Regresní analýza

jednoduchá lineární regrese
mnohonásobná lineární regrese
logistická regrese

Regresní analýza

- korelační koeficient říká, že mezi dvěma proměnnými existuje souvislost - jsme schopni vyslovit určitou předpověď, predikci
- Např. pohlaví – příjem: ale nejsme schopni vyvodit, o kolik více muži vydělávají více než ženy --- nutná regresní analýza
- Jednoduchá lineární regrese, podobně jako bivariační korelační analýza, zkoumá vztah mezi dvěma proměnnými.
- Na rozdíl od korelace však dokáže nejenom popsat těsnost mezi dvěma proměnnými, ale dokáže také říci, jak velký vliv má nezávisle proměnná X na proměnnou závislou Y, a jakou konkrétní hodnotu bude mít závisle proměnná Y, když budeme vědět, jakou hodnotu má proměnná X – dokáže tedy z hodnot nezávisle proměnné predikovat hodnoty závisle proměnné.

Podmínky pro užití regresní analýzy

- (1) Vztah mezi analyzovanými proměnnými musí být lineární,
- (2) závisle proměnná Y je měřena na intervalové úrovni a nezávisle proměnná X je buď intervalová, nebo dichotomická,
- (3) obě proměnné by měly být přibližně normálně rozloženy – při dostatečně velkém souboru (např. N > 100) se však nemusíme tímto předpokladem příliš trápit, neboť díky centrální limitní větě platí, že v takové situaci nenormální rozložení nemá na výsledky velký účinek.

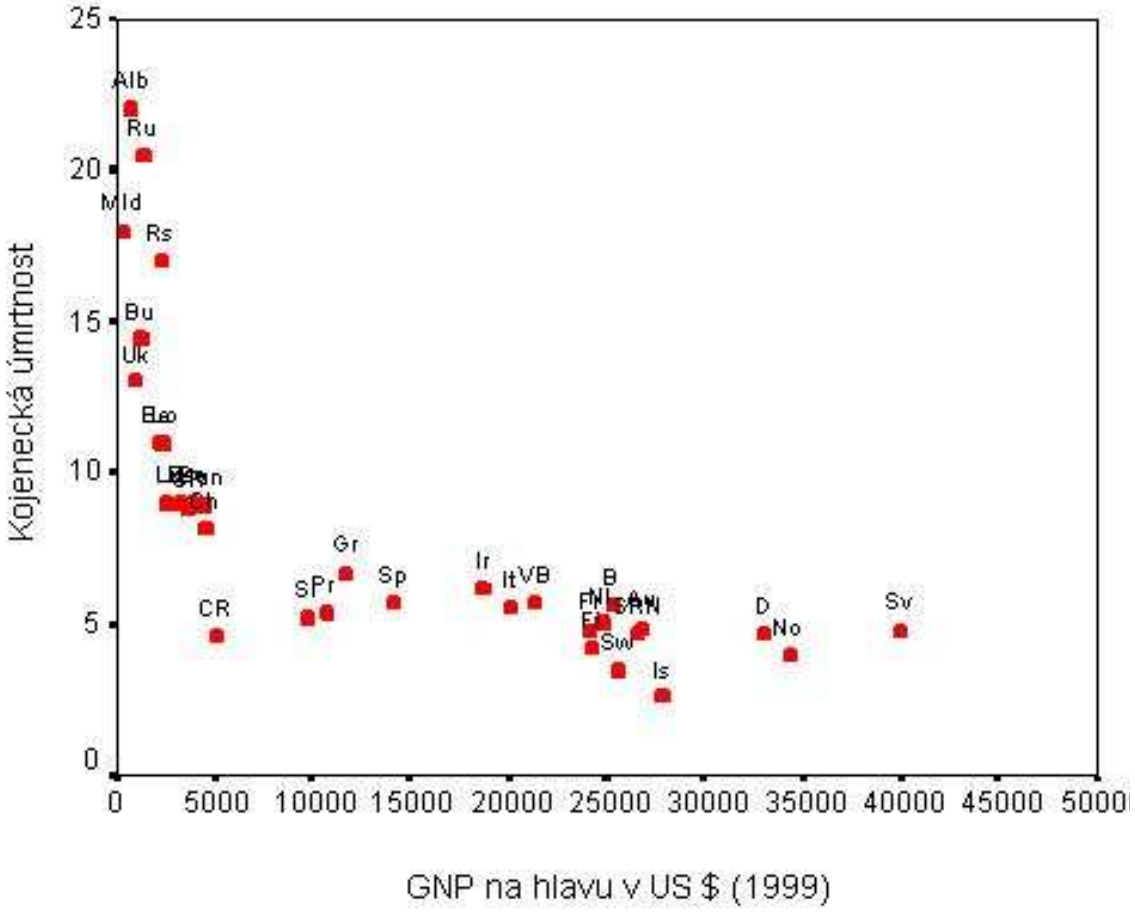
- Základním smyslem jednoduché lineární regrese je sumarizovat vztah mezi dvěma proměnnými tím způsobem, že se určí přímka, která nejlépe vystihuje průběh vztahu. Jakmile je tato přímka stanovena, mohou se vypočítat její parametry, to je může se stanovit rovnice této přímky:

$$y = a + bx$$

- kde *y* je hodnota závisle proměnné, *x* je hodnota nezávisle proměnné, *a* je parametr, který říká, v jakém bodě přímka protíná vertikální osu Y, *b* je hodnota, která určuje směr přímky a v regresní analýze se jí říká regresní koeficient.

