

MATH 60604A
Statistical modelling
§ 6b - Clustered data example

HEC Montréal
Department of Decision Sciences

Example: worker motivation

We consider an example of clustered data.

- A large business collected data about its employees through a questionnaire.
- In this example, the response variable is **worker motivation**.
- The level of worker motivation is defined as the **sum** of three items, namely
 - I share several of the company's values.
 - I feel loyal to the company.
 - I'm proud to tell people what company I work for.

measured on a Likert scale, ranging from strongly disagree (1) to strongly agree (5).

- These data originates from *Lee, H.-J. and Peccei, R. (2007). Organizational-Level Gender Dissimilarity and Employee Commitment. British Journal of Industrial Relations, 45, 687–712.*

The `motivation` data set contains the variables

- `nunit`: number of employees in the unit (department).
- `idunit`: id of unit in which the employee worked.
- `idemployee`: id of employee (within unit).
- `yrserv`: years of service of employee.
- `sex`: sex of employee, either male (0) or female (1).
- `agemanager`: age of the unit manager (in years).
- `motiv`: worker motivation score.

- One possible source of correlation between observations is the `unit` (or `department`) (group variable).
- It's possible that motivation is affected by what unit an employee belongs to due to factors such as temperature or type of work, among other things.

Summary statistics

In the `motivation` data, units have different number of employees. For example, unit 1 has nine employees (three women and six men).

Obs.	nunit	idunit	idemployee	yrserv	sex	motiv	agemanager
1	9	1	1	16	0	8	40
2	9	1	2	18	0	6	40
3	9	1	3	17	1	7	40
4	9	1	4	16	0	8	40
5	9	1	5	13	0	9	40
6	9	1	6	3	0	13	40
7	9	1	7	10	0	9	40
8	9	1	8	4	1	14	40
9	9	1	9	13	1	10	40

There are two types of variables, both of which can be included in the model:

1. those fixed for all individuals in the unit (`nunit` and `agemanager`)
2. others that vary within unit (`yrserv` and `sex`).

Grouping variable and study objective

- In the longitudinal `revenge` example, the group variable was the individual and we only had explanatory variables that were fixed for each person, i.e., fixed in time, with the exception of the time variable itself.
- Here there are 100 groups in total, and 1016 observations in the file.
- The goal is to study the impact of sex, years of service, unit size, and of the age of the manager on worker mobilisation.
- However, we must account for the potential within-unit correlation. There is no natural ordering for the observations within a unit (contrary to the `revenge` example which contained repeated measurements over time).

Covariance structure for worker motivation

- We can consider the compound symmetry covariance structure to account for within-unit correlation.
 - This means that we assume that the (conditional) correlation between a pair of observations in the same unit is always the same.

SAS code for a linear model with equicorrelated errors

```
proc mixed data=statmod.motivation method=reml;  
class idunit;  
model motiv = sex yrserv agemanager nunit / solution;  
repeated / subject=idunit type=cs r=1 rcorr=1;  
run;
```

Covariance matrix within-unit (unit 1)

Estimated R Matrix for idunit 1

Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9
1	1.3709	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448
2	0.2448	1.3709	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448
3	0.2448	0.2448	1.3709	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448
4	0.2448	0.2448	0.2448	1.3709	0.2448	0.2448	0.2448	0.2448	0.2448
5	0.2448	0.2448	0.2448	0.2448	1.3709	0.2448	0.2448	0.2448	0.2448
6	0.2448	0.2448	0.2448	0.2448	0.2448	1.3709	0.2448	0.2448	0.2448
7	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	1.3709	0.2448	0.2448
8	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	1.3709	0.2448
9	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	1.3709

Compound symmetry model

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
CS	idunit	0.2448
Residual		1.1261

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
1	79.49	<.0001

- The estimated covariance parameter from the compound symmetry model is $\hat{\tau} = 0.2448$. It is significantly different from 0, suggesting a positive correlation between the worker motivation score between employees in the same unit, after adjusting for the effects of the explanatory variables.
- This estimated correlation between workers within-unit is $\hat{\rho} = 0.1785$.

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	13.7633	0.3955	97	34.80	<.0001
sex	0.5622	0.06835	914	8.23	<.0001
yrserv	-0.4722	0.006015	914	-78.50	<.0001
agemanager	0.01929	0.006801	97	2.84	0.0056
nunit	0.006470	0.02019	97	0.32	0.7493

- The effects of three explanatory variables are significant: **sex**, **yrserv** and **agemanager**.
 - women are more motivated than men, on average.
 - The longer a person has been employed at the company, the less (s)he is motivated.
 - The older the manager, the more motivated the employee.
- However, the size of the unit is not significant.

Example: worker motivation

- It might be interesting to include an effect for the unit variable in the model.
- But, as we've already seen, when we include a fixed effect for each group, we lose the ability to estimate the effects of variables which are fixed within group.
- This would mean we could not include the variables `agemanager` or `nunit`.
- However, there is still a way to include a "group effect" while also keeping the possibility of including variables which are fixed within group.
- We need to use **random effects** instead of fixed effects, which can also be used to model the covariance structure.