Regresní a korelační analýza

Regresní analýza

- zkoumání jednostranné závislosti numerické proměnné *y* (závislá, vysvětlovaná) na numerické proměnné *x* (nezávislá, vysvětlující)
- nezávislá proměnná = příčina, závislá proměnná = důsledek
- důležitý je přitom směr závislosti (která proměnná je závislá a která nezávislá)
- závislost většinou modelujeme nějakou matematickou funkcí (tzv. regresní funkce).

Korelační analýza

- zabývá se především intenzitou vzájemného vztahu numerických proměnných
- na intenzitu závislosti je kladen větší důraz než na její směr
- zahrnuje míry intenzity závislosti
- „correlatió“ = vzájemná souvislost (z lat.)
- z výpočetních a interpretačních hledisek se regresní a korelační analýza prolínají.

Regresní modely

- matematické modely, které vyjadřují představu o průběhu závislosti proměnných
- umožňují odhady neznámých hodnot závisle proměnné ze známých hodnot nezávisle proměnné.

Obecný tvar modelu:

$$y_i = \eta_i + \varepsilon_i = \eta(x_i) + \varepsilon_i \ , \quad i = 1, 2, \ldots, n.$$

Symbolika: η_i … deterministická složka

ε_i … náhodná (rušivá) složka.

Typy modelů:

- **aditivní (součtový)** – jeho složky se skládají sčítáním, je nejběžnější
- **multiplikativní (součinový)** – jeho složky se skládají násobením.

Teoretická regresní funkce: $\eta = \eta(x)$

- existují různé typy regresních funkcí
- nejčastější jsou lineární regresní funkce
- linearita se může hodnotit jak z hlediska proměnných, tak z hlediska parametrů
- každá regresní funkce má určitý počet parametrů (jejich počet značíme *p*).

Parametry regresní funkce:

- neznámé konstanty; symbolicky je značíme řeckými písmeny $(\beta_0, \beta_1, \ldots, \beta_m)$
- jejich hodnoty lze odhadnout z výběrových dat
- je třeba k jejich odhadu zvolit takovou metodu, aby odhady měly co nejlepší vlastnosti.

1) Funkce lineární z hlediska parametrů

přímka	$\eta = \beta_0 + \beta_1 x$
rovina	$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
nadrovina	$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$
parabola	$\eta = \beta_0 + \beta_1 x + \beta_2 x^2$
hyperbola	$\eta = \beta_0 + \beta_1 x^{-1}$
logaritmická funkce	$\eta = \beta_0 + \beta_1 \ln x$
polynom	$\eta = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_m x^m$

2) Funkce nelineární z hlediska parametrů

exponenciální funkce	$\eta = \beta_0 \beta_1^x$
mocninná funkce	$\eta = \beta_0 x^{\beta_1}$
Tőrnquistova křivka	$\eta = \dfrac{\beta_0 x}{x + \beta_1}$

Jednoduchá lineární regrese

- regresní funkce je lineární z hlediska parametrů
- má jednu vysvětlující proměnnou (regresor) x.

Teoretická (hypotetická) regresní funkce: $\eta = \beta_0 + \beta_1 x$

- β_0, β_1 … *parametry*; x … *regresor*
- nutno provést odhad neznámých parametrů β_0, β_1
- odhad parametrů lineární regresní funkce provádíme **metodou nejmenších čtverců**
- když odhadneme parametry, získáme tzv. výběrovou regresní funkci.

Empirická (výběrová) regresní funkce: $\hat{\eta} = Y = b_0 + b_1 x$

- b_0, b_1 … *odhady parametrů*; $b_0 = \hat{\beta}_0$; $b_1 = \hat{\beta}_1$

Metoda nejmenších čtverců

- lze ji použít pouze k odhadu parametrů funkcí lineárních v parametrech (v lineární regresi)
- *princip:* parametry odhadujeme tak, aby pro ně byl minimální součet čtverců reziduí.

$y_i = \eta_i + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\quad i = 1, 2, \ldots, n$

$y_i = Y_i + \hat{\varepsilon}_i = b_0 + b_1 x_i + \hat{\varepsilon}_i$

Reziduum: $\hat{\varepsilon}_i = y_i - Y_i = y_i - b_0 - b_1 x_i = e_i$

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i \right)^2 \quad \Rightarrow \quad \text{minimalizovat}$$

1. stanovíme parciální derivace a položíme je rovny 0
2. vznikne soustava dvou rovnic o dvou neznámých (tzv. normální rovnice)
3. vyřešíme ji a získáme vzorce pro výpočet b_0 *a* b_1.

