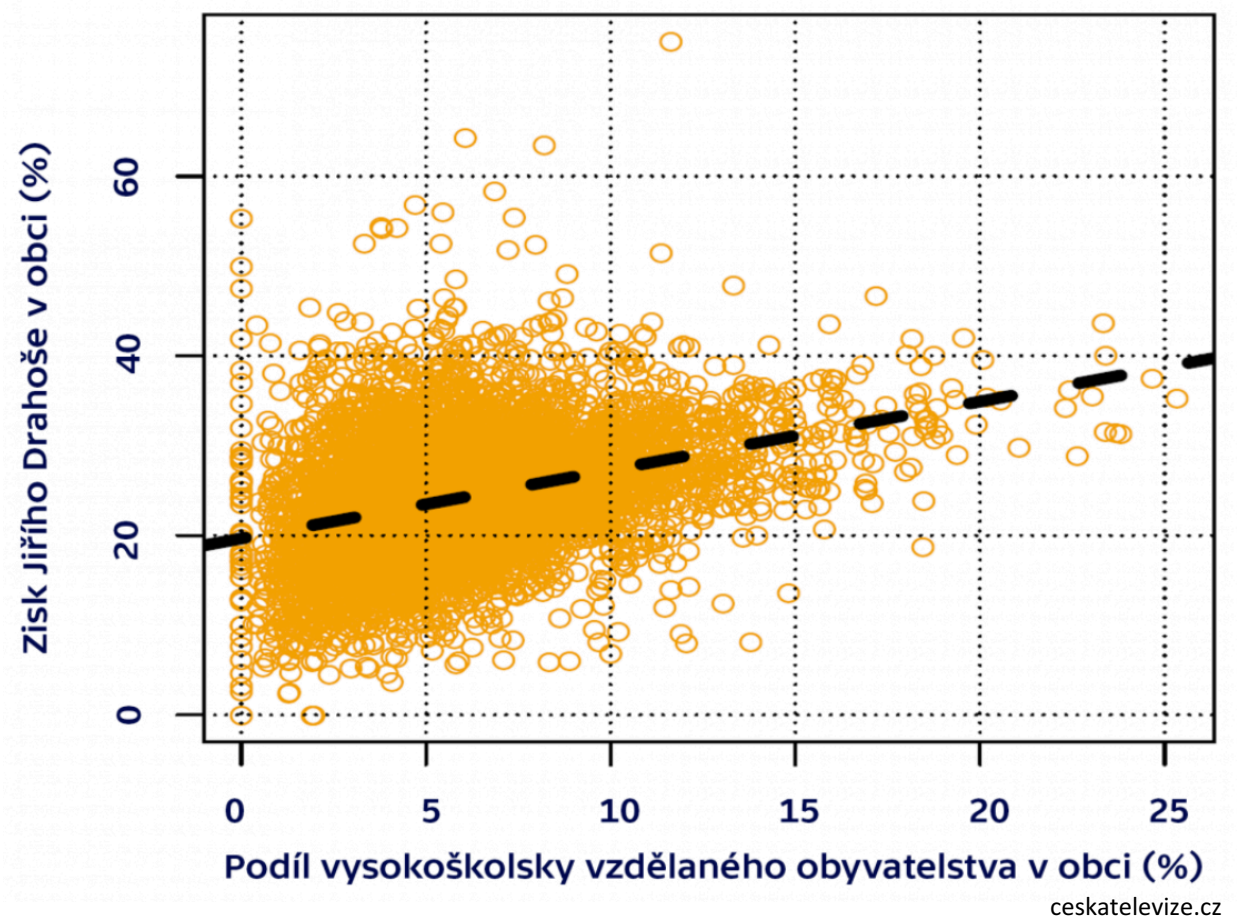
Regresní analýza – úvod a principy

METODOLOGICKÝ PROSEMINÁŘ II

TÝDEN 8 | 11. DUBNA 2018

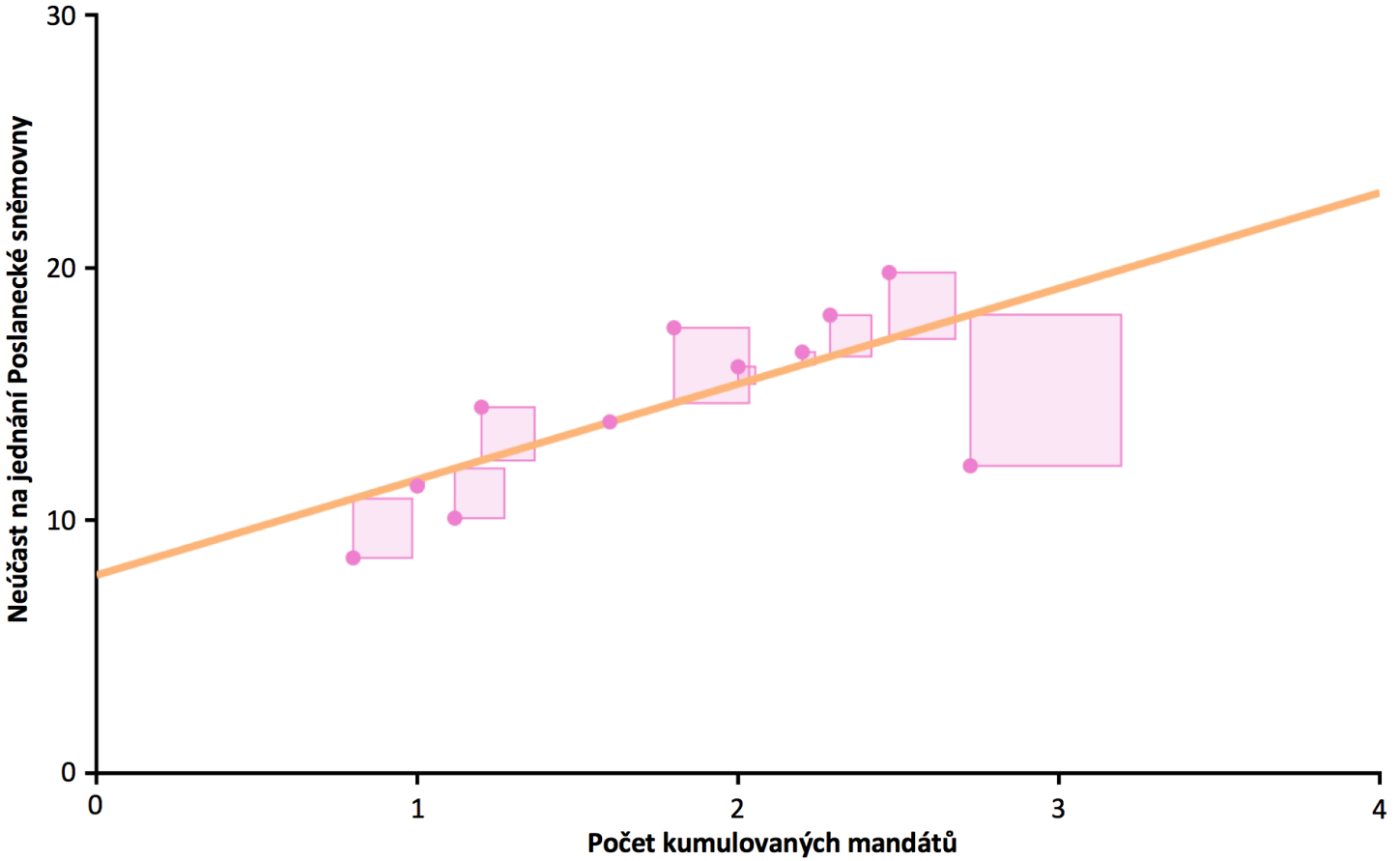
Logika regresní analýzy



ceskatelevize.cz

Regresní analýza

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



students.brown.edu

Jednoduchá regresní analýza

- vyjadřuje vztah mezi dvěma proměnnými
