Regression Analysis

Methodology of Conflict and Democracy Studies

January 11

Aim of this lecture

- How to do regression analysis

- Which regression analysis to choose:
  - Linear (OLS – Ordinary Least Squares) regression
  - Logistic regression

- Interpretation of the results

Regression Analysis

- A variety of techniques with the same aim

- Identification of effects of one or more IVs on DV

- What it allows:
  - Identify effect of each independent variable
  - Control of effects of other independent/control variables
  - Predict values of DV based on specific values of IVs

Which Regression?

- Everything depends on your dependent variable

- Linear (OLS) regression:
  - Scale variable (or long ordinal)

- Logistic regression:
  - Binary variable (0/1) – binary logistic regression
  - Nominal (0/1/2/3) – multinomial logistic regression

- No limits on independent variables (all types allowed)

Examples

- OLS regression:
  - How do age, gender and education affect income of people?
  - Does attendance on lectures increase % amount of obtained points in your courses?

- Logistic regression:
  - Do men have higher chances to end up in jail than women?
  - Does attendance on lectures increase your chances of avoiding F in a course?

OLS Regression - Requirements

- Dependent variable:
  - Exactly one variable, normal distribution

- Independent variable:
  - One or more variables, all types without limits

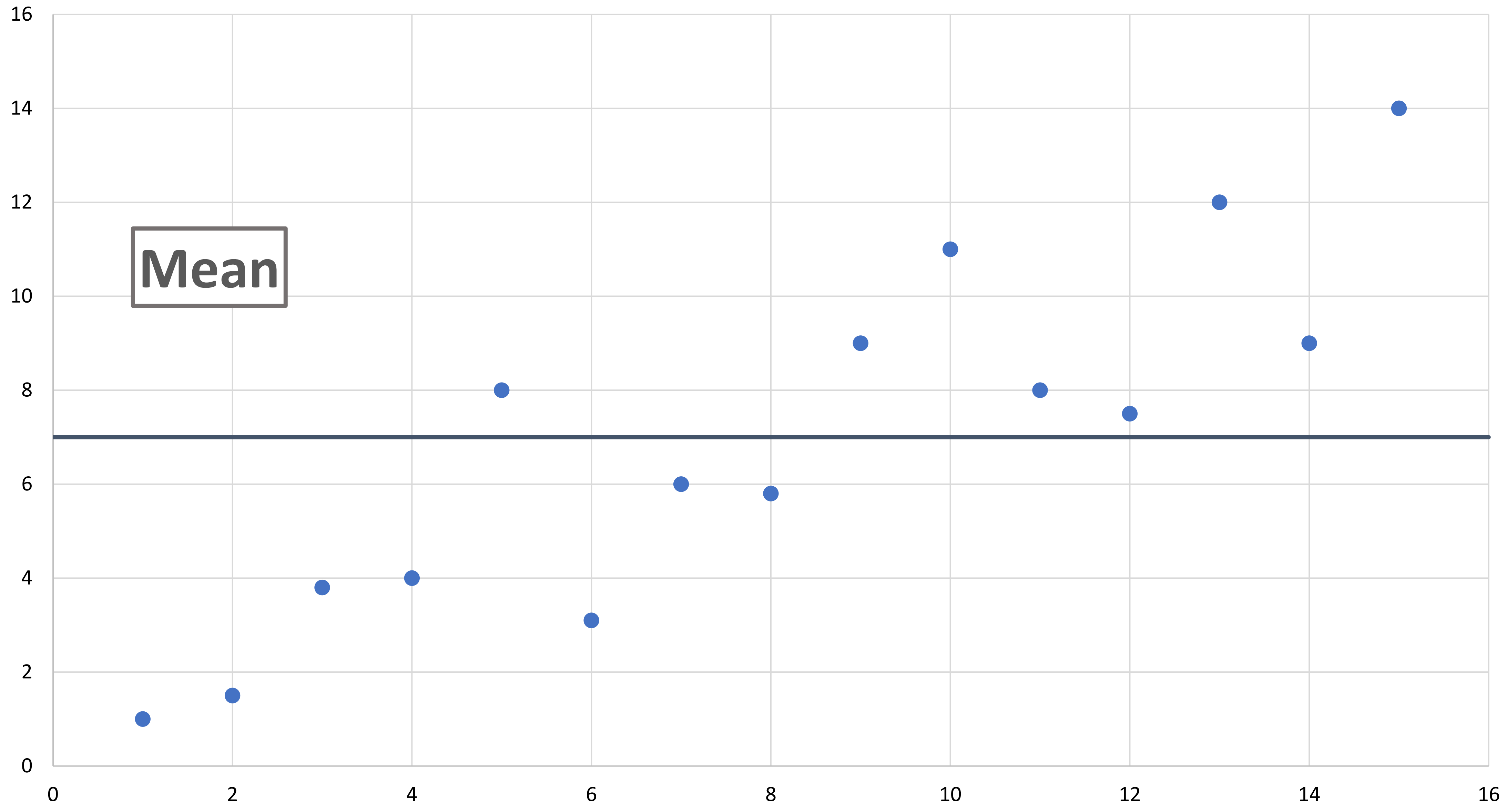
- Further requirements:
  - Independence of observations
  - No collinearity between independent variables
  - Linear relationship between IVs and DV
  - Homogeneity of distribution of residuals

