

17. Статистические свойства оценок по методу наименьших квадратов параметров множественной регрессии. Коэффициент детерминации и скорректированный коэффициент детерминации.

Модель множественной регрессии (многомерная регрессивная модель)

$$y_t = \beta_1 + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + \varepsilon_t, \quad t = \overline{1, n} \quad \text{или} \quad y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + \varepsilon_t, \quad t = \overline{1, n},$$

где x_{tp} — значение регрессора x_p в наблюдении t , а $x_{t1} = 1$, $t = \overline{1, n}$. С учетом этого различаются модели со свободным и без свободного члена.

Если выполнены следующие гипотезы, то модель называется **нормальной линейной регрессионной**:

1. $y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + \varepsilon_t$, $\overline{1, n}$ — спецификация модели;
2. $x_{t1}, \dots, x_{tk}$ — детерминированные величины. Вектор $x_s = (x_{1s}, \dots, x_{ns})^T$, $s = \overline{1, k}$ линейно независимые в R^n ;
3. $\mathbb{E}\varepsilon_t = 0$, $\mathbb{E}(\varepsilon_t^2) = \text{Var}(\varepsilon_t) = \sigma^2$ — не зависит от t ;
4. $\mathbb{E}(\varepsilon_t \varepsilon_s) = 0$ при $t \neq s$ — статистическая независимость (некоррелированность) ошибок для разных наблюдений;
5. Ошибки ε_t , $t = \overline{1, n}$ имеют совместное нормальное распределение: $\varepsilon_t \sim N(0, \sigma^2)$.

Матричный вид гипотез

Пусть y — вектор-столбец $(y_1, \dots, y_n)^T$, $\beta = (\beta_1, \dots, \beta_k)^T$ — вектор коэффициентов; $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ — вектор ошибок. $X = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nk} \end{bmatrix}$ — $n \times k$ матрица объясняющих переменных.

Столбцами матрицы X являются $n \times 1$ векторы регрессоров $x_s = (x_{1s}, \dots, x_{ns})^T$, $s = \overline{1, k}$

1. $y = X\beta + \varepsilon$ — спецификация модели;
2. X — детерминированная матрица, имеет максимальный ранг k ;
3. $\mathbb{E}(\varepsilon) = 0$, $\text{Var}(\varepsilon) = \varepsilon \varepsilon^T = \sigma^2 I_n$;
4. $\varepsilon \sim N(0, \sigma I_n)$, то есть ε — нормально распределенный случайный вектор со средним 0 и матрицей ковариации $\sigma^2 I_n$ (нормальная линейная регрессионная модель).

Теорема Гаусса-Маркова для множественной регрессии

Предположим, что выполнены гипотезы 1-3. Тогда оценка методом наименьших квадратов $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$ является наиболее эффективной (в смысле наименьшей дисперсии) оценкой в классе линейных (по y) несмещённых оценок (Best Linear Unbiased Estimator, BLUE).

Статистические свойства МНК оценок

Введем вектор прогнозируемых значений $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$

Введем вектор остатков регрессии: $e = y - \hat{y} = y - X\hat{\beta} = (I - X(X^T X)^{-1} X^T)y = My$, где $M = (I - X(X^T X)^{-1} X^T) = (I - N)$.

Вычислим мат. ожидание и матрицу ковар. e :

$$\mathbb{E}(e) = (I - X(X^T X)^{-1} X^T) \mathbb{E}(y) = (I - X(X^T X)^{-1} X^T) X\beta = X\beta - X\beta = 0. \text{Var}(e) = \text{Var}(My) = M \text{Var}(y) M^T = M \sigma^2 I M^T = \sigma^2 M$$

Оценка дисперсии ошибок: $\mathbb{E}(e^T e) = \text{tr}(\text{Var}(e)) = \sigma^2 \text{tr}(I_n - N) = (n - k) \sigma^2$ (tr — след матрицы).

Следовательно $s^2 = \hat{\sigma}^2 = \frac{e^T e}{n - k} = \frac{\sum_{i=1}^n e_i^2}{n - k}$ — несмещенная оценка дисперсии ошибок, то есть $\mathbb{E}s^2 = \sigma^2$.

