MATH 60604
Modélisation statistique
§ 5b - Exemple de données longitudinales

Léo Belzile

HEC Montréal
Département de sciences de la décision

Exemple: évolution dans le temps du désir de vengeance

- Cet exemple traite du phénomène de vengeance des consommateurs qui est en forte croissance avec les possibilités qu'offre internet.
- Nous nous limiterons à étudier l'impact de certaines variables sur le désir de vengeance et aussi de voir comment le désir de vengeance évolue dans le temps.
- Les données utilisées ici sont fictives mais dans l'étude, elles provenaient de personnes qui s'étaient plaintes d'une firme sur les sites `ConsumerAffairs.com` et `RipOffReport.com`.
- Cinq vagues de questionnaires ont été envoyés à ces personnes, à des intervalles de deux semaines.

Exemple: évolution dans le temps du désir de vengeance

- **Variable dépendante d'intérêt**: le désir de vengeance, mesuré dans chaque questionnaire.
  - Moyenne de cinq items sur une échelle de Likert allant de pas du tout d'accord (1) à tout à fait d'accord (7).
  - Par exemple, l'un des items est « Je voulais prendre des mesures pour causer des problèmes à l'entreprise ».
- **Variables explicatives**: seulement mesurées lors de la première vague — le sexe, l'âge et deux variables mesurant des comportements de vengeance :
  - **plainte vindicative**, basée sur des items tel que « je me suis plaint des services de l'entreprise pour faire passer des moments difficiles au service à la clientèle ».
  - **bouche-à-oreille négatif**, basée sur trois items tel que « j'ai colporté des commentaires négatifs sur l'entreprise par bouche-à-oreille ».

Données `vengeance`

- Un échantillon de 80 personnes ont participé à l'étude.
- Les données sont dans le fichier `vengeance.sas7bdat`.
- Les variables sont
  - `id`: identification de la personne (entier allant de 1 à 80).
  - `t`: temps de mesure (1 à 5).
  - `vengeance`: désir de vengeance (variable dépendante).
  - `sexe`: homme (`0`) ou femme (`1`).
  - `age`: âge (en années).
  - `vc`: score pour items liées aux plaintes vindicatives
  - `wom`: score pour items liés au bouche-à-oreille négatif.

Données des trois premiers individus

Code SAS pour imprimer un sous-ensemble des données

```
proc print data=modstat.vengeance(where=(id<4));
run;
```

Obs.	sexe	age	vc	wom	id	t	vengeance
1	1	38	1	5.6666666667	1	1	4.6
2	1	38	1	5.6666666667	1	2	4
3	1	38	1	5.6666666667	1	3	3.6
4	1	38	1	5.6666666667	1	4	2.4
5	1	38	1	5.6666666667	1	5	2.4
6	0	28	1	1.3333333333	2	1	1.2
7	0	28	1	1.3333333333	2	2	1
8	0	28	1	1.3333333333	2	3	1.8
9	0	28	1	1.3333333333	2	4	1
10	0	28	1	1.3333333333	2	5	1
11	1	40	4	3	3	1	5
12	1	40	4	3	3	2	4.6
13	1	40	4	3	3	3	3.6
14	1	40	4	3	3	4	4.2
15	1	40	4	3	3	5	1.2

Évolution dans le temps du désir de vengeance

- Il est important de comprendre la structure des données, ici une ligne par observation.
- Dans le fichier, on a cinq lignes par personne:
  - aucune valeur manquante,
  - chacune des cinq lignes correspond à un temps de mesure `t`.
- La seule variable qui évolue est `vengeance`.
  - Les variables `sexe`, `age`, `vc` et `wom` ont seulement été mesurées au temps `t=1`, elles sont reportées pour chaque temps de mesure.
- Lorsqu'on a des modèles longitudinaux, il est souvent nécessaire de formater les données pour avoir une ligne par mesure (format long).

Statistiques descriptives avec données répétées

Il faut être prudent si on calcule des statistiques descriptives pour les variables qui sont fixes, comme `sexe`, `age`, `vc` et `wom`.

- La moyenne empirique sera identique uniquement parce qu'on a pas le même nombre d'observations par personne, $T = 5$.
- L'estimée de l'erreur-type de la moyenne sera trop petite parce que le nombre réel de mesures uniques $N = 80$ n'est pas égale au nombre de lignes de la base de données $NT = 400$.

Exemple de calcul

L'âge moyen des $N = 80$ participants est $\overline{\text{age}} = 42.075$ ans.

