Statistical Data Analysis with Stata

Additional Topics: Comparing distributions, Multilevel models

Katrin Auspurg & Thomas Hinz

Workshop at Taras Shevchenko National University, Kyiv
September 2015
Day 2

Structure

1. State of projects (short presentations)
2. Next steps
3. Comparing distributions
4. Hierarchical data / Multilevel data
5. Exercises

State of projects

- Thank you for sending us updates.
- Meanwhile, the possibility to include factorial survey modules into (larger) surveys becomes clearer. More on this from Andrii Gorbachyk.
- Working papers, conference abstracts etc. are highly welcome.
- We should include short presentations into our workshop. Now!

Next steps

- Decision 1: what projects to be included to what samples (depending on research proposals and available resources)

- Decision 2: selection of guest researcher coming for about 4-6 weeks in Germany fall 2015 (depending on progress of projects and time constraints)

Comparing groups and distributions

- One is frequently interested in the question if groups come from the same population or if differences between groups are significant. For example:
  - Are the vignette judgments influenced by deck
  - Do respondents differ in vignette judgments by subgroups?

- For this purpose measures of central tendency can be examined (e.g. mean, median); additionally the variation (e.g. variance) should be compared too.
- Furthermore, it is suggested to compare the entire distributions. There are graphical methods and statistical distributional tests for this.

Comparing groups with Stata

- 2 groups:
  - Comparing means: t test `(ttest)`
  - Comparing variances: `(sdtest)`
  - Comparing medians : Wilcoxon Mann Whitney test (`ranksum`)
- More than 2 groups:
  - Comparing means: Analysis of variance `(anova; oneway)`
  - Comparing variances: e.g. Bartlett's test `(can be found in the output with oneway)`
  - Comparing medians: Kruskal Wallis test `(kwallis)`
- Influence of metric independent variable: regression analysis `(regress)`

Comparing entire distributions with Stata

- Graphical comparison of distributions of metric variables: e.g. with `kdensity`:

```
kdensity vig_judge1 if r_sex == 1, ///
plot(kdensity vig_judge1 if r_sex == 2) ///
legend(label(1 male) label(2 female)rows(1))
```

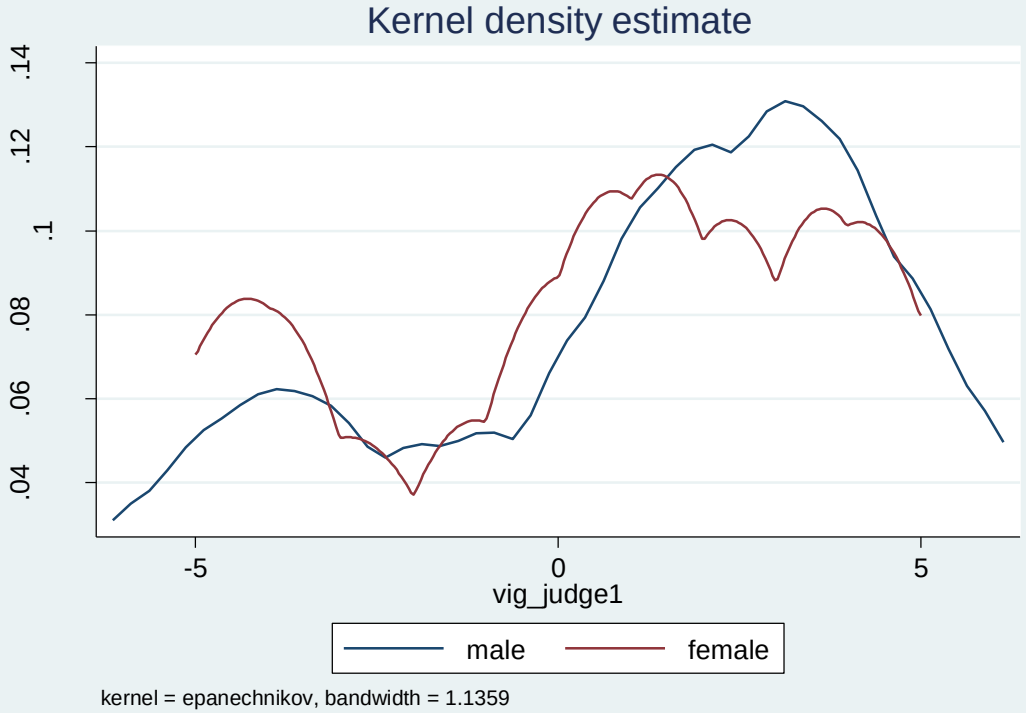
- Statistical testing: e.g. Kolmogorov Smirnov test `(ksmirnov)`

```
ksmirnov vig_judge1,  by(r_sex)
```

