Лекція 22. Множинна лінійна регресія

На практиці здебільшого залежна змінна y_i пов'язана з впливом не одного, а кількох аргументів (факторів). У цьому разі регресію називають *множинною*. При цьому якщо аргументи в функції регресії в першій степені, то множинна регресія називається *лінійною*, у противному разі — множинною *нелінійною регресією*.

Нехай між показником Y та факторами $X_1, X_2, ..., X_m$ існує лінійний зв'язок

$$y = b_0^0 + b_1^0 x_{1i} + b_2^0 x_{2i} + ... + b_m^0 x_{mi} \qquad (22.1)$$

Для вибірки обсягом n матимемо систему лінійних рівнянь

$$\begin{aligned}
y_1 &= b_0 + b_1 x_{11} + b_2 x_{12} + ... + b_m x_{1m} + \varepsilon_1, \\
y_2 &= b_0 + b_1 x_{21} + b_2 x_{22} + ... + b_m x_{2m} + \varepsilon_2, \\
y_3 &= b_0 + b_1 x_{31} + b_2 x_{32} + ... + b_m x_{3m} + \varepsilon_3, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
y_n &= b_0 + b_1 x_{n1} + b_2 x_{n2} + ... + b_m x_{nm} + \varepsilon_n
\end{aligned}$$

де ε_i - випадкова величина, що має нормальний закон розподілу з числовими характеристиками $M(\varepsilon_i) = 0$, $D(\varepsilon_i) = M(\varepsilon_i^2) = \sigma_\varepsilon^2$ і при цьому $K_{ij} = 0$. У векторно-матричній формі система рівнянь набуває вигляду:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}; \quad B = \begin{pmatrix} b_0 \\ b_1 \\ ... \\ b_m \end{pmatrix}; \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{pmatrix}; \quad X = \begin{pmatrix} 1 & x_{11} & ... & x_{1m} \\ 1 & x_{21} & ... & x_{2m} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{nm} \end{pmatrix} \qquad (22.2)$$

де n – кількість проведених спостережень, m – кількість пояснюючих змінних. Матрицю X розміром $(m+1)\cdot n$ називають регресійною, а елементи x_{ij} цієї матриці — регресорами. Параметри рівняння (22.1) є величинами сталими, але невідомими. Ці параметри оцінювання статистичними точковими оцінками $b_0, b_1, b_2, ..., b_m$, які дістають шляхом обробки результатів вибірки, і є

величинами випадковими. Таким чином, рівнянню (22.1) відповідає статистична оцінка

$$\hat{y}_x = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_m x_{mi} + \varepsilon \qquad (22.3)$$

де y та $\hat{y}_x$ - відповідно фактичні та розрахункові (статистичні) значення ознаки Y, ε - похибка.

Побудова рівняння регресії зводиться до оцінки його параметрів в (22.1). Для оцінки параметрів множинної лінійної регресії застосуємо метод найменших квадратів (МНК) в матричній формі.

$$B = \left(X^T X\right)^{-1} X^T Y, \qquad (22.4)$$

де X^T - транспонована матриця X, $\left(X^T X\right)^{-1}$ - обернена матриця, що отримана після добутків матриць $X^T X$.

В результаті, з урахуванням знайдених коефіцієнтів, запишемо рівняння регресії

$$\hat{y}_x = b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_m$$

По заданим значенням $x_{i1}, x_{i2}, ..., x_{im}$ знайдемо $\hat{y}_{x_i}$ за формулою $\hat{Y} = XB$.

Вектор похибок визначимо як $e = Y - \hat{Y}$

$$e = \begin{pmatrix} y_1 - \hat{y}_{x_1} \\ y_2 - \hat{y}_{x_2} \\ ..... \\ y_n - \hat{y}_{x_n} \end{pmatrix} \qquad (22.5)$$

Коефіцієнт $b_j \quad j = \overline{1,m}$ показує на скільки одиниць зміниться $\hat{y}_x$, якщо x_j зміниться(збільшиться або зменшиться) на одну одиницю.