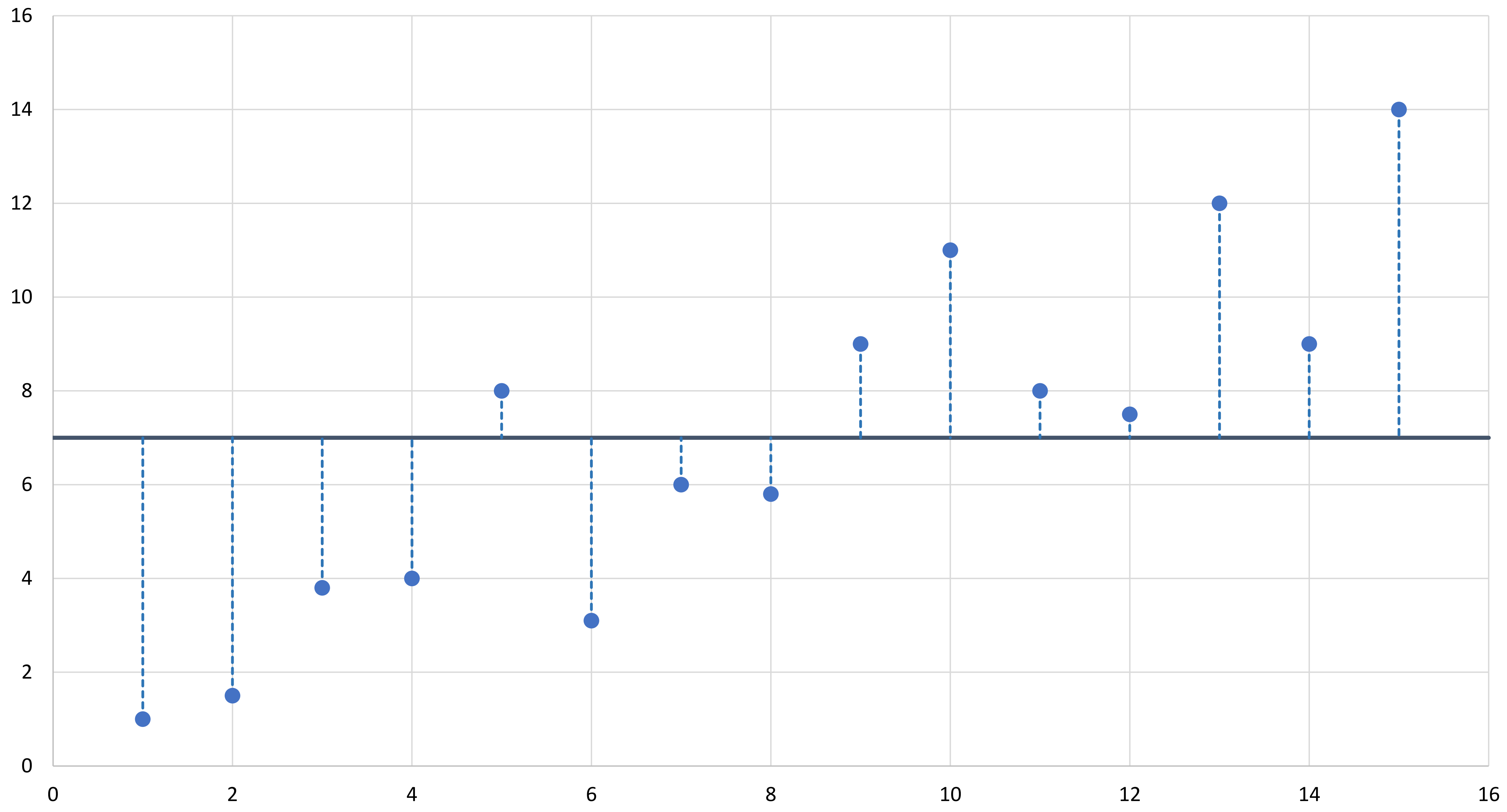
What is OLS Regression about?