- In a similar way it is possible to test if the data follows a given distribution (e.g. normal distribution)
- More information: http://www.ats.ucla.edu/stat/stata/faq/eq_dist.htm

Comparing entire distributions - Example

- Distribution of vignette judgments by respondents' sex




```
. ksmirnov vig_judge1,  by(r_sex)

Two-sample Kolmogorov-Smirnov test for equality of

 Smaller group       D       P-value
 ---------------------------------------
 male:             0.0000     1.000
 female:          -0.1386     0.019
 Combined K-S:     0.1386     0.039

Note: Ties exist in combined dataset;
      there are 11 unique values out of 620 observ
```

Summary

1. Comparing distributions
2. Hierarchical data / Multilevel analysis
3. Outlook

Hierarchical data

- Also called "clustered data" or "multi level data"
- Data are hierarchically structured or related e.g.:
  - Pupils from separate classes or schools
  - Interviews from separate neighbourhoods which again can be combined to different districts
  - Vignettes in respondents!
- Particularly, for sociology the influence of social context is interesting. For example:
  - Does the social composition of working teams or neighbourhoods have an influence on actions or attitudes of people working or living there?
  - Does the composition of cohorts influence the chance to find employment?
  - Durkheim's study about suicide represents a classical example for this kind of context-based hypothesis.

Data structure for factorial survey

10 judgments of vignettes per respondent

→ nested observations

ID_Befragter: 3						
ID_Vignette	Geschlecht	Alter	Abschluss	Beruf	Einkommen	Urteil
1	2	50	3	7	500	-2
2	1	60	2	5	900	-5
3	2	30	2	8	3800	3

ID_Befragter: 2							
ID_Vignette	Geschlecht	Alter	Abschluss	Beruf	Einkommen	Urteil	5
1	2	40	3	6	1200	-2	0
2	2	30	1	8	500	-4	1
3	1	40	1	5	3800	3	0
4	2	50	2	4	500	-5	-2

ID_Befragter: 1								-3
ID_Vignette	Geschlecht	Alter	Abschluss	Beruf	Einkommen	Urteil	2	0
1	1	30	1	2	500	-5	3	1
2	2	40	3	3	950	3	4	
3	2	30	2	1	15000	5	5	
4	1	50	3	9	10000	3	5	
5	2	60	3	6	2500	0	0	
6	2	50	3	8	3800	-1		
7	1	40	2	5	500	-5		
8	2	30	1	3	6800	3		
9	1	30	1	9	1200	-3		
10	2	30	2	5	10000	3		

Data matrix

ID_Befragte	ID_Vignette	Geschlecht	Alter	Abschluss	Beruf	Einkommen	Urteil	Alter_befr	Geschl_befr
1	1	1	30	1	2	500	-5	27	2
1	2	2	40	3	3	950	3	27	2
1	3	2	30	2	1	15000	5	27	2
1	4	1	50	3	9	10000	3	27	2
1	5	2	60	3	6	2500	0	27	2
1	6	2	50	3	8	3800	-1	27	2
1	7	1	40	2	5	500	-5	27	2
1	8	2	30	1	3	6800	3	27	2
1	9	1	30	1	9	1200	-3	27	2
1	10	2	30	2	5	10000	3	27	2
2	1	2	40	3	6	1200	-2	23	1

Hierarchical Data – Statistical aspects

- If the special data structure is not considered, parameter estimation is biased: The assumption of independent units within the sample is probably not true.
- This is given if an "intra class correlation" ρ exists: The units within a single context are more similar to each other than in different contexts. (Consequence: Less efficiency in comparison to simple random samples, sample size has to be increased by the design effect to keep estimation precision constant.)
- If the context is neglected, the estimation of standard deviations and regression coefficients will be biased (normally under-estimated), so that assessing statistical significance might be wrong (normally too optimistic).
- Remedy: Regression with "robust" standard errors (in Stata: add option `cluster(clustervar)` to regression; or explicit modelling with multilevel analysis)

