EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Analýza dat v programu SPSS pro začátečníky a mírně pokročilé

III.

Ondřej Brom

2019

Strategický projekt UTB ve Zlíně, reg. č. CZ.02.2.69/0.0/0.0/16_015/0002204

Informace o autorech:

Ing. Ondřej Brom
ACREA CR, spol. s r.o.
obrom@acrea.cz

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OBSAH

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

ÚVOD

Text podává informace o základních procedurách, které budou na kurzu probrány, a to o korelační analýze a lineární regresi. Slouží jako podpůrný materiál pro účastníky kurzu. Důraz je kladen na popis jednotlivých metod v IBM SPSS Statistics.

Text vychází z obecných statistických postupů, které jsou prezentovány v učebních textech (Schroeder, Sjoquist & Stephan, 1986; Cohen et al, 2002; Chen, Popovich, 2002; Hebák, Hustopecký & Malá, 2005) a také z lektorských zkušeností autora.

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

1 KORELACE

Procedura Bivariate Correlations umožňuje vypočítat Pearsonův lineární korelační koeficient a dva typy neparametrických koeficientů korelace – Kendallovo tau-b a Spearmanovo ρ. K dispozici je rovněž test nulovosti korelačních koeficientů. Kromě toho lze zobrazit kovarianční matici vstupních proměnných a některé další statistiky.

Korelace charakterizuje vzájemný vztah dvou číselných nebo ordinálních proměnných. Tento vztah vyjadřujeme korelačním koeficientem. Sledujeme-li vztahy většího počtu proměnných současně, tabelujeme korelační koeficienty do tzv. korelační matice – čtvercového schématu zobrazujícího hodnoty koeficientů pro všechny dvojice vstupních proměnných. Vzhledem k tomu, že korelační koeficient je symetrickou mírou (nezáleží na pořadí proměnných), je rovněž korelační matice symetrická.

Pearsonův lineární korelační koeficient vyjadřuje míru lineární závislosti dvou číselných proměnných. Před jeho výpočtem je třeba ověřit, zda data neobsahují odlehlá pozorování, která by mohla získané závěry zkreslit. Tento typ koeficientu není vhodný tam, kde mezi proměnnými existuje jiný typ závislosti než lineární.

Pearsonův lineární korelační koeficient nabývá hodnot z intervalu [-1,1]. Je-li jeho absolutní hodnota rovna jedné, data leží přesně na přímce. Korelační koeficient roven jedné charakterizuje přímou úměrnost (přímka je rostoucí), korelační koeficient roven mínus jedné odpovídá nepřímé úměrnosti (přímka je klesající). Při zkoumání reálných dat se však s těmito hraničními hodnotami korelačního koeficientu téměř nesetkáváme (data neleží přesně na přímce), ale zajímá nás, do jaké míry se přímce přibližují. Čím blíže jedné je absolutní hodnota koeficientu, tím lépe přímka data vystihuje a tím silnější lineární závislost mezi proměnnými existuje. Jestliže neexistuje lineární vztah mezi zkoumanými proměnnými, je jejich korelační koeficient roven nule.

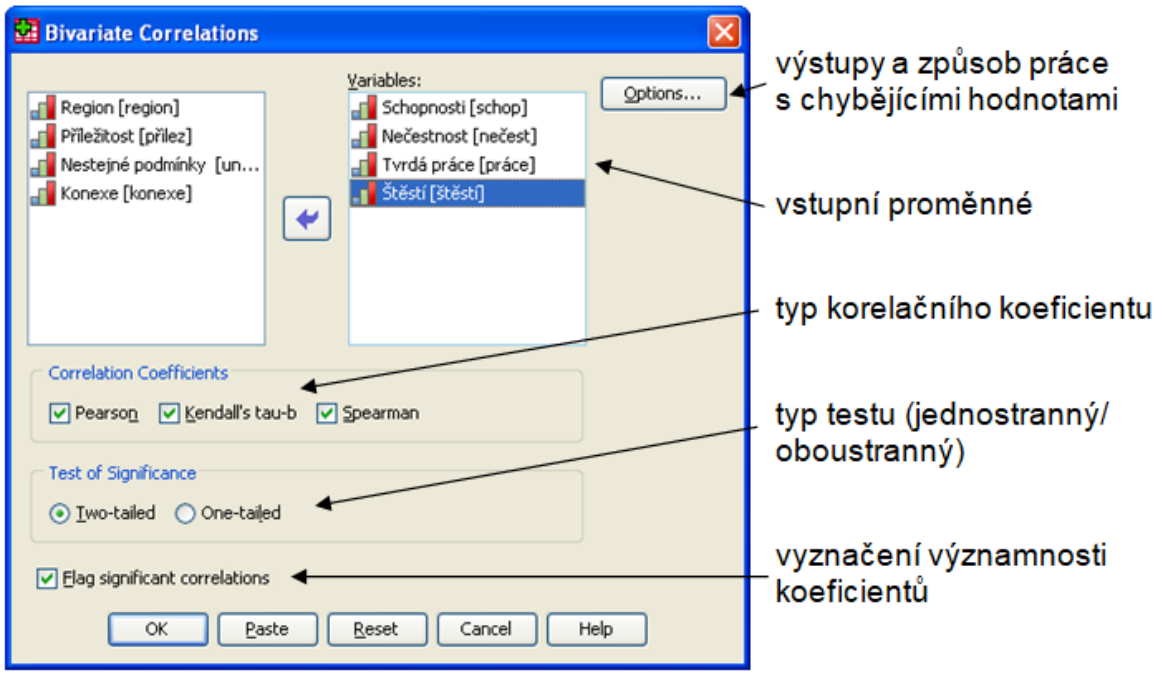
V některých případech je vhodnější užít místo Pearsonova lineárního korelačního koeficientu neparametrické korelační koeficienty – například tam, kde je rozložení některé z proměnných výrazně šikmé, data obsahují vzdálená pozorování nebo máme k dispozici pouze pořadí hodnot. SPSS nabízí dva typy neparametrických korelačních koeficientů: **Spearmanovo ρ** a **Kendallovo tau-b**. Tyto koeficienty neměří lineární závislost, ale vyja-

dřují, do jaké míry si odpovídají pořadí hodnot dvou proměnných. Z tohoto důvodu jsou také méně citlivé na odlehlá pozorování. Oba koeficienty nabývají hodnot z intervalu [-1,1]. Čím je absolutní hodnota koeficientu větší, tím je popisovaný vztah mezi proměnnými silnější.

