Analiza wielowymiarowa

Analiza korelacji

Zmienne kanoniczne

Analiza składowych głównych

Analiza czynnikowa

Analiza skupień

Analiza dyskryminacyjna

Analiza korelacji

Obserwujemy wektor losowy $(Y, X_1, \ldots, X_k)$

Macierz kowariancji

$$\mathcal{C} = \begin{bmatrix} C_{YY} & C_{YX_1} & C_{YX_2} & \cdots & C_{YX_k} \\ C_{YX_1} & C_{X_1X_1} & C_{X_1X_2} & \cdots & C_{X_1X_k} \\ \vdots & \vdots & \vdots & & \vdots \\ C_{YX_k} & C_{X_1X_k} & C_{X_2X_k} & \cdots & C_{X_kX_k} \end{bmatrix}$$

Macierz korelacji

$$\mathcal{R} = \begin{bmatrix} 1 & r_{YX_1} & r_{YX_2} & \cdots & r_{YX_k} \\ r_{YX_1} & 1 & r_{X_1X_2} & \cdots & r_{X_1X_k} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{YX_k} & r_{X_1X_k} & r_{X_2X_k} & \cdots & 1 \end{bmatrix}$$

$$\frac{1}{\sqrt{(2\pi)^n \det \mathcal{C}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\mathcal{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

Elipsoida koncentracji

$$(\boldsymbol{x}-\boldsymbol{\mu})'\mathcal{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \mathrm{const}$$

$$\mathcal{C} = \begin{bmatrix} \sigma_{11}^2 & \varrho\sigma_{11}\sigma_{22} \\ \varrho\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{bmatrix}$$

$$\frac{1}{2}\left(\sigma_{11}^2+\sigma_{22}^2-\sqrt{\left(\sigma_{11}^2-\sigma_{22}^2\right)^2+4\varrho^2\sigma_{11}^2\sigma_{22}^2}\right),$$

$$\frac{1}{2}\left(\sigma_{11}^2+\sigma_{22}^2+\sqrt{\left(\sigma_{11}^2-\sigma_{22}^2\right)^2+4\varrho^2\sigma_{11}^2\sigma_{22}^2}\right),$$

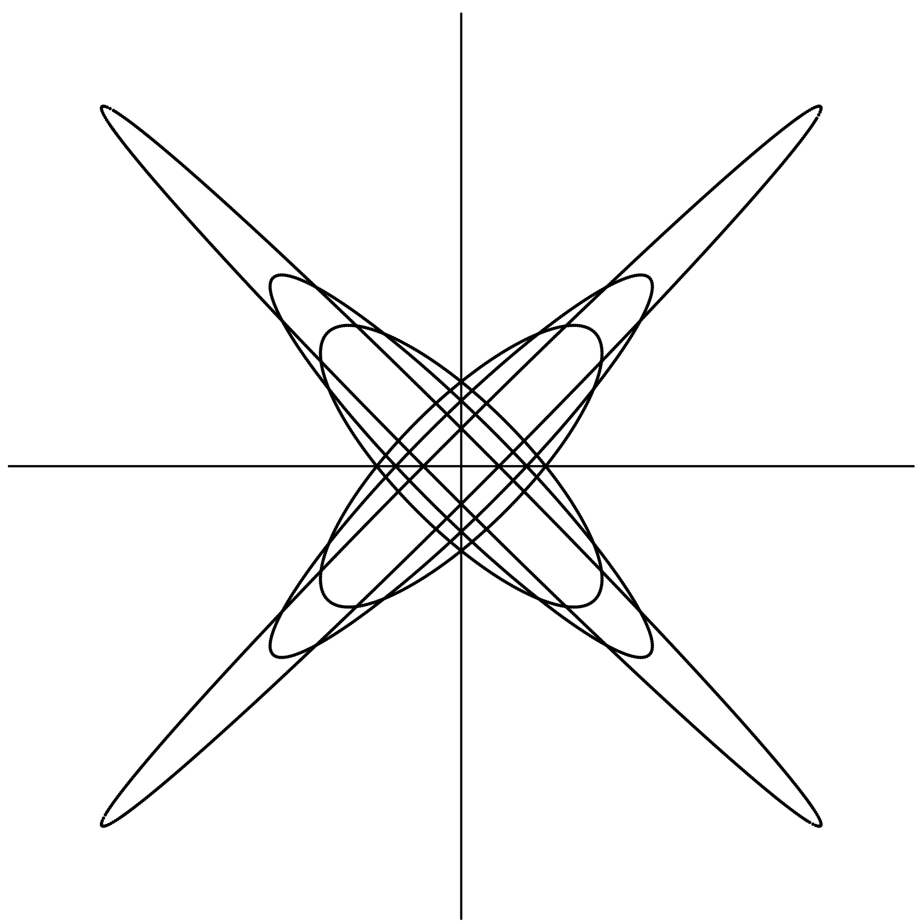
$$\left[\frac{\sigma_{11}^2-\sigma_{22}^2-\sqrt{\left(\sigma_{11}^2-\sigma_{22}^2\right)^2+4\varrho^2\sigma_{11}^2\sigma_{22}^2}}{2\varrho\sigma_{11}\sigma_{22}}, 1\right],$$

