LEKCE 10
ZÁKLADY LINEÁRNÍ REGRESE
(s dodatkem o odhadu regresní křivky)

Zjistíme-li prostřednictvím korelačního koeficientu, že mezi dvěma proměnnými existuje souvislost, to je že změny hodnot jedné proměnné jsou doprovázeny konsistentními změnami hodnot v druhé proměnné, jsme schopni vyslovit určitou předpověď, predikci. Např. zjistíme-li, že mezi pohlavím a příjmem existuje poměrně silná souvislost, kdy muži mají většinou vyšší příjem než ženy, jsme schopni dělat predikce – bude-li nám např. představen nový manželský pár, budeme moci s jistou pravděpodobností předpokládat, že muž bude mít vyšší příjem než žena.[1] Co však z korelace nejsme schopni vyvodit, je o kolik více muži vydělávají více než ženy. Technikou analýzy, která nám umožní tento druh výpovědi formulovat, je regresní analýza.

Regresní analýza je jednou z nejpoužívanějších statistických technik analýzy dat ve společenských vědách. 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 – a v tom je její síla – také říci

1. jak velký vliv má nezávisle proměnná X na proměnnou závislou Y,
2. 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.

Jaké jsou podmínky pro užití lineární 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.

Ukažme si vše na příkladu. Vyjděme z demografických dat z roku 1999, která jsou obsažena v tab. 10. 1. Uvádějí kojeneckou úmrtnost, tedy 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*) v amerických dolarech. Budeme se zajímat o to, do jaké míry je v Evropě kojenecká úmrtnost podmíněna ekonomickou vyspělostí země. Řečeno jinými slovy, budeme hledat vztah mezi ekonomickou vyspělostí země (což je naše nezávisle proměnná X) a mírou kojenecké úmrtnosti (proměnná závislá Y). Data k tomuto příkladu jsou převzata ze souboru dmg-data.sav.

Jak je z tabulky vidět, kojenecká úmrtnost v evropských zemích značně variuje a v roce 1999 se pohybovala mezi 2,6 úmrtími kojenců na 1000 obyvatel (Island) až po 22 zemřelých kojenců (Albánie). Evropský průměr byl v roce je 8,3. Již pouhá „okometrická“ analýza naznačuje, že země s vyšším GNP mají nižší kojeneckou úmrtnost a naopak..

[1] Výraz „s jistou pravděpodobností“ je zcela na místě. Pokud bude korelace vysoká, bude tato pravděpodobnost vyšší, než když tato korelace bude pouze střední síly. Stoprocentní jistotu předpovědi ale z korelační souvislosti nemůžeme odvodit nikdy, neboť jak jsme již uvedli dříve, ani vysoká korelace ještě neznamená, že mezi sledovanými jevy je příčinný vztah.

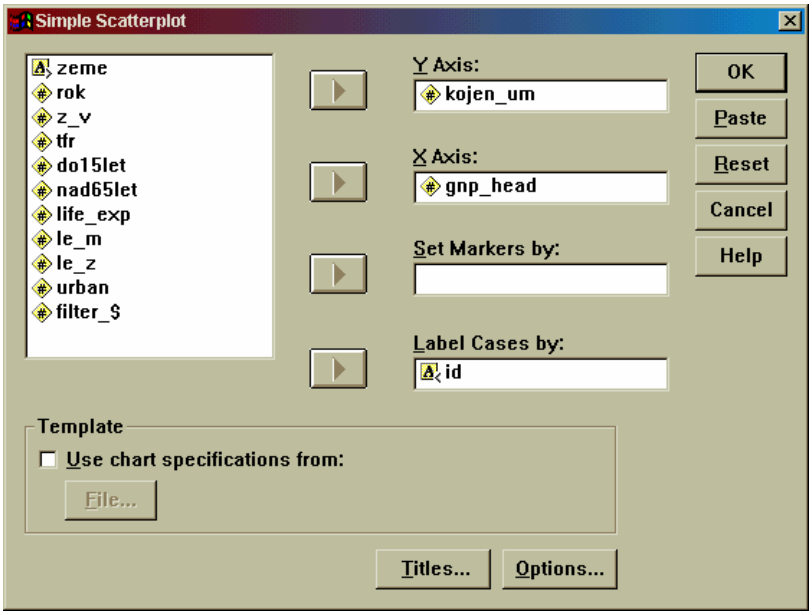
Jakýkoliv příklad regresní analýzu je vždy dobré začít nejdříve řešit nejdříve graficky, abychom zjistili, zdali vztah mezi proměnnými má lineární charakter. Použijeme k tomu bodový graf, který zadáme tak, jak je zobrazeno na obr. 10.1.

Tab. 10. 1: Kojenecká úmrtnost a hrubý národní produkt (GNP) na hlavu v evropských zemích (1999)

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

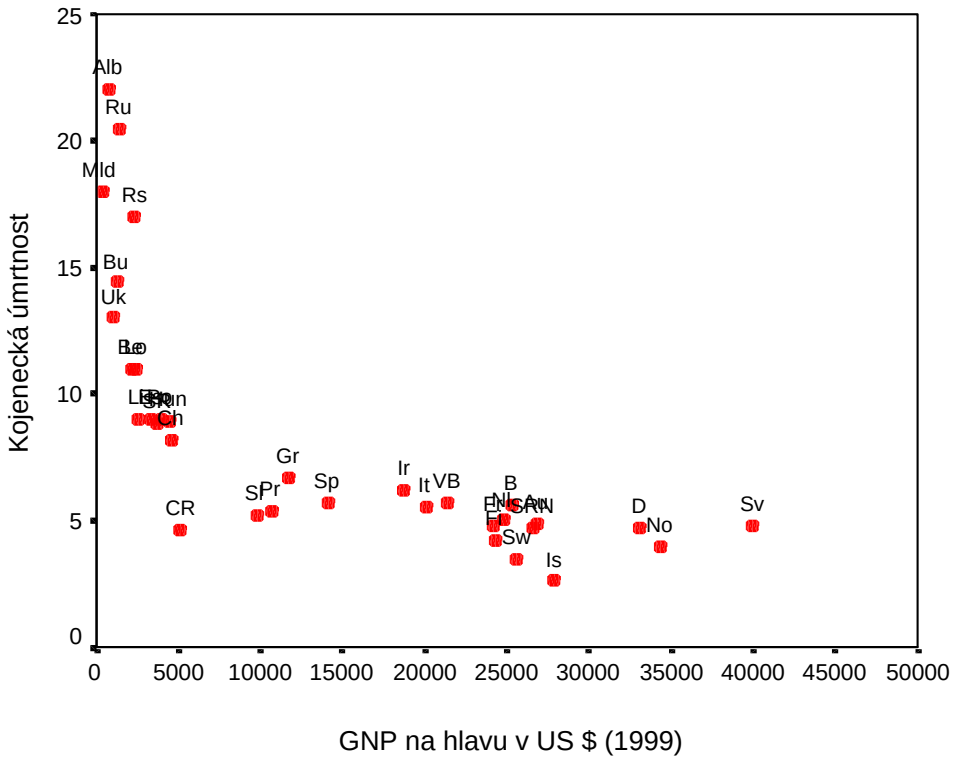
Graphs – Scatter – Simple – Define – (do obdélníčku *Y Axis* umístíme jméno závisle proměnné, v našem případě má jméno *kojen_um*, do obdélníčku *X Axis* umístíme jméno nezávisle proměnné, v našem případě *gnp_head*. Do obdélníčku *Label Cases by* vložíme jména zemí v jejich zkratce, což je proměnná *Id* – viz obr. 10.1). Abychom k jednotlivým bodům dostali jejich popisek, musíme si ještě kliknout na *Options* a v nich zaškrtnout okénko *Display chart with case labels*.

