Statistical Data Analysis with Stata

Katrin Auspurg & Thomas Hinz

Workshop at Taras Shevchenko National University, Kyiv
September 2015
Day 1

Summary

1. Introducing the dataset and Stata
2. Descriptive statistics
3. Recoding variables, data management
4. Cross tabulation and Chi² test
5. Scatter plots and simple linear regression
6. Multiple linear regression
7. Group differences and interaction terms

Appendix

- Correlations
- Comparing mean values
- Analysis of variance
- Logic of controlling for confounding
- Logistic regression
- Non-linear effects

Summary

1. Introducing the dataset and Stata
2. Descriptive statistics
3. Recoding variables, data management
4. Cross tabulation and Chi² test
5. Scatter plots and simple linear regression
6. Multiple linear regression
7. Group differences and interaction terms

Appendix

- Correlations
- Comparing mean values
- Analysis of variance
- Logic of controlling for confounding
- Logistic regression
- Non-linear effects

Pretest data from a factorial survey project

- We will use pretest data from Irina's project
- Subject: The public's perception of the justice of governmental support distribution for socially excluded groups
- 650 vignettes from 65 subjects (students)

Further information about the dataset

- Stored in dta format
- Descriptions can be taken from the questionnaire

#	Name	Label	Type	Format	Value label	Notes
	id		float	%9.0g		
	vignr		byte	%9.0g		
	deck		double	%12.0g		
	gender	gender of vignette person	double	%12.0g	gender	
	labormarket	position in the labor market...	double	%35.0g	labormarket	
	maritalstatus	marital status of vignette pe...	double	%31.0g	maritalstatus	
	children	The presens of children	double	%18.0g	children	
	idp	status of IDPs	double	%26.0g	idp	
	identity	self-identification of person	double	%33.0g	identity	
	income	income of vignette person	double	%12.0g		
	faculty	faculty	float	%9.0g		
	course	course	float	%9.0g		
	vig_judge1		float	%14.0g	support	
	vig_judge2		float	%9.0g		
	complexity	complexity	float	%9.0g		
	r_sex	SEX	float	%9.0g	sex	
	r_year	YBIRTH	float	%9.0g		
	r_income	income	float	%9.0g		
	r_minwage	minimum wage in ukraine	float	%9.0g		
	r_residence	type of house	float	%23.0g	residence	
	r_incomesource	incomesource	float	%45.0g	incomesource	
	r_labourmarket	labour market position	float	%9.0g	labourmarket	
	r_comments	comments	float	%9.0g		
	end_hours	Finishing time: hours	float	%9.0g		
	id_numeric		float	%9.0g		
	id_vignette		byte	%8.0g		

Basic information about Stata (1/3)

- Basic information about Stata and overview of important commands (see Stata Introduction, pdf document, September 2014)
- Always save commands in a do-file! (comments can be indicated with *; commands are written as one line)
- Commands for statistical analysis will be explicates in more detail

- Some fundamental commands to start with:

`log using NAME LOG FILE`	Results will be written into a log file (*log)
`log close`	Close the log file
`tab varname`	Frequency tables
`fre varname`	Frequency tables with labels

Basic information about Stata (2/3)

`recode` *varname* `(value = value)`	Recode a variable
`gen` *newvar* `= ...`	Generate a new variable
`help` *command*	Getting help concerning the respective command
`command if` *condition*	Selection of cases if condition is true

- Commands must always be in separate lines, otherwise a line break can be manually inserted using "///"

```
command ///
 continuation command
```

Indication for Stata: command continues in the next line

Basic information about Stata (3/3)

- Details about particular commands can be looked up in Stata help (help…). Moreover, extensive possibilities for looking up information is available at the Stata website on the Internet.
- Options go after a comma. But ATTENTION: if-statements always come before a comma (they are written directly following the main command):

  `command if condition, option`

- Missing values are indicated with a point (.). Internally these values are stored as +∞! Pay attention to this for recoding, case selection, etc.
- Stata provides excellent graphical tools. They are not covered in this brief overview. Only basic illustrations will be given.
- Stata is case sensitive. It considers the use of small or upper (initial) letters!

Summary

1. Introducing the dataset and Stata
2. Descriptive statistics
3. Recoding variables, data management
4. Cross tabulation and Chi² test
5. Scatter plots and simple linear regression
6. Multiple linear regression
7. Group differences and interaction terms

Appendix

- Correlations
- Comparing mean values
- Analysis of variance
- Logic of controlling for confounding
- Logistic regression
- Non-linear effects

Univariate description: Introduction

- Data analysis always starts with getting a summary of univariate frequency distributions of the variables concerned :
  - Graphical illustrations of frequencies,
    e.g. bar or pie charts, histograms, etc.
  - Numerical descriptions of distributions:
    - Measures of central tendency, e.g. mode, median, arithmetic mean
    - Measures of dispersion, e.g. range, interquartile range, variance, standard deviation
    - Measures of skewness and kurtosis

Level of measurement and measures of central tendency and dispersion

- The **selection of measure** depends on the level of measurement of the variables:

scale of measurement	**central tendency**	**dispersion**
nominal	mode	Simpson's D
ordinal	median	quartile's distance
interval	arithmetric mean	variance, standard deviation
ratio	geometric mean	variation coefficient Gini coefficient

Inspiration for slide design

Univariate descriptions with Stata: Examples (1/5)

- **One-way table of frequencies** are obtained by the command **tab**ulate *var*

```
. tab r_year if vignr==1
```

YBIRTH	Freq.	Percent	Cum.
1991	1	1.54	1.54
1992	2	3.08	4.62
1993	5	7.69	12.31
1994	16	24.62	36.92
1995	21	32.31	69.23
1996	7	10.77	80.00
1997	13	20.00	100.00
Total	65	100.00	

Univariate descriptions with Stata: Examples (2/5)

- The command „frequencies“ often is more illustrative (to be installed as additional package or ado)

```
ssc install fre, replace
```

fre *var*

```
. fre r_residence if vignr==1

r_residence ── type of house
```

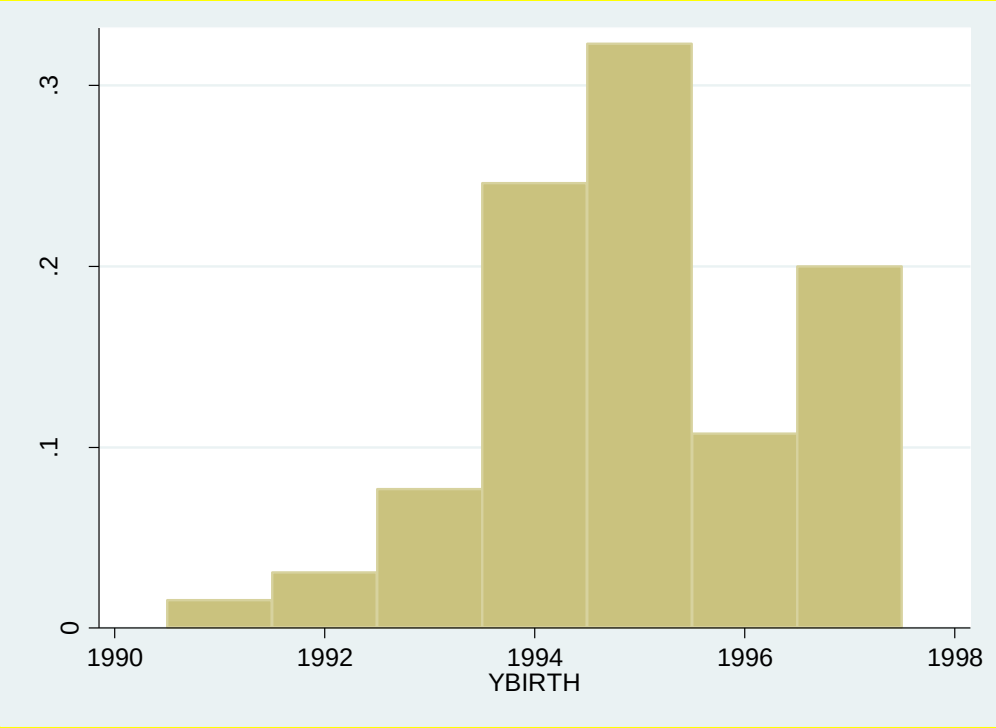
		Freq.	Percent	Valid	Cum.
Valid	1 living with parents	38	58.46	58.46	58.46
	2 rented apartment	3	4.62	4.62	63.08
	3 living in own apartment	3	4.62	4.62	67.69
	4 living in a dormitory	21	32.31	32.31	100.00
	Total	65	100.00	100.00	

Univariate descriptions with Stata: Examples (3/5)

- **Histograms:**

  hist *var*, discrete

  ```
  hist r_year, discrete
  ```



Univariate descriptions with Stata: Examples (4/5)

- For metric variables, measures that describe distributions and graphical illustrations usually are better suitable than frequency tables.
- **Measures of central tendency and dispersion** for metric data:

sum *var*, detail

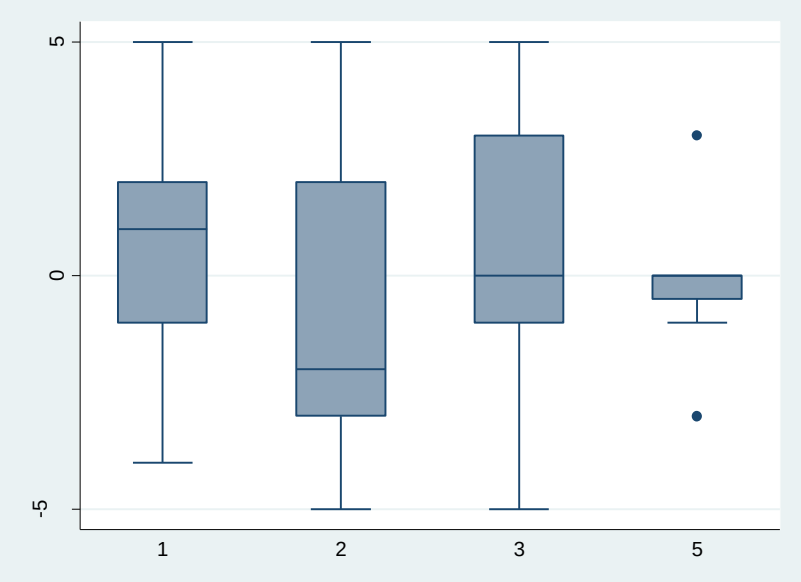
```
. sum r_income if vignr==1, detail

                          income
-------------------------------------------------------------
      Percentiles      Smallest
 1%           4              4
 5%         700            400
10%        1000            700       Obs                  54
25%        1200            800       Sum of Wgt.          54

50%        2000                      Mean            2950.537
                         Largest      Std. Dev.       3784.052
75%        3000             8000
90%        4000            10000      Variance        1.43e+07
95%       10000            20000      Skewness        3.592514
99%       20000            20000      Kurtosis        16.17439
```

