MATH 60604A
Statistical modelling
§ 5h - Group heteroscedasticity

HEC Montréal
Department of Decision Sciences

Block covariance structure for group heteroscedasticity

- Assume that observations within groups have the same covariance structure, but the parameters of the latter differ between groups.
- Assuming consecutive grouped measurements, the covariance matrix of all the measurements is

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

- We assume that $\boldsymbol{\Sigma}_1 \neq \cdots \neq \boldsymbol{\Sigma}_m$.

Group heteroscedasticity

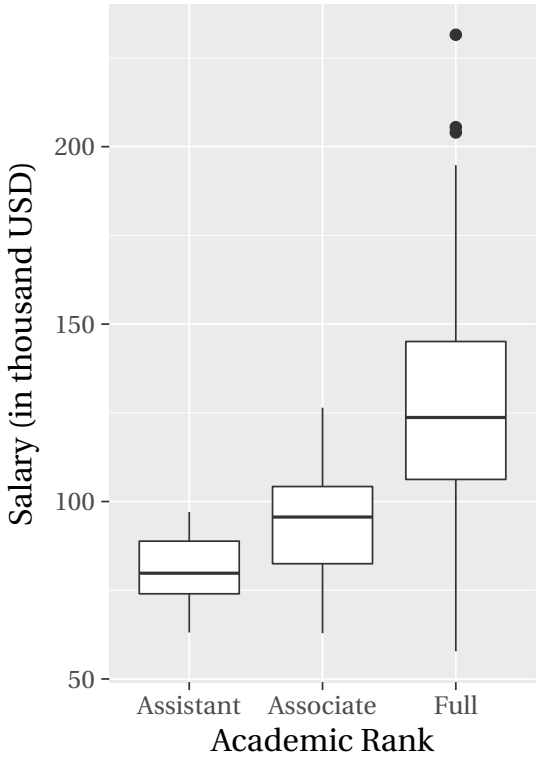
- If the data are independent (within and between group), but heteroscedastic between groups, the matrix $\mathbf{\Sigma}_i = \sigma_i^2 \mathbf{I}$, where $\mathbf{I}$ is the identity matrix with ones on the diagonal and zero for off-diagonal entries.
- In this case, there are m variance parameters to estimate (one per group).
- We could use a different structure for $\mathbf{\Sigma}_i$. SAS allows this, but the blocks cannot share parameters, so we get m times the number of parameters in $\mathbf{\Sigma}_i$. There must be enough observations in each group to reliably estimate the covariance parameters.

Wage discrimination in a US college

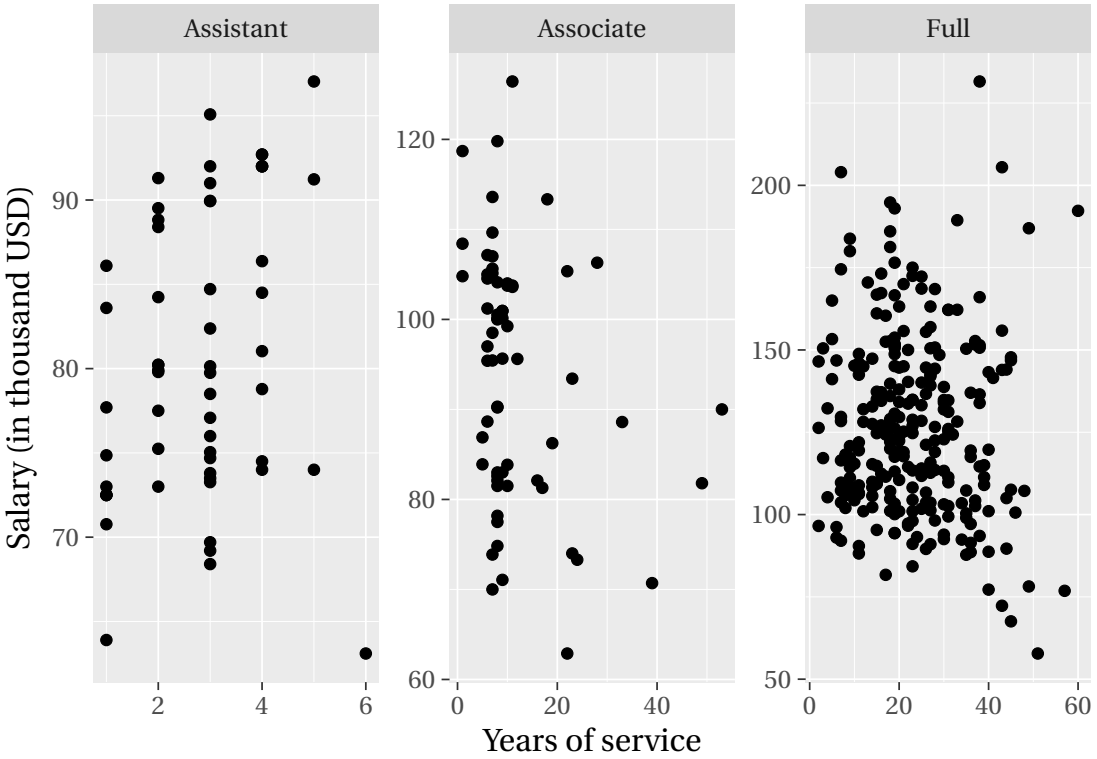
The `college` data set provides the nine-month academic salary (in thousand dollars) in 2008–2009 of professors in a college in the USA.

