Ш. *Множинний t-метод.* Нехай M – кількість *a priori* обраних для аналізу контрастів. (Зазвичай контрасти вибирають для дослідження після експерименту.) Тоді наближена довірча область для множини цих вибраних контрастів $\left\{\sum_{i=1}^{I} c_i^{(j)} a_i\right\}_{j=1}^{M}$ з рівнем довіри, не менше ніж $(1-\gamma), \gamma > 0$, задається системою довірчих інтервалів для кожного контрасту $\sum_{i=1}^{I} c_i^{(j)} a_i$ з рівнем довіри $\left(1-\frac{\gamma}{M}\right), \gamma > 0$ і визначається таким чином:

$$\left|\sum_{i=1}^{I} c_i^{(j)} a_i - \sum_{i=1}^{I} c_i^{(j)} \overline{y}_{i\bullet}\right| \le \sqrt{\overline{S}_e \left(\sum_{i=1}^{I} \frac{\left(c_i^{(j)}\right)^2}{N_i}\right)} t_{\frac{\gamma}{2M}} (N-I) = \Delta_3^j, j = \overline{1,M}, \qquad (18)$$

де $t_\gamma(\nu)$ – 100γ відсоткова точка t-розподілу Стьюдента з ν ступенями свободи.

рівнем довіри $(1-\gamma), \gamma > 0$, в результаті чого отримали (для методів Шеффе та Тьюкі) відповідний довірчий інтервал виду $\left[\hat{k}-\Delta_i, \hat{k}+\Delta_i\right], \Delta_i > 0, i = \overline{1,2}$, де $\hat{k} = \sum_{i=1}^{I} c_i \overline{y}_{i\bullet}$.

Це дозволяє на другому кроці область прийняття гіпотези (13) для цих методів записати в такому вигляді:

$$\left(\hat{k}-\Delta_i\right)\left(\hat{k}+\Delta_i\right) \le 0 \Leftrightarrow \hat{k}^2 \le \Delta_i^2 \Leftrightarrow \left|\hat{k}\right| \le \Delta_i, i = 1,2,$$

тобто справедливість останньої нерівності означає, що відповідний контраст слід вважати таким, що незначимо відхиляється від нуля із

статистичної точки зору з рівнем значущості $\gamma > 0$, у протилежному випадку його треба вважати таким, що істотно відхиляється від нуля.

Остаточно область прийняття відповідної гіпотези матиме такий вигляд:

I. для методу Шеффе з урахуванням (15) одержуємо

$$\left|\sum_{i=1}^{I} c_i \overline{y}_{i\cdot}\right| \le \sqrt{\overline{S}_e \left(\sum_{i=1}^{I} \frac{c_i^2}{N_i}\right)(I-1) F_\gamma (I-1, N-I)} = \Delta_1,$$

II. для методу Тьюкі, якщо взяти до уваги (17), отримуємо

$$\left|\sum_{i=1}^{I} c_i \overline{y}_{i\cdot}\right| \le \frac{1}{2}\sum_{i=1}^{I} |c_i| \sqrt{\frac{\overline{S}_e}{N_0}}\, q_\gamma (I, N-I) = \Delta_2,$$

III. для множинного t-методу врахування (19) дозволяє записати

$$\left|\sum_{i=1}^{I} c_i^{(j)} \overline{y}_{i\cdot}\right| \le \sqrt{\overline{S}_e \left(\sum_{i=1}^{I} \frac{\left(c_i^{(j)}\right)^2}{N_i}\right)} t_{\frac{\gamma}{2M}} (N-I) = \Delta_3^j, \; j = \overline{1, M}.$$

η - залежна кількісна скалярна змінна,

а незалежні якісні скалярні змінні:

ζ_1 - фактор A, який набуває своїх значень з I_1 градацій,

ζ_2^{I} - фактор B, який набуває своїх значень з I_2 градацій.

Необхідно за спостереженнями над залежною змінною η при активних різних сполученнях градацій незалежних змінних ζ_1 та ζ_2 побудувати математичну модель залежності змінної η від змінних ζ_1 та ζ_2.

Фон: Приклад 2: η – врожайність зернової культури,

ζ_1 – сорт зернової культури, всього I_1 сортів,

ζ_2 – вид добрива, всього I_2 видів добрива.

