4 БАГАТОФАКТОРНА ЛІНІЙНА РЕГРЕСІЯ

Багатофакторна лінійна регресія є узагальненням простої лінійної регресії, яка розглядалась вище. У багатофакторній регресії ми припускаємо, що на залежну змінну *y* може впливати більше, ніж один фактор.

Основні припущення моделі багатофакторної лінійної регресії:

1) випадкова величина u (неспостережувана випадкова величина) має нормальний розподіл з нульовим математичним сподіванням і сталою дисперсією;

2) значення випадкової величини u не залежить від величини m, яка визначає кількість незалежних змінних, що входять в модель;

3) випадкові величини u_i і є статистично незалежними;

4) факторні ознаки x_i є лінійно незалежними (у випадку лінійної залежності отримуємо явище мультиколінеарності).

Методи побудови багатофакторної регресійної моделі:

1. Метод усіх можливих регресій.

Метод усіх можливих регресій – історично перший метод побудови лінійних регресійних моделей і найбільш громіздкий серед усіх методів. Ідея методу полягає у побудові множини регресійних рівнянь, які містять усі можливі комбінації попередньо відібраних факторів, і у порівнянні цих рівнянь за трьома критеріями: коефіцієнтом детермінації R^2, стандартною похибкою $\hat{\sigma}_u$ і критерієм Меллоуза C_p. У загальному випадку для m відібраних факторів (пояснюючих змінних) можна побудувати 2^m рівнянь регресії і виконати їх порівняння.

Побудова і аналіз усіх можливих регресійних рівнянь є доволі громіздка і ненадійна процедура, тому цей метод рекомендується використовувати при невеликій кількості відібраних факторів.

2. Покроковий регресійний метод.

Цей метод є найпоширенішим на практиці і більш економним у порівнянні з попереднім. Ідея методу полягає у послідовному включенні до моделі факторів (пояснюючих змінних) до тих пір, поки модель не стане задовільною. Порядок включення факторів до моделі вибирається на основі значень коефіцієнтів парної кореляції між пояснюючими і залежною змінною моделі. Алгоритм методу покрокової кореляції можна подати у наступному вигляді:

1) розраховується кореляційна матриця для усіх змінних моделі, які планується включити до моделі;

2) спочатку з кореляційної матриці вибирається і включається до моделі той фактор x_i, $i=\overline{1,m}$, якому у кореляційній матриці відповідає найбільший за модулем коефіцієнт парної кореляції з залежною змінної моделі *y* (нехай це буде змінна x_1). Будується регресійне рівняння з однією незалежною змінною $\hat{y}=a_0+a_1x_1$ і для нього обчислюється коефіцієнт

детермінації. Після цього перевіряється чи значима ця змінна за коефіцієнтом детермінації і за частковим *F*– критерієм. Якщо ні, то приймаємо $y = \bar{y}$ і процес побудови моделі припиняється. Якщо так, то переходимо до наступного кроку;

3) на основі аналізу кореляційної матриці серед тих пояснюючих змінних, що залишились, шукаємо нову змінну, яка має найбільший за модулем коефіцієнт кореляції з *y* і включаємо її до моделі (нехай це буде змінна x_2);

4) будується нове рівняння регресії:

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2$$

і для нього розраховується звичайний R^2 оцінений $\bar{R}^2$ коефіцієнти детермінації. Аналізується зміна цих показників у порівнянні з попередньою моделлю. Потім розраховуються часткові *F*– критерії для кожного фактора. Серед них обирається найменше значення і порівнюється із заздалегідь обраним критичним значенням *F*– критерію. В залежності від результатів перевірки додана на цьому кроці змінна або залишається у моделі, або відкидається;

5) після цього модель перераховується в залежності від факторів, які залишились і здійснюється перехід до кроку 3.

Процес побудови моделі за наведеним алгоритмом припиняється, якщо жодний фактор, що знаходиться у поточному рівнянні , не вдається виключити, а новий претендент на включення не відповідає частковому *F*– критерію.

3. Метод виключень.

Метод виключень діє у зворотному порядку порівняно з методом покрокової регресії і є також досить поширеним. Загальний алгоритм методу складається з 5 кроків:

1) будується рівняння регресії, яке включає всі відібрані фактори. Якщо попередньо було відібрано *m* факторів, то вихідне базове рівняння має вигляд:

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m;$$

2) для кожного фактора (пояснюючої змінної) x_i, $i = \overline{1, m}$ обчислюється значення часткового *F*– критерію;

3) серед розрахованих значень часткового *F*– критерію вибирається найменше $F_{\min}$ і порівнюється із заздалегідь обраним критичним значенням розподілу Фішера $F_{кр}$;

4) якщо $F_{\min} < F_{кр}$, то відповідний фактор виключається з рівняння. Проводиться новий розрахунок регресійного рівняння вже без виключеного фактора і здійснюється перехід знову до кроку 2;

5) якщо $F_{\min} > F_{кр}$, то регресійне рівняння залишається без змін.

Невідомі параметри багатофакторної регресії обчислюються за допомогою методу найменших квадратів, суть якого полягає в мінімізації суми квадратів помилок.

Вплив кожного фактора (кожної незалежної змінної) на залежну змінну характеризується частковими коефіцієнтами регресії (параметрами). Частковий коефіцієнт регресії показує, на скільки одиниць зміниться значення залежної

змінної при збільшенні значення відповідного фактора на одиницю при умові, що значення всіх інших факторів залишатимуться постійними.

Значущість всіх параметрів одночасно перевіряється за допомогою *F* – критерію, а кожного окремо – за допомогою *t* –критерію (для простої лінійної регресії *F* – та *t* –тести були еквівалентними, для багатофакторної моделі це не так).

Основні формули

1. *Узагальнена багатофакторна регресійна модель:*

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_m x_m + u .$$

2. *Вибіркова багатофакторна регресійна модель:*

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m .$$

3. *Оцінювання невідомих параметрів у багатофакторній регресії.*

З матричної форми вибіркової багатофакторної регресійної моделі:

$$Y = XA + U ,$$

будемо мати:

$$A = \left(X^t \cdot X \right)^{-1} \cdot \left(X^t \cdot Y \right),$$

де $Y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_m \end{pmatrix}$ – вектор-стовпець значень залежної змінної *y*;

$X = \begin{pmatrix} 1 & x_{11} & ... & x_{m1} \\ 1 & x_{12} & ... & x_{m2} \\ ... & ... & ... & ... \\ 1 & x_{1n} & ... & x_{mn} \end{pmatrix}$ – матриця значень факторних ознак $x_1, x_2, ..., x_m$ (матриця спостережень), X^t – матриця, транспонована до матриці *X*;

$A = \begin{pmatrix} a_0 \\ a_1 \\ ... \\ a_m \end{pmatrix}$ – вектор-стовпець невідомих параметрів;

$U = \begin{pmatrix} u_0 \\ u_1 \\ ... \\ u_m \end{pmatrix}$ – вектор-стовпець помилок.

4. *Обчислення середніх значень* $\bar{Y}, \bar{X}_i$, *дисперсій* $\sigma_Y^2, \sigma_{X_i}^2$ *та стандартних відхилень* σ_Y, σ_{X_i} залежної змінної *Y* та незалежних змінних X_1, X_2 проводимо, використовуючи функції відповідно СРЗНАЧ, ДИСПР та СТАНДОТКЛОНП

5. Кореляційна матриця.

Елементи матриці обчислюємо за формулами:

$$r_{YX_i} = \frac{\left(Y^*\right)^t X_i^*}{n}, \qquad r_{X_i X_j} = \frac{\left(X_i^*\right)^t X_j^*}{n},$$

де Y^*, X_i^* $(i = \overline{1, m})$ – нормалізовані змінні:

$$Y^* = \frac{Y - \bar{Y}}{\sigma_Y}, \qquad X_1^* = \frac{X_1 - \bar{X}_1}{\sigma_{X_1}}, \qquad X_2^* = \frac{X_2 - \bar{X}_2}{\sigma_{X_2}};$$

Кореляційна матриця симетрична, її елементи характеризують ступені залежності між відповідними змінними.

Якщо $r_{X_i X_j} > 0$, то відповідні змінні пов'язані додатною кореляцією, тобто при зростанні однієї змінної – друга також має тенденцію до зростання.

