HSS 8005

Quantitative analysis +

2024 Dr Chris Moreh

Week 6

Hierarchies

Multilevel models

Click and press F for full screen

View on ReCap

A brief review of single-level regression

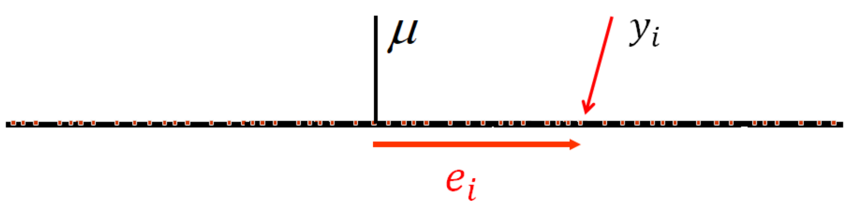
A brief review of single-level regression

Let's start with a very simple example

Observations (n=24)

e.g. 

- We aim to model an **outcome** measurement, our *"estimand"*: Y
- We have data on a number (n) of **observations** (i) (e.g. survey respondents; pupils; students; factory workers; events; etc.): $i_{1...n}$
- We assume that observations are **independent** of each other (e.g. different respondents randomly sampled from a population)
- The outcome measurement has a grand mean across all observations (μ), and each observation (i) has some deviation (*"error"*) from this mean (e_i)
- $$y_i = \mu + e_i$$



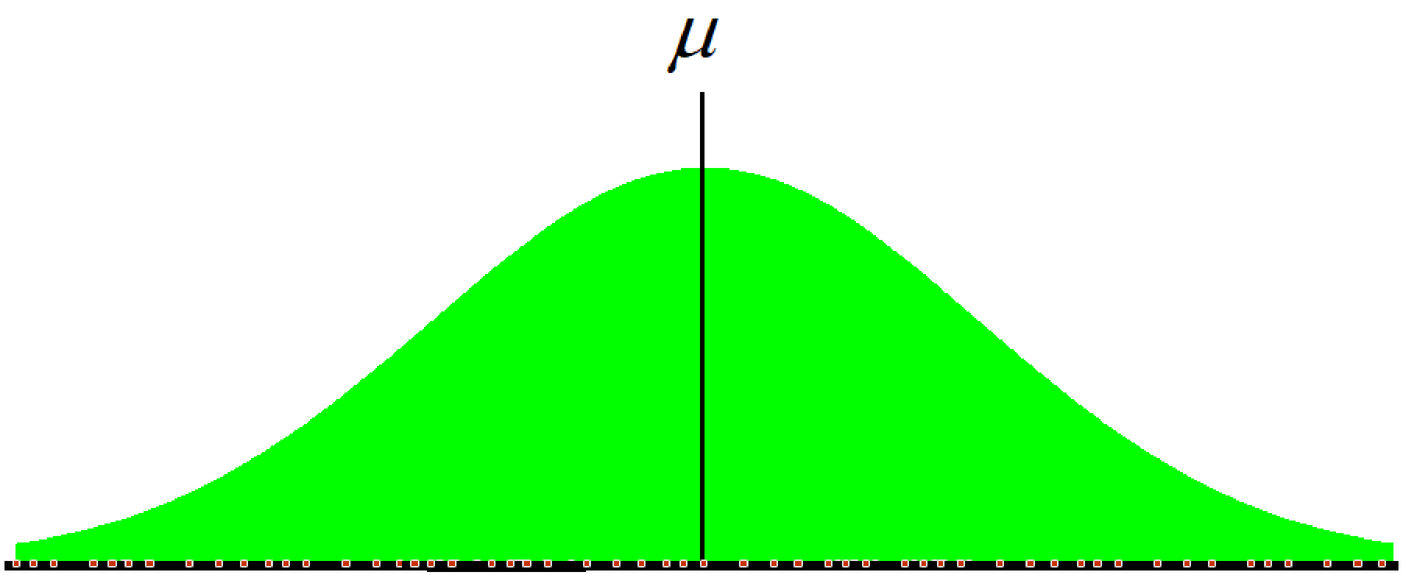
A brief review of single-level regression

$$\underbrace{y_i}_{\textbf{Observed}} = \underbrace{\mu}_{\textbf{Fixed}} + \underbrace{e_i}_{\textbf{Random}}$$