A2. What was your income last month per one family member? (If you live in a family, the income of one of its members is determined by dividing the total monthly income by the number of family members, including minors). _____ UAH

Univariate descriptions with Stata: Examples (5/5)

- An informative illustration of metric data is given by boxplots.:
  graph box *var* ,over (*groupvar*)

```
graph box complexity if vignr==1, over(course)
```



course — course

		Freq.	Percent	Valid	Cum.
Valid	1	13	20.00	20.31	20.31
	2	17	26.15	26.56	46.88
	3	26	40.00	40.63	87.50
	5	8	12.31	12.50	100.00
	Total	64	98.46	100.00	
Missing	.	1	1.54		
Total		65	100.00		

B1. Overall, how difficult or easy was to evaluate the fictional descriptions of people?

Very difficult — Neither difficult nor easy — Very easy

-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Univariate descriptions with Stata: exercises (to be done in the afternoon session)

1. How many percent of the respondents are older than 20 years?
2. Choose a suitable graphic illustration for respondents' income (histograms for continuous variables are drawn with the discrete option, you may choose the number of categories by defining "bin(#)"; histogram VAR, bin(#) frequency).
3. Is there a difference by age group (older than 20 years vs. 20 years and younger) concerning the complexity evaluation? (You need to recode the variable!)

Summary

1. Introducing the dataset and Stata
2. Descriptive statistics
3. Recoding variables, data management
4. Cross tabulation and Chi² test
5. Scatter plots and simple linear regression
6. Multiple linear regression
7. Group differences and interaction terms

Appendix

- Correlations
- Comparing mean values
- Analysis of variance
- Logic of controlling for confounding
- Logistic regression
- Non-linear effects

Content-based categorization (1/3)

- For many analyses a restricted number of categories is useful (as regards content or for statistical reasons). For example it can be interesting to look only at the students living with their parents yes/no.
- To check the codes for the original variables, use the commands *fre*, *tabulate* or *codebook*:

```
codebook r_residence
```

```
r_residence
-------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  residence

                 range:  [1,4]                        units:  1
         unique values:  4                    missing .:  0/650

            tabulation:  Freq.   Numeric  Label
                           380         1  living with parents
                            30         2  rented apartment
                            30         3  living in own apartment
                           210         4  living in a dormitory
```

Content-based categorization (2/3)

- A dichotomous variable (out of a continuous variable) can be generated in the following way:

  1. Generate a new variable with two categories which correspond to the content-based categories „r_income below median“ or „r_income median and above“:

```
generate r_income_hgroup= 0
replace r_income_hgroup= 1 if r_income > 2000
replace r_income_hgroup = . if r_income ==.
```

  2. Assign a useful name and value label:

```
lab var r_income_hgroup „high income (median split)“
label define yesno 0 “no“ 1 “yes“
lab val r_income_hgroup yesno
```

  3. Check using a table of frequencies:

```
fre r_income_hgroup
```

- By combining logic operators (& = and; | = or) more complicated recoding can be done

Content-based categorization (3/3)

- If you only want to sum up categories of variables that already exist, you can work with the command recode:

```
recode var numlist  = num
```

For our example:

```
gen r_dormitory =  r_residence
recode r_dormitory 1 2 3=0    4=1
```

„/" stands for „to", here categories 1 to 3

or even easier:

```
recode r_dormitory 1/3=0  4=1
```

Further proceeding as before (assign variable labels; check using table of frequencies)

Categorization by shares

- Alternatively, variables can be split by shares, for example at the median or the quartiles. For example it might be interesting to divide the respondents by their income (r_income) into three groups. For that:

1. First get a summary of the variable (for instance: missing values?)

```
fre r_income
```

2. Generate a new variable (r_income_cat3) with three shares at about the same size (more precisely lowest 33%, the medium 33%, the highest 33%):

```
xtile r_income_cat3 = r_income, nquantiles(3)
```

3. Assign labels:

```
lab var r_income_cat3 „resp_income_categorial"
lab define lowhigh3 1 "low" ///
2 "middle" 3 "high"
lab val r_income_cat3 lowhigh3
```

line break

4. Check using table of frequencies:

```
fre r_income_cat3
```

More useful command for recoding:

- Define values as missing:

```
mvdecode varlist, mv(numlist)
```

e.g.

```
mvdecode r_income, mv(4)
```

For the pretest data no recoding required – but if ones want define an extremely low income as missing one could do it as shown

- Change string variables (variables consisting of text) into numerical variables:

```
destring var, replace
```

Pretest data do not contain any string variables. There are further commands for data management

Summary of important commands

- Univariate statistics
  - `fre ...`
  - `tab...`
  - `sum ...`
- Graphs
  - `hist ...`
  - `graph box`
- Restrictions/sub-groups
  - `if...`

Summary of important commands

- Recode variables
  - `gen ...`
  - `xtile ...`

- Edit variables (labels etc.)
  - `lab var ...`
  - `lab define ...`
  - `lab val ...`
  - `mvdecode ...`
  - `destring`

Always check your recoding afterwards! (e.g. with `fre` or `sum`)

Categorization of variables: Exercises (to be done in the afternoon)

1. Generate a variable which informs you whether the respondents are living with their parents yes/no.
2. Generate a variable which categorizes age of respondents into three groups of equal size.

Summary

1. Introducing the dataset and Stata
2. Descriptive statistics
3. Recoding variables, data management
4. Cross tabulation and Chi² test
5. Scatter plots and simple linear regression
6. Multiple linear regression
7. Group differences and interaction terms

Appendix

- Correlations
- Comparing mean values
- Analysis of variance
- Logic of controlling for confounding
- Logistic regression
- Non-linear effects

Cross tabulations: Basic information

- Also referred to as „contingency table".
- Joint frequency distribution of two discrete characteristics/variables. If characteristics are metric, there is the possibility of categorization (for an example, see next slide).
- Give the opportunity to test hypothesis about relationships:
  - First you look at differences in the percentages.
  - Then you decide on the certainty under which you can infer from sample to population: Chi-square test for independence.
  - Moreover you can check the strength of the relationship between independent (IV) and dependent variable (DV) using measures of association such as Cramer's V.

Drawing up a cross tabulation

- Conventions you should pay attention to:
  - The dependent variable (DV, Y) is represented in the table's rows.
  - The independent variable (IV, X) in the table's columns.
  - Both variables should consist of a manageable number of categories.
- Command for cross tabulation:

  tab *depvar indepvar*

  with conditional frequencies:

  …, ***col****umn*

  with Chi-square test for independence:

  …, ***col****umn chi2*

Chi-square test for independence

- Aim: Check if the relationship between IV and DV discovered in the sample is „more than just by chance"→ **statistical significance testing.**
- We test the **null hypothesis (H_0):**
  there is no relation between IV and DV
  against the **alternative hypothesis (H_1):**
  there is relation between IV and DV.
- General logic behind statistical significance testing:
  Determining statistical measures (which follow a certain distribution) that allow calculating probability of error (here: Chi-square value).
- Properties Chi-square:
  - Value is always positive, that's why you can't tell about the direction of the association.
  - Specific values depend on size of table and sample!

Chi-square test for independence (2/2)

- Stata doesn't only calculate the test statistic, but also tests for significance and displays the probability for the test statistic under the null hypothesis (probability of error α or p-value).
- Convention for social sciences: If $p \leq .05$, then H_0 rejected.
- Help for interpretation:

Probability of error	***Meaning*** *(Symbol)*
$p > 0.05$	not significant (ns)
$p \leq 0.05$	significant (*)
$p \leq 0.01$	highly significant (**)
$p \leq 0.001$	most significant (***)

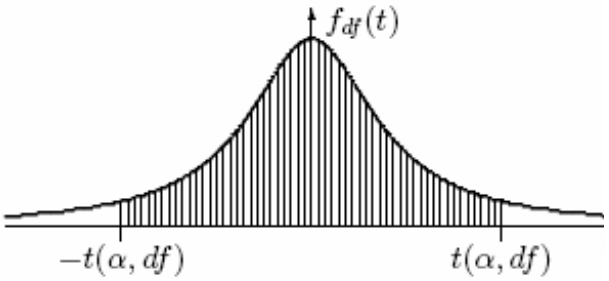
Excursion significance testing: Logic und procedure

- Following questions are answered by statistical hypothesis testing: Is an observation or estimation consistent with a formulated hypothesis, is it close enough to the hypothesis?
- We test the null hypothesis H_0 against the alternative hypothesis H_1; testing is based on rules which decide whether H_0 can be rejected or not.
- For testing we calculate the test statistic using sample data: This might be a point estimation for an unknown population parameter about which value the H_0 can report something (e.g. mean).
- For the test we determine the probability distribution of the test statistic:=*sampling distribution*. (Results theoretically if we draw infinite times samples of the same size from a population and each time determine the value of the test statistic.)
- Determining rejection area: Compare the test statistic with the critical value of the distribution.

