# Co potřebuje klinik vědět o (bio)statistice?

## Korelační a regresní analýza

**Pravoslav Stránský**

Ústav lékařské biofyziky a Oddělení výpočetní techniky Univerzita Karlova v Praze, Lékařská fakulta v Hradci Králové, Česká republika / Department of Medical Biophysics and Division of Computer Science, Charles University in Prague, School of Medicine at Hradec Králové, Czech Republic

*Stránský P. Co potřebuje klinik vědět o (bio)statistice? Korelační a regresní analýza. Folia Gastroenterol Hepatol 2005; 3 (3): 110 – 113.*

***Souhrn.*** *Korelace a regrese jsou statistické metody umožňující kvantifikovat vztah mezi dvěma nebo více veličinami. Korelace umožňuje určit sílu vztahu mezi proměnnými, regrese určení typu závislosti.*

***Klíčová slova:*** *korelace, regrese, asociace mezi nezávisle a závisle proměnnými veličinami, kovariance, Pearsonův korelační koeficient, regresní koeficienty, metoda nejmenších čtverců, koeficient determinace*

*Stránský P. What should a clinician know about (bio)statistics? Correlation and regression analyses. Folia Gastroenterol Hepatol 2005; 3 (3): 110 – 113.*

***Abstract.*** *Correlation and regression are statistical methods that enable to quantify relationship between two or more variables. Correlation makes possible to determine the power of association among independent and dependent variables, regression enables to set a type of dependence. .*

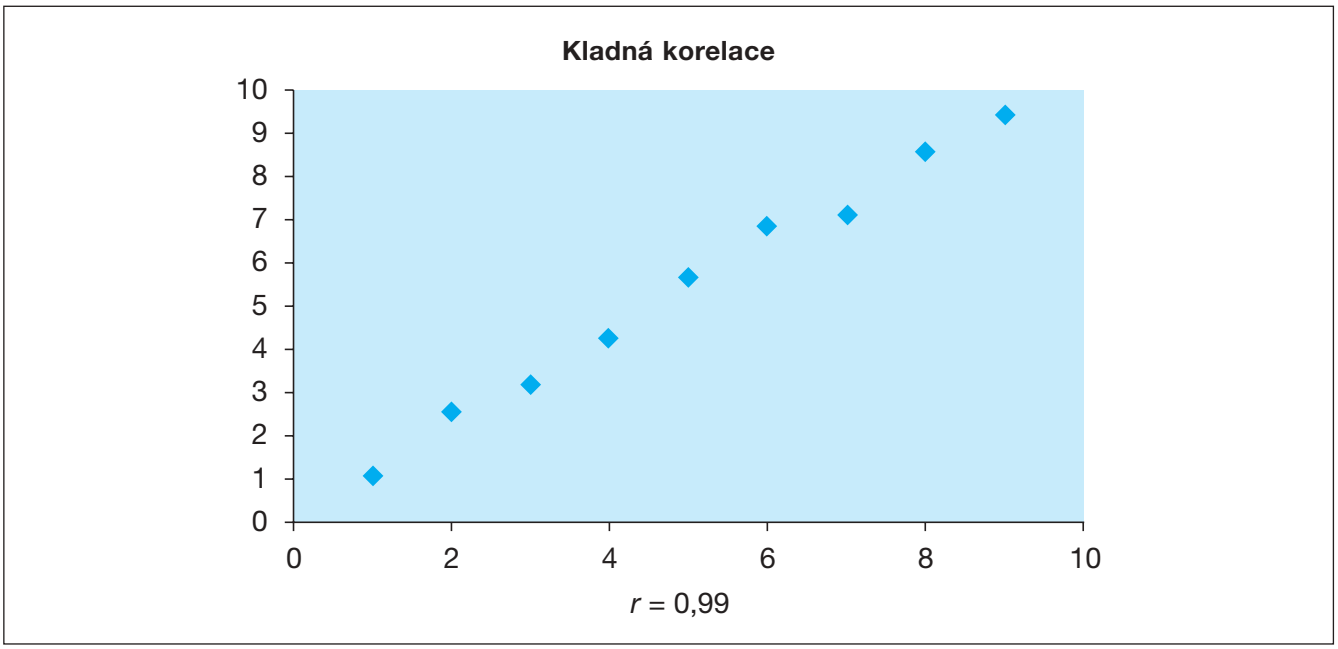
***Key words:*** *correlation, regression, association among independent and dependent variables, covarince, Pearsons' coefficient of correlation, regression coefficients, least square method, coefficient of determination*

