M1 Statistique
2e partie

I Vecteurs gaussiens
II Modèle linéaire gaussien
III Estimation non-paramétrique

I Vecteurs gaussiens

1- Propriétés des vecteurs gaussiens

Définition

Une variable aléatoire Z est dite de loi normale centrée réduite, ce qu'on note $Z \sim \mathcal{N}(0,1)$ si la loi a pour densité $x \mapsto \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Une v.a. Y est dite de loi normale de paramètres (μ, σ^2) si il existe une v.a. $Z \sim \mathcal{N}(0,1)$ telle que $Y = \mu + \sigma Z$.

Lorsque $\sigma = 0$, on dit que Y est une v.a. gaussienne (normale) dégénérée

Remarque

Si $\sigma > 0$, la densité de la v.a. $Y \sim \mathcal{N}(\mu, \sigma^2)$ est $x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Théorème Soit $Y \sim \mathcal{N}(\mu, \sigma^2)$
Pour tout $\xi \in \mathbb{C}$, on a $\varphi_Y(\xi) = e^{i\xi\mu - \frac{\sigma^2\xi^2}{2}} = \mathbb{E}(e^{i\xi Y})$

Définition

Un vecteur aléatoire X à valeurs dans $\mathbb{R}^d$ est un vecteur gaussien si toute combinaison linéaire de ses composantes est une va. gaussienne.

Le vecteur des moyennes de X est le vecteur $(\mathbb{E}(X_1), \dots \mathbb{E}(X_d))$ et sa matrice de covariance est

$$\Sigma_{i,j} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}\left((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j)) \right)$$

$$= \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\,\mathbb{E}(X_j)$$

Théorème

La loi d'un vecteur gaussien est caractérisée par son vecteur des moyennes m et sa matrice de covariance Σ. On note cette loi $\mathcal{N}(m, \Sigma)$, et si $X \sim \mathcal{N}(m, \Sigma)$, alors

$$\xi \in \mathbb{C}^d \qquad \varphi_X(\xi) = \mathbb{E}\left(e^{i \xi \cdot X}\right) = \exp\left(i m \cdot \xi - \frac{1}{2}\, {}^t\xi \Sigma \xi \right)$$

preuve

La loi d'un vecteur aléatoire est caractérisée par sa fonction caractéristique. Soit X un vecteur gaussien, on pose m son vecteur des moyennes et Σ sa matrice de covariance.

Pour $\lambda \in \mathbb{R}^d$, on calcule $\mathbb{E}\left(e^{i \lambda \cdot X}\right)$

Puisque X est un vecteur gaussien, $\lambda_1 X_1 + \dots + \lambda_d X_d = \lambda \cdot X$ est une v.a. gaussienne. De plus, par linéarité

$$\mathbb{E}(\lambda \cdot X) = \lambda_1 \mathbb{E}(X_1) + \lambda_2 \mathbb{E}(X_2) + \dots + \lambda_d \mathbb{E}(X_d)$$

$$= \lambda \cdot m$$

$$\mathrm{Var}(\lambda \cdot X) = \mathrm{Var}(\lambda_1 X_1 + \dots + \lambda_d X_d)$$

$$= \sum_{i=1}^{d} \mathrm{Var}(\lambda_i X_i) + 2 \sum_{i=1}^{d} \sum_{j=i+1}^{d} \mathrm{Cov}(\lambda_i X_i, \lambda_j X_j)$$

$$\mathrm{Var}(\lambda . X) = \sum_{i=1}^{d} \lambda_i^2 \mathrm{Var}(X_i) + 2 \sum_{i=1}^{d} \sum_{j=i+1}^{d} \lambda_i \lambda_j \mathrm{Cov}(X_i, X_j)$$

$$= \sum_{i=1}^{d} \sum_{j=1}^{d} \lambda_i \lambda_j \Sigma_{i,j}$$

$$= {}^t\lambda \Sigma \lambda$$

Autrement dit, $\lambda . X \sim \mathcal{N}(\lambda . m, {}^t\lambda \Sigma \lambda)$

donc $\mathbb{E}\left(e^{i \lambda . X}\right) = e^{i \lambda . m - \frac{1}{2} {}^t\lambda \Sigma \lambda}$

La fonction caractéristique de X ne dépend que de m et de Σ, donc la loi de X est caractérisée par m et par Σ.

La fonction caractéristique s'étendant de façon unique en une fonction holomorphe, on a également

$$\mathbb{E}\left(e^{i \lambda . X}\right) = e^{i \lambda . m - \frac{1}{2} {}^t\lambda \Sigma \lambda} \quad \text{pour } \lambda \in \mathbb{C}^d$$

Propriétés

* Σ est une matrice symétrique positive

$$\Sigma_{i,j} = \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i) = \Sigma_{j,i}$$

Pour tout $\lambda \in \mathbb{R}^d$, ${}^t\lambda \Sigma \lambda = \mathrm{Var}(\lambda . X) \geq 0$

* Σ est diagonalisable en base orthonormée

* Si $(X_1, \dots X_d)$ sont des vac gaussiennes indépendantes alors $X = (X_1, \dots X_d)$ est un vecteur gaussien

* X gaussien $\Rightarrow$ toutes les coordonnées de X sont gaussiennes

Exemple

$Y \sim \mathcal{N}(0,1)$, $Z \sim \mathcal{B}(\frac{1}{2})$ $Y \perp\!\!\!\perp Z$

On pose $X = (Y, (2Z-1)Y)$

$X_1 \sim \mathcal{N}(0,1)$

$$\mathbb{E}\left(e^{i\xi(2Z-1)Y}\right) = \mathbb{E}\left(\mathbb{E}\left(e^{i\xi(2Z-1)Y} \mid Y\right)\right)$$

$$= \mathbb{E}\left(\frac{1}{2}e^{i\xi Y} + \frac{1}{2}e^{-i\xi Y}\right)$$

$$= \frac{1}{2}e^{-\frac{\xi^2}{2}} + \frac{1}{2}e^{-\frac{(-\xi)^2}{2}} = e^{-\frac{\xi^2}{2}}$$

donc $X_2 \sim \mathcal{N}(0,1)$

mais X n'est pas gaussien

$X_1 + X_2 = 2ZY$ n'est pas gaussienne

$$\mathbb{P}(X_1 + X_2 = 0) = \frac{1}{2}$$

Proposition

Soit X un vecteur gaussien

Si $\mathrm{Cov}(X_i, X_j) = 0$, alors $X_i \perp\!\!\!\perp X_j$

preuve

Le vecteur (X_i, X_j) est un vecteur gaussien

de matrice de covariance $\Sigma = \begin{pmatrix} \mathrm{Var}(X_i) & 0 \\ 0 & \mathrm{Var}(X_j) \end{pmatrix}$

donc $\mathbb{E}\left(e^{i(\lambda X_i + \mu X_j)}\right) = e^{i\lambda \mathbb{E}(X_i) + i\mu \mathbb{E}(X_j) - \frac{1}{2}\lambda^2 \mathrm{Var}(X_i) - \frac{1}{2}\mu^2 \mathrm{Var}(X_j)}$

$= \mathbb{E}\left(e^{i\lambda X_i}\right)\mathbb{E}\left(e^{i\mu X_j}\right)$

donc $X_i \perp\!\!\!\perp X_j$

Proposition

Si $X \sim \mathcal{N}(m, \Sigma)$ alors pour tout $A \in \mathcal{M}_{d',d}$ et pour tout $b \in \mathbb{R}^{d'}$, $AX + b \sim \mathcal{N}(Am+b, A\Sigma {}^tA)$

preuve

Toute combinaison linéaire des coordonnées de $AX+b$ est une combinaison linéaire des coordonnées de X donc est gaussienne

On calcule son vecteur des moyennes. Soit $i \leq d'$

$$\mathbb{E}\left((AX+b)_i\right) = \mathbb{E}\left((AX)_i + b_i\right)$$

$$= \mathbb{E}\left((AX)_i\right) + b_i$$

$$= \mathbb{E}\left(\sum_{k=1}^{d} A_{ik} X_k\right) + b_i$$

$$= \sum_{k=1}^{d} A_{i,k}\, \mathbb{E}(X_k) + b_i$$

$$= \sum_{k=1}^{d} A_{i,k}\, m_k + b_i = (Am+b)_i$$

On calcule la matrice de covariance.
Soit $i, j \leq d$ et on calcule

$$\mathrm{Cov}\left((AX+b)_i, (AX+b)_j \right)$$

$$= \mathrm{Cov}\left((AX)_i + b_i, (AX)_j + b_j \right)$$

$$= \mathrm{Cov}\left((AX)_i, (AX)_j \right)$$

$$= \mathrm{Cov}\left(\sum_{k=1}^{d} A_{ik} X_k, \sum_{\ell=1}^{d} A_{j\ell} X_\ell \right)$$

$$= \sum_{k=1}^{d} \sum_{\ell=1}^{d} A_{ik} A_{j\ell} \mathrm{Cov}(X_k, X_\ell)$$

$$= \sum_{k=1}^{d} \sum_{\ell=1}^{d} A_{ik} A_{j\ell} \Sigma_{k\ell}$$

$$= \sum_{\ell=1}^{d} (A\Sigma)_{i\ell} {}^tA_{\ell j} = (A\Sigma {}^tA)_{ij}$$

Exemple

$X \sim \mathcal{N}(0, I_2)$

$Y = AX + b$, avec $A = \begin{pmatrix} 1 & 2 \\ 0 & 3 \\ 2 & 1 \end{pmatrix}$ $b = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}$

$$Y \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 & 6 & 4 \\ 6 & 9 & 3 \\ 4 & 3 & 5 \end{pmatrix} \right)$$

$$\begin{pmatrix} 1 & 0 & 2 \\ 2 & 3 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 0 & 3 \\ 2 & 1 \end{pmatrix}$$

Algorithme de génération d'un vecteur gaussien

Soit $m \in \mathbb{R}^d$ et Σ une matrice symétrique positive

$$Y \sim \mathcal{N}(m, \Sigma)$$

Σ est diagonalisable en base orthonormée, donc

$$\Sigma = P D {}^tP$$

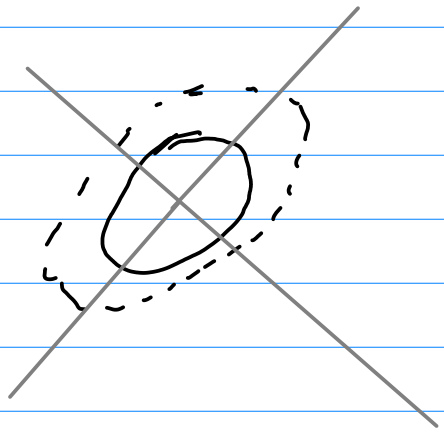
$$\text{Donc } X = {}^tP(Y - m) \sim \mathcal{N}(0, {}^tP \Sigma P) \sim \mathcal{N}(0, D)$$

La matrice de covariance de X est diagonale, donc les coordonnées $X_1, \dots X_d$ sont indépendantes

Par conséquent, si on pose $Z_1, \dots Z_d$ des v.a. iid de loi $\mathcal{N}(0,1)$, on définit ensuite

$$X_i = \sqrt{D_i}\, Z_i$$

puis on pose $Y = PX + m$



Théorème

Soit $X \sim \mathcal{N}(m, \Sigma)$.