Velmi často provádíme také test nulovosti korelačního koeficientu. Nulová hypotéza testu je formulována tak, že korelační koeficient je na celém základním souboru roven nule. Na základě získané signifikance potom rozhodujeme o zamítnutí/nezamítnutí této hypotézy.

1.1 Volání procedury v SPSS

Nastavení dialogu

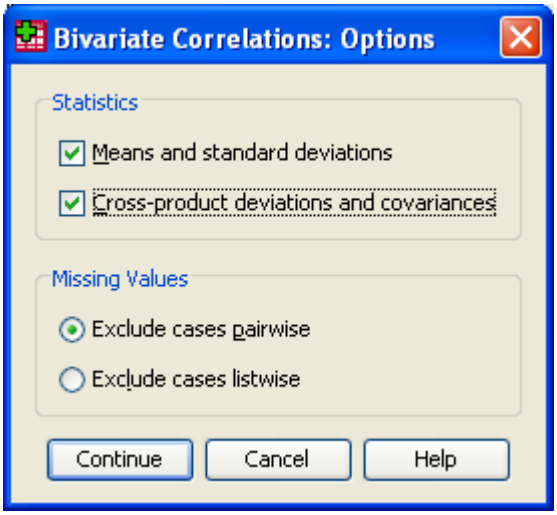


Obrázek 1: Nastavení dialogu procedury Bivariate Correlations

- Do pole *Variables* přeneseme nejméně dvě proměnné, jejichž vzájemné vztahy sledujeme. Zadáme-li proměnných více, získáme korelační matici s hodnotami koeficientů pro všechny dvojice proměnných.

- V části *Correlation Coefficients* označíme požadované typy korelačního koeficientu (Pearsonův lineární korelační koeficient, Kendallovo tau-b, Spearmanovo ρ).
- Procedura rovněž automaticky provádí test nulovosti korelačních koeficientů. V části *Test of Significance* volíme mezi oboustrannou a jednostrannou alternativní hypotézou.
- Zaškrtávací políčko *Flag significant correlations* určuje, zda mají být v korelační matici hvězdičkami vyznačeny koeficienty, které jsou na základě předchozího testu významně různé od nuly.

1.1.1 Tlačítko Options



Obrázek 2: Tlačítko Options v proceduře Bivariate Correlations

Tlačítkem *Options* volíme další požadované výstupy a způsob práce s chybějícími hodnotami.

Statistics (Statistiky)

- *Means and standard deviations* – průměry a směrodatné odchylky vstupních proměnných.

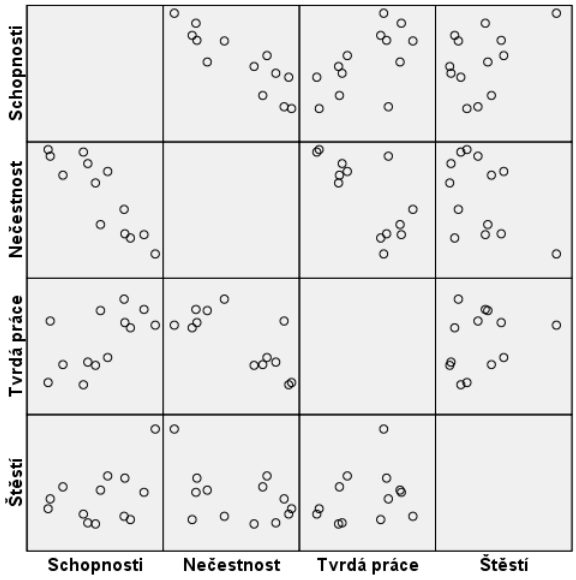
- *Cross-product deviations and covariances* – kovarianční matice vstupních proměnných a součiny odchylek od průměrů (tj. hodnoty součtů, které tvoří čitatel vzorce pro výpočet Pearsonova korelačního koeficientu).

Missing Values (Chybějící hodnoty)

- *Exclude cases pairwise* – případy s chybějícími hodnotami u některé ze vstupních proměnných jsou vynechány z výpočtu korelačních koeficientů pouze tam, kde je to nezbytně nutné.
- *Exclude cases listwise* – případy s chybějící hodnotami u některé ze vstupních proměnných jsou vyloučeny z výpočtu všech korelačních koeficientů.

1.2 Výstupy

Datový soubor obsahuje vybrané informace z mezinárodního výzkumného projektu Sociální spravedlnost, realizovaného v roce 1991. Ze třinácti zemí, které se výzkumu účastnily, máme k dispozici souhrnné údaje – průměrná hodnocení důležitosti vybraných položek pro získání bohatství. Naším cílem je podrobněji prozkoumat vzájemný vztah těchto proměnných.



Obrázek 3: Grafické znázornění vztahů proměnných

Před výpočtem korelačního koeficientu je užitečné nejprve zobrazit vztahy mezi proměnnými graficky. V případě, že je vstupních proměnných více můžeme využít maticový bodový graf (*Graphs, Legacy Dialogs, Scatter/Dot*), který se skládá z bodových grafů pro všechny dvojice zadaných proměnných.

Z grafu ověříme, zda se v datech nevyskytují odlehlá pozorování nebo jiný typ problému, který by mohl další výsledky zkreslit. Zároveň si vytvoříme orientační představu o vztazích mezi proměnnými. Zde je například patrný lineární vztah (protiběžnost) mezi proměnnými Schopnosti a Nečestnost, zatímco proměnná Štěstí se nezdá být s ostatními příliš svázaná.

Vzhledem k tomu, že v datech nemáme žádná výrazně odlehlá pozorování, můžeme užít Pearsonův lineární korelační koeficient.

Následující výstupy byly získány pomocí nabídky *Bivariate Correlations*.

1.2.1 Popisné statistiky

Descriptive Statistics

	Mean	Std. Deviation	N
Schopnosti	3.58	.20	13
Nečestnost	3.59	.48	13
Tvrdá práce	3.28	.38	13
Štěstí	3.11	.24	13

Tabulka 1: Tabulka popisných statistik z procedury Bivariate Correlations

Tabulka *Descriptive Statistics* poskytuje přehled o průměrech, směrodatných odchylkách a počtech platných pozorování všech vstupních proměnných.

Zde se například jenom nepatrně liší průměrné hodnoty proměnných Schopnosti a Nečestnost, zatímco mezi jejich směrodatnými odchylkami je výraznější rozdíl – názory na vliv nečetného chování pro získání bohatství se v jednotlivých zemích více diferencují než názory na vliv schopností, ačkoliv se jejich hodnocení v průměru pohybuje okolo stejné hodnoty.

