

13. Regrese

Jedním z hlavních úkolů matematické statistiky je hledání a studium závislostí mezi dvěma či více proměnnými. Závisle proměnná se zpravidla označuje Y a nezávisle proměnné $X_1, \dots, X_i, i \geq 1$. Závislosti mezi Y a skupinou proměnných X mohou být funkční nebo statistické.

V případě přímé funkční závislosti je náhodná část ve vyjádření této závislosti nulová, proto se jí nebudeme samozřejmě zabývat. V běžných situacích však nastává situace, kdy Y má povahu náhodné veličiny. Nezávisle proměnné $X_1, \dots, X_i$ mohou být nenáhodnými (pevnými) veličinami nebo náhodnými veličinami.

Statistickou závislostí mezi náhodnou veličinou Y a veličinami $X_1, \dots, X_i$ rozumíme předpis, který každé uspořádané i – tici $x_1, \dots, x_i$ přiřazuje právě jedno podmíněné rozdělení náhodné veličiny Y . Náhodná veličina Y je tedy statisticky závislá na veličinách $X_1, \dots, X_i$, jestliže změny v hodnotách $x_1, \dots, x_i$ mají za následek změnu podmíněného rozdělení náhodné veličiny Y .

13.1 Teoretické základy

V praxi není většinou zákon rozdělení (typ náhodné veličiny Y) znám. Máme většinou k dispozici pouze data ve formě uspořádaných $i+1$ – tic, kde prvních i složek odpovídá hodnotám x a poslední složka odpovídá hodnotě y . Na základě experimentálních údajů se provede výběr funkce, která má co nejlépe vystihovat rozložení skutečných údajů, tuto funkci nazýváme teoretickou regresní funkcí a je uváděna ve tvaru

$$y = f(x_1, \dots, x_i; \gamma_0, \dots, \gamma_p) \quad (13.1),$$

kde hodnoty $\gamma_0, \dots, \gamma_p$ jsou nazývány teoretickými regresními parametry (koeficienty). Tyto parametry mají povahu konstant, které nejsou známy. V modelovaných vztazích je budeme nahrazovat jejich bodovými odhady $a_0, \dots, a_p$ tak, aby funkce

$$\hat{Y} = f(x_1, \dots, x_i; a_0, \dots, a_p) \quad (13.2)$$

co nejlépe aproximovala naměřené hodnoty Y . Funkce (13.2) se nazývá empirickou regresní funkcí a veličiny $a_0, \dots, a_p$ empirickými regresními parametry. Empirické regresní parametry jsou náhodné veličiny. Jednou z metod jak je určit je použít metodu nejmenších čtverců.

V dalším textu bude dále závisle proměnná Y (**vysvětlovaná** proměnná) náhodnou veličinou, naopak nezávisle proměnné (**vysvětlující** proměnné) $X_1, \dots, X_i$ jsou nenáhodnými veličinami.

Regresní závislosti bude pro naše potřeby jednostranná závislost dané náhodné veličiny (např. Y) na jiné veličině – ne nutně náhodné veličině (např. $X_1, \dots, X_i$). Pro další účely třídění a studování regresních závislostí se rozlišují následující situace:

a. Náhodná veličina závisí na jedné další veličině – párová regrese. Například :

$$y = \gamma_0 \cdot x + \frac{\gamma_1}{x + \gamma_3}$$

b. Náhodná veličina –závisí na dvou a více veličinách – mnohonásobná regrese

$$\text{.Například : } y = \gamma_1 \cdot x_1 + \frac{\gamma_2 \cdot x_2}{\gamma_3 + x_3} + \gamma_0$$

Z pohledu tvaru funkce (13.1) můžeme dále dělit jednotlivé případy regresních závislostí na:

1. Lineární regresi – uvažovaná funkce f je lineární vzhledem k parametrům $\gamma_0, \dots, \gamma_p$. Například : $y = \gamma_1 \cdot x_1 + \gamma_2 \cdot x_2 + \gamma_0$

2. Nelineární regresi – uvažovaná funkce f je nelineární funkcí vzhledem k parametrům $\gamma_0, \dots, \gamma_p$. Například : $y = \frac{\gamma_1 \cdot x_1 + \gamma_2 \cdot x_2}{\gamma_3 \cdot x_3 + \gamma_4 \cdot x_4}$.

Regresní analýzou nazveme tedy tu část matematické statistiky , která se zabývá studiem a konstrukcí regrese. Nejdříve určíme podmínky regresního modelu.

