ОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ

Н. В. ПЕРЦЕВ

ЛЕКЦИИ по эконометрике

Часть II. Вычислительные аспекты

Омск 2003

Аннотация

Часть II лекций посвящена вычислительным аспектам задач эконометрики. Здесь приводятся основные формулы, таблицы и т.д., используемые в регрессионном анализе и обработке временных рядов. Материал по системам эконометрических уравнений в компактной и доступной форме изложен в следующих учебниках:

1) Елисеева И.И. и др. Эконометрика. М.: Финансы и статистика, 2001 (глава 4);

2) Елисеева И.И. и др. Практикум по эконометрике. М.: Финансы и статистика, 2001 (раздел 3);

3) Кремер Н.Ш., Путко Б.А. Эконометрика. М.: Юнити, 2002 (глава 9).

Эти же учебники рекомендуются для изучения и других материалов курса.

Для студентов заочной, заочно-ускоренной и вечерне-ускоренной форм обучения ОмГУ.

ЧАСТЬ 1. РЕГРЕССИОННЫЙ АНАЛИЗ

1.1. ПОСТАНОВКА ЗАДАЧИ

Пусть изучается некоторый объект V, который характеризуется величинами

$$x,\, y, \ldots, z,\, w,$$

отражающими его свойства. Нас будут интересовать зависимости между этими величинами и те формулы, которые их задают. Такие зависимости можно представить в виде двух основных форм.

Первая форма зависимости – это функциональная зависимость, когда одна из величин явно (неявно, параметрически и т.д.) выражается через остальные. Здесь, как правило, имеется вполне конкретная формула, связывающая между собой рассматриваемые величины. Часто бывает так, что вид формулы известен с точностью до входящих в нее коэффициентов, и тогда эти коэффициенты требуется найти по результатам наблюдений (измерений). В более сложном варианте конкретный вид формулы может вызывать определенные трудности, и тогда следует рассматривать набор формул и выбирать какую-то одну из них.

Вторая форма зависимости – это стохастическая зависимость, которая, как правило, не описывается конкретной формулой. Здесь зависимость между величинами проявляется в том, что изменение одной из величин влияет на возможные значения оставшихся величин. Если же зависимость между величинами отсутствует, то изменение одной из них никаким образом не отражается на возможных значениях остальных. Более точно такая зависимость проявляется в изменении закона распределения одной величины под влиянием конкретных значений других величин. Если же зависимость между величинами отсутствует, то изменение одной из них не отражается на законах распределения остальных.

Существуют также и другие варианты зависимостей, сочетающие в себе функциональную и стохастическую зависимости. Кроме того, возможен вариант зависимостей, когда значения одной, двух или трех величин достаточно хорошо описываются одной так называемой объясняющей переменной.

В простейшем случае зависимость между двумя величинами y и x строится в виде

$$y = f(x) + \varepsilon, \tag{1}$$

где $f(x)$ – некоторая функция. Величина ε учитывает погрешность приближенной связи $y \approx f(x)$ и включает в себя все неучтенные или неизвестные факторы. Очевидно, что выбор функции $f(x)$ представляет собой довольно трудную задачу, для решения которой необходимо уметь оценивать свойства погрешности ε. Обычно функцию $f(x)$ выбирают так, чтобы дисперсия погрешности $D(\varepsilon) = D(y - f(x))$ была бы как можно меньше, то есть $D(y - f(x)) \to \min$. Как известно, решение данной задачи дает функция

$$f(x) = M(y/x), \tag{2}$$

где выражение $M(y/x)$ означает условное математическое ожидание величины y при фиксированном значении величины x. Функция $f(x)$ называется регрессией y на x. На практике нахождение $f(x)$ по формуле (2) довольно затруднительно или вообще

невозможно, поскольку необходимо иметь информацию о совместном распределении пары $(x,\, y)$ в соответствующей генеральной совокупности. Поэтому, как правило, $f(x)$ подбирают среди некоторого класса достаточно простых функций и затем по выборочным данным определяют ее коэффициенты. В конкретных задачах часто используют линейные, квадратичные, показательные, тригонометрические и др. функции. Например, формула связи $y = a_0 + a_1 x + a_2 x^2 + \varepsilon$, содержащая параметры a_0, a_1, a_2, отражает квадратичную зависимость y от x.

В более общем случае может изучаться зависимость величины y от многомерной величины $u = (x_1,\, x_2, \ldots,\, x_k)$, и эта зависимость строится в виде

$$y = f(b,\, u) + \varepsilon, \tag{3}$$

где $b = (b_1,\, b_2, \ldots, b_\ell)$ – вектор неизвестных параметров. Функция $f(b,\, u)$ называется множественной регрессией y на u. В уравнении (3) величина y называется зависимой переменной, а $x_1,\, x_2, \ldots,\, x_k$ – объясняющими переменными.

Выбор функции $f(b,\, u)$ и оценка ее параметров опирается на набор данных, представленных в следующей таблице.

Таблица 1

Набор данных для регрессионного анализа

N	y	x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{21}	$\cdots$	x_{k1}
2	y_2	x_{12}	x_{22}	$\cdots$	x_{k2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
i	y_i	x_{1i}	x_{2i}	$\cdots$	x_{ki}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	y_n	x_{1n}	x_{2n}	$\cdots$	x_{kn}

Предполагается, что значения величин $(y_i,\, x_{1i},\, x_{2i}, \ldots, x_{ki})$ получены одновременно при конкретном наблюдении (измерении), $1 \le i \le n$. Кроме того, считается, что каждый столбец таблицы 1 задает выборку значений соответствующей величины. Параметр n означает объем выборки.

Для оценки вектора параметров b зависимости (3) по данным табл. 1 используется метод наименьших квадратов (МНК). Сущность этого метода заключается в следующем. Составляется функция $L(b)$, которая описывает меру рассеивания данных по переменной y относительно приближенных значений по переменной u

$$L(b) = \sum_{i=1}^{n} \big(y_i - f(b,\, u_{[i]})\big)^2, \tag{4}$$

где $u_{[i]} = (x_{1i},\, x_{2i}, \ldots, x_{ki})$, $1 \le i \le n$. В некоторых случаях выражение для $L(b)$ имеет более сложный вид. Искомый вектор оценок параметров $\bar{b}$ находится как решение задачи на экстремум

$$L(b) \to \min. \tag{5}$$

Решение задачи (5) может быть найдено аналитически либо численно с помощью специализированных пакетов программ.

Предварительный вид функции $f(x)$ или $f(b,\, u)$ может быть установлен, исходя из графического анализа данных. Наиболее удобно изучать парную зависимость, т.е.

зависимость y от какой-то одной из объясняющих переменных, например, от x_1. Здесь используют графическое представление пар точек (x_{1i}, y_i) на плоскости, $1 \leq i \leq n$. При нанесении этих пар на плоскость получается некоторое «облако» точек, форма которого может говорить о наличии или отсутствии зависимостей. Если «облако» точек имеет вполне конкретную, выраженную форму, то можно вполне уверенно говорить о наличии зависимости между переменными x_1 и y. В противном случае зависимости может и не быть (см. рис. 1, 2).

Графическое представление данных позволяет сделать определенный качественный вывод о возможной зависимости между рассматриваемыми переменными x_1 и y. Вместе с тем установление факта их зависимости или независимости требует привлечения количественных методов, которые излагаются в следующих разделах.

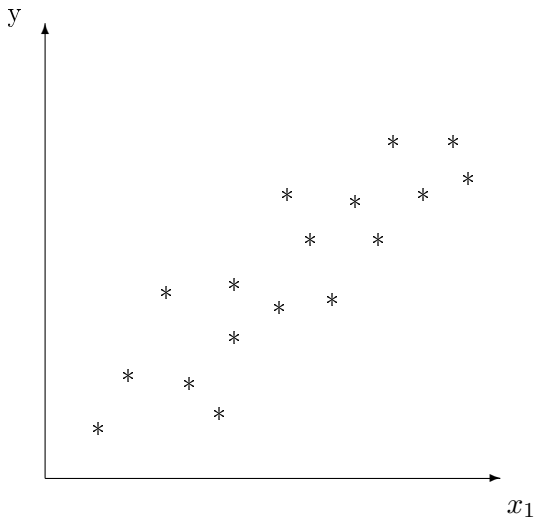

Рис. 1. Облако точек. Имеется зависимость

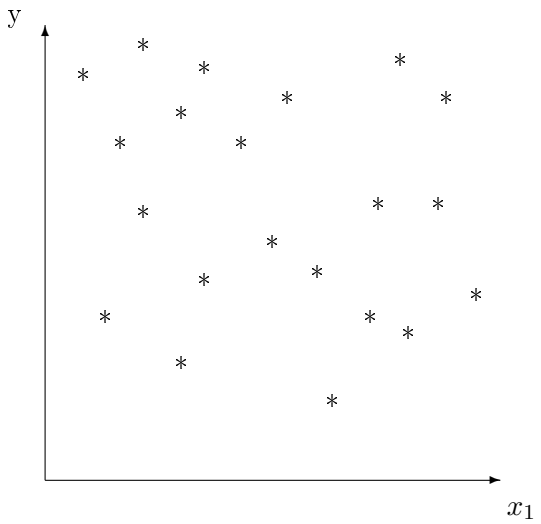

Рис. 2. Облако точек. Нет зависимости

Графическое представление данных позволяет сделать определенный качественный вывод о возможной зависимости между рассматриваемыми величинами x_1 и y. Вместе с тем установление факта их зависимости или независимости требует привлечения количественных методов, которые излагаются в следующих разделах.