* Si Σ est inversible, alors
X a pour densité $\frac{1}{(2\pi)^{d/2}\sqrt{\det\Sigma}} \exp\left(-\frac{1}{2}{}^t(x-m)\Sigma^{-1}(x-m)\right)$

* Si Σ n'est pas inversible, alors
la loi de $(X-m)$ est supportée par $\mathrm{Ker}(\Sigma)^{\perp} = \mathrm{Vec}(e_i)$
où (e_i) les vep de Σ
de vap >0

preuve

- Si Σ est inversible, alors
$X = PZ + m$, avec $\Sigma = PD{}^tP$
et $Z \sim \mathcal{N}(0, D)$
Z a pour densité $\prod_{i=1}^{d} \frac{1}{\sqrt{2\pi D_i}} e^{-\frac{z_i^2}{2D_i}}$ $\quad D_i > 0$

d'où, par changement de variables, on obtient
la densité de X, où on utilise
$\prod_{i=1}^{d} D_i = \det\Sigma$

- Si Σ n'est pas inversible
alors $X = PZ + m$ p.s. avec $Z \sim \mathcal{N}(0, D)$
et le support de Z est donné par
$\{(x_1, \dots x_d) \in \mathbb{R}^d : x_i = 0 \text{ si } D_i = 0\}$

Proposition

Si $(Y, X_1, \dots X_d)$ est un vecteur gaussien

alors $\mathbb{E}(Y \mid X_1, \dots X_d)$ est une fonction affine de $X_1, \dots X_d$

$$\mathbb{E}(Y \mid X_1, \dots X_d) = a + b_1 X_1 + \dots b_d X_d$$

$Y - \mathbb{E}(Y \mid X_1, \dots X_d) \perp\!\!\!\perp (X_1, \dots X_d)$

Exemple

Calculer $\mathbb{E}(Y_1 \mid Y_2, Y_3)$

$$\mathbb{E}(Y_1 \mid Y_2, Y_3) = a + bY_2 + cY_3$$

avec $$\mathbb{E}\left(\mathbb{E}(Y_1 \mid Y_2, Y_3)\right) = a + b\mathbb{E}(Y_2) + c\mathbb{E}(Y_3)$$

$$Y \sim \mathcal{N}\left(\begin{pmatrix}0\\-1\\0\end{pmatrix}, \begin{pmatrix}5 & 6 & 4\\6 & 9 & 3\\4 & 3 & 5\end{pmatrix}\right)$$

$$\Rightarrow \quad 0 = a - b$$

$$\mathbb{E}\left(\mathbb{E}(Y_1 \mid Y_2, Y_3) Y_2\right) = a\mathbb{E}(Y_2) + b\mathbb{E}(Y_2^2) + c\mathbb{E}(Y_2 Y_3)$$

$$\Rightarrow \quad 6 = -a + 9b + 3c$$

$$\mathbb{E}\left(\mathbb{E}(Y_1 \mid Y_2, Y_3) Y_3\right) = a\mathbb{E}(Y_3) + b\mathbb{E}(Y_2 Y_3) + c\mathbb{E}(Y_3^2)$$

$$\Rightarrow \quad 4 = 3b + 5c$$

$$\begin{cases} a - b = 0 \\ -a + 9b + 3c = 6 \\ 3b + 5c = 4 \end{cases} \Rightarrow \begin{cases} a = \dots \\ b = \dots \\ c = \dots \end{cases}$$

Théorème Central Limite

Soit $(V_1, \dots V_n)$ un n-échantillon de vecteurs aléatoires de vecteur de moyenne m et de matrice de covariance Σ

alors $$\sqrt{n}\left(\overline{V}_n - m\right) \Rightarrow \mathcal{N}(0, \Sigma)$$

2 - Le modèle gaussien

Soit $(X_1, \dots X_n)$ un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$ sous la loi $\mathbb{P}_{\mu,\sigma^2}$

Proposition

L'estimateur du maximum de vraisemblance de (μ, σ^2) est $\left(\overline{X}_n, \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \right)$

preuve

$$L(X_1, \dots X_n; \mu, \sigma^2) = \sum_{i=1}^{n} \text{Log} \left(\frac{e^{-\frac{(X_i - \mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \right)$$

$$= \left(\sum_{i=1}^{n} - \frac{(X_i - \mu)^2}{2\sigma^2} \right) - \frac{n}{2} \text{Log}(2\pi\sigma^2)$$

On calcule le gradient de L

$$\begin{cases} \frac{dL}{d\mu} = \sum_{i=1}^{n} - \frac{(\mu - X_i)}{\sigma^2} = -n\frac{\mu}{\sigma^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n} X_i \\ \frac{dL}{d\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2(\sigma^2)^2} - \frac{n}{2\sigma^2} \end{cases}$$

$\Rightarrow$ L'estimateur du maximum de vraisemblance est $(\hat{\mu}_n, \hat{\sigma}_n^2)$ où $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2$$

Remarque

$\mathbb{E}(\hat{\mu}_n) = \mu$ $\qquad$ $\mathbb{E}(\hat{\sigma}_n^2) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left((X_i - \overline{X}_n)^2\right)$

$$= \frac{n-1}{n} \sigma^2$$

donc on préfère souvent $S_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$

car c'est un estimateur sans biais

Théorème de Cochran

Soit X un vecteur gaussien de loi $\mathcal{N}(m, \sigma^2 I)$

On note $\mathbb{R}^d = E_1 \oplus E_2 \oplus \dots \oplus E_r$ en sous-espaces orthogonaux

Les projections orthogonales $\Pi_{E_1} X \dots \Pi_{E_r} X$ sont des vecteurs gaussiens indépendants.

preuve

On construit une bon de $\mathbb{R}^d$ $e_1, \dots e_d$

telle que $e_1, \dots e_{d_1}$ est une bon de E_1

$e_{d_{r-1}}, \dots e_d$ est une bon de E_r

On note $U = (e_1 | \dots | e_d)$ $\quad (U\,{}^tU = I)$

On observe que ${}^tUX \sim \mathcal{N}({}^tUm, \underbrace{{}^tU\, \sigma^2 I\, U}_{\sigma^2 I})$

La matrice de cov de ${}^tUX$ est diagonale, donc ses coordonnées sont indép

donc $(e_1, X), \dots (e_{d_1}, X)$ est indépendant de

$(e_{d_1+1}, X), \dots (e_{d_2}, X)$ — —

$(e_{d_{r-1}+1}, X), \dots (e_{d_r}, X)$

d'où $\Pi_{E_1} X, \Pi_{E_2} X, \dots \Pi_{E_r} X$ sont indépendants

Corollaire

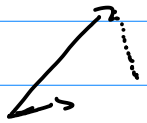
$\hat{\mu}_n \perp\!\!\!\perp \hat{\sigma}_n$

preuve

On décompose $\mathbb{R}^n = \underbrace{\text{Vect}(1,..1)}_{E_1} \oplus \underbrace{\text{Vect}(1,..1)^\perp}_{E_2}$

Posons $u = (1,..1)$

Soit $x \in \mathbb{R}^n$, $\pi_{E_1} x = \frac{(u,x)}{(u,u)} u = \frac{x_1+x_2+\dots+x_n}{n} u$



$\pi_{E_1} X = \bar{X}_n \, u$

$\pi_{E_2} x = x - \pi_{E_1} x = x - \frac{x_1+\dots+x_n}{n} u$

$$\pi_{E_2} X = \begin{pmatrix} X_1 - \bar{X}_n \\ X_2 - \bar{X}_n \\ \vdots \\ X_n - \bar{X}_n \end{pmatrix}$$

Par théorème de Cochran, $\pi_{E_1} X$ et $\pi_{E_2} X$

sont indépendants.

De plus $\hat{\sigma}_n^2 = \frac{1}{n} \| \pi_{E_2} X \|^2$

$\hat{\mu}_n = \frac{u \cdot \pi_{E_1}(X)}{n}$

On en déduit que $\hat{\mu}_n$ et $\hat{\sigma}_n^2$ sont indépendantes.

Définition

Loi du χ^2 à d degrés de liberté

X suit une loi du χ^2 à d degrés de liberté ($X \sim \chi^2(d)$) si il existe un vecteur gaussien $Y \sim \mathcal{N}(0, I_d)$ tel que $X = \|Y\|_2^2$

$$= Y_1^2 + Y_2^2 + \dots + Y_d^2$$

Propriété

Si $X_1 \sim \chi^2(d)$ et $X_2 \sim \chi^2(d')$ et $X_1 \perp\!\!\!\perp X_2$ alors $X_1 + X_2 \sim \chi^2(d+d')$

Proposition

Si $X \sim \chi^2(d)$ alors X admet pour densité

$$z \longmapsto \mathbb{1}_{z>0} \frac{1}{2^{\frac{d}{2}} \Gamma(\frac{d}{2})} z^{\frac{d}{2}-1} e^{-\frac{z}{2}}$$

preuve

$$\mathbb{E}(f(X)) = \mathbb{E}(f(Y_1^2 + \dots + Y_d^2))$$

$$= \int_{\mathbb{R}^d} f(y_1^2 + \dots + y_d^2) \frac{e^{-\frac{y_1^2 + \dots + y_d^2}{2}}}{(2\pi)^{\frac{d}{2}}} dy_1 \dots dy_d$$

$$= c \int_0^{+\infty} r^{d-1} f(r^2) e^{-\frac{r^2}{2}} dr$$

$z = r^2 \qquad dz = 2r\,dr$

$$= \frac{c}{2} \int_0^{+\infty} z^{\frac{d}{2}-1} f(z) e^{-\frac{z}{2}} dz$$

Définition

Loi de Student à d ddl

La v.a. Z est de loi de Student à d ddl (ce qu'on note $Z \sim \mathcal{T}(d)$) si on peut écrire

$$Z \sim \frac{X}{\sqrt{Y/d}} \quad \text{avec} \quad \begin{cases} X \sim \mathcal{N}(0,1) \\ Y \sim \chi^2(d) \\ X \perp\!\!\!\perp Y \end{cases}$$

Proposition

La densité de la loi $\mathcal{T}(d)$ est donnée par $x \mapsto \frac{1}{\sqrt{\pi d}} \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \left(1 + \frac{x^2}{d}\right)^{-\frac{d+1}{2}}$

En particulier, la loi $\mathcal{T}(d)$ n'admet pas de moment d'ordre d

Proposition

Soit X un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$

* $\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
* $\frac{n \hat{\sigma}_n^2}{\sigma^2} \sim \chi^2(n-1)$
* $\frac{\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}}{\sqrt{\frac{n\hat{\sigma}_n^2}{\sigma^2(n-1)}}} \sim \mathcal{T}(n-1)$

$$\sqrt{n-1}\frac{\overline{X}_n - \mu}{\sqrt{\hat{\sigma}_n^2}} \sim \mathcal{T}(n-1) \sim \sqrt{n}\frac{\overline{X}_n - \mu}{\sqrt{S_n^2}}$$

Remarque

Quand $n \to \infty$, $\frac{Y_1^2 + \dots + Y_n^2}{n} \longrightarrow 1$ p.s.

$$\frac{\chi^2(n)}{n} \to 1$$

$$\mathcal{C}(n) \Rightarrow \mathcal{N}(0,1)$$

pour de grandes valeurs de n, on fait

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sqrt{S_n^2}} \approx \mathcal{N}(0,1)$$

Intervalles de confiance

Intervalle de confiance de niveau $1-\alpha$ sur μ

Sous $\mathbb{P}_{\mu,\sigma^2}$ $\left(\sqrt{n-1}\,\left|\frac{\bar{X}_n - \mu}{\sqrt{S_n^2}}\right| > z\right) = \mathbb{P}\left(|\mathcal{C}_{n-1}| > z\right)$

donc en posant $t_{1-\alpha}^{(n-1)}$ le quantile d'ordre $1-\alpha$ de la loi $\mathcal{C}(n-1)$

on construit $IC_{1-\alpha}(\mu) = \left[\bar{X}_n \pm t^{(n-1)}_{1-\frac{\alpha}{2}} \sqrt{\frac{S_n^2}{n-1}}\right]$

De même, pour construire un IC sur la variance

$$\mathbb{P}_{\mu,\sigma^2}\left(\frac{(n-1)S_n^2}{\sigma^2} \in [a,b]\right) = \mathbb{P}\left(\chi^2(n-1) \in [a,b]\right)$$

d'où $IC_{1-\alpha}(\sigma^2) = \left[\frac{(n-1)S_n^2}{\chi^{(n-1)}_{1-\alpha/2}}, \frac{(n-1)S_n^2}{\chi^{(n-1)}_{\frac{\alpha}{2}}}\right]$

