

MATH 60604
Modélisation statistique
§ 5c - Formulation du modèle

HEC Montréal
Département de sciences de la décision

- Commençons par ajuster un modèle de régression ordinaire, qui nous servira de point de départ pour la suite.
- Ce modèle néglige la possible corrélation intra-personne et fait comme si les observations étaient indépendantes.
 - Le désir de vengeance d'une personne à un temps donné est fort possiblement corrélé avec celui des autres temps car, justement, il s'agit de la même personne.
 - Si c'est le cas, le postulat d'indépendance n'est pas vérifié; il s'ensuit que l'inférence n'est pas valide.
- Le modèle linéaire est

$$\text{vengeance} = \beta_0 + \beta_1 \text{sexe} + \beta_2 \text{age} + \beta_3 \text{vc} + \beta_4 \text{wom} + \beta_5 \text{t} + \varepsilon,$$

où les termes d'erreur ε sont supposés indépendants.

Modéliser la progression chronologique

- Il y a deux façons naturelles de modéliser la variable temps:
 - soit on suppose un effet linéaire entre t et *vengeance* (variable continue).
 - Soit on traite t comme variable catégorielle.
- Nous allons utiliser `proc mixed` pour nous familiariser avec sa syntaxe.

Code SAS pour ajuster un modèle linéaire

```
proc mixed data=modstat.vengeance method=reml;  
model vengeance = sexe age vc wom t / solution;  
run;
```

Sortie du modèle linéaire avec proc mixed

Informations sur le modèle	
Table	INFSTAT.VENGEANCE
Variable dépendante	vengeance
Structure de covariance	Diagonal
Méthode d'estimation	REML
Méthode de variance résiduelle	Profil
Méthode SE des effets fixes	Basé(e) sur le modèle
Méthode des degrés de liberté	Résidu

Dimensions	
Paramètres de covariance	1
Colonnes dans X	6
Colonnes dans Z	0
Sujets	1
Max. obs. par sujet	400

Valeur estimée du paramètre de covariance	
Param. de cov.	Estimation
Residual	0.3791

Tests d'ajustement	
-2 log-vraisemblance restreinte	776.7
AIC (préférer les petites valeurs)	778.7
AICC (préférer les petites valeurs)	778.7
BIC (préférer les petites valeurs)	782.6

La sortie de mixed est plus compliquée que celle de glm.

Estimés des coefficients pour la moyenne

Solution pour effets fixes

Effet	Estimation	Erreur type	DDL	Valeur du test t	Pr > t
Intercept	-0.1689	0.2249	394	-0.75	0.4532
sexe	0.1357	0.06748	394	2.01	0.0450
age	0.04586	0.004507	394	10.18	<.0001
vc	0.5225	0.01951	394	26.78	<.0001
wom	0.3989	0.02474	394	16.12	<.0001
t	-0.5675	0.02177	394	-26.07	<.0001

Les effets de toutes les variables semblent significatifs, quoique tout juste pour sexe.

- Plus la personne a eu initialement des comportements de type v_c ou w_m , plus le désir de vengeance est élevé.
- L'effet du temps est particulièrement d'intérêt ici. On voit qu'il est négatif. À chaque vague successive, la valeur de vengeance diminue de 0.568, en moyenne, lorsque les autres variables demeurent inchangées. C'est ce qu'on avait vu dans les graphes.
- **Mais peut-on se fier aux conclusions des tests d'hypothèse?** La réponse est non. L'inférence (tests et intervalles de confiance) est faussée lorsqu'on néglige la dépendance entre les observations.

