SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

3/2017
ročník/volume 27

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

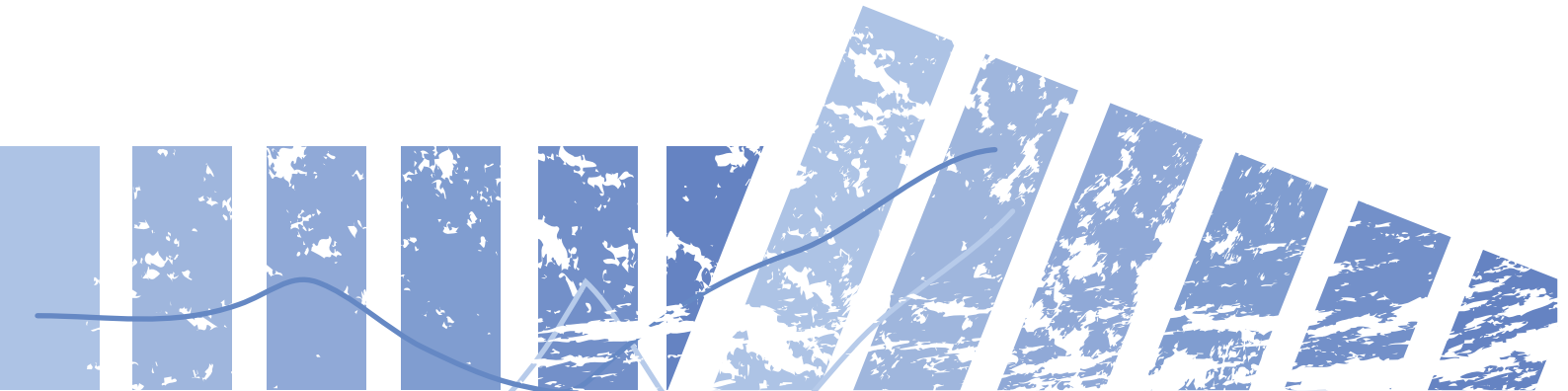
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 3

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 20 – 33

Dátum vydania/Publication date: 15. júl 2017/July 15, 2017



Gábor SZŰCS
Fakulta matematiky, fyziky a informatiky Univerzity Komenského v Bratislave

VIACROZMERNÁ ANALÝZA ROZPTYLU A JEJ APLIKÁCIE

MULTIVARIATE ANALYSIS OF VARIANCE AND ITS APPLICATIONS

ABSTRAKT
Analýza rozptylu bola navrhnutá predovšetkým na posúdenie podobností a odlišností medzi viacerými súbormi, pričom jej prostredníctvom možno testovať efekt jedného alebo viacerých faktorov. Ak sa na pozorovaných objektoch súčasne sledujú viaceré štatistické znaky, odporúča sa použiť viacrozmernú analýzu rozptylu, ktorá ponúka komplexnejšie možnosti testovania odlišností medzi súbormi alebo efektmi. Tento článok poskytuje krátky úvod do teórie jednorozmernej analýzy rozptylu a zovšeobecnenie modelu pre prípad viacerých rozmerov. Uvádza teoretický aparát na testovanie hypotéz pomocou viacrozmernej analýzy rozptylu a ukážky jej použitia na reálnych dátach v softvéri R.

ABSTRACT
Analysis of variance was primarily designed to assess similarities and differences between more groups and by means of which the effect of one or more factors can be tested. If various statistical features are being pursued simultaneously on the observed objects, it is recommended to use the multivariate analysis of variance offering more complex possibilities of the difference testing between two or more groups or effects. This paper provides a brief introduction to the theory of one-dimensional analysis of variance and the generalization of the model to the multivariate case. It presents the theoretical apparatus for hypothesis testing using the multivariate analysis of variance and its illustrations on real data sets in R software.

KĽÚČOVÉ SLOVÁ
analýza rozptylu, viacrozmerná analýza rozptylu, testovanie rovnosti vektorov stredných hodnôt, testovanie významnosti vplyvu vysvetľujúcich premenných

KEY WORDS
analysis of variance, multivariate analysis of variance, testing the equality of mean vectors, significance test of the impact of explanatory variables

1. ÚVOD
Analýza rozptylu (známa aj pod skratkou ANOVA, Analysis of Variance) patrí medzi najpoužívanejšie parametrické metódy matematickej štatistiky. Je zovšeobecnením dvojvýberového Studentovho t-testu, pretože pomocou nej môžeme porovnať nielen dva, ale aj viaceré súbory a zistiť podobnosti a odlišnosti medzi nimi. Model analýzy rozptylu je založený na lineárnom regresnom modeli, v ktorom kvantitatívnu vysvetľovanú premennú popisujú neznáme regresné koeficienty, matica plánu a náhodný šum. ANOVA má široké spektrum praktického využitia, napr. vo farmácii, v biológii, genetike, poľnohospodárstve či priemysle.

Model ANOVA vychádza zo základného predpokladu, že rozdelenie pravdepodobnosti vysvetľovanej premennej je normálne (Gaussovo rozdelenie). Tento predpoklad je jednou z najväčších slabín analýzy rozptylu z hľadiska praktickej aplikácie metódy, pretože sledované štatistické znaky sa v každom prípade neriadia normálnym rozdelením. Pri silnom porušení predpokladu normality sa odporúča, aby sa štatistický test a model analýzy rozptylu nahradil jeho neparametrickým ekvivalentom, takzvaným Kruskalovým-Wallisovým testom, ktorý nevyžaduje, aby sa vysvetľované premenné správali podľa určitého rozdelenia pravdepodobnosti. Pomocou Kruskalovho-Wallisovho neparametrického testu môžeme testovať aj hypotézu o rovnosti stredných hodnôt v jednotlivých súboroch, teda v konečnom dôsledku môžeme dospieť k podobnému štatistickému vyhodnoteniu ako pri analýze rozptylu.

V praxi sa stretávame aj s takými situáciami, v ktorých je potrebné súčasne vysvetliť viaceré štatistické znaky. Vtedy už základná jednorozmerná analýza rozptylu nemusí stačiť na adekvátny opis premenných prostredníctvom vysvetľujúcich efektov. Tým, že pri analýze by sa naraz bral do úvahy iba jeden štatistický znak, mohli by sa stratiť cenné informácie zo simultánneho sledovania všetkých vysvetľovaných premenných. Viacrozmerná analýza rozptylu (v medzinárodnej literatúre nazývaná ako Multivariate Analysis of Variance, MANOVA) práve pre takéto situácie ponúka čiastočné riešenie a komplexnejšie testovanie odlišností medzi viacerými súbormi, na ktorých sa súčasne sledujú minimálne dva štatistické znaky.

