Mnohonásobná lineární regrese

(pracovní verze, bud ještě doplněno)

Ladislav Rabušic

Sledovat analyticky vztahy mezi dvěma proměnnými je v sociálněvědním výzkumu velmi často přílišným zjednodušením skutečnosti. Sociální svět je charakteristický provázaností jednotlivých proměnných, jejich vzájemným působením, multideterminací. Aby naše analýza byla sofistikovanější, zavádíme modely, které obsahují několik nezávisle proměnných.

Velmi účinnou metodou pro analýzu vztahů mezi sadou nezávisle proměnných a jednou závisle proměnnou je vícenásobná lineární regrese.

V analýze založené na vícenásobné regresi hledáme hodnoty závisle proměnné z lineární kombinace hodnot několika (dvou a více nezávisle proměnných). Vzorec pro výpočet je podobný jako v případě jednoduché regrese:

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3$$

Y je závisle proměnná, jejíž hodnoty se snažíme predikovat, *a* je konstanta, hodnoty b_1, b_2, b_3, jsou regresní koeficienty (říká se jim také parciální regresní koeficienty) a X_1, X_2, X_3, jsou hodnoty nezávisle proměnné.

Cíle mnohonásobné regrese jsou stejné jako u regrese jednoduché:
(1) vysvětlit rozptyl v závisle proměnné *Y* . K tomu slouží statistika R^2;
(2) 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.
(3) s pomocí sestavené regresní rovnice predikovat pro jednotlivé případy hodnoty závisle proměnné.

Dříve ale, než provedeme jakoukoliv regresní analýzu, musíme si být jisti, že naše data splňují několik podmínek (předpokladů) k tomu, aby mohla být do regresní analýzy vpuštěna. Samotné analýze tedy musí předcházet podrobná inspekce dat – tento krok ostatně platí pro jakékoliv statistické analýzy.

Co tedy musíme konkrétně udělat?

1. Musíme prozkoumat, zdali naše data splňují předpoklady pro regresní analýzu.
2. Pokud je nesplňují, musíme se rozhodnout, jak vážné jsou jejich prohřešky proti těmto předpokladům;
3. pokud jsou vážné, musíme s daty provést některé operace, abychom je ostranili.

Předpoklady regresní analýzy

Data musejí naplnit sedm hlavních předpokladů regresní analýzy

1. Závisle proměnná Y musí být proměnná metrická (měřena na intervalové úrovni). Pokud není, musíme použít logistickou regresi.
2. 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.
3. 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.
4. V datech nesmějí být odlehlé hodnoty (*outliers*), neboť na ty je regresní analýza citlivá. Odlehlé hodnoty mohou vážně narušit odhady parametrů rovnice.
5. Proměnné musejí být v lineárním vztahu. Vícenásobná lineární regrese je založena Pearsonově korelačním koeficientu, takže neexistence linearity způsobuje, že i důležité vztahy mezi proměnnými, pokud nejsou lineární, zůstanou neodhaleny.
6. Proměnné jsou normálně rozloženy, jinak hrozí nepřesnost výsledků. V ideálním případě bychom měli vícenásobné rozložení, ale vzhledem k tomu, že toto není tak úplně jednoduché zjistit, je nejlepším řešením je prozkoumat rozložení každé proměnné, která vstupuje do analýzy. 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.
7. 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 testovat některé z předpokladů

Jak odhalit multikolinearitu a jak s ní naložit?

a) Prozkoumejte jednotlivé bivariační korelace. Vysoké vzájemné korelace jsou zdrojem multikolinearity.

b) 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?

Detailní vysvětlení naleznete v textu *Lekce-3_normroz*. Zde tedy jen stručně:

a) prozkoumejte šikmost a špičatost rozložení jednotlivých proměnných

b) nechejte si udělat histogram s proloženou křivkou normálního rozložení

c) použijte Kolmogorov-Smirnovův test

d) 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.

Jak zjistit mnohonásobnou normalitu rozložení?

Porušení mnohonásobné normality rozložení zjistíme analýzou reziduí a jejich grafickým zobrazením. SPSS umí vytvořit dva druhy grafů reziduí:

a) Graf, kdy na ose X jsou vyneseny standardizované predikované hodnoty (*ZPRED) a na ose Y jsou hodnoty standardizovaných reziduí (*standardized predicted values* x *standardized residual values*).[1] Pokud jsou předpoklady mnohonásobného normálního rozložení splněny, distribuce reziduí v datech nebude vykazovat žádný jasný vzorec a případy budou rovnoměrně zobrazeny po obou stranách středové osy Y. Pokud tomu tak nebude, je to indikátor nenormality rozložení.

b) Histogram standardizovaných reziduí. Histogram by měl mít podobu přibližně normálního rozložení.

Test linearity

a) Bivariační linearitu můžeme odhadnout pomocí bodového grafu. Ten je však neúčinný v případě, že náš soubor obsahuje velké množství jednotek

b) 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.

Odlehlé hodnoty

I jedna jediná odlehlá hodnota (v případě nevelkého výběrového souboru), může způsobit problém spolehlivosti vypočtených odhadů. Odlehlé hodnoty odhalíme jednak tím, že prozkoumáme rozložení hodnot proměnné a její směrodatnou odchylku nebo prozkoumáme bodové grafy. Pro vyřešení problému odlehlých hodnot máme několik možností: (1) prověříme hodnoty dané proměnné, zdali při jejich nahrávání nedošlo k překlepu; (2) statisticky proměnnou transformujeme; (3) upravíme hodnotu odlehlého případu; (4) odstraníme případy (respondenty) s odlehlou hodnotou; (5) proměnnou vymažeme.

