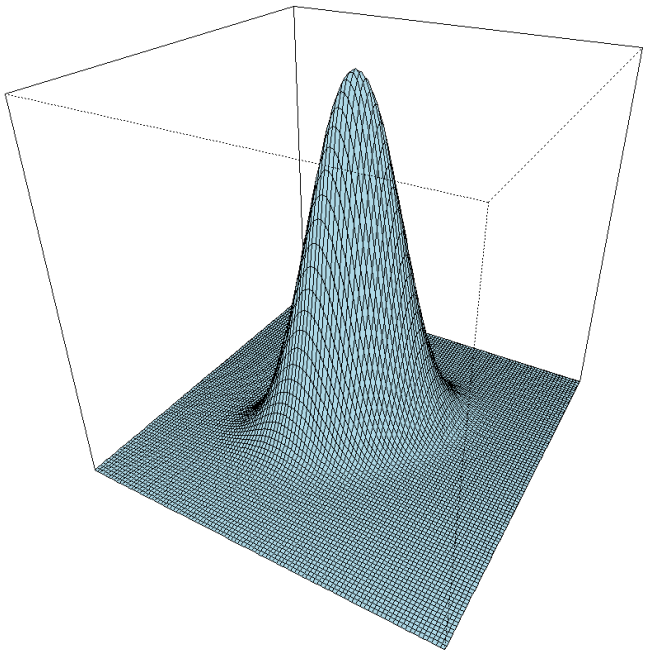
Melandrium

Metody vědecké práce II.



2020

Tomáš Löster, Jakub Danko, Věra Radváková

Ing. Tomáš Löster, Ph. D.
Ing. Jakub Danko, PhD.
PhDr. Věra Radváková, Ph. D.

METODY VĚDECKÉ PRÁCE II.

Recenzenti:
doc. Ing. Silvia Megyesiová, PhD.
doc. Ing. Diana Bílková, Dr.

Autoři:

Vydání této monografie-učebnice bylo podpořeno projektem IRS VŠE v Praze č. 10/2019 s názvem: *Zvýšení úrovně kvalifikačních prací - učebnice Metody vědecké práce II.*

Tato kniha byla doporučena k vydání vědeckou radou nakladatelství.

Libuše Macáková, MELANDRIUM, 2020
Fügnerova 691
274 01 Slaný
IČO: 48709395
ISBN: 978-80-87990-23-0

Obsah

Předmluva **4**
1 Vícenásobná regresní analýza **10**
1.1 Regresní analýza 10
1.2 Zpracování dat z oblasti vícenásobné regresní analýzy 15
2 Metody shlukové analýzy **27**
2.1 Hierarchické metody shlukování 28
2.2 Metody pro pevné C-shlukování 31
2.3 Metody pro fuzzy shlukování 32
2.4 Dvoukroková shluková analýza 32
2.5 Zpracování dat z oblasti shlukové analýzy 33
3 Analýza (metoda) hlavních komponent **46**
3.1 Zpracování dat z oblasti analýzy hlavních komponent 49
4 Faktorová analýza **58**
4.1 Zpracování dat z oblasti faktorové analýzy 61
5 Diskriminační analýza **69**
5.1 Zpracování dat z oblasti diskriminační analýzy 72
6 Zpracování vybraných vícerozměrných statistických analýz v prostředí R **83**
6.1 Zpracování dat z oblasti vícenásobné regresní analýzy 83
6.2 Zpracování dat z oblasti shlukové analýzy 93
6.3 Zpracování dat z oblasti analýzy hlavních komponent 101
6.4 Zpracování dat z oblasti faktorové analýzy 109
6.5 Zpracování dat z oblasti diskriminační analýzy 113
7 Kvalitativní výzkum **118**
7.1 Základní principy a zaměření kvalitativního výzkumu 118
7.2 Metody kvalitativního šetření 122
7.2.1 Řízený rozhovor 122
7.2.2 Případová studie 124
7.2.3 Pozorování 127
7.2.4 Tematická analýza 129
7.2.5 Dotazník 132
7.3 Kvalitativní dotazování 134
7.3.1 Výzkumná otázka 134
7.3.2 Techniky dotazování 135
7.4 Analýza a interpretace dat kvalitativního výzkumu 141
Seznam literatury **148**
Rejstřík **150**
Summary **153**

Předmluva

Vážení čtenáři,

publikace *Metody vědecké práce II* navazuje na učebnici *Metody vědecké práce* (Löster, Mazouch, Radváková, Sigmund, Vltavská, 2018), ve které se autoři zaměřili na kvantitativní metody pro zpracování dat, tedy nejvyužívanější exaktní metody vědecké práce. V předkládané knize čtenář nalezne především náročnější vícerozměrné statistické metody, ale i zpracování kvalitativní linie výzkumného šetření. Vybrané vícerozměrné statistické metody můžete využít nejen pro své vědecké práce (ať už bakalářské, diplomové či doktorské disertační), ale také v praxi. Jedná se o oblíbené metody, které jsou zástupci metod různých oblastí využití. Základním pojítkem pro všechny metody je jejich použitelnost v případě, že jsou jednotlivá pozorování (objekty) charakterizovány alespoň třemi kvantitativními proměnnými. Navážeme na metody a postupy, které byly postupně představovány v prvním díle naší knihy. Zde navazujeme vícenásobnou regresní analýzou, u které se budeme soustředit na možné využití a problémy, které oproti jednoduché regresní analýze mohou vzniknout. Ze zástupců klasifikačních metod, u kterých chceme zařazovat jednotlivé objekty do shluků, se seznámíme se shlukovou a diskriminační analýzou. Z metod sloužících ke snižování dimenze se seznámíme s faktorovou analýzou a analýzou hlavních komponent.

Stejně jako v prvním díle naší knihy s názvem nebudou uváděny všechny výpočetní tvary beze zbytku. Bude představena daná metoda, vyjádřeno její použití a následně na praktickém příkladu ukázána aplikace, včetně komentářů získaných výstupů. Stejně jako v prvním díle knihy také budou všechny metody představeny v systému IBM SPSS verze 21, včetně postupů, jak výstupy získat. Budeme však předpokládat, že přípravu dat a základní popisné

statistiky v systému SPSS již čtenář ovládá. V opačném případě odkážeme na náš první díl. Nově zde také budou ukázány postupy, jak získat stejné výsledky za pomocí systému R, který je v současné době velmi oblíbený. Poslední kapitola se zabývá kvalitativní linií výzkumného šetření. Autorem kapitol jedna až pět je Tomáš Löster, kapitoly šesté Jakub Danko a sedmou kapitolu zpracovala Věra Radváková.

V posledních letech se zvýšil počet a komplexita výzkumných metod, metod pro sběr a zpracování dat, vědeckých přístupů. Bohatost metod poskytuje každému autorovi nejen větší možnost lépe vybrat výzkumný prostředek k dosažení vytyčeného cíle, ale také autoři stojí před nutností volby. Určitá výzkumná metoda nebo strategie není dobrá nebo špatná v absolutním měřítku. Je pouze natolik dobrá, jak adekvátně se hodí pro řešení daného problému. Výzkumný cíl často řídí volbu metody. Jedině správně zvolená metoda může vést k užitečným a důvěryhodným výsledkům. Samozřejmě vedle jasně vymezeného cíle a účelu celé vědecké práce.

Empirický výzkum vždy znamená systematické zkoumání. Je to proces vytváření nových poznatků – systematická a pečlivě naplánovaná činnost. Má nás přiblížit k objevení hledané podstaty, k nalezení nových či zásadních stránek zkoumaného jevu. Musí však být založen na přesném metodickém i metodologickém postupu a na objektivním teoretickém podkladu. Při přípravě je nutné, aby si autor přesně stanovil cíl a metody výzkumu. Jednou z nejobtížnějších částí každé vědecké práce bývá právě začátek – oblast výzkumu, výzkumný problém, účel výzkumu, výzkumná otázka, hypotézy apod. Nutné jsou základní fáze:

- příprava - volba tématu, určení metodiky celé práce a přesně naformulovat cíl,
- plán výzkumu - návrh výzkumu, časový průběh, metody sbírání a

zpracování dat,

- provedení studie - sběr dat, analýza a interpretace,
- zpráva o výsledcích - závěry, limity, doporučení a prezentace různými způsoby.

V odborné literatuře je možné zaznamenat různé pohledy na rozdělení výzkumných metod, řadu definic a vymezení základních pojmů. Existuje mnoho způsobů, jak se o světě něco dozvědět, avšak při vědeckých výzkumech používáme pouze několik hlavních postupů, výzkumných strategií. Podle metodologického přístupu bývá tradiční dělení na základní typy:

- kvalitativní,
- kvantitativní,
- smíšené.

Kvantitativní a kvalitativní metodologie byly často prezentovány jako dvě odlišná a vzájemně soupeřící paradigmata (paradigma chápeme jako přijímané teorie, definice, přístupy). V současné době na ně nahlížíme více jako na dva rozdílné diskursy. Metody a jejich aplikace na zkoumaný problém vždy tvoří jeden celek, což znamená, že různá zaměření výzkumu ovlivňují výběr různých kvalitativních i kvantitativních metod. Také nelze vytvořit jednu metodu stejně dobrou pro zkoumání veškerých otázek daného šetření. Výběr metod a jejich zpracování především závisí na tom, co chceme výzkumem zjistit.

Obě metodologie vycházejí z jiných předpokladů, většinou zkoumají jiné problémy a dávají odlišné závěry. Tyto závěry však nejsou lepší, horší, ani soupeřící, naopak například při postupu použití kvalitativních metod a posléze kvantitativních metod, lze vytvořit dostatečně hlubokou teorii potvrzenou na širokém vzorku. Obě metodologie lze tedy kombinovat. Kvantitativní metodologie se také používá pro potvrzení kvalitativního zkoumání. Pro

doplnění a zjištění různých aspektů je možné použít oba přístupy. Nejedná se o soupeřící programy. Výběr jednoho či druhého typu zkoumání se řídí především výzkumným záměrem a vědeckou orientací autora zkoumání.

Kvalitativní a kvantitativní metodologie se neliší pouze svými metodami, ale také způsobem analýzy. **Kvantitativní** průzkum je zejména založen na sbírání dat, která jsou pak analyzována různými statistickými technikami. Soustřeďuje se na ověřování vztahů mezi proměnnými, nebo na zjištění, jakým způsobem se proměnné spojují. Ještě před vlastním výzkumem je nezbytné znát proměnné a postup interpretace dat podle zvolené kvantitativní metody. Základními metodami jsou různá statistická šetření, testy, dotazníky, experiment, analýza oficiálních statistik, obsahová analýza, strukturované pozorování. **Kvalitativní** metodologie se zabývá specifickým případem v jeho vlastním kontextu, který popisuje v procesu. Jejím cílem je získat vhled a porozumění problému v celé jeho šíři a hloubce. Základními metodami jsou řízené rozhovory, případové studie, pozorování (delší, intenzivní kontakt s terénem), audio či videozáznamy, interpretace a porozumění textům, dokumentům (tematická analýza), otevřené otázky v dotazníkovém šetření.

Kvantitativní přístup testuje předem stanovené hypotézy. Cílem je potvrdit či vyvrátit ověřovanou teorii vědeckého zkoumání. Kvalitativní přístup se snaží porozumět situacím, které jsou často výsledkem neznámého (nepopsaného) procesu.

Učebnice, kterou vám předkládáme, si nedává za cíl přinášet výsledky primárního výzkumu, ale snaží se o maximální srozumitelnost výkladu především pro studenty. Při jejím zpracování autoři využívali známé i méně známé publikační zdroje, které jsou vždy uvedeny v seznamu literatury. Pro jejich vysokou erudovanost a přesně zavedené pojmy jsou v učebnici v rámci

zpracování jednotlivých podkapitol využity tak, aby nedošlo k porušení již nadefinovaných a běžně používaných termínů.

Prosinec 2019 Autoři

Přehled použitých symbolů

X	statistická proměnná
n	počet pozorování
k	počet proměnných
η	regresní funkce
ε	nesystematická složka
I^2	index determinace
R^2	koeficient determinace
β_i	i-tý regresní parametr
S_y	celkový součet čtverců
$S_{y,\mathrm{T}}$	teoretický (modelový) součet čtverců
$S_{y,\mathrm{R}}$	reziduální součet čtverců
p	počet parametrů v regresní analýze, příp. p-hodnota
$\mathbf{x}_i$	i-tý objekt
q	počet nově vytvářených proměnných
s	počet vytvářených diskriminačních funkcí
α	hladina významnosti
$\mathbf{D}$	matice vzdáleností
r	počet shluků

1 Vícenásobná regresní analýza

V prvním dílu naší knihy jsme se seznámili se základními postupy, které jsou součástí jednoduché regresní analýzy. Jak jsme si uvedli, i při využití regresní analýzy je vždy třeba ověřit splnění určitých předpokladů použití dané metody, což bývá v reálných úlohách velmi často opomíjeno. Stejně jako v prvním dílu si popíšeme postupy, zaměříme se na získání výstupů praktických úloh a interpretaci získaných výsledků. Zájemce o podrobný popis metod, včetně všech výpočetních vzorců odkážeme na některý zdroj z uváděné literatury.

V případě analýzy závislosti a hledání modelu pro popis vztahu mezi kvantitativními proměnnými lze za určitých předpokladů využít regresní a korelační analýzu.

1.1 Regresní analýza

Jestliže se budeme snažit modelovat jednostrannou závislost mezi kvantitativními proměnnými, využívá se k tomu regresní analýza. Na rozdíl od jednoduché regresní analýzy, se kterou jsme se seznámili v předchozím dílu knihy, budeme nyní uvažovat vícenásobnou regresní analýzu. V tomto případě budeme mít jednu vysvětlovanou (závislou) proměnnou a celou řadu vysvětlujících (nezávislých) proměnných. Základním cílem je najít takový vztah mezi proměnnými, kdy se pomocí hodnot všech vysvětlujících proměnných (alespoň dvou) budeme snažit pochopit chování hodnot vysvětlované proměnné.

Označme spojitou vysvětlovanou proměnnou Y a vysvětlující proměnné $X_1, ..., X_K$ (spojité nebo diskrétní). Definujme teoretický regresní model, který popisuje závislost proměnné Y na lineární kombinaci proměnných $X_1, ..., X_K$, jako

$$y_i = \eta_i + \varepsilon_i, \qquad (1)$$

kde η_i je teoretická regresní funkce,

ε_i je náhodná složka.

Teoretickou regresní funkce η_i můžeme definovat jako podmíněnou střední hodnotu vysvětlované proměnné *Y* pro jednotlivé kombinace hodnot vysvětlujících proměnných X_i. Jedná se o deterministickou (systematickou) složku modelu.

Nesystematická (náhodná) **složka** ε_i představuje stochastickou část regresního modelu a představuje výsledek působení dalších nesystematických vlivů. Při odhadování regresních modelů si na její chování klademe určité požadavky. Požadujeme, aby náhodná složka se řídila normálním rozdělením s nulovou střední hodnotou a konstantním rozptylem. Formálně je samozřejmě vhodné tyto požadavky pomocí vhodných nástrojů ověřovat. Ty jsou implementovány do softwarových produktů a pomocí *p*-hodnoty, s níž jsme se seznámili v prvním dílu je třeba uvedené předpoklady vyhodnotit. Daný předpoklad v sobě skrývá několik souvisejících požadavků. Požadujeme také, aby všechny dvojice náhodných složek byly nekorelované, tj. aby byly vzájemně lineárně nezávislé. Pokud by jednotlivé náhodné složky byly korelované, regresní modely se komplikují a hovoříme o tzv. *autokorelaci*, kterou je třeba z modelu odstranit. Z daného předpokladu o pravděpodobnostním chování náhodné složky dále vyplývá, že požadujeme, aby jejich rozptyl byl konstantní (a byl roven konstantě σ^2). Pokud by požadavek na konstantní rozptyl nebyl dodržen, jedná se o tzv. *heteroskedasticitu.* Ověřování přítomnosti autokorelace v modelu se provádí například pomocí Durbin-Watsonovy statistiky, jejíž hodnota blízká číslu dva indikuje, že se v modelu nevyskytuje problém s autokorelací prvního řádu.

Problém s autokorelací se obvykle vyskytuje u časových řad a problém s heteroskedasticitou se obvykle vyskytuje u prostorových dat.

Regresní funkci můžeme v našem případě vyjádřit následujícím způsobem

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K, \qquad (2)$$

kde symboly β_i představují neznámé regresní parametry.

Odhady neznámých regresních parametrů lze získat pomocí různých metod, nicméně v souladu s prvním dílem za dodržení určitých předpokladů využijeme *metodu nejmenších čtverců.*

Kvalitu regresního modelu budeme opět hodnotit pomocí *indexu* (resp. *koeficientu*) *determinace*, který je možné určit podle vzorce

$$R^2 = \frac{S_{y,T}}{S_y} = 1 - \frac{S_{y,R}}{S_y}, \qquad (3)$$

kde S_y je variabilita vysvětlované proměnné vyjádřená pomocí tzv. celkového součtu čtverců,

$S_{y,T}$ je část celkové variability vysvětlované proměnné, která je vysvětlena použitým modelem, tzv. teoretický součet čtverců,

$S_{y,R}$ je ta část celkové variability vysvětlované proměnné, kterou se nepodařilo regresním modelem vysvětlit, tzv. reziduální součet čtverců.

Index, resp. koeficient determinace tedy vyjadřuje, jaký podíl z celkové variability vysvětlované proměnné se pomocí daného regresního modelu podařilo vysvětlit díky všem vysvětlujícím proměnným dohromady. Po vynásobení stem udává, kolik procent z celkové variability se podařilo vysvětlit pomocí regresního modelu, tedy hodnotí kvalitu modelu.

V případě, že se srovnávají modely s nestejným počtem parametrů (tedy lišící se například počtem vysvětlujících proměnných), je třeba využít *upravený index* (resp. *koeficient*) *determinace*, který se počítá podle vzorce:

$$R^2_{upr.} = 1 - (1 - R^2)\frac{n-1}{n-p}, \qquad (4)$$

kde p je počet parametrů dané regresní funkce, n je celkový počet pozorování.

Index (resp. koeficient) determinace, stejně jako upravený koeficient determinace nabývá hodnot od nuly do jedné. Hodnota nula znamená, že se pomocí daného modelu nepodařilo vysvětlit žádnou část variability vysvětlované proměnné. Hodnota jedna znamená, že se podařilo vysvětlit daným modelem 100 % variability vysvětlované proměnné.

Zejména u vícenásobné regrese je třeba při vybírání proměnných, s jejichž pomocí se snažíme vysvětlit hodnoty vysvětlované proměnné předpokládat určité teoretické a věcné souvislosti. I v případě vícenásobné regresní analýzy je třeba pro zhodnocení výsledného modelu využívat i další diagnostické nástroje, kterými jsou *dílčí t testy* a ke zhodnocení modelu jako celku *celkový F test*.

DÍLČÍ T TESTY

Jak již bylo uvedeno v prvním dílu naší knihy, testovaná hypotéza všech dílčích t testů, je formulována jako nulová hodnota daného regresního parametru. V případě, že by byla nulová hodnota regresního koeficientu, znamená to, že příslušná vysvětlující proměnná je v modelu nevýznamná a je ji možné z modelu vyřadit. Vyhodnocení provedeme pochopitelně pomocí *p*-hodnoty v softwaru.

CELKOVÝ F TEST

Pomocí *celkového F testu* je ověřována významnost všech regresních koeficientů současně. Testovaná hypotéza *celkového F testu* je formulována jako nulová hodnota všech regresních koeficientů (tzn. parametrů s výjimkou konstanty, která se předpokládá nenulová). V případě nezamítnutí testované hypotézy celkového F testu by to znamenalo, že nemá smysl modelovat daný vztah pomocí vybraných vysvětlujících proměnných a je třeba hledat jiný model. Testové kritérium je založeno na rozkladu celkové variability na tu část, kterou se podařilo pomocí modelu vysvětlit a tu část, kterou se vysvětlit nepodařilo. Vyhodnocení se obvykle provede pomocí *p*-hodnoty ze softwaru.

Poznámka:

Při modelování vztahu pomocí vícenásobné regresní analýzy je třeba ještě ověřovat, zda v modelu není problém s tzv. *multikolinearitou*. Ta představuje vzájemnou škodlivou lineární závislost mezi vysvětlujícími proměnnými. Za tu je považována situace, kdy absolutní hodnota Pearsonova korelačního koeficientu mezi libovolnou dvojicí vysvětlujících proměnných je větší, než 0,75 (v některých zdrojích je uváděna hodnota 0,8). V takovém případě je třeba tento problém odstranit a z modelu vyřadit tu proměnnou, jejíž vliv na vysvětlující proměnnou je nižší. To lze zjistit opět zjistit pomocí korelačního koeficientu. První impulz, jež na možný problém s multikolinearitou může upozornit, je logický „rozpor“ mezi celkovým F testem a dílčími t testy. V případě, že je zamítnuta testovaná hypotéza celkového F testu a nezamítnuty testované hypotézy všech (případně drtivé většiny t testů), analytik by měl ověřit, zda to není způsobeno právě multikolinearitou.

V softwarových produktech bývají implementovány procedury, které slouží k výběru vysvětlujících proměnných. Bývají označeny jako *Forward* a *Backward selection*. Jsou to metody, které slouží k postupnému přidávání,

resp. odstraňování vysvětlujících proměnných. Za proměnnou, kterou je vhodné do modelu zařadit je považována ta proměnná, jejíž zařazení povede ke statisticky významnému zvýšení kvality daného modelu, resp. při vyřazování proměnných se postupuje opačně. Takovýmto způsobem se postupuje do doby, kdy již není žádná další proměnná, která by významně zvýšila kvalitu regresního modelu, resp. žádná proměnná, jejíž vyřazení by nezpůsobilo významné snížení kvality modelu. Tyto metody výběru proměnných taktéž umí odstranit problém s multikolinearitou a nezařadí do modelu současně dvě vysvětlující proměnné, které by tento stav způsobily.

1.2 Zpracování dat z oblasti vícenásobné regresní analýzy

V následující části textu bude demonstrováno použití vícenásobné regresní analýzy na konkrétním příkladu. K řešení bude využit systém SPSS, kde jsou tyto postupy implementovány. Následně bude provedena interpretace dílčích výsledků a závěrů.

Příklad č. 1:

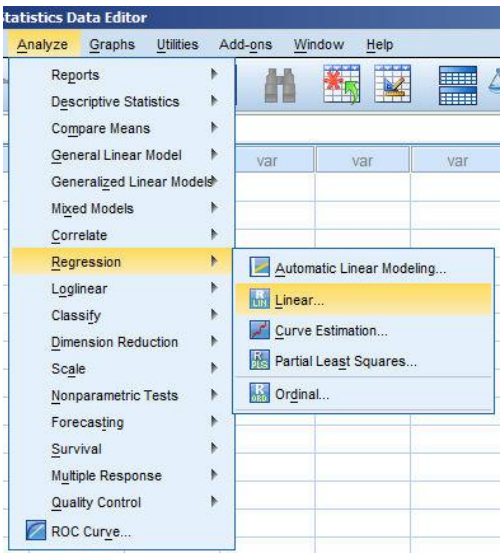
V tabulce 1 jsou zaznamenány prodejní ceny v Korunách, počty ujetých kilometrů a stáří v měsících za 14 automobilů značky Alfa u stejného modelu. Pomocí regresní analýzy nalezněte vhodný model závislosti ceny automobilu značky Alfa na počtu ujetých kilometrů a stáří vozidla. Zároveň vyjádřete vhodnost a kvalitu zvoleného modelu. Při analýzách uvažujte obvyklou 5% hladinu významnosti.

Tabulka 1: Vstupní údaje pro příklad 1

Cena	Počet kilometrů	Stáří
222 934	20 119	15
200 632	26 009	19
198 497	25 003	21
198 486	25 016	20
207 158	23 755	19
202 570	23 756	19
194 419	26 257	21
190 343	27 508	22
210 707	21 271	17
218 957	22 510	18
190 290	27 570	22
186 205	28 831	23
186 264	28 763	23
190 338	27 514	22

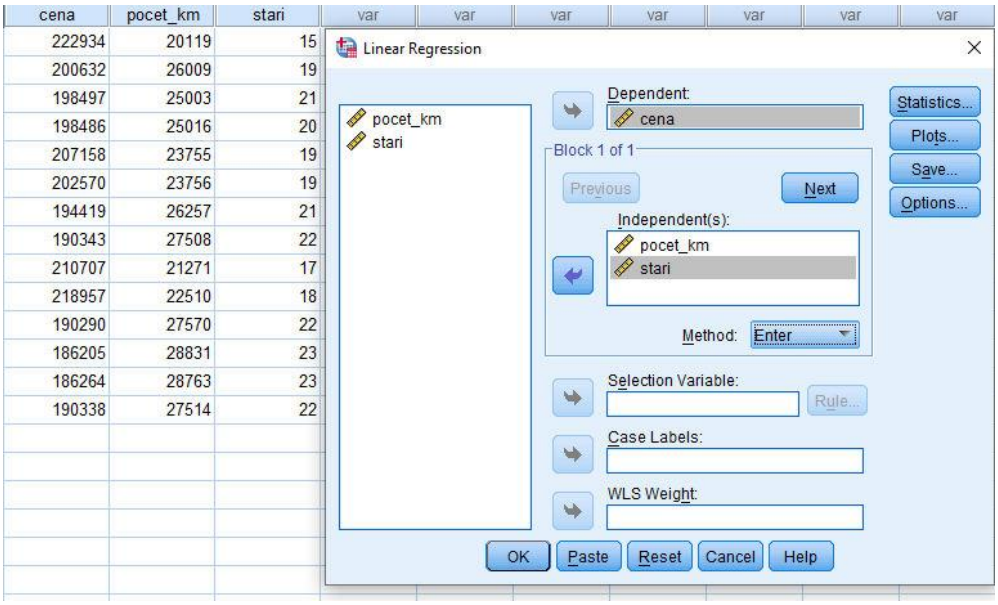
Řešení:

Vzhledem k faktu, že je úkolem prozkoumat závislost ceny (kvantitativní spojitá proměnná) na dvou kvantitativních proměnných, vhodným nástrojem je vícenásobná regresní analýza. Spuštění regresní analýzy v SPSS je, stejně jako v případě jednoduché regresní analýzy následující:



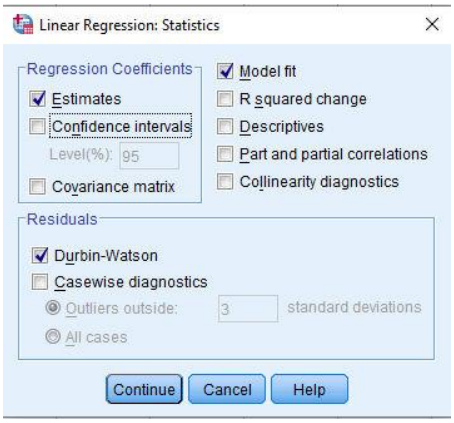
Obrázek č. 1: Spuštění regresní analýzy v SPSS

Po vyvolání menu pro regresní analýzu je třeba správně vložit vysvětlovanou proměnnou do části okna „*Dependent*“, jejíž hodnoty se snažíme vysvětlit (v našem případě se jedná o cenu) a současně obě vysvětlující proměnné (počet ujetých kilometrů a stáří automobilu) do části okna „*Independents*(*s*)“, viz obrázek 2.



Obrázek č. 2: Nastavení vstupních parametrů regrese v SPSS

Ve vstupním okně bude třeba v záložce „*Statistics*“ nastavit příslušnou spolehlivost v případě, že požadujeme intervaly spolehlivosti pro odhady regresních parametrů, resp. „*Durbin-Watsonovu*“ statistiku pro ověření vhodnosti daného modelu.



Obrázek č. 3: Nastavení vstupních parametrů regrese v SPSS

Po zadání uvedených parametrů, jak je uvedeno na výše uvedených obrázcích, po spuštění procesu regresní analýzy obdržíme výstup, který je patrný na obrázku 4.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,967[a]	,936	,924	3195,220	2,803

a. Predictors: (Constant), stari, pocet_km

b. Dependent Variable: cena

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1641142352	2	820571175,9	80,374	,000[b]
	Residual	112303756,0	11	10209432,36		
	Total	1753446108	13			

a. Dependent Variable: cena

b. Predictors: (Constant), stari, pocet_km

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	300626,010	8416,093		35,720	,000
	pocet_km	-2,172	1,240	-,510	-1,751	,108
	stari	-2286,264	1427,674	-,466	-1,601	,138

a. Dependent Variable: cena

Obrázek č. 4: Výstup regresní analýzy z SPSS

Hlavní výstup regresní analýzy je v tomto případě rozdělen do tří tabulek. V první tabulce jsou uvedeny základní charakteristiky modelu. Jsou zde zejména uvedeny důležité charakteristiky, jako jsou index determinace označený „*R square*“, upravený index determinace označený „*Adjusted R square*“, hodnota Durbin-Watsonovy statistiky „*Durbin-Watson*“ a odmocnina z indexu determinace označená „*R*“. Pod uvedenou tabulkou je poznamenáno v části „*predictors*“, že je v modelu uvažována konstanta a obě vysvětlující proměnné. V našem případě stáří i počet ujetých kilometrů.

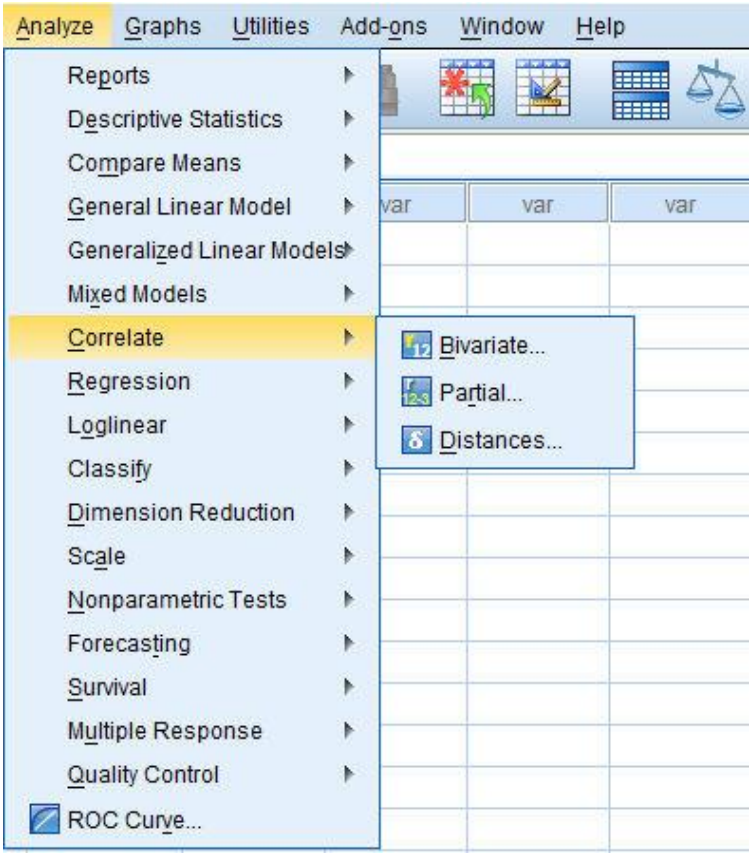
Ve druhé tabulce je uveden průběh celkového F testu. Teoretický (modelový) součet čtverců je označen jako „*Regression*“, reziduální součet čtverců je označen jako „*Residual*“, jejich součet je celkový součet čtverců, který je označen jako „*Total*“. Hodnoty jednotlivých součtů čtverců jsou uvedeny ve druhém sloupci „*Sum of Squares*“. Příslušné stupně volnosti jsou uvedeny ve třetím sloupci označeném jako „*df*“. Hodnoty jsou určeny jako (p – 1), kde p představuje počet parametrů regresního modelu (v našem případě 3), resp. (n – p), kde n je počet objektů (automobilů). Ve sloupci „*Mean Square*“ je uveden předvýpočet testového kritéria. Obě hodnoty se určí jako podíl příslušného součtu čtverců a stupňů volnosti. Hodnota testového kritéria „*F*“ v dalším sloupci se určí jako podíl dvou průměrných čtverců z předchozího sloupce. V posledním sloupci „*Sig.*“ Je uvedena p-hodnota příslušného F testu, která bude následně srovnána se zvolenou hladinou významnosti.

Ve třetí tabulce jsou uvedeny ve sloupci označeném „*B*“ hodnoty odhadnutých regresních parametrů pomocí metody nejmenších čtverců. Dále jsou zde rozepsány průběhy dílčích t testů, kterých je celkem p. Ve sloupci „*Std. Error*“ je uvedena směrodatná chyba odhadu, která se využívá pro výpočet testového kritéria ve sloupci „*t*“. To se stanoví jako podíl příslušného odhadu parametru a směrodatné chyby odhadu. V posledním sloupci „*Sig.*“ jsou uvedeny p-

hodnoty všech dílčích t testů, které jsou, stejně jako v případě celkového F testu, porovnávány s předem zvolenou hladinou významnosti.

