

Эконометрика

Полковников Александр Александрович

Волжский политехнический институт (филиал)
ФГБОУ ВПО "Волгоградский государственный технический университет"

Конспект лекций для студентов направления
"Экономика"

Дисперсионный анализ

Проанализируем дисперсию результата y

$$S_y^2 = \frac{1}{n} \sum_{k=1}^n (y_k - \bar{y})^2.$$

Для любой точки $(x_k; y_k)$ на корреляционном поле можем записать

$$y_k - \bar{y} = (y_k - \hat{y}_k) + (\hat{y}_k - \bar{y}),$$

где $\hat{y}_k = a + bx_k$ — ордината точки “наилучшей прямой”, имеющей абсциссу x_k .

Возведем обе части последнего равенства в квадрат и просуммируем по всем k от 1 до n :

$$\sum_{k=1}^n (y_k - \bar{y})^2 = \sum_{k=1}^n (y_k - \hat{y}_k)^2 + \sum_{k=1}^n (\hat{y}_k - \bar{y})^2 + 2 \sum_{k=1}^n (y_k - \hat{y}_k) (\hat{y}_k - \bar{y}). \quad (1)$$

Докажем, что третья сумма в правой части формулы (1) равна нулю.

$$\begin{aligned} & \sum_{k=1}^n (y_k - \hat{y}_k) (\hat{y}_k - \bar{y}) = \\ &= \sum_{k=1}^n (y_k - a - bx_k) (a + bx_k - \bar{y}) = \\ &= \sum_{k=1}^n (y_k - \bar{y} + b\bar{x} - bx_k) (\bar{y} - b\bar{x} + bx_k - \bar{y}) = \end{aligned}$$

$$\begin{aligned} &= \sum_{k=1}^n ((y_k - \bar{y}) - b(x_k - \bar{x})) b(x_k - \bar{x}) = \\ &= b \left(\sum_{k=1}^n (y_k - \bar{y})(x_k - \bar{x}) - b \sum_{k=1}^n (x_k - \bar{x})^2 \right) = \\ &= b (n(\overline{xy} - \bar{x} \cdot \bar{y}) - bnS_x^2) = bn((\overline{xy} - \bar{x} \cdot \bar{y}) - bS_x^2) = 0. \end{aligned}$$

Таким образом, доказали равенство:

$$\sum_{k=1}^n (y_k - \bar{y})^2 = \sum_{k=1}^n (y_k - \hat{y}_k)^2 + \sum_{k=1}^n (\hat{y}_k - \bar{y})^2. \quad (2)$$

Определение

Величину, стоящую в левой части соотношения (2),

$$TSS = n \cdot S_y^2 = \sum_{k=1}^n (y_k - \bar{y})^2$$

будем называть **полной суммой квадратов** (total sum of squares)

Определение

Первое слагаемое правой части, как и раньше, обозначим RSS

$$RSS = \sum_{k=1}^n (y_k - \hat{y}_k)^2 = \sum_{k=1}^n e_k^2$$

и будем называть **остаточной суммой квадратов**.

Определение

Второе слагаемое правой части соотношения (2) будем называть **суммой квадратов, объясненной моделью** (*explained sum of squares*), и обозначать ESS , так что

$$ESS = \sum_{k=1}^n (\hat{y}_k - \bar{y})^2 = b^2 n S_x^2.$$

Иначе говоря, равенство (2) представляет собой разложение полной суммы квадратов на сумму квадратов, объясненную моделью, и остаточную сумму квадратов:

$$TSS = ESS + RSS.$$

В терминах сумм квадратов удобно записать коэффициент детерминации

$$R^2 = \frac{b^2 S_x^2}{S_y^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Таким образом, для коэффициента детерминации справедливо соотношение:

$$0 \leq R^2 \leq 1.$$

Полная сумма квадратов отклонений результата y от среднего значения вызвана влиянием множества причин, которые условно можно разделить на две группы: 1) причины, вызванные фактором x , 2) вызванные прочими факторами. Если фактор x не оказывает влияния на результат y , то линия регрессии параллельна оси Ox :

$$\hat{y} = \bar{y}.$$

В этом случае,

$$ESS = 0, \quad TSS = RSS, \quad R^2 = 0.$$

Если же прочие факторы не влияют на результат, то y и x связаны функционально и все точки $(x_k; y_k)$ лежат на линии регрессии $\hat{y}_k = y_k$. В этом случае,

$$RSS = 0, \quad TSS = ESS, \quad R^2 = 1.$$

Степень изменчивости результата y от наблюдения к наблюдению может быть охарактеризована исправленной выборочной дисперсией

$$\text{Var}(y) = \frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2 = \frac{nS_y^2}{n-1} = \frac{TSS}{n-1}$$

Используя величину $\text{Var}(y)$, можно выражение (2) записать иначе:

$$\text{Var}(y) = \frac{TSS}{n-1} = \frac{ESS}{n-1} + \frac{RSS}{n-1}.$$

Раннее при составлении системы нормальных уравнений доказали, что

$$\sum_{k=1}^n e_k = 0, \quad \sum_{k=1}^n (y_k - \hat{y}_k) = 0, \quad \sum_{k=1}^n y_k = \sum_{k=1}^n \hat{y}_k, \quad \bar{y} = \bar{\hat{y}},$$

где \hat{y} — величина, принимающая в k -м наблюдении значение \hat{y}_k . Отсюда,

$$\frac{ESS}{n-1} = \frac{\sum_{k=1}^n (\hat{y}_k - \bar{y})^2}{n-1} = \frac{\sum_{k=1}^n (\hat{y}_k - \bar{\hat{y}})^2}{n-1} = \text{Var}(\hat{y}).$$

Кроме того,

$$\begin{aligned} \frac{RSS}{n-1} &= \frac{\sum_{k=1}^n (y_k - \hat{y}_k)^2}{n-1} = \frac{\sum_{k=1}^n e_k^2}{n-1} = \frac{\sum_{k=1}^n (e_k - \bar{e})^2}{n-1} = \\ &= \frac{\sum_{k=1}^n (e_k - \bar{e})^2}{n-1} = \text{Var}(e). \end{aligned}$$

В итоге получаем разложение

$$\text{Var}(y) = \text{Var}(\hat{y}) + \text{Var}(e),$$

которое часто называют **дисперсионным анализом** (analysis of variance — ANOVA).

В терминах дисперсий можно выразить коэффициент детерминации

$$R^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(y)} = 1 - \frac{\text{Var}(e)}{\text{Var}(y)}.$$

Рассмотрим коэффициент корреляции между величинами y и \hat{y} :

$$\begin{aligned} r_{y\hat{y}} &= \frac{\text{Cov}(y; \hat{y})}{\sqrt{\text{Var}(y) \text{Var}(\hat{y})}} = \frac{\text{Cov}(\hat{y} + e; \hat{y})}{\sqrt{\text{Var}(y) \text{Var}(\hat{y})}} = \frac{\text{Cov}(\hat{y}; \hat{y}) + \text{Cov}(e; \hat{y})}{\sqrt{\text{Var}(y) \text{Var}(\hat{y})}} = \\ &= \frac{\text{Var}(\hat{y}) + 0}{\sqrt{\text{Var}(y) \text{Var}(\hat{y})}} = \sqrt{\frac{\text{Var}(\hat{y})}{\text{Var}(y)}}. \end{aligned}$$

Получили что,

$$r_{y\hat{y}}^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(y)} = R^2.$$

Величину $r_{y\hat{y}}$ называют **множественным коэффициентом корреляции**.