Multilevel analysis versus robust standard error

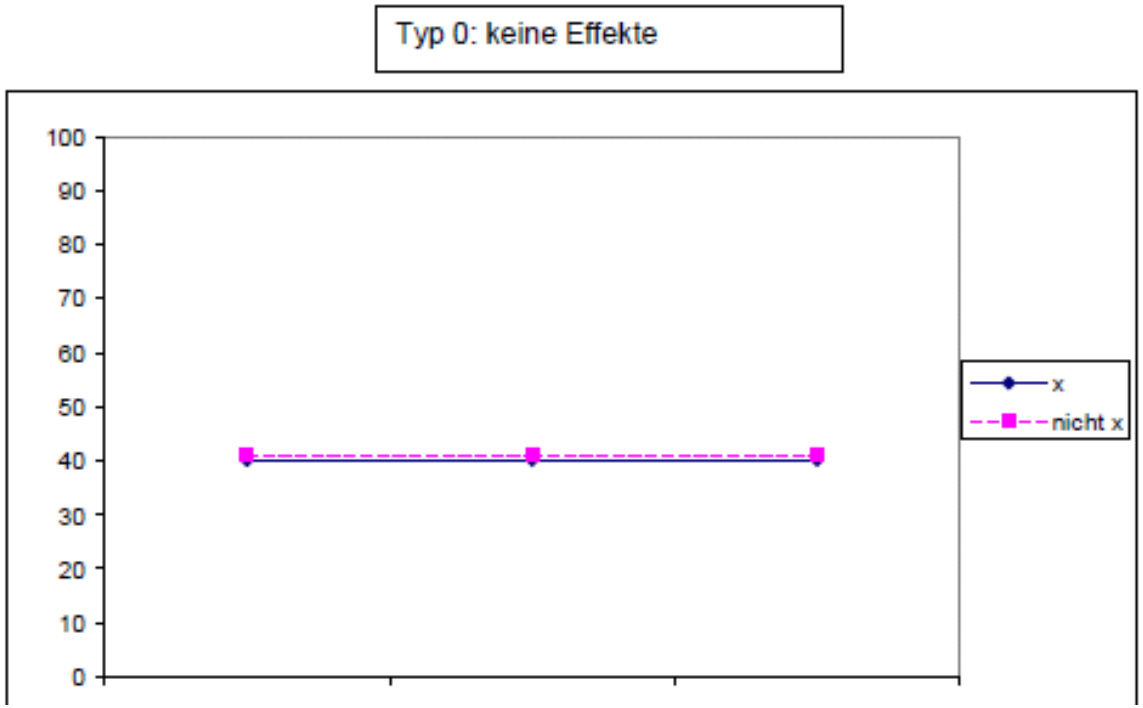
- Multilevel analyses have the aim and the advantage to estimate context effects as exact as possible and separate individual from collective effects.
- An exact modelling of context effects is only possible if there is enough information: if there are too little observations for every context unit, the method might estimate adequate standard errors; but context effects are estimated with high uncertainty, so that there is little gain of information compared to regression with robust standard errors.
- Our analyses with factorial survey data show that there is almost no gain of information for multilevel analysis with up to 20 vignettes per person asked compared to regression with robust standard error, especially if the vignettes have many dimensions. In these cases regressions, with robust standard errors are better because less assumptions have to be satisfied.

Hierarchical data - Hypotheses

- For many analyses in social sciences it is essential to separate individual, collective and context hypotheses.
  - Individual hypothesis: An individual characteristic is influenced by an individual characteristic
  - Collective hypothesis: A collective characteristic is influenced by a collective characteristic
  - Context hypothesis: An individual characteristic is influenced by a collective characteristic
- Ecological fallacy must be refused!
- Statistical modelling with multilevel analyses makes it possible to differentiate between the effects and to estimate them correctly. Furthermore interactions between the levels can be analysed ("cross level interactions"): Do certain individual effects appear more often if certain context characteristics exist?

