Problems and opportunities: when observations are nested

Silje Synnøve Lyder Hermansen

2026-02-20

Where are we in the course?

Where are we in the course?

We are entering the core of this course

1. R-skills and regression recap (week 1-3)
2. Data structures (week 5-6, 14)
3. Limited and categorical outcome variables (GLMs) (week 7-13)

Recap: R-skills

Recap: R-skills

Our work flow until now

1. **R-skills** and regression recap (week 1-3)
2. Data structures (week 5-6, 14)
3. Limited and categorical outcome variables (GLMs) (week 7-13)

Recap of the last three weeks

I've introduced new concepts in class, you've honed them at home

week 1

- in class: core concepts in R: objects, functions, syntax, subsetting (guessing game + indexation)
- at home: build knowledge of the base R language, workflow

week 2

- in class: two new dialects (ggplot2, tidyverse)
- at home: more base R + new vocabulary

week 3

- in class
  - little new vocabulary, but new applications of it
  - core modeling concepts:
    - equations are expressions of a theory
    - prediction for interpretation
- at home: hone these skills

Where are we going?

Two core assumptions in ordinary regression

Linear models (OLS) rely on two overarching assumptions that are often violated.

1. **Assumption 1:** outcomes (y) conditional on the predictors (x) are normally distributed (week 6-13)
2. **Assumption 2:** observations are independent and identically distributed (iid) (week 4-5, 14)

$\Rightarrow$ *this course looks at strategies for when these are not satisfied*

Core assumption 1: outcomes (y) conditional on the predictors (x) are normally distributed

- problem: limited and categorical outcome variables are not continuous
- solution:
  - recode the dependent variable and describe the data generating process w/probability distribution
  - choice of model depends on the data generating process - e.g. logit, multinomial, ordinal, poisson, neg.bin, zero-inflated, coxph...

$\Rightarrow$ *a topic for later*

Assumption 2: Observations are not iid:

- problem: observations do not have equal probability of arriving in the sample
- solution:
  - a mindful strategy for how to leverage variation: hierarchical/nested data
  - strategies when our sample does not reflect the population: missing data

$\Rightarrow$ *today: what do we do when observations are not iid?*

We are entering the core of this course

1. R-skills (week 1-3)
2. **Data structures (when observations are not iid)** (week 5-6, 14)
3. Limited and categorical outcome variables (GLMs) (week 7-13)

The purpose of this course

$\Rightarrow$ *Take 1 (negative): find solutions when the assumptions of the linear model are not satisfied*

$\Rightarrow$ *Take 2 (positive): pick models that are tailored to the data generating process*

Negative take: Three assumptions of the linear model

Our example: MEPs' local staff size

Our example: MEPs' local staff size

Let's express a theory that MEPs hire local staff to offset electoral disadvantages.

$$y_i = a + bx_i$$

- y: number of local assistants (`LocalAssistants`)
- x: national party's seat share in national parliament (`SeatsNatPal.prop`)
- unit of observation: MEPs observed every 6th month (`MEP.rda`)
- Hypothesis: $b < 0$

Interpreting: setting a scenario using descriptive statistics

Use descriptive statistics to find a reasonable partial scenario for interpretation

```r
#Summarize the results
summary(df$x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00000 0.08511 0.25714 0.24601 0.39692 0.67876     195
```

```r
#Calculate the inter-quartile range (25th to 75th percentile)
IQR(df$x, na.rm = T)
```

```
## [1] 0.3118167
```

- The party with the lowest support got less than 1% of the votes, while the party with the strongest support received 1%.
- The inter-quartile range gives the difference between typical small vs typical large parties.

Interpreting: Applying the scenario for substantive effect

Here, the marginal effect and first difference is the same (all effects are linear).

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.561 -1.561 -0.519  0.652 40.470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.56125    0.05743  44.596   <2e-16 ***
## x           -0.44420    0.18926  -2.347    0.019 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.803 on 6946 degrees of freedom
##   (195 observations deleted due to missingness)
## Multiple R-squared:  0.0007924,  Adjusted R-squared:  0.0006486
## F-statistic: 5.509 on 1 and 6946 DF,  p-value: 0.01895
```

- The predicted difference in staff size between the two is 0.1 employees (-0.44 * 0.3)

⇒ *how valid are these results (any omitted variable bias?)*

Linear models are BLUE

"Best Linear Unbiased Estimators" (BLUE) makes sure that the parameters (regression coefficients) and standard errors describe the mean and spread in a normal distribution.

- Unbiased: residuals sum up to 0. The model is "on average right"
- Efficient: several combinations of parameters could be possible; the model picks the ones that generate the fewest errors (least spread).

Three assumptions of the linear model

The traditional way of assessing the linear model, is to check the residuals

1. residuals are normally distributed (unique to the OLS)
2. residuals are equally distributed over the range of y (homoscedasticity) (unique to the OLS)
3. residuals are not correlated with x (no omitted variable bias) (common for all regressions)

What are residuals?

Residuals are the difference between what we observed and expected (predicted)

$$y_i = a + bx_i + \epsilon_i$$

