Losowe czynniki

$$m = 1 \qquad p = 2$$

Obserwujemy trzy cechy: X_1, X_2, X_3
Obiekt $\longrightarrow (X_1, X_2, X_3)$

1. Czy cechy X_1, X_2, X_3 są niezależne?
   Czy X_1 zależy jednocześnie od X_2 oraz X_3?
   Czy X_1 zależy od X_2?
   Czy X_1 zależy od X_3?

2. Opis ilościowy zależności.

3. Wnioski.

Założenie:
Łączny rozkład cech X_1, X_2, X_3 jest normalny

Współczynnik korelacji wielokrotnej
między X_1 a parą (X_2, X_3)

$$\varrho_{1|2,3} = \max_{\lambda_1,\lambda_2} \varrho(X_1, \lambda_1 X_2 + \lambda_2 X_3)$$

1. $\varrho_{1|2,3} \in \langle 0, 1\rangle$
2. $\varrho_{1|2,3} = 0$: X_1 nie jest zależne od (X_2, X_3)
3. $\varrho_{1|2,3} = 1$: X_1 jest liniową funkcją (X_2, X_3)

$$\varrho_{1|2,3} = \sqrt{1 - \frac{|\mathcal{C}|}{|\mathcal{C}_{11}|}}$$

$|\mathcal{C}|$: wyznacznik macierzy $\mathcal{C} = \begin{pmatrix} 1 & \varrho_{12} & \varrho_{13} \\ \varrho_{21} & 1 & \varrho_{23} \\ \varrho_{31} & \varrho_{32} & 1 \end{pmatrix}$

$|\mathcal{C}_{11}|$: wyznacznik macierzy $\mathcal{C}_{11} = \begin{pmatrix} 1 & \varrho_{23} \\ \varrho_{32} & 1 \end{pmatrix}$

$$H_0 : \text{Cecha } X_1 \text{ jest niezależna od } X_2, X_3$$
$$\Updownarrow$$
$$H_0 : \varrho_{1|2,3} = 0$$

Test współczynnika korelacji wielokrotnej
(poziom istotności α)

Próba: $(X_{1i}, X_{2i}, X_{3i}), i = 1, \ldots, n$
Statystyka testowa

$$R_{1|2,3} = \sqrt{1 - \frac{|C|}{|C_{11}|}}$$

$|C|$, $|C_{11}|$: próbkowe odpowiedniki $|\mathcal{C}|$ oraz $|\mathcal{C}_{11}|$

Wartość krytyczna $r(\alpha; n, 3)$

Jeżeli $R_{1|2,3} > r(\alpha; n, 3)$,
to hipotezę odrzucamy.

Współczynnik korelacji cząstkowej
między X_1 a X_2
lub
między X_1 a X_3

$$\varrho_{12|3} = \frac{\varrho_{12} - \varrho_{13}\varrho_{23}}{\sqrt{(1-\varrho_{13}^2)(1-\varrho_{23}^2)}}$$

Miernik indywidualnego wpływu zmiennej X_2 na X_1

$$\varrho_{13|2} = \frac{\varrho_{13} - \varrho_{12}\varrho_{23}}{\sqrt{(1-\varrho_{12}^2)(1-\varrho_{23}^2)}}$$

Miernik indywidualnego wpływu zmiennej X_3 na X_1

$$H_0 : \text{Cecha } X_1 \text{ jest niezależna od } X_2$$
$$\Updownarrow$$
$$H_0 : \varrho_{12|3} = 0$$

Test współczynnika korelacji cząstkowej
(poziom istotności α)

Próba: $(X_{1i}, X_{2i}, X_{3i}), i = 1, \ldots, n$
Statystyka testowa

$$R_{12|3} = \frac{R_{12} - R_{13}R_{23}}{\sqrt{(1 - R_{13}^2)(1 - R_{23}^2)}}$$

Wartość krytyczna $r(\alpha; n-1)$

Jeżeli $R_{12|3} > r(\alpha; n-1)$,
to hipotezę odrzucamy.

