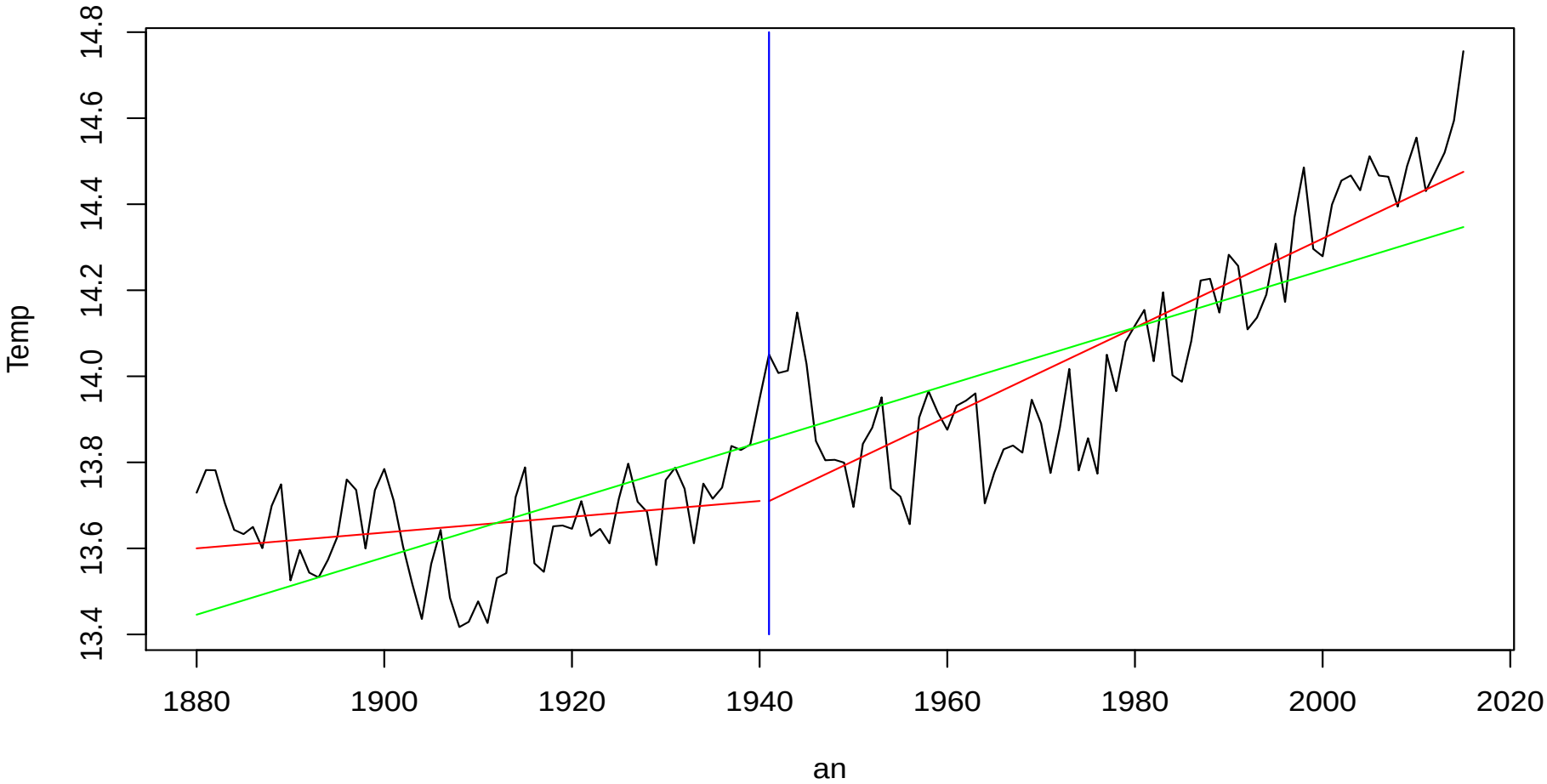
Cours d'Econométrie 2

Master M.A.E.F. Première Année

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Année 2021-2022

Importance du Modèle linéaire



Importance du Modèle linéaire (2)

Problème : Quelle variété de carottes en fonction du sol ?

	Sol	Vari	Temps
1	1	1	6
2	1	1	10
3	1	1	11
4	1	2	13
5	1	2	15
6	1	3	14
7	1	3	22
8	2	1	12
9	2	1	15
⋮	⋮	⋮	⋮

Importance du Modèle linéaire (3)

Problème : Faut-il manger du chocolat pour obtenir un prix Nobel ?

Importance du Modèle linéaire (4)

Problème : Modèliser et/ou prédire $\mathtt{Y}$ (à valeurs dans $\mathbb{R}$) en fonction de $\mathtt{X}$ (à valeurs dans $\mathbb{R}^d$) quand on observe $(X_i, Y_i)_{1 \leq i \leq n}$

$\Longrightarrow$ Le modèle linéaire et les moindres carrés

- Sont simples et explicites
- Ils fournissent une référence comparative
- Ils sont facilement interprétables

Mais ils...

- Peuvent être optimisés par l'apprentissage statistique
- Ne sont pas souples et prescrivent à l'avance un comportement
- Ne sont pas très robustes

Organisation du cours

1. Cours de 1h30 TP de 2h00
2. 2 Contrôles Continus (CC1 et CC2) de 1h30 présentiel mars et avril
3. Examen final en mai de 3h (Par)
4. Note finale $= \max\left(Par\,,\ \frac{1}{2}(CC + Par)\right)$ où $CC = \max(CC1\,,\ CC2)$

Plan du cours

Plan du cours

Le cadre général du modèle linéaire

Dans toute la suite, on supposera que :

$$(X_i, Y_i)_{1 \leq i \leq n} \text{ est } \textbf{observé}, \text{ avec } X_i \in \mathbb{R}^d, \ Y_i \in \mathbb{R},$$

- `X` est relatif aux variables exogènes (explicatives)

  Exemple : $X = (1, X_1, \ldots, X_p)$ régression multiple

- `Y` est la variable endogène (à expliquer)

  $\Longrightarrow$ Apprentissage supervisé

Le cadre général du modèle linéaire (2)

On suppose qu'existe un **modèle linéaire** liant les Y_i aux X_i :

$$Y = Z\,\theta^* + \varepsilon$$

avec $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$, $\theta^* = \begin{pmatrix} \theta_0^* \\ \theta_1^* \\ \vdots \\ \theta_p^* \end{pmatrix}$ et $Z = \begin{pmatrix} {}^tX_1 \\ {}^tX_2 \\ \vdots \\ {}^tX_n \end{pmatrix}$.

Dans la suite :

- Y et $Z \in \mathcal{M}_{(n,p+1}(\mathbb{R})$ ont été observés et sont connus ;
- ε n'est pas observé, $\theta^* \in \mathbb{R}^{p+1}$ est inconnu.

Le cadre général du modèle linéaire (3)

Exemples :

- Modèle linéaire simple : $Y_i = \theta_0^* + \theta_1^* Q_i + \varepsilon_i$ pour $i = 1, \ldots, n$ où Q variable quantitative réelle : Température en fonction de l'année

$$\Longrightarrow \text{Alors } Z = \begin{pmatrix} 1 & Q_1 \\ 1 & Q_2 \\ \vdots & \vdots \\ 1 & Q_n \end{pmatrix} \text{ avec } p = 1 \text{ et } X_i = \begin{pmatrix} 1 \\ Q_i \end{pmatrix}$$

- Modèle linéaire multiple : $Y_i = \theta_0^* + \theta_1^* X_i^{(1)} + \cdots + \theta_p^* X_i^{(p)} + \varepsilon_i$ pour $i = 1, \ldots, n$ où $X^{(1)}, \ldots, X^{(p)}$ sont p variables quantitatives réelles : Température/Année+population mondiale+éruptions volcaniques

$$\Longrightarrow \text{Alors } Z = \begin{pmatrix} 1 & X_1^{(1)} & \ldots & X_1^{(p)} \\ 1 & X_2^{(1)} & \ldots & X_2^{(p)} \\ \vdots & \vdots & & \\ 1 & X_n^{(1)} & \ldots & X_n^{(p)} \end{pmatrix} \text{ et } X_i = \begin{pmatrix} 1 \\ X_i^{(1)} \\ \vdots \\ X_i^{(p)} \end{pmatrix}$$

Le cadre général du modèle linéaire (4)

Exemples :

- Modèle à un facteur : $Y_i = \theta_1^* \mathbb{I}_{F_i=x_1} + \cdots + \theta_J^* \mathbb{I}_{F_i=x_J} + \varepsilon_i$ pour $i = 1, \ldots, n$ où F variable qualitative prenant J modalités (x_j) : Décès par Covid/continent

$$\Longrightarrow \text{Alors } Z = \begin{pmatrix} \mathbb{I}_{F_1=x_1} & \cdots & \mathbb{I}_{F_1=x_J} \\ \mathbb{I}_{F_2=x_1} & \cdots & \mathbb{I}_{F_2=x_J} \\ \vdots & & \vdots \\ \mathbb{I}_{F_n=x_1} & \cdots & \mathbb{I}_{F_n=x_J} \end{pmatrix} \text{ avec } p+1 = J$$

- Modèle linéaire simple avec facteur : $Y_i = \sum_{j=1}^{J} \theta_{0,j}^* \mathbb{I}_{F_i=x_j} + \sum_{j=1}^{J} \theta_{1,j}^* \mathbb{I}_{F_i=x_j} Q_i + \varepsilon_i$ pour $i = 1, \ldots, n$ où Q variable quantitative et F variable qualitative prenant J modalités (x_j) : réelles : Temps germination/Heures soleil+variété

But et méthode

But : • Estimer θ^*

- Tester si un modèle linéaire est légitime
- Interpréter et prédire avec un modèle linéaire.

Méthode : Moindres carrés :

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^{p+1}}{\text{Argmin}} \left\| Y - Z\,\theta \right\|^2 \quad \text{où} \quad \|U\|^2 = {}^t U\, U \quad \text{pour } U \in \mathbb{R}^n$$

Remarque : Choix de critère arbitraire ! Autre choix : moindres valeurs absolues $\widetilde{\theta} = \underset{\theta \in \mathbb{R}^{p+1}}{\text{Argmin}} \left| Y - Z\,\theta \right|$ où $|U| = \sum_{i=1}^{n} |U_i|$ pour $U \in \mathbb{R}^n$

Plan du cours

Les hypothèses "classiques"

Hypothèses pour le modèle linéaire : $Y = Z\,\theta^* + \varepsilon$

0. **(A0)** $n \geq p+1$ et Z de rang $p+1$
   $\Longrightarrow$ la matrice $(p+1)\times(p+1)$ ${}^tZ\,Z$ est inversible ;
1. **(A1)** $\mathbb{E}(\varepsilon_i) = 0$ pour tout $i = 1, \ldots, n$;
2. **(A2) Homoscédasticité :** $\mathrm{var}(\varepsilon_i) = \sigma^2$ pour tout $i = 1, \ldots, n$;
3. **(A3) Non corrélation :** pour tout $i \neq j$, $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$;
4. **(A4) Gaussianité :** ε vecteur gaussien ;

Remarque : De **(A1)-(A4)**, on a $\varepsilon \overset{\mathcal{L}}{\sim} \mathcal{N}_n\big(0\,,\,\sigma^2\,I_n\big)$

Les prémices

Propriété

Pour le produit scalaire "classique" dans $\mathbb{R}^n$, $< U, V >= {}^t U\, V$, *si* A *sev de* $\mathbb{R}^n$, *on note* P_A *la projection orthogonale sur* A. *Alors si* P_A *dénote aussi la matrice de* P_A *dans la base canonique :*

- $P_A = {}^t P_A$, $(P_A)^2 = P_A$ *et* $I_n - P_A = P_{A^\perp}$;
- *Les valeurs propres de* P_A *sont* $\begin{cases} 1 \text{ avec sev propre } A \\ 0 \text{ avec sev propre } A^\perp \end{cases}$.

Proposition

Pour le modèle $Y = Z\,\theta^* + \varepsilon$ *et* $\widehat{\theta} = \underset{\theta \in \mathbb{R}^{p+1}}{\text{Argmin}} \left\| Y - Z\,\theta \right\|^2$ *et sous* **(A0)**, *avec* $[Z] = \left\{ Z\,\gamma,\ \gamma \in \mathbb{R}^{p+1} \right\}$ *sev de* $\mathbb{R}^n$,

$$Z\,\widehat{\theta} = P_{[Z]}\,Y = Z\left({}^t Z\, Z\right)^{-1} {}^t Z\, Y \quad \Longrightarrow \quad \widehat{\theta} = \left({}^t Z\, Z\right)^{-1} {}^t Z\, Y.$$

Les prémices (2)

Démonstration.

On remarque que $\min_{\theta \in \mathbb{R}^{p+1}} \left\| Y - Z\,\theta \right\|^2 = \min_{U \in [Z]} \left\| Y - U \right\|^2$. Si $U = P_{[Z]}\,Y + V$ avec $V \in [Z]$, alors $\left\| Y - U \right\|^2 = \left\| P_{[Z]^\perp}\,Y - V \right\|^2 = \left\| P_{[Z]^\perp}\,Y \right\|^2 + \|V\|^2$ par Pythagore. Donc minimum pour $V = 0$.
Montrons que $P_{[Z]} = Z\left({}^tZ\,Z\right)^{-1}{}^tZ$. Si $U \in [Z]$, d'où $U = Z\,\gamma$, alors $P_{[Z]}U = Z\left({}^tZ\,Z\right)^{-1}{}^tZ\,Z\,\gamma = Z\,\gamma = U$. De plus, pour $Y \in \mathbb{R}^n$, alors on doit avoir $Y - P_{[Z]}Y \in [Z]^\perp$. Or ${}^tZ\left(Y - Z\left({}^tZ\,Z\right)^{-1}{}^tZ\,Y\right) = {}^tZ\,Y - {}^tZ\,Z\left({}^tZ\,Z\right)^{-1}{}^tZ\,Y = 0$ donc pour tout $Y \in \mathbb{R}^n$, on a bien $Y - Z\left({}^tZ\,Z\right)^{-1}{}^tZ\,Y \in [Z]^\perp$.
Enfin, sous **(A0)**, Z de rang $p+1$, alors $Z\,\theta = 0 \Longrightarrow \theta = 0$. Donc si $Z\,\widehat{\theta} = Z\left(\left({}^tZ\,Z\right)^{-1}{}^tZ\,Y\right)$ alors $\widehat{\theta} = \left({}^tZ\,Z\right)^{-1}{}^tZ\,Y$. □

Premiers résultats

Propriété

Sous **(A0)-(A3)**, *alors :*

1. $\widehat{\theta} = \theta^* + \left({}^t Z Z\right)^{-1} {}^t Z \varepsilon$ *d'où* $\mathbb{E}[\widehat{\theta}] = \theta^*$ *(sans biais),* $cov(\widehat{\theta}) = \sigma^2 \left({}^t Z Z\right)^{-1}$
2. *Avec* $\widehat{Y} = Z\,\widehat{\theta} = P_{[Z]}\,Y$ *(prédiction),* $\mathbb{E}[\widehat{Y}] = Z\,\theta^*$ *et* $cov(\widehat{Y}) = \sigma^2\,P_{[Z]}$
3. *Avec* $\widehat{\varepsilon} = Y - \widehat{Y} = P_{[Z]^\perp}\,Y = P_{[Z]^\perp}\,\varepsilon$, $\mathbb{E}[\widehat{\varepsilon}] = 0$ *et* $cov(\widehat{\varepsilon}) = \sigma^2\,P_{[Z]^\perp}$

Démonstration.

1. On a $\widehat{\theta} = \left({}^t Z\,Z\right)^{-1} {}^t Z\,Y = \left({}^t Z\,Z\right)^{-1} {}^t Z\,(Z\,\theta^* + \varepsilon) = \theta^* + \left({}^t Z\,Z\right)^{-1} {}^t Z\,\varepsilon$. D'où $\mathbb{E}[\widehat{\theta}] = \theta^* + \mathbb{E}\left[\left({}^t Z\,Z\right)^{-1} {}^t Z\,\varepsilon\right] = \theta^*$ car $\mathbb{E}[\varepsilon] = 0$ par **(A1)**. Et $\text{cov}(\widehat{\theta}) = \text{cov}\left(\left({}^t Z\,Z\right)^{-1} {}^t Z\,\varepsilon\right) = \left({}^t Z\,Z\right)^{-1} {}^t Z\,\text{cov}(\varepsilon)\,{}^t\left(\left({}^t Z\,Z\right)^{-1} {}^t Z\right) = \left({}^t Z\,Z\right)^{-1} {}^t Z\,\sigma^2\,I_n\,Z\,\left({}^t Z\,Z\right)^{-1}$ en utilisant $\text{cov}(A + B\,X) = B\,\text{cov}(X)\,{}^t B$ pour A et B matrices composées de nombres réels. D'où $\text{cov}(\widehat{\theta}) = \sigma^2\,\left({}^t Z\,Z\right)^{-1} {}^t Z\,Z\,\left({}^t Z\,Z\right)^{-1} = \sigma^2\,\left({}^t Z\,Z\right)^{-1}$.
2. $\mathbb{E}[\widehat{Y}] = Z\,\mathbb{E}[\widehat{\theta}] = Z\,\theta^*$ et $\text{cov}(\widehat{Y}) = \text{cov}\left(P_{[Z]}\,Y\right) = P_{[Z]}\text{cov}(Y)\,{}^t P_{[Z]} = \sigma^2\,P_{[Z]}$.
3. On a bien $P_{[Z]^\perp}\,Y = P_{[Z]^\perp}\left(Z\,\theta^* + \varepsilon\right) = P_{[Z]^\perp}\,\varepsilon$ d'où $\mathbb{E}[\widehat{\varepsilon}] = P_{[Z]^\perp}\,\mathbb{E}[\varepsilon] = 0$. $\text{cov}(\widehat{\varepsilon}) = \text{cov}\left(P_{[Z]^\perp}\,\varepsilon\right) = P_{[Z]^\perp}\text{cov}(\varepsilon)\,{}^t P_{[Z]^\perp} = \sigma^2\,P_{[Z]^\perp}$.

Premiers résultats (3)

Pourquoi choisir les moindres carrés ? Trois raisons :

1. Les formules sont explicites et immédiatement calculables ;
2. $\widehat{\theta}$ est l'estimateur du MV de θ^* sous **(A0)-(A4)** ;
3. Le Théorème de Gauss-Markov donne une optimalité à cet estimateur :

Théorème

Sous **(A0)-(A4)**, *si* $\widetilde{\theta}$ *autre estimateur linéaire* ($\widetilde{\theta} = MY$, M *matrice*) *et sans biais de* θ^* ($\mathbb{E}[\widetilde{\theta}] = \theta^*$) *alors* $cov(\widehat{\theta}) \leq cov(\widetilde{\theta})$.

Démonstration.

Voir TD !

□

Premiers résultats (4)

Conséquence : Si on rajoute **(A4)**, on a $\begin{cases} \widehat{\theta} \overset{\mathcal{L}}{\sim} \mathcal{N}_{p+1}\left(\theta^* , \sigma^2 \left({}^t ZZ\right)^{-1}\right) \\ \widehat{Y} \overset{\mathcal{L}}{\sim} \mathcal{N}_n\left(Z\,\theta^* , \sigma^2 P_{[Z]}\right) \\ \widehat{\varepsilon} \overset{\mathcal{L}}{\sim} \mathcal{N}_n\left(0 , \sigma^2 P_{[Z]^\perp}\right) \end{cases}$

Démonstration.

Sous **(A4)**, ε est un vecteur gaussien, Y aussi, ainsi que $A + B\,Y$ pour A et B matrices réelles. □

Obtenir des intervalles de confiance ou tests sur $\widehat{\theta}$, $\widehat{Y}$ ou $\widehat{\varepsilon} \Longrightarrow$ Estimer σ^2 !

Définition

Sous **(A0)-(A3)**, *on définit l'estimateur* $\widehat{\sigma^2} = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$

Remarque : Sous **(A0)-(A4)**, $\left(\widehat{\theta}, \frac{1}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2\right)$ estimateur du MV de (θ^*, σ^2).

Premiers résultats (5)

Propriété

Sous **(A0)-(A3)**, *alors :* $\mathbb{E}\left[\widehat{\sigma^2}\right] = \sigma^2$ *(estimateur sans biais)*

Démonstration.

