5

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Экономические переменные обычно зависят не от одного, а от многих факторов. Так, на равновесную цену товара влияют издержки производителей, доходы покупателей, цены взаимозаменяющих и взаимодополняющих товаров, время года и другие факторы.

5.1. Модель множественной линейной регрессии

5.1.1. Множественная регрессия. Линейная модель регрессии с несколькими объясняющими переменными является обобщением модели парной линейной регрессии и записывается в следующем виде:

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon, \qquad (5.1)$$

где y – зависимая переменная, x_1, x_2,…, x_k – объясняющие переменные, β_1, β_2,…, β_k – коэффициенты регрессии, ε – случайный член, включение которого в уравнение регрессии обусловлено теми же причинами, что и в случае парной регрессии. Модель в форме (5.1) не может содержать постоянный член. Но на самом деле это не так. Формально полагая, что во всех наблюдениях $x_1 = 1$ получим модель с постоянным членом:

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon, \qquad (5.2)$$

которая содержит $k-1$ объясняющих переменных $x_2, ..., x_k$ и постоянный член, обозначенный через β_1. Запись в форме (5.1)

удобна тем, что не нужно без необходимости каждый раз оговаривать, содержит или не содержит модель постоянный член. При $k = 2$ и $x_1 = 1$ уравнение (5.1) переходит в уравнение парной линейной регрессии.

5.1.2. Матричная форма записи модели. Пусть имеется выборка, состоящая из n наблюдений зависимой и объясняющих переменных $y_i, x_{i1}, x_{i2}, ..., x_{ik}$, $i = 1,2,...,n$, для которых уравнение регрессии (5.1) запишется в виде системы уравнений:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i, \quad i = 1,2,...,n. \tag{5.3}$$

Определим вектор-столбцы и матрицу:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{22} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Используя эти обозначения, систему уравнений (5.3) можно записать в компактной матричной форме:

$$y = X\beta + \varepsilon. \tag{5.4}$$

П р и м е р 5.1. Рассмотрим модель зависимости дохода индивидуума от различных факторов:

$$income_i = \beta_1 + \beta_2 edu_i + \beta_2 age_i + \beta_3 sex_i + \varepsilon_i, \quad i = 1,2,...,n,$$

где обозначено $income_i$ – месячный доход, edu_i – уровень образования, age_i – возраст, sex_i – пол i-го индивидуума.

Для записи модели в матричной форме (5.4) определим:

$$y = \begin{pmatrix} income_1 \\ income_2 \\ \vdots \\ income_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & edu_1 & age_1 & sex_1 \\ 1 & edu_2 & age_2 & sex_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & edu_n & age_n & sex_n \end{pmatrix}, \; \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

В этой модели 3 объясняющие переменные, n наблюдений, которые дают информацию для формирования столбца y и матрицы *X*. Модель содержит постоянный член β_1 и три коэффициента $\beta_{2,}\beta_3,\beta_4$ при объясняющих переменных.

5.2. Метод наименьших квадратов

Для модели множественной линейной регрессии принцип наименьших квадратов так же, как и для парной линейной регрессии, состоит в оценке коэффициентов регрессии из условия минимума суммы квадратов остатков. Оцененное уравнение множественной линейной регрессии имеет вид:

$$\hat{y} = b_1 x_1 + b_2 x_2 + ... + b_k x_k \tag{5.5}$$

Запишем его для всех наблюдений:

$$\hat{y}_i = b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik}, \qquad i = 1,2,...,n.$$

Их также можно записать в матричной форме:

$$\hat{y} = Xb, \tag{5.6}$$

где столбцы $\hat{y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)'$, $b = (b_1, b_2, ..., b_k)'$. Здесь и далее штрих означает транспонирование. *Транспонированная строка* –

это столбец с элементами строки, расположенными в том же порядке. При транспонировании матрицы ее строки становятся соответствующими столбцами.

Определим столбец остатков – отклонений фактических значений y_i зависимой переменной от ее значений $\hat{y}_i$, вычисленных по оцененному уравнению регрессии (5.6):

$$e = y - \hat{y} = y - Xb, \qquad e = (e_1, e_2, \ldots e_n)'.$$

Метод наименьших квадратов заключается в определении коэффициентов оцененного уравнения (5.6) из условия минимума суммы квадратов остатков:

$$e'e = \sum_{i=1}^{n} e_i^2 \to \min.$$

Если столбцы матрицы X линейно независимы, то существует обратная матрица $(X'X)^{-1}$, и можно найти оценку вектора β коэффициентов регрессии:

$$b = (X'X)^{-1} X'y. \tag{5.7}$$

Он также обозначается через $\hat{\beta}_{OLS}$. В случае парной линейной регрессии отсюда можно получить оценки коэффициентов из раздела 4.