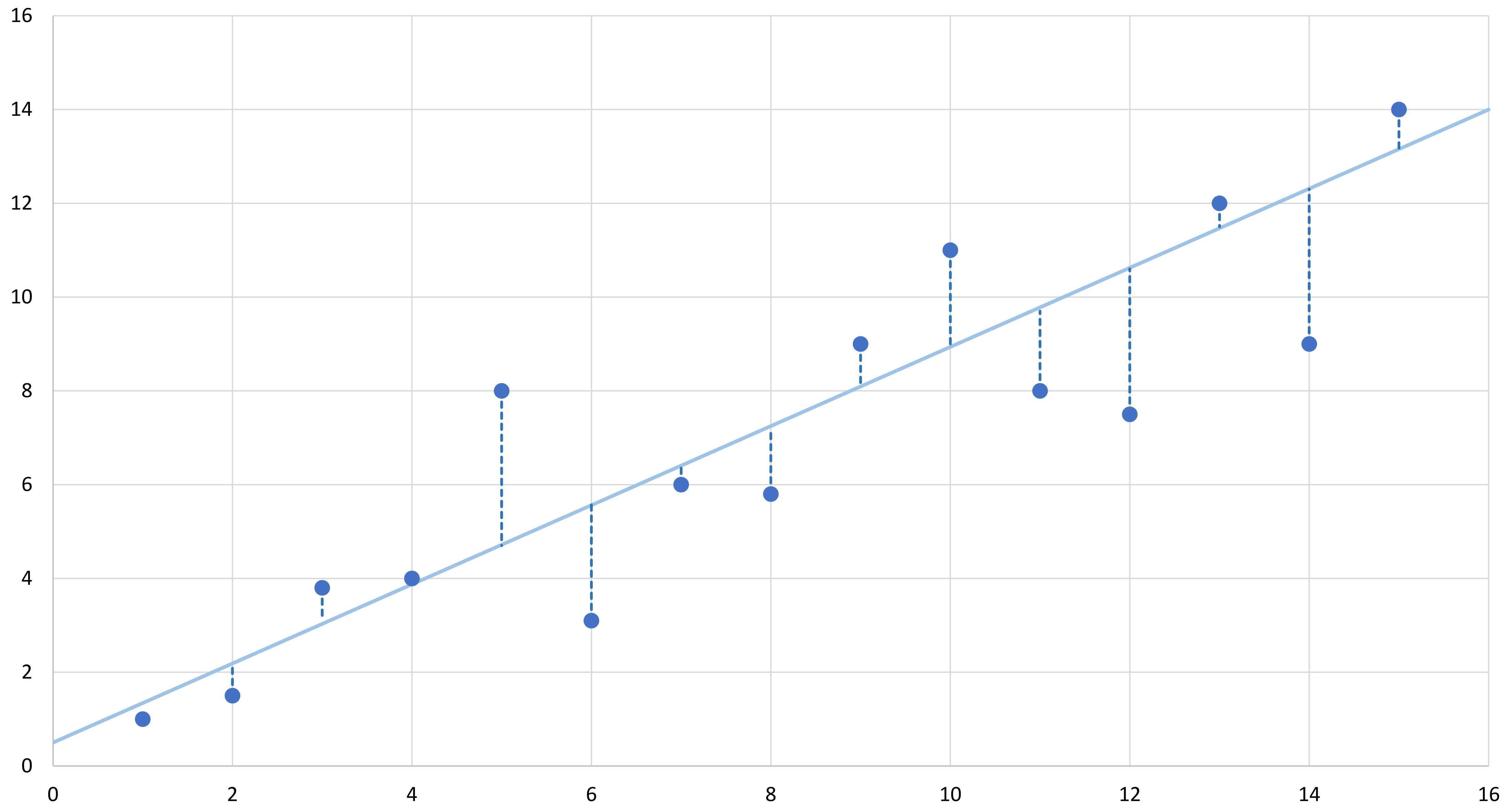
- Basically OLS regression is about searching for ideal lines that best describe the relationship between independent and dependent variable