- Supposons qu'on dispose de m groupes d'observations tels que:
 1. Il y a n_i observations dans le groupe i ($i = 1, \dots, m$).
 2. Deux observations d'un même groupe sont possiblement corrélées.
 3. Deux observations provenant de deux groupes distincts sont indépendantes.
- Les groupes peuvent être formés de plusieurs manières:
 - Plusieurs mesures peuvent être prises sur un même sujet (mesures répétées) et chaque individu forme alors un groupe.
 - Un groupe peut aussi être formé d'individus d'une même école, d'une même unité ou département au travail ou d'une même famille.
- Comme par le passé, nous allons supposer que nous avons une variable réponse et un ensemble de p variables explicatives.
- Pour simplifier la notation, on dénote par \mathbf{X}_i l'ensemble des variables aléatoires pour le groupe i .

- On utilise l'indice i pour indiquer le groupe et j pour indiquer la j e observation au sein du groupe i .
 - Si le groupe est une entreprise, i dénote l'entreprise, et j dénote le sujet.
 - Pour des données longitudinales, i représente le sujet et j la mesure à un **temps** donné.
- On note $Y_i = (Y_{i1}, \dots, Y_{in_i})$ l'**ensemble des observations** de la variable d'intérêt pour le groupe i .
- Pour les variables explicatives, on a maintenant besoin de trois indices:
 - i pour le groupe
 - j pour le numéro d'observation au sein du groupe
 - k pour la variable explicative.
- On note donc $\mathbf{X}_{ij} = (1, X_{ij1}, \dots, X_{ijp})$ l'ensemble des p variables explicatives pour l'observation j du groupe i .

Le modèle de régression linéaire peut s'écrire

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp} + \varepsilon_{ij}$$

pour $i = 1, \dots, m$ et $j = 1, \dots, n_i$, où ε_{ij} est le terme d'erreur de l'observation j du groupe i .

- Comme avant, on suppose que $E(\varepsilon_{ij} | \mathbf{X}_{ij}) = 0$ et donc

$$E(Y_{ij} | \mathbf{X}_i) = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp}.$$

Structure de covariance/corrélation

- Puisqu'on assume les variables explicatives \mathbf{X} fixes dans le modèle, parler de structure de corrélation des erreurs ε est équivalent à parler de la structure de corrélation des observations Y .
- Nous allons permettre la dépendance entre les observations d'un même groupe.
- On suppose les groupes indépendants, donc $\text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$ si $i \neq i'$.
- On modélise la corrélation **intra-groupe** en assumant que la matrice de covariance des Y du groupe i est

$$\text{Cov}(Y_i | \mathbf{X}_i) = \boldsymbol{\Sigma}_i,$$

ou, de manière équivalente,

$$\text{Cov}(\varepsilon_i | \mathbf{X}_i) = \boldsymbol{\Sigma}_i,$$

où $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})$ est le vecteur des erreurs du groupe i .

Covariance par blocs pour données longitudinales

- Supposons sans perte de généralité que les données sont ordonnées par groupe.
- On suppose que les observations d'un groupe i sont corrélées, mais que les observations de groupes différents sont indépendantes.
- Alors, la matrice de covariance de toutes les **observations** est **diagonale par blocs**:

$$\text{Cov}(Y) = \begin{pmatrix} \Sigma_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma_2 & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma_m \end{pmatrix}.$$

- Dans notre exemple *vengeance*, on a $n = 80 \times 5 = 400$ observations.
- La matrice de covariance **intra-groupe**, Σ_i , est 5×5 parce que notre échantillon équilibré ($n_1 = \cdots = n_m = 5$). Le bloc Σ_i est identique peu importe le groupe.
- La covariance **inter-groupe** est **nulle** ($\mathbf{0}$) parce qu'on suppose que les données d'individus différents sont indépendantes les unes des autres.

- En général, la structure de covariance dépendra de quelques paramètres qui seront estimés au même titre que les paramètres β .
- La structure de covariance sera habituellement spécifiée par l'analyste. Il n'est pas rare qu'il faille essayer plus d'une structure de covariance pour déterminer celle qui convient le mieux aux données.
- Nous reviendrons sur le problème du choix de la structure de covariance plus tard, mais nous présentons dès maintenant une structure de covariance, parmi les plus simples.