Regresní analýza

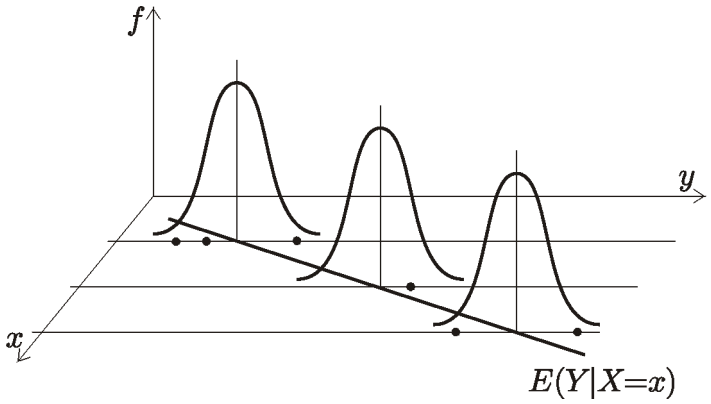
1 Regresní funkce

Důležitou statistickou úlohou je hledání a zkoumání závislostí proměnných, jejichž hodnoty získáme při realizaci experimentů. Vzhledem k jejich náhodnému charakteru reprezentuje nezávisle proměnné náhodný vektor $\mathbf{X} = (X_1, \ldots, X_k)$ a závisle proměnnou náhodná veličina Y. Vektor $\mathbf{X}$ může být i nenáhodný, jak bývá v aplikacích časté, anebo jsou rozptyly všech složek $X_1, \ldots, X_k$ zanedbatelné vůči rozptylu náhodné veličiny Y.

1. Pojmy K popisu a vyšetřování závislosti Y na $\mathbf{X}$ užíváme **regresní analýzu**, přičemž tuto závislost vyjadřuje **regresní funkce**

$$y = \varphi(\mathbf{x}, \beta) = E(Y|\mathbf{X} = \mathbf{x}),$$

kde $\mathbf{x} = (x_1, \ldots, x_k)$ je vektor nezávisle proměnných (hodnota náhodného vektoru $\mathbf{X}$), y je závisle proměnná (hodnota náhodné veličiny Y), $\beta = (\beta_1, \ldots, \beta_m)$ je vektor parametrů, tzv. **regresních koeficientů** β_j, $j = 1, \ldots, m$, a $E(Y|\mathbf{X} = \mathbf{x})$ je podmíněná střední hodnota.



Obrázek 1: Závislost Y na $\mathbf{X}$ pro $k = 1$

2. Poznámka Při vyšetřování závislosti Y na $\mathbf{X}$ získáme realizací n experimentů $(k+1)$-rozměrný statistický soubor $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$ s rozsahem n, kde y_i je pozorovaná hodnota náhodné veličiny Y_i a $\mathbf{x}_i$ je pozorovaná hodnota vektoru nezávisle proměnných $\mathbf{X}$, $i = 1, \ldots, n$. Na Obrázku 1 je znázorněn případ pro $k = 1$, tedy pro $\mathbf{x} = x_1 = x$ (místo x_1 stačí psát x), a s opakovanými pozorováními. Opakování pozorování pro danou hodnotu nezávisle proměnné $\mathbf{x}$ však v regresní analýze není nezbytné.

3. Pojmy Pro určení odhadů neznámých regresních koeficientů β_j minimalizujeme tzv. **reziduální součet čtverců**

$$S^* = \sum_{i=1}^{n} [y_i - \varphi(\mathbf{x}_i, \beta)]^2$$

a hovoříme o tzv. **metodě nejmenších čtverců**.

Pro aplikaci regresní analýzy je nezbytné znát tvar (předpis) regresní funkce. Obvykle jej volíme tak, aby co nejvíce odpovídal vyšetřované nebo uvažované závislosti. Bývá zvykem volit regresní funkci s co nejmenším počtem regresních koeficientů, avšak dostatečně flexibilní a s požadovanými vlastnostmi: monotonie, předepsané hodnoty, asymptoty aj. Vychází se přitom povětšinou ze zkušenosti, avšak v současné době se při realizaci regresní analýzy na PC dají často úspěšně použít vhodné databáze regresních funkcí.

2 Lineární regresní funkce

4. Pojmy **Lineární regresní funkce** (lineární vzhledem k regresním koeficientům) má tvar

$$y = \sum_{j=1}^{m} \beta_j f_j(\mathbf{x}),$$

kde $f_j(\mathbf{x})$ jsou známé funkce neobsahující regresní koeficienty $\beta_1, \ldots, \beta_m$.

