8

NELINEÁRNÍ REGRESNÍ MODELY

8.1 Tvorba nelineárního regresního modelu

Postup tvorby nelineárního regresního modelu se dá rozčlenit do těchto kroků:

1. Návrh regresního modelu. Obvykle se jako nelineární regresní model používá nějaká fyzikální nebo empirická závislost.

2. Odhadování parametrů. Na rozdíl od lineárních regresních modelů je třeba pro hledání minima kritéria regrese použít iterativních algoritmů. V naprosté většině případů se používá kritérium minima součtu čtverců odchylek (reziduí).

3. Posouzení kvality odhadů. Kvalita nalezených odhadů se standardně posuzuje podle jejich intervalů spolehlivosti nebo pouze jejich rozptylů $D(b_j)$. Příčinou vysokých rozptylů parametrů bývá také předčasné ukončení minimalizačního procesu před dosažením minima.

4. Grafické posouzení vhodnosti modelu. Zahrnuje řadu metod a charakteristik. Grafická analýza reziduí využívá grafu reziduí vs. predikce, ve kterém lze snadno odhalit:

a) odlehlé hodnoty,

b) trend v reziduích,

c) nedostatečné střídání znaménka u reziduí,

d) heteroskedasticitu.

K ověření normality rozdělení reziduí lze užít i rankitových grafů a vyčíslení koeficientu šikmosti $g_1(\hat{e})$ a špičatosti $g_2(\hat{e})$.

5. Základní statistické charakteristiky. O přiblížení navrženého modelu k experimentálním datům informuje hodnota sumy čtverců reziduí v minimu $U(\boldsymbol{b})$, ze které se vyčíslí *reziduální rozptyl* $s^2 = U(\boldsymbol{b})/(n - m)$. Jednoduchou charakteristikou, založenou na hodnotě $U(\boldsymbol{b})$, je *koeficient determinace* D, který je pro lineární regresní modely čtvercem vícenásobného korelačního koeficientu,

$$D = 1 - \frac{U(\boldsymbol{b})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \quad \text{kde} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i ,$$

Stonásobek koeficientu determinace se nazývá *regresní rabat*, 100 *D* [%]. V literatuře se často nesprávně užívá *Hamiltonův R-faktor*

$$R\&faktor \;=\; \sqrt{\frac{U(\boldsymbol{b})}{\sum_{i=1}^{n} y_i^2}} \;,$$

Pro $\bar{y} = 0$ platí, že R^2*-faktor* = 1 - *D*. Pro $\bar{y} \neq 0$ platí mezi *R-faktorem* a koeficientem determinace *D* vztah

$$R\&faktor \;=\; \sqrt{(1-D) - \frac{(1-D)\, n\, \bar{y}^2}{\sum_{i=1}^{n} y_i^2}} \;.$$

Hamiltonův R-faktor ukazuje na rozdíl mezi modelem $y = f(x, \boldsymbol{\beta})$ a modelem $y = 0$, což u modelů s absolutním členem nemá smysl a hodnoty *Hamiltonova R-faktoru* vycházejí v těchto případech *nesprávně nízké*. Je třeba upozornit, že *D* i *R-faktor* jsou funkcí počtu parametrů modelu, a to *D* je funkcí rostoucí s počtem parametrů, zatímco *Hamiltonův R-faktor* klesající. Ani *D*, ani *R-faktor* není proto vhodným rozlišovacím kritériem k porovnání modelů o různém počtu parametrů.

K rozlišení mezi modely je vhodnější užít *Akaikova informačního kritéria AIC,* pro které platí $AIC = -L(\boldsymbol{b}) + 2m$. Za optimální se považuje model, pro který dosahuje *AIC* minimální hodnoty. Při použití metody nejmenších čtverců a modelů nepatřících do téže třídy je

$$AIC \;=\; n \ln\left[\frac{U(\boldsymbol{b})}{n}\right] + 2m \;.$$

6. Regresní diagnostika. Obsahuje stejně jako u lineárních regresních modelů pomůcky a postupy analýzy regresního tripletu, tj. pro *kritiku dat*, *kritiku modelu* a *kritiku metody*. Analýzou vlivných bodů se identifikují body, které silně ovlivňují odhadované regresní parametry v modelu, což umožňuje určit vybočující pozorování nebo extrémy. Pro aditivní modely měření a užívanou metodu nejmenších čtverců jsou rezidua definována vztahem $\hat{e}_i = y_i - f(x_i, \mathbf{b})$. Popis je uveden v 6. kapitole.

A. Analýza klasických reziduí. Kritika dat se skládá z analýzy několika druhů grafických diagnostik a tabulek různých druhů reziduí. V řadě programů aplikované nelineární regrese je analýza reziduí hlavní diagnostickou pomůckou při rozlišení chemického modelu, a navíc těsnost dosaženého proložení experimentálními body je mírou vhodnosti navrženého modelu. Mezi nejčastěji užívané charakteristiky patří *směrodatná odchylka reziduí s(ê)*, která by se měla rovnat velikosti šumu závisle proměnné *y*, *koeficient šikmosti* $g_1(\hat{e})$ a *koeficient špičatosti* $g_2(\hat{e})$ reziduí.

