Regresní analýza v „prostorové“ analýze

Petr Voda

Regrese - připomenutí

- Nástroj k analýze vlivu více nezávisle proměnných na jednu závisle proměnné
  - Jak různé vlastnosti obcí ovlivňují to, kolik procent hlasů v ní kandidát dostane
  - Testování teorie
  - Musíme mít nějaký předpoklad, jaké proměnné, proč a jak by měly ovlivňovat závisle proměnnou
  - Proč = mechanismus
  - Jak = hypotéza

Podmínky použití

- Pro prostorovou analýzu můžeme některé jinak důležité podmínky ignorovat
  - Normalita závisle proměnné
- Některé musíme ignorovat
  - Nezávislost pozorování
- Podmínky
  - Jedna závisle proměnná
  - Předpoklad lineárního vztahu
  - **Nezávislost nezávisle proměnných mezi sebou**
  - **Opakem je multikolinearita**

Specifika v prostorové analýze

- Nezávislost pozorování
  - Často narušeno
  - V blízkých lokalitách často podobné hodnoty
- Normalita závisle proměnné
  - Velmi důležitá zejména pro hodnoty inferenční statistiky
  - V analýze zahrnující všechny případy není nutná taková přísnost
  - Rozdělení by se ale normálnímu mělo alespoň přibližovat
- Multikolinearita
  - Častý problém
- Nestacionarita
  - V různých místech mohou být vztahy mezi proměnnými různé

Teorie

- Teorie konfliktních linií
  - Strany vznikly, protože některé sociální skupiny chtěly být reprezentovány
  - Strany jako odraz sociální struktury
  - Vlastníci x pracující
  - Církev x stát
  - Město x venkov
  - Centrum x periferie
- Politická socializace
  - Prostředí může ovlivňovat předávané hodnoty
  - Např. sudety

Geografické efekty

- Sousedský efekt
- Efekt nákazy
- Efekt sporného bodu

Kontextuální vs kompozitní efekty

- Kompozitní efekt
  - Strana má v místě se silným zastoupením skupiny vysokou podporu, protože členové skupiny stranu volí
  - Např. katolící a KDU-ČSL
- Kontextuální efekt
  - Přítomnost skupiny vytváří příhodný kontext pro to, aby stranu volili jiní lidé
  - Např. efekt nezaměstnanosti na podporu levicových stran
  - Nezaměstnaní se velice často neúčastní voleb
  - Vysoká nezaměstnanost aktivizuje zaměstnance pocitově ohrožené nezaměstnaností

Proměnné

- Výběr proměnných je určen teorií
- Obvykle v procentech – jinak zkoumáte jen efekt velikosti obce
- Správně spočítaná procenta
  - Jinak se měří nezaměstnanost a jinak religiozita
- Proměnné, které měří různé věci
- Proměnné, které mají smysl
  - - teorie
  - nebo nějaká vyargumentovaná úvaha

  - **Formulace hypotéz**

Problematické proměnné

- Volební účast
- Kombinace proměnných měřících tutéž věc:
  - Podíl zš, podíl sš, podíl vš
  - Mladí, střední, starší
  - …
- Velikost obce
  - Volba nějakých funkčních kategorií
    - 1) Velkoměsto (Obce nad 1000 obyvatel)
    - 2) Maloměsto (Obce od 251-1000 obyvatel)

Problém se senátními volbami

- Je kandidát totéž co strana?
  - Korelace s podporou strany
  - Vysoká korelace: ok
  - Nízká korelace: předpoklady o podpoře kandidáta nemohou být totožné jako předpoklady o podpoře strany
    - Sousedský efekt?
    - Koalice stran
    - Představitel nějakého „křídla“ strany?
    - Kampaň?
    - Něco jiného?

Základ: „jednoduchá regrese“

- Závisle proměnná: podpora kandidáta
- Nezávisle proměnné: indikátory konfliktních linií

- Příklad: podpora Czernina v obvodu jičín
- Np:
  - vlastníci/pracující: nezaměstnanost
  - Město/venkov: velikost obce (dummy)
  - Navíc – okres, podíl důchodců

Co regrese dělá

- **Odhad parametrů** přímky (při 1 nezávisle proměnné), roviny (při 2) či nadroviny (při více)
- Parametry: **sklon** (pro každou proměnnou) a ***konstanta*** (jedna pro celý model)
- Parametry popisují vztah mezi nezávisle a závisle proměnnou
- Hodnota závisle proměnné (y) = konstanta (a)+ sklon(b)*hodnota nezávisle proměnné (x)
- $y = a + b*x$
- $y = a + b1*x + b2*x + b3*x + \ldots$

Co nám výpočet poskytne?

