Elektronická podpora pre časť: Štatistické výpočty v MATLABe

(Doplnenia a komentár textu skrípt, riešenia niektorých príkladov a úloh)

Doplnenia textu skrípt sú pre lepšiu orientáciu označené symbolmi ♠(začiatok) ♣(koniec)

Kap. 2 Opisná štatistika viacrozmerného štatistického súboru

V tejto kapitole predmetom skúmania nebude jeden stĺpec údajov štandardného tvaru dát, ale dva, resp. viac stĺpcov. Každý z nich zvlášť síce dokážeme charakterizovať spôsobmi, ktoré sme preberali v prvej kapitole, avšak tie nám neprezradia nič o asociácii medzi nimi (t.j. o väzbe, o súvislosti). To nové, čo je jadrom tejto kapitoly, to je spôsob opísania väzby, charakterizovania asociácie medzi skúmanými znakmi.

2.1 Stochastická väzba kvalitatívnych znakov

♠Teraz si dovolíme začať reakciou na námietku jedného študenta z minulého školského roku. Text tohoto článku (článku 2.1 v skriptách) sa mu videl zbytočne komplikovaný. Vraj, prečo tabuľka obsahuje tie čísla, ktoré obsahuje, t.j. také, ktoré potom v úvahe meníme, aby sme mohli hovoriť o pomeroch 4:6:3, resp. 3:1:1. Prečo jej riadky neobsahujú to, čo úvaha "vyžaduje", jednoducho, prečo od začiatku nie je to tabuľka, kde v prvom riadku sú čísla: 12 – 18 – 9 a v druhom riadku: 36 – 12 – 12 ?

Nuž, odpoveď je, že reálne dáta veľmi, veľmi zriedka kopírujú "ideálne" pomery! Situáciu môžeme prirovnať k takej, v ktorej ide o hádzanie mincou. Povedzme, nech ide o 20 násobné hádzanie mincou (riadnou, nepoškodenou, napr. dvojkorunovou mincou). Hoci vieme, že oba možné výsledky hodu (znak Z, resp. číslo Č) majú pravdepodobnosť 0.5, predsa považujeme za veľmi nepravdepodobné, že výsledok 20 násobného opakovaného hádzania by bol: Z, Č, Z, Č, Z, Č, Z, Č, Z, Č, Z, Č, Z, Č, Z, Č, Z, Č, Z, Č.

Koho táto odpoveď neuspokojuje, prosím, obráťte sa na svojho učiteľa. Tento text ho nemôže nahradiť.

Poznamenávame, že pochopenie myšlienok článku 2.1 sa rýchle zúročí, veď v ďalšej kapitole bude reč o stochastickej väzbe medzi dvomi diskrétnymi náhodnými veličinami (str. 139). Ubezpečujeme čitateľa, že "pozadie" týchto vecí je rovnaké.♣

2.2 Stochastická väzba kvantitatívnych znakov

Uvažujme o situácii z prvého príkladu článku 1.1 (záznam výsledkov práce krúžku 16-tich študentov)

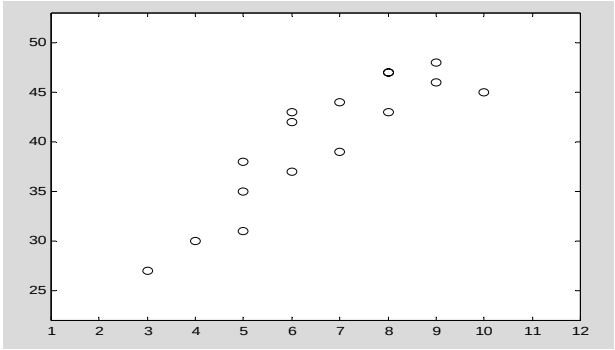
študent	*aktivita*	*test 1*	*test 2*	*zadania*	*skúška*	*body celkom*
1	8	13	14	7	39	81
2	4	9	11	6	42	72
..	..	..	..	..	..	..
15	8	11	12	6	37	74
16	10	14	13	8	43	88

Predstavme si na chvíľu, že tabuľka je kompletná. Znak "zadania" predstavuje hodnotenie domácej práce a je na mieste otázka, či hodnotenie domácej práce (ktorá asi nie je vždy samostatná) a hodnotenie získané na skúške, sú veličiny závislé, alebo nie. Pre jednoduchosť, označme premennou x hodnotenie zadaní a premennou y hodnotenie zo skúšky. Aby sme ušetrili priestor, číselné hodnoty z príslušných stĺpcov (t.j. hodnoty piateho a šiesteho stĺpca) uveďme ako riadkové vektory:

```
x = [7   6   4   8   5  10   9   7   6   8   5   5   3   9   6   8];
y = [39 42  30  47  38  45  48  44  43  47  35  31  27  46  37  43];
```

