VYSOKÁ ŠKOLA BÁŇSKÁ – TECHNICKÁ UNIVERZITA OSTRAVA

Hornicko – geologická fakulta

Katedra geoinformatiky

PROSTOROVÁ ANALÝZA DAT

Regresní analýza

Lucie Orlíková

Ostrava, 2020

ÚVOD DO REGRESNÍ ANALÝZY

Prostorové regresní modelování představuje slibné rozšíření klasických regresních modelů s využitím prostorové struktury dat. Snaží se tím adresovat jeden z hlavních nedostatků klasických regresních modelů pracujících s prostorovými daty, tedy jejich nedostatek nezávislosti mezi pozorováními v blízkých místech, což způsobuje porušení jednoho ze základních principů řešení regresních rovnic. Současná nabídka programových implementací prostorového regresního modelování je poměrně široká a stále více uživatelů ji využívá pro své analýzy. Ne všichni si ale uvědomují vhodnost některých postupů a jejich výsledky tím mohou být nepříjemně poznamenány.

Postup přípravy regresního modelu

Na začátku přípravy každého regresního modelu je potřebné si dobře rozmyslet, co je závisle proměnná Y, jaké jsou nezávisle proměnné X_1 až X_n, resp. vysvětlující proměnné. To nemusí být tak jednoduché, jak vypadá, viz další text k provedení EDA, kdy se zvažuje, které proměnné a v jaké podobě se použijí v regresním vztahu. Klíčové je také rozhodnutí, jaký typ modelu připravujeme. Zjednodušeně lze říci, že rozlišujeme:

- Exploratorní (vysvětlující) s pevně danými proměnnými, u nichž nás zajímají jejich vztahy a vliv na Y
- Prediktivní (předpovědní) kde je cílem je co nejlépe vypočítat (odhadnout) Y ze sady proměnných.

Na začátku každého modelování se používá průzkumová analýza dat (Exploratory Data Analysis EDA), jejímž cílem je poznat vlastnosti datových sad. U prostorových dat se pak více mluví o ESDA (Exploratory Spatial Data Analysis), která zkoumá i prostorové vlastnosti dat.

EDA zahrnuje analýzu distribuce každé proměnné, zejména ocenění její asymetrie a provedení vzájemné korelační a regresní analýzy všech proměnných. Cílem EDA je odhalit problémy proměnných s heteroskedascitou a odstranit multikolinearitu vznikající díky úzkým korelacím nezávisle proměnných. Je vhodné také počítat sekundární proměnné, jako jsou poměry (míry, kvocienty, indexy) a kvůli problému MAUP také hustoty (přepočty na plochu).

Pro odstranění asymetrie je zejména u predikčních modelů potřebné provést vhodnou transformaci, která zlepší symetrii distribuce proměnné, jinak mohou být vychýlené výsledné odhady závisle proměnné. U vysvětlujících modelů se ale po transformaci můžeme potýkat s vhodnou interpretací, protože daleko snáze se interpretují původní veličiny (zejména s využitím jejich standardizovaných regresních koeficientů). Pro predikční modely, zejména u proměnných s velmi odlišnými rozsahy hodnot, je zpravidla potřebné provést standardizaci proměnných, nejběžnější je Z-standardizace. Opět platí, že standardizace v případě vysvětlujících modelů, může komplikovat interpretaci vlivu jednotlivých nezávisle proměnných.

Postup řešení v Excelu a SPSS

Nejdříve začneme bivariační lineární regresí, tedy variabilitu závislé proměnné budeme vysvětlovat pouze jednou nezávislou proměnnou.

Příklad č.1

V rámci řešeného cvičení budeme pracovat s daty z voleb v roce 2017 na úrovni okresů a jako závislou proměnnou zvolíme procento hlasů pro KDU-ČSL a jako nezávislou proměnnou zvolíme podíl věřících.

