MATH60619A
Statistical analysis and inference
§ 7 - Introduction to mixed effects models

Léo Belzile

HEC Montréal
Department of Decision Sciences

Chapter overview

Chapter 7 - Introduction to mixed effects models

Inclusion of group effect for the mean

Random effect models

Prediction for mixed effect models

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 $\mathbf{1}_{\mathrm{g}=i}$ for $i=1,\ldots,m-1$ if there are m groups.
- Suppose that we only include the categorical variable g representing groups,

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

  - for the baseline (group m), the intercept is β_0,
  - the group effect for $\mathrm{g}=i$ is β_i $(i=1,\ldots,m-1)$, and the overall group-specific “intercept” is $\beta_0+\beta_i$.

Linear model with a group effect for the `revenge` data

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.

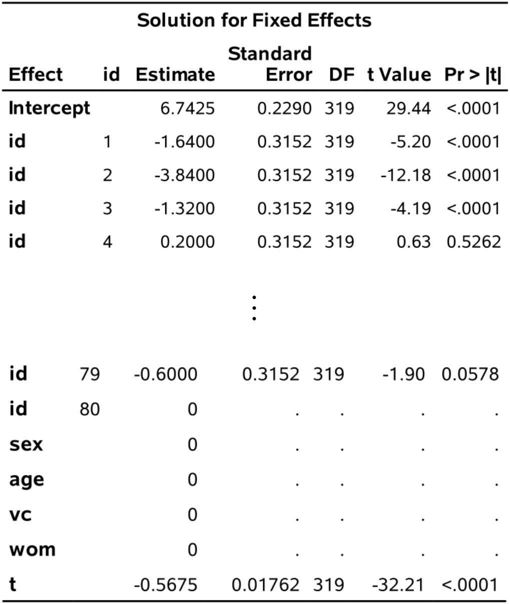
Model with fixed effects for subject

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).

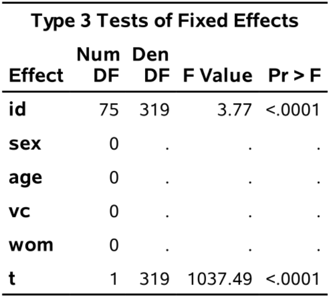
Estimates of fixed effects



Solution for Fixed Effects

Effect	id	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		6.7425	0.2290	319	29.44	<.0001
id	1	-1.6400	0.3152	319	-5.20	<.0001
id	2	-3.8400	0.3152	319	-12.18	<.0001
id	3	-1.3200	0.3152	319	-4.19	<.0001
id	4	0.2000	0.3152	319	0.63	0.5262
			⋮			
id	79	-0.6000	0.3152	319	-1.90	0.0578
id	80	0	.	.	.	.
sex		0	.	.	.	.
age		0	.	.	.	.
vc		0	.	.	.	.
wom		0	.	.	.	.
t		-0.5675	0.01762	319	-32.21	<.0001

Parameter significance



Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
id	75	319	3.77	<.0001
sex	0	.	.	.
age	0	.	.	.
vc	0	.	.	.
wom	0	.	.	.
t	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).

Collinearity

- 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 ε.