Súvis, väzbu, zatiaľ nevidíme, preto body zobrazme:

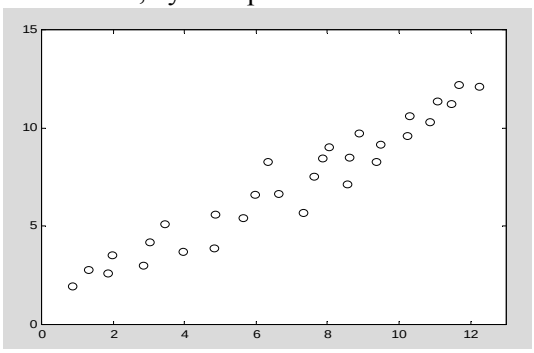
```
okno = [min(x)-2, max(x)+2, min(y)-5, max(y)+5]; plot(x,y,'ko'), axis(okno)
```

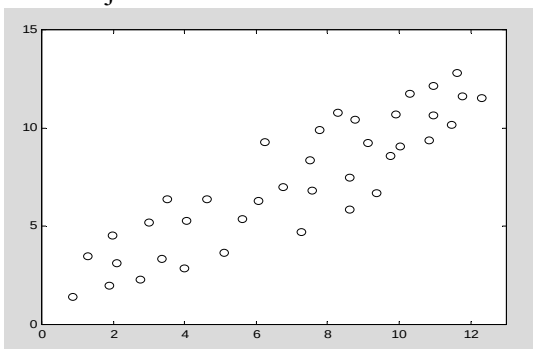


Na osi x sú body získané zo zadaní (domáce úlohy) a na osi y sú body získané na skúške.

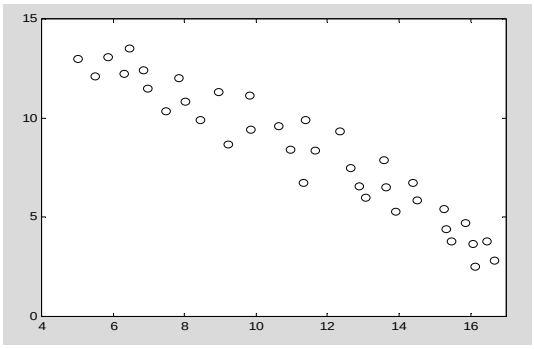
Je zrejmé, že väzba medzi premennými *x, y* tu je, aj keď nie je funkcionálna, veď napr. hodnote $x = 5$ odpovedá nie jedna, ale 3 hodnoty premennej *y* (sú nimi: 31, 35, 38; analogicky je to aj s ďalšími hodnotami premennej *x*). Ďalej vidieť, že s rastúcim *x* rastie *y*, hoci nie v bežnom zmysle (viď napr. dvojice [6, 42], [7, 39]). Berieme do úvahy, že tu ide o vzťah *stochastický*, vzťah, ktorý (voľne povedané) znamená, že nejaká vlastnosť je prítomná na významnej väčšine prvkoch, avšak nie na všetkých. Nasledujúce obrázky ilustrujú možné prípady.

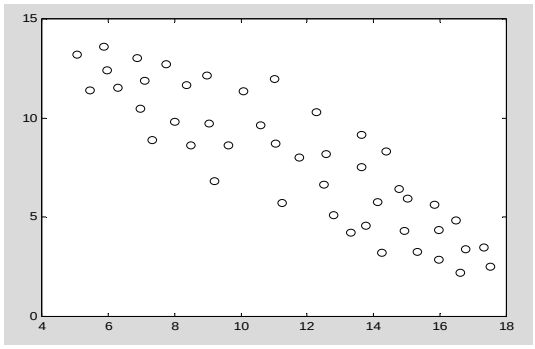
Prvá dvojica obrázkov predstavuje situáciu, v ktorej na ľavom obrázku s rastúcim *x* rastie *y*, pričom väzba je dosť silná, kým na pravom s rastúcim *x* rastie *y*, ale väzba je slabšia.



