MATH 60604
Modélisation statistique
§ 5h - Hétéroscédasticité de groupe

HEC Montréal
Département de sciences de la décision

Structure de covariance pour données groupées hétéroscédastiques

- On peut supposer que la structure de covariance est la même pour tous les groupes, mais que ces paramètres diffèrent pour chaque groupe.
- Si les données groupées sont consécutives, la matrice de covariance de toutes les observations est

$$\text{Cov}\,(Y) = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_m \end{pmatrix}.$$

- On suppose que $\boldsymbol{\Sigma}_1 \neq \cdots \neq \boldsymbol{\Sigma}_m$.

Hétéroscédasticité de groupes

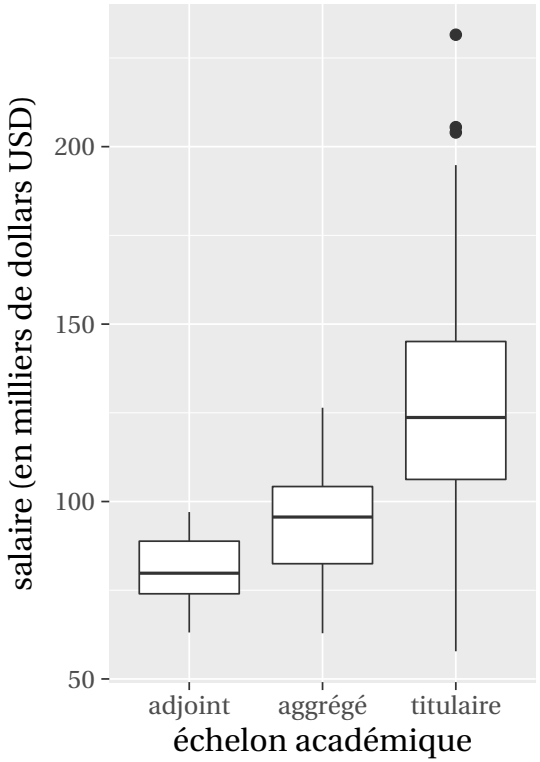
- Si toutes les mesures sont indépendantes (intra- et inter-groupes), mais qu'elles sont hétéroscédastiques par groupe, la matrice $\mathbf{\Sigma}_i = \sigma_i^2 \mathbf{I}$, où $\mathbf{I}$ est la matrice identité composée de uns sur la diagonale et de zéros hors diagonale.
- Il y a m paramètres de variance à estimer (une par groupe).
- On pourrait envisager une structure plus complexe pour $\mathbf{\Sigma}_i$. SAS permet cela, mais les blocs ne peuvent pas partager de paramètres et donc on aura m fois le nombre de paramètres de $\mathbf{\Sigma}_i$ à estimer. Pour cela, il faut suffisamment d'observations dans chaque groupe pour estimer les paramètres de covariance de manière fiable.

