Výstavba regresního modelu regresním tripletem

Prof. RNDr. Milan Meloun, DrSc.,
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice

Souhrn: *Postup hledání regresního modelu je popsán obecně a dokumentován na 3 úlohách analytické laboratoře. Skládá se z těchto kroků: 1. Návrh modelu začíná vždy od nejjednoduššího modelu, lineárního. 2. Předběžná analýza dat sleduje proměnlivost proměnných na rozptylových diagramech, indexových grafech. Vyšetřuje se multikolinearita, heteroskedasticita, autokorelace a vlivné body. 3. Odhadování parametrů se provádí klasickou metodou nejmenších čtverců, následuje testování významnosti parametrů Studentovým t-testem. Střední kvadratická chyba predikce MEP a Akaikovo informační kritérium AIC jsou rozhodčí kritéria při hledání modelu. 4. Regresní diagnostika provádí identifikaci vlivných bodů a ověření předpokladů metody nejmenších čtverců. V případě více vysvětlujících proměnných se posoudí vhodnost proměnných pomoci parciálních regresních grafů a parciálních reziduálních grafů. 5. Konstrukce zpřesněného modelu: parametry zpřesněného modelu jsou odhadovány s využitím (a) metody vážených nejmenších čtverců (MVNČ) při nekonstantnosti rozptylu, (b) metody zobecněných nejmenších čtverců (MZNČ) při autokorelaci, (c) metody podmínkových nejmenších čtverců (MPNČ) při omezení kladených na parametry, (d) metody racionálních hodností u multikolinearity, (e) metody rozšířených nejmenších čtverců (MRNČ) pro případ, že všechny proměnné jsou zatížené náhodnými chybami, a konečně (f) robustních metod pro jiná rozdělení než normální a data s vybočujícími hodnotami a extrémy.*

Při výstavbě regresních modelů se běžně užívá metody nejmenších čtverců. Metoda nejmenších čtverců poskytuje postačující odhady parametrů jenom při současném splnění všech předpokladů o datech a o regresním modelu. Pokud tyto předpoklady nejsou splněny, ztrácí metoda nejmenších čtverců své vlastnosti.

Základní předpoklady metody nejmenších čtverců (MNČ): Statistické vlastnosti odhadů $\hat{\boldsymbol{y}}_P, \hat{\boldsymbol{e}}, \boldsymbol{b}$ závisí na splnění jistých předpokladů. Pokud platí předpoklady I až IV, jsou odhady **b** parametrů **β** **n**ejlepší, **n**estranné a **l**ineární (NNLO). Navíc mají asymptoticky normální rozdělení. Pokud platí ještě předpoklad VII, mají odhady **b** normální rozdělení i pro konečné výběry.

I. *Regresní parametry β mohou nabývat libovolných hodnot.* V praxi však často existují omezení parametrů, která vycházejí z jejich fyzikálního smyslu.

II. *Regresní model je lineární v parametrech a platí aditivní model měření.*

III. *Matice nenáhodných, nastavovaných hodnot vysvětlujících proměnných* $\boldsymbol{X}$ *má hodnost rovnou právě m*. To znamená, že žádné její dva sloupce $\boldsymbol{x}_j$, $\boldsymbol{x}_k$ nejsou *kolineární*, tj. rovnoběžné vektory. Tomu odpovídá i formulace, že matice $\boldsymbol{X}^T\boldsymbol{X}$ je symetrická regulární matice, ke které existuje inverzní matice a jejíž determinant je větší než nula.

IV. *Náhodné chyby* ε_i *mají nulovou střední hodnotu* $E(\varepsilon_i) = 0$. To musí u korelačních modelů platit vždy. U regresních modelů se může stát, že $E(\varepsilon_i) = K$, $i = 1, ..., n$, což znamená, že model neobsahuje absolutní člen. Po jeho zavedení bude $E(\varepsilon_i') = 0$, kde $\varepsilon_i' = y_i - \hat{y}_{P,i} - K$.

V. *Náhodné chyby* ε_i *mají konstantní a konečný rozptyl* $E(\varepsilon_i^2) = \sigma^2$. Také podmíněný rozptyl $D(y/\boldsymbol{x}) = \sigma^2$ je konstantní a jde o *homoskedastický* případ.

VI. *Náhodné chyby* ε_i *jsou vzájemně nekorelované a platí* $cov(\varepsilon_i\ \varepsilon_j) = E(\varepsilon_i\ \varepsilon_j) = 0$. Pokud mají chyby normální rozdělení, jsou nezávislé. Tento požadavek odpovídá požadavku nezávislosti měřených veličin *y*.