```
df <-
  df %>%
  mutate(
    #Predicted values
    predicted = predict(mod, df),
    #Difference between predicted and observed
    residuals = y - predicted,
    #Standardized spread is measured as standard deviations
    residuals_s = residuals/sd(residuals, na.rm = T)
  )
```

- We often standardize them by dividing them by their own standard deviation.

Assumption 1: Residuals are normally distributed

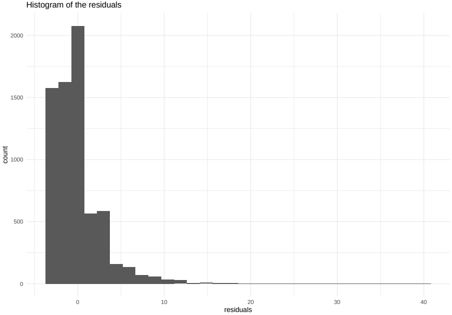
Assumption 1: Residuals are normally distributed

Normally distributed errors allow you to do hypotheses tests

- limitations to the limitation:
  - categorical predictors: parameters are group averages
  - many predictors: the model ends up with normal errors
  - self-restraint in the interpretation: use scenarios that actually exist in the data

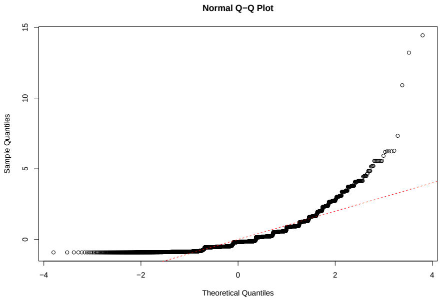
$\Rightarrow$ *mostly important in small samples; least important overall*

Distribution of my residuals



- histograms give a first impression

Compare with a standard normal distribution



- another way is to compare the standardized residuals to a standard normal distribution
- a perfect correlation would follow the diagonal; here, we see the tails are off

$\Rightarrow$ *normality is not strictly necessary for OLS to be unbiased; only for hypothesis testing and confidence intervals in small samples.*

Assumption 2: Residuals are homoskedastic

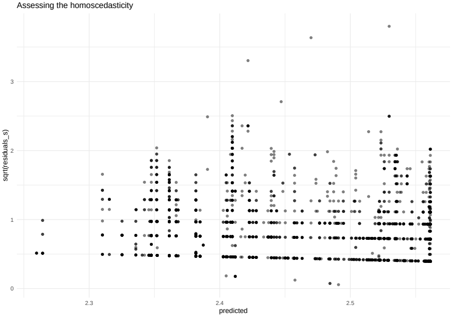
Assumption 2: Residuals are homoskedastic

The residuals have an equal spread over the entire range of xs (i.e. your predicted y)

- are the **standard errors** correct
  - if not, they will be too high in some range, and too low elsewhere
  - does not relate to the **parameter**
- potential fix for heteroskedastic errors:
  - robust standard errors
  - more control variables
  - varying intercept model
  - GLMs

$\Rightarrow$ *If violated, you'll be over-confident in your results*

Spread of my residuals



- we can plot the residuals against the predicted y; there should be no "fan"
- there's a bit of that going on here (bigger spread on high predicted values)

⇒ *estimation is unbiased (regression coefficients are correct), but inefficient (standard errors might be wrong).*

Early warning

a violation is often an early warning that the third assumption is violated as well

Assumption 3: Residuals are not correlated with x

Assumption 3: Residuals are not correlated with x

Residuals contain all the variation in y that could be explained by other covariates that are *not* currently in your model

A correlation is a sign of:

- misspesification of the y ~ x relationship (might actually be non-linear)
- omitted variable bias (spurious relationship/open backdoors): when z (omitted) causes both x and y.

Correlation between x and residuals: in numbers

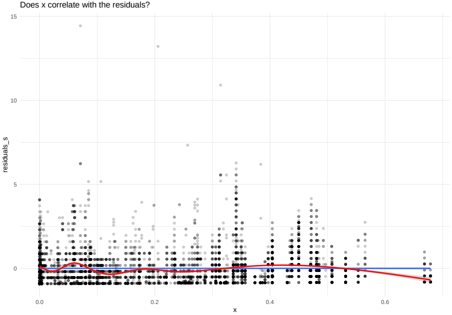
- testing linear relationship with Pearson's R does not gives room for worry

