ***  |     |      |      |     |     |   |
| Salinity       | -4.79646  | 0.09031    | -53.11  | <2e-16   | ***  |     |      |      |     |     |   |
| ---            |           |            |         |          |      |     |      |      |     |     |   |
| Signif. codes: | 0         | '***'      | 0.001   | '**'     | 0.01 | '*' | 0.05 | '. ' | 0.1 | ' ' | 1 |

- Estimate :  $\hat{\beta}_1 = -4.79646$
- Std. Error :  $\hat{\sigma} \sqrt{(\mathbb{X}'\mathbb{X})_{2,2}^{-1}} = 0.09031$
- t value :  $T = \frac{\hat{\beta}_1}{\hat{\sigma} \sqrt{(\mathbb{X}'\mathbb{X})_{2,2}^{-1}}} = -53.11.$
- $Pr(> |t|)$  : p-value du test de nullité.

# Inférence sur $\sigma^2$

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



## Inférence sur $\sigma^2$

Intervalle de confiance de niveau  $1 - \alpha$  sur  $\sigma^2$

$$\left[ \frac{(n-p)\hat{\sigma}^2}{c_{n-p}(1-\alpha/2)} ; \frac{(n-p)\hat{\sigma}^2}{c_{n-p}(\alpha/2)} \right]$$

Test de niveau  $\alpha$  sur  $\sigma^2$

- Soit le test

$$\begin{cases} H_0 & : \sigma^2 = a \quad \text{pour } a > 0 \\ H_1 & : \sigma^2 \neq a \end{cases}$$

- Règle de décision : Rejeter  $H_0$  si

$$\frac{(n-p)\hat{\sigma}^2}{a} < c_{n-p}(\alpha/2) \quad \text{ou} \quad \frac{(n-p)\hat{\sigma}^2}{a} > c_{n-p}(1-\alpha/2).$$

## Prévision

---

# Prévision

- Modèle :  $Y_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$ .
- Nouvelle observation :  $x_{n+1} = (x_{n+1,0}, \dots, x_{n+1,p-1})$  (vecteur ligne).
- Variable réponse (non observée) :  $Y_{n+1} = x_{n+1}\beta + \epsilon_{n+1}$ , où :
  - $\epsilon_{n+1} \perp\!\!\!\perp (\epsilon_1, \dots, \epsilon_n)$ ,
  - $\epsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$ .
- Objectif : prédire une nouvelle observation de  $Y_{n+1}$ .
- Valeur prédictive  $\hat{Y}_{n+1}^{(p)} = x_{n+1}\hat{\beta}$ .
- Erreur de prévision :  $Y_{n+1} - \hat{Y}_{n+1}^{(p)}$ .

## Propriétés de l'erreur de prévision

- $\mathbb{E} [Y_{n+1} - \hat{Y}_{n+1}^{(p)}] = x_{n+1}\beta - x_{n+1} \mathbb{E} [\hat{\beta}] = 0.$

$$\begin{aligned}\mathbb{V} (Y_{n+1} - \hat{Y}_{n+1}^{(p)}) &= \mathbb{V} (Y_{n+1}) + \mathbb{V} (\hat{Y}_{n+1}^{(p)}) \quad (\text{car décorrélée}) \\ &= \sigma^2 + \sigma^2 x_{n+1} (\mathbb{X}' \mathbb{X})^{-1} x'_{n+1} \\ &= \sigma^2 (1 + x_{n+1} (\mathbb{X}' \mathbb{X})^{-1} x'_{n+1}).\end{aligned}$$

- $Y_{n+1} - \hat{Y}_{n+1}^{(p)} = x_{n+1}\beta + \epsilon_{n+1} - x_{n+1}\hat{\beta} = x_{n+1}(\beta - \hat{\beta}) + \epsilon_{n+1}.$

Or  $\hat{\beta} \perp\!\!\!\perp \epsilon_{n+1}$  par hypothèse sur les bruits.

$$\rightsquigarrow Y_{n+1} - \hat{Y}_{n+1}^{(p)} \sim \mathcal{N} (0, \sigma^2 (1 + x_{n+1} (\mathbb{X}' \mathbb{X})^{-1} x'_{n+1})) .$$