VII. *Chyby* ε_i *mají normální rozdělení* $N(0, \sigma^2)$. Vektor $\boldsymbol{y}$ má pak vícerozměrné normální rozdělení se střední hodnotou $\boldsymbol{X}\boldsymbol{\beta}$ a kovarianční maticí $\sigma^2\ \boldsymbol{E}$, kde $\boldsymbol{E}$ je jednotková matice.

Regresní diagnostika

Metoda nejmenších čtverců nezajišťuje obecně nalezení přijatelného modelu, a to jak ze statistického, tak i z fyzikálního hlediska. Musí být splněny podmínky, odpovídající složkám tzv. *regresního tripletu* [data, model, metoda odhadu].

Regresní diagnostika obsahuje postupy k identifikaci

a) vhodnosti dat pro navržený regresní model (složka *data*),

b) vhodnosti modelu pro daná data (složka *model*),

c) splnění základních předpokladů MNČ (složka *metoda*).

Základní rozdíl mezi regresní diagnostikou a klasickými testy spočívá v tom, že u regresní diagnostiky není třeba přesně formulovat alternativní hypotézu. Tímto pojetím se regresní diagnostika blíží spíše k *exploratorní regresní analýze*, která vychází z faktu, že "uživatel ví o analyzovaných datech přece jenom více než počítač". Počítač slouží jako nástroj analýzy dat, modelu a metody odhadu. Model je navrhován v interakci uživatele s programem. Tím by měl být omezen vznik formálních regresních modelů, které nemají fyzikální smysl a jsou v technické praxi obyčejně jen omezeně použitelné.

1. Data: mezi základní techniky diagnostiky patří stanovení rozmezí dat, jejich variability a přítomnosti vybočujících pozorování. K tomu lze využít grafů rozptýlení s kvantily a řady postupů průzkumové analýzy jednorozměrných dat. Přes svoji jednoduchost umožňuje diagnostika identifikovat ještě před vlastní regresní analýzou

a) *nevhodnost dat* (malé rozmezí nebo přítomnost vybočujících bodů),

b) *nesprávnost navrženého modelu* (skryté proměnné),

c) *multikolinearitu*,

d) *nenormalitu* v případě, kdy jsou vysvětlující proměnné náhodné veličiny.

Kvalita dat úzce souvisí s užitým regresním modelem. Při posuzování se sleduje především výskyt *vlivných bodů* (VB), které mohou být hlavním zdrojem řady problémů, jako je zkreslení odhadů a růst rozptylů až k naprosté nepoužitelnosti regresních modelů. Podle toho, kde se vlivné body vyskytují, lze provést dělení na

1. *Vybočující pozorování* (outliers), které se liší v hodnotách vysvětlované (závisle) proměnné *y* od ostatních, a

2. *Extrémy* (high leverage points), které se liší v hodnotách vysvětlujících (nezávisle) proměnných $\boldsymbol{x}$ nebo v jejich kombinaci (v případě multikolinearity) od ostatních bodů.

Vyskytují se však i body, které jsou jak vybočující, tak i extrémní. K identifikaci vlivných bodů typu vybočujícího pozorování se využívá zejména různých typů reziduí a k identifikaci extrémů pak diagonálních prvků H_{ii} projekční matice $\boldsymbol{H}$.

2. Model: kvalitu regresního modelu lze posoudit v případě jedné vysvětlující proměnné *x* přímo z rozptylového grafu závislosti *y* na *x*. V případě více vysvětlujících proměnných a multikolinearity mohou však rozptylové grafy *mylně indikovat* nelineární trend i u lineárního modelu. Z řady různých grafů k posouzení vztahu *y* a x_j se omezíme na a) parciální regresní grafy, a b) parciální reziduální grafy.

Parciální regresní grafy byly Belseyem zařazeny mezi základní nástroje počítačové interaktivní analýzy regresních modelů. Umožňují nejenom posouzení kvality navrženého regresního modelu, ale indikují i přítomnost vlivných bodů a nesplnění předpokladů klasické metody nejmenších čtverců. Parciální regresní graf pro posouzení vztahu mezi y a i-tou vysvětlující proměnnou x_i je závislost *reziduí* $\boldsymbol{v}$ regrese y na sloupcích matice $\boldsymbol{X}_{(i)}$ a reziduí u regrese x_i na sloupcích matice $\boldsymbol{X}_{(i)}$**.** Přitom matice $\boldsymbol{X}_{(i)}$ vznikne z matice $\boldsymbol{X}$ vynecháním i-tého sloupce $\boldsymbol{x}_i$, odpovídajícího i-té vysvětlující proměnné. Parciální regresní grafy mají tyto vlastnosti:

a) Směrnice přímky v parciálním regresním grafu je stejná jako odhad b_j v neděleném modelu a úsek je roven nule. Tato lineární závislost platí pouze v případě, že navržený model je správný.

b) Korelační koeficient mezi oběma proměnnými parciálního regresního grafu odpovídá parciálnímu korelačnímu koeficientu $\boldsymbol{R}_{yx_j(x)}$.

