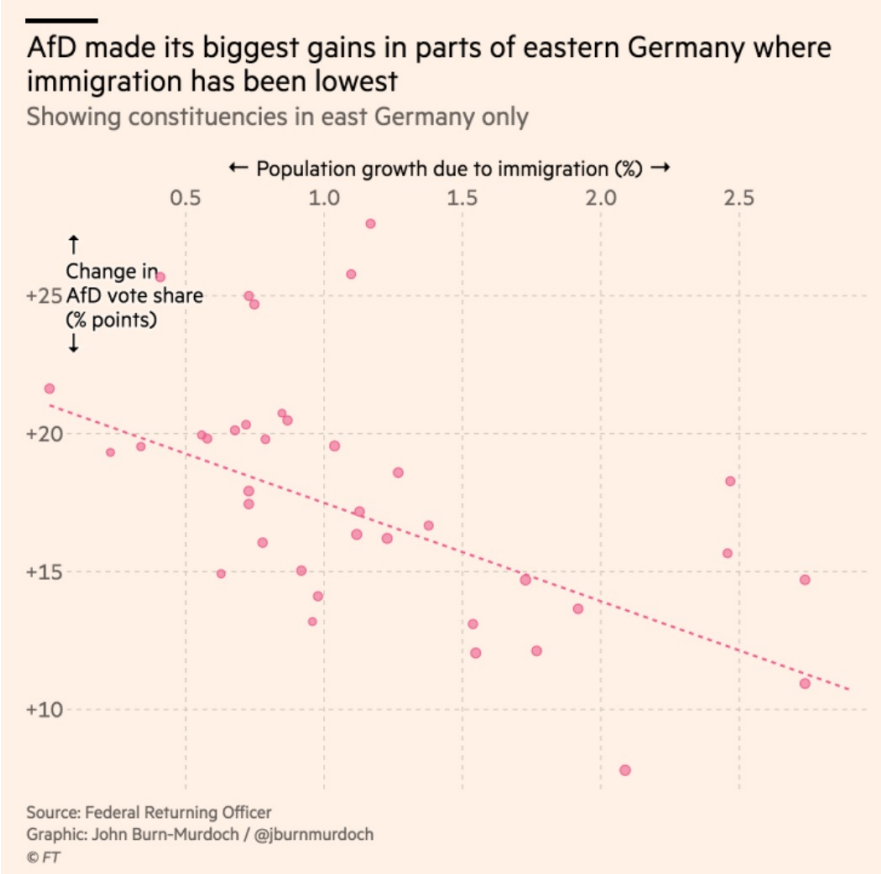
Regresní analýza

ANALÝZA DAT

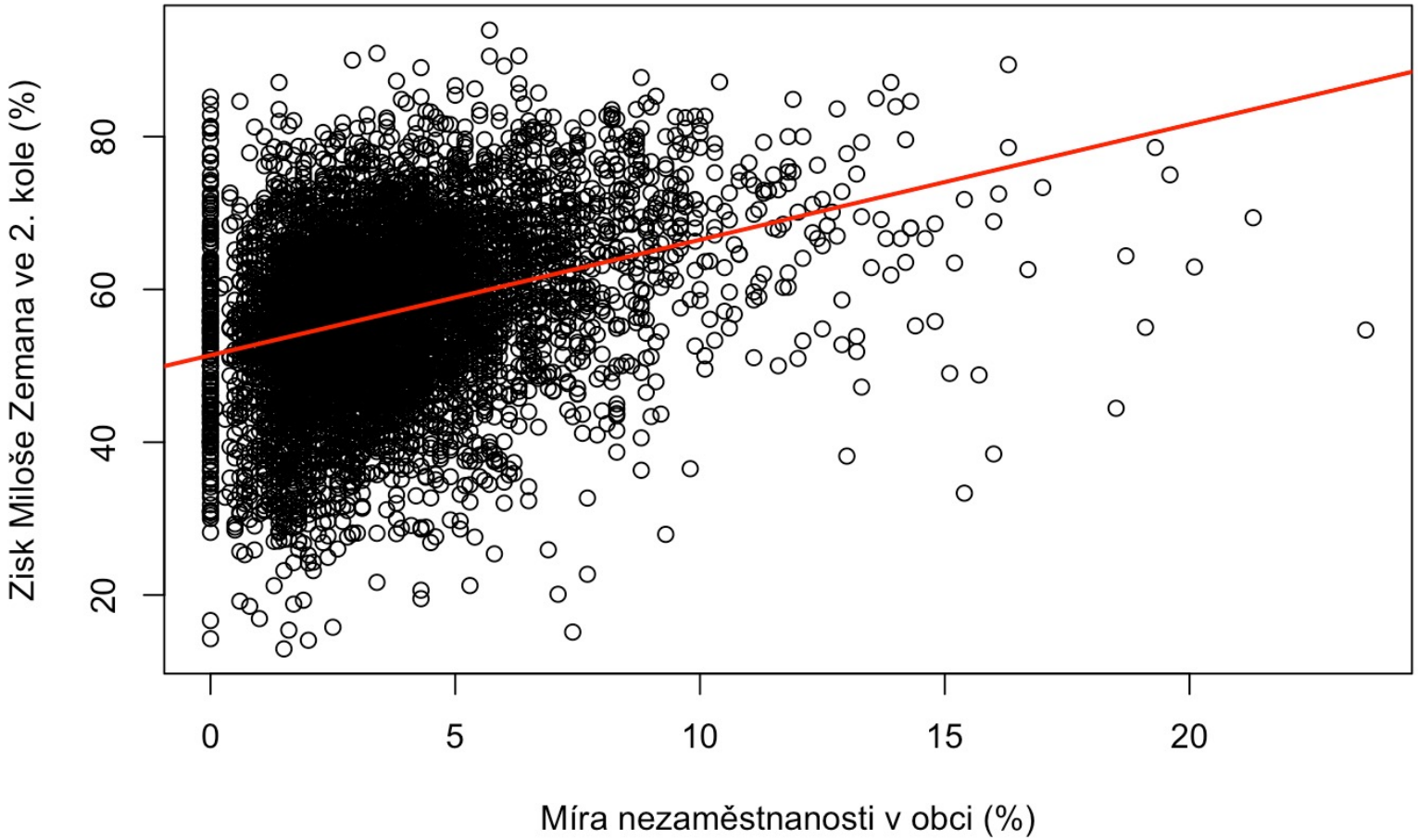
1. DUBNA 2021

Regresní analýza v médiích



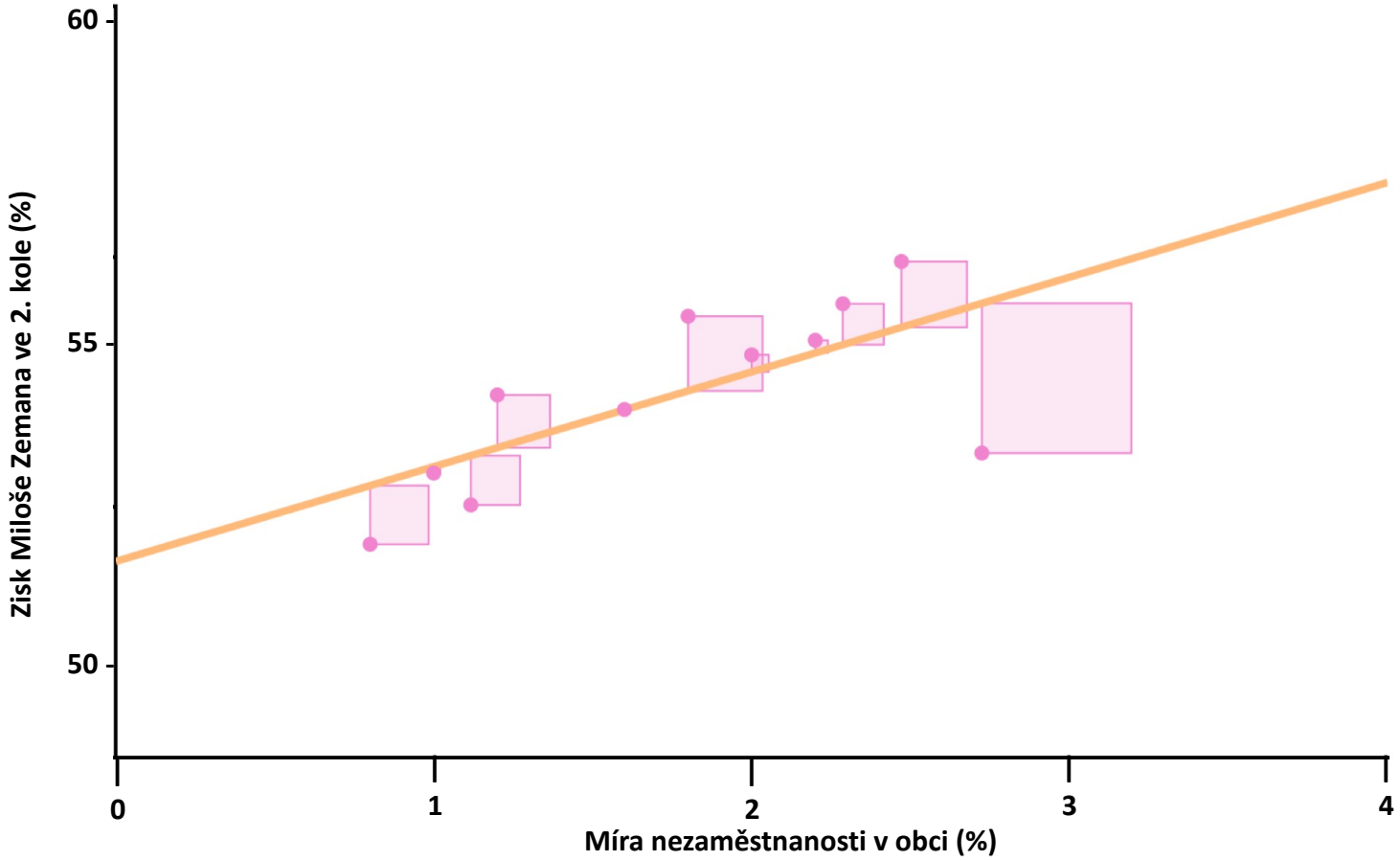
ft.com

Logika regresní analýzy



Regresní analýza

OLS (*ordinary least squares*, metoda nejmenších čtverců)



students.brown.edu

Jednoduchá regresní analýza

- vyjadřuje vztah mezi dvěma proměnnými
  - nezávisle proměnná X – proměnná, která způsobuje změnu (nezaměstnanost)
  - závisle proměnná Y – proměnná, která je měněna (zisk Miloše Zemana)
- využívá k tomu přímku – ta se velmi lehce popisuje (je třeba popsat (1) průnikem a (2) sklonem)