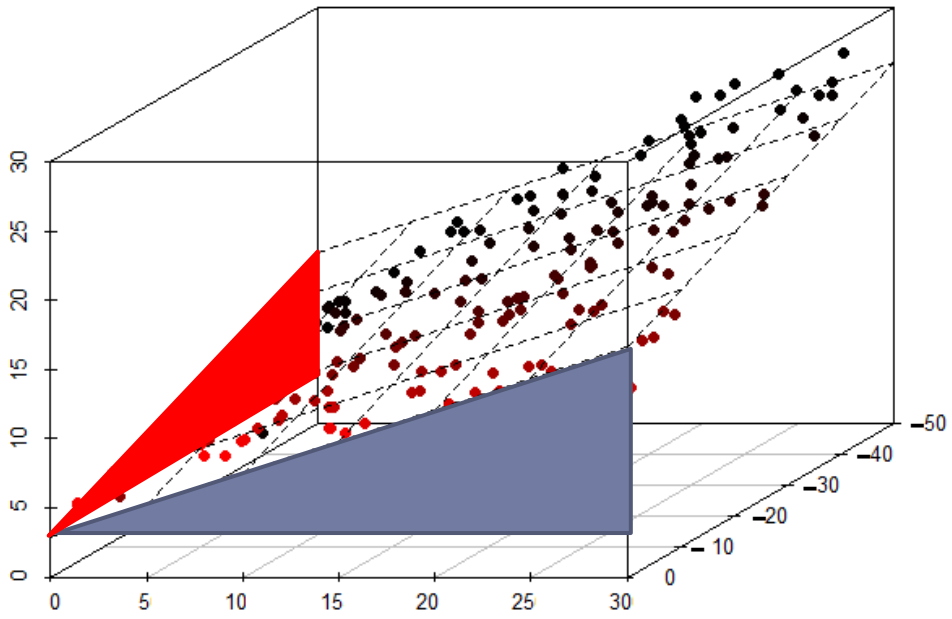
- R-square (česky index determinace)
  - Ukazuje jak dobře model sedí na data
- Parametry
  - Unstandardized beta (nestandardizovaný beta koeficient)
    - Standardizovaný beta koeficient
  - Constant (konstanta)
- Hodnoty signifikance

Co je to R-square?

- Ukazuje, kolik procent rozptylu závisle proměnné je vysvětleno přidáním nezávisle proměnných
- Původní rozptyl je vypočten jako suma kvadratických odchylek mezi průměrem a jednotlivými hodnotami závisle proměnné
- „nový“ rozptyl je vypočten jako suma odchylek od regresní přímky/roviny
- Rozdíl mezi původním a novým rozptylem vydělený původní variabilitou = R-square
- Čím víc proměnných, tím nižší R-square
  - Řešeno pomocí adjusted R-squre

Nestandardizovaný Beta koeficient

- efekt nezávisle proměnné na závisle proměnnou
- „**o kolik se změní hodnota závisle proměnné, pokud se hodnota nezávisle proměnné změní o jednotku**“ pokud vše ostatní zůstává shodné
- Různé proměnné se mohou změnit o různý počet jednotek
  - Pro srovnání síly proměnných v modelu – standardizovaný koeficient beta (jakou změnu v počtu směrodatných odchylek závisle proměnné způsobí změna o směrodatnou odchylku nezávisle proměnné)


0
5
10
15
20
25
30
0
5
10
15
20
25
30
0
−10
−20
−30
−40
−50

Interpretace efektu dummy proměnné

- Podpora Kandidáta je v obcích nad 1500 obyvatel o 0,3procentního bodu nižší než v obcích do 1500 obyvatel
- Nebo též
- Pokud je vše ostatní shodné, pak rozdíl v podpoře kandidáta mezi vesnicí (obce do 1500 obyvatel) a maloměsty/městy (obce na d 1500 obyvatel) je 0,3 procentního bodu. V menších obcích je podpora vyšší.

- **Kategorie K se liší o +- XX od referenční kategorie** (pokud je vše ostatní shodné)
- Interpretace musí obsahovat:
  - Identifikaci kategorie proměnné, ke které je koeficient vztažen
  - Identifikaci referenční kategorie
  - Informaci o velikosti rozdílu

