

Section 1 :

Définition des modèles linéaires

STAT-F-406

Master en sciences mathématiques, Master en statistique

ACTU-F4001

Master en sciences actuarielles

Davy Paindaveine

Université libre de Bruxelles

2024–2025

Définition des modèles linéaires

Soit (X, Y) un vecteur aléatoire, où Y est à valeurs dans \mathbb{R} et $X = (X_1, \dots, X_k)$ est à valeurs dans \mathbb{R}^k .

On supposera toujours que $E[Y^2] < \infty$ et $E[X_j^2] < \infty$ pour tout j .

Cette hypothèse assure que toutes ces v.a. appartiennent à

$$L_2(\Omega, \mathcal{A}, P) := \{Z : \Omega \rightarrow \mathbb{R} \text{ telles que } E[Z^2] < \infty\},$$

qui est un espace de Hilbert muni du produit scalaire $\langle Z_1, Z_2 \rangle = E[Z_1 Z_2]$ (et donc de la norme $\|Z\| = \sqrt{\langle Z, Z \rangle} = \sqrt{E[Z^2]}$). Cette structure va nous permettre de définir ce qu'on entend par β et par ε dans

$$Y = \beta' X + \varepsilon.$$

Identification par projection

En l'état, pour tout $v \in \mathbb{R}^k$, on peut écrire

$$Y = \beta' X + \varepsilon = (\beta + v)' X + (\varepsilon - v' X) = \tilde{\beta}' X + \tilde{\varepsilon},$$

de sorte que les quantités β et ε ne sont pas **identifiables**
(quel β devons-nous estimer?)

Puisqu'on veut que l'approximation $Y \approx \beta' X$ soit aussi bonne que possible, il est naturel de choisir

$$\beta = \arg \min_{\alpha \in \mathbb{R}^k} \|Y - \alpha' X\|,$$

ce qui mène à

$$\beta' X = \Pi(Y; X_1, \dots, X_k),$$

la projection orthogonale de Y sur l'espace vectoriel engendré par X_1, \dots, X_k .

Indépendance linéaire

La projection est univoquement définie. Et β ?

Pour que β le soit également, on impose que X_1, \dots, X_k soient des vecteurs linéairement indépendants de $L_2(\Omega, \mathcal{A}, P)$: autrement dit, on suppose qu'il n'existe pas $v \in \mathbb{R}^k \setminus \{0\}$ tel que $v'X = 0$ p.s.

Par conséquent, on ne pourra pas considérer

$$X = (X_1, X_2, X_3) = (U, V, U - 25V),$$

mais on pourra considérer

$$X = (X_1, X_2, X_3) = (U, V, U^2).$$

Indépendance linéaire

Lemme 1

β est univoquement défini

$\stackrel{(i)}{\Leftrightarrow} X_1, \dots, X_k$ sont des vecteurs linéairement indépendants

$\stackrel{(ii)}{\Leftrightarrow} E[XX'] > 0$, au sens où $E[XX']$ est une matrice définie positive.

Preuve:

$\stackrel{(i)}{\Leftrightarrow}$ suit directement de la théorie des espaces vectoriels.

$\stackrel{(ii)}{\Rightarrow}$ Pour tout $v \in \mathbb{R}^k \setminus \{0\}$, $v'E[XX']v = E[(v'X)(X'v)] = E[(v'X)^2] \geq 0$.
Si cette espérance est nulle, alors $(v'X)^2 = 0$ p.s., c'est-à-dire $v'X = 0$ p.s., ce qui contredit l'indépendance linéaire des X_j .

$\stackrel{(ii)}{\Leftarrow}$ Par l'absurde, supposons qu'il existe $v \in \mathbb{R}^k \setminus \{0\}$ tel que $v'X = 0$ p.s.
On a alors $v'E[XX']v = E[v'XX'v] = E[(v'X)^2] = 0$, ce qui contredit le fait que $E[XX'] > 0$. □

Expression explicite pour β

Théorème 1

Supposons que $E[XX'] > 0$. Alors $\beta = (E[XX'])^{-1}E[XY]$.

Preuve: Posons $\beta_0 = (E[XX'])^{-1}E[XY]$. Bien sûr, $\beta_0'X$ est dans l'espace vectoriel engendré par X_1, \dots, X_k . Par ailleurs, pour tout $\alpha \in \mathbb{R}^k$,

$$\begin{aligned}\langle \alpha'X, Y - \beta_0'X \rangle &= E[\alpha'X(Y - \beta_0'X)] = \alpha'E[XY - X\beta_0'X] \\ &= \alpha'(E[XY] - E[XX'\beta_0]) = \alpha'(E[XY] - E[XX']\beta_0) = 0.\end{aligned}$$

Donc $\beta_0'X = \Pi(Y; X_1, \dots, X_k)$, ce qui montre que $\beta = \beta_0$. □

Remarque: puisque $E[XX'] > 0$, l'inverse $(E[XX'])^{-1}$ existe bien.

