Linear Mixed Models

and their applications to multilevel and repeated measures designs

Marcello Gallucci
Univerisity of Milano-Bicocca

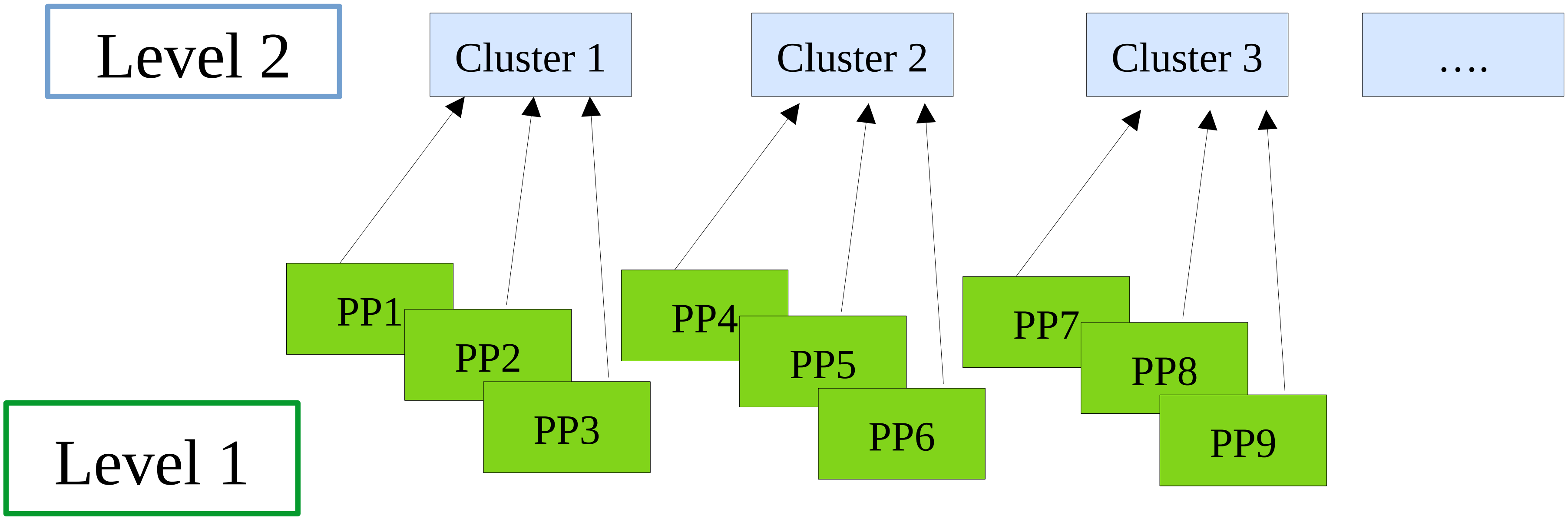
Multilevel Designs

Multilevel designs are research designs in which the **sampling** of cases is done in different, hierarchical steps

Level 2	A sample of clusters	*A sample of countries*
	For each cluster	*For each country*
Level 1	A sample of cases	A sample of participants

Multilevel Designs

Multilevel designs are research designs in which the **sampling** of cases is done in different, hierarchical steps



Multilevel Designs

Multilevel designs are research designs in which the **sampling** of cases is done in different, hierarchical steps

Level 3	A sample of clusters	*A sample of countries*
	Within each cluster	*Within each country*
Level 2	A sample of cases	A sample of participants
	Within each case	*Within each participant*
Level 1	A sample of measures	Repeated-measures over time

Multilevel designs in psychology

- Samples of individuals within countries (or cities, or regions)
- Samples of pupils within classes within schools
- Samples of participants within experimental groups
- Samples of individuals within families (or couples)
- Samples of individuals within communities
- Samples of measures within participants (repeated-measures designs)

Multilevel Designs

Multilevel designs are research designs in which the **sampling** of cases is done in different, hierarchical steps

Level 1

A sample of cases

A sample of participants

Effects due to difference among cases

Differences across people

Level 2

A sample of clusters

A sample of countries

Influence on variables relationships

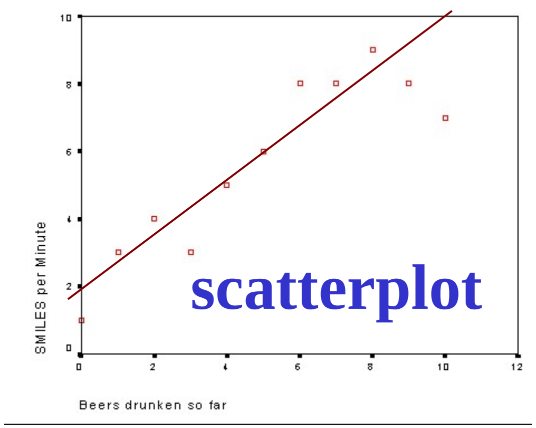
Different relationships across countries

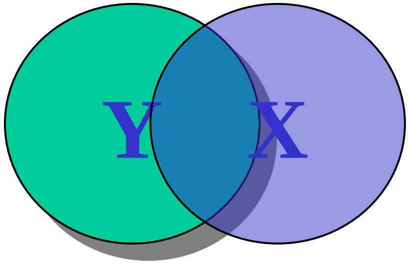
Mixed models

Design	Statistical Model
Multilevel	**The mixed model** Aka: *random coefficients linear model* Aka: *hierarchical linear model* Aka: *multilevel model*

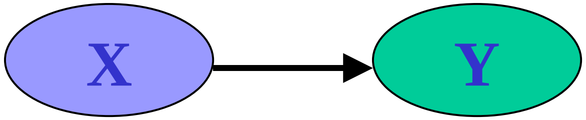
A statistical model

A simple **statistical model** is an **efficient** and **concise** representation of the data describing an empirical phenomenon

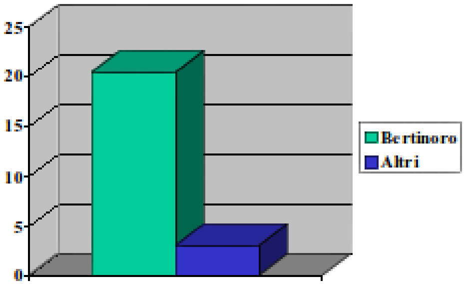




Path Diagram



Difference in mean



Software

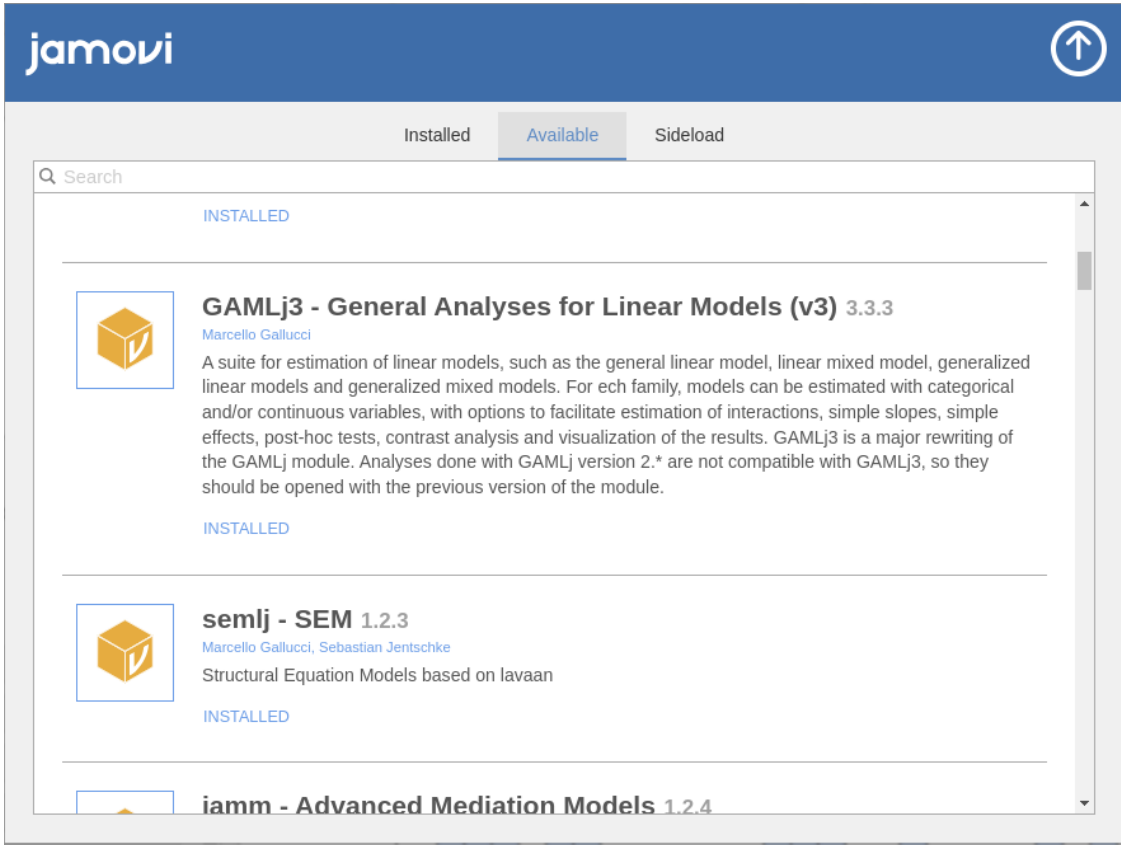
jamovi Stats. Open. Now.

SPSS

R

jamovi

- In jamovi mixed models can be estimated with the GAMLj3 module

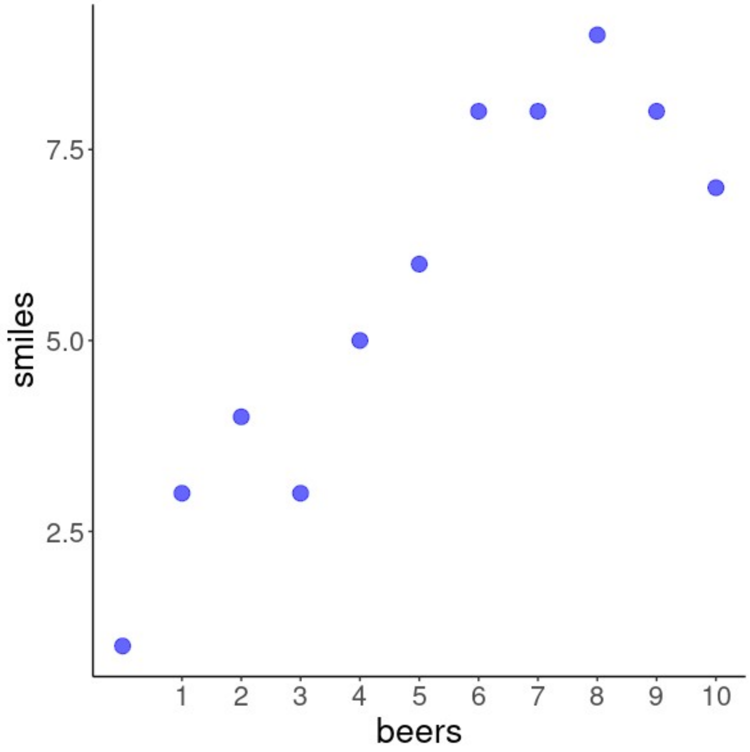


Docs and examples: https://gamlj.github.io/

Example "beers & smiles"

Consider this hypothetical research, where we went to a bar and measured (in a given time) the number of beers drunk and number of smiles smiled by participants

	beers	smiles
1	0	1
2	1	3
3	2	4
4	3	3
5	4	5
6	5	6
7	6	8
8	7	8
9	8	9
10	9	8
11	10	7
12		

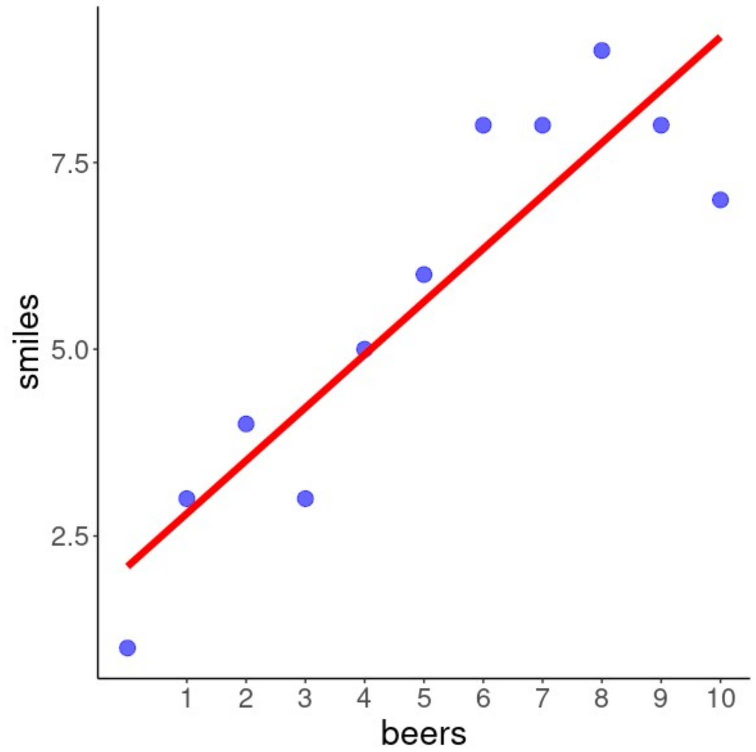


Example “beers & smiles”

We represent the relationship between the two variable with a **efficient** and **concise** set of coefficients: **a straight line**

The regression line

$$\hat{y}_i = a + b \cdot x_i$$

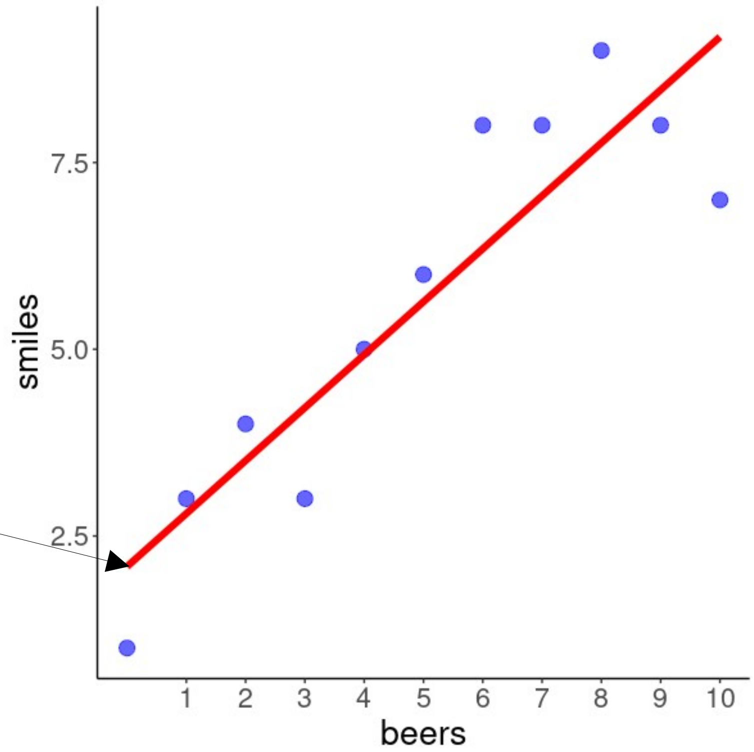


Intercept

We represent the relationship between the two variable with a concise and efficient set of coefficients: **a straight line**

The regression line

$$\hat{y}_i = a + b \cdot x_i$$



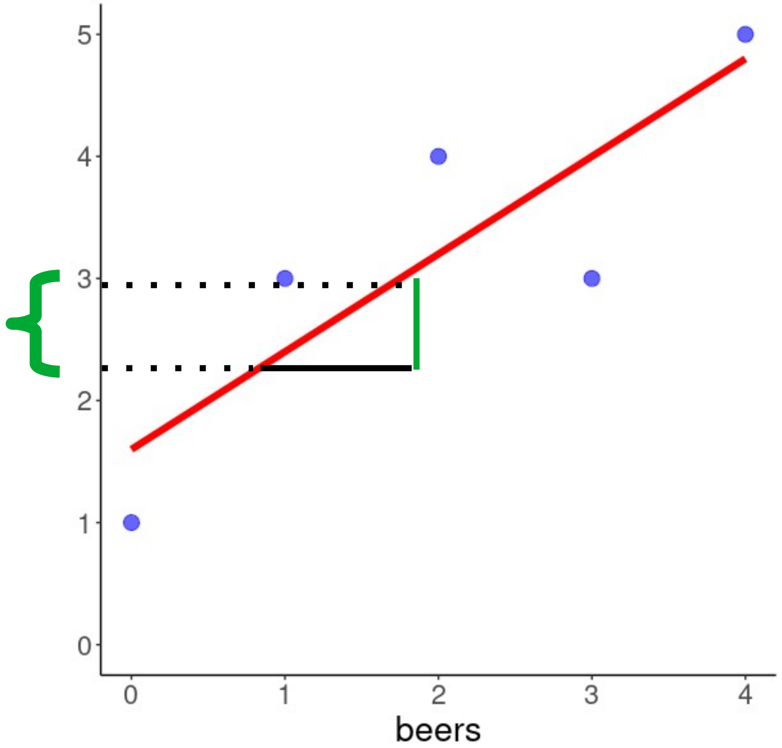
a: Intercept = the expected value of Y for X=0

Slope

We represent the relationship between the two variable with a concise and efficient set of coefficients: **a straight line**

The regression line

$$\hat{y}_i = a + b \cdot x_i$$



a: slope = the expected change in Y for one unit more in X

Results

If we run the analysis (with any software, here **jamovi**) we obtain the parameters

The regression line

$$\hat{y}_i = a + b \cdot x_i$$

Parameter Estimates (Coefficients)

Names	Estimate	SE	95% Confidence Intervals Lower	Upper	β	df	t	p
(Intercept)	2.091	0.684	0.543	3.638	−0.000	9	3.057	0.014
beers	0.709	0.116	0.448	0.971	0.898	9	6.132	< .001

Goodness of fit

Proportion of variance explained

Model Fit

R^2	Adj. R^2	df	df (res)	F	p
0.807	0.785	1	9	37.6	< .001

[4]

GLM

- The regression line is a an application of the **general linear model**

General Linear Model

$$y_i = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \,..\, b_k \cdot x_{ki} + e_i$$

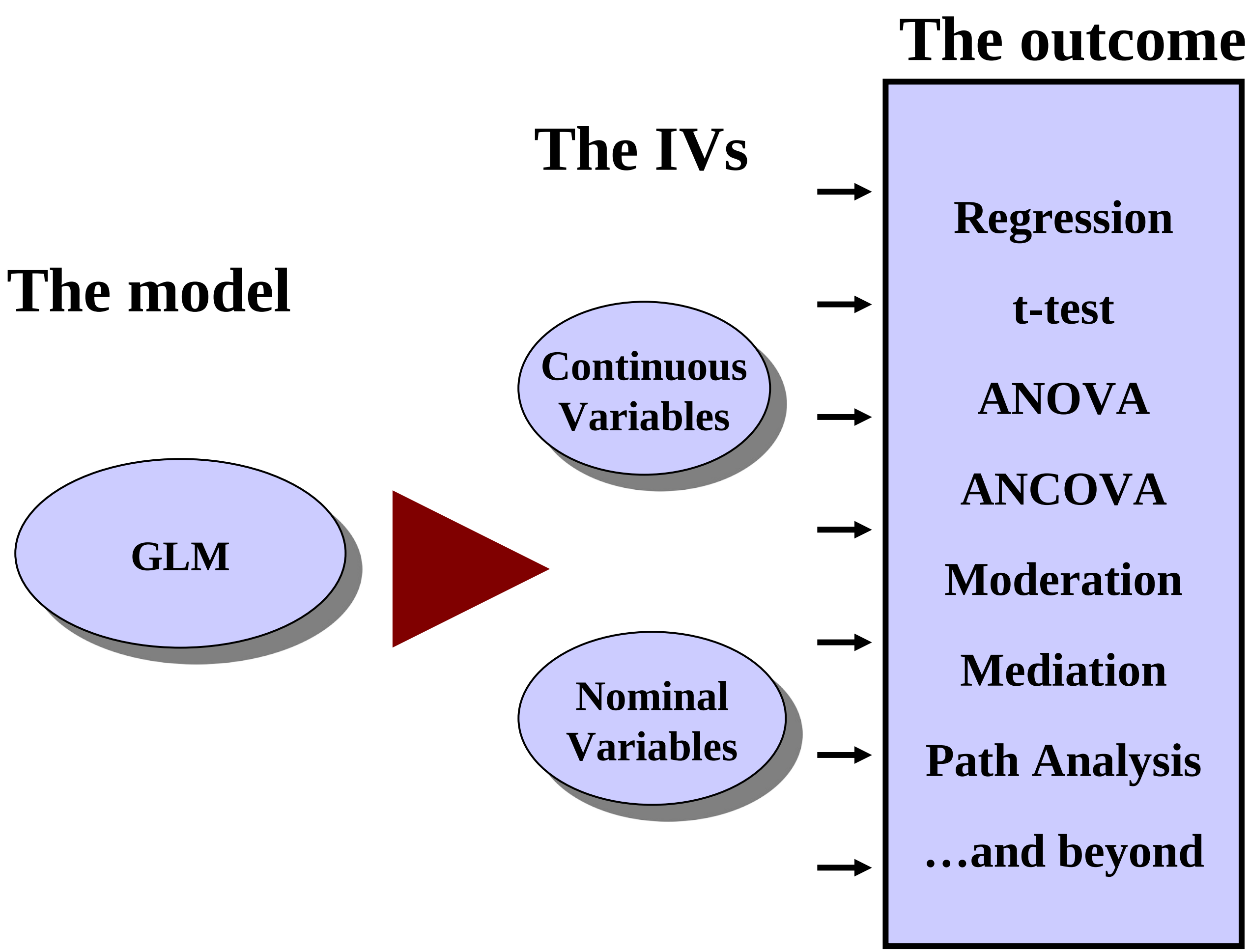
Dependent variable

Independent variables

Errors

GLM

The GLM is the model we almost always use in classical analyses



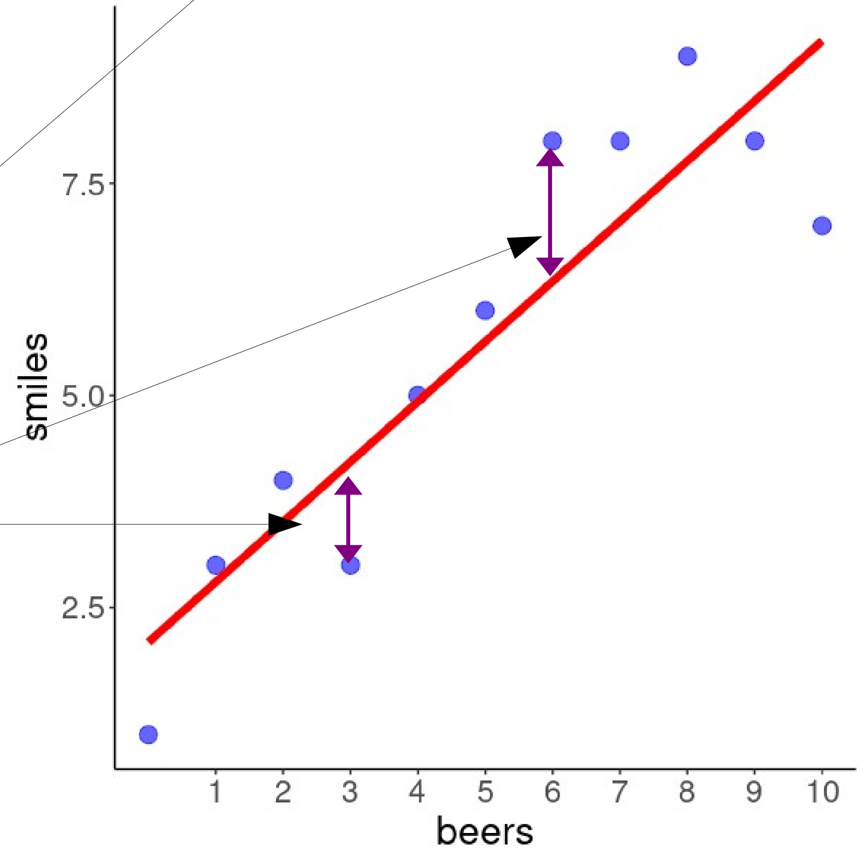
Some GLM Assumptions

$$\hat{y}_i = a + b \cdot x_i$$

1) There exists one and only one value of each parameter (e.g. the slope) in the population: **Fixed effects**

2) Any observed deviation from the predicted values is deemed to be error (residuals)

3) The random deviations from the model are independent of each other



Independence of cases

The random deviations from the model are independent of each other

- To abide by this assumption, the sample of cases (participants) should be a random sample of independent cases
- No dependency of cases, no connection
- No clustering

Violations are likely when

- Samples of individuals within countries (or cities, or regions)
- Samples of pupils within classes within schools
- Samples of participants within experimental groups
- Samples of individuals within families (or couples or communities)
- Samples of measures within participants (repeated-measures designs)

GLM

When the assumptions are NOT met because of multilevel data,

we generalize the GLM to the **Linear Mixed Model**

Linear Mixed Model

GLM

Regression

T-test

ANOVA

ANCOVA

Moderation

Mediation

Path Analysis

LMM

Random coefficients models

Random intercept regression models

One-way ANOVA with random effects

One-way ANCOVA with random effects

Intercepts-and-slopes-as-outcomes models

Mixed Linear Models

- With the mixed model one can take into the account dependency among cases (within clusters) almost in any situation
- It allows applying the GLM logic to a broader range of designs
- Any kind of independent variables
- Efficient handling of missing values
- Multi-level research designs
- Repeated measures designs
- Generalizes to the generalized linear model (logistic etc)

Example “beers”

Let's consider the case where the beer-smile research was conducted by gathering data in **several different bars**