On a $\widehat{\sigma^2} = \frac{1}{n-(p+1)} \left\|\widehat{\varepsilon}\right\|^2 = \frac{1}{n-(p+1)} {}^t\widehat{\varepsilon}\widehat{\varepsilon} = \frac{1}{n-(p+1)} {}^t\left(P_{[Z]^\perp}\,\varepsilon\right)\left(P_{[Z]^\perp}\,\varepsilon\right) = \frac{1}{n-(p+1)} {}^t\varepsilon\, P_{[Z]^\perp}\varepsilon$.
Comme $\widehat{\sigma^2}$ est à valeurs réelles, on a $\text{Trace}(\widehat{\sigma^2}) = \widehat{\sigma^2} = \frac{1}{n-(p+1)} \text{Trace}({}^t\varepsilon\, P_{[Z]^\perp}\varepsilon)$. Or pour A et B 2 matrices telles que $A\,B$ et $B\,A$ existent, alors $\text{Trace}(A\,B) = \text{Trace}(B\,A)$. D'où $\text{Trace}({}^t\varepsilon\, P_{[Z]^\perp}\varepsilon) = \text{Trace}(P_{[Z]^\perp}\varepsilon\,{}^t\varepsilon)$. Les opérateurs $\mathbb{E}$ et Trace étant tous deux linéaires, on a $\mathbb{E}\left[\text{Trace}(U)\right] = \text{Trace}\left(\mathbb{E}\left[U\right]\right)$ pour U une matrice aléatoire. D'où $\mathbb{E}\left[\widehat{\sigma^2}\right] = \frac{1}{n-(p+1)} \text{Trace}\left(\mathbb{E}\left[P_{[Z]^\perp}\varepsilon\,{}^t\varepsilon\right]\right) = \frac{1}{n-(p+1)} \text{Trace}\left(P_{[Z]^\perp}\mathbb{E}\left[\varepsilon\,{}^t\varepsilon\right]\right)$. Mais $\mathbb{E}\left[\varepsilon\,{}^t\varepsilon\right] = \text{cov}(\varepsilon) = \sigma^2\, I_n$. Donc $\mathbb{E}\left[\widehat{\sigma^2}\right] = \frac{\sigma^2}{n-(p+1)} \text{Trace}\left(P_{[Z]^\perp}\right)$. Enfin, avec la propriété sur les valeurs propres d'une matrice de projection, on en déduit que $\text{Trace}\left(P_{[Z]^\perp}\right) = \dim\left([Z]^\perp\right) = n-(p+1)$. D'où $\mathbb{E}\left[\widehat{\sigma^2}\right] = \sigma^2$. □

Le cas gaussien

Théorème (Théorème de Cochran)

Soit $(A_1, \dots, A_\ell)$ *des sev de* $\mathbb{R}^n$ *tels que* $A_1 \overset{\perp}{\oplus} A_2 \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} A_\ell$. *Soit* $\varepsilon \overset{\mathcal{L}}{\sim} \mathcal{N}\left(0\,,\,\sigma^2\, I_n\right)$. *Alors :*

1. *Les vecteurs* $\left(P_{A_i}\varepsilon\right)_i$ *sont gaussiens et indépendants ;*
2. *Les variables* $\left\|P_{A_i}\varepsilon\right\|^2$ *suivent la loi* $\sigma^2\,\chi^2\left(\dim(A_i)\right)$.

Démonstration.

1. Comme P_{A_i} est une matrice de réels et ε un vecteur gaussien, alors $P_{A_i}\varepsilon$ est un vecteur gaussien, centré car ε est centré. Donc $\left(P_{A_i}\varepsilon\right)_i$ est une famille indépendante si $\mathrm{cov}\left(P_{A_i}\varepsilon\,,\,P_{A_j}\varepsilon\right) = 0$ pour $i \neq j$. Or $\mathrm{cov}\left(P_{A_i}\varepsilon\,,\,P_{A_j}\varepsilon\right) = \mathbb{E}\left[P_{A_i}\varepsilon\,{}^t\left(P_{A_j}\varepsilon\right)\right] = \mathbb{E}\left[P_{A_i}\varepsilon\,{}^t\varepsilon^t P_{A_j}\right] = P_{A_i}\mathrm{cov}(\varepsilon)^t P_{A_j} = \sigma^2\, P_{A_i}P_{A_j}$. Mais $A_i \perp A_j$ donc $P_{A_i}P_{A_j} = 0$, d'où le résultat.

□

Le cas gaussien (2)

Démonstration.

2. $\left\|P_{A_i}\varepsilon\right\|^2 = {}^t\varepsilon\,{}^tP_{A_i}P_{A_i}\varepsilon = {}^t\varepsilon\,P_{A_i}\,\varepsilon$. Mais P_{A_i} est une matrice réelle symétrique donc diagonalisable et on peut écrire que $P_{A_i} = Q_i\,D_i\,{}^tQ_i$ avec Q_i une matrice orthogonale et D_i une matrice diagonale avec les valeurs propres de P_{A_i}, donc par exemple d'abord $\dim(A_i)$ uns sur la diagonale puis dessous $n - \dim(A_i)$ zéros. D'où $\left\|P_{A_i}\varepsilon\right\|^2\,{}^t\left(Q_i\,\varepsilon\right)\,D_i\,\left({}^tQ_i\,\varepsilon\right)$. Soit $\varepsilon' = {}^tQ_i\,\varepsilon$. Alors $\varepsilon' \overset{\mathcal{L}}{\sim} \mathcal{N}_n\left(0\,,\,\sigma^2\,{}^tQ_i\,Q_i\right) = \mathcal{N}_n\left(0\,,\,\sigma^2\,I_n\right)$ car Q_i est une matrice orthogonale (donc $Q_i^{-1} = {}^tQ_i$). Donc ε' est un vecteur gaussien centré standard. Comme ${}^t\varepsilon'\,D_i\,\varepsilon' = \sum_{j=1}^{\dim(A_i)}(\varepsilon'_j)^2$, les ε'_j étant des variables gaussiennes centrées indépendantes de même variance σ^2, on a donc $\sum_{j=1}^{\dim(A_i)}(\varepsilon'_j)^2 \overset{\mathcal{L}}{\sim} \sigma^2\,\chi^2(\dim(A_i))$.

□

Le cas gaussien (3)

Proposition

Sous **(A0)-(A4)**,

1. *On a* $\widehat{Y}$ *et* $\widehat{\varepsilon}$ *indépendants ;*
2. *On a* $\widehat{\sigma^2} \overset{\mathcal{L}}{\sim} \frac{\sigma^2}{n-(p+1)} \chi^2(n-(p+1))$ *et* $\widehat{\sigma^2}$ *indépendant de* $\widehat{Y}$ *et de* $\widehat{\theta}$.

Démonstration.

1. On a $\widehat{Y} = Z\theta^* + P_{[Z]}\,\varepsilon$ et $\widehat{\varepsilon} = P_{[Z]^\perp}\,\varepsilon$. Avec Cochran, $P_{[Z]}\,\varepsilon$ et $P_{[Z]^\perp}\,\varepsilon$ sont indépendants.
2. On a $\widehat{\sigma^2} = \frac{1}{n-(p+1)} \left\| P_{[Z]^\perp}\,\varepsilon \right\|^2$, donc d'après Cochran, $\widehat{\sigma^2} \overset{\mathcal{L}}{\sim} \frac{\sigma^2}{n-(p+1)} \chi^2\left(\dim([Z]^\perp)\right)$ d'où le résultat. Mais $\widehat{\sigma^2}$ fonction mesurable de $P_{[Z]^\perp}\,\varepsilon$ indépendant de $\widehat{Y}$, donc $\widehat{\sigma^2}$ indépendant de $\widehat{Y}$. Enfin, $\widehat{\theta} = ({}^tZ\,Z)^{-1}\,{}^tZ\,Y = ({}^tZ\,Z)^{-1}\,{}^tZ\,Z\,({}^tZ\,Z)^{-1}\,{}^tZ\,Y = ({}^tZ\,Z)^{-1}\,{}^tZ\,\widehat{Y}$. Comme $\widehat{\sigma^2}$ indépendant de $\widehat{Y}$, alors $\widehat{\sigma^2}$ indépendant de $\widehat{\theta}$.

□

Remarque : Sous **(A0)-(A3)**, on a juste $\widehat{Y}$ et $\widehat{\varepsilon}$ non corrélés puisque $\operatorname{cov}\left(\widehat{Y}\,,\,\widehat{\varepsilon}\right) = \operatorname{cov}\left(Z\theta^* + P_{[Z]}\,\varepsilon\,,\,P_{[Z]^\perp}\,\varepsilon\right) = P_{[Z]}\operatorname{cov}(\varepsilon)\,{}^tP_{[Z]^\perp} = 0$.

Applications aux tests statistiques

Rappels :

1. Si $X \overset{\mathcal{L}}{\sim} \mathcal{N}(0,1)$ indépendante de $Z \overset{\mathcal{L}}{\sim} \chi^2(\ell)$, $\dfrac{X}{\sqrt{\frac{1}{\ell} Z}} \overset{\mathcal{L}}{\sim} t(\ell)$: Student
2. Si $Z_1 \overset{\mathcal{L}}{\sim} \chi^2(\ell_1)$ indépendante de $Z_2 \overset{\mathcal{L}}{\sim} \chi^2(\ell_2)$, $\dfrac{\frac{1}{\ell_1} Z_1}{\frac{1}{\ell_2} Z_2} \overset{\mathcal{L}}{\sim} F(\ell_1, \ell_2)$: Fisher

Proposition

Sous **(A0)-(A4)**, *et pour* $C \in \mathbb{R}^{p+1}$ *fixé, on a :*

$$\frac{1}{\sqrt{\widehat{\sigma^2}}} \left({}^t C \left({}^t Z\, Z \right)^{-1} C \right)^{-1/2} \left({}^t C \left(\widehat{\theta} - \theta^* \right) \right) \overset{\mathcal{L}}{\sim} t(n-(p+1))$$

Démonstration.

(A0)-(A4) $\Longrightarrow {}^t C(\widehat{\theta} - \theta^*)$ est une variable gaussienne, avec $\mathbb{E}\left[{}^t C(\widehat{\theta} - \theta^*) \right] = 0$ et $\operatorname{cov}\left({}^t C(\widehat{\theta} - \theta^*) \right) = {}^t C \operatorname{cov}(\widehat{\theta})\, C = \sigma^2 {}^t C \left({}^t Z Z \right)^{-1} C \Longrightarrow {}^t C(\widehat{\theta} - \theta^*) \overset{\mathcal{L}}{\sim} \mathcal{N}\left(0, \sigma^2 {}^t C \left({}^t Z Z \right)^{-1} C \right)$.

Ceci entraîne $\frac{1}{\sigma} \left({}^t C \left({}^t Z Z \right)^{-1} C \right)^{-1/2} {}^t C(\widehat{\theta} - \theta^*) \overset{\mathcal{L}}{\sim} \mathcal{N}(0,1)$. Ensuite on utilise le fait que $\dfrac{\widehat{\sigma^2}}{\sigma^2} \overset{\mathcal{L}}{\sim} \chi^2\left(n-(p+1)\right)/(n-(p+1))$ et que $\widehat{\sigma^2}$ est indépendante de ${}^t C(\widehat{\theta} - \theta^*)$ □

Applications aux tests statistiques (2)

Test de Student : Soit le problème de test $\begin{cases} H_0 : {}^t C\,\theta^* = 0 \\ H_1 : {}^t C\,\theta^* \neq 0 \end{cases}$

Proposition

Sous **(A0)-(A4)**, *et sous l'hypothèse* H_0 *alors :*

$$\widehat{T} = \frac{1}{\sqrt{\widehat{\sigma^2}}} \left({}^t C \left({}^t Z\, Z\right)^{-1} C\right)^{-1/2} {}^t C\,\widehat{\theta} \overset{\mathcal{L}}{\approx} t(n-(p+1))$$

Exemples d'utilisation :

- Avec ${}^t C = (0, \ldots, 0, 1, 0, \ldots, 0)$, on teste $H_0 : \theta_j^* = 0$
  $\Longrightarrow$ On teste ainsi si $X^{(j)}$ est significative
- Avec ${}^t C = (0, \ldots, 0, \pm 1, 0, \ldots, 0, \pm 1, 0, \ldots, 0)$, on teste $H_0 : \theta_j^* = \theta_\ell^*$
  $\Longrightarrow$ On teste ainsi la différence d'effet entre 2 modalités

Applications aux tests statistiques (3)

On s'intéresse à plusieurs combinaisons linéaires des paramètres :

Proposition

Sous **(A0)-(A4)**, *et pour* $C \in \mathcal{M}_{p+1,q}(\mathbb{R})$ *fixé avec* $1 \leq q \leq p+1$ *tel que le rang de* C *est* q, *on a :*

$$\frac{1}{\sigma^2}\,{}^t(\widehat{\theta}-\theta^*)C\left({}^tC\left({}^tZ\,Z\right)^{-1}C\right)^{-1}{}^tC\,(\widehat{\theta}-\theta^*) \overset{\mathcal{L}}{\approx} \chi^2(q)$$

Démonstration.

Utilisant la proposition précédente, $\frac{1}{\sigma}\left({}^tC\left({}^tZZ\right)^{-1}C\right)^{-1/2}{}^tC(\widehat{\theta}-\theta^*) \overset{\mathcal{L}}{\approx} \mathcal{N}_q(0\,,\,I_q)$. D'où $\left\|\frac{1}{\sigma}\left({}^tC\left({}^tZZ\right)^{-1}C\right)^{-1/2}{}^tC(\widehat{\theta}-\theta^*)\right\|^2 \overset{\mathcal{L}}{\approx} \chi^2(q)$. On finit la démonstration en montrant que $\left\|\frac{1}{\sigma}\left({}^tC\left({}^tZZ\right)^{-1}C\right)^{-1/2}{}^tC(\widehat{\theta}-\theta^*)\right\|^2 = \frac{1}{\sigma^2}\,{}^t(\widehat{\theta}-\theta^*)C\left({}^tC\left({}^tZ\,Z\right)^{-1}C\right)^{-1}{}^tC\,(\widehat{\theta}-\theta^*)$. □

Applications aux tests statistiques (4)

Test de Fisher : Soit le problème de test $\begin{cases} H_0 : {}^t C\,\theta^* = {}^t(0,\dots,0) \\ H_1 : {}^t C\,\theta^* \neq {}^t(0,\dots,0) \end{cases}$

Proposition

Sous **(A0)-(A4)**, *et sous l'hypothèse* H_0 *alors :*

$$\widehat{F} = \frac{\frac{1}{q}\,{}^t\widehat{\theta} C \left({}^t C \left({}^t Z\, Z\right)^{-1} C\right)^{-1} {}^t C\,\widehat{\theta}}{\widehat{\sigma^2}} \overset{\mathcal{L}}{\approx} F\left(q\,,\, n-(p+1)\right)$$

Démonstration.

On utilise la convergence précédente en remarquant que sous H_0 alors ${}^t C\,\theta^* = 0$, donc θ^* n'apparaît pas dans la statistique $\widehat{F}$. De plus, on a bien $\widehat{\sigma^2} \overset{\mathcal{L}}{\approx} \frac{\sigma^2}{n-(p+1)}\,\chi^2\left(n-(p+1)\right)$ d'après Cochran. La preuve sera établie si l'on montre que le numérateur et le dénominateur de $\widehat{F}$ sont indépendants. Mais le numérateur est une fonction déterministe de $\widehat{\theta}$ donc de $P_{[Z]}\varepsilon$ quand le dénominateur est une fonction déterministe de $P_{[Z]^\perp}\varepsilon$. Or d'après Cochran, comme $[Z]^\perp \perp [Z]$ alors $P_{[Z]}\varepsilon$ et $P_{[Z]^\perp}\varepsilon$ sont des vecteurs aléatoires indépendants. □

Applications aux tests statistiques (5)

Exemples d'utilisation :

- Avec ${}^tC = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$, on teste $H_0 : \ \theta_j^* = 0,\ j = 1, \dots, p$

  $\Longrightarrow$ On teste ainsi grossièrement le modèle

- Si ${}^tC = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & -1 \end{pmatrix}$, test de $H_0 : \ \theta_j^* = \theta_1^*,\ j = 1, \dots, p$

  $\Longrightarrow$ On teste ainsi la significativité d'un facteur

- Plus généralement, on peut tester la significativité d'une interaction de facteurs, de possibles regroupement de facteurs ou modalités...

Applications aux tests statistiques (6)

Autre expression du test de Fisher :

Soit le modèle linéaire $Y = Z\,\theta^* + \varepsilon$ et un "sous-modèle" linéaire $Y = Z_{(0)}\,\theta^*_{(0)} + \varepsilon$ avec $[Z_{(0)}] \subset [Z]$ avec $\dim([Z_{(0)}]) = 1 + p_0$ et $p_0 < p$.

On considère le problème de test : $\begin{cases} H_0 : \text{ Vrai modèle } Y = Z_{(0)}\,\theta^*_{(0)} + \varepsilon \\ H_1 : \text{ Vrai modèle } Y = Z\,\theta^* + \varepsilon \end{cases}$

Proposition

$$\widehat{F} = \frac{\frac{1}{p-p_0}\left(\|P_{[Z_0]^\perp} Y|^2 - \|P_{[Z]^\perp} Y\|^2\right)}{\widehat{\sigma^2}} \Longrightarrow \widehat{F} \overset{\mathcal{L}}{\approx} F\big(p - p_0, n - (p+1)\big) \text{ sous } H_0$$

Exemples d'application : • Régression polynomiale : $\begin{cases} H_0 : \text{ Degré} = p_0 \\ H_1 : \text{ Degré} = p > p_0 \end{cases}$

• Analyse de la variance : $\begin{cases} H_0 : \text{ Facteur non significatif} \\ H_1 : \text{ Facteur significatif} \end{cases}$

Applications aux tests statistiques (7)

Démonstration.

Comme $[Z_{(0)}] \subset [Z]$, alors $[Z]^\perp \subset [Z_{(0)}]^\perp$. Notons $A = [Z] \cap [Z_{(0)}]^\perp$, d'où $[Z]^\perp \overset{\perp}{\oplus} A = [Z_{(0)}]^\perp$ et ainsi $\left\|P_{[Z_0]^\perp} Y\right\|^2 - \left\|P_{[Z]^\perp} Y\right\|^2 = \left\|P_A Y\right\|^2 = \left\|P_A \varepsilon\right\|^2$ par Pythagore et sous l'hypothèse H_0, donc $Z\,\theta^* \in [Z_{(0)}]$. Comme ε vecteur gaussien, avec Cochran, on a donc $\frac{1}{p-p_0}\left\|P_A \varepsilon\right\|^2 \overset{\mathcal{L}}{\sim} \frac{\sigma^2}{p-p_0}\chi^2(p-p_0)$ car $\dim(A) = \dim([Z]) - \dim([Z_{(0)}])$.

Par ailleurs, au dénominateur, on a vu que $\widehat{\sigma^2} = \frac{1}{n-(p+1)}\left\|P_{[Z]^\perp}\varepsilon\right\|^2 \overset{\mathcal{L}}{\sim} \frac{\sigma^2}{n-(p+1)}\chi^2(n-(p+1))$ d'après Cochran. Mais on a également $A \perp [Z]^\perp$ puisque $A = [Z] \cap [Z_{(0)}]^\perp \subset [Z]$. Donc d'après Cochran également, $P_{[Z]^\perp}\varepsilon$ et $P_A\varepsilon$ sont deux vecteurs gaussiens de $\mathbb{R}^n$ indépendants.