Teoretické základy modelu MANOVA a možnosti jeho využitia boli vysvetlené vo viacerých výborných publikáciách, napríklad v [4], [5], [8], [9], [13]. S konkrétnymi aplikáciami MANOVA sa môžeme stretnúť v rôznych odvetviach vedeckého výskumu, napríklad v genetike [17], hydrológii [18], neurológii alebo v oblasti potravinárstva. Postupy viacrozmernej analýzy rozptylu by sa dali využiť napríklad aj pri práci s dátami z oblasti antropometrie [6], psychológie [15] a pri tvorbe ekonomických štúdií [10]. Náš ilustračný príklad, ktorý prezentujeme v 4. časti tohto článku, je z oblasti papierenského priemyslu [1], [7].

2. MODEL JEDNOROZMERNEJ ANALÝZY ROZPTYLU

Ako sme už v úvode spomínali, model jednorozmernej analýzy rozptylu je vlastne lineárnym regresným modelom, ktorý môžeme zapísať v tvare $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{\varepsilon}$, kde $\boldsymbol{Y}$ je náhodný vektor vysvetľovaných premenných dĺžky n, $\boldsymbol{X}$ je tzv. matica plánu, $\boldsymbol{B}$ je vektor efektov (prípadne vektor stredných hodnôt) a $\boldsymbol{\varepsilon}$ je náhodný vektor dĺžky n, s nezávislými, rovnako rozdelenými zložkami, ktoré majú normálne rozdelenie $N(0, \sigma^2)$, pričom σ^2 je rozptyl (alebo disperzia) zložiek náhodného vektora $\boldsymbol{Y}$, aj zložiek náhodného vektora $\boldsymbol{\varepsilon}$.

Prípad dvoch súborov a jedného faktora

V prípade, keď potrebujeme porovnať dva súbory, vektor $\boldsymbol{Y}$ je stĺpcovým náhodným vektorom $\boldsymbol{Y} = \left(Y_{11}, Y_{12}, \ldots, Y_{1n_1}, Y_{21}, Y_{22}, \ldots, Y_{2n_2}\right)$, kde zložky $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ sú hodnoty sledovaného štatistického znaku na prvom objekte, $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ sú hodnoty sledovaného štatistického znaku na druhom objekte a $n_1 + n_2 = n$. V súlade s vyššie uvedenou konštrukciou predpokladáme, že $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ je náhodný výber z rozdelenia $N(\mu_1, \sigma^2)$ s neznámou strednou hodnotou $\mu_1 \in \mathbb{R}$, a analogicky

$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ je náhodný výber z rozdelenia $N(\mu_2, \sigma^2)$ s neznámou strednou hodnotou $\mu_2 \in \mathbb{R}$. Tiež predpokladáme, že všetky zložky vektora $\boldsymbol{Y}$ sú navzájom nezávislé. Maticu plánu $\boldsymbol{X}$ v tomto prípade môžeme zapísať v tvare:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}.$$

Vektor efektov $\boldsymbol{B}$ je v tomto prípade dvojzložkovým stĺpcovým vektorom neznámych stredných hodnôt, môžeme teda písať v tvare: $\boldsymbol{B} = (\mu_1, \mu_2)$. Poznamenáme, že počet jednotiek v hornej časti matice $\boldsymbol{X}$ je n_1, kým v dolnej časti matice ich počet je n_2. Matica plánu $\boldsymbol{X}$ vlastne „zapína“ a „vypína“ efekty stredných hodnôt na 1., resp. 2. súbor.

V praxi obvykle hľadáme odpoveď na otázku, či sú medzi dvoma sledovanými súbormi štatisticky významné rozdiely alebo nie. Presnejšie, či sa stredná hodnota 1. súboru štatisticky významne líši od strednej hodnoty 2. súboru alebo nie. Nulová hypotéza príslušného štatistického testu sa zvyčajne formuluje v tvare $H_0: \mu_1 = \mu_2$ a testuje sa oproti alternatívnej hypotéze $H_1: \mu_1 \neq \mu_2$. Opísaný model je najjednoduchším prípadom analýzy rozptylu, ktorý sa nazýva (jednorozmernou) jednofaktorovou analýzou rozptylu pre prípad dvoch súborov a je totožný s modelom dvojvýberového Studentovho t-testu. Ďalšie detaily modelu a odvodenie testovej štatistiky sú dostupné napr. v [8], [9].

Ako jednoduchý príklad by sme mohli uviesť aplikáciu z poľnohospodárstva, v ktorej potrebujeme porovnať hektárové výnosy dvoch odrôd pšenice. V tomto prípade by premenná Y_{11} predstavovala hektárový výnos prvej odrody pšenice na prvom poli (pri prvom meraní), premenná Y_{12} by znamenala hektárový výnos prvej odrody pšenice na druhom poli (pri druhom meraní) atď., až veličina Y_{1n_1} by špecifikovala hektárový výnos prvej odrody pšenice pri n_1-om meraní. Analogicky premenné $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ by postupne určovali hektárový výnos druhej odrody pšenice pri prvom, druhom až n_2-om meraní.

Predpokladajme, že hektárové výnosy v prípade oboch odrôd pšenice sa správajú podľa normálneho rozdelenia s rovnakou smerodajnou odchýlkou $\sigma > 0$ a merania sú navzájom nezávislé. Pomocou testu jednorozmernej jednofaktorovej analýzy rozptylu by sme mohli testovať, či sú štatisticky významné rozdiely medzi dvoma odrodami pšenice, čo sa týka ich očakávaného (priemerného) hektárového výnosu. Táto analýza je *jednorozmerná*, pretože pri každom meraní sledujeme jediný štatistický znak – hektárový výnos. Keby sme potrebovali súčasne vysvetliť aj ďalšie štatistické znaky, napríklad dĺžku stebla pšenice, hmotnosť slamy alebo dĺžku klasu, tak by sme používali viacrozmernú analýzu rozptylu (pozri v 3. časti tohto príspevku). Ďalej, náš pôvodný príklad je *jednofaktorový*, pretože priemerný hektárový výnos vysvetľujeme len pomocou jedného faktora – odrody pšenice. Ak by sme pri opise hektárového výnosu chceli brať do úvahy efekt viacerých vysvetľujúcich premenných, napríklad efekt typu pôdy alebo aj efekt hnojenia, tak by sme aplikovali dvoj- či trojfaktorovú analýzu rozptylu (pozri záverečnú časť tejto kapitoly).