For each participant we measured # of beers and # of smiles

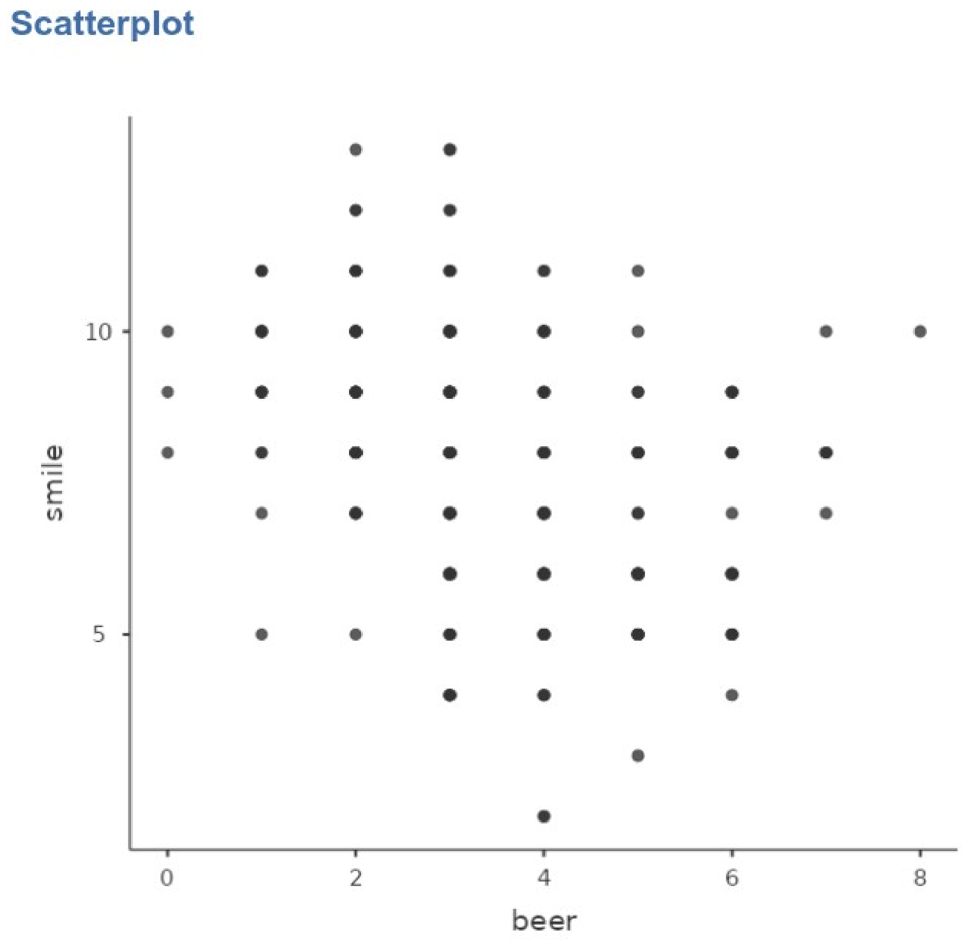
bar

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	a	3	1.3	1.3	1.3
	b	14	6.0	6.0	7.3
	c	22	9.4	9.4	16.7
	d	21	9.0	9.0	25.6
	e	14	6.0	6.0	31.6
	f	20	8.5	8.5	40.2
	g	24	10.3	10.3	50.4
	h	12	5.1	5.1	55.6
	i	16	6.8	6.8	62.4
	l	22	9.4	9.4	71.8
	m	21	9.0	9.0	80.8
	n	15	6.4	6.4	87.2
	o	16	6.8	6.8	94.0
	p	11	4.7	4.7	98.7
	q	3	1.3	1.3	100.0
	Total	234	100.0	100.0	

For a total of 234 participants

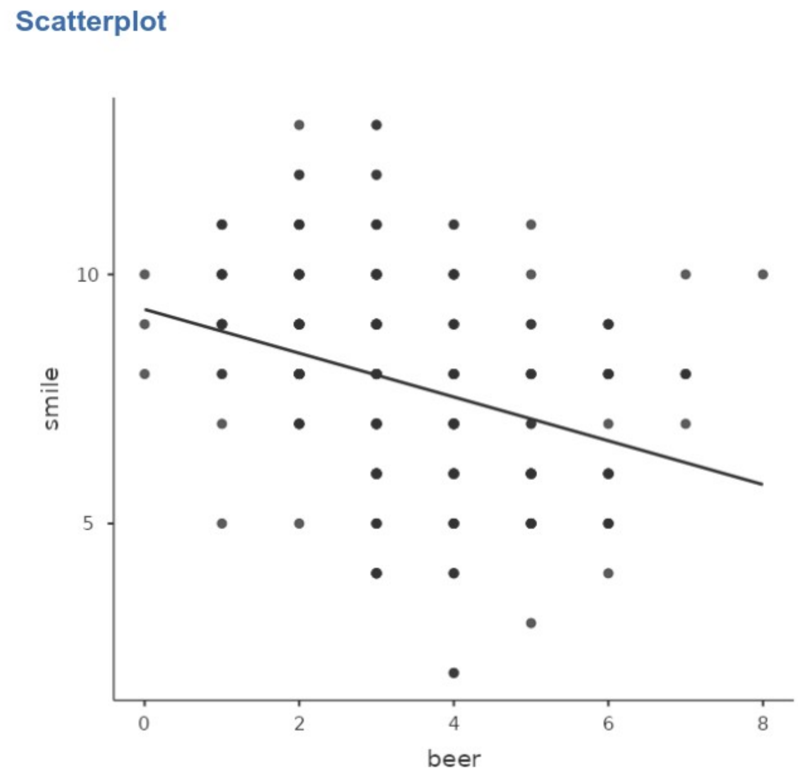
Example “beers” 2

As compared with the example with a few participants, now we have a very different scatterplot



Example “beers & smiles” 2

A simple regression confirms that results are indeed different



Negative effect

Fixed Effects Parameter Estimates

Names	Estimate	SE	95% Confidence Interval Lower	95% Confidence Interval Upper	β	df	t	p
(Intercept)	7.765	0.130	7.508	8.022	0.000	232	59.503	< .001
beer	−0.440	0.085	−0.608	−0.271	−0.320	232	−5.147	< .001

Why

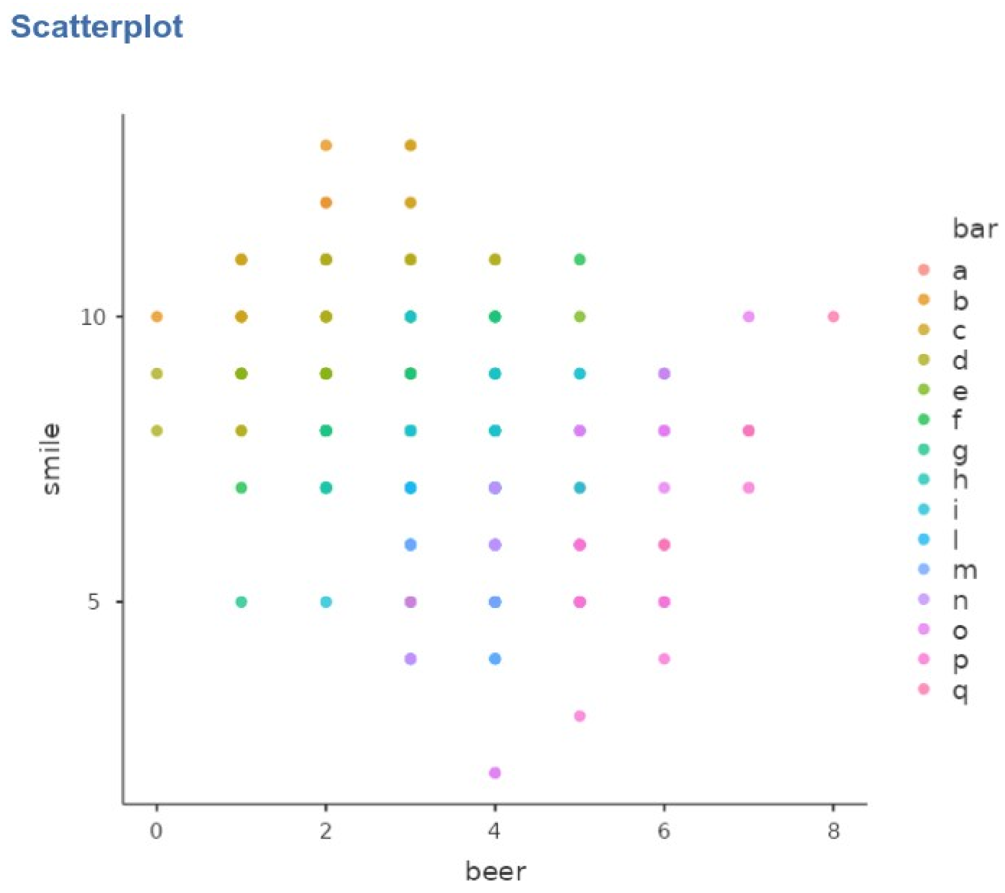
Results may be biased by a mis-specification of the model, where the structure of the data is not taken into account

- In fact:
  - Subjects are sampled in clusters specified by **bars**
  - Each bar may have specific characteristics (quality, entertainment, etc) that may affect the measured variables
  - Subjects within the same bar may be more similar than across bars

Scatterplot by Bar

Let's see the data broken down by bar

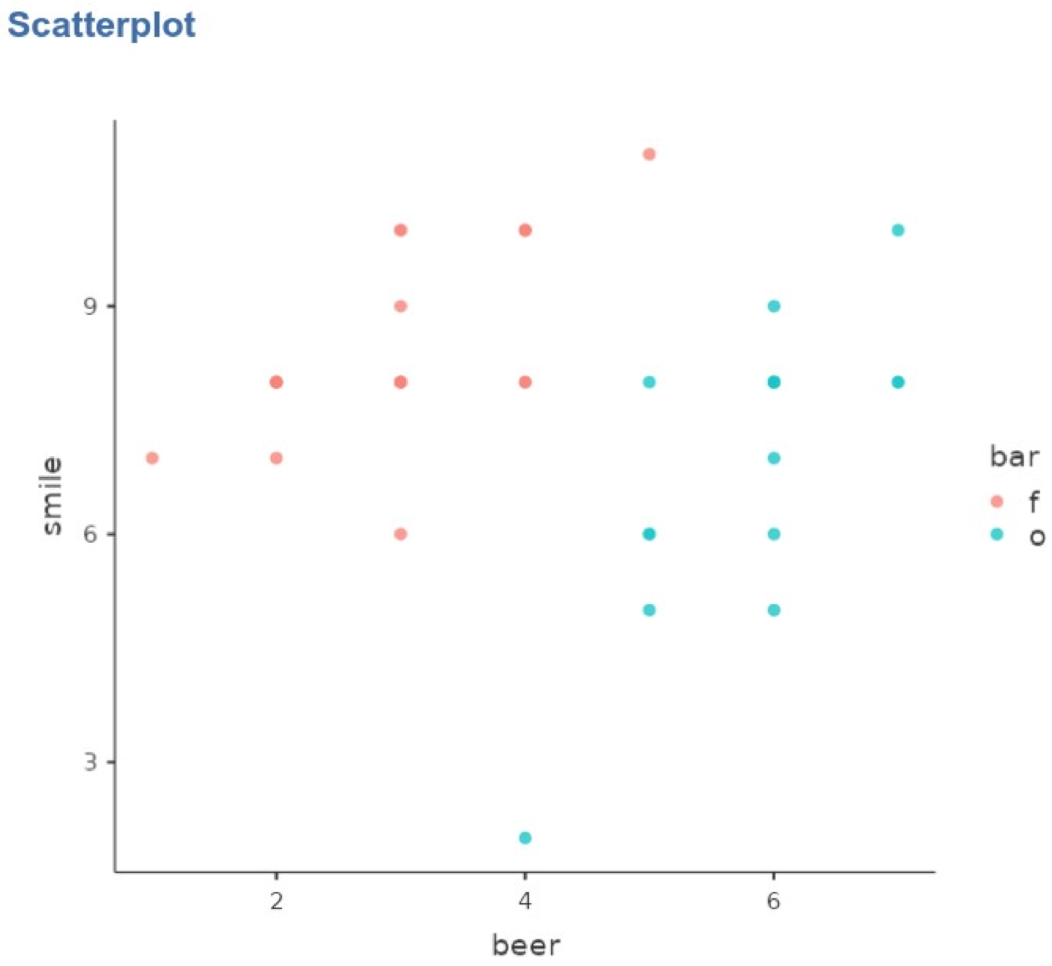
Bar



Scatterplot by Bar

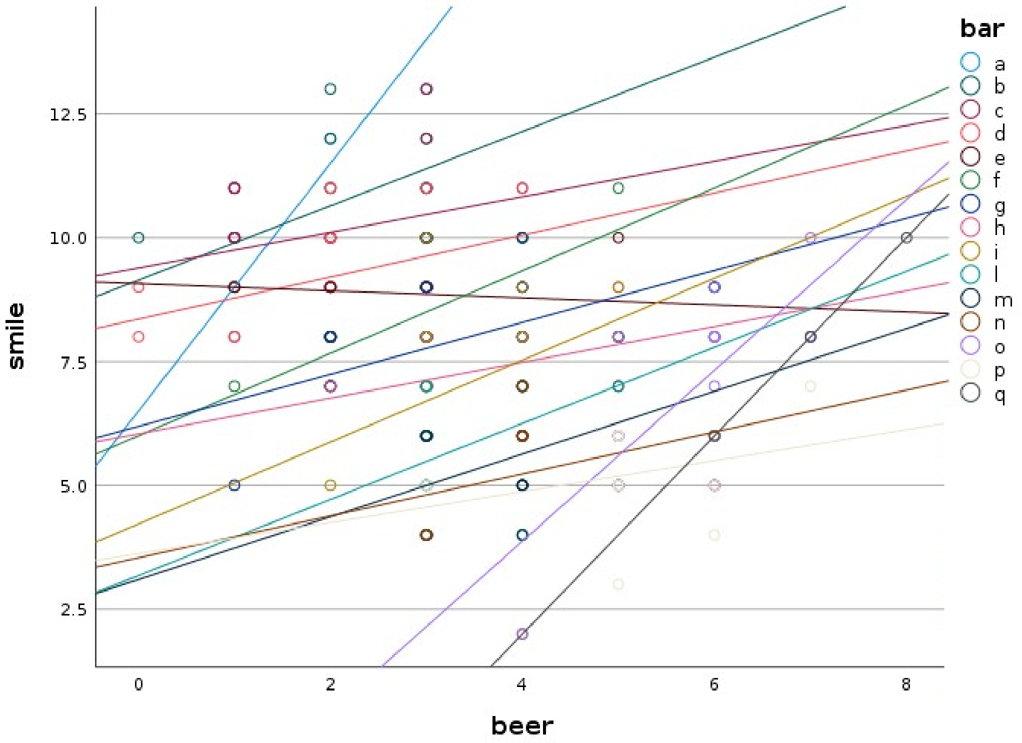
Let's see the data only for bar “f” and “o”

Bar



Scatterplot by Bar

It seems that the relations between IV and DV is positive, but within each bar

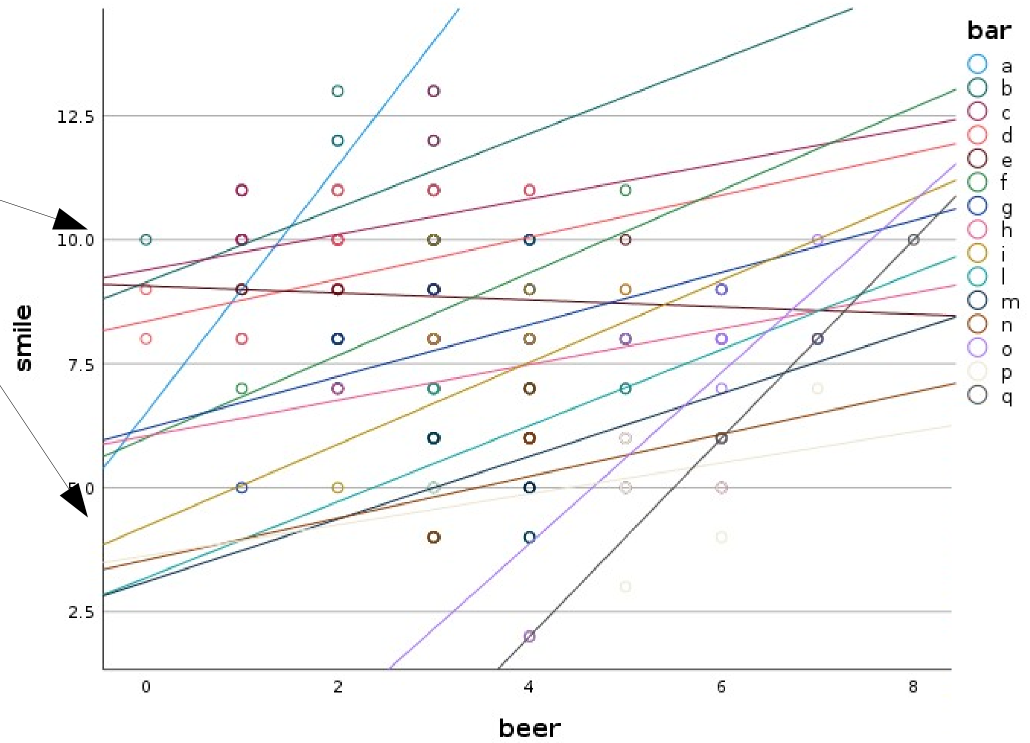


Bar

Scatterplot by Bar

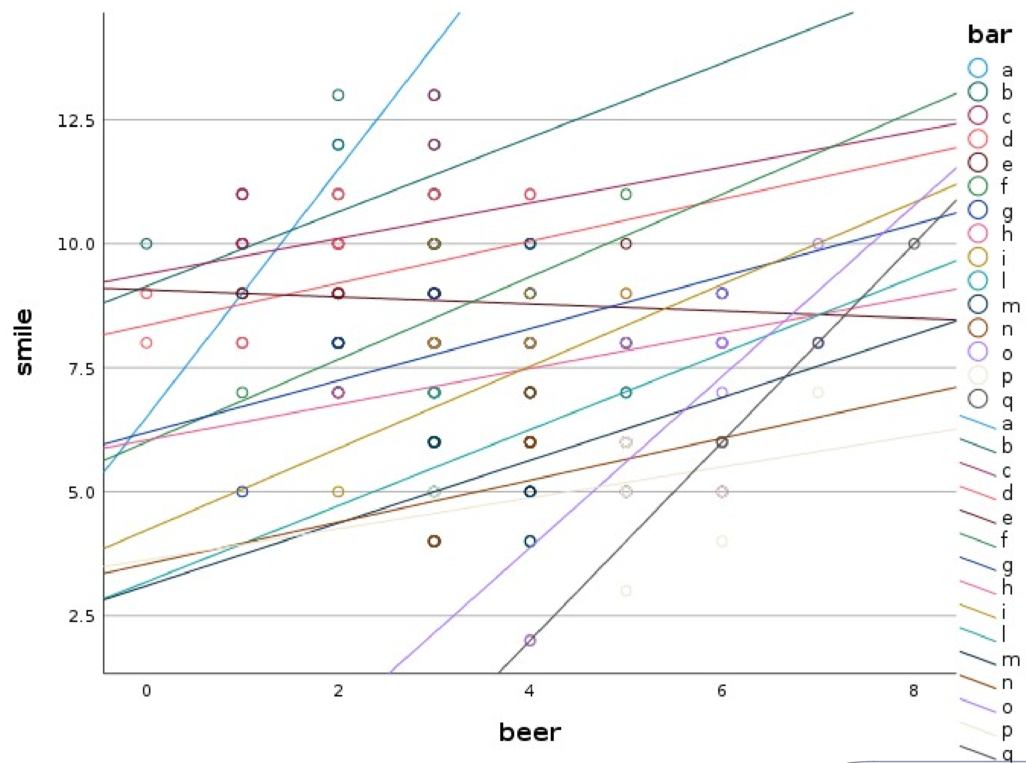
Bar

Intercepts seem to be different from bar to bar



Scatterplot by Bar

Bar



Slopes are all positive

Slopes seem to vary across bars

The Model

- It seems that considering the participants as all equivalent and independent of each other (GLM assumption) does not fit our data

- It seems that a better model should allow each bar (each cluster) to have a different regression line (a different intercept and **b** coefficient)

The Model

- Let's define a model with a regression line for each cluster

y_{ij} Smiles of subject i in the cluster j

$$\hat{y}_{ia} = a_a + b_a \cdot x_{ia}$$
$$\hat{y}_{ib} = a_b + b_b \cdot x_{ib}$$
$$\hat{y}_{ic} = a_c + b_c \cdot x_{ic}$$

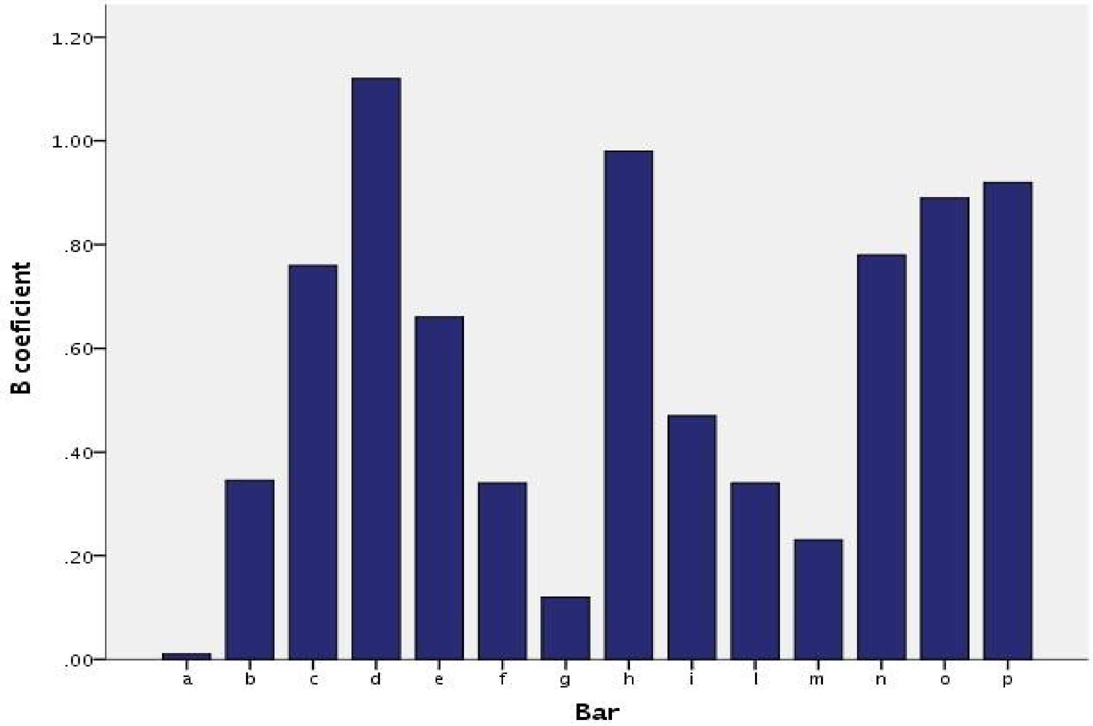
$$\hat{y}_{ij} = a_j + b_j \cdot x_{ij}$$

In these regressions the coefficients may vary from cluster to cluster: **they are not Fixed**

Varying coefficients

- If coefficients may vary, they will have a distribution

A possible distribution of coefficients b estimated for different clusters

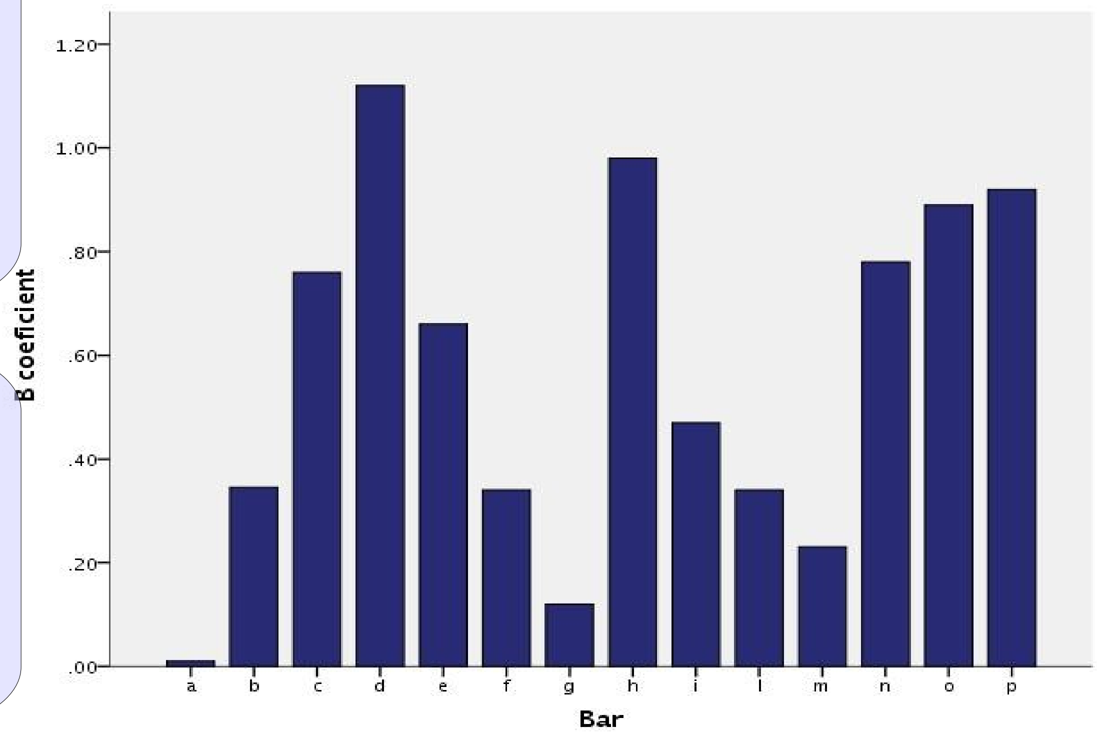


Random coefficients

- Varying coefficients are called random coefficients

Coefficients will exhibit variability: Coefficients are **random**

That is: in the **population** there exist different coefficients, a sample of which we estimated using the clustered data

