MATH 60604
Modélisation statistique
§ 5d - Équicorrélation

HEC Montréal
Département de sciences de la décision

Covariance du modèle d'équicorrélation

- Supposons que les observations d'un même groupe sont interchangeables. Plus précisément, supposons que la corrélation (conditionnelle aux variables explicatives) entre deux observations Y d'un groupe est toujours la même et que la variance (conditionnelle) de Y est constante.
- S'il y a cinq observations dans le groupe i , la matrice de covariance est

$$\mathbf{\Sigma}_i = \begin{pmatrix} \sigma^2+\tau & \tau & \tau & \tau & \tau \\ \tau & \sigma^2+\tau & \tau & \tau & \tau \\ \tau & \tau & \sigma^2+\tau & \tau & \tau \\ \tau & \tau & \tau & \sigma^2+\tau & \tau \\ \tau & \tau & \tau & \tau & \sigma^2+\tau \end{pmatrix}.$$

- Cette paramétrisation est utilisée par SAS.
- Il est important de comprendre que la covariance (conditionnelle) entre deux observations quelconques est τ et que la variance (conditionnelle) de chaque observation est $\sigma^2+\tau$.

Corrélation du modèle d'équicorrélation

La matrice de corrélation correspondante du modèle d'équicorrélation est

$$\mathbf{R}_i = \begin{pmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{pmatrix},$$

où $\rho = \tau/(\sigma^2 + \tau)$.

- La corrélation (conditionnelle) entre deux observations au sein d'un même groupe est toujours ρ. Cette structure de covariance porte le nom d'équicorrélation (*compound symmetry* en anglais) et dépend de deux paramètres σ^2 et τ.

Procédure `mixed` pour les données corrélées

Code SAS pour ajuster le modèle d'équicorrélation

```
/* Créer une copie de t */
data vengeance;
set modstat.vengeance;
tcat=t;
run;

proc mixed data=vengeance method=reml;
class id tcat;
model vengeance = sexe age vc wom t / solution;
repeated tcat / subject=id type=cs r=1 rcorr=1;
run;
```

Spécification de la structure de dépendance dans `proc mixed`

La commande **`repeated`** permet de définir la structure de dépendance des données.

- Le premier argument de la commande **`repeated`** spécifie l'ordre des observations au sein du groupe (utile pour les données longitudinales). Cette variable doit être catégorielle (déclarée via **`class`**).
- L'option **`subject`** définit la variable qui identifie les groupes d'observations.
- L'option **`type`** spécifie le modèle de covariance intra-groupe.
- L'option `r=1` (`rcorr=1`) demande que l'estimation de la matrice de covariance (corrélation) de la personne 1 soit présentée dans la sortie.

Comme nous voulons aussi utiliser la variable `t` dans le modèle en tant que variable continue, nous avons créé une autre variable `t` (`tcat` ici), afin de pouvoir l'utiliser comme argument de **`repeated`**.

Disgression technique

- Comme la structure utilisée est `cs`, nous n'aurions pas eu besoin de donner l'argument premier `tcat` à **`repeated`** car cette structure n'utilise pas l'ordre des observations à l'intérieur d'un groupe.
- L'ordre est en revanche important pour d'autres types de structure, comme le modèle autorégressif. Ça ne fait pas de tort de conserver l'information.

Spécification du modèle

Informations sur le modèle	
Table	WORK.TEMP
Variable dépendante	vengeance
Structure de covariance	Symétrie composée
Effet du sujet	id
Méthode d'estimation	REML
Méthode de variance résiduelle	Profil
Méthode SE des effets fixes	Basé(e) sur le modèle
Méthode des degrés de liberté	Between-Within

Informations sur les niveaux de classe

Classe	Niveaux	Valeurs
id	80	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
tcat	5	1 2 3 4 5

Matrices de covariance et de corrélation pour l'individu 1

Matrice R estimée pour id 1					
Ligne	Col1	Col2	Col3	Col4	Col5
1	0.3858	0.1374	0.1374	0.1374	0.1374
2	0.1374	0.3858	0.1374	0.1374	0.1374
3	0.1374	0.1374	0.3858	0.1374	0.1374
4	0.1374	0.1374	0.1374	0.3858	0.1374
5	0.1374	0.1374	0.1374	0.1374	0.3858

Matrice de corrélation R estimée pour id 1					
Ligne	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.3563	0.3563	0.3563	0.3563
2	0.3563	1.0000	0.3563	0.3563	0.3563
3	0.3563	0.3563	1.0000	0.3563	0.3563
4	0.3563	0.3563	0.3563	1.0000	0.3563
5	0.3563	0.3563	0.3563	0.3563	1.0000

Avec la structure d'équicorrélation, la corrélation est la même pour toutes les mesures répétées de l'individu 1.

Paramètres de la structure de covariance

Valeur estimée du paramètre de covariance		
Param. de cov.	**Sujet**	**Estimation**
CS	id	0.1374
Residual		0.2483

- La covariance intra-groupe du modèle d'équicorrélation est paramétrisé avec
  - Var $(Y_{ij}) = \sigma^2 + \tau$;
  - Cov $(Y_{ij}, Y_{ij'}) = \tau$.
- L'estimé de la covariance conditionnelle entre observations pour une personne donnée est $\widehat{\tau} = 0.137$.
- L'estimé de la covariance conditionnelle d'une observation est $\widehat{\tau} + \widehat{\sigma}^2 = 0.386$.