A brief review of single-level regression

- We also assume that the error term (e) is Normally distributed around a mean of 0 and has some variance (σ^2) that we are estimating

$$y_i = \mu + e_i$$



$$e_i \sim N(0, \sigma^2)$$

A brief review of single-level regression

An applied example

Data from the @Osterman2021CanWeTrustEducationFostering article *"Can We Trust Education for Fostering Trust? Quasi-experimental Evidence on the Effect of Education and Tracking on Social Trust"*:

```
# Import the data
osterman <- data_read("https://cgm
```

- cumulative European Social Survey (ESS) data, consisting of nine rounds from 2002 to 2018
- data are weighted using ESS design weights (we will disregard this, so we can expect our results to differ somewhat)
- follows *"the established approach of using a validated three-item scale"* to study generalised social trust

A brief review of single-level regression

An applied example

The outcome variable of interest is an eleven-point scale measure of *"social trust"*:

- The scale consists of the classic trust question, an item on whether people try to be fair, and an item on whether people are helpful:
    - *"Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?"*
    - *"Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair?"*
        - *"Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves?"*
- All of the items may be answered on a scale from 0 to 10 (where 10 represents the highest level of trust) and the scale is calculated as the mean of the three items
- The three-item scale improves measurement reliability and cross-country validity compared to using a single item, such as the classic trust question.

A brief review of single-level regression

An applied example

We’ll select a few variables of interest to keep:

```
osterman <- osterman %>%
  select("trustindex3", "cntry", "
```

And we’ll do some data wrangling; we’ll also reduce the dataset for the purpose of our demonstrations to make it run faster.

```
set.seed(1234)

osterman <- osterman %>%
  labelled::unlabelled() %>% as_ti
  filter(cntry %in% c("GB", "IE",
  group_by(cntry) %>% slice_sample
  mutate(cntry = as_factor(cntry),
         fmnoncntr = ifelse(facntr
  sjlabelled::var_labels( trustind
                          eduyrs25
                          paredu_a
                          fmnoncnt
                        )
```

A brief review of single-level regression

An applied example

Our variables of interest look like this:

	Mean	Std.Dev	Median	Min	Max	N.Valid
agea	52.49	12.89	54.00	25.00	80.00	400
eduyrs25	12.63	4.20	12.00	0.00	24.00	395
facntr	0.96	0.19	1.00	0.00	1.00	400
female	0.54	0.50	1.00	0.00	1.00	400
fmnoncntr	0.05	0.21	0.00	0.00	1.00	400
mocntr	0.97	0.17	1.00	0.00	1.00	400
paredu_a_high	0.32	0.47	0.00	0.00	1.00	379
trustindex3	4.89	1.79	5.00	0.00	9.00	400

A brief review of single-level regression

An applied example

The *“country”* variable

		Valid		Total	
cntry	Freq	%	% Cum.	%	% Cum.
DE	50	12.50	12.50	12.50	12.50
ES	50	12.50	25.00	12.50	25.00
FR	50	12.50	37.50	12.50	37.50
GB	50	12.50	50.00	12.50	50.00
HU	50	12.50	62.50	12.50	62.50
IE	50	12.50	75.00	12.50	75.00
PL	50	12.50	87.50	12.50	87.50
PT	50	12.50	100.00	12.50	100.00
<NA>	0			0.00	100.00
Total	400	100.00	100.00	100.00	100.00

A brief review of single-level regression

An applied example

Let's start by fitting a single-level model of social trust as a function of education, age, gender, parental education and whether either of the parents were born abroad (i.e. the variable we computed earlier).