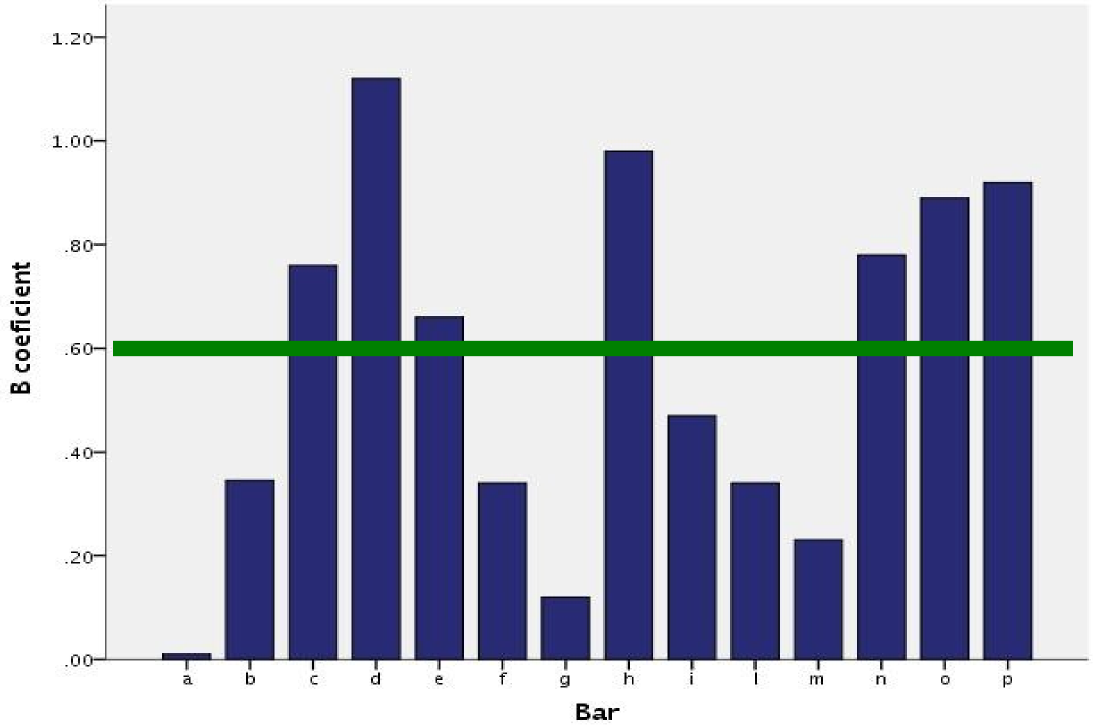


Average of the coefficients

- If coefficients vary as a variable in the population, they will have a mean and a variance, that we can estimate in our data

Mean of coefficients

$$\bar{b} = \frac{\sum_j b_j}{k}$$

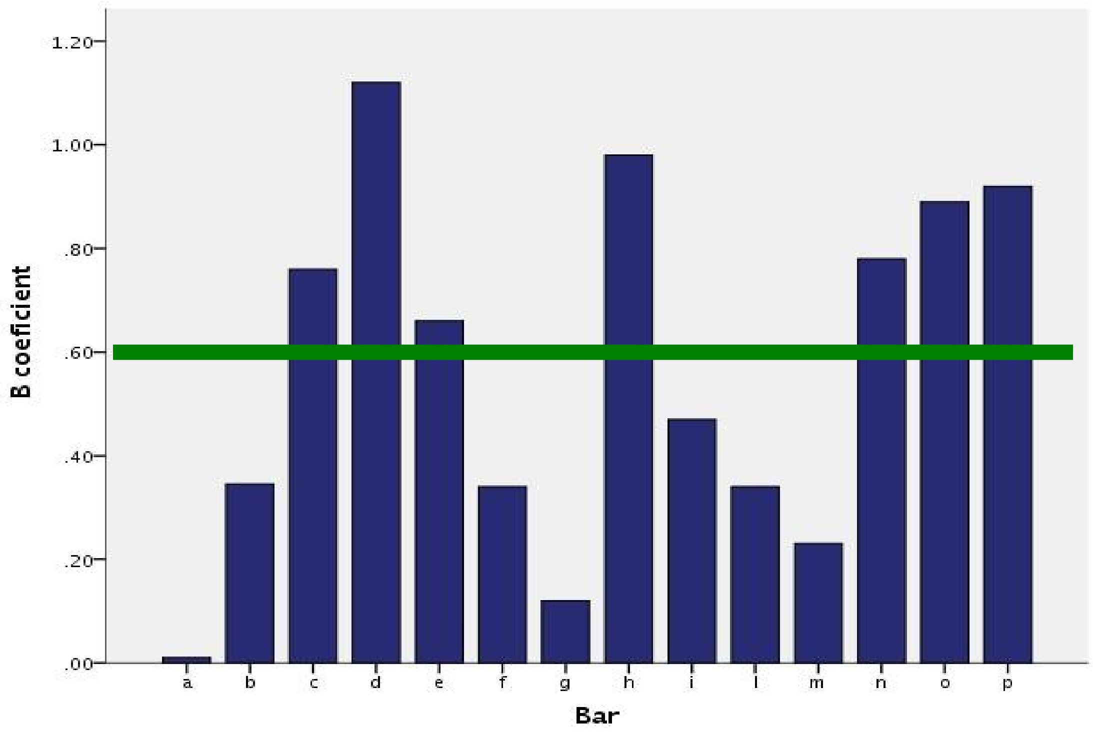


Fixed coefficients

- If coefficients vary as a variable in the population, they will have a mean and a variance, that we can estimate in our data

Mean of coefficients

$$\bar{b} = \frac{\sum_j b_j}{k}$$



Recall the mean is a fixed parameter for a distribution, and so is the mean of the coefficients: it is a **fixed effect**

The Model

- We can now define a model with a regression for each cluster and the mean values of coefficients

One regression per cluster

$$\hat{y}_{ij} = a_j + b_j \cdot x_{ij}$$

Each coefficient is defined as the deviation from the mean coefficient

$$b'_j = b_j - \bar{b}$$

Overall model

$$\hat{y}_{ij} = a_j + b'_j \cdot x_{ij} + \bar{b} \cdot x_{ij}$$

The Model

- We can now define a model with a regression for each cluster and the mean value of coefficients

Overall model

$$\hat{y}_{ij} = \underbrace{a_j + b'_j}_{\text{Random coefficients}} \cdot x_{ij} + \underbrace{\bar{b}}_{\text{Fixed coefficient}} \cdot x_{ij}$$

The mixed model

- The same goes for the intercepts

One regression per cluster

$$\hat{y}_{ij} = a_j + b_j \cdot x_{ij}$$

Intercepts as deviations from the average intercept

$$a'_j = a_j - \bar{a}$$

Overall model

$$\hat{y}_{ij} = \bar{a} + a'_j + \bar{b} \cdot x_{ij} + b'_j \cdot x_{ij}$$

The mixed model

- We can now define a model with a regression for each cluster and the mean values of coefficients

Overall model

$$\hat{y}_{ij} = \bar{a} + a'_j + \bar{b} \cdot x_{ij} + b'_j \cdot x_{ij}$$

Random coefficients

Fixed coefficients

A linear model which contains both fixed and random effects is called a
Linear Mixed Model

GLM as a special case

It is clear that everything we know for the GLM applies here: the GLM is in fact a special case of the LMM, where there are not random effects

LMM

$$\hat{y}_{ij} = \bar{a} + a'_j + \bar{b} \cdot x_{ij} + b'_j \cdot x_{ij}$$

GLM

$$\hat{y}_{ij} = \bar{a} + \bar{b} \cdot x_{ij}$$

The mixed model

- In practice, mixed models allow to estimate the kind of effects we can estimate with the GLM, but they allow the effects to vary across clusters.

- Effects that vary across clusters are called **random effects**

- Effects that do not vary (the ones that are the same across clusters) are said to be **fixed effects**

The mixed model

- To specify a correct model, we only need to understand if there are **clusters of cases** (measures or participants) and decide which coefficients (intercepts or b coefficients) may vary across those clusters

- The fixed effects of the model are interpreted like in the GLM (regression/ANOVA)

- **Random effects** are generally not interpreted, but we can look at their variance to decide to keep them as random (variance>0) or fix them.

- In this way we take into the account the dependence among data

Building a model

To build a model in a simple way, we need to answer very few questions:

- What is (are) the cluster variable(s)?
- What are the fixed effects?
- What are the random effects?

A clustering variable

- **What is (are) the cluster variable(s)?**
- What are the fixed effects?
- What are the random effects?

  - Any variable that groups observations (cases or measurements) such that scores may be more similar within each group than across groups
  - Any variable whose levels (groups) are a sample of a larger population of levels (groups)
  - Example: bars created groups of scores (participants) that may be more similar within the bar that across bars

Fixed effects

- What is (are) the cluster variable(s)?
- **What are the fixed effects?**
- What are the random effects?
  - Any effect that we are interested in on average (as in a standard ANOVA/Regression)
  - Example: the effect of beer on smiles in general

Fixed effects

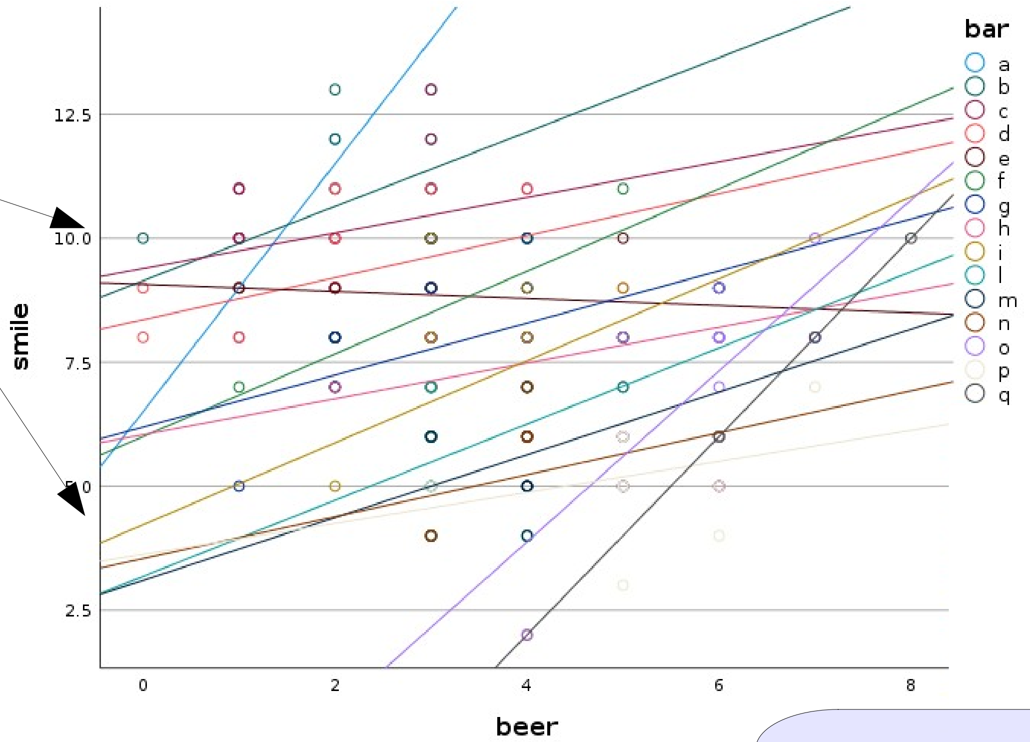
- What is (are) the cluster variable(s)?
- What are the fixed effects?
- **What are the random effects?**
  - Any effect that may vary from cluster to cluster
  - (Thus:) **Any effect that can be computed within each cluster**
  - Example: the intercepts and the effect of beer on smiles each bar

Beers at the bar

We start with a simple model

Bar

Intercepts seem to be different from bar to bar



Also the slopes may be different from bar to bar

Beers at the bar

We can now try a model where also the **b** coefficients are allow to vary across clusters

$$y_{ij} = \bar{a} + a_j + \bar{b} \cdot x_{ij} + b \cdot x_{ij} + e_{ij}$$

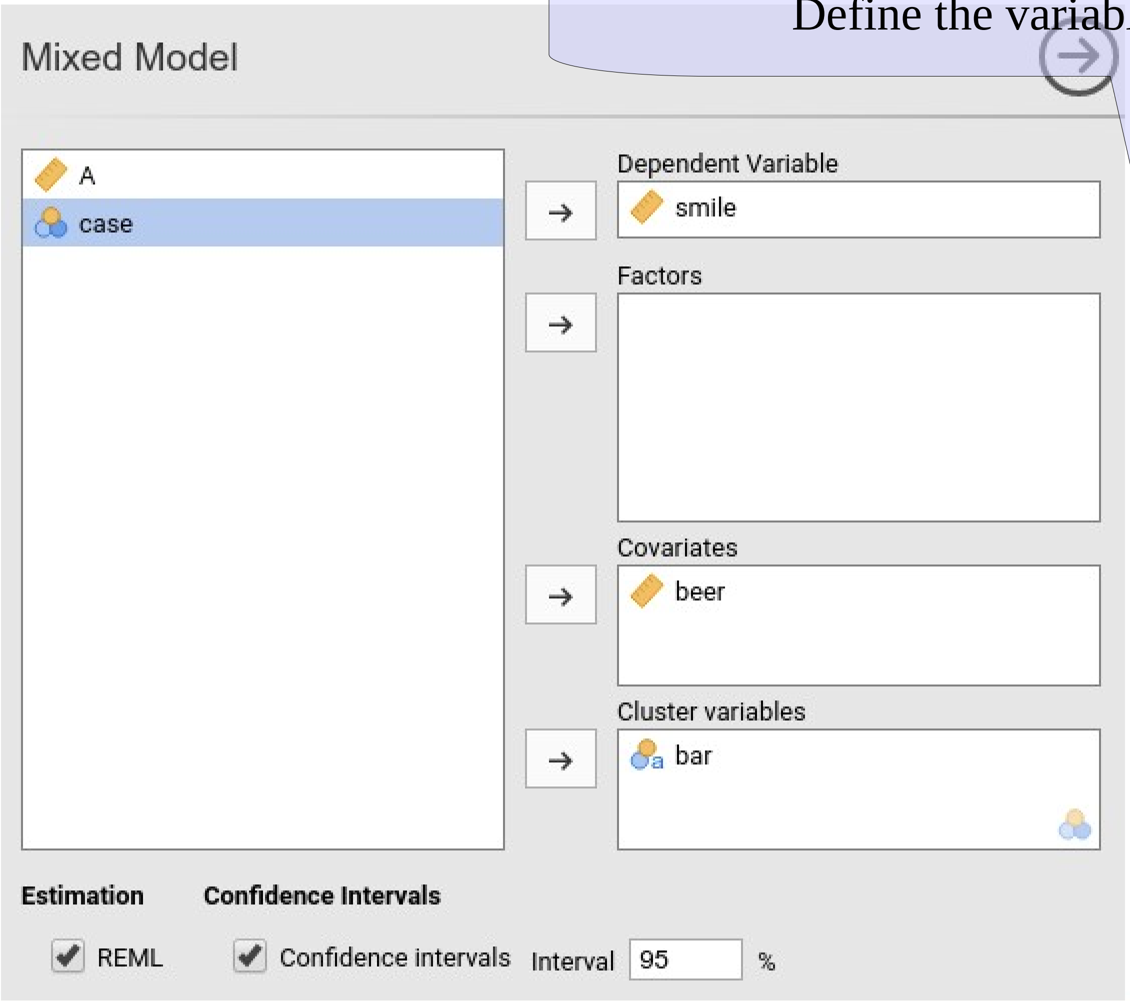
- Fixed effects? Intercept and beer effect
- Random effects? Intercepts and b coefficients
- Clusters? bar

Some authors may call this model:
Random-coefficients regression
or
Intercepts- and Slopes-as-outcomes model

jamovi

- We setup the variables

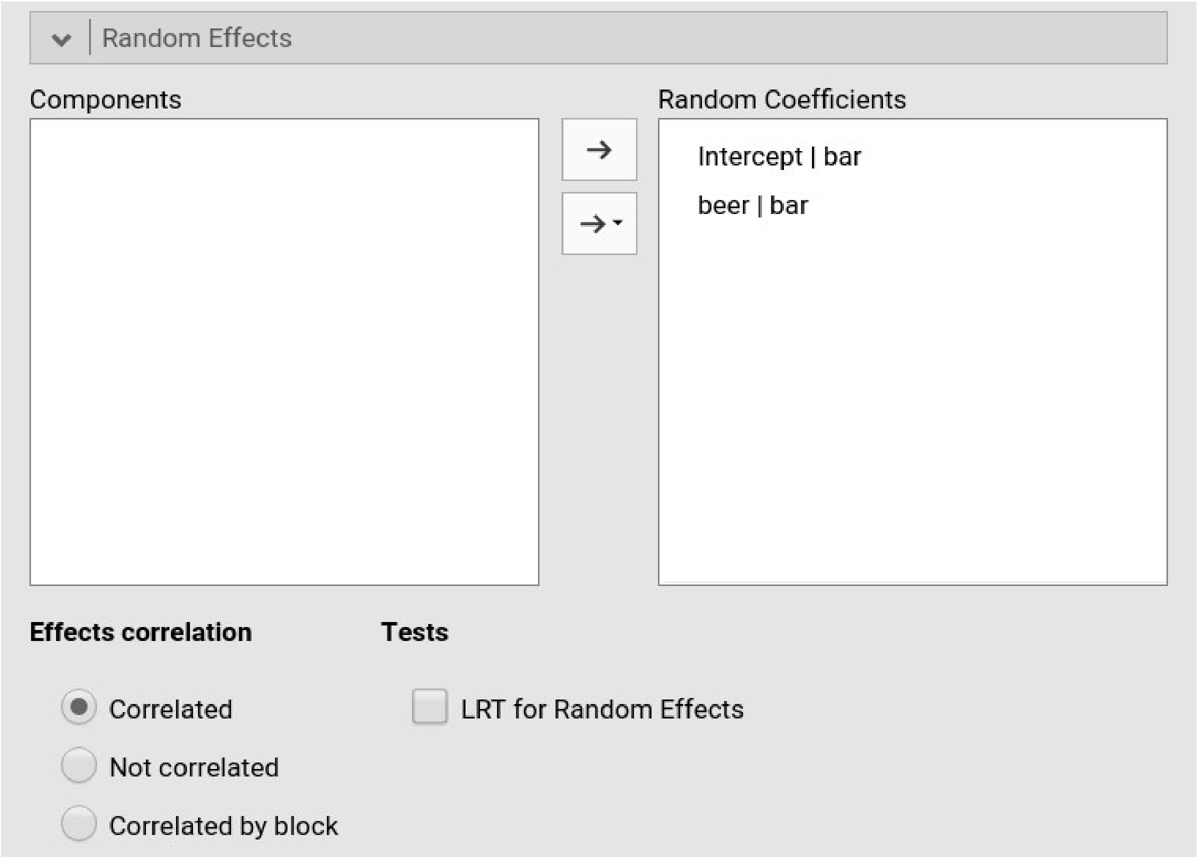
Define the variables role



Jamovi

- We define the intercept and the effect of beer as a random coefficients

Define the random component



Random component

- As soon as you define the random component, you get the results

Random Components

Groups	Name	SD	Variance	ICC
bar	(Intercept)	2.417	5.842	0.803
	beer	0.167	0.028	
Residual		1.196	1.431	

Note. Number of Obs: 234 , groups: bar 15

Random coefficients variances

Random Parameters correlations

Groups	Param.1	Param.2	Corr.
bar	(Intercept)	beer	−0.766

Random coefficients correlation

ICC

- The intra-class correlation indicates the proportion of variance explained by the differences across clusters

σ_a

Random Components				
Groups	Name	SD	Variance	ICC
bar	(Intercept)	2.417	5.842	0.803
	beer	0.167	0.028	
Residual		1.196	1.431	

Note. Number of Obs: 234 , groups: bar 15

σ

ICC=Intra-class correlation

$$ICC = \frac{\sigma_a}{\sigma_a + \sigma}$$

$$ICC = \frac{5.842}{5.842 + 1.431} = .803$$

Variance

Variances of random coefficients inform us on the variability of the effects

- As long as a coefficient has variability, we keep it in the random component

- When variances are exactly zero (and software gives a general warning), effects should be set only as fixed

Results

- Model goodness of fit

Model Results

Model Fit

Type	R^2	df	LRT X^2	p
Conditional	0.822	4	203.003	< .001
Marginal	0.090	1	17.016	< .001

[4]

>

R-squared Marginal: How much variance can the fixed effects alone explain of the overall variance

R-squared Conditional: How much variance can the fixed and random effects together explain of the overall variance

Results

- Ominibus Tests (like in ANOVA)

F-test for the main effect of beer

Model Results

Fixed Effect Omnibus tests

	F	Num df	Den df	p
beer	36.057	1	7.234	< .001

Note. Satterthwaite method for degrees of freedom

Results

- Coefficients (like in Regression)

Coefficients for the main effect of beer

Fixed Effects Parameter Estimates

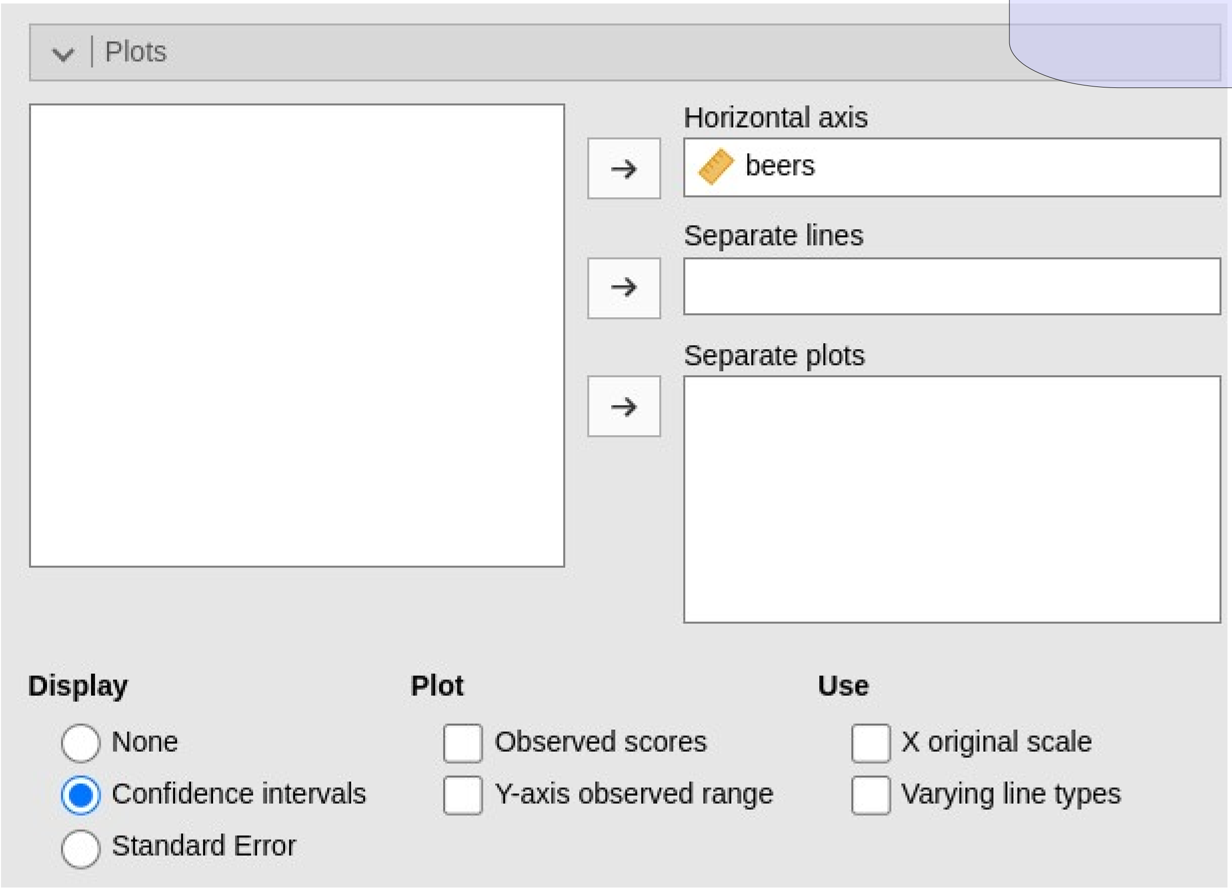
Names	Estimate	SE	95% Confidence Interval Lower	95% Confidence Interval Upper	df	t	p
(Intercept)	7.610	0.633	6.368	8.851	12.928	12.013	< .001
beer	0.555	0.093	0.374	0.737	7.234	6.005	< .001

Intercept: On average, as beers increase on 1 unit, we expect smile to increase of .555 smiles