$$Y = b_0 + b_1 * X$$

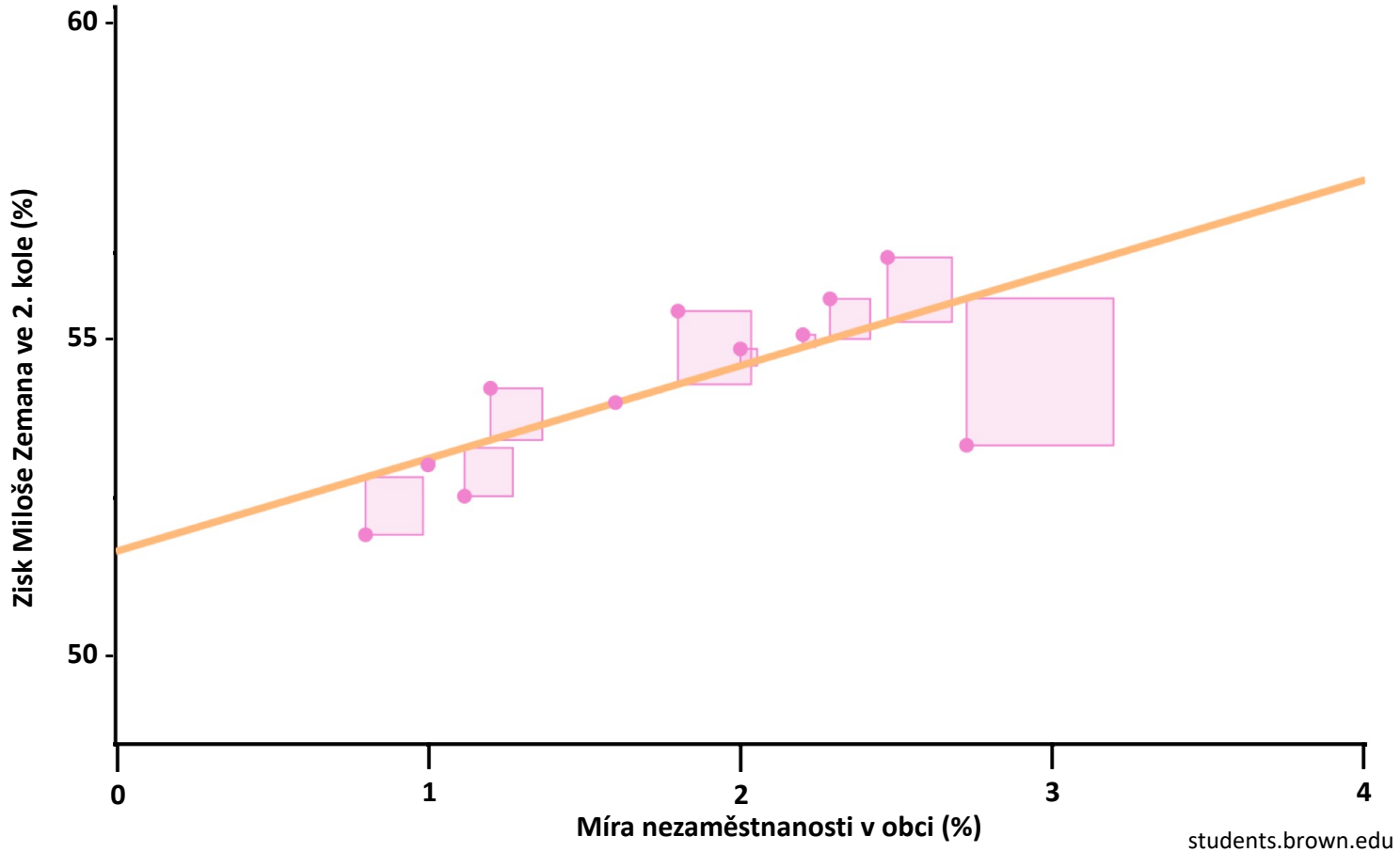
  - b_0 = průnik (hodnota Y, když X je rovno nule)
  - b_1 = sklon (změna v hodnotě Y v případě navýšení hodnoty X o jednu jednotku)
- tento model by byl perfektním lineárním vztahem
- v aktuálním výzkumu toto ale nikdy nenastane – proto potřebujeme chybu predikce

$$Y = b_0 + b_1 * X + e$$

- chyba predikce *e* reprezentuje další nepozorované faktory (vedle proměnné X ovlivňující Y)

Regresní analýza

OLS (*ordinary least squares*, metoda nejmenších čtverců)



students.brown.edu

Příklad jednoduché regrese

- zkoumáme vztah mezi ziskem Miloše Zemana a mírou nezaměstnanosti v obci
- pro každé pozorování *i* můžeme napsat:

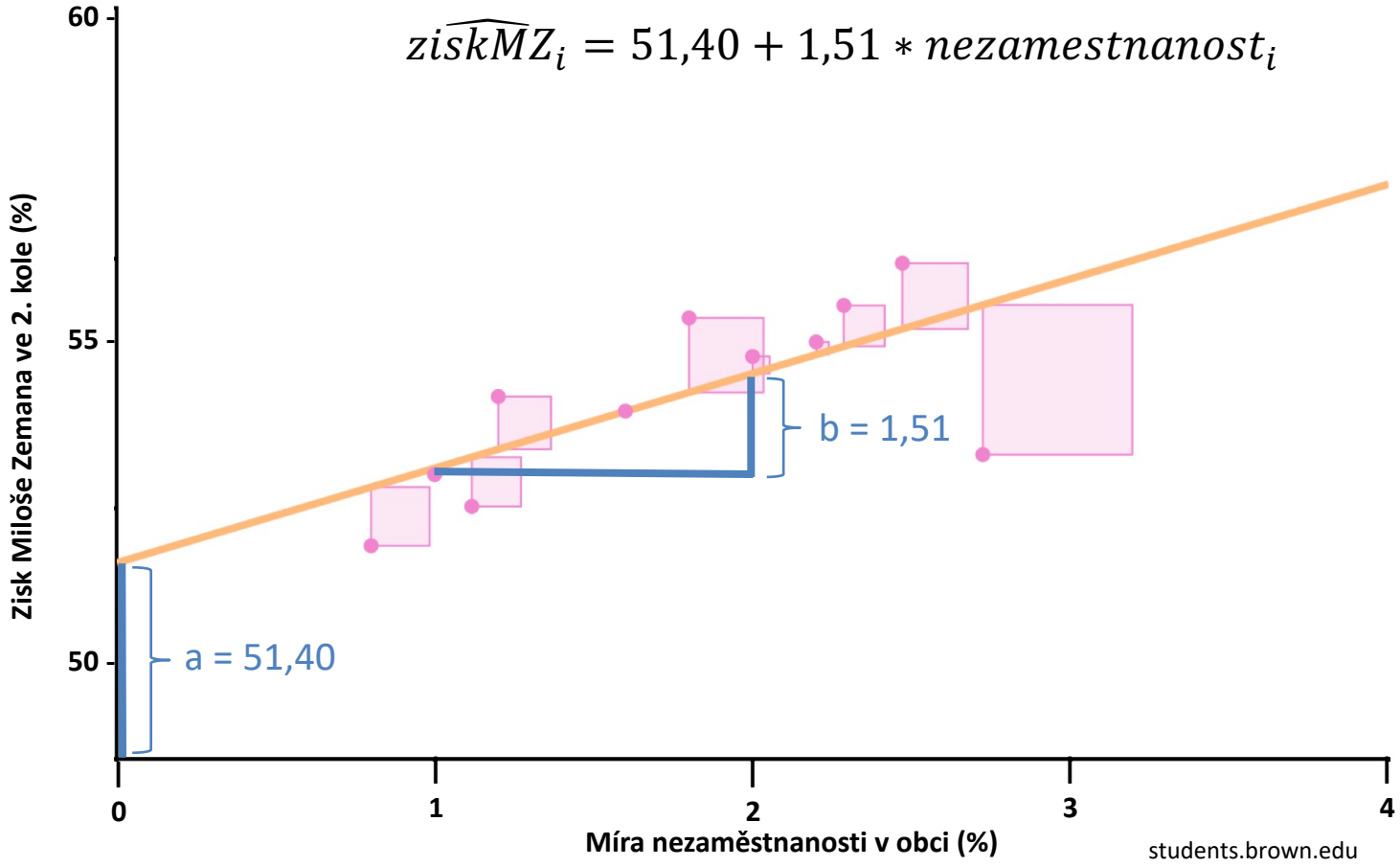
$$ziskMZ_i = b_0 + b_1 * nezamestnanost_i + e_i$$

- jak vybereme nejlepší linku (neboli jak určíme parametry průniku b_0 a sklonu b_1)?
  - vybereme takovou linku, aby výsledná chybovost *e* byla co nejmenší
  - jednou metodou pro odhalení nejmenší výsledné chybovosti je metoda nejmenších čtverců (OLS)

- https://www.geogebra.org/m/XUkhCJRj

Regresní analýza

OLS (*ordinary least squares*, metoda nejmenších čtverců)



Interpretace jednoduché regresní analýzy

$$\widehat{ziskMZ}_i = 51{,}40 + 1{,}51 * nezamestnanost_i$$

- „při zvýšení nezaměstnanosti v obci o jeden procentní bod se zisk pro Miloše Zemana v obci zvýší o 1,51 procentního bodu“
- „pokud je nezaměstnanost v obci nulová, zisk Miloše Zemana zde činí 51,40 %“

Hodnocení koeficientů

- vypočítaný koeficient vztahu mezi závisle a nezávisle proměnnou má určitou chybu vyplývající z reziduí a počtu měření
- tuto chybu testujeme proti nulové hypotéze, že koeficient je nulový (tedy že žádný vztah mezi oběma proměnnými neexistuje)
- pomocí t-hodnoty spočítáme vzdálenost koeficientů od nuly (počet směrodatných odchylek)
- od této hodnoty odvodíme p-hodnotu
  - pravděpodobnost platnosti nulové hypotézy, že koeficient je nulový (tedy že zde není žádný vztah)
- čím je p-hodnota nižší, s tím vyšší pravděpodobností můžeme zamítnout nulovou hypotézu
  - tedy narůstá pravděpodobnost platnosti alternativní hypotézy, že zde vztah mezi nezávisle a závisle proměnnou existuje
- většinou je pro nás zásadní konvenční úroveň 95% jistoty (případně 99 %)

