

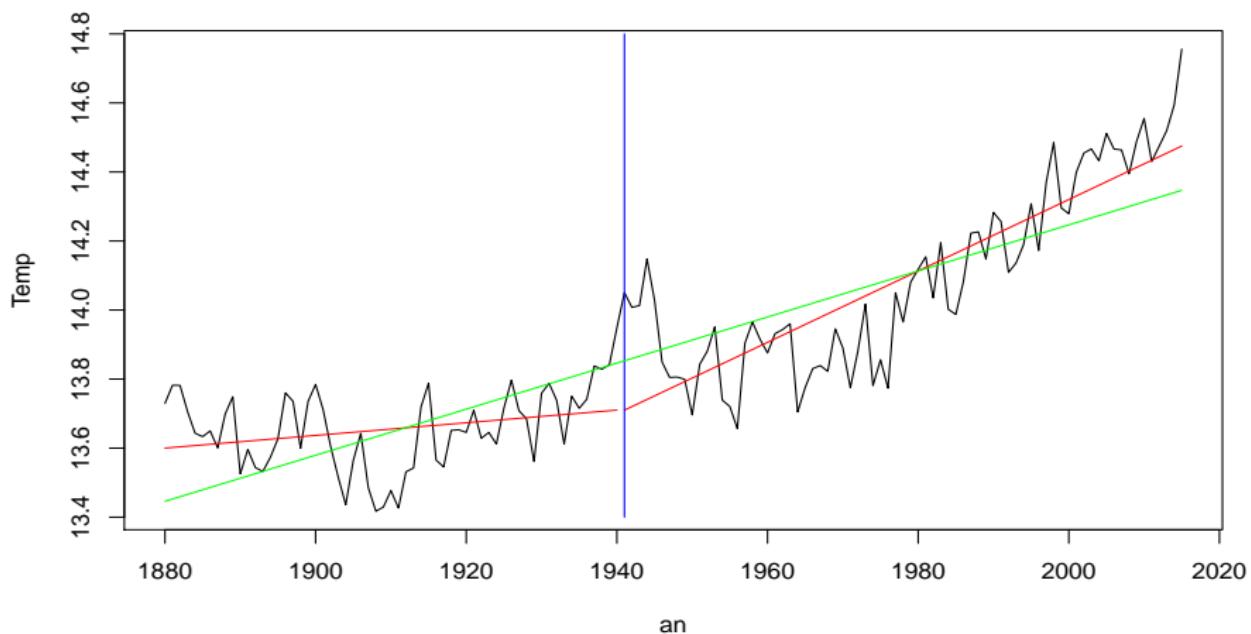
Cours d'Econométrie 2

Master M.A.E.F. Première Année



Année 2023-2024

Importance du Modèle linéaire



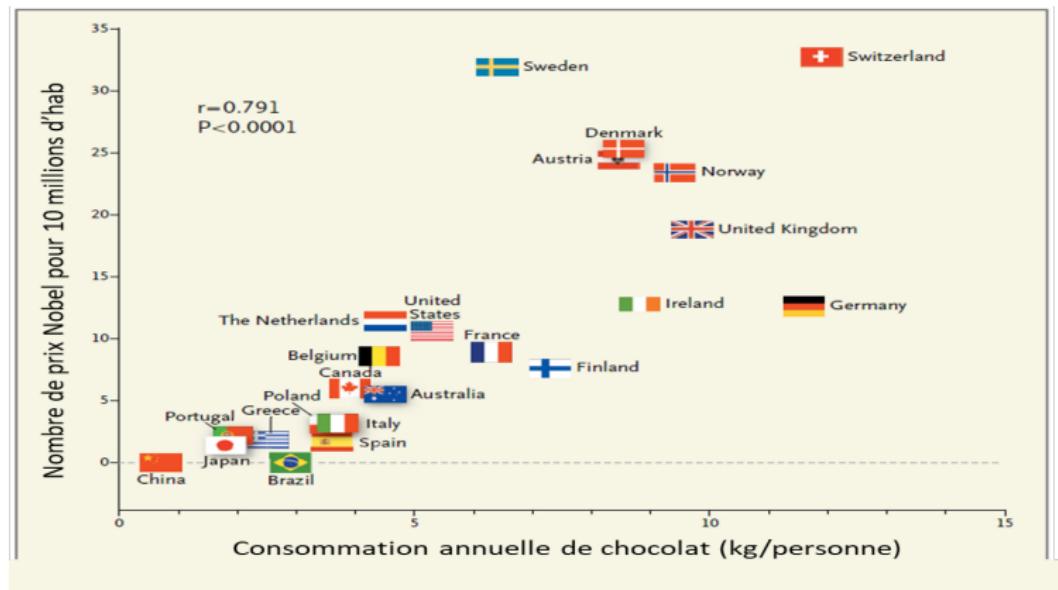
Importance du Modèle linéaire (2)

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

⇒ 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 = $\max(Par, \frac{1}{2}(CC + Par))$ où $CC = \max(CC1, CC2)$

Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires
- Comportement asymptotique des estimateurs et tests

Le cadre général du modèle linéaire

Dans toute la suite, on supposera que :

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

- X est relatif aux variables exogènes (explicatives)

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

- Y est la variable endogène (à expliquer)

\implies 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 $Z = \begin{pmatrix} {}^t X_1 \\ {}^t X_2 \\ \vdots \\ {}^t X_n \end{pmatrix}$.

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 $i = 1, \dots, n$ où Q variable quantitative réelle : Température en fonction de l'année

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

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

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

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 $i = 1, \dots, n$ où F variable qualitative prenant J modalités (x_j) : Décès par Covid/continent

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

- Modèle linéaire simple avec facteur :

$Y_i = \sum_{j=1}^J \theta_{0,j}^* \mathbb{I}_{F_i=x_j} + \sum_{j=1}^J \theta_{1,j}^* Q_i \mathbb{I}_{F_i=x_j} + \varepsilon_i$ pour $i = 1, \dots, n$ 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 :

$$\hat{\theta} = \operatorname{Argmin}_{\theta \in \mathbb{R}^{p+1}} \|Y - Z\theta\|^2 \quad \text{où} \quad \|U\|^2 = {}^t U U \quad \text{pour } U \in \mathbb{R}^n$$

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

Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires
- Comportement asymptotique des estimateurs et tests

Les hypothèses "classiques"

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

- ① **(A0)** $n \geq p + 1$ et Z de rang $p + 1$
 \implies la matrice $(p + 1) \times (p + 1)$ ${}^t Z Z$ est inversible ;
- ② **(A1)** $E(\varepsilon_i) = 0$ pour tout $i = 1, \dots, n$;
- ③ **(A2) Homoscédasticité** : $\text{var}(\varepsilon_i) = \sigma^2$ pour tout $i = 1, \dots, n$;
- ④ **(A3) Non corrélation** : pour tout $i \neq j$, $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$;
- ⑤ **(A4) Gaussianité** : ε vecteur gaussien ;

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

Les prémisses

Propriété

Pour le produit scalaire "classique" dans \mathbb{R}^n , $\langle U, V \rangle = {}^t U V$, 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 $\widehat{\theta} = \operatorname{Argmin}_{\theta \in \mathbb{R}^{p+1}} \|Y - Z\theta\|^2$ et sous **(A0)**, avec $[Z] = \{Z\gamma, \gamma \in \mathbb{R}^{p+1}\}$ sev de \mathbb{R}^n ,

$$Z\widehat{\theta} = P_{[Z]} Y = Z({}^t Z Z)^{-1} {}^t Z Y \implies \widehat{\theta} = ({}^t Z Z)^{-1} {}^t Z Y.$$

Les prémisses (2)

Démonstration.

On remarque que $\min_{\theta \in \mathbb{R}^{p+1}} \|Y - Z\theta\|^2 = \min_{U \in [Z]} \|Y - U\|^2$. Si $U = P_{[Z]} Y + V$ avec $V \in [Z]^\perp$, alors $\|Y - U\|^2 = \|P_{[Z]^\perp} Y - V\|^2 = \|P_{[Z]^\perp} Y\|^2 + \|V\|^2$ par Pythagore. Donc minimum pour $V = 0$.

Montrons que $P_{[Z]} = Z ({}^t Z Z)^{-1} {}^t Z$. Si $U \in [Z]$, d'où $U = Z\gamma$, alors

$P_{[Z]} U = Z ({}^t Z Z)^{-1} {}^t Z Z\gamma = Z\gamma = U$. De plus, pour $Y \in \mathbb{R}^n$, alors on doit avoir

$Y - P_{[Z]} Y \in [Z]^\perp$. Or ${}^t Z (Y - Z ({}^t Z Z)^{-1} {}^t Z Y) = {}^t Z Y - {}^t Z Z ({}^t Z Z)^{-1} {}^t Z Y = 0$ donc pour tout $Y \in \mathbb{R}^n$, on a bien $Y - Z ({}^t Z Z)^{-1} {}^t Z Y \in [Z]^\perp$.