Jamovi

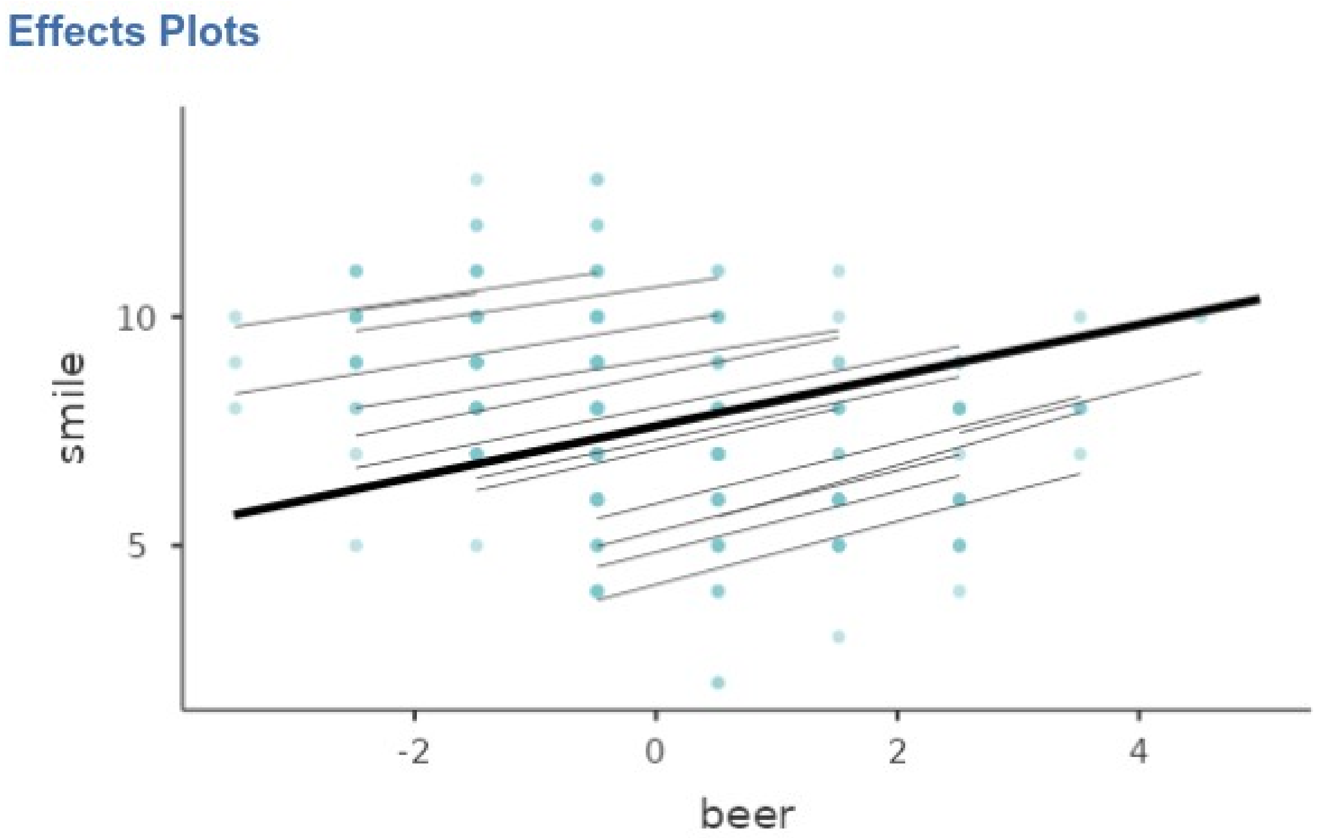
- Jamovi can plot any order of interaction

Plot



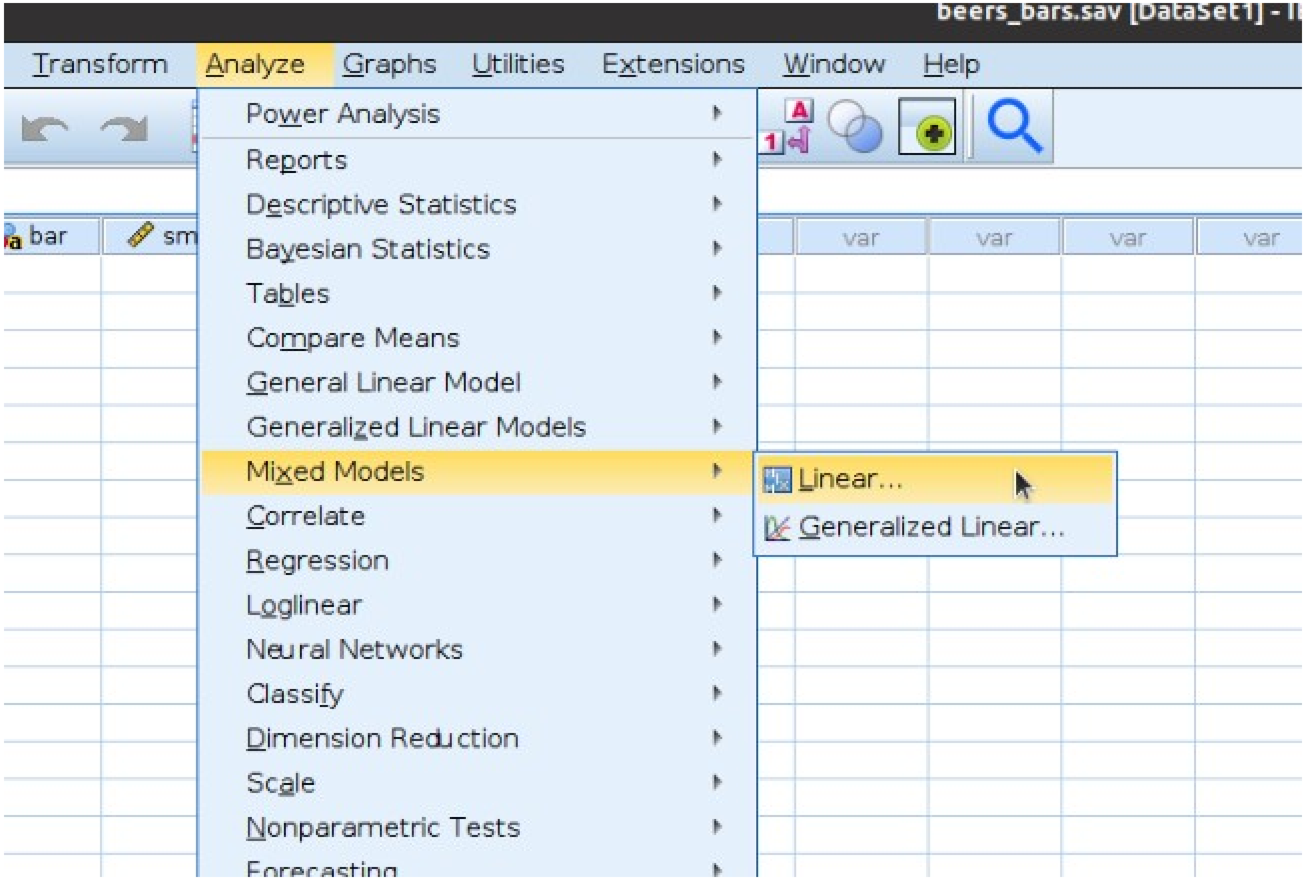
Jamovi

- In this case is fixed and random effects regression lines



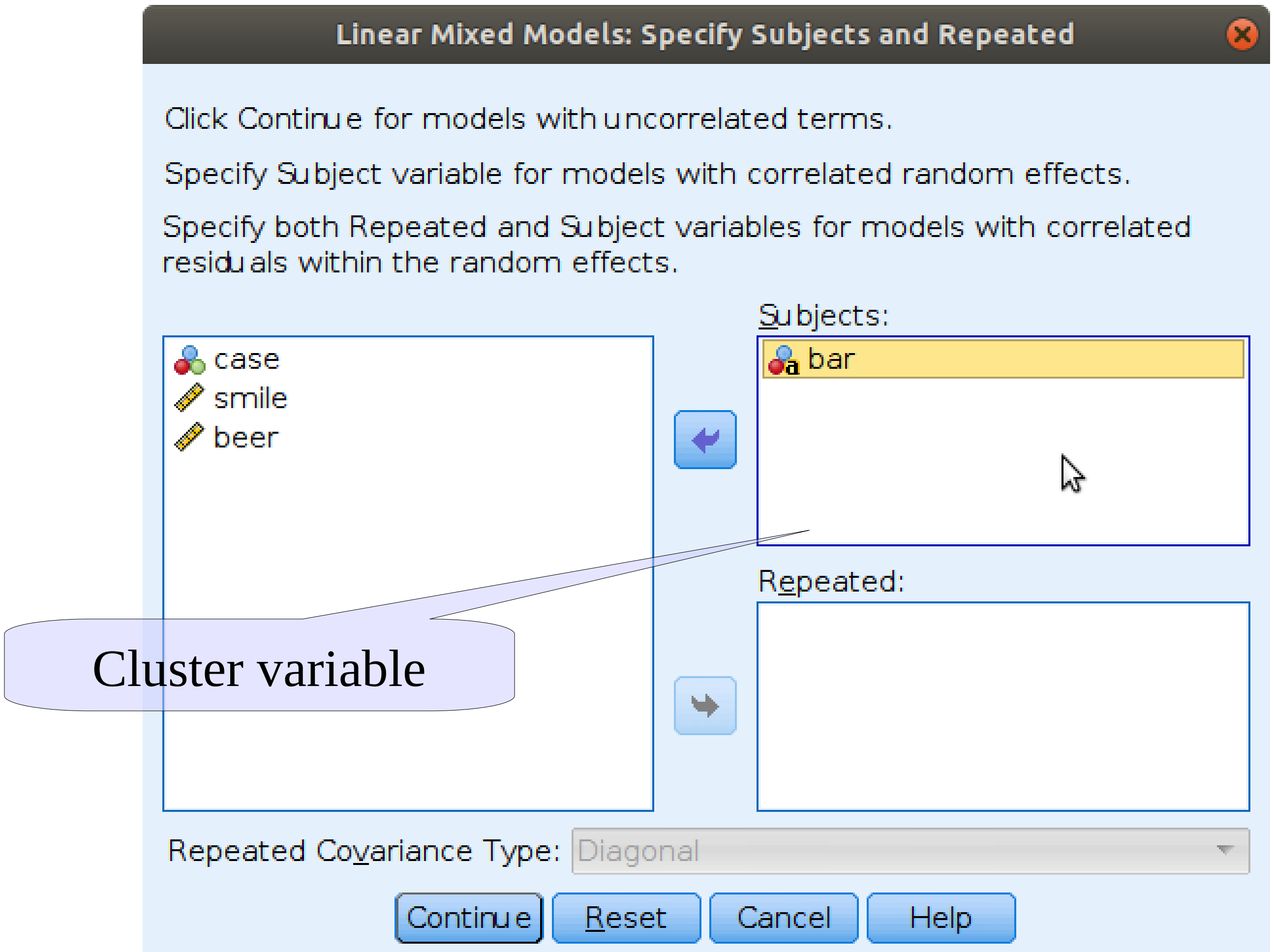
SPSS

Analyze → Mixed Models → Linear



SPSS Input

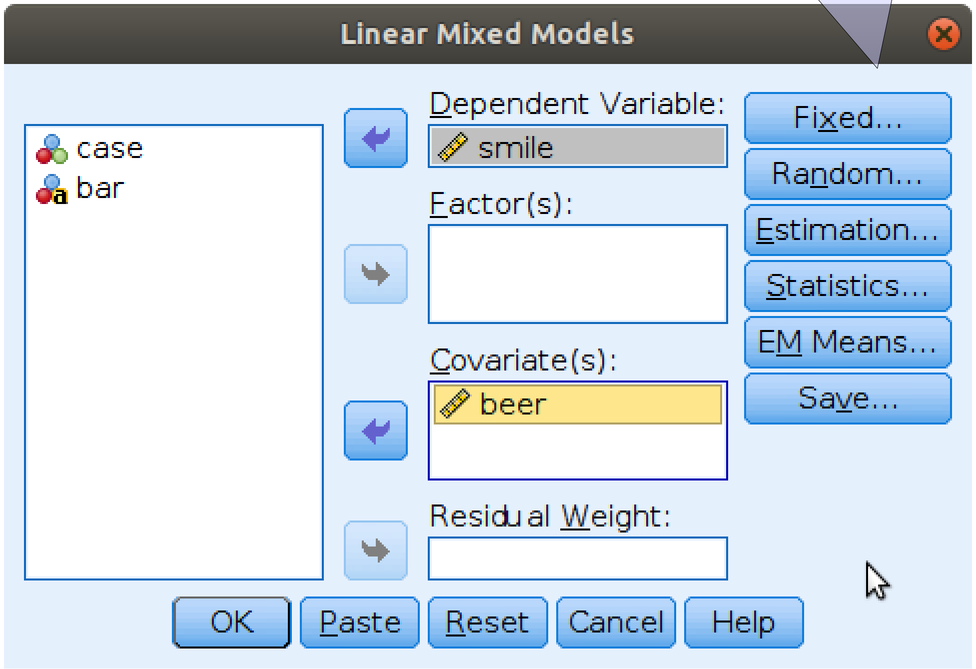
Analyze → Mixed Models → Linear



SPSS Input

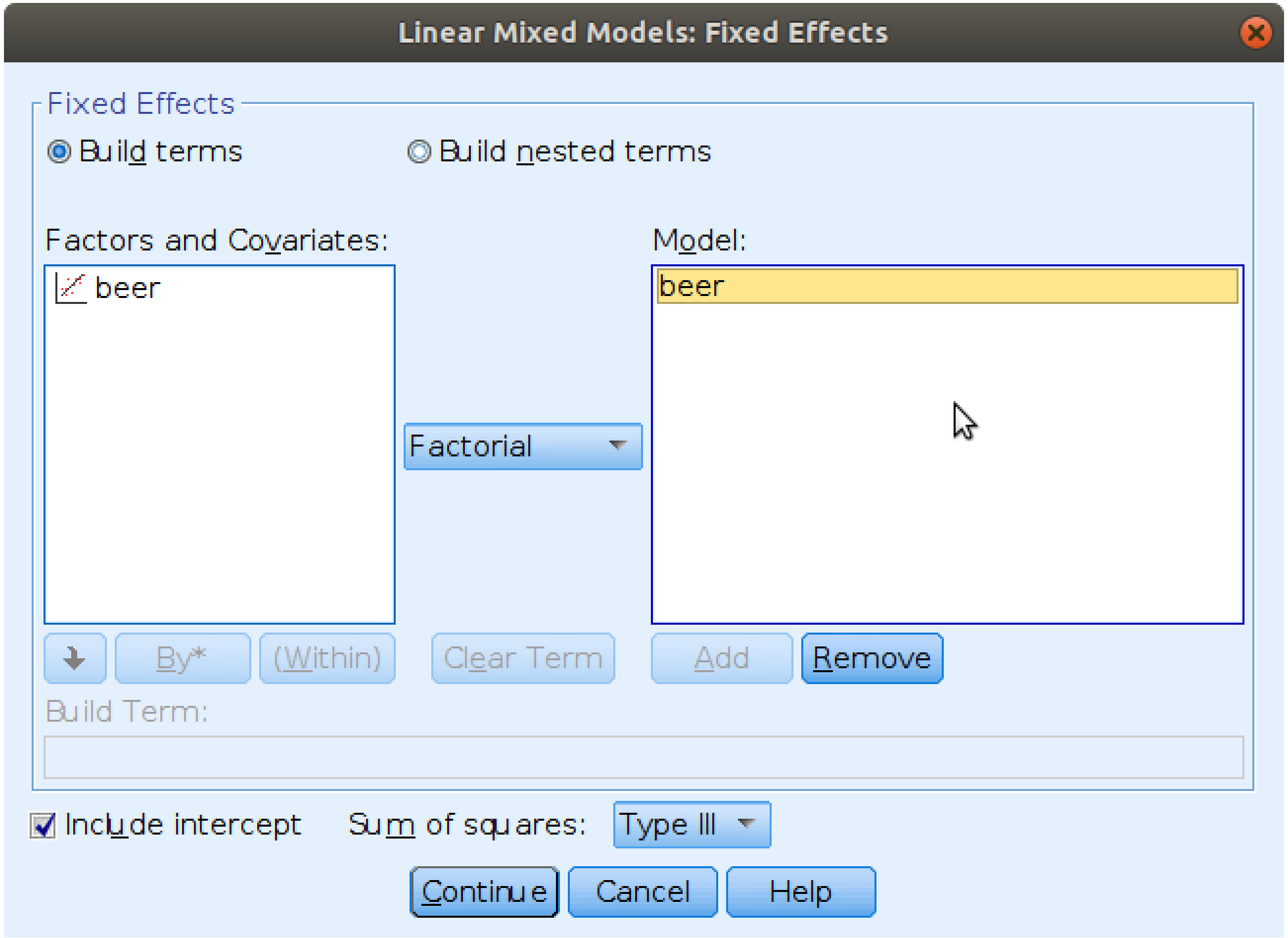
Analyze → Mixed Models → Linear

Then select Fixed



SPSS Input

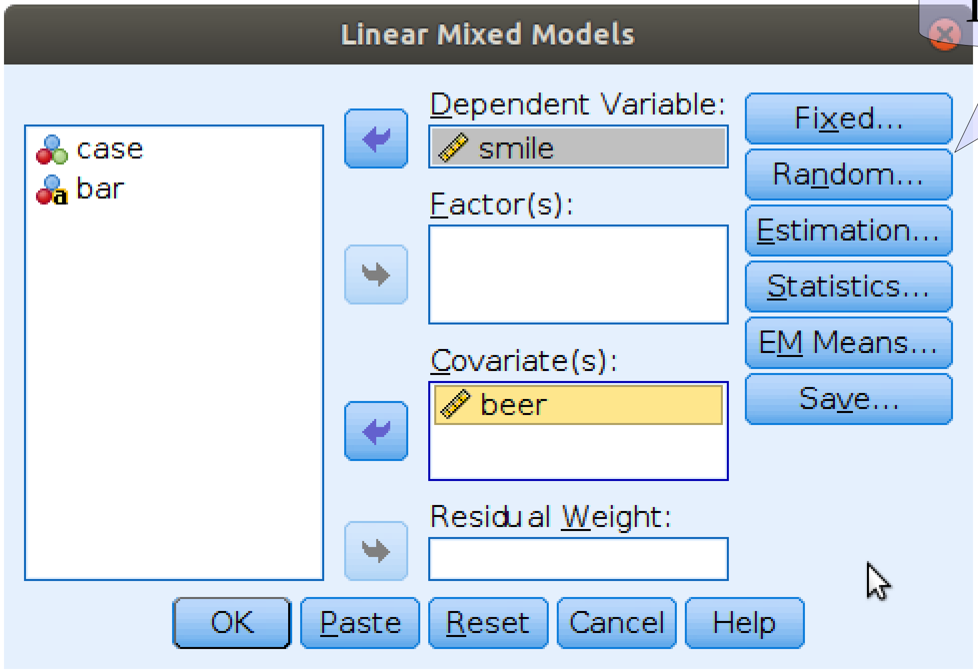
Analyze → Mixed Models → Linear



SPSS Input

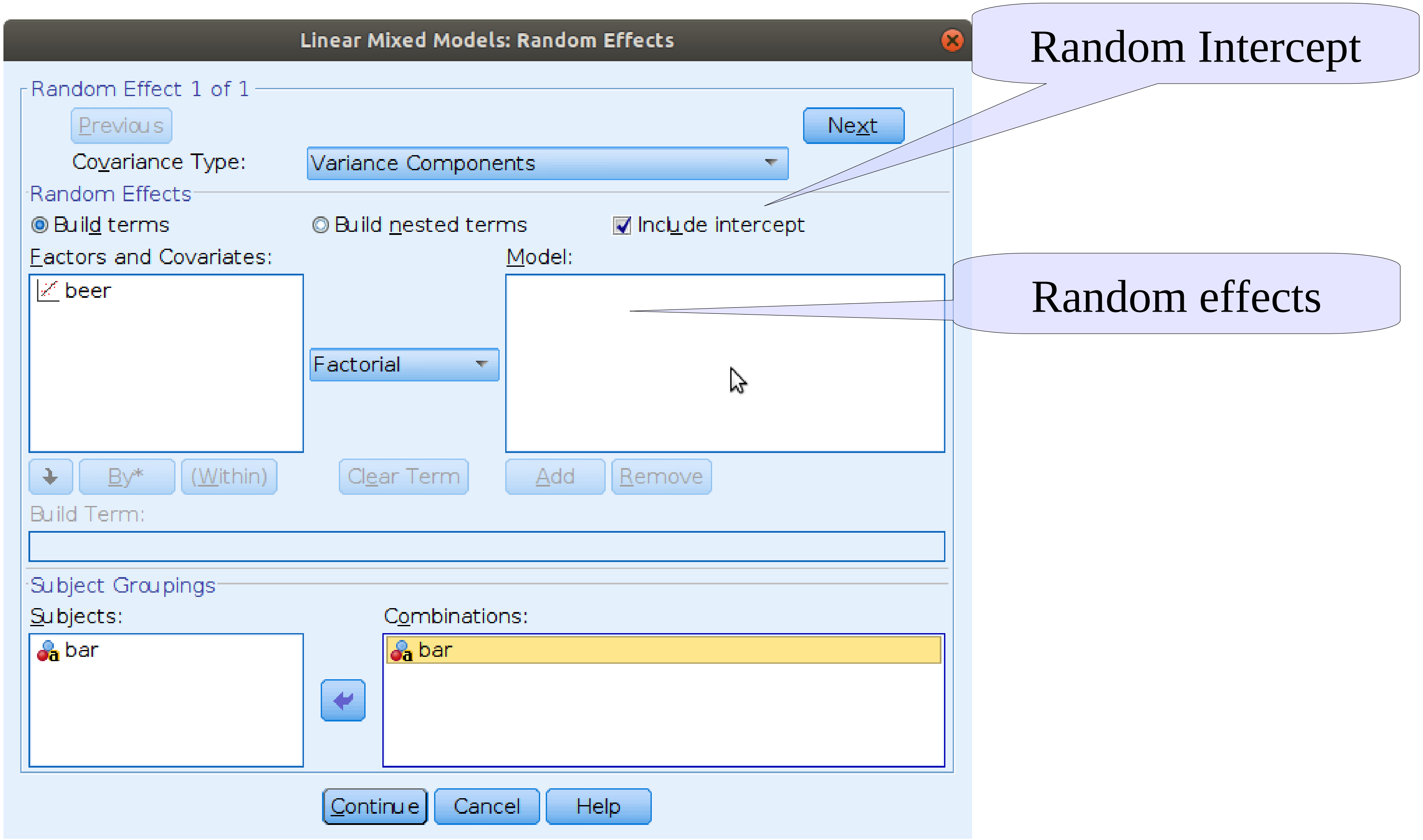
Analyze → Mixed Models → Linear

Then select Random



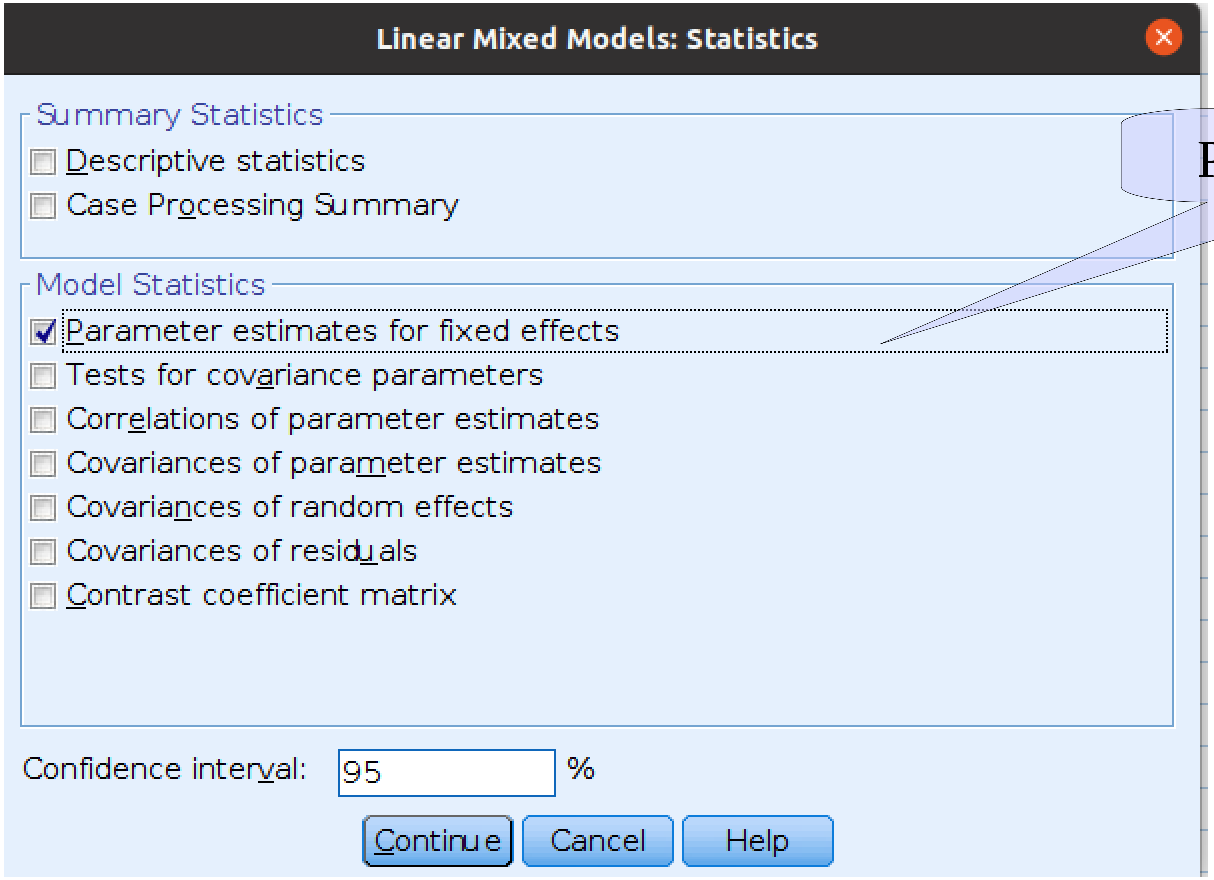
SPSS Input

Analyze → Mixed Models → Linear



SPSS Input

Analyze → Mixed Models → Linear



Print coefficients

R syntax

Load the required libraries

```
library(lme4)
library(lmerTest)

dat<-read.csv2('../../../Forge/jamovi/gamlj/gamlj/data/beers_bars.csv')

mm1<-lmer(smile~1+beer+(1+beer|bar),data=dat)

summary(mm1)
```

Random effects

Cluster variable

Fixed effects (intercept can be omitted as ti is included by default)

R syntax

Test random coefficients

Omnibus Tests

```
rand(mm1)

anova(mm1)

plot(dat$smile~dat$beer,col=dat$bar)
```

Plot

Mixed Linear Models

- With the mixed model one can take into the account dependency among measures (within clusters) almost in any situation
- It allows applying the GLM logic to a broader range of designs
- Any kind of independent variables
- Generalizes to the generalized linear model (logistic etc)
- Efficient handling of missing values
- **Multi-level research designs**
- **Repeated measures designs**

Multi-level Questions

Multi-level question

In psychology, often the research question regards variables that vary at different levels of the sampling structure

- **The multi-level questions are answered estimating a mixed model**
- What is peculiar**:**
  - The importance of the clustering variables (higher levels)
  - The research questions
  - The cluster level is called group level (*group=cluster in this terminology*)

Example

- Assume we have a multi-country research, in which we measured individuals (people) *charity contribution* and their *income* (*individual level*).