1.2.2 Korelační matice – Pearsonův lineární korelační koeficient

Correlations

		Schopnosti	Nečestnost	Tvrdá práce	Štěstí
Schopnosti	Pearson Correlation	1	-.879**	.591*	.414
	Sig. (2-tailed)		.000	.033	.160
	N	13	13	13	13
Nečestnost	Pearson Correlation	-.879**	1	-.737**	-.476
	Sig. (2-tailed)	.000		.004	.100
	N	13	13	13	13
Tvrdá práce	Pearson Correlation	.591*	-.737**	1	.268
	Sig. (2-tailed)	.033	.004		.375
	N	13	13	13	13
Štěstí	Pearson Correlation	.414	-.476	.268	1
	Sig. (2-tailed)	.160	.100	.375	
	N	13	13	13	13

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Tabulka 2: Tabulka korelací z procedury Bivariate Correlations

Korelační matice obsahuje pro každou dvojici vstupních proměnných hodnotu Pearsonova lineárního korelačního koeficientu (Pearson Correlation) a significance testu nulovosti tohoto koeficientu (Sig. (2-tailed)). V případě že byla pro práci s chybějícími hodnotami použita metoda pairwise, je v každé buňce tabulky uveden také počet případů, ze kterých byl korelační koeficient spočítán. Při použití metody listwise je počet případů zapsán pod tabulkou.

Korelační koeficienty významně odlišné od nuly jsou v tabulce označené hvězdičkami (jedna hvězdička odpovídá nenulovosti na hladině spolehlivosti 95 %, dvě hvězdičky hladině spolehlivosti 99 %).

Pro větší přehlednost upravíme tabulku pivotací tak, aby v horní vrstvě byly zobrazeny pouze hodnoty korelačních koeficientů. Takto získáme čtvercovou symetrickou matici s jedničkami na diagonále.

Tabulku můžeme ještě zpřehlednit skriptem Obarvení tabulky, který je volně k dispozici na stránkách www.acrea.cz. Skript podbarví jednotlivé buňky podle hodnot korelačního koeficientu. Záporným koeficientům odpovídá modrá barva, kladným červená. Barva je tím sytější, čím je absolutní hodnota koeficientu vyšší.

Correlations

Pearson Correlation

	Schopnosti	Nečestnost	Tvrdá práce	Štěstí
Schopnosti	1	-.879**	.591*	.414
Nečestnost	-.879**	1	-.737**	-.476
Tvrdá práce	.591*	-.737**	1	.268
Štěstí	.414	-.476	.268	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Tabulka 3: Obarvená tabulka korelací z procedury Bivariate Correlations

Z korelační matice vyplývá, že například mezi dvojicí proměnných Schopnosti a Nečestnost je silná negativní závislost. To znamená, že v zemích, kde mají schopnosti větší vliv na úspěch, má nečestné jednání vliv menší. Naopak proměnné Schopnosti a Tvrdá práce jsou korelovány kladně – v zemích, kde jsou pro získání bohatství důležité schopnosti je důležitější i tvrdá práce. Korelační koeficienty proměnné Štěstí s ostatními proměnnými jsou na 95% hladině spolehlivosti nevýznamné, lze tedy s 5% rizikem tvrdit, že mezi vlivem štěstí a ostatních uvedených faktorů není lineární závislost.

1.3 Kovarianční matice a součiny odchylek od průměrů

Correlations

		Schopnosti	Nečestnost	Tvrdá práce	Štěstí
Schopnosti	Pearson Correlation	1	-,879**	,591*	,414
	Sig. (2-tailed)		,000	,033	,160
	Sum of Squares and Cross-products	,461	-,998	,524	,235
	Covariance	,038	-,083	,044	,020
	N	13	13	13	13
Nečestnost	Pearson Correlation	-,879**	1	-,737**	-,476
	Sig. (2-tailed)	,000		,004	,100
	Sum of Squares and Cross-products	-,998	2,794	-1,608	-,665
	Covariance	-,083	,233	-,134	-,055
	N	13	13	13	13
Tvrdá práce	Pearson Correlation	,591*	-,737**	1	,268
	Sig. (2-tailed)	,033	,004		,375
	Sum of Squares and Cross-products	,524	-1,608	1,703	,293
	Covariance	,044	-,134	,142	,024
	N	13	13	13	13
Štěstí	Pearson Correlation	,414	-,476	,268	1
	Sig. (2-tailed)	,160	,100	,375	
	Sum of Squares and Cross-products	,235	-,665	,293	,697
	Covariance	,020	-,055	,024	,058
	N	13	13	13	13

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Tabulka 4: Kovarianční matice z procedury Bivariate Correlations

Tabulku můžeme dále doplnit o kovarianční matici (Covariance) a součiny odchylek od průměrů (Sum of Squres and Cross-products), tj. hodnoty součtů z čitatele vzorce pro výpočet Pearsonova korelačního koeficientu.

1.4 Neparametrické korelace

Correlations

			Schopnosti	Nečestnost	Tvrdá práce	Štěstí
Kendall's tau_b	Schopnosti	Correlation Coefficient	1,000	-,795**	,410	,179
		Sig. (2-tailed)	.	,000	,051	,393
		N	13	13	13	13
	Nečestnost	Correlation Coefficient	-,795**	1,000	-,359	-,179
		Sig. (2-tailed)	,000	.	,088	,393
		N	13	13	13	13
	Tvrdá práce	Correlation Coefficient	,410	-,359	1,000	,154
		Sig. (2-tailed)	,051	,088	.	,464
		N	13	13	13	13
	Štěstí	Correlation Coefficient	,179	-,179	,154	1,000
		Sig. (2-tailed)	,393	,393	,464	.
		N	13	13	13	13
Spearman's rho	Schopnosti	Correlation Coefficient	1,000	-,912**	,555*	,286
		Sig. (2-tailed)	.	,000	,049	,344
		N	13	13	13	13
	Nečestnost	Correlation Coefficient	-,912**	1,000	-,577*	-,280
		Sig. (2-tailed)	,000	.	,039	,354
		N	13	13	13	13
	Tvrdá práce	Correlation Coefficient	,555*	-,577*	1,000	,253
		Sig. (2-tailed)	,049	,039	.	,405
		N	13	13	13	13
	Štěstí	Correlation Coefficient	,286	-,280	,253	1,000
		Sig. (2-tailed)	,344	,354	,405	.
		N	13	13	13	13

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Tabulka 5: Neparametrické korelace z procedury Bivariate Correlations

Matice neparametrických korelačních koeficientů mají obdobnou strukturu jako matice Pearsonova lineárního korelačního koeficientu, oba dva typy koeficientů jsou však uvedeny v jedné tabulce.

