6

LINEÁRNÍ REGRESNÍ MODELY

Při budování 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 výsledky metodou nejmenších čtverců své vlastnosti. Užití lineární regresní analýzy se týkají následujících možností:

1. *Popis dat:* hledáme vztah, lineární regresní model, který sumarizuje soubor dat. 2 *Určení parametrů:* nejběžnějším cílem regresní analýzy je vyčíslení nejlepších odhadů neznámých parametrů regresního modelu. Uživatel navrhne regresní model a regresní analýzou se snaží model prokázat. Často tento cíl překrývá i ostaní záměry regresní analýzy.

3. *Predikce:* nejdůležitějším cílem regresní analýzy je predikce, vyčíslení hodnot závisle proměnných. Bývá to často cena, dodací lhůta, účinnost, obložnost v nemoc-nici, výtěžek reakce, síla kovu, atd. Predikce jsou důležité i v plánování, monitoringu, vyhodnocování chemických procesů atd. Je však řada předpokladů a kvalifikací, které se musí respektovat v regresním modelu a datech. Často se například nesmí extrapolovat mimo rozsah dat. Intervalové odhady vyžadují dodržení předpokladu normality. Metoda nejmenších čtverců (MNČ) má svých sedm důležitých předpo-kladů, které je třeba respektovat a dodržet.

4. *Řízení:* regresní modely lze využít také k monitoringu a řízení systémů, například ke kalibraci měřicího systému. Když využijeme regresní model k řídicím účelům, nezávisle proměnné musí být vztaženy k závisle proměnné kauzálním způsobem.

5. *Výběr (volba) proměnných:* volba proměnných sleduje ty nezávisle proměnné, které vysvětlují významný objem proměnlivosti závisle proměnné. V řadě aplikací nejde o jednorázový proces, ale o spojitý proces výstavby modelu.

Základní předpoklady metody nejmenších čtverců: statistické vlastnosti odhadů $\hat{\boldsymbol{y}}_P, \hat{\boldsymbol{e}}$, $\boldsymbol{b}$ závisí na splnění jistých základních předpokladů.

Předpoklady metody nejmenších čtverců:

I. *Regresní parametry β mohou nabývat libovolných hodnot.* V praxi však existují často 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 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 kore-lačních 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* $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0$. 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 $\boldsymbol{X\beta}$ a kovarianční maticí $\sigma^2\boldsymbol{E}$, kde $\boldsymbol{E}$ je jednotková matice.

Pokud platí předpoklady I až VI, jsou odhady $\boldsymbol{b}$ parametrů $\boldsymbol{\beta}$ **n**ejlepší, **n**estranné a **l**ineární (NNLO). Navíc mají asymptoticky normální rozdělení. Pokud platí ještě předpoklad VII, mají odhady $\boldsymbol{b}$ normální rozdělení i pro konečné výběry.

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* (kritika dat, kritika modelu a kritika metody 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 H_A. 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 zde slouží pouze 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 regresní 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 z kap. 2. Přes svoji

jednoduchost umožňuje regresní 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 náhodnými veličinami.

Kvalita dat úzce souvisí s užitým regresním modelem. Při posuzování se sleduje především výskyt *vlivných bodů*, které mohou být hlavním zdrojem řady problémů, jako je zkreslení odhadů a růst rozptylů až k naprosté nepoužitelnosti regresních modelů. Vlivné body lze rozdělit do tří skupin:

a) *Hrubé chyby*, které jsou způsobeny měřenou veličinou (*vybočující pozorování*) nebo nevhodným nastavením vysvětlujících proměnných (*extrémy*). Hrubé chyby jsou obyčejně důsledkem chyb při manipulaci s daty.

b) *Body s vysokým vlivem* (tzv. golden points) jsou speciálně vybrané body, které byly přesně změřeny, a které obvykle rozšiřují predikční schopnosti modelu.

c) *Zdánlivě vlivné body* vznikají jako důsledek nesprávně navrženého regresního modelu.

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. O jejich výsledném vlivu však především rozhoduje to, že jsou extrémy. K identifikaci vlivných bodů typu vybočujícího pozorování se využívá zejména různých typů reziduí a k identifikaci extrémů pak diagonálních prvků H_{ii} projekční matice $\boldsymbol{H}$ (detaily v učebnici[72]).

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 $\hat{\boldsymbol{R}}_{yx_j(\boldsymbol{x})}$.

c) Rezidua v parciálním regresním grafu jsou shodná s klasickými rezidui $\hat{e}_i$ pro nedělený model.

d) V grafu jsou indikovány vlivné body a i některá porušení předpokladů metody nejmenších čtverců (heteroskedasticita).

