KORELACE, JEJÍ VYUŽITÍ A ZNEUŽITÍ

JIŘÍ MILITKÝ , Katedra textilních materiálů, Technická universita v Liberci, 461 17 Liberec
MILAN MELOUN , Katedra analytické chemie, Universita Pardubice, Pardubice

1. Úvod

Mezi nejfrekventovanější úlohy statistického zpracování dat patří regrese a korelace [1]. Bohužel také u těchto úloh dochází velmi často k mylné interpretaci a (ne)záměrnému zneužití. V této práci je pozornost zaměřena na formálně nejjednodušší problém tj. párový korelační koeficient ρ a jeho odhad r. Jsou uvedeny základní pojmy a souvislosti korelačního koeficientu. Je učiněn pokus o objasnění některých překvapivých vlastností tohoto koeficientu v případech, kdy se použije na obecná data neodpovídající základním předpokladům (dvourozměrná normalita).

Sám objev koncepce regrese a korelace Francisem Galtonem na přelomu 20 - tého století patří k základním milníkům rozvoje statistiky. Jak je i u řady dalších základních objevů obvyklé, je dnes standardně klasický párový korelační koeficient označován jménem Pearsonův podle Karla Pearsona, který v r. 1920 precizoval myšlenky Galtona z roku 1888 (viz diskuse v [2]).

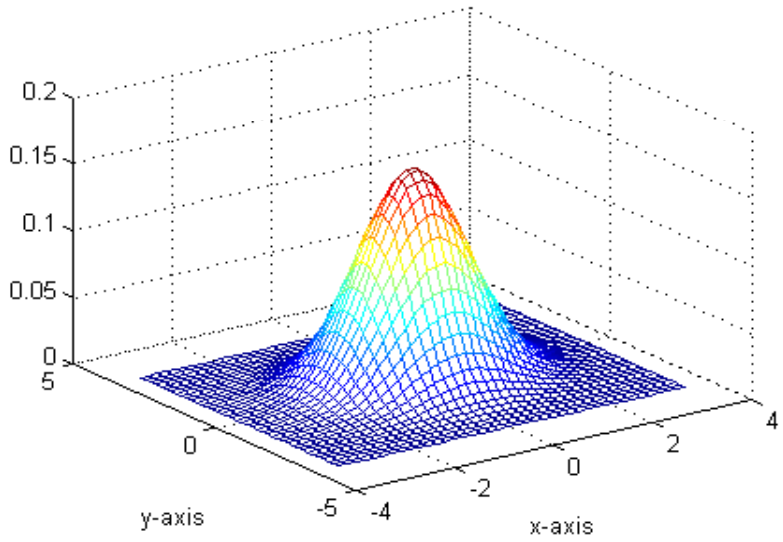
Galton ve svých pracích o dědičnosti zavedl elipsy konstantní hustoty a navrhl korelační koeficient ρ jako míru lineární regrese. Korelační koeficient identifikoval jako směrnici regresní přímky ve standardizovaných proměnných.

Pearson zpřesnil řadu Galtonových myšlenek, např. že pokud je dvourozměrná hustota pravděpodobnosti normální je regrese Y na X lineární se směrnicí $\rho \sqrt{D(Y)}/\sqrt{D(X)}$. Zde D(Y) a D(X) jsou rozptyly náhodných proměnných Y a X. Pearson také navrhl výpočet odhadů pro ρ a jejich chyb. Jeho žák Yule ukázal, že postačuje aby regrese byla lineární s konstantním rozptylem a směrnice bude mít stále tvar $\rho\sqrt{D(Y)}/\sqrt{D(X)}$. Pearson sám se bezúspěšně snažil o zobecnění korelačního koeficientu na případ nelineární heteroskedastické regrese. To se podařilo až Bjervemu a Doksumovi v r. 1993 [3]. Z dalších, kteří se zasloužili o rozvoj koncepce korelačního koeficientu je zřejmě nejvýznamnější R. A. Fisher, který určil rozdělení výběrového korelačního koeficientu. Ten také považoval korelační koeficient za přirozenou míru asociace mezi dvěma proměnnými a zavedl jeho použití ve studiích o genetice.

Řada prací z nejrůznějších oborů využívá a často zneužívá korelačního koeficientu pro vyjádření souvislostí mezi dvěma veličinami. Je to způsobeno hlavně tím, že většina nestatistiků chápe vysokou korelaci za potvrzení příčinné souvislosti a nulovou korelaci za nezávislost. I lidé poučení o základních souvislostech však často stojí před problémem typu jak interpretovat korelační koeficient ve speciálních případech a proč se vlastně chová tak zvláštně. Cílem této práce je pokus o odhalení některých známých a některých překvapivých vlastností korelačního koeficientu.

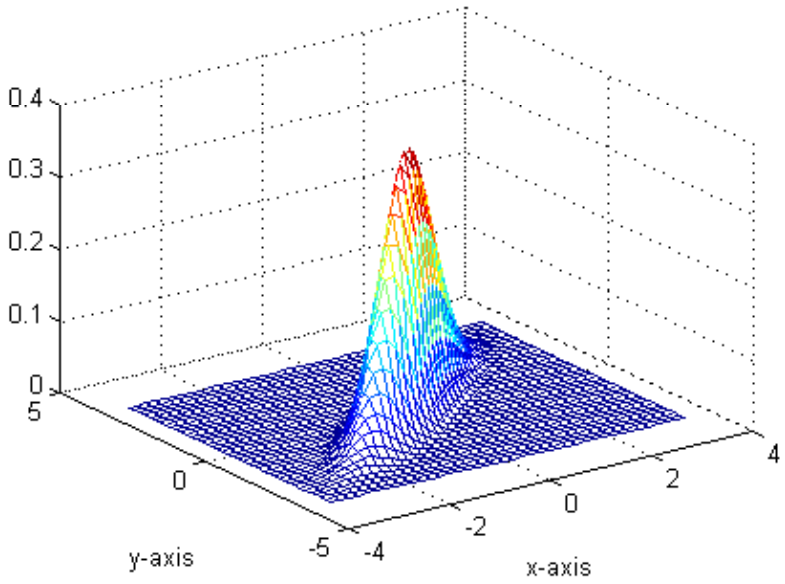
2. Charakterizace dvourozměrné náhodné veličiny

V této kapitole jsou shrnuty základní pojmy a charakteristiky týkající se dvourozměrné náhodné veličiny. Podrobnosti a odvození lze nalézt v knize [1]. Uvažujme dvě náhodné veličiny Y a X s hustotami pravděpodobnosti f(y) a f(x) a sdruženou hustotou pravděpodobnosti f(y,x). Sdružená hustota pravděpodobnosti pro dvě nezávislé náhodné veličiny se standardizovaným normálním rozdělením je na obr 1.



Obr 1. Dvourozměrné normální rozdělení pro nezávislé náhodné veličiny (nulový korelační koeficient)

Na obr. 2 je sdružená hustota pravděpodobnosti pro dvě silně korelované náhodné veličiny se standardizovaným normálním rozdělením (korelační koeficient 0.9).