Hodnoty různých typů korelačních koeficientů není vhodné navzájem srovnávat, můžeme však porovnat jejich významnost. Zde vychází Kendallův koeficient pro proměnné Tvrdá práce a Schopnosti nebo Nečestnost na rozdíl od Pearsonova koeficientu nevýznamný. Spearmanův koeficient pro proměnné Nečestnost a Tvrdá práce je významný na 5% hladině významnosti, ale Pearsonův koeficient dokonce na 1% hladině. Tyto výsledky jsou v souladu s obecněji platným tvrzením, že při splnění předpokladů je vhodnější užít parametrickou variantu koeficientu (Pearsonův korelační koeficient), neboť s testováním nulovosti u neparametrických koeficientů je spojena větší pravděpodobnost chyby druhého druhu (nulovou hypotézu zamítáme méně často).

2 LINEÁRNÍ REGRESE

Regresní analýza umožňuje vyjádřit statistickou závislost zkoumané číselné proměnné na jedné nebo více nezávislých proměnných. Nezávislé proměnné by měly být rovněž číselné, jestliže však potřebujeme do analýzy zařadit také nominální proměnné (například region, náboženství apod.), je třeba převést je na indikátory nebo jiný typ kontrastů.

Model lineární regrese předpokládá, že mezi závislou proměnnou a nezávislými proměnnými existuje lineární vztah, a hledá co možná nejlepší vyjádření analyzované proměnné na základě lineární kombinace prediktorů.

Zkoumaná závislost má tedy tvar:

$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k + \varepsilon$, kde

Y … závislá proměnná

X_i … nezávislé proměnné

b_i … koeficienty regresní rovnice

ε … náhodná chyba

Cílem analýzy je nalézt koeficienty této rovnice tak, aby závislost co nejpřesněji vystihovala data.

Z geometrického hlediska se pokoušíme daty proložit přímku, rovinu nebo jinou lineární nadrovinu (podle počtu dimenzí problému). Kritérium pro rozhodnutí, která z možných přímek (rovin apod.) charakterizuje data nejlépe, obvykle vychází z tzv. metody nejmenších čtverců – požadujeme, aby součet druhých mocnin odchylek jednotlivých bodů od jejich předpovědi byl minimální.

Model vychází z těchto základních předpokladů:

- pozorování jsou mezi sebou navzájem nezávislá
- skutečné hodnoty a chyby jsou navzájem nezávislé
- rezidua mají normálního rozdělení s nulovou střední hodnotou a konstantním rozptylem

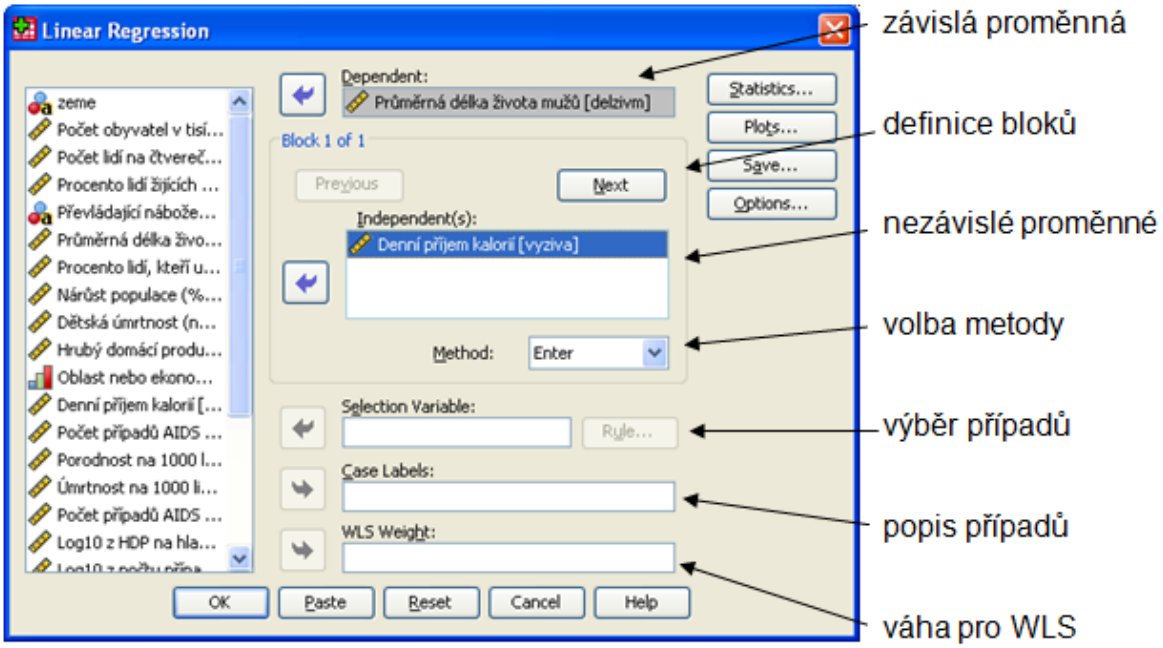
- matice X má plnou hodnost (tj. mezi nezávislými proměnnými není funkční lineární závislost)
- náhodné složky jsou navzájem nekorelované

Procedura nabízí široký výběr nástrojů pro ověření předpokladů, vyhodnocení kvality modelu, odhalování extrémních hodnot apod. K dispozici je celá řada statistik, testů i různých typů grafů od základních až po velmi speciální. Vybrané informace lze rovněž uložit do datové matice nebo celý model exportovat do formátu XML.