Оцінка статистичної значущості коефіцієнтів регресії

Для оцінки статистичної значущості коефіцієнтів рівняння регресії обчислимо незміщену оцінку дисперсії

$$S^2_{залиш} = \frac{\sum e_i^2}{n-m-1}, \qquad (22.6)$$

де n - число спостережень, m - число пояснюючих змінних. Тоді вибіркові дисперсії емпіричних коефіцієнтів регресії визначимо за формулою

$$S^2_{b_{j-1}} = S^2_{залиш} \cdot z_{jj}, \ j = \overline{1, m+1}, \qquad (22.7)$$

де z_{jj} - діагональний елемент матриці

$$Z = \left(X^T X\right)^{-1}. \qquad (22.8)$$

Стандартні помилки коефіцієнтів регресії визначимо як $S_{b_j} = \sqrt{S^2_{b_j}}$. Знаходимо t - статистики за формулою

$$t_{b_j} = \frac{b_j}{S_{b_j}}, \ j = \overline{0, m} \qquad (22.9)$$

За таблицею розподілу критичних точок Стьюдента визначимо $t_{крит} = t\left(\frac{\delta}{2}, n-m-1\right)$, де δ рівень значущості, n - кількість спостережень, m - число пояснюючих змінних $\left(\nu = k = n-m-1\right)$ - число ступенів свободи.

Якщо $\left|t_{b_j}\right| > t_{крит}$, то коефіцієнт статистично значимий, а якщо ця умова не виконується, тобто $\left|t_{b_j}\right| < t_{крит}$, то коефіцієнт статистично не значимий. Статистично не значима і відповідна змінна, при якій він знаходиться. Ця змінна може бути виключена з моделі.

Побудова довірчих інтервалів для коефіцієнтів рівняння регресії

На основі заданого рівня значущості δ побудуємо критерій потужності $\gamma = 1-\delta$ для якого виконується умова $P\left(\left|t_{факт}\right| < t_\gamma\right) = \gamma = 1-\delta$, де $t_{факт} = \frac{b_j - b_j^0}{S_{b_j}}$. Далі отримаємо:

$$P\left(-t\left(\frac{\delta}{2},n-m-1\right)<\frac{b_j-b_j^0}{S_{b_j}}<t\left(\frac{\delta}{2},n-m-1\right)\right)=\gamma-1-\delta,\ j=\overline{0,m}$$

Звідки довірчі інтервали для коефіцієнтів регресії знаходимо за формулами

$$b_j-t\left(\frac{\delta}{2},n-m-1\right)\cdot S_{b_j}<b_j^0<b_j+t\left(\frac{\delta}{2},n-m-1\right)\cdot S_{b_j},\ j=\overline{0,m} \qquad (22.10)$$

Таким чином, з надійністю $\gamma=(1-\delta)\cdot 100\%$, можна стверджувати, що коефіцієнти теоретичного рівняння (22.1) належать інтервалам:

$$b_j^0\in\left(b_j-t\left(\frac{\delta}{2},n-m-1\right)\cdot S_{b_j};b_j+t\left(\frac{\delta}{2},n-m-1\right)\cdot S_{b_j}\right),\ j=\overline{0,m}.$$

Знаходження стандартизованих коефіцієнтів регресії

Стандартизовані коефіцієнти регресії визначаються за формулою:

$$b_j^s=b_j\frac{\sigma_{x_j}}{\sigma_y},\ j=\overline{1,m}, \qquad (22.11)$$

де $\sigma_{x_j}^2=\frac{\sum_{i=1}^{n}\left(x_{ji}-\overline{x}_j\right)^2}{n-1}$, $\sigma_y^2=\frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}{n-1}$.

Тоді стандартизоване рівняння регресії набуде вигляду:

$$\hat{y}_x=b_1^s x_1+b_2^s x_2+\ldots+b_m^s x_m+e.$$

Стандартизовані коефіцієнти показують на скільки сігм σ_y (середніх квадратичних відхилень) зміниться в середньому результат, якщо відповідний фактор $x_j\ \ j=\overline{1,m}$ зміниться на одну сигму σ_{x_j} при умові, що інші фактори залишаться без змін.

Порівняння між собою стандартизованих коефіцієнтів дозволяє ранжувати фактори $x_j\ \ j=\overline{1,m}$ по рівню впливу на результативну змінну.

Знаходження коефіцієнтів еластичності

Коефіцієнт еластичності визначається за формулою

$$E_j = b_j \frac{x_j^*}{y^*} \quad j = \overline{1,m}, \tag{22.12}$$

де x_j^* та y^* значення пояснюючої та залежної змінних в точці обчислення еластичності. Коефіцієнт еластичності показує на скільки відсотків зміниться залежна змінна *y*, якщо пояснююча змінилась на 1%. На практиці часто знаходять середній коефіцієнт еластичності

$$\bar{E}_j = b_j \frac{\bar{x}_j}{\bar{y}} \quad j = \overline{1,m},$$

Обчислення коефіцієнта детермінації та скорегованого коефіцієнта детермінації

Коефіцієнт детермінації визначимо за формулою

$$R^2 = \frac{B^T X^T Y - n(\bar{y})^2}{Y^T Y - n(\bar{y})^2}, \tag{22.13}$$

де Y^T - транспонований вектор-стовбець Y, B^T - транспонований вектор-стовбець B.

