Statistique
Cours 9

Modèles linéaires

Y $\quad x_1, x_2, \dots, x_m$

$y = a_1 x_1 + a_2 x_2 + \dots + a_m x_m + b$ (terme poubelle) $\quad a_i \in \mathbb{R}$, $b_i \in \mathbb{R}$

Inconnues a_i, $i \in 1, \dots, m$, b_i

⇓

Estimation $\hat{a}$, $\hat{b}$ ⇒ Régression multi-linéaire

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ & & & \\ x_{m1} & & \cdots & x_{mm} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \\ b \end{bmatrix} \qquad y = Xa$$

hyp : $n \gg m$

2° $\underbrace{x_1 \dots x_m}$ $\forall a_i$

variables explicatives

3° y dépend linéairement des $x_1, \dots, x_m$

↳ variable à expliquer

$$y = Xa$$

$$\hat{a} = X^+ y$$

où $X^+ = (X^T X)^{-1} X^T$

$$\hat{y} = X \hat{a}$$

Analyse des résidus

$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_m \end{bmatrix}$, $e_i = y_i - \hat{y}_i$ erreur de l'estimation

$4° \; E(e) = 0$

$V = E\left[(X - E(X))(X - E(X))^T\right]$

$= \begin{bmatrix} V_{xx} & V_{xy}^T \\ V_{xy} & V_{xy} \end{bmatrix}$ matrice des variance - covariances.

$V_{xx}(i,i) \Rightarrow$ variance de x_i avec x_j

$V_{y,y} \Rightarrow$ ——— de y

$V_{xx}(i,j) \Rightarrow$ covariance entre x_i et x_j.

Corrélation : $R^2 = \dfrac{V_{xy}^T V_{xx}^{-1} V_{xy}}{V_{yy}}$

Variance de régression $s^2 = \dfrac{n}{n-m-1} V_{y,y} (1 - R^2)$

Tests de signification

Si $\dfrac{n-m-1}{m} \dfrac{R^2}{1-R^2} \geq F_{\alpha, m, n-m-1} \Rightarrow$ il y a au moins un coef parmis les a_i qui est significatif ($a_i \neq 0$)

$Idc(\hat{a}_i) = \left[\hat{a}_i - t_{\alpha, n-m-1} \sqrt{V(\hat{a}_i)}, \; \hat{a}_i + t_{\alpha, n-m-1} \sqrt{V(\hat{a}_i)}\right]$

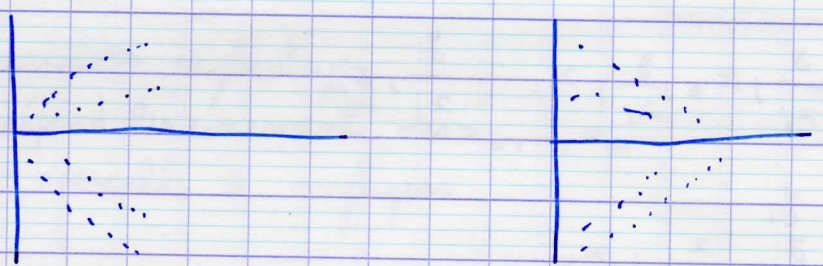
avec $V(\hat{a}_i) = S = (X^T X)^{-1}$

$x_1 \dots x_q$ x_q sous matrice de x créée à partir de x_1, x_i, x_q.

Linéarisation des données

Fonct	Transf.	Fonct
$y = \alpha x^{\beta}$	$y' = \log y$ $x' = \log x$	$y' = \log \alpha + \beta x'$
$y = \alpha e^{\beta x}$	$y' = \ln y$	$y' = \ln \alpha + \beta x$
$y = \alpha + \beta \log(x)$	$x' = \log(x)$	$y' = \alpha + \beta x'$
$y = \frac{x}{\alpha x - \beta}$	$y' = \frac{1}{y}$ $x' = \frac{1}{x}$	$y' = \alpha - \beta x'$
$y = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$	$y' = \ln\left(\frac{y}{1-y}\right)$ $x' = \alpha + \beta x$	

Détection de non-linéarité par examen des résidus



pb grandes valeurs de y — variance non constante — pb petites valeurs de y

Transformation sur les valeurs de y

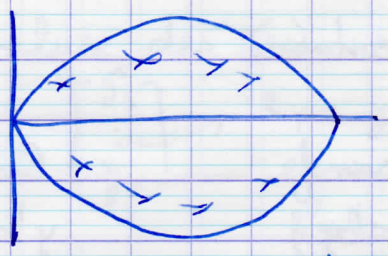
$y' = \sqrt{y}$

$y' = \log(y)$

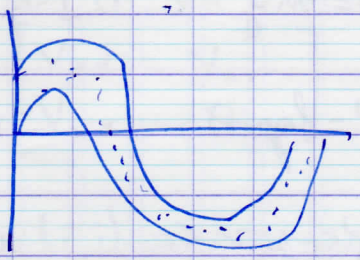
$y' = \log(y+1)$

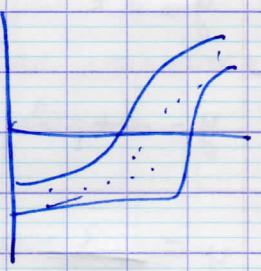
$y' = \frac{1}{y}$

$y' = \frac{1}{y+1}$



$y' = \log(y)$ $\quad y' = \log(y+1)$





variance non constante
pas de réponse.

Algo de la régression multi-linéaire

$y \mid x_1, \dots, x_m$

1° $\frac{y}{x_1}, \frac{y}{x_2}, \dots, \frac{y}{x_m}$

x_i test de signification $F_{\alpha,1,m-2}$ et le meilleur

2° $\frac{y}{x_i}, x_1, \frac{y}{x_i}, x_2 \dots \frac{y}{x_i}, x_m$

(D) y_i

3° $\frac{y}{x_i}, x_j, x_s \dots$