Jozef Kušnier

GEOMETRICKÁ INTERPRETÁCIA MANOVA MODELU

Úvod

Analýza rozptylu (ANOVA) je základná štatistická analýza, ktorá zovšeobecňuje testy porovnania stredných hodnôt dvoch skupín na porovnanie stredných hodnôt viacerých skupín. Má uplatnenie v rôznych oblastiach ekonomického bádania. V behaviorálnej ekonómii sa často vykonávajú kontrolované experimenty s cieľom zistiť, ako nastaviť ekonomické prostredie tak, aby to viedlo k lepším ekonomickým výsledkom. Napríklad môžeme skúmať vplyv finančného odmeňovania a rôznych foriem nefinančných odmien na pracovné výsledky. V tomto prípade sú jednotlivé testované formy odmien úrovňami faktora, ktoré môžu ovplyvniť závislú premennú. Vplyvy iných potenciálnych faktorov sú ošetrené randomizáciou experimentu. Teda pri veľkých vzorkách očakávame približne rovnaké úrovne iných vysvetľujúcich faktorov a jednotlivé faktory nemusia byť ani známe. Podobné experimenty sú však v praxi často nemožné, napríklad vplyv pohlavia na výšku finančnej odmeny za prácu nemožno týmto spôsobom testovať, keďže pohlavie nemožno randomizovať a teda nedá sa náhodne priradiť subjektu v experimente. V iných oblastiach ekonómie sme preto odkázaní na observačné štúdie, pri ktorých sa pasívne sledujú jednotlivé premenné. Avšak aj v observačných štúdiach nachádza ANOVA model uplatnenie ako základný, východiskový model z ktorého sú odvodené zložitejšie modely, ktoré zohľadňujú aj vplyv iných premenných na závislú premennú, ako napríklad analýza kovariancie, lineárna regresia, zovšeobecnené lineárne modely.

Podobne je ANOVA dôležitá v teórii náhodných výberov. V praxi sme z finančných a časových dôvodov prakticky odkázaní na výberové šetrenia základných súborov. Pri práci s údajmi z výberu vzniká nevyhnutne tzv. výberová chyba z dôvodu neúplnosti údajov. Vhodným spôsobom navrhnutia realizácie výberu však vieme ovplyvňovať veľkosť tejto chyby. Stratifikácia výberu je základná technika, ktorá vedie k znižovaniu výberovej chyby a ktorá je založená na ANOVA analýze. Podobne zhlukovanie výberových jednotiek možno ilustrovať ANOVA modelom s náhodnými efektami.

Keďže ANOVA ma také centrálne postavenie v rôznych oblastiach štatistickej inferencie v ekonomickej praxi, ukážeme v tomto článku jej goemetrickú interpretáciu v priestore pozorovaní. Ukážeme to však aj pre viacrozmernú analýzu rozptylu (MANOVA), pri ktorej sú použité vzorce ešte náročnejšie na pochopenie. Pri MANOVA navyše ukážeme aj geometrickú interpretáciu v priestore premenných. Tieto MANOVA geometrické predstavy vedú k lepšiemu porozumeniu vzorcov aj v zložitejších viacrovnicových modeloch v ekonometrii. Výpočet zrealizujeme na konkrétnom príklade bez použitia špeciálnych štatistických softvérov, jednoducho iba za pomoci MS Excelu.

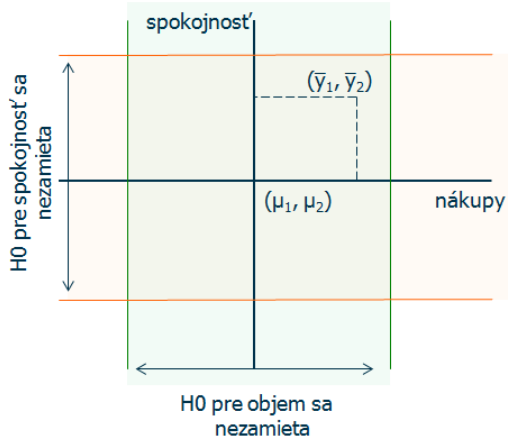
1 PRÍPADOVÁ ŠTÚDIA - OBCHODNÝ REŤAZEC

Obchodný reťazec chce zmerať výkonnosť jednotlivých predajní a identifikovať veľmi výkonné a málo výkonné predajne. Kedže podnikanie reťazca je motivované tvorbou zisku, tak ho prirodzene zaujíma veľkosť obratu v danej predajni. Ale reťazec myslí aj na budúcnosť a vie, že jednorázový vysoký obrat mu nemusí garantovať budúce vysoké obraty. Preto sa zaujíma aj o lojalitu zákazníkov, ich spokojnosť a aký imidž má vo verejnosti, teda svoju reputáciu. Urobí si preto prieskum spokojnosti na vzorke svojich zákazníkov. My sa pre jednoduchosť zameriame na dva ukazovateľe: spokojnosť zákazníka s predajňou a objem jeho nákupov v danej predajni za sledovaný mesiac. Spokojnosť je číslo od 0 do 100, je to index spokojnosti vyrátaný z prieskumu spokojnosti. Objem nákupov je v eurách, je to suma nákupov zo zákazníckej karty. Naša závislá premenná je teda dvojrozmerná a zaujímajú nás rozdiely medzi predajňami. Budeme ich testovať pomocou MANOVA modelu.

1.1 Prečo viacrozmerný model?

Predstavme si, že by sme testovali veľkosť strednej hodnoty pre každú závislú premennú zvlášť. Množina akceptácie nulovej hypotézy je interval pre obe testovacie štatistiky. Preto je oblasť akceptácie obdĺžnik v grafe na obrázku 1. Na osi x máme objem nákupov, na osi y spokojnosť. Graf je vycentrovaný do bodu testovanej nulovej hypotézy (μ_1, μ_2). Testovaná predajňa je zobrazená bodom so súradnicami rovnajúcimi sa výberovému priemeru objemu nákupov a spokojnosti $(\bar{y}_1,\bar{y}_2)$.

Obrázok 1: Obdĺžnik nezamietania nulovej hypotézy
(Zdroj: vlastné spracovanie)

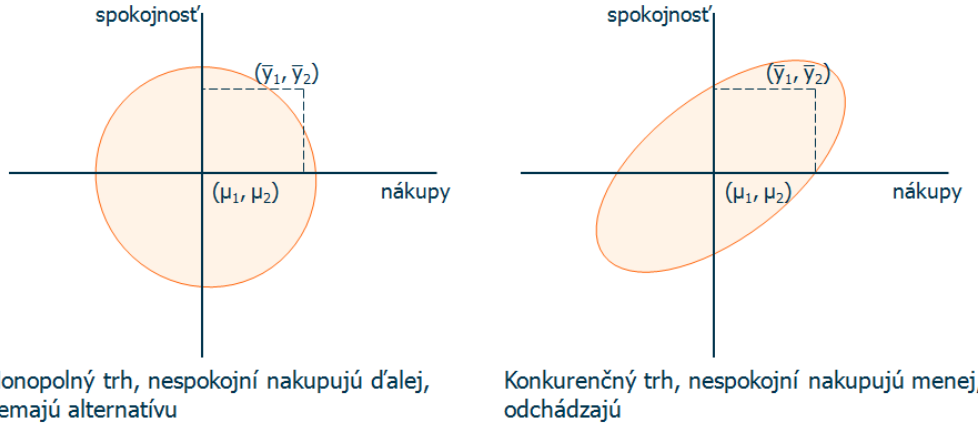


Ak označíme α hladinu významnosti pre každý test, tak za platnosti nulovej hypotézy je bod reprezentujúci testovanú predajňu v oblasti akceptácie s pravdepodobnosťou $(1-\alpha)(1-\alpha) \cong 1-2\alpha$. Teda oblasť zamietania má pravdepodobnosť približne 2α a chyba prvého druhu je zvýšená. Toto zvýšenie chyby prvého druhu je tým

väčšie, čím väčšia je dimenzia závislej premennej, ktorá je ignorovaná jednorozmerným prístupom. Preto aj používanie viacerých jednorozmerných ANOVA testov namiesto jedného MANOVA testu vedie k zvýšeniu chyby prvého druhu.