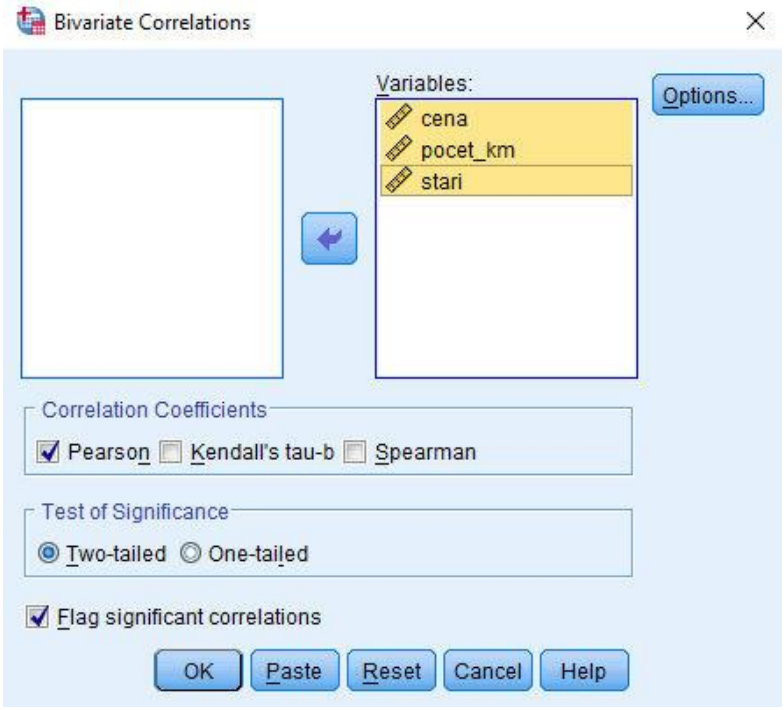
V našem konkrétním případu je z první tabulky zřejmé, že pomocí zvoleného regresního modelu se podařilo vysvětlit 93, 6 % variability hodnot proměnné cena automobilu. Při hodnocení modelu jako celku využijeme celkový F test ve druhé tabulce. Ze stanovené *p*-hodnoty testu je zřejmé (její hodnota je 0,000), že je menší, než zvolená hladina významnosti (0,05) a tak na této hladině významnosti zamítáme testovanou hypotézu. Bylo tedy prokázáno, že alespoň jeden regresní parametr je statisticky významně odlišný od nuly a tím pádem alespoň jedna vysvětlující proměnná má v modelu opodstatnění.

Odhady jednotlivých koeficientů jsou patrné ze třetí tabulky ve sloupci B. Jak je však patrné u obou dílčích t testů u vysvětlujících proměnných, že *p*-hodnoty jsou však větší, než je zvolená hladina významnosti. To znamená, že ani u jedné vysvětlující proměnné není testovaná hypotéza (o nulové hodnotě příslušného regresního parametru) zamítnuta. Tato skutečnost je způsobena problémem tzv. *multikolinearity*, což je, jak bylo uvedeno výše, vzájemná závislost mezi vysvětlujícími proměnnými. Takovýto rozpor mezi závěrem celkového F testu a dílčích t testů je obvykle způsobem právě multikolinearitou. Zda je multikolinearita v daném modelu přítomna lze ověřit pomocí korelační matice, kterou vyvoláme podle návodu na obrázku 5.



Obrázek č. 5: Spuštění korelace v SPSS

V našem případě si vybereme párové korelační koeficienty označené jako „*Bivariate*“ a zadáme všechny vysvětlující proměnné. Zároveň si pro ušetření jednoho kroku rovnou do okna vložíme i vysvětlovanou proměnnou. Z té však nebudeme identifikovat multikolinearitu, ale ponecháme ji pro další účely. Ve spodní části okna lze zvolit příslušný typ korelačního koeficientu, přičemž v našem případě ponecháme automaticky zvolenou volbu Pearsonova korelačního koeficientu. Zároveň je třeba vyznačit podobu alternativní hypotézy testu o nulové hodnotě korelačního koeficientu. V našem případě ponecháme dvoustrannou alternativní hypotézu.



Obrázek č. 6: Zadání parametrů korelace v SPSS

Po zadání vstupních parametrů podle obrázků 5 a 6 získáme výslednou korelační matici, která je patrná z obrázku 7.

Correlations

		cena	pocet_km	stari
cena	Pearson Correlation	1	-,960**	-,958**
	Sig. (2-tailed)		,000	,000
	N	14	14	14
pocet_km	Pearson Correlation	-,960**	1	,965**
	Sig. (2-tailed)	,000		,000
	N	14	14	14
stari	Pearson Correlation	-,958**	,965**	1
	Sig. (2-tailed)	,000	,000	
	N	14	14	14

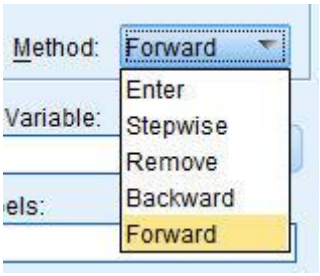
**. Correlation is significant at the 0.01 level (2-tailed).

Obrázek č. 7: Výstup korelační matice v SPSS

Z uvedeného výstupu je zřejmé, že hodnota korelačního koeficientu mezi stářím vozidla a počtem ujetých kilometrů je velmi vysoká, konkrétně jde o hodnotu 0,965. Podle výše popsaného je za škodlivou závislost považována absolutní hodnota korelačního koeficientu větší, než 0,8, což v tomto případě je jednoznačně splněno. Příslušný test prokazuje statisticky významnou závislost mezi danou dvojicí proměnných, což poznáme podle *p*-hodnoty, kterou nalezneme v příslušné buňce. Protože je daná hodnota prakticky nulová, znamená to, že na všech rozumných hladinách významnosti zamítáme testovanou hypotézu a závislost je považována za statisticky významnou.

Je tedy zřejmé, že obě vysvětlující proměnné v modelu nemohou být současně a jednu bude třeba z modelu odstranit. Logicky to bude ta proměnná, jejíž vliv na vysvětlující proměnnou je menší. To zjistíme pomocí korelačních koeficientů právě těchto proměnných ve vztahu k vysvětlované proměnné. Z uvedené korelační matice je zřejmá nižší síla závislosti stáří a ceny, proto bude zřejmě vyřazována stáří automobilu.

V případě, že bychom měli složitější model, kde by bylo více vysvětlujících proměnných než dvě a byla-li by identifikována multikolinearita, jednalo by se o složitější proces výběru proměnných. V systému SPSS jsou implementovány metody, které pomáhají proměnné vybírat. Ve vstupním okně jsou při spouštění regrese v dolní části výstupu označeném „*Method*“ jednotlivé způsoby výběru uvedeny. Dosud v modelu bylo označeno „*Enter*“, kdy do modelu byly zahrnuty všechny vysvětlující proměnné.



Obrázek č. 8: Metody výběru proměnných v SPSS

Nyní si ukážeme průběh metody postupného vyřazování proměnných označené „*Backward*“. Po již známém spuštění regresní analýzy získáme následující výstup, který je uveden na obrázku 9 až 11.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	stari, pocet_km[b]	.	Enter
2	.	stari	Backward (criterion: Probability of F-to-remove >= ,100).

a. Dependent Variable: cena

b. All requested variables entered.

Obrázek č. 9: Výběr proměnných „*Backward*“ část. 1 v SPSS

Jak je patrné z obrázku 9, na počátku byly do modelu vloženy obě vysvětlující proměnné a vysvětlovanou proměnnou je cena automobilu. V dalším kroku, který je označen jako „*Model 2*“ je z modelu vyloučena proměnná stáří (u níž byl výše prokázán nižší vliv) na vysvětlovanou proměnnou. Takovýmto způsobem by bylo postupováno až do situace, kdy by z modelu byly odstraněny jednotlivé „problematické“ proměnné.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,967[a]	,936	,924	3195,220
2	,960[b]	,921	,914	3397,123

a. Predictors: (Constant), stari, pocet_km

b. Predictors: (Constant), pocet_km

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1641142352	2	820571175,9	80,374	,000[b]
	Residual	112303756,0	11	10209432,36		
	Total	1753446108	13			
2	Regression	1614960760	1	1614960760	139,939	,000[c]
	Residual	138485348,0	12	11540445,67		
	Total	1753446108	13			

a. Dependent Variable: cena

b. Predictors: (Constant), stari, pocet_km

c. Predictors: (Constant), pocet_km

Obrázek č. 10: Výběr proměnných „*Backward*“ část. 2 v SPSS

Z obrázku 10 je patrný průběh celkového F testu pro oba dva modely. První je při přítomnosti obou vysvětlujících proměnných, druhý již pouze s vysvětlující proměnnou počet kilometrů. Tyto skutečnosti jsou zobrazeny pod tabulkou. Je zřejmé, že oba modely jsou statisticky významné, protože testovaná hypotéza (na základě hodnot ve sloupci „*Sig*.“) byla zamítnuta.

Příslušné odhady parametrů, včetně konstanty a průběhy dílčích t testů jsou zřejmé pro oba modely z obrázku 11.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	300626,010	8416,093		35,720	,000
	pocet_km	-2,172	1,240	-,510	-1,751	,108
	stari	-2286,264	1427,674	-,466	-1,601	,138
2	(Constant)	303195,202	8783,813		34,517	,000
	pocet_km	-4,089	,346	-,960	-11,830	,000

a. Dependent Variable: cena

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
2	stari	-,466[b]	-1,601	,138	-,435	,069

a. Dependent Variable: cena

b. Predictors in the Model: (Constant), pocet_km

Obrázek č. 11: Výběr proměnných „*Backward*“ část. 3 v SPSS

Z výstupů pro model 1 je zřejmé, že dílčí t testy u obou vysvětlujících proměnných jsou nevýznamné, tedy takto postavený model je nevhodný. Model číslo 2 je již pouze s vysvětlující proměnnou počet kilometrů, a vychází jako významný. Hodnota koeficientu determinace je v tomto případě 0,921, což lze považovat za velmi uspokojivou hodnotu a model v tuto chvíli můžeme prohlásit za kvalitní a zapsat si jeho odhadnutý tvar, jehož průběhem je regresní přímka:

Cena_automobilu = 303 195,2 – 4,085*počet_kilometrů.

Z uvedeného modelu je zřejmé, že s každým dalším ujetým kilometrem se cena automobilu snižuje v průměru o 4,085 Korun a úplně nový automobil by stál 303 195,2 Korun.

2 Metody shlukové analýzy

V současné době je velmi oblíbená vícerozměrná klasifikační metoda, kterou používají i odborníci z praxe, nazvaná shluková analýza. Základním cílem všech klasifikačních metod je zařazování objektů do skupin na základě jejich vlastností. V závislosti na tom, zda jsou skupiny dopředu známé či neznámé, se ke klasifikaci využívá shluková či diskriminační analýza. U shlukové analýzy obvykle nebývají skupiny dopředu známé, naopak u diskriminační analýzy bývají skupiny předem známé. Použití shlukové analýzy je velmi široké. Lze jej najít v medicíně při klasifikaci pacientů, marketingu, při klasifikaci zákazníků. Využít ji lze také ke klasifikaci dokumentů, zemí, podniků či lidí.

Při aplikaci shlukové analýzy je třeba rozlišovat, pomocí jakých typů proměnných jsou jednotlivá pozorování (tzv. objekty) charakterizovány. Proměnné mohou být kvantitativní, binární či kvalitativní. Pro analýzy lze využít také proměnné různých typů. V současné literatuře existuje mnoho metod a postupů, které využívají výše uvedené proměnné, avšak nejvíce propracované jsou postupy, kdy jsou proměnné čistě kvantitativní. V odborné literatuře jsou uváděny různé způsoby klasifikace metod shlukové analýzy. Ty je možné rozčlenit na tzv. *tradiční* metody, které jsou hojně využívány, a které jsou dostatečně implementovány v softwarových produktech a *moderní* metody, k jejichž rozmachu dochází díky rozvoji výpočetní techniky. Obvyklé členění tradičních metod, které je uváděné ve většině pramenů, je členění na *hierarchické* a *nehierarchické* metody shlukování. Hierarchické metody směřují k vytváření tzv. stromů (stromovité struktury shluků) a *nehierarchické* metody se zaměřují na rozklad jednotlivých objektů do určitého počtu shluků. Výhodou hierarchických metod shlukování je, že dopředu nebývá nutné znát počet shluků, do kterých se objekty klasifikují. Tato informace může být právě

jejich výsledkem. Mezi moderní metody shlukování lze zařadit metody založené na mřížce, metody založené na modelu a modely založené na hustotě.

Dalším důležitým pohledem, podle kterého lze dělit metody shlukování je na tzv. *disjunktní*, u kterých je každý objekt jednoznačně zařazen do právě jednoho shluku a metody *překrývající se*, u kterých mohou být objekty zařazeny do více shluků. Jiný pohled na shlukování dělí metody na *pevné shlukování*, u kterých je zařazení objektu do shluku vyjádřeno hodnotou nula nebo jedna a tzv. *fuzzy shlukování*, u kterého je pro každý objekt a každý shluk stanovena tzv. *míra příslušnosti*. Ta představuje číslo z intervalu od nuly do jedné a vyjadřuje jakousi „pravděpodobnost“ klasifikace daného objektu do příslušného shluku.

V rámci naší knihy bude uvažováno pouze pevné shlukování s jednoznačným přiřazením do jednotlivých shluků a tradiční hierarchické metody shlukování, které jsou velmi dobře propracované a implementované do mnoha softwarových produktů, včetně IBM SPSS.

2.1 Hierarchické metody shlukování

Tyto metody jsou založeny na hierarchickém uspořádání objektů do shluků. Jejich výhodou je, že před samotnou analýzou není nutné stanovit počet shluků, což je považováno za jejich hlavní výhodu oproti většině ostatních metod shlukování. Hierarchické shlukování je reprezentováno dvěma základními typy shlukování, a to *aglomerativním shlukováním* (postupně se spojují jednotlivé objekty nebo skupiny objektů do shluků až do situace, kdy jsou všechny objekty spojeny do jednoho shluku) a *divizivním shlukováním* (postupuje se opačným způsobem, tj. na počátku jsou všechny objekty v jediném shluku a jeho postupným rozdělováním se získá hierarchický systém podmnožin shluků).

Aplikace různých metod shlukování na stejné objekty popsané identickými vlastnostmi mohou přinášet různé výsledky. Jak se uvádí v literatuře, například Löster, Pavelka (2013) či Gan (2007) „Nelze apriori říci, která z metod je nejlepší pro daný problém“. Důležitou vlastností hierarchických metod shlukování je skutečnost, že výsledky předešlého kroku jsou vždy přiřazeny k získaným výsledkům v následujícím kroku, a je tak vytvářena stromová struktura. Grafickým vyjádřením postupného procesu shlukování jednotlivých objektů je *dendrogram,* což je znázornění procesu spojování nebo rozdělování dílčích shluků v jednotlivých krocích v závislosti na jejich vzdálenosti. Při shlukové analýze je využita matice vzdáleností, a tak i volba typu vzdáleností ovlivňuje výslednou klasifikaci.

Aglomerativní způsob shlukování může být popsán následujícím způsobem:

1. Na základě vybrané míry vzdáleností je stanovena matice vzdáleností **D**, v níž jsou zaznamenány vzdálenosti (odlišností) pro všechny dvojice objektů. V počátečním kroku je každý objekt samostatným shlukem.

2. Z matice vzdáleností **D** se vyberou se dva shluky, jejichž vzdálenost je minimální.

3. Tyto dva shluky jsou následně spojeny do nového shluku. V matici vzdáleností **D** je vynechán řádek a sloupec reprezentující vzdálenost těchto spojovaných shluků a je nahrazen novým řádkem, resp. sloupcem, který představuje jejich společnou vzdálenost od ostatních shluků, podle typu zvolené metody. V tomto kroku se rozměr matice vzdáleností **D** snižuje o jednotku.

4. Postup se opakuje od kroku číslo 2 do okamžiku, kdy se všechny objekty nachází v jednom shluku.

Nejznámější a nejčastěji používanou mírou je Euklidova vzdálenost, která se počítá podle vzorce

$$D_{\mathrm{E}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{t=1}^{k} \left(x_{it} - x_{jt} \right)^2}, \tag{5}$$

kde x_{it} je hodnota t-té proměnné u i-tého objektu.

Výpočet této míry vzdálenosti mezi i-tým a j-tým objektem je založen na Pythagorově větě podle výše uvedeného vzorce. V literatuře se uvádí, že tato míra vzdáleností není vhodná pro případ, že jednotlivé proměnné, které charakterizují objekty, jsou velmi silně závislé.

Mezi další známé míry vzdálenosti lze zařadit například Manhattanskou, Minskovského, Mahalanobisovu či tětivovou vzdálenost. Jejich popis nalezne čtenář v příslušné literatuře.

Mezi nejznámější metody shlukování patří metody nejbližšího souseda, nejvzdálenějšího souseda, metoda průměrné vzdálenosti, centroidní, mediánová a Wardova.

Metoda nejbližšího souseda (jednoduché spojení) představuje nejstarší a nejjednodušší metodu. Mezi její nevýhody patří, že dochází k tzv. řetězení objektů. Výsledná klasifikace objektů do shluků obvykle neposkytuje nejlepší výsledky.

Metoda nejvzdálenějšího souseda (úplné spojení) odstraňuje problém řetězení a její výhodou je, že vytváří malé, kompaktní a dobře oddělené shluky.

Při **metodě průměrné vzdálenosti** nejsou výsledky ovlivněny extrémními hodnotami, jako je tomu například u metody nejbližšího a nejvzdálenějšího souseda.

Výhodou **Centroidní metody** je, že není významně ovlivňována odlehlými objekty.
U této metody se mohou objevit také tzv. *zmatečné shluky*.

Cílem **Mediánové metody** je snaha odstranit nedostatek centroidní metody, který byl popsán výše. V literatuře se uvádí, že „... rozdílné počty objektů u shluků způsobí rozdílnou váhu prvních dvou složek rekurzivního předpisu centroidní metody, a tak se stává, že vlastnosti malých shluků se ve výsledném sjednocení ztrácejí". Mediánová metoda je obdobou centroidní metody. Rozdíl spočívá v tom, že místo vzdáleností mezi centroidy shluků používá vzdálenost mezi mediány těchto shluků.

Wardova metoda (Wardova-Wishartova metoda) řeší princip shlukování odlišným způsobem než výše uvedené metody, které se zabývají optimalizací vzdáleností mezi jednotlivými shluky. Při této metodě se minimalizuje heterogenita shluků, tj. shluky se vytvářejí pomocí maximalizace vnitroshlukové homogenity. Měrou homogenity shluků je vnitroshlukový součet čtverců odchylek hodnot od průměru shluku. V literatuře se uvádí, že bývá používána ve spojení s druhou mocninou Euklidovy vzdálenosti definované výše.

2.2 Metody pro pevné C-shlukování

V této části jsou uvedeny metody shlukování do předem stanoveného počtu shluků, jejichž výsledkem je pevné přiřazení objektů do disjunktních shluků. Metody *C*-shlukování představují iterativní postupy, které začínají stanovením

počátečních centroidů, což představuje klíčový problém těchto metod shlukování. Podle typu zvoleného centroidu jsou rozlišovány například metody *C*-průměrů a *C*-medoidů.

Metoda *C*-průměrů je vhodná v případě, že jsou objekty charakterizovány pouze kvantitativními proměnnými. Jedná se o případ metody *C*-shlukování, kde jako centroidy shluků jsou využity vektory aritmetických průměrů.

Metoda *C*-medoidů je také vhodná pro případ, kdy jsou objekty charakterizovány kvantitativními proměnnými. Nejprve se provede počáteční rozdělení objektů do *r* shluků. Každý shluk je reprezentován *medoidem*, což je objekt, pro který platí, že průměrná vzdálenost k ostatním objektům v tomto shluku je minimální.

2.3 Metody pro fuzzy shlukování

Při fuzzy shlukové analýze je pro každý *i*-tý objekt a *h*-tý shluk určena *míra příslušnosti* u_{ih}, která představuje příslušnost, že daný objekt *i* je zařazen do *h*-tého shluku. Jedná se o postup, který patří k nehierarchickým metodám shlukování. Jeho výhodou je, že na rozdíl od výše uvedených technik může indikovat zařazení jednoho objektu do více shluků. Existuje mnoho metod, které mohou být aplikovány pro fuzzy shlukování. Mezi ně lze zařadit například algoritmus fuzzy *C*-průměrů, fuzzy *C*-medoidů či algoritmus FANNY.

2.4 Dvoukroková shluková analýza

Pro shlukování objektů může být využita tzv. **dvoukroková shluková analýza**, která je implementována v systému IBM SPSS. Tato metoda může být využita pro shlukování objektů, které jsou charakterizovány i samotnými nominálními proměnnými, případně proměnnými různých typů. Obecně však

jde o metodu určenou pro shlukování velkého množství objektů. Jako míra vzdálenosti, resp. nepodobnosti může být využita buď Euklidova míra (pouze pro případ kvantitativních proměnných) nebo věrohodnostní míra (pro proměnné různých typů).

2.5 Zpracování dat z oblasti shlukové analýzy

V následující části textu bude demonstrováno použití shlukové na konkrétním příkladu. K řešení bude opět nejprve využit systém SPSS, kde jsou tyto postupy implementovány. Následně bude provedena interpretace dílčích výsledků a závěrů.

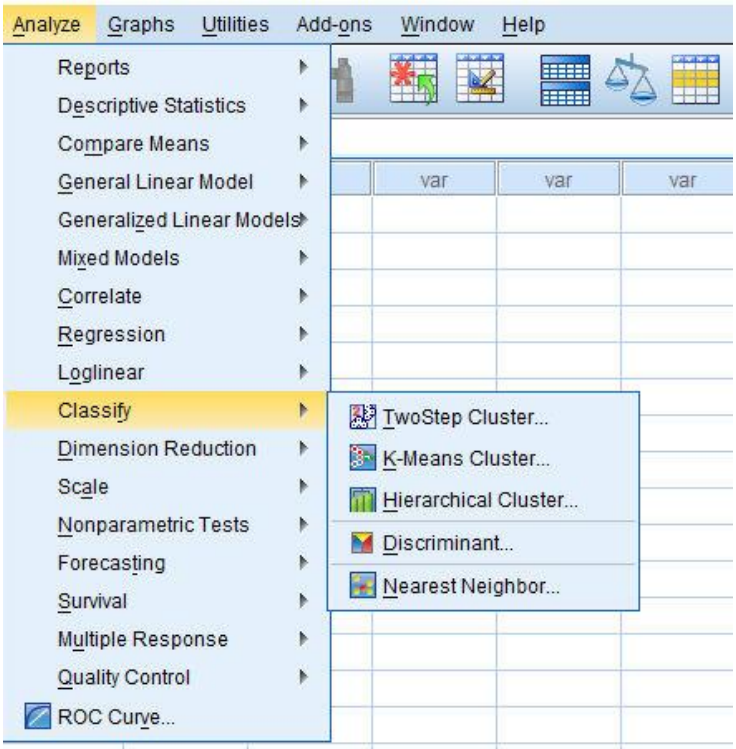
Příklad č. 1:

K demonstraci shlukové analýzy bude využit velmi známý soubor určený ke klasifikaci týkající se rostlin – Kosatců s názvem „*Iris*“. Soubor a podrobné informace k němu je možné najít na webové adrese https://archive.ics.uci.edu/ml/datasets/iris. Soubor určený ke klasifikaci obsahuje celkem 150 kosatců tří druhů (*Iris Setosa, Iris Versicolour, Iris Virginica*). Každý druh je zastoupen shodně 50 objekty a je charakterizován pomocí čtyř kvantitativních proměnných, jako jsou délka a šířka okvětního lístku a délka a šířka kališního lístku, vždy uváděno v centimetrech. Pomocí hierarchické shlukové analýzy proveďte klasifikaci jednotlivých objektů (Kosatců) a porovnejte se skutečnou příslušností do dané skupiny.

Řešení:

Vzhledem k faktu, že je úkolem klasifikovat jednotlivé objekty do skupin, úloha bude demonstrována pomocí shlukové analýzy. U té nebývá počet shluků (tříd) dopředu známý, nicméně, v našem případě tuto informaci

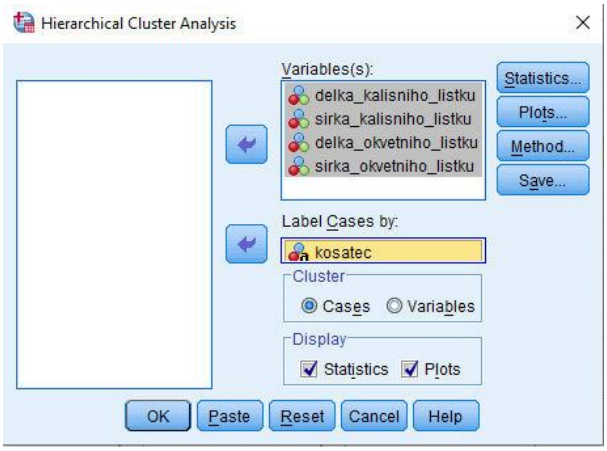
využijeme k následnému hodnocení metod shlukování. Aktivace shlukové analýzy v SPSS se provádí jednoduchým způsobem, který je uveden na obrázku 12.



Obrázek č. 12: Spuštění shlukové analýzy v SPSS

V danou chvíli je třeba zvolit skupinu metod shlukování. „*TwoStep Cluster*“ je dvoukroková shluková analýza, která je vhodná pro shlukování velmi rozsáhlých datových souborů nebo pro případ shlukování, kdy jsou objekty charakterizovány pomocí proměnných různých typů (kvalitativní, případně jejich kombinace s kvantitativními proměnnými). „*K-Means Cluster*“ je jeden ze způsobů nehierarchického způsobu shlukování.

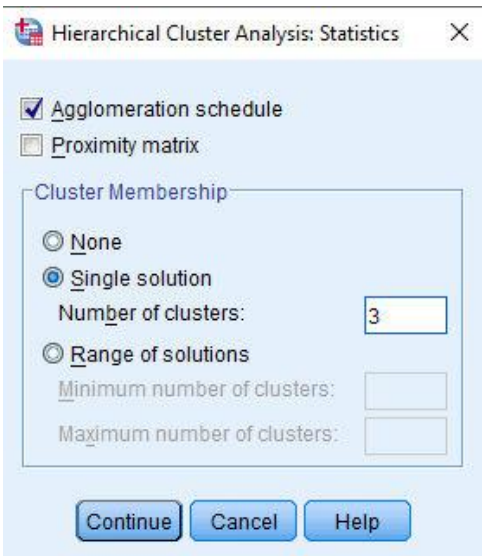
Vzhledem k faktu, že je v našem případě úkolem shlukovat pomocí hierarchických metod shlukování, provedeme tuto volbu tak, jak je patrné z obrázku 13.



Obrázek č. 13: Spuštění shlukové analýzy v SPSS

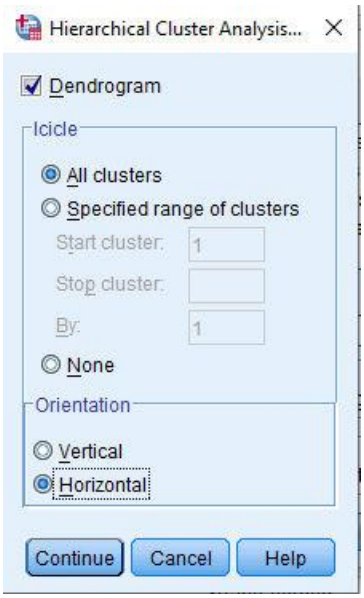
Již připravené a všechny čtyři charakterizující proměnné vložíme do okna „*Variable*(s)“, a protože u každého objektu známe jeho název, vložíme označení kosatec do „*Label Cases by*“. Dále ponecháme standardně vybrané zobrazení jak statistik (*Statistics*), tak grafů (*Plots*).

V záložce „*Statistics*“ je třeba zejména určit v části „*Cluster Membership*“, zda je dopředu známé zařazení do určitého počtu shluků „*Single solution*“ a nebo nikoliv „*None*“.



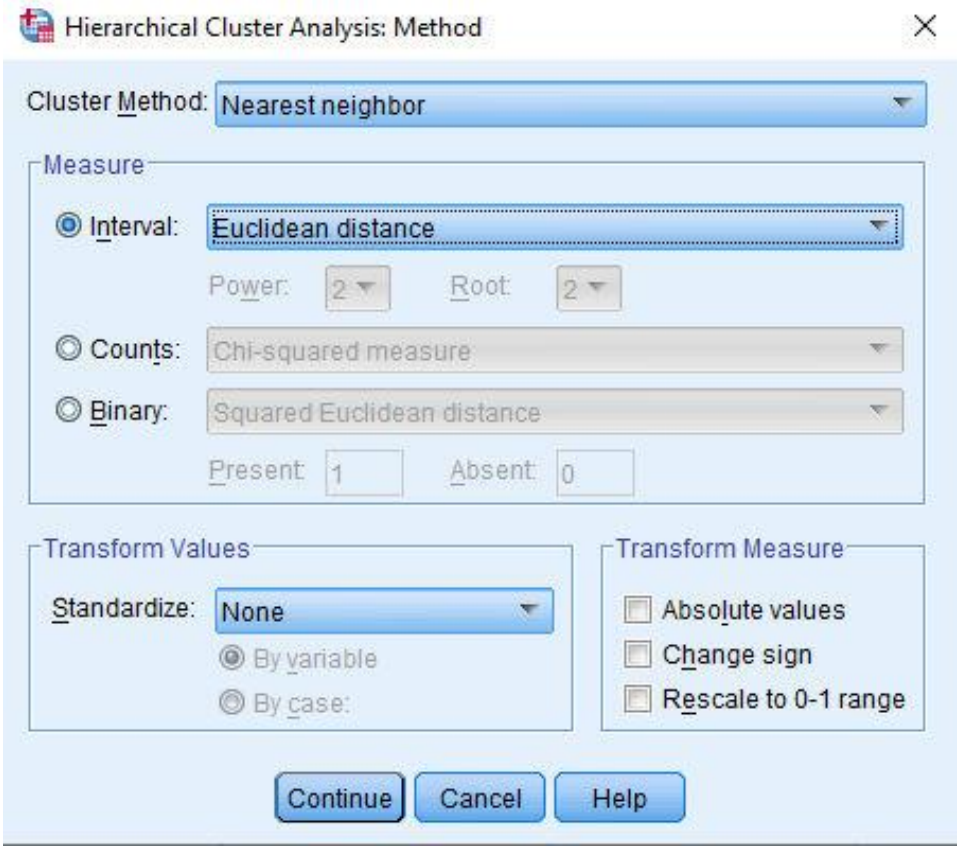
Obrázek č. 14: Nastavení počtu shluků v SPSS

Výběr grafu provedeme zaškrtnutím „*Dendrogram*“ v okně *Plot,* jak je uvedeno v obrázku 15.



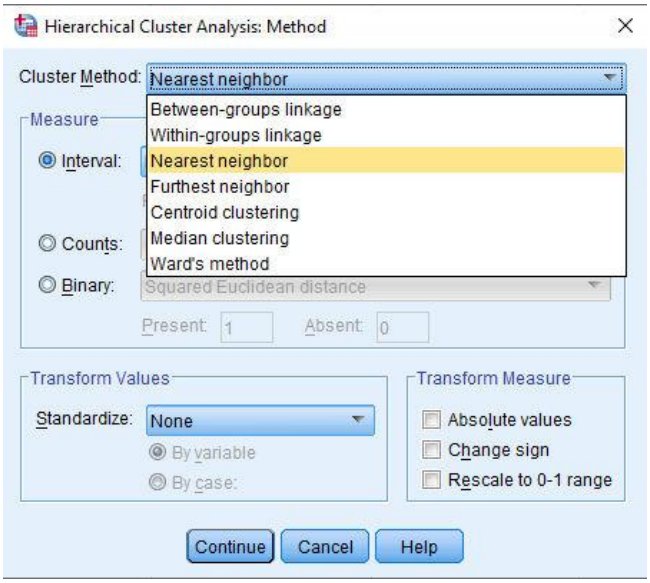
Obrázek č. 15: Nastavení počtu shluků v SPSS

Velmi důležitým krokem je nastavení metody shlukování a míry, která bude při shlukování použita. To se provádí výběrem záložky „*Cluster Method*“, jak je patrné z obrázku 16.



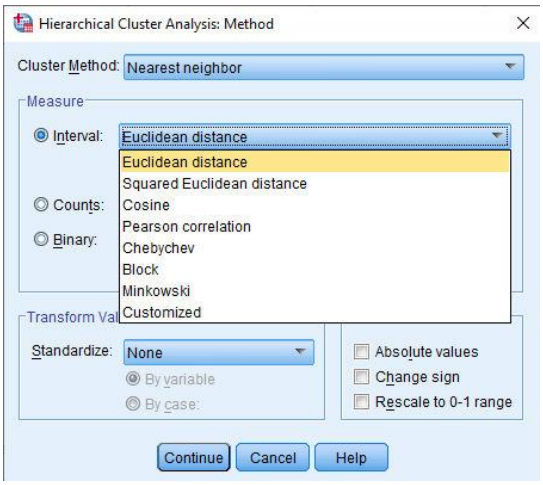
Obrázek č. 16: Nastavení metod shlukování v SPSS

Na výběr je celá řada metod shlukování, jako je například metoda nejbližšího souseda, nejvzdálenějšího souseda, centroidní, mediánová či Wardova metoda. V našem případě začneme nejstarší metodou, tedy metodou nejbližšího souseda.



