VERGLEICHENDE SOZIALFORSCHUNG MIT MEHREBENENMODELLEN IN R

Forschungspraktikum I und II
Dr. Christian Czymara
Hierarchical linear models

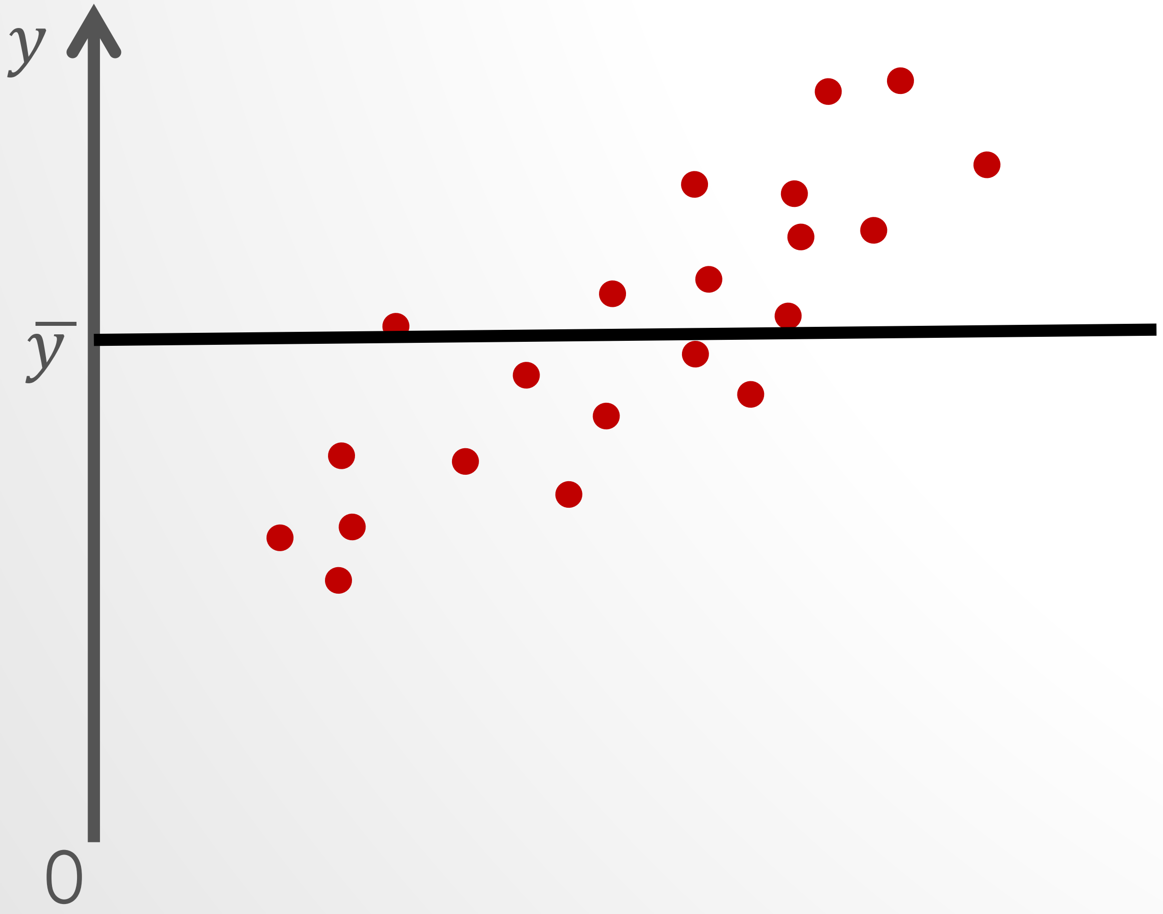
AGENDA

- We learned about linear and logistic regression
- These are excellent estimators when certain assumptions hold
- One assumption: independence of observations
- Today:
  - What if this assumption is violated?
  - And… What does it even mean?

