Régression linéaire

M8 – Chapitre 3

I. Méthodes d'études possibles

Variable a expliquer x	Variables explicative X	Méthode
Quantitative	Quantitative	Régression
Quantitative	Qualitative	Analyse de la variance
Qualitative	Quantitative	Analyse discriminante
Qualitative	Qualitative	

II. Régression simple

1. Le modèle « classique »

$$\begin{cases} y_1 = ax_1 + b + \varepsilon_1 \\ \vdots \\ y_n = ax_n + b + \varepsilon_n \end{cases}$$

Ecriture scalaire

$$\boldsymbol{y} = a\boldsymbol{x} + b\boldsymbol{e} + \boldsymbol{\varepsilon}$$

Ecriture vectorielle

$$\boxed{\boldsymbol{y} = \boldsymbol{X\alpha} + \boldsymbol{\varepsilon}}$$

Ecriture matricielle

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{e} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \quad \boldsymbol{\alpha} = \begin{pmatrix} a \\ b \end{pmatrix}$$

2. Solution scalaire

On cherche a et b tel que $\min\limits_{a,b} \sum_{i=1}^{n} \varepsilon_i^2 \quad \Rightarrow \quad \boxed{\hat{a} = \frac{cov(x,y)}{V_x}} \quad \boxed{\hat{b} = \bar{x}\hat{a} + \bar{y}}$

III. Régression multiple

1. Le modèle

$$\begin{cases} y_1 = \mathrm{a}_0 + a_1 x_{11} + \cdots + a_p x_{1p} + \varepsilon_1 \\ \vdots \\ y_n = \mathrm{a}_0 + a_1 x_{n1} + \cdots + a_p x_{np} + \varepsilon_1 \end{cases}$$

Ecriture scalaire

$$\boldsymbol{y} = \mathrm{a}_0\boldsymbol{e} + a_1\boldsymbol{x_1} + \cdots + a_p\boldsymbol{x_p} + \boldsymbol{\varepsilon}$$

Ecriture vectorielle

$$\boxed{\boldsymbol{y} = \boldsymbol{X\alpha} + \boldsymbol{\varepsilon}}$$

Ecriture matricielle

$$\boldsymbol{x_i} = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{ni} \end{pmatrix} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{e} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \quad \boldsymbol{\alpha} = \begin{pmatrix} a_0 \\ \vdots \\ a_p \end{pmatrix}$$

2. Solution

$$\underline{X^{\mathrm{T}}\varepsilon} = 0 \Rightarrow X^{\mathrm{T}}(y - X\hat{\alpha}) = 0 \Rightarrow (X^{\mathrm{T}}X)\hat{\alpha} = X^{\mathrm{T}}y \Leftrightarrow \boxed{\hat{\alpha} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y} \Leftrightarrow \underline{\hat{\alpha} = (X^{\mathrm{T}}X) \setminus (X^{\mathrm{T}}y)}$$

3. Prévision

$$z = X\alpha = \underbrace{X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}}_{H}y = Hy$$

IV. Diagnostic de la régression

Résultat	Formule		
R²	$\boxed{\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SC\ Total} = \underbrace{\sum_{i=1}^{n}(y_i-z_i)^2}_{SC\ Résiduels} + \underbrace{\sum_{i=1}^{n}(z_i-\bar{y})^2}_{SC\ Expliqué}}$ $R = \lvert cor(\boldsymbol{y},\boldsymbol{z})\rvert$ coefficient de corrélation multiple	$\boxed{R^2 = \frac{SC\ Expliqué}{SC\ Total}}$ $0 \leq R^2 \leq 1$ $R = 1$: modèle bon $R = 0$: modèle mauvais $R^2 = r_{XY}^2$ en régression simple	
Matrice d'influence	$z = \underbrace{X\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}}_{H}y \ = Hy$	$z_i = H_{i\bullet}^T y$	$H_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{j=1}^{n}\left(x_j-\bar{x}\right)^2}$ Pour la régression simple
Variances estimées	$s^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i$	$\left(s^{(-i)}\right)^2 = \frac{1}{n-p-1}\left(\sum_{i=1}^{n}\hat{\varepsilon}_i - \frac{\hat{\varepsilon}_i^2}{1-H_{ii}}\right)$	
Résidus	$\hat{\varepsilon} = y - z$ $\hat{\varepsilon} = (I-H)y$ **Résidus** Sans structure / distrib. normal / pas d'aberrants	$r_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1-H_{ii}}} = \frac{\hat{\varepsilon}_i}{\sqrt{V(\hat{\varepsilon}_i)}}$ **Résidus standardisés**	
Divergence	$\hat{\varepsilon}_i^{(-i)} = \frac{\hat{\varepsilon}_i}{1-H_{ii}} = y_i - z_i^{-i}$ **Résidus de validation croisée**	$err_{VC} = \sum_{i=1}^{n}\left(\hat{\varepsilon}_i^{(-i)}\right)^2$ **Erreur de validation croisée**	$t_i = \frac{\hat{\varepsilon}_i}{s^{(-i)}\sqrt{1-H_{ii}}} = \frac{\hat{\varepsilon}_i^{(-i)}}{\sqrt{V\left(\hat{\varepsilon}_i^{(-i)}\right)}}$ **Résidus studentisés** = résidus de VC normalisés
Levier et contribution	$\boxed{levier_i = H_{ii}} = \lVert H_{i\bullet}\rVert^2$ Important si $H_{ii} > \frac{2(p+1)}{n}$ **Levier**	$c_i = \frac{H_{ii}}{p(1-H_{ii})}\frac{\hat{\varepsilon}_i^2}{s^2}$ Suspect si $c_i > \frac{4}{n}$ **Contribution**	
Cp de Mallows	$\boxed{Cp = \frac{1}{s^2}\sum_{i=1}^{n}\left(y_i - z_i^{(-i)}\right)^2 - n + 2p}$ Conserver la combinaison de variable avec le plus faible Cp		