- We can have information about country taxes regulations (*country level*)

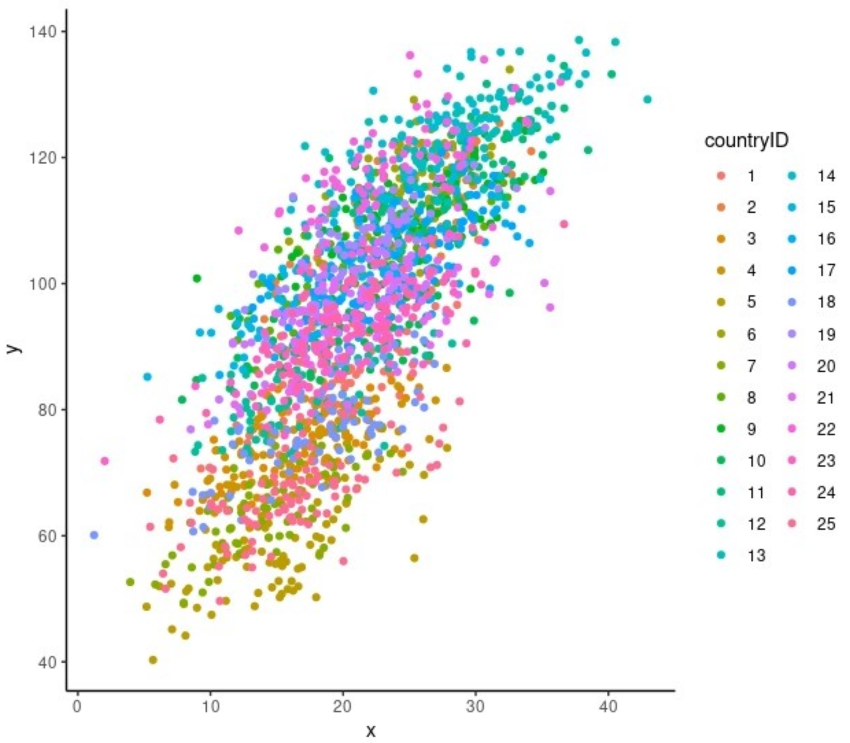
- We are interested on the relationship between *contribution* and *income* at the **individual level and at the country level**

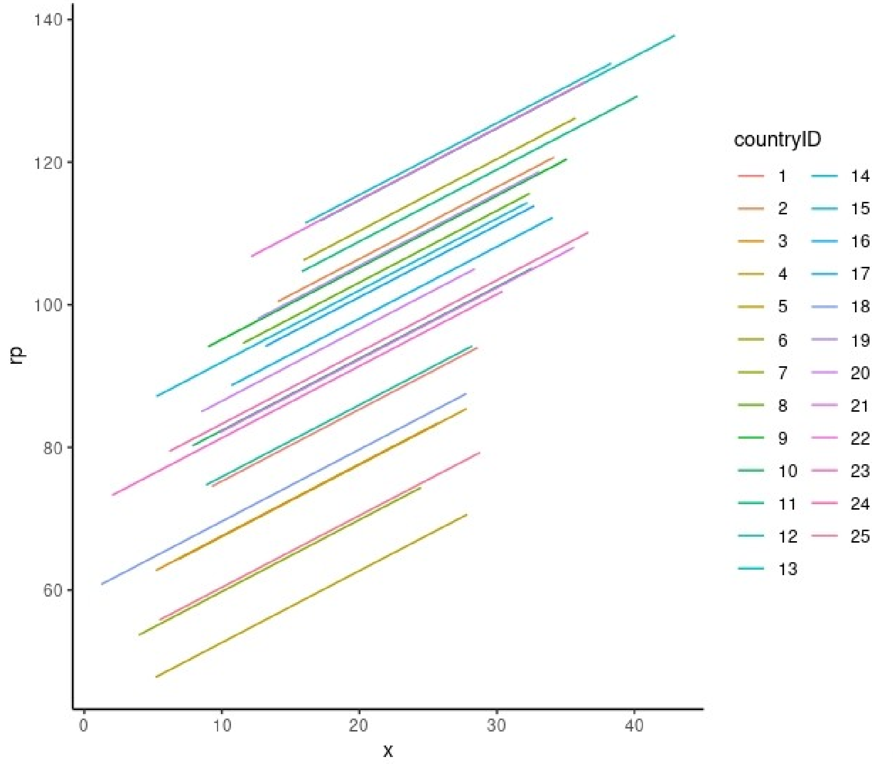
Questions

- We are interested on the relationship between charity donations and income at the **individual levels and at the country level**
- Independently of the country: Does people with higher income contribute more? (*individual level effect*)
- Independently of the people: Do countries with average higher income show higher average contribution? (*country level effect*)

Structure vs aim

- If we only look at the data structure, we have the same structure of the beers at bars example
- But the research aim is different

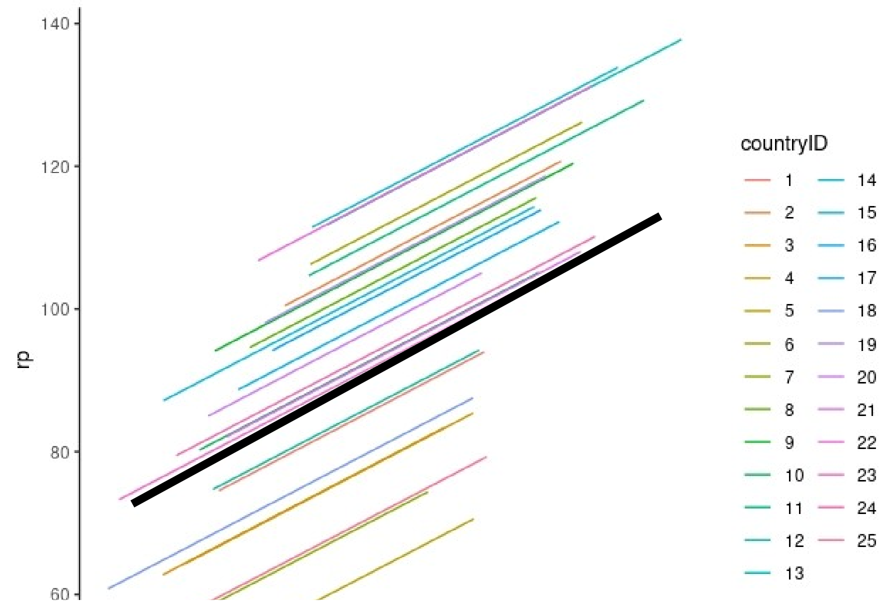




Individual level

- If we fit a model like *beer at bars*, we only get the individual level effect, averaged across countries

Independently of the country: Does people with higher income contribute more? (*individual level effect*)



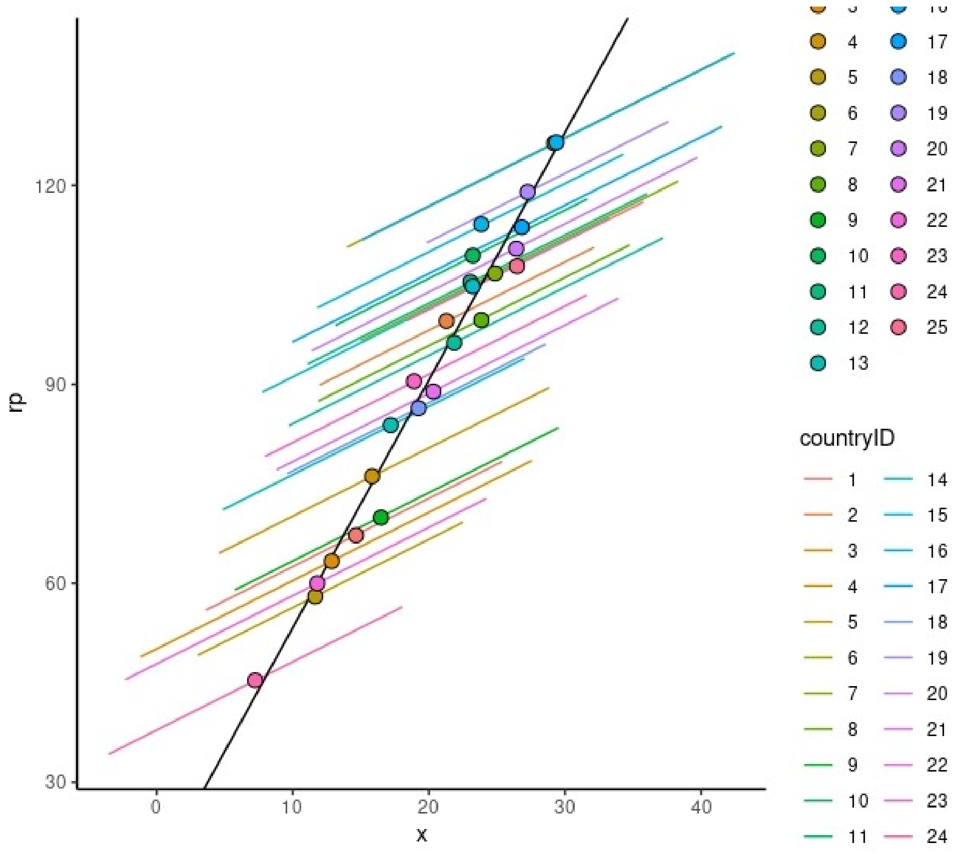
Fixed Effects Parameter Estimates

			95% Confidence Interval				
Names	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	95.73	3.0129	89.830	101.64	24.0	31.8	< .001
x	1.01	0.0235	0.964	1.06	1928.6	43.0	< .001

Country level

- But income may have an effect also at the country (second) level

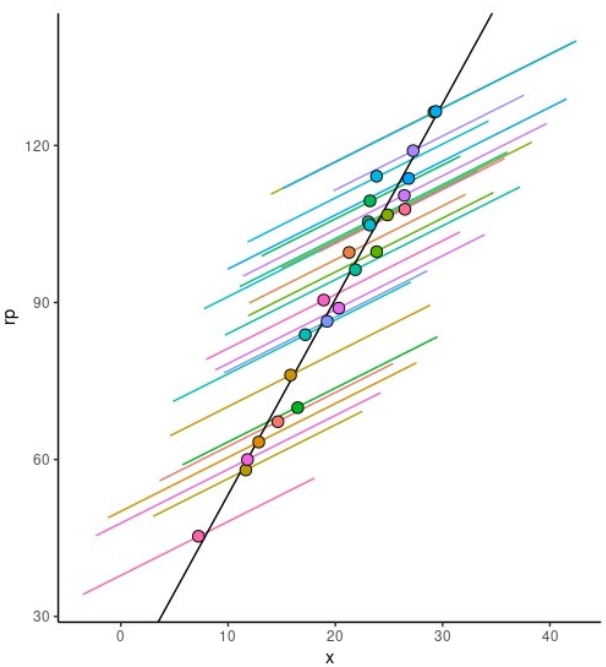
Independently of the people: Do countries with average higher income show higher average contribution? (*country level effect*)



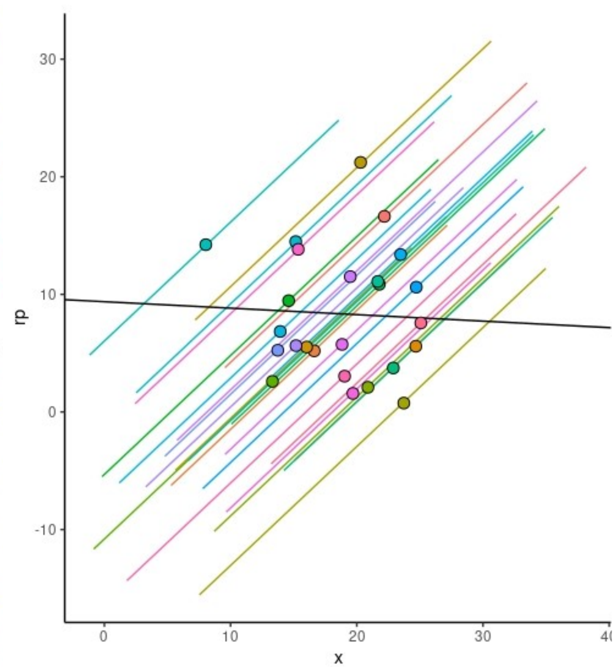
Country vs individual level

- The effect of a variable at each level is **independent** of the effect at any other level

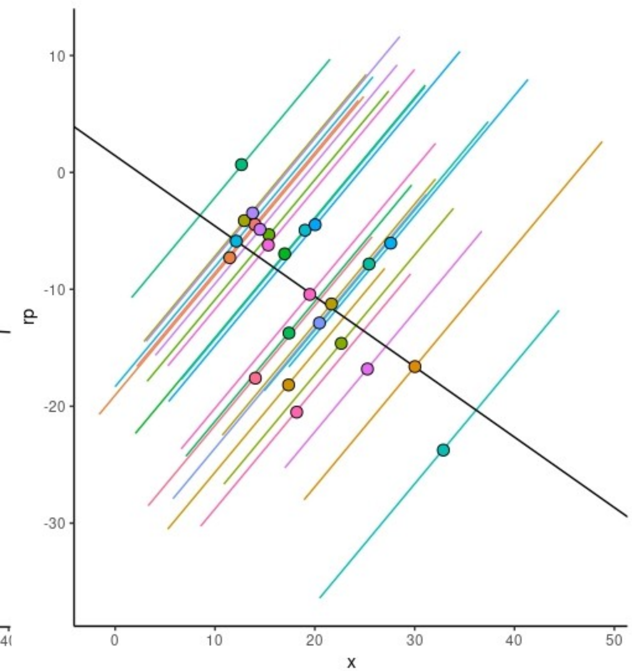
Individual +, country +



individual +, country 0



individual +, country -



The mixed model

- To capture the effect of countries (second level) we should include the country levels means
- To make it independent of people levels, we group-center individual level **x**

$$\hat{y}_{ij} = \bar{a} + a'_j + b_1 \cdot (x_{ij} - \bar{x}_j) + b_2 \cdot \bar{x}_j$$

Group centered x (country centered)

Group mean (country mean)

Coefficients

- The model returns the effects at level 1 (individuals) and level 2 (country)

$$\hat{y}_{ij} = \bar{a} + a'_j + b_1 \cdot (x_{ij} - \bar{x}_j) + b_2 \cdot \bar{x}_j$$

Independently of the country: Does people with higher income contribute more?

Do countries with average higher income show higher average contribution?

Data

- We simply compute two new variables: the group centered x and the group means

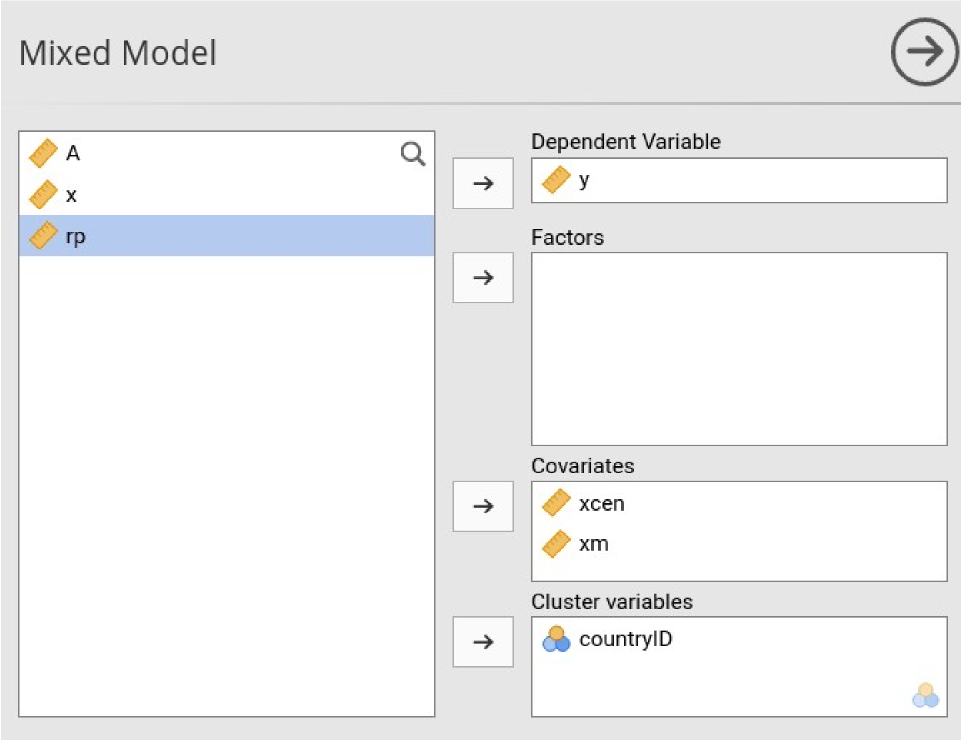
$$xcen=\left(x_{ij}-\bar{x}_j\right)$$

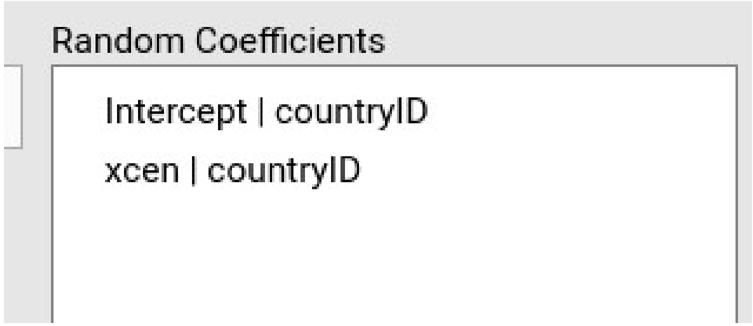
$$xm=\bar{x}_j$$

countryID	x	y	xm	xcen
10	12.919	86.400	20.547	-7.628
10	27.128	98.083	20.547	6.581
10	21.709	99.612	20.547	1.163
10	20.586	94.523	20.547	0.040
10	14.580	81.559	20.547	-5.967
10	28.697	106.248	20.547	8.150
10	8.937	84.683	20.547	-11.609
10	17.241	94.658	20.547	-3.306
10	18.252	84.677	20.547	-2.295
10	18.913	90.324	20.547	-1.633
11	33.202	129.347	28.333	4.870
11	23.598	114.079	28.333	-4.735
11	23.746	109.661	28.333	-4.587
11	26.870	113.275	28.333	-1.463
11	25.305	116.171	28.333	-3.028
11	31.789	114.460	28.333	3.456
11	18.956	119.912	28.333	-9.377
11	32.524	123.181	28.333	4.191

Model

- We use them as independent variables





Results (fixed effect)

- And interpret the coefficients accordingly

Fixed Effects Parameter Estimates

			95% Confidence Interval				
Names	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	95.65	1.3013	93.104	98.21	23.1	73.5	< .001
xcen	1.01	0.0278	0.954	1.06	21.8	36.3	< .001
xm	4.37	0.3126	3.757	4.98	23.1	14.0	< .001

Within each country: as people income increases of 1 unit, contribution increases of **1.01**

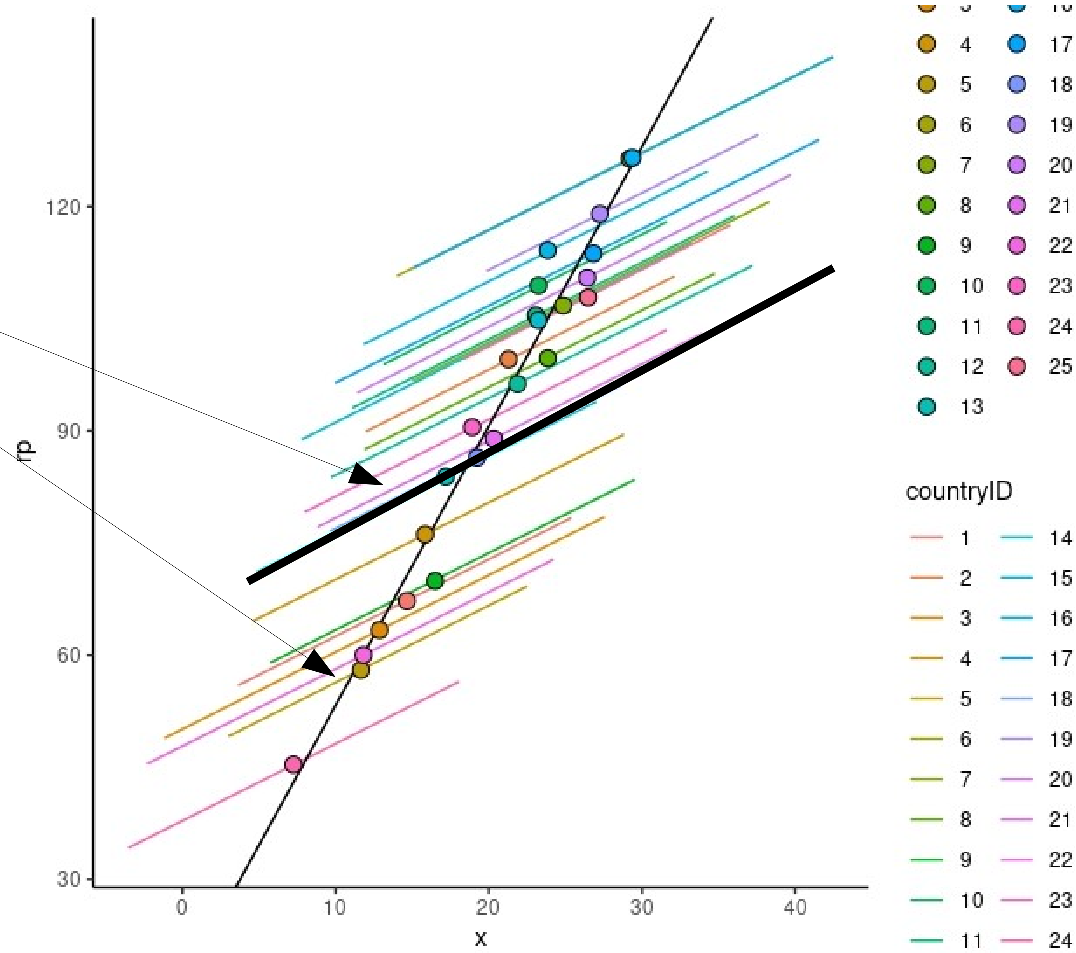
Across countries: As the average income of a country increases 1 unit, the average contribution increases of **4.37** units

Results (fixed effect)

- And interpret the coefficients accordingly

Fixed Effects Parameter Estimates

Names	Estimate	SE
(Intercept)	95.65	1.3013
xcen	1.01	0.0278
xm	4.37	0.3126



Multi-level models

- **The multi-level model is estimated using a mixed model**

- What is peculiar**:**

  - We want to estimate predictors effects at each level

  - We want to estimate higher level effect over and beyond lower level effect

Mixed Linear Models

- With the mixed model one can take into the account dependency among measures (within clusters) almost in any situation
- It allows applying the GLM logic to a broader range of designs
- Any kind of independent variables
- Generalizes to the generalized linear model (logistic etc)
- Efficient handling of missing values
- Multi-level research designs
- **Repeated measures designs**

Repeated Measures ANOVA as a linear mixed model

A repeated measures design

- Consider now a classical repeated measure design (within-subjects) the levels of the WS IV (5 different conditions) are represented by different measures taken on the same person

Condition

Participants	1	2	3	4	5
1	Y11	Y21	Y31	Y41	Y51
2	Y12	Y22	Y32	Y42	Y52
3	Y13	Y23	Y33	Y43	Y53
….					
N	Y1n	Y2n	Y3n	Y4n	Y5n

Standard file format

- For many applications of the repeated-measure design, each level of the WS-factor is represented by a column in the file

One participant, one row

	id	y.1	y.2	y.3	y.4	y.5
1	1	0.140	0.220	0.439	0.271	0.009
2	2	0.431	0.518	0.492	0.483	0.433
3	3	0.612	0.431	0.446	0.509	0.573
4	4	0.291	0.702	1.000	0.892	0.751
5	5	0.156	0.494	0.500	0.564	0.286
6	6	0.700	0.364	0.573	0.572	0.690
7	7	0.346	0.513	0.572	0.460	0.766
8	8	0.446	0.493	0.545	0.406	0.429
9	9	0.052	0.553	0.333	0.535	0.531
10	10	0.103	0.347	0.358	0.567	0.668
11	11	0.141	0.453	0.373	0.252	0.287
12	12	0.043	0.736	0.541	0.534	0.348
13	13	0.622	0.727	0.529	0.305	0.483
14	14	0.154	0.223	0.101	0.167	0.167
15	15	0.715	0.545	0.568	0.527	0.575
16	16	0.928	0.625	0.506	0.418	0.185
17	17	0.578	0.245	0.417	0.489	0.630
18	18	0.262	0.555	0.417	0.470	0.390
19	19	0.725	0.353	0.310	0.170	0.404
20	20	0.037	0.172	0.267	0.328	0.225
21	21	0.101	0.632	0.490	0.175	0.464
22	22	0.733	0.515	0.474	0.599	0.703
23	23	0.225	0.530	0.427	0.386	0.364
24	24	0.221	0.542	0.576	0.795	0.673

