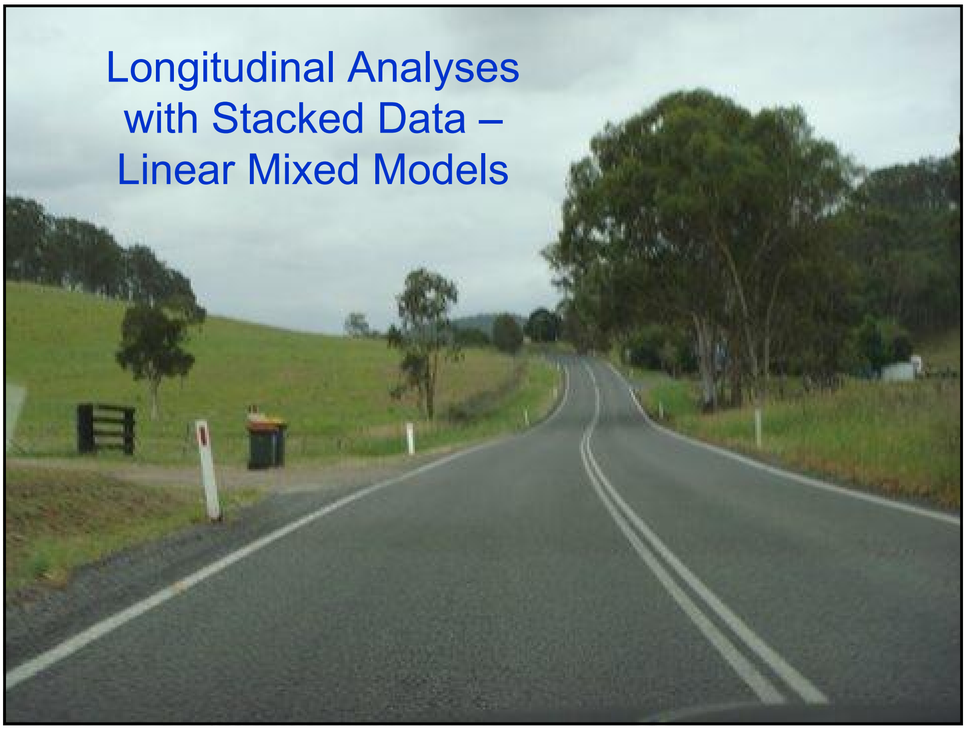
Longitudinal Analyses with Stacked Data – Linear Mixed Models



Outline

- Distinguish wide and long (stacked) data
- Use *restructure data* to stack repeated measures data
- Carry out a repeated measures analysis with stacked data using GLM
- Distinguish the fixed, random and nested parts of the model
- Implications of using the stacked form when there are missing data
- Implications of different forms of 'missingness'
- The same analysis using the *linear mixed model* procedure
- Advantages of the linear mixed model procedure
  - Maximum likelihood estimation
  - Model different correlations between
  - Allow random variation around coefficients as well as the intercept
  - Can be seen as multi-level models
- A between-subject by within-subject analysis
- An individual growth (random coefficient) model

subid	chsexw1	relp1w1a	relp1w2a	relp1w3a
11110002	2	22	16	17
11110014	2	20	16	20
11120008	2	15	14	14
11190001	2	16	16	18
11200006	2	18	12	15
11220054	2	15	11	16
11220156	2	10	10	11
11350010	2	10	12	8
11440010	2	12	11	12
22081017	2	12	14	12
22111005	2	13	13	11
22111013	1	23	20	16
22111017	2	8	7	7
22161034	1	26	24	24
22181002	2	9	11	10
22201001	2	12	13	12
22231019	1	16	17	10
22291023	1	20	14	14
22301011	1	18	15	10
22331005	2	19	16	15
22341002	1	14	13	11
22361001	1	17	9	9

Number of cases read: 22 Number of cases listed: 22

Data in *wide* format – one line per subject

In *long* (*stacked*) format, each subject will be represented by three lines

1=male, 2=female

Stacking data

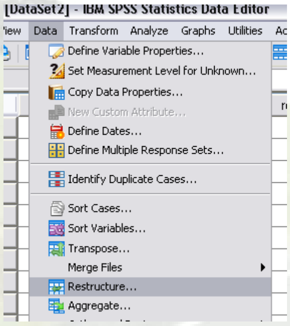
varstocases make relp from relp1w1a relp1w2a relp1w3a/
index=time(3)/
keep=subid chsexw1.

