

519.2:57+61

STATISTICKÉ TESTY REGRESNÍCH ZÁVISLOSTÍ: VÝHLEDY TESTOVÁNÍ ČASOVÝCH ŘAD PROTEINOVÝCH MAP

Jiří KNÍŽEK, Jiří STULÍK, Jan ÖSTERREICHER, Aleš MACELA

Ústav radiobiologie a imunologie Vojenské lékařské akademie J. E. Purkyně, Hradec Králové

Souhrn

Práce uvádí možnosti testování vzájemných vztahů regresních závislostí. Těmito metodami je možné testovat, zda dvě přibližně paralelní závislosti jsou shodné, nebo jestli jsou neshodné v kterémkoli bodě, v libovolném intervalu nebo v celém rozsahu měření. Pro hodnocení regresních závislostí platí, že testování rozdílnosti (neshodnosti) má několikanásobně vyšší informační validitu než testování shodnosti.

Klíčová slova: Testy regresních závislostí; Dvouzávislostní t-testy; Lineární regrese; Testy časových řad; Proteinové mapy.

Statistical Tests of Regression Dependences: Prospects of Time Series of Protein Map Testing

Summary

The study states the possibilities of testing of regression dependence mutual relations. It is possible to test by these methods whether two approximately parallel dependences are equal or unequal in any point, in any interval or the whole extent of measurement. It holds true for the evaluation of regression dependences that the testing of inequality has a few times higher information validity than the testing of equality.

Key words: Tests of regression dependences; Two-dependence t-tests; Linear regression; Tests of time series; Protein maps.

Úvod

Pro některá hodnocení biologických vztahů a zákonitostí (např. závislost srdeční frekvence na dávce sympatomimetik (adrenalin, noradrenalin atd.), závislost účinnosti léčiva v populaci na koncentraci léčiva apod.) je nezbytné používat statistického prokládání křivek empirickými hodnotami metodou regresní analýzy. Pro tyto účely je třeba křivky vzájemně testovat. Předložená metoda v budoucnosti nalezne pravděpodobně velmi účinné uplatnění pro hodnocení a testování časových řad proteinových map z dvourozměrné gelové elektroforézy.

Regresní úloha

Obecný regresní model lze vyjádřit ve tvaru

$$y_i = \eta_i + e_i = \eta(x_{1,i}, x_{2,i}, \dots, x_{k,i}; \alpha_0, \alpha_1, \dots, \alpha_p) + e_i; \\ i = 1, 2, \dots, n;$$

kde y je náhodnou veličinou, zatímco $x_1, x_2, \dots, x_k$ mohou být buď pevnými (nenáhodnými) proměnnými,

nebo také náhodnými veličinami.

Symbol e pak představuje náhodnou chybu a indexy $i = 1, 2, \dots, n$ pořadová čísla měření (pozorování) těchto veličin. Regresní koeficienty, tj. hledané parametry, jsou označeny symboly $\alpha_0, \alpha_1, \dots, \alpha_p$. Obecný lineární regresní model (1, 2) lze vyjádřit ve tvaru

$$y_i = \eta_i + e_i = \alpha_0 + \alpha_1 x_{1,i} + \dots + \alpha_p x_{p,i} + e_i; \\ i = 1, 2, \dots, n;$$

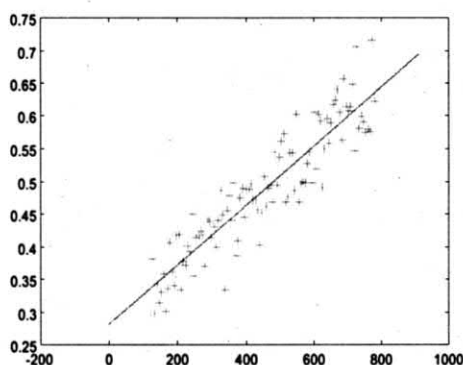
a model regresní přímky (graf 1) ve tvaru

$$y_i = \eta_i + e_i = \alpha_0 + \alpha_1 x_i + e_i; i = 1, 2, \dots, n.$$

Prakticky: regresní přímka se vyznačuje tím, že hodnoty regresních koeficientů α_0 a α_1 mají takovou optimální hodnotu, že součet druhých mocnin náhodných chyb

$$\sum_{i=1}^n e_i^2 = \min$$

je minimální možný. Z tohoto důvodu tato metoda optimalizace byla nazvána a je všeobecně známa jako *metoda nejmenších čtverců*.



