

LINEÁRNÍ REGRESNÍ MODEL

Dvourozměrný náhodný výběr: nezávislé dvojice $(X_1, Y_1), \dots, (X_n, Y_n)$, jsou nezávislé kopie náhodného vektoru (X, Y) ,
 X a Y jsou spojité náhodné veličiny,
např. výška X a hmotnost Y vysokoškolského studenta.

Data: pozorované (naměřené) číselné hodnoty $(x_1, y_1), \dots, (x_n, y_n)$, například výška a hmotnost n studentů.

Sílu lineární závislosti veličin X a Y měří **korelační koeficient** ρ_{XY} .
Platí $|\rho_{XY}| \leq 1$,

$\rho_{XY} = 1 \Leftrightarrow Y = \alpha + \beta X, \beta > 0, \rho_{XY} = -1 \Leftrightarrow Y = \alpha + \beta X, \beta < 0$,
 X, Y nezávislé $\Rightarrow X, Y$ nekorelované ($\rho_{XY} = 0$).

Výběrový korelační koeficient $r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$,

$$s_{XY} = \frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1} (\sum X_i Y_i - n \bar{X} \bar{Y}),$$

$$s_X^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2 = \frac{1}{n-1} (\sum X_i^2 - n \bar{X}^2), s_Y^2 \text{ analogicky.}$$

Teoreticky je r_{XY} náhodná veličina, prakticky číslo vypočítané z dat.

Nabývá hodnot z intervalu $\langle -1, 1 \rangle$, $|r_{XY}| \approx 1 \Rightarrow Y \approx \alpha + \beta X$,
neboli $Y = \alpha + \beta X + \varepsilon$, kde Y je závisle proměnná,
 X je nezávisle proměnná a ε je náhodná chyba.

Model **LINEÁRNÍ REGRESE** (regresní přímka) má **předpoklady**:

(i) pro jednotlivé dvojice $(X_1, Y_1), \dots, (X_n, Y_n)$ je

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad i = 1, \dots, n,$$

(ii) náhodné chyby ε_i jsou nezávislé, normálně rozdělené
s nulovou střední hodnotou a konstantním rozptylem σ^2 ,

(iii) $X_i = x_i, i = 1, \dots, n$ jsou nenáhodné.

Z předpokladů plyne:

$Y_i = \alpha + \beta X_i + \varepsilon_i$, $i = 1, \dots, n$ mají podmíněně

při známých $X_i = x_i$, $i = 1, \dots, n$ normální rozdělení

se střední hodnotou $EY_i = \alpha + \beta x_i$, neboť $E\varepsilon_i = 0$, $i = 1, \dots, n$

a s rozptylem $\text{var } Y_i = \sigma^2$, neboť přičtení konstanty $\alpha + \beta x_i$

k náhodné veličině ε_i rozptyl neovlivní.

Práce s modelem regresní přímky:

- odhad parametrů α, β ,
- testování nulovosti parametrů,
- ověření předpokladů,
- posouzení kvality modelu.

Odhady parametrů α, β : metoda nejmenších čtverců ...

minimalizujeme součet čtverců odchylek $\varepsilon_1, \dots, \varepsilon_n$

bodů o souřadnicích $(x_1, Y_1), \dots, (x_n, Y_n)$

od proložené přímky $y = \alpha + \beta x$, tedy řešíme úlohu

$$\min \sum (Y_i - \alpha - \beta x_i)^2.$$

To lze učinit derivováním podle α, β a položením derivací rovných 0.

Řešení: odhad parametru β je $b = \frac{S_{xY}}{S_x^2}$,

odhad parametru α je $a = \bar{Y} - b\bar{x}$.

Předpovědi: $\hat{Y}_i = a + bx_i$... odhady pro $EY_i = \alpha + \beta x_i$, $i = 1, \dots, n$.

Rezidua: $e_i = Y_i - \hat{Y}_i$, $i = 1, \dots, n$.

Odhady a, b stejně jako předpovědi $\hat{Y}_i$ a rezidua e_i jsou teoreticky náhodné veličiny, prakticky čísla spočítaná z dat.

Body $(x_i, \hat{Y}_i)$, $i = 1, \dots, n$ leží na odhadnuté regresní přímce.

$\hat{Y}_i$ lze chápat jako předpověď hodnoty veličiny Y_i lineárním regresním modelem při známém x_i . To se uplatní např. při příchodu nového studenta s výškou x_{n+1} , jehož hmotnost nemáme možnost změřit.

Testy hypotéz o parametrech:

$H_0^{(\alpha)}$: $\alpha = 0$... regresní přímka prochází počátkem,

$H_0^{(\beta)}$: $\beta = 0$... **Y nezávisí formou modelu lineární regrese na X .**

(platí: $\beta = 0 \Leftrightarrow \rho_{XY} = 0$.)

$H_0^{(\beta)}$ zamítáme na hladině 5 %, když:

- $|b|$ je dostatečně velké (odtud testová statistika a kritický obor),

- 0 neleží v 95 % - ním intervalovém odhadu pro β ,

- p-hodnota je menší než 5 %.

Pro $H_0^{(\alpha)}$ analogicky.

Analýza reziduí vypočítaných z dat $e_i = y_i - a - bx_i, i = 1, \dots, n$ znamená grafické **ověření předpokladů**:

- bodové grafy dvojic $(x_i, e_i), (i, e_i), i = 1, \dots, n$ by měly mít body rovnoměrně roztroušené v rovině kolem nulové úrovně,

jinak graf (x_i, e_i) naznačuje nelineární závislost EY_i na x_i ,

případně nekonstantní rozptyl veličin ε_i a Y_i ,

graf (i, e_i) někdy umožní odhalit závislost mezi veličinami ε_i ,

a tedy i mezi veličinami Y_i ,

- normalita veličin ε_i a $Y_i, i = 1, \dots, n$:

histogram reziduí - měl by být jednovrcholový symetrický,

normální diagram (Q-Q-plot) - měl by mít lineární průběh,

případně testy normality aplikované na rezidua (p-hodnota $> 0,05$).

Koeficient determinace: nástroj pro posouzení shody modelu s daty.

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}.$$

Teoreticky náhodná veličina, prakticky číslo vypočítané z dat.

Nabývá hodnot z intervalu $\langle 0,1 \rangle$, interpretuje se jako **procento**

variability závisle proměnné Y , které se podařilo vysvětlit modelem. V modelu regresní přímky platí: $R^2 = r_{XY}^2$.

Dobrá shoda modelu s daty $\Rightarrow$ předpovědi $\hat{Y}_i$ jsou blízké pozorovaným hodnotám $Y_i \Rightarrow R^2$ **blízký 1.**

Čím menší je R^2 , tím spíše působí na Y ještě jiné vlivy než X nebo se Y chová zcela náhodně.

MNOHONÁSOBNÁ LINEÁRNÍ REGRESE

Rozšiřuje model regresní přímky na více nezávisle proměnných.
Uvažujme n nezávislých kopií náhodných veličin $Y, X_1, \dots, X_k$:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} \dots & \dots X_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} \dots & \dots X_{nk} \end{pmatrix}.$$

Pracujeme s daty, která představuje
 n pozorovaných realizací závisle proměnné Y
a nezávisle proměnných (regresorů) $X_1, \dots, X_k$.

Např. Y ... krevní tlak, X_1 ... věk, X_2 ... hmotnost,
 X_3 ... hladina cholesterolu v krvi apod., vše měřeno u n osob.

Předpoklady:

(i) $Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i, i = 1, \dots, n,$

(ii) náhodné chyby ε_i jsou nezávislé, normálně rozdělené
s nulovou střední hodnotou a konstantním rozptylem σ^2 ,

(iii) $X_{ij} = x_{ij}, i = 1, \dots, n, j = 1, \dots, k$ jsou nenáhodné,

(iv) matice $\mathbf{X}$ má lineárně nezávislé sloupce.

Y je spojitá náhodná veličina, regresory $X_1, \dots, X_k$ většinou spojité,
ojediněle diskrétní. Např. X_4 ... pohlaví (0 muž, 1 žena)
 $\Rightarrow$ model má pro muže konstantu β_0 a pro ženy $\beta_0 + \beta_4$.

Z předpokladů plyne: náhodné veličiny

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i, i = 1, \dots, n$$

podmíněně při známých $X_{ij} = x_{ij}$ normálně rozdělené,

$$EY_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}, \text{ var } Y_i = \sigma^2.$$

Odhady parametrů: metoda nejmenších čtverců ...

$$\min \sum (Y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik})^2.$$

Řešení lze spočítat maticově:
$$\begin{pmatrix} b_0 \\ \vdots \\ b_k \end{pmatrix} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{Y}).$$

Testy hypotéz o parametrech:

$H_0^{(0)}: \beta_0 = 0$... model neobsahuje konstantu (intercept),

$H_0^{(j)}: \beta_j = 0$... **Y nezávisí formou modelu**

mnohonásobné lineární regrese na $X_j, j = 1, \dots, k$.

Předpovědi: $\hat{Y}_i = b_0 + b_1 x_{i1} + \dots + b_k x_{ik}, i = 1, \dots, n.$

Rezidua: $e_i = Y_i - \hat{Y}_i, i = 1, \dots, n.$

$\hat{Y}_i$ jsou odhadem středních hodnot $EY_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$, reprezentují předpověď hodnoty veličiny Y_i modelem.

Analýza reziduí = ověření předpokladů, probíhá analogicky jako v modelu regresní přímky, k odhalení nelinearity závislosti EY_i na x_i a nekonstantního rozptylu ε_i a Y_i se kreslí graf dvojic $(\hat{Y}_i, e_i), i = 1, \dots, n.$

Normalita náhodných chyb ε_i , a tedy i veličin $Y_i, i = 1, \dots, n$ není potřeba pro odhad parametrů modelu metodou nejmenších čtverců, ale **je potřeba pro testování hypotéz o parametrech** k určení kritických oborů, intervalových odhadů parametrů a p-hodnot.

Nesplnění normality reziduí: výsledky **testování nulovosti parametrů $\beta_1, \dots, \beta_k$ lze zkontrolovat pomocí koeficientu determinace R^2** (počítá se stejně jako v modelu regresní přímky): malé snížení koeficientu determinace v modelu s vyloučeným regresorem X_j oproti původnímu modelu $\Rightarrow X_j$ v modelu mnohonásobné lineární regrese nepřispívá zásadně k vysvětlení variability závisle proměnné Y (β_j lze považovat za nulový).