RECAP: ORDINARY LEAST SQUARES ESTIMATOR

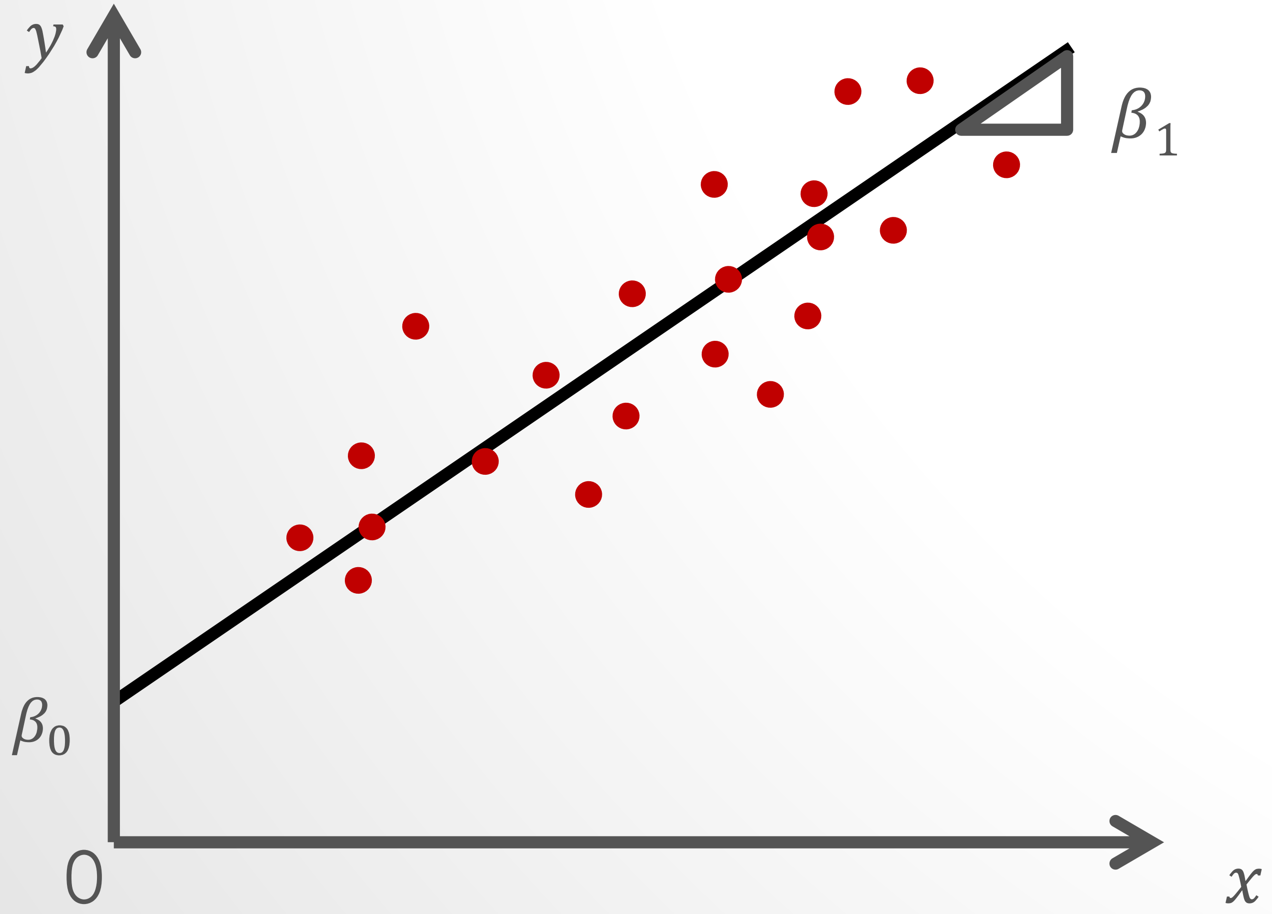
OLS REGRESSION: NULL MODEL

$$y = \beta_0 + e$$



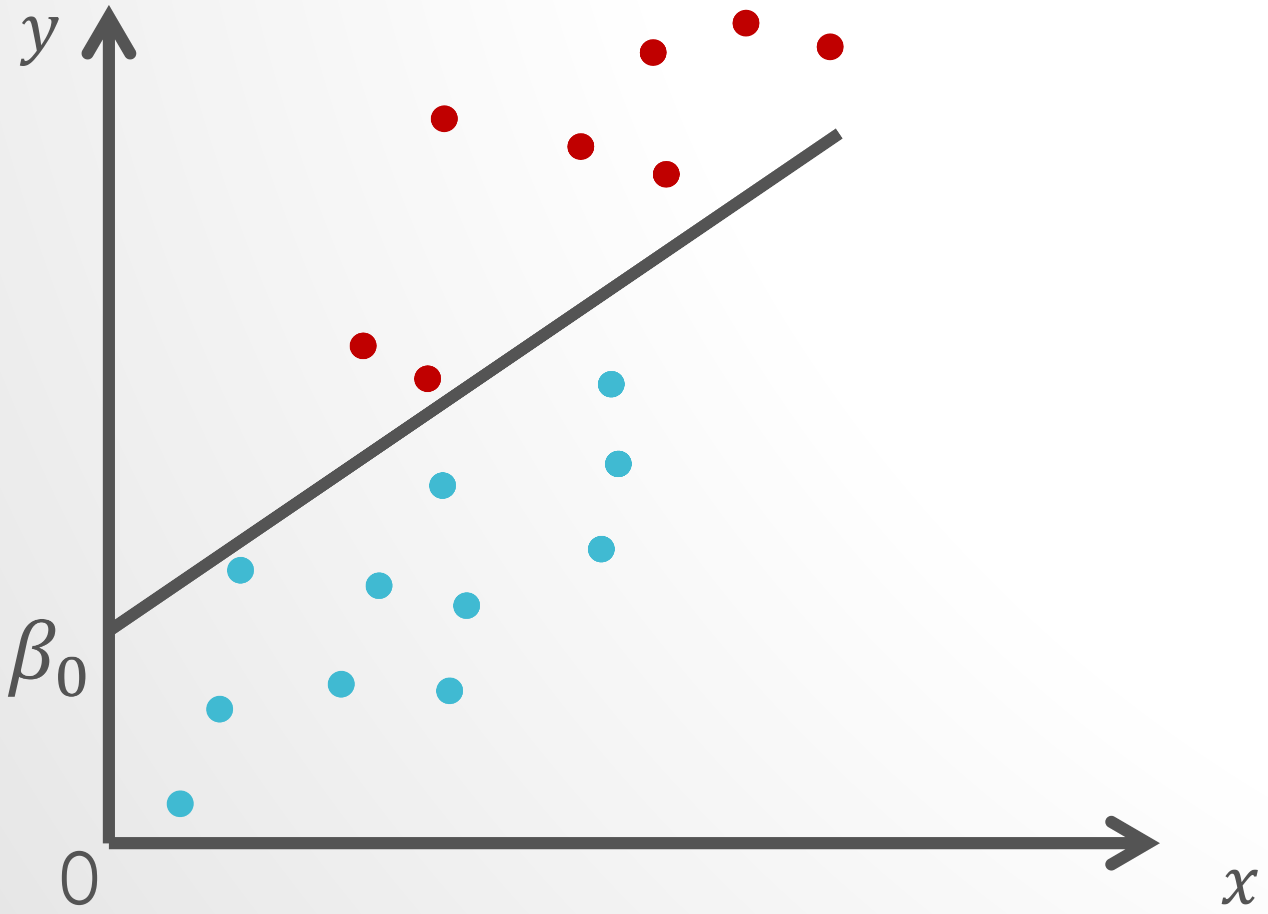
MULTIPLE OLS REGRESSION

$$y = \beta_0 + \beta_1 x + e$$



OLS REGRESSION WITH CLUSTERED DATA

- $y = \beta_0 + \beta_1 x + e$



OLS ASSUMPTIONS

- Correct model specification (no relevant x missing)
- Strict exogeneity (x not correlated with error term)
- Linear independency (x no linear functions of one another)
- Uncorrelated errors
- Homoscedasticity (errors equal across all x)
- Normality (errors normally distributed given x)

