Zadania ze statystyki cz.4 I– miary związków między zmiennymi

Zadanie 1

Poniższa tabela prezentuje rozkład wydatków na kulturę i wieku głowy gospodarstwa domowego. Proszę wyznaczyć model regresji liniowej wpływu wieku na wydatki oraz ustalić wielkość błędu standardowego (reszt) tego modelu.

y	**x**
300	32
250	28
320	29
200	44
280	52
140	60
230	40

Zadanie 2

Dla danych z zadania 1! Należy wyrangować te cechy i następnie obliczyć współczynnik korelacji rang Spearmana.

Zadanie 3

Tabela poniżej prezentuje szereg korelacyjny dwóch cech: x – liczba godzin przeznaczonych na naukę; y – ocena ze sprawdzianu. Proszę zbadać siłę i kierunek związku miedzy tymi zmiennymi, wykorzystując współczynnik korelacji rang Spearmana.

x	y
3	3
4	3,5
5	4
6	4
6	4,5
7	3,5
7	4
7	4,5
9	5

PODSUMOWANIE

Regresja liniowa – błąd standardowy pomiaru

- Po wyprowadzeniu funkcji liniowej, każdemu X możemy przypisać dwie wartości Y - wartość empiryczną (zbadaną) *yi* oraz teoretyczną *y'*(wynikającą z danej funkcji y=ax+b).

- Błąd standardowy modelu regresji „e”, to różnice między wartościami empirycznymi a teoretycznymi.

$$e = \sqrt{\frac{\sum(y_i - \hat{y})^2}{N - 1 - k}}$$

$k - liczba\ zmiennych\ niezależnych$

$\hat{y} - wartości\ teoretyczne\ Y$

Współczynnik korelacji rang Spearmana

- Jest szczególnym przypadkiem współczynnika korelacji liniowej Pearsona. Różnica między tymi współczynnikami jest tym większa, im większa jest krzywoliniowość badanego związku.
- Dedykowany jest związkom krzywoliniowym.
- Opiera się na rangach wartości zmiennej niezależnej i zależnej.
- Gdy każda ze zmiennych posiada wartości niepowtarzalne (nie ma dwóch obserwacji o takich samych wartościach dwóch wśród zmiennych), a tym samym rangi w obu zmiennych są niepowtarzalne, korelację między zmiennymi obliczymy”:

$$r_s = 1 - \frac{6 * \sum(Rx_i - Ry_i)^2}{N * (N^2 - 1)} = 1 - \frac{6 * \sum {d_i}^2}{N * (N^2 - 1)}$$

$Rx_i - ranga\ jednostki\ ze\ względu\ na\ zmienną\ X$

$Ry_i - ranga\ jednostki\ ze\ względu\ na\ zmienną\ Y$

${d_i}^2 - różnica\ między\ rangami\ x\ i\ y\ do\ kwadratu$

$N - liczebność\ próby$

Nadawanie rang – rangowanie cech

- Rangowanie to inaczej uporządkowywanie cech ze względu na wartości cechy w taki sposób, że każdej obserwacji nadaje się numer porządkowy, wskazujący na miejsce w strukturze wyrażające natężenie danej cechy.

X	ranga X
90	1
100	2
120	3
125	4
135	5
140	6
145	7
150	8
160	9
170	10
177	11
189	12
200	13

Rangowanie zmiennej – rangi wiązane

- Jeżeli w szeregu występują jednakowe pomiary (obserwacje, które miały identyczne wartości danej zmiennej) przypisane różnym obserwacjom, to każdej z takich obserwacji musimy przypisać jednakową rangę (numer porządkowy).
- Sytuację, w której co najmniej dwie różne obserwacje otrzymały taką samą rangę nazywamy wiązaniem, taki rozkład szeregiem powiązanym rangami.
- W przypadku takich rozkładów musimy posługiwać się współczynnikami dla zmiennych powiązanych rangami.