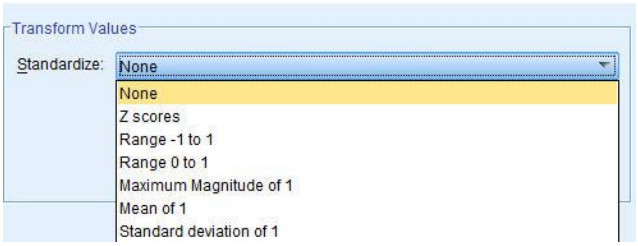
Obrázek č. 17: Nabízené metody shlukování v SPSS

Kromě metody shlukování je třeba také zvolit příslušnou míru, kterou daná metoda při shlukování využívá. Opět je na výběr celá řada, například měr vzdáleností, jako jsou Euklidova vzdálenost, případně její čtverec, kosinová míra, korelační koeficient, Čebyševova vzdálenost, atd. V našem příkladu využijeme velmi často používanou a oblíbenou Euklidovu míru vzdálenosti.



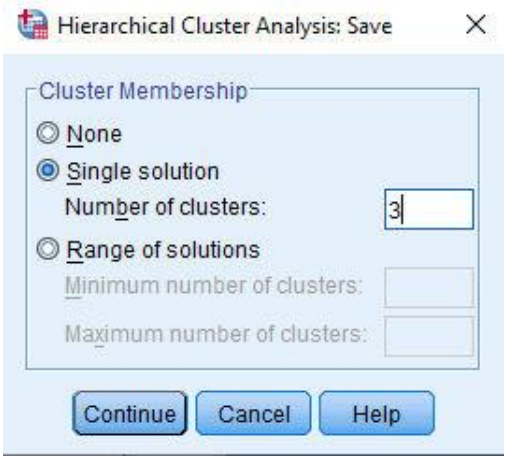
Obrázek č. 18: Nabízené míry při shlukování v SPSS

Pro případ nestejných měrných jednotek bývá doporučena transformace proměnných, kterých je v SPSS taktéž nabízena celá řada, jak je patrné z obrázku 19. Protože jsou všechny proměnné ve stejných měrných jednotkách, v našem příkladu nebude prováděna žádná transformace.



Obrázek č. 19: Nabízené transformace v SPSS

Pokud si budeme přát vyznačit do původní datové matice příslušnost objektů do shluků, kterou určí příslušná metoda shlukování, v záložce „*Save*“ vybereme „*Single Solution*“ a následně vyznačíme počet shluků, je-li dopředu znám. V našem případě požadujeme zařazení do jednoho ze tří známých shluků.



Obrázek č. 20: Volba zápisu příslušnosti objektu do shluku v SPSS

Po spuštění takto zadaných hodnot v SPSS je v první tabulce výstupu shrnuto, kolik objektů bylo shlukováno, pomocí jaké metody shlukování a jaká byla

využita míra vzdálenosti. Pokud by v datové matici byly chybějící hodnoty, tato skutečnost by byla také ve výstupu na obrázku 21 zachycena.

Case Processing Summary[a,b]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
150	100,0	0	,0	150	100,0

a. Euclidean Distance used

b. Single Linkage

Obrázek č. 21: Obecný výstup shlukování v SPSS

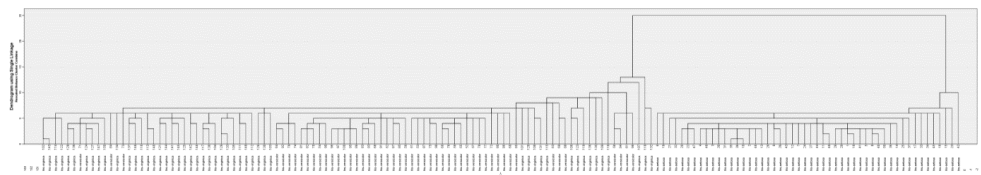
V další části výstupu je zachyceno, jak je patrné z obrázku 22, jak jsou postupně jednotlivé objekty (rostliny) zařazovány do shluků.

Cluster Membership

Case	3 Clusters
1:Iris-setosa	1
2:Iris-setosa	1
3:Iris-setosa	1
4:Iris-setosa	1
5:Iris-setosa	1
6:Iris-setosa	1
7:Iris-setosa	1

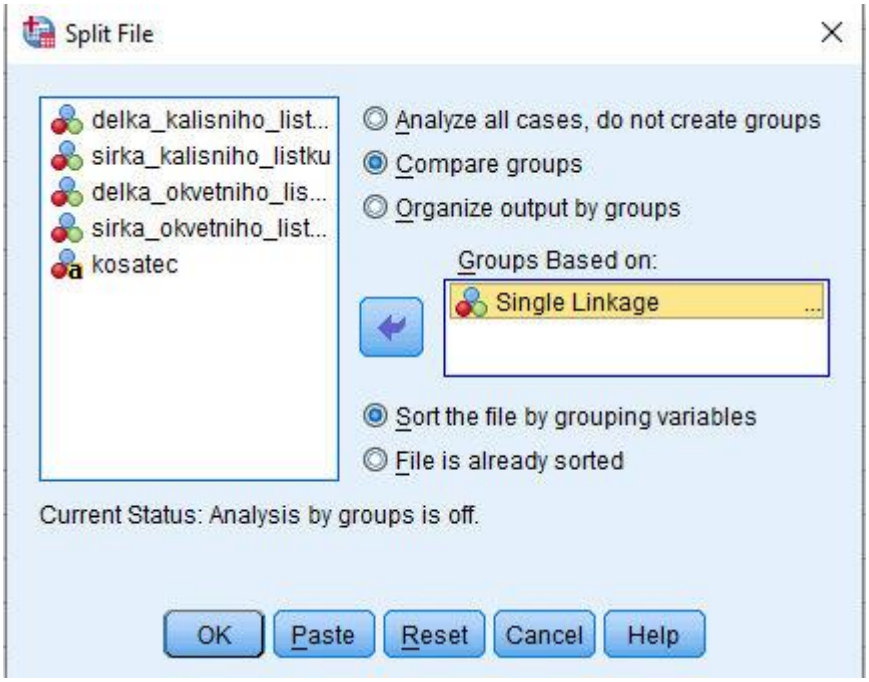
Obrázek č. 22: Přiřazení objektů do shluků v SPSS

Z obrázku 23 je patrný dendrogram, který jsme po zaškrtnutí volby získali. Je z něj patrné, jak jsou jednotlivé druhy postupně zařazovány do jednotlivých shluků.



Obrázek č. 23: Získaný dendrogram z SPSS

Pokud bychom chtěli získat charakteristiky jednotlivých získaných shluků, je třeba vyjádřit „štěpení“ analýz v kartě *Data*. Třídící faktor v našem případě bude vytvořená příslušnost ke shlukům určená metodou nejbližšího souseda „*Single Linkage*“. Protože chceme mít výstup v jedné tabulce, abychom jednotlivé skupiny mohli porovnat, vybereme možnost „*Compare groups*“. Pokud bychom chtěli mít informace o vytvořených shlucích v samostatných tabulkách, zvolili bychom „*Organize output by groups*“. To vše je patrné z obrázku 24.



Obrázek č. 24: Nastavení štěpení výstupů pro jednotlivé skupiny v SPSS

Z výstupu na obrázku 25 jsou zřejmé vybrané popisné charakteristiky pro jednotlivé shluky (třídy). Je zřejmé, že metoda nejbližšího souseda správně identifikovala první druh (iris-setosa). Ostatní dvě skupiny byly odděleny velmi špatně, protože do druhého shluku bylo navíc nesprávně zařazeno 48 rostlin ze třetího shluku. To může být způsobeno jednou z nevýhod této jednoduché, nejstarší metody shlukování s názvem „řetězení“.

Descriptive Statistics

Single Linkage		N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
1	delka_kalisniho_listku	50	1,50	4,30	5,80	5,0060	,35249	,124
	sirka_kalisniho_listku	50	2,10	2,30	4,40	3,4180	,38102	,145
	delka_okvetniho_listku	50	,90	1,00	1,90	1,4640	,17351	,030
	sirka_okvetniho_listku	50	,50	,10	,60	,2440	,10721	,011
	Valid N (listwise)	50						
2	delka_kalisniho_listku	98	2,80	4,90	7,70	6,2306	,63122	,398
	sirka_kalisniho_listku	98	1,60	2,00	3,60	2,8531	,30804	,095
	delka_okvetniho_listku	98	3,90	3,00	6,90	4,8724	,79894	,638
	sirka_okvetniho_listku	98	1,50	1,00	2,50	1,6673	,42445	,180
	Valid N (listwise)	98						
3	delka_kalisniho_listku	2	,20	7,70	7,90	7,8000	,14142	,020
	sirka_kalisniho_listku	2	,00	3,80	3,80	3,8000	,00000	,000
	delka_okvetniho_listku	2	,30	6,40	6,70	6,5500	,21213	,045
	sirka_okvetniho_listku	2	,20	2,00	2,20	2,1000	,14142	,020
	Valid N (listwise)	2						

Obrázek č. 25: Charakteristiky shluků získaných metodou nejbližšího souseda v SPSS

Velmi oblíbenou metodou shlukování, která řeší proces shlukování odlišným způsobem, jak bylo uvedeno výše (pomocí minimalizace vnitroshlukové a maximalizace mezishlukové variability), je Wardova metoda. Proto si její výsledky také stanovíme a vzájemně porovnáme. Postup spuštění je obdobný jako v předchozím případě, nyní však zvolíme Wardovu metodu a k ní vybereme čtverec Euklidovy vzdálenosti. Při praktickém navázání na předchozí příklad je třeba upozornit, že v tuto chvíli je třeba deaktivovat štěpení souboru a opět analyzovat všechny objekty současně. Následně je možné spustit proces shlukování.

Case Processing Summary[a,b]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
150	100,0	0	,0	150	100,0

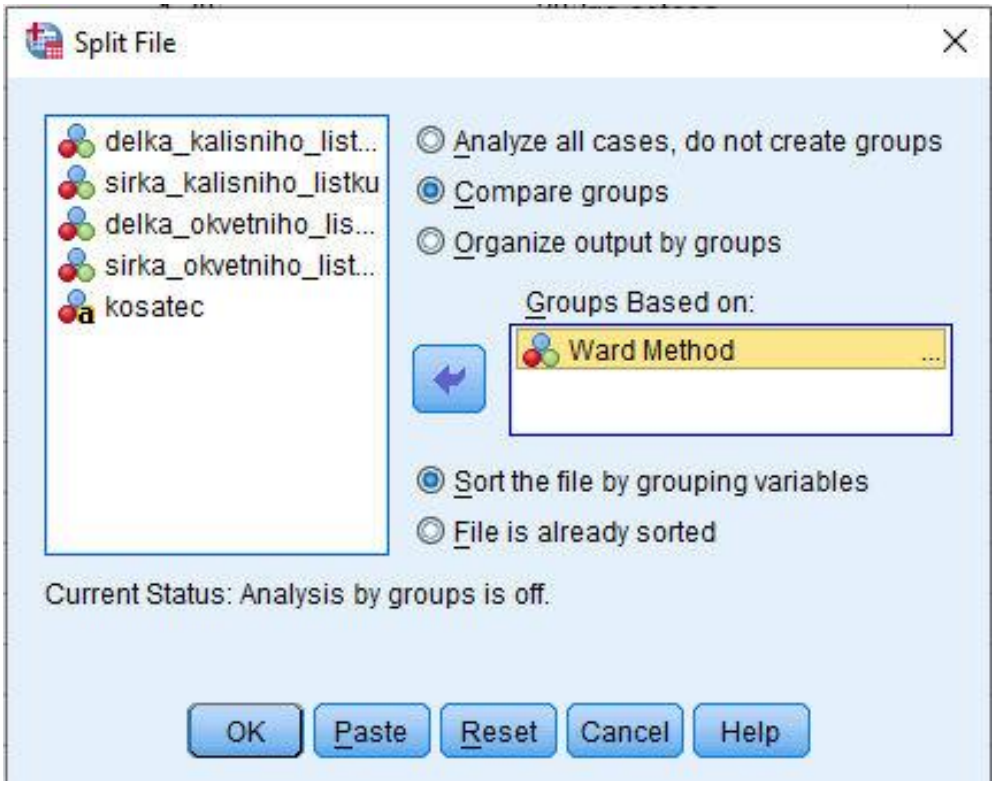
a. Squared Euclidean Distance used

b. Ward Linkage

Obrázek č. 26: Základní charakteristiky ve výstupu při použití Wardovy metody v SPSS

Výstupy, které bychom následně získali, budou vzhledově identické, avšak s rozdílnou klasifikací objektů.

Abychom mohli porovnat úspěšnost obou metod shlukování, je opět třeba správně nastavit štěpení souboru, tentokrát na základě výsledků Wardovy metody, viz obrázek 27.



Obrázek č. 27: Štěpení analýz pro Wardovu metodu v SPSS

Po spuštění popisné statistiky a nastavení vybraných charakteristik obdržíme ve výstupu tabulku s jednotlivými hodnotami tak, jak je patrné z obrázku 28.

Descriptive Statistics

Ward Method		N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
1	delka_kalisniho_listku	50	1,50	4,30	5,80	5,0060	,35249	,124
	sirka_kalisniho_listku	50	2,10	2,30	4,40	3,4180	,38102	,145
	delka_okvetniho_listku	50	,90	1,00	1,90	1,4640	,17351	,030
	sirka_okvetniho_listku	50	,50	,10	,60	,2440	,10721	,011
	Valid N (listwise)	50						
2	delka_kalisniho_listku	64	2,10	4,90	7,00	5,9203	,47616	,227
	sirka_kalisniho_listku	64	1,40	2,00	3,40	2,7516	,29546	,087
	delka_okvetniho_listku	64	2,60	3,00	5,60	4,4203	,52650	,277
	sirka_okvetniho_listku	64	1,40	1,00	2,40	1,4344	,29289	,086
	Valid N (listwise)	64						
3	delka_kalisniho_listku	36	1,70	6,20	7,90	6,8694	,49154	,242
	sirka_kalisniho_listku	36	1,30	2,50	3,80	3,0861	,28701	,082
	delka_okvetniho_listku	36	1,90	5,00	6,90	5,7694	,48037	,231
	sirka_okvetniho_listku	36	,90	1,60	2,50	2,1056	,24371	,059
	Valid N (listwise)	36						

Obrázek č. 28: Charakteristiky shluků získaných Wardovou metodou v SPSS

Z výstupu na obrázku 28 je zřejmé, že i v tomto případě byl první shluk (iris-setosa) bezezbytku správně určen a všech 50 rostlin je v něm obsaženo. Na rozdíl od metody nejbližšího souseda, je u dalších dvou zbývajících shluků správně klasifikováno (zařazeno) větší množství objektů, a tak je zřejmé, že Wardova metoda je při shlukování objektů mnohem úspěšnější.

Abychom porovnali výsledky obou metod shlukování se skutečnými charakteristikami jednotlivých shluků, vygenerujeme si popisné charakteristiky podle předem známého zařazení objektů shluků do jednotlivých druhů tak, jak je zřejmé z obrázku 29. Je patrné, že vypočtené charakteristiky druhu iris-setosa plně odpovídají skutečným charakteristikám, tzn. opravdu byly všechny objekty správně klasifikovány. Je dále zřejmé, že 14 rostlin iris-virginica bylo nesprávně zařazeno do shluku iris-versicolor. Tím, že jsme dopředu znali příslušnost objektů do shluků, nyní můžeme zhodnotit úspěšnost obou metod shlukování. Je zřejmé, že použití Wardovy metody vedlo k vysoké, 90,67% úspěšnosti při klasifikaci, což je možné považovat za velmi uspokojivý výsledek.

Descriptive Statistics

kosatec		N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Iris-setosa	delka_kalisniho_listku	50	1,50	4,30	5,80	5,0060	,35249	,124
	sirka_kalisniho_listku	50	2,10	2,30	4,40	3,4180	,38102	,145
	delka_okvetniho_listku	50	,90	1,00	1,90	1,4640	,17351	,030
	sirka_okvetniho_listku	50	,50	,10	,60	,2440	,10721	,011
	Valid N (listwise)	50						
Iris-versicolor	delka_kalisniho_listku	50	2,10	4,90	7,00	5,9360	,51617	,266
	sirka_kalisniho_listku	50	1,40	2,00	3,40	2,7700	,31380	,098
	delka_okvetniho_listku	50	2,10	3,00	5,10	4,2600	,46991	,221
	sirka_okvetniho_listku	50	,80	1,00	1,80	1,3260	,19775	,039
	Valid N (listwise)	50						
Iris-virginica	delka_kalisniho_listku	50	3,00	4,90	7,90	6,5880	,63588	,404
	sirka_kalisniho_listku	50	1,60	2,20	3,80	2,9740	,32250	,104
	delka_okvetniho_listku	50	2,40	4,50	6,90	5,5520	,55189	,305
	sirka_okvetniho_listku	50	1,10	1,40	2,50	2,0260	,27465	,075
	Valid N (listwise)	50						

Obrázek č. 29: Charakteristiky skutečných skupin kosatců v SPSS

3 Analýza (metoda) hlavních komponent

Mezi vícerozměrné statistické metody, které jsou velmi oblíbené, a tím pádem často používané, lze také zařadit tzv. analýzu (metodu) hlavních komponent, anglicky *Principal Component Analysis* (PCA). Jejím primárním cílem nejčastěji bývá uváděna redukce (zjednodušení) počtu proměnných takovým způsobem, aby došlo k co nejmenší ztrátě informace z původních proměnných. Jestliže má k tomuto účelu být vhodná právě analýza hlavních komponent, je třeba předpokládat vzájemnou závislost původních proměnných, kterou jsme v kapitole o regresní analýze označili jako multikolinearitu. Před jejím počátkem tedy v rámci počáteční analýzy je nutné stanovit korelační matici a z ní zjistit, zda jsou proměnné vzájemně korelované. Bez uvedeného předpokladu by použití této metody ztrácelo význam.

Princip této metody je možné popsat jako tvorbu lineární transformace původních proměnných na nové tak, že nově vytvořené proměnné jsou vzájemně nekorelované. Tyto nové proměnné se označují se jako tzv. *hlavní komponenty*. Transformace, které se využívá ke tvorbě proměnných se, podle (Meloun, 2005), nazývá jako Karhunen-Loevova nebo Hotellingova tranfsormace. Z myšlenky této metody vyplývá, že počet hlavních komponent je (výrazně) nižší, než je počet původních proměnných. Pokud by počet hlavních komponent byl stejný jako počet původních proměnných, použití metody by opět ztrácelo smysl. Základním atributem každé hlavní komponenty je charakteristika její variability (měnlivosti), kterou jsme popsali v prvním dílu naší knihy, a označili ji jako rozptyl. Při analýze hlavních komponent jsou jednotlivé komponenty seřazovány sestupně podle své důležitosti z hlediska podílu na vysvětlené variabilitě a představují jakýsi skrytý (latentní) atribut, který je hledán tak, aby informace z původních proměnných byla co nejvíce zachována. Na rozdíl od původních proměnných, které byly využité například

ve vícerozměrné regresní analýze, nejsou nové komponenty přímo pozorovatelné a jejich věcná interpretace není bezpodmínečně nutná. První hlavní komponenta vysvětluje (definuje) největší část variability původních proměnných. Druhá hlavní komponenta se snaží popsat největší část zbylé variability, která není zahrnuta v první komponentně. Důležitou vlastností je, že je kolmá na první hlavní komponentu. Třetí, a pak následně i další komponenty, se snaží vysvětlit největší část dosud nevysvětleného rozptylu v předchozích komponentách a opět platí, že jsou kolmé na předchozí komponenty. Při použití této metody bývá uváděno, že odlehlé a extrémní objekty je třeba odstranit. V případě, že jsou původní proměnné ve stejných měrných jednotkách a v případě, že mají alespoň přibližně stejnou variabilitu, je východiskem této metody výběrová kovarianční matice původních proměnných (náhodných veličin). Pokud tomu tak není, východiskem je následně korelační matice.

V praxi lze metodu hlavních komponent použít k samotnému snížení počtu původních proměnných (bez toho, aniž bychom ztratili velké množství informace), tedy ke snížení rozměru (dimenze) dané úlohy. Dalším vhodným a častým použitím bývá například grafická presentace vícerozměrných dat do prostoru roviny, opět bez velkého poklesu množství informace obsažené v původních proměnných. Neméně častým použitím je odstranění závažného problému, který jsme popsali u regresní analýzy, tedy problému multikolinearity o které jsme uvedli, že v krajním případě způsobuje až nepoužitelnost metody nejmenších čtverců, a tedy nemožnost odhadu parametrů regresního modelu. V neposlední řadě bývá tato metoda využívána v teorii jakosti.

Velmi důležitou součástí této úlohy je stanovení počtu hlavních komponent, do kterých mají být původní proměnné transformovány. Jak bylo uvedeno

výše, je zřejmé, že jejich počet by měl být nižší, než počet původních proměnných, jinak by úloha ztrácela smysl. Při stanovení jejich počtu se obvykle může vycházet z podílu vysvětlené variability původních proměnných. Jak uvádí (Meloun, 2005), obvykle se volí tolik hlavních komponent, aby procento vysvětlené variability bylo z intervalu 70 až 90 %. Počet komponent je možné stanovit také podle vlastních čísel, která bývají větší než 1, případně podle grafu vlastních čísel komponent, kde je nalezen zlom v grafu. V (Meloun, 2005) se dále uvádí, že bývají vyloučeny hlavní komponenty, pro které jsou vlastní vektory menší než 1, resp. je vhodnější vypouštět ty, jejichž vlastní čísla jsou nižší než 0,7. Při praktických úlohách bývá zvykem stanovit graf, tzv. *Scree plot,* v němž se na osu *X* nanáší hodnota příslušné komponenty (s maximálním počtem rovným počtu původních proměnných) a na osu *Y* se nanáší hodnota příslušného vlastního čísla.

Samotné parametry u jednotlivých hlavních komponent není možné porovnávat. Aby jim bylo možné přiřadit logický význam, stanovují se tzv. *komponentní zátěže*, což jsou korelační koeficienty, které vyjadřují míru lineární závislosti hlavních komponent s původními proměnnými.

Vzhledem k tomu, že výstup analýzy hlavních komponent je mnohdy využíván k další analýze (například k již zmíněné regresi či ke shlukové analýze), bývá vhodné určit pro každé z *n* pozorování, hodnotu komponent, které se označují jako tzv. *komponentní scóre*.

Pokud bychom použili metodu hlavních komponent v situaci, kdy proměnné nejsou vzájemně lineárně závislé, výsledné řešení v podobě hlavních komponent by odpovídalo původním proměnným, a tedy nedošlo k žádné redukci proměnných.

3.1 Zpracování dat z oblasti analýzy hlavních komponent

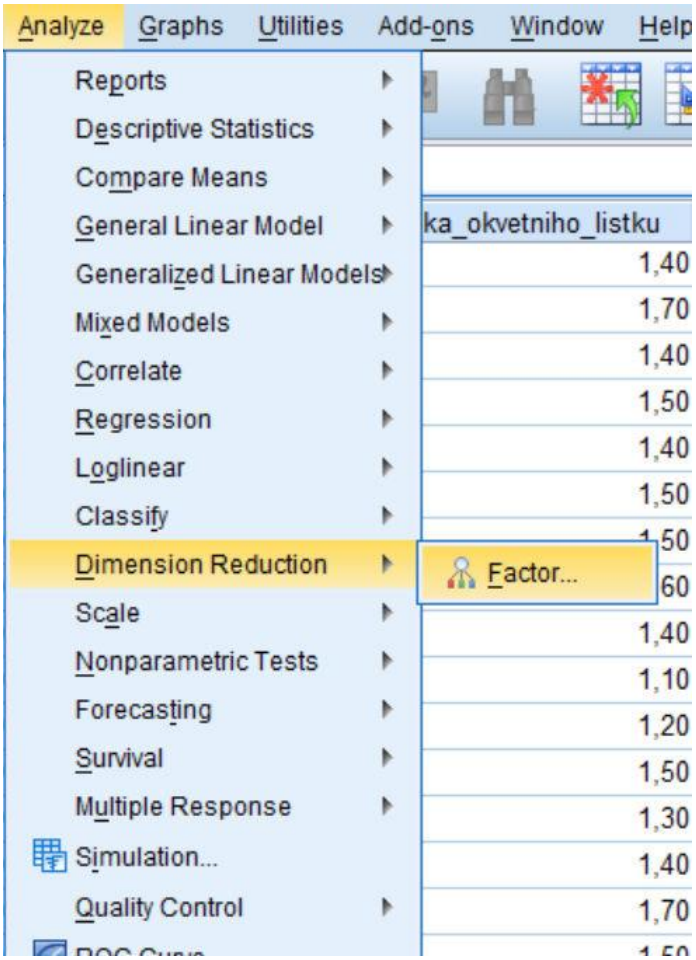
V následující části textu bude demonstrováno použití metody hlavních komponent na konkrétním příkladu. K řešení bude opět nejprve využit systém SPSS, kde jsou tyto postupy implementovány. Následně bude provedena interpretace dílčích výsledků a závěrů.

Příklad č. 1:

K demonstraci metody hlavních komponent, stejně jako v případě shlukové analýzy, bude využit velmi známý soubor Kosatců s názvem „*Iris*“. Soubor sice neobsahuje velké množství proměnných (obsahuje pouze 4), avšak z důvodu provázanosti a pochopení souvislostí jej využijeme i zde. Pro připomenutí uvádíme, že soubor obsahuje celkem 150 objektů.

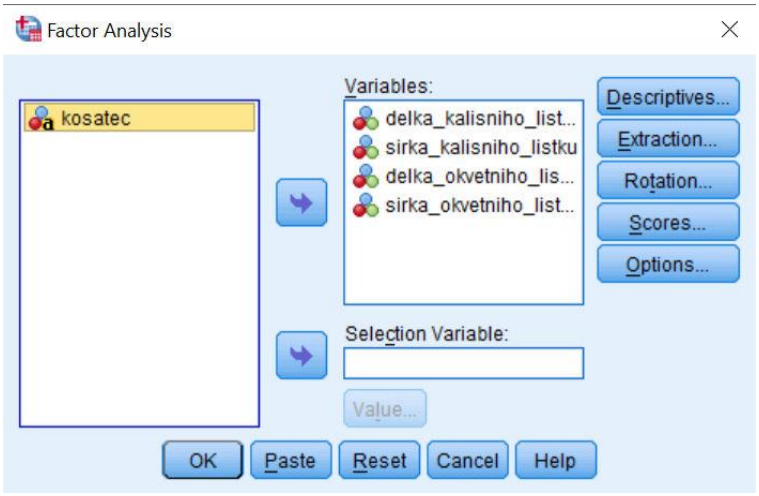
Řešení:

Spuštění metody hlavních komponent v SPSS se provádí způsobem, který je patrný z obrázku 30.



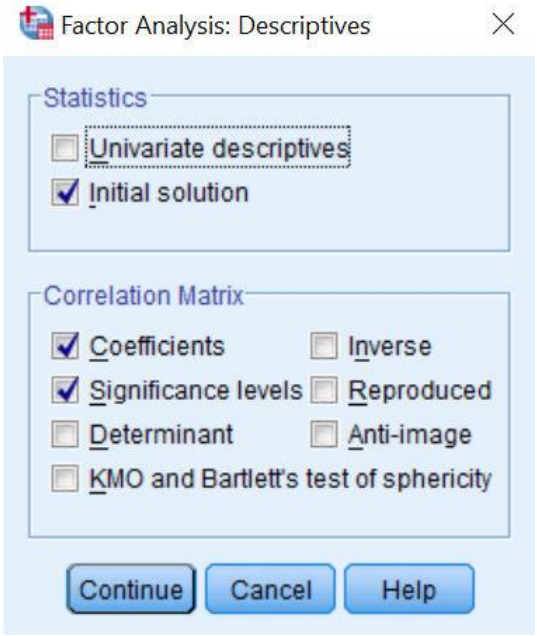
Obrázek č. 30: Spuštění metody hlavních komponent analýzy v SPSS

Jak je z obrázku 30 patrné, spuštění metody hlavních komponent se provádí přes výběr tzv. faktorové analýzy (viz následující kapitola). Po jejím výběru je třeba definovat původní proměnné, které do analýzy hlavních proměnných vstupují. To je zřejmé z obrázku 31.



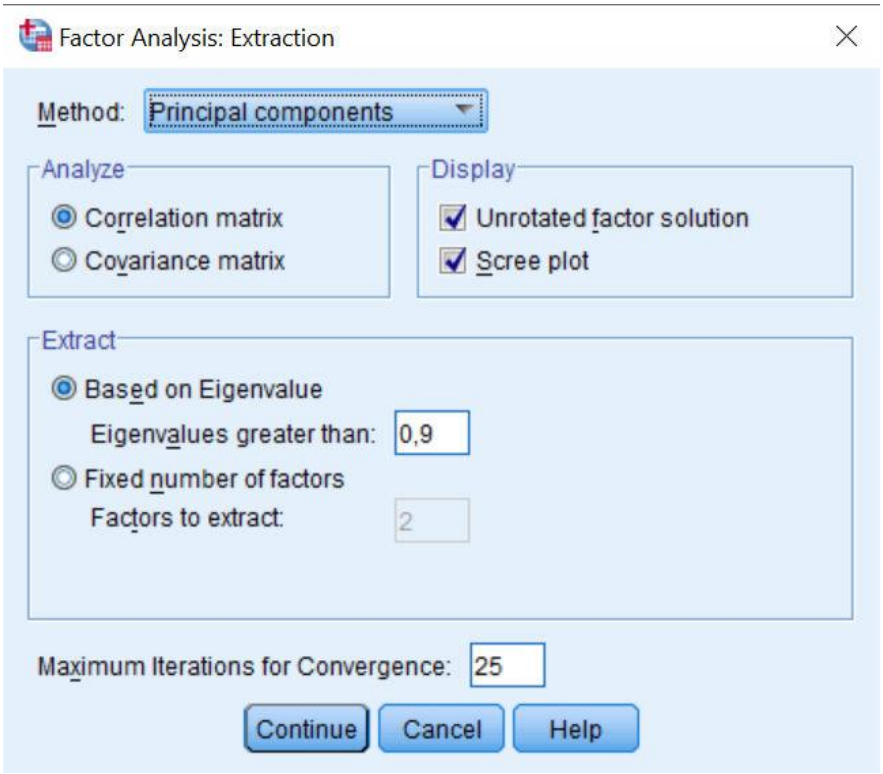
Obrázek č. 31: Spuštění metody hlavních komponent analýzy v SPSS

V záložce *Descriptives* si můžeme nechat vygenerovat korelační matici mezi původními proměnnými a také příslušné testy pro ověření statistické významnosti daných závislosti, jak je patrné z obrázku 32.



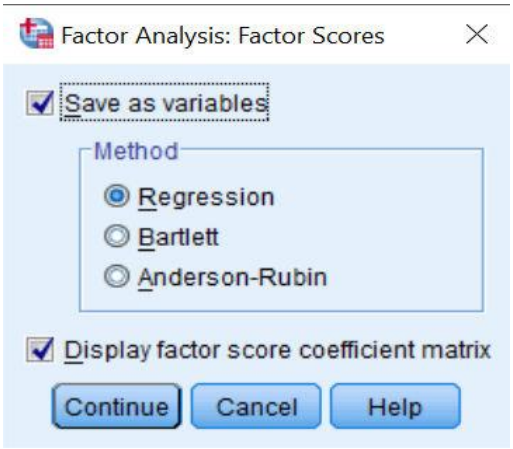
Obrázek č. 32: Spuštění metody hlavních komponent analýzy v SPSS

V záložce *Extraction* vybereme metodu hlavních komponent (*Principal components*), dále si necháme vygenerovat graf označený jako *Scree plot* a v části *Exact* je třeba vyznačit hodnotu vlastních čísel, která bude hraniční pro tvorbu hlavních komponent. V našem případě zvolíme například číslo 0,9.



Obrázek č. 33: Spuštění metody hlavních komponent analýzy v SPSS

V případě, že chceme do původní datové matice přidat nové sloupce, jejichž počet bude odpovídat počtu nově vytvářených hlavních komponent, zvolíme v záložce *Scores* položku *Save as variables*, případně *Display factor score coefficient matrix*.



Obrázek č. 34: Spuštění metody hlavních komponent analýzy v SPSS

Po spuštění procedur podle výše uvedeného způsobu získáme následující výstupy, které nalezneme v příslušném výstupovém okně SPSS. Z obrázku 35 je patrná výběrová korelační matice mezi původními proměnnými a *p*-hodnoty příslušných testů. Vyplývá z ní vysoká velmi silná přímá závislost mezi délkou kališních lístku a délkou okvětního lístku a šířkou okvětního lístku. Statistická významnost je prokázána na všech rozumných hladinách významnosti (0,01; 0,05, atd.), a tak je použití metody hlavních komponent možné považovat za rozumné.

Correlation Matrix

		delka_kalisniho_listku	sirka_kalisniho_listku	delka_okvetniho_listku	sirka_okvetniho_listku
Correlation	delka_kalisniho_listku	1,000	-,109	,872	,818
	sirka_kalisniho_listku	-,109	1,000	-,421	-,357
	delka_okvetniho_listku	,872	-,421	1,000	,963
	sirka_okvetniho_listku	,818	-,357	,963	1,000
Sig. (1-tailed)	delka_kalisniho_listku		,091	,000	,000
	sirka_kalisniho_listku	,091		,000	,000
	delka_okvetniho_listku	,000	,000		,000
	sirka_okvetniho_listku	,000	,000	,000	

Obrázek č. 35: Výstup metody hlavních komponent analýzy v SPSS

Z obrázku 36 jsou patrná vlastní čísla a podíly vysvětlené variability pomocí dané komponenty. Je zřejmé, že nejvyšší počet hlavních komponent je čtyři (odpovídá počtu původních proměnných). Dále je z výstupu zřejmé kumulativní procento vysvětlené variability. V pravé části tabulky je zřejmý „výsledek“ pro vybraný počet hlavních komponent, který v tomto případě odpovídá počtu 2 (hraniční hodnota u vlastních čísel byla v našem případě zvolena 0,9). Z výstupu je zřejmé, že pomocí dvou komponent bylo vysvětleno 95,80 % variability. Pokud by byly zvoleny čtyři komponenty (což jistě postrádá smysl), byla by vysvětlena všechna variabilita původních proměnných.