Enfin, sous **(A0)**, Z de rang $p+1$, alors $Z\theta = 0 \implies \theta = 0$. Donc si $Z\widehat{\theta} = Z(({}^t Z Z)^{-1} {}^t Z Y)$ alors $\widehat{\theta} = ({}^t Z Z)^{-1} {}^t Z Y$. □

Premiers résultats

Propriété

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

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

Démonstration.

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

Premiers résultats (3)

Pourquoi choisir les moindres carrés ? Trois raisons :

- ① Les formules sont explicites et immédiatement calculables ;
- ② $\hat{\theta}$ 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 $\tilde{\theta}$ autre estimateur linéaire ($\tilde{\theta} = M Y$, M matrice) et sans biais de θ^* ($E[\tilde{\theta}] = \theta^*$) alors $cov(\hat{\theta}) \leq cov(\tilde{\theta})$.

Démonstration.

Voir TD !

Premiers résultats (4)

Conséquence : Si on rajoute **(A4)**, on a

$$\left\{ \begin{array}{l} \hat{\theta} \stackrel{\mathcal{L}}{\sim} \mathcal{N}_{p+1}(\theta^*, \sigma^2 ({}^t Z Z)^{-1}) \\ \hat{Y} \stackrel{\mathcal{L}}{\sim} \mathcal{N}_n(Z \theta^*, \sigma^2 P_{[Z]}) \\ \hat{\varepsilon} \stackrel{\mathcal{L}}{\sim} \mathcal{N}_n(0, \sigma^2 P_{[Z]^\perp}) \end{array} \right.$$

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 $\hat{\theta}$, \hat{Y} ou $\hat{\varepsilon} \implies$ Estimer σ^2 !

Premiers résultats (5)

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)**, $(\widehat{\theta}, \frac{1}{n} \sum_{i=1}^n \widehat{\varepsilon}_i^2)$ estimateur du MV de (θ^*, σ^2) .

Remarque : Sous **(A0)-(A4)**, le risque quadratique de $\widetilde{\sigma^2} = \frac{1}{n} \sum_{i=1}^n \widehat{\varepsilon}_i^2$ est < à celui de $\widehat{\sigma^2}$.

Premiers résultats (6)

Propriété

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

Démonstration.

On a $\widehat{\sigma^2} = \frac{1}{n-(p+1)} \|\widehat{\varepsilon}\|^2 = \frac{1}{n-(p+1)} {}^t \widehat{\varepsilon} \widehat{\varepsilon} = \frac{1}{n-(p+1)} {}^t (P_{[Z]^\perp} \varepsilon) (P_{[Z]^\perp} \varepsilon) = \frac{1}{n-(p+1)} {}^t \varepsilon P_{[Z]^\perp} \varepsilon$.

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 AB et BA existent, alors $\text{Trace}(AB) = \text{Trace}(BA)$. 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

$\mathbb{E}[\text{Trace}(U)] = \text{Trace}(\mathbb{E}[U])$ pour U une matrice aléatoire. D'où

$\mathbb{E}[\widehat{\sigma^2}] = \frac{1}{n-(p+1)} \text{Trace}(\mathbb{E}[P_{[Z]^\perp} \varepsilon {}^t \varepsilon]) = \frac{1}{n-(p+1)} \text{Trace}(P_{[Z]^\perp} \mathbb{E}[{}^t \varepsilon \varepsilon])$. Mais

$\mathbb{E}[{}^t \varepsilon \varepsilon] = \text{cov}(\varepsilon) = \sigma^2 I_n$. Donc $\mathbb{E}[\widehat{\sigma^2}] = \frac{\sigma^2}{n-(p+1)} \text{Trace}(P_{[Z]^\perp})$. Enfin, avec la propriété sur les valeurs propres d'une matrice de projection, on en déduit que

$\text{Trace}(P_{[Z]^\perp}) = \dim([Z]^\perp) = n - (p + 1)$. D'où $\mathbb{E}[\widehat{\sigma^2}] = \sigma^2$.



Le cas gaussien

Théorème (Théorème de Cochran)

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

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

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 $(P_{A_i}\varepsilon)_i$ est une famille indépendante si $\text{cov}(P_{A_i}\varepsilon, P_{A_j}\varepsilon) = 0$ pour $i \neq j$ car $(P_{A_1}\varepsilon, \dots, P_{A_\ell}\varepsilon)$ est un vecteur gaussien (de taille ℓn). Or $\text{cov}(P_{A_i}\varepsilon, P_{A_j}\varepsilon) = \mathbb{E}[P_{A_i}\varepsilon^t (P_{A_j}\varepsilon)] = \mathbb{E}[P_{A_i}\varepsilon^t \varepsilon^t P_{A_j}] = P_{A_i} \text{cov}(\varepsilon)^t P_{A_j} = \sigma^2 P_{A_i} P_{A_j}$. Mais $A_i \perp\!\!\!\perp A_j$ donc $P_{A_i} P_{A_j} = 0$, d'où le résultat.



Le cas gaussien (2)

Démonstration.

② $\|P_{A_i}\varepsilon\|^2 = {}^t\varepsilon{}^t P_{A_i} P_{A_i}\varepsilon = {}^t\varepsilon{}^t P_{A_i} \varepsilon$. Mais P_{A_i} est une matrice réelle symétrique donc diagonalisable et on peut écrire que $P_{A_i} = Q_i D_i {}^t Q_i$ 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 $\|P_{A_i}\varepsilon\|^2 = {}^t({}^t Q_i \varepsilon) D_i ({}^t Q_i \varepsilon)$. Soit $\varepsilon' = {}^t Q_i \varepsilon$. Alors $\varepsilon' \stackrel{\mathcal{L}}{\sim} \mathcal{N}_n(0, \sigma^2 {}^t Q_i Q_i) = \mathcal{N}_n(0, \sigma^2 I_n)$ car Q_i est une matrice orthogonale (donc $Q_i^{-1} = {}^t Q_i$). 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 $\sum_{j=1}^{\dim(A_i)} (\varepsilon'_j)^2 \stackrel{\mathcal{L}}{\sim} \sigma^2 \chi^2(\dim(A_i))$.



Le cas gaussien (3)

Proposition

Sous (A0)-(A4),

- ① On a \widehat{Y} et $\widehat{\varepsilon}$ indépendants ;
- ② On a $\widehat{\sigma^2} \stackrel{\mathcal{L}}{\sim} \frac{\sigma^2}{n-(p+1)} \chi^2(n - (p + 1))$ 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 $\widehat{\sigma^2} = \frac{1}{n-(p+1)} \|P_{[Z]^\perp} \varepsilon\|^2$, donc d'après Cochran, $\widehat{\sigma^2} \stackrel{\mathcal{L}}{\sim} \frac{\sigma^2}{n-(p+1)} \chi^2(\dim([Z]^\perp))$ 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, $\widehat{\theta} = ({}^t Z Z)^{-1} {}^t Z Y = ({}^t Z Z)^{-1} {}^t Z Z ({}^t Z Z)^{-1} {}^t Z Y = ({}^t Z Z)^{-1} {}^t Z \widehat{Y}$. 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 $\text{cov}(\widehat{Y}, \widehat{\varepsilon}) = \text{cov}(Z\theta^* + P_{[Z]} \varepsilon, P_{[Z]^\perp} \varepsilon) = P_{[Z]} \text{cov}(\varepsilon, P_{[Z]^\perp} \varepsilon) = 0$.

Applications aux tests statistiques

Rappels :

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

Proposition

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

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

Démonstration.

(A0)-(A4) $\implies {}^t C (\hat{\theta} - \theta^*)$ est une variable gaussienne, avec $\mathbb{E}[{}^t C (\hat{\theta} - \theta^*)] = 0$ et $\text{cov}({}^t C (\hat{\theta} - \theta^*)) = {}^t C \text{cov}(\hat{\theta}) C = \sigma^2 {}^t C ({}^t Z Z)^{-1} C \implies {}^t C (\hat{\theta} - \theta^*) \stackrel{\mathcal{L}}{\sim} \mathcal{N}(0, \sigma^2 {}^t C ({}^t Z Z)^{-1} C)$.

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



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 :

$$\widehat{T} = \frac{1}{\sqrt{\widehat{\sigma}^2}} ({}^t C ({}^t Z Z)^{-1} C)^{-1/2} {}^t C \widehat{\theta} \stackrel{\mathcal{L}}{\sim} t(n - (p + 1))$$

Exemples d'utilisation :

- Avec ${}^t C = (0, \dots, 0, 1, 0, \dots, 0)$, on teste $H_0 : \theta_j^* = 0$
 \implies On teste ainsi si $X^{(j)}$ est significative
- Avec ${}^t C = (0, \dots, 0, \pm 1, 0, \dots, 0, \pm 1, 0, \dots, 0)$, on teste $H_0 : \theta_j^* = \theta_\ell^*$
 \implies 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 :

$$\frac{1}{\sigma^2} {}^t(\hat{\theta} - \theta^*) C ({}^t C ({}^t Z Z)^{-1} C)^{-1} {}^t C (\hat{\theta} - \theta^*) \stackrel{\mathcal{L}}{\sim} \chi^2(q)$$

Démonstration.