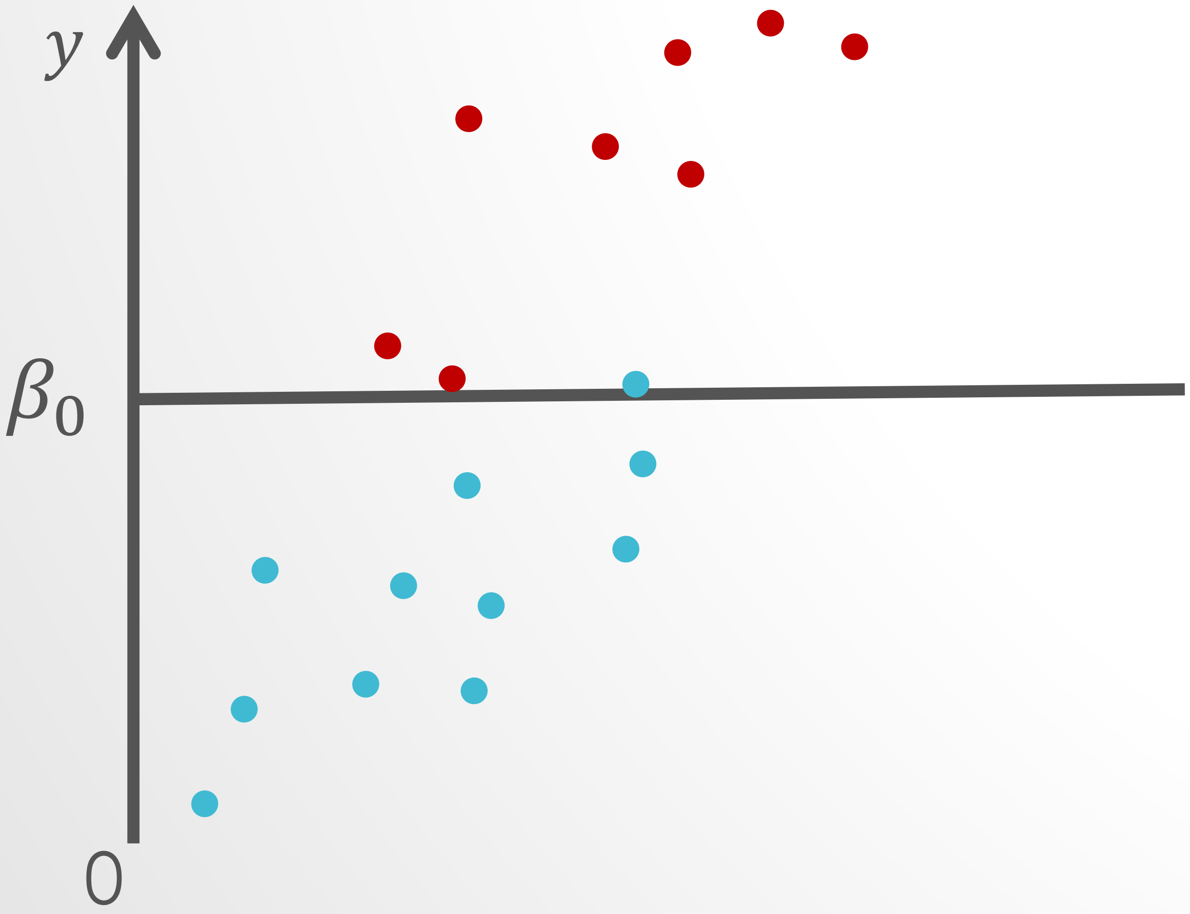
CORRELATED ERRORS IN CROSS-NATIONAL RESEARCH

- Observations within each country may have something in common
- … which separates them from observations in other countries
- For example, income of two random Frenchmen is likely more similar than income between a random Frenchman and a random Romanian
- This means observations are not statistically independent
- This violation of uncorrelated errors assumption
- Biased standard errors leading to wrong p-values and confidence intervals