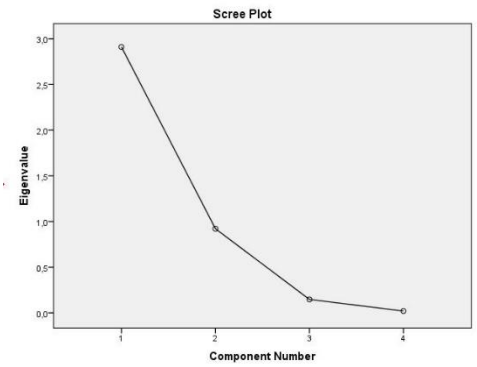
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,911	72,770	72,770	2,911	72,770	72,770
2	,921	23,031	95,801	,921	23,031	95,801
3	,147	3,684	99,485			
4	,021	,515	100,000			

Extraction Method: Principal Component Analysis.

Obrázek č. 36: Výstup metody hlavních komponent analýzy v SPSS

Z grafu na obrázku 37 jsou patrné hodnoty vlastních čísel (na ose Y) a příslušné komponenty (na ose X).



Obrázek č. 37: Výstup metody hlavních komponent analýzy v SPSS

Pro výsledný počet komponent je, jak je patrné z obrázku 38, možné vyčíst hodnoty korelačních koeficientů původních proměnných k nově vytvářeným komponentám (*komponentní zátěže*). Je zřejmé, že první tři proměnné (délka a šířka okvětního lístku a délka kališního lístku) vykazují velmi silné korelace s první komponentou (hodnota korelačního koeficientu neklesá pod hodnotu 0,89) a proměnná šířka kališního lístku vykazuje velmi silnou korelaci s druhou komponentou (hodnota korelačního koeficientu je 0,888).

Component Matrix[a]

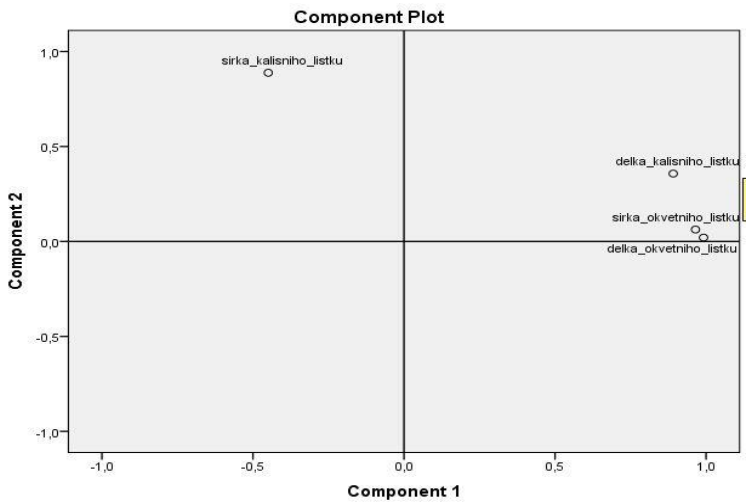
	Component	
	1	2
delka_okvetniho_listku	,992	,020
sirka_okvetniho_listku	,965	,063
delka_kalisniho_listku	,891	,357
sirka_kalisniho_listku	-,449	,888

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Obrázek č. 38: Výstup metody hlavních komponent analýzy v SPSS

Grafické zachycení vazby původních proměnných k vytvořeným komponentám je zřejmé z obrázku 39.



Obrázek č. 39: Výstup metody hlavních komponent analýzy v SPSS

Odhadnuté parametry příslušené lineární kombinace pro jednotlivé původní proměnné u výsledných komponent je možné vyčíst z tabulky na obrázku 40. Ty je dále možné využít pro získání komponentních scórů u jednotlivých objektů. Pokud bychom postupně dosazovali hodnoty původních proměnných pronásobených danými parametry u dané komponenty z tabulky 40, získali bychom příslušná komponentní scóre, která jsou naznačena na obrázku 41.

Component Score Coefficient Matrix

	Component	
	1	2
delka_kalisniho_listku	,306	,388
sirka_kalisniho_listku	-,154	,964
delka_okvetniho_listku	,341	,022
sirka_okvetniho_listku	,332	,068

Extraction Method: Principal Component Analysis.
Component Scores.

Obrázek č. 40: Výstup metody hlavních komponent analýzy v SPSS

Výsledná komponentní scóre (získaná postupných dosazováním parametrů z obrázku 40), jsou zachycena na obrázku 41 a splňují podmínku vzájemné lineární nezávislosti, a proto je možné jejich využití pro další vícerozměrné analýzy, například pro vícerozměrnou regresní analýzu či shlukovou analýzu, jak bylo uvedeno výše.

FAC1_1	FAC2_1
-1,38329	-,33071
-1,34605	-,59746
-1,39545	,70068
-1,20955	1,57686
-1,42871	,07743
-1,30495	,25712
-1,36810	-1,13720
-1,27856	-,46586
-1,26385	1,11171
-1,35959	,16468
-1,29924	-,73635
-1,54205	-,97431
-1,28067	1,96256
-1,31524	2,82692
-1,28678	1,57188
-1,27944	,53405
-1,10646	1,48607
-1,36693	1,20250
-1,11843	,44700

Obrázek č. 41: Výstup metody hlavních komponent analýzy v SPSS

4 Faktorová analýza

Faktorová analýza je, stejně jako metoda hlavních komponent, která byla popsána v předchozí kapitole, primárně zaměřena na snížení dimenze vícerozměrné statistiky, tedy na zjednodušení rozměru úlohy. I v tomto případě se předpokládá, že existuje velké množství kvantitativních proměnných, které chceme nahradit mnohem menším počtem nových proměnných tak, abychom ztratili co nejmenší množství informace. Hlavním předpokladem použití této metody je, stejné jako u metody hlavních komponent, že existují významné korelace (multikolinearita) mezi původními proměnnými. Je tedy zřejmé, že před samotným zahájením této metody je opět vhodné realizovat průzkumovou analýzu dat a zjistit korelace mezi původními proměnnými, jinak by použití této metody ztrácelo smysl. Faktorová analýza bývá považována jako rozšíření metody hlavních komponent, přičemž, na rozdíl od ní, se snaží o vysvětlení existujících korelací mezi původními proměnnými. V tomto případě se z původních proměnných pomocí lineárních kombinací stanou nové proměnné (tzv. *faktory*), které také nejsou v praxi přímo měřitelné. Nové faktory je pak také následně možné využít pro další analýzy, protože jsou opět vzájemně nezávislé. Jednotlivé parametry modelu dané lineární kombinace původních proměnných se nazývají *faktorové zátěže*. Mezi nevýhodu této metody je, že její výsledek nemusí přinášet jednoznačné řešení a je poměrně subjektivní a výsledná interpretace, jak se uvádí například v (Stankovičová, 2007), může být mlhavá. Faktorová analýza částečně odstraňuje některé problémy metody hlavních komponent, avšak kromě výše uvedeného má nedostatky, které spočívají zejména ve vysoké subjektivitě a nejednoznačnosti řešení. I při použití této metody je potřeba určit počet faktorů, do kterých budou původní proměnné transformovány.

Při matematickém vyjádření modelu faktorové analýzy se opět předpokládá, že máme n objektů charakterizovaných pomocí k původních proměnných, které mají vícerozměrné rozdělení s k-členným vektorem středních hodnot a kovarianční maticí hodnosti k. Ve výsledku pak model předpokládá, že existuje celkem q skrytých, neměřitelných společných faktorů, kterých je výrazně méně, než je počet původních proměnných. Při aplikaci této metody hledáme pro každou původní zjištěnou proměnnou takový „funkční předpis“, kde odhadujeme hodnoty tzv. *faktorových zátěží*, které popisují vliv i-tého faktoru na j-tou původní proměnnou. Jak uvádí (Stankovičová, 2007), jednotlivé faktorové zátěže představují regresní koeficienty mezi původními skutečně naměřenými proměnnými a novými nepozorovatelnými faktory. Dále se uvádí, že při splnění určitých podmínek se jedná o kovariance (míry závislostí) mezi nimi. Pokud jsou původní proměnné vyjádřeny ve stejných měrných jednotkách, dílčí faktorové zátěže je možné interpretovat jako příspěvek i-tého odhadovaného faktoru na j-tou vysvětlovanou proměnnou. Pokud bychom využili proces normování, tj. nepracovali bychom s hodnotami původních proměnných, ale s normovanými daty (od původních hodnot bychom odečetli jejich střední hodnotu a vydělili bychom směrodatnou odchylkou), odhadnuté faktorové zátěže by pak představovaly korelační koeficienty původních proměnných a nově vytvořených faktorů. Stejně jako v případě metody hlavních komponent, jsou jednotlivé faktory nezávislé se stejným pravděpodobnostním rozdělením. Takovýto model bývá označován jako ortogonální faktorový model.

Při formulaci ortogonálního faktorového modelu, za určitých předpokladů, které jsou uvedeny například ve (Stankovičová, 2007), je možné vyjádřit určité důsledky:

a) Rozptyl každé *j*-té původní proměnné je možné vyjádřit jako součet tzv. *komunalit* (rozptyl *j*-té původní proměnné, který je vysvětlen pomocí *q* nových faktorů) a reziduálního rozptylu *j*-té proměnné, který se nepodařilo vysvětlit pomocí *q* faktorů.
b) Kovarianci mezi libovolnou dvojicí původních proměnných je možné vyjádřit pomocí faktorových zátěží.

Komunalita *j*-té proměnné u ortogonálního faktorového modelu vzniká jako součet čtverců jednotlivých faktorových zátěží (vah) získaných odhadem z faktorového modelu. V softwarových produktech bývá uvedeno více způsobů, jak získávat odhady parametrů faktorového modelu. Jedna z nich, která bývá velmi často v různých softwarových produktech (včetně SPSS) implementována, je zmíněna v předchozí kapitole. Jedná se o metodu hlavních komponent. Jak uvádí různí autoři, každá metoda přináší jiné výsledky a nelze jednoznačně říci, která metoda přináší nejlepší řešení. Jak uvádí (Stankovičová, 2007) pro dostatečně velké vzorky a pro vysoký počet vysvětlujících proměnných poskytují metody podobné výsledky.

Důležitým vstupem před započetím samotného procesu faktorové analýzy je stanovení počtu faktorů *q*, které budou z původních proměnných vytvářeny. Často bývá uváděno, že před počátkem faktorové analýzy je vhodné začít metodou hlavních komponent a určit počáteční odhad počtu faktorů. Výsledné řešení však bývá často obtížně interpretovatelné, a tak se uvádí, že je vhodné provádět tzv. *rotaci* (též transformaci) *faktorů*. Cílem je získat takové řešení, aby ve výsledku bylo co nejvíc faktorových zátěží (vah) blízkých nule a zároveň co nejvíc zbývajících (ostatních) vah blízké jedné. Rotace může být ortogonální (pravoúhlá či kolmá) či kosoúhlá (šikmá). Jak uvádí (Stankovičová, 2007), ortogonální rotace vede k řešení s nekorelovanými faktory a kosoúhlá rotace vede k získání výsledných závislých faktorů. Oba

postupy jsou různými autory kritizovány i vychvalovány. Při stanovení počtu faktorů je vhodné vycházet i z teoretických aspektů, a to i s ohledem na následnou interpretaci, která bývá někdy požadována, a volit jejich počet tak, aby došlo k co největšímu vysvětlení celkového rozptylu. Kromě cíle v podobě „interpretace" je vhodné stanovit pro každou jednotku tzv. *faktorová skóre*, což jsou hodnoty nově vytvořených, přímo neměřitelných proměnných. Tyto hodnoty pak následně mohou vstupovat do dalších analýz, jak již bylo uvedeno, například do vícerozměrné regresní analýzy či do shlukové analýzy. Metody pro odhad faktorových skórů jsou různé. Jak uvádí (Stankovičová, 2007), může jít například o vícenásobnou regresní metodu, Bartlettovu metodu či Harmanovu metodu, atd. Při aplikaci faktorové analýzy však bývá vhodné určit, zda původní data, která máme k dispozici, jsou pro faktorovou analýzu vhodná. K tomu existují různé způsoby, například stanovení KMO statistiky (*Kaiser´s Measure of Sampling Adequancy*), která vyjadřuje celkovou míru vhodnosti dat pro faktorovou analýzu. Jak uvádí (Stankovičová, 2007), je vhodné, aby daná data splňovala podmínku, že jejich KMO statistika je vyšší než 0,8. Hodnoty přibližně rovné číslu 0,6 jsou uváděny jako „v toleranci". Dalším možností, jak ověřit vhodnost dat, je například Bartlettův test, jehož testovaná hypotéza tvrdí nevhodnost dat pro faktorovou analýzu.

4.1 Zpracování dat z oblasti faktorové analýzy

V této textu si ukážeme aplikaci faktorové analýzy na praktickém datovém souboru. K řešení bude opět využit systém SPSS, kde jsou tyto postupy vhodně implementovány.

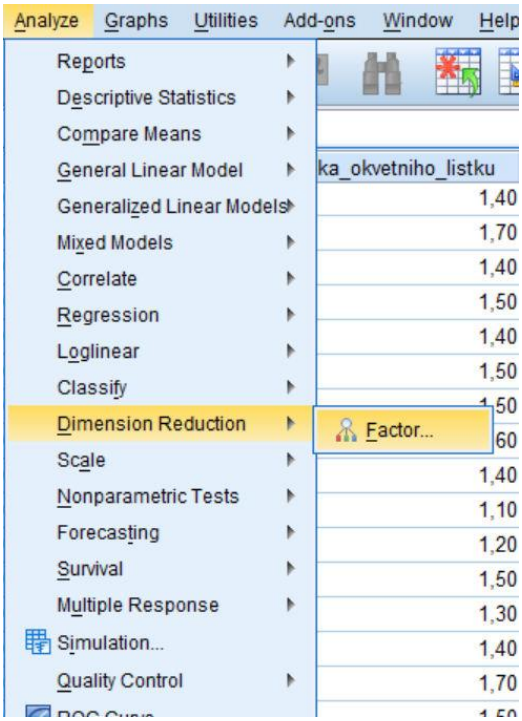
Příklad č. 1:

K demonstraci faktorové analýzy opět použijeme soubor, který jsme využívali v předchozích kapitolách, tj. soubor Kosatců s názvem „*Iris*". Jak je již známo,

soubor obsahuje čtyři proměnné a každé ze tří skupin je stejně velká a obsahuje 50 objektů. Celkový počet objektů je, jak jsme si již uvedli výše, 150.

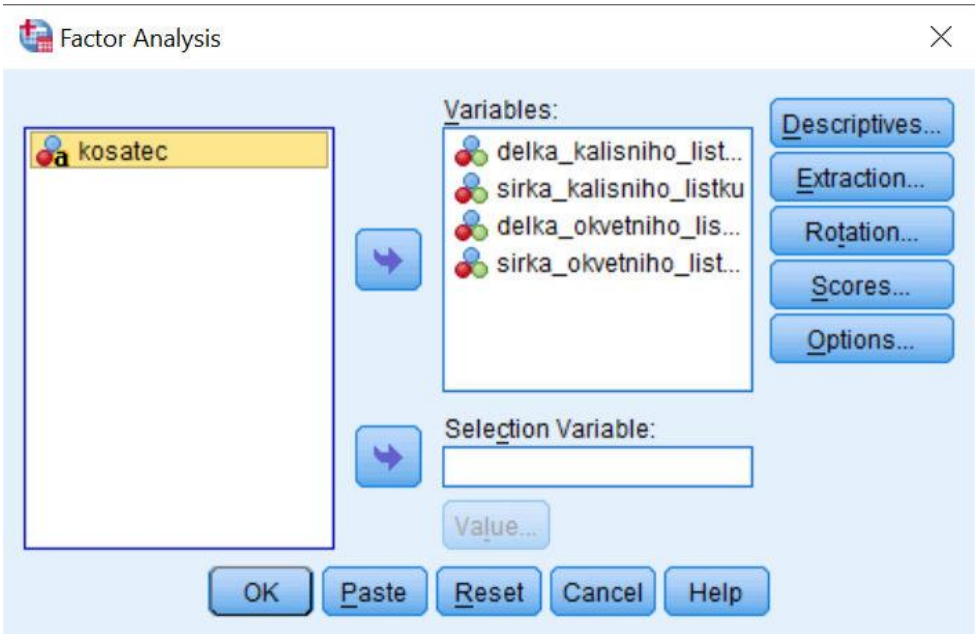
Řešení:

Spuštění faktorové analýzy v SPSS je patrné z obrázků 42 až 47.



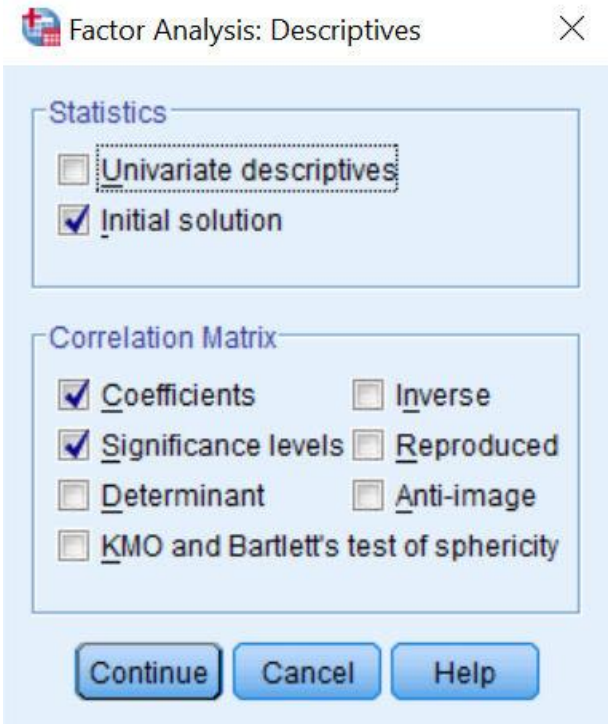
Obrázek č. 42: Spuštění faktorové analýzy v SPSS

Z obrázku 42 je zřejmé, kde nalezneme proceduru faktorové analýzy. V dalším kroku, který je zřejmý z obrázku 43, je nutné vybrat původní proměnné, na jejichž základě budou vytvářeny nové faktory. V našem případě zvolíme všechny čtyři číselné charakteristiky Kostaců, tedy délku a šířku kališního lístku, délku a šířku okvětního lístku.



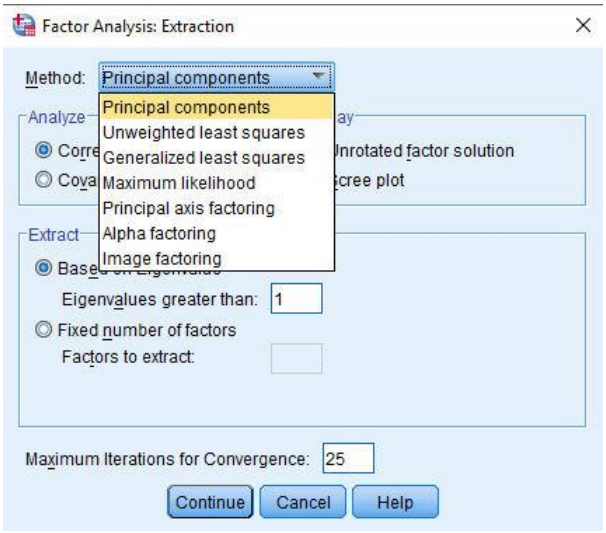
Obrázek č. 43: Spuštění faktorové analýzy v SPSS

Popisné charakteristiky, stejně jako v předchozí části textu věnované metodě hlavních komponent, získáme výběrem a nastavením podle obrázku 44. Pokud bychom chtěli získat představu o vhodnosti použití faktorové analýzy pro daná konkrétní data, můžeme zvolit KMO statistiku, případně další typ testu nazvaný *Bartlett´s test of sphericity*.



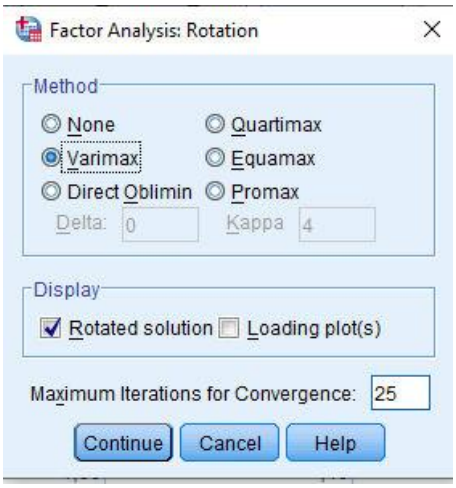
Obrázek č. 44: Spuštění faktorové analýzy v SPSS

V dalším okně zobrazeném na obrázku 45 označeném „*Extraction*“ zvolíme metodu, s jejíž pomocí bude model vytvářen. Zároveň je třeba nastavit pravidlo pro počet vytvářených faktorů. V tomto případě, na rozdíl od předcházející kapitoly, můžeme zvolit například hraniční hodnotu vlastního čísla pro tvorbu faktorů 1. Druhou možností je vybrat fixně počet vytvářených faktorů, například s ohledem na teoretické znalosti dané problematiky, v části označené jako „*Fixed number of factors*“.



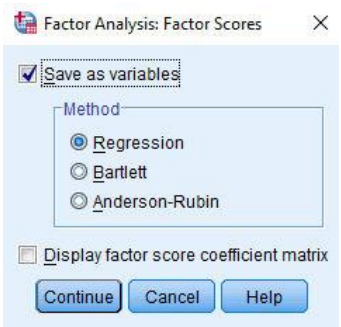
Obrázek č. 45: Spuštění faktorové analýzy v SPSS

V okně „*Rotation*“, jak je patrné z obrázku číslo 46, můžeme nastavit výše zmíněnou rotaci faktorů pomocí vybraných metod s cílem získání takových výsledků, které nám pomohou lépe interpretovat získané výsledky. To je samozřejmě smysluplné v případě, že získáme alespoň dva výsledné faktory.



Obrázek č. 46: Spuštění faktorové analýzy v SPSS

Pokud chceme, aby ve výsledné datové matici byly zachyceny hodnoty odhadnutých faktorových scórů pro jednotlivé objekty, vybereme možnost „*Save as variables*“ podle návodu na obrázku 47. K tomu jsou opět k dispozici různé metody. My si zvolíme první označenou, tedy metodu využívající regresní analýzu. Z okna je patrné, že existuje i například Bartlettova metoda.



Obrázek č. 47: Spuštění faktorové analýzy v SPSS

Pokud postupujeme podle výše zvolených obrázků, po spuštění faktorové analýzy získáme výstupy, které jsou zřejmé z obrázků 48 a 49.

Z obrázku 48 jsou patrné hodnoty jednotlivých charakteristik a parametrů. Z výstupu je zřejmé, že výsledný vytvářený faktor, který splňuje zvolenou podmínku vlastních čísel větších než jedna, je pouze jediný. Z tohoto důvodu také nebyla využita rotace faktorů. Z první tabulky jsou k dispozici hodnoty výše popsaných komunalit, které představují hodnoty rozptylů pro dané původní proměnné vysvětlené pomocí q nových faktorů (v našem případě s pomocí jediného faktoru). Pomocí výsledného faktoru, jak je patrné z druhé tabulky na obrázku 48, se podařilo vysvětlit 72, 77 % celkové variability. Pokud bychom zvolili podmínku vlastních čísel stejně jako v předchozí kapitole, počet vytvářených faktorů by byl roven hodnotě dva a došlo by ke zvýšení podílu vysvětlené variability. Ze třetí tabulky na obrázku 48 jsou

patrné hodnoty faktorových zátěží, které bychom mohli využít pro získání konkrétních faktorových scórů jednotlivých objektů.

Communalities

	Initial	Extraction
delka_kalisniho_listku	1,000	,794
sirka_kalisniho_listku	1,000	,202
delka_okvetniho_listku	1,000	,983
sirka_okvetniho_listku	1,000	,931

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,911	72,770	72,770	2,911	72,770	72,770
2	,921	23,031	95,801			
3	,147	3,684	99,485			
4	,021	,515	100,000			

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component
	1
delka_kalisniho_listku	,891
sirka_kalisniho_listku	-,449
delka_okvetniho_listku	,992
sirka_okvetniho_listku	,965

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

Obrázek č. 48: Výstup faktorové analýzy v SPSS

Z obrázku 49 jsou patrné získané hodnoty jednotlivých nově vytvořených faktorových scórů (za pomoci faktorových zátěží z obrázku 48), které je možné využít pro potřeby dalších analýz.

FAC1_1
-1,32288
-1,21883
-1,38329
-1,34605
-1,39545
-1,20955
-1,42871
-1,30495
-1,36810
-1,27856
-1,26385
-1,35959

Obrázek č. 49: Výstup faktorové analýzy v SPSS

5 Diskriminační analýza

Mezi oblíbené klasifikační metody, jejichž cílem je zařazování objektů do předem známých skupin, lze zařadit diskriminační analýzu. Diskriminační analýzu lze také chápat jako metodu zkoumání závislostí, kdy uživatel hledá závislost jedné kvalitativní proměnné na skupině *k* kvantitativních proměnných, které se nazývají diskriminátory. Tato metoda je vhodná pro případ, kdy jsou dopředu známé výsledné skupiny a uživatel chce nalézt jakési „pravidlo“, podle něhož bude moci následně zařazovat další (nové) objekty. Původně nezařazený objekt je zařazen do dané skupiny na základě největší míry podobnosti, resp. nejmenší vzdálenosti od existující třídy (skupiny). Znalost či neznalost příslušnosti objektů do skupiny před počátkem analýzy je základní rozdíl v použití shlukové a diskriminační analýzy.

Při úlohách diskriminační analýzy je tedy nutné najít diskriminační (přiřazovací) pravidlo, které bývá označováno jako *diskriminační funkce,* s jehož pomocí se zařazují objekty do skupin. Při tvorbě diskriminační funkce je třeba rozlišit, zda analytik přesně ví, jaké proměnné jsou vhodné pro tvorbu funkce, či zda je třeba v rámci analýzy vybrat vhodné proměnné pro tvorbu hledané diskriminační funkce. V prvním případě je vhodná tzv. *kanonická diskriminační analýza,* ve druhém případě je vhodná tzv. *kroková diskriminační analýza*. Volba krokové diskriminační analýzy bývá vhodná zejména v případě, kdy je k dispozici velký počet proměnných a analytik má za úkol s posloupností krokové diskriminační analýzy vybrat pouze ty nejlepší.

Mezi nejznámější metody kanonické diskriminační analýzy (do úlohy vstupují všechny předem vybrané proměnné) bývá zařazována Fisherova lineární diskriminační funkce, která je vhodná v případě vícerozměrného normálního rozdělení se stejnými kovariančními maticemi. Vzhledem k robustnosti této

metody je možné její použití i při nesplnění těchto předpokladů. Při využití Fisherovy lineární diskriminační funkce dochází k maximalizaci Fisherova kanonického diskriminačního kritéria, které představuje poměr mezi meziskupinovou a vnitroskupinovou variabilitou. Výstupem bude tolik klasifikačních (diskriminačních) funkcí, kolik je celkový počet skupin. Před počátkem řešení úlohy by měla být otestována shoda rozptylů, resp. kovariačních matic (v členění podle skupin) a shoda středních hodnot uvnitř jednotlivých skupin. Pro případ Fisherovy lineární diskriminační funkce je vhodné, aby vnitroskupinové kovarianční matice byly shodné, v opačném případě je vhodné zvolit kvadratickou diskriminační funkci či logistickou diskriminaci, jak je uvedeno například v (Stankovičová, 2007).

Při použití lineární diskriminační funkce jsou výsledkem tzv. Fisherovy lineární diskriminační funkce, s jejichž cílem jsou vytvářeny hodnoty tzv. *kanonických proměnných*, které jsou vzájemně nezávislé a vyjadřují celkovou variabilitu původních proměnných. Tyto funkce bývají uspořádány sestupně podle hodnoty vlastního čísla a každá z nich slouží k vysvětlení celkové variability původních proměnných. Pro více než jednu diskriminační funkci je nutné ověřit, zda na rozlišení všech hledaných skupin je třeba využít všech s diskriminačních funkcí, či je možné využít menší počet. Číslo s představuje menší číslo z počtu proměnných a počtu vytvářených skupin mínus jedna. Pro jednotlivé původní proměnné je vhodné stanovit tzv. *diskriminační koeficienty* (váhy), které slouží k určení jejich příspěvku pro tvorbu (oddělení) q skupin. Jejich interpretace je analogická, jako je interpretace dílčích regresních koeficientů. Dále bývá vhodné stanovit korelační koeficienty (*strukturní koeficienty*), které vyjadřují jednoduchou korelaci mezi původní proměnnou a vytvořenou diskriminační funkcí. I v tomto případě je důležité znaménko tohoto koeficientu.

Zařazování nových objektů do skupin q pomocí diskriminační úlohy je možné řešit různými způsoby. Podle (Stankovičová, 2007) se buď jedná o zařazování podle kritické hodnoty diskriminačního skóre nebo podle bayesovské teorie rozhodování, případně jde o zařazování objektů s využitím klasifikačních funkcí či se jedná o klasifikaci podle Mahalanobisovy vzdálenosti.

Při využití klasifikačních funkcí (například Fisherových lineárních diskriminačních funkcí) stanovených pro každou z q skupin se postupuje tak, že se vypočítá klasifikační skóre každé nově zařazované statistické jednotky. Ta je zařazena do té skupiny, pro kterou byla dosažena nejvyšší hodnota tohoto klasifikačního skóre. Při klasifikaci pomocí Mahalanobisovy vzdálenosti se postupuje tak, že se stanoví Mahalanobisova vzdálenost každé zařazované statistické jednotky od centroidu každé skupiny. Nově klasifikovaná jednotka je zařazena do té skupiny, jejíž vzdálenost od centroidu dané skupiny je nejmenší.

Před započetím klasifikace nových jednotek je vhodné, aby byla ověřena její přesnost. K tomu existují různé metody. Jednou z nich je, při dostatečně velkém rozsahu výběrového vzorku, rozdělení původního souboru na dvě části. První část souboru se využije k odhadu klasifikačního kritéria a na druhé části (nevyužitých jednotek) souboru se provede kontrola klasifikace. Výsledkem tohoto postupu bývá získání nezkresleného odhadu přesnosti klasifikace. Při rozdělování souboru na dvě takovéto části nebývá obvykle jednoznačně definováno, v jakém poměru mají být soubory rozděleny na část pro tvorbu diskriminační funkce a jaká část souboru má být využita pro její kontrolu. Jinou možností je tzv. *křížové ověření přesnosti*, kdy se postupuje takovým způsobem, že se při odhadu klasifikačního kritéria vynechá jeden objekt. Následně se klasifikuje daný objekt do některé ze skupin. Tímto způsobem se postupuje tak dlouho, dokud není klasifikováno všech n objektů. Na základě

tohoto způsobu zkoumání přesnosti je podle (Stankovičová, 2007) získán nejméně vychýlený odhad míry přesnosti.

5.1 Zpracování dat z oblasti diskriminační analýzy

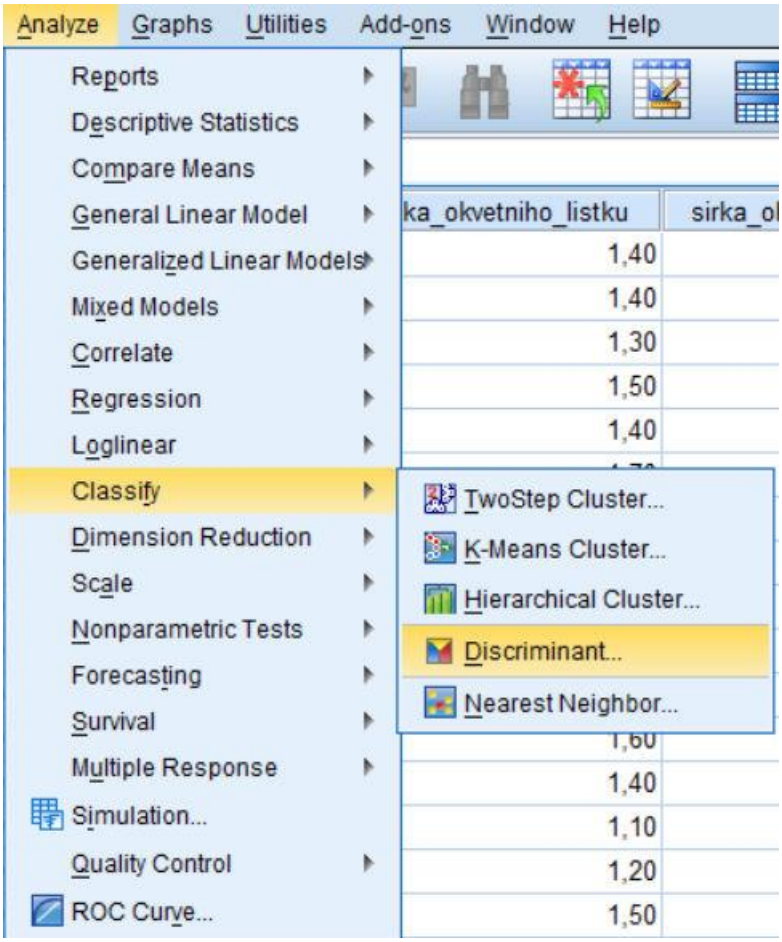
V této textu si ukážeme aplikaci diskriminační analýzy opět na již známém a dobře popsaném datovém souboru Iris. K ilustraci diskriminační analýzy v této části textu opět využijeme nejprve systém SPSS.