Parciální reziduální grafy se označují také jako grafy *"komponenta + reziduum"*. Parciální reziduální grafy však poskytují poněkud odlišné informace než parciální regresní grafy:

a) Směrnice lineární závislosti je rovna b_j a úsek je nulový. Lineární závislost pak ukazuje na vhodnost navržené proměnné x_j v modelu.

b) Rezidua regresní přímky jsou přímo rezidua $\hat{e}_i$ pro nedělený model.

c) Pokud je úhel mezi $\boldsymbol{x}_j$ a některými sloupci matice $\boldsymbol{X}_{(j)}$ malý (*multikolinearita*), ukazuje parciální reziduální graf nesprávně malý rozptyl kolem regresní přímky $b_j \boldsymbol{x}_j$ a dochází navíc i k potlačení efektu vlivných bodů.

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), popř. 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 zatíženy 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

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

kde

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

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

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

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

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

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

Pro případ stejných rozptylů, tj. $K = 1$, vede dosazení do uvedených vztahů k odhadům minimalizujícím kolmé vzdálenosti (*ortogonální 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*, (**G**eneral **P**rincipal **C**omponent **R**egression) vedoucí k systému vychýlených odhadů, kde vychýlení je řízeno jedním parametrem.

Postup výstavby lineárního regresního modelu (ADSTAT)

(definice regresních diagnostik a ostatních statistických pojmů jsou v učebnici[72]):

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 $x_j x_k$. Pouze v případech, u kterých je předem známo, že model má obsahovat funkce vysvětlujících proměnných, může být výchozí model dle těchto požadavků upraven.

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 i vlivné body, které způsobují multikolinearitu.

Podle volby uživatele se provedou požadované transformace původních proměn-ných. Zadává se, zda model obsahuje absolutní člen. Uživatel může volit polyno-mickou transformaci zadáním stupně polynomu, Taylorův rozvoj do 2. stupně a lineární model s interakcemi. Uživatel může zadat libovolnou mocninu původních proměnných včetně logaritmů. Ostatní typy transformací se provádějí při přípravě dat k výpočtu v datovém editoru. K odstranění případné heteroskedasticity, vzniklé nelineární transformací proměnné y, je možné zadat nestatistické váhy, jež odpovídají kvazilinearizaci.

Provádí se sestavení korelační matice $\boldsymbol{R}$ a její rozklad na vlastní čísla a vlastní vektory. Jsou vypočteny faktory *VIF* (variation inflation factor) k indikaci multikolinearity a dále jsou vyčíslena setříděná vlastní čísla. K určení inverzní matice $\boldsymbol{R}^{-1}$ se užívá metoda racionálních hodností GPCR 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 intervalu $10^{-5} \leq P \leq 10^{-3}$.

3. Odhadování parametrů: odhadování parametrů modelu se provádí metodou racionálních hodnostní GPCR s volbou $P = 10^{-5}$, což je vlastně MNČ. 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*, popř. posoudit linearitu modelu.

4. Regresní diagnostika: identifikace vlivných bodů je prováděna využitím pěti rozličných grafů, a to *grafů Wiliamsova, Pregibonova, McCullohova-Meeterova, L-R,* a *grafu predikovaných reziduí*. Dále musíme ověřit 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 využitím parciálních regresních grafů nebo grafů "komponenta + reziduum". Obvykle jsou využívány následující tabulky:

Tabulka výsledků obsahuje hodnoty predikce $\hat{y}_i$, rozptylů predikce $D(\hat{y}_i)$ a relativní odchylky predikce od experimentálních dat. Je uvedena i průměrná absolutní, resp. relativní odchylka a reziduální suma čtverců *RSC*. Následuje statistická analýza klasických reziduí.

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 $\hat{\kappa}_1$.

Tabulka vlivných bodů obsahuje veličiny H_{ii}, $H_{ii}^{(}$, D_i, A_i, DF_i, $LD_i(\boldsymbol{b})$, $LD_i(\Phi^2)$ a $LD_i(\boldsymbol{b}, \Phi^2)$. Hvězdičkou bývají označeny hodnoty silně vlivných bodů.

5. Konstrukce zpřesněného modelu: při využití

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, kladených na parametry,

d) *metody racionálních hodnostní GPCR* u multikolinearity,

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

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

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

7. Kalibrační modely: u kalibračních modelů se pro daný signál y^* vypočte hodnota x^* spolu se svým konfidenčním intervalem. Před vlastním užitím kalibračního modelu je vhodné určit limitu detekce a limitu stanovení, které určují použitelnou dolní hranici kalibračního modelu nebo odpovídající metody. Postup obsahuje

(a) Návrh modelu.

(b) Statistickou analýzu reziduí.

(c) Výpočet derivací a integrálů.

(d) Určení kalibračních mezí.

(e) Sestavení kalibrační tabulky.

8. Testování různých hypotéz: ve zvláštních případech, jako je porovnání několika přímek atd., se provádí testování pomocí dalších testů k ověřování rozličných typů hypotéz.

Uživatel může při interaktivní práci s počítačem některé tabulky nebo grafy vynechat. Na základě analýzy vlivných bodů a reziduí lze provést i vypuštění některých bodů a výpočty pak zopakovat. Podrobný popis, vzorce a statistické testy najde čtenář v doporučované učebnici[72].