OLS REGRESSION WITH CLUSTERED DATA

OLS WITH CLUSTERED DATA: NULL MODEL

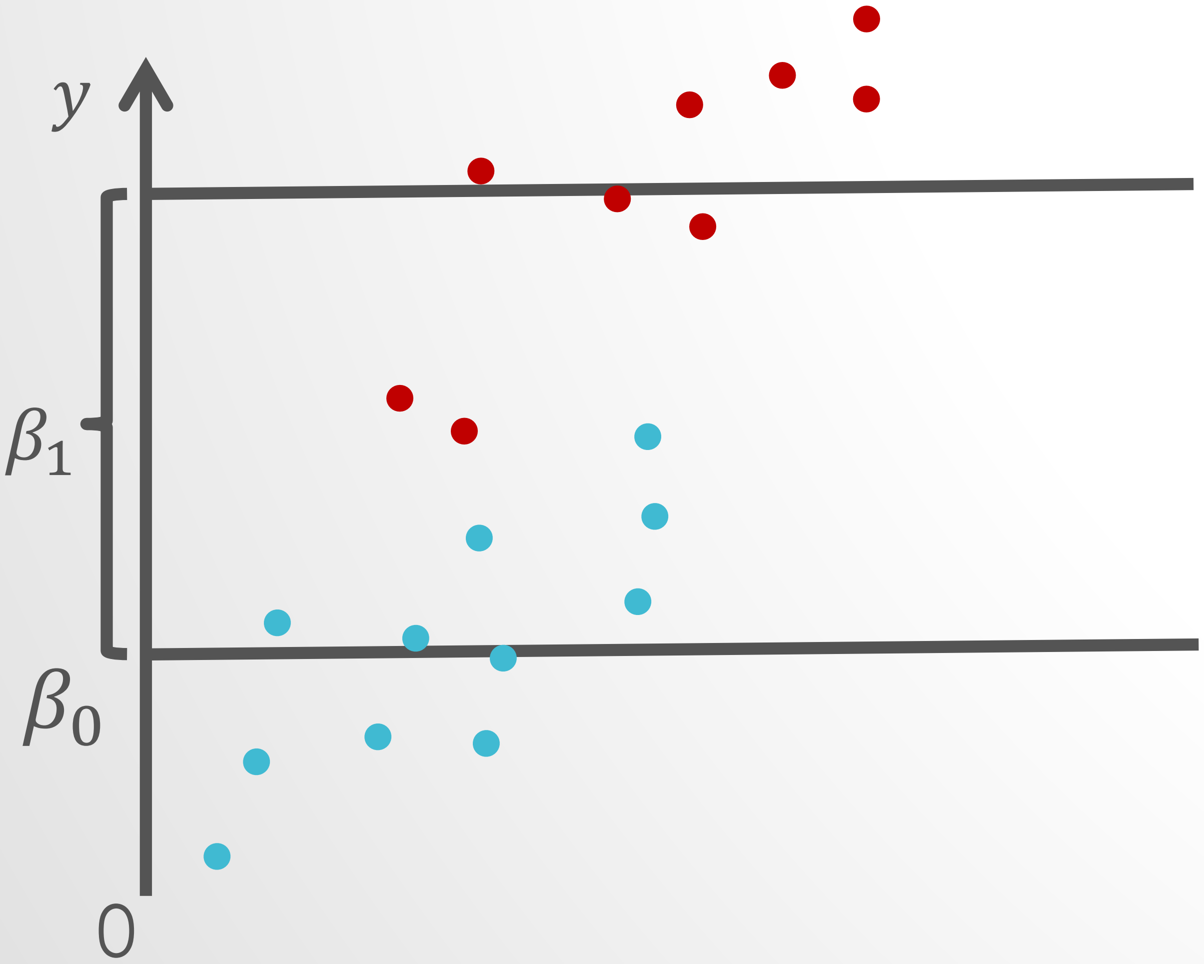
- $y = \beta_0 + e$



OLS WITH CLUSTERED DATA: COUNTRY FIXED EFFECTS

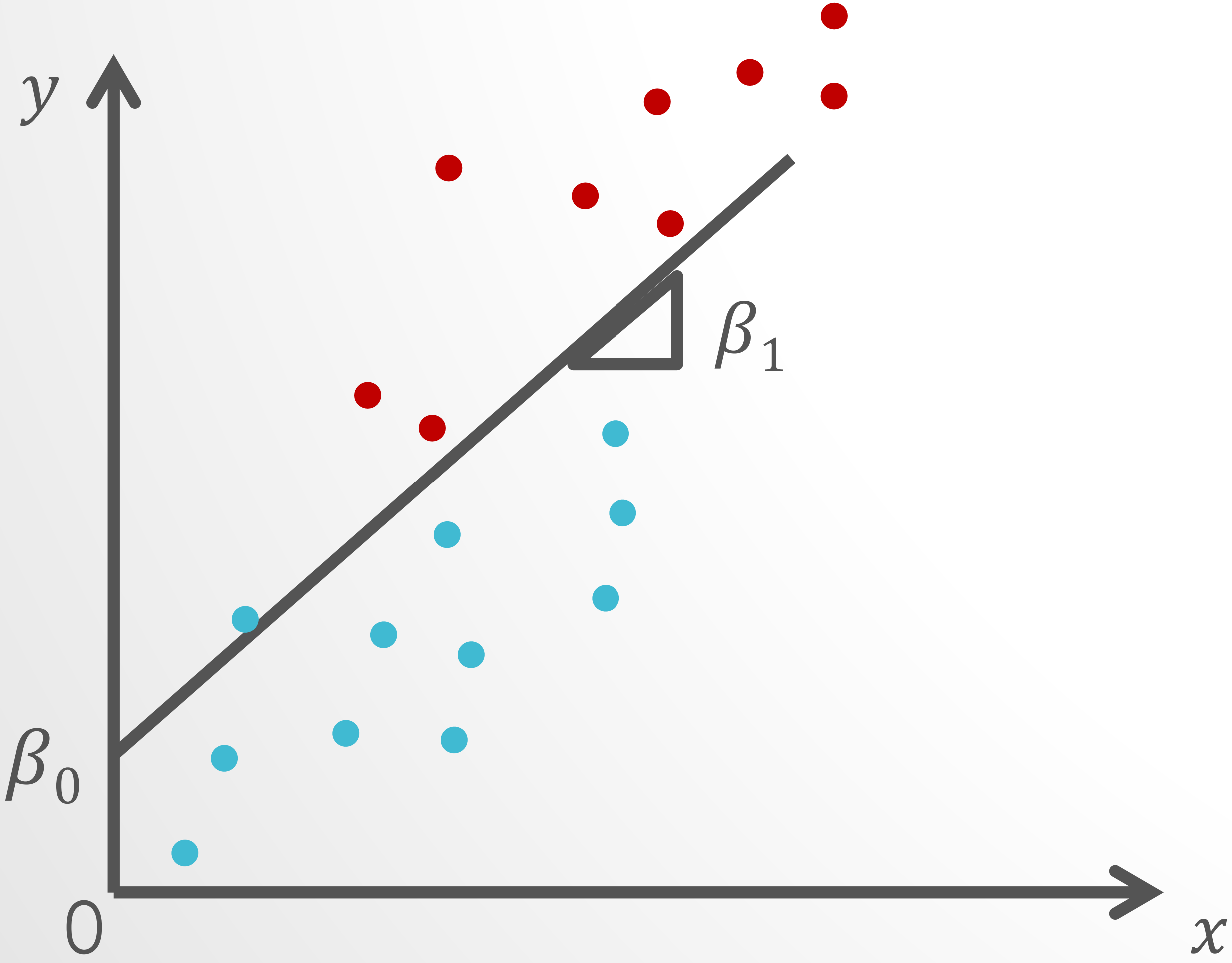
- A dummy variable for a country captures *everything* that is specific about this country
- One way to account for differences between: Statistical control (add dummy variables for each country)
- In this case, the residuals within each country are statistically independent by design (remember statistical control)
- Model estimates one effect for each country
- For example: "*Attitudes are 1.4 units more positive in the UK compared to Denmark*"
- These effects are fixed values
- So-called *Country Fixed Effects Model*