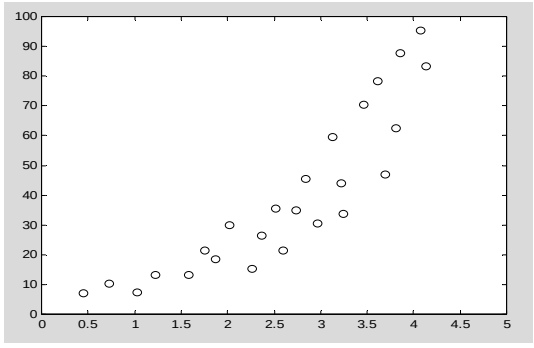


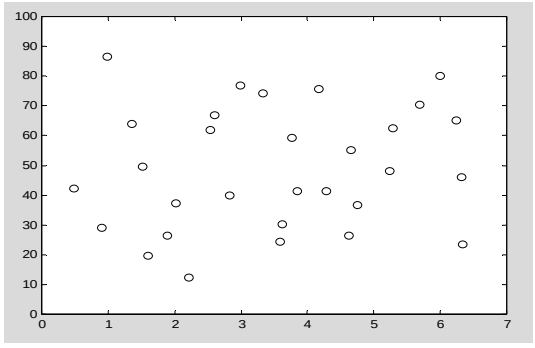
Druhá dvojica zobrazuje situáciu, v ktorej na ľavom obrázku s rastúcim *x* klesá *y*, pričom väzba je silná, kým na pravom s rastúcim *x* klesá *y*, avšak sila väzby je slabšia.





Tretia dvojica predstavuje jednak situáciu, keď väzba nie je lineárna (vľavo) a jednak situáciu, keď o väzbe nie je možné hovoriť (obrázok vpravo):





Takýmto obrázkom sa v Helpe MATLABu (resp. v anglicky písaných knižkách) hovorí "scatter plot". Aj keď veľa naznačujú, je zrejmé, že posudzovanie sily väzby len pohľadom na zobrazenie bodov je subjektívne. V nasledujúcom článku hovoríme o číselných charakteristikách stochastickej väzby, teda o mierach, ktoré tú väzbu zachytávajú objektívne. Zatiaľ o väzbe medzi znakmi (veličinami) hovoríme v intuitívnej rovine a chápeme ju ako závislosť medzi nimi.

Vezmime na vedomie, že termín *stochastická nezávislosť* je presne definovaný pojem v teórii pravdepodobnosti (časť z nej je obsahom ďalších kapitol). V štatistike tento pojem vystupuje buď ako predpoklad, teda *ad hoc* alebo ako *hypotéza* (ktorú štatistickými metódami zamietame, alebo nezamietame). Poznamenávame, že problematika štatistických testov je mimo rámca nášho kurzu.

2.3 Výberová kovariancia a výberový korelačný koeficient

Kovariancia a korelačný koeficient sú číselné charakteristiky, popisujúce silu lineárneho vzťahu dvojice náhodných veličín (dvojrozmerného náhodného vektora). Ich definícia a spôsob ich výpočtu patrí do TP (Teórie Pravdepodobnosti). Nám pôjde o definovanie ich výberových protipólov, t.j. ich odhadov, ktoré je možné vyčísliť z daného súboru údajov (x_i, y_i), teda z náhodného výberu. Vo výbere nám ide o dvojicu kvantitatívnych znakov (teda o dvojicu číselných veličín X, Y). Údaje (x_i) sú namerané hodnoty veličiny X na jednotkách výberu a sú to čísla jedného stĺpca štandardného tvaru dát, kým údaje (y_i) sú zistené (namerané) hodnoty veličiny Y (a sú to čísla druhého, resp. iného stĺpca uvažovanej matice štandardnej formy dát. Pritom dvojica (x_i, y_i), pochádza z i-tej jednotky výberu. Nech počet jednotiek výberu (t.j. počet riadkov matice) sa rovná n.

Číselnou charakteristikou lineárnej väzby veličín X, Y je *výberová kovariancia*:

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

Ľahko sa ukáže, že platí:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{x}\,\bar{y}$$

a preto výberová kovariancia sa rovná nule práve vtedy, ak *priemer súčinu sa rovná súčinu priemerov*.

Kovariancia postihuje lineárnu väzbu v tom zmysle, že keď sa rovná nule, tak lineárna väzba medzi X, Y nie je. Na to, aby sme jej nenulové hodnoty mohli považovať za mieru lineárnej závislosti medzi X a Y, má kovariancia jeden zrejmý nedostatok: ak zmeníme mierku (resp. mierky), v ktorej uvádzame hodnoty veličín X a Y, tak sa zmení jej hodnota (čo je zrejme nevhodné, veď sila väzby medzi X, Y by nemala byť závislou na zvolenej mierke).

