Statistika

Regresní a korelační analýza – Úvod do problému

Roman Biskup

Jihočeská univerzita v Českých Budějovicích
Ekonomická fakulta (Zemědělská fakulta)
Katedra aplikované matematiky a informatiky

2008/2009

Obsah

Závislost statistických znaků

Pohledy na typy závislostí

- Dle typu vazby
  - bezprostřední kauzální závislost
  - zprostředkovaná kauzální závislost
  - náhodná souvislost
- Dle typu statistických znaků
  - závislost dvou alternativních statistických znaků – asociační tabulky
  - závislost dvou nominálních statistických znaků – kontingenční tabulky
  - závislost dvou diskrétních (intervalově setříděných) numerických statistických znaků – korelační tabulky
  - závislost spojitého numerického znaku na alternativním/nominálním znaku – rozkladové tabulky
  - závislost alternatiho/nominálního znaku na spojitých numerických i nominálních statistických znacích – diskriminační analýza, ...
  - závislost skupiny spojitých numerických znaků na skupině jiných spojitých numerických znaků – regresní a korelační analýza
  - ...

„Závislost“ statistických znaků (proměnných)

Směr závislosti/souvislosti

- Jednostranná závislost
  - jedna skupina proměnných závisí na jiných
  - nezávislé a závislé proměnné
- Oboustranná závislost
  - dá se předpokládat souvislost proměnných, nedá se však určit, co je příčina a co je následek
  - vysvětlující a vysvětlované proměnné
    - vysvětlující: snadno měřitelné, dopředu známé, ...
    - vysvětlované: hůře měřitelné, měřitelné až následně, ...
  - exogenní a endogenní proměnné
    - exogenní: proměnné mimo systém, vysvětlují chování systému, ...
    - endogenní: proměnné daného systém, popisují chování systému, ...

Závislost statistických znaků (proměnných) I

Logická/věcná síla závislosti

- Pevná (funkční) závislost
  - konkrétním l hodnotám jedné skupiny proměnných $(x_{1i}, x_{2i}, \ldots, x_{li})$ odpovídá právě q-tice hodnot druhé skupiny proměnných $(y_{1i}, y_{2i}, \ldots, y_{qi})$
  - závislost mezi nimi lze vyjádřit beze zbytku funkčním předpisem $(y_{1i}, y_{2i}, \ldots, y_{qi}) = f(x_{1i}, x_{2i}, \ldots, x_{li})$, kde f je vícerozměrná funkce l proměnných – funkční závislost
  - důsledek lze jednoznačně určit jednou, nebo několika málo příčinami – neexistují žádné další neznámé a/nebo náhodné vlivy
- Volná (stochastická) závislost,
  - konkrétním l hodnotám jedné skupiny proměnných $(x_{1i}, x_{2i}, \ldots, x_{li})$ může odpovídat více q-tic hodnot druhé skupiny proměnných $(y_{1i}, y_{2i}, \ldots, y_{qi})$
  - Změny hodnot jedné skupiny proměnných jsou doprovázeny změnami:
    - podmíněných průměrů druhé skupiny proměnných – korelační závislost
    - podmíněného pravděpodobnostního rozdělení druhé skupiny proměnných – statistická závislost

Závislost statistických znaků (proměnných) II

Logická/věcná síla závislosti

- Důsledek je určen velkým počtem příčin, které:
  - nelze přesně (funkčně) postihnout a/nebo
  - všechny nejsou známé a/nebo
  - působí náhodné vlivy

Regresní a korelační analýza

- Regresní analýza:
  - slouží k popisu závislosti dvou a více numerických proměnných – hledáme matematický model – regresní funkci, která by měla:
    - vyjadřovat charakter závislosti a co nejvěrněji zobrazovat průběh změn podmíněných průměrů závisle proměnné/proměnných,
    - vysvětlovat složku hodnoty závisle proměnné/proměnných která je funkcí nezávisle proměnné/proměnných (deterministická složka) – druhá (nevysvětlená) složka je výsledkem dalších (vedlejších a náhodných) vlivů (náhodná složka)
  - slouží k odhadu hodnot nebo středních hodnot proměnné/proměnných podmíněných hodnotami jedné či většího počtu vysvětlujících proměnných
  - odpovídá na otázku: *Jak vypadá závislost mezi proměnnými?*
- Korelační analýza:
  - slouží k vyjádření síly závislosti/těsnosti dvou a více numerických proměnných, respektive porovnání vhodnosti různých regresních modelů
  - odpovídá na otázku: *Jak silná je závislost mezi proměnnými, respektive jak moc odpovídá model skutečnosti?*

Regresní analýza

Dělení dle počtu závislých a nezávislých proměnných

- Jednoduchá regresní analýza,
  - slouží k popisu závislosti dvou numerických proměnných – hledáme matematický model – regresní funkci $y = f(x)$
  - funkce f (tj. model závislosti) je nejčastěji: lineární, polynomiální (kvadratická, kubická), hyperbolická, exponenciální, mocninná, odmocninná, logaritmická, ...
- Vícenásobná regresní analýza,
  - slouží k popisu závislosti jedné numerické proměnné na skupině jiných numerických proměnných – hledáme matematický model – regresní funkci $y = f(x_1, x_2, \ldots, x_l)$
  - tj. model závislosti je nejčastěji tzv.: adititivní, multiplikativní, model s interakcemi, ...
- Vícerozměrná regresní analýza,
  - slouží k popisu závislosti skupiny více numerických proměnných na skupině jiných numerických proměnných – hledáme matematický model – regresní funkci $(y_1, y_2, \ldots, y_q) = f(x_1, x_2, \ldots, x_l)$
  - Ani se neptejte :-o.

