Modèle linéaire et moindre carré

I Cadre et Hypothèses

- Variable d'intérêt $y \in \mathbb{R}$ (variable endogène)
  Variable explicative $x \in \mathbb{R}^p$ (variable exogène) observée
- But : prédire y en fonction de x
  + précisément on cherche $E(y|x)$
- Modèle linéaire $E(y|x) = x^* \beta_0$ avec $\beta_0 \in \mathbb{R}^p$

  Soit en posant $\varepsilon = y - E(y|x)$

  $y = x^* \beta + \varepsilon$ avec $E(\varepsilon|x) = 0$.

  Hypothèse forte de structure !
- Pb : inférence sur β_0 à partir de $(y_i = x_i^* \beta + \varepsilon_i)_{1 \le i \le n}$
- Hypothèse H_ε sur les résidus

$$H_\varepsilon = \begin{cases} E(\varepsilon_i | x_i) = 0 & (1) \\ \mathrm{Var}(\varepsilon_i) = \sigma_0^2 & (2) \\ \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ si } i \neq j & (3) \end{cases}$$

(1) résidus centrés conditionnellement à x (inexplicable)
(2) variance finie + (3) décorrélation (ε bruit blanc)

- $H_\varepsilon \Rightarrow$ $E(y_i) = x^* \beta$
  $\mathrm{Var}(y_i) = \sigma^2$
  $\mathrm{Cov}(y_i, y_j) = 0$ si $i \neq j$

  ⚠ conditionnel à x_i implicite.

Pb inférence sur β et σ_0 (les paramètres du modèle)

- Notation matricielle $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $X = \begin{pmatrix} x_1^* \\ \vdots \\ x_n^* \end{pmatrix}$ $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

Observation : $Y = X\beta + \varepsilon$

$H_\varepsilon \Rightarrow \quad E(\varepsilon) = 0 \qquad Cov(\varepsilon) = \sigma^2 I_n$

$E(Y) = X\beta \qquad Cov(Y) = \sigma^2 I_n$

Rq Si on note $X = \begin{pmatrix} x_1^t \\ x_n^t \end{pmatrix} = (X_1 \dots X_p)$

$$X\beta = \sum X_\ell \beta_\ell = \sum \beta_\ell X_\ell$$

- Identifiabilité si il existe une solution unique à $X\beta = X\beta_0$.
  si l'on veut un résultat uniforme en β
  - $\Rightarrow X$ est de rang $\geq p$ (le maximum possible $X \in \mathbb{R}^{n \times p}$)

  Rq : X de rang $p \Rightarrow$ Im X est un sev de $\mathbb{R}^n$ de dimension p.

- Si pas identifiabilité, pb mal posé estimer β
  $\Rightarrow$ dégt de paramétrisation (suppression de variables, régression) ...

- Matrice X a un caractère arbitraire : ordre des variables / ordre des observations / exemples

  $Y = X\beta + \varepsilon$ et $Y = \tilde{X}\tilde{\beta} + \varepsilon$

  ont 2 paramétrisations équivalentes si l'on peut passer de β_0 à $\tilde{\beta}_0$ de manière univoque.

  Ex $\tilde{X} = XZ^{-1}$ et $\tilde{\beta} = Z\beta$, avec Z inversible

  Même prédiction mais interprétation et valeurs peuvent être $\neq$

Ex : régression affine simple

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & \vdots \\ 1 & x_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Param équivalente $y_i = \beta'_0 + \beta_1 (x_i - \bar{x}) + \varepsilon_i$

$(\beta'_0 - \beta_1 \bar{x} = \beta_0)$

Régression affine multiple :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_{i,p} x_{i,p} + \varepsilon_i$$

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & & & \\ 1 & & & \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_1 \end{pmatrix}$$

$\Rightarrow$ $p+1$ paramètres $(+\sigma !)$

Analyse de la variance

$$y_{i,k} = m_i + \varepsilon_{i,k} \qquad \begin{matrix} 1 \le k \le n_i \\ 1 \le i \le p \end{matrix} \qquad (n_i > 0)$$ identifiable

$$X = \begin{pmatrix} 1 & & & \\ \vdots & & & \\ 1 & & & \\ & 1 & & \\ & \vdots & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ & & & \vdots \\ & & & 1 \end{pmatrix} \begin{matrix} \} n_1 \\ \\ \\ \} n_2 \\ \\ \\ \\ \} n_p \\ \\ \end{matrix} \qquad \beta = \begin{pmatrix} m_1 \\ \vdots \\ m_p \end{pmatrix}$$

p paramètres

Analyse de la covariance

$$y_{i,k} = a_i + b_i x_{i,k} + \varepsilon_i \qquad \begin{matrix} 1 \le k \le n_i \\ 1 \le i \le p \end{matrix}$$

$$X = \begin{pmatrix} 1 & & & x_{1,1} & & \\ \vdots & & & \vdots & & \\ 1 & & & x_{1,n_1} & & \\ & 1 & & & x_{2,1} & \\ & \vdots & & & \vdots & \\ & 1 & & & x_{2,n_2} & \\ & & \ddots & & & \\ & & & & & x_{p,1} \\ & & & & & \vdots \\ & & & & & x_{p,n_p} \end{pmatrix}$$

$$\beta = \begin{pmatrix} a_1 \\ \vdots \\ a_p \\ b_1 \\ \vdots \\ b_p \end{pmatrix} \qquad 2p \text{ paramètres}$$

Modèle de courbe : $y_i = \sum \alpha_k \beta_k(x_i) + \varepsilon_i$ avec β_k connue

$$X = \begin{pmatrix} \beta_1(x_1) & \cdots & \beta_p(x_1) \\ \vdots & & \vdots \\ \beta_1(x_n) & & \beta_p(x_n) \end{pmatrix} \qquad \beta = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} \qquad p \text{ paramètres}$$

II Moindre carrée (ordinaire) MCO

(least squares) Hyp. X de rang plein (identifiabilité)

Idée : Chercher $\tilde{\beta}$ qui minimise la somme des carrés résiduels

(SCR en français SSR en anglais)

$$SCR(\tilde{\beta}) = \|Y - X\tilde{\beta}\|^2 = \sum (y_i - x_i^* \tilde{\beta})^2$$

Terme d'attache aux données qui privilégie des prédictions proches des valeurs observées

Rq. $E(SCR(\beta)) = E(\|\varepsilon\|^2) = Tr(Cov(\varepsilon)) = \sigma^2 n$

. on aurait pu remplacer la norme 2 par une autre norme (ou une autre fonction) ($\|\ \|_1$ ou mélange $\|\ \|_1$ et $\|\ \|_2$ à la Huber) mais calcul facile et bonne propriété théorique

Propr 1) $\nabla SCR(\hat{\beta}) = -2X^*(Y - X\tilde{\beta})$

2) $\frac{\partial^2 SCR}{\partial \tilde{\beta}^2}(\hat{\beta}) = 2X^*X$

3) SCR est strictement convexe.

Preuve

$$SCR(\hat{\beta}) = \|Y - X\hat{\beta}\|^2$$

$$SCR(\hat{\beta}+h) = \|(Y - X\hat{\beta}) - Xh\|^2$$

$$= \|Y - X\hat{\beta}\|^2 - 2\langle Xh, Y - X\hat{\beta}\rangle + \|Xh\|^2$$

$$= \|Y - X\hat{\beta}\|^2 - 2\langle h, X^*(Y - X\hat{\beta})\rangle + \|Xh\|^2$$

$$\Rightarrow \nabla SCR(\hat{\beta}) = -2X^*(Y - X\hat{\beta})$$

$$\Rightarrow \frac{\partial^2 SCR}{\partial \hat{\beta}^2}(\hat{\beta}) = 2X^*X$$

or $\forall \alpha \in \mathbb{R}^p \quad \alpha^*(2X^*X)\alpha$

$$= 2(X\alpha)^* X\alpha = 2\|X\alpha\|_2^2 \geq 0$$

et même > 0 car X de rang plein
$\alpha \neq 0$

$\Rightarrow 2X^*X$ définie positive $\Rightarrow$ SCR strictement convexe

Cor : $\hat{\beta} = \underset{\tilde{\beta}}{\operatorname{argmin}}\ SCR(\tilde{\beta}) = \underset{\tilde{\beta}}{\operatorname{argmin}}\ \|Y - X\tilde{\beta}\|^2$

existe et est unique.

De plus $\hat{\beta} = (X^*X)^{-1} X^* Y$

Preuve : le premier point découle de la stricte convexité

. le second s'obtient en utilisant la condition d'ordre 1 du minimum

$$\nabla SCR(\hat{\beta}) = 0 = -2X^*(Y - X\hat{\beta})$$

$$\Rightarrow X^*X\hat{\beta} = X^*Y$$

or X^*X définie positive $\Rightarrow X^*X$ inversible

$$\Rightarrow \hat{\beta} = (X^*X)^{-1} X^* Y$$

Proposition Estimation de β par moindre carré

L'estimateur des moindres carrés $\hat{\beta} = \underset{\tilde{\beta}}{\operatorname{argmin}}\ SCR(\tilde{\beta})$ est donné

par $\hat{\beta} = (X^*X)^{-1} X^* Y$

et vérifie 1) $E(\hat{\beta}) = \beta$

2) $Cov(\hat{\beta}) = \sigma^2 (X^*X)^{-1}$

Preuve : . la première partie est une reduite du Th précédent

$$E(\hat{\beta}) = E\left((X^*X)^{-1} X^* Y\right)$$

$$= (X^*X)^{-1} X^* E(Y) = (X^*X)^{-1} X^* X \beta$$

$$= \beta$$

$$Cov(\hat{\beta}) = Var\left((X^*X)^{-1} X^* Y\right)$$

$$= (X^*X)^{-1} X^* Cov(Y) \left((X^*X)^{-1} X^*\right)^*$$

$$= (X^*X)^{-1} X^* \sigma^2 I_n \times (X^*X)^{-1}$$

$$= \sigma^2 (X^*X)^{-1}$$

Rq si $\tilde{\beta} = Z\beta$ alors par les moindres carrés $\hat{\beta} = Z\hat{\beta}$.

Prop (Gauss-Markov) Parmi les estimateurs linéaires sans biais de β, l'estimateur des moindres carrés $\hat{\beta}$ est celui de variance minimale (BLUE Best Unbiased Linear Estimator)

Preuve. $\beta^* = LY$

$E(\beta^*) = \beta \Rightarrow LX\beta = \beta$ Vrai $\forall \beta$

$\Rightarrow LX = Id$

$Cov(\beta^*) = L\,Cov(Y)\,L^* = \sigma^2 LL^*$

$= \sigma^2 (M+\Delta)(M+\Delta)^*$ où $M = (X^*X)^{-1}X^*$

et $\Delta = L - M$

$= \sigma^2 [MM^* + \Delta\Delta^* + M\Delta^* + \Delta M^*]$

or $\Delta M^* = \underbrace{(L-M)}_{0 \text{ car } L \text{ et } M \text{ sont inverses}} X (X^*X)^{-1}$

$= 0 \quad = M\Delta^*$

$Cov(\beta^*) = \sigma^2 (X^*X)^{-1} + \sigma^2 \underbrace{\Delta\Delta^*}_{\geq 0}$

$\Rightarrow Cov(\beta^*) \geq Cov(\hat{\beta})$

Prédiction : $\hat{Y} = X\hat{\beta}$ prédiction de $E(y|x)$ pour les x_i observés

$\hat{y} = x^*\hat{\beta}$ prédiction de $E(y|x)$ pour un nouvel x

Prop : $E(\hat{Y}) = X\beta$ $\quad Cov(\hat{Y}) = \sigma^2 X(X^*X)^{-1}X^*$

$E(\hat{y}) = x^*\beta$ $\quad Var(\hat{y}) = \sigma^2 x^*(X^*X)^{-1}x$

Preuve : immédiat !