Graf 1: Regresní přímka

Lineární parametrická funkce v regresní analýze

Regresní analýza zavádí tzv. *lineární parametrickou funkci* (1)

$$\gamma = c_0\alpha_0 + c_1\alpha_1 + \dots + c_p\alpha_p; \quad \sum_{j=0}^p c_j^2 > 0;$$

kde $c_0, c_1, \dots, c_p$ jsou známá reálná čísla. Sumace značí to, že alespoň jedno z těchto čísel je nenulové.

Příklady *lineárních parametrických funkcí* jsou:

1. regresní parametr α_j (v tomto případě $c_j = 1$,
 $c_0 = c_1 = \dots = c_{j-1} = c_{j+1} = \dots = c_p = 0$,
 $j = 0, 1, \dots, p$),
2. rozdíl $\alpha_1 - \alpha_2$ (v tomto případě, $c_1 = 1$, $c_2 = -1$,
 $c_0 = c_3 = c_4 = \dots = c_p = 0$),
3. hodnota regresní funkce $\eta = \alpha_0 + \sum_{i=1}^p x_i \alpha_i$
(v tomto případě $c_0 = 1$; $c_i = x_i$; $i = 0, 1, \dots, p$).

Tyto *lineární parametrické funkce* pak lze libovolně testovat:

$$H_0: \gamma = \gamma_0$$

$$H_A: \gamma < \gamma_0, \quad H_A: \gamma \neq \gamma_0, \quad H_A: \gamma > \gamma_0,$$

kde γ_0 je zvolená konstanta.

Sdružená lineární parametrická funkce pro dva paralelní regresní modely

Pro dva paralelní lineární regresní modely (označené horními indexy I a II)

$$y_i^I = \eta_i^I + e_i^I = \alpha_0^I + \alpha_1^I x_{1,i}^I + \dots + \alpha_{p'}^I x_{p',i}^I + e_i^I;$$

$$i = 1, 2, \dots, n^I;$$

$$y_i^{II} = \eta_i^{II} + e_i^{II} = \alpha_0^{II} + \alpha_1^{II} x_{1,i}^{II} + \dots + \alpha_{p''}^{II} x_{p'',i}^{II} + e_i^{II};$$

$$i = 1, 2, \dots, n^{II};$$

kde platí zcela analogický význam příslušných symbolů jako v části **Regresní úloha**, lze analogicky formulovat (1) *sdruženou lineární parametrickou funkci*

$$\gamma = \sum_{j=0}^{p'} c_j^I \alpha_j^I + \sum_{j=0}^{p''} c_j^{II} \alpha_j^{II}; \quad \sum_{j=0}^{p'} (c_j^I)^2 > 0; \quad \sum_{j=0}^{p''} (c_j^{II})^2 > 0;$$

která analogicky umožňuje testovat vzájemné vztahy mezi těmito lineárními modely.

Obdoba dvouvýběrového t-testu pro regresní závislosti

Zavedením (např. časových) závislostí

$$x_{1,i}^I(t_i^I), x_{2,i}^I(t_i^I), \dots, x_{p',i}^I(t_i^I); \quad i = 1, 2, \dots, n^I;$$

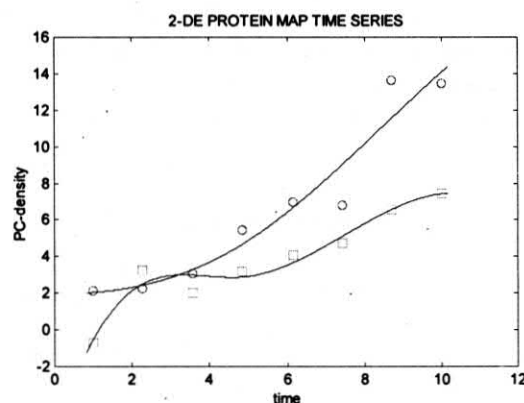
$$x_{1,i}^{II}(t_i^{II}), x_{2,i}^{II}(t_i^{II}), \dots, x_{p'',i}^{II}(t_i^{II}); \quad i = 1, 2, \dots, n^{II};$$

např. ortogonálních polynomů pro neekvidistantní nezávisle proměnnou (3), lze zcela korektně statisticky aproximovat jakékoli křivkové závislosti (4):