Parciální reziduální grafy se označují také jako grafy *"komponenta + reziduum"*. Parciální reziduální grafy však poskytují poněkud odlišné informace než parciální regresní grafy. Směrnice lineární závislosti je rovna b_j a úsek je nulový. Lineární závislost pak ukazuje na vhodnost navržené proměnné x_j v modelu.

Parciální reziduální grafy se doporučují především k indikaci rozličných typů nelinearity v případě nesprávně navrženého regresního modelu.

3. Metoda: V praxi bývají některé předpoklady MNČ porušeny, což vede k použití jiných kritérií. K porušení předpokladů dochází v těchto základních případech:

a) Na parametry jsou kladena omezení, což vede na užití *metody podmínkových nejmenších čtverců (MPNČ).*

b) Kovarianční matice chyb není diagonální (autokorelace), příp. data nemají stejný rozptyl (heteroskedasticita), což vede na užití *metody zobecněných nejmenších čtverců (MZNČ)*, resp. *metody vážených nejmenších čtverců (MVNČ).*

c) Rozdělení dat nelze považovat za normální nebo se v datech vyskytují vlivné body. V takovém případě se místo kritéria metody nejmenších čtverců užije *robustního* kritéria, které je na porušení předpokladu o rozdělení chyb a na vlivné body málo citlivé. Z robustních kritérií jsou nejznámější *M-odhady*. Jedná se o maximálně věrohodné odhady pro vhodnou hustotu pravděpodobnosti chyb. Pro odhad parametrů $\boldsymbol{b}$ se užívá *iterační metody vážených nejmenších čtverců* (*IVNČ*).

d) Také proměnné x mohou být zatížené náhodnými chybami, což vede na užití *metody rozšířených nejmenších čtverců (MRNČ).* Pro případ regresní přímky je použití metody rozšířených nejmenších čtverců velmi jednoduché. Postačuje znalost poměru rozptylu σ_y^2 (vysvětlovaná proměnná) a σ_x^2 (vysvětlující proměnné), $K = \sigma_y^2/\sigma_x^2$. Pro odhad směrnice regresní přímky $y = a\,x + b$ pak platí

$$a = L + \operatorname{sign}(S_{yx})\,\sqrt{K + L^2}$$

kde

$$L = \frac{S_{yx} - K\,S_x}{2S_x}$$

a sign S_{yx} je znaménková funkce. Symboly S označují součty čtverců, odpovídajících proměnných

$$S_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 \ , \quad S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 \ , \quad S_{yx} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Při znalosti odhadu směrnice $\hat{a}$ se snadno určí odhad úseku $\hat{b}$ ze vztahu

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

Pro případ stejných rozptylů, tj. $K = 1$ vede dosazení do výše uvedených vztahů k odhadům minimalizujícím kolmé vzdálenosti (*orthogonální regrese)*. Pro odhady rozptylů odhadů $\hat{a}$, $\hat{b}$ se pak používá speciálních vztahů.

e) Pro špatně podmíněné matice $\boldsymbol{X^T X}$ se používá *metoda racionálních hodností*, vedoucí k systému vychýlených odhadů, kde vychýlení je řízeno jedním parametrem.

Postup výstavby lineárního regresního modelu:

1. Návrh modelu: začíná se vždy od nejjednoduššího modelu, u kterého vystupují jednotlivé vysvětlující proměnné v prvních mocninách a nevyskytují se žádné interakční členy typu $x_j\, x_k$.

2. Předběžná analýza dat: sleduje se proměnlivost jednotlivých proměnných a možné párové vztahy. Užívá se proto rozptylových diagramů závislosti x_j na x_k nebo indexových grafů závislosti x_j na j. Posuzuje se významnost proměnných s ohledem na jejich proměnlivost a přítomnost multikolinearity. Přibližně lineární vztah mezi proměnnými v rozptylových grafech závislosti x_j na x_k indikuje multikolinearitu. Lze rovněž odhalit vlivné body, které způsobují multikolinearitu.

Podle volby uživatele se provedou požadované transformace původních proměnných. Zadává se, zda model obsahuje absolutní člen. Uživatel může volit polynomickou transformaci zadáním stupně polynomu.