Mathematically, we fit the following model:

$$trustindex3 = \beta_0 + \beta_1 * eduyears25 + \beta_2 * agea + \beta_3 * female + \\ + \beta_4 * paredu + \beta_5 * fmnoncntr + error$$

A brief review of single-level regression

An applied example

Model summary:

Observations	376 (24 missing obs. deleted)
Dependent variable	trustindex3
Type	OLS linear regression

F(5,370)	7.593
R^2	0.093
Adj. R^2	0.081

	Est.	2.5%	97.5%	t val.	p
(Intercept)	2.551	1.454	3.648	4.572	0.000
eduyrs25	0.110	0.062	0.157	4.537	0.000
agea	0.020	0.005	0.034	2.656	0.008
female	-0.204	-0.559	0.151	-1.131	0.259
paredu_a_high	0.229	-0.185	0.643	1.088	0.277
fmnoncntr	-0.954	-1.833	-0.075	-2.134	0.033

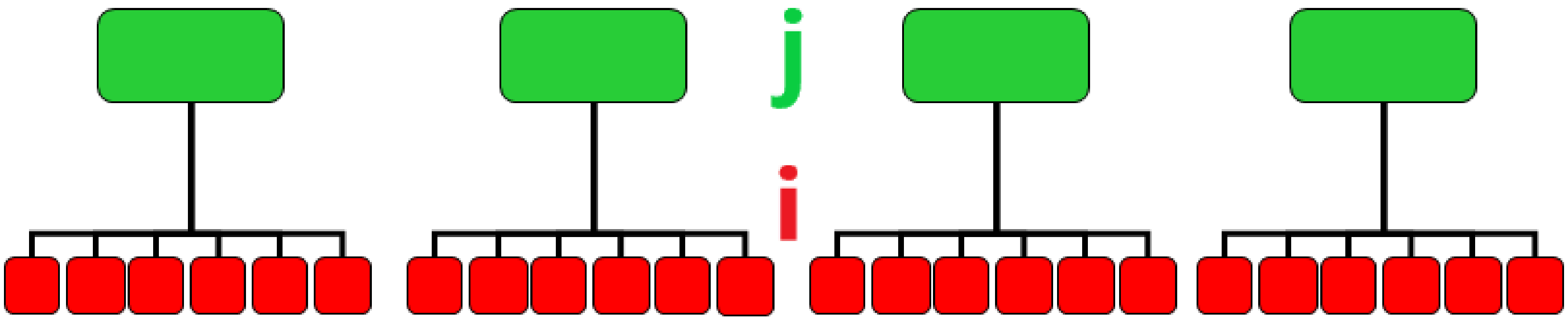
Standard errors: OLS

We have interpreted this model in earlier weeks. Our interest now is in extending it to account for the nesting of cases within different countries.

