13. prednáška

Korelačná a regresná analýza

V predchádzajúcich úvahách sme na jednotkách štatistického súboru sledovali jeden znak X. Ak na každom prvku daného súboru pozorujeme dva znaky X a Y, hovoríme o tzv. *dvojrozmernom rozdelení*. V tejto časti prednášky nás bude zaujímať, či medzi pozorovanými znakmi existuje štatistická závislosť, a ak áno, aký je jej stupeň.
Napr. pri meraní výšky žiakov deviatych ročníkov vybranej školy a ich hmotnosti sme získali údaje:

Žiak	$Výška(X)$	$Hmotnosť(Y)$
1	164	49
2	175	68
3	177	72
4	168	55
5	172	53
$\vdots$	$\vdots$	$\vdots$

Získané údaje môžeme znázorniť v súradnicovom systéme:

Závislosť medzi pozorovanými znakmi X a Y môže mať rôzny charakter, napr.

Predpokladajme, že máme náhodný výber o rozsahu n, pričom sme na každom prvku merali dva kvantitatívne znaky X, Y. Je prirodzené definovať ako mieru štatistickej závislosti znakov X, Y **výberový koeficient korelácie premenných** X, Y ako náprotivok $\rho(X, Y)$, ktorý poznáme z pravdepodobnosti. Tam sme definovali korelačný koeficient vzťahom:

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

Teoretické charakteristiky odhadneme z empirických hodnôt:

$$\widehat{cov(X, Y)} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$\widehat{\sigma_1^2} = S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{a} \quad \widehat{\sigma_2^2} = S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Výberový koeficient korelácie potom vyzerá takto:

$$r(X, Y) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}},$$

po úprave

$$r(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Poznámka:
Ak (X, Y) je náhodný vektor s dvojrozmerným **normálnym** rozdelením, potom z rovnosti $\rho(X, Y) = 0$ vyplýva, že náhodné premenné X a Y sú nezávislé. Vo všeobecnosti to neplatí!!!!!

Test významnosti koeficienta korelácie

Budeme predpokladať, že pozorujeme dva kvantitatívne znaky X, Y na n prvkoch náhodného výberu. Ďalej predpokladáme, že (X, Y) má dvojrozmerné normálne rozdelenie pravdepodobnosti a nech $\rho(X, Y)$ je koeficient korelácie.
Budeme testovať hypotézu:
$H_0 : \rho = 0$
proti
$H_a : \rho \neq 0$, resp. $H_a : \rho > 0$, resp. $H_a : \rho < 0$

Testovaná hypotéza H_0 je hypotézou o nezávislosti pozorovaných znakov X a Y a alternativna hypotéza H_a je hypotézou, že medzi znakmi X a Y existuje signifikantná štatistická závislosť. Testovacím kritériom je premenná

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$$

Oblasti zamietnutia hypotézy H_0 sú uvedené v nasledujúcej tabuľke:

H_a	$\rho \neq 0$	$\rho > 0$	$\rho < 0$
W_α	$(-\infty; -t_\alpha > \cup < t_\alpha; \infty)$	$< t_\alpha; \infty)$	$(-\infty; -t_\alpha >$
krit. hodnota	$F_{t(n-2)}(t_\alpha) = 1 - \frac{\alpha}{2}$	$F_{t(n-2)}(t_\alpha) = 1 - \alpha$	$F_{t(n-2)}(t_\alpha) = 1 - \alpha$

Jednoduchá (párová) lineárna regresia

Nech je daný náhodný výber o rozsahu n, pričom na každom jeho prvku sme merali dva kvantitatívne znaky X, Y a predpokladáme, že náhodný vektor (X, Y) má dvojrozmerné normálne rozdelenie. Výsledkom merania je postupnosť usporiadaných dvojíc reálnych čísel $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
Ak je koeficient korelácie medzi skúmanými znakmi X, Y štatisticky významný, bude nás zaujímať závislosť medzi nimi z hľadiska **regresie**. Teda našou snahou bude odhadnúť hodnoty znaku Y (tzv. *závislej premennej*) na základe daných hodnôt znaku X (tzv. *nezávislej premennej*). To znamená, že pri skúmaní štatistickej závislosti Y na X sa pokúsime nájsť vhodný matematický model (funkciu), v ktorom je vyjadrená predstava o tejto závislosti.
Z pravdepodobnosti vieme, že regresná funkcia je podmienená stredná hodnota náhodnej premennej Y, t.j. $\overline{y}(x) = E(Y/X = x) = f(x; a_0, a_1, \ldots, a_k)$.
Predpokladáme, že tvar funkcie $f(x; a_0, a_1, \ldots, a_k)$ poznáme a to, čo nepoznáme, sú koeficienty $a_0, a_1, \ldots, a_k$.
V rámci nášho kurzu sa budeme zaoberať tým najjednoduchším prípadom, keď **regresnou krivkou** je **priamka**.
Keby sa nám podarilo odstrániť spolupôsobenie vedľajších vplyvov na vzťah premenných X a Y, ležali by všetky body $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ na priamke $f(x) = a_0 + a_1 x$, čo je deterministický model.
Na premennú Y však okrem premennej X pôsobia aj iné faktory, preto body $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ neležia na priamke, ale kolíšu okolo nej. To sa snažíme zachytiť aj v matematickom modeli. Preto každú hodnotu závislej premennej Y rozložíme na dve zložky, na **deterministickú** a **náhodnú**, t.j.

