

Linear Regression – lineární regrese

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:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k + \varepsilon, \text{ kde}$$

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 data proložit přímkou, 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:

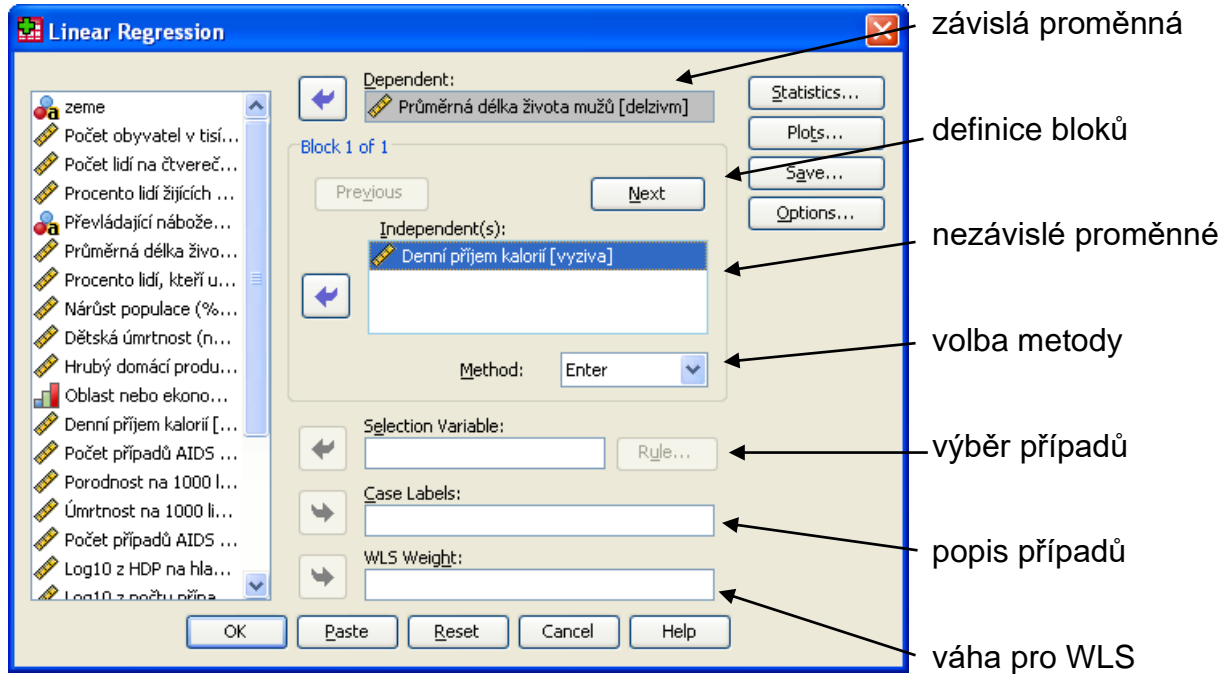
- pozorování jsou mezi sebou navzájem nezávislá
- skutečné hodnoty a chyby jsou navzájem nezávislé
- rezidua mají normálního rozdělení s nulovou střední hodnotou a konstantním rozptylem
- matice X má plnou hodnotu (tj. mezi nezávislými proměnnými není funkční lineární závislost)
- náhodné složky jsou 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.