2.1 Volání procedury v IBM SPSS Statistics

Analyze → Regression → Linear

Nastavení dialogu

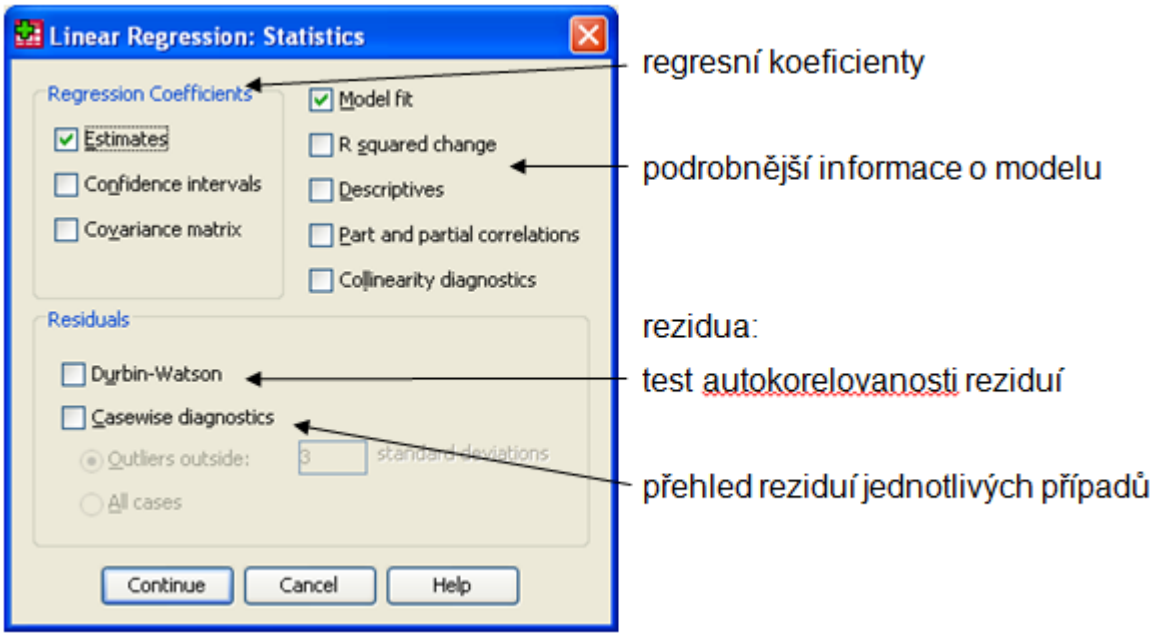


Obrázek 4: Nastavení dialogu procedury Linear Regression

- Do pole *Dependent* přeneseme závislou proměnnou.
- Do pole *Independent(s)* zadáme nezávislé proměnné.

- V rozbalovacím seznamu *Method* zvolíme některou z metod pro výběr prediktorů, které pomáhají zajistit, aby model obsahoval všechny důležité prediktory, ale nebyl přeurčen:
  - *Enter* – model se všemi zadanými nezávislými proměnnými.
  - *Forward* – v každém kroku se do modelu postupně přidá jedna proměnná, která model nejvíce zlepší. V okamžiku, kdy zlepšení již nepřekročí určitou hranici, proces končí.
  - *Backward* – z modelu se všemi zadanými prediktory se v každém kroku ubírá nejvíce nadbytečná proměnná tak dlouho, dokud zhoršení modelu nepřekročí stanovenou hranici.
  - *Stepwise* – jedná se o kombinaci metod Forward a Backward. Do modelu se postupně přidávají proměnné a současně se v každém kroku kontroluje, zda není možné odebrat některou z již zařazených proměnných jako nadbytečnou.
  - *Remove* – metoda pracuje s definovanými bloky proměnných (viz následující bod). Všechny proměnné bloku jsou vždy odebrány společně v jednom kroku.
- Pomocí tlačítek *Previous* a *Next* můžeme definovat bloky (skupiny) proměnných. Pro každý blok nastavujeme samostatně metodu výběru prediktorů.
- V poli *Selection Variable* lze zadat proměnnou, která určuje případy vstupující do modelu. To je užitečné především tehdy, když je třeba data rozdělit na dvě skupiny – na první z nich model vytvoříme a na druhé testujeme jeho kvalitu.
- V poli *Case Labels* definujeme popis jednotlivých případů.
- *WLS Weight* umožňuje zadat proměnnou určující váhu případu pro váženou metodu nejmenších čtverců. Aby bylo toto pole aktivní, je nutné mít k dispozici modul *Regression Models*.

2.1.1 Tlačítko Statistics

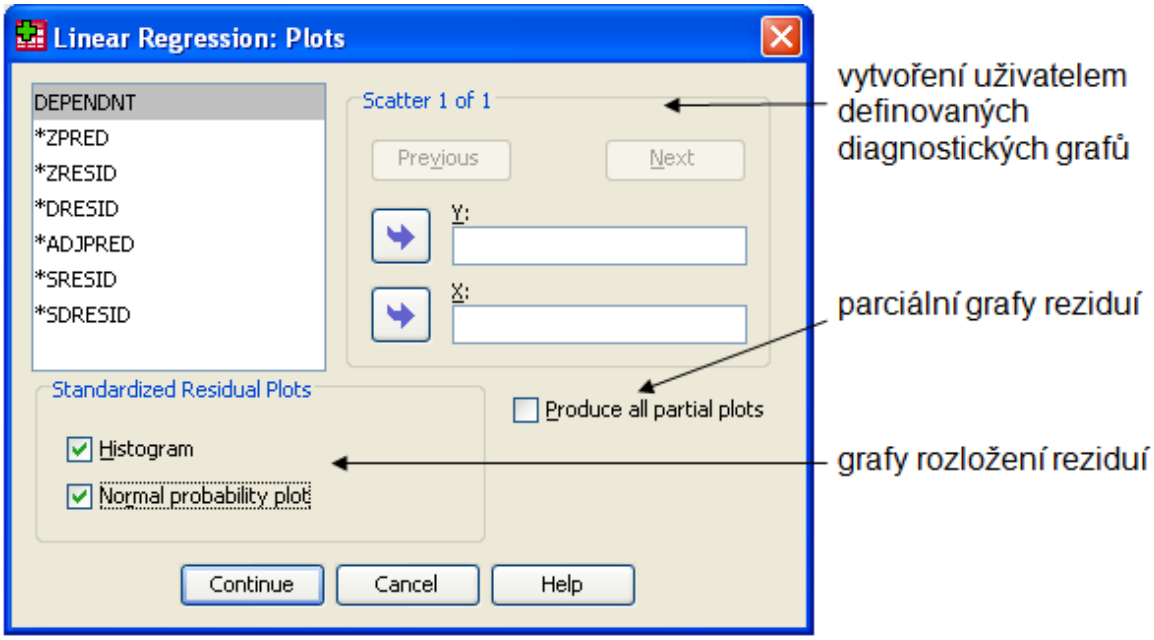


Obrázek 5: Tlačítko Statistics procedury Linear Regression

Tlačítkem *Statistics* nastavíme požadované tabulkové výstupy:

- V části *Regression Coefficients*: odhady regresních koeficientů včetně standardní chyby, standardizovaných koeficientů a testu významnosti (*Estimates*), intervaly spolehlivosti pro regresní koeficienty (*Confidence intervals*), korelační a kovarianční matice regresních koeficientů (*Covariance matrix*).
- Přehled základních informací o modelu včetně koeficientu determinace a tabulky ANOVA (*Model fit*), změny koeficientu determinace při přidání nebo odebrání nezávislé proměnné (*R squared change*), popisné statistiky a korelační matice prediktorů (*Descriptives*), Pearsonův lineární korelační koeficient, částečné a parciální korelace (*Part and partial correlations*), diagnostika kolinearity (*Collinearity diagnostics*).
- Podrobnější informace o reziduích (*Residuals*): výpočet Durbin-Watsonovy statistiky pro testování autokorelovanosti reziduí (*Durbin-Watson*) a diagnostika reziduí všech případů nebo jen případů, kde hodnota rezidua překročí zvolený násobek směrodatné odchylky (*Casewise diagnostics*).

