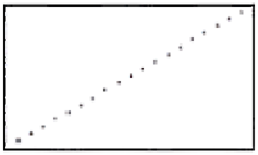
3. KORELAČNÍ KOEFICIENT

3.2. Věta

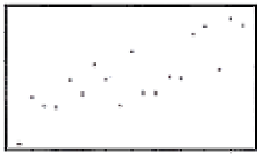
Pro koeficient korelace platí $-1 \leq r \leq 1$ a rovnosti je dosaženo právě když mezi hodnotami $x_1, \ldots, x_n$ a $y_1, \ldots, y_n$ existuje úplná lineární závislost, tj. existují konstanty a, b tak, že $y_i = a + bx_i$, $i = 1, \ldots, n$, přičemž znaménko + platí pro $b > 0$, znaménko – pro $b < 0$.

3.3. Poznámka

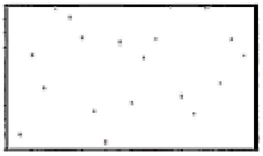
Představu o významu hodnot koeficientu korelace podávají následující dvourozměrné tečkové diagramy.



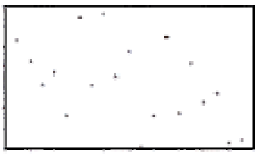
r = 1,00



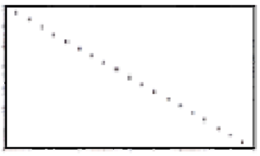
r = 0,76



r = 0,00



r = -0,37



r = -1,00

3.4. Příklad

Pro datový soubor z příkladu OCEL.STA vypočtěte

a) aritmetické průměry znaků X, Y
b) rozptyly a směrodatné odchylky znaků X, Y
c) koeficient korelace znaků X, Y Statistika – základní statistika/tabulky – korelační matice – 1seznam proměnných

Řešení:

ad a) $m_1 = 95{,}9$, $m_2 = 114{,}4$.
ad b) $s_1^2 = 1070{,}24$, $s_2^2 = 1075{,}12$, $s_1 = 32{,}71$, $s_2 = 32{,}79$.
ad c) $r = 0{,}936$.
Koeficient korelace svědčí o tom, že mezi oběma znaky existuje velmi silná přímá lineární závislost – čím je vyšší mez plasticity, tím je vyšší mez pevnosti a čím je nižší mez plasticity, tím je nižší mez pevnosti.

Při výpočtu číselných charakteristik se v řadě situací uplatní věta shrnující některé jejich vlastnosti. Pro lepší pochopení uvedených vlastností slouží následující příklad.

4. Regresní přímka

(Jak vyjádřit závislost mezi dvěma znaky?)

Po prostudování této kapitoly budete umět:

- stanovit odhady parametrů regresní přímky a znát jejich význam
- posoudit kvalitu proložení regresní přímky dvourozměrným tečkovým diagramem
- vypočítat regresní odhady závisle proměnného znaku
- stanovit odhady parametrů druhé regresní přímky
- znát vztahy mezi parametry první a druhé regresní přímky.

Pro zvládnutí této kapitoly budete potřebovat 3 – 4 hodiny studia.

Budeme se zabývat speciálním případem, kdy hodnoty znaku Y závisejí na hodnotách znaku X přibližně lineárně.Ukážeme si, jak tuto závislost popsat regresní přímkou, jak odhadnout její parametry metodou nejmenších čtverců na základě znalosti dvourozměrného datového souboru a jak posoudit kvalitu regresní přímky pomocí indexu determinace. Vysvětlíme si význam regresních parametrů a v příkladu se budeme zabývat regresní přímkou meze pevnosti na mez plasticity.

4.1. Motivace

Cílem regresní analýzy je vystižení závislosti hodnot znaku Y na hodnotách znaku X. Při tom je nutné vyřešit dva problémy: jaký typ funkce použít k vystižení dané závislosti a jak stanovit konkrétní parametry zvoleného typu funkce? Typ funkce určíme buď logickým rozborem zkoumané závislosti nebo se snažíme ho odhadnout pomocí dvourozměrného tečkového diagramu. Zde se omezíme na lineární závislost $y = \beta_0 + \beta_1 x$. Odhady b_0 a b_1 neznámých parametrů β_0, β_1 získáme na základě dvourozměrného datového souboru $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$ metodou nejmenších čtverců. Požadujeme, aby průměr součtu čtverců odchylek skutečných a odhadnutých hodnot byl minimální, tj. aby výraz $\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ nabýval svého minima vzhledem k β_0 a β_1. Tento výraz je minimální, jsou-li jeho první derivace podle β_0 a β_1 nulové. Stačí tyto derivace spočítat, položit je rovny 0 a řešit systém dvou rovnic o dvou neznámých, tzv. systém normálních rovnic.

4.2. Definice

Nechť je dán dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$ a přímka $y = \beta_0 + \beta_1 x$. Výraz $q(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ se nazývá *rozptyl hodnot znaku Y kolem* *přímky* $y = \beta_0 + \beta_1 x$. Přímka $y = \beta_0 + \beta_1 x$, jejíž parametry minimalizují rozptyl $q(\beta_0, \beta_1)$ v celém dvourozměrném prostoru, se nazývá *regresní přímka znaku Y na znak X. Regresní odhad* i-té hodnoty znaku Y značíme $\hat{y}_i = b_0 + b_1 x_i$, $i = 1, ..., n$. Kvadrát koeficientu korelace znaků X, Y se nazývá *index determinace* a značí se ID^2 (Index determinace udává, jakou část variability hodnot znaku Y vystihuje regresní přímka. Nabývá hodnot z intervalu $\langle 0,1 \rangle$. Čím je bližší 1, tím lépe vystihuje regresní přímka závislost Y na X.)

