Regrese

Úloha: z **N** bodů dat $(\vec{x}_i,\ Y_i)$ určit **M** neznámých parametrů $a_1, \ldots, a_M$ závislosti

$$y = y(\vec{x}; a_1, \ldots, a_M) = y(\vec{x}; \vec{a})$$

$\vec{x}$ - 1 nebo více nezávislých proměnných - vysvětlující proměnné
y - vysvětlovaná proměnná

Hledaná závislost je nazývána **modelem**. Budeme se většinou zabývat lineárními modely, tj. modely lineárními vzhledem ke koeficientům a_j.

- **Lineární regrese (lineární modely)**
  - prostá (lineární závislost) $y = a_1 + a_2 \cdot x$
  - zobecněná (zobecněný polynom) $y = \sum_{j=1}^{M} a_j \cdot X_j(x)$
  - vícenásobná (lineární s více proměnnými) $y = a_1 + \sum_{j=2}^{M} a_j \cdot x_{j-1}$
  - zobecněná vícenásobná $y = \sum_{j=1}^{M} a_j \cdot X_j(\vec{x})$
- **Nelineární regrese (nelineární modely)**
  - linearizovatelné - např. $y = a_1 \cdot \exp(-a_2 x)$, $y = a_1 \cdot x^{a_2}$
  - nelinearizovatelné - např. $y = \sum_{j=1}^{M/2} a_{2j-1} \exp(-a_{2j} x)$

Předpoklady

- Hodnoty vysvětlující proměnné $\vec{x}_i$ jsou známy přesně (neobsahují náhodnou chybu)
- Vysvětlovaná proměnná y_i obsahuje náhodnou složku (je změřena s chybou) a tedy pro i = 1,. . .,N

  $$y_i = y(\vec{x}_i; a_1, \ldots, a_M) + e_i$$

  kde e_i je náhodná chyba měření.
- Střední hodnota náhodné chyby je nulová $E(e_i) = 0$ pro $\forall i = 1, \ldots, N$.
- Náhodné chyby jsou navzájem nekorelované $\mathrm{Cov}(e_i, e_k) = 0$ pro $i \neq k,\ i, k = 1, \ldots, N$.
- Náhodné chyby e_i stejné rozptyly (směrodatné odchylky) = homoskedasticita. Klasický případ – neznámé rozptyly.

Lineární regrese (zobecněná)

Zobecněný lineární model

$$Ey_i = \sum_{j=1}^{M} a_j \cdot X_j(\vec{x}_i) = \sum_{j=1}^{M} a_j x_{ij}$$

Náhodné veličiny y_i uspořádáme do náhodného vektoru $\vec{y} = (y_1, y_2, \ldots, y_N)$ a platí

$$E\vec{y} = \boldsymbol{X}\vec{a} \qquad\qquad \boldsymbol{X} = (x_{ij})$$

kde **regresní (konstrukční) matice** $\boldsymbol{X}$ má N řádků, M sloupců. Funkce $X_j(x)$ (resp. sloupce regresní matice) nazýváme **bázové funkce** zobecněného lineárního modelu. Dále předpokládáme, že $Dy_i = \sigma^2$, kde σ je neznámé. Model značíme $\vec{y} \sim (\boldsymbol{X}\vec{a}, \sigma^2\boldsymbol{I})$.

Vektor odchylek $\vec{e} = \vec{y} - \boldsymbol{X}\vec{a}$, kde $E\vec{e} = \vec{0}$ a kovarianční matice $\boldsymbol{\mathcal{D}}_{\vec{e}} = \sigma^2\boldsymbol{I}$.

Metoda nejmenších čtverců

Minimum vzhledem k $\vec{a}$ sumu kvadrátů odchylek měření Y_i od modelu $y(x_i)$

$$\min S(\vec{a}) = \min_{\vec{a}} \sum_{i=1}^{N} [Y_i - y(\vec{x}_i; \vec{a})]^2 = \sum_{i=1}^{N} \left(Y_i - \sum_{j=1}^{M} x_{ij} a_j \right)^2$$

Jednou z možností, jak minimum hledat je položit

$$\left.\frac{\partial S}{\partial a_j}\right|_{\widetilde{\vec{a}}} = 0 = -2 \sum_{i=1}^{N} x_{ij} \left(Y_i - \sum_{k=1}^{M} x_{ik} \widetilde{a_k} \right)$$

což vede k řešení systému M lineárních rovnic (pro $j = 1, \ldots, M$)

$$\sum_{k=1}^{n} \left(\widetilde{a_k} \sum_{i=1}^{n} x_{ij} x_{ik} \right) = \sum_{i=1}^{N} x_{ij} Y_i$$

který lze zapsat vektorově ve tvaru

$$\boldsymbol{X}^T\boldsymbol{X}\widetilde{\vec{a'}} = \boldsymbol{X}^T\vec{Y}$$

Toto je **systém normálních rovnic** s maticí $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}$ řádu $M \times M$.