```
##
##  Pearson's product-moment correlation
##
## data:  df$residuals_s and df$x
## t = 3.1632e-15, df = 6946, p-value = 1
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02351429  0.02351429
## sample estimates:
##          cor
## 3.795386e-17
```

- this is the same as saying the mean of the residuals is 0

```
## [1] 3.549449e-14
```

Correlation between x and residuals: visual



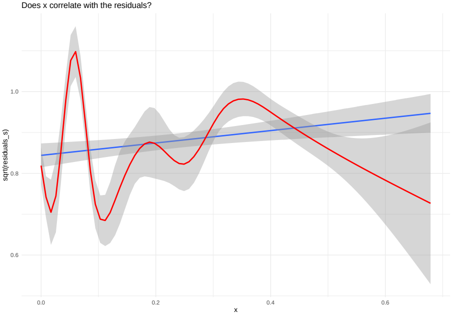
- a bivariate model seems to indicate a flat slope

Is this enough?

- relation between x and residuals may be non-linear
- joint correlation of several covariates is hard to check
- endogeneity may still exist!

$\Rightarrow$ *are there confounders lurking somewhere?*

Let's check that relationship again



- correlating x with the square root of the residuals give a positive correlation
- the *dispersion* of outcomes depend on x (assumption 2 is related to 3)

Time to think

Time to think

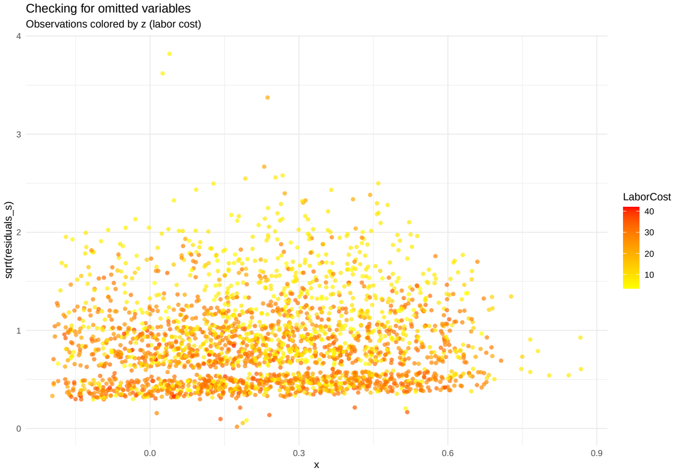
If you find signs of heteroskedasticity and/or correlation between ϵ and x, you should consider

- **observables:** are there control variables that I've omitted?
- **non-observables:** are there groups of observations that share the same "identity"?

What is a confounder?

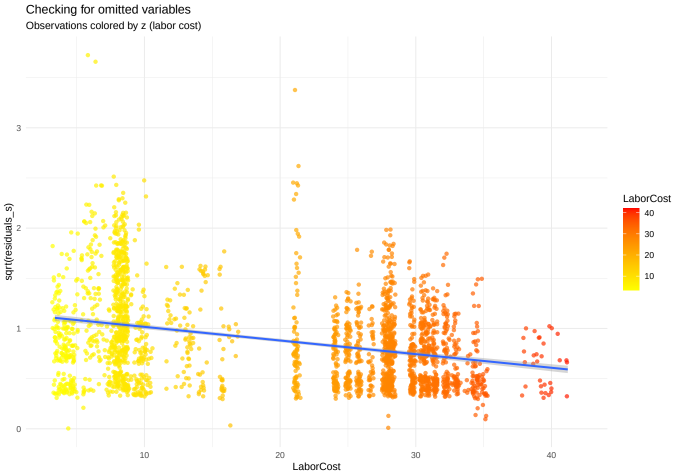
- statistics: variable that correlates with both x and y
- theory: variable that causes x and y; not a “mediator”

Suggestion for omitted, but observable confounder: Labor cost



- Would labor cost impact both vote share of a party and staff size?

Correlation of labor cost with residuals



- Correlating labor cost directly with the residuals reveals a pattern

Correlation of labor cost: in numbers

- statistics: Labor cost is correlated with x, y and thus residuals.