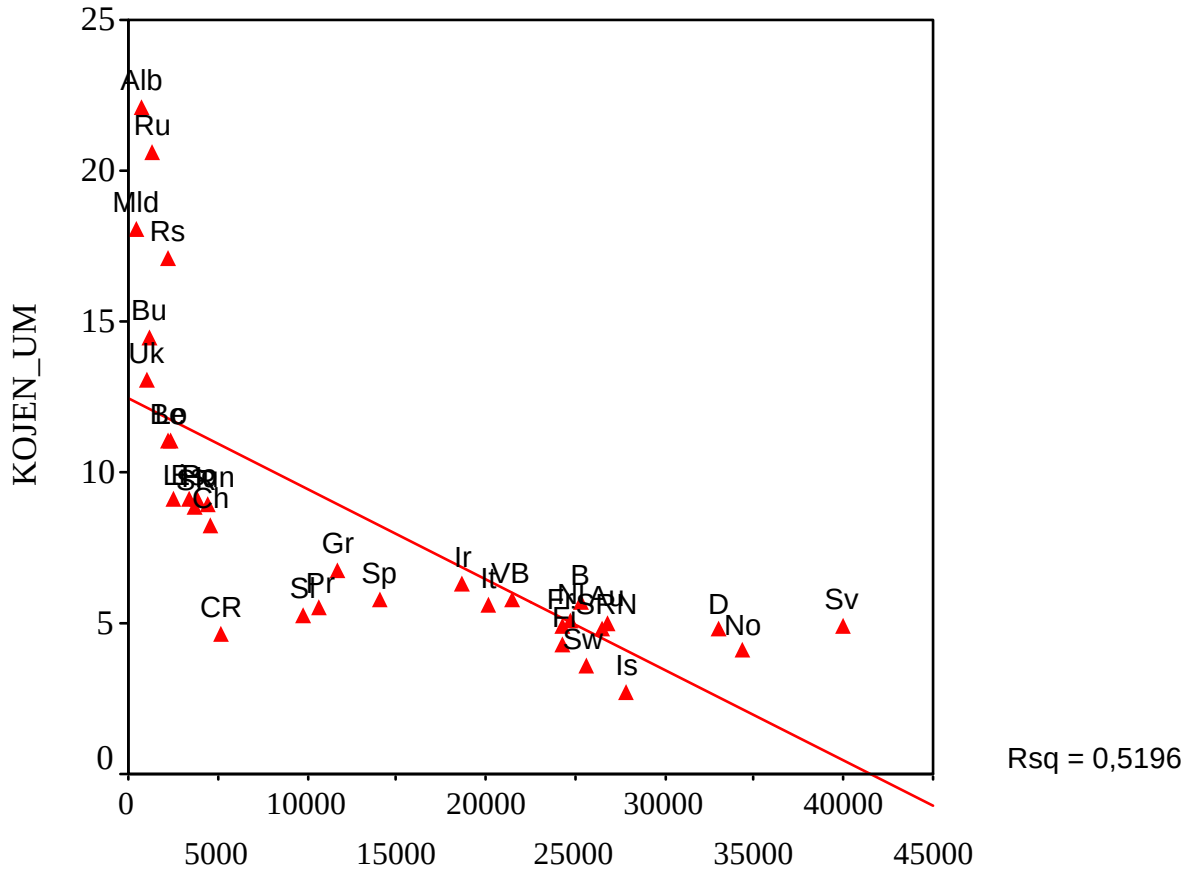
Příklad:

- Vztah mezi kojeneckou úmrtností (počet zemřelých kojenců během prvního roku života na 1000 živě narozených), a ekonomickou vyspělost země indikovanou hrubým národním produktem na hlavu (*Gross National Product – GNP*)
- Do jaké míry je v Evropě kojenecká úmrtnost podmíněna ekonomickou vyspělostí země. Budeme hledat vztah mezi ekonomickou vyspělostí země (což je naše nezávisle proměnná X) a mírou kojenecké úmrtnosti (proměnná závislá Y).



Země	Kojen. úmrt.	GNP na hlavu
Albánie	22	810
Belgie	5,6	25 380
Bělorusko	11,0	2 180
Bulharsko	14,4	1 220
Česko	4,6	5 115
Dánsko	4,7	33 040
Estonsko	9,0	3 360
Finsko	4,2	24 280
Francie	4,8	24 210
Chorvatsko	8,2	4 620
Irsko	6,2	18 710
Island	2,6	27 830
Itálie	5,5	20 090
Litevsko	9,0	2 540
Lotyšsko	11,0	2 420
Maďarsko	8,9	4 510
Moldávie	18,0	380
Německo	4,7	26 570
Nizozemsko	5,0	24 780
Norsko	4,0	34 310
Polsko	9,0	3 910
Portugalsko	5,4	10 670
Rakousko	4,9	26 830
Rumunsko	20,5	1 360
Rusko	17,0	2 260
Řecko	6,7	11 740
Slovensko	8,8	3 700
Slovinsko	5,2	9 780
Španělsko	5,7	14 100
Švédsko	3,5	25 580
Švýcarsko	4,8	39 980
Ukrajina	13,0	980
V. Británie	5,7	21 410

Regresní přímka popisující vztah mezi kojeneckou úmrtností a GNP



Analyze – Regression – Linear - Dependent (vložíme příslušnou závisle proměnnou) – *Independent* (vložíme nezávisle proměnnou)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,721[a]	,520	,504	3,5502

a. Predictors: (Constant), GNP_HEAD GNP na hlavu v US $ (1998)

- Hlavními ukazateli vhodnosti modelu pro naše data jsou údaje o velikosti R a R2 (*R Square*).