*ccc.sav [DataSet1] - IBM SPSS Statistics Data Editor

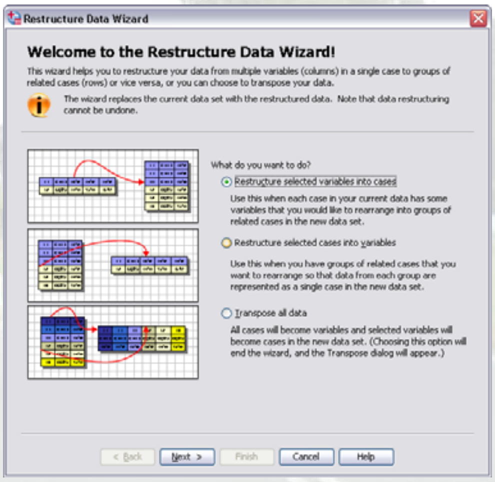
File Edit View Data Transform Analyze Graphs Utilities Add-ons Windo

	subid	chsexw1	time	relp
1	11110002	2	1	22
2	11110002	2	2	16
3	11110002	2	3	17
4	11110014	2	1	20
5	11110014	2	2	16
6	11110014	2	3	20
7	11120008	2	1	15
8	11120008	2	2	14
9	11120008	2	3	14

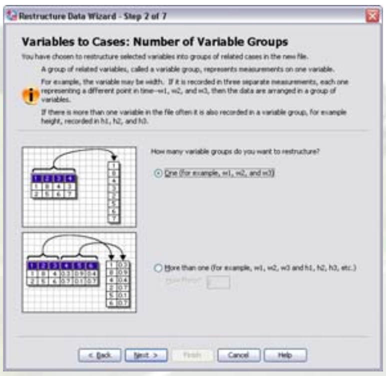
Stacking data – point-and-click



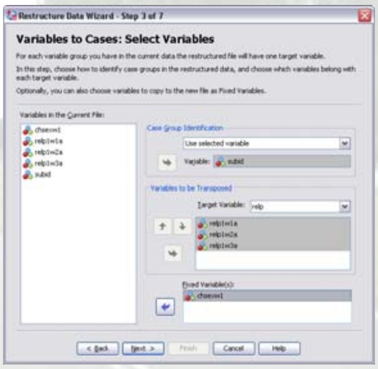
➔



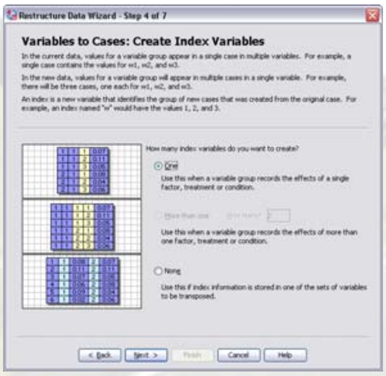
Stacking data – point-and-click (2)



➔



Stacking data – point-and-click (3)



➔

A GLM analysis with stacked data

Subject will be a random factor. Need to provide numbering for subject factor levels within each group.

```
sort cases by chsexw1 subid.
compute subwgrp = subwgrp + 1.
if (subid eq lag(subid))subwgrp = lag(subwgrp).
if (chsexw1 ne lag(chsexw1))subwgrp = 1.
leave subwgrp.
execute.
```

- The *lag* function allows reference to the value of a variable on the previous line of data
- The *leave* command tells SPSS not to initialise the variable to system-missing for the next case

A GLM analysis with stacked data (2)

subid	chsexw1	time	relp	subwgrp
22111013	1	1	23	1.00
22111013	1	2	20	1.00
22111013	1	3	16	1.00
22161034	1	1	26	2.00
22161034	1	2	24	2.00
22161034	1	3	24	2.00
22231019	1	1	16	3.00
22231019	1	2	17	3.00
22231019	1	3	10	3.00
22291023	1	1	20	4.00
22291023	1	2	14	4.00
22291023	1	3	14	4.00
22301011	1	1	18	5.00
22301011	1	2	15	5.00
22301011	1	3	10	5.00
22341002	1	1	14	6.00
22341002	1	2	13	6.00
22341002	1	3	11	6.00
22361001	1	1	17	7.00
22361001	1	2	9	7.00
22361001	1	3	9	7.00
11110002	2	1	22	1.00
11110002	2	2	16	1.00
11110002	2	3	17	1.00
11110014	2	1	20	2.00

The subject number is now included, starting at 1 for each group (sex)

A GLM analysis with stacked data (3)

```
glm relp by time chsexw1 subwgrp/
  random=subwgrp/
  design=time chsexw1 subwgrp(chsexw1)
     time by chsexw1.
```

- *subwgrp* is specified to be a random factor
  - this means that the levels of the factor (subjects in this case) are a random sample of all possible levels
- *subwgrp(chsexw1)* specifies that *subwgrp* is *nested* within *chsexw1.* Nesting occurs when (for example) the levels of factor A which occur under level 1 of factor B are different from those which occur under level 2 of factor B.