Multilevel models

Multilevel models

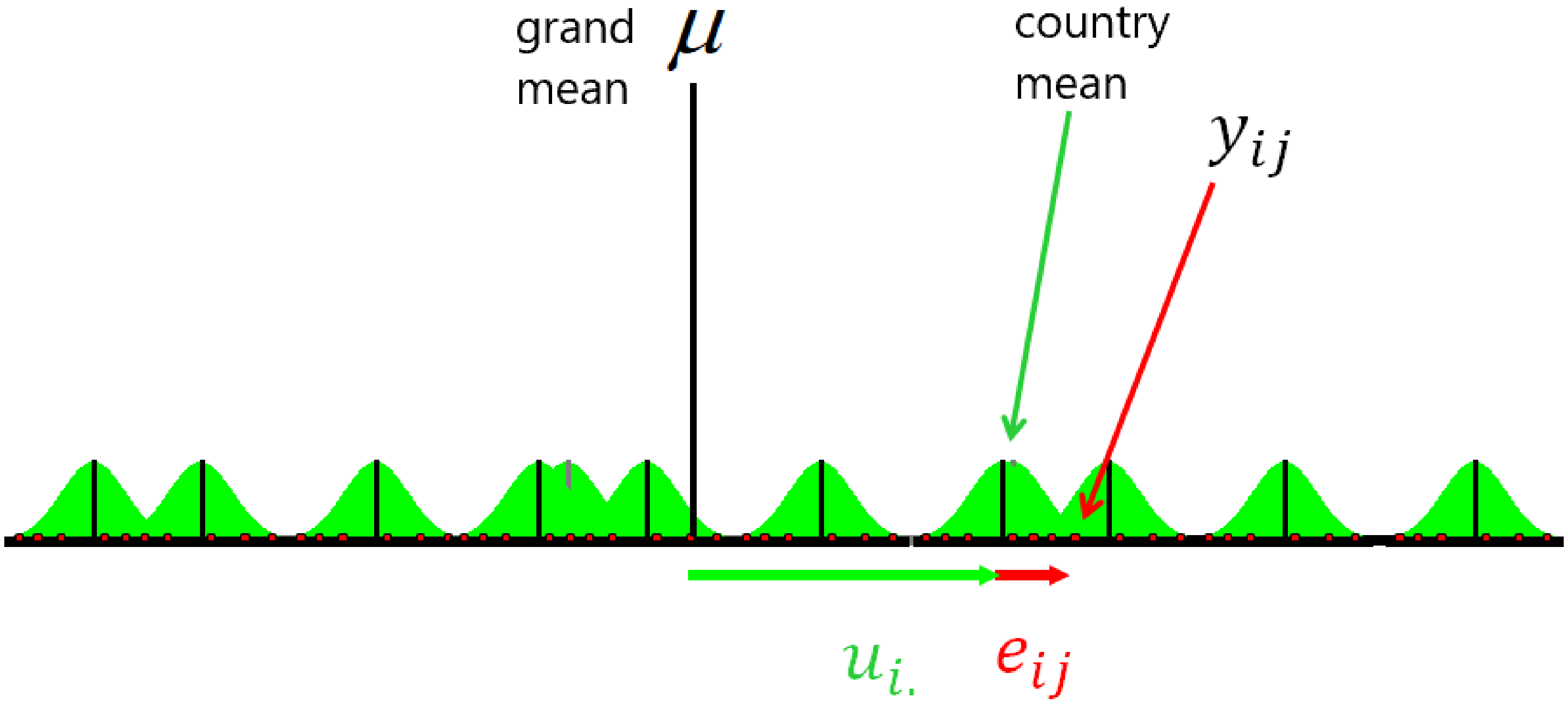
In our dataset we cannot assume that the observations are fully independent (or that the errors are independently distributed). We know that observations were sampled from within selected countries, so the countries are cluster variables that may have a group-level influence on the behaviour, opinions, conditions etc. of our individual observations.



