ULB

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

2023–2024

Définition des modèles linéaires

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

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

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

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

qui est un espace de Hilbert muni du produit scalaire $\langle Z_1, Z_2 \rangle = \mathrm{E}[Z_1 Z_2]$ (et donc de la norme $\|Z\| = \sqrt{\langle Z, Z \rangle} = \sqrt{\mathrm{E}[Z^2]}$). 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 devrons-nous estimer?)

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

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

ce qui mène à

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

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

Indépendance linéaire

La projection est univoquement définie. Et β?

Pour que β le soit également, on impose que $X_1, \ldots, X_k$ 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
$\overset{(i)}{\Leftrightarrow}$ *$X_1, \ldots, X_k$ sont des vecteurs linéairement indépendants*
$\overset{(ii)}{\Leftrightarrow}$ *$\mathrm{E}[XX'] > 0$, au sens où $\mathrm{E}[XX']$ est une matrice définie positive.*

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

$\overset{(ii)}{\Rightarrow}$ Pour tout $v \in \mathbb{R}^k \setminus \{0\}$, $v'\mathrm{E}[XX']v = \mathrm{E}[(v'X)(X'v)] = \mathrm{E}[(v'X)^2] \geq 0$. 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.

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

Expression explicite pour β

Théorème 1

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

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

$$\langle \alpha' X, Y - \beta_0' X \rangle = \mathrm{E}[\alpha' X (Y - \beta_0' X)] = \alpha' \mathrm{E}[XY - X\beta_0' X]$$

$$= \alpha'(\mathrm{E}[XY] - \mathrm{E}[XX'\beta_0]) = \alpha'(\mathrm{E}[XY] - \mathrm{E}[XX']\beta_0) = 0.$$

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

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

$$\mathrm{E}[\varepsilon X] = 0$$

dans ce modèle (pourquoi?)

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

Une hiérarchie de modèles linéaires

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

Notons que

$$\mathrm{E}[\varepsilon X] = \mathrm{E}[\mathrm{E}[\varepsilon X|X]] = \mathrm{E}[\mathrm{E}[\varepsilon|X]X] = \mathrm{E}[0] = 0,$$

ce qui est compatible avec le modèle faible.

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

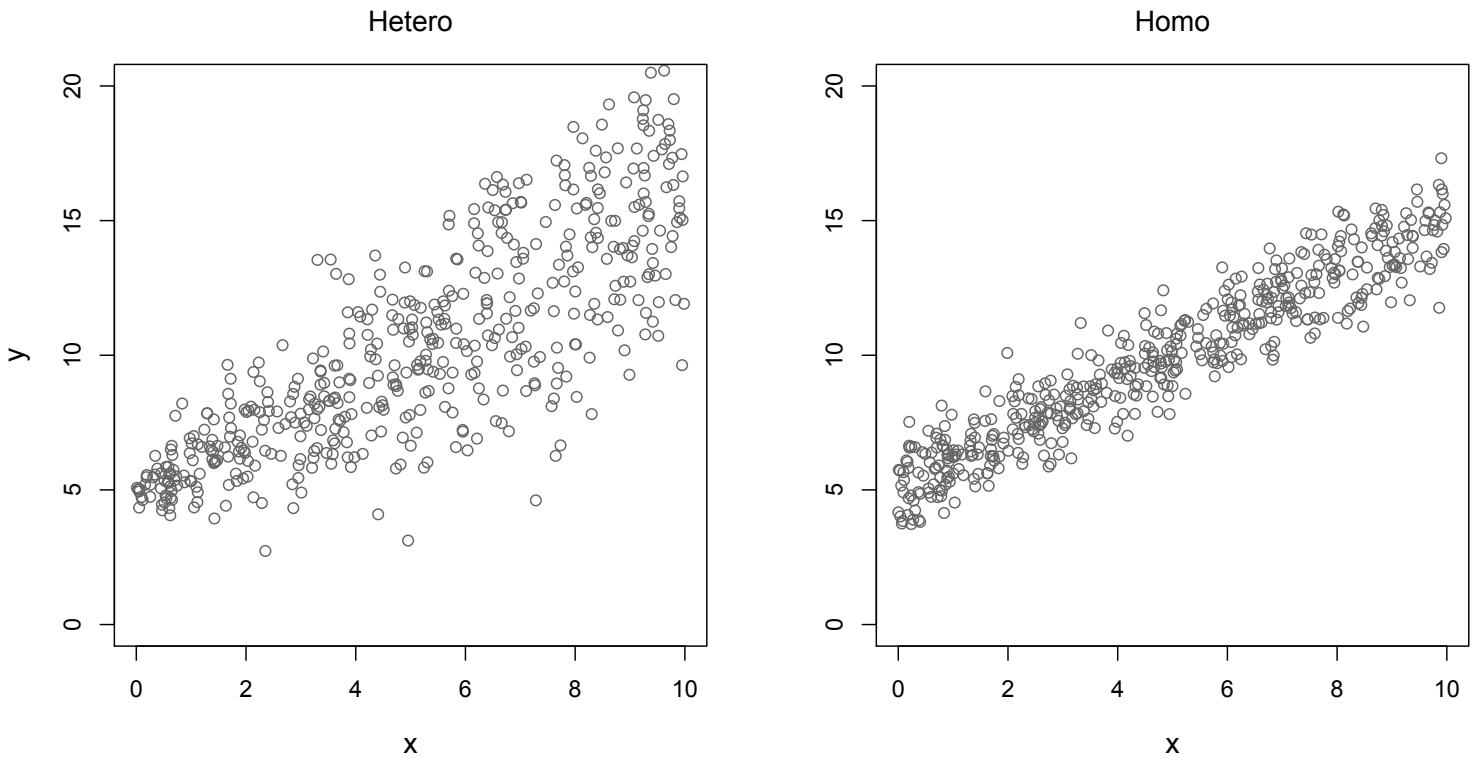
$$\mathrm{E}[Y|X] = \mathrm{E}[\beta' X + \varepsilon|X] = \mathrm{E}[\beta' X|X] + \mathrm{E}[\varepsilon|X] = \beta' X,$$

ce qui s'interprète aisément.

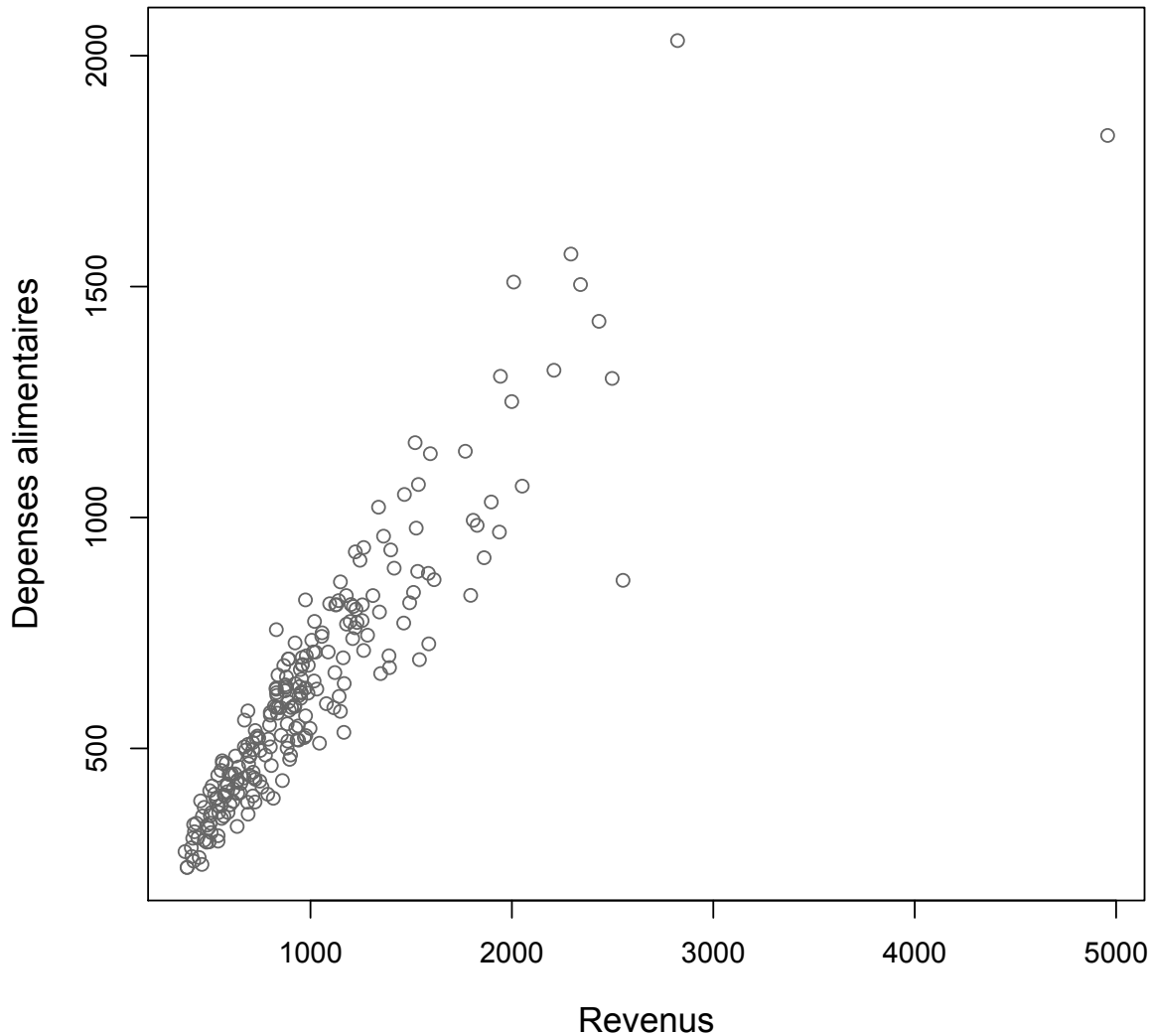
Une hiérarchie de modèles linéaires

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

- modèle semi-fort hétéroscédastique: $\sigma^2(X)$ quelconque
- modèle semi-fort homoscédastique: $\sigma^2(X) = \sigma^2$ p.s.



Une hiérarchie de modèles linéaires



Une hiérarchie de modèles linéaires

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

On a alors

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

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 $X = (X_1, \ldots, X_k)$ ne dépend pas de β ni (à partir du modèle semi-fort homoscédastique) de σ^2.

On dira que $X_1, \ldots, X_k$ sont des variables exogènes.