Volání procedury v IBM SPSS Statistics

Analyze → Regression → Linear

Nastavení dialogu

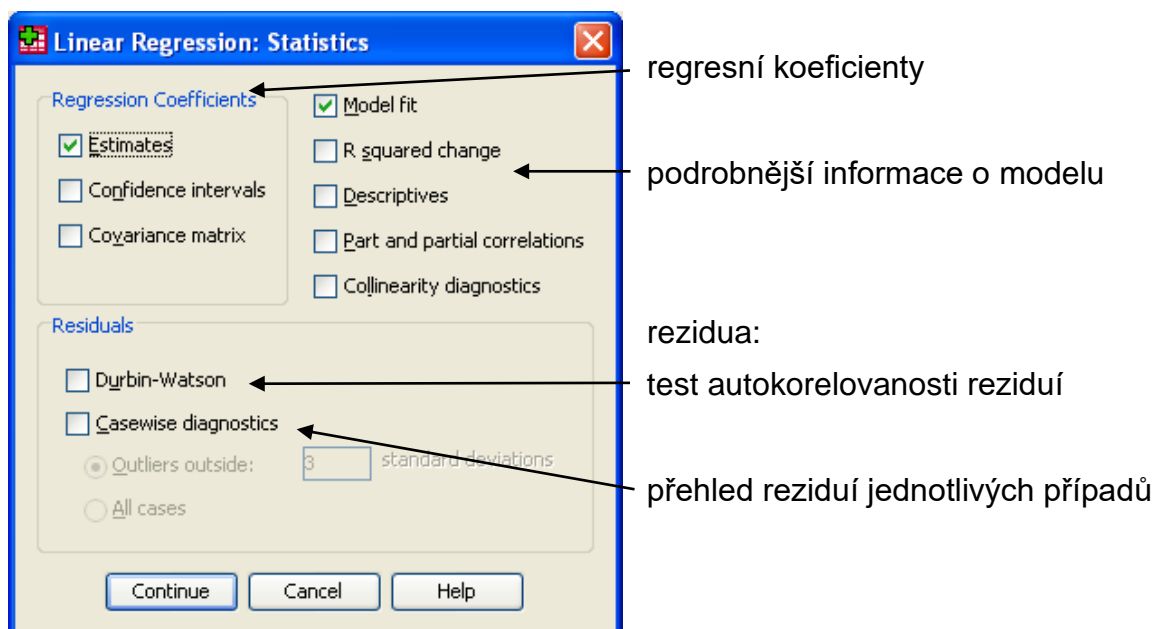


- Do pole *Dependent* přeneseme závislou proměnnou.
- Do pole *Independent(s)* zadáme nezávislé proměnné.
- V rozbalovacím seznamu *Method* zvolíme některou z metod pro výběr prediktorů, které pomáhají zajistit, aby model obsahoval všechny důležité prediktory, ale nebyl přeuročen:
 - *Enter* – model 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.

Listy procedur IBM SPSS Statistics

- 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*.