♠ Napríklad, nech $U = \alpha X$, $V = \beta Y$. Počítajme kovarianciu dvojice (U,V):

$$s_{uv} = \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v}) = \frac{1}{n}\sum_{i=1}^{n}(\alpha x_i - \alpha\bar{x})(\beta y_i - \beta\bar{y}) = \alpha\beta\, s_{xy}$$

♣

Tento nedostatok je eliminovaný modifikáciou vzorca pre kovarianciu a modifikácia vedie k pojmu *výberového korelačného koeficientu*:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(je treba predpokladať, že oba činitele v menovateli sú kladné, čo je splnené, akonáhle aspoň dve hodnoty súboru čísel tak (x_i) ako aj (y_i) sú rôzne). Dá sa dokázať, že platí:

1. Hodnota r_{xy} je vždy číslo z uzavretého intervalu [-1, 1].
2. Ak $r_{xy} = 1$, tak existujú reálne $a, b, b > 0$, že platí: $y_i = a + b x_i$ pre všetky $i = 1, 2, ..., n$.
3. Ak $r_{xy} = -1$, tak existujú reálne $a, b, b < 0$, že platí: $y_i = a + b x_i$ pre všetky $i = 1, 2, ..., n$.
4. Ak $u_i = \alpha x_i$, $v_i = \beta y_i$, tak $r_{uv} = r_{xy}$ (zmena mierky nezmení hodnotu korelačného koeficienta).

♠ Dôkazy bodov 1, 2 a 3 nie sú jednoduché, avšak dôkaz bodu 4 je naozaj ľahký. Stačí zvážiť, ako sa zmenia výrazy v menovateli, pretože ako sa zmení čitateľ už vieme z toho, čo je napísané povyše.

Zrejme platí:

$$\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})=\sum_{i=1}^{n}(\alpha x_i-\alpha\bar{x})(\beta y_i-\beta\bar{y})=\alpha\beta\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

a pre výrazy v menovateli máme

$$\sum_{i=1}^{n}(u_i-\bar{u})^2=\sum_{i=1}^{n}(\alpha x_i-\overline{\alpha x})^2=\alpha^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

a analogicky

$$\sum_{i=1}^{n}(v_i-\bar{v})^2=\sum_{i=1}^{n}(\beta y_i-\overline{\beta y})^2=\beta^2\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

Po dosadení pre korelačný koeficient dostávame

$$r_{uv}=\frac{\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i-\bar{u})^2.\sum_{i=1}^{n}(v_i-\bar{v})^2}}=\frac{\alpha\beta\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{|\alpha||\beta|\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{\alpha\beta}{|\alpha||\beta|}r_{xy}$$

Teraz je evidentné, že ak súčin $\alpha\beta > 0$, tak $r_{uv} = r_{xy}$, avšak, ak $\alpha\beta < 0$, tak $r_{uv} = -r_{xy}$. Ako vidíme, v skriptách je nepresnosť. Na ospravedlnenie autora skrípt: Na str. 127 ide o kontext, v ktorom sa hovorí o transformáciách $u = \alpha x$, $v = \beta y$ ako o *zmenách mierky* pre veličiny *X*, resp. *Y* a vtedy, samozrejme, aj α aj β sú kladné čísla, a preto naozaj $r_{uv} = r_{xy}$.

♣

Ak sa výberový korelačný koeficient rovná nule, hovoríme o *nekorelovanosti* znakov (ten prípad nastane práve vtedy, keď sa kovariancia rovná nule). Tieto slová sú platné v teórii pravdepodobnosti; pri praktických výpočtoch sa zriedka stane, že výberová kovariancia a teda aj výberový korelačný koeficient sa rovná nule. Očakávame, že malé hodnoty výberového korelačného koeficienta (teda, lepšie povedané, v absolútnej hodnote malé) signalizujú, že lineárny vzťah medzi X a Y môžeme vylúčiť. Lenže, čo znamená "malé" (teda, "temer" nulové) a čo znamená *významne* nenulové? Tento moment je dotyk so samotným jadrom štatistiky.

Aj keď tu, žiaľ, nie je priestor na kompletnú odpoveď, pre podporenie štatistického uvažovania uveďme:

Ak rozsah výberu $n = 20$, tak významne nenulové sú hodnoty (v absolútnej hodnote) väčšie ako 0,44.

Ak napr. n = 30, tak sú nimi už hodnoty (v absolútnej hodnote) väčšie ako 0,36.

Ak napr. n = 50, tak už hodnoty väčšie ako 0,27 sú považované za významne nenulové.

To, čo sme práve povedali, platí za predpokladu, že základný súbor, z ktorého výber pochádza, má *normálne* rozdelenie (trochu predbiehame, lebo pojem normálneho rozdelenia bude vysvetlený v kap. 4).