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



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 :

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

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 $\widehat{\sigma}^2 \stackrel{\mathcal{L}}{\sim} \frac{\sigma^2}{n - (p + 1)} \chi^2(n - (p + 1))$ 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 ${}^t C = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$, on teste $H_0 : \theta_j^* = 0, j = 1, \dots, p$

\implies On teste ainsi grossièrement le modèle

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

\implies 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

$$\hat{F} = \frac{\frac{1}{p-p_0} (\|P_{[Z_{(0)}]^\perp} Y\|^2 - \|P_{[Z]^\perp} Y\|^2)}{\widehat{\sigma}^2} \implies \hat{F} \stackrel{\mathcal{L}}{\sim} F(p - p_0, n - (p + 1)) \text{ sous } H_0$$

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 $[Z]^\perp \oplus A = [Z_{(0)}]^\perp$ et ainsi $\|P_{[Z_0]^\perp} Y\|^2 - \|P_{[Z]^\perp} Y\|^2 = \|P_A Y\|^2 = \|P_A \varepsilon\|^2$ par Pythagore et sous l'hypothèse H_0 , donc $Z \theta^* \in [Z_{(0)}]$. Comme ε vecteur gaussien, avec Cochran, on a donc

$$\frac{1}{p-p_0} \|P_A \varepsilon\|^2 \stackrel{\mathcal{L}}{\sim} \frac{\sigma^2}{p-p_0} \chi^2(p-p_0) \text{ car } \dim(A) = \dim([Z]) - \dim([Z_{(0)}]).$$

Par ailleurs, au dénominateur, on a vu que $\widehat{\sigma^2} = \frac{1}{n-(p+1)} \|P_{[Z]^\perp} \varepsilon\|^2 \stackrel{\mathcal{L}}{\sim} \frac{\sigma^2}{n-(p+1)} \chi^2(n-(p+1))$ 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 $\widehat{F} \stackrel{\mathcal{L}}{\sim} F(p-p_0, n-(p+1))$ sous H_0 .

□

Remarque : Test de Fisher cas particulier du précédent : rang de $[Z_{(0)}] = p_0 + 1 \implies \exists p_0 + 1$ combi. lin. des colonnes de Z engendrant $[Z_{(0)}]$. Avec $[Z_{(0)}^c]$ sev de \mathbb{R}^n tel que $[Z_{(0)}] \oplus [Z_{(0)}^c] = [Z]$, alors $[Z_{(0)}^c]$ engendré par $p - p_0$ combi. lin. de colonnes de $Z \implies H_0$ équivalente au fait que ces $p - p_0$ combi. lin. sont nulles.

Applications aux intervalles de confiance

On peut déduire des résultats précédents :

- ① Sur les $\hat{\theta}_j^*$: avec q_α quantile de niveau $1 - \alpha/2$ d'une $t(n - (p + 1))$:

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

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

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

Coefficient de détermination R^2

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

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{Y}_n \mathbb{I}_n\|^2} \quad \text{avec } \mathbb{I}_n = {}^t(1, \dots, 1)$$

R^2 mesure le "pouvoir de prédition" 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 $\begin{cases} H_0 : \theta_i^* = 0, \ 1 \leq i \leq p \\ H_1 : \exists i_0, \ \theta_{i_0}^* \neq 0 \end{cases} \implies \text{statistique de Fisher}$

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

Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires
- Comportement asymptotique des estimateurs et tests

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 $+\infty$

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

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 :

$$(\mathbb{E}[|X_0|] < \infty) \iff (\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{p.s.} \mathbb{E}[X_0])$$

Théorème (Théorème de la limite centrale)

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

$$\sqrt{n} (\bar{X}_n - \mathbb{E}[X_0]) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, \text{var}(X_0)).$$

Un nouveau théorème central limite

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

- $\widehat{\theta}^{(n)} = \theta^* + ({}^t Z Z)^{-1} {}^t Z \varepsilon$ 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 Feller-Lindeberg)

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

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

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

Un nouveau théorème central limite (2)

Démonstration.

Soit $Z_n = \sum_{j=1}^n a_j^{(n)} U_j$. Alors pour tout $u \in \mathbb{R}$, $\phi_{Z_n}(u) = \mathbf{E}[e^{iuZ_n}]$, d'où

$\phi_{Z_n}(u) = \prod_{j=1}^n \mathbf{E}[e^{iu a_j^{(n)} U_j}]$ car les (U_i) sont indépendantes. Une v.a. X ayant des moments 1 et 2 admet un DL d'ordre 2 de sa fonction caractéristique : $\phi_X(t) = 1 + i \mathbf{E}[X] t - \frac{t^2}{2} \mathbf{E}[X^2] + o(t^2)$ quand $t \rightarrow 0$. Comme $\max_{1 \leq j \leq n} |a_j^{(n)}| \xrightarrow[n \rightarrow +\infty]{} 0$, $u a_j^{(n)} \xrightarrow[\eta \rightarrow +\infty]{} 0$ pour $1 \leq j \leq n$. $\forall u \in \mathbb{R}$,

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

en utilisant les hypothèses sur les U_j . Comme **heuristiquement**

$\prod_{j=1}^n \mathbf{E}[e^{iu a_j^{(n)} U_j}] = \exp \left(\sum_{j=1}^n \log (\mathbf{E}[e^{iu a_j^{(n)} U_j}]) \right)$, et $\log(1+x) = x + o(x)$ pour $x \rightarrow 0$, on obtient :

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

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



Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires
- Comportement asymptotique des estimateurs et tests

Premiers résultats

Propriété

Sous **(A0)-(A3)**, et si $(^t Z Z)^{-1} \xrightarrow[n \rightarrow +\infty]{} 0$, alors $\widehat{\theta}^{(n)} \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \theta^*$

Démonstration.

On a $\mathbb{E}[\widehat{\theta}_i^{(n)}] = \theta_i^*$ et $\text{var}(\widehat{\theta}_i^{(n)}) = \sigma^*{}^2 [(^t Z Z)^{-1}]_{i+1, i+1} \xrightarrow[n \rightarrow +\infty]{} 0$. Donc avec l'inégalité de Bienaym -Tchebychev, $\widehat{\theta}_i^{(n)} \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \theta_i^*$ pour tout $i \in \{0, \dots, p\}$. D'o  le r sultat puisque p fix . □

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 $\widehat{\theta}_0^{(n)} \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \theta_0^*$ 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 $\hat{\sigma}_n^2 \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \sigma^{*2}$.

Démonstration.

On a $\hat{\sigma}_n^2 = \frac{1}{n-(p+1)} \|P_{[Z]^\perp} \varepsilon\|^2$. Comme un projeté d'un vecteur a toujours une norme inférieure à celle de ce vecteur, $U_n = \frac{1}{n-(p+1)} (\|\varepsilon\|^2 - \|P_{[Z]^\perp} \varepsilon\|^2) \geq 0$. De plus

$\mathbf{E}[U_n] = \frac{1}{n-(p+1)} \mathbf{E}[\|P_{[Z]} \varepsilon\|^2] = \sigma^{*2} \frac{p+1}{n-(p+1)} \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} 0$. D'après l'inégalité de Markov, on en déduit que $U_n \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} 0$. Enfin, par la loi forte des grands nombres $\frac{1}{n} \|\varepsilon\|^2 \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \sigma^{*2}$, donc par Slutsky, comme p fixé, $\frac{1}{n-(p+1)} \|\varepsilon\|^2 \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \sigma^{*2}$. Finalement, $\hat{\sigma}_n^2 = \frac{1}{n-(p+1)} \|\varepsilon\|^2 - U_n$ on en déduit donc que $\hat{\sigma}_n^2 \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \sigma^{*2}$. □

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

Normalité asymptotique

Propriété

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

Démonstration.

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

Remarque : On a donc aussi $\frac{(^t Z Z)^{1/2}}{\widehat{\sigma}_n} (\widehat{\theta}^{(n)} - \theta^*) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}_{p+1}(0, I_{p+1})$

Question : Généralisable si on remplace **(A2)-(A4)** par **(A23)** ?

Normalité asymptotique (2)

Propriété

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

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

Démonstration.