Provádí se sestavení korelační matice $\boldsymbol{R}$ a její rozklad na vlastní čísla a vlastní vektory. Jsou vypočteny *VIF* k indikaci multikolinearity a tisknuta setříděná vlastní čísla. K určení inverzní matice $\boldsymbol{R}^{-1}$ se užívá metoda racionálních hodností pro standardně zadávané vychýlení $P = 10^{-15}$. Uživatel může zadat jinou hodnotu parametru vychýlení P, což však vede pro vyšší hodnoty P k vychýleným odhadům. Bývá proto vhodné volit P z tohoto intervalu $10^{-5} \le P \le 10^{-3}$.

3. Odhadování parametrů: odhadování parametrů modelu se provádí metodou racionálních hodností s volbou $P = 10^{-5}$. Ze zobecněné inverzní matice $\boldsymbol{R}^{-1}$ jsou určovány odhady parametrů $\boldsymbol{b}$, jejich směrodatné odchylky $\sqrt{D(b_j)}$ a velikosti testačních statistik Studentova t-testu významnosti pro $\beta_j = 0$. Dále jsou provedeny testy významnosti odhadů b_j, vícenásobného korelačního koeficientu R a koeficientu determinace D. Je vhodné sledovat souhrnné charakteristiky regrese jako je střední kvadratická chyba predikce *MEP* a Akaikovo informační kritérium *AIC*, případně posoudit linearitu modelu.

4. Regresní diagnostika: s využitím pěti rozličných grafů je prováděna identifikace vlivných bodů, a to *grafy Wiliamsovým, Pregibonovým, McCulloh-Meeterovým, L-R,* a *grafem predikovaných reziduí*. Dále pak ověření splnění předpokladů metody nejmenších čtverců jako je homoskedasticita, nepřítomnost autokorelace a normalita rozdělení chyb. Pokud dojde k úpravě dat, je třeba provést znovu regresní diagnostiku se zaměřením na porušení předpokladů metody nejmenších čtverců a posouzení vlivu multikolinearity. V případě více vysvětlujících proměnných se posoudí vhodnost jednotlivých proměnných a jejich funkcí s využitím parciálních regresních grafů nebo grafů "komponenta + reziduum". *Tabulka reziduí*

obsahuje klasická rezidua $\hat{e}_i$, normovaná rezidua $\hat{e}_{Ni}$, standardizovaná rezidua $\hat{e}_{Si}$ a Jackknife rezidua $\hat{e}_{Ji}$. Je uveden odhad autokorelačního koeficientu reziduí prvního řádu $\hat{\varrho}_1$. *Tabulka vlivných bodů* obsahuje veličiny H_{ii}, $\boldsymbol{H}^*_{ii}$, D_i, A_i, DF_i, $LD_i(\boldsymbol{b})$, $LD_i(\hat{\sigma}^2)$ a $LD_i(\boldsymbol{b}, \sigma^2)$. Hvězdičkou jsou označeny hodnoty silně vlivných bodů.

5. Konstrukce zpřesněného modelu: s využitím

a) *metody vážených nejmenších čtverců (MVNČ)* při nekonstantnosti rozptylů,

b) *metody zobecněných nejmenších čtverců (MZNČ)* při autokorelaci,

c) *metody podmínkových nejmenších čtverců (MPNČ)* při omezeních na parametry,

d) *metody racionálních hodností RH* u multikolinearity,

e) *metody rozšířených nejmenších čtverců (MRNČ)* pro případ, že všechny proměnné jsou zatížené náhodnými chybami,

f) *robustní metody* pro jiná rozdělení dat než normální a data s vybočujícími hodnotami a extrémy jsou odhadovány parametry zpřesněného modelu.

6. Zhodnocení kvality modelu: s využitím klasických testů, postupů regresní diagnostiky a doplňkových informací o modelované soustavě se provede posouzení kvality navrženého lineárního regresního modelu.

Vzorová úloha: *Model teplotní závislosti přechodového tlaku bismutu* (J6.01)

Ukážeme postup analýzy jednorozměrného lineárního regresního modelu. Byl studován přechodový tlak bismutu I - II *p* jako funkce teploty *t.* Nalezněte lineární regresní model, který bude adekvátní daným datům. Vyšetřete regresní triplet a indikujte vlivné body.