# Intervalle de confiance en prévision

- $Y_{n+1} - \hat{Y}_{n+1}^{(p)} \sim \mathcal{N}(0, \sigma^2 (1 + x_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x'_{n+1}))$ .
- $\hat{\sigma}^2 \perp\!\!\!\perp Y_{n+1} - \hat{Y}_{n+1}^{(p)}$  car  $\hat{\sigma}^2 \perp\!\!\!\perp \hat{\beta}$  et  $\hat{\sigma}^2 \perp\!\!\!\perp \epsilon_{n+1}$ .

$$\rightsquigarrow \frac{Y_{n+1} - \hat{Y}_{n+1}^{(p)}}{\sqrt{\hat{\sigma}^2 (1 + x_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x'_{n+1})}} \sim \mathcal{T}(n-p).$$

Un intervalle de confiance de niveau  $1 - \alpha$  sur  $Y_{n+1}$  est donné par :

$$\left[ \hat{Y}_{n+1}^{(p)} \pm t_{n-p}(1 - \alpha/2) \sqrt{\hat{\sigma}^2 (1 + x_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x'_{n+1})} \right]$$

Encore des tests

---

# Motivation

- Modèle :  $\mathbf{Y} = \mathbb{X}\beta + \epsilon$ .
- On sait apprécier la significativité individuelle d'une variable au moyen d'un test de nullité du coefficient  $\beta_j$  associé.
- Que dire de la significativité **simultanée** de plusieurs variables ?
- Applications :
  - Tester la validité globale d'un modèle,
  - Tester la validité d'un sous-modèle,
  - Domaines de confiance sur plusieurs composantes de  $\beta$  : ellipsoïdes.

## Test de validité globale du modèle

- Modèle avec constante :

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i.$$

- Si  $\beta_1 = \cdots = \beta_{p-1} = 0$ , alors les variables explicatives n'influencent pas la réponse.

### Test de validité globable

$$\begin{array}{l|l} H_0 & : \beta_1 = \cdots = \beta_{p-1} = 0 \\ H_1 & : \exists 1 \leq k \leq p-1 : \beta_k \neq 0. \end{array}$$

- Remarque : attention, la nullité du coefficient de la constante n'est pas testée.

## Test de validité d'un sous-modèle

- Peut-on supprimer certaines variables explicatives ?
- Équivalent à tester la nullité simultanée des coefficients associés.

### Test de validité d'un sous-modèle

$$\begin{array}{ll} H_0 & : \beta_{i_1} = \dots = \beta_{i_q} = 0 \\ H_1 & : \exists 1 \leq k \leq q : \beta_{i_k} \neq 0. \end{array}$$

### Exemple

- Modèle :  $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i$ .
- Peut-on considérer que  $Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i,1} + \tilde{\epsilon}_i$  ?
- On teste  $H_0 : \beta_2 = \beta_3 = 0$  contre son contraire.

## Formalisme unifié

Test de nullité de  $\beta_j : H_0 : \beta_j = 0$

$$\beta_j = 0 \iff [0, \dots, 0, \underbrace{1}_{j+1}, 0, \dots, 0] \beta = 0.$$

Test de validité globale :  $H_0 : \beta_1 = \dots \beta_{p-1} = 0$

$$H_0 \iff \underbrace{\left(0, \mathbf{I}_{p-1}\right)}_{(p-1) \times p} \beta = 0_{\mathbb{R}^{p-1}}.$$

Test d'un sous-modèle : exemple  $p = 4$  et  $H_0 : \beta_1 = \beta_3 = 0$

$$H_0 \iff \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \beta = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

$$\rightsquigarrow H_0 : A\beta = 0, \quad A \in \mathcal{M}_{q,p}(\mathbb{R}).$$

# Construction d'une statistique de test

$$H_0 : A\beta = 0, \quad A \in \mathcal{M}_{q,p}(\mathbb{R}).$$

$$H_0 \iff \|A\beta\| = 0 \iff \|A\beta\|^2 = 0.$$

- $\|A\beta\|^2$  peut être estimée par  $\|\hat{A}\beta\|^2$ .  
~ $\rightsquigarrow$  Loi de  $\|\hat{A}\beta\|^2$  ?