$$\left[\frac{\sigma_{11}^2-\sigma_{22}^2+\sqrt{\left(\sigma_{11}^2-\sigma_{22}^2\right)^2+4\varrho^2\sigma_{11}^2\sigma_{22}^2}}{2\varrho\sigma_{11}\sigma_{22}}, 1\right],$$

$$\mathcal{C} = \begin{bmatrix} 1 & \varrho \\ \varrho & 1 \end{bmatrix}$$

$$1 - \varrho \qquad 1 + \varrho$$

$$[-1, 1] \qquad [1, 1]$$



Współczynnik korelacji wielokrotnej
między Y a $(X_1, X_2, \ldots, X_k)$

$$r_{Y|X_1,X_2,\ldots,X_k} = \max_{\lambda_1,\lambda_2,\ldots,\lambda_k} r(Y, \lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k)$$

1. $r_{Y|X_1,X_2,\ldots,X_k} \in \langle 0, 1 \rangle$

2. $r_{Y|X_1,X_2,\ldots,X_k} = 0$:
Y średnio nie jest zależne od $(X_1, X_2, \ldots, X_k)$

3. $r_{Y|X_1,X_2,\ldots,X_k} = 1$:
Y jest liniową funkcją $(X_1, X_2, \ldots, X_k)$

$$r_{Y|X_1,X_2,\ldots,X_k} = \sqrt{1 - \frac{|\mathcal{R}|}{|\mathcal{R}_{11}|}}$$

. . . .

$$\frac{r^2_{Y|X_1,X_2,\ldots,X_k}}{1 - r^2_{Y|X_1,X_2,\ldots,X_k}} \cdot \frac{n-k+1}{k-1} \sim F_{k-1,n-k+1}$$

Współczynnik korelacji cząstkowej

$r_{YX_1(X_2 \ldots X_k)}$ opisuje zależność pomiędzy Y oraz X_1 z wyłączeniem pozostałych zmiennych

$$
\begin{aligned}
e_i &= Y_i - (\hat\beta_0 + \hat\beta_2 X_{2i} + \hat\beta_3 X_{3i} + \cdots + \hat\beta_k X_{ki}) \\
e_i^* &= X_{1i} - (\hat\alpha_0 + \hat\alpha_2 X_{2i} + \hat\alpha_3 X_{3i} + \cdots + \hat\alpha_k X_{ki}) \\
& \quad i = 1, 2, \ldots, n
\end{aligned}
$$

$$r_{YX_1(X_2 \ldots X_k)} = r_{ee^*}$$

$$\mathcal{R}^{-1} = \begin{bmatrix} r^{YY} & r^{YX_1} & r^{YX_2} & \ldots & r^{YX_k} \\ r^{YX_1} & r^{X_1X_1} & r^{X_1X_2} & \ldots & r^{X_1X_k} \\ \vdots & \vdots & & \vdots & \\ r^{YX_k} & r^{X_1X_k} & r^{X_2X_k} & \ldots & r^{X_kX_k} \end{bmatrix}$$

$$r_{YX_1(X_2 \ldots X_k)} = -\frac{r^{YX_1}}{\sqrt{r^{YY} r^{X_1X_1}}}$$

....