Vzorce pro výpočet parametrů výběrové regresní přímky:

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

b_1 … ***výběrový regresní koeficient (směrnice výběrové regresní přímky)***
udává průměrnou změnu proměnné *y* odpovídající zvýšení proměnné *x* o jednotku.

Sdružené regresní přímky

$Y = a_{yx} + b_{yx} x$ popisuje závislost *y* na *x*
$X = a_{xy} + b_{xy} y$ popisuje závislost *x* na *y*

1. $b_{yx} = b_{xy} = 0$

- ***x a y jsou korelačně nezávislé***
- sdružené regresní přímky svírají pravý úhel.

2. $b_{yx} = \frac{1}{b_{xy}}$

- ***x a y jsou perfektně závislé***
- sdružené regresní přímky svírají nulový úhel, tedy splývají.

Míry těsnosti lineární závislosti

Koeficient determinace: $r_{yx}^2 = r_{xy}^2 = b_{yx} \cdot b_{xy} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_{xy}}{s_y^2} = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2}$; $\quad r_{xy}^2 \in \langle 0;1 \rangle$

Koeficient korelace: $r_{yx} = r_{xy} = \sqrt{r_{yx}^2} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\left(\overline{x^2} - \bar{x}^2\right) \cdot \left(\overline{y^2} - \bar{y}^2\right)}}$; $\quad r_{xy} \in \langle -1;1 \rangle$

- měří pouze sílu lineární závislosti, nikoli závislosti obecně
- tento koeficient je symetrický.

Interpretace koeficientu korelace:

1. znaménko (+/–) udává směr závislosti:

$r_{xy} > 0 \Rightarrow$ přímá závislost

$r_{xy} < 0 \Rightarrow$ nepřímá závislost

2. $|r_{xy}|$ udává sílu závislosti:

$r_{xy} = 0 \quad \Rightarrow$ lineární nezávislost

$|r_{xy}| = 1 \quad \Rightarrow$ funkční (perfektní) závislost

$|r_{xy}| \to 0 \quad \Rightarrow$ slabá lineární závislost

$|r_{xy}| \to 1 \quad \Rightarrow$ silná lineární závislost

Test hypotézy o nulové hodnotě korelačního koeficientu

1) $H_0 : \rho_{yx} = 0$ (lineární nezávislost *x* a *y*)

$H_1 : non\ H_0$

2) ***Testové kritérium:***

$$t = \frac{r_{yx} \cdot \sqrt{n-2}}{\sqrt{1-r_{yx}^2}} \; ; \quad \text{Statistika } t \text{ má při platnosti } H_0 \text{ rozdělení } t(n-2)$$

3) ***Kritický obor:***

$$W \equiv \left\{ t;\, t \le t_{\frac{\alpha}{2}}(n-2)\ a\ t \ge t_{1-\frac{\alpha}{2}}(n-2) \right\}$$

4) Závěr testu:

Pokud leží hodnota testového kritéria v kritickém oboru, zamítáme H_0 a přijímáme H_1, tedy prokázali jsme hypotézu o lineární závislosti proměnných *x* a *y*.

Spearmanův koeficient pořadové korelace: $r_s = 1 - \frac{6\sum_{i=1}^{n}(a_i - b_i)^2}{n(n^2-1)}$; $\quad r_s \in \langle -1;1 \rangle$

- vychází pouze z pořadí naměřených hodnot
- neodráží pouze lineární závislost (jako klasický koeficient korelace), ale měří, jak dobře popisuje vhodná monotónní (tedy i nelineární) funkce závislost proměnných
- ***interpretace a test hypotézy o nulové hodnotě:*** stejné jako u korelačního koeficientu.

Míry těsnosti závislosti

- obecné míry, nezávislé na typu regresní funkce
- lze použít i pro měření nelineární závislosti
- tyto míry nejsou symetrické.

Index determinace: $I^2 = \frac{S_T}{S_y}$; $I^2 \in \langle 0;1 \rangle$

- udává, jaký podíl variability proměnné *y* lze vysvětlit zvolenou regresní funkcí
- lze ho vyjádřit v %

Index korelace: $I = \pm\sqrt{I^2}$; $I \in \langle -1;1 \rangle$

Rozklad celkového součtu čtverců

$S_y = S_T + S_R$

S_y … ***celkový součet čtverců***

S_T … ***teoretický součet čtverců***

část variability, kterou lze vysvětlit zvolenou regresní funkcí

S_R … ***reziduální součet čtverců***

část variability, kterou nelze vysvětlit zvolenou regresní funkcí.