COUNTRY FIXED EFFECTS



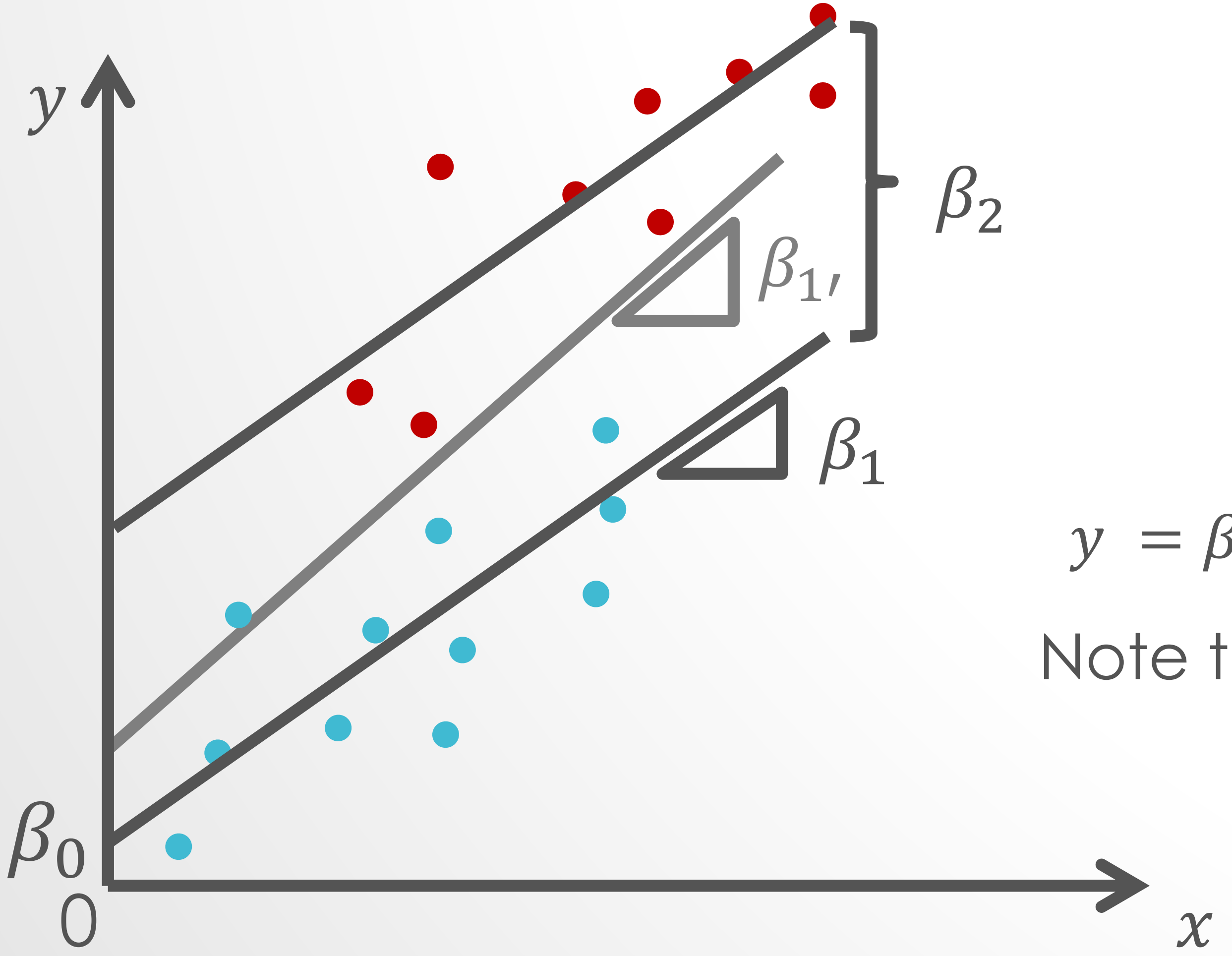
- Red dots belong to country 1, blue dots to country 2
- $y = \beta_0 + \beta_1 country + e$
- $country$ = dummy variable for country (1: red; 0: blue)

