

Korelační analýza a jednoduchá regresní analýza

Kvantitativní metody zkoumání politiky (JPB283)

Michal Parížek

14. 4. 2025

Institut politologických studií FSV UK

Obsah

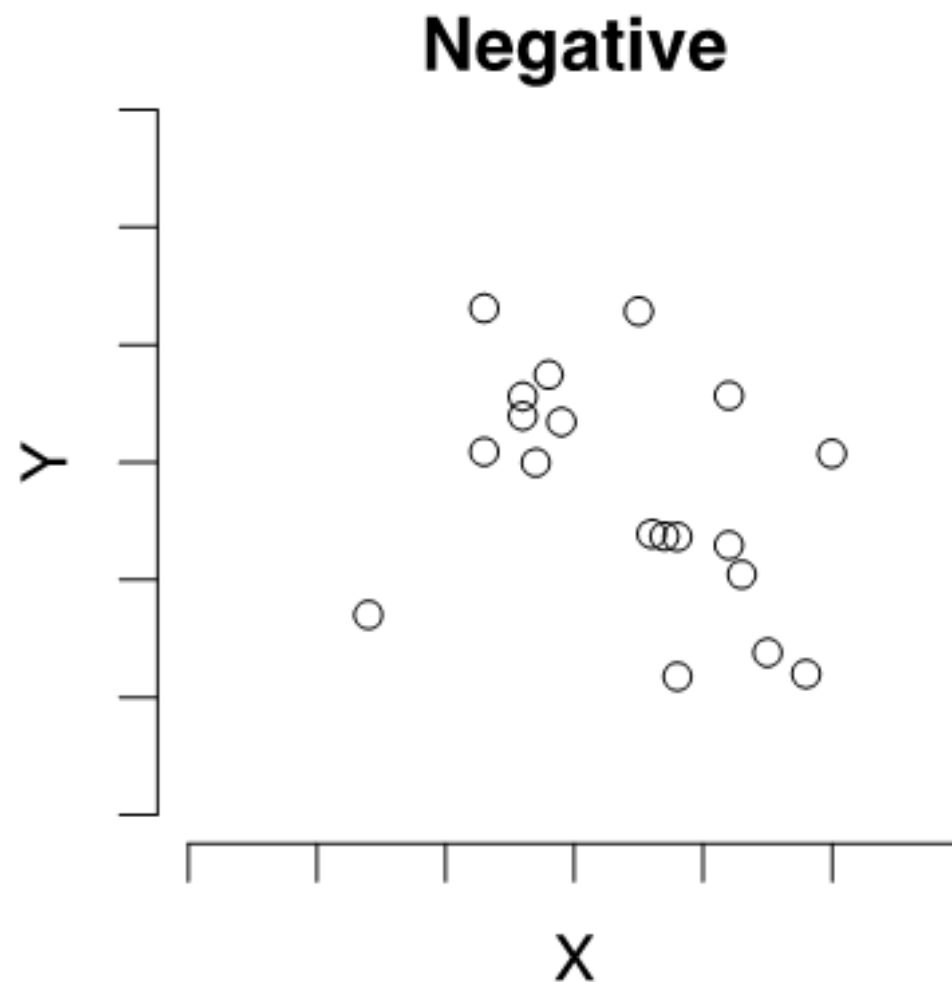
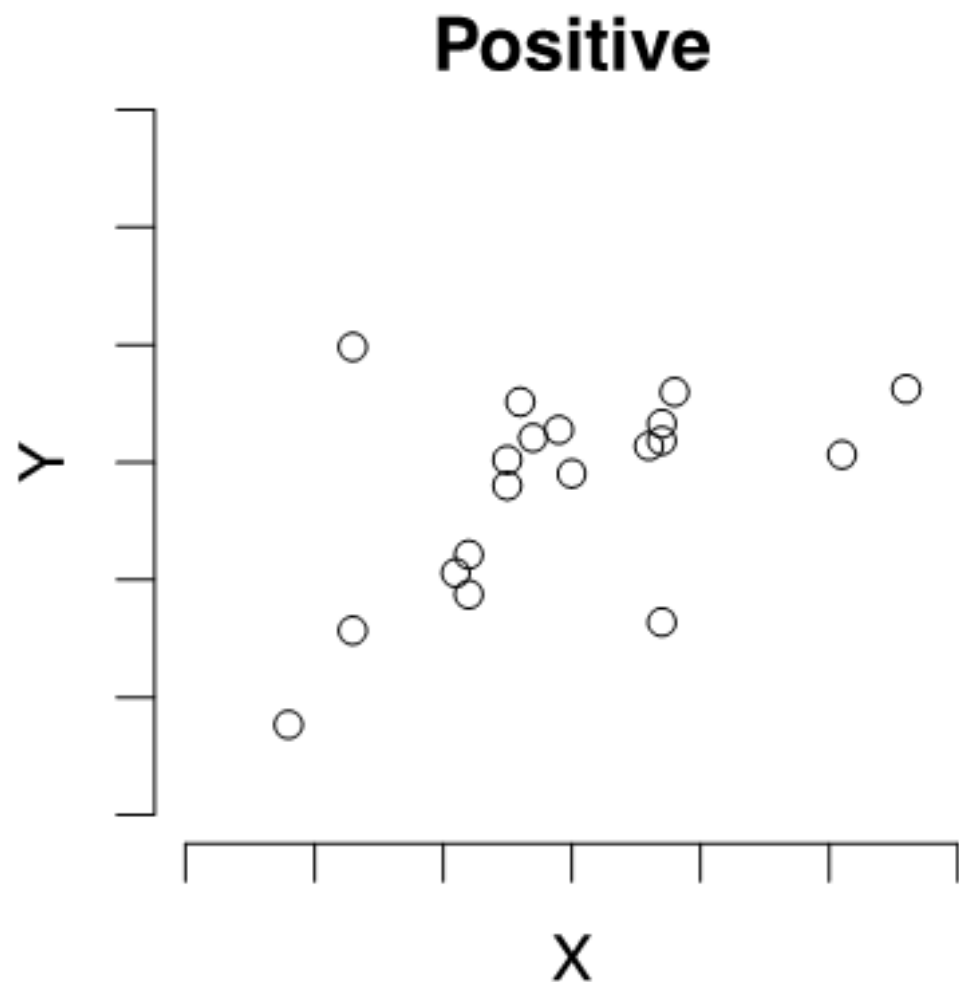
- Korelační analýza
- Principy regresní analýzy
- Jak dobrá je naše regresní přímka?

Přehled testů pro vztah různých druhů proměnných

Nezávislá proměnná	Závislá proměnná	
	Diskrétní (nominální, ordinální)	Spojitá (intervalová, poměrová)
Diskrétní (nominální, ordinální)	Chí-kvadrát	T-test (rozdíl průměrů)
Spojitá (intervalová, poměrová)	Logit, Probit	Korelace, Regrese