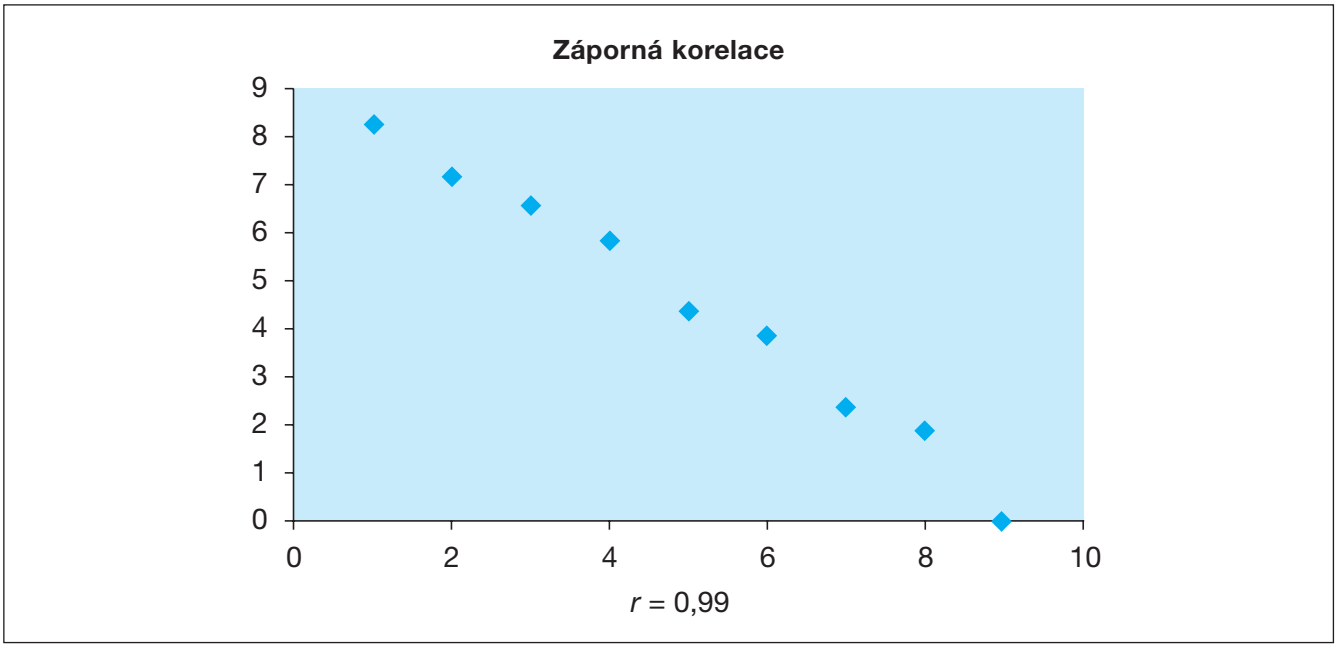
Korelace a regrese jsou statistické metody, které umožňují kvantifikovat vztah mezi dvěma nebo více veličinami. Za splnění určitých předpokladů (v prvním přiblížení to je Gaussovo rozložení studovaných veličin) korelace umožňuje určit **sílu** vztahu mezi proměnnými. Regrese umožňuje určení typu závislosti, tedy zjistit, jaká funkce je nejvhodnější k vyjádření zkoumaného vztahu. Známe-li danou funkci, můžeme z hodnoty jedné proměnné určit s jistou pravděpodobností hodnotu druhé. Relativně jednoduchá interpretace výsledků korelační a regresní analýzy přispěla k tomu, že mluvíme o korelaci mezi veličinami i tehdy, když ty nesplňují požadavek Gaussova rozložení (nemohou ho splnit, protože se například jedná o veličiny ordinální). Jako příklad takovéto asociace uveďme vztah mezi úspěšností chirurgické léčby dané nemoci a typem zdravotnického zařízení.