1.2. ЛИНЕЙНАЯ РЕГРЕССИОННАЯ ЗАВИСИМОСТЬ

1.2.1. Основные предположения.

Примем, что связь между зависимой и объясняющими переменными имеет следующий вид:

$$y = b_0 + b\,u + \varepsilon = b_0 + b_1\,x_1 + \ldots + b_i\,x_i + \ldots + b_k\,x_k + \varepsilon. \tag{6}$$

Здесь b_0, $b_1, \ldots, b_i, \ldots, b_k$ – параметры линейной зависимости (линейной регрессии), величина ε – случайная ошибка наблюдений (измерений). Все эти параметры являются, вообще говоря, неизвестными и подлежат определению по выборочным данным. Если для некоторого $1 \le i \le k$ окажется, что $b_i \ne 0$, то формула (6) будет говорить о существовании зависимости между переменными x_i и y. При $b_i = 0$ нельзя говорить о зависимости между y и x_i, выраженной в линейной форме. Если же для всех $1 \le i \le k$ окажется, что $b_i = 0$, то будем говорить, что линейная зависимость между y и объясняющими переменными отсутствует. Зависимость y от x_i может иметь место, но в другой, более сложной форме.

Нахождение оценок параметров и обоснование зависимости (6) опирается на следующие предположения относительно случайной составляющей ε:

Н1) математическое ожидание и дисперсия величины ε таковы, что

$$M(\varepsilon) = 0, \quad D(\varepsilon) = \sigma^2 = const > 0. \tag{7}$$

Н2) любые пары значений ε_i, ε_j величины ε являются некоррелированными, т.е. при $i \ne j$ имеет место равенство $M(\varepsilon_i\,\varepsilon_j) = 0$, в частности, это верно и для пар $\varepsilon_i = y_i - b_0 - b\,u_i$ и $\varepsilon_j = y_j - b_0 - b\,u_j$, где y_i, y_j, u_i, u_j взяты из таблицы 1;

Н3) величина ε имеет нормальное распределение с параметрами, заданными формулой (7).

Выполнение предположений Н1) и Н2) позволяет применить метод наименьших квадратов (МНК) и получить формулы для оценок параметров зависимости (6). Предположения Н1) и Н2) называют основными предположениями МНК.

Выполнение предположения Н3) дает возможность обосновать наличие или отсутствие зависимости между переменной y и переменными $x_1, x_2, \ldots, x_k$, заданной формулой (6).

1.2.2. Формулы для одной объясняющей переменной.

Изучаем зависимость вида

$$y = b_0 + b_1\,x_1 + \varepsilon. \tag{8}$$

Полагаем, что у нас имеются выборки значений переменной y и переменной x_1, представленные в соответствующих столбцах таблицы 1. Обозначим

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{9\,a}$$

$$Q_{x_1x_1} = \sum_{i=1}^{n} x_{1i}^2 - n\,(\bar{x}_1)^2, \quad Q_{x_1y} = \sum_{i=1}^{n} (x_{1i}\,y_i) - n\,\bar{x}_1\,\bar{y}, \tag{9\,b}$$

Оценки параметров $b_0,\ b_1$, входящих в уравнение регрессии (8), равны

$$\bar{b}_1 = \frac{Q_{x_1 y}}{Q_{x_1 x_1}},\quad \bar{b}_0 = \bar{y} - \bar{b}_1\,\bar{x}_1. \tag{10}$$

Используя полученные оценки, сформируем выборку остатков

$$e_1,\ e_2, \ldots, e_i, \ldots, e_n, \tag{11}$$

которые задают отклонения наблюдаемых значений y_i от их предсказанных значений по формуле (8), т.е.

$$e_i = y_i - \bar{b}_0 - \bar{b}_1\,x_{1i},\ 1 \le i \le n. \tag{12}$$

По выборке (11) вычислим остаточную сумму квадратов

$$Q_{ee} = \sum_{i=1}^{n} e_i^2, \tag{13}$$

которая будет использована в последующих расчетах.

Оценки (10) дают приближенные значения параметров $b_0,\ b_1$, т.е. $b_0 \approx \bar{b}_0,\ b_1 \approx \bar{b}_1$. На основании этих приближенных равенств нельзя получить уверенного заключения о точном значении искомого параметра b_1 и проверить неравенство $b_1 \neq 0$. Именно это неравенство и будет говорить о наличии или отсутствии линейной зависимости между y и x_1. Точность оценивания параметра b_1 зависит от объема выборки n и характеризуется стандартной ошибкой оценки $\bar{b}_1$. Стандартная ошибка σ_1 оценки $\bar{b}_1$ находится по формуле

$$\sigma_1 = \sqrt{\frac{Q_{ee}}{(n-2)\ Q_{x_1 x_1}}}. \tag{14}$$

Зафиксируем значение объясняющей переменной $x_1 = x_1^*$. Тогда выражение

$$\hat{y} = \bar{b}_0 + \bar{b}_1\,x_1^* \tag{15}$$

будет задавать приближенное значение зависимой переменной y, т.е. $y \approx \hat{y}$. Точность оценивания y характеризуется стандартной ошибкой полученной оценки и обозначается σ_y. Стандартная ошибка σ_y задается формулой

$$\sigma_y = \sqrt{\frac{Q_{ee}}{n-2}\left(1 + \frac{1}{n} + \frac{(x_1^* - \bar{x}_1)^2}{Q_{x_1 x_1}}\right)}. \tag{16}$$

Перейдем к обоснованию наличия или отсутствия линейной зависимости (8) между переменными y и x_1. Выдвигаем гипотезу H_0 об отсутствии такой зависимости. Это равносильно тому, что $b_1 = 0$. Зафиксируем уровень значимости $\alpha \cdot 100\%$. Число α задает вероятность ошибки первого рода. Ошибка первого рода означает, что представленные данные и результаты их обработки не согласуются с принятой гипотезой H_0, и мы ее отвергаем. Проверка гипотезы H_0 опирается на два способа.

При первом способе вычисляем величину

$$F = \frac{\bar{b}_1^2\ Q_{x_1 x_1}\,(n-2)}{Q_{ee}}. \tag{17}$$

Эту величину будем сравнивать с критическим значением F_α распределения Фишера с $f_1 = 1$ и $f_2 = n - 2$ степенями свободы на уровне значимости α (таблица П1). Если окажется, что выполнено неравенство $F \le F_\alpha$, то гипотезу H_0 принимаем, и линейную зависимость (8) называем не значимой. Если же $F > F_\alpha$, то гипотезу H_0 отклоняем и считаем, что между переменными x_1 и y имеется линейная зависимость, и эту зависимость будем называть значимой.

При втором способе строим границы доверительного интервала для параметра b_1 по формуле

$$b_1 \in I_1 = (\bar{b}_1 - t_\alpha \sigma_1,\, \bar{b}_1 + t_\alpha \sigma_1), \tag{18}$$

где t_α – критическое значение распределения Стьюдента с $n - 2$ степенями свободы на уровне значимости α (таблица П2). Интервал I_1 накрывает параметр b_1 с вероятностью $p = 1 - \alpha$. Если окажется, что доверительный интервал I_1 содержит в себе число ноль, то считается, что $b_1 = 0$ и, как следствие, гипотеза H_0 принимается. Если же доверительный интервал I_1 не содержит в себе число ноль, то полагается, что $b_1 \neq 0$ и поэтому гипотеза H_0 отклоняется.