Navyše používanie viacerých jednorozmerných ANOVA modelov ignoruje korelácie medzi závislými premennými a teda nemôže zodpovedať na otázky ohľadom vzťahu medzi lineárnymi kombináciami závislých premenných a vysvetľujúcimi faktormi. Korelácia medzi závislými premennými je modelovaná MANOVA modelom a ovplyvňuje výsledok testovania. Oblasť nezamietnutia nulovej hypotézy závisí od združenej distribúcie závislých premenných a teda tie isté marginálne empirické distribúcie môžu viesť k rozličným záverom ako vidno na obrázku 2.

Obrázok 2: Závislosť oblasti nezamietania H0 od korelačnej štruktúry
(Zdroj: vlastné spracovanie)



2 PRIESTOR POZOROVANÍ

Budeme sa teraz venovať samotnej geometrickej interpretácii analýzy rozptylu. Začneme najskôr so zakladnými pojmami ako priemer, variancia a korelácia.

Nech $y_1, y_2, \ldots, y_n$ je realizácia náhodného výberu, kde n je počet pozorovaní. Označme stĺpcový vektor $\boldsymbol{y}=(y_1,y_2,\ldots,y_n)^T$. Môžeme si ho predstaviť ako bod v n-rozmernom euklidovskom priestore. Na obrázku 3 je zobrazený príklad pre výber rozsahu 3, teda pre 3-rozmerný priestor pozorovaní.

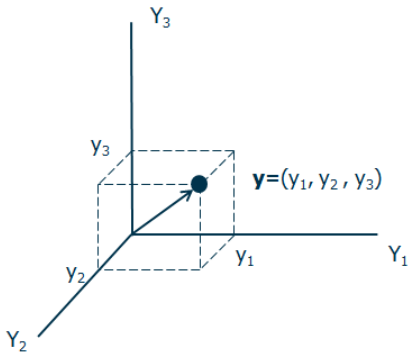
Pre úplnosť, v euklidovskom priestore sú definované nasledovné funkcie:

skalárny súčin dvoch vektorov **x** a **y** $\quad \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{n} \boldsymbol{x_i}\, \boldsymbol{y_i} = \boldsymbol{x^T y},$

norma vektora **x** $\quad \|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = \sqrt{\sum_{i=1}^{n} \boldsymbol{x_i^2}} = \sqrt{\boldsymbol{x^T x}},$

kosínus uhla medzi vektormi **x** a **y** $\quad \boldsymbol{\cos}(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle / (\|\boldsymbol{x}\| \|\boldsymbol{y}\|).$

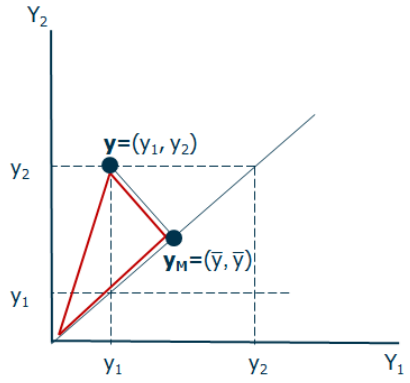
Obrázok 3: Náhodný výber v 3-rozmernom priestore pozorovaní
(Zdroj: vlastné spracovanie)



2.1 Priemer ako súradnice ortogonálnej projekcie na diagonálu

Označme diagonálnu, teda vektor pozostávajúci zo samých *1*, ako $\mathbf{1}=(1,1,...,1)^T$. Označme ortogonálnu projekciu realizácie náhodného výberu $\boldsymbol{y}$ na diagonálu $\mathbf{1}$ ako $\boldsymbol{y}_M$. Pre 2-rozmerný prípad je to zobrazené na obrázku 4.

Obrázok 4: Ortogonálna projekcia na diagonálu
(Zdroj: vlastné spracovanie)



Z definície ortogonálnej transformácie dostávame pravouhlý trojuholník s vrcholmi v bodoch *(0,0)*, (y_1,y_2), $\boldsymbol{y}_M$ a pravým uhlom pri vrchole $\boldsymbol{y}_M$. Súradnice ortogonálnej projekcie $\boldsymbol{y}$ na diagonálu $\mathbf{1}$, teda $\boldsymbol{y}_M$, sú priemerom náhodného výberu $\boldsymbol{y}$. Môžeme to nasledovne zrátať:

$$\boldsymbol{y_M} = \frac{\langle \boldsymbol{y}, \mathbf{1} \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle} \mathbf{1} = \bar{y}\mathbf{1} = (\bar{y}, \bar{y})^T$$

2.2 Smerodajná odchýlka ako dĺžka centrovaného vektora

Označme $\boldsymbol{y}_C$ reziduálnu zložku realizácie náhodného výberu po ortogonálnej transformácii na diagonálu, teda $\boldsymbol{y}_C = \boldsymbol{y} - \boldsymbol{y}_M$. Z Pytagorovej vety dostávame

$$\|\boldsymbol{y}\|^2 = \|\boldsymbol{y_M}\|^2 + \|\boldsymbol{y_C}\|^2$$

a keďže $\|\boldsymbol{y}\|^2 = \sum_i \boldsymbol{y_i^2}$, $\|\boldsymbol{y_M}\|^2 = \boldsymbol{n\bar{y}^2}$, $\|\boldsymbol{y_C}\|^2 = \sum_i(\boldsymbol{y_i} - \boldsymbol{\bar{y}})^2$, tak dostávame

$$\sum_i \boldsymbol{y_i^2} = \sum_i(\boldsymbol{y_i} - \boldsymbol{\bar{y}})^2 + \boldsymbol{n\bar{y}^2}$$

teda známy vzťah, ktorý sa odvádza algebraicky v základných kurzoch štatistiky:
$var(Y) = E(Y^2) - E^2(Y)$.

Všimnime si, že dĺžka reziduálneho vektora $\boldsymbol{y}_C$ je úmerná výberovej smerodajnej odchýlke realizácie náhodného výberu $\boldsymbol{y}$, teda

$$\|\boldsymbol{y_C}\|^2 = \|\boldsymbol{y} - \boldsymbol{\bar{y}1}\|^2 = \sum_i(\boldsymbol{y_i} - \boldsymbol{\bar{y}})^2 \approx \boldsymbol{sample.var(y)}$$

V štatistike nás zaujíma premenlivosť, teda takéto odchýlky od ukazovateľa centrálnej tendencie, preto v analýzach vychádzame zo sumy štvorcov odchýlok od priemeru a nie zo samotných súm štvorcov. Ak by sme nepozorovali žiadnu premenlivosť, teda všetky naše namerané dáta by boli konštantné, tak štatistika nemá čo vysvetliť.