II Le modèle linéaire gaussien

On suppose qu'on dispose d'un vecteur d'observations $Y = (Y_1, \dots Y_n)$ associé à une matrice de données $X = \begin{pmatrix} X_{1,1} & \dots & X_{1,d} \\ \vdots & & \\ X_{n,1} & \dots & X_{n,d} \end{pmatrix}$

On suppose qu'il existe une relation linéaire entre X et Y, qu'on essaie de déterminer

1 - Définition du modèle linéaire gaussien

Définition vectorielle

Soit Y un vecteur d'observations réelles, on dit que Y suit un modèle linéaire gaussien si Y s'écrit $Y = m + \varepsilon$, où $\varepsilon \sim \mathcal{N}(0, \sigma^2 Id)$

On suppose que $m \in V$ un sous-espace vectoriel fixé et connu de $\mathbb{R}^n$

$$\left(\mathbb{P}_{m,\sigma^2} \sim \mathcal{N}(m, \sigma^2 I),\ m \in V,\ \sigma^2 \in \mathbb{R}_+\right)$$

Définition matricielle

Soit Y un vecteur d'observations réelles, on dit que Y suit un modèle linéaire gaussien si Y s'écrit

$$Y = X\beta + \varepsilon \quad \text{où} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 Id)$$

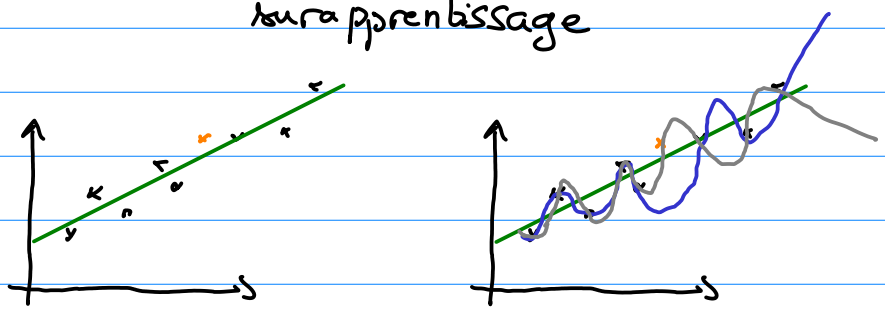
$$\forall j \leq n,\ Y_j = X_{j,1}\beta_1 + \dots + X_{j,d}\beta_d + \varepsilon_j$$

$$\left(\mathbb{P}_{\beta,\sigma^2} \sim \mathcal{N}(X\beta, \sigma^2 I),\ \beta \in \mathbb{R}^d,\ \sigma^2 \in \mathbb{R}_+\right)$$

Remarque

* L'espace vectoriel V dans la définition vectorielle correspond à $\mathrm{Im}(X) = \{X\beta, \beta \in \mathbb{R}^d\}$ dans la définition matricielle

* On suppose toujours que $\dim(V) = d \ll n$
Si d est trop grand, on risque le surapprentissage



La matrice X des données est souvent appelée matrice des variables explicatives et Y est le vecteur des variables à expliquer

Exemple

Y_t le prix d'achat d'une voiture
$X_{t,1}$ le revenu de l'acheteur
$X_{t,2}$ le nombre de km parcouru par l'acheteur

Un TLG pour Y serait

$$Y_t = \mu + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \varepsilon_t$$

On pourrait aussi imaginer

$$Y_t = \mu' + \beta_1' X_{t,1}^2 + \beta_2' \sqrt{X_{t,2}} + \varepsilon_t$$

Objectif du modèle linéaire gaussien

→ Expliquer les données,
expliciter l'existence d'une relation entre Y et X
corrélation n'est pas causalité

→ Prédire les mesures des futures données

→ Objectif statistique
Identifier β et σ^2

Tester l'hypothèse $\{\beta_1 = 0\}$

→ Deux types de données explicatives
Variables quantitatives $X \in \mathbb{R}$
Variables qualitatives $X \in E$ ensemble fini
Il est souvent intéressant
d'identifier E avec $\{(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)\}$
ANOVA

Remarque

Identifiabilité du modèle
$\theta \mapsto \mathbb{P}_\theta$ est injective
Le modèle linéaire gaussien est identifiable
ssi la matrice X est injective (de rang d)
(Pour tout $m \in \mathrm{Im}(X)$, il existe un unique
$\beta \in \mathbb{R}^d$ tel que $m = X\beta$)

Exemple

On teste une nouvelle combinaison
engrais - variété

On construit 4 types de terrains
ancien engrais - ancienne variété
ancien engrais - nouvelle variété
nouvel engrais - ancienne variété
nouvel engrais - nouvelle variété

Y_t le rendement de la parcelle,

le rendement moyen de chaque parcelle est donné par un rendement de référence μ et d'un ecart de rendement $c_1, \dots c_4$ dépendant de la méthode

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 \\ \vdots & 1 & 1 & 0 \\ & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{pmatrix} \qquad Y = X(\mu, c_1, \dots c_4) + \mathcal{N}(0, I_{\sigma^2})$$

ce modèle est non-identifiable

Pour le rendre identifiable, on peut ajouter une condition, par exemple
- $\mu = 0$, c_i le rendement moyen d'une parcelle de type i
- $c_1 + c_2 + c_3 + c_4 = 0$

2- Estimation de paramètre

On considère $\langle . \rangle$ le produit scalaire Euclidien sur $\mathbb{R}^n$, $\|.\|_2$ la norme associée à ce produit scalaire et Π_V est la projection orthogonale sur V

On note aussi Π_V la matrice de projection orthogonale sur V

Théorème

On considère le MLG $Y = m + \varepsilon$
où $m \in V$ et $\varepsilon \sim \mathcal{N}(0, I_n \sigma^2)$

L'estimateur du maximum de vraisemblance de (m, σ^2) est $(\hat{m}, \hat{S}_n^2)$, où

$$\rightarrow \hat{m} = \Pi_V Y$$

$$\rightarrow \hat{S}_n^2 = \frac{1}{n} \| Y - \Pi_V Y \|_2^2 = \frac{1}{n} \| \Pi_{V^\perp} Y \|_2^2$$

Les variables $\hat{m}$ et $\hat{S}_n^2$ sont indépendantes et de loi $\hat{m} \sim \mathcal{N}(m, \sigma^2 \Pi_V)$

$$\text{et } \frac{n \hat{S}_n^2}{\sigma^2} \sim \chi^2(n - \dim V)$$

Remarque

On a en particulier $\mathbb{E}\left[\frac{n \hat{S}_n^2}{\sigma^2}\right] = n - \dim(V)$

Si cet estimateur était non-biaisé, on aurait obtenu n, cet estimateur est donc biaisé

On peut utiliser $\hat{\sigma}_n^2 = \frac{1}{n - \dim V} \| Y - \Pi_V Y \|_2^2$

preuve

Sous la loi $\mathbb{P}_{m,\sigma^2}$ Y est de loi $\mathcal{N}(m, \sigma^2 I_d)$

Sa densité est donc

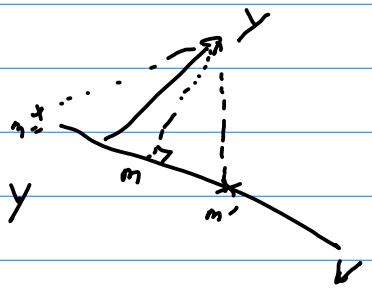
$$f(y; m, \sigma^2) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{(\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}(y-m)(y-m)\right)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|y-m\|^2}{2\sigma^2}\right)$$

La log-vraisemblance de Y est donc

$$L(y; m, \sigma^2) = -\frac{\|Y-m\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)$$

$\rightarrow$ Quelle que soit la valeur de σ^2, cette quantité est maximale pour m minimisant $\|Y-m\|$, c'est-à-dire $m = \Pi_V Y$



On fixe $m = \Pi_V(Y)$ et on étudie

$$\sigma^2 \mapsto L(Y; m, \sigma^2) = -\frac{\|Y-m\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)$$

En dérivant, cette quantité est maximale pour σ^2 tel que

$$\frac{\|Y-m\|^2}{2(\sigma^2)^2} - \frac{n}{2\sigma^2} = 0$$

$$\Rightarrow \quad \sigma^2 = \frac{1}{n}\|Y-m\|^2$$

Par théorème de Cochrane, on a

$$\Pi_V Y \perp\!\!\!\perp \Pi_{V^\perp} Y$$

donc $\hat{m} = \Pi_V Y \perp\!\!\!\perp \hat{\sigma}_n^2 = \frac{1}{n} \|\Pi_{V^\perp} Y\|_2^2$

$$\hat{m} \sim \mathcal{N}\left(\Pi_V m, \Pi_V (\sigma^2 I) {}^t\Pi_V\right) \qquad {}^t\Pi_V = \Pi_V \text{ et } \Pi_V \circ \Pi_V = \Pi_V$$

$$\sim \mathcal{N}\left(m, \sigma^2 \Pi_V\right)$$

et $\Pi_{V^\perp} Y \sim \mathcal{N}\left(0, \sigma^2 \Pi_{V^\perp}\right)$

$\dim(V^\perp) = n - \dim(V)$

donc $\dfrac{\|\Pi_{V^\perp} Y\|_2^2}{\sigma^2} \sim \chi^2(n - \dim V)$ □

On souhaite reformuler ce théorème sous forme matricielle pour une implémentation plus facile.
Une première étape est de déterminer la projection orthogonale sur $\mathrm{Im}(X)$

Lemme

Soit X une matrice $n \times d$ avec $n \geq d$

X injective $\iff {}^tX X$ inversible

preuve

Si ${}^tX X$ est inversible, alors elle est bijective

donc X est injective (et ${}^tX$ est surjective)

i.e est de rang maximum d

On suppose X injective, et on suppose par l'absurde que ${}^tXX$ n'est pas inversible

Il existe $u \in \mathbb{R}^d$ non nul tel que ${}^tXXu = 0$

Donc ${}^tu.({}^tXXu) = 0$

i.e. ${}^t(Xu)(Xu) = 0$

d'où $\|Xu\|^2 = 0$ ce qui implique $Xu = 0$

Mais $Xu = 0$ montre que X n'est pas inversible on a une contradiction. ∎

Proposition

Soit X une matrice $n \times d$ avec $n \geq d$ injective

La projection orthogonale de $\mathbb{R}^n$ sur $Im(X)$ est donnée par $X({}^tXX)^{-1}{}^tX$

Remarque

Posons $P = X({}^tXX)^{-1}{}^tX$, alors

- $P^2 = P$

$X({}^tXX)^{-1}{}^tXX({}^tXX)^{-1}{}^tX = X({}^tXX)^{-1}I{}^tX = P$

$\Rightarrow$ P est une matrice de projection

- ${}^tP = P$

${}^t(X({}^tXX)^{-1}{}^tX) = ({}^tX){}^t(({}^tXX)^{-1}){}^tX$

${}^t({}^tYX) = {}^tXX$

$= X({}^t({}^tXX))^{-1}{}^tX$

$= X({}^tXX)^{-1}{}^tX$

$\Rightarrow$ P est une matrice de projection orthogonale

preuve

La projection orthogonale sur $\mathrm{Im}(X)$ est la projection orthogonale sur $W = \mathrm{Vect}(e_1, \dots e_d)$

où $X = (e_1 | e_2 | \dots | e_d)$

X est injective donc $e_1, \dots e_d$ sont libres

Soit $u \in \mathbb{R}^n$, on cherche à déterminer $\hat{u} \in W$ tel que $u - \hat{u} \perp W$

Cela revient à imposer $\langle u - \hat{u}, e_i \rangle = 0 \quad \forall i \leq d$