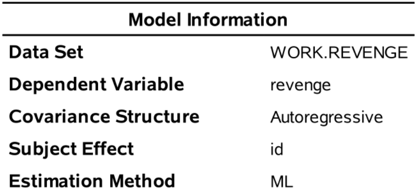
SAS code to include a group effect with AR(1) correlation

```
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 χ^2_1). 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 $\mathrm{AR}(1)$ structure for the errors using maximum likelihood.

Model	**AIC**	**BIC**	$\widehat{\rho}$ (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 β 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.

Introduction to random effects models

Random effects give another way of accounting for within-group correlation and allows prediction of group-level effects in addition to population-level effects.

- The main characteristic of the linear mixed model is to allow certain variables to have random effects, i.e., to have parameters that vary from one group to another (from one person to another in repeated measures data).
- While each group is allowed an individual effect, the overall average of these effects is zero.

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.*

motivation data

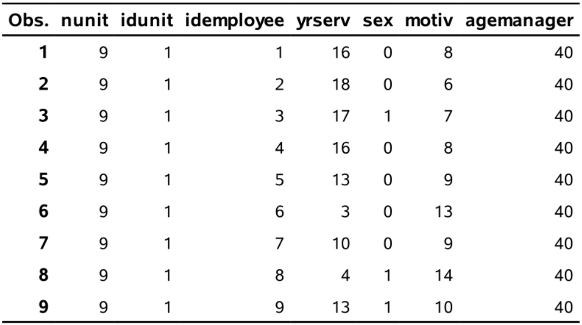
The data are found in the file motivation.sas7bdat and contains the variables

- 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).

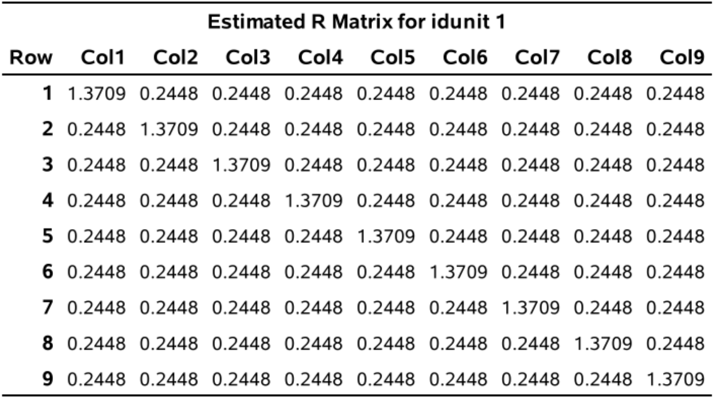
Covariance structure for worker motivation

- 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;
```

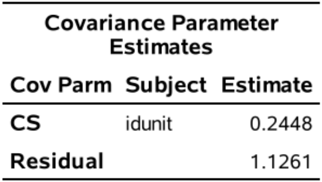
Covariance matrix within-unit (unit 1)



Estimated R Matrix for idunit 1

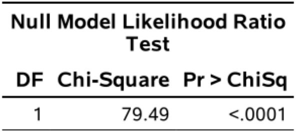
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9
1	1.3709	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448
2	0.2448	1.3709	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448
3	0.2448	0.2448	1.3709	0.2448	0.2448	0.2448	0.2448	0.2448	0.2448
4	0.2448	0.2448	0.2448	1.3709	0.2448	0.2448	0.2448	0.2448	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

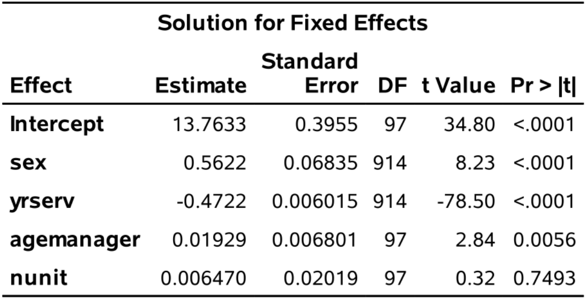


Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	79.49	<.0001

- The estimated covariance parameter from the compound symmetry model is $\widehat{\tau} = 0.2448$. 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 $\widehat{\rho} = 0.1785$.

Fixed effect estimates



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.

Random effects models

- When an explanatory variable is modeled with a random effect, we assume that the total effect of this variable is a combination of
  1. a common effect for the entire population and
  2. a within-group effect.
- For example, when considering repeated measures from the same individuals, the effect of a variable can be split into a common effect for all individuals in the population, and a unique effect for each individual.
- In the example of worker motivation, the effect of years of service could be split up into a common effect for all employees (in all units) and a unique effect in each unit for employees.

Random effects models

- The simplest random effects model is one with only a group-specific intercept, assumed random.
- The equation of the linear regression model is

$$Y_{ij} = \beta_0 + \beta_1 \mathsf{X}_{ij1} + \beta_2 \mathsf{X}_{ij2} + \cdots + \beta_p \mathsf{X}_{ijp} + \varepsilon_{ij}$$

  for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$ and where Y_{ij} is observation j from group i.
- Assume for now that the $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.
- Suppose that we want to allow the intercept β_0 to vary from one group to another, but with a random effect.

Random effects models

- The model becomes

$$Y_{ij} = (\beta_0 + b_i) + \beta_1 \mathsf{X}_{ij1} + \beta_2 \mathsf{X}_{ij2} + \cdots + \beta_p \mathsf{X}_{ijp} + \varepsilon_{ij}$$

- The intercept specific to group i is $\beta_0 + b_i$. It consists of
  - A common effect over all groups, β_0;
  - A group-specific effect, b_i.