Věta Nechť matice $\boldsymbol{A}$ normálních rovnic je regulární a tedy existuje právě jedno řešení $\widetilde{\vec{a}}$ systému normálních rovnic. Pak $\widetilde{\vec{a}}$ je nejlepším nestranným lineárním odhadem skutečných parametrů $\vec{a}$ modelu $y \sim (\boldsymbol{X}\vec{a}, \sigma^2\boldsymbol{I})$.

Dosadíme výsledek do definice modelu a dostáváme odhad $\widehat{y_i} = y(x_i)$ hodnoty Ey_i, vektorově ve tvaru

$$\hat{\vec{y}} = \boldsymbol{H}\vec{Y} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{Y}$$

Matice $\boldsymbol{H}$ je **projekční matice**.

Vysvětlení Na lineární regresi se mohu dívat jako na lineární projekci z N-rozměrného prostoru $\vec{y}$ do jeho M-rozměrného podprostoru, daného bázovými funkcemi (vektory = sloupci regresní matice).

Věta Vektor $\hat{\vec{y}}$ je nejlepším nestranným lineárním odhadem $E\vec{y}$.

Vektor $\vec{u} = \vec{Y} - \hat{\vec{y}}$ je vektor **reziduí** (klasická rezidua).

Věta Rezidua mají střední hodnotu $E\vec{u} = \vec{0}$
a kovarianční matici $\boldsymbol{\mathcal{D}}_{\vec{u}} = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$.

Pozn. Klasická rezidua nemají stejný rozptyl a nejsou navzájem nezávislá. Proto se konstruují další typy reziduí.

Kvadrát Eukleidovské normy $||\vec{u}||^2 = \sum_{i=1}^{N} u_i^2 = S(\widetilde{\vec{a}})$ se nazývá **reziduální součet čtverců** RSS a

$$S_y^2 = \frac{RSS}{N - M}$$

je nestranným odhadem rozptylu dat σ^2.

Věta Kovarianční matice odhadu $\widetilde{\vec{a}}$ parametrů modelu je

$$\boldsymbol{\mathcal{D}}_{\widetilde{\vec{a}}} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \simeq S_y^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Věta Vrstevnice (izočáry) $S(\vec{a})$ ohraničují konfidenční oblasti (oblasti spolehlivosti) v prostoru parametrů $\vec{a}$.

Prostá lineární regrese

Lineární model - přímka $y = a_1 + a_2 x$. Data $(Y_1, Y_2, \ldots, Y_N)$ získaná pro $(x_1, x_2, \ldots, x_N)$.
Bázové funkce $X_1(x) = 1$, $X_2(x) = x$. Regresní matice

$$\boldsymbol{X} = X_j(x_i) = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}$$

Systém normálních rovnic je

$$\begin{pmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{pmatrix} \begin{pmatrix} \widetilde{a_1} \\ \widetilde{a_2} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N Y_i \\ \sum_{i=1}^N x_i Y_i \end{pmatrix}$$

Determinant matice $\boldsymbol{A} = \boldsymbol{X}^T \boldsymbol{X}$ normálních rovnic je $\det(\boldsymbol{A}) =$
$= N \sum_{i=1}^N x_i^2 - (\sum_{i=1}^N x_i)^2$ a inverzní matice je

$$\boldsymbol{A}^{-1} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} = \frac{1}{\det(\boldsymbol{A})} \begin{pmatrix} \sum_{i=1}^N x_i^2 & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & N \end{pmatrix}$$

Oblasti spolehlivosti v prostoru parametrů (a_1, a_2) jsou dány rovnicí

$$N(a_1 - \widetilde{a_1})^2 + \sum_{i=1}^N x_i^2 \times (a_2 - \widetilde{a_2})^2 + 2 \sum_{i=1}^N x_i \times (a_1 - \widetilde{a_1})(a_2 - \widetilde{a_2}) = C - S(\widetilde{a_1}, \widetilde{a_2})$$

Pozn. Model $y = a'_1 + a'_2(x - \overline{x})$ má kovarianci $\mathrm{Cov}(\widetilde{a'}_1, \widetilde{a'}_2) = d_{\widetilde{a'}_1 \widetilde{a'}_2} = 0$.

Příklad Nechť $\vec{x} = (1, 2, 3, 4, 5)$ a $\vec{Y} = (2.1, 2.9, 4.1, 4.8, 6.1)$. Stanovte parametry prostého lineárního modelu.

Řešení Regresní matice

$$\boldsymbol{X} = X_j(x_i) = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix}$$

Normální rovnice

$$\begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix} \begin{pmatrix} \widetilde{a_1} \\ \widetilde{a_2} \end{pmatrix} = \begin{pmatrix} 20 \\ 69.9 \end{pmatrix}$$