On peut réécrire cette condition
$$\begin{aligned} {}^t e_1 \cdot (u - \hat{u}) &= 0 \\ {}^t e_2 \cdot (u - \hat{u}) &= 0 \\ &\vdots \\ {}^t e_d \cdot (u - \hat{u}) &= 0 \end{aligned}$$

i.e ${}^t X (u - \hat{u}) = 0$

Donc ${}^t X u = {}^t X \hat{u}$

De plus, comme $\hat{u} \in \mathrm{Im}(X)$, on peut écrire $\hat{u} = X\hat{v}$

On a donc ${}^t X u = {}^t X X \hat{v}$

d'où $\hat{v} = ({}^t X X)^{-1} {}^t X u$

On a donc $\hat{u} = X\hat{v} = X ({}^t X X)^{-1} {}^t X u$

On a donc bien montré que $\pi_{\mathrm{Im}(X)} = X ({}^t X X)^{-1} {}^t X$ ∎

Théorème Estimation de paramètre (forme matricielle)

X est une matrice injective $n \times d$

L'estimateur du maximum de vraisemblance du MLG $Y = X\beta + \varepsilon$, $\beta \in \mathbb{R}^d$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

est $\hat{\beta}_n = ({}^tXX)^{-1}{}^tXY$

$$\hat{S}_n^2 = \frac{1}{n}\| Y - X\hat{\beta}_n \|^2$$

De plus, $\hat{\beta}_n \perp\!\!\!\perp \hat{S}_n^2$ et $\beta_n \sim \mathcal{N}\left(\beta, \sigma^2({}^tXX)^{-1}\right)$

$$\frac{n\hat{S}_n^2}{\sigma^2} \sim \chi^2(n-d)$$

preuve

Le modèle vectoriel correspond au modèle matriciel avec $V = \mathrm{Im}(X)$

donc $\hat{m}_n = X\hat{\beta}_n$

$$\Pi_V Y = X\hat{\beta}_n$$

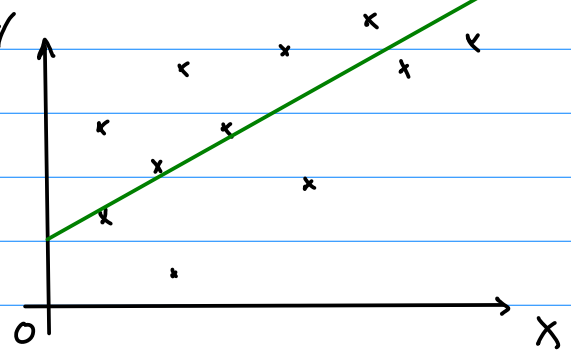
(Prop) $$X({}^tXX)^{-1}{}^tXY = X\hat{\beta}_n$$

(injectivité de X) $$({}^tXX)^{-1}{}^tXY = \hat{\beta}_n$$

et $$\hat{S}_n^2 = \frac{1}{n}\| Y - \hat{m}_n \|^2 = \frac{1}{n}\| Y - X\hat{\beta}_n \|^2$$ ∎

3 - Régression linéaire simple

On suppose qu'on a un vecteur Y de données à expliquer et un vecteur X de données explicatives



On cherche une relation affine entre X et Y,

c'est-à-dire $$Y_j = a + b X_j + \varepsilon_j$$

où (ε_j) iid de loi $\mathcal{N}(0, \sigma^2)$

Déterminer l'EMV de (a, b, σ^2)

Observons pour commencer que ce modèle statistique s'écrit $Y = \bar{X} \cdot \beta + \varepsilon$

où $\beta = \begin{pmatrix} a \\ b \end{pmatrix}$, $\bar{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim \mathcal{N}(0, \sigma^2 I_n)$

donc, par le théorème précédent, on a

$$\hat{\beta}_n = \left({}^t\bar{X}\,\bar{X}\right)^{-1} {}^t\bar{X}\, Y = \begin{pmatrix} \hat{a}_n \\ \hat{b}_n \end{pmatrix}$$

avec $${}^t\bar{X}\,\bar{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x}_n \\ \bar{x}_n & \overline{x^2}_n \end{pmatrix}$$

$$\left({}^t\bar{X}\,\bar{X}\right)^{-1} = \frac{1}{n\left(\overline{x^2}_n - \bar{x}_n^2\right)} \begin{pmatrix} \overline{x^2}_n & -\bar{x}_n \\ -\bar{x}_n & 1 \end{pmatrix}$$

$$\frac{X \cdot Y}{n} = \frac{1}{n}\sum X_i Y_i = \overline{XY_n}$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

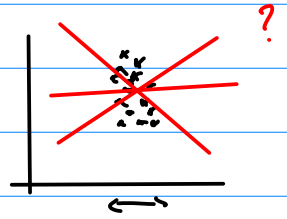
$${}^t\overline{X}\,Y = \begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{pmatrix}\begin{pmatrix} n\overline{Y}_n \\ X \cdot Y \end{pmatrix} = n\begin{pmatrix} \overline{Y}_n \\ \frac{X \cdot Y}{n} \end{pmatrix}$$

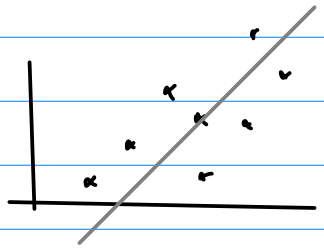
donc $$\hat{\beta}_n = \frac{1}{\overline{X}_n^2 - \overline{X}_n^2}\begin{pmatrix} \overline{X_n^2}\,\overline{Y}_n - \overline{X}_n\,\overline{XY}_n \\ \overline{XY}_n - \overline{X}_n\,\overline{Y}_n \end{pmatrix}$$

d'où $$\hat{a}_n = \frac{\overline{X_n^2}\,\overline{Y}_n - \overline{X}_n \times \overline{XY}_n}{\overline{X_n^2} - \overline{X}_n^2}$$

et $$\hat{b}_n = \frac{\overline{XY}_n - \overline{X}_n\,\overline{Y}_n}{\overline{X_n^2} - \overline{X}_n^2}$$

<u>Remarque</u> Le dénominateur de $\hat{a}_n$, $\hat{b}_n$ est $\hat{S}^2(X)$, ce dénominateur est d'autant plus grand que les données sont variées





On en déduit l'ETMV de σ^2 comme étant

$$\frac{1}{n}\|Y - \overline{X}\hat{\beta}_n\|^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{a}_n - \hat{b}_n X_i)^2$$

$\Rightarrow$ Cela revient à dire qu'on cherche les valeurs de a et b qui minimisent

$$\sum_{i=1}^{n}(Y_i - a - bX_i)^2$$

On retombe sur la méthode des moindres carrés

Proposition

L'estimation du maximum de vraisemblance du modèle

$$Y_j = a + b X_j + \varepsilon_j \quad , (\varepsilon_j) \text{ iid } \mathcal{N}(0, \sigma^2)$$

est donnée par $\hat{b}_n = \dfrac{\overline{Cov}(X,Y)}{\overline{Var}(X)}$, $\hat{a}_n = \overline{Y}_n - \hat{b}_n \overline{X}_n$

et $\hat{S}_n^2 = \dfrac{1}{n} \sum_{j=1}^{n} (Y_j - \hat{a}_n - X_j \hat{b}_n)^2$

où $\overline{Var}(x) = \dfrac{1}{n} \sum_{j=1}^{n} (x_j - \overline{x}_n)^2 = \overline{x^2_n} - \overline{x}_n^2$

et $\overline{Cov}(x,y) = \dfrac{1}{n} \sum (x_j - \overline{x}_n)(y_j - \overline{y}_n) = \overline{xy_n} - \overline{x}_n \overline{y}_n$

Remarque

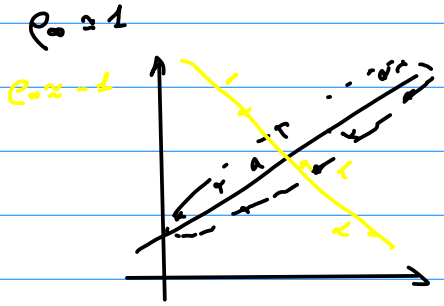
Le coefficient de corrélation de X et Y

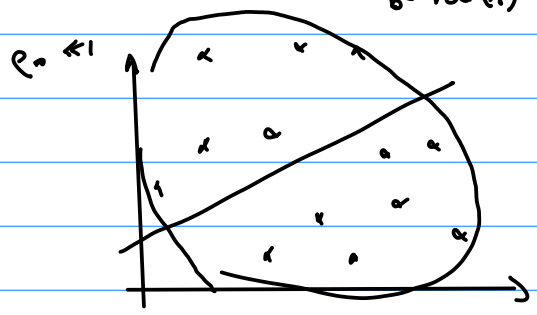
$$\rho_n = \frac{\overline{Cov}(X,Y)}{\sqrt{\overline{Var}(X)\,\overline{Var}(Y)}}$$

$$\rho_n \xrightarrow[n\to\infty]{} \rho_\infty = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} = \frac{Cov(X, a + bX + N)}{\sqrt{Var(X)\,Var(a+bX+N)}}$$

où $N \sim \mathcal{N}(0, \sigma^2) \perp\!\!\!\perp X$

$$\rho_\infty = \frac{b\,Var(X)}{\sqrt{Var(X)(b^2 Var\,X + \sigma^2)}} = \frac{sgn(b)}{\sqrt{1 + \dfrac{\sigma^2}{b^2\,Var(X)}}}$$





⚠ Coefficient de corrélation petit ne correspond pas exactement à relation linéaire inexistante

Test de relation linéaire $Y = a + bX + \varepsilon$

On cherche à tester si l'ordonnée à l'origine a est nulle ou non.

$H_0 = \{ a = 0 \}$

$H_1 = \{ a \neq 0 \}$

Estimation du MLG

l'hypothèse alternative est l'existence de données supplémentaires à prendre en compte

L'estimation de a est (par Théorème d'estimation matriciel)

$$\hat{a}_n = \frac{\overline{X_n^2}\,\overline{Y_n} - \overline{X_n} \times \overline{XY_n}}{\overline{X_n^2} - \overline{X_n}^2} \sim \mathcal{N}\left(a, \frac{\sigma^2 \overline{X_n^2}}{n\left(\overline{X_n^2} - \overline{X_n}^2\right)}\right)$$

"preuve"

$$\hat{a}_n = \frac{1}{n \overline{Var}(X)} \sum_{j=1}^{n} \overline{X_n^2} Y_j - \overline{X}_n X_j Y_j$$

est combinaison linéaire de coordonnées du vecteur gaussien Y

$$\hat{a}_n = \frac{1}{n \overline{Var}(X)} \sum_{j=1}^{n} \left(a + b X_j + \varepsilon_j\right)\left(\overline{X_n^2} - \overline{X}_n X_i\right)$$

$$= \underbrace{\frac{1}{n \overline{Var}(X)} \sum_{j=1}^{n} \left(a + b X_j\right)\left(\overline{X_n^2} - \overline{X}_n X_j\right)}_{c} + \underbrace{\frac{1}{n \overline{Var} X} \sum_{j=1}^{n} \varepsilon_j \left(\overline{X_n^2} - X_j \overline{X}\right)}_{Z}$$

avec $c = \frac{1}{n \overline{Var}(X)} \left(an\left(\overline{X_n^2} - \overline{X_n}^2\right) + bn\underbrace{\left(\overline{X_n^2}\,\overline{X}_n - \overline{X}_n \overline{X_n^2}\right)}_{0} \right) = a$

et $Z \sim \mathcal{N}\left(0, \frac{1}{n^2 \overline{Var}(X)^2} \sigma^2 \sum_{j=1}^{n} \overline{X}_n^4 - 2 X_j \overline{X}_n \overline{X_n^2} + X_j^2 \overline{X_n^2}\right) \sim \mathcal{N}\left(0, \frac{\sigma^2 \overline{X_n^2}}{n\left(\overline{X_n^2} - \overline{X_n}^2\right)}\right)$

□

Sous H_0, $\hat{a}_n \sim \mathcal{N}\left(0, \frac{\sigma^2 \overline{X_n^2}}{n \overline{Var_n}(X)}\right)$