On a donc bien $\widehat{F} \overset{\mathcal{L}}{\sim} F(p-p_0, n-(p+1))$ sous H_0. □

Remarque : Ce test de Fisher est un cas particulier du précédent : comme le rang de $[Z_{(0)}]$ est p_0+1, il existe p_0+1 colonne de la matrice Z engendrant $[Z_{(0)}]$. Les $p-p_0$ autres colonnes, d'indice $i_1, \ldots, i_{p-p_0}$ engendrent $A^\perp$. Aussi H_0 est équivalente à $\theta^*_{i_1} = \cdots = \theta^*_{i_{p-p_0}} = 0$

Applications aux intervalles de confiance

On peut déduire des résultats précédents :

1. Sur les θ_j^* : avec q_α quantile de niveau $1-\alpha/2$ d'une $t(n-(p+1))$:

$$\left[\widehat{\theta}_j - q_\alpha \sqrt{\widehat{\sigma^2}\left(({}^tZ\,Z)^{-1}\right)_{j+1,j+1}}\,,\ \widehat{\theta}_j + q_\alpha \sqrt{\widehat{\sigma^2}\left(({}^tZ\,Z)^{-1}\right)_{j+1,j+1}}\,\right]$$

2. Sur une prédiction $\widehat{Y}_{n+1}$ quand $z_{n+1} = {}^t(1, X_{n+1}^{(1)}, \ldots, X_{n+1}^{(p)})$ est observé : avec q_α quantile de niveau $1-\alpha/2$ d'une $t(n-(p+1))$:

$$\left[\widehat{Y}_{n+1} - q_\alpha \sqrt{\widehat{\sigma^2}\left({}^tz_{n+1}({}^tZ\,Z)^{-1}z_{n+1}+1\right)}\,,\right.$$
$$\left.\widehat{Y}_{n+1} + q_\alpha \sqrt{\widehat{\sigma^2}\left({}^tz_{n+1}({}^tZ\,Z)^{-1}z_{n+1}+1\right)}\,\right]$$

Coefficient de détermination R^2

On associe à une régression le coefficient de détermination R^2 défini par :

$$R^2 = \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y}_n)^2}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2} = 1 - \frac{\left\|\widehat{\varepsilon}\right\|^2}{\left\|Y - \overline{Y}_n \mathbb{1}_n\right\|^2} \qquad \text{avec} \quad \mathbb{1}_n = {}^t(1, \ldots, 1)$$

R^2 mesure le "pouvoir de prédiction" du modèle : plus $R^2 \uparrow 1$ meilleur il est

Attention ! Il est possible d'avoir $R^2 = 0.99$ sans avoir le bon modèle et avoir $R^2 = 0.01$ alors que le modèle est le bon !

Propriété

Pour le problème de test $\begin{cases} H_0 : \ \theta_i^* = 0, \ 1 \leq i \leq p \\ H_1 : \ \exists i_0, \ \theta_{i_0}^* \neq 0 \end{cases} \Longrightarrow$ *statistique de Fisher*

$$\widehat{F} = \frac{\frac{1}{p}\left(\|P_{[\mathbb{1}_n]^\perp} Y|^2 - \|P_{[Z]^\perp} Y\|^2\right)}{\widehat{\sigma^2}} = \frac{n-(p+1)}{p}\left(\frac{R^2}{1-R^2}\right).$$

Plan du cours

Quelle asymptotique ?

Dans la suite on conserve le modèle $Y = Z\,\theta^* + \varepsilon$.

Asymptotique choisie : Le nombre p de variables est fixé mais le nombre n d'invidus tend vers $+\infty$

$$\Longrightarrow \begin{cases} Y = Y^{(n)} \text{ : vecteur de taille} \to \infty \\ Z = Z^{(n)} \text{ : matrice dont le nombre de lignes} \to \infty \\ \varepsilon = \varepsilon^{(n)} \text{ : vecteur de taille} \to \infty \end{cases}$$

Mais θ^* vecteur de taille $p+1$ constant

Comportements asymptotiques de $\widehat{\theta}^{(n)}$, $\widehat{T}^{(n)}$ et $\widehat{F}^{(n)}$?

Résultats classiques

Théorème (Loi forte des grands nombres)

Si $(X_i)_{i\in\mathbb{N}}$ *suite de v.a.i.i.d. alors :*

$$\left(\mathbb{E}\left[|X_0|\right] < \infty\right) \Longleftrightarrow \left(\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p.s.} \mathbb{E}[X_0]\right)$$

Théorème (Théorème de la limite centrale)

Si $(X_i)_{i\in\mathbb{N}}$ *suite de v.a.i.i.d. telle que* $\mathbb{E}\left[X_0^2\right] < \infty$ *alors*

$$\sqrt{n}\left(\overline{X}_n - \mathbb{E}[X_0]\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0\,,\, var(X_0)\right).$$

Un nouveau théorème limite

Problème : • Les (Y_i) ne sont pas une suite i.i.d.

• $\widehat{\theta}^{(n)} = \theta^* + \left({}^t Z\, Z\right)^{-1} {}^t Z\, \varepsilon$ sous **A0-A1**, d'où $\widehat{\theta}_i^{(n)} = \theta_i^* + \sum_{k=1}^n \alpha_{i,k}\, \varepsilon_k$.
Quelle limite ?

Théorème (**Théorème de Lindeberg**)

$(U_i)_{i\in\mathbb{N}}$ suite de v.a.i.i.d. avec $\mathbb{E}[U_0] = 0$ et $\mathbb{E}\left[U_0^2\right] = 1$, $(a_j^{(n)})_{1\leq j\leq n,\, n\in\mathbb{N}^*}$ tableau triangulaire de réels tel que $\sum_{j=1}^n \left(a_j^{(n)}\right)^2 = 1$ pour tout $n \in \mathbb{N}^*$. Alors

$$\max_{1\leq j\leq n} |a_j^{(n)}| \underset{n\to+\infty}{\longrightarrow} 0 \iff \sum_{j=1}^n a_j^{(n)}\, U_j \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0\,,1).$$

Exemple d'utilisation : $a_j^{(n)} = \frac{1}{\sqrt{n}} \implies \sqrt{n}\,\overline{U}_n \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0\,,1)$.

Un nouveau théorème limite (2)

Démonstration.

Soit $Z_n = \sum_{j=1}^n a_j^{(n)}\, U_j$. Alors pour tout $u \in \mathbb{R}$, $\phi_{Z_n}(u) = \mathbb{E}\big[e^{i\,u\,Z_n}\big]$, d'où $\phi_{Z_n}(u) = \prod_{j=1}^n \mathbb{E}\big[e^{i\,u\,a_j^{(n)}\,U_j}\big]$ car les (U_i) sont indépendantes. Si $\max_{1\leq i\leq n} |a_j^{(n)}| \underset{n\to+\infty}{\longrightarrow} 0$, pour tout $1 \leq j \leq n$, $u\,a_j^{(n)}\,U_j \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} 0$. Développement limité de Taylor en 0, pour tout $u \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\big[e^{i\,u\,a_j^{(n)}\,U_j}\big] &= 1 + i\,u\,\mathbb{E}\big[a_j^{(n)}\,U_j\big] - \frac{u^2}{2}\,\mathbb{E}\big[\big(a_j^{(n)}\,U_j\big)^2\big] + o\big(\mathbb{E}\big[u^2\,\big(a_j^{(n)}\,U_j\big)^2\big]\big)\\ &= 1 + 0 - \frac{u^2}{2}\,\big(a_j^{(n)}\big)^2 + o\big(\big(a_j^{(n)}\big)^2\big)\end{aligned}$$

en utilisant les hypothèses sur les U_j. Comme $\displaystyle\prod_{j=1}^n \mathbb{E}\big[e^{i\,u\,a_j^{(n)}\,U_j}\big] = \exp\Big(\sum_{j=1}^n \log\big(\mathbb{E}\big[e^{i\,u\,a_j^{(n)}\,U_j}\big]\big)\Big)$, et avec $\log(1+x) = x + o(x)$ pour $x \to 0$, on obtient :

$$\phi_{Z_n}(u) = \exp\Big(\sum_{j=1}^n -\frac{u^2}{2}\,\big(a_j^{(n)}\big)^2 + o\big(\big(a_j^{(n)}\big)^2\big)\Big)$$

car $\max_{1\leq i\leq n} |a_j^{(n)}| \underset{n\to+\infty}{\longrightarrow} 0$. Avec la condition $\sum_{j=1}^n \big(a_j^{(n)}\big)^2 = 1$, on en déduit que $\phi_{Z_n}(u) \underset{n\to+\infty}{\longrightarrow} \exp\big(-\frac{1}{2}\,u^2\big)$ qui est la fonction caractéristique d'une $\mathcal{N}(0,1)$. □

Plan du cours

Premiers résultats

Propriété

Sous **(A0)-(A3)**, *et si* $\left({}^tZ\,Z\right)^{-1} \underset{n\to+\infty}{\longrightarrow} 0$, *alors* $\widehat{\theta}^{(n)} \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} \theta^*$

Démonstration.

On a $\mathbb{E}\left[\widehat{\theta}_i^{(n)}\right] = \theta_i^*$ et $\text{var}(\widehat{\theta}_i^{(n)}) = \sigma^{*2}\left[\left({}^tZ\,Z\right)^{-1}\right]_{i+1,i+1} \underset{n\to+\infty}{\longrightarrow} 0$. Donc avec l'inégalité de Bienaymé-Tchebychev, $\widehat{\theta}_i^{(n)} \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} \theta_i^*$ pour tout $i \in \{0,\ldots,p\}$. D'où le résultat puisque p fixé. □

Attention ! L'estimateur $\widehat{\theta}^{(n)}$ ne converge pas toujours !

Exercice : $Z = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}$.

Premiers résultats (2)

Démonstration.

Alors $({}^t Z\, Z)^{-1} = \frac{1}{(n-1)} \begin{pmatrix} 1 & -1 \\ -1 & n \end{pmatrix}$, ne converge pas vers 0. On a

$$\widehat{\theta}^{(n)} = \theta^* + \frac{1}{(n-1)} \begin{pmatrix} 1 & -1 \\ -1 & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n \varepsilon_i \\ \varepsilon_1 \end{pmatrix} = \theta^* + \begin{pmatrix} \frac{1}{n-1} \sum_{i=2}^n \varepsilon_i \\ \varepsilon_1 - \frac{1}{n-1} \sum_{i=2}^n \varepsilon_i \end{pmatrix}$$

Donc $\widehat{\theta}_0^{(n)} \xrightarrow[n\to+\infty]{\mathcal{P}} \theta_0^*$ mais $\widehat{\theta}_1^{(n)}$ ne converge pas. □

Par la suite, nous remplacerons **(A2)** et **(A3)** par :

(A23) : $(\varepsilon_i)_i$ est une suite de v.a.i.i.d. et $\mathbb{E}[\varepsilon_0^2] < \infty$.

Remarque : Sous **(A23)** alors **(A2)** et **(A3)** sont vérifiées.

Premiers résultats (3)

Propriété

Sous **(A0)-(A1)-(A23)**, *alors* $\widehat{\sigma}_n^2 \xrightarrow[n\to+\infty]{\mathcal{P}} \sigma^{*2}$.

Démonstration.

On a $\widehat{\sigma}_n^2 = \frac{1}{n-(p+1)} \left\| P_{[Z]^\perp} \varepsilon \right\|^2$. Comme un projeté d'un vecteur a toujours une norme inférieure à celle de ce vecteur, $U_n = \frac{1}{n-(p+1)} \left(\left\| \varepsilon \right\|^2 - \left\| P_{[Z]^\perp} \varepsilon \right\|^2 \right) \geq 0$. De plus $\mathbb{E}[U_n] = \frac{1}{n-(p+1)} \mathbb{E}\left[\left\| P_{[Z]} \varepsilon \right\|^2 \right] = \sigma^{*2} \frac{p+1}{n-(p+1)} \xrightarrow[n\to+\infty]{} 0$. D'après l'inégalité de Markov, on en déduit que $U_n \xrightarrow[n\to+\infty]{\mathcal{P}} 0$. Enfin, par la loi forte des grands nombres $\frac{1}{n} \left\| \varepsilon \right\|^2 \xrightarrow[n\to+\infty]{\mathcal{P}} \sigma^{*2}$, donc par Slutsky, comme p fixé, $\frac{1}{n-(p+1)} \left\| \varepsilon \right\|^2 \xrightarrow[n\to+\infty]{\mathcal{P}} \sigma^{*2}$. Finalement, $\widehat{\sigma}_n^2 = \frac{1}{n-(p+1)} \left\| \varepsilon \right\|^2 - U_n$ on en déduit donc que $\widehat{\sigma}_n^2 \xrightarrow[n\to+\infty]{\mathcal{P}} \sigma^{*2}$. □

Remarque : Pas de condition sur Z pour la convergence hormis **(A0)** !

Normalité asymptotique

Propriété

Sous **(A0)-(A4)**, *alors* $\left({}^t Z\, Z\right)^{1/2}\left(\widehat{\theta}^{(n)} - \theta^*\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{p+1}\left(0\,,\, \sigma^{*2}\, I_{p+1}\right)$.

Démonstration.

On a montré que $\widehat{\theta}^{(n)} \overset{\mathcal{L}}{\approx} \mathcal{N}_{p+1}\left(\theta^*\,,\, \sigma^{*2}\left({}^t Z\, Z\right)^{-1}\right)$. □

Question : Ce résultat se généralise-t-il si on remplace **(A2)-(A4)** par **(A23)** ?

Normalité asymptotique (2)

Propriété

Sous **(A0)-(A1)-(A23)**, *et si* $\mathbb{E}\left[\varepsilon_1^4\right] = \mu_4^* < \infty$, *alors*

$$\sqrt{n}\left(\widehat{\sigma}_n^2 - \sigma^{*2}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0\,,\,\mu_4^* - \sigma^{*4}\right).$$

Démonstration.

On reprend la preuve précédente. On a par le TLC classique $\sqrt{n}\left(\frac{1}{n}\left\|\varepsilon\right\|^2 - \sigma^{*2}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0\,,\,\mu_4^* - \sigma^{*4}\right)$. Par Slutsky, on en déduit pareillement que $\sqrt{n}\left(\frac{1}{n-(p+1)}\left\|\varepsilon\right\|^2 - \sigma^{*2}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0\,,\,\mu_4^* - \sigma^{*4}\right)$. On écrit ensuite $\widehat{\sigma}_n^2 = \frac{1}{n-(p+1)}\left\|\varepsilon\right\|^2 - U_n$. On a montré que $U_n \geq 0$ et $\mathbb{E}[U_n] = \frac{p+1}{n-(p+1)}\,\sigma^{*2}$ donc $\mathbb{E}[\sqrt{n}\,U_n] \xrightarrow[n\to+\infty]{} 0$. Par suite, par l'Inégalité de Markov, $\sqrt{n}\,U_n \xrightarrow[n\to+\infty]{\mathcal{P}} 0$. D'où le résultat par Slutsky encore. □

Remarque : Si on rajoute **(A4)**, alors $\sqrt{n}\left(\widehat{\sigma}_n^2 - \sigma^{*2}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0\,,\,2\,\sigma^{*4}\right)$

Normalité asymptotique (3)

Définition

Une suite de vecteurs aléatoires $(Z^n)_n$ dont la taille peut dépendre de n est asymptotiquement gaussienne si pour toute suite de combinaisons linéaires ${}^t(C^n)' Z^n$ non nulles, on a

$$U^n = \frac{{}^t(C^n) Z^n - \mathbb{E}\left[{}^t(C^n) Z^n\right]}{\sqrt{var\left({}^t(C^n) Z^n\right)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Proposition

Sous **(A0)-(A1)-(A23)**, *et si* $\max_{1\leq i\leq n} \left|\left(Z \left({}^t Z Z\right)^{-1} {}^t Z\right)_{ii}\right| \xrightarrow[n\to+\infty]{} 0$, *alors* $(\widehat{Y}^{(n)})_{n\in\mathbb{N}} = Z\,\widehat{\theta}^{(n)}$ *est asymptotiquement normal.*

Normalité asymptotique (4)

Démonstration.

On va utiliser la caractérisation de la normalité asymptotique et le théorème de Lindeberg pour cela. Soit $C \in \mathbb{R}^n$. Alors :

$$\frac{{}^tC\,(\widehat{Y}^{(n)} - Z\,\theta^*)}{\sqrt{\mathrm{var}\big({}^tC\,\widehat{Y}^{(n)}\big)}} = \frac{{}^tC\,P_{[Z]}\varepsilon}{\sigma\,\sqrt{{}^tC\,P_{[Z]}\,C}} = {}^tD\,\varepsilon',$$

avec $\varepsilon' = \frac{\varepsilon}{\sigma}$ et $D = \frac{P_{[Z]}\,C}{\sqrt{{}^tC\,P_{[Z]}\,C}} = (D_i)_{1\leq i\leq n}$. On peut appliquer le Théorème de Lindeberg ((ε'_i) est une suite de v.a.i.i.d. telle que $\mathbb{E}[\varepsilon'] = 0$ et $\mathrm{var}(\varepsilon'_i) = 1$ et $\sum_{i=1}^n D_i^2 = 1$ par construction) dès que l'on montre que $\max_{1\leq i\leq n} |D_i| \underset{n\to+\infty}{\longrightarrow} 0$. Pour cela, on peut écrire que $D = P_{[Z]}D$, d'où :

$$D_i^2 = \big((P_{[Z]}D)_i\big)^2 = \Big(\sum_{k=1}^n (P_{[Z]})_{ik} D_k\Big)^2 \leq \sum_{k=1}^n (P_{[Z]})_{ik}^2 \sum_{k=1}^n D_k^2 = \sum_{k=1}^n (P_{[Z]})_{ik}(P_{[Z]})_{ki} = (P_{[Z]})_{ii}$$

par l'inégalité de Cauchy-Schwarz, la symétrie de $P_{[Z]}$ et le fait que $P_{[Z]}P_{[Z]} = P_{[Z]}$. □

Normalité asymptotique (5)

Corollaire

Sous **(A0)-(A1)-(A23)**, *et si* $\max_{1\leq i\leq n} \left|\left(Z\left({}^tZ\,Z\right)^{-1}{}^tZ\right)_{ii}\right| \underset{n\to+\infty}{\longrightarrow} 0$, *alors*

$$\left({}^tZ\,Z\right)^{1/2}\left(\widehat{\theta}^{(n)}-\theta^*\right) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}\left(0\,,\,I_{p+1}\right)$$

Démonstration.

On veut montrer que pour tout $C\in\mathbb{R}^{p+1}$, alors ${}^tC\,\widehat{\theta}^{(n)}$ est asymptotiquement gaussien. Mais $\widehat{\theta}^{(n)}=\left({}^tZ\,Z\right)^{-1}{}^tZ\,\widehat{Y}^{(n)}$. Or pour tout $C'\in\mathbb{R}^n$, alors ${}^tC'\,\widehat{Y}^{(n)}$ est asymptotiquement gaussien. Donc en particulier pour ${}^tC'={}^tC\,\left({}^tZ\,Z\right)^{-1}{}^tZ$. D'où le résultat. □

Conséquence : Dès que le nombre d'individus devient grand (au moins 20 ou 30), l'estimateur $\widehat{\theta}^{(n)}$ a une loi $\simeq$ gaussienne sans l'hypothèse **(A4)**.

Comportement asymptotique des tests

Propriété

On a $\frac{1}{n}\chi^2(n) \xrightarrow[n\to+\infty]{\mathcal{P}} 1$, *d'où* $t(n) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$ *et* $F(p_0,n) \xrightarrow[n\to\infty]{\mathcal{L}} \frac{1}{p_0}\chi^2(p_0)$

Démonstration.

On a $\frac{1}{n}\chi^2(n) \overset{\mathcal{L}}{\sim} \frac{1}{n}\sum_{i=1}^n Z_i^2$ où $Z=(Z_i) \overset{\mathcal{L}}{\sim} \mathcal{N}(0,I_n)$. Avec la loi forte des grands nombres, $\frac{1}{n}\sum_{i=1}^n Z_i^2 \xrightarrow[n\to+\infty]{\mathcal{P}} \mathbb{E}[Z_0^2]=1$. Ensuite, on utilise $t(n) \overset{\mathcal{L}}{\sim} \frac{\mathcal{N}(0,1)}{\sqrt{\frac{1}{n}\chi^2(n)}}$ et $F(p_0,n) \overset{\mathcal{L}}{\sim} \frac{\frac{1}{p_0}\chi^2(p_0)}{\frac{1}{n}\chi^2(n)}$ □

Corollaire

Sous **(A0)-(A1)-(A23)**, *et si* $\max_{1\le i\le n}\left|\left(Z\left({}^tZ\,Z\right)^{-1}{}^tZ\right)_{ii}\right| \xrightarrow[n\to+\infty]{} 0$, *alors sous les hypothèses* H_0 *respectives, les statistiques de test de Student* $\widehat{T}^{(n)}$ *et Fisher* $\widehat{F}^{(n)}$ *tendent respectivement vers des lois* $\mathcal{N}(0,1)$ *et* $\frac{1}{q}\chi^2(q)$.

Comportement asymptotique des tests (2)

Démonstration.

Pour $\widehat{T}^{(n)} = \dfrac{1}{\sqrt{\widehat{\sigma^2}}} \left({}^t C \left({}^t Z\, Z\right)^{-1} C\right)^{-1/2} {}^t C\, \widehat{\theta}^{(n)}$, comme $\widehat{\theta}^{(n)}$ est asymptotiquement gaussien ${}^t C\, \widehat{\theta}^{(n)}$ l'est également. Et $\widehat{\sigma^2} \xrightarrow[n\to+\infty]{\mathcal{P}} \sigma^{*2}$. D'où le résultat.