Assumptions of the random intercept model

The model equation is

$$Y_{ij} = (\beta_0 + b_i) + \beta_1 \mathsf{X}_{ij1} + \beta_2 \mathsf{X}_{ij2} + \cdots + \beta_p \mathsf{X}_{ijp} + \varepsilon_{ij}$$

- The $b_1, \ldots, b_m$ quantities are assumed to be independent random variables (and also independent from the ε terms and the explanatory variables).
- We assume both random intercepts and errors are normally distributed,
  - $b_i \sim \mathcal{N}(0, \sigma_b^2)$ for all i ($i = 1, \ldots, m$).
  - $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.
- The ε_{ij} terms could be correlated within group i, but they are assumed independent for the time being.

Random effects models: mean

In this model, we still have the so-called marginal mean of Y_{ij},

$$\mathsf{E}\left(Y_{ij} \mid \mathbf{X}_i\right) = \beta_0 + \beta_1 \mathrm{X}_{ij1} + \beta_2 \mathrm{X}_{ij2} + \cdots + \beta_p \mathrm{X}_{ijp}.$$

- At the population level, the mean of Y_{ij} is only a function of the fixed effects.

We also have the conditional mean of Y_{ij}, which depends on the group-specific effect,

$$\mathsf{E}\left(Y_{ij} \mid \mathbf{X}_i, b_i\right) = \beta_0 + b_i + \beta_1 \mathrm{X}_{ij1} + \beta_2 \mathrm{X}_{ij2} + \cdots + \beta_p \mathrm{X}_{ijp}$$

- It is possible to get predicted values for the random variables b_i.

Random effects models

- With this kind of model, we can estimate the mean $\mathsf{E}\left(Y_{ij} \mid \mathbf{X}_i, b_i\right)$ which can be thought of as a prediction of the value of Y_{ij} after accounting for the group-specific effects.
- In repeated measures data, this allows us to predict an individual's trajectory while accounting for specific effects of the individual.
- One interesting fact is that when we add a random effect for the group variable, we can still estimate effects of variables that are fixed within a group.

Random effects models: variance/covariance

Since it's random, the term b_i introduces a within-group correlation in the model. Because ε_{ij} is independent of b_i for all i, j, the (conditional) variance of an observation is

$$\mathrm{Var}\left(Y_{ij} \mid \mathbf{X}_i\right) = \mathrm{Var}\left(b_i\right) + \mathrm{Var}\left(\varepsilon_{ij}\right) = \sigma_b^2 + \sigma^2$$

The covariance between two individuals in the same group is

$$\mathrm{Cov}\left(Y_{ij}, Y_{ik} \mid \mathbf{X}_i\right) = \sigma_b^2, \qquad j \neq k.$$

Consequently, the correlation between two individuals in the same group is

$$\mathrm{Corr}\left(Y_{ij}, Y_{ik} \mid \mathbf{X}_i\right) = \frac{\sigma_b^2}{\sigma^2 + \sigma_b^2}, \qquad j \neq k.$$

This quantity is often called the intra-class correlation.

Mathematical aside

Both β_j and explanatories are assumed non-random, thus

$$\begin{aligned}\mathsf{Cov}\left(Y_{ij}, Y_{ik} \mid \mathbf{X}_i\right) &= \mathsf{Co}\left(\beta_0 + b_i + \beta_1 \mathrm{X}_{ij1} + \cdots + \varepsilon_{ij},\right.\\ &\qquad\qquad \left.\beta_0 + b_i + \beta_1 \mathrm{X}_{ik1} + \cdots + \varepsilon_{ik} \mid \mathbf{X}_i\right)\\ &= \mathsf{Cov}\left(b_i + \varepsilon_{ij}, b_i + \varepsilon_{ik}\right)\\ &= \mathsf{Var}\left(b_i\right) + \mathsf{Cov}\left(\varepsilon_{ij}, \varepsilon_{ik}\right) = \sigma^2_b.\end{aligned}$$

where the last step follows from independence of b_i and ε's and because $\mathsf{Cov}\left(Y_{ij}, Y_{ij}\right) = \mathsf{Var}\left(Y_{ij}\right)$.

Variability of $\boldsymbol{Y}$

Alternatively,

$$\begin{aligned}\mathsf{Var}\left(\boldsymbol{Y}_i \mid \mathbf{X}_i\right) &= \mathsf{Var}\begin{pmatrix} b_i \\ \vdots \\ b_i \end{pmatrix} + \mathsf{Var}\begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix} \\ &= \begin{pmatrix} \sigma^2_b & \sigma^2_b & \cdots & \sigma^2_b \\ \sigma^2_b & \sigma^2_b & \cdots & \sigma^2_b \\ \vdots & \ddots & \ddots & \vdots \\ \sigma^2_b & \sigma^2_b & \cdots & \sigma^2_b \end{pmatrix} + \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}.\end{aligned}$$

Compound symmetry correlation induced by the random intercept model

- We can see that introducing a random effect for the intercept implies that the observations within the same group are correlated, and the correlation is the same regardless of which individuals are considered (equicorrelation).
- The difference is that now the correlation must be non-negative, since σ_b^2 is a variance whereas the correlation for the compound symmetry model was

$$-\frac{1}{\max(n_i)+1} \leq \rho \leq 1.$$

- This limitation is not usually of consequence, because within-group correlations tend to be positive.
- **Adding a random intercept is the same as using the compound symmetry structure.**

Adding a random intercept with the `random` command

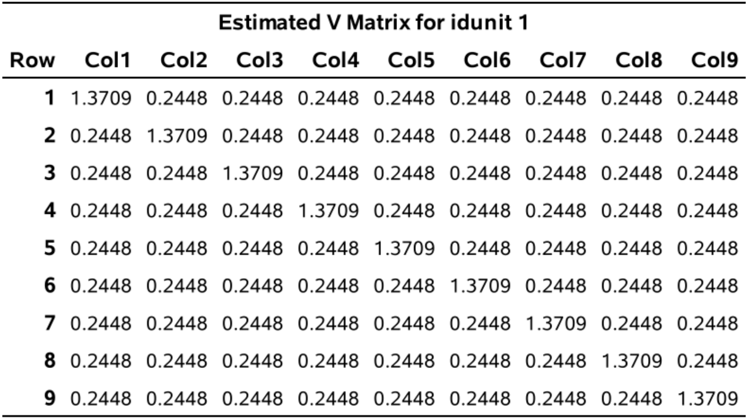
- The command `repeated` allows us to specify the covariance structure for the errors in `proc mixed`.
- If we don't use the `repeated` command, the errors are assumed to be independent.

SAS code for a random intercept model with independent errors

```
proc mixed data=statmod.motivation;
class idunit;
model motiv = sex yrserv agemanager nunit / solution;
random intercept / subject=idunit v=1 vcorr=1;
run;
```

Including a random intercept naturally induces a compound symmetry correlation structure. Therefore, we do not need to specify anything for the covariance structure of the errors.

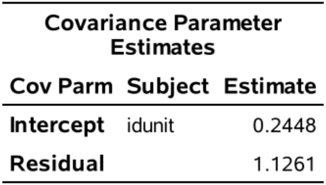
Covariance matrix specified by the random intercept model



Estimated V 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

Covariance parameter estimates



Covariance Parameter Estimates

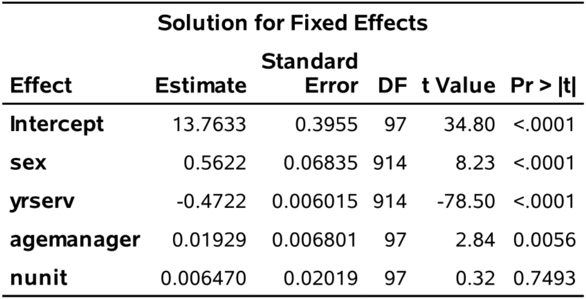
Cov Parm	Subject	Estimate
Intercept	idunit	0.2448
Residual		1.1261

- The variance estimate for the random intercept is $\widehat{\sigma}_b^2 = 0.2448$, whereas the estimate of the variance of the error term is $\widehat{\sigma}^2 = 1.1261$.
- Consequently, the estimate of the within-unit correlation is

$$\widehat{\rho} = \frac{\widehat{\sigma}_b^2}{\widehat{\sigma}_b^2 + \sigma^2} = 0.1785.$$

- This is exactly the same correlation for the observation than that obtained from compound symmetry covariance model for the errors (command `repeated`).

Fixed effect estimates



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 the explanatory variables (and their standard errors) are also the same as for the compound symmetry structure — both models are equivalent for the response assuming the within-unit correlation is positive.

Longitudinal data

- In the worker motivation example, the natural correlation structure for the data is the equicorrelation structure, which assumes that each pair of observations within a group has the same correlation, because the groups represent units in a company.
- We've seen that it's not necessary to model the covariance structure by assuming a structure on the errors of the model; we just need to include a random intercept which will automatically induce a compound symmetry structure on the data.
- However, this kind of structure might not be valid in longitudinal data, and we might want to use an $\mathrm{AR}(1)$ structure, as we saw in the customer revenge example.

Longitudinal data

- In longitudinal data, specifying an $\mathrm{AR}(1)$ structure for the data can be done by adding a structure on the errors of the data, as well as through the specification of random effects.
- The model is

$$Y_{ij} = \beta_0 + b_i + \beta_1 \mathrm{X}_{ij1} + \beta_2 \mathrm{X}_{ij2} + \cdots + \beta_p \mathrm{X}_{ijp} + \varepsilon_{ij}$$

  but we now assume
  - for the random intercept term, $b_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2)$,
  - for the errors, $\varepsilon_i \sim \mathcal{N}\left(\mathbf{0}_{n_i}, \boldsymbol{\Sigma}_i\right)$,
  - ε_{ij} are not independent within each subject i — $\boldsymbol{\Sigma}_i$ is not diagonal.
- We can assume an $\mathrm{AR}(1)$ structure (or a structure other than compound symmetry) on the covariance matrix of the errors $\boldsymbol{\Sigma}_i$.

Side remark: while the matrices $\boldsymbol{\Sigma}_i$ may share the same parameter values for each subject i, they need not be the same size, hence the subscript.

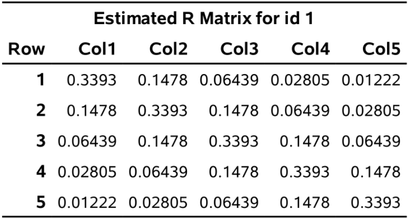
Random intercept with AR(1) structure

SAS code for the random intercept model with AR(1) errors

```
proc mixed data=revenge;
class id tcat;
model revenge = sex age vc wom t / solution;
random intercept / subject=id v=1 vcorr=1;
repeated tcat / subject=id type=ar(1) r=1 rcorr=1;
run;
```

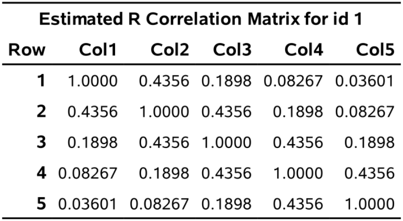
- The option `v=1 vcorr=1` tells SAS to output the covariance/correlation matrix for the $\boldsymbol{Y}$ observations for subject 1.
- The option `r=1 rcorr=1` tells SAS to output the covariance/correlation matrix for the errors ε for subject 1.

Covariance and correlation matrices for the **errors**



Estimated R Matrix for id 1

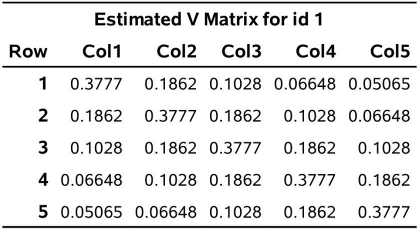
Row	Col1	Col2	Col3	Col4	Col5
1	0.3393	0.1478	0.06439	0.02805	0.01222
2	0.1478	0.3393	0.1478	0.06439	0.02805
3	0.06439	0.1478	0.3393	0.1478	0.06439
4	0.02805	0.06439	0.1478	0.3393	0.1478
5	0.01222	0.02805	0.06439	0.1478	0.3393



Estimated R Correlation Matrix for id 1

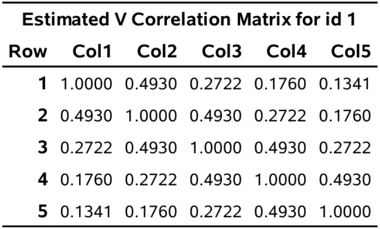
Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.4356	0.1898	0.08267	0.03601
2	0.4356	1.0000	0.4356	0.1898	0.08267
3	0.1898	0.4356	1.0000	0.4356	0.1898
4	0.08267	0.1898	0.4356	1.0000	0.4356
5	0.03601	0.08267	0.1898	0.4356	1.0000

Covariance and correlation matrices for the **observations**



Estimated V Matrix for id 1

Row	Col1	Col2	Col3	Col4	Col5
1	0.3777	0.1862	0.1028	0.06648	0.05065
2	0.1862	0.3777	0.1862	0.1028	0.06648
3	0.1028	0.1862	0.3777	0.1862	0.1028
4	0.06648	0.1028	0.1862	0.3777	0.1862
5	0.05065	0.06648	0.1028	0.1862	0.3777

