

MATH 60604A  
Statistical modelling  
§ 6a - Group effects

HEC Montréal  
Department of Decision Sciences

# Inclusion of group effects

- So far, we have only accounted for group structure by modelling the within-group correlation.
- We may also want to include a **group effect** in the mean model, i.e., a different intercept for each group.
- This is done by adding the categorical group variable  $g$  as explanatory variable in the mean model, which translates into  $m - 1$  indicator variables  $\mathbf{1}_{g=i}$  for  $i = 1, \dots, m - 1$  if there are  $m$  groups.

# Equation for group effect

- Suppose that we only include the categorical variable  $g$  representing groups,

$$Y_{ij} = \beta_0 + \sum_{i=1}^{m-1} \beta_i \mathbf{1}_{g=i} + \varepsilon_{ij},$$

- for the baseline (group  $m$ ), the intercept is  $\beta_0$ ,
- the group effect for  $g = i$  is  $\beta_i$  ( $i = 1, \dots, m - 1$ ), and the overall group-specific "intercept" is  $\beta_0 + \beta_i$ .

We consider a regression model for `revenge` with a group effect to illustrate the challenges.

- The idea here is to model the fact that desire for revenge can vary between subjects.
- In the current example, there are only five observations per person to estimate the group effect.
- The model will ignore the within-person correlation for now.

## SAS code to fit a linear model via REML

```
proc mixed data=revenge method=reml;  
class id;  
model revenge = id sex age vc wom t / solution;  
run;
```

In addition to the categorical variable `id`, the model includes the same explanatory variables as before. Each person has his/her own "intercept" parameter (`id=80` is the baseline category).



Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
<b>id</b>	75	319	3.77	<.0001
<b>sex</b>	0	.	.	.
<b>age</b>	0	.	.	.
<b>vc</b>	0	.	.	.
<b>wom</b>	0	.	.	.
<b>t</b>	1	319	1037.49	<.0001

There are **no** parameters estimates or tests for the variables **sex**, **age**, **vc** or **wom**, but there is for the time variable **t**. Because some covariates are fixed over time, their effect are not uniquely estimable (perfect collinearity). If we remove **id** from the model, we can however estimate their effects (hence 75 df rather than 79 in the *F*-table).

- Once we've included a fixed effect for each person, **it is impossible to include any variable that does not vary in time for a single person.**
- The variables `sex`, `age`, `vc` and `wom` are fixed in time for each person (`vc` and `wom` were only measured once, at time 1).
- These variables are already implicitly included in the individual effect. There is **perfect collinearity** between a variable fixed in time, and the `id` variable.
- This means that we can perfectly predict the value of `sex` (and the three others) by only looking at the `id` variable.
- Therefore, we cannot have a fixed effect for each individual while simultaneously including variables that are fixed in time for each subject.



# Challenges arising from the inclusion of a group effect

- Group is a categorical variable: we need enough observations in each group to reliably estimate the group effects.
- If the number of groups  $m$  is large relative to the overall sample size, there may also be too many parameters in the model.
- We cannot estimate the effect of variables that do not vary within group if we add group effects.

# Model with group effect and correlation structure

The model fitted next includes only `id` and the time variable `t` as explanatory variables in the mean model, but we specify in addition an AR(1) correlation structure within-individual for the errors  $\varepsilon$ .

## SAS code to include a group effect with AR(1) correlation

```
proc mixed data=revenge method=reml;
class id tcat;
model revenge = id t / solution;
repeated tcat / subject=id type=ar(1);
run;
```

The effect of the AR(1) correlation parameter is significant (likelihood ratio test statistic of 21.68, negligible  $p$ -value under  $\chi_1^2$ ). The estimate of the time effect is  $-0.5684$ , very close to that we got in the model including `sex`, `age`, `vc` and `wom`, and the AR(1) structure model in the previous chapter.

# Remark on model comparison

- We have to be careful **not to use** the AIC and BIC reported in the output to compare this model to the earlier one including `sex`, `age`, `vc` and `wom`, since we used the REML estimation method (the default).
- AIC and BIC obtained through REML, are **not comparable** if the "mean" parts of the models (fixed effects) are not the same.
- If we want to compare these models, we must use the maximum likelihood estimator (option `method=ml` when calling `proc mixed`).

Model Information	
Data Set	WORK.REVENGE
Dependent Variable	revenge
Covariance Structure	Autoregressive
Subject Effect	id
Estimation Method	ML

## Remark on model comparison

We fit both models with an AR(1) structure for the errors using maximum likelihood.

Model	AIC	BIC	$\hat{\rho}$ ( $p$ -value)
sex, age, vc, wom, t	666.1	685.1	0.48 ( $10^{-20}$ )
id, t	653.4	851.1	-0.013 (0.83)

- The preferred model according to AIC includes `id`, but AIC tends to select complicated models.
- The preferred model according to BIC includes `sex`, `age`, `vc` and `wom` and throws away the `id` variable.
- Once we include an individual effect for group, the correlation structure seems to be unnecessary — the estimated coefficient is even negative, which is counter-intuitive and suggests the model is over-parametrized.

## Remark on model comparison

- The choice of covariates depends on the type of study. If we're interested in studying the effects of one or more of the variables `sex`, `age`, `vc` or `wom`, then we don't have any choice: we must choose a model that contains all of them.
- If we're only interested in the time effect, then the two models will come to the same conclusion either way.
- Often, the optimization routine fails — we cannot estimate both the  $\beta$  and the covariance matrix parameters.
- It is possible to include variables that are fixed within group (within person in our example) **and** group effects (`id` in our example) at the same time by using [random effects](#).