Prostá regresní a korelační analýza [1]

[1]Tyto materiály byly vytvořeny za pomoci grantu FRVŠ číslo 1145/2004.

Problematika závislosti

V podstatě lze rozlišovat mezi závislostí nepodstatnou, čili náhodnou a závislosti příčinnou čili kauzální. V případě kauzální závislosti lze odlišit závislost jednostrannou a závislost oboustrannou.

Dle složitosti můžeme rozlišovat jednodušší formy příčinné, tedy kauzální závislosti a složitější formy kauzální závislosti.

Z hlediska statistická teorie pak rozlišujeme dva typy závislostí. Prvním z nich je tzv. statistická závislost. Tu lze popsat následujícím způsobem.Sledujeme-li statistické znaky $y, x_1, x_2, \cdots, x_p$ a mění-li se určitým způsobem podmíněné rozdělení znaku y při změnách $x_1, x_2, \cdots, x_p$, pak mluvíme o **statistické závislosti** znaku y na $x_1, x_2, \cdots, x_p$. Speciálním typem této statistické závislosti je tzv. **korelační závislost**, při které se mění podmíněné střední hodnoty znaku y.

Cíle regresní a korelační analýzy

Cíle regresní a korelační analýzy lze spatřovat ve dvou hlavních bodech:

- ve vystižení směru korelační závislosti. Tím odpovídáme na otázku, jak se změní závisle proměnná, jestliže změníme nezávisle proměnnou o jednotku. Směr korelační závislosti vyjadřujeme pomocí regresní čáry. Ta je spojnicí vyrovnaných hodnot závisle proměnné, odpovídajícím hodnotám nezávisle proměnné. Statistické metody, které řeší tento úkol, shrnujeme pod společný název regresní analýza.
- v posouzení toho, do jaké míry jsou pozorované hodnoty v blízkém okolí regresní čáry, či zda se pozorované hodnoty od regresní čáry značně vzdalují. Čím jsou pozorované hodnoty blíže k regresní čáře, tím daná regresní čára poskytuje hodnotnější odhad, a naopak, čím se pozorované hodnoty více odchylují od regresní čáry, tím je mezi proměnnými menší statistická závislost. Odhady pořízené na základě takovéto regresní čáry jsou pak méně hodnotné. Shrneme-li výše uvedené, lze říci, že dalším úkolem korelační a regresní analýzy je posouzení těsnosti korelační závislosti. Podstatou je tedy posouzení variability pozorovaných hodnot kolem regresní čáry. Tento problém řeší korelační analýza.

Závisle proměnná $\times$ nezávisle proměnná

Znak y nazýváme vysvětlovanou nebo závisle proměnnou, znaky $x_1, x_2, \cdots, x_p$ vysvětlujícími nebo nezávisle proměnnými.

Prostá lineární regrese

Nejjednodušším modelem se kterým se lze v regresní analýze setkat je prostá lineární regrese. Jejím parametrickým vyjádřením je funkce $y = \alpha + \beta x$, tedy rovnice přímky. Zabývejme se tímto modelem blíže.

Při regresní analýze se snažíme najít neznámé parametry α a β tak, aby výsledný (odhadnutý) model $\hat{y} = a + bx$ co nejlépe vystihoval námi pozorované data. K odhadu neznámých parametrů, tzv. regresních koeficientů zpravidla používáme metodu nejmenších čtverců. Podstatou této metody je minimalizace součtu čtverců reziduí, tedy:

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow \text{ min } . \tag{1}$$

Předpoklady modelu

- Střední hodnota reziduí je nulová. Nebo-li $E(\epsilon_i) = 0$
- Rozptyl reziduí je konstantní pro všechny pozorování, tedy $Var(\epsilon_i) = \sigma^2$
- Rezidua sledují normální rozdělení $\epsilon_i \sim N(0, \sigma^2)$
- Jednotlivé pozorování závislé proměnné y_i jsou navzájem nezávislé. V důsledku toho pak i jednotlivé ϵ_i
- Jednotlivé úrovně -hodnoty regresorů jsou pevné, pokud jsou náhodné, pak jsou navzájem nezávislé.
- Funkce je lineární kombinací regresních koeficientů.

Odhady a a b neznámých regresních koeficientů α a β jsou tedy určeny z podmínky:

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow \text{ min } , \tag{2}$$

kde po dosazení za $\hat{y}_i$ získáme v případě fitování modelu prosté lineární regrese:

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2 \rightarrow \text{ min.} \tag{3}$$

Nalézt minimum této kvadratické funkce znamená položit parciální derivace podle a a b nule a řešit vzniklou soustavu rovnic. Po několika úpravách získáme tzv. normální rovnice:

$$\begin{array}{rcl} na + b\sum_{i=1}^{n} x_i & = & \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 & = & \sum_{i=1}^{n} x_i y_i \ . \end{array} \tag{4}$$

Řešením soustavy rovnic lze získat vzorce pro výpočet regresních koeficientů:

$$a = \frac{\sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2}{(\sum_{i=1}^{n} x)^2 - n\sum_{i=1}^{n} x_i^2} \tag{5}$$

a

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \ . \tag{6}$$

Interpretace regresního koeficientu plyne především z toho, že odhadnutý regresní koeficient b je formálně směrnicí regresní přímky. Udává tedy, jak velká bude změna závislé y, změní-li se nezávisle proměnná x o jednotku. Kladná hodnota regresního koeficentu ($b > 0$) znamená, že s růstem nezávisle proměnné poroste i hodnota závisle proměnné. Záporná hodnota pak vyjadřuje pokles závisle proměnné při růstu hodnot nezávisle proměnné.