Аналіз його здійснимо на підставі F-статистики Фішера

$$F_{факт} = \frac{R^2}{1-R^2} \cdot \frac{n-m+1}{m},$$

де n - число спостережень, m - число пояснюючих змінних.

Для визначення статистичної значущості коефіцієнта детермінації порівняємо статистику $F_{факт}$ з критичною точкою розподілу Фішера $F_{крит} = F(\delta;\nu_1;\nu_2)$, $\nu_1 = m$, $\nu_2 = n-m-1$ число ступенів свободи, δ-рівень значущості. Перевіряються дві тісно пов'язані гіпотези:

$$H_0 : R^2 = 0 \text{ та } H_1 : R^2 \neq 0.$$

Якщо $F_{факт} > F_{крит}$, то коефіцієнт детермінації статистично значимий, гіпотеза H_0 відхиляється на користь гіпотези H_1 і це означає, що сукупний вплив пояснюючих змінних на залежну змінну істотний. Якщо ж $F_{факт} < F_{крит}$, H_0 то гіпотеза приймається і рівняння регресії визнається статистично не значущим і не надійним.

Коефіцієнт детермінації показує долю розкиду залежної змінної y, що пояснюється рівнянням регресії.

Скоригований коефіцієнт детермінації знайдемо за формулою

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-m-1} \qquad (22.14)$$

і порівняємо його з R^2.

Знаходження прогнозного значення $\hat{y}_{x\,прогн}$ і побудова довірчого інтервалу для прогнозного значення

Для визначення прогнозного значення $\hat{y}_{x\,прогн}$ залежної змінної y, необхідно в отримане по МНК рівняння регресії (22.3) підставити прогнозні значення пояснюючих змінних $x_{j\,прогн}$, $j = \overline{1,m}$, тобто

$$\hat{y}_{x\,прогн} = b_0 + b_1 x_{1\,прогн} + b_2 x_{2\,прогн} + ... + b_m x_{m\,прогн}.$$

Для побудови довірчого інтервалу для прогнозного значення визначається середня стандартна помилка прогнозу

$$S_{\hat{y}\,прогн} = S_{залиш} \cdot \sqrt{1 + X_0^T \left(X^T X\right)^{-1} X_0}\,, \qquad (22.15)$$

де $S_{залиш} = \sqrt{S^2_{залиш}} = \sqrt{\frac{\sum e_i^2}{n-m-1}}$ (22.6), вектор X_0 утворений із прогнозованих значень пояснюючих змінних $X_0 = \begin{pmatrix} 1 \\ x_{1\,прогн} \\ x_{2\,прогн} \\ ..... \\ x_{m\,прогн} \end{pmatrix}$, X_0^T - транспонований вектор X_0.

Визначається гранична помилка $\Delta_{\hat{y}\,прогн} = t\left(\frac{\delta}{2};\nu\right) \cdot S_{\hat{y}\,прогн}$, де $t\left(\frac{\delta}{2};\nu\right)$ - критичне значення розподілу Стьюдента. Будується довірчий інтервал для прогнозу:

$$y^0_{x\,прогн} \in \left(\hat{y}_{x\,прогн} - \Delta_{\hat{y}\,прогн}; \hat{y}_{x\,прогн} + \Delta_{\hat{y}\,прогн}\right).$$

Отже з надійністю $\gamma = (1-\delta)\cdot 100\%$, можна стверджувати, що $y^0_{x\,прогн}$ належить до визначеного прогнозного інтервалу.

Перевірка моделі на мультиколінеарність

Важливу роль в економетричних дослідженнях займає проблема: з'ясувати чи існують між пояснюючими факторами моделі взаємозв'язки, так звана *мультиколінеарність*. Мультиколінеарність означає наявність тісного лінійного зв'язку між двома або кількома пояснюючими змінними моделі. Розглянемо алгоритм Феррара-Глобера перевірки моделі на мультиколінеарність. Алгоритм містить три види статистичних критеріїв:

- Перевірка всього масиву пояснюючих змінних (χ^2-критерій)
- Перевірка незалежності змінної з усіма іншими змінними (F-критерій Фішера);
- Перевірка кожної пари пояснюючих змінних (t-критерій Стьюдента).

