Exercises linear mixed modelling: mixed design

Stéphanie M. van den Berg

February 28, 2020

There are two sections with exercises.

0.1 Exercises

Below we see data from a study on the effects of the financial crisis on the number of employees in specific Dutch companies. The companies are distinguised into food and non-food related companies. The number of employees are recorded in January 2008 and January 2011.

company	food	2008	2011
1	nonfood	42	63
2	food	104	126
3	nonfood	76	58
4	food	65	131

1. These data are in wide format. Rewrite the datamatrix in such a way that we have the same data in long format. Provide column (variable) names.

...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

2. Do we need to use a linear mixed model, or can we analyse these data with an ordinary linear model?

3. We want to test the null-hypothesis that the effects of the financial crisis in 2008 has the same effect on the number of employees in the food sector as in the non-food sector. Provide the syntax that helps you test this hypothesis.

4. Suppose the output in Figure 4 results from an analysis done by a colleague:

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	998.000	70192.133	.000
food	1	998.000	3389.819	.000
year	1	998	1122.117	.000
food * year	1	998	.437	.509

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	81.574000	.664366	1989.879	122.785	.000	80.271074	82.876926
[food=1.00]	39.312000	.939556	1989.879	41.841	.000	37.469384	41.154616
[food=2.00]	0[b]	0	.	.	.	.	.
[year=2008.00]	-22.056000	.913130	998	-24.154	.000	-23.847874	-20.264126
[year=2011.00]	0[b]	0	.	.	.	.	.
[food=1.00] * [year=2008.00]	.854000	1.291360	998	.661	.509	-1.680093	3.388093
[food=1.00] * [year=2011.00]	0[b]	0	.	.	.	.	.
[food=2.00] * [year=2008.00]	0[b]	0	.	.	.	.	.
[food=2.00] * [year=2011.00]	0[b]	0	.	.	.	.	.

b. This parameter is set to zero because it is redundant.

Estimates of Covariance Parameters

Parameter		Estimate	Std. Error
Residual		208.451418	9.331567
Intercept [subject = company]	Variance	12.239802	6.996594

Figure 1: Output of a MIXED analysis done by a colleague.

She provides you with the information that food=1 means the food sector and food=2 is the nonfood sector.

What does the model predict regarding the number of employees in 2008 in the non-food sector?

5. What does the model predict regarding the number of employees in 2011 in the non-food sector?

6. What does the model predict regarding the number of employees in 2008 in the food sector?

7. What does the model predict regarding the number of employees in 2011 in the food sector?

8. How large is the effect of the crisis in the food sector?

9. How large is the effect of the crisis in the non-food sector

10. How large is the intraclass correlation (ICC)? Give the computation.

11. Could we have done the analysis with an ordinary linear model? Explain your answer.

12. Can we reject the null-hypothesis that the effects of the crisis were the same in the food and non-food sectors? Explain your answer.

0.2 Answers

1. It could look like this:

company	sector	year	NEmployees	...
1	nonfood	2008	42	...
1	nonfood	2011	63	...
2	food	2008	104	...
2	food	2011	126	...
3	nonfood	2008	76	...
3	nonfood	2011	58	...
4	food	2008	65	...
4	food	2011	131	...

2. The data are clustered into companies: for each company we have two data points, so we should at least try a linear mixed model. Only if the variance of the company random effects is extremely small, we could use a linear model without random effects.

3. One option is to let SPSS construct the dummy variables:

```
MIXED employees BY year sector
  /FIXED=year sector year*sector
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(company) COVTYPE(VC).
```

   Or you do the dummy coding yourself, for example like this:

```
RECODE year (2008=0) (2011=1) INTO year2011.
RECODE sector ('Nonfood'=0) ('food'=1) INTO food.
EXECUTE.

COMPUTE food2011=year2011*food.
EXECUTE.

MIXED employees WITH year2011 food food2011
  /FIXED= year2011 food food2011
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(company) COVTYPE(VC).
```

4. the nonfood sector is food=2, so the predicted number of employees in 2008 in the nonfood sector is equal to $81.57 + 0 - 22.056 + 0 = 59.514$

5. the nonfood sector is food=2, so the predicted number of employees in 2011 in the nonfood sector is equal to $81.57 + 0 + 0 + 0 = 81.57$

6. the food sector is food=1, so the predicted number of employees in 2008 in the food sector is equal to $81.57 + 39.31 - 22.056 + 0.85 = 99.674$

7. the food sector is food=1, so the predicted number of employees in 2011 in the food sector is equal to $81.57 + 39.31 + 0 + 0 = 120.88$

8. in the food sector the effect is a $120.88 - 99.674 = 21.206$ increase in number of employees

9. in the non-food sector the effect is a $81.57 - 59.514 = 22.056$ increase in number of employees

10. the ICC is $\frac{12}{12+208} = 0.05$

11. we have clustering, with multiple data point per company, so in general a linear mixed model is better than an ordinary linear model. However, since the intraclass correlation is rather low, the results would be very similar if we would use an ordinary linear model.