- $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(\mathbb{X}'\mathbb{X})^{-1})$ .
- $A\hat{\beta}$  est un vecteur gaussien.

$$A\hat{\beta} \sim \mathcal{N}_q(A\beta, \sigma^2 A(\mathbb{X}'\mathbb{X})^{-1} A').$$

## Loi de $\|\hat{\beta}\|^2$

- Centrage :

$$A(\hat{\beta} - \beta) \sim \mathcal{N}_q \left( 0, \underbrace{\sigma^2 A(\mathbb{X}'\mathbb{X})^{-1} A'}_{\Sigma_A} \right).$$

- Réduction :  $\Sigma_A = \Sigma_A^{1/2} \Sigma_A^{1/2}$ .

$$\Sigma_A^{-1/2} (A(\hat{\beta} - \beta)) \sim \mathcal{N}_q(0, \mathbf{I}_q).$$

- Donc :

$$\left\| \Sigma_A^{-1/2} A(\hat{\beta} - \beta) \right\|^2 \sim \chi^2(q).$$

## Loi de $\|A\hat{\beta}\|^2$

$$\begin{aligned} \left\| \Sigma_A^{-1/2} A(\hat{\beta} - \beta) \right\|^2 &= (\hat{\beta} - \beta)' A' \Sigma_A^{-1} A (\hat{\beta} - \beta) \\ &= \frac{1}{\sigma^2} (\hat{\beta} - \beta)' A' [A(\mathbb{X}'\mathbb{X})^{-1} A']^{-1} A (\hat{\beta} - \beta). \end{aligned}$$

↔ Remplacer  $\sigma$  par  $\hat{\sigma}$  ?

**Rappel : Loi de Fisher**

Si  $U \sim \chi^2(\nu_1)$  et  $V \sim \chi^2(\nu_2)$  et  $U \perp\!\!\!\perp V$ , alors

$$\frac{U/\nu_1}{V/\nu_2} \sim F(\nu_1, \nu_2),$$

Loi de Fisher à  $\nu_1$  et  $\nu_2$  degrés de liberté.

# Statistique de test

- $\frac{1}{\sigma^2}(\hat{\beta} - \beta)' A' [A(\mathbb{X}'\mathbb{X})^{-1}A']^{-1} A(\hat{\beta} - \beta) \sim \chi^2(q)$  : aléatoire au travers de  $\hat{\beta}$ .
- $\hat{\sigma}^2$  et  $\hat{\beta}$  sont indépendants.
- $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$ .

Conclusion :

$$\frac{1}{q\hat{\sigma}^2}(\hat{\beta} - \beta)' A' [A(\mathbb{X}'\mathbb{X})^{-1}A']^{-1} A(\hat{\beta} - \beta) \sim F(q, n-p).$$

## Test de nullité d'un coefficient

$$\begin{array}{ll} H_0 & : \beta_j = 0 \\ H_1 & : \beta_j \neq 0 \end{array}$$

- Avec  $A = [0, \dots, 0, \underset{j+1}{\cancel{1}}, 0, \dots, 0]$ , on obtient la statistique de test :

$$F(\mathbf{Y}) = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2(\mathbb{X}'\mathbb{X})_{j+1,j+1}^{-1}},$$

qui suit une loi de Fisher  $F(1, n - p)$  sous  $H_0$ .

- Le test consistant à rejeter  $H_0$  si  $F(\mathbf{Y}) > f_{1,n-p}(1 - \alpha)$  est de niveau  $\alpha$ .
- Remarque :  $F(\mathbf{Y}) = T(\mathbf{Y})^2$ . Les tests de nullité de Student et de Fisher sont équivalents. En effet, si  $U \sim \mathcal{T}(n - p)$ , alors  $U^2 \sim F(1, n - p)$ .

## Test de validité globale du modèle (avec constante) i

$$\begin{array}{lcl} H_0 & : & \beta_1 = \cdots = \beta_{p-1} = 0 \\ H_1 & : & \exists 1 \leq k \leq p-1 : \beta_k \neq 0 \end{array}$$

- Statistique de test :

$$F(Y) = \frac{1}{(p-1)\hat{\sigma}^2} \hat{\beta}' A' [A(\mathbb{X}'\mathbb{X})^{-1} A']^{-1} A \hat{\beta}, \quad A = \underbrace{(0, \mathbf{I}_{p-1})}_{(p-1) \times p},$$

qui suit une loi  $F(p-1, n-p)$  sous  $H_0$ .