V medicíně se nejčastěji setkáváme s touto situací: hodnota jedné veličiny, které se říká **závisle proměnná** (vysvětlovaná) a v pravoúhlých souřadnicích se vynáší na osu *y*, je určena hodnotou veličiny druhé, **nezávisle proměnné** (vysvětlující), vynášené na osu *x*.

Příkladem mohou být vztahy mezi množstvím vypitého alkoholu (nezávisle proměnná) a hladinou alkoholu v krvi (závisle proměnná) nebo pravděpodobnost vzniku infarktu myokardu (závisle proměnná) při znalosti hodnot systolického krevního tlaku, BMI, cholesterolemie, množství vykouřených cigaret (nezávisle proměnné).

Určení, která z proměnných je závislá a která nezávislá, je z matematického hlediska nepodstatné, nerozlišitelné. Z logického hlediska je to často věcí konvence. Například u dospívajících jedinců můžeme při znalosti jejich věku (považovaného za nezávisle proměnnou veličinu) a pohlaví odhadnout jejich pravděpodobnou výšku či hmotnost (závisle proměnné) nebo naopak z výšky a/nebo hmotnosti odhadnout jejich věk. Podobně z hodnoty atmosférického tlaku můžeme určit nadmořskou výšku, ve které se nacha-




zíme, a obráceně znalost nadmořské výšky umožňuje vypočítat hodnotu atmosférického tlaku a následně hodnotu parciálního tlaku kyslíku v dané nadmořské výšce.

Při splnění předpokladu gaussovského rozložení zkoumaných veličin a předpokladu, že závislost mezi veličinami je lineární, je míra vazby mezi veličinami vyjádřena tzv. **kovariancí**. Tato statistická charakteristika je formální obdobou rozptylu (1). Je to součet součinů odchylek nezávisle a závisle proměnné veličiny dělený počtem pozorování, zmenšeným o jedničku. Vzhledem k tomu, že ve většině případů jsou nezávisle a závisle proměnné různé veličiny, tj. mají jiný (fyzikální) rozměr, má však také kovariance jiný rozměr než rozptyl. Aby bylo možné srovnávat výsledky získané při různých pozorováních, provádí se standardizace kovariance tak, že se vydělí druhou odmocninou součinu rozptylů nezávisle a závisle proměnné. Výsledná hodnota pak leží v intervalu od -1 do +1, je bezrozměrná (má rozměr rovný 1) a nazývá se **Pearsonův korelační koeficient**, *r.* Adjektivum Pearsonův zdůrazňuje skutečnost, že se jedná o statistickou charakteristiku vypočtenou za předpokladu normality dat, tedy pomocí tzv. parametrické metody. V neparametrické statistice, které se bude věnovat další článek z této série, je jedním ze způsobů určení vztahu výpočet Spearmanova koeficientu pořadové korelace.

