APLIKÁCIA MATLABU PRI RIEŠENÍ NIEKTORÝCH ÚLOH KORELAČNEJ A REGRESNEJ ANALÝZY

PhDr. Eva Ostertagová, PhD.
Technická univerzita v Košiciach
Fakuta elektrotechniky a informatiky
Katedra matematiky
B. Němcovej 32, 042 00 Košice
eva.ostertagova@tuke.sk

Anotácia

V príspevku sú uvedené niektoré problémy spojené so štatistickými závermi v prípade *dvojrozmerného* základného súboru (*X*, *Y*).

Kľúčové slová

Výberová kovariancia, modifikovaná výberová kovariancia, výberový korelačný koeficient, lineárna regresná funkcia, regresné koeficienty.

Náhodný výber a jeho charakteristiky

V štatistike sa používa tzv. *výberová metóda*, ktorá spočíva v tom, že popis určitého štatistického súboru a závery o ňom opierame o údaje, týkajúce sa len vybraných jednotiek tohto súboru. Teda základným princípom výberovej metódy je usudzovanie z časti na celok. Súbor, ktorý je predmetom výskumu potom nazývame *základný súbor* a súbor jednotiek, ktoré boli z neho určitým spôsobom vybrané, nazývame *výberový súbor*.

Pri výberovej metóde sa vyskytujú dve skupiny problémov. Prvú skupinu problémov, ktorá spočíva v odhade určitých neznámych parametrov základného súboru na základe charakteristík výberu, rieši tzv. *teória odhadu*. Druhú skupinu problémov tvoria otázky porovnávania parametrov toho istého druhu v rôznych základných súboroch, porovnávanie niektorého parametra základného súboru s nejakou určenou hodnotou, prípadne porovnávanie celých rozdelení. Túto skupinu problémov rieši *teória testovania hypotéz*.

Dostatočná kvalita odhadov a testov musí byť zabezpečená výberom, ktorý by dobre reprezentoval základný súbor. Reprezentatívnosť výberu sa dosahuje najčastejšie náhodným výberom, ktorý možno realizovať rôznymi spôsobmi, ako sú napr. prostý náhodný výber, oblastný výber a viacstupňový výber. Najjednoduchším typom je *prostý náhodný výber*, ktorý je charakteristický tým, že jednotky sa vyberajú priamo z neroztriedeného základného súboru v zásade tak, aby každá jednotka mala rovnakú možnosť byť vybraná. Je potrebné rozlišovať medzi *náhodným výberom s vrátením (s opakovaním)*, pri ktorom sa každá vybraná jednotka vracia do základného súboru a *náhodným výberom bez vrátenia (bez opakovania)*, pri ktorom sa vybraná jednotka do základného súboru nevracia. Pri vysvetľovaní štatistických metód budeme v tomto príspevku vychádzať z prostého náhodného výberu s opakovaním.

Neznáme číselné charakteristiky základného súboru sa odhadujú pomocou príslušných výberových charakteristík. Ak vyberieme zo základného súboru *n* jednotiek, t.j. ak urobíme náhodný výber V_n o rozsahu *n*, a ak zistíme u každej vybranej jednotky hodnotu znaku *X*, potom získame *n výberových hodnôt* (realizácií, nameraných hodnôt) $x_1, x_2, \ldots, x_n$.

Teraz uvedieme definície *výberových charakteristík*, a to výberového priemeru, výberového rozptylu, modifikovaného výberového rozptylu, výberovej smerodajnej odchýlky a modifikovanej výberovej smerodajnej odchýlky.

Definícia 1.
Nech je daný náhodný výber V_n o rozsahu *n*, kde $x_1, x_2, \ldots, x_n$ sú namerané hodnoty. *Výberový priemer* náhodného výberu V_n sa pre prosté triedenie definuje ako číslo $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$, pričom v prípade triedenia početností alebo intervalového triedenia ho môžeme zapísať v tvare $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot n_i$, kde n_i sú absolútne početnosti a *k* je počet tried.

Definícia 2.
Nech je daný náhodný výber V_n o rozsahu *n*, kde $x_1, x_2, \ldots, x_n$ sú namerané hodnoty. *Výberový rozptyl (disperzia)* náhodného výberu V_n sa pre prosté triedenie definuje ako číslo $s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$, pričom v prípade triedenia početností alebo intervalového triedenia ho môžeme zapísať v tvare $s^2 = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i$, kde n_i sú absolútne početnosti a *k* je počet tried.

Modifikovaný výberový rozptyl (disperzia) náhodného výberu V_n sa pre prosté triedenie definuje ako číslo $s^{*2} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$, pričom v prípade triedenia početností alebo intervalového triedenia ho môžeme zapísať v tvare $s^{*2} = \frac{1}{n-1} \cdot \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i$.