5. Poznámka Při lineární regresní analýze, kdy hledáme lineární regresní funkci, aplikujeme tzv. *lineární regresní model* založený na předpokladech:

1. Vektor $\mathbf{x}$ je nenáhodný, takže funkce nabývají nenáhodných hodnot $f_{ji} = f_j(\mathbf{x}_i)$ pro $j = 1, \ldots, m$ a $i = 1, \ldots, n$.
2. Matice $\mathbf{F} = \begin{pmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mn} \end{pmatrix}$ typu (m, n) s prvky f_{ji} má hodnost $m < n$.
3. Náhodná veličina má střední hodnotu $E(Y_i) = \sum_{j=1}^{m} \beta_j f_{ji}$ a konstantní rozptyl $D(Y_i) = \sigma^2 > 0$ pro $i = 1, \ldots, n$.
4. Náhodné veličiny Y_i jsou nekorelované a mají normální rozdělení pravděpodobnosti pro $i = 1, \ldots, n$.

6. Poznámka V části literatury se místo popsaného lineárního regresního modelu také uvádí ekvivalentní lineární model ve tvaru

$$Y_i = \sum_{j=1}^{m} \beta_j f_j(\mathbf{x}_i) + E_i, i = 1, \ldots, n,$$

kde E_i jsou nekorelované náhodné veličiny (vyjadřující např. náhodné chyby měření) s normálním rozdělením pravděpodobnosti $\mathrm{N}(0, \sigma^2)$.

Odhady regresních koeficientů, rozptylu a funkčních hodnot, a také testy statistických hypotéz o regresních koeficientech provádíme pomocí následujících vztahů. V nich jsou použita označení matic:

$$\mathbf{H} = \mathbf{F}\mathbf{F}^T = \begin{pmatrix} \sum_{i=1}^{n} f_{1i} f_{1i} & \cdots & \sum_{i=1}^{n} f_{1i} f_{mi} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} f_{mi} f_{1i} & \cdots & \sum_{i=1}^{n} f_{mi} f_{mi} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{g} = \mathbf{F}\mathbf{y} = \begin{pmatrix} \sum_{i=1}^{n} f_{1i} y_i \\ \vdots \\ \sum_{i=1}^{n} f_{mi} y_i \end{pmatrix},$$

kde $\mathbf{F}^T$ značí transponovanou matici.

7. Vlastnosti Platí:

1. **Bodový odhad regresního koeficientu** je b_j, $j = 1, \ldots, m$, kde matice $\mathbf{b}$ je řešení soustavy lineárních algebraických rovnic (tzv. *soustavy normálních rovnic*)

$$\mathbf{Hb} = \mathbf{g}.$$

2. **Bodový odhad lineární regresní funkce** je

$$y = \sum_{j=1}^{m} b_j f_j(\mathbf{x}).$$

3. **Bodový odhad rozptylu** σ^2 je

$$s^2 = \frac{S^*_{\min}}{n-m},$$

kde $S^*_{\min} = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} b_j f_{ji} \right)^2 = \sum_{i=1}^{n} y_i^2 - \sum_{j=1}^{m} b_j g_j$ je minimální hodnota reziduálního součtu čtverců a g_j je prvek matice $\mathbf{g}$.

4. **Intervalový odhad regresního koeficientu** β_j se spolehlivostí $1-\alpha$, je

$$\left\langle b_j - t_{1-\alpha/2} s\sqrt{h^{jj}}; b_j + t_{1-\alpha/2} s\sqrt{h^{jj}} \right\rangle,$$

kde h^{jj} je j-tý diagonální prvek matice $\mathbf{H}^{-1}$ a $t_{1-\alpha/2}$ je $\left(1-\frac{\alpha}{2}\right)$-kvantil Studentova rozdělení s $n-m$ stupni volnosti - viz tabulku **T2**.

5. **Intervalový odhad střední funkční hodnoty regresní funkce** y pro libovolné pevné $\mathbf{x}$ se spolehlivostí je

$$\left\langle \sum_{j=1}^{m} b_j f_j(\mathbf{x}) - t_{1-\alpha/2} s\sqrt{h^*}; \sum_{j=1}^{m} b_j f_j(\mathbf{x}) + t_{1-\alpha/2} s\sqrt{h^*} \right\rangle,$$

kde $h^* = \mathbf{f}(\mathbf{x})^T \mathbf{H}^{-1} \mathbf{f}(\mathbf{x})$, přičemž $\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix}$ a $t_{1-\alpha/2}$ je $\left(1-\frac{\alpha}{2}\right)$-kvantil Studentova rozdělení s $n-m$ stupni volnosti - viz tabulku **T2**. *Intervalový odhad individuální funkční hodnoty regresní funkce* y pro libovolné pevné $\mathbf{x}$ se spolehlivostí $1-\alpha$ obdržíme analogicky, avšak místo h^* vezmeme $1+h^*$.

6. **Test hypotézy** $H: \beta_j = \beta_{j0}$ proti alternativní hypotéze $\bar{H}: \beta_j \neq \beta_{j0}$ na hladině významnosti α, kde j je jeden pevně zvolený index,$j = 1, \ldots, m$, provádíme pomocí pozorované hodnoty testového kritéria

$$t = \frac{b_j - \beta_{j0}}{s\sqrt{h^{jj}}},$$

$\overline{W}_\alpha = \langle -t_{1-\alpha/2}; t_{1-\alpha/2} \rangle$ a $t_{1-\alpha/2}$ je $\left(1-\frac{\alpha}{2}\right)$-kvantil Studentova rozdělení s $n-m$ stupni volnosti - viz tabulku **T2**. Tento test je možno také provést pomocí výše uvedeného intervalového odhadu koeficientu β_j se spolehlivostí $1-\alpha$.

8. Poznámka Z intervalových odhadů střední funkční hodnoty, resp. individuální funkční hodnoty, se konstruuje **pás spolehlivosti pro střední hodnotu** (viz užší pás kolem regresní přímky na obr. 2), resp. **pás spolehlivosti pro individuální hodnotu** (viz širší pás kolem regresní přímky na obr. 2). Test hypotézy se týká jen jednoho (i když libovolného) regresního koeficientu. Současný test více regresních koeficientů je nutno provést pomocí tzv. **sdružené hypotézy**.

9. Poznámka Orientační mírou vhodnosti vypočtené regresní funkce pro získaná data je **koeficient vícenásobné korelace**

$$r = \sqrt{1 - \frac{S^*_{\min}}{\sum y_i^2 - n\left(\bar{y}\right)^2}},$$

resp. **index** (**koeficient**) **determinace** r^2, které nabývají hodnot z intervalu $\langle 0; 1\rangle$. Číslo $r^2 100$ % vyjadřuje (dle často užívané konvence) procentuální podíl z rozptylu hodnot y_i „vysvětlený“ vypočtenou regresní funkcí. Hodnoty r (a tím také r^2) blízké 1 naznačují vhodnost zvoleného tvaru regresní funkce. Pro bližší posouzení vhodnosti vypočtené regresní funkce se provádí její grafický rozbor vzhledem k pozorovaným bodům $[\mathbf{x}_1, y_1], \ldots, [\mathbf{x}_n, y_n]$. Pro rigorózní závěr je však nutné provést tzv. **regresní diagnostiku** a testovat další statistické hypotézy.

Regresní funkce rozdělujeme na *lineární* a *nelineární* (vzhledem k regresním koeficientům). Některé nelineární regresní funkce můžeme vhodnou linearizací převést na lineární (např. mocninnou nebo exponenciální funkci logaritmujeme). Jde sice o běžně používaný postup, kdy však řešíme jiný regresní model nežli původně uvažovaný.

10. Poznámka Nejvíce užívanou lineární regresní funkcí pro pozorovaný dvourozměrný statistický soubor $(x_1, y_1), \ldots, (x_n, y_n)$ je funkce

$$y = \beta_1 + \beta_2 x,$$

jejímž grafem je tzv. **regresní přímka**. Pro tuto funkci je $\mathbf{x} = x_1 = x$ (místo x_1 píŮeme x), $m = 2$, $f_1(x) = 1$, $f_2(x) = x$, takže

$$\mathbf{F} = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \quad , \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Při ručním výpočtu lze pro regresní funkci použít následující **explicitní vztahy**, kde pro jednoduchost $\sum$ značí $\sum_{i=1}^{n}$.