Různé formy mnohonásobné regrese

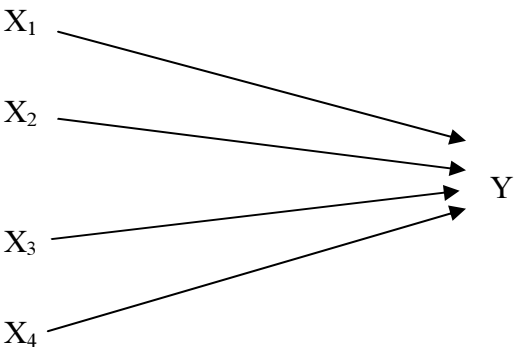
Máme-li data v takové podobě, že vyhovují podmínkám regrese, můžeme regresní analýzu spustit. Jelikož ale existují různé možnosti, jak tuto analýzu provést, musíme ještě učinit několik rozhod-

[1] Viz obr. 10.10 v textu o jednoduché regresi.

nutí.

Především si musíme ujasnit, zdali výsledky, které získáme, nám mají posloužit k deskripci problému, anebo k testování hypotézy o lineárních (a snad i kauzálních) vztazích. Základem mnohonásobné regrese je model. Deskriptivní model má následující podobu (viz obr. 1).

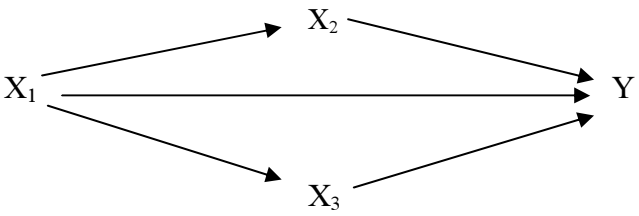
Obr. 1: Deskriptivní model mnohonásobné regrese



V tomto modelu nepředpokládáme žádnou strukturu vztahů mezi nezávisle proměnnými, mnohonásobná regrese nám „pouze“ sdělí sílu vlivu jednotlivých nezávisle proměnných na proměnnou závislou a dále nám řekne, jak velký podíl rozptylu závisle proměnné je vysvětlen našimi proměnnými nezávislými.

Model s kauzální strukturou má podobu jinou (viz obr. 2). Zde již máme představu jednak o vlivu nezávisle proměnných jednak mezi sebou, jednak o vlivu na proměnnou závislou.

Obr. 2: Kauzální model mnohonásobné regrese



Sestavení modelu, ať již jednoduchého deskriptivního nebo složitějšího kauzálního, vyžaduje vždy rozvahu o počtu proměnných, které necháme do vstoupit do mnohonásobné regrese. Samotné adjektivum „mnohonásobná“ by mohlo nezkušeného analytika svádět k tomu, aby pracoval s co možná největším počtem proměnných – s vírou, že čím více proměnných do regrese zahrne, tím vyšší podíl rozptylu vysvětlí.

To je samozřejmě špatný přístup. Ve vědě, stejně jako v životě, platí princip efektivity, tedy snaha dosáhnout s minimálními vstupy maximálně možného efektu. Proto jak v deskriptivním, tak v testovacím způsobu regrese se snažíme počet proměnných redukovat na maximálně možnou míru. Do rovnice zavádíme pouze takové proměnné, o nichž víme z teorie nebo z empirických zobecnění vyplývajících z analýzy jiných autorů, že jsou pro daný problém relevantní.

V deskriptivním postupu, kde jde o maximalizaci vysvětleného rozptylu (R^2), je samozřejmě pokušení zahrnout do modelu větší množství proměnných větší než v modelu testovacím. Odolejme však tomuto vábení. V případě, kdy testujeme hypotézu o kauzálním modelu, je počet proměnných jasně diktován existující teorií.

O počtu proměnných ovšem rozhodují ještě i další, nemeritorní aspekty regresní analýzy. Tato analýza je citlivá na poměr mezi počtem případů a počtem proměnných. Velikost analyzovaného souboru, tedy počet jeho jednotek, případů či respondentů je důležitým faktorem při rozhodování o počtu proměnných, s nimiž můžeme v mnohonásobné regresi pracovat. Značný počet proměnných totiž např. malých souborech neúměrně zvyšuje hodnotu R^2.

Z těchto důvodů statistikové stanovili pravidla, která by nám měla pomoci se v této situaci zorientovat. Pravidla vycházejí z počtu případů, které připadají na jednu proměnnou. Ačkoliv se u různých autorů poněkud odlišují, znějí následovně:

- při regresi založené na metodě *ENTER* (osvětlíme si za chvíli) by mělo na každou proměnnou připadat minimálně dvacet případů (poměr tedy 1:20). Budeme-li tedy chtít pracovat např. se čtyřmi proměnnými, náš soubor by měl mít minimálně 80 jednotek.
- při regresi počítané metodu *STEPWISE* nebo metodou hierarchickou, potřebujeme větší počet případů, neboť poměr by zde měl být 1:40. Pro model se čtyřmi proměnnými tak budeme potřebovat minimálně 160 případů.
- Nejnižší možný poměr proměnná/počet případů je 1:5. V tom případě ale platí silný požadavek na normalitu – graf rezudí by měl vykazovat jasné znaky normálního rozložení. (de Vaus 2002:357).
- Velmi striktní pravidlo, které ve své učebnici několikrát opakuje, razí Field (2000). Tvrdí, že v sociálních vědách bychom měli mít v regresi na jednu nezávisle proměnnou (na jeden predikotr) 15 případů.

Tato pravidla je třeba chápat jako orientační.[2]

Samotná výpočetní procedura regrese může být provedena buď v jednom kroku, nebo ve více krocích. Pokud provedeme regresi v jednom kroku, vrhneme všechny proměnné do výpočtu najednou a ve výsledku se zajímáme o R^2 a sílu jednotlivých regresních koeficientů. Regresi je ale možné provádět i tak, že nejdříve do ní vložíme jednu proměnnou, abychom zjistili, jak velký podíl rozptylu závislé proměnné vysvětlí, pak ve druhém kroku přidáme další proměnnou a sledujeme, kolik variance navíc tato proměnná vysvětlí, pak přidáme třetí atd.