Obr 2. Dvourozměrné normální rozdělení pro závislé náhodné veličiny (korelační koeficient 0.9)

Pro charakterizaci jednotlivých náhodných veličin se standardně používají momentové charakteristiky jako jsou střední hodnoty E(Y), E(X) a rozptyly D(Y), D(X).

Základní vlastnosti středních hodnot E(.) a rozptylů D (.) je možno nalézt v každé příručce matematické statistiky resp. v knihách o zpracování dat (např. [1]). Při charakterizaci rozdělení náhodného vektoru je nutné také definovat míry intenzity vztahu mezi jeho složkami Y a X . Základní takovou mírou je druhý smíšený centrální moment nazývaný kovariance cov(Y,X), která je jednoduše definována vztahem

$$\mathrm{cov}(Y,X) = E(Y*X) - E(X)*E(Y) \qquad (1)$$

Kovariance má tyto základní vlastnosti:

1) *cov(Y,X) = 0, tj. nekorelovanost znamená, že*
- **a)** E(Y*X) = 0 a zároveň E(Y) = E(X) = 0,
- **b)** E(Y*X) = E(Y) . E(X)

Případ ad a) platí pro situaci, kdy jsou Y a X centrované náhodné veličiny a zároveň ortogonální, (tj. skalární součin n-tice realizací těchto proměnných je nulový.).
Případ ad b) platí pro situaci, kdy jsou náhodné veličiny Y a X , nezávislé (pak je také sdružená distribuční funkce F(Y , X) = F(Y) * F(X) atd.)
Pozor z výše uvedeného plyne, že nekorelovanost neznamená obecně nezávislost! Pouze pro případ vícerozměrného normálního rozdělení skutečně nekorelovanost odpovídá nezávislosti.

2) *Čím je stupeň intenzity vztahu mezi* Y a X *a vyšší, tím je vyšší i kovariance*. Limitou je když jsou Y a X lineárně závislé. Pak $Y = \alpha X+\beta$, $D(Y) = \alpha^2 * D(X)$ resp. $\alpha = \sqrt{D(Y)} / \sqrt{D(X)}$ a

$$\mathrm{Cov}(Y,X) = E(\alpha X^2+\beta X) - E(X)*E(\alpha X+\beta) = \alpha D(X) = D(Y)/\alpha = \sqrt{D(X)}* \sqrt{D(Y)} \qquad (2)$$

Pojem intenzita vztahu se týká míry linearity mezi Y a X . Pro nelineární vztahy může vyjít kovariance nulová. To je patrné z příkladu 1.

3) *Kovariance je symetrickou funkcí svých argumentů*, tj.

$$\mathrm{cov}(Y,X) = \mathrm{cov}(X,Y)$$

4) *Kovariance dvou náhodných proměnných Y a X se nemění posunem počát*ku. Tedy pro dvě deterministické konstanty a,b je

$$\mathrm{cov}(Y+a, X+b) = \mathrm{cov}(Y,X)$$

5) *Změna měřítka náhodných proměnných se projeví změnou kovariance o stejné faktory*. Tedy pro dvě deterministické konstanty a, b je

$$\mathrm{cov}(Y*a, X*b) = a*b \,.\, \mathrm{cov}(Y, X)$$

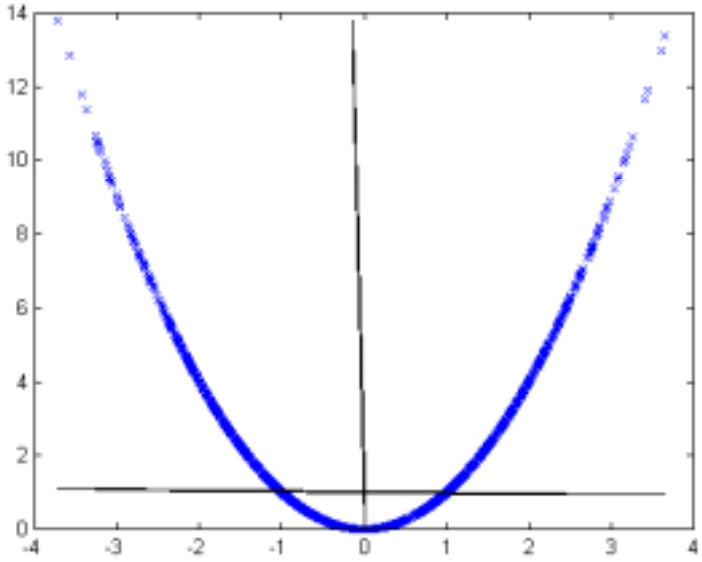
Z prvního tvrzení plyne, že existují ***různé situace***, kdy jsou dvě náhodné veličiny nekorelované. V praxi se běžně zaměňuje pojem nezávislost a nekorelovanost a často i ortogonalita.Objasněme si tyto důležité pojmy na případě, kdy je k dispozici dvourozměrný náhodný výběr (y_i , x_i) pro i = 1…N z rozdělení náhodných veličin Y a X tedy máme dva vektory $\underline{x}$, $\underline{y}$. Obecně jsou tyto vektory lineárně nezávislé, pokud nelze nalézt nenulovou konstantu C takovou, že $C* \underline{x} = \underline{y}$. Pro ortogonální vektory platí, že jejich skalární součin je

nulový, tedy $\underline{x}^T \underline{y} = 0$, a pro nekorelované náhodné veličiny platí, že $\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = 0$, kde $\bar{x}$ a $\bar{y}$ jsou aritmetické průměry.

Platí, že lineárně nezávislé vektory nemusí být ani ortogonální ani nekorelované. Na druhé straně jsou nekorelované náhodné veličiny vždy lineárně nezávislé. Pokud jsou Y a X náhodné veličiny mohou však být přesně nelineárně závislé. Pokud jsou náhodné veličiny centrované (tj. mají nulové střední hodnoty), je ortogonalita shodná s nekorelovaností.

Příklad 1

Určeme kovarianci mezi centrovanými náhodnými veličinami Y a Xc = X – E(X) , mezi kterými je přesná kvadratická závislost $Y = Xc^2$. Protože Xc je centrovaná náhodná veličina s nulovou střední hodnotou platí pro kovarianci C(Y,Xc) vztah $cov(Y,Xc) = E(Xc^2*Xc) = E(Xc^3)$. Pro případ, že má náhodná veličina Xc symetrické unimodální rozdělení jsou její liché centrální momenty nulové. Tedy i cov(Y,Xc) = 0 i když mezi nimi platí kvadratická závislost. Na obr. 3 je znázorněno 10 000 bodů kde x-ové souřadnice jsou generovány z normálního standardizovaného rozdělení a y-nové souřadnice jsou x^2. Jsou tam znázorněny také regresní přímky E(Y/x) = horizontála a E(X/y) = vertikála. Nulová směrnice odpovídá nulovému korelačnímu koeficientu.