Uviedli sme tieto skutočnosti preto, aby sme si uvedomili, že bez ďalšieho štúdia štatistiky nie je možné dať odpoveď na otázky, ktoré vznikajú celkom prirodzene pri elementárnej analýze dát.

♠

Na tomto mieste sa prenášame do článku 2.4 a pokúsime sa o

- zdôraznenie rozhodujúcich momentov pri odvádzaní parametrov regresnej priamky
- objasnenie významu koeficientu determinácie a jeho súvisu s korelačným koeficientom

Výberovú regresnú priamku (teda regresnú priamku vzťahujúcu sa k uvažovanému náhodnému výberu) definujeme ako priamku $y = a^o + b^o x$, ktorá minimalizuje súčet štvorcov odchýliek, t.j. pre ktorú platí:

$$\sum[y_i - a^\circ - b^\circ x_i]^2 \le \sum[y_i - a - b x_i]^2$$

pre všetky $a, b \in R$.

Majme na zreteli, že (x_i, y_i), $i = 1,2,..,n$ sú dané vstupné dáta a našou úlohou je nájsť a^o, b^o, teda nájsť parametre regresnej priamky. Hore uvedená nerovnosť znamená, že chceme nájsť bod (a^o, b^o), v ktorom funkcia

$$\mathrm{SSO}(a,\,b) \;=\; \sum_{i=1}^{n} [\,y_i - a - bx_i\,]^2$$

ako funkcia dvoch premenných, nadobúda globálne minimum. Na rozdiel od článku 6.2 (kap. 6), teraz použijeme elementárnejší postup odvodenia vzťahov pre neznáme a^o, b^o a na získané vzťahy sa pozrieme (aspoň čiastočne) štatistickým pohľadom. Umelé, ale jednoduché úpravy dávajú:

$$\mathrm{SSO}(a,\,b) \;=\; \Sigma\,[\,y_i - a - b\,x_i\,]^2 \;=\; \Sigma\,[\,(y_i - \overline{y}) + \overline{y} - a - b(x_i - \overline{x}) - b\,\overline{x}\,]^2 \;=$$

$$= \Sigma\,[\,(y_i - \overline{y}) - b(x_i - \overline{x}) + (\overline{y} - a - b\,\overline{x})\,]^2$$

Umocnime trojčlen a zaveďme označenia:

$SS_{xx} = \Sigma\,(x_i - \overline{x})^2$, $SS_{yy} = \Sigma\,(y_i - \overline{y})^2$, $SS_{xy} = \Sigma\,(x_i - \overline{x})(y_i - \overline{y})$. Potom platí:

$$\mathrm{SSO}(a,\,b) \;=\; n(\overline{y} - a - b\,\overline{x})^2 + SS_{xx}\,b^2 \;-\; 2\,SS_{xy}\,b + SS_{yy} \;=$$

$$= n(\overline{y} - a - b\,\overline{x})^2 + \left(\sqrt{SS_{xx}}\,b - \frac{SS_{xy}}{\sqrt{SS_{xx}}}\right)^2 + \left(SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}\right)$$

Označme $b^o = \dfrac{SS_{xy}}{SS_{xx}}$ a tiež, nech $a^o = \overline{y} - b^o\,\overline{x}$. Ak vo výraze pre $\mathrm{SSO}(a,\,b)$ položíme $b = b^o$, tak prostredný člen výrazu sa rovná nule. Ak súčasne za a položíme $a = a^o$, tak aj prvý člen sa rovná nule a je preto zrejmé, že v argumente $(a,\,b) = (a^o,\,b^o)$ nadobúda funkcia SSO globálne minimum. Hodnotou toho minima je hodnota tretieho člena, ktorý nie je závislý od a, resp. b. Sformulujme, čo sme dostali.

VETA. Nech $(x_i\,,\,y_i)$, pre $i = 1,\,2,\,\ldots,\,n$, sú také, že nie všetky (x_i) sú rovnaké. Potom rovnica *výberovej regresnej priamky* má tvar:

$$y \;=\; a^o + b^o x$$

kde $b^o = \dfrac{SS_{xy}}{SS_{xx}}$, $a^o = \overline{y} - b^o\,\overline{x}$.