A GLM analysis with stacked data (4)

Between-Subjects Factors

		Value Label	N
time	1		22
	2		22
	3		22
chsexw1	1	Male	21
	2	Female	45
subwgrp	1.00		6
	2.00		6
	3.00		6
	4.00		6
	5.00		6
	6.00		6
	7.00		6
	8.00		3
	9.00		3
	10.00		3
	11.00		3
	12.00		3
	13.00		3
	14.00		3
	15.00		3

Tests of Between-Subjects Effects

Dependent Variable:relp

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	12671.079	1	12671.079	305.985	.000
	Error	828.216	20	41.411[a]		
time	Hypothesis	107.335	2	53.668	17.025	.000
	Error	126.089	40	3.152[b]		
chsexw1	Hypothesis	115.072	1	115.072	2.779	.111
	Error	828.216	20	41.411[a]		
subwgrp(chsexw1)	Hypothesis	828.216	20	41.411	13.137	.000
	Error	126.089	40	3.152[b]		
time * chsexw1	Hypothesis	57.032	2	28.516	9.046	.001
	Error	126.089	40	3.152[b]		

a. MS(subwgrp(chsexw1))
b. MS(Error)

Note that GLM automatically chooses the error terms, as long as the appropriate terms are included In the design. Note that the error term for *time* and *time* by *chsexw1* is the *time* by *subwgrp(chsexw1)* interaction.

A GLM analysis with stacked data (5)

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
time	Sphericity Assumed	107.335	2	53.668	17.025	.000
	Greenhouse-Geisser	107.335	1.810	59.316	17.025	.000
	Huynh-Feldt	107.335	2.000	53.668	17.025	.000
	Lower-bound	107.335	1.000	107.335	17.025	.001
time * chsexw1	Sphericity Assumed	57.032	2	28.516	9.046	.001
	Greenhouse-Geisser	57.032	1.810	31.517	9.046	.001
	Huynh-Feldt	57.032	2.000	28.516	9.046	.001
	Lower-bound	57.032	1.000	57.032	9.046	.007
Error(time)	Sphericity Assumed	126.089	40	3.152		
	Greenhouse-Geisser	126.089	36.191	3.484		
	Huynh-Feldt	126.089	40.000	3.152		
	Lower-bound	126.089	20.000	6.304		

Note that the results are the same as those obtained with the wide dataset, using

glm relp1w1a relp1w2a relp1w3a by chsexw1/ wsfactor=time 3/ design.

Tests of Between-Subjects Effects

Measure:MEASURE_1
Transformed Variable:Average

Source	df	F	Sig.
chsexw1	1	2.779	.111
Error	20		

Tests of Between-Subjects Effects

Dependent Variable:relp

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	12671.079	1	12671.079	305.985	.000
	Error	828.216	20	41.411[a]		
time	Hypothesis	107.335	2	53.668	17.025	.000
	Error	126.089	40	3.152[b]		
chsexw1	Hypothesis	115.072	1	115.072	2.779	.111
	Error	828.216	20	41.411[a]		
subwgrp(chsexw1)	Hypothesis	828.216	20	41.411	13.137	.000
	Error	126.089	40	3.152[b]		
time * chsexw1	Hypothesis	57.032	2	28.516	9.046	.001
	Error	126.089	40	3.152[b]		

a. MS(subwgrp(chsexw1))
b. MS(Error)

Missing Data (1)

An important advantage of using stacked data is that we don't lose a whole case when it has missing observations at one or two time points.

But what about bias due to the missing observations?

Here we need to consider formulations regarding missing data, developed by Rubin.