12. The null-hypothesis cannot be reject as the year by sector interaction effect is not signifcantly different from 0, $t(998) = 0.66, p = 0.51$. (alternatively, $F(1, 998) = 0.44, p = 0.51$). Note however that the statistical results are in terms of absolute number of employees. These data show that the average number of employees in 2008 is larger in the food sector than in the non-food sector. Perhaps it would be wiser to look at percentage increase in number of employees: A change from 100 to 102 reflects a larger impact than a change from 1000 to 1002.

0.3 Exercises

1. A psychologist studies whether age affects math performance. In 2017, she measures math performance (one score) in a group of 80-year-olds and she measures math performance (one score) in a group of 90-year-olds.
   1. In this design, is the age variable a between-participants variable or a within-participant variable?
   2. Would you analyze these data with a linear model, or with a linear mixed model? Explain.

2. A psychologist studies whether age affects math performance. She measures math performance (one score) in a group of 7-year-olds and she measures math performance again when the same children are 8 years old.
   1. In this design, is the age variable a between-participants variable or a within-participant variable?
   2. Would you analyze these data with a linear model, or with a linear mixed model? Explain.

3. Look at the data table below.

ID	Nationality	Sex	Mathscore
1	Dutch	Male	67
2	Dutch	Female	88
3	German	Male	50
4	German	Female	98
...	...	...	...

   In this data set on Math performance, we see two variables, nationality and sex. 1. What kind of variables are these: within-participant variables or between-participants variables? Explain.
   2. Would you call this a mixed design? Explain.
   3. Would you analyze this data set with a linear model or with a linear mixed model? Explain.

4. Look at the data table below.

ID	Nationality	Age	Mathscore
1	Dutch	3	67
1	Dutch	5	88
2	German	4	50
2	German	6	98
...	...	...	...

1. In this data set on Math performance, we see two variables, nationality and age. What kind of variables are these: within-participant variables or between-participants variables? Explain.
2. Would you call this a mixed design? Explain.
3. Would you analyze this data set with a linear model or with a linear mixed model? Explain.

5. Look at the data table below.

ID	Subject	Sex	Mood
1	Psychology	Male	67
1	Psychology	Female	88
2	Sociology	Female	50
2	Sociology	Male	98
...	...	...	...

1. In this data set on mood in transsexuals, we see two variables, the subject they have a Master's degree in, and sex. What kind of variables are these: within-participant variables or between-participants variables? Explain.
2. Would you call this a mixed design? Explain.
3. Would you analyze this data set with a linear model or with a linear mixed model? Explain.

6. Look at the data table below.

SchoolID	Country	Year	Avarage Mathscore
1	The Netherlands	2010	67
1	The Netherlands	2011	88
1	The Netherlands	2012	50
1	The Netherlands	2013	98
2	Germany	2010	67
2	Germany	2011	88
2	Germany	2012	50
2	Germany	2013	98
...	...	...	...

1.In this data set on average Math performance in schools, we see two variables, country of the school and year of data collection. What kind of variables are these: within-school variables or between-schools variables? Explain.
2. Would you call this a mixed design? Explain.
3. Would you analyze this data set with a linear model or with a linear mixed model? Explain.

1 Answers

1. 1. The age variable is a between-participants variable: some of the participants are 80 years old and some are 90 years old: none are both at the same time. Age discriminates between two sets of participants, so it is a between-participants variable.
   2. Two groups of participants were studied. Because we only have one measure for each participant, there is no clustering, and we use an ordinary linear model.

2. 1. The age variable is a within-participants variable: children are studied twice and scores can therefore be compared within an individual.
   2. One group of participants was studied and for each participant we have two math scores. Because we have more than one measure for each participant, we have to use a linear mixed model to account for clustering.

3. 1.Each participant is either Dutch or German. This is a between-participants variable. Each participant is either male or female, sex discriminates between separate groups of participants, so sex is a between-participants variable.
   2. This is *not* a mixed design as it does not have both within-participant and between-participants independent variables.
   3. Because we only have one measure for each participant, there is no clustering, and we use an ordinary linear model.

4. 1. Each participant is either Dutch or German. This is a between-participants variable. On measurement 1 the same participants have a different age than on measurement 2. Age is therefore a within-participant variable.
   2. This is a mixed design as it has both a within-participant and a between-participants independent variable.
   3. For each participant we have two math scores, so we would have to use a linear mixed model to account for clustering.

5. 1. Each participant has only one Masters degree. This is a between-participants variable. Between the two measurements, participants change their sex. This is a within-participant variable: we can compare people's mood when they are male and when they are female.
   2. This is a mixed design as it has both a within-participant and a between-participants independent variable.
   3. For each participant we have two mood scores, so we would have to use a linear mixed model to account for clustering.

6. 1. Each school is based in only one country and has measurements across four years. Country is a between-schools variable and year is a within-school variable.
   2. This is a mixed design as it has both a within-school and a between-schools independent variable.

3. For each school we have four average math scores, so we would have to use a linear mixed model to account for clustering.