- The best line is the one that is the least inaccurate of all possible lines

- Accuracy measured using sum of squares of vertical differences between predicted and observed data


Mean
0
2
4
6
8
10
12
14
16
0
2
4
6
8
10
12
14
16


16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16


16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16

The Outcomes of OLS Regression

- OLS regression estimates:
  - Intercept
  - Effects of each independent variable

- $y = b_0 + b_1 * x + b_2 * y + b_3 * z + \ldots$

- ***y*** stands for predicted value of dependent variable
- $\boldsymbol{b_0}$ stands for intercept
- $\boldsymbol{b_1}$, $\boldsymbol{b_2}$, $\boldsymbol{b_3}$ etc. stand for slopes of independent variables ***x, y, z*** etc.

R square

- Provides information about the overall fit of the model
- How well our model (= our IVs) explain the dependent variable
- Comparison of improvement of regression line compared to mean

- Ranges from 0 to 1 (zero to hundred per cent)

- Show how much of the variance of dependent variable we are able to explain using our set of independent variables
- Use Adjusted R square to control for inflation of number of IVs

Intercept (Constant)

- The predicted value of dependent variable if the values of all independent variables are zero

- $y = \mathbf{b_0} + b_1 * x + b_2 * y + b_3 * z$

- If x, y, z etc. = 0 then

- $y = \mathbf{b_0} + b_1 * 0 + b_2 * 0 + b_3 * 0$
- $y = \mathbf{b_0}$

Outcomes Concerning Independent Variables

- **Unstandardized B coefficient:**
  - Shows how the value of dependent variable changes if the value of an independent variable increases by one unit
  - For example if IV is measured in hours – the B coefficient shows how the DV changes if the value of IV increases by one hour

- $y = b_0 + \mathbf{b_1}*x + \mathbf{b_2}*y + \mathbf{b_3}*z$

Outcomes Concerning Independent Variables

- **Unstandardized B coefficient:**
  - Shows how the value of dependent variable changes if the value of an independent variable increases by one unit
  - For example if IV is measured in hours – the B coefficient shows how the DV changes if the value of IV increases by one hour

- **Standardized Beta coefficient:**
  - Compares the importance of IVs
  - Higher distance from zero shows higher importance of IV

- **Significance:**
  - Shows whether the found effect of IV can be applied to population

Example

- Is turnout in local elections affected by town population?