Tlačítko *Statistics*

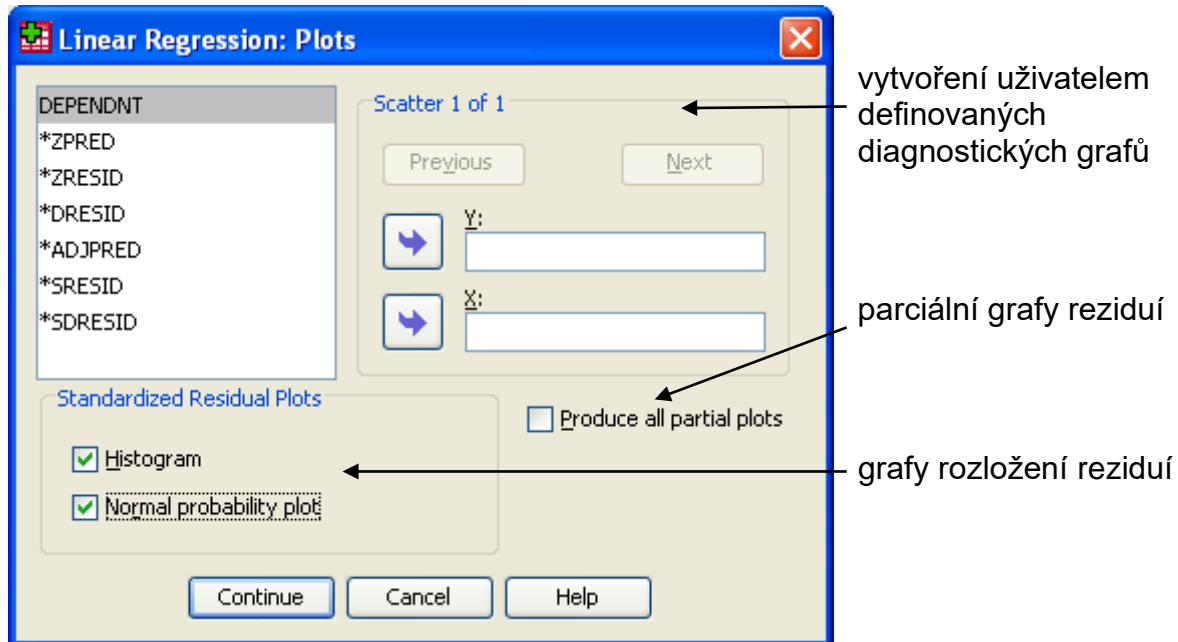


Tlačítkem *Statistics* nastavíme požadované tabulkové výstupy:

- 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*).