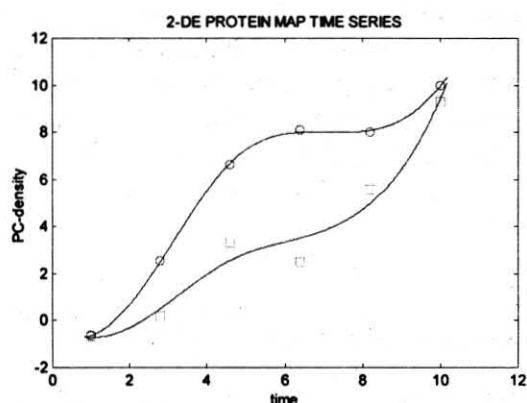
$$y_i^I(t_i^I); \quad i = 1, 2, \dots, n^I;$$

$$y_i^{II}(t_i^{II}); \quad i = 1, 2, \dots, n^{II}.$$

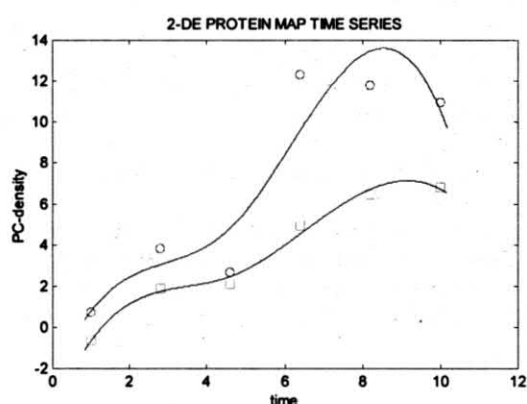
Pomocí výše naznačených matematických prostředků je pak možné, jako obdoba dvouvýběrového t-testu (5), provádět testování vzájemných vztahů regresních závislostí v obecném intervalu $\langle t_1; t_2 \rangle$. V případě, že $t_1 = t_2 = t$, se jedná o dvouzávislostní test v bodě t . Je možné např. testovat, zda dvě přibližně paralelní závislosti jsou shodné, nebo zda jsou neshodné v kterémkoli bodě, v libovolném intervalu nebo v celém rozsahu měření (grafy 2-5).



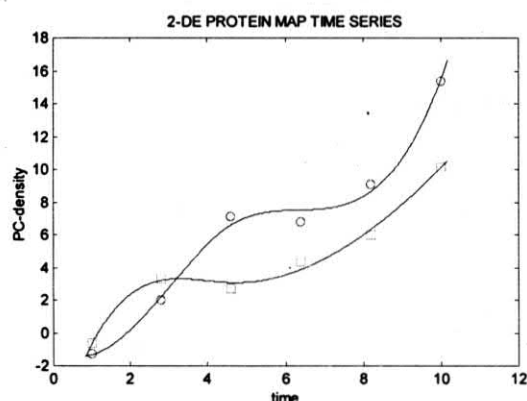
Graf 2: Dvě regresní závislosti



Graf 3: Dvě regresní závislosti



Graf 4: Dvě regresní závislosti



Graf 5: Dvě regresní závislosti

I zde platí, že testování rozdílnosti (neshodnosti) má několikanásobně vyšší informační validitu než testování shodnosti (5). I tady pro experimentální biologii vyplývá pravidlo, že pokusy by měly být koncipovány takovým způsobem, aby klíčové statistické hodnocení celé experimentální práce vedlo spíše na testování rozdílnosti (neshodnosti) než na testování shodnosti.

Počítačový modul pro informační analýzu časových řad odezev autonomních proteinových systémů (MIAPS)

Je navrhován jako pravděpodobně velmi účinná pracovní pomůcka pro rozhodování mezi různými hypotézami o biologicko-dynamických vlastnostech proteinových systémů zkoumaných pomocí dvou-rozměrné gelové elektroforézy.

Jednou z metod studia zákonitostí autonomních biologických systémů je metoda jejich identifikace, tj. určování biologicko-dynamických vlastností, způsobem informační analýzy časových odezev na přesně definovaný vstupní biologický signál.

