

# Statistiques (12/12/2011)

## Ch1: Analyse bivariée

↳ Deux variables au plus, on les analyse.

↳ Distribution statistique des variables  
Dans une population (taille pooids)

↳ sur  $\mathbb{R}$

$$\hookrightarrow Z = (x, y) \text{ sur } \mathbb{R}$$

$x$  et  $y$  {  
↳ qualitatif  
↳ quantitatif  
↳ mixte (qualitatif & quantitatif)

Objectif: possibilité d'estimer une variable en utilisant l'autre variable  
→ chercher une variable en fonction d'autres variables

$Z(x_1, \dots, x_n)$ : multivariée:

- $x$ : variable à expliquer
- $x$ : variable explicative

Deux types:  
1. Exogénat: plusieurs variables explicatives étude seule à expliquer

$x_1, \dots, x_n$ : variable explicative  
 $\hookrightarrow y = f(x)$ : fonction d'expliquer  
en fonction de  $x$ : (peut en fait détailler)

modèle statistique  
pas forcément bijective  
l'existence de  $f$ :  
corrélation:  $y$  dépend de  $x$  si linéaire

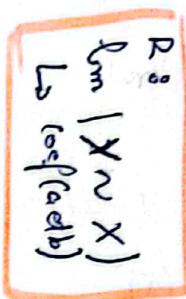
Selon le type de  $y$ : la corrélation devient

↳ Régression: si  $y$  quantitatif

↳ Classification: si  $y$  qualitatif  
• principe de la classification et de la régression  
• Corrélation et régression  
• En corrélation, on cherche le type de régression

→ Les modèles de régression:

↳ Régression linéaire simple (quantitatif)



$$\hookrightarrow y = a + bx$$

↳  $f$  est fonction bijective

↳  $y$  corrélé à  $x$  et  $x$  corrélé à  $y$

↳ Régression exponentielle:

$$Q: \ln(Z \sim X) \quad \hookrightarrow y = \ln(x)$$

$$\hookrightarrow y = \ln(x)$$

↳ Régression logarithmique:

$$\hookrightarrow x = \log(a + bx)$$

$$\hookrightarrow y = a + b \log(x)$$

↳ Régression polynomiale:

$$\hookrightarrow y = a + b_1 x + b_2 x^2 + \dots + b_n x^n$$

↳ Régression multiple:

Q:  $\ln$   
↳ Exogénat:  
Simple → multivarié

$$\hookrightarrow y = a + b_1 x_1 + \dots + b_n x_n$$

① ↳ Régression logistique

↳  $x$  qualitatif

$$\hookrightarrow x = a + bx \hookrightarrow 1$$

R<sub>0</sub>  
gln

## 2) expo

$$Z = \ln(x)$$

Dans R

$$\ln(Z \sim x)$$

$$Z = a + bX$$

$$\ln(Y) = a + bX$$

$$Y = e^a \cdot e^{bX}$$

## 3) log a:

$$\text{chgt: } Z = e^y \quad \text{su} \quad Z = \ln(x)$$

devenable

$$\ln(Z \sim x)$$

↓  
x non représenté  
de y

$$Z = a + bX$$

$$e^X = a + bX$$

$$X = \ln(a + bX)$$

$$\ln(Y, Z)$$

$$Z = \ln(X)$$

$$\ln(Z \sim X)$$

$$Y = \ln(Z)$$

$$X = a + bZ$$

$$X = a + b \ln(X)$$

"+", Dans R est (et, AND)

$$\ln(x) \quad X \sim x + x^2 + x^3$$

n'compte pas comme puissance

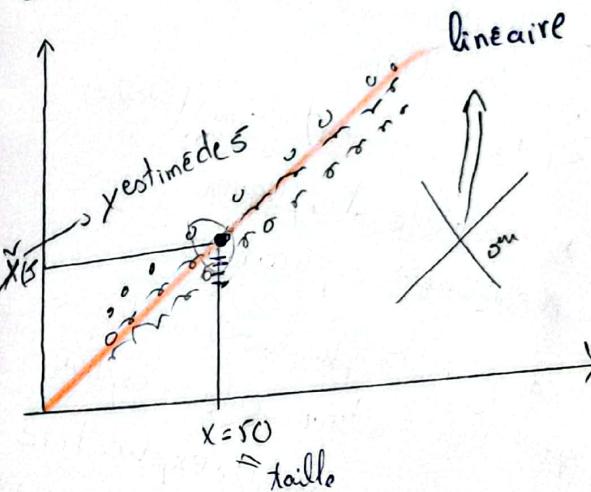
$$\ln(x) \quad X \sim x + I(x^2) + I(x^3)$$