Hledaný vektor koeficientů je $\vec{\widetilde{a}} = (1.03, 0.99)$, hodnoty vypočtené z modelu jsou $\vec{\widehat{y}} = (2.02, 3.01, 4, 4.99, 5.98)$.
Vektor reziduí $\vec{u} = \vec{Y} - \vec{\widehat{y}} = (0.08, -0.11, 0.1, -0.19, 0.12)$, reziduální suma čtverců $RSS = 0.079$, počet stupňů volnosti $\nu = 5 - 2 = 3$, odhad rozptylu $S_y^2 = 0.079/3 = 0.026$, odhad směrodatné chyby dat $S_y = 0.16$. Inverzní matice

$$\boldsymbol{A}^{-1} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} = \begin{pmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{pmatrix}$$

a tedy $\sigma_{\widetilde{a_1}} \simeq \sqrt{0.026 \times 1.1} = 0.17$ a $\sigma_{\widetilde{a_2}} \simeq \sqrt{0.026 \times 0.1} = 0.05$. Kovariance $\mathrm{Cov}(\widetilde{a_1}, \widetilde{a_2}) \simeq 0.026 \times -0.3 = -0.0079$, lineární korelační koeficient $r_{\widetilde{a_1},\widetilde{a_2}} \simeq -0.0079/(0.17 \times 0.053) = -0.88$.

Oblasti spolehlivosti v prostoru parametrů (a_1, a_2) jsou dány rovnicí

$$5(a_1 - 1.03)^2 + 55(a_2 - 0.99)^2 + 30(a_1 - 1.03)(a_2 - 0.99) = C - 0.079$$

Po diagonalizaci

$$0.78(a_1 - 0.277a_2 - 0.756)^2 + 54.93(a_2 + 0.277a_1 - 1.275)^2 = C - 0.079$$

Projekční matice $\boldsymbol{H}$ je

$$\boldsymbol{H} = \begin{pmatrix} 0.6 & 0.4 & 0.2 & 0 & -0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 & 0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0 & 0.1 & 0.2 & 0.3 & 0.4 \\ -0.2 & 0 & 0.2 & 0.4 & 0.6 \end{pmatrix}$$

Směrodatné odchylky reziduí jsou $\sigma_{u_i} \simeq \sqrt{S_y^2(1 - H_{ii})}$
a tedy $\vec{\sigma}_{\vec{u}} \simeq (0.10, 0.14, 0.15, 0.14, 0.10)$.

Normální lineární model

Předpokládáme, že odchylky e_i mají normální rozdělení $e_i \sim N(0, \sigma^2)$ a tedy vektor $\vec{e} \sim N(\vec{0}, \sigma^2 \boldsymbol{I})$. Potom i $\vec{y} \sim N(\boldsymbol{X}\vec{a}, \sigma^2 \boldsymbol{I})$.

Věta Metoda nejmenších čtverců je pro normální lineární model odhadem pomocí metody maximální věrohodnosti.

Odvození Podmíněná hustota pravděpodobnosti vektory $\vec{Y}$ za podmínky, že koeficienty modelu jsou $\vec{a}$ je

$$f(Y_1, \ldots, Y_N | \vec{a}) \sim \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\frac{Y_i - y(\vec{x}_i; \vec{a})}{\sigma}\right)^2\right]$$

kde $y(\vec{x}_i; \vec{a})$ je hodnota Ey_i odvozená z modelu. Poloha maxima f je shodná s $\ln f$ a také s polohou minima $-2\sigma^2 \ln f$. Hledáme tedy minimum funkce

$$S(\vec{a}) = \sum_{i=1}^{N} (Y_i - y(\vec{x}_i; \vec{a}))^2$$

Pro hodnoty parametrů $\vec{a}$, odpovídající minimu součtu kvadrátů reziduí, je hustota pravděpodobnosti naměření zpracovávaných dat maximální (maximální věrohodnost).

Pro normální lineární model lze využívat znalosti rozdělení statistik výběrů z normálního rozdělení. V praxi se často používá předpokladu, že jde o normální lineární model, pokud to není v rozporu s daty a nevede to k problémům.

Věta Pro normální lineární model má veličina $RSS/\sigma^2 = (N - M)S_y^2/\sigma^2$ rozdělení $\chi^2(N - M)$ (o $N - M$ stupních volnosti).

Pozn. To dovoluje intervalový odhad σ a tudíž lze zkontrolovat experimentálně udávanou chybu měření.

Pro normální lineární model lze pro oblasti spolehlivosti v prostoru $\vec{a}$ parametrů omezené vrstevnicemi $S(\vec{a}) = C = RSS + C'$ lze z hodnoty C' určit parametr spolehlivosti $(1 - \alpha)$.

Nekonstantní rozptyl dat (heteroskedasticita)

Pokud je známá závislost chyb měření na i a/nebo x_i ve tvaru $\sigma_i^2 = C \times f(i, x_i)$, kde konstanta C známa být nemusí, používáme **váhy** w_i

$$w_i \sim \frac{1}{\sigma_i^2}, \qquad \text{např.} \qquad w_i = \frac{1}{f(i, x_i)}$$

Metodu nazýváme **váženou metodou nejmenších čtverců** a hledáme

$$\min S(\vec{a}) = \sum_{i=1}^{N} w_i \left(Y_i - y(\vec{x}_i; \vec{a})\right)^2$$

Vysvětlení Hledáme vlastně minimum součtu čtverců relativních odchylek

$$S(\vec{a}) = \sum_{i=1}^{N} \left(\frac{Y_i - y(\vec{x}_i; \vec{a})}{\sigma_i}\right)^2$$

Regresní matice má pro váženou metodu nejmenších čtverců tvar

$$\boldsymbol{X} = \left(w_i^{1/2} X_j(x_i)\right) \simeq \left(\frac{X_j(x_i)}{\sigma_i}\right)$$

Pozn. Pokud jsou směrodatné odchylky dat přesně známy, pro odhady konfidenčních intervalů by měla být použita poněkud jiná rozdělení, než při směrodatných odchylkách spočtených z experimentu.