Různé způsoby vkládání proměnných do výpočtu regrese může přinést i značně odlišné výsledky. Pořadí vložených proměnných je důležité, neboť může vést k různým odhadům relativní síly jejich důležitosti. Proto je při mnohonásobné regresi vždy nutné si dobře rozmyslet, jakou metodu vkládání proměnných volíme, což se odvíjí od toho, jaký druh výsledků od regrese požadujeme.

V mnohonásobné lineární regresi máme tři možnosti, jak do výpočtu vkládat proměnné:

1. Metoda standardní (tzv. metoda *Enter*). Všechny proměnné jsou do výpočtu vloženy najednou
2. Metoda postupného vkládání (*Stepwise*). Proměnné jsou vkládány do výpočtu regrese postup-

[2] Jiné pravidlo např. zní: když nás zajímá R^2, pak velikost souboru by měla být přinejmenším $50 + 8k$, kde k je počet nezávisle proměnných. Při regrese se čtyřmi nezávisle proměnnými bychom tak měli mít minimálně 50 + (8x4) = 82 případů. Pokud nás ale zajímá výpočet regresních koeficientů, měl by se počet případů ovíjet od pravidla $104 + k$. Při čtyřech proměnných bychom tak měli mít soubor se 108 jednotkami.

ně 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.

3. 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.

1. Metoda Enter

Tuto metodu použijeme tehdy, když chceme popsat, jak velký podíl variance závisle proměnné je vysvětlen nezávisle proměnnými (R^2), 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). Tato metoda naopak nedovoluje testovat hypotézu o kauzálních vztazích v modelu.

Ilustrace metody (příklad je převzat od de Vause, 2002): V jednom výzkumu byla měřena sociální izolace studentů středních škol. Předpokládalo se, že na tuto závisle proměnnou budou působit tyto nezávisle proměnné: úzkost studenta, sociální dovednosti studenta, jeho symptomy psychózy, míra jeho deprese, výsledky ve škole (prospěch) a míra jeho aktivity. Jelikož výzkumníci neměli žádnou teorii, jak nezávisle proměnné uspořádat do kauzálního modelu, spokojili se s deskripcí problému. Proto pro výpočet regresních statistik použili metodu *Enter*. Výsledky jsou uvedeny v tab. 1.

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

Co regrese ukazuje? Především, nezávisle proměnné vysvětlily 59 % variance sociální izolace, což je docela hodně. Dále vidíme, že pro vysvětlení sociální úzkosti jsou relevantními proměnnými úzkost, symptomy psychózy, deprese a aktivita – všechny jsou statisticky signifikantní (ve sloupci Sig. mají signifikaci nižší než 0,05). Co se týče relativního vlivu, nejsilněji působí proměnná deprese (standardizovaný beta koeficient je 0,72 – čím vyšší míra deprese, tím vyšší skóre sociální izolace), dále působí aktivita (–0,29: čím vyšší aktivita, tím nižší deprese), úzkost (0,28) a psychóza. Vzhledem k tomu, že ani prospěch ani sociální dovednosti nehrály žádnou roli (jejich signifikance byla vyšší než 0,05 a také betakoeficnety jsou nízké), z této regrese by vyplývalo, že na sociální izolaci mají především vliv psychické vlastnosti, sociální již tak vlekou roli nehrají.

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. Jejich charakteristiky jsou uvedeny v tab. 2.

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

Model 1 obsahuje závisle proměnnou (sociální izolace), konstantu a nezávisle proměnnou depresi (viz pozn. *a* pod tabulkou). Model 2 obsahuje navíc ještě proměnnou aktivita, model 3 přidal proměnnou úzkost atd.

Deprese vstoupila do modelu jako první, neboť vysvětluje největší podíl variance, 45 % (čteme sloupec *Adjusted R Square*). Přidáme-li do rovnice proměnnou aktivita (viz model 2), R^2 se zvýší o 4 % na 49 %. Není to velké zvýšení, nicméně, jak ukazuje sloupec *Sig. F Change*, je to zvýšení signifikantní. I další dvě proměnné zvýšily R^2 a to signifikantním způsobem. Poslední model, model 4, vysvětluje celkem 56 % variance závisle proměnné. Ve srovnání s předchozím výsledkem, k němuž jsme dospěli metodou enter, máme tak model, který je „sevřenější“, regrese nám pomohla nalézt model s malým počtem čtyř signifikantních nezávisle proměnných, které (z původních šesti) nejlépe predikují hodnoty proměnné závislé.

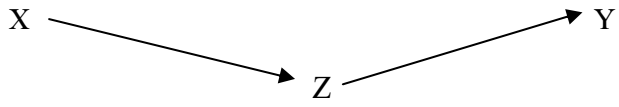
Metoda Stepwise má ještě dvě varianty. Metodu *Forvard* (metoda dopředná), která s nezávisle proměnnými pracuje tím způsobem, že v prvním kroku vypočítá model pouze s konstantou. Pak hledá první proměnnou, která nejlépe předikuje závisle proměnnou – soudí tak podle jednoduchého korelačního koeficientu. Když ji najde a zjistí, že je statisticky signifikantní, vloží ji do modelu a hledá další. Tímto způsobem pak sestaví model, která obsahuje jen ty „nejlepší“ proměnné. Opakem této metody je metoda *Backward* (metoda zpětná), kdy jsou do modelu vsunuty nejdříve všechny nezávisle proměnné a algoritmus výpočtu pak postupně eliminuje krok za krokem ty proměnné, které nejsou statisticky signifikantní. Výsledkem je opět model s těmi „nejlepšími“, to je statisticky signifikantními proměnnými.