```
##                y     x residuals LaborCost
## y           1.00 -0.03      1.00     -0.30
## x          -0.03  1.00      0.00     -0.15
## residuals   1.00  0.00      1.00     -0.29
## LaborCost -0.30 -0.15     -0.29      1.00
```

- theory: it causes hiring decisions (budgetary limits), but not really vote share?

-> -> ->

Implementation

- Let's control for labor cost anyways.

```
##
## Call:
## lm(formula = y ~ x + LaborCost, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.183 -1.691 -0.517  1.036 39.034
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.491255   0.093758  47.902  < 2e-16 ***
## x           -1.162326   0.183244  -6.343 2.39e-10 ***
## LaborCost   -0.075100   0.002956 -25.403  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.681 on 6945 degrees of freedom
##   (195 observations deleted due to missingness)
## Multiple R-squared:  0.08574,    Adjusted R-squared:  0.08548
## F-statistic: 325.7 on 2 and 6945 DF,  p-value: < 2.2e-16
```

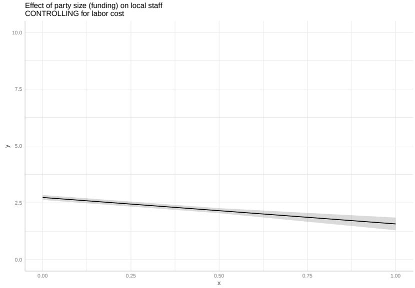
Compare results

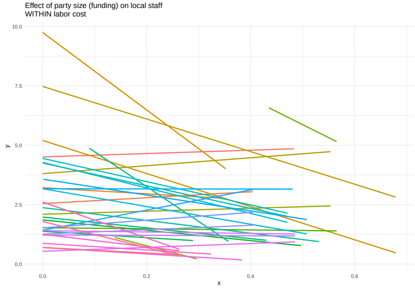
- What happened?
- Can you make a new interpretation of the marginal effect?

Table 1:

	Dependent variable:	
	y	
	(1)	(2)
x	-0.444^{**}	-1.162^{***}
	(0.189)	(0.183)
LaborCost		-0.075^{***}
		(0.003)
Constant	2.561^{***}	4.491^{***}
	(0.057)	(0.094)
Observations	6,948	6,948
R^2	0.001	0.086
Adjusted R^2	0.001	0.085
Residual Std. Error	2.803 (df = 6946)	2.681 (df = 6945)
F Statistic	5.509^{**} (df = 1; 6946)	325.674^{***} (df = 2; 6945)
Note:		*p<0.1; **p<0.05; ***p<0.01

Assumption

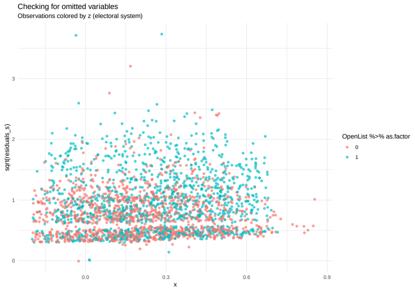


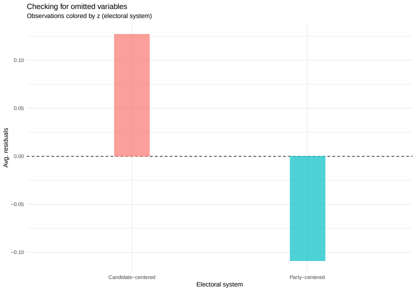


- the regression slope is an average of all slopes within each level of labor cost
- if you don't think that is the best description, you need an interaction effect

Hunting for confounders: your turn!

Suggestion for omitted control variables: Electoral system





- Would electoral system impact both vote share of a party and staff size?

Correlation of electoral system

- statistics: Electoral system is correlated with y and x.