Data: Teplota t [°C], tlak *p* [bar]:

20.8	25276,	20.9	25256,	21.0	25216,	21.9	25187,	22.1	25217,
22.1	25187,	22.4	25177,	22.5	25177,	24.8	25098,	24.8	25093,
25.0	25088,	34.0	24711,	34.0	24701,	34.1	24716,	42.7	24374,
42.7	24394,	42.7	24384,	49.9	24067,	50.1	24057,	50.1	24057,
22.5	25147,	23.1	25107,	23.0	25077				

Řešení:

1. Odhadování parametrů: klasickou metodou nejmenších čtverců (MNČ) byly nalezeny nejlepší odhady úseku β_0 a směrnice β_1. Studentův *t*-test ukázal, že úsek (absolutní člen) β_0 je statisticky významný a směrnice β_1 je statisticky významná.

	Odhad	Směrodatná odchylka	Test H0: B[j] = 0 vs. HA: B[j] <> 0 t-kriterium	hypoteza H0 je	Hlad. význam.
B[0]	2.6068E+04	1.6169E+01	1.6122E+03	Zamítnuta	0.000
B[1]	-3.9874E+01	5.0419E-01	-7.9084E+01	Zamítnuta	0.000

2. Regresní diagnostika: absolutní hodnota *párového korelačního koeficientu R* ukazuje, že navržený lineární regresní model je statisticky významný. Vysoká hodnota *koeficientu determinace* $D = R^2$ (99.67%), představuje procento variability, vysvětlené modelem. *Predikovaný koeficient determinace* R^2_p ukazuje na predikční schopnost modelu, je však vyčíslen jinak než R^2, místo *RSC* se ve vztahu užije *MEP*. *Střední kvadratická chyba predikce MEP* a *Akaikovo informační kritérium AIC* se užívají k rozlišení mezi několika navrženými modely. Za optimální se považuje model, pro který dosahuje *MEP* a *AIC* minimální hodnotu.

Vícenásobný korelační koeficient, R	: 9.9833E-01
Koeficient determinace, *D*	: 9.9665E-01
Predikovaný koeficient determinace, Rp^2	: 9.9804E-01

Střední kvadratická chyba predikce, *MEP*	: 6.8546E+02
Akaikeho informační kritérium, *AIC*	: 1.5054E+02

3. Konstrukce zpřesněného modelu:

(a) Po odstranění bodů č. 23 (*kritika dat*) byly nalezeny nové odhady parametrů zpřesněného modelu. Zpřesněný model (v závorce je uveden vždy odhad směrodatné odchylky parametru) y = 26 078 (13) - 40.1 (0.4) x_1 je doložen statistickými charakteristikami: *párový korelační koeficient* R = 0.9990, *koeficient determinace* D = 99.808% a *predikovaný korelační koeficient* R_P = 0.99885 dosáhly vesměs vysokých hodnot. *Střední kvadratická chyba predikce MEP* = 414.22 a *Akaikeho informační kritérium AIC* = 132.62 dosáhly nižších hodnot než u předešlého modelu, což dokazuje, že zpřesněný model je lepší. Rezidua nyní vykazují normální rozdělení a nevykazují trend, stále však vykazují heteroskedasticitu, a proto lze doporučit použít metodu vážených nejmenších čtverců.

(b) Užitím statistické váhy ($w_i = 1/y_i^2$) kompenzujeme heteroskedasticitu v datech. Obdržíme nové odhady parametrů. Opravený model má tvar, (v závorce je uveden odhad směrodatné odchylky parametru) y = 26 079 (13) - 40.1 (0.4) x_1. Jelikož došlo ke snížení rozhodujících kritérií, tj. *střední kvadratické chyby predikce MEP* = 410.29 a *Akaikeho informačního kritéria AIC* = 132.39, lze považovat tyto odhady za lepší než předešlé.

4. Zhodnocení kvality modelu: porovnáním hodnot regresní diagnostiky lze snadno provést zhodnocení *regresního tripletu* u dosaženého lineárního regresního modelu pro upravená data, zbavená odlehlých hodnot a metodou vážených nejmenších čtverců. Nalezený a prokázaný model teplotní závislosti přechodového tlaku bizmutu má tvar, (v závorce je vždy uveden odhad směrodatné odchylky parametru)

$$y = 26\,079\ (13) - 40.1\ (0.4)\ x_1.$$

Poděkování: Práce vznikla za podpory grantu Ministerstva zdravotnictví NS9831-4/2008 a vědeckých záměrů MSMT0021627502.

Doporučená literatura

[1] Meloun M., Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 1994, Academia Praha 2004..

[2] Meloun M., Militký J., Hill M., *Počítačová analýza vícerozměrných dat v příkladech,* Academia Praha 2005.

[3] Meloun M., Militký J., *Kompendium statistického zpracování experimentálních dat*, Academia Praha 2002, 2006.