Коваріаційний аналіз

Приклад 1. Залежна змінна η – кількісний показник ризику зараження та важкість перебігу COVID-19,
незалежна якісна змінна ζ_1 – група крові пацієнта,
незалежна якісна змінна ζ_2 – раса пацієнта,
незалежна якісна змінна ζ_3 – стать пацієнта,
незалежна кількісна змінна ξ_1 – вік,
незалежна кількісна змінна ξ_2 – рівень цукру у крові,
незалежна кількісна змінна ξ_3 – середня кількість викурених сигарет за добу, … .

Вступ

Коваріаційний аналіз — розділ аналізу даних, який займається побудовою математичних моделей істотних зв'язків між залежною кількісною змінною η та вектором незалежних якісних змінних $\vec{\zeta}$ та вектором незалежних кількісних змінних $\vec{\xi}$. Даний розділ аналізу даних є об'єднанням регресійного та дисперсійного аналізу даних

Постановка задачі

Нехай $\vec{\eta}$ має розмірність q', а $\vec{\xi}$ має розмірність p', k-те спостереження над η є $y(k)$, тоді математична модель коваріаційного аналізу матиме наступний вигляд:

$$y(k) = x_g^T(k)\,\alpha_g + x^T(k)\,\alpha + e(k),\ k=\overline{1,N} \qquad (1)$$

Зауваження. α має розмірність p, що дорівнює кількості регресорів, а α_g ("альфа-ге") має розмірність q, що дорівнює кількості невідомих параметрів дисперсійної моделі.

α, $e(k)$ може зал. від $x_g(k)$

$$y = X_g\alpha_g + X\alpha + e \qquad (2)$$

$$y = \begin{pmatrix} y(1) \\ \vdots \\ y(N) \end{pmatrix},\quad X_g = \begin{pmatrix} x_g^T(1) \\ \vdots \\ x_g^T(N) \end{pmatrix},\quad X = \begin{pmatrix} x^T(1) \\ \vdots \\ x^T(N) \end{pmatrix},\quad e = \begin{pmatrix} e(1) \\ \vdots \\ e(N) \end{pmatrix}$$

Класичний коваріаційний аналіз

Припущення класичного коваріаційного аналізу:

1. $e \sim N(\Theta, \sigma^2 E_N)$, $\sigma^2 > 0$
2. $\operatorname{rank}(X_g) = q$ (враховані лінійні обмеження)
   $\operatorname{rank}(X) = p$ (3)
3. $\alpha_g \in R^q$, $\alpha \in R^p$
4. Матриця X не залежить від матриці X_g

Двокроковий метод найменших квадратів

I крок:

$$y - X\alpha = X_g \alpha_g + e$$

Нехай $\alpha = \theta_p$

$$\hat{\alpha}_g(\theta) = (X_g^T X_g)^{-1} X_g^T y$$

S_e – залишкова сума квадратів

$$S_e(\theta) = \| y - X_g \hat{\alpha}_g \|^2 = \| y - X_g (X_g^T X_g)^{-1} X_g^T y \|^2 =$$

$Q = E - X_g (X_g^T X_g)^{-1} X_g^T$

Q – симетрична, ідемпотентна (на с/р).

$$= \| (E - X_g (X_g^T X_g)^{-1} X_g^T) y \|^2 = y^T Q^T Q y = y^T Q^2 y =$$
$$= y^T Q y = \| y \|_Q^2$$

II крок:

$$y \to y - X\alpha$$

$$S_e(\alpha) = (y - X\alpha)^T Q (y - X\alpha)$$

$$\hat{\alpha} = \arg\min_{\alpha} S_e(\alpha) = (X^T Q X)^{-1} X^T Q y \quad (4)$$

$$\hat{\alpha}_g(\hat{\alpha}) = (X_g^T X_g)^{-1} X_g (y - X\hat{\alpha}) \quad (5)$$

$$S_e(\hat{\alpha}) = (y - X\hat{\alpha})^T Q (y - X\hat{\alpha})$$