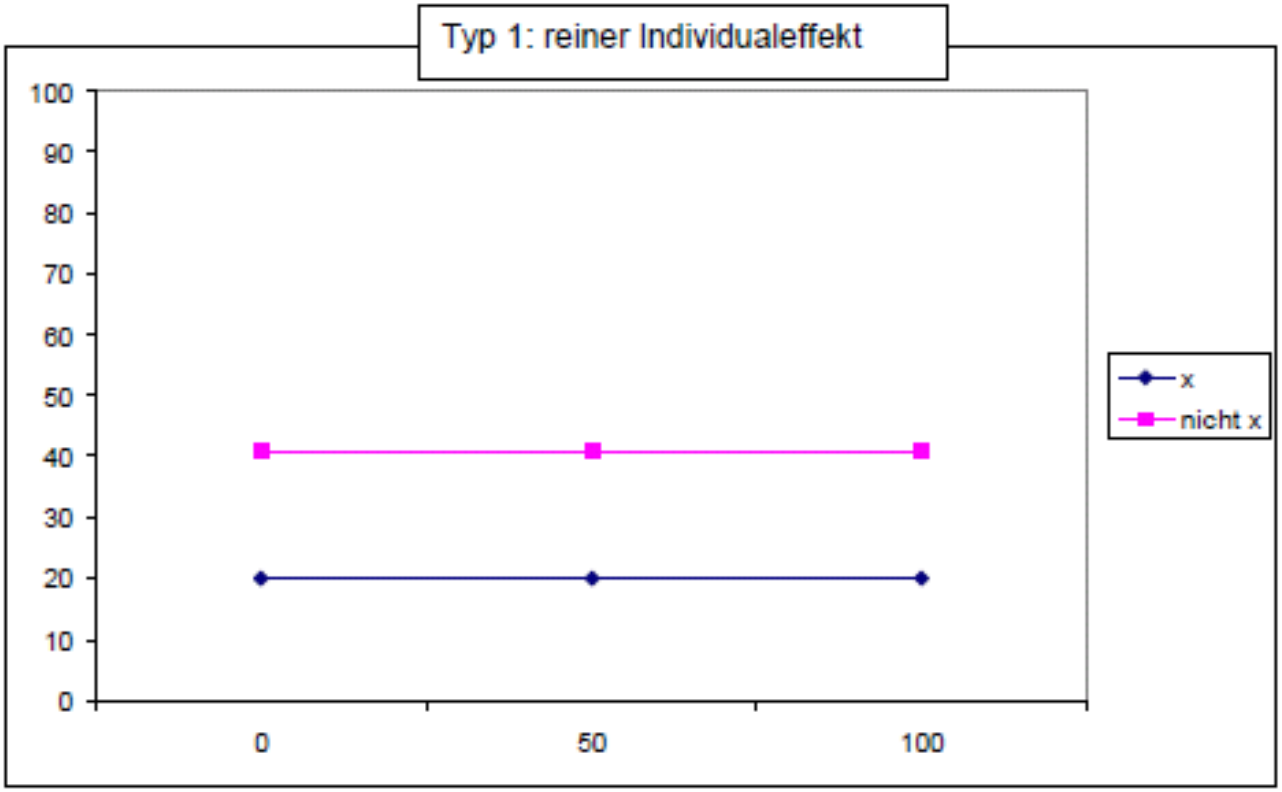
Illustration – no effect

- For the purpose of illustration the share value P is always indicated on the x-axis. The two lines show the relationship between the share value P and Y separately for both individual characteristics (x and not x)