- Le test consistant à rejeter  $H_0$  si  $F(\mathbf{Y}) > f_{p-1, n-p}(1 - \alpha)$  est de niveau  $\alpha$ .

## Test de validité globale du modèle (avec constante) ii

- On montre que :

$$F(\mathbf{Y}) = \frac{(\text{SCT} - \text{SCR})/(p-1)}{\text{SCR}/(n-p)} = \frac{n-p}{p-1} \frac{R^2}{1-R^2}.$$

- Interprétation des cas limite du  $R^2$  ?

## Test de validité d'un sous-modèle i

$$\begin{array}{l|l} H_0 & : \beta_{j_1} = \dots, \beta_{j_q} = 0 \\ H_1 & : \exists 1 \leq k \leq q : \beta_{j_k} \neq 0. \end{array}$$

- On pose

$$A = \begin{pmatrix} \text{constante} & & & & j_1+1 & & & \\ \hat{0} & 0 & \dots & 0 & \hat{1} & 0 & \dots & 0 \\ \vdots & \vdots \end{pmatrix}.$$

- La statistique de test est :

$$F(\mathbf{Y}) = \frac{1}{q\hat{\sigma}^2} \hat{\beta}' A' [A(\mathbb{X}'\mathbb{X})^{-1} A']^{-1} A \hat{\beta} \sim F(q, n-p) \quad \text{sous } H_0.$$

- Le test consistant à rejeter  $H_0$  si  $F(\mathbf{Y}) > f_{q,n-p}(1 - \alpha)$  est de niveau  $\alpha$ .

## Test de validité d'un sous-modèle ii

$$\begin{array}{lcl} H_0 & : & \beta_{j_1} = \dots, \beta_{j_q} = 0 \\ H_1 & : & \exists 1 \leq k \leq q : \beta_{j_k} \neq 0. \end{array}$$

$\Leftrightarrow$  tester si les observations peuvent être issues du sous-modèle

$$\mathbf{Y} = \tilde{\mathbb{X}}\tilde{\beta} + \tilde{\epsilon},$$

où  $\tilde{\mathbb{X}}$  est composée des colonnes de  $\mathbb{X}$  privées des colonnes  $j_1 + 1, \dots, j_q + 1$  et où  $\tilde{\beta} \in \mathbb{R}^{p-q}$ .

- Ajustement **modèle complet**  $\mathbf{Y} = \mathbb{X}\beta + \epsilon : \rightsquigarrow \text{SCE}, \text{SCR et } R^2$ .
- Ajustement **sous-modèle**  $\mathbf{Y} = \tilde{\mathbb{X}}\tilde{\beta} + \tilde{\epsilon} : \rightsquigarrow \widetilde{\text{SCE}}, \widetilde{\text{SCR}} \text{ et } \widetilde{R^2}$ .

$$F(\mathbf{Y}) = \frac{(\widetilde{\text{SCR}} - \text{SCR})/q}{\text{SCR}/(n-p)} = \frac{n-p}{q} \frac{R^2 - \widetilde{R^2}}{1-R^2}.$$