Exemple: structure de corrélation

- Par conséquent, l'estimation de la corrélation (conditionnelle) entre deux observations d'une même personne (corrélation intra-classe) est donc :

$$\widehat{\rho} = \frac{\widehat{\tau}}{\widehat{\tau} + \widehat{\sigma}^2} = \frac{0.137}{0.137 + 0.248} = 0.356.$$

- On peut retrouver ces valeurs dans les matrices de covariance et de corrélation fournies pour la première personne.
- **Vous devez être en mesure de reconstruire la corrélation sur la base de la sortie (d'où l'importance de comprendre la formule).**

Test du rapport de vraisemblance pour le paramètre τ

Test du rapport de vraisemblance du modèle nul		
DDL	khi-2	Pr > khi-2
1	67.25	<.0001

- On peut tester l'hypothèse $\mathscr{H}_0 : \tau = 0$ versus l'alternative $\mathscr{H}_1 : \tau \neq 0$ à l'aide d'un **test du rapport de vraisemblance**.
- Le tableau ci-dessus donne la statistique de test pour $\mathscr{H}_0 : \tau = 0$, qui correspond au modèle de covariance $\sigma^2\mathbf{I}$ du modèle de régression linéaire classique (modèle réduit), ajusté par REML.
- On conclut que le modèle réduit sans corrélation n'est pas une **simplification adéquate** du modèle complet avec une structure d'équicorrélation.
- **Le test du rapport de vraisemblance produit par SAS compare toujours le modèle ajusté au modèle linéaire homoscédastique sans corrélation.**

Test du rapport de vraisemblance à la mitaine

Tests d'ajustement	
-2 log-vraisemblance restreinte	709.4
AIC (préférer les petites valeurs)	713.4
AICC (préférer les petites valeurs)	713.4
BIC (préférer les petites valeurs)	718.2

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

- On peut obtenir la valeur de la statistique de test manuellement en comparant les valeurs de la vraisemblance REML des deux modèles, ici $-2\ell_{\text{reml}}(\widehat{\boldsymbol{\theta}}_0) = 776.7$ et $-2\ell_{\text{reml}}(\widehat{\boldsymbol{\theta}}) = 709.4$. La statistique est 67.3.
  - C'est la même valeur que dans le tableau, à arrondi près.
- La distribution nulle du test de rapport de vraisemblance est χ^2_1 (pourquoi?).
- On peut comparer la valeur du test au quantile 95% de la loi χ^2_1, 3.84. Puisque la valeur de la statistique est supérieure à 3.84, on rejette $\mathscr{H}_0$ à niveau $\alpha = 0.05$.

Estimés des paramètres de moyenne

Solution pour effets fixes

Effet	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|
Intercept	-0.1689	0.3422	75	-0.49	0.6231
sexe	0.1357	0.1060	75	1.28	0.2044
age	0.04586	0.007080	75	6.48	<.0001
vc	0.5225	0.03065	75	17.05	<.0001
wom	0.3989	0.03887	75	10.26	<.0001
t	-0.5675	0.01762	319	-32.21	<.0001

Le désir de vengeance décroît avec le temps, une fois qu'on contrôle pour l'effet des autres variables.

Estimés des coefficients

- Le modèle ajusté est toujours un modèle linéaire,

$$\widehat{\texttt{vengeance}} = -0.169 + 0.136\texttt{sexe} + 0.0459\texttt{age} + 0.523\texttt{vc} + 0.399\texttt{wom} - 0.568\texttt{t}.$$

- Les estimés des paramètres $\boldsymbol{\beta}$ sont exactement les mêmes que ceux du modèle de régression linéaire ordinaire.
- C'est vrai seulement pour le cas spécial du modèle d'équicorrélation avec le même nombre d'observations dans chaque groupe.
- En général, cependant, les estimés d'autres modèles ne seront pas très différents de ceux du modèle linéaire ordinaire s'ils ont les mêmes régresseurs, mais des structures de covariance différentes.

Comparaison des coefficients du modèle

Effet	Estimation	Erreur type	Effet	Estimation	Erreur type
Intercept	-0.1689	0.2249	**Intercept**	-0.1689	0.3422
sexe	0.1357	0.06748	**sexe**	0.1357	0.1060
age	0.04586	0.004507	**age**	0.04586	0.007080
vc	0.5225	0.01951	**vc**	0.5225	0.03065
wom	0.3989	0.02474	**wom**	0.3989	0.03887
t	-0.5675	0.02177	**t**	-0.5675	0.01762

- La précision des estimés $\widehat{\beta}$ change (à gauche modèle linéaire ordinaire, à droite le modèle avec équicorrélation).
- Les erreurs-types des paramètres sont plus grandes dans le modèle d'équicorrélation. Les conclusions sur la significativité des paramètres ne changent pas, hormis pour la variable `sexe` qui n'est plus significative.
- La corrélation intra-groupe rend les observations partiellement redondantes: on a moins d'information qu'avec un échantillon de même taille avec uniquement des données indépendantes, donc les estimés des paramètres sont moins précis.