$$y_{ij} = y_{\text{respondent, country}}$$

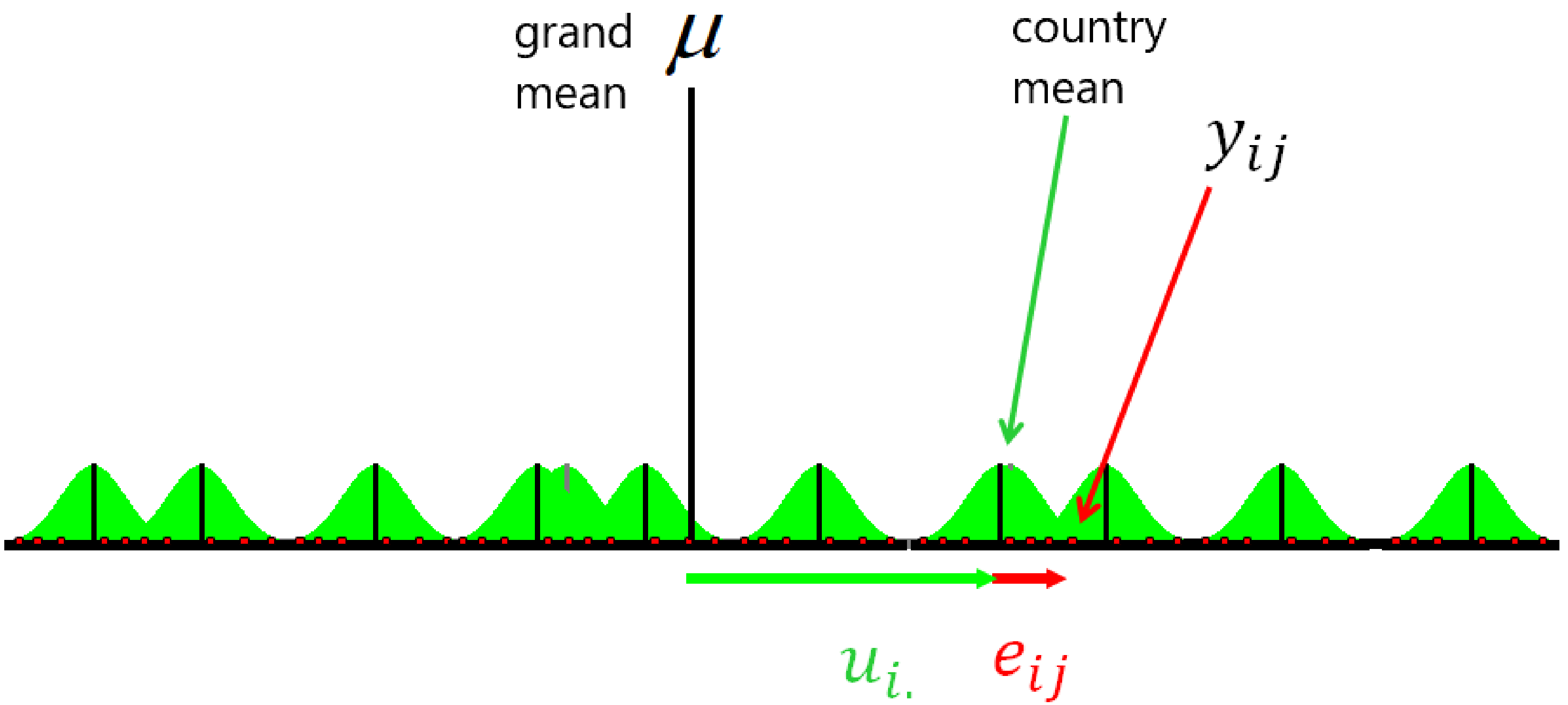
Multilevel models

In our dataset we cannot assume that the observations are fully independent (or that the errors are independently distributed). We know that observations were sampled from within selected countries, so the countries are cluster variables that may have a group-level influence on the behaviour, opinions, conditions etc. of our individual observations.



Multilevel models

In our dataset we cannot assume that the observations are fully independent (or that the errors are independently distributed). We know that observations were sampled from within selected countries, so the countries are cluster variables that may have a group-level influence on the behaviour, opinions, conditions etc. of our individual observations.



$$y_{ij} = \mu + u_{i.} + e_{ij}$$

Multilevel models

In our dataset we cannot assume that the observations are fully independent (or that the errors are independently distributed). We know that observations were sampled from within selected countries, so the countries are cluster variables that may have a group-level influence on the behaviour, opinions, conditions etc. of our individual observations.

$$\underbrace{y_{ij}}_{\textbf{Observed}} = \underbrace{\mu}_{\textbf{Fixed}} + \underbrace{u_{i.} + e_{ij}}_{\textbf{Random}}$$

$$u_{i.} \sim N(0, \tau^2)$$
$$e_{ij} \sim N(0, \sigma^2)$$

Multilevel models

http://mfviz.com/hierarchical-models/

Multilevel models

In our dataset we cannot assume that the observations are fully independent (or that the errors are independently distributed). We know that observations were sampled from within selected countries, so the countries are cluster variables that may have a group-level influence on the behaviour, opinions, conditions etc. of our individual observations.