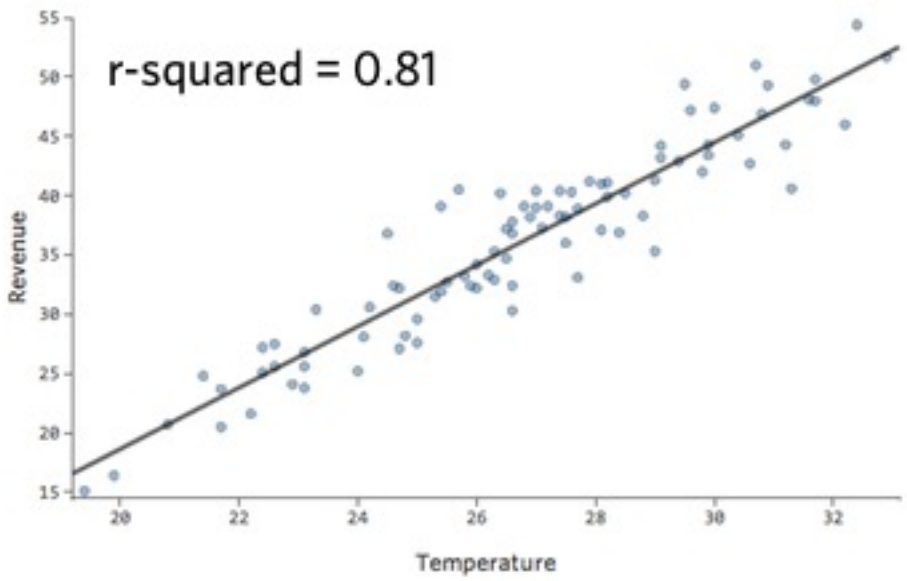
Kvalita modelu

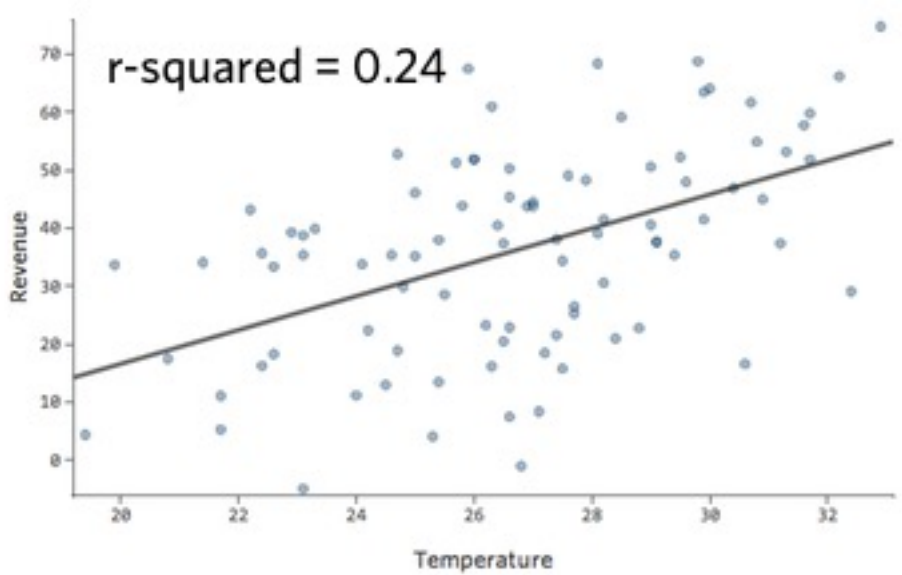
- měříme pomocí procenta rozptylu závisle proměnné, který je vysvětlen regresním modelem
- výpočet pracuje s chybovostí naivního modelu predikce podle průměru závisle proměnné Y (celkový součet čtverců) a regresního modelu (součet čtverců reziduí)

$$R^2 = \frac{celkový\ součet\ čtverců\ -\ součet\ čtverců\ reziduí}{celkový\ součet\ čtverců} = \frac{vysvětlený\ rozptyl}{celkový\ rozptyl}$$

- $0 \leq R^2 \leq 1$
- interpretace: Část celkového rozptylu dat vysvětlená aplikovaný modelem.
- R^2 se vždy navyšuje s přidáním další nezávisle proměnné
- proto se používá i upravené R^2, které penalizuje vysvětlovací schopnost modelu na základě počtu aplikovaných proměnných

Kvalita modelu





docs.statwing.com

Vícenásobná regresní analýza

- v praxi nikdy nedochází k tomu, že závisle proměnnou Y ovlivňuje jenom jedna nezávisle proměnná X
- při vytváření skutečně výstižných analytických modelů je třeba zahrnout i další vlivné proměnné
- v rámci vícenásobné regresní analýzy tak odhalujeme sílu efektu hned několika nezávisle proměnných (X_1, X_2, X_3 atd.) na závisle proměnnou Y
- většinou stále existuje jedna hlavní nezávisle proměnná X_1 a ostatní proměnné X_2, X_3 atd. považujeme za tzv. kontrolní proměnné
- nezávisle (kontrolní) proměnné nevkládáme do analytického modelu nikdy (!) náhodně, ale vždy na základě předchozího výzkumu a předpokladu, co má skutečně určitý vliv

