Klasická nebo robustní lineární regrese?

Luděk Dohnal[1], Paul Faigl[2]
[1]Referenční laboratoř pro klinickou biochemii, ÚKBLD 1.LF UK a VFN, Karlovo nám. 32, 12111 Praha 2, ludek.dohnal@lf1.cuni.cz
[2]FCDD, University of Southern Queensland, Toowoomba, Australie, faiglp@usq.edu.au

Souhrn
Stručná charakteristika a srovnání klasické lineární regrese metodou nejmenších čtverců a robustnějších postupů nazvaných "standardizovaná metoda hlavních komponent" a regrese podle Passinga-Babloka.

Klíčová slova
lineární regrese, robustní regrese, Passing, Bablok

1. Úvod

Regresní postupy se používají už dlouho, ale teprve s rozšířením nejdříve programovatelných kalkulátorů a později minipočítačů či stolních počítačů a jejich příslušného programového vybavení se začaly používat ve velkém. Bohužel nejčastěji užívaná klasická lineární regrese se aplikuje nepříliš vhodným způsobem nebo se její význam přeceňuje (1). V dalším se omezíme pouze na regresi s přímkovým modelem.

"Klasická" lineární regrese pracuje na principu minimalizace součtu čtverců odchylek naměřených hodnot od hodnot vypočtených ve směru vysvětlované neboli závisle proměnné. Jeden ze základních předpokladů použití tohoto způsobu je, aby hodnoty vysvětlující (nezávisle) proměnná měly neurčitosti řádově menší, než hodnoty vysvětlované (závisle) proměnné. Pokud se v dřívějších dobách klasická regrese používala převážně při konstrukci např. kalibračních křivek, koncentrace analytu v jednotlivých kalibračních roztocích byly získány přesným navážením a rozpuštěním čisté substance a přesným ředěním tohoto základního roztoku a uvedený přepoklad byl automaticky splněn. Jakmile se však rozmohlo její použití např. pro porovnání výsledků dvou analytických postupů (třebas testování možné zaměnitelnosti metod) a oba analytické postupy měly srovnatelně velkou neurčitostí stanovení, začala selhávat. To, že se stále s tímto nedobrým použitím setkáváme, je způsobeno m.j. nedostatečnou informovaností a také tím, že na rozdíl od jiných druhů regrese je klasická lineární regrese metodou nejmenších čtverců zmiňována ve všech příručkách, implementována ve všech statistických programech a v řadě tabulkových kalkulátorů (spreadsheets). Jiné regresní postupy jsou na stolních počítačích dostupné mnohem obtížněji. Pro případy, kdy není použití klasické regrese vhodné, lze použít robustnější postupy, tedy takové, které nejsou tolik citlivé na volbu vysvětlující a vysvětlované proměnné (požadavek mnohem menší neurčitosti u vysvětlující proměnné), na odlehlé hodnoty, na "nenormalitu" apod. Budou zde uvedeny dva z nich a to tzv. standardizovaná metoda hlavních komponent (5) a postup uveřejněný Passingem a Bablokem (4,5,8).

2. "Klasická" lineární regrese

Tato regrese je podrobně popsána např. v (2). Praktické postřehy a postup výpočtu ještě z doby, kdy výpočetní technikou v laboratoři byla logaritmická pravítka a kalkulačky,

jsou srozumitelně popsány např. v (6). Jednou z jejich typických vlastností je, že pokud vzájemně vyměníme vysvětlující a vysvětlovanou proměnnou, vyjde nám ve většině případů dost jiná regresní přímka.