- využívá k tomu přímku – ta se velmi lehce popisuje (je třeba popsat (1) průnikem a (2) sklonem)

$$Y = a + b * X$$

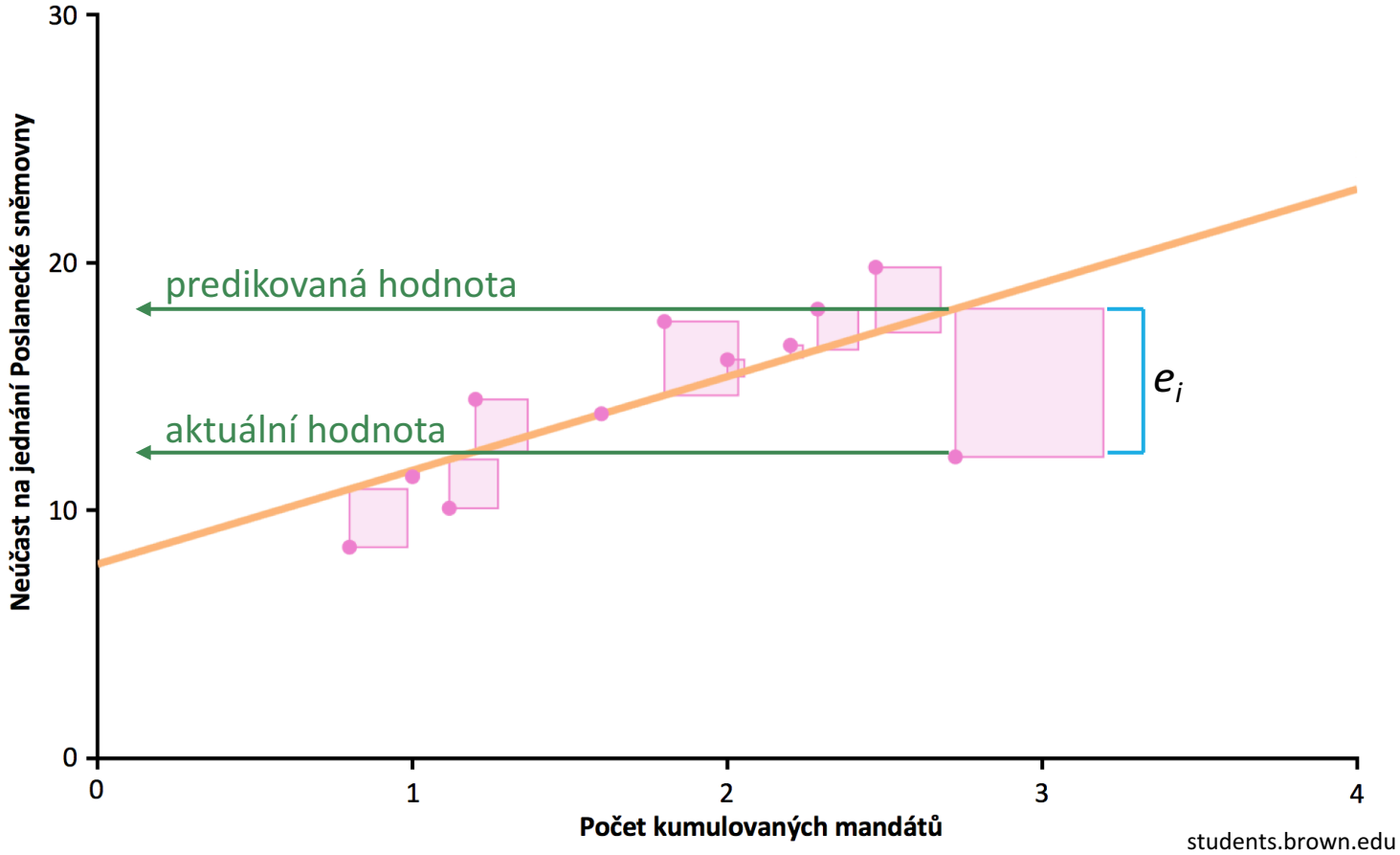
  - *a* = průnik (hodnota Y, když X je rovno nule)
  - *b* = sklon (změna v hodnotě Y v případě navýšení hodnoty X o jednu jednotku)
- tento model by byl perfektním lineárním vztahem
- v aktuálním výzkumu toto ale nikdy nenastane – proto potřebujeme chybu predikce