Obr. 10.1: Zadání pro bodový graf



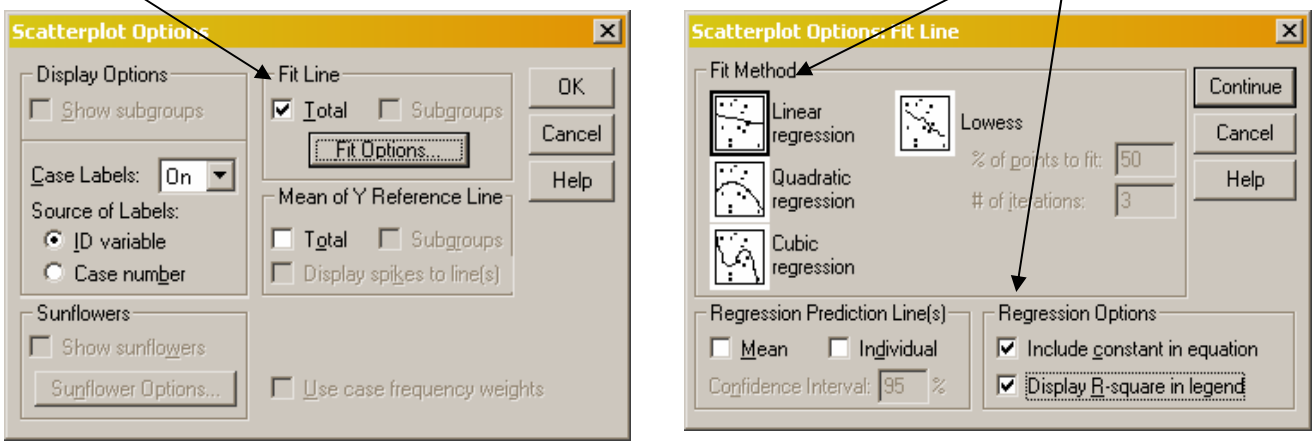
Výsledkem tohoto zadání je graf, který je zobrazen jako obr. 10.2. Z obrázku je patrné, že tvar vztahu mezi GNP a kojeneckou úmrtností nemá lineární charakter, takže bychom za normálních okolností neměli techniku lineární regrese použít a měli bychom hledat jiný způsob, co s tím udělat – ukázku, jak to udělat naleznete v doplňku na konci tohoto textu. Nicméně dovolte z didaktických důvodů regresní analýzu použít a daty proložit regresní přímku.

Obr. 10. 2: Kojenecká úmrtnost v závislosti na GNP na hlavu v evropských zemích (data z roku 1999)



Úkolem regresní analýzy je proložit body grafu regresní přímku. Vzhledem k tomu, že počet možných přímek, jimiž by mohly být body proloženy, je nesmírně velký (resp. nekonečný), používá se k této úloze metoda nejmenších čtverců, která nalezne takovou přímku, aby součet druhých mocnin vzdáleností jednotlivých bodů od přímky byl nejmenší (v tomto případě je výsledkem jediná přímka). V našem případě je výsledek metody nejmenších čtverců přímka, kterou ukazuje obrázek 10.3.

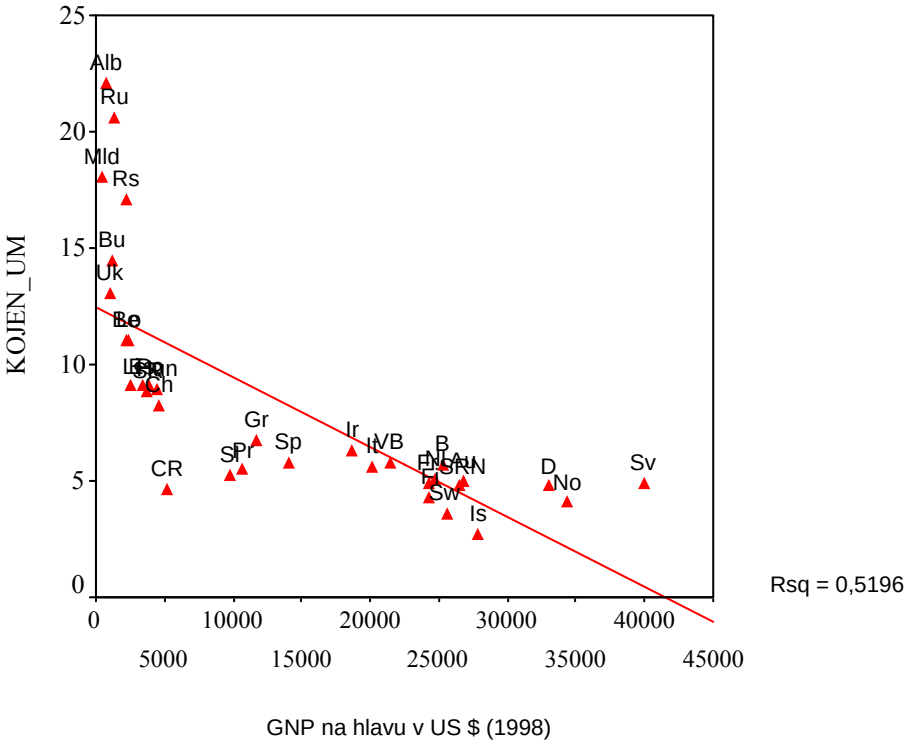
Regresní přímku do grafu dostaneme tak, že dvakrát klikneme na graf, abychom jej mohli editovat. Pak zvolíme z nabídky menu zvolíme *Chart – Options* a budeme požadovat, aby do grafu byla zanesena Fit Line (zaškrtneme políčko *Total*). Potom ještě klikneme na *Fit Options*, abychom zvolili metodu výpočtu – v novém dialogovém okně, které se objeví, budeme klinutím požadovat lineární regresi.



Na obrázku 10.3 je zřetelně vidět, že část bodů je od přímky poměrně značně vzdálena – pokud by regresní přímka měla být dobrým modelem vztahu těchto dvou proměnných, měly by se všechny body k přímce těsně přimykat, což však není např. případ Albánie (Alb), Rumunska (Ru), ČR nebo Švýcarska

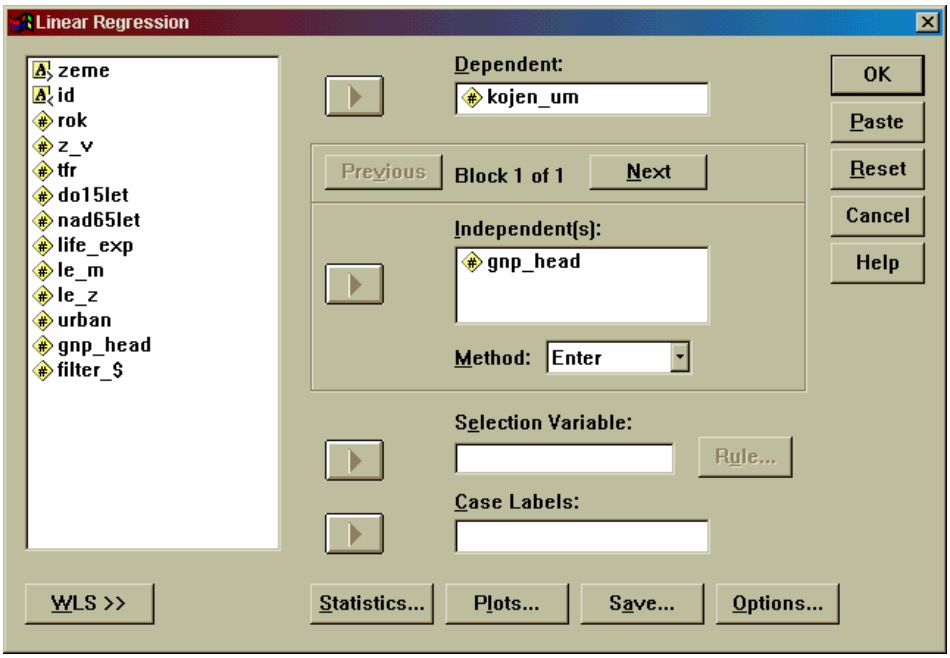
(SV). Kdyby regresní model platil beze zbytku, musela být mít např. ČR vzhledem ke svému GNP více než dvojnásobnou kojeneckou úmrtnost, než jakou ve skutečnosti má, a naopak Švýcarsko by muselo mít kojeneckou úmrtnost asi pětinásobně nižší, než jakou má (O rozdílu mezi skutečnými hodnotami a hodnotami, vypočtenými z modelu budeme ještě hovořit).