Metodu Stepwise je vhodné použít tehdy, když naším cílem je maximalizovat predikci s pokud možná co nejmenším počtem relevantních proměnných. Může také sloužit jako první, explorativní, krok k budování modelu.

3. Metoda hierarchická (Blocks)

Tato metoda umožňuje výzkumníkovi, aby on sám a nikoliv program, určil pořadí vstupu jednotlivých proměnných. Toto pořadí vychází z hypotézy o kauzálním modelu. Mějme např. tento jednoduchý model se dvěma nezávisle proměnnými:

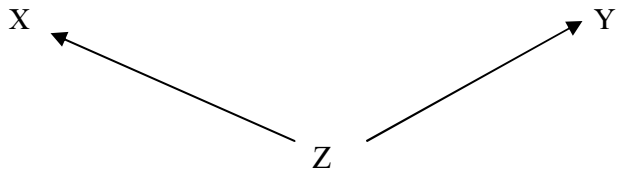
Obr. 3: Kauzální model o třech proměnných, v němž *Z* plní úlohu intervenující proměnné



Zde testujeme hypotézu, že variance Y závisí na proměnné Z a proměnná Z je ovlivňována působením X. Pokud je model platný, neměli bychom nalézt mezi X a Y žádný vztah. Abychom tuto hypotézu otestovali, necháme nejdříve vypočítat model regrese mezi Z a Y a podíváme se na R^2. Pak pustíme, v druhém kroku, do regrese proměnnou X. Pokud je naše hypotéza správná, tímto krokem bychom neměli nijak zvýšit hodnotu R^2, neboť model vztah mezi X a Y nepředpokládá.

Předpokládejme nyní, že máme opět model se dvěma nezávisle proměnnými, ale uspořádaný kauzálně poněkud jinak (viz obr. 4)

Obr. 4: Kauzální model o třech proměnných, kde *Z* je vnější proměnnou



V tomto modelu je Z vnější proměnnou, která způsobuje korelaci mezi X a Y. Testovat tento model by znamenalo pustit do regrese nejdříve proměnnou Z. R^2 by mělo mít jistou netriviální velikost. Po přidání proměnné X bychom opět, jako v předchozím modelu, neměli zaznamenat zvýšení R^2.

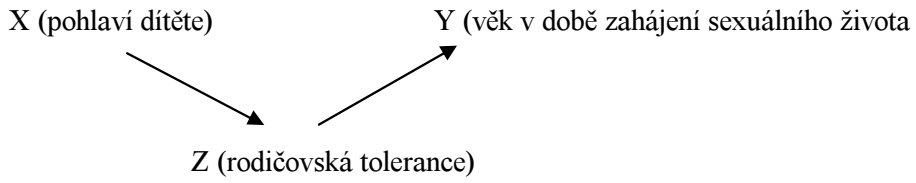
Testy obou modelů by tedy přinesly stejný výsledek, jak tedy rozpoznat, který z nich platí?

Postup:

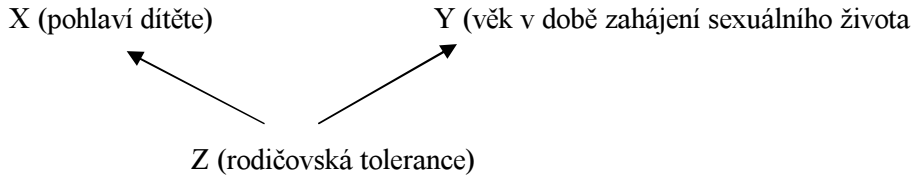
1. Namalujte si diagramy obou modelů a místo symbolů použijte reálná jména proměnných.
2. Zkontrolujte, zda směr působení proměnných odpovídá logice nebo časovému sledu působení proměnných.

Testujeme hypotézu, že na věk v době zahájení pohlavního života má sice vliv pohlaví dítěte, ale nepřímo skrze intervenující proměnnou, jíž je míra rodičovské tolerance. Tu ovlivňuje pohlaví, neboť rodiče chlapců jsou tolerantnější k jeho „párovacím“ aktivitám než rodiče děvčat.

Z jako intervenující proměnná:



Z jako vnější proměnná:



Jelikož musí platit logická a časová sekvence ve vztahu mezi proměnnými, je v našem případě zřejmé, že nemůže platit model vnější proměnné, neboť rodičovská tolerance (Z) nemůže mít vliv na pohlaví dítěte. Proto bychom na základě výsledků regrese museli přijmout model intervenující proměnné.

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

Nejdříve hlavní způsob práce: Pokud nemáme hypotézu o kauzálních vztazích mezi proměnnými, je vždy dobré začít regresi metodou ENTER, tedy se všemi nezávisle proměnnými. Z výpočtů zjistíme, které proměnné přispívají podstatným způsobem k tomu, abychom byli schopni predikovat hodnoty závisle proměnné, a které můžeme vyloučit. Znovu nechejte spočítat regresi, nyní již pouze s relevantními proměnnými a použijte vypočtených koeficientů k sestavení regresního modelu (tedy rovnice). Pak není na škodu použít metodu *forward stepwise* a zjistit přínos jednotlivých prediktorů (nezávisle proměnných).

A nyní základní upozornění: SPSS dokáže vypočítat mnoho komplikovaných operací a nabídnout analytikovi množství výstupů. nedokáže však rozhodnout, zdali náš model je kvalitní, zdali je adekvátní, zkrátka zdali má smysl. To dokáže pouze lidský mozek a v případě mnohonásobné regrese raději ještě lépe mozek kvalifikovaný. Naštěstí mezi údaji, které SPSS nabízí je řada těch, které nám v ohodnocení kvality modelu budou nápomocny.