Ilościowy opis zależności

X_1: zmienna zależna; X_2, X_3: zmienne niezależne
Założenie: $\varrho_{1|2,3} \neq 0$

Ilościowy opis zależności X_1 od X_2, X_3:

$$E(X_1|X_2 = x_2, X_3 = x_3) = f(x_2, x_3)$$

Funkcja f nosi nazwę **funkcji regresji**

Przy założeniu normalności

$$f(x_2, x_3) = \beta_0 + \beta_1 x_2 + \beta_2 x_3$$

β_1, β_2 — współczynniki regresji
β_0 — stała regresji

Zadanie: oszacować parametry funkcji regresji

$$\varrho_{12|3} \neq 0 \text{ oraz } \varrho_{13|2} = 0$$

$$\Downarrow$$

$$f(x_2, x_3) = \beta_0 + \beta_1 x_2$$

$$\varrho_{12|3} = 0 \text{ oraz } \varrho_{13|2} \neq 0$$

$$\Downarrow$$

$$f(x_2, x_3) = \beta_0 + \beta_2 x_3$$

$$\varrho_{12|3} \neq 0 \text{ oraz } \varrho_{13|2} \neq 0$$

$$\Downarrow$$

$$f(x_2, x_3) = \beta_0 + \beta_1 x_2 + \beta_2 x_3$$

Ocena parametrów funkcji regresji

$$f(x_2, x_3) = \beta_0 + \beta_1 x_2 + \beta_2 x_3$$

Próba: $(X_{1i}, X_{2i}, X_{3i}), i = 1, \ldots, n$

Niezbędne rachunki

$$\bar{X}_1,\ \mathrm{var}X_1,\ \bar{X}_2,\ \mathrm{var}X_2,\ \bar{X}_3,\ \mathrm{var}X_3$$

$$\mathrm{cov}(X_1, X_2),\ \mathrm{cov}(X_1, X_3),\ \mathrm{cov}(X_2, X_3)$$

$$\begin{cases} \widehat{\beta}_1 \mathrm{var}X_2 + \widehat{\beta}_2 \mathrm{cov}(X_2, X_3) = \mathrm{cov}(X_1, X_2) \\ \widehat{\beta}_1 \mathrm{cov}(X_2, X_3) + \widehat{\beta}_2 \mathrm{var}X_3 = \mathrm{cov}(X_1, X_3) \\ \bar{X}_1 - \widehat{\beta}_1 \bar{X}_2 - \widehat{\beta}_2 \bar{X}_3 = \widehat{\beta}_0 \end{cases}$$

Wariancja resztowa

$$S^2 = \frac{\mathrm{var}X_1 - \widehat{\beta}_1 \mathrm{cov}(X_1, X_2) - \widehat{\beta}_2 \mathrm{cov}(X_1, X_3)}{n-3}$$

Przedziały ufności (poziom ufności $1-\alpha$)

Wariancje estymatorów $\widehat{\beta}_0$, $\widehat{\beta}_1$ oraz $\widehat{\beta}_2$

$$S_{\beta_0}^2 = S^2\left(\frac{1}{n} + \frac{\frac{\bar{X}_2^2}{\mathrm{var}X_2} + \frac{\bar{X}_3^2}{\mathrm{var}X_3} - \frac{R_{23}^2}{\mathrm{cov}(X_2, X_3)}}{1 - R_{23}^2}\right)$$

$$S_{\beta_1}^2 = \frac{S^2}{(1-R_{23}^2)\mathrm{var}X_2}, \quad S_{\beta_2}^2 = \frac{S^2}{(1-R_{23}^2)\mathrm{var}X_3}$$

$$\beta_0 \in (\widehat{\beta}_0 - t(\alpha; n-3)S_{\beta_0}, \widehat{\beta}_0 + t(\alpha; n-3)S_{\beta_0})$$

$$\beta_1 \in (\widehat{\beta}_1 - t(\alpha; n-3)S_{\beta_1}, \widehat{\beta}_1 + t(\alpha; n-3)S_{\beta_1})$$

$$\beta_2 \in (\widehat{\beta}_2 - t(\alpha; n-3)S_{\beta_2}, \widehat{\beta}_2 + t(\alpha; n-3)S_{\beta_2})$$

Przykład. Badano zależność wydatków na artykuły spożywcze (spo), papierosy (pap) oraz alkohol (alk). Zbadać istnienie zależności między obserwowanymi cechami. Jeżeli taka zależnośc istnieje, to dokonać ilościowego opisu zależności wydatków na artykuły spożywcze od wydatków na papierosy i alkohol.