Rq on peut montrer que $\hat{y} = x^*\hat{\beta}$ est optimal pour la variance parmi les estimateurs linéaires sans biais de $E(C^*X) = C^*X\beta = x^*\beta$

$Var(C^*Y) = \sigma^2 C^*C \geq \sigma^2 C^*X(X^*X)^{-1}X^*C = \sigma^2 x^*(X^*X)^{-1}x = Var(\hat{y})$

Interprétation géométrique $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2$

$\Rightarrow \hat{Y} = \underset{\hat{Y} = X\beta}{\operatorname{argmin}} \|Y - \hat{Y}\|^2$

$= \underset{\hat{Y} \in Im\,X}{\operatorname{argmin}} \|Y - \hat{Y}\|^2$

$\Rightarrow$ $\hat{Y}$ est la projection orthogonale de Y sur $Im\,X$

(lien avec l'espérance conditionnelle L^2 et la meilleure approximation linéaire)

$\hat{Y} = P_{Im\,X} Y$ où P est une matrice de projection $\perp$

Rq $\hat{Y} = X\beta = \underbrace{X(X^*X)^{-1}X^*}_{? P_{\mathrm{Im}X} ?} Y$

$H = X(X^*X)^{-1}X^*$ Hat Matrix = matrice chapeau = projecteur

- $H^* = H$
- $HH = X(X^*X)^{-1}X^* \times X(X^*X)^{-1}X^* = X(X^*X)^{-1}X^* = H$
- $H = X(X^*X)^{-1}X^* \Rightarrow \mathrm{Im}\,H \subset \mathrm{Im}\,X$
- $\forall z \in \mathrm{Im}\,X$, $\exists \tilde{\beta}, z = X\tilde{\beta} \Rightarrow Hz = HX\beta = X(X^*X)^{-1}X^*X\beta = X\beta = z$

$\Rightarrow$ on retrouve par le calcul la propriété précédente.

■ Résidu : $\hat{\varepsilon} = Y - \hat{Y}$ "estimation" des ε_i

Prop : $E\begin{pmatrix}\hat{Y}\\ \hat{\varepsilon}\end{pmatrix} = \begin{pmatrix}X\beta\\ 0\end{pmatrix}$ et $\mathrm{Cov}\left(\begin{pmatrix}\hat{Y}\\ \hat{\varepsilon}\end{pmatrix}\right) = \sigma^2\begin{pmatrix} H = P_{\mathrm{Im}X} & 0\\ 0 & I-H = P_{\mathrm{Im}X^\perp}\end{pmatrix}$

Preuve : . $E(\hat{Y}) = X\beta$ a déjà été vérifié

$E(\hat{\varepsilon}) = E(Y - \hat{Y}) = E(Y) - E(\hat{Y}) = 0$

. $\hat{Y} = HY = P_{\mathrm{Im}X}Y$

$\hat{\varepsilon} = Y - \hat{Y} = (I-H)Y = (I - P_{\mathrm{Im}X})Y = P_{\mathrm{Im}X^\perp}Y$

$\begin{pmatrix}\hat{Y}\\ \hat{\varepsilon}\end{pmatrix} = \begin{pmatrix}P_{\mathrm{Im}X}\\ P_{\mathrm{Im}X^\perp}\end{pmatrix} Y$

$\Rightarrow \mathrm{Cov}\left(\begin{pmatrix}\hat{Y}\\ \hat{\varepsilon}\end{pmatrix}\right) = \begin{pmatrix}P_{\mathrm{Im}X}\\ P_{\mathrm{Im}X^\perp}\end{pmatrix}\left[\sigma^2 I_n\right]\begin{pmatrix}P^*_{\mathrm{Im}X} & P^*_{\mathrm{Im}X^\perp}\end{pmatrix}$

$= \sigma^2\begin{bmatrix} P_{\mathrm{Im}X}P^*_{\mathrm{Im}X} & P_{\mathrm{Im}X}P^*_{\mathrm{Im}X^\perp}\\ P_{\mathrm{Im}X^\perp}P^*_{\mathrm{Im}X} & P_{\mathrm{Im}X^\perp}P_{\mathrm{Im}X^\perp}\end{bmatrix}$

$= \sigma^2\begin{bmatrix} P_{\mathrm{Im}X} & 0\\ 0 & P_{\mathrm{Im}X^\perp}\end{bmatrix}$

Cor $E\begin{pmatrix}\hat\beta\\ \hat\varepsilon\end{pmatrix} = \begin{pmatrix}\beta\\ 0\end{pmatrix}$ et $Cov\left(\begin{pmatrix}\hat\beta\\ \hat\varepsilon\end{pmatrix}\right) = \begin{pmatrix}(X^*X)^{-1} & 0\\ 0 & P_{Im X^\perp}\end{pmatrix}$

Preuve : Il suffit de noter que $\hat\beta = (X^*X)^{-1}X^*\hat Y$.

Prop : $\hat\sigma^2 = \dfrac{SCR(\hat\beta)}{n-p} = \dfrac{\|Y - X\hat\beta\|_2^2}{n-p}$ est un estimateur sans biais de σ^2

Preuve.

$$SCR(\hat\beta) = \|Y - X\hat\beta\|^2 = \|\hat\varepsilon\|^2$$

or comme $E(\hat\varepsilon) = 0$ $\quad E(\|\hat\varepsilon\|^2) = E\,Tr(\hat\varepsilon\,\hat\varepsilon^*)$

$$= Tr\left(E(\hat\varepsilon\,\hat\varepsilon^*)\right)$$

et comme $E(\hat\varepsilon) = 0$ $E(\|\hat\varepsilon\|^2) = Tr(Cov(\hat\varepsilon))$

$$\text{or } Tr(Cov(\hat\varepsilon)) = Tr(\sigma^2 P_{Im X^\perp})$$

$$= \sigma^2(n-p) \quad \text{car } \dim Im X^\perp = n-p$$

Exc : régression linéaire simple

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$Y = X\beta + \varepsilon$$

avec $X = \begin{pmatrix}1 & x_1\\ | & |\\ 1 & x_n\end{pmatrix}$ et $\beta = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}$

$$X^*X = \begin{pmatrix}n & \sum x_i\\ \sum x_i & \sum x_i^2\end{pmatrix} = n\begin{bmatrix}1 & \bar x\\ \bar x & \overline{x^2}\end{bmatrix}$$

$$(X^*X)^{-1} = \frac{1}{n}\frac{1}{\overline{x^2} - (\bar x)^2}\begin{bmatrix}\overline{x^2} & -\bar x\\ -\bar x & 1\end{bmatrix} =$$

$$X^*Y = \begin{pmatrix}\sum y_i\\ \sum x_i y_i\end{pmatrix} = n\begin{pmatrix}\bar y\\ \overline{xy}\end{pmatrix}$$

$$\hat{\beta} = (X^*X)^{-1} X^* Y = \frac{1}{\overline{x^2} - \bar{x}^2} \begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix}$$

$$= \frac{1}{\overline{x^2} - \bar{x}^2} \begin{bmatrix} \overline{x^2}\bar{y} - \bar{x}\,\overline{xy} \\ -\bar{x}\bar{y} + \overline{xy} \end{bmatrix} \begin{matrix} \\ \\ \end{matrix} \quad \begin{matrix} \\ \\ \end{matrix} \quad \begin{matrix} -\bar{x}^2\bar{y} + \bar{x}(\bar{x}\bar{y}) \\ \\ \end{matrix}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{bmatrix} \bar{y} - \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}\,\bar{x} \\ \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \end{bmatrix} = \begin{bmatrix} \bar{y} - \dfrac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}\,\bar{x} \\ \dfrac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} \end{bmatrix}$$

$$\rho(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}} = \frac{\mathrm{cov}(x,y)}{\sigma(x)\sigma(y)}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{y} - \dfrac{\rho(x,y)\sigma(y)}{\sigma(x)}\,\bar{x} \\ \dfrac{\rho(x,y)\sigma(y)}{\sigma(x)} \end{pmatrix}$$

$$\hat{y} = \bar{y} + \frac{\rho(x,y)\sigma(y)}{\sigma(x)}(x - \bar{x}) \qquad \frac{\hat{y} - \bar{y}}{\sigma(x)} = \rho(x,y)\,\frac{x - \bar{x}}{\sigma(x)}$$

$$\hat{Y} = \bar{y}\mathbb{1} - \frac{\rho(x,y)\sigma(y)}{\sigma(x)}\left(X - \bar{x}\mathbb{1}\right)$$

$$Y - \hat{Y} = \left(Y - \bar{y}\mathbb{1}\right) - \frac{\rho(x,y)\sigma(y)}{\sigma(x)}\left(X - \bar{x}\mathbb{1}\right)$$

$$\frac{1}{n}\|Y - \hat{Y}\|^2 = \frac{1}{n}\|Y - \bar{y}\mathbb{1}\|^2 - 2\,\frac{\rho(x,y)\sigma(y)}{\sigma(x)}\,\frac{1}{n}\langle Y - \bar{y}\mathbb{1},\, X - \bar{x}\mathbb{1}\rangle + \frac{\rho^2(x,y)\sigma^2(y)}{\sigma^2(x)}\,\frac{1}{n}\|X - \bar{x}\mathbb{1}\|^2$$

$$= \sigma^2(y) - 2\,\frac{\rho(x,y)\,\mathrm{cov}(x,y)\sigma(y)}{\sigma(x)} + \rho^2(x,y)\sigma^2(y)$$

$$= \sigma^2(y)\left(1 - \rho^2(x,y)\right)$$

$SCR = \|Y - \hat{Y}\|^2 = n\,\sigma^2(y)\,(1 - \rho^2(x,y))$

$\hat{\sigma}^2 = \frac{1}{n-2}\|Y - \hat{Y}\|^2 = \frac{n}{n-2}\,\sigma^2(y)\,(1 - \rho^2(x,y))$

$$Cov\left(\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}\right) = (X^*X)^{-1}$$

$$= \frac{1}{n\,Var(x)}\begin{bmatrix}\overline{x^2} & -\bar{x}\\ -\bar{x} & 1\end{bmatrix}$$