Obr. 10. 3: Regresní přímka popisující vztah mezi urbanizací a populačním stárnutím



Vypočítejme si nyní parametry regresní rovnice. Regresi vypočítáme v SPSS následujícím sledem příkazů.

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



Výsledkem výpočtu je několik tabulek. Interpretaci regresní analýzy začínáme vždy tím, že zhodnotíme, zdali je regresní přímka adekvátním modelem pro příslušná data. Proto nejdříve zkoumáme tabulku s údaji o R, R Square (tab. 10.2) a tabulku ANOVy (10.3)

Tab. 10. 2:

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 R^2 (*R Square*). Hodnota R je v případě jednoduché lineární regrese vlastně hodnotou Pearsonova korelačního koeficientu[2] (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.

R^2 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 R^2. Pokud data budou těsně přimykat k regresní přímce, chyba predikce bude malá a R^2 bude vysoké.

R^2 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 R^2 = 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[3].

[2] Jde o tzv. vícenásobný korelační koeficient, který udává jaká je závislost mezi napozorovanými hodnotami závislé proměnné a odhadnutými hodnotami na základě regresního modelu.

[3] Upozorněme též, že přímka není zcela vhodným modelem, a proto v případě některé nelineární křivky by mohla být hodnota koeficientu determinace ještě vyšší.

Tab. 10. 3:

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 (tab. 10.3) rovněž říká, zdali je model vhodný pro data, nebo ne, to je 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 a jeho signifikance (Sig. by měla být nižší než 0,05). F je v našem případě signifikantní, což značí, že vypočítaný regresní model je vhodný.

Máme-li tedy důvěru v to, že má smysl pracovat s lineárním modelem regrese,[4] podívejme se nyní na parametry regresní přímky (viz tab. 10.4). Vyčteme je ve sloupci *Unstandardized Coefficients* (nestandardizované koeficienty). Z údajů v tabulce 11.4 pak můžeme sestavit regresní rovnici.:

$$k.ú. = 12,47 + (-0,00037 \times GNP)$$

Tab. 10. 4:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
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

Jak tuto rovnici čteme? Hodnoty závisle proměnné, což je kojenecká úmrtnost, vzniknou jako součin hodnoty regresního koeficientu B (B = -0,0003) [5] a hodnoty GNP. Konstanta, která má v našem případě hodnotu 12,47, zase říká, v jaké výšce protíná regresní přímka osu Y, když hodnota závisle proměnné je nula. Tato konstanta je důležitá interpretačně. Říká totiž, jak vysoká bude hodnota závisle proměnné, když hodnota nezávisle proměnné bude nulová. Řečeno jinými slovy, kdyby teoreticky byl GNP nulový, pak by kojenecká úmrtnost byla 12,5 (12,47).

Regresní rovnice umožňuje analyticky několik věcí. Tak především 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 (samozřejmě za předpokladu, že všechno ostatní zůstane konstantní). V našem příkladě

[4] Jak si ukážeme později, SPSS umí data proložit i jinou funkcí než lineární, např. kvadratickou, exponenciální či logaritmickou.

[5] Výraz -3,007E-0,4 je zkráceným matematickým výrazem, který umožňuje, abychom nemuseli psát dlouhou řadu nul za desetinnou čárkou. Poslední část tohoto výrazu, to je E-0,4 říká, o kolik pozic máme posunout desetinnou čárku doleva. V našem případě musíme posunout o čtyři desetinná místa.

má regresní koeficient hodnotu -0,00037, což umožňuje formulovat následující výrok.[6] 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

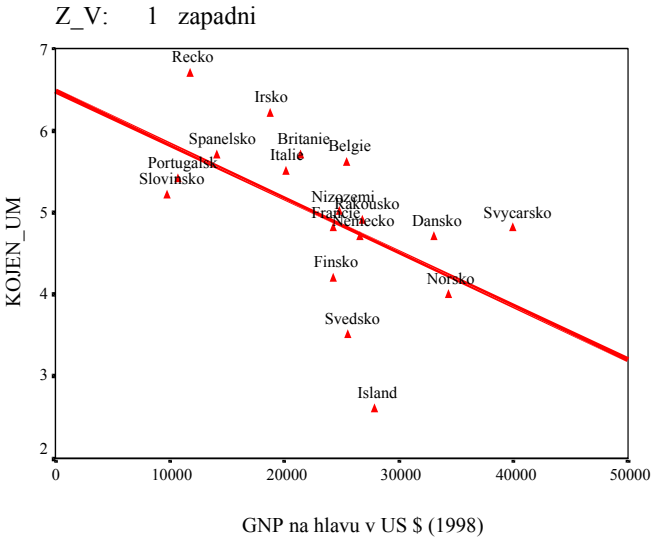
Takže při GNP 30 000 dolarů na hlavu by měla být kojenecká úmrtnost velmi nízká, pouhých 1,37 zemřelých kojenců na 1000 živě narozených dětí.

* * *

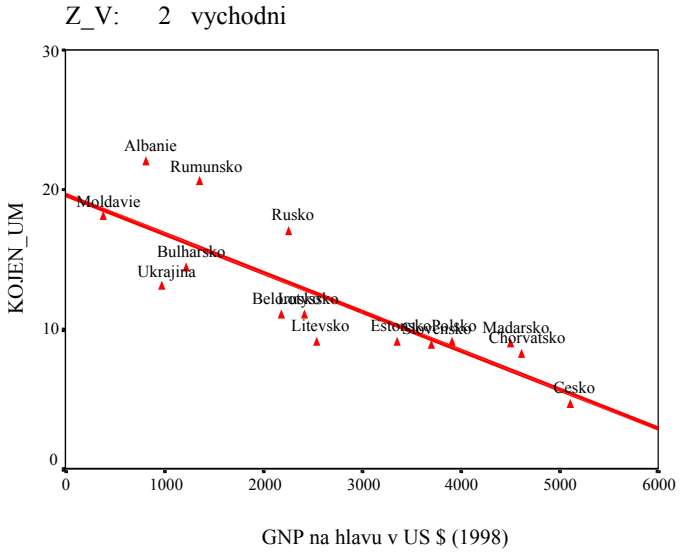
Při pozorném pohledu na obrázek 10. 2 se možná mnohým může zdát, že rozložení zemí v sobě vlastně skrývá dva modely. Jeden je modelem pro bývalé země východoevropské, druhý pak pro země západoevropské. Zdá se tedy, že by mohlo mít smysl neanalyzovat vztah mezi kojeneckou úmrtností a GNP pro všechny evropské země najednou, nýbrž rozdělit celou analýzu do dvou podsouborů: podsoubor zemí západních, k nimž přičleníme i Slovinsko[7] a podsoubor zemí východních. Zkusme to.

Analýzy provedeme přes proceduru Data – Split file. Použijeme k tomu dichotomickou proměnnou z_v, která rozděluje země na západní a východní. Nejdříve uděláme grafy a v nich necháme zobrazit regresní přímku (viz obr. 10.4 a obr. 10. 5).

Obr. 10. 4. Západní Evropa



Obr. 10. 5. Východní Evropa



[6] Až dosud jsme se v analýze dat 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.

[7] Důvod tohoto kroku vám ozřejmíme na konci této kapitoly.

A nyní necháme opět spočítejme pro oba podsoubory lineární regresi. Viz výstupy v tab. 10.6a 10.6b.

Tab. 10.6a:

Model Summary

Z_V	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1 zapadni	1	,553[a]	,306	,262	,8334
2 vychodni	1	,838[a]	,702	,680	2,8577

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

Tab. 10.6b:

ANOVA[b]

Z_V	Mo		Sum of Squares	df	Mean Square	F	Sig.
1 zapadni	1	Regression	4,891	1	4,891	7,042	,017[a]
		Residual	11,113	16	,695		
		Total	16,004	17			
2 vychodni	1	Regression	250,609	1	250,609	30,689	,000[a]
		Residual	106,160	13	8,166		
		Total	356,769	14			

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

b. Dependent Variable: KOJEN_UM

Srovnáme-li výsledky s regresí pro celý soubor, pak lineární model je pro západní země horším modelem. R^2 se snížilo na 0,31 (vysvětluje tedy pouze 31 % variance kojenecké úmrtnosti) a síla vztahu (R = 0,55) je také samozřejmě nižší. Zato pro východoevropské země nastalo zlepšení. R se zvýšilo na 0,84, R^2 se zvýšilo na 0,70, takže plných 70 % variance kojenecké úmrtnosti je vysvětleno velikostí GNP na hlavu.