Interpretace efektu kardinální proměnné

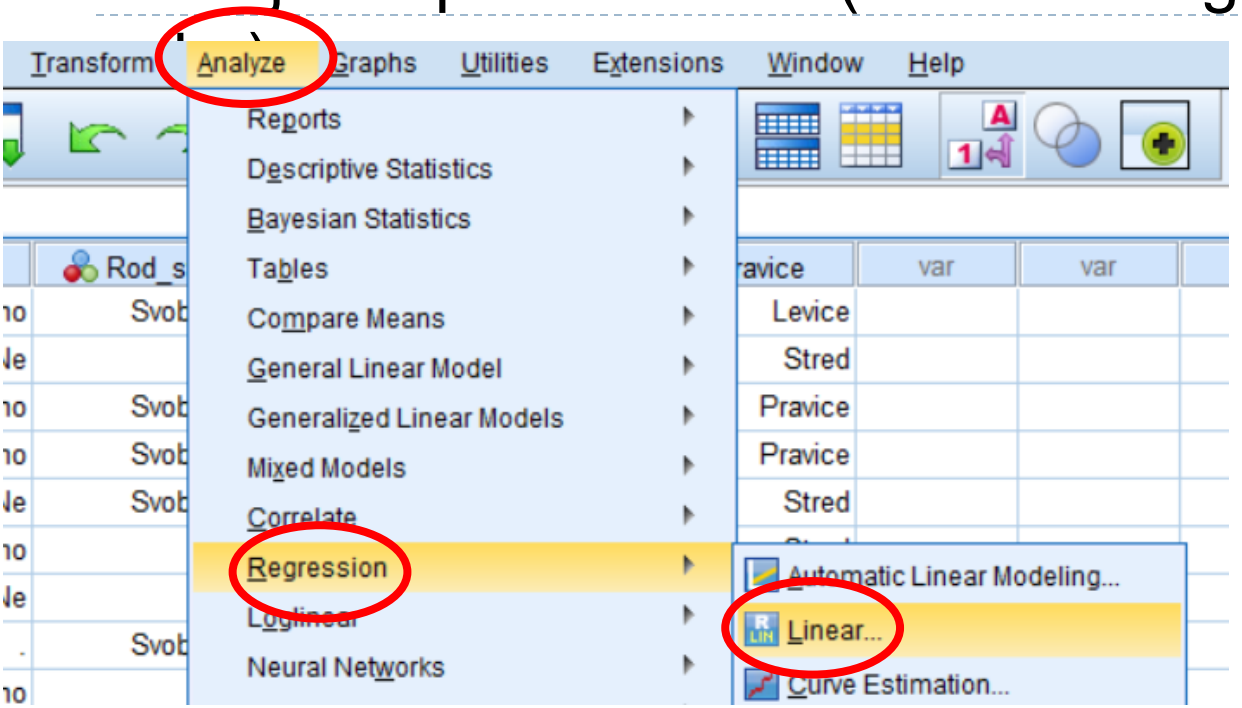
- Pokud je v obci A o 1 pb vyšší podíl vš obyvatelstva než v obci B a vše ostatní je shodné, pak v obci A je podpora kandidáta o 0,5 pb vyšší
- Nebo též
- S růstem podílu obyvatel s VŠ vzděláním o 1 pb (pokud vše ostatní zůstává shodné) podpora kandidáta roste o 0,5 pb
- Lze násobit
  - Pokud je podíl VŠ obyvatelstva vyšší o 2 pb, pak je zisk kandidáta vyšší o 1 pb