$$Y = a + b * X + e$$

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

Regresní analýza

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



Příklad jednoduché regrese

- zkoumáme vztah mezi počtem kumulovaných mandátů a neúčastí na jednání Poslanecké sněmovny
- pro každé pozorování *i* můžeme napsat:

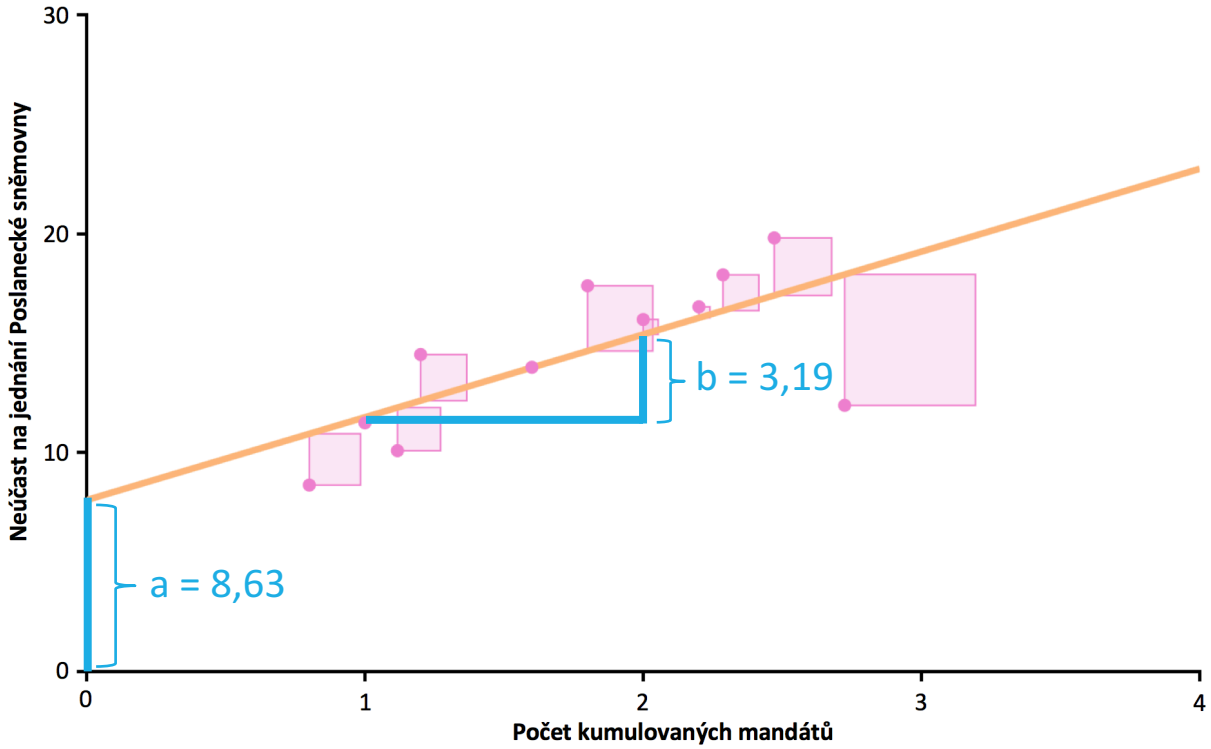
$$neúčast_i = a + b * mandátů_i + e_i$$

- jak vybereme nejlepší linku (neboli jak určíme parametry průniku *a* a sklonu *b*)?
  - vybereme takovou linku, aby výsledná chybovost *e* byla co nejmenší
  - jednou metodou pro odhalení nejmenší výsledné chybovosti je metoda nejmenších čtverců (OLS)
- http://www.dangoldstein.com/regression.html

