Statistika - vícerozměrné metody

Mgr. Martin Sebera, Ph.D.

Katedra kineziologie
Masarykova univerzita
Fakulta sportovních studií

Brno 2012

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Oblast sportu je jednou z mnoha oblastí, kde je zřejmá poptávka po uplatňování exaktních metod a to v interakci s vědou a výzkumem. Ani vědní obor Kinantropologie není výjimkou. Velmi často je nutné řešit problémy vedoucí k vícerozměrným statistickým metodám. Lidská představivost o datech končí už v 3D prostoru, vyšší dimenze je již složité nikoliv zobrazit, ale spíše pochopit a interpretovat. Vícerozměrné metody pak mohou pomoci zejména při redukci dimenze dat na „rozumné“ množství, resp. najít vztahy, které situaci zjednoduší a umožní lepší pochopení. Ne vždy je to však možné a účelné.

Předložená studijní text začíná vysvětlením pojmů i tak je požadována alespoň základní znalost statistiky. Vybraná témata jsou zpracovaná s cílem popsat základní teoretické aspekty jednotlivých metod společně s praktickými příklady, které poskytnou návod a adekvátní postup řešení ve statistickém sw Statistica 10 firmy Statsoft. V textu je obsažena teorie, zájemce o přesnější informace odkážeme na literaturu, kde jsou rozebrány jednotlivé metody s maximální podrobností. Teoretická část je kompilací českých statistiků věnujících se vícerozměrné problematice ve statistice (Hebák, Cyhelský, Meloun, Militký, Hindls, Hendl) a zahraniční autorů (Hair, Johnson, Tabachnick). Ze 4 probíraných oblasti (analýza rozptylu; regresní analýza; analýza hlavních komponent a faktorová analýza; shluková analýza) věnujeme největší pozornost regresní analýze a to pro její dostupnost skrz mnoho statistického software, zároveň obtížnost, mnohoznačnost a nutnost splňovat jednotlivé předpoklady.

Elektronická forma studijní pomůcky umožňuje a počítá s rozšiřováním o další oblasti statistiky, vylepšení teoretických partií a také o další řešené příklady.

Projekt vznikl za podpory Fondu rozvoje vysokých škol FRVŠ/0478/2010.

V Brně 23. 1. 2012

Pojmy

Náhodné veličiny

Za *náhodné veličiny* označujeme proměnné, u kterých nejsme schopni určit hodnotu. Opačně, proměnné, u kterých hodnotu známe nebo je daná, označujeme za *nenáhodné*.

Typy proměnných

Při statistické analýze potřebujeme u každé proměnné určit její typ. Můžeme se setkat s několika způsoby klasifikace proměnných, v našem textu popisujeme přístup, který za hlavní kritérium považuje *typy vztahů mezi hodnotami*. Podle Řezánkové, Marka & Vrabce (2001) u tohoto hlediska rozlišujeme proměnné:

- **Nominální.** Hodnotou je číslo nebo text. U těchto proměnných můžeme provádět jen rozdělení četností, případně operaci porovnání. Příklad: student absolvoval motorický test „běh na 50 m“ s výkonem 7,4 s a motorický test „leh-sed s výsledkem 50 opakování za minutu. Číselné hodnoty 7,4 a 50 určují jen odlišné výsledků motorických testů, nic jiného se vyčíst nedá
- **Ordinální znaky** umožňuje provádět srovnání a tím určit pořadí. V případě textových proměnných je nutné tyto převést na čísla. Příklad: v dotaznících vyjadřujeme míru souhlasu s daným tvrzením. Svou kondicí hodnotím jako: *vynikající – velmi dobrou – dobrou – slabou – špatnou*. Výroky respondentů můžeme určit pořadí, jak který respondent souhlasí s tvrzením. Však netvrdíme, že rozdíl mezi odpověďmi *vynikající* a *velmi dobrou* je stejný jako mezi *slabou* a *špatnou*.
- **Intervalové** kromě porovnání můžeme provádět operaci součtu a rozdílu. Příklad: výška a hmotnost jedince. Naměříme-li u batolete výšku v cm po čtyřech měsících hodnoty 60, 62, 64, 66, znamená to, že každým měsícem dítě vyrostlo o 2 cm.
- **Poměrové znaky** umožňují interpretovat kromě operace rovnosti, uspořádání a rozdílu ještě operace podílu a součinu. Příklad: zaběhne-li atlet 100 m za 11 s a druhý atlet za 22 s, je možné prohlásit, že první je dvakrát rychlejší než druhý.

Nominální a ordinální proměnné jsou souhrnně označovány jako kvalitativní; intervalové a poměrové proměnné jsou souhrnně označovány jako kvantitativní (numerické, kardinální). Kvantitativní proměnné můžeme podle jiného hlediska dělit na

- *diskrétní*, které nabývají pouze celočíselných obměn (*počet permanentek do posilovny*), a
- *spojité* (*metrické*), jež mohou nabývat libovolných hodnot z určitého intervalu (věk respondenta, výkon ve vrhu koulí).

Nominální, ordinální a kvantitativní diskrétní proměnné můžeme souhrnně označit jako **kategoriální** (obměny těchto proměnných nazýváme kategoriemi).

- *dichotomické* (*alternativní*), které nabývají pouze dvou kategorií (ekonomicky aktivní a neaktivní, kuřák a nekuřák), a
- *vícekategoriální* (*množné*), jež nabývají více než dvou kategorií (rodinný stav, obor).

Důležitá jsou primární data, každou transformací původních dat do skupin, kategorií, intervalů ztrácíme informace v nich obsažené. Pro statistickou analýzu jsou původní data nejvhodnější.

Členění datové matice ze provést zejména horizontálně. Rozčlenění souboru do skupin je někdy dáno a cílem je porovnání skupin (analýza rozptylu), jindy je hledání rozčleněné samotným cílem analýzy (shluková analýza). Data budeme předkládat ve formě *datové matice* typu $n \times p$, kde

řádky reprezentují případy, objekty, testované osoby. Sloupce představují proměnné, tedy jednotlivé zkoumané vlastnosti.

Odhady a testy hypotéz

Statistická hypotéza je předpoklad o hodnotě neznámého parametru nebo o zákonu rozdělení sledované veličiny. Statistické hypotézy jsou tedy domněnky o populaci, jejichž pravdivost lze ověřovat prostřednictvím statistických testů.

Hypotézu, jejíž platnost ověřujeme, nazýváme *testovanou (nulovou) hypotézou* a značíme ji *H (H_0)*. Proti testované hypotéze stanovíme *alternativní hypotézu A (H_1),* která hypotézu H popírá. Testování sledované hypotézy H proti alternativní hypotéze A je postup, podle něhož na základě náhodného výběru rozhodneme mezi dvěma tvrzeními – sledovanou hypotézou H a alternativní hypotézou A. Testové kritériem je statistika T(X), jejíž rozdělení známe. Testy (výběrové statistiky) jsou náhodné veličiny (funkce náhodného výběru), pomocí kterých na základě výsledků z náhodného výběru rozhodneme, zda má být ověřovaná hypotéza zamítnuta či nikoliv.

Kritický obor W_α, je interval, který je ohraničený tzv. kritickými hodnotami, což jsou kvantily rozdělení příslušného testového kritéria. Kritický obor W_α tvoří doplněk k 100 (1- α) %-nímu intervalu spolehlivosti. Jestliže hodnota testové statistiky $T(X) \in W_\alpha$, potom hypotézu H zamítáme (Seberová & Sebera, 1999).

Výsledkem testování je buď zamítnutí hypotézy H ve prospěch alternativy A či nezamítnutí hypotézy H. **Skutečnost, že hypotézu H nezamítáme, neznamená že naměřená data tuto hypotézu potvrzují, ale pouze to, že ji nevyvracejí.**

Číslo α se nazývá *hladina statistické významnosti testu*. Hladina statistické významnosti α tedy určuje pravděpodobnost, že testovací charakteristika padne mimo obor přijetí. Obvykle nabývá hodnot od 0,001 do 0,15 v závislosti na povaze zkoumaného problému (tedy nemusí to být jen hodnota 0,05, jak je v mnoha učebních textech doporučováno).

Při testování hypotéz se můžeme dopustit chyby dvěma způsoby: Buď zamítneme hypotézu, která platí – to je chyba prvního druhu α - nebo naopak tuto hypotézu nezamítneme, i když je nesprávná – v tomto případě se jedná o chybu druhého druhu β.

Mezi základní nedostatky statistické významnosti patří:

- použití je možné jen v případě reprezentativního vzorku pomocí náhodného výběru.
- závislost a na počtu pozorování (měření, respondentů)
- statisticky významné neznamená důležité

Tab. 1 Možné výsledky testování hypotézy

Skutečnost	**Rozhodnutí**	
	nezamítáme H	**zamítáme H**
Hypotéza H platí	správné rozhodnutí pravděpodobnost = 1-α	chyba I. druhu pravděpodobnost = α
Platí alternativa A	chyba II. druhu pravděpodobnost = β	správné rozhodnutí pravděpodobnost = 1-β (síla testu)

- Jestliže snížíme α, zvýší se β
- Snížení chyby II. druhu bez toho abychom ovlivnili chybu I. druhu je možné pouze zvýšením rozsahu výběru.

Věcná významnost

- „selský rozum“, neboli logické stanovení např. rozdílu, který budeme považovat vzhledem k povaze problému za významný. Úsudek vychází z předchozích zkušeností, ale i z chyb měření
- používání nestatistického hodnocení velikosti rozdílu či vztahu ve výzkumných výsledcích, tzv. „size of effect“, zvláště pomocí tzv. koeficientu η^2 (eta^2) jakožto podílu, resp. procenta vysvětleného rozptylu (např. u ANOVY). $\eta 2 = SS_b / SS_T$, kde SS_b je meziskupinový součet čtverců a SS_T je celkový součet čtverců
- Např. ke kvantifikování velikosti účinku, tj. k hodnocení věcné významnosti je možné použít *Cohenův koeficient účinku d.* Jednou z hlavních výhod koeficientu je jeho nezávislost na rozsahu výběru. Platí pro něj konvenční hodnoty, jež usnadňují rozhodnutí, kdy lze hovořit o velkém efektu. Pokud je *d* větší než 0,8, je efekt velký; pro *d* z intervalu 0,5 – 0,8 je efekt střední; efekt pod hodnotou 0,2 lze považovat za malý.

Tab. 2 Možné výsledky při srovnání statistické a věcné testování hypotézy

	věcná	
statistická	ano	ne
ano	jednoznačné potvrzení	spíše nepřijmout, výsledek může být ovlivněn velkým výběrem souboru dat
ne	spíše nepřijmout, výsledek je neprůkazný, může být náhodným jevem	jednoznačné potvrzení

Postup při práci s hypotézami by měl vypadat následovně: 1. nejprve zhodnotit věcnou významnost jak absolutně (v jednotkách měření), tak i relativně k podílu vlivu ostatních faktorů (např. pomocí ω^2), a jen jde-li o randomizovaný výzkum pak 2. použít statistickou významnost α jakožto riziko zobecnění.

Testování statistické významnosti pak probíhá tak, že vypočítáme hodnotu testové statistiky, porovnáme ji s kritickými hodnotami (kvantily), odpovídajícími hladině významnosti α, a rozhodneme o zamítnutí či nezamítnutí hypotézy H. Při testování pomocí statistických programů se používá jiný postup: Spočte se hodnota testové statistiky a k ní nejmenší kritický obor, při kterém bychom ještě mohli na základě této hodnoty zamítnout hypotézu H_0 proti dané alternativě. Hladina významnosti, odpovídající tomuto kritickému oboru, se nazývá minimální hladina významnosti (p-hodnota). **Pokud je $p > \alpha$, pak hypotézu H_0 nezamítáme.V opačném případu, kdy $p \leq \alpha$, pak hypotézu H_0 zamítáme.**

Problémy ověřování normality

Předpoklad normality je často vyžadován pro použití většiny statistických metod. U vícerozměrných statistik se jedná o vícerozměrné normální rozdělení sledovaných proměnných, jehož lze (někdy) dosáhnout v případě nesplnění transformací dat, resp. je možnost použít neparametrické metody. K ověření normality lze použít grafické posouzení nebo testy: chí-kvadrát dobré shody, Kolmogorov-Smirnovov a Shapiro-Wilksův test. Tyto testy jsou neparametrické.

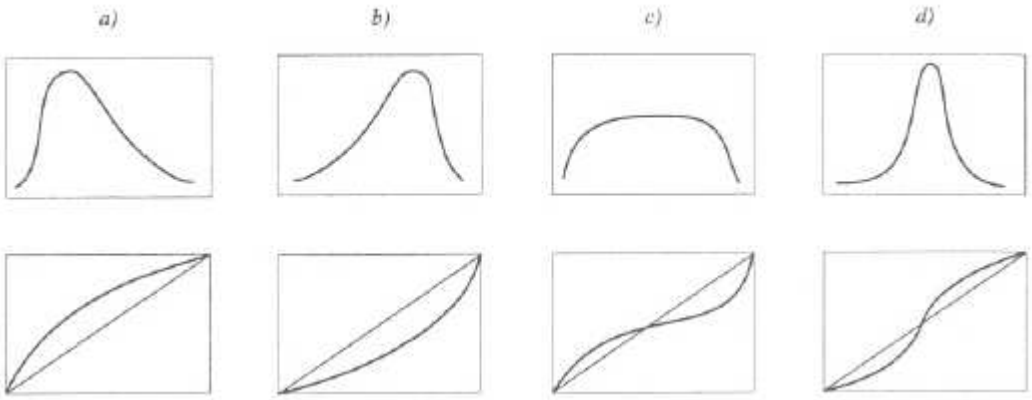
- Chí-kvadrát test dobré shody je založen na srovnání očekávaných a skutečných četností ve třídách.
- U Kolmogorov-Smirnovova testu je testovým kritériem maximální rozdíl mezi předpokládanou *(teoretickou)* plně specifikovanou distribuční funkcí a výběrovou *(empirickou)* distribuční funkcí, jejichž, hodnoty určujeme jako kumulativní relativní četnosti ve výběru.
- Shapiro-Wilkův test porovnává naměřené hodnoty s kvantily normovaného normálního rozdělení pro pravděpodobnosti výběrové distribuční funkce. Ve srovnání v testem K-S má větší

sílu neboli menší pravděpodobnost chyby II. druhu.

- Grafické posouzení jednorozměrné normality. Lze použít u max. závislosti 2 proměnných, při větším počtu proměnných jsou grafy již hůře zobrazitelné a hůře interpretovatelné
    1) Histogram rozdělení četností, který by se v ideálním případě blížil Gaussově křivce.
    2) *Q-Q diagram,* kde se na ose vynášejí kvantily sledované funkce s kvantily normálního rozdělení

Výhodou grafického posouzení je přesnější určení důvodů porušení normality (několik odlehlých hodnot, resp. rozdělení je opravdu zcela odlišné od normálního).

Q-Q diagramy pro normální rozdělení umožňují posoudit více než jen optické posouzení normality a existenci odlehlých pozorování. Průběh bodů indikuje i odchylky od předpokládané šikmosti a špičatosti: Průběh:
a) konkávní ukazuje *kladnou šikmost* s větší variabilitou vyšších hodnot,
b) konvexní ukazuje *zápornou šikmost* s větší variabilitou nižších hodnot,
c) konkávně konvexní naznačuje rozdělení s *dlouhými konci,* menší špičatost.
d) konvexně konkávní naznačuje rozdělení s *krátkými konci,* větší špičatost.



Obr. 1 Vztah histogramu a Q-Q grafu pro různá narušení normality
a) kladné sešikmení, b) záporné sešikmení. c) nižší špičatost, d) vyšší špičatost
(Hebák, Hustopecký, Jarošová & Pecáková, Vícerozměrné statistické metody 1, p. 104)

Transformace

Jak bylo uvedeno výše, jednou z možností, jak si pomocí, pokud data nesplňují podmínku normality, je provést transformaci na rozdělení *normální* nebo jemu blízké. Je zřejmé, že půjde o nelineární transformaci, neboť lineární transformace by zachovala původní tvar rozdělení. Použitelné algoritmy jsou:
a) *odmocninová transformace* $t = \sqrt{x}$, mají-li data charakter četností
b) *logitová transformace* $t = \frac{1}{2}\ln\left(\frac{x}{1-x}\right)$, jde-li o podíly (relativní četnosti)
c) *logaritmická transformace* $t = \ln x$, mají-li data charakter logaritmicko-normálnímu rozdělení

V mnoha případech výše uvedené transformace nepomohou a musí se vyzkoušet náročnější způsoby. Např. Boxův-Coxův systém transformací nebo plošnou (nelineární) transformací.

Vícerozměrné normální rozdělení

Mnoho statistických metod vyžaduje splnění podmínka *normality*, přesněji sledované proměnné musí splňovat podmínku normality. Ze zkušeností s reálnými daty vyplývá, že podmínka normality nebývá vždy splněna, resp. mnohdy není vůbec lehké najít data, která by podmínku

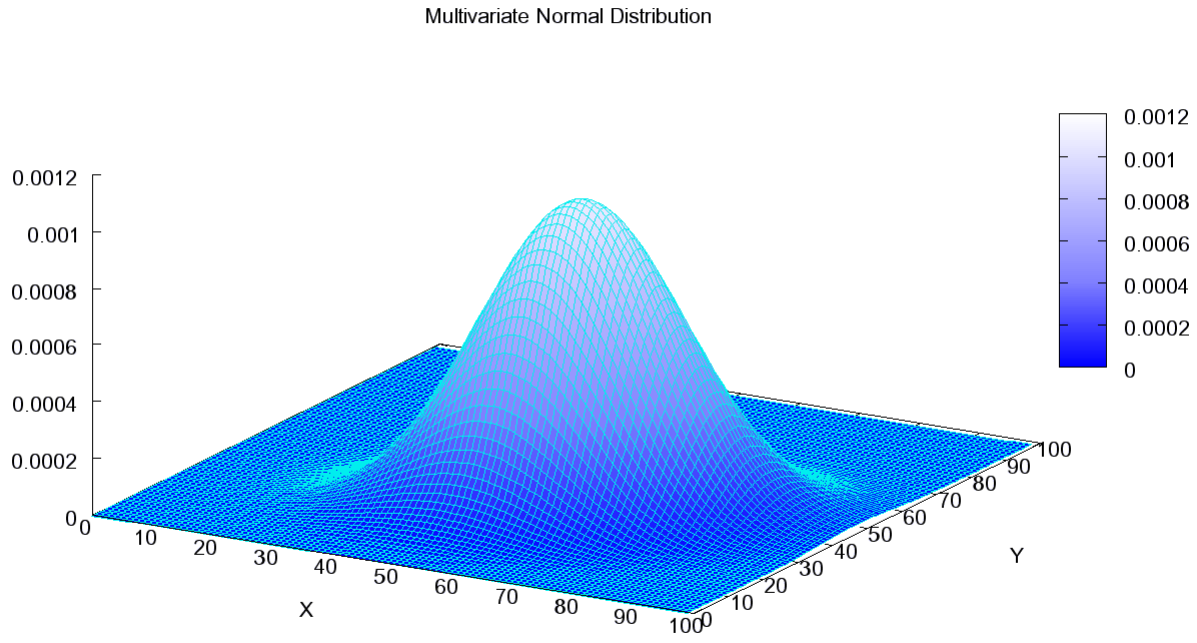
normality splňovala.

Pro naše potřeby nadefinujme *normalitu* jako *simultánního normálního rozdělení* dvou a více náhodných veličin. Mnohé statistické metody vycházejí z předpokladu, že dala byla vybrána z *vícerozměrného normálního rozdělení.* Vícerozměrné normální rozdělení je rozšířením jednorozměrného normálního rozdělení pro případ p ≥ 2 náhodných veličin. Náhodný vektor x má vícerozměrné normální rozdělení, má-li jeho hustota pravděpodobnosti tvar

$$f(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} e^{(x-\mu)^T \Sigma^{-1} (x-\mu)/2},$$

kde μ je vektor p středních hodnot veličin $X_1, X_2, \ldots, X_p$,
E je kovariančí matice C(x) a $-\infty < x_j < \infty$, j = 1, 2, …, p.

Dvourozměrné normální rozdělení je případem p-rozměrného normálního rozdělení pro p = 2. Jeho charakteristické tvar je znázorněn na obr. 2.



Obr. 2. Charakteristický tvar dvourozměrného normálního rozdělení

Srovnání rozptylů K normálních rozdělení

Pro K ≥ 2 výběrů jedné veličiny X s normálním rozdělením uvažujme střední hodnoty označené jako $\mu_1, \mu_2, \ldots, \mu_k$ a rozptyly $\sigma_1^2, \sigma_2^2, \ldots \sigma_k^2$

testujeme hypotézu H: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

vyjadřující že aspoň v jednom případě rovnost neplatí, se označuje za *test homoskedasticity.*

Zamítnutí hypotézy na hladině významnosti znamená nestejné rozptyly veličiny X. Tento jev, *heteroskedasticita*, má pro mnohé statistické postupy závažné důsledky. Některé statistické procedury, jsou založeny na předpokladu homoskedasticity (např. regresní analýza či analýza rozptylu), jsou citlivé na nestejnou variabilitu ve skupinách pozorování. Jedním z univerzálních testů je *Bartlettův test.*

Parametrické – neparametrické (testy, data)

Parametrické:

- testy normality nezamítnou hypotézu o normálním rozložení dat
- mají vyšší sílu testu (schopnost rozpoznal platnost alternativní hypotézy) než testy neparametrické
- při zamítnutí hypotézy o normalitě dat je možné provést buď transformaci dat a přiblížit se tak normalitě nebo přejít na neparametrické testy

Neparametrické testy

- Lze použít při malém rozsahu dat, nezávisle na rozdělení nebo pokud tvar rozdělení nelze upravit transformacemi
- Síla testu klesá z důvodu ztráty původní informace o datech, která jsou nahrazena jejich pořadím, proto pořadové statistiky.

Analýza rozptylu

Pomocí analýzy rozptylu lze využít při zkoumání vztahu mezi nezávislými a závislými proměnnými, zejména při vyhodnocování experimentálních dal. Zkoumáme-li vliv jediného faktoru na jednu či více závislých proměnných, jde o jednofaktorovou analýzu rozptylu. Při více faktorech mluvíme o vícefaktorové analýze rozptylu. Jednorozměrná analýza rozptylu (ANOVA) předpokládá jedinou vysvětlovanou proměnnou, při vícerozměrné analýze rozptylu (MANOVA) můžeme mít i více vysvětlovaných proměnných současně.

Pro zjištění, zda pozorovaná variabilita proměnné Y závisí na příslušnosti hodnot ve skupinách rozkládáme celkovou variabilitu na složky odpovídající různým zdrojům variability (odtud název analýza rozptylu). Variabilitu vyjadřujeme v jednorozměrném případě pomocí součtů čtverců, ve vícerozměrném případě pomocí matic, u nichž součty čtverců tvoří hlavní diagonálu. Model analýzy rozptylu je speciálním případem obecného lineárního modelu (GLM) a hypotézy o vlivu faktorů, jsou speciálním případem obecné lineární hypotézy o parametrech modelu (Hebák, Hustopecký, Jarošová & Pecáková, Vícerozměrné statistické metody 1, p. 160).

Elementární popis závislosti

Základní představu o závislosti mezi dvěma jevy charakterizovanými znaky X a Y získáme uspořádáním empirických údajů, tj. dvojic $[x_i, y_i]$, do dvourozměrné tabulky. Údaje můžeme uspořádat podle variant znaku X, tak podle variant znaku Y a dostaneme **klasickou korelační tabulku** - viz tab. 3 - kde n_{ij} jsou sdružené četnosti, $n_{i.}$ a $n_{.j}$ jsou okrajové četnosti.

Tab. 3 Schéma klasické korelační tabulky

x_i	y_j						$n_{i.}$
	y_1	y_2	…	y_j	…	y_s	
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	$n_{2.}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	$n_{i.}$
…	…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{ks}	$n_{k.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.s}$	n

Podobně jako u jednorozměrného rozdělení četností počítáme z dvourozměrné tabulky následující průměry a rozptyly:

podmíněný průměr
$$\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = \frac{\sum_{j=1}^{s} y_j n_{ij}}{n_{i.}} \qquad (1.1)$$

podmíněný rozptyl
$$s^2_{y,i} = \frac{\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2}{n_i - 1} = \frac{\sum_{j=1}^{s} \left(y_j - \bar{y}_i\right)^2 n_{ij}}{n_{i.} - 1} \qquad (1.2)$$

celkový průměr
$$\bar{y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}{n} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{s} y_j n_{ij}}{n} = \frac{\sum_{i=1}^{k} \bar{y}_i n_i}{n} \qquad (1.3)$$

celkový rozptyl $$s_y^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2}{n-1} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{s}\left(y_j-\bar{y}\right)^2 n_{ij}}{n-1} \qquad (1.4)$$

rozptyl podmíněných průměrů $$s_{\bar{y}_i}^2 = \frac{\sum_{i=1}^{k}\left(\bar{y}_i-\bar{y}\right)^2 n_i}{n-1} \qquad (1.5)$$

průměr podmíněných rozptylů $$\overline{s_{y,i}^2} = \frac{\sum_{i=1}^{k} s_{y,i}^2 n_i}{n} \qquad (1.6)$$

Jednofaktorová ANOVA

ANOVA (z anglického Analysis of Variance), se v praxi používá buď jako samostatná technika nebo jako postup umožňující analýzu zdrojů variability u lineárních statistických modelů. Ze statistického hlediska lze analýzu rozptylu chápat jako speciální případ regresní analýzy, kdy vysvětlující (nezávisle) proměnná má pouze binární charakter, čili může nabývat pouze hodnot 0 nebo 1. Podle konkrétního uspořádání experimentu existuje celá řada variant analýzy rozptylu - viz např. Meloun & Militký (2004).