Upozornění Pokud jsou směrodatné odchylky závislé na hodnotě y (např. konstantní relativní odchylky nebo např. u četností odchylky $\sigma \sim \sqrt{y}$), pak váhy w_i jsou funkcí $Ey_i = y(x_i)$ a jde o úlohu na nelineární regresi! Často se úloha převádí přibližně na lineární regresi tím, že se při vyjádření $w_i = f(y(x_i)) \simeq f(Y_i)$. Převod implicitně předpokládá malé relativní chyby měření. Výsledky lze zpřesnit postupnou iterací vah, ale ani to není úplně přesné!

Kontrola modelu

Podmínky Pokud studovaná závislost musí splňovat určité podmínky (např. normalizační), model musí splnit tytéž podmínky a tím je oblast přípustných parametrů $\vec{a}$ omezena vazebnými podmínkami. Je třeba hledat buď podmíněný extrém nebo případně zmenšit počet parametrů tak, aby model vždy podmínky splnil.

Upozornění Pokud jsou metodou nejmenších čtverců nalezeny parametry $\widetilde{\vec{a}}$ takové, že po jejich dosazení model nesplňuje vazebné podmínky, je tento výsledný model zcela bezcenný!!

Přípustnost modelu Kontrola, zda model není v rozporu s daty. Jakmile některé kritérium zamítne statistickou hypotézu "model popisuje naměřenou závislost", pak $\Rightarrow$ jiný model (případně $\Rightarrow$ nesplněný předpoklad).

Grafická kontrola modelu Kromě vynesení dat a vypočtené funkce $y(x;\widetilde{\vec{a}})$ do xy grafu, vynáším vždy graf reziduí $u(x)$. Rezidua mají být náhodná a nekorelovaná! Pokud graf reziduí vykazuje pravidelnou závislost $\Rightarrow$ problém. Buď jde o projev korelace reziduí nebo model není schopen úplně vysvětlit závislost $y(x)$.

Znaménkový test přípustnosti modelu

Přípustnost modelu lze testovat na základě předpokládané nekorelovanosti odchylek dat. Rezidua by měla často měnit znaménko.
Znaménkový test - test frekvence změn znaménka.
Počet n_+ kladných reziduí, n_- záporných reziduí ($n_+ + n_- = N$) a počet sekvencí reziduí se stejným znaménkem n_u
(např. posloupnost -1,1,3,1,-2,-1,1 obsahuje 4 sekvence - $n_u = 4$).
Střední hodnota a rozptyl počtu sekvencí dán vztahy

$$\begin{aligned} En_u &= 1 + \frac{2n_+n_-}{n_+ + n_-} \simeq 1 + \frac{N}{2} \\ Dn_u &= \frac{2n_+n_-(2n_+n_- - n_+ - n_-)}{(n_+ + n_-)^2(n_+ + n_- - 1)} \simeq \frac{N}{4} \end{aligned}$$

Pro $n_+ > 10$ a $n_- > 10$ má veličina

$$U = \frac{n_u - En_u + 0.5}{\sqrt{Dn_u}}$$

přibližně normální rozdělení $\mathcal{N}$(0,1), pro menší hodnoty n_+, n_- jsou pravděpodobnosti U tabelovány. Pokud $P(U' \leq U) \leq \alpha$ (hladina významnosti α) zamítneme hypotézu, že model odpovídá datům.
Např. z 11 naměřených hodnot ($n_+ = 7$, $n_- = 4$) jsou pouze 3 sekvence reziduí se stejným znaménkem ($n_u = 3$), pravděpodobnost P($n_u \leq 3$) = 0.036 $\Rightarrow$ model není přípustný (za předpokladu nekorelovaných chyb měření v sousedních bodech).

Pozn. Pro malá N $(N \leq 15)$ znaménkový test často nedokáže zamítnout špatný model, ale graf reziduí jej může odhalit.

Koeficient determinace je veličina

$$R^2 = \frac{\sum_{i=1}^{N} \left(\widehat{y_i} - \overline{Y}\right)^2}{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2} = 1 - \frac{RSS}{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2}$$

kde $\overline{Y}$ je průměr Y_i.

Pozn. Veličina $CSS = \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2$ se nazývá celkový součet kvadrátů odchylek.