$\hat{a}_n$ n'est pas une statistique libre, car sa loi dépend de σ^2

On construit alors une statistique libre avec $\hat{S}_n^2$

rappelons que $\hat{a}_n \perp\!\!\!\perp \hat{S}_n^2$ et $\frac{n \hat{S}_n^2}{\sigma^2} \sim \chi^2(n-2)$

$\frac{\mathcal{N}(0,1)}{\sqrt{\chi^2(d)/d}} \sim \mathcal{C}(d)$

$$\frac{\hat{a}_n \times \sqrt{\frac{n \overline{Var_n}(X)}{\sigma^2 \overline{X_n^2}}}}{\sqrt{\frac{n}{\sigma^2}\hat{S}_n^2/n-2}} \sim \mathcal{C}(n-2)$$

$$\sqrt{\frac{(n-2)\overline{Var_n}(X)}{\overline{X_n^2}}} \frac{\hat{a}_n}{\sqrt{\hat{S}_n^2}} \sim \mathcal{C}(n-2)$$ est une statistique libre

On construit donc la statistique de test de niveau α de H_1 contre H_0 par

$$T = \mathbb{1}_{\left| \sqrt{\frac{(n-2)\overline{Var_n}(X)}{\overline{X_n^2}}} \frac{\hat{a}_n}{\sqrt{\hat{S}_n^2}} \right| \geq s}$$

en choisissant s tel que $\mathbb{P}_{H_0}(T = 1) = 1 - \alpha$

i.e. $\mathbb{P}_{H_0}\left(\left| \sqrt{\frac{(n-2)\overline{Var_n}(X)}{\overline{X_n^2}}} \frac{\hat{a}_n}{\sqrt{\hat{S}_n^2}} \right| \geq s \right) = 1 - \alpha$

$(n-2)$

on prend $s = t_{1 - \frac{1-\alpha}{2}}$

et on a $T = \mathbb{1}_{\left| \frac{\hat{a}_n}{\sqrt{\hat{S}_n^2}} \right| \geq t_{\frac{1+\alpha}{2}}^{(n-2)} \sqrt{\frac{\overline{X_n^2}}{(n-2)\overline{Var_n}(X)}}}$

on a toujours intérêt à choisir $\overline{Var_n}(X) \gg 1$

Test de significativité de X $\quad Y = a + bX + c$

$H_0 = \{ b = 0 \}$ contre $H_1 = \{ b \neq 0 \}$

Ce qu'on teste, c'est si X est effectivement une variable explicative du modèle

$$\hat{b}_n = \frac{\overline{X_n^2} \times \overline{XY_n} - \overline{X_n}\,\overline{Y_n}}{\overline{X_n^2} - \overline{X_n}^2} \sim \mathcal{N}\left(b, \frac{\sigma^2}{n} \frac{1}{\overline{X_n^2} - \overline{X_n}^2} \right)$$

Construisons le test de niveau $1-\alpha$ de H_1 contre H_0

Sous H_0, on a $\hat{b}_n \sim \mathcal{N}\left(0, \frac{\sigma^2}{n} \frac{1}{\overline{X_n^2} - \overline{X_n}^2}\right)$

On observe que $\dfrac{\hat{b}_n \sqrt{\frac{n\,\mathrm{Var}(X)}{\sigma^2}}}{\sqrt{\frac{n\hat{S}_n^2}{\sigma^2} \frac{1}{n-2}}} \sim \mathcal{T}(n-2)$

$\mathcal{N}(0,1)$ au numérateur ; $\chi^2(n-2)$ sous la racine au dénominateur

La statistique de test est donc

$$\mathbb{1}\left\{ \sqrt{(n-2)\mathrm{Var}(X)} \left| \frac{\hat{b}_n}{\sqrt{\hat{S}_n^2}} \right| \geq t_{\frac{1+\alpha}{2}}^{(n-2)} \right\}$$

qu'on peut réécrire $\mathbb{1}\left\{ \left| \dfrac{\overline{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\hat{S}_n^2}} \right| \geq \dfrac{t_{\frac{1+\alpha}{2}}^{(n-2)}}{\sqrt{n-2}} \right\}$

coefficient de corrélation de X et Y

Prédiction pour une nouvelle donnée $Y = a + bx + \varepsilon$

avec les estimateurs $\hat{a}_n$, $\hat{b}_n$ et $\hat{S}_n^2$ calculés

on récupère une nouvelle donnée x, et on souhaite donner un intervalle de confiance sur la quantité y associée

un estimateur naturel de y est

$$\hat{y}_n = \hat{a}_n + \hat{b}_n x$$

de plus
$$y - \hat{y}_n = a + bx + \varepsilon - \hat{a}_n - \hat{b}_n x$$
$$= \underbrace{a - \hat{a}_n}_{\text{erreur sur } a} + \underbrace{(b - \hat{b}_n)}_{\text{erreur sur } b} x + \varepsilon$$
$$\sim \mathcal{N}\left(0,\ \sigma^2 \left(1 + \frac{(x - \bar{x}_n)^2}{n \widehat{\mathrm{Var}}(x)}\right)\right)$$

On montre que

$$\frac{\hat{y}_n - y}{\sqrt{\hat{S}_n^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x}_n)^2}{\widehat{\mathrm{Var}}(x)}\right)}} \sim \mathcal{T}(n-2)$$

donc un $IC_{1-\alpha}(y) = \left[\hat{y}_n \pm t_{1-\frac{\alpha}{2}}^{(n-2)} \sqrt{\hat{S}_n^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x}_n)^2}{\widehat{\mathrm{Var}}(x)}\right)}\right]$

Exercice Vérifier les calculs ci-dessus

4 - Régions de confiance et tests fondamentaux

a- Intervalle de confiance et tests pour la variance

Sous le modèle linéaire gaussien, l'estimateur du maximum de vraisemblance de σ^2 est

$$\hat{S}_n^2 = \frac{1}{n} \| Y - \Pi_V Y \|_2^2 = \frac{1}{n} \| Y - X\hat{\beta}_n \|^2$$

et une version débiaisée de cet estimateur est

$$\hat{\sigma}_n^2 = \frac{1}{n - \dim(V)} \| Y - \Pi_V Y \|_2^2 = \frac{1}{n - \mathrm{Rg}(X)} \| Y - X\hat{\beta}_n \|^2$$

On sait que $\hat{\sigma}_n^2 \perp\!\!\!\perp (\hat{m}_n, \hat{\beta}_n)$

et que $$\frac{n-d}{\sigma^2} \hat{\sigma}_n^2 \sim \chi^2(n-d)$$

où $d = \dim(V) = \mathrm{Rg}(X)$

Par conséquent, l'intervalle de confiance unilatère à droite de niveau $1-\alpha$ pour σ^2 est

$$IC_{1-\alpha}(\sigma^2) = \left[0, \frac{(n-d)}{\chi^2_\alpha (n-d)} \hat{\sigma}_n^2 \right[$$

Une statistique de test pour $H_0 : \{\sigma > \sigma_{ref}\}$ contre $H_1 : \{\sigma \leq \sigma_{ref}\}$ est donnée par

$$T = \mathbb{1}_{\hat{\sigma}^2 < s} \quad \text{avec } s \text{ tel que}$$

$$\sup_{m, \sigma^2 \in H_0} \mathbb{P}_{m,\sigma^2}(T = 1) < \alpha$$

$$= \mathbb{P}_{m, \sigma^2_{ref}}(T = 1) < \alpha$$

d'où on obtient le test de niveau α

$$T = \mathbb{1}\left\{ \hat{\sigma}_n^2 \leq \frac{\sigma_{ref}^2}{n-p} \chi_\alpha^{2\,(n-p)} \right\}$$

En effet $\mathbb{P}_{m,\sigma_{ref}^2}\left(\hat{\sigma}_n^2 \leq s \right) = \mathbb{P}_{m,\sigma_{ref}^2}\left(\frac{\hat{\sigma}_n^2 (n-d)}{\sigma_{ref}^2} \leq \frac{s\,(n-d)}{\sigma_{ref}^2} \right)$

$$= \mathbb{P}\left(\chi^2(n-d) \leq \frac{s\,(n-d)}{\sigma_{ref}^2} \right) = \alpha$$

$$\rightarrow \frac{s\,(n-d)}{\sigma_{ref}^2} = \chi_\alpha^{2\,(n-d)} \Rightarrow s = \frac{\sigma_{ref}^2}{n-d} \chi_\alpha^{2\,(n-d)}$$

b. Test d'une relation affine

On se place dans le cadre du modèle matriciel

Pour $c \in \mathbb{R}^d$ et $a \in \mathbb{R}$, on cherche à tester l'hypothèse $H_0 : \{ c.\beta = a \}$ contre $H_1 : \{ c.\beta \neq a \}$

où $c.\beta = c_1\beta_1 + c_2\beta_2 + \dots + c_d\beta_d$

Exemple

On peut souhaiter tester $\beta_1 = \beta_2$

On peut souhaiter tester l'hypothèse $\beta_j = 0$, c'est à dire la significativité de la donnée X_j.

La statistique de test naturelle à considérer est la variable $T_n = \dfrac{c.\hat{\beta}_n - a}{\sqrt{\hat{\sigma}_n^2\, {}^t c ({}^t X X)^{-1} c}}$

En effet, sous H_0, $c.\beta_n - a \sim \mathcal{N}\left(\underbrace{c.\beta - a}_{0}, \sigma^2\, {}^t c ({}^t X X)^{-1} c \right)$

En utilisant que $\frac{(n-d)\hat{\sigma}_n^2}{\sigma^2} \sim \chi^2(n-d)$ $\perp\!\!\!\perp$ $c\hat{\beta}_n - a$

donc $$\frac{c\cdot\hat{\beta}_n - a}{\sqrt{\hat{\sigma}_n^2\, t_c({}^tX X)^{-1} c}} = \frac{c\cdot\hat{\beta}_n - a}{\sqrt{\sigma^2\, t_c({}^tX X)^{-1} c}} \cdot \frac{1}{\sqrt{\frac{(n-d)\hat{\sigma}_n^2}{\sigma^2}/n-d}}$$

$\mathcal{N}(0,1)$ $\qquad$ $\chi^2(n-d)$

$$\sim \mathcal{C}(n-d)$$

Donc T_n est une statistique libre du modèle

On propose donc le test de niveau α de H_0 contre H_1

$$\text{ll} \quad |T_n| > t_{1-\frac{\alpha}{2}}^{(n-d)}$$

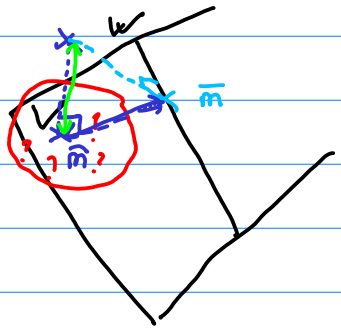
car $$\mathbb{P}_{H_0}\left(|T_n| > t_{1-\frac{\alpha}{2}}^{(n-d)}\right) = \mathbb{P}\left(T_n > t_{1-\frac{\alpha}{2}}^{(n-d)}\right) + \mathbb{P}\left(T_n < -t_{1-\frac{\alpha}{2}}^{(n-d)}\right) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$$

Avec les mêmes méthodes, on construit

$$IC_{1-\alpha}(c\cdot\beta) = \left[c\cdot\hat{\beta}_n \pm t_{1-\frac{\alpha}{2}}^{(n-d)}\ \hat{\sigma}_n^2 \sqrt{t_c\left({}^tX X\right)^{-1} c}\right]$$

c. Test de Fisher d'un sous-modèle

On se place dans le modèle vectoriel $Y = m + \varepsilon$
où $m \in V \subset \mathbb{R}^n$ et $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$
On se donne un sous-espace vectoriel W de V
et on souhaite tester $H_0 : \{m \in W\}$ contre $H_1 : \{m \notin W\}$



$$\frac{\|\cdot\|^2}{\sigma^2} \sim \chi^2(n - \dim V)$$

$$\frac{\|\cdot\|^2}{\sigma^2} \sim \chi^2(\dim V - \dim W)$$

Définition

La loi de Fisher à k et ℓ degrés de liberté notée $F(k, \ell)$ est la loi de $Z = \frac{U/k}{V/\ell}$, où

