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

Модель множественной регрессии (многомерная регрессивная модель)

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

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

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

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

Матричный вид гипотез

Пусть y — вектор-столбец $(y_1, ..., y_n)^T$, $\beta = (\beta_1, ..., \beta_k)^T$ — вектор коэффициентов; $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ — вектор ошибок.

$$X = \begin{bmatrix} x_{11} & ... & x_{1k} \\ ... & ... & ... \\ x_{n1} & ... & x_{nk} \end{bmatrix} - n \times k \text{ матрица объясняющих переменных.}$$

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

1. $y = X\beta + \varepsilon$ — спецификация модели;

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

Теорема Гаусса-Маркова для множественной регресии

Предположим, что выполнены гипотезы 1-3.

Тогда оценка методом наименьших квадратов $\widehat{\beta}_{OLS} = (X^T X)^{-1} X^T y$ является наиболее эффективной (в смысле наименьшей дисперсии) оценкой в классе линейных (по y) несмещённый 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:

$\mathbb{E}(e) = (I - X(X^T X)^{-1} X^T)\mathbb{E}(y) = (I - X(X^T X)^{-1} X^T)X\beta = X\beta - X\beta = 0.$

$Var(e) = Var(My) = MVar(y)M^T = M\sigma^2 IM^T = \sigma^2 M$

Оценка дисперсии ошибок:

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

Следовательно $s^2 = \widehat{\sigma}^2 = \frac{e^T e}{n-k} = \frac{\sum e_i^2}{n-k}$ — 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

$AM = (X^TX)^{-1}X^T(I - X(X^TX)^{-1}X^T) = 0$

Тогда (так как $\mathbb{E}e = 0$): $Cov(\hat{\beta}, e) = \mathbb{E}((\hat{\beta} - \beta)e^T) = \mathbb{E}(A\varepsilon\varepsilon^TM) = \sigma^2AM = 0$

Так как s^2 является функцией $e \Rightarrow$ оценки $\hat{\beta}$ и s^2 независимы.

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

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

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

Тоже самое в векторной форме:

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

Третье слагаемое равно нулю:

$(y - \hat{y})^T(y - \overline{y}i) = e^T(X\hat{\beta} - \overline{y}i) = e^TX\hat{\beta} - \overline{y}e^Ti = 0$, так как $e^TX = 0$, $e^Ti/n = 0$.

Поэтому: $||y - \overline{y}i||^2 = ||y - \hat{y}||^2 + ||\hat{y} - \overline{y}i||^2 (TSS + ESS + RSS)(*)$

TSS – вся дисперсия. ESS – необъясненная часть дисперсии. RSS – объясненная часть дисперсии.

Положим: $y_* = y - \overline{y}i, \quad \hat{y_*} = \hat{y} - \overline{y}i \Rightarrow y_*^Ty_* = e^Te + \hat{y_*}^T\hat{y_*}$

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

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

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 возрастании числа регрессоров, является коррекция R^2 на число регрессоров.

Скорректированным (adjusted) R^2 называется

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

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