  Hodnota R je v případě jednoduché lineární regrese vlastně hodnotou Pearsonova korelačního koeficientu (ale pozor, zde nabývá pouze kladných hodnot, takže nemůže sloužit pro vyjádření korelačního vztahu – k tomu slouží standardizovaný koeficient beta, jehož výpočet je součástí výstupu z regresní analýzy). Čím vyšší je v regresi hodnota R, tím více si můžeme být jisti, že regresní model vyhovuje našim datům. V našem případě je R = 0,72, což není špatný výsledek.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,721[a]	,520	,504	3,5502

a. Predictors: (Constant), GNP_HEAD GNP na hlavu v US $ (1998)

- R2 signalizuje, jak **přesná bude predikce** hodnot podle naší regresní rovnice. Pokud data budou rozložena daleko od regresní přímky, chyba predikce bude velká a to vyústí v nízké R2. Pokud data budou těsně přimykat k regresní přímce, chyba predikce bude malá a R2 bude vysoké.
- R2 tak vlastně indikuje, jak silný je regresní vztah mezi dvěma proměnnými. Vynásobíme-li jej 100, získáme vlastně koeficient determinace, jak jsme o něm hovořili v předchozí kapitole. Pro naše data je R2 = 0,52 což značí, že rozptyl v datech je z 52 % způsoben chováním proměnné GNP na hlavu. Zbylých 48 % variance je třeba hledat v dalších, pravděpodobně neekonomických faktorech. Nicméně ekonomický vliv se zdá být pro úroveň kojenecké úmrtnosti v evropských zemích poměrně značný.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	422,577	1	422,577	33,527	,000[a]
	Residual	390,730	31	12,604		
	Total	813,307	32			

a. Predictors: (Constant), GNP_HEAD GNP na hlavu v US $ (1998)

b. Dependent Variable: KOJEN_UM

- Tabulka analýzy rozptylu, která je druhým výstupem z regresní analýzy, rovněž říká, zdali je model vhodný pro data, nebo ne, neboť měří rozdíl mezi skutečnými daty a daty, které vzniknou na základě aplikace regresního modelu.
- Z tabulky jsou pro praktickou práci nejdůležitější údaje o hodnotě F (mělo by být vyšší než 1) a jeho signifikance (Sig. by měla být nižší než 0,05).
- F je v našem případě mnohem větší než 1 a je signifikantní. Což značí, že vypočítaný regresní model je vhodný.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12,470	,950		13,124	,000
	GNP_HEAD GNP na hlavu v US $ (1998)	-3,007E-04	,000	-,721	-5,790	,000

a. Dependent Variable: KOJEN_UM

- Máme-li tedy důvěru v to, že má smysl pracovat s lineárním modelem regrese, podívejme se na parametry regresní přímky z tabulky, která je třetím základním výstupem z regresní analýzy.
- Vidíme, že obsahuje ve sloupcích údaje o nestandardizovaném koeficientu B a o standardizovaném koeficientu Beta. V jednoduché regresi pracujeme především s nestandardizovaným regresním koeficientem B. Standardizované koeficienty Beta se používají převážně v mnohonásobné regresi.
- V korelační analýze dat jsme se setkávali s koeficienty, které byly standardizovány, a proto nabývaly hodnot v rozsahu <0;1> nebo <-1;1>. Nestandardizovaný regresní koeficient může v podstatě nabýt hodnoty jakékoliv.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12,470	,950		13,124	,000
	GNP_HEAD GNP na hlavu v US $ (1998)	-3,007E-04	,000	-,721	-5,790	,000

a. Dependent Variable: KOJEN_UM

- Pro interpretaci našich dat je dobré vnímat regresní koeficient B dohromady spolu s korelačním koeficientem R2. Regresní koeficient B nám dává informaci o tom, jak velký vliv má nezávisle proměnná X na závisle proměnnou Y a současně umožňuje predikci Y pro jednotlivé případy. Jelikož však tato predikce bude nepřesná, R2 nám pomáhá odhadnout, jak velká nepřesnost v našich odhadech bude.
- V prvním řádku máme údaje o hodnotě *a*, což je naše konstanta (*Constant*). V našem případě má hodnotu 12,47. V průsečíku druhého řádku a sloupce B je nestandardizovaný regresní koeficient (-3,007E-04), a v průsečíku se sloupcem Beta máme údaj o standardizovaném koeficientu (-0,721). Údaje o signifikanci (*Sig.*) říkají, zdali náš odhad je dílem výběrové chyby nebo ne. Signikance menší než 0,05 (což ne nyní případ) značí, že náš výsledek není výsledkem výběrové chyby a že jej tedy můžeme očekávat i v základním souboru.

Sestavme nyní z údajů v tabulce 10.4 regresní rovnici. Má tuto podobu:

$$kojen.úmr. = 12{,}47 + (-0{,}00037 \times GNP)$$

- Hodnoty závisle proměnné, což je kojenecká úmrtnost, vzniknou jako součin hodnoty regresního koeficientu B (B = -0,0003) a hodnoty GNP.
- Konstanta, která má v našem případě hodnotu 12,47, zase říká, v jak vysoká bude hodnota závisle proměnné, když hodnota nezávisle proměnné bude nulová. Kdyby teoreticky byl GNP nulový, pak by kojenecká úmrtnost byla 12,5 (12,47) – takže konstanta ukazuje průměr proměnné Y.
- Hodnota regresního koeficientu B říká, o kolik se změní hodnota závisle proměnné y, když se hodnota nezávisle proměnné zvýší o jednotku, v níž je měřena. V našem příkladě má regresní koeficient hodnotu -0,00037, což umožňuje formulovat následující výrok. Zvýší-li se GNP na hlavu o jeden dolar, sníží se kojenecká úmrtnost o 0,00037. Zvýší-li se o GNP na hlavu o 1000 dolarů, kojenecká úmrtnost se sníží o ,0003*1000 = 0,37.
- Regresní rovnice dále umožňuje z hodnot nezávisle proměnné predikovat hodnotu proměnné závislé. Předpokládejme např., že by v nějaké zemi byl GNP na hlavu 30 000 dolarů. Jaká by v takové zemi byla kojenecká úmrtnost (k. ú.)? Pro zodpovězení této otázky stačí dosadit příslušné hodnoty do regresní rovnice:

k. ú. = 12,47 + (-0,00037 x 30 000)

k. ú. = 12,47 + (-11,1)

k. ú. = 1,37

Takže při GNP 30 000 dolarů na hlavu by měla být kojenecká úmrtnost velmi nízká, pouhých 1,37 zemřelých kojenců na 1000 živě narozených dětí.