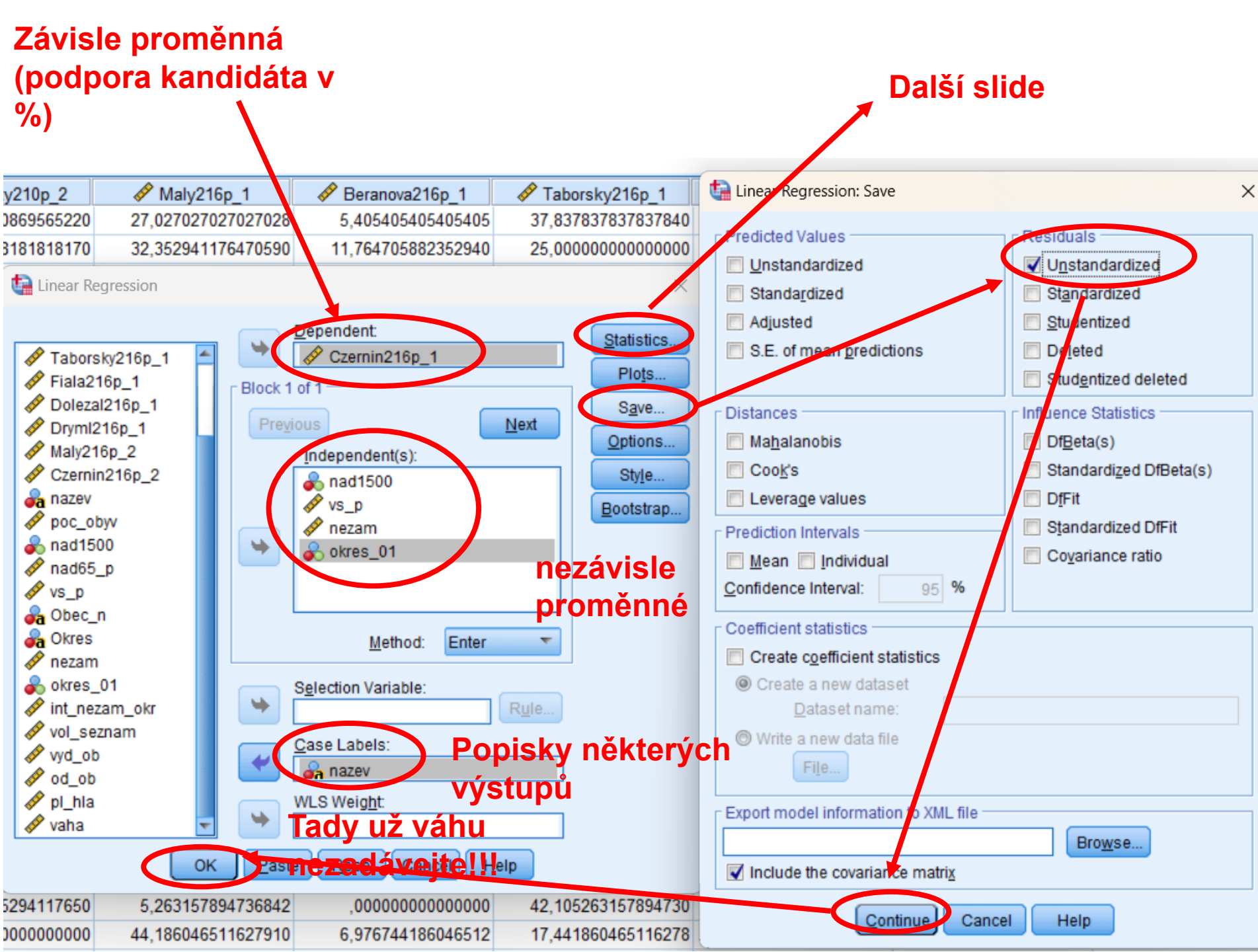
Rezidua

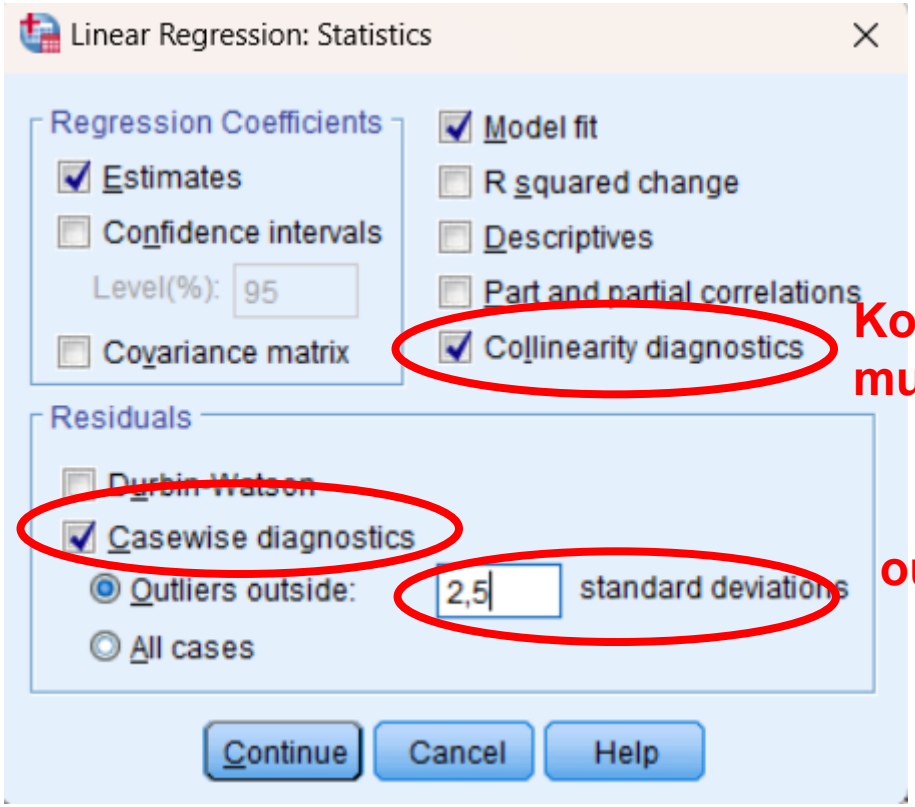
- Jak se skutečná podpora liší oproti očekávané
- Kde se kandidátovi dařilo nad poměry a kde pohořel, i když by neměl?
  - Očekávaná hodnota: dána kombinací konstanty, hodnot koeficientů a hodnot příslušných proměnných v obci
- Nestandardizovaná: ukazují rozdíl v hodnotách závisle proměnné (v našem případě procentní body)
- Standardizovaná: statistický rozdíl – umožňují zhodnotit relativní význam odchylky
- **Možnost zobrazit v mapě**

Jak naklikat v spss

- Mějte zapnutou váhu (data -> weight cases-> weight




Závisle proměnná (podpora kandidáta v %)
Další slide
Linear Regression
Dependent:
Czernin216p_1
Block 1 of 1
Previous
Next
Independent(s):
nad1500
vs_p
nezam
okres_01
Method: Enter
Selection Variable:
Rule...
Case Labels:
nazev
WLS Weight:
Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Taborsky216p_1
Fiala216p_1
Dolezal216p_1
Dryml216p_1
Maly216p_2
Czernin216p_2
nazev
poc_obyv
nad1500
nad65_p
vs_p
Obec_n
Okres
nezam
okres_01
int_nezam_okr
vol_seznam
vyd_ob
od_ob
pl_hla
vaha
nezávisle proměnné
Popisky některých výstupů
Tady už váhu nezadávejte!!!
Linear Regression: Save
Predicted Values
Unstandardized
Standardized
Adjusted
S.E. of mean predictions
Residuals
Unstandardized
Standardized
Studentized
Deleted
Studentized deleted
Distances
Mahalanobis
Cook's
Leverage values
Influence Statistics
DfBeta(s)
Standardized DfBeta(s)
DfFit
Standardized DfFit
Covariance ratio
Prediction Intervals
Mean
Individual
Confidence Interval: 95 %
Coefficient statistics
Create coefficient statistics
Create a new dataset
Dataset name:
Write a new data file
File...
Export model information to XML file
Browse...
Include the covariance matrix
Continue
Cancel
Help