On a $\widehat{F}^{(n)} = \dfrac{\frac{1}{q}\, {}^t\widehat{\theta}^{(n)}\, C \left({}^t C \left({}^t Z\, Z\right)^{-1} C\right)^{-1} {}^t C\, \widehat{\theta}^{(n)}}{\widehat{\sigma^2}}$. Comme $\widehat{\theta}^{(n)}$ est asymptotiquement gaussien la loi asymptotique de $\frac{1}{q}\, {}^t\widehat{\theta}^{(n)}\, C \left({}^t C \left({}^t Z\, Z\right)^{-1} C\right)^{-1} {}^t C\, \widehat{\theta}^{(n)}$ est la même que celle obtenue dans le cas gaussien, donc $\frac{\sigma^2}{q}\, \chi^2(q)$. Et comme précédemment $\widehat{\sigma^2} \xrightarrow[n\to+\infty]{\mathcal{P}} \sigma^{*2}$. D'où le résultat.

□

Conséquence : Concrètement, dès que le nombre d'individus devient grand (au moins 20 ou 30), très souvent les tests du modèle linéaire peuvent être utilisés sans avoir l'hypothèse gaussienne **(A4)**

Plan du cours

Le problème de la sélection d'un modèle linéaire

Cadre : On a observé $Y_1, \ldots, Y_n$, que l'on veut modéliser par des variables exogènes $Z^{(1)}, \ldots, Z^{(p)}$ prenant les valeurs observées $(Z_i^{(j)})_{1 \leq i \leq n}$

Questions :

- Les variables exogènes interviennent-elles toutes dans un modèle linéaire ?
- Comment choisir parmi elles et par rapport à quel critère ?

Une première réponse ... insatisfaisante

Première réponse : Utilisation de tests de Student ou de Fisher !

$\Longrightarrow$ On teste si la variable $Z^{(p_0)}$ est significative

$\Longrightarrow$ Mais comment tester si plusieurs variables sont significatives ?

- Tests de Student successifs. Comment et pour quel risque ?
- Tests de Fisher. Mais comment choisir les tests ? Tout tester ? Quel risque choisir ?

Nouvelle formalisation

Notation : Pour $Z^{(1)}, \ldots, Z^{(p)}$ les p variables exogènes, on note m un modèle composé de certaines variables parmi elles.

$\Longrightarrow m \in \mathcal{P}(\{1, \ldots, p\})$ et on notera $|m| = \text{Card}(m)$

Exemple : $p = 5$ et $m = \{2, 3, 5\}$, donc $Z^{(2)}$, $Z^{(3)}$ et $Z^{(5)}$ compose m

Notation : On notera $\mathcal{M}$ une famille de modèle. Deux cas :

- Le plus souvent, la famille exhaustive : $\mathcal{M} = \mathcal{P}(\{1, \ldots, p\})$
- Parfois la famille hiérarchique : $\mathcal{M} = \big\{\emptyset, \{1\}, \{1, 2\}, \ldots, \{1, 2, \ldots, p\}\big\}$

  $\Longrightarrow$ Régression polynomiale par exemple

Nouvelle formalisation (2)

On dispose donc de $|m|$ variables potentiellement explicatives. On considérera la matrice de taille $(n, |m|+1)$:

$$Z_{(m)} = \begin{pmatrix} 1 & Z_1^{(i_1)} & Z_1^{(i_2)} & \dots & Z_1^{(i_{|m|})} \\ 1 & Z_2^{(i_1)} & Z_2^{(i_2)} & \dots & Z_2^{(i_{|m|})} \\ \vdots & \vdots & & \vdots & \\ 1 & Z_n^{(i_1)} & Z_n^{(i_n)} & \dots & Z_n^{(i_{|m|})} \end{pmatrix} \quad \text{lorsque} \quad m = \{i_1, \cdots, i_{|m|}\}.$$

On supposera que pour tout $m \in \mathcal{M}$ le rang de la matrice $Z_{(m)}$ est $|m|+1$.

Nouvelle formalisation (4)

Hypothèses sur le vrai modèle : Il existe $m^* \in \mathcal{M}$, inconnu, tel que :

$$Y = \mu^* + \varepsilon^* = Z_{(m^*)}\,\theta_{(m^*)} + \varepsilon^*,$$

sous **(A0)-(A3)**, $\theta_{(m^*)}$ vecteur de taille $|m^*| + 1$ ne contenant pas de 0

Modèles d'analyse : On utilise la famille de modèles :

$$Y = \mu + \varepsilon = Z_{(m)}\,\theta_{(m)} + \varepsilon$$

avec $m \in \mathcal{M}$, $\mu \in \mathbb{R}^n$, $\theta_{(m)} \in \mathbb{R}^{|m|+1}$. Par la suite, on appellera aussi modèle m le modèle d'analyse.

Définitions

Définition

Avec le modèle d'analyse est $m \in \mathcal{M}$:

- *si $m = m_p = \{1, \cdots, p\}$, le modèle est complet,toutes les variables explicatives disponibles sont considérées.*
- *si $m^* \subset m$ avec $m \neq m^*$, le modèle est sur-ajusté (overfitting).*
- *si $m \subset m^*$ avec $m \neq m^*$, le modèle est faux (misspecified).*

Rappels : pour le modèle d'analyse $m \in \mathcal{M}$

1. $\widehat{\theta}_{(m)} = (({}^t Z_{(m)})\, Z_{(m)})^{-1} (Z_{(m)})'\, Y$ d'où $\widehat{Y}_{(m)} = Z_{(m)}\, \widehat{\theta}_{(m)}$.
2. Deux estimateurs de σ^2 :

$$\widehat{\sigma}^2_{(m)} = \frac{1}{(n - |m| - 1)} \left\| Y - \widehat{Y}_{(m)} \right\|^2 \quad \text{et} \quad \tilde{\sigma}^2_{(m)} = \frac{1}{n} \left\| Y - \widehat{Y}_{(m)} \right\|^2.$$

CP de Mallows : minimisation du risque quadratique

On veut déterminer $m \in \mathcal{M}$ qui minimise le **risque quadratique :**

$$\begin{aligned} R(m, m^*) &= \mathbb{E}\left[\|\widehat{Y}_{(m)} - \mu^*\|^2\right] = \mathbb{E}\left[\|Z_{(m)}\,\widehat{\theta}_{(m)} - \mu^*\|^2\right] \\ &= \mathbb{E}\left[\|Z_{(m)}\,\widehat{\theta}_{(m)} - \mu^*_{(m)}\|^2\right] + \mathbb{E}\left[\|\mu^* - \mu^*_{(m)}\|^2\right] \quad \text{(Pythagore)} \end{aligned}$$

avec $\mu^*_{(m)} = P_{[Z_{(m)}]}\mu^*$. On a $\mathbb{E}\left[\widehat{Y}_{(m)}\right] = \mathbb{E}\left[P_{[Z_{(m)}]}\,Y\right] = \mu^*_{(m)}$ et

$$\begin{aligned} \mathbb{E}\left[\|\widehat{Y}_{(m)} - \mu^*\|^2\right] &= \mathbb{E}\left[\|P_{[Z_{(m)}]}\varepsilon^*\|^2\right] + \mathbb{E}\left[\|\mu^* - \mu^*_{(m)}\|^2\right] \\ &= \mathbb{E}\left[\mathrm{Tr}\left({}^t\varepsilon^* P_{[Z_{(m)}]}\varepsilon^*\right)\right] + \mathbb{E}\left[\|\mu^* - \mu^*_{(m)}\|^2\right] \\ &= \sigma_*^2\,\mathrm{Tr}\left(P_{[Z_{(m)}]}\right) + \|\mu^* - \mu^*_{(m)}\|^2 \\ &= (|m|+1)\,\sigma_*^2 + \|\mu^* - \mu^*_{(m)}\|^2. \end{aligned}$$

CP de Mallows : minimisation du risque quadratique (2)

$$\begin{aligned}
\mathbb{E}\left[\|Y-\widehat{Y}_{(m)}\|^2\right] &= \mathbb{E}\left[\|Y-\mu^*_{(m)}\|^2\right] - \mathbb{E}\left[\|\widehat{Y}_{(m)}-\mu^*_{(m)}\|^2\right] \quad \text{(Pythagore)} \\
&= \mathbb{E}\left[\|Y-\mu^*+\mu^*-\mu^*_{(m)}\|^2\right] - \mathbb{E}\left[\|\widehat{Y}_{(m)}-\mu^*_{(m)}\|^2\right] \\
&= \mathbb{E}\left[\|Y-\mu^*\|^2\right] + \|\mu^*-\mu^*_{(m)}\|^2 - (|m|+1)\,\sigma^2_* \\
&= (n-(|m|+1))\,\sigma^2_* + \|\mu^*-\mu^*_{(m)}\|^2
\end{aligned}$$

$$\begin{aligned}
\text{D'où} \qquad R(m,m^*) &= (|m|+1)\,\sigma^2_* + \mathbb{E}\left[\|Y-\widehat{Y}_{(m)}\|^2\right] - (n-(|m|+1)) \cdot \\
\Longrightarrow \quad \frac{R(m,m^*)}{n\,\sigma^2_*} &= \frac{2(|m|+1)}{n} - 1 + \frac{1}{n\,\sigma^2_*}\,\mathbb{E}\left[\|Y-\widehat{Y}_{(m)}\|^2\right] \\
\Longrightarrow \quad \frac{R(m,m^*)}{n\,\sigma^2_*} &\simeq \frac{2(|m|+1)}{n} - 1 + \frac{\tilde{\sigma}^2_{(m)}}{\tilde{\sigma}^2_{(m_p)}}
\end{aligned}$$

$$\text{en estimant} \begin{cases} \frac{1}{n}\,\mathbb{E}\left[\|Y-\widehat{Y}_{(m)}\|^2\right] \\ \sigma^2_* \end{cases} \quad \text{par} \begin{cases} \tilde{\sigma}^2_{(m)} = \frac{1}{n}\,\|Y-\widehat{Y}_{(m)}\|^2 \\ \tilde{\sigma}^2_{(m_p)} \end{cases}$$

CP de Mallows : minimisation du risque quadratique (3)

Définition

Le Cp **de Mallows** *valant pour* $m \in \mathcal{M}$ *:*

$$\widehat{Cp}(m) = \frac{\tilde{\sigma}^2_{(m)}}{\tilde{\sigma}^2_{(m_p)}} + 2\,\frac{|m|}{n}.$$

On sélectionnera donc un modèle $\widehat{m}$ tel que :

$$\widehat{m}_{CP} = \text{Arg}\min_{m \in \mathcal{M}} \{\widehat{Cp}(m)\}.$$

Ce critère a été introduit par Mallows en 1967.

Critère R^2 ajusté : autre minimisation de risque quadratique

Rappel : Avec le critère R^2, on a $R^2(m) = 1 - \dfrac{\|Y - \widehat{Y}_{(m)}\|^2}{\|Y - \overline{Y}\|^2}$.

Remarque : $\|Y - \widehat{Y}_{(m)}\|^2 = \|P_{[X_{(m)}]^\perp} Y\|^2$ décroît pour une suite emboîtée croissante de modèle : maximiser le R^2 conduit à choisir m_p. Entre modèles de même cardinal $|m|$, R^2 peut être utilisé pour choisir entre eux

Idée du R^2_{Aju} : Dans le R^2, diviser le numérateur par $n - |m| - 1 \Longrightarrow$ espérance du numérateur est σ^2, ne dépend plus de m. D'où :

$$\widehat{R^2_{Aju}}(m) = 1 - \frac{\frac{1}{n-|m|-1}\|Y - \widehat{Y}_{(m)}\|^2}{\frac{1}{n-1}\|Y - \overline{Y}\|^2}.$$

On **maximise** $\widehat{R^2_{Aju}}(m)$ et $\widehat{m}_{R^2_{Aju}} = \text{Argmax}_{m\in\mathcal{M}}\{\widehat{R^2_{Aju}}(m)\}$.

Remarque : Maximiser le $\widehat{R^2_{Aju}}(m)$ revient à minimiser $\frac{1}{n-|m|-1}\|Y - \widehat{Y}_m\|^2$.

Critère *AIC* : minimisation la dissemblance de Kullback

Autre écart entre mesures de proba : Si la même mesure domine m et m^*

$$d(m, m^*) = \int_{\mathbb{R}^n} f_m(x) \log\left(\frac{f_m(x)}{f_{m*}(x)}\right) dx \quad (f \text{ densité de Y})$$

$\Longrightarrow$ Minimiser la dissemblance de Kullback revient à celle du critère *AIC* (pour Akaiké Information Criterion, 1973), tel que

$$\begin{aligned}\widehat{AIC}(m) &= -2 \log L(Y \mid \widehat{\gamma}_{(m)}) + 2\big(|m| + 2\big) \\ &= -2 \times \log(\text{Vraisemblance maximisée}) + 2 \times \text{Nombre de paramètres.}\end{aligned}$$

Suivant ce critère, on choisira m tel que $\widehat{m}_{AIC} = \text{Argmin}_{m \in \mathcal{M}}\{\widehat{AIC}(m)\}$.

Remarque : Dans le cas gaussien, le critère *AIC* pourra s'écrire :

$$\widehat{AIC}(m) = n \log\left(\tilde{\sigma}^2_{(m)}\right) + 2\left(|m| + 2\right),$$

puisqu'alors $\log(\text{Vraisemblance maximisée}) = -n \log \tilde{\sigma}_{(m)} - \frac{n}{2} \log 2\pi - \frac{n}{2}$.

Le critère *BIC* comme maximisation d'une proba

Idée : Pour chaque m, on estime asymptotiquement $\mathbb{P}\big(Y \mid m\big)$

Après calcul, pour n grand

$$-2\,\log\big(\mathbb{P}\big(Y \mid m\big)\big) \simeq -2\,\log L(Y \mid \widehat{\gamma}_{(m)}) + \log(n)\,\big(|m|+1\big)$$

Le critère *BIC* (Bayesian Information Criterium) introduit en 1978 par Schwarz :

$$\begin{aligned}\widehat{BIC}(m) &= -2 \times \log(\text{Vraisemblance maximisée}) + \log(n) \times \text{Nombre de para} \\ &= n\,\log\big(\tilde{\sigma}^2_{(m)}\big) + \log(n)\,\big(|m|+1\big).\end{aligned}$$

Suivant ce critère, on choisira m tel que $\widehat{m}_{BIC} = \text{Argmin}_{m\in\mathcal{M}}\{\widehat{BIC}(m)\}$.

Les différents critères

Conclusion : Dans le cadre du modèle linéaire, on dispose de 3 critères à minimiser et un critère à maximiser $\widehat{R^2_{Aju}}$:

- $\widehat{Cp}(m) = \dfrac{\|Y - \widehat{Y}_m\|^2}{\|Y - \widehat{Y}_{(m_p)}\|^2} + 2\,\dfrac{(|m|+1)}{n}$
- $\widehat{R^2_{Aju}}(m) = 1 - \dfrac{n-1}{n-|m|-1}\,\dfrac{\|Y - \widehat{Y}_{(m)}\|^2}{\|Y - \overline{Y}\|^2}$
- $\widehat{AIC}(m) = n \log\left(\|Y - \widehat{Y}_{(m)}\|^2\right) + 2(|m|+1)$
- $\widehat{BIC}(m) = n \log\left(\|Y - \widehat{Y}_{(m)}\|^2\right) + \log n\,(|m|+1).$

Remarque : On a $\|Y - \widehat{Y}_{(m)}\|^2 = n\,\widetilde{\sigma}^2_{(m)} = (n - |m| - 1)\,\widehat{\sigma}^2_{(m)}$

Plan du cours

Choix entre 2 modèles

Soit $m_1, m_2 \in \mathcal{M}$. On cherche $\mathbb{P}\big(\widehat{\text{Crit}}(m_1) \leq \widehat{\text{Crit}}(m_2)\big)$, probabilité de m_1 plutôt que m_2 pour $\widehat{Cp}$, $\widehat{AIC}$ ou $\widehat{BIC}$, l'inverse pour $\widehat{R^2_{Aju}}$.

Remarque : Si $|m_1| = |m_2| \Longrightarrow$ On préfère le modèle le mieux ajusté, celui avec somme des carrés des résidus la plus faible, ce que choisit le R^2...

On fixe désormais $|m_1| < |m_2|$, et on définit la statistique de type Fisher suivante :

$$\widehat{F}(m_1, m_2) := \left(\frac{n - |m_2| - 1}{|m_2| - |m_1|}\right) \frac{\|Y - \widehat{Y}_{(m_1)}\|^2 - \|Y - \widehat{Y}_{(m_2)}\|^2}{\|Y - \widehat{Y}_{(m_2)}\|^2}.$$

Choix entre 2 modèles (2)

On a :

- $$\mathbb{P}\left[\widehat{\mathrm{Cp}}(m_1) \leq \widehat{\mathrm{Cp}}(m_2)\right] = \mathbb{P}\Big[\|Y - \widehat{Y}_{(m_1)}\|^2 - \|Y - \widehat{Y}_{(m_2)}\|^2 \leq 2\,\frac{|m_2| - |m_1|}{n}\,\|Y - \widehat{Y}_{(m_p)}\|^2\Big]$$
$$= \mathbb{P}\left[\widehat{F}(m_1, m_2) \leq 2\,\frac{n - |m_2| - 1}{n}\,\frac{\|Y - \widehat{Y}_{(m_p)}\|^2}{\|Y - \widehat{Y}_{(m_2)}\|^2}\right]$$
- $$\mathbb{P}\left[\widehat{R^2_{Aju}}(m_1) \leq \widehat{R^2_{Aju}}(m_2)\right] = \mathbb{P}\left[\widehat{F}(m_1, m_2) \geq 1\right]$$
- $$\mathbb{P}\left[\widehat{\mathrm{AIC}}(m_1) \leq \widehat{\mathrm{AIC}}(m_2)\right]$$
$$= \mathbb{P}\left[\widehat{F}(m_1, m_2) \leq \left(\frac{n - |m_2| - 1}{|m_2| - |m_1|}\right)\left(\exp\left(\frac{2(|m_2| - |m_1|)}{n}\right) - 1\right)\right]$$
- $$\mathbb{P}\left[\widehat{\mathrm{BIC}}(m_1) \leq \widehat{\mathrm{BIC}}(m_2)\right]$$
$$= \mathbb{P}\left[\widehat{F}(m_1, m_2) \leq \left(\frac{n - |m_2| - 1}{|m_2| - |m_1|}\right)\left(\exp\left((|m_2| - |m_1|)\,\frac{\log n}{n}\right) - 1\right)\right].$$

Probabilité de préférer un sur-modèle

Proposition

Sous les conditions de normalité asymptotique, si m_2 contient strictement m^*,

$$\mathbb{P}\left[\widehat{Cp}(m^*) \geq \widehat{Cp}(m_2)\right] \underset{n\to+\infty}{\longrightarrow} \mathbb{P}\left[\chi^2(|m_2| - |m^*|) \geq 2(|m_2| - |m^*|)\right]$$

$$\mathbb{P}\left[\widehat{R^2_{Aju}}(m^*) \leq \widehat{R^2_{Aju}}(m_2)\right] \underset{n\to+\infty}{\longrightarrow} \mathbb{P}\left[\chi^2(|m_2| - |m^*|) \geq (|m_2| - |m^*|)\right]$$

$$\mathbb{P}\left[\widehat{AIC}(m^*) \geq \widehat{AIC}(m_2)\right] \underset{n\to+\infty}{\longrightarrow} \mathbb{P}\left[\chi^2(|m_2| - |m^*|) \geq 2(|m_2| - |m^*|)\right]$$

$$\mathbb{P}\left[\widehat{BIC}(m^*) \geq \widehat{BIC}(m_2)\right] \underset{n\to+\infty}{\longrightarrow} 0.$$

On remarque que **le critère R^2_{Aju} a plus tendance asymptotiquement à sur-ajuster que les critères Cp ou AIC.**

Probabilité de préférer un sur-modèle (2)

Démonstration.