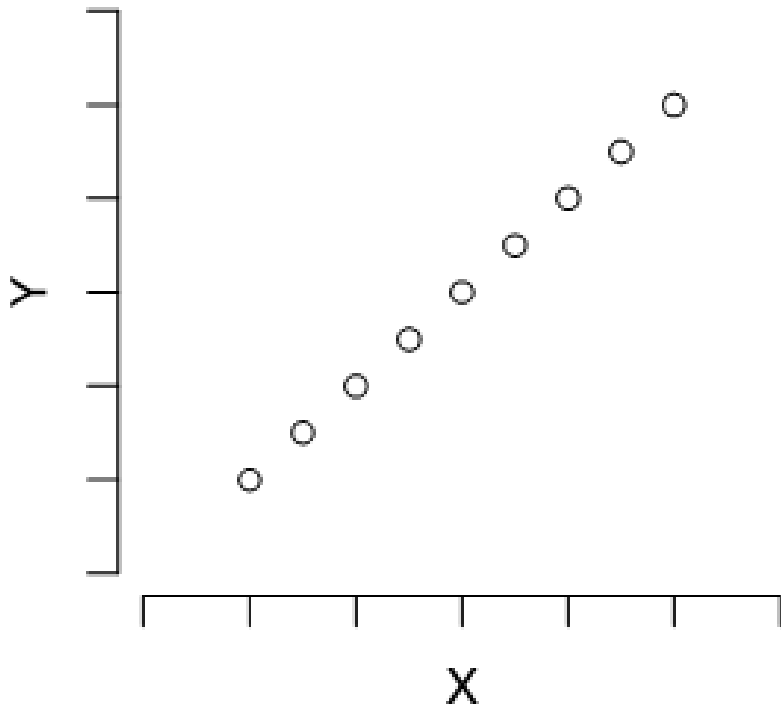
Korelační analýza

Pozitivní a negativní vztah mezi proměnnými

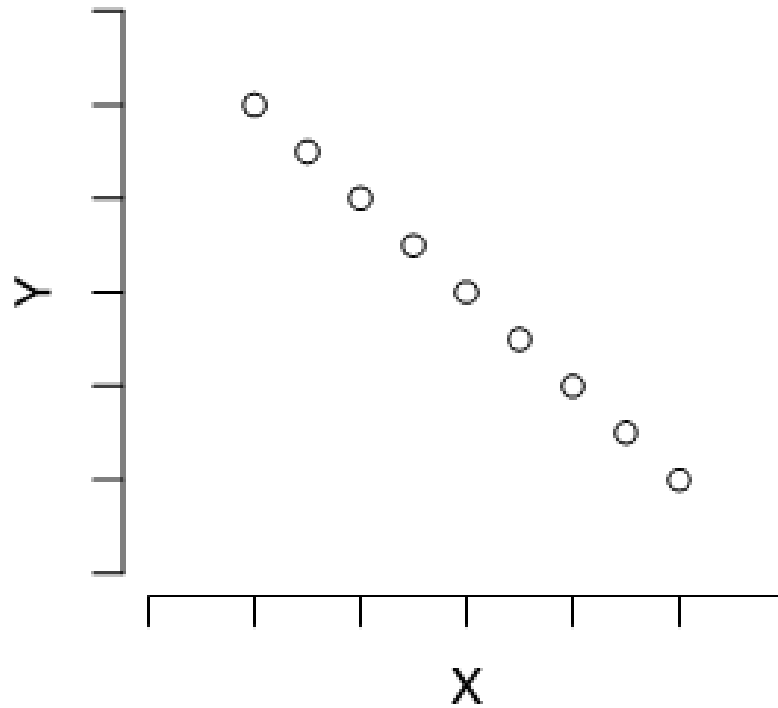


Síla a směr korelace

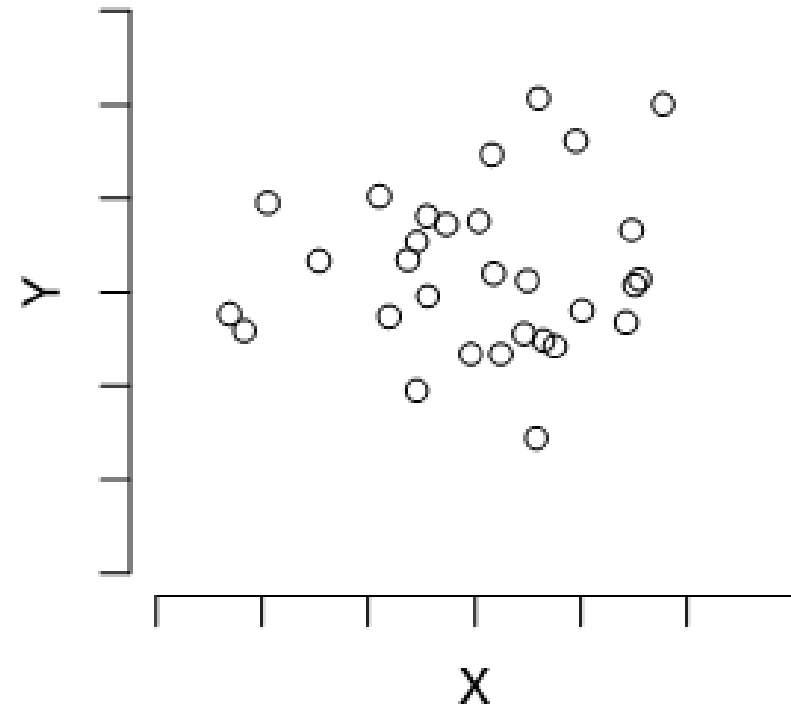
Strong



Strong



Weak



Dokážeme to nějak změřit?

- Zkonstruujeme měřítko síly vztahu mezi spojitými proměnnými
- ideálně měřítko *standardizované*, umožňující aplikaci napříč různými kontexty

Pearsonův korelační koeficient r

- $|r| = 1$ označuje perfektní vztah
- $r = 0$ označuje absenci vztahu
- $r = 1$ označuje perfektní pozitivní vztah
- $r = -1$ označuje perfektní negativní vztah

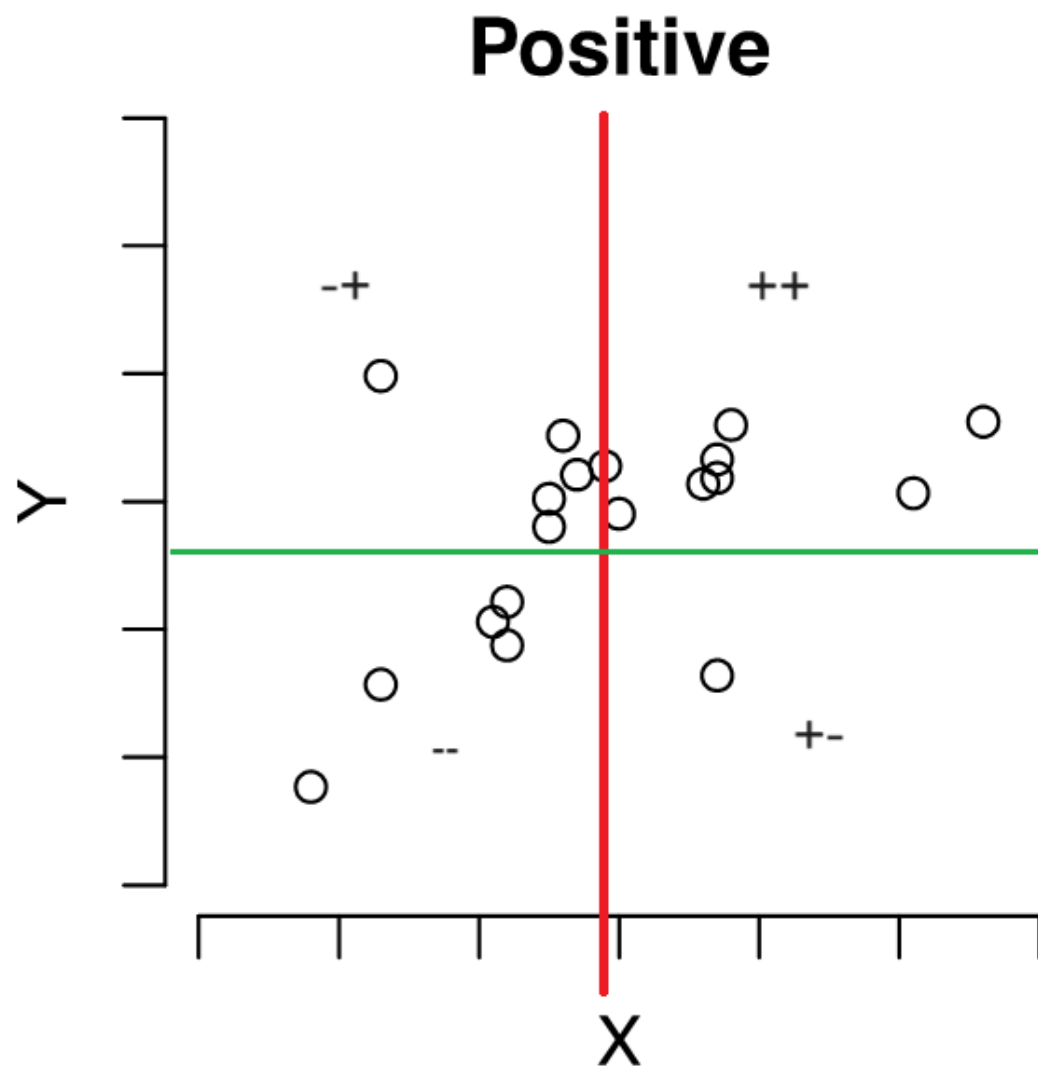
Pearsonův korelační koeficient r

- Jádrem koeficientu je výpočet kovariance

$$\text{Cov} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

Pearsonův korelační koeficient r

- Mění se hodnoty proměnných pro jednotlivá pozorování
- spolu (++)/--)
- proti sobě(+--/--+)
- nebo nezávisle na sobě?



Pearsonův korelační koeficient r

- Ale hodnota kovariance záleží na (rozdílné) míře rozptylu jednotlivých proměnných
- hodnotu kovariance poměříme pomocí měřítka rozptýlenosti obou proměnných, tj. pomocí jejich směrodatných odchylek

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{N}}$$

Pearsonův korelační koeficient r

$$r = \frac{\textit{Kovariance } x \textit{ a } y}{\textit{Součin směrodatných odchylek } x \textit{ a } y}$$

$$r = \frac{\frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N-1}}{\sqrt{\frac{\Sigma(X - \bar{X})^2}{N}} \times \sqrt{\frac{\Sigma(Y - \bar{Y})^2}{N}}}$$

Příklad: měření informační provázanosti světa v rámci projektu GLOWIN (a Peace Research Center Prague)



GLOWIN
Global Flows of Political Information

[HOME](#) [ABOUT](#) [TEAM](#) [DATA & TOOLS](#) [PUBLICATIONS](#)

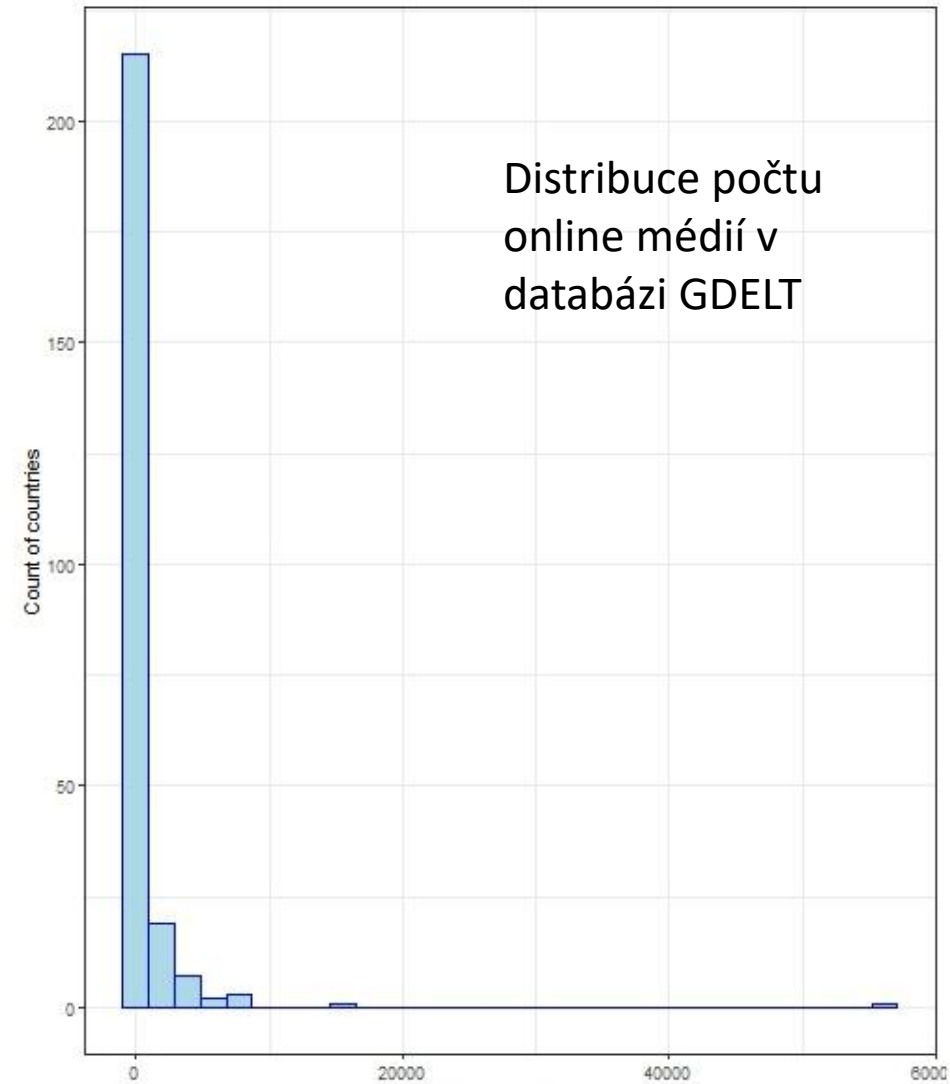
Innovative agenda, unique data and cutting-edge methodology

Global Flows of Political Information (GLOWIN) is a research project that creates the **first map of the flows of political information across countries**. This is a **pioneering research agenda**, as social sciences simply do not know the extent to which politically relevant information, as reported in mass media, travels across nations. GLOWIN opens the space for systematic exploration of this **important dimension of globalization** and international politics long unaddressed.