• comment choisir le type de regression convenable

↳ calcul des coef

① ↳ su représentation graphique

② plot(X, Y) : nuage des pts de x en fonction de x



$$\begin{cases} X = a + b_1 X_1 + b_2 X_2 + \dots + b_n X^n \\ X = a + b_1 X_1 + b_2 X_2 + \dots + b_n X^n \\ \ln(X \sim X_1 + X_2) \quad \begin{cases} a, b_1, b_2 \\ Y = a + b_1 X_1 + b_2 X_2 \end{cases} \end{cases}$$

chgt devenable

$$X_1 = X, X_2 = X^2, X_3 = X^3$$

$$\ln(X \sim X + X^2 + X^3)$$



(2)  $y^{\text{d'ordre 2}} \propto P(\deg) = 3$

$R_{Y/X}$  = coeff de corrélat de  $y$  en  $X$

$\text{cor}(y, x) = \text{cor}(x, y)$  : coeff de

corrélation linéaire de  $x$  et  $y$

$R_{X/Y}$  : coeff de corrélat de  $x$  en  $y$

$R_{X/Y} \neq R_{Y/X}$  à ne pas confondre

$$R_{Y/X} = \frac{S_b^2(x)}{S_a^2(y)}$$

La déviance cond:  $S_w^2(x) + S_b^2(x)$

$$\tilde{S}(x) =$$

$$0 < R_{Y/X} < 1$$

Si  $R_{Y/X} = 1$  Il y a une corrélation

Si  $\approx 0$  pas de corrélat.

de  $y$  en  $x$

$\begin{cases} x \\ y \end{cases}$  sensibles

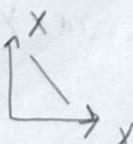
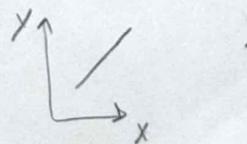
$\begin{cases} x \\ y \end{cases}$  pas fonctionnel de  $x$

$\begin{cases} x \\ y \end{cases}$  pas de corrélat.

$S_b^2(y)$  variance de  $y$  conditionné par  $x$

cas particulier de corrélation

La corrélation linéaire



$$y = a + bx$$
$$(y - a) \frac{1}{b} = x$$

$\approx -1 < \text{cor}(x, y) < 1$

inversement proportionnelle

$\approx 1$  proportionnelle à  $x$

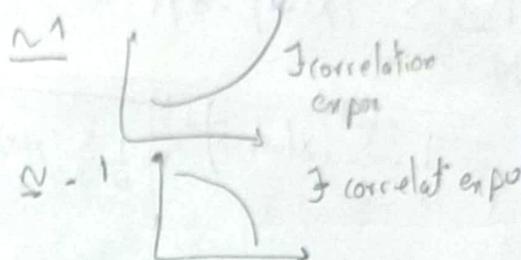
proche de pas de corrélation linéaire



• corré exp:  $y = \ln(x)$

$$-1 < \text{cor}(x, \ln(x)) < 1$$

Le coeff de corréat = exponentiel



$\text{cor}(x, x)$

Le en R:  $\text{cor}(x, y)$

Le manuellement:

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\sigma(x) \sigma(y)}$$

• covariance:

$$\text{cov}(x, y)$$

+  $\sigma(x)$ : écart type

• covariance: généralisation de l'avariance

Le moyen des produits (?) vocal

$$\text{cov}(x, y) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^n n_{ij} (x_i - \bar{x})(y_j - \bar{y})$$

définis

$$\text{cov}(x, y) = \frac{1}{N} \sum_{\omega} (x(\omega) - \bar{x})(y(\omega) - \bar{y})$$

$$(\text{cov}(x, y)) = \overline{xy} - \overline{x}\overline{y}$$

Relation

dist st: parabolétiques (vecteur)