MULTIPLE OLS WITH CLUSTERED DATA



$$y = \beta_0 + \beta_1 x + e$$

… WITH COUNTRY FIXED EFFECTS



$$y = \beta_0 + \beta_1 x + \beta_2 country + e$$

Note that $\beta_1 \neq \beta_{1'}$

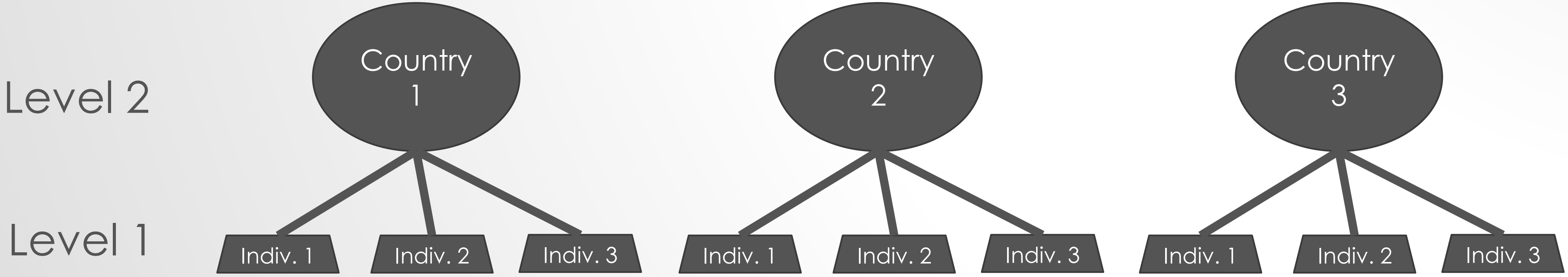
LIMITS OF COUNTRY FIXED EFFECTS

- Dummies account for *all differences* in y between countries
- Identification of effects variables country-level impossible
- Country dummies “control away“ all variance between countries
- Not possible to test which *particular* differences may be important
- For example: Let us say we find a significant difference in welfare state support between Sweden and Spain
  - Is this difference due to differences in national wealth?
  - Or difference in welfare state regime?
  - …
- Methodologically, there is *no variance between countries left* that independent variables on the country level could explain
- However, testing individual level associations (or cross-level interactions) still possible

HIERARCHICAL LINEAR MODELS

HIERARCHICAL LINEAR MODELS

- General application: clustered (nested, hierarchical) data



- Often used for spatial data (e. g., individuals nested in countries)
- But also applicable to, e. g., temporal clustering (observations nested in individuals)

RANDOM INTERCEPT

- Clustered data violates assumption of independence
- HLMs allow the intercept to vary between countries
- Accounts for average differences in y between countries
- Similar to idea of country fixed effects
- However, no point estimates but *variance* estimated (${\delta_u}^2$)

→Additional parameter (a second error term: u)

FIXED & RANDOM EFFECTS

- $\delta_u{}^2$ is called a *random effect (random intercept)*
- It captures the variation of y across countries
- Point estimates of the explanatory variables in a model are called *fixed effects*
- Fixed because values are usually the same for all observations
- Because HLMs estimate random and fixed effects simultaneously, these models are also called *mixed models*
- HLMs model the differences between countries not as fixed effects (by using country dummies) but as random effects (additional error term)

A CAUTIONARY NOTE ON TERMINOLOGY

- The terms fixed effects (FE) and random effects (RE) are used differently in HLM and panel data analysis
- For HLMs:
  - Fixed effects: Point estimates of explanatory variables
  - Random effects: Varying intercepts and slopes → No point estimates but(co-)variances
- For panel data:
  - FE: Models that account for all level 2 variance (like country FE models)
  - RE: Models that only account for serial correlation

TERMINOLOGY

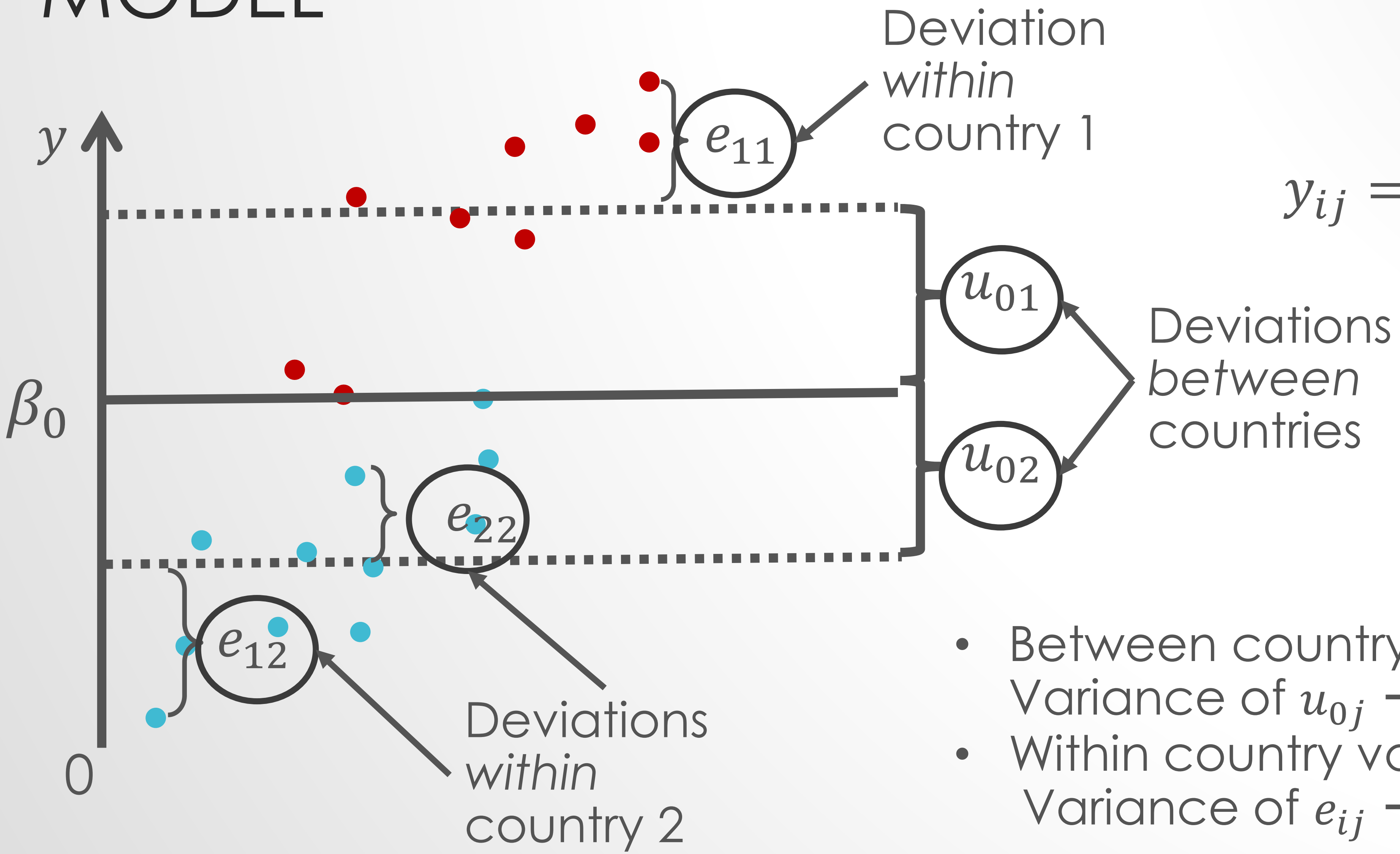
RESIDUALS OF HLMS

- There are residuals on two (or more) levels
  - Unexplained variance between countries
    - Level 2 residuals: ${\delta_u}^2$
    - The random intercept
  - Unexplained variance within countries
    - Level 1 residuals: ${\delta_e}^2$
    - Equivalent to the stochastic error term in basic OLS regression