Definícia 3.
Nech je daný náhodný výber V_n o rozsahu n, kde x_1, x_2, ... , x_n sú namerané hodnoty. *Výberová smerodajná odchýlka*, resp. *modifikovaná výberová smerodajná odchýlka* náhodného výberu V_n sa definuje ako číslo $s = \sqrt{s^2}$, resp. $s^* = \sqrt{s^{*2}}$.

Korelačná a regresná analýza

Pri skúmaní vzájomnej závislosti viacerých náhodných premenných riešime dve základné úlohy:

1. *Korelačná úloha*: Aká je miera závislosti?
2. *Regresná úloha*: Aký je tvar závislosti?

Definícia 4.
Nech (x_1, y_1), ... , (x_n, y_n) sú namerané hodnoty nezávislého náhodného výberu V_n o rozsahu n systému dvoch náhodných premenných (X, Y) a nech $\bar{x}$ a $\bar{y}$ sú ich výberové priemery, s_x a s_y sú ich výberové smerodajné odchýlky a nech s_x^* a s_y^* sú ich modifikované výberové smerodajné odchýlky. Definujeme:

1. *Výberová kovariancia*

$$k_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}).$$

2. *Modifikovaná výberová kovariancia*.

$$k_{xy}^* = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}).$$

3. *Výberový korelačný koeficient*

$$r_{xy} = \frac{k_{xy}}{s_x \cdot s_y} = \frac{k_{xy}^*}{s_x^* \cdot s_y^*}.$$

Po zavedení označení $\overline{x^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2$, $\overline{y^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i^2$, $\overline{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i$ a po úpravách dostaneme, že $s_x^2 = \overline{x^2} - (\bar{x})^2$, $s_y^2 = \overline{y^2} - (\bar{y})^2$ a $k_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}$.

Potom pre výberový korelačný koeficient platí vzťah:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \cdot \sqrt{\overline{y^2} - (\bar{y})^2}}.$$

Výberový korelačný koeficient r_{xy} meria *tesnosť* lineárnej závislosti medzi premennými X a Y, a to obojstranne, t.j. $r_{xy} = r_{yx}$. Výberový korelačný koeficient nadobúda hodnoty z intervalu $\langle -1, 1 \rangle$. Čím je hodnota $|r_{xy}|$ bližšia k 1, tým je lineárna závislosť silnejšia a čím je hodnota r_{xy} bližšia k 0, tým je lineárna závislosť slabšia. Ak $r_{xy} = 0$, tak hovoríme, že lineárna závislosť medzi X a Y neexistuje (môže však existovať iná závislosť).

V tomto príspevku sa budeme zaoberať len *lineárnou regresnou funkciou* tvaru $y = a_0 + a_1 \cdot x$. Pomocou *metódy najmenších štvorcov* sa dá odvodiť, že odhady *regresných koeficientov* a_0 a a_1 je možné získať riešením dvoch lineárnych rovníc o dvoch neznámych:

$$a_0 \cdot n + a_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i ,$$

$$a_0 \cdot \sum_{i=1}^{n} x_i + a_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i .$$

Príklad: V tabuľke sú výsledky meraní výkonu Y [kW] a otáčok X [min^{-1}] benzínového motora:

Otáčky	2 000	2 500	3 000	3 500	4 000
Výkon	29	43	55	64	71

Riešte nasledujúce úlohy:
a) Určte výberový korelačný koeficient r_{xy}.
b) Stanovte odhady regresných koeficientov regresnej priamky $y = a_0 + a_1 \cdot x$.
c) Znázornite danú regresnú priamku spolu s bodmi $[x_i, y_i]$.

Riešenie:

a) *Výberový korelačný koeficient* r_{xy} môžeme vypočítať na základe vzťahu:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \cdot \sqrt{\overline{y^2} - (\bar{y})^2}} .$$

Počítame postupne:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{15\,000}{5} = 3\,000 ,$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{262}{5} = 52,4 ;$$

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i = \frac{838\,500}{5} = 167\,700 ,$$

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \frac{47\,500\,000}{5} = 9\,500\,000 ,$$

$$\overline{y^2} = \frac{1}{n} \sum_{i=1}^{n} y_i^2 = \frac{14\,852}{5} = 2{,}9704 \cdot 10^3 ,$$

$$s_x = \sqrt{\overline{x^2} - (\bar{x})^2} = 707{,}1068 ,$$

$$s_y = \sqrt{\overline{y^2} - (\bar{y})^2} = 14{,}988 .$$

Dostaneme výsledok:

$$r_{xy} = \frac{167\,700 - 3\,000 \cdot 52{,}4}{707{,}1068 \cdot 14{,}988} = 0{,}9907 .$$