Sous des hypothèses, $\widehat{F}(m^*, m_2) \xrightarrow[n\to\infty]{\mathcal{L}} \frac{1}{|m_2|-|m^*|}\chi^2(|m_2|-|m^*|)$. De plus $\frac{\|Y-\widehat{Y}_{(m_p)}\|^2}{\|Y-\widehat{Y}_{(m_2)}\|^2} \xrightarrow[n\to\infty]{\mathcal{L}} 1$. Avec l'aide de développements limités on obtient bien les résultats de la Proposition à partir des probabilités en fonction de $\widehat{F}(m^*, m_2)$ obtenues pour les 3 premiers critères.
De plus pour tout $C > 0$, il existe n_0 tel que pour tout $n \geq n_0$, $\left(\frac{n-|m_2|-1}{|m_2|-|m^*|}\right)\left(\exp\left((|m_2|-|m^*|)\frac{\log n}{n}\right)-1\right) \simeq \log n > C$. Ainsi pour $n \geq n_0$, en utilisant la probabilité obtenue pour le critère BIC, on a :

$$\mathbb{P}\left[\widehat{F}(m^*, m_2) \geq \left(\frac{n-|m_2|-1}{|m_2|-|m^*|}\right)\left(\exp\left((|m_2|-|m^*|)\frac{\log n}{n}\right)-1\right)\right] \leq \mathbb{P}\left[\widehat{F}(m^*, m_2) \geq C\right].$$

Comme $\widehat{F}(m^*, m_2) \xrightarrow[n\to\infty]{\mathcal{L}} \frac{\chi^2(|m_2|-|m^*|)}{|m_2|-|m^*|}$, on a

$$\mathbb{P}\left[\mathrm{BIC}(m^*) \geq \mathrm{BIC}(m_2)\right] \leq \mathbb{P}\left[\widehat{F}(m^*, m_2) \geq C\right] \xrightarrow[n\to+\infty]{} \mathbb{P}\left[\frac{\chi^2(|m_2|-|m^*|)}{|m_2|-|m^*|} \geq C\right] \leq \frac{1}{C},$$

d'après l'Inégalité de Markov. Mais comme ceci est vrai pour tout $C > 0$, on obtient bien le résultat. □

Probabilité de préférer un sur-modèle (3)

Conséquence : Sous les mêmes hypothèses :

$$\mathbb{P}\left(m^* \subset \widehat{m}_{BIC},\ m^* \neq \widehat{m}_{BIC}\right) \underset{n\to+\infty}{\longrightarrow} 0$$

Démonstration.

On a grâce aux propriétés précédentes, puisque $\left|\mathcal{M}\right| \leq 2^p$ ne dépend pas de n,

$$\begin{aligned}\mathbb{P}\left(m^* \subset \widehat{m}_{BIC},\ m^* \neq \widehat{m}_{BIC}\right) &\leq \mathbb{P}\left(\exists m \in \mathcal{M},\ m^* \subset m,\ m \neq m^*,\ \widehat{\mathrm{BIC}}(m) < \widehat{\mathrm{BIC}}(m^*)\right) \\ &\leq \sum_{m^* \subset m,\ m \neq m^*} \mathbb{P}\left(\widehat{\mathrm{BIC}}(m) < \widehat{\mathrm{BIC}}(m^*)\right) \underset{n\to+\infty}{\longrightarrow} 0.\end{aligned}$$

□

Conclusion : le critère BIC évite asymptotiquement de sur-ajuster

Remarque : Résultat également valable si $\log n$ du BIC est remplacé par c_n, avec $c_n \underset{n\to+\infty}{\longrightarrow} \infty$ et $c_n = o(n)$

$$\Longrightarrow \text{ critère GIC : } \quad \widehat{\mathrm{GIC}}(m) = n \log\left(\|Y - \widehat{Y}_{(m)}\|^2\right) + c_n\,|m|.$$

Probabilité de préférer un faux-modèle

Proposition

Sous les mêmes hypothèses et avec $\left(Z_{(m_p)}\right)_n$ *vérifiant* $\frac{1}{d(n)}\,{}^t Z_{(m_p)}\, Z_{(m_p)} \underset{n\to+\infty}{\longrightarrow} M$, *où* $d(n)\log^{-1}(n) \underset{n\to+\infty}{\longrightarrow} \infty$ *et* M *matrice définie positive. Alors pour Crit=Cp,* R^2_{Aju}, *AIC ou BIC,*

$$\mathbb{P}\left[\text{"Crit choisit un faux-modèle"}\right] \underset{n\to+\infty}{\longrightarrow} 0.$$

Démonstration.

Preuve trop complexe, voir le livre... □

Probabilité de préférer un faux-modèle (2)

Conséquences : Sous les mêmes hypothèses :

1. $\inf_{n\in\mathbb{N}} \mathbb{P}\left(m^* \subset \widehat{m}_{Crit},\ m^* \neq \widehat{m}_{Crit}\right) \geq p_{Crit} > 0$ pour $Crit = \begin{cases} Cp \\ R^2_{Aju} \\ AIC \end{cases}$
2. $\mathbb{P}\left(\widehat{m}_{BIC} = m^*\right) \underset{n\to+\infty}{\longrightarrow} 1$

Conclusion :

1. Le critère BIC (ou GIC) est **asymptotiquement consistant**
2. Les **critères Cp, R^2_{Aju} AIC ont une probabilité positive de sur-ajuster asymptotiquement**.
3. R^2_{Aju} est le moins intéressant et nous ne recommandons pas son utilisation.

Etude asymptotique du risque quadratique

Le risque quadratique du modèle sélectionné est :

$$R_n^{Crit} = \mathbb{E}_{m^*}\left(\|\mu^* - \widehat{Y}_{(\widehat{m}_{Crit})}\|^2\right).$$

Il est clair que :

$$R_n^{Crit} = \sum_{m \in \mathcal{M}} R_n(m) \quad \text{et} \quad R_n(m) = \mathbb{E}_{m^*}\left(\|\mu^* - \widehat{Y}_{(m)}\|^2 \mathbb{I}_{\{\widehat{m}_{Crit}=m\}}\right).$$

Propriété

Sous les hypothèses précédentes, lorsque le critère utilisé est GIC avec $c_n = o(n)$, *et si* $\mathbb{E}[\varepsilon_0^4] < \infty$ *alors* $R_n^{GIC} \underset{n\to+\infty}{\longrightarrow} R_n(m^*) = (|m^*| + 1)\,\sigma_*^2$.

Interprétation : En utilisant le modèle choisi par GIC, le risque quadratique est asymptotiquement le même que si on connaissait a priori le vrai modèle.

Etude asymptotique du risque quadratique (2)

Démonstration.

On écrit $R_n(m) = \mathbb{E}_{m^*}\left(\|\mu^* - \widehat{Y}_{(m)}\|^2 \mathbb{I}_{\{\widehat{m}=m\}}\right)$, d'où avec $p_n(m) = \mathbb{P}(\widehat{m} = m)$,

$$\begin{aligned} R_n(m) &= \mathbb{E}_{m^*}\left(\|\mu^* - \mu^*_{(m)}\|^2 \mathbb{I}_{\{\widehat{m}=m\}} + \|\mu^*_{(m)} - \widehat{Y}_{(m)}\|^2 \mathbb{I}_{\{\widehat{m}=m\}}\right) \\ &= \|P_{[X^{(m)}]^\perp} X^{(m_p)}\theta^{(m_p)}\|^2 \, p_n(m) + \mathbb{E}_{m^*}\left(\|P_{[X^{(m)}]}\varepsilon\|^2 \mathbb{I}_{\{\widehat{m}=m\}}\right) \\ &= J_1(m) + J_2(m). \end{aligned}$$

Concernant $J_1(m)$, si m est un sur-modèle ou m^* alors $J_1(m) = 0$ car $P_{[X^{(m)}]^\perp} X^{(m_p)}\theta^{(m_p)} = 0$.
Mais pour un faux-modèle m, pour n suffisamment grand, avec $t(n, m_1) = P_{[X^{(m)}]^\perp} X^{(m^*)}\theta^{(m^*)} \simeq C(m)\, d(n)$ lorsque $n \to \infty$ avec $C(m) > 0$ ne dépendant pas de n,

$$J_1(m) = \|t(n,m)\|^2 \, p_n(m) \leq 2\, C(m)\, d_n \exp\left(-\frac{1}{8}\,\frac{C(m)}{|m| - |m^*|}\, d_n\right),$$

donc $J_1(m) \underset{n\to+\infty}{\longrightarrow} 0$ car on a supposé $\log n = o(d_n)$. □

Etude asymptotique du risque quadratique (3)

Démonstration.

Si $m = m^*$, alors $J_2(m) \underset{n\to+\infty}{\longrightarrow} (|m^*|+1)\,\sigma_*^2$ car $p_n(m^*) \underset{n\to+\infty}{\longrightarrow} 1$. Si $m \neq m^*$, on majore $J_2(m)$ avec Cauchy-Schwarz, puisque :

$$J_2 \leq \left[\mathbb{E}_{m^*}\left((\|P_{[X^{(m)}]}\varepsilon\|^2)^2\right)\right]^{1/2} [p_n(m)]^{1/2},$$

d'après la définition de $p_n(m)$. Or on montre que

$$\mathbb{E}_{m^*}\left((\|P_{[X^{(m)}]}\varepsilon\|^2)^2\right) = \mathbb{E}_{m^*}\left(\sum_{i,j,i',j'=1}^{n} \pi_{ij}\pi_{i'j'}\varepsilon_i\varepsilon_j\varepsilon_{i'}\varepsilon_{j'}\right).$$

En dénombrant les cas où deux ou 4 indices sont égaux, on montre que $\mathbb{E}_{m^*}\left((\|P_{[X^{(m)}]}\varepsilon\|^2)^2\right)$ est bornée. Comme pour $m \neq m^*$, $p_n(m) \underset{n\to+\infty}{\longrightarrow} 0$, on en déduit que $R_n(m) \underset{n\to+\infty}{\longrightarrow} 0$ tandis que $R_n(m) \underset{n\to+\infty}{\longrightarrow} (|m^*|+1)\,\sigma_*^2$. D'où le résultat final. □

Plan du cours

Les hypothèses "classiques"

Hypothèses pour le modèle linéaire : $Y = Z\,\theta^* + \varepsilon$

1. **(A1)** $\mathbb{E}(\varepsilon_i) = 0$ pour tout $i = 1, \ldots, n$;
2. **(A2) Homoscédasticité :** $\mathrm{var}(\varepsilon_i) = \sigma^2$ pour tout $i = 1, \ldots, n$;
3. **(A3) Non corrélation :** pour tout $i \neq j$, $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$;
4. **(A4) Gaussianité :** ε vecteur gaussien ;

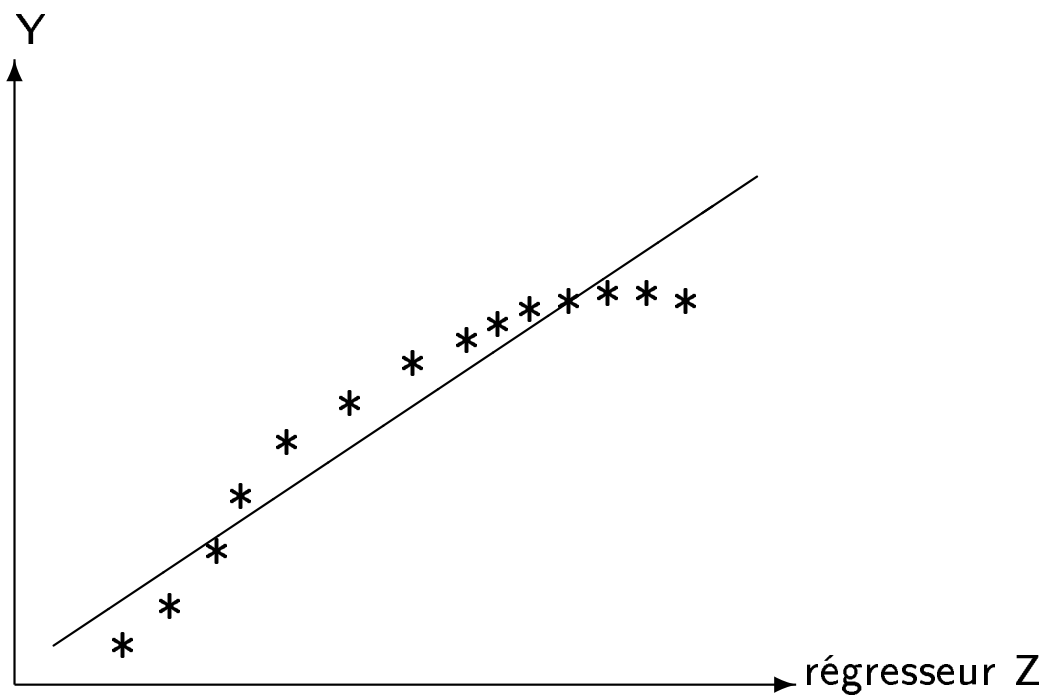
Même si les tests de Student et Fisher sont positifs, R^2 proche de 1...

Diagnostics graphiques $\Longrightarrow$ des conditions ne sont pas vérifiées

Remise en question de **(A1)**

- En régression linéaire simple : Nuage de points et droite de régression donnent une information quasi exhaustive.

Exemple :



$\Longrightarrow$ On voit une courbure de la "vraie" courbe de régression de Y, le modèle n'est pas adéquat $\Longrightarrow$ **(A1)** n'est pas vérifiée.

Remise en question de **(A1)**

- En régression linéaire multiple, impossible d'utiliser le nuage de points car il y a plusieurs régresseurs.

  $\Longrightarrow$ On travaille avec les $\widehat{\varepsilon}_i = Y_i - \widehat{\theta}_0 - \widehat{\theta}_1 Z_i^{(1)} - \cdots - \widehat{\theta}_p Z_i^{(p)}$

  Rappels : $\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i = 0$ et $\widehat{\varepsilon} = P_{[Z]^\perp}\varepsilon = (I_n - (p_{ij}))\,\varepsilon$

  ainsi que $\mathbb{E}[\widehat{\varepsilon}_i] = 0$, $\mathrm{var}(\widehat{\varepsilon}_i) = \sigma^2\,(1 - p_{ii})$ et $\mathrm{cov}(\widehat{\varepsilon}_i, \widehat{\varepsilon}_j) = -\sigma^2\,p_{ij}$

  $\Longrightarrow$ $\mathrm{cov}(\widehat{Y}_i\,,\,\widehat{\varepsilon}_i) = 0$ pour tout $i = 1, \ldots, n$ sous **(A1)**.

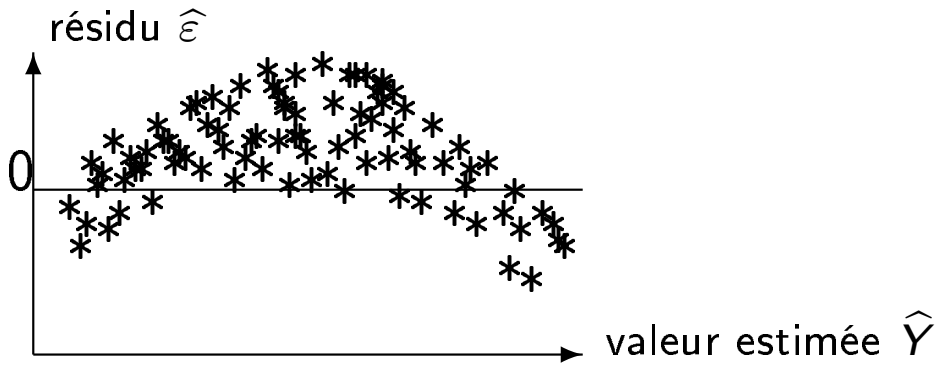
  $\Longrightarrow$ **Graphique des résidus** $(\widehat{\varepsilon}_i)_i$ en fonction des **valeurs prédites** $(\widehat{Y}_i)_i$.

  Le nuage de points doit être "équilibré" autour de l'axe des abscisses

Remise en question de **(A1)**

Concrètement, si on ne voit rien de notable sur le graphique (nuage de points centré et aligné quelconque), c'est très bon signe !

Contre-exemple :



$\Longrightarrow$ Modèle inadapté aux données, les $\widehat{\varepsilon}_i$ dépendent des $\widehat{Y}_i$.

Solutions possibles

1. On n'a pas considéré certaines variables explicatives et on les rajoute !

2. On transforme les régresseurs $Z^{(1)}, \cdots, Z^{(p)}$ par des fonctions, pourvu que le nouveau modèle reste interprétable. Par exemple en utilisant une :

   Transformation de Box-Cox

   $$\tau(x, \lambda) := \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{pour } \lambda \neq 0 \\ \log x & \text{pour } \lambda = 0 \end{cases}$$

   où λ est un réel a priori inconnu.

   Idée : on transforme les variables explicatives quantitatives par Box-Cox en cherchant le paramètre λ qui maximise le R^2.

3. On décompose les variables explicatives sur des classes (donc on obtient des v.a. qualitatives) pour voir si l'évolution est linéaire ou autre