Предположим, что между переменными x_1 и y установлена значимая линейная зависимость. Тогда по заданному $x_1 = x_1^*$ можно указать границы для ожидаемого значения y с учетом влияния случайной составляющей ε. Эти границы устанавливаются в форме доверительного интервала

$$y \in I_y = (\hat{y} - t_\alpha \sigma_y,\, \hat{y} + t_\alpha \sigma_y), \tag{19}$$

который содержит значение переменной y с вероятностью $p = 1 - \alpha$. Критическое значение t_α описано выше, величины $\hat{y}$, σ_y заданы формулами (15) и (16). Границы доверительного интервала $c_1 = \hat{y} - t_\alpha \sigma_y$, $c_2 = \hat{y} + t_\alpha \sigma_y$, как функции от величины x_1^*, приведены на рис. 3. Из него видно, что точность предсказания возможных значений y убывает по мере удаления точки x_1^* от точки $\bar{x}_1$.

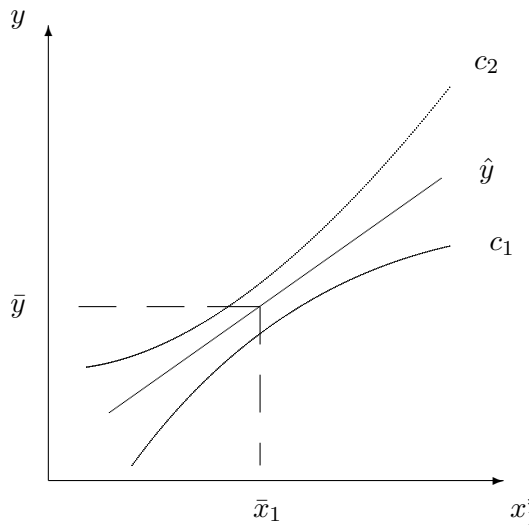

Рис. 3. Границы доверительного интервала для y

1.2.3. Формулы для двух объясняющих переменных.

В этом разделе будем изучать зависимость y от двух объясняющих переменных x_1 и x_2, заданную в линейной форме

$$y = b_0 + b_1\, x_1 + b_2\, x_2 + \varepsilon. \tag{20}$$

Предполагается, что между x_1 и x_2 нет линейной связи, то есть $x_2 \neq a_1\, x_1 + a_2,\ a_1 \neq 0$. В противном случае соотношение (20) с точностью до обозначений совпадает с (8). Полагаем, что у нас имеются выборки значений переменной y и переменных x_1, x_2, представленные в соответствующих столбцах таблицы 1. Обозначим

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i}, \quad \bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{2i}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{21 a}$$

$$\Delta = Q_{x_1x_1}\cdot Q_{x_2x_2} - Q^2_{x_1x_2}, \quad Q_{x_1x_1} = \sum_{i=1}^{n} x^2_{1i} - n\,(\bar{x}_1)^2, \tag{21 b}$$

$$Q_{x_2x_2} = \sum_{i=1}^{n} x^2_{2i} - n\,(\bar{x}_2)^2, \quad Q_{yy} = \sum_{i=1}^{n} y^2_i - n\,(\bar{y})^2, \tag{21 c}$$

$$Q_{x_1y} = \sum_{i=1}^{n}(x_{1i}\, y_i) - n\,\bar{x}_1\,\bar{y}, \quad Q_{x_2y} = \sum_{i=1}^{n}(x_{2i}\, y_i) - n\,\bar{x}_2\,\bar{y}, \tag{21 d}$$

$$Q_{x_1x_2} = \sum_{i=1}^{n}(x_{1i}\, x_{2i}) - n\,\bar{x}_1\,\bar{x}_2. \tag{21 e}$$

Тогда оценки параметров линейной зависимости (20) задаются соотношениями

$$\bar{b}_1 = \frac{Q_{x_2x_2}\cdot Q_{x_1y} - Q_{x_1x_2}\cdot Q_{x_2y}}{\Delta}, \tag{22 a}$$

$$\bar{b}_2 = \frac{Q_{x_1x_1}\cdot Q_{x_2y} - Q_{x_1x_2}\cdot Q_{x_1y}}{\Delta}, \tag{22 b}$$

$$\bar{b}_0 = \bar{y} - \bar{b}_1\,\bar{x}_1 - \bar{b}_2\,\bar{x}_2. \tag{22 c}$$

Заметим здесь, что при проведении вычислений по формулам (21), (22) часто получаются неверные результаты. Это вызвано ошибками округлений, к которым очень чувствительны эти формулы. Поэтому все вычисления должны проводиться очень аккуратно и продуманно.

Используя полученные оценки, сформируем выборку остатков

$$e_1,\, e_2, \ldots, e_i, \ldots, e_n, \tag{23}$$

которые задают отклонения наблюдаемых значений y_i от их предсказанных значений по формуле (20), т.е.

$$e_i = y_i - \bar{b}_0 - \bar{b}_1\, x_{1i} - \bar{b}_2\, x_{2i},\ 1 \leq i \leq n. \tag{24}$$

По выборке (24) вычислим остаточную сумму квадратов

$$Q_{ee} = \sum_{i=1}^{n} e^2_i, \tag{25}$$

которая будет использована в последующих расчетах. Величина Q_{ee} также может быть найдена по формуле

$$Q_{ee} = Q_{yy} - (\bar{b}_1\, Q_{x_1 y} + \bar{b}_2\, Q_{x_2 y}). \tag{26}$$

Это позволяет проверить правильность вычислений по формулам (21)–(25).

Оценки (22) дают приближенные значения параметров $b_0,\ b_1,\ b_2$, т.е.

$$b_0 \approx \bar{b}_0, \quad b_1 \approx \bar{b}_1, \quad b_2 \approx \bar{b}_2.$$

На основании этих приближенных равенств нельзя получить уверенного заключения о точных значениях параметров b_1 и b_2, а также проверить неравенства $b_1 \neq 0$ или $b_2 \neq 0$. Эти неравенства указывают на наличие или отсутствие линейной зависимости между y и объясняющими переменными x_1 и x_2. Точность оценивания параметров b_1, b_2 зависит от объема выборки n и характеризуется стандартными ошибками оценок $\bar{b}_1,\ \bar{b}_2$. Стандартные ошибки $\sigma_1,\ \sigma_2$ оценок $\bar{b}_1,\ \bar{b}_2$ находятся по формулам

$$\sigma_1 = \sqrt{\frac{Q_{ee}}{n-3} \cdot \frac{Q_{x_2 x_2}}{\Delta}}, \quad \sigma_2 = \sqrt{\frac{Q_{ee}}{n-3} \cdot \frac{Q_{x_1 x_1}}{\Delta}}. \tag{27}$$

Зафиксируем значения объясняющих переменных $x_1 = x_1^*,\ x_2 = x_2^*$. Тогда выражение

$$\hat{y} = \bar{b}_0 + \bar{b}_1\, x_1^* + \bar{b}_2\, x_2^* \tag{28}$$

будет задавать приближенное значение зависимой переменной y, т.е. $y \approx \hat{y}$. Точность оценивания y характеризуется стандартной ошибкой полученной оценки и обозначается σ_y. Стандартная ошибка σ_y задается формулой

$$\sigma_y = \sqrt{\frac{Q_{ee}}{n-3} \cdot \left(1 + \frac{1}{n} + \frac{1}{\Delta} \cdot \delta_y\right)}, \tag{29}$$

где величина δ_y равна

$$\delta_y = Q_{x_2 x_2}(x_1^* - \bar{x}_1)^2 - 2\, Q_{x_1 x_2}(x_1^* - \bar{x}_1)(x_2^* - \bar{x}_2) + Q_{x_1 x_1}(x_2^* - \bar{x}_2)^2. \tag{30}$$

Перейдем к обоснованию наличия или отсутствия линейной зависимости (20) между переменными y и $x_1,\ x_2$. Выдвигаем гипотезу H_0 об отсутствии такой зависимости. Это равносильно тому, что $b_1 = 0$ и $b_2 = 0$. Зафиксируем уровень значимости $\alpha \cdot 100\%$. Число α задает вероятность ошибки первого рода. Ошибка первого рода означает, что представленные данные и результаты их обработки не согласуются с принятой гипотезой H_0, и мы ее отвергаем. Для проверки гипотезы H_0 вычислим величину

$$F = \frac{(\bar{b}_1\, Q_{x_1 y} + \bar{b}_2\, Q_{x_2 y})\,(n-3)}{2\, Q_{ee}} \tag{31}$$

и будем сравнивать ее с критическим значением F_α распределения Фишера с $f_1 = 2$ и $f_2 = n - 3$ степенями свободы на уровне значимости α (таблица П1). Если окажется, что выполнено неравенство $F \leq F_\alpha$, то гипотезу H_0 принимаем, а линейную зависимость (20) называем не значимой. Если же $F > F_\alpha$, то гипотезу H_0 отклоняем и считаем, что между y и $x_1,\ x_2$ имеется линейная зависимость, и эта зависимость называется значимой.