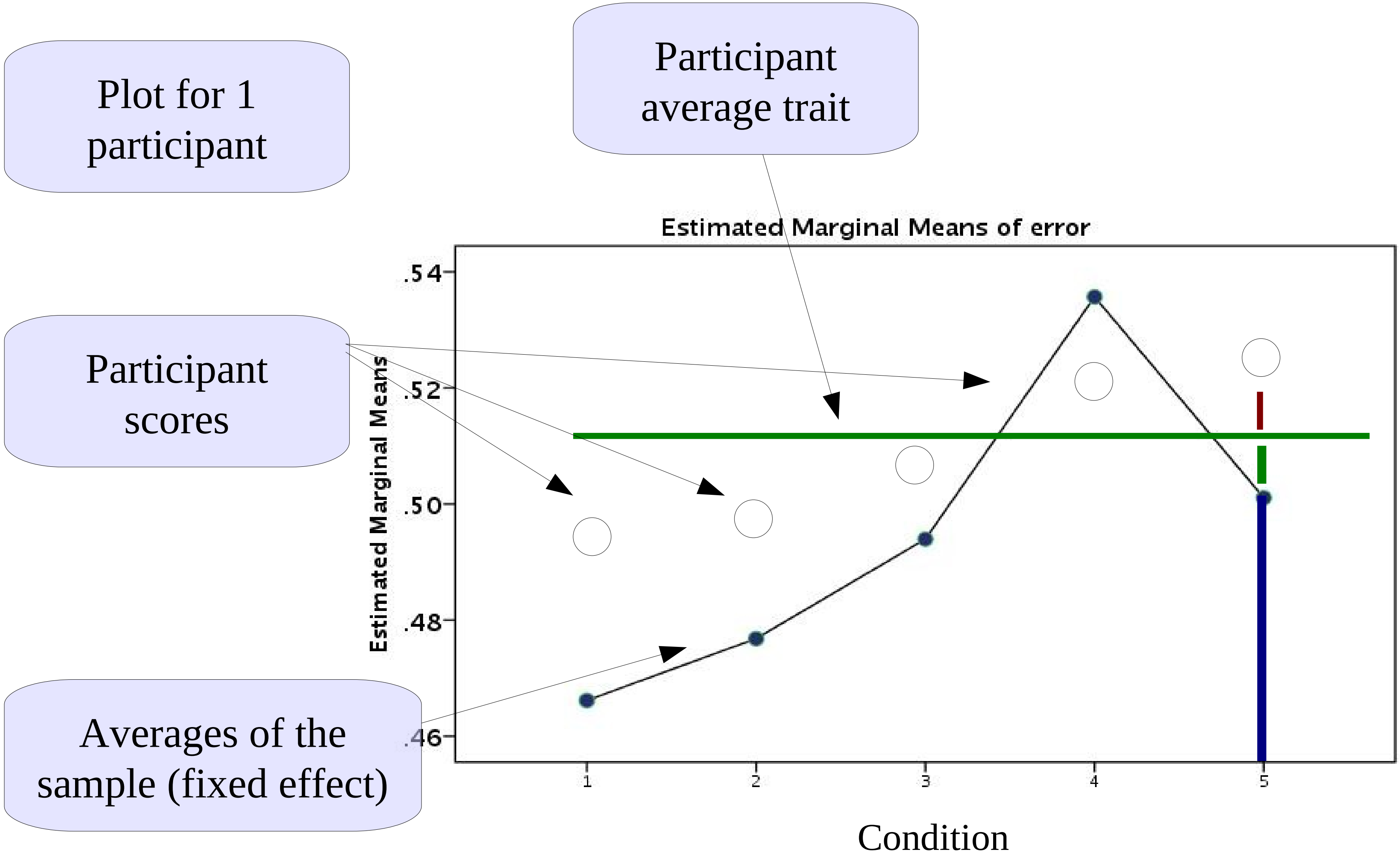
Long file format

- For the mixed model we need to tabulate the data as if they came from a between-subject design

One measure, one row

	id	group	condition	y
1	1	1	1	0.140
2	1	1	2	0.220
3	1	1	3	0.439
4	1	1	4	0.271
5	1	1	5	0.009
6	2	1	1	0.431
7	2	1	2	0.518
8	2	1	3	0.492
9	2	1	4	0.483
10	2	1	5	0.433
11	3	1	1	0.612
12	3	1	2	0.431
13	3	1	3	0.446
14	3	1	4	0.509
15	3	1	5	0.573
16	4	0	1	0.291
17	4	0	2	0.702
18	4	0	3	1.000
19	4	0	4	0.892
20	4	0	5	0.751
21	5	1	1	0.156
22	5	1	2	0.494
23	5	1	3	0.500
24	5	1	4	0.564
25	5	1	5	0.286

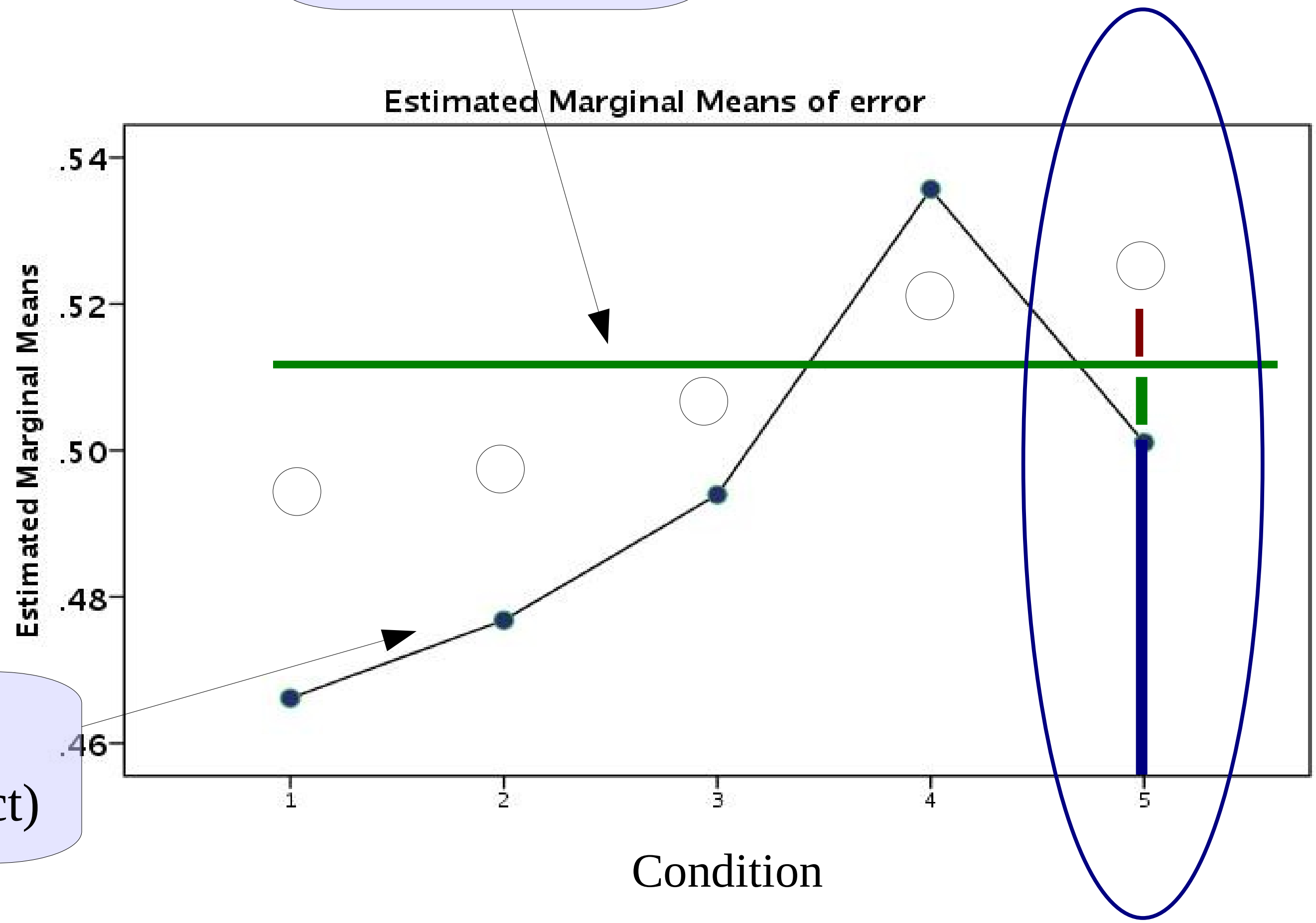
Participant scores



Where does the score come from?

Plot for 1 participant

Participant average trait



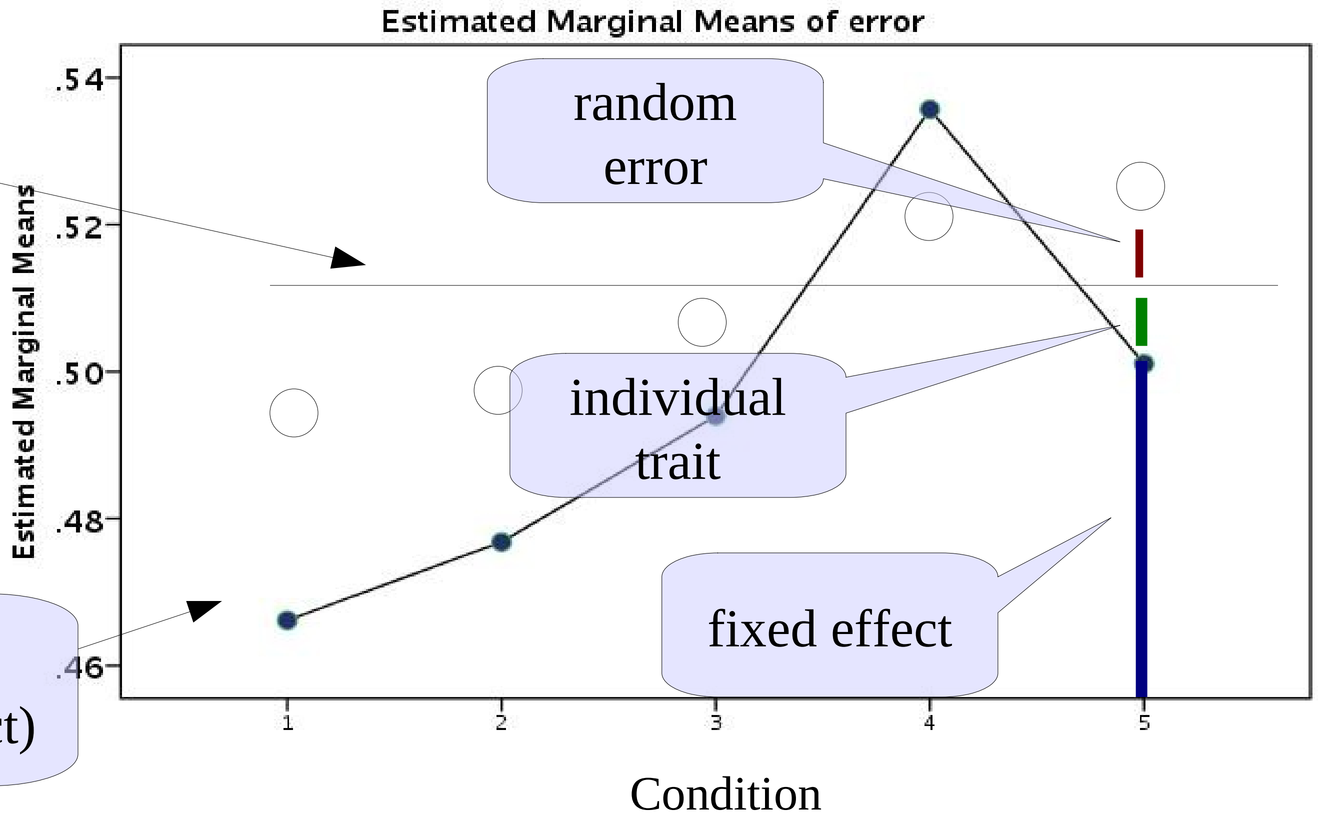
Averages of the sample (fixed effect)

Participant component

Plot for 1 participant

Participant individual trait

Averages of the sample (fixed effect)



Solution

Thus, we should consider an extra residual term which represents participants individual characteristic. This term is the same within each participant

$$Y_{11} = a + b_1 \cdot T_1 + u_1 + e_{11}$$
$$Y_{21} = a + b_2 \cdot T_2 + u_1 + e_{21}$$
$$Y_{31} = a + b_3 \cdot T_3 + u_1 + e_{31}$$

one participant
one trait

Each score,
one residual

Average effects
of trials

....

$$Y_{1j} = a + b_1 \cdot T_1 + u_j + e_{1j}$$
$$Y_{2j} = a + b_2 \cdot T_2 + u_j + e_{2j}$$
$$Y_{3j} = a + b_3 \cdot T_3 + u_j + e_{3j}$$

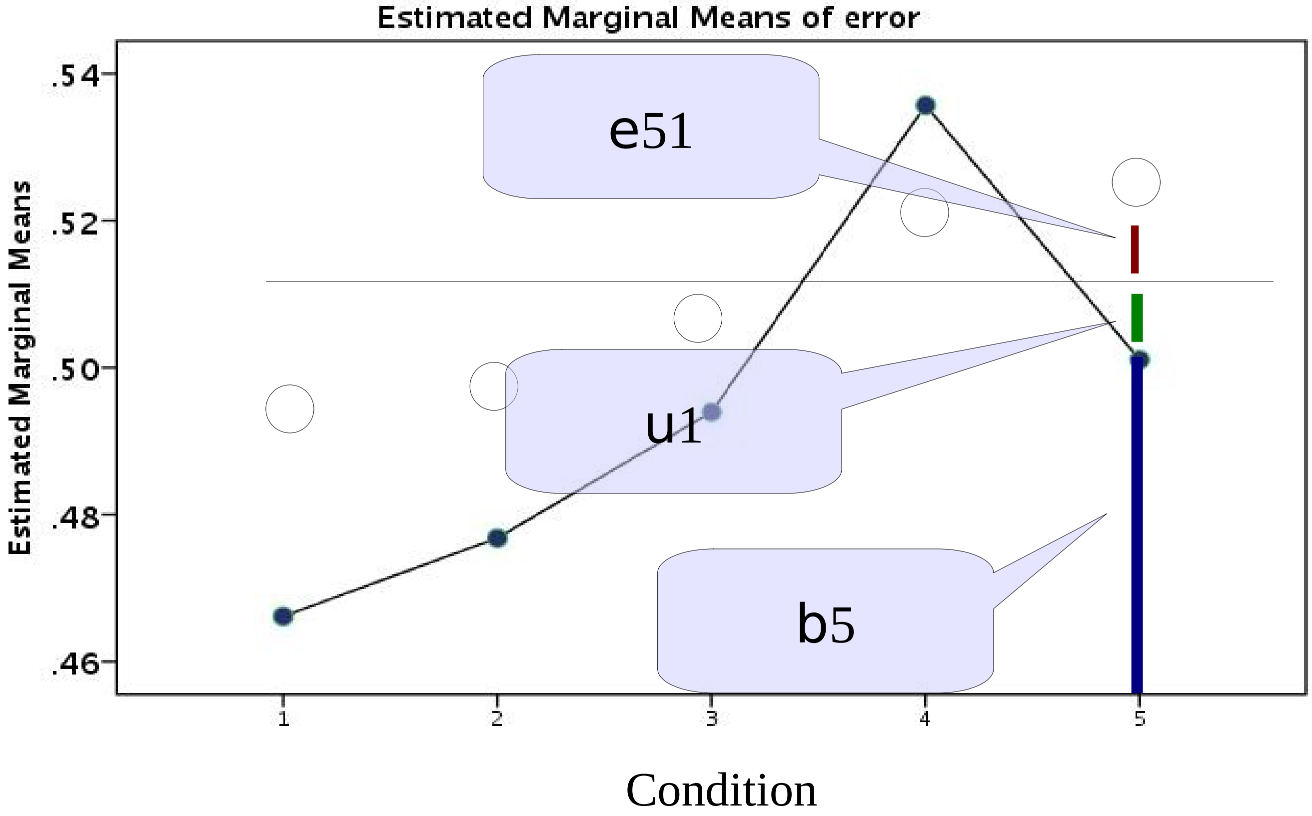
Each score,
one error

One participant
one trait

We assume the 5 trials are dummy coded

Participant component

$$Y_{51} = a + b \cdot T_5 + u_1 + e_{51}$$



Building the model

We translate this in the standard mixed model

$$Y_{ij} = a + b' \cdot T_i + u_j + e_{ij}$$

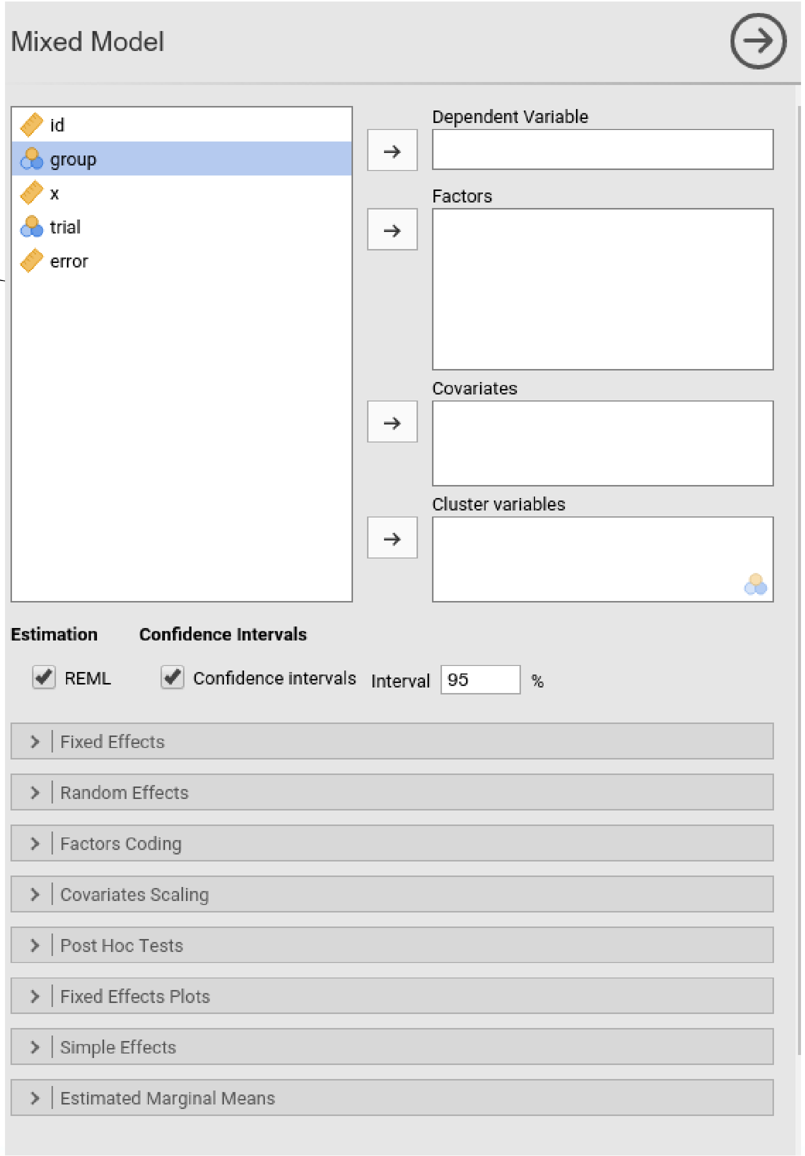
$$y_{ij} = \bar{a} + a_j + \bar{b} \cdot x_{ij} + e_{ij}$$

- Fixed effects? Intercept and trial effect
- Random effects? Intercepts
- Clusters? participants

GAMLj: General mixed models

Variables

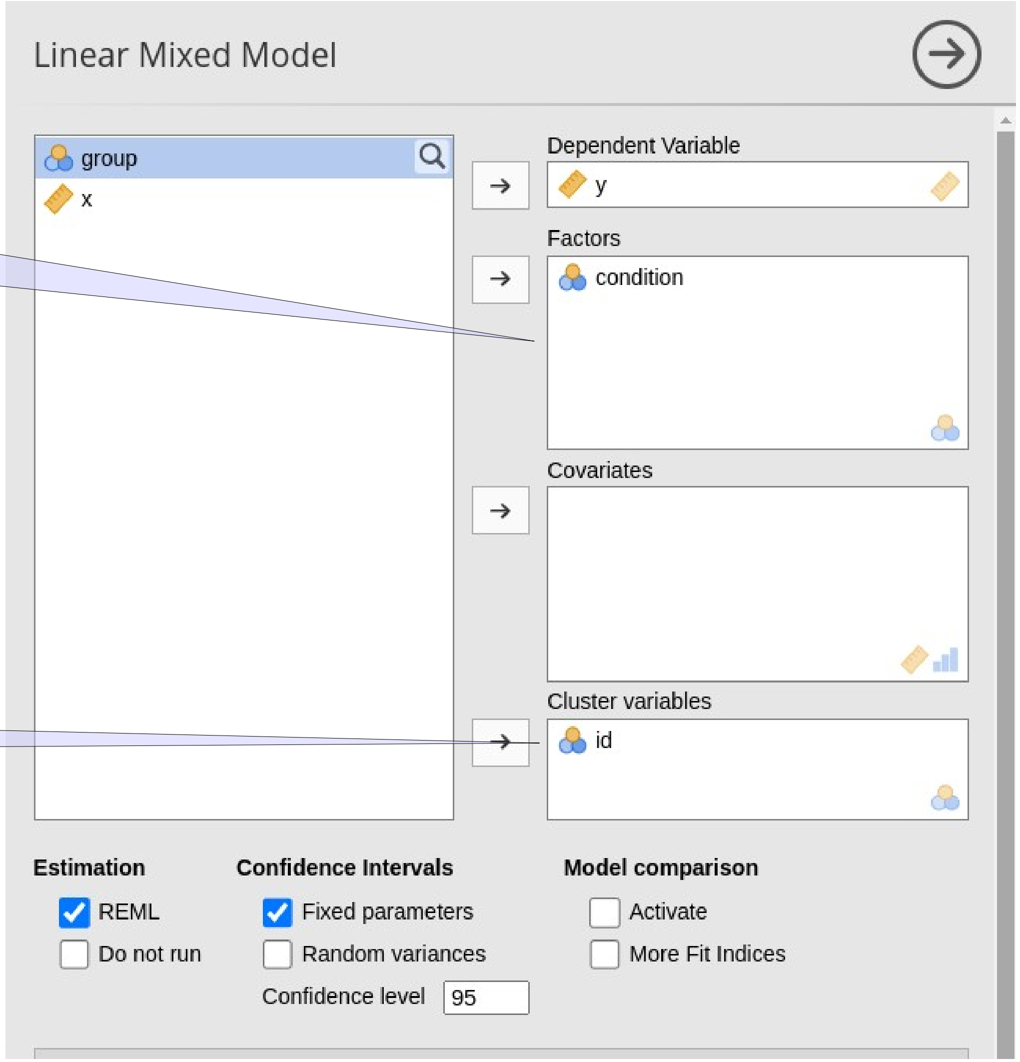
Options



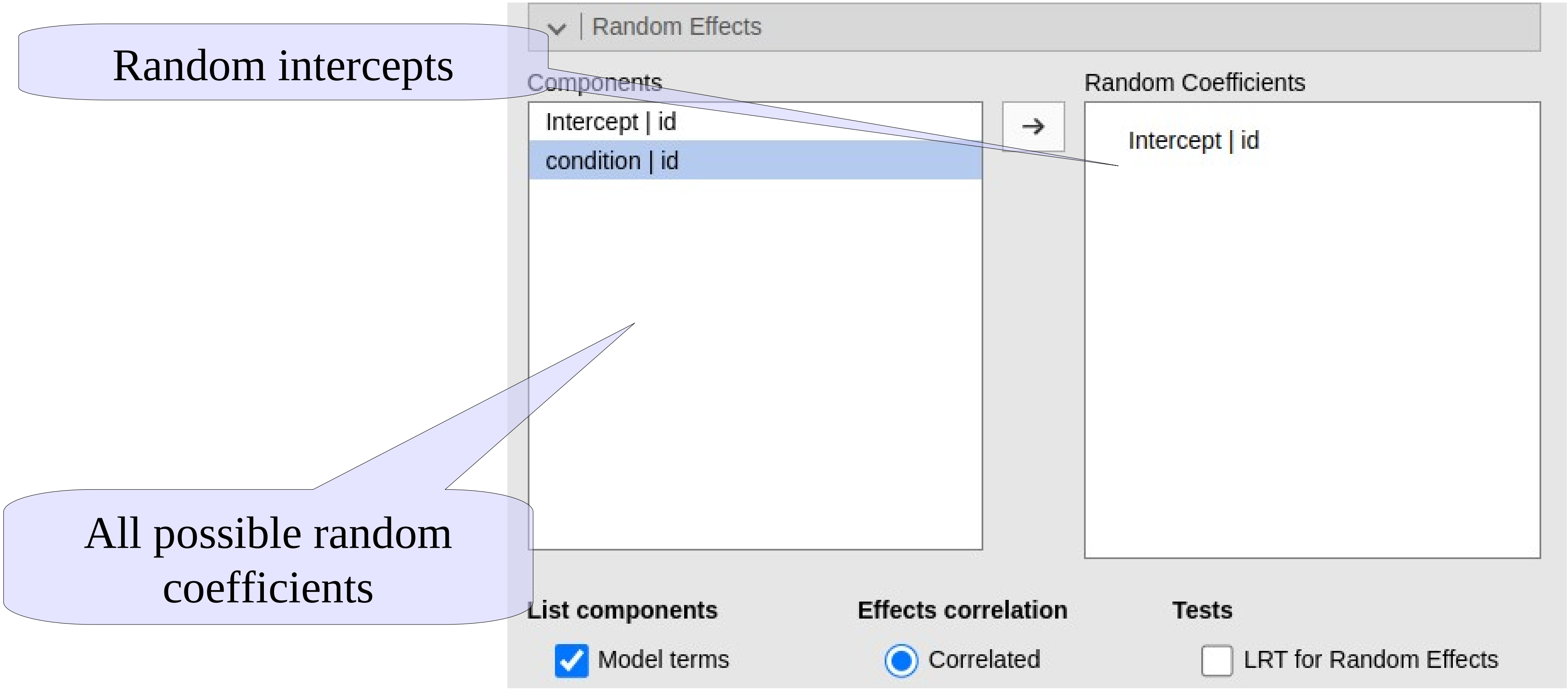
GAMLj: General mixed models

Categorical independent variable

Clustering variable(s)