Podkladem pro jednofaktorovou analýzu rozptylu jsou hodnoty y_{ij} (i = 1, …, k a j = 1, …, s) proměnné Y roztříděné do k skupin podle úrovní (variant) x_1, x_2, …, x_k faktoru X. Podstatou analýzy rozptylu je rozklad celkového rozptylu na složku objasněnou (známý zdroj variability) a složku neobjasněnou (reziduální, chybovou), o níž se předpokládá, že je náhodná.

Ze vztahu (1.4) pro celkový rozptyl plyne, že celkovou variabilitu charakterizuje součet

$$S_y = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2 , \qquad (1.7)$$

jemuž přísluší (n – 1) stupňů volnosti, $\bar{y}$ je celkový průměr (1.3).

Ze vztahu (1.2) plyne, že variabilitu uvnitř skupin charakterizuje součet

$$S_y = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2 , \qquad (1.8)$$

jemuž přísluší (n - k) stupňů volnosti, $\bar{y}_i$ je podmíněný průměr (1.1).

Variabilitu (1.5) podmíněných průměrů, čili variabilitu mezi skupinami, charakterizuje součet

$$S_{y,v} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2 , \qquad (1.9)$$

jemuž přísluší (k – 1) stupňů volnosti.

Mezi uvedenými součty platí vztah

$$S_y = S_{y,m} + S_{y,v} . \qquad (1.10)$$

Při malých rozdílech mezi výběrovými podmíněnými rozptyly (1.2) lze předpokládat, že variabilita (1.5) podmíněných průměrů kolem celkového průměru (1.3) je způsobena závislostí Y na X. Základním předpokladem použití analýzy rozptylu je, že každý z k nezávislých výběrů znaku Y

pochází z normálního rozdělení $N(\mu_i, \sigma_i^2)$ se stejným rozptylem σ^2. Předpoklad normality lze ověřit např. testem dobré shody. V praxi se od toho často upouští a posuzuje se pouze, zda se ve skupinách hodnot proměnné Y, zjištěných na jednotlivých úrovních faktoru X, nevyskytují vysloveně extrémní hodnoty a zda se hodnoty blízké podmíněným průměrům vyskytují častěji než hodnoty, jejichž vzdálenost od podmíněných průměrů je větší.

K ověření hypotézy o stejných rozptylech k normálních rozdělení lze použít Bartlettův test. Nevýhodou Bartlettova testu je to, že je velmi citlivý na porušení předpokladu normality. Jsou-li četnosti všech tříd stejné, tj. $n_1 = n_2 = \ldots = n_k$, používá se k testování hypotézy o rovnosti rozptylů také Hartleyův nebo Cochranův test. I od něj se v praxi často upouští a vychází se pouze z intuitivního posuzování rozdílnosti podmíněných rozptylů. Nejsou-li hodnoty σ_i^2 příliš rozdílné a nevykazují-li s rostoucím X vzestupnou ani sestupnou tendenci, považujeme předpoklad o stejných rozptylech normálních rozdělení $N(\mu_i, \sigma_i^2)$, kde i = 1, … k, za přijatelný.

Při testování hypotézy H, že znak (faktor) X neovlivňuje znak Y vlastně testujeme hypotézu, že rozdělení proměnné Y mají na různých úrovních faktoru X stejné střední hodnoty μ_i. Alternativní hypotéza tvrdí, že alespoň jedna ze středních hodnot μ_i se liší od ostatních, čili H: X neovlivňuje Y, A: H neplatí.

K testu hypotézy H se používá testové kritérium

$$F = \frac{S_{y,m}/(k-1)}{S_{y,v}/(n-k)}. \qquad (1.11)$$

Kritický obor je vymezen nerovností

$$W_\alpha: F > F_{1-\alpha}(k-1, n-k), \qquad (1.12)$$

kde $F_{1-\alpha}(k-1, n-k)$ je $100(1-\alpha)\%$ kvantil F-rozdělení o $\nu_1 = k - 1$ a $\nu_2 = n - k$ a stupních volnosti.

Padne-li hodnota testového kritéria do tohoto kritického oboru, přijímáme na hladině významnosti α hypotézu o statisticky významné závislosti proměnné Y na proměnné X.

Místo porovnání vypočtené hodnoty testového kritéria F s hodnotou kvantilu $F_{1-\alpha}(k-1, n-k)$ nabízí statistický software minimální hladina významnosti p, při které lze hypotézu H ještě zamítnout. Je-li $p \leq \alpha$, zamítáme testovanou hypotézu H o nezávislosti proměnné Y na proměnné X.

Tab. 4 Tabulka pro jednofaktorovou analýzu rozptylu

Variabilita	Součty čtverců	Počty stupňů volnosti	Průměrné čtverce	Testové kritérium	Hladina významnosti
Meziskupinová (vysvětlená)	$S_{y,m}$	$\nu_1 = k-1$	$S_{y,m}/\nu_1$	$F = \frac{S_{y,m}/\nu_1}{S_{y,v}/\nu_2}$	p
Vnitroskupinová (reziduální, chybová)	$S_{y,v}$	$\nu_2 = n-k$	$S_{y,v}/\nu_2$	---	---
Celková	S_y	$\nu = n-1$	---	---	---

Jak již bylo výše uvedeno, při jednofaktorové analýze rozptylu se předpokládá, že k nezávislých výběrů hodnot znaku Y pochází z normálních rozdělení se stejnými rozptyly. To znamená, že před vlastním testem by měl být ověřen předpoklad o normalitě a předpoklad o stejných rozptylech.

Předpoklad normality rozdělení a shody rozptylů v různých skupinách lze ověřovat pomocí testů, v praxi se často užívají grafy, které jsou součásti výstupu počítačových procedur. F-test není příliš citlivý na porušení předpokladu normality (určité opatrnosti je třeba jen při existenci odlehlých hodnot), a pokud jsou data vyvážená, tj. v každé skupině je stejný počet hodnot, není

příliš citlivý ani na porušení předpokladu homoskedasticity (Hebák, Hustopecký, Jarošová & Pecáková, Vícerozměrné statistické metody 1, p. 162)

Prokážeme-li existenci vlivu faktoru, následuje hlubší analýza výsledků, při níž zjišťujeme, mezi kterými skupinami existují rozdíly. Porovnáváme dvojice středních hodnot, tj. testujeme hypotézy H: $\mu_i - \mu_j = 0$ pro různá i, j.

Bylo odvozeno mnoho metod, které umožňují kontrolu chyby I. druhu a které se označují jako metody mnohonásobného porovnávání. Uvedeme zde metody nejčastěji zastoupené ve statistických paketech. Může se také stát, že výsledky mnohonásobného porovnávání jsou v konfliktu s výsledky F-testu analýzy rozptylu. Např. všechny intervaly při párovém porovnávání mohou obsahovat nulu, ačkoliv F-test složené hypotézy H: $\mu_1 = \mu_2 = ... = \mu_k$ zamítnul testovanou hypotézu.

LSD (Fisher)

Použijeme-li metodu nejmenšího významného rozdílu (LSD) při porovnávání různých dvojic hodnot současně, není již riziko chyby I. druhu α dodrženo. Nejedná se tedy vlastně o metodu mnohonásobného porovnávání. Protože jsou intervaly spolehlivosti úzké, stává se, že porovnání vyjde významné i v případě, kdy F-test analýzy rozptylu nezamítnul hypotézu H: $\mu_1 = \mu_2 = ... = \mu_k$. Proto Fisher doporučuje konstruovat interval jen v případě, kdy hypotéza H byla F-testem zamítnuta.

Bonferroni

Bonferroniho metoda patří ke konzervativním testům, zvláště při větším počtu porovnávání, to znamená, že intervaly jsou široké a celková chyba I. druhu je menší než α.

Scheffé

Test je odvozen pro porovnání všech možných kontrastů a proto je rovněž konzervativní.

Jednorozměrné úlohy s více faktory

Při analýze experimentálních výsledků se často výsledky třídí podle více než jednoho faktoru, buď přímo zkoumáme vliv několika faktorů na závislou kvantitativní proměnnou, nebo můžeme mít zkoumaný faktor jen jeden, ale vzhledem ke způsobu realizace experimentu vstupuje do modelu jeden nebo více *blokových* faktorů. Zde se omezíme jen na případ dvou faktorů. Pro zkoumání vlivu jednoho faktoru použijeme model bez interakce. Vyhodnocení úplného faktoriálního experimentu provedeme pomocí modelu s interakcí.

Model pro dva faktory s interakcí má tvar

$$y_{kgi} = \mu + \alpha_k + \beta_g + (\alpha\beta)_{kg} + \varepsilon_{kgi},$$
$$k= 1, 2, \ldots, K, g = 1, 2 \ldots, G, i = 1, 2, \ldots, r,$$

v něm μ vyjadřuje obecnou konstantu, α_k efekt k-té úrovně jednoho faktoru, β_g efekt g-té úrovně druhého faktoru, $(\alpha\beta)_{kg}$ efekt interakce, tj. efekt kombinace daných úrovní obou faktorů a ε_{kgi} náhodnou složku splňují cí obvyklé předpoklady.

Testujeme jednak hypotézy o tzv. hlavních efektech faktorů, tj. hypotézy o tom, že efekty všech úrovní daného faktoru (bez ohledu na úroveň druhého faktoru) jsou nulové

$$H: \alpha_l = \alpha_2 = \ldots = \alpha_k = 0, \text{ resp. } H: \beta_l = \beta_2 = \ldots = \beta_k = 0$$

jednak hypotézu o efektu interakce

$$H: (\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{ij} = 0$$

to znamená hypotézu o tom, že velikost efektu změny úrovně jednoho faktoru nezávisí na konkrétní

úrovni druhého faktoru

Tab. 5 Dvoufaktorová analýza rozptylu, model s interakcí

Zdroj variability	Součet čtverců	Stupně volnosti	Průměrný čtverec
Faktor A	Q_{B-A}	$v_A = K-1$	$Q_{B,A} / v_A$
Faktor B	Q_{B-B}	$v_B = G-1$	$Q_{B,B} / v_B$
Interakce	Q_{B-AB}	$v_{AB} = (K-1)*(G-1)$	$Q_{B,AB} / v_{AB}$
Reziduální	Q_E	$v_E = KG(r-1)$	Q_E / v_E
Celkový	Q_T	n-1	

Vícerozměrné úlohy s jedním faktorem

Místo jednoho pozorování na experimentální jednotce budeme nyní uvažovat vektor p pozorování a úvahy zobecníme pro p-rozměrný případ. Pro vícerozměrnou analýzu rozptylu použijeme model

$$\mathbf{y}_{ki} = \boldsymbol{\mu}_k + \boldsymbol{\varepsilon}_{ki}$$

Testovanou hypotézu zamítneme na hladině významnosti α, překročí-li hodnota testové statistiky F kvantil $f_{l-\alpha}$ (v_1, v_2). Výpočet hodnot statistik včetně uvedených transformací a příslušných p-hodnot je běžnou součástí počítačových programů pro vícerozměrnou analýzu rozptylu, např. ve statistických paketech SPSS nebo STATISTICA. Podrobný teoretický popis přesahuje rámec tohoto studijního textu, čtenáře odkážeme na (Hebák, Hustopecký, Jarošová & Pecáková, Vícerozměrné statistické metody 1, p. 178).

Obecný postup při analýze rozptylu

V úvodu má výzkumník určit na základě dat a povahy problému o jaký model ANOVY se bude jednat: s pevnými, náhodnými nebo smíšenými efekty. Jsou definovány hypotézy a vypočítány parametry ANOVY. Následuje interpretace:

1. Odhadu parametrů základního modelu ANOVA.
2. Ověřování významnosti a konstrukce různých submodelů u modelů s pevnými efekty.
3. Vyjádření složek rozptylů u modelů s náhodnými efekty a testování jejich významnosti.
4. Ověření předpokladů normality, homogenity rozptylů a přítomnosti silně vybočujících pozorování.
5. Interpretace výsledků s ohledem na zadání dat a jejich případné úpravy.

(Meloun & Militký, 2004, p. 560)

Příklad 1 Jednorozměrná ANOVA

Zadání: Pro porovnání tří hodnotitelů A_1, A_2, A_3 byl proveden tento experiment: Každé respondent byl změřen 3 hodnotiteli. V tabulce 6 jsou uvedeny naměřené hodnoty motorického testu v běhu na 1 km. Hodnoty jsou uvedené v sekundách. Zjistěte, zda existují významné rozdíly mezi výsledky jednotlivých hodnotitelů.

Data: n = 10.

Tab. 6 Vstupní data

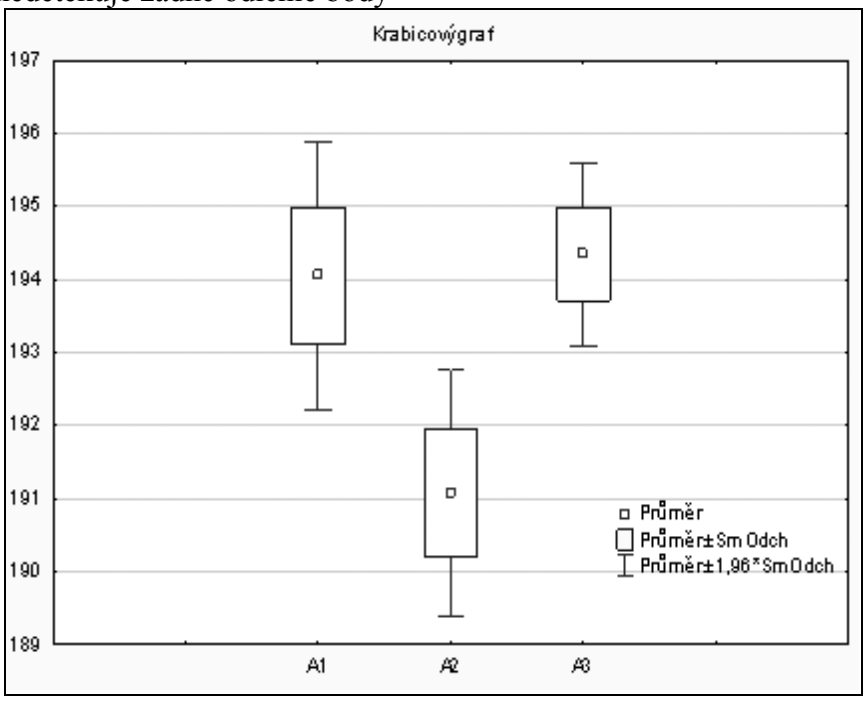
A_1	A_2	A_3
194,6	190,2	194,5
193,5	191,3	195,2
194,6	192,4	194,5
194,6	191,3	195,2
192,4	192,4	193,6
194,6	190,2	194,7
194,6	190,2	193,6
192,4	191,3	194,3
194,6	190,2	194,5
194,6	191,3	193,4

Řešení: Z údajů v tabulce 6 byly určeny následující sloupcové charakteristiky (tab. 7):

Tab. 7 Sloupcové základní charakteristiky

	Popisné statistiky				
Proměnná	**N platných**	Průměr	Minimum	Maximum	Sm.odch.
A1	10	194,0500	192,4000	194,6000	0,934820
A2	10	191,0800	190,2000	192,4000	0,867692
A3	10	194,3500	193,4000	195,2000	0,638140

Krabicový graf nedetekuje žádné odlehlé body



Obr. 3 Krabicový graf

Splnění předpokladů:

- Nezávislost výběrů – je dána podstatou experimentu
- Normalita – ANOVA není citlivá na porušení předpokladu normality, pokud se jedná o vyvážená data (stejný počet hodnot ve skupinách). Pozn. v případě porušení normality můžeme použít Kruskall-Walissův test
- Shoda rozptylů – nezamítáme hypotézu o rovnosti rozptylů – tab. 8

Statistiky → ANOVA → Jednofaktorová ANOVA → více výsledků → Předpoklady

Tab. 8 Testování shody rozptylů

	Testy homogenity rozptylu Efekt: "hodnotitel"				
	Hartleyů F-max	Cochranů C	Bartlett Chí-kv.	ův SV	p
čas 1 km	2,145975	0,429641	1,296865	2	0,522865

Jednotlivé součty čtverců a složky rozptylu jsou uvedeny v tabulkách 9.

Statistiky → ANOVA → Jednofaktorová ANOVA→ Velikost efektů

Tab. 9 Výsledky analýzy rozptylu

Zdroj rozptylu	**Součet čtverců S**	**Stupně volnosti ν**	**Průměrný čtverec S / ν**	**Testovací kritérium F_e**
Zkušebny	$S_A = 65{,}345$	2	32,673	48,190
Reziduální	$S_R = 18{,}306$	27	0,678	-
Celkový	$S_C = 83{,}652$	29	2,885	-

	Jednorozměrné testy významnosti, velik. efektů a síly pro čas 1 km Sigma-omezená parametrizace Dekompozice efektivní hypotézy							
Efekt	SČ	Stupně volnosti	PČ	F	p	Parciál. éta-kvadr.	Výstřednost	Pozor. síla (alfa=0,05)
Abs. člen	1119324	1	1119324	1650920	0,000000	0,999984	1650920	1,000000
hodnotitel	65	2	33	48	0,000000	0,781165	96	1,000000
Chyba	18	27	1					

Protože podíl $F_e = 32{,}673 / 0{,}678 = 48{,}190$ vysoko překračuje kvantil $F_{0,95}(2, 27) = 5{,}448$, zamítáme hypotézu o rovnosti efektů úrovní A_1, A_2, A_3. Scheffého procedura vícenásobného porovnání (tab. 10) ukázala, že rozdíly mezi průměry $\hat{\mu}_1$ a $\hat{\mu}_2$ jsou významné. Rovněž rozdíly mezi průměry $\hat{\mu}_2$ a $\hat{\mu}_3$ nemůžeme považovat za statisticky nevýznamné.

Tab. 10 Výsledek Scheffeho metody mnohonásobného pozorování

Č. buňky	Scheffeho test; Pravděpodobnosti pro post-hoc testy Chyba: meziskup. PČ = ,67800, sv = 27,000			
	zkušebna	{1} 194,05	{2} 191,08	{3} 194,35
1	A1		0,000000	0,720475
2	A2	0,000000		0,000000
3	A3	0,720475	0,000000	

Závěr: Jednofaktorová analýza rozptylu s pevnými efekty ukázala, že rozdíly mezi výsledky jednotlivých hodnotitelů jsou statisticky významné. Zatímco rozdíly mezi výsledky hodnotitelů A_1 a A_3 jsou náhodné, hodnotitel A_2 měří systematicky odlišné (nižší) hodnoty než hodnotitelé A_1 a A_3.

Příklad 2. Dvojrozměrná ANOVA bez opakování

Zadání: Bylo sledováno, zda čas potřebný k vyřešení určité úlohy závisí na době a na hlučnosti okolí. Dvanáct vybraných studentů majících stejné studijní výsledky bylo rozděleno do tří skupin. První skupina řešila úlohu ráno, druhá v poledne a třetí večer. V každé skupině vždy jeden student pracoval v tichém prostředí, druhý poslouchal reprodukovanou hudbu, třetí rozhlasovou hru a čtvrtý silný pouliční hluk. Počet minut potřebných k vyřešení úlohy je uveden v tabulce 11. Zjistěte, zda doba potřebná k vyřešení úlohy závisí na denní době a na hlučnosti okolí.

Tento příklad byl zařazen z důvodu, že na něm statistický software STATISTICA 10 „havaruje“. Tzn. nedokáže ve svých výstupech provést vyhodnocení požadovaného modelu.

Tab. 11 Počet minut potřebných k vyřešení úlohy

	faktor B			
faktor A	**ticho**	**hudba**	**hra**	**hluk**
ráno	6	7	8	6
v poledne	8	5	10	5
večer	7	6	12	7

Řešení:

Nejprve vypočítáme základní statistické charakteristiky a graficky znázorníme průměry jednotlivých efektů (tab. 12 a 13).

Tab. 12 Základní statistické charakteristiky faktoru A

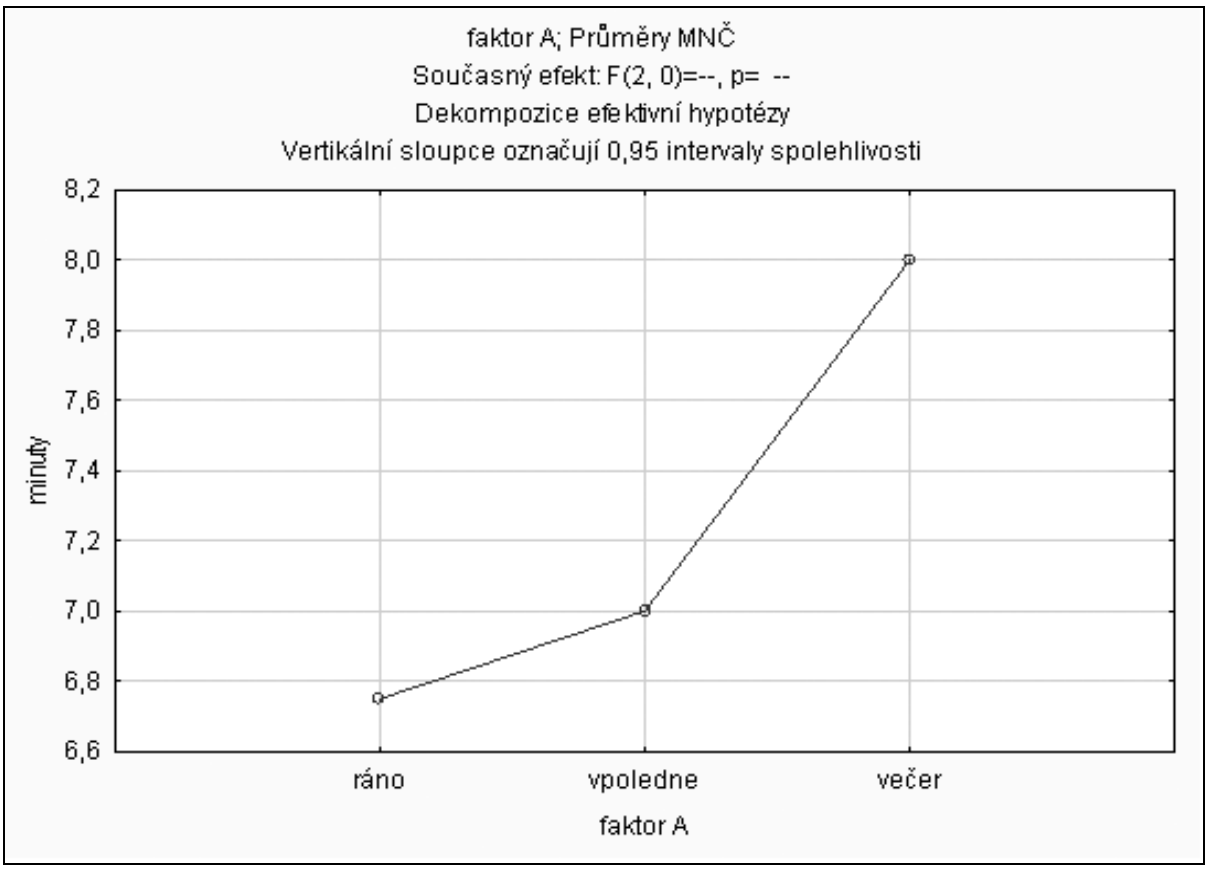
Rozkladová tabulka popisných statistik (anova-2-minuty)
N=12 (Sezn. záv. prom. bez ChD)

faktor A	minuty (ticho) Průměry	minuty (ticho) N	minuty (ticho) Sm.odch.	minuty (hudba) Průměry	minuty (hudba) N	minuty (hudba) Sm.odch.	minuty (hra) Průměry	minuty (hra) N	minuty (hra) Sm.odch.	minuty (hluk) Průměry	minuty (hluk) N	minuty (hluk) Sm.odch.
ráno	6,000000	1	0,00	7,000000	1	0,00	8,00000	1	0,00	6,000000	1	0,00
vpoledne	8,000000	1	0,00	5,000000	1	0,00	10,00000	1	0,00	5,000000	1	0,00
večer	7,000000	1	0,00	6,000000	1	0,00	12,00000	1	0,00	7,000000	1	0,00

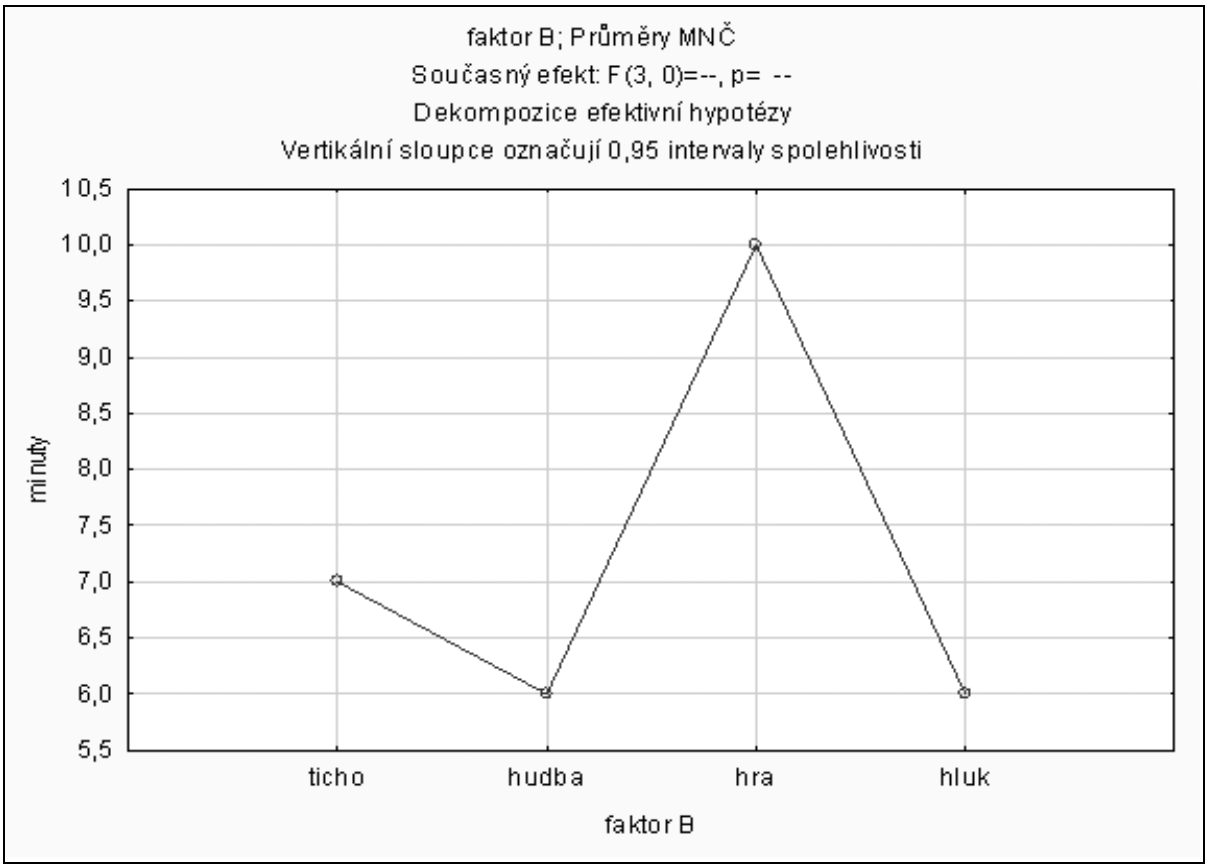
Tab. 13 Základní statistické charakteristiky faktoru B