$\Rightarrow Var(\hat{b}_1) = \frac{1}{n\,Var(x)}$ $\rightsquigarrow$ x dispersé

$Var(\hat{b}_0) = \frac{1}{n}\left[1 + \frac{\overline{x}^2}{\overline{x^2} - \overline{x}^2}\right]$ $\rightsquigarrow$ $\bar{x}$ proche de 0

Rq estimateur non corrélés si $\bar{x} = 0$

III Le modèle linéaire gaussien

on passe de $H_\varepsilon = \begin{cases} E(\varepsilon) = 0 \\ Cov(\varepsilon) = \sigma^2 I_n \end{cases}$

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

Hypothèse beaucoup plus forte $\Rightarrow$ résultats plus forts caractérisation des lois

Modèle : $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$

Prop :
- $\hat{\beta} = (X^*X)^{-1}X^*Y$, $\hat{Y} = X\hat{\beta} = P_{Im\,X}Y$ et $\hat{\varepsilon} = Y - \hat{Y} = P_{Im\,X^\perp}Y$
- $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^*X)^{-1})$, $\hat{Y} \sim \mathcal{N}(X\beta, \sigma^2 P_{Im\,X})$ et $\hat{\varepsilon} \sim \mathcal{N}(0, \sigma^2 P_{Im\,X^\perp})$
- $(\hat{\beta}, \hat{Y}) \perp\!\!\!\perp \hat{\varepsilon}$

Preuve : . le premier point est un rappel

. le second utilise le fait que l'image par une application linéaire d'un vecteur gaussien est un vecteur gaussien et que celui-ci est entièrement caractérisé par sa moyenne et sa covariance que l'on a déjà calculé

- le troisième point s'obtient en notant que $\begin{pmatrix} \hat{\beta} \\ \hat{\varepsilon} \end{pmatrix}$ est un vecteur gaussien dont la covariance a une structure par bloc $\left(\begin{array}{c|c} \cdot & 0 \\ \hline 0 & \cdot \end{array}\right)$

Prop 1 $\|\hat{\varepsilon}\|^2 \sim \sigma^2 \chi^2(n-p)$

Preuve : $\hat{\varepsilon} \sim \mathcal{N}(0, P_{\mathrm{Im}\, X^\perp})$

donc $\exists$ B matrice orthogonale telle que $B P_{\mathrm{Im}\, X^\perp} B^* = \left(\begin{array}{c|c} I_{n-p} & 0 \\ \hline 0 & 0 \end{array}\right)$

$\Rightarrow B\hat{\varepsilon} \sim \mathcal{N}\left(0, \sigma^2 \left(\begin{array}{c|c} I_{n-p} & 0 \\ \hline 0 & 0 \end{array}\right)\right)$

$\Rightarrow (B\hat{\varepsilon})_i = \sigma e_i \mathbb{1}_{i \le n-p}$ avec e_i iid

$$\|B\hat{\varepsilon}\| = \sum_i |(B\hat{\varepsilon})_i|^2 = \sum_{i=1}^{n-p} \sigma^2 |e_i|^2$$

$$\sim \sigma^2 \chi^2(n-p)$$

Rq : propriété générale des gaussiennes

$\varepsilon_1 \sim \mathcal{N}(0, P)$ et $\varepsilon_2 \sim \mathcal{N}(0, Q)$ avec P et Q deux projecteurs orthogonaux orthogonaux.

$\Rightarrow \|\varepsilon_1\|^2 \sim \chi^2(\mathrm{rg}\, P)$ $\|\varepsilon_2\|^2 \sim \chi^2(\mathrm{rg}\, Q)$

et $\|\varepsilon_1\|^2 \perp\!\!\!\perp \|\varepsilon_2\|^2$

Lien avec le maximum de vraisemblance

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

$$\Rightarrow dP(\tilde{Y} = Y ; \tilde{\beta}^2, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\|Y - X\beta\|^2}{2\sigma^2}} dY$$

$\Rightarrow$ La log vraisemblance des observations est donnée

$$L_n(\beta, \sigma^2) = -\frac{n}{2}\left[\log 2\pi + \log \sigma^2\right] - \frac{1}{2\sigma^2}\|Y - X\beta\|^2$$

$\Rightarrow$ Estimateur du maximum de vraisemblance

$$\frac{\partial L_n}{\partial \tilde{\beta}} = X^*(Y - X\tilde{\beta}) \quad \Rightarrow \hat{\beta}^{MV} = (X^*X)^{-1}X^*Y = \hat{\beta}$$

$$\frac{\partial L_n}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|Y - X\beta\|^2$$

$$\Rightarrow \widehat{\sigma^2}^{MV} = \frac{\|Y - X\hat{\beta}\|^2}{n} = \frac{n-p}{n}\widehat{\sigma^2}$$

biaisé mais converge en espérance !

Prop Parmi les estimateurs sans biais, $\hat{\beta}$ est l'estimateur de moindre variance (BUE Best Unbiased Estimator)

Preuve La matrice de Fisher est donnée (si β assez régulier)

$$J(\beta) = E\left(-\left(\frac{\partial \ln(\beta\sigma^2)}{\partial \beta}\right)^* \left(\frac{\partial \ln(\beta\sigma^2)}{\partial \beta}\right)\right)$$

$$= E\left(-\frac{\partial^2 \ln(\beta\sigma^2)}{\partial \beta^2}\right) \overset{\text{calcul déjà fait pour minimiser SSR}(\beta)}{=} E\left[\frac{1}{\sigma^2} X^* X\right]$$

$$= \frac{1}{\sigma^2}(X^* X)$$

- le Théorème de Cramer-Rao indique qu'un estimateur sans biais T de β doit satisfaire $\mathrm{Var}(T) \geq (J(\beta))^{-1} = \sigma^2 (X^* X)^{-1}$

L'estimateur des moindres carrés atteint cette borne, il est donc optimal

Prédiction $\hat{y} = x^* \hat{\beta} \sim \mathcal{N}(x^*\beta, x^*(X^*X)^{-1}x)$

IV Tests dans le modèle linéaire gaussien → Voir plus loin

V Analyse asymptotique du modèle linéaire

$$y_i = x_i^* \beta + \varepsilon_i \quad 1 \leq i \leq n$$

que se passe t'il quand n grandit ? $X_n = \begin{pmatrix} x_1^* \\ x_{n+1}^* \end{pmatrix}$ joue un rôle important

Sous H_ε

Prop Sous H_ε, $\hat{\beta}_n \xrightarrow{L^2} \beta$ dès que $\mathrm{Tr}\left[(X_n^* X_n)^{-1}\right] \to 0$

Preuve $E(\hat{\beta}_n) = \beta$ et $\mathrm{Var}(\hat{\beta}_n) = (X_n^* X_n)^{-1}$

$$E\|\hat{\beta}_n - \beta\|^2 = \mathrm{Tr}\left(X_n^* X_n\right)^{-1} \to 0$$

Cor Sous H_ε, $\mathrm{Tr}\left((X_n^* X_n)^{-1}\right) \to 0 \Rightarrow \hat{\beta}_n \xrightarrow{P} \beta$

Rq $\mathrm{Tr}\left((X_n^* X_n)^{-1}\right) \to 0 \Rightarrow (X_n^* X_n)^{-1} \to 0$ on matrice de covariance

Ex • Si X_i iid de loi Z telle que $Q = E(Z^* Z)$ dp

$$\Rightarrow \frac{1}{n} X_n^* X_n \to Q \text{ par la loi faible des grands nombres}$$

$$\Rightarrow n(X_n^* X_n^{-1}) \to Q^{-1}$$

$$\Rightarrow (X_n^* X_n^{-1}) \to 0$$

Sous H_ε + ε_i indépendant + Condition techniques

- $\frac{1}{n} X_n^* X_n \to Q$ définie positive ("Logique")

$H_{\varepsilon X}^+$: $\sum \|x_i\|^{2+\delta} E|\varepsilon_i|^{2+\delta} = o(n^{1+\delta/2})$ ("technique")

(ok si ε_i iid $E(|\varepsilon_1|^{2+\delta}) < +\infty$ et $\sum \|x_i\|^{2+\delta} \leq o(n^{1+\frac{\delta}{2}})$

(ok si $\|x_i\| < c$ ou $E(\|x_1\|^{2+\delta}) < +\infty$)

Pro1 Sous $H_{\varepsilon X}^+$, $\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 Q^{-1})$

Preuve

- $E(\hat{\beta}_n) = \beta$ et $Cov(\hat{\beta}_n) = \sigma^2 (X_n^* X_n)^{-1}$

donc $E(\sqrt{n}(\hat{\beta}_n - \beta)) = 0$ et $Cov(\sqrt{n}(\hat{\beta}_n - \beta)) = \sigma^2 \left(\frac{1}{n}(X_n^* X_n)\right)^{-1}$

- $\sqrt{n}(\hat{\beta}_n - \beta) = \sqrt{n}\left[(X_n^* X_n)^{-1} X_n^* Y - \beta\right]$

$= \sqrt{n}(X_n^* X_n)^{-1}\left[X_n^* Y - X_n^* X_n \beta\right]$

$= \sqrt{n}(X_n^* X_n)^{-1} X_n^* [Y - X_n \beta]$

$= \sqrt{n}(X_n^* X_n)^{-1} X_n^* \varepsilon_n$

$= \left(\frac{1}{n} X_n^* X_n\right)^{-1} \underbrace{\frac{1}{\sqrt{n}} X_n^* \varepsilon_n}_{Z_n}$

$Q_n^{-1} \to Q^{-1}$

$a^* Z_n = \frac{1}{\sqrt{n}} a^* X_n^* \varepsilon_n = \frac{1}{\sqrt{n}} (X_n a)^* \varepsilon_n$

$= \frac{1}{\sqrt{n}} \sum_{i=1}^n \underbrace{(x_i^* a)}_{\beta_i} \varepsilon_i \times \frac{1}{\sqrt{n}}$

- $\beta_i \perp\!\!\!\perp$
- $E(\beta_i) = 0$

$S_n^2 = \sum_{i=1}^n E(|\beta_i|^2)$

$K_n = \sum_{i=1}^n E(|\beta_i|^{2+\delta})$

$S_n^2 = \sum_{i=1}^n \frac{1}{n}(x_i^* a)^2 E(\varepsilon_i)^2 = \sigma^2 \frac{1}{n} a^* X_n^* X_n a \to \sigma^2 a^* Q a > 0$ $= \sigma^2 a^* Q_n a$

$K_n = \sum \frac{1}{n^{1+\delta/2}} (x_i^* a)^{2+\delta} E|\varepsilon_i|^{2+\delta} = o(1) \Rightarrow \frac{K_n}{(S_n)^{2+\delta}} \to 0$