Tlačítko Plots

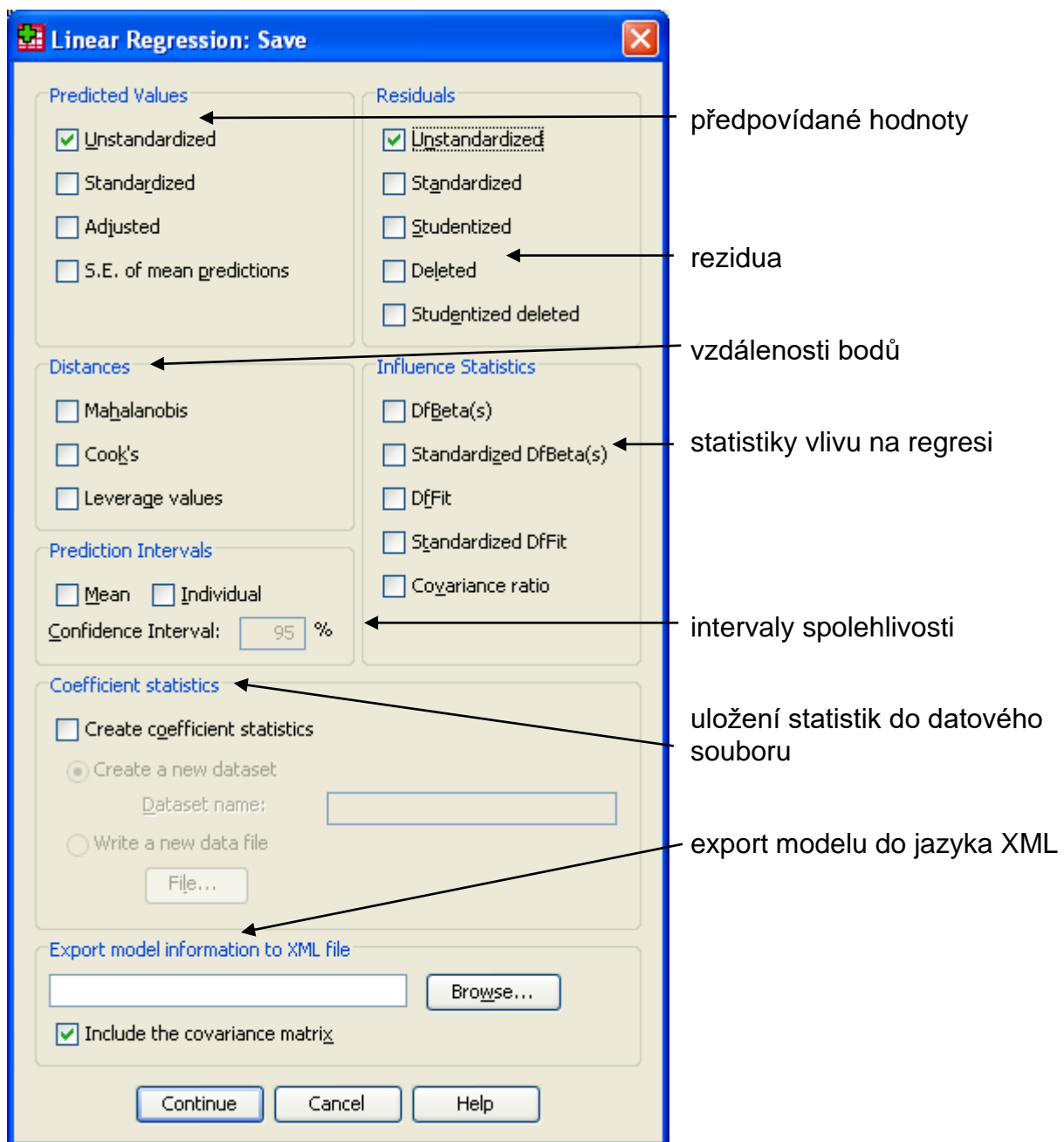


Tlačítkem *Plots* zadáváme požadované typy grafů, které jsou určeny především pro diagnostiku reziduí.

- 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 (*DEPENDNT*), 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 Residual Plots* rozhodujeme o zobrazení histogramu (*Histogram*) a grafu pro ověřování normality (*Normal probability plot*).

Zaškrtnutím políčka *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).

Tlačítko Save



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

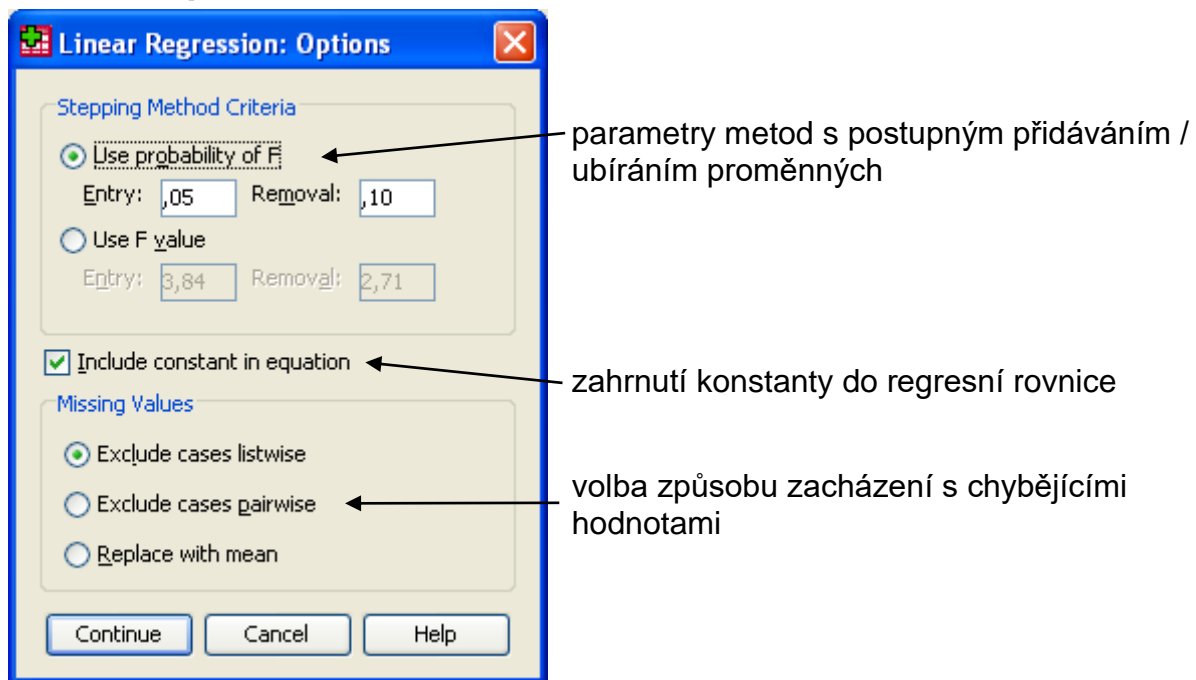
- Předpovídané hodnoty (*Predicted Values*): nestandardizované (*Unstandardized*), standardizované na z-skóry (*Standardized*), adjustované, tj. předpovídané hodnoty získané při vyloučení daného případu z výpočtu regresních koeficientů (*Adjusted*) a standardní chyby předpovídaných hodnot (*S.E. of mean 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,