Příklad č. 1:

Na základě datového souboru „*Iris*“ se pokusíme odhadnout Fisherovu lineární diskriminační funkci a ověřit účinnost diskriminace pro každou ze skupin, jejichž původní velikosti jsou 50 objektů. Využijeme k tomu všechny původní proměnné.

Řešení:

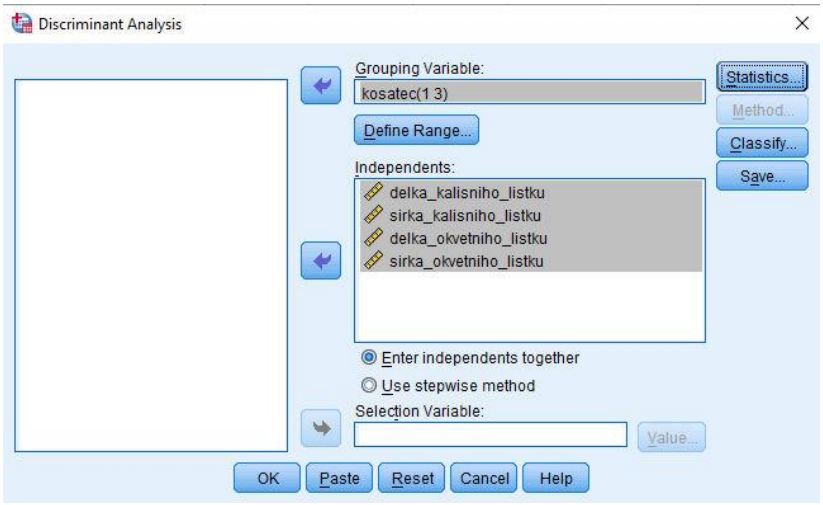
Proces aktivace diskriminační analýzy je patrný z obrázků 50 až 54. Z obrázku 50 je zřejmé umístění diskriminační analýzy v SPSS.



Obrázek č. 50: Spuštění diskriminační analýzy v SPSS

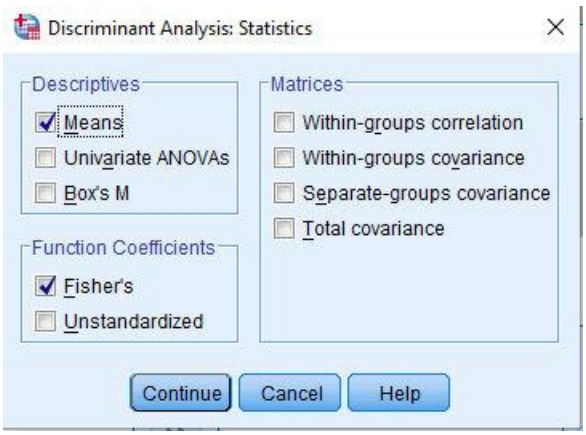
Do hlavního menu, po výběru diskriminační analýzy, je třeba vyznačit všechny proměnné, s jejichž pomocí bude vytvářena diskriminační funkce. V našem případě se opět jedná o proměnné délka a šířka kališního lístku, délka a šířka okvětního lístku. Dále je třeba vyznačit klasifikační proměnnou, která přiřazuje jednotlivým skutečným objektům jejich reálné přiřazení do známých skupin. Výběrem „*Define Range*“, jak je patrné na obrázku 51, musíme vyznačit čísla jednotlivých přiřazených skupin podle reálného zařazení *n* původních objektů. V tomto okně je také třeba vyznačit, zda bude využita kanonická diskriminační analýza „*Enter independents together*“, a nebo bude využita kroková

diskriminační analýza, s jejíž pomocí budou vybírány vhodné proměnné pro diskriminaci. Další nastavení ke krokové diskriminační analýze by následovalo v záložce „*Method*“. V našem případě zvolíme všechny proměnné ke tvorbě výsledných diskriminačních funkcí.



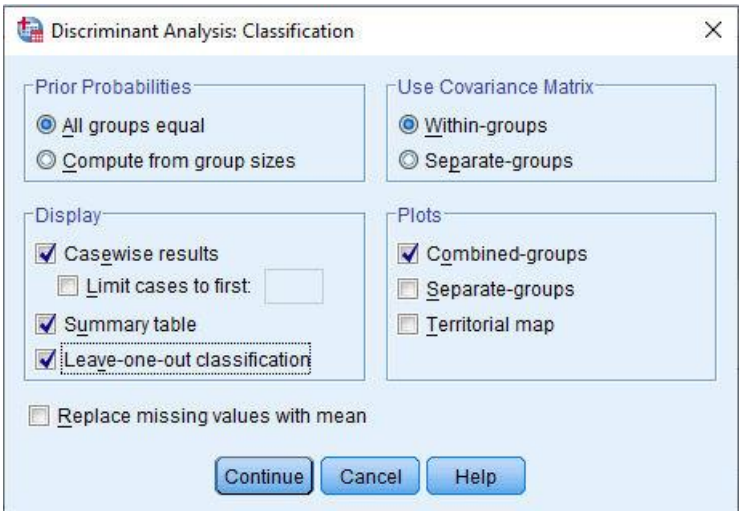
Obrázek č. 51: Spuštění diskriminační analýzy v SPSS

V okně na obrázku 52 vyznačíme, jaké popisné charakteristiky chceme vygenerovat. Dále vyznačíme, že využíváme Fisherovu lineární diskriminační funkci a chceme získat odhady jednotlivých koeficientů Fisherovy lineární diskriminační funkce.



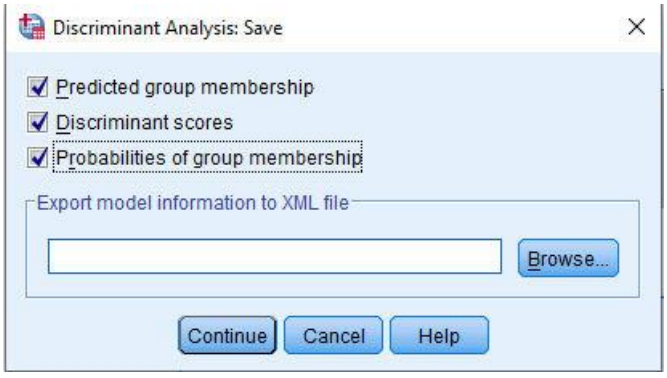
Obrázek č. 52: Spuštění diskriminační analýzy v SPSS

Doplňkové charakteristiky pro hodnocení výsledné klasifikace vybereme podle návodu na obrázku 53. Stejně tak si v daném okně můžeme zvolit graf, který bude následně vygenerován do výstupu. Pokud chceme získat představu o přesnosti výsledné klasifikace, necháme si zobrazit křížové ověření přesnosti klasifikace pomocí „*Leave-one-out classification*“.



Obrázek č. 53: Spuštění diskriminační analýzy v SPSS

Pokud máme zájem, aby do původní datové matice byla vygenerována „hypotetická“ přiřazení objektů do skupin na základě dané diskriminační funkce, zvolíme „*Predicted group membership*“, pokud chceme zobrazit hodnoty diskriminačních skórů zvolíme „*Discriminant scores*“. Máme-li zájem o zobrazení pravděpodobností zařazení objektů do jednotlivých skupin ve výstupu, zvolíme „*Probabilities of group membership*“ tak, jak je zachyceno na obrázku 54.



Obrázek č. 54: Spuštění diskriminační analýzy v SPSS

Po spuštění diskriminační analýzy podle výše popsaného návodu získáme výstupy, které jsou patrné z obrázků 55 – 62. Z obrázku 55 jsou patrné popisné charakteristiky pro jednotlivé kosatce, které již známe z přechozích kapitol. Je zde uveden průměr (*mean*), směrodatná odchylka (*Std. Deviation*) pro každý druh kosatců. V tomto výstupu jsou uvedeny také jednotlivé velikosti daných skupin a jim přiřazené váhy.

Group Statistics

kosatec		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
Iris-setosa	delka_kalisniho_listku	5,0060	,35249	50	50,000
	sirka_kalisniho_listku	3,4180	,38102	50	50,000
	delka_okvetniho_listku	1,4640	,17351	50	50,000
	sirka_okvetniho_listku	,2440	,10721	50	50,000
Iris-versicolor	delka_kalisniho_listku	5,9360	,51617	50	50,000
	sirka_kalisniho_listku	2,7700	,31380	50	50,000
	delka_okvetniho_listku	4,2600	,46991	50	50,000
	sirka_okvetniho_listku	1,3260	,19775	50	50,000
Iris-virginica	delka_kalisniho_listku	6,5880	,63588	50	50,000
	sirka_kalisniho_listku	2,9740	,32250	50	50,000
	delka_okvetniho_listku	5,5520	,55189	50	50,000
	sirka_okvetniho_listku	2,0260	,27465	50	50,000
Total	delka_kalisniho_listku	5,8433	,82807	150	150,000
	sirka_kalisniho_listku	3,0540	,43359	150	150,000
	delka_okvetniho_listku	3,7587	1,76442	150	150,000
	sirka_okvetniho_listku	1,1987	,76316	150	150,000

Obrázek č. 55: Výstup diskriminační analýzy v SPSS

Pokud jsme postupovali podle výše uvedených obrázků, získáme jednotlivé výstupy diskriminační analýzy, které jsou zachyceny na obrázku 56.

Z výstupu je zřejmé, že pro dostatečné odlišení námi definovaných skupin by postačily dvě diskriminační funkce. V první tabulce jsou uvedeny hodnoty vlastních čísel a příslušné procento vysvětlené variability. Ve třetí tabulce uvedeného výstupu jsou zobrazeny hodnoty standardizovaných diskriminačních koeficientů (vah) a ve čtvrté tabulce hodnoty korelačních koeficientů mezi danou proměnnou a vytvořenou diskriminační funkcí (strukturní koeficienty).

Analysis 1

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	32,272[a]	99,1	99,1	,985
2	,278[a]	,9	100,0	,466

a. First 2 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,024	545,577	8	,000
2	,783	35,641	3	,000

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
delka_kalisniho_listku	-,422	,017
sirka_kalisniho_listku	-,527	,734
delka_okvetniho_listku	,940	-,400
sirka_okvetniho_listku	,585	,575

Structure Matrix

	Function	
	1	2
delka_okvetniho_listku	,705*	,166
sirka_kalisniho_listku	-,116	,867*
sirka_okvetniho_listku	,632	,740*
delka_kalisniho_listku	,222	,314*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

Obrázek č. 56: Výstup diskriminační analýzy v SPSS

Z výstupu na obrázku 57 jsou patrné centroidy pro případné měření vzdálenosti pomocí Mahalanobisovy vzdálenosti.

Functions at Group Centroids

kosatec	Function 1	Function 2
Iris-setosa	-7,616	,213
Iris-versicolor	1,822	-,718
Iris-virginica	5,793	,505

Unstandardized canonical discriminant functions evaluated at group means

Obrázek č. 57: Výstup diskriminační analýzy v SPSS

Z výstupu na obrázku 58 jsou patrné klasifikační charakteristiky. V první tabulce je zřejmé, že bylo klasifikováno 150 objektů.

Classification Statistics

Classification Processing Summary

Processed		150
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		150

Prior Probabilities for Groups

kosatec	Prior	Cases Used in Analysis Unweighted	Weighted
Iris-setosa	,333	50	50,000
Iris-versicolor	,333	50	50,000
Iris-virginica	,333	50	50,000
Total	1,000	150	150,000

Classification Function Coefficients

	kosatec		
	Iris-setosa	Iris-versicolor	Iris-virginica
delka_kalisniho_listku	23,466	15,703	12,491
sirka_kalisniho_listku	23,568	6,954	3,444
delka_okvetniho_listku	-16,203	5,284	12,822
sirka_okvetniho_listku	-18,025	6,298	21,063
(Constant)	-86,053	-72,770	-104,295

Fisher's linear discriminant functions

Obrázek č. 58: Výstup diskriminační analýzy v SPSS

Ze třetí tabulky na obrázku 58 jsou patrné koeficienty příslušné odhadnuté Fisherovy lineární diskriminační funkce pro každou z q existujících skupin. V našem případě máme tři skupiny podle druhu kosatců. Ty bychom využili pro klasifikaci nově zařazované jednotky. Vypočetli bychom hodnotu klasifikačního scóre dané jednotky pro každou skupinu a zařadili bychom ji do té skupiny, u níž by hodnota tohoto klasifikačního skóre byla nejvyšší.

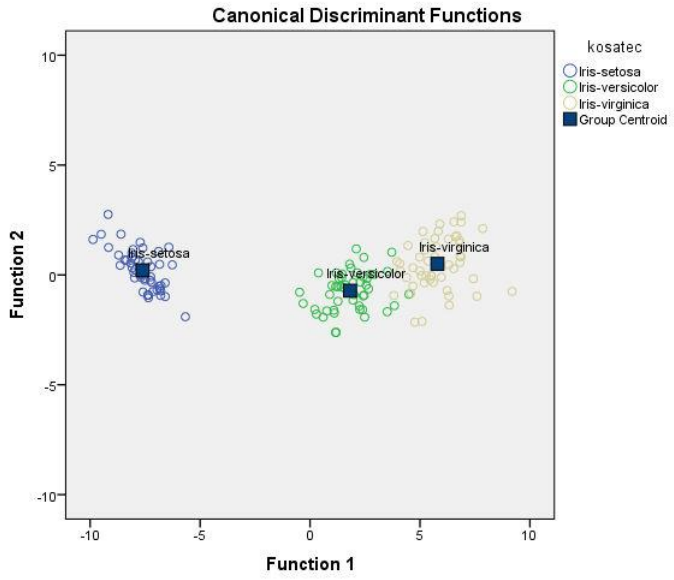
Z výstupu na obrázku 59 jsou patrné jednotlivé hodnoty diskriminačních skórů a postupné zařazování původních objektů do vytvářených skupin, případně hodnoty Mahalanobisových vzdáleností od výše uvedených centroidů pro daný objekt.

Casewise Statistics

			Highest Group					Second Highest Group			Discriminant Scores	
				P(D>d | G=g)			Squared Mahalanobis Distance to Centroid			Squared Mahalanobis Distance to Centroid		
	Case Number	Actual Group	Predicted Group	p	df	P(G=g | D=d)		Group	P(G=g | D=d)		Function 1	Function 2
Original	1	1	1	,890	2	1,000	,234	2	,000	99,252	-8,085	,328
	2	1	1	,561	2	1,000	1,157	2	,000	80,455	-7,147	-,755
	3	1	1	,899	2	1,000	,214	2	,000	87,351	-7,511	-,238
	4	1	1	,512	2	1,000	1,337	2	,000	75,003	-6,838	-,643
	5	1	1	,818	2	1,000	,402	2	,000	101,190	-8,158	,541
	6	1	1	,444	2	1,000	1,624	2	,000	95,969	-7,724	1,482
	7	1	1	,918	2	1,000	,172	2	,000	83,239	-7,235	,377
	8	1	1	,981	2	1,000	,039	2	,000	89,884	-7,630	,017
	9	1	1	,286	2	1,000	2,507	2	,000	70,720	-6,583	-,987
	10	1	1	,514	2	1,000	1,329	2	,000	84,518	-7,369	-,914

Obrázek č. 59: Výstup diskriminační analýzy v SPSS

Z obrázku 60 je zřejmé vytváření q skupin za pomocí zvoleného počtu diskriminačních funkcí (v našem případě dva). Zároveň je vždy zobrazen příslušný centroid dané skupiny, jejichž hodnoty byly zachyceny v tabulce na obrázku 57.



Obrázek č. 60: Výstup diskriminační analýzy v SPSS

Výsledky křížové klasifikace objektů podle postupu, který byl popsán výše, jsou zřejmé z obrázku 61. Je zřejmé, že všechny kosatce *Iris-setosa* byly správně klasifikovány. Jeden kosatec *Iris-virginica* byl nesprávně klasifikován jako *Iris-versicolor* a dva kosatce *Iris-versicolor* byly nesprávně označeny jako *Iris-virginica*. Celkem tedy byly nesprávně klasifikovány tři objekty, což při celkovém počtu 150 kosatců představuje 98% úspěšnost s využitím Fisherovy lineární diskriminační funkce, což lze jistě považovat za velmi uspokojivý výsledek.

Classification Results[a,c]

		kosatec	Predicted Group Membership			Total
			Iris-setosa	Iris-versicolor	Iris-virginica	
Original	Count	Iris-setosa	50	0	0	50
		Iris-versicolor	0	48	2	50
		Iris-virginica	0	1	49	50
	%	Iris-setosa	100,0	,0	,0	100,0
		Iris-versicolor	,0	96,0	4,0	100,0
		Iris-virginica	,0	2,0	98,0	100,0
Cross-validated[b]	Count	Iris-setosa	50	0	0	50
		Iris-versicolor	0	48	2	50
		Iris-virginica	0	1	49	50
	%	Iris-setosa	100,0	,0	,0	100,0
		Iris-versicolor	,0	96,0	4,0	100,0
		Iris-virginica	,0	2,0	98,0	100,0

a. 98,0% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 98,0% of cross-validated grouped cases correctly classified.

Obrázek č. 61: Výstup diskriminační analýzy v SPSS

Pokud jsme vybrali vygenerování charakteristik (hodnoty diskriminačních scórů a pravděpodobnosti klasifikace), získáme do původní matice výstup zobrazený na obrázku 62.

kosatec	Dis_1	Dis1_1	Dis2_1	Dis1_2	Dis2_2	Dis3_2
Iris-setosa	Iris-setosa	-8,08495	,32845	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-7,14716	-,75547	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-7,51138	-,23808	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-6,83768	-,64288	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-8,15781	,54064	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-7,72363	1,48232	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-7,23515	,37715	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-7,62974	,01667	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-6,58274	-,98737	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-7,36884	-,91363	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-8,42181	,67623	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-7,24740	-,08292	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-7,35062	-1,03936	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-7,59647	-,77672	1,00000	,00000	,00000
Iris-setosa	Iris-setosa	-9,86937	1,61486	1,00000	,00000	,00000

Obrázek č. 62: Výstup diskriminační analýzy v SPSS

6 Zpracování vybraných vícerozměrných statistických analýz v prostředí R

V této kapitole si uvedeme, jak zpracovat příklady, které byly řešeny v předchozích kapitolách v prostředí IBM SPSS verze 21 prostřednictvím statistického programovacího jazyka *R*. Velkou výhodou tohoto programovacího jazyka je to, že je volně dostupný a v současnosti velmi často používán. Pokud v tomto prostředí pracujete poprvé, je třeba si nejprve stáhnout samotný statistický programovací jazyk *R* (https://cran.rapporter.net/) v závislosti na operačním systému a na následné analýzy a zpracování dat doporučujeme používat editor *RStudio,* který je také třeba stáhnout a nainstalovat (https://rstudio.com/products/rstudio/download/).

Všechny příkazy a potřebné vysvětlení jsou uvedeny při konkrétních příkladech. V případě, že je třeba k provedení analýzy využití nějaké speciální knihovny (balíčky funkcí), které obsahují soubory specifických funkcí určených k řešení konkrétní problematiky, je třeba si tyto knihovny nejprve nainstalovat prostřednictvím příkazu: `install.packages("názevBalíčku")`, například: `install.packages("ggplot2")`. Následně je třeba tento balíček spustit pomocí příkazu: `library(názevBalíčku)`, například: `library(ggplot2)`.

6.1 Zpracování dat z oblasti vícenásobné regresní analýzy

V následující části textu bude demonstrováno použití vícenásobné regresní analýzy na konkrétním příkladu. K řešení bude využit statistický programovací jazyk *R*, kde jsou tyto postupy implementovány.

Příklad č. 1:

Cena	Počet kilometrů	Stáří
222 934	20 119	15
200 632	26 009	19
198 497	25 003	21
198 486	25 016	20
207 158	23 755	19
202 570	23 756	19
194 419	26 257	21
190 343	27 508	22
210 707	21 271	17
218 957	22 510	18
190 290	27 570	22
186 205	28 831	23
186 264	28 763	23
190 338	27 514	22

V tabulce jsou zaznamenány prodejní ceny v Korunách, počty ujetých kilometrů a stáří v měsících za 14 automobilů značky Alfa u stejného modelu. Pomocí regresní analýzy nalezněte vhodný model závislosti ceny automobilu značky Alfa na počtu ujetých kilometrů a stáří vozidla. Zároveň vyjádřete vhodnost a kvalitu zvoleného modelu. Při analýzách uvažujte obvyklou 5% hladinu významnosti.

Řešení v *R*:

Při práci ve statistickém programovacím jazyce *R* je třeba nejprve načíst údaje, které chceme analyzovat. Často používaným a v prostředí *R* dobře pracujícím datovým formátem je formát *.csv* (*comma separated value* – hodnoty oddělené čárkou). Údaje potřebné k provedení analýzy máme uložené v souboru **priklad01.csv** a na jeho načtení použijeme funkci `read.csv2()`. Funkce `read.csv2()` se využívá, když načítáme údaje, v nichž desetinná místa oddělujeme čárkou a samotné sloupce jsou v souboru odděleny středníkem,

což je typické pro středoevropské databáze. V případě analýzy amerických dat, pro které je typické použití desetinné tečky a sloupce jsou odděleny čárkou používáme funkci `read.csv()`. Některé vybrané argumenty těchto funkcí jsou:

- `file` je název souboru s příponou uváděný v uvozovkách, v našem případě `"priklad01.csv"`.
- `header = TRUE` je příkaz, který poukazuje na to, že v prvním řádku našeho souboru se nenacházejí údaje, ale jde o záhlaví sloupců. Tento argument je nastaven automaticky, proto pokud se v prvním řádku nenacházejí názvy sloupců, je třeba tento argument změnit na `header = FALSE`.
- `sep = ";"` je v případě funkce `read.csv2()` primárně nastavený oddělovač sloupců. Pokud jsou sloupce v našem souboru odděleny jinak, je třeba v tomto argumentu změnit příslušný oddělovač.
- `dec = ","` je v případě funkce `read.csv2()` primárně nastavený oddělovač desetinných míst. Pokud v datasetu nepoužíváme desetinnou čárku, je třeba v tomto argumentu změnit příslušný oddělovač.
- `row.names` může být vektor, ve kterém se nacházejí jmenovky řádků nebo může jít o číslo, které udává pořadí sloupce, ve kterém se jmenovky řádků nacházejí nebo může jít o slovní pojmenování sloupce, ve kterém se jmenovky řádků nacházejí. S popisky řádků se setkáváme například, pokud v řádcích jsou uloženy informace o konkrétních respondendech a v tomto případě by popisky řádků mohly být jejich jména. Nejčastěji bývají jmenovky řádků uloženy v prvním sloupci databáze, proto velmi často při načítání používáme argument `row.names = 1`. V našem případě, však v datasetu `"priklad01.csv"` nemáme žádné jmenovky sloupců, proto tento argument vynecháváme.

Abychom si mohli načíst data z externího souboru, je třeba mít tento soubor uložený v pracovním adresáři (working directory). Nastavení pracovního adresáře můžeme provést v *RStudiu* buď prostřednictvím sekvence příkazů **Session - Set Working Directory - Choose Directory**, klávesové zkratky **Ctrl + Shift + H** nebo přímo prostřednictvím funkce `setwd()`, jejíž argumentem je adresa pracovního adresáře.

Načteme si údaje a uložíme je do objektu **priklad01**. Na přiřazení používáme symbol šipky `"<-"`, případně `"="`. Objekt se uloží ve formáte takzvaného dataframe.

```
priklad01<-read.csv2("priklad01.csv")
```

Pokud bychom si chtěli zobrazit prvních nebo posledních několik pozorování, můžeme na to použít funkci `head()` respektive `tail()`. Tyto funkce jsou vhodné zejména v případě, pokud pracujeme s datovými maticemi s velkým počtem řádků, při kterých zobrazení všech pozorování nepřipadá v úvahu.

```
head(priklad01)
```

```
##     cena pocet_km stari
## 1 222934    20119    15
## 2 200632    26009    19
## 3 198497    25003    21
## 4 198486    25016    20
## 5 207158    23755    19
## 6 202570    23756    19
```

```
tail(priklad01)
```

```
##      cena pocet_km stari
## 9  210707    21271    17
## 10 218957    22510    18
## 11 190290    27570    22
## 12 186205    28831    23
## 13 186264    28763    23
## 14 190338    27514    22
```

Pokud chceme vědět s jakými proměnnými pracujeme můžeme si je vypsat prostřednictvím funkce `colnames()`, která nám vypíše jmenovky sloupců.

```
colnames(priklad01)
```

```
## [1] "cena"     "pocet_km" "stari"
```

Základní souhrnné charaktristiky analyzované databáze můžeme zjistit prostřednictvím funkce `summary()`.

```
summary(priklad01)
```

```
##       cena           pocet_km         stari      
##  Min.   :186205   Min.   :20119   Min.   :15.00  
##  1st Qu.:190339   1st Qu.:23755   1st Qu.:19.00  
##  Median :198492   Median :25513   Median :20.50  
##  Mean   :199843   Mean   :25277   Mean   :20.07  
##  3rd Qu.:206011   3rd Qu.:27513   3rd Qu.:22.00  
##  Max.   :222934   Max.   :28831   Max.   :23.00
```

Lineární regresní analýzu v prostředí *R* provádíme prostřednictvím funkce `lm()`. Nejprve je třeba si definovat regresní rovnici, kterou chceme odhadovat. Uložíme ji do objektu **form01**. Na levé straně této regresní rovnice se nachází vysvětlována proměnná, což v našem případě je cena automobilu. Na pravé straně této regresní rovnice se nacházejí vysvětlující proměnné (v našem případě počet ujetých kilometrů a stáří automobilu), které oddělujeme symbolem `"+"`. Levou a pravou stranu regresní rovnice oddělujeme symbolem `"~"`. V případě, že bychom nechtěli odhadovat úrovňovou konstantu (intercept), na konec pravé strany napíšeme `-1` nebo `+0`.

```
form01<-cena~pocet_km+stari
```

Funkce `lm()` má několik argumentů, přičemž nejdůležitější z nich jsou `formula`, který představuje předpis odhadované regresní rovnice a `data`, který představuje název datasetu odkud pocházejí data, z nichž rovnici odhadujeme. Výsledky získané touto funkcí si uložíme do objektu **model01**.

```
model01<-lm(formula = form01, data = priklad01)
```

Pro zobrazení výsledků vícenásobné lineární regrese použijeme funkci `summary()`.

```
summary(model01)
```

```
##
## Call:
## lm(formula = form01, data = priklad01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4858.4  -941.5  -118.0   588.5  8368.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 300626.010   8416.093  35.720 9.96e-13 **
*
## pocet_km        -2.172      1.240  -1.751    0.108
## stari        -2286.264   1427.674  -1.601    0.138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
##
## Residual standard error: 3195 on 11 degrees of freedo
m
## Multiple R-squared:  0.936,  Adjusted R-squared:  0.9
243
## F-statistic: 80.37 on 2 and 11 DF,  p-value: 2.727e-0
7
```

V úvodu výstupu regresní analýzy máme uvedeno, jakou funkci jsme použili a základní popisné charakteristiky reziduí. Následuje tabulka odhadnutých regresních parametrů modelu (sloupec `Estimate`) spolu se směrodatnými chybami odhadů těchto regresních parametrů (sloupec `Std. Error`). Podílem odhadu regresních parametrů a směrodatných chyb odhadu dostáváme hodnotu testového kritéria *t* ve sloupci `t value`. Toto kritérium slouží k testování hypotézy o významnosti regresních koeficientů a porovnává se s příslušným

kvantilem Studentova t rozdělení. V posledním sloupci `Pr (>|t|)` jsou uvedeny *p*-hodnoty všech dílčích t testů, které jsou porovnávány s předem zvolenou hladinou významnosti. Nulovou hypotézu o nevýznamnosti příslušného regresního parametru zamítáme, pokud *p*-hodnota je menší nebo rovna zvolené hladině významnosti.

V dolní části výstupu jsou uvedeny hodnoty indexu determinace (`Multiple R-squared`) a upraveného indexu determinace (`Adjusted R-squared`) a *F* statistiky, kterou využíváme při testování hypotézy o statistické významnosti modelu jako celku. Jak můžeme vidět, *p*-hodnota tohoto testu je menší než jakákoli rozumná hladina významnosti, což znamená, že model jako celek je statisticky významný (alespoň jeden jeho parametr je statisticky významný).

Funkce `summary()` nám zobrazuje pouze odhady regresních parametrů, směrodatné chyby odhadů, hodnoty testového kritéria *t* a *p*-hodnoty dílčích t testů. Pokud bychom však chtěli provést klasický rozklad celkového součtu čtverců (S_y) na teoretický ($S_{y,T}$) a reziduální součet čtverců ($S_{y,R}$), jak jsme si to uvedli na příkladu v SPSS, musíme k tomu použít funkci `anova()` – analýzu rozptylu, která však v prostředí statistického programovacího jazyka *R* zobrazuje svůj výstup trochu jinak jako jsme na to zvyklí z prostředí SPSS.

```
anova(model01)
```

```
## Analysis of Variance Table
##
## Response: cena
##           Df     Sum Sq    Mean Sq  F value    Pr(>F)
## pocet_km   1 1614960760 1614960760 158.1832 7.168e-08
***
## stari      1   26181592   26181592   2.5645    0.1376
## Residuals 11  112303756   10209432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

Při interpretaci této tabulky začínáme od spodního řádku, kde máme uvedenou tu část variability, kterou se nepodařilo modelem vysvětlit, která je vyjádřena reziduálním součtem čtverců a je rovna 112 303 756 ($S_{y,R}$: reziduální součet čtverců). Ta část variability, která je vysvětlena modelem (vyjádřena prostřednictvím teoretického součtu čtverců) je rovna 1 614 960 760 (teoretický součet čtverců pro proměnnou `pocet_km`) + 26 181 592 (teoretický součet čtverců pro proměnnou `stari`) = 1 641 142 352 ($S_{y,T}$: teoretický součet čtverců). Celkový součet čtverců představuje součet reziduálního a teoretického součtu čtverců: 112 303 756 + 1 641 142 352 = 1 753 446 108 (S_y: celkový součet čtverců). Pro srovnání a lepší přehlednost uvádíme výstup z prostředí SPSS uvedený v Kapitole 1.2 na Obrázku 4:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,967[a]	,936	,924	3195,220	2,803

a. Predictors: (Constant), stari, pocet_km

b. Dependent Variable: cena

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1641142352	2	820571175,9	80,374	,000[b]
	Residual	112303756,0	11	10209432,36		
	Total	1753446108	13			

a. Dependent Variable: cena

b. Predictors: (Constant), stari, pocet_km

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	300626,010	8416,093		35,720	,000
	pocet_km	-2,172	1,240	-,510	-1,751	,108
	stari	-2286,264	1427,674	-,466	-1,601	,138

a. Dependent Variable: cena

Pokud bychom vydělili hodnotu teoretického součtu čtverců hodnotou celkového součtu čtverců, dostali bychom hodnotu indexu, respektive koeficientu determinace:

$$\frac{1\,641\,142\,353}{1\,753\,446\,108} = 0{,}9359525$$

Vraťme se však k výstupu z funkce `summary()`. Jak můžeme vidět, *p*-hodnota F testu o statistické významnosti modelu jako celku je menší než jakákoli rozumná hladina významnosti, což znamená, že model jako celek je statisticky významný (alespoň jeden jeho parametr je statisticky významný). Pokud se však podíváme na *p*-hodnoty dílčích t testů, s výjimkou úrovňové konstanty, ani jeden regresní parametr není statisticky významný. Tento rozporuplný závěr je způsoben problémem multikolinearity, která představuje vzájemnou závislost mezi vysvětlujícími proměnnými. Ověření přítomnosti multikolinearity je možné pomocí korelační matice, na její výpočet používáme funkci `cor()`. Důležitým argumentem této funkce je `method`, která je automaticky nastavena na Pearsonovu korelační metriku, můžeme si však vybrat ze tří různých korelačních koeficientů: Pearsonův (`method = "pearson"`), Kendallův (`method = "Kendall"`) či Spearmanův (`method = "Spearman "`).

```
cor(priklad01, method = "pearson")
```

```
##                cena   pocet_km      stari
## cena      1.0000000 -0.9596984 -0.9581789
## pocet_km -0.9596984  1.0000000  0.9650480
## stari    -0.9581789  0.9650480  1.0000000
```

Z uvedeného výstupu je zřejmé, že hodnota korelačního koeficientu mezi stářím vozidla a počtem ujetých kilometrů je velmi vysoká (0,965). Za škodlivou závislost je považována absolutní hodnota korelačního koeficientu alespoň 0,8, což je právě náš případ. Pokud bychom chtěli tuto statisticky významnou lineární závislost mezi danou dvojicí proměnných testovat, můžeme na to použít funkci `cor.test()`, jejímž argumentem je dvojice vektorů, mezi nimiž chceme vypočítat vzájemnou lineární závislost danou

korelačním koeficientem. Pokud chceme pracovat pouze s konkrétním sloupcem nějakého dataframe, můžeme se na něj odkazovat prostřednictvím symbolu dolaru `"$"` a názvu sloupce.