TH de Lyapunov simplifié

$$\frac{1}{S_n} \sum_{i=1}^{n} \beta_i \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

$$\frac{1}{S_n} \frac{a^*}{\|a\|} Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

$$S_n^2 = \sigma^2 a^* Q_n a \to \sigma^2 a^* Q a$$

$$\Rightarrow a^* Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 a Q a)$$

$$\Rightarrow Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 Q)$$

or $\Rightarrow \sqrt{n}(\hat{\beta}_n - \beta) = Q_n^{-1} Z_n$

avec $Q_n^{-1} \to Q^{-1}$

$$\Rightarrow \sqrt{n}(\hat{\beta}_n - \beta) \to \mathcal{N}(0, \sigma^2 Q^{-1})$$

Rq En pratique, on utilise Q_n au lieu de Q

$$\sqrt{n}\, Q_n^{1/2} (\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 I_p)$$

■ Convergence pour $\hat{\sigma}^2$

Prop : Sous $H_{\varepsilon,X}$, $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$

+ ε_i iid

Preuve : $\hat{\sigma}^2 = \frac{\|Y - \hat{Y}\|^2}{n-p} = \frac{n}{n-p} \frac{\|(I - P_{Im X})Y\|^2}{n}$

$$= \frac{n}{n-p} \frac{\|(I - P_{Im X})\varepsilon\|^2}{n}$$

$$= \frac{n}{n-p} \frac{\varepsilon^* (I - P_{Im X}) \varepsilon}{n}$$

$$= \frac{n}{n-p} \left[\frac{\varepsilon^* \varepsilon}{n} - \frac{\varepsilon^* P_{Im X} \varepsilon}{n} \right]$$

$\frac{n}{n-p} \to 1$; $\frac{\varepsilon^* \varepsilon}{n} \xrightarrow{P} \sigma^2$ (loi des grands nbre)

$$\frac{\varepsilon^* P_{X_m} \varepsilon}{n} = \frac{\varepsilon^* X_m (X_m^* X_m)^{-1} X_m^* \varepsilon_m}{n}$$

$$= \frac{1}{\sqrt{n}} (X_m^* \varepsilon)^* (X_m^* X_m)^{-1} \frac{1}{\sqrt{n}} (X_m^* \varepsilon)$$

$$= Z_m (X_m^* X_m)^{-1} Z_m$$

$$= \frac{1}{n} Z_m Q_m^{-1} Z_m$$

$$= \frac{1}{n} \| \underbrace{Q_m^{-1/2} Z_m}_{\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 I_p)} \|_2^2 \xrightarrow{P} 0$$

Cor Si $E(|\varepsilon_i|^4) < +\infty$,

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \longrightarrow \mathcal{N}(0, \mathrm{Var}(|\varepsilon_1|^2))$$

Preuve

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \sqrt{n}\left[\frac{\varepsilon^* \varepsilon}{n} - \sigma^2\right] + \frac{p}{\sqrt{n}} \frac{\varepsilon^* \varepsilon}{n} + \sqrt{n} \frac{\varepsilon^* P_{X_m} \varepsilon}{n-p}$$

$\sqrt{n}\left[\frac{\varepsilon^* \varepsilon}{n} - \sigma^2\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}(|\varepsilon_1|^2))$

$\frac{p}{\sqrt{n}} \frac{\varepsilon^* \varepsilon}{n} \xrightarrow{P} 0$ car $\frac{\varepsilon^* \varepsilon}{n} \xrightarrow{P} \sigma^2$

$\sqrt{n} \frac{p}{n-p} \xrightarrow{P} 0$

IV Test dans le modèle linéaire gaussien

1 Tests simples

a Test de Student

(H_0) $\beta_q = a$ contre (H_1) $\beta_q \neq 0$

Prop : Sous H_0 $\frac{\hat{\beta}_q - a}{\hat{\sigma}\sqrt{(X^*X)^{-1}_{qq}}} \sim T(n-p)$ Student

Preuve Sous H_0 $\beta_q \sim \mathcal{N}(a, \sigma^2 (X^*X)^{-1}_{qq})$ et $\beta_q \perp\!\!\!\perp \hat{\sigma}^2$

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

Cor : Test $\left| \frac{\hat{\beta}_q - a}{\hat{\sigma}\sqrt{(X^*X)^{-1}_{qq}}} \right| > Q_{T(n-p)}\left(1 - \frac{\alpha}{2}\right) = \tilde{Q}_{T(n-p)}\left(\frac{\alpha}{2}\right)$

est de niveau α

Rq $\Leftrightarrow$ avec IC

Rq $F_T(Q_\alpha) = \alpha$ $\quad \tilde{Q}(\alpha) = Q(1-\alpha)$

Rq P value

$2 \tilde{Q}_{T(n-p)}\left(\left|\frac{\hat{\beta}_q - a}{\hat{\sigma}\sqrt{(X^*X)^{-1}_{qq}}}\right|\right)$

$= 2(1 - F_{T(n-p)}(\ldots))$

$\geq \alpha$ faut rien de la zone de rejet

- Test sur une combinaison linéaire (Généralisation)

$H_0 : b^* \beta = a$ contre H_1 $b^* \beta \neq a$

Prop sous H_0 $\dfrac{b^* \hat{\beta} - a}{\hat{\sigma} \sqrt{b^* (X^* X)^{-1} b}} \sim T(n-p)$

$\Rightarrow$ Test similaire au précédent

- Test sur σ^2

$H_0 : \sigma = \sigma_0$ contre $H_1 = \sigma \neq \sigma_0$

Prop sous H_0 $= (n-p) \dfrac{\hat{\sigma}^2}{\sigma_0^2} = \dfrac{1}{\sigma_0^2} \sum | y_i - x_i^* \hat{\beta} |^2 \sim \chi^2(n-p)$

$= \dfrac{SSR(\hat{\beta})}{\sigma_0^2}$

$\Rightarrow$

Test $\leq \dfrac{SSR(\hat{\beta})}{\sigma_0^2} \geq Q_{\chi^2(n-p)}\left(1 - \frac{\alpha}{2}\right)$

ou $\dfrac{SSR(\hat{\beta})}{\sigma^2} \leq Q_{\chi^2(n-p)}\left(\frac{\alpha}{2}\right)$

p values $= \min\left(2\left(1 - F_{\chi^2}\left(\dfrac{SSR(\hat{\beta})}{\sigma_0^2}\right)\right), 2 F_{\chi^2}\left(\dfrac{SSR(\hat{\beta})}{\sigma^2}\right) \right)$

$\Rightarrow$ Intervalle de confiance

- Test de Fisher

$\mathcal{M}_1 : Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ $\beta \in \mathbb{R}^p$

$\mathcal{M}_0 : Y \sim \mathcal{N}(Z\gamma, \sigma^2 I_n)$ $\gamma \in \mathbb{R}^q$ $q < p$

on dit que $\mathcal{M}_0$ est un sous modèle de $\mathcal{M}_1$ (ou que les modèles sont emboîtés)

si $\operatorname{Im} Z \subset \operatorname{Im} X$

Ex : • $\beta = H\gamma$ cf analyse de la variance

• importance de $r = p - q$ contraintes linéaires
$\rightsquigarrow$ sous espace de dimension q à paramètres.

Prédiction par ordre de complexité

$$Y \quad \hat{Y}^{(M_1)} \quad \hat{Y}^{(M_0)}$$
$$\| \quad \|$$
$$P_{Im X} Y \quad P_{Im Z} Y$$

comme $Im Z \subset Im X$ on en déduit

que $I - P_{Im X} \perp P_{Im X} - P_{Im Z}$

Conclusion gaussien $\underset{Y - \hat{Y}^{(M_1)}}{(I - P_{Im X})Y} \perp \underset{\hat{Y}^{(M_1)} - \hat{Y}^{(M_2)}}{(P_{Im X} - P_{Im Z})Y}$

Sous H_0 : M_0 est valide

$$Y - \hat{Y}^{(M_1)} \sim \mathcal{N}(0, \sigma^2 P_{Im X^\perp}) \quad \text{(c'est un également un } \mathcal{N}$$
$$\hat{Y}^{(M_1)} - \hat{Y}^{(M_0)} \sim \mathcal{N}(0, \sigma^2 P_{\text{Im } Z \text{ dans } Im X})$$

$$\Rightarrow \| Y - \hat{Y}^{(M_1)} \|^2 \sim \sigma^2 \chi^2(n-p) \quad + \perp$$
$$\| \hat{Y}^{(M_1)} - \hat{Y}^{(M_0)} \|^2 \sim \sigma^2 \chi^2(p-q)$$

$$\Rightarrow \frac{\| \hat{Y}^{(M_1)} - \hat{Y}^{(M_0)} \|^2 / (p-q)}{\| Y - \hat{Y}^{(M_1)} \|^2 / (n-q)} \sim \mathcal{T}(p-q, n-p)$$

Rq : $$\frac{[SSR(M_0) - SSR(M_1)] / (p-q)}{SSR(M_1) / (n-q)} \sim \mathcal{T}(p-q, n-q)$$

Rem : $\| Y - \hat{Y}^{(M_1)} \|^2 = SSR(M_1)$

$\| Y - \hat{Y}^{(M_0)} \|^2 = SSR(M_0) = \| Y - Y^{(M_1)} \|^2 + \| Y^{(M_1)} - Y^{(M_0)} \|^2$

$\Rightarrow \| \hat{Y}^{(M_1)} - \hat{Y}^{(M_0)} \|^2 = SSR(M_0) - SSR(M_1)$

$\Rightarrow$ Test de Fisher.

0 Test de Wald

$\mathcal{N}(X\beta, \sigma^2 I_m)$ avec $\beta \in \mathbb{R}^p$

Soit C $\mathcal{M}_{p-q \times p}$ de rang $(p-q)$ $r \in p-q$ entiers avec $p \leq q < p$

et $c \in \mathbb{R}^{p-q}$

$H_0 : C\beta = c$ contre $H_1 : C\beta \neq c$

Sous H_0 $C\hat{\beta} - c \sim \mathcal{N}(0, \sigma^2 C (X^*X)^{-1} C^*)$

Lem $Z \sim \mathcal{N}_m(0, \Gamma)$ avec Γ inversible

$\Rightarrow$ $Z^* \Gamma^{-1} Z \sim \chi^2_m$

Preuve $\Gamma^{-1} = PDP$ avec Γ inversible et matrice de covariance
$D \geq 0$

$= {}^tP D^{1/2} D^{1/2} P$

$Z^* \Gamma^{-1} Z = \| D^{1/2} P Z \|^2$

or $D^{1/2} P Z \sim \mathcal{N}(0, \underbrace{D^{1/2} P \Gamma^{-1} P D^{1/2}}_{D^{1/2} P P^{-1} D^{-1} P^{-1} P D^{1/2}})$

I_m

Prop $\dfrac{\left[(C\hat{\beta} - c)^* \left[C (X^*X)^{-1} C^* \right]^{-1} (C\hat{\beta} - c)\right] / (p-q)}{\hat{\sigma}^2} \sim \mathcal{F}(p-q, n-p)$

Preuve Lemme + ind de $\hat{\beta}$ et $\hat{\sigma}^2$

"Cor" $C = Id$ $\hat{\beta} = \beta$

$$\frac{\left[(\hat{\beta} - \beta)^* (X^*X) (\hat{\beta} - \beta)\right] / p}{\hat{\sigma}^2} \sim \mathcal{F}(p, n-p)$$