Excursion significance testing: Graphical illustration

two sided

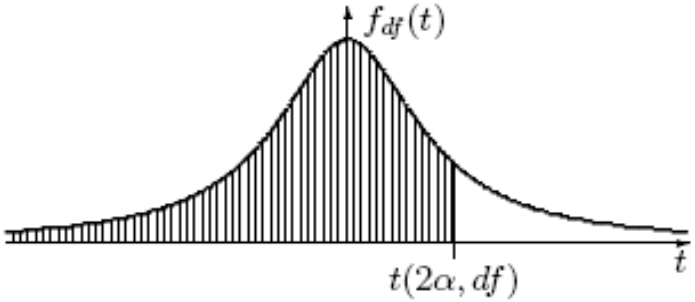
Zweiseitiger Test: $H_0 : \beta = 0$



$$P(-t(\alpha, df) < t_{emp} < t(\alpha, df)) \; = \; 1 - \alpha$$

one sided

Rechts-einseitiger Test: $H_0 : \beta \leq 0$



$$P(t_{emp} \leq t(2\alpha, df)) \; = \; 1 - \alpha$$

Source: Engel/Möhring/Troitzsch 1995: 112

Excursion significance testing

- When deciding whether H_0 can be rejected, two kind of errors can appear:
  - Type I error: H_0 is rejected although it is true („false alarm").
  - Type II error: H_0 is kept although it is false.

unknown reality	Decision for H_0	Decision for H_1/against H_0
H_0 is true	1-α probabilty of correct decision	α probablity that H_0 is rejected while being true type 1 error
H_1 is true/H_0 is false	β probablity that H_0 is accepted (not rejected) while being false type 2 error	1-β probablity of correct decision (power of test, probability of detecting a wrong H_0 as wrong)

→ The validity of an hypothesis can never be proven with last certainty by statistical testing!

Strength of association: Cramer‘s V

- With the χ^2coefficient and the corresponding significance you can determine if a correlation between two variables is more than just by chance.
- But: no comparison of the χ^2 values possible if sample and table size different, because:
  - Coefficient value depends on sample size
  - Coefficient value depends on number of cells in the table
- A measure independent of this is Cramer's V. Now, statements about the strength of the relationship are possible.
- Properties of Cramer's V:
  - Values between [0,1]
  - Interpretation:
    - Cramer's V = 0 → no association between X and Y
    - Cramer's V = 1 → perfect association between X and Y

Crosstables: Example (1/3)

- Testing a hypothesis: Respondents (which happen to be students) live or do not live with parents depending on sex.
- Recode the kind of residences into a dummy variable:

1. two categories yes/no "lives with parents":

```
generate r_parents = r_residence
recode r_parents  1 = 1  2/4 = 0
```

2. allocate name and labels:

```
lab var r_parents lives_with_parents
label define yesno 0 no 1 yes
lab val r_parents  yesno
```

3. checking by frequencies/tabulate:

```
tab r_parents
```

Crosstables: Example (2/3)

- Getting a cross table with column percentages and a Chi²-test of I independence

```
tab r_parents r_sex if vignr==1, col chi V
```

lives_with _parents	SEX male	female	Total
no	2	22	24
	15.38	44.90	38.71
yes	11	27	38
	84.62	55.10	61.29
Total	13	49	62
	100.00	100.00	100.00

```
          Pearson chi2(1) =   3.7720   Pr = 0.052
               Cramér's V =  -0.2467
```

Crosstables: Example (3/3)

- Interpretation:
  - Column percentage: 85 percent of male respondents live with their parents while (only) 55 percent of female respondents. Difference in percentage points is at about 30 percentage points.
  - Chi² square value is 3.77 with one degree of freedom.
  - p value is .052.
  - The null hypothesis can be rejected at a significance level of 0.10.
  - There is an association of medium size. (Cramers V = -.26).
  - Note: In cases of 2x2 tables Cramer's V is equivalent to phi coefficient. Depending on the table structure (main/secondary diagonal) positive or negative values are calculated.

Cross tabulation: Exercises

1. Check whether "living with parents" is depending on age. Formulate a reasonable research hypothesis and perform a chi square test of independence.
2. Think about an association of "living with parents" and a dummy variable for the "income level (median split)". Test a possible hypothesis.

Summary

1. Introducing the dataset and Stata
2. Descriptive statistics
3. Recoding variables, data management
4. Cross tabulation and Chi² test
5. Scatter plots and simple linear regression
6. Multiple linear regression
7. Group differences and interaction terms

Appendix

- Correlations
- Comparing mean values
- Analysis of variance
- Logic of controlling for confounding
- Logistic regression
- Non-linear effects

Scatterplots: Basic information

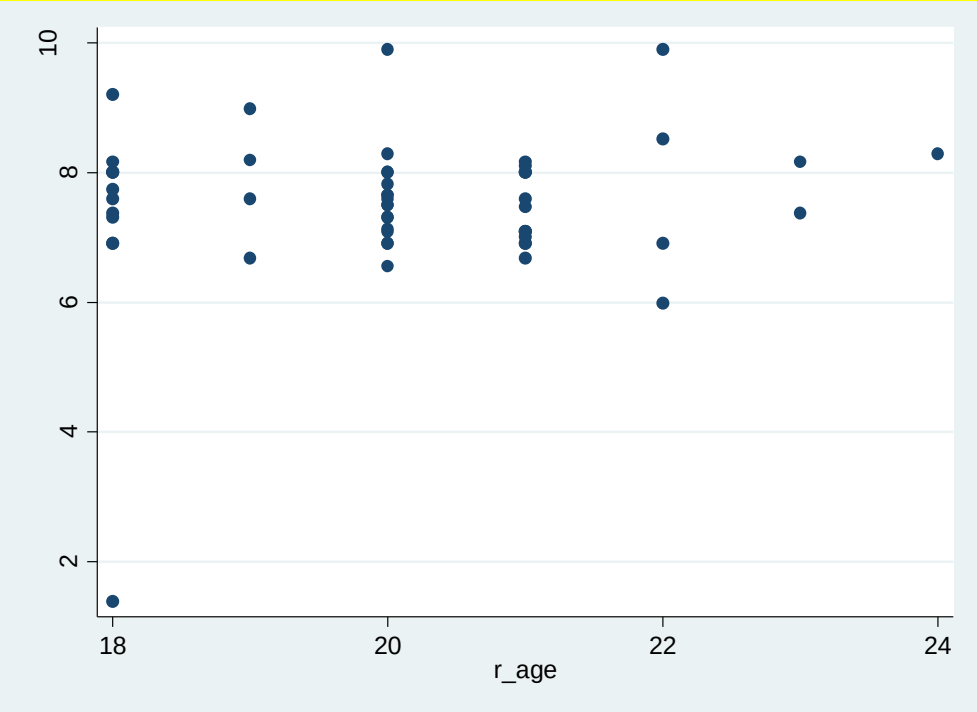
- Graphical illustration of bivariate distributions of metric variables..
- Conventions:
  - IV on the horizontal (X-)axis.
  - DV on the vertical (Y-)axis.
  - The pairs consisting of the values of the two variables are depicted as a collection of points (scatterplot) within the coordinate system.
- The pattern of the points gives information about the **nature**, **direction** and **strength** :
  - Linear / curvilinear
  - Positive / negative
  - Strong / week
- Stata command:

```
sca tter depvar indepvar
```

Scatterplot: Example

- Does age positively "influence" the respondents' income? Often income is logged. [Age is calculated by the difference between 2015 and r_year]

```
scatter  r_ln_income r_age  if vignr==1
```



Scatterplot does not point to a linear correlation.

Note the outlier at the left bottom. It is very influential! Also double check such cases!

Regression analysis: Basic information

- Regression analysis offers following advantages compared to the methods presented so far:
  - Predicting individual values of the dependent variable on the basis of individual values of the independent variable becomes possible.
  - Controlling for one or more confounding variables. One is frequently interested in *several* independent variables which jointly influence the dependent variable.
- Application and form:
  - Simple linear regression analyzes the correlation between a metric dependent and one metric or dichotomous independent variable.
  - Multiple linear regression is an extension of simple regression in which more than one independent variable is considered. This gives the opportunity to control for confounding factors.
  - Logistic regression is an alternative model which is used if the dependent variable is dichotomous.

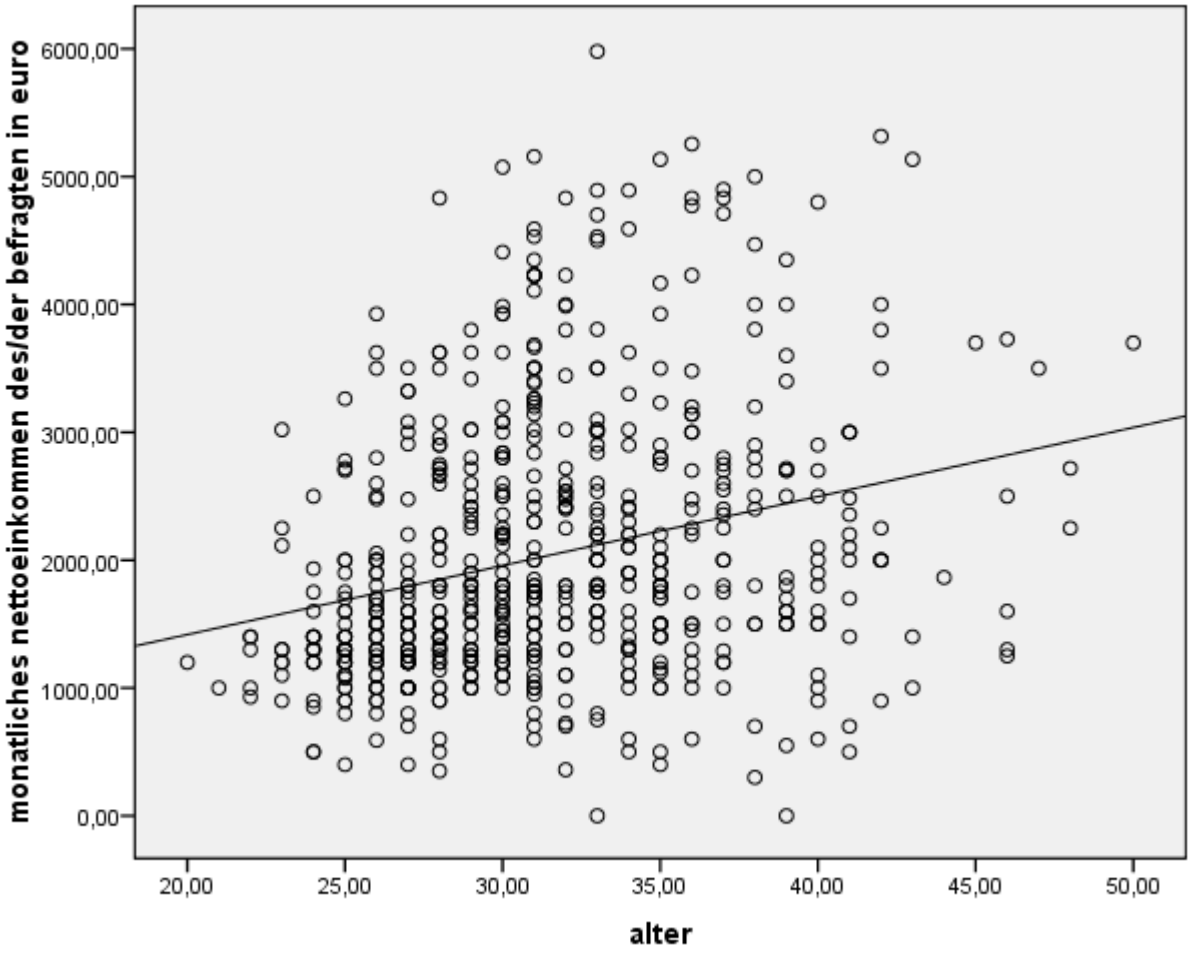
Simple linear regression: Basic idea (1/2)