Definice 13.1

Určujeme regresi Y na i nenáhodných veličinách $x_1, \dots, x_i$. Nechť je dále dáno n ($n > p+1$) uspořádaných $i+1$ – tic $(x_{1k}, \dots, x_{ik}, y_k)$, kde $k = 1, \dots, n$. Tyto údaje byly postupně získány empiricky. Na základě těchto údajů byla zvolena teoretická regresní funkce s neznámými parametry $\gamma_0, \dots, \gamma_p$. Dále předpokládejme , že pro náhodné veličiny Y_k , $k=1, \dots, n$ platí:

$$1. Y_k = f(x_1, \dots, x_i; \gamma_0, \dots, \gamma_p) + e_k \quad (13.2)$$

Veličiny e_k představují náhodné chyby měření a jsou zřejmě také náhodnými veličinami. Tyto náhodné veličiny musí splňovat následující podmínky:

$$2. E(e_k) = 0 \quad (13.3)$$

$$3. \text{Var}(e_k) = \sigma^2 \quad (13.4)$$

$$4. \text{cov}(e_k, e_l) = 0 \quad (13.5)$$

Dále platí

$$5. E(Y_k) = f(x_{1k}, \dots, x_{ik}; \gamma_0, \dots, \gamma_p)$$

$$6. \text{VAR}(Y_k) = \sigma^2.$$

K řešení výše popsaného modelu používáme nejčastěji metodu **nejmenších čtverců**. Součástí řešení je nalezení odhadů parametrů $\gamma_0, \dots, \gamma_p$. Tyto odhady budeme dále označovat symboly $c_1, \dots, c_p$. Metoda spočívá v nalezení minima čtverců hodnot e_k . Obecněji budeme studovat tuto metodu aplikovanou na případ lineární regrese.

Hledáme tedy minimum funkce

$$S_r(\gamma_0, \dots, \gamma_p) = \sum_{k=1}^n e_k^2 = \sum_{k=1}^n (Y_k - f(x_1, \dots, x_i; \gamma_0, \dots, \gamma_p))^2, \quad (13.6)$$

požadujeme tedy , aby součet čtverců odchylek naměřených hodnot Y_k od teoretických hodnot $f(x_1, \dots, x_i; \gamma_0, \dots, \gamma_p)$ byl minimální. Najdeme tedy nejdříve podezřelé body (stacionární body S) :

$$\frac{\partial S(\gamma_0, \dots, \gamma_p)}{\partial \gamma_k} = 0, \quad k = 0, \dots, p \quad (13.7)$$

Poté ověříme pomocí klasických metod, zda v nalezených bodech se nachází minimum. Soustava rovnic (13.6) se nazývá **soustavou normálních rovnic**. Řešení této soustavy je jednoznačné , jestliže alespoň $p + 1$ z uspořádaných k – tic $[x_{1k}, \dots, x_{ik}]$, $k=1, \dots, n$ je navzájem různých. Řešením soustavy (13.6) budou potom hodnoty $a_1, \dots, a_p$, které budou bodovým odhadem parametrů $\gamma_0, \dots, \gamma_p$. Odtud tedy získáme bodové odhady teoretické regresní funkce (13.1) ve tvaru

$$\hat{Y}_k = f(x_{1k}, \dots, x_{ik}; a_0, \dots, a_p), \quad k = 1, \dots, n$$

Rozdíl

$$\hat{e}_k = Y_k - \hat{Y}_k \quad (13.8)$$

se nazývají rezidui a veličina

$$S_r = \sum_{k=1}^n \hat{e}_k^2 = \sum_{k=1}^n (Y_k - \hat{Y}_k)^2 \quad (13.9)$$

se nazývá **reziduálním součtem čtverců**.

Studiem řešení soustavy normálních rovnic a reziduálního součtu čtverců se budeme zabývat dále v následujících částech této kapitoly a v kapitole následující.

13.2 Model lineární regrese

V této části se budeme zabývat nejjednodušším případem lineární regrese – případem , kdy odhadujeme jen dva parametry α a β . Takovýto speciální vztah označujeme jako regresní přímku a je vyjádřen

$$Y = \alpha + \beta \cdot x + e \quad (13.10)$$

K odvození bodových odhadů skutečných parametrů α a β použijeme klasickou metodu nejmenších čtverců. Bodové odhady označíme pro naše potřeby a a b . Tedy podle (13.6) je

$$S(a, b) = \sum_{k=1}^n e_k^2 = \sum_{k=1}^n (Y_k - a - b \cdot x_k)^2 \quad (13.11)$$

Tento výraz budeme derivovat podle proměnných a , b . Výsledná soustava normálních rovnic je potom následující

$$\frac{\partial S(a, b)}{\partial a} = -2 \cdot \sum_{k=1}^n (Y_k - a - b \cdot x_k) = 0 \quad (13.12)$$

$$\frac{\partial S(a, b)}{\partial b} = -2 \cdot \sum_{k=1}^n x_k \cdot (Y_k - a - b \cdot x_k) = 0 \quad (13.13)$$

Přepíšeme – li tuto soustavu klasickým způsobem máme:

$$\begin{aligned} n \cdot a + b \cdot \sum_{k=1}^n x_k &= \sum_{k=1}^n Y_k, \\ a \cdot \sum_{k=1}^n x_k + b \cdot \sum_{k=1}^n x_k^2 &= \sum_{k=1}^n x_k Y_k \end{aligned} \quad (13.14)$$

Z této soustavy vyplývá , že řešení existuje , je – li determinant soustavy různý od nuly. Tedy $n \cdot \sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k \right)^2 \neq 0$, ale tento vztah nastane, jestliže hodnoty x_k budou navzájem různé (stačí dokonce , aby aspoň dva byly navzájem různé). Soustava tedy bude mít právě jedno řešení. Nejprve upravíme první rovnici v (13.14) tím , že ji vydělíme hodnotou n , získáme tak

$$a + b \cdot \bar{x} = \bar{Y} \quad (13.15)$$

Budeme – li znát jednu z hodnot a , b pomocí tohoto vztahu dopočteme druhou . Určíme hodnotu b , dosazením (13.15) do druhého vztahu v (13.14).

$$\begin{aligned} (\bar{Y} - b \cdot \bar{x}) \cdot n \cdot \bar{x} + b \cdot \sum_{k=1}^n x_k^2 &= \sum_{k=1}^n x_k Y_k \Rightarrow \\ b \cdot \left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right) &= \sum_{k=1}^n x_k Y_k - n \cdot \bar{x} \cdot \bar{Y} \Rightarrow \\ b &= \frac{\sum_{k=1}^n x_k Y_k - n \cdot \bar{x} \cdot \bar{Y}}{\left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)} \end{aligned} \quad (13.16)$$

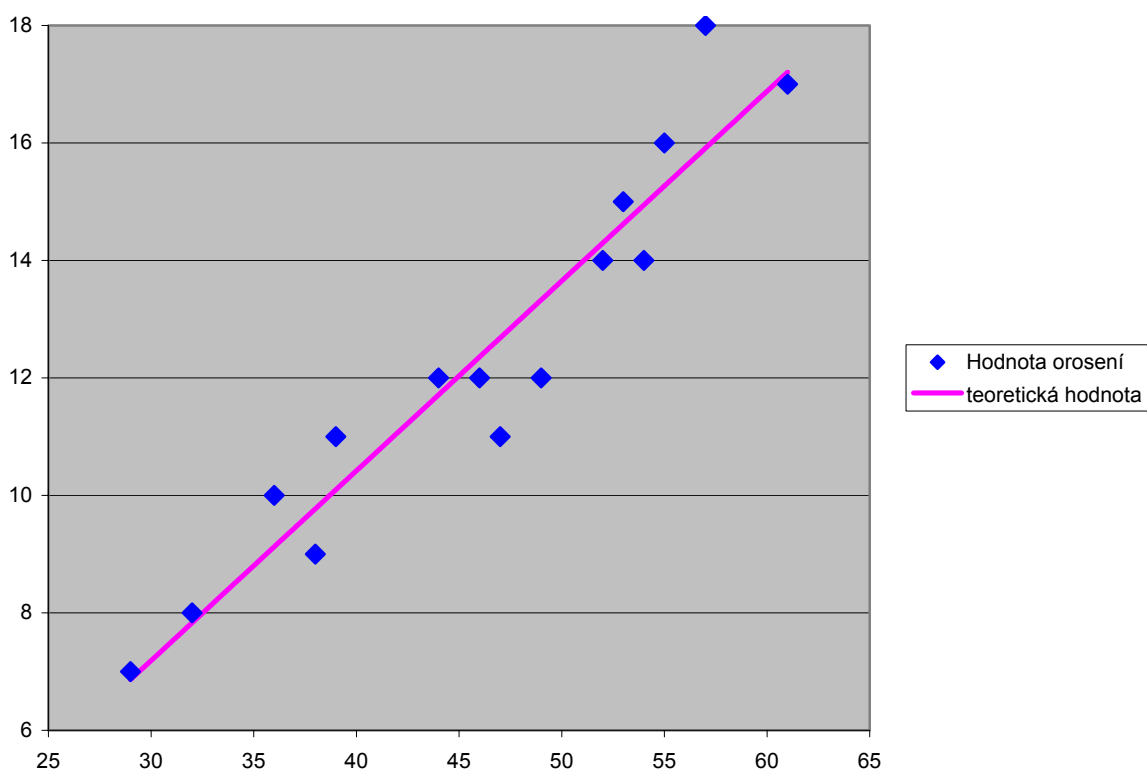
Ze vztahů (13.15) a (13.16) určíme samozřejmě hodnoty a , b .

Příklad 13.2

V 15 za sebou následujících dnech byla zaznamenána hodnota relativní vlhkosti v % a hodnota orosení v mm^3 . Naměřené údaje jsou uvedeny v tabulce dále

Relativní vlhkost	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Hodnota orosení	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

Provedeme – li výpočet získáme následující hodnoty $a = -2,51$ a hodnota $b = 0,32$. Výsledek můžeme také zobrazit graficky



Z výše uvedeného grafu je patrné , že žádná ze skutečných naměřených hodnot Y na přímce neleží , přesto však je vyjádření vztahu mezi x a Y pomocí lineárního vztahu dobré.

Podle předpokladů lineárního modelu předpokládáme , že střední hodnota chyb e_k je rovna nule, jejich rozptyl je konstantní a roven σ^2 . Dále předpokládáme, že hodnoty Y_k (tzv. vysvětlované proměnné) jsou typu normální rozdělení. Odtud tedy vyplývá , že $E(Y_k) = \alpha + \beta \cdot x_k$ a dále $VAR(Y_k) = \sigma^2$. Celkově tedy $Y_k \sim N(\alpha + \beta \cdot x_k, \sigma^2)$. Budeme dále vyšetřovat vlastnosti bodových odhadů parametrů α a β . Tedy

$$E(b) = E \left(\frac{\sum_{k=1}^n x_k Y_k - n \bar{x} \bar{Y}}{\left(\sum_{k=1}^n x_k^2 - n \bar{x}^2 \right)} \right) = \frac{\sum_{k=1}^n (x_k - \bar{x}) E(Y_k)}{\left(\sum_{k=1}^n x_k^2 - n \bar{x}^2 \right)} = \frac{\sum_{k=1}^n (x_k - \bar{x}) \cdot (\alpha + \beta \cdot x_k)}{\left(\sum_{k=1}^n x_k^2 - n \bar{x}^2 \right)} =$$

$$= \frac{\alpha \cdot \sum_{k=1}^n (x_k - \bar{x}) + \beta \cdot \left(\sum_{k=1}^n x_k \cdot (x_k - \bar{x}) \right)}{\left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)} = \beta \cdot \frac{\left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)}{\left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)} = \beta.$$

Protože je $E(b) = \beta$, je podle definice 11.12 z kapitoly 11. tento odhad nevychýlený. Zjistíme dále hodnotu rozptylu tohoto odhadu

$$\begin{aligned} VAR(b) &= VAR \left(\frac{\sum_{k=1}^n x_k Y_k - n \cdot \bar{x} \cdot \bar{Y}}{\left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)} \right) = \frac{\sum_{k=1}^n (x_k - \bar{x})^2 \cdot VAR(Y_k)}{\left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)^2} = \frac{\sigma^2 \cdot \sum_{k=1}^n (x_k - \bar{x})^2}{\left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)^2} = \\ &= \frac{\sigma^2}{\left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)} \end{aligned} \quad (13.17)$$

Při výpočtu jsme využili následující rovnost

$$\sum_{k=1}^n (x_k - \bar{x})^2 = \sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \quad (13.18)$$

Podobně nyní zjistíme střední hodnotu a rozptyl pro odhad **a**.

$$E(a) = E \left(\frac{\sum_{k=1}^n Y_k}{n} - b \cdot \bar{x} \right) = \sum_{k=1}^n \frac{\alpha + \beta \cdot x_k}{n} - \bar{x} \cdot E(b) = \alpha + \beta \cdot \bar{x} - \bar{x} \cdot \beta = \alpha,$$

tedy i bodový odhad parametru α je nevychýlený. Rozptyl hodnoty **a** je roven

$$VAR(a) = \frac{\sum_{k=1}^n VAR(Y_k)}{n^2} + (\bar{x}^2) \cdot VAR(b) = \frac{\sigma^2}{n} + \frac{\sigma^2 \cdot \bar{x}^2}{\left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)} = \sigma^2 \cdot \frac{\sum_{k=1}^n x_k^2}{n \cdot \left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)}$$

Závěrem můžeme tedy konstatovat, že náhodné veličiny **a** a **b** jsou typu normální rozdělení (jsou lineární kombinací nezávislých náhodných veličin typu normálního rozdělení) a můžeme navíc určit jejich parametry tedy

$$a \sim N \left(\alpha, \frac{\sigma^2 \cdot \sum_{k=1}^n x_k^2}{n \cdot \left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)} \right) \text{ a dále } b \sim N \left(\beta, \frac{\sigma^2}{\left(\sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \right)} \right).$$

Tyto dva vztahy budou klíčové pro další určení intervalových odhadů skutečných parametrů α a β .

Věta 13.3

Nechť S_r je reziduální součet čtverců definovaný vztahem (13.6). Potom $\frac{S_r}{\sigma^2} \sim \chi_{n-2}^2$.

Důkaz:

Proveden například v [1].

Poznámka 13.4

Výše uvedený důkaz budeme motivovat následující úvahou: Náhodné veličiny Y_k jsou nezávislé, jsou dále normálního typu. Pokud je znormujeme máme

$$\frac{Y_k - E(Y_k)}{\sqrt{VAR(Y_k)}} = \frac{Y_k - \alpha - \beta \cdot x_k}{\sigma} \quad (13.19)$$

Proto tedy

$$\sum_{k=1}^n \left(\frac{Y_k - \alpha - \beta \cdot x_k}{\sigma} \right)^2 \sim \chi_n^2$$

Protože ale nahradíme skutečné hodnoty α a β jejich nevychýlenými odhady $\mathbf{a}$ a $\mathbf{b}$, sníží se počet stupňů volnosti o 2 – tedy o počet parametrů, které jsme odhadovali.

Pro další práci označme

$$S_{xy} = \sum_{k=1}^n (x_k - \bar{x}) \cdot (Y_k - \bar{Y}) = \sum_{k=1}^n x_k Y_k - n \cdot \bar{x} \cdot \bar{Y} \quad (13.20)$$

$$S_{xx} = \sum_{k=1}^n (x_k - \bar{x})^2 = \sum_{k=1}^n x_k^2 - n \cdot \bar{x}^2 \quad (13.21)$$

$$S_{yy} = \sum_{k=1}^n (Y_k - \bar{Y})^2 = \sum_{k=1}^n Y_k^2 - n \cdot \bar{Y}^2 \quad (13.22)$$

Jestliže použijeme tuto notaci můžeme koeficienty $\mathbf{a}$ a $\mathbf{b}$ zapsat také takto

$$b = \frac{S_{xy}}{S_{xx}} ; \quad a = \bar{Y} - b \cdot \bar{x}.$$

Dále odvodíme následující výpočetní identitu :

$$S_r = \frac{S_{xx} \cdot S_{yy} - S_{xy}^2}{S_{xx}} \quad (13.23)$$

$$\begin{aligned} S_r &= \sum_{k=1}^n (Y_k - a - b \cdot x_k)^2 = \sum_{k=1}^n \left(Y_k - \bar{Y} + \frac{S_{xy}}{S_{xx}} \cdot \bar{x} - \frac{S_{xy}}{S_{xx}} \cdot x_k \right)^2 = \\ &= \sum_{k=1}^n \left(\frac{S_{xx} \cdot (Y_k - \bar{Y}) - S_{xy} \cdot (x_k - \bar{x})}{S_{xx}} \right)^2 = \frac{S_{xx}^2 \cdot S_{yy} - 2 \cdot S_{xx} \cdot S_{xy} \cdot S_{xy} + S_{xy}^2 \cdot S_{xx}}{S_{xx}^2} = \\ &= \frac{S_{xx} \cdot S_{yy} - S_{xy}^2}{S_{xx}} \end{aligned}$$

Ze všech výše uvedených rovností vyplývají následující tvrzení:

$$1. \quad \frac{S_r}{\sigma^2} \sim \chi_{n-2}^2 \quad (13.24)$$

$$2. \quad \frac{\frac{(b - \beta)}{\sigma}}{\sqrt{\frac{S_{xx}}{S_r}}} = \sqrt{\frac{(n-2) \cdot S_{xx}}{S_r}} \cdot (b - \beta) \sim t_{n-2} \quad (13.25)$$

$$3. \sqrt{\frac{n \cdot (n-2) \cdot S_{xx}}{\sum_{k=1}^n x_k^2 \cdot S_r}} \cdot (a - \alpha) \sim t_{n-2} \quad (13.26)$$

Toto jsou základní vztahy, pomocí nichž je možno konstruovat jak intervalové odhady neznámých parametrů α a β , ale i neznámé hodnoty σ . Dále jsou na nich založeny klasické statistické hypotézy o těchto parametrech.

Často jsme postaveni před otázku nalezení odhadu (prognózy) hodnot v nějakém bodě x_0 , na základě znalosti vstupních dat $\{x_k, Y_k\}$. Podle způsobu naší konstrukce je zřejmé, že přirozeným bodovým odhadem je v tomto případě hodnota $a + b \cdot x_0$ (je opět nevychýleným odhadem skutečné neznámé hodnoty – dokažte!). Podobně jako v předchozích krocích nás bude především zajímat typ náhodné veličiny.

$$a + b \cdot x_0 = \frac{\sum_{k=1}^n Y_k}{n} - \frac{\sum_{k=1}^n Y_k \cdot (x_k - \bar{x}) \cdot (\bar{x} - x_0)}{S_{xx}} = \sum_{k=1}^n Y_k \cdot \left(\frac{1}{n} - \frac{(x_k - \bar{x}) \cdot (\bar{x} - x_0)}{S_{xx}} \right) \quad (13.27)$$

Z předchozího vztahu (13.25) vyplývá (proč?), že náhodná veličina $a + b \cdot x_0$ je typu normálního rozdělení. Je tedy nutno zjistit jeho střední hodnotu a rozptyl.

$$A. E(a + b \cdot x_0) = E(a) + x_0 \cdot E(b) = \alpha + \beta \cdot x_0$$

B.

$$\begin{aligned} VAR(a + b \cdot x_0) &= \sum_{k=1}^n \left(\frac{1}{n} - \frac{(x_k - \bar{x}) \cdot (\bar{x} - x_0)}{S_{xx}} \right)^2 \cdot VAR(Y_k) = \sigma^2 \cdot \sum_{k=1}^n \left(\frac{1}{n^2} - 2 \cdot \frac{(x_k - \bar{x}) \cdot (\bar{x} - x_0)}{n \cdot S_{xx}} + \frac{(x_k - \bar{x})^2 \cdot (\bar{x} - x_0)^2}{S_{xx}^2} \right) = \\ &= \sigma^2 \cdot \left(\frac{1}{n} - \frac{2 \cdot (\bar{x} - x_0)}{S_{xx}} \cdot \sum_{k=1}^n \frac{(x_k - \bar{x})}{n} + (\bar{x} - x_0)^2 \cdot \frac{\sum_{k=1}^n (x_k - \bar{x})^2}{S_{xx}^2} \right) = \sigma^2 \cdot \left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right). \end{aligned}$$

Z těchto rovností tedy vyplývá, že

$$a + b \cdot x_0 \sim N \left(\alpha + \beta \cdot x_0, \sigma^2 \cdot \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \right) \quad (13.28)$$

Pro intervalový odhad skutečných hodnot využijeme vztahů (13.24) a (13.28). Z těchto vztahů vyplývá

$$\frac{a + b \cdot x_0 - \alpha - \beta \cdot x_0}{\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \cdot \sqrt{\frac{S_r}{n-2}}} \sim t_{n-2} \quad (13.29)$$

Pomocí tohoto vztahu je možno odhadovat chování $\alpha + \beta \cdot x$ v daném bodě x_0 (odhadujeme chování parametrů !).

Na druhou stranu je někdy důležité vyšetřit chování hodnoty Y v konkrétním bodě x_0 . V tomto případě stanovujeme budoucí hodnotu náhodné veličiny Y . Opět platí nám známé vztahy v bodě x_0

$$Y \sim N(\alpha + \beta \cdot x_0, \sigma^2) \quad (13.30)$$

a vztah (13.29). Z nich můžeme stanovit rozdělení následujícího rozdílu

$$Y - a - b \cdot x_0 \sim N\left(0, \sigma^2 \cdot \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right) \text{ nebo ekvivalentně}$$

$$\frac{Y - a - b \cdot x_0}{\sigma \cdot \sqrt{\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim N(0,1) \quad (13.31)$$

Využijeme – li dále vztahu (13.24) získáváme

$$\frac{Y - a - b \cdot x_0}{\sqrt{\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \cdot \sqrt{\frac{S_r}{(n-2)}}} \sim t_{n-2} \quad (13.32)$$

Tento vztah je rozhodující pro stanovení odhadu budoucí hodnoty Y v bodě x_0 .

13.3 Mnohonásobná regrese

V této kapitole uvedeme základní myšlenky řešení případů, kdy vysvětlovaná proměnná Y je vysvětlována pomocí více než jedné proměnné x . Podržíme dále předpoklady obecných lineárních modelů a navíc regresní funkce bude lineární vzhledem k neznámým parametrům.

Zavedeme tedy označení

$$Y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_k \cdot x_k + e, \quad (13.33)$$

$\beta_0, \dots, \beta_k$ jsou neznámé parametry (budeme je odhadovat z dat), e je náhodná chyba se střední hodnotou rovnou nule a rozptylem rovným σ^2 . Parametry i hodnota σ^2 jsou neznámé hodnoty, podobně jako v předchozí kapitole je budeme odhadovat z dat (budou tedy záviset na náhodné veličině Y). Podobně jako v předchozí části bude tedy platit

$$E(Y_i) = \beta_0 + \beta_1 \cdot x_{i1} + \dots + \beta_k \cdot x_{ik} \quad (13.34)$$

Jestliže dále označíme $B_0, \dots, B_k$ jako bodové odhady skutečných parametrů $\beta_0, \dots, \beta_k$, pak můžeme stejně jako v předchozí části použít metodu nejmenších čtverců a hledat tyto odhady jako hodnoty, které minimalizují následující funkci

$$S(\beta_0, \dots, \beta_k) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 \cdot x_{i1} - \dots - \beta_k \cdot x_{ik})^2 \quad (13.35)$$