Если зависимость (20) признана значимой, то следует установить какая из объясняющих переменных x_1 и (или) x_2 оказывает влияние на y. Зафиксируем уровень значимости $\alpha \cdot 100\%$ и построим доверительные интервалы для параметров b_1 и b_2:

$$b_1 \in I_1 = (\bar{b}_1 - t_\alpha \sigma_1,\, \bar{b}_1 + t_\alpha \sigma_1), \quad b_2 \in I_2 = (\bar{b}_2 - t_\alpha \sigma_2,\, \bar{b}_2 + t_\alpha \sigma_2), \tag{32}$$

где t_α – критическое значение распределения Стьюдента с $n-3$ степенями свободы на уровне значимости α (таблица П2), а величины $\sigma_1,\ \sigma_2$ заданы формулами (27). Интервалы $I_1,\ I_2$ накрывают параметры $b_1,\ b_2$ с вероятностью $p = 1 - \alpha$. Если окажется, что доверительный интервал I_1 содержит в себе число ноль, то полагаем тогда, что $b_1 = 0$, и принимаем, что переменная y не зависит от x_1. Аналогично, если I_2 содержит в себе число ноль, то полагаем, что $b_2 = 0$ и, принимаем, что переменная y не зависит от x_2. В противном случае, когда I_1 и I_2 не содержат число ноль, полагаем, что обе объясняющие переменные оказывают влияние на y. Очевидно, что при $\bar{b}_1 > 0$ возрастание x_1 будет вести к возрастанию y, а при $\bar{b}_1 < 0$ – к ее уменьшению. Аналогичный вывод о зависимости y от x_2 можно сделать по знаку $\bar{b}_2$.

Предположим, что зависимость (20) между y и $x_1,\ x_2$ признана значимой. Тогда по заданным $x_1 = x_1^*$ и $x_2 = x_2^*$ можно указать границы для ожидаемого значения y с учетом влияния случайной составляющей ε. Эти границы устанавливаются в форме доверительного интервала

$$y \in I_y = (\hat{y} - t_\alpha \sigma_y,\, \hat{y} + t_\alpha \sigma_y), \tag{33}$$

который содержит значение переменной y с вероятностью $p = 1 - \alpha$. Критическое значение t_α берется так же, как и в формуле (32), а величины $\hat{y},\ \sigma_y$ задаются формулами (28) и (29). Границы доверительного интервала $c_1 = \hat{y} - t_\alpha \sigma_y,\ c_2 = \hat{y} + t_\alpha \sigma_y$, как функции от величин x_1^* и x_2^* образуют некоторую поверхность, сечения которой напоминают фигуру, представленную на рис. 3. Точность предсказания возможных значений y убывает по мере удаления точки $(x_1^*,\, x_2^*)$ от точки $(\bar{x}_1,\, \bar{x}_2)$.

1.2.4. Проверка предположений МНК по выборке остатков.

Все приведенные выше формулы получены при условии, что выполнены предположения Н1, Н2, Н3 метода наименьших квадратов. Проверка этих предположений является необходимым этапом установления зависимости между изучаемыми переменными. Отсутствие такой проверки делает бессмысленным все вычисления, поскольку они будут не обоснованными. Проверка предположений МНК опирается на изучение свойств выборки остатков, которые задаются формулами (12) или (24). С практической точки зрения можно рекомендовать следующую последовательность действий.

Этап А. Проверка предположения Н2.

Данное предположение напрямую связано со всеми формулами, которые позволяют вычислить оценки параметров линейной зависимости. Для проверки предположения Н2) можно использовать критерий Дарбина–Уотсона. Этот критерий применяется с целью проверки гипотезы H_0 об отсутствии автокорреляции остатков. Автокорреляция остатков означает, что любая последовательная пара остатков связана между собой линейной зависимостью. В этом случае формулы (10) и (22) использовать не рекомендуется, особенно, если объем выборки не очень большой. Получаемые по этим формулам оценки $\bar{b}_0,\ \bar{b}_1,\ \bar{b}_2$ могут значительно отклоняться от $b_0,\ b_1,\ b_2$ и

не давать их истинных значений. Следствием этого может являться неправильная интерпретация результатов обработки данных и необоснованное применение формул (19), (33) для предсказания значений переменной y.

Алгоритм проверки гипотезы H_0 об отсутствии автокорреляции остатков состоит в следующем. Зафиксируем уровень значимости $\alpha \cdot 100\%$. Число α задает вероятность ошибки первого рода. Ошибка первого рода означает, что результаты обработки данных не согласуются с принятой гипотезой H_0, и мы ее отвергаем. Из таблицы распределения Дарбина–Уотсона (таблица П3) выбираются значения статистик d_L, d_U для уровня значимости α, заданного объема выборки n и числа объясняющих переменных k (у нас $k = 1$ или $k = 2$. Вычисляется величина

$$d = \frac{\sum\limits_{i=2}^{n}(e_i - e_{i-1})^2}{Q_{ee}}, \tag{34}$$

значения которой заключены в промежутке $[0, 4]$. Величина Q_{ee} задается формулами (13) и (25).

Возможны четыре случая: 1) если $d < d_L$, то гипотеза H_0 отвергается в пользу гипотезы о положительной автокорреляции; 2) если $d > 4 - d_L$, то гипотеза H_0 отвергается в пользу гипотезы об отрицательной автокорреляции; 3) если $d_U \le d \le 4 - d_U$, то гипотеза H_0 принимается; 4) если $d_L \le d \le d_U$ или $4 - d_U \le d \le 4 - d_L$, то нельзя сделать определенный вывод (требуется привлечение дополнительных данных или применение других критериев).

Случаи 1) и 2) означают, что все вычисления по приведенным формулам не являются достаточно обоснованными и никаких содержательных выводов об изучаемой зависимости сделать нельзя. Случай 3) говорит об обоснованности применения формул (10) и (22) для нахождения оценок параметров и о возможности дальнейшего анализа изучаемой зависимости. В случае 4) можно применять формулы (10) и (22), но последующие вычисления и выводы об изучаемой зависимости требуют определенной осторожности.

Этап В. Проверка предположения Н3.

Это предположение позволяет проверять значимость рассматриваемой зависимости по формулам (17), (33), строить доверительные интервалы для параметров по формулам (18), (32) и использовать формулы (19) и (33) для предсказания значений переменной y. Рассмотрим гипотезу H_0, состоящую в том, что выборка остатков извлечена из генеральной совокупности с нормальным распределением. Зафиксируем уровень значимости $\alpha \cdot 100\%$. Число α задает вероятность ошибки первого рода. Ошибка первого рода означает, что результаты обработки данных не согласуются с принятой гипотезой H_0, и мы ее отклоняем. Последнее говорит о том, что распределение генеральной совокупности отличается от нормального.

Если гипотеза H_0 отклонена, то проверка значимости модели и предсказание значений переменной y не могут опираться на указанные формулы. Исключение здесь составляет случай так называемых больших выборок, т.е. выборок, объем которых n достаточно велик, $n \sim 100$. Для таких выборок можно использовать все формулы, опирающиеся на доверительные интервалы, но формулы (17) и (33) применять не рекомендуется.

Ниже приводятся два сравнительно простых способа, которые позволяют сделать вывод о нормальности распределения без существенных вычислительных затрат.

Первый способ опирается на критерий хи-квадрат. Алгоритм работы по этому способу следующий. По выборке остатков находим величины

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i, \quad s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e})^2, \tag{35}$$

которые означают выборочное среднее и выборочную дисперсию остатков. Заметим, что величина $\bar{e}$ должна быть близка к нулю, т.е. $\bar{e} \approx 0$, так как ее теоретическое значение равно нулю. Существенное отличие $\bar{e}$ от нуля связано либо с ошибками в вычислениях, либо с грубым округлением результатов вычислений. Для нахождения s_e^2 более удобно использовать формулу

$$s_e^2 = \frac{Q_{ee} - n\,(\bar{e})^2}{n-1}, \tag{36}$$

где величина Q_{ee} задается формулами (13) и (25). Квадратный корень из s_e^2, т.е. величина $s_e = \sqrt{s_e^2}$ называется выборочным среднеквадратическим отклонением. Далее строим промежутки:

$$(-\infty,\, \bar{e} - s_e), \quad [\bar{e} - s_e,\, \bar{e} - 0.5\, s_e), \quad [\bar{e} - 0.5\, s_e,\, \bar{e}),$$

$$[\bar{e},\, \bar{e} + 0.5\, s_e), \quad [\bar{e} + 0.5\, s_e,\, \bar{e} + s_e), \quad [\bar{e} + s_e,\, +\infty).$$

Затем определяем, сколько элементов из выборки остатков попадает в эти промежутки. Их количество обозначим соответственно через n_1, n_2, n_3, n_4, n_5, n_6. Для контроля правильности подсчетов следует проверить выполнение соотношения

$$\sum_{i=1}^{6} n_i = n. \tag{37}$$

Примем, что объем выборки $n > 35$ и вычислим величину

$$N^2 = \frac{n_1^2 + n_6^2}{0.1587\, n} + \frac{n_2^2 + n_5^2}{0.1498\, n} + \frac{n_3^2 + n_4^2}{0.1915\, n} - n. \tag{38}$$

Если окажется, что $N^2 \le 7.81$, то можно считать, что выборка остатков получена из генеральной совокупности с нормальным распределением. В случае выполнения неравенства $N^2 > 11.3$ предположение о нормальном распределении в выборке остатков, т.е. гипотеза H_0, отвергается.