Regresní analýza

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

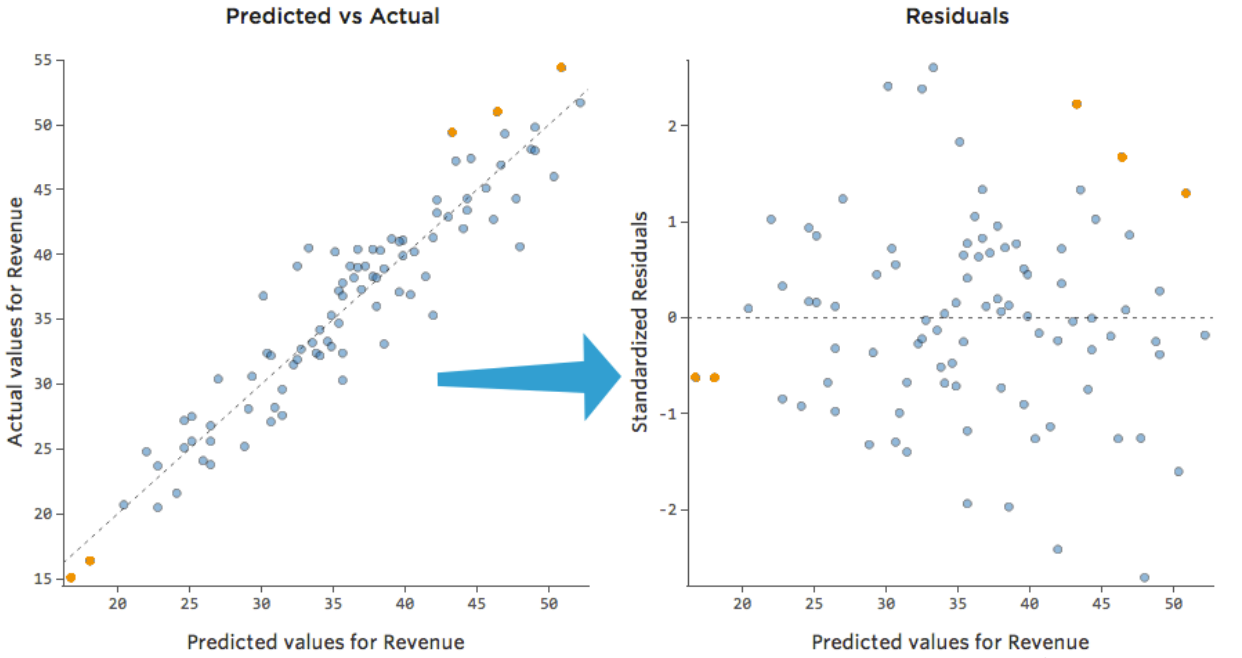
$$\widehat{neúčast}_i = 8{,}63 + 3{,}19 * mandátů_i$$