Obr 3. Případ nulového korelačního koeficientu pro funkční závislost

Nevýhodou kovariance pro vyjádření intenzity lineárního vztahu mezi je fakt, že její hodnoty závisejí na rozptylech D(Y) a D(X), a tedy i na měřítku, ve kterém jsou vyjádřeny Y a X. Proto je vhodné zavést standardizovanou kovarianci ρ(Y,X) = ρ, která se nazývá korelační koeficient. Ten je definován poměrem

$$\rho = \frac{\mathrm{cov}(Y,X)}{\sqrt{D(Y)*D(X)}} \qquad (3)$$

Korelační koeficient leží v rozmezí (-1,1). Pro ρ = 0 jde o nekorelované náhodné veličiny. Pokud je ρ > 0, jde o pozitivně korelované, a pokud je ρ < 0, jde o negativně korelované náhodné veličiny. Pro pozitivně korelované náhodné veličiny platí, že zvětšování (zmenšování)

jedné náhodné veličiny vede v průměru ke zvětšování (zmenšování) druhé náhodné veličiny. Pro negativně korelované náhodné veličiny je tomu naopak.

Korelační koeficient má tyto základní vlastnosti:

1) *rovnost $\rho = 1$ ukazuje, že mezi Y a X je přesně lineární závislost*. Pokud je ρ blízké jedné, je mezi odpovídajícími náhodnými veličinami intenzivní lineární asociace.

2) *pokud jsou Y a X nezávislé, je vždy $\rho = 0$.* Také pro nelineárně funkčně závislé veličiny však může být ρ = 0 (viz příklad 1)

3) *pokud mají Y a X dvourozměrné normální rozdělení, znamená $\rho = 0$, že jsou nezávislé*

4) *korelační koeficient $\rho(X,X)$ a $\rho(Y,Y)$ je roven jedné.*

5) *hodnota korelačního koeficientu je invariantní vůči lineární transformaci náhodných proměnných* Y a X . Tedy pro deterministické konstanty a, b, c, d platí, že
ρ (a*Y+c, b*X+d) = ρ(Y,X)*sign(a*b) , kde sign(a*b) je znaménková funkce (sign (x) = 1 pro x > 0, sign (x) = 0 pro x =0 a sign (x) = -1 pro x < 0)

6) *Pokud jde o lineární závislost $Y = a*X + b$, je vždy $\rho = sign(a)$*

7) *Korelační koeficient je symetrickou funkcí argumentů.*

Vlastnost 5) se často využívá při analýze dat, protože umožňuje práci v různých soustavách jednotek. Také ukazuje na jednu interpretaci korelačního koeficientu. Pro standardizované náhodné veličiny Ys = (Y-E(Y))/ √D(Y) a Xs = (X-E(X))/ √D(X), s nulovými středními hodnotami a jednotkovými rozptyly platí, že ρ (Ys, Xs) = ρ(Y,X) = E(Ys, Xs). *Z toho plyne, že korelační koeficient je roven střední hodnotě součinu standardizovaných náhodných veličin.*
Korelační koeficient velmi úzce souvisí s pojmem regrese. Statisticky je regrese Y na X podmíněná střední hodnota E(Y/x) a lze ji tedy odvodit ze znalosti podmíněné hustoty pravděpodobnosti f(y/x) resp. sdružené hustoty pravděpodobnosti f(y,x). Pro případ že f(y,x) a f(y) jsou hustoty pravděpodobnosti normálního rozdělení má teoretická regrese tvar

$$E(Y/x) = E(X) + \rho\sqrt{\frac{D(Y)}{D(X)}}(x - E(X)) \qquad (4)$$

Rov. (4) je zřejmě rovnicí přímky E(Y/x) = a*X + b se směrnicí a = ρ*√D(Y) /√D(X) a úsekem , b = E(Y) - a*E(X), která prochází těžištěm o souřadnicích E(Y), E(X). Je tedy zřejmé, že teoretické regresní funkce pro případ, že Y, X mají dvourozměrné normální rozdělení, je skutečně lineární. Pro odpovídající podmíněný rozptyl D(Y /x) platí, že

$$D(Y/x) = D(Y)*(1-\rho^2) \qquad (5)$$

Podmíněný rozptyl tedy nezávisí na podmínce a teoretická regresní funkce pro případ, že Y, X mají dvourozměrné normální rozdělení, je homoskedastická. Dá se snadno dokázat, že

$$D(Y/x) = E[(Y - a*x - b)^2] \tag{6}$$

Z tohoto vztahu a rov. (5) lze odvodit, že korelační koeficient je také roven

$$\rho = \sqrt{1 - \frac{E(D(Y/x))}{D(Y)}} = \sqrt{1 - \frac{D(Y/x)}{D(Y)}} \tag{7}$$

Druhá rovnost platí jen pro případ homoskedasticity. Vztah (7) - první rovnost lze použít i pro obecnější vyjádření závislosti mezi náhodnými veličinami, kdy D(Y/x) se nepočítá z rov. (5). Pak se ρ označuje jako korelační poměr η. Pomocí rov. (7) lze snadno určit korelační koeficienty (poměry) i pro různé empirické regrese.
Analogicky platí pro regresi E(X/y) v případě dvourozměrného normálního rozdělení lineární vztah

$$E(X/y) = E(Y) + \rho\sqrt{\frac{D(X)}{D(Y)}}(y - E(Y)) \tag{8}$$

procházející opět těžištěm. Směrnice této regrese c = ρ*√D(X) /√D(Y). Odpovídající podmíněný rozptyl D(X/y) má tvar

$$D(X/y) = D(X)*(1 - \rho^2) \tag{9}$$

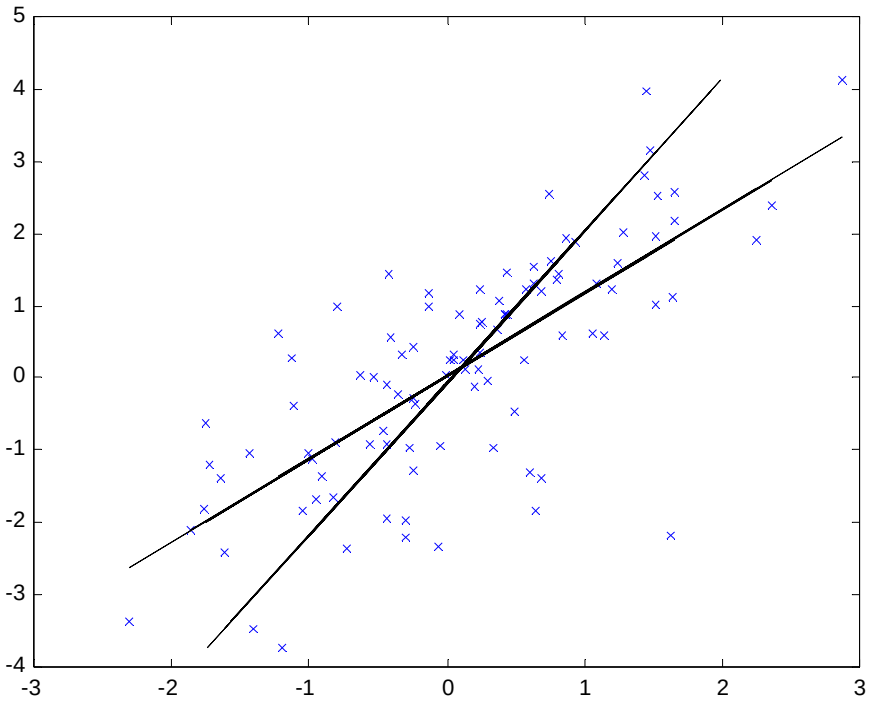
Je patrné, že se směrnice obou teoretických regresí liší. Lze snadno zjistit, že c = ρ^2/a. *Čtverec korelačního koeficientu je pak roven součinu obou směrnic.*