При объеме выборки $25 \le n \le 35$ вычисляем величину

$$M^2 = \frac{(n_1 + n_2)^2 + (n_5 + n_6)^2}{0.3085\, n} + \frac{n_3^2 + n_4^2}{0.1915\, n} - n. \tag{39}$$

Тогда, если $M^2 \le 3.84$, то считается, что выборка остатков получена из генеральной совокупности с нормальным распределением. Если же будет верно $M^2 > 6.63$, то данное предположение (гипотеза H_0) отвергается.

Если же будут выполнены неравенства $7.81 < N^2 \le 11.3$ или $3.84 < M^2 \le 6.63$, то гипотезу H_0 можно принять (при уровне ошибки первого рода $\alpha = 0.01$), но желательно провести дополнительное исследование с помощью других критериев, описанных в специальной литературе.

Второй способ ориентируется на характерные особенности графика плотности распределения случайных величин с нормальным распределением. Эти особенности проявляются в генеральных коэффициентах асимметрии и эксцесса, которые для нормального распределения равны нулю. Алгоритм работы по данному способу следующий. По выборке остатков вычисляем $\bar{e},\ s_e$, а также выборочные коэффициенты асимметрии – A_e и эксцесса – E_e:

$$A_e = \frac{1}{s_e^3} \cdot \frac{1}{n} \sum_{i=1}^{n} (e_i - \bar{e})^3, \quad E_e = \frac{1}{s_e^4} \cdot \frac{1}{n} \sum_{i=1}^{n} (e_i - \bar{e})^4 - 3. \tag{40}$$

Находим дисперсии этих коэффициентов

$$D(A_e) = \frac{6\,(n-1)}{(n+1)\,(n+3)}, \quad D(E_e) = \frac{24\,n\,(n-2)\,(n-3)}{(n+1)^2\,(n+3)\,(n+5)}. \tag{41}$$

Если окажется, что выборочные коэффициенты асимметрии и эксцесса удовлетворяют неравенствам

$$|A_e| \le 3\sqrt{D(A_e)}, \quad |E_e| \le 5\sqrt{D(E_e)}, \tag{42}$$

то можно считать, что выборка остатков получена из генеральной совокупности с нормальным распределением. В противном случае предположение о нормальности следует отвергнуть или считать сомнительным. Необходимо учитывать тот факт, что вывод, полученный по данному способу, является весьма приближенным. Применение этого способа рекомендуется в случаях, когда объем выборки сравнительно мал (n лежит в промежутке 10–25) и невозможно применить критерий хи-квадрат или другие критерии.

Этап С. Проверка предположения Н1.

Данное предположение связано с однородностью результатов наблюдений (измерений) зависимой и объясняющих переменных. Это означает, что случайная составляющая ε и ее значения $\varepsilon_1,\ \varepsilon_2, \ldots,\ \varepsilon_n$, формирующие выборку значений зависимой переменной $y_1,\ y_2, \ldots,\ y_n$, сохраняют неизменное математическое ожидание и дисперсию. Другими словами, все наблюдения (измерения) проводятся с одинаковой точностью, задаваемой дисперсией $D(\varepsilon_i) = \sigma^2 = const > 0,\ 1 \le i \le n$. Это свойство называется гомоскедастичностью (однородностью) результатов наблюдений (измерений). В некоторых случаях дисперсия $D(\varepsilon_i)$ величин ε_i может зависеть от номера наблюдения (измерения) i, значений переменных y_i, x_{1i}, x_{2i} и т.д. В этом случае говорят о том, что результаты наблюдений (измерений) являются неоднородными, и имеет место гетероскедастичность наблюдений (измерений).

Последствия гетероскедастичности проявляются в следующем. Формулы (10) и (22) можно использовать для нахождения оценок $\bar{b}_0,\ \bar{b}_1,\ \bar{b}_2$. Однако стандартные ошибки оценок и зависимой переменной уже не могут вычисляться по формулам (14), (16), (27), (29). Это означает, что проверка значимости влияния объясняющих переменных и предсказание значений зависимой переменной по формулам (18), (19), (32), (33) становится невозможным.

Рассмотрим гипотезу H_0, состоящую в том, что гетероскедастичность отсутствует. Зафиксируем уровень значимости $\alpha \cdot 100\%$. Число α задает вероятность ошибки первого рода. Ошибка первого рода означает, что результаты обработки данных не

согласуются с принятой гипотезой H_0, и мы ее отклоняем. Отклонение этой гипотезы свидетельствует о том, что результаты наблюдений (измерений) не являются гомоскедастичными.

Ниже приводятся два теста, позволяющие проверять гипотезу H_0. В обоих тестах изучается взаимосвязь между значениями остатков e и значениями одной из объясняющих переменных x_1 или x_2. Если рассматривается зависимость y от двух переменных, то каждый из тестов нужно применять последовательно к паре $(e,\, x_1)$ и паре $(e,\, x_2)$. Для определенности возьмем пару $(e,\, x_1)$. Из таблицы 1 формируем выборку значений переменной x_1. Записываем эту выборку и выборку остатков в следующей форме

$$x_{11},\, x_{12}, \ldots, x_{1i}, \ldots, x_{1n}, \tag{43}$$

$$e_1,\, e_2, \ldots, e_i, \ldots, e_n, \tag{44}$$

Тест Голдфелда–Квандта.

Этот тест применяется в предположении, что выборка остатков извлечена из генеральной совокупности с нормальным распределением. Заметим, что именно такое предположение проверяется на этапе В. Алгоритм работы по тесту следующий. Упорядочиваем выборку (43) в порядке возрастания значений x_{1i}. В соответствии с новым порядком переставляем элементы выборки (44). Полагаем, что она будет записана в форме

$$\bar{e}_1,\, \bar{e}_2, \ldots, \bar{e}_j, \ldots, \bar{e}_n, \tag{45}$$

где $\bar{e}_j$ – это остатки e_i, расставленные в новом порядке. Задаем число m как целую часть дроби $n/3$. Вычисляем суммы

$$G_1 = \sum_{j=1}^{m} (\bar{e}_j)^2\,, \quad G_2 = \sum_{j=n-m+1}^{n} (\bar{e}_j)^2\,, \tag{46}$$

и из них выбираем наибольшую G_{max} и наименьшую G_{min}. Далее находим величину

$$G = \frac{G_{max}}{G_{min}} \tag{47}$$

и сравниваем ее с критическим значением F_α распределения Фишера с $f_1 = f_2 = m-k$ степенями свободы на уровне значимости α (таблица П1). Параметр k означает число объясняющих переменных (у нас $k = 1$ или $k = 2$). Если окажется, что выполнено неравенство $G \le F_\alpha$, то гипотезу H_0 принимаем и считаем, что имеет место гомоскедастичность результатов наблюдений (измерений). Если же $G > F_\alpha$, то гипотезу H_0 отклоняем в пользу гетероскедастичности этих результатов. Ошибка первого рода равна α.

Тест ранговой корреляции Спирмена.

Этот тест не требует никаких специальных предположений относительно выборки остатков, за исключением того, что объем выборки $n \ge 9$. Тест проверяет отсутствие монотонной зависимости между модулями остатков и значениями переменной x_1. Алгоритм работы по тесту следующий.

Вместо выборки (44) рассматриваем выборку из модулей остатков, а именно

$$|e_1|,\, |e_2|, \ldots, |e_i|, \ldots, |e_n|. \tag{48}$$

Выборку (43) упорядочиваем по возрастанию и находим ранги ее элементов. Рангом элемента x_{1i} называется номер того места, на который x_{1i} встанет в упорядоченной выборке. Для элементов, имеющих одинаковые значения, ранг берется одинаковым и равным среднему арифметическому мест, на которые они встали. Ранг элемента x_{1i} обозначим через β_{1i}. Аналогично упорядочиваем по возрастанию выборку (48) и находим ранги ее элементов. Ранг элемента $|e_i|$ обозначим через γ_{1i}.