students.brown.edu

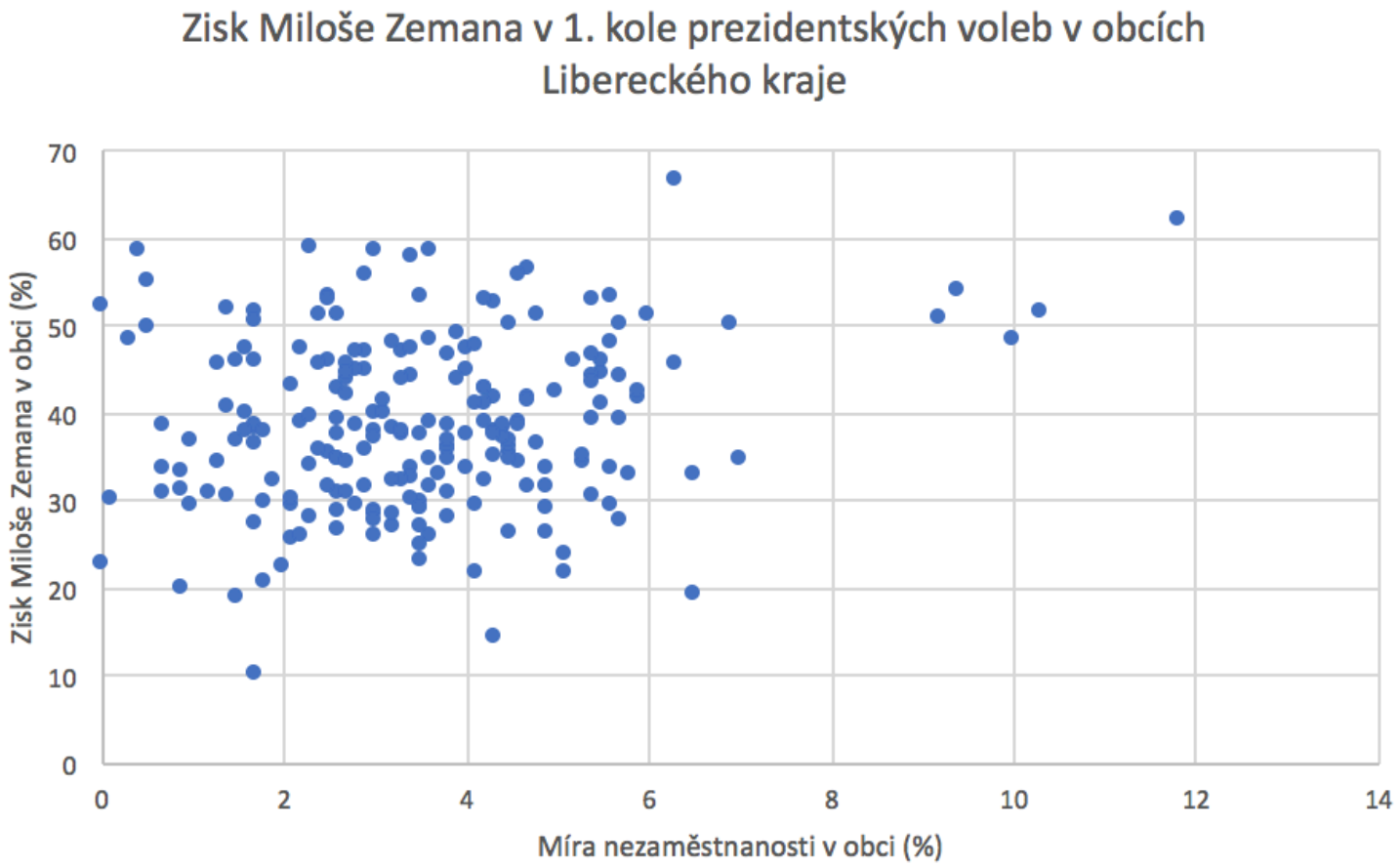
Regresní diagnostika – analýza reziduí

- graf reziduí je bodový graf, kde na ose X je zakreslena nezávisle proměnná X a na ose Y je vyobrazena hodnota reziduí *e*
- pomáhá odhalit systematické vzorce v chybovosti
- součet reziduí je vždy roven nule



docs.statwing.com

Regresní analýza v Excelu



Regresní analýza v Excelu

1. využít doplněk „Analýza dat“ (*Data Analysis*) – návod na případnou instalaci
2. v Excelu v záložce Data vybereme dlaždici „Analýza dat“
3. zvolit nástroj „Regrese“
4. vybereme rozsah proměnné Y (závisle proměnná)
   - včetně názvu proměnné a vybrat možnost popisků
5. vybereme rozsah proměnné X (nezávisle proměnná)
   - včetně názvu proměnné a vybrat možnost popisků
6. vybereme hladinu spolehlivosti (typicky 95 %)
7. zvolíme možnost zakreslení na nový list (ten pojmenujeme)
8. v nabídce navíc vybereme možnost reziduí, grafu s rezidui a grafu regresní přímky

Regresní analýza v Excelu
Interpretace výsledků analýzy

		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
průnik	Intercept	34,74457634	1,43225382	24,25867249	3,24537E-63	31,92136933	37,56778334
nezávisle proměná	nezamestnanost	1,08097976	0,360383559	2,999525737	0,003025991	0,370604704	1,791354817

parametry regresní přímky *a* (34,75) a *b* (1,08)

standardní chyba parametrů

koeficient vydělený standardní chybou (slouží ke statistickým testům nulové vs. alternativní hypotézy)

pravděpodobnost nulové hypotézy, že parametr je roven 0 (tedy bez efektu); čím je toto číslo nižší, tím máme větší jistotu, že nezávisle proměnná má skutečně vliv (rozhodující konvenční hranice je 0,05)

intervaly spolehlivosti okolo parametrů
(interval obsahuje pravý regresní koeficient populace v 95 % případů hypoteticky opakovaných výběrů vzorků)

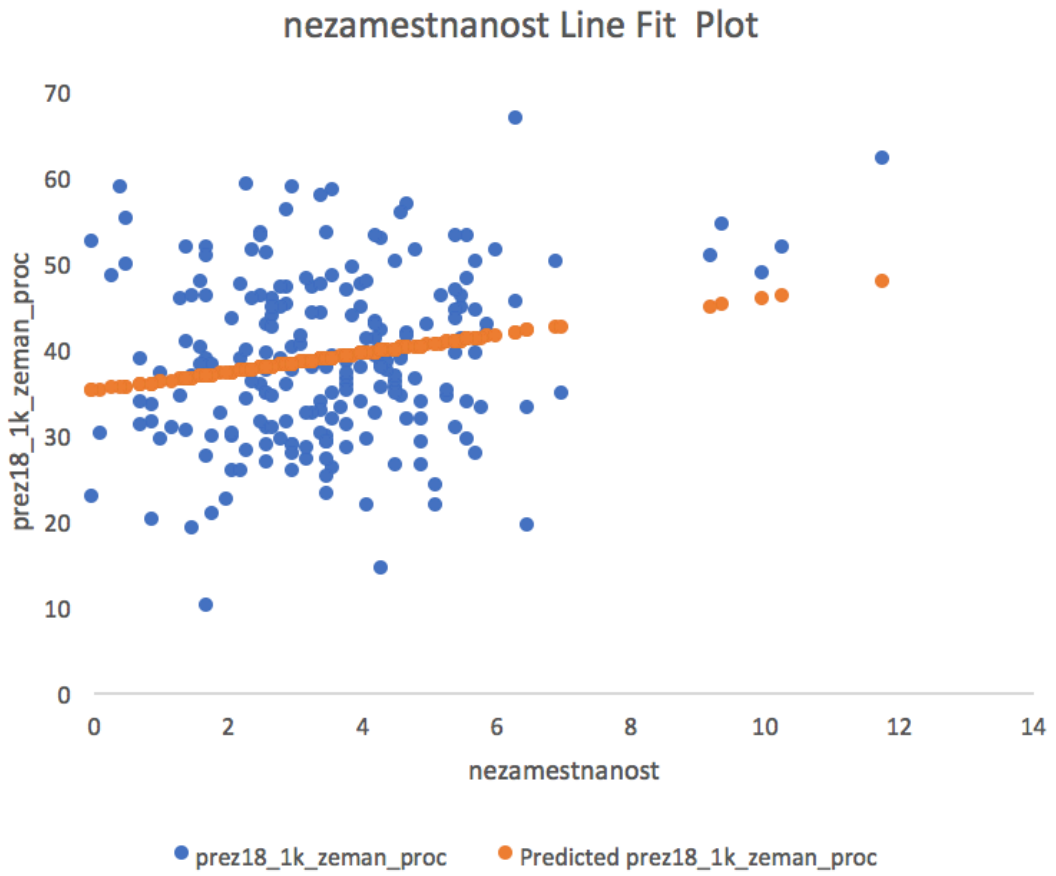
Regresní analýza v Excelu
Interpretace výsledků analýzy