Koeficient významnosti modelu je veličina

$$F_R = \frac{(CSS - RSS)(N - M)}{RSS\ (M - 1)} = \frac{\left[\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2 - \sum_{i=1}^{N} \left(\widehat{y_i} - \overline{Y}\right)^2\right](N - M)}{\sum_{i=1}^{N} \left(\widehat{y_i} - \overline{Y}\right)^2\ (M - 1)}$$

Významnost modelu Pokud zjištěná hodnota F_R je statisticky významná, tj. pravděpodobně nevznikla náhodou při y nezávisejícím na $\vec{x}$, pak říkáme, že model je statisticky významný. Statisticky významný model = data aproximuje podstatně lépe než konstanta.

Věta Pro normální lineární model $\vec{y} \sim N(\boldsymbol{X}\vec{a}, \sigma^2\boldsymbol{I})$ má koeficient F_R významnosti modelu Fischerovo (Fischerovo-Snedecorovo) rozdělení.

Test χ^2 - jen při známých směrodatných odchylkách dat. Pro váženou metodu nejmenších čtverců a normální rozdělení odchylek se známými směrodatnými odchylkami σ_i má veličina

$$S(\widetilde{\vec{a}}) = \sum_{i=1}^{N} \left(\frac{Y_i - y(\vec{x}_i; \vec{a})}{\sigma_i} \right)^2 \qquad \sim \chi^2(N - M)$$

a pokud je $P(S' \geq S) \leq \alpha$ lze testem S vyloučit přípustnost modelu.

Pozn. Test je velmi citlivý na předpoklad normálního rozdělení odchylek a velikost chyby měření.
Pokud $S(\widetilde{\vec{a}})$ vyjde příliš malé ($P(S' \leq S) \leq \alpha$) $\Rightarrow$ chyby jsou menší než udávané.

Vlastnosti modelu

Hodnost modelu Pokud je hodnost regresní matice $\boldsymbol{X}$ menší než M, nejsou bázové funkce vzájemně nezávislé - nejsou nezávislé sloupcové vektory $X_j(x_i)$. Je třeba snížit počet bázových funkcí.

Multikolinearita modelu je situace, kdy jsou sice bázové funkce nezávislé, ale jsou blízké k vektorům vzájemně závislým. V takové situaci je velká lineární korelace navzájem mezi parametry a_j modelu. Multikolinearita vede ke zvětšení směrodatných odchylek parametrů a_j.

Pozn. U prosté lineární regrese dochází k multikolinearitě, pokud jsou všechna data soustředěna v relativně úzké oblasti daleko od osy.

Upozornění Při polynomiální regresi s bázovými funkcemi $1, x, x^2, \ldots, x^{M-1}$ a rovnoměrném rozložení x_i dochází při $m \gtrsim 6$ téměř vždy k multikolinearitě modelu.

Boj s multikolinearitou Nejspolehlivější metodou je **ortogonalizace** bázových funkcí.
Metoda nejmenších čtverců určuje skalární součin funkcí $f(x)$ a $g(x)$ ve tvaru

$$(f, g) = \sum_{i=1}^{N} w_i f(x_i) g(x_i)$$

Ortogonalizaci provedeme Gramm-Schmidtovou metodou. Parametry modelu jsou pak nekorelované.

Konfidenční interval parametru modelu Pro normální lineární model s neznámou směrodatnou odchylkou má statistika $T(a_j) = (a_j - \widetilde{a_j})/S_{a_j}$ Studentovo t-rozdělení o $(N - M)$ stupních volnosti. Označme β kvantil Studentova rozdělení $t_{N-M}(\beta)$. Pak intervalový odhad a_j se spolehlivostí $1 - \alpha$ je

$$\widetilde{a_j} - S_{a_j} t_{N-M}(1 - \alpha/2) \leq a_j \leq \widetilde{a_j} + S_{a_j} t_{N-M}(1 - \alpha/2)$$

Významnost parametru a_j Parametr a_j je významný na hladině významnosti α, pokud lze zamítnout hypotézu $a_j = 0$, tj. 0 leží mimo příslušný konfidenční interval.

Doporučení Pokud je některý parametr modelu nevýznamný, zkuste ho vynechat. Modelu s menším M se obvykle dává přednost - větší počet stupňů volnosti. Při regresi by mělo platit $M \leq \min(N/2, \sqrt{N})$.

Upozornění Pokud je model významný a všechny parametry nevýznamné, jde o projev multikolinearity modelu!

Kontrola kvality dat

Jedná se o posouzení vhodnosti dat pro model. Sleduje se především výskyt **vlivných bodů**, které mají mnohem větší vliv na odhady než ostatní.

Podle vlivu se dělí na 3 skupiny

- **Golden points** – body, které byly speciálně změřeny s vyšší přesností. Obvykle zlepšují vlastnosti modelu.
- **Hrubé chyby** – způsobené nevhodným nastavením vysvětlujících proměnných (**extrémy**) nebo měřenou veličinou (**vybočující pozorování**).
- **Zdánlivé vlivné body** - důsledek nesprávně navrženého modelu

Analýza prvků projekční matice - Pokud jsou v projekční matici $\boldsymbol{H}$ některé prvky H_{ii} přibližují k 1, odpovídající prvky podstatně ovlivní odhad parametrů. Pokud nejde o golden points, jde o nevhodný výběr dat (extrémy) nebo o chybnou konstrukci modelu.
Pozn. Pokud $H_{ii} = 1$, vypočtená závislost prochází Y_i.