Составим разности между рангами по формуле: $w_i = \beta_{1i} - \gamma_{1i},\ 1 \le i \le n$. Вычислим коэффициент ранговой корреляции по Спирмену

$$R_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} w_i^2. \tag{49}$$

Используя R_s, найдем величину

$$T_s = |R_s| \cdot \sqrt{\frac{n-2}{1 - R_s^2}}, \tag{50}$$

которую будем сравнивать с критическим значением t_α распределения Стьюдента с $n-2$ степенями свободы на уровне значимости α (таблица П2). Если окажется, что выполнено неравенство $T_s \le t_\alpha$, то гипотеза H_0 принимается и, следовательно, имеет место гомоскедастичность результатов наблюдений (измерений). При выполнении неравенства $T_s > t_\alpha$ гипотеза H_0 отвергается и можно говорить о том, что имеет место гетероскедастичность результатов наблюдений (измерений). Ошибка первого рода равна α.

1.2.5. Выбор уровня значимости для проверки гипотезы H_0.

Уровень значимости α задают заранее, до начала обработки данных. В научных исследованиях обычно используют следующие значения:

$$\alpha = 0.05, \quad 0.01, \quad 0.001.$$

В соответствии с принятой терминологией, отклонение гипотезы H_0 на уровне $\alpha = 0.05$ называют значимым, на уровне $\alpha = 0.01$ – статистически значимым и на уровне $\alpha = 0.001$ – высоко статистически значимым. В некоторых задачах, связанных, например, с исследованием технических изделий или производственных процессов, иногда используется уровень $\alpha = 0.1$.

В ходе обработки данных может возникнуть ситуация, когда мы отвергаем гипотезу H_0 на уровне $\alpha = 0.05$, но вместе с тем мы может ее принять на уровне $\alpha = 0.01$. Аналогичная ситуация возникает и для уровней значимости $\alpha = 0.01$ и $\alpha = 0.001$. В таких случаях рекомендуется провести дополнительное исследование, собрать новые данные и провести их обработку. При этом можно объединить имеющиеся и вновь полученные данные, либо обработку провести отдельно. При невозможности дополнительных исследований требуется привлечь другие статистические методы для проверки гипотезы H_0.

1.3. НЕЛИНЕЙНАЯ РЕГРЕССИОННАЯ ЗАВИСИМОСТЬ

1.3.1. Нелинейные зависимости и их преобразования.

При изучении взаимосвязей между переменными часто возникают регрессионные зависимости

$$y = f(b, u) + \varepsilon = f(b,\, x_1,\, x_2, \ldots, x_k) + \varepsilon, \tag{51}$$

в которых функция f не линейна по одной или нескольким объясняющим переменным, но линейна по параметрам. На практике используются степенная, логарифмическая, показательная и обратная зависимости. Так, например, степенная зависимость задается формулой $y = b_0 + b_1\, x + b_2\, x^2 + \ldots + b_k\, x^k + \varepsilon$. Изучение этой зависимости может быть сведено к множественной линейной регрессии, в которой полагаем, что $x_1 = x$, $x_2 = x^2, \ldots, x_k = x^k$. Для логарифмической зависимости $y = b_0 + b_1 \ln x + \varepsilon$ можно использовать новую переменную $x_1 = \ln x$. Показательная зависимость $y = b_0 + b_1 \exp(x) + \varepsilon$ переходит в линейную при замене $x_1 = \exp(x)$. Обратная зависимость $y = b_0 + b_1/x + \varepsilon$ преобразуется в линейную путем замены $x_1 = 1/x$.

К нелинейным зависимостям относятся регрессионные зависимости, задаваемые с помощью аналогов производственных функций $y = \alpha\, x_1^{\beta}\, x_2^{\gamma} \cdot \varepsilon_+$, где $\varepsilon_+ > 0$ – случайная составляющая. Эта зависимость сводится к линейной путем логарифмирования обеих частей уравнения $\ln y = \ln \alpha + \beta \ln x_1 + \gamma \ln x_2 + \ln \varepsilon_+$ и переобозначения переменных.

В общем случае преобразование и замена переменных не всегда упрощает нелинейные зависимости и сводит их к линейным. Более того, нужно иметь в виду, что указанные замены могут привести к нарушениям основных предпосылок МНК. Это в свою очередь может поставить под сомнение результаты расчетов.

1.3.2.Сравнение и выбор наиболее подходящей зависимости.

При изучении зависимости (51) может быть выбрано несколько функций f, которые устанавливают связь между переменными y и x_1, $x_2, \ldots, x_k$. Возникает задача: какую зависимость выбрать? Решение этой задачи предполагает несколько этапов. На первом этапе отбрасываем все зависимости, для которых не выполняются предположения МНК или те зависимости, которые не являются значимыми. На втором этапе сравниваем оставшиеся зависимости, используя соотношение

$$Q_{yy} = Q_{rr} + Q_{ee}, \tag{52}$$

где $Q_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ – сумма квадратов отклонений зависимой переменной от ее средневыборочного значения, $Q_{rr} = \sum_{i=1}^{n} (y_i - f(\bar{b}, u_{[i]}))^2$ – сумма квадратов отклонений зависимой переменной от ее средневыборочного значения, которая объясняется с помощью уравнения регрессии, Q_{ee} – остаточная сумма квадратов, отражающая влияние неучтенных факторов и случайных составляющих. Формула (52) определяет коэффициент детерминации

$$R^2 = \frac{Q_{rr}}{Q_{yy}} = 1 - \frac{Q_{ee}}{Q_{yy}}. \tag{53}$$

Заметим, что $0 \le R^2 \le 1$. Чем ближе коэффициент R^2 к единице, тем лучше выбранная зависимость (51) объясняет значение Q_{yy}. Из двух зависимостей (51), предложенных для описания переменной y, выбирают ту, для которой больше R^2. Если число

объясняющих переменных в зависимостях существенно различно, то для их сравнения рекомендуется использовать скорректированный коэффициент детерминации, который описан в специальной литературе.

Укажем на важную взаимосвязь коэффициента детерминации R^2 с величиной F статистики Фишера, которая вычисляется по формулам (17) и (31) при проверке значимости линейных зависимостей. Эта взаимосвязь задается соотношением

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k}, \tag{54}$$

где k – число объясняющих переменных (у нас $k = 1$ или $k = 2$). Из этой формулы, например, видно, что не значимость изучаемой зависимости (случай $F \le F_p$) обусловлена малыми значениями коэффициента детерминации R^2.

1.4. ПРИМЕРЫ

Пример 1. Изучается зависимость стоимости квартиры y (тыс. у.е.) от ее общей площади x_1 (кв.м.) Для удобства анализа данные (x_1, y) упорядочены по переменной x_1 и записаны в следующей таблице.

x_1	28	29	32	35	40	44	45	51	53
y	5.3	9.2	15.2	20.7	21.7	36.5	39.3	52.7	55.4
x_1	58	64	65	73	75	80	83	93	-
y	64.3	76	79.1	94.8	101	89.5	114.8	137.4	-

Из этой таблицы видно, что практически для всего набора данных росту x_1 соответствует возрастание y. Требуется установить зависимость между y и x_1 в линейной форме (уровень значимости равен 5%). Здесь объем выборки $n = 17$. Опираясь на таблицу данных и формулы раздела 1.2.2, находим, что

$$\bar{x}_1 = 55.76, \quad \bar{y} = 59.55, \quad Q_{x_1x_1} = 6557.06, \quad Q_{x_1y} = 12714.79,$$
$$Q_{ee} = 27.09, \quad \bar{b}_0 = -48.4705, \quad \bar{b}_1 = 1.9377.$$

Для обоснованности применения указанных формул рассмотрим выборку остатков. Используя найденные оценки $\bar{b}_0$, $\bar{b}_1$, получим набор чисел e_1, $e_2, \ldots, e_{17}$:

- 0.48, 1.48, 1.67, 1.35, - 7.34, - 0.29, 0.58, 2.35, 1.17, 0.39, 0.46, 1.62, 1.82, 4.15, - 17.04, 2.44, 5.67.

В качестве примера вычислим e_2. Имеем:

$$e_2 = y_2 - (\bar{b}_0 + \bar{b}_1\, x_{12}) = 9.2 - (-48.4705 + 1.9377 \cdot 29) = 1.4772 \approx 1.48.$$

Обратимся к проверке гипотезы H_0 об отсутствии автокорреляции остатков, используя критерий Дарбина–Уотсона. Задавая $\alpha = 5\%$, $k = 1$, $n = 17$, из таблицы П3 находим значения статистик $d_L = 1.13$, $d_U = 1.38$. Вычисляя величину d по формуле (34), получаем, что $d = 1.67$ и верно неравенство $d_U \le d \le 4 - d_U$. Поэтому гипотеза H_0 принимается.