K testování reziduí lze užít všech statistik, známých z lineárních regresních modelů. Potíže zde činí pouze určení rozdělení testačních statistik, které jsou závislé na nelinearitě modelu.

B. Analýza vlivných bodů. U lineárních regresních modelů (viz 6. kapitola) jsou k dispozici všechny charakteristiky k odhalení vlivných bodů pomocí reziduí $\hat{e}_i$ a diagonálních prvků P_{ii} projekční matice $\boldsymbol{P} = \boldsymbol{X}\,(\boldsymbol{X}^{\mathrm{T}}\,\boldsymbol{X})^{-1}\,\boldsymbol{X}^{\mathrm{T}}$, zatímco u nelineárních modelů je rozdíl v matici $\boldsymbol{P}$. Matice $\boldsymbol{P} = \boldsymbol{J}\,(\boldsymbol{J}^{\mathrm{T}}\,\boldsymbol{J})^{-1}\,\boldsymbol{J}^{\mathrm{T}}$ totiž obsahuje $\boldsymbol{J}$ Jakobián čili derivaci modelové funkce podle jednotlivých parametrů v daných bodech.

U nelineárních regresních modelů je situace komplikována tím, že již nelze vyjádřit odhady parametrů a rezidua jako lineární kombinaci experimentálních dat. Pokud se užije linearizace nelineárního modelu, je možné užít přímo všech technik odhalení vlivných bodů v lineárních modelech. Vychází se z jednokrokové aproximace odhadu $\boldsymbol{b}_{(i)}$, pro kterou platí

$$\boldsymbol{b}_{(i)}^{1} \;=\; \boldsymbol{b} \;-\; \frac{(\boldsymbol{J}^{T}\,\boldsymbol{J})^{-1}\,\boldsymbol{J}_{i}\,\hat{e}_{i}}{1 \;-\; P_{ii}} \quad ,$$

kde P_{ii} jsou prvky projekční matice $\boldsymbol{P}$. Lze vyčíslit *charakteristiku* DFS_{ij}, která vyjadřuje vliv i-tého bodu na odhad j-tého parametru, vztahem

$$DFS_{ij} \;=\; \frac{b_{j} \;-\; b_{j(i)}^{1}}{\hat{s}_{(i)}\,\sqrt{V_{ii}}} \quad ,$$

kde $\hat{s}_{(i)}^{2}$ je odhad rozptylu vyčíslený při vynechání i-tého bodu, pro který platí

$$\hat{s}_{(i)}^{2} \;=\; \frac{U(\boldsymbol{b}) \;-\; \dfrac{\hat{e}_{i}^{2}}{1 \;-\; P_{ii}}}{n \;-\; m \;-\; 1} \quad ,$$

Symbol V_{ii} značí prvky matice $\boldsymbol{V} = (\boldsymbol{J}^{\mathrm{T}}\,\boldsymbol{J})^{-1}$. Při testování se považuje i-tý bod za vlivný, pokud je $DFS_{ij} > 2/\sqrt{n}$.

Vlivné body lze také identifikovat na základě jednokrokové aproximace *Jackknife reziduí* $\hat{e}_{Ji}$, pro kterou platí vztah

$$\hat{e}_{Ji} \;=\; \frac{\hat{e}_{i}}{\hat{s}_{(i)}\,\sqrt{1 \;-\; P_{ii}}} \quad ,$$

K vyjádření vlivu jednotlivých bodů na odhady parametrů lze použít i kvadratického rozvoje regresního modelu a vyčíslovat změny vektoru vychýlení $\boldsymbol{h}_{(i)}$ při vynechání

i-tého bodu nebo změny střední hodnoty i-tého rezidua při vynechání i-tého bodu. Mezi nelineární míry vlivu i-tého bodu na odhady parametrů patří *věrohodnostní vzdálenost*

$$LD_{i} \;=\; 2\,[\ln L(\boldsymbol{b}) \;-\; \ln L(\boldsymbol{b}_{(i)})] \quad .$$

Pro případ metody nejmenších čtverců bude věrohodnostní vzdálenost ve tvaru

$$LD_i = n \ln\left[\frac{U(\boldsymbol{b}_{(i)})}{U(\boldsymbol{b})}\right] .$$

Do obou vztahů lze dosadit buď odhady $\boldsymbol{b}_{(i)}$, určené regresí při vynechání i-tého bodu, nebo $\boldsymbol{b}^1_{(i)}$, určené z jednokrokové aproximace. Je-li $LD_i > \chi^2_{1-\alpha}(2)$, je daný bod silně vlivný. Obyčejně se volí $\alpha = 0.05$.