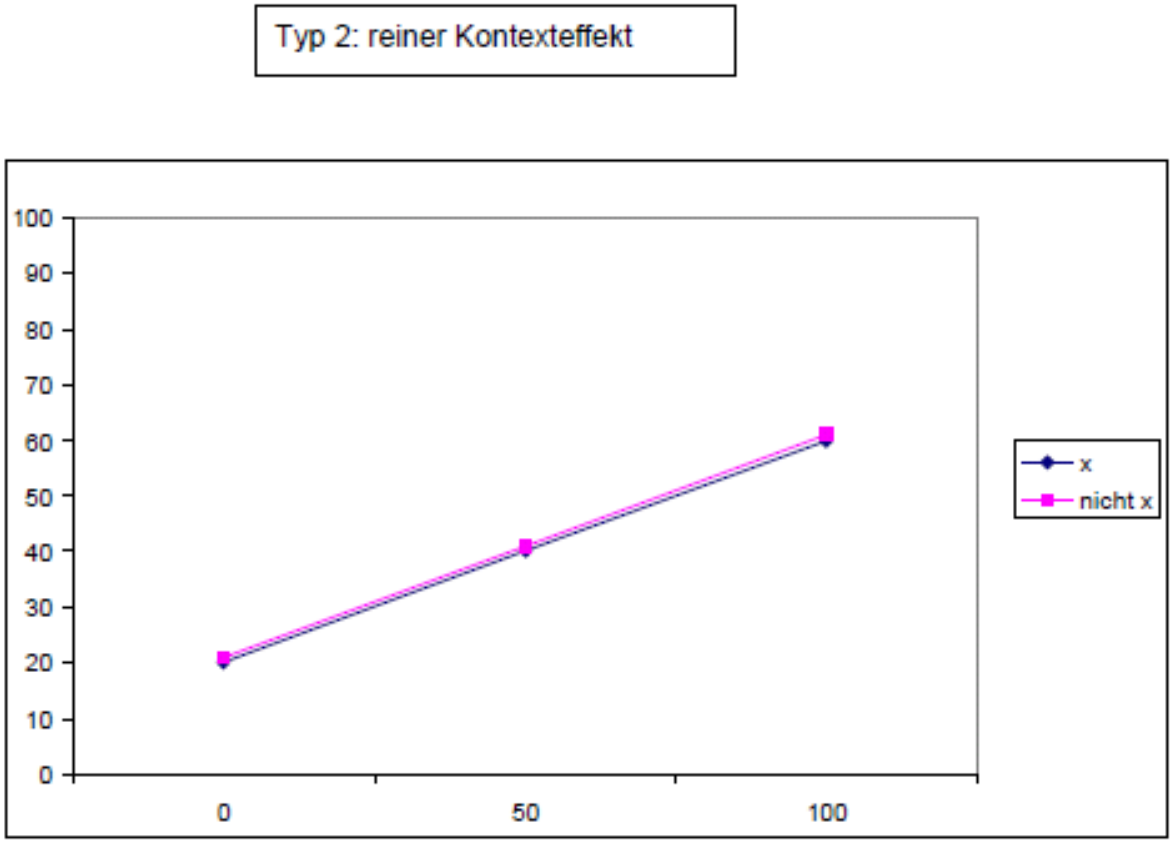


Individual effect only

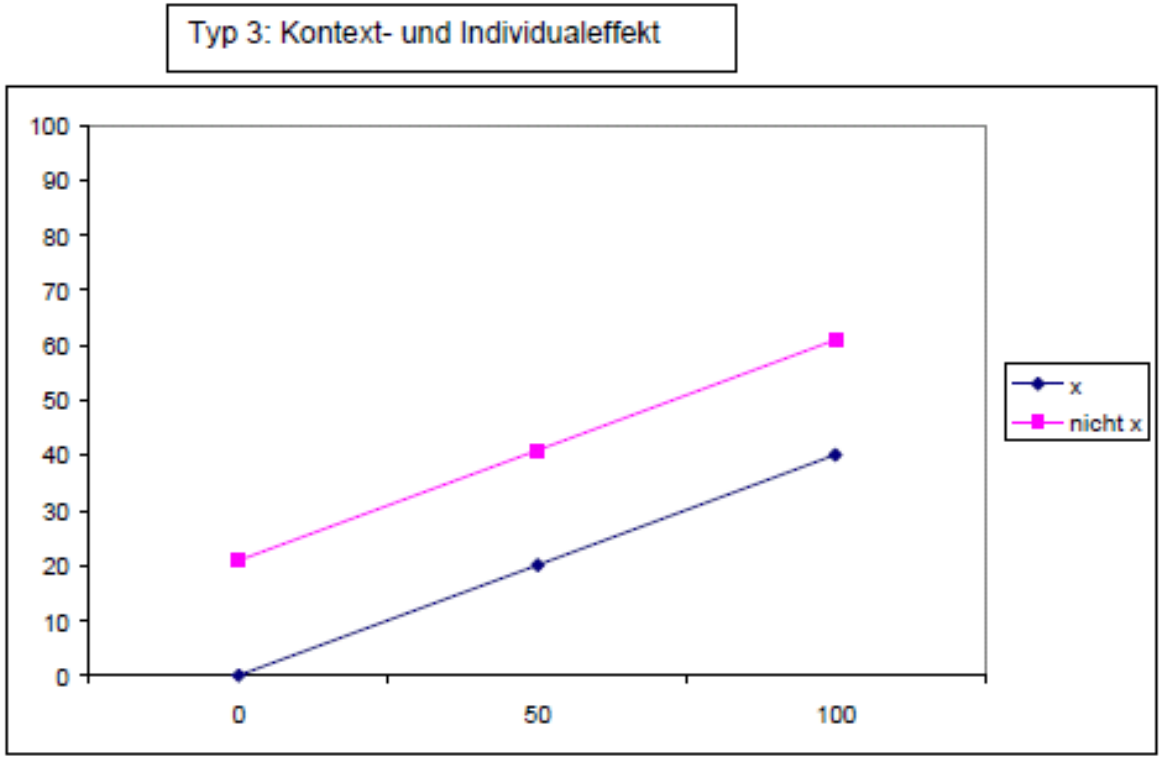
Individual effect only



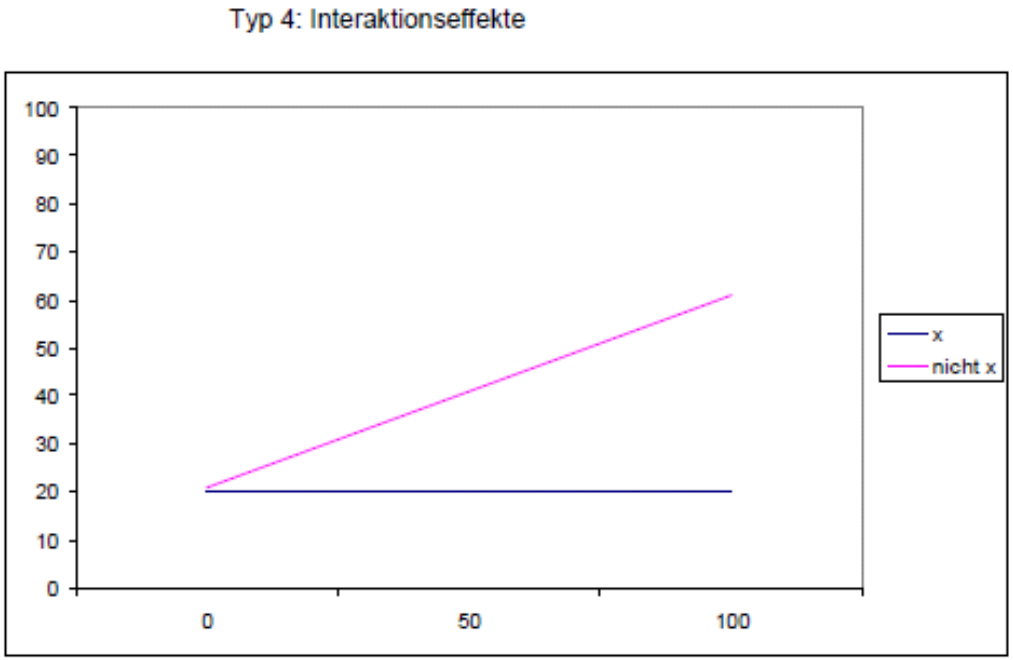
Context effect only



Context and individual effect



Interaction effect



Interaction of individual and context effects

Multilevel models – general information

- Multilevel models are characterized by separately estimating error terms for each level (random variables, expected value 0). Therefore, it is possible to model the variance of the mean (grand means) between and within both levels.
- For an "empty" model without covariates and 2 levels:

$$Y_{ij} = \beta_{0ij} + u_{0j} + e_{0ij}$$

(in which i indicates the lower level, j upper level.)

- For multilevel analyses it is important to estimate the variance for both error terms u_{0j} (upper level) and e_{0ij} (lower level). They are also called *random part* of the model, whereas the estimation of the constant and the coefficients is called *fixed part*.

Multilevel models – Estimation, Rho

- Multilevel models are estimated by Maximum Likelihood and similar methods. The empty model supplies estimator for the intercept β_{0ij}, and for the variance from u_{0j} and e_{0ij}.
- The intraclass correlation coefficient ρ can be calculated as a ratio of Var u_{0j} and Var e_{0ij}.
- It indicates how much of the variance of Y around the *grand mean* can be explained by the affiliation to a context unit. The higher the value ρ, the more information about the values of Y is gained by the multilevel design.

Multilevel models – inclusion of covariates

- The design can be extended by explanatory variables on the lower level, here noted X:

$$Y_{ij} = \beta_{0ij} + \beta_1 X_{ij} + u_{0j} + e_{0ij}$$

- This is a so called "Random Intercept Model". In opposite to simple OLS regressions it includes a more complex modelling of the error term.
- Variables from the upper level, here noted Z, can be added too- this might be for example the share of foreigners within the unit.

$$Y_{ij} = \beta_{0ij} + \beta_1 X_{ij} + \beta_2 Z_j + u_{0j} + e_{0ij}$$

- If variable X contributes to explaining Y, the variances of the error term should decrease while adding these variables. This can be used to determine the explained variance on both levels (within and between), analogously to R^2 values in OLS regression.

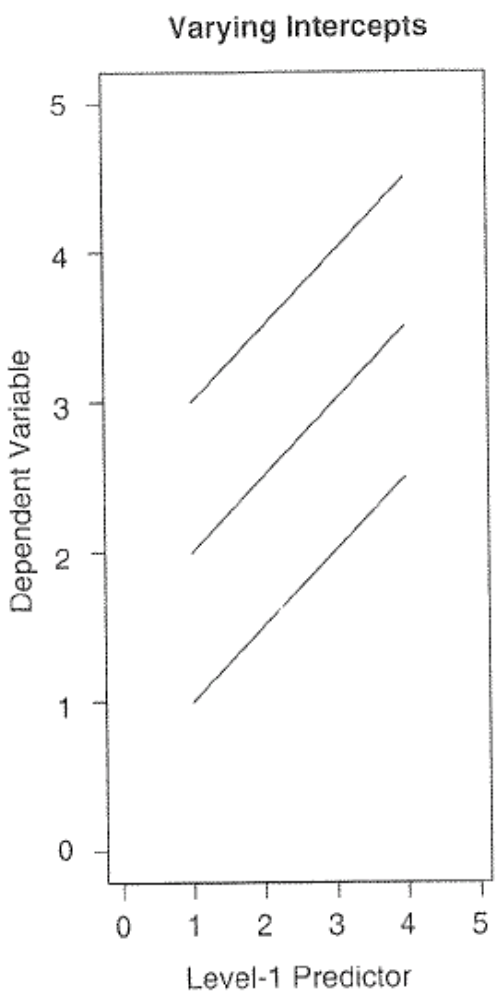
Multilevel models – random slopes, t tests

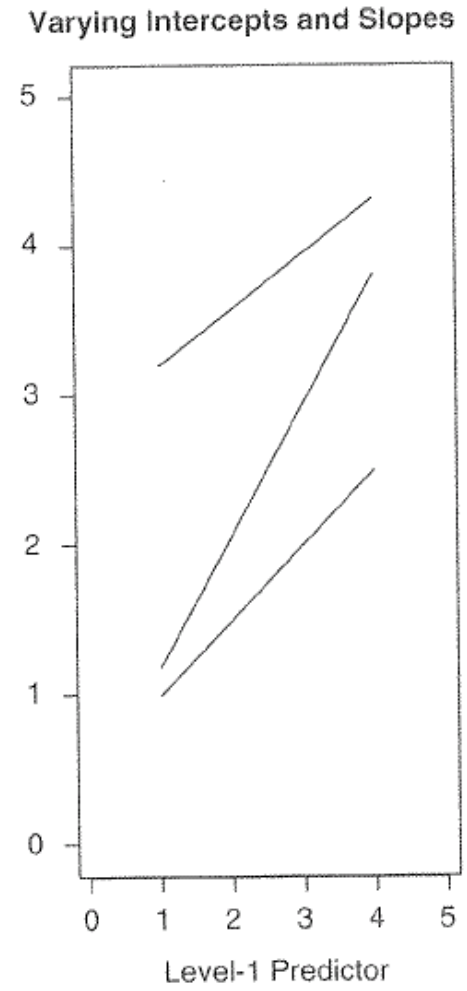
- Complex models allow for different coefficients of the explanatory variable between the context units. Differently said: The influence of X on Y fluctuates between the variable, randomly varying contexts. In technical terms further random terms are added to the estimation equation:

$$Y_{ij} = \beta_{0ij} + \beta_{1j}X_{ij} + \beta_2 Z_j + u_{0j} + u_{1j} + e_{0ij}$$

- You talk about "random slopes" then. The application can't be treated here. You can read about it in specialised literature.
- For fixed effects the significance of the coefficients can be tested with t-tests as in simple OLS regression. But attention: The rule of thumb (Comparing the t-value with standard normal distribution at certain significance level) can only be applied to the significance of variables with higher units if the number of context units is large enough (approximately 40).

