KORELAČNÁ ZÁVISLOSŤ

REGRESNÁ ÚLOHA

- ak chceme charakterizovať korelačnú závislosť medzi kvantitatívnymi parametrami musíme riešiť regresnú úlohu, teda charakterizovať regresiu:
  - správne vystihnúť charakter závislosti medzi závisle premennou a nezávisle premennou veličinou, teda zvoliť vhodný typ regresnej funkcie
  - odhadnúť jej parametre

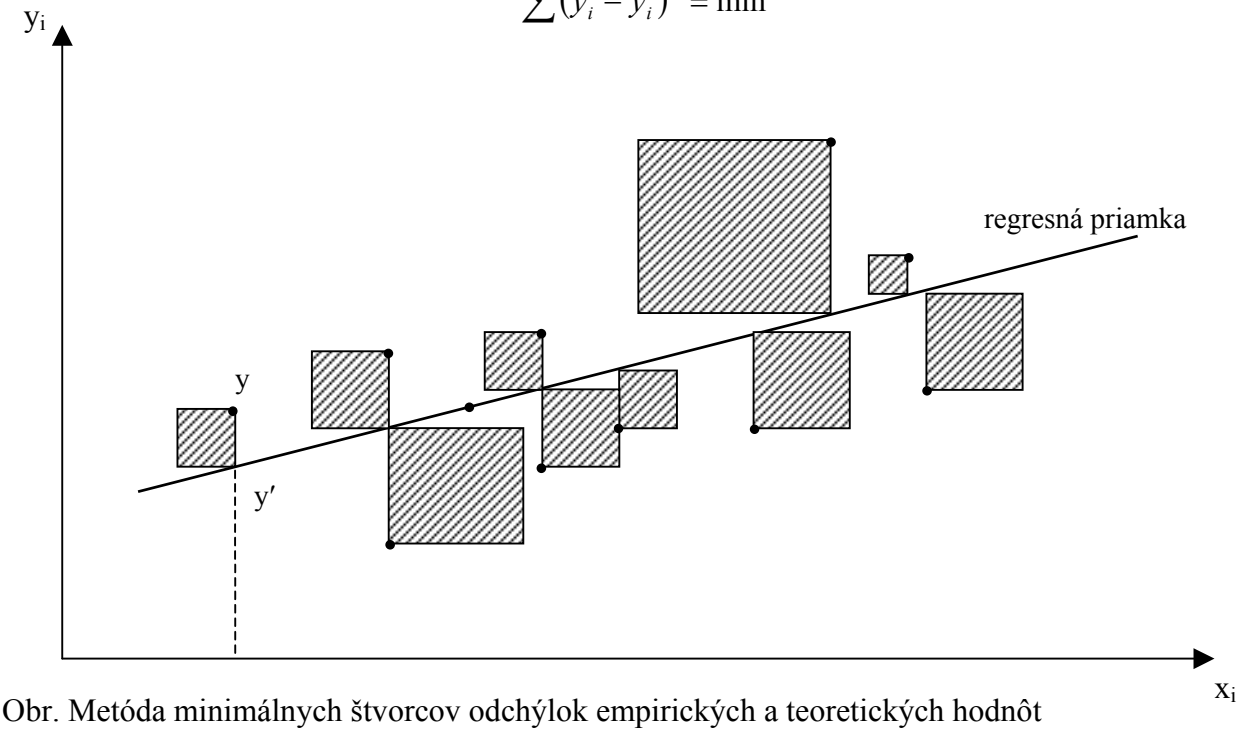
♦ *METÓDA NAJMENŠÍCH ŠTVORCOV*

- najjednoduchšou formou korelácie je lineárna korelácia medzi dvoma kvantitatívnymi znakmi, teda jednoduchá lineárna korelácia
- najčastejšie používaná metóda odhadu regresných funkcií pre takúto jednoduchú (lineárnu) korelačnú závislosť je **metóda najmenších štvorcov**
- východiskom pre odhad parametrov regresnej funkcie sú empirické údaje, pričom hodnoty závislej premennej sa označujú ako y_i
- teoretické (ideálne) hodnoty - vyrovnané hodnoty, t.j. hodnoty ležiace na priamke sa označujú ako y'_i
- z rôznych možností, ktorými možno preložiť priamku cez body v korelačnom diagrame je najvhodnejšia tá alternatíva, pri ktorej sa súčet odchýlok empirických hodnôt od teoretických hodnôt bude rovnať 0:

$$\sum (y_i - y'_i) = 0$$

- pre použitie vo všeobecnosti sa táto podmienka upravila – súčet štvorcov odchýlok empirických (skutočných) hodnôt od teoretických hodnôt má byť minimálny:

$$\sum (y_i - y'_i)^2 = \min$$



Obr. Metóda minimálnych štvorcov odchýlok empirických a teoretických hodnôt

- pretože rovnica priamky má tvar:

$$y' = a + bx_i$$

môžeme príslušnú funkciu prepísať na tvar:

$$f(a,b) = \sum \left(y_i - a - bx_i\right)^2$$

- parametre rovnice priamky a,b je možné vypočítať z rovníc (úpravou, ktorá je založená na tom, že prvá derivácia funkcie F(a,b) podľa obidvoch veličín je rovná nule a z následných normálnych rovníc s použitím determinantov):

$$a = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

- keď vypočítame parametre a, b, konkrétne hodnoty potom dosadíme do všeobecného tvaru rovnice priamky: $y' = a + bx_i$

- táto **vyrovnávajúca** priamka sa nazýva **regresná priamka** a umožňuje odhadovať z veľkosti jednej premennej veľkosť druhej premennej

- Veličina *b* z regresnej priamky je vlastne smernicou priamky a označuje sa ako **regresný koeficient**. Je to základný ukazovateľ pri uskutočňovaní korelačnej analýzy, pretože podáva informácie o priebehu závislosti, teda o koľko sa v priemere zmení závisle premenná veličina y_i pri zmene nezávisle premennej veličiny x_i. V prípade priamej závislosti je regresný koeficient kladný (rastúca priamka) a v prípade nepriamej závislosti je záporný (priamka je klesajúca).

- Konštanta *a* pri grafickom zobrazení regresnej priamky určuje bod, v ktorom priamka pretína os *y*. Jej hlavný zmysel je, že posúva regresnú priamku v priestore, preto sa označuje aj ako **lokujúca konštanta**.

KORELAČNÁ ÚLOHA

- vzťah medzi premennými veličinami môže mať rôznu intenzitu (od úplnej nezávislosti až po úplnú závislosť)Ak chceme určiť stupeň závislosti medzi premennými musíme riešiť korelačnú úlohu.

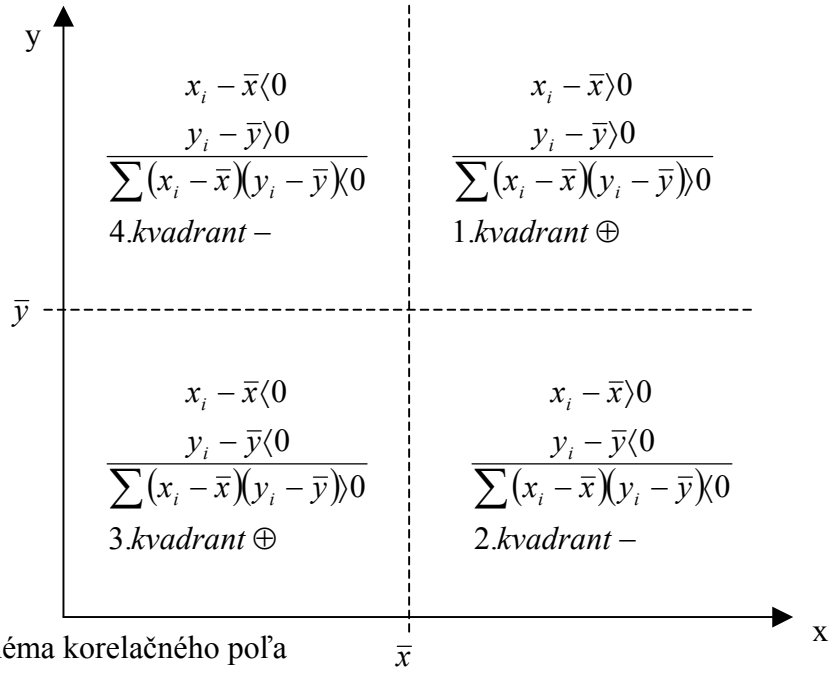
- stupeň závislosti medzi premennými charakterizujú **MIERY TESNOSTI ŠTATISTICKEJ ZÁVISLOSTI**:

♦ *KORELAČNÝ* KOEFICIENT *r*

- určuje mieru tesnosti (stupňa) závislosti

- jeho definovanie vychádza z úvah o súčte odchýlok jednotlivých hodnôt obidvoch korelovaných znakov od ich priemerov

- ak sú body rozptýlené rovnomerne vo všetkých kvadrantoch, celkový súčet je blízky nule a znaky sú teda na sebe nezávislé. Ak sú body rozložené okolo uhlopriečky, ide o závislosť medzi znakmi, ktorá je tým tesnejšia čím viac budú body priliehať k uhlopriečke. Rozmiestnenie v 3. a 1. kvadrante svedčí o kladnom smere závislosti, v 4. a 2. kvadrante o zápornej závislosti.



Obr. Schéma korelačného poľa

- na výpočet koeficientu korelácie sa používajú viaceré vzťahy:

$$r = \frac{\sum x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left(\sum x_i^2 - n \cdot \bar{x}^2\right) \cdot \left(\sum y_i^2 - n \cdot \bar{y}^2\right)}}$$

- Ak je r = 1, závislosť je úplná priama; r = -1 korelácia je úplná nepriama ak je r = 0, medzi veličinami je nezávislosť. Presnejšie:
r<0,3 nízka tesnosť
0,3≤r<0,5 mierna tesnosť
0,5≤r<0,7výrazná tesnosť
0,7≤r<0,9 vysoká tesnosť
0,9≤r veľmi vysoká tesnosť

- na zistenie spoľahlivosti hodnoty koeficientu korelácie sa používa tzv. **stredná chyba koeficientu korelácie σ_r :**

$$\sigma_r = \frac{1 - r^2}{\sqrt{n}}$$

n – počet dvojíc hodnôt znakov medzi ktorými meriame závislosť
- koeficient korelácie je spoľahlivou mierou tesnosti závislosti vtedy keď je väčší ako trojnásobok teoretickej strednej chyby, teda:

$$3 \cdot \sigma_r \langle r$$

♦ ***KOEFICIENT DETERMINÁCIE*** r^2
- je veľmi dôležitý pre hodnotenie stupňa závislosti
- vyjadruje podiel rozptylu teoretických hodnôt závisle premennej z rozptylu empirických hodnôt závisle premennej
- teoretické hodnoty sú odhadnuté na základe regresnej priamky
- koeficient korelácie je druhou odmocninou koeficientu determinácie
- stupne tesnosti závislosti podľa koeficientu determinácie:

$r^2 < 10\%$ nízka tesnosť
$10\% \le r^2 < 25\%$ mierna tesnosť
$25\% \le r^2 < 50\%$ výrazná tesnosť
$50\% \le r^2 < 80\%$ vysoká tesnosť
$80\% \le r^2$ veľmi vysoká tesnosť

♦ ***KORELAČNÝ POMER*** η_{yx}

- najvšeobecnejšia miera tesnosti závislosti, ktorú možno vypočítať bez ohľadu na to, či sa už riešila regresná úloha:

$$\eta_{yx} = \sqrt{\frac{\sum_{i=1}^{m} \bar{y}_i^2 n_i - n\bar{y}^2}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij}^2 - n\bar{y}^2}}$$

- korelačný pomer nadobúda hodnoty (0,1)
- v prípade lineárnej závislosti korelačný koeficient a korelačný pomer sú približne rovnaké, v prípade nelineárnej závislosti korelačný pomer je vyšší ako koeficient korelácie