11. Vlastnosti Platí:

1. $\mathbf{H} = \begin{pmatrix} \sum 1 & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \quad \mathbf{g} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}, \quad \sum 1 = n,$
2. $\det \mathbf{H} = n \sum x_i^2 - \left(\sum x_i\right)^2, \quad b_2 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\det \mathbf{H}} \quad , \quad b_1 = \bar{y} - b_2 \bar{x},$
3. $S^*_{\min} = \sum \left(y_i - b_1 - b_2 x_i\right)^2 = \sum y_i^2 - b_1 \sum y_i - b_2 \sum x_i y_i, \quad s^2 = \frac{S^*_{\min}}{n-2},$
4. $h^{11} = \frac{\sum x_i^2}{\det \mathbf{H}}, \quad h^{22} = \frac{n}{\det \mathbf{H}},$
5. $h^* = \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x_i^2 - n(\bar{x})^2} = \frac{1}{n} + \frac{n(x-\bar{x})^2}{\det \mathbf{H}},$
6. $r = |\, r(x, y)\,|$, kde $r(x, y)$ je koeficient korelace (viz kapitolu Popisná statistika).

12. Příklad U osmi náhodně vybraných firem poskytujících konzultace v oblasti jakosti výroby byly v roce 1993 zjištěny počty zaměstnanců x a roční obraty y (mil. Kč) jak je uvedeno v Tabulce 1:

x_i	3	5	5	8	9	11	12	15
y_i	0,8	1,2	1,5	1,9	1,8	2,4	2,5	3,1

Tabulka 1: Počty zaměstnanců a roční obraty

Vyjádřete závislost ročního obratu firmy na počtu zaměstnanců ve tvaru $y = \beta_1 + \beta_2 x$, vypočtěte intervalový odhad β_2 se spolehlivostí 0,95, testujte na hladině významnosti 0,05 hypotézu $H: \beta_1 = 0, 2,$

určete bodový a intervalový odhad $y(10)$ se spolehlivostí 0,95. Pomocí grafu a koeficientu korelace r posuďte vhodnost regresní funkce. Předpokládejte, že roční obrat má podmíněné normální rozdělení s konstantním rozptylem vzhledem k počtu zaměstnanců.

Řešení V následující Tabulce 2 jsou pomocné výpočty:

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	3	0,8	9	2,4	0,64
2	5	1,2	25	6,0	1,44
3	5	1,5	25	7,5	2,25
4	8	1,9	64	15,2	3,61
5	9	1,8	81	16,2	3,24
6	11	2,4	121	26,4	5,76
7	12	2,5	144	30,0	6,25
8	15	3,1	225	46,5	9,61
Σ	68	15,2	694	150,2	32,80

Tabulka 2: Pomocné výpočty

Vlastní výpočty provedeme v následujících krocích:

(1) Jde o regresní přímku, takže s využitím výše uvedených vzorců obdržíme pro $n = 8$ z tabulky matici $\mathbf{H} = \begin{pmatrix} 8 & 68 \\ 68 & 694 \end{pmatrix}$, jejíž determinant je $\det \mathbf{H} = 8 \cdot 694 - 68^2 = 928$, takže bodový odhad je

$$b_2 = \frac{8 \cdot 150,2 - 68 \cdot 15,2}{928} = 0,1810344 \doteq 0,181.$$

Dále je $\bar{x} = 68/8 = 8,5$, $\bar{y} = 15,2/8 = 1,9$, takže bodový odhad β_1 je

$$b_1 = 1,9 - 0,1810344 \cdot 8,5 = 0,3612068 \doteq 0,361.$$

Potom bodový odhad regresní funkce je $y = 0,361 + 0,181x$.

(2) Minimální hodnota reziduálního součtu čtverců je

$$S^*_{\min} = 32,80 - 0,3612068 \cdot 15,2 - 0,1810344 \cdot 150,2 \doteq 0{,}1182758$$

a bodový odhad rozptylu σ^2, resp. směrodatné odchylky σ , je

$s^2 = 0,1182758/(8-2) = 0,0197126$, resp. $s = \sqrt{0,0197126} \doteq 0,1404017$.

(3) Diagonální prvky matice jsou

$$h^{11} = 694/928 \doteq 0,7478448, \quad h^{22} = 8/928 \doteq 0,00862069.$$

Z tabulky **T2** je pro $8 - 2 = 6$ stupňů volnosti $t_{0,975} = 2,447$, takže intervalový odhad regresního koeficientu β_2 je

$$\beta_2 \in \Big\langle 0,1810344 - 2,447 \cdot 0,1404017\sqrt{0,00862069};$$

$$0,1810344 + 2,447 \cdot 0,1404017\sqrt{0,00862069}\Big\rangle =$$

$$= \langle 0,1491353; 0,2129334\rangle \doteq \langle 0,149; 0,213\rangle .$$

Bodový odhad přírůstku ročního obratu odpovídajícího zvýšení stávajícího počtu zaměstnanců firmy o jednoho zaměstnance je tedy 181 000 Kč a intervalový odhad tohoto přírůstku se spolehlivostí 0,95 je 149 000 Kč až 213 000 Kč.