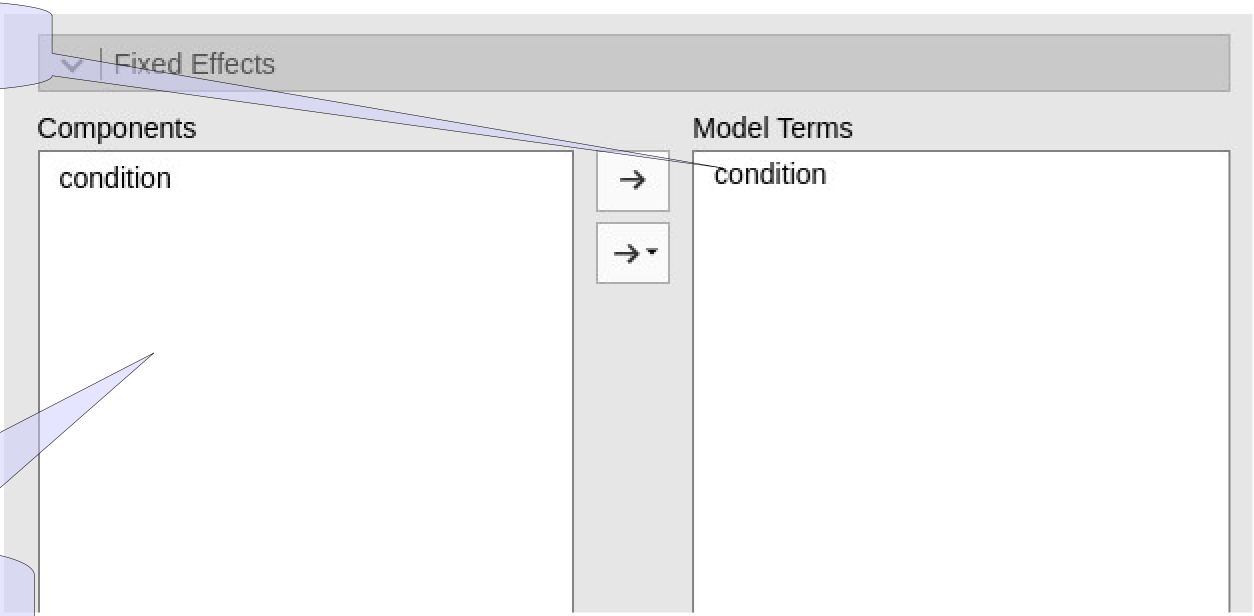


GAMLj: random coefficients



GAMLj: fixed coefficients

Fixed effect



All possible fixed coefficients

GAMLj: Results: model

Model Results

Model Fit

Type	R^2	df	LRT X^2	p
Conditional	0.217	5	78.931	< .001
Marginal	0.015	4	18.773	< .001

[4]

R-squared Conditional: How much variance can the fixed and random effects together explain of the overall variance

R-squared Marginal: How much variance can the fixed effects alone explain of the overall variance

GAMLj: Results: random

Variance of intercepts

Random Components

Groups	Name	Variance	SD	ICC
id	(Intercept)	0.00780	0.0883	0.205
Residual		0.03020	0.1738	

Note. Number of Obs: 1000 , Number of groups: id 200

As long as the variance is non-zero, we are fine

GAMLj: Results: fixed

F-tests

Fixed Effects Omnibus Tests

	F	df	df (res)	p
condition	4.72	4	796	< .001

Coefficients

Parameter Estimates (Fixed coefficients)

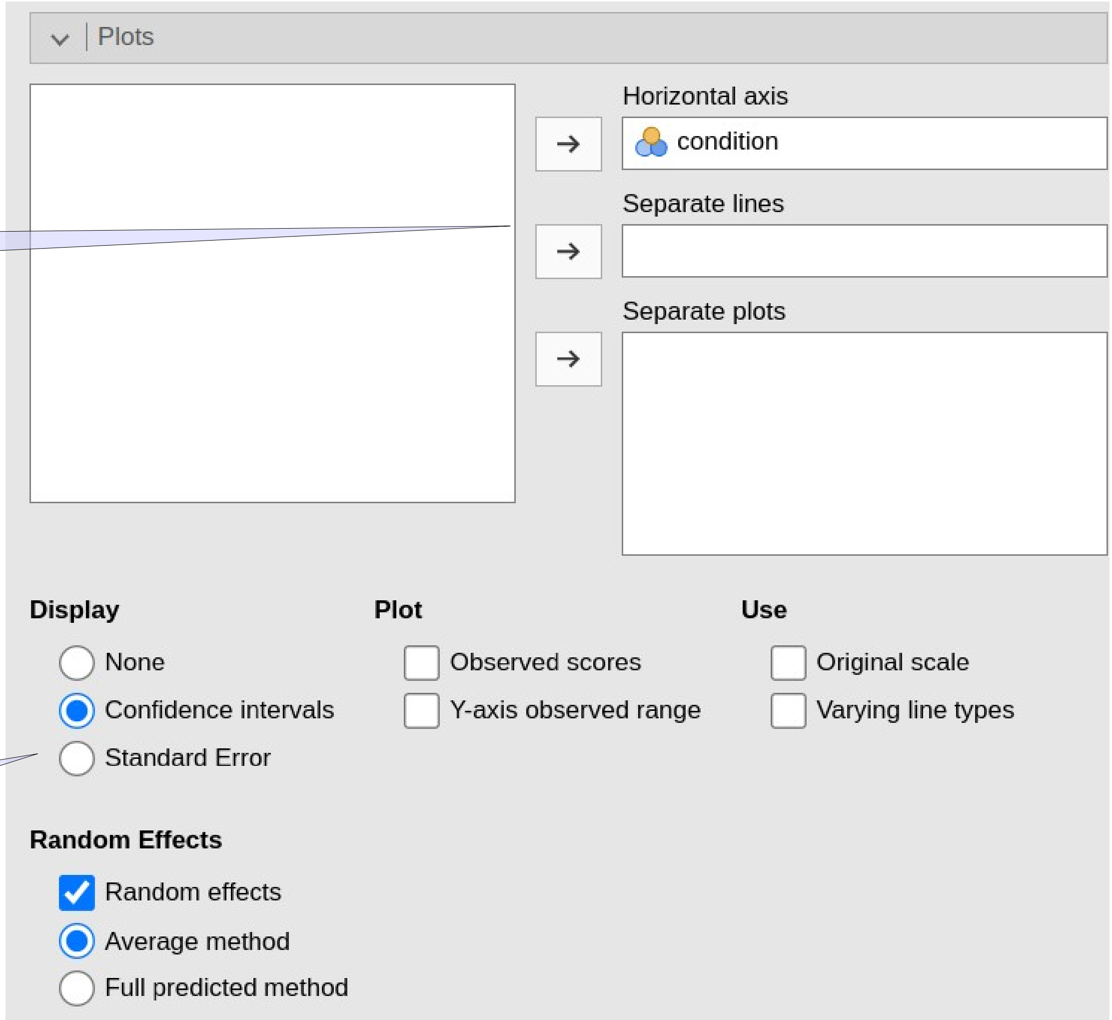
				95% Confidence Intervals				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	0.4947	0.00832	0.47841	0.5111	199	59.462	< .001
condition1	2 - 1	0.0107	0.01738	−0.02344	0.0448	796	0.613	0.540
condition2	3 - 1	0.0278	0.01738	−0.00632	0.0619	796	1.598	0.110
condition3	4 - 1	0.0695	0.01738	0.03541	0.1036	796	4.000	< .001
condition4	5 - 1	0.0349	0.01738	8.03e−4	0.0690	796	2.009	0.045

[5]

Contrasts used to cast the categorical IV

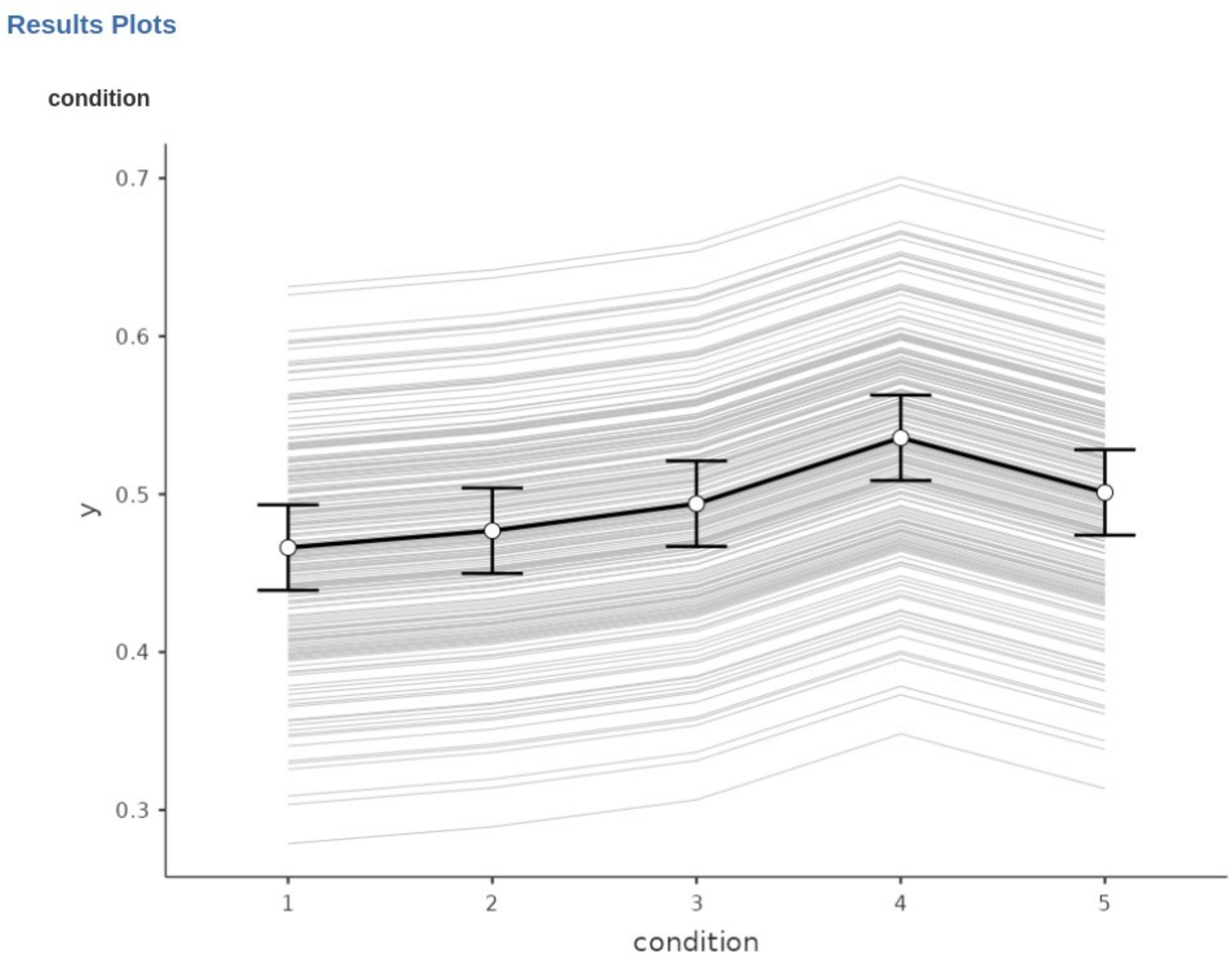
GAMLj: plot

Fixed effects to plot



Options

GAMLj: plot



Between and Repeated Measures Anova

linear mixed model

Standard design

- There are two groups - a Control group and a Treatment group, measured at 4 times. These times are labeled as 1 (pretest), 2 (one month posttest), 3 (3 months follow-up), and 4 (6 months follow-up).

- The dependent variable is a depression score (e.g. Beck Depression Inventory) and the treatment is drug versus no drug. If the drug worked about as well for all subjects the slopes would be comparable and negative across time. For the control group we would expect some subjects to get better on their own and some to stay depressed, which would lead to differences in slope for that group (*)

*) https://www.uvm.edu/~dhowell/StatPages/More_Stuff/Mixed-Models-Repeated/Mixed-Models-for-Repeated-Measures1.html

Standard design

- There are two groups - a Control group and a Treatment group, measured at 4 times. These times are labeled as 1 (pretest), 2 (one month posttest), 3 (3 months follow-up), and 4 (6 months follow-up).

Contingency Tables

Contingency Tables

	group		
time	1	2	Total
0	12	12	24
1	12	12	24
3	12	12	24
6	12	12	24
Total	48	48	96

96 observations
24 subjects

Standard design: data

- Data are in the long format

One participant 4 rows

Data Analyses

Exploration T-Tests ANOVA Regression Frequencies Factor Linear Models

	subj	time	group	dv
1	1	0	1	296
2	1	1	1	175
3	1	3	1	187
4	1	6	1	192
5	2	0	1	376
6	2	1	1	329
7	2	3	1	236
8	2	6	1	76
9	3	0	1	309
10	3	1	1	238
11	3	3	1	150
12	3	6	1	123
13	4	0	1	222
14	4	1	1	60
15	4	3	1	82
16	4	6	1	85
17	5	0	1	150
18	5	1	1	271
19	5	3	1	250

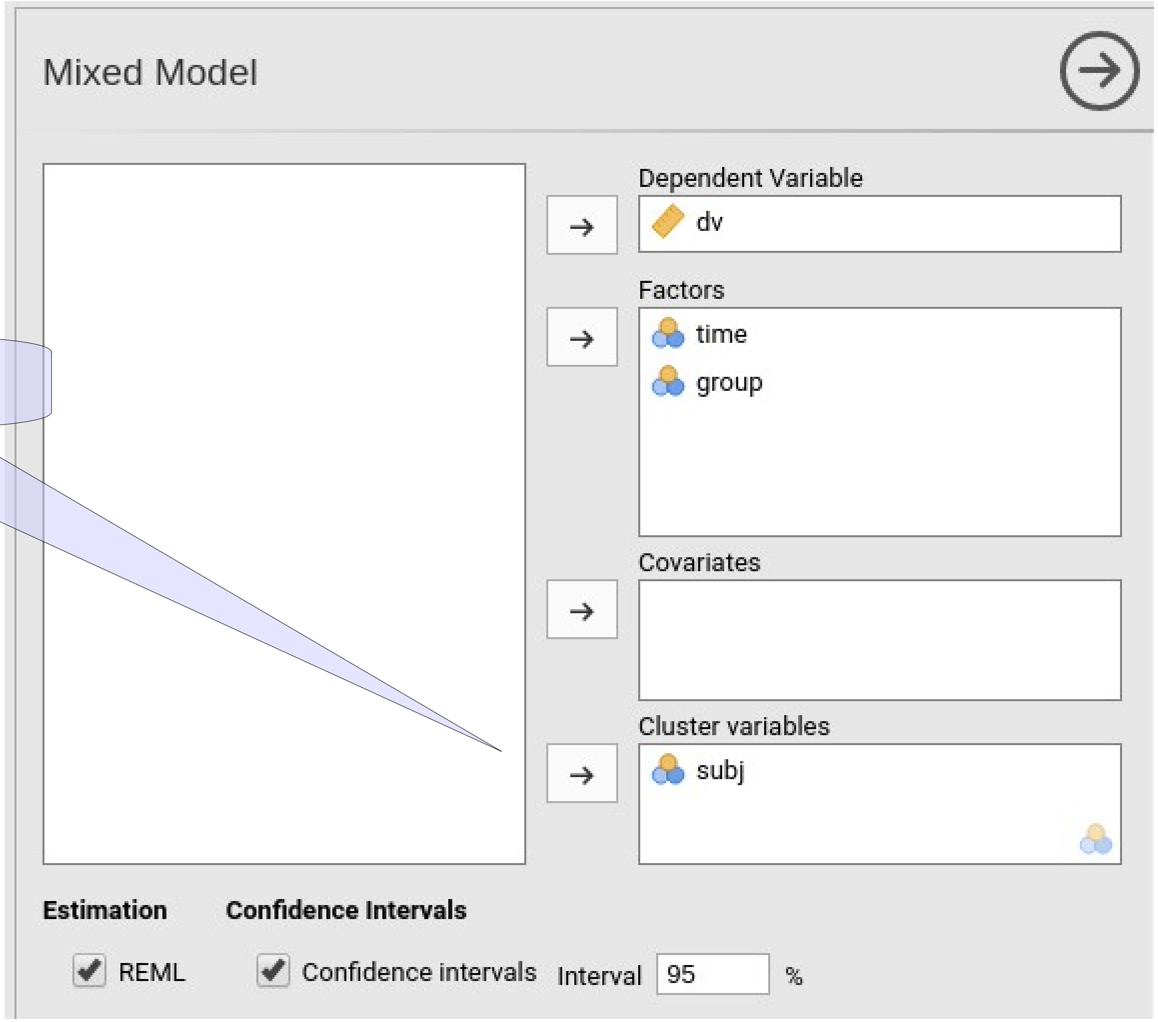
Mixed model

We can translate this in a standard mixed model

- Fixed effects? Intercept and group,time, and interaction effect
- Random effects? Intercepts
- Clusters? subjects

Variables

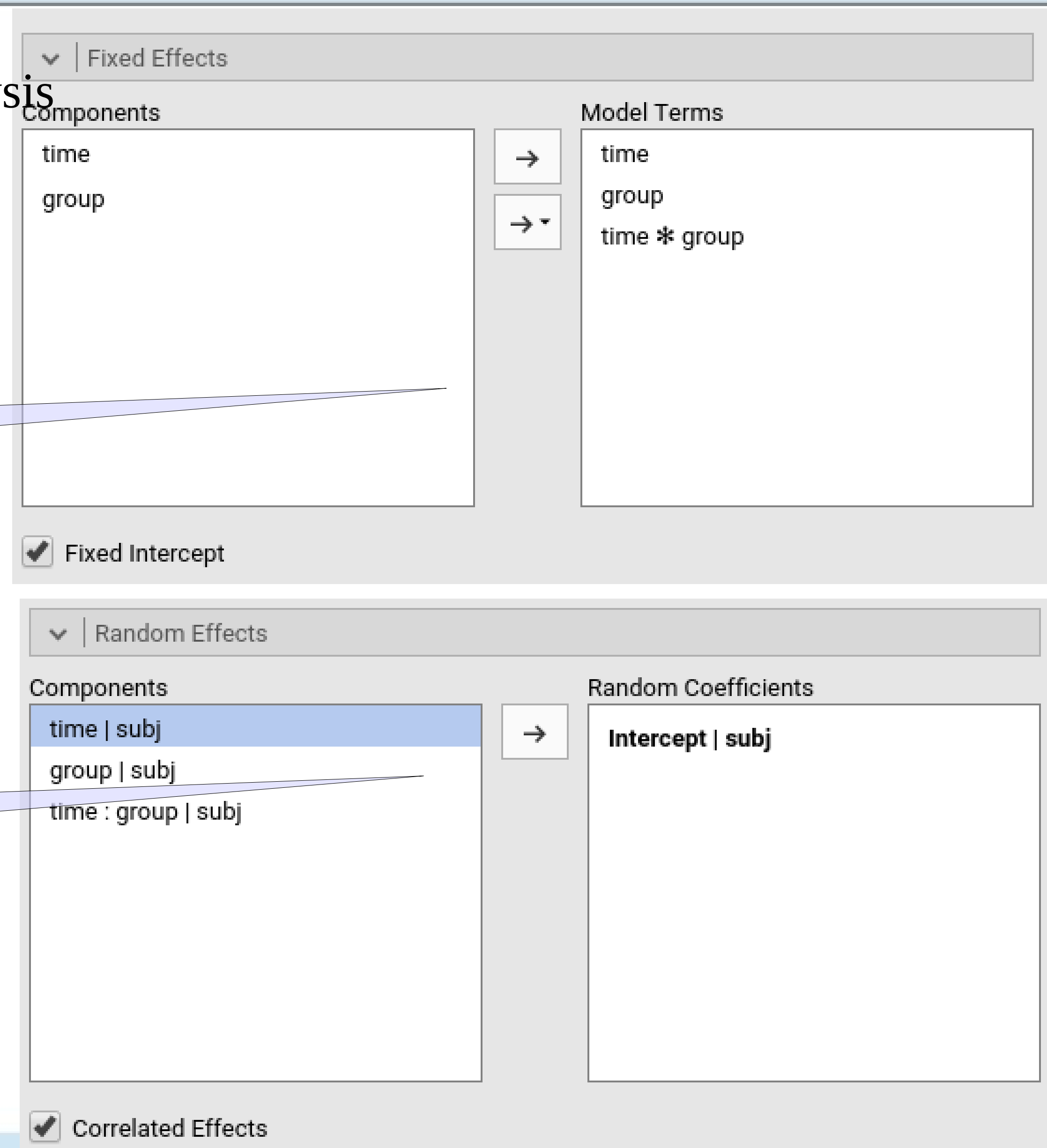
- Definition of the analysis



Clustering variable

Model

- Definition of the analysis



Fixed effects

Random effects

Results

- Interpretation of results

Model fit

Random effects

Model Results

Model Fit

Type	R^2	df	LRT X^2	p
Conditional	0.768	8	107.251	< .001
Marginal	0.554	7	101.043	< .001

[4]

Random Components

Groups	Name	Variance	SD	ICC
subj	(Intercept)	2539	50.4	0.479
Residual		2761	52.5	

Note. Number of Obs: 96 , Number of groups: subj 24

Results

- Interpretation of results

Fixed F-tests

Fixed Effect ANOVA

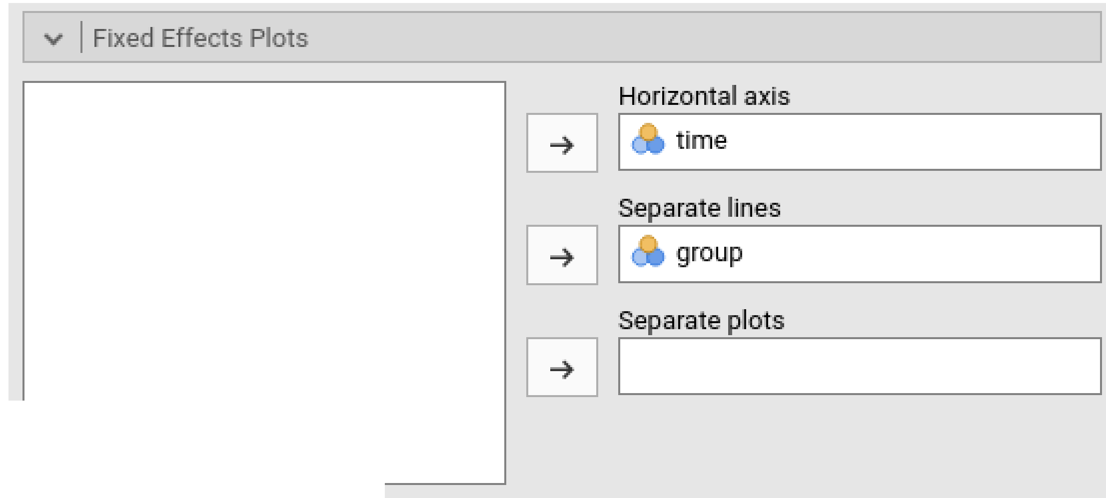
	F	Num df	Den df	p
time	45.14	3	66.0	< .001
group	13.71	1	22.0	0.001
time:group	9.01	3	66.0	< .001

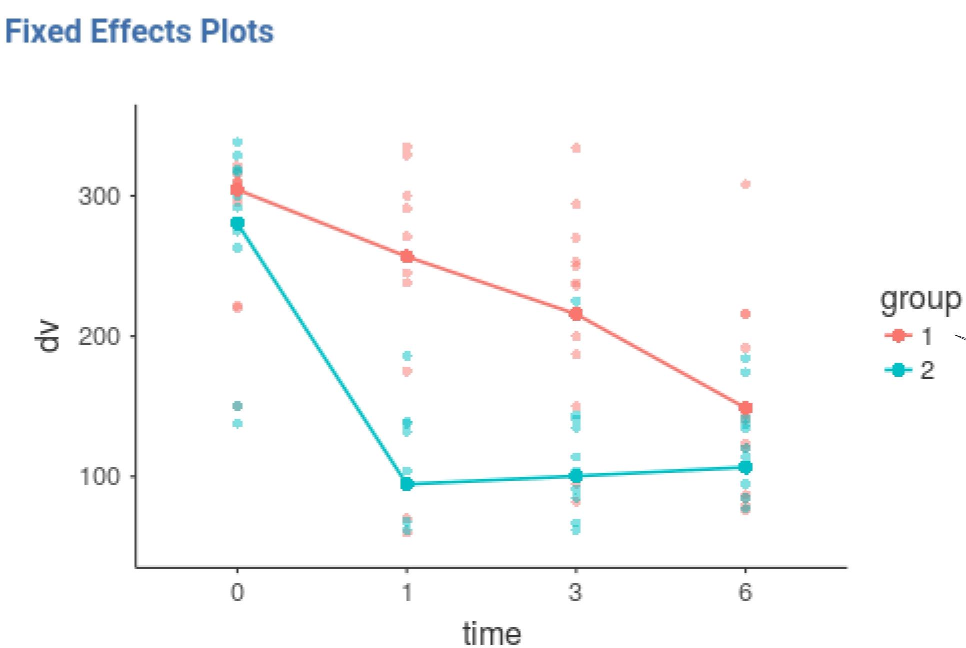
Note. Satterthwaite method for degrees of freedom