Pre rovnicu regresnej priamky platí:

$$y - \overline{y} \;=\; \frac{SS_{xy}}{SS_{xx}}\,(x - \overline{x})$$

a to znamená, že môžeme písať:

$$y - \overline{y} \;=\; \frac{s_{xy}}{s_x^2}\,(x - \overline{x})$$

Získali sme tvar, v ktorom vystupujú výberová kovariancia a výberový rozptyl. Predelením oboch strán rovnice výberovou smerodajnou odchýlkou s_y, dosiahneme úplne symetrický tvar rovnice regresnej priamky, v ktorej bude vystupovať výberový korelačný koeficient:

$$\frac{y - \overline{y}}{s_y} \;=\; r_{xy}\,\frac{x - \overline{x}}{s_x}$$

♠ Teraz krátko diskutujeme štatistický obsah výrazov, ktoré vystupujú v procese odvádzania koeficientov výberovej regresnej priamky. Objasníme význam koeficientu determinácie a poukážeme na jeho súvis s korelačným koeficientom. Ide o zdôvodnenie (v štatistike veľmi zásadného) faktu:

Výberový korelačný koeficient je mierou lineárnej väzby medzi premennými x a y.

Začnime sumarizáciou vzťahov, ktoré sme získali v procese odvádzania rovnice regresnej priamky. Globálne minimum funkcie SSO sa realizuje v bode (a^o, b^o), kde $b^o = \frac{SS_{xy}}{SS_{xx}}$, $a^o = \bar{y} - b^o \bar{x}$. Hodnota globálneho minima, t.j. hodnota $SSO(a^o, b^o)$ sa označuje symbolom SSE

$$SSE = SSO(a^o, b^o) = \sum [y_i - a^o - b^o x_i]^2$$

Hodnoty $a^o + b^o x_i$ sa nazývajú *opravené* hodnoty a označujeme ich y_i^o: $y_i^o = a^o + b^o x_i$. Potom zrejme

$$SSE = SSO(a^o, b^o) = \sum [y_i - y_i^o]^2 .$$

Hodnotu SSE (ktorá je zrejme vždy nezáporná) môžeme považovať za mieru nesúladu dát s lineárnym modelom. Ak by totiž lineárny model bol tým nepochybne správnym (ideálnym) modelom pre naše dáta a keby (v procese merania, resp. získavania dát) žiadne náhodné chyby prítomné neboli, tak body (x_i, y_i) by ležali na priamke a $SSE = SSO(a^o, b^o) = 0$.

Avšak, SSE treba modifikovať ("normovať"). Je zrejmé, že bez modifikácie, SSE má vadu – predstavme si, že namiesto v n bodoch, závislosť meriame ešte v ďalších n bodoch. Hoci by sa na okolnostiach nič nemenilo, SSE by sa (približne) zdvojnásobila!

Tou modifikáciou SSE bude podiel SSE/SS_{yy}. Všimnime si totiž, že z odvodeného vzťahu

$$SSO(a, b) = n(\bar{y} - a - b\bar{x})^2 + \left(\sqrt{SS_{xx}}\, b - \frac{SS_{xy}}{\sqrt{SS_{xx}}}\right)^2 + \left(SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}\right)$$

plynie

$$SSE = SSO(a^o, b^o) = \left(SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}\right)$$

odkiaľ zrejme máme: $SSE \le SS_{yy}$, a preto

$$0 \le \frac{SSE}{SS_{yy}} \le 1$$

Malé hodnoty SSE/SS_{yy} hovoria v prospech lineárneho modelu, kým hodnoty blízke 1 v neprospech. Ak zavedieme *koeficient determinácie* vzťahom

$$R^2 = 1 - \frac{SSE}{SS_{yy}}$$

tak platí: Malé hodnoty R^2 hovoria v neprospech lineárneho modelu, kým hodnoty blízke 1 v prospech lineárneho modelu. Preto

R^2 je koeficient, ktorého hodnoty informujú o miere súladu dát s lineárnym modelom.

(R^2 nie je mocnina, je to skrátka symbol pre výraz, ktorého hodnota je vždy z intervalu [0, 1]). Nakoniec, všimnime si súvis medzi R^2 a r_{xy}:

$$R^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{\frac{SS_{xy}^2}{SS_{xx}}}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}} = r_{xy}^2$$

a práve toto je argument, ktorý zdôvodňuje, prečo výberový korelačný koeficient r_{xy} môžeme považovať za mieru lineárnej väzby medzi x a y.