Například u autonomního biologického systému: *tkáň, buněčná kultura, bakteriální suspenze* apod. mohou být vstupním biologickým signálem: kancerogeny, ionizující nebo UV záření, xenobiotika, léky, stresové signály, regulátory buněčného cyklu apod. Jako odezva jsou relativní změny (množství toho kterého proteinu v jednotkách počítačové hustoty) v proteinových mapách (dvourozměrových gelových elektroforetogramech) proměřované v různých časových intervalech po uskutečnění biologického signálu. Tyto relativní změny budou pomocí MIAPS podrobovány hloubkové informační analýze. MIAPS bude pracovat na principu statistického porovnávání lineárních (nebo variantně nelineárních) parametrických funkcí představujících (především) buď rychlost, nebo zrychlení přibývání, resp. ubývání proteinů. Takovýto způsob analýzy časových odezev autonomních proteinových systémů nebyl dosud použit.

MIAPS by se mohl stát velmi účinnou pracovní pomůckou při studiu a analýze autonomních proteinových systémů. Existují představy o tom, že proteinové mapy by byly používány také pro rutinní klinická laboratorní vyšetření. MIAPS by umožnil značné zvýšení informační účinnosti těchto postupů.

Konkrétněji: proteinová mapa obsahuje asi 2500 proteinů (jsou popsány případy, kdy proteinů je až 10 000). Při čísle 2500 je myslitelných

$$\binom{2500}{2} = 3123750 \text{ vzájemných vztahů (dvojic)}$$

časových průběhů PC-denzit proteinů, které je třeba systematicky otestovat.

Toto hodnocení bude možné pod dvěma celkovými přístupy:

1. Přístup, kdy jsou porovnávány časové průběhy PC-denzit proteinů v původních naměřených jednotkách.
2. Přístup, kdy jsou porovnávány časové průběhy PC-denzit proteinů v normalizovaných jednotkách.

Normalizace: maximální hodnota ≈ 1 (v rámci robustnosti pro daný průběh). Tím bude zajištěno, že do porovnatelného vztahu budou dány i proteiny, u nichž změna s časem je proporcio-nální, avšak řádově odlišná.

V další fázi tohoto projektu by mohlo být přistoupeno i k simultánnímu vzájemnému testování 3 a více časových průběhů PC-denzit proteinů, v rámci technických možností (výkonnosti PC). Základní teoretické vztahy potřebné pro odvození příslušných testačních algoritmů již byly formulovány (6, 7). Počty kombinací u vícečlených skupin závislosti dosahují velmi vysokých hodnot:

$$\begin{aligned} \binom{2500}{3} &\approx 2.6 \cdot 10^9, \\ \binom{2500}{4} &\approx 1.6 \cdot 10^{12}, \\ \binom{2500}{5} &\approx 8.1 \cdot 10^{14}. \end{aligned}$$

Do budoucnosti lze jistě předpokládat další zvýšení rychlosti počítačů. V současnosti již existují po-

čítače pracující na úrovni PC Pentium 1100 apod., což je umožněno spřažením dvou procesorů v centrální početní jednotce. Tento fakt dává naději, že v budoucnosti bude možné řešit i takové náročné úkoly, které jsou nyní nemyslitelné.

Literatura

1. HÁTLE, J. - LIKEŠ, J.: Základy počtu pravděpodobnosti a matematické statistiky. Praha, SNTL, Bratislava, ALFA, 1972. 463 s.
2. MELOUN, M. - MILITKÝ, J.: Statistické zpracování experimentálních dat. Praha, PLUS, 1994. 839 s.
3. RALSTON, A.: Základy numerické matematiky. Praha, Academia, 1978. 635 s.
4. WILKS, SS.: Mathematical Statistics. New York, John Wiley & Sons, 1962. 644 s.
5. KNÍŽEK, J. - ÖSTERREICHER, J. - MACELA, A.: Pravděpodobnost v t-testu v excelu. Voj. zdrav. Listy, 2000, vol. 69, č. 5, s. 206-213.
6. ZVÁRA, K.: Regresní analýza. Praha, Academia, 1989. 245 s.
7. HAHN, GJ. - HENDRICKSON, RW.: A table of percentage points of the distributions of the largest absolute value of k Student t variates and its application. Biometrika, 1971, vol. 58, p. 323-332.

Korespondence: Ing. Jiří Knížek, CSc.

Vojenská lékařská akademie J. E. Purkyně
PS 35/ÚRI
500 01 Hradec Králové
e-mail: knizek@pmfhk.cz

Do redakce došlo 26. 6. 2000