On reprend la preuve précédente. On a par le TLC classique

$$\sqrt{n} \left(\frac{1}{n} \|\varepsilon\|^2 - \sigma^{*2} \right) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, \mu_4^* - \sigma^{*4}). \text{ Par Slutsky, on en déduit pareillement que}$$

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

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

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

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

Proposition

Sous **(A0)-(A1)-(A23)**, et si $\max_{1 \leq i \leq n} |(Z (t Z Z)^{-1} t Z)_{ii}| \xrightarrow{n \rightarrow +\infty} 0$, alors $(\hat{Y}^{(n)})_{n \in \mathbb{N}} = Z \hat{\theta}^{(n)}$ 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 :

$$\frac{{}^t C (\widehat{Y}^{(n)} - Z \theta^*)}{\sqrt{\text{var}({}^t C \widehat{Y}^{(n)})}} = \frac{{}^t C P_{[Z]} \varepsilon}{\sigma \sqrt{{}^t C P_{[Z]} C}} = {}^t D \varepsilon',$$

avec $\varepsilon' = \frac{\varepsilon}{\sigma}$ et $D = \frac{P_{[Z]} C}{\sqrt{{}^t C P_{[Z]} C}} = (D_i)_{1 \leq i \leq n}$. On peut appliquer le Théorème de Lindeberg ((ε'_i)

est une suite de v.a.i.i.d. telle que $\mathbb{E}[\varepsilon'_i] = 0$ et $\text{var}(\varepsilon'_i) = 1$ et $\sum_{i=1}^n D_i^2 = 1$ par construction) dès que l'on montre que $\max_{1 \leq i \leq n} |D_i| \xrightarrow{n \rightarrow +\infty} 0$. Pour cela, on peut écrire que $D = P_{[Z]} D$, d'où :

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

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 $\max_{1 \leq i \leq n} |(Z ({}^t Z Z)^{-1} {}^t Z)_{ii}| \xrightarrow[n \rightarrow +\infty]{} 0$, alors

$$({}^t Z Z)^{1/2} (\widehat{\theta}^{(n)} - \theta^*) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, \sigma^{*2} I_{p+1})$$

Démonstration.

On veut montrer que pour tout $C \in \mathbb{R}^{p+1}$, alors ${}^t C \widehat{\theta}^{(n)}$ est asymptotiquement gaussien. Mais $\widehat{\theta}^{(n)} = ({}^t Z Z)^{-1} {}^t Z \widehat{Y}^{(n)}$. Or pour tout $C' \in \mathbb{R}^n$, alors ${}^t C' \widehat{Y}^{(n)}$ est asymptotiquement gaussien. Donc en particulier pour ${}^t C' = {}^t C ({}^t Z Z)^{-1} {}^t Z$. 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 $\frac{1}{n} \chi^2(n) \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} 1$, d'où $t(n) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1)$ et $F(p_0, n) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \frac{1}{p_0} \chi^2(p_0)$

Démonstration.

On a $\frac{1}{n} \chi^2(n) \stackrel{\mathcal{L}}{\approx} \frac{1}{n} \sum_{i=1}^n Z_i^2$ où $Z = (Z_i) \stackrel{\mathcal{L}}{\approx} \mathcal{N}(0, I_n)$. Avec la loi forte des grands nombres,

$\frac{1}{n} \sum_{i=1}^n Z_i^2 \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \mathbb{E}[Z_0^2] = 1$. Ensuite, on utilise $t(n) \stackrel{\mathcal{L}}{\approx} \frac{\mathcal{N}(0, 1)}{\sqrt{\frac{1}{n} \chi^2(n)}}$ et $F(p_0, n) \stackrel{\mathcal{L}}{\approx} \frac{\frac{1}{p_0} \chi^2(p_0)}{\sqrt{\frac{1}{n} \chi^2(n)}}$ □

Corollaire

Sous **(A0)-(A1)-(A23)**, et si $\max_{1 \leq i \leq n} |(Z ({}^t Z Z)^{-1} {}^t Z)_{ii}| \xrightarrow[n \rightarrow +\infty]{} 0$, 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 $\widehat{T}^{(n)} = \frac{1}{\sqrt{\sigma^2}} ({}^t C ({}^t Z Z)^{-1} C)^{-1/2} {}^t C \widehat{\theta}^{(n)}$, comme $\widehat{\theta}^{(n)}$ est asymptotiquement gaussien ${}^t C \widehat{\theta}^{(n)}$ l'est également. Et $\widehat{\sigma}^2 \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \sigma^{*2}$. D'où le résultat.

On a $\widehat{F}^{(n)} = \frac{\frac{1}{q} {}^t \widehat{\theta}^{(n)} C ({}^t C ({}^t Z Z)^{-1} C)^{-1} {}^t C \widehat{\theta}^{(n)}}{\widehat{\sigma}^2}$. Comme $\widehat{\theta}^{(n)}$ est asymptotiquement gaussien la loi asymptotique de $\frac{1}{q} {}^t \widehat{\theta}^{(n)} C ({}^t C ({}^t Z Z)^{-1} C)^{-1} {}^t C \widehat{\theta}^{(n)}$ est la même que celle obtenue dans le cas gaussien, donc $\frac{\sigma^2}{q} \chi^2(q)$. Et comme précédemment $\widehat{\sigma}^2 \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \sigma^{*2}$. 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

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires
- Comportement asymptotique des estimateurs et tests

Le problème de la sélection d'un modèle linéaire

Cadre : On a observé Y_1, \dots, Y_n , que l'on veut modéliser par des variables exogènes $Z^{(1)}, \dots, Z^{(p)}$ 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 !

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

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

$\implies m \in \mathcal{P}(\{1, \dots, p\})$ 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 : $\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$
- Parfois la famille hiérarchique : $\mathcal{M} = \{\emptyset, \{1\}, \{1, 2\}, \dots, \{1, 2, \dots, p\}\}$

\implies Régression polynomiale par exemple

Nouvelle formalisation (2)

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

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

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 $m = m_p = \{1, \dots, p\}$, 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 $m^* \not\subset m$, 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 :

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

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 $\widehat{Y}_{(m)} = P_{[Z_{(m)}]} Y = \mu_{(m)}^* + P_{[Z_{(m)}]} \varepsilon$, d'où

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

CP de Mallows : minimisation du risque quadratique (2)

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

D'où

$$\begin{aligned}R(m, m^*) &= (|m| + 1) \sigma_*^2 + \mathbb{E} \left[\|Y - \hat{Y}_{(m)}\|^2 \right] - (n - (|m| + 1)) \cdot \\ \implies \frac{R(m, m^*)}{n \sigma_*^2} &= \frac{2(|m| + 1)}{n} - 1 + \frac{1}{n \sigma_*^2} \mathbb{E} \left[\|Y - \hat{Y}_{(m)}\|^2 \right] \\ \implies \frac{R(m, m^*)}{n \sigma_*^2} &\simeq \frac{2(|m| + 1)}{n} - 1 + \frac{\tilde{\sigma}_{(m)}^2}{\tilde{\sigma}_{(m_p)}^2}\end{aligned}$$

en estimant $\left\{ \begin{array}{l} \frac{1}{n} \mathbb{E} \left[\|Y - \hat{Y}_{(m)}\|^2 \right] \\ \sigma_*^2 \end{array} \right\}$ par $\left\{ \begin{array}{l} \tilde{\sigma}_{(m)}^2 = \frac{1}{n} \|Y - \hat{Y}_{(m)}\|^2 \\ \tilde{\sigma}_{(m_p)}^2 \end{array} \right\}$

CP de Mallows : minimisation du risque quadratique (3)

Définition

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

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

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

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

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

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 $n - |m| - 1 \implies$ espérance du numérateur est σ^2 , ne dépend plus de m . D'où :

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

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^*

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

⇒ Minimiser la dissemblance de Kullback revient à celle du critère *AIC* (pour Akaike Information Criterion, 1973), tel que

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

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

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

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

puisque 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 $\mathbb{P}(Y | m)$

Après calcul, pour n grand

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

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

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

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

Les différents critères

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

- $\widehat{Cp}(m) = \frac{\|Y - \widehat{Y}_m\|^2}{\|Y - \widehat{Y}_{(m_p)}\|^2} + 2 \frac{(|m| + 1)}{n}$
- $\widehat{R}_{Aju}^2(m) = 1 - \frac{n - 1}{n - |m| - 1} \frac{\|Y - \widehat{Y}_{(m)}\|^2}{\|Y - \bar{Y}\|^2}$
- $\widehat{AIC}(m) = n \log (\|Y - \widehat{Y}_{(m)}\|^2) + 2(|m| + 1)$
- $\widehat{BIC}(m) = n \log (\|Y - \widehat{Y}_{(m)}\|^2) + \log n (|m| + 1).$

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

Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires
- Comportement asymptotique des estimateurs et tests

Choix entre 2 modèles

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

Remarque : Si $|m_1| = |m_2| \implies$ 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 :

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

Probabilité de préférer un sur-modèle

Proposition

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

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

$$\mathbb{P} \left[\widehat{R_{Aju}^2}(m^*) \leq \widehat{R_{Aju}^2}(m_2) \right] \xrightarrow{n \rightarrow +\infty} \mathbb{P} \left[\chi^2(|m_2| - |m^*|) \geq (|m_2| - |m^*|) \right]$$

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

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

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

Probabilité de préférer un sur-modèle (2)

Démonstration.

Sous des hypothèses, $\widehat{F}(m^*, m_2) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \frac{1}{|m_2| - |m^*|} \chi^2(|m_2| - |m^*|)$. De plus

$\frac{\|Y - \widehat{Y}_{(m_p)}\|^2}{\|Y - \widehat{Y}_{(m_2)}\|^2} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} 1$. 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 $\widehat{F}(m^*, m_2) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \frac{\chi^2(|m_2| - |m^*|)}{|m_2| - |m^*|}$, on a

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

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 :

$$\mathbb{P}(m^* \subset \widehat{m}_{BIC}, m^* \neq \widehat{m}_{BIC}) \xrightarrow{n \rightarrow +\infty} 0$$

Démonstration.