Analýza reziduí K odhalení **vlivných bodů** může sloužit graf závislosti prostých reziduí na predikovaných reziduích. **Predikované reziduum** e_{Pi} je odchylka hodnoty Y_i od výsledku modelu, kde je dvojice (x_i, Y_i) vynechána.

Kontrola předpokladů modelu

Homoskedasticita (Rozptyl dat stejný pro všechny i) – Nejčastěji nastává závislost náhodných odchylek e_i na velikosti střední hodnoty $y(x_i) = Ey_i$. Tuto závislost odhalí graf závislosti kvadrátu u_i^2 reziduí na velikosti odhadu $\widehat{y_i}$ střední hodnoty Ey_i. Ještě lepší je použít kvadrátu standardizovaných reziduí

$$e_{Si} = \frac{u_i}{S_y\sqrt{1 - H_{ii}}}$$

protože tato rezidua by měla mít stejný rozptyl $De_{Si} = 1$.

Pokud zjistíme, že předpoklad neplatí, je třeba stanovit vhodné váhy w_i a použít metodu vážených nejmenších čtverců.

Nekorelované náhodné odchylky - Korelaci náhodných odchylek byla už posuzována při testování vhodnosti modelu (znaménkové kritérium, závislost u_i na x_i). Korelaci sousedních odchylek lze odhalit na grafu závislosti u_i na u_{i-1}.
Pokud jsou náhodné odchylky e_i korelované a je známa (odhadnuta) kovarianční matice

$$\mathcal{D}_{\vec{e}} = \sigma^2 \boldsymbol{K}$$

lze metodu nejmenších čtverců zobecnit tak, že systém normálních rovnic mají tvar

$$\boldsymbol{X}^T \boldsymbol{K}^{-1} \boldsymbol{X} \widetilde{\vec{a}} = \boldsymbol{X}^T \boldsymbol{K}^{-1} \vec{Y}$$

a projekční matice $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{K}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{K}^{-1}$.

Normalita modelu (normální rozdělení odchylek e_i) Seřadíme rezidua u_i podle velikosti $u_{(i)}$. Spočítáme μ_P kvantily $i/(N+1)$ normálního rozdělení $N(0,1)$. Tvar závislosti $u_{(i)}$ na μ_P může odhalit diferenci rozdělení odchylek e_i od normálního rozdělení.

Pásy spolehlivosti

Rozptyl vypočtených hodnot $\hat{y}$ - Kovarianční matice vypočtených hodnot $\hat{\vec{y}} = \boldsymbol{H}\vec{y}$ je

$$\mathcal{D}_{\hat{\vec{y}}} = \sigma^2 \boldsymbol{H} \simeq S_y^2 \boldsymbol{H}$$

Tedy $\sigma^2_{\hat{y}_i} \simeq S_y^2 H_{ii}$.

Označme sloupcový vektor $\vec{\Xi}(x) = (X_1(x), X_2(x), \ldots, X_M(x))^T$, $\vec{\Xi}_i = \vec{\Xi}(x_i)$, pak $\vec{\Xi}_i^T$ je i-tý řádek regresní matice. Pak $H_{ii} = \vec{\Xi}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \vec{\Xi}_i$.

Pro libovolné x je rozptyl vypočtené hodnoty $\hat{y}(x) = y(x; \widetilde{\vec{a}})$ je dán vztahem

$$D\tilde{y}(x) = \sigma^2 \vec{\Xi}^T(x)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\vec{\Xi}(x) \simeq S_y^2 \vec{\Xi}^T(x)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\vec{\Xi}(x)$$

Pro normální lineární model má veličina

$$\frac{Ey(x) - \tilde{y}(x)}{\sqrt{D\tilde{y}(x)}} \qquad \sim t_{N-M}$$

Studentovo rozdělení s $(N - M)$ stupni volnosti.

Konfidenčním intervalem $Ey(x)$ s koeficientem spolehlivosti α je interval

$$\tilde{y}(x) - \sqrt{D\tilde{y}(x)} t_{N-M}(1 - \alpha/2) \leq Ey(x) \leq \tilde{y}(x) + \sqrt{D\tilde{y}(x)} t_{N-M}(1 - \alpha/2)$$

Oblast na grafu $y(x)$ vyhovující výše uvedeném nerovnostem se nazývá **pás spolehlivosti** (konfidenční pás).

Pás spolehlivosti pro predikci je oblast, kde by se měla při libovolném pevném x data (výsledek měření) nalézat s pravděpodobností $1 - \alpha$. Při predikci je $S^2_{y_p} = S_y^2 + S_{\tilde{y}}^2$ a dál lze postupovat stejně jako u pásu spolehlivosti modelu.

Toleranční pás Oblast, v níž s pravděpodobností $(1 - \alpha)$ bude $\geq \beta$ % dat.

Příklad Pro prostou lineární regresi najděte $D\hat{y}(x)$ a pás spolehlivosti.