$U \sim \chi^2(k)$ $\quad V \sim \chi^2(\ell)$ et $U \perp\!\!\!\perp V$

Remarque

On a $\lim_{\ell \to \infty} Z_{k,\ell} = U/k$

donc pour $\ell \gg 1$, $F(k, \ell) \approx \frac{\chi^2(k)}{k}$

Théorème

$d = \dim(V)$ $\quad d' = \dim(W)$

Lorsque $m \in W$, la statistique

$$F = \frac{\|\Pi_W Y - \Pi_V Y\|^2/(d-d')}{\|Y - \Pi_V Y\|^2/(n-d)} \sim \mathcal{F}(d-d', n-d)$$ et est indépendante de $\Pi_W Y$

preuve Par théorème de Cochan

on a $\Pi_W Y$, $\Pi_{W^{\perp V}} Y$ et $\Pi_{V^\perp} Y$ qui

sont des v.a. gaussiens indépendants

de plus $\Pi_W Y \sim \mathcal{N}(m, \sigma^2 I_W)$ car $\Pi_W {}^t\Pi_W$

$\Pi_{W^{\perp V}} Y \sim \mathcal{N}(0, \sigma^2 I_{W^{\perp V}})$

$\Pi_{V^\perp} Y \sim \mathcal{N}(0, \sigma^2 I_{V^\perp})$

Or $\Pi_V Y - \Pi_W Y = \Pi_{W^{\perp V}} Y$

$Y - \Pi_V Y = \Pi_{V^\perp} Y$

Donc $F = \dfrac{\| \Pi_{W^{\perp V}} Y \|^2 / \dim(W^{\perp V})}{\| \Pi_{V^\perp} Y \|^2 / \dim(V^\perp)} \sim \mathcal{F}\left(\dim(W^{\perp V}), \dim(V^\perp)\right)$

et est $\perp\!\!\!\perp$ de $\Pi_W Y$ □

On en déduit le test suivant de niveau α
de $H_0 = \{ m \in W \}$ contre $H_1 = \{ m \in V \setminus W \}$

$$\mathbb{1}_{F > \mathcal{F}_{1-\alpha}^{(d-d', n-d)}}$$

Remarque

Si $d - d' = 1$, i.e. W est un hyperplan de V
on retrouve le test de Student bilatéral

$$T = \frac{N}{\sqrt{K_{kk}/k}} \quad \text{donc} \quad T^2 = \frac{N^2}{K_{kk}/k} \sim \mathcal{F}(1, k)$$

d. Test de Wald de plusieurs relations affines

Considérons un MLG $Y = X.\beta + \varepsilon$

On teste l'hypothèse $H_0 = \{ c.\beta = a \}$

contre H_1

où $c \in M_{d',d}$ injective et $a \in \mathbb{R}^{d'}$

Exemple $c = \begin{bmatrix} -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ et $a = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

dans ce cas on teste sous H_0 si $2\beta_1 = \beta_2$ et $\beta_3 = 0$

Rappelons que ${}^tXX$ est symétrique définie positive on va noter $T^2 = ({}^tXX)^{-1}$

donc $c\,({}^tXX)^{-1}\,{}^tc = {}^t(T\,{}^tc)(T\,{}^tc)$ est symétrique définie positive

où $W = \{\beta : c\beta = a\}$

$\| \Pi_W Y - \underbrace{\Pi_V Y}_{X\hat\beta_n} \|^2$

Théorème

Sous H_0, $W = \dfrac{[{}^t(c.\hat\beta_n - a)\, c\, ({}^tXX)^{-1}\, {}^tc\, (c.\hat\beta_n - a)]/d'}{\| Y - X\hat\beta_n \|^2/n-d}$

suit une loi $F(d', n-d)$ degrés de liberté

preuve

Sous H_0, $\hat{\beta}_n \sim \mathcal{N}\left(\beta, \sigma^2 ({}^tXX)^{-1}\right)$

donc $c\hat{\beta}_n - a \sim \mathcal{N}\left(\underbrace{c\beta - a}_{0}, \sigma^2 c ({}^tXX)^{-1} {}^tc\right)$

or $c ({}^tXX)^{-1} {}^tc$ est symétrique définie positive

il existe Δ telle que $\Delta^2 = c ({}^tXX)^{-1} {}^tc$

on obtient donc $\Delta^{-1}(c\hat{\beta}_n - a) \sim \mathcal{N}(0, \sigma^2 I_{d'})$

$$\frac{1}{\sigma^2} {}^t(c\hat{\beta}_n - a)\, c ({}^tXX)^{-1} {}^tc\, (c\hat{\beta}_n - a) = \frac{1}{\sigma^2} \| \Delta^{-1}(c\hat{\beta}_n - a) \|^2$$

$$\sim \chi^2(k)$$

$\hat{\beta}_n$ étant indépendant de $\hat{\sigma}_n^2 = \|Y - X\hat{\beta}_n\|^2 / n$

on obtient

$$\frac{\frac{1}{\sigma^2} {}^t(c\hat{\beta}_n - a)\, c ({}^tXX)^{-1} {}^tc\, (c\hat{\beta}_n - a)/d'}{\frac{1}{\sigma^2} \|Y - X\hat{\beta}_n\|^2/(n-d)}$$

$$\sim F(d', n-d)$$

Le test de $H_0 : \{c\beta = a\}$ contre H_1 s'écrit alors

$$\| \{ W \geq f_{1-\alpha}^{(d', n-d)} \}$$

c'est un test de niveau α

Le test de Wald permet la création de régions de confiance pour $\hat{\beta}_n$

Pour tout c matrice injective, on peut définir

$$\mathcal{E}_c = \left\{ a \in \mathbb{R}^{d'} : \frac{{}^t(c\hat{\beta}_n - a)\, c\, ({}^tXX)^{-1}\, {}^tc\, (c\hat{\beta}_n - a)/d'}{\|Y - X\hat{\beta}_n\|^2/(n-d)} \leq f_{\ldots} \right]$$

est une région de confiance de niveau exactement $1-\alpha$ pour $c.\beta$

$\mathcal{E}_c$ est un ellipsoïde centré en $c\,\hat{\beta}_n$

En particulier $c = I_d$

$$\mathcal{E} = \left\{ a \in \mathbb{R}^k : \frac{{}^t(\hat{\beta}_n - a)\, ({}^tXX)^{-1}\, (\hat{\beta}_n - a)/d}{\|Y - X\hat{\beta}_n\|^2/(n-d)} \leq f^{(d,n-d)}_{1-\alpha} \right\}$$

est une région de confiance de β

5 - Variables quantitatives : La régression linéaire multiple

On suppose qu'on souhaite expliquer une variable Y à partir de plusieurs variables quantitatives $X_1, \dots X_d$

Définition

Une variable quantitative est une variable dont la valeur numérique a un sens : taille, aire, volume, température,

Dans cette situation, on pose

$$Y_i = \mu + \beta_1 X_1 + \dots + \beta_d X_d$$

donc les paramètres sont $(\mu, \beta_1, \dots \beta_d) \in \mathbb{R}^{d+1}$

La matrice des données $X = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,d} \\ 1 & & & \\ \vdots & & & \\ 1 & x_{n,1} & \dots & x_{n,d} \end{pmatrix}$

On appelle $(X_1, \dots X_d)$ les régresseurs et $(\mu, \beta_1, \dots \beta_d)$ les coefficients de régression.

Par exemple, on peut utiliser un test de Student pour déterminer la significativité de la j^e variable X_j

$$H_0 = \{ \beta_j = 0 \} \quad \text{contre} \quad H_1$$

On se fixe $d' \leq d$, et on teste

$H_0 = \{ \beta_1 = \dots = \beta_{d'} = 0 \}$ contre H_1 : il existe un coefficient non-nul

$$C = \begin{pmatrix} 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{pmatrix} \qquad a = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$X_{rep} = \begin{pmatrix} 1 & X_{1,d'+1} & \dots & X_{1,d} \\ 1 & & & \\ \vdots & \vdots & & \\ 1 & X_{n,d'+1} & \dots & X_{n,d} \end{pmatrix}$$

La matrice des données significatives sous H_0

On pose $V = \{ X.\beta \,,\, \beta \in \mathbb{R}^d \} = \mathrm{Im}(X)$

$W = \{ X_{rep}.\beta \,,\, \beta \in \mathbb{R}^{d-d'} \} = \mathrm{Im}(X_{rep})$

$H_0 = \{ X.\beta \in W \}$ $\qquad$ $H_1 = \{ X.\beta \notin W \}$

donc la statistique de test sera

$$F = \frac{\| \Pi_V Y - \Pi_W Y \|^2 / d'}{\| Y - \Pi_V Y \| / n-d}$$

$$= \frac{\| X({}^tXX)^{-1}{}^tXY - X_{rep}({}^tX_{rep}X_{rep})^{-1}{}^tX_{rep}Y \|^2 / d'}{\| Y - X({}^tXX)^{-1}{}^tXY \|^2 / n-d} \sim \mathcal{F}(d', n-d)$$

On construit le test $\mathbb{1}\{ F \geq q_{1-\alpha}^{(d',n-d)} \}$

6- Application au cas des variables qualitatives : l'analyse de variance ANOVA

On s'intéresse à un modèle linéaire gaussien dont les données sont des variable qualitatives avec un nombre fini de modalités

Définition

Une variable qualitative est une variable ne permettant pas de traduction chiffrée immédiate. On appelle modalités l'ensemble des valeurs possible de cette variable

a. ANOVA à un facteur

On considère un modèle linéaire gaussien où X prends des valeurs dans un ensemble I avec $|I| \geq 2$

Le modèle d'analyse de variance s'écrit

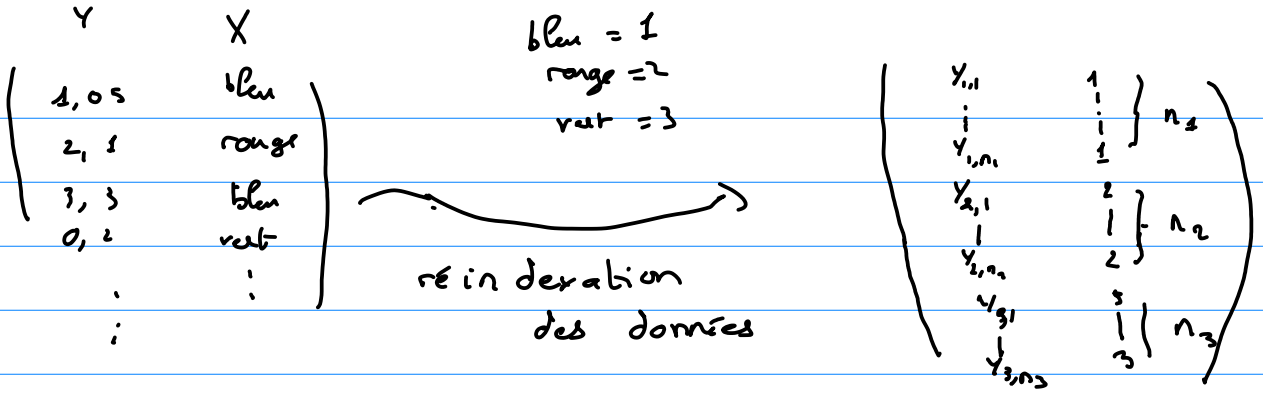
$$Y_k = \mu + a_i + \varepsilon_k \quad \text{si} \quad X_k = i$$

où $\varepsilon_k \sim \mathcal{N}(0, \sigma^2)$ i.i.d

Quitte à renuméroter, on exprime le modèle sous la forme

$$\begin{cases} Y_{i,k} = \mu + a_i + \varepsilon_{i,k} & \varepsilon_{i,k} \sim \mathcal{N}(0,\sigma^2) \text{ i.i.d} \\ i \in \{1,2,\dots |I|\} \quad k \in \{1,\dots n_i\} \end{cases}$$

n_i : nombre de données avec la modalité i et $n = n_1 + \dots + n_I$



La matrice des données de ce GLG s'écrit

$$X = \begin{matrix} n_1 \updownarrow \\ \\ \\ \\ \\ \\ n_I \updownarrow \end{matrix} \begin{pmatrix} 1 & 1 & 0 & \\ \vdots & \vdots & \vdots & \ddots \\ 1 & 0 & 0 & \\ & & & \\ & & & \\ & & & \end{pmatrix} \quad \text{et} \quad \beta = \begin{pmatrix} \mu \\ a_1 \\ a_2 \\ \vdots \\ a_I \end{pmatrix}$$

cette matrice n'est pas injective, le modèle est surparamétré. On imposera donc $\mu = 0$, pour retrouver une formulation matricielle

$$\begin{cases} Y_{i,k} = a_i + \varepsilon_{i,k} & \varepsilon_{i,k} \text{ iid } \mathcal{N}(0, \sigma^2) \\ i \leq I, \quad k \leq n_i \end{cases}$$