Prípad viacerých súborov a jedného faktora

Predchádzajúci jednoduchý model bol navrhnutý na porovnanie dvoch súborov (v našom ilustračnom príklade dvoch odrôd pšenice), ľahko ho však môžeme zovšeobecniť pre prípad ℓ súborov (napríklad pre ℓ rôznych odrôd pšenice). Potom náhodný vektor $\boldsymbol{Y}$ možno napísať v tvare stĺpcového vektora dĺžky n ako $(Y_{11}, Y_{12}, \dots, Y_{1n_1}, Y_{21}, Y_{22}, \dots, Y_{2n_2}, \dots, Y_{\ell 1}, Y_{\ell 2}, \dots, Y_{\ell n_\ell})$, kde $n_1, n_2, \dots, n_\ell$ sú počty meraní v jednotlivých súboroch a $n_1 + n_2 + \cdots + n_\ell = n$, pričom náhodná premenná Y_{ik} vyjadruje hodnotu štatistického znaku k-teho objektu (k-teho pozorovania) v i-tom súbore pre $k = 1,2, \dots, n_i$ a $i = 1,2, \dots, \ell$. Opäť predpokladajme, že $Y_{i1}, Y_{i2}, \dots, Y_{in_i}$ je náhodný výber z rozdelenia $N(\mu_i, \sigma^2)$ s neznámou strednou hodnotou $\mu_i \in \mathbb{R}$ pre $i = 1,2, \dots, \ell$ a zložky vektora $\boldsymbol{Y}$ sú nezávislé. Vektor efektov $\boldsymbol{B}$ v lineárnom regresnom modeli $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{\varepsilon}$ je v tomto prípade ℓ-zložkovým stĺpcovým vektorom neznámych stredných hodnôt, t. j. $\boldsymbol{B} = (\mu_1, \mu_2, \dots, \mu_\ell)$, kým maticu plánu $\boldsymbol{X}$ môžeme písať v tvare

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

kde počty jednotiek v jednotlivých blokoch matice $\boldsymbol{X}$ sú postupne $n_1, n_2, \dots, n_\ell$.

Základnou úlohou je overiť platnosť nulovej hypotézy o rovnosti stredných hodnôt sledovaného štatistického znaku v rámci všetkých ℓ súborov, t. j. $H_0: \mu_1 = \mu_2 = \cdots = \mu_\ell$. Táto nulová hypotéza sa testuje oproti alternatívnej hypotéze $H_1: \exists i, h \in \{1,2, \dots, \ell\}, i \neq h$, pre ktoré $\mu_i \neq \mu_h$. V nulovej hypotéze tvrdíme, že medzi súbormi nie sú štatisticky významné rozdiely v zmysle strednej hodnoty sledovaného štatistického znaku. Ak túto spoločnú strednú hodnotu z nulovej hypotézy označíme symbolom μ, tak model jednofaktorovej analýzy rozptylu môžeme prepísať do tvaru

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik} = \mu_i + \varepsilon_{ik},$$

kde α_i je čistý efekt i-teho súboru na hodnotu sledovaného štatistického znaku ($i \in \{1,2, \dots, \ell\}$) a platí vzťah $\mu_i = \mu + \alpha_i$. Formálne sme teda pôvodný model iba reparametrizovali: strednú hodnotu i-teho súboru (μ_i) sme nahradili súčtom spoločnej strednej hodnoty (spoločného efektu μ) a čistého efektu i-teho súboru (α_i). Potom nulovú hypotézu o rovnosti stredných hodnôt môžeme zapísať aj v ekvivalentnom tvare $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_\ell$.

Vzniká prirodzená otázka: prečo sa táto parametrická štatistická metóda nazýva analýzou rozptylu, ak primárne slúži na testovanie rovnosti stredných hodnôt? Odpoveď treba hľadať v odvodení a vo výslednom tvare testovej štatistiky spomínaného testu (pozri napríklad v [8], [9]), ktorá dáva do pomeru sumu štvorcov

odchýlok vysvetleného modelu (akýsi rozptyl alebo variabilitu medzi súbormi) a celkovú sumu štvorcov odchýlok (celkový rozptyl vysvetľovanej premennej). Testovú štatistiku v reči matematickej štatistiky zapisujeme v tvare

$$F = \frac{\frac{\mathrm{ESS}}{\ell - 1}}{\frac{\mathrm{TSS}}{n - \ell}} = \frac{\frac{\mathrm{TSS} - \mathrm{RSS}}{\ell - 1}}{\frac{\mathrm{TSS}}{n - \ell}},$$

kde $\mathrm{ESS} = \sum_{i=1}^{\ell} n_i (\bar{Y}_{i\cdot} - \bar{Y})^2$ je suma štvorcov odchýlok vysvetleného modelu (*explained sum of squares*, ESS), $\bar{Y}_{i\cdot}$ je aritmetický priemer vo vnútri i-teho súboru, $\bar{Y}$ je celkový aritmetický priemer vektora $\boldsymbol{Y}$, $\mathrm{RSS} = \sum_{i=1}^{\ell} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_{i\cdot})^2$ je suma štvorcov rezíduí (*residual sum of squares*, RSS) a $\mathrm{TSS} = \sum_{i=1}^{\ell} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y})^2$ je celková suma štvorcov odchýlok (*total sum of squares*, TSS). Medzi zavedenými veličinami platí vzťah $\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$. Za predpokladu normality, nezávislosti premenných Y_{ik} a platnosti nulovej hypotézy platí, že testová štatistika F má Fisherovo F-rozdelenie so stupňami voľnosti $\ell - 1$ a $n - \ell$. Spomínanú F-štatistiku môžeme interpretovať nasledovne: ak medzi strednými hodnotami sledovaných súborov sú len malé rozdiely, tak rozptyl medzi súbormi (ESS) bude malý a celá hodnota F-štatistiky bude nízka. Ak hodnota štatistiky je menšia ako kritická hodnota F-rozdelenia, tak hypotézu o rovnosti stredných hodnôt nezamietame. V opačnom prípade, ak už len jediný súbor je iný ako ostatné, tak hodnota čitateľa F-štatistiky bude vysoká (pre vysokú variabilitu medzi súbormi), teda F-hodnota pravdepodobne presiahne kritickú hodnotu F-rozdelenia a hypotézu H_0 zamietneme.