$$\frac{r_{YX_1(X_2 \ldots X_k)}}{\sqrt{1 - r^2_{YX_1(X_2 \ldots X_k)}}} \sqrt{n-k+1} \sim t_{n-k+1}$$

Liniowa funkcja regresji

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

$$\boldsymbol{\beta}' = [\beta_0, \beta_1, \ldots, \beta_k]$$

$$\hat{\boldsymbol{\beta}} = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\boldsymbol{y}$$

$$\boldsymbol{y}' = [Y_1, Y_2, \ldots, Y_n]$$

$$\mathcal{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \ldots & X_{k1} \\ 1 & X_{12} & X_{22} & \ldots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \ldots & X_{kn} \end{bmatrix}$$

Wariancja resztowa

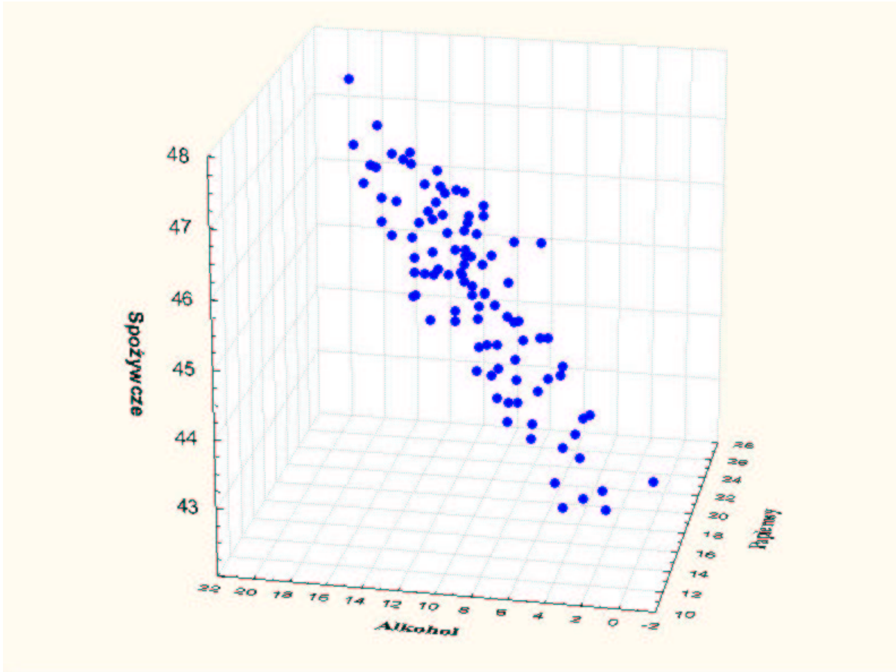
$$\frac{1}{n-k}\boldsymbol{y}'(\boldsymbol{I} - \mathcal{X}(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X})\boldsymbol{y}$$

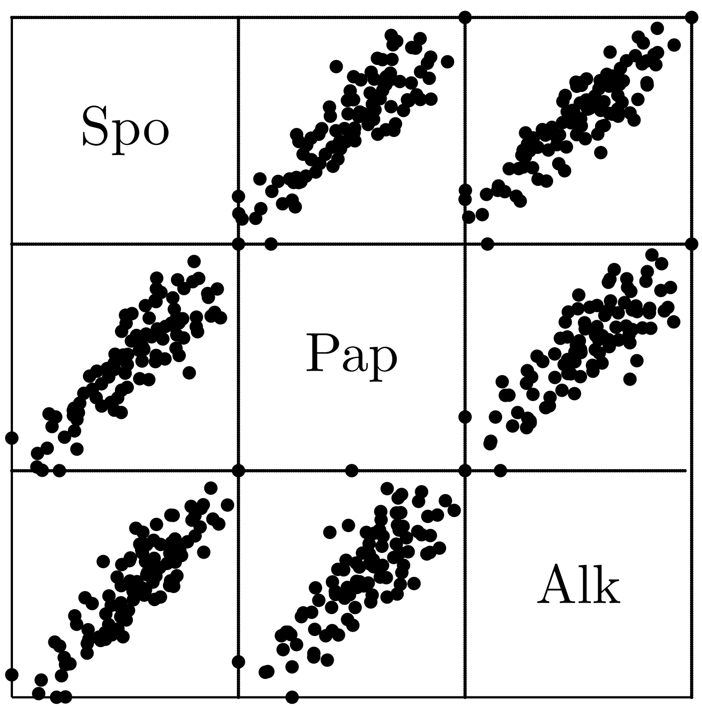
Przykład. Badano zależność wydatków na artykuły spożywcze, papierosy oraz alkohol. Zbadać istnienie zależności między obserwowanymi cechami. Jeżeli taka zależność istnieje, to dokonać ilościowego opisu zależności wydatków na artykuły spożywcze od wydatków na papierosy i alkohol.