Dílčí kroky řešení:

1. V prvním kroku testovat normalitu proměnných a případně navrhnout normalizaci – zjistíme, že nemají normální rozdělení a bylo by vhodné data normalizovat (v příkladu níže jsou výsledky bez normalizace, abyste mohli výsledky porovnat)
2. Vlastní lineární regrese se v SPSS spouští v nabídce Analyze/Regression/Linear
3. Vyberte závislou a nezávislou proměnnou a přesuňte je do patřičných kolonek
4. V nabídce Save je možné vybrat možnost pro uložení celou řadu vypočtených ukazatelů, mimo jiné predikovanou hodnotu, rezidua apod.
5. V nabídce Plots je možné vybrat si např. histogram standardizovaných reziduí
6. Výstupem regresní analýzy jsou tři tabulky. První sumarizuje kvalitu modelu, kde nejdůležitější je koeficient korelace a koeficient determinace, kdy jsme vysvětlili 88,9 % variability závislé proměnné. Také směrodatná chyba odhadu je velice nízká vzhledem k velikosti závislé proměnné. V druhé tabulce jsou výsledky ANOVA pro zhodnocení přispění vysvětlení variability nezávislou proměnnou (v případě většího počtu proměnných je to výhodné). V poslední tabulce jsou uvedeny koeficienty pro regresní rovnici. Je vidět, že míra věřících je statisticky významná a napomáhá výrazně k vysvětlení variability.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,943[a]	,889	,888	,997880518

a. Predictors: (Constant), mira_veric

b. Dependent Variable: kdu_csl

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	598,165	1	598,165	600,709	,000[b]
	Residual	74,682	75	,996		
	Total	672,848	76			

a. Dependent Variable: kdu_csl

b. Predictors: (Constant), mira_veric

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-,858	,238		-3,604	,001
	mira_veric	,383	,016	,943	24,509	,000

a. Dependent Variable: kdu_csl

V Excelu se regrese spouští opět z nabídky Analýza dat/Regrese. Do formuláře je potřeba definovat oblast, kde se vyskytují data pro závislou a nezávislou proměnnou a dále je možnost vybrat, co všechno bude součástí výsledku a to hlavně grafické zobrazení reziduí. Excel má implementovány také funkce pro výpočet sklonu regresní přímky – SLOPE(data závislé proměnné, data nezávislé proměnné) a průsečík s osou y – INTERCEPT(data závislé proměnné, data nezávislé proměnné).

Zadání – Z datové matice si vyberte vlastní závislou proměnnou a pokuste se vysvětlit její variabilitu vámi vybranou nezávislou proměnnou s využití lineární regrese. Využijte SPSS nebo Excel a řiďte se postupem uvedeným u řešeného cvičení.

Častěji však vysvětlujeme závislou proměnnou větším počtem nezávislých proměnných a využívá se tak vícenásobná regrese. Největším problémem v případě vícenásobné regrese je volba sady nezávislých proměnných. Na přednášce jsme si představili čtyři základní strategie. V dalším řešeném příkladu si ukážeme jak teoreticky postupovat/nepostupovat.

Příklad č.2

Budeme pokračovat s tematikou voleb a budeme vysvětlovat výsledky jiné strany, tentokráte si vyberete Piráty (opět výsledky voleb do poslanecké sněmovny z roku 2017) a hodnotit budeme procentuální podíl hlasů na úrovni okresů. Jako nezávislé proměnné máme možnost pracovat se 7 proměnnými podílem vysokoškoláků, mírou nezaměstnanosti, podílem obyvatel nad 64 let, podílem věřících, mírou podnikatelů, podíl rodáků a podíl mužů.

Jak vybrat nezávislé proměnné? Jednou ze strategií je tzv. kitchen-sink přístup, tedy vybrat všechny a uvidíme, jak to dopadne. Tento přístup NENÍ vhodný, protože samozřejmě s každou další proměnnou bude narůstat koeficient determinace, ale je důležité si uvědomit, že není cílem ho maximalizovat. Nicméně podívejme se na výsledky. Podařilo se vysvětlit 22 % variability. A vzhledem k tomu, že průměrný podíl hlasů za okres je 0,8 % hlasů, tak směrodatná chyba je docela vysoká. Ze všech 7 nezávislých proměnných je statisticky významná jen jedna – podíl vysokoškoláků se záporným koeficientem a blízko hranice je také podíl věřících. Obecně je vhodnější strategie vymazat proměnné, které nejsou významné nebo přidávat jen ty proměnné, které významně zvýší hodnotu koeficientu determinace a výrazně tak zvýší vysvětlenou variabilitu.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,469[a]	,220	,140	,155338623

a. Predictors: (Constant), podil_muzu, mira_nezam, mira_veric, podil65vic, mira_podni, podil_vs, podil_roda

b. Dependent Variable: pirati

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,468	7	,067	2,774	,013[b]
	Residual	1,665	69	,024		
	Total	2,133	76			

a. Dependent Variable: pirati

b. Predictors: (Constant), podil_muzu, mira_nezam, mira_veric, podil65vic, mira_podni, podil_vs, podil_roda