Prípad viacerých súborov a viacerých faktorov

Najkomplexnejším prípadom jednorozmernej analýzy rozptylu je ten, v ktorom porovnávame viaceré súbory a vysvetľované premenné opisujeme pomocou viacerých efektov a ich interakcií. Idea a predpoklady ANOVA aj v tomto prípade zostávajú rovnaké, ako sme ich opísali predtým, avšak lineárny regresný model a testové štatistiky budú z matematického hľadiska už o niečo zložitejšie. Ako príklad dvojfaktorovej analýzy rozptylu by sme mohli uviesť istú modifikáciu ilustračnej štúdie z oblasti poľnohospodárstva (uvedenej na začiatku tejto kapitoly). Predpokladajme teda, že potrebujeme porovnať hektárové výnosy siedmich odrôd pšenice, ktoré sa pestovali v troch rôznych typoch pôdy. Nech Y_{ijk} je hektárový výnos i-tej odrody pšenice siatej do j-teho typu pôdy v prípade k-teho pozorovania (pri danej kombinácii odrody pšenice a typu pôdy), pričom v našom príklade $i = 7$ odrôd pšenice, $j = 3$ a $k = 1,2, \ldots, n_{ij}$, kde n_{ij} je počet meraní v prípade i-tej odrody a j-teho typu pôdy. Podrobný všeobecný popis dvojfaktorovej analýzy rozptylu a odvodenie testových štatistík na testovanie hypotézy o rovnosti stredných hodnôt je možné nájsť napríklad v knihe [9].

Možnosti používania modelu ANOVA a testovanie hypotéz v tomto modeli ponúka viacero matematických či štatistických softvérov, ako napríklad SPSS, Microsoft Excel, SAS alebo R. V prípade posledného menovaného softvéru možno aplikovať napríklad funkciu `aov`, ktorá je súčasťou základného štatistického balíka programu [12].

3. MODEL VIACROZMERNEJ ANALÝZY ROZPTYLU

Model viacrozmernej analýzy rozptylu vznikol zovšeobecnením jednorozmerného modelu pre situácie, keď sa na skúmaných objektoch súčasne sleduje p štatistických znakov, kde p je prirodzené číslo väčšie ako 1. Podkladový lineárny regresný model viacrozmernej analýzy rozptylu (tzv. všeobecný model MANOVA) môžeme zapísať v tvare

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E},$$

kde $\boldsymbol{Y}$ je náhodná matica typu $n \times p$, pričom n je počet všetkých sledovaných objektov (počet riadkov matice $\boldsymbol{Y}$), p je počet sledovaných štatistických znakov (počet stĺpcov matice $\boldsymbol{Y}$), $\boldsymbol{X}$ je známa nenáhodná matica plánu rozmerov $n \times \ell$, pričom ℓ je počet regresných parametrov (efektov alebo ošetrení) používaných pre každý sledovaný štatistický znak, $\boldsymbol{B}$ je nenáhodná neznáma matica efektov typu $\ell \times p$ a $\boldsymbol{E}$ je náhodná matica rozmerov $n \times p$, ktorej každý riadok je náhodným výberom z p-rozmerného normálneho rozdelenia $N_p(\boldsymbol{0},\boldsymbol{\Sigma})$, kde $\boldsymbol{\Sigma}$ je kovariančnou maticou p-tice sledovaných štatistických znakov.

Prípad jedného faktora

Ak sledovanú závislú premennú vysvetľujeme pomocou jedného faktora, tak vyššie uvedený všeobecný model viacrozmernej analýzy rozptylu môžeme napísať v tvare

$$\boldsymbol{Y}_{ki} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ki},$$

kde $\boldsymbol{Y}_{ki}$ je p-rozmerný vektor hodnôt sledovaných štatistických znakov v prípade k-teho pozorovania v i-tom súbore (alebo v prípade i-teho ošetrenia) pre $k = 1,2,\dots,n_i$ a $i = 1,2,\dots,\ell$, pričom n_i je počet pozorovaní (meraní) vykonaných v i-tom súbore a $n = \sum_{i=1}^{\ell} n_i$ je celkový počet pozorovaní. Ďalej, symbolom $\boldsymbol{\mu}$ označujeme vektor stredných hodnôt sledovaných štatistických znakov nezávislý od súborov ($\boldsymbol{\mu}$ je tzv. spoločný alebo celkový efekt). Podobne ako pri jednorozmernom modeli ANOVA, aj v tomto prípade $\boldsymbol{\alpha}_i$ označuje čistý efekt i-teho súboru na hodnoty sledovaných štatistických znakov. Poznamenajme, že $\boldsymbol{\alpha}_i$ je tiež p-rozmerným vektorom, pre ktorý platí vzťah $\boldsymbol{\mu} + \boldsymbol{\alpha}_i = \boldsymbol{\mu}_i$, kde $\boldsymbol{\mu}_i$ je vektorom stredných hodnôt i-teho súboru. Posledným členom modelu jednofaktorovej viacrozmernej analýzy rozptylu je p-rozmerný náhodný šum $\boldsymbol{\varepsilon}_{ki}$, o ktorom predpokladáme, že pochádza z p-rozmerného normálneho rozdelenia $N_p(\boldsymbol{0},\boldsymbol{\Sigma})$.

Hlavnou úlohou MANOVA je testovanie vplyvu súborov na hodnoty sledovaných štatistických znakov, t. j. či jednotlivé súbory majú vplyv na výsledné hodnoty alebo nie. Nulovú hypotézu o rovnosti vektorov stredných hodnôt sledovaného štatistického znaku v rámci všetkých ℓ súborov môžeme písať v tvare $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_\ell$. Ak platí nulová hypotéza, tak to znamená, že medzi súbormi nie sú štatisticky významné rozdiely v zmysle strednej hodnoty a vysvetľované premenné možno popísať len pomocou spoločného efektu $\boldsymbol{\mu}$. Uvedená nulová hypotéza sa obvykle testuje oproti alternatívnej hypotéze $H_1: \exists i,h \in \{1,2,\dots,\ell\}, i \neq h$, pre ktoré $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_h$. Tvrdí, že medzi súbormi je aspoň jeden, ktorý sa štatisticky významne líši od ostatných súborov v zmysle vektora stredných hodnôt. Na definovanie testových štatistík pre uvedenú

nulovú hypotézu je potrebné zaviesť niekoľko ďalších označení a definovať rozdelenia pravdepodobnosti pre viacrozmerné náhodné výbery.