5.4. Проверка значимости коэффициентов регрессии

При практическом построении модели линейной регрессии существенен вопрос о значимости ее коэффициентов, вычисленных по конкретной выборке. Пусть выполнены условия $1 - 2'$. В частности, предполагается, что $\varepsilon_i \sim N(0, \sigma^2)$. Для заданного числа β_i^0, сформулируем две гипотезы:

нулевая $\quad H_0: \ \beta_i = \beta_i^0,$

альтернативная $H_1: \ \beta_i \neq \beta_i^0.$

Пусть b_i – оценка коэффициента β_i, а s_{b_i} – стандартная ошибка оценки b_i. Величина

$$\frac{b_i - \beta_i^0}{s_{b_i}} \sim t_{(n-k)},$$

т.е. имеет t – распределение Стьюдента с $n - \kappa$ степенями свободы, где n – число наблюдений, а κ – число оцениваемых коэффициентов уравнения (5.1), включая и свободный член.

Неизвестная дисперсия σ^2 заменяется на ее несмещенную оценку s^2. Стандартная ошибка оцененного коэффициента b_i вычисляется по формуле:

$$s_{b_i} = \sqrt{s^2 (X'X)^{-1}_{ii}},$$

где s – стандартная ошибка регрессии,

$$s^2 = \frac{1}{n-k} \sum_{i=1}^{n} e_i^2 .$$

Для выбранного числа δ (уровня значимости) по таблице t – распределения определяется критическое значение $t_c = t_{\delta/2, n-k}$, для которого вероятность реализации $t \sim t_{(n-k-1)}$, такого, что $-t_c \leq t \leq t_c$, равна $1 - \delta$. Затем проверяется условие:

$$-t_c \leq \frac{b_i - \beta_i^0}{s_{b_i}} \leq t_c .$$

Если оно не выполняется, гипотеза H_0 отвергается, а при его выполнении H_0 не отвергается (принимается).

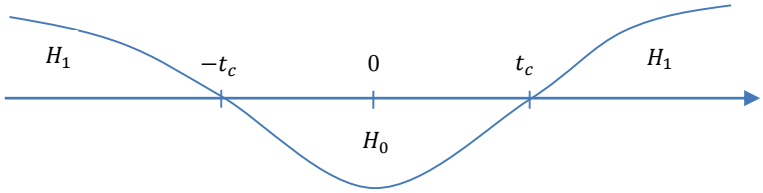

Рис. 5.2. Критическая область и область принятия гипотезы

На рис. 5.2 показаны на числовой оси критическая область, на которой не отвергается гипотеза H_0, и область, на которой принимается гипотеза H_1. Обычно коэффициент b_i сравнивают с $\beta_i^0 = 0$.

В случае если, например, при $\delta = 0.05$ гипотеза H_0 отвергается, то говорят, что коэффициент b_i значимый на 5%-ном уровне.

Значения стандартной ошибки s_{b_i} коэффициента b_i и соответствующая *t*-статистика t_{b_i} вычисляются в эконометрических пакетах.

П р и м е р 5.2. Результат оценивания может быть, например, записан так:

$$\hat{y} = \underset{(8.3)}{5.32} - \underset{(0.021)}{0.112w} + \underset{(11.7)}{21.4r} - \underset{(1.12)}{3.02u},$$

где под коэффициентами в скобках указаны соответствующие им стандартные ошибки. Допустим, что выборка, по которой проводилось оценивание, содержит n = 12 наблюдений.

По таблице распределения Стьюдента для $\delta = 0.05$ и $\delta = 0.01$ найдем критические значения:

$$t_{0.05/2,8} = 2.306\,, \qquad t_{0.01/2,8} = 3.355\,.$$

Отношения коэффициентов к своим стандартным ошибкам равны соответственно: 0.64, –5.33, 1.83, –2.7. Следовательно, сравнивая с критическими значениями, заключаем, что свободный член 5.32 и коэффициент 21.4 при переменной r незначимы на 5%-ном уровне, коэффициент –0.112 при переменной w значим на 1%-ном уровне, а коэффициент –2.7 при переменной u значим на 5%-ном уровне и незначим на 1%-ном уровне.