$\Rightarrow$ Ellipsoïde de confiance

Ex : Test de rupture : analyse de la covariance avec variance

Période 1 $\mathbb{R}^p$ $Y_1 \sim \mathcal{N}(X_1 \beta_1, \sigma_1^2 \Sigma_n)$ $\beta_1 \in \mathbb{R}^p$ n_1 obs

Période 2 $Y_2 \sim \mathcal{N}(X_2 \beta_2, \sigma_2 \Sigma_n)$

Test d'absence de rupture

- Test $\sigma_1^2 = \sigma_2^2$

$$\frac{SCR(\hat{\beta}_1)/(n_1-p)}{SCR(\hat{\beta}_2)/(n_2-p} \sim F(n_1-p, n_2-p)$$

- Modèle "complet"

Test $\beta_1 = \beta_2$ par la méthode de Fisher

- Test de Fischer du modèle

H_0 $E(Y) = \beta_0 \beta$ H_1 $E(Y) = X\beta$

Sous H_0 : $$\frac{[SCR(H_0) - SCR(H_1)]/p-1}{SCR(H_1)/(n-p)} \underset{\text{ou}}{\sim} F(p-1, n-p)$$

$$= \frac{\|X\hat{\beta} - \bar{Y}\|^2 / p-1}{\|X\hat{\beta} - Y\|^2/(n-p)}$$

- Coefficient $R^2 = \frac{\|\hat{Y} - \bar{Y}\|^2}{\|Y - \bar{Y}\|^2} \leq 1$

mesure d'ajustement de $\hat{Y}$

$$R^2 = 1 - \frac{\|Y - \hat{Y}\|^2}{\|Y - \bar{Y}\|^2}$$

on a utilisé $\langle \hat{Y} - \bar{Y}, Y - \hat{Y} \rangle = 0$

on a également $\langle \hat{Y} - \bar{Y}, Y - \bar{Y} \rangle = \|\hat{Y} - \bar{Y}\|^2$

$$\Rightarrow R^2 = \frac{\langle \hat{Y} - \bar{Y}, Y - \bar{Y} \rangle^2}{\|\hat{Y} - \bar{Y}\|^2 \|Y - \bar{Y}\|^2} = \rho(Y, \hat{Y}_2)$$

- On vérifie que Fisher $\Omega_0 \subset \Omega_1$

$$F = \frac{n - \dim(\Omega_1)}{\dim(\Omega_1) - \dim(\Omega_0)} \frac{R^2_{\Omega_1} - R^2_{\Omega_0}}{1 - R^2_{\Omega_1}}$$

VI Moindres carrés généralisés

$$H'_\varepsilon \begin{cases} E(\varepsilon) = 0 \\ Cov(\varepsilon) = \sigma^2 \Sigma \quad \text{avec } \Sigma \neq c\, I_n \end{cases}$$

2 cas classiques : . $Cov(\varepsilon) = Diag(\sigma_i^2)$: hétéroscédasticité

. $Cov(\varepsilon_i, \varepsilon_j) \neq 0$ ~ ε : processus temporel

- Cas Σ connue

on note $\Sigma^{1/2}$ une racine de Σ qui existe

$\Sigma = ODO \Rightarrow \Sigma^{1/2} = OD^{1/2}$ orient

$\Rightarrow \Sigma^{-1/2} = OD^{-1/2}O$

$$\tilde{Y} = \Sigma^{-1/2} Y$$

$$\Rightarrow = \Sigma^{-1/2} X \beta + \underbrace{\Sigma^{-1/2} \varepsilon}_{\tilde{\varepsilon}}$$

$$\Rightarrow E(\tilde{\varepsilon}) = 0$$

$$Cov(\tilde{\varepsilon}) = \sigma^2 Id$$

$\Rightarrow$ on s'est ramené au cas ordinaire

$$\Rightarrow \hat{\beta} = (X^* \Sigma^{-1} X)^{-1} X \Sigma^{-1} Y$$

Interprétation géométrique : Σ^{-1} ⇒ forme quadratique définie positive

⇒ norme sur $\mathbb{R}^n$

$$\|Y\|^2_{\Sigma^{-1}} = Y^t \Sigma^{-1} Y$$

β = argmin $\|\hat{Y} - \hat{X}\beta\|^2$

= argmin $\|\Sigma^{-1/2}(Y - X\beta)\|^2$

= argmin $\|Y - X\beta\|^2_{\Sigma^{-1}}$

↝ Projection Σ^{-1} orthogonale !

Prop : • $E(\hat{\beta}) = \beta$ $\quad Var(\hat{\beta}) = (X^t \Sigma^{-1} X)^{-1}$

• (Gauss-Markov) $\hat{\beta}$ est le meilleur estimateur linéaire sans biais de β

• $\hat{Y} = X\hat{\beta}$ est le projeté Σ^{-1} orthogonal de Y sur $\mathrm{Im}\, X$

• $\hat{\sigma}^2 = \dfrac{\|\hat{Y} - \hat{X}\beta\|^2}{n-p} = \dfrac{\|Y - X\beta\|^2_{\Sigma^{-1}}}{n-p}$ estime sans biais σ^2 $\quad \left(\neq \dfrac{\|Y - X\beta\|^2_{\Sigma}}{n-p}\right)$

Preuve : Immédiat

Prop : • Si Y est gaussien : $\hat{\beta}$ est gaussien $\perp\!\!\!\perp \hat{\sigma}^2 \sim \sigma^2 \dfrac{\chi^2_{(n-p)}}{n-p}$.

• Pour β coïncide avec le maximum de vraisemblance

• $\hat{\beta}$ est optimal parmi les estimateurs sans biais.

Preuve $L_n(y; \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(Y - X\beta)^t \Sigma^{-1} (Y - X\beta)\right)$

Rq : Pour calculer $\hat{\beta}$, on peut • soit par $\hat{Y}$ et le calcul de $\Sigma^{-1/2}$

• soit par la forme et le calcul de Σ

→ calcul des SCR par des formules ≠

■ Cas Σ inconnu

$\rightsquigarrow$ on utilise les moindres carrés ordinaires :

$$\hat{\beta} = \operatorname{argmin} \|Y - X\beta\|_2^2$$

$$\Leftrightarrow \hat{\beta} = (X^*X)^{-1} X^* Y$$

Prop $\quad E(\hat{\beta}) = \beta$

$$\operatorname{Cov}(\hat{\beta}) = \sigma^2 \left[(X^*X)^{-1} X^* \Sigma X (X^*X)^{-1} \right]$$

⚠ on perd tout le reste ⚠

Prop $\quad$ si $\lambda_M(\Sigma_n) = O(1)$

$\lambda_m(X^*X) \xrightarrow[n\to\infty]{} +\infty \quad \left[\lambda_M\left[(X^*X)^{-1}\right] = O(1) \right]$

$\Rightarrow \operatorname{Tr}\left((X^*X)^{-1} X^* \Sigma X (X^*X)^{-1}\right) \longrightarrow 0$

$\Rightarrow \hat{\beta}$ est asymptotiquement convergent

Preuve :

$$\operatorname{Tr}\left((X^*X)^{-1} X^* \Sigma X (X^*X)^{-1} \right)$$

$$= \operatorname{Tr}\left(\Sigma X (X^*X)^{-2} X^* \right)$$

$$* \leq \lambda_M(\Sigma) \operatorname{Tr}\left(X (X^*X)^{-2} X^* \right) \leq \lambda_M(\Sigma) \operatorname{Tr}\left((X^*X)^{-1} \right)$$

$$\leq \lambda_M(\Sigma)\, p\, \lambda_M\left((X^*X)^{-1}\right) \longrightarrow 0$$

$*$: $\Sigma = O^* D O$ avec O orthonormée

$$\operatorname{Tr}(\Sigma V) = \operatorname{Tr}(O^* D O V)$$

$$= \operatorname{Tr}(D O V O^*)$$

$$\leq \left(\max_i D_{ii}\right) \operatorname{Tr}(O V O^*)$$

$$\leq \lambda_M(\Sigma) \operatorname{Tr}(V)$$

- Estimation de Σ impossible : n^2 paramètres
  mais estimation possible si Σ a une forme paramétrique simple ...

VII Etude des résidus

1/ Tests

- résidu $\hat{\varepsilon} = Y - \hat{Y} \quad \simeq \quad \varepsilon = Y - E(Y)$

$E(\hat{\varepsilon}) = 0$	$E(\varepsilon) = 0$
$Cov(\hat{\varepsilon}) = \sigma^2 (1-H)$	$Cov(\varepsilon) = \sigma^2 I$
$\hookrightarrow Var\ \hat{\varepsilon}_i = \sigma^2 (1-h_{ii})$	$\hookrightarrow Var(\varepsilon_i) = \sigma^2$

- Modification des résidus pour les rendre de même variance

$$\tilde{\epsilon}_i = \frac{\hat{\varepsilon}_i}{\sqrt{1-h_{ii}}} \qquad \tilde{\tilde{\epsilon}}_i = \frac{\hat{\varepsilon}_i}{\sigma\sqrt{1-h_{ii}}}$$

$\Rightarrow$ Standardisation en utilisant une estimation de la variance

$$\epsilon_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \qquad \text{ou} \qquad \epsilon_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

où $\hat{\sigma}_{(i)}$ estimation dans le modèle où l'on enlève la $i^{ème}$ observation

Rq $\epsilon_i \simeq \mathcal{N}(0,1)$ $\qquad \epsilon_i^* \sim \mathcal{F}(n-1-p)$

Pb pour calculer $\hat{\sigma}_{(i)}$ $\Rightarrow$ Formule "magique"

$$\frac{\hat{\sigma}^2}{\hat{\sigma}^2_{(i)}} = \frac{n-p}{n-1-p-\epsilon_i^2}$$

$$\Rightarrow \epsilon_i^* = \epsilon_i \sqrt{\frac{n-p-}{n-1-p-\epsilon_i^2}}$$

⚠ Pas de décorrélation entre les $\hat{\epsilon}_i$, ϵ_i, ϵ_i^*. ⚠

En pratique, on fait les tests sur ϵ_i ou ϵ_i^*

⇒ Test de symétrie

- Wilcoxon

basé sur $\sum_i \text{sign}(\epsilon_{(i)})$

⇒ Test de symétrie / indépendance

- Wald-Wolfowitz Test des séqu

$\text{sign}(\epsilon_{(i)}) = + + - - + + + - + - -$

⇒ on compte le nombre de "suite de même signe"