$$\rho^2 = a*c \tag{10}$$

Rov. (10) ukazuje, že čím bude vyšší korelační koeficient, tím více se budou obě regrese blížit a pro ρ = 1 splynou. Pro ρ = 0 bude a =c = 0, tj. obě regrese budou rovnoběžné s odpovídajícími osami souřadného systému a budou mít rovnice E(Y/x) = E(X) resp. E(X/y) =E(Y). Úhel mezi těmito regresemi bude roven γ = 90°.
Pokud je korelační koeficient nenulový a kladný, platí pro úhel mezi oběma regresemi vztah

$$tg\gamma = \frac{(1 - \rho^2)\sqrt{D(X)*D(Y)}}{\rho*(D(X) + D(Y))} \tag{11}$$

Z uvedeného je patrné, že pro případ nenulového korelačního koeficientu existují dvě různé regrese.
Na obr. 4 je znázorněno 100 bodů kde x-ové souřadnice jsou generovány z normálního rozdělení se střední hodnotou 0 a jednotkovým rozptylem tj. x = N(0,1) a y-nové souřadnice jsou generovány jako y = 1*x+ N(0,1). Jsou tam vyneseny také regresní přímky y = 0.997*x + 0.1 a x = 0.4743*y - 0.054. Korelační koeficient je roven 0.69. Teoretické hodnoty jsou D(X) = 1, D(Y) = 2, cov(X,Y) =1, ρ =1/√2=0.707 , E(Y/x) = 1*x a E(X/y) = 0.5*y. Je patrný rozdíl mezi regresemi E(Y/x) a E(X/y) a vliv menšího počtu dat na výsledky regrese.



Obr 4. Lineární regrese pro normálně rozdělená data ($\rho = 1/\sqrt{2}$).

V některých úlohách je problémem volba typu regrese, protože neexistuje zdůvodnění pro výběr E(Y/x) ani E(X/y). Pak je možno volit pro vyjádření vztahu mezi a mají hlavní osu elipsy konstantní hustoty [1].

Při praktické aplikaci uvedených vztahů se teoretické střední hodnoty, teoretické rozptyly a populační korelační koeficient nahrazují odhady počítanými na základě výběru velikosti N. Je tedy patrné, že ze základních charakteristik náhodných veličin lze určit všechny parametry korelačního modelu.

Výrazného zjednodušení vztahů pro lineární regresi je možno docílit zavedením standardizovaných proměnných Ys = (Y-E(Y))/ √D(Y) a Xs = (X-E(X))/ √D(X).
Snadno lze zjistit, že platí

$$E(Ys/xs) = \rho * Xs \text{ resp. } E(Xs/ys) = \rho * Ys \qquad (12)$$

Rov. (12) jsou odvozeny s využitím faktů, že standardizované proměnné mají nulový průměr a jednotkový rozptyl a korelační koeficient je nezávislý na lineární transformaci argumentů. Z rov. (12) plyne geometrický význam korelačního koeficientu jako směrnice úhlu, který svírá regresní přímka ve standardizovaných souřadnicích s osou x (pokud jde o E(Y/x)), resp. s osou y (pokud jde o E(X /y). Regresní koeficient a , resp. c (směrnice regresní přímky) je ve standardizovaných souřadnicích
roven párovému korelačnímu koeficientu. To usnadňuje jeho interpretaci. Z uvedeného plyne, že i při regresi v původních proměnných je korelační koeficient úzce spjat se směrnicí regresní přímky.
Směrnice regresní přímky je tedy vážený korelační koeficient, kde „váhy“ jsou úměrné podílu směrodatných odchylek obou náhodných proměnných.

Poznámka:
To je důvod, proč se i při empirické regresi jeví korelační koeficient závislý na velikosti směrnice regresní přímky. Přesněji jde o závislost směrnice regresní přímky na velikosti korelačního koeficientu. V případě výběrového korelačního koeficientu je situace ještě komplikována pokud existuje omezené rozmezí dat resp. další omezení neodpovídající náhodnému výběru..

Proto, aby byla teoretická regrese lineární nepostačuje pouze normalita, marginálních rozdělení.
Také podmíněné, resp. sdružené rozdělení musí být normální. To je třeba mít na paměti při ověřování typu teoretické regrese.

3. Výběrový korelační koeficient

Jak již bylo uvedeno je párový (Galton-Pearsonův) korelační koeficient vlastně standardizovaná kovariace. Na základě N-tice experimentálních dat (x_i , y_i) o nichž se v rámci dosavadních úvah předpokládá, že jde o dvourozměrný náhodný výběr je možné stanovit výběrový korelační koeficient r , při náhradě populačních charakteristik jejich odhady. Veličina r je náhodná proměnná a pro daný rozsah výběru záleží její rozdělení na hodnotě teoretického korelačního koeficientu ρ .

Omezme se nejdříve na případ, kdy je simultánní rozdělení veličin (Y,X), dvourozměrné normální a $\rho = 0$. Pak je hustota pravděpodobnosti f(r) symetrická kolem nuly. Dá se ukázat, že transformovaná náhodná veličina

$$t = \frac{r * \sqrt{N-2}}{\sqrt{1-r^2}} \qquad (13)$$

má Studentovo rozdělení s (n-2) stupni volnosti.

Toho lze využít pro test nezávislosti (pro dvourozměrné normální rozdělení je nekorelovanost totožná s nezávislostí), tedy testování hypotézy $H_o : \rho = 0$ proti různým alternativám H. Vyjde-li hodnota t z rov. (13) větší než kvantil Studentova rozdělení, zamítá se H_o na zvolené hladině významnosti. Uvedený test nezávislosti je silně nerobustní. Platí pouze v případě simultánní normality Y, X. Pro urychlení konvergence vhodné funkce výběrového korelačního koeficientu k normálnímu rozdělení je možno použít např. známé Fisherovy transformace z = arctgh(r). Pro N>50 má náhodná veličina z přibližně normální rozdělení se střední hodnotou E(z) = arctgh(ρ) a rozptylem D(z) = 1/(N-3).

Jednoduchá je Harleyova transformace $H(r) = (n - 2)^{0.5}$ arcsin (r) a Rubenova transformace $R(r) = (n - 2{,}5)^{0.5} r / (1 - 0{,}5r^2)^{0.5}$. Náhodné veličiny H(r) i R(r) již mají i pro menší výběry normované normální rozdělení.