Missing Data (2)

MCAR – missing completely at random
- 'missingness' unrelated to values of *Xi* and *Y*

MAR - missing at random
- 'missingness' unrelated to *Y,* adjusted for *X*i

MNAR – missing not at random
- 'missingness' related to *Y*
- also referred to as *non-ignorable missing*

For details, see Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147-177

Missing Data (3)

An example given by Schafer & Graham (p. 152-153):

All patients have BP measured at Time 1. At Time 2, some patients do not have a BP measurement (i.e., it is missing). The T1 BP is X, the T2 BP is Y.

MCAR: Those with missing Y values are a random sample of patients at T1.

MAR: Those who returned at T2 had BP > 140 at T1

MNAR: Only patients whose BP was > 140 at T2 had their BPs recorded.

Missing Data (4) and Maximum Likelihood

Rubin showed that we can make valid inferences about the population in the presence of MAR (as well as MCAR) if we use analyses based on <u>maximum-likelihood</u> methods.

Keith refers briefly to ML (p. 394) – it is the default method for most SEM programs.

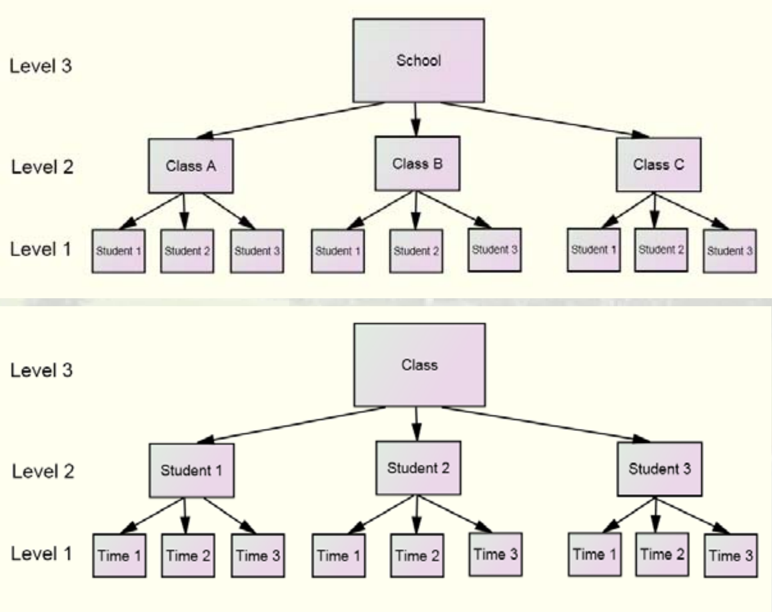
In a nutshell, ML is an iterative method which seeks to find estimates of the population parameters (coefficients, for example) which maximise the likelihood of obtaining the observed sample. (Originated by R. A. Fisher)

GLM does not use ML, the linear mixed model does.

Therefore, for analyses which involve missing data or are unbalanced, we turn to the *linear mixed model*.

Mixed and multilevel models

- Mixed models are so-called because they contain both *fixed* and *random* factors.
- In repeated measures analyses, *subject* is a random factor.
- We'll consider mixed models as multi-level models, because this is leads to a framework which is useful for more complex models.
- Judith Singer wrote an article which made the link between mixed and multi-level models. Peugh and Enders based their article on Singer's.



Multilevel structures

A two-level model (observations within subjects)

Based on the notation of Peugh & Enders (p. 721)

i = ith observation, j = jth subjects, γ_{00} = grand mean

Level 1 (observation): $Y_{ij} = B_{0j} + r_{ij}$

Level 2 (subject): $B_{0j} = \gamma_{00} + U_{0j}$

Combined: $Y_{ij} = \gamma_{00} + U_{0j} + r_{ij}$

Note: γ = gamma

A two-level model (observations within subjects)

Based on the notation of Peugh & Enders (p. 721)

i = ith observation, j = jth subject, γ_{00} = grand mean

Level 1 (observations): $Y_{ij} = B_{0j} + r_{ij}$

Score for observation *i* for subject *j*

Difference between observation *i* and the mean for subject *j*

Mean (intercept) for subject *j*

A two-level model (observations within subjects)

Based on the notation of Peugh & Enders (p. 721)

i = ith observation, j = jth subject, γ_{00} = grand mean

Level 2 (subject): $B_{0j} = \gamma_{00} + U_{0j}$

Mean (intercept) for subject j

Grand mean

Difference between mean for subject j and grand mean

A two-level model (observations within subjects)