⇒ Test d'indépendance

- Durbin-Watson

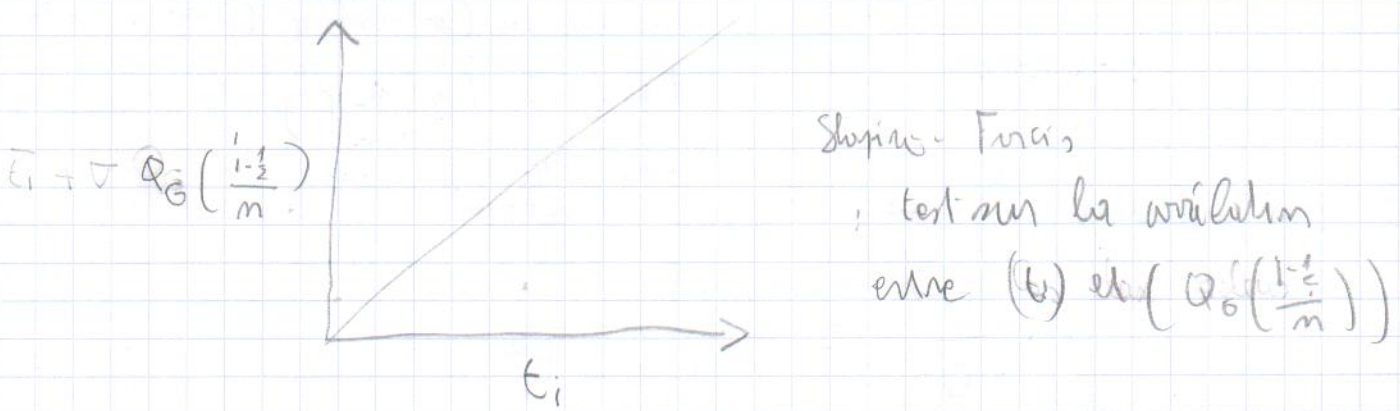
→ Estimation de ρ dans un modèle AR(1)

$$\epsilon_i = \rho \epsilon_{i-1} + \sqrt{1-\rho^2}\, a_i$$

+ Test de $\rho = 0$

⇒ Test de normalité

- Pearson : Test du χ^2 sur des boités équiprobables
- sQQplot de Shapiro-Francia



- Lilliefors (Kolmogorov Smirnov)

$$D_+ = \max\left(i/n - F_G\left(\frac{t_{(i)} - \bar{t}}{\hat{\sigma}}\right)\right)$$

$$D_- = \max\left(F_G\left(\frac{t_{(i)} - \bar{t}}{\hat{\sigma}}\right) - \frac{i-1}{n}\right)$$

$$D = \max(D_+, D_-)$$

$\Rightarrow \| F_n - F_G \|_\infty$: Kolmogorov - Smirnov

- Cramer-von-Mises

$$\sum_{i=1}^{n} \left(F_G\left(\frac{x_{(i)} - \bar{x}}{\hat{\sigma}}\right) - \frac{i - \frac{1}{2}}{n}\right)$$

- Anderson - Darling

$$\sum_{i=1}^{n/2} (2i-1)\left[\ln\left(F_G\left(\frac{t_{(i)} - \bar{t}}{\hat{\sigma}}\right)\right) + \ln\left[1 - F_G\left(\frac{t_{(n-i+1)} - \bar{t}}{\hat{\sigma}}\right)\right]\right.$$

plus de poids que le test de Cramer von Mises

- Shapiro Wilk

$$\frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n/2} a_i \left(t_{[n-i+1]} - t_{(i)}\right)^2$$ (la raie de Shapiro-Francia)

- Test de d'Agostino / Test de Jarque - Bera

Tests basés sur des relations entre les moments des gaussiennes.

$\Rightarrow$ Test de symétricité

2/ Diagnostics

Tests de comportement "bizarre" (différent) pour les données

- levier = "levier"

  . $h_{ii} = x_i^t (X^t X)^{-1} x_i = X^t (X^t X)^{-1} X = H$

  ≠ distance au centre du nuage des (x_i) mesuré par la distance de Mahalanobis associée au nuage de points.

  . $\hat{y}_i = h_{ii}\, y_i + \sum_{i \neq j} h_{ij}\, y_j$

Prop : $\sum h_{ii} = p$

Preuve $\sum h_{ii} = Tr(H) = p$

Heuristique : Si tous les points ont aussi importance, $h_{ii} \leq \frac{p}{n}$

Points à étudier si $h_i > c\,\frac{p}{n}$ $\quad c = 2$ ou 3

- Résidu standardisé $t_i = \dfrac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$

Heuristique : si t_i est grand, il faut comparer si c'est un point mal prédit à un point atypique

$|t_i| > t_\alpha \quad t_\alpha \simeq 2\text{-}3$

- Résidu studentisé $t_i^* = \dfrac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = t_i \sqrt{\dfrac{n-1-p}{n-p-t_i^2}}$

Heuristique similaire $|t_i^*| > t_\alpha \quad t_\alpha \simeq 2\text{-}3$ " $\simeq F(n-1-p)$ "

Rq Si le point est mal prédit ou atypique il a une influence fort sur $\hat{\sigma} \Rightarrow |t_i^*| >> |t_i| \rightarrow$ plus sensible

Rq : on a toujours $|t_i^*| \geq |t_i|$

■ Variabilité de la prédiction $\hat{y}_i$ quand on retire i :

Différence of fits

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i),i}}{\hat{\sigma}_{(i)}\sqrt{h_i}} = t_i^* \sqrt{\frac{h_i}{1-h_i}} \quad \left(\sim \mathcal{N}(0,1)\right)$$

Relation entre t_i^* et h_i

Test classique $|DFFITS_i| > c\sqrt{\frac{p}{n}}$ $\quad c \simeq 2-3$

■ Variabilité de la prédiction quand on retire i

Distance de Cook

$$D_i = \frac{\|\hat{Y} - \hat{Y}_{(i)}\|^2}{\hat{\sigma}^2 p} = \frac{t_i^2}{p}\frac{h_i}{1-h_i} \quad \left(\tilde{\sim} \frac{\chi^2(p)}{p}\right)$$

Test $D_i > 1$

ou $D_i > \frac{4}{n-p}$

■ Variabilité de l'estimation des coefficients quand on retire i :

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{(i),j}}{\hat{\sigma}_{(i)}\sqrt{(X^tX)^{-1}_{j,j}}} \quad \left(\sim \mathcal{N}(0,1)\right)$$

$$= t_i^* \left[\frac{[(X^tX)^{-1}X^t]_{j,i}}{\sqrt{(X^tX)^{-1}_{j,j}}\,(1-h_{ii})}\right]$$

$DFBETAS_{j,i} > 1$

$DFBETAS_{j,i} > \frac{c}{\sqrt{n}}$

⇒ Mesure l'influence composée par composante

■ Variabilité de la matrice

$$V(\hat{\beta}) = \hat{\sigma}^2 \det(X^tX)^{-1} \Rightarrow \text{taille des boules}$$

$$COVRATIO_i = \frac{V(\hat{\beta}_{(i)})}{V(\hat{\beta})}$$

$$= \left[\frac{n-1-p}{n-p} + \frac{t_i^{*2}}{n-p}\right]^p (1-h_{ii})^{-1}$$

Comparer au repère à 1

< 1 déstabilisante

> 1 améliorante

Test $|COVRATIO_i - 1| > \epsilon' \frac{p}{n}$ $\epsilon \sim 2 \to 3$

3/ Résidus partiels,

Outil de diagnostique

$$y_0 = x_0^* \beta + \varepsilon_0$$

$$\Rightarrow y_0 = x_{0(i)}^* \beta_{(i)} + x_{0i} \beta_i + \varepsilon_0$$

$$\varepsilon_{0,i} = y_0 - x_{(i)}^* \beta_{(i)} = x_{0i} \beta_i + \varepsilon_0$$

$\Rightarrow$ on se ramène à 1 pb à une variable

En pratique

$$\widehat{\varepsilon_{i}} = Y - X_{(i)} \widehat{\beta}_{(i)}$$

avec $\widehat{\beta}_{(i)}$ estimé soit dans le modèle complet soit dans le modèle sans la variable X_i

L'observation de $\widehat{\varepsilon}_{i,j} = y_j - x_{j,(i)}^* \widehat{\beta}_{(i)}$

en fonction de $\sim x_{j,(i)}$ permet de visualiser le reste de dépendance en X_i

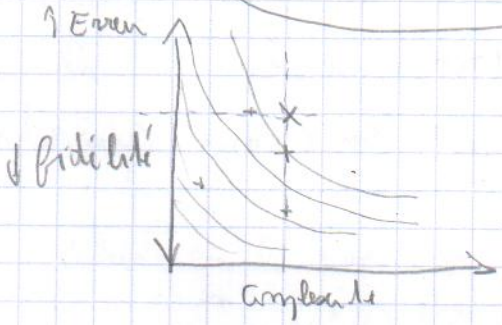
$\Rightarrow$ Estimation "visuelle" de l'information restante

VIII Sélection de Modèles / Sélection de variables

- Même question : comment choisir les régresseurs à utiliser ?
  Objectifs ≠ : prédiction → sélection de modèles
  interprétation → sélection de variables

- Principe de sélection = principe de parcimonie
  Principe du rasoir d'Occam
  Recherche d'un compromis entre fidélité aux observations et complexité.

⇒ Biais / Variance



⇒ A même complexité, on choisit le modèle le plus fidèle

⇒ A même fidélité, on choisit le modèle le moins complexe

Pb Comment donner un ordre "total" sur $\mathbb{R}^2$?

- Ordre partiel : Test de Fischer

$\mathcal{M}_0 \subset \mathcal{M}_1$

$$\Rightarrow \frac{[RSS(0) - RSS(1)]/(p_1 - p_0)}{RSS(1)/(n - p_1)} > t_\alpha$$

$$RSS(0) + t_\alpha \frac{RSS(1)}{n - p_1} p_0 > RSS(1) + t_\alpha \frac{RSS(1)}{n - p_1} p_1$$

⇒ choix du modèle $\mathcal{M}_1$

Critère $C(\mathcal{R}) = RSS(\mathcal{M}) + t_\alpha \hat{\sigma}^2 p$

- $R^2 = 1 - \frac{\|y - \hat{y}\|^2}{\|y - \bar{y}\|^2}$ augmente avec la fidélité mais ne diminue pas avec la complexité

⇒ pas de compromis

- $R^2_q = 1 - \dfrac{\|Y-\hat{Y}\|^2/(n-p)}{\|Y-\bar{Y}\|^2/(n-1)}$

  prise en compte de la complexité avec le terme en $(n-p)$

  $R^2_q(\mathcal{M}_0) \leq R^2_q(\mathcal{M}_1)$

  $\dfrac{\|Y-\hat{Y}^0\|^2}{n-p_0} \geq \dfrac{\|Y-\hat{Y}^1\|^2}{n-p_1}$

  $\hat{\sigma}_0^2 \geq \hat{\sigma}_1^2$ $\rightsquigarrow$ = Variance dans le modèle précis

  $\Rightarrow$ Difficile à utiliser.