Co z toho všeho vyplývá? Dospěli jsme zajímavému závěru. Evropa se z hlediska kojenecké úmrtnosti dělí do dvou oblastí, západoevropské a východoevropské. Zatímco v západoevropských zemích není příliš vhodné modelovat vztah mezi GNP a kojeneckou úmrtností prostřednictvím lineární přímky, a nemá proto velký smysl na něj aplikovat model lineární regrese, ve střední a východní Evropě lineární podobu tento vztah má. Což jinými slovy znamená, že onen poměrně výrazný ekonomický vliv na kojeneckou úmrtnost, který jsme nalezli v naší analýze pro všechny evropské země, byl způsoben velkou váhou tohoto faktoru v zemích východoevropských. Zdá se tedy, že od jisté úrovně ekonomické vyspělosti tento lineární vliv přestává platit.

Rovnice lineární regrese je pro východoevropské země následující (viz tab. 10.7):

$$k.ú. = 19{,}63 - 0{,}0028*GNP$$

Tab. 10.7:

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	19,629	1,516		12,949	,000
	GNP_HEAD GNP na hlavu v US $ (1998)	-,0028	,001	-,838	-5,540	,000

a. Dependent Variable: KOJEN_UM

Což znamená, že pokud se zvýší GNP na hlavu o 1000 dolarů, sníží se kojenecká úmrtnost o 2,8 zemřelých kojenců na 1000 živě narozených dětí. A to už je vskutku velmi zajímavý výsledek.

Predikované hodnoty a rezidua

Predikční schopnosti regresní přímky můžeme využít také k tomu, abychom se podívali, do jaké míry regresní model – neboť vypočtená regresní přímka je skutečně pouze modelem, který ve zhuštěné a elegantní podobě sumarizuje vztah mezi sadou hodnot dvou proměnných – zpětně reprodukuje výchozí hodnoty závisle proměnné. Pokud je bude reprodukovat relativně věrně, můžeme si být jisti, že regresní rovnice je dobrým modelem dat.

To, jak rovnice reprodukuje výchozí hodnoty závisle proměnné, je možné zjistit dvěma způsoby. Tím prvním je ruční způsob, kdy bychom do vypočtené rovnice postupně dosazovali hodnoty GNP v jednotlivých zemích a kontrolovali, zdali se vypočtená (říká se jí také predikovaná) kojenecká úmrtnost shoduje s příslušným statistickým údajem. Z tab. 10. 1 například víme, že na Slovensku bylo GNP na hlavu 3700 dolarů. Po dosazení tohoto údaje do regresní rovnice pro východoevropské země vypočítáme, že kojenecká úmrtnost (k.ú.) by měla být:

$$\text{k.ú. SR} = 19{,}63 - 0{,}0028*3700 = 9{,}27$$

Skutečná kojenecká úmrtnost přitom byla 8,8. Oba údaje jsou si docela blízké, což nás ale nemůže překvapit, neboť když se podíváme na obrázek 11.5, tak uvidíme, že Slovensko se velmi těsně přimyká regresní přímce. Udělejme stejný pokus pro Albánii, která, jak ukazuje obr. 11.5, je od přímky poměrně dosti vzdálena.

$$\text{k.ú. Albánie} = 19{,}63 - 0{,}0028*810 = 17{,}36$$

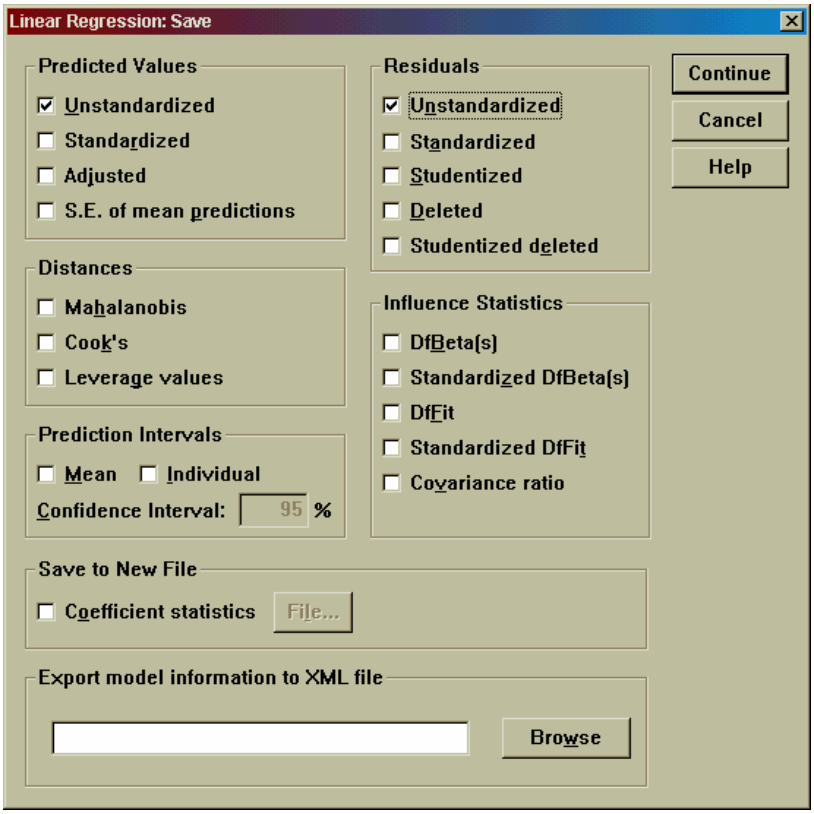
Skutečná kojenecká úmrtnost byla v Albánii 22. Zde je tedy již rozdíl větší.

Program SPSS umožňuje, abychom tento způsob kontroly nemuseli provádět ručně. Vypočítá za nás (podle příslušné regresní rovnice, kterou si pamatuje) predikované hodnoty kojenecké úmrtnosti pro všechny východoevropské země a tyto hodnoty umístí jako novou proměnnou na konec matice pod názvem *PRE_1*. Dále vypočítá tzv. rezidua neboli rozdíl mezi skutečnou hodnotou kojenecké úmrtnosti a její predikovanou hodnotou. Ty umístí jako další proměnnou do matice pod názvem RES_1. Procedura k celé operaci je následující:

V našem případě musíme nastavit analýzu pro podsoubor východoevropských zemí, neboť jenom pro ty má, jak jsme viděli, regresní analýza význam. Pokud jsme ještě neodstranili nastavení ve *Split file* požadavek na výpočty pro podsoubor západních a východních zemí, učiníme tak nyní. Poté si v nastavíme v *Data – Select cases* výpočty pouze pro východoevropské země. Pak:

Analyze – Regression – Linear – (vložíme jména nezávisle a závisle proměnné) – *Save* – (ve sloupečku Predicted Values zaškrtneme políčko Unstandardized a ve sloupečku Residuals zaškrtneme políčko Unstandardized – viz obr. 10.6) – *Continue – OK*