Распределение: $\frac{e^T e}{\sigma^2} \sim \chi^2(n - k)$ или $(n - k) \frac{s^2}{\sigma^2} \sim \chi^2(n - k)$

Оценки $\hat{\beta}$ и s^2 независимы в предположении нормальной линейной множественной регрессионной модели $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T (X\beta + \varepsilon) = \beta + (X^T X)^{-1} X^T \varepsilon = \beta + A\varepsilon$

Вектор $\hat{\beta}$ и e имеют совместное многомерное нормальное распределение $\Rightarrow$ нужно доказать их некоррелированность

$AM = (X^T X)^{-1} X^T (I - X(X^T X)^{-1} X^T) = 0$. Тогда (так как $\mathbb{E}e = 0$): $\text{Cov}(\hat{\beta}, e) = \mathbb{E}((\hat{\beta} - \beta)e^T) = \mathbb{E}(A\varepsilon \varepsilon^T M) = \sigma^2 AM = 0$. Так как s^2 является функцией $e \Rightarrow$ оценки $\hat{\beta}$ и s^2 независимы.

Коэффициент детерминации R^2 и скорректированный коэффициент детерминации R_{adj}^2

Разобьем вариацию $\sum (y_t - \bar{y})^2$ на две части: объясняемую и не объясняемую.

$$\sum (y_t - \bar{y})^2 = \sum (y_t - \hat{y}_t)^2 + \sum (\hat{y}_t - \bar{y})^2 + 2 \sum (y_t - \hat{y}_t)(\hat{y}_t - \bar{y})$$

То же самое в векторной форме:

$$(y - \bar{y}i)^T (y - \bar{y}i) = (y - \hat{y})^T (y - \hat{y}) + (\hat{y} - \bar{y}i)^T (\hat{y} - \bar{y}i) + 2(y - \hat{y})^T (\hat{y} - \bar{y}i), \text{ где } i - \text{единичный вектор.}$$

Третье слагаемое равно нулю: $(y - \hat{y})^T (\hat{y} - \bar{y}i) = e^T (X\hat{\beta} - \bar{y}i) = e^T X\hat{\beta} - \bar{y}e^T i = 0$, так как $e^T X = 0$, $e^T i/n = 0$.

Поэтому: $\|y - \bar{y}i\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - \bar{y}i\|^2$ ($TSS = ESS + RSS$)(*) TSS — вся дисперсия. ESS — необъясненная часть дисперсии. RSS — объясненная часть дисперсии.

Положим: $y_* = y - \bar{y}i$, $\hat{y}_* = \hat{y} - \bar{y}i \Rightarrow y_*^T y_* = e^T e + \hat{y}_*^T \hat{y}_*$.

Коэффициентом детерминации называют $R^2 = 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS} = 1 - \frac{e^T e}{y_*^T y_*} = \frac{\hat{y}_*^T \hat{y}_*}{y_*^T y_*}$, $R^2 \in [0, 1]$.

R^2 корректно определен, только если вектор $i = (1, \dots, 1)^T$ принадлежит линейной оболочке векторов $x_1, \dots, x_k$.

Свойства R^2 :

1. R^2 возрастает при добавлении еще одного регрессора.
2. R^2 изменяется даже при простейшем преобразовании зависимой переменной, поэтому сравнивать по значению R^2 можно только регрессии с одинаковыми зависимыми переменными.

Попыткой устранить эффект, связанный с ростом R^2 при возраст. числа регресс., является коррекция R^2 на число регресс.

Скорректированным (adjusted) R^2 называется $R_{adj}^2 = 1 - \frac{e^T e / (n - k)}{y_*^T y_* / (n - 1)}$

Заметим, что нет никакого существенного оправдания такого способа коррекции (то есть можно по другому корректировать, главное устранять недостаток возрастания при увеличении числа регрессоров).

Свойства скорректированного R^2

1. $R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-k}$
2. $R^2 \geq R_{adj}^2$, $k > 1$.
3. $R_{adj}^2 \leq 1$, но может принимать значения < 0 .