- `salary`: nine month income (in thousand dollars).
- `rank`: academic rank of the professor (assistant , associate or full).
- `field`: categorical variable indicating whether research field is applied or theoretical.
- `sex`: sex of individual, either man or woman.

Box-and-whiskers plots



Relationship between number of years of service and salary



The explanatory data analysis shows clear heteroscedasticity within academic rank.

Dealing with group heteroscedasticity

SAS code for a different variance per group

```
proc mixed data=statmod.college plots=studentpanel;
class field rank sex;
model salary = sex field rank;
repeated / group = rank;
run;
```

The argument `repeated / group` specifies the group structure.

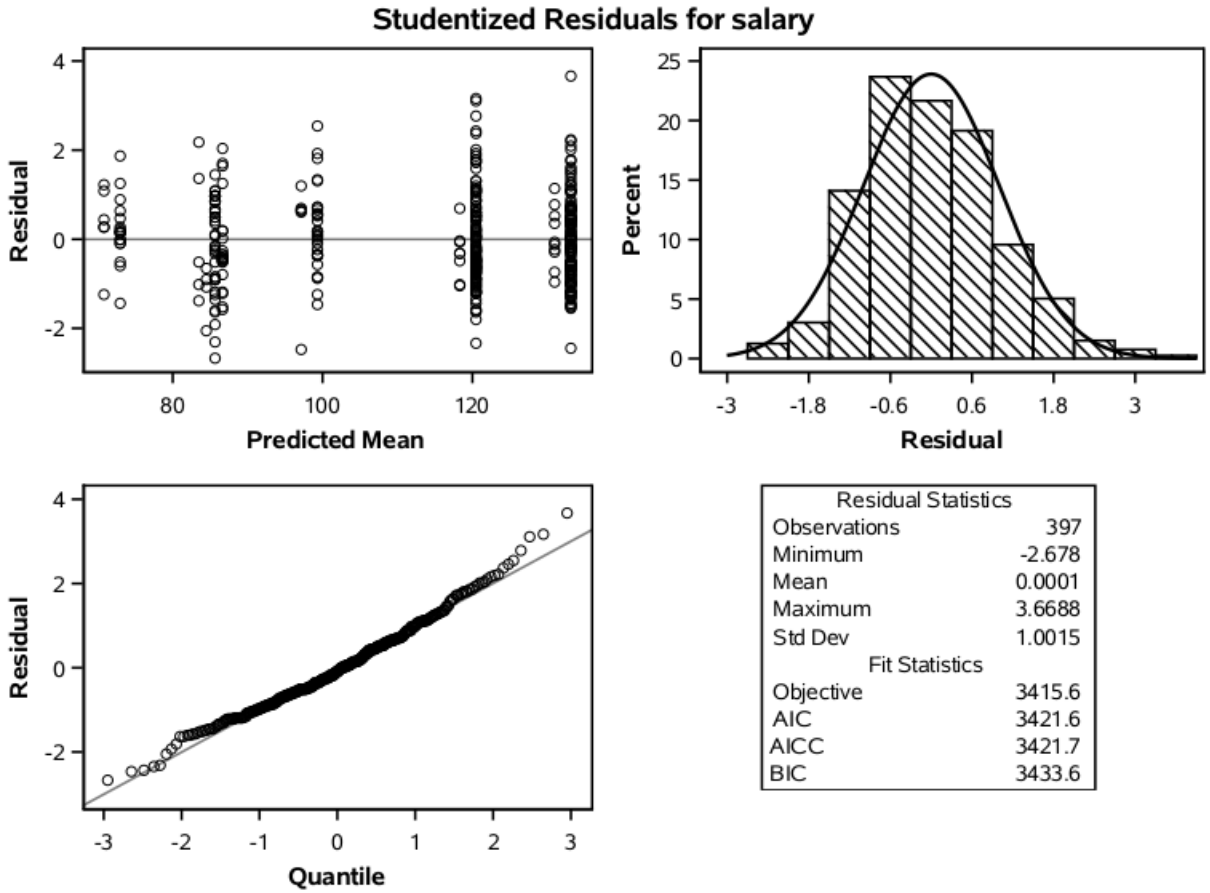
Variance estimates per group and significance test

Covariance Parameter Estimates		
Cov Parm	**Group**	**Estimate**
Residual	rank assistant	42.4817
Residual	rank associate	115.29
Residual	rank full	722.44

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
2	164.78	<.0001

The variance increases with rank. The likelihood ratio test shows that the model with a different group for each rank is significantly better than the linear model which assumes a constant variance for every observation.

Diagnostic plots for the `profsalaries` data



The residual plots show that the model captures most features well. We can be confident in our inference.

Mean parameter estimates

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
sex	1	392	1.55	0.2141
field	1	392	92.85	<.0001
rank	2	392	334.46	<.0001

- Solely comparing the salary of mean and women academics using a two-sample test is wrong, because rank is an important explanatory variable.
- Moreover, the proportion of full professor that are women (7%) is much lower than for assistant or associated professors (16%)
- After accounting for rank and dealing with group heteroscedasticity, there is no evidence of gender gap.