МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра «Теория рынка»

Тимофеев В.С.

ОСНОВЫ ЭКОНОМЕТРИКИ
(Раздел 3. парная регрессия)

теоретические материалы для студентов ОФиП

Новосибирск, 2008

3. МОДЕЛЬ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

3.1. Постановка задачи

Пусть имеются наблюдения за двумя признаками: *X* и *Y*. Допустим, что проведенный ранее корреляционный анализ показал наличие значимой линейной связи между этими признаками. Каждое *i*-е наблюдение можно изобразить на плоскости (*X*,*Y*) в виде точки с координатами (x_i, y_i), как это представлено на рис.5. Наша задача будет состоять в том, чтобы найти такое уравнение линейной зависимости, которое бы «наилучшим образом» описывало исходные данные.

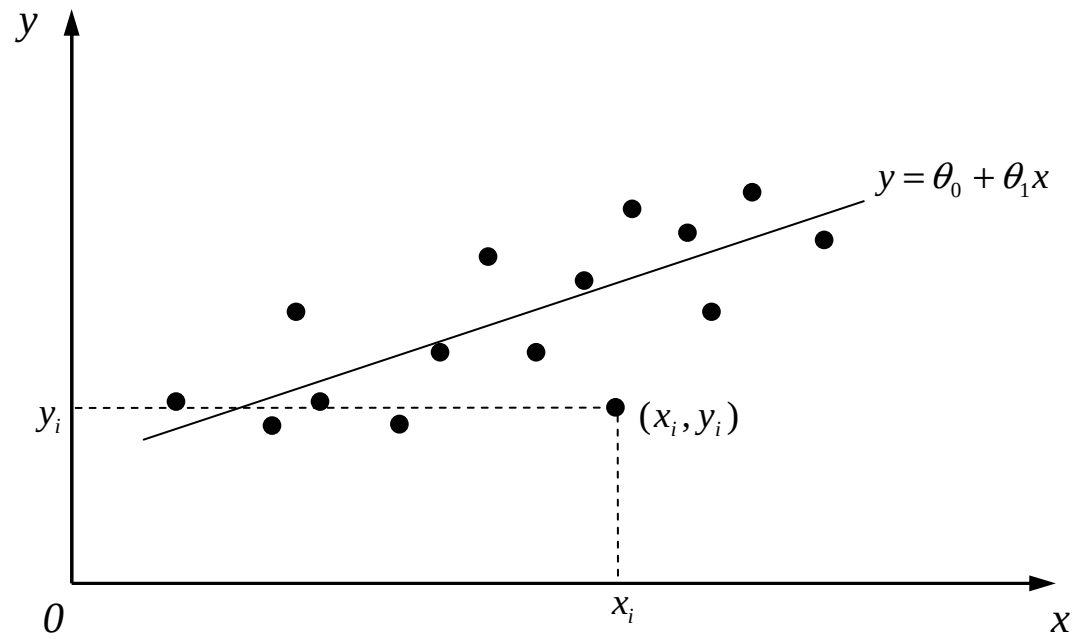


Рисунок 5. Графическое представление исходных данных при построении парной регрессии

Для определенности будем считать, что *Y* – это зависимая переменная, а *X* – независимая и постараемся найти аналитическое уравнение, математически описывающее эту зависимость. Тогда общий вид модели для

каждого наблюдения исходных данных, можно записать следующим образом:

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i,\ i = 1, 2, \ldots, N, \qquad (9)$$

где N – общее число наблюдений; y_i – значение признака (переменной) Y в i-м наблюдении; x_i – значение признака (переменной) X в i-м наблюдении; ε_i – случайная величина (ошибка наблюдения); θ_0, θ_1 – неизвестные параметры.

Уравнение (9) называется уравнением парной регрессии или в общем случае регрессионным уравнением. Переменные x_i являются детерминированными (неслучайными) величинами, а y_i – стохастическими (случайными) величинами в силу случайности ε_i. Присутствие в регрессионном уравнении случайной величины ε_i является обязательным. Этот факт можно объяснить целым рядом причин. Одна из них связана с тем, что наша модель является упрощением действительности, и, следовательно, в ней могут отсутствовать факторы, оказывающие влияние на отклик. Например, оборот предприятия может зависеть не только от цен на продукцию, но и от экономической ситуации в стране, действий конкурентов и ряда других факторов, которые либо не поддаются измерению, либо оказались упущены на постановочном этапе анализа. Другая причина может быть связана с наличием ошибок (погрешностей) при сборе и регистрации статистических данных.

Соотношением (9) определяется уравнение не одной прямой, а целое семейство. При этом количество возможных уравнений в семействе бесконечно и каждое уравнение из этого семейства отличается от любого другого своими значениями параметров θ_0, θ_1. Таким образом, выбор «наилучшего» уравнения сводится к выбору конкретных значений параметров θ_0 и θ_1.

3.2. Методы решения

Чтобы формализовать процесс выбора «наилучших» значений неизвестных параметров, необходимо ввести некоторый критерий, позволяющий однозначно определять степень отклонения модели от исходных данных. Естественно, что таких критериев может быть очень много. Приведем критерии, получившие наибольшее распространение.

1. Сумма квадратов отклонений (принцип Лежандра [ ])

$$F = \sum_{i=1}^{N} \left(y_i - (\theta_0 + \theta_1 x_i) \right)^2 .$$

Согласно этому критерию наилучшими считаются такие значения неизвестных параметров, которые соответствуют минимальному значению суммы квадратов отклонений значений зависимой переменной, рассчитанных по уравнению регрессии, от наблюдаемых значений зависимой переменной. Метод определения значений неизвестных параметров на основе этого критерия получил название «Метод наименьших квадратов (МНК)».

К преимуществам этого метода обычно относят простоту вычислительных процедур, положенных в основу метода, а также хорошие статистические свойства получаемых оценок. Недостатками можно считать чувствительность оценок к появлению грубых ошибок в исходных данных и необходимость постулирования нормального распределения случайной величины ε_i для получения некоторых дополнительных результатов.

2. Сумма модулей отклонений

$$F = \sum_{i=1}^{N} \left| y_i - (\theta_0 + \theta_1 x_i) \right| .$$

Согласно этому критерию наилучшими считаются такие значения неизвестных параметров, которые соответствуют минимальному значению суммы абсолютных величин отклонений значений зависимой переменной, рассчитанных по уравнению регрессии, от наблюдаемых значений зависимой переменной. Процесс определения значений неизвестных параметров на

основе этого критерия получил название «Метод наименьших модулей (МНМ)».

К преимуществам этого метода обычно относят значительно меньшую чувствительность результатов к появлению грубых ошибок в исходных данных, а к недостаткам – сложность вычислительных процедур, сопровождающих поиск оценок неизвестных параметров и возможность возникновения ситуаций, когда нет однозначного решения.

3. Обобщенный критерий

В общем случае для решения задачи определения неизвестных параметров возможно построение других критериев:

$$F = \sum_{i=1}^{N} g\left(y_i - (\theta_0 + \theta_1 x_i)\right),$$

где $g(u)$ – некоторая функция меры (как правило, чётная), определяющая величину отклонения значений зависимой переменной, рассчитанных по уравнению регрессии, от наблюдаемых значений зависимой переменной.

Преимущества и недостатки этого метода зависят от свойств функции меры $g(u)$.