SPSS vypočítává v mnohonásobné lineární regresi tři hlavní typy výstupů:

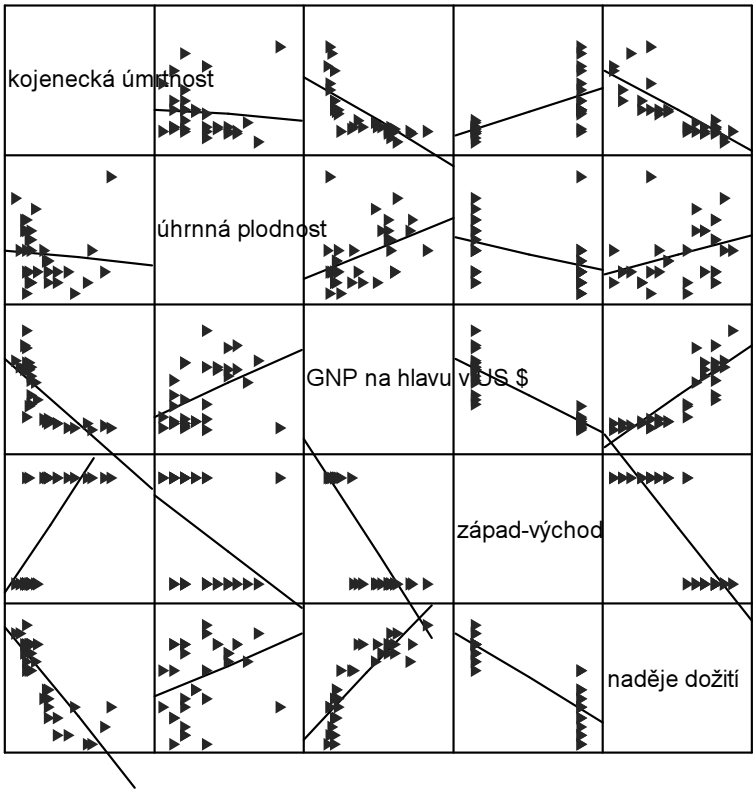
- adekvátnost modelu – R^2
- tabulku ANOVA – test signifikance pro R^2
- regresní koeficenty pro jednotlivé nezávisle proměnné

Ukažme si vše na konkrétním příkladu:

Chceme zjistit, jaký vliv mají na střední délku života proměnné GNP na hlavu, kojenecká úmrtnost, úhrnná plodnost a typ země (dichotomická proměnná: 0 = západ, 1 = východ).

Začněme nejdříve s kontrolou linearity. K tomu si necháme udělat bodový maticový graf (viz obr. 5).

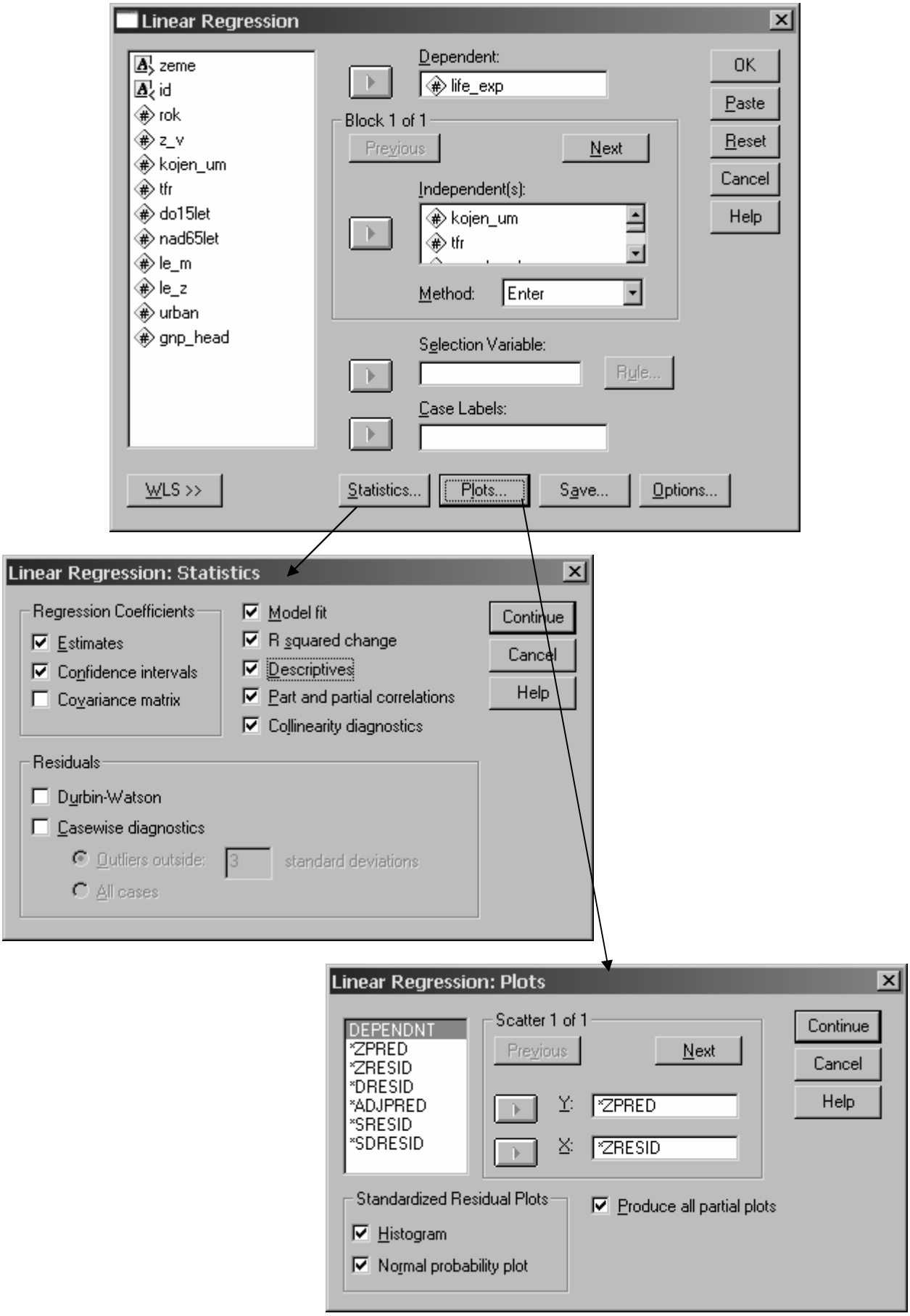
Obr. 5: Bodový maticový graf mezi proměnnými vstupujícími do mnohonásobné regrese