Listy procedur IBM SPSS Statistics

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, significance 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 zaškrťovacího políčka *Include the covariance matrix* rozhodneme, zda má být rovněž exportována kovarianční matice.

Tlačítko Options



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

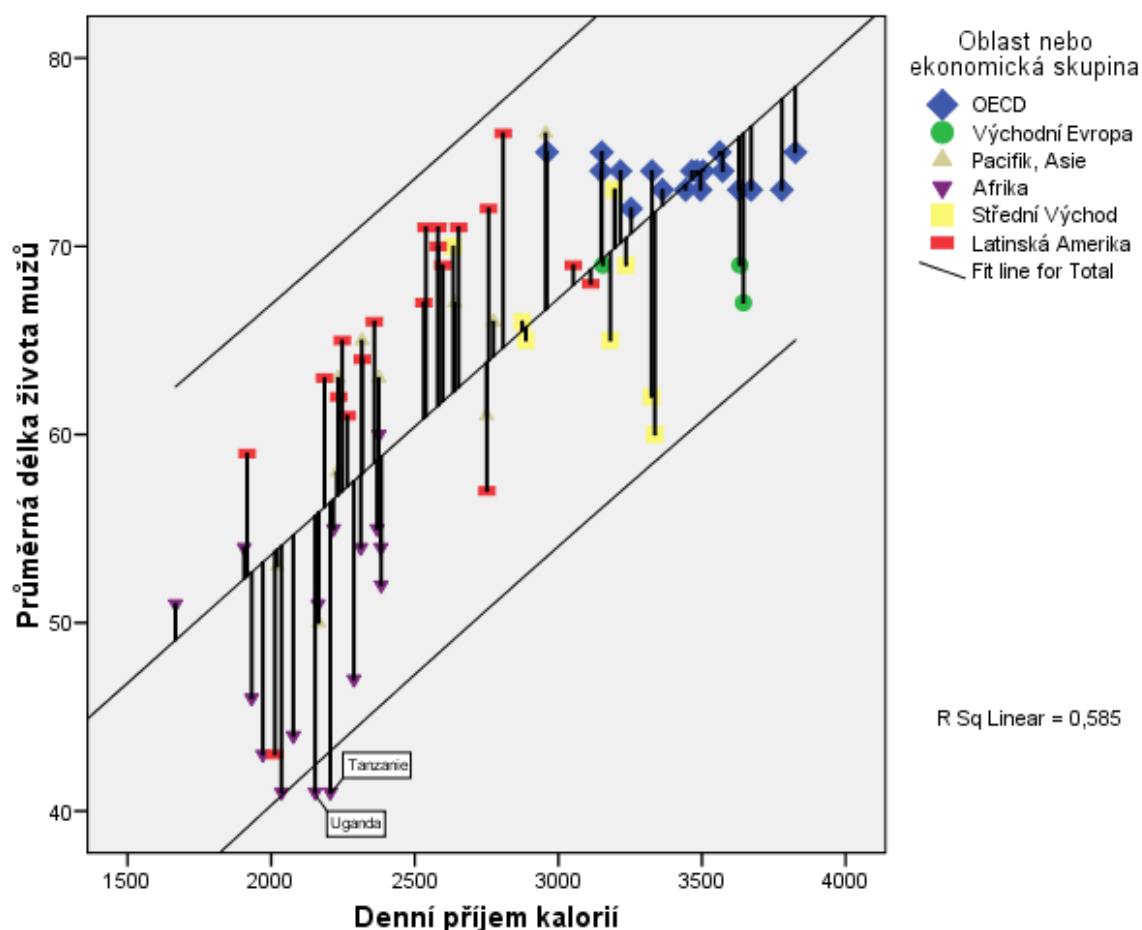
- V části *Stepping Method Criteria* zadáváme parametry pro vstup a výstup proměnných do/z modelu pro případ, že jsme zvolili některou z metod postupného výběru proměnných. Kritérium může být založeno na hodnotě *F* nebo na její dosažené hladině významnosti (*significance*).
- Pomocí zaškrtnutí políčka *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*).