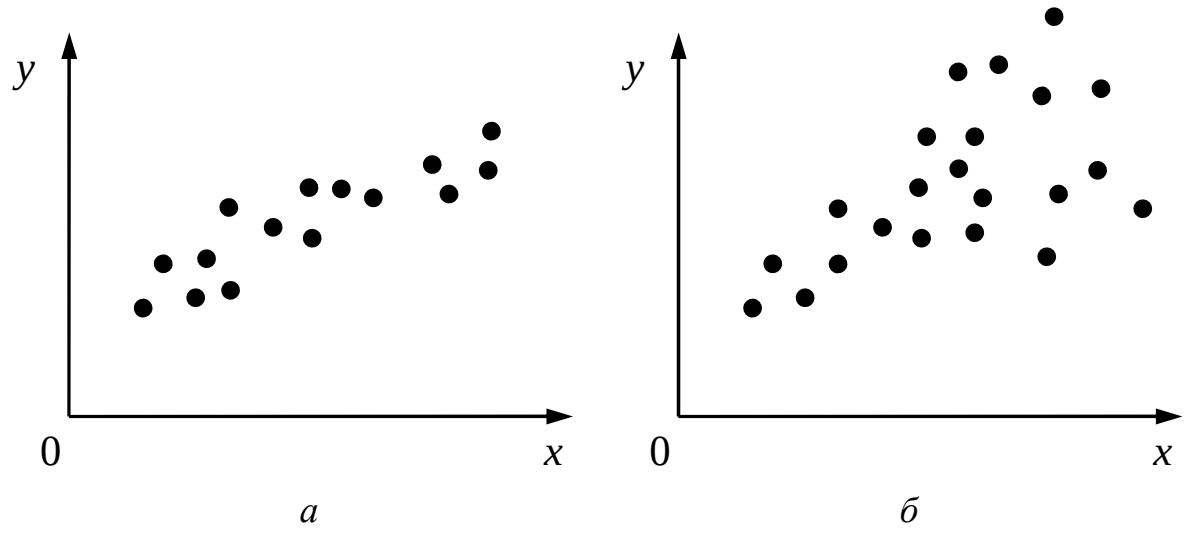
Следует отметить, что на сегодняшний день наиболее широкое распространение на практике получил метод наименьших квадратов, поэтому рассмотрим его более подробно.

3.3. Оценивание параметров методом наименьших квадратов

Для того чтобы МНК приводил к получению корректных результатов, необходимо выполнение ряда предположений о случайной составляющей уравнения (9) – ε_i.

1. $E[\varepsilon_i] = 0$, $i = 1, 2, \ldots, N$. Это условие означает, что $E[y_i] = \theta_0 + \theta_1 x_i$, т.е. при фиксированном x_i среднее ожидаемое значение отклика равно $\theta_0 + \theta_1 x_i$.

2. $E\left[\varepsilon_i^2\right] = D[\varepsilon_i] = \sigma^2$, $i = 1,2,\ldots,N$. Это условие говорит о том, что дисперсия ошибки для всех наблюдений одинакова. Это условие называется условием гомоскедастичности. Случай, когда это условие не выполняется, называется гетероскедастичностью (см. рис. 6). Также из данного условия следует, что $D[y_i] = \sigma^2$, $i = 1,2,\ldots,N$.



а *б*

Рис. 6. Вид исходных данных при гомоскедастичности (а) и гетероскедастичности (б)

3. $E\left[\varepsilon_i \varepsilon_j\right] = 0$, $i \neq j$. Это условие указывает на некоррелированность ошибок для разных наблюдений. В случае, когда это условие не выполняется, говорят об автокорреляции ошибок.

Этих трех предположений достаточно для того, чтобы можно было использовать метод наименьших квадратов для оценивания параметров модели (9). Но если необходимо проводить более полный эконометрический анализ, то потребуется еще одно предположение – о нормальном распределении ошибок.

4. $\varepsilon_i \sim N\left(0, \sigma^2\right)$, $i = 1,2,\ldots,N$. При выполнении этого условия модель (9) называют нормальной линейной регрессионной моделью.

В соответствии с основной идеей метода наименьших квадратов, поиск неизвестных параметров уравнения (9) производится как решение задачи на экстремум [3]:

$$F = \sum_{i=1}^{N} \left(y_i - (\theta_0 + \theta_1 x_i) \right)^2 \to \min_{\theta_0, \theta_1}. \tag{10}$$

Запишем необходимое условие экстремума:

$$\begin{cases} \dfrac{\partial F}{\partial \theta_0} = -2 \sum\limits_{i=1}^{N} \left(y_i - \theta_0 - \theta_1 x_i \right) = 0, \\ \dfrac{\partial F}{\partial \theta_1} = -2 \sum\limits_{i=1}^{N} x_i \left(y_i - \theta_0 - \theta_1 x_i \right) = 0. \end{cases} . \tag{11}$$

После раскрытия скобок и приведения подобных получим систему нормальных уравнений:

$$\begin{cases} \theta_0 N + \theta_1 \sum\limits_{i=1}^{N} x_i = \sum\limits_{i=1}^{N} y_i, \\ \theta_0 \sum\limits_{i=1}^{N} x_i + \theta_1 \sum\limits_{i=1}^{N} x_i^2 = \sum\limits_{i=1}^{N} x_i y_i. \end{cases} . \tag{12}$$

Решением этой системы являются оценки неизвестных параметров:

$$\hat{\theta}_1 = \frac{\sum\limits_{i=1}^{N} \left(x_i - \overline{x} \right) \left(y_i - \overline{y} \right)}{\sum\limits_{i=1}^{N} \left(x_i - \overline{x} \right)^2} = \frac{N \sum\limits_{i=1}^{N} x_i y_i - \left(\sum\limits_{i=1}^{N} x_i \right) \cdot \left(\sum\limits_{i=1}^{N} y_i \right)}{N \sum\limits_{i=1}^{N} x_i^2 - \left(\sum\limits_{i=1}^{N} x_i \right)^2} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \left(\overline{x} \right)^2}, \tag{13}$$

$$\hat{\theta}_0 = \frac{1}{N} \sum_{i=1}^{N} y_i - \frac{1}{N} \sum_{i=1}^{N} x_i \hat{\theta}_1 = \overline{y} - \overline{x} \hat{\theta}_1. \tag{14}$$

Здесь следует особое внимание обратить на разницу между такими понятиями, как «неизвестные параметры θ_0 и θ_1» и «оценки неизвестных параметров $\hat{\theta}_0$ и $\hat{\theta}_1$». Оценки параметров по своей сути являются случайными величинами и зависят от исходных данных. Каждое значение, полученное по соотношениям (13) и (14), можно рассматривать как отдельную реализацию случайной величины. Истинные же значения

параметров всегда остаются неизвестными. На практике мы имеем дело ТОЛЬКО с оценками, степень доверия к которым может быть различной.

3.4. Интерпретация параметров уравнения парной регрессии

Возможность четкой экономической интерпретации неизвестных параметров уравнения регрессии сделало модель парной регрессии достаточно распространенной в эконометрических исследованиях.

Величина параметра θ_1 показывает среднее изменение отклика с изменением регрессора на одну единицу.

Что касается параметра θ_0, то формально можно сказать, что это значение y при $x = 0$. Однако часто этот параметр вообще может не иметь экономического содержания. Если регрессор по своему смыслу не имеет и не может иметь нулевого значения, то такая трактовка не имеет смысла. Попытки экономической интерпретации параметра θ_0 иногда могут приводить к абсурдным результатам (особенно при $\theta_0 < 0$).

Однако в любом случае можно проинтерпретировать знак параметра θ_0. Так, если $\theta_0 < 0$, то относительное изменение отклика оказывается меньше относительного изменения регрессора. И, наоборот, если $\theta_0 > 0$, то относительное изменение отклика оказывается больше относительного изменения регрессора. Для доказательства этого утверждения достаточно рассмотреть относительные изменения x и y. Пусть $\frac{dy}{y} < \frac{dx}{x}$, тогда

$$\frac{dy}{dx} < \frac{y}{x},\ \frac{\theta_1 dx}{dx} < \frac{\theta_0 + \theta_1 x}{x},\ \theta_1 x < \theta_0 + \theta_1 x,$$

откуда $\theta_0 > 0$. Для случая $\theta_0 < 0$ доказательство проводится аналогично.

Рассмотрим пример. Пусть зависимость оборота фирмы от затрат на рекламу описывается уравнением $\hat{y}_i = 10 + 2.8 x_i$ (y – оборот (млн. руб.), x – затраты на рекламу (тыс. руб.)). Тогда значение оценки $\hat{\theta}_1 = 2.8$ означает, что

увеличение рекламных затрат на одну тысячу рублей влечет за собой увеличение оборота в среднем на 2.8 млн. руб. Значение оценки $\hat{\theta}_0 = 10$ может означать среднюю величину оборота при полном отсутствии рекламы. Положительный знак $\hat{\theta}_0$ означает, что относительное изменение затрат на рекламу больше относительного изменения оборота. Действительно, пусть затраты на рекламу увеличились с 10 тыс. руб. до 11 тыс. руб., тогда относительное изменение затрат на рекламу есть

$$\frac{dx}{x} = \frac{11-10}{10} = 0.1.$$

Аналогичным образом можно вычислить относительное изменение оборота

$$\frac{dy}{y} = \frac{(10 + 2.8 \cdot 11) - (10 + 2.8 \cdot 10)}{(10 + 2.8 \cdot 10)} = \frac{40.8 - 38}{38} = 0.07.$$

Как видно из полученных результатов, увеличение затрат на рекламу на 10 % влечет за собой среднее увеличение оборота на 7 %.

3.5. Свойства МНК-оценок коэффициентов регрессии

Как уже отмечалось выше, оценки параметров $\hat{\theta}_0$ и $\hat{\theta}_1$ являются непрерывными случайными величинами, которые обладают всеми характеристиками случайных величин. Сейчас нас будут интересовать такие характеристики как математическое ожидание и дисперсия.

Нетрудно заметить, что математическое ожидание оценок параметров совпадает с истинными значениями этих параметров, т.е. МНК-оценки являются несмещенными:

$$E\left(\hat{\theta}_1\right) = E\left(\frac{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}\right) = \frac{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)E\left(y_i - \overline{y}\right)}{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2} =$$

$$= \frac{\sum_{i=1}^{N}(x_i - \overline{x})(\theta_0 + \theta_1 x_i - \theta_0 - \theta_1 \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})\theta_1 (x_i - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \theta_1,$$

$$E\left(\hat{\theta}_0\right) = E\left(\overline{y} - \overline{x}\hat{\theta}_1\right) = E\left(\overline{y}\right) - \overline{x}E\left(\hat{\theta}_1\right) = \theta_0 + \theta_1\overline{x} - \theta_1\overline{x} = \theta_0.$$

Для дальнейшего удобства изложения введем дополнительное обозначение:

$$z_i = \frac{(x_i - \overline{x})}{\sum_{k=1}^{N}(x_i - \overline{x})^2},$$

с учетом которого можно легко записать

$$\hat{\theta}_1 = \sum_{i=1}^{N} z_i (y_i - \overline{y}). \qquad (15)$$

Тогда дисперсии оценок будут определяться как

$$D\left(\hat{\theta}_1\right) = D\left(\sum_{i=1}^{N} z_i (y_i - \overline{y})\right) = D\left(\sum_{i=1}^{N} z_i y_i\right) = \sum_{i=1}^{N} {z_i}^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}, \quad (16)$$

$$D\left(\hat{\theta}_0\right) = D\left(\sum_{i=1}^{N}\left(\frac{1}{N} - \overline{x}z_i\right)y_i\right) = \sigma^2 \sum_{i=1}^{N}\left(\frac{1}{N} - \overline{x}z_i\right)^2 = \sigma^2\left(\frac{1}{N} + \frac{\overline{x}^2}{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}\right) =$$

$$= \sigma^2 \frac{\sum_{i=1}^{N} {x_i}^2}{N\sum_{i=1}^{N}(x_i - \overline{x})^2}. \qquad (17)$$

Как видно из этих соотношений, дисперсии оценок $\hat{\theta}_0$ и $\hat{\theta}_1$ зависят от дисперсии ошибки σ^2, кроме этого, на точность определения параметров также оказывает влияние и расположение точек, в которых проводились измерения.

Знание дисперсий оценок позволяет проводить их сравнение по точности – более точным оценкам соответствует меньшая дисперсия. С этой точки зрения МНК-оценки обладают рядом полезных свойств, сформулированных в виде теоремы Гаусса-Маркова [7, 14, 20].

Теорема Гаусса-Маркова. *Для регрессионной модели*

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i,\ i = 1,2,\ldots,N,$$

где для случайной ошибки выполнены условия:

$$E[\varepsilon_i] = 0,\ i = 1,2,\ldots,N,$$

$$E\left[\varepsilon_i^2\right] = D[\varepsilon_i] = \sigma^2,\ i = 1,2,\ldots,N,$$

$$E[\varepsilon_i \varepsilon_j] = 0,\ i \neq j$$

оценки $\hat{\theta}_0$ *и* $\hat{\theta}_1$*, полученные по МНК, имеют наименьшую дисперсию в классе всех линейных (по* y_i*) несмещенных оценок.*

Доказательство теоремы можно найти в [14].

Следует обязательно отметить, что здесь речь идет только о линейных оценках. Другими словами, никто не гарантирует, что не существует нелинейной несмещенной оценки с дисперсией меньшей, чем у МНК-оценки. То же самое можно сказать и о смещенных оценках.

3.6. Оценка дисперсии ошибки

Список неизвестных параметров модели парной регрессии не ограничивается только коэффициентами уравнения (9). Есть еще один неизвестный параметр – это дисперсия ошибок σ^2. Без информации об этом параметре проведение эконометрического анализа в полном объеме становится невозможным.

На практике значение дисперсии ошибки бывает известно крайне редко, поэтому, как и в случае с коэффициентами регрессии, приходится работать с ее оценкой.

Для оценивания дисперсии ошибки нам потребуется ввести еще одно понятие – остатки модели. Под остатками модели будем понимать отклонения наблюдаемых значений отклика от предсказанных по уравнению модели (рис. 7).