Avšak suma švorcov a teda aj veľkosť vektorov rastie s dimenziou priestoru, v ktorom sa pohybujeme. Dokonca ak nepozorujeme žiadnu premenlivosť v dátach, teda máme konštantné pozorovania, tak pridaním ďalšieho konštantného pozorovania narastie suma štvorcov, teda dĺžka vektora. Možno si to ukázať na konštantnom n-rozmernom vektore $\boldsymbol{1}_n$. Jeho suma štvorcov je n, dĺžka je $n^{1/2}$. Ak pridáme ďalšie pozorovanie a uvažujeme $n+1$ rozmerný vektor $\boldsymbol{1}_{n+1}$, tak suma štorcov narastie na $n+1$ a dĺžka na $(n+1)^{1/2}$. Preto je vhodné sumy štvorcov a teda dĺžky vektorov normalizovať na veľkosť dimenzie. Dostávame takzvané stredné sumy štvorcov, lebo ich normalizujeme dĺžkou konštantného jednotkového vektoru v danom priestore.

Tieto úvahy vedú k zavedeniu počtu stupňov voľnosti ako počtu dimenzií, s ktorými pracujeme. Napríklad reziduálny vektor $\mathbf{y}_C$, teda centrovaný pôvodný vektor $\mathbf{y}$, vznikol po ortogonálnej projekcii na diagonálu. Teda z jeho definície je otogonálny na diagonálu a teda sa vyskytuje v nadrovine s normálou diagonálou. Preto sa vyskytuje v priestore s dimenziou $n-1$ a je vhodné ho normovať číslom $n-1$ a nie číslom n. Potom však už rozklad štvorcov neplatí, ale takéto normovanie je vhodnejšie na testovanie hypotéz, keďže očisťuje testovacie štatistiky od efektu veľkosti dimenzie.
Podobne, ak je model zložitejší, ako uvidíme v prípade ANOVA, tak piestor stredných hodnôt modelovaných modelom má viac ako jednu dimenziu a preto reziduálna zložka leží v priestore s dimenziou zmenšenou o dimenziu modelu.

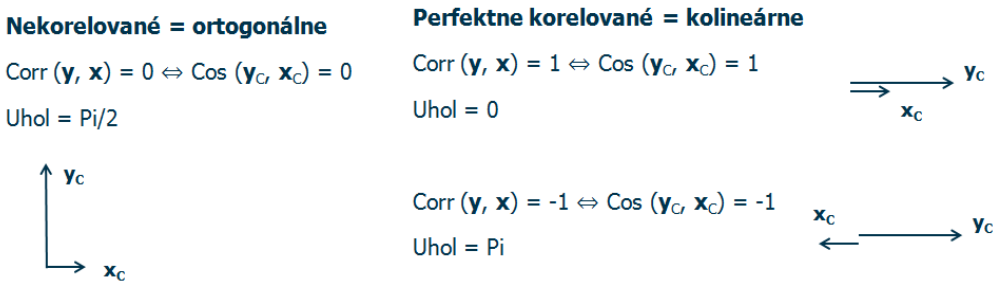
2.3 Korelácia ako kosínus centrovaných vektorov

Označme realizácie dvoch náhodných výberov ako $\boldsymbol{x}=(x_1, x_2, ..., x_n)^T$ a $\boldsymbol{y}=(y_1, y_2, ..., y_n)^T$. Potom priamo z definície korelácie a kosínusu vyplýva nasledovná rovnosť:

$$\boldsymbol{sample.corr(x,y)} = \sum_i(x_i-\bar{x})(y_i-\bar{y})/\sqrt{\sum_i(x_i-\bar{x})^2\sum_i(y_i-\bar{y})^2} = \\ = \langle \boldsymbol{x}-\bar{x}\boldsymbol{1}, \boldsymbol{y}-\bar{y}\boldsymbol{1}\rangle/\|\boldsymbol{x}-\bar{x}\boldsymbol{1}\|\|\boldsymbol{y}-\bar{y}\boldsymbol{1}\| = \boldsymbol{cos(x_C, y_C)}$$

Teda hovoríme o ortgonálnych a kolineárnych nahodných premenných tak, ako je to zobrazené na obrázku 5.

Obrázok 5: Korelácia ako kosínus
(Zdroj: vlastné spracovanie)



2.4 ANOVA algebraicky

Pripomenieme si teraz ANOVA analýzu algebraicky. Model má tvar

$$\boldsymbol{Y_{ij} = \mu_j + \sigma\varepsilon_{ij}}$$ kde $i=1,...,n_j$ $j=1,...,J$ $\varepsilon_{ij} \sim N(0,1)$ navzájom nezávislé.

Teda pozorujeme náhodné premenné Y_{ij}, ktorých variabilitu chceme vysvetliť pomocou kategoriálnej premennej, faktora s J úrovňami. Index j označuje úroveň faktora alebo skupinu, z ktorej pochádza pozorovanie Y_{ij}. Index i označuje i-té pozorovanie vrámci j-tej skupiny. Zaujíma nás, či sú jednotlivé stredné hodnoty μ_j rovnaké, teda konštantné pre každú skupinu, alebo nie a teda prispievajú k variabilite vysvetľovanej premennej.

Algebraická dekompozícia celkovej sumy štvorcov TSS pre konkrétnu realizáciu stratifikovaného náhodného výberu je nasledovná:

$$\boldsymbol{TSS} = \sum_j\sum_i\left(y_{ij}-\bar{y}_{..}\right)^2 = \sum_j\sum_i((y_{ij}-\bar{y}_{.j})-(\bar{y}_{.j}-\bar{y}_{..}))^2 = \sum_j\sum_i\left(y_{ij}-\bar{y}_{.j}\right)^2 + \\ 2\sum_j\sum_i(y_{ij}-\bar{y}_{.j})(\bar{y}_{.j}-\bar{y}_{..}) + \sum_j\sum_i(\bar{y}_{.j}-\bar{y}_{..})^2 = \boldsymbol{RSS + 0 + ESS}$$

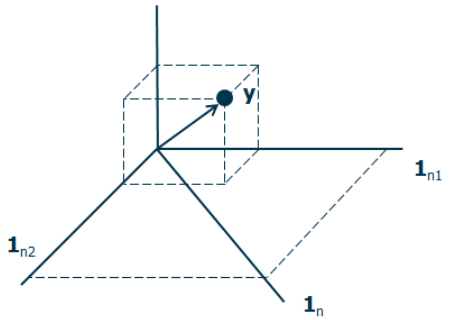
Teda $TSS = RSS + ESS$, kde
RSS – reziduálna suma štvorcov – nevysvetlená – vnútri skupín,
ESS – modelová suma štvorcov – vysvetlená – medzi skupinami.

Opäť nám to pripomína Pytagorovu vetu, keď sme využili vlastnosť priemeru, že suma odchýlok od priemeru je nulová, teda diagonála je kolmá na centrovaný vektor. Ukážeme si to teraz geometricky.