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.

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*.



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.

Listy procedur IBM SPSS Statistics

Dále je zde znázorněna regresní přímka a další dvě čáry (pod a nad regresní přímku), 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.

Následující výstupy byly získány pomocí nabídky *Analyze, Regression, Linear*.

Přehled základních informací o modelu

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	Denní příjem _a kalorií	.	Enter

a. All requested variables entered.

b. Dependent Variable: Průměrná délka života mužů

Tabulka *Variables Entered/Removed* podává přehled základních informací o modelu: v tomto případě byla užita metoda *Enter* – tj. model je vytvořen na základě všech zadaných vstupních proměnných bez dalšího výběru prediktorů, vysvětlovanou proměnnou je *Průměrná délka života mužů* (viz poznámka pod tabulkou) a vysvětlující proměnnou je *Denní příjem kalorií*.

Míry kvality modelu

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.765 ^a	.585	.580	6.555

a. Predictors: (Constant), Denní příjem kalorií

b. Dependent Variable: Průměrná délka života mužů

Tabulka *Model Summary* poskytuje informace o kvalitě modelu. První tři statistiky – *R*, *R Square* (koeficient determinace) a *Adjusted R Square* (adjustovaný koeficient determinace) – vyjadřují shodu modelu se skutečností. *R* je korelační koeficient

predikované proměnné s předpovědí vytvořenou modelem (při jedné nezávislé proměnné se jedná dokonce také o korelační koeficient vysvětlující a vysvětlované proměnné). Koeficient determinace (druhá mocnina R) vyjadřuje procento variability závislé proměnné vysvětlené modelem. Nabývá hodnot od nuly do jedné – je-li blízký jedné, jde o model s vynikající predikční schopností, zatímco blízkost k nule signalizuje špatný model. Standardní chyba odhadu (*Std. Error of the Estimate*) udává velikost typické chyby, které se při použití regresního modelu dopouštíme (skutečná velikost chyby se případ od případu náhodně mění). Toto číslo vyjadřuje odmocninu z *Residual Mean Square* (viz následující tabulka) a mělo by být co nejmenší.

ANOVA

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4427.015	1	4427.015	103.023	.000 ^a
	Residual	3136.905	73	42.971		
	Total	7563.920	74			

a. Predictors: (Constant), Denní příjem kalorií

b. Dependent Variable: Průměrná délka života mužů

Tabulka ANOVA umožňuje testovat, zda takto definovaný model celkově má smysl, tedy zda vysvětlovaná proměnná závisí na lineární kombinaci vysvětlujících proměnných. Nulová hypotéza je formulovaná tak, že všechny koeficienty b_1, \dots, b_k v regresní rovnici jsou nulové. Testujeme ji proti alternativní hypotéze, že alespoň jeden z těchto koeficientů je nenulový. Pro rozhodnutí o zamítnutí nebo nezamítnutí nulové hypotézy je podstatný poslední sloupec tabulky (*Sig.*), který udává dosaženou hladinu významnosti testu. V našem případě je tato hodnota menší než 0.05, a tedy zamítáme nulovou hypotézu. Model má takto smysl.