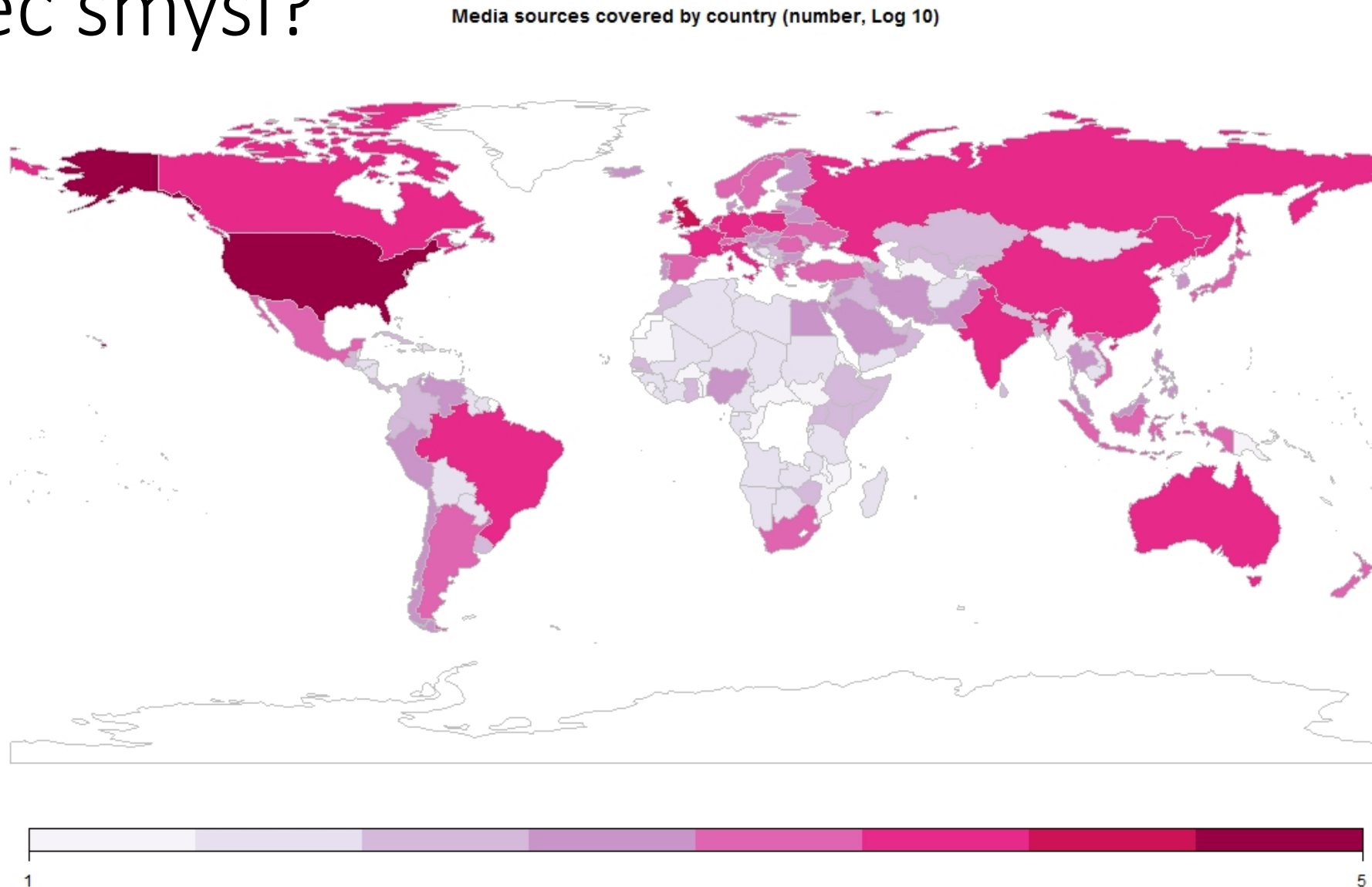
We combine deep **inter-disciplinary theorization** with **cutting-edge methodology** and **data collection techniques**. Unique empirical data and advanced automated text analysis allow for coverage of news from virtually **all countries of the world** and **more than 30,000 online media outlets**. Moreover, to estimate the amount of political information contained in this news, GLOWIN derives an **original measure of political information** based on insights from information theory.

Our approach in application: Covid-19

Here is a small teaser, showcasing the kind of data and insight we work on in GLOWIN. This video shows how the online media coverage of Covid-19, in connection to individual countries, spread in the early period of the pandemic.



Mediální mapa světa: dává měřítko/zdroj dat vůbec smysl?



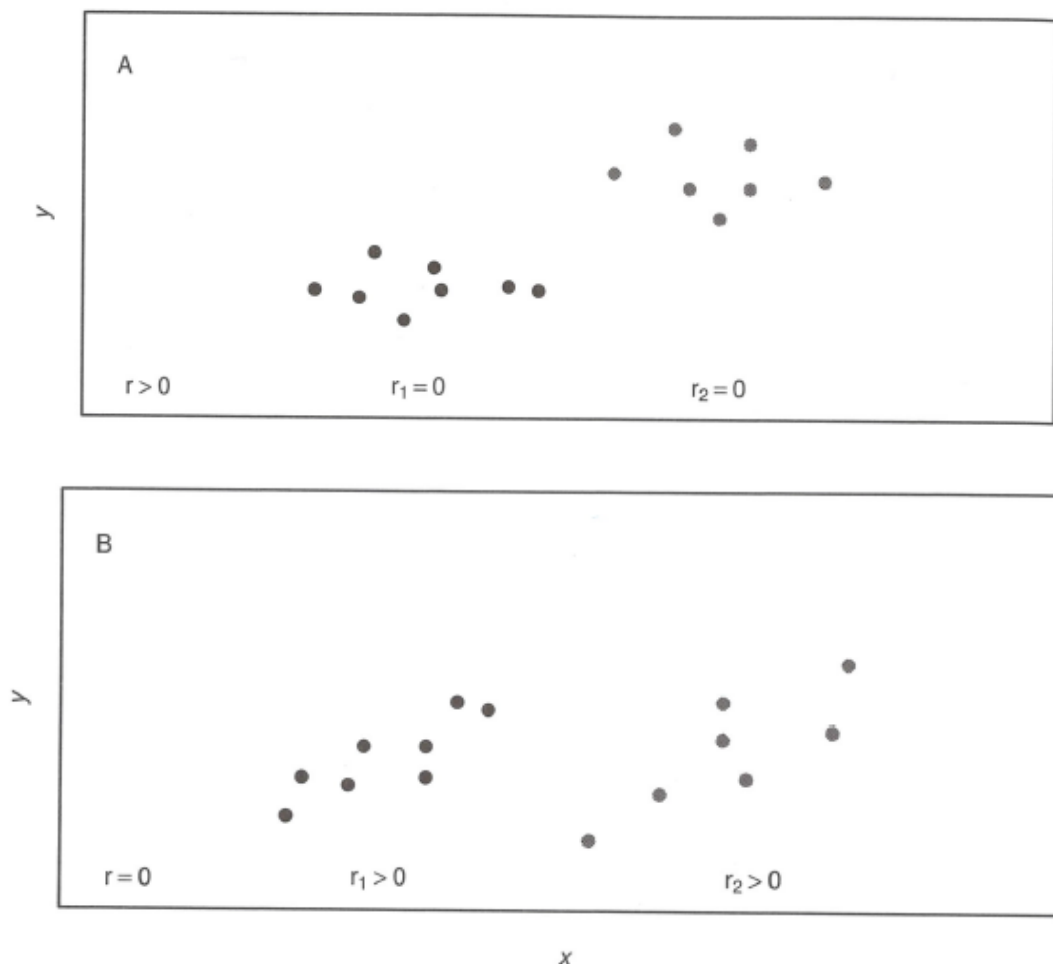
Mediální mapa světa: dává měřítko/zdroj dat vůbec smysl?