Les estimateurs de $\beta = (a_1, \dots a_I)$ et σ^2

$$^tX X = \begin{pmatrix} n_1 & & 0 \\ & n_2 & \\ 0 & & \ddots \\ & & & n_I \end{pmatrix}$$

donc $\hat{\beta}_n = \left(\frac{\overline{Y}_i}{n_i}, \ 1 \leq i \leq I \right)$

où $\overline{Y}_i = \frac{Y_{i,1} + \dots Y_{i,n_i}}{n_i}$

$$\hat{\beta}_n \sim \mathcal{N}\left(\beta, \sigma^2 \begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n_I} \end{pmatrix} \right)$$

$$\hat{\sigma}_n^2 = \frac{1}{n - I} \sum_{i=1}^{I} \sum_{k=1}^{n_I} \left(Y_{i,k} - \overline{Y}_i \right)^2$$

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

Le test statistique classique est le test de pertinence de la factorisation

$$H_0 = \{ a_1 = a_2 = \dots = a_I \} \quad \text{contre} \quad H_1 := \{ \exists i \neq j \text{ avec } a_i \neq a_j \}$$

On pose c la projection orthogonale sur ${}^t(1, 1 \dots 1)$

$$H_0 = \{ c\beta = 0 \} \quad \text{contre} \quad H_1$$

Sous H_0, la projection de Y sur ${}^t(1, \dots 1)$ est $\frac{1}{n} \sum_{i=1}^{I} \sum_{k=}^{n_i} Y_{i,k} = \overline{Y}_{\cdot}$.

Dans ce cas, on a

$$F = \frac{\| X ({}^tX X)^{-1} {}^tX Y - X_0 ({}^tX_0 X_0)^{-1} {}^tX_0 Y \|^2 / (I-1)}{\| Y - X ({}^tX X)^{-1} {}^tX Y \|^2 / (n-I)}$$

$$= \frac{n-I}{I-1} \frac{\sum_{i=1}^{I} n_i (\overline{Y}_i - \overline{Y}_0)^2}{\sum_{i=1}^{I} \sum_{k=1}^{n_i} (Y_{i,k} - Y_{i,\cdot})^2}$$

qui suit sous H_0 la loi $F(I-1, n-I)$

Le test est alors $\mathbb{1}_{F \geq \delta}$

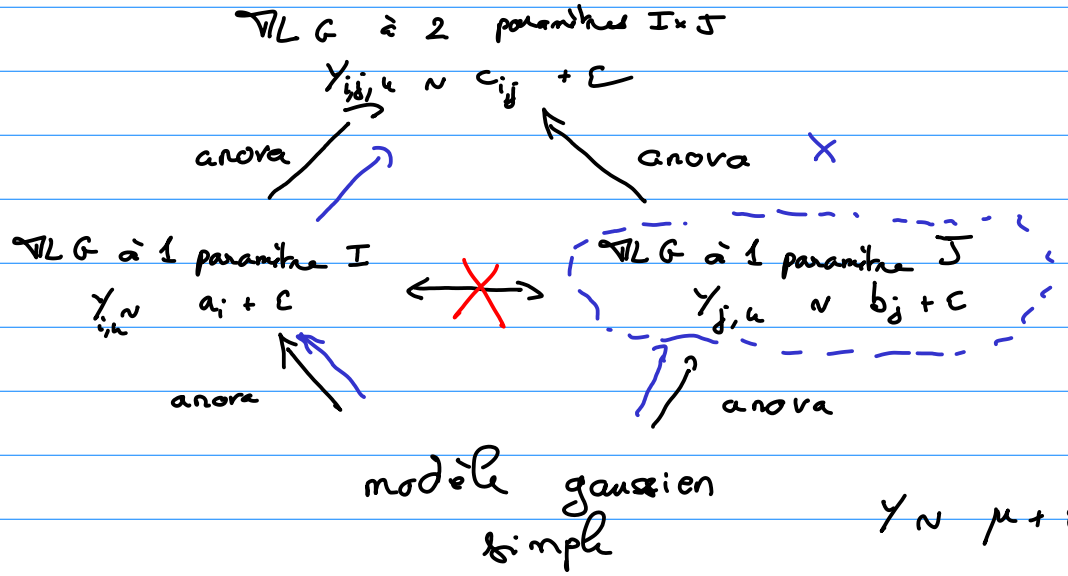
b. Extension à l'ANOVA à deux facteurs

Si on a plusieurs modalités, on écrit le modèle sous la forme $I \times J$

$$\begin{cases} Y_{i,j,u} = a_{i,j} + \varepsilon_{i,j,k} \\ 1 \leq i \leq I, \ 1 \leq j \leq J \quad \text{et} \quad k \leq n_{i,j} \end{cases}$$

On peut se poser la question

$H_0 = \{ a_{i,1} = a_{i,2} = \dots a_{i,I} \text{ pour tout } 1 \le i \le I \}$



modèle gaussien simple $\quad Y \sim \mu + \varepsilon$

7 - Analyse des résidus

Quand on étudie un modèle linéaire gaussien, on fait l'hypothèse que $Y \sim \mathcal{N}(m, \sigma^2 I_n)$

Il est souvent important de tester cette hypothèse par l'analyse des résidus

$$\hat{\varepsilon} = Y - \hat{m}_\lambda = Y - X\hat{\beta}_\lambda$$

- $\mathbb{E}(\hat{\varepsilon}) = 0$? $\quad \bar{\hat{\varepsilon}}$ comparée à 0
- $\mathrm{Var}(\hat{\varepsilon}) = \sigma^2 \quad \widehat{\mathrm{Var}}(\hat{\varepsilon})$ comparé à $\hat{\sigma}^2_\lambda$
- Variables qualitatives tester l'indépendance de ε
- Loi de ε est normale ? $\quad$ q-q plot

III Estimation de densité

On a un n-échantillon $(X_1, \dots X_n)$ d'une variable aléatoire et on cherche à déterminer la loi de X

1. Consistance de la fonction de répartition empirique

On suppose que la loi de $(X_1, \dots X_n)$ admet une densité par rapport à la mesure de Lebesgue

On pose $F : x \mapsto \mathbb{P}(X_1 \leq x)$ leur fonction de répartition

Rappelons que F est continue, croissante et

$$\lim_{x \to \infty} F(x) = 1 \qquad \lim_{x \to -\infty} F(x) = 0$$

Définition

La fonction de répartition empirique de $X = (X_1, \dots X_n)$ est la v.a. définie par

$$F_n : x \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$$: La proportion de données inférieures ou égales à x

Remarque

F_n est une fonction en escaliers continue à droite avec des limites à gauche en tout point, dont les discontinuités sont aux points

$$X^{(1)} \leq X^{(2)} \leq \dots \leq X^{(n)}$$

où $(X^{(1)} \dots X^{(n)})$ est la réorganisation de $(X_1, \dots X_n)$ dans l'ordre croissant

$$F_n(x) = \frac{i}{n} \quad \text{pour} \quad X^{(i)} \leq x \leq X^{(i+1)}$$

Proposition

On a $\lim_{n \to \infty} F_n(x) = F(x)$ p.s. pour tout $x \in \mathbb{R}$

Autrement dit, $F_n(x)$ est un estimateur fortement consistant de $F(x)$ qui est asymptotiquement normal

$$\lim_{n \to \infty} \sqrt{n}\left(F_n(x) - F(x)\right) = \mathcal{N}\left(0, F(x)(1-F(x))\right) \text{ en loi}$$

preuve

Soit $x \in \mathbb{R}$, $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$

et $\left(\mathbb{1}_{X_i \leq x}, i \in \mathbb{N}\right)$ est une suite de v.a. iid

de loi Bernoulli $(F(x))$

Par LGN, on a $F_n(x) \to F(x)$ p.s.

Par TCL on a $\sqrt{n}(F_n(x) - F(x)) \rightsquigarrow \mathcal{N}(0, F(x)(1-F(x)))$ en loi

Corollaire Théorème de Glivenko-Cantelli

F_n converge vers F uniformément

preuve Théorème de Dini

Définition

Les quantiles empiriques de X sont les quantités définies par $F_n^{(-1)}(q) = \inf\{x \in \mathbb{R} : F_n(x) \geq q\}$

$= X^{(\lfloor nq \rfloor)}$

Propriété

Les quantiles empiriques sont des estimateurs fortement consistants des quantiles de la loi de X

2- Test d'ajustement à une loi ou à une famille de lois

a. Test d'ajustement à une loi unique

On teste l'hypothèse H_0 : { X a pour fdr F } contre H_1

On a un estimateur fortement consistant de la fdr de X, un test naturel est de la forme

$$h(X) = \mathbb{1}_{\|F_n - F\|_\infty > s}$$

On remarque que F_n est une fonction en escaliers

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

$$= \max_{1 \le i \le n} \max \left(\left| \frac{i-1}{n} - F(x^{(i)}) \right| , \left| \frac{i}{n} - F(x^{(i)}) \right| \right)$$

Sous H_1 $\quad \liminf \|F_n - F\|_\infty > 0$

sous H_0 $\quad \lim_{n \to +\infty} \|F_n - F\|_\infty = 0$

On souhaite fixer le seuil s tel que

$$\mathbb{P}_{H_0} \left(\|F_n - F\|_\infty > s \right) = 1 - \alpha$$

Théorème

La variable aléatoire $\|F_n - F\|_\infty$ ne dépend pas de la valeur de F (c'est une statistique libre)

preuve

Soit $(U_1, \dots U_n)$ n v.a. i.i.d de loi Uniforme sur $[0,1]$

$(F^{-1}(U_1), F^{-1}(U_2) \dots F^{-1}(U_n))$ est un n-échantillon de fonction de répartition F

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$$

Dès lors

$$\|F_n - F\|_\infty = \max_{1 \leq i \leq n} \max \left(\left| \frac{i-1}{n} - F\left(F^{-1}(U^{(i)})\right) \right|, \left| \frac{i}{n} - F(F^{-1}(U^{(i)})) \right| \right)$$

$$= \max_{1 \leq i \leq n} \max \left(\left| \frac{i-1}{n} - U^{(i)} \right|, \left| \frac{i}{n} - U^{(i)} \right| \right)$$

ne dépend pas de F ∎

Cela permet donc de définir des tables de la

Lemme Soit F_n la fdr empirique d'un échantillon de loi uniforme sur $[0,1]$

On a $\lim_{n \to \infty} (\sqrt{n}(F_n(x) - x), x \in [0,1]) = (B_x, x \in [0,1])$

où B est un pont brownien (mouvement brownien conditionné à $B_1 = 0$)