Выше при выводе соотношения для TSS доказали равенство

$$\sum_{k=1}^n (y_k - \hat{y}_k) (\hat{y}_k - \bar{y}) = 0.$$

Учитывая

$$\sum_{k=1}^n (y_k - \hat{y}_k) = 0,$$

получаем

$$\sum_{k=1}^n (y_k - \hat{y}_k) \hat{y}_k = 0,$$

$$\sum_{k=1}^n e_k \hat{y}_k = 0.$$

Доказали, что величины e и \hat{y} не коррелированы.

Доказали следующие свойства остатков:

- 1 средняя величина остатков равна нулю $\bar{e} = 0$,
- 2 остатки e и фактор x не коррелированы,
- 3 остатки e и предсказанные значения \hat{y} не коррелированы.

Также предполагается, что остатки e удовлетворяют следующим дополнительным условиям:

- 4 в выборке остатки не коррелированы между собой,
- 5 остатки имеют постоянную дисперсию, так называемая **гомоскедастичность остатков**. Если дисперсии различны, то говорят о **гетероскедастичности**,
- 6 остатки имеют нормальное распределение.

Основные соотношения метода наименьших квадратов имеют наглядную геометрическую интерпретацию.

Введем в рассмотрение следующие n -мерные векторы:

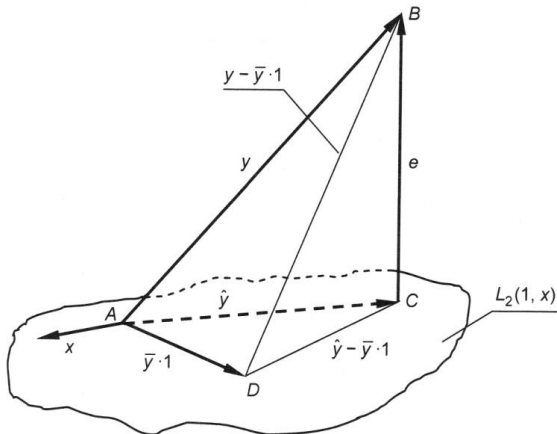
$$y = (y_1, \dots, y_n)^T, \quad \hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T, \quad \mathbb{1} = (1, \dots, 1)^T,$$

$$x = (x_1, \dots, x_n)^T, \quad e = (e_1, \dots, e_n)^T.$$

В терминах этих векторов свойства остатков можно записать следующим образом

$$y = \hat{y} + e, \quad e^T \mathbb{1} = 0, \quad e^T x = 0.$$

Так как вектор e ортогонален единичному вектору $\mathbb{1}$ и вектору x , то он ортогонален порожденному векторами $\mathbb{1}$ и x двумерному линейному подпространству $L_2(\mathbb{1}; x)$. Вектор \hat{y} является линейной комбинацией векторов $\mathbb{1}$ и x , а потому также принадлежит подпространству $L_2(\mathbb{1}; x)$. Вектор e ортогонален $L_2(\mathbb{1}; x)$, а значит, ортогонален любому вектору из $L_2(\mathbb{1}; x)$, в том числе и вектору \hat{y} .



Изображенный на рисунке треугольник ABC — прямоугольный. При этом \hat{y} является ортогональной проекцией вектора y на подпространство $L_2(1; x)$.

Отложим от точки A вектор $\bar{y}1$. Обозначим конец этого вектора D . Треугольник BCD является прямоугольным. Теорема Пифагора для этого треугольника имеет вид

$$|BD|^2 = |BC|^2 + |CD|^2,$$

т. е.

$$|y - \bar{y}1|^2 = |y - \hat{y}|^2 + |\hat{y} - \bar{y}1|^2$$

или в координатной форме

$$\sum_{k=1}^n (y_k - \bar{y})^2 = \sum_{k=1}^n (y_k - \hat{y}_k)^2 + \sum_{k=1}^n (\hat{y}_k - \bar{y})^2.$$

Это и есть доказанное ранее разложение

$$TSS = RSS + ESS.$$

СПАСИБО ЗА ВНИМАНИЕ!