Cas particulier du changement de structure

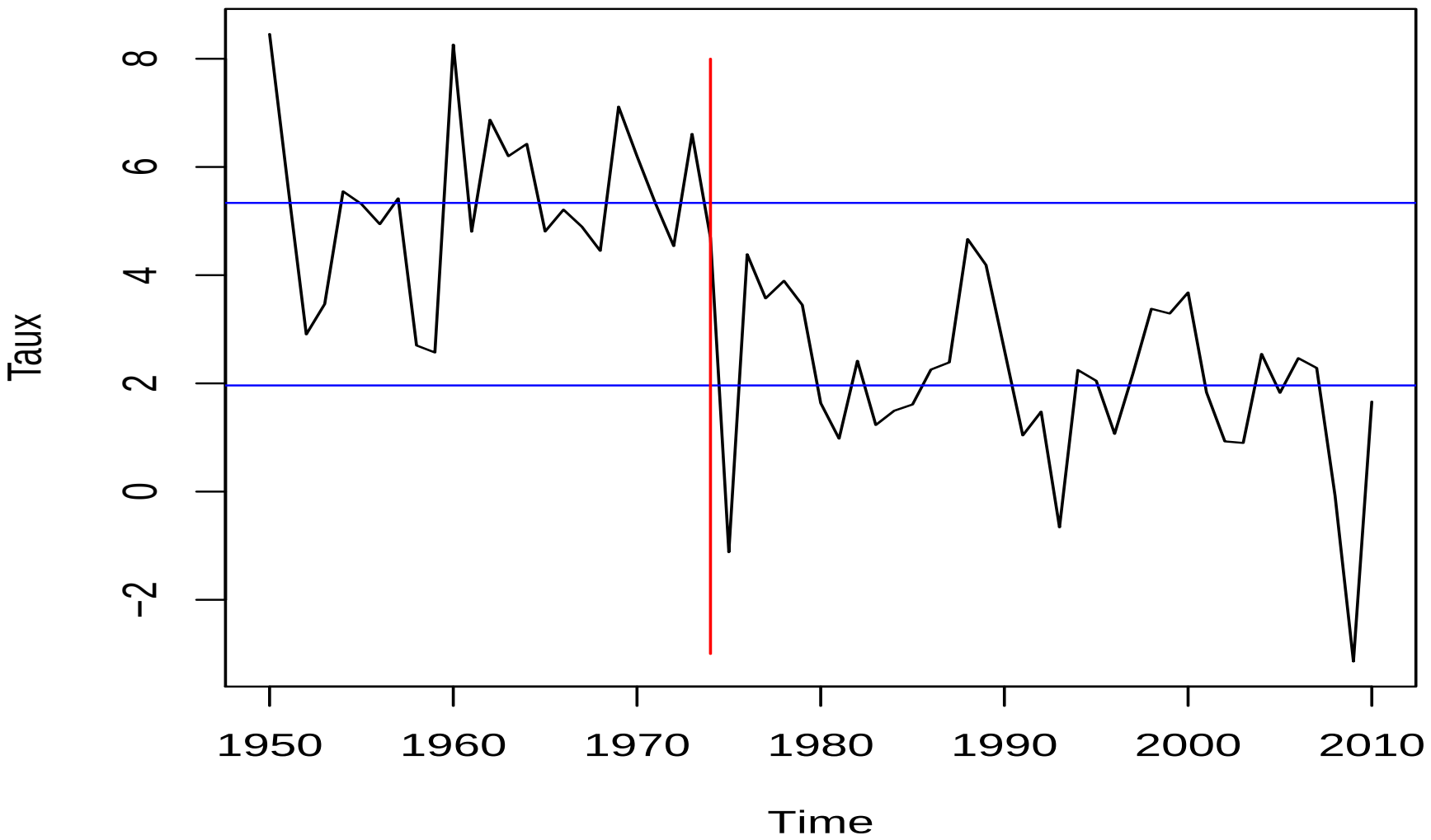


Figure – Taux de croissance annuel en France de 1950 à 2010

Détection de rupture : Températures annuelles du Globe

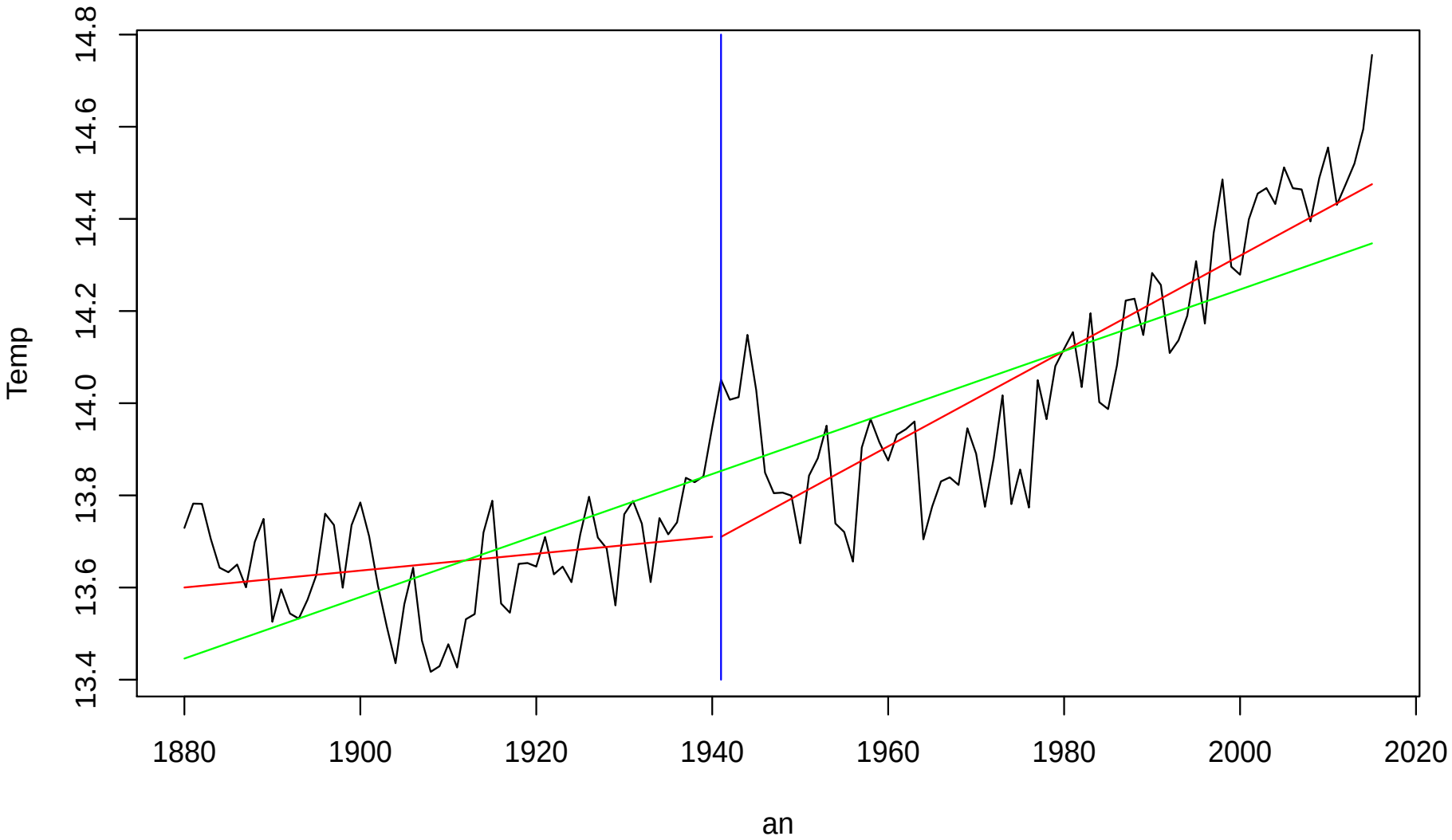


Figure – Détection de rupture linéaire pour les températures annuelles du globe
$\Longrightarrow R^2 \simeq 0.82$ et $\widehat{a}_2 \simeq 0.01$ (pval$\simeq 2\,10^{-16}$) et $\widehat{C} \simeq 55$: **Rupture ! !**

Détection de ruptures dans des modèles linéaires simples

On suppose $(X_1, \cdots, X_n)$ une série chronologique et :

- Pour $1 \leq t \leq t^*$, $X_t = a_1 + b_1 t + \varepsilon_t$;
- Pour $t^* + 1 \leq t \leq n$, $X_t = a_2 + b_2 t + \varepsilon_t$.

On suppose également que t^* est "suffisamment" éloigné de 1 et de n et que (ε_t) n'est pas trop dépendante.

Détection de ruptures dans des modèles linéaires simples

On utilise l'estimateur suivant :

$$\widehat{t} = \underset{10 \leq t \leq n-10}{\operatorname{Arg\,min}} \left(\sum_{k=1}^{t} \left(X_k - \widehat{a}_1(t) - \widehat{b}_1(t)\, k\right)^2 + \sum_{k=t+1}^{n} \left(X_k - \widehat{a}_2(t) - \widehat{b}_2(t)\, k\right)^2 \right)$$

où $\begin{cases} (\widehat{a}_1(t), \widehat{b}_1(t)) \text{ est l'estimateur MCO sur } \{1, \cdots, t\} \\ (\widehat{a}_2(t), \widehat{b}_2(t)) \text{ est l'estimateur MCO sur } \{t+1, \cdots, n\}) \end{cases}$

Détection de rupture

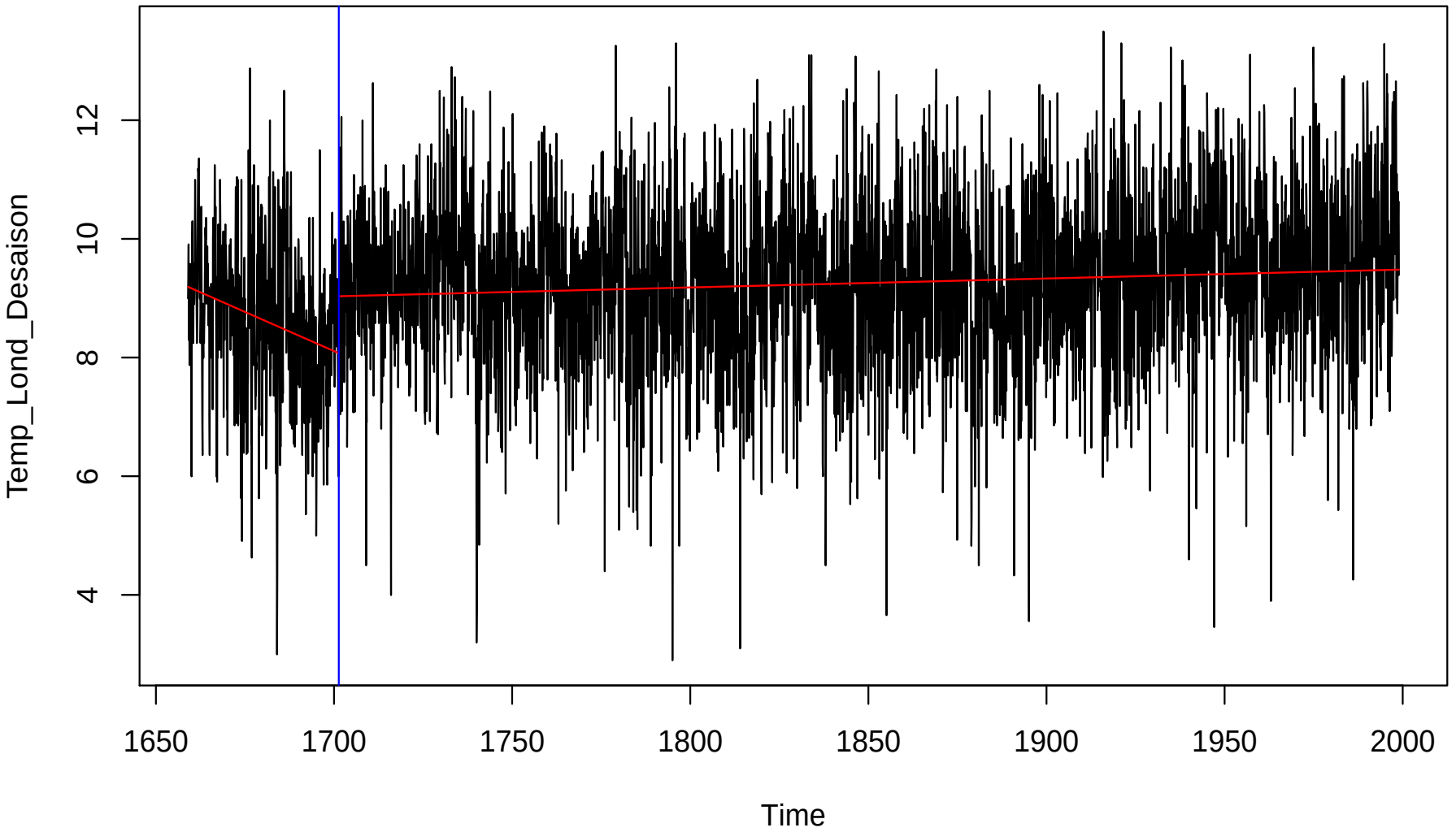


Figure – Détection de rupture linéaire pour les températures mensuelles de Londres désaisonnalisée $\Longrightarrow R^2 \simeq 0.036$ et $\widehat{a}_2 \simeq 0.0015$

Détection de rupture (suite)

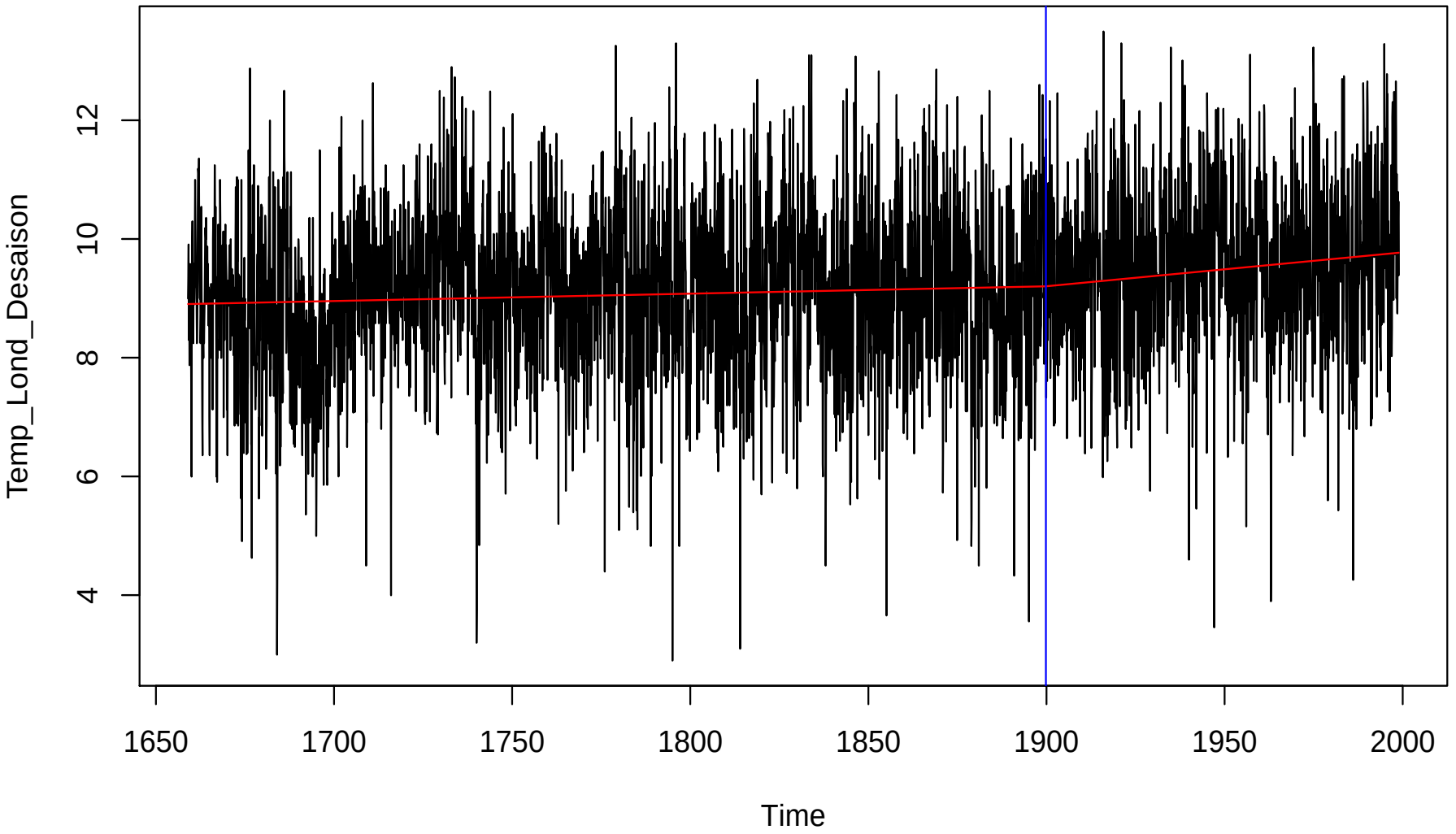


Figure – Détection de rupture dans le "slope" pour les températures mensuelles de Londres désaisonnalisée $\Longrightarrow R^2 \simeq 0.026$ et $\widehat{b_2} \simeq 0.0056$

Remarque théorique

Proposition

D'après Bai et Perron (1998) ou Lavielle et Moulines (2000), sous certaines hypothèses assez faibles :

- $t^* = [n\tau^*]$;
- $(\varepsilon_t)_t$ *suite de variables mélangeantes ;*

Alors, quand $n \to \infty$, *pour toute suite* (a_n) *telle que* $a_n \underset{n\to+\infty}{\longrightarrow} \infty$

$$\left|\widehat{t}_n - t^*\right| = O_{\mathbb{P}}(a_n).$$

$$\Longrightarrow \qquad \left|\widehat{\tau}_n - \tau^*\right| = O_{\mathbb{P}}(a_n).$$

Test de Chow pour modèle linéaire simple

Soit le problème de test suivant :

$$\begin{cases} H_0 : \text{il n'y a pas de rupture ;} \\ H_1 : \text{il y a une rupture.} \end{cases}$$

On utilise la statistique de test de Chow :

$$\widehat{C} = \frac{\frac{1}{2}\left(SCR_0 - SCR_1\right)}{\frac{1}{n-4} SCR_1} \xrightarrow[n\to\infty]{\mathcal{L}} \frac{1}{2}\chi^2(2).$$

où $\begin{cases} SCR_0 = \sum_{k=1}^{n} \left(X_k - \widehat{a}_1(n) - \widehat{b}_1(n)\, k\right)^2 \\ SCR_1 = \sum_{k=1}^{\widehat{t}} \left(X_k - \widehat{a}_1(\widehat{t}) - \widehat{b}_1(\widehat{t})\, k\right)^2 + \sum_{k=\widehat{t}+1}^{n} \left(X_k - \widehat{a}_2(\widehat{t}) - \widehat{b}_2(\widehat{t})\, k\right)^2 \end{cases}$

$\Longrightarrow$ Températures : valide la présence de rupture dans les 2 cas...

Cadre général

Test de Chow

Le test de Chow teste un éventuel changement de structure dans l'écriture du modèle (individu = temps !) temporelle). Il s'agit donc de tester :

$$H_0 : \; Y = Z\,\theta + \varepsilon \qquad \text{contre} \qquad H_1 : \begin{cases} Y^{(1)} = Z^{(1)}\,\theta_1 + \varepsilon^{(1)} \\ Y^{(2)} = Z^{(2)}\,\theta_2 + \varepsilon^{(1)}, \end{cases}$$

où ${}^tY = \left({}^tY^{(1)}, {}^tY^{(2)}\right)$. Sous H_0 on a un sous-modèle de H_1, et le modèle sous H_1 s'écrit comme un modèle linéaire. On peut donc définir un test de Fisher de sous-modèle et on a, sous H_0 :

$$\widehat{F} = \frac{\frac{1}{p+1}}{\frac{1}{n-2(p+1)}} \frac{SC_0 - SC_1}{SC_1} \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \frac{1}{p+1}\,\chi^2(p+1).$$

Si la date de rupture j^* telle que $Y^{(1)} = {}^t(Y_1, \ldots, Y_{j^*})$, $Y^{(2)} = {}^t(Y_{j^*+1}, \ldots, Y_n)$ est inconnue : on choisit celle qui maximise la statistique de Chow ! Cela revient à minimiser :

$$SC_1(j) = \sum_{i=0}^{j} (Y_i - (X^{(1)}\,\widehat{\theta}_1)_i)^2 + \sum_{i=j+1}^{n} (Y_i - (X^{(2)}\,\widehat{\theta}_2)_i)^2$$

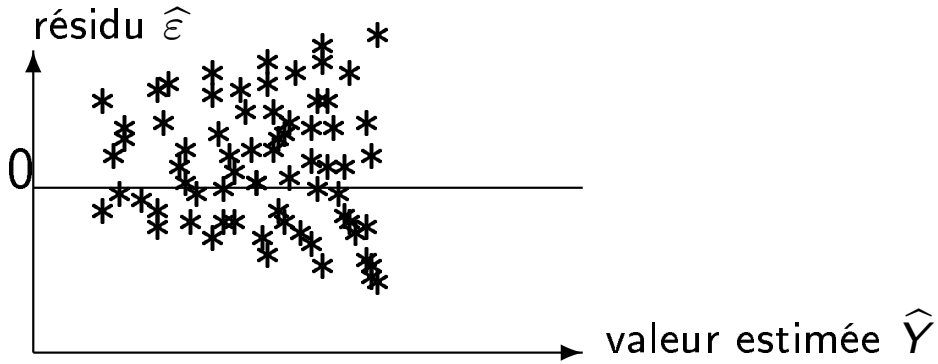
D'où $\widehat{j} = \operatorname{Arg\,min}_j SC_1(j)$.

On montre sous les conditions précédentes que $\frac{\widehat{j}}{j^*} \xrightarrow[n\to+\infty]{\mathcal{P}} 1$ et

$$\widehat{F}(\widehat{j}) = \frac{\frac{1}{p+1}}{\frac{1}{n-2(p+1)}} \frac{SC_0 - SC_1(\widehat{j})}{SC_1(\widehat{j})} \xrightarrow[n\to\infty]{\mathcal{L}} \frac{1}{p+1}\chi^2(p+1).$$

Remise en question de **(A2)** : Hétéroscédasticité

Diagnostic d'héteroscédasticité : on représente les $\widehat{Y}_i$ en fonction des $\widehat{\varepsilon}_i$. Par exemple :



$\Longrightarrow$ variance des résidus semble inhomogène

mais sous **(A1-A3)** $\text{var}(\widehat{\varepsilon}_i) = \sigma^2(1 - p_{ii})$ non constante

$\Longrightarrow$ Tracé des résidus dit "Studentisés"

$\widetilde{\varepsilon}_i = \dfrac{\widehat{\varepsilon}_i}{\sqrt{1 - p_{ii}}\,\widehat{\sigma}^{(i)}}$ où $\widehat{\sigma}^{(i)}$ obtenue à partir de $(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$.