Řešení Označme matici normálních rovnic $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}$. Označme $\overline{x} = \sum_{i=1}^N x_i/N$ a $Dx = \sum_{i=1}^N (x_i - \overline{x})^2/N$. Pak

$$D\tilde{y}(x) \simeq S_y^2\, \vec{\Xi}^T(x)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\vec{\Xi}(x) = \frac{S_y^2}{\det(A)}\,(1\ x)\begin{pmatrix} \sum_{i=1}^N x_i^2 & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & N \end{pmatrix}\begin{pmatrix} 1 \\ x \end{pmatrix}$$

$$D\tilde{y}(x) \simeq \frac{S_y^2}{N}\left(1 + \frac{(x-\overline{x})^2}{Dx}\right)$$

Pro data z výše uvedeného příkladu je $N = 5$, $S_y^2 = 0.026$, $\overline{x} = 3$ a $Dx = 2$. Pak

$$D\tilde{y}(x) = 0.0052\left(1 + \frac{(x-3)^2}{2}\right)$$

Pro koeficient spolehlivosti $1 - \alpha = 0.95$ je $t_3(0.975) = 3.18$ a pás spolehlivosti je dán nerovnostmi

$$1.03{+}0.99x{-}0.23\sqrt{1 + \frac{(x-3)^2}{2}} \le y(x) \le 1.03{+}0.99x{+}0.23\sqrt{1 + \frac{(x-3)^2}{2}}$$

Vysvětlující proměnná zatížena náhodnými odchylkami

Často měříme nejen y, ale také x. Pokud náhodné odchylky v x nejsou zanedbatelné, je třeba metodu nejmenších čtverců modifikovat.

Předpokládáme

- Náhodná proměnná x_i, naměřená hodnota $X_i = Ex_i + \xi_i$
- Náhodná odchylka ξ_i má $E\xi_i = 0$ a rozptyl $D\xi_i = \tau_i$.
- Nekorelované odchylky $\mathrm{Cov}(\xi_i, \xi_j) = 0$ pro $i \neq j$.
- Vzájemně nekorelované odchylky x a y, tedy $\mathrm{Cov}(\xi_i, e_j) = 0$ pro $\forall i, j$.

Pro prostou lineární regresi $y = a_1 + a_2 x$ je rozptyl y_i je dán výrazem

$$Dy_i = \sigma_i^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\Big|_{x_i} \right)^2 \tau_i^2 = \sigma_i^2 + a_2^2 \tau_i^2$$

Minimalizujeme tedy

$$S(a_1, a_2) = \sum_{i=1}^{N} \frac{(Y_i - a_1 - a_2 X_i)^2}{\sigma_i^2 + a_2^2 \tau_i^2}$$

Pozn. Jde o nelineární regresi. Při rovnosti obou odchylek $\tau = \sigma$, jde o minimalizaci sumy kvadrátů kolmých vzdáleností.

Pozn. I při předpokladu konstantních σ^2 a τ^2 je nutno znát poměr σ^2/τ^2.

Linearizovatelné nelineární modely

Nechť $y = y(x; \vec{a})$ je nelineární funkce parametrů $\vec{a}$. Model je linearizovatelný, pokud existují transformace $z = g(y)$, $v = h(x)$ a $\vec{b} = G(\vec{a})$ takové, že model $z = z(v, \vec{b})$ je lineární.

Všechny předpoklady o původních proměnných je nutno převést na předpoklady o nových proměnných a naopak všechny výsledky je třeba převést na závěry o původním modelu.

Upozornění Nezapomenout na vztah mezi Dy_i a Dz_i.

Příklad Model $y = a_1 x^{a_2}$ je linearizovatelný, protože $\ln y = \ln a_1 + a_2 \ln x$.
Tedy $z = \ln y$, $v = \ln x$, $b_1 = \ln a_1$, $b_2 = a_2$.
Při nekorelovaných odchylkách y_i rozptyl $Dz_i \simeq (\mathrm{d}z/\mathrm{d}y)^2 Dy_i = Dy_i/(Ey_i)^2$.
Předpoklad o konstantních absolutních odchylkách z ($\sigma_{z_i} = C$) je ekvivalentní předpokladu o konstantních relativních odchylkách y ($\sigma_{y_i} = CEy_i = Cy(x_i)$).

Příklady linearizovatelných modelů

$$y(x) = a_1 \ \exp\left(- \sum_{j=2}^{M} a_j x^{j-1} \right)$$

$$y(x) = \frac{a_1}{1 + \sum_{j=2}^{M} a_j x^{j-1}}$$

Nelineární modely

Často se používají modely, které linearizovat nelze, například suma několika Gaussových funkcí

$$y(x;\vec{a}) = \sum_{j=1}^{K} a_{3j-2} \exp\left[-\frac{(x-a_{3j-1})^2}{2a_{3j}^2}\right]$$

kde $M = 3K$.

<u>*Pozn.*</u> Pro výše uvedenou funkci je třeba definovat $a_2 < a_5 < \ldots < a_{3M-2}$, aby ke každému minimu neexistovalo symetrické minimum dané přehozením parametrů.