- Ale co když ty extrémní rozdíly napříč zeměmi prostě odpovídají velikosti mediálního trhu?
- Aproximace velikosti mediálního trhu velikostí ekonomiky - GDP (ok, trochu divoké)
- Extrémně šikmé proměnné – logaritmická transformace

	A	B	C	D	E	F	G	H	I
1		log10gni	logsources						
2	1	9.40312052	0.95424251						
3	2	10.2931437	1.90848502						
4	3	11.0733003	1.62324929						
5	5	10.1186654	1.98227123		r=	=PEARSON(B2:B184;C2:C184)			
6	7	11.5858576	2.45484486			0.8666314			
7	8	11.7931168	3.23829707						
8	9	10.0791782	2.25767857		n=	=COUNT(C2:C184)			
9	10	9.15902608	1.5797836			183			
10	11	12.1098821	3.77974075						
11	12	11.6205691	2.90579588		t=	=(E5*SQRT(F9-2))/(SQRT(1-E5^2))			
12	13	10.5909348	2.07918125			23.367786			
13	14	9.50073251	1.07918125						
14	15	11.6961921	3.06032003		df=	=F9-2			
15	16	9.96286088	1.38021124			181			
16	17	10.0763026	1.44715803						
17	18	11.4157327	2.34242268		sig	=T.DIST.2T(ABS(F12);F15)			
18	19	10.7707404	2.5575072			1.521E-56			

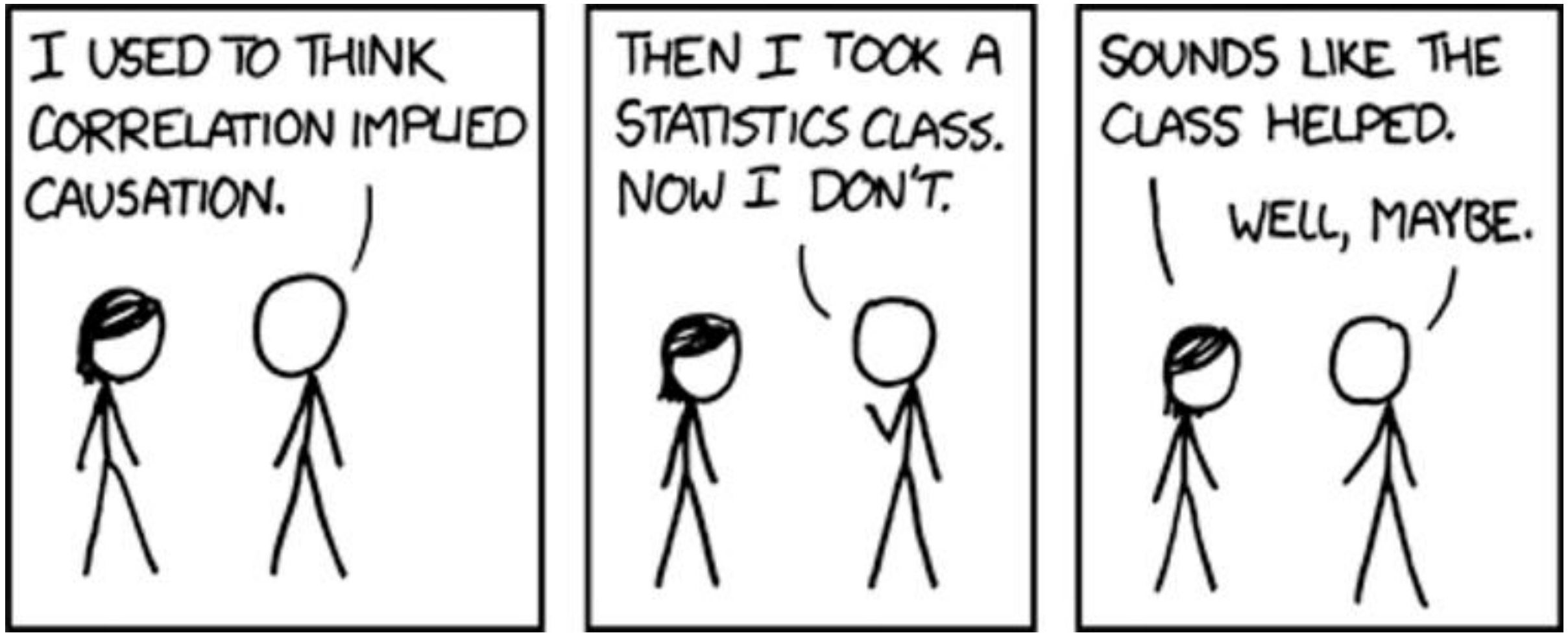
	A	B	C	D	E	F	G	H	I
1		log10gni	logsources						
2	1	9.40312052	0.95424251						
3	2	10.2931437	1.90848502						
4	3	11.0733003	1.62324929						
5	5	10.1186654	1.98227123	r=	=PEARSON(B2:B184;C2:C184)				
6	7	11.5858576	2.45484486		0.8666314				
7	8	11.7931168	3.23829707						
8	9	10.0791782	2.25767857	n=	=COUNT(C2:C184)				
9	10	9.15902608	1.5797836		183				
10	11	12.1098821	3.77974075						
11	12	11.6205691	2.90579588	t=	=(E5*SQRT(F9-2))/(SQRT(1-E5^2))				
12	13	10.5909348	2.07918125		23.367786				
13	14	9.50073251	1.07918125						
14	15	11.6961921	3.06032003	df=	=F9-2				
15	16	9.96286088	1.38021124		181				
16	17	10.0763026	1.44715803						
17	18	11.4157327	2.34242268	sig	=T.DIST.2T(ABS(F12);F15)				
18	19	10.7707404	2.5575072		1.521E-56				

Nepravá korelace



- Další proměnné způsobují vztah
- Nehomogenita dat:
 - v rámci skupiny vztah existuje, ale napříč skupinami není viditelný
 - napříč skupinami pozorujeme vztah, ale ten je ve skutečnosti dán odlišností skupin jako celků, ne jevy napříč pozorováními

Korelace není kauzalita

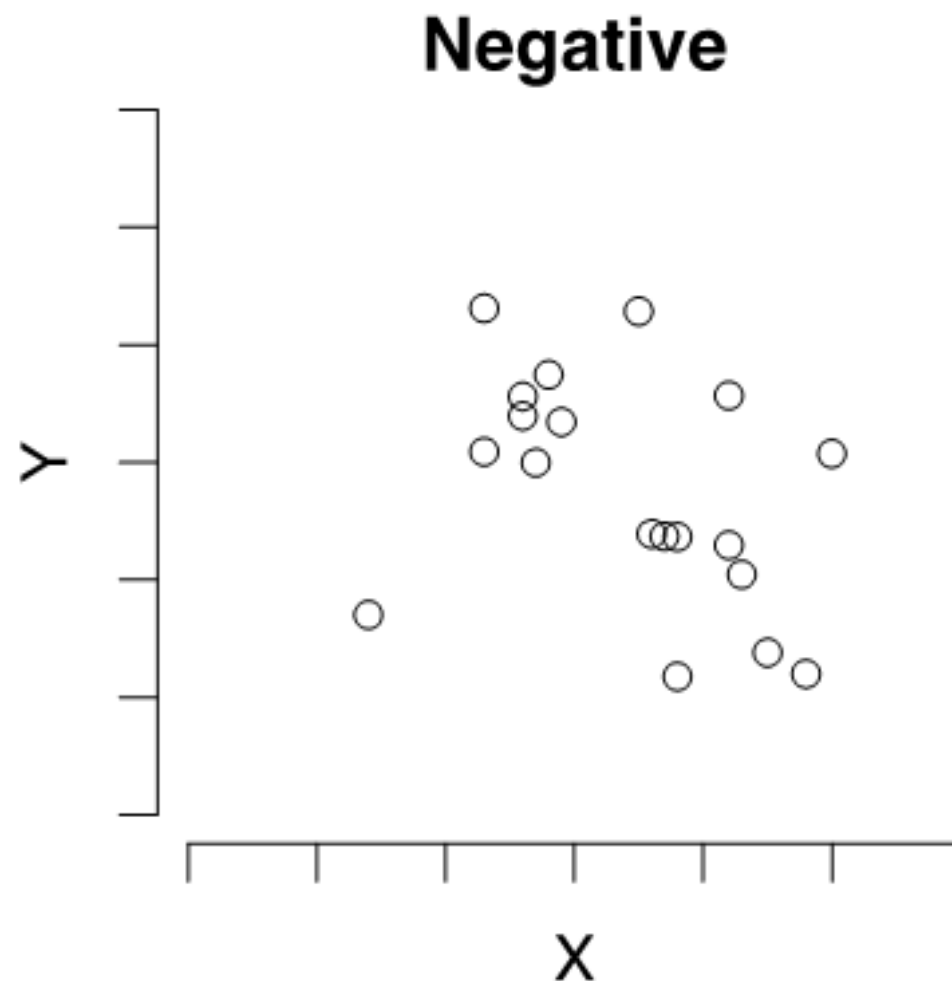
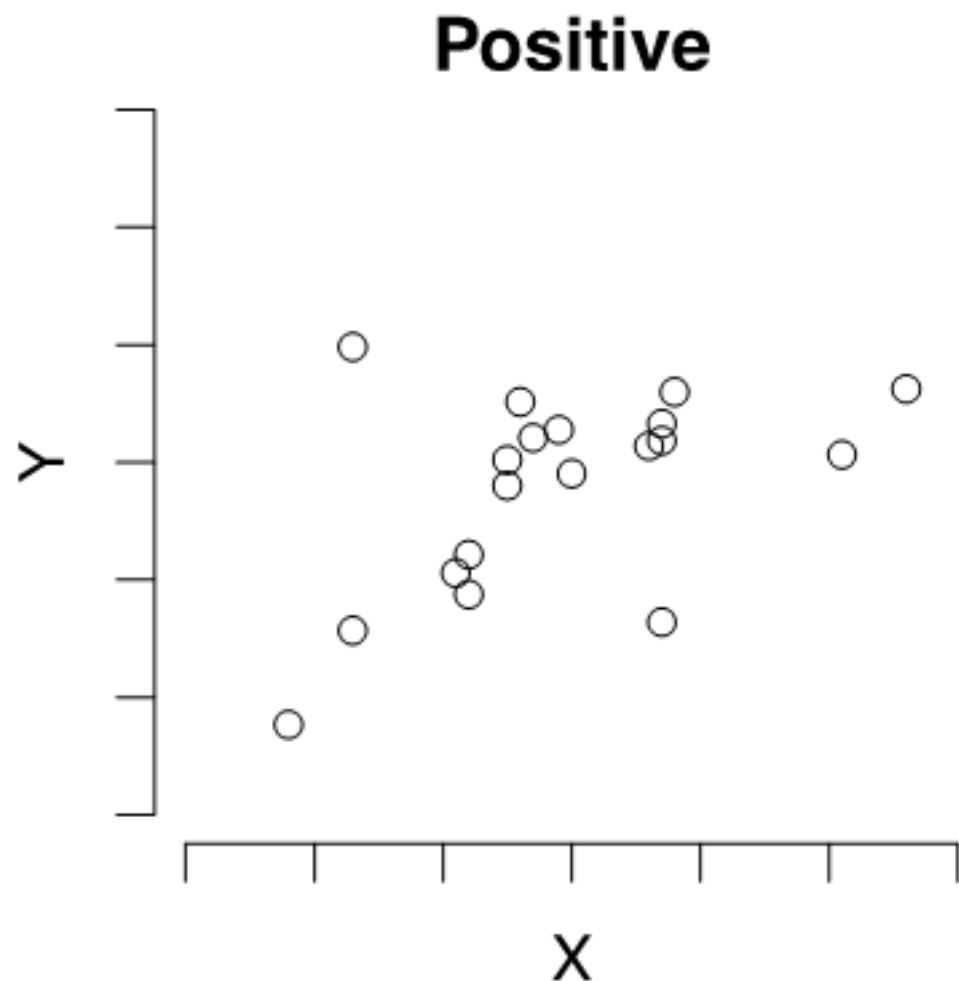


Co nám tedy v korelaci chybí?