Volba regresní funkce

Při volbě regresní funkce je nutné znát její základní vlastnosti, tj. znát jednotlivé funkce, jejich analytické vyjádření, jejich průběh, definiční obor a obor hodnot.

V prvé řadě má regresní model co nejlépe zobrazit reálné vztahy mezi jevy a odrážet je v jejich podstatných rysech. Z tohoto důvodu je třeba vycházet z posouzení věcné podstaty zkoumaných jevů a jejich souvislostí.

V mnoha případech však není možno volit regresní funkci apriorně. Pak volíme regresní funkci na základě posouzení závislosti v pozorovaných datech. Tento přístup však nemusí vést k nalezení regresní funkce (problém malého počtu pozorování) vhodné pro popis závislosti v základním souboru.

Pro empirické posouzení závislosti je možno použít bodový diagram nebo čáru podmíněných průměrů. Obvykle se však postupuje takto:

- Vymezíme množinu regresních funkcí - pokud možno jednoduchých
- Určíme odhady jednotlivých regresních parametrů pro jednotlivé typy regresních funkcí
- Na základě různých kritérií zkoumáme která z regresních funkcí nejlépe vyhovuje empirickým datům.

Regresní modely

Dle tvaru regresní funkce lze rozlišovat různé typy regresních modelů:

- Modely lineární z hlediska parametrů mají regresní funkci tvaru:

$$\hat{y} = \beta_0 + \beta_1 f_1 + \beta_2 f_2 + \cdots + \beta_p f_p \ , \tag{7}$$

  kde regersory f jsou libovolné známé funkce vysvětlujících proměnných. Speciálním případem jsou modely typu:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \ , \tag{8}$$

  kde regresory jsou přímo vysvětlující proměnné, tj. modely lineární z hlediska parametrů i z hlediska vysvětlujících proměnných. Příkladem můžou být tyto modely:

$$\hat{y} = \beta_0 + \beta_1 x \tag{9}$$

  nebo hyperbolický regresní model

$$\hat{y} = \beta_0 + \beta_1 \frac{1}{x} \tag{10}$$

$$\hat{y} = \beta_0 + \beta_1 \log_{10} x \tag{11}$$

$$\hat{y} = \beta_0 + \beta_1 \log_e x \tag{12}$$

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{13}$$

  a samozřejmě mnohé další ...

- Modely nelineární jak v parametrech, tak vzhledem k vysvětlujícím proměnným, které se však transformací dají převést na lineární tvar z hlediska regresních parametrů. Příkladem může být mocninná funkce

$$\hat{y} = \alpha x^{\beta} \tag{14}$$

nebo exponenciální funkce

$$\hat{y} = \alpha \beta^{x} \tag{15}$$

- Nelineární modely které se nedají jednoduše transformovat na lineární tvar např:

$$\hat{y} = \alpha \beta^{x} + \gamma \tag{16}$$

Otázka vhodnosti modelu

Jedním ze základních kritérií pro posouzení kvality regresní funkce je tzv. součet čtverců reziduí, definovaný jako

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{17}$$

Na základě tohoto kritéria dáváme přednost tomu regresnímu modelu pro nějž nabývá tato statistika nižší hodnoty.

V případě, že porovnáváme regresní modely s různým počtem regresních parametrů, musíme si uvědomit, že u regresní funkce s větším počtem parametrů bude reziduální součet čtverců nižší než u regresní funkce s menším počtem regresních parametrů. Z tohoto důvodu využíváme pro srovnání tzv. reziduální rozptyl definovaný jako

$$s_e^2 = \frac{S}{n - p} \tag{18}$$

Index determinace

Další velice důležitou charakteristikou vhodnosti regresní funkce je tzv. index determinace. Jeho konstrukce vychází z rozkladu součtu čtvercových odchylek hodnot vysvětlované proměnné od jejich aritmetického průměru

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{19}$$

na dvě složky. A to na součet čtverců reziduí:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{20}$$

a na součet čtvercových odchylek teoretických hodnot od aritmetického průměru

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{21}$$

Součet čtvercových odchylek teoretických hodnot od průměru představuje tu část součtu čtverců , kterou je možno vysvětlit zvolenou regresní funkcí. Podíl

$$I^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{22}$$

se nazývá **index determinace**. Tato míra nabývá hodnot z uzavřeného intervalu $\langle 0, 1 \rangle$.
Index determinace nám udává z kolika procent variabilita nezávisle proměnné vysvětluje variabilitu závisle proměnné.

Korelační koeficient

Pro posuzování vhodnosti regresní funkce a těsnosti závislosti vysvětlované proměnné y na uvažovaných vysvětlujících proměnných se používá také druhá odmocnina indexu determinace. Ta se nazývá **index korelace** (**koeficient korelace**). V případě prosté lineární regrese jej lze definovat například takto:

$$r_{yx} = \frac{cov(x, y)}{\sigma_x \sigma_y} \tag{23}$$

Tato statistika vyjadřuje stupeň lineární statistické závislosti. Symbol $cov(x, y)$ v čitateli představuje kovarianci proměnných x a y. Ve jmenovateli pak vystupuje součin směrodatných odchylek nezávisle a závisle proměnné.