U dichotomické proměnné západ-východ jsou samozřejmě obrázky nesmyslné. S linearitou to nevypadá tak špatně, snad jen vtah mezi kojeneckou úmrtností a úhrnnou plodností není úplně čistý a podobně vztah mezi kojeneckou úmrtností a GNP.

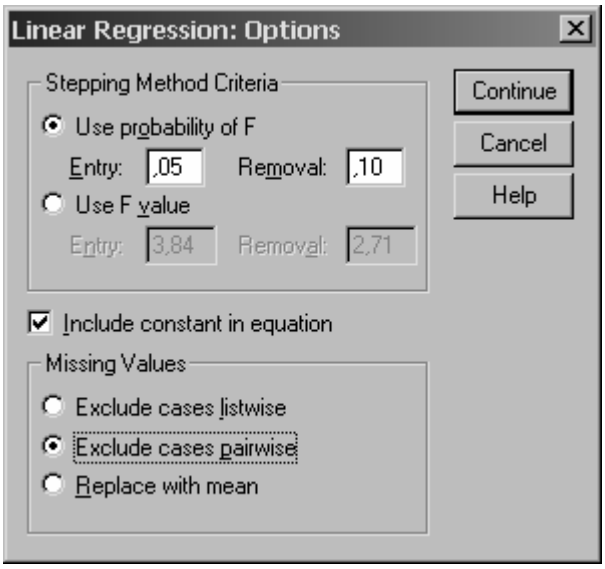
Jak zadat výpočet

Při zadávání výpočtu musíme samozřejmě nejdříve specifikovat závisle a nezávisle proměnné, dále volíme statistiky (*Statistics*), které chceme nechat spočítat, pak grafy (*Plots*) a nakonec se rozhodujeme mezi některými nabídkami z *Options*.


Linear Regression
zeme
id
rok
z_v
kojen_um
tfr
do15let
nad65let
le_m
le_z
urban
gnp_head
Dependent:
life_exp
Block 1 of 1
Previous
Next
Independent(s):
kojen_um
tfr
Method:
Enter
Selection Variable:
Rule...
Case Labels:
OK
Paste
Reset
Cancel
Help
WLS >>
Statistics...
Plots...
Save...
Options...
Linear Regression: Statistics
Regression Coefficients
Estimates
Confidence intervals
Covariance matrix
Model fit
R squared change
Descriptives
Part and partial correlations
Collinearity diagnostics
Continue
Cancel
Help
Residuals
Durbin-Watson
Casewise diagnostics
Outliers outside:
3
standard deviations
All cases
Linear Regression: Plots
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Scatter 1 of 1
Previous
Next
Y:
*ZPRED
X:
*ZRESID
Continue
Cancel
Help
Standardized Residual Plots
Histogram
Normal probability plot
Produce all partial plots

Regrese umí kreslit mnoho grafů, pro nás je nejdůležitější graf standardizovaných predikovaných hodnot (ZPRED) na ose X proti standardizovaným residuím (ZRESID) na ose Y. Umožní nám totiž kontrolovat předpoklad homoskedasticity.

Options:



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 nadije dožití	KOJEN_UM kojenecká úmrtnost	TFR úhrnná plodnost	GNP_HEAD GNP na hlavu v US $ (1998)	Z_V západ-východ
Pearson Correlation	LIFE_EXP nadije 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 nadije 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 nadije 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 – R^2

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 nadije dožití

V této tabulce nás zajímají dva údaje, *R Sguare* (R^2) a *Adjusted* R^2. R^2 říká, jak velké množství variance závisle proměnné (naděje dožití) je vysvětleno sadou námi zvolených nezávisle proměnných. V tomto případě je R^2 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 R^2 může být uměle zvýšena počtem proměnných, které vstupují do analýzy – a právě *Adjusted R Square* bere počet proměnných v úvahu a velikost R^2 na základě toho upravuje (adjustuje). Je to důležité především pro malé soubory, ve velkých souborech se obě statistiky budou dosti podobat.

Někteří statistikové tvrdí, že SPSS počítá tuto charakteristiku nevhodně (podle Wherryho rovnice) a navrhují lepší výpočet podle Steinovy rovnice (viz Field 2000:130).
Jen pro úplnost, R je údaj o mnohonásobném korelačním koeficientu mezi závisle proměnnou a všemi nezávisle proměnnými.

Tabulka ANOVA

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 nadìje dožití

V této tabulce se dozvídáme, zdali platí nulová hypotéza, že $R^2 = 0$. 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 R^2 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).

Regresní koeficienty

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 v US $ (1998)	6,305E-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

Nejdříve nás zajímá konstanta. Jelikož závisle proměnná je měřena v letech, jsou její hodnoty jakož i hodnoty B koeficientů rovněž v letech. Konstanta má hodnotu 79,968, což říká, že průměrná naděje dožití by měla být podle našeho modelu v námi analyzovaných zemích asi 80 roků.