Y: wydatki na artykuły spożywcze
X_1: wydatki na papierosy
X_2: wydatki na alkohol

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

$$\mathcal{R}^{-1} = \begin{bmatrix} 6.7557 & -2.8763 & -3.6442 \\ -2.8763 & 4.3835 & -1.0599 \\ -3.6442 & -1.0599 & 5.1247 \end{bmatrix}$$


Spożywcze
Alkohol
Papierowy
48
47
46
45
44
43
22 20 18 16 14 12 10 8 6 4 2 0 -2
10 12 14 16 18 20 22 24 26 28


Spo
Pap
Alk

Współczynnik korelacji wielokrotnej

$$|\mathcal{R}| = 0.0469$$

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

$$r_{Y|X_1,X_2} = \sqrt{1 - \frac{|\mathcal{R}|}{|\mathcal{R}_{11}|}} = \sqrt{1 - \frac{0.0469}{0.3165}} = 0.9230$$

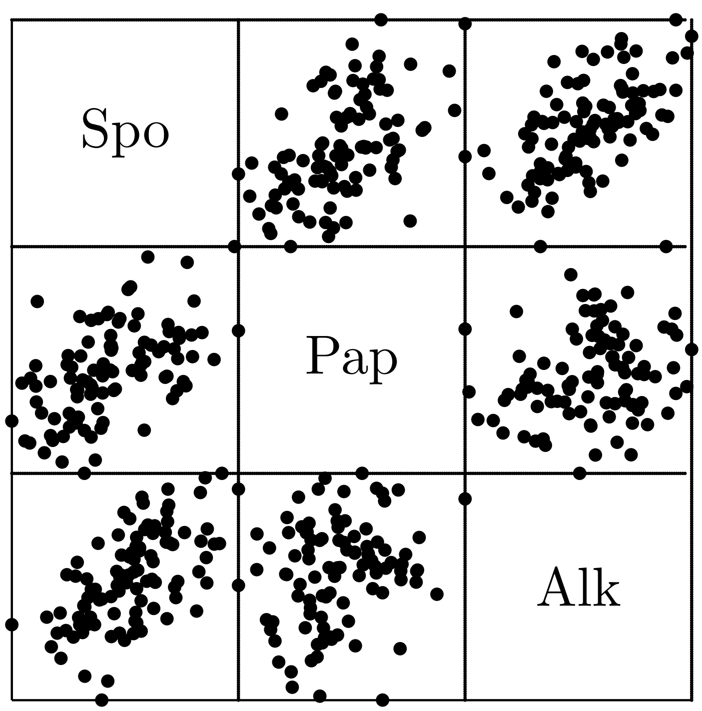
$$p - \text{value} \approx 0.003577$$

Współczynniki korelacji cząstkowych

$$r_{YX_1(X_2)} = -\frac{-2.8763}{\sqrt{6.7557 \cdot 4.3835}} = 0.5284$$

$$r_{YX_2(X_1)} = -\frac{-3.6442}{\sqrt{6.7557 \cdot 5.1247}} = 0.6957$$

$$r_{X_1X_2(Y)} = -\frac{-1.0599}{\sqrt{4.3835 \cdot 5.1247}} = 0.2236$$



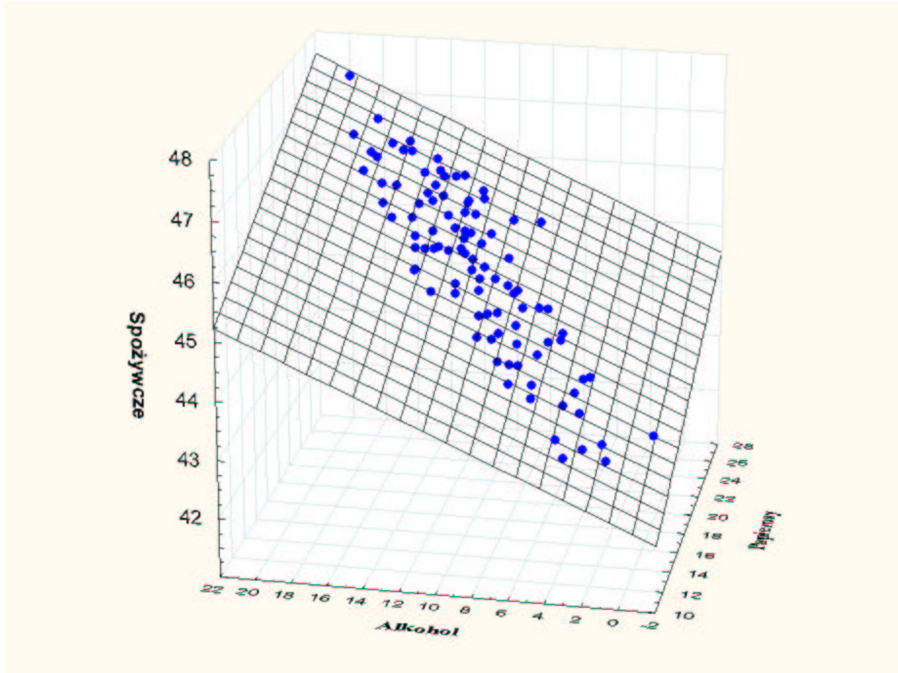
Ocena parametrów funkcji regresji

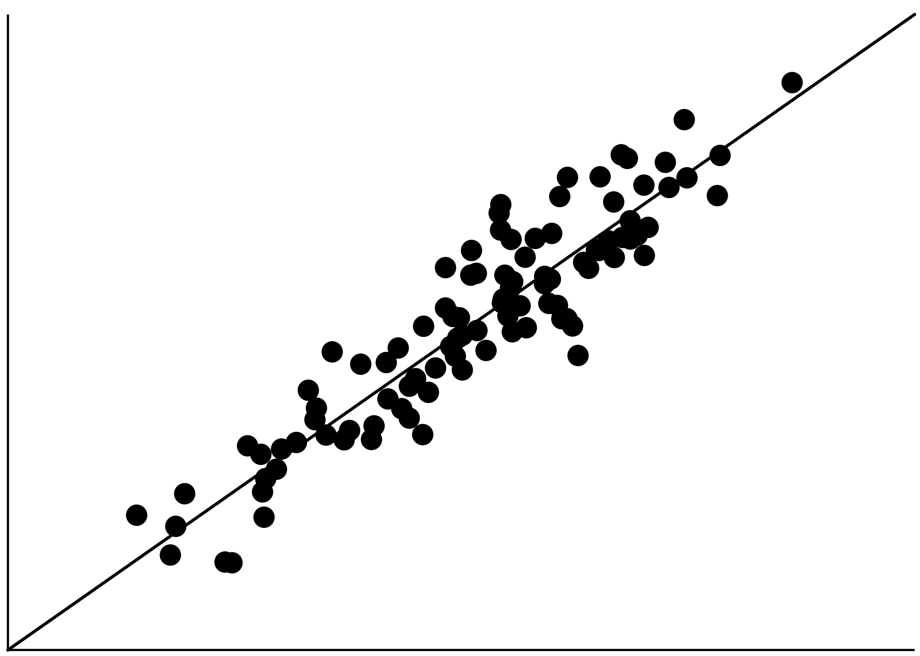
$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\widehat{\beta}_0 = 41.072954$$
$$\widehat{\beta}_1 = 0.131759$$
$$\widehat{\beta}_2 = 0.129004$$