(4) Pozorovaná hodnota testového kritéria pro je

$$t = \frac{0,3612068 \ - 0,2}{0,1404017\sqrt{0,7478448}} \doteq 1,3277.$$

Pro alternativní hypotézu $\bar{H}: \ \beta_1 \neq 0,2$ je $\overline{W}_{0,05} = \langle -2,447; 2,447\rangle$. Vzhledem k tomu, že $t \in \overline{W}_{0,05}$, hypotézu $H: \ \beta_1 = 0,2$ na hladině významnosti 0,05 nezamítáme. Na dané hladině významnosti vlastně

nezamítáme hypotézu, že firma bez zaměstnanců (pracují jen majitelé), neboť $y(0) = \beta_1$, bude mít roční obrat okolo 200 000 Kč.

(5) Bodový odhad střední i individuální hodnoty ročního obratu firmy pro 10 zaměstnanců je

$$y(10) = 0,3612068 + 0,1810344 \cdot 10 = 2,1715508 \doteq 2,172.$$

U dané firmy lze tedy očekávat roční obrat okolo 2 172 000 Kč. Protože

$$h^* = \frac{1}{8} + \frac{8(10 - 8,5)^2}{928} = 0,1443965,$$

je intervalový odhad se spolehlivostí 0,95 střední hodnoty ročního obratu firmy s 10 zaměstnanci

$$y(10) \in \Big\langle 2,1715508 - 2,447 \cdot 0,1404017\sqrt{0,1443965};$$

$$2,1715508 + 2,447 \cdot 0,1404017\sqrt{0,1443965}\Big\rangle =$$

$$= \langle 2,0409985; 2,3021031\rangle \doteq \langle 2,041; 2,302\rangle .$$

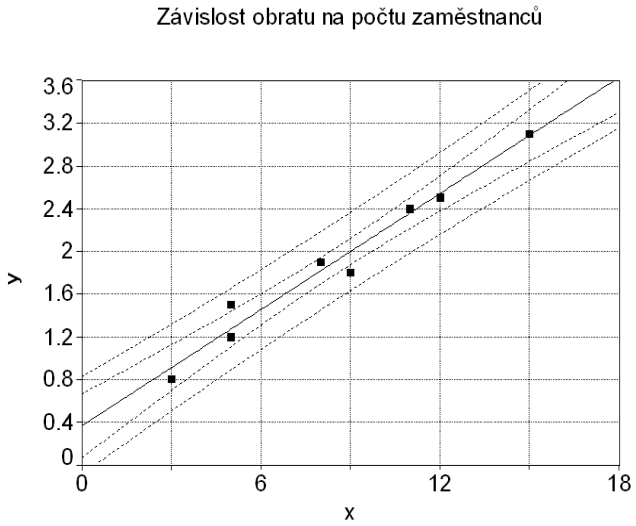
Se spolehlivostí 0,95 lze očekávat, že střední hodnota ročního obratu takové firmy bude od 2 040 000 Kč do 2 302 000 Kč. Jestliže použijeme ve výpočtu místo h^*, dostaneme intervalový odhad se spolehlivostí 0,95 individuální hodnoty ročního obratu firmy s 10 zaměstnanci

$$y(10) \in \Big\langle 2,1715508 - 2,447 \cdot 0,1404017\sqrt{1 + 0,1443965};$$

$$2,1715508 + 2,447 \cdot 0,1404017\sqrt{1 + 0,1443965}\Big\rangle =$$

$$= \langle 1,8040193; 2,5390823\rangle \doteq \langle 1,804; 2,539\rangle .$$

Se spolehlivostí 0,95 lze očekávat, že roční obrat (individuální hodnota ročního obratu) takové firmy bude od 1 804 000 Kč do 2 539 000 Kč, viz Obrázek 2.



Obrázek 2: Graf regresní přímky a pásů spolehlivosti

(6) Koeficient korelace je $r = 0,984798$, takže index determinace je $r^2 = 0,969827$. Z grafu na Obrázku 2 a velikosti koeficientu korelace vidíme, že zvolený tvar regresní funkce vcelku dobře vystihuje danou závislost. Podle často používané konvence lze říci, získaná regresní funkce vyjadřuje celkem r^2 100 % = 96,98 % změn (variability) pozorovaného obratu firmy.