Проверим гипотезу H_0 о том, что остатки получены из генеральной совокупности с нормальным распределением. Для этого вычислим выборочные коэффициенты асимметрии и эксцесса, которые будут равны $A_e = -2.65$, $E_e = 8.05$. Получим, что

$$|A_e| = 2.65 > 3\sqrt{D(A_e)} = 3\sqrt{0.266} = 1.55,$$
$$|E_e| = 8.05 > 5\sqrt{D(E_e)} = 5\sqrt{0.601} = 3.88.$$

Отсюда делаем вывод о том, что генеральное распределение остатков скорее всего отличается от нормального. Поскольку объем выборки небольшой, то нет оснований применять формулы (17), (18) и (19) для проверки значимости линейной зависимости, границ доверительного интервала неизвестного параметра b_1 и предсказания возможных значений переменной y.

Рассмотрим гипотезу H_0 об отсутствии гетероскедастичности. Зададим $\alpha = 5\%$. Поскольку нормальность распределения остатков не подтверждена, то на это свойство опираться нельзя. Поэтому применим здесь тест ранговой корреляции Спирмена. В исходной таблице данных переменные x_{1i} упорядочены по возрастанию и среди них нет совпадающих элементов. Потому ранг x_{1i} равен $\beta_{1i} = i,\ 1 \le i \le 17$. Выборка, содержащая модули остатков, имеет вид

$$0.48,\ 1.48,\ 1.67,\ 1.35,\ 7.34,\ 0.29,\ 0.58,\ 2.35,\ 1.17,\ 0.39,\ 0.46,\ 1.62,\ 1.82,\ 4.15,\ 17.04,\ 2.44,\ 5.67.$$

Упорядочивая эту выборку по возрастанию, находим ранги ее элементов:

$$\gamma_{11} = 4,\ \gamma_{12} = 8,\ \gamma_{13} = 10,\ \gamma_{14} = 7,\ \gamma_{15} = 16,\ \gamma_{16} = 1,\ \gamma_{17} = 5,\ \gamma_{18} = 12,\ \gamma_{19} = 6,$$

$$\gamma_{110} = 2,\ \gamma_{111} = 3,\ \gamma_{112} = 9,\ \gamma_{113} = 11,\ \gamma_{114} = 14,\ \gamma_{115} = 17,\ \gamma_{116} = 13,\ \gamma_{117} = 15.$$

Используя формулу (49), вычислим коэффициент ранговой корреляции по Спирмену. Получим $R_s = 0.45$. По формуле (50) находим, что $T_s = 1.952$. Сравним T_s с критическим значением t_α распределения Стьюдента с $n - 2 = 15$ степенями свободы (таблица П2). Имеем, что $T_s = 1.952 < t_\alpha = 2.131$. Следовательно, гипотеза H_0 принимается, и можно считать, что имеет место гомоскедастичность результатов наблюдений (измерений).

В итоге можно утверждать, что величина y находится в монотонно возрастающей зависимости от величины x_1 и эта зависимость приближенно описывается формулой $y \approx -48.47 + 1.94\,x_1$. Естественно, что более детальный анализ изучаемой зависимости требует привлечения и других объясняющих переменных. К ним могут относиться, например, этаж, тип дома и т.д.

Пример 2. Изучается зависимость урожайности зерновых культур y (ц/га) от количества внесенных удобрений на 1 га пашни x_1 (т). Обследовано десять хозяйств, по которым получены следующие данные:

x_1	3.5	5.0	6.5	10.5	13.0
y	16.4	15.2	14.6	20.8	26.6
x_1	4.0	7.5	8.5	6.0	12.5
y	12.7	15.5	17.0	14.2	25.9

Требуется построить и обосновать линейную зависимость между y и x (уровень значимости равен 5%). В этой задаче объем выборки $n = 10$. Используя формулы раздела 1.2.2, находим, что $\bar{b}_0 = 7.72,\ \ \bar{b}_1 = 1.32$. Следовательно, зависимость y от x_1 приближенно описывается формулой $y \approx 7.72 + 1.32\,x_1$.

Как и в предыдущем примере, рассмотрим выборку остатков. Используя найденные оценки $\bar{b}_0,\ \bar{b}_1$, получим (с учетом округления) набор чисел $e_1,\ e_2, \ldots, e_{10}$:

$$4.06,\ 0.88,\ -1.71,\ -0.79,\ 1.71,\ -0.30,\ -2.13,\ -1.95,\ -1.45,\ 1.67.$$

Проверим, что эти остатки получены из генеральной совокупности с нормальным распределением. Вычислим выборочные коэффициенты асимметрии и эксцесса, которые будут соответственно равны $A_e = 0.86,\ E_e = 0.003$. Получим, что

$$|A_e| = 0.86 < 3\sqrt{D(A_e)} = 3\sqrt{0.377} = 1.84,$$
$$|E_e| = 0.003 < 5\sqrt{D(E_e)} = 5\sqrt{0.569} = 3.77.$$

Поэтому будем считать, что генеральное распределение остатков описывается нормальным законом.

Обратимся к проверке гипотезы H_0 об отсутствии автокорреляции остатков, используя критерий Дарбина–Уотсона. Выбирая $\alpha = 5\%,\ k = 1,\ n = 10$, из таблицы П3 находим значения статистик $d_L = 0.88,\ d_U = 1.32$. Вычисляя величину d по формуле (34), получаем, что $d = 1.12$ и верно неравенство $d_L \le d \le d_U$. Поэтому гипотеза H_0 не может быть принята или отвергнута. Здесь требуется привлечение дополнительных данных или других критериев проверки этой гипотезы.

Рассмотрим теперь гипотезу H_0 об отсутствии гетероскедастичности. Поскольку гипотеза о нормальном распределении остатков подтверждена, то можно использовать тест Голдфелда–Квандта. Имеем, что

$$x_{11} = 3.5,\ x_{12} = 5.0,\ x_{13} = 6.5,\ x_{14} = 10.5,\ x_{15} = 13.0,$$

$$x_{16} = 4.0,\ x_{17} = 7.5,\ x_{18} = 8.5,\ x_{19} = 6.0,\ x_{110} = 12.5.$$

Упорядочим значения объясняющей переменной x_1 в порядке возрастания. Получим следующий порядок их расположения:

$$x_{11},\ x_{16},\ x_{12},\ x_{19},\ x_{13},\ x_{17},\ x_{18},\ x_{14},\ x_{110},\ x_{15}.$$

В соответствии с этим порядком выборка остатков запишется в такой последовательности

$$e_1,\ e_6,\ e_2,\ e_9,\ e_3,\ e_7,\ e_8,\ e_4,\ e_{10},\ e_5,$$

иначе,

$$4.06,\ \ -0.30,\ \ 0.88,\ \ -1.45,\ \ -1.71,\ \ -2.13, -1.95,\ \ -0.79,\ \ 1.67,\ \ 1.71.$$

Зададим число m как целую часть от дроби $n/3 = 10/3$. Округляя в большую сторону, возьмем $m = 3$. Тогда $n - m + 1 = 8$. Для вычисления величины G по формулам (46), (47) из упорядоченной выборки остатков выберем первый, второй, третий остатки (они дадут сумму G_1), а также восьмой, девятый, десятый остатки (они дадут сумму G_2). Получим

$$G_1 = (4.06)^2 + (-0.30)^2 + (0.88)^2 = 17.348,\ \ G_2 = (-0.79)^2 + (1.67)^2 + (1.71)^2 = 6.3371,$$

$$G_{max} = 17.348,\ \ G_{min} = 6.3371,\ \ G = \frac{17.348}{6.3371} = 2.74.$$

Критическое значение F_α распределения Фишера выбирается для уровня значимости $\alpha = 5\%$ и числа степеней свободы $f_1 = f_2 = m - k = 3 - 1 = 2$, т.е. $F_\alpha = 19.0$. Так как выполнено неравенство $G \le F_\alpha$, то гипотезу H_0 принимаем и считаем, что имеет место гомоскедастичность результатов наблюдений (измерений).

Оценим значимость линейной зависимости на уровне 5%. Сначала применим первый способ. По формуле (17) вычислим величину F. Она будет равна $F = 38.18$. Эту величину сравним с критическим значением F_α распределения Фишера с $f_1 = 1$ и $f_2 = 8$ степенями свободы на уровне значимости $\alpha = 0.05$ (см. таблицу П1). Это значение равно $F_{0.05} = 5.32$. Поскольку выполнено неравенство $F > F_{0.05}$, то считаем, что между переменными x_1 и y действительно имеется линейная зависимость. Применим далее второй способ и построим границы доверительного интервала для параметра b_1. Используя формулы (14) и (18), уровень значимости 5%, получим, что $t_\alpha = t_{0.05} = 2.306$ и $b_1 \in I_1 = (0.84, 1.8)$. Как видно, доверительный интервал I_1 не накрывает число ноль, поэтому и здесь принимаем решение о том, что между x_1 и y имеется линейная зависимость.