- Hypothesis: Turnout decreases as population increases
- Null hypotheses: There is no relation between population size and turnout

- Dependent variable:
  - Turnout – turnout in % (scale)

- Independent variable:
  - Population_th - town population in thousands of people (scale)

How to Perform the OLS Regression

- Analyze > Regression > Linear

- Select the variables:
  - Turnout into ‘Dependent’
  - Population_th in the section for independent variables

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,265[a]	,070	,070	12,58928

a. Predictors: (Constant), Population_1000

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35006,761	1	35006,761	220,877	,000[b]
	Residual	462315,124	2917	158,490		
	Total	497321,885	2918			

a. Dependent Variable: Turnout

b. Predictors: (Constant), Population_1000

- Model Summary:
  - Our model explains 7 per cent (0,07 * 100) of variance of dependent variable
- ANOVA:
  - Our model is a significant improvement in predicting the dependent variable and our results can be applied to the population

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	60,800	,244		248,812	,000
	Population_th	-,591	,040	-,265	-14,862	,000

a. Dependent Variable: Turnout

- Intercept (Constant):
  - Predicted value of dependent variable if all independent variables = 0
  - In a (non-existing) town with zero population the turnout in local election is predicted as 60.8 per cent

- $y = \mathbf{b_0} + b_1 * x$
- $y = 60.8 + b_1 * x$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	60,800	,244		248,812	,000
	Population_th	-,591	,040	-,265	-14,862	,000

a. Dependent Variable: Turnout

- Unstandardized B:
  - Shows how the value of DV changes if the value of an IV increases by one unit
  - Population_th is measured in thousands of people
  - Interpretation – for each thousand people living in a town the turnout drops by 0.591 percentage points
  - This effect is significant at 99.9 % and so it can be applied to population (we reject the null hypothesis about absence of relationship between IV and DV)

- $y = b_0 + \mathbf{b_1} * x$
- $y = 60.8 + (-0.591) * x$
- $y = 60.8 – 0.591 * x$

Predictions Based on Results

- $y = b_0 + b_1 * x$
- Turnout = 60.8 – 0.591*Population_th

	Population	Population in thousands	Formula	Predicted turnout
Town 1	500	0.5	60.8 – 0.591*0.5 = 60.8 – 0.296	**60.5**
Town 2	1,000	1	60.8 – 0.591*1 = 60.8 – 0.591	**60.2**
Town 3	5,000	5	60.8 – 0.591*5 = 60.8 – 2.955	**57.8**
Town 4	10,000	10	60.8 – 0.591*10 = 60.8 – 5.91	**54.9**
Town 5	25,000	25	60.8 – 0.591*25 = 60.8 – 14.775	**46.0**

Example 2

- Is turnout in local elections affected by town population, the local financial situation and whether there is a true competition?

- Dependent variable:
  - Turnout – turnout in % (scale)

- Independent variables:
  - Population_th - town population in thousands of people (scale)
  - Fin_Index – indicator of financial situation in town (1-6; 1 = worst, 6 = best) (scale)
  - Competition – 1 for at least two competitors or 0 for only one competitor (binary)

How to Perform the OLS Regression

- Analyze > Regression > Linear

- Select the variables:
  - Turnout into ‘Dependent’
  - Population_th, Fin_index and Competition in the section for independent variables

- Because we have more than one IV:
  - Statistics > Collinearity Diagnostics

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,675[a]	,456	,455	9,62055

a. Predictors: (Constant), Competition, Fin_Index, Population_th

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	225580,308	3	75193,436	812,419	,000[b]
	Residual	269150,062	2908	92,555		
	Total	494730,370	2911			

a. Dependent Variable: Turnout

b. Predictors: (Constant), Competition, Fin_Index, Population_th

- Our model explains 45.5 per cent of variance of dependent variable
- Substantial improvement compared to model that included only one independent variable