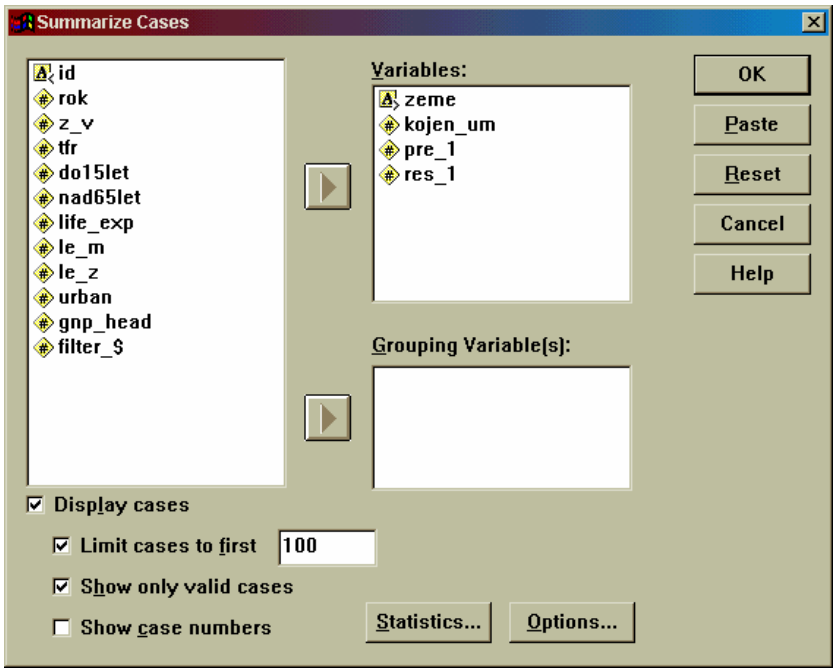
Obr. 10.6: Dialogové okno pro zpětný výpočet hodnot závisle proměnné, pro výpočet reziduí a pro jejich uložení jako nových proměnn;ých



Tím jsme provedli příslušné výpočty a v matici dat se objevily nové dva sloupce nazvané PRE_1 a RES_1. Jejich hodnoty si nyní zobrazíme do tabulky. Učiníme tak následovně (viz obr. 10.7):

Analyze – Reports – Case Summaries (do políčka Variables vložíme jméno identifikující proměnné ”zeme” , dále proměnnou ”kojen_um” a nově uložené proměnné ”pre_1” a ”res_1”) – *OK*

Obr. 10.7: Dialogové okno pro zobrazení hodnot nezávisle a závisle proměnné, rekonstruovaných hodnot závisle proměnné a reziduí



Výsledek uvádí tab. 10.8.

Tab. 10.8: Skutečné hodnoty kojenecké úmrtnosti, jejich vypočtené hodnoty a rezidua

Case Summaries[a]

	ZEME	KOJEN_UM	PRE_1 Unstandardized Predicted Value	RES_1 Unstandardized Residual
1	Albanie	22,00	17,36	4,64
2	Belorusko	11,00	13,54	-2,54
3	Bulharsko	14,40	16,22	-1,82
4	Cesko	4,60	5,33	-,73
5	Estonsko	9,00	10,24	-1,24
6	Chorvatsko	8,20	6,72	1,48
7	Litevsko	9,00	12,53	-3,53
8	Lotyssko	11,00	12,86	-1,86
9	Madarsko	8,90	7,02	1,88
10	Moldavie	18,00	18,57	-,57
11	Polsko	9,00	8,70	,30
12	Rumunsko	20,50	15,83	4,67
13	Rusko	17,00	13,31	3,69
14	Slovensko	8,80	9,29	-,49
15	Ukrajina	13,00	16,89	-3,89

a. Limited to first 100 cases.

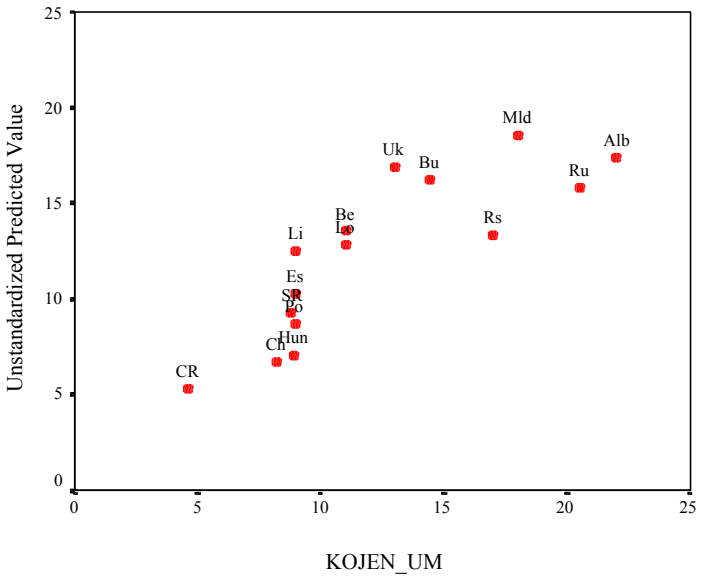
Srovnáním sloupce ”KOJEN_UM“ se sloupcem ”PRE_1” můžeme zjistit, jak se vypočtené hodnoty (PRE_1) odlišují od pozorovaných (skutečných).[8] Nejjednodušším způsobem ale, jak tyto rozdíly zjistit, je podívat se do sloupce ”RES_1”, v němž jsou uvedeny rozdíly mezi skutečnou a vypočtenou hodnotou. Kladná hodnota rezidua znamená, že vypočtená (predikovaná) hodnota je vyšší, než skutečná, zatímco záporná hodnota rezidua naznačuje, že vypočtená hodnota je nižší než skutečná.

[8] Náš ruční výpočet pro Slovensko přinesl údaj predikované kojenecké úmrtnosti 9,27. Rozdíl byl způsoben naším zaokrouhlením parametrů v rovnici.

Z tabulky 10.8 je patrné, že v některých případech (zemích), se predikované i skutečné hodnoty příliš neodlišují (např. v případě Česka, Moldávie či Polska), jindy je rozdíl poměrně značný (např. pro Ukrajinu, Rusko, Rumunsko či Albánii). Tento velký rozdíl naznačuje, že uvedené země se „chovají" – z hlediska vztahu mezi ekonomickou vyspělostí a kojeneckou úmrtností jinak, než by naznačoval model.

Predikovaných hodnot můžeme ještě využít pro kontrolu, zdali je regresní model adekvátní reálným datům. Je zřejmé, že pokud vyneseme do grafu hodnoty kojenecké úmrtnosti (závisle proměnné), kterou umístíme na osu *x* a na osu *y* hodnoty predikované, měly být by v případě, že model data dobře uchopil, body grafu na přímce. Jak ukazuje obr. 10.8, body grafu nevytvářejí zcela jasnou přímku, ale nejsou od ní daleko.

Obr. 10. 8: Graf predikovaných hodnot (pre_1) oproti hodnotám závisle proměnné

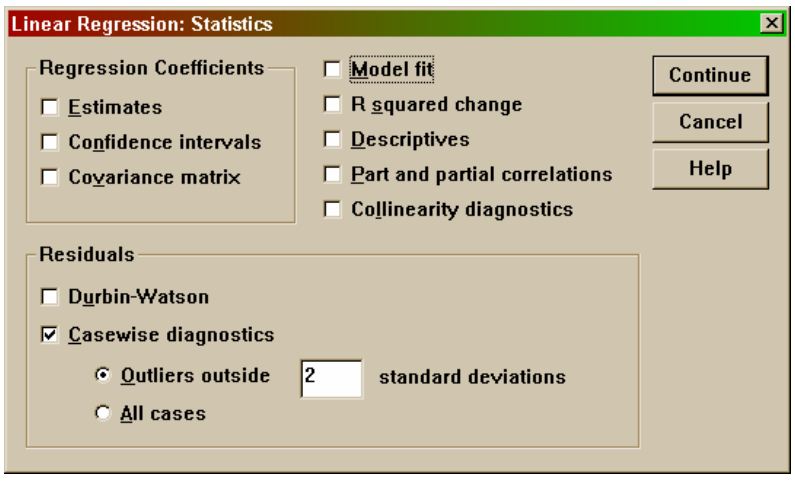


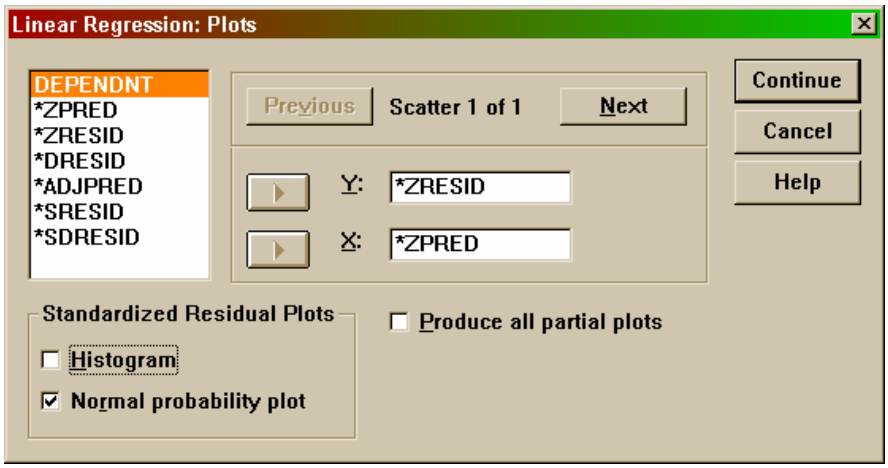
Pozn. Upravte měřítka na obou osách tak, aby byla stejně dlouhá.

