2.4.1 Regresní závislost

V matematice vyjadřujeme závislost hodnot jedné proměnné na hodnotách druhé proměnné funkčním vztahem. V praktických úlohách je však situace složitější. Při měření hodnot sledované veličiny, při jejíž realizaci působí řada dalších (náhodných) vlivů, dostáváme soubor naměřených hodnot, které vykazují často jisté odchylky proti hodnotám, které bychom očekávali z teoretického rozboru sledovaného jevu nebo z jakési očekávané pravidelnosti.

Příklad 2.4.1 *Při soustružení vzniká v místě obrábění na nástroji teplota, závislá na rychlosti posuvu nástroje. Mezi teplotou θ měřenou ve stupních Celsia a rychlostí posuvu v v metrech za minutu byl odvozen teoretický vztah $\theta = \alpha v^{\beta}$, kde α a β jsou konstanty, závisející na dalších podmínkách experimentu. Hodnoty, které byly naměřeny při laboratorním měření, však tomuto vztahu odpovídají jen velmi přibližně, jak lze vidět z grafu.*

Předpokládejme, že sledovanou náhodnou veličinu Y lze vyjádřit jako funkci (zpravidla nenáhodných) veličin $X_1, \ldots, X_r$ a náhodné odchylky ϵ jako

$$Y = f(X_1, \ldots, X_r; \theta_1, \ldots, \theta_s) + \epsilon.$$

Funkce f se nazývá **regresní funkce** a $\theta_1, \ldots, \theta_s$ nazýváme **parametry regrese**. O náhodné veličině ϵ, která se často nazývá neprávem „chybou“, předpokládáme, že má symetrické rozdělení se střední hodnotou 0 a rozptylem σ^2. Obvyklý je předpoklad normálního rozdělení $N(0, \sigma^2)$. Uvedený vztah se nazývá **regresní model**. Podle druhu závislosti regresní funkce na neznámých parametrech $\theta_1, \ldots, \theta_s$ potom hovoříme buď o *lineárním regresním modelu* nebo o *nelineárním regresním modelu*. Nadále se budeme zabývat pouze lineárním modelelm.

Střední hodnota $E(Y)$ je potom funkcí hodnot veličin $X_1, \ldots, X_r$ a neznámých parametrů $\theta_1, \ldots, \theta_s$. Tuto vlastnost vyjádříme vztahem

$$E(Y) = f(x_1, \ldots, x_r; \theta_1, \ldots, \theta_s),$$

kde $x_1, \ldots, x_r$ jsou naměřené hodnoty veličin $X_1, \ldots, X_r$ a $\theta_1, \ldots, \theta_s$ jsou parametry.

Náhodné veličině Y se říká **vysvětlovaná proměnná**, veličinám $X_1, \ldots,$ X_r budeme říkat **vysvětlující proměnné**. Podle tvaru regresní funkce budeme mluvit o **přímkové, exponenciální, kvadratické, polynomické** a jiných regresích. V případě přímkové regrese rozlišujeme podle počtu vysvětlujících proměnných tzv. **jednoduchou regresi** s jednou vysvětlující proměnnou a **vícenásobnou regresi** s více vysvětlujícími proměnnými.

V zásadě zde máme dva problémy: určit tvar (typ) regresní funkce a Při vyšetřování regresní závislosti je regresní funkce zpravidla známa (vyplývá z teoretických vztahů) nebo se její tvar odhaduje (opticky, například podle X-Y grafu rozptýlenosti). Proto se v dalším textu omezíme na úlohu odhadu regresních parametrů předpokládané regresní funkce. K tomu nejčastěji používáme tzv. **metodu nejmenších čtverců**. Tato metoda spočívá v provedení n nezávislých měření hodnot veličin $Y, X_1, \ldots, X_r$ a v nalezení hodnot $\hat{\theta}_1, \ldots, \hat{\theta}_s$, při nichž funkce

$$S(\theta_1, \ldots, \theta_s) = \sum_{i=1}^{n} \left[y_i - f(x_{1_i}, \ldots, x_{r_i}; \theta_1, \ldots, \theta_s) \right]^2$$

nabývá svého minima. Vektory $y_i, x_{1_i}, \ldots, x_{r_i}$ označují i-té pozorování vektoru $Y, X_1, \ldots, X_r, i = 1, \ldots, n$.