Mnohonásobná lineární regrese

Cíle mnohonásobné regrese jsou stejné jako u regrese jednoduché:

- vysvětlit rozptyl v závisle proměnné *Y* . K tomu slouží statistika R2;
- odhadnout (vypočítat) vliv každé z nezávisle proměnných X na proměnnou závislou. Sílu tohoto vlivu sdělují nestandardizované regresní koeficienty *b*. Vliv každé nezávisle proměnné je odhadován tak, že je kontrolováno působení ostatních nezávisle proměnných, které vstupují do modelu. Mnohonásobná regrese prostřednictvím standardizovaných regresních koeficientů (*beta*) také pomáhá určit relativní sílu vlivu jednotlivých proměnných na proměnnou závislou – my tak zjistíme, které proměnné mají na rozptyl závisle proměnné největší vliv a které mají naopak vliv nejmenší.
- s pomocí sestavené regresní rovnice predikovat pro jednotlivé případy hodnoty závisle proměnné.

Předpoklady regresní analýzy

- Závisle proměnná Y musí být proměnná metrická (měřena na intervalové úrovni). Pokud není, musíme použít logistickou regresi.
- Nezávisle proměnné jsou měřeny rovněž na intervalové úrovni. Mohou to být i proměnné neintervalové, ale pouze dichotomické. Jelikož mnoho důležitých nezávislých proměnných nemá tuto vlasnost, překonáváme tento problém tím, že vytváříme *dummy* proměnné.
- Nezávisle proměnné by neměly být mezi sebou příliš vysoce korelovány, neboť to je porušením požadavku na absenci multikolinearity. Pokud v datech existuje multikolinearita, výsledky regrese jsou nespolehlivé. Vysoká multikolinearita zvyšuje pravděpodobnost, že a dobrý prediktor (= nezávisle proměnná) bude shledán statisticky nevýznamný a bude vyřazen z modelu.
- V datech nesmějí být odlehlé hodnoty (*outliers*), neboť na ty je regresní analýza citlivá. Odlehlé hodnoty mohou vážně narušit odhady parametrů rovnice.
- Proměnné musejí být v lineárním vztahu. Vícenásobná lineární regrese je založena na Pearsonově korelačním koeficientu, takže neexistence linearity způsobuje, že i důležité vztahy mezi proměnnými, pokud nejsou lineární, zůstanou neodhaleny.
- Proměnné jsou normálně rozloženy, jinak hrozí nepřesnost výsledků. Máme-li dostatečně velký vzorek, tento předpoklad nás nemusí příliš trápit z důvodů platnosti centrálního limitního teorému. Ten zaručuje, že porušení normality ve velkých výběrových souborech nemá příliš vážné následky.
- Vztahy mezi proměnnými vykazují homoskedascitu, tedy homogenitu rozptylu. Což znamená, že rozptyl v datech jedné proměnné bude víceméně shodný pro všechny hodnoty druhé proměnné. Např. pokud bude rozptyl v příjmech shodný pro všechny věkové skupiny, pak mezi věkem a příjmem bude existovat homoskedasticita. Opakem homoskedasticity je heteroskedasticita.

  Převzato od: de Vauss, David. 2002. Analyzing Social Scinece Data. SAGE, London., str. 343–344.)

Jak odhalit multikolinearitu a jak s ní naložit?

- Prozkoumejte jednotlivé bivariační korelace. Vysoké vzájemné korelace jsou zdrojem multikolinearity.
- Prozkoumejte test multikolinearity, který je jedním z výstupů vícenásobné regrese: k diagnóze poslouží jednak údaje o *variable inflation factor* (VIF), jednak údaje o toleranci (*tolerance*). Hrubé pravidlo říká, že pokud je ukazatel tolerance 0,2 a menší, pak v našich datech existuje multikolinearita. Stejně tak, pokud ukazatel VIF bude na úrovni hodnoty 5 a vyšší, máme v datech multikolinearitu.
- Pokud zjistíme, že multikolinearitu způsobuje vysoká bivariační korelace, je namístě vypustit problematickou proměnnou z analýzy. Nedopustíme se tím žádného zločinu, neboť když máme v datech dvě vysoce vzájemně korelované proměnné, velmi často to znamená, že obě indikují podobný jev. Tím, že jednu z těchto proměnných z regresního modelu vyřadíme, nijak jej neoslabíme. Pokud je multikolinearita zapříčiněna vzájemnou interkorelovaností několika proměnných, nabízí se řešení zkombinovat je do jedné nové proměnné. Tu vytvoříme např. s pomocí analýzy hlavních komponent (faktorové analýzy).

Jak prověřit normalitu?

- prozkoumejte šikmost a špičatost rozložení jednotlivých proměnných
- nechejte si udělat histogram s proloženou křivkou normálního rozložení
- použijte Kolmogorov-Smirnovův test
- podívejte se na rozložení dichotomické proměnné – pokud asi 80-90 % případů jsou v jedné kategorii dichotomie, musíme takovou dichotomii považovat za rozložení, které je vychýlené, a tudíž není normální.

Test linearity

- Bivariační linearitu můžeme odhadnout pomocí bodového grafu. Ten je však neúčinný v případě, že náš soubor obsahuje velké množství jednotek
- Prozkoumáme graf standardizovaných skutečných hodnot Y a predikovaných residuí Y (jak se to dělá si ukážeme za chvíli). Pokud graf vykazuje nelineární podobu, pak si můžeme být jisti, že buď jedna z nezávisle proměnných nebo kombinace nezávisle proměnných mají nelineární vztah s proměnnou závislou (Y). Tento graf nám také pomůže odhalit případnou *heteroskedasticitu* v datech.
- Pokud vztahy mezi našimi proměnnými nejsou lineární, musíme se pokusit ty proměnné, u nichž jsme detektovali nelinearitu, statisticky transformovat (např. ji logaritmujeme, nebo odmocníme apod.) tak, abychom požadavek linearity naplnili. Nepomůže-li tento postup, musíme použít jiný typ regrese – nelineární regresi), která není na linearitu citlivá.