3. Standardizovaná metoda hlavních komponent

V anglickém originálu "Standardized principial component analysis" (5). Názorně se dá charakterizovat třeba takto. Provedeme "klasickou" regresi, pak vzájemně vyměníme vysvětlující a vysvětlovanou proměnnou a provedeme znovu "klasickou" regresi. Získáme tak dvě regresní přímky

$$y = a_1 + b_1 x \quad y = a_2 + b_2 x$$

Výslednou regresní přímku získáme jako jakousi "střední hodnotu" obou původních přímek. Její směrnici b (slope) vypočteme jako geometrický průměr obou původních směrnic

$$b = \sqrt{\frac{b_1}{b_2}}$$

a její úsek a (intercept) vypočteme

$$a = \bar{y} - b\bar{x}$$

přičemž $\bar{y}$ **a** $\bar{x}$ **jsou aritmetické průměry všech hodnot y a x. Výsledná regresní přímka**

$$y = a + bx$$

není tedy závislá na volbě vysvětlující a vysvětlované proměnné. Podmínkou pro použití této "průměrující" regrese je, aby v rozmezí, v němž ji provádíme, byla alespoň přibližně splněna rovnost

$$\frac{se_x}{s_x} = \frac{se_y}{s_y}$$

v níž se_x **,**se_y **jsou směrodatné odchylky, které popisují neurčitost hodnot x a y. Získají se z dat o opakovatelnosti resp. reprodukovatelnosti, která nejsou součástí hodnot zpracovávaných prováděnou regresí.** s_x**,** s_y **jsou směrodatné odchylky vypočítané z dat x a y, se kterými provádíme regresi. Jsou tedy mírou rozptýlení těchto dat, chceme-li mírou šíře intervalu, v němž regresi provádíme.**

4. Passing-Bablokova regrese

Jedná se o regresi (4,5,8), jejíž výsledky jsou prý m.j. velmi odolné vůči odlehlým hodnotám (outliers). Je založena na myšlence spojit každý bod x_i**,**y_i **s každým bodem** x_j**,**y_j **přímkou a vypočítat všechny směrnice takto vzniklých přímek.**

$$b_{ij} = \frac{y_i - y_j}{x_i - x_j} \quad \text{pro} \quad 1 \le i \le j \le n$$

Jak učí kombinatorika (8), počet takto vzniklých přímek se rovná počtu kombinací 2. třídy (vždy dva body) z n prvků (n=počet měření) bez opakování (v dané kombinaci je daný bod použit nanejvýš jednou). Uvedený počet kombinací je roven $\binom{n}{2}$ **.**
Obecně platí, že

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

přičemž n! = $n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$

S rostoucím n roste poměrně rychle počet kombinací - tedy počet všech dvoubodovývh přímek. Několik příkladů počtů těchto kombinací je uvedeno v tab.1.

Tab.1 Kombinace druhé třídy z n prvků s opakováním

n	2	3	4	5	10	100	1000
počet kombinací	1	3	6	10	45	4950	499500

Směrnice výsledné (celkové) regresní přímky b se pak vypočte jako medián všech jednotlivých směrnic b_{ij}

$$b = \tilde{b}_{ij}$$

Úsek výsledné regresní přímky a se vypočítá jako medián $\tilde{a}_i$ **rozdílů**

$$a_i = y_i - bx_i$$

Procedura podle Passing-Babloka může zahrnovat konstrukci intervalů spolehlivost úseku a a směrnice b, dále testování hypotézy b=1 a hypotézy a=0.

5. Závěr

Pro porovnání jsou data dle (5) uvedená v tab.2 zpracovaná třemi diskutovanými regresními postupy. Výsledky jsou sumarisovány v tab.3 a na obr.1. Úsudek nechť si každý udělá svůj.