- Our model is a significant improvement in predicting the dependent variable and our results can be applied to the population

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	55,569	1,912		29,069	,000		
	Population_th	-,770	,031	-,347	-25,077	,000	,980	1,020
	Fin_Index	-1,382	,361	-,053	-3,831	,000	,994	1,006
	Competition	17,995	,397	,625	45,308	,000	,984	1,016

a. Dependent Variable: Turnout

- Intercept (Constant):
  - Predicted value of dependent variable if all independent variables = 0
  - In a (non-existing) town with zero population, financial index of 0 and with only a single competitor the turnout in local election is predicted as 55.569 per cent

- $y = \mathbf{b_0} + b_1*x + b_2*y + b_3*z$
- $y = 55.569 + b_1*x + b_2*y + b_3*z$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	55,569	1,912		29,069	,000		
	Population_th	-,770	,031	-,347	-25,077	,000	,980	1,020
	Fin_Index	-1,382	,361	-,053	-3,831	,000	,994	1,006
	Competition	17,995	,397	,625	45,308	,000	,984	1,016

a. Dependent Variable: Turnout

- Unstandardized B:
  - Shows how the value of DV changes if the value of an IV increases by one unit
  - **Population_th** is measured in thousands of people
  - Interpretation – for each thousand people living in a town the turnout drops by 0.77 percentage points
  - This effect is significant at 99.9 % and so it can be applied to population (we reject the null hypothesis about absence of relationship between IV and DV)

- $y = b_0 + \mathbf{b_1}{*}x + b_2{*}y + b_3{*}z$
- $y = 55.569 – \mathbf{0.77}{*}x + b_2{*}y + b_3{*}z$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	55,569	1,912		29,069	,000		
	Population_th	-,770	,031	-,347	-25,077	,000	,980	1,020
	Fin_Index	-1,382	,361	-,053	-3,831	,000	,994	1,006
	Competition	17,995	,397	,625	45,308	,000	,984	1,016

a. Dependent Variable: Turnout

- Unstandardized B:
    - Shows how the value of DV changes if the value of an IV increases by one unit
    - **Fin_Index** is measured on a scale from 1 to 6
    - Interpretation – for each increase on the financial scale by one the turnout drops by 1.382 percentage points
    - This effect is significant at 99.9 % and so it can be applied to population (we reject the null hypothesis about absence of relationship between IV and DV)

- $y = b_0 + b_1 * x + \mathbf{b_2} * y + b_3 * z$
- $y = 55.569 - 0.77 * x - \mathbf{1.382} * y + b_3 * z$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	55,569	1,912		29,069	,000		
	Population_th	-,770	,031	-,347	-25,077	,000	,980	1,020
	Fin_Index	-1,382	,361	-,053	-3,831	,000	,994	1,006
	Competition	17,995	,397	,625	45,308	,000	,984	1,016

a. Dependent Variable: Turnout

- Unstandardized B:
  - Shows how the value of DV changes if the value of an IV increases by one unit
  - **Competition** is a binary variable (0 = no competition; 1 = at least two candidates)
  - Interpretation – if there is a competition, the turnout in town increases by 17.995 percentage points
  - This effect is significant at 99.9 % and so it can be applied to population (we reject the null hypothesis about absence of relationship between IV and DV)

- $y = b_0 + b_1*x + b_2*y + \mathbf{b_3}*z$
- $y = 55.569 - 0.77*x - 1.382*y + \mathbf{17.995}*z$

Unstandardized B Coefficient

- Scale v. Binary Variables

- Same definition for scale and binary variables:
  - Shows how the value of DV changes if the value of an IV increases by one unit

BUT

- Binary (dummy) variables have only two values – 0 and 1
  - Unlike scale variables, there is only one possible increase by one unit
  - The estimated effect is thus completely exhausted by this one increase