Jiný způsob, jak testovat adekvátnost modelu, je prostřednictvím reziduí. Pokud je model pro data adekvátní, měla by být rezidua normálně rozložena. To testuje tzv. *Normal probability plot* neboli graf pravděpodobnosti normality. Získáme ho s pomocí zadání:

Analyze – Regression – Linear – Statistics – Casewise diagnostics *(+ viz obr. 10.9)* – Plots – Normal probabilty plot *(+ viz obr. 10.9)*

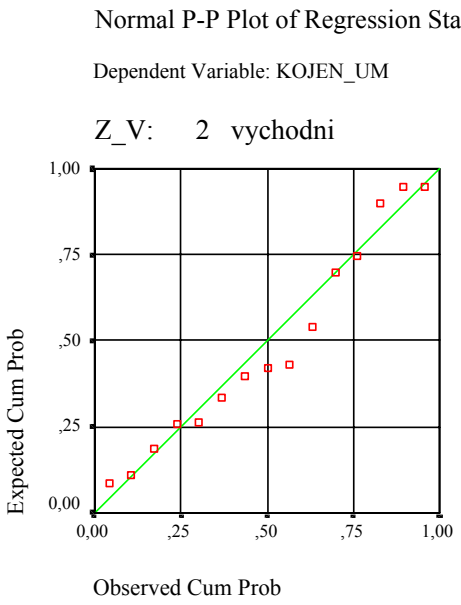
Obr. 10. 9: Zadání pro analýzu reziduí





Výsledkem je graf na obr. 10. 10. I zde by měly jednotlivé případy tvořit přímku. Naše data ideální přímku netvoří, nicméně nejsou od přímky příliš vzdálena. Což nevyvrací naše přesvědčení, že regresní lineární model je ve východoevropských zemích vhodným modelem pro vztah mezi GNP na hlavu a kojeneckou úmrtností.

Obr. 10. 10: Graf pravděpodobnosti normality rozložení reziduí



Závěrečné shrnutí

Lineární regrese umí určit vztah mezi dvěma proměnnými prostřednictvím regresní rovnice. Pokud je tato rovnice dobrým modelem vztahu, dokáže pro nové případy předvídat hodnoty závisle proměnné (y).

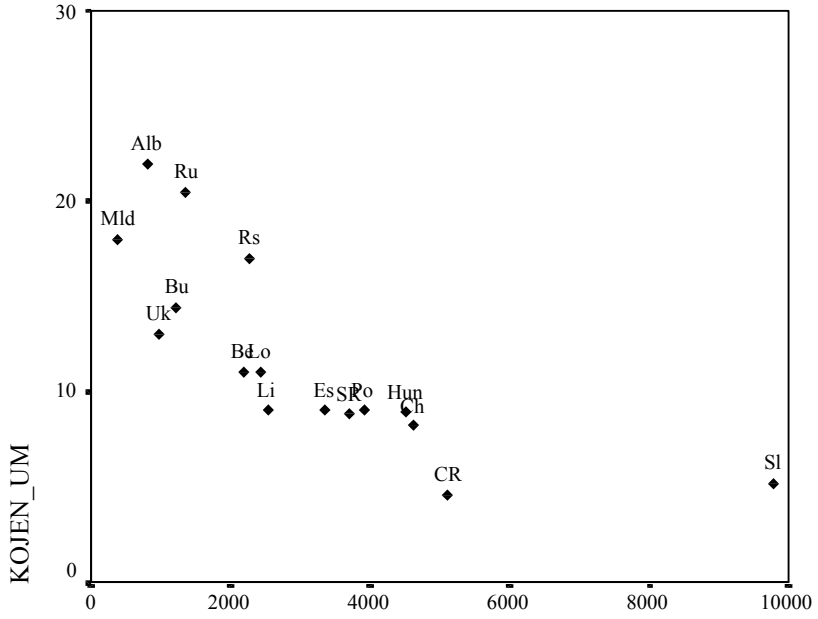
Výsledky regresní analýzy mohou být znehodnoceny:

1. nelinearitou vztahu (potom je možné použít nelineární regresi, ale to už přesahuje náplň našeho kurzu)
2. odlehlými hodnotami (outliers) (řešením může být vyřazení těchto pozorování nebo vhodná transformace proměnné)
3. přítomností subpopulací (když je soubor velmi heterogenní) - řešením je výpočet regrese zvlášť v dílčích subpopulacích nebo užití sofistikovaných víceúrovňových modelů (ty však již přesahují rámec tohoto kurzu)

V našem příkladě, kdy jsme hledali vztah mezi GNP a kojeneckou úmrtností, jsme narazili na problém druhého a třetího typu. Subpopulace jsme odhalili dvě: západní a východní Evropu, outlier bylo Slovin-

sko (Sl) – viz obr. 10.11, který ukazuje odlehlost Slovinska mezi východoevropskými zeměmi. To byl také hlavní důvod, proč jsme Slovinsko z analytických důvodů nezařadili mezi země východoevropské.

Obr. 10.11:



Regresní analýza je mocným statistickým nástrojem. V jejím užití však musíme být obezřetnými a pečlivě kontrolovat, zdali je předpoklad linearity naplněn. Doplňme, že jsme si ukázali nejjednodušší případ regrese s jednou nezávislou proměnnou (jednoduchá regrese). V praxi však konstruujeme zpravidla modely s více nezávislými proměnnými. Tato metoda je náplní kurzu v magisterském studiu.

V praxi se také setkáváme s tím, že závislá proměnná není intervalová, ale ordinální nebo nominální. Pro tyto případy existují různé varianty tzv. logistické regrese, s nimiž je opět možné se seznámit v magisterském studiu. Nedočkavé studenty můžeme odkázat například na učebnici Hendla, druhý a třetí díl učebnic napsaných kolektivem pod vedením Hebáka a knihy Militkého a Melouna. Samozřejmě existuje nepřeberné množství zahraniční literatury, nicméně její cenová dostupnost je značně nižší. Vyučující však zájemcům může samozřejmě vhodné tituly (dostupné v knihovně fakulty) doporučit.

Literatura:

Field, A. 2000. *Discovering Statistics using SPSS for Windows*. SAGE, London (v knihovně FSS).

Hebák, Hustopecký, Malá. 2005: Vícerozměrné statistické metody (2), Informatorium.

Hebák a kol. 2005: Vícerozměrné statistické metody (3), Informatorium.

Hendl J. 2004. Přehled statistických metod zpracování dat. Praha: Portál

Vaus D. A. de. 2002. *Analyzing Social Science Data*. London: Sage, str. 368-373. (v knihovně FSS).

Meloun, M., Militký, J.: *Statistické zpracování experimentálních dat*. Academia, Praha 2004.