Comme β est bien défini, on peut maintenant poser $\varepsilon \stackrel{\text{def}}{=} Y - \beta'X$.

Une hiérarchie de modèles linéaires

Le modèle obtenu en définissant β et ε comme ci-dessus, sans faire d'hypothèses supplémentaires, est le **modèle faible**.

Notons qu'on a toujours

$$E[\varepsilon X] = 0$$

dans ce modèle (pourquoi?)

Dans la suite, on utilisera $\sigma^2(X) \stackrel{\text{def}}{=} E[\varepsilon^2|X]$.

Une hiérarchie de modèles linéaires

On appellera **modèle semi-fort** le modèle obtenu en imposant en plus que $E[\varepsilon|X] = 0$ p.s.

Notons que

$$E[\varepsilon X] = E[E[\varepsilon X|X]] = E[E[\varepsilon|X]X] = E[0] = 0,$$

ce qui est compatible avec le modèle faible.

De façon plus importante, le modèle semi-fort prévoit que

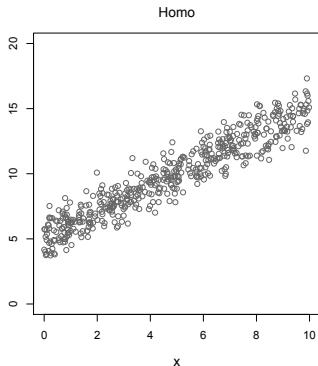
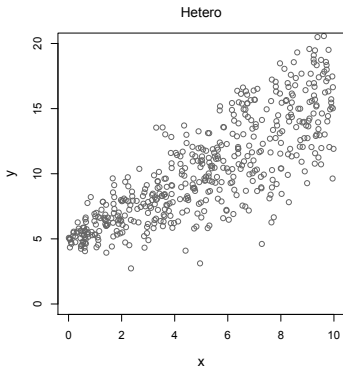
$$E[Y|X] = E[\beta'X + \varepsilon|X] = E[\beta'X|X] + E[\varepsilon|X] = \beta'X,$$

ce qui s'interprète aisément.

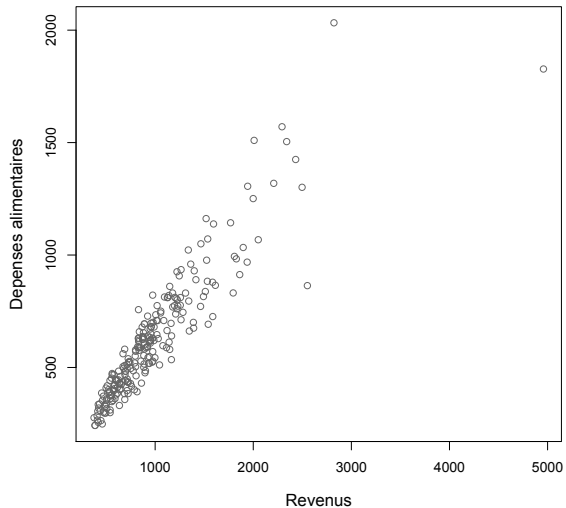
Une hiérarchie de modèles linéaires

Dans ce modèle, $\sigma^2(X) = \text{Var}[\varepsilon|X]$. On différenciera entre

- ▶ modèle semi-fort **hétéroscéastique**: $\sigma^2(X)$ quelconque
- ▶ modèle semi-fort **homoscéastique**: $\sigma^2(X) = \sigma^2$ p.s.



Une hiérarchie de modèles linéaires



Une hiérarchie de modèles linéaires

On parlera de **modèle fort** si $\varepsilon \perp\!\!\!\perp X$ et $E[\varepsilon] = 0$.

On a alors

$$E[\varepsilon|X] = E[\varepsilon] = 0 \text{ et } \text{Var}[\varepsilon|X] = \text{Var}[\varepsilon] \stackrel{\text{def}}{=} \sigma^2,$$

de sorte que ce modèle est semi-fort homoscédastique.

Dans le modèle fort, toute la distribution conditionnelle $\varepsilon | [X = x]$ est la même pour les différentes valeurs de x (et pas seulement E et Var)

Enfin, on parlera de **modèle fort avec normalité** si $\varepsilon \perp\!\!\!\perp X$ et $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
Dans ce modèle, on a donc $Y|[X = x] \sim \mathcal{N}(\beta'x, \sigma^2)$.

Variables exogènes

On supposera que la distribution de $X = (X_1, \dots, X_k)$ ne dépend pas de β ni (à partir du modèle semi-fort homoscédastique) de σ^2 .

On dira que X_1, \dots, X_k sont des **variables exogènes**.