Wariancja resztowa

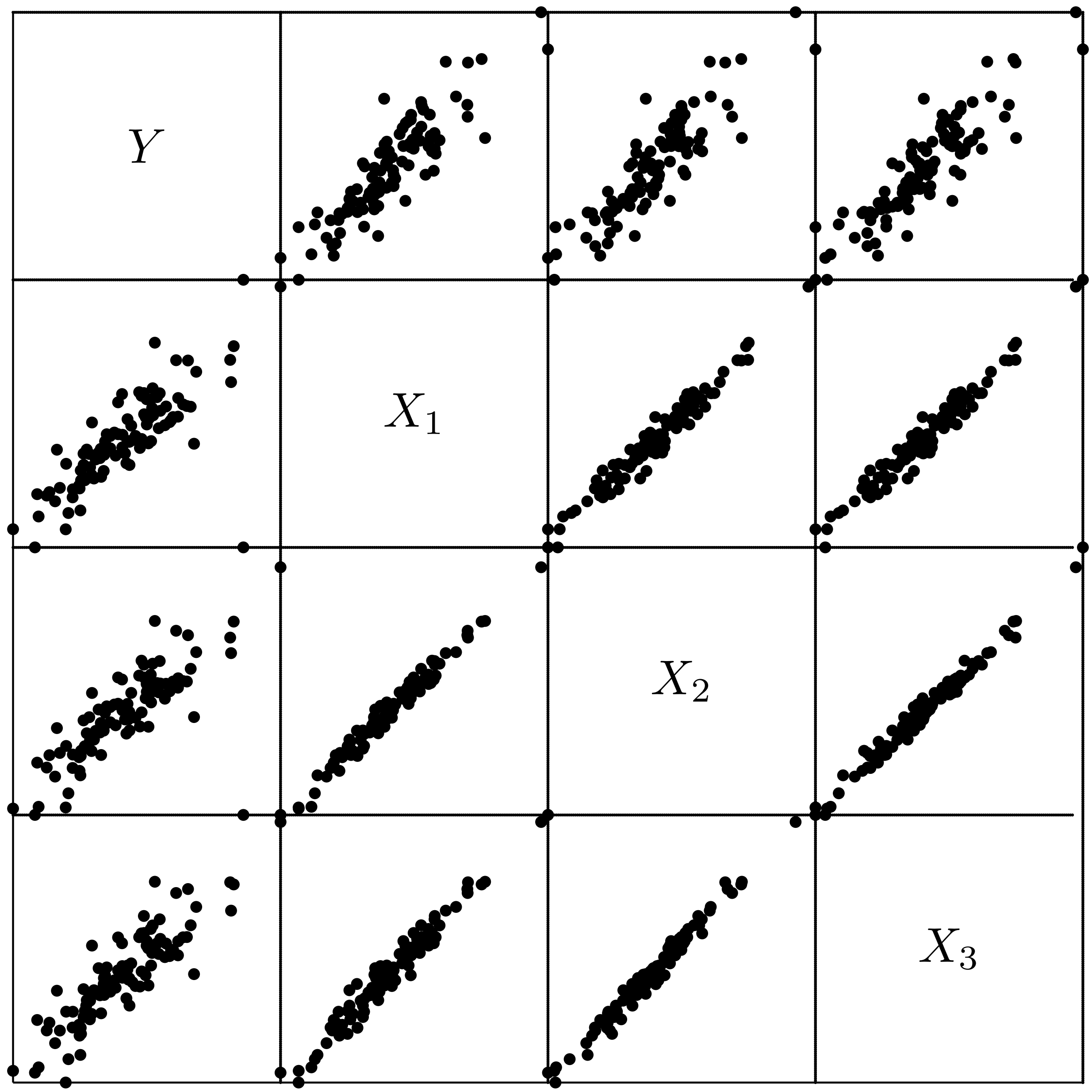
$$S^2 = 0.151115$$


Spożywcze
Alkohol
48
47
46
45
44
43
42
22 20 18 16 14 12 10 8 6 4 2 0 -2
10 12 14 16 18 20 22 24 26 28



. . . .

Przykład. Badano zależność między zmienną Y a zmiennymi X_1, X_2, X_3. Dokonano $n = 100$ obserwacji.



$$\mathcal{R} = \begin{bmatrix} 1.000 & 0.884 & 0.885 & 0.887 \\ 0.884 & 1.000 & 0.988 & 0.980 \\ 0.885 & 0.988 & 1.000 & 0.990 \\ 0.887 & 0.980 & 0.990 & 1.000 \end{bmatrix}$$