Definícia 1. Nech $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n$ je náhodný výber z p-rozmerného normálneho rozdelenia $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, kde $\boldsymbol{\Sigma}$ je pozitívne definitná matica typu $p \times p$. Nech $\boldsymbol{Z}^\top = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n)$ je náhodná matica typu $p \times n$ a nech $\boldsymbol{M} = \boldsymbol{Z}^\top \boldsymbol{Z}$ je náhodná matica typu $p \times p$. Rozdelenie pravdepodobnosti matice $\boldsymbol{M}$ nazývame Wishartovým rozdelením s parametrom $\boldsymbol{\Sigma}$ a stupňami voľnosti n; označujeme to zápisom $\boldsymbol{M} \sim W_p(\boldsymbol{\Sigma}, n)$.

Definícia 2. Nech $\boldsymbol{M}$ a $\boldsymbol{N}$ sú nezávislé náhodné matice typu $p \times p$ a nech $\boldsymbol{M} \sim W_p(\boldsymbol{I}_p, m)$, $\boldsymbol{N} \sim W_p(\boldsymbol{I}_p, n)$, pričom $m \geq p$ a symbolom $\boldsymbol{I}_p$ označujeme identickú maticu typu $p \times p$ (maticu, ktorá má na svojej hlavnej diagonále samé jednotky a mimo hlavnej diagonály samé nuly). Definujme náhodnú premennú Λ predpisom

$$\Lambda = \frac{\det(\boldsymbol{M})}{\det(\boldsymbol{M} + \boldsymbol{N})} = \frac{1}{\det(\boldsymbol{I}_p + \boldsymbol{M}^{-1}\boldsymbol{N})},$$

kde $\det(\boldsymbol{M})$ označuje determinant matice $\boldsymbol{M}$ a $\boldsymbol{M}^{-1}$ je inverznou maticou matice $\boldsymbol{M}$. Rozdelenie pravdepodobnosti náhodnej premennej Λ nazývame Wilksovým Lambda-rozdelením s parametrami p, m, n; označujeme to zápisom $\Lambda \sim \Lambda(p, m, n)$.

Definujme teraz viacrozmerné ekvivalenty veličín ESS, RSS a TSS zavedených v 2. časti tohto príspevku pri modeli jednorozmernej jednofaktorovej analýzy rozptylu.

Definícia 3. Nech pre náhodnú maticu $\boldsymbol{Y}$ platia všetky predpoklady modelu MANOVA a nech

$$\boldsymbol{H} = \sum_{i=1}^{\ell} n_i (\overline{\boldsymbol{Y}}_{\cdot i} - \overline{\boldsymbol{Y}})(\overline{\boldsymbol{Y}}_{\cdot i} - \overline{\boldsymbol{Y}})^\top$$

je matica súčtov štvorcov a súčinov odchýlok medzi súbormi, kde $\overline{\boldsymbol{Y}}_{\cdot i}$ je p-rozmerný vektor aritmetických priemerov hodnôt sledovaných štatistických znakov vo vnútri i-teho súboru a $\overline{\boldsymbol{Y}}$ je p-rozmerný vektor celkových aritmetických priemerov počítaných po stĺpcoch matice $\boldsymbol{Y}$. Ďalej nech

$$\boldsymbol{E} = \sum_{i=1}^{\ell} \sum_{k=1}^{n_i} (\boldsymbol{Y}_{ki} - \overline{\boldsymbol{Y}}_{\cdot i})(\boldsymbol{Y}_{ki} - \overline{\boldsymbol{Y}}_{\cdot i})^\top$$

je matica súčtov štvorcov a súčinov odchýlok vo vnútri súborov a

$$\boldsymbol{T} = \sum_{i=1}^{\ell} \sum_{k=1}^{n_i} (\boldsymbol{Y}_{ki} - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_{ki} - \overline{\boldsymbol{Y}})^\top$$

je matica celkových súčtov štvorcov a súčinov odchýlok.

Poznamenáme, že medzi vyššie definovanými maticami platí vzťah $\boldsymbol{T} = \boldsymbol{E} + \boldsymbol{H}$, ktorý je viacrozmernou analógiou vzťahu $\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$. Tiež dodávame, že za

predpokladu, že riadky matice $\boldsymbol{Y}$ pochádzajú z náhodného výberu z rozdelenia $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, náhodné matice $\boldsymbol{E}$ a $\boldsymbol{H}$ majú Wishartovo rozdelenie.

Veta 1. Nech platia všetky vyššie zavedené označenia a predpoklady. Potom testová štatistika pomerom vierohodností pre hypotézu o rovnosti vektorov stredných hodnôt sledovaného štatistického znaku v rámci ℓ súborov v modeli MANOVA ($H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_\ell$) má tvar

$$\Lambda = \frac{\det(\boldsymbol{E})}{\det(\boldsymbol{T})} = \frac{\det(\boldsymbol{E})}{\det(\boldsymbol{E}+\boldsymbol{H})} = \frac{1}{\det(\boldsymbol{I}_p + \boldsymbol{E}^{-1}\boldsymbol{H})},$$

kde testová štatistika Λ má Wilksovo Lambda-rozdelenie s parametrami $p, \nu_{\boldsymbol{E}}, \nu_{\boldsymbol{H}}$, kde $\nu_{\boldsymbol{E}} = n - \ell$ je hodnosť náhodnej matice $\boldsymbol{E}$ a $\nu_{\boldsymbol{H}} = \ell - 1$ je hodnosť náhodnej matice $\boldsymbol{H}$.

Dôkaz Vety 1 možno nájsť napríklad v [9].

Test zavedený vo Vete 1 nazývame Wilksovým testom pomerom vierohodností a interpretujeme ho podobne ako F-test definovaný pre jednorozmernú analýzu rozptylu. Ak determinant matice $\boldsymbol{E}^{-1}\boldsymbol{H}$ je malý, tak to znamená, že medzi súbormi nie sú veľké odchýlky. Potom aj determinant matice $\boldsymbol{I}_p + \boldsymbol{E}^{-1}\boldsymbol{H}$ bude malý, teda to znamená, že nulovú hypotézu nezamietame pri vysokých hodnotách Λ-testovej štatistiky. Hodnoty testovej štatistiky sa v praxi obvykle porovnávajú s kritickými hodnotami aproximatívneho Fisherovho F-rozdelenia alebo s kritickými hodnotami aproximatívneho χ^2-rozdelenia (ďalšie detaily sú uvedené v publikáciách [8], [9], [14]).