Nyní se podíváme na nestandardizované regresní koeficienty. B koeficient vyjadřuje vliv nezávisle proměnné na proměnnou závislou očištěnou od vlivu působení ostatních proměnných a říká, o kolik se změní hodnota závisle proměnné, pokud se nezávisle proměnná zvýší o jednotku. Nabývají kladných, nebo záporných hodnot. Kladná hodnota koeficientu znamená, že mezi touto nezávisle proměnnou a závisle proměnnou je pozitivní vztah, záporná hodnota indikuje negativní vztah. V našem případě např. zvýšení kojenecké úmrtnosti o jednotku (to je zvýšení o jednotku na 1000 obyvatel) znamená snížení naděje dožití o 0,32 roku, neboť hodnota tohoto koeficientu je záporná -0,32. Podobně zvýšení GNP na hlavu o jeden dolar zvýší naději dožití o 0,000063 roku. Zvýšení o 1000 dolarů na hlavu pak přidá naději dožití 0,063 roku. Pozor ale, na míru vlivu jednotlivých proměnných nelze z B koeficientů usuzovat, neboť jsou měřeny v různých jednotkách. Zajímavý je údaj o vlivu dichotomie východ-západ. Zvýšení o jednotku (tedy při přechodu ze západu na východ) se sníží naděje dožití o 3,2 roku. Hodnoty B koeficientů jsou důležité především pro predikce. My se musíme kromě jejich velikosti zajímat také o jejich signifikanci. Nevýznamný koeficient, tedy takový, jehož signifikance je vyšší než 0,05 indikuje, že výsledek vznikl s velkou pravděpodobností díky výběrové chybě a nemůžeme jej tedy očekávat i v základním souboru.
Z našich koeficientů nejsou signifikantní úhrnná plodnost a GNP.

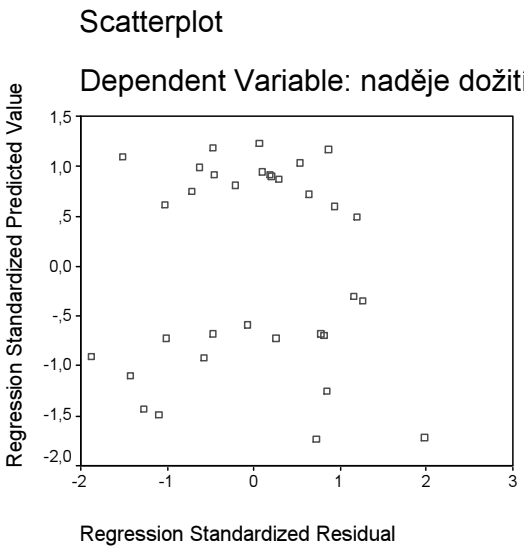
Sloupec *Std. error* slouží k výpočtu intervalu spolehlivosti koeficinetů B, které jsou uvedeny ve sloupci *95% Confidence Interval for B*. Je to klasický interval spolehlivosti, který indikuje, jaký rozsah bude s 95% jistotou mít hodnota B v základním souboru. Čím širší je interval spolehlivosti, tím bude i nepřesnější predikce. To se většinou děje tehdy, když R^2 je nízké – nízké R^2 tak vždy indikuje problém s predikcí.

V dalším kroku se zajímáme o standardizované regresní koeficienty. Ty již umožňují srovnávat míru vlivu jednotlivých nezávisle proměnných. Nejvyšší beta je u dichotomie západ x východ (-0.41) a kojenecká úmrtnost (-0,40).

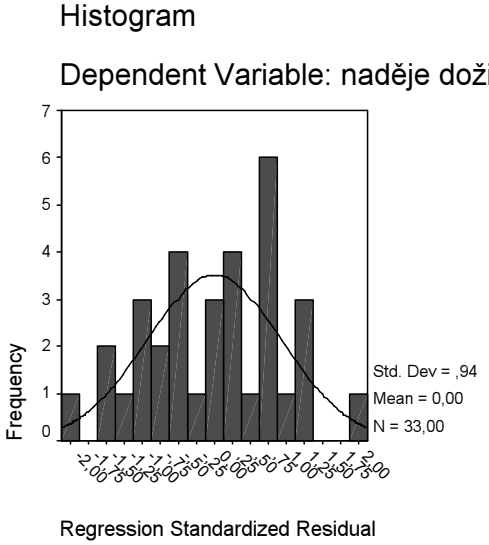
V dalších sloupcích jsou údaje o korelacích (*Correlations*). *Zero-order* korealce (korelace nultého řádu) je běžný Pearsonův koeficient lineární korelace, který měří korelaci mezi proměnnými aniž by kontroloval působení ostatních nezávisle proměnných. Jeho hodnota je v naší tabulce vesměs vysoká, což je dáno vzájemnou interkorelovaností mezi proměnnými. Parciální korelace je korelace mezi Y a X při odstranění vlivu zbylých proměnných. Je to tedy jakýsi čistý vliv dané proměnné na proměnnou závislou v rámci daného regresního modelu. *Part correlation* představuje množství, o které by došlo k redukci R^2, kdyby proměnná X byla odstraněna z regresní rovnice. Pokud bychom tedy z rovnice odstranili úhrnnou plodnost, R^2 by se snížilo o 0,035. Další ze signálů, že tato proměnná nijak k vysvětlení variance v naději dožití nepřispívá.

Poslední dva sloupečky v tabulce 3 uvádějí diagnostické údaje o kolinearitě, konkrétně údaje o *VIF* a *toleranci.* VIF by podle některých nemělo být věší než 10, podle jiných již hodnota 5 naznačuje problém. Jiní říkají, že pokud průměr z VIF je větší než 1, může multikolinearita znehodnotit model. Průměrné VIF vypočítáme tak že prostě sečteme jeho jendotlivé hodnoty apodělíme počtem nezávisle proměnných (tedy: (2,525 + 1,451 + 5,475 + 4,787) / 4 = 3,55.
S VIF souvisí tolerance, je de facto reciproční hodnotou tolerance (1/VIF). Tolerance se pohybuje v intervalu od 0 do 1. Pokud je tolerance menší než 0,2, může tato proměnná způsobovat problém multikolinearity, pokud je menší než 0,1, pak je to již vážné. Mezi našimi proměnnými je to GNP, které má hodnoty nižší, než by měly být.