```
##               y     x residuals LaborCost OpenList
## y          1.00 -0.03      1.00     -0.30     0.12
## x         -0.03  1.00      0.00     -0.15    -0.11
## residuals  1.00  0.00      1.00     -0.29     0.12
## LaborCost -0.30 -0.15     -0.29      1.00    -0.09
## OpenList   0.12 -0.11      0.12     -0.09     1.00
```

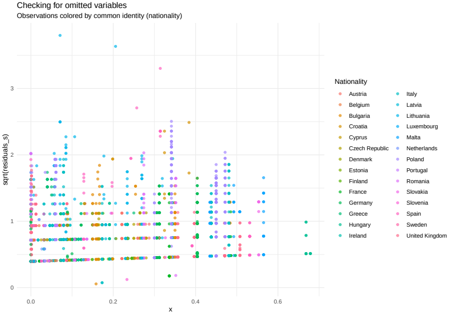
- theory: it causes hiring decisions (electoral incentives), but what about party size in national parliament?

Implementation

```
##
## Call:
## lm(formula = y ~ x + OpenList, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.871 -1.859 -0.804  0.910 40.146
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.19975    0.06762  32.533   <2e-16 ***
## x           -0.23406    0.18912  -1.238    0.216
## OpenList     0.67078    0.06740   9.952   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.783 on 6945 degrees of freedom
##   (195 observations deleted due to missingness)
## Multiple R-squared:  0.01484,    Adjusted R-squared:  0.01456
## F-statistic: 52.32 on 2 and 6945 DF,  p-value: < 2.2e-16
```

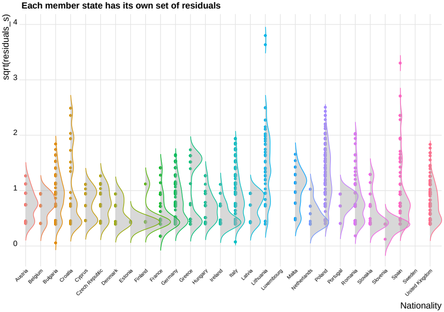
Hunting for omitted variables: common identity

Do these covariates have a common "location"?



- Would nationality impact both vote share of a party and staff size (and labor cost and electoral system)?

Hunting for omitted variables: common identity



- instead of thinking of the residuals as one common distribution, we can think of it as a set of distributions, one for each country

⇒ *Varying-intercept models do this by "labeling" the residuals according to group identities.*

Hierarchical models

Overview

- varying-intercepts: control for group identity
  - fixed effects (mostly this week)
  - random effects (mostly next week)
- varying slopes: (next week)
  - different effects of x per group
- handle standard errors (next week)
  - fixed effects + robust standard error
  - random effects + 2-level variables

Positive take: Strategic leverage of variation

Positive take: Strategic leverage of variation

Phenomena are sometimes observed within a shared context

- we suspect that there are unobserved covariates that influence
  - the outcome and our predictors → *spurious relationships/confounders*
  - our standard error → *observations are too similar/too many*
- examples:
  - geographic context:
    - patients in hospitals: same administrative procedures
    - unemployed in municipalities: same job market/economy
    - conflicts in countries: same competition for resources/power
  - time:
    - patients/unemployed/conflicts: years
  - time and space:
    - time-series cross-sectional/panel data
    - e.g. MEPs in years from countries

Data contains variation

Analysis is about strategically leveraging variation

- information (β)
- noise:
  - random noise: lack of precision (σ^2)
  - bias: confounders:
    - as control variables (λz)
    - or labelled residuals (${\sigma_j}^2$)

$\Rightarrow$ *hierarchical models are very explicit about this*

Our example: MEPs and their local investments

All Members of the European Parliament have the same budget for local staff

- time-series cross-section data with three groups:
  - MEPs are observed every 6 months (`MEP`)
  - there is variation in nationality (`Nationality`)
  - there is variation over time (`Period`)
- covariates at the group-level:
  - MEP: gender, nationality
  - Nationality: electoral system
  - Period: election, reform
- covariates across groups:
  - MEP/time: age
  - Nationality/time: labor cost

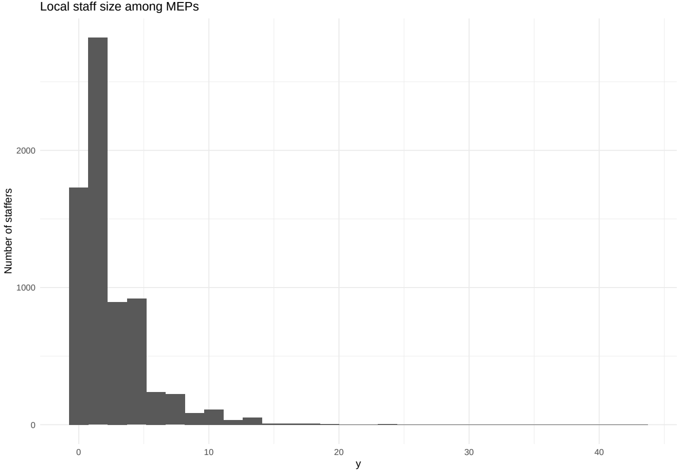
Nesting

We sometimes distinguish between nested and non-nested observations

- nested observations share group identity
  - observations in MEPs never change personal identity
  - MEPs never change nationality (almost)
- non-nested observations have cross-cutting identities
  - time is neither nested in nationality nor MEP

Our dependent variable: Local staff size

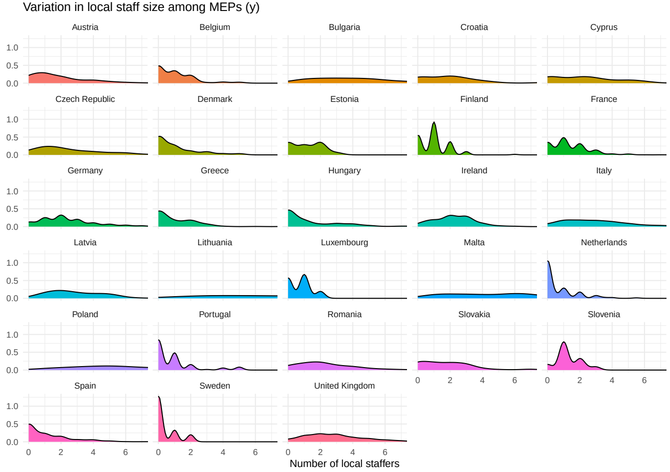
There is variation in the size of MEPs' local staff. What part of this variation am I interested in?



Groups of observations

Groups of observations

Let's consider the distribution of local staff *within* and *between* each member state.



Variation and group averages

Let's consider the distribution of local staff in light of one of the groupings (individual)