Pri testovaní hypotézy o rovnosti vektorov stredných hodnôt v modeli MANOVA sa okrem Wilksovho testu pomerom vierohodností používajú aj ďalšie testy, napríklad Pillaiov test, Lawleyov-Hotellingov test alebo Royov test.

Nech platia všetky vyššie zavedené označenia a predpoklady. Definujme testovú štatistiku V ako

$$V = \mathrm{st}((\boldsymbol{E}+\boldsymbol{H})^{-1}\boldsymbol{H}),$$

kde st označuje stopu matice, teda súčet diagonálnych prvkov matice. Testová štatistika V sa nazýva Pillaiovou stopou [11], [14]. Ďalej, definujme testovú štatistiku U predpisom

$$U = \mathrm{st}(\boldsymbol{E}^{-1}\boldsymbol{H}),$$

ktorá sa nazýva Lawleyovou-Hotellingovou stopou [8], [14]. Označme symbolom λ_1 najväčšie vlastné číslo matice $\boldsymbol{E}^{-1}\boldsymbol{H}$, ktoré sa nazýva Royovo najväčšie vlastné číslo. Potom testová štatistika definovaná ako

$$\theta = \frac{\lambda_1}{1+\lambda_1}$$

sa nazýva Royovou testovou štatistikou pomocou najväčšieho vlastného čísla [8], [14]. V teoretických prácach autorov týchto testových štatistík sa ukázalo, že všetky

tri sú aplikovateľné pri testovaní hypotézy $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_\ell$ v modeli MANOVA. Hoci všetky tri testové štatistiky (náhodné premenné V, U a θ) majú odvodené svoje rozdelenie pravdepodobnosti, v praxi sa obvykle aplikujú aproximácie založené na Fisherovom F-rozdelení. Ďalšie podrobnosti o aproximáciách a ich použiteľnosti obsahuje napríklad práca [14].

4. PRAKTICKÁ APLIKÁCIA VIACROZMERNEJ ANALÝZY ROZPTYLU

Praktickú ukážku testovania hypotéz v modeli MANOVA sme zvolili z oblasti papierenského priemyslu. Pôvodné dáta boli publikované v článku [1] a ďalej sa podrobnejšie analyzovali v knihe [7]. V oboch prípadoch však išlo o analýzy zamerané na použiteľnosť tzv. zovšeobecnených lineárnych modelov (*generalized linear models*, GLM), takže naše skúmania o aplikovateľnosti MANOVA sa od nich odlišujú.

Experiment sa uskutočnil v roku 1985 v papierni *Norske Skog* v nórskom meste Skogn a bol zameraný na meranie kvality papiera pri rôznych nastaveniach výrobného procesu. Trval 30 hodín, pričom výskumníci v hodinových intervaloch merali a zaznamenávali 13 rôznych ukazovateľov kvality papiera, ktoré teraz pracovne označíme ako $Y_1, Y_2, \ldots, Y_{12}, Y_{13}$. Počas experimentu sa podarilo zaznamenať 29 úspešných meraní, jedno meranie sa nepodarilo. Cieľom pôvodného experimentu bolo zistiť, ako vplývajú 3 ovplyvniteľné nastavenia procesu výroby papiera (označíme ich ako X_1, X_2, X_3) na kvalitu papiera. Ukázalo sa, že výstupné premenné $Y_1, Y_2, \ldots, Y_{12}, Y_{13}$ majú zložitú multivariačnú štruktúru, a preto už autor pôvodného článku [1] na vyhodnotenie experimentu používal metodiky viacrozmerných štatistických analýz.

V našich analýzach sa zameriame na testovanie hypotézy o rovnosti vektorov stredných hodnôt v modeli MANOVA, t. j. otestujeme, či rôzne nastavenia troch vstupov výroby majú alebo nemajú vplyv na 13 súčasne sledovaných štatistických znakov (na ukazovatele kvality papiera). Dodatočne vykonáme aj testovanie významnosti vplyvu vysvetľujúcich premenných (nastavení výrobného procesu) na sledované vysvetľované premenné (ukazovatele kvality papiera). Všetky naše analýzy sa uskutočnili v prostredí softvéru R [12], značnú časť postupu preto uvádzame heslovito a v programátorskom formáte.

```
# načítanie dát (zdroj dát: [1])
np <- read.table(
"http://www.iam.fmph.uniba.sk/ospm/Szucs/data/norwaypaper.csv",
header=TRUE); attach(np);

# názvy používaných premenných:
# Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 Y9 Y10 Y11 Y12 Y13 X1 X2 X3
# vypísanie matice plánu:
cbind(rep(1,length(X1)), X1, X2, X3)
```

Keďže model MANOVA je založený na predpoklade normality vysvetľovaných premenných, v prvom kroku vykonáme zovšeobecnený Shapirov-Wilkov test na viacrozmerné normálne rozdelenie publikovaný v článku [16]. Pri tomto teste v softvéri R používame balík `mvShapiroTest` [3].

```
install.packages("mvShapiroTest"); library(mvShapiroTest);
mvShapiro.Test(cbind(Y1,Y2,Y3,Y4,Y5,Y6,Y7,Y8,Y9,Y10,Y11,Y12,Y13))
# Výstup:
# Generalized Shapiro-Wilk test for Multivariate Normality by
# Villasenor-Alva and Gonzalez-Estrada
# MVW = 0.96663, p-value = 0.698
```

Skratkou `MVW` sa v predošlom výstupe označuje hodnota testovej štatistiky zovšeobecneného Shapirovho-Wilkovho testu na viacrozmerné normálne rozdelenie (matematické vyjadrenie testovej štatistiky možno nájsť napríklad v článku [16]). Nulovou hypotézou testu je, že dáta pochádzajú z normálneho rozdelenia. P-hodnota testu (`p-value`) vyšla pomerne vysoká (0,698), takže nulovú hypotézu by sme nezamietli na bežne používaných hladinách významnosti, teda hodnoty našich ukazovateľov kvality papiera by sa mohli riadiť podľa 13-rozmerného normálneho rozdelenia. Výsledok tohto testu však musíme brať s určitou rezervou, pretože v dátovom súbore sa nachádza len 29 pozorovaní, čo je pomerne malý počet na adekvátne posúdenie normality.

