

Эконометрика

Полковников Александр Александрович

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

Конспект лекций для студентов направления
"Экономика"

§2. Множественная регрессия

Модель множественной регрессии — это уравнение, отражающее корреляционную связь между результатом и несколькими факторами. В общем виде оно может быть записано в виде:

$$y = f(x_0, \dots, x_m, e),$$

где

y — зависимая переменная (результат),

x_0, \dots, x_m — независимые переменные (факторы),

f — некоторая функция,

e — случайный остаток.

При проведении регрессионного анализа всегда предполагается, что наблюдения, на основе которых он проводится, были получены по однородной совокупности единиц. Для обеспечения статистической достоверности модели количество наблюдений должно быть в 8–10 раз больше количества параметров при факторах.

При построении модели предполагается, что факторы оказывают влияние на результат, причем влияние одного фактора не зависит от влияния других факторов. Однако, корреляционная связь может существовать и между факторами (т. е. присутствует, т. н. **мультиколлинеарность**).

Существование корреляционной связи между факторами может быть выявлено с помощью коэффициентов корреляции $r_{x_i x_j}$ между ними, которые можно записать, например, в виде **корреляционной матрицы**

$$r_{xx} = \begin{pmatrix} 1 & & & \\ r_{x_1 x_2} & 1 & & \\ \vdots & \vdots & \ddots & \\ r_{x_1 x_m} & r_{x_2 x_m} & \dots & 1 \end{pmatrix}.$$

Наличие мультиколлинеарности можно подтвердить, найдя определитель корреляционной матрицы. Если связь между факторами отсутствует, то определитель равен единице. Если связь между факторами является сильной, то определитель близок к нулю.

Дублирующие факторы с коэффициентом корреляции $|r_{x_i x_j}| \geq 0,7$ исключаются из модели. Предпочтение отдается тому фактору, который наименее связан с другими факторами. Существует несколько методов спецификации модели:

метод последовательного включения факторов предполагает, что сначала будет построена модель с фактором, наиболее тесно связанным с результатом, а затем поочередно добавляются другие факторы;

метод последовательного исключения факторов предполагает, что сначала строится модель с максимально большим количеством факторов, из которой затем поочередно исключаются незначимые факторы;

шаговый регрессионный анализ является продолжением метода включения. Построение модели начинается с расчета парной регрессии с фактором, наиболее тесно связанным с результатом. А затем добавление каждого нового фактора сопровождается не только оценкой значимости включения данного фактора, но и проверкой значимости факторов уже включенных в модель. Выявленные незначимые факторы исключаются из модели.

ступенчатый регрессионный анализ начинается также с расчета парной регрессии с фактором, наиболее тесно связанным с результатом. Затем вычисляются регрессионные остатки и строится уравнение их зависимости от следующего по степени влияния на результат фактора. По этому уравнению опять вычисляются остатки и процедура повторяется до тех пор пока получаются значимые уравнения.

Говоря о **линейных** эконометрических моделях с несколькими объясняющими переменными, мы фактически исходим о существовании теоретического соотношения

$$y = \theta_0 x_0 + \dots + \theta_m x_m + e$$

между переменными y и x_0, \dots, x_m . Если в правую часть такого соотношения включается константа, то в качестве x_0 тождественно берется единица.

Обращенная к статистическим данным линейная эконометрическая модель с $(m + 1)$ -й объясняющей переменной имеет вид:

$$y_i = \theta_0 x_{i0} + \dots + \theta_m x_{im} + e_i, \quad i = 1, \dots, n,$$

где

y_i — значение результата в i -м наблюдении,

x_{ij} — значение j -го фактора в i -м наблюдении,

θ_j — коэффициент при j -м факторе,

n — количество наблюдений,

e_i — случайная ошибка в i -м наблюдении.

Запишем линейную модель в матричной форме. Для этого обозначим через

$$y = (y_1, \dots, y_n)^T, \quad \theta = (\theta_0, \dots, \theta_m)^T, \quad e = (e_1, \dots, e_n)^T,$$

$$X = \begin{pmatrix} x_{10} & x_{11} & \dots & x_{1m} \\ x_{20} & x_{21} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \dots & x_{nm} \end{pmatrix}, \quad X^T = \begin{pmatrix} x_{10} & x_{20} & \dots & x_{n0} \\ x_{11} & x_{21} & \dots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1m} & x_{2m} & \dots & x_{nm} \end{pmatrix}.$$

Тогда линейную модель можно записать в матричном виде:

$$y = X \cdot \theta + e.$$

При $m = 1$ получаем модель парной линейной регрессии с объясняющими переменными:

$$x_0 \equiv 1, \quad x_1 = x.$$

и матрицей X :

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad X^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix}.$$

Оценивание неизвестных коэффициентов модели можно провести методом наименьших квадратов

$$\sum_{i=1}^n \left(y_i - \sum_{j=0}^m \theta_j x_{ij} \right)^2 = (y - X\theta)^T (y - X\theta) \rightarrow \min .$$