Plan działania

1. Istnienie zależności (globalne)

2. Istnienie zależności (szczegółowe)

3. Ilościowy opis

4. Wnioski

spo	pap	alk	spo	pap	alk	spo	pap	alk
44.700	17.219	9.007	46.111	20.942	10.999	45.404	20.009	12.531
46.276	25.147	17.896	45.711	22.822	16.530	45.525	23.036	7.398
42.816	14.058	3.966	46.444	25.741	15.733	45.464	23.497	9.083
43.913	16.138	2.994	45.002	20.720	10.082	45.045	22.251	13.428
43.153	13.745	1.762	43.945	14.510	7.168	45.239	20.838	14.094
42.882	13.699	1.531	45.444	22.108	12.129	44.957	18.856	11.239
45.135	19.768	9.826	43.973	19.331	6.090	44.990	21.613	10.985
44.630	21.172	8.871	44.160	16.562	6.022	46.453	23.518	13.543
44.814	18.716	12.596	44.547	18.017	9.695	43.461	12.029	3.969
45.866	25.955	12.608	45.513	20.351	13.938	45.299	20.736	11.418
46.661	21.698	16.489	43.239	16.374	3.604	45.258	22.986	11.462
45.902	25.328	13.603	45.042	19.075	9.032	43.883	14.958	5.947
44.476	19.728	7.633	45.849	23.434	14.054	45.951	20.049	11.978
45.758	19.657	10.890	45.691	24.335	13.439	45.524	20.994	11.236
43.478	15.544	4.372	45.589	24.425	12.026	44.837	18.890	7.913
45.028	19.563	13.800	46.038	22.876	15.742	45.645	21.738	14.515
43.258	14.392	0.000	44.852	18.392	11.117	45.262	20.759	11.350
46.448	21.677	13.751	44.418	19.357	8.986	45.214	20.536	8.660
45.758	22.670	14.357	44.684	18.497	6.391	45.235	20.834	12.183
45.596	19.358	9.863	46.993	24.193	17.175	44.264	19.093	7.888
45.694	22.280	10.962	44.766	23.872	12.055	45.114	21.661	13.746
44.759	19.592	10.129	45.881	24.349	13.843	43.692	14.715	5.933
45.139	18.918	13.511	45.542	20.990	9.781	46.191	21.654	10.359
45.487	21.357	13.203	46.267	22.777	12.233	44.102	18.026	7.559
45.862	21.197	11.346	45.764	23.187	13.694	44.015	16.131	7.026
45.122	19.459	10.473	44.175	18.641	8.736	46.216	21.593	16.206
44.058	16.185	8.181	44.931	19.738	10.106	44.438	17.409	4.812
46.627	23.092	15.384	44.648	20.565	8.123	46.375	21.986	17.361
45.113	23.196	11.913	47.343	27.154	19.683	43.833	16.376	3.436
46.655	24.503	18.693	44.019	18.870	9.191	46.590	24.103	16.300
44.356	16.834	9.488	44.269	17.659	4.979	45.976	25.565	13.919
45.919	19.618	15.049	46.353	25.101	15.462	43.970	16.474	7.606
44.800	17.701	5.745	42.809	15.205	3.151	45.872	23.032	10.712
						43.604	16.210	3.852

Populacja:

Cechy:
X_1: wydatki na artykuły spożywcze
X_2: wydatki na papierosy
X_3: wydatki na alkohol