Různé formy mnohonásobné regrese

- Metoda standardní (tzv. metoda *Enter*). Všechny proměnné jsou do výpočtu vloženy najednou
- Metoda postupného vkládání (*Stepwise*). Proměnné jsou vkládány do výpočtu regrese postupně podle předem zadaných matematických kritérií. V této metodě výzkumník nekontroluje pořadí proměnných, jak postupně vstupují do analýzy, o pořadí rozhoduje SPSS – to je algoritmus výpočtu a kritéria vkládání. Je to metoda, které se s trochou nadsázky říká metoda pro nalezení „nejlepšího“ modelu.
- Metoda hierarchická (*Blocks*). Pořadí, v němž proměnné vstupují do výpočtu řídí výzkumník a odvíjí se od jeho kauzálního modelu, který testuje.

Každá metoda přináší interpretačně odlišné výsledky !

Metoda Enter

- Tuto metodu použijeme tehdy, když chceme popsat, jak velký podíl variance závisle proměnné je vysvětlen nezávisle proměnnými (R2), dále jak velký vliv má každá z nezávisle proměnných na proměnnou závislou při kontrole vlivu působení ostatních proměnných (nestandardizované regresní koeficienty) a konečně jaký je relativní důležitost každé z nezávisle proměnných (standardizované regresní koeficienty beta).

Tab. 1. Výsledky regrese metodou Enter

Proměnná	B	Beta	Sig
X_1 úzkost	2,5	0,28	0,01
X_2 sociální dovednosti	-1.1	-0,09	0,24
X_3 symptomy psychózy	1,4	0,21	0,04
X_4 deprese	6,1	0,72	0,00
X_5 prospěch	1,3	0,09	0,26
X^6 skóre aktivity	-2,3	-0,29	0,00

R^2 = 0,59, Sig. = 0,001
Dependent variable: sociální izolace

2. Metoda Stepwise

- Metoda stepwise je metodou k nalezení „nejlepšího“ modelu. Mějme stejné proměnné, které ale do regrese vložíme postupně, nikoliv najednou. Jelikož máme šest nezávisle proměnných, může regrese vypočítat v této metodě až šest různých modelů. Každý model se bude od toho předchozího lišit v tom, že v něm bude o jednu nezávisle proměnnou více. Do výpočtu a do modelu vstupují pouze ty proměnné, které jsou statisticky významně vztaženy s proměnnou závislou. My už víme z výpočtu metodou *enter*, že pouze čtyři proměnné statisticky signifikantní ve svém působení na proměnnou Y, takže metoda stepwise vypočítá pouze čtyři modely.

Tab. 2. Výsledky regrese metodou Stepwise

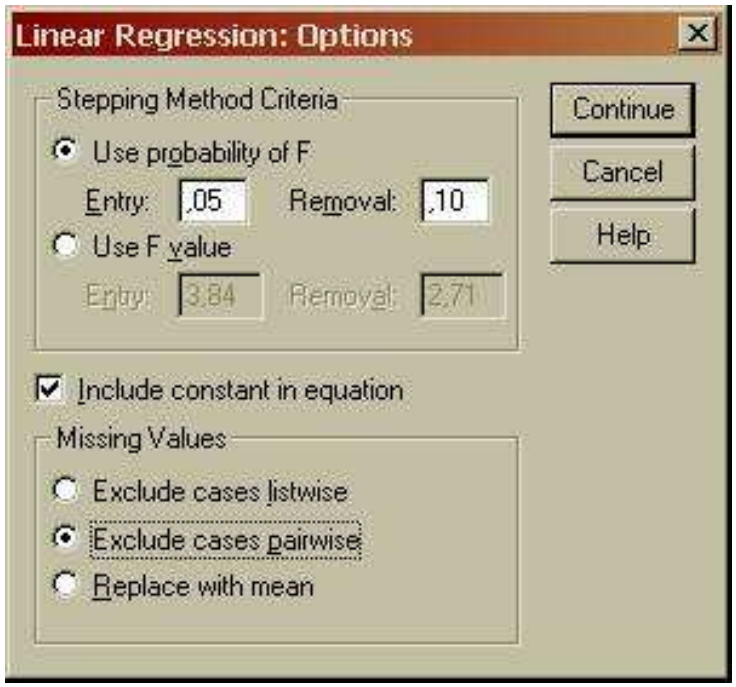
				Change statistics	
Model	R	R Square	Adjusted R Square	R Square Change	Sig. F Change
1	0,68	0,46	0,45	0,46	0,00
2	0,71	0,50	0,49	0,04	0,00
3	0,74	0,55	0,54	0,05	0,00
4	0,76	0,58	0,56	0,03	0,00

a Predictors: (Constant), deprese
b Predictors: (Constant), deprese, aktivita
c Predictors: (Constant), deprese, aktivita, úzkost
d Predictors: (Constant), deprese, aktivita, úzkost, psychóza

Jak provést regresi a jak rozumět výstupům z regresní analýzy v SPSS

- SPSS vypočítává v mnohonásobné lineární regresi tři hlavní typy výstupů:
- adekvátnost modelu – R2
- tabulku ANOVA – test signifikance pro R2
- regresní koeficenty pro jednotlivé nezávisle proměnné

Důležitý je způsob práce zacházení s chybějícími hodnotami (missing vlaues). Default je v SPSS *Exclude cases listwise*, což není příliš výhodné. Znamená to, že pokud některý případ bude mít chybějící hodnotu v některé z proměnných, které vstupují do analýzy, bude z analýzy vyloučen. *Pairwise* způsob dělá to, že případ s chybějící hodnotou vynechává pouze ve výpočtech s tou proměnno, kde nemá hodnoty, ale ve všech ostatních výpočtech případ vrací do hry. Není tedy z analýzy úplně ztracen, jako je tomu u způsobu listwise.