Нехай при активному сполученні i-ї градації змінної ζ_1 та j-ї градації змінної ζ_2 доступно N_{ij} спостережень y_{ijk} над η, $i=\overline{1,I_1}$, $j=\overline{1,I_2}$, $k=\overline{1,N_{ij}}$ $\left(N_{ij}\geq 1\right)$. Тоді математичну *модель двофакторного дисперсійного аналізу* будемо шукати в такому вигляді:

$$y_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+e_{ijk},\qquad i=\overline{1,I_1},\ j=\overline{1,I_2},\ k=\overline{1,N_{ij}}\ \left(N_{ij}\geq 1\right),\quad (20)$$

де

y_{ijk} – k-те спостереження над η при активному сполученні i-ї градації змінної ζ_1 та j-ї градації змінної ζ_2,

μ – загальне середнє всіх спостережень у деякому розумінні,

α_i – *кількісний вираз відносного впливу i-ї градації змінної ζ_1 на η відносно* μ (або, іншими словами, *головний ефект i-го рівня фактора A*),

β_j – *кількісний вираз відносного впливу j-ї градації змінної ζ_2 на η відносно* μ (або, іншими словами, *головний ефект j-го рівня фактора B*),

γ_{ij} – *кількісний вираз відносного впливу взаємодії i-ї градації змінної ζ_1 та j-ї градації змінної ζ_2 на η відносно* μ (або, іншими словами, *взаємодія i-го рівня фактора A та j-го рівня фактора B*),

e_{ijk} – похибка моделі k-го спостереження над η при активному сполученні i-ї градації змінної ζ_1 та j-ї градації змінної ζ_2.

Причому всього спостережень доступно в кількості

$$N=\sum_{i=1}^{I_1}\sum_{j=1}^{I_2}N_{ij}.$$

У свою чергу, *кількісний вираз абсолютного впливу i-ї градації змінної* ζ_1 **на** η **дорівнює**

$$a_i = \mu + \alpha_i, \quad i = \overline{1, I_1},$$

кількісний вираз абсолютного впливу j-ї градації змінної ζ_2 **на** η **має вигляд**

$$b_j = \mu + \beta_j, \quad j = \overline{1, I_2},$$

а *кількісний вираз абсолютного впливу взаємодії i-ї градації змінної* ζ_1 *та j-ї градації змінної* ζ_2 *на* η **визначається як**

$$c_{ij} = \mu + \gamma_{ij}, \quad i = \overline{1, I_1}, j = \overline{1, I_2}.$$

Припустимо, що похибки e_{ijk} моделі (20) є:

- $e_{ijk} \sim \mathcal{N}\left(0, \sigma^2\right), \sigma^2 > 0, \quad \forall i, j, k;$
- $\left\{e_{ijk}\right\}$ - незалежні.

Необхідно за доступними скалярними спостереженнями $\left\{y_{ijk},\ i = \overline{1, I_1},\ j = \overline{1, I_2},\ k = \overline{1, N_{ij}}\ \left(N_{ij} \geq 1\right)\right\}$ знайти оцінки невідомих параметрів:

$$\begin{aligned} &\mu, \\ &\alpha_1, \alpha_2, \ldots, \alpha_{I_1}, \\ &\beta_1, \beta_2, \ldots, \beta_{I_2}, \\ &\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1I_2}, \\ &\gamma_{21}, \gamma_{22}, \ldots, \gamma_{2I_2}, \\ &\quad \ldots \\ &\gamma_{I_1 1}, \gamma_{I_1 2}, \ldots, \gamma_{I_1 I_2} \end{aligned} \tag{21}$$

математичної моделі (20). Шукаємо за допомогою МНК. Загальна кількість невідомих параметрів буде дорівнювати

$$1 + I_1 + I_2 + I_1 I_2 = \left(1 + I_1\right)\left(1 + I_2\right).$$

Далі схема розв'язання задачі двофакторного дисперсійного аналізу повністю аналогічна процедурі розв'язання задачі однофакторного дисперсійного аналізу.

Спочатку модель (20) переписується, як і раніше, а саме:

$$y = X\alpha + e,$$

причому

y – вектор-стовпчик з усіх спостережень y_{ijk},

α – вектор-стовпчик з усіх невідомих параметрів (21),

X – матриця відповідної розмірності, елементи кожного рядка якої всі дорівнюють нулю, окрім першого та трьох інших, що відповідають місцезнаходженню відповідних параметрів головних ефектів та попарної взаємодії у векторі α,

e – вектор-стовпчик з усіх похибок моделі

$\left\{e_{ijk},\ i = \overline{1, I_1},\ j = \overline{1, I_2},\ k = \overline{1, N_{ij}}\ \left(N_{ij} \geq 1\right)\right\}$.

У цьому випадку матриця X буде мати неповний ранг. Тому, щоб скористатися МНК при визначенні оцінок вектора α, необхідно врахувати додаткові лінійні обмеження, які справедливі для нього.

