Podstawy regresji wielorakiej — ściąga

(por. skrypt, rozdział 6.3.1).

W oparciu o próbę statystyczną prostą postaci $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \dots (X_n, Y_n, Z_n)$ modelujemy zależność

$$Z \approx \hat{Z} = aX + bY + c. \tag{1}$$

Zależność (1) można przedstawić w postaci:

$$\mathbb{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \dots \\ Z_n \end{bmatrix} = \mathbb{X} \begin{bmatrix} c \\ b \\ a \end{bmatrix} + \epsilon = \begin{bmatrix} 1 & Y_1 & X_1 \\ 1 & Y_2 & X_2 \\ & \dots & \\ 1 & Y_n & X_n \end{bmatrix} \begin{bmatrix} c \\ b \\ a \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix}, \tag{2}$$

Jeśli kolumny macierzy $\mathbb{X}$ są **niewspółliniowe** tzn. liniowo niezależne w sensie algebraicznym (stochastycznym niekoniecznie) wówczas macierz

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} n & \sum_{i=1}^n Y_i & \sum_{i=1}^n X_i \\ \sum_{i=1}^n Y_i & \sum_{i=1}^n Y_i^2 & \sum_{i=1}^n X_i Y_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i Y_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \tag{3}$$

jest nieosobliwa, a poprzez jej odwrócenie można wyznaczyć estymatory parametrów regresji

$$\begin{bmatrix} \hat{c} \\ \hat{b} \\ \hat{a} \end{bmatrix} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Z}. \tag{4}$$

Zauważmy, że jest to łatwo wyliczalne dla próbki $(x_1, y_1, z_1), (x_2, y_2, z_2), \dots (x_n, y_n, z_n)$ będącej realizacją wyjściowej próby statystycznej — można więc wyznaczyć wartości tych estymatorów.

Jako ocenę jakości dopasowania modelu oszacować należy współczynnik determinacji $\rho^2 = \frac{\mathrm{V}ar\hat{Z}}{\mathrm{V}arZ} = 1 - \frac{\mathrm{V}ar\epsilon}{\mathrm{V}arZ}$, a konkretniej — wyznaczyć jego estymator R^2.

Przy oznaczeniach (skąd znanych?) $\hat{z}_i = ax_i + by_i + c$,

$$\mathrm{SST} = \sum_{i=1}^n (z_i - \overline{z})^2, \quad \mathrm{SSR} = \sum_{i=1}^n (\hat{z}_i - \overline{z})^2, \quad \mathrm{SSE} = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (z_i - \hat{z}_i)^2$$

współczynnik R^2 wyznacza się ze wzoru

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}. \tag{5}$$

Względnie, ponieważ nieobciążonym estymatorem $\mathrm{V}ar\epsilon$ jest

$$S_\epsilon^2 = \sum_{i=1}^n \frac{\epsilon_i^2}{n-k-1} \approx \mathrm{SSE}/(n-k-1),$$

stosuje się **współczynnik skorygowany** $\bar{R}^2$,

$$\bar{R}^2 = 1 - \frac{\mathrm{SSE}/(n-k-1)}{\mathrm{SST}/(n-1)}, \tag{6}$$

gdzie k jest liczbą zmiennych w modelu — w opisanym przykładzie $k = 2$.

Idąc dalej, macierz $\mathbb{V}$ wariancji/kowariancji estymatorów regresji oblicza się przez

$$\mathbb{V} = (\mathbb{X}^T\mathbb{X})^{-1}S_\epsilon^2. \tag{7}$$

Umożliwia ona wyznaczanie przedziałów ufności dla parametrów regresji oraz testy statystyczne.

Testując hipotezę $H\colon a = 0$ przeciw kontrhipotezie $K\colon a \neq 0$ stosujemy test t, gdzie statystyka testowa $T = \frac{\hat{a}}{S_a}$ ma rozkład t-Studenta o $n-k-1$ stopniach swobody i obustronny zbiór krytyczny. Z kolei przedział ufności, na poziomie ufności $1-\alpha$ dla parametru a wyznacza się jako

$$\left[\hat{a} - t(1-\frac{\alpha}{2}, n-k-1)S_a, \hat{a} + t(1-\frac{\alpha}{2}, n-k-1)S_a\right]. \tag{8}$$

Posługując się analizą wariancji można również wnioskować o regresji jako całości testując hipotezę $H: \forall_i a_i = 0$ przeciw $K: \exists_i a_i \neq 0$. Wykonuje się wówczas test F, gdzie statystyka testowa $F = \frac{\mathrm{SSR}/k}{\mathrm{SSE}/(n-k-1)}$ ma rozkład Fischera-Snedecora o $(k, n-k-1)$ stopniach swobody.