Based on the notation of Peugh & Enders (p. 721)

i = ith observation, j = jth subject, γ_{00} = grand mean

Combined: $Y_{ij} = [\gamma_{00}] + [U_{0j} + r_{ij}]$

Fixed part

Random part

This is an unconditional random intercept model

A two-level model – introducing two fixed factors

i = ith observation, j = jth subject, γ_{00} = grand mean

Level 1 (observations): $Y_{ij} = B_{0j} + B_{1j}\text{time2}_{ij} + B_{2j}\text{time3}_{ij} + r_{ij}$

Level 2 (subjects): $B_{0j} = \gamma_{00} + \gamma_{01}\text{sex}_j + U_{0j}$

Model: $Y_{ij} = [\gamma_{00} + \gamma_{01}\text{sex} + \gamma_{10}\text{time2}_{ij} + \gamma_{20}\text{time3}_{ij} +$

$\gamma_{11}(\text{sex} * \text{time2}_{ij}) + \gamma_{12}(\text{sex} * \text{time3}_{ij})] + [U_{0j} + r_{ij}]$

This is still only a <u>random intercept</u> model

A two-level model – two fixed factors

It's easier to set up the model in SPSS than to write the equations!

These commands refer to the stacked dataset. We don't need the codes for subjects within groups.

```
mixed relp by chsexw1 time/
  fixed=intercept time chsexw1 time*chsexw1/
  random=intercept | subject(subid)/
  print=solution testcov.
```

A two-level model – two fixed factors

Explanation of syntax (1)

mixed relp by chsexw1 time/

DV

categorical variables follow *by*

Time could be treated as a numeric variable, in which case it would be entered after *with.*

A two-level model – two fixed factors

Explanation of syntax (2)

This subcommand specifies the fixed part of the model:

fixed=intercept time chsexw1 time*chsexw1/

A fixed intercept is added by default, but I like to specify it explicitly.

Main effects and interaction

A two-level model – two fixed factors

Explanation of syntax (3)

This subcommand specifies the random part of the model:

random=intercept | subject(subid)/

Specifies that there Is random variation around the intercept

Note use of vertical bar

Indicates that observations with the same *subid* are not independent; those with different values of *subid* are independent.

A two-level model – two fixed factors

Explanation of syntax (4)

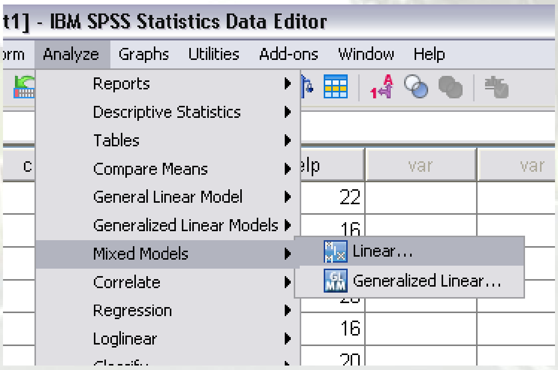
This subcommand specifies the output:

print=solution testcov.

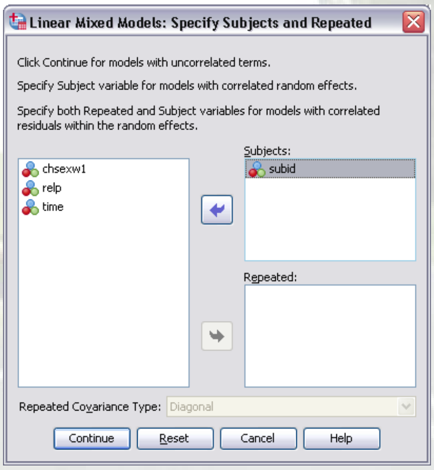
Asks for parameter estimates (as well as the default ANOVA table).

Asks for the variance estimates.

A two-level model – two fixed factors
Point-and-click (1)



A two-level model – two fixed factors
Point-and-click (2)

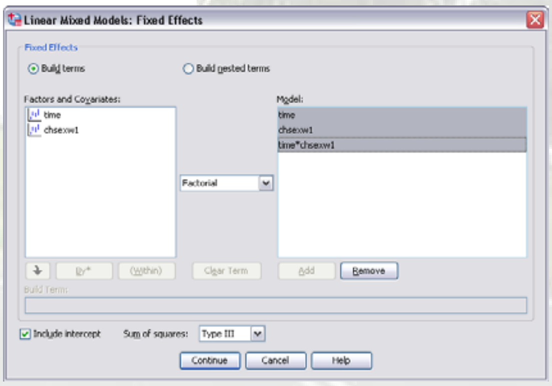


A two-level model – two fixed factors Point-and-click (3)