2.1.2 Tlačítko Plots



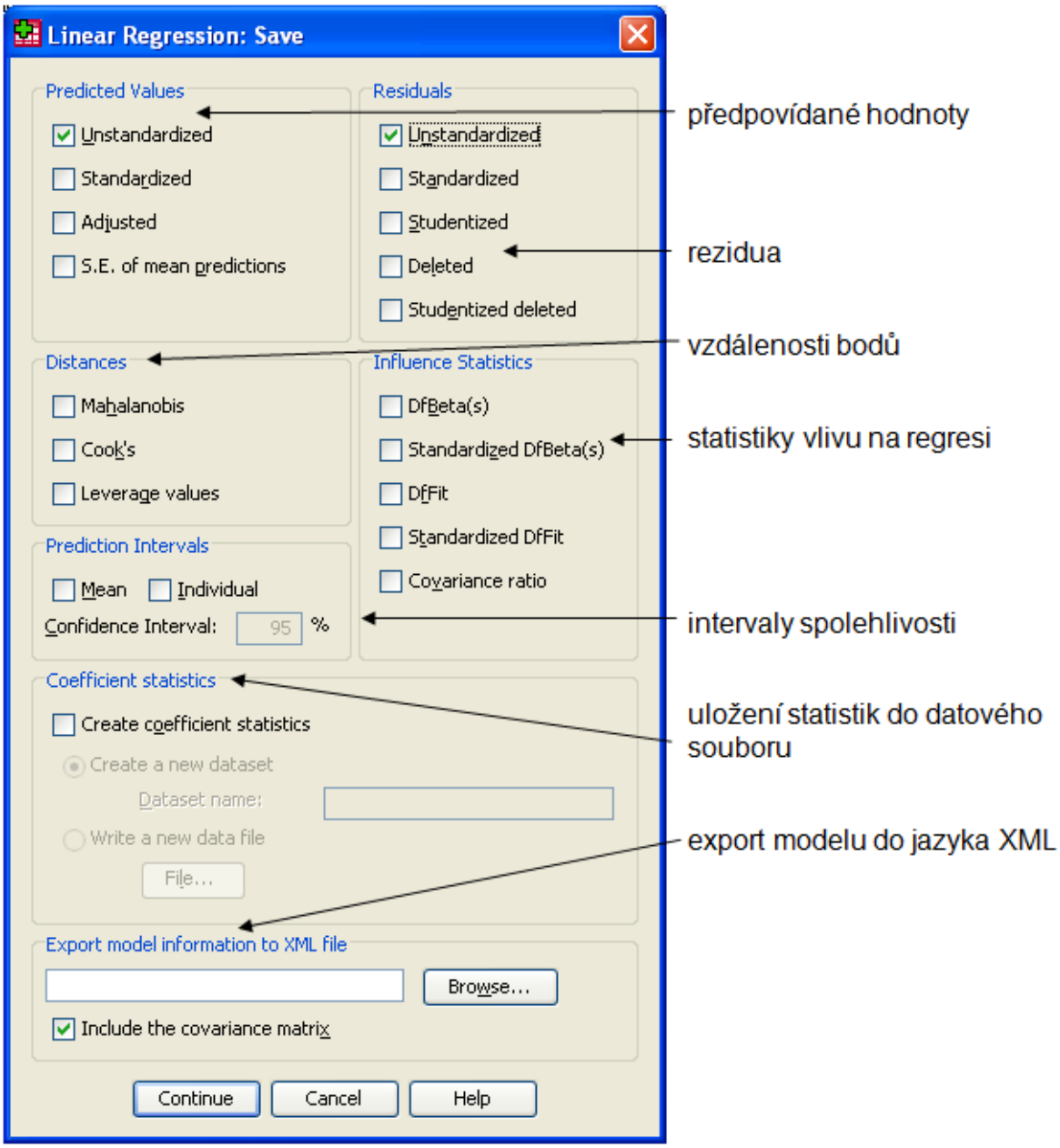
Obrázek 6: Tlačítko Plots procedury Linear Regression

Tlačítkem *Plots* zadáváme požadované typy grafů, které jsou určeny především pro diagnostiku reziduí.

- V části *Scatter 1 of 1* můžeme zadat bodový graf určením os X a Y. Ze seznamu v levé části okna zvolíme požadované charakteristiky a šipkami je přeneseme do vybraného políčka. K dispozici jsou tyto možnosti: závislá proměnná (*DEPENDENT*), standardizovaná předpovídaná hodnota (*ZPRED*), standardizovaná rezidua (*ZRESID*), vynechávaná rezidua (*DRESID*), adjustovaná předpovídaná hodnota (*ADJPRED*), studentizovaná rezidua (*SRESID*), studentizovaná vynechávaná rezidua (*SDRESID*) – podrobnější informace viz popis tlačítka *Save*.
- V poli *Standardized Residua Plots* rozhodujeme o zobrazení histogramu (*Histogram*) a grafu pro ověřování normality (*Normal probability plot*).

Zaškrtávacím políčkem *Produce all partial plots* volíme parciální grafy reziduí. (Vodorovná osa odpovídá vždy jednomu z prediktorů, svislá osa závislé proměnné. Do bodového grafu jsou proti sobě vynášeny hodnoty reziduí pro model, kdy je daná proměnná vysvětlována pomocí ostatních prediktorů. Graf tedy vyjadřuje vztah mezi závislou proměnnou a jednou z nezávislých proměnných očištěný od vlivu ostatních prediktorů).