2.5 ANOVA geometricky

Aby sme si vedeli vytvoriť geometrickú predstavu, budeme pracovať iba s dvomi úrovňami faktora, ktoré majú početnosti n_1 a n_2 pozorovaní. Budeme opäť indexovať iba jedným indexom, tak ako pri predošlých geometrických úvahach. Teda máme $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ vektor realizácií, označme ďalej
$\boldsymbol{1}_{n1} = (1,\ldots,1,0,\ldots,0)^T$ vetor n_1 jednotiek zodpovedajúcich prvej úrovni faktora
$\boldsymbol{1}_{n2} = (0,\ldots,0,1,\ldots,1)^T$ vetor n_2 jednotiek zodpovedajúcich druhej úrovni faktora.
Model má tvar: $\boldsymbol{Y} = (\boldsymbol{1}_{n1}, \boldsymbol{1}_{n2})\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, kde $\boldsymbol{\beta} = (\mu_1, \mu_2)^T$ s obvyklými predpokladmi.
Na obrázku 6 sú zobrazené tieto 3 vektory v priestore pozorovaní.

Obrázok 6: ANOVA priestor pozorovaní
(Zdroj: vlastné spracovanie)



Je zrejmé, že $\boldsymbol{1}_{n1}$ a $\boldsymbol{1}_{n2}$ sú navzájom kolmé, kedže $\boldsymbol{1}_{n1}^T\boldsymbol{1}_{n2} = 0$. Ďalej je zrejmé, že diagonála $\boldsymbol{1}_n$ je súčtom $\boldsymbol{1}_{n1}$ a $\boldsymbol{1}_{n2}$, teda $\boldsymbol{1}_n = \boldsymbol{1}_{n1} + \boldsymbol{1}_{n1}$.

2.5.1 Projekcia na model

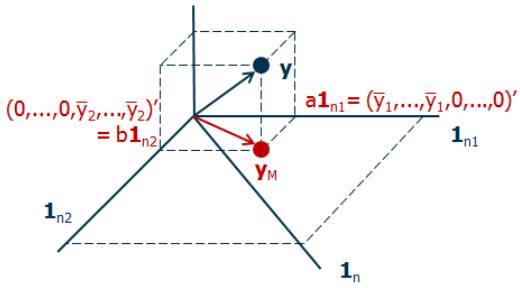
Označme opať $\boldsymbol{y}_M$ ortogonálnu projekciu $\boldsymbol{y}$ na rovinu generovanú vektormi $\boldsymbol{1}_{n1}$ a $\boldsymbol{1}_{n2}$, teda $\boldsymbol{y}_M = a\boldsymbol{1}_{n1} + b\boldsymbol{1}_{n2}$, kde a,b sú neznáme koeficienty. Ďalej označme reziduálny vektor $\boldsymbol{y}_R = \boldsymbol{y} - \boldsymbol{y}_M$ a zobrazíme situáciu na obrázku 7.

Ukážeme, že $a=\bar{y}_1$ a $b=\bar{y}_2$. Projekcia $\boldsymbol{y}$ na $\boldsymbol{1}_{n1}$ je

$$\bar{\boldsymbol{y}}_1 = \langle \boldsymbol{y}, \boldsymbol{1}_{n1}\rangle / \langle \boldsymbol{1}_{n1}, \boldsymbol{1}_{n1}\rangle = (\langle \boldsymbol{y}_R, \boldsymbol{1}_{n1}\rangle + \langle \boldsymbol{y}_M, \boldsymbol{1}_{n1}\rangle)/\boldsymbol{n}_1 =$$
$$(\langle \boldsymbol{y}_R, \boldsymbol{1}_{n1}\rangle + \boldsymbol{a}\langle \boldsymbol{1}_{n1}, \boldsymbol{1}_{n1}\rangle + \boldsymbol{b}\langle \boldsymbol{1}_{n2}, \boldsymbol{1}_{n1}\rangle)/\boldsymbol{n}_1 = (\boldsymbol{0} + \boldsymbol{a}\boldsymbol{n}_1 + \boldsymbol{0})/\boldsymbol{n}_1 = \boldsymbol{a}$$

Podobne pre $b=\bar{y}_2$ a preto $\boldsymbol{y}_M = \bar{y}_1\boldsymbol{1}_{n1} + \bar{y}_2\boldsymbol{1}_{n2}$.

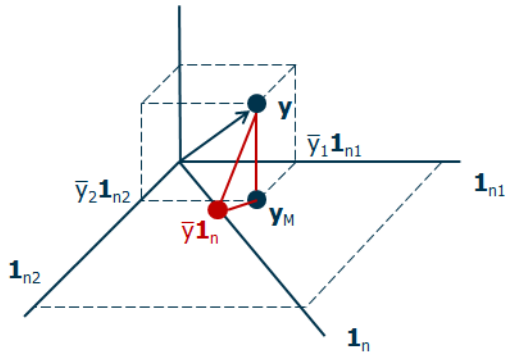
Obrázok 7: ANOVA projekcia na priestor stredných hodnôt
(Zdroj: vlastné spracovanie)



2.5.2 *ANOVA pravouhlý trojuholník*

Opäť projekcia $\boldsymbol{y}$ na diagonálu je $\bar{y}\boldsymbol{1_n}$. Označme centrovaný vektor $\boldsymbol{y_C} = \boldsymbol{y} - \bar{y}\boldsymbol{1_n}$. Zobrazíme to na obrázku 8.

Obrázok 8: ANOVA pravouhlý tojuholník
(Zdroj: vlastné spracovanie)



Dostaneme pravouhlý trojuholník s preponou $\boldsymbol{y_C} = \boldsymbol{y} - \bar{y}\boldsymbol{1_n}$ a odvesnami $\boldsymbol{y_R} = \boldsymbol{y} - \boldsymbol{y_M}$ a $\boldsymbol{y_M} - \bar{y}\boldsymbol{1_n}$. Podľa Pytagora dostaneme:

$$\|\boldsymbol{y_C}\|^2 = \|\boldsymbol{y_M} - \bar{\boldsymbol{y}}\boldsymbol{1_n}\|^2 + \|\boldsymbol{y_R}\|^2$$

Teda $\sum_j \sum_i (\boldsymbol{y_{ij}} - \bar{\boldsymbol{y}}_{..})^2 = \sum_j \boldsymbol{n_j} (\bar{\boldsymbol{y}}_{.j} - \bar{\boldsymbol{y}}_{..})^2 + \sum_j \sum_i (\boldsymbol{y_{ij}} - \bar{\boldsymbol{y}}_{.j})^2$

2.5.3 *ANOVA tabuľka ako zobrazenie Pytagorovej vety*

Z geometrickej interpretácie vyplýva, že tzv. ANOVA tabuľka je spôsob zobrazenia Pytagorovej vety. ANOVA tabuľka udáva rozklad sumy štvorcov nasledovne:

TSS: celková *SS*, variabilita, ktorú chceme vysvetliť
Dimenzia/Stupne voľnosti: *n-1* (nadrovina s normálou diagonálou)

$$\boldsymbol{TSS} = \|\boldsymbol{y_C}\|^2 = \sum_j \sum_i \left(\boldsymbol{y_{ij}} - \overline{\boldsymbol{y}}_{..}\right)^2$$

ESS: *SS* vysvetlená modelom
Dimenzia/Stupne voľnosti: *J-1* (priestor vektorov v priestore modelu kolmých na diagonálu)

$$\boldsymbol{ESS} = \|\boldsymbol{y_M} - \overline{\boldsymbol{y}}\boldsymbol{1_n}\|^2 = \sum_j \boldsymbol{n_j}\left(\overline{\boldsymbol{y}}_{.j} - \overline{\boldsymbol{y}}_{..}\right)^2$$