Rozkladová tabulka popisných statistik (anova-2-minuty)
N=12 (Sezn. záv. prom. bez ChD)

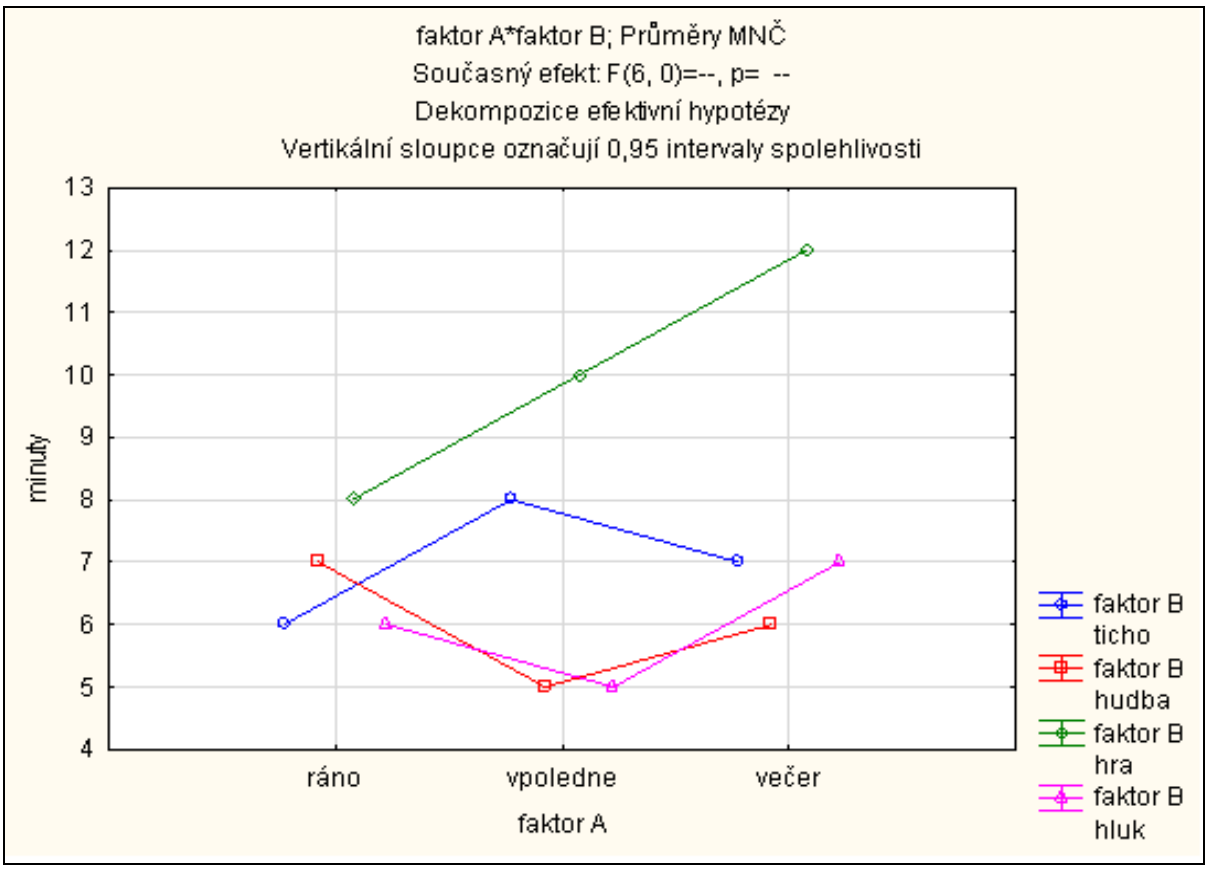
faktor B	minuty (ráno) Průměry	minuty (ráno) N	minuty (ráno) Sm.odch.	minuty (vpoledne) Průměry	minuty (vpoledne) N	minuty (vpoledne) Sm.odch.	minuty (večer) Průměry	minuty (večer) N	minuty (večer) Sm.odch.
ticho	6,000000	1	0,00	8,00000	1	0,00	7,00000	1	0,00
hudba	7,000000	1	0,00	5,00000	1	0,00	6,00000	1	0,00
hra	8,000000	1	0,00	10,00000	1	0,00	12,00000	1	0,00
hluk	6,000000	1	0,00	5,00000	1	0,00	7,00000	1	0,00



Obr. 4 Grafické znázorněn vlivu faktoru A



Obr. 5 Grafické znázorněn vlivu faktoru B



Obr. 6 Grafické znázorněn vlivu interakce faktorů A a B

Tab. 14 Výstup analýzy rozptylu v počtu minut potřebných k vyřešení úlohy

Zdroj rozptylu	**Součet čtverců S**	**Stupně volnosti ν**	**Průměrný čtverec S / ν**	**Testovací kritérium F_e**
Úrovně faktoru A	$S_A = 3{,}50$	2	1,75	0,833
Úrovně faktoru B	$S_B = 32{,}25$	3	10,75	5,119
Interakce Tukey	$S_T = 3{,}67$	1	3,67	1,747
Reziduální	$S_R = 10{,}50$	5	2,10	-
Celkový	$S_C = 46{,}25$	11	4,20	-

Statistiky → ANOVA → Vícefaktorová ANOVA→ Velikost efektů

Efekt	Jednorozměrné testy významnosti, velik. efektů a síly pro minuty Sigma-omezená parametrizace Dekompozice efektivní hypotézy							
	SČ	Stupně volnosti	PČ	F	p	Parciál. éta-kvadr.	Výstřednost	Pozor. síla (alfa=0,05)
Abs. člen	630,7500	1	630,7500					
faktor A	3,5000	2	1,7500					
faktor B	32,2500	3	10,7500					
faktor A*faktor B	10,5000	6	1,7500					
Chyba		0						

Byly testovány hypotézy (tab. 14) o nulovosti efektů faktoru A. Srovnání kvantilu $F_{0,95}(2, 5) = 5,787$ s hodnotou $F = 0,833$ vede k závěru, že efekt faktoru A je nevýznamný. Efekt faktoru B, $F_{0,95}(3, 5) = 5,409 > 5,119$, je sice nevýznamný, ale blízkost hodnot 5,409 a 5,119 signalizuje, že hlučnost z části ovlivňuje dobu potřebnou k vyřešení úlohy. Nevýznamný je rovněž efekt interakce, neboť $F_{0,95}(1, 5) = 6,608 > 1,747$.

Závěr: Dvoufaktorová analýza rozptylu bez opakování pozorování ukázala, že denní doba neovlivňuje čas potřebný k vyřešení úlohy. Na druhé straně se nepodařilo prokázat, že hlučnost okolí ovlivňuje dobu potřebnou k řešení příkladu.

Příklad 3 Dvojrozměrná ANOVA s opakováním

Zadání: Byl zkoumán výsledný čas v motorickém testu v závislosti na typu suplementace sportovce (faktor A) a na způsobu tréninku (faktor B). Každá kombinace byla realizována čtyřikrát nezávisle na sobě. Výsledky jsou uvedeny v tabulce 15. Zjistěte, jak ovlivňuje výsledný čas druh suplementace a způsob tréninku.

Data: n = 24

Tab. 15 Výsledný čas

	Způsob tréninku											
Suplementace	**Bez tréninku**				**Aerobní**				**Anaerobní**			
výrobce 1	2,8	3,2	3,0	3,0	3,7	3,6	3,9	3,6	3,4	3,8	3,7	3,6
výrobce 2	3,1	2,7	3,0	2,9	3,4	3,4	3,0	3,8	4,2	4,0	4,1	3,9

Řešení:

Na základě výsledků z programu Statistica 10 byla sestavena tabulka 16 a ručním výpočtem tabulka 17

.

Statistiky → ANOVA → Vícefaktorová ANOVA→ Velikost efektů

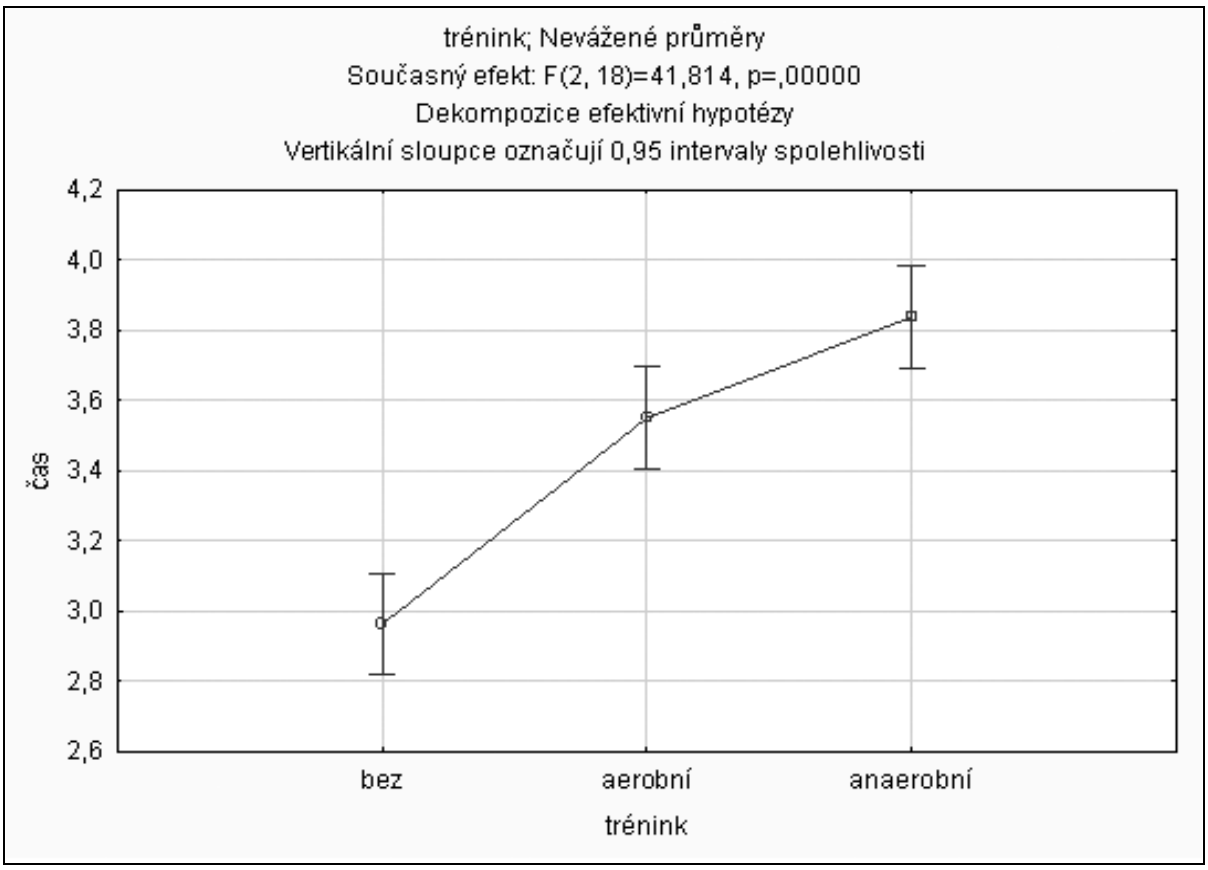
Tab. 16 Analýza rozptylu výsledku motorického testu

	Jednorozměrné testy významnosti, velik. efektů a síly pro čas (anova-3-test) Sigma-omezená parametrizace Dekompozice efektivní hypotézy							
Efekt	SČ	Stupně volnosti	PČ	F	p	Parciál. éta-kvadr.	Výstřednost	Pozor. síla (alfa=0,05)
Abs. člen	285,6600	1	285,6600	7506,394	0,000000	0,997608	7506,394	1,000000
suplem	0,0017	1	0,0017	0,044	0,836585	0,002427	0,044	0,054519
trénink	3,1825	2	1,5913	41,814	0,000000	0,822883	83,628	1,000000
suplem*trénink	0,5508	2	0,2754	7,237	0,004938	0,445718	14,474	0,887984
Chyba	0,6850	18	0,0381					

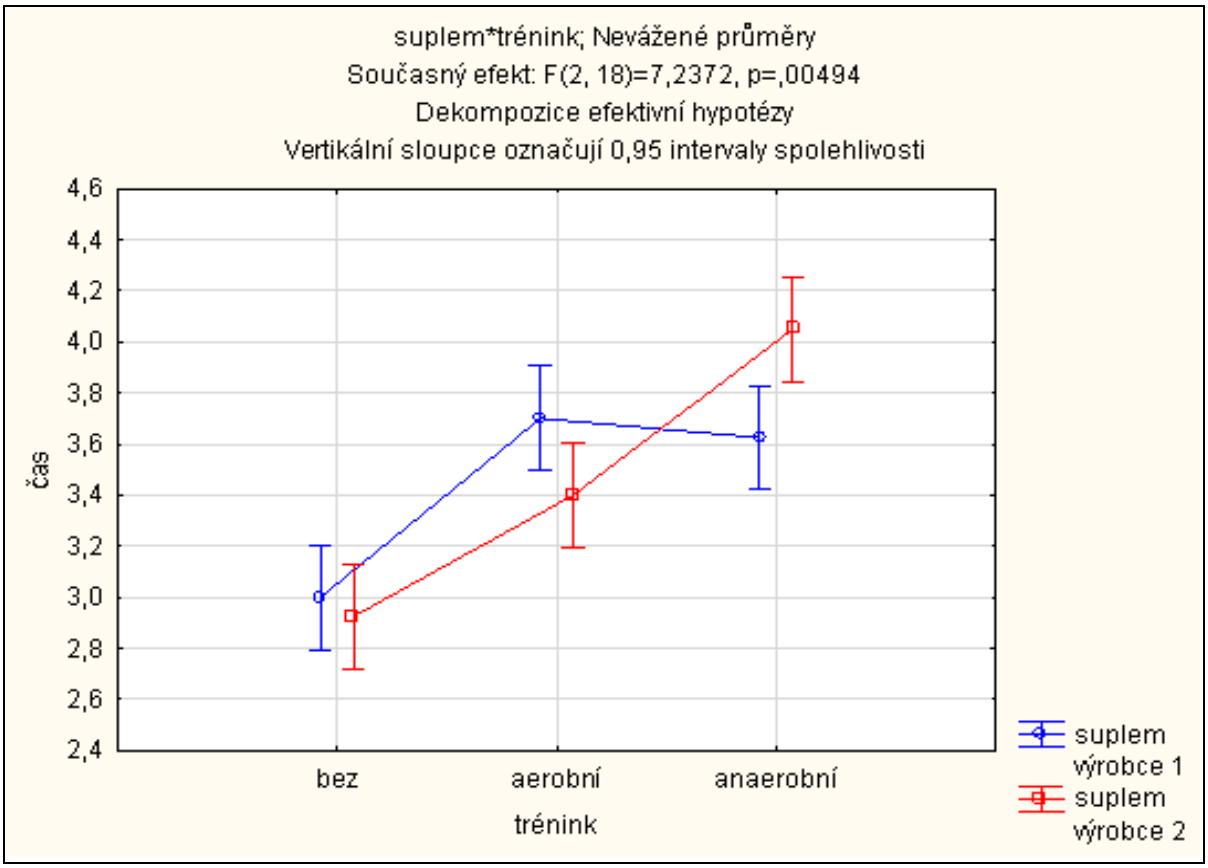
Tab. 17 Analýza rozptylu výsledku motorického testu

Zdroj rozptylu	**Součet čtverců S**	**Stupně volnosti ν**	**Průměrný čtverec S / ν**	**Testovací kritérium F_e**
Úrovně faktoru A	$S_A = 0,0017$	1	0,0017	0,044
Úrovně faktoru B	$S_B = 3,1825$	2	1,5912	41,814
Interakce AB	$S_{AB} = 0,5508$	2	0,2754	7,237
Reziduální	$S_R = 0,6850$	17	0,0381	-
Celkový	$S_C = 4,4200$	23	0,1922	-

Srovnáme-li hodnoty testovacích kritérií z tabulky 16 a 17 s příslušnými kvantily F-rozdělení zjistíme, že efekt faktoru A je nevýznamný ($0,004 < 4,414 = F_{0,95}(1, 17)$). Vliv faktoru B je statisticky významný ($41,814 > 3,555 = F_{0,95}(2, 17)$). Rovněž vliv interakce AB je významný ($7,237 > 3,555 = F_{0,95}(2, 17)$).



Obr. 7 Grafické znázorněné vlivu efektu „trénink“



Obr. 8 Grafické znázorněné vlivu interakce efektů „trénink“ a „suplementace“

Závěr: Nepodařilo se prokázat závislost výsledného času na druhu suplementace. Je však prokázán vliv tréninku (obr. 7). Rovněž byla prokázána přítomnost interakcí. To znamená, že všechny způsoby tréninku neovlivňují oba typy suplementace stejným způsobem (obr. 8).

Lineární regrese

Statistické modelováni závislosti

Získáme-li v našem výzkumném šetření proměnné, mezi nimiž lze zdůvodnit hledání vzájemného lineárního vztahu, můžeme použít metodu lineární regrese. Regresní analýza je statistická metoda pro modelování závislosti jedné nebo několika (nejlépe měřitelných spojitých) vysvětlovaných *náhodných veličin* (závisle proměnných) Y_1, Y_2, …, Y_G na jedné nebo více vysvětlujících veličinách (nezávisle proměnných) X_l, X_2, …, X_K. Základním úkolem regresní analýzy je pomocí matematické funkce vysvětlit proměnné Y pomocí vysvětlujících proměnných X.

Příčinnost nemůže být statistickou analýzou prokázána, dostáváme totiž jen informaci o závislosti mezi proměnnými. K prokázání příčinnosti je potřeba sestavit komplexní výzkumný plán, ve které budeme minimalizovat všechny aspekty vyplývající z předmětné oblasti. V hierarchii plánů výzkumu z hlediska validity závěru vzhledem k průkazu příčinnosti stojí nejvýše randomizované klinické studie a metaana1ytické studie (Hendl, 2004, p. 75). Analýzu nikdy nelze provádět bez obsahového významu proměnných a jen na základě případové studie, i s např. rozsáhlým výběrovým souborem. Statistický popis závislosti dvou proměnných neznamená přítomnost příčinného vztahu (Hebák, Malá & Hustopecký, Vícerozměrné statistické metody 2, p. 11).

Lineární - funkce lineární v parametrech či funkce, které lze na lineární v parametrech převést vhodnou transformací (např. logaritmováním)

Příklady regresních funkcí

a) $Y = \beta_0 + \beta_1 X + \beta_2 Z + ... + \beta_k Q$

b) $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

c) $Y = \beta_0 \beta_1^X \beta_2^Z$, kterou lze přepsat do lineárního tvaru (lineárního v parametrech)

$$ln(Y) = ln(\beta_0) + X\, ln(\beta_1) + Z\, ln(\beta_2)$$

Nelineární - do této skupiny budeme zařazovat funkce nelineární v parametrech (a linearitu nelze dosáhnout ani vhodnou transformací)

Příklady regresních funkcí

a) $Y = \beta_0 + \beta_1 \beta_2 X$

b) $Y = \beta_{0+} \beta_1^X$

Regrese a korelace

Pojem regrese pochází z prací antropologa a meteorologa Francise Galtona, které předložil veřejnosti v letech 1877 až 1885. Galton se zabýval obecnými otázkami dědičnosti a konkrétně se zajímalo vztah mezi výškou otců a jejich prvorozených synů. Pozorováním a analýzou údajů došel k rovnici, ze které vyplývá, že vysocí otcové sice mají i vysoké syny, ale v průměru jsou větší než jejich synové, a podobně i malí otcové mají i malé syny, ale v průměru jsou menší než jejich synové. Tuto tendenci návratu následující generace směrem k průměru nazval Galton regresi (původně tomuto jevu říkal reversion, což později změnil na regression = krok zpět). Současné pojetí regresní analýzy má sice jen málo společného s původním záměrem Galtona, nicméně myšlenka přístupu k empirickým údajům zůstala zachována a pojem regrese se natolik vžil, že se používá dodnes (Hebák, Malá & Hustopecký, Vícerozměrné statistické metody 2, p. 20).

Korelace znamená vzájemný vztah mezi dvěma procesy nebo veličinami. Pokud se mezi dvěma procesy ukáže korelace, je pravděpodobné, že na sobě závisejí, nelze z toho však ještě usoudit,

že by jeden z nich musel být příčinou a druhý následkem. To samotná korelace nedovoluje rozhodnout.

V určitějším slova smyslu se pojem korelace užívá ve statistice, kde znamená vzájemný lineární vztah mezi znaky či veličinami x a y. Tento vztah může být kladný, pokud (přibližně) platí $y = kx$, nebo záporný ($y = -kx$). Míru korelace pak vyjadřuje korelační koeficient, který může nabývat hodnot od −1 až po +1.

Hodnota korelačního koeficientu −1 značí zcela nepřímou závislost, tedy čím více se zvětší hodnoty v první skupině znaků, tím více se zmenší hodnoty v druhé skupině znaků, např. vztah mezi uplynulým a zbývajícím časem. Hodnota korelačního koeficientu +1 značí zcela přímou závislost, např. vztah mezi rychlostí běhu a běžeckou frekvencí kroků sprintera. Pokud je korelační koeficient roven 0, pak mezi znaky není žádná statisticky zjistitelná lineární závislost. Je dobré si uvědomit, že i při nulovém korelačním koeficientu na sobě veličiny mohou záviset, pouze tento vztah nelze vyjádřit lineární funkcí, a to ani přibližně. Může jít např. o nelineární závislost. Z nekorelovanosti náhodných veličin striktně nevyplývá jejich nezávislost, ale naopak z jejich nezávislosti vyplývá i jejich nekorelovanost (Zvonař, Pavlík, Sebera, Vespalec & Štochl, 2010).

Mezi nevýhody korelačního koeficientu patří jeho citlivost k náhodné chybě. Proto se používá ve srovnávacím experimentu. Je též citlivý také k rozmezí měření. Zvětšením rozsahu měření lze zvýšit hodnotu korelačního koeficientu blízko k 1. Závažná je skutečnost, že korelační koeficient neodhaluje ani přítomnost proporcionální chyby ani chyby konstantní (Hendl, 2004, p. 285). Doporučuje se nahradit/doplnit posouzení korelačního koeficientu, který je pouze mírou lineární závislosti výsledků, jinými postupy, např. Bland-Altmanovým rozdílovým grafem.

Jednoduché, dílčí, vícenásobné i podmíněné korelační koeficienty jsou mírami vzájemné lineární závislosti náhodných veličin. Rozdíl mezi nimi je v tom, zda vyjadřují vzájemný lineární vztah dvou náhodných veličin při neuvažování všech ostatních veličin (jednoduché), závislost mezi jednou náhodnou veličinou a lineární funkcí všech nebo některých ostatních veličin (vícenásobné), vzájemný lineární vztah dvou náhodných veličin při statistickém vyloučení všech nebo některých ostatních veličin (dílčí) nebo vzájemný vztah dvou nebo více veličin pro dané hodnoty jiných veličin (podmíněné). (Hebák, Malá & Hustopecký, Vícerozměrné statistické metody 2, p. 24).

Regresní modely a jejich klasifikace

Obtížnost konstrukce regresního modelu souvisí s řadou nejistot zcela zásadního charakteru. Z věcné analýzy i z konkrétních dat můžeme získat mnoho informací, ale nakonec je nutné předpokládat:

- součtový nebo součinový vliv uvažovaných i neuvažovaných činitelů;
- určitý typ regresní funkce;
- pravděpodobnostní chování a rozdělení rušivé složky;
- konkrétní okruh *rozhodujících* vysvětlujících proměnných $X_1, X_2, \ldots, X_K$.

Většinou se předpokládá, že zkoumanou závislost znaku $\boldsymbol{Y}$ na znaku $\boldsymbol{X}$ popisuje **aditivní regresní model**

$$\boldsymbol{Y} = f(\boldsymbol{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \tag{2.1}$$

kde vektor $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)'$ je náhodný vektor pozorovaných hodnot, $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)'$ je nenáhodný vektor vysvětlujících hodnot, funkce $f(\boldsymbol{X},\boldsymbol{\beta})$ je **teoretická regresní funkce,** vektor $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ je **vektor regresních koeficientů (parametrů)** a $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ je **vektor chyb**, čili vektor nezávislých náhodných veličin s rozdělením $N(0, \sigma^2)$.