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

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

Conclusion : le critère BIC évite asymptotiquement de sur-ajuster

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

$$\implies \text{critère GIC : } \widehat{\text{GIC}}(m) = n \log (\|Y - \widehat{Y}_{(m)}\|^2) + c_n |m|.$$

Probabilité de préférer un faux-modèle

Proposition

Sous les mêmes hypothèses et avec $(Z_{(m_p)})_n$ vérifiant

$\frac{1}{d(n)} {}^t Z_{(m_p)} Z_{(m_p)}$ $\xrightarrow[n \rightarrow +\infty]{} M$, où $d(n) \log^{-1}(n) \xrightarrow[n \rightarrow +\infty]{} \infty$ et M matrice définie positive. Alors pour Crit=Cp, R_{Aju}^2 , AIC ou BIC,

$$\mathbb{P} ["Crit \text{ choisit un faux-modèle}"] \xrightarrow[n \rightarrow +\infty]{} 0.$$

Démonstration.

Preuve trop complexe, voir le livre... □

Probabilité de préférer un faux-modèle (2)

Conséquences : Sous les mêmes hypothèses :

- ① $\inf_{n \in \mathbb{N}} \mathbb{P}(m^* \subset \hat{m}_{Crit}, m^* \neq \hat{m}_{Crit}) \geq p_{Crit} > 0$ pour $Crit = \begin{cases} Cp \\ R^2_{Aju} \\ AIC \end{cases}$
- ② $\mathbb{P}(\hat{m}_{BIC} = m^*) \xrightarrow{n \rightarrow +\infty} 1$

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 :

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

Il est clair que :

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

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 $R_n^{GIC} \xrightarrow[n \rightarrow +\infty]{} R_n(m^) = (|m^*| + 1) \sigma_*^2$.*

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 $R_n(m) = \mathbb{E}_{m^*} \left(\|\mu^* - \widehat{Y}_{(m)}\|^2 \mathbf{1}_{\{\widehat{m}=m\}} \right)$, d'où avec $p_n(m) = \mathbb{P}(\widehat{m} = m)$,

$$\begin{aligned} R_n(m) &= \mathbb{E}_{m^*} \left(\|\mu^* - \mu_{(m)}^*\|^2 \mathbf{1}_{\{\widehat{m}=m\}} + \|\mu_{(m)}^* - \widehat{Y}_{(m)}\|^2 \mathbf{1}_{\{\widehat{m}=m\}} \right) \\ &= \|P_{[X^{(m)}]^\perp} X^{(m_p)} \theta^{(m_p)}\|^2 p_n(m) + \mathbb{E}_{m^*} \left(\|P_{[X^{(m)}]} \varepsilon\|^2 \mathbf{1}_{\{\widehat{m}=m\}} \right) \\ &= J_1(m) + J_2(m). \end{aligned}$$

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 $n \rightarrow \infty$ 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 $J_1(m) \xrightarrow[n \rightarrow +\infty]{} 0$ car on a supposé $\log n = o(d_n)$. □

Etude asymptotique du risque quadratique (3)

Démonstration.

Si $m = m^*$, alors $J_2(m) \xrightarrow{n \rightarrow +\infty} (|m^*| + 1) \sigma_*^2$ car $p_n(m^*) \xrightarrow{n \rightarrow +\infty} 1$. 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 $m \neq m^*$, $p_n(m) \xrightarrow{n \rightarrow +\infty} 0$, on en déduit que $R_n(m) \xrightarrow{n \rightarrow +\infty} 0$ tandis que $R_n(m) \xrightarrow{n \rightarrow +\infty} (|m^*| + 1) \sigma_*^2$. D'où le résultat final. □

Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires
- Comportement asymptotique des estimateurs et tests

Les hypothèses "classiques"

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

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

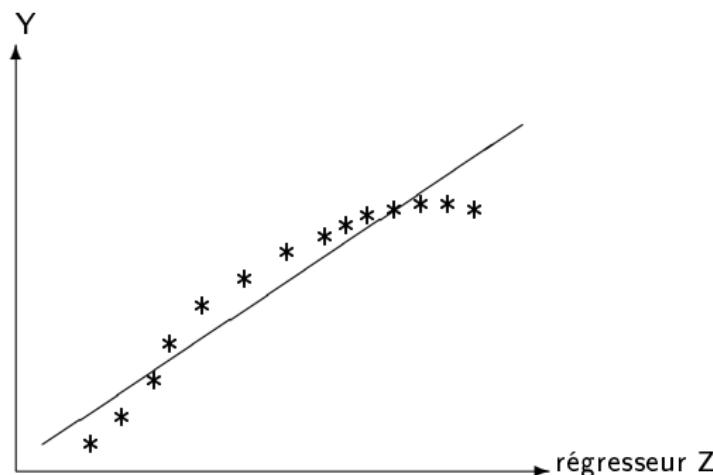
Même si les tests de Student et Fisher sont positifs, R^2 proche de 1...

Diagnostics graphiques \implies 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 :



⇒ On voit une courbure de la "vraie" courbe de régression de Y , le modèle n'est pas adéquat ⇒ (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.

⇒ On travaille avec les $\hat{\varepsilon}_i = Y_i - \hat{\theta}_0 - \hat{\theta}_1 Z_i^{(1)} - \dots - \hat{\theta}_p Z_i^{(p)}$

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

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

⇒ $\text{cov}(\hat{Y}_i, \hat{\varepsilon}_i) = 0$ pour tout $i = 1, \dots, n$ sous (A1).

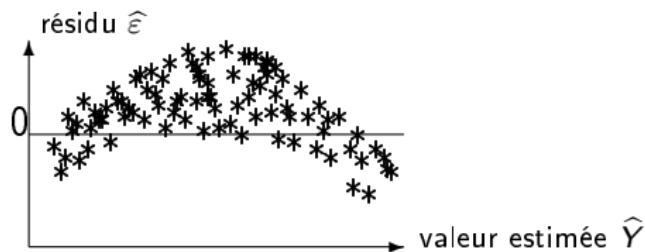
⇒ **Graphique des résidus** ($\hat{\varepsilon}_i$) en fonction des **valeurs prédictes** (\hat{Y}_i).

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 :



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

Solutions possibles

- ➊ On n'a pas considéré certaines variables explicatives et on les rajoute !
- ➋ On transforme les régresseurs $Z^{(1)}, \dots, Z^{(p)}$ par des fonctions de ces régresseurs :
 - ▶ Utilisation de polynômes de ces variables ;
 - ▶ Utilisation de fonctions de ces variables : avec paramètres (MC non linéaires) ou sans paramètres (régression non paramétrique).

Idée : on utilise le critère BIC pour valider le nouveau modèle
- ➌ 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