Úlohy

2.5.1 Uvažujte o kontingenčnej tabuľke rozmeru [2, 3], keď zachytáva empirické údaje získané prieskumom, ktorý vyšetroval záujem stredoškolákov o typ ďalšieho štúdia (veličina Y). Nech znakom X je pohlavie a znak Y má štyri kategórie: (zamerania) prírodovedné, technické, humanitné a iné. Predpokladajme, že prieskum sa urobil na náhodnom výbere 200 stredoškolákov. Zostavte (vymyslite, navrhnite) kontingenčné tabuľky tak, aby

a) znaky X, Y boli skôr nezávislé
b) znaky X, Y boli skôr závislé

(vzhľadom na to, že otázku závislosti/nezávislosti nevieme jednoznačne zodpovedať, ide len o intuitívne riešenie, ktorého hodnovernosť je otázna, a preto sme použili slovíčko "skôr").

Riešenie: V nasledujúcej kontingenčnej tabuľke, pomery v prvom riadku sa približne rovnajú pomerom v druhom a tiež v treťom riadku, čo odpovedá situácii, keď pohlavie neovplyvňuje rozloženie záujmu o typ ďalšieho štúdia a to znamená, že na základe takýchto dát by sme znaky X, Y považovali za nezávislé.

	prírodov.	technické	humanitné	iné	
chlapci	18	32	39	11	100
dievčatá	22	28	41	9	100
	40	60	80	20	

V nasledujúcej tabuľke to tak, zrejme, nie je. Pôjde o situáciu, v ktorej evidentne pohlavie ovplyvňuje záujem o typ štúdia, a teda v prípade takýchto dát by sme znaky považovali za stochasticky závislé.

	prírodov.	technické	humanitné	iné	
chlapci	30	40	20	10	100
dievčatá	20	10	60	10	100
	50	50	80	20	

2.5.3 V tejto úlohe pôjde o závislosť, či nezávislosť kvantitatívnych znakov X a Y. V tabuľkách sú uvedené početnosti n_{ij} tried, do ktorých boli zatriedené hodnoty veličín X a Y. Analyzujte hodnoty v kontingenčnej tabuľke a vyslovte hypotézu o závislosti, či nezávislosti znakov (veličín) X a Y.

(a) X (priemer stĺpika v cm), triedy: (1): menej ako 1.95 cm; (2): [1.95—2.05]; (3): viac ako 2.05 cm. Analogicky tri triedy pre hodnoty Y (výška stĺpika v cm).

X\Y	1	2	3
1	3	10	4
2	9	48	8
3	2	12	4

Danú tabuľku upravíme, rozšírime o súčty. Tak dostaneme rozloženie znaku bez uvažovania toho druhého. Rozloženie 14 – 70 – 16 je rozloženie znaku Y bez uvažovania hodnôt znaku X.

X\Y	1	2	3	
1	3	10	4	17
2	9	48	8	65
3	2	12	4	18
	14	70	16	

Môžeme povedať: hodnoty znaku Y sú (približne) v pomere 1: 5: 1 (lebo početnosti sú: 14 – 70 – 16) neuvažujúc hodnoty znaku X.
Pre X = 2 (a tiež, pre X = 3) sú početnosti (približne) v rovnakom pomere a to signalizuje nezávislosť znakov. Avšak pre X = 1, sme na vážkach, namiesto početnosti 10 by sme tam chceli vidieť viac: 13 až 17. Je evidentné, že v reálnych situáciách potrebujeme nejaký postup, objektívny postup, ktorý vylučuje takéto subjektívne posudzovanie. Taký postup ponúkajú štatistické testy, ktorými sa objektívne testuje nezávislosť znakov.

(b) Zostavte kontingenčnú tabuľku z údajov o priemere kmeňov javorov D a ich výšky V z úlohy 1.6.7
Kategorizujte D do tried: [16 – 18), [18 – 20), [20 – 22), [22 – 24), [24 – 26) a
V do tried: [5 – 6), [6 – 7), [7 – 8).

Táto časť úlohy 2.5.3 má za cieľ poukázať na prácu, ktorú je treba vynaložiť na získanie kontingenčnej tabuľky z daného súboru dát. Pri "ručnom" riešení tejto úlohy si uvedomíte, že zatrieďovanie daných údajov (v našom prípade len 35 dvojíc) je dosť prácne. Samozrejme, že v MATLABe môžeme tento proces algoritmizovať. Nasledujúca funkcia priradí vektoru dát, vektor jeho "kódov" (každému údaju priradíme index triedy, do ktorej patrí). Potom môžeme využiť M-funkciu "crosstab" (štatistického toolboxu), ktorej hodnotou je kontingenčná tabuľka.

```
function xx = kodujdata(x,ht)
% argin  x  je vektor dat;
%        ht je vektor hranic tried: ht(1)<= xmin, xmax < ht(k)
% argout xx - vektor kodov: xx(i)=j prave ak  ht(j)<= x(i)< ht(j+1)
% ------------------------------------------------------------
n = length(x);
k = length(ht);
xx = zeros(size(x));
for i= 1:n
    for j= 1:k-1
           if ht(j)<= x(i)
              xx(i)= xx(i) + 1;
           end
    end
end
```