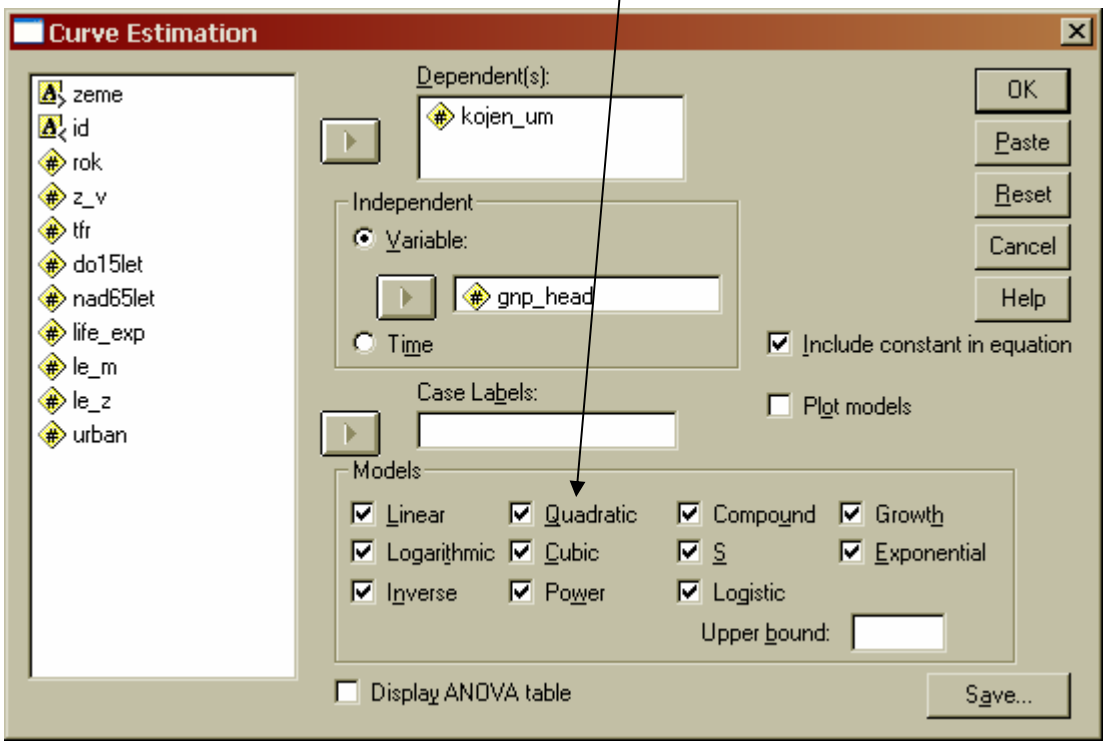
Doplněk: Co dělat, když data není vhodné proložit přímkou?

Odhad regresní funkce*

V situaci, kdy data naznačují, že lineární regrese není tím pravým postupem, je možné se pokusit o nalezení modelu nějaké křivky, která by nejlépe odpovídala vztahu mezi nezávisle a závisle proměnnou. V SPSS je to procedura *Curve estimation (Analyze – Regression – Curve estimation)*.

Praktický postup, jak na to:

1. Zvolte odhad všech možných křivek, které SPSS nabízí:



Syntax:

```
* Curve Estimation.
TSET NEWVAR=NONE .
CURVEFIT /VARIABLES=kojen_um  WITH gnp_head
  /CONSTANT
  /MODEL=LINEAR LOGARITHMIC INVERSE QUADRATIC CUBIC COMPOUND POWER S GROWTH
EXPONENTIAL LGSTIC
  /PLOT NONE.
```

2. Po výpočtu prohlédnu výstup a hledám funkci, která má největší F a nejnižší signifikanci. V našem příkladadu to jasně funkce POWER.

* Děkuji kolegovi Janu Spoustovi za konzultaci při vytváření tohoto návodu.

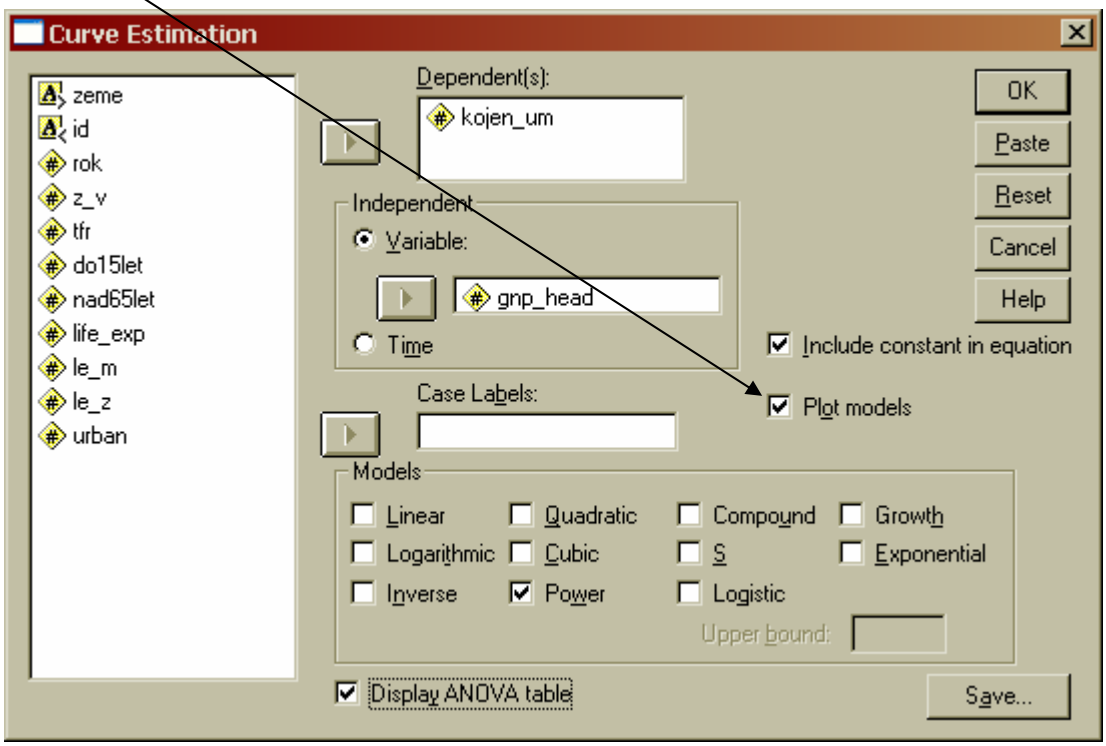
Model Summary and Parameter Estimates

Dependent Variable: kojen_um kojenecká úmrtnost

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	,520	33,527	1	31	,000	12,470	,000		
Logarithmic	,778	108,415	1	31	,000	38,582	-3,392		
Inverse	,609	48,183	1	31	,000	5,863	7627,957		
Quadratic	,662	29,404	2	30	,000	14,572	-,001	1,68E-008	
Cubic	,757	30,106	3	29	,000	17,132	-,002	9,186E-008	E-012
Compound	,649	57,272	1	31	,000	11,731	1,000		
Power	,834	156,246	1	31	,000	200,518	-,373		
S	,541	36,603	1	31	,000	1,722	764,761		
Growth	,649	57,272	1	31	,000	2,462	-3,6E-005		
Exponential	,649	57,272	1	31	,000	11,731	-3,6E-005		
Logistic	,649	57,272	1	31	,000	,085	1,000		

The independent variable is gnp_head GNP na hlavu v US $ (1998).

Provedu ještě jednou analýzu odhadu křivky metodou POWER s tím, že si nechám vytvořit i graf:



Výsledek:

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,913	,834	,829	,222

The independent variable is gnp_head GNP na hlavu v US $ (1998).

Koeficient determinace (R^2) je 0,83, což je výborný výsledek.