$$Y = b_0 + b_1 * X_1 + b_2 * X_2 + b_2 * X_2 + \cdots + e$$

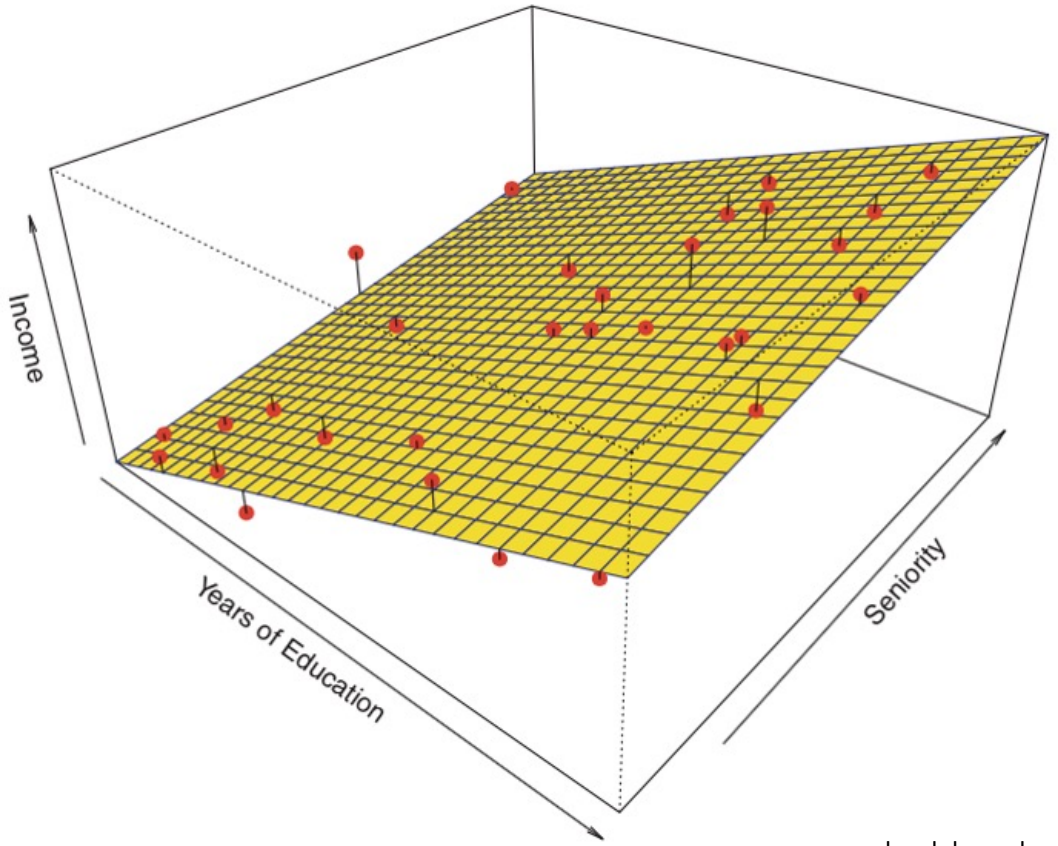
Příklad vícenásobné regrese

- stále zkoumáme vztah mezi ziskem Miloše Zemana a nezaměstnaností
- z již proběhlých výzkumů ale víme, že volební zisky v obci ovlivňuje také průměrný věk nebo místní podíl vysokoškoláků
- pro každé pozorování *i* můžeme napsat:

$$ziskMZ_i = b_0 + b_1 * nezamestnanost_i + b_2 * vek_i + b_3 * podilVS_i + e_i$$

- nyní už nevybíráme nejlepší linku, ale vícedimenzionální prostory, které prostupují body takovým způsobem, aby chybovost byla opět co nejmenší
  - logika je tedy velmi podobná jednoduché regresi, jen se pohybujeme ve větším množství dimenzí
  - i když si toto obtížně představujeme, pro statistické programy to není v podstatě žádný rozdíl

Vícenásobná regresní analýza



sphweb.bumc.bu.edu

Interpretace vícenásobné regresní analýzy

$$\widehat{ziskMZ}_i = 37{,}84 + 0{,}96 * nezamestnanost_i + 0{,}60 * vek_i + (-1{,}60) * podilVS_i$$

- „při zvýšení nezaměstnanosti o jeden procentní bod a stálosti všech ostatních parametrů (věku a podílu vysokoškoláků) se zisk pro Miloše Zemana v obci zvýší o 0,96 procentního bodu“
- „při zvýšení průměrného věku v obci o jeden rok a stálosti všech ostatních parametrů (nezaměstnanosti a podílu vysokoškoláků) se zisk pro Miloše Zemana v obci zvýší o 0,60 procentního bodu“
- „při zvýšení podílu vysokoškoláků o jeden procentní bod a stálosti všech ostatních parametrů (nezaměstnanosti a věku) se zisk pro Miloše Zemana v obci sníží o 1,60 procentního bodu“
- „pokud je nezaměstnanost nulová, průměrný věk v obci je nulový a nežije zde ani jeden vysokoškolák je zisk Miloše Zemana 37,84 %“

Regresní analýza v praxi

Table 4. Ordinary least squares regression analyses of plenary sessions attendance, plenary sessions voting activity, and committee meetings attendance.