Volba regresního modelu

- Apriorní volba regresního modelu/modelů
  - volba druhu souvislosti/závislosti (závislé-nezávislé, vysvětlované-vysvětlující, . . .)
  - výběr modelu dle věcné souvislosti (volba funkce), respektive analytické řešení problému (např. diferenciální rovnice, apod.)
  - inspirace daty (korelační pole, . . .)
- Posteriorní volba regresního modelu
  - ověřování předpokladů pro použití modelu a odhad jeho parametrů
  - síla těsnosti (korelační analýza)
  - interpretovatelnost výsledků

Konstrukce regresního modelu

Způsob odhadů regresních koeficientů

- Příklady modelů
  - $y = \beta_0 + \beta_1 x$, *lineární regrese*
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_l x_l$, *adititvní model*
  - $x, x_1, x_2, \ldots x_l, y$ – proměnné modelu; $\beta, \beta_1, \beta_2, \ldots \beta_l$ – parametry modelu – tzv. regresní koeficienty
- Souvislost modelu a reality
  - model: $\mathbf{y} = f(\mathbf{x})$ – neznáme regresní koeficienty
  - realita $\mathbf{y}_i = f(\mathbf{x}_i) + \varepsilon_i$, pro $i = 1, \ldots, n$ – regresní koeficienty nastaveny tak, aby co nejlépe odpovídaly realitě, ale i tak zůstává nevysvětlená chyba – ε
  - odhad dle modelu: $\hat{\mathbf{y}} = f(\mathbf{x})$ – na základě odhadnutých koeficientů vypočteny hodnoty, které by měly odpovídat jak modelu tak reálné situaci
- Způsoby odhadu regresních koeficientů
  - metoda nejmenších čtverců
  - iterační metody
  - metoda maximální věrohodnosti
  - metoda vybraných bodů
  - ...

Metoda nejmenších čtverců – MNČ I

Demonstrace pro jednoduchou regresi

- Myšlenka MNČ
  - Součet druhých mocnin reziduí je pro danou regresní funkci f a data minimální
$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min$$
    - kde $[x_i; y_i]$, pro $i = 1, \ldots, n$ jsou empirické (naměřené) hodnoty nezávislé a závislé proměnné,
    - $\hat{y}_i$ je hodnota teoretická/vypočtená $\hat{y}_i = f(x_i)$ – zkráceně $\hat{y}(x_i)$,
- Poznámky k MNČ
  - Podmínku výše lze splnit vhodnou volbou regresních koeficientů β_j, pro $j = 0, \ldots, r$.
  - $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ je tedy funkcí $(r + 1)$ proměnných – označme ji $S(\beta_0, \beta_1, \ldots, \beta_r)$.
  - Je-li funkce f z pohledu regresních koeficientů lineárně separabilní, nebo dá-li se transformací na takovou funkci převést, lze tyto koeficienty získat jednoznačně pomocí prostředků matematické analýzy:
    - $\frac{\partial S}{\partial \beta_j} = 0$, pro $j = 0, \ldots, r$;

Metoda nejmenších čtverců – MNČ II

Demonstrace pro jednoduchou regresi

- řešení soustavy $(r+1)$ lineárních rovnic o $(r+1)$ neznámých, které má pro alespoň $(r+1)$ dvojic $[x_i; y_i]$ s různými x_i jednoznačné řešení;

- Za předpokladu, že model obsahuje absolutní člen, se odchylky teoretických a empirických hodnot (tj. reziduí) se v součtu „vynulují“: $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$;
- Sledují-li proměnné X a Y normální rozdělení, jsou odhady regresních koeficientů získané MNČ shodné s odhady získanými metodou maximální věrohodnosti.

Korelace a kovariace I

Posouzení lineární závislosti dvou numerických proměnných

- Variabilita jednotlivých proměnných X a Y:
  - rozptyl proměnné X: $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$,
  - rozptyl proměnné Y: $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$.
- Společná variabilita proměnných X a Y:
  - kovariance proměnných X a Y:

$$cov_{yx} = cov_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

    - $cov_{xx} = \sigma_x^2$, $cov_{yy} = \sigma_y^2$
    - $cov_{yx} \in \mathbb{R}$
  - (lineární) korelace proměnných X a Y (korelační koeficient, koeficient korelace):

$$r_{yx} = r_{xy} = \frac{cov_{yx}}{\sigma_x \sigma_y}$$

Korelace a kovariace II

Posouzení lineární závislosti dvou numerických proměnných

- $r_{yx} = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}$
- $r_{yx} = \dfrac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] \left[n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}$
- $r_{yx} \in \langle -1; 1 \rangle$
- $r_{yx} > 0$ – pozitivní (lineární) korelační závislost
- $r_{yx} < 0$ – negativní (lineární) korelační závislost
- $r_{yx} = 0$ – (lineární) nekorelovanost
- $|r_{yx}| = 1$ – matematická/funkční závislost
- r_{yx} – slovní hodnocení v biologii

	$\lvert r_{yx}\rvert = 0$	(lineární) korelační nezávislost
	$\lvert r_{yx}\rvert < 0{,}3$	nízký stupeň korelační závislosti
$0{,}3 \leq$	$\lvert r_{yx}\rvert < 0{,}5$	mírný stupeň korelační závislosti
$0{,}5 \leq$	$\lvert r_{yx}\rvert < 0{,}7$	střední stupeň korelační závislosti
$0{,}7 \leq$	$\lvert r_{yx}\rvert < 0{,}9$	vysoký stupeň korelační závislosti
$0{,}9 \leq$	$\lvert r_{yx}\rvert < 1$	velmi vysoký stupeň korelační závislosti
	$\lvert r_{yx}\rvert = 1$	matematická/funkční závislost

- Nejen hodnota r_{yx}, ale i rozsah souboru (n), vypovídá o síle závislosti!