A two-level model – two fixed factors Point-and-click (4)

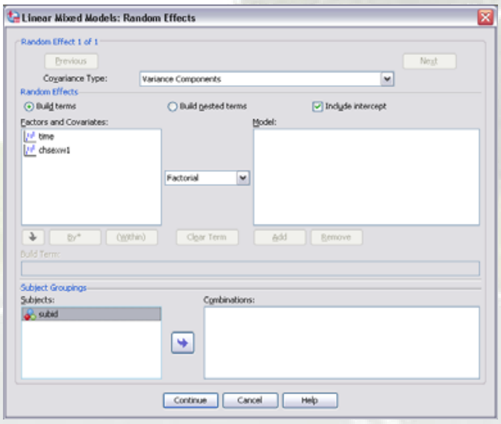
Click on *Fixed*, select both variables, and *Factorial,* and click on *Add.*

This specifies a fixed model with both main effects and their interaction.



A two-level model – two fixed factors
Point-and-click (5)

Click on *Continue*, then *Random*, and check *Include intercept.*



A two-level model – two fixed factors
Output (1)

Model Dimension[a]

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed Effects	Intercept	1		1	
	time	3		2	
	chsexw1	2		1	
	chsexw1 * time	6		2	
Random Effects	Intercept	1	Variance Components	1	subid
Residual				1	
Total		13		8	

a. Dependent Variable: relp.

This shows the number of parameters which were estimated

A two-level model – two fixed factors Output (2)

Information Criteria[a]

-2 Restricted Log Likelihood	304.630
Akaike's Information Criterion (AIC)	308.630
Hurvich and Tsai's Criterion (AICC)	308.840
Bozdogan's Criterion (CAIC)	314.818
Schwarz's Bayesian Criterion (BIC)	312.818

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: relp.

The likelihood referred to here is the probability of obtaining the current sample if the population had the parameter values given in the output.

This likelihood is maximised during the estimation of the parameter values.

The likelihood (probability) is typically small.

For example, the probability represented by -2 Log Likelihood is $7.09 * 10^{-67}$.

A two-level model – two fixed factors Output (3)

Information Criteria[a]

-2 Restricted Log Likelihood	304.630
Akaike's Information Criterion (AIC)	308.630
Hurvich and Tsai's Criterion (AICC)	308.840
Bozdogan's Criterion (CAIC)	314.818
Schwarz's Bayesian Criterion (BIC)	312.818

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: relp.

Under specific conditions, the fit of different models can be compared by referring the difference in the -2LL for the two models to a chi-squared distribution.

If ML is used (as opposed to the default, REML), models differing in fixed or random effects, or both, can be compared.

If REML is used, the models compared must have the same fixed effects but can differ in random effects.

Note: You can switch to maximum-likelihood (ML) from the default restricted or residual ML (REML) by including this sub-command: /method=ml

A two-level model – two fixed factors Output (4)

An example of comparing model fit

(1) This shows the fit for the first model:

Information Criteria[a]

-2 Restricted Log Likelihood	304.630

(2) We then fit a model <u>without</u> the random effect:

```
mixed relp by chsexw1 time/
  fixed=intercept time chsexw1 time*chsexw1/
  print=solution testcov.
```

(3) Fit for the second model:

Information Criteria[a]

-2 Restricted Log Likelihood	350.233

(4) The difference between the -2LL values is 350.233 – 304.63 = 45.6. We can refer this to a chi-squared distribution with one degree of freedom (the models differ only by the random intercept factor, which accounts for 1 df). The p-value is < .0005, which means that the model without the random factor has a worse fit.

Note that:

The information criteria are displayed in smaller-is-better forms.

A two-level model – two fixed factors Output (5)

The ANOVA table

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	20	301.837	.000
time	2	40.000	17.025	.000
chsexw1	1	20	2.779	.111
chsexw1 * time	2	40.000	9.046	.001

a. Dependent Variable: relp.

By default, Type III tests are performed; Type I (sequential) can be requested.