Założenie:
normalność rozkładów badanych cech

Techniki statystyczne:
1. Weryfikacja hipotezy $H_0 : \varrho_{1|2,3} = 0$
2. Weryfikacja hipotez

$$H_0 : \varrho_{12|3} = 0 \text{ oraz } H_0 : \varrho_{13|2} = 0$$

3. Dopasowanie odpowiedniej funkcji regresji

Poziom istotności $\alpha = 0.05$
Poziom ufności $1 - \alpha = 0.95$

Obliczenia

$$\bar{x}_1 = 45.090 \qquad \bar{x}_2 = 20.215 \qquad \bar{x}_3 = 10.496$$

$$\sum x_{1i}^2 = 203414.266951$$

$$\sum x_{2i}^2 = 41897.260651$$

$$\sum x_{3i}^2 = 12750.029343$$

$$\sum x_{1i}x_{3i} = 47697.150706$$

$$\sum x_{2i}x_{3i} = 22324.107946$$

$$\sum x_{1i}x_{2i} = 91428.748091$$

$$C = \begin{bmatrix} 1.0000 & 0.8717 & 0.8914 \\ 0.8717 & 1.0000 & 0.8267 \\ 0.8914 & 0.8267 & 1.0000 \end{bmatrix}$$

$$H_0 : \varrho_{1|2,3} = 0$$

Test współczynnika korelacji wielokrotnej

$$|C| = 0.0469$$

$$|C_{11}| = \left| \begin{bmatrix} 1.0000 & 0.8267 \\ 0.8267 & 1.0000 \end{bmatrix} \right| = 0.3165$$

$$R_{1|2,3} = \sqrt{1 - \frac{|C|}{|C_{11}|}} = \sqrt{1 - \frac{0.0469}{0.3165}} = 0.9230$$

Wartość krytyczna $r(0.05; 100, 3) = 0.2447$

Hipotezę odrzucamy

$$H_0 : \varrho_{12|3} = 0$$

Test współczynnika korelacji cząstkowej

$$\begin{aligned} R_{12|3} &= \frac{R_{12} - R_{13}R_{23}}{\sqrt{(1 - R_{13}^2)(1 - R_{23}^2)}} \\ &= \frac{0.8717 - 0.8914 \cdot 0.8267}{\sqrt{(1 - 0.8914^2)(1 - 0.8267^2)}} \\ &= 0.5284 \end{aligned}$$

Wartość krytyczna $r(0.05; 100 - 1) = 0.245$

$$H_0 : \varrho_{13|2} = 0$$

$$\begin{aligned} R_{13|2} &= \frac{R_{13} - R_{12}R_{23}}{\sqrt{(1 - R_{12}^2)(1 - R_{23}^2)}} \\ &= \frac{0.8914 - 0.8717 \cdot 0.8267}{\sqrt{(1 - 0.8717^2)(1 - 0.8267^2)}} \\ &= 0.6957 \end{aligned}$$

Ocena parametrów funkcji regresji

$$f(x_2, x_3) = \beta_0 + \beta_1 x_2 + \beta_2 x_3$$

$$\begin{cases} 1033.082880\widehat{\beta}_1 + 1106.175319\widehat{\beta}_2 = 278.818551 \\ 1106.175319\widehat{\beta}_1 + 1733.028891\widehat{\beta}_2 = 369.315683 \\ 45.090 - 20.215\widehat{\beta}_1 - 10.496\widehat{\beta}_2 = \widehat{\beta}_0 \end{cases}$$

$$\widehat{\beta}_1 = 0.131759 \qquad \widehat{\beta}_2 = 0.129004$$

$$\widehat{\beta}_0 = 41.072954$$

Wariancja resztowa

$$S^2 = \frac{99.038107 - 278.818551\widehat{\beta}_1 - 369.315683\widehat{\beta}_2}{n-3}$$

$$= 0.151115$$

Przedziały ufności (poziom ufności 0.95)

Wariancje estymatorów $\widehat{\beta}_0$, $\widehat{\beta}_1$ oraz $\widehat{\beta}_2$

$$S^2_{\beta_0} = 0.309071$$

$$S^2_{\beta_1} = 0.021496 \qquad S^2_{\beta_2} = 0.016597$$

$$\beta_0 \in (40.45953, 41.68637)$$

$$\beta_1 \in (0.089094, 0.174423)$$

$$\beta_2 \in (0.096063, 0.161944)$$