```
## # A tibble: 28 x 6
##    Nationality      y_j  sd_j   n_j m_means sd_means
##    <chr>          <dbl> <dbl> <int>   <dbl>    <dbl>
##  1 Austria        1.79  1.65    170    2.34     1.76
##  2 Belgium        0.971 1.15    210    2.34     1.76
##  3 Bulgaria       4.13  2.77    169    2.34     1.76
##  4 Croatia        3.17  4.15     75    2.34     1.76
##  5 Cyprus         2.19  1.91     57    2.34     1.76
##  6 Czech Republic 2.45  2.04    198    2.34     1.76
##  7 Denmark        1.01  1.31    122    2.34     1.76
##  8 Estonia        1.12  0.961    50    2.34     1.76
##  9 Finland        1.02  0.917   131    2.34     1.76
## 10 France         1.38  1.28    611    2.34     1.76
## # i 18 more rows
```

each member state has

- a mean staff size (average staff): e.g.1.79
- a group size (number of observations): e.g.170

***within*-national variation**

- a standard deviation for each distribution: e.g.1.65

***between*-national variation**

- a mean of means (grand mean): 2.3381962
- the standard deviation of the group means: 1.76

→ *we group and label the variation*

⇒ *Which of the variations do I want to leverage?*

Which of the variations do I leverage?

- within-group variation
  - calculate group means to factor out/control away between-group variation
  - regress residuals/remaining variation on within-group predictors

→ *fixed effects (e.g. on member states)*

- between-group variation
  - calculate group means
  - regress the group means on group-level predictors (e.g. electoral system)

→ *an aggregated data frame (e.g. using reframe())*

- both
  - linear model (pooled model)
  - hierarchical models
    - random intercepts account "label"
    - random intercepts with 2-level predictors

→ *hierarchical models leverage both within- and between-group variation*

Why care?

Why care?

When observations have these group identities (are nested), we run the risk of:

- too small standard errors (the sample N is too high, given that observations are not iid.)
- leveraging the "wrong" variation (e.g. the Simpson's paradox, not testing our theory)

Recap

Recap

What did I want to convey with this session?

- confounders:
  - theoretical: z causes x and y
  - statistical: z correlates with x and y
- confounders live in the residuals
- sometimes we can "catch" them using "group identities"
- group identities can enter as varying intercepts:
  - a way to label the residuals
  - control for confounders

$\Rightarrow$ *a warmup for fixed- and random-effects models*