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 $\vec{a}$.

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.