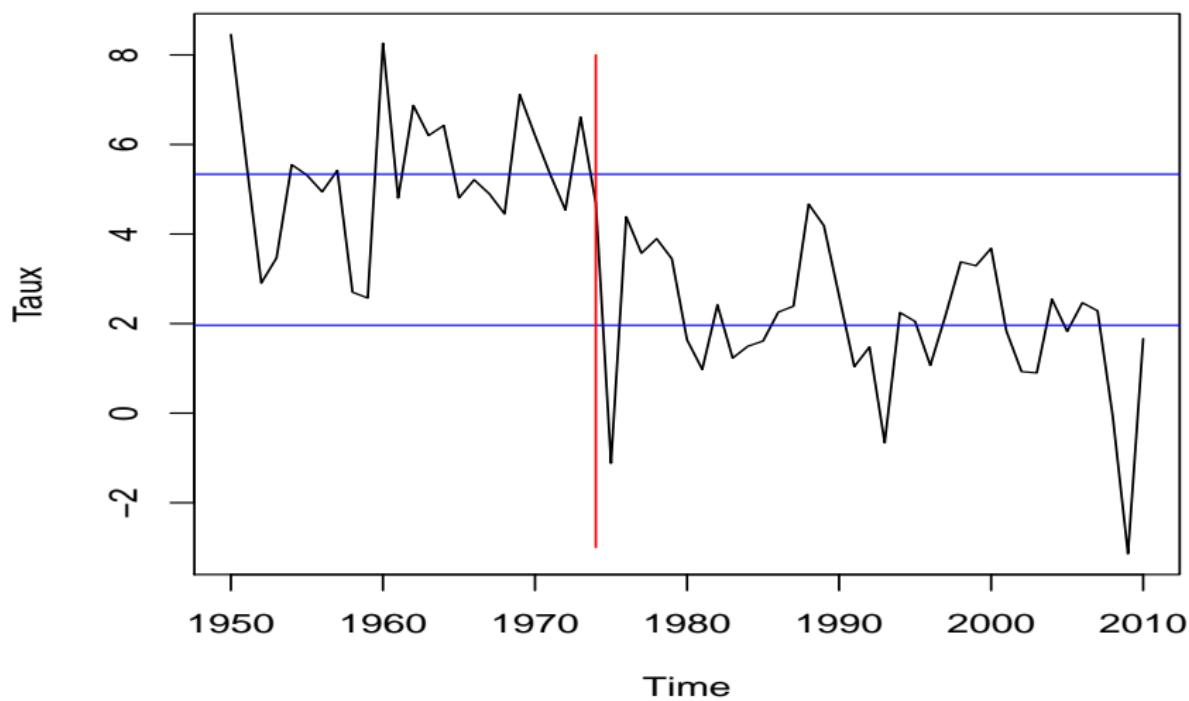


Figure – Taux de croissance annuel en France de 1950 à 2010

Détection de rupture : Températures annuelles du Globe

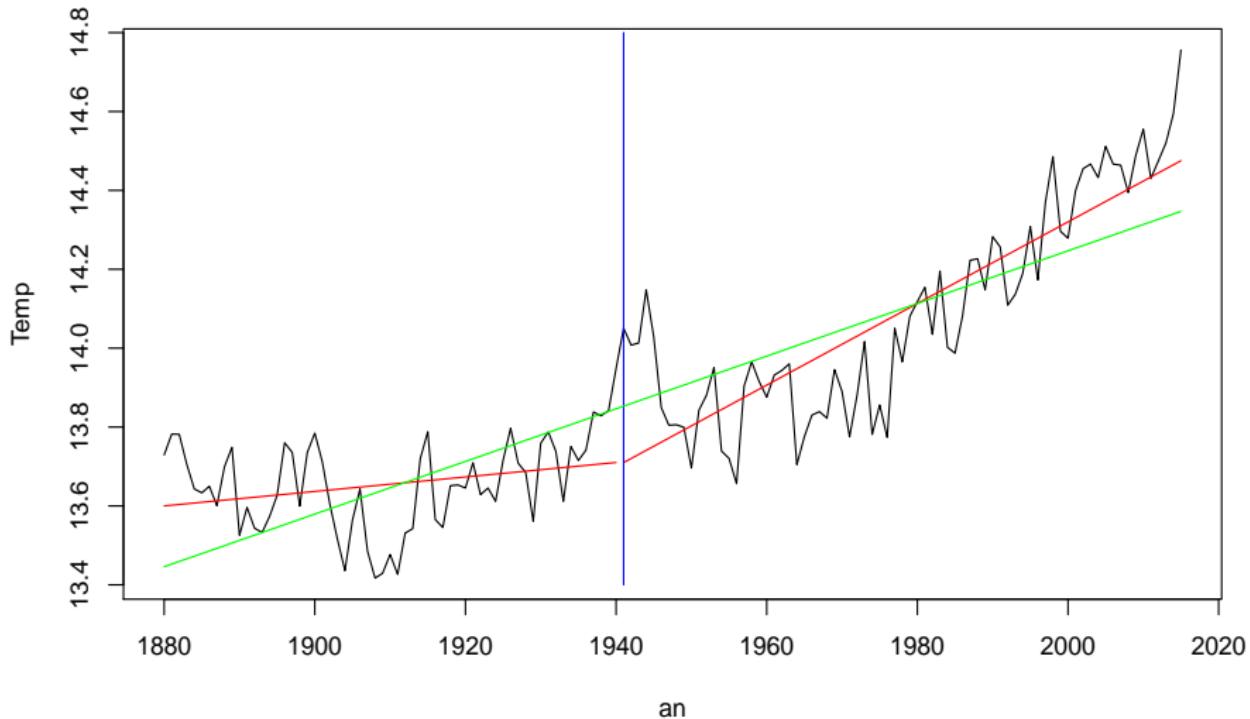


Figure – Détection de rupture linéaire pour les températures annuelles du globe
⇒ $R^2 \simeq 0.82$ et $\hat{a}_2 \simeq 0.01$ ($pval \simeq 2 \cdot 10^{-16}$) et $\hat{C} \simeq 55$: **Rupture !!**

Détection de ruptures dans des modèles linéaires simples

On suppose (X_1, \dots, 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 :

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

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

Détection de rupture

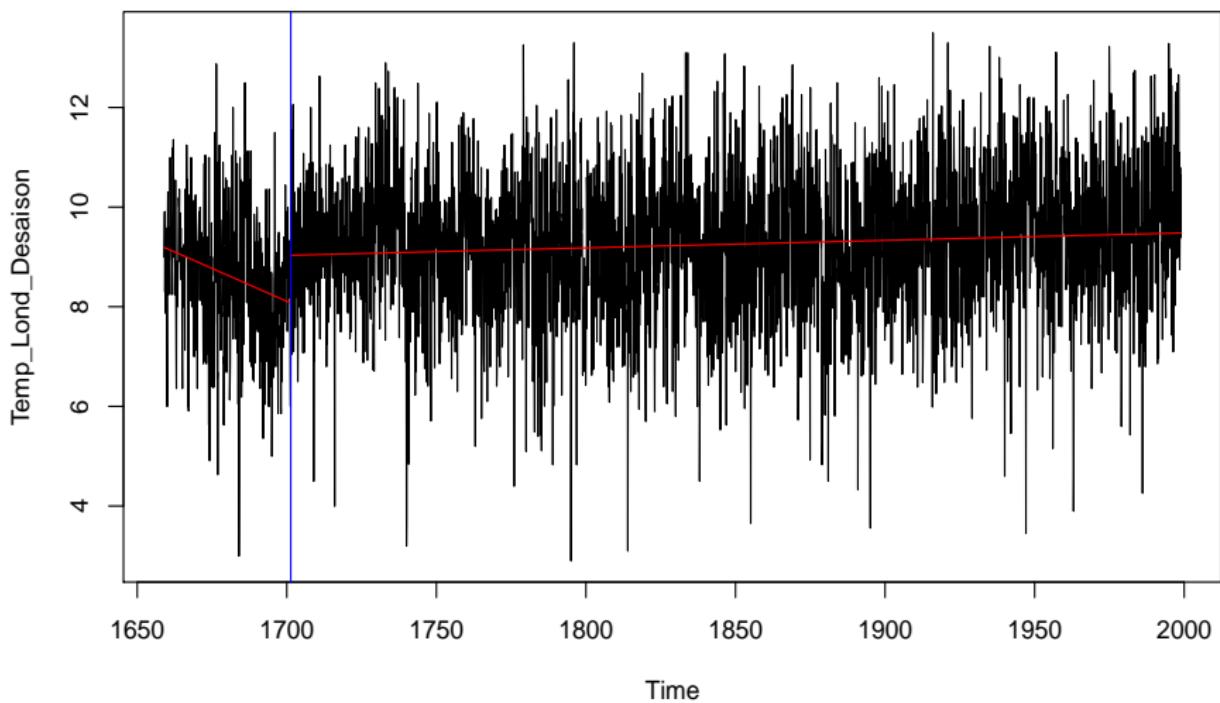


Figure – Détection de rupture linéaire pour les températures mensuelles de Londres désaisonnalisée $\implies R^2 \simeq 0.036$ et $\hat{a}_2 \simeq 0.0015$

Détection de rupture (suite)

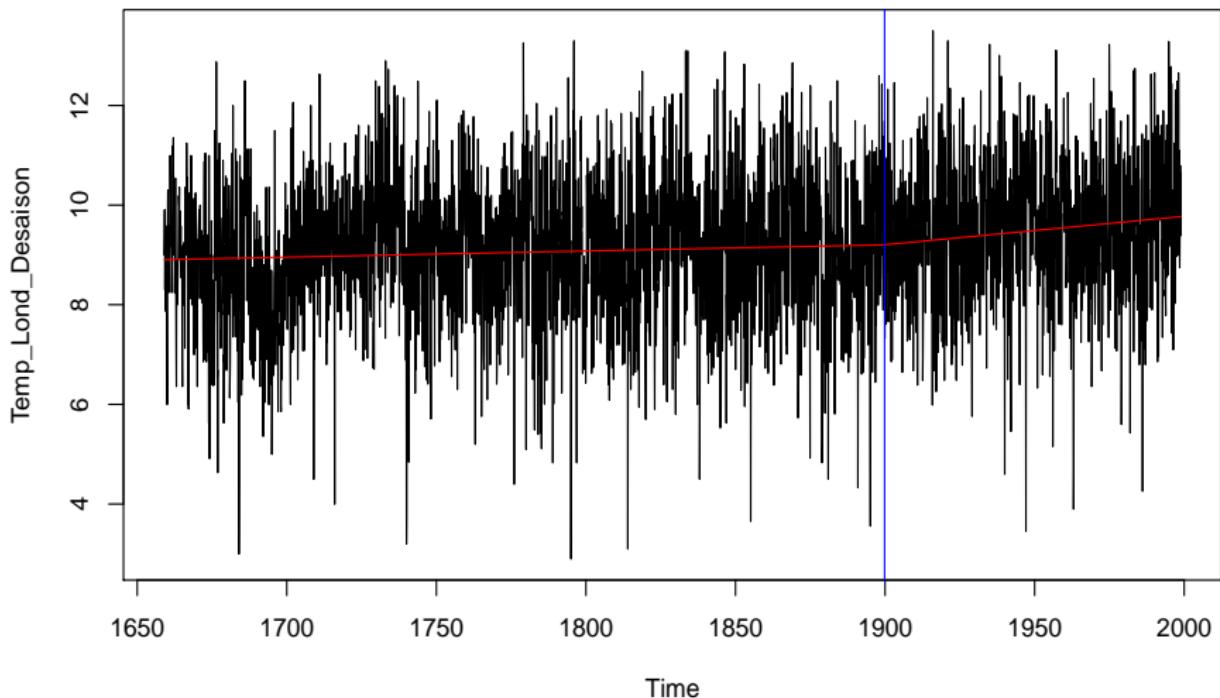


Figure – Détection de rupture dans le "slope" pour les températures mensuelles de Londres désaisonnalisée $\Rightarrow R^2 \simeq 0.026$ et $\hat{b}_2 \simeq 0.0056$