- l'écart-type est $S = 7.49$ et $\text{se}(\overline{\text{age}}) = 0.837$, où

$$S^2 = \frac{\sum_{i=1}^{N}(\text{age}_i - \overline{\text{age}})^2}{N-1}, \qquad \text{se}(\overline{\text{age}}) = \frac{S}{\sqrt{80}}.$$

Comparez avec le calcul suivant **qui n'est pas correct**.

$$S_*^2 = \frac{\sum_{i=1}^{NT}(\text{age}_i - \overline{\text{age}})^2}{(NT-1)} \approx S^2, \qquad \text{se}(\overline{\text{age}}) \neq \frac{S_*}{\sqrt{400}} = 0.37$$

Statistiques descriptives avec SAS

On peut par contre se limiter aux mesures pour un instant donné.

Code SAS pour calculer les statistiques descriptives

```
proc means data=modstat.vengeance(where=(t=1));
var sexe age vc wom;
run;
proc corr data=modstat.vengeance(where=(t=1));
var sexe age vc wom;
run;
```

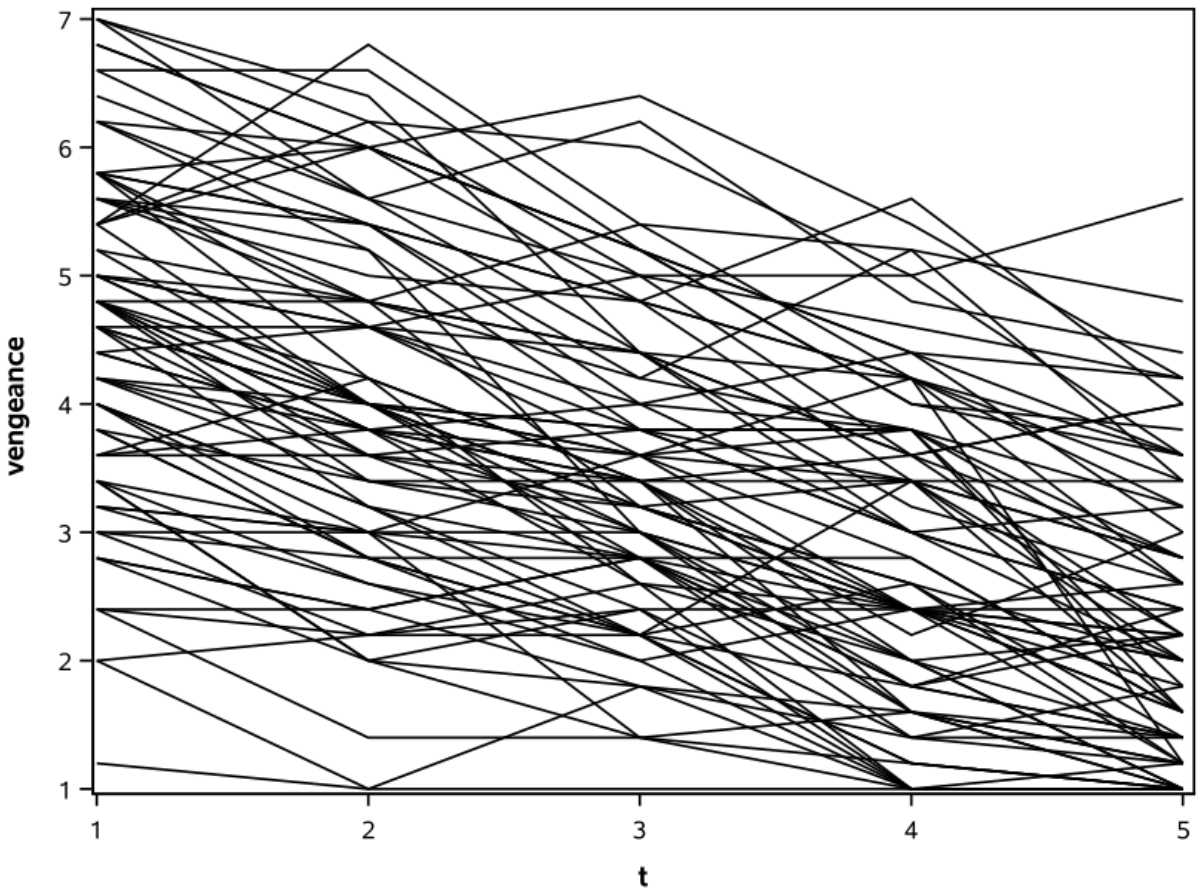
Visualisation de l'effet temporel

- Il est plausible que le désir de vengeance varie en fonction du temps.
- Une façon simple de visualiser l'évolution du désir de vengeance est de tracer un graphique de `vengeance` en fonction de t pour chaque personne.

Code SAS pour dessiner un graphique spaghetti

```
proc sgplot data=modstat.vengeance;
series x=t y=vengeance / group=id;
run;
```

Graphique spaghetti



Ce graphe contient 80 courbes (une par personne) qui s'entrecroisent. Bien que difficile à interpréter, on peut déceler une tendance: en moyenne, la valeur du désir de vengeance tend à décroître au fil du temps.