Keďže viacrozmernú normalitu vysvetľovaných premenných sme nezamietli, môžeme pristúpiť k hlavným testom v modeli MANOVA. V nulovej hypotéze testu o rovnosti vektorov stredných hodnôt predpokladáme, že rôzne nastavenia troch vstupov výrobného procesu nemajú štatisticky významný vplyv na výslednú kvalitu papiera, t. j. $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$ alebo $H_0: \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_3$. Pri testovaní používame všetky štyri testové štatistiky uvedené na konci predchádzajúcej časti a výpočty vykonáme pomocou balíka `car` [2].

```
# vytvorenie lineárneho regresného modelu,
# ktorý obsahuje aj celkový efekt
model1 <- lm(cbind(Y1,Y2,Y3,Y4,Y5,Y6,Y7,Y8,Y9,Y10,Y11,Y12,Y13) ~
             1+X1+X2+X3, data=np)
#
install.packages("car"); library(car);
# CAR = Companion to Applied Regression [2]
#
# používame príkaz lht = Linear Hypothesis Test
lht(model1, c("X1", "X2", "X3"))
#
# Výstup:
# Multivariate Tests:
#              Df test stat approx F num Df  den Df     Pr(>F)
# Pillai        3  2.269414  3.584183     39 45.00000 2.6920e-05 ***
# Wilks         3  0.006956  4.380605     39 39.24367 5.2064e-06 ***
# Hot.-Lawley   3 17.176653  5.138315     39 35.00000 1.7985e-06 ***
# Roy           3 11.897814 13.728247     13 15.00000 4.8994e-06 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Stĺpce výstupu sú nasledovné:
# názov testovej štatistiky (Pillai, Wilks, Hot.-Lawley, Roy),
# Df = počet stupňov voľnosti (počet testovaných premenných),
# test stat = hodnota testovej štatistiky,
# approx F = hodnota aproximatívnej F-štatistiky,
# num Df, den Df = počty stupňov voľnosti (parametre) F-štatistiky,
```

```
# Pr(>F) = p-hodnota testu počítaná z aproximatívnej F-štatistiky.
#
# V poslednom riadku výstupu sú uvedené vizuálne kódy
# signifikantnosti testov medzi rôznymi hladinami významnosti.
```

Z posledného stĺpca výstupu, označeného ako `Pr(>F)`, je zrejmé, že p-hodnota všetkých štyroch testov vyšla veľmi nízka. To znamená, že pri testoch nastalo významné porušenie nulovej hypotézy, ktorú preto s veľkou istotou zamietame. To znamená, že aspoň jedno z troch nastavení vstupov výrobného procesu má štatisticky významný vplyv na výslednú kvalitu papiera, teda efekt aspoň jedného ošetrenia sa líši od ostatných dvoch efektov.

V predchádzajúcom teste sa dáta testovali ako jeden celok a zistilo sa, že niektoré nastavenia vstupných premenných by mohli mať vplyv na kvalitu papiera. Môžeme preto položiť prirodzené otázky: Ktorá vstupná premenná z X_1, X_2, X_3 má najväčší vplyv na 13 pozorovaných ukazovateľov? Všetky vysvetľujúce premenné majú štatisticky významný vplyv alebo len niektoré z nich? Odpoveď na tieto otázky nám dá tzv. test významnosti vplyvu vysvetľujúcich premenných (pozri [8], [9]), ktorý tiež vykonáme v softvéri R pomocou balíka `car` [2] a príkazu `Anova`. Nulovou hypotézou testu je, že testovaná vstupná premenná nemá štatisticky významný vplyv na hodnoty výstupnej premennej (t. j. má nulový efekt na hodnoty výstupnej premennej).

```
Anova(model1, test.statistic="Wilks")
Anova(model1, test.statistic="Pillai")
Anova(model1, test.statistic="Hotelling-Lawley")
Anova(model1, test.statistic="Roy")
#
# výstupy týchto príkazov sú podobné, preto uvádzame len prvý z nich
# Type II MANOVA Tests: Wilks test statistic
#    Df test stat approx F num Df den Df   Pr(>F)
# X1  1  0.080818  11.3734     13     13 4.73e-05 ***
# X2  1  0.247041   3.0479     13     13  0.02717 *
# X3  1  0.283753   2.5242     13     13  0.05369 .
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Stĺpce výstupu sú skoro rovnaké, ako v prípade výstupu
# pri príkaze lht (viď vyššie).
```

Výsledok testu založený na Wilksovej testovej štatistike preukázal, že najvýznamnejšou vstupnou premennou je premenná X_1, pretože p-hodnota testu významnosti (uvedená v poslednom stĺpci výstupu) vyšla veľmi nízka ($0{,}0000473$), teda nulovú hypotézu o nulovom efekte premennej X_1 by sme zamietli na všetkých bežne používaných hladinách významnosti. Ďalej na hladine významnosti $0{,}05$ by aj premenná X_2 mohla mať štatisticky významný efekt na kvalitu papiera, pretože p-hodnota testu vyšla $0{,}02717 < 0{,}05$, teda nulovú hypotézu o nevýznamnom vplyve premennej X_2 by sme zamietli. Z výstupu tiež vidíme, že najslabší efekt na kvalitu papiera má premenná X_3, pri ktorej by sme nulovú hypotézu na hladine významnosti $0{,}05$ nezamietli (p-hodnota $= 0{,}05369 > 0{,}05$). Pre úplnosť dodávame, že tieto testy významnosti vplyvu vysvetľujúcich premenných by sme mohli vykonať aj pomocou

funkcie `manova`, ktorá je súčasťou základného štatistického balíka softvéru R. Použitie a výstup funkcie uvádzame ďalej.

```
m <- manova(model1)
summary(m)
#           Df  Pillai approx F num Df den Df   Pr(>F)
# X1         1 0.91918  11.3734     13     13 4.73e-05 ***
# X2         1 0.76630   3.2790     13     13  0.02048 *
# X3         1 0.71625   2.5242     13     13  0.05369 .
# Residuals 25
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Stĺpce výstupu sú skoro rovnaké, ako v prípade výstupu
# pri príkaze lht (viď vyššie).
```

Výsledok tohto testu pri použití Pillaiovej testovej štatistiky vyšiel veľmi podobne ako v predchádzajúcom výstupe. Môžeme teda zhrnúť, že najvýznamnejšou premennou je prvá vstupná premenná výrobného procesu, kým zvyšné dve vstupné nastavenia majú pravdepodobne len menší vplyv na výslednú kvalitu vyrábaného papiera.