Nejsme-li si jisti a rozhodujeme-li se mezi několika modely, potom zpravidla volíme ten, v němž je hodnota funkce $S(\theta_1, \ldots, \theta_s)$ – takzvaný **reziduální součet čtverců** – nejmenší.

V případě lineární regresní funkce $f(x, \alpha, \beta) = \alpha + \beta x$ budeme minimalizovat funkci $S(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$. Nutnou podmínkou pro extrém funkce dvou proměnných je nulovost obou parciálních derivací

$$\frac{\partial S}{\partial \alpha} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial S}{\partial \beta} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = 0,$$

což vede k takzvané **soustavě normálních rovnic**

$$n\alpha + \beta\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\alpha\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

jejímž řešením dostaneme bodové odhady a a b parametrů α a β

$$\begin{aligned} b &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ a &= \overline{y}-\frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\overline{x}=\overline{y}-b\overline{x} \end{aligned}$$

kde $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i, \overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$. Podmínku postačující není třeba vyšetřovat, neboť funkce $S(\alpha,\beta)$ je ryze konvexní.

2.4.2 Jednoduchá přímková regrese

Velmi častým případem regresní závislosti je **přímková regrese**. Předpokládejme regresní vztah $Y = \alpha + \beta X + \epsilon$, kde X je náhodná veličina a ϵ je náhodná veličina s normálním rozdělením $N(0, \sigma^2)$.

Bodové odhady a a b parametrů α a β získáme metodou nejmenších čtverců ve tvaru uvedeném v příkladě VIII.1.6. Naměřené hodnoty $y_1, \ldots, y_n$ lze považovat za hodnoty realizací nezávislých náhodných veličin $Y_1, \ldots, Y_n$ s normálním rozdělením $N(a + bx_i, \sigma^2)$. Z tohoto hlediska jsou bodové odhady a a b odhadovými statistikami, a tedy náhodnými veličinami. Hodnoty $e_i = y_i - a - bx_i, i = 1, \ldots, n$ se nazývají **rezidua** a lze je považovat za odhady hodnot chybového členu ϵ. Číslo $\hat{y}_i = a + bx_i$ je odhadem hodnoty náhodné veličiny Y_i.

Označme $S_R = S(a, b)$ takzvaný **reziduální součet čtverců**

$$S_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i - a - bx_i{}^2 = \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i.$$

Bodový odhad s^2 rozptylu σ^2 chybového členu ϵ je potom dán vztahem $s^2 = \frac{S_R}{(n-2)}$ a nazývá se **reziduální rozptyl**.

Pomocí s^2 lze vyjádřit odhady rozptylu obou regresních parametrů

$$S_a^2 = \frac{s^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}, \quad S_b^2 = \frac{s^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}$$

Statistiky $T_\alpha = \frac{(a-\alpha)}{S_a}$ a $T_\beta = \frac{(b-\beta)}{S_b}$ mají Studentovo t-rozdělení o $(n-2)$ stupních volnosti. Intervalové odhady pro parametry α a β jsou potom dány nerovnostmi

$$b - S_b t_{n-2}(1-\frac{\gamma}{2}) \leq \beta \leq b + S_b t_{n-2}(1-\frac{\gamma}{2})$$

kde $(1-\gamma)$ je koeficient spolehlivosti a $t_{n-2}(1-\frac{\gamma}{2})$ je $(1-\frac{\gamma}{2})$-kvantil t-rozdělení o $(n-2)$ stupních volnosti.

V některých případech nás zajímá, zda hodnota některého z parametrů se liší významně od nulové hodnoty nebo ne a zda jej lze tudíž v regresní funkci vynechat. Oboustranné testy nulovosti regresních koeficientů lze založit na odhadových statistikách $T_a = \frac{a}{S_a}$ resp. $T_b = \frac{b}{S_b}$ a jim odpovídajícím kritickým oborům tak, že při splnění nerovnosti

$$|T_a| \geq t_{n-2}(1 - \frac{\gamma}{2}), \text{ resp. } |T_b| \geq t_{n-2}(1 - \frac{\gamma}{2}),$$

zamítneme hypotézu o nulovosti parametru α, resp. β, na hladině významnosti γ.