Discrimination salaire dans un collège américain

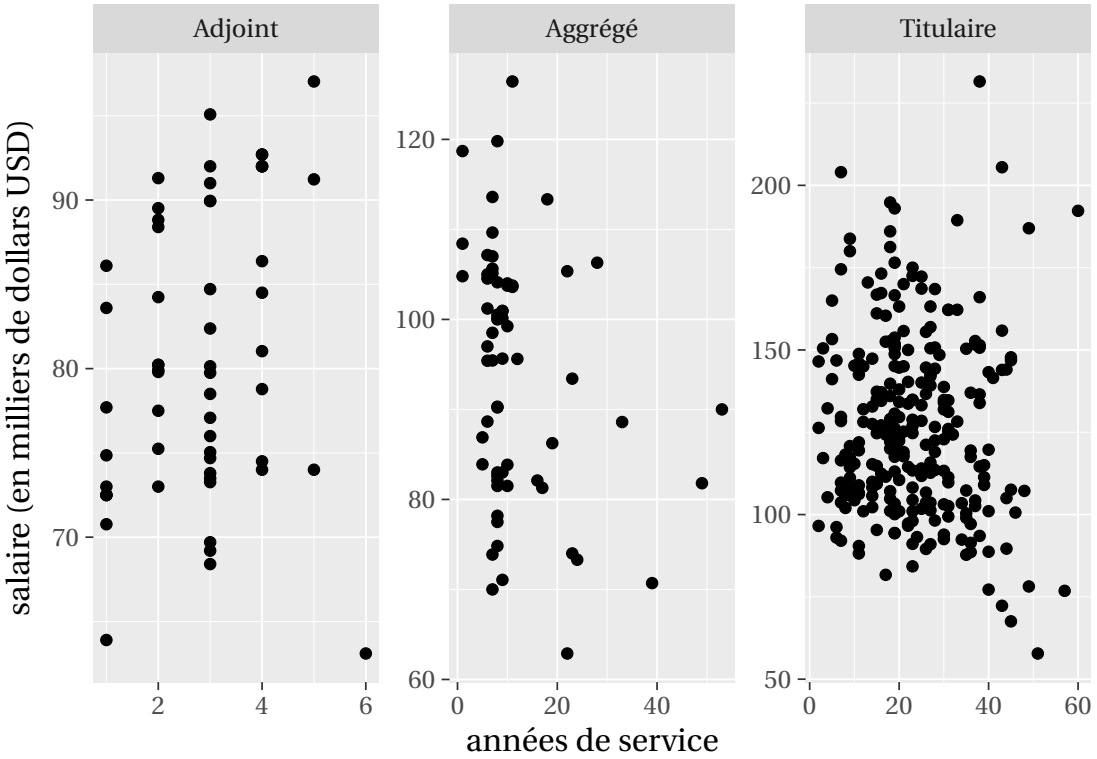
Les données `college` contiennent le salaire sur neuf mois (en milliers de dollars) pour 2008-2009 dans un collège américain.

- `salaire`: salaire de professeurs pendant l'année académique 2008–2009 (en milliers de dollars USD).
- `echelon`: échelon académique, soit adjoint, aggrégé ou titulaire.
- `domaine`: variable catégorielle indiquant le champ d'expertise du professeur, soit appliqué ou théorique.
- `sexe`: indicateur binaire pour le sexe, soit homme ou femme.

Boîte à moustache



Relation entre nombre d'années de service et salaire



L'analyse exploratoire montre clairement que la variance au sein des échelons diffère.

Modéliser l'hétéroscédasticité de groupe

Code SAS pour spécifier une variance différente pour chaque groupe

```
proc mixed data=modstat.college plots=studentpanel;
class sexe domaine echelon;
model salaire = sexe domaine echelon;
repeated / group = echelon;
run;
```

L'argument `repeated / group` renseigne SAS sur la structure des groupes.