2.1.3 Tlačítko Save



Obrázek 7: Tlačítko Save procedury Linear Regression

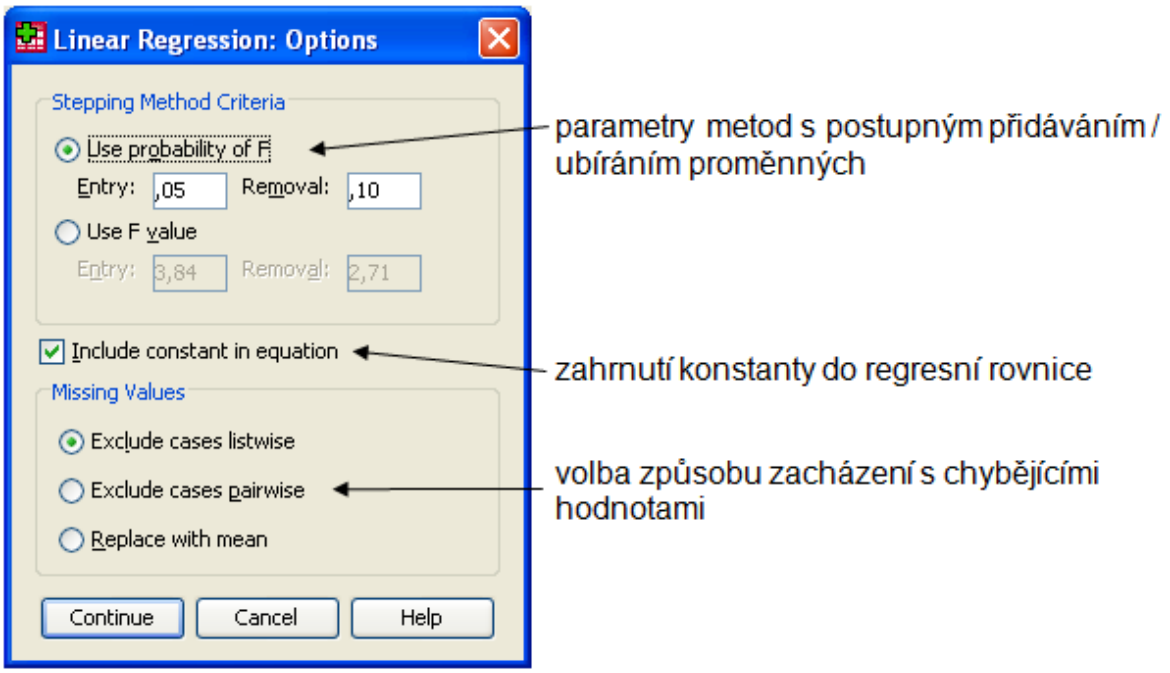
Tlačítkem *Save* volíme, které proměnné a další informace budou uloženy do datového souboru, a případně rozhodujeme o exportu celého modelu do jazyka XML.

- Předpovídané hodnoty (*Predicted Values*): nestandardizované (*Unstandardized*), standardizované na z-skóry (*Standardized*), adjustované, tj. předpovídané hodnoty získané při vyloučení daného případu z výpočtu regresních koeficientů (*Adjusted*) a standardní chyby předpovídaných hodnot (*S.E of mean predictions*).
- Vzdálenosti (*Distances*): *Mahalanobis* – vyjadřuje, jak se liší hodnoty nezávislých proměnných daného případu od průměru všech případů, *Cook's* – měří, jak se změní rezidua všech případů při vyloučení daného případu, *Leverage values* – charakterizují vliv konkrétního případu na průběh regresní přímky.
- Intervaly a pásy spolehlivosti (*Prediction Intervals*): pás spolehlivosti pro regresní přímku (*Mean*) a intervaly spolehlivosti pro jednotlivá pozorování (*Individual*), volba hladiny spolehlivosti (*Confidence Interval*).
- Rezidua (*Residuals*): nestandardizovaná (*Unstandardized*); standardizovaná (*Standardized*); studentizovaná, tj. rezidua dělená odhadem směrodatné odchylky, která se však liší případ od případu podle vzdálenosti hodnoty závislé proměnné od průměru závislé proměnné (*Studentized*); vynechávaná, tj. rezidua získaná při vyloučení daného případu z odhadu regresních koeficientů (*Deleted*); studentizovaná vynechávaná rezidua, tj. rezidua standardizovaná metodou *Deleted*, dělená svojí standardní chybou (*Studentized deleted*).
- Statistiky vlivu (*Influence Statistics*): *DfBeta(s)* – charakterizuje rozdíly v odhadech regresních koeficientů při vyloučení případu; *Standardized DfBeta(s)* – standardizované rozdíly v odhadech regresních koeficientů při vyloučení případu; *DfFit* – změna v předpovídané hodnotě při vyloučení případu; *Standardized DfFit* – standardizované vyjádření změny v předpovídané hodnotě při vyloučení případu; *Covariance ratio* – podíl determinantu kovarianční matice s vyloučeným případem vzhledem k determinantu počítanému ze všech případů.
- Uložení statistik koeficientů modelu do nového datového souboru (*Coefficient Statistics*): kovarianční matice regresních koeficientů, odhady koeficientů, jejich standardní chyba, signifikance a stupně volnosti testu nulovosti koeficientu.

Možnost uložení do nového datového okna (*Create a new dataset*, okno *Dataset name* specifikuje název datového okna), nebo do nového datového souboru (*Write a new data file*, tlačítko *File* specifikuje název a umístění souboru).

- Export modelu do XML (*Export model information to XML file*): tlačítkem *Browse* definujeme název a umístění souboru a pomocí zaškrtávacího políčka *Include the covariance* matrix rozhodneme, zda má být rovněž exportována kovarianční matice.

2.1.4 Tlačítko Options



Obrázek 8: Tlačítko Options procedury Linear Regression

Tlačítko *Options* je určeno k nastavení parametrů modelu a způsobu práce s chybějícími hodnotami u nezávislých proměnných.

- V části *Stepping Method Criteria* zadáváme parametry pro vstup a výstup proměnných do/z modelu pro případ, že jsme zvolili některou z metod postupného

výběru proměnných. Kritérium může být založeno na hodnotě F nebo na její dosažené hladině významnosti (*significance*).