Častější je případ, kdy je simultánní rozdělení veličin Y, X dvourozměrné normální a teoretický korelační koeficient $\rho \neq 0$. Pak je pro malé rozsahy výběru hustota pravděpodobnosti f(r/ρ) výběrového korelačního koeficientu silně nesymetrická. Pro velké rozsahy výběru (od cca N=500) je možno hustotu pravděpodobnosti f(r/ρ) aproximovat normálním rozdělením se střední hodnotou E(r) = ρ a rozptylem $D(r) = (1-\rho^2)^2 / (N-1)$. Normalita platí i pro širší třídu eliptických rozdělení se špičatostí g_2, kdy se pouze rozptyl násobí faktorem $1+ g_2$.
Pro případ, že teoretický korelační koeficient je ρ, platí výběrový korelační koeficient r určený z N-tice dat tyto relace:

Střední hodnota

$$E(r) \approx \rho + (1-\rho^2) * [-\frac{\rho}{2(N-1)} + \frac{\rho - 9\rho^3}{8(N-1)^2} - .... \quad (14)$$

Rozptyl

$$D(r) \approx \frac{1}{N}(1-\rho^2)^2 [1 + \frac{11\rho^2}{2(N-1)} + ...] \quad (15)$$

Šikmost

$$g_1(r) \approx -\frac{6\rho}{\sqrt{N-1}}[....]$$

Špičatost

$$g_2(r) \approx -\frac{6(12\rho^2 - 1)}{N-1}[.....]$$

Pro malé výběry je výběrový korelační koeficient r vychýlený. Vychýlení b = E(r)- ρ je tedy v prvním přiblížení úměrné

$$b \approx E(r) - \rho \approx -\frac{\rho(1-\rho^2)}{2(N-1)} \quad (16)$$

To znamená, že pro ρ>0 je výběrový korelační koeficient r *podhodnocený*.
Např. pro ρ = 0.8 a N = 10 vyjde E(r) = 0.78. Lze použít také korigovaný korelační koeficient, který se však od N = 15 liší od nekorigovaného o méně než 5%. Často se místo r používá jeho čtverec r^2 označovaný jako koeficient determinace. Pro případ, že ρ = 0 je střední hodnota

$$E(r^2) \approx \frac{1}{N-1} \quad (17)$$

Koeficient determinace vychází tedy pro ρ = 0 a malá N nadhodnocený. Např. pro ρ = 0 a N = 5 vyjde $E(r^2)$ = 0.25. Pro zlepšení statistických vlastností r se používají různé transformace. Velmi jednoduchá je např. **Nairova** transformace

$$u = \frac{r - \rho}{1 - r * \rho} \quad (18)$$

Veličina u má přibližně normální rozdělení se střední hodnotou E(u) = 0 a rozptylem $D(u) = (N-1)^{-1}$.

4. Interpretace korelačního koeficientu

Pro ilustraci chování výběrového korelačního koeficientu v různých situacích, kdy např. nejsou splněny podmínky náhodného výběru, je výhodné uvažovat, že data (x_i, y_i) i =1..N, představují dva N rozměrné vektory **x** a **y.** Pro odstranění závislosti na posunu počátku lze provést centrování tj. odečtení příslušných aritmetických průměrů. Pak $\mathbf{x}_c = \mathbf{x} - \mathbf{I} * \bar{x}$ a $\mathbf{y}_c = \mathbf{y} - \mathbf{I} * \bar{y}$. Zde **I** je jednotkový vektor. Snadno lze zjistit, že korelační koeficient je roven

$$r = \frac{\langle \mathbf{x}_c, \mathbf{y}_c \rangle}{d(\mathbf{x}_c) * d(\mathbf{y}_c)} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{d(\mathbf{x}) * d(\mathbf{y})} = \cos(\alpha) \tag{18}$$

Zde $\langle \mathbf{x}, \mathbf{y} \rangle$ je skalární součin a $d(\mathbf{x}) = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ je délka vektoru. Korelační koeficient je tedy kosínus úhlu mezi vektorem **y** a **x.** Již z tohoto vyjádření je patrné, že úhel mezi **x** a **y** ukazuje na míru lineární závislosti . V lineární algebře se pro vyjádření míry linearity mezi dvěma vektory používá determinant Gramovy matice Ta má pro případ sejných délek $\|\mathbf{x}\| = \|\mathbf{y}\|$ tvar [4]

$$G = \begin{pmatrix} d(\mathbf{x})^2 & \langle \mathbf{x}, \mathbf{y} \rangle \\ \langle \mathbf{x}, \mathbf{y} \rangle & d(\mathbf{y})^2 \end{pmatrix} = \|\mathbf{x}\|^2 * \begin{pmatrix} 1 & \cos\alpha \\ \cos\alpha & 1 \end{pmatrix}$$

Vlastní čísla jsou $\lambda_{1,2} = 1 \pm \cos\alpha$ a determinant Gramovy matice je $\det(G) = \lambda_1 * \lambda_2$.
Z vyjádření korelačního koeficientu ve tvaru (18) pak vyjde

$$r = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}$$

Vlastní čísla tedy souvisí se změnami r s ohledem na závislost obou vektorů. Pokud $\lambda_2 = 0$ oba vektory jsou kolineární a pro $\lambda_1 = \lambda_2$ jsou vektory ortogonální což může být i případ kdy jsou body (x_i ,y_i) maximálně rozptýlené v tom smyslu, že všechny možné přímky mají stejný reziduální součet čtverců.

Často se korelační koeficient využívá v souvislosti s lineárním regresním modelem

$$Y = b + a * X + \varepsilon \tag{19}$$

kde $E(\varepsilon) = 0$ a X je nezávislé na chybách ε. Po určení odhadů $\tilde{a}$, $\tilde{b}$ parametrů a b metodou nejmenších čtverců lze definovat vektor predikce $\tilde{\mathbf{y}} = \tilde{b} + \tilde{a} * \mathbf{x}$ a vektor reziduí $\mathbf{e} = \mathbf{y} - \tilde{\mathbf{y}}$.
Lze snadno ověřit, že čtverec délky centrovaného vektoru $\mathbf{y}_c$ je roven součtu čtverců délky centrovaného vektoru predikce $\tilde{\mathbf{y}}_c = \tilde{\mathbf{y}} - \mathbf{I} * \bar{y}$ a vektoru reziduí. Tyto veličiny jsou totožné s rozkladem celkového součtu čtverců $CSC = \sum_i (y_i - \bar{y})^2$ na

teoretický součet čtverců objasněný regresním modelem $TSC = \sum_i (\tilde{y}_i - \bar{y})^2$ a

reziduální součet čtverců $RSC = \sum_i (\tilde{y}_i - y_i)^2$.