Regression Statistics		
Multiple R	0,201316276	korelační koeficient ukazující sílu lineárního vztahu
R Square	0,040528243	koeficient determinace <0,1>; říká, kolik bodů je na regresní přímce a jak dobrá je tedy prediktivní schopnost modelu
Adjusted R Square	0,036023681	koeficient determinace upravený vzhledem k počtu nezávisle proměnných
Standard Error	9,526153362	standardní chyba regrese
Observations	215	počet případů v regresi

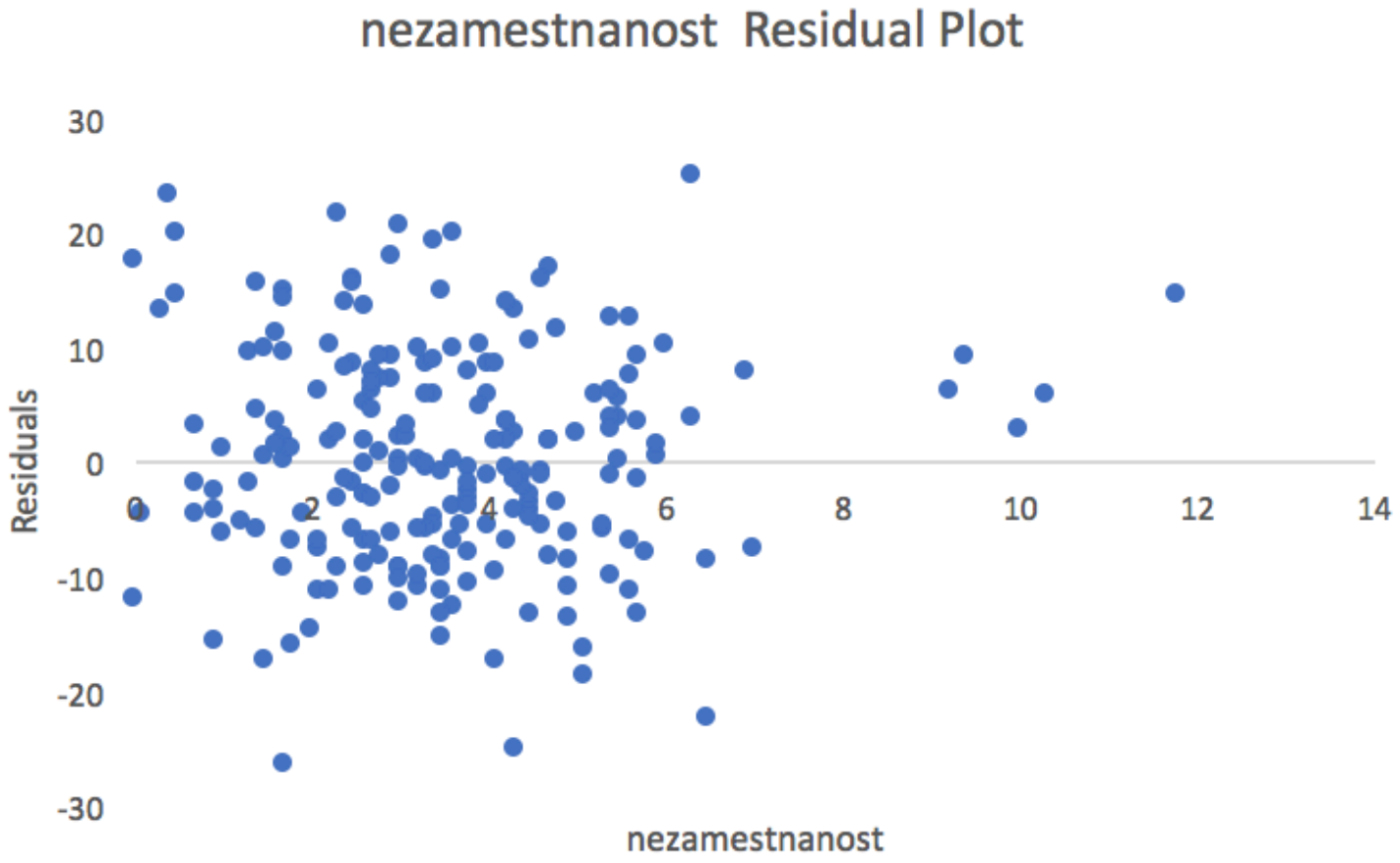
$$\widehat{ziskMZ}_i = 34{,}75 + 1{,}08 * nezaměstnanost_i$$