Předpokládejme, že přesné hodnoty $x_i \in \mathcal{X}$ a náhodné veličiny

$$y_i = y(x_i;\vec{a}) + e_i$$

kde $Ee_i = 0$, $De_i = \sigma^2$ a $Cov(e_i, e_j) = 0$ pro $i \neq j$. Pak lze hledat odhad $\widetilde{\vec{a}}$ skutečných parametrů $\vec{a}$ minimalizací funkce

$$S(\vec{a}) = \sum_{i=1}^{N} \left(Y_i - y(x_i;\vec{a})\right)^2$$

Pro normální rozdělení e_i jde o odhad maximální věrohodnosti.
Hledáme absolutní minimum funkce $S(\vec{a})$.

<u>**Věta**</u> Nechť

- Parametry $\vec{a}$ jsou z M-rozměrné konvexní množiny Ω.
- Derivace regresní funkce

  $$\frac{\partial}{\partial a_j} y(x;\vec{a}) \qquad \text{a} \qquad \frac{\partial^2}{\partial a_j\, \partial a_k} y(x;\vec{a}) \qquad j,k = 1,2,,\ldots,M$$

  jsou spojitými funkcemi $\vec{a} \in \Omega$ pro $\forall x \in \mathcal{X}$.
- Nechť matice řádu $N \times M$

  $$\boldsymbol{X}(\vec{a}) = \frac{\partial}{\partial a_j} y(x_i;\vec{a})$$

  má hodnost M pro $\forall \vec{a} \in \Omega$.

Pak za poměrně obecných předpokladů regularity je poloha minima $\widetilde{\vec{a}}$ **konzistentním** odhadem skutečných parametrů $\vec{a}$ a $S_y^2 = S(\widetilde{\vec{a}})/(N-M) = RSS/(N-M)$ je **konzistentním** odhadem σ^2.

Přibližným odhadem kovarianční matice parametrů nelineární regrese je matice

$$\boldsymbol{C} = S_y^2 \left(\boldsymbol{X}^T(\widetilde{\vec{a}})\ \boldsymbol{X}(\widetilde{\vec{a}}) \right)^{-1}$$

Lze stanovit přibližná (platná pro velká N) rozdělení statistik

$$\frac{\widetilde{a_j} - a_j}{\sqrt{C_{jj}}} \sim t_{N-M} \qquad \frac{RSS}{\sigma^2} \sim \chi^2_{N-M}$$

a tedy přibližně stanovit intervalový odhad pro skutečný parametr a_j a rozptyl dat σ^2.

Metoda Monte Carlo se používá pro přesnější intervalový odhad parametrů nelineárního modelu. Ze skutečných dat získáme odhady $\widetilde{\vec{a}}$ a S_y^2. Pro tyto parametry vygenerujeme syntetická data a stanovíme parametry pro syntetická data.

Robustní odhady parametrů

Metoda nejmenších čtverců je silně ovlivněna body s velkou chybou měření (užívá se termín "vybočující pozorování"). Pokud máme podezření na hrubou chybu měření, bod vynecháme, případně lze jeho vliv zmenšit vhodnou vahou. Pokud není rozdělení odchylek dat normální a má dlouhá "křídla", je pravděpodobnost vybočujících pozorování velká, vybočujících pozorování je tedy více a nelze je obvykle vynechat. V této situaci může metoda nejmenších čtverců dát chybný odhad parametrů modelu.

Metoda nejmenší sumy absolutních hodnot minimalizuje funkci

$$M(\vec{a}) = \sum_{i=1}^{N} |Y_i - y(x_i; \vec{a})|$$

Tato metoda je méně ovlivněna vybočujícími pozorováními (jde tedy o robustní odhad parametrů). I pro lineární model metoda vede k řešení nelineárních rovnic.

Pozn. Pro prostou lineární regresi je výpočet odhadu parametrů metodou nejmenší sumy absolutních hodnot jednoduchý.

Postup regrese

1. Sestavení modelu
   - musí splňovat normalizační podmínky kladené na data
2. Určení hodnot neznámých parametrů - kritérium výběru
   - metoda nejmenších čtverců - bez vah, s vahami
   - robustní metody - suma absolutních odchylek
3. Nalézt směrodatnou odchylku měření (*pokud není dopředu známa*)
4. Nalézt rozptyl neznámých parametrů a korelace mezi nimi, *příp. oblast spolehlivosti (množinu v prostoru parametrů, kde se parametry nacházejí s pravděpodobností* $1 - \alpha$)
5. Významnost koeficientů - a_j *je významný, pokud lze vyloučit hypotézu* $a_j = 0$
   - snížení počtu koeficientů
   - multikolinearita a boj s ní
6. Kontrola kvality modelu
   - významnost modelu
   - náhodnost reziduí, graf reziduí, znaménkový test
   - χ^2 test
7. Kvalita dat pro model
   - rozsah dat, odlehlé body
   - vlivné body - hrubé chyby, zlaté body, zdánlivé vlivné body
8. Kontrola předpokladů modelu
   - rozptyl chyb měření (homoskedacita) a jejich korelace
   - rozdělení chyb
9. Pásy spolehlivosti - modelu, pro predikci