

Modèles de régression linéaire

Le but d'un modèle de régression est d'exprimer comment une variable X dépend d'autres variables $(A_i)_{i \in [1, p]}$.

X s'appelle alors variable à expliquer tandis que les $(A_i)_{i \in [1, p]}$ sont les variables explicatives.

Puisqu'on a souvent plusieurs échantillons ou plusieurs mesures les variables sont à valeurs dans $\mathbb{R}^n$ (n échantillons ou n mesures).

Dans un modèle linéaire on doit exprimer X comme une fonction linéaire des $(A_i)_{i \in [1, p]}$, ainsi en fixant p coefficients on peut définir un modèle. Les coefficients sont appelés paramètres du modèle.

23.1 **Déf** Un modèle de régression linéaire à p variables dans $\mathbb{R}^n$ s'écrit $\tilde{X} = A\theta$ où $A = [A_1, \dots, A_p] \in \mathcal{M}_{n,p}(\mathbb{R})$ et $\theta = (\theta_i)_{i \in [1, p]} \in \mathbb{R}^p$.
 $(A_i)_{i \in [1, p]}$ est la famille des p variables explicatives du modèle.
 $(\theta_i)_{i \in [1, p]}$ est la famille des p paramètres du modèle.

En pratique on dispose souvent de X contenant les mesures effectives d'une certaine quantité et des $(A_i)_{i \in [1, p]}$ contenant les mesures effectives des p facteurs qu'on pense explicatifs. Il s'agit alors de fixer les paramètres du modèle afin qu'il soit le plus juste possible. Ici on considèrera que le modèle le plus juste est celui qui colle au plus près de la réalité, c'est à dire celui qui minimise $\|X - \tilde{X}\|^2$, on l'appelle solution au problème des moindres carrés.

23.2 **Déf** $\tilde{X} = A\theta$ est un modèle régulier $\Leftrightarrow \text{rg}(A) = p \Leftrightarrow$ les var. explicatives sont linéairement indép.

23.3 **Pk** $\hat{\theta} \in \mathbb{R}^p$ tel que $A\hat{\theta} = P_{\text{Im}(A)}^+ X$ est solution au problème des moindres carrés.
 $\tilde{X} = A\theta$ est régulier $\Rightarrow \hat{\theta} = (A'A)^{-1} A'X$ est l'unique solution au problème des moindres carrés.

23.4 **Déf** Soit $\hat{\theta} \in \mathbb{R}^p$. $\hat{\theta}$ est un estimateur de paramètre $\Leftrightarrow A\hat{\theta} = P_{\text{Im } A}^\perp X$

Rq Dans le cas où le modèle n'est pas régulier il y a plusieurs estimateurs de paramètre possibles : le projeté de X sur $\text{Im } A$ est unique mais comme A n'est pas inj. ...

23.5 **Déf** Soit $\hat{x} \in \mathbb{R}^n$. $\hat{x}$ est une valeur prédite par le modèle $\Leftrightarrow$ il existe $\hat{\theta}$ un estimateur de paramètre tel que $\hat{x} = A\hat{\theta}$

23.6 **Déf** Soit le modèle $\tilde{x} = A\theta$ et $\hat{x} = P_{\text{Im } A}^\perp(X)$. On garde ces notations jusqu'à la fin.

SCE = Somme des Carrés des Erreurs = $\|X - \hat{x}\|^2$

VEH = Valeur Explicative du modèle = $1 - \frac{\text{SCE}}{\|X\|^2}$

23.7 **Pl'** $\text{SCE} = \sum_{i=1}^n \left(X_i - \sum_{k=1}^p A_{ik} \hat{\theta}_k \right)^2$ où $\hat{\theta}$ est un estimateur de paramètre. $\text{VEH} = \frac{\|\hat{x}\|^2}{\|X\|^2}$

$\exists \hat{\theta}$ tel que $\hat{x} = A\hat{\theta}$. $\|X - \hat{x}\|^2 = \sum_{i=1}^n (X_i - \hat{x}_i)^2$ or $\hat{x}_i = \sum_{k=1}^p A_{ik} \hat{\theta}_k$