Kontrola multikolinearity

outlieři

R2

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,797[a]	,635	,626	7,008109456

a. Predictors: (Constant), okres_01, nezam, vs_p, nad1500

b. Dependent Variable: Czernin216p_1

- model má/nemá vysvětlovací sílu
- 63,5 % případů tento model vysvětluje
- Index determinace vysvětluje „jen“ 63,5 %.
- 63 % variability výsledků kandidáta je vysvětleno použitými nezávisle proměnnými

Coefficients[a]

Double-click to activate

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	33,024	2,548		12,961	,000		
	nad1500	-,152	2,169	-,005	-,070	,944	,511	1,955
	vs_p	,514	,186	,165	2,759	,006	,59	1,671
	nezam	-,508	,347	-,089	-1,466	,144	,58	1,725
	okres_01	-17,214	1,308	-,749	-13,157	,000	,665	1,505

a. Dependent Variable: Czernin216p_1

- Nejsilnější je efekt proměnné okres (standard. Beta je největší číslo)
- V obcích okresu Jičín je podpora kandidáta o téměř 20 pb vyšší než v obcích okresu Nymburk, pokud vše ostatní zůstává shodné
- S každým procentním bodem podílu vysokoškolsky vzdělaného obyvatelstva roste podpora kandidáta o 0,5 pb
- Nemáme problém s multikolinearitou

Nestandardizované koeficienty

- Dummy
  - Nejvíce pozitivní vazbu tak můžeme vidět u výsledku kandidátky s obyvateli z vesnic pod 1000 obyvatel, kdy s každým obyvatelem takovéto vesnice vzrostl elektorát Bochníčkové o 0,07 pr. Bodu
  - Kandidáta v okolí domácí obce podpořilo o 3 p.b. více voličů (chybí „než v ostatních obcích“)

Efekty kardinálních proměnných

- nedá se tedy jednoznačně potvrdit, že obce s vyšším podílem seniorů výrazně více volí ANO
- je patrné, že u zaměstnanců je volba kandidáta ODS méně pravděpodobná
- S každým procentem podnikatelů v populaci klesala pravděpodobnost volby kandidáta
- Na zisky Staňka působí nejsilněji výskyt lidí v důchodovém věku a mladí lidé
- proměnné podíl nezaměstnaných osob má na kandidáta X pozitivní efekt
- Kladný vztah měly výsledky kandidáta také k zemědělcům
- vyšší podíl vysokoškolsky vzdělaných v obci kandidátovi škodil

Příklad problému s multikolinearitou

	koeficient		koeficient	kolinearita
	B	Std. Error	Beta	VIF
Konstanta	-5,309	12,403		
volební účast_p	0,072	0,113	0,059	2,210
nezaměstnanost_říjen16p	1,842	0,372	0,443	2,067
mládež_p	0,400	0,345	0,137	3,596
dospělý_p	0,461	0,215	0,217	2,646
důchodce_p	0,125	0,277	0,049	2,992
základní_p	-0,158	0,313	-0,104	11,020
Soused_CK	1,020	3,088	0,021	1,031
střední_p	-0,112	0,251	-0,070	6,422
vysoké_p	-0,746	0,316	-0,403	7,577

Outlieři

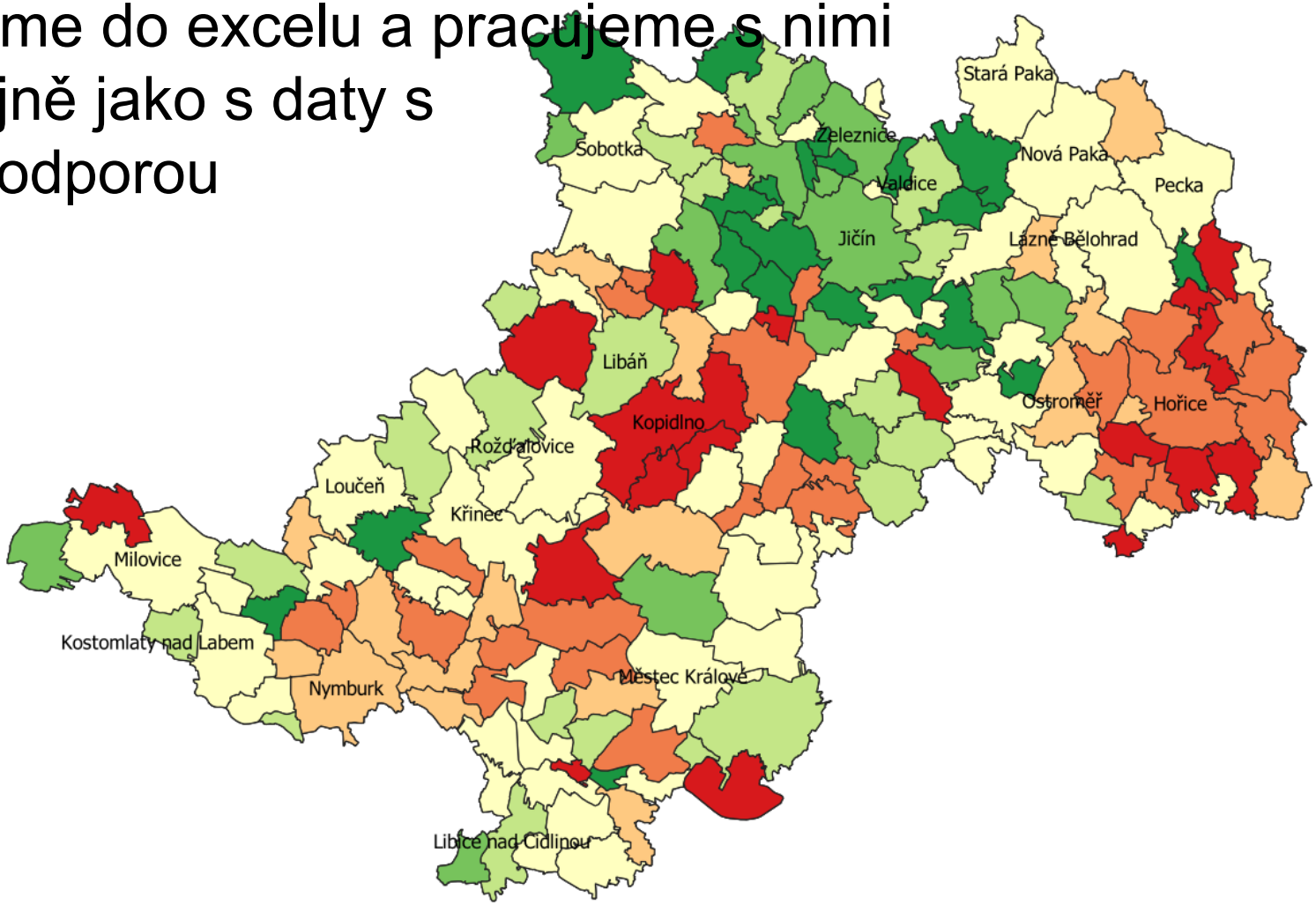
Casewise Diagnostics[a]