With such data, it makes sense to allow regression coefficients to vary by group.

Such variation can already be achieved by simply including group indicators in a least squares regression framework.

In other words, we could extend our previous model like such:

$$trustindex3 = \beta_0 + \beta_1 * eduyears25 + \beta_2 * agea + \beta_3 * female + \\ + \beta_4 * paredu + \beta_5 * fmnoncntr + \beta_6 * cntry + error$$

Multilevel models

Very often, simply including group indicators in a least squares regression gives unacceptably noisy estimates.

Instead, we use *“multilevel regression”*, a method of partially pooling varying coefficients, equivalent to Bayesian regression where the variation in the data is used to estimate prior distribution on the variation of intercepts and slopes.

The terminology surrounding multilevel models can be confusing. Different disciplines use various names for them, for example:

- Variance components
- Random intercepts and slopes
- Random effects
- Random coefficients
- Varying coefficients
- Intercepts- and/or slopes-as-outcomes
- Hierarchical linear models
- Multilevel models (implies multiple levels of hierarchically clustered data)
- Growth curve models (possibly Latent GCM)
- Mixed effects models

Multilevel models

Some of these terms might be more historical, others are more often seen in a specific discipline, others might refer to a certain data structure, and still others are special cases (e.g. *“null”* models with no explanatory variables).

Though you will hear many definitions, **random effects** are simply those specific to an observational unit, however defined. In our examples We will mostly encounter the case where the observational unit is the level of some grouping factor, but this is only one possibility.

Mixed effects - or simply *“mixed”* - models generally refer to a mixture of fixed and random effects. This is probably the most general term, with no specific data structure implied.

Fitting multilevel models

In `R` we can fit multilevel models using the `lmer` function from the ***lme4*** package.

Initially, it is advisable to first fit some simple, preliminary models, in part to establish a baseline for evaluating larger models. Then, we can build toward a final model for description and inference by attempting to add important covariates, centering certain variables, and checking model assumptions.

The standard first step is to model only the outcome measurement, without any predictors, to get a sense for the effect of the clusters; this is often called a ***random intercepts model*** or ***null model***:

```
mod_null <- lmer(trustindex3 ~ 1 +
```

The second step is then to fit the full covariate model:

```
mod_mixed = lmer(trustindex3 ~ edu
```

Fitting multilevel models

Results: “*null*” model:

Observations	400
Dependent variable	trustindex3
Type	Mixed effects linear regression

AIC	1590.16
BIC	1602.13
Pseudo-R^2 (fixed effects)	0.00
Pseudo-R^2 (total)	0.09

Fixed Effects

	Est.	S.E.	t val.	d.f.	p
(Intercept)	4.89	0.21	23.39	7.00	0.00

p values calculated using Satterthwaite d.f.

Random Effects

Group	Parameter	Std. Dev.
cntry	(Intercept)	0.54
Residual		1.72

Grouping Variables

Group	# groups	ICC
cntry	8	0.09

The intra-class correlation (ICC) tells us the percentage of variation in the outcome variable attributable to differences between countries.

Fitting multilevel models

Results: Covariate model:

Observations	376
Dependent variable	trustindex3
Type	Mixed effects linear regression

AIC	1500.20
BIC	1531.64
Pseudo-R² (fixed effects)	0.08
Pseudo-R² (total)	0.15

Fixed Effects

	Est.	S.E.	t val.	d.f.	p
(Intercept)	2.52	0.62	4.03	184.00	0.00
eduyrs25	0.10	0.02	3.86	364.33	0.00
agea	0.02	0.01	2.81	324.25	0.01
female	-0.14	0.18	-0.77	368.19	0.44
paredu_a_high	0.23	0.21	1.09	369.34	0.28
fmnoncntr	-0.81	0.44	-1.85	368.30	0.06

p values calculated using Satterthwaite d.f.

Random Effects

Group	Parameter	Std. Dev.
cntry	(Intercept)	0.49
Residual		1.69

Grouping Variables

Group	# groups	ICC