Полученные результаты говорят о том, что для анализа данных можно применять формулу линейной зависимости $y \approx 7.72 + 1.32\,x_1$ и формулу (19) для оценки урожайности y при заданном значении $x_1 = x_1^*$ внесенных удобрений (с указанной выше оговоркой, вызванной значением коэффициента d). Отметим здесь, что коэффициент $\bar{b}_0 = 7.72$ указывает на нижнюю границу урожайности без применения удобрений. Коэффициент $\bar{b}_1 = 1.32$ отражает прирост урожайности на одну тонну внесенных удобрений.

Пример 3. Изучается зависимость товарооборота магазина y (тыс. руб./нед) от численности работающих x_1 (чел.) и площади подсобных помещений x_2 (кв.м). Результаты исследований по $n = 8$ магазинам представлены в следующей таблице:

x_1	31	34	35	41	38	32	29	34
x_2	29.5	14.2	18.0	21.3	47.5	10.0	21.0	36.5
y	22.0	14.0	23.0	43.0	66.0	7.6	12.0	36.0

На основании этих данных требуется установить наличие зависимости величины y от x_1 и x_2, а также оценить их относительный вклад в величину товарооборота. Уровень значимости возьмем равным 5%. Примем, что данные в таблице таковы, что выполнены все предположения МНК относительно случайных ошибок наблюдений $\varepsilon_i = y_i - b_0 - b_1\,x_{1i} - b_2\,x_{2i}$, $1 \le i \le 8$ (их проверять не нужно).

Для решения поставленной задачи вычислим промежуточные величины, используемые в формулах раздела 1.2.3. Используя данные из таблицы, получаем, что

$$\sum_{i=1}^{8} x_{1i} = 274, \quad \sum_{i=1}^{8} x_{2i} = 198, \quad \sum_{i=1}^{8} y_i = 223.6,$$

$$\sum_{i=1}^{8} x_{1i}^2 = 9488, \quad \sum_{i=1}^{8} x_{2i}^2 = 5979.08, \quad \sum_{i=1}^{8} y_i^2 = 8911.76,$$

$$\sum_{i=1}^{8} (x_{1i}\,x_{2i}) = 6875.6, \quad \sum_{i=1}^{8} (x_{1i}\,y_i) = 8049.2, \quad \sum_{i=1}^{8} (x_{2i}\,y_i) = 6954.7.$$

Отсюда по формулам (22) находим, что

$$\bar{b}_0 = -94.55, \quad \bar{b}_1 = 2.80, \quad \bar{b}_2 = 1.07.$$

Применяя формулу (31), проверим гипотезу H_0 о не значимости линейной модели. Получаем, что $F = 151.7$, а критическое значение распределения Фишера при $f_1 = 2$

и $f_2 = 5$ степенях свободы равно $F_{0.05} = 5.79$. Поскольку выполнено неравенство $F > F_{0.05}$, то гипотезу H_0 отвергаем и линейную зависимость между y и объясняющими переменными $x_1,\ x_2$ считаем значимой.

Построим далее доверительные интервалы для параметров b_1, b_2 и установим влияние каждой переменной x_1 и x_2 на y. По формулам (27), (32) находим, что $I_1 = (2.03,\ 3.57),\ I_2 = (0.83,\ 1.31)$, причем оба интервала не содержат число ноль. В итоге получаем, что линейная зависимость имеет место, а значения товарооборота y можно находить по приближенной формуле: $y \approx -94.55 + 2.80\,x_1 + 1.07\,x_2$. Из этой формулы следует, что вклад в товарооборот каждого работающего примерно в два раза больше, чем одного квадратного метра подсобных помещений.

ЧАСТЬ 2. ВРЕМЕННЫЕ РЯДЫ

2.1. ПОСТАНОВКА ЗАДАЧИ

При проведении различных исследований приходится изучать объекты, характеристики которых изменяются во времени. Эти характеристики могут иметь определенные (детерминированные) тенденции, на которые накладываются случайные составляющие. Одна из простейших зависимостей, описывающих указанную ситуацию, задается соотношением

$$y(t) = f(t,\, x(t)) + \varepsilon(t),\ t = 0,\, 1,\, 2, \ldots,\, T. \tag{55}$$

Здесь переменная t означает время, выраженное в условных единицах (месяц, год и т.д.), $y(t)$ задает наблюдаемую переменную, которая зависит от детерминированной составляющей $x(t)$ и случайной составляющей $\varepsilon(t)$. Предполагается, что вид функции $f(t,\, x(t))$ известен, быть может, с точностью до входящих в нее параметров, а $M(\varepsilon(t)) = 0$. В общем случае могут иметь место более сложные зависимости, описывающие изменение переменной $y(t)$ во времени. В некоторых ситуациях форма зависимости вообще может быть неизвестной, и ее следует получить путем предварительного анализа результатов наблюдений (измерений).

Примем, что имеется набор данных

$$y(t_1),\ y(t_2),\ y(t_3), \ldots, y(t_n), \tag{56}$$

которые получены в результате наблюдения объекта V в моменты времени

$$t = t_1 < t_2 < t_3 < \ldots < t_n. \tag{57}$$

Данные (56) являются исходной информацией, на основании которой можно решать различные задачи относительно переменной $y(t)$. Основными из этих задач являются следующие:

1) установление наиболее простой зависимости, достаточно хорошо описывающей набор данных (56);

2) оценка параметров установленной зависимости, построение доверительных интервалов для них;

3) проведение расчетов с целью получения возможных значений переменной $y(t)$ в заданный период времени $t > T = t_n$.

Первые две задачи решаются в рамках анализа временных рядов, который опирается на определенный набор математических моделей, описывающих данные (56). Типичными представителями здесь являются трендовая модель, модель распределенных лагов, модель взаимосвязи двух рядов. Третья задача связана с предсказанием значений переменной $y(t)$ при $t > T$. Сущность предсказания состоит в том, что для переменной $y(t)$ указываются верхняя и нижняя границы ее возможных значений с заданным уровнем вероятности. Знание таких границ позволяет оценить риск при принятии решения на основе сделанного прогноза. Построение верхней и нижней границ опирается на определенные свойства переменной $y(t)$, которые в конкретных задачах могут и не выполняться.

Особенность переменной $y(t)$ состоит в том, что она является случайной функцией времени (в другой терминологии – случайным процессом). Это означает, что в

каждый фиксированный момент $t = t_*$ переменная $y(t_*)$ представляет собой случайную величину с заданной функцией распределения

$$F_*(x) = P\{y(t_*) \le x\}. \tag{58}$$

Однако этим не исчерпывается описание $y(t)$. Главная информация находится в совместной функции распределения значений $y(t)$ в заданные моменты времени t_1, $t_2, \ldots,\ t_k$, то есть в функции

$$F(x_1, x_2, \ldots, x_k) = P\{y(t_1) \le x_1,\, y(t_2) \le x_2, \ldots,\, y(t_k) \le x_k\}. \tag{59}$$

Знание такой функции позволяет, например, предсказывать значения переменной $y(t_k)$ по известным значениям

$$y(t_1),\, y(t_2), \ldots, y(t_{k-1}), \quad t_1 < t_2 < \cdots < t_{k-1} < t_k. \tag{60}$$

Множество всех значений составной величины $\{y(t_1),\, y(t_2), \ldots, y(t_k)\}$ образует генеральную совокупность, соответствующую функции распределения (59). Каждая конкретная составная величина $\{y(t_1),\, y(t_2), \ldots, y(t_k)\}$ называется реализацией случайной функции $y(t)$. В ходе изучения объекта V мы можем иметь дело с любой реализацией $\{y(t_1),\, y(t_2), \ldots, y(t_k)\}$, но фактически будем располагать только какой-то одной из них, а именно той, что представлена в наборе данных (56). Эта одна реализация, как правило, и применяется для предсказания значений $y(t)$ при заданных $T < t \le T_1$ (см. рис. 4, 5).

2.2. ТРЕНДОВЫЕ МОДЕЛИ

Примем, что переменная $y(t)$ описывается зависимостью

$$y(t) = g(t) + \varepsilon(t), \tag{61}$$

в которой $g(t)$ – заданная функция, а переменная t принимает значения $t_1 < t_2 < \ldots < t_n$. Функция $g(t)$ называется трендом временного ряда. В приложениях часто используются функции вида:

$$a + b\,t,\ a + b\,t + c\,t^2,\ e^{a+bt},\ a\,t^{\,b},\ a + b\,\ln t. \tