Regresná analýza

Základné pojmy

Regresná analýza

skúma funkčný vzťah (priebeh závislosti), podľa ktorého sa mení závisle premenná $\boldsymbol{Y}$ pri zmenách nezávislých veličín $\boldsymbol{x_1, x_2, ..., x_k}$.

$$\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ni})^{\mathrm{T}}$$

$$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^{\mathrm{T}}$$

Regresná funkcia

Priebeh závislosti odhadujeme vhodnými funkciami (tzv. vyrovnávajúcimi):

$$Y = f\left(x_1, x_2, \ldots, x_k, \beta_1, \beta_2, \ldots, \beta_p, \varepsilon\right)$$

kde

$\beta_1, \beta_2, \ldots, \beta_p$ sú neznáme odhadované parametre regresnej funkcie

$\varepsilon = (\ \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\ )^{\mathrm{T}}$ je náhodná odchýlka (náhodná premenná)

Odhadované funkcie musia byť lineárne z hľadiska parametrov.

Označenia

$b_1, b_2, \ldots, b_p$ sú odhadmi $\beta_1, \beta_2, \ldots, \beta_p$

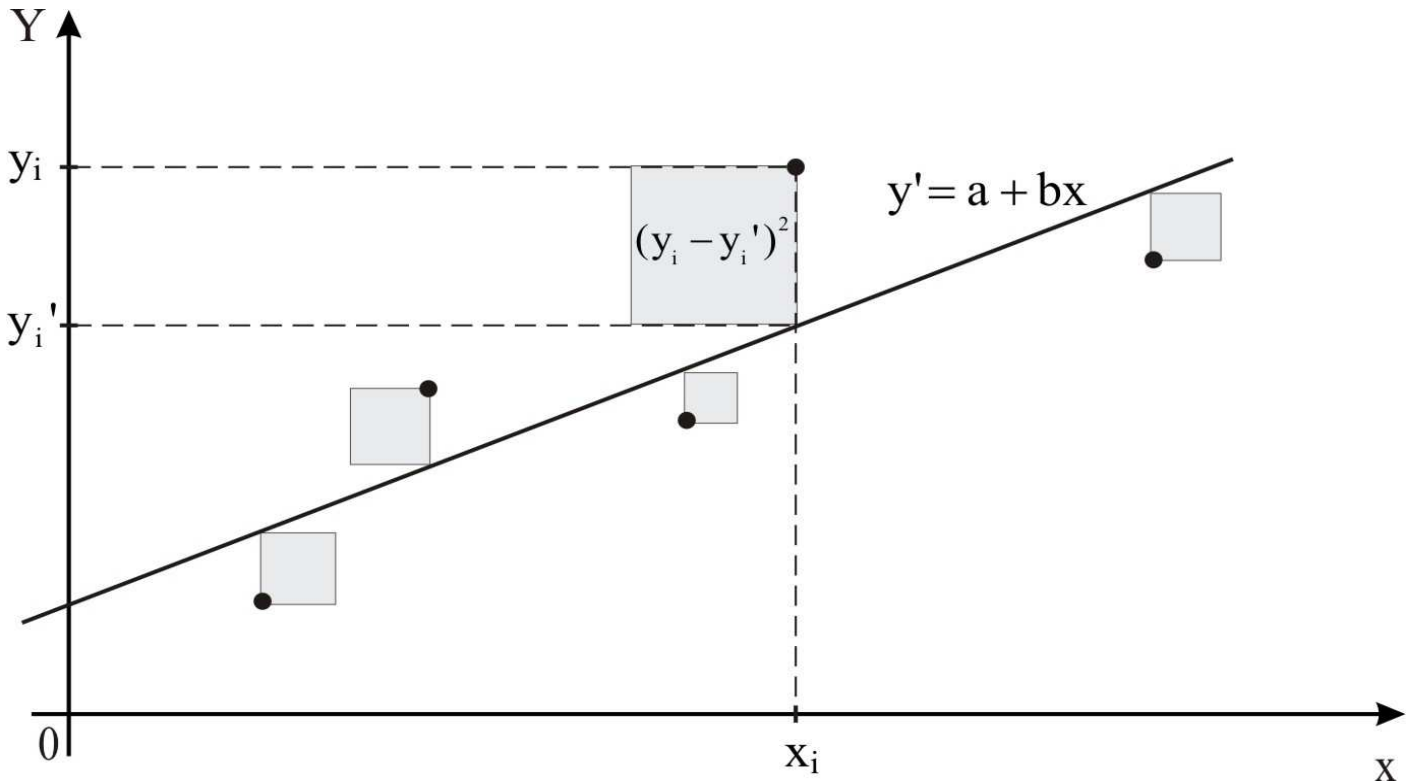
$$\hat{Y} = f(x_1, x_2, \ldots, x_k, b_1, b_2, \ldots, b_p)$$

$e_i = Y_i - \hat{Y}_i$ je odhad ε_i

Metóda najmenších štvorcov

- Princíp MNŠ: minimalizujeme výraz

$$S\left(\beta_1,\beta_2,\ldots,\beta_p\right)=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-\mathrm{f}\left(x_{i1},x_{i2},\ldots,x_{ik},\beta_1,\beta_2,\ldots,\beta_p\right)\right)^2$$



Podmienky MNŠ

- parametre β_j , $j = 1, 2, ..., p,$ sú nenáhodné a neznáme
- $E(\varepsilon_i) = 0$,
- $D(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$
- ε_i sú nekorelované, t.j. $Cov(\varepsilon_i, \varepsilon_j) = 0$ pre $i \neq j$,
- $\boldsymbol{x}_i$ sú lineárne nezávislé,

navyše predpokladajme, že

- ε_i majú normálne rozdelenie.

Vlastnosti regresnej funkcie získanej MNŠ

- $\sum_{i=1}^{n} e_i = 0$ t.j. súčet reziduálnych odchýlok je rovný nule,
- $\sum_{i=1}^{n} e_i^2$ je minimálny,
- regresná funkcia prechádza bodom

$$(\overline{y}, \overline{x}_1, \ldots, \overline{x}_k)$$

- odhad regresnej funkcie je najlepším lineárnym nevychýleným odhadom

Podľa **počtu premenných**, ktorých závislosť skúmame, hovoríme o

- jednoduchej (párovej) regresii
- viacnásobnej regresii

Regresná priamka

- regresný model má tvar:

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \; i = 1, 2, \ldots, n$$

- odhadom regresnej funkcie $\alpha + \beta x$ je

$$\hat{y} = a + bx$$

- odhady parametrov α, β regresnej funkcie a, b vypočítame metódou najmenších štvorcov:

$$(a, b) = \arg\min_{(\alpha,\beta)} S(\alpha, \beta) = \arg\min_{(\alpha,\beta)} \sum_{i=1}^{n} \left(Y_i - (\alpha + \beta x_i)\right)^2$$

Vypočítame prvé parciálne derivácie podľa neznámych parametrov, položíme ich rovné nule a dostaneme nasledujúcu sústavu dvoch rovníc o dvoch neznámych:

$$\sum_{i=1}^{n} y_i n_i = na + b\sum_{i=1}^{n} x_i n_i$$

$$\sum_{i=1}^{n} y_i x_i n_i = a\sum_{i=1}^{n} x_i n_i + b\sum_{i=1}^{n} x_i^2 n_i$$

- Riešením tejto sústavy získame nasledujúcu realizáciu **odhadu** jednotlivých parametrov:

$$b = \frac{\mathrm{cov}(x, y)}{s_x^{\,2}} \qquad a = \bar{y} - b\bar{x}$$

- **Interpretácie**:

a – lokujúca konštanta, nemá ekonomickú interpretáciu

b – regresný koeficient,

ekonomická interpretácia - udáva, o koľko merných jednotiek sa v priemere zmení závisle premenná Y, ak sa nezávisle premenná x zmení o jednu mernú jednotku.

Príklad:

Predajňa spoločnosti XXX robí zimný výpredaj. O dennej tržbe v tis. Sk (znak Y) a výške zľavy v percentách (znak X) máte nasledujúce informácie:

Denná tržba	Výška zľavy
- 200	10
- 250	20
- 320	30
- 380	40
- 470	50
- 550	60

Úloha:

Za predpokladu lineárnej závislosti medzi výškou tržieb a výškou zliav, odhadnite výšku tržby, ak by v predajni bola zľava na tovar 45 %.

Regresná parabola

- Model je:

$$Y_i = \alpha + \beta X_i + \gamma X_i^2 + \varepsilon_i \ , \ i = 1, \ldots, n$$

- odhadom vyrovnávajúcej funkcie je

$$\hat{y} = a + bx + cx^2$$

- koeficienty a, b, c neinterpretujeme

Pomocou MNŠ dostávame sústavu troch rovníc o troch neznámych, ktorej riešením získame hľadané parametre:

$$\sum_{i=1}^{n} y_i n_i = na + b\sum_{i=1}^{n} x_i n_i + c\sum_{i=1}^{n} x_i^2 n_i$$

$$\sum_{i=1}^{n} x_i y_i n_i = a\sum_{i=1}^{n} x_i n_i + b\sum_{i=1}^{n} x_i^2 n_i + c\sum_{i=1}^{n} x_i^3 n_i$$

$$\sum_{i=1}^{n} x_i^2 y_i n_i = a\sum_{i=1}^{n} x_i^2 n_i + b\sum_{i=1}^{n} x_i^3 n_i + c\sum_{i=1}^{n} x_i^4 n_i$$

Exponenciála

- model:

$$Y_i = \beta_1 \beta_2^{x_i} \varepsilon_i', \qquad i = 1, 2, \ldots, n$$

- funkcia nie je v parametroch lineárna, musíme ju najskôr zlogaritmovať, aby sme mohli použiť na odhad parametrov MNŠ:

$$\ln \hat{y} = \ln a + \ln bx$$

označenia: $A = \ln a$, $B = \ln b$

- sústava rovníc z MNŠ:

$$\sum_{i=1}^{n} \ln y_i n_i = n \ln a + \ln b \sum_{i=1}^{n} x_i n_i$$

$$\sum_{i=1}^{n} x_i n_i \ln y_i = \ln a \sum_{i=1}^{n} x_i n_i + \ln b \sum_{i=1}^{n} x_i^2 n_i$$

- riešením sústavy získame $A = \ln a$ a $B = \ln b$
- parametre a, b potom vypočítame nasledovne:

  $b = e^B$, $a = e^A$

Interpretácia:

- a – nemá ekonomickú interpretáciu
- b – koľkokrát sa zmení v priemere závisle premenná, ak sa nezávisle premenná zmení o jednu mernú jednotku

Združená regresia

- Združené regresné funkcie dostaneme, ak navzájom zameníme závislú a nezávislú premennú.
- Uhol, ktorý zvierajú grafy lineárnych združených regresných funkcií (priamky)

$$Y = \alpha_{yx} + \beta_{yx} x + \varepsilon_{yx} \quad \text{a} \quad X = \alpha_{xy} + \beta_{xy} y + \varepsilon_{xy}$$

je riešením **koeficientu korelácie** (čím je uhol menší, tým je závislosť silnejšia) a zároveň platí

$$\left| r_{yx} \right| = \sqrt{b_{yx} b_{xy}}$$

Viacnásobná lineárna regresia

$(Y, x_1, x_2, ..., x_k)$

- Skúma vplyv dvoch alebo viacerých nezávisle premenných na jednu závisle premennú Y.
- Regresný model môžeme zapísať v maticovom tvare:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

jeho odhad: $\hat{\boldsymbol{Y}} = \boldsymbol{Xb}$

Označenie v maticovom tvare:

$\boldsymbol{Y}$ je vektor závisle premennej s rozmerom $n \times 1$

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

β je vektor regresných koeficientov s rozmerom $(k+1) \times 1$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

$\boldsymbol{X}$ je matica nezávisle premenných rozmeru $n \times (k+1)$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

- Na odhad parametrov sa používa metóda najmenších štvorcov.
- Maticový zápis rovníc pre odhad parametrov je nasledujúci:

$$\boldsymbol{b} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

- Interpretácia:

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} \;,\; i = 1, 2, \ldots, n$$

b_i – o koľko merných jednotiek sa v priemere zmení závisle premenná, ak sa nezávisle premenná x_i zmení o jednu mernú jednotku,

za predpokladu, že ostatné nezávislé premenné zostanú konštantné

Normálne rovnice pre odhad parametrov modelu s dvomi vysvetľujúcimi veličinami

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

$$\sum_{i=1}^{n} y_i n_i = n b_0 + b_1 \sum_{i=1}^{n} x_{1i} n_i + b_2 \sum_{i=1}^{n} x_{2i} n_i$$

$$\sum_{i=1}^{n} x_{1i} y_i n_i = b_0 \sum_{i=1}^{n} x_{1i} n_i + b_1 \sum_{i=1}^{n} x_{1i}^{2} n_i + b_2 \sum_{i=1}^{n} x_{1i} x_{2i} n_i$$

$$\sum_{i=1}^{n} x_{2i} y_i n_i = b_0 \sum_{i=1}^{n} x_{2i} n_i + b_1 \sum_{i=1}^{n} x_{1i} x_{2i} n_i + b_2 \sum_{i=1}^{n} x_{2i}^{2} n_i$$

Interval spoľahlivosti pre regresný koeficient β

- je odvodený z náhodnej premennej, ktorá ma Studentovo rozdelenie s n-2 stupňami voľnosti

$$T = \frac{b - \beta}{S_b}$$

kde b je výberový regresný koeficient,

$$S_b = \frac{S_e}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}} \quad \text{a} \quad S_e = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

interval spoľahlivosti

$$(b - t_{1-\alpha/2,n-2}.S_b ; b + t_{1-\alpha/2,n-2}.S_b)$$

Test hypotézy o regresnom koeficiente β

- test, či je regresný koeficient štatisticky významný:

  1^0 H_0: $\beta = 0$ H_1: $\beta \neq 0$

  2^0 α

  3^0 testovacia štatistika

$$T = \frac{b - 0}{S_b}$$

  4^0 kritická oblasť je daná kvantilmi Studentovho rozdelenia, t.j. ak $|T| > t_{1-\alpha/2,n-2}$