Якщо ж $r_{X_i X_j} < 0$, то відповідні змінні пов'язані від'ємною кореляцією, тобто при зростання однієї змінної – друга має тенденцію до спадання.

6. Залишкова дисперсія:

$$\sigma_u^2 = \frac{U^t \cdot U}{n - m - 1} = \frac{\sum_{i=1}^{n} u_i^2}{n - m - 1}.$$

7. Загальна дисперсія:

$$\sigma_y^2 = \frac{1}{n-1} \cdot \left(Y^t \cdot Y - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} y_i \right)^2 \right).$$

8. Коефіцієнт множинної кореляції:

$$R = \sqrt{1 - \frac{\sigma_u^2}{\sigma_y^2}}.$$

9. Довірчі границі коефіцієнта множинної кореляції R:

$$\tilde{R} = R \pm t_{\alpha,k} \cdot \sigma_R,$$

де $\sigma_R = \frac{1 - R^2}{\sqrt{n - m - 1}}$,

$t_{\alpha,k}$ – табличне значення t –статистики Стьюдента для рівня значущості α і числа ступенів свободи $k = n - m - 1$.

10. Значущість коефіцієнта множинної кореляції:

$$t_R = \frac{R}{\sigma_R}.$$

Якщо $t_R > t_{\alpha,k}$, то з ймовірністю p можна стверджувати, що коефіцієнт множинної кореляції значущий.

11. Коефіцієнт детермінації:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}.$$

12. Оцінений коефіцієнт детермінації:

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\frac{n-1}{n-m-1}.$$

13. Спостережуване значення критерію Фішера для перевірки моделі на адекватність

$$F_p = \frac{\delta_y^2}{\sigma_u^2} = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 \Big/ (m-1)}{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \Big/ (n-m-1)}$$

порівнюється з табличним значенням F_{α, k_1, k_2}, де $k_1 = m-1$, $k_2 = n-m-1$. Якщо $F_p > F_{\alpha, k_1, k_2}$, то з ймовірністю $p = 1-\alpha$ можна стверджувати, що економетрична модель адекватно описує математичне явище.

14. Довірчий інтервал для окремого значення $\hat{y}_k$:

$$\hat{y}_k \pm \Delta y_k,$$

де $\Delta y_k = t_{\alpha,k} \cdot \sigma_u \cdot \sqrt{1 + X_k \cdot \left(X^t \cdot X\right)^{-1} \cdot X_k^t}$.

15. Довірчий інтервал для прогнозованого значення $\hat{y}_p$:

$$\hat{y}_p \pm \Delta y_p,$$

де $\Delta y_p = t_{\alpha,k} \cdot \sigma_u \cdot \sqrt{1 + X_p \cdot \left(X^t \cdot X\right)^{-1} \cdot X_p^t}$,

$\hat{y}_p = X_p \cdot A.$

16. Значущість параметрів моделі за t-критерієм Стьюдента.

Для цього обчислимо

$$\sigma_{a_0} = \sigma_u \cdot \sqrt{b_{11}};\ t_{a_0} = \frac{|a_0|}{\sigma_{a_0}};$$

$$\sigma_{a_1} = \sigma_u \cdot \sqrt{b_{22}};\ t_{a_1} = \frac{|a_1|}{\sigma_{a_1}};$$

…,

$$\sigma_{a_m} = \sigma_u \cdot \sqrt{b_{(m+1)(m+1)}};\ t_{a_m} = \frac{|a_m|}{\sigma_{a_m}},$$

де b_{ii} – елементи матриці $\left(X^t \cdot X\right)^{-1}$, $i = \overline{1, m+1}$.

Якщо t_{a_0}, t_{a_1}, ..., $t_{a_m} > t_{\alpha,k}$ то параметри a_0, a_0, a_1, ..., a_m значущі.

17. Коефіцієнти еластичності:

$$KE_1 = a_1 \cdot \frac{\overline{x}_1}{\overline{y}},\ KE_2 = a_2 \cdot \frac{\overline{x}_2}{\overline{y}},\ \ldots,\ KE_m = a_m \cdot \frac{\overline{x}_m}{\overline{y}}.$$

Коефіцієнти еластичності показують, на скільки % зростає показник *y* при зростанні фактора x_i на 1%.