(a) Vlivné body ovlivňují nejenom odhady parametrů, ale také relativní vychýlení $\boldsymbol{h}_R$, které je značně citlivé na jejich výskyt.

(b) Charakteristiky založené na linearizaci nebo kvadratické aproximaci nelineárního modelu neindikují vždy správně přítomnost vlivných bodů. Hodí se především pro málo nelineární modely.

(c) Nejlepší indikaci vlivných bodů poskytuje věrohodnostní vzdálenost LD_i. Pouze tato charakteristika umožňuje indikaci celé skupiny vlivných bodů, kde může dojít k jejich vzájemnému "maskování".

(d) U praktických úloh postačuje aproximace LDS_i.

7. Mapa citlivostní funkce. Na rozdíl od lineárních regresních modelů je třeba u nelineárních modelů počítat s řadou komplikací, jako je neodhadnutelnost někte-rých parametrů, existence minima funkce $U(\boldsymbol{\beta})$ jen pro některé regresní modely, výskyt lokálních minim a existence sedlových bodů, ovlivňujících kriteriální funkci $U(\boldsymbol{\beta})$ a špatnou podmíněnost parametrů v regresním modelu. Tyto problémy lze částečně indikovat na základě analýzy *normalizovaných citlivostních koeficientů*

$$C_{j(i)} = \beta_j \frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_j} \qquad \begin{matrix} j = 1, \ldots, m \\ i = 1, \ldots, n \end{matrix} .$$

Pro vizuální posouzení špatné podmíněnosti, vzniklé jako důsledek přibližné multikolinearity mezi parametry β_j, β_h, se konstruují *citlivostní grafy*. Obyčejně jde o závislosti $C_{j(i)}$ a $C_{h(i)}$ na x_i, $i = 1, ..., n$. Lze také vynášet závislost normalizovaných citlivostních koeficientů přímo na indexu i.

Pro vyjádření citlivosti regresních modelů na změnu parametru β_j je možné využít celkové *citlivostní funkce*

$$C_{cj} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_j}\right]^2 .$$

Tato citlivostní funkce je nekonstantní pro takové parametry β_j, které jsou v modelu $f(x, \boldsymbol{\beta})$ nelineární.

Citlivostní grafy parametrů jsou pak závislosti C_{cj} na β_j v okolí bodů $\beta_j^{(0)}$ nebo b_j. Pokud jsou citlivostní grafy parametrů přibližně konstantní, indikuje to malou citlivost regresního modelu ke změnám j-tého parametru, nebo je model $f(x, \boldsymbol{\beta})$ vzhledem k parametru β_j *lineární*.

8. Predikční schopnost modelu. Predikční schopnost se může posoudit postupem “cross-validation”: data se rozdělí na dvě podskupiny M_1 (s indexy $i = 1, ..., \text{int}(n/2)$) a M_2

(s indexy i = int(n/2) + 1, ..., n). Označí se odhady parametrů z bodů podskupiny M_1 jako $\boldsymbol{b}(M_1)$ a z bodů podskupiny M_2 jako $\boldsymbol{b}(M_2)$. Predikční schopnost modelu lze pak vyjádřit kritériem

$$K = \frac{U(\boldsymbol{b})}{\sum_{i \in M_1} [y_i - f(x_i, \boldsymbol{b}(M_2))]^2 + \sum_{i \in M_2} [y_i - f(x_i, \boldsymbol{b}(M_1))]^2} .$$

Predikční schopnost modelu je tím vyšší, čím víc se hodnota K blíží k jedné. Mezi další kritéria patří *střední kvadratická chyba predikce*

$$MEP = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i, \boldsymbol{b}_{(i)}))^2 .$$

Místo odhadu $\boldsymbol{b}_{(i)}$ lze použít také jednokrokové aproximace $\boldsymbol{b}^1_{(i)}$. Čím je *MEP* nižší, tím je model věrohodnější a má lepší predikční schopnost.

9. Souhlas s požadavky fyzikálního smyslu. U navržených modelů jsou na odhady parametrů kladena omezení, vycházející z fyzikálního smyslu odpovídajících parametrů. Standardně se vyžaduje, aby odhady ležely v jisté předpokládané oblasti (např. koncentrace v oblasti kladných čísel, molární absorpční koeficienty ε v oboru čísel 10 až 10^6, konstanty stability log β_{pqr} v oboru čísel 0 až 50 atd.).

Program ADSTAT umožňuje numerickou a statistickou analýzu nelineárního regresního modelu $f(x, \boldsymbol{\beta})$ s využitím minimalizační hybridní strategie "double dog-leg". Vstupem je soubor experimentálních dat $\{x_i, y_i\}$, i = 1, ..., n, a nulté přiblížení odhadovaných parametrů $\boldsymbol{\beta}^{(0)}$. Uživatel zadává regresní model a může volit, zda se vybrané parametry zkonstantní.