

MATH 60604  
Modélisation statistique  
§ 5a - Introduction aux données corrélées et  
longitudinales

HEC Montréal  
Département de sciences de la décision

- Le but de ce chapitre est de voir comment prendre en compte la dépendance entre observations.
- On se cantonne à la modélisation de la matrice de covariance pour prendre en compte la dépendance entre observations (pour les données longitudinales et groupées) et l'hétéroscédasticité de groupe.

# Quand les données ne sont pas indépendantes

- Si les observations sont positivement corrélées, les erreurs-type estimés sont **trop petites**.
- On détecte des différences significatives qui ne le sont pas en réalité (erreur de type I enflée, ou faux positifs plus fréquents).

Généralement, la corrélation entre observations provient de

- dépendance temporelle, catégorisée en
  - données longitudinales: mesures répétées sur des individus (séries courtes)
  - séries chronologiques: observations à plusieurs périodes (séries longues). Ces données nécessitent des modèles adaptés qui ne sont pas couverts dans ce cours.
- données groupées: données sur des sujets qui ne sont pas indépendants (familles, groupes, etc.)

- Soit un vecteur aléatoire  $Y$  de dimension  $n$ .
  - Dans la situation qui nous intéresse, un tel vecteur sera habituellement composé des mesures répétées sur un individu ou bien d'observations d'un groupe d'individus.
- L'espérance (ou moyenne théorique) d'un tel vecteur est  $E(Y)$  calculée terme par terme,  $E(Y) = (E(Y_1), \dots, E(Y_n))$ .
- On note aussi la variance de la  $i$ ème composante  $\sigma_{ii} = \sigma_i^2 = \text{Var}(Y_i)$ .
- De même, la covariance entre la  $i$ ème et la  $j$ ème composante est  $\sigma_{ij} = \text{Cov}(Y_i, Y_j)$ .

Pour un vecteur aléatoire  $Y$ , on définit la **matrice de covariance** comme étant la matrice symétrique  $n \times n$

$$\text{Cov}(Y) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2n} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \ddots & \sigma_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \cdots & \sigma_n^2 \end{pmatrix} .$$

- Le  $i$ e élément de la diagonale de  $\text{Cov}(Y)$  est la variance de  $Y_i$ .
- Cette matrice est symétrique, avec  $\sigma_{ij} = \sigma_{ji}$ .

- La corrélation entre  $Y_i$  et  $Y_j$  est donnée par:

$$\rho_{ij} = \text{Corr}(Y_i, Y_j) = \frac{\sigma_{ij}}{\sigma_i \sigma_j}.$$

- La **matrice de corrélation** de  $Y$  est définie comme étant la matrice symétrique  $n \times n$  qui contient un sur la diagonale et les corrélations hors diagonale,

$$\text{Corr}(Y) = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \rho_{31} & \rho_{32} & 1 & \ddots & \rho_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & 1 \end{pmatrix}.$$

Un des traits principaux des données corrélées et longitudinales est la nécessité de tenir compte de la corrélation intra-classe.

- cela reviendra souvent à modéliser la matrice de covariance des observations d'un même groupe (ou d'un même individu dans le cas de mesures répétées).

# Études longitudinales sur des sujets indépendants

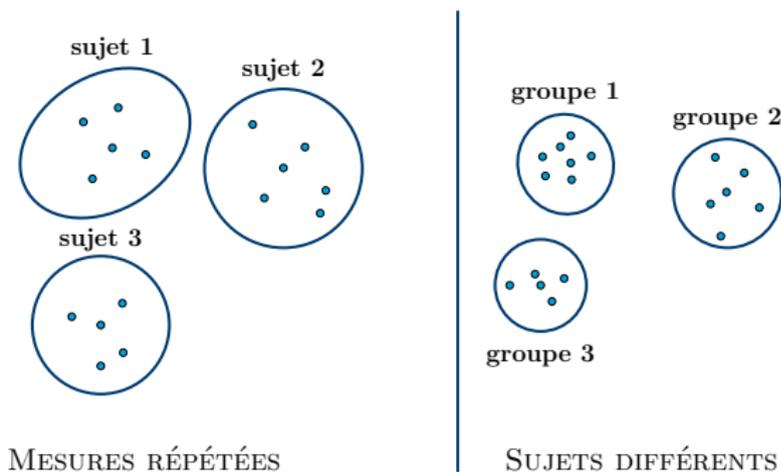
- Dans ce type d'études, plusieurs mesures (habituellement à différents moments dans le temps) sont prises sur les mêmes individus.
  - on nomme ces données **mesures répétées** ou **données longitudinales**; les économètres parlent plutôt de **données de panel**.
- Les individus sont **indépendants** les uns des autres, mais les mesures pour un même sujet ne sont pas indépendantes.
- Un fichier de données pour de telles études a typiquement ce format:

sujet	temps	score	sexe
1	1	5	0
1	2	6	0
1	3	4	0
2	1	2	1
2	2	4	1
2	3	7	1

- Dans ce type d'étude, les sujets sont échantillonnés à l'intérieur d'un **groupe**.
- Voici plusieurs exemples:
  - sujets échantillonnés dans plusieurs ménages (familles),
  - sujets échantillonnés dans plusieurs entreprises,
  - sujets échantillonnées dans des écoles, dans des hôpitaux, etc.
- Dans tous ces cas, il y a de la corrélation entre les mesures des sujets appartenant au même groupe (famille, école, entreprise).

# Les données corrélées sont des données groupées

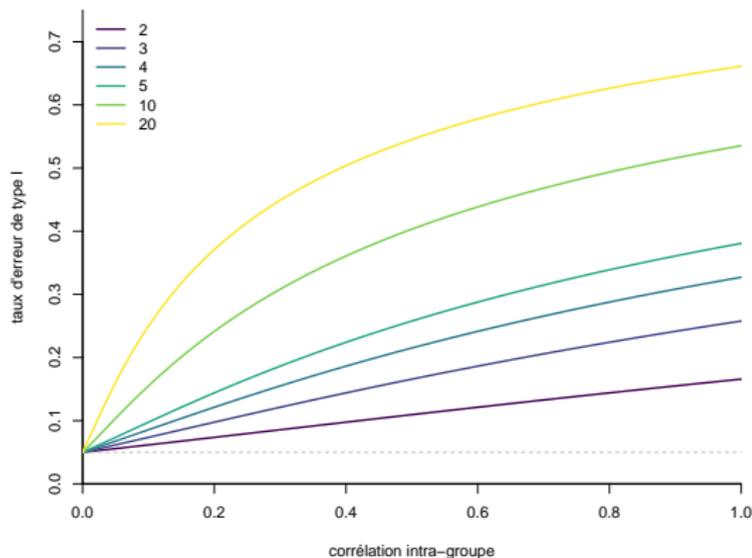
- On peut toujours considérer les données corrélées comme étant des données groupées avec de la corrélation intra-groupe.
- Dans le cas de données longitudinales, le groupe est le sujet lui-même et on a donc plusieurs observations par groupe.
- Dans les autres cas, les groupes sont les ménages, les écoles, les hôpitaux, les entreprises, etc.



Dans tous les cas, un point égale une ligne du fichier de données.

# Que se passe-t-il si on ignore la corrélation intra-groupes?

- Supposons que nous avons des données groupées et que nous voulons faire un test- $t$  pour un échantillon sur ces données à niveau 5%.
- La figure suivante montre quelle est la vraie probabilité d'erreur de type I (qu'on pense être de 5%) en fonction de la corrélation intra-groupe pour différentes valeurs de la taille des groupes  $m$ .



- Il est frappant de voir à quel point la probabilité d'erreur de type I augmente rapidement avec la corrélation et qu'elle augmente d'autant plus vite que le nombre d'observations par groupe est grand.
- La conclusion tirée du test- $t$  n'est pas valide si on ne tient pas compte de la corrélation intra-groupe.
- La distortion du niveau illustre que l'inférence statistique n'est généralement plus valide lorsqu'elle découle d'une méthode supposant l'indépendance entre les observations et que ce n'est pas vérifiée.