Riešenie. Ak x je vektor priemerov kmeňov stromov, tak výstup funkcie "kodujdata" dáva vektor xx

```
xx
xx =  2     3     2     1     3     2     4     4     1     3     2     4
      3     5     4     3     3     4     2     3     2     3     2     3
      1     2     3     2     1     3     5     1     3     3     4
```

Analogicky, pre y, vektor výšok stromov, resp. pre vektor yy zakódovaných dát máme:

```
yy
yy =  2     2     1     1     2     2     3     3     1     2     2     3
      2     3     3     2     2     3     2     2     2     2     2     2
      1     2     2     2     1     3     3     1     2     2     3
```

Výstup štatistickej funkcie "crosstab" dáva maticu kontingenčnej tabuľky:

```
crosstab(xx, yy)

ans =
     5     0     0
     1     8     0
     0    12     1
     0     0     6
     0     0     2
```

a to znamená, že hľadaná kontingenčná tabuľka má tvar:

	[5, 6)	[6, 7)	[7, 8)	
[16, 18)	5	0	0	5
[18, 20)	1	8	0	9
[20, 22)	0	12	1	13
[22, 24)	0	0	6	6
[24, 26)	0	0	2	2
	6	20	9	

2.5.4 Nasledujúce údaje predstavujú hodnoty veličín X, Y, kde X je váha osobného automobilu a Y je jeho spotreba (reálne dáta pochádzajú z náhodného výberu uskutočneného v 80-tych rokoch minulého storočia). Ukážte, že závislosť Y na X môže byť dobre popísaná regresnou priamkou. Nájdite jej rovnicu, graficky znázornite rezíduá a vypočítajte hodnotu SSE. Vyjadrite sa k vhodnosti lineárneho modelu.

váha	1550	860	1225	905	1315	1185	995	1860	1720	1580
spotreba	11.5	7.3	8.5	6.7	10.8	8.5	7.8	15.4	14.0	13.0

```
vaha
vaha =
       1550     860    1225     905    1315    1185     995    1860    1720    1580
spot
spot =
      11.50    7.30    8.50    6.70   10.80    8.50    7.80   15.40   14.00   13.00
kovmat = cov(vaha, spot)
kovmat =
         1.0e+005 *
                  1.2098    0.0103
                  0.0103    0.0001

b0 = kovmat(1,2)/var(vaha)
b0 =
     0.0086

a0 = mean(spot) - b0*mean(vaha)
a0 =
    -0.9327
```

Rovnica výberovej regresnej priamky má tvar $y = -0.9327 + 0.0086\,x$
Reziduálny súčet štvorcov získame napr. takto:

```
SSE = sum( (spot - (a0 + b0*vaha) ).^2 )
SSE =
      3.7354
```

to znamená, že sme postupovali podľa vzťahu SSE $=$ SSO$(a^o, b^o) = \sum [y_i - a^o - b^o x_i]^2$. Na SSE môžeme nazerať ako na druhú mocninu vzdialenosti medzi dvomi vektormi:
vektorom $y = (y_i)$ pôvodných dát a
vektorom $y^o = (y_i^o)$, $y_i^o = a^o + b^o x_i$, t.j. vektorom opravených dát.

Druhá odmocnina z SSE je teda norma (euklidovská norma) vektora rezíduí: (r_i), $r_i = y_i - a^o - b^o x_i$. MATLAB túto hodnotu uvádza ako *normu rezíduí* (k jej hodnote sa dostanete cez Figure window – Tools – Basic fitting – linear fitting, vyplnením dialogového okna). V našej jednoduchej situácii ju môžeme získať jednoducho:

```
sqrt(SSE)
ans =
      1.9327
```

Vhodnosť lineárneho modelu potvrdzuje hodnota výberového korelačného koeficienta:

```
corrcoef(vaha, spot)
ans =
      1.0000    0.9773
      0.9973    1.0000
```

Výberový korelačný koeficient sa rovná 0.977 a taká hodnota potvrdzuje adekvátnosť lineárneho modelu.
Zobrazenie situácie dáva obrázok "cars80.fig" (otvoríte ho MATLABom), alebo "cars80.jpg".

Rovnako ako 2.5.4, sa riešia úlohy 2.5.5 a 2.5.6. Úlohy 2.5.7 a 2.5.8 sú jednoduché.