Case Number	nazev	Std. Residual	Czernin216p_1	Predicted Value	Residual
8	Sedliště	2,995	40,00000000	19,01002131	20,98997869
25	Kostelec	2,556	36,84210526	18,92744453	17,91466073
42	Borek	5,557	53,33333333	14,38636159	38,94697174
100	Choteč	2,593	36,36363636	18,19061709	18,17301927
104	Soběraz	2,535	37,50000000	19,73421275	17,76578725
107	Březina	2,764	41,86046512	22,49026457	19,37020054
108	Zámostí-Blata	6,740	68,75000000	21,51411023	47,23588977
127	Netřebice	2,710	50,87719298	31,88429415	18,99289884
138	Chotěšice	3,166	54,87804878	32,69229555	22,18575323
140	Kouty	2,810	56,57894737	36,88556551	19,69338186
147	Činěves	3,412	56,92307692	33,01456905	23,90850787
149	Vlkov pod Oškobrhem	4,213	62,50000000	32,97260074	29,52739926
157	Čilec	-2,996	10,93750000	31,93396552	-20,9964655
166	Dymokury	6,857	81,53310105	33,47996151	48,05313954
174	Velenice	2,619	52,38095238	34,02786847	18,35308391

a. Dependent Variable: Czernin216p_1

Rezidua v mapě

- Pokud jsme zatrhli v save rezidua, tak si data exportujeme do excelu a pracujeme s nimi úplně stejně jako s daty s volební podporou



Prezentace výsledků

- Postačuje R2, B a Beta (+ konstanta)
- Signifikance není nutná, nikam se nezobecňuje
- Stejně tak není nutné ukazovat multikolinearitu v každé tabulce, pokud používáte stále stejné nezávisle proměnné
- Pro prezentaci výsledku je vhodné přejmenovat proměnné
- Desetinná místa
- Přeložení do češtiny

	B	Beta	VIF
konstanta	33.62		
nezaměstanost	-0.51	-0.09	1.48
vš	0.50	0.16	1.56
nad 65	-0.03	-0.01	1.14
okres jičín	-17.13	-0.75	1.25
R	0.60		

Přidání interakce

- Interakce = proměnná x proměnná
- Jak se mění **EFEKT** jedné proměnné při změně hodnoty druhé proměnné o jednotku
- Např. efekt nezaměstnanosti je větší na periferii než v centru
  - V našem případě v okrese Nymburk – paří do středočeského kraje, zatímco Jičín do Královehradeckého kraje, kvůli většímu navázání na Prahu můžeme očekávat větší citlivost i na menší rozdíly v nezaměstnanosti