Pokud je absolutní hodnota *r* rovna jedné, je vztah mezi veličinami deterministický, funkční. Jsou-li veličiny *x* a *y* nezávislé, pak se *r* = 0. Problémem je, že neplatí opačná implikace. Obecně nelze říci, že jestliže se *r* = 0, jsou x a y nezávislé. Nulová hodnota korelačního koeficientu znamená nezávislost jen v případě, že vztah mezi veličinami je lineární a veličiny mají Gaussovo rozložení. Na základě zkušenosti lze sílu závislosti mezi veličinami a absolutní hodnotou korelačního koeficientu *r* vyjádřit takto:

$0 < r < 0{,}1$ - veličiny jsou nezávislé
$0{,}2 < r < 0{,}3$ - závislost mezi veličinami je slabá
$0{,}4 < r < 0{,}6$ - závislost je střední
$0{,}7 < r < 0{,}8$ - závislost je silná
$r > 0{,}9$ - závislost je velmi silná

Korelace může být kladná nebo záporná. V prvním případě se vzrůstem nezávisle proměnné roste i hodnota závisle proměnné (obr. 1). Záporná korelace znamená, že zvětšování hodnoty nezávisle proměnné má za následek pokles závisle proměnné (obr. 2). Je třeba mít na mysli, že vztah je lineární. Často se záporná korelace zaměňuje s nepřímou úměrou, při které je vztah mezi veličinami stejný, ale grafickým vyjádřením nepřímé úměry je hyperbola.

Nulová hypotéza $H_0$ v korelační analýze předpokládá, že *r* = 0 (veličiny jsou nezávislé). Můžeme-li ji zamítnout, přijmeme hypotézu alternativní $H_A$, která říká, že veličiny jsou závislé. Zvolená hodnota pravděpodobnosti *p* udává pravděpodobnost chyby I. druhu (2), tedy že zamítneme nulovou hypotézu, i když ve skutečnosti platí. Skutečnost, že můžeme nulovou hypotézu zamítnout, neznamená, že závislost mezi veličinami je příčinná. Pravděpodobnost odpovídající hodnotě testového kritéria v případě korelační analýzy silně závisí na počtu pozorování. Pro *r* = 0,3 (tedy existence slabé závislosti mezi veličinami) zamítneme $H_0$ na 5% hladině významnosti již tehdy, když se počet pozorování rovná 45.

Závěrem zdůrazněme, že o korelaci mezi veličinami můžeme mluvit pouze tehdy, když tyto mají gaussovské rozložení a vztah je lineární.

Podaří-li se nám prokázat, že mezi veličinami existuje významná korelace, stává se zajímavou otázka, jaký je přesný tvar vztahu popisujícího tuto závislost. Protože korelace přepokládá, že vztah je lineární, znamená to, že body, které odpovídají naměřeným hodnotám, můžeme v pravoúhlých souřadnicích proložit přímku, jejíž rovnice je

$y = bx + a$

Regresní analýza umožňuje vypočíst odhady hodnot koeficientů *b* a *a*, které se nazývají parametry regrese, regresní koeficienty. Koeficient *b* je směrnice regresní přímky a určuje její sklon, úhel který svírá s osou *x* (přesněji je tangentou tohoto úhlu). Koeficient a je tzv. absolutní člen a je to hodnota závisle proměnné veličiny odpovídající nulové hodnotě veličiny nezávisle proměnné (úsek vytnutý regresní přímkou na ose *y*). Odhady regresních koeficientů se označují řeckými písmeny $\alpha$ a $\beta$.

Známe-li hodnoty odhadů $\alpha$ a $\beta$, můžeme pro každou hodnotu nezávisle proměnné vypočíst odpovídající hodnotu závisle proměnné. Vypočtená hodnota se obvykle liší od hodnoty naměřené a rozdíl těchto hodnot se nazývá **residuum**. To se používá při výpočtu odhadů koeficientů regresní rovnice. Nejčastěji používanou metodou je ta, která požaduje, aby součet druhých mocnin (čtverců) residuí byl minimální. Podle tohoto požadavku se metoda označuje jako **metoda nejmenších čtverců** (least square method).