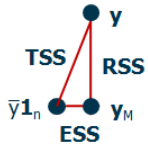
RSS: nevysvetlená *SS*
Dimenzia/Stupne voľnosti: *n-J* (priestor vektorov kolmých na priestor modelu)

$$\boldsymbol{RSS} = \|\boldsymbol{y_R}\|^2 = \sum_j \sum_i \left(\boldsymbol{y_{ij}} - \overline{\boldsymbol{y}}_{.j}\right)^2$$

2.5.4 F-test a t-test ako tangens

Sumy štvorcov v ANOVA tabuľke majú χ^2 rozdelenie, preto aj jednotlivé členy Pytagorovej vety majú toto rozdelenie. Chceme testovať model, teda či model vysvetľuje variabilitu v závislej premennej, v našom prípade či sa stredné hodnoty líšia pre jednotlivé skupiny alebo nie. Nulová testovaná hypotéza je, že stredné hodnoty sú rovnaké medzi skupinami, teda že model nevysvetľuje variabilitu závislej premennej. Z geometrickej interpretácie je jasné, že v prípade rovnakých stredných hodnôt je vysvetlená suma štvorcov *ESS* relatívne malá vzhľadom k reziduálnej sume *RSS*. Je nenulová iba v dôsledku náhody v náhodnom výbere. Preto je aj pomer *ESS/RSS* malý, ako vidno z trojuholníka na obrázku 9.

Obrázok 9: ANOVA pravouhlý tojuholník detailne pre F-test
(Zdroj: vlastné spracovanie)



Ak sumy štvorcov očistíme od vplyvu veľkosti dimenzií, v ktorých sa nachádzajú, tak dostávame tzv. F-rozdelenie. Teda v našom prípade s *J=2* skupinami je

$$F = (n\text{-}2)\, ESS/RSS \sim F(1,n\text{-}2) \sim t(n\text{-}2)^2.$$

Preto je *t*-štatistika úmerná tangensu uhla medzi vektormi $\boldsymbol{y}_R$ a $\boldsymbol{y}_C$. Presne

$$t = (n\text{-}2)^{1/2}\, tan(\boldsymbol{y}_R, \boldsymbol{y}_C).$$

Všeobecne H0 zamietame, ak $F > F(J-1,n-J,\alpha)$, kde α je hranica významosti testu. Teda F si možno predstaviť ako štvorec tangensu uhla z pravouhlého trojuholníka vzniknutého z ortogonálnej projekcie na priestor stredných hodnôt z modelu.

3 MANOVA GEOMETRICKY

Ukážeme si teraz geometrickú interpretáciu viacrozmernej analýzy rozptylu. Okrem priestoru pozorovaní zobrazíme aj priestor premenných. Začne najskôr s formálnou definíciou modelu.

Uvažujme teda viac vysvetľovaných premenných, označme ich ${}^1Y, {}^2Y, ..., {}^pY$. Každú modelujeme ANOVA modelom. Ale navyše uvažujeme aj o koreláciach medzi nimi. Teda pre každú úroveň faktora j modelujeme p-rozmerný náhodný vektor nasledovne:

$$({}^1\boldsymbol{Y}, {}^2\boldsymbol{Y}, \dots, {}^p\boldsymbol{Y}) = ({}^1\boldsymbol{\mu}, {}^2\boldsymbol{\mu}, \dots, {}^p\boldsymbol{\mu}) + \boldsymbol{\varepsilon}^T,$$ kde $\boldsymbol{\varepsilon} \sim N_p(0,\Sigma)$ a Σ nemusí byť diagonálna.

Pre konkrétnu realizáciu vektorového náhodného výberu sú MANOVA odhady $\widehat{{}^k\boldsymbol{\mu}_j} = {}^k\overline{\boldsymbol{y}}_j$, teda tie isté ako pri ANOVA. Pri ANOVA boli získané minimalizáciou reziduálneho vektora, teda MNŠ metódou, ortogonálnou projekciou. Tu však máme viacero reziduálnych vektorov, ktoré môžu byť korelované. Teda namiesto reziduálnej sumy štvorcov máme tzv. reziduálnu maticu sumy štvorcov a sumy súčinov.

Ukážeme, že pri použití ANOVA odhadov strednej hodnoty

- maticu sumy štvorcov a sumy súčinov možno dekomponovať na modelovú maticu a reziduálnu maticu
- MANOVA reziduálna matica je "minimálna" v istom zmysle.

3.1 MANOVA matice

Označme:

- $n \times p$ maticu realizácií Y, tj. $Y = ({}^1\boldsymbol{y}, {}^2\boldsymbol{y}, \dots, {}^p\boldsymbol{y})$,
- $n \times 1$ vektor pozorovaní zodpovedajúce j-tej úrovni faktora $\boldsymbol{1}_{nj}$,
- $n \times J$ maticu X všetkých úrovní faktora, tj. $X = (\boldsymbol{1}_{n1}, \boldsymbol{1}_{n2}, \dots, \boldsymbol{1}_{nJ})$.

Potom model má tvar $Y = XM + E$, kde

- M je $J \times p$ matica neznámych stredných hodnôt,
- E je $n \times p$ matica realizácie nahodného výberu vektora $\sim N_p(0,\Sigma)$.

Ďalej označme:

- centrovanú maticu $Y_C = Y - \boldsymbol{1}_n({}^1\overline{y}, {}^2\overline{y}, \dots, {}^p\overline{y})$,
- modelovú maticu $\boldsymbol{Y_M} = \boldsymbol{X}\widehat{\boldsymbol{M}}$, kde $\widehat{\boldsymbol{M}}$ je matica priemerov, tj. MNŠ odhadov stredných hodnôt,
- reziduálnu maticu $Y_R = Y - Y_M$.

Potom matice súčtov štvorcov a súčtov súčinov sú:

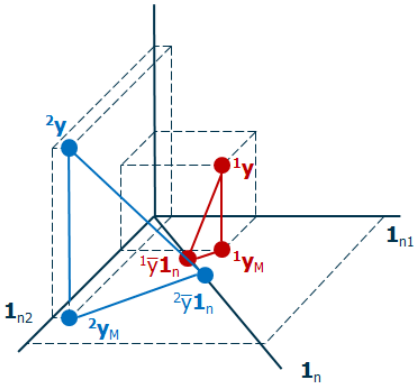
- T matica p x p celkového súčtu štvorcov a produktov, ktorú chceme modelom vysvetliť, $T = Y_C^T Y_C = (Y - \mathbf{1}_n({}^1\bar{y}, {}^2\bar{y}, \ldots, {}^p\bar{y}))^T (Y - \mathbf{1}_n({}^1\bar{y}, {}^2\bar{y}, \ldots, {}^p\bar{y}))$
- B matica p x p súčtov štvorcov a produktov vysvetlených modelom, medzi skupinami matica $B = (Y_M - \mathbf{1}_n({}^1\bar{y}, {}^2\bar{y}, \ldots, {}^p\bar{y}))^T (Y_M - \mathbf{1}_n({}^1\bar{y}, {}^2\bar{y}, \ldots, {}^p\bar{y}))$
- W maticu p x p reziduálnych súčtov štvorcov a produktov, vnútri skupín matica $W = Y_R^T Y_R = (Y - Y_M)^T (Y - Y_M)$.

Ukážeme, že obdoba Pytagorovej vety platí aj pre tieto matice, teda že $T = W + B$.

3.2 MANOVA rozklad matice súčtov štvorcov a súčtov súčinov

Majme teda viac závislých premenných. Pre jednoduchosť dve premenné a ich realizácie vo vektoroch ${}^1\boldsymbol{y}$ a ${}^2\boldsymbol{y}$. Každá premenná izolovane má svoj ANOVA model, teda máme 2 pravouhlé trojuholníky s rozkladom sumy štvorcov. Zobrazené sú na obrázku 10.