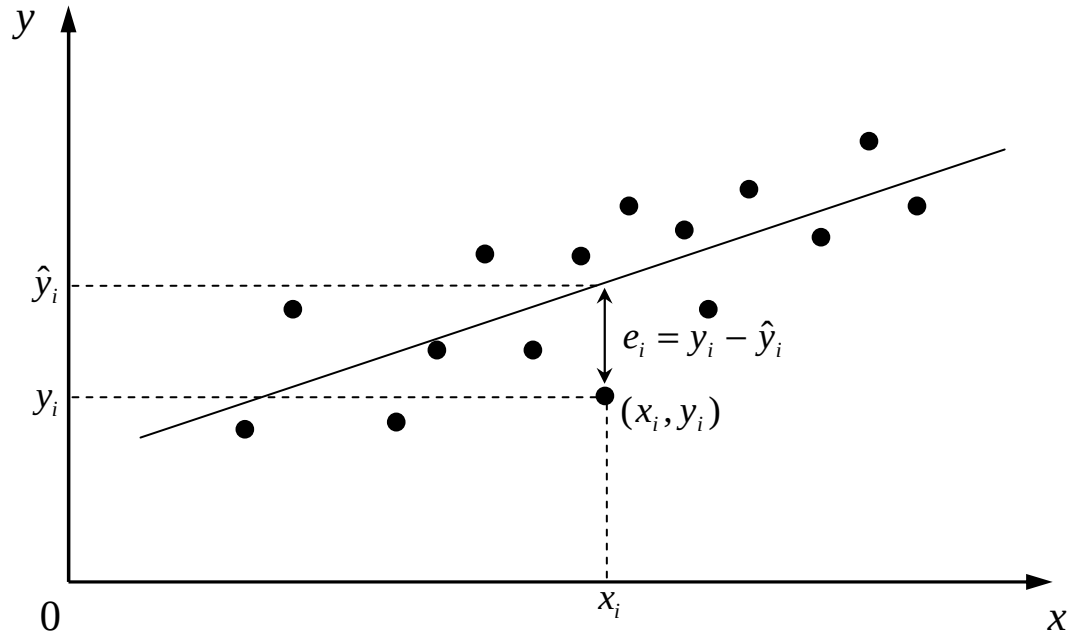


Рис. 7. Остатки уравнения регрессии

Через $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$ обозначен прогноз значения y_i в точке x_i, и остатки регрессии e_i определяются из уравнения

$$y_i = \hat{y}_i + e_i = \hat{\theta}_0 + \hat{\theta}_1 x_i + e_i .$$

Остатки могут быть определены для каждой точки исходных данных. Однако не следует путать остатки регрессии со случайными ошибками в уравнении (9). В отличие от ошибок остатки наблюдаемы, т.е. могут быть вычислены.

Одной из особенностей процедуры метода наименьших квадратов является тот факт, что сумма остатков модели всегда нулевая

$$\sum_{i=1}^{N} e_i = 0 .$$

Остатки e_i также можно рассматривать и как оценки случайных ошибок ε_i, что позволяет их использовать при оценивании дисперсии ошибки σ^2. Рассмотрим, каким образом можно представить сумму квадратов остатков:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i \right)^2 = \left| \begin{matrix} x_i' = x_i - \overline{x} \\ y_i' = y_i - \overline{y} \end{matrix} \right| = \sum_{i=1}^{N} \left(\overline{y} + y_i' - \hat{\theta}_0 - \hat{\theta}_1 \overline{x} - \hat{\theta}_1 x'_i \right)^2 =$$

$$\sum_{i=1}^{N} \left(y_i' - \hat{\theta}_1 x_i' \right)^2 = \sum_{i=1}^{N} \left(\theta_1 x_i' + \varepsilon_i - \overline{\varepsilon} - \hat{\theta}_1 x_i' \right)^2 = \sum_{i=1}^{N} \left(\left(\theta_1 - \hat{\theta}_1 \right) x_i' + \left(\varepsilon_i - \overline{\varepsilon} \right) \right)^2 =$$

$$= \sum_{i=1}^{N} x_i'^2 \left(\theta_1 - \hat{\theta}_1 \right)^2 + 2 \left(\theta_1 - \hat{\theta}_1 \right) \sum_{i=1}^{N} x_i' \left(\varepsilon_i - \overline{\varepsilon} \right) + \sum_{i=1}^{N} \left(\varepsilon_i - \overline{\varepsilon} \right)^2 .$$

Теперь вычислим математическое ожидание каждого из полученных слагаемых.

$$E\left(\sum_{i=1}^{N} x_i'^2 \left(\theta_1 - \hat{\theta}_1 \right)^2 \right) = D\left(\hat{\theta}_1 \right) \sum_{i=1}^{N} x_i'^2 = \frac{\sigma^2}{\sum_{i=1}^{N} x_i'^2} \sum_{i=1}^{N} x_i'^2 = \sigma^2$$

$$E\left(2 \left(\theta_1 - \hat{\theta}_1 \right) \sum_{i=1}^{N} x_i' \left(\varepsilon_i - \overline{\varepsilon} \right) \right) = -2E\left(\sum_{j=1}^{N} \varepsilon_j z_j \sum_{i=1}^{N} x_i' \left(\varepsilon_i - \overline{\varepsilon} \right) \right) =$$

$$= -2E\left(\sum_{i=1}^{N} \sum_{j=1}^{N} x_i' \varepsilon_i z_j \varepsilon_j - \overline{\varepsilon} \sum_{j=1}^{N} z_j \varepsilon_j \sum_{i=1}^{N} x_i' \right) = -2 \sum_{i=1}^{N} x_i' z_i \sigma^2 = -2\sigma^2$$

$$E\left(\sum_{i=1}^{N} \left(\varepsilon_i - \overline{\varepsilon} \right)^2 \right) = E\left(\sum_{i=1}^{N} \varepsilon_i{}^2 - 2\overline{\varepsilon} \sum_{i=1}^{N} \varepsilon_i + \overline{\varepsilon}^2 N \right) =$$

$$= N\sigma^2 - 2N \frac{1}{N} \sigma^2 + N \frac{1}{N} \sigma^2 = (N-1)\sigma^2 .$$

Отсюда очевидно, что

$$\sum_{i=1}^{N} e_i^2 = \sigma^2 - 2\sigma^2 + (N-1)\sigma^2 = (N-2)\sigma^2 ,$$

следовательно, несмещенная оценка дисперсии ошибок принимает вид

$$\hat{\sigma}^2 = S^2 = \frac{1}{N-2}\sum_{i=1}^{N} e_i^2 . \qquad (18)$$

Кроме того, можно показать (см. [14, 17]), что оценка дисперсии ошибки S^2 и МНК-оценки $\hat{\theta}_0$ и $\hat{\theta}_1$ статистически независимы.

3.7. Статистические свойства оценок параметров. Распределения основных статистик

Дальнейшее изложение будет справедливо только для нормальной линейной регрессионной модели, т.е. для случая выполнения условия о нормальном распределении случайных ошибок. В этом случае величины y_i также имеют нормальное распределение:

$$y_i \sim N\left(\theta_0 + \theta_1 x_i; \sigma^2\right),\ i = 1, 2, \ldots, N .$$

Так как МНК-оценки (13), (14) являются линейными функциями от y_i, то их распределение также будет нормальным

$$\hat{\theta}_0 \sim N\left(\theta_0, \sigma^2 \frac{\sum_{i=1}^{N} x_i^2}{N\sum_{i=1}^{N} x_i'^2}\right), \qquad (19)$$

$$\hat{\theta}_1 \sim N\left(\theta_1, \sigma^2 \frac{1}{\sum_{i=1}^{N} x_i'^2}\right). \qquad (20)$$

Если по каким-либо причинам предположение о нормальности ошибок не выполняется, то утверждения (19), (20), вообще говоря, неверны. Однако если исходные данные удовлетворяют некоторым специальным условиям регулярности, то распределение оценок $\hat{\theta}_0$ и $\hat{\theta}_1$ будет асимптотически нормальным. То есть при $N \to \infty$ распределение оценок будет стремиться к нормальному закону [1, 17].