- přesný popis vztahu mezi proměnnými
- schopnost predikce hodnot závislé proměnné na základě hodnot nezávislé proměnné
- ... a spousta možností, o kterých teď ani nevíme, že bychom je mohli chtít

Principy regresní analýzy

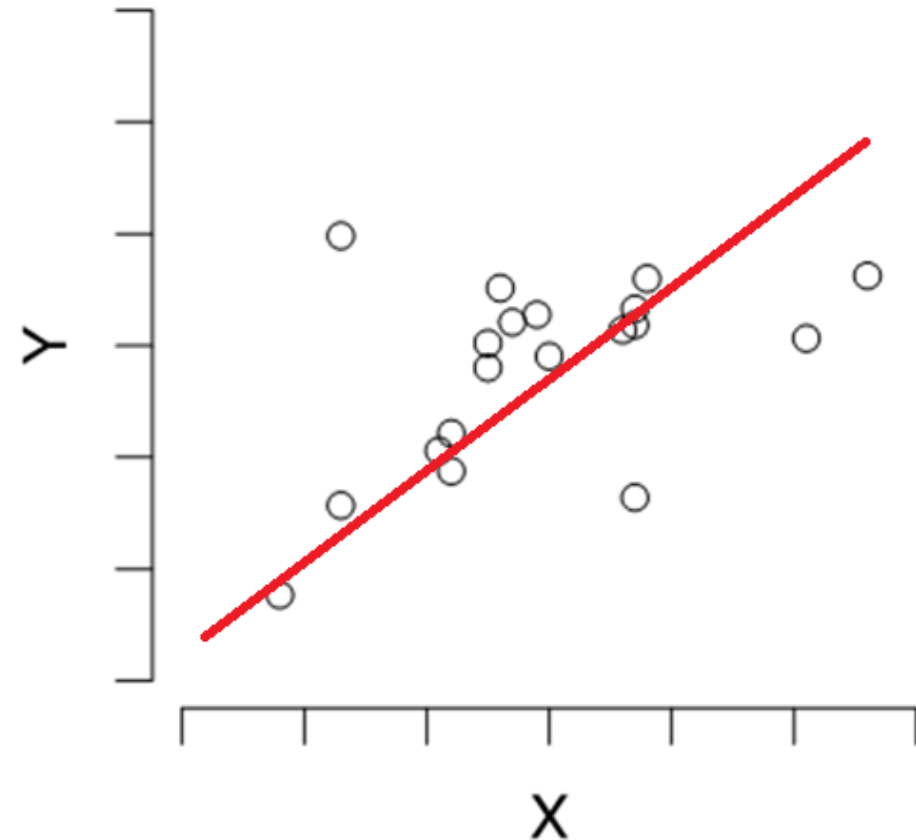
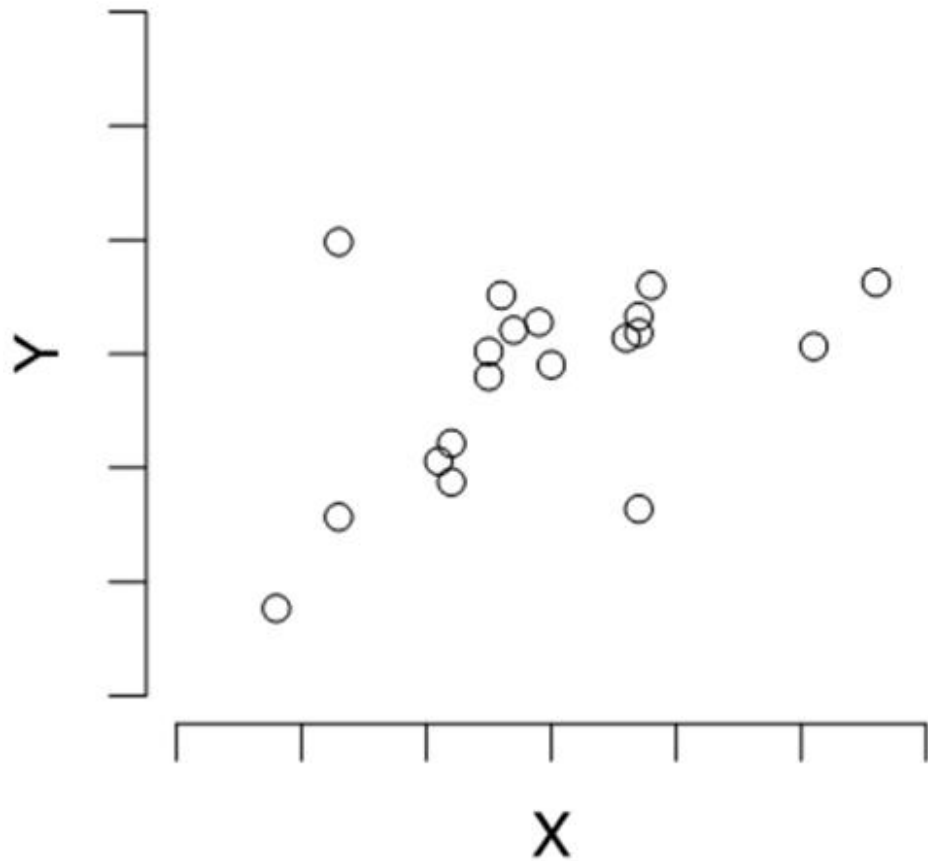
Pozitivní a negativní vztah mezi proměnnými



Dokážeme nějak smysluplně numericky zachytit vztah mezi proměnnými?

- Co takhle zkusit proložit daty přímkou?
- Přímka je geometricky velice jednoduchý objekt
- $Y = a + bX$

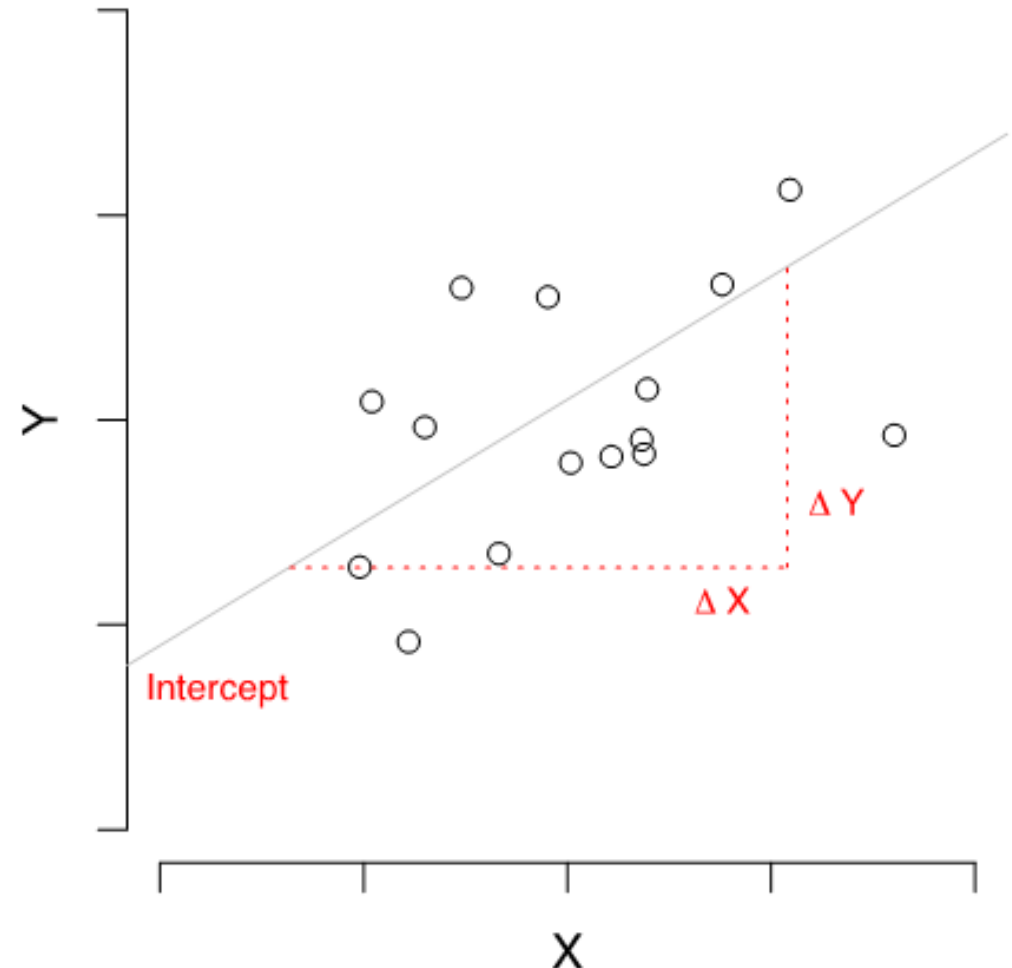
Nezachycuje náhodou ta přímka vztah, který hledáme?