- „při zvýšení nezaměstnanosti o jeden procentní bod se zisk pro Miloše Zemana v obci Libereckého kraje zvýší o 1,08 procentního bodu“
- „pokud je nezaměstnanost nulová, je zisk Miloše Zemana 34,75 procent“

Regresní analýza v Excelu
Interpretace výsledků analýzy



Regresní analýza v Excelu
Interpretace výsledků analýzy



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

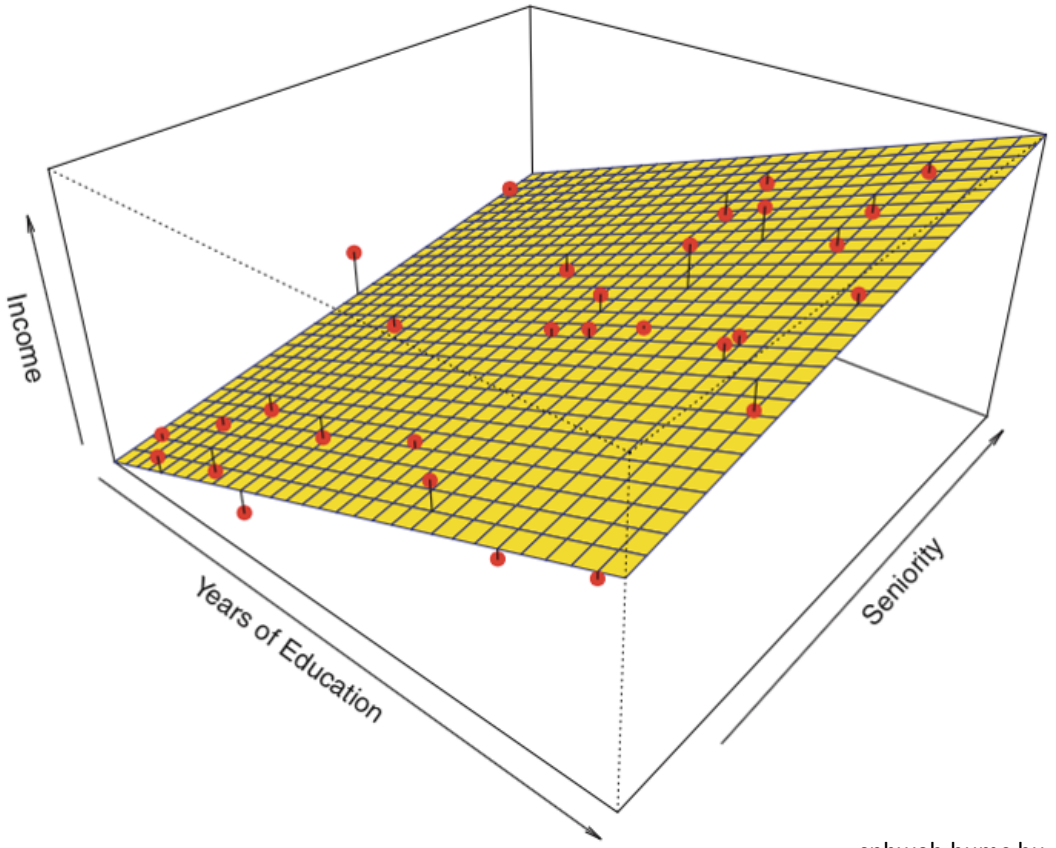
Příklad vícenásobné regrese

- 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:

$$ziskMZ_i = a + b_1 * nezaměstnanost_i + b_2 * věk_i + b_3 * podílVŠ_i + e_i$$

- 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

Vícenásobná regresní analýza v Excelu

1. v Excelu v záložce Data vybereme dlaždici „Analýza dat“
2. zvolit nástroj „Regrese“
3. vybereme rozsah proměnné Y (závisle proměnná)
   - včetně názvu proměnné a vybrat možnost popisků
4. vybereme rozsah proměnných X, které musí být vedle sebe ve sloupcích (nezávisle proměnné)
   - včetně názvu proměnné a vybrat možnost popisků
5. vybereme hladinu spolehlivosti (typicky 95 %)
6. zvolíme možnost zakreslení na nový list (ten pojmenujeme)
7. v nabídce navíc vybereme možnost reziduí, grafu s rezidui a grafu regresní přímky

Vícenásobná regresní analýza v Excelu
Interpretace výsledků analýzy

		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
průnik	Intercept	53,85393656	8,726315603	6,171440389	3,41387E-09	36,65200676	71,05586635
nezávisle proměná X_1	vek	-0,061865919	0,209379207	-0,295473082	0,76792307	-0,474609004	0,350877165
nezávisle proměná X_2	vs_proc	-2,303024162	0,193330849	-11,91234703	2,33433E-25	-2,684131589	-1,921916735
nezávisle proměná X_3	nezamestnanost	0,147848043	0,29191606	0,506474509	0,61305239	-0,427597517	0,723293604