Regresní model (2.1) vyjadřuje, že empirické údaje y_i se budou více či méně lišit od teoretických hodnot Y_i, čili platí

$$y_i = Y_i + \varepsilon_i = f(x_i, \beta_0, \beta_1, \ldots, \beta_p) + \varepsilon_i, \tag{2.2}$$

Z předpokladů o rozdělení rušivých složek ε_i bezprostředně vyplývá, že pozorované hodnoty y_i náhodné veličiny Y mají normální rozdělení $N(Y_i, \sigma^2)$. Nejsou tedy zatížené systematickými chybami, měření jsou prováděna se stejnou přesností a jsou nekorelované.

Popíšeme nejpoužívanějších typy jednorovnicových regresních modelů se zvláštním zaměřením na modely lineární:

Lineární model

V lineárním modelu se předpokládá součtový vliv všech činitelů a regresní funkci

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \qquad (2.3)$$

ve kterém β_0 je absolutní člen a $\beta_1 \ldots \beta_k$ jsou dílčí regresní koeficienty. Například parametr β_l je interpretován jako očekávaná změna veličiny Y při jednotkovém růstu veličiny X_1 za předpokladu už uvažovaného, a tudíž statisticky konstantního vlivu vysvětlujících proměnných X_2, X_3, …, X_K, a analogicky je hodnocen význam ostatních dílčích regresních koeficientů.

Racionální celistvé a lomené funkce

Velmi často se používá regresní model, který je lineární z hlediska všech parametrů, ale nelineární z hlediska vysvětlujících proměnných. Oblíbené jsou především modely s jednou vysvětlující proměnnou. V této skupině je asi nejznámější *model regresní paraboly s-tého stupně*

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_s X^s + \varepsilon$$

a zvláště *regresní parabola druhého stupně,* kdy s = 2.

Modely převoditelné transformací na lineární model

Pro exponenciální, mocninné, různě kombinované a další regresní funkce je rozumnější předpokládat obecně *součinový (multiplikativní)* typ regresního modelu ve tvaru

$$Y = \eta \varepsilon$$

ve kterém η je regresní funkce a ε rušivá složka. Časté je použití *lineární exponenciální* regresní funkce $\eta = \beta_0 \beta_{\square}^{X}$. Oblíbené jsou rovněž různé formy mocninných regresních funkcí nebo další kombinace uvedených i jiných typů.

Modely nelineární z hlediska parametrů

V opačném případě, kdy regresní funkce má tvar rozdílný od (2.3), mluvíme o nelineární regresní funkci. Podle toho, zda regresní funkce $f(X,\beta)$ je či není lineární funkcí regresních parametrů, rozlišujeme lineární a nelineární regresi. Rozdíl mezi oběma typy spočívá především ve způsobu výpočtu bodových odhadů regresních parametrů. Lineární modely jsou pro svou jednoduchost velmi oblíbené, ale skutečné vztahy mezi veličinami bývají většinou nelineární. V přírodních, technických i společenských vědách se používají nejrůznější typy nelineárních modelů. Například v ekonomické literatuře najdeme téměř 20 věcně zdůvodněných nelineárních produkčních funkcí a podobně je tomu v oblasti spotřeby, poptávky, investic a dalších. Touto problematikou se však zabývat nebudeme.

Vyrovnávací kritéria

Vyrovnáním experimentálních dat se rozumí proložení regresní funkcí takovou, při kterém je celková chyba nejmenší. Celkovou chybou můžeme popsat jako:

- Minimalizace kritéria nejmenšího součtu čtverců
- Minimalizace maximální hodnoty rezidua
- Minimalizace součtu absolutních hodnot reziduí

Nemusí být automaticky nejlepší výsledek, který získáme použitím nejznámější a nejpoužívanější *metody nejmenších čtverců.* Tato metoda vychází z požadavku, aby součet čtverců odchylek pozorovaných hodnot y_i od hodnot $\hat{Y}_i$ ležících na regresní křivce byl minimální, čili hledáme minimum funkce

$$S_R = \sum_{i=1}^{n} \left[y_i - \hat{Y}_i \right]^2 = \sum_{i=1}^{n} \left[y_i - f(x_i, \mathbf{b}) \right]^2 , \tag{2.4}$$

kde

$$\hat{Y}_i = f(x_i, \boldsymbol{b}) \tag{2.5}$$

je **odhad teoretické regresní funkce** (2.1) a rozdíly

$$e_i = \left(y_i - \hat{Y}_i \right), \; i = 1, \ldots, n, \tag{2.6}$$

jsou tzv. **rezidua**.

Rezidua e_i považujeme za odhady chyby ε_i. Součet (2.4) se nazývá **reziduální součet čtverců** a funkce (2.5) se nazývá **empirická (výběrová) regresní funkce.**

Lineární regresní model má tedy tvar

$$Y = \beta_0 f_0 + \beta_1 f_1 + \beta_2 f_2 + \ldots + \beta_p f_p + \boldsymbol{\varepsilon} \tag{2.7}$$

kde $\beta_0, \beta_1, \ldots, \beta_p$ jsou neznámé parametry, regresory f_j, j = 0, 1, …, *p,* jsou známé funkce proměnné X a $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ je vektor nezávislých náhodných veličin s rozdělením $N(0, o^2)$.

Mezi nejužívanější lineární regresní funkce (2.7) patří **přímka** (2.11) a **parabola** (2.12)**,** které jsou vlastně nejjednodušší případy **polynomické regrese** s regresní funkcí

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p . \tag{2.8}$$

Další regresní funkce lineární z hlediska parametrů je logaritmická funkce

$$Y = \beta_0 + \beta_1 \ln X , \tag{2.9}$$

která představuje **logaritmickou regresi.**

Bodové odhady a intervaly spolehlivosti

Bodové odhady v lineárním regresním modelu

Ve statistické literatuře věnované bodovým odhadům mají tradičně některé požadavky přednost před jinými. Na prvním místě se požaduje nezkreslenost (nestrannost, nevychýlenost) odhadu s nejmenším rozptylem. Například při platnosti podmínek klasického lineárního modelu je nejlepším lineárním nezkresleným odhadem odhad b pořízený metodou nejmenších čtverců. Kvalita zvolené statistiky je dána nejen oprávněností učiněných předpokladů a podmínek, ale i volbou hodnotícího kritéria.

Základní metodou odhadu parametrů lineárních regresních funkcí je metoda nejmenších čtverců, tj. požadavek, aby reziduální součet (2.4) byl minimální. Dostaneme **soustavu (p+1) lineárních (normálních) rovnic**

$$\partial \, SR / \partial \, b0 = 0, \partial \, SR / \partial \, b1 = 0, \ldots, \partial \, SR / \partial \, bp = 0. \tag{2.10}$$

Řešením soustavy (2.10) získáme odhady b_0, b_1, …, b_p parametrů β_0, β_1, …, β_p. Při výpočtu odhadů parametrů regresní přímky a regresní paraboly řešíme následující soustavy rovnic:

přímka $\hat{Y} = b_0 + b_1 x$

$$\begin{aligned} nb_0 + b_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned} , \qquad (2.11)$$

parabola $\hat{Y} = b_0 + b_1 x + b_2 x^2$

$$\begin{aligned} nb_0 + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 + b_2 \sum_{i=1}^{n} x_i^3 &= \sum_{i=1}^{n} y_i x_i \\ b_0 \sum_{i=1}^{n} x_i^2 + b_1 \sum_{i=1}^{n} x_i^3 + b_2 \sum_{i=1}^{n} x_i^4 &= \sum_{i=1}^{n} y_i x_i^2 \end{aligned} , \qquad (2.12)$$

logaritmická funkce $\hat{Y} = b_0 + b_1 \ln x$

$$\begin{aligned} nb_0 + b_1 \sum_{i=1}^{n} \ln x_i &= \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} \ln x_i + b_1 \sum_{i=1}^{n} (\ln x_i)^2 &= \sum_{i=1}^{n} y_i \ln x_i \end{aligned} \qquad (2.13)$$

Intervaly spolehlivosti pro regresní parametry

100(1-α)%- ní dvoustranný interval spolehlivosti pro regresní parametr β_j je vymezen nerovnostmi

$$b_j - t_{1-\alpha/2}(\nu) s(b_j) < \beta_j < b_j + t_{1-\alpha/2}(\nu) s(b_j) \text{ , } j = 0, 1, 2, \ldots, p, \qquad (2.14)$$

kde b_j je bodový odhad parametru β_j, $t_{1-\alpha/2}(\nu)$ je kvantil t-rozdělení s $\nu = n - (p+1)$ stupni volnosti a **$s(b_j)$ je směrodatná chyba bodového odhadu** b_j, pro kterou platí

$$s(b_j) = \sqrt{s_R^2 h_{jj}} \text{ , } \qquad (2.15)$$

s_R^2 je **reziduální rozptyl**

$$s_R^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{Y}_i\right)^2}{n - (p+1)} , \qquad (2.16)$$

a h_{jj} je diagonální prvek matice

$$\boldsymbol{H = (X'X)^{-1}}, \qquad (2.17)$$

kde matice $\boldsymbol{X}$ je tzv. **matice regresorů,**

$$\mathbf{X} = \begin{pmatrix} 1 & f_1(x_1) & \ldots & f_p(x_1) \\ 1 & f_1(x_2) & \ldots & f_p(x_2) \\ \ldots & \ldots & \ldots & \ldots \\ 1 & f_1(x_n) & \ldots & f_p(x_n) \end{pmatrix} \qquad (2.18)$$

Většina statistických programů počítá vedle reziduálního rozptylu (2.14) také **směrodatnou odchylku reziduí s_R**, tj.

$$s_R = \sqrt{s_R^2} \tag{2.19}$$

Testy hypotéz o hodnotách regresních parametrů

Individuální t-test o nulové hodnotě regresního parametru testuje hypotézu

H: $\beta_j = 0, j = 1, 2, ..., p,$ proti alternativě A: $\beta_j \neq 0.$ (2.20)

Testovým kritériem je náhodná veličina

$$t = \frac{b_j}{s(b_j)}, \tag{2.21}$$

kde b_j je bodový odhad regresního koeficientu β_j a $s(b_j)$ je směrodatná chyba (2.15) tohoto odhadu.

Kritický obor W_α je vymezen nerovností

$$|t_j| > t_{1-\alpha/2}(n-c), \tag{2.22}$$

kde $t_{1-\alpha/2}(n-c)$ je kvantil t-rozdělení s n – c = n – (p+1) stupni volnosti.

Celkový F-test je test hypotézy

H: $\beta_0 = k, \beta_1 = \beta_2 = ... = \beta_p = 0$ proti A: $\beta_j \neq 0, j = 1, 2, ..., p.$ (2.23)

Testovým kritériem je náhodná veličina

$$F = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{y}\right)^2}{p} : \frac{\sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2}{n-(p+1)}, \tag{2.24}$$

kde

$$\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{y}\right)^2 = S_T \tag{2.25}$$

je **teoretický součet** a součet $\sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2 = S_R$ je **reziduální součet** (2.4).

Kritický obor W_α je vymezen nerovností

$$F \geq F_{1-\alpha}(c-1, n-c), \tag{2.26}$$

kde $F_{1-\alpha}(c-1, n-c)$ je kvantil F-rozdělení s $\nu_1 = c-1$ a $\nu_2 = n-c$ stupni volnosti, $c = p + 1$.

Vede-li celkový F-test k zamítnutí hypotézy H a většina t-testů rovněž, považujeme zvolenou regresní funkci za vyhovující. Při vyšetřování regresní závislosti konstruujeme často takzvané **pásy spolehlivosti.** Statistické programy většinou kreslí kolem regresní přímky dva pásy: Užší pás pro podmíněnou střední hodnotu a širší pás spolehlivosti pro predikci.

Interval spolehlivosti pro podmíněnou střední hodnotu

100(1-α)%-ní dvoustranný interval spolehlivosti pro podmíněnou střední hodnotu Y_i (**pás spolehlivosti kolem regresní funkce**) je vymezen nerovnostmi

$$\hat{Y}_i - t_{1-\alpha/2}(\nu)\, s\left(\hat{Y}_i\right) < Y_i < \hat{Y}_i + t_{1-\alpha/2}(\nu)\, s\left(\hat{Y}_i\right), \tag{2.27}$$

kde $\hat{Y}_i$ je hodnota regresní funkce odpovídající zvolené hodnotě x_i vysvětlující proměnné X, $t_{1-\alpha/2}(\nu)$ je kvantil t-rozdělení s $\nu = n-(p+1)$ stupni volnosti a $s\left(\hat{Y}_i\right)$ je směrodatná chyba (2.28) bodového odhadu $\hat{Y}_i$.

Směrodatná chyba $s\left(\hat{Y}_i\right)$ bodového odhadu $\hat{Y}_i$

$$s\left(\hat{Y}_i\right) = \sqrt{s_R^2 \mathbf{x}_i' \mathbf{H} \mathbf{x}_i}\,, \tag{2.28}$$

kde s_R^2 je reziduální rozptyl (2.16), vektor

$$\mathbf{x}_i' = [\,1,\ f_1(x_i),\ f_2(x_i),\ \ldots,\ f_p(x_i)\,] \tag{2.29}$$

je vektor hodnot regresorů pro danou hodnotu x_i, $\mathbf{x}_i$ je vektor transponovaný k $\mathbf{x}_i'$ a matice $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1}$ je matice (2.17).

Interval spolehlivosti pro individuální předpověď

100(1-α)%- ní dvoustranný interval spolehlivosti pro predikovanou hodnotu proměnné Y_{i0}, odpovídající dané hodnotě x_i vysvětlující proměnné X (**pás spolehlivosti pro predikci**), je vymezen nerovnostmi

$$\hat{Y}_i - t_{1-\alpha/2}(\nu) s\left(\hat{Y}_{i0}\right) < Y_{i0} < \hat{Y}_i + t_{1-\alpha/2}(\nu) s\left(\hat{Y}_{i0}\right), \tag{2.30}$$

kde $\hat{Y}_i$ je hodnota regresní funkce odpovídající zvolené hodnotě x_i vysvětlující proměnné X, $t_{1-\alpha/2}(\nu)$ je kvantil t-rozdělení s $\nu = n-(p+1)$ stupni volnosti a $s\left(\hat{Y}_{i0}\right)$ je směrodatná chyba (2.31) odhadu individuální hodnoty.

Směrodatná chyba $s\left(\hat{Y}_{i0}\right)$ odhadu individuální hodnoty

$$s\left(\hat{Y}_{i0}\right) = \sqrt{s_R^2 (1 + \mathbf{x}_i' \mathbf{H} \mathbf{x}_i)}\,, \tag{2.31}$$

kde s_R^2 je reziduální rozptyl (2.16), $\mathbf{x}_i'$ je vektor (2.29), $\mathbf{x}_i$ je vektor transponovaný k $\mathbf{x}_i'$ a matice ***H*** je matice (2.17).

Analýza rezidui a vlivná pozorování

Rezidua jsou základním *diagnostickým nástrojem,* a to nejen při hodnocení kvality regresní funkce, ale i obecněji při posuzování oprávněnosti předpokladů zvoleného regresního modelu. Jakákoli systematičnost *(nenáhodnost)* zjištěná u rezidui indikuje nějaký (zatím *neidentifikovaný)* nedostatek odhadnutého regresního modelu. Může to být chybně zvolený typ regresní funkce, nevhodný plán experimentu, nenáhodný výběr, nesprávně zvolené vysvětlující proměnné, nesplnění předpokladů metody, špatné představy o modelu, chybná nebo příliš vlivná pozorování, silná vzájemná závislost vysvětlujících proměnných, ale i jiná narušení regresní úlohy (Hebák, Malá & Hustopecký, Vícerozměrné statistické metody 2, p. 92).

Klasická rezidua

popisují rozdíly mezi skutečnými a odhadnutými hodnotami vysvětlované proměnné.

$$e_i = \left(y_i - \hat{Y}_i\right),$$

kde y_i je experimentální hodnota a $\hat{Y}_i$ je vyrovnaná hodnota.

Rezidua e_i by měla především ***vyhovovat předpokladu normality*** **a** ***nezávislosti***.

Nejpoužívanější test, jímž ověřujeme nezávislost reziduí v modelu, je *Durbinův-Watsonův test autokorelace*. Durbinův-Watsonův test používá statistiku

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e^2{}_t}, \tag{2.32}$$

kde $e_i = \left(y_i - \hat{Y}_i\right)$ je reziduum (2.6).

Statistika (2.32) nabývá hodnot z intervalu (0; 4). V případě, že hodnota DW se pohybuje kolem 2, nelze zamítnou hypotézu o nezávislosti náhodných poruch. Blíží-li se hodnota DW 0 nebo 4, jsou rezidua závislá.

Všechny programy nabízejí ***grafy reziduí***. Rezidua e_i zobrazená v závislosti na hodnotách x_i umožňují zhruba ověřit nezávislost reziduí. Je-li regresní funkce správně určena, pak jsou body náhodně rozmístěny kolem vodorovné osy. Jestliže rezidua vykazují určitý trend, je to známka nesprávně zvolené regresní funkce.

Detekce vlivných bodů

Vlivné body zkreslují odhady a zvyšují rozptyl. Lze je rozdělit do dvou skupin:

- ***odlehlé body*****,** které se liší od ostatních v y-ové složce a
- ***extrémy,*** které se liší od ostatních v x-ové složce.

Tyto body ovlivňují výrazně výsledky regrese a uživatel musí rozhodnout, zda jde o hrubé chyby, které je třeba vyloučit, nebo naopak o body, které zlepšují kvalitu a stabilitu regrese.

Statistické programy při identifikaci vlivných bodů využívají vedle klasických reziduí (2.6), která obecně nemají stejný rozptyl, následující rezidua:

Standardizovaná rezidua e_{SI} mají tvar

$$e_{Si} = \frac{e_i}{s_R\sqrt{1 - p_{ii}}}, \tag{2.33}$$

kde e_i je klasické reziduum (2.6), S_R je reziduální směrodatná odchylka (2.19) a p_{ii} jsou diagonální prvky projekční matice (2.34).

Projekční matice P má tvar

$$P = X(X'X)^{-1}X', \tag{2.34}$$

kde $\boldsymbol{X}$ je matice (2.18).

Poněkud lepší diagnostické vlastnosti než standardizovaná rezidua mají **Jackknife rezidua** e_{JI}.

$$e_{Ji} = \frac{e_i}{s_{R(i)}\sqrt{1 - p_{ii}}}, \tag{2.35}$$

kde e_i je klasické reziduum (2.6), $s_{R(i)} = \sqrt{s^2_{R(i)}}$ je reziduální směrodatná odchylka (2.19) při vynechání i-tého pozorování a p_{ii} je prvek matice (2.34).

Kvalita modelu

Vystihneme-li průběh závislosti regresní funkcí (2.5), zajímají nás velikosti odchylek experimentálních hodnot y_i od vyrovnaných hodnot $\hat{Y}_i$ (hodnot ležících na výběrové regresní křivce). Přichází-li v úvahu více typů regresní funkce, můžeme při výběru využít následující kritéria:

- **Reziduální rozptyl $\mathbf{s_R^2}$** (2.16)

$$s_R^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2}{n-(p+1)}, \tag{2.36}$$

Za vhodnější se považuje ta regresní funkce, u níž má reziduální rozptyl menší hodnotu.

- **Index determinace $\mathbf{i_{yx}^2}$**

$$i_{yx}^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}, \tag{2.37}$$

kde součet $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{y}\right)^2 = S_T$ je **teoretický součet** (2.25) a součet

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = S_y \tag{2.38}$$

je **celkový součet**.

Výběrovou regresní funkci považujeme za tím výstižnější, čím je index determinace bližší jedné. Vztah (2.37) pro malé výběry odhad indexu determinace nadhodnocuje. Navíc index závisí na počtu parametrů regresní funkce. Proto statistické programy uvádějí *upravenou hodnotu indexu determinace* i_{kor}^2, kde

$$i_{kor}^2 = 1-(1-i_{yx}^2)\frac{n-1}{n-c}, \tag{2.39}$$

kde n je počet pozorování a $c = p + 1$ je počet parametrů regresní funkce.

V některých statistických programech je index determinace označován jako výběrový **koeficient determinace $\boldsymbol{R^2}$**. Odmocnina z výrazu (2.37) je v programech označována jako **vícenásobný korelační koeficient $\boldsymbol{R}$**.

$$R = \sqrt{i_{yx}^2} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}. \tag{2.40}$$

Některé statistické pakety uvádějí **Akaikeho informační kritérium**

$$AIC = n \ln\left(\frac{S_R}{n}\right) + 2c, \tag{2.41}$$

kde n je rozsah výběru, S_R je reziduální součet čtverců (2.4) a $c = p+ 1$ je počet parametrů regresní funkce. Za vhodnější je považován ten model, pro který je AIC minimální.

Výběr vysvětlujících proměnných

V mnoha případech je účelné zmenšit množinu vysvětlujících proměnných a to např. pro ulehčení interpretace. Metody hledání nejlepšího modelu na základě postupného přidávání proměnných do regresní funkce (forward a stepwise) vycházejí z přírůstku regresního součtu čtverců, jehož velikost je hodnocena pomocí sekvenčních F-testů nebo na základě zvýšení indexu determinace, přičemž použít lze ekvivalentně i hodnoty a testy dílčích korelačních koeficientů. Metoda *forward* se od používanější metody *stepwise* liší jen tím, že při metodě stepwise se po každém zařazení nové proměnné zkoumá, zda by se dříve zařazené proměnné dostaly do modelu při obráceném pořadí zařazování. Při použití metody *backward* je postup obrácený. Začíná se od modelu se všemi vysvětlujícími proměnnými, pak se na základě velikosti poklesu regresního součtu čtverců, indexu determinace nebo pomoci dílčích korelačních koeficientů zkoumá, které proměnné lze z modelu vypustit (Hebák, Malá & Hustopecký, Vícerozměrné statistické metody 2, p. 105).

Postup při lineární regresní analýze:

- Návrh modelu, kdy volíme vhodný tvar regresní funkce, která respektuje teoretický model závislosti. Není–li teoretický model znám, provádíme analýzu bodového diagramu a grafu podmíněných průměrů.
- Odhad regresních parametrů a testy jejich významnosti.
- Regresní diagnostika, kdy provádíme analýzu reziduí a identifikaci vlivných bodů.
- Konstrukce zpřesněného modelu, kdy vycházíme z výsledků regresní diagnostiky, např. vyloučíme vlivné body a podobně.
- Zhodnocení kvality modelu vychází ze statistických charakteristik, testů a regresní diagnostiky. Výsledkem je buď přijetí navrženého modelu, nebo návrh modelu dalšího.

Příklad 1 Korelace

Máme k dispozici data 50 nejlepších světových výkonů desetibojařů (a k nim přepočtené body). Vypočítejte matici korelací mezi všemi proměnnými a určete, která disciplína nejvíce koreluje s celkovým bodovým ziskem.

Tab. 18 Vstupní data

100 m	dálka	koule	Výška	400 m	110m př.	disk	tyčka	oštěp	1500 m	celkem
943	1089	810	915	919	985	827	849	892	798	9026
966	1035	899	840	905	1010	836	880	925	698	8994
992	1081	894	868	885	977	840	910	777	667	8891
989	1063	834	831	960	932	799	910	817	712	8847
929	1000	877	868	907	965	857	880	736	814	8832
1001	908	800	878	889	1007	928	910	898	613	8832
952	967	784	831	997	924	734	1035	844	747	8815
847	1007	870	831	888	891	799	957	926	776	8792
987	1017	810	794	903	993	830	972	859	625	8790
910	1050	869	963	899	958	784	972	749	630	8784
885	932	811	887	939	975	806	849	900	778	8762
975	876	854	906	927	998	864	880	743	711	8735
975	1012	847	887	968	978	905	790	671	692	8725
870	952	851	776	875	984	829	880	924	768	8709
952	1079	704	840	893	1044	768	849	842	735	8706
870	918	766	840	900	953	868	998	843	749	8705
890	915	853	896	826	907	895	880	845	791	8698
1020	1000	816	973	860	972	938	790	762	563	8694
931	1030	788	896	911	958	745	941	708	772	8680
956	1010	819	803	907	985	778	790	836	784	8667
845	915	752	887	926	955	789	819	1004	762	8654
883	1002	809	840	809	978	867	941	824	691	8644
947	935	933	776	934	986	920	849	642	722	8644
874	866	811	831	933	869	871	849	867	863	8634
947	1066	724	850	874	995	755	880	757	779	8627
810	990	909	776	876	878	931	880	908	668	8626
892	1035	722	944	935	968	657	910	803	751	8617
924	957	815	944	998	931	807	819	657	751	8603
924	1002	740	896	953	977	754	910	659	759	8574
931	1043	806	831	827	903	826	957	722	728	8574
919	960	853	831	926	944	825	941	734	641	8573
938	915	831	831	878	978	799	972	686	743	8571
867	950	853	896	896	934	754	972	697	747	8566
943	871	834	973	903	865	780	804	792	791	8554
959	952	817	944	922	895	782	790	756	731	8548
943	987	847	813	898	871	882	731	850	726	8547
839	970	840	831	836	867	812	910	870	760	8534
872	932	891	878	802	950	838	941	797	627	8528
924	922	763	731	904	997	790	910	819	766	8526
971	955	821	794	880	969	887	895	749	605	8526
850	940	742	803	809	913	760	1067	874	766	8524
943	1073	749	749	828	948	728	910	799	796	8522
845	945	796	925	858	857	878	790	763	862	8519
850	898	838	803	905	905	913	790	764	840	8506
812	967	834	944	872	815	807	849	818	782	8500

100 m	dálka	koule	Výška	400 m	110m př.	disk	tyčka	oštěp	1500 m	celkem
892	816	818	831	857	894	956	941	802	692	8497
888	900	882	896	872	965	746	849	860	638	8496
867	871	792	878	888	890	789	819	883	814	8491
913	970	773	982	830	965	796	910	808	544	8490
839	990	805	822	855	926	781	880	774	814	8485

Statistiky → Základní statistiky/tabulky → Korelační matice

Tab. 19 Korelační matice

Proměnná	Korelace (data-50 desetiboju) Označ. korelace jsou významné na hlad. p < ,05000 N=50 (Celé případy vynechány u ChD)										
	100 m	dálka	koule	výška	400 m	110m př	disk	tyčka	oštěp	1500 m	celkem
Body 100 m	1,000	0,329	-0,011	0,062	0,321	0,570	0,083	-0,094	-0,307	-0,423	0,458
Body skok dálka	0,329	1,000	-0,131	-0,027	0,027	0,335	-0,326	0,086	-0,084	-0,105	0,455
Body koule	-0,011	-0,131	1,000	-0,047	-0,006	-0,157	0,480	-0,125	-0,084	-0,230	0,216
Body výška	0,062	-0,027	-0,047	1,000	0,119	-0,119	-0,144	-0,238	-0,192	-0,163	0,047
Body 400 m	0,321	0,027	-0,006	0,119	1,000	0,163	-0,134	-0,215	-0,190	0,153	0,378
Body překážky	0,570	0,335	-0,157	-0,119	0,163	1,000	-0,100	0,144	-0,085	-0,369	0,460
Body disk	0,083	-0,326	0,480	-0,144	-0,134	-0,100	1,000	-0,264	-0,013	-0,239	0,109
Body tyčka	-0,094	0,086	-0,125	-0,238	-0,215	0,144	-0,264	1,000	0,002	-0,203	0,114
Body oštěp	-0,307	-0,084	-0,084	-0,192	-0,190	-0,085	-0,013	0,002	1,000	0,067	0,256
Body 1500 m	-0,423	-0,105	-0,230	-0,163	0,153	-0,369	-0,239	-0,203	0,067	1,000	-0,091
celkem	0,458	0,455	0,216	0,047	0,378	0,460	0,109	0,114	0,256	-0,091	1,000

Závěr:

Můžeme konstatovat, že s celkovým bodovým ziskem nejvíce korelují 3 proměnné (tab. 19): běh na 100 m, skok do dálky a 110 m př. s hodnotou korelačního koeficientu 0,45-0,46.