4.3. Věta

Nechť $y = b_0 + b_1 x$ je regresní přímka znaku Y na znak X. Přitom úsek b_0 regresní přímky udává velikost jejího posunutí na svislé ose (tj.udává, jaký je regresní odhad hodnoty znaku Y, nabývá-li znak X hodnoty 0) a směrnice b_1 udává, o kolik jednotek se změní hodnota znaku Y, změní-li se hodnota znaku X o jednotku. Je-li $b_1 > 0$, dochází s růstem X k růstu Y a hovoříme o přímé závislosti hodnot znaku Y na hodnotách znaku X. Je-li $b_1 < 0$, dochází s růstem X k poklesu Y a hovoříme o nepřímé závislosti hodnot znaku Y na hodnotách znaku X.

4.4. Příklad

Pro datový soubor z příkladu OCEL.STA

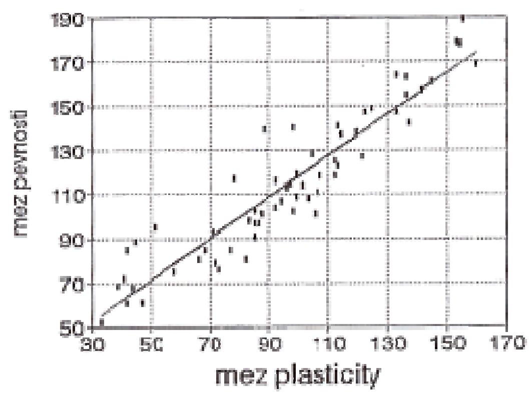
a) určete regresní přímku meze pevnosti na mez plasticity.
b) Zakreslete regresní přímku do dvourozměrného tečkového diagramu.
c) Jak se změní mez pevnosti, vzroste-li mez plasticity o jednotku?
d) Najděte regresní odhad meze pevnosti pro mez plasticity = 60.
e) Vypočtěte index determinace a interpretujte ho.

Řešení:

STATISTIKA - vícerozměrná regerse – proměnné - výběr závislé a nezávislé proměnné – výpočet výsledky regrese

ad a) $y = 24{,}5 + 0{,}937x$.

ad b)



Povšimněte si, že koeficient korelace znaků X a Y vypočtený v příkladě OCEL.STA. činil 0,936. Tato hodnota je blízká 1, což svědčí silné přímé lineární závislosti mezi znaky X a Y. Tečky v dvourozměrném tečkovém diagramu nejsou příliš rozptýleny kolem regresní přímky.

ad c) Mez pevnosti vzroste o 0,937 $kpcm^{-2}$ – viz parametr b_1 vypočtený v bodě (a) ad d) $\hat{y} = 24{,}5 + 0{,}937 \times 60 = 80{,}72$.

ad e) $ID^2 = r_{12}^2 = 0{,}936^2 = 0{,}876$. Znamená to, že 87,6% variability hodnot meze pevnosti je vysvětleno regresní přímkou.

Shrnutí

Pokud vzhled dvourozměrného tečkového diagramu svědčí o existenci určitého stupně lineární závislosti znaku Y na znaku X, můžeme tímto diagramem proložit **regresní přímku** znaku Y na znak X. (Pozor – nelze se spokojit pouze s výpočtem korelačního koeficientu, je nutné grafické posouzení závislosti.) Její parametry (tj. posunutí a směrnici) odhadujeme **metodou nejmenších čtverců**. Kvalitu proložení posuzujeme pomocí **indexu determinace** – čím je tento index bližší 1, tím je regresní přímka výstižnější a čím je bližší 0, tím je regresní přímka nevhodnější pro vystižení závislosti Y na X. Dosadíme-li danou hodnotu znaku X do rovnice regresní přímky, získáme **regresní odhad** příslušné hodnoty znaku Y.

Má-li smysl zkoumat též opačný směr závislosti, tj. X na Y, hledáme **druhou regresní přímku**. 1. a 2. regresní přímka se označují jako **sdružené regresní přímky**.

Kontrolní otázky a úkoly

1. V čem spočívá princip metody nejmenších čtverců?
2. Uveďte příklad dvourozměrného datového souboru z ekonomické praxe vhodný pro použití regresní přímky.
3. Co vyjadřuje index determinace a jak se počítá?
4. (S) U osmi náhodně vybraných studentů byly zjišťovány jejich matematické a verbální schopnosti. Výsledky matematického testu udává znak X, výsledky verbálního Y.

X	80	50	36	58	72	60	56	68
Y	65	60	35	39	48	44	48	61

   a) Vypočtěte koeficient korelace a interpretujte ho.
   b) Najděte rovnice sdružených regresních přímek.
   c) Zlepší-li se výsledek v matematickém testu o 10 bodů, o kolik bodů se zlepší výsledek ve verbálním testu?
   d) Zlepší-li se výsledek ve verbálním testu o 10 bodů, o kolik bodů se zlepší výsledek v matematickém testu?