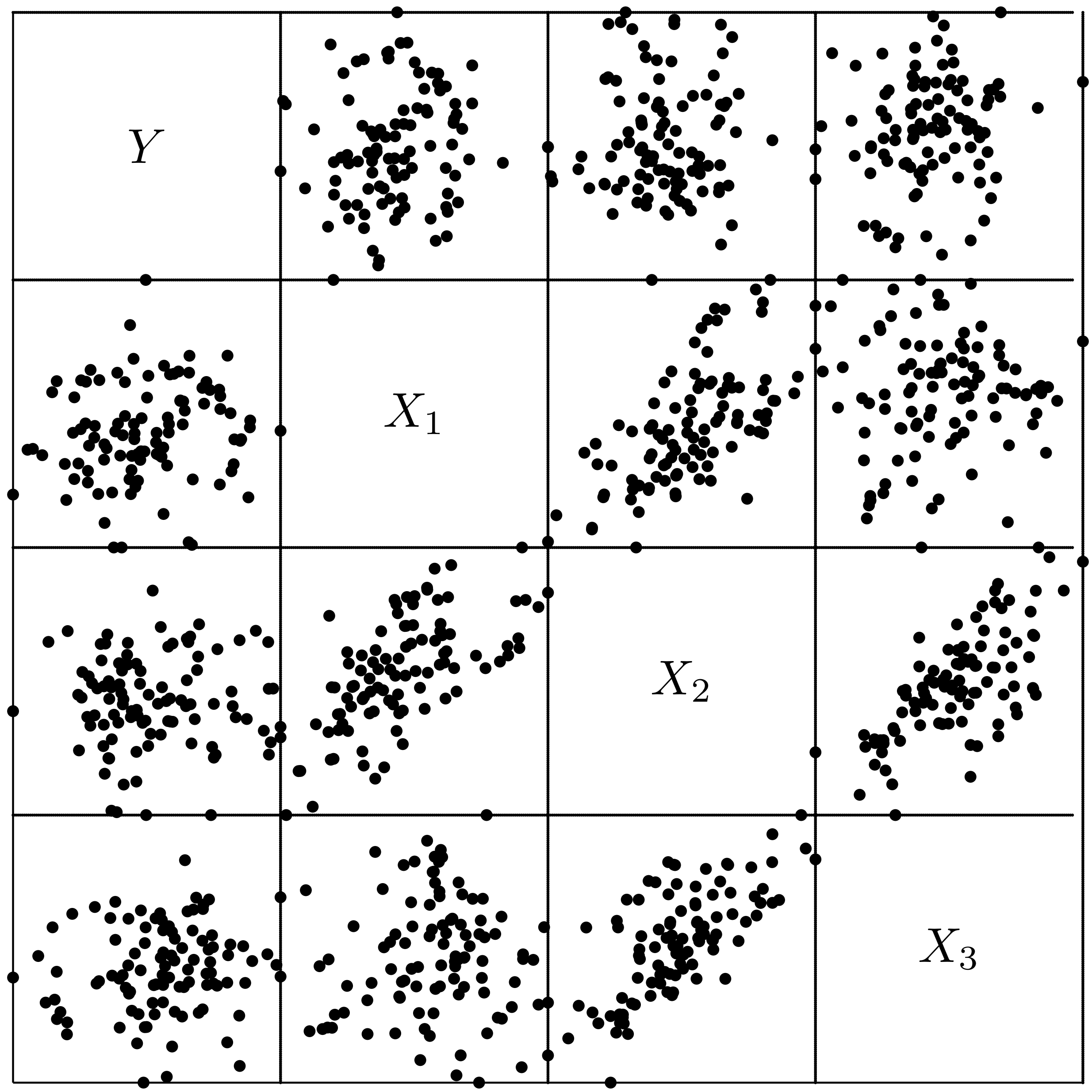
$$r_{Y|X_1,X_2,X_3} = 0.8902$$

$$\mathcal{R}^{-1} = \begin{bmatrix} 4.819 & -1.775 & -0.083 & -2.453 \\ -1.775 & 44.187 & -40.182 & -1.940 \\ -0.083 & -40.182 & 87.816 & -47.497 \\ -2.453 & -1.940 & -47.497 & 52.103 \end{bmatrix}$$

$$r_{YX_1(X_2,X_3)} = 0.1217$$

$$r_{YX_2(X_1,X_3)} = 0.0040$$

$$r_{YX_3(X_1,X_2)} = 0.1548$$


Y
X1
X2
X3