Multilevel models – random intercepts and slopes





Source: Douglas A. Luke (2004): Multilevel Modeling. Thousand Oaks: Sage University Press

Multilevel models – further information

- There are a lot of special testing methods and statistical measures: for example to justify the use of a multilevel model or to measure the goodness of fit. To get more information, look up specialised literature.
- Models can be extended to more than 2 levels: this kind of model means more work and takes more time.
- Interactions can be used analogously to OLS regression. E.g. interactions between levels can be generated as product of the corresponding variables (X*Z) and then be added to the regression model.
- Multilevel models are used for panel data too. The structure is very similar: Several observations at different time periods are clustered within one unit (e.g. person).

Fixed effect models

- Apart from random effects models, fixed effects models are especially used for panel data: They correspond to regression models in which for every higher level a new – fixed – constant is estimated (as if one would add k-1 dummy variables to an OLS regression, to estimate different constants for k districts).
- These models should be used if there is correlation between context variables and the error term u_{oj} – or differently said, there is *an omitted variable bias*.
- An advantage of these models is to exclude constant unobserved heterogeneity completely. Only influences of *within-variation* are taken into account.
- Many economists prefer fixed effects models for causal analyses. Though it is not possible to test the influence of constant characteristics of the higher level (for example for panel data: sex) with these models because of perfect collinearity to the constant.

Fixed vs. random intercept model?

criteria	fixed effects	random effects
statistical conclusions from sample to population intended?	no (clusters are unique, e.g. national states)	yes (clusters are randomly selected)
minimum number of clusters required	no	in order to estimate the variance components: 10 to 20 recommended
assumption	no distributional assumption for the fixed effects	error term is normally distributed, variance constant, exogeneous covariates
estimation of cluster characteristics possible	no (only with interaction effects)	yes
mininal cluster size	no restriction, if a lot of clusters are greater than 2, but large clusters needed for a precise estimation of fixed effects	no restriction, if a lot of clusters are greater than 2, but large clusters needed for a reliable estimation of random effects
unbiased estimation for within effects	yes	yes if mean values of clusters are included in the model
thumb rule N<10 N=10 and n_j =100	 X	 X

Rabe Hesketh 2008, Snijders/Bosker 2012

Multilevel regression with Stata

- Can be requested:

```
xtreg depvar [indepvars] [if] [in] ,i(clustervar) ///
[re RE_options]
```

  e.g. empty model for estimating average judgement:

```
xtreg vig_judge1, i(id)
```

- More complex models van be estimated with `gllamm` or `xtmixed`. More information: e.g. Stata book from Rabe-Hesketh/Skrondal.
- Furthermore there is special software for multilevel models, as for example MLwiN which can partly be accessed via Stata (as far as the software is installed).

Summary

1. Comparing distributions
2. Hierarchical data / Multilevel analysis
3. Exercises

Excercises (for the afternoon)

1. Re-estimate the OLS model with vig_judge1 as dependent variable. Include all vignette variables. Additionally, estimate with robust standard errors adjusting for the clustering within respondents.

2. Estimate a random effects/multi model using `xtreg`. Compare the models from exercise 1.

3. Think of interesting interaction effects between vignette variables and respondents' characteristics. Maybe ask Irina what effects she likes you to analyze.

Excercises (for the afternoon)

4. How could one calculate the amount of UAH that a displaced person from the Donbass should ideally receive more/less than other vignette persons? Maybe use the `wtp` command.

5. If one is interested in determining a social minimum of social support one could check if the amount of UAH in the vignettes has a linear effect on the vig_judge1. Any ideas to check for?

6. Think of an additional hypothesis you want to address. Suggest a way of testing your hypothesis.