Příklad 2 Parciální a mnohonásobná korelace

Jevy vedle sebe neexistují izolovaně, ale téměř vždy na naše sledované proměnné působí další proměnné, o kterých nevíme nebo které neumíme změřit. Naše sledované proměnné jsou tak ovlivněny dalšími proměnnými. Může se jednat např. o výšku a váhu. Korelace ostatních proměnných budou pravděpodobně pozitivní. Po jejich vyloučení se směr závislosti může zcela otočit. Ke zjištění použijeme výpočet parciálních korelačních koeficientů.

Známe-li všechny tři korelační součinitele mezi třemi parametry téhož souboru, které označíme r_{xy}, r_{xz}, r_{yz}, pak můžeme stanovit částečnou (parciální) korelaci mezi kterýmikoliv dvěma parametry s vyloučením vlivu třetího, tedy za předpokladu, že třetí parametr je konstantní. Vzorce pro parciální korelační součinitele jsou

$$r_{xy.z} = \frac{(r_{xy} - r_{xz} \cdot r_{yz})}{\sqrt{(1 - r_{xz}^2) \cdot (1 - r_{yz}^2)}}$$

$$r_{xz.y} = \frac{(r_{xz} - r_{xy} \cdot r_{yz})}{\sqrt{(1 - r_{xy}^2) \cdot (1 - r_{yz}^2)}}$$

$$r_{yz.x} = \frac{(r_{yz} - r_{xy} \cdot r_{xz})}{\sqrt{(1 - r_{xy}^2) \cdot (1 - r_{xz}^2)}}$$

Příklad: u skupiny dětí byly vypočítány korelační součinitele mezi
tělesnou výškou a hmotností $r_{xy} = 0{,}91$
výškou a výkonem ve skoku vysokém $r_{xz} = 0{,}86$
hmotností a výkonem ve skoku vysokém $r_{yz} = 0{,}69$

Korelace mezi hmotností a výkonem ve skoku vysokém je překvapivě vysoká a kladná. Uvědomíme-li si ale, že těžší dítě bývá také vyšší, je zřejmé, že vazbu hmotnost/výkon zprostředkuje tělesná výška, kterou bychom měli vyloučit. Pak parciální korelační součinitel mezi hmotností a výkonem ve skoku vysokém tuto vazbu vylučuje:

$$r_{yz.x} = \frac{(0{,}69 - 0{,}91 \cdot 0{,}86)}{\sqrt{(1 - 0{,}91^2) \cdot (1 - 0{,}86^2)}} = -0{,}438$$

Místo původní kladné korelace jsme dostali zápornou parciální korelaci, protože byl vyloučen zprostředkující vliv tělesné výšky. S rostoucí hmotností při stálé tělesné výšce výkon ve skoku vysokém klesá.

Mnohonásobný koeficient korelace se používá v situacích, kdy chceme zjistit celkovou sílu vztahu mezi zvolenou proměnnou na jedné straně a několika dalšími (predikujícími) proměnnými X_2, X_3, …, X_k na straně druhé. Hodnotí se jím význam kumulativního vlivu více proměnných na zvolenou cílovou proměnnou. Mnohonásobný korelační koeficient, který pro tři proměnné značíme $r_{x.yz}$ je roven

$$r_{x.yz} = \sqrt{\frac{(r_{xy}^2 + r_{xz}^2 - 2r_{xz} \cdot r_{xy} \cdot r_{yz})}{1 - r_{yz}^2}} = \sqrt{\frac{(0{,}91^2 + 0{,}86^2 - 2 \cdot 0{,}86 \cdot 0{,}91 \cdot 0{,}69)}{1 - 0{,}69^2}} = 0{,}96$$

Mnohonásobný korelační koeficient mezi výkonem ve skoku vysokém jako cílovou proměnnou a dvěma prediktory má hodnotu 0,96.

Příklad 3 Kanonická korelace

Zadání: Na základě údajů z příkladu 1 zjistěte vztah mezi vektory **x**(X_1, X_2) a **y**(X_4, X_5).

Statistiky → Vícerozměrné průzkumné techniky → Kanonická analýza

Řešení:

V kanonické korelační analýze se zkoumá povaha vztahů mezi dvěma množinami proměnných. Vztahy vyjadřujeme pomocí komponent, což jsou lineární kombinace proměnných z dané množiny proměnných. Komponenty hledáme po dvojicích. V dvojici odpovídá vždy jedna komponenta jedné množině z obou skupin proměnných. První dvojice má mít největší možnou korelaci. Druhá dvojice je tvořena nezávislými (ortogonálními) komponentami k první dvojici a má druhou největší možnou korelaci. Tak postupujeme, až jsou obě množiny proměnných i jejich vzájemné vztahy popsány dvěma systémy nezávislých komponent. V této analýze podobně jako při popisu vztahu jednoduchým korelačním koeficientem se nerozlišuje mezi nezávislými a závislými proměnnými (Hendl, 2004, p. 422).

Kanonické korelace tedy měří intenzitu lineární závislosti mezi dvěma skupinami lineárních funkcí vektorů **x** a **y**. Pomocí programu STATISTICA 10 byly zjištěny kanonické korelační koeficienty a koeficienty kanonických proměnných pro vektor **x** a vektor **y**.

Tab. 20 Výsledky kanonické korelace pro vektor x

	Kanonické váhy, pravá sada (domacnosti)			
Proměnná	**Kořen 1**	Kořen 2		
Počet členů X1	-2,46239	2,15149		
Počet dětí X2	1,68645	-2,80145		

Koeficienty kanonických proměnných pro vektor **x**:

X_1: -1,68621 -2,80160

X_2: 2,46221 2,15170

Tab. 21 Výsledky kanonické korelace pro vektor y

	Kanonické váhy, pravá sada (domacnosti)			
Proměnná	**Kořen 1**	Kořen 2		
Příjem X4	-0,492771	1,12131		
Vydání X5	-0,630966	-1,04978		

Koeficienty kanonických proměnných pro vektor **y**:

X_4: 0,49269 -1,12142

X_5: 0,63102 1,04982

Lineární kombinace složek náhodného vektoru **x** = (X_1, X_2) jsou kanonické proměnné

$$U_1 = - 2{,}46239\, X_1 + 1{,}68645\, X_2$$

a

$$U_2 = 2{,}15149\, X_1 - 2{,}80145\, X_2.$$

Lineární kombinace složek náhodného vektoru **y** = (X_4, X_5) jsou kanonické proměnné

$$V_1 = - 0{,}492771\, X_4 - 0{,}630966\, X_5$$

a

$$V_2 = 1{,}12131\, X_4 - 1{,}04978\, X_5.$$

Tab. 22 Souhrn kanonické korelace

	Souhrn kanonické analýzy (domacnosti) Kanonické R: ,88041 Chi2(4)=46,660 p=0,0000	
N=34	**L sada**	P sada
Počet proměnných	2	2
Získaný rozptyl	100,000%	100,000%
Celková redundance	61,7427%	46,7633%
Proměnné: 1	Příjem X4	Počet členů X1
2	Vydání X5	Počet dětí X2

Intenzitu lineární závislosti mezi dvěma skupinami lineárních funkcí vektorů **x** a **y**, tj. mezi kanonickými náhodnými proměnnými **u** a **v** měří kanonický korelační koeficient $R_{XY} = 0{,}88041$.

Závěr:

Z lineárních rovnic

$$-1{,}68621\ X_1 + 2{,}46221\ X_2 = 0{,}49269\ X_4 + 0{,}63102\ X_5$$

$$-\ 2{,}80160\ X_1 + 2{,}15170\ X_2 = -1{,}12142\ X_4 + 1{,}04982\ X_5$$

získaných metodou kanonických korelací plyne, že při zvýšení počtu členů domácnosti a nepatrném snížení veličiny X_2 (počet dětí) vede ke snížení veličiny X_4 (příjem) a zvýšení veličiny X_5 (vydání).

Hodnota skupinového korelačního koeficientu $R_{XY} = 0{,}88$ signalizuje silnou lineární závislost mezi vektory **x** a **y**. Uvedené rovnice tedy popisují 78% variability dat.

Příklad 4 Vícerozměrný lineární model

Zadání: U dvaceti vybraných domácností byly zjištěny údaje o čtvrtletních výdajích na potraviny a nápoje (y), čtvrtletním příjmu domácnosti (x_1), počtu dětí (x_2), průměrném věku vydělávajících členů domácnosti (x_3) a počtu členů domácnosti (x_4). Rozhodněte, které proměnné významně přispívají k vysvětlení variability hodnot čtvrtletních výdajů a zkonstruujte lineární regresní model s nejlepší podmnožinou vysvětlujících proměnných.

Data: n = 20, x_j = vysvětlující proměnná, y závislá proměnná.

Tab. 23 Vstupní data

příjem [Kč] x_1	počet dětí x_2	průměrný věk x_3	počet členů x_4	výdaje [Kč] y
11172	0	55	1	3464
8868	0	21	1	1982
17414	0	49	1	3228
10730	0	22	1	3034
24110	0	62,5	2	10146
38530	0	57	2	8202
22902	0	54,5	2	9332
25448	0	57,5	2	7096
20326	0	28	2	6248
39186	1	38,5	3	13816
28758	1	45,5	3	10328
33658	1	28,5	3	4786
24272	1	36	3	9710
30386	2	35	4	10778
31750	2	30,5	4	10568
39456	2	32,5	4	14260
48458	2	38	4	10934
37990	2	37	4	6388
24920	2	33,5	4	8584
40064	3	47	5	16950

Řešení:

1. Nejprve zařadíme do regresního modelu všechny vysvětlující proměnné. Klasickou metodou nejmenších čtverců byla určena regresní funkce (tab. 24).

Statistiky → Vícenásobná regrese

Tab. 24 Výsledky regrese

N=20	Výsledky regrese se závislou proměnnou : y (regrese-priklad4) R= ,84107828 R2= ,70741268 Upravené R2= ,62938939 F(4,15)=9,0667 p<,00063 Směrod. chyba odhadu : 2448,5					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(15)	p-hodn.
Abs.člen			-4027,04	2981,415	-1,35071	0,196813
x1	0,114384	0,257774	0,04	0,095	0,44374	0,663568
x2	-0,334792	0,561725	-1348,29	2262,207	-0,59601	0,560057
x3	0,258589	0,159904	84,19	52,060	1,61716	0,126676
x4	1,043321	0,643472	3353,42	2068,233	1,62139	0,125760

neboli

$$\hat{y} = -4027 + 0{,}042063\, x_1 - 1348{,}3\, x_2 + 84{,}188\, x_3 + 3353{,}4\, x_4$$

s upraveným koeficientem determinace, který bere v potaz počet nezávislých proměnných, $\hat{R}^2 = 0{,}629$ a reziduální směrodatnou odchylkou $s_l = 2448{,}5$.

Podle výsledků t -testů nemůžeme zamítnout žádnou z hypotéz H_0: $\beta_j = 0$ pro j = 1, 2, 3, 4. Podle výsledků F - testu naopak alespoň jeden z regresních koeficientů je nenulový. Příčinou je existence multikolinearity mezi proměnnými. *Párové korelační koeficienty* (tab. 25) mezi dvojicemi vysvětlujících proměnných signalizují, že silná závislost je především mezi proměnnými x_2 a x_4, tedy mezi počtem dětí a členů domácnosti. ($r_{24} = 0{,}9581$) a rovněž mezi proměnnými x_1 a x_4, tedy mezi příjmem a počtem členů domácnosti ($r_{14} = 0{,}7884$). Na druhé straně je závislost velmi slabá mezi věkem a počtem členů domácnosti ($r_{34} = -0{,}17322$).

Tab. 25 Korelační matice

Proměnná	Korelace (regrese-priklad4) Označ. korelace jsou významné na hlad. p < ,05000 N=20 (Celé případy vynechány u ChD) **Průměry**	Sm.odch.	x1	x2	x3	x4
x1	27919,90	10937,07	1,0000	0,6819	0,0895	0,7884
x2	0,95	1,00	0,6819	1,0000	-0,2797	0,9581
x3	40,42	12,35	0,0895	-0,2797	1,0000	-0,1732
x4	2,75	1,25	0,7884	0,9581	-0,1732	1,0000

2. O tom, které z proměnných v modelu ponecháme rozhodneme pomocí dopředné krokové regrese (tab. 26).

Statistiky → Vícenásobná regrese → Detailní nastavení → Další možnosti (kroková nebo hřebenová regrese)

Tab. 26 Výsledky dopředné regrese

N=20	Výsledky regrese se závislou proměnnou : y (regrese-priklad4) R= ,83178121 R2= ,69185998 Upravené R2= ,65560821 F(2,17)=19,085 p<,00005 Směrod. chyba odhadu : 2360,3 **b***	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-3063,80	2393,396	-1,28011	0,217701
x4	0,824137	0,136699	2648,92	439,374	6,02886	0,000014
x3	0,324514	0,136699	105,65	44,505	2,37393	0,029647

3. Byly vybrány 2 proměnné x_3 a x_4. Dostaneme tak regresní rovnici ve tvaru

$$\hat{y} = -3063{,}8 + 2648{,}92\, x_3 + 105{,}65\, x_4$$

s upraveným koeficientem determinace $\hat{R}^2 = 0{,}655$ a reziduální směrodatnou odchylkou $s_l = 2360{,}3$.

Závěr: Ze srovnání jednotlivých modelů plyne, že nejlepším modelem popisujícím závislost výdajů za potraviny na příjmu, počtu dětí, věku a počtu členů domácnosti je model

$$\hat{y} = -3063{,}8 + 2648{,}92\, x_3 + 105{,}65\, x_4$$

Příklad 5 Validizace nové metody

Zadání: Osm respondentů se zúčastnilo experimentu spojeného s diagnostikou a analýzou složení lidského těla pomocí 2 přístrojů různých výrobců. Zjistěte, zda mezi výsledky uvedených přístrojů je podstatný rozdíl. Uvedená data představují procentuální zastoupení tělesného tuku.

Tab. 27 Výsledky dopředné regrese

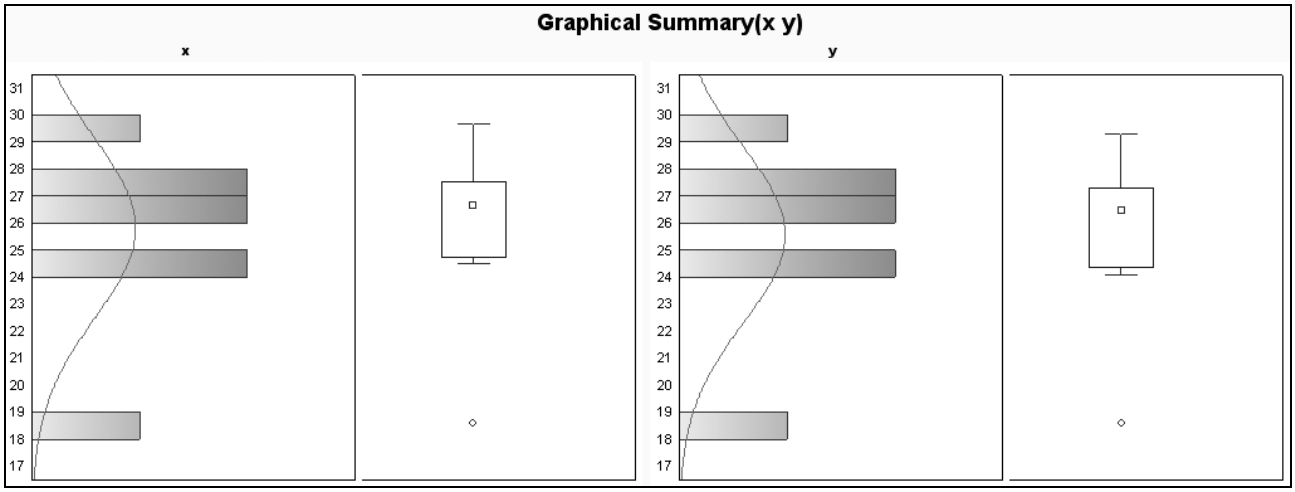
Číslo	metoda 1 - (x)	metoda 2 - (y)
1	18,6	18,58
2	27,6	27,37
3	27,5	27,27
4	25,0	24,64
5	24,5	24,10
6	26,8	26,33
7	29,7	29,33
8	26,5	26,63

Řešení:

Pokud obě metody poskytují stejné výsledky, bude závislost y na x lineární $y = \beta_1 x + \beta_0$ s jednotkou směrnicí $\beta_1 = 1$ a nulovým úsekem $\beta_0 = 0$.

Pro porovnání výsledků y vůči výsledkům x určíme odhady b_0 a b_1 a zkonstruujeme 95% -ní interval spolehlivosti pro úsek i pro směrnici.

1. Protože rozsah výběru je malý (n = 8), omezíme se na grafickou analýzu, tj. na histogram a krabicový graf



Obr. 9 Histogram a krabicový graf

Krabicový graf (obr. 9) ukazuje, že vlivným bodem je bod č. 1. Po vynechání vlivných bodů jsme získali následující odhady - viz tab. 28.

Tab. 28 Změna úseku a směrnice

	b_0	**$s(b_0)$**	**b_1**	**$s(b_1)$**
původní data	0,3333	0,61184	0,97761	0,023568
vynechání č. 1	-0,43952	1,37251	1,00611	0,051121

Srovnáme-li hodnoty z tabulky 6 s odhadem úseku $b_0 = 0{,}3333$ ($\pm$ 0,61184) a s odhadem směrnice $b_1 = 0{,}97761$ ($\pm$ 0,023568) vidíme, že bod č. 1 ovlivňuje úsek i směrnici původní regresní přímky.

2. Metodou nejmenších čtverců byly vypočítány odhady parametrů a směrodatné odchylky odhadů. Odhad úseku $b_0 = 0{,}3333$ ($\pm$ 0,61184), $b_1 = 0{,}97761$ ($\pm$ 0,023568). Určíme jednoduché 95% -ní intervaly spolehlivosti pro parametry β_0 a β_1.

$$b_0 - t_{1-\alpha/2}\,(n-m)\; s(b_0) \le \beta_0 \le b_0 + t_{1-\alpha/2}\,(n-m)\; s(b_0),$$

po dosazení

$$0{,}3333 - 2{,}447 \cdot 0{,}61184 \le \beta_0 \le 0{,}3333 + 2{,}447 \cdot 0{,}61184$$

a po vyčíslení

$$-1{,}16387 \le \beta_0 \le 1{,}83047$$

Protože *95% interval spolehlivosti pro úsek regresní přímky* zahrnuje nulu, nelze úsek β_0 považovat za významně odchýlený od nuly. Změny úseku regresní přímky vynecháním vlivného bodu leží v uvedeném intervalu spolehlivosti, proto je považujeme za statisticky nevýznamné.

Analogicky určíme 95% -ní interval spolehlivosti pro parametr β_1.

$$b_1 - t_{1-\alpha/2}\,(6)\; s(b_1) \le \beta_1 \le b_1 + t_{1-\alpha/2}\,(6)\; s(b_1),$$

po dosazení

$$0{,}97761 - 2{,}447 \cdot 0{,}02357 \le \beta_1 \le 0{,}97761 + 2{,}447 \cdot 0{,}02357$$

a po vyčíslení

$$0{,}91993 \le \beta_1 \le 1{,}03529.$$

Protože *95% interval spolehlivosti pro směrnici regresní přímky* obsahuje jedničku, můžeme směrnici β_1 považovat za jednotkovou. Rovněž změny směrnice z tabulky 6 leží v uvedeném intervalu spolehlivosti a jsou také statisticky nevýznamné.

Závěr: Intervaly spolehlivosti úseku a směrnice indikují, že úsek regresní přímky lze považovat za nulový, tj. $\beta_0 = 0$ a také směrnice β_1 se významně neliší od jedničky. Rozdíly mezi výsledky získanými oběma přístroji jsou statisticky nevýznamné a přístroje můžeme považovat za rovnocenné.

Příklad 6 Porovnání dvou regresních přímek

Zadání: U dvaceti prodaných ojetých automobilů určité značky byla zjištěna cena, stáří auta a počet ujetých kilometrů. Závislost ceny na stáří automobilu popište regresní přímkou. Rovněž závislost ceny automobilu na počtu ujetých kilometrů charakterizujte regresní přímkou a obě přímky porovnejte.

Data: n = 20, y = cena auta [tis. Kč], x_1 = stáří auta [roky], x_2 = ujeté kilometry [tis. km]

Tab. 29 Vstupní data

i	x_{i1}	x_{i2}	y_i	i	x_{i1}	x_{i2}	y_i
1	0,6	1,1	55,0	11	5,0	36,0	34,0
2	1,0	2,5	54,6	12	5,1	66,2	31,0
3	1,1	10,4	50,6	13	5,2	44,5	29,0
4	2,0	4,5	51,1	14	5,6	42,0	31,6
5	2,3	31,4	47,0	15	5,9	36,4	34,0
6	2,5	8,6	50,0	16	6,0	82,6	25,6
7	3,0	32,4	43,6	17	6,1	64,5	28,0
8	4,1	25,3	41,3	18	6,3	70,8	24,6
9	4,4	16,0	43,0	19	6,8	78,7	27,0
10	4,8	54,0	39,9	20	7,5	90,2	17,6

Řešení:

1. Určíme obě *regresní přímky*.

Statistiky → Vícenásobná regrese

Tab. 30 Odhady parametrů, reziduální součty čtverců, odhady reziduálních rozptylů.

Výsledky regrese se závislou proměnnou : cena (auta)
R= ,96260583 R2= ,92660999 Upravené R2= ,92253277
F(1,18)=227,26 p<,00000 Směrod. chyba odhadu : 3,1047

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(18)	p-hodn.
Abs.člen			59,95239	1,617693	37,0604	0,000000
stáří	-0,962606	0,063853	-5,16469	0,342592	-15,0753	0,000000

Výsledky regrese se závislou proměnnou : cena (auta)
R= ,93210080 R2= ,86881190 Upravené R2= ,86152367
F(1,18)=119,21 p<,00000 Směrod. chyba odhadu : 4,1509

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(18)	p-hodn.
Abs.člen			52,55444	1,629991	32,2422	0,000000
km	-0,932101	0,085371	-0,36661	0,033578	-10,9182	0,000000

	b_{0j}	b_{1j}	RSC_j	s_j^2
stáří	59,952	-5,1647	173,50	9,639
ujeté km	52,554	-0,3666	310,14	17,230

Mezi stářím automobilu a jeho cenou je velmi těsná nepřímá lineární závislost, kterou charakterizuje regresní přímka

$$y' = 59{,}952\ (\pm\ 1{,}6177) - 5{,}1674\ (\pm\ 0{,}3426)\ x_1$$

a korelační koeficient $R_1 = -0{,}9626$

Závislost ceny automobilu na počtu ujetých kilometrů je rovněž nepřímá a velmi těsná. Tuto lineární závislost mezi naměřenými hodnotami charakterizuje regresní přímka

$$y' = 52{,}559\ (\pm\ 1{,}63) - 0{,}3666\ (\pm\ 0{,}0336)\ x_2$$

a korelační koeficient $R_2 = -0{,}9321$

2. Před vlastním testováním úseků a směrnic regresních přímek ověříme *rovnost reziduálních rozptylů* pomocí F - testu. Hodnotu statistiky

$$F_2 = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)} = \frac{17{,}7230}{9{,}639} = 1{,}7875$$

porovnáme s kvantilem $F_{0,95}(18,18) = 2{,}23$.

Protože $1{,}7875 < 2{,}23$ nezamítáme předpoklad rovnosti rozptylů.