Obrázok 10: MANOVA dva pravouhlé tojuholníky
(Zdroj: vlastné spracovanie)



Rozložíme teraz aj sumu súčinov:

$$\begin{aligned}\langle {}^1\boldsymbol{y_C}, {}^2\boldsymbol{y_C}\rangle &= \langle {}^1\boldsymbol{y_R} + {}^1\boldsymbol{y_M} - {}^1\bar{\boldsymbol{y}}\mathbf{1_n}, {}^2\boldsymbol{y_R} + {}^2\boldsymbol{y_M} - {}^2\bar{\boldsymbol{y}}\mathbf{1_n}\rangle \\ &= \langle {}^1\boldsymbol{y_R}, {}^2\boldsymbol{y_R}\rangle + \langle {}^1\boldsymbol{y_R}, {}^2\boldsymbol{y_M} - {}^2\bar{\boldsymbol{y}}\mathbf{1_n}\rangle + \langle {}^2\boldsymbol{y_R}, {}^1\boldsymbol{y_M} - {}^1\bar{\boldsymbol{y}}\mathbf{1_n}\rangle \\ &+ \langle {}^1\boldsymbol{y_M} - {}^1\bar{\boldsymbol{y}}\mathbf{1_n}, {}^2\boldsymbol{y_M} - {}^2\bar{\boldsymbol{y}}\mathbf{1_n}\rangle \\ &= \langle {}^1\boldsymbol{y_R}, {}^2\boldsymbol{y_R}\rangle + \mathbf{0} + \mathbf{0} + \langle {}^1\boldsymbol{y_M} - {}^1\bar{\boldsymbol{y}}\mathbf{1_n}, {}^2\boldsymbol{y_M} - {}^2\bar{\boldsymbol{y}}\mathbf{1_n}\rangle \\ &= ({}^1\boldsymbol{y_R})^T({}^2\boldsymbol{y_R}) + ({}^1\boldsymbol{y_M} - {}^1\bar{\boldsymbol{y}}\mathbf{1_n})^T({}^2\boldsymbol{y_M} - {}^2\bar{\boldsymbol{y}}\mathbf{1_n})\end{aligned}$$

Teda $T = W + B$, keďže

$$Y_C^T Y_C = Y_R^T Y_R + (Y_M - \mathbf{1}_n({}^1\bar{y}, {}^2\bar{y}, \ldots, {}^p\bar{y}))^T (Y_M - \mathbf{1}_n({}^1\bar{y}, {}^2\bar{y}, \ldots, {}^p\bar{y})).$$

3.3 Minimálnosť MANOVA reziduálnej matice

Ukážeme si, že reziduálna matica $W = Y_R^T Y_R$ je v istom zmysle minimálna. V jednorozmernom prípade je minimálna suma štvorcov reziduií. Vo viacrozmernom prípade budeme požadovať, aby akákoľvek lineárna kombinácia reziduálnych vektorov mala minimálnu varianciu.

Ľubovoľná lineárna kombinácia $\boldsymbol{k}$ MANOVA reziduálnych vektorov je $Y_R\boldsymbol{k}$. Jej variancia je úmerná $\boldsymbol{k}^T Y_R^T Y_R \boldsymbol{k}$. Ukážeme, že pre maticu G ľubovoľných odhadov stredných hodnôt nie je variancia lineárnej kombinácie reziduií menšia ako pri MANOVA odhadoch, teda že

$$\boldsymbol{k}^T(\boldsymbol{Y}-\boldsymbol{XG})^T(\boldsymbol{Y}-\boldsymbol{XG})\boldsymbol{k} \geq \boldsymbol{k}^T(\boldsymbol{Y}-\boldsymbol{Y_M})^T(\boldsymbol{Y}-\boldsymbol{Y_M})\boldsymbol{k} = \boldsymbol{k}^T\boldsymbol{W}\boldsymbol{k}$$

Úpravami postupne dostávame:

$$\begin{aligned}&\boldsymbol{k}^T(\boldsymbol{Y}-\boldsymbol{XG})^T(\boldsymbol{Y}-\boldsymbol{XG})\boldsymbol{k}\\ &= \boldsymbol{k}^T((\boldsymbol{Y}-\boldsymbol{Y_M})+(\boldsymbol{Y_M}-\boldsymbol{XG}))^T((\boldsymbol{Y}-\boldsymbol{Y_M})+(\boldsymbol{Y_M}-\boldsymbol{XG}))\boldsymbol{k}\\ &= \boldsymbol{k}^T(\boldsymbol{Y}-\boldsymbol{Y_M})^T(\boldsymbol{Y}-\boldsymbol{Y_M})\boldsymbol{k} + 2\boldsymbol{k}^T(\boldsymbol{Y}-\boldsymbol{Y_M})^T(\boldsymbol{Y_M}-\boldsymbol{XG})\boldsymbol{k}\\ &+ \boldsymbol{k}^T(\boldsymbol{Y_M}-\boldsymbol{XG})^T(\boldsymbol{Y_M}-\boldsymbol{XG})\boldsymbol{k}\end{aligned}$$

Keďže MANOVA reziduá sú kolmé na rovinu modelu tak

$$2\boldsymbol{k}^T(\boldsymbol{Y}-\boldsymbol{Y_M})^T(\boldsymbol{Y_M}-\boldsymbol{XG})\boldsymbol{k} = \boldsymbol{0}$$

Súčty štvorcov sú nezáporné, teda

$$\boldsymbol{k}^T(\boldsymbol{Y_M}-\boldsymbol{XG})^T(\boldsymbol{Y_M}-\boldsymbol{XG})\boldsymbol{k} \geq \boldsymbol{0}$$

Preto dostávame

$$\boldsymbol{k}^T(\boldsymbol{Y}-\boldsymbol{XG})^T(\boldsymbol{Y}-\boldsymbol{XG})\boldsymbol{k} \geq \boldsymbol{k}^T\boldsymbol{W}\boldsymbol{k}$$

3.4 MANOVA v priestore premenných

Na obrázku 10 sú reziduálne vektory kolineárne, teda perfektne korelované. Všeobecne sa reziduálne vektory nachádzajú v priestore s dimenziou *n-2*. To si však už nevieme predstaviť a na obrázku 10 sa nedajú zobraziť. Preto sa sústredíme iba na priestor, v ktorom sa nachádzajú reziduá, aby sme mohli zobraziť viac dimenzií. A aby sme obohatili geometrickú interpretáciu, prejdeme do priestoru premenných. V priestore premenných sú jednotlivé osi premenné a každé pozorovanie možno zakresliť ako bod v tomto priestore so súradnicami zodpovedajúcimi realizáciam premenných na osiach.

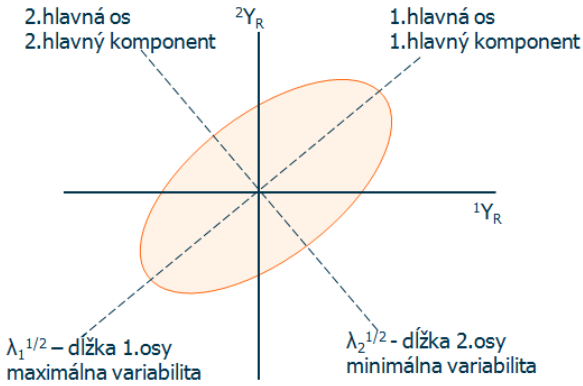
Všetky MANOVA matice, *T, B* a *W*, sú z definície symetrické a pozitívne definitné, keď uvažujeme plnú hodnoť matíc. Sú to kovariačné matice centrovaných vektorov, modelom predikovaných vektorov a reziduálnych vektorov, keď zanedbávame konštantu *n*. Možno si ich preto predstaviť ako elipsoidy v priestore premenných. Konkrétne matica *W* je zobrazená na obrázku 11 ako elipsa.