Из (19), (20) следует, что

$$\hat{\theta}_0 - \theta_0 \sim N\left(0, \sigma^2_{\hat{\theta}_0}\right), \tag{21}$$

$$\hat{\theta}_1 - \theta_1 \sim N\left(0, \sigma^2_{\hat{\theta}_1}\right), \tag{22}$$

следовательно,

$$\frac{\hat{\theta}_0 - \theta_0}{\sigma_{\hat{\theta}_0}} \sim N(0,1), \tag{23}$$

$$\frac{\hat{\theta}_1 - \theta_1}{\sigma_{\hat{\theta}_1}} \sim N(0,1), \tag{24}$$

где $\sigma^2_{\hat{\theta}_0} = D(\hat{\theta}_0)$, $\sigma^2_{\hat{\theta}_1} = D(\hat{\theta}_1)$, $N(0,1)$ – стандартное нормальное распределение.

Оценки дисперсий оценок $\hat{\theta}_0$ и $\hat{\theta}_1$ могут быть получены из формул (16), (17) после замены дисперсии ошибок σ^2 на ее оценку (18):

$$\hat{\sigma}^2_{\hat{\theta}_1} = S^2_{\hat{\theta}_1} = \frac{S^2}{\sum_{i=1}^{N} x_i'^2}, \tag{25}$$

$$\hat{\sigma}^2_{\hat{\theta}_0} = S^2_{\hat{\theta}_0} = S^2 \frac{\sum_{i=1}^{N} x_i^2}{N\sum_{i=1}^{N} x_i'^2}. \tag{26}$$

Из теории математической статистики [1, 14] известно, что

$$\frac{(N-2)S^2}{\sigma^2} \sim \chi^2(N-2), \tag{27}$$

где $\chi^2(N-2)$ – распределение «хи-квадрат» с $(N-2)$ степенями свободы.

С учетом статистической независимости параметров линейной регрессионной модели, а также соотношений (23), (24), (27) будет справедливо записать, что

$$\frac{\left(\hat{\theta}_1 - \theta_1\right)/\sigma_{\hat{\theta}_1}}{S/\sigma} \sim t\left(N-2\right), \qquad (28)$$

где $t\left(N-2\right)$ – распределение Стьюдента с $(N-2)$ степенями свободы.

Выражение (28) является основой для построения процедур проверки статистических гипотез относительно параметров регрессионной модели. Если в (28) заменить неизвестные истинные значения дисперсий их оценками, получим

$$\frac{\left(\hat{\theta}_1 - \theta_1\right)}{S_{\hat{\theta}_1}} \sim t\left(N-2\right). \qquad (29)$$

Величина

$$t = \frac{\left(\hat{\theta}_1 - \theta_1\right)}{S_{\hat{\theta}_1}} \qquad (30)$$

известна в математической статистике как «статистика Стьюдента» [9,14,20,21].

Аналогичным образом можно показать, что величина

$$t = \frac{\left(\hat{\theta}_0 - \theta_0\right)}{S_{\hat{\theta}_0}} \qquad (31)$$

также распределена по закону Стьюдента $t\left(N-2\right)$.

Следует отметить такой важный факт, что при выполнении условий регулярности исходных данных и при большом количестве наблюдений, введенные *t*-статистики (30), (31) будут распределены по закону Стьюдента $t\left(N-2\right)$ и без предположения о нормальности случайных ошибок [14].

3.8. Проверка статистических гипотез о параметрах. Доверительные интервалы

Относительно параметров θ_0 и θ_1 линейной модели можно проверять некоторые утверждения, оформленные в виде статистических гипотез.

Основные гипотезы регрессионного анализа относительно параметров θ_0 и θ_1 линейной модели можно условно разделить на два типа. К первому типу будем относить гипотезы о значении параметров. Для параметра θ_1 в общем случае эта гипотеза записывается в виде

$$H_0 : \theta_1 = \theta_{10}, \qquad (32)$$

где θ_{10} – некоторое заданное значение.

Если вернуться к примеру из п.3.4, то при исследовании зависимости величины оборота от затрат на рекламу нас может интересовать вопрос: «Можно ли считать эффективность рекламной кампании такой, что каждая вложенная тысяча средств дает прирост оборота на 3 млн. рублей?» В виде гипотезы это можно записать следующим образом:

$$H_0 : \theta_1 = 3.$$

Для остальных параметров гипотезы о значении записываются аналогичным образом.

Ко второму типу гипотез будем относить гипотезы о незначимости параметров, которые позволяют оценить факт влияния отдельных факторов на зависимую переменную. По своей сути гипотеза о незначимости является частным случаем гипотезы о значении, в которой идет проверка на равенство нулевому значению.

Например, для параметра θ_1 в общем случае эта гипотеза записывается в виде

$$H_0 : \theta_1 = 0. \qquad (33)$$

Если эта гипотеза не отвергается, то можно считать, что независимая переменная не оказывает существенного (значимого) влияния на отклик. Другими словами, при справедливости гипотезы (33) независимая переменная исключается из модели. Графически это означает, что линия регрессии лишь случайно отклоняется от линии, параллельной оси *x*. В этом случае можно смело записать, что $E(y) = \theta_0$.

Проверка гипотез о значении и о незначимости параметров проводится по одной и той же схеме с помощью критерия Стьюдента [14, 20]. Поэтому рассмотрим процедуру проверки только для гипотезы (32).

Пусть заданы гипотезы H_0 и ее альтернатива H_1:

$$H_0 : \theta_1 = \theta_{10},$$

$$H_1 : \theta_1 \neq \theta_{10}.$$

Предположим, что верна гипотеза H_0 и, исходя из этого, вычислим значение *t*-статистики, подставив в (30) вместо истинного значения θ_1 гипотетическое значение θ_{10}:

$$t = \frac{\left(\hat{\theta}_1 - \theta_{10}\right)}{S_{\hat{\theta}_1}}.$$

Как уже отмечалось выше, найденное значение *t* является реализацией случайной величины, распределенной по закону Стьюдента с $(N-2)$ степенями свободы. Как и раньше, гипотеза H_0 отвергается, если $|t| > t_{êð}\left((1-\alpha), N-2\right)$, где $t_{êð}\left((1-\alpha), N-2\right)$ – критическое значение, найденное по таблицам квантилей распределения Стьюдента. Если $|t| \leq t_{êð}\left((1-\alpha), N-2\right)$, то гипотеза H_0 не отвергается.

Нетрудно заметить, что при проверке гипотезы о незначимости параметра θ_1, малые абсолютные значения *t*-статистики соответствуют отсутствию достоверной статистической связи между регрессором и откликом.

Проверка гипотез для параметра θ_0 проводится аналогично, с той разницей, что вместо статистики (30) используется статистика (31).