```
cor.test(priklad01$stari, priklad01$pocet_km)
```

```
##
##  Pearson's product-moment correlation
##
## data:  priklad01$stari and priklad01$pocet_km
## t = 12.756, df = 12, p-value = 2.441e-08
## alternative hypothesis: true correlation is not equal
to 0
## 95 percent confidence interval:
##  0.8903675 0.9891489
## sample estimates:
##      cor
## 0.965048
```

Vidíme, že i na základě tohoto testu, můžeme danou lineární závislost považovat za statisticky významnou. Je proto jasné, že obě vysvětlující proměnné v modelu nemohou být současně a jednu z nich bude třeba z modelu odstranit. Bude to zřejmě ta proměnná, jejíž vliv na vysvětlující proměnnou je menší. To zjistíme pomocí korelačních koeficientů těchto proměnných ve vztahu k vysvětlováné proměnné. Z uvedené korelační matice je zřejmá nižší síla lineární závislosti stáří a ceny, proto zřejmě vyřadíme stáří.

V praxi se však často setkáváme se složitějšími modely, kde máme několik vysvětlujících proměnných a v případě identifikace multikolinearity by byl tento proces výběru a odstraňování vysvětlujících proměnných složitější. Jednou z možností jak řešit tento problém je využít metodu postupného vyřazování proměnných (Backward Elimination) nebo postupného zařazování proměnných (Forward Selection), které jsou v *R* implementovány funkcí `()` s argumentem `step = "backward"`, respektive `step = "forward"`.

```
step(model01, direction = "backward")
```

```
## Start:  AIC=228.57
## cena ~ pocet_km + stari
##
##            Df Sum of Sq       RSS    AIC
## <none>                  112303756 228.57
## - stari     1  26181592 138485348 229.50
## - pocet_km  1  31291435 143595191 230.01

##
## Call:
## lm(formula = cena ~ pocet_km + stari, data = priklad0
1)
##
## Coefficients:
## (Intercept)     pocet_km        stari
##  300626.010       -2.172    -2286.264
```

Uvádíme příklad metody postupného vyřazování proměnných (Backward elimination), při níž algoritmus začal při úplném modelu se všemi vysvětlujícími proměnnými (`cena ~ pocet_km + stari`), u kterého bylo stanoveno Akaikeho informační kritérium tohoto modelu (`AIC = 228.57`) a dále postupně z modelu byl odstraněn nejprve věk (`- stari`), vypočítáno Akaikeho informační kritérium (`229.50`) a následně byl odstraněn počet kilometrů (`- pocet_km`) a bylo stanoveno Akaikeho informační kritérium (`230.01`). Pokud se při volbě specifikace modelu rozhodujeme pouze na základě Akaikeho informačního kritéria, vybíráme model, který má tuto hodnotu nejnižší, proto daný algoritmus vybral za nejvhodnější model s oběma vysvětlujícími proměnnými.

6.2 Zpracování dat z oblasti shlukové analýzy

Příklad č. 2:

K demonstraci shlukové analýzy bude využit velmi známý soubor určený ke klasifikaci týkající se rostlin – Kosatců s názvem „*Iris*“. Soubor a podrobné

informace k němu je možné najít na webové adrese https://archive.ics.uci.edu/ml/datasets/iris. Soubor určený ke klasifikaci obsahuje celkem 150 kosatců tří druhů (Iris Setosa, Iris Versicolour, Iris Virginica). Každý druh je zastoupen shodně 50 objekty a je charakterizován pomocí čtyř kvantitativních proměnných, jako jsou délka a šířka okvětního lístku a délka a šířka kališního lístku, vždy uváděno v centimetrech. Pomocí hierarchické shlukové analýzy proveďte klasifikaci jednotlivých objektů (Kosatců) a porovnejte se skutečnou příslušností do dané skupiny.

Řešení v *R*:

Dataset *Iris* je v prostředí *R* přímo implementován, a proto jej není potřeba načítat z externího zdroje. Uložme si tento dataset do objektu **data02** a vypíšeme si základní charakteristiky tohoto datasetu.

```
data02<-iris

head(data02)

##   Sepal.Length Sepal.Width Petal.Length Petal.Width S
pecies
## 1          5.1         3.5          1.4         0.2
setosa
## 2          4.9         3.0          1.4         0.2
setosa
## 3          4.7         3.2          1.3         0.2
setosa
## 4          4.6         3.1          1.5         0.2
setosa
## 5          5.0         3.6          1.4         0.2
setosa
## 6          5.4         3.9          1.7         0.4
setosa

tail(data02)

##     Sepal.Length Sepal.Width Petal.Length Petal.Width
Species
```

```
## 145          6.7         3.3          5.7         2.5
virginica
## 146          6.7         3.0          5.2         2.3
virginica
## 147          6.3         2.5          5.0         1.9
virginica
## 148          6.5         3.0          5.2         2.0
virginica
## 149          6.2         3.4          5.4         2.3
virginica
## 150          5.9         3.0          5.1         1.8
virginica
```

```
summary(data02)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Pet
al.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.
:0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st
Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Medi
an :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean
:1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd
Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.
:2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
```

Jak můžeme vidět, při základních charakteristikách se v případě kvantitativních proměnných zobrazily popisné statistiky a v případě kvalitativních proměnných absolutně četnosti těchto kategorií.

Na úvod musíme nadefinovat matici vzdáleností, na níž se bude následně realizovat samotná shluková analýza. Matici vzdáleností vytvoříme z původní

datové matice prostřednictvím funkce `dist()`, při které máme na výběr z více druhů vzdáleností, které do funkce přidáváme jako argument method: `method = "euklidean"`, `method = "maximum"`, `method = "manhattan"`, `method = "canberra"`, `method = "binary"` a `method = "minkowski"`.

Pokud máme objekty definovány proměnnými v různých měrných jednotkách často se ještě před samotným výpočtem matice vzdáleností údaje vhodně transformují. Jednou z možností transformace údajů je takzvané normování, které můžeme provést prostřednictvím funkce `scale()`, jejímž argumentem je datová matice, jejíž sloupce chceme normovat. V případě tohoto datasetu *Iris* to však není nutné, protože všechny hodnoty jsou v centimetrech a hodnoty jsou srovnatelné, proto toto normování můžeme přeskočit.

Nyní chceme vypočítat matici vzdáleností pro dataset *Iris*. V tomto datasetu máme 150 pozorování, 4 proměnné a v posledním pátém sloupci máme informace o konkrétním druhu (přiřazení do konkrétní skupiny). Proto budeme počítat matici vzdáleností pouze z prvních 4 sloupců a ze všech řádků. Takový výběr v *R* provádíme prostřednictvím hranatých závorek, kde první pozice poukazuje na řádky a druhá pozice na sloupce. Pokud chceme všechny řádky, necháme první pozici prázdnou a když chceme první čtyři sloupce využijeme na to sekvenci od 1 do 4, na což v R používáme symbol dvojtečky (`":"`). Použijeme Euklidovskou vzdálenost.

```
euklidovskaVzdalenost<-dist(data02[,1:4], method="euclid
ean")
```

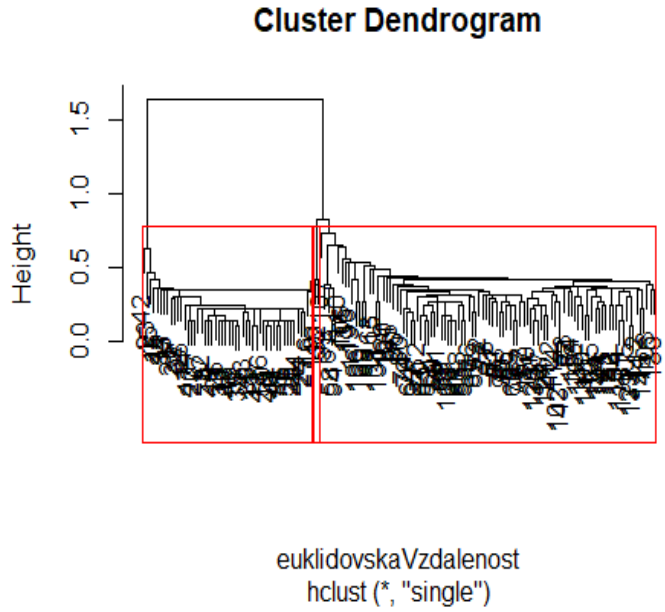
V objektu **euklidovskaVzdalenost** máme nyní uloženou matici euklidovské vzdáleností objektů z datasetu *Iris* uloženého v objektu **data02**. Jelikož původní dataset má 150 pozorování, matice euklidovské vzdálenosti má 150 řádků a 150 sloupců, je symetrická s nulami na hlavní diagonále. Tato matice nám poslouží k realizaci shlukové analýzy. Budeme realizovat hierarchické

shlukování, proto používáme funkci `hclust()`. Na výběr máme z více shlukovacích metod, které do funkce přidáváme jako argument method: `method = "ward.D"`, `method = "ward.D2"`, `method = "single"`, `method = "complete"`, `method = "average"`, `method = "mcquitty"`, `method = "median"` a `method = "centroid"`. Začneme nejstarší metodou nejbližšího souseda, která je v R označena jako `method = "single"`. Výsledky si můžeme uložit do objektu **fitSingle**.

```
fitSingle<-hclust(euklidovskaVzdalenost, method = "singl
e")
```

Výsledky hierarchického shlukování můžeme zobrazit prostřednictvím takzvaného dendrogramu pomocí funkce `plot()`. Pokud si chceme zobrazit i ohraničení jednotlivých skupin, použijeme funkci `rect.hclust()`, jejíž argumenty jsou samotný objekt vytvořen hierarchickým shlukováním, `k`, který představuje počet skupin (v našem případě 3, neboť máme tři druhy kosatců) a argumentem `border` nastavujeme barvu ohraničení jednotlivých skupin.

```
plot(fitSingle)
rect.hclust(fitSingle,k = 3, border = "red")
```



Z grafu je jasně vidět, že daná metoda není velmi vhodná, protože víme, že skutečné přiřazení je takové, že máme 50 pozorování u všech tří druhů kosatců. Zde však vidíme, že přiřazení je výrazně odlišné, což může být způsobeno takzvaným problémem řetězení, které může být v případě této metody přítomný. Pokud bychom si chtěli zobrazit přiřazení do *k* shluků při aplikaci této metody, využíváme k tomu funkci `cuttree()` s argumenty samotného objektu vytvořeného hierarchickým shlukováním a počtem shluků (v našem případě 3). Uložme toto přiřazení do objektu **prirazeniSingle**.

```
prirazeniSingle<-cutree(fitSingle,k = 3)
prirazeniSingle
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [71] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [106] 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

```
2 2 3 2 2 2 2 2 2 2 2
## [141] 2 2 2 2 2 2 2 2 2 2
```

I na základě tohoto výstupu vidíme, že většinou metoda přiřazovala objekty do druhého a prvního shluku, třetí shluk se v přiřazení vyskytuje pouze dvakrát. Pokud chceme toto přiřazení konfrontovat se skutečnou příslušností do skupin, která je uvedena ve sloupci `species` objektu **data02**, můžeme tak učinit prostřednictvím jednoduché kontingenční tabulky s využitím funkce `table()` následujícím způsobem.

```
table(data02$Species, prirazeniSingle)
```

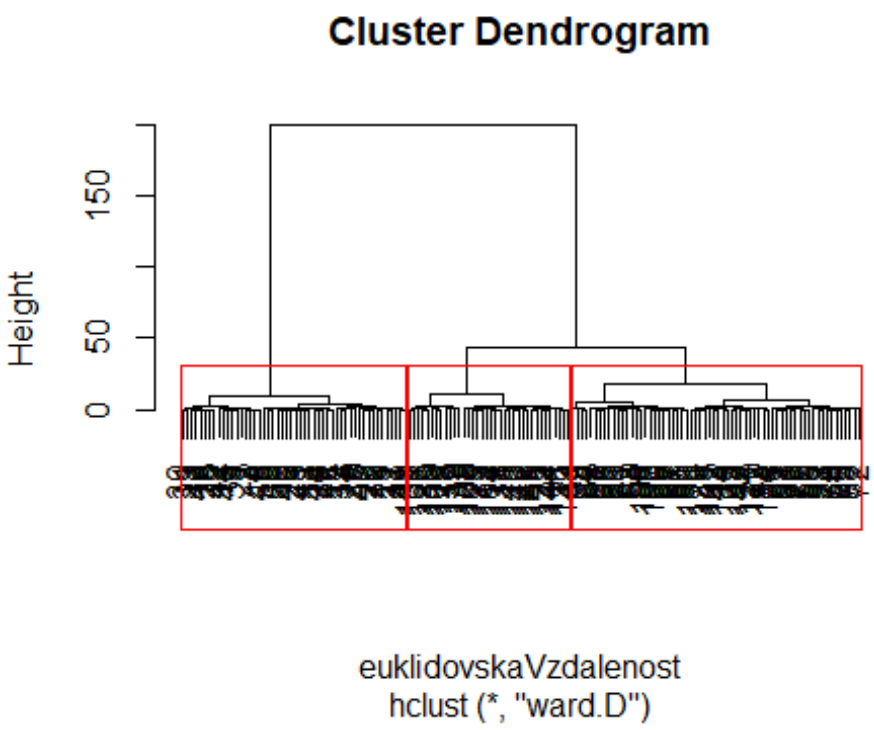
```
##              prirazeniSingle
##                1  2  3
##   setosa      50  0  0
##   versicolor   0 50  0
##   virginica    0 48  2
```

Vidíme, že metodě nejbližšího souseda se podařilo správně identifikovat první druh (iris-setosa). Ostatní dvě skupiny byly odděleny velmi špatně, protože do druhého shluku bylo zařazeno navíc nesprávně dalších 48 rostlin z třetího shluku. Je proto rozumné použít jinou metodu shlukování. V praxi se nejčastěji používá kombinace euklidovské vzdálenosti a Wardové metody shlukování, jejíž výhodou je, že většinou na výstupu nabízí shluky relativně podobné velikosti. Postup shlukování je stejný jako v případě metody nejbližšího souseda.

```
fitWard<-hclust(euklidovskaVzdalenost, method ="ward.D")
```

Výsledky shlukování si opět vykreslíme dendrogramem.

```
plot(fitWard)
rect.hclust(fitWard,k = 3, border = "red")
```




```
prirazeniWard<-cutree(fitWard,k = 3)
prirazeniWard
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [71] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3
## [106] 3 2 3 3 3 3 3 3 2 2 3 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 3 2 3 3 3 3 2 3 3 3 3 2 3
## [141] 3 3 2 3 3 3 2 3 3 2
```

Na první pohled vypadají takto získané výsledky vhodnější, než v případě metody nejbližšího souseda. Porovnání úspěšnosti přiřazení můžeme opět vidět v kontingenční tabulce.

```
table(data02$Species, prirazeniWard)
```

```
##              prirazeniWard
##                1  2  3
##   setosa      50  0  0
##   versicolor   0 50  0
##   virginica    0 14 36
```

V tomto případě byly správně přiřazeny objekty nacházejí na hlavní diagonále kontingenční tabulky a nesprávné přiřazeny objekty nad, respektive pod touto hlavní diagonálou. Jak můžeme vidět, druh iris-setosa byl přiřazen správně, iris-versicolor také a iris-virginica byl přiřazen správně v 36 případech, přičemž v zbývajících 14 případech byl nesprávně identifikován jako iris-versicolor.

6.3 Zpracování dat z oblasti analýzy hlavních komponent

Příklad č. 3:

K demonstraci metody hlavních komponent, stejně jako v případě shlukové analýzy, bude využit velmi známý soubor Kosatců s názvem „*Iris*“. Soubor sice neobsahuje velké množství proměnných (obsahuje pouze 4), avšak z důvodu provázanosti a pochopení souvislostí jej využijeme i zde. Pro připomenutí uvádíme, že soubor obsahuje celkem 150 objektů.

Řešení v *R*:

Při řešení tohoto příkladu využijeme skutečnost, že datový soubor již máme uložený v objektu **data02**, který jsme využívali při shlukové analýze.

Na úvod si vypočítáme korelační matici, abychom mohli identifikovat, zda je vůbec použití metody hlavních komponent na redukci dimenzí vhodné. Korelační matici počítáme opět pouze pro první 4 sloupce objektu **data02**.

```
cor(data02[,1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Pe
tal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538
0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -
0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000
0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654
1.0000000
```

Z korelační matice vyplývá vysoká pozitivní závislost mezi délkou kališních lístků (Sepal.Length) a délkou okvětního lístku (Petal.Length) a šířkou okvětního lístku (Petal.Width). O tom, že jde o statisticky významnou lineární závislost se můžeme přesvědčit pomocí funkce `cor.test()`.

```
cor.test(data02$Sepal.Length,data02$Petal.Length)

##
##  Pearson's product-moment correlation
##
## data:  data02$Sepal.Length and data02$Petal.Length
## t = 21.646, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal
to 0
## 95 percent confidence interval:
##  0.8270363 0.9055080
## sample estimates:
##       cor
## 0.8717538

cor.test(data02$Sepal.Length,data02$Petal.Width)

##
##  Pearson's product-moment correlation
##
## data:  data02$Sepal.Length and data02$Petal.Width
## t = 17.296, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal
to 0
## 95 percent confidence interval:
```

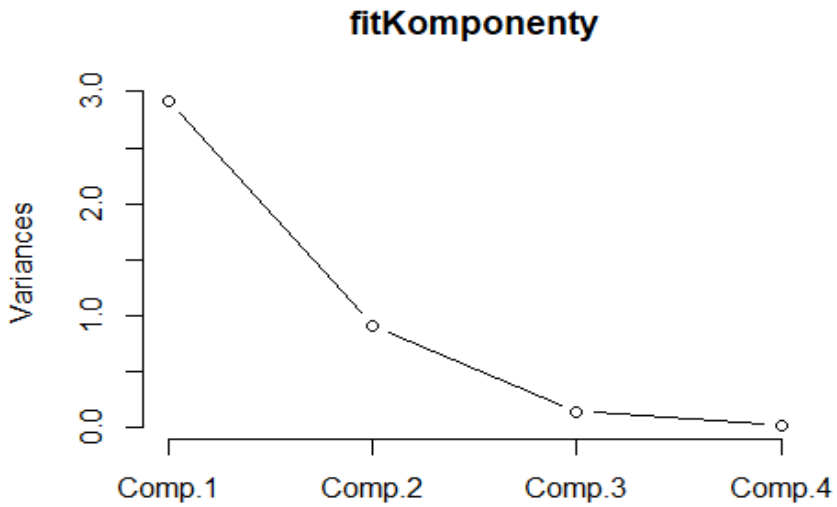
```
##  0.7568971 0.8648361
## sample estimates:
##       cor
## 0.8179411
```

Analýzu hlavních komponent v *R* provádíme prostřednictvím funkce `princomp()`. Důležitými argumenty této funkce je jednak datová matice, ze které chceme vypočítávat hlavní komponenty a zda chceme při jejich výpočtu vycházet z korelační nebo kovarianční matice. Pokud chceme vycházet z korelační matice tak nastavíme argument `cor = TRUE`. Uložme si výsledek této analýzy do objektu **fitKomponenty**.

```
fitKomponenty<-princomp(data02[,1:4], cor = TRUE)
```

Graf vlastních čísel korelační matice si můžeme zobrazit prostřednictvím funkce `plot()`. Argumentem `type = "lines"` pouze nastavíme, že chceme, aby šlo o spojnicový graf.

```
plot(fitKomponenty, type = "lines")
```



Prostřednictvím funkce `summary()` je možné si zobrazit souhrnnou tabulku za jednotlivé komponenty. V řádku `Standard deviation` se nacházejí směrodatné odchylky pro každou komponentu.

```
summary(fitKomponenty)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3
Comp.4
## Standard deviation     1.7083611 0.9560494 0.38308860
0.143926497
## Proportion of Variance 0.7296245 0.2285076 0.03668922
0.005178709
## Cumulative Proportion  0.7296245 0.9581321 0.99482129
1.000000000
```

Vlastní čísla korelační matice představují rozptyly komponent, tedy druhé mocniny těchto hodnot, které jsou zobrazeny i v předchozím grafu. Pokud bychom si chtěli vypočítat vlastní čísla korelační matice můžeme použít funkci `Eigen()`, která kromě vlastních čísel vypočítá i vlastní vektory korelační matice. Pokud nás zajímají pouze vlastní čísla můžeme použít symbol dolaru (`"$"`) a z výstupu získat pouze hodnoty vlastních čísel (`values`).

```
eigen(cor(data02[,1:4]))
```

```
## eigen() decomposition
## $values
## [1] 2.91849782 0.91403047 0.14675688 0.02071484
##
## $vectors
##            [,1]        [,2]       [,3]       [,4]
## [1,]  0.5210659 -0.37741762  0.7195664  0.2612863
## [2,] -0.2693474 -0.92329566 -0.2443818 -0.1235096
## [3,]  0.5804131 -0.02449161 -0.1421264 -0.8014492
## [4,]  0.5648565 -0.06694199 -0.6342727  0.5235971
```

```
eigen(cor(data02[,1:4]))$values
```

```
## [1] 2.91849782 0.91403047 0.14675688 0.02071484
```

Tato vlastní čísla opravdu představují druhé mocniny směrodatných odchylek (rozptyly) jednotlivých komponent, o čemž se můžeme jednoduše přesvědčit.

```
summary(fitKomponenty)$sdev
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4
## 1.7083611 0.9560494 0.3830886 0.1439265
```

```
summary(fitKomponenty)$sdev^2
```

```
##     Comp.1     Comp.2     Comp.3     Comp.4
## 2.91849782 0.91403047 0.14675688 0.02071484
```

V řádku `Proportion of Variance` jsou uvedeny podíly variability, které danou komponentou víme vysvětlit. V řádku `Cumulative proportion` jsou kumulativní podíly variability (kolik procent variability víme vysvětlit, pokud bychom postupně brali v úvahu další a další komponentu). Jak můžeme vidět, první komponentou bychom vysvětlili vysvětlit téměř 73 % variability. První a druhou komponentou bychom vysvětlili téměř 96 % variability, což považujeme za dostatečné, a proto další dvě komponenty nepřidáváme.

Koeficienty vlastních vektorů získáme prostřednictvím funkce `loadings()`. Jde o odhadnuté parametry příslušné lineární kombinace pro jednotlivé původní proměnné u výsledných komponent.

```
loadings(fitKomponenty)
```

```
##
## Loadings:
##              Comp.1 Comp.2 Comp.3 Comp.4
## Sepal.Length  0.521  0.377  0.720  0.261
## Sepal.Width  -0.269  0.923 -0.244 -0.124
## Petal.Length  0.580         -0.142 -0.801
## Petal.Width   0.565         -0.634  0.524
##
##                Comp.1 Comp.2 Comp.3 Comp.4
## SS loadings      1.00   1.00   1.00   1.00
```

```
## Proportion Var   0.25   0.25   0.25   0.25
## Cumulative Var   0.25   0.50   0.75   1.00
```

Chybějící koeficienty ve výpisu jsou malé, nikoli však nulové o čemž se můžeme přesvědčit tak, že si je vypíšeme jako sloupcové vektory.

```
loadings(fitKomponenty)[,1]

## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.5210659   -0.2693474    0.5804131    0.5648565

loadings(fitKomponenty)[,2]

## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##   0.37741762   0.92329566   0.02449161   0.06694199

loadings(fitKomponenty)[,3]

## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.7195664   -0.2443818   -0.1421264   -0.6342727

loadings(fitKomponenty)[,4]

## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.2612863   -0.1235096   -0.8014492    0.5235971
```

První dvě komponenty proto v našem případě mají tvar:

$$Z_1 = 0.5210659 \times \text{Sepal.Length} - 0.2693474 \times \text{Sepal.Width} + 0.5804131 \times \text{Petal.Length} + 0.5648565 \times \text{Petal.Width}$$

$$Z_2 = 0.37741762 \times \text{Sepal.Length} + 0.92329566 \times \text{Sepal.Width} + 0.02449161 \times \text{Petal.Length} + 0.06694199 \times \text{Petal.Width}$$

Tyto hodnoty můžeme následně použít pro výpočet takzvaných komponentních skóre u jednotlivých objektů. Ty lze zobrazit prostřednictvím příkazu `fitKomponenty$scores`.

```
fitKomponenty$scores
```

```
##              Comp.1       Comp.2       Comp.3       Co
mp.4
##   [1,] -2.26470281  0.480026597  0.127706022  0.02416
8204
##   [2,] -2.08096115 -0.674133557  0.234608854  0.10300
6775
##   [3,] -2.36422905 -0.341908024 -0.044201485  0.02837
7053
##   ...
## [149,]  1.37278779  1.011254419 -0.933395241  0.02612
8648
## [150,]  0.96065603 -0.024331668 -0.528248807 -0.16307
8032
```

Nás však nezajímají všechny 4 komponentní skóre, poněvadž se snažíme o redukci dimenzí a z výše uvedeného výstupu vyplývá, že dvě dimenze nám postačují, a proto nám stačí zobrazit pouze první dva sloupce předchozího výpisu.

```
fitKomponenty$scores[,1:2]
```

```
##              Comp.1       Comp.2
##   [1,] -2.26470281  0.480026597
##   [2,] -2.08096115 -0.674133557
##   [3,] -2.36422905 -0.341908024
##   ...
## [149,]  1.37278779  1.011254419
## [150,]  0.96065603 -0.024331668
```

Na jejich výpočet však program nepoužívá matici původních údajů ale standardizovanou matici původních údajů, které násobí s příslušnými koeficienty hlavních komponentů.

Na standardizaci používáme klasické normování (z-skóre):

$$z = \frac{x - \mu}{\sigma}$$

Pokud chceme normovat sloupce datové matice, využíváme k tomu funkci `scale()`. Princip výpočtu si můžeme ukázat na příkladu prvního objektu.

Skutečné hodnoty prvního objektu jsou:

```
data02[1,1:4]
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1         3.5          1.4         0.2
```

Hodnoty prvního objektu po normování sloupců jsou:

```
scale(data02[,1:4])[1,1:4]
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##   -0.8976739    1.0156020   -1.3357516   -1.3110521
```

Koeficienty pro první komponentu jsou:

```
loadings(fitKomponenty)[,1]
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.5210659   -0.2693474    0.5804131    0.5648565
```

Komponentní skóre pro první objekt – první komponenta:

```
sum(scale(data02[,1:4])[1,1:4]*loadings(fitKomponenty)[,
1])
```

```
## [1] -2.257141
```

Koeficienty pro druhou komponentu jsou:

```
loadings(fitKomponenty)[,2]
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##   0.37741762   0.92329566   0.02449161   0.06694199
```

Komponentní skóre pro první objekt – druhá komponenta:

```
sum(scale(data02[,1:4])[1,1:4]*loadings(fitKomponenty)[,
2])
```

```
## [1] 0.4784238
```

Výsledná komponentní skóre splňují podmínku vzájemné lineární nezávislosti a ty je možné použít pro další vícerozměrné analýzy.

Na konec je ještě možné vypočítat si takzvanou matici komponentních zátěží, jde o matici korelačních koeficientů původních proměnných k nově vytvořeným komponentám.

```
cor(cbind(fitKomponenty$scores[,1:2], data02[,1:4]))[,1:
2]
```

```
##                              Comp.1                C
omp.2
## Comp.1        1.000000000000000000 -0.00000000000000
09585
## Comp.2       -0.000000000000009585  1.00000000000000
00000
## Sepal.Length  0.8901687648612949255  0.36082988811302
39666
## Sepal.Width  -0.4601427064479081674  0.88271626916238
39930
## Petal.Length  0.9915551834193611080  0.02341518837916
51333
## Petal.Width   0.9649789606692488197  0.06399984704374
62413
```

Z matice je zřejmé, že se splnil předpoklad o tom, že výsledné komponenty jsou vzájemně lineárně nezávislé a jak můžeme vidět v daném výstupu, první komponenta velmi dobře vysvětluje délku okvětního lístku (Petal.Length), šířku okvětního lístku (Petal.Width) a délku kališních lístků (Sepal.Length). Druhá komponenta je vhodná na vysvětlení šířky kališních lístků (Sepal.Width).

6.4 Zpracování dat z oblasti faktorové analýzy

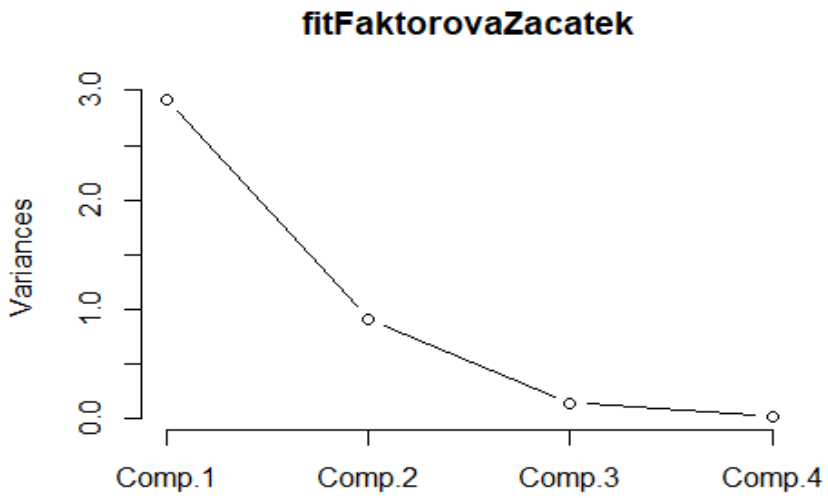
Příklad č. 4:

K demonstraci faktorové analýzy opět použijeme soubor, který jsme využívali v předchozích kapitolách, tj. soubor Kosatců s názvem „Iris“. Jak je již známo, soubor obsahuje čtyři proměnné a každé ze tří skupin je stejně velká a obsahuje 50 objektů. Celkový počet objektů je tedy 150.

Řešení v *R*:

Nyní budeme podobně, jako při metodě hlavních komponent, vycházet z korelační matice. Pro výpočet faktorové matice použijeme podobně jako v případě metody hlavních komponent funkci `princomp()`. Opět chceme vědět kolik faktorů potřebujeme na vysvětlení dostatečného procenta variability. Vhodný počet faktorů můžeme určit například na základě grafu vlastních čísel.

```
fitFaktorovaZacatek<-princomp(data02[,1:4], cor = TRUE)
plot(fitFaktorovaZacatek, type = "lines")
```




```
summary(fitFaktorovaZacatek)

## Importance of components:
##                           Comp.1 Comp.2  Comp.3   Comp.4
## Standard deviation     1.7084 0.9560 0.38309 0.143926
```

```
## Proportion of Variance 0.7296 0.2285 0.03669 0.005179
## Cumulative Proportion  0.7296 0.9581 0.99482 1.000000
```

Z daného výstupu vidíme, že jedním faktorem bychom vysvětlili téměř 73 % variability, což považujeme za dostačující, a proto budeme extrahovat 1 faktor (`nfactors = 1`). V případě, pokud bychom extrahovali více faktorů, doporučuje se takzvaná rotace faktorů, na kterou se nejčastěji používá metoda *varimax* (`rotate = "varimax"`). Využíváme funkci `principal()` z knihovny (balíčku) `psych`.

```
library(psych)

fitFaktorovaFinal<-principal(data02[,1:4], nfactors = 1)
fitFaktorovaFinal

## Principal Components Analysis
## Call: principal(r = data02[, 1:4], nfactors = 1)
## Standardized loadings (pattern matrix) based upon cor
relation matrix
##                PC1   h2    u2 com
## Sepal.Length  0.89 0.79 0.208   1
## Sepal.Width  -0.46 0.21 0.788   1
## Petal.Length  0.99 0.98 0.017   1
## Petal.Width   0.96 0.93 0.069   1
##
##                 PC1
## SS loadings    2.92
## Proportion Var 0.73
##
## Mean item complexity =  1
## Test of the hypothesis that 1 component is sufficient
.
##
## The root mean square of the residuals (RMSR) is  0.13
##  with the empirical chi square  28.19  with prob <  0
.00000076
##
## Fit based upon off diagonal values = 0.97
```

Pokud bychom chtěli extrahovat 2 faktory (příklad rotace faktorů metodou *varimax*):

```
fitFaktorovaFinal2faktory<-principal(data02[,1:4], nfact
ors = 2,           rotate = "varimax")
fitFaktorovaFinal2faktory
```

```
## Principal Components Analysis
## Call: principal(r = data02[, 1:4], nfactors = 2, rota
te = "varimax")
## Standardized loadings (pattern matrix) based upon cor
relation matrix
##                RC1   RC2   h2     u2 com
## Sepal.Length  0.96  0.05 0.92 0.0774 1.0
## Sepal.Width  -0.14  0.98 0.99 0.0091 1.0
## Petal.Length  0.94 -0.30 0.98 0.0163 1.2
## Petal.Width   0.93 -0.26 0.94 0.0647 1.2
##
##                        RC1  RC2
## SS loadings           2.70 1.13
## Proportion Var        0.68 0.28
## Cumulative Var        0.68 0.96
## Proportion Explained  0.71 0.29
## Cumulative Proportion 0.71 1.00
##
## Mean item complexity =  1.1
## Test of the hypothesis that 2 components are sufficie
nt.
##
## The root mean square of the residuals (RMSR) is  0.03
##  with the empirical chi square  1.72  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

V daných výstupech vidíme takzvanou matici faktorových saturací (v druhém případě po rotaci), podíly vysvětlené variability připadající na jednotlivé faktory, komunality a test, zda jsou jeden respektive dva extrahované faktory dostačující (`Test of the hypothesis that 1 component is sufficient`). I na základě *p*-hodnoty testu vidíme, že jeden extrahovaný faktor je dostatečný.

Uvažujme tedy jeden extrahovaný faktor, pracujeme s objektem **fitFaktorovaFinal**. Komunality vidíme i v sloupci `h2` matice faktorových saturací. Tyto komunality představují podíl rozptylu vysvětleného společnými faktory na celkovém rozptylu.