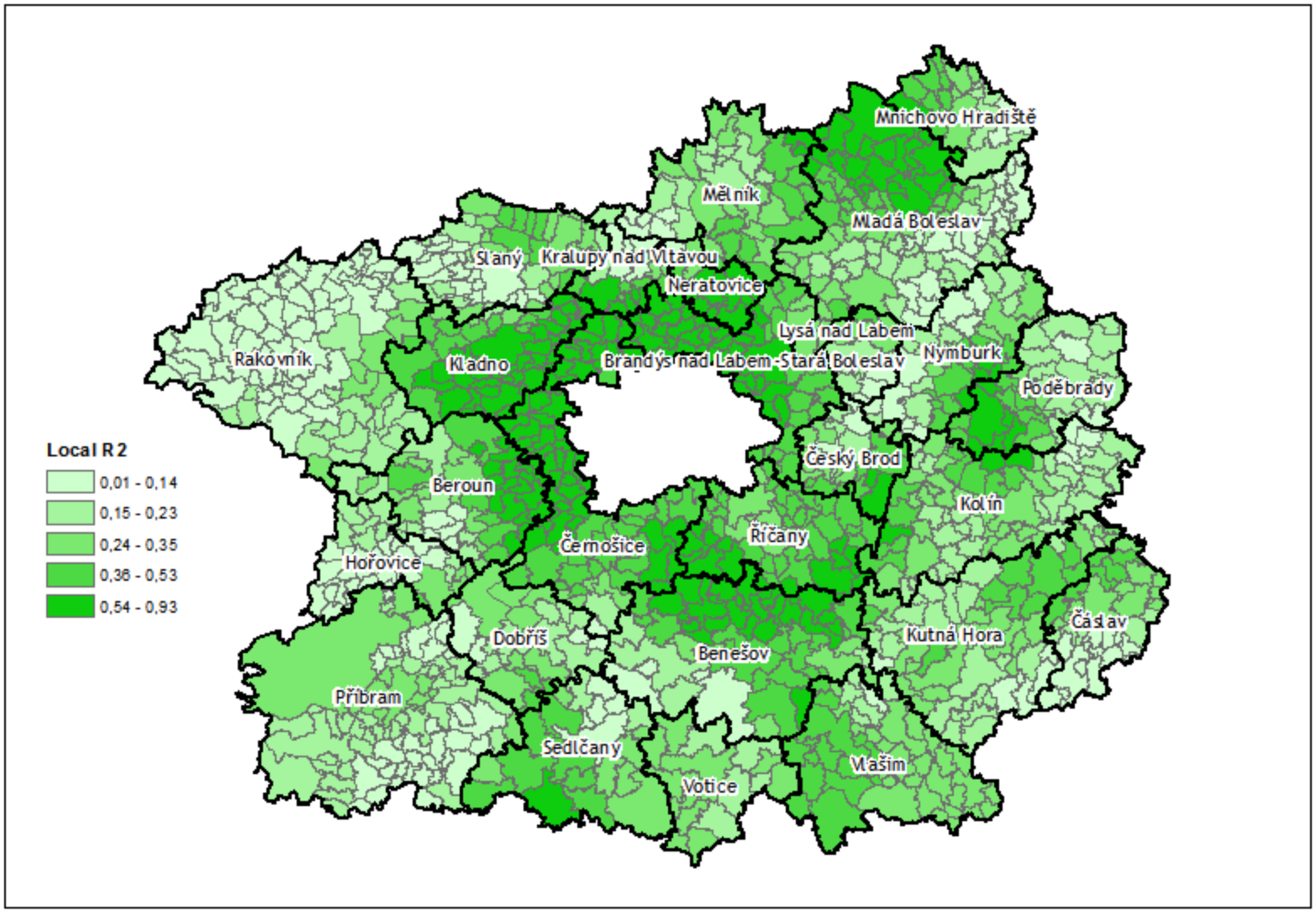
Model		Unstandardized Coefficients		Coefficients
		B	Std. Error	Beta
1	(Constant)	32,733	2,681	
	nad1500	,203	2,390	,006
	vs_p	,498	,193	,160
	nezam	-,432	,408	-,076
	okres_01	-16,254	2,991	-,707
	int_nezam_okr	-,250	,699	-,042