Здійснимо реалізацію алгоритму по кроках.

Крок 1. Нормалізуємо змінні $x_1, x_2, ..., x_m$ за формулою

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{\sqrt{n\sigma_{x_j}^2}}, \tag{22.16}$$

де n - число спостережень; m - число пояснюючих змінних; $\overline{x}_j$ - середнє арифметичне j-ої пояснюючої змінної; $\sigma_{x_j}^2$ - дисперсія j-ої пояснюючої змінної, яка обчислюється за формулою $\sigma_{x_j}^2 = \dfrac{\sum_{i=1}^{n}\left(x_{ji} - \overline{x}_j\right)^2}{n}$.

Крок 2. Побудуємо нову матрицю $X^{*} = \left\{x_{ij}^{*}\right\}_{i=\overline{1,m}}^{j=\overline{1,m}}$, елементами якої є нормалізовані змінні x_{ij}^{*}. Обчислимо кореляційну матрицю R.

$$R = \left(X^{*}\right)^{T} \cdot X^{*} = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1m} \\ r_{21} & 1 & \ldots & r_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ r_{m1} & r_{m2} & \ldots & r_{mm} \end{pmatrix}, \tag{22.17}$$

де $\left(X^{*}\right)^{T}$ - транспонована матриця X^{*}, $r_{ij} = r_{x_i x_j}$ - парні коефіцієнти кореляції які визначають силу зв'язку між пояснюючими змінними x_i та x_j, $i = \overline{1,m}$, $j = \overline{1,m}$.

Якщо діагональні елементи матриці *R* не рівні одиниці, то на головній діагоналі ставлять одиниці, а до інших елементів рядка додаємо різницю між діагональним елементом і одиницею.

Крок 3. Знаходимо визначник $|R|$ матриці *R.*

Застосовуємо критерій χ^2, для цього обчислюємо $\chi_{факт}^2 = -\left(n - 1 - \frac{1}{6}(2m+5)\right)\ln|R|$. Порівнюємо його з $\chi_{табл}^2$, знайдений за допомогою таблиць розподілу критичних точок χ^2 при ступені свободи $\nu = 0{,}5m(m-1)$ і при рівні значущості δ. Якщо $\chi_{табл}^2 < \chi_{факт}^2$, то в масиві

пояснюючих змінних мультиколінеарність присутня, в іншому випадку коли $\chi^2_{табл} > \chi^2_{факт}$ мультиколінеарність відсутня.

Крок 4. Знаходимо матрицю помилок C за формулою:

$$C = R^{-1} = \left(\left(X^* \right)^T \cdot X^* \right)^{-1} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \dots & \dots & \dots & \dots \\ c_{m1} & c_{m2} & \dots & c_{mm} \end{pmatrix}, \qquad (22.18)$$

де R^{-1} - обернена матриця до R.

Знаходимо $F_{k\ факт} = \dfrac{(c_{kk} - 1)(n - m)}{m - 1}$, $k = \overline{1,m}$, (22.19)

де c_{kk} - діагональні елементи матриці C. Значення $F_{k\ факт}$ порівнюємо з табличним $F_{табл}$ взятим при рівні значущості δ і ступенями свободи $\nu_1 = n - m$ і $\nu_2 = m - 1$. Якщо $F_{k\ факт} > F_{табл}$, то відповідна k -та пояснююча змінна мультиколінеарна з усіма іншими. В іншому випадку, $F_{k\ факт} < F_{табл}$, то k -та пояснююча змінна не мультиколінеарна з іншими.

Крок 5. Знаходимо часткові коефіцієнти кореляції, які характеризують тісноту зв'язку між двома змінними x_i та x_j, $i = \overline{1,m}$, $j = \overline{1,m}$, $i \neq j$ за умови, що інші змінні не впливають на цей зв'язок (досліджується існування попарної мультиколінеарності):

$$r_{ij.12\dots(i-1)(i+1)\dots(j-1)(j+1)\dots m} = \frac{-c_{ij}}{\sqrt{c_{ii} c_{jj}}}, \qquad (22.20)$$

де c_{ij} - елементи матриці помилок, матриці C.

Крок 6. Для оцінки статистичної значущості часткових коефіцієнтів кореляції розраховується *t*-критерій Стьюдента за формулою:

$$t_{ij} = \left| r_{ij.12\dots} \right| \cdot \frac{\sqrt{n - m}}{\sqrt{1 - r^2_{ij.12\dots}}}. \qquad (22.21)$$

Значимо критерію t_{ij} порівнюємо з табличним $t_{табл} = t(0{,}5\delta; \nu)$, яке обчислюється при рівні значущості δ та $\nu = n - m$ ступенях свободи.