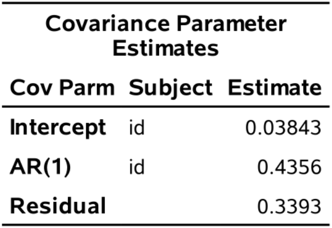


Estimated V Correlation Matrix for id 1

Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.4930	0.2722	0.1760	0.1341
2	0.4930	1.0000	0.4930	0.2722	0.1760
3	0.2722	0.4930	1.0000	0.4930	0.2722
4	0.1760	0.2722	0.4930	1.0000	0.4930
5	0.1341	0.1760	0.2722	0.4930	1.0000

- Note that these matrices are different than the ones for the errors in the previous slide.
- The covariance of the data is the sum of that of the random effects and of the covariance on the errors.

Covariance parameters



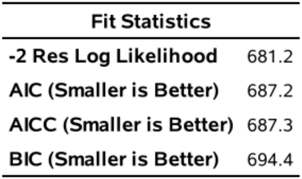
Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	id	0.03843
AR(1)	id	0.4356
Residual		0.3393

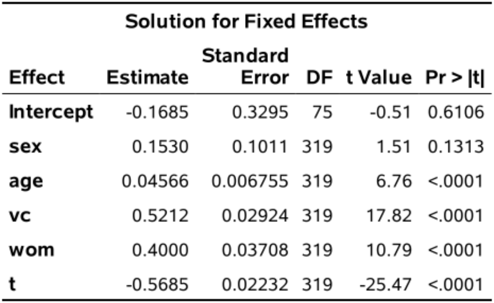
There are three covariance parameters in the model, the estimates of which are

- $\widehat{\sigma}_b^2 = 0.038$ for the variance of the random effect,
- $\widehat{\sigma}^2 = 0.33$ for the variance of the errors ε,
- $\widehat{\rho} = 0.43$ for the lag-one correlation of the $\mathrm{AR}(1)$ model.

Information criteria and fixed effects



Fit Statistics	
-2 Res Log Likelihood	681.2
AIC (Smaller is Better)	687.2
AICC (Smaller is Better)	687.3
BIC (Smaller is Better)	694.4



Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-0.1685	0.3295	75	-0.51	0.6106
sex	0.1530	0.1011	319	1.51	0.1313
age	0.04566	0.006755	319	6.76	<.0001
vc	0.5212	0.02924	319	17.82	<.0001
wom	0.4000	0.03708	319	10.79	<.0001
t	-0.5685	0.02232	319	-25.47	<.0001

These estimates are for the mixed model with a random intercept for `id` and an AR(1) structure for the errors. We can compare it using information criteria to the model that includes does not include the random intercept from Chapter 6.

Fixed or random effect?

- The theory we cover generalizes to more complex setting (e.g., we could allow random slopes for each groups).
- Loosely speaking, the main difference between fixed and random effects is
  - **fixed effects** for group are used when we have few groups and lots of replicates and we care about the effect of the group (small m, large n_i).
  - **random effects** are used when there are enough levels of the factor group to estimate the variance σ^2_b reliably; we are not interested in the effects per say (large m, small n_i).
- Testing whether a random effect is needed or not is equivalent to testing if the variance of the random effect $\sigma^2_b = 0$; this is a non-standard testing problem, which is beyond the scope of this course…

Conditional and marginal means

- At the population level, the marginal mean of Y_{ij} is, as usual,

$$\mathsf{E}\left(Y_{ij} \mid \mathbf{X}_i\right) = \beta_0 + \beta_1 \mathrm{X}_{ij1} + \beta_2 \mathrm{X}_{ij2} + \cdots + \beta_p \mathrm{X}_{ijp}.$$