- For the moment we ignore the coefficients of the parameter estimates

Results: plot

- Interpretation of results

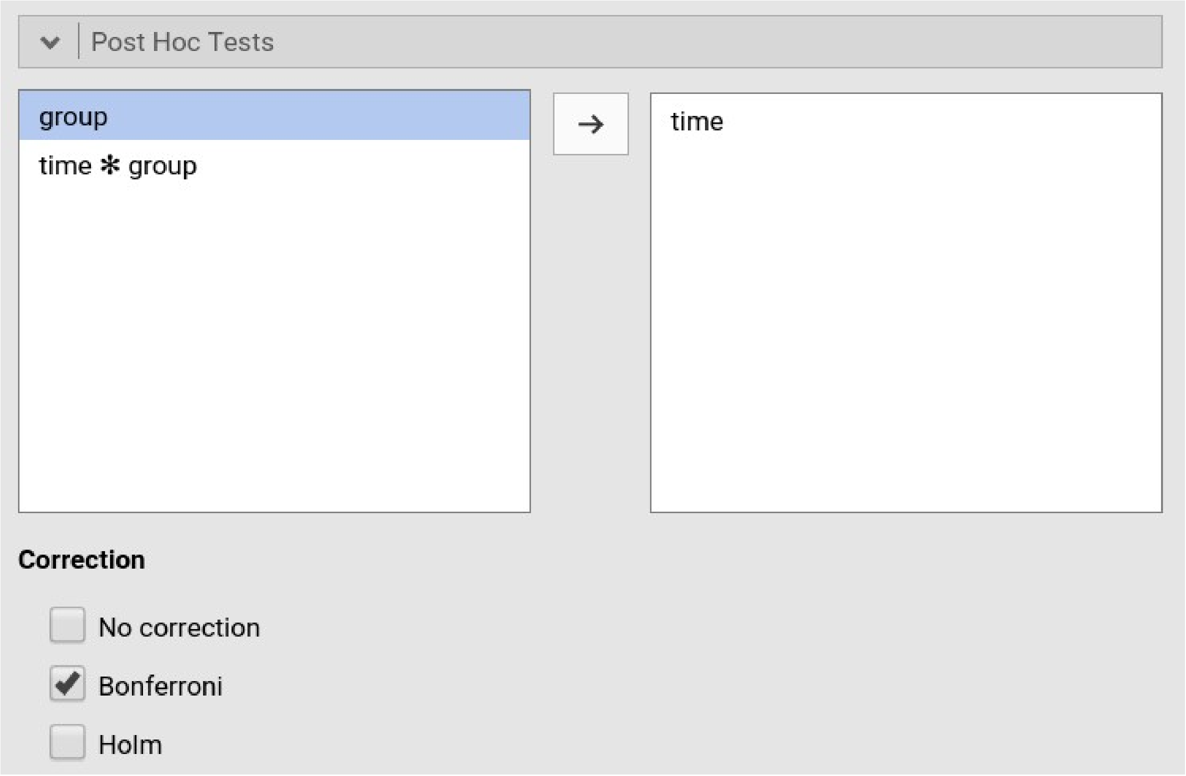




Red is control group

Post-hoc tests

- As for the GLM, post-hoc tests compare all possible pairs of means and correct for inflated Type-I error



Post-hoc tests

- As for the GLM, post-hoc tests compare all possible pairs of means and correct for inflated Type-I error

Post Hoc Tests

Post Hoc Comparisons - time

Comparison							
time		time	Difference	SE	test	df	$p_{bonferroni}$
0	-	1	116.8	15.2	7.70	66.0	< .001
0	-	3	134.3	15.2	8.86	66.0	< .001
0	-	6	164.6	15.2	10.85	66.0	< .001
1	-	3	17.5	15.2	1.16	66.0	1.000
1	-	6	47.8	15.2	3.15	66.0	0.015
3	-	6	30.3	15.2	2.00	66.0	0.300

Generalized Mixed Models

Generalized Linear Models

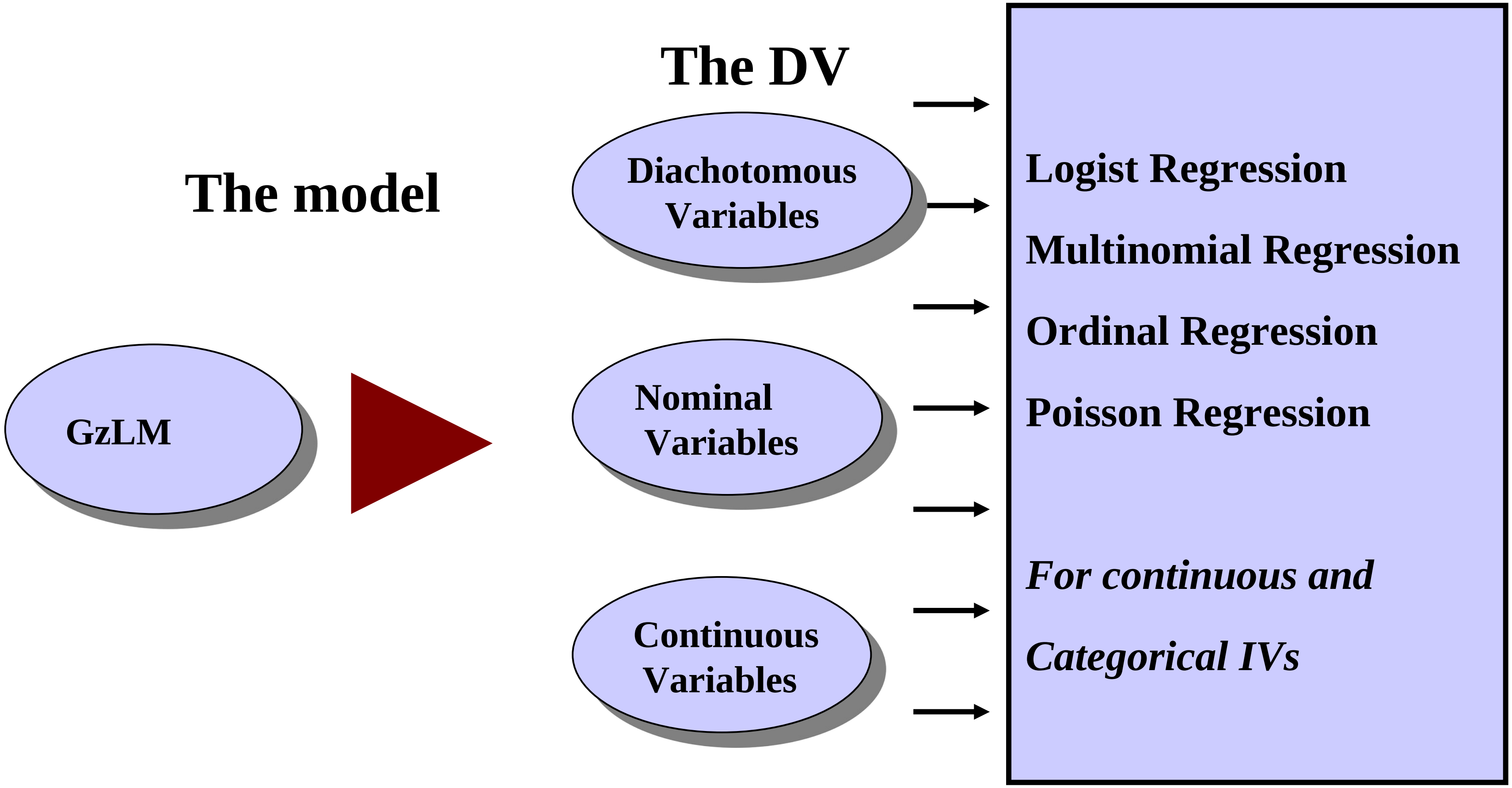
- There are many situations where the dependent variable is not normally distributed:
  - Predicting groups
  - Predicting choices (yes/no, left/right, etc.)
  - Predicting frequencies of behavior

GLM

When the assumptions are NOT met because the dependent variable is not normally distributed (dichotomous, frequencies, categorical etc), we generalize the GLM to the

Generalized Linear Model (GzLM)

Generalized Linear Model



GzLM

- The generalized linear model is a linear model with the dependent variable modeled with a specific function (link function) and with specific error distribution

Generalized Linear Model

$$f\left(y_i\right)=a+b_1\cdot x_{1i}+b_2\cdot x_{2i}+..b_k\cdot x_{ki}+e_i$$

Dependent variable

Specify a distributio n shape

Generalized mixed model

- Applying this logic we obtain a large set of possible statistical techniques

$$f(y_i) = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + .. b_k \cdot x_{ki} + e_i$$

Dependent Variable	**Model**
Continuous	**Linear**
Dichotomous	**Logistic**
Categorical	**Multinomial**
Ordinal	**Ordinal**
Frequencies	**Poisson**

Logic

- When the DV is categorical, we can use the **Generalized Linear model** which allows to apply regression/ANOVA technique to categorical dependent variables

- For multilevel designs (or in general dependent data), we use the **Generalized Linear Mixed model** to allow coefficients to vary randomly across clusters, thus taking dependency into the account

An example: logistic mixed model

- Imagine a study conducted in 70 schools. In each school the same exam is taken by students of equivalent age and grade. For each student, we recorded whether the student passed the exam, **pass**, the student's score in math test, **math**, and the number of extracurricular **activities** the student undertook during the semester.

- The researcher wants to estimate the effect of the math test on the probability of passing the exam, together with the amount of extracurricular activities may moderate the math effect.

- Each school has a different number of students, ranging from 51 to 100. Each student presents three values: the score in the math test, the number of activity undertaken and whether the exam was passed pass=1 or not, pass=0.

Design

- Schools are the clusters

Because we have a dichotomous dependent variable, we need a logistic regression

Because we have clustered data, we need a logistic **mixed** model

Frequencies

Frequencies of pass

Levels	Counts	% of Total	Cumulative %
0	2479	49.2 %	49.2 %
1	2562	50.8 %	100.0 %

Frequencies of school

Levels	Counts	% of Total	Cumulative %
1	95	1.9 %	1.9 %
2	62	1.2 %	3.1 %
3	60	1.2 %	4.3 %
4	56	1.1 %	5.4 %
5	90	1.8 %	7.2 %
6	72	1.4 %	8.6 %
7	82	1.6 %	10.3 %
8	89	1.8 %	12.0 %
9	100	2.0 %	14.0 %
10	59	1.2 %	15.2 %

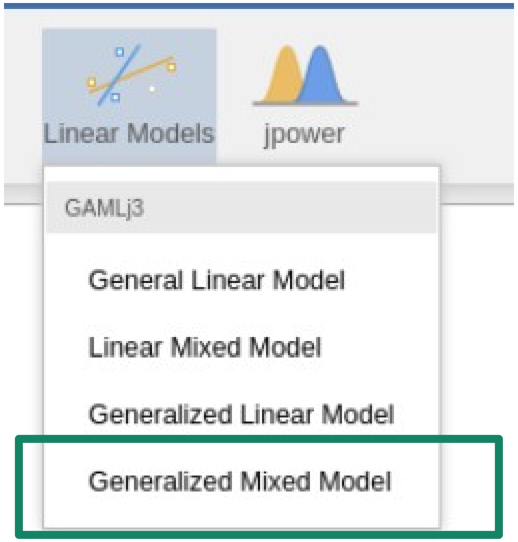
DATA

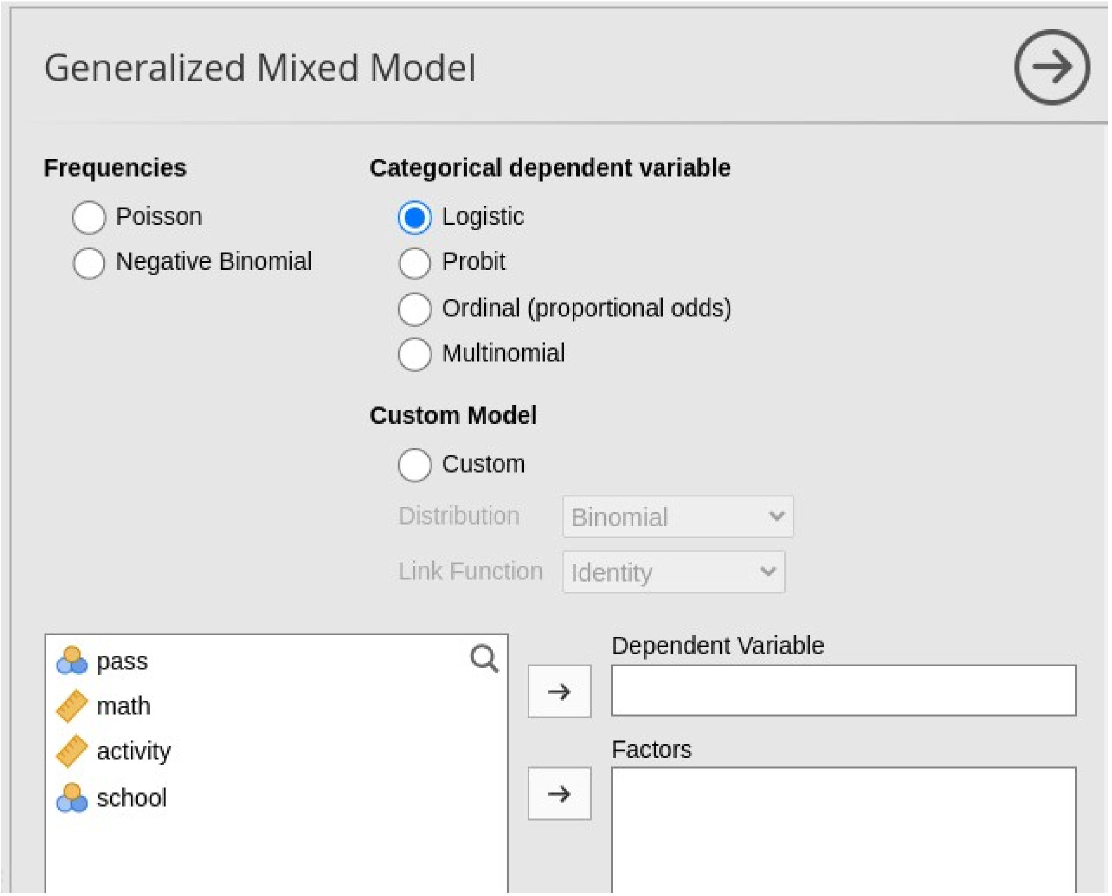
- Data are again in the long format

	pass	math	activity	school
1	1	46.72	3	1
2	1	55.54	5	1
3	1	77.18	6	1
4	1	67.98	4	1
5	0	36.35	5	1
6	1	42.86	2	1
7	1	55.19	3	1
8	1	46.99	3	1
9	1	61.20	2	1
10	1	43.89	3	1
11	1	51.67	3	1
12	1	48.88	3	1
13	1	63.98	4	1
14	1	53.17	2	1
15	1	43.63	4	1
16	1	45.64	2	1
17	1	46.08	5	1
18	1	38.60	3	1
19	0	42.51	5	1
20	1	75.52	1	1
21	1	59.76	3	1
22	1	41.59	3	1
23	1	52.96	3	1

GzLMM

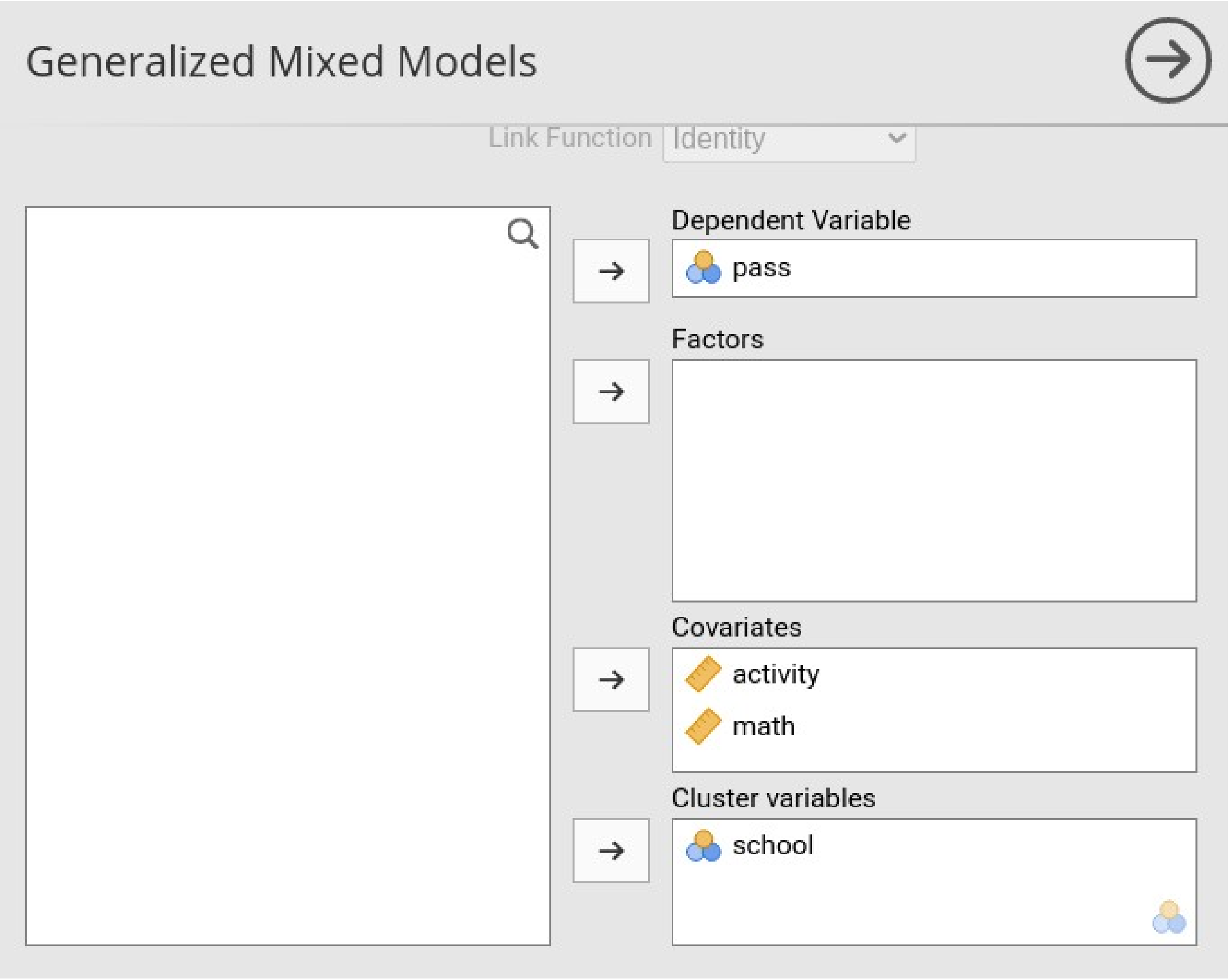
- We launch the module





GzLMM

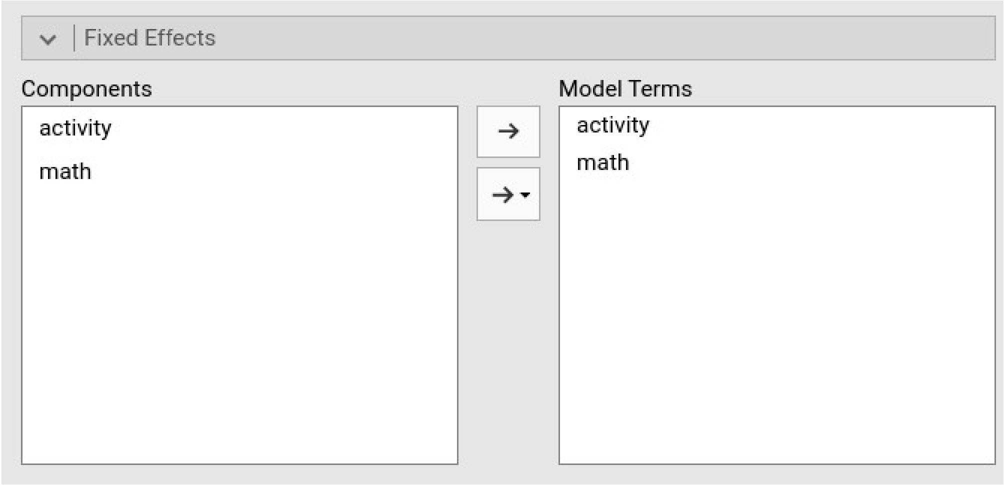
- We select the variables role



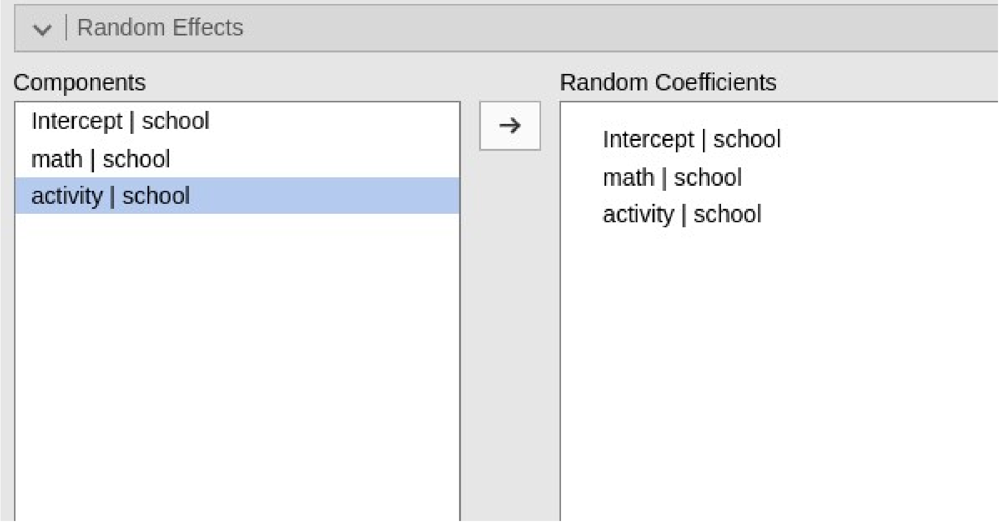
GzLMM

- Define the model parameters

Fixed Effects



Random Effects



Results

- Info table

Direction of the model: What are we predicting?

Model Info

Info		
Model Type	Logistic Model	Model for binary y
Model	lme4::glmer	pass ~ 1 + math + activity + (1 + math + activity | school)
Distribution	Binomial	Dichotomous event distribution of y
Link function	Logit	Log of the odd of y
Direction	P(y=1)/P(y=0)	P(pass = 1) / P(pass = 0)
Sample size	5041	
Converged	yes	
C.I. method	Wald	

[3]

R-squared for the whole model and for the fixed effects

Model Fit

Type	R^2	df	LRT X^2	p
Conditional	0.650	8	2416.726	< .001
Marginal	0.028	2	71.134	< .001

[4]

>

Results

- Random component

Random Components

Groups	Name	Variance	SD	ICC
school	(Intercept)	3.06	1.7486	0.482
	math	1.12e−4	0.0106	
	activity	2.60	1.6133	
Residual		3.29	1.8138	

Note. Number of Obs: 5041 , Number of groups: school 70

Proportion of variance accounted for by the intercepts

Results

- Fixed effects: **Omnibus Tests**

GAMLj uses the Chi-Squared

Fixed Effects Omnibus Tests

	X^2	df	p
math	118.85	1.00	< .001
activity	1.24	1.00	0.266

Results

- Fixed effects: **coefficients**

Here we found the exp(B)

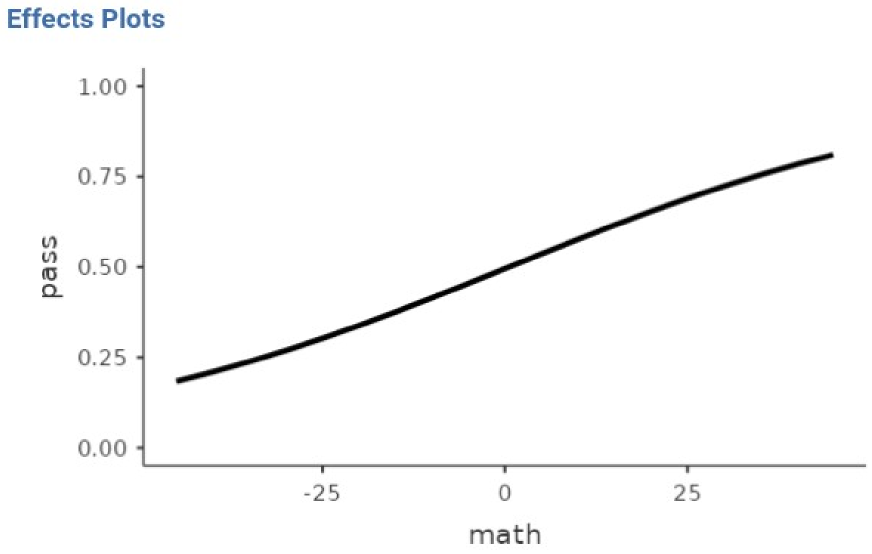
Parameter Estimates (Fixed Coefficients)

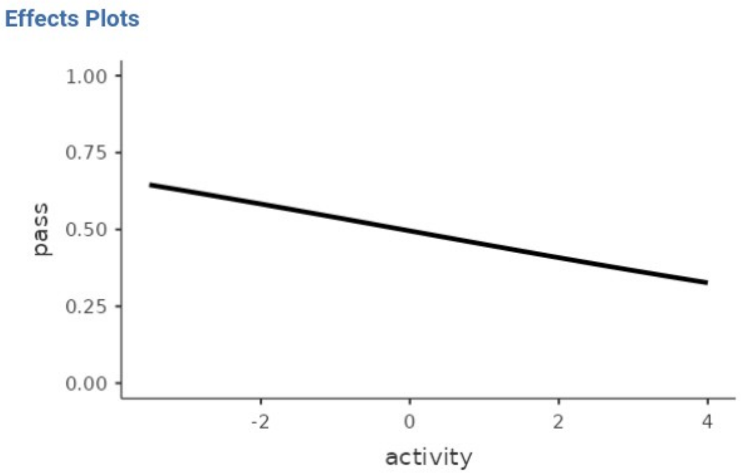
				Exp(B) 95% Confidence Intervals			
Names	Estimate	SE	Exp(B)	Lower	Upper	z	p
(Intercept)	0.0462	0.21435	1.047	0.688	1.59	0.216	0.829
math	0.0464	0.00425	1.047	1.039	1.06	10.902	< .001
activity	−0.2223	0.19982	0.801	0.541	1.18	−1.112	0.266

[5]

Results

- Plot: The probability to pass the exam as a function of the independent variables





Logic

- **General linear model** allows for analyzing a variety of design with normally distributed DV by apply regression/ANOVA tecniques
- For multilevel (or in general clustered data), we use the **Linear Mixed model** to allow coefficients to vary randomly across clusters, thus taking dependency into the account
- When the DV is categorical, we can use the **Generalized Linear model** which allows to apply regression/ANOVA tecniques to categorical dependent variables
- Formultilevel (or in general dependent data), we use the **Generalized Linear Mixed model** to allow coefficients to vary randomly across clusters, thus taking dependency into the account

The end