$\Longrightarrow$ Sous **(A1-4)**, $\widetilde{\varepsilon}_i \overset{\mathcal{L}}{\sim} t(n - p - 2)$

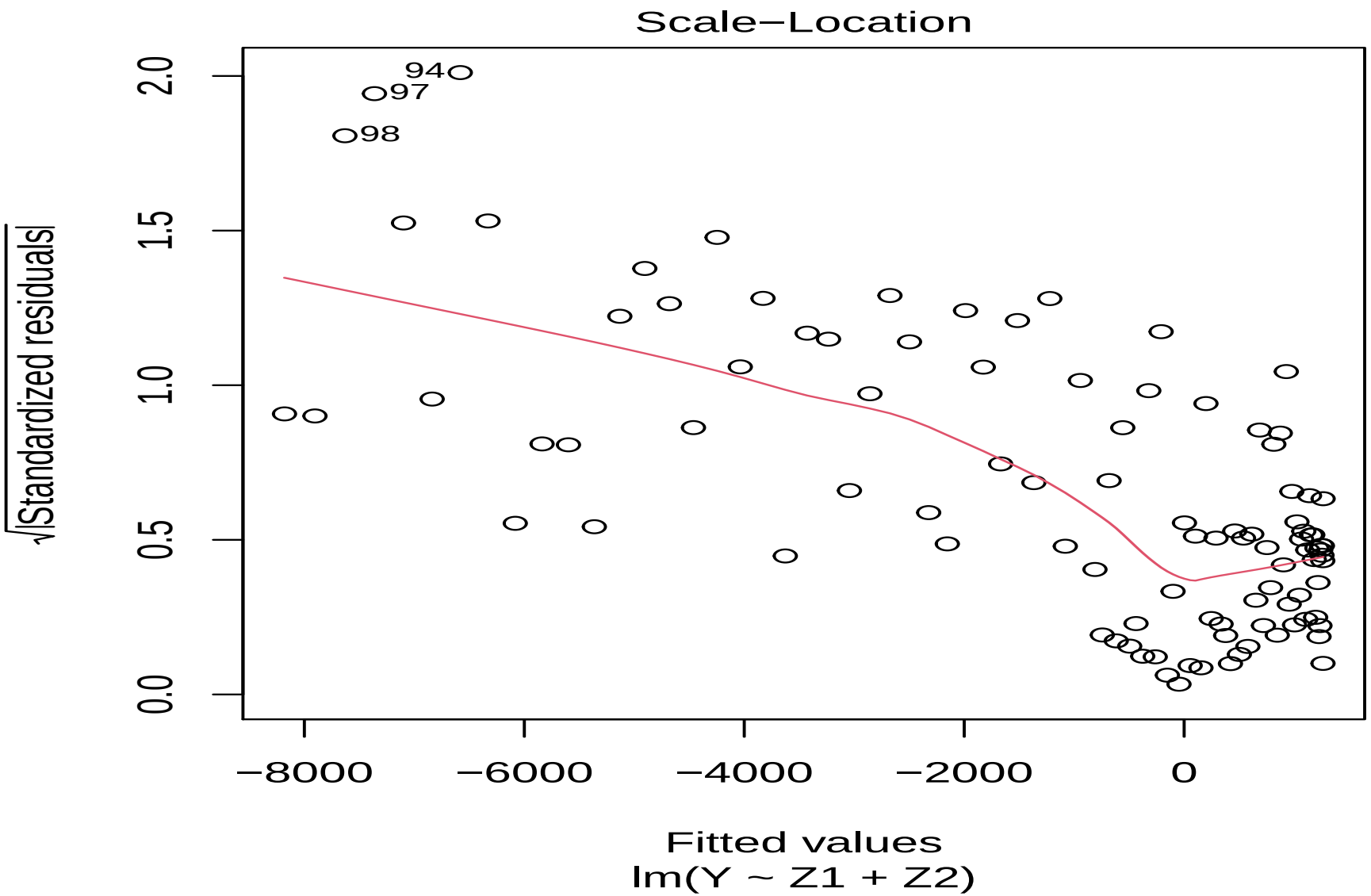


Figure – Exemple d'hétéroscédasticité

Remise en question de **(A2)** : Hétéroscédasticité

Modifications possibles à apporter au modèle :

Transformation de Box-Cox sur Y Lorsque les Y_i sont des variables positives, on peut utiliser :

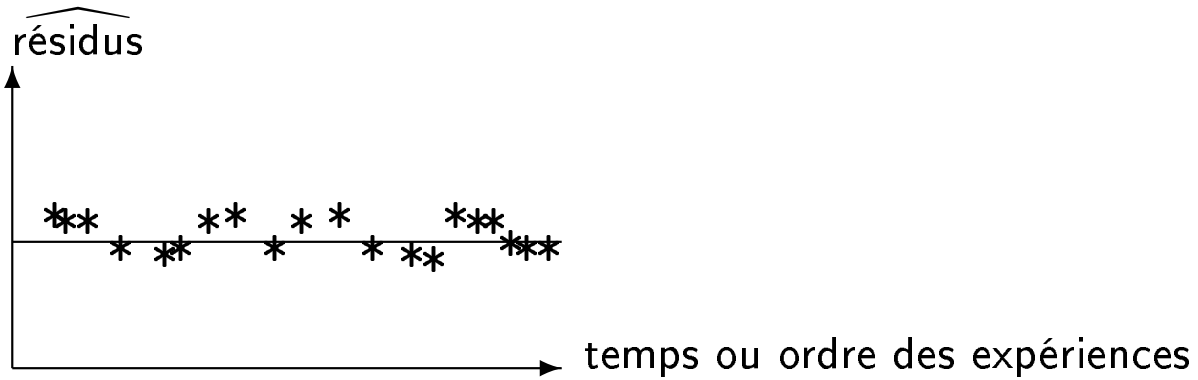
$$\tau(Y_i, \lambda) := \begin{cases} \frac{Y_i^\lambda - 1}{\lambda} & \text{pour } \lambda \neq 0 \\ \log Y_i & \text{pour } \lambda = 0 \end{cases}$$

où λ est un réel a priori inconnu.
Numériquement, à partir d'une grille de valeurs de λ on calcule la variance des résidus $\widehat{\sigma}^2_{\widehat{\varepsilon},\lambda}$ pour chaque valeur de λ. On choisira alors :

$$\widehat{\lambda} = \text{Argmin}_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \left\{ \widehat{\sigma}^2_{\widehat{\varepsilon},\lambda} \right\}.$$

Remise en question de **(A3)** : corrélation du bruit

Graphe des $(\widehat{Y}_i, \widehat{\varepsilon}_i)$ ou $(\widehat{Y}_i, \widetilde{\varepsilon}_i)$ permet de visualiser les éventuels problèmes de dépendance :



$\Longrightarrow$ Les résidus ont tendance à rester par paquets lorsqu'ils se trouvent d'un côté ou de l'autre de 0 ou change de signe trop fréquemment

Remise en question de **(A3)** : corrélation du bruit

- **Test de runs** : basé sur le nombre de runs, c'est-à-dire sur le nombre de paquets de résidus consécutifs de même signe (8 sur le graphe) runs.

  On peut montrer que si $\widehat{N}$ est le nombre de runs, et si on teste :

$$\begin{cases} H_0 : & \text{Indépendance du bruit} \\ H_1 : & \text{Non indépendance du bruit} \end{cases}$$

  alors sous H_0 on montre que $$\frac{\widehat{N} - \frac{1}{2}(n+1)}{\frac{1}{2}\sqrt{n-1}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

- **Test portemanteau** : Si $p_{ij} \underset{n\to+\infty}{\longrightarrow} 0$ suffisamment vite,

$$n \sum_{k=1}^{K_{\max}} \widehat{\rho}^2_{\widehat{\varepsilon}}(k) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(K_{\max}) \quad \text{sous } H_0$$

Utilisation des moindres carrés généralisés

Définition

On suppose le modèle linéaire général

$$Y = X\,\theta + \varepsilon, \qquad \textit{avec} \qquad \mathbb{E}[\varepsilon] = 0.$$

On suppose $\Sigma = \mathbb{E}\left[\varepsilon\,{}^t\varepsilon\right]$ *de rang* n *(*X *est supposée de rang* $p+1$*).*
On considère la distance définie par la norme

$$\|U - V\|_\Sigma^2 = {}^t(U - V)\,\Sigma^{-1}\,(U - V).$$

Distance associée au produit scalaire dans $\mathbb{R}^n$, $< Z_1, Z_2 >= {}^t Z_1\,\Sigma^{-1}\,Z_2$.
L'estimateur $\widehat{\theta}_G$ *de* θ *par* **MCG** *minimise* $\|Y - X\cdot\theta\|_\Sigma$ *pour* $\theta \in \mathbb{R}^{p+1}$ *et*

$$\widehat{\theta}_G = \left({}^t X\,\Sigma^{-1}\,X\right)^{-1}\,{}^t X\,\Sigma^{-1}\,Y.$$

Remarque : Si ε suit une loi $\mathcal{N}_n(0,\Sigma)$, alors $\widehat{\theta}_G$ estimateur MV de θ.

Utilisation des MCG

En utilisant le Théorème de Gauss-Markov, on peut montrer :

Proposition

$\widehat{\theta}_G$ est l'estimateur de θ non biaisé et linéaire ayant la plus petite matrice de variance-covariance ($\leq$ à celle de l'estimateur par MCO).

Problème : En pratique on ne connaît pas a priori la matrice Σ !

$\Longrightarrow$ Estimer Σ à partir des résidus $\widehat{\varepsilon}$ obtenus par MCO.

$\Longrightarrow$ Si $\widehat{\Sigma} \to \Sigma$ avec $\widehat{\Sigma}$ inversible, on approche $\widehat{\theta}_G$ par estimateur $\widetilde{\theta}_G$ appelé **estimateur par moindres carrés pseudo-généralisés** défini par

$$\widetilde{\theta}_G = \left({}^t X \, \widehat{\Sigma}^{-1} \, X\right)^{-1} {}^t X \, \widehat{\Sigma}^{-1} \, Y.$$

Utilisation des MCG (2)

Cas général : comment estimer Σ matrice symétrique de taille n contenant $(n(n+1)/2$ termes à partir d'un échantillon de taille n ? Difficile...

Deux cas particuliers :

- **Cas particulier ARMA :** (ε_n) ARMA(p,q) Démarche en 3 temps :
  1. Par MCO, si $\max_{1\leq i\leq n}|p_{ii}| \underset{n\to+\infty}{\longrightarrow} 0$, $\widehat{\theta} \underset{n\to+\infty}{\overset{\mathcal{P}}{\longrightarrow}} \theta$ et $\widehat{\varepsilon}_i \underset{n\to+\infty}{\overset{\mathbf{L^2}}{\longrightarrow}} \varepsilon_i$ pour $i \in \mathbb{N}$.
  2. Estimation des paramètres de l'ARMA(p,q) par MV à partir de $(\widehat{\varepsilon}_1, \cdots, \widehat{\varepsilon}_n)$. Si $\alpha = (a_1, \cdots, a_p, b_1, \cdots, b_q, \sigma^2)$ les paramètres, on montre que $\widehat{\alpha} \underset{n\to+\infty}{\overset{\mathcal{P}}{\longrightarrow}} \alpha$.
  3. On définit alors $\widehat{\Sigma} = \Sigma(\widehat{\alpha})$ où $\text{cov}(\varepsilon) = \Sigma(\alpha)$ puis l'estimateur par MCPG $\widetilde{\theta}_G$, qui asymptotiquement a une covariance inférieure à celle de $\widehat{\theta}$.

Utilisation des MCG (3)

- **Cas particulier rupture :** (ε_n) tel que :

$$\text{var}(\varepsilon_i) = \sigma_j^2 \text{ si } t_j \leq i < t_{j+1} \quad \text{et} \quad \text{cov}(\varepsilon_k, \varepsilon_\ell) = 0$$

1. On estime les $\widehat{\sigma}_j^2$ par MCO sur chaque zone $\{t_j, t_j + 1, \ldots, t_{j+1} - 1\}$
2. On utilise les MCG avec $\widehat{\Sigma} = \text{diag}(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \ldots, \widehat{\sigma}_m^2)$
3. Estimation des t_j ?

Plan du cours

Points aberrants et points aberrants

Qu'appelle-t-on point aberrant ? Deux sortes :

1. Une donnée qui est issu d'une **erreur**. Typiquement erreur de saisie

   Exemple : Intervertion prix et km dans base de véhicules d'occasion

2. Une donnée qui n'est pas une erreur mais un individu **atypique**

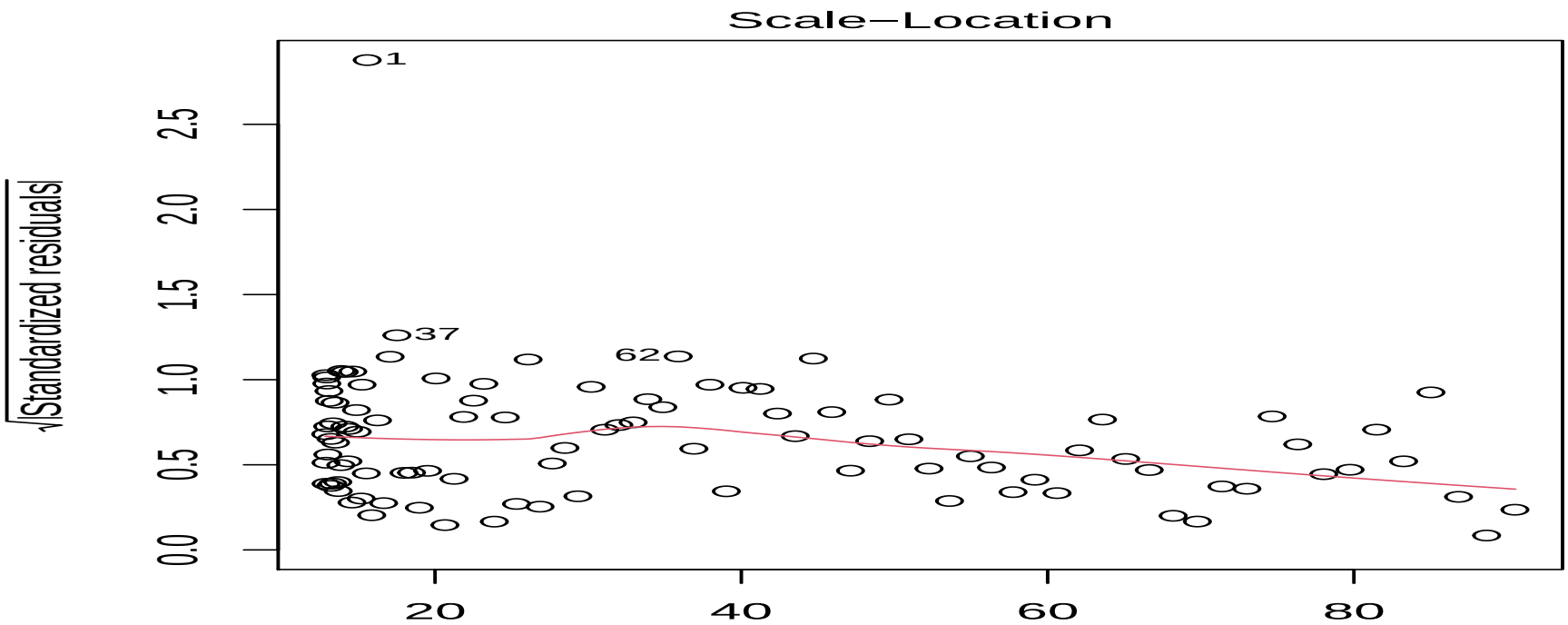
   Exemple : Véhicules de taxis dans base de véhicules d'occasion

Utilisation des résidus studentisés

Tracé des résidus dit "Studentisés"

$\widetilde{\varepsilon}_i = \dfrac{\widehat{\varepsilon}_i}{\sqrt{1-p_{ii}}\,\widehat{\sigma}^{(i)}}$ où $\widehat{\sigma}^{(i)}$ obtenue à partir de $(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$.

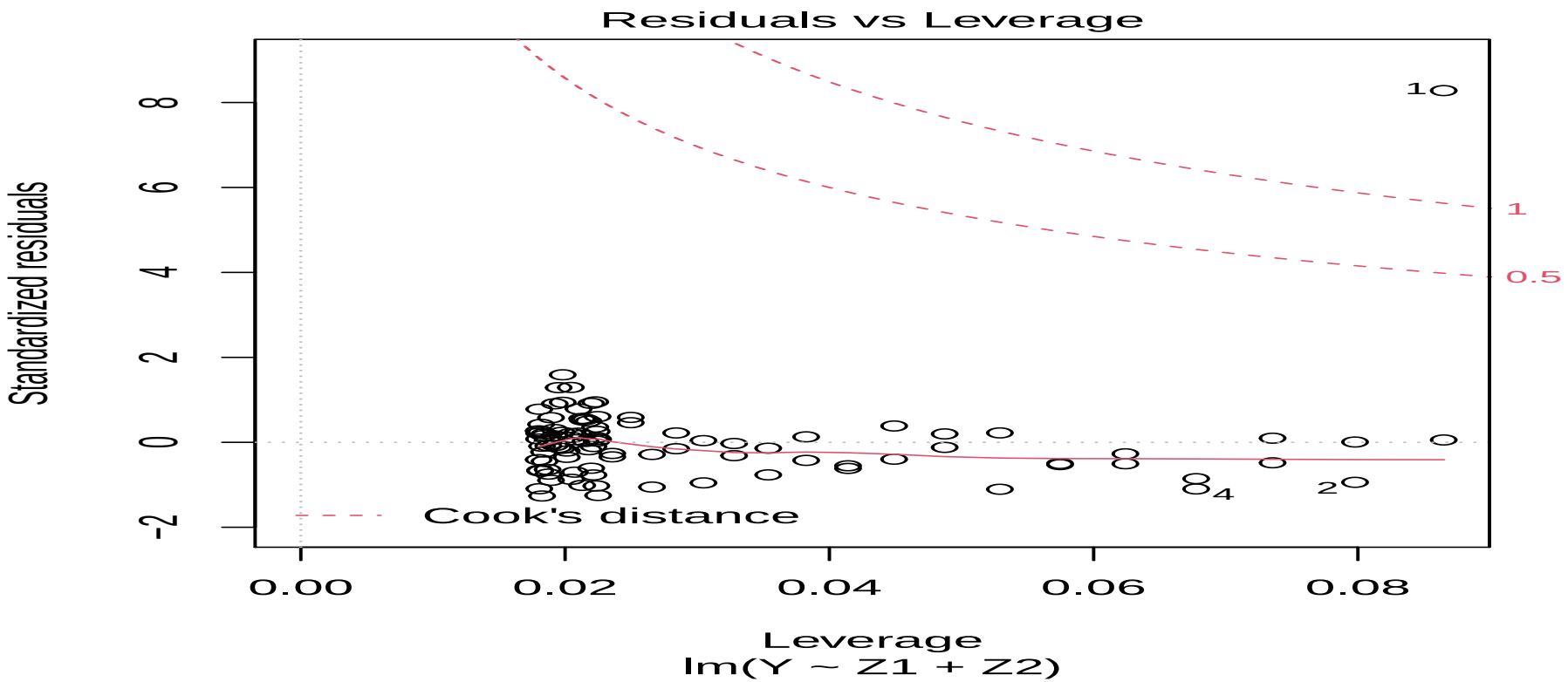
$$\Longrightarrow \text{Sous } \textbf{(A1-4)},\ \widetilde{\varepsilon}_i \overset{\mathcal{L}}{\sim} t(n-p-2)$$



Utilisation de la distance de Cook

On appelle distance de Cook pour la $i$ème observation, la statistique

$$\widehat{D}_i = \frac{\sum_{j=1}^{n}(\widehat{Y}_j - \widehat{Y}_j^{(-i)})^2}{(p+1)\,\widehat{\sigma}^2} = \frac{p_{ii}}{(p+1)(1-p_{ii})}\,\frac{\widehat{\varepsilon}_i^2}{(1-p_{ii})\,\widehat{\sigma}^2}$$



Plan du cours

Régression logistique

Les Y_i variables qualitatives, et pour commencer les Y_i variables à deux modalités, que nous noterons 0 et 1.

Exemple : Y_i mesure l'obtention ou non d'un crédit (économie), la mort ou la vie (biologie, pharmacologie),...

$(Y_1, \cdots, Y_n)$ observés et variables $X^{(1)}, \cdots, X^{(p)}$ variables potentiellement explicatives de Y (quantitatives ou qualitatives).

Comme les Y_i ne prennent pour valeurs que 0 ou 1, on ne peut utiliser un modèle linéaire "habituel", $Y = X\theta + \varepsilon$. On cherchera un modèle reliant les probabilités que $Y = 0$ et $Y = 1$ avec les variables explicatives. Plus concrètement, on note

$$p_i = P(Y_i = 1) \qquad \text{et donc} \qquad 1 - p_i = P(Y_i = 0).$$

L'idée sera ainsi d'écrire que :

$$g(p_i) = \theta_0 + \theta_1 X_i^{(1)} + \cdots + \theta_p X_i^{(p)} \qquad \text{pour tout } i \in \{1, \cdots, n\},$$

où g est une fonction réelle monotone qui va de $[0,1]$ dans $\mathbb{R}$.