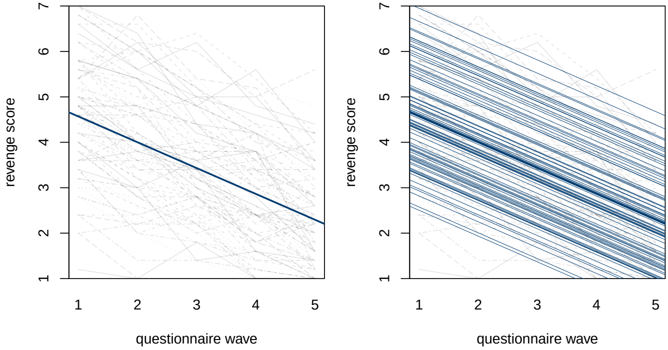
- However, the conditional mean of Y_{ij}, given by the group specific effects, is

$$\mathsf{E}\left(Y_{ij} \mid \mathbf{X}_i, b_i\right) = (\beta_0 + b_i) + \beta_1 \mathrm{X}_{ij1} + \cdots + \beta_p \mathrm{X}_{ijp}.$$

- By predicting the random effect b_i, we can predict the value of Y_{ij} while accounting for the group-specific random effect for the intercept.

Graphical illustration

We can display the random intercept for the `revenge` data. Both models are fitted with $\mathrm{AR}(1)$ covariance for the errors and `t` as fixed effect.



Model without (left) and with random intercept for `id` (right).

Prediction

- According to the model specification presented in the last section, the terms, $\boldsymbol{b}$, are random variables and not parameters (i.e., fixed quantities but unknown).
- We can always get predictions for these random variables.
- Be careful to not confuse prediction with estimation. Prediction is for random variables.
- Once we have predicted values for the $\boldsymbol{b}$ terms and estimates for the fixed effect parameters, β, we can get predictions for the outcome variables $\boldsymbol{Y}$.

Prediction: model **without** random effects

- If there are no random effects in the model (for example, if we had fitted a model that directly specified the covariance structure using `repeated`), then we make predictions in the same way as we did for ordinary linear regression.
- That is, the prediction for Y_{ij} is

$$\widehat{Y}_{ij} = \widehat{\beta}_0 + \widehat{\beta}_1 \mathrm{X}_{ij1} + \ldots + \widehat{\beta}_p \mathrm{X}_{ijp}.$$

- This quantity is also the estimate of the mean (at the population level) of the response variable.

Prediction: model **with** random effect

- If there are random effects in the model, the estimation of the mean (at the population level) of the response variable for an individual with the characteristics of individual j from group i is

$$\widehat{Y}_{ij} = \widehat{\beta}_0 + \widehat{\beta}_1 \mathsf{X}_{ij1} + \ldots + \widehat{\beta}_p \mathsf{X}_{ijp}.$$

- But we can also get predictions of the values of the response variable for individual j in group i
- For example, in a model with a random intercept b_i,

$$\widehat{Y}_{ij} = \widehat{\beta}_0 + \widehat{b}_i + \widehat{\beta}_1 \mathsf{X}_{ij1} + \ldots + \widehat{\beta}_p \mathsf{X}_{ijp}.$$

- If, however, we want to get predictions for a new individual that was not included in the original dataset, then we have no choice but to use the mean prediction, because the random effect estimate of this group is not available.

Predictions for random effects

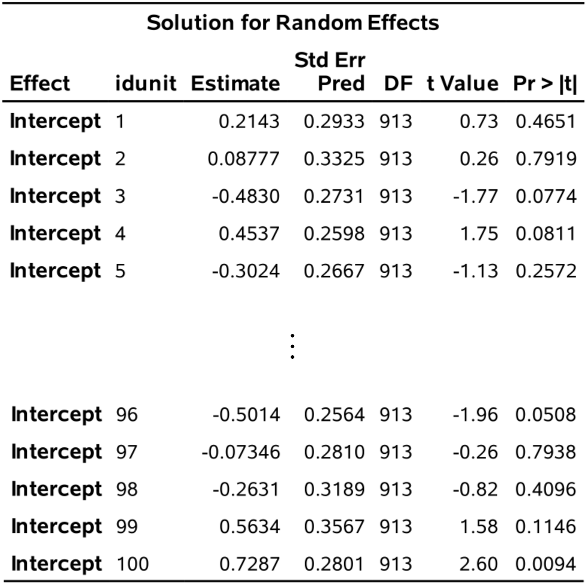
Let’s revisit the last model for this example, with a random intercept.

SAS code for the random intercept model

```
proc mixed data=statmod.motivation;
class idunit;
model motiv = sex yrserv agemanager nunit / solution;
random intercept / subject=idunit type=vc solution;
ods output Mixed.SolutionR=re;
run;
```

The option `solution` in the command `random` is used to get predictions of the random effects. The command `ods output` saves these in order to make diagnostic plots for the random effets.