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

$$S_T = \sum_{i=1}^{n} (Y_i - \bar{y})^2 \; ; \quad S_R = \sum_{i=1}^{n} (y_i - Y_i)^2 \; .$$

Testování vhodnosti regresního modelu

Celkový F – test

- testuje vhodnost modelu jako celku
- analýza rozptylu.

1) $H_0 : \beta_0 = c, \; \beta_1, \beta_2, ..., \beta_m = 0$ (regresní funkce nemá žádný význam, tj. není vhodná)
$H_1 : non \; H_0$

2) ***Testové kritérium:***

$F = \frac{S_T / p-1}{S_R / n-p}$; Statistika *F* má při platnosti H_0 rozdělení $F(p-1; n-p)$.

3) *Kritický obor:*

$$W \equiv \{F;\ F > F_{1-\alpha}(p-1;\ n-p)\}$$

4) Závěr testu:

Pokud leží hodnota testového kritéria v kritickém oboru, zamítáme H_0 a přijímáme H_1. Model lze na dané hladině významnosti považovat za vhodný.

Dílčí t – testy

- testy o nulové hodnotě jednotlivých regresních parametrů
- počet testů je roven počtu parametrů modelu.

1) $H_0 : \beta_h = 0,\ h = 1,2,\ldots,m$

$H_1 : non\ H_0$

2) *Testové kritérium:*

$t_h = \dfrac{b_h}{s(b_h)}$, $h = 1,2,\ldots,m.$ Statistika t_h má při platnosti H_0 rozdělení $t(n-p)$.

3) *Kritický obor:*

$$W \equiv \left\{ t_h;\ t_h \le t_{\frac{\alpha}{2}}(n-p)\ a\ t_h \ge t_{1-\frac{\alpha}{2}}(n-p) \right\}$$

4) Závěr testu:

Pokud leží hodnota testového kritéria v kritickém oboru, zamítáme H_0 a přijímáme H_1. Testovaný parametr lze na dané hladině významnosti považovat v regresní funkci za přínosný.

Jednoduchá nelineární regrese

- není-li regresní funkce lineární v parametrech, nelze její parametry odhadnout metodou nejmenších čtverců
- pro odhad parametrů se používá řada různých metod, například metoda linearizující transformace (logaritmická apod.) nebo metoda částečných součtů
- většinou následují další metody pro zlepšení vlastností odhadů
- výpočetně značně náročné (využití statistických programů).

Volba vhodného typu regresní funkce

- volba by se měla v prvé řadě opírat o věcný rozbor vztahů proměnných
- při volbě nejvhodnější regresní funkce lze kombinovat různá kritéria
- vždy se snažíme o jednoduchost modelu (ne příliš mnoho parametrů)
- úspěšnost modelu je nezbytné ověřit vhodným testem
- dále je třeba změřit přilnavost regresní funkce k datům vhodnou mírou.

1. Index determinace

- za vhodnější je považována ta regresní funkce, u které je hodnota I^2 vyšší.

***Při srovnávání funkcí s nestejným počtem parametrů je třeba hodnotu I^2 upravit (penalizovat)**, neboť u funkcí s vyšším počtem parametrů vychází hodnota I^2 automaticky vyšší.*

Existují různé formy penalizace, například:

$$I^2_{adj} = 1 - \left(1 - I^2\right) \cdot \frac{n-1}{n-p} = 1 - \frac{(n-1)S_R}{(n-p)S_y} \ .$$

Pozn.: adjusted = upravený.

2. Testové kritérium F

- testové kritérium celkového *F*-testu vhodnosti modelu
- za vhodnější je považována funkce, u níž je hodnota statistiky *F* vyšší
- toto kritérium lze použít vždy, bez ohledu na to, kolik parametrů mají srovnávané regresní funkce.

$$F = \frac{S_T / p - 1}{S_R / n - p}$$

3. Reziduální součet čtverců: $S_R = \sum_{i=1}^{n} \left(y_i - Y_i\right)^2$

- za vhodnější je považována funkce, která má reziduální součet čtverců nižší
- reziduální součet čtverců lze použít pouze tehdy, když srovnáváme funkce se stejným počtem parametrů.

4. Reziduální rozptyl: $S_R^2 = \frac{S_R}{n-p}$

- za vhodnější je považována funkce, která má reziduální rozptyl nižší
- reziduální rozptyl lze použít vždy, bez ohledu na to, kolik parametrů mají srovnávané regresní funkce.