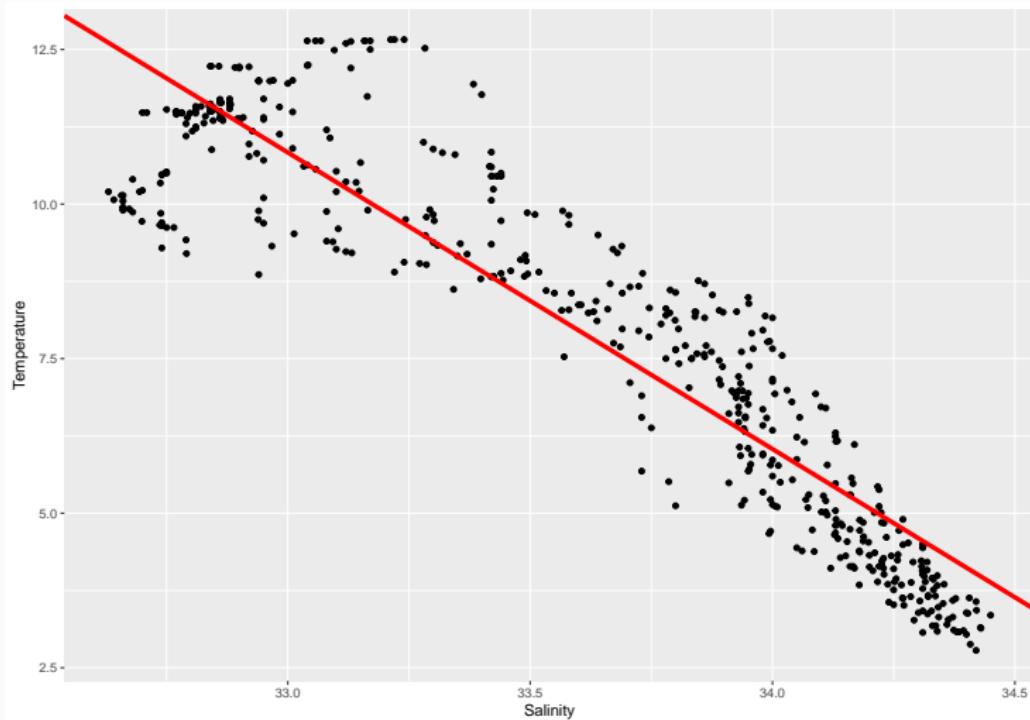
## Test d'hypothèse affine sur $\beta$

$$\left| \begin{array}{ll} H_0 & : A\beta = a \\ H_1 & : A\beta \neq a \end{array} \right. , \quad A \in \mathcal{M}_{q,p}(\mathbb{R}), \quad \text{Rk}(A) = q.$$

$$F(\mathbf{Y}) = \frac{1}{q\hat{\sigma}^2} (\hat{A}\beta - a)' [A(\mathbb{X}'\mathbb{X})^{-1}A']^{-1} (\hat{A}\beta - a) \sim F(q, n-p) \quad \text{sous } H_0.$$

- Le test consistant à rejeter  $H_0$  si  $F(\mathbf{Y}) > f_{q,n-p}(1 - \alpha)$  est de niveau  $\alpha$ .
- Propriétés : Ces tests de Fisher coïncident avec les tests de vraisemblance maximale.

## Exemple : Salinité et Température (CalCOFI) i



Droite de régression :  $y = 169.1178 - 4.79646x$ .

## Exemple : Salinité et Température (CalCOFI) ii

```
> m <- lm(Temperature~Salinity,data=bottle)
> summary(m)
```

Call:

```
lm(formula = Temperature ~ Salinity, data = bottle)
```

Residuals:

| Min      | 1Q       | Median   | 3Q      | Max     |
|----------|----------|----------|---------|---------|
| -2.79153 | -0.75022 | -0.06611 | 0.66100 | 3.04295 |

Coefficients:

|                | Estimate  | Std. Error | t value  | Pr(> t )   |         |   |
|----------------|-----------|------------|----------|------------|---------|---|
| (Intercept)    | 169.11780 | 3.03735    | 55.68    | <2e-16 *** |         |   |
| Salinity       | -4.79646  | 0.09031    | -53.11   | <2e-16 *** |         |   |
| ---            |           |            |          |            |         |   |
| Signif. codes: | 0 ‘***’   | 0.001 ‘**’ | 0.01 ‘*’ | 0.05 ‘.’   | 0.1 ‘ ’ | 1 |

Residual standard error: 1.123 on 491 degrees of freedom

Multiple R-squared: 0.8517, Adjusted R-squared: 0.8514

F-statistic: 2821 on 1 and 491 DF, p-value: < 2.2e-16

## Exemple : Salinité et Température (CalCOFI) iii

```
Residual standard error: 1.123 on 491 degrees of freedom
Multiple R-squared:  0.8517, Adjusted R-squared:  0.8514
F-statistic:  2821 on 1 and 491 DF,  p-value: < 2.2e-16
```

- $\hat{\sigma} = 1.123$
- $R^2 = 0.8517.$
- Test de validité globale du modèle (ici  $p = 2$ ) :
  - $F(\mathbf{Y}) = 2821$
  - Sous  $H_0$   $F(\mathbf{Y})$  suit une loi de Fisher à 1 et 491 degrés de liberté (donc  $n = 493$ ).
  - Au seuil  $\alpha = 5\%$ , on rejette  $H_0$ .