- Principe d'optimisation en moyenne

  qualité d'un estimateur mesuré par son erreur quadratique moyenne:

$$E\left(\|E(Y|X) - \hat{Y}\|^2\right)$$

$$= \underbrace{\|E(Y|X) - E(\hat{Y})\|^2}_{\text{biais}^2} + \underbrace{Var\left(\|\hat{Y}\|^2\right)}_{\text{Variance}}$$

$$= \|E(Y) - P_{Im X} E(Y)\|^2 + Var\left(\|P_{Im X} Y\|^2\right)$$

$$= \|E(Y) - P_{Im X} E(Y)\|^2 + \sigma^2 p$$

Pb : Terme de biais impossible à mesurer

$\Rightarrow$ Estimation à l'aide de $\|Y-\hat{Y}\|^2$

$$E\left(\|Y-\hat{Y}\|^2\right) = E\left(\|Y - P_{Im X} Y\|^2\right)$$

$$= \|E(Y) - P_{Im X} E(Y)\|^2 + Var\left(\|Y - P_{Im X} Y\|^2\right)$$

$$= \|E(Y) - P_{Im X} E(Y)\|^2 + \sigma^2 (n-p)$$

$\Rightarrow$ $E\left(\|E(Y) - \hat{Y}\|^2\right) = E\left(\|Y-\hat{Y}\|^2 + (2p-n)\sigma^2\right)$

$\Rightarrow$ Utilisation du critère : $\|Y-\hat{Y}\|^2 + (2p-n)\sigma^2$

$\Rightarrow$ pb estimation de σ^2.

$\Rightarrow$ choix $\hat{\sigma}^2 = \dfrac{\|Y-\hat{Y}\|^2}{n-p}$ $\Rightarrow$ $\|Y-\hat{Y}\|^2 + (2p-n)\hat{\sigma}^2 = \dfrac{p}{n-p}\|Y-\hat{Y}\|^2$

no biais en Bon critère si pas de biais $= p\, R^2_q$

⇒ plus strict que R^2_a

⇒ Forme classique : C_p de Mallows

$$C_p = \frac{\|Y - \hat{Y}\|^2}{\sigma^2} + 2p - n$$

si l'estimateur est sans biais $C_p \geq p$

Règle classique : . minimiser C_p

Règle du bon sens : choisir le modèle le plus simple tel que $C_p \leq p$.

- Vraisemblance et pénalisation

$$\ln(Y ; (\beta, \sigma^2) ; \mathcal{M}) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\|Y - X\beta\|^2$$

Valeur en le maximum de vraisemblance

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad \hat{\sigma}^2 = \frac{\|Y - X\beta\|^2}{n} \neq MCO$$

$\ln(\mathcal{M})$:

$$\ln(Y ; (\hat{\beta}, \hat{\sigma}^2) ; \mathcal{M}) = -\frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2}\left[\log 2\pi + 1\right]$$

$$= -\frac{n}{2}\log\left(\frac{\|Y - \hat{Y}\|^2}{n}\right) - \frac{n}{2}\left[\log 2\pi + 1\right]$$

ne dépend que de l'attache aux données (comme le R^2)

⇒ nécessité d'une pénalisation

Choix classique : $-2\ln(\mathcal{M}) + 2\lambda_n |p|$

Variante sur le choix de λ_n

AIC : $\lambda_n = 2$

BIC : $\lambda_n = \log n$

AIC : An Information Criterion
Akaike Information Criterion

Idée : Bon modèle est un modèle tel que $P(X; \hat{\theta}(Y); \mathcal{M})$ soit proche de $P_0(X)$ en moyenne

⇒ Mesure par la divergence de Kullback-Leibler

$$KL(P_0, P(\cdot; \hat{\theta}(Y); \mathcal{M}))$$

$$= E_{X \sim P_0}\left[\ln \frac{P_0(X)}{P(\hat{\theta}, \mathcal{M})}\right]$$

$$= - E_{X \sim P_0}[-\ln P(X; \hat{\theta}(Y); \mathcal{M})] + E_{X \sim P_0}[\ln P_0(X)]$$

$$E_{X \sim P_0} KL(P_0, P(\cdot; \hat{\theta}(Y); \mathcal{M})$$

$$= - E_{X \sim P_0} E_{Y \sim P_0}[\ln P(X; \hat{\theta}(Y); \mathcal{M})]$$

$$+ \underbrace{E_{X \sim P_0}[\ln P_0(X)]}_{\text{cste}}$$

Pb : Comment estimer $E_{Y \sim P_0} E_{X \sim P_0}[-\ln P(X; \hat{\theta}(Y); \mathcal{M})]$

à partir de $-\ln P(Y; \hat{\theta}(Y); \mathcal{M})$

Idée de la preuve

$$\theta_0 = \operatorname{argmin} E_{X \sim P_0}[-\ln P(X, \theta; \mathcal{M})]$$

$$E_{Y \sim P_0} E_{X \sim P_0}[-\ln P(X; \hat{\theta}(Y); \mathcal{M}] - E_X[-\ln P(X, \theta_0; \mathcal{M})]$$

$$\simeq d/2 \quad (\text{q } \chi^2)$$

$$E_Y[-\ln P(Y; \theta_0; \mathcal{M})] - [-\ln P(Y; \theta_0(Y); \mathcal{M}]$$

$$\simeq d/2$$

⇒ χ^2 qui traînent ...

$$E_{Y\sim P_0} E_{X\sim P_0}\left[-\ln P(X;\hat\theta(Y);M)\right]$$

$$\simeq -\ln P(Y;\hat\theta(Y);M) + d = AIC(M)/2$$

$$AIC(M) = -2\ln P(Y;\hat\theta(Y);M) + 2d$$

* Validation croisée.

- BIC : Bayesian Information Criterion

Idée : Modélisation bayésienne et choix du modèle le plus probable étant donné Y

$$P(Y;\theta;M) = P(Y|\theta;M)\,P(\theta|M)\,P(M)$$

$$= P(M,\theta|Y)\,P(Y)$$

$$P(M|Y) = \int P(M,\theta|Y)\,dP_\theta$$

$$= \int \frac{P(Y|\theta;M)\,P(\theta|M)P(M)}{P(Y)}\,P(\theta|M)\,d\theta$$

$$= \frac{P(M)}{P(Y)}\int P(Y|\theta;M)\,P^2(\theta|M)\,d\theta$$

Si $P(M) = \text{cte}$ il suffit de calculer l'intégrale

$$\int P(Y|\theta;M)\,P^2(\theta|M)\,d\theta = \int e^{-n\left[\frac{1}{n}\left[-\ln P(Y|\theta,M) - 2\ln P(\theta|M)\right]\right]}\,d\theta$$

Lm : $\int e^{-nL(u)}\,du \simeq e^{-nL(u^*)}\left(\frac{2\pi}{n}\right)^{\frac{d}{2}}\left|-L''(u)\right|^{-1/2}$

$$L = \frac{1}{n}\left[-\ln P(Y|\theta;M) - 2\ln P(\theta|M)\right]$$

$$-L''(\hat\theta) = +\frac{1}{n}\left[\frac{\partial^2}{\partial\theta^2}\left(-\ln P(Y|\theta;M)\right) + 2\,\frac{\partial^2}{\partial\theta^2}\ln P(\theta|M)\right]$$

$$\hookrightarrow A(\theta^*_M) = \qquad = O\left(\frac{1}{n}\right)$$

$$-\ln P(m|y)$$

$$\simeq -\ln P(y|m,\theta^*) - 2\ln P(\theta^*|m) + \frac{d}{2}\log n - \frac{d}{2}\log 2\pi - \ln|A_n(\theta^*)|$$

$$\theta^* = \operatorname{argmin} L(\theta) = \operatorname{argmin}\left[-\frac{1}{n}\ln P(y|\theta,m) - \frac{2}{n}\ln P(\theta,m)\right]$$

$$\rightarrow \hat{\theta} = \operatorname{argmin}\left[\frac{1}{n}\ln P(y|\theta;M)\right]$$

$$\Rightarrow -\ln P(m|y) \simeq -\ln P(y|m,\hat{\theta}) + \frac{d}{2}\log n \underbrace{-2\ln P(\theta^*|m) + \frac{d}{2}\log 2\pi - \frac{1}{2}\ln|A_n\theta^*)|}_{O(1)}$$

$$\Rightarrow BIC(m) = -2\ln P(y|m,\hat{\theta}) + \log n\, d$$

$$\simeq -\ln P(m|y)$$

Rq : ne dépend pas de $P(\theta|m)$

Prise en compte d'a priori sur les modèles

en ajoutant un terme $-\ln P(m)$

Comparaison

R_a (C_p) AIC BIC

$\longleftarrow$ complexité des modèles croissante

Rq : AIC $\rightarrow$ C_p si σ^2 est connu.

lien avec Fisher sur les modèles emboîtés.

- Exploration exhaustive
  → Essayer tous les modèles
  et choisir celui qui minimise le critère choisi.

1b : Algorithmique combinatoire → coût explose avec le nb de variables

- Exploration intelligente.
  → Explorer l'ensemble des modèles en ajoutant / enlevant des variables explicatives en fonction du critère.
  → Stratégie "gloutonne" : moins coûteux mais difficile de mettre l'optimalité.

- Partial least Square
  on regarde les régresseurs au fur et à mesure en fonction de leur corrélation avec le résidu partiel
  on choisit ensuite dans la famille explorée le meilleur

- Moindre carrée pénalisée : si X^tX n'est pas inversé

$$\|Y - X\beta\|^2 + \lambda \|\beta\|_p$$

Si $p=0$: ½ critère de sélection de modèle
Si $p=2$: critère de régularisation de type Tykhonov
$p=1$: sélection + sélection et algorithme efficace
⇒ LASSO sélection + résultat théorique

- Validation croisée : Test des modèles avec une approche similaire à AIC en découpant les données en apprentissage et vérification.

Modèle linéaire généralisés

⇒ Généralisation du modèle linéaire

Famille de loi P_θ : gaussienne, binomiale, poisson

$\eta_i = x_i^* \beta$

$\Rightarrow y_i \sim P_{\theta_i} \Rightarrow g(E(y_i)) = x_i^* \beta$

↑ fonction de lien

ex. P_θ = gaussien, $g(x) = x \Rightarrow$ Modèle linéaire gaussien

. P_θ Bernoulli, $g(x) = \ln \frac{x}{1-x} \Rightarrow$ logit

. P_θ Poisson, $g(x) = \ln x$

⇒ Famille "exponentielle", Maximum de vraisemblance, Tests

I Famille exponentielle

Y_i variables indépendantes admettant des distributions issues d'une même structure exponentielle.