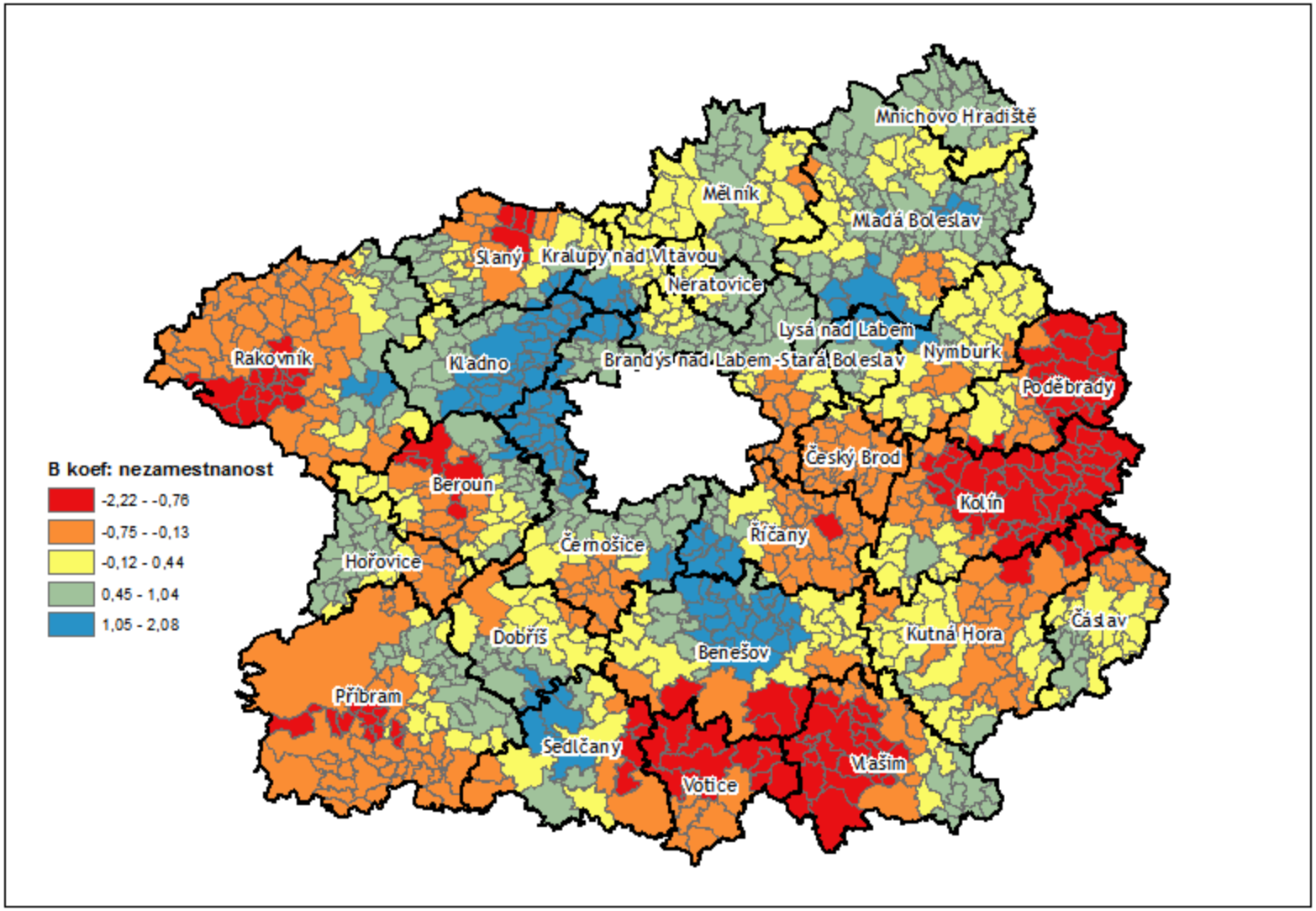
a. Dependent Variable: Czernin216p_1

- V okrese 1 (Nymburk) je efekt nezaměstanosti o 0,25 silnější než v okrese Jičín
- V okrese Jičín s každým pb nezaměstnanosti klesá podpora kandidáta o 0,5 pb, v okrese Nymburk klesá o 0,7 pb (-,432 + -,25 = -,682)

Další možnosti

- Prostorově vážená regrese
  - Přidává informaci o nestacionaritě vztahů
  - Spíše explorativní charakter
  - Často obtižné najít ve výsledcích nějaký smysl
- Víceúrovňové modelování
  - Závisle proměnnou ovlivňují proměnné z různých úrovní
  - Volební chování jedince je ovlivněno jeho vlasnostmi a vlastnostmi prostředí
  - Různé vlastnosti voliče v různém prostředí vedou k různým volbám
  - Obvyklý problém: nedostatek dat


Local R2
0,01 - 0,14
0,15 - 0,23
0,24 - 0,35
0,36 - 0,53
0,54 - 0,93
Mnichovo Hradiště
Mělník
Mladá Boleslav
Slaný
Kralupy nad Vltavou
Neratovice
Lysá nad Labem
Nymburk
Rakovník
Kladno
Brandýs nad Labem-Stará Boleslav
Poděbrady
Český Brod
Beroun
Kolín
Říčany
Černošice
Hořovice
Čáslav
Kutná Hora
Dobříš
Benešov
Příbram
Sedlčany
Vlašim
Votice


B koef: nezamestnanost
-2,22 - -0,76
-0,75 - -0,13
-0,12 - 0,44
0,45 - 1,04
1,05 - 2,08
Mnichovo Hradiště
Mělník
Mladá Boleslav
Slaný
Kralupy nad Vltavou
Neratovice
Lysá nad Labem
Nymburk
Rakovník
Kladno
Brandýs nad Labem-Stará Boleslav
Poděbrady
Český Brod
Beroun
Kolín
Černošice
Říčany
Hořovice
Dobříš
Benešov
Kutná Hora
Čáslav
Příbram
Sedlčany
Vlašim
Votice

Data za městské části

- Jen pro členěná statutární města
  - Praha, Brno, Ostrava, Plzeň, Ústí nad Labem, Pardubice, Liberec, Opava
- https://www.czso.cz/csu/czso/vysledky-scitani-2021-otevrena-data
- https://www.czso.cz/csu/czso/otevrena_data_pro_vysledky_scitani_lidu_domu_a_bytu_2011_sldb_2011
- Otevřená data volby.cz obsahují v položce obec rovnou data za městské části
  - V případě „běžných“ měst potřeba pracovat s volebními okrsky

SLDB 2011

- Seznam území
- Vyfiltrovnání městských částí
- Kód + sloupec městská část
- Připojení k datům

- Pro „běžná“ města lze vytvořit z dat na úrovni ZSJ
  - Zdá se že již/ještě nejsou veřejně dostupná
  - Možné získat na žádost

SLDB 2021

- Data v csv
- vložení dat do Excelu
  - Data -> Načíst externí data -> Z textu a nastavit správný oddělovač (čárka) a typ souboru (UTF-8)
- Městské části filtr -> cis=44