	Dependent variable:					
	Plenary sessions attendance		Plenary sessions voting activity		Committee meetings attendance	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Cumulative MOH	1.304		0.853		−1.161	
	(1.065)		(0.573)		(1.550)	
MC		3.786*		3.271***		7.032**
		(2.213)		(1.210)		(3.065)
MMC		0.566		−1.050		−7.337
		(3.229)		(1.766)		(4.472)
RC		0.853		−0.639		−2.921
		(2.485)		(1.359)		(3.441)
MRC		−15.763**		−1.659		−32.653***
		(6.277)		(3.433)		(8.694)
Female	1.068	1.122	0.779	0.877	−3.697	−3.674
	(2.274)	(2.228)	(1.224)	(1.218)	(3.311)	(3.086)
Age	0.033	0.012	0.082	0.067	0.115	0.046
	(0.096)	(0.095)	(0.051)	(0.052)	(0.139)	(0.131)
Education	−3.063	−2.271	0.030	0.163	2.027	3.442
	(2.430)	(2.374)	(1.308)	(1.298)	(3.537)	(3.288)
Parliamentary	−2.168***	−2.272***	−1.216***	−1.323***	−1.183	−1.461
Experience	(0.821)	(0.804)	(0.442)	(0.440)	(1.195)	(1.114)
Geographic	−0.181	−0.286	0.714	0.691	−0.494	−0.693
Area	(0.923)	(0.897)	(0.497)	(0.490)	(1.344)	(1.242)
VV	2.094	1.645	9.565***	9.418***	−0.028	−0.670
	(3.549)	(3.468)	(1.910)	(1.896)	(5.167)	(4.803)
KSČM	11.858***	10.904***	−5.539***	−5.712***	11.529**	9.750**
	(3.027)	(2.956)	(1.630)	(1.617)	(4.407)	(4.094)
ODS	3.226	3.405	7.462***	7.913***	6.903*	8.346**
	(2.525)	(2.548)	(1.359)	(1.393)	(3.676)	(3.529)
TOP 09	7.159***	6.909**	10.087***	10.711***	4.571	5.756
	(2.674)	(2.786)	(1.439)	(1.524)	(3.893)	(3.859)
Constant	81.337***	81.888***	77.212***	77.479***	68.658***	70.046***
	(5.567)	(5.410)	(2.997)	(2.959)	(8.106)	(7.493)
N	132	132	132	132	132	132
R^2	0.221	0.285	0.574	0.596	0.092	0.245
Adjusted R^2	0.157	0.206	0.539	0.552	0.017	0.162
F Statistic	3.441***	3.610***	16.325***	13.406***	1.228	2.946***
	(df = 10;121)	(df = 13;118)	(df = 10;121)	(df = 13;118)	(df = 10;121)	(df = 13;118)

Note: *p*-values: ***$p < .01$, **$p < .05$, *$p < .1$.

nezávisle proměnné

průnik

index determinace

koeficient

směrodatná chyba

p-hodnoty označené počtem hvězdiček (větší množství hvězd znamená větší jistotu vlivu proměnné)

Hájek, L. (2017). The effect of multiple-office holding on the parliamentary activity of MPs in the Czech Republic. *The Journal of Legislative Studies, 23*(4), pp. 484-507.

Regresní analýza v praxi

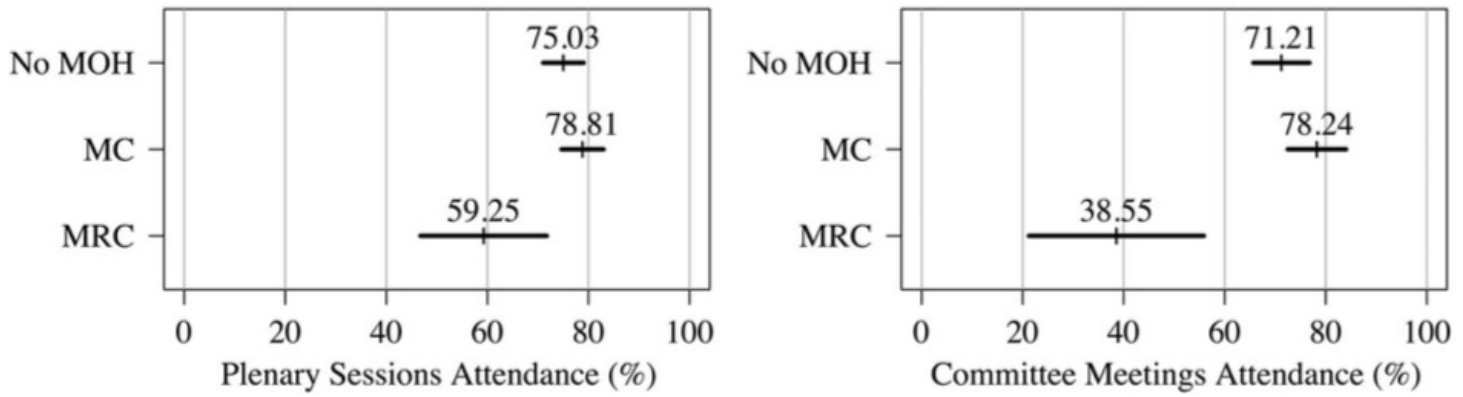


Figure 1. The effect of multiple-office holding on plenary sessions attendance and committee meetings attendance.

Note: The expected values and associated 95 per cent confidence intervals are simulated using the Zelig package in R (R Core Team, 2007). The simulations are conducted for male deputies with a university education, average age, parliamentary experience, geographic proximity, and affiliated to ČSSD as the largest (opposition) party. All other mandates are held at zeros and the only change is between zero and one in the case of the analysed mandates.

Hájek, L. (2017). The effect of multiple-office holding on the parliamentary activity of MPs in the Czech Republic. *The Journal of Legislative Studies, 23*(4), pp. 484-507.