```
fitFaktorovaFinal$communality
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##       0.7924       0.2117       0.9832       0.9312
```

Výsledkem faktorové analýzy je, kromě odhadu matice faktorových saturací, také výpočet hodnot faktorových skóre, které následně může vstupovat do dalších analýz. Tato faktorová skóre se odhadují pomocí různých metod, primárně je nastavena ve funkci `principal()` regresní metoda (`method = "regression"`). Hodnoty faktorových skóre si můžeme zobrazit prostřednictvím příkazu `fitFaktorovaFinal$scores`.

```
fitFaktorovaFinal$scores
```

```
##               PC1
##   [1,] -1.321232
##   [2,] -1.214037
##   [3,] -1.379296
##   ...
## [149,]  0.800887
## [150,]  0.560449
```

6.5 Zpracování dat z oblasti diskriminační analýzy

Příklad č. 5:

Na základě datového souboru „*Iris*“ se pokusíme odhadnout Fisherovu lineární diskriminační funkci a ověřit účinnost diskriminace pro každou ze skupin, jejichž původní velikosti jsou 50 objektů.

Řešení v *R*:

Budeme používat lineární diskriminační analýzu. Na její aplikaci použijeme funkci `lda()` z knihovny (balíčku) `MASS`.

Do objektu **modelDiskriminacni** si nadefinujeme Fisherovou lineární diskriminační funkci, kde klasifikační proměnnou je druh kosatců. Proměnné, na jejichž základě se bude diskriminační funkce vytvářet jsou délka a šířka kališního lístku a délka a šířka okvětního lístku.

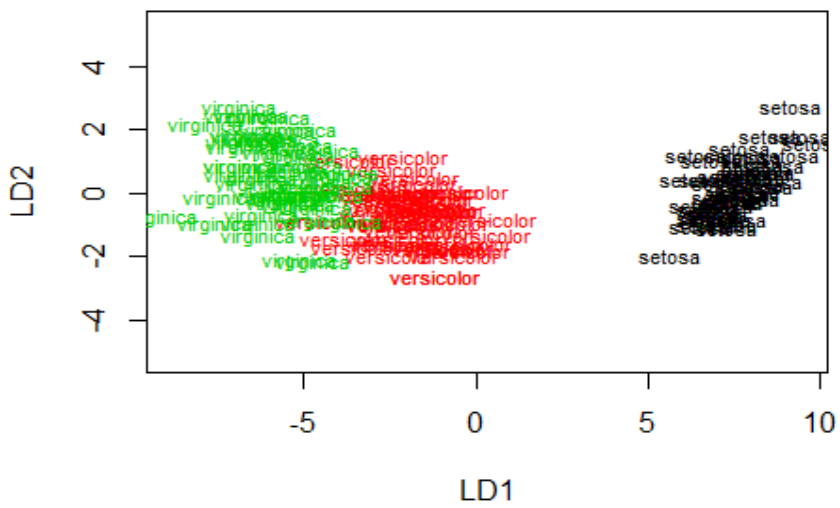
Ve výstupu je uvedena apriorní pravděpodobnost příslušnosti do konkrétní skupiny (máme tři skupiny po 50 objektů, proto jde v každé skupině o hodnotu 33,3 %), průměry jednotlivých skupin a z výstupu také vidíme, že pro dostatečné odlišení námi definovaných skupin by postačily dvě diskriminační funkce. Ve výstupu `Coefficients of linear discriminants` máme zobrazené hodnoty diskriminačních koeficientů. Ve výstupu `Proportional of trace` je vyjádřeno procento vysvětlené variability.

```
library(MASS)

modelDiskriminacni<-lda(Species~Sepal.Length+Sepal.Width
+Petal.Length+Petal.Width, data = data02)
modelDiskriminacni

## Call:
## lda(Species ~ Sepal.Length + Sepal.Width + Petal.Leng
th + Petal.Width,
##     data = data02)
##
## Prior probabilities of groups:
##     setosa versicolor  virginica
##     0.3333     0.3333     0.3333
##
## Group means:
##            Sepal.Length Sepal.Width Petal.Length Peta
l.Width
## setosa            5.006       3.428        1.462
0.246
## versicolor        5.936       2.770        4.260
1.326
## virginica         6.588       2.974        5.552
2.026
##
```

```
## Coefficients of linear discriminants:
##                    LD1     LD2
## Sepal.Length  0.8294  0.0241
## Sepal.Width   1.5345  2.1645
## Petal.Length -2.2012 -0.9319
## Petal.Width  -2.8105  2.8392
##
## Proportion of trace:
##    LD1    LD2
## 0.9912 0.0088
```

Dané objekty si můžeme vizualizovat prostřednictvím funkce `plot()`, přičemž jsme použili argument `col = as.integer(data02$Species)` z důvodu, aby se nám jednotlivé objekty zabarvily na základě příslušnosti do konkrétní skupiny. Na ose x máme hodnotu diskriminačních skóre pro první diskriminační funkci. Na ose y máme hodnotu diskriminačních skóre pro druhou diskriminační funkci.

```
plot(modelDiskriminacni, col = as.integer(data02$Species
))
```



Pomocí hodnot diskriminačních koeficientů můžeme vytvořit predikční model a ověřit jím zařazení jednotlivých objektů do skupin. Použijeme na to stejnou funkci jako v předchozím případě, v níž doplníme argument `CV = TRUE`.

```
modelDiskriminacniPredikce<-lda(Species~Sepal.Length+Sep
al.Width+Petal.Length+Petal.Width, data = data02, CV = T
RUE)
modelDiskriminacniPredikce

## $class
##   [1] setosa     setosa     setosa     setosa     set
osa     setosa
##   [7] setosa     setosa     setosa     setosa     set
osa     setosa
## ...

## [145] virginica  virginica  virginica  virginica  vir
ginica  virginica
## Levels: setosa versicolor virginica
##
## $posterior
##        setosa versicolor virginica
## 1   1.000e+00  5.087e-22 4.385e-42
## 2   1.000e+00  9.588e-18 8.888e-37
## 3   1.000e+00  1.984e-19 8.607e-39
## ...
## 149 3.097e-40  1.339e-05 1.000e+00
## 150 3.586e-33  2.059e-02 9.794e-01
```

Daná funkce nabízí několik výstupů, nás však hlavně zajímá úspěšnost klasifikace. Přiřazení jednotlivých objektů na základě diskriminačních funkcí můžeme vidět ve výstupu `modelDiskriminacniPredikce$class`, které vychází z takzvané posteriorní pravděpodobnosti (`modelDiskriminacniPredikce$posterior`) příslušnosti ke konkrétní skupině. Každý objekt se zařadí do konkrétní skupiny na základě maxima této posteriorní pravděpodobnosti.

Pokud si chceme vypočítat úspěšnost klasifikace, opět nám k tomu může pomoci kontingenční tabulka. Porovnáváme skutečné přiřazení (`data02$Species`) a přiřazení námi uvažovanou diskriminační funkcí (`modelDiskriminacniPredikce$class`), tuto kontingenční tabulku si uložme jako objekt s názvem **tabulka**.

```
tabulka<-table(modelDiskriminacniPredikce$class, data02$
Species)
tabulka
```

```
##
##              setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         48         1
##   virginica       0          2        49
```

Z uvedeného výstupu vidíme, že se nám správně podařilo klasifikovat všechny druhy setosa, 48 z 50 druhů versicolor (zbylé dva jsme nesprávně klasifikovaly jako virginica) a 49 z 50 druhů virginica (jeden jsme nesprávně klasifikovaly jako versicolor).

Prostřednictvím funkce `prop.table()` změníme tyto absolutní četnosti v kontingenční tabulce na četnosti relativní. Na základě funkce `diag()` si můžeme zobrazit relativní úspěšnost naší klasifikace na základě jednotlivých skupin a celkovou diskriminační schopnost modelu.

```
prop.table(tabulka)
```

```
##
##                setosa versicolor virginica
##   setosa     0.333333   0.000000  0.000000
##   versicolor 0.000000   0.320000  0.006667
##   virginica  0.000000   0.013333  0.326667
```