Predictions of the random effects

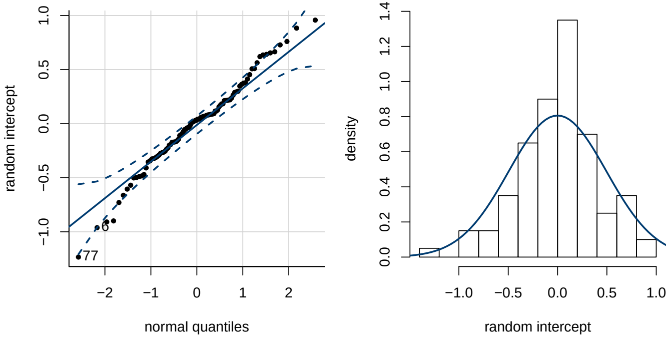


Solution for Random Effects

Effect	idunit	Estimate	Std Err Pred	DF	t Value	Pr > |t|
Intercept	1	0.2143	0.2933	913	0.73	0.4651
Intercept	2	0.08777	0.3325	913	0.26	0.7919
Intercept	3	-0.4830	0.2731	913	-1.77	0.0774
Intercept	4	0.4537	0.2598	913	1.75	0.0811
Intercept	5	-0.3024	0.2667	913	-1.13	0.2572
		⋮				
Intercept	96	-0.5014	0.2564	913	-1.96	0.0508
Intercept	97	-0.07346	0.2810	913	-0.26	0.7938
Intercept	98	-0.2631	0.3189	913	-0.82	0.4096
Intercept	99	0.5634	0.3567	913	1.58	0.1146
Intercept	100	0.7287	0.2801	913	2.60	0.0094

Histogram of random effects

We can plot histograms and quantile-quantile plots of the predicted random intercepts.



These can help us check the normality assumption of the random effects (think of these as further residual diagnostics). Note that (by construction), the average of these random effects is always zero.

Predictions for observations Y

- With `proc mixed`, we can save the values for all observations in the data file:
  - Predictions for the mean of the population (fixed effects),
  - Individual predictions (fixed and random effects).
- This is done using the options `outpm` and `outp`, respectively, in the `model` command.
- **Trick**: If you want predictions for new individuals, you can just include these observations in the data file, while leaving the Y variable blank (with "." in SAS). These individuals will not be used in estimation of the model, but we will retrieve their predicted values.

Prediction for new employees

We assume that we want to get predictions for two new employees, one of whom is part of a unit already present in the dataset (`idunit=1`) and one that is part of a unit not in the original dataset (`idunit=101`).

SAS code to input two new observations

```
data newdata;
input nunit idunit idemployee yrserv sex
     motiv agemanager;
cards;
9 1 10 5 0 . 40
9 101 1 5 0 . 40;
run;

/* Merge observations with database */
data motivation;
set statmod.motivation newdata;
run;
```

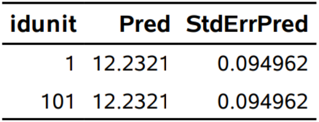
Code to fit the model and get predictions

SAS code to output predictions from a mixed model

```
proc mixed data=motivation;
class idunit;
model motiv = sex yrserv agemanager nunit
     / solution outp=prediction outpm=mean;
random intercept / subject=idunit type=vc;
run;
```

- The data file used is `data=motivation`, which contains the 1018 observations, but only the 1016 observations from the original file are used in fitting the model.
- However, predictions will be made for all 1018 observations in the files `mean` and `prediction`.

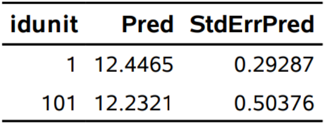
Population mean of the two new subject (file `mean`)



idunit	Pred	StdErrPred
1	12.2321	0.094962
101	12.2321	0.094962

- The fitted mean (12.23) is the same in both cases because only the fixed effects were used and the two employees have the same values for the explanatory variables.

Predictions for the two new subjects (file `prediction`)



idunit	Pred	StdErrPred
1	12.4465	0.29287
101	12.2321	0.50376

- This time, the random effects are used if they're available. Since unit 1 was present in the model fitting, its random effect is used in making the prediction (12.45).
- However, the unit 101 was absent when fitting the model. Therefore, the prediction for the employee in unit 101 is only based on the fixed effects in the model, meaning that we get the same predicted value (12.23) as before.
- The standard errors for the individual predictions are larger, reflecting the added individual uncertainty arising from the errors and the random effects.