Часто в задачу исследования может входить не только проверка соответствия найденных оценок параметров некоторым заданным значениям, но и определения с заданной вероятностью $(1-\alpha)$ диапазона, в котором эти параметры могут изменяться. Для этого достаточно разрешить относительно θ_1 неравенство

$$P\left\{\frac{\left|\hat{\theta}_1-\theta_1\right|}{S_{\hat{\theta}_1}}<t_{\hat{e}\eth}\right\}=(1-\alpha).$$

В результате получим

$$P\left\{\hat{\theta}_1-t_{\hat{e}\eth}S_{\hat{\theta}_1}<\theta_1<\hat{\theta}_1+t_{\hat{e}\eth}S_{\hat{\theta}_1}\right\}=(1-\alpha),$$

т.е. с вероятностью $(1-\alpha)$ истинное значение параметра θ_1 находится в интервале $\left[\hat{\theta}_1-t_{\hat{e}\eth}S_{\hat{\theta}_1}\ ;\ \hat{\theta}_1+t_{\hat{e}\eth}S_{\hat{\theta}_1}\right]$.

При построении доверительного интервала получаем намного больше информации, чем при проверке гипотез о значении. Любая гипотеза на равенство значению, лежащему внутри интервала, не будет отвергаться, и наоборот – гипотеза о значении, лежащем вне интервала, будет всегда отвергаться.

В продолжение нашего примера допустим, что для параметра θ_1 был получен 95 %-й доверительный интервал [1.9; 3.7]. Этот факт позволяет говорить о том, что при увеличении рекламных расходов на одну тысячу рублей ожидаемый прирост оборота с вероятностью 0.95 может принимать любые значения в диапазоне от 1.9 млн. руб. до 3.7 млн. руб.

3.9. Разложение суммы квадратов и проверка значимости уравнения регрессии

После того как найдено уравнение парной регрессии, необходимо провести оценку его адекватности исследуемому процессу. Проверить на адекватность (на значимость) уравнение регрессии – это значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, наблюдаемым процессам или явлениям. Оценка значимости уравнения регрессии может проводиться с помощью *F*-критерия Фишера [14, 20]. Непосредственному расчету *F*-критерия предшествует анализ дисперсии. При этом основная роль отводится разложению общей суммы

квадратов отклонений переменной y от среднего значения $\overline{y}$ на две части – «объясненную» и «необъясненную (остаточную)».

Общая сумма квадратов

$$\sum_{i=1}^{N}(y_i-\overline{y})^2$$

характеризует величину разброса значений зависимой переменной. Этот разброс может быть вызван, с одной стороны, изменениями входных факторов, а с другой – случайными воздействиями или неучтенными в модели факторами. Если неучтенных факторов нет, и случайные воздействия отсутствуют, то все изменения отклика должны объясняться моделью.

Для того чтобы выделить объясненную и необъясненную части, подставим в общую сумму квадратов очевидное тождество

$$y_i-\overline{y}=(y_i-\hat{y}_i)+(\hat{y}_i-\overline{y}).$$

В результате получим

$$\sum_{i=1}^{N}(y_i-\overline{y})^2=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\sum_{i=1}^{N}(\hat{y}_i-\overline{y})^2+2\sum_{i=1}^{N}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y}).$$

В силу специфических свойств вектора остатков [14] последнее слагаемое обращается в ноль, т.е.

$$2\sum_{i=1}^{N}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})=0.$$

Поэтому справедливо следующее равенство:

$$\sum_{i=1}^{N}(y_i-\overline{y})^2=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\sum_{i=1}^{N}(\hat{y}_i-\overline{y})^2. \qquad (34)$$

Введем обозначения:

$TSS=\sum_{i=1}^{N}(y_i-\overline{y})^2$ – общая сумма квадратов (total sum of squares);

$ESS=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$ – остаточная сумма квадратов, называемая иногда суммой квадратов остатков (error sum of squares);

$RSS = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$ – сумма квадратов, обусловленная регрессией или объясненная сумма квадратов (regression sum of squares).[1]

В результате получим

$$TSS = ESS + RSS . \qquad (35)$$

Любая сумма квадратов отклонений связана с так называемым «числом степеней свободы». Число степеней свободы зависит от количества наблюдений N и количества определяемых по ним величин. Применительно к общей сумме квадратов число степеней свободы показывает, сколько независимых отклонений из N возможных

$$(y_1 - \bar{y}), (y_2 - \bar{y}), \ldots, (y_N - \bar{y})$$

требуется для вычисления данной суммы квадратов. Наличие известного среднего значения $\bar{y}$ дает нам возможность произвести вычисление TSS, используя только $(N-1)$ независимых отклонений. Например, имеем пять наблюдений

$$y_1 = 1,\ y_2 = 2,\ y_3 = 3,\ y_4 = 4,\ y_5 = 5.$$

Среднее значение $\bar{y} = 3$. Отклонения от среднего будут равны соответственно

$$-2; -1; 0; 1; 2 .$$

Так как $\sum_{i=1}^{5} (y_i - \bar{y}) = 0$, то свободно изменяться могут только 4 отклонения, а оставшееся всегда может быть выражено через них. Таким образом, число степеней свободы общей суммы квадратов равняется $(N-1)$.

Аналогичным образом можно показать, что число степеней свободы остаточной суммы квадратов равно $(N-2)$, а число степеней свободы объясненной суммы квадратов равно 1.

[1] К сожалению, эти сокращения не являются общепринятыми. В некоторых литературных источниках первое слагаемое в правой части (34) обозначается через RSS (residual sum of squares), а второе – через ESS (explained sum of squares).

Зная все три суммы квадратов, можно делать некоторые предварительные выводы о качестве регрессионного уравнения. Например, если остаточная сумма квадратов намного превышает объясненную, то это говорит о том, что остатки регрессии очень велики. Это может свидетельствовать либо о вычислительных ошибках, либо о том, что построенная регрессионная модель плохо описывает данные. Если же остаточная сумма квадратов много меньше объясненной, то это говорит о малых остатках и хорошем качестве модели.

Однако такой анализ является поверхностным и не лишен субъективизма. Для объективной оценки качества регрессионного уравнения необходимо использовать некоторые специальные критерии.

Одной из наиболее эффективных оценок качества уравнения регрессии, характеристикой прогностической силы анализируемой регрессионной модели является коэффициент детерминации

$$R^2 = 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS}. \qquad (36)$$

Коэффициент детерминации показывает долю объясненной дисперсии в общей дисперсии зависимой переменной и может принимать значения в диапазоне $0 \le R^2 \le 1$.

Если $R^2 = 0$, то это означает, что общая сумма квадратов равна остаточной, т.е. уравнение регрессии совершенно не объясняет изменения зависимой переменной.

Если $R^2 = 1$, то общая сумма квадратов равна объясненной и все наблюдаемые точки лежат точно на линии регрессии, а все остатки нулевые.

Чем ближе значение коэффициента детерминации к единице, тем более точно уравнение регрессии описывает данные. Однако зная только конкретное значение коэффициента детерминации, нельзя делать никаких однозначных выводов о пригодности уравнения к практическому использованию. В зависимости от ситуации значимыми могут признаваться уравнения с коэффициентом детерминации, равным 0.87, и уравнения,

коэффициент детерминации которых равен 0.33. А для того чтобы четко определять пригодность уравнения регрессии в каждом конкретном случае, необходима формальная проверка на значимость.