```
diag(prop.table(tabulka,2))
```

```
##     setosa versicolor  virginica
##     1.0000     0.9600     0.9800
```

```
sum(diag(prop.table(tabulka)))
```

```
## [1] 0.98
```

Jak můžeme vidět, podařilo se nám správně klasifikovat 100% druhů setosa, 96% druhů versicolor a 98% druhů virginica. Celková diskriminační schopnost modelu je velmi vysoká (na úrovni 98%), což je velmi dobrý výsledek, musíme si však uvědomit, že jsme ji určovaly ověřováním na původní datové množině a proto je nadhodnocena. V praxi často postupujeme tak, že si datovou množinu rozdělíme v nějakém vhodném poměru na dvě části (trénovací a testovací množinu). Na trénovací množině odhadneme koeficienty diskriminační funkce a její úspěšnost ověřujeme na testovací množině.

7 Kvalitativní výzkum

7.1 Základní principy a zaměření kvalitativního výzkumu

V této kapitole jsou popsány základní znaky kvalitativního přístupu, vysvětleny metody a principy, objasněna důležitost formulování výzkumné otázky v kvalitativním výzkumném šetření. Kapitola se věnuje nejen sběru dat, ale i jejich analýzou a interpretací, zvláštnostem hodnocení a závěrům z kvalitativního zkoumání. Při zpracování kapitoly vycházíme z českých i světových erudovaných zdrojů a především se opíráme o zavedenou teorii zpracovanou docentem Hendlem.

Kvalitativní výzkum má za cíl hlubší porozumění zkoumané skutečnosti. Je používán tehdy, kdy kvantitativní výzkum nepostačuje daným účelům a nelze jej proto pro řešení výzkumného problému použít, nebo je potřeba některou část kvantitativního šetření více prohloubit. Jednotná definice kvalitativního výzkumu neexistuje, jelikož se jedná o velice široký pojem, zahrnující značnou škálu přístupů. Může však být definován jako výzkum, který se zaměřuje na získání hlubokého pochopení zkoumané reality. Výzkumný problém může být

sledován na relativně malém počtu respondentů. Metodika obvykle nespoléhá na zapojování statistických analýz. Kvalitativní přístup je vhodný pro situace, kde se málo ví o subjektu, který má být analyzován. Role autora je především získat náhled kontextu, který je studován: jeho logika, uspořádání a jeho explicitní i implicitní pravidla. Provádí se pomocí delšího a intenzivního kontaktu s jedincem, institucí, skupinou apod.

Kvalitativní výzkum je zaměřen na interpretace subjektivních významů, popis kontextů získávání zkušeností, jednání nebo chování, přičemž se zajímá především o subjektivní teorie jedinců v daném prostředí. Kvalitativní výzkum je zejména vhodný, jestliže je cílem:

- porozumět subjektivním zkušenostem jedinců nebo skupin,
- porozumět působení různých faktorů (sociálních, kulturních, politických apod.) a interakcím mezi jedinci a prostředím,
- první seznámení s novou nebo složitou oblastí,
- podpořit kvantitativní výzkum, či získat hlubší vhled do problematiky, kterou odhalilo nejčastěji dotazníkové šetření.

Základní zaměření kvalitativního výzkumu je dáno relativně obecnými otázkami a ne hypotézami, které se mají testovat. Jak se výzkum rozvíjí, otázky se postupně upřesňují nebo se generují nové, které probíhající výzkum podrobněji specifikují. To může vést k potřebě pozměnit plán výzkumu a k zacílenému sběru dat. V tomto smyslu má plán kvalitativního výzkumu naléhavý charakter, jeho plánování je pružné, aby reagoval na okolnosti výzkumu a dosavadní výsledky. Při plánování kvalitativního výzkumu je nutné:

- určit účel a zaměření studie, vymezit hranice a kritéria pro to, které informace bude studie zahrnovat nebo vylučovat (hranice se ale mohou během výzkumu měnit),
- rozhodnout se, zda se kvalitativní přístup zvolí jako hlavní výzkumná strategie,
- určit, kde, kdy a od koho se budou sbírat data,
- určit fáze výzkumu,
- určit metody pro sběr dat,
- navrhnout organizaci sběru dat (zadávání otázek, přepis dat apod.),
- naplánovat, jak se budou data analyzovat, určit programový systém pro ukládání a zpracování dat s přihlédnutím ke zvolené organizaci sběru dat,
- naplánovat logistiku, přesný časový rozpis a popřípadě financování,
- věnovat se opatřením pro zajištění kvality celého postupu.

Na začátku kvalitativního výzkumného procesu bývá pozorování a sběr dat. Potom pátráme po pravidelnostech existujících v těchto datech, pátráme po významu těchto dat, formulujeme interpretaci, děláme předběžné závěry a výstupem mohou být nově formulované výzkumné otázky. Vytváříme si komplexní obraz o realitě tím, že vyhledáváme a analyzujeme jakékoliv informace, které přispívají k osvětlení výzkumných otázek. Kvalitativní výzkum se vždy provádí pomocí delšího intensivního kontaktu s terénem nebo situací jedince či skupiny jedinců. Autor se snaží získat integrovaný pohled na předmět studie, na jeho kontextovou logiku. Používají se většinou relativně málo standardizované metody získávání dat. Hlavním úkolem je objasnit, jak lidé v daném prostředí dojdou k pochopení toho, co se děje, proč jednají či nejednají určitým způsobem. Data se induktivně analyzují. Standardizace v kvalitativním výzkumu je slabá, a proto má kvalitativní výzkum poměrně

nízkou reliabilitu. Slabá standardizace výzkumu, volná forma otázek a odpovědí nevynucuje taková omezení jako kvantitativní výzkum. Proto potenciálně může mít vysokou validitu.

Kvalitativní výzkum vychází z interpretativního příkladu, ve kterém se dává důraz na porozumění významům lidského jednání a zkušenosti a na získání podrobných zpráv o pohledech zkoumaných jedinců. Výzkum vychází z představy, že svět je konstituován interaktivním jednáním a je pro jedince i skupiny významově strukturovaný. Proto je nutné ho vidět nejdříve očima sledovaného jedince nebo zkoumané skupiny. Tato primární rovina popisu a rozboru se na vyšší úrovni překračuje hledáním pravidel, norem a struktur, které jednotlivé interpretace ovlivňují a jichž si není jedinec obvykle vědom. Zvláštní roli hraje přitom koncept subjektivní teorie. Subjektivní teorie jsou vždy určené vlastními zkušenostmi a učením jedince v sociálním prostředí. Umožnuje jedinci orientaci a zvládání úkolů v každodenním životě. V nich jedinec disponuje repertoárem interpretací a řešení, které lze bezprostředně uplatnit a které se s velkou pravděpodobností osvědčí. Kvalitativní výzkum tak může být použit na prohloubení dosavadní teorie, nebo pro vytvoření teorie nové. Díky používání kvalitativních metod byla objevena nová témata, která byla doposud nezměřitelná kvantitativními metodami. V následující tabulce jsou uvedeny některé rozdíly mezi kvantitativním a kvalitativním výzkumným šetřením.

Kvalitativní výzkum	**Kvantitativní výzkum**
výzkumník je uvnitř, přítomen situaci	výzkumník je vně situace
osobní rozhovory, případové studie	dotazníky, testy, měření
menší počet respondentů	větší počet respondentů
časově náročný sběr dat	časově méně náročný sběr dat

méně strukturovanosti	vysoká míra strukturovanosti
indukce z výsledků	dedukce z výsledků
zkoumání problému do hloubky	zkoumání více okrajové
vytváření teorií	testování teorií
často nemožnost zobecnění	možné zobecnění

Tabulka 2: Rozdíly mezi kvantitativním a kvalitativním výzkumem

zdroj: autor

7.2 Metody kvalitativního šetření

Kvalitativními metodami realizujeme sběr dat, jejich kódování, vyhodnocování, interpretaci, ale i hodnocení kvality kvalitativního výzkumného šetření.

Základní metody kvalitativního výzkumu jsou:

- řízené rozhovory,
- případová studie,
- pozorování,
- tematická analýza (interpretace textů a dokumentů),
- audio či video záznam (interpretace záznamu),
- dotazník s otevřenými otázkami.

7.2.1 Řízený rozhovor

Rozhovor patří k nejčastějším metodám kvalitativního zkoumání. Nejčastěji se uvádí strukturovaný otevřený rozhovor, neformální rozhovor, fenomenologický rozhovor, narativní rozhovor, skupinová diskuze. „Tyto přístupy odlišuje rozsah určenosti a standardizace pořadí otázek při dotazování, počet osob, které se zúčastní rozhovoru, forma informací, jež se při dotazování získají, i situace rozhovoru. Každý z nich má slabiny a přednosti a vyžaduje poněkud odlišnou přípravu“ (Hendl, 2005, s. 168). U kvalitativních

rozhovorů jsou typy dat přepis z těchto rozhovorů, audio i videozáznamy a osobní komentáře. Sběr dat tvoří naslouchání vyprávění, kladení otázek a získávání odpovědí. Důležité je získat pravdivé odpovědi od respondenta. Rozhovor provádí pouze jedna osoba, je důležité, aby se jednalo o odborníka a aby se na začátku rozhovoru odstranily psychické zábrany. Důležitý je i způsob kladení otázek, otázky by měly být jasné, citlivé, neutrální a otevřené. Správná otázka dává dotazovanému možnost použít vlastní slova, bez toho, aby mu byla vnucována nějaká odpověď. Dotazovaný musí vyjádřit svůj vlastní názor a pocit. Zároveň může samostatně navrhovat vztahy a souvislosti. Nepředkládáme předem určené formulace odpovědí. Pokud je rozhovor veden správně, tak tazatel i respondent cítí, že jde o dvoustrannou rovnocennou komunikaci. Pro udržení důvěry by měl tazatel zpovídanému poskytnout příslušnou informaci o účelu otázky. Tazatel zároveň musí podněcovat respondenta, aby mu svěřoval další podrobnosti.

Velmi důležité je celkové vedení rozhovoru. Do přípravy před rozhovorem samotným patří mimo vypracování návodu i výběr vhodného místa nebo dohodnutý čas. Případné nahrávací zařízení je lepší mít nainstalované před začátkem rozhovoru. Rozhovor začne, až je vše připraveno.

Má následující strukturu:

- ***Úvod (introduction)*** – na začátku se tazatel představí a vysvětlí cíl studie, požádá o povolení k nahrávání, zodpoví všechny respondentovy otázky o povaze studie.
- ***Rozehřátí (warmup)*** – v této fázi se buduje vztah mezi tazatelem a respondentem. Může se začít otázkami o prostředí, kde se rozhovor odehrává.
- ***Hlavní rozhovor (main body of the interview)*** – zde se tazatel začíná ptát podle návodu.

- ***Zchladnutí (cool-off)*** – přijde na řadu, když se hlavní rozhovor chýlí ke konci. Tazatel může nasměrovat rozhovor do neformální roviny, aby se lépe ukončoval.
- ***Uzavření (closure)*** – poděkování a rozloučení.

V průběhu rozhovoru by si měl tazatel dát pozor i na tón hlasu. Jelikož jde o rozhovor z očí do očí, měl by si také dávat pozor na způsob, jak se ptá i jak reaguje na odpovědi, aby nevnesl do sběru dat nějaké předsudky. Je nutné, aby rozhovor byl v celém svém průběhu neutrální.

7.2.2 Případová studie

Nazývána též kazuistika (case study). Jedná se o detailní studium jednoho či několika málo případů. Účelem případové studie je aplikace získaných poznatků při porozumění případům obdobným. Předpokladem je soustředit se na jeden objekt

Je to metoda, která umožnuje zachycení složitostí, detailů, vztahů a procesů probíhajících v daném prostředí. Předpokládá, že podrobný výzkum jednoho případu přispěje k lepšímu porozumění a pochopení jiných, obdobných případů. Tyto případy je ovšem třeba vnímat a chápat v širším kontextu, eventuálně je srovnat s dalšími případy. Zkoumá, jaké jsou charakteristiky daného případu nebo skupiny porovnávaných případů. Na rozdíl od statistického šetření, které shromažďuje relativně omezené množství dat od jednoho nebo několika málo jedinců. Jde v ní o zachycení složitosti zkoumaného případu, o popis vztahů. Případová studie musí splňovat určité podmínky:

- stanovit typy otázek, na něž bude hledat odpovědi odkrýváním zkoumaného případu v terénu,
- vymezit roli výzkumníka,

- zvážit, zda bude zkoumat současný stav, nebo historii případu.

Každá případová studie má svůj vlastní logický rámec. Vývoj teorie je součástí záměru, zpravidla jsou také stanoveny určité komponenty nebo jednotky analýzy, okruhy nebo typy zdrojů, získání dat a způsoby jejich záznamu. Podstatný je výběr případu, který bude zkoumán tak, aby reprezentoval určitý typ nebo skupinu obdobných případů. Různorodost reality daných případů poskytuje více interpretací, na nichž se, jak již bylo uvedeno, podílejí aktéři, role výzkumníka a jeho asertivita je podstatnou formou analytické generalizace. Případové studie mohou zkoumat jednotlivé osoby, malé skupiny či sociální skupiny, organizace, instituce, události nebo vztahy.

Typy případových studií jsou:

- **Osobní případová studie**

Jedná se o podrobný výzkum určitého aspektu u jedné osoby. Výzkumník se věnuje minulosti, kontextovým faktorům a postojům, které zkoumané události předcházely. Zkoumá možné příčiny, determinanty, faktory a procesy, jež s ní měly souvislost.

- **Studie malé skupiny**

Výzkumník zkoumá jednu komunitu v určité lokalitě, případně lokalitu samotnou. Popisuje a analyzuje vzorce hlavních aspektů života této skupiny, zvláštnosti lokality apod.

- **Studium sociálních skupin**

Výzkumník se zabývá zkoumáním malých přímo komunikujících skupin (např. zaměstnanci jednoho podniku). Popisuje a analyzuje vztahy, rozdíly, aktivity ve skupině.

- **Studium organizací a institucí**

Výzkumník zkoumá firmy, školy a jiné organizace, uskutečnění programů, kulturu organizací, procesy změn a adaptací. Hledá nejlepší vzorce zavedení určitého typu řízení, chování, evaluace.

- **Zkoumání událostí, rolí a vztahů**

V tomto typu se výzkumník zaměřuje na určitou událost. Částečně se může překrývat se studiem sociálních skupin a organizací. Vytváří analýzu interakce členů skupiny, stereotypy apod.

Ve všech uvedených typech se může jednat o různé výzkumné záměry. Nejčastěji jde o objasnění pouze jednoho daného případu a proniknutí do hloubky problému v dané situaci nebo organizaci se záměrem porozumět okolnostem tohoto případu. Může jít i o objasnění určitého jevu vyskytujícího

se v realitě na jednom nebo několika případech jev reprezentujících, nebo zkoumá více případů s cílem ověřit určitou koncepci nebo inovaci a získat poznatky prokazující určité společné rysy poskytující vhodnost či nevhodnost této koncepce nebo inovace.

7.2.3 Pozorování

Vědecké pozorování hraje důležitou roli v rámci kvalitativního i kvantitativního výzkumu. Je to cílevědomé, soustavné a plánovité vnímání procesů a jevů, které směřuje k odhalení podstatných souvislostí a vztahů sledované skutečnosti. Vědecké pozorování se na rozdíl od laického pozorování dá definovat jako technika sběru informací založená na zaměřeném, systematickém a organizovaném sledování aspektů, které jsou předmětem zkoumání.

Pozorování musí splňovat několik podmínek:

- přesná organizovanost,
- průběh podle stanoveného plánu,
- přesná registrace pozorovaných jevů a procesů.

Další podmínky musí splňovat také pozorovatel. Měl by mít zdravé smyslové orgány, schopnost přesného odhadu, schopnost koncentrace pozornosti, oproštění od negativních vlivů, předsudků a zaujatosti, schopnost přesného vnímání, vedení bezprostředních a přesných záznamů.

V rámci metodiky vědecko-výzkumné činnosti stojí pozorování nejčastěji na začátku výzkumného procesu. Zúčastněné pozorování patří rozhodně mezi nejstarší metody zkoumání.

Pozorování jako metodu výzkumu dělíme na několik typů, které se mezi sebou mohou vzájemně kombinovat. První možnost dělení je podle míry standardizace, tedy podle toho, jak moc je pozorování formalizované:

- **Nestandardizované pozorování** má velmi nízký nebo žádný stupeň formalizace. Velmi často bývá u tohoto typu určen pouze cíl nebo předmět pozorování a ostatní aspekty pozorování jsou doplňovány průběžně.
- **Standardizované pozorování** má striktně danou formu. Předem je stanoven cíl a přesná podoba výzkumu, včetně například času nebo místa. Pozorovatel připraví záznamový arch s předem určenými pozorovacími kategoriemi, díky tomu může být výzkum kompatibilnější a vzniká tím i možnost spolupráce více výzkumníků na jednom projektu
- **Polostandardizované pozorování** je určitý kompromis mezi oběma hraničními metodami – některé projevy jsou zaznamenány formalizovaným záznamem, jiné aspekty výzkumu jsou zaznamenány nestandardizovanou formou.

Další dělení může být podle pozice pozorovatele ke zkoumaným osobám:

- **Otevřené pozorování** (někdy nazývané zjevné) – zkoumané osoby jsou informovány o výzkumu. Nebezpečím je možnost zkreslení výsledků kvůli narušení přirozeného chování pozorovaného subjektu či skupiny.
- **Skryté pozorování** – umožňuje záznam poznatků, aniž by o tom studované subjekty věděly. Zároveň s sebou přináší nebezpečí odhalení a s tím související komplikace v pokračování výzkumu. Poslední možností je rozdělení podle toho, zda pozorovatel vstupuje či

nevstupuje do skupiny.

- **Nezúčastněné pozorování** – pozorovatel zůstává mimo sledovanou skupinu
- **Zúčastněné pozorování** – pozorovatel se zapojuje do skupiny a spolupodílí se na jejím životě a aktivitách.

Důležitý je **plán pozorování**, protože je velmi jednoduché nevědomě filtrovat, co vidíme, v závislosti na našich očekáváních. Je tedy dobré použít pozorovací plán nebo kontrolní seznam, aby údaje z pozorování byly nezaujaté. Hlavní výhodou této metody je detailní popis pozorované skutečnosti. Dojde tak k pochopení problému v jeho hloubce, tím se také otevírá možnost rozsáhlé komparace dvou jevů. Pozorování probíhá v přirozeném prostředí, výzkumník zachycuje bezprostřední zkušenost s pozorovaným jevem, závěry z výzkumu proto bývají opravdu vypovídající. Další výhodou metody je, že pozorování ověří, zda se lidé chovají doopravdy tak, jak si myslí. Můžeme ověřit, zda informace, které respondenti vyplnili v dotazníku jsou pravdivé. Můžeme zaznamenávat i audio či videozáznamem.

7.2.4 Tematická analýza

Zpracování dat prostřednictvím metody tematické analýzy znamená dokonale určit výzkumné otázky, neboť jednotlivé kroky jsou vždy v souladu s těmito otázkami. Jedná se o výzkumný záměr otázky, výběr výzkumného vzorku, provedení analýzy a ověření validity. Prostřednictvím otázky zkoumáme, jaká témata jsou charakteristická pro daný problém či pro přístup respondentů k danému výzkumnému problému. Skrze tematickou analýzu ze získaných dat od respondentů vydefinujeme několik témat, které tvoří rámec teorie o daném

problému, popřípadě do jaké míry se názory respondentů shodují s těmito tématy.

Cílem analýzy je z čerstvých dat od přímých aktérů vytvořit právě tu rámcovou teorii např. o určitém přístupu někoho k něčemu, pomocí které v práci testujeme např. konkrétní opatření.

Jedná se o jednu z nejstarších metod obsahové analýzy při kvalitativním výzkumu, kterému nabízí snadný a flexibilní přístup. (Obsahová analýza /content analysis/ nebo také věcná analýza spadá však do kvantitativní analýzy.) Témata zkoumaného textu v rámci tematické analýzy jsou řazena do kategorií, které jsou definovány ještě před samotným zkoumáním. Díky redukci informací do vytvořených kategorií je velmi úspornou metodou. Tematická analýza může být použita ke spojení zdánlivě neslučitelných materiálů tak, aby dohromady dávaly smysl na základě společných témat. Bývá užívána k analyzování a získání kvalitativních informací o osobě, organizaci, určité situaci apod., bývá aplikována, aby se výzkum mohl posunout od pouhého čtení získaných dat k objevení vzoru a zformování specifické výzkumné otázky. Autoři využívají tematickou analýzu jako prostředek k získání hlubšího porozumění informacím, které v průběhu výzkumu získali (Merten, 2017).

Tematická analýza využívá koncept takzvaných ukotvených teorií, tedy teorií, které mají své předpoklady, vycházejí přímo z empirických dat skrze hledání vhodných témat, jejich kontrastů a výsledkem je vytváření tematických modelů. „Tematická analýza je flexibilním přístupem ke kvalitativním datům, který je využitelný napříč obory a výzkumnými otázkami" (Braun, Clarke, 2006, s.14).

Autorky pro co nejefektivnější zpracování dat navrhují následující strukturu. Pro spolehlivost doporučují pravidelnou revizi.

- **Seznámení se s daty**

Logický první krok zpracovávání dat. V případě rozhovorů autor opatří jejich autentický přepis. Autorky doporučují následně pečlivě a opakovaně přečíst všechna dostupná textová data. Počáteční opakující se vzorce si výzkumník zaznamenává, aby je mohl následně využít při tvorbě počátečních kódů. Nejdůležitějším aspektem prvního kroku je porozumění obsahu.

- **Generování počátečních kódů**

Ve druhém kroku výzkumník zaznamenává vzorce, které mu přijdou zajímavé významné, nebo které se opakují. Tyto počáteční kódy jsou spíše úzkými specifiky textu než tématy, ale umožňují zasazení výpovědí respondentů do kontext.

- **Vyhledání témat**

Součástí třetího kroku je interpretace nabytých kódů do témat. Relevantní data jsou roztříděna do odpovídajících témat. Myšlenkový proces výzkumníka by měl sledovat jednotlivé kódy, témata a případně i podtémata.

- **Revize témat**

Výzkumník se vrací k předchozím krokům a reviduje je. Zvažuje separaci myšlenek do více kódů, roztřízení témat i jejich případné vyřazení. Mezi tématy by měly být zřejmé rozdíly a data v jednotlivých tématech by spolu měla souviset. Revize se dělá jak ve vztahu k jednotlivým kódům (krok č. 2), tak ve vztahu k tématům (krok č. 3).

- **Definování a pojmenování témat**

Výzkumník definuje pevná témata, pojmenuje je a případně definuje existující podtémata. Tento krok po výzkumníkovi vyžaduje určení toho, co jednotlivá témata definuje a co je rozlišuje tak, aby každé téma bylo v celku konzistentní a byla zřejmá jeho podstata.

- **Sepsání zprávy**

Posledním krokem je přetvoření analýzy do srozumitelné zprávy. Výzkumník využívá příklady související s jednotlivými tématy, výzkumnou otázkou a dalšími prameny. Výzkumná zpráva musí být pro čtenáře přesvědčivá, její podstata zřejmá a validní. Zpráva by neměla být pouhým popisem témat a poskytnout analýzu podpořenou empirickou evidencí adresující výzkumnou otázku (Braun, Clarke, 2006).

7.2.5 Dotazník

Dotazníkové šetření je metoda především kvantitativního výzkumu. Je nutné přesně formulovat cíl a úlohu dotazníku ve vztahu ke zkoumanému výzkumnému problému. Autor vytvoří otázky, které daným způsobem popisují výzkumný problém. Pro sestavení efektivního dotazníku se používají tři typy otázek:

- uzavřené,
- polouzavřené,
- otevřené.

V rámci dotazníkového šetření je možné zařadit právě kvalitativní linii výzkumu pomocí **otevřených otázek**. Otevřené otázky nabízejí jako odpověď pouze textové odpovědi. Jsou určeny ke zkoumání specifických odpovědí respondentů, které není možné vyjádřit jinak než slovním popisem. V odpovědi na otevřenou otázku se může respondent vyjádřit svými slovy podle vlastního uvážení. Nedostává na výběr z připravených variant odpovědí. Výhodou otevřených otázek je především skutečnost, že umožňují získat odpověď, která autora dotazníku nemusela napadnout. Dále podněcují respondenta k hlubšímu zamyšlení nad tématem a věrněji zachycují pohled

respondenta na otázku, jelikož není omezen variantami odpovědi. Autor dotazníku tak může získat originální myšlenky.

Efektivnost otevřené otázky zpravidla přímo souvisí s tím, nakolik respondent pochopil otázku a s jeho kompetentností soudit o dané věci. Proto **otázky musí být jasně formulované**, všichni respondenti by jim měli rozumět stejným způsobem. Jasná otázka pro autora dotazníku nemusí být jasná pro respondenty. Otevřené otázky je velmi vhodné použít v předvýzkumu a po vyhodnocení získaných odpovědí přeformulovat na otázku uzavřenou nebo otázku polootevřenou tím, že se provede obsahová analýza získaných odpovědí a jejich kategorizace. Podobná kategorizace, resp. klasifikace se provádí i v případě zařazení otevřené otázky do hlavního šetření. Získané odpovědi se kódují na základě vypracovaných kódových klíčů.

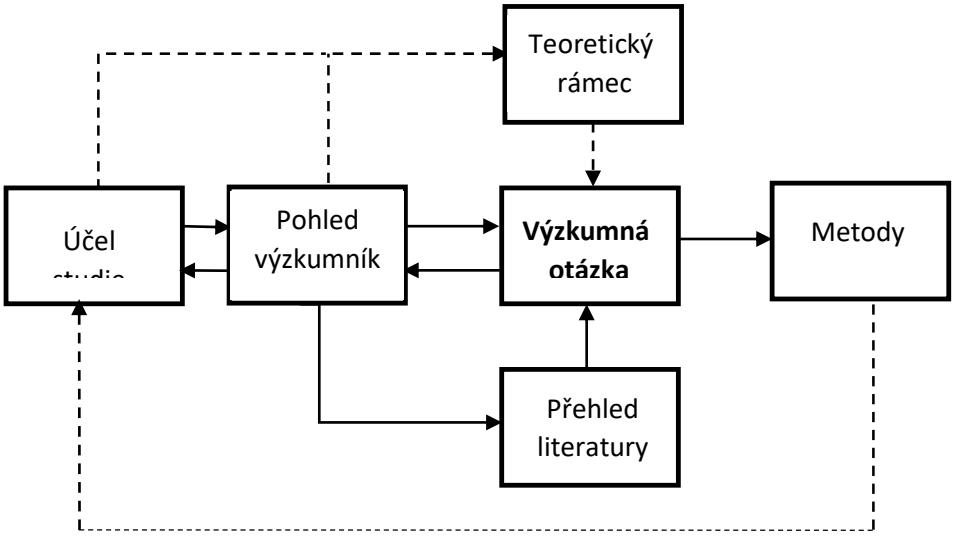
Otevřené otázky začínají často slovem: **proč, jak, čím, kdo, co**. Jde o to zjistit chybějící informace. Otázky by neměly být sugestivní, to znamená, že by neměly vyjadřovat postoj tazatele k tomu, na co se ptá (autor do otázky zjevně či skrytě dává požadovanou odpověď). Správně postavená otázka má velký vliv na odpověď respondenta.

7.3 Kvalitativní dotazování

7.3.1 Výzkumná otázka

Účel každé vědecké studie musí být propojen s metodami, které se použijí k řešení, a zároveň s výzkumnou otázkou. Výzkumná otázka je způsob, jakým uvažujeme o daném tématu ze svého vlastního hlediska. Následující obrázek ukazuje návaznost mezi určením účelu studie, pohledem výzkumníka, teoretickým rámcem, zvolenými metodami a výzkumnou otázkou.

Výzkumné otázky začínají obecným problémem, který je postupně zužován,



Obrázek 63: Návaznost od určení studie k výzkumné otázce zdroj: Hendl, 2005

vycházejí tedy z obecného určení cíle výzkumu, který převádějí do konkrétnější podoby. Tento úzce vymezený pohled pak soustředí zájem autora na zcela určitou oblast zkoumání. Výzkumné otázky udávají výzkumu směr, poskytují rámec pro další interpretaci a závěry. Obvykle autor formuluje jednu zastřešující výzkumnou otázku, ke které lze stanovit dílčí výzkumné otázky. Je dobré formulovat si otázku jako výzkumný problém. **Výzkumný problém**

znamená, co chce autor řešit a na které otázky chce odpovědět. Nejen samotná volba, ale i formulace předpokládá zkušenosti a přehled v dané. Definitivní formulace výzkumné otázky by měla dozrávat a vznikat postupně. Musí předcházet prostudování literatury, zmapování toho, co se zjistilo a popsalo v dané oblasti a také jakými metodami.

Častým nedostatkem při formulaci výzkumné otázky či výzkumného problému je, že autor si stanovil téma (oblast, kterou chce zkoumat), nikoli výzkumný problém. Problém tedy vymezí příliš široce. V otázce musí být zcela konkrétně a přesně, co chceme zkoumat. Výzkumný problém musí prohlubovat poznání problematiky, nesmí být triviální, jednoduchý, nemůže se stát, aby odpověď byla jasná – ano, ne – a nemá smysl to dále složitě dokazovat. I při psaní kratších textů (třeba seminárních prací) je proto dobré na výzkumné otázce pracovat – i zde může autor udělat zásadní chyby, například zabývat se příliš mnoha problémy najednou a nedospět tak k žádnému konkrétnímu závěru.

Výzkumné otázky v kvalitativním výzkumu se většinou týkají třech oblastí:

- popisu a interpretaci významů, které přisuzují jedinci situacím a jednáním,
- oblasti jazyka jako prostředku komunikace,
- vytváření teorií zkoumáním konfigurací a kontingencí v kvalitativních datech.

7.3.2 Techniky dotazování

Mezi základní metody a techniky dotazování při sběru dat patří různé typy rozhovorů, dotazník s otevřenými otázkami, stimulované vzpomínání,

komentování videozáznamů, myšlení nahlas, technika kritických událostí. Otázky by vždy měly být neutrální, citlivé a jasně formulované.

Mezi základní taktiky kvalitativního dotazování řadíme **šest typů otázek:**

- otázky vztahující se ke zkušenostem,
- otázky vztahující se k názorům,
- otázky vztahující se k pocitům,
- otázky vztahující se k chování,
- otázky vztahující se ke znalostem,
- otázky vztahující se k vnímání.

Jde o otázky typu „Co myslíte…Jaký je váš názor na…Co byste si přál…Co tedy cítíte…Na co se vás ptali.“ Pro správné řazení otázek neexistují pravidla. Tazatel pouze musí dbát nato, aby jeho otázky byly srozumitelné a jasné, stejně tak se nesmí pokládat více otázek najednou.

K dotazování patří různé typy a druhy otázek**,** ať už ve formě mluvené (rozhovor) či psané (dotazník). Kladení otázek není jednoduchá záležitost. **Nejde o to, na co se ptáme, ale jak se ptáme.** Existuje několik zásad pro tvorbu otázek:

- **srozumitelnost otázky** – je lepší nepoužívat žádné cizí a odborné výrazy, aby otázce porozuměli všichni respondenti, další problém může nastat u pojmů, které si každý jedinec může vyložit podle svého (spravedlnost, svoboda apod).,
- **jednoznačnost otázky** – respondent musí přesně vědět, na co se ho ptáme,
- **psychologická přijatelnost otázky** – jde o užívání taktních formulací v souvislosti s nepříjemným sdělením a událostmi,
- **formulace dotazu** – dotaz samotný nesmí sám respondentovi vnucovat správnou odpověď,

- **neverbální projevy** – tazatel si musí dávat pozor na své neverbální projevy, které by mohly ovlivnit odpovědi respondenta,
- je důležité zamyslet se, než položíme respondentovi otázku Proč?, abychom se nedostali do oblastí, které jsou pro respondenta nepříjemné odhalovat, či úplně mimo náš výzkum.

Rozeznáváme několik typů a druhů otázek. Typ otázky vyjadřuje její konkrétní podoba a druh otázky určuje její funkce v rámci konkrétního dotazování.

A/ Podle toho, jak je otázka standardizovaná:

- **Otázky volné** – nenabízejí respondentovi žádnou variantu odpovědi, proto má úplně volný prostor ke svému vyjádření. Nevýhodou je velmi náročné zpracování takového dotazníku.
- **Otázky uzavřené** – odpovědi z výběru z nabízených variant, tento druh otázek se používá hlavně v dotazníku a dobře se zpracovává, musíme však počítat s mírným zkreslením výsledků, protože respondenti mohou volit odpověď na základě jakéhosi kompromisu, pokud nenajdou jednoznačně správnou odpověď.
- **Otázky polouzavřené** – u těchto otázek najdeme nabízené varianty i místo pro vlastní odpověď.

B/ Dělení podle počtu variant a způsobu práce s nimi:

- **Otázky dichotomické** – mají dvě možné odpovědi.
- **Otázky polytomické** – mají více variant odpovědí než dvě. Dále se dělí na:
  - otázky výběrové – respondenti musí z nabízených variant vybrat pouze jednu jedinou

- otázky výčtové – respondent vybírá všechny možnosti, které považuje za adekvátní

C/ Dělení podle funkce, kterou otázky mají v dotazování:

- **Výzkumné otázky** – prostřednictvím nich získáváme potřebné informace a údaje.
- **Filtrační otázky** – rozčleňují respondenty podle toho, jak dalece se mohou ke zkoumané oblasti vyjádřit.
- **Grafické otázky** – slouží ke zpestření dotazování a zpřesnění odpovědí.

D/ Další možné typy otázek:

- **Otázky kontrolní** – určitý dotaz může být několikrát jinak formulovaný a tyto odpovědi se poté porovnávají (v dotazníku i rozhovoru).
- **Otázky kontaktní** – otázky mají za úkol navázat s respondentem vztah, bývají na začátku dotazování a snaží se respondenta „naladit“.

Důležité je i řazení otázek. Otázky se kladou v předem přesně daném pořadí, tomuto pevnému stanovení otázek se říká **dramaturgie dotazování**. Na začátku rozhovoru nebo dotazníku by se měly nacházet úvodní otázky, které navazují kontakt s respondentem. První čtvrtinu by pak měly tvořit snadné otázky, pro respondenta neproblémové, zajímavé a nekonfliktní. Druhou čtvrtinu by měly tvořit ty nejtěžší otázky, které jsou pro výzkum klíčové. Ve třetí čtvrtině bychom měli nalézt jednodušší otázky, přesto pořád důležité pro výzkum. Zde je vhodné dotazování nějakým způsobem zpestřit, abychom nepřišli o pozornost respondenta. Na závěr je vhodné zařadit otázky lehčího charakteru a méně důležité (Hendl, 2005).

Individuální hloubkový rozhovor

Individuální rozhovor může být řízený rozhovor pomocí návodu, strukturovaný rozhovor s otevřenými otázkami, neformální rozhovor, narativní rozhovor, fenomenologický rozhovor. Cílem je odhalit často velmi hluboce zakořeněné příčiny respondentových názorů nebo chování či jednání. Tazatel vždy podněcuje respondenta k diskuzi bez zábran na daná témata, umožňuje respondentovi vyjádřit své názory a pocity k předmětu výzkumu. Důležité jsou nejen odpovědi respondenta, ale zaznamenáváme také jeho reakce. Hloubkový rozhovor trvá zpravidla více jak 1 hodinu a může být se svolením respondenta nahráván pro pozdější analýzu.

Individuální rozhovor rozdělujeme na několik typů:

- rozhovor pomocí návodu – přesně daný seznam otázek, které je nutné v rámci rozhovoru probrat. Takový seznam pomáhá udržet zaměření rozhovoru. Někdy se jedná o problémově zaměřený rozhovor, kdy je rozhovor orientován na jeden konkrétní problém,
- strukturovaný rozhovor s otevřenými otázkami – cílem je co nejvíce minimalizovat efekt tazatele na kvalitu rozhovoru,
- neformální rozhovor (nestrukturovaný rozhovor) - spontánní generování otázek, velká míra volnosti, některé otázky jsou přirozeně vytvářeny až na základě vyprávění respondenta,
- polostrukturovaný rozhovor – stojí mezi strukturovaným a nestrukturovaným rozhovorem. Tento typ má předem daný soubor témat a volně přidružených otázek, ale jejich pořadí, volba slov a formulace může být pozměněna, případně může být něco dovysvětleno. Konkrétní otázky mohou být vynechány, nebo naopak přidány,
- narativní rozhovor – specifická podoba rozhovoru, kdy je respondent vybídnut ke zcela volnému vyprávění nějakého tématu (v biografii

zejména – životní události, zážitky), určujeme předmět vyprávění a zároveň se hledá subjekt, u něhož je jistota, že bude schopen vyprávět. 4 fáze – stimulace, vyprávění, kladení otázek pro vyjasnění nejasností, zobecňující otázky. Předpokládá se, že volné vyprávění odhalí subjektivní zkušenosti, což pomocí přímého dotazování nejde,

- fenomenologický rozhovor-rozhovor zaměřený na historii dotazovaného. Rozděluje se na tři části – první rozhovor (historie života jedince – jak k tomu u vás došlo?), druhý rozhovor (podrobnosti zkušeností – popište svoji zkušenost…), třetí rozhovor (reflexe jeho zkušenosti – jaký smysl má váš život…).

Skupinová diskuze (focus group)

Jedná se o moderátorem řízený rozhovor malé skupiny vybraných osob na stanovené téma.

Příspěvky diskutující neadresují moderátorovi, ale sobě navzájem a vstupují tak do vzájemných vztahů a přesvědčování. Cílem je konfrontace názorů diskutujících a pozorování způsobů jejich modifikace vzájemným ovlivňováním členy skupiny navzájem. Účastníci jsou příslušníky stejné sociální skupiny. Jejich homogenita zaručí podobné vnímání a zkušenosti a komunikační dovednosti na srozumitelné úrovni. Diskutující mají s diskutovaným problémem osobní zkušenost. Diskuse není ovlivňována jinými faktory. Účastníci nejsou ve vzájemném příbuzenském či jiném vztahu a moderátor je připraven "zkrotit" případné dominantní jedince. Skupinový rozhovor trvá poměrně dlouho 1,5-2 hodiny. Zúčastňuje se ho 8-12 „spotřebitelů“, nebo 6-7 odborníků-expertů.

Existují dva základní principy skupinové diskuze:

- skupinové interview a vyprávění - uvolňují se zde racionalizační schémata, psychické zábrany, diskutující lépe odhalují své myšlení a

postoje v běžném životě,

- brainstorming - cílem je najít nové pohledy na řešení daného problému a najít tato řešení. Je to typ skupinové diskuze, které se účastní nejčastěji experti. Výhodný jako nástroj podněcování tvořivosti. Nápady se necenzurují a všechny se zaznamenávají. Následuje vyhodnocení a zpracování získaných nápadů.

Dotazník

Dotazník je základní metodou kvantitativního šetření, ale dá se využít i v kvalitativním výzkumu. Dotazník by měl na první pohled upoutat pozornost. Vedle už zmíněné srozumitelnosti a přehlednosti je nutná:

- jazyková korektnost,
- jednoduchost vyplňování,
- typografická a grafická úprava.

Délka dotazníku by neměla překročit 20 otázek a doba vyplňování 10 minut. U velmi dobře motivovaného respondenta může délka dotazníku mít až 40 otázek, doba vyplňování však nesmí překročit 20 minut. U formulace otázek se držíme zásad pro tvorbu otázek, přidáme pouze stručnost – používat krátké, stručné věty. Jak už bylo řečeno nepoužívat sugestivní otázky, tj. takové, které svou formulací napovídají odpověď, a vyvarovat se haló-efektu, tj. řadě příbuzných otázek za sebou, kde se odpověď z první otázky přenáší i do ostatních.

7.4 Analýza a interpretace dat kvalitativního výzkumu

V kvalitativním výzkumu se většinou používá několik zdrojů dat. Cílem je získat co nejkomplexnější porozumění tématu. Význam pro lepší kvalitu dat spočívá v myšlence, že sběr informací pomocí více zdrojů (může jít o

respondenty, organizace, možné události) lépe vysvětlí různé stránky situace a zkušenosti, aby bylo možné zobrazit jejich komplexitu. Pro vyhodnocení toho, zda jsou data odpovídající a přiměřená, je nutné posoudit, jestli zvolené zdroje umožňují autorovi dobře prozkoumat subjektivní významy, jednání a sociální kontext relevantní k výzkumné otázce.

Data, která autor získal, zaznamenává tak, aby to umožnilo jejich analýzu a zároveň se uchovaly v nich obsažené subjektivní významy a sociální kontext. Znamená to tedy, že autor často přistupuje k doslovnému přepisování rozhovorů, přičemž se používá systém transkripčních značek pro zachycení jejich stylu a průběhu. Taková transkripce není vždy možná, protože obvykle vede k ohromnému množství dat. Elektronické zaznamenávání rozhovorů nebo terénních poznámek, poznámkování je užitečnou volbou, jež umožňuje analýzu materiálu jako celku. Existují však i případy, kde není elektronické nahrávání rozhovorů možné, protože působí rušivě. Ve zprávě o výzkumu je nutné přiblížit způsob záznamu, aby tento proces byl zcela transparentní. K průhlednosti výzkumného procesu přispívá i podrobná zpráva o všech metodologických krocích.

V rámci kvalitativního výzkumu je charakteristické intenzivní vzájemné působení autora výzkumu a respondentů. Toto hledisko je důležité, protože významy jsou dané do kontextu a není možné jim bez kontextu porozumět. Zaujatost výzkumníka přispívá k tomu, že výzkum bude lépe reagovat na projevy respondentů a požadavky situace. Toto může i pomoci výsledkům, které nebudou ovlivněny názory a předsudky autora. Důležité jsou také úvahy a přemýšlení výzkumníka o vlastních vstupních názorech a zkušenostech, aby je skutečně dokázal odlišit od nových informací z výzkumu.

Autoři kvalitativního výzkumu nejčastěji používají analýzu. **Analýza** materiálu má vést k odhalení a k popisu témat. Témata je možné odhalit v

procesu induktivního kódování nebo deduktivně pomocí literatury, ale také na základě vlastní empirie autora a v závislosti na položené otázce. Autoři se v analýze se zabývají získanými daty na několika úrovních. Zajímají se o jednotlivá slova, koncepty, verbální i nonverbální výrazy. Určitá slova potom mohou navést k tématům, jestliže se zkoumá jejich opakování v určitém kontextu a způsobu použití. Potom, co se prozkoumá text tematicky, přechází se k hledání vztahů uvnitř a mezi tématy. Hledáme například, jestli jsou určitá slova nebo koncepty korelovány s jinými slovy a tématy. Nebo se ptáme, zda určitá slova jsou asociována s nějakými neverbálními signály nebo citovými stavy. Také se můžeme přesvědčit, existují-li nějaké vztahy mezi neverbálními signály. Rovněž autory zajímá, jak respondenti s určitými vlastnostmi používají určité výrazy apod. Mnohdy se používá **induktivní analýza**, při které se témata pomalu odhalují z nasbíraného materiálu. Kvalitativní analýza vyžaduje kreativitu, aby bylo možné nestrukturovaná kvalitativní data významově uspořádat a propojit pomocí holistického vyprávění o sledovaném případě. Základními prvky většiny strategií vyhodnocení kvalitativních dat je **tematická analýza** a hledání vazeb mezi jevy (Hendl, 2002).

Kvalitativní analýza dat není numerická, nestojí na kvantifikacích, tj. cílem není spočítat, kolik respondentů prohlásilo určité tvrzení a také se nejedná o pouhou obsahovou analýzu, cílem není identifikovat témata a určit, kolikrát se ve výpovědi respondentů objevila.

Kvalitativní výzkum má kruhový charakter. Podrobný rozbor dat je dobré začít dělat bezprostředně ve chvíli, kdy máme první přepis rozhovoru či první terénní poznámky. Data pozorně procházíme a čteme bez ohledu na předpoklady, které jsme získali studiem literatury či bez ohledu na předem připravené kategorie a co možná nejvíce nad nimi přemýšlíme. Směřujeme tedy k tomu, abychom s daty nemanipulovali, abychom se nepokoušeli svoje

předpoklady a teoretická východiska aplikovat na data. Proto data čteme opakovaně a první čtení by mělo být co možná nejvíce induktivní. Obzvláště v prvních fázích analýzy výpovědí (ale i komentářů) je nutné dát pozor, abychom do komentářů neprojektovali své vlastní předpoklady. Zkusme se držet toho, co nám text říká. Držme se co nejvíce dat, vycházejme z výpovědí, komentářů a jejich vzájemných propojení. Analýza dat je časově velmi náročná, data je nutné číst opakovaně.

V kvalitativní analýze je častým termínem kód nebo krátká fráze. Vyjadřují nejvýznamnější nebo shrnující znak určité skupiny textových nebo vizuálních dat. **Kódy** dovolují uspořádat velké množství výpovědí, umožňují je porovnávat. Pod jedním kódem jsou vedle sebe výpovědi, které spolu tematicky souvisí. **Text** je tak rozdělen na jednotky, těmto jednotkám jsou přidělena jména a s takto pojmenovanými jednotkami výzkumník dále pracuje. Kód reprezentuje téma dané významové jednotky.

Autoři mnohdy jako základ své analýzy uvádějí tzv. **zakotvenou teorii** (grounded theory). Tento postup, jehož autory jsou Glaser a Strauss (1967) a Strauss a Corbinová (1999) je metodou navržení teorie, která se induktivně generuje pomocí dat. Výzkumníci shromažďují a analyzují data pomocí metody nepřetržitého porovnávání, aby postupně vystihly konfigurace v datech na stále větší úrovně abstrakce pomocí induktivně navržených obecných kategorií a vztahů mezi nimi. Tato metoda návrhu teorie patří mezi nejlépe propracované a popsané metody v kvalitativním výzkumu. Někdy se však zdá, že odkaz na zakotvenou teorii se stává pouhým stereotypem. Názor, že každá analýza kvalitativních dat musí vést k induktivně navržené teorii, je chybný. I v kvalitativním výzkumu pracujeme někdy s apriorními hypotézami nebo s předem definovanými kategoriemi.

V rámci kvalitativní analýzy neexistuje striktní hranice mezi analýzou a interpretací. Interpretací myslíme analýzu + syntézu dohromady. Při výkladu a vysvětlování dat využíváme odbornou teoretickou literaturu, zjištění z jiných výzkumů, vlastní zkušenosti apod. Je třeba jasně (i graficky) odlišit, kde končí data a kde začíná výklad autora. Interpretace je aktem hledání významu daného výroku či akce. Autor ji obvykle umisťuje hned za část s přesnými daty. Nestačí konstatovat, že to, co říkají naši respondenti, je zajímavé, ani jednoduše vlastními slovy opakovat či variovat datový úryvek. Cílem interpretace je udělat určitý myšlenkový skok. Víme, že se něco děje a je třeba nabídnout vysvětlení, proč se to děje. Ke každé výpovědi z dat by měla být připojena rozpracovaná analýza i syntéza. Pokud nejsme s to výpověď uspokojivě interpretovat, nemá v našem textu co dělat. Z hlediska etických doporučení nesmí být publikována žádná data, která by mohla umožnit identifikaci respondentů. Autor musí respondenty informovat o principu důvěrnosti a o způsobech anonymizace dat. Výzkum nikdy nesmí způsobit nikomu z účastníků výzkumu žádnou profesní ani psychickou újmu.

Postup kvalitativní analýzy dat:

- třídění dat, jejich kódování a kategorizace,
- formulace základních tvrzení,
- interpretace,
- komparace,
- teoretická generalizace.

Výsledek kvalitativní analýzy:

- seznam a podrobný popis klíčových témat,

- teorie, hypotézy k dalšímu ověřování,
- chronologie, sekvence,
- schéma, model,
- typologie, kategorizace.

Výsledky kvalitativního výzkumu mají být dostatečně podrobné, aby bylo možné porozumět jednání popisovaných účastníků a jejím zkušenostem v daném kontextu. Autor by měl v této části citlivě používat odbornou terminologii. Výklad postupuje od popisu prostředí a interakcí, využitím citací z projevů účastníků a popisu exemplárních situací, v diskuzi jejich významů a důležitosti. Provázání mezi všemi výsledky a daty mají být srozumitelné. Kvalitativní výzkum většinou zdůrazňuje, že výsledky je nutné uvažovat lokálně v daném kontextu. Obvykle se neusiluje o zobecnění na větší populaci. Zobecňuje se vzhledem k teorii a je úkolem každého čtenáře, zda se z výsledků poučí a teorii použije, či ji bude aplikovat. Výsledky se obvykle podávají textem, který vysvětluje subjektivní definice významů fenoménů v kontextu daného prostředí a situace. Cílem je přiblížit čtenáři zkušenosti, které se popisují z pohledu účastníků. Proto popisy prostředí výzkumu, výsledků a interpretací bývají natolik podrobné, aby čtenář mohl určit jejich aplikovatelnost pro vlastní situaci.

Autoři kvalitativních výzkumů vysvětlují především subjektivní zkušenosti, jednání a kontext zkoumaných jedinců. Dávají přednost zachycení stanovisek a pohledů respondentů. Hlavním kritériem kvalitativního výzkumu je míra, jak se podařilo autenticky zachytit a prezentovat pohledy zkoumaných jedinců. Zpráva má také přiblížit mocenské vztahy mezi výzkumníkem a zkoumanými jedinci a míru reflexivity, kterou výzkumník vnáší do procesu výzkumu a do svých interpretací.

Každý kvalitativní výzkum chce porozumět zkoumanému případu do větší hloubky. Uznává jeho složitost, protože každý případ může vygenerovat mnoho dat. V dané studii lze použít kvalitativní postup pouze pro jeden nebo výjimečně pro několik málo případů.

V poslední době se věnuje kvalitativní linii výzkumů velká pozornost. To však vyžaduje velké porozumění jak od výzkumníků, respondentů, tak i od konzumentů.

Kvalitativní výzkum se snaží interpretovat pohledy subjektů na zkoumaný předmět tím, že výzkumník přejímá jejich perspektivu. Využívá se podrobný popis každodenních situací. Jde o porozumění akcím a významům v jejich sociálním kontextu. Při kvalitativním výzkumu se neomezuje počet proměnných ani vztahy mezi nimi, o jejich redukci rozhodují samy zkoumané subjekty. Jsou upřednostňovány otevřené a nestrukturované výzkumné plány, analýza vychází z velkého množství informací o malém počtu jedinců. Převažuje zájem o reálné celky, individuální osudy a vzájemné působení všech zúčastněných. Úkolem kvalitativního výzkumu je vytvoření holistického obrazu zkoumaného předmětu, zachycení toho, jak jednotliví účastníci situace interpretují a zachycení nové interpretace těchto interpretací.

Seznam literatury

Braun, V., Clarke, V. (2018*). Qualitative Research in Psychology. Using thematic analysis in psychology* [online]. 2018, 77-101 [cit. 2018-03-21]. Dostupné z: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa

Corbin, J., Strauss, A. L. (1999). *Základy kvalitativního výzkumu*. Boskovice: Albert.

Gan, G., Ma, Ch., Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM, Philadelphia.

Glaser, B. G., Strauss, A. L. (1967). *The Discovery of grounded theory*. New York: Aldine.

Hendl, J. (2005*). Kvalitativní výzkum – základní metody a aplikace*. Praha: Portál.

Hendl, J., Blahuš, P. (2002). *Metodologie výzkumné práce.* Praha: UK.

Hindls, R., Hronová, S., Seger, J., Fischer, J. (2007). *Statistika pro ekonomy*. Praha: PROFESSIONAL PUBLISHING.

Král, P. (2009). *Viacrozmerné štatistické metódy so zameraním na riešenie problémov ekonomickej praxe*. Banska Bystrica: Univerzita Mateja Bela.

Löster, T., Řezanková, H., Langhamrová, J. (2009). *Statistické metody a demografie*. Praha: VŠEM.

Löster, T., Pavelka, T. (2013). Hodnocení výsledků shlukování v ekonomických úlohách. *Forum statisticum slovacum*. roč. 9, č. 7.

Löster, T. (2014). *Metody shlukové analýzy a jejich hodnocení*. Slaný: Melandrium.

Löster, T. (2016). *Příklady ze statistiky*. Slaný: Melandrium.

Marek, L., Malá, I., Pecáková, I., Löster, T., Čabla, A. (2015). *Statistika v příkladech*. Praha: Kamil Mařík – Professional Publishing.

Meloun, M., Militký, J., Hill, M. (2005). *Počítačová analýza vícerozměrných dat v příkladech.* Praha: Academia.

Merten, K. (2017). *Typologie metod obsahové analýzy* [online]. [cit. 2019-7-15]. Dostupné z: archiv pořízený dne 2018-08-07.

Miles, B., Huberman, A.M. (2015). *Qualitative Data Analysis for Health.* Dostupné z: https://books.google.cz/url?id=Um7JWua8LD4C&pg=PR17&q=http://www.springer com/shop&clientid=ca-print-springer.

Pecáková, I. (2011). *Statistika v terénních průzkumech*. Praha: PROFESSIONAL PUBLISHING.

Radváková, V., Löster, T., Mazouch, P., Sigmund, T., Vltavská, K. (2018). *Metody vědecké práce.* Praha: Oeconomica.

R Core Team (2017). R: A language and environment for statistical computing.

R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Řezanková, H., Húsek, D., Snášel, V. (2009). *Shluková analýza dat*. Praha: PROFESSIONAL PUBLISHING.

Řezanková, H. (2011). *Analýza dat z dotazníkových šetření*. Praha: PROFESSIONAL PUBLISHING.

Řezanková, H., Löster, T. (2013). *Základy statistiky*. Praha: Oeconomica.

Stankovičová, I., Vojtková, M. (2007). *Viacrozmerné štatistické metódy s aplikáciami.* Bratislava: Ekonómia.

Český statistický úřad: www.czso.cz

Švaříček, R. (2015) *Je zakotvená teorie teorií? Studia minora Facultatis philosophicae universitatis Brunensis.* Vol. 10, No. 1, s. 21-44.

Walker, I., (2013) *Výzkumné metody a statistika*. Praha: Grada.

Rejstřík

A

aglomerativní shlukování 26
analýza dat 140, 145
autokorelace 9

C

celkový F test 11

D

dendrogram 27, 38
dílčí t test 11, 24
diskriminační analýza 67
diskriminační funkce 67, 68, 69, 71, 72, 73, 75, 78, 79, 110, 114
diskriminační koeficienty 68
divizivní shlukování 26
dotazník 119, 132
dramaturgie dotazování 134

E

Euklidova vzdálenost 29, 40

F

faktor 39, 64, 107, 109
faktorová analýza 56
faktorová zátěž 56, 57
fuzzy shlukování 26, 30

H

heteroskedasticit 10
hierarchické metody shlukování 26

I

individuální rozhovor 135

induktivní analýzu 139
interpretace dat 6, 138

K

kanonická diskriminační analýza 67, 71
koeficient determinace 7, 10, 11
komponenta 45, 104, 105
komponentní scóre 46, 54
komponentní zátěže 46, 53
komunalita 58
korelační koeficient 36
kroková diskriminační analýza 67, 72
kvalitativní výzkum 115, 117, 118, 140, 142, 143, 144
kvantitativní výzkum 115, 116, 117

M

metoda hlavních komponent 56
metody shlukování 26, 28, 29, 34, 36, 37, 39, 95
  centroidní metoda 29
  dvoukroková shluková analýza 30
  fuzzy shlukování 30
  k-medoidů 30
  k-průměrů 30
  mediánová metoda 29
  nejbližšího souseda 28
  nejvzdálenějšího souseda 28
  pevné k-shlukování 29
  průměrné vzdálenosti 29
  Wardova metoda 29
míra příslušnosti 26, 30
multikolinearita 18, 21, 56

N

nehierarchické metody shlukování 25

P

pozorování 3, 6, 7, 11, 25, 46, 84, 92, 93, 94, 117, 118, 123, 124, 125, 136
případová studie 118, 121, 122

R

regresní analýza 8, 14
regresní funkce 7, 9, 11
rozhovor 119, 120, 132, 135, 136, 137

S

shluková analýza 25, 32, 92
skupinová diskuze 119
strukturní koeficienty 68, 75

T

technika dotazování 132
tematická analýza 6, 119, 139

V

výzkumná otázka 4
výzkumný problém 4, 128, 131

Summary

The textbook *Methods of Academic Work II* contains seven chapters. The book's individual chapters are chiefly devoted to multi-dimensional statistical methods. These are popular methods that are representatives of methods of various fields of use. The essential link among all the methods is their applicability in the case that individual observations (objects) are characterised by at least three quantitative variables. The book builds on the methods and approaches that were successively outlined in the preceding volume *Methods of Academic Work* (Löster, Mazouch, Radváková, Sigmund, Vltavská, 2018). The authors build on that via multiple regression analysis that is focused on possible applications and the problems that may arise in comparison with simple regression analysis. With regard to representatives of classification methods, where individual objects are organised into clusters, readers are introduced to cluster and discriminatory analysis. Here factor analysis and analysis of main components are replaced by dimension reduction methods.

For the first time, approaches under which the same results can be obtained with the help of the R system, which is very popular at present, are shown. However, because that system was not employed in the previous book and the literature in the Czech language is as yet limited, a whole chapter is dedicated to essential approaches to working in system R in this book. The final, seventh chapter explores the qualitative line of research, from fundamental principles and focus through specific methods of qualitative research to analysis and interpretation of qualitative research data.

In recent years there has been an increase in the number and complexity of research methods, methods for the collection and processing of data and approaches to research. The wealth of methods doesn't only provide increased opportunity to better select research means for reaching identified targets – the

authors of academic essays are also faced with the necessity of making a choice. A certain research method or strategy is not good or bad in absolute terms. It is only as good as its suitability for resolving a specific problem. The research aim frequently determines the choice of method.

Empirical research always means systematic exploration. It is a process of creating new findings – a systematic and carefully planned endeavour. It should approach the discovery of a sought essence, the uncovering of new or essential aspects of the subject of study. However, it must be based on a precise systematic and methodological procedure and on an objective theoretical basis. Authors need to set the aims and methods of their research in the preparatory stage.