Ve sloupci (*Sum of Squares*) dále nalezneme rozklad celkové variability analyzované proměnné (*Total*) na složku vysvětlenou regresním modelem (*Regression*) a složku odpovídající náhodným chybám (*Residual*). Ve sloupci *df* jsou uvedeny příslušné stupně volnosti, ve sloupci (*Mean Square*) průměrný čtverec, tj. podíl odpovídajícího součtu čtverců a stupňů volnosti. V následujícím sloupci nalezneme testovou statistiku F , která je získána jako podíl průměrných čtverců.

Tabulka regresních koeficientů

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	26.369	3.772		6.990	.000
	Denní příjem kalorií	.014	.001	.765	10.150	.000

a. Dependent Variable: Průměrná délka života mužů

Listy procedur IBM SPSS Statistics

Tabulka *Coefficients* obsahuje informace o odhadech regresních koeficientů a testech nulovosti.

V části *Unstandardized Coefficients* jsou uvedeny odhady regresních koeficientů (*B*) a jejich standardní chyby (*Std. Error*). V tomto případě má tedy rovnice tvar:

Průměrná délka života mužů = 26.369 + 0.014* *Denní příjem kalorií* + náhodná chyba

Absolutní člen vyjadřuje posunutí regresní přímky ve směru osy y (tj. hodnotu y-nové souřadnice v bodě x=0). V tomto případě by tedy hodnota 26.369 odpovídala odhadované *Průměrné délce života mužů* pro případ, že by *Denní příjem kalorií* byl nulový. Při odhadech je však třeba přistupovat opatrně k „protažení“ přímky příliš daleko, kde vztah již nemusí mít smysl nebo se mění jeho charakter. Hodnota 0.014 vyjadřuje, o kolik by se podle modelu změnila *Průměrná délka život mužů*, jestliže by se *Denní příjem kalorií* zvýšil o jednotku.

Ve sloupci *Standardized Coefficients*, *Beta* nalezneme příslušné standardizované koeficienty. Jestliže z původních proměnných přejdeme na jejich z-skóry, získáme analogickou rovnici s novými koeficienty, z níž však "vypadne" absolutní člen. Absolutní hodnoty standardizovaných koeficientů umožňují porovnat míru vlivu jednotlivých nezávislých proměnných na závislou proměnnou.

Poslední dva sloupce tabulky se vztahují k testování hypotézy o nulovosti jednotlivých regresních koeficientů. Zde nulovou hypotézu na 95% hladině spolehlivosti v obou případech zamítáme, a tedy oba koeficienty mají v regresní rovnici smysl.

Statistiky reziduí

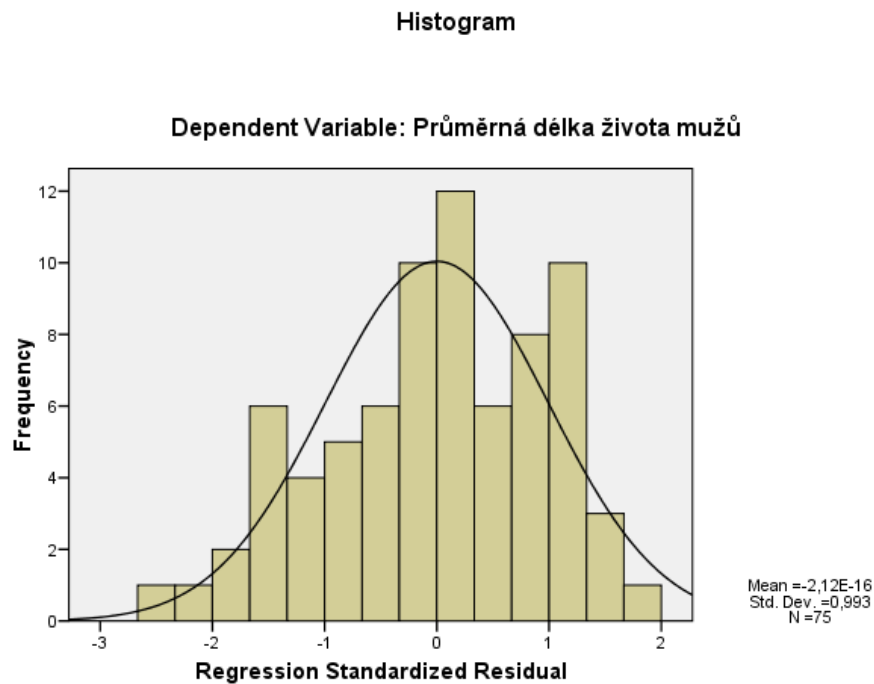
Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	49.08	78.47	63.88	7.735	75
Residual	-15.418	11.382	.000	6.511	75
Std. Predicted Value	-1.914	1.886	.000	1.000	75
Std. Residual	-2.352	1.736	.000	.993	75