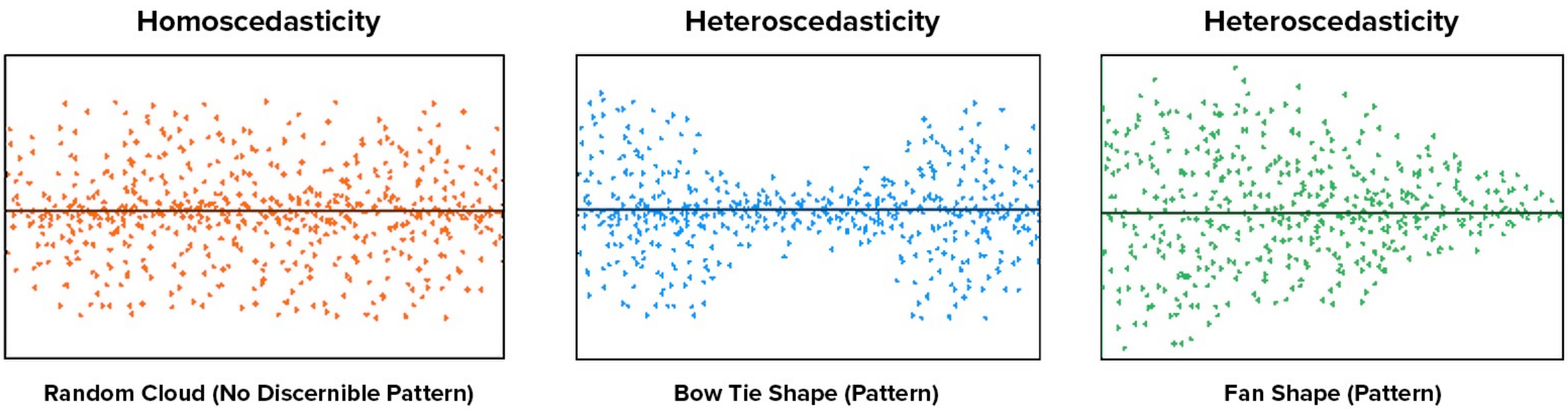
Předpoklady regresní analýzy

1. typ proměnných
   - závisle proměnná je intervalová nebo poměrová, spojitá a co nejméně omezená
   - nezávisle proměnná je intervalová nebo poměrová; může být i nominální, ale jen dichotomická
2. lineární vztah mezi závisle proměnnou a nezávisle proměnnými
   - v jiném případě může být mezi proměnnými vztah (například kvadratický) a OLS regrese ho neodhalí
   - pokud vztah není lineární, je třeba sáhnout k jiné podobě modelu
3. multikolinearita
   - nezávisle proměnné by mezi sebou neměly být příliš vysoce korelovány
   - testujeme pomocí VIF skóre – pokud je nižší než 5, je vše v pořádku
   - pokud problém nastane, některé proměnné by měly být vyřazeny, aby nedocházelo k duplikaci

Předpoklady regresní analýzy

4. pozor na odlehlé hodnoty
   - mohou velmi značně ovlivnit podobu regresní přímky
   - jde o odlehlé hodnoty jak na straně závisle proměnné, tak nezávisle proměnných
   - zjišťujeme pomocí Cookovy vzdálenosti – pokud je vyšší než 1, jde o odlehlou hodnotu
   - řešením je vyřazení z modelu – je to ale třeba nějak obhájit
5. homoskedasticita
   - rozptyl reziduálních hodnot je podobný na všech místech hodnot závisle proměnné Y
   - odhalujeme skrze bodový graf závisle proměnné Y na ose x a hodnot reziduí na ose y
   - pokud vidíme heteroskedasticitu (rozložení reziduí vykazuje nějaký vzorec), pravděpodobně nejde o lineární vztah nebo nám uniká nějaká další proměnná

Homoskedasticita vs. heteroskedasticita



clevertap.com

Předpoklady regresní analýzy

6. nezávislost reziduí
   - pro jakákoliv dvě pozorování nesmí být vztah mezi jejich rezidui (nesmí být autokorelace)
   - odhalujeme pomocí Durbin-Watsonova testu – hodnoty mezi 1 a 3 jsou v pořádku
   - porušení této podmínky je obvyklé v modelech s časovými proměnnými – když je podmínka porušena, je třeba časové proměnné upravit
7. normální distribuce reziduí s nulovým průměrem
   - jinými slovy distribuce hodnot proměnných by se měla blížit normálnímu rozložení
   - odhalujeme pomocí „q-q plot“ – teoretické kvantily proti standardizovaným reziduím
   - v ideálním případě data leží na diagonální ose – jiná podoba značí jiný než lineární vztah
   - v případě porušení podmínky je třeba opět uvažovat nad aplikací jiného modelu

- pro pochopení principů fungování OLS regrese - http://students.brown.edu/seeing-theory/regression-analysis/index.html

Shrnutí

- regresní analýza je jedním z nejlepších nástrojů pro popis vztahu mezi proměnnými
- data prokládá přímkou (plochou atd.) a hledá nejlepší vyjádření vztahu
- jednou z metod hledání ideálního vztahu je metoda nejmenších čtverců
- v případě prezentace výsledků jsou zásadní koeficienty nezávisle proměnných a jejich p-hodnota
- při vícenásobné regresy využíváme kontrolní proměnné
- pozor na správnou interpretaci koeficientů!
- pozor na splnění předpokladů pro regresní analýzu!