Regresním modelem se snažíme vysvětlit změny - variabilitu - vysvětlované veličiny Y pomocí změn vysvětlující veličiny X. Podíl části variability Y vysvětlené modelem ku celkové variabilitě Y, zpravidla vyjádřený v procentech, se nazývá **koeficient determinace** R^2 a je dán vztahy

$$R^2 = \frac{\sum_{i=1}^{n}(a + bx_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \left(1 - \frac{S_R}{\sum_{i=1}^{n}(y_i - \overline{y}^2}\right),$$

kde S_R je reziduální součet čtverců.

K úplné regresní analýze patří i **analýza reziduí**. Především by měly vyhovovat předpokladu normality, za kterého byly všechny předchozí výsledky odvozeny. Pokud tomu tak není, nelze výsledky považovat za důvěryhodné. K ověření shody hodnot reziduí s normálním rozdělením lze použít některý z testů, uvedených v odstavci VII.3, nebo pravděpodobnostní papír, který je popsán v kapitole X. Z analýzy reziduí lze detekovat i takzvaná **odlehlá pozorování**. To znamená ty hodnoty, které byly chybně naměřeny nebo indikují nesrovnalosti v modelu, a jimž je třeba věnovat zvláštní pozornost. Ke zjišťování těchto hodnot lze použít například krabicové gragy, popsané v kapitole X.

Model lineární regrese lze použít i v některých případech, kdy závislost mezi veličinami X a Y není lineární. Jsou to případy, kdy lze provést takzvanou **linearizaci modelu**. Vhodnou transformací převedeme nelineární závislost na lineární a použijeme lineární regresní model. Přitom však musíme být velmi opatrní, neboť vše, co bylo odvozeno pro lineární regresní model za předpokladu normality chybového členu ϵ platí pouze pro "linearizovaný model", nikoli pro model původní, a to opět za předpokladu, že náhodná veličina, odpovídající transformovanému chybovému členu v linearizovaném modelu, má normální rozdělení.

Příklad 2.4.2 *Závislost mezi teplotou θ a rychlostí posuvu v v příkladu 2.4.1. lze považovat za regresní závislost ve tvaru $\theta = \alpha.v^{\beta}.\epsilon$, kde α a β jsou regresní koeficienty a ϵ je náhodná veličina se střední hodnotou 1. Provedeme-li transformaci $Y = ln\theta, X = lnv, a = ln\alpha, e = ln\epsilon ab = \beta$, dostaneme $Y = a + bX + e$, tedy lineární vztah.*

Podobně jako v předchozím příkladu lze linearizovat i jiné modely, např. logaritmický, tj. $Y = ln(\alpha + \beta.X)$, reciproký $Y = \frac{1}{\alpha+\beta.X}$ a další.

Lineární regresí rozumíme regresní model s regresní funkcí, která je lineární kombinací parametrů modelu,

$$Y_i = x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + e_i, \quad i = 1, \ldots, n,$$

tj. zapsáno maticově

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

kde

$$\boldsymbol{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{X}_{n\times k} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix},$$

$$\boldsymbol{\beta}_{k\times 1} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{e}_{n\times 1} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

jsou vektor vysvětlovaných pozorování, regresní matice, vektor parametrů a chybový vektor s $\mathrm{E}\,\boldsymbol{e} = 0$.

jsou vektor vysvětlovaných pozorování, regresní matice, vektor parametrů a chybový vektor s $\mathrm{E}\,\boldsymbol{e} = 0$.

Předpokládáme, že počet parametrů je menší než počet pozorování, $k < n$, a že hodnost matice $\boldsymbol{X}$ je k (tj. $\boldsymbol{X}$ má lineárně nezávislé sloupce — nezavádíme nadbytečné vysvětlující proměnné).

Pokud některý sloupec matice $\boldsymbol{X}$ tvoří jednotkový vektor, hovoříme o modelu **s absolutním členem** (jímž je příslušná složka vektoru β).

Odhadem parametrů regresního modelu

$$Y_i = f(x_{i1}, \ldots, x_{i\ell}; \beta_1, \ldots, \beta_k) + e_i, i = 1, \ldots, n,$$

metodou nejmenších čtverců nazýváme ty hodnoty $\boldsymbol{b} = (b_1, \ldots, b_k)'$ parametrů $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, které minimalizují **součet čtverců**

$$S(\beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (Y_i - f(x_{i1}, \ldots, x_{i\ell}; \beta_1, \ldots, \beta_k))^2.$$

Za předpokladu hladkosti S je tento odhad řešením soustavy **normálních rovnic** $\partial S / \partial \beta_i = 0, i = 1, \ldots, k.$

Hodnoty $\widehat{Y}_i = f(x_{i1}, \ldots, x_{i\ell}; b_1, \ldots, b_k), i = 1, \ldots, n,$ označujeme jako **vyrovnané hodnoty** pozorování Y_i, $i = 1, \ldots, n.$

Reziduálním součtem čtverců S_e rozumíme minimální hodnotu součtu čtverců S při odhadování parametrů regresního modelu pomocí metody nejmenších čtverců,

$$S_e = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})'(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}),$$

kde $\boldsymbol{Y}$ a $\widehat{\boldsymbol{Y}}$ jsou pozorované, resp. vyrovnané hodnoty závislé proměnné $\boldsymbol{Y}$.

S_e udává modelem nevysvětlenou část z původního **celkového součtu čtverců**

$$S_t = (\boldsymbol{Y} - \overline{Y})'(\boldsymbol{Y} - \overline{Y}),$$

kde $\overline{Y}$ je aritmetický průměr složek $\boldsymbol{Y}$.

V modelu lineární regrese $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ má odhad parametrů $\boldsymbol{\beta}$ metodou nejmenších čtverců tvar

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

a reziduální součet čtverců je

$$S_e = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y}.$$

Jsou-li složky vektoru chyb $\boldsymbol{e}$ nekorelované a se stejným rozptylem σ^2, tj. $\mathrm{E}\,\boldsymbol{e} = 0$, $\mathrm{var}\,\boldsymbol{e} = \sigma^2 I$, je

$$\mathrm{E}\,\boldsymbol{b} = \boldsymbol{\beta}, \quad \mathrm{var}\,\boldsymbol{b} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}, \quad \mathrm{E}\,\frac{S_e}{n-k} = \sigma^2,$$

tedy $\boldsymbol{b}$ a $s^2 = S_e/(n-k)$ jsou nestrannými odhady $\boldsymbol{\beta}$ a σ^2. Odhad $\boldsymbol{b}$ je dokonce nejlepší mezi lineárními (ve složkách vektoru $\boldsymbol{Y}$) nestrannými odhady $\boldsymbol{\beta}$.

Koeficientem determinace rozumíme veličinu

$$R^2 = 1 - \frac{S_e}{S_t},$$

kde S_e je reziduální a S_t celkový součet čtverců.

V modelu lineární regrese s absolutním členem leží hodnota R^2 v intervalu $\langle 0, 1 \rangle$ a udává, jaký podíl rozptylu v pozorování závislé proměnné se podařilo regresí vysvětlit (větší hodnoty znamenají větší úspěšnost regrese).

Předpokládejme, že v modelu lineární regrese má chybový vektor $(e_1, \ldots, e_n)$ normální rozdělení $\mathrm{N}_n(0, \sigma^2 I)$.

V modelu s absolutním členem β_1, kde $\beta_2, \ldots, \beta_k$ jsou ostatní parametry, lze pro test hypotézy

$$H_0 \colon \beta_2 = \cdots = \beta_k = 0$$

proti

$$H_1 \colon \beta_j \neq 0 \text{ pro některé } j = 2, \ldots, k$$

použít statistiku

$$F = \frac{R^2}{1 - R^2} \frac{n - k}{k - 1},$$

kde R^2 je koeficient determinace. Hypotézu H_0 zamítneme na hladině významnosti α, když

$$F > F_{1-\alpha}(k - 1, n - k),$$

kde $F_{1-\alpha}(k - 1, n - k)$ je $(1 - \alpha)$-kvantil rozdělení $F_{k-1,n-1}$ (což je rozdělení statistiky F za platnosti H_0).