On en déduit donc que

$$p_i = p_i(\theta) = g^{-1}\big(\theta_0 + \theta_1 X_i^{(1)} + \cdots + \theta_p X_i^{(p)}\big).$$

Les modèles les plus utilisés de cette fonction g sont les suivants :

1. Fonction **probit** : $g^{-1}(x) = \displaystyle\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$;
2. Fonction **logit** : $g^{-1}(x) = \dfrac{e^x}{1+e^x} \quad \Longrightarrow \quad g(p) = \ln\left(\dfrac{p}{1-p}\right)$;
3. Fonction **log-log** : $g^{-1}(x) = 1 - \exp(-e^x) \implies g(p) = \ln\big(-\ln(1-p)\big)$

On peut trouver des légitimations à l'utilisation des 2 premières fonctions :

1. Fonction **probit** : Si on considère le modèle classique $Z = X\theta + \varepsilon$ avec $\varepsilon \overset{\mathcal{L}}{\sim} \mathcal{N}(0, I_n)$, Z étant une variable dite latente, et $Y = \mathbb{1}_{Z \geq 0}$. Alors on peut montrer que $p_i = g^{-1}((X\theta)_i)$.

2. Fonction **logit** : Si on considère le modèle classique $Z = X\theta + \varepsilon$ avec $\varepsilon_i \overset{\mathcal{L}}{\sim} \mathcal{L}$ (loi logisitique), Z étant une variable dite latente, et $Y = \mathbb{1}_{Z \geq 0}$. Alors on peut montrer que $p_i = g^{-1}((X\theta)_i)$.

Avec le modèle linéaire, on aurait envie d'écrire un modèle de régression suivant :

$$(g(Y_i))_{1\leq i\leq n} = X\theta + \varepsilon$$

et on estime θ par moindres carrés ordinaires. Les $g(Y_i)$ ne prennent que 2 valeurs, mais on peut supposer que X est de rang p donc on peut estimer θ. On en déduit alors les valeurs prédites :

$$\widehat{p}_i = g^{-1}\left((X\widehat{\theta})_i\right),$$

ce qui permet d'avoir une estimation de la probabilité que Y_i soit égale à 1. On en déduit également pour un nouvel individu $n+1$ une prédiction de $P(Y_{n+1} = 1)$:

$$\widehat{p}_{n+1} = g^{-1}\left((X\widehat{\theta})_{n+1}\right).$$

Maximum de vraisemblance

Cette méthode est limitée car la variable observée ne prend que 2 valeurs.

On va plutôt revenir à une estimation par maximum de vraisemblance : les Y_i sont des variables de Bernoulli **indépendantes** de paramètres $p_i(\theta)$. On a la relation $p_i(\theta) = g^{-1}\big((X\theta)_i\big)$ et la vraisemblance s'écrit :

$$L_\theta(Y_1, \cdots, Y_n) = \prod_{i=1}^{n} p_i(\theta)^{Y_i}(1 - p_i(\theta))^{1-Y_i},$$

(on a un modèle de type binomial). En passant au logarithme, on obtiendra ainsi :

$$\widehat{\theta} = \text{Argmax}_{\theta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} Y_i \log\left[g^{-1}\big((X\theta)_i\big)\right] + (1 - Y_i) \log\left[1 - g^{-1}\big((X\theta)_i\big)\right].$$

Approximation par Newton-Raphson

$\Longrightarrow \widehat{\theta}$ non explicite !

$\Longrightarrow$ **Méthode de Newton-Raphson** pour résoudre $f(x_0) = 0$:

Suite récurrente (u_n) définie par :

$$u_{n+1} = u_n - \frac{f(u_n)}{f'(u_n)}.$$

Avec $M_1 = \min_{x \in I} |f'(x)|$ et $M_2 = \max_{x \in I} |f''(x)|$, alors

$$|u_n - x_0| \leq \frac{M_2}{2M_1} |u_{n-1} - x_0|^2 \leq \left(\frac{M_2}{2M_1}\right)^{2^n - 1} |u_0 - x_0|^{2^n}.$$

Convergence quadratique si u_0 proche de x_0 (si $\frac{M_2}{2M_1}|u_0 - x_0| < 1$).

Utilisation de Newton-Raphson

Extremum local d'une fonction régulière, soit $\partial_\theta f(\theta_0) = 0$.

Comme θ est de dimension $p+1$, on cherche une solution de l'équation :

$$\frac{\partial}{\partial \theta_i} \log(L_\theta(Y_1, \cdots, Y_n) = 0 \qquad \text{pour tout } i = 0, \cdots, p.$$

Ainsi on définira la suite $(\theta^{(n)})$ telle que :

$$\theta^{(n+1)} = \theta^{(n)} - \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(L_{\theta^{(n)}}(Y_1, \cdots, Y_n) \right)^{-1}_{ij} \left(\frac{\partial}{\partial \theta_i} \log(L_{\theta^{(n)}}(Y_1, \cdots, Y_n) \right)_i$$

Pénible à calculer (dépendantes de la fonction g choisie, ainsi que de X), mais calcul possible conduisant à un MCG dans le cas Logit (voir le logiciel).

TLC

Théorème

Si les Y_i sont indépendantes, g de classe $\mathcal{C}^2$ et X de rang $(p+1)$, on peut montrer que :

$$\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(L_{\widehat{\theta}}(Y_1,\cdots,Y_n)\right)_{ij}^{1/2}\left(\widehat{\theta}-\theta\right)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}_{p+1}\left(0\,,\,I_{p+1}\right)\qquad(1)$$

Remarques :

1. Les variables Y_i ne sont pas identiquement distribuées
2. On a utilisé le Lemme de Slutsky.

$\Longrightarrow$ Intervalles de confiance et tests.

Validation et tests

Une fois que l'on obtient $\widehat{\theta}$ comment savoir si :

1. le modèle de régression est acceptable ?
2. comment choisir les variables explicatives ?
3. comment choisir entre `logit`, `probit`,...

Pour $\widehat{\theta}$ calculé, on obtient $\widehat{Y}_i = \mathbb{I}_{p_i(\widehat{\theta}) \geq 1/2}$: **valeur prédite**

Matrice de confusion Pas spécifique à la méthode de régression logistique :

$$M = \begin{pmatrix} \text{Card}\big(1 \text{ prédit, } 1 \text{ en réalité}\big) & \text{Card}\big(0 \text{ prédit, } 1 \text{ en réalité}\big) \\ \text{Card}\big(1 \text{ prédit, } 0 \text{ en réalité}\big) & \text{Card}\big(0 \text{ prédit, } 0 \text{ en réalité}\big) \end{pmatrix}.$$

$\Longrightarrow$ Comparer différentes méthodes ou différentes bases d'apprentissage.

Tests

Matrice de confusion : niveau d'erreur à partir duquel le modèle est oui ou non satisfaisant ?

On préférera plutôt des tests issus de la statistique inférentielle.

Test de rapport de vraisemblance : Tester des hypothèses telles que :

$H_0 : \theta_{i_1} = 0, \theta_{i_2} = 0, \cdots, \theta_{i_m} = 0$ contre H_1 : Le modèle est complet,

Exemple : Tester significativité d'une variables ou du modèle global.

Problème de test revenant à tester un sous-modèle contre le modèle complet :

$$\widehat{T} = \frac{\text{Vraisemblance maximisée sous } H_0}{\text{Vraisemblance maximisée sous } H_1}.$$

Test RV

Théorème

Lorsque les variables sont indépendantes les unes les autres on montre que pour un modèle régulier (ce qui est le cas ici lorsque la fonction g est de classe $\mathcal{C}^2$) et sous H_0 :

$$-2 \log\left(\widehat{T}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(p+1-m).$$

Exemple : Si $p = 0$, cela revient test Neyman-Pearson

Test de Wald

Test de Wald : Utilise la normalité asymptotique de l'estimateur pour tester si des coefficients sont nuls ou non

$H_0 : \theta_{i_1} = 0, \theta_{i_2} = 0, \cdots, \theta_{i_m} = 0$ contre $H_1 :$ Le modèle est complet

On définit ainsi :

$$\widetilde{T} = (\widehat{\theta}_{i_1}, \cdots, \widehat{\theta}_{i_m}) \left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log(L_{\widehat{\theta}}(Y_1, \cdots, Y_n)) \right)_{i,j=i_1,\cdots,i_m} \begin{pmatrix} \widehat{\theta}_{i_1} \\ \vdots \\ \widehat{\theta}_{i_m} \end{pmatrix}.$$

Théorème

Sous l'hypothèse H_0, on déduit facilement de (1),

$$\widetilde{T} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(m).$$

Sélection de modèles

Utiliser les critères de sélection de modèles déjà vus pour choisir un meilleur modèle possible.

En particulier, on utilisera le **critère** BIC défini par :

$$\widehat{BIC}(m) = -2 \log(\text{Vraisemblance maximisée pour le modèle } m) + \log n \, |m|,$$

et l'on choisira $\widehat{m} = \text{Argmin}_{m \in \mathcal{M}} BIC(m)$.

Plan du cours

Régression polytomique

Y variable qualitative pouvant prendre J modalités

Exemple : Choix d'un vin blanc, rouge ou rosé

$\longrightarrow$ Estimer en fonction de X la probabilité

$$p_{ij} = P(Y_i = j \,|\, X) \qquad \text{pour } j = 1, \ldots, J \text{ et } i = 1, \ldots, n.$$

On met part une modalité, par exemple la modalité J (choix sans incidence)

On effectue $J - 1$ régressions logistiques avec **logit**

$$\log\left(\frac{p_{ij}}{p_{iJ}}\right) = (X\theta^{(j)})_i \qquad \text{pour } j = 1, \ldots, J-1 \text{ et } i = 1, \ldots, n,$$

avec la condition $\sum_{j=1}^{J} p_{ij} = 1$, ce qui donne toutes les p_{ij}.

Régression polytomique (suite)

Remarque : Pour chaque modalité $j \neq J$ on associe un vecteur $\theta^{(j)}$

On a :

$$p_{ij} = \frac{e^{(X\theta^{(j)})_i}}{1 + \sum_{k=1}^{J-1} e^{(X\theta^{(k)})_i}}$$

Cela signifie concrètement que l'on va estimer

$$\widehat{p}_{ij} = \frac{e^{(X\widehat{\theta}^{(j)})_i}}{1 + \sum_{k=1}^{J-1} e^{(X\widehat{\theta}^{(k)})_i}},$$

et $\widehat{p}_{iJ} = 1 - \sum_{j=1}^{J-1} \widehat{p}_{ij}$.

Pour la prédiction d'une valeur on utilise la règle :

$$\widehat{Y}_i = \text{Argmax}_{j=1,\ldots,J} \widehat{p}_{i,j}.$$

Régression polytomique (suite)

Comment estimer les différents vecteurs $\theta^{(j)}$?

$\Longrightarrow$ Estimateur par maximum de vraisemblance.

Sous l'hypothèse (Y_i) indépendantes, la vraisemblance s'écrit :

$$L_{(\theta^{(1)},\dots,\theta^{(J-1)})}(Y_1,\cdots,Y_n) = \prod_{i=1}^{n}\left(\prod_{j=1}^{J} p_{ij}^{\mathbb{1}_{Y_i=j}}\right).$$

Cela correspond à la vraisemblance d'une loi multinomiale.

$\Longrightarrow$ Algorithme de Newton-Raphson pour approcher $\left(\widehat{\theta}^{(1)},\dots,\widehat{\theta}^{(J-1)}\right)$

$\Longrightarrow$ Tests du rapport de vraisemblance, de Wald, sélection de modèle.

Régression polytomique pour données ordonnées

Si modalités ordonnées, modèle utilisant la variable latente $Z = X\theta + \varepsilon$:

$$Y = \sum_{j=1}^{J} j \, \mathbb{1}_{Z \in [a_{j-1}, a_j[} \quad \text{où} \quad a_0 = -\infty, \; a_J = \infty,$$

et $\alpha = (a_1, \ldots, a_{J-1})$ inconnus et à estimer comme $\theta = {}^t(\theta_0, \ldots, \theta_p)$.

$\Longrightarrow \mathbb{P}(Y_i = j) = p_{ij}(\theta, \alpha) = F\big(a_j - (X\theta)_i\big) - F\big(a_{j-1} - (X\theta)_i\big)$

$\Longrightarrow$ Sous l'hypothèse (Y_i) indépendantes, la vraisemblance s'écrit :

$$L_{\theta,\alpha}(Y_1, \cdots, Y_n) = \prod_{i=1}^{n} \Big(\prod_{j=1}^{J} \big(p_{ij}(\theta, \alpha)\big)^{\mathbb{1}_{Y_i = j}} \Big).$$

$\Longrightarrow$ Estimation numérique + tests et sélection de modèle.

Plan du cours

Moindres carrés non linéaires

Cadre : $(Y_i)_{1\leq i\leq n}$ quantitative à expliquer par $(X_i^{(j)})_{1\leq i\leq n}$, $1 \leq j \leq p$.

MCO ou MCG : lien linéaire entre les Y_i et les $X_i^{(j)}$.

$\Longrightarrow$ Modèle sous une forme plus générale :

$$Y_i = g_\theta(X_i^{(1)}, \cdots, X_i^{(p)}) + \varepsilon_i$$

où Y_i, $X_i^{(k)}$, g_θ fonction connue, mais θ est un vecteur inconnu, les ε_i inconnus, mais indépendants, centrés, de variance finie et constante.

$\Longrightarrow$ On cherche à estimer θ : cadre semi-paramétrique.

Remarque : Cadre non-paramétrique plus général avec un modèle :

$$Y_i = g(X_i^{(1)}, \cdots, X_i^{(p)}) + \varepsilon_i$$

où g est inconnue et $\in$ à un espace fonctionnel ($\mathbb{L}^2$). Approches possibles : estimation par noyaux, par splines, par régressions locales, par projection...

Cas unidimensionnel

On se contentera ici d'évoquer le cas unidimensionnel suivant :

$$Y_i = g_\theta(X_i) + \varepsilon_i.$$

On suppose impossible de linéariser le modèle, comme pour $g_\theta(x) = \theta_1\, x^{\theta_2}$ par un passage au logarithme.

Exemples : Citons par exemple :

- La fonction de Lorentz : $g_\theta(x) = \dfrac{\theta_1}{1 + \left(\frac{x-\theta_2}{\theta_3}\right)^2}$.
- Un modèle de type loi de puissance : $g_\theta(x) = \theta_1 + \theta_2\, x^{\theta_3}$.

Méthode

Le but est donc d'estimer $\theta \Longrightarrow$ méthode des moindres carrés.

Il s'agira donc de minimiser (en θ) :

$$S(\theta) = \sum_{i=1}^{n} \left(Y_i - g_\theta(X_i)\right)^2,$$

$$\Longrightarrow \qquad \widehat{\theta} = \text{Argmin}_{\theta \in \mathbb{R}^d} S(\theta) : \quad \text{(MV dans le cas gaussien)}$$

Contrairement au cas linéaire, il peut y avoir plusieurs solutions.

Exemple : Cas de la fonction de Lorentz $g_\theta(x) = \dfrac{\theta_1}{1 + \left(\frac{x-\theta_2}{\theta_3}\right)^2}$: θ_3 ou $-\theta_3$.

$\Longrightarrow$ Bijectivité de $\theta \mapsto g_\theta$ clé de l'unicité de $\widehat{\theta}$.

Minimisation

Pour minimiser $S(\theta)$ pas de formule explicite.

$\Longrightarrow$ Minimisation d'une fonction à plusieurs variables.

Solution des équations normales :

$$0 = \sum_{i=1}^{n} \frac{\partial g_\theta}{\partial \theta}(X_i)\big(Y_i - g_\theta(X_i)\big)$$

Ceci s'écrit matriciellement :

$$0 = {}^t\dot{G}_\theta(X)\left(Y - G_\theta(X)\right)$$

avec $G_\theta(X) = \big(g_\theta(X_i)\big)_i$, $Y = (Y_i)_i$ et $\dot{G}_\theta(X)$ la matrice $\left(\frac{\partial g_\theta}{\partial \theta_j}(X_i)\right)_{ij}$.

Si $\widehat{\theta}$ une solution de l'équation normale supposée unique, alors

$$0 = {}^t\dot{G}_{\widehat{\theta}}(X)\left(Y - G_{\widehat{\theta}}(X)\right). \qquad (2)$$

Minimisation (suite)

Deux possibilités :

1. On utilise une méthode de résolution de type Newton-Raphson ;

2. On "linéarise" le système avec suite $(\widetilde{\theta}^k)_k$ et formule de Taylor

$$\begin{aligned} G_{\widehat{\theta}}(X) &\simeq G_{\widetilde{\theta}^k}(X) + \dot{G}_{\widetilde{\theta}^k}(X)\left(\widehat{\theta} - \widetilde{\theta}^k\right) \\ &\simeq G_{\widetilde{\theta}^k}(X) + \dot{G}_{\widetilde{\theta}^k}(X)\left(\widetilde{\theta}^{k+1} - \widetilde{\theta}^k\right) \end{aligned}$$

On peut écrire que $Y - G_{\widehat{\theta}}(X) \simeq Y - G_{\widetilde{\theta}^k}(X) - \dot{G}_{\widetilde{\theta}^k}(X)\left(\widetilde{\theta}^{k+1} - \widetilde{\theta}^k\right)$.

Les équations normales s'approchent alors par les équations :

$$0 = {}^t\dot{G}_{\widetilde{\theta}^k}(X)\left(Y - G_{\widetilde{\theta}^k}(X) - \dot{G}_{\widetilde{\theta}^k}(X)\left(\widetilde{\theta}^{k+1} - \widetilde{\theta}^k\right)\right).$$

Si $\Delta^{k+1} = \widetilde{\theta}^{k+1} - \widetilde{\theta}^k$ et $Z^k = Y - G_{\widetilde{\theta}^k}(X)$ (connu) :

$$0 = {}^t\dot{G}_{\widetilde{\theta}^k}(X)\left(Z^k - {}^t\dot{G}_{\widetilde{\theta}^k}(X)\,\Delta^{k+1}\right)$$

$$\Longrightarrow \quad \Delta^{k+1} = \left({}^t\dot{G}_{\widetilde{\theta}^k}(X)\,\dot{G}_{\widetilde{\theta}^k}(X)\right)^{-1} {}^t\dot{G}_{\widetilde{\theta}^k}(X)\,Z^k.$$

Plan du cours

TCL pour l'estimateur des moindres carrés non-linéaires

Soit $\widehat{\theta}$, solution des MCNL.

Proposition

On suppose que :

1. Θ *est un ouvert borné de* $\mathbb{R}^d$;
2. *Pour* $\theta_1, \theta_2 \in \Theta$, $\frac{1}{n}\sum_{i=1}^{n}\left(g_{\theta_1}(X_i) - g_{\theta_2}(X_i)\right)^2 \underset{n\to+\infty}{\longrightarrow} K(\theta_1,\theta_2) < \infty$ *et* $K(\theta_1,\theta_2) \neq 0$ *si* $\theta_1 \neq \theta_2$;
3. *Pour tout* $x = X_i$, g_θ, $\left(\frac{\partial^2 g_\theta}{\partial\theta_i \partial\theta_j}(x)\right)$ *est continue au voisinage de* θ^* ;
4. *Pour tout* θ *dans un voisinage de* θ^*, *la matrice* $I(\theta) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\dot{G}_\theta(X_i)^t\dot{G}_\theta(X_i)$ *existe et est inversible ;*
5. $(\varepsilon_i)_i$ *suite de vaiid centrées, telles que* $\mathrm{var}\varepsilon_i = \sigma^2$ *et* $\mathbb{E}(\varepsilon_i^4) < \infty$.

Alors :

$$\sqrt{n}\left(\widehat{\theta} - \theta^*\right) \underset{r\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(0\,,\,\sigma^2\, I(\theta^*)^{-1}\right).$$

Idée de preuve

On montre que :

1. $\frac{1}{n} S(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - g_\theta(X_i)\right)^2 \xrightarrow[n\to+\infty]{\mathcal{P}} \sigma^2 + K(\theta^*, \theta)$ minimisé en θ^*.
2. $\frac{1}{n} \frac{\partial}{\partial \theta} S(\theta^*)$ vérifie un TLC :

$$\sqrt{n} \frac{1}{n} \frac{\partial}{\partial \theta} S(\theta^*) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0\,,\, I(\theta^*)\right)$$

3. Formule de Taylor pour $\frac{1}{n} \frac{\partial}{\partial \theta} S(\widehat{\theta})$:

$$\frac{1}{n} \frac{\partial}{\partial \theta} S(\theta^*) \simeq \left(\frac{1}{n} \frac{\partial^2}{\partial \theta^2} S(\theta^*)\right) (\widehat{\theta} - \theta^*)$$

Conséquence

Grâce au TLC, on peut en déduire :

- Intervalles de confiance pour les θ_i
- Tests de Wald et tests du rapport de vraisemblance
- Sélection de modèle avec un BIC (log-vraisemblance gaussienne)