$\text{VEH} = 1 - \frac{\text{SCE}}{\|X\|^2} = \frac{\|X\|^2 - \|X - \hat{x}\|^2}{\|X\|^2}$ Or puisque $(X - \hat{x}) \perp \hat{x}$ le théorème de Pythagore donne $\|X\|^2 = \|X - \hat{x}\|^2 + \|\hat{x}\|^2$ d'où $\|X\|^2 - \|X - \hat{x}\|^2 = \|\hat{x}\|^2$ et $\text{VEH} = \frac{\|\hat{x}\|^2}{\|X\|^2}$

La VEH ne permet que de comparer des modèles pour un même problème ; plus elle est proche de 1 meilleur c'est

23.8 **Pl'** $\mathbb{1} \in \text{Im } A \Leftrightarrow \forall X \in \mathbb{R}^n, E[X] = E[\hat{x}]$

• Si $\mathbb{1} \in \text{Im } A$: $E[X] - E[\hat{x}] = E[X - \hat{x}]$ or puisque $X - \hat{x} \in \text{Im } A^\perp$ on a $\mathbb{1} \perp X - \hat{x}$ donc $E[X - \hat{x}] \cdot \mathbb{1} = P_{\mathbb{1}}^\perp(X - \hat{x}) = 0$ d'où $E[X] = E[\hat{x}]$.

• Réc : Si $\forall X \in \mathbb{R}^n, E[X] = E[\hat{x}]$, $\forall X \in \mathbb{R}^n, P_{\mathbb{1}}^\perp X = P_{\mathbb{1}}^\perp \hat{x} = P_{\mathbb{1}}^\perp \times P_{\text{Im } A}^\perp X$ donc

$P_{\mathbb{1}}^\perp = P_{\mathbb{1}}^\perp \times P_{\text{Im } A}^\perp$ soit $\frac{\mathbb{1} \cdot \mathbb{1}'}{n} = \frac{\mathbb{1} \cdot \mathbb{1}'}{n} P_{\text{Im } A}^\perp$ donc $\mathbb{1}' \cdot \mathbb{1} = \mathbb{1}' \cdot \mathbb{1} P_{\text{Im } A}^\perp$ et en transposant $\mathbb{1} = (P_{\text{Im } A}^\perp)'(\mathbb{1})$ or la transposée d'un proj. ortho est lui-même (cf 23.3) donc $P_{\text{Im } A}^\perp(\mathbb{1}) = \mathbb{1}$ d'où $\mathbb{1} \in \text{Im } A$

23.9 **Pl'-Déf** Si $\mathbb{1} \in \text{Im } A$ on définit le coefficient de détermination

$$R^2 = 1 - \frac{\text{SCE}}{n \text{Var}(X)} = \frac{\text{var}(\hat{x})}{\text{var}(X)}$$

$\left. \begin{array}{l} \hat{x} \in \text{Im } A \\ \mathbb{1} \in \text{Im } A \end{array} \right\} \underbrace{(\hat{x} - \bar{x} \mathbb{1})}_{\in \text{Im } A} \perp \underbrace{X - \hat{x}}_{\in \text{Im } A^\perp}$. Donc le th de Pythagore donne $\|\hat{x} - \bar{x} \mathbb{1}\|^2 + \|X - \hat{x}\|^2 = \|\hat{x} - \bar{x} \mathbb{1} + X - \hat{x}\|^2$ soit $\|\hat{x} - \bar{x} \mathbb{1}\|^2 = \|X - \bar{x} \mathbb{1}\|^2 - \|X - \hat{x}\|^2$

$$\text{donc } R^2 = 1 - \frac{\|X - \hat{x}\|^2}{\|X - \bar{x} \mathbb{1}\|^2} = \frac{\|X - \bar{x} \mathbb{1}\|^2 - \|X - \hat{x}\|^2}{\|X - \bar{x} \mathbb{1}\|^2} = \frac{\|\hat{x} - \bar{x} \mathbb{1}\|^2 \times \frac{1}{n}}{\|X - \bar{x} \mathbb{1}\|^2 \times \frac{1}{n}} = \frac{\|\hat{x} - \bar{x} \mathbb{1}\|^2 \times \frac{1}{n}}{\|X - \bar{x} \mathbb{1}\|^2 \times \frac{1}{n}} = \frac{\text{Var}(\hat{x})}{\text{Var}(X)}$$

d'après 23.8.