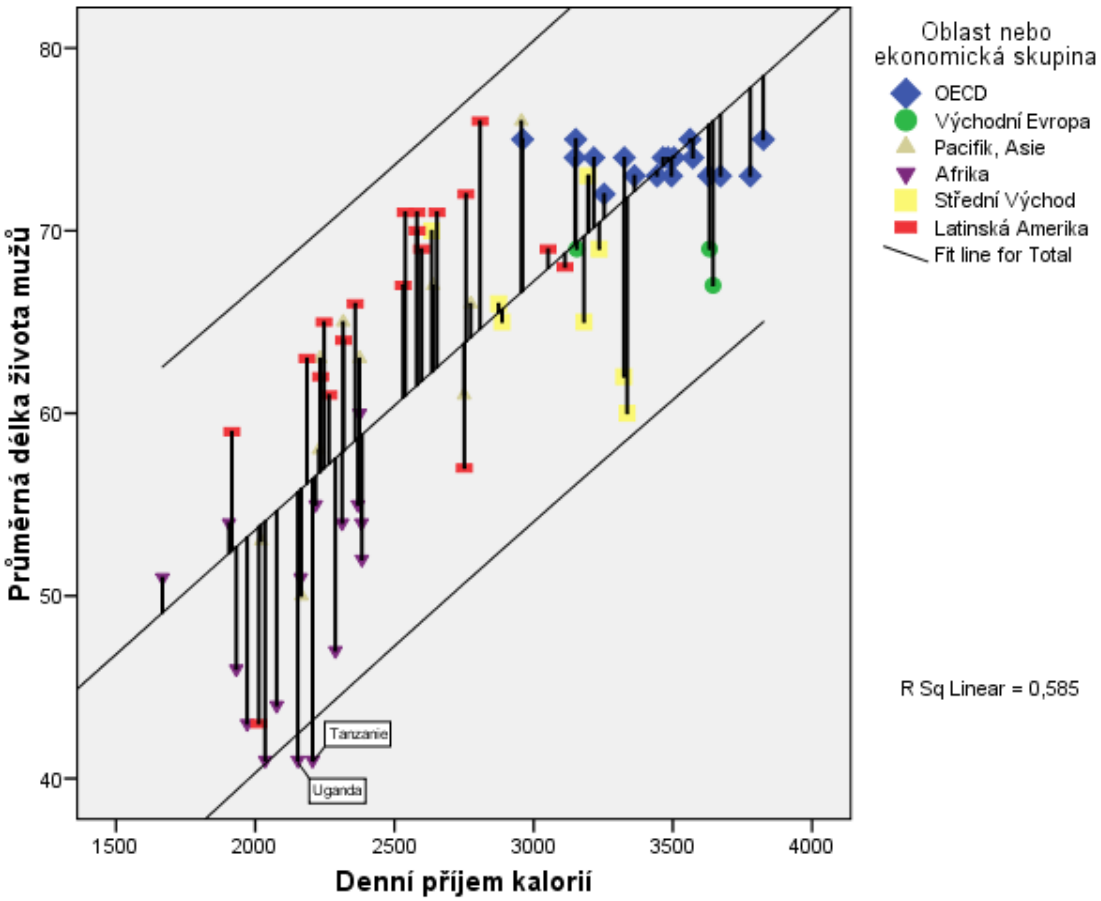
- Pomocí zaškrtávacího políčka *Include constant in equation* rozhodujeme, zda má být do modelu zahrnuta konstanta.
- V poli *Missing Values* volíme způsob práce s chybějícími hodnotami u prediktorů: z výpočtu jsou vyloučeny všechny případy s chybějící hodnotami u některé z nezávislých proměnných (*Exclude cases listwise*), korelační matice, ze které výpočet vychází, je odvozena ze všech platných případů vždy pro danou dvojici proměnných (*Exclude cases pairwise*), chybějící hodnoty jsou nahrazeny průměrem proměnné (*Replace with mean*).

2.2 Výstupy

Datový soubor obsahuje základní informace o vybraných zemích světa (každý případ představuje jednu zemi). Pomocí regresní analýzy se pokusíme vyjádřit závislost Průměrné délky života mužů v dané zemi na Denním příjmu kalorií připadajícím na osobu.

2.3 Bodový graf

Nejprve zobrazíme data do bodového grafu s využitím procedury *Graphs, Legacy Dialogs, Scatter/Dot, Simple Scatter*. Na vodorovnou osu zadáme Denní příjem kalorií, na svislou osu zobrazíme Průměrnou délku života mužů. Dále doplníme informaci o proměnné, která charakterizuje případy (proměnnou země přeneseme do pole *Label Cases By*). Aby mohly být jednotlivé oblasti barevně odlišeny, zadáme ještě proměnnou oblast do pole *Set Markers by*.



Obrázek 9: Graf z procedury Linear Regression

Každý bod v grafu odpovídá jednomu případu, souřadnice vyjadřují hodnoty Denního příjmu kalorií a Průměrné délky života pro danou zemi. Oblasti jsou od sebe odlišeny barvou i typem značky.

Můžeme pozorovat, že vyšším hodnotám jedné proměnné odpovídají rovněž vyšší hodnoty druhé proměnné a naopak. Tento doutníkový tvar grafu je vhodný pro užití lineární regrese.

Dále je zde znázorněna regresní přímka a další dvě čáry (pod a nad regresní přímkou), které vyznačují 95% pás spolehlivosti pro individuální hodnoty – uvnitř pásu by mělo ležet přibližně 95 % pozorování.

Pro každou zemi nalezneme odhad Průměrné délky života mužů vytvořený modelem na regresní přímce v bodě, který odpovídá zjištěnému Dennímu příjmu kalorií. Mezi skutečnou hodnotou a odhadem modelu jsou drobné rozdíly – tzv. rezidua (přímka nevystihuje závislost dokonale, zbývá zde určitá nevysvětlená část variability). Rezidua jsou v grafu znázorněna svislými úsečkami. Dobrý model by měl mít rezidua přibližně normálně rozložena a s nulovou střední hodnotou – odchylky na obě dvě strany jsou stejně pravděpodobné a čím větší odchylka, tím méně je pravděpodobná. V našem případě nalezneme výraznější odchylky od modelu především pro Ugandu a Tanzanii, kde je pozorovaná hodnota již mimo 95% pás spolehlivosti.

R Sq Linear (tzv. koeficient determinace) charakterizuje sílu lineárního vztahu. Nabývá hodnot od nuly do jedné, a čím vyšší je tento údaj, tím silnější je vztah.

SHRNUTÍ

V textu byl kladen důraz na popis jednotlivých metod v IBM SPSS Statistics takovým způsobem, aby účastník byl schopen danou metodu aplikovat na konkrétní problém.

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

SEZNAM POUŽITÉ LITERATURY

Cohen J., Cohen P., West S.G., & L. S. Aiken. (2002). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. UK: Routledge.

Hebák P., Hustopecký J. & I. Malá. (2005). Vícerozměrné statistické metody (2). Praha: Informatorium.

Chen P. Y. & P. M. Popovich. (2002). Correlation: Parametric and Nonparametric Measures. Los Angeles: SAGE.

Schroeder L. D., Sjoquist D. L. & P. E. Stephan. (1986). Understanding Regression Analysis. Los Angeles: SAGE.

SEZNAM OBRÁZKŮ

SEZNAM TABULEK