(3)

$$x = (x_1, x_2, \dots, x_N) \in \mathbb{R}^N$$

Les sont des observations non pondérées  
modalités

$$x = (y_1, y_2, \dots, y_N) \in \mathbb{R}^N$$

$$\text{spécifique}$$

$$\text{cor}(x,y) = \frac{1}{N} \sum_{w=1}^N (x(w) - \bar{x})(y(w) - \bar{y})$$

$$\text{cor}(x,y) = \frac{\sqrt{\sum_{w=1}^N (x(w) - \bar{x})^2} \sqrt{\sum_{w=1}^N (y(w) - \bar{y})^2}}{N}$$

$$x - \bar{x} : \text{tendeur}$$

$$\hookrightarrow = (x_1 - \bar{x}, x_2 - \bar{x}, \dots)$$

$$y - \bar{y} : (x_1 - \bar{x}, \dots, x_n - \bar{x})$$

$$\text{cor}(x,y) = \frac{x \cdot \bar{x} \circ y \cdot \bar{y}}{\|x - \bar{x}\| \cdot \|y - \bar{y}\|}$$

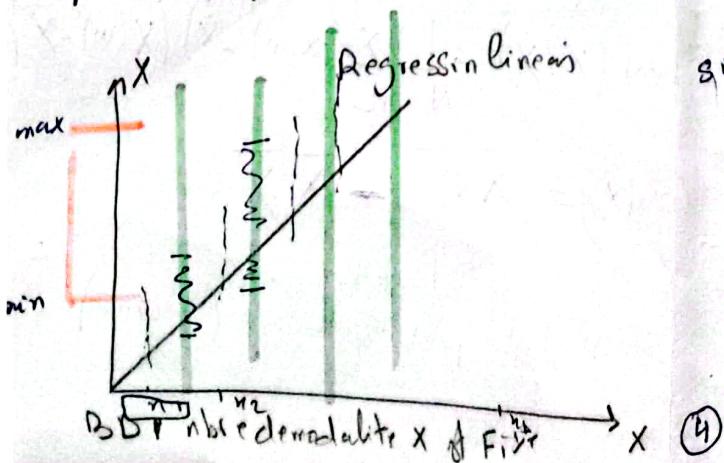
$$\text{cor}(x,y) = \cos(x, y)$$

$$\text{cor}(x,y) = 1 \quad \text{et} \quad \text{cor}(x,y) = 0$$

$\hookrightarrow$  Il y a des liens latents non

pas linéaire

$$\bullet \text{plot}(x,y) =$$



$$\text{cor}(x,y) = 10 \text{ pts}$$

$$\text{cor}(x,y) = -10 \text{ pts}$$

$$\text{cor}(x,y) = 0 \text{ pts}$$

$$z = (x, y)$$

$\hookrightarrow y$  en fonction de  $x$

$$\text{① cor}(x,y) = \rightarrow |\text{cor}(x,y)| \xrightarrow{\text{corrélation}}$$

$\xrightarrow{\text{ou}} 0$   
peut être  
corrélation  
linéaire

② Choisir une corrélation au choix de la variable

③  $\text{cor}(x,y) \neq 0$  pas de corrélation  
(pas la peine de chercher)

$$\bullet \text{cor}(x,y) \approx 1$$

$x$  est corrélé en  $x$  donc  
on fait le retour à ①  
correlation pas linéaire

$$\cos: |\text{cor}(x,y)| \approx 1 : \text{linéaire}$$

$$y = a + bx$$

calcul de  $a$  et  $b$

$$\hookrightarrow \text{Im}(x \sim x) \xrightarrow{a, b}$$

On peut prédire  $y$

sinon

$\hookrightarrow$  formule:

$$b = \frac{\text{cov}(x,y)}{\text{var}(y)}$$

ptile coor. clmme moyenne

$$a = \bar{y} - b\bar{x}$$