a. Dependent Variable: Průměrná délka života mužů

Při označení grafů reziduí se ve výstupu rovněž objeví tabulka s přehledem základních statistik předpovídané hodnoty (*Predicted Value*), reziduí (*Residual*), předpovídané hodnoty standardizované na z-skóry (*Std. Predicted Value*) a standardizovaných reziduí (*Std. Residual*). (Při standardizaci reziduí se vychází z nevychýleného odhadu chybové variance $s^2 = ESS/(N-2)$, z toho důvodu nemá standardizovaná proměnná směrodatnou odchylku 1, ale o něco málo menší.

Histogram reziduí

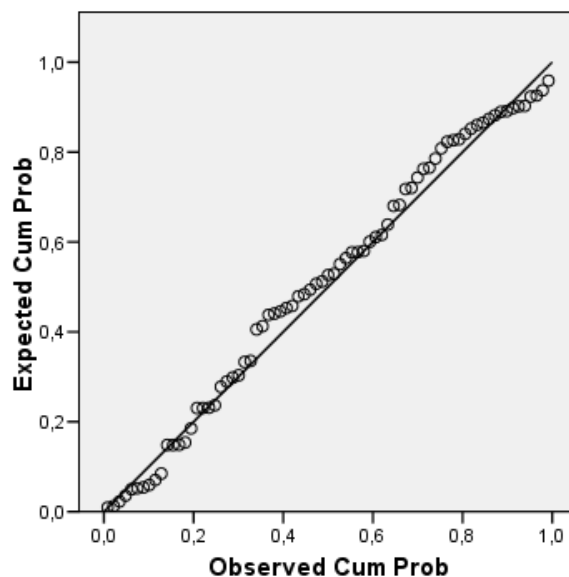


Histogram reziduí umožňuje získat základní přehled o rozložení reziduí a provést srovnání s normálním rozložením. Graf nás rovněž upozorní na případný problém s extrémními hodnotami reziduí. Z našeho obrázku je vidět, že rozložení je o něco více zešikmené doleva a vyskytuje se zde několik větších (záporných) reziduí. Bude se jednat především o Tanzanii a Ugandu, u kterých jsme již v bodovém grafu zjistili, že leží mimo 95% pás spolehlivosti pro individuální hodnoty.

Graf normality reziduí: P-P plot

Normal P-P Plot of Regression Standardized Residual

Dependent Variable: Průměrná délka života mužů



Listy procedur IBM SPSS Statistics

Graf pro ověřování normality dat (P-P plot) slouží k optickému posouzení, zda data pocházejí z normálního rozdělení. Graf zachycuje vztah mezi kumulativními proporcemi zkoumaného a normálního rozdělení (procentem hodnot, které jsou menší než daná hodnota). Každý bod se vztahuje k jednomu případu v datech. Souřadnice na ose x znázorňují kumulativní proporce pozorovaného rozložení pro daný případ, na ose y jsou vyneseny kumulativní proporce očekávaného normálního rozložení. Čím blíže referenční přímce se body nacházejí, tím lepší je shoda s normálním rozdělením.