Výstupy – metoda ENTER

Variables Entered/Removed [b]

Model	Variables Entered	Variables Removed	Method
1	Z_V západ-východ, TFR úhrnná plodnost, KOJEN_UM kojenecká úmrtnost, GNP_HEAD GNP na hlavu v US $ (1998) [a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: LIFE_EXP nadeje dožití

Descriptive Statistics

	Mean	Std. Deviation	N
LIFE_EXP nadije dožití	74,30	4,004	33
KOJEN_UM kojenecká úmrtnost	8,2909	5,04141	33
TFR úhrnná plodnost	1,4576	,27160	33
GNP_HEAD GNP na hlavu v US $ (1998)	13898,6364	12086,42183	33
Z_V západ-východ	,48	,508	33

- Toto je výpočet průměrů všech proměnných, které vstoupily do regrese a jejich směrodatných odchylek. Pro samotnou interpretaci výsledků regrese nejsou důležité, ale *Descriptives* současně tisknou i matici korelací (Pearsonovy koeficienty lineární korelace) a ta je už regresi důležitá – především pro prvotní kontrolu multikolinearity – mezi proměnnými by neměla být žádná korelace větší než 0,9.

Correlations

		LIFE_EXP nadìje dožití	KOJEN_UM kojenecká úmrtnost	TFR úhrnná plodnost	GNP_HEAD GNP na hlavu v US $ (1998)	Z_V západ-východ
Pearson Correlation	LIFE_EXP nadìje dožití	1,000	-,826	,328	,859	-,874
	KOJEN_UM kojenecká úmrtnost	-,826	1,000	-,085	-,721	,696
	TFR úhrnná plodnost	,328	-,085	1,000	,433	-,413
	GNP_HEAD GNP na hlavu v US $ (1998)	,859	-,721	,433	1,000	-,883
	Z_V západ-východ	-,874	,696	-,413	-,883	1,000
Sig. (1-tailed)	LIFE_EXP nadìje dožití	.	,000	,031	,000	,000
	KOJEN_UM kojenecká úmrtnost	,000	.	,319	,000	,000
	TFR úhrnná plodnost	,031	,319	.	,006	,008
	GNP_HEAD GNP na hlavu v US $ (1998)	,000	,000	,006	.	,000
	Z_V západ-východ	,000	,000	,008	,000	.
N	LIFE_EXP nadìje dožití	33	33	33	33	33
	KOJEN_UM kojenecká úmrtnost	33	33	33	33	33
	TFR úhrnná plodnost	33	33	33	33	33
	GNP_HEAD GNP na hlavu v US $ (1998)	33	33	33	33	33
	Z_V západ-východ	33	33	33	33	33

Adekvátnost modelu – R2

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,931[a]	,867	,848	1,562	,867	45,540	4	28	,000

a. Predictors: (Constant), Z_V západ-východ, TFR úhrnná plodnost, KOJEN_UM kojenecká úmrtnost, GNP_HEAD GNP na hlavu v US $ (1998)

b. Dependent Variable: LIFE_EXP nadìje dožití

- V této tabulce nás zajímají dva údaje, *R Sguare* (R2) a *Adjusted R2*. R2 říká, jak velké množství variance závisle proměnné (naděje dožití) je vysvětleno sadou námi zvolených nezávisle proměnných. V tomto případě je R2 0,87 neboli 87 % variance závisle proměnné je vysvětleno nezávisle proměnnými. Učebnice ale doporučují, abychom se dívali spíše na údaj o *Adjusted R Square*. Je to z toho důvodu, že velikost R2 může být uměle zvýšena počtem proměnných, které vstupují do analýzy – a právě *Adjusted R Square* bere počet proměnných v úvahu a velikost R2 na základě toho upravuje (adjustuje). Je to důležité především pro malé soubory, ve velkých souborech se obě statistiky budou dosti podobat.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	444,626	4	111,156	45,540	,000[a]
	Residual	68,344	28	2,441		
	Total	512,970	32			

a. Predictors: (Constant), Z_V západ-východ, TFR úhrnná plodnost, KOJEN_UM kojenecká úmrtnost, GNP_HEAD GNP na hlavu v US $ (1998)

b. Dependent Variable: LIFE_EXP nadíje dožití

- V této tabulce se dozvídáme, zdali platí nulová hypotéza, že R2 = 0. To nám ozřejmí F test a jeho signifikance. Je-li signifikance menší než 0,5, nemůžeme nulovou hypotézu zamítnout a máme jistotu, že námi zjištěné R2 můžeme očekávat také v populaci (v našem školním příkladu, kdy máme vzorek evropských zemí, které nebyly vybrány náhodou, tato inference není tak úplně na místě).

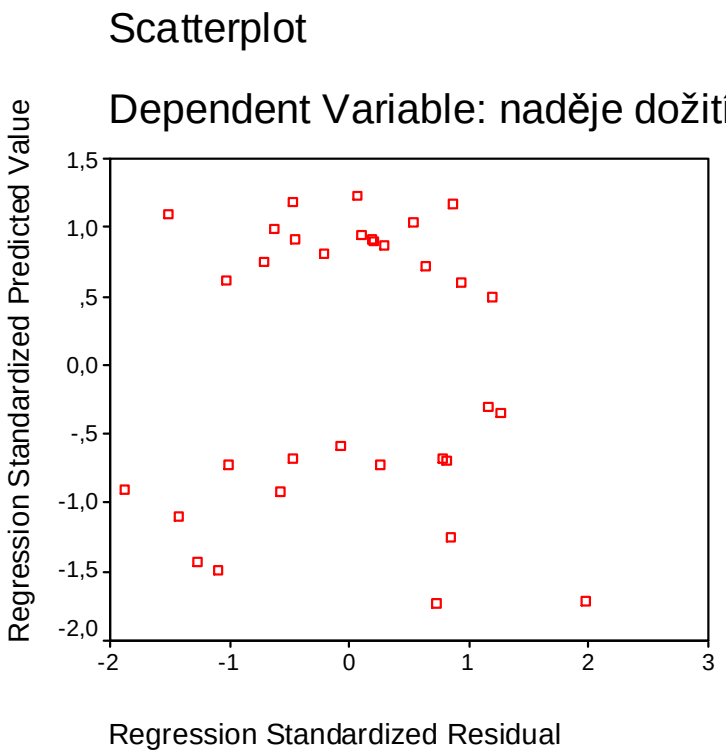
Tab. 3: Regresní koeficienty a další statistiky mnohonásobné regerse