Якщо $t_{ij} > t_{табл}$, то між змінними x_i та x_j, $i = \overline{1,m}$, $j = \overline{1,m}$, $i \neq j$ існує мультиколінеарність. В іншому випадку, коли $t_{ij} < t_{табл}$, то між змінними x_i та x_j мультиколінеарнеарності немає.

Крок 7. Висновки:

Якщо $F_{k\,факт} > F_{табл}$,то змінна x_k залежить від інших пояснюючих змінних і необхідно вирішувати питання про її вилучення з моделі.

Якщо $t_{ij} > t_{табл}$, то змінні x_i та x_j тісно пов'язані між собою.

Аналізуючи *F* і *t* критерії необхідно зробити висновок, яку зі змінних необхідно виключити з даної моделі (звичайно, тут потрібно керуватися в першу чергу економічними міркуваннями).

Перевірка моделі на гетероскедастичність

Якщо властивість дисперсії залишків не змінюється від спостереження до спостереження, то це *називається гомоскедастичністю*. Якщо ж дисперсія змінюється від спостереження до спостереження, то ця властивість називається гетероскедастичністю. Вона приводить до того, що оцінки МНК параметрів моделі, стають неефективними, залишаючись при цьому незміщеними і обґрунтованими. Якщо кількість спостережень невелика, то використовують параметричний тест Гольдфельда-Квандта, який також представимо по крокам.

Крок 1. Вихідні дані, по тій пояснюючій змінній, яка може викликати зміну дисперсії залишків (обраної заздалегідь, скажімо x_1), ранжуються або по зростанню, або по спаданню. В подальшому цю процедуру повторюємо по кожній з пояснюючих змінних.

Крок 2. Викинемо l спостережень, які знаходяться в середині векторів ранжируваних вихідних даних $l \approx \frac{4n}{15}$, де n - число спостережень, побудуємо дві нові моделі за новоствореним сукупностями спостережень, розмірністю

$\frac{n-l}{2}$ кожна, за умови, що $\frac{n-l}{2} \geq m$, де m - кількість пояснюючих змінних. Застосовуючи МНК, знаходимо коефіцієнти регресії для кожної сукупності окремо за формулою $B = \left(X^T X\right)^{-1} X^T Y$.

Крок 3. Будуємо рівняння регресії для кожної сукупності. Визначимо вектори похибок для кожної групи спостережень як $e = Y - \hat{Y}$, тобто

$e_1 = \begin{pmatrix} y_1 - \hat{y}_{x_1} \\ y_2 - \hat{y}_{x_2} \\ ..... \\ y_l - \hat{y}_{x_l} \end{pmatrix}$ - для першої сукупності, $e_2 = \begin{pmatrix} y_{n-l} - \hat{y}_{x_{n-l}} \\ y_{n-l+1} - \hat{y}_{x_{n-l+1}} \\ ..... \\ y_n - \hat{y}_{x_n} \end{pmatrix}$ - для другої.

Знаходимо суму квадратів відхилень для кожної сукупності

$$S_1^2 = e_1^T e_1,\ S_2^2 = e_2^T e_2.$$

Крок 4. Застосовуємо F-статистику Фішера. Знаходимо $F_{факт} = \frac{S_1^2}{S_2^2}$. Висувається основна гіпотеза H_0 про наявність гомоскедастичності та альтернативна гіпотеза H_1 про наявність гетероскедастичності. Якщо $F_{факт} \leq F_{крит}$, то справедлива H_0 гіпотеза про наявність гомоскедастичності, противному випадку, якщо $F_{факт} > F_{крит}$, то приймається гіпотеза про наявність гетероскедастичності.

$F_{крит} = F\left(\delta; \nu_1; \nu_2\right)$ - це критична точка розподілу Фішера, обчислена при заданому рівні значущості δ та $\nu_1 = 0{,}5 \cdot \left(n - l - 2 \cdot m\right)$, $\nu_2 = 0{,}5 \cdot \left(n - l - 2 \cdot m\right)$ числу ступенів свободи. Де n – число спостережень, m – кількість пояснюючих змінних, l – кількість вилучених спостережень.

Після завершення дослідження по одній пояснюючій змінній переходять до наступної і так повторюють процедуру до тих пір поки не дослідять всі підозрілі на гетероскедастичність змінні.