3. Za předpokladu homoskedasticity testujeme hypotézu o *homogenitě úseků*, tzn.

$$H_0: \beta_{01} = \beta_{02} \text{ proti } H_A: \beta_{01} \neq \beta_{02}.$$

Nejprve určíme podle vztahu (6.132) v Meloun & Militký (2004, p. 607) váhové koeficienty w_{B1} a w_{B2} odpovídající úsekům obou přímek.

$$w_{B1} = \frac{20\sum_{i=1}^{20}(x_{i1} - 4{,}265)^2}{\sum_{i=1}^{20} x_{i1}^2} = \frac{20821{,}255}{44593} = 3{,}68$$

$$w_{B2} = \frac{20\sum_{i=1}^{20}(x_{i2} - 39{,}905)^2}{\sum_{i=1}^{20} x_{i2}^2} = \frac{20.15282{,}49}{47130{,}67} = 6{,}48.$$

Ze vztahu (6.138) v Meloun & Militký (2004, p. 609) vypočítáme sdružený odhad úseku.

$$b_{oC} = \frac{w_{B1} b_{01} + w_{B2} b_{02}}{w_{B1} + w_{B2}} = 55{,}233.$$

Dosadíme do testační statistiky F_I - viz (6.140) v Meloun, Militký (2004, p. 609)

$$F_I = \frac{3{,}68\,(59{,}952 - 55{,}233)^2 + 6{,}48\,(52{,}554 - 55{,}233)^2}{\frac{483{,}64}{36}} = 9{,}562.$$

Protože kvantil $F_{0,95}(1,36) = 4,128$ je menší než $F_I = 9,562$, nelze na hladině významnosti $\alpha = 0,05$ považovat úseky regresních přímek za shodné.

4. Zda mají uvažované regresní přímky stejnou směrnici ověříme testem *homogenity směrnic*. Platí-li hypotéza H_0: $\beta_{11} = \beta_{12}$, pak hodnota testační statistiky $F_S < F_{0,95}(1,36)$.

   Podle vztahu (6.143a) v Meloun & Militký (2004, p. 610) určíme sdružený odhad celkové směrnice

$$b_{1C} = \frac{82,1255.(-5,1647) + 15282,49.(-0,3666)}{15364,615} = -0,3922.$$

   Testační statistika F_S - viz (6.144) v Meloun, Militký (2004, p. 610) má hodnotu

$$F_S = \frac{82,1255(-5,1647 + 0,3922)^2 + (-0,3666 + 0,3922)^2}{\frac{483,64}{36}} = 139,236.$$

Protože hodnota testačního kritéria je podstatně větší než kvantil $F_{0,95}(1,36) = 4,128$, nelze na hladině významnosti $\alpha = 0,05$ považovat regresní přímky za rovnoběžné.

Závěr:

Výsledky testů homogenity úseků a homogenity směrnic ukázaly, že rozdíly u srovnávaných regresních přímek jsou statisticky významné. Tento závěr také potvrzuje *test shody regresních přímek*, čili regresní přímky nelze pokládat za totožné.

Odhad ceny ojetého auta na základě stáří je odlišný od odhadu ceny na základě ujetých kilometrů. Na odhady mají vliv další vysvětlující proměnné.

Příklad 7 Kvadratický regresní model

Zadání: Bylo změřeno procentuální zlepšení startovní reakce (*r)* v závislosti na počtu tréninkových jednotek (*t*). Sestavte regresní model, a vyjádřete přesnost modelu.

Tab. 31 Vstupní data

Pořadové číslo	počet tréninků	reakce
1	0	1,000
2	10	1,000
3	20	0,997
4	30	0,996
5	40	0,993
6	50	0,985
7	60	0,983
8	70	0,978
9	80	0,973
10	90	0,961
11	100	0,958

Řešení: 1. *Sestavení regresního modelu*

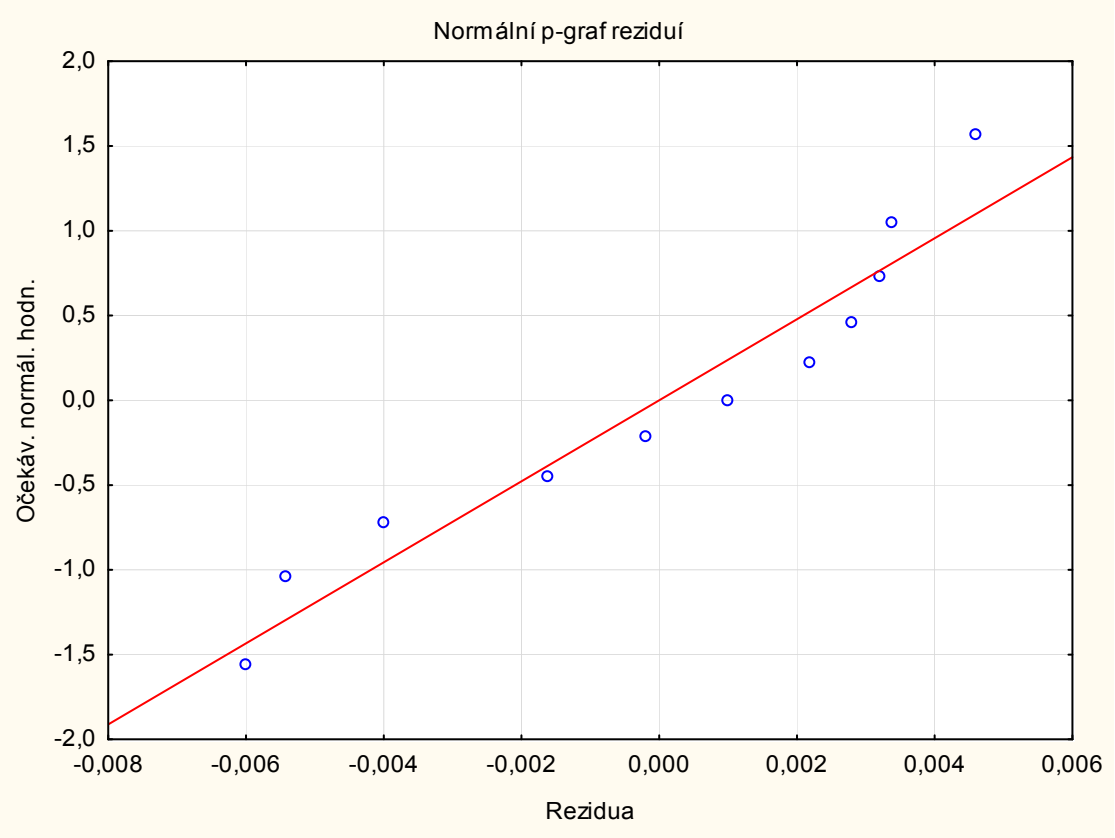
Nejprve byl určen ze všech 11 bodů lineární model (tab. 32)

Tab. 32 Výsledky regrese

	Výsledky regrese se závislou proměnnou : reakce R= ,96858132 R2= ,93814978 Upravené R2= ,93127753 F(1,9)=136,51 p<,00000 Směrod. chyba odhadu : ,00395					
N=11	**b***	Sm.chyba z b*	b	Sm.chyba z b	t(9)	p-hodn.
Abs.člen			1,006000	0,002228	451,5418	0,000000
počet tréninků	-0,968581	0,082899	-0,000440	0,000038	-11,6839	0,000001

$$r = 1{,}006\ (\pm\ 2{,}23.10^{-3}) - 0{,}00044 \pm 3{,}7.10^{-5}\ t$$

s koeficientem determinace R^2 = 0,93815, střední kvadratická chyba MEP = $0{,}17.10^{-5}$, s_e = 0,00395. Testování kvality modelu však ukázalo, že rezidua vykazují trend. Tento závěr potvrdil také p-graf reziduí (obr. 10). Trend v reziduích signalizuje potřebu zavedení kvadratického členu.



Obr. 10 P-graf reziduí

Klasickou metodou nejmenších čtverců byl ze všech 11 bodů určen kvadratický regresní model (tab. 33).

Statistiky → Pokročilé lineární/nelineární modely → Obecné lineární modely → Polynomická regrese →

Tab. 33 Výsledky kvadratické regrese

Efekt	Odhady parametrů (reakce) Sigma-omezená parametrizace									
	reakce Param.	reakce Sm.Ch.	reakce t	reakce p	-95,00% LmtSpol.	+95,00% LmtSpol.	reakce Beta (ß)	reakce Sm.Ch. ß	-95,00% LmtSpol.	+95,00% LmtSpol.
Abs. člen	1,000545	0,001398	715,5962	0,000000	0,997321	1,003770				
počet tréninků	-0,000076	0,000065	-1,1739	0,274204	-0,000226	0,000074	-0,168101	0,143201	-0,49832	0,162122
počet tréninků^2	-0,000004	0,000001	-5,8038	0,000403	-0,000005	-0,000002	-0,831113	0,143201	-1,16134	-0,500890

Závislé Proměnná	Test SČ celého modelu vs. SČ reziduí (reakce)										
	Vícenás. R	Vícenás. R2	Upravené R2	SČ Model	sv Model	PČ Model	SČ Rezid.	sv Rezid.	PČ Rezid.	F	p
reakce	0,994047	0,988130	0,985162	0,002243	2	0,001122	0,000027	8	0,000003	332,9771	0,000000

$$r = 1{,}0005\ (\pm\ 1{,}3982.10^{-3}) - 3{,}64.10^{-6}\ (6{,}3.10^{-7})\ t^2$$

s koeficientem determinace R^2= 0,98813, MEP = $5{,}0464.10^{-6}$, $s_e = 1{,}8353.10^{-3}$

Závěr:

Použitím kvadratického modelu došlo nejen ke zlepšení statistických charakteristik, než při použití lineárního modelu

Metoda hlavních komponent

Víme, že je nutně třeba identifikovat odlehlá. resp. příliš vlivná pozorování, protože někdy až dramaticky ovlivňuji výsledky nejen regresní analýzy či analýzy rozptylu. Rovněž předpoklad homoskedasticity či vícerozměrného normálního rozdělení není pouhým konstatováním, ale vážně míněným varováním i doporučením pro případné transformace dat nebo modelu. Podobně vzájemná silná lineární závislost *(multikolinearita)* vysvětlujících proměnných je vážným nebezpečím pro interpretaci regresních charakteristik. Je to přirozený důsledek obecnější situace, ve které *rozměr* prostoru, v němž se data nacházejí,je ve skutečnosti *nižší* než je počet sledovaných veličin.

Analýzou hlavních komponent (Principal Component Analysis – PCA) se doporučuje začít téměř každou vícerozměrnou úlohu.

Cíle metody hlavních komponent

U mnoha výzkumných úloh se lze setkat se situací, kdy výchozí počet proměnných, sledovaných u zkoumaných jevů a procesů, je značný a pro interpretaci nepřehledný. Pro zjednodušení analýzy a snadnější hodnocení výsledků je často vhodné zkoumat, zda by studované vlastnosti pozorovaných objektu nebylo možné nahradit menším počtem jiných (třeba i umělých) proměnných s co nejmenší ztrátou informace.

Můžeme použít 2 metody: metodu hlavních komponent (PCA) a její rozšíření faktorovou analýzu. Cílem obou metod je nalézt v *pozadí stojící* a tedy *skryté* (umělé, neměřitelné, latentní) proměnné (dále *komponenty*), které dostatečně vysvětlují původní variabilitu. Tyto nově vytvořené proměnné jsou lineární kombinací původních měřitelných proměnných. Formálně můžeme popsat variabilitu původních proměnných lineární kombinací jednotlivých faktorů, ale obsahově se jedná o jinou interpretaci. Pracujeme totiž s komponentami, které nejsou přímo pozorovatelné proměnné jak je tomu u regrese. Vztah mezi původní proměnnou a novým faktorem se popisuje pomocí korelačního koeficientu, který se nazývá *faktorová zátěž*.

Analýza hlavních komponent je často využívána u vícerozměrných metod jako první krok při velkém počtu měření (případů) nebo proměnných typicky s úkolem provést jejich redukci. Pro splnění úkolu se postupuje následovně:

1. Komponenty jsou zařazovány v pořadí takovém, že první vysvětluje největší procento celkové variability, a jsou řazeny v pořadí podle vysvětlení původní variability.
2. Každá další komponenta vysvětluje co nejvíce ze zbývající celkové variability.
3. Komponenty již nejsou vzájemně korelované.

Počet hlavních komponent

Ve statistických software jsou k dispozici 3 pomocná kritéria, podle nichž rozhodneme o počtu komponent:

1. počet vlastních čísel komponent větších než 1
2. použijeme tolik komponent, které vysvětlují určité procento (např. 90 %) původní variability
3. použijeme Scree graf (je vytvořen sestupně z vlastních čísel komponent), kde hledáme bod zlomu od rychlého klesání k pozvolnému.

Faktorová analýza

Faktorová analýza je další statistická metoda, která je zaměřená na vytváření *nových* proměnných a na snížení rozsahu *(redukci)* dal s co nejmenší ztrátou informace. Nové proměnné jsou latentní, skryté, nepřímo pozorovatelné. Ve srovnání s metodou hlavních komponent hledá vzájemné souvislosti vstupních proměnných.

Jedním ze základních cílů faktorové analýzy je posoudit strukturu vztahů sledovaných proměnných a zjistit tak, zda dovoluje jejich rozdělení do skupin, ve kterých by studované proměnné ze stejných skupin spolu nekorelovaly než proměnné z různých skupin. Těmto skupinám říkejme faktory, které by měly umožnit lepší pochopení vstupních proměnných.

Povaha faktorové analýzy je spíše heuristická a průzkumná (explorativní) než ověřovací *(konfirmační)*. Konfirmační faktorová analýza v tomto textu není blíže vysvětlena, čtenář tuto problematiku může nalézt např. knize prof. Hendla Přehled statistických metod. Faktorová analýza je často kritizovaná Pochybnosti se týkají nejednoznačnosti řešení v důsledku subjektivity mnoha kroků i cílů, přibližnost výsledků a mlhavé interpretace.

Jednoduchá struktura a rotace faktorů

Vlastní faktorovou analýzu provádíme ve 4 krocích:

1. Určíme počet faktorů. Např. pomocí metody hlavních komponent.
2. Určíme faktorové zátěže mezi faktory a původními proměnnými.
3. Pro lepší interpretovatelnost provedeme rotaci matice faktorových zátěží.
4. Odhadneme faktorová skóre

Pojem rotace faktorů označuje transformaci matice faktorových zátěži. Takových transformací že existuje nekonečně mnoho a je otázkou, která budeme považovat za optimální. Literatura a statistické software nabízejí celou řadu rotačních algoritmů. Pro lepší interpretaci je dobré, aby faktorové zátěže byly buď blízko 1 nebo 0, což znamená, že korelační koeficient vztahu mezi původní proměnnou a faktorem je buď silný nebo slabý (žádný). Tak zajistíme zařazení původních proměnných k některým faktorům.

Nejznámější rotace

Ouartimax

Kritériem je funkce, která je součtem čtvrtých mocnin faktorových zátěží. Metoda *quartimax* produkuje *obecný faktor,* protože na rozdíl od doporučované metody *varimax* je rozptyl je počítán přes celou matici a nikoli postupně pro všechny sloupce. Zátěže zbývajících faktorů pak bývají nižší než při použití metody varimax.

Varimax

Požadavkům jednoduché struktury se nejvíce přibližuje metoda varimax. Tato metoda volí transformační matici takovou, aby součet rozptylů druhých mocnin faktorových zátěží v jednotlivých sloupcích byl co největší. Metoda varimax je nejpoužívanější metoda pro rotaci faktorů. Produkuje ortogonální faktory, které splňují představy o jednoduché struktuře.

Odhadnuté hodnoty společných faktorů pro jednotlivé případy (respondenty, objekty), nazývané *faktorové skóre*. Tyto jsou nejen užitečným nástrojem diagnózy dat, ale zároveň případným důležitým *vstupem* do dalších analýz a postupů (např. zpětná rekonstrukce korelační matice dat).

Příklad 1 Metody s latentními proměnnými

Zadání: U třiceti čtyř vybraných domácností byly zjištěny údaje o počtu členů domácnosti (X_1), počtu dětí (X_2), průměrném věku (X_3), měsíčnímu příjmu domácnosti (X_4) a měsíčních výdajích za potraviny a nápoje (X_5) - viz. tabulku 34. Pomocí metod s latentními proměnnými zredukujte počet vstupních proměnných.

Tab. 34 Údaje o domácnostech

Domác-nosti	Počet členů X_1	Počet dětí X_2	Průměr. věk X_3	Příjem X_4	Vydání X_5
1	2	0	60,5	14290	5177
2	2	0	60,0	13459	5840
3	1	0	55,0	5586	1732
4	3	1	38,5	19593	6908
5	4	2	35,0	15193	5389
6	4	2	34,5	14741	5683
7	2	0	62,5	12055	5073
8	2	0	57,0	19265	4101
9	2	0	60,5	7908	4584
10	2	0	54,5	11451	4666
11	3	1	45,5	14379	5164
12	4	2	34,5	16236	5178
13	5	2	47,0	20032	8475
14	4	2	30,5	15875	5284
15	2	0	57,5	12724	3548
16	3	1	28,5	16829	2393
17	4	2	33,5	13998	5155
18	4	2	32,5	19728	7130
19	3	1	36,0	12136	4855
20	2	0	28,0	20484	2357
21	2	0	59,5	19187	4339
22	6	4	33,5	20462	3786
23	1	0	21,0	4434	991
24	1	0	21,0	4089	1936
25	1	0	49,0	8707	1614
26	4	2	38,0	24229	5467
27	2	0	39,0	13737	2451
28	1	0	22,0	5365	1517
29	1	0	23,0	4810	832
30	4	2	37,0	18995	3194
31	4	2	33,5	12460	4292
32	4	2	36,0	20146	3597
33	2	0	28,0	10163	3124
34	1	0	28,0	13998	1552

Řešení:

Pod označením metody s latentními proměnnými rozumíme skupinu metod, které popisují a v jistém smyslu vysvětlují pozorovaná data pomocí jejich závislosti na nepozorované charakteristice, kterou lze za určitých předpokladů matematicky zkonstruovat. Nejznámější jsou dvě příbuzné metody, a to analýza hlavních komponent a faktorová analýza.

Obě metody se snaží o vyjádření původních proměnných pomocí menšího počtu latentních proměnných. Od latentních proměnných se v obou metodách požaduje, aby maximálně reprezentovaly (vysvětlovaly) původní proměnné. Konkretizace tohoto požadavku je v obou metodách odlišná. V metodě PCA latentní proměnné (komponenty) vysvětlují maximum celkového rozptylu původních proměnných. V metodě FA latentní proměnné (faktory) vysvětlují především vzájemné souvislosti mezi pozorovanými proměnnými.

1. Nejprve provedeme korelační analýzu. Z korelační matice **R** zjistíme hodnoty jednoduchých korelačních koeficientů, tj. těsnost párových závislostí mezi jednotlivými náhodnými veličinami.

Statistiky → Základní statistiky/tabulky → Korelační matice

Tab. 35 Barevná korelační matice

	Barevná matice abolut. hodnoty r (domacnosti) N=34 (Celé případy vynechány u ChD) abs(r)>= 0	0,10	0,20	0,30	0,40	0,5	0,60	0,70	0,80	0,90	1
Proměnná	**Počet členů X1**	Počet dětí X2	Průměr. věk X3	Příjem X4	Vydání X5						
Počet členů X1	1,000	0,952	0,114	0,698	0,651						
Počet dětí X2	0,952	1,000	0,291	0,571	0,472						
Průměr. věk X3	0,114	0,291	1,000	0,117	0,363						
Příjem X4	0,698	0,571	0,117	1,000	0,577						
Vydání X5	0,651	0,472	0,363	0,577	1,000						

Z korelační matice (tab. 35) plyne, že $r_{12} = 0{,}952$ signalizuje velmi těsnou lineární závislost mezi veličinami X_1 (počet členů domácnosti) a X_2 (počet dětí). Koeficient korelace $r_{14} = 0{,}698$ signalizuje těsnou závislost mezi veličinami X_1 a X_4., tj. závislost mezi počtem členů domácnosti a příjmem. Rovněž těsná lineární závislost je mezi počtem členů domácnosti a měsíčním vydáním domácnosti ($r_{15} = 0{,}651$). Na druhé straně je velmi slabá závislost mezi věkem a počtem členů ($r_{13} = -0{,}114$).

2. Další procedura, kterou modul Vícerozměrná analýza nabízí je matice parciálních koeficientů

Tab. 36 Matice parciálních koeficientů

	X_1	X_2	X_3	X_4	X_5
X_1	1	0,94962	0,164513	0,47399	0,58107
X_2		1	-0,33066	-0,33311	-0,43831
X_3			1	0,05432	0,32100
X_4				1	-0,07498
X_5					1

Parciální korelační koeficienty vyjadřují závislost mezi dvěma náhodnými veličinami ze skupiny náhodných veličin, přičemž vliv ostatních veličin se považuje za konstantní.

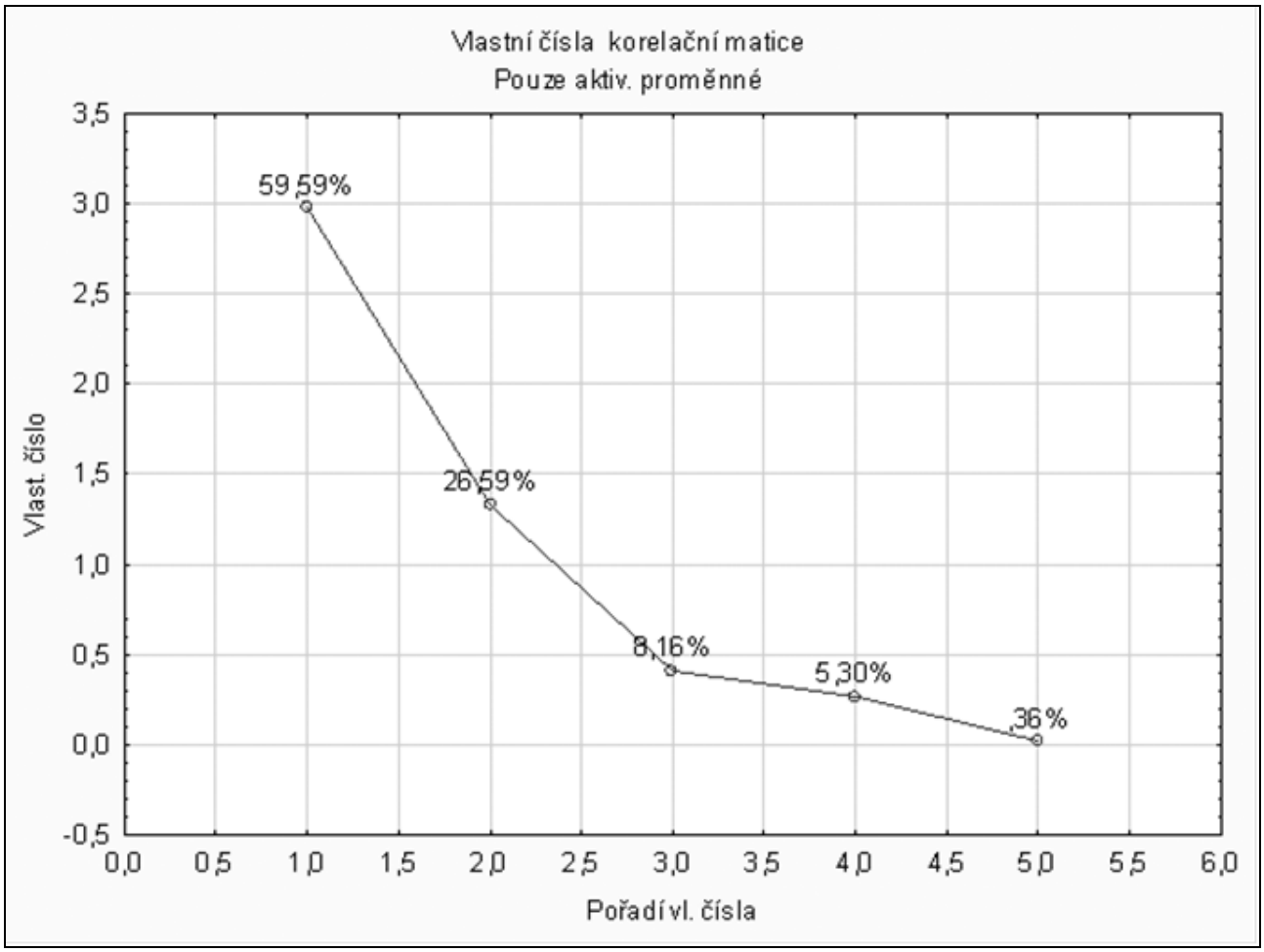
V matici (tab. 36) vidíme rovněž těsnou závislost mezi veličinami X_1 a X_2. Vysoké hodnoty jednoduchého i parciálního korelačního koeficientu mezi počtem veličin členů domácnosti a počtem dětí signalizují první redukci počtu proměnných, tj. vynechání veličiny X_2.

3. K redukci výchozího počtu původních proměnných (X_1, X_2, X_3, X_4, X_5) užijeme metodu PCA, tj. analýzu hlavních komponent - viz tab. 37.

Statistiky → Vícerozměrné průzkumné techniky → Hlavní komponenty a klasifikační analýza → Vlastní čísla

Tab. 37 Metoda PCA

Pořadí vl.č.	Vlastní čísla korelační matice a související statistiky (domacnosti) Pouze aktiv. proměnné					
	vl. číslo	% celk. rozptylu	Kumulativ. vl. číslo	Kumulativ. %		
1	2,980	59,591	2,980	59,591		
2	1,330	26,594	4,309	86,185		
3	0,408	8,162	4,717	94,348		
4	0,265	5,296	4,982	99,643		
5	0,018	0,357	5,000	100,000		



O

Obr. 11 Scree graf

Z tabulky 37 čteme, že první komponenta vysvětluje 60 % celkové variability, druhá komponenta vysvětluje téměř 27 % celkové variability. Protože podíl zbývajících tří komponent na celkové variabilitě je poměrně malý, můžeme konstatovat, že pro vysvětlení původních proměnných budou stačit dvě komponenty, které vysvětlují více než 86 % celkové variability

Tabulku 37 doplňuje scree graf - viz obr. 11.