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	76,725	2,012		38,139	,000	72,604	80,846					
	KOJEN_UM kojenecká úmrtnost	-,317	,087	-,399	-3,644	,001	-,496	-,139	-,826	-,567	-,251	,396	2,525
	TFR úhrnná plodnost	,620	1,225	,042	,506	,617	-1,889	3,130	,328	,095	,035	,689	1,451
	GNP_HEAD GNP na hlavu US $ (1998)	,305E-05	,000	,190	1,179	,248	,000	,000	,859	,218	,081	,183	5,475
	Z_V západ-východ	-3,243	1,191	-,411	-2,724	,011	-5,682	-,805	-,874	-,458	-,188	,209	4,787

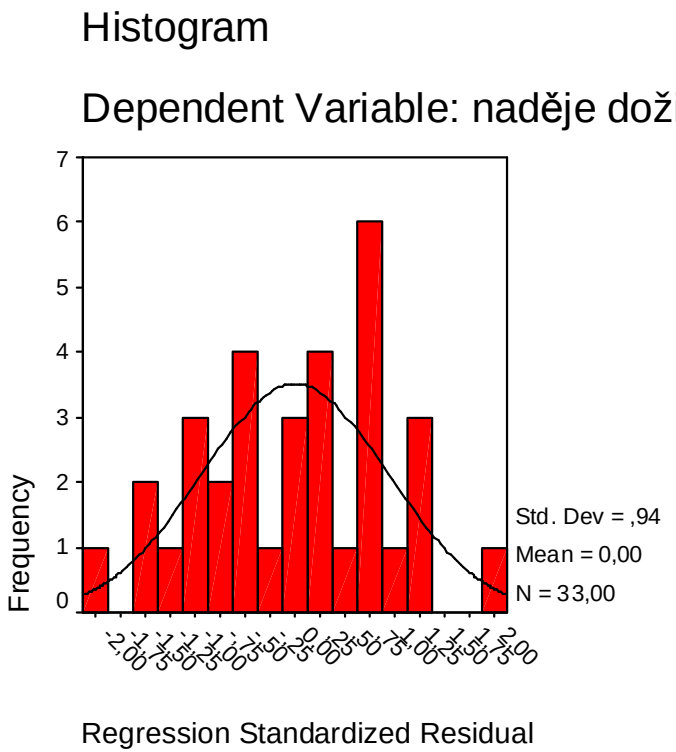
a. Dependent Variable: LIFE_EXP nadìje dožití

Kontroly předpokladů
– zda je užití lineární regresní analýzy vhodné

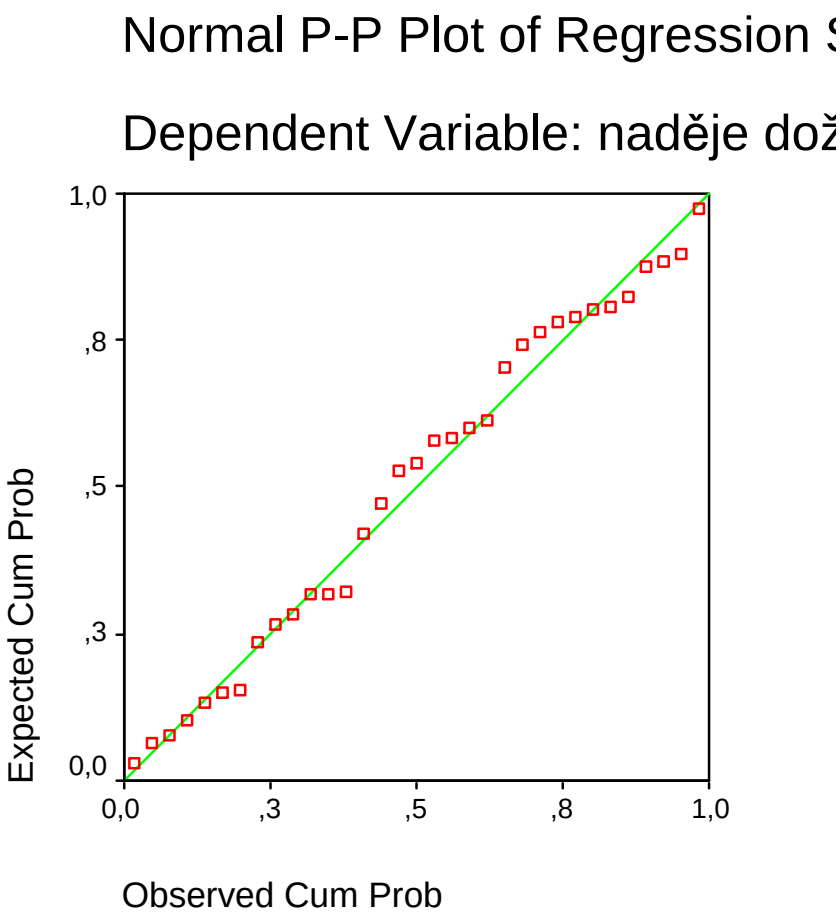


Graf by neměl vykazovat žádný vzorec v uspořádání proměnných: Náš bohužel ukazuje, což je signálem, že předpoklad lienarity a homoskedasticity není naplněn.