- The idea of linear regression is to put a straight line through the collection of data points which best „summarizes" a linear relationship. In other words: which explains the correlation between the two variables as good as possible. A straight line is chosen if you presume the correlation to be **linear**.
- Example: Association between age and income in a study on labor market success of university graduates.

- Y axis: monthly income
- X axis: age in years

Simple linear regression: Basic idea (2/2)

Net income/month in Euro



Age in years

Simple linear regression: Formalization

- Such a straight line can easily be described mathematically:

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

- As one can see, not all the points (or even only a few) are to be found on the line. That's why a so called error term (often ε) is added to the equation (also designated as residual):

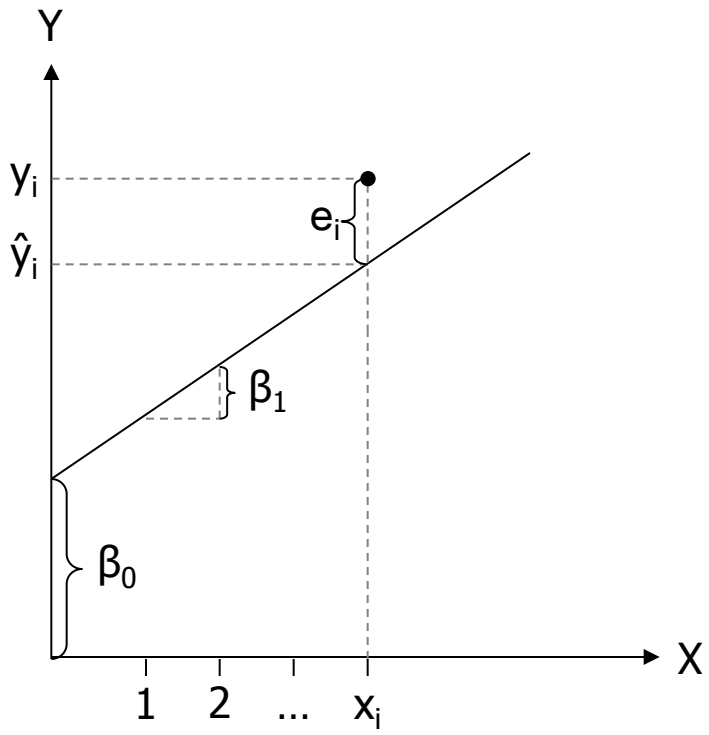
$$y = \beta_0 + \beta_1 x + e$$

- The deviance results from measurement errors as well as from the fact that the dependent variable Y can almost never be completely explained by the independent variable X (because there are other variables that have an influence on Y too).
- The line is laid through the collection of points so that the sum of squared residuals is as small as possible. The equation to solve this problem is:

$$\min \sum_{i=1}^{n} e_i^2$$

- The residuals have to be squared because otherwise positive and negative deviance would cancel out. Furthermore larger deviance should have a stronger influence on the calculation.
- For that, the expected value of the residuals is equal to 0, meaning that the error's average equals to 0: E(e)=0.

Simple linear regression: Proceeding and notation



e_i: Residual of studied object i, for example the difference between estimated and real income of person i.

X: Independent variable (regressor), e.g. age

Y: Dependent variable (regressand), e.g. income

x_i: Value of the independent variable for studied object i, for example person i's age

y_i: Value of the dependent variable for studied object i, for example person i's income

$\hat{y}_i$: With the regression function estimated value of Y for studied object i, for example expected income of a person at a certain age

β_0: Intercept corresponding to the value of the dependent variable if the independent variable equals 0, e.g. income of a person at age 0

β_1: slope parameter, corresponding to the value the dependent variable changes if the independent variables changes one unit.

Simple linear regression: Goodness of fit (1/2)

- After estimating the parameters of the regression equation, one question appears: How well does the straight line fit the data?
- For that: calculation of coefficient of determination R^2.
- R^2 is based on splitting up the total sum of squares: the total variation of the real Y-values can be split into variation explained by the regression and variation not explained by the regression

TSS	= ESS	+ RSS
(*Total Sum of Squares*	*= Explained Sum of Squares*	*+ Residual Sum of Squares)*
$TSS = \Sigma(Y_i - \overline{Y})^2$	$ESS = \Sigma(\hat{Y}_i - \overline{Y})^2$	$RSS = \Sigma e_i^{\,2}$

- R^2 results as the share of the explained variation in the total variation.

Simple linear regression: Goodness of fit (2/2)

- $R^2 = \frac{ESS}{TSS} = \frac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2} = 1 - \frac{\Sigma e_i^2}{\Sigma(Y_i - \overline{Y})^2} = 1 - \frac{RSS}{TSS}$
- R^2 is a share. Therefore R^2 values are between 0 and 1.
- Interpretation of R^2:
  - $R^2 = 0$ → the regression function does not contribute to explaining Y
  - $R^2 = 1$ → the regression function explains Y perfectly (all points lay exactly on the straight line)
  - Example: $R^2 = 0{,}35$ → 35% of the variance are explained by the independent variable.

Simple linear regression: Interpretation

- Significance F-Test: → The entire model contributes somewhat to explaining the dependent variable.
- R^2: Share of variation which is explained by the independent variable; says something about the model's goodness of fit.
- Positive (negative) sign in front of the coefficient: Independent variable has a positive (negative) influence on the dependent variable.
- Significance of the coefficients: If the values of the coefficients are significant, they differ significantly („more than just by chance") from zero; a significant correlation between this variable and the dependent variable exists.
- Value of the coefficient: In simple linear regression, unstandardized coefficients indicate by how many units **in average** the dependent variable increases (or decreases) if the independent variable increases by one unit.
- Intercept: indicates the expected value of the dependent variable if all independent variables are zero.

Summary

1. Introducing the dataset and Stata
2. Descriptive statistics
3. Recoding variables, data management
4. Cross tabulation and Chi² test
5. Scatter plots and simple linear regression
6. Multiple linear regression
7. Group differences and interaction terms

Appendix

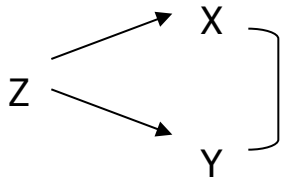
- Correlations
- Comparing mean values
- Analysis of variance
- Logic of controlling for confounding
- Logistic regression
- Non-linear effects

Multiple linear regression: Application

- Multiple linear regression is an extension of simple linear regression which allows for including several independent variables simultaneously into the analysis.
- This has the advantage to „control" for confounding factors: Does an correlation remain if supplementary independent variables are included? Or is it possibly a spurious relationship or „interpretation" (-> next slide)?
- The basic idea is the same as for simple regression, however, we now look at several dimensions: If there are two independent variables one can imagine regression as laying a surface throughout three-dimensional collection of points.

Possible influences of confounding factors

1



= Explanation

(Z leads to X, Y)

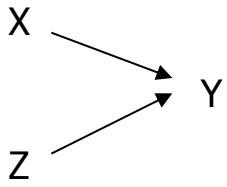
2

X → Z → Y

= Interpretation

(X leads to Z, Z leads to Y)

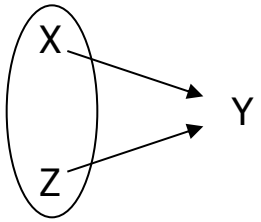
3



= Multicausality

(X, Z lead to Y)

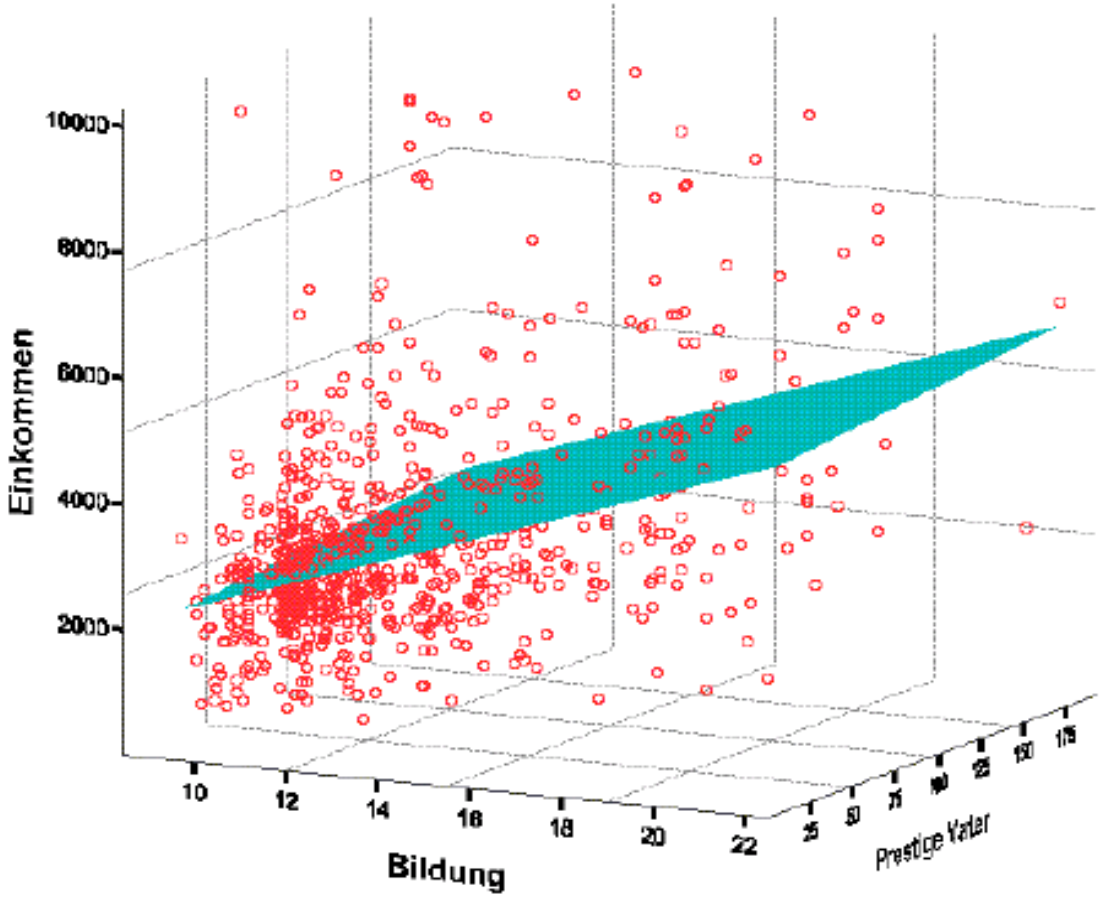
4



= Interaction

(X, Z lead to Y)

Multiple linear regression: Illustration



(Quelle: Brüderl, Josef (2004): Skript zur Vorlesung multivariate Analyseverfahren: 18)

Y: income X_1: years of education X_2: SES of father

Multiple linear regression: Proceeding

- Assumption: Dependent variable (y) and all independent variables (x_1, x_2, x_3,…) are additively and linearly linked:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- As for simple regression an additional error term e (epsilon) is introduced because y can't be perfectly predicted on the basis of the independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

- The problem to solve remains:

$$\min \sum_{i=1}^{n} e_i^2$$

- To keep expressions simple, estimators are derived in matrix notation. The derivation will not be covered here. You can read about it in appropriate textbooks (e.g. Greene 1993).

Multiple linear regression: Interpretation (1/2)

- Interpretation is mostly analogous to simple linear regression :
  - F-Test significant: The entire model contributes somewhat to explaining the dependent variable.
  - R^2 indicates how many percent of the variation in the dependent variable are explained by the independent variables.
  - But: R^2 increases with every additional variable, even if this variable doesn't contribute to the explanation→ adjusted R^2 takes this into account. That's why adjusted R^2 is used as measure for goodness of fit if we have many independent variables (you can't interpret this as explained variation anymore!).
  - Regression coefficients indicate the influence of a specific independent variable **while controlling for all other independent variables** (that's the advantage of multiple regression analysis: Many confounding factors can be controlled for at the same time).

Multiple linear regression: Interpretation (2/2)

- Standardized regression coefficients (so called „beta coefficients“) measure by how many standard deviations the dependent variable changes if the concerned independent variable changes by one standard deviation (better for comparison of the influence‘s size when variables have different units)
- Note: Regressions are based on several assumptions (see appendix). If these assumptions are not met the results might possibly be biased!

Multiple linear regression with pretest data

- Stata- command:

  regression *depvar varlist*

- Categorial variables are included by:

  xi: reg *depvar i.catvar*

A categorial variable with k levels will be used as k-1 dummies

For the first time, we use the vignette judgments of fairness evaluation as dependent variables and the vignette variables as independent variables.

Exercise and refresh (for the afternoon session)

1. Cross check the correlation structure of vignette variables. Why is this important?
2. Where does the correlation come from?

V10. A single woman. She is unemployed and not looking for work. She has one child aged 4 years. She lives in your region since many years. She identifies herself strongly as Ukrainian.Her level of income is 1600 UAH per month.

Multiple linear regression: Example (1/3)

```
. reg vig_judge1 i.gender i.labormarket i.maritalstatus i.children i.idp i.identity i.income
```

Source	SS	df	MS		
Model	1849.05556	18	102.725309	Number of obs	= 650
Residual	5549.4429	631	8.79467972	F(18, 631)	= 11.68
				Prob > F	= 0.0000
Total	7398.49846	649	11.3998435	R-squared	= 0.2499
				Adj R-squared	= 0.2285
				Root MSE	= 2.9656

vig_judge1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gender						
female	.0492403	.2446524	0.20	0.841	-.4311911	.5296718
labormarket						
unemployed and looking for work	1.220862	.3983296	3.06	0.002	.4386503	2.003074
unemployed and not looking for work	-.8362481	.3967929	-2.11	0.035	-1.615443	-.0570537
disabled, not able to work	3.183453	.4192851	7.59	0.000	2.36009	4.006816
student	.7503753	.4275147	1.76	0.080	-.0891484	1.589899
retired	1.265863	.4633442	2.73	0.006	.3559793	2.175746

Multiple linear regression: Example (2/3)

- Interpretation:
  - Pretest data show that around 22 percent of respondents' judgements are explained by vignette variables. Here, only a **part of the results** is shown.
  - The gender of vignette person seems to be irrelevant for the judgement. The coefficient is tiny and not significant.
  - Recall that higher values the DV mean that the state should help the vignette person by social support.
  - Recall as well that "working full time" is the category of reference category for labor market status. A vignette person who is unemployed but looking for work is seen as more deserving the social support compared to a person working full-time. The effect size is 1.22 (significant).
  - An unemployed person not looking for work however should not qualify for social support. The coefficient has a negative sign.
  - The most deserving group is "disabled not able to work".

Multiple linear regression: Example (3/3)

- Interpretation (continuation):
  - F-Test shows a significant fit of the model (overall).
  - Again: Explained variance is quite OK.
  - For the moment, we do not consider the nested data structure. Please wait for tomorrow.

Linear regression: Exercises for afternoon

1. Please go through all vignette variables and describe their influence on the vignette judgements. Consider the reference categories carefully.
2. What does the "identity" condition with "no information" mean?

maritalstatus						
married to an employed spouse	-.3712017	.2959277	-1.25	0.210	-.952324	.2099206
single	.1856502	.2982868	0.62	0.534	-.4001049	.7714052
children						
child 4 years old	-.1326691	.3533848	-0.38	0.707	-.8266217	.5612835
child 14 years old	-.3597435	.3783054	-0.95	0.342	-1.102633	.3831464
without children	-1.548962	.3540723	-4.37	0.000	-2.244265	-.8536597
idp						
moved for personal reasons	-.7521151	.310811	-2.42	0.016	-1.362464	-.1417661
lives in region	-.8861742	.2905676	-3.05	0.002	-1.456771	-.3155777
identity						
identifies barely as Ukrainian	-.869634	.2955863	-2.94	0.003	-1.450086	-.2891821
	.0352046	.2956089	0.12	0.905	-.5452916	.6157008
income						
1600	-.2446138	.3401598	-0.72	0.472	-.912596	.4233684
2400	-1.348433	.3526262	-3.82	0.000	-2.040896	-.6559698
3200	-1.57428	.3427031	-4.59	0.000	-2.247257	-.9013035
_cons	1.926029	.5579058	3.45	0.001	.8304522	3.021606

Regression: wrap up

- Purpose of multiple regression: Testing the influence of several metric or dichotomous independent variables on metric dependent variable (→ controlling for confounding factors) .
- General proceeding: Extending an equation of a straight line (linear regression) by additional independent variables, minimizing the sum of squared residuals.
- Interpretation:
  - R^2: Share of explained variation, adjusted R^2 takes the inclusion of additional independent variables into account.
  - F-Test: Tests H_0 that all IV have no influence versus H_1 that at least one IV does have an influence.
  - Coefficients: While changing the IV by one unit the dependent variable (DV) changes by ß units.
  - Sign of coefficient indicates direction of the influence; using coefficient and standard error you can calculate the t value which gives information about the significance of specific IV.

Summary

1. Introducing the dataset and Stata
2. Descriptive statistics
3. Recoding variables, data management
4. Cross tabulation and Chi² test
5. Scatter plots and simple linear regression
6. Multiple linear regression
7. Group differences and interaction terms

Appendix

- Correlations
- Comparing mean values
- Analysis of variance
- Logic of controlling for confounding
- Logistic regression
- Non-linear effects

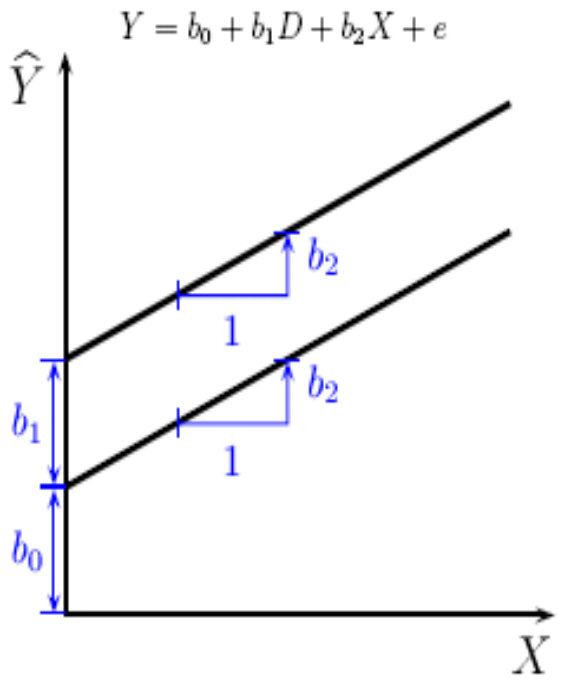
Interactions (1/3)

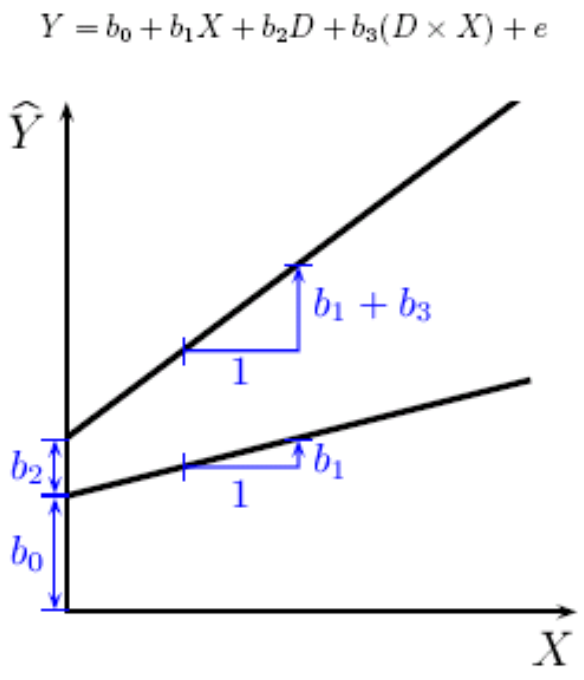
- „Interaction”: Effect of one variable depends on the value of another variable; with interaction terms one can for example test if sub-groups differ regarding the influence of a certain variable (e.g. if the influence of vignette variables is dependent on respondents’ characteristics).
- This can be tested by separate estimations for sub groups of respondents (male vs. female).
- Though there is the disadvantage of not being able to report about the significance of a possible difference.
- Adding “interaction terms” to the regression is a solution.
- Technically speaking: Adding a new variable which is generated as the product of the variables concerned (vignette variable*respondents’ sex).

Interactions (2/3)

- One has to check if the interactions term (vignette variable*respondents' sex) is significant, additionally to the "main effects" of vignette variables and respondents' sex.
- Depending on the level of measurement of the two variables one can distinguish different interaction terms (and interpretations). If it's the interaction between a metric and a binary variable, both slope and intercept differ. On the next slide this case is graphically illustrated.

Interactions (3/3)





Quelle: Stocker 2004: Einführung in die angewandte Ökonometrie. Unveröffentlichtes Vorlesungsskript der Universität Innsbruck: 181f.

Interaction Terms/Exercise for the afternoon

Please analyze whether there is an interaction effect (group difference) between the vignette dimension (children) and respondents' sex. You can use the possibility to easily integrate interaction terms in the regression models with the "#".

e.g. xi: reg *depvar i.catvar i.r_sex i.catvar#i.r_sex*

Summary

1. Introducing the dataset and Stata
2. Descriptive statistics
3. Recoding variables, data management
4. Cross tabulation and Chi² test
5. Scatter plots and simple linear regression
6. Multiple linear regression
7. Group differences and interaction terms

Appendix

- Correlations
- Comparing mean values
- Analysis of variance
- Logic of controlling for confounding
- Logistic regression
- Non-linear effects

Pearson's correlation coefficient (1/2)

- If a scatter plot indicates a **linear** correlation, Pearson's correlation coefficient ("Pearson's r") is the common measure to determine direction and strength of the correlation.
- Informs about the direction and the degree in which changes in one variable change the other variable (joint covariation, measured with the product of the standard deviations).
- Assumptions:
  - Both variables must be metric.
  - Linear correlation (scatter plot).
  - X und Y ought to be approximately normally distributed.
- Exception: point-biserial correlation coefficient (one of the variables dichotomous, that is a dummy variable).

Pearson's correlation coefficient (2/2)

- Properties:
  - Values: [-1; +1]
  - Interpretation:
    - 0 no *linear* correlation
    - 0 < |r| < 0.2 very weak
    - 0.2 < |r| < 0.4 weak
    - 0.4 < |r| < 0.6 modest
    - 0.6 < |r| < 0.8 strong
    - 0.8 < |r| < 1.0 very strong
    - 1 perfect linear correlation

Sign:

r > 0	positive linear correlation
r < 0	negative linear correlation

- Stata command (same for the point-biserial correlation coefficient):

```
pwcorrelations var var, sig
```

Pearson's correlation coefficient (example with pretest-data)

- Check if the "complexity" rating is correlated with respondents own labor market experience. Recode labor market status.

```
. corr complexity r_lm_exp
(obs=620)

             | comple~y r_lm_exp
-------------+------------------
  complexity |   1.0000
  r_lm_exp   |   0.1950   1.0000
```

Correlation is relatively small. Having labor market experience makes the vignette evaluation on average easier. Using the pwcorr command (with option `sig`) would show that this correlation is not significant in the pretest data.

Comparing means: Basic information (1/2)

- Question: Do the means differ significantly between groups? For example, do the average graduation marks differ between children with/without academic background?
- If the differences in the means are significant, one assumes the differences to exist in the corresponding population (statistical inference).
- In the following we will have a look at methods for metric, approximately normally distributed variables (for corresponding "non-parametric" alternatives: see textbooks on statistics).

Comparing means: Basic information (2/2)

- The choice of method depends on the attributes of the variables:
  - Two parameter values: t-Test
  - More than two parameter values: Analysis of variance

t-Test for independent samples: Example (1/2)

- Analyzing a sub group difference of vignette judgements by labor market experience (yesno).
- Stata-command: **`ttest`** *depvar*, **`by`**(*groupvar*)

```
. ttest vig_judge1,by(r_lm_exp)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	510	.772549	.1431247	3.23221	.4913612	1.053737
1	130	-.0384615	.3343581	3.812269	-.6999972	.6230741
combined	640	.6078125	.1332431	3.370813	.3461653	.8694597
diff		.8110106	.3298834		.163222	1.458799

```
    diff = mean(0) - mean(1)                                      t =   2.4585
Ho: diff = 0                     degrees of freedom =      638

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9929         Pr(|T| > |t|) = 0.0142          Pr(T > t) = 0.0071
```

t-Test for independent samples: Example (2/2)

- Interpretation:
  - Respondents without labor market experience have an average judgement of .77. Respondents with labor market experience show a lower average value (-.04). In general, labor market experience changes the evaluation towards a less friendly social support.
  - t-test is significant (t=2.4585; p=.0142).
  - Confidence interval: With a probability of 95% percent the interval from .16 until 1.46 contains the difference of means.
- Note that the conducted t-test assumes equal variances in both groups. This could be checked using a Levene-test. In case the assumption is not met, several modifications of the t-test exist (see Stata-help command `help ttest`).

Analysis of variance: Application

- Analysis of variance tests if the distribution of a **metrically scaled dependent variable Y** differs within different groups of the **nominally or ordinal scaled independent variable**. One has a look at the groups' mean values of variable Y (as well as the variance between and within the groups).
- Analysis of variance is a tool frequently used for experimental or quasi experimental studies. Groups result from the division into an experimental and a control group.
- Analysis of variance can be used to compare **social contexts** (classes, schools, federal states).

Analysis of variance: test of significance within the framework of ANOVA

- The following **null hypothesis** is tested: The **mean values** of all groups (g = 1, 2, …, G) are **the same**:

  H_0: $\mu_1 = \mu_2 = \mu_g$

- **Alternative hypothesis**: Fot at least two groups **the mean values are different**:

  H_1: $\mu_g \neq \mu_{g'}$

- For calculating the **test statistic** (F value) , one relates **the explained to the unexplained variance** (for details see e.g. Fahrmeir et al. 2000):

$$F_{emp} = \frac{MS_b}{MS_w}$$

- If $F_{emp} > F_{1-\alpha}$ for given df_b and df_w, we have to reject H_0 at significance level α.

One factorial ANOVA: Example (1/2)

- Stata command: **on**eway *depvar indepvar*, tab
- We analyze if judgements differ by (three) age groups.

```
. oneway vig_judge1 r_age_cat3, tab
```

3 quantiles of r_age	Summary of vig_judge1 Mean	Std. Dev.	Freq.
1	.18536585	3.2910608	410
2	1.50625	3.2330776	160
3	.7375	3.737405	80
Total	.57846154	3.3763654	650

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	203.105016	2	101.552508	9.13	0.0001
Within groups	7195.39345	647	11.1211645		
Total	7398.49846	649	11.3998435		

Bartlett's test for equal variances: chi2(2) = 2.6494 Prob>chi2 = 0.266

One factorial ANOVA: Example (2/2)

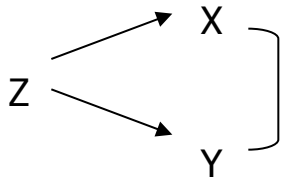
- Interpretation:
  - The average judgement is highest for the second age group. It is lowest in the youngest group.
  - F-Test shows that at least one mean difference is significant ($F = 9.13$ $df_b = 2$, $df_w = 647$; $p = .000$).
  - Bartlett-Test assumes the null hypothesis that the variance in all three groups is equal. This is a prequisite for an ANOVA model. Null hypothesis has not to be rejected (p=.266).

Comparing means: Exercises in the afternoon

1. Does the respondents' income vary by age groups (three levels)?
2. Does the respondents' age significantly differ by course?
3. Does the vignette evaluation differ by vignette position?

Possible influences of confounding factors

1



= Explanation

(Z leads to X, Y)

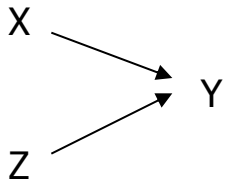
2

X → Z → Y

= Interpretation

(X leads to Z, Z leads to Y)

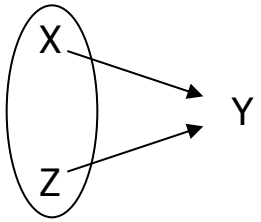
3



= Multicausality

(X, Z lead to Y)

4



= Interaction

(X, Z lead to Y)

Logic of controlling for confounding

- **Assumption:** Coefficient a describes strength and direction of the correlation between X and Y (e.g. Cramer's V or regression coefficient).
- The theoretically assumed confounding factor Z has for illustrative purposes two parameter values (z=1 and z=2).
- One compares the **amount a_0 (a_{xy})** between X and Y **without considering the confounding factor** with the **amount of a_1 (a_{xy} l z=1) and a_2 (a_{xy} l z=2)**, the correlation between X und Y **while considering each time just one category of Z**.
- For example one can look at regression estimations for separate groups.
- (Confounding can easily be taken into account while adding variables to a multiple regression: does the original correlation remain stable if one controls for further variables?)

Logic of controlling for confounding

	a_1 (a_{xy} | z=1)	a_2 (a_{xy} | z=2)	
Case 1: $a_0 \neq 0$	= 0	= 0	Explanation or interpretation
	< a_0	< a_0	Partial effect/ partial explanation, partial interpretation
	= a_0	= a_0	Multicausality
	> a_0 / < a_0	or < a_0 / > a_0	Interaction
Case 2: $a_0 = 0$	$a_1 \neq 0$	$a_1 \neq 0$	Suppression

Logistic regression

- So far: Linear regression to examine correlation between metric dependent and metric or dichotomous independent variables.
- The dependent variable is often not metric, but dichotomous:
  - Successful job hunting
  - Change of the employer
  - Participation in election
  - Breaking off the survey yes/no
- In this case OLS-regression is not appropriate.
- Logistic regression represents a method which allows for a suitable analysis of dichotomous dependent variables.
- Independent variables must be metric or dichotomous.

Logistic regression: Deriving the model (1/7)

- OLS regression is not used for the analysis of dichotomous dependent variables for different reasons, although it might seem appropriate at first glance.
- The dichotomous random variable Y_i, takes value 1 if a certain event takes place or a certain characteristic is present and value 0 if the characteristic does not take place or is not present.

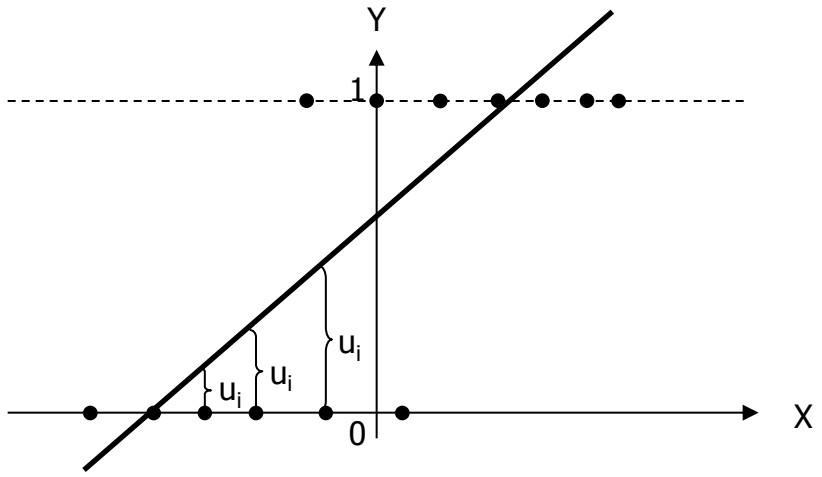
$$Y_i = \begin{cases} 1 \text{ event takes place} \\ 0 \text{ event does not take place} \end{cases}$$

- The probability distribution of such a variable is given by $P(Y_i = 1)$ if the parameter value 1 and by $1 - P(Y_i = 1)$ if the parameter value 0.
- By definition the expected value can be calculated as follows :

$$E(Y_i) = 0 \cdot [1 - P(Y_i = 1)] + 1 \cdot P(Y_i = 1) = P(Y_i = 1)$$

Logistic regression: Deriving the model (2/7)

- This expected value can be modelled as a linear function of independent variable, therefore use the known linear regression model (the so called linear probability model, LPM).
- Several problems emerge:
  - The linear regression model is defined for values from $[-\infty,+\infty]$, though probability can only take values between 0 and 1.
  - The error term is not homoscedastic as illustrated below:



Logistic regression: Deriving the model (3/7)

- Problems (to be continued):
  - The residuals are not normally distributed.
  - Furthermore there are content-based reasons for which the assumption of a linear correlation is not plausible. It can rather be assumed that there is a central area of X for which changes do have an especially strong impact (in this area the question for the values of Y_i is virtually answered) meanwhile changes at the poles have a much smaller impact.

Logistic regression: Deriving the model (4/7)

- The logistic distribution function offers a solution. It is defined by:

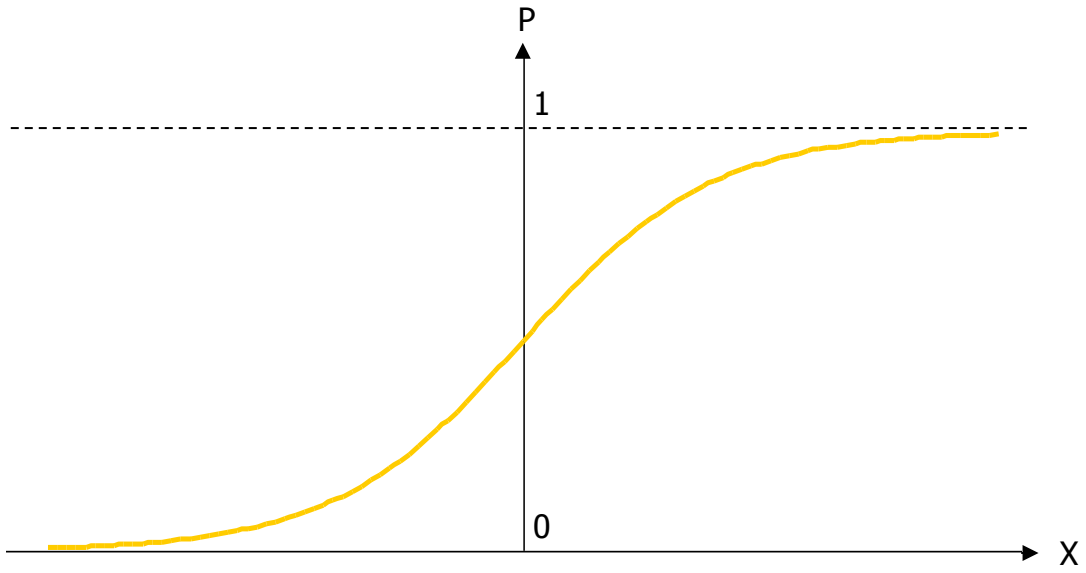
$$Y = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

- If one applies this to the estimation of the expected value and parameterises the course of the function, one gets the following model:

$$E(Y_i \mid X_{i1},\ldots, X_{im}) = P_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{i1}+\ldots+\beta_m X_{im})}} = \frac{1}{1+e^{-Z_i}}$$

Logistic regression: Deriving the model (5/7)

- Graphically the function can be drawn like this:



Logistic regression: Deriving the model (6/7)

- The equation is not linear which makes estimation difficult. Though it can be linearized:

$$P_i = \frac{1}{1+e^{-Z_i}} \Rightarrow 1-P_i = 1-\frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{Z_i}}$$

Converting leads to:

$$e^{Z_i} = \frac{P_i}{1-P_i}$$ } Odds

resp. $$\frac{P_i}{1-P_i} = e^{Z_i} = e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_m X_{im}}$$

Taking the logarithm gives us:

$$\underbrace{\ln\left(\frac{P_i}{1-P_i}\right)}_{\text{Log-Odds/ Logit}} = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_m X_{im} = Z_i$$

Logistic regression: Deriving the model (7/7)

- The expression $P_i/(1-P_i)$ is called *Odds*, the expression $ln(P_i/(1-P_i))$ *Log-Odds* or *Logit,* shortly *L.*
- This model can't be correctly estimated with OLS.
- Consequently another estimation method is used: Maximum Likelihood (ML).
- The ML principle of estimating parameters is more frequently applicable than OLS estimation. It is the most common method of estimation.

Excursion: Maximum likelihood estimation MLE

- Basic principle of the ML method: Determine the unknown estimation parameters thus the probability of getting the observed values of the dependent variable is as high as possible
- ML estimators are
  - Asymptotically unbiased,
  - Asymptotically consistent and,
  - Asymptotically efficient.
- Asymptotic means that the properties only apply for many observations. Unfortunately there is no clear reference in common textbooks what "many" means. Some authors recommend to calculate ML estimators with only more then 50 cases. Others recommend more then 100 or 30-50 cases per independent variable.
- The exact estimation method is to be found in common statistics' textbooks. As a rule it is not possible to solve the equations analytically, so iterative methods are applied.

Logistic regression: Interpretation (1/2)

- The interpretation of the coefficients isn't as easy as for OLS regression.
- Different approaches can be useful:
  - **Direction of the effect:** A positive sign for the coefficient means a positive influence on the log odds and thereby on the probability of the event to happen.
  - **Odds Ratios:** By using the e-function of the coefficient, one gets so called odds ratios which are frequently reported. One can calculate the percentage changing of the odds (not probabilities!) for a one unit change in an independent variable. This is given by:

$$(e^{\beta_j} - 1) \cdot 100$$

  - The advantage of odds ratios is that they are independent from the values of the independent variables. Though they aren't very easy to interpret neither. That's different for marginal effect.

Logistic regression: Interpretation (2/2)

- **Marginal effects:** They indicate the change of probability if there is a little change in the independent variables. This is (at least for continuous variables) easy to interpret. But the values are dependent from the values of the independent variable. Marginal effects are frequently calculated for the mean value of all independent variables.
- **Probability effects:** First the expected probability that the dependent variable takes value 1 (event does take place) is calculated for a certain combination of values of the independent variables. Then you calculate the expected probability after changing the interesting independent variable for one unit while the other variables are kept constant. The difference of the two expected probabilities is called probability effect.

Logistic regression: Goodness of fit

- There is no measure such as R^2 which exactly corresponds to OLS regression.
- But there are measures which follow in certain ways the definition of the determination coefficient. In SPSS outputs you find Cox-Snell-R^2 and Nagelkerkes R^2.
  - Cox and Snell´s R^2:

$$R^2_{CS} = 1 \left(\frac{L_0}{L_1} \right)^{\frac{2}{n}}$$

with L_0 : Likelihood of the so called null model, model without covariates

L_1 : Likelihood of the model under consideration

  - Nagelkerkes R^2:

$$R^2_N = \frac{R^2_{CS}}{1 \quad (L_0)^{\frac{2}{n}}}$$

  - R^2 according to Cox and Snell takes values between $[0; 1-(L_0)^{2/n}]$, so it can't reach value 1; Nagelkerkes R^2 adjusts the values to [0,1].

Logistic regression with our data

- Stata command for logistic regression:

  `logit` `depvar indepvar`

Logistic regression: example (1/2)

- We analyze if "living with parents" is determined by age, sex and income of respondents' household. Think about clear-cut hypotheses!

```
. logit r_parents r_age r_female r_ln_income if vignr==1

Iteration 0:   log likelihood = -35.105014
Iteration 1:   log likelihood = -29.235393
Iteration 2:   log likelihood = -29.036286
Iteration 3:   log likelihood = -29.034354
Iteration 4:   log likelihood = -29.034354

Logistic regression                               Number of obs     =         51
                                                  LR chi2(3)        =      12.14
                                                  Prob > chi2       =     0.0069
Log likelihood = -29.034354                       Pseudo R2         =     0.1729

------------------------------------------------------------------------------
   r_parents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       r_age |  -.4027296   .2251225    -1.79   0.074    -.8439616    .0385023
    r_female |  -1.135728   .8984616    -1.26   0.206     -2.89668    .6252248
 r_ln_income |   .9962432   .5070978     1.96   0.049     .0023498    1.990137
       _cons |   1.775749   5.512716     0.32   0.747    -9.028976    12.58047
------------------------------------------------------------------------------
```

Logistic regression: example (2/2)

- Here are the "odds ratios"!

```
. logit r_parents r_age r_female r_ln_income if vignr==1,or

Iteration 0:   log likelihood = -35.105014
Iteration 1:   log likelihood = -29.235393
Iteration 2:   log likelihood = -29.036286
Iteration 3:   log likelihood = -29.034354
Iteration 4:   log likelihood = -29.034354

Logistic regression                               Number of obs     =         51
                                                  LR chi2(3)        =      12.14
                                                  Prob > chi2       =     0.0069
Log likelihood = -29.034354                       Pseudo R2         =     0.1729

------------------------------------------------------------------------------
   r_parents | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       r_age |   .6684928   .1504927    -1.79   0.074     .4300037    1.039253
    r_female |   .3211883   .2885754    -1.26   0.206     .0552062    1.868666
 r_ln_income |   2.708089   1.373266     1.96   0.049     1.002353    7.316533
       _cons |   5.904703   32.55095     0.32   0.747     .0001199    290824.2
------------------------------------------------------------------------------
```

Logistic regression: example (3/3)

- Interpretation:
  - By rising of respondent's age the probability of living with parents decreases (significant at 10 percent level).
  - There is a non-significant tendency that female respondents do not live with their parents.
  - The higher the income per household member, the higher the probability of living with parents. But think about causality!
  - Pseudo-R^2 (a measure which gives information on model fit to data) is about 0,17 – not that bad. But we have only 51 respondents…

Linear regression: Assumptions (1/2)

- Linear regression is based on several assumptions, e.g.:
  - Linearity in the parameters (but not especially in the variables).
  - Variance of the residuals the same for all values of X (homoscedasticity), that means $var(u_i|X_i)=\sigma^2$ for all *i.*
  - For every two values X_i and X_j the correlation between the corresponding errors is zero (no autocorrelation), meaning that $cov(u_i,u_j|X_i,X_j)= 0$ for all $i \neq j$.
  - Regressor (X) und error terms are not correlated, meaning $cov(u_i,X_i) =0$ for all *i.*
  - The number of observations must be bigger than the number of parameters.
  - The model must be correctly specified, e.g. all relevant explaining variables have to be included.
  - For models with more than one independent variable there is no perfect multicollinearity, that means no perfect connection between the independent variables.
  - For statistical testing we need the assumption of normally distributed error terms (but not for estimation itself).

Linear regression: Assumptions (2/2)

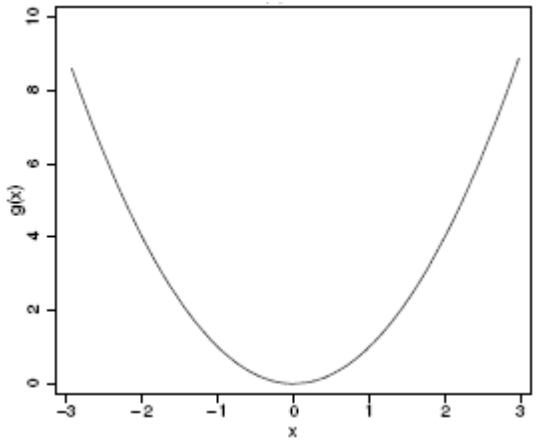
- If all assumptions are satisfied, OLS estimators are BLUE (Best Linear Unbiased Estimator), which means
  - They are unbiased, correspond "on average" the true parameters of the population
  - Among all given estimators they have the smallest variation (efficiency).
- Furthermore the estimators are consistent, that means they approach the true population parameter with increasing sample size.
- If the assumptions are not satisfied the values of the coefficients might be estimated wrong or statistical testing can lead to wrong conclusions. There are some remedies for this which can't be explained here because of lack of time. You can read about them in common textbooks (e.g. Gujarati 2002; Kohler/Kreuter 2006).

Modelling of non-linear relationships (1/2)

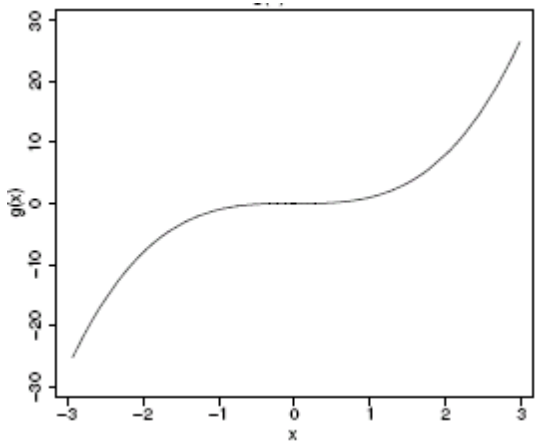
- By adding polynomials (e.g. squared variables) to regression, non-linear relationships can be modelled.
- In terms of proceeding this means to transform the corresponding variable (e.g. square it) and add this transformed variable to the regression
- Take into account: The effect of the corresponding variable isn't linear anymore, but depends on its concrete values.

Abb.1: nichtlineare Transformationen mittels Polynomen

$g(X) = X^2$



$g(X) = X^3$

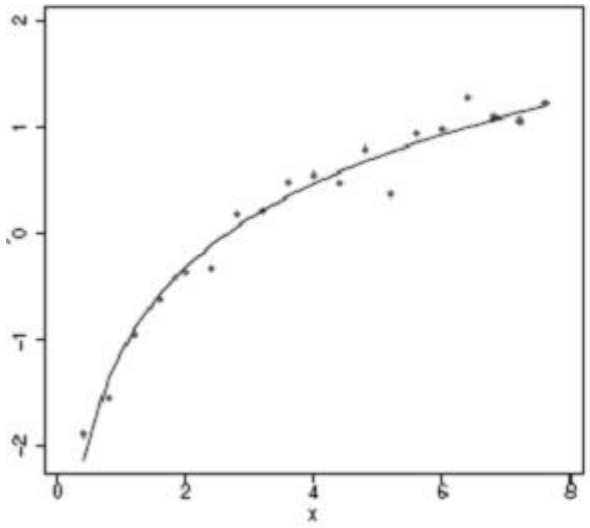


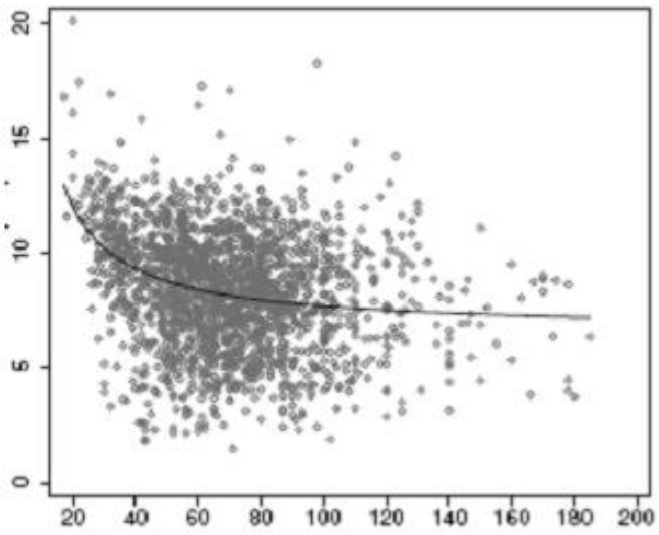
Modelling of non-linear relationships (2/2)

Squared terms: Quite frequently, one is interested in modeling (inverted) u-shape relations.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X^2_1$$

If $\beta_1 > 0$ and $\beta_2 < 0$: inverted u-shape with decreasing rate
If $\beta_1 < 0$ and $\beta_2 > 0$: u-shape with decreasing rate





References

- Brüderl, Josef (2004): Multivariate Verfahren. Unveröffentlichtes Skript zur gleich lautenden Vorlesung an der Universität Mannheim. Available at: http://www.sowi.uni-mannheim.de/lehrstuehle/lessm/veranst/MultiVorlesung.pdf.
- Diekmann, Andreas (2007): Empirische Sozialforschung. Rowohlt: Reinbek bei Hamburg.
- Jann, Ben (2002): Einführung in die Statistik. München, Wien: Oldenburg.
- Fahrmeir, Ludwig/ Künstler, Rita/ Pigeot, Iris/ Tutz, Gerhard (2000): Statistik. Der Weg zur Datenanalyse. Berlin u.a.: Springer.
- Gujarati, Damodar N. (2002): Basic Econometrics. 4. Aufl. New York: Mcgraw-Hill.
- Kohler, Ulrich/ Kreuter, Frauke (2006): Datenanalyse mit Stata. Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung. München, Wien: Oldenburg.
- Urban, Dieter/Mayerl, Jochen (2008): Regressionsanalyse. Theorie, Technik und Anwendung. Wiesbaden: Verlag für Sozialwissenschaften.
- Tutz, Gerhard (2000): Die Analyse kategorialer Daten. Anwendungsorientierte Einführung in Logit-Modellierung und kategoriale Regression. München, Wien: Oldenburg.