RANDOM INTERCEPT MODEL: NULL MODEL

- Null model does not include any explanatory variables
- In contrast to basic OLS, HLMs includes an estimate for the variance of the intercept to account for differences between countries
- Hence, the null model is sometimes also called a *random intercept only model*
- Put differently, the null model decomposes the total variance of y into a *between country part* (${\delta_u}^2$) and a *within country part* (${\delta_e}^2$)

RANDOM INTERCEPT MODEL: NULL MODEL



$$y_{ij} = \beta_0 + u_j + e_{ij}$$

- Between country variance = Variance of $u_{0j} \rightarrow \delta_u{}^2$
- Within country variance = Variance of $e_{ij} \rightarrow \delta_e{}^2$

RANDOM INTERCEPT MODEL: NULL MODEL

- $y_{ij} = \beta_0 + u_j + e_{ij}$
- β_0: Mean value across all countries → does not vary (it is the constant)
- u_j: Deviation of country mean from β_0 → Varies only between countries (j)
- e_{ij}: Deviations of individuals from respective country mean → Varies within countries / between individuals (i)

INTRA-CLASS-CORRELATION COEFFICIENT

- The Intra-Class-Correlation Coefficient (ICC) indicates the correlation within the clusters (countries in our case)
- For the null model, this is the *share of variance between the countries*
- $\mathrm{ICC} = \frac{{\delta_u}^2}{{\delta_u}^2 + {\delta_e}^2}$
- The larger the ICC, the more important it is to model this data structure

EXAMPLE: NULL MODEL

- Data: ESS 2002/03
- Outcome: life satisfaction (`stflife`)
- Variable which uniquely identifies higher-level units: `cntry`
- Question: how large is share of between country variance?

→estimate null model

→In R: `lmer(stflife ~ 1 + (1 | cntry), = ESS02)`

Random intercept for countries

EXAMPLE: ICC

Life satisfaction	**Model 0**
Intercept	**7.02*** **
Random effects	
Intercept	**0.617**
Residual	**4.559**

* p<0.05, ** p<0.01, *** p<0.001

$$\mathrm{ICC} = \frac{0.617}{0.617 + 4.559} = 0.119$$

→ About 12 percent of the overall variance is between country variance

ADDING INDIVIDUAL-LEVEL PREDICTORS

- Equation of random intercept model:
  - $y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_j + e_{ij}$
- As two equations:
  - Individual level: $y_{ij} = \beta_0 + \beta_1 x_{1ij} + e_{ij}$
  - Country level: $\beta_0 = \gamma_0 + u_j$

Grand mean (mean over whole sample)

