KORELAČNÍ ANALÝZA

Korelační analýza

- Měření *intenzity závislosti* mezi proměnnými
- Úzká návaznost na regresní analýzu, neboť se v ní využívá teorie lineárních regresních modelů
- Nehledá formu vztahu mezi proměnnými, neboť už primárně vychází z předpokladu, že tento vztah je lineární (dokonce nejen z hlediska parametrů, ale i z hlediska proměnných), a soustředí se na konstrukci měr závislostí mezi těmito proměnnými.

Populační párový korelační koeficient

- **V nejjednodušším případě se sleduje závislost dvou náhodných veličin *Y* a *X*. V tomto případě lze použít jako míru lineární závislosti těchto veličin *(párový) koeficient korelace* ρ_{xy}, definovaný vztahy:**

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma(X).\sigma(Y)} \text{ pro } \sigma(X) > 0, \sigma(Y) > 0,$$

  - **Jinak $\rho_{xy}=0$**
- **Kde $Cov(x,y)$ značí kovarianci náhodných veličin *X* a *Y*, $\sigma(x)$ a $\sigma(y)$ je směrodatná odchylka veličiny *X*, respektive *Y*, symbol *E* značí střední hodnotu náhodné veličiny.**

Hodnoty koeficientu korelace

- Pro párový koeficient korelace platí, že ρ_{xy} je z intervalu [-1,1].
- Je-li $\rho_{xy} = 0$, říkáme, že veličiny *X* a *Y* jsou nezkorelované.
- Je-li $\rho_{xy} = 1$ nebo $\rho_{xy} = -1$, existuje přesná funkční závislost mezi veličinami *X* a *Y* v podobě přímky.
- Tato přímka je rostoucí v prvním případě a klesající ve druhém případě.
- Je-li $\rho_{xy} = 0$, je třeba se omezit pouze na konstatování, že obě veličiny jsou nezkorelované. Nelze tvrdit, že jsou (statisticky) nezávislé.

Příklad

- Vypočítejme koeficient korelace , jsou-li dány tyto údaje:

X	-2	-1	0	1	2
Y	4	1	0	1	4

- Všechny páry hodnot nastávají se stejnou pravděpodobností p.

Řešení příkladu

- Mezivýpočty pro vyhodnocení koeficientu korelace:

X_i	Y_i	$X_i.Y_i$
-2	4	-8
-1	1	-1
0	0	0
1	1	1
2	4	8
$\Sigma x_i = 0$	$\Sigma y_i = 10$	$\Sigma x_i.y_i = 0$

- Kovariance: $Cov(X,Y) = E(XY) - E(X) \cdot E(Y) = p\sum_i x_i y_i - p^2 \sum_i x_i \sum_i y_i = 0$
- $\rho_{xy}=0$

- Přitom tyto veličiny rozhodně nejsou nezávislé. Dokonce mezi nimi existuje přesná funkční závislost v podobě kvadratické funkce.

Výběrový párový korelační koeficient

- Populační párový korelační koeficient aproximujeme výběrovým párovým korelačním koeficientem r_{xy}, získaným na základě realizace náhodného výběru:

$$r_{xy} = \frac{n.\sum x_i.y_i - \sum x_i.\sum y_i}{\sqrt{\left[n.\sum x_i^2 - (\sum x_i)^2\right]\left[n.\sum y_i^2 - (\sum y_i)^2\right]}}$$

Test statistické významnosti korelačního koeficientu

1. Nulová hypotéza H_0: $\rho_{xy} = 0$,
   alternativní hypotéza H_1: $\rho_{xy} \neq 0$.
2. Testové kritérium:

$$T = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

3. Kritická hodnota testu K = $t_{n-2}(\alpha)$
   - Studentov0 rozdělení s *n*-2 stupni volnosti.

4. Je-li $|T| \geq K$, pak se H_0 zamítá, co znamená že korelační koeficient není statisticky významný, existuje lineární závislost mezi X a Y. V opačném případě se H_0 nezamítá, Y není lineárně závislé na X.

Příklad

- Mějme hodnoty x_i a y_i získané náhodným výběrem:

x_i	y_i
-2	-5
-1	-3
0	0
1	1
2	4

- Pro tyto hodnoty vypočítejte hodnotu korelačního koeficientu a testujte jeho statistickou významnost na hladině významnosti 0,01.

Příklad - řešení

- **Pro výpočet potřebujeme hodnoty:**

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
-2	-5	10	4	25
-1	-3	3	1	9
0	0	0	0	0
1	1	1	1	1
2	4	8	4	16
$\Sigma = 0$	$\Sigma = -3$	$\Sigma = 22$	$\Sigma = 10$	$\Sigma = 51$

- **Test statistické významnosti:**

$$r_{xy} = \frac{5.22 - 0.(-3)}{\sqrt{(5.10 - 0).(5.51 - (-3)^2)}} = 0,9918.$$

1. $H_0: \rho_{xy} = 0,\ \rho_{xy} \neq 0.$
2. $$T = \frac{0,9918.\sqrt{5-2}}{\sqrt{1-0,9918^2}} = \frac{1,718}{\sqrt{0,016}} = 13,443$$
3. Kritická hodnota testu $K = t_{n-2}(\alpha) = t_{5-2}(0,01) = 5,84$
4. Protože $T > K$, je lineární závislost Y na X významná.

Index korelace

- Není-li regresní funkcí, podle níž se posuzuje korelace veličin, přímka, ale jiná, i **nelineární** funkce, je možné k odhadu závislosti *X* a *Y* použít *index korelace*:

$$I_{xy} = \sqrt{\frac{S_{\hat{Y}}}{S_Y}}$$

- Kde:

$$S_{\hat{Y}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 ,$$

$$S_Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 .$$

Koeficient determinace

- Koeficient determinace určuje přiléhavost dat ke zvolenému modelu

- $R^2 = I^2 = \frac{S_{\hat{Y}}}{S_Y}$
- Index determinace (nebo taky koeficient determinace) tedy udává kvalitu regresního modelu, přesněji vyjádřeno udává, kolik procent rozptylu vysvětlované proměnné je vysvětleno modelem a kolik zůstalo nevysvětleno;
- nabývá hodnot od nuly do jedné (teoreticky i včetně těchto krajních mezí), přičemž hodnoty blízké nule značí špatnou kvalitu regresního modelu; hodnoty blízké jedné značí dobro kvalitu regresního modelu;
- udává se většinou v procentech.

SPEARMANŮV KORELAČNÍ KOEFICIENT

- Jsou-li hodnoty veličin *X* ,*Y* zadány **pořadím**, používá se k odhadu míry závislosti těchto veličin *Spearmanův koeficient (pořadové) korelace*, který se počítá dle vzorce

$$r_S = 1 - \frac{6 \cdot \sum_i d_i^2}{n(n^2 - 1)}.$$

- Kde d_i diference i-tého pořadí *X* a *Y* a *n* je počet párů hodnot *X* a *Y* , tedy rozsah výběru.

Příklad

- Výrobky byly seřazeny dle jakosti dvěma komisemi, z nichž jednu tvořili odborníci a druhou zástupci laické veřejnosti. Rozhodněte, zda se výsledky hodnocení obou komisí shodují ve smyslu korelace.

Výrobek	Laické pořadí	Odborné pořadí
1	7	8
2	9	9
3	8	7
4	10	10
5	6	6
6	5	4
7	3	5
8	4	3
9	2	2
10	1	1

Příklad - řešení

- V levé části níže uvedené tabulky jsou pořadí, v pravé části této tabulky jsou spočteny rozdíly v pořadí.

Výrobek	Laické pořadí	Odborné pořadí	d_i	d_i^2
1	7	8	-1	1
2	9	9	0	0
3	8	7	1	1
4	10	10	0	0
5	6	6	0	0
6	5	4	1	1
7	3	5	-2	4
8	4	3	1	1
9	2	2	0	0
10	1	1	0	0

$$r_S = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} = 1 - \frac{6.8}{10.99} = 0,95.$$

Test statistické významnosti pořadového koeficientu korelace

1. Testovaná hypotéza H_0: X,Y jsou nezávislé vs. alternativní hypotéza H_1: X,Y nejsou nezávislé.

2. Testové kritérium má tvar:

$$T = (n-1) \cdot r_S .$$

3. Kritická hodnota testu K = kritická hodnota rozdělení $N(0,1)$ na hladině významnosti alfa = NORMSINV(1-alfa).

4. Je-li $|T| \geq K$, zamítáme hypotézu H_0. V opačném případě přijímáme H_0.

Přijmeme-li H_0, víme, že jsou veličiny nezávislé, a tedy i nezkorelované. Pokud hypotézu zamítneme, víme, že veličiny nejsou nezávislé, nejsme ale schopni rozhodnout v takovém případě, zda jsou nezkorelované. Test platí přibližně pro $n \geq 30$.