5. ZÁVER

Viacrozmerná analýza rozptylu má široké spektrum využitia od lekárskej vedy cez poľnohospodárstvo až po priemyselné aplikácie. Jej najväčšou výhodou je, že komplexne, v rámci jediného modelu, zohľadňuje všetky vstupné a výstupné premenné, a to aj v prípade, keď sú viacrozmerné. V tomto článku sme zhrnuli najdôležitejšie teoretické poznatky o modeli viacrozmernej analýzy rozptylu a testovaní hypotéz v tomto modeli. Poukázali sme aj na možné slabé miesta testovania pomocou modelu MANOVA, najmä čo sa týka prípadného porušenia predpokladu normality kvantitatívnych vysvetľovaných premenných. V článku sme uviedli štyri štatistické testy vhodné na testovanie rovnosti vektorov stredných hodnôt a ich použitie sme ilustrovali reálnym príkladom z oblasti výroby papiera. Do riešenia ilustračnej úlohy sme doplnili aj test významnosti vplyvu vysvetľujúcich premenných v modeli MANOVA. Jedným z hlavných výstupov tohto príspevku je ukážka postupu testovania hypotéz v modeli MANOVA v prostredí softvéru R. Tento postup je do veľkej miery všeobecný, a preto sa dá použiť aj pri iných výskumoch a dátových súboroch.

Tento článok vznikol s podporou grantov VEGA 2/0047/15, VEGA 1/0251/16 a APVV-0465-12.

LITERATÚRA

[1] ALDRIN, M.: Moderate projection pursuit regression for multivariate response data. In: Computational Statistics & Data Analysis, Elsevier, 1996, No. 5, p. 501-531.

[2] FOX, J. – WEISBERG, S.: An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage, 2011. Dostupné na: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion [prístup k 24. 4. 2017].

[3] GONZALEZ-ESTRADA, E. – VILLASENOR-ALVA, J. A.: mvShapiroTest: Generalized Shapiro-Wilk test for multivariate normality. R package version 1.0,

2013. Dostupné na: https://CRAN.R-project.org/package=mvShapiroTest [prístup k 23. 11. 2016].
[4] HÄRDLE, W. K. – HLÁVKA, Z.: Multivariate Statistics: Exercises and Solutions. New York: Springer, 2007. 368 p. ISBN 978-0-387-73508-5.
[5] HÄRDLE, W. K. – SIMAR, L.: Applied Multivariate Statistical Analysis. Heidelberg: Springer, 2012. 516 p. ISBN 978-3-642-17229-8.
[6] HEINZ, G. – PETERSON, L. J. – JOHNSON, R. W. – KERK, C. J.: Exploring Relationships in Body Dimensions. In: Journal of Statistics Education, 2003, No. 2.
[7] IZENMAN, A. J.: Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. 1st Edition. New York: Springer, 2008. ISBN 978-0-387-78189-1.
[8] JOHNSON, R. A. – WICHERN, D. W.: Applied Multivariate Statistical Analysis. 6th Edition. Harlow: Pearson Education Limited, 2014. 770 p. ISBN 13: 978-1-292-02494-3.
[9] LAMOŠ, F. – POTOCKÝ, R.: Pravdepodobnosť a matematická štatistika: Štatistické analýzy. Bratislava: Univerzita Komenského, 1998. 343 s. ISBN 80-223-1262-2.
[10] MURA, L. – BULECA, J. – HAJDUOVA, Z. – ANDREJKOVIC, M.: Quantitative financial analysis of small and medium food enterprises in a developing country. In: Transformations in business & economics. 2015, No 1, p. 161-173.
[11] PILLAI, K. C. S.: On the Distribution of the Largest Root of a Matrix in Multivariate Analysis. In: Ann. Math. Statist., 1967, No. 2, p. 616-617.
[12] R CORE TEAM: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2016. Dostupné na: https://www.R-project.org/ [prístup k 10. 4. 2017].
[13] RENCHER, A. C.: Multivariate Statistical Inference and Applications. New York: Wiley, 1998. 592 p. ISBN: 978-0-471-57151-3.
[14] SVETLÍKOVÁ, B.: Rôzne spôsoby testovania v MANOVA. Bratislava: Fakulta matematiky, fyziky a informatiky Univerzity Komenského v Bratislave, 2015.
[15] TIMM, N. H.: Multivariate Analysis with Applications in Education and Psychology. Wadsworth, 1975.
[16] VILLASENOR-ALVA, J. A. – GONZALEZ-ESTRADA, E.: A generalization of Shapiro-Wilk's test for multivariate normality. In: Communications in Statistics: Theory and Methods, 2009, No. 11, p. 1870-1883.
[17] YANG, Q. – WANG, Y.: Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies. In: Journal of Probability and Statistics, 2012.
[18] WANG, X. – MELESSE, A. M. – YANG, W.: Development of a Multivariate Regression Model for Soil Nitrate Nitrogen Content Prediction. In: Journal of Spatial Hydrology, 2006, No. 2.

RESUME

Multivariate analysis of variance (MANOVA) has a wide range of applications, from medical science through agriculture to industrial applications. The main advantage of MANOVA is the simultaneous consideration of all input variables, even if the variables follow a complex multivariate distribution. This paper presents the theoretical background of the univariate and multivariate analysis of variance and hypothesis testing in these models. We have also pointed out possible limitations of MANOVA, especially the violation of normality assumption of explanatory variables of the linear regression model. To test the equality of mean vectors the following four

test statistics were used: Wilk's lambda, Pillai's trace, Hotelling-Lawley's trace and Roy's largest eigenvalue. To illustrate the usage of these test statistics, a research study from the paper industry was mentioned. The significance test of the impact of explanatory variables in the MANOVA model was carried out as well. The main contribution of this paper is to demonstrate the MANOVA model construction and testing of hypotheses in R software. Our approach is general enough to be applied to other studies and datasets as well.

PROFESIJNÝ ŽIVOTOPIS

***Mgr. Gábor Szűcs, PhD.**, vyštudoval pravdepodobnosť a matematickú štatistiku na Fakulte matematiky, fyziky a informatiky Univerzity Komenského v Bratislave a podnikové hospodárstvo a manažment na Univerzite J. Selyeho v Komárne. Doktorandské štúdium absolvoval na Fakulte matematiky, fyziky a informatiky Univerzity Komenského v odbore aplikovaná matematika a od roku 2015 pôsobí na tejto fakulte ako odborný asistent. Vo svojej výskumnej činnosti sa venuje oceňovaniu poistných dôchodkov, skúmaniu rozdelení pravdepodobnosti v neživotnom poistení a ďalším oblastiam aplikovanej štatistiky.*

KONTAKT

Gabor.Szucs@fmph.uniba.sk