Tedy CSC=TSC+RSC.
Vektory $\mathbf{y}_c$ a $\tilde{\mathbf{y}}_c$ svírají úhel jehož kosínus je korelační koeficient. Pak

$$r = \cos\alpha = \sqrt{\frac{TSC}{CSC}} = \sqrt{1 - \frac{RSC}{CSC}} \tag{21}$$

Korelační koeficient je tedy relativní míra variability objasněné regresním modelem. Toto vyjádření má smysl, pokud se v regresi uvažuje absolutní člen. Pro případ přímky procházející počátkem je třeba stále počítat s hodnotou $\bar{y}$ jako nenulovou. Pokud by se dosadilo $\bar{y} = 0$ vyšel by korelační koeficient nesprávně vyšší než pro regresi s absolutním členem. (To je jedna z častých chyb řady statistických programů).
Po přechodu od výběru k populaci lze psát

$$\rho = \sqrt{\frac{D(a*x+b)}{D(a*x+b+\varepsilon)}} = \sqrt{\frac{D(x)*a^2}{D(x)*a^2+D(\varepsilon)}} \qquad (22)$$

Z výběrových hodnot (x_i ,y_i) i =1..N lze snadno určit odhady potřebných veličin, které umožňují výpočet r.
Výběrový korelační koeficient má řadu analogických vlastností jako populační korelační koeficient a některé zcela speciální. V práci [5] je shromážděno celkem třináct způsobů vyjádření r a v práci [6] je doplněn způsob čtrnáctý. Jsou to:

1. Korelace jako podíl momentů (smíšeného a centrálních).
2. Korelace jako standardizovaná kovariance.
3. Korelace jako směrnice regresní přímky ve standardizovaných proměnných.
4. Korelace jako geometrický průměr směrnic regresí E(Y/x) a E(X/y).
5. Korelace jako poměr teoretického a celkového součtu čtverců.
6. Korelace jako průměr standardizovaných dat
7. Korelace jako funkce úhlu mezi regresemi E(Y/x) a E(X/y).
8. Korelace jako funkce úhlu mezi vektory **y** a **x**.
9. Korelace jako funkce rozptylu diference mezi standardizovanými proměnnými
10. Korelace jako funkce „balonu“ aproximujícího elipsu konstantní hustoty.
11. Korelace jako funkce elipsy konstantní hustoty.
12. Korelace jako funkce tetovací statistiky u plánovaných experimentů.
13. Korelace jako poměr dvou průměrů.
14. korelace jako podíl shody proměnných (objektů).

Většina vyjádření korelace již byla diskutována nebo jde o různé kuriozity. Za zmínku stojí vyjádřeni č. 9, 11, 13 a 14.

Vyjádření č. 9
Pro standardizované náhodné veličiny Ys = (Y-E(Y))/√D(Y) a Xs = (X-E(X))/√D(X) se může definovat rozptyl jejich diference

$$D(Ys-Xs) = D(Ys)+D(Xs)+2*D(Xs)*D(Ys)*\rho = 1+1+2*\rho$$

Tedy $\rho = 1- D(Ys-Xs)/2$

Vyjádření č. 11
Toto je jedno z prvních vyjádření vycházející z elips konstantních hustot tj. řezů sdružené hustoty f(x,y) = c, pro zvolená c. Pokud se pracuje s centrovanými proměnnými a délka hlavní poloosy elipsy konstantních hustot je D resp. vedlejší poloosy je d platí, že

$$\rho = \frac{D^2 - d^2}{D^2 + d^2} \tag{23}$$

Vyjádření č. 13

Toto je také z prvních vyjádření použitých Galtonem a využívané v sledování dědičnosti. Obdobné argumenty byly použity při zavedení pojmu regrese Gossetem. Nechť je definována libovolná hodnota x_p. Pak je možno za předpokladu dvourozměrné normality odvodit, že

$$\rho = \frac{E(Y / X > x_p)}{E(X / X > x_p)} \tag{24}$$

Prahová hodnota x_p se používá zejména u různých psychologických testů (typu: mají chytří rodiče chytré děti?).

Vyjádření č. 14

V řadě praktických úloh se korelace používá k posouzení přístrojů, metod, zpracování atd. V těchto případech je zajímavé jak souvisí korelace se podílem shodných výsledků (odlišných pouze co do lineární transformace). Je vhodné použít opět standardizované náhodné veličiny Ys = (Y-E(Y))/ √D(Y) a Xs = (X-E(X))/ √D(X). Nechť lze data rozdělit na skupinu shodných (velikosti k) pro které jsou $y_{si}=x_{si}$ a skupinu nesouhlasných (velikosti N-k), kde mezi y_{si} a x_{si} neexistuje žádná vazba. Korelační koeficient je pak

$$r = \left(\frac{1}{N-1}\right) * \left[\sum_{i=1}^{k} x^2{}_{si} + \sum_{i=k+1}^{N} y_{si} * x_{si}\right] \tag{25}$$

a jeho střední hodnota je

$$E(r) = \frac{k}{N-1}$$

Tedy korelační koeficient odpovídá podílu shodných výsledků. Protože úplná shoda je nereálná je vhodnější uvažovat případ kdy se za shodu považuje pokud y_{si} leží v jistém rozmezí (-A.A) veličiny Xs. Odvození pro tento případ je uvedeno v práci [6]. Pokud se uvažuje rovnoměrné rozdělení k shodných prvků v intervalu (-A, A) má korelační koeficient tvar

$$E(r) \approx \frac{k/N}{\sqrt{1 + \frac{k}{N}\frac{A^2}{3}}} \tag{26}$$

Z rov (26) je patrné že k růstu korelačního koeficientu vede:

- rozšíření intervalu shody tedy růst A (při možném menším poklesu k).
- zvětšení k při konstantním A.

je tedy patrné, že vyšší korelační koeficient ještě nemusí znamenat vyšší podíl shody.

Pokud se při sběru dat provedou omezení vzrůstá nebezpečí nesprávné interpretace. Mantel [7] se zabýval problémem, kdy se vybírají body (objekty) podle jistých omezení. Např. při

testování kvality prostředků pro zlepšení vlastností materiálů se vybírají jen ty, které na základě předběžných informací dobře fungují. Vychází se opět ze standardizovaných proměnných Ys a Xs s korelační m koeficientem ρ. Nechť jsou omezení na Xs taková, že E(Xs) =A, E(Xs ^{2}) =B a tedy D(Xs) = C = B-A^2. Korelační koeficient pro tato omezení ρ_R má pak tvar

$$\rho_R = \rho * \sqrt{\frac{C}{1-\rho^2 * (1-C)}} \qquad (27)$$

Koeficient C se dá interpretovat jako poměr rozptylu Xs omezeného ku neomezenému. Obyčejně se provádí omezení tak, že C <0 . Pak je ρ_R menší než ρ. Pokud je C >0 (vypouštějí se prostřední hodnoty a zůstávají extrémy) je ρ_R vyšší než ρ. .Způsoby výpočtu C jsou uvedeny v práci [7]. Připomeňme pouze, že je třeba určit první a druhý obecný moment pro useknuté normální rozdělení.