Hore uvedený výpočet pomocou kalkulačky môžeme efektívne zjednodušiť použitím MATLABu:

x=[2000,2500,3000,3500,4000],y=[29,43,55,64,71]
r=corrcoef(x,y),rxy=r(1,2)

b) *Regresnú priamku* hľadáme v tvare $y = a_0 + a_1 \cdot x$, pričom odhady regresných koeficientov a_0 a a_1 získame riešením sústavy dvoch lineárnych rovníc o dvoch neznámych:

$$a_0 \cdot n + a_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i ,$$

$$a_0 \cdot \sum_{i=1}^{n} x_i + a_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i .$$

V našom prípade to bude táto sústava rovníc:

$5 \cdot a_0 + 15\,000 \cdot a_1 = 262$,
$15\,000 \cdot a_0 + 47\,500\,000 \cdot a_1 = 838\,500$.

Riešením danej sústavy rovníc dostaneme tieto výsledky: $a_0 = -10{,}6000$ a $a_1 = 0{,}0210$.
Teda $y = -10{,}6000 + 0{,}0210 \cdot x$.

Teraz ukážeme riešenie tejto úlohy použitím MATLABu dvojakým spôsobom. Prvým spôsobom je riešenie danej sústavy dvoch lineárnych rovníc o dvoch neznámych (regresné koeficienty a_0 a a_1) pomocou inverznej matice:

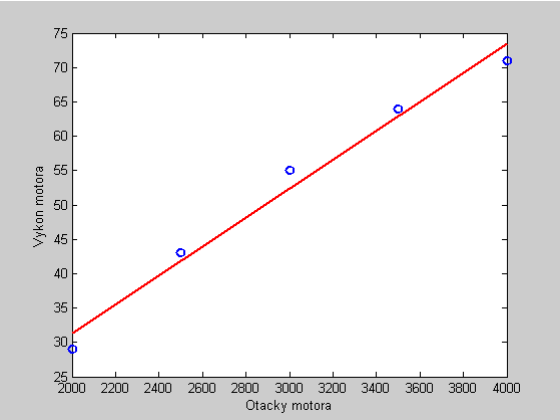
A=[length(x),sum(x);sum(x),sum(x.^2)],
B=[sum(y);sum(x.*y)],koef=inv(A)*B

Druhým spôsobom je jednoduché riešenie priamo použitím štandardnej funkcie MATLABu *polyfit(x,y,n)*, kde *n* udáva stupeň polynómu (pozrite si to podrobnejšie použitím príkazu *help polyfit*):

koef=polyfit(x,y,1),a1=koef(1),a0=koef(2)

c) Použitie MATLABu pre grafické znázornenie:

plot(x,y,'o','linewidth',2);b=axis;hold on,
plot(b(1:2),polyval(koef,b(1:2)),'r','linewidth',2)
xlabel('Otacky motora'),ylabel('Vykon motora')



Záver

Príkladom v príspevku je uvedený účinný spôsob realizácie numerických výpočtov a ich grafického znázornenia softvérovým produktom MATLAB.

Literatúra

1. Cyhelský L., Kaňoková J., Novák I.: Základy teorie statistiky pro ekonomy. SNTL/Alfa, Praha 1979.
2. Daňo I., Ostertagová E.: Numerické metódy, pravdepodobnosť a matematická štatistika v počítačovom prostredí MATLABu. Elfa, Košice 2009.
3. Gavalec M., Kováčová N., Ostertagová E., Skřivánek J.: Pravdepodobnosť a matematická štatistika v počítačovom prostredí MATLABu. Elfa, Košice 2002.
4. Hátle J., Likeš J.: Základy počtu pravdepodobnosti a matematické statistiky. SNTL/Alfa, Praha 1974.
5. Ostertagová E.: Pravdepodobnosť a matematická štatistika v príkladoch. Elfa, Košice 2005.
6. Ostertagová, E.: Využitie Matlabu pri riešení niektorých typov úloh z pravdepodobnosti a matematickej štatistiky. Sborník 29. konference o matematice na VŠTEZ: Matematika v inženýrském vzdělávání. Univerzita Tomáše Bati ve Zlíně, 2006.
7. Pirč, V., Ostertagová, E.: Matematika s MATLABom. FEI TU v Košiciach, 2007. http://www.tuke.sk/fei-km/soft2/uLern_Viewer.htm

Príspevok bol vypracovaný v rámci riešenia grantového projektu VEGA č. 1/0679/08 Integrovaný systém pre inovované projektovanie, plánovanie, organizovanie a riadenie výroby.