$$y_i = a_0 + a_1 x_i + e_i \qquad i = 1, 2, \ldots, n,$$

kde

- y_i ... je i–ta hodnota premennej Y,
- a_0 ... je priesečník osi y s regresnou priamkou,
- a_1 ... je **regresný koeficient** (smernica regresnej priamky) a udáva, o koľko sa zmení Y, ak sa X zmení o jednotku,
- x_i ... je i–ta hodnota premennej X,
- e_i ... je i–ta **náhodná chyba** premennej X.

Bodmi $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ treba preložiť tzv.**vyrovnávajúcu priamku**, ktorá je daná vzťahom

$$\widehat{y_i} = \widehat{a_0} + \widehat{a_1} x_i \qquad i = 1, 2, \dots, n,$$

kde

- $\widehat{y_i}$ ...je očakávaná (vyrovnaná)hodnota premennej Y pre danú hodnotu premennej X,
- $\widehat{a_0}$ ...je bodový odhad koeficienta a_0,
- $\widehat{a_1}$ ...je bodový odhad koeficienta a_1,
- x_i ...je i–ta hodnota premennej X.

Rozdiely medzi medzi y_i a $\widehat{y_i}$ označujeme $\widehat{e_i}$, nazývame ich **rezíduami regresnej priamky** a interpretujeme ich ako bodové odhady náhodných chýb modelu $y_i = a_0 + a_1 x_i + e_i, \quad i = 1, 2, \dots, n$.
Našim cieľom je regresnú priamku voliť tak, aby rozdiely $\widehat{e_i} = y_i - \widehat{y_i}$ boli minimálne. A to je podstata **metódy najmenších štvorcov**. Minimalizujeme:

$$S\check{S} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))^2$$

Odhady $\widehat{a_0}$, $\widehat{a_1}$ dostaneme riešením sústavy rovníc:

$$\frac{\partial S\check{S}}{\partial a_0} = 0, \qquad \frac{\partial S\check{S}}{\partial a_1} = 0.$$

Nájdeme stacionárne body a "máme to šťastie", že v nich SŠ nadobúda absolútne minimum. ☺

$$\frac{\partial S\check{S}}{\partial a_0} = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) = 0$$

$$\frac{\partial S\check{S}}{\partial a_1} = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)\, x_i = 0$$

Po úprave dostaneme:

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

$$\sum_{i=1}^{n} x_i y_i = a_0 \sum_{i=1}^{n} x_i + a_i \sum_{i=1}^{n} x_i^2$$

To je systém dvoch rovníc o dvoch neznámych a_0, a_1 a vyriešime ho Cramerovým pravidlom. ☺

Po nájdení rovnice regresnej priamky je potrebné overiť, či tento model je "kvalitný", či dobre vystihuje závislosť medzi X, Y. Pri riešení regresnej úlohy často prichádza do úvahy viacero typov regresných funkcií (kvadratická, logaritmická, exponenciálna, ...) preto sa skúma, ktorá z týchto funkcií "lepšie prilieha" výberovým údajom. To sa dá merať rôznymi charakteristikami. Jednou z nich je **výberový koeficient determinácie** R^2, ktorý je druhou mocninou výberového korelačného koeficienta, t.j. $R^2 = r^2(X, Y)$. Z viacero modelov "kvalitnejší" model je ten s vyšším koeficientom determinácie.

Príklad:
U deviatich náhodne vybraných otcov bola zistená ich výška a výška ich dospelých synov s týmito výsledkami:

Výška otcov x_i	174	180	176	168	182	188	176	177	174
Výška synov y_i	177	182	176	173	180	191	179	181	176

a) Nájdite výberový korelačný koeficient.
b) Za predpokladu normálneho rozdelenia (X, Y) otestujte na hladine významnosti $\alpha = 0,05$, či je rozdiel medzi výškou otcov a synov štatisticky významný.
c) Nájdite odhad regresnej priamky.