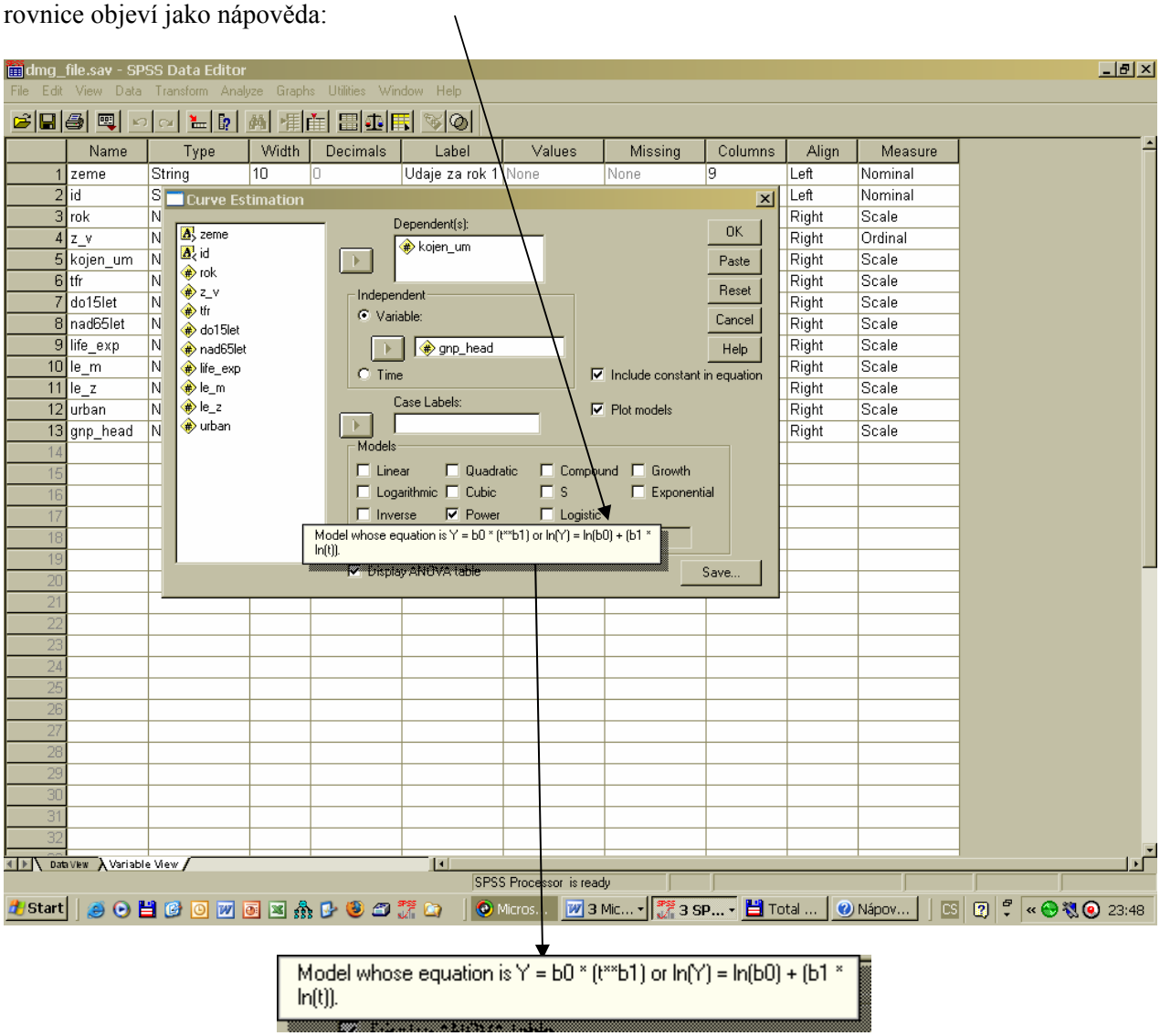
Model Summary and Parameter Estimates

Dependent Variable: kojen_um kojenecká úmrtnost

	Model Summary					Parameter Estimates	
Equation	R Square	F	df1	df2	Sig.	Constant	b1
Power	,834	156,246	1	31	,000	200,518	-,373

The independent variable is gnp_head GNP na hlavu v US $ (1998).

Abych mohl napsat příslušnou rovnici, v níž vepíšu vypočtené parametry, musím zjistit, jak rovnice pro Power vypadá. Po kliknutí pravým tlačítkem myši v dialogovém okně na nápis Power, se mně příslušná rovnice objeví jako nápověda:



Model whose equation is Y = b0 * (t**b1) or ln(Y) = ln(b0) + (b1 * ln(t)).

Rovnice má tvar: $y = b_0 * t^{b_1}$, kde $\mathbf{b_0}$ je konstanta, **t** je nezávisle proměnná (GNP na hlavu) a $^{b}{}_{1}$ je vypočtený regresní koeficient. Po dosazení příslušných parametrů z tabulky vypadá rovnice následovně:

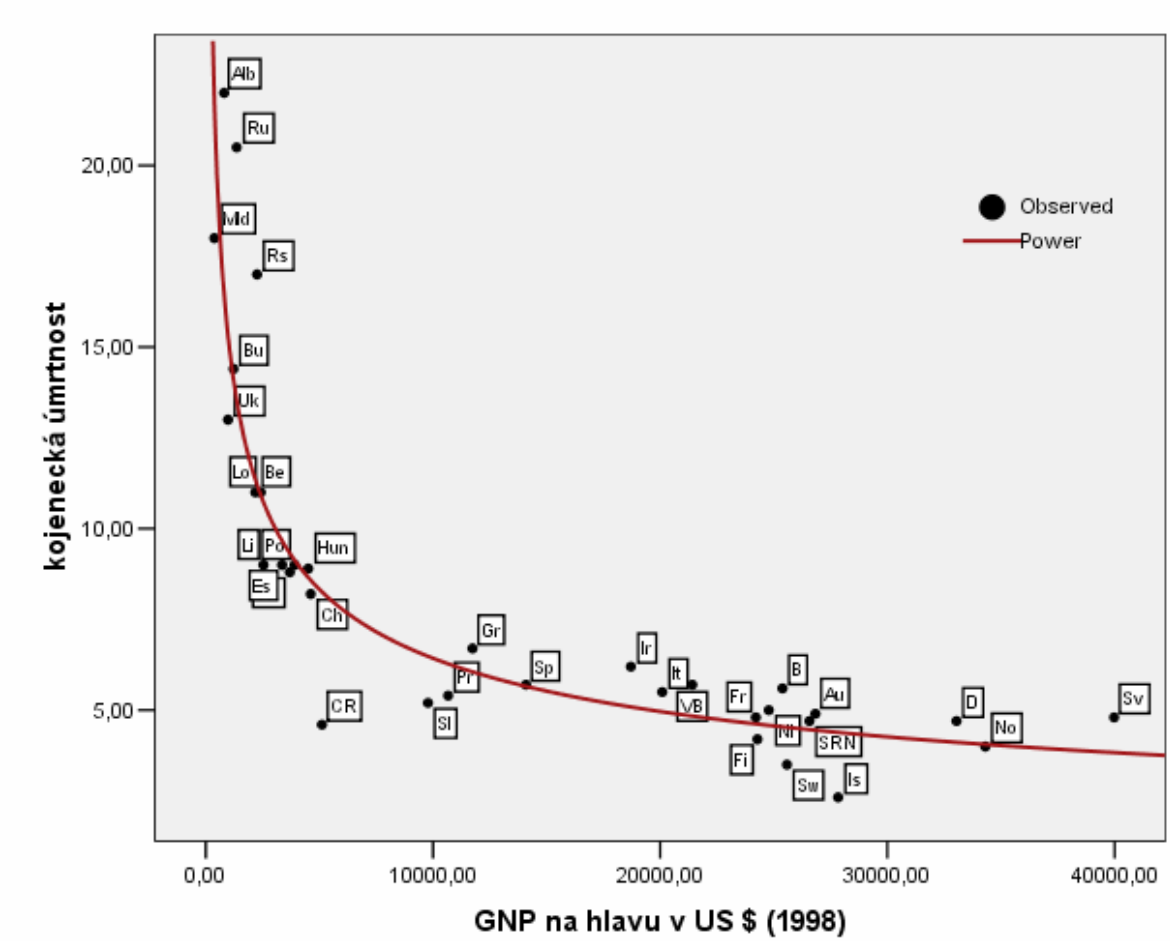
*kojenecká úmrtnost = 200,5 * (GDP na hlavu v dolarech)* $^{-0,37}$

Pohled na graf (viz níže) nás utvrzuje v přesvědčení, že jsme zvolili dobrou funkci. Pro kontrolu vypočítejme modelové hodnoty kojenecké úmrtnosti pro Belgii, které měla GNP = 25 380 dolarů, a Běloruska, jež mělo GNP 2 180 dolarů.

*kojenecká úmrtnost Belgie= 200,5 * 25 380* $^{-0,37}$ = 4,7

*kojenecká úmrtnost Běloruska= 200,5 * 2 180* $^{-0,37}$ = 11,7

Skutečná úmrtnost Belgie byla 5,6 a Běloruska 11,0. Rezidua tedy nejsou skutečně velká.



Závěr: Daný model můžeme klidně přijmout.