Приравнивание к нулю частных производных этой суммы по каждой из переменных $\theta_0, \dots, \theta_m$ приводит к системе нормальных уравнений, которую можно записать в матричном виде

$$X^T X \theta = X^T y.$$

Система имеет единственное решение, которое указывает на точку минимума, если выполнено условие **идентификации**

$$\det(X^T X) \neq 0.$$

Решение системы нормальных уравнений имеет вид:

$$\hat{\theta} = (X^T X)^{-1} X^T y.$$

Обозначая через

$$\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T$$

вектор прогнозных значений результата, получим

$$\hat{y} = X\hat{\theta} = X(X^T X)^{-1} X^T y.$$

Вектор остатков e можно записать

$$e = y - \hat{y} = y - X(X^T X)^{-1} X^T y = \left(I_n - X(X^T X)^{-1} X^T \right) y.$$

Для остаточной суммы квадратов получаем

$$RSS = |e|^2 = e^T e = (y - X\hat{\theta})^T (y - X\hat{\theta}) =$$

$$= y^T y - \hat{\theta}^T X^T y - y^T X \hat{\theta} + \hat{\theta}^T X^T X \hat{\theta} =$$

(поскольку $y^T X \hat{\theta}$ — скаляр, то $y^T X \hat{\theta} = \hat{\theta}^T X^T y$)

$$= |y|^2 - \hat{\theta}^T X^T y - \hat{\theta}^T X^T y + \hat{\theta}^T X^T X \hat{\theta} =$$

$$= |y|^2 - \hat{\theta}^T X^T y + \hat{\theta}^T (X^T X \hat{\theta} - X^T y).$$

Из соотношения (система нормальных уравнений)

$$X^T X \hat{\theta} = X^T y$$

имеем

$$\begin{aligned} RSS &= |y|^2 - \hat{\theta}^T X^T y = |y|^2 - \hat{y}^T (\hat{y} + e) = \\ &= |y|^2 - |\hat{y}|^2 - \hat{y}^T e = |y|^2 - |\hat{y}|^2 \end{aligned}$$

теорему Пифагора в \mathbb{R}^n

$$|y|^2 = |\hat{y}|^2 + |e|^2.$$

При $m = 1$ матрица $(X^T X)^{-1}$ имеет вид:

$$(X^T X)^{-1} = \frac{1}{n \sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k \right)^2} \begin{pmatrix} \sum_{k=1}^n x_k^2 & - \sum_{k=1}^n x_k \\ - \sum_{k=1}^n x_k & n \end{pmatrix}.$$

Уравнение регрессии, построенное по исходным данным, называется моделью **в натуральной форме**.

Если же провести предварительную стандартизацию переменных, входящих в модель, т. е. выполнить нормировку

$$y^* = \frac{y - \bar{y}}{\sqrt{S_y^2}}, \quad x_k^* = \frac{x_k - \bar{x}_k}{\sqrt{S_{x_k}^2}}, \quad k = 1, \dots, m$$

а затем построить по новым переменным модель множественной линейной регрессии

$$\hat{y}^* = \beta_1 x_1^* + \dots + \beta_m x_m^*$$

то такая модель будет называться **моделью в стандартизированной форме**.

Применение метода наименьших квадратов к стандартизированной форме модели дает следующую систему нормальных уравнений

$$\left\{ \begin{array}{ccccccc} \beta_1 & + & \beta_2 r_{x_1 x_2} & + & \dots & \beta_m r_{x_1 x_m} & = & r_{y x_1}, \\ \beta_1 r_{x_1 x_2} & + & \beta_2 & + & \dots & \beta_m r_{x_2 x_m} & = & r_{y x_2}, \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_1 r_{x_1 x_m} & + & \beta_2 r_{x_2 x_m} & + & \dots & \beta_m & = & r_{y x_m}. \end{array} \right.$$

Последняя система уравнений может быть записана в матричном виде

$$r_{xx}\beta = r_{yx},$$

где

$$r_{xx} = \begin{pmatrix} 1 & & & \\ r_{x_1x_2} & 1 & & \\ \vdots & \vdots & \ddots & \\ r_{x_1x_m} & r_{x_2x_m} & \dots & 1 \end{pmatrix}$$

— матрица межфакторной корреляции,

$\beta = (\beta_1, \dots, \beta_m)^T$ — вектор стандартизированных коэффициентов,

$r_{yx} = (r_{yx_1}, \dots, r_{yx_m})^T$ — вектор коэффициентов корреляции между результатом y и факторами.

Зная коэффициенты β_j стандартизированного уравнения можно найти коэффициенты “чистой” регрессии θ_j :

$$\theta_j = \beta_j \sqrt{\frac{S_y^2}{S_{x_j}^2}}, \quad j = 1, \dots, m,$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}_1 - \dots - \theta_m \bar{x}_m.$$

Отсюда, стандартизированные коэффициенты регрессии β_j показывают на сколько среднеквадратических отклонений S_y в среднем изменится результат y при изменении фактора x_j на одно среднеквадратичное отклонение S_{x_j} при фиксированном уровне других факторов, включенных в модель.

СПАСИБО ЗА ВНИМАНИЕ!