♦ ***INDEX KORELÁCIE*** i_{yx}

- určuje tesnosť závislosti v prípade nelineárnej korelácie

$$i_{yx} = \sqrt{\frac{\sum_{j=1}^{n} \left(y'_j - \bar{y}\right)^2}{\sum_{j=1}^{n} \left(y_j - \bar{y}\right)^2}}$$

- nadobúda hodnoty z intervalu (0,1), pričom čím viac sa blíži jeho hodnota k jednej, tým ide o tesnejšiu závislosť

VÝPOČET KORELAČNEJ ZÁVISLOSTI Z ÚDAJOV USPORIADANÝCH V KORELAČNEJ TABUĽKE

- v tomto prípade musíme brať do úvahy početností jednotlivých hodnôt znaku a vzhľadom na upraviť výrazy pre výpočet jednotlivých charakteristík (niektoré súčty sa zapíšu inou formou)

➢ korelačný pomer

$$\eta_{yx} = \sqrt{\frac{n\sum_{i=1}^{k} \frac{1}{n_i}\left(\sum_{j=1}^{l} y_j n_{ij}\right)^2 - \left(\sum_{j=1}^{l} y_j n_j\right)^2}{n\sum_{j=1}^{l} y_j^2 n_j - \left(\sum_{j=1}^{l} y_j n_j\right)^2}}$$

➢ index korelácie

$$i_{yx} = \sqrt{\frac{n\sum_{i=1}^{k} y_i'^2 n_i - \left(\sum_{j=1}^{l} y_j n_j\right)^2}{n\sum_{j=1}^{l} y_j^2 n_j - \left(\sum_{j=1}^{l} \left(y_j n_j\right)\right)^2}}$$

➢ koeficient korelácie

$$r_{yx} = r_{xy} = \frac{n\sum_{i=1}^{k}\sum_{j=1}^{l} x_i y_j n_{ij} - \left(\sum_{i=1}^{k} x_i n_i\right)\left(\sum_{j=1}^{l} y_j n_j\right)}{\sqrt{\left[n\sum_{i=1}^{k} x_i^2 n_i - \left(\sum_{i=1}^{k} x_i n_i\right)^2\right]\left[n\sum_{j=1}^{l} y_j^2 n_j - \left(\sum_{j=1}^{l} y_j n_j\right)^2\right]}}$$

kde

$$\sum_{i=1}^{k}\sum_{j=1}^{l} x_i y_j n_{ij} = \sum_{i=1}^{k}\left(x_i \sum_{j=1}^{l} y_j n_{ij}\right) = \sum_{j=1}^{l}\left(y_j \sum_{i=1}^{k} x_i n_{ij}\right)$$

pričom

$$\sum_{j=1}^{l} y_j n_{1j} = y_1 n_{11} + y_2 n_{12} + \ldots + y_l n_{1l} \text{ atď}$$

- pri výpočte charakteristík z korelačnej tabuľky sa využíva metóda vhodne zvoleného počiatku
- ak transformujeme

$$v_i = \frac{x_i - a}{h} \qquad\qquad v'_j = \frac{y_j - b}{g}$$

kde *a,h, b,g* sú voľne zvolené konštanty
stačí vypočítať korelačný koeficient medzi pomocnými premennými:

$$r_{vv'} = \frac{n\sum_{i=1}^{k}\sum_{j=1}^{l} v_i v'_j n_{ij} - \left(\sum_{i=1}^{k} v_i n_i\right)\left(\sum_{j=1}^{l} v'_j n_j\right)}{\sqrt{\left[n\sum_{i=1}^{k} v_i^2 n_i - \left(\sum_{i=1}^{k} v_i n_i\right)^2\right]\left[n\sum_{j=1}^{l} v'^2_j n_j - \left(\sum_{j=1}^{l} v'_j n_j\right)\right]}}$$

kde

$$\sum_{i=1}^{k}\sum_{j=1}^{l} v_i v'_j n_{ij} = \sum_{i=1}^{k}\left(v_i \sum_{j=1}^{l} v'_j n_{ij}\right) = \sum_{j=1}^{l}\left(v'_j \sum_{i=1}^{k} v_i n_{ij}\right)$$