Regression Statistics	
Multiple R	0,654828609
R Square	0,428800507
Adjusted R Square	0,420679187
Standard Error	7,384888659
Observations	215

Srovnání jednoduché a vícenásobné regresní analýzy

Regression Statistics	
Multiple R	0,201316276
R Square	**0,040528243**
Adjusted R Square	0,036023681
Standard Error	9,526153362
Observations	215

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	**34,74457634**	1,43225382	24,25867249	3,24537E-63	31,92136933	37,56778334
nezamestnanost	**1,08097976**	0,360383559	2,999525737	**0,003025991**	0,370604704	1,791354817

jednoduchá regresní analýza

vícenásobná regresní analýza

Regression Statistics	
Multiple R	0,654828609
R Square	**0,428800507**
Adjusted R Square	0,420679187
Standard Error	7,384888659
Observations	215

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	**53,85393656**	8,726315603	6,171440389	3,41387E-09	36,65200676	71,05586635
vek	-0,061865919	0,209379207	-0,295473082	0,76792307	-0,474609004	0,350877165
vs_proc	-2,303024162	0,193330849	-11,91234703	2,33433E-25	-2,684131589	-1,921916735
nezamestnanost	**0,147848043**	0,29191606	0,506474509	**0,61305239**	-0,427597517	0,723293604

Interpretace vícenásobné regresní analýzy

$$\widehat{ziskMZ}_i = 53{,}85 + 0{,}15 * nezaměstnanost_i + (-0{,}06) * věk_i + (-2{,}30) * podílVŠ_i$$

- „při zvýšení nezaměstnanosti o jeden procentní bod a stálosti všech ostatních parametrů (věku a podílu vysokoškoláků) se zisk pro Miloše Zemana v obci Libereckého kraje zvýší o 0,15 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 Libereckého kraje sníží o 0,06 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 Libereckého kraje sníží o 2,30 procentního bodu“
- „pokud je nezaměstnanost nulová, průměrný věk v obci je nulový a nežije zde ani jeden vysokoškolák je zisk Miloše Zemana 53,85 procent“

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

standardní 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

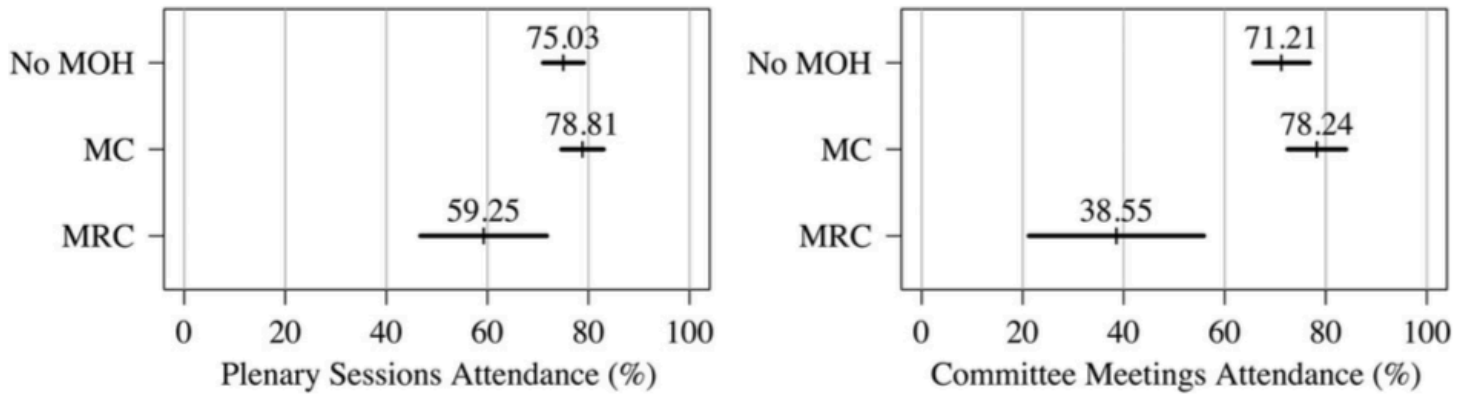


Figure 1. The effect of multiple-office holding on plenary sessions attendance and committee meetings attendance.

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
   - nezávisle proměnná je intervalová nebo poměrová; může být i nominální, ale jen dichotomická
2. multikolinearita
   - nezávisle proměnné by mezi sebou neměly být příliš vysoce korelovány
3. pozor na odlehlé hodnoty!
   - mohou velmi značně ovlivnit podobu regresní přímky
4. normální distribuce reziduí s nulovým průměrem
   - jinými slovy distribuce hodnot proměnných by se měla blížit normálnímu rozložení
5. homoskedasticita
   - rozptyl reziduálních hodnot je podobný na všech místech hodnot závisle proměnné Y
6. nezávislost reziduí
   - souvisí s homoskedasticitou – v chybovosti nesmí existovat vzorec
7. lineární vztah mezi závisle proměnnou a nezávisle proměnnými
   - v jiném případě může být mezi proměnnými vztah a OLS regrese ho neodhalí

- 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!