Označíme-li řešení (13.35) tak, jak jsme dříve uvedli $B_0, \dots, B_k$, dostaneme derivováním vztahu (13.35) soustavu rovnic

$$\begin{aligned} \sum_{i=1}^n (Y_i - B_0 - B_1 \cdot x_{i1} - \dots - B_k \cdot x_{ik}) &= 0 \\ \sum_{i=1}^n x_{i1} \cdot (Y_i - B_0 - B_1 \cdot x_{i1} - \dots - B_k \cdot x_{ik}) &= 0 \\ \sum_{i=1}^n x_{i2} \cdot (Y_i - B_0 - B_1 \cdot x_{i1} - \dots - B_k \cdot x_{ik}) &= 0, \\ &\vdots \\ \sum_{i=1}^n x_{ik} \cdot (Y_i - B_0 - B_1 \cdot x_{i1} - \dots - B_k \cdot x_{ik}) &= 0 \end{aligned} \quad (13.36)$$

která se nazývá jako v jednorozměrném případě soustava normálních rovnic nebo normální soustava. Tuto soustavu můžeme upravit například takto

$$\begin{aligned} \sum_{i=1}^n Y_i &= n \cdot B_0 - B_1 \cdot \sum_{i=1}^n x_{i1} + \dots + B_k \cdot \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} \cdot Y_i &= n \cdot \sum_{i=1}^n x_{i1} \cdot B_0 + B_1 \cdot \sum_{i=1}^n x_{i1}^2 + \dots + B_k \cdot \sum_{i=1}^n x_{i1} \cdot x_{ik} \\ \sum_{i=1}^n x_{i2} \cdot Y_i &= n \cdot \sum_{i=1}^n x_{i2} \cdot B_0 + B_1 \cdot \sum_{i=1}^n x_{i1} \cdot x_{i2} + \dots + B_k \cdot \sum_{i=1}^n x_{i2} \cdot x_{ik} \\ &\vdots \\ \sum_{i=1}^n x_{ik} \cdot Y_i &= n \cdot \sum_{i=1}^n x_{ik} \cdot B_0 + B_1 \cdot \sum_{i=1}^n x_{i1} \cdot x_{ik} + \dots + B_k \cdot \sum_{i=1}^n x_{ik}^2 \end{aligned} \quad (13.37)$$

Dříve než budeme tuto soustavu řešit, zavedeme označení , které v dalším textu budeme dodržovat.

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} B_0 \\ \vdots \\ B_k \end{pmatrix}$$

Model mnohonásobné regrese je tedy možno zapsat pomocí předchozích symbolů takto

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{e} \quad (13.38)$$

Normální rovnice (13.37) se zapíše v maticové formě

$$\mathbf{X}^T \cdot \mathbf{X} \cdot \mathbf{B} = \mathbf{X}^T \cdot \mathbf{Y} \quad (13.39)$$

Symbol $\mathbf{X}^T$ znamená transponovanou matici k matici $\mathbf{X}$. Nalezení řešení (13.39) znamená řešit tuto soustavu klasickými prostředky . Dá se ukázat , že matice $\mathbf{X}^T \cdot \mathbf{X}$ je pozitivně definitní a regulární viz [1] str. .Existuje k ní inverzní matice a je

$$\mathbf{B} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{Y} \quad (13.40)$$

Příklad

V tabulce jsou uvedeny hodnoty lokalit s počtem obyvatel , mírou rozvodovosti a mírou sebevražd. Z daných údajů zjistíme hodnoty odhadů parametrů β_0 , ..., β_k .

Řešení:

Tabulka 13.1

Počet obyvatel v tisících	Rozvodovost na 100 000 obyvatel	Počet sebevražd na 100 000 obyvatel
679,00	30,40	11,60
1 420,00	34,10	16,10
1 349,00	17,20	9,30
296,00	26,80	9,10
6 975,00	29,10	8,40
323,00	18,70	7,70
4 200,00	32,60	11,30
633,00	32,50	8,40

Budeme zjišťovat hodnotu odhadů koeficientů β_0 , ..., β_2 . Náhodnou veličinou Y bude počet sebevražd , proměnnými hodnotami x_1 a x_2 budou postupně počet obyvatel a rozvodovost v regionu. Řešením získáme hodnoty $B_0 = 3,50735$; $B_1 = -0,00024771$ a $B_2 = 0,260947$. Za předpokladů lineární regrese je vztah mezi výše uvedenými hodnotami

vyjádřen : $Y = 3,5073 - 0,0002 \cdot x_1 + 0,2609 \cdot x_2$. Z tohoto vztahu je zřejmé , že proměnná x_1 nehraje v určení hodnoty Y významnější roli. Tento příklad budeme ještě vyšetřovat dále.

V dalším textu uvedeme souhrnně výsledky , kterých lze dosáhnout analýzou výše uvedených tvrzení.

Souhrnné výsledky:

1. Pro odhady $\mathbf{B}$ parametrů $\boldsymbol{\beta}$ platí :

$$1.1. \quad E(\mathbf{B}) = \boldsymbol{\beta}, \quad VAR(\mathbf{B}) = \sigma^2 \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1}. \quad (13.41)$$

$\mathbf{B}$ je vícerozměrné normální rozdělení!

Obdoba tvrzení (13.7) !