Minimálna variancia $\boldsymbol{k}^T\boldsymbol{W}\boldsymbol{k}$ znamená, že každým smerom $\boldsymbol{k}$ je elipsoid minimálny. Jeho obsah je minimálny. $Obsah = c(\lambda_1\lambda_2\ldots\lambda_p)^{1/2}$, kde c je konštanta, ktorá

závisí od dimenzie p a čísla π. Čísla $\lambda_1^{1/2},...,\lambda_p^{1/2}$ sú dĺžky osí elipsoidu. V štatistike majú interpretáciu ako smerodajné odchýlky hlavných komponentov. Ich štvorce, teda variancie, alebo presne v našom prípade sumy švorcov, sú vlastné čísla matice W. Matica W sa dá rozložiť na ortogonálnu tranformáciu, teda rotáciu v prípade 2 dimenzií a ponaťahovanie v smere osí. Teda $W=U\Lambda U^T$, kde stĺpce matice U sú ortogonálne vlastné vektory matice W a matica Λ je diagonálna matica s vlastnými číslami λ_i matice W na diagonále. Preto MANOVA minimalizuje determinant matice W, ktorý je úmerný štvorcu obsahu elipsoidu prislúchajúceho matici W, keďže

$$det(W) = det(U\Lambda U^T) = det(\Lambda U^T U) = det(\Lambda) = \lambda_1\lambda_2...\lambda_p.$$

Obrázok 11: Zobrazenie pozitívne definitnej matice W
(Zdroj: vlastné spracovanie)



3.5 MANOVA štatistické rozdelenia

Ukázali sme, že celkovú maticu súčtov štvorcov a súčinov T možno rozdeliť na modelovú B a reziduálnu maticu W, t.j. $T=B+W$, kde detereminant matice W je minimálny spomedzi determinantov reziduálnych matíc lineárnych odhadov stredných hodnôt. Opäť možno reportovať MANOVA tabuľku rozkladu matice T podobne ako v ANOVA prípade. A opäť pri testovaní nulovej hypotézy o rovnosti stredných hodnôt medzi skupinami možno vychádzať z takéhoto rozkladu. Na rozdiel od jednorozmerného prípadu však testovacia štatistika nie je založená na tangense, teda pomere odvesien, ale na pomere k prepone, teda $det(W)/det(T)$. Ak je reziduálna matica W malá vzhľadom k celkovej matici T, tak zamietame hypotézu o rovnosti stredných hodnôt. Podiel týchto determinantov má tzv. Wilksovo lambda rozdelenie a píšeme $det(W)/det(B+W) \sim \Lambda(p,n-J,J-1)$. Je pomerom determinantov dvoch Wishartových rozdelení. Matica W má Wishartove rozdelenie, ak suma štvorcov $\boldsymbol{k}'W\boldsymbol{k}$ má χ^2 rozdelenie pre každý smer $\boldsymbol{k}$, značíme $W \sim W_p(\Sigma,n-J)$.
Wilksovo lambda vieme ďalej upraviť nasledovne:

$$det(W)/det(B+W) = det(I + W^{-1}B)^{-1} = \prod_{i=1..p} 1/(1+\lambda_i),$$

kde λ_i sú vlastné hodnoty matice $W^{-1}B$. Tu už vystupuje priamo pomer model verzus reziduá ako v prípade F-štatistiky.

Namiesto determinantu, teda obsahu, môžeme uvažovať celkovú variabilitu ako súčet variancií jednotlivých premenných. V tomto prípde je "veľkosť" kovariančnej matice jej stopa, teda súčet diagonálnych prvkov. Opäť zo spektrálneho rozkladu:

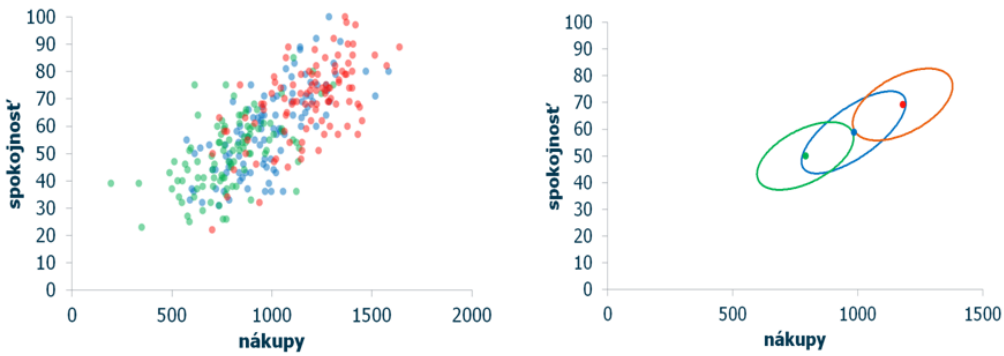
$$Tr(\Sigma) = Tr(U^T diag(\boldsymbol{\lambda})U) = Tr(diag(\boldsymbol{\lambda})) = \sum_{i=1..p} \lambda_i,$$

teda stopa je aj súčtom variancií hlavných komponentov. Ak teda nahradíme determinant stopov, tak dostávame ďalšiu možnosť testovacej štatistiky. V tomto prípade $tr((I + W^{-1}B)^{-1}) = \sum_{i=1..p} \lambda_i/(1+\lambda_i)$, prípadne sa používa priamo $tr(W^{-1}B) = \sum_{i=1..p} \lambda_i$.

4 VÝPOČET V MS EXCELI

V prípadovej štúdii z obchodného reťazca testujeme rovnosť stredných hodnôt spokojnosti a objemu nákupov pre 3 predajne na základe 100 respondentov v každej predajni. Na obrázku 12 sú zobrazené jednotlivé merania a prvé dva momenty po skupinách.

Obrázok 12: Jednotlivé pozorovania a prvé dva momenty pre predajne obchodného reťazca
(Zdroj: vlastné spracovanie)



Z nameraných hodnôt vyrátame centrované dáta a pre každú predajňu jej priemerné hodnoty. Jednoducho dostávame predikované hodnoty a reziduálne hodnoty. Matice súm štvorcov a súčinov sa tiež priamočiaro zrátajú.