Дійсно, враховуючи зміст невідомих параметрів (21), можна стверджувати, що:

$$\exists \left\{v_i\right\}_{i=1}^{I_1}, \left\{w_j\right\}_{j=1}^{I_2} : \forall i\ \ v_i > 0, \forall j\ \ w_j > 0,$$

$$\begin{cases} \sum\limits_{i=1}^{I_1} v_i \alpha_i = 0, \\ \sum\limits_{j=1}^{I_2} w_j \beta_j = 0, \\ \sum\limits_{i=1}^{I_1} v_i \gamma_{ij} = 0, \quad j = \overline{1, I_2}, \\ \sum\limits_{j=1}^{I_2} w_j \gamma_{ij} = 0, \quad i = \overline{1, I_1}. \end{cases} \tag{22}$$

Визначення вагових коефіцієнтів $\{v_i\}_{i=1}^{I_1}$ та $\{w_j\}_{j=1}^{I_2}$ здійснюється відповідно до змісту конкретної постановки задачі.

Врахування лінійних обмежень (22) дозволяє однозначно визначити оцінку $\hat{\alpha}$ методом найменших квадратів у математичній моделі (20) за спостереженнями $\left\{y_{ijk}, i=\overline{1,I_1},\ j=\overline{1,I_2}, k=\overline{1,N_{ij}}\ \left(N_{ij}\geq 1\right)\right\}$.

Окрім цього, цікавою є перевірка на значимість параметрів моделі двофакторного дисперсійного аналізу, і насамперед – перевірка з деяким рівнем значущості $\gamma > 0$ таких гіпотез:

$$H_0^A : \alpha_1 = \alpha_2 = \ldots = \alpha_{I_1} = 0, \tag{23}$$

$$H_0^B : \beta_1 = \beta_2 = \ldots = \beta_{I_2} = 0, \tag{24}$$

$$H_0^{AB} : \gamma_{ij} = 0, \quad i=\overline{1,I_1},\ j=\overline{1,I_2}. \tag{25}$$

Для розв'язання цих задач достатньо використати той самий математичний апарат, що й при розв'язанні відповідних задач в однофакторному дисперсійному аналізі. Проте формули стануть більш громіздкими. Для їх спрощення наведемо розв'язок для випадку, коли справедливо

$$N_{ij} = N_0\ \left(N_0 \geq 1\right), \quad i=\overline{1,I_1},\ j=\overline{1,I_2}, \tag{26}$$

а відповідні вагові коефіцієнти $\{v_i\}_{i=1}^{I_1}, \{w_j\}_{j=1}^{I_2}$ усі однакові

$$v_i = \frac{1}{I_1}, \quad i=\overline{1,I_1}; \qquad w_j = \frac{1}{I_2}, \quad j=\overline{1,I_2}. \tag{27}$$

Тепер загальна кількість спостережень визначатиметься таким чином:

$$N = I_1 I_2 N_0.$$

У результаті використання методу найменших квадратів для визначення невідомих параметрів (21) у математичній моделі (20) за наявності лінійних обмежень (22) при справедливості припущень (26), (27) отримуємо такі їх оцінки:

$$\hat{\mu} = \overline{y},$$

$$\hat{\alpha}_i = \overline{y}_{i\cdot\cdot} - \overline{y}, \quad i = \overline{1, I_1},$$

$$\hat{\beta}_j = \overline{y}_{\cdot j\cdot} - \overline{y}, \quad j = \overline{1, I_2},$$

$$\hat{\gamma}_{ij} = \overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}, \quad i = \overline{1, I_1}, j = \overline{1, I_2},$$

де

$$\overline{y} = \frac{1}{N} \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \sum_{k=1}^{N_0} y_{ijk},$$

$$\overline{y}_{ij\cdot} = \frac{1}{N_0} \sum_{k=1}^{N_0} y_{ijk}, \quad i = \overline{1, I_1}, \; j = \overline{1, I_2},$$

$$\overline{y}_{i\cdot\cdot} = \frac{1}{I_2 N_0} \sum_{j=1}^{I_2} \sum_{k=1}^{N_0} y_{ijk}, \quad i = \overline{1, I_1},$$

$$\overline{y}_{\cdot j\cdot} = \frac{1}{I_1 N_0} \sum_{i=1}^{I_1} \sum_{k=1}^{N_0} y_{ijk}, \quad j = \overline{1, I_2}.$$

Таблиця двофакторного дисперсійного аналізу

Проведемо аналіз *повної суми квадратів відхилень спостережень* y_{ijk} *від загального середнього* $\overline{y}$. Дійсно,

$$\sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \sum_{k=1}^{N_0} \left(y_{ijk} - \overline{y}\right)^2 = \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \sum_{k=1}^{N_0} \Big[\left(y_{ijk} - \overline{y}_{ij\cdot}\right) +$$

$$+ \left(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}\right) + \left(\overline{y}_{i\cdot\cdot} - \overline{y}\right) + \left(\overline{y}_{\cdot j\cdot} - \overline{y}\right) \Big]^2 =$$

$$= \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \sum_{k=1}^{N_0} \left(y_{ijk} - \overline{y}_{ij\cdot}\right)^2 +$$

$$+ N_0 \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \left(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}\right)^2 + \qquad \text{(на с/р)}$$

$$+ I_2 N_0 \sum_{i=1}^{I_1} \left(\overline{y}_{i\cdot\cdot} - \overline{y}\right)^2 + I_1 N_0 \sum_{j=1}^{I_2} \left(\overline{y}_{\cdot j\cdot} - \overline{y}\right)^2.$$

Останнє перетворення справедливо в силу того, що в усіх подвійних добутках квадратні дужки дорівнюють нулеві. Отриманий результат скорочено можна записати таким чином:

$$S = S_e + S_A + S_B + S_{AB}, \tag{28}$$

де

$$S = \sum_{i=1}^{I_1}\sum_{j=1}^{I_2}\sum_{k=1}^{N_0}\left(y_{ijk} - \bar{y}\right)^2, S_e = \sum_{i=1}^{I_1}\sum_{j=1}^{I_2}\sum_{k=1}^{N_0}\left(y_{ijk} - \bar{y}_{ij\cdot}\right)^2,$$

$$S_A = I_2 N_0 \sum_{i=1}^{I_1}\left(\bar{y}_{i\cdot\cdot} - \bar{y}\right)^2, S_B = I_1 N_0 \sum_{j=1}^{I_2}\left(\bar{y}_{\cdot j\cdot} - \bar{y}\right)^2,$$

$$S_{AB} = N_0 \sum_{i=1}^{I_1}\sum_{j=1}^{I_2}\left(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}\right)^2.$$

Отже, для S, *повної суми квадратів відхилень спостережень* y_{ijk} *від загального середнього* $\bar{y}$, у випадку двофакторної моделі дисперсійного аналізу отримали розклад (28), аналогічний розкладу (12), отриманому для однофакторної моделі.

Результати двофакторного дисперсійного аналізу також заносять у відповідну *таблицю 2.1 двофакторного дисперсійного аналізу.*

Таблиця 2.1 двофакторного дисперсійного аналізу

Джерело варіації	**Сума квадратів**	**Кількість ступенів свободи**	**Середня сума квадратів**	***F*-статистика**	$\gamma_{\max}$
головні ефекти фактора A	S_A	$I_1 - 1$	$\bar{S}_A = \dfrac{S_A}{I_1 - 1}$	$F_A = \dfrac{\bar{S}_A}{\bar{S}_e}$	γ_A
головні ефекти фактора B	S_B	$I_2 - 1$	$\bar{S}_B = \dfrac{S_B}{I_2 - 1}$	$F_B = \dfrac{\bar{S}_B}{\bar{S}_e}$	γ_B
взаємодії факторів A та B	S_{AB}	$(I_1 - 1)(I_2 - 1)$	$\bar{S}_{AB} = \dfrac{S_{AB}}{(I_1 - 1)(I_2 - 1)}$	$F_{AB} = \dfrac{\bar{S}_{AB}}{\bar{S}_e}$	γ_{AB}
помилки	S_e	$I_1 I_2 (N_0 - 1)$	$\bar{S}_e = \dfrac{S_e}{I_1 I_2 (N_0 - 1)}$		
	S	$N - 1$			

В останньому рядку табл. 2.1, таблиці двофакторного дисперсійного аналізу, наведено суми по другому та третьому стовпчикам відповідно. Зауважимо, що результат у другому стовпчику збігається з уже отриманим результатом (28). Підраховані в таблиці значення $\gamma_A, \gamma_B, \gamma_{AB}$ – значення максимальних рівнів значущості γ, за яких гіпотези (23), (24), (25) будуть справедливі, а використання статистик F_A, F_B, F_{AB} дозволяє записати області прийняття гіпотез H_0^A, H_0^B, H_0^{AB}, відповідно:

для гіпотези H_0^A:

$$F_A < F_\gamma\left(I_1 - 1, I_1 I_2 (N_0 - 1)\right),$$

для гіпотези H_0^B:

$$F_B < F_\gamma\left(I_2 - 1, I_1 I_2 (N_0 - 1)\right),$$

для гіпотези H_0^{AB}:

$$F_{AB} < F_\gamma\left((I_1 - 1)(I_2 - 1), I_1 I_2 (N_0 - 1)\right),$$

де $F_\gamma(\upsilon_1, \upsilon_2)$ – 100γ відсоткова точка F-розподілу з параметрами υ_1 та υ_2.