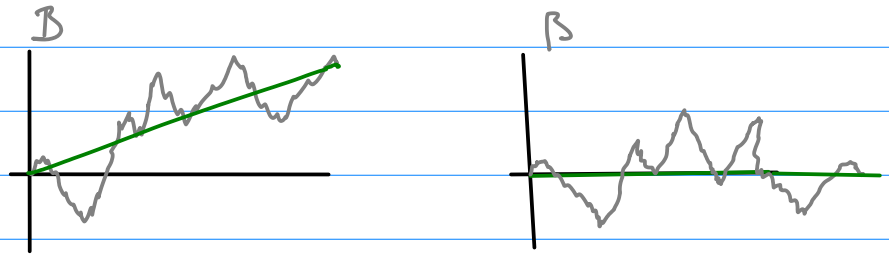
Corollaire $\lim_{n \to \infty} \mathbb{P}(\sqrt{n} \|F_n - F\|_\infty \geq \lambda) = \mathbb{P}(\sup_{x \in [0,1]} |B_x| \geq \lambda)$

$$= 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2\lambda^2}$$

Remarque

Soit $(B_t, t \geq 0)$ un mouvement Brownien

On pose $\beta_s = B_s - s B_1$ pour tout $s \in [0,1]$



Calculons $\mathrm{Cov}(B_s - s B_1, B_1)$

$$\mathrm{Cov}(B_s, B_1) - s\,\mathrm{Cov}(B_1, B_1) = s - s = 0$$

$\mathrm{Cov}(B_t, B_s) = \min(s,t)$ $\quad s \leq t$ $\quad B_t - B_s \perp\!\!\!\perp B_s$ $\quad B_t = B_s + B_t - B_s$

Donc comme $\mathrm{Cov}(\beta_s, B_1) = 0$, $\beta_s \perp\!\!\!\perp B_1$

Calculons $\mathrm{Cov}(\beta_s, \beta_u) = \mathrm{Cov}(B_s - s B_1, B_u - u B_1)$

$$= \mathrm{Cov}(B_s, B_u) - s\,\mathrm{Cov}(B_1, B_u) - u\,\mathrm{Cov}(B_s, B_1) + us\,\mathrm{Cov}(B_1, B_1)$$

$$= \min(s,u) - su.$$

Le pont brownien a même loi que $(\beta_s, s \in [0,1])$

et est un processus gaussien de matrice de covariance

$(\min(s,u) - su, \quad s, u \in [0,1])$

preuve du Lemme

Soit $(U_1, \dots U_n)$ un n-échantillon de loi uniforme sur $[0,1]$ et F_n la fonction de répartition empirique de cet échantillon

$$\left(\sqrt{n}\,(F_n(x) - x),\ x \in [0,1]\right) \xrightarrow[n\to\infty]{} (\beta_x,\ x \in [0,1])$$

Pour tout x fixé, $\sqrt{n}\,(F_n(x) - x) \xrightarrow[n\to\infty]{\text{en loi}} \mathcal{N}(0, x(1-x))$

et $x(1-x)$ est bien la variance de β_x

Soit $(x_1, \dots x_k) \in [0,1]^k$, supposons $x_1 < x_2 < \dots < x_k$

On montre $\left(\sqrt{n}\,(F_n(x_j) - x_j),\ j \le k\right) \xrightarrow[n\to\infty]{\text{en loi}} (\beta_{x_j},\ j \le k)$

par Théorème Central Limite multidimensionnel

$$\sqrt{n} \times \frac{1}{n} \sum_{p=1}^{n} \left(\begin{pmatrix} \mathbb{1}_{U_p \le x_1} \\ \mathbb{1}_{U_p \le x_2} \\ \vdots \\ \mathbb{1}_{U_p \le x_k} \end{pmatrix} - \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix} \right)$$

$$= \frac{1}{\sqrt{n}} \sum_{p=1}^{n} \begin{pmatrix} Y_1(p) \\ \vdots \\ Y_k(p) \end{pmatrix} - \begin{pmatrix} \mathbb{E}(Y_1(p)) \\ \vdots \\ \mathbb{E}(Y_k(p)) \end{pmatrix}$$

$$\xrightarrow[\text{loi}]{} \mathcal{N}(0, \Sigma)$$

où $\Sigma_{ij} = \mathrm{Cov}\left(Y_i(1), Y_j(1)\right)$

$$= \mathrm{Cov}\left(\mathbb{1}_{U_1 \le x_i}, \mathbb{1}_{U_1 \le x_j}\right)$$

$$= \mathbb{E}\left(\mathbb{1}_{U_1 \le x_i,\, U_1 \le x_j}\right) - \mathbb{E}\left(\mathbb{1}_{U_1 \le x_i}\right)\mathbb{E}\left(\mathbb{1}_{U_1 \le x_j}\right)$$

$$\Sigma_{i,j} = \min(x_i, x_j) - x_i x_j$$

Or la matrice de covariance de β est

$$\begin{pmatrix} \beta_{x_1} \\ \vdots \\ \beta_{x_n} \end{pmatrix} \sim \mathcal{N}(0, \Sigma)$$

On a donc bien $\left(\sqrt{n}\,(F_n(x_j) - x_j),\ j \le k\right) \to (\beta_{x_j},\ j \le k)$

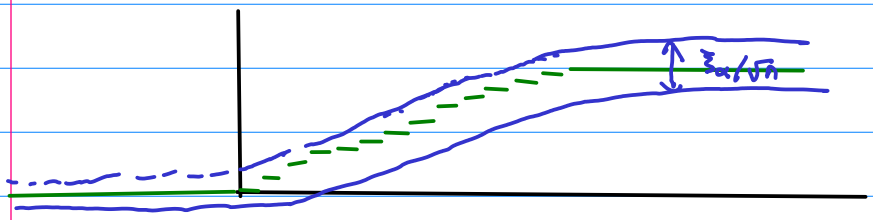
En prenant $k \gg 1$, $x_j = \frac{j}{k}$

Théorème Test de Kolmogorov

Soit $\alpha \in [0,1]$ on pose ξ_α le quantile d'ordre α de la variable $\sup_{s \in [0,1]} |B_s|$, le test $\mathbb{1}_{\|F_n - F\|_\infty > \frac{\xi_\alpha}{\sqrt{n}}}$

est un test asymptotique de niveau α de $H_0 = \{F \text{ est la fdr}\}$ contre H_1

De plus, on peut construire des intervalles de confiance pour la fonction de répartition F



<u>Remarque</u> Si X prend un nombre fini K de valeurs, on utilise un test du χ^2 pour tester l'adéquation à une loi.

Toutefois $\left(\sqrt{n}\left(F_n(j) - \sum_{i=1}^{j} p_i\right),\ j \le K\right) \to \left(\beta_{\sum_{i=1}^{j} p_i},\ j \le K\right)$

donc le test de Kolmogorov reste applicable

b. Adéquation à une famille de lois

$\mathbb{X} = (X_1, \dots X_n)$ un n-échantillon

On souhaite tester l'hypothèse, par exemple

$H_0 = \{ \mathbb{X}$ est de loi exponentielle$\}$

contre H_1

De façon générale, on test $H_0 = \{ \mathbb{X}$ est de loi $\in \mathbb{P}_\theta, \theta \in \Theta \}$

contre H_1

Une façon de réaliser ce test est de déterminer un estimateur $\hat{\theta}_n$ de θ, et d'utiliser la statistique

$$\| F_n - F_{\hat{\theta}_n} \|_\infty$$

Mais cette statistique n'est pas en général libre, on utilise des estimations asymptotiques ad-hoc.

Toutefois, il existe deux familles de lois pour lesquelles il est possible d'obtenir des statistiques libres

→ Ajustement à une famille de lois exponentielles

$$\mathcal{E} = \{ F_\lambda : x \longmapsto (1 - e^{-\lambda x}) \mathbb{1}_{x \geq 0}, \quad \lambda > 0 \}$$

L'estimateur du maximum de vraisemblance de λ

$$f(x_1, \dots x_n, \lambda) = \lambda^n e^{-\lambda (x_1 + \dots + x_n)}$$

$$\log f(x_1, \dots x_n; \lambda) = n \log \lambda - \lambda (x_1 + \dots + x_n)$$

est $\hat{\lambda}_n = \frac{1}{\bar{x}_n}$

Lemme

La variable $\sup_{x>0} |F_n(x) - F_{\frac{1}{\bar{X}_n}}(x)|$ est une statistique libre du paramètre λ

preuve Si $(X_1, \dots X_n)$ est un n-échantillon de loi $\text{Exp}(\lambda)$, on peut écrire $X_j = \frac{e_j}{\lambda}$ où $(e_1, \dots e_n)$ n-échantillon de loi $\text{Exp}(1)$

$$\bar{X}_n = \frac{\bar{e}_n}{\lambda}, \quad \text{et}$$

$$\sup_{x>0} |F_n(x) - F_{\frac{1}{\bar{X}_n}}(x)| = \max_{i \le n} \max \left| \frac{i-1}{n} - F_{\frac{1}{\bar{X}_n}}(X^{(i)}) \right|, \left| F_{\frac{1}{\bar{X}_n}}(X^{(i)}) - \frac{i}{n} \right|$$

$$= \max_{i \le n} \max \left| \frac{i-1}{n} - F_{\frac{\lambda}{\bar{e}_n}}\left(\frac{e^{(i)}}{\lambda}\right) \right|, \left| F_{\frac{\lambda}{\bar{e}_n}}\left(\frac{e^{(i)}}{\lambda}\right) - \frac{i}{n} \right|$$

or $F_{\frac{\lambda}{\bar{e}_n}}\left(\frac{e^{(i)}}{\lambda}\right) = 1 - e^{-\frac{\lambda}{\bar{e}_n}\left(\frac{e^{(i)}}{\lambda}\right)} = F_{\frac{1}{\bar{e}_n}}\left(e^{(i)}\right)$

$$= \max_{i \le n} \max \left| \frac{i-1}{n} - F_{\frac{1}{\bar{e}_n}}(e^{(i)}) \right|, \left| F_{\frac{1}{\bar{e}_n}}(e^{(i)}) - \frac{i}{n} \right|$$

qui ne dépend pas de λ. ∎

On peut donc construire des test exact d'adéquation à la famille de lois exponentielles

→ Test d'ajustement à la loi normale

$H_0 = \{(X_1, \dots X_n)$ n-échantillon de la loi normale$\}$ contre H_1

On a également la statistique libre donnée par

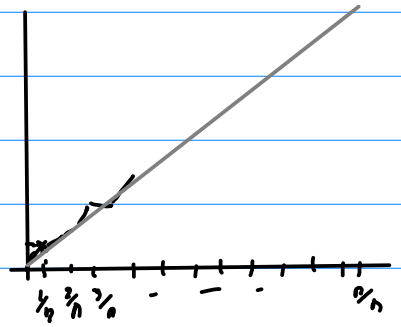
$$\sup_{x \in \mathbb{R}} \left| F_n(x) - \Phi\left(\frac{x - \bar{X}_n}{\sqrt{S_n^2}}\right) \right|$$

$$\| \Phi_{\bar{X}_n, \sqrt{S_n^2}}(x)$$

Ce test est très utilisé pour l'analyse des résidus d'un modèle linéaire gaussien

$(Y_j - (X\hat{\beta}_n)_j, j \leq n)$ suit un n-échantillon de loi normale si l'analyse du TLG est justifié

On représente ce test sous la forme d'un "q-q plot"



3- Test d'homogénéité de Kolmogorov - Smirnov

Ce test permet de vérifier que deux jeux de données suivent la même loi.

$(X_1, \dots X_n)$ et $(Y_1, \dots Y_m)$

de loi F de loi G

On souhaite tester $H_0 = \{F = G\}$ contre H_1

On définit la statistique $\mathbb{1}_{\|F_n - G_m\|_\infty > s}$

Cette statistique dépend uniquement de $Im(F)$ et $Im(G)$ donc si X et Y sont à densité par rapport à la mesure de Lebesgue, c'est une statistique libre.