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3,088	3,443		,897	,373
	mira_nezam	-,006	,011	-,085	-,526	,601
	mira_podni	,018	,014	,230	1,337	,186
	mira_veric	-,007	,004	-,300	-1,692	,095
	podil65vic	-,002	,022	-,012	-,079	,938
	podil_vs	-,027	,010	-,495	-2,700	,009
	podil_roda	,001	,008	,043	,186	,853
	podil_muzu	-,043	,063	-,113	-,678	,500

a. Dependent Variable: pirati

Tím, že se rovnou vrháme na konstrukci modelu, tak můžeme často přehlédnout nedostatky v datech, jako jsou např. chybějící hodnoty, se kterými se musíme vypořádat. Dalším problémem jsou odlehlé hodnoty a multikolinearita. Odlehlé hodnoty ovlivňují kvalitu modelu a pokud je z-skóre této proměnné vyšší než 3 nebo menší než -3, tak se tento záznam většinou vynechává z modelu. Dalším možným přístupem je využití tzv. leverage hodnot. Pokud jsou větší než 2p/n, kde p je počet nezávislých proměnných, tak by měla hodnota být považována jako odlehlá. Tuto proměnnou je možné vypočítat v nabídce Save a zaškrtnutím Leverage values. Hraniční hodnota pro náš případ je 2*7/77 = 0,1818. Tímto jsme identifikovali 7 odlehlých hodnot a 5 z nich je výrazně nad touto hranicí – Praha-západ, Praha, Jeseník, Brno-město, Praha-východ, Karviná a Plzeň-jih. Multikolinearita je často problém a jako hraniční se považuje korelační koeficient, který převýší nevysvětlenou variabilitu nezávislé proměnné (1 – r2). Podobným ukazatelem je také variance inflation index (VIF) a pokud přesáhne hodnotu

5, tak je významný problém s multikolinearitou. Tento ukazatel je možné přidat do výstupu v nabídce Statistics zaškrtnutím Collienarity Diagnostics.

V dalším modelu vymažeme proměnné Praha-západ, protože má velmi vysokou leverage hodnotu (0,41). Neměli bychom bezhlavě mazat odlehlé hodnoty, ale jen tehdy, pokud je to opodstatněné. Odstranili jsme nezávislou proměnnou podíl rodáků, protože vykazuje vysokou multikolinearitu a také podíl obyvatel na 65 let, protože významnost obou proměnných je velice nízká. Výsledky jsou shrnuty v tabulkách níže.

Koeficient determinace nepatrně poklesl, směrodatná chyba klesla a přibyla další statisticky významná nezávislá proměnná a tou je podíl věřících. Problémy s multikolinearitou již nejsou patrné.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,464[a]	,216	,160	,153889404

a. Predictors: (Constant), podil_muzu, mira_nezam, mira_veric, mira_podni, podil_vs

b. Dependent Variable: pirati

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,456	5	,091	3,847	,004[b]
	Residual	1,658	70	,024		
	Total	2,113	75			

a. Dependent Variable: pirati

b. Predictors: (Constant), podil_muzu, mira_nezam, mira_veric, mira_podni, podil_vs

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2,793	2,945		,948	,346		
	mira_nezam	-,005	,010	-,067	-,473	,638	,551	1,814
	mira_podni	,020	,013	,221	1,505	,137	,518	1,931
	mira_veric	-,006	,003	-,279	-2,486	,015	,892	1,121
	podil_vs	-,025	,010	-,431	-2,575	,012	,400	2,502
	podil_muzu	-,037	,057	-,098	-,655	,515	,501	1,995

a. Dependent Variable: pirati

Pro vytvoření finálního modelu je vhodné odstranit všechny statisticky nevýznamné nezávislé proměnné a výsledný model vysvětluje 17,2 % variability závislé proměnné. Regresní rovnice je: pirati = 1,061 – 0,007(mira vericich) – 0,12(podil vs).

V SPSS je možné otestovat řadu **nelineárních regresních modelů** a to funkcí Regression/Curve Estimation. Je možné vybrat z celé řady různých nelineárních vztahů. V našem případě však vychází zdaleka nejlépe právě lineární vztah.

Program ArcMap má implenetován nástroj **Explanatory Regression**, který je zástupce brutal force přístupu, kdy vyzkouší všechny kombinace všech nezávislých proměnných a nabídne nejlepší řešení pro jednu až všechny proměnné.