VÍCENÁSOBNÁ ZÁVISLOST – PŘÍPAD DVOU VYSVĚTLUJÍCÍCH PROMĚNNÝCH

- **Chceme-li zjistit lineární závislost proměnné *Y* na větším počtu vysvětlujících proměnných $X_1, X_2, \ldots X_p$, používáme k měření těsnosti závislosti buďto:**
- **a. koeficienty dílčí (parciální) korelace,**
- **b. koeficient vícenásobné korelace.**

Koeficient dílčí (parciální) korelace

- Měří kupříkladu intenzitu lineární závislosti proměnné *Y* na vysvětlující proměnné X_1 za předpokladu, že je jistým způsobem odstraněn vliv ostatních proměnných . Značí se:

$$r_{yx_1 \bullet x_2,\ldots,x_p}$$

- Jde o proměnné, které jsou uvedeny za symbolem „•“.
- Důvod výpočtu tohoto ukazatele jen ten, že vliv proměnné X_1 může být zkreslen současným působením proměnných X_2, …X_p. Omezíme se nyní na případ $p = 2$.

Výběrový koeficient parciální korelace p=2

- Koeficient dílčí korelace vystupuje opět ve dvou podobách: buďto jde o populační koeficient nebo jeho odhad – výběrový koeficient parciální korelace. Výběrový koeficient parciální korelace se v **případě dvou vysvětlujících proměnných**

$$r_{yx_1 \bullet x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}} \qquad r_{yx_2 \bullet x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right)\left(1 - r_{x_1x_2}^2\right)}}.$$

- Koeficienty mohou nabývat hodnot z intervalu [-1,1].

Test statistické významnosti koeficientu parciální korelace

1. $H_0: \quad \rho_{yx_1 \bullet x_2} = 0$ (není přítomna korelační závislost), $H_1: \quad \rho_{yx_1 \bullet x_2} \neq 0$.

2. Testové kritérium:

4. Pokud $|T| \geq t_{n-3}(\alpha)$, $T = \dfrac{r_{yx_1 \bullet x_2}\sqrt{n-3}}{\sqrt{1-r^2_{yx_1 \bullet x_2}}}$.

3. Kritická hodnota na hladině alfa = K = $t_{n-3}(\alpha)$ = TINV(α,*n-3*).

4. Pokud $|T| \geq t_{n-3}(\alpha)$,

pak je koeficient parciální korelace statisticky významný, tj. nenulový.

Koeficient vícenásobné korelace

- ***Koeficient vícenásobné korelace*** **měří závislost proměnné *Y* na všech vysvětlujících proměnných $X_1, X_2, \ldots X_p$ dohromady.**
- **Pro 2 vysvětlující proměnné se výběrová verze tohoto koeficientu spočte dle vztahu:**

$$r_{y\bullet x_1x_2} = \sqrt{\frac{r_{yx_1}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2} + r_{yx_2}^2}{1 - r_{x_1x_2}^2}}, \quad 0 \le r_{y\bullet x_1x_2} \le 1.$$

Test statistické významnosti koeficientu vícenásobné korelace:

1. $H_0: \quad \rho_{y \bullet x_1 x_2} = 0$ (není závislost) vs. $H_1: \quad \rho_{y \bullet x_1 x_2} \neq 0$.

2. Testové kritérium: $$T = \frac{r^2_{y \bullet x_1 x_2} \cdot (n-3)}{2 \cdot \left(1 - r^2_{y \bullet x_1 x_2}\right)}$$

3. Kritická hodnota K se tentokrát týká Fisherova rozdělení se stupni volnosti 2 a *n*-3: $F_{2,n-3}$(alfa)

4. Pokud je $T \geq F_{2,n-3}(\alpha)$,

pak je koeficient vícenásobné korelace statisticky významný na dané hladině významnosti. V opačném případě není statisticky významný.

Děkuji za pozornost