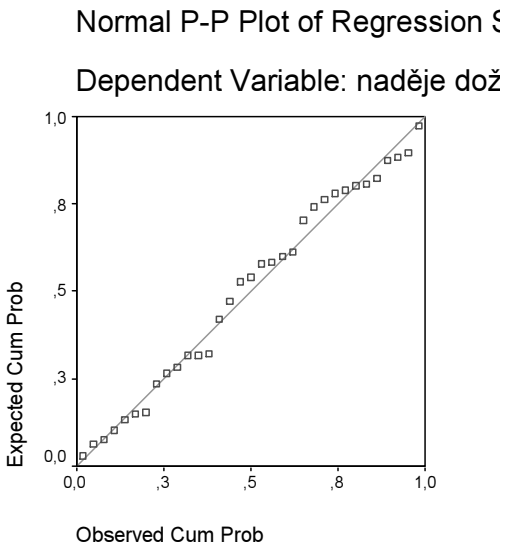
Kontrola předpokladů:



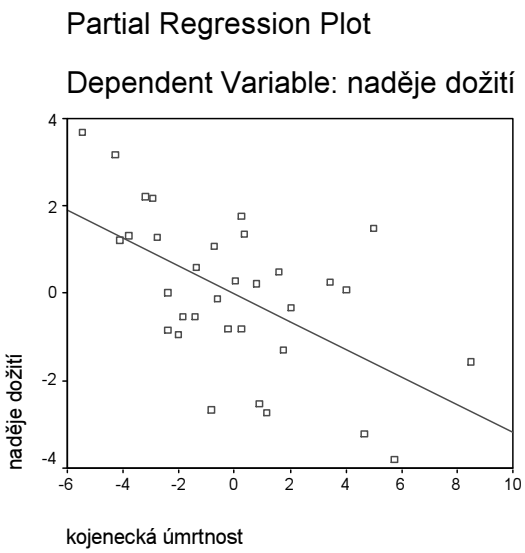
Graf by neměl vykazovat žádný vzorec v uspořádání proměnných: Náš bohužel ukazuje, což je signálem, že předpoklad lienarity a homoskedasticity není naplněn.



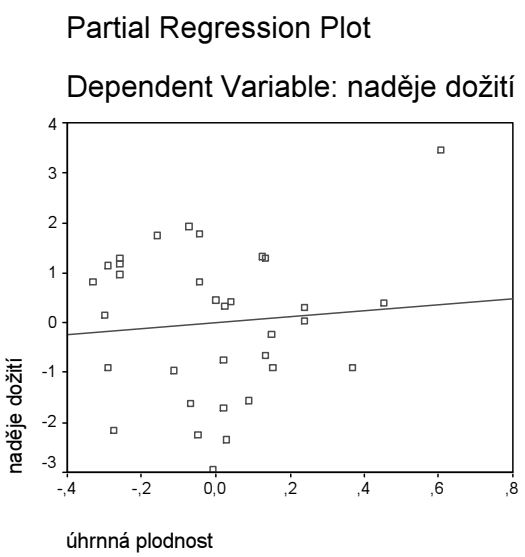
Histogram reziduí ukazuje, že rezidua nejsou normálně rozložena, což znamená že požadavek na mnhonásobnou normalitu je porušen. Což naznačuje i Q-Q graf (viz níže).



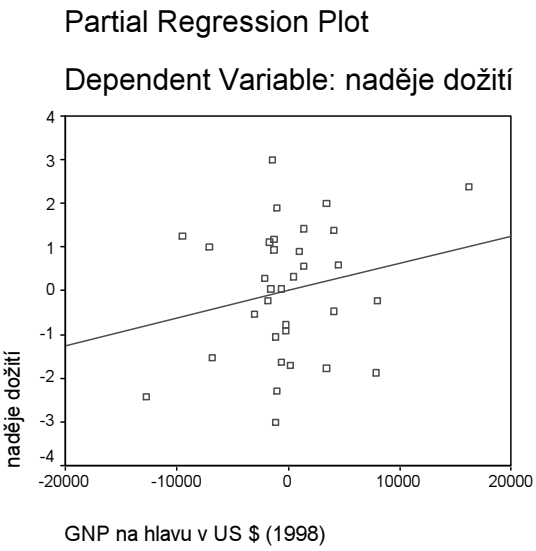
Grafy Partial Regression Plots testují homoskedasticitu:



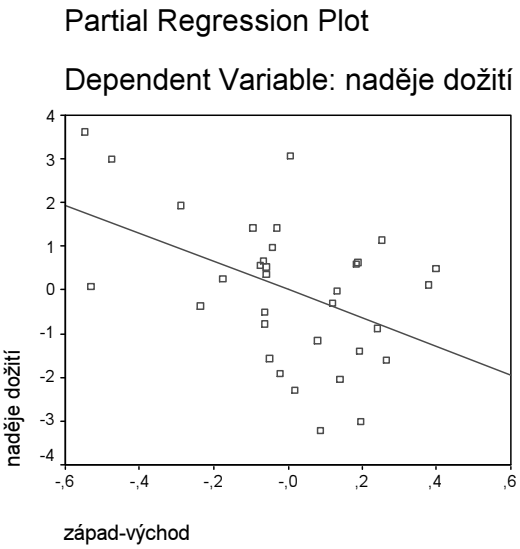
Ok, body jsou rovnoměrně rozloženy kolem přímky.



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



Rovněž špatně



OK

Zkusit metodu Stepwise a metodu hierachickou