ZÁVISLOST KVANTITATIVNÍCH VELIČIN

Zabýváme se závislostí dvou náhodných veličin (např. výška a hmotnost dospělého muže). Máme k disposici n realizací náhodných veličin X, Y, tedy $(X_1, Y_1), \ldots, (X_n, Y_n)$. Pokud se omezíme na jejich lineární závislost, máme k disposici dva modely, které však při prokazování závislosti vedou ke stejnému kritickému oboru.

Korelační koeficient.

Předpoklady. $(X_1, Y_1), \ldots, (X_n, Y_n)$ jsou nezávislé realizace náhodného vektoru (X, Y). Náhodný vektor (X, Y) má **dvourozměrné normální rozdělení**. K tomu je nutné (ale nemusí stačit), aby každá z veličin X, Y měla normální rozdělení. Označme

$$X \sim \mathsf{N}(\mu_X, \sigma_X^2), \qquad Y \sim \mathsf{N}(\mu_Y, \sigma_Y^2), \qquad \rho_{XY} = \mathsf{cov}(X, Y)/(\sigma_X \cdot \sigma_Y).$$

Korelační test. $\mathsf{H}_0 : \rho_{XY} = 0, \qquad \mathsf{H}_1 : \rho_{XY} \neq 0$
Odhadem ρ_{XY} je **(výběrový) korelační koeficient**:

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}},$$

kde je

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}), \quad S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Na hladině α nulovou hypotézu zamítáme, když platí některé z ekvivalentních tvrzení

- $|T| \geq t_{n-2}(1 - \alpha/2)$, kde $T = \frac{r_{XY}}{1 - r_{XY}^2} \sqrt{n-2}$;
- $0 \notin I$, kde I je $100(1-\alpha)\%$ interval spolehlivosti pro ρ_{XY};
- p-hodnota je nejvýše 5 %.

Platí $|r_{XY}| \leq 1$, podobně $|\rho_{XY}| \leq 1$. Jsou-li náhodné veličiny X, Y nezávislé, potom platí $\rho_{XY} = 0$. Obrácené tvrzení neplatí, $\rho_{XY} = 0$ ještě nemusí znamenat, že X, Y jsou nezávislé. S náhodnými veličinami X, Y zachází korelační koeficient **symetricky**.

Regresní přímka.

Předpoklady. Pro nezávislé náhodné veličiny $Y_1, \dots, Y_n$ platí $Y_i \sim \mathsf{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ (střední hodnota Y lineárně závisí na x, rozptyl je konstantní), kde $\beta_0, \beta_1, \sigma > 0$ jsou neznámé nenáhodné parametry.

Regresní přímka zachází se dvěma náhodnými veličinami **nesymetricky**. Modelem je **podmíněné rozdělení** náhodné veličiny Y (závisle proměnná) pro danou hodnotu x nezávisle proměnné X. Když má náhodný vektor (X, Y) dvourozměrné normální rozdělení jak předpokládá korelační test, pak podmíněným rozdělením náhodné veličiny Y při daném $X = x$ je normální rozdělení $\mathsf{N}\Big((\mu_Y - \rho_{XY}\frac{\sigma_Y}{\sigma_X}\mu_X) + \rho_{XY}\frac{\sigma_Y}{\sigma_X}x, (1-\rho_{XY}^2)\sigma_Y^2)\Big)$, což při zjednodušeném značení můžeme zapsat jako $\mathsf{N}(\beta_0 + \beta_1 x, \sigma^2)$. Model lze tedy psát jako $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, kde $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$ jsou nezávislé náhodné veličiny (náhodné chyby, které nelze přímo pozorovat). Parametry β_0, β_1 odhadneme **metodou nejmenších čtverců**, tj. minimalizací výrazu

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 ,$$

tedy minimalizací součtu druhých mocnin odchylek bodů (x_i, Y_i) od přímky $y = \beta_0 + \beta_1 x$. Derivováním minimalizovaného výrazu podle β_0 a β_1 a položením derivací rovným nule dostaneme odhady parametrů β_1, β_0

$$b_1 = S_{XY}/S_X^2, \qquad b_0 = \bar{Y} - b_1\bar{x}.$$

Bodový graf (x_i, Y_i) můžeme doplnit **regresní přímkou** $y = b_0 + b_1 x$, na které leží body $(x_i, \hat{Y}_i)$, kde $\hat{Y}_i = b_0 + b_1 x_i$ jsou **odhady středních hodnot** Y_i. Rozdíly $e_i = Y_i - \hat{Y}_i$ se nazývají **rezidua**.

Test nezávislosti. Střední hodnota $\mathsf{E}\,Y = \beta_0 + \beta_1 x$ nezávisí na x právě tehdy, když je $\beta_1 = 0$. Zvolíme tedy $\mathsf{H}_0 : \beta_1 = 0$ a $\mathsf{H}_1 : \beta_1 \neq 0$.
Nulovou hypotézu na hladině α zamítáme, když platí některé z ekvivalentních tvrzení

- $|T| \geq t_{n-2}(1-\alpha/2)$, kde $T = \frac{b_1}{S}\sqrt{\sum(x_i - \bar{x})^2}$ a $S = \sqrt{\frac{1}{n-2}\sum_{i=1}^n e_i^2}$;
- $0 \notin I$, kde I je $100(1-\alpha)\%$ interval spolehlivosti pro β_1;
- p-hodnota je nejvýše 5 %.

Protože lze ukázat, že v modelu regresní přímky je statistika T korelačního testu stejná jako v testu nezávislosti, jsou oba dosud popsané kritické obory totožné.

Koeficient determinace $R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum(Y_i - \bar{Y})^2}$ nabývá hodnot z intervalu $\langle 0, 1\rangle$ a vyjadřuje jaký díl variability závisle proměnné se podařilo vysvětlit modelem. V případě regresní přímky platí $R^2 = r_{XY}^2$.

Ověření předpokladů se provádí jako analýza reziduí, zpravidla graficky.

- Bodový graf (x_i, e_i) umožňuje ověřit linearitu závislosti střední hodnoty Y na x, případně nezávislost rozptylu Y na x.
- Bodový graf (i, e_i) někdy umožní odhalit závislost mezi hodnotami závisle proměnné.
- Normální diagram (qq-plot) reziduí e_i slouží k ověření normality.

Mnohonásobná lineární regrese.

Přepoklady. Pro nezávislé náhodné veličiny $Y_1, \dots, Y_n$ platí

$$Y_i \sim \mathsf{N}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \sigma^2),$$

kde $\beta_0, \beta_1, \dots, \beta_k, \sigma > 0$ jsou neznámé konstanty. Model umožňuje vyšetřit závislost **závisle proměnné** Y (náhodná veličina) na několika (zpravidla spojitých) nenáhodných **nezávisle proměnných** $x_1, \dots, x_k$.

Model lze psát také jako $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$, kde $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$ jsou opět nezávislé náhodné veličiny (náhodné chyby, které nelze přímo pozorovat). Parametry $\beta_0, \beta_1, \dots, \beta_k$ odhadneme **metodou nejmenších čtverců**. Řešení lze vyjádřit maticově. Označme

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

Odhadem je vektor

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Dobrý statistický program zpravidla poskytne pro každý parametr β_j statistiku T_j s odpovídající p-hodnotou pro test hypotézy $\beta_j = 0$, která vyjadřuje možnost vypustit z modelu j-tou nezávisle proměnnou. Podobně poskytne **odhady středních hodnot** závisle proměnné

$$\hat{Y}_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}$$

a **rezidua** $e_i = Y_i - \hat{Y}_i$.

Koeficient determinace se počítá stejně jako pro regresní přímku, má také stejnou interpretaci, jen není vázán s korelačními koeficienty $r_{Yx_1}, \dots, r_{Yx_k}$.

Ověření předpokladů se provádí jako analýza reziduí, zpravidla graficky.

- Bodový graf $(\hat{Y}_i, e_i)$ umožňuje ověřit linearitu závislosti střední hodnoty Y na lineární funkci nezávisle proměnných, případně nezávislost rozptylu na zmíněné lineární funkci.
- Bodový graf (i, e_i) někdy umožní odhalit závislost mezi hodnotami závisle proměnné.
- Normální diagram (qq-plot) reziduí slouží k ověření normality.