Regrese je úzce spojena s korelací a druhá mocnina *r* (*r* square) se nazývá **koeficient determinace** a vysvětluje, jaká část variability závisle proměnné je způsobena variabilitou nezávisle proměnné veličiny. Po vynásobení 100 udává procento rozptylu závisle proměnné vysvětlené regresí. Zbytek do sta procent je způsoben náhodnými chybami, kterými je zatížena závisle proměnná veličina.

Na rozdíl od korelační analýzy se požadavek na gaussovské rozdělení týká pouze veličiny závisle proměnné. U veličiny nezávisle proměnné se naopak předpokládá, že to není veličina náhodná, ale že může být stanovena experimentátorem. V medicíně takovýto požadavek může být splněn spíše výjimečně (dávka léku, tepová frekvence při kardiostimulaci) a ve většině případů se tento předpoklad diskrétně přechází. Druhým úskalím regresní analýzy je snaha určit hodnoty závisle proměnné veličiny i pro takové hodnoty nezávisle proměnné, které neleží v intervalu hodnot, ze kterých se výpočet odhadů regresních koeficientů prováděl. To je absolutně nepřípustné a může to vést ke zcela mylným závěrům.

V regresní analýze lze formulovat dvě nulové hypotézy. První předpokládá, že *b* = 0 a znamená, že pro jakoukoli hodnotu *x* je *y* rovno *a* (regresní přímka je rovnoběžná s osou *x*). Jinými slovy tato $H_0$ je ekvivalentní $H_0$ v korelační analýze (*r* = 0) a znamená, že mezi veličinami není žádný vztah. Můžeme-li nulovou hypotézu zamítnout, přijmeme hypotézu alternativní,

která tvrdí, že vztah mezi veličinami existuje. Sílu vztahu v tomto případě můžeme odhadnout z hodnoty pravděpodobnosti *p* odpovídající vypočtené hodnotě testového kritéria (čím je *p* menší, tím je síla vztahu větší).

Druhá nulová hypotéza předpokládá, že se hodnota absolutního členu rovná nule ($a = 0$), tedy že regresní přímka prochází počátkem. Tento předpoklad je někdy teoreticky zdůvodnitelný, příkladem ve fotometrii kalibrační přímka vyjadřující závislost absorbance na koncentraci látky v roztoku (při nulové koncentraci by měla být i nulová absorbance). V takovýchto případech zamítnutí nulové hypotézy upozorňuje na skutečnost, že nejsou splněny všechny předpoklady platné pro danou metodu (světlo procházející roztokem je absorbováno rozpouštědlem). V uvedeném příkladu jde o tzv. lineární kalibraci, při které vlastně určujeme hodnotu nezávisle proměnné z naměřené hodnoty závisle proměnné veličiny.

Velmi často se stává, že není splněn předpoklad linearity. Ukázalo se, že je to možné napravit transformací jedné či obou veličin v regresi. Jestliže je hodnota závisle proměnné veličiny exponenciální funkcí veličiny nezávisle proměnné (intenzita prošlého rentgenového záření vyjádřená jako funkce absorpčního koeficientu objemové jednotky tkáně ve výpočetní tomografii), vztah zlinearizujeme zlogaritmováním hodnoty závisle proměnné.

LITERATURA

1. Stránský P. Co potřebuje klinik vědět o (bio)statistice. Jak popsat data. Folia Gastroenterol Hepatol 2005; 3: 42 – 46.
2. Stránský P. Co potřebuje klinik vědět o (bio)statistice. Testování hypotéz. Folia Gastroenterol Hepatol 2005; 3: 74 – 76.

**Adresa pro korespondenci / correspondence to:**
Prof. MUDr. Pravoslav Stránský, Ústav lékařské biofyziky a Oddělení výpočetní techniky, Univerzita Karlova v Praze, Lékařská fakulta v Hradci Králové, Šimkova 870, P.O. Box 38, 500 38 Hradec Králové, Česká republika / Czech Republic.
E-mail: str@lfhk.cuni.cz