X	Ranga X
2	1,5
2	1,5
3	3,5
3	3,5
5	5,5
5	5,5
6	7
7	8,5
7	8,5
9	10
10	11
11	12

- Sytuację gdy obserwacje posiadają powtarzające się wartości zmiennej X lub Y, a tym samym powtarzają się rangi tych zmiennych, nazywamy pomiarami **powiązanymi rangami**, lub **rangami wiązanymi**.
- W takiej sytuacji musimy uwzględnić ilość obserwacji, które są powiązane rangami.

$$r_{hos} = \frac{T_x + T_y - \sum d_i^2}{2 * \sqrt{T_x * T_y}}$$

$$T_x = \frac{N^3 - N - \sum(t_{xi}^3 - t_{xi})}{12}$$

$$T_y = \frac{N^3 - N - \sum(t_{yi}^3 - t_{yi})}{12}$$

$t_{xi} lub\ t_{yi} -$ *liczba obserwacji powiązanych rangą w grupie tych samych zmiennych*

Przykład

Zmienna X prezentuje dochód na osobę w rodzinie, zmienna Y przeciętne wydatki na używki. Czy wysokość dochodu jest związana z kwotą przeznaczaną na używki?

$$d_i^2$$

X	Y	ranga X	ranga Y	Rxi - Ryi	(Rxi - Ryi)²
900	120	1	1	0	0
1000	150	2	2	0	0
1200	300	3	5	-2	4
1250	340	4	7	-3	9
1350	320	5	6	-1	1
1400	370	6	8	-2	4
1450	420	7	10	-3	9
1500	450	8	11	-3	9
1600	500	9	13	-4	16
1700	470	10	12	-2	4
1770	400	11	9	2	4
1890	290	12	4	8	64
2000	220	13	3	10	100
				Σ	**224**

Ponieważ pomiary są niepowtarzalne, to również rangi są niepowiązane, a zatem stosujemy:

$$r_s = 1 - \frac{6 * \sum {d_i}^2}{N * (N^2 - 1)}$$

$$r_s = 1 - \frac{6 \cdot 224}{13 \cdot (13^2 - 1)}$$

$$r_s = 1 - \frac{1344}{2184} = 1 - 0{,}615 = 0{,}385$$

Przykład 2

Zmienna X prezentuje dochód na osobę w rodzinie, zmienna Y przeciętne wydatki na używki. Czy wysokość dochodu jest związana z kwotą przeznaczaną na używki?

Tym razem występują powtarzalne pomiary (np. dwa razy kwota 1200 i dwa razy kwota 1400 oraz wydatki 300 zł). Występują więc rangi wiązane.

X	Y	rangaX	rangaY	d²
900	120	1	1	0
1000	150	2	2	0
2000	220	13	3	100
1200	300	3,5	4,5	1
1890	300	12	4,5	56,25
1200	340	3,5	6,5	9
1350	340	5	6,5	2,25
1400	370	6,5	8	2,25
1770	400	11	9	4
1400	420	6,5	10	12,25
1500	450	8	12	16
1600	450	9	12	9
1700	450	10	12	4
				216

Występują rangi wiązane, a zatem rhos obliczamy:

$$r_{hos} = \frac{T_x + T_y - \sum d_i^2}{2 * \sqrt{T_x * T_y}}$$

Uwzględniając obserwacje powiązane rangami w każdej zmiennej (X i Y)

$$T_x = \frac{N^3 - N - \sum(t_{xi}^3 - t_{xi})}{12}$$
$$T_y = \frac{N^3 - N - \sum(t_{yi}^3 - t_{yi})}{12}$$

$$T_x = \frac{N^3 - N - \sum(t_{xi}^3 - t_{xi})}{12}$$

xi	ranga Xi	txi	txi^3	txi^3-txi
1200	3,5	2	8	6
1400	6,5	2	8	6
			Σ	**12**

$$T_x = \frac{13^3 - 13 - 12}{12} = \frac{2197 - 13 - 12}{12} = \frac{2172}{12} = 181$$

$$T_y = \frac{N^3 - N - \sum(t_{yi}^3 - t_{yi})}{12}$$

yi	rangaYi	tyi	tyi^3	tyi^3-tyi
300	4,5	2	8	6
340	6,5	2	8	6
450	12	3	27	24
			Σ	**36**

$$T_y = \frac{13^3 - 13 - 36}{12} = \frac{2197 - 13 - 36}{12} = \frac{2148}{12} = 179$$

A zatem:

$$r_{hos} = \frac{T_x + T_y - \sum d_i^2}{2 * \sqrt{T_x * T_y}}$$

$$r_{hos} = \frac{181 + 179 - 216}{2 \cdot \sqrt{181 \cdot 179}} = \frac{144}{2 \cdot \sqrt{32399}} = \frac{144}{2 \cdot 179{,}99} = 0{,}4$$