Проверку на значимость регрессионного уравнения можно проводить с помощью статистической гипотезы, например, относительно величины коэффициента детерминации:

$$H_0 : R^2 = 0, \tag{37}$$

$$H_1 : R^2 \neq 0.$$

Проверка этой гипотезы может проводиться с помощью F-критерия. Этот критерий опирается на тот факт, что отношение объясненной дисперсии зависимой переменной к остаточной дисперсии подчиняется распределению Фишера. Поскольку на практике истинные значения этих дисперсий остаются неизвестными, то приходится их оценивать с помощью соответствующих сумм квадратов отклонений.

Из разложения сумм квадратов (34) может быть получено разложение для дисперсий [21]:

$$D(y) = D(\hat{y}) + D(y - \hat{y}),$$

где $D(y)$ – полная дисперсия зависимой переменной y, $D(\hat{y})$ – объясненная дисперсия или дисперсия расчетных значений $\hat{y}$, $D(y - \hat{y})$ – остаточная или необъясненная дисперсия.

В качестве оценок полной, остаточной и объясненной дисперсий используются отношения соответствующих сумм квадратов к своим числам степеней свободы:

$$\hat{D}(y) = S_y^2 = \frac{TSS}{(N-1)} = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{(N-1)}, \tag{38}$$

$$\hat{D}(y - \hat{y}) = S^2 = \frac{ESS}{(N-2)} = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{(N-2)}, \tag{39}$$

$$\hat{D}(\hat{y}) = S_{\hat{y}}^2 = \frac{RSS}{1} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2}{1}. \qquad (40)$$

Если сравнить соотношения (18) и (39), то можно заметить, что дисперсия случайной ошибки и остаточная дисперсия оцениваются одинаковым образом.

Из математической статистики известно [20, 21], что величина

$$F = \frac{S_{\hat{y}}^2}{S^2} \qquad (41)$$

подчиняется распределению Фишера с 1 и $(N-2)$ степенями свободы. При проверке гипотезы (37) вычисленное по формуле (41) значение F-статистики сравнивают с критическим значением $F_{\hat{e}\eth}\left((1-\alpha),1,(N-2)\right)$. Критическое значение F-статистики – это максимальная величина отношения дисперсий, которая для заданного уровня доверительной вероятности $(1-\alpha)$ может иметь место при случайном отклонении от нулевой гипотезы. Это значение определяется по специальным статистическим таблицам [4].

Если оказывается, что $F > F_{\hat{e}\eth}\left((1-\alpha),1,(N-2)\right)$, то гипотеза (37) отвергается и уравнение регрессии признается значимым, т.е. пригодным для практического использования.

Если оказывается, что $F \le F_{\hat{e}\eth}\left((1-\alpha),1,(N-2)\right)$, то гипотеза (37) не отвергается и уравнение регрессии не признается значимым.

В случае парной линейной регрессии коэффициент детерминации равен квадрату парного коэффициента корреляции $R^2 = r_{xy}^2$. Это позволяет установить связь между статистикой Фишера и R^2 [20]:

$$F = (N-2)\frac{R^2}{1-R^2}. \qquad (42)$$

3.10. Таблица дисперсионного анализа

Результаты проведенного регрессионного анализа в компактной форме можно представить в виде так называемой «Таблицы дисперсионного анализа». Исторически эта таблица была разработана для представления результатов дисперсионного анализа наблюдений [8, 19]. Однако такая форма представления результатов оказалась очень удачной и стала использоваться также в регрессионном и ковариационном анализах. В настоящее время представление результатов в виде таблицы дисперсионного анализа является общепринятым для большинства статистических и эконометрических пакетов прикладных программ (*Statistica, SAS, SPSS, Econometric Views* и др.) [14, 18]. Стандартная таблица дисперсионного анализа выглядит следующим образом (табл.1).

Таблица 1

Таблица дисперсионного анализа

Источник дисперсии	Число степеней свободы	Сумма квадратов	Средняя сумма квадратов	*F*-статистика	*F*-критическое	Значимость
Регрессия	1	RSS	$\frac{RSS}{1}$	F	$F_{\hat{e}\eth}$	Да / Нет
Ошибка	$N-2$	ESS	$\frac{ESS}{N-2}$	–	–	–
Итог	$N-1$	TSS	$\frac{TSS}{N-1}$	–	–	–

В эту таблицу включены все основные результаты расчетов, проводимых при проверке значимости регрессионной модели, а именно: TSS, RSS и ESS – элементы из разложения (35); F – статистика Фишера, вычисляемая по формуле (42); $F_{\hat{e}\eth}$ – критическое значение статистики Фишера для соответствующего уровня доверительной вероятности. Кроме этого, в таблице дисперсионного анализа часто присутствует ответ на вопрос:

«Значимо ли уравнение регрессии?» (иногда последний столбец таблицы дисперсионного анализа может опускаться).

3.11. Прогнозирование в регрессионных моделях

Если после проверки уравнения регрессии на значимость было установлено, что оно является пригодным для практического использования, то оно может применяться для построения различных прогнозов. Следует отметить, что обычно термин «прогнозирование» используется в тех случаях, когда требуется предсказать состояние наблюдаемого объекта в будущем. Для регрессионных моделей он имеет более широкое значение. Здесь может возникнуть потребность оценить значение зависимой переменной для некоторого заданного, часто отсутствующего в исходных данных, значения независимой переменной. Именно в этом смысле – как построение оценки зависимой переменной – и следует понимать прогнозирование в эконометрике.

Обычно различают точечное и интервальное прогнозирование. В первом случае требуется вычислить конкретное число – ожидаемое значение зависимой переменной, во втором случае требуется указать интервал, в котором с известной доверительной вероятностью находится истинное значение зависимой переменной. Выделяют также безусловное и условное прогнозирование в зависимости от того, задано ли интересующее нас значение объясняющей переменной точно или с погрешностью.

Допустим, нам задано известное значение независимой переменной x_{i} и при таком значении требуется спрогнозировать значение зависимой переменной. Точечный прогноз может быть очень просто вычислен по уравнению регрессии путем подстановки x_{i} вместо независимой переменной:

$$\hat{y}_{i} = \hat{\theta}_0 + \hat{\theta}_1 x_{i} \ .$$

В силу того что вероятность попадания в любую точку равна нулю, фактическое значение зависимой переменной y_i при заданном значении x_i никогда в точности не совпадет с точечным прогнозом $\hat{y}_i$. Поэтому на практике точечный прогноз всегда следует дополнять расчетом стандартной ошибки прогноза $\sigma_{\hat{y}_i}$ и интервальным прогнозом.

Для заданного уровня доверительной вероятности $(1-\alpha)$ при известной стандартной ошибке $\sigma_{\hat{y}_i}$ истинное значение зависимой переменной будет находиться в интервале

$$\hat{y}_i - t_{êð}\sigma_{\hat{y}_i} \le y_i \le \hat{y}_i + t_{êð}\sigma_{\hat{y}_i}, \qquad (43)$$

где $t_{êð}$ – критическое значение, найденное по таблицам квантилей распределения Стьюдента.