Nechť je např. omezení, že X leží nad 0.7 směrodatné odchylky (75.8% ním percentilem). Pak je omezení (x_d = 0.7 a x_h = ∞) a

$$E(X) = \int_{0.7}^{\infty} xf(x)dx / \int_{0.7}^{\infty} f(x)dx \qquad E(X^2) = \int_{0.7}^{\infty} x^2 f(x)dx / \int_{0.7}^{\infty} f(x)dx$$

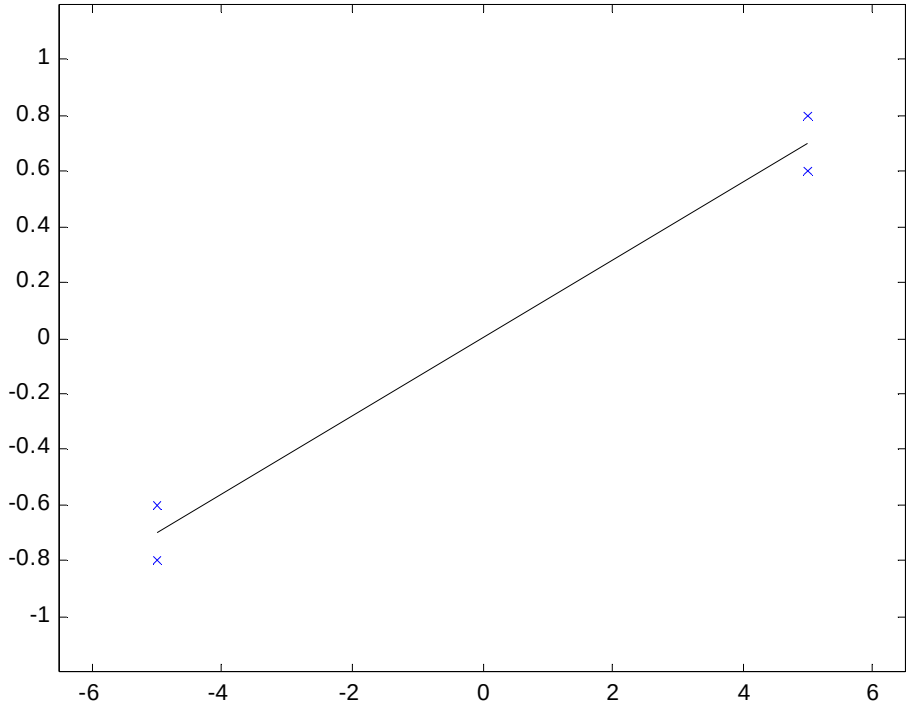
Pro normální rozdělení je pak E(X) = 1.2905 E(X^2)=1.9033a C = 0.2379. Pro případ, že ρ=0.707 vyjde z rov. (27), že ρ_R = 0.438,tedy silné podhodnocení.

Naopak, pokud se provede omezení, že X leží v mezích –1.0 až 1.5 směrodatné odchylky je (x_d = -1.5 a x_h = 1.0). Pak E(X) = -0.498 E(X^2) = 2.935 a C = 2.681. Pro případ, že ρ=0.707 vyjde z rov. (27), že ρ_R = 0.854 ,tedy silné nadhodnocení.

Je tedy zřejmé, že omezení na data může vést k neočekávaným výsledkům. Např. nechť se v prvním experimentu s prostředky pro zlepšení vlastností materiálu (Y=cena, X=kvalita) použije všech prostředků a nechť vyjde korelace ρ=0.707. Pokud se v druhém kole, vyberou jen prostředky vyšší kvality ležící nad 0.7 směrodatné odchylky vyjde korelace pouze ρ_R = 0.438. Analogicky lze zdánlivě "vylepšovat" korelaci stratifikovaným výběrem.

Pro objasnění některých dalších vlastností korelačního koeficientu je výhodné zavést data s předem definovanými vlastnostmi . Jedna taková data jsou znázorněna na obr 5. Při jejich generaci se vyšlo z předpokladu, že diskrétní náhodný vektor (Y,X) má sdruženou hustotu pravděpodobnosti odpovídající rovnoměrnému rozdělení ve čtyřech bodech (-k, -p-e) , (-k, -p+e) , (k, p+e) , (k, p-e). Uvažujme, že k>0, p >0 , e>0 , p-e>0. Pak mají jednotlivé parametry speciální interpretaci:

- **k** určuje rozmezí dat na x-ové ose,
- **p** jsou hodnoty predikce $\tilde{y}$ určené MNČ v jednotlivých bodech. (y- nové hodnoty regresní přímky),
- **e** jsou rezidua v jednotlivých bodech $e = y - \tilde{y}$.



Obr 5. Data pro charakterizaci vlastností korelačního koeficientu (k=5,p=0.7,e=0.1, korelační koeficient 0.9899).

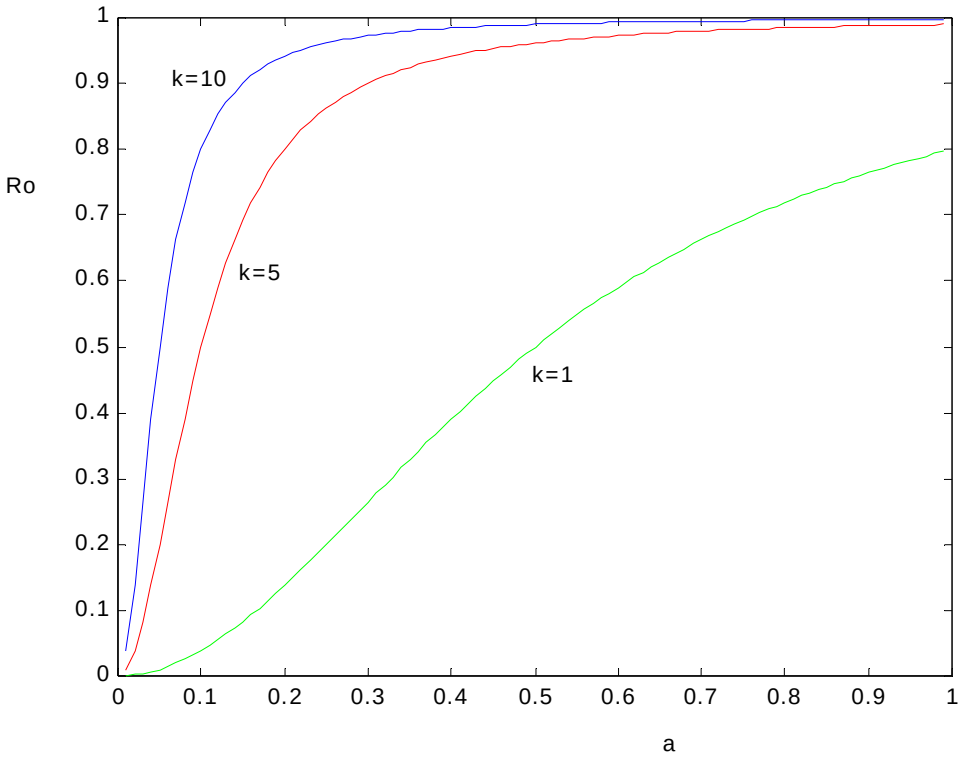
Lze jednoduše odvodit, že pro takto definované náhodné veličiny Y,X platí E(X)=E(Y)= 0, D(X)= k^2, D(Y)= p^2+e^2 , cov(X,Y))= p*k. Korelační koeficient je tedy ve tvaru

$$\rho = \frac{p}{\sqrt{p^2 + e^2}} \tag{28}$$

a nezávisí zde na velikosti parametru k. Regresní přímka prochází počátkem a má směrnici a = p /k. Z této rovnice lze snadno určit např. parametr p tak aby pro dané e nabýval korelační koeficient hodnoty ρ = c. Platí, že