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 $n \rightarrow \infty$, pour toute suite (a_n) telle que $a_n \xrightarrow[n \rightarrow +\infty]{} \infty$

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

$$\implies |\widehat{\tau}_n - \tau^*| = O_{\mathbb{P}}(a_n).$$

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](#) :

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

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

⇒ 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 :

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

où ${}^tY = ({}^tY^{(1)}, {}^tY^{(2)})$. 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 :

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

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

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

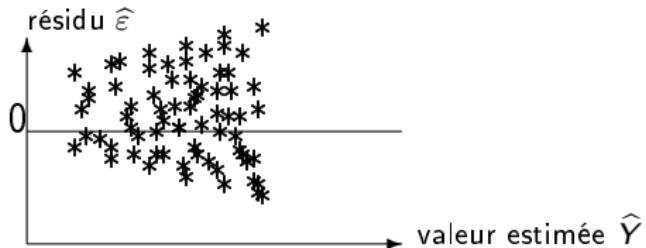
D'où $\hat{j} = \operatorname{Arg} \min_j SC_1(j)$.

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

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

Remise en question de (A2) : Hétéroscédasticité

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



⇒ variance des résidus semble inhomogène
mais sous (A1-A3) $\text{var}(\widehat{\varepsilon}_i) = \sigma^2(1 - p_{ii})$ non constante

⇒ Tracé des résidus dit "Studentisés"

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

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

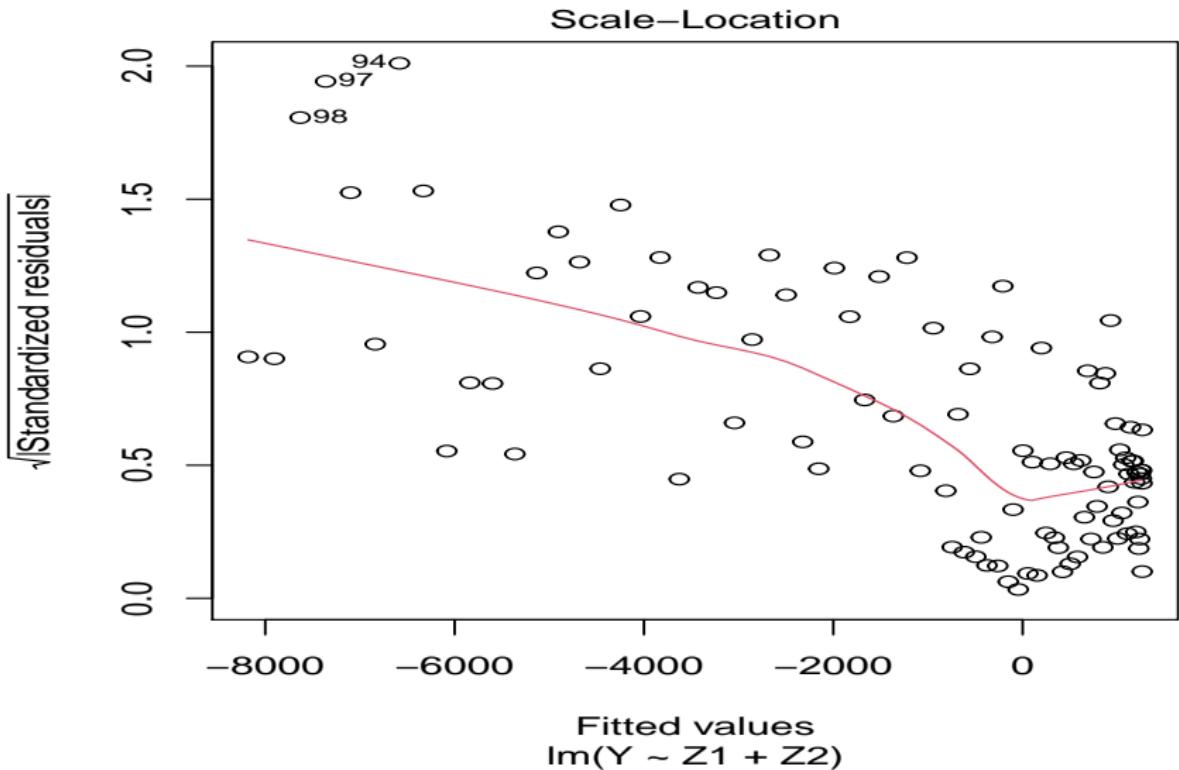


Figure – Exemple d'hétéroscédasticité

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 :

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

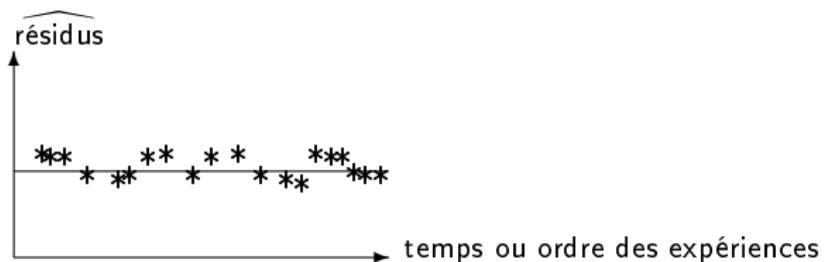
où λ est un réel a priori inconnu.

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

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

Remise en question de (A3) : corrélation du bruit

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



⇒ 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 \hat{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

$$\frac{\hat{N} - \frac{1}{2}(n+1)}{\frac{1}{2}\sqrt{n-1}} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

- **Test portemanteau** : Si $p_{ij} \xrightarrow[n \rightarrow +\infty]{} 0$ suffisamment vite,

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

Utilisation des moindres carrés généralisés

Définition

On suppose le modèle linéaire général

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

On suppose $\Sigma = \mathbb{E}[\varepsilon^t \varepsilon]$ 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 , $\langle Z_1, Z_2 \rangle = {}^tZ_1\Sigma^{-1}Z_2$.

L'estimateur $\widehat{\theta}_G$ de θ par MCG minimise $\|Y - X \cdot \theta\|_{\Sigma}$ pour $\theta \in \mathbb{R}^{p+1}$ et

$$\widehat{\theta}_G = ({}^tX\Sigma^{-1}X)^{-1}{}^tX\Sigma^{-1}Y.$$

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

$\hat{\theta}_G$ 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 !

⇒ Estimer Σ à partir des résidus $\hat{\varepsilon}$ obtenus par MCO.

⇒ Si $\hat{\Sigma} \rightarrow \Sigma$ avec $\hat{\Sigma}$ inversible, on approche $\hat{\theta}_G$ par estimateur $\tilde{\theta}_G$ appelé **estimateur par moindres carrés pseudo-généralisés** défini par

$$\tilde{\theta}_G = ({}^t X \hat{\Sigma}^{-1} X)^{-1} {}^t X \hat{\Sigma}^{-1} Y.$$

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 $\max_{1 \leq i \leq n} |p_{ii}| \xrightarrow[n \rightarrow +\infty]{} 0$, $\widehat{\theta} \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \theta$ et $\widehat{\varepsilon}_i \xrightarrow[n \rightarrow +\infty]{\mathbf{L}^2} \varepsilon_i$ pour $i \in \mathbb{N}$.
- ② Estimation des paramètres de l'ARMA(p, q) par MV à partir de $(\widehat{\varepsilon}_1, \dots, \widehat{\varepsilon}_n)$. Si $\alpha = (a_1, \dots, a_p, b_1, \dots, b_q, \sigma^2)$ les paramètres, on montre que $\widehat{\alpha} \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} \alpha$.
- 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 $\{t_j, t_j + 1, \dots, t_{j+1} - 1\}$
- ② On utilise les MCG avec $\widehat{\Sigma} = \text{diag}(\widehat{\sigma}_1^2, \dots, \widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \dots, \widehat{\sigma}_m^2)$
- ③ Estimation des t_j ?

Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires
- Comportement asymptotique des estimateurs et tests

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 : Intervention prix et km dans base de véhicules d'occasion

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

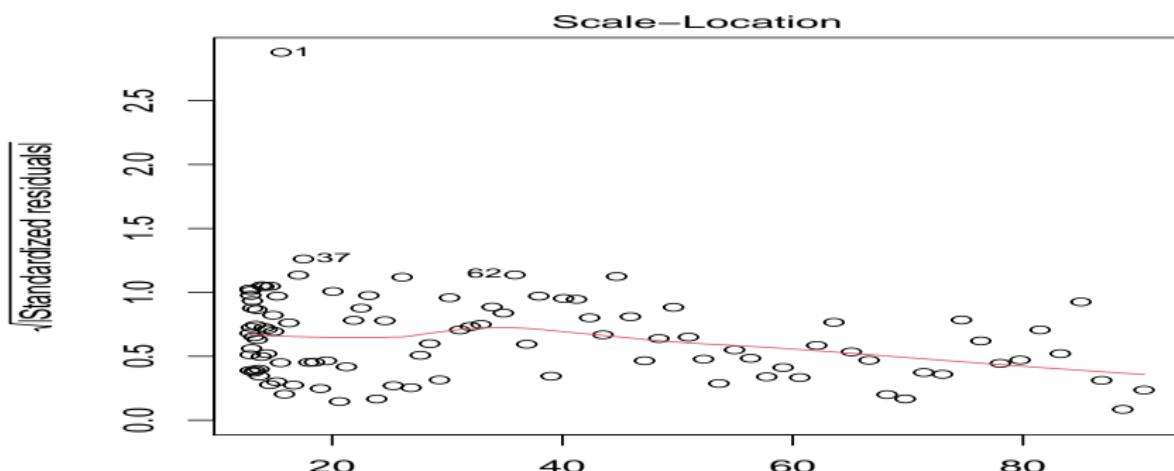
Exemple : Véhicules de taxis dans base de véhicules d'occasion

Utilisation des résidus studentisés

Tracé des résidus dit "Studentisés"

$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{\sqrt{1 - p_{ii}} \hat{\sigma}^{(i)}} \text{ où } \hat{\sigma}^{(i)} \text{ obtenue à partir de } (Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n).$$

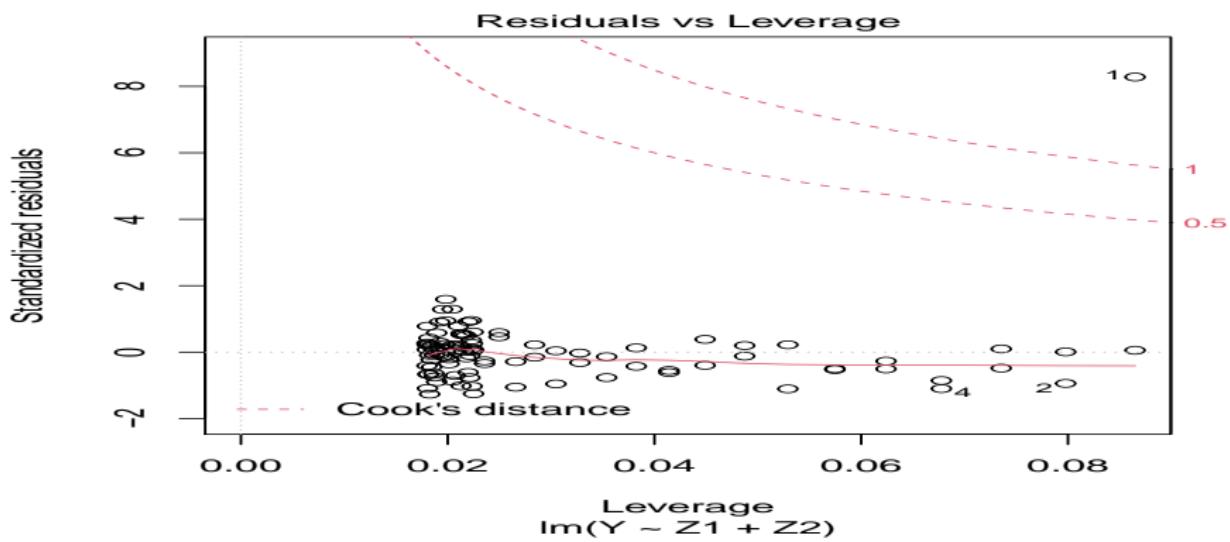
$\implies \text{Sous (A1-4), } \tilde{\varepsilon}_i \stackrel{\mathcal{L}}{\sim} t(n-p-2)$



Utilisation de la distance de Cook

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

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



Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires
- Comportement asymptotique des estimateurs et tests

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, \dots, Y_n) observés et variables $X^{(1)}, \dots, X^{(p)}$ 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

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

L'idée sera ainsi d'écrire que :

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

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

On en déduit donc que

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

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

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

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 $\varepsilon \stackrel{\mathcal{L}}{\sim} \mathcal{N}(0, I_n)$, Z étant une variable dite latente, et $Y = \mathbb{I}_{Z \geq 0}$. 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 $\varepsilon_i \stackrel{\mathcal{L}}{\sim} \mathcal{L}$ (loi logistique), Z étant une variable dite latente, et $Y = \mathbb{I}_{Z \geq 0}$. 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 :

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

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 $P(Y_{n+1} = 1)$:

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

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 $p_i(\theta) = g^{-1}((X\theta)_i)$ et la vraisemblance s'écrit :

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

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

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

Approximation par Newton-Raphson

⇒ $\hat{\theta}$ non explicite !

⇒ 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 :

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

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

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

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).

Théorème

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

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

Remarques :

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

⇒ Intervalles de confiance et tests.

Validation et tests

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

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

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

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

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

⇒ 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 :

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

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

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

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

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 :

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

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

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

On définit ainsi :

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

Théorème

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

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

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 $\widehat{m} = \operatorname{Argmin}_{m \in \mathcal{M}} BIC(m)$.

Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires
- Comportement asymptotique des estimateurs et tests

Régression polytomique

Y variable qualitative pouvant prendre J modalités

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

→ Estimer en fonction de X la probabilité

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

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

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

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

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

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

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

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

$$\hat{Y}_i = \operatorname{Argmax}_{j=1,\dots,J} \hat{p}_{i,j}.$$

Régression polytomique (suite)

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

⇒ Estimateur par maximum de vraisemblance.

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

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

Cela correspond à la vraisemblance d'une loi multinomiale.

⇒ Algorithme de Newton-Raphson pour approcher $(\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(J-1)})$

⇒ Tests du rapport de vraisemblance, de Wald, sélection de modèle.

Régression polytomique pour données ordonnées

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

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

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

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

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

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

\implies Estimation numérique + tests et sélection de modèle.

Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires
- Comportement asymptotique des estimateurs et tests

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 $X_i^{(j)}$.

⇒ Modèle sous une forme plus générale :

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

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.

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

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

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

où g est inconnue et appartient à un espace fonctionnel (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 : $g_\theta(x) = \frac{\theta_1}{1 + \left(\frac{x - \theta_2}{\theta_3}\right)^2}.$
- Un modèle de type loi de puissance : $g_\theta(x) = \theta_1 + \theta_2 x^{\theta_3}.$

Méthode

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

Il s'agira donc de minimiser (en θ) :

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

$$\implies \hat{\theta} = \operatorname{Argmin}_{\theta \in \mathbb{R}^d} S(\theta) : \text{ (MV dans le cas gaussien)}$$

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

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

\implies Bijectivité de $\theta \mapsto g_\theta$ clé de l'unicité de $\hat{\theta}$.

Minimisation

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

⇒ Minimisation d'une fonction à plusieurs variables.

Solution des équations normales :

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

Ceci s'écrit matriciellement :

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

avec $G_\theta(X) = (g_\theta(X_i))_i$, $Y = (Y_i)_i$ 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

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

Minimisation (suite)

Deux possibilités :

- ① On utilise une méthode de résolution de type Newton-Raphson ;
- ② On "linéarise" le système avec suite $(\tilde{\theta}^k)_k$ et formule de Taylor

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

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

Les équations normales s'approchent alors par les équations :

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

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

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

$$\implies \Delta^{k+1} = ({}^t \dot{G}_{\tilde{\theta}^k}(X) \dot{G}_{\tilde{\theta}^k}(X))^{-1} {}^t \dot{G}_{\tilde{\theta}^k}(X) Z^k$$

Plan du cours

1 Rappels sur le modèle linéaire

- Le cadre général du modèle linéaire
- Les hypothèses et leurs conséquences

2 Comportement asymptotique des statistiques

- Quelques théorèmes limite
- Conséquences sur les estimateurs et tests de la régression linéaire

3 Sélection de modèle en régression

- Critères de sélection de modèles
- Comportement asymptotique des modèles choisis

4 Les possibles problèmes et leurs solutions

- Faux modèle, hétéroscédasticité, dépendance
- Points aberrants

5 Régression logistique et polytômique

- Régression logistique
- Régression polytômique

6 Moindres carrés non linéaires

- Le cadre des moindres carrés non linéaires

- Comportement asymptotique des estimateurs et tests

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

Alors :

$$\sqrt{n} (\widehat{\theta} - \theta^*) \xrightarrow[r \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2 I(\theta^*)^{-1}).$$

Idée de preuve

On montre que :

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

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

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

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

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)