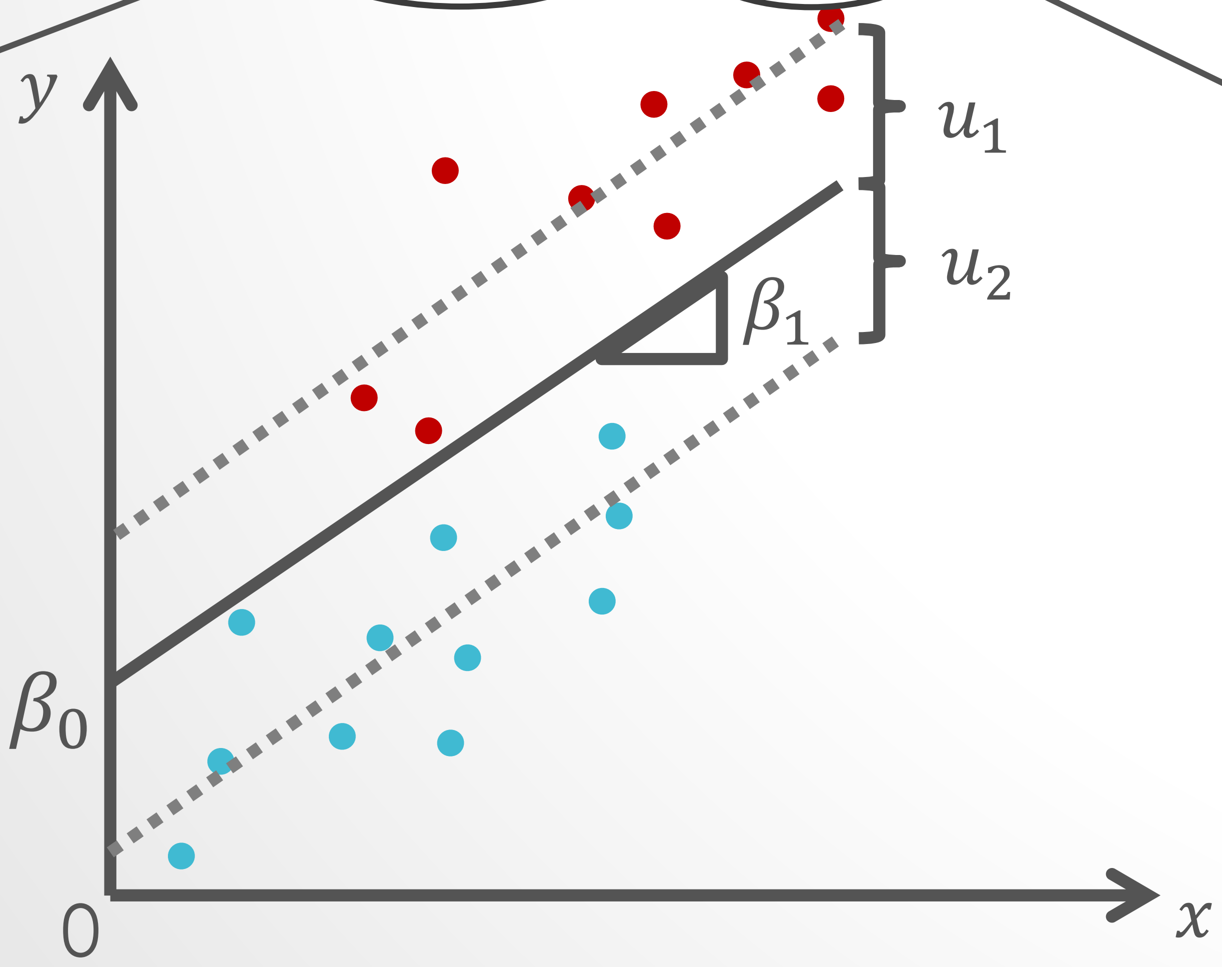
Variance of intercept between countries

ADDING INDIVIDUAL-LEVEL PREDICTORS

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_j + e_{ij}$$

Fixed part / fixed effects

Random part / random effects



COMPOSITION EFFECTS

- Differences between countries may be explained by both individual-level and country-level variables
- When differences between countries are explained by individual level characteristics this is called *composition effects*
- In other words: when adding individual level variables to the model reduces between country variance
- In this case the differences between countries are due to differences of the individuals living in each country and not due to idiosyncrasies of the countries themselves

EXAMPLE: ADDING INDIVIDUAL-LEVEL PREDICTORS

- Explanatory variable income (`hinctnt`, level 1)
- Question 1: Do income and life satisfaction correlate?
- Question 2: How large is the share of between country variance that can be explained by income (composition effect)?

→`lmer(stflife ~ hinctnt + (1 | cntry), = ESS02)`

EXAMPLE: BETWEEN COUNTRY VARIANCE

Life satisfaction	Model 0	**Model 1**
Income		0.18 ***
Intercept	7.02***	5.99 ***
Random effects		
Intercept	**0.617**	**0.319**
Residual	4.559	4.262

* p<0.05, ** p<0.01, *** p<0.001

- Explained *between* country variance: $1 - \frac{0.319}{0.617} = 0.48$

→ Almost half of the cross-national differences in life satisfaction explained by differences in individual income (composition effect)

EXAMPLE: WITHIN COUNTRY VARIANCE

Life satisfaction	Model 0	**Model 1**
Income		0.18 ***
Intercept	7.02***	5.99 ***
Random effects		
Intercept	0.617	0.319
Residual	**4.559**	**4.262**

* $p<0.05$, ** $p<0.01$, *** $p<0.001$

- Explained *within* country variance: $1 - \frac{4.262}{4.559} = 0.065$

→ 6.5 percent of differences in life satisfaction within countries is explained by income differences

ADDING COUNTRY LEVEL PREDICTORS

- Individual level: $y_{ij} = \beta_0 + \beta_1 x_{1ij} + e_{ij}$
- Country level: $\beta_0 = \gamma_0 + \gamma_1 z_{1j} + u_j$

z_{1j}: variable on country level

γ_1: effect of variable z_{1j} → varies neither across countries nor individuals (fixed effect)

EXAMPLE: ADDING COUNTRY LEVEL PREDICTORS

- Explanatory variable GDP/c (`rgdpc`, level 2)
- Question: do national wealth and life satisfaction correlate?

→`lmer(stflife ~ hinctnt + rgdpc + (1 | cntry), = ESS02)`

EXAMPLE: FULL MODEL

Life satisfaction	Model 0	Model 1	**Model 2**
Income		0.18 ***	**0.18 *****
GDP/c			**0.02 *****
Intercept	7.02***	5.99 ***	**5.10 *****
Random effects			
Intercept	0.617	0.319	**0.210**
Residual	4.559	4.262	**4.262**

* p<0.05, ** p<0.01, *** p<0.001

EXAMPLE: BETWEEN COUNTRY VARIANCE REDUCTION

- Since we have estimated three models (null, individual level, country level), we can compare between country variance between…
- The full (country level-) model and the null model: $1 - \frac{0.210}{0.617} = 0.66$

→Together, income and GDP account for almost two third of between country variance in life satisfaction (composition + country effect combined)

- The full model and the individual level model: $1 - \frac{0.210}{0.319} = 0.34$

→GDP explains about a third of between country variance *that is not due to income differences* (country effect)

- Of course, comparison of different models also possible for within country variance

COMPARING MODELS

- To compare models, they must be *nested*

→One model must be a specific form of another more general model (one where parameters are set to zero)

→Practically, this means both models must be based on the same sample

- Likelihood ratio test uses the values of the likelihood function to test which model fits the data better
- To be precise, it compares models' *deviance* values ($\mathbf{deviance} = -2 * \mathbf{ln}(likelihood)$), based on a χ^2-distribution
- When test result is statistically significant, the more complex model (the model with more parameters) performs better
- `lrtest()` function of `lmtest` package

SUMMARY

- People within countries are likely to be more similar
- … which makes them somewhat less similar to people between countries
- From a methodological point of view, the model needs to account for this *statistical dependence*
- Only interested in “broader” differences across countries or not really interested in these differences at all? Use country fixed effects
- Interested in which country characteristics drive the cross-national differences? Use random effects / hierarchical linear models
- These models estimate the variation between countries

LITERATURE

- Schmidt-Catran, Fairbrother & Andreß (2019). Multilevel models for the analysis of comparative survey data: Common problems and some solutions. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 71 (1), 99-128.
- Hox (2002): pages 1 to 32 in "Multilevel Analysis. Techniques and Applications." Routledge.