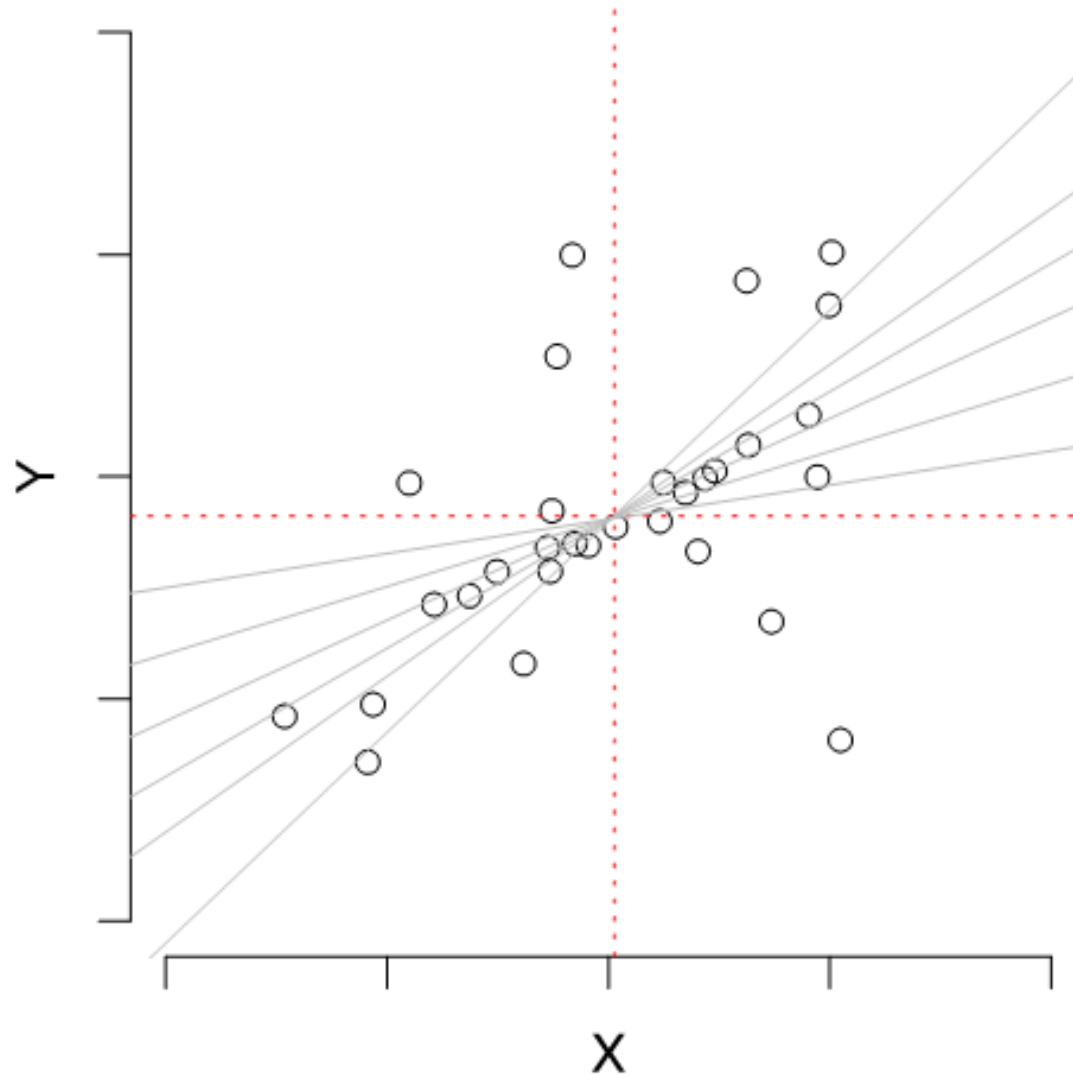
Sklon přímky určí, jestli je mezi proměnnými vztah

$$Y = a + bX + e$$

- a je konstanta
- b je sklon přímky, tedy **změna hodnoty závislé proměnné Y v důsledku jednotkové změny X**
- e je chyba, rozdíl mezi přímkou a pozorováním (error)

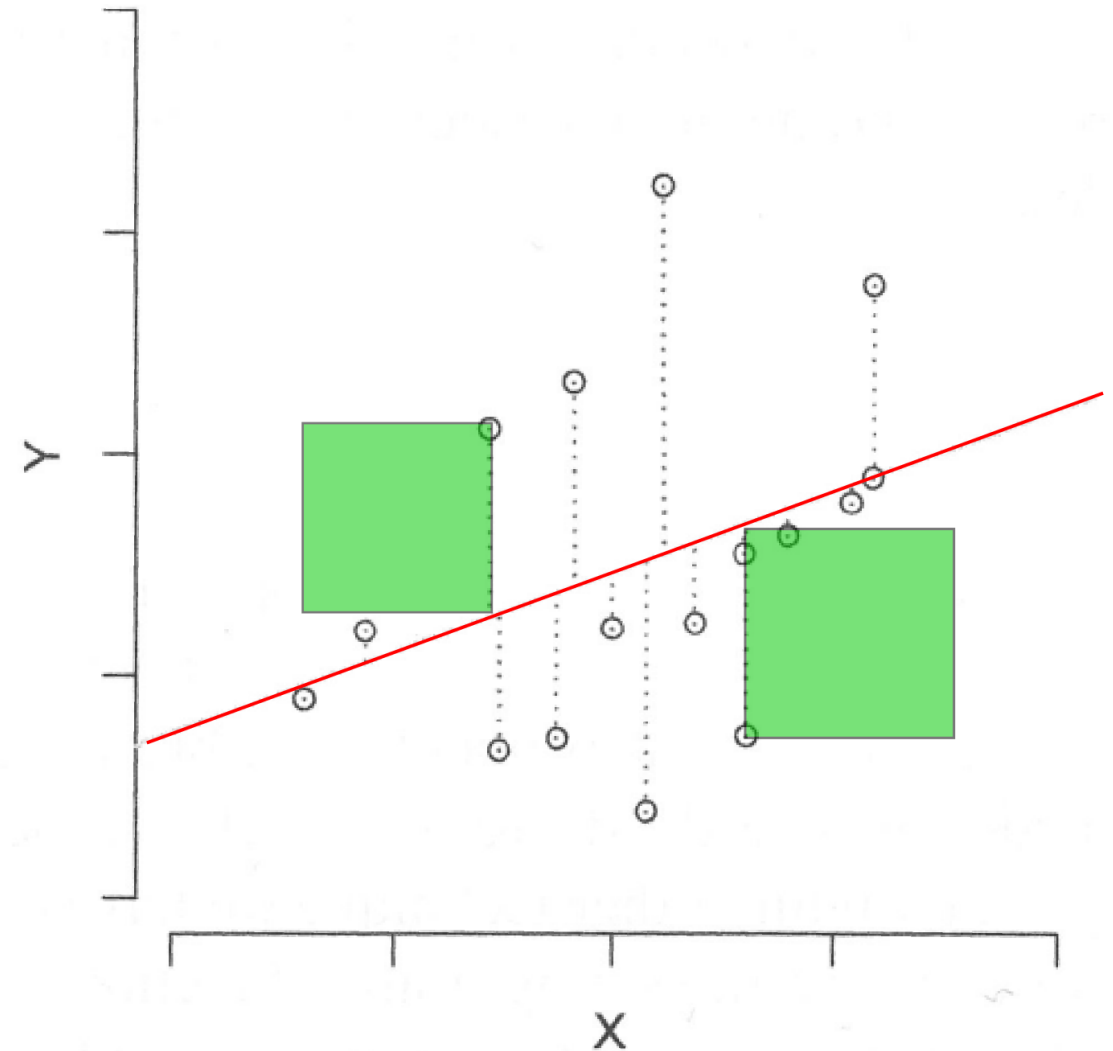


No jo, ale jakou přímku zvolit?



Metoda nejmenších čtverců (ordinary least squares, OLS)

- Volíme tu přímku, která **minimalizuje rezidua**, tj. rozdíl na Y mezi přímkou samotnou (kterou hledáme) a jednotlivými pozorováními
- Pro zamezení krácení $+/-$ chyb počítáme s druhými mocninami (čtverci)