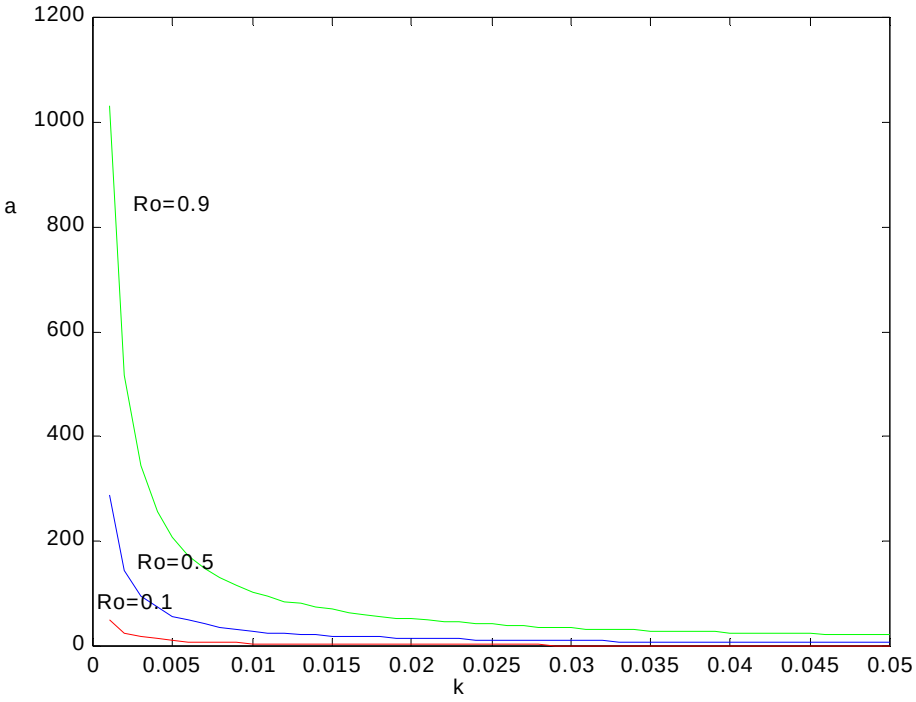
$$|p| = \frac{|c * e|}{\sqrt{1 - c^2}}$$

Je patrné, že k růstu korelačního koeficientu povede zvětšování velikosti parametru p. Pro c →1 roste p nade všechny meze ale rezidua stále zůstávají velikosti e, tedy perfektní korelace rovné jedné nelze docílit. Zajímavější je však zjištění, že zároveň roste směrnice regresní přímky. Tedy pro fixované rozmezí na x-ové ose roste s růstem směrnice regresní přímky také korelační koeficient. To je patrné z rov (28) při dosazení za p= a*k. Na obr. 6 je závislost korelačního koeficientu ρ na směrnici regresní přímky a pro různá k a e=0.5.



Obr. 6 Závislost korelačního koeficientu ρ na směrnici regresní přímky *a* (e=0.5).

Na druhé straně lze změnou rozmezí k docílit situace kdy se mění směrnice regresní přímky při stejném korelačním koeficientu. To je patrné z obr 7, kde je závislost směrnice regresní přímky na rozmezí k na ose x a pro různá ρ a e = 0.5.



Obr 7. Závislost směrnice regresní přímky a na rozmezí k na ose x (e = 0.5).

Je patrná hyperbolické závislost, která je tím povlovnější, čím je korelační koeficient nižší. Velikost korelačního koeficientu tedy může nebo nemusí růst s růstem směrnice regresní přímky. Záleží silně na rozmezí dat a jejich vzájemné poměru. V praktických úlohách je tedy třeba mít na paměti, že každé omezení může způsobit potíže při interpretaci korelačního koeficientu.

4. Korelační křivka

Korelační křivka je přihozeným zobecněním korelačního koeficientu na případ nelineární resp. heteroskedastické regrese. Vychází se z rov (22), která se zobecňuje pro nelineární a heroskedastické modely $y = f(x,\mathbf{a}) + \sigma(x) * N(0,1)$ nahrazením směrnice a lokální směrnicí $a(x) = df(x,\mathbf{a})/dx$ a použitím podmíněného rozptylu $\sigma^2(x)$ místo reziduálního rozptylu. Rov (22) pak přechází na tvar

$$\rho(x) = \sqrt{\frac{D(x) * a^2(x)}{D(x) * a^2(x) + \sigma^2(x)}}$$

Takto lokálně definovaný korelační koeficient je přirozenou mírou nelineárních závislostí. Také tento korelační koeficient je invariantní vůči lineární transformaci. Poměrně jednoduše lze určit i neparametrickou versi korelačního koeficientu $\rho(x)$ při náhradě derivace diferencí. Další podrobnosti o korelační křivce lze nalézt v práci [2].

5. Závěr

Z uvedeného je patrné, že existuje řada různých situací, které mohou způsobit, že korelační koeficient poskytuje kuriózní výsledky. Řada dalších problémů se vyskytuje u případu zanedbání významných proměnných nebo korelací více proměnných. Některé typové problémy jsou řešeny v práci [1]. ***Obecně však platí, že vysoký korelační koeficient ještě neznamená, že mezi proměnnými je silná lineární vazba natož pak příčinná souvislost***.

Poděkování

Tato práce vznikla s podporou grantu GAČR č. 106/99/1184, grantu MŠMT č. VS 97084 a výzkumného záměru MŠMT J11/98:244101113

6. Literatura

[1] Meloun M., Militký J., : Zpracování experimentálních dat, East Publishing, Praha 1998
[2] Blyth S.: Int. Stat. Review, **62**, 393 (1994).
[3] Bjrerve S., Doksum K.A.: Annals of Stsatistcs , **21**, 890 (1993)
[4] Kas S.: Amer. Math. Monthly **96**, 910, (1989)
[5] Rovine M. J., von Eye A. : Amer. Statistician **51**, 42, (1997)
[6] Rodgers J. L. Nicewander W.A. : Amer. Statistician **52**, 59, (1998)
[7] Mantel N.: Biometrics **22**, 182 , (1966)

Název souboru:	korelace
Adresář:	E:\Pom
Šablona:	D:\Program Files\Microsoft Office\Sablony\Normal.dot
Název:	Korelace
Předmět:	
Autor:	katedra textilních materiálů
Klíčová slova:	
Komentáře:	
Datum vytvoření:	14.09.00 13:37
Číslo revize:	2
Poslední uložení:	14.09.00 13:37
Uložil:	Milan Meloun
Celková doba úprav:	0 min.
Poslední tisk:	14.09.00 13:44
Jako poslední úplný tisk	
Počet stránek:	16
Počet slov:	4 620 (přibližně)
Počet znaků:	26 337 (přibližně)