Для оценки стандартной ошибки прогноза подставим оценку (14) в уравнение регрессии и получим

$$\hat{y}_i = \overline{y} - \hat{\theta}_1\overline{x} + \hat{\theta}_1 x_i$$

или

$$\hat{y}_i = \overline{y} + \hat{\theta}_1(x_i - \overline{x}).$$

Нетрудно заметить, что дисперсия прогноза может быть выражена следующим образом

$$\sigma^2_{\hat{y}_i} = \sigma^2_{\overline{y}} + \sigma^2_{\hat{\theta}_1}\left(x_i - \overline{x}\right)^2. \qquad (44)$$

Из математической статистики известно, что

$$\sigma^2_{\overline{y}} = \frac{\sigma^2}{N}.$$

Заменяя неизвестное значение σ^2 на ее оценку S^2, получим

$$\hat{\sigma}^2_{\overline{y}} = \frac{S^2}{N}. \qquad (45)$$

После подстановки (25) и (45) в (44) и извлечения квадратного корня получим оценку стандартной ошибки прогноза отклика в точке x_i

$$\hat{\sigma}_{\hat{y}_{\ddot{I}}} = \sqrt{\frac{S^2}{N} + \frac{S^2}{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}\left(x_{\ddot{I}} - \overline{x}\right)^2}\,. \qquad (46)$$

Из (46) видно, что значение стандартной ошибки прогноза явно зависит от точки $x_{\ddot{I}}$, в которой строится прогноз, а если говорить более точно, то от разности $\left(x_{\ddot{I}} - \overline{x}\right)$. Чем больше эта разность, тем больше стандартная ошибка прогноза и ниже точность прогнозирования. Наименьшее значение стандартной ошибки соответствует точке $x_{\ddot{I}} = \overline{x}$, именно в окрестности этой точки и достигается наивысшая точность прогнозирования. Что касается доверительного интервала для $y_{\ddot{I}}$, то, как видно из (43), его ширина напрямую зависит от стандартной ошибки прогноза. На рис. 8 показано, как будут изменяться границы доверительного интервала в зависимости от прогнозной точки при фиксированном уровне доверительной вероятности (именно с этой вероятностью истинное значение отклика будет находиться в указанных границах).

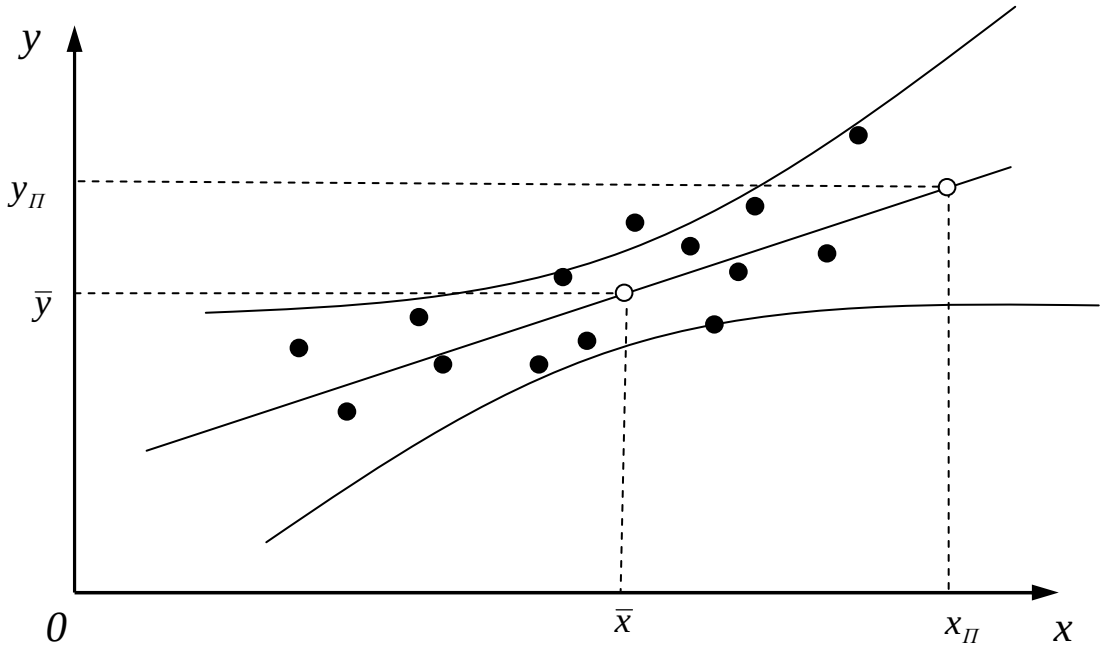


Рис. 8. Доверительный интервал уравнения регрессии

Изменение доверительной вероятности также существенно влияет на интервальный прогноз. Так, при увеличении доверительной вероятности ширина доверительного интервала должна увеличиваться, а при уменьшении – уменьшаться.

От только что рассмотренной ситуации принципиально отличается случай, когда требуется сделать прогноз при условии, что нет твердой уверенности в справедливости структуры модели. Такое возможно, если в регрессионной модели под случайной ошибкой ε подразумевается не только ошибка измерения, но и воздействие некоторых неучтенных факторов.

Отличие состоит в том, что оценка стандартной ошибки прогноза должна вычисляться следующим образом:

$$\hat{\sigma}_{\hat{y}_i} = \sqrt{1+\frac{S^2}{N}+\frac{S^2}{\sum_{i=1}^{N}\left(x_i-\overline{x}\right)^2}\left(x_{\ddot{I}}-\overline{x}\right)^2}\,. \qquad (47)$$

Нетрудно заметить, что при расчете доверительного интервала с использованием формулы (47) вместо (46) его ширина увеличивается.

Контрольные вопросы

1. Опишите модель парной регрессии.
2. Для чего в модель парной регрессии включается случайная составляющая?
3. Перечислите известные вам методы поиска неизвестных параметров модели парной регрессии.
4. В чем заключаются основные предположения метода наименьших квадратов?
5. Для чего выдвигается предположение о нормальном распределении случайных ошибок?
6. В чем разница между θ_0, θ_1 и $\hat{\theta}_0$, $\hat{\theta}_1$?
7. Что показывают величины θ_0 и θ_1?

8. Сформулируйте теорему Гаусса-Маркова.
9. Каким образом можно оценить дисперсию случайной ошибки модели?
10. В чем разница между остатками и ошибками модели?
11. По какому закону распределены оценки параметров модели парной регрессии?
12. Что показывают величины $S^2_{\hat{\theta}_0}$ и $S^2_{\hat{\theta}_1}$?
13. Что такое «гипотеза о значимости параметра»? Каким образом она проверяется?
14. Чем отличается проверка значимости параметра θ_0 и параметра θ_1?
15. Как можно определить доверительный интервал для параметра регрессионного уравнения?
16. Какова зависимость ширины доверительного интервала от дисперсии случайной ошибки?
17. Что такое *TSS*, *RSS*, *ESS*? Как связаны между собой эти величины?
18. Как определяется и что показывает коэффициент детерминации?
19. Назовите основные свойства коэффициента детерминации.
20. Как проводится проверка значимости регрессионного уравнения?
21. Как строится таблица дисперсионного анализа?
22. Чем отличается точечное и интервальное прогнозирование?
23. В чем разница между условным и безусловным прогнозированием?
24. От чего зависит точность прогнозирования по уравнению регрессии?