⇒ Densité ayant une forme particulière

$$f(y_i ; \theta_i, \phi) = e^{\frac{y_i \theta_i - b(\theta_i)}{\phi} + w(y_i, \phi)}$$

θ_i : paramètre naturelle ; ϕ : paramètre de nuisance ; $\frac{y_i\theta_i - b(\theta_i)}{\phi}$ dépend de θ_i ; $w(y_i, \phi)$ ne dépend pas de θ_i

Rq $E(y_i ; \theta, \phi) = b'(\theta)$

$\mathrm{Var}(y_i ; \theta, \phi) = \phi\, b''(\theta)$

$\frac{\partial \ln f}{\partial \theta} = \frac{y_i - b'(\theta_i)}{\phi}$ $\quad E\left(\frac{\partial \ln f}{\partial \theta}(x)\right) = 0$

$\frac{\partial^2 \ln f}{\partial \theta^2} = -\frac{b''(\theta_i)}{\phi}$ $\quad -E\left(\frac{\partial^2 \ln f}{\partial \theta^2}\right) = E\left(\left(\frac{\partial \ln f}{\partial \theta}\right)^2\right)$

Rq L'estimateur du maximum de vraisemblance de $Y_1, \dots, Y_n \sim f(\cdot ; \theta ; \phi)$ est $\hat{\theta}$ défini par $b'(\hat{\theta}) = \frac{1}{n} \sum Y_i$

⇒ Structure simple

⚠ estimation de ϕ plus complexe...

Ex • Gaussienne

$$f(y; m, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-m)^2}$$

$$= e^{\left(\frac{ym - \frac{m^2}{2}}{\sigma^2} + \left[-\frac{y^2}{2\sigma^2} - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2\right]\right)}$$

$\theta = m$

$\nu(m) = \frac{m^2}{2}$ $\quad \phi = \sigma^2$

$\nu'(m) = m \quad \Rightarrow E(y) = m$

$\nu''(m) = 1 \quad \Rightarrow V(y) = \sigma^2$

• Bernoulli

$$f(3, \pi) = \pi^3 (1-\pi)^{1-3} = 1-\pi \left(\frac{\pi}{1-\pi}\right)^3$$

$$= e^{3 \ln \frac{\pi}{1-\pi} + \ln(1-\pi)}$$

$\theta = \ln \frac{\pi}{1-\pi} \Rightarrow e^\theta = \frac{\pi}{1-\pi} \Rightarrow (1-\pi) e^\theta = \pi$

$\phi = 1$ $\quad \Rightarrow \pi = \frac{e^\theta}{1+e^\theta}$

$$f(3, \theta) = e^{3\theta - \ln(1+e^\theta)}$$

$\nu(\theta) = \ln(1+e^\theta) \quad (= \ln(1-\pi))$

$\nu'(\theta) = \frac{e^\theta}{1+e^\theta} \quad (= \pi = E(3))$

$$\nu''(\theta) = \frac{e^\theta(1+e^\theta) - e^\theta e^\theta}{(1+e^\theta)^2} = \frac{e^\theta}{(1+e^\theta)^2} = \frac{e^\theta}{1+e^\theta} \frac{1}{1+e^\theta} \left(= \pi \times (1-\pi)\right)$$

$= V(3)$

• Poisson

$$f(3, \lambda) = \frac{\lambda^3 e^{-\lambda}}{3!} = e^{3\ln\lambda - \lambda + \ln 3!}$$

$\theta = \ln\lambda \qquad \phi = 1$

$\nu(\theta) = e^\theta \quad (= \lambda)$

$\nu'(\theta) = e^\theta \quad (= \lambda = E(3))$

$\nu''(\theta) = e^\theta \quad (= \lambda = E(3))$

Binomiale, Gamma ...

Rq $v''(\theta) = \mathrm{Var}(y) \Rightarrow v$ strictement convexe

$\Rightarrow E(y) = b'(\theta)$ est en bijection avec θ

Gaussien $E(y) = m = \theta$

Bernoulli $E(y) = \pi = \frac{e^\theta}{1+e^\theta} \Leftrightarrow \theta = \ln \frac{\pi}{1-\pi} = \ln \frac{E(y)}{1+E(y)}$ ← fonction canonique

Poisson $E(y) = \lambda = e^\theta \Leftrightarrow \theta = \ln \lambda = \ln E(y)$

II Modèles linéaires généralisés

$\begin{cases} \text{. Famille exponentielle pour } y_i \text{ indépendants} \\ \text{. Existence de régresseurs } x_i \text{ et d'un coefficient } \beta \text{ tel que} \\ \quad g(E(y_i|x_i)) = x_i^* \beta \quad \text{où } g \text{ est une fonction croissante} \end{cases}$

Ex
- Gaussienne : $g = Id$ (= fonction unique) $\Rightarrow$ modèle linéaire gaussien
- Bernoulli :
  - $g = \ln \frac{x}{1-x}$ (= fonction unique) $\Rightarrow$ modèle logit
    $g^{-1}(x) = \frac{e^x}{1+e^x}$
  - $g = \Phi^{-1}(x)$ (= probit gaussien) $\Rightarrow$ modèle probit
    $g^{-1}(x) = \Phi(x)$
  - $g(x) = \ln(-\ln(1-x))$ $\Rightarrow$ modèle log log
    $g^{-1}(x) = 1 + e^{e^{-x}}$
- Poisson
  - $g(x) = \ln x$ $\Rightarrow$ modèle poissonien avec lien logarithmique
    $g^{-1}(x) = e^x$

$\Rightarrow g(v'(\theta_i)) = x_i^* \beta \Leftrightarrow \theta_i = (v')^{-1}(g^{-1}(x_i^* \beta))$

Rq si g fonction unique $g(v'(\theta)) = \theta = x_i^* \beta$

$\Rightarrow \theta_i = x_i^* \beta$

- Prédiction $\hat{\beta} \Rightarrow E(y) = g^{-1}(x^* \hat{\beta})$
- Vraisemblance

$$\ell_n(y_i ; \theta \beta, \phi)$$

$$= \sum_{i=1}^{n} \ell_n(y_i ; \theta(x_i \beta), \phi)$$

$$= \sum_{i=1}^{n} \left[\frac{y_i \theta_i(\beta, x_i) - v(\theta_i(x, \beta))}{\phi} + w(y_i ; \phi) \right]$$

$$\frac{\partial \ell_n}{\partial \beta_j} = \sum \frac{\partial \ell_n}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta}$$

$$= \sum \frac{y_i - v'(\theta_i(x, \beta))}{\phi} \frac{\partial \theta_i(x, \beta)}{\partial \beta}$$

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{v''(\theta_i)} x_{ij} (g^{-1})'(x_i^* \beta)$$

$$\frac{\partial \ell_n}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - g^{-1}(x_i^* \beta)) x_{ij}}{\phi v''((v')^{-1} g^{-1}(x^* \beta))} (g^{-1})'(x_i^* \beta)$$

$\uparrow v''(\theta_i)$

$\Rightarrow$ Equations de maximum de vraisemblance

$$\frac{\partial \ell_n}{\partial \beta_j} = 0 \quad \forall j$$

$\Rightarrow$ Résolution numérique possible

Rq : si $g = (v')^{-1}$ $(g^{-1})' = v''$

$$\Rightarrow \frac{\partial \ell_n}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - v'(x_i^* \beta)) x_{ij}}{\phi}$$

$$\Rightarrow X^* (Y - \mu) = 0 \text{ où } \mu = \begin{pmatrix} v'(x_1^* \beta) \\ v'(x_i^* \beta) \end{pmatrix} = \begin{pmatrix} E(y_1) \\ E(y_n) \end{pmatrix}$$

III Qualités d'ajustement et Tests

Qualités d'ajustement

- Deviance $D = -2(\mathcal{L} - \mathcal{L}_{sat})$ ($\sim$ RSS)

  - $\mathcal{L}$: vraisemblance au maximum de vraisemblance dans le modèle considéré à p paramètres
  - $\mathcal{L}_{sat}$: vraisemblance dans le modèle saturé à n paramètres

Asymptotiquement sous H_0 le modèle est vrai : $D \sim \chi^2(n-p)$

- Pearson

$$\text{Sous } H_0 \quad X^2 = \sum_{i=1}^{n} \frac{(y_i - g^{-1}(x_i^* \hat{\beta}))^2}{\phi\, v''((v')^{-1} g^{-1}(x_i^* \hat{\beta}))} \sim \chi^2(n-p)$$

Tests

- Rapports de vraisemblance

  Pour deux modèles emboîtés $M_0 \subset M_1$

  Sous H_0, $D_1 - D_0 = -2(\mathcal{L}_1 - \mathcal{L}_0) \sim \chi^2(p_1 - p_0)$

  si ϕ est connu

  $\sim F(p_1 - p_0, n-p)$

  Gaussien

- Test de Wald

  H_0 : $C\beta = d$ avec r = nb de contrainte.

$$\Rightarrow \text{Sous } H_0 \quad (C\beta - d)^* \left(C (X^* W X)^{-1} C^* \right)^{-1} (C\beta - d) \sim \chi^2(r)$$

où W est diagonale et définie $W_{ii} = \dfrac{1}{\phi\, v''((v')^{-1} g^{-1}(x_i^* \beta))} \left[(g^{-1})'(x_i^* \beta) \right]^2$

Pb Estimer f à partir de l'observation

où $y_i = f(x_i) + \varepsilon_i$

- on a vu comment faire si $f(x_i) = \beta_0 + \beta_1 x_i$

ou plus généralement si $f(x_i) = \beta_0 \sum_{k=1}^{p} \beta_k \varphi_k(x_i)$

Ex classique $\varphi_k(x_i) = x_i^k$

⇒ Approche paramétrique

Hyp $f \in \{ f_\theta \text{ avec } \theta \in \mathbb{R}^p \} = \mathcal{M}$

⇒ $f_{\hat\theta}$ avec $\hat\theta = \operatorname{argmin} \sum |y_i - f(x_i)|^2$

- on a également étudié le choix de la première famille ...

- Approche non paramétrique

Hyp $f \in \mathcal{C}^0$ ⇒ Trop de solutions possible !

≃ on va chercher $\hat f_n$ dans un modèle $\mathcal{M}_n$

dont la complexité augmente avec $n \longrightarrow \mathcal{C}^0$

- Ex . $\mathcal{M}_n$ = polynôme de degré $k(n)$ avec $k(n) \longrightarrow +\infty$

Sélection du degré par pénalisation

$L^2([0,1])$ par validation croisée

- $\mathcal{M}_n$ = polynôme trigonométrique | Biais = régularité

↕ Même principe | Variance = dimension

$L^2([0,1])$

- Binning = découpage en boîte
  - ⇒ solution ne modèle linéaire simple dans chaque boîte
  - ⇒ estimé / polynomial / - ..
  - ⇒ Difficulté taille des boîtes : biais vs variance.

- Pb : bord des limites ⇒ discontinuité
  - ⇒ Solution de la régression locale
  - ⇒ régression avec les poids voisins
  - ⇒ Pb discontinuité ensuite avec l'entrée et la sortie des poids
  - ⇒ Noyau à support compact estimé / seulement
  - ⇒ Régression locale pondérée (less régression robuste)
  - ⇒ Pb choix de la largeur de boîte Biais du paramètre h Variance

- Boîte qui se "perd" : critère de raccord
  - ⇒ spline
  - ⇒ choix des nœuds ⇒ sélection de modèle.
  - ⇒ Problématique biais variance
  (Formulation variationnelle ?)
- ondelettes : ex de la base de Haar..
  "Pb de la taille des bins"
  - ⇒ modèles emboîtés + Bon exemple en $L^2(\mathbb{R})$