

Régression linéaire

M8 – Chapitre 3

I. Méthodes d'études possibles

Variable à 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

$$y = ax + be + \varepsilon$$

Ecriture vectorielle

$$y = X\alpha + \varepsilon$$

Ecriture matricielle

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

2. Solution scalaire

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

III. Régression multiple

1. Le modèle

$$\begin{cases} y_1 = a_0 + a_1x_{11} + \dots + a_px_{1p} + \varepsilon_1 \\ \vdots \\ y_n = a_0 + a_1x_{n1} + \dots + a_px_{np} + \varepsilon_n \end{cases}$$

Ecriture scalaire

$$y = a_0e + a_1x_1 + \dots + a_px_p + \varepsilon$$

Ecriture vectorielle

$$y = X\alpha + \varepsilon$$

Ecriture matricielle

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ni} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \quad \alpha = \begin{pmatrix} a_0 \\ \vdots \\ a_p \end{pmatrix}$$

2. Solution

$$X^T \varepsilon = 0 \Rightarrow X^T(y - X\hat{a}) = 0 \Rightarrow (X^T X)\hat{a} = X^T y \Leftrightarrow \hat{a} = (X^T X)^{-1} X^T y \Leftrightarrow \hat{a} = (X^T X) \setminus (X^T y)$$

3. Prédiction

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

Régression linéaire

M8 – Chapitre 3

IV. Diagnostic de la régression

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