Regresní rovnice

- Koeficienty b (sklon) a a (konstanta, průsečík s Y) se dají snadno spočítat

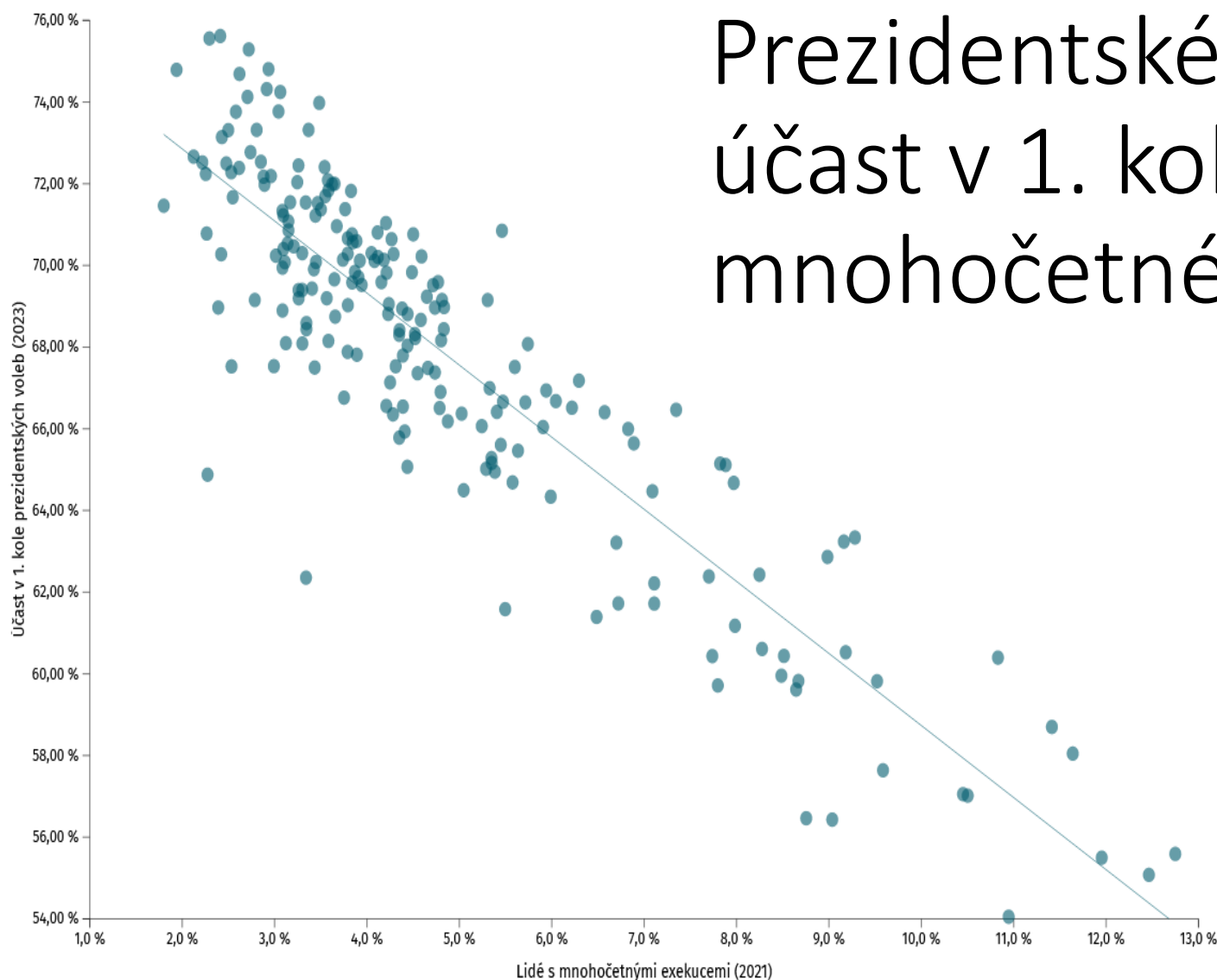
- $$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

- $$a = \bar{Y} - b\bar{X}$$

- Výsledkem je regresní rovnice zachycující pro každé X predikovanou hodnotu $\hat{Y}$

$$\hat{Y} = a + bX$$

Koeficient determinace (R^2) = 80 %



Prezidentské volby, PAQ
účast v 1. kole vs.
mnohočetné exekuce

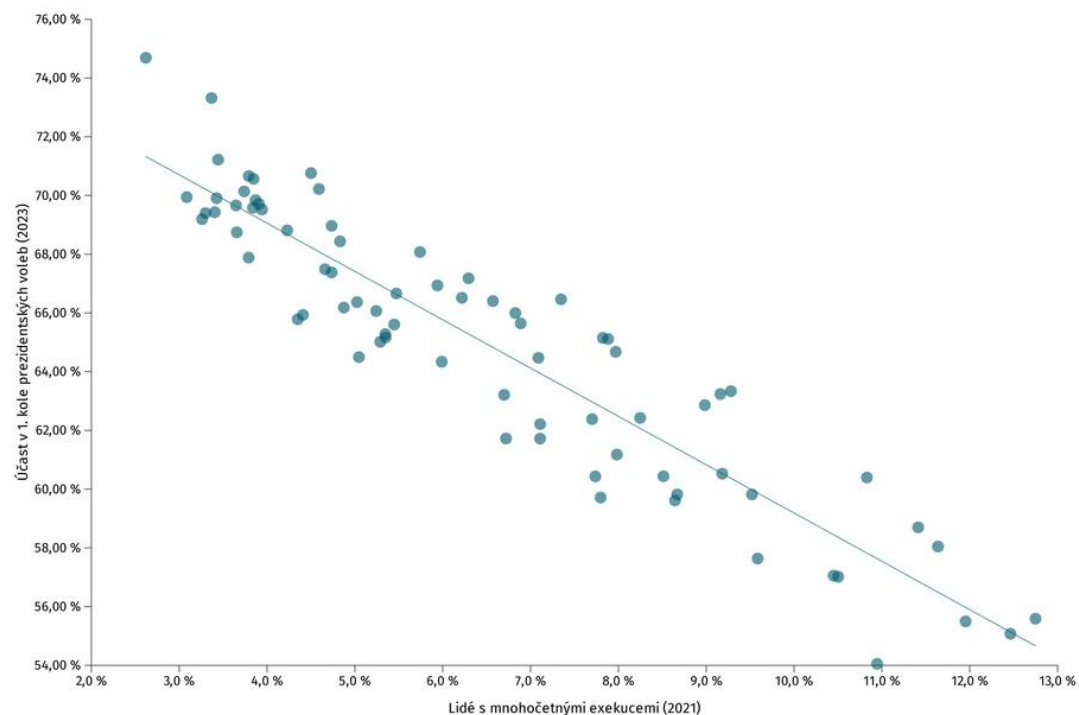
Zdroj dat: ČSÚ, Exekutorská komora ČR, Czech Household Panel Study

https://www.datapaq.cz/?v1=volby_prez2023_1kolo_ucast_2023&v2=podil_lidi_s_mnohocetnymi_exekucemi_2021&vis=scatterplot&t_sortby=volby_prez2023_1kolo_ucast_2023&t_sortdir=desc
https://twitter.com/dan_prokop/status/1616152554013851656/photo/3

Prezidentské volby, účast v 1. kole a mnohočetné exekuce: Sudety

Vztah účasti v 1. kole (2023) a Lidé s mnohočetnými exekucemi (2021) na úrovni ORP - Sudety

Koeficient determinace (R^2) = 86 %



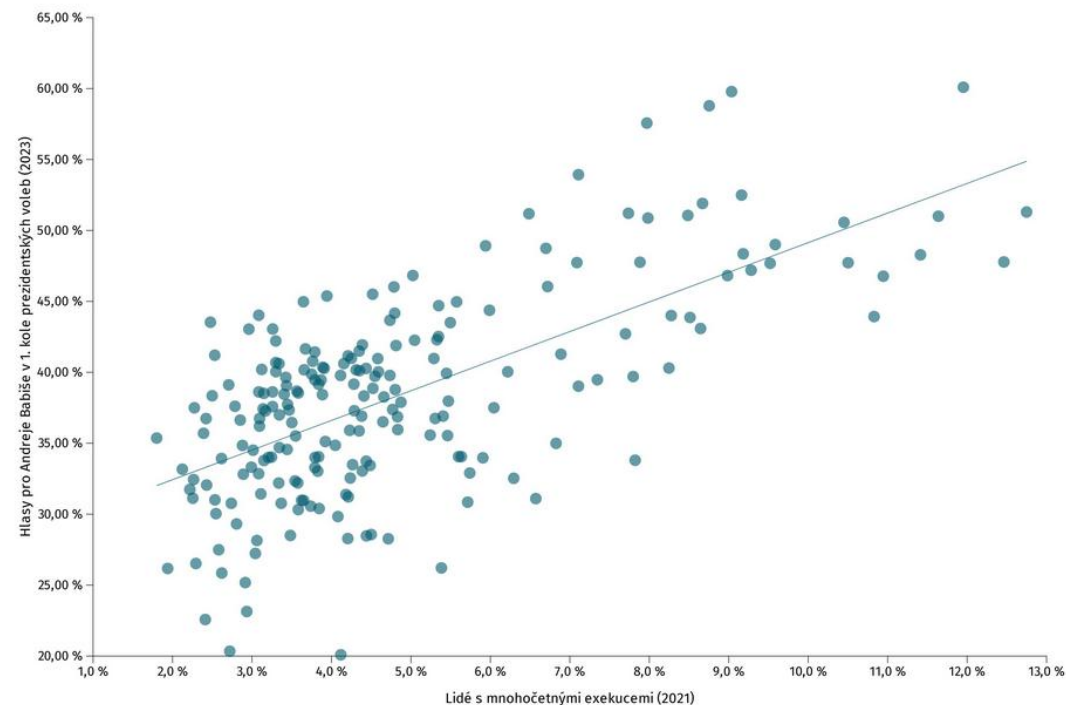
Vytvořeno v DataPAQ · Licence CC BY-NC 4.0 · Zdroj dat: ČSÚ, Exekutorská komora ČR, Czech Household Panel Study

https://twitter.com/dan_prokop/status/1616152554013851656/photo/3

Prezidentské volby, mnohočetné exekuce a podpora A. Babiše v 1. kole

Vztah ukazatelů Hlasy pro Andreje Babiše v 1. kole prezidentských voleb (2023) a Lidé s mnohočetnými exekucemi (2021) na úrovni ORP