Kontroly předpokladů
– zda je užití lineární regresní analýzy vhodné



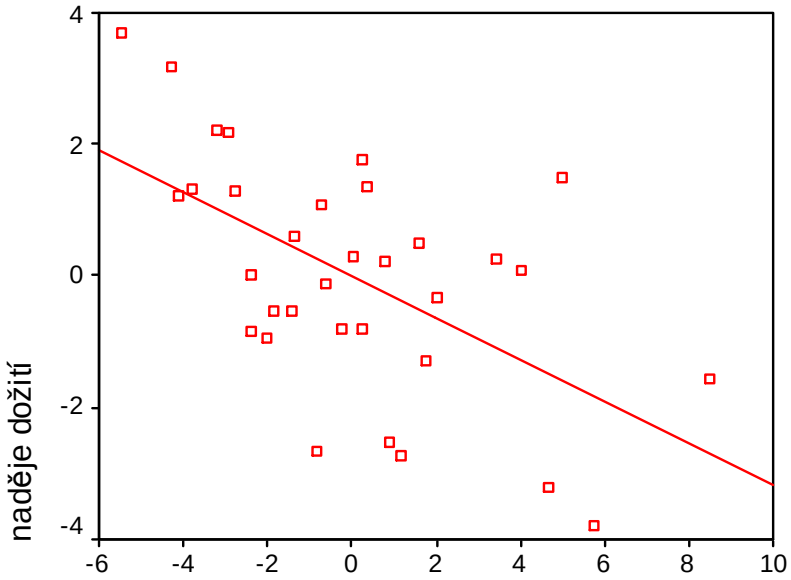
Histogram reziduí ukazuje, že rezidua nejsou normálně rozložena, což znamená že požadavek na mnhonásobnou normalitu je porušen. Což naznačuje i Q-Q graf (viz níže).


Normal P-P Plot of Regression S
Dependent Variable: naděje dož
Expected Cum Prob
Observed Cum Prob
0,0
,3
,5
,8
1,0

Grafy Partial Regression Plots testují homoskedasticitu:

Partial Regression Plot

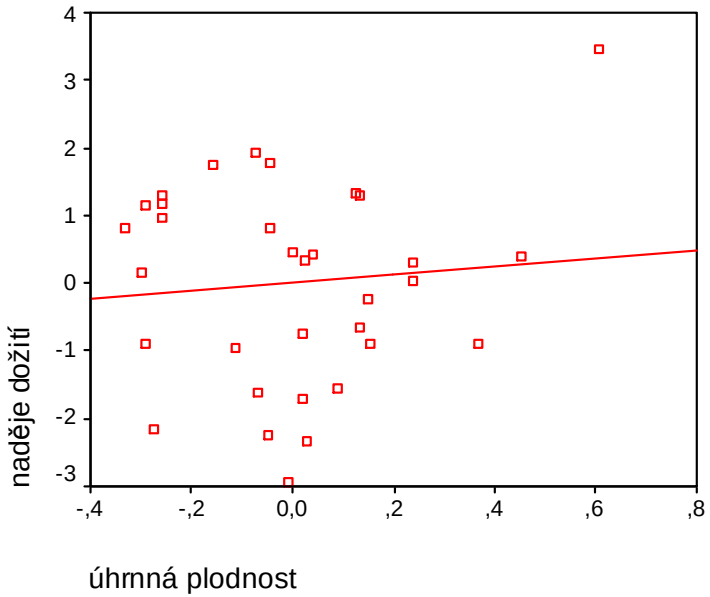
Dependent Variable: naděje dožití



Ok, body jsou rovnoměrně rozloženy kolem přímky.

Partial Regression Plot

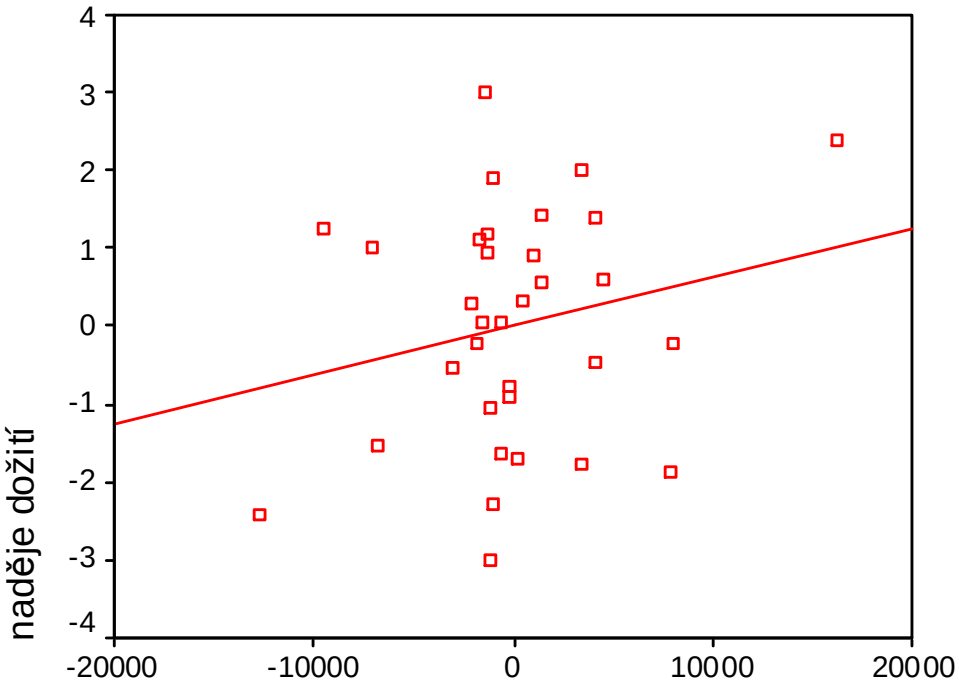
Dependent Variable: naděje dožití



Toto je problém, je tam zužující se trend. Heteroskedasticita.

Partial Regression Plot

Dependent Variable: naděje dožití



Rovněž špatně

- V případě, že testy využití vychází špatně, jsou možnosti:

  - použít metodu lineární regrese „Stepwise“ (postupné vkládání proměnných do modelu)

  - použít metodu logistické regrese