The coding underlying this table is the same as in GLM (deviation)

A two-level model – two fixed factors Output (6)

Parameter estimates

Estimates of Fixed Effects[b]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	13.200000	1.029727	26.249	12.819	.000	11.084342	15.315658
[time=1]	.866667	.648303	40.000	1.337	.189	-.443602	2.176935
[time=2]	-.400000	.648303	40.000	-.617	.541	-1.710268	.910268
[time=3]	0[a]	0	.	.	.	.	.
[chsexw1=1]	.228571	1.825512	26.249	.125	.901	-3.522090	3.979233
[chsexw1=2]	0[a]	0	.	.	.	.	.
[chsexw1=1] * [time=1]	4.847619	1.149318	40.000	4.218	.000	2.524761	7.170477
[chsexw1=1] * [time=2]	2.971429	1.149318	40.000	2.585	.013	.648571	5.294286
[chsexw1=1] * [time=3]	0[a]	0	.	.	.	.	.
[chsexw1=2] * [time=1]	0[a]	0	.	.	.	.	.
[chsexw1=2] * [time=2]	0[a]	0	.	.	.	.	.
[chsexw1=2] * [time=3]	0[a]	0	.	.	.	.	.

a. This parameter is set to zero because it is redundant.
b. Dependent Variable: relp.

The fixed factors are dummy- or indicator-coded and the highest-numbered category is the reference category (as in GLM).

A two-level model – two fixed factors Output (7)

Variances

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Residual		3.152222	.704858	4.472	.000	2.033679	4.885975
Intercept [subject = subid]	Variance	12.752857	4.371400	2.917	.004	6.513798	24.967825

a. Dependent Variable: relp.

- The residual is r_{ij} - the variance of observations within subjects
- The intercept variance is U_{0j} – the between-subject variance
- The *intraclass correlation*, rho (ρ) – the proportion of the total variance due to individuals – is U_{0j} / (U_{0j} + r_{ij}) = 12.75 / (12.75 + 3.15) = .80

A two-level model – two fixed factors
Further options

```
mixed relp by chsexw1 time/
  fixed=intercept time chsexw1 time*chsexw1/
  random=intercept | subject(subid)/
  print=solution testcov/
  emmeans=table(time*chsexw1) compare(time)/
  emmeans=table(time*chsexw1) compare(chsexw1)/
  test="test of linear and quadratic time by sex" chsexw1*time 1 0 -1 -1 0 1;
     chsexw1*time -1 2 -1 1 -2 1.
```

- The estimated marginal means subcommands test the simple effects of *time* and *sex*.
- The *test* command tests the significance of the linear *time* by *sex* and quadratic *time* by *sex* interaction contrasts.

The contrast for linear *time* is -1 0 1, and for quadratic *time* 1 -2 1. These are multiplied by the *sex* contrasts (-1 1) to obtain the interaction contrasts.

A two-level model – two fixed factors
Further options (2)

```
test="test of linear and quadratic time by sex" chsexw1*time 1 0 -1 -1 0 1;
    chsexw1*time -1 2 -1 1 -2 1.
```

L1 is the linear *time by sex* interaction contrast
L2 is the quadratic *time by sex* interaction contrast

Contrast Estimates[a,b]

Contrast	Estimate	Std. Error	df	Test Value	t	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
L1	4.847619	1.149318	40.000	0	4.218	.000	2.524761	7.170477
L2	1.095238	1.990677	40.000	0	.550	.585	-2.928070	5.118546

a. test of linear and quadratic time by sex
b. Dependent Variable: relp.

A two-level model – two fixed factors

Further options (3)

Correlations between residuals

The commands specified so far assume *compound symmetry*, which is stronger than *circularity*, in that it assumes that the variances of the DVs are equal (see the footnote on p. 49 of *GLM*).

We can specify different correlations between the residual errors by using the *repeated* subcommand rather than the *random* subcommand:

Subcommand	-2LL*	*p* for sex*time
repeated=time | subject(subid) covtype(cs)/	313.44	.00027
repeated=time | subject(subid) covtype(hf)/	310.08	.00027
repeated=time | subject(subid) covtype(ar1)/	325.88	.012

The fit is slightly better if circularity rather than CS is assumed, but the *p*-value is unchanged. The assumption of autocorrelation Is not supported. Note the change in the *p*-value.

* ML rather than REML was used to allowed comparison of models with different random structures.

A two-level model – two fixed factors

A linear growth model

- Treat the effect of *time* as linear rather than categorical
- Allow individual random variation around the change over time as well as around the intercept.

```
mixed relp by chsexw1 with time/
  fixed=intercept time chsexw1 time*chsexw1/
  random=intercept time | subject(subid)/
  print=solution testcov.
```

Time is specified after *with* and is included in the *random* statement.