4. Další metodou redukce proměnných je faktorová analýza. Maximální počet faktorů nastavíme na hodnotu 2 metodou PCA (viz předchozí výpočet)

Statistiky → Vícerozměrné průzkumné techniky → Faktorová analýza

Po zadání počtu společných faktorů (2), dostáváme tabulku 38, ve které jsou odhady faktorových zátěží.

Tab. 38 Faktorové zátěže

Proměnná	Faktor. zátěže (Bez rot.) (domacnosti) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000) **Faktor 1**	Faktor 2
Počet členů X1	-0,966651	0,164130
Počet dětí X2	-0,884202	0,376960
Průměr. věk X3	-0,004101	-0,958157
Příjem X4	-0,820112	-0,157796
Vydání X5	-0,768588	-0,466603
Výkl.roz	2,979555	1,329719
Prp.celk	0,595911	0,265944

Výsledné řešení vykazovalo velmi dobrou interpretovatelnost, zde není potřeba provést rotaci faktorové matice. Přesto v jiných příkladech to bude téměř nezbytné, zkusmo provedeme rotaci metodou varimax a dostáváme tabulku 39.

Tab. 39 Faktorová rotace

Proměnná	Faktor. zátěže (Varimax pr.) (domacnosti) Extrakce: Hlavní komponenty (Označené zatěže jsou >,690000) **Faktor 1**	Faktor 2
Počet členů X1	0,979918	-0,033381
Počet dětí X2	0,926672	-0,255325
Průměr. věk X3	-0,124075	0,950098
Příjem X4	0,791642	0,266056
Vydání X5	0,699283	0,565199
Výkl.roz	2,950048	1,359226
Prp.celk	0,590010	0,271845

V tabulce 39 vidíme, že první faktor je silně korelován s proměnnými X_1, X_2, X_4 a X_5, druhý faktor je korelován s proměnnou X_3.

Závěr:

Metodou faktorové analýzy jsme určili dva hlavní faktory (latentní proměnné)

$$F_1 = 0{,}9979918\ X_1 + 0{,}926672\ X_2 - 0{,}124075\ X_3 + 0{,}791642\ X_4 + 0{,}699283\ X_5,$$

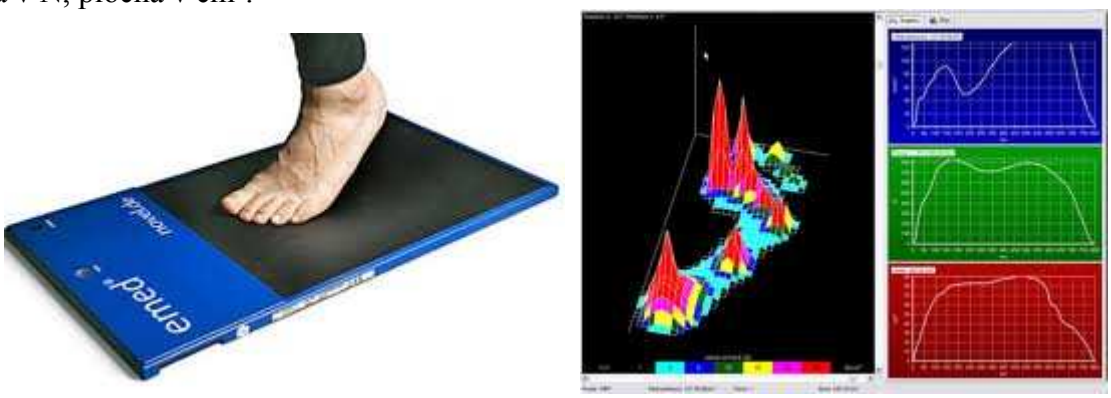
$$F_2 = -0{,}033381\ X_1 - 0{,}255325\ X_2 + 0{,}950098\ X_3 + 0{,}66056\ X_4 + 0{,}565199\ X_5.$$

Protože první faktor významně koreluje s proměnnými X_1 (počet členů domácnosti), X_2 (počet dětí), X_4 (příjmy) a X_5 (výdaje), možno jej označit jako *faktor sociální úrovně*, druhý faktor (méně významný) můžeme označit jako *faktor věkové struktury domácnosti*.

Příklad 2 Redukce proměnných

Pomocí plantografie, metody, kdy za pomocí tlakové plošiny (obr. 12) je snímáno rozložení tlaku nohy, obvykle při chůzi, je možné získat statickou a dynamickou charakteristiku rozložení tlaku. Při výzkumu rozložení tlaku chůze jsme sledovali celkem 28 parametrů u 55 respondentů a to zejména silových, časových a prostorových. Plošina EMED i dodaný software je od firmy Novel. (novel.de/novelcontent/emed). Pro redukci počtu proměnných byl použit sw Statistica 10 firmy Statsoft (www.statsoft.cz). Následují názvy proměnných a ilustrační obrázky.

Legenda: A-aktivní část kroku, P-pasivní část kroku, síla (aktuální, celková, maximální) měřena v N, plocha v cm^2.



Obr. 12 Tlaková deska EMED a graf rozložení tlaku

Tab. 40 Sledované parametry

Síla-A část[N]-levá	Síla-Fmax P-levá	čas-aktivní-levá
Síla-A část[N]-pravá	Síla-Fmax P-pravá	čas-aktivní-pravá
Síla-P část (N)-levá	Síla-F a (N)-levá	čas-celková délka-levá
Síla-P část (N)-pravá	Síla-F a (N)-pravá	čas-celková délka-pravá
Síla-celk.F(N)-levá	Plocha-A m (cm2)-levá	čas-poměr A/P-levá
Síla-celk.F(N)-pravá	Plocha-A m (cm2)-pravá	čas-poměr A/P-pravá
Síla-Max síla (N)-levá	Plocha-A a (cm2)-levá	čas-% A části-levá
Síla-Max síla (N)-pravá	Plocha-A a (cm2)-pravá	čas-% A části-pravá
Síla-Fmax A-levá	čas-pasivní-levá	
Síla-Fmax A-pravá	čas-pasivní-pravá	

Pracovat s tímto počtem proměnných je složité, již na první pohled je zřetelné, že lze předpokládat, že mnohé spolu budu korelovat, tudíž by stálo za úvahu vybrat jen typické reprezentanty z množiny původních proměnných.

Metodou PCA vypočítáme vlastní čísla korelační matice.

Tab. 41 Výpočet vlastních čísel

	vlastní číslo	**% celk. rozptylu**	**Kumulat. % celk. rozptylu**			**vlastní číslo**	**% celk. rozptylu**	**Kumulat. % celk. rozptylu**
1	13,663	48,796	48,796		**15**	0,015	0,055	99,891
2	6,646	23,736	72,531		**16**	0,011	0,040	99,931
3	2,950	10,536	83,068		**17**	0,010	0,036	99,966
4	1,581	5,647	88,715		**18**	0,003	0,012	99,978
5	0,852	3,043	91,758		**19**	0,002	0,007	99,985
6	0,698	2,492	94,250		**20**	0,001	0,005	99,990
7	0,635	2,269	96,518		**21**	0,001	0,003	99,994

8	0,362	1,295	97,813		**22**	0,001	0,003	99,996
9	0,276	0,986	98,799		**23**	0,001	0,002	99,998
10	0,099	0,355	99,154		**24**	0,000	0,001	100,000
11	0,088	0,315	99,469		**25**	0,000	0,000	100,000
12	0,048	0,172	99,641		**26**	0,000	0,000	100,000
13	0,036	0,130	99,771		**27**	0,000	0,000	100,000
14	0,019	0,066	99,837					

Na základě tab. 41 stanovíme počet faktorů na 4. Tyto 4 faktory vysvětlují 88,7 % celkové variability a s tím můžeme být spokojeni. Poté provedeme rotaci Varimax, která volí transformační matici takovou, aby součet rozptylů druhých mocnin faktorových zátěží v jednotlivých sloupcích byl co největší. Výsledkem je tab. 42, kde jsou zvýrazněny proměnné s faktorovou zátěží vyšší než 0,7.

Tab. 42 Matice faktorových zátěží po rotaci Varimax

	F1	**F2**	**F3**	**F4**
Síla- A část[N]-levá	**0,907**	0,250	0,042	-0,037
Síla- A část[N]-pravá	**0,936**	0,213	0,181	-0,055
Síla-P část (N)-levá	**0,903**	-0,100	-0,140	0,033
Síla-P část (N)-pravá	**0,880**	-0,174	-0,247	-0,022
Síla-celk.F(N)-levá	**0,974**	0,157	0,028	-0,035
Síla-celk.F(N)-pravá	**0,981**	0,123	0,046	-0,059
Séla-Max síla (N)-levá	**0,904**	0,223	-0,026	-0,134
Séla-Max síla (N)-pravá	**0,922**	0,167	0,009	-0,114
Síla-Fmax A-levá	**0,898**	0,231	-0,007	-0,133
Síla-Fmax A-pravá	**0,916**	0,176	0,047	-0,116
Séla-Fmax P-levá	**0,862**	-0,109	-0,184	-0,006
Séla-Fmax P-pravá	**0,883**	-0,146	-0,283	-0,065
Séla-F a (N)-levá	**0,908**	0,086	0,175	0,135
Séla-F a (N)-pravá	**0,861**	0,134	0,344	0,041
Plocha-A m (cm2)-levá	-0,156	0,035	0,105	**0,964**
Plocha-A m (cm2)-pravá	-0,018	0,025	0,194	**0,972**
Plocha-A a (cm2)-levá	**0,740**	0,150	0,329	0,103
Plocha-A a (cm2)-pravá	0,672	0,189	0,463	0,181
čas-pasivní-levá	-0,067	**-0,829**	0,475	-0,015
čas-pasivní-pravá	-0,284	**-0,874**	0,148	0,106
čas-aktivní-levá	-0,033	0,683	0,617	0,116
čas-aktivní-pravá	0,182	**0,700**	0,622	0,099
čas-celková délka-levá	-0,055	0,184	**0,909**	0,122
čas-celková délka-pravá	0,002	0,188	**0,921**	0,214
čas-poměr A/P-levá	-0,017	**0,917**	0,265	0,049
čas-poměr A/P-pravá	0,190	**0,855**	0,357	-0,060
čas-% A části-levá	0,046	**0,921**	0,174	0,094
čas-% A části-pravá	0,258	**0,856**	0,299	0,013

Výsledkem faktorové analýzy je výběr zástupců (pokud to výzkumníkovi dává smysl, nikoliv automaticky) jen jedné proměnné z každého faktoru. Asi nejtěžší část faktorové analýzy je pojmenování nových faktorů. V našem případě se jedná o faktor síly (F1), faktor časových charakteristik (F2), faktor časo-délkových vlastností (F3) a faktor plošných vlastností (F4). Pokud však výzkumník nedokáže smysluplně pojmenovat faktory a najít pro ně interpretaci v souvislosti původními daty, nelze faktorovou analýzu a její výsledky použít.

Příklad 3 Konfirmační faktorová analýza

Data pocházejí z dotazníku „Inventory of Learning Styles“ u vysokoškolských studentů (Vermunt & Rijswijk, 1988). Zajímá nás, zda výsledky sledované skupiny budou vymezovat stejné rysy učení jako u originálních dat. Data a výsledky jsou součástí dizertační práce (Sebera, 2009).

Tab. 43 Popis proměnných a vstupní data

p1	Hledání vztahů a strukturování	p11	Získání diplomu
p2	Kritická aktivita, nezávislost	p12	Profesní motivace
p3	Memorování a vybavování	p13	Testování sebe sama, svých možností
p4	Analyzování	p14	Osobní zájmy, záliby
p5	Konkretizování a dodávání osobnostního smyslu	p15	Ambivalentní motivace
p6	Autoregulace průběhu a výsledku učení	p16	Absorbování znalostí
p7	Autoregulace obsahové stránky učení	p17	Konstruování znalostních struktur
p8	Vnější regulace průběhu učení	p18	Používání znalostí
p9	Vnější regulace výsledků učení	p19	Stimulování sebevzdělávání
p10	Absence řízení vedoucí k problémům	p20	Kooperování

č.	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	p16	p17	p18	p19	p20
1	26	15	15	14	23	16	15	16	13	15	11	23	7	20	11	27	38	26	27	24
2	25	12	12	16	16	15	7	10	12	8	11	22	17	21	5	24	40	26	34	20
3	20	14	16	13	14	16	11	17	17	14	11	19	17	21	12	35	32	24	27	27
4	27	9	17	21	17	14	14	16	15	14	14	23	11	23	7	32	39	28	32	21
5	18	7	16	14	14	17	8	10	11	11	12	23	15	19	7	29	33	22	29	21
6	25	7	20	18	12	17	15	13	20	14	15	21	19	17	12	35	33	23	29	24
7	18	6	20	8	15	16	7	17	12	11	12	21	10	22	8	37	40	28	37	22
8	22	8	11	13	14	15	9	16	15	11	14	21	10	19	6	28	37	24	29	30
9	27	12	15	9	21	17	11	16	14	13	11	22	21	23	10	23	36	28	27	22
10	29	10	17	18	18	21	10	19	19	12	13	19	16	21	8	27	38	25	29	27
11	15	10	11	16	10	13	8	10	12	15	17	20	12	14	14	26	27	17	24	24
12	13	9	10	9	13	8	5	16	9	20	16	15	11	16	11	25	30	26	28	12
13	18	8	16	13	13	16	5	14	11	18	18	19	10	16	15	37	31	25	21	21
14	21	14	13	14	16	23	11	18	15	14	14	13	13	16	13	29	31	18	25	30
15	21	13	13	14	16	20	11	16	15	15	13	16	13	16	14	29	31	18	25	30
16	18	9	9	11	14	14	6	15	20	15	16	19	18	21	8	28	34	23	29	35
17	17	6	15	10	11	13	8	18	19	12	16	23	13	20	8	33	35	26	29	30
18	20	8	16	16	15	15	13	14	18	11	14	21	19	21	9	33	33	22	29	24
19	23	7	15	14	15	15	15	11	12	11	19	23	18	24	9	31	39	27	32	31
20	22	6	12	14	14	12	10	13	13	12	9	19	13	21	8	32	34	25	29	31
21	22	11	13	20	18	16	12	16	16	11	12	15	12	19	11	30	37	24	31	32
22	28	11	19	19	19	24	16	21	16	13	14	23	16	21	10	31	35	23	30	27
23	17	6	15	12	12	18	6	12	11	10	13	14	15	11	13	27	25	20	26	19
24	19	11	16	16	15	22	10	15	13	16	16	19	14	19	15	32	31	27	36	28
25	22	5	18	16	16	15	13	12	16	15	14	23	20	17	13	25	36	27	32	27
26	29	9	22	14	13	18	14	12	18	10	13	19	20	18	11	24	36	28	31	24
27	23	13	15	17	18	19	10	18	16	17	11	19	16	19	11	33	34	26	27	24
28	21	4	11	10	14	15	4	10	8	14	10	21	10	13	11	26	29	18	23	18
29	21	4	11	10	14	15	4	10	8	14	11	21	10	15	11	26	28	17	23	19
30	29	9	12	17	19	22	8	14	17	8	16	25	25	24	6	34	40	29	28	22
31	31	9	13	17	17	21	8	14	17	8	15	25	25	24	6	33	39	29	27	22
32	22	8	13	16	17	24	8	16	15	14	15	20	13	19	9	34	30	21	33	28
33	28	10	13	14	16	23	9	11	16	16	14	21	14	19	13	21	35	19	30	25
34	17	7	19	16	14	17	5	22	18	13	15	19	17	19	11	36	33	22	28	19
35	19	9	14	15	13	15	8	18	16	13	18	21	18	17	17	35	35	25	30	32
36	24	11	10	13	16	20	16	11	15	14	10	20	15	17	15	22	40	25	33	25

č.	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	p16	p17	p18	p19	p20
37	18	5	15	18	14	14	5	15	13	16	15	23	21	17	9	37	33	28	22	28
38	22	6	18	15	15	13	9	12	12	17	17	18	18	16	15	36	37	24	34	28
39	21	5	18	11	9	13	10	11	15	9	14	22	10	18	11	30	33	21	28	21
40	22	8	17	17	9	16	5	17	19	15	20	20	23	22	12	39	36	28	34	21
41	13	6	8	12	14	14	9	14	16	9	22	15	12	11	12	34	30	15	24	31
42	25	10	19	21	18	23	13	17	12	10	16	24	25	22	9	30	23	21	23	24
43	23	9	14	16	20	18	7	19	20	11	13	24	12	19	8	31	34	26	32	23
44	20	8	13	16	13	9	8	16	16	15	16	16	18	16	13	35	30	23	25	23
45	15	8	17	17	15	9	10	13	14	12	9	15	7	21	6	26	27	23	16	25
46	20	6	11	13	17	16	8	14	13	24	14	24	18	20	10	38	34	27	33	25
47	28	8	12	20	19	18	15	13	20	22	18	20	21	21	15	36	38	29	37	36
48	20	7	15	19	12	12	9	14	20	14	15	20	16	21	14	33	33	25	25	24
49	14	6	8	11	7	10	9	16	7	15	17	21	18	16	13	38	35	25	33	31
50	19	8	16	18	22	17	17	19	16	25	15	18	18	21	14	36	32	25	26	31
51	27	9	13	17	18	22	10	15	12	11	12	20	9	19	15	26	37	24	30	23
52	27	7	18	16	14	15	11	9	16	19	13	19	21	19	5	27	35	24	35	22
53	22	9	19	17	16	17	7	16	13	14	11	21	14	17	12	31	32	23	29	20
54	27	9	17	17	17	19	9	15	14	22	15	22	11	19	12	31	39	25	33	17
55	24	11	12	14	17	22	7	17	12	13	8	17	12	19	7	23	41	25	34	23
56	24	8	20	19	14	18	9	14	18	15	15	20	14	20	12	29	33	23	29	29
57	9	5	9	8	10	10	6	9	10	7	13	20	21	23	6	27	33	24	26	14
58	25	10	20	20	18	26	15	20	21	17	18	21	21	20	12	33	34	24	25	23
59	31	9	14	17	20	14	13	13	12	23	13	23	17	22	8	29	36	26	28	30
60	27	9	17	23	19	12	14	17	15	14	13	23	11	23	7	32	39	28	32	21
61	25	14	10	9	17	15	5	15	12	17	16	18	18	19	10	28	33	27	30	27
62	27	10	17	16	20	25	5	18	15	19	17	17	14	18	13	31	35	25	33	27
63	16	7	15	12	13	17	7	14	16	18	16	18	16	19	15	34	36	24	28	34
64	19	10	18	16	21	18	17	10	14	15	17	24	16	17	11	41	38	30	35	33
65	21	10	18	14	10	19	6	12	19	15	12	20	19	20	8	33	32	22	31	24
66	18	7	12	13	12	14	5	15	15	13	12	19	15	19	12	26	33	22	27	26
67	25	10	18	20	12	19	10	16	18	12	13	19	18	21	9	29	36	22	28	24
68	20	7	13	14	16	14	14	10	16	11	14	21	11	19	6	30	36	22	30	28
69	31	8	10	9	13	19	10	9	11	12	10	17	9	20	7	18	34	21	26	21
70	10	6	17	14	17	14	9	15	14	14	10	22	12	15	8	34	32	25	26	24
71	10	7	17	14	17	14	9	15	14	14	10	22	12	15	8	34	32	25	26	24
72	24	8	11	15	16	16	12	14	18	9	13	21	19	21	12	29	31	25	27	24
73	16	9	16	15	13	19	9	13	11	15	13	23	20	22	10	25	37	28	34	32
74	18	6	17	22	18	16	11	17	17	13	9	24	16	23	5	33	37	26	34	27
75	19	15	16	13	17	11	18	10	11	16	13	25	22	22	5	31	36	29	36	13
76	21	14	13	14	16	22	11	17	15	14	14	16	13	16	13	29	31	18	25	30
77	22	5	16	17	19	16	10	17	18	14	14	21	11	23	12	28	34	25	26	22
78	23	5	15	14	15	16	15	11	12	9	14	23	16	24	9	32	39	30	32	33
79	22	7	18	23	14	15	7	14	13	22	17	20	19	14	14	36	30	21	24	27
80	24	9	15	16	19	19	12	19	17	11	16	18	21	19	16	30	37	25	30	29
81	22	10	13	9	12	15	9	14	16	23	11	18	13	20	14	31	28	24	25	25
82	22	10	13	9	12	16	9	13	16	22	13	18	13	20	14	31	28	24	25	25
83	28	13	19	19	19	22	16	20	16	13	15	23	16	21	10	31	35	23	30	27
84	17	6	15	11	11	17	6	12	11	10	13	14	15	11	13	30	26	20	26	20
85	11	5	16	10	8	11	4	15	15	23	16	17	11	16	18	39	26	24	21	19
86	27	10	19	21	16	19	15	18	19	14	14	20	20	21	12	30	37	24	30	28
87	22	10	10	19	21	19	8	17	21	16	14	21	9	16	18	29	38	29	35	19
88	26	14	20	23	21	22	13	15	21	18	17	25	20	25	9	34	37	25	33	23
89	19	9	18	14	13	20	9	15	16	12	13	22	20	21	5	34	40	25	33	35
90	19	10	20	14	13	21	9	17	16	12	14	24	23	23	5	34	40	25	33	35
91	24	9	21	18	16	18	8	16	20	12	15	18	20	21	13	41	41	29	34	29
92	23	11	19	15	17	20	11	19	17	22	17	23	19	20	14	39	38	27	34	32

č.	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	p16	p17	p18	p19	p20
93	22	7	15	17	15	16	10	14	14	11	13	20	10	21	15	29	33	24	28	25
94	25	8	16	16	18	17	9	15	13	21	11	19	20	16	15	30	30	25	28	21
95	26	8	16	16	18	16	9	15	13	24	11	19	20	16	16	30	30	25	28	21
96	17	5	18	9	11	12	5	13	14	16	14	22	19	21	6	29	26	29	30	30

Tab. 44 Analýza hlavních komponent

	vlastní číslo	**% celkového rozptylu**	**Kumulativní % rozptylu**
1	4,96	24,82	24,82
2	2,54	12,71	37,53
3	2,24	11,22	48,75
4	1,33	6,65	55,40
5	1,24	6,19	61,59
6	1,03	5,15	66,74

K zlepšení interpretovatelnosti faktorů jsme provedli rotaci faktorů metodou Varimax. Vybrané faktorové zátěže proměnných ILS u jednotlivých faktorů obsahuje tabulka (tab. 45). Pro přehlednost jsou znázorněny jen zátěže větší než ± 0,5.

Statistiky → Vícerozměrné průzkumné techniky → Faktorová analýza

Tab. 45 Faktorové zátěže proměnných a faktorů (po rotaci)

		F1	**F2**	**F3**	**F4**	**F5**	**F6**
zpracování učiva	Hledání vztahů a strukturování			0,79			
	Kritická aktivita, nezávislost			0,66			
	Memorování a vybavování						0,72
	Analyzování						0,57
	Konkretizování a dodávání osobnostního smyslu			0,63			
řízení učení	Autoregulace průběhu a výsledku učení			0,72			
	Autoregulace obsahové stránky učení			0,56			
	Vnější regulace průběhu učení		0,87				
	Vnější regulace výsledků učení		0,53				
	Absence řízení vedoucí k problémům				0,85		
studijní motivace	Získání diplomu					0,76	
	Profesní motivace	0,61					
	Testování sebe sama, svých možností					0,51	0,50
	Osobní zájmy, záliby	0,62					
	Ambivalentní motivace				0,69		
přístupy ke studiu	Absorbování znalostí						
	Vytváření, konstruování znalostních struktur	0,78					
	Používání znalostí	0,85					
	Stimulování sebevzdělávání	0,74					
	Kooperování					0,65	

V prvním faktoru jsou nejvyšší náboje u proměnných *Používání znalostí, Vytváření, konstruování znalostních struktur* a *Stimulování sebevzdělávání.* U druhého faktoru je nejvyšší náboj u *Vnější regulace průběhu učení*. Výrazně nižší, ale vzájemně blízká je *Vnější regulace výsledků učení*. Třetí komponenta nese vysoké náboje u *Hledání vztahů a strukturování* a u *Autoregulace průběhu a výsledku učení*. Čtvrtá komponenta má výrazně vyšší náboj u *Absence řízení vedoucí k problémům*, pátá u *Získání diplomu* a šestá nejvíce u *Memorování a vybavování*. Charakter komponenty s rovnoměrně rozloženým nábojem má nejvíce první komponenta, do určité

míry i druhá. Výsledky faktorové analýzy jsme srovnali s výsledky studie Vermunta & Rijswijka (1988). Konstatujeme značnou podobnost ve vytvořených faktorech.

Srovnání výsledků faktorové analýzy

Srovnávání výsledků dosažených pomocí faktorové analýzy je pouze orientační, protože neexistuje metodika pro srovnávání jednotlivých faktorů a faktorových zátěží. Proto porovnáváme a interpretujeme jen základní podobnosti, jakými jsou počty nalezených faktorů, přítomnost proměnných v jednotlivých faktorech a velikost skórů (tab. 46).