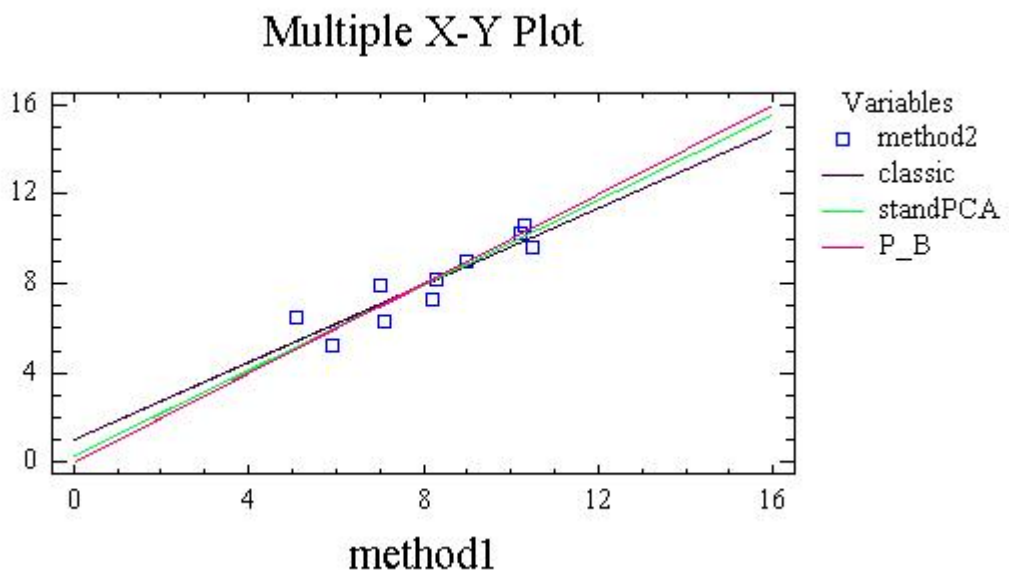
Tab.2 Data pro srovnání regresních postupů, n=10 (1)

proměnná	data									
metoda 1 (x)	7.0	8.3	10.5	9.0	5.1	8.2	10.2	10.3	7.1	5.9
metoda 2 (y)	7.9	8.2	9.6	9.0	6.5	7.3	10.2	10.6	6.3	5.2

Tab.3 Výsledky srovnávaných regresních postupů, n=10, y=a+bx (1)

regresní postup	směrnice b	úsek a
"klasická" lineární regrese	0.862	1.049
standardizovaná metoda hlavních komponent	0.947	0.352
Passing-Bablokova regrese	1.000	-0.050

Obr.1 Výsledky srovnávaných regresních postupů, n=10, y=a+bx (1)



6. Literatura

1) Dohnal,L.: Lesk a bída lineární regrese, FONS No.3, 33-34, 1997 nebo http://www.lf1.cuni.cz/~ldohna/dohnal/leskbida.htm

2) Meloun,M.,Militký,J.: Statistické zpracování experimentálních dat, 839 stran, PLUS spol. s r.o., Praha (1994)

3) Passing,H.,Bablok,W.: A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for methods comparison studies in clinical chemistry. Part I. J.Clin.Chem.Clin.Biochem. 21, 709-720 (1983)

4) Passing,H.,Bablok,W.: Comparison of several regression procedures for method comparison studies and determination of sample size. Application of linear regression procedures for methods comparison studies in clinical chemistry. Part II. J.Clin.Chem.Clin.Biochem. 22, 431-445 (1984)

5) Leonhardt,W.: Regression und gepaarter t-Test beim Methodenvergleich. Z.med.Labor-Diagn. 24, 168-172 (1983)

6) Eckschlager,K.,Horsák,I.,Kodejš,Z.: Vyhodnocování analytických výsledků a metod. 223 stran, SNTL, Praha 1980

7) Bartsch,H.-J.: Matematické vzorce, 578 str., SNTL Praha 1965

8) Bablok,W.,Passing,H.,Bender,R.,Scheider,B.: A general regression procedure for method transformation. Application of linear regression procedures for method comparison studies in clinical chemistry. Part III. J.Clin.Chem.Clin.Biochem. 26, 783-790 (1988)

9) Marquis,P.: Method validator 1.1.9.2 (freeware), http://www.marquis-soft.com/ (1999)