T matica 2x2 celkového súčtu štvorcov a produktov, ktorú chceme modelom vysvetliť, má hodnoty

$$T = Y_C^T Y_C = (Y - \mathbf{1}_n({}^1\overline{y}, {}^2\overline{y}, \ldots, {}^p\overline{y}))^T (Y - \mathbf{1}_n({}^1\overline{y}, {}^2\overline{y}, \ldots, {}^p\overline{y})) = \begin{matrix} 19646578 & 919903 \\ 919903 & 80024 \end{matrix}$$

B matica 2x2 vysvetlených modelom súčtov štvorcov a produktov, má hodnoty

$B = (Y_M - 1_n({}^1\overline{y}, {}^2\overline{y}, \ldots, {}^p\overline{y}))^T (Y_M - 1_n({}^1\overline{y}, {}^2\overline{y}, \ldots, {}^p\overline{y})) = \begin{matrix} 7582518 & 374889 \\ 374889 & 18571 \end{matrix}$

W matica 2 x 2 reziduálnych súčtov štvorcov a produktov, má hodnoty

$W = Y_R' Y_R = (Y - Y_M)^T (Y - Y_M) = \begin{matrix} 12064060 & 545014 \\ 545014 & 61453 \end{matrix}$

Teda ľahko overíme, že $T = W + B$.
Môžeme to zapísať aj do tabuľky zdrojov rozptylu, ktorej prvky sú matice:

Disperzia	*Matica*			*d.f.*
	a11	*a12*	*a22*	
Vnútri supermarketov	*12064060*	*545014*	*61453*	*2*
Medzi supermarketami	*7582518*	*374889*	*18571*	*297*
Celková	*19646578*	*919903*	*80024*	*299*

Determinanty možno zrátať použitím funkcie *MDETERM*, dostávame:
Det(W) = 444 327 704 452, Det(T) = 725 974 618 315
a teda *Wilksovo Lambda = Det(W)/Det(T) = 0,612043* je realizácia z $\Lambda(2,297,2)$ pri platnosti nulovej hypotézy.
Prevedieme Λ na F-rozdelenie, ktoré možno v Exceli zrátať. Platí, že
$(1-\Lambda(2,297,2)^{1/2}) / \Lambda(2,297,2)^{1/2} \sim F(4,592)*2/296$.
Teda dostaneme *F=41,17804691* a pre kritickú hodnotu použijeme funkciu *FINV(0.05,4,592). F(4,592,0.05)=2,386985554*, teda kritická hodnota je nízka v porovnaní s našou nameranou hodnotou.
Môžeme tiež zrátať p-hodnotu ako *FDIST(41.18,4,592)*, dostávame 1.81763E-30, teda veľmi nízku hodnotu. Preto zamietame nulovú hypotézu a hovoríme, že supermarkety sú odlišné vzhľadom na spokojnosť a objem nákupov.

Záver

Ukázali sme geometrickú interpretáciu základných štatistických pojmov. Priemer ako projekciu na diagonálu, varianciu ako štvorec dĺžky vektora, koreláciu ako kosínus, stupňe voľnosti ako počet dimenzií vektorových priestorov, F-test ako tangens uhla v pravouhlom trojuholníku vzniknutom pri ortogonálnej projekcii na priestor stredných hodnôt. Pre viacrozmerné analýzy sme ukázali geometrickú interpretáciu kovariančnej matice ako elipsoidu, determinantu ako objemu. Tieto isté geometrické predstavy možno ďalej aplikovať na:

- Testovanie lineárnych hypotéz o parametroch modelu. Jediný rozdiel v tomto prípade je, že priestor stredných hodnôt je limitovaný testovanou hypotézou. Zostrojenie pravouhlého trojuholníka a teda aj testovanie hypotéz má obdobnú geometriu ako sme tu prezentovali.

- Rozšírenie modelu na viac faktorov. Viac faktorov možno skombinovať do jedného a postupovať rovnako. Počet úrovní finálneho jedného faktora rastie geometricky rýchlo s počtom uvažovaných faktorov a preto aj počet stupňov voľnosti modelu rastie. Môže ľahko prekročiť počet pozorovaní a teda model nebude identifikovateľný. Preto sa opäť obmedzuje priestor stredných hodnôt, teda dimenzia modelu. Odhadujú sa najmä hlavné efekty faktorov, prípadne párové interakcie, ak je dosť pozorovaní.
- Lineárnu regresiu. Nemusíme byť obmedzení na kategoriálne premenné ako v prípade faktorov v analýze rozptylu.

V neposlednom rade prispeje takéto geometrické pochopenie aj k lepšiemu porozumeniu iných viacrozmerných techník, v ktorých sa vyskytujú matice ortogonálnych projekcií či determinanty pozitívne definitných matíc.

Tento článok vznikol s prispením grantovej agentúry VEGA v rámci projektu číslo 1/0092/15: Moderné prístupy k navrhovaniu komplexných štatistických prieskumov.

Kľúčové slová

MANOVA, ANOVA

Klasifikácia JEL

C3

LITERATÚRA

[1] Lamoš, F. - Potocký, R. 1989. *Pravdepodobnosť a matematická štatistika.* Bratislava: Alfa, 1989. 344 s. ISBN 80-05-00115-0
[2] Rao R. C. 1978. *Lineární metody statistické indukce a jejich aplikace.* Praha: Academia, 1978. 666 s.
[3] Härdle, W.K. - Simar, L. 2012. *Applied Multivariate Statistical Analysis*. Springer-Verlag Berlin Heidelberg, 2012. 516 s. ISBN 978-3-642-17229-8
[4] Cox, T. F. - Cox, M. A. A. 2000. *Multidimensional Scaling*. Chapman and Hall/CRC, 2000. 328 s. ISBN 1-58488-094-5

RESUMÉ

V článku prezentujeme geometrickú interpretáciu viacrozmernej analýzy rozptylu. Zadefinujeme priestor pozorovaní a ukážme geometrickú interpretáciu základných štatistických ukazovateľov. Priemer je interpretovaný ako súradnice ortogonálnej projekcie vektora pozorovaní na diagonálu, smerodajná odchýlka ako dĺžka vektora a korelácia ako kosínus uhla vektorov. ANOVA tabuľka rozkladu variancie je interpretovaná ako Pytagorova veta, F-test ako tangens v pravouhlom trojuholníku a počet stupňov voľnosti ako počet dimenzií priestoru, v ktorom sa vektor nachádza. Ďalej je definovaný priestor premenných, v ktorom je kovariančná matica interpretovaná ako

elipsoid a jej determinant ako objem. Ukázaný je MANOVA rozklad matice súčtov štvorcov a súčtu súčinov obdobný Pytagorovej vete. Minimálnosť determinantu reziduálnej matice súčtu štvorcov a súčtu súčinov je ukázaná. Výpočet MANOVA testu je prezentovaný na konkrétnom príklade testovania rozdielnosti výkonnosti predajní obchodného reťazca. Test je implementovaný v MS Exceli bez nutnosti použitia špeciálneho štatistického softvéru.

SUMMARY

A geometrical interpretation of multivariate analysis of variance is presented in the article. Space of observations in defined and a geometric interpretation of basic statistics in shown. Average is interpreted as coordinates of the orthogonal projection of a vector of observations on the diagonal, standard deviation as the length of the vector and correlation as the cosine of an angle of two vectors. ANOVA table of the variance decomposition is interpreted as Pythagoras theorem, F-test as the tangent in a right-angled triangle and the number of degrees of freedom as the number of dimensions of a vector space. Then also the space of variables is defined, where the covariance matrix is interpreted as an ellipsoid and the determinant of the matrix as the volume of the ellipsoid. The decomposition of MANOVA sum of squares and cross products and the minimality of the residual sum of squares and cross product matrix are shown. The computation of MANOVA is presented using an example of testing a similarity of performance measures of retail chain outlets. The test is implemented in MS Excel, no special statistical software is needed.

Kontakt

Mgr. Jozef Kušnier, Katedra štatistiky, Fakulta hospodárskej informatiky, Ekonomická univerzita v Bratislave, Dolnozemská cesta 1b, 852 35 Bratislava, email: kusnier.jozef@gmail.com