Tab. 46 Srovnání výsledků faktorové analýzy

	F1	F2	F3	F4	F5	F6	V1	V2	V3	V4
Hledání vztahů a strukturování			0,79				0,72			
Kritická aktivita, nezávislost			0,66				0,70			
Memorování a vybavování						0,72		0,73		
Analyzování						0,57		0,76		
Konkretizování a dodávání osobn. smyslu			0,63				0,65			
Autoregulace průběhu a výsledku učení			0,72				0,74			
Autoregulace obsahové stránky učení			0,56				0,72			
Vnější regulace průběhu učení		0,87						0,73		
Vnější regulace výsledků učení		0,53						0,54		
Absence řízení vedoucí k problémům				0,85					0,74	
Získání diplomu					0,76					
Profesní motivace	0,61									-0,80
Testování sebe sama, svých možností					0,51	0,50				
Osobní zájmy, záliby	0,62						0,54			
Ambivalentní motivace				0,69					0,65	
Absorbování znalostí								0,54		
Vytváření, konstruování znal. struktur	0,78						0,75			
Používání znalostí	0,85									-0,74
Stimulování sebevzdělávání	0,74								0,73	
Kooperování					0,65				0,61	

Legenda:

- F1-6 – faktorové zátěže (vyšší než 0,5) zjištěné analýzou hlavních komponent z dat ILS v experimentální a kontrolní skupině
- V1-4 – faktorové zátěže na komponentách zjištěné analýzou hlavních komponent z práce Vermunt 1998, ILS se 120 položkami administrován studentům práva, literatury, ekonomie a bankovnictví

Význam provedené faktorové analýzy je doplňkový. Je potřeba brát v úvahu malý počet respondentů v této studii (faktorová analýza byla provedena na všech vstupních datech adaptovaného ILS z obou skupin).

Struktura nalezených faktorů a zařazení jednotlivých proměnných do faktorů se jeví na první pohled jako zcela odlišná. Nicméně společné zákonitosti zde pozorujeme:

- Faktor F3 jednoznačně identifikuje svým složením, včetně hodnot faktorových skóru faktor V1, kterým Vermunt & Rijswijk (1988) definuje styl učení orientovaný **na smysl**.

- Faktor F4 se podobá původnímu faktoru V3, který definuje **nezaměřený** styl.
- Původní faktor V2 (definující **reprodukční** styl) se v naší struktuře nevyskytuje. Stačí však provést spojení našich faktorů F2 a F6 a přiblížíme se faktoru V2. Takovéto spojení je možné provést jen „nezávazně“, neboť z podstaty konstrukce faktorů v analýze hlavních komponent je zřejmé, že námi spojované faktory F2 a F6 jsou nezávislé.
- Původní faktor V4, který je přítomností proměnných *Profesní motivace* a *Používání znalostí* identifikován jako aplikační studijní styl, nacházíme obsažen v našem faktoru F1. Zde jsou sdruženy proměnné z oblasti „studijní motivace“ a „přístupy ke studiu“. Faktor je sycen i jinými proměnnými než V4, nicméně lze v něm najít jasnou souvislost s **aplikačním** studijním stylem.

Je nutné opět připomenout, že srovnání našich faktorů s prací Vermunta & Rijswijka (1988) slouží jen pro orientaci, nicméně i tak zde můžeme identifikovat a popsat všechny 4 definované styly učení – aplikační, reprodukční, nezaměřený a styl orientovaný na smysl.

Shluková analýza

Shluková analýza (cluster analysis) seskupuje, shlukuje data do společných skupin a to na základě podobnosti (ne podobnosti, vzdálenosti). O datech toho většinou víme velmi málo. Data jsou reprezentována svými *k* charakteristikami, dostáváme k-rozměrný vektor. Výsledkem shlukové analýzy je vytvoření *dendogramu* (hierarchický strom shluků), kde platí, že podobné případy budou ve stejném nebo blízkém shluku a rozdílné případy (a shluky do kterých padnou) budou od sebe vzdáleny.

Standardizace dat

Vzdálenost samotná je závislá na měřítkách jednotlivých veličin. V případě nesourodosti (statisíce vs. jednotky) je možné data standardizovat, protože jinak by celá analýza závisela nejvíce na proměnné s největším rozsahem. Neexistuje však pravidlo, zda standardizaci použít nebo ne.

Vzdálenost objektů

Postup pří shlukové analýze probíhá v zásadě ve dvou krocích. V prvním kroku se vypočtou vzdálenosti objektů (proměnných nebo případů) a uloží se do matice vzdáleností. Ve druhém kroku se na základě této matice objekty postupně slučují do shluků. Shluky nahradí sloučené objekty a podrobují se novému vypočtu vzdáleností podle stejných principů.
V statistických software je možné zvolit z několika různých způsobů výpočtu vzdálenosti:

- Euklidovké vzdálenosti - $d(x,y) = \{\Sigma_i (x_i - y_i)\}^{1/2}$, - klasická míra vzdálenosti, která pro dva body v prostoru určuje délku „nejkratší cesty“ z jednoho bodu do druhého
- Blokové vzdálenosti (Manhattan) - $d(x,y) = \Sigma_i |x_i - y_i|$ - suma vzdáleností v jednotlivých dimenzích. Název i výpočet je inspirován vzdáleností, kterou na Manhattanu člověk urazí při cestě z jednoho bodu do druhého. Nelze jít po spojnici, musí se jít po kolmých ulicích
- Čebyševovy vzdálenosti - $d(x,y) = \text{Max} |x_i - y_i|$ - maximum ze vzdáleností v jednotlivých dimenzích
- Mocninné vzdálenosti - $d(x,y) = (\Sigma_i |x_i - y_i|^p)^{1/r}$ - uživatelem definovaná míra vzdáleností. Čím vyšší parametr p, tím vyšší váha se přikládá větším vzdálenostem v jednotlivých dimenzích a snižuje se význam malých vzdáleností. Vysoké p nejvíce „propaguje“ body hodně vzdálené ve všech dimenzích. Parametr r působí opačným směrem, čím vyšší r, tím menší váha se přikládá větším vzdálenostem. r ovšem působí celkově bez ohledu na dimenze
- Procentuální neshoda - $d(x,y) = (\text{počet } x_i \neq y_i) / i$ - je vhodná pouze pro kategorické proměnné. Pro dva objekty se spočte jako podíl počtu dimenzí, v nichž se jejich hodnota liší, ku celkovému počtu dimenzí
- 1- Pearsonův r - $d(x,y) = 1 - r(x,y)$ - míra založená na korelaci. Největší vzdálenost přiřazuje negativně korelovaným objektům, nejmenší naopak pozitivně korelovaným objektům. Nevhodná pro malý počet dimenzí.

Kterou míru vzdálenosti vybrat? Procentuální neshoda je určena pro kategorické proměnné. Pokud neprovedeme standardizaci dat, pak některé proměnné mají větší rozptyl jiné menší. Pokud bychom zvolili Čebyševovu míru vzdálenosti, bude o zařazení do clusterů rozhodovat právě proměnná s největším rozptylem a vliv ostatních proměnných bude zanedbatelný. Všechny ostatní míry jsou přijatelné a lze je postupně vyzkoušet. Nejběžnější je použití Euklidovské vzdálenosti.

POZOR. Obecně ale nemusí být proměnná s největším rozptylem skutečně ta, která nejvíce odlišuje objekty!

Pravidla slučování

V matici vzdáleností se nalezne minimum a objekty, jimž tato vzdálenost přísluší se spojí do shluku. Dojde k výpočtu nové matice vzdáleností. Celý cyklus se opakuje až do vytvoření jediného velkého shluku.

I v pravidlech pro spojování je možné vybrat několik možností.

- Jednoduché spojení - vzdálenost dvou shluků se určí jako vzdálenost dvou nejbližších objektů (případů/proměnných). Rozumí se dvou nejbližších objektů z různých shluků! Tento algoritmus má tendenci spojovat objekty do dlouhých „řetízků“
- Úplné spojení - vzdálenost shluků je naopak dána vzdáleností těch dvou objektů, které jsou nejdále od sebe. Algoritmus je vhodný pro případy, kdy jsou objekty přirozeně rozdělené do určitých skupin. Má tendenci spíše tvořit skupiny s podobným počtem objektů
- Nevážený průměr skupin dvojic - vzdálenost shluků je prostým průměrem vzdáleností všech párů objektů, které lze vytvořit tak, že z každého shluku vezmeme jeden objekt. Tato varianta algoritmu pracuje lépe v případech, kdy vstupní objekty mají spíš charakter oddělených skupin. Lze ale použít i pro objekty mající „řetízkovou“ strukturu
- Vážený průměr skupin dvojic - obdoba předchozího algoritmu. Při výpočtu průměru se navíc berou jako váhy počty objektů v jednotlivých clusterech
- Nevážený centroid skupin dvojic - vzdálenost shluků je určí jako vzdálenost mezi centroidy shluků. (Centroid je bodem definovaným průměry v jednotlivých dimenzích)
- Vážený centroid skupin dvojic (medián) - vážená varianta předchozího algoritmu

Využití shlukové analýzy

Použitím jiného způsobu výpočtu vzdálenosti objektů a pravidel slučování lze dojít k různým hierarchickým stromům. Každý takový výsledek může být způsoben jinou vlastností původních dat. Výsledek shluková analýzy musí být potvrzen i jiným úsudkem a znalostí vědeckého pracovníka, bez ní (ostatně jako v celé statistice) se jedná pouze o „hru čísel“.

Příklad 1 Shluková analýza

Zadání: Přirozený pohyb obyvatelstva v České republice v letech 1985 - 1995 je uveden v tabulce 47. Vytvořte shluky poměrně stejnorodých ročníků.

Tab. 47 Vstupní data

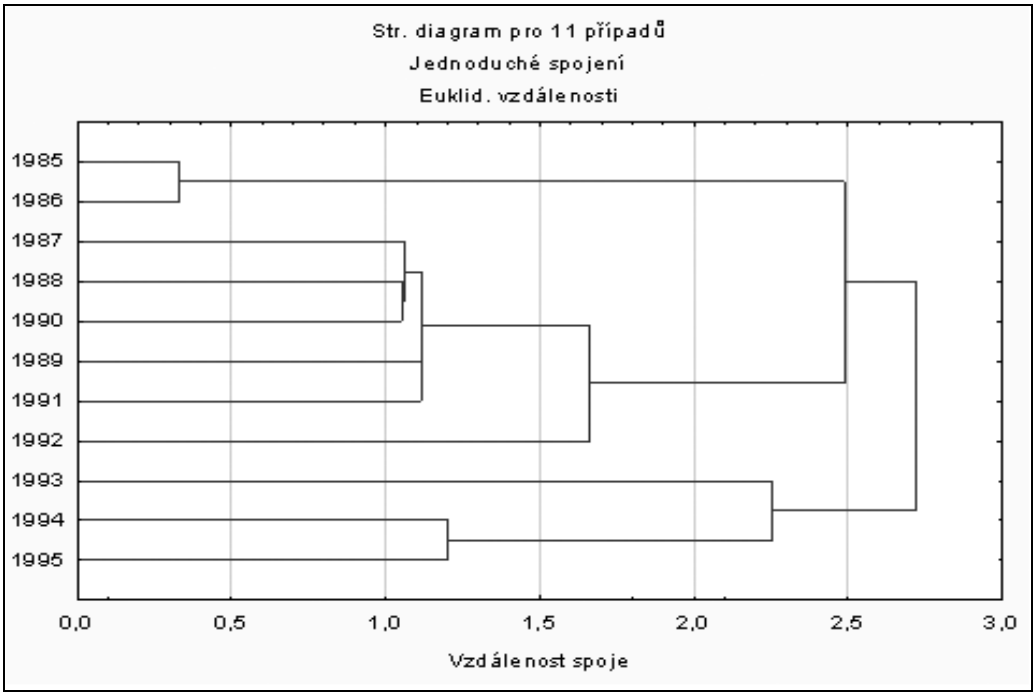
Rok	Sňatky na 1000 obyv. X_1	Rozvody na 1000 obyv. X_2	Živě naroz. na 1000 obyv. X_3	Zemřelí na 1000 obyv. X_4	Kojenecká úmrtnost v ‰ X_5	Potraty na 1000 obyv. X_6
1985	7,8	2,95	13,1	12,7	12,5	9,6
1986	7,9	2,86	12,9	12,8	12,3	9,6
1987	8,1	3,00	12,7	12,3	12,0	12,0
1988	7,9	2,96	12,8	12,1	11,0	12,2
1989	7,8	3,03	12,4	12,3	10,0	12,0
1990	8,8	3,09	12,6	12,5	10,8	12,0
1991	7,0	2,85	12,5	12,1	10,4	11,4
1992	6,2	2,77	11,8	11,7	9,9	10,3
1993	6,4	2,93	11,7	11,4	8,5	8,0
1994	5,7	2,99	10,3	11,4	7,9	6,5
1995	5,3	3,01	9,3	11,4	7,7	6,0

Statistiky → Vícerozměrné průzkumné statistiky → Shluková analýza

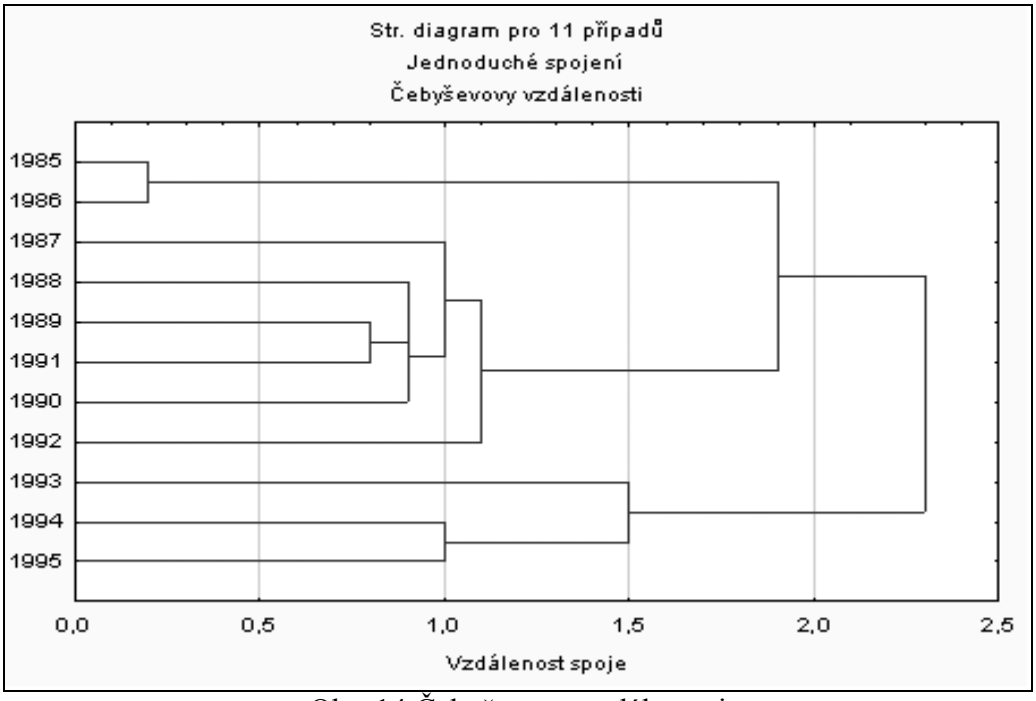
Řešení:

Shluková analýza je metodou vyhledávání homogenních skupin v rámci souboru nějakých statistických jednotek, které bývají v této souvislosti označovány jako objekty. Shluková analýza rozkládá toto „mračno“ do jistého počtu disjunktních podmnožin - shluků. Pravidla tohoto shlukování bývají konstruována tak, aby shlukování vedlo k vytvoření shluků, v jejichž rámcích jsou jednotky co nejvíce homogenní, zatímco jednotky z různých shluků se ve svých vlastnostech co nejvíce odlišují. Shluková analýza je spíše než metodou, rozsáhlým komplexem metod typologického třídění, které můžeme klasifikovat z ně-kolika různých pohledů. Podle typu výpočetního algoritmu dělíme metody shlukové analýzy na hierarchické, paralelní a sekvenční, přičemž hierarchické metody můžeme dále podrobněji klasifikovat na aglomerativní a divizní. Využijeme výhradně metody z hierarchickým aglomerativním algoritmem.

Při tvorbě shluků využijeme nabídky programu STATISTICA 10, který pro měření vzdálenosti objektů využívá několik způsobů výpočtu vzdálenosti



Obr. 13 Euklidovké vzdálenosti



Obr. 14 Čebyševovy vzdálenosti

Závěr:

Všechny výše uvedené metody pro výpočet vzdálenosti mezi shluky daly stejné výsledky, jak pro dva shluky, tak pro tři shluky. Viz obr. 13 a obr. 14. První shluk podobných objektů tvoří roky (1985, 1986), (1989, 1991) a (1994, 1995) a ve druhém shluku jsou zbývající roky.

Příklad 2 Shluková analýza

Máme 5 objektů. Jeden objekt je charakterizován metrickými znaky (2, 10), druhý (3, 8), třetí (4, 9), čtvrtý (10, 4) a pátý (11, 5). Vypočtěte matici vzdáleností v Euklidově metrice a proveďte shlukování metodou jednoduchého spojení. Výsledky interpretujte graficky.

Řešení:

Euklidova metrika je definována vztahem

$$d_E(X_k, X_l) = \left[\sum_{p=1}^{P} \left(x_{kp} - x_{lp} \right)^2 \right]^{1/2} \qquad (2.1)$$

vzdálenost podle kriteria průměrné vazby se vypočte podle vztahu

$$D_{AB}(S_1, S_2) = \frac{1}{N_1 N_2} \sum_{x_k \in S_1} \sum_{x_l \in S_2} d_E(x_k, x_l) \qquad (2.2)$$

kde N_1 a N_2 jsou počty objektů ve třídách S_1 a S_2.

Matice vzdáleností **D**, do níž sestavujeme vypočtené vzdálenosti všech možných dvojic objektů x_k, x_l, je čtvercová symetrická matice řádu N (počet objektů), s nulami na hlavní diagonále. V našem příkladu je matice vzdáleností

Tab. 48 „ruční“ a software výpočet matice vzdáleností

$$D = \begin{Vmatrix} 0 & \sqrt{5} & \sqrt{5} & 10 & \sqrt{106} \\ \sqrt{5} & 0 & \sqrt{2} & \sqrt{65} & \sqrt{73} \\ \sqrt{5} & \sqrt{2} & 0 & \sqrt{61} & \sqrt{65} \\ 10 & \sqrt{65} & \sqrt{61} & 0 & \sqrt{2} \\ \sqrt{106} & \sqrt{73} & \sqrt{65} & \sqrt{2} & 0 \end{Vmatrix}$$

	Euklid. vzdálenosti (shlukova)				
Případ	**S1**	S2	S3	S4	S5
S1	0,0	2,24	2,24	10,0	10,3
S2	2,2	0,00	1,41	8,1	8,5
S3	2,2	1,41	0,00	7,8	8,1
S4	10,0	8,06	7,81	0,0	1,4
S5	10,3	8,54	8,06	1,4	0,0

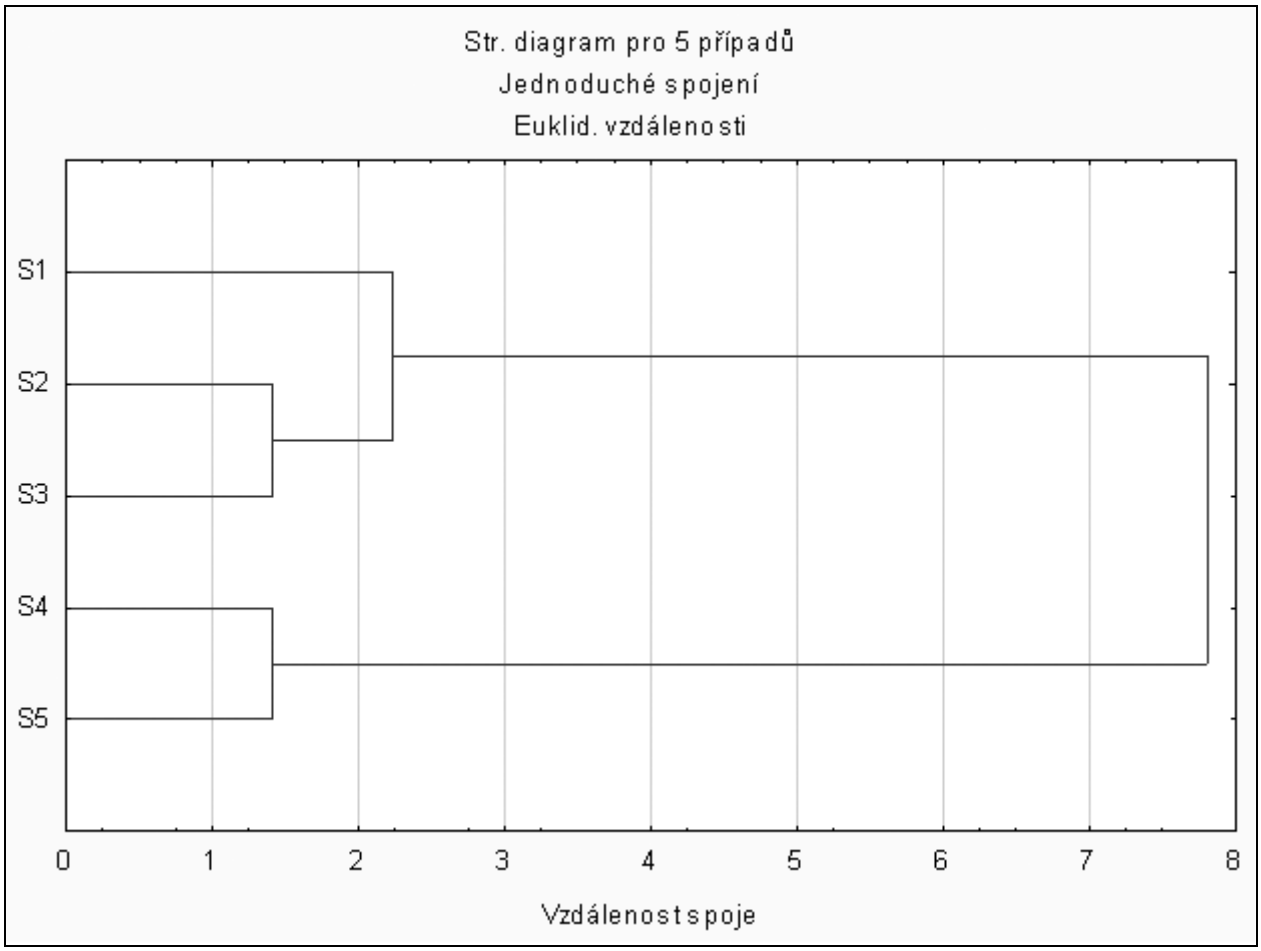
Z matice **D** plyne, že objekty, jejichž vzdálenost (2.1) je rovna $\sqrt{2}$ (objekty S2 a S3; S4 a S5) tvoří dva shluky – řádky 1 a 2 (tab. 4). Další shlukování provedeme podle vztahu (15). Shluk z objektů S2 a S3 vytvoří nový shluk A = (S1, S2, S3), nikoliv shluk B = (S1, S4, S5) protože vzdálenost $D_{AB}(S_{2,3}, \text{"1"}) = \sqrt{5} < D_{AB}(S_{4,5}, \text{"1"})$. Procedura shlukování končí vytvořením shluku, který zahrnuje všechny objekty.

Tab. 49 Rozvrh shlukování

spojení vzdálen.	Rozvrh slučování (shlukova) Jednoduché spojení Euklid. vzdálenosti Obj. č. 1	Obj. č. 2	Obj. č. 3	Obj. č. 4	Obj. č. 5
1,414214	S2	S3			
1,414214	S4	S5			
2,236068	S1	S2	S3		
7,810250	S1	S2	S3	S4	S5

Závěr:

Proces shlukování metodou jednoduchého spojení je znázorněn v tab. 49 a na obr. 15 je odpovídající dendrogram.



Obr. 15 Dendrogram

Literatura

Cyhelský, L. (1974) *Úvod do teorie popisné statistiky*. (1st ed.). Praha: SNTL.

Cyhelský, L., Kahounová, J. & Hindls, R. (1996) *Elementární statistická analýza*. Praha: Management Press.

Hair, J. F., Black, B., Babin, B., Anderson, R. E. & Tatham, R. L. (2006) *Multivariate Data Analysis*. (6th ed.). Prentice Hall.

Hebák, P., Hustopecký, J., Jarošová, E. & Pecáková, I. (2007) *Vícerozměrné statistické metody 1*. (2nd ed.). Praha: Informatorium.

Hebák, P., Malá, I. & Hustopecký, J. (2006) *Vícerozměrné statistické metody 2*. Praha: Informatorium.

Hebák, P., Hustopecký, J., Pecáková, I., Plašil, M., Průša, M., Řezanková, H., Svobodová, A. & Vlach, P. (2007) *Vícerozměrné statistické metody 3*. Praha: Informatorium.

Hendl, J. (2004) *Přehled statistických metod zpracování dat*. Portál.

Hindls, R., Hronová, S. & Novák, I. (2000) *Metody statistické analýzy pro ekonomy*. (2nd ed.). Praha: Management Press.

Johnson, R. A. & Wichern, D. W. (2007) *Applied Multivariate Statistical Analysis*. (6th ed.). Pearson.

Meloun, M. & Militký, J. (2004) *Statistická analýza experimentálních dat*. (2nd ed.). Praha: Academia.

Řezanková, H., Marek, L. & Vrabec, M. (2001) *Interaktivní učebnice statistiky* [online]. [cit. 2011-09-09]. Dostupné na www: http://iastat.vse.cz/

Sebera, M. (2009) *E-learning a ovlivňování učebních stylů*. Disertační práce. Brno: Masarykova univerzita.

Seberová, H. & Sebera, M. (1999) *Počítačové zpracování dat II*. (1st ed.). Vyškov: VVŠ PV.

Tabachnick, B. G. & Fidell, L. S. (2006) *Using Multivariate Statistics*. (5th ed.). Allyn & Bacon.

Zvonař, M., Pavlík, J., Sebera, M., Vespalec, T. & Štochl, J. (2010) *Vybrané kapitoly z antropomotoriky*. 1st ed. Brno: Masarykova univerzita.

Rejstřík