2. Podobně jako v jednorozměrném případě nazveme viz např. (13.6)

$$S_r = (\mathbf{Y} - \mathbf{X} \cdot \mathbf{B})^T \cdot (\mathbf{Y} - \mathbf{X} \cdot \mathbf{B}) = \sum_{i=1}^n (Y_i - B_0 - B_1 \cdot x_{i1} - B_2 \cdot x_{i2} - \dots - B_k \cdot x_{ik})^2 \quad (13.42)$$

reziduálním součtem čtverců . Pro tuto hodnotu platí:

$$2.1. \quad S_r = \mathbf{Y}^T \cdot \mathbf{Y} - \mathbf{B}^T \cdot \mathbf{X}^T \cdot \mathbf{Y} \quad (13.43)$$

$$2.2. \quad \frac{S_r}{\sigma^2} \sim \chi^2_{n-(k+1)} \quad (13.44)$$

3. V dalším textu označujeme symbolem $s^2 = \frac{S_r}{n-(k+1)}$ tzv. reziduální rozptyl . Tato náhodná veličina je jak je vidět z předchozího tvrzení nestranným odhadem parametru σ^2

4. Necht' v_{ij} je (i,j) – tý prvek matice $(\mathbf{X}^T \cdot \mathbf{X})^{-1}$. Pak pro každé $i = 1, \dots, k$ má náhodná veličina

$$T_i = \frac{B_i - \beta_i}{\sqrt{s^2 \cdot v_{ii}}} \quad (13.45)$$

rozdělení $t_{n-(k+1)}$.

5. Podle obecných požadavků musí platit $E(Y/x) = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_k \cdot x_k$, bodovým odhadem této hodnoty je $\sum_{i=1}^n B_i \cdot x_i$, kde $x_0 = 1$. Jestliže chceme určit intervalový odhad

výše uvedené hodnoty , musíme určit rozdělení $\sum_{i=1}^n B_i \cdot x_i$. Protože tento výraz můžeme zapsat jako lineární kombinaci náhodných veličin Y_i , které jsou typu normální rozdělení , bude i naše hledané rozdělení typu normálního. Spočtíme tedy střední hodnotu:

$$E\left(\sum_{i=1}^n x_i \cdot B_i\right) = \sum_{i=1}^n x_i \cdot E(B_i) = \sum_{i=1}^n x_i \cdot \beta_i \quad (13.46)$$

Spočtěme dále rozptyl:

$$VAR\left(\sum_{i=1}^n x_i \cdot B_i\right) = \text{cov}\left(\sum_{i=1}^n x_i \cdot B_i, \sum_{j=1}^n x_j \cdot B_j\right) = \mathbf{x}^T \cdot (\mathbb{X}^T \cdot \mathbb{X})^{-1} \cdot \mathbb{X} \cdot \sigma^2 \quad (13.47),$$

kde $\mathbf{x} = \begin{pmatrix} x_0 \\ \vdots \\ x_k \end{pmatrix}$ je vektor vstupních hodnot. Výše uvedený výpočet je uveden v detailech

například v [1]. Jestliže tedy využijeme vztahy (13.46) a (13.47) získáme následující

$$\frac{\sum_{i=1}^n B_i \cdot x_i - \sum_{i=1}^n x_i \cdot \beta_i}{\sigma \cdot \sqrt{\mathbf{x}^T \cdot (\mathbb{X}^T \cdot \mathbb{X})^{-1} \cdot \mathbb{X}}} \sim N(0,1) \quad (13.48)$$

6. Využijeme – li vztahu (13.48) a (13.44) získáváme

$$\frac{\sum_{i=1}^n B_i \cdot x_i - \sum_{i=1}^n x_i \cdot \beta_i}{\sqrt{\frac{S_r}{n-(k+1)}} \cdot \sqrt{\mathbf{x}^T \cdot (\mathbb{X}^T \cdot \mathbb{X})^{-1} \cdot \mathbb{X}}} \sim t_{n-k-1} \quad (13.49)$$

Odtud můžeme získat intervalové odhady pro odhady vzdáleností $\sum_{i=1}^n B_i \cdot x_i - \sum_{i=1}^n x_i \cdot \beta_i$, neboli pro budoucí predikované hodnoty.

7. Podobně jako v případě jednorozměrném se zajímáme také o případ odhadu vzdáleností $Y(x)$ na $\sum_{i=1}^n B_i \cdot x_i$. Pokud provedeme podobnou analýzu jako v předchozím, získáme výsledky velmi podobné jednorozměrnému případu viz (13.32)

$$\frac{Y(x) - \sum_{i=1}^n B_i \cdot x_i}{\sqrt{\frac{S_r}{n-(k+1)}} \cdot \sqrt{1 + \mathbf{x}^T \cdot (\mathbb{X}^T \cdot \mathbb{X})^{-1} \cdot \mathbb{X}}} \sim t_{n-k-1} \quad (13.50),$$

odtud můžeme odvodit intervalové odhady resp využít tento vztah pro konstrukci statistik ve vhodných statistických hypotézách.