Estimés des variances des groupes et test de significativité

Valeur estimée du paramètre de covariance

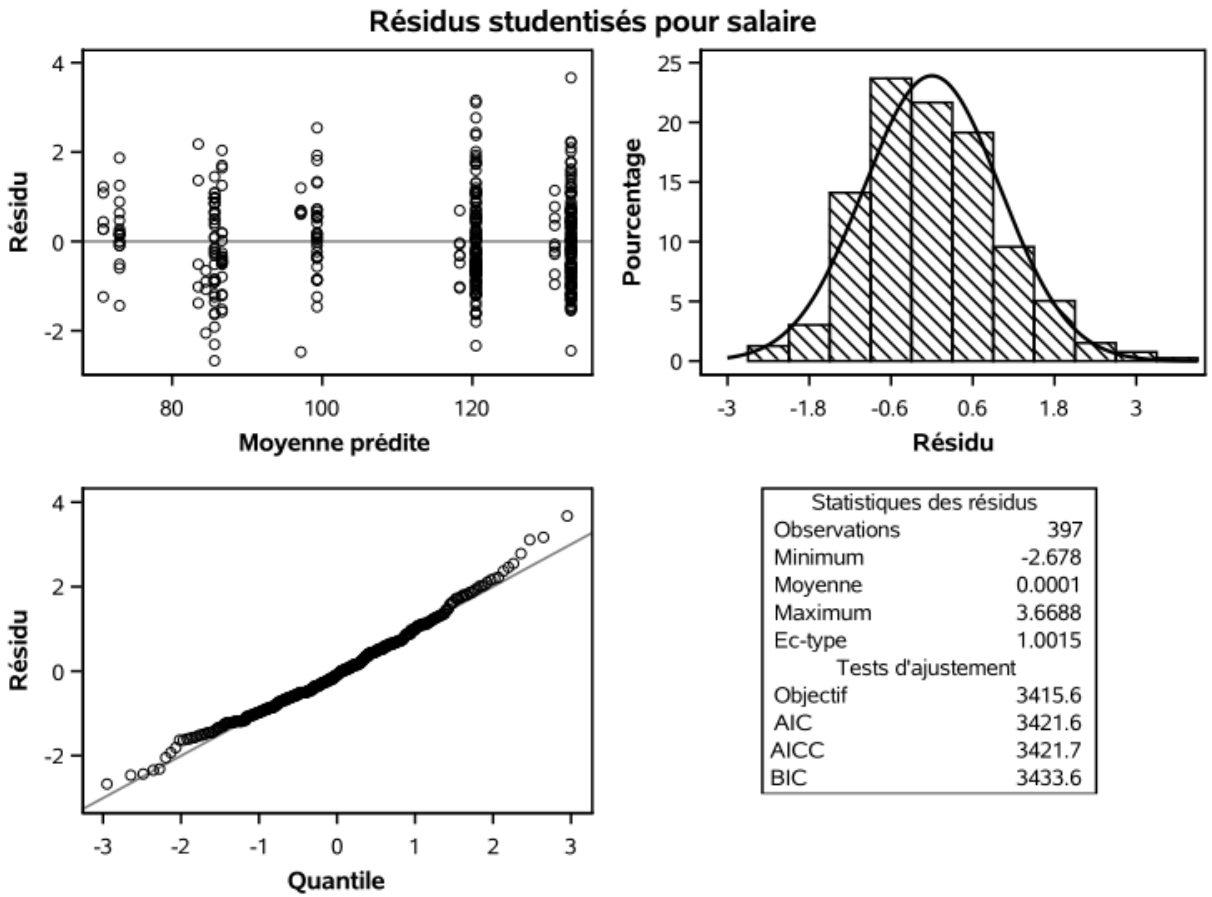
Param. de cov.	Groupe	Estimation
Residual	echelon adjoint	42.4817
Residual	echelon aggrege	115.29
Residual	echelon titulaire	722.44

Test du rapport de vraisemblance du modèle nul

DDL	khi-2	Pr > khi-2
2	164.78	<.0001

La variance croît avec l'échelon. Le test du rapport de vraisemblance montre que le modèle qui suppose des variances différentes au sein de chaque échelon est significativement meilleur que le modèle linéaire ordinaire, qui suppose une variance constante pour toutes les observations.

Diagnostics graphiques pour les données `salaireprofs`



Le diagramme des résidus studentisés versus les valeurs ajustées est conforme aux attentes. On peut être confiant pour notre inférence sur les paramètres de la moyenne.

Estimés des paramètres de la moyenne

Tests des effets fixes de type 3

Effet	DDL num.	DDL den.	Valeur F	Pr > F
sexe	1	392	1.55	0.2141
domaine	1	392	92.85	<.0001
echelon	2	392	334.46	<.0001

- Comparer le salaire des hommes et des femmes à l'aide d'un test-*t* est incorrect, car le rang à un impact important sur le salaire.
- Cela est dû à la plus faible proportion de femmes qui sont titulaires (7%) que pour les adjointes et les aggrégées (16%).
- Une fois que l'on a pris en compte l'hétéroscédasticité de groupe et l'effet de l'échelon, il n'y a pas de preuve de discrimination salariale et les écarts observés sont explicables.