Багатофакторний дисперсійний аналіз

...

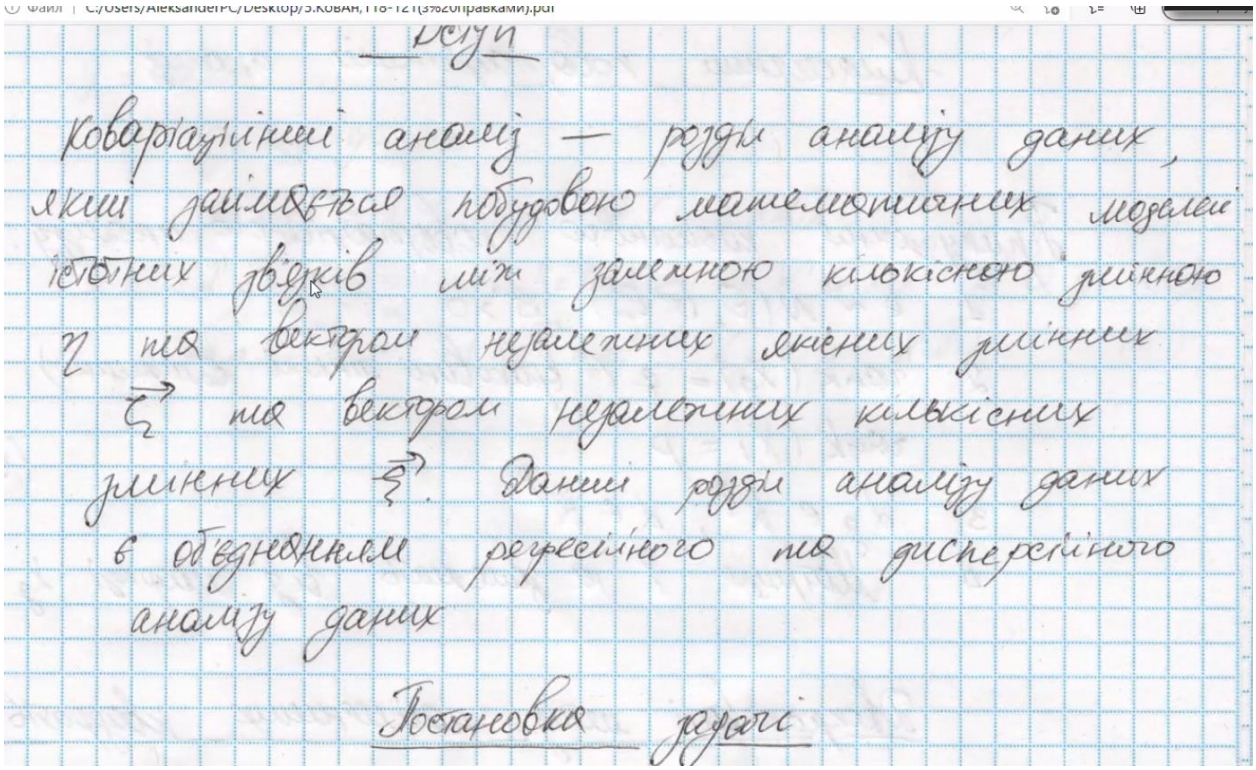
Самостійна робота №6. З навчального посібника «Слабоспицький О.С. Дисперсійний аналіз даних, 2013» пропрацювати матеріал наведений у **Розділі 3.**: «Багатофакторний дисперсійний аналіз».

(пропустити 5 стор.)

[redacted]

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, … .


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

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

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

$\alpha, e(k)$ можуть зал. від $x_g(k)$

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

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

Класичний коваріаційний аналіз

Припущення класичного коваріаційного аналізу:

1. $e \sim N(0, \sigma^2 E_N)$, $\sigma^2 > 0$
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

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

$$\left| \begin{array}{l} Q = E - X_g (X_g^T X_g)^{-1} X_g^T \\ Q - \text{симетрична, ідемпотентна (на с/р).} \end{array} \right.$$

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

$$= y^T Q y = \| y \|_Q^2$$

II крок:

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

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

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}}\ S_e(\alpha) = (X^T Q X)^{-1} X^T Q y \quad (4)$$

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

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