This analysis would give rise to individual growth curves.

A two-level model – two fixed factors

A linear growth model (2)

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	38.739	309.038	.000
time	1	42	32.536	.000
chsexw1	1	38.739	13.975	.001
chsexw1 * time	1	42.000	17.654	.000

a. Dependent Variable: relp.

What is being tested here?

A significant *time* by *sex* interaction as before

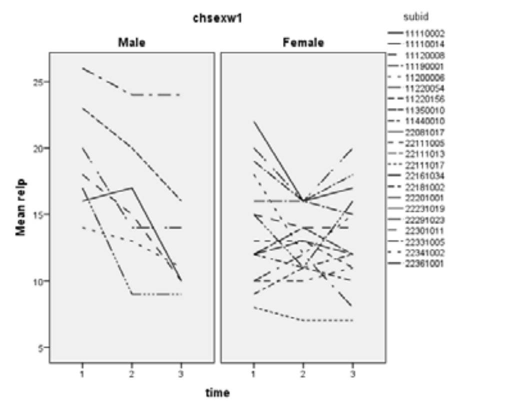
Estimates of Covariance Parameters[b]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Residual		3.176531	.693176	4.583	.000	2.071122	4.871923
Intercept [subject = subid]	Variance	12.744754	4.371192	2.916	.004	6.507087	24.961824
time [subject = subid]	Variance	.000000[a]	.000000	.	.	.	.

a. This covariance parameter is redundant. The test statistic and confidence interval cannot be computed.
b. Dependent Variable: relp.

However, there is no significant individual variation around the change over time that's not accounted for by the interaction.

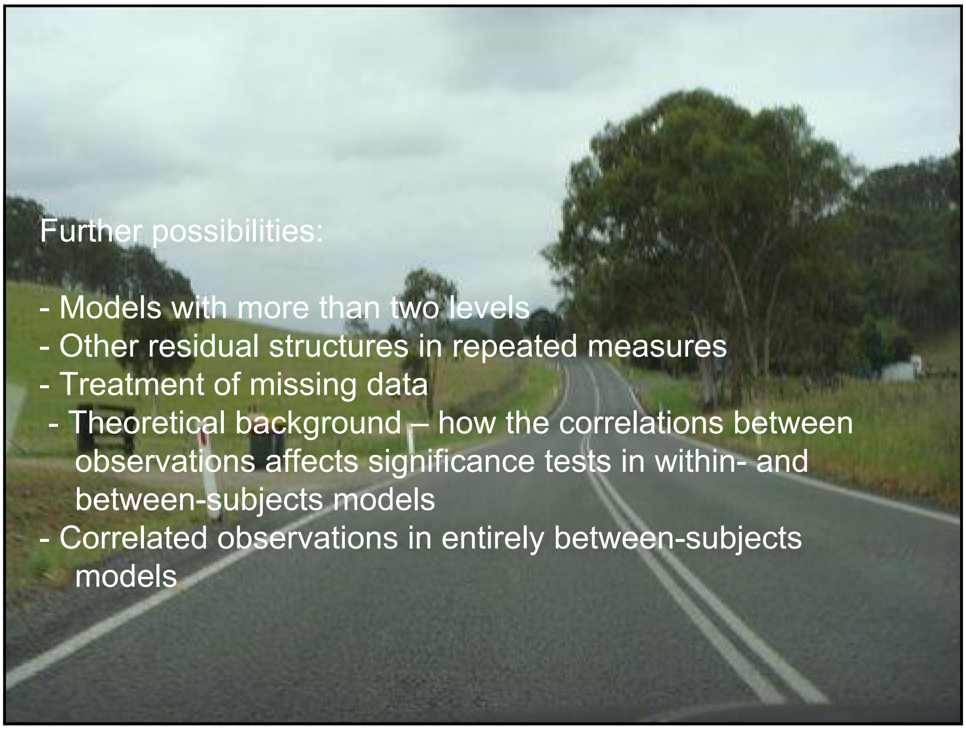
A two-level model – two fixed factors
A linear growth model (3)



Change over time for

Males : -2.86 (p<.0005)
Females: -.43 (p=.19)

Plots over time for male and female individual subjects. Note large differences in intercepts.


Further possibilities:
- Models with more than two levels
- Other residual structures in repeated measures
- Treatment of missing data
- Theoretical background – how the correlations between observations affects significance tests in within- and between-subjects models
- Correlated observations in entirely between-subjects models