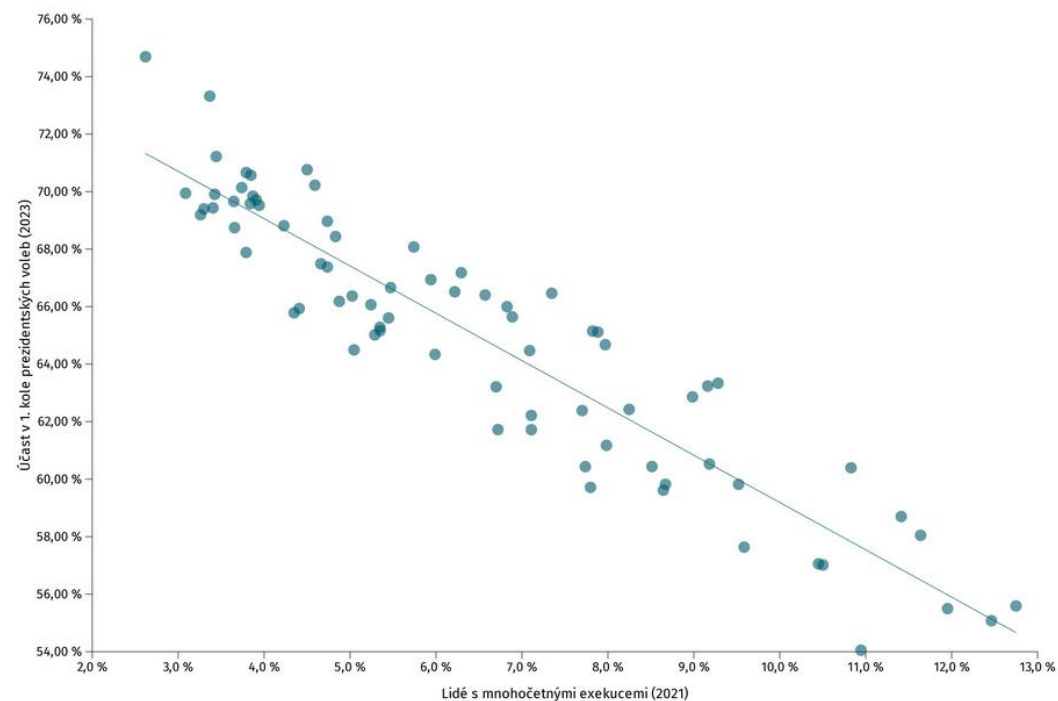
Koeficient determinace (R^2) = 44 %



Vytvořeno v DataPAQ · Licence CC BY-NC 4.0 · Zdroj dat: ČSÚ, Exekutorská komora ČR, Czech Household Panel Study

Vztah účasti v 1. kole (2023) a Lidé s mnohočetnými exekucemi (2021) na úrovni ORP - Sudety

Koeficient determinace (R^2) = 86 %



Vytvořeno v DataPAQ · Licence CC BY-NC 4.0 · Zdroj dat: ČSÚ, Exekutorská komora ČR, Czech Household Panel Study

https://twitter.com/dan_prokop/status/1616152554013851656/photo/3

Jak dobrá je naše regresní
přímka?

Kolik nám toho o závislé proměnné naše analýza řekne?

- rovnice $\hat{Y} = a + bX$ zachycuje nejlepší možnou přímku proloženou daty
- Ale bude to stačit, myslíš? Jak moc dobré to “nejlepší” vlastně je?



<http://memecreator.cz/meme/ptm3x9>

Kolik nám toho o závislé proměnné naše analýza řekne?

1) Kolik z variability na závislé proměnné vlastně přímka vysvětluje?

2) A můžeme věřit tomu, že sklon b je skutečně nenulový? Tedy že je mezi proměnnými skutečně vztah?



<http://memecreator.cz/meme/ptm3x9>

Ad 1) Koeficient determinace r^2 (většinou R^2)

- Jak zhodnotit míru vysvětlující síly regresní přímky vůči realitě?

- porovnáním **objemu vysvětlené variace Y** vůči **celkové variaci Y**

- $ESS_y = \sum(\hat{Y} - \bar{Y})^2$

- $SS_y = \sum(Y - \bar{Y})^2$

- ESS, explained sum of squares

- SS, total sum of squares

Ad 1) Koeficient determinace r^2 (většinou R^2)

$$r^2 = \frac{ESS_y}{SS_y}$$

- $r^2 = 0$ znamená, že regresní model nijak nezlepšuje naše chápání Y
- $r^2 = 1$ znamená, že model perfektně vysvětluje Y

Přestávka: trocha toho laskavého českého humoru



Ad 2) Můžeme už prosím *konečně* říct něco o vztahu mezi X a Y ?

- Mezi proměnnými existuje vztah, pokud má identifikovaná nejlepší přímka nenulový sklon, tj. $b \neq 0$
- Ovšem pracujeme s nedokonalými odhady na základě vzorku a nikoliv dokonalou znalostí populace (reality)
- Platí $b \neq 0$ s dostatečnou mírou jistoty vzhledem k tomu, kolik je v našich datech prostoru pro náhodné chyby?

Ad 2) Můžeme už prosím *konečně* říct něco o vztahu mezi X a Y ?

- Celková průměrná míra chyby naší predikce Y ...

$$MSe = \frac{\sum (Y - \hat{Y})^2}{N - 2}$$

- vztažená k celkové varianci (rozptylu) na X

$$s_b = \sqrt{\frac{MSe}{s_X^2 \times N - 1}}$$

- *Dává...* směrodatnou chybu odhadu sklonu b

Ad 2) Můžeme už prosím *konečně* říct něco o vztahu mezi X a Y ?

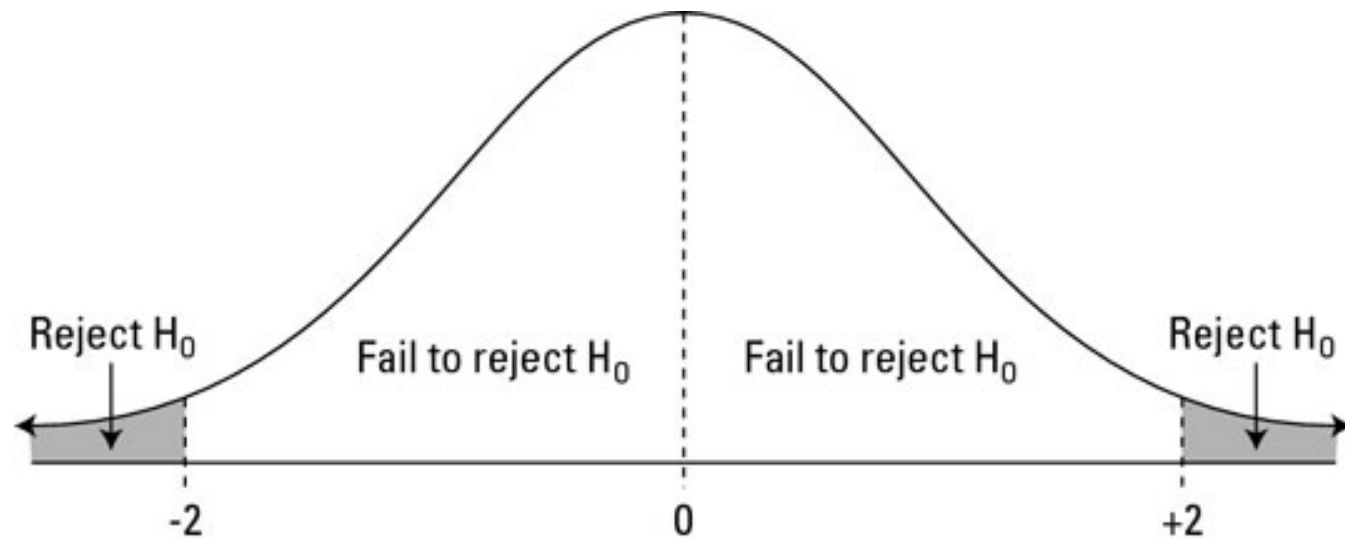
- a konečně spočítáme t-statistiku koeficientu b (jako v t-testu)

$$t = \frac{b}{s_b}$$

- a podle statistické tabulky zjistíme, s jakou pravděpodobností je toto získané b skutečně odlišné od 0

Co to sakra znamená?

- Je hodnota koeficientu b dostatečně daleko od 0? Pokud ano, je mezi nezávislou a závislou proměnnou pravděpodobně **statistický vztah a vyvrácíme H_0**
- Jak měříme „dostatečnost“ vzdálenosti od 0? Porovnáním hodnoty b s mírou chyby v odhadu velikosti b .



Aplikace: predikce vztahu k EU na základě vztahu strany k imigraci

- Predikuje pozice strany vůči imigraci (na škále 0-10, kde 10 je silná restrikce imigrace) pozici strany vůči EU (na škále 1-7, kde 7 je silná podpora integraci)?
- Data z Chapel Hill Expert Survey (2014)

	A	B	C
1	eu_position	immigrate_policy	
2	6	2.8	
3	6	3.4000001	
4	6.25	1.8	
5	6.1999998	2.2	
6	6.4000001	6.1999998	
7	6.5999999	6.5999999	
8	6.4000001	4.1999998	
9	6.5999999	5.4000001	
10	5	7.5999999	
11	6.4000001	4.4000001	
12	2.5999999	9.6000004	
13	3.4000001	1.8	
14	2.5	8.5	
15	6	5.5	
16	7	2.5999999	
17	5.5	7.0999999	
18	4.6999999	2.8	

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.296859493							
R Square	0.088125559							
Adjusted R Square	0.084684523							
Standard Error	1.703059704							
Observations	267							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	74.28009692	74.2801	25.610185	7.82155E-07			
Residual	265	768.6092746	2.900412					
Total	266	842.8893715						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>ower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	5.977427739	0.252457526	23.67696	2.4065E-67	5.480349908	6.47450557	5.48035	6.47450557
immigrate_policy	-0.221459424	0.043761058	-5.06065	7.8216E-07	-0.307623033	-0.135295815	-0.30762	-0.135295815

[illegible]

Regression Statistics

Multiple R	0.296859493
R Square	0.088125559
Adjusted R Square	0.084684523
Standard Error	1.703059704
Observations	267

[illegible]

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	74.28009692	74.2801	25.610185	7.82155E-07
Residual	265	768.6092746	2.900412		
Total	266	842.8893715			

a

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	5.977427739	0.252457526	23.67696	2.4065E-67	5.480349908	6.47450557	5.48035	6.47450557
immigrate_policy	-0.221459424	0.043761058	-5.0606	7.8216E-07	-0.307623033	-0.135295815	-0.30762	-0.135295815

b

 $P(b=0)$



Odkazy

- Gerring, John, and Dino Christenson. *Applied Social Science Methodology: An Introductory Guide*. Cambridge, United Kingdom ; New York: Cambridge University Press, 2017.
- Hendl, Jan. *Přehled statistických metod zpracování dat*. Praha: Portál, 2013.