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	55,569	1,912		29,069	,000		
	Population_th	-,770	,031	-,347	-25,077	,000	,980	1,020
	Fin_Index	-1,382	,361	-,053	-3,831	,000	,994	1,006
	Competition	17,995	,397	,625	45,308	,000	,984	1,016

a. Dependent Variable: Turnout

- **Competition:**
  - 0 – no competition (only one candidate)
  - 1 – competition (at least two candidates)
  - Shift from 0 to 1 means that towns with competition are predicted to have a nearly 18 percentage points higher turnout than towns without competition

- **Population_th:**
  - Shift of population from 1 thousand to 2 thousand leads to drop of turnout by 0.77 percentage points
  - Shift of population from 1 thousand to 5 thousand leads to drop of turnout by 3.08 percentage points (4 times decrease of 0.77)
  - Shift of population from 5 thousand to 12 thousand leads to drop of turnout by 5.39 percentage points (7 times decrease of 0.77)

Standardized Beta Coefficient

- Provide information about importance of independent variables
- Measured in standard deviation units → allow to easily compare the IVs
- Higher distance from zero (both positive and negative) indicates higher importance of the independent variables

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	55,569	1,912		29,069	,000		
	Population_th	-,770	,031	-,347	-25,077	,000	,980	1,020
	Fin_Index	-1,382	,361	-,053	-3,831	,000	,994	1,006
	Competition	17,995	,397	,625	45,308	,000	,984	1,016

a. Dependent Variable: Turnout

- Results show that Competition is the most important predictor of all three independent variables

- Population_th is less important and Fin_Index is the least important

Predictions Based on Results

- $y = b_0 + b_1*x + b_2*y + b_3*z$
- Turnout = 55.569 – 0.77*Population_th – 1.382*Fin_Index + 17.995*Competition

	Population	Fin_Index	Competition	Formula	Predicted turnout
Town 1	1,000	3	0	55.569 – 0.77*1 – 1.382*3 + 17.995*0	**50.7**
Town 2	1,000	3	1	55.569 – 0.77*1 – 1.382*3 + 17.995*1	**68.6**
Town 3	5,000	3	0	55.569 – 0.77*5 – 1.382*3 + 17.995*0	**47.6**
Town 4	10,000	6	1	55.569 – 0.77*10 – 1.382*6 + 17.995*1	**57.6**
Town 5	25,000	6	0	55.569 – 0.77*25 – 1.382*6 + 17.995*0	**28.0**

Control of Assumptions

- Outliers – cases with extreme values
- Heteroscedasticity – variance of residuals
- Collinearity – association between independent variables

- How to do that:
  - Analyze > Regression > Linear
  - Statistics > Collinearity diagnostics + casewise diagnostics (2.5)
  - Plots > Y: ZRESID, X: ZPRED

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	55,569	1,912		29,069	,000		
	Population_th	-,770	,031	-,347	-25,077	,000	,980	1,020
	Fin_Index	-1,382	,361	-,053	-3,831	,000	,994	1,006
	Competition	17,995	,397	,625	45,308	,000	,984	1,016

a. Dependent Variable: Turnout

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Variance Proportions Population_th	Variance Proportions Fin_Index	Variance Proportions Competition
1	1	2,933	1,000	,00	,02	,00	,03
	2	,858	1,849	,00	,96	,00	,00
	3	,205	3,786	,01	,01	,01	,97
	4	,004	25,770	,99	,01	,99	,00

a. Dependent Variable: Turnout

- VIF above 5 (10) or Tolerance below 0.2 (0.1) constitutes a problem
- Similarly more higher values on same dimensions indicate collinearity
- Solution – more models or dropping one of the variables

Outliers

- The data should contain up to:
  - 5 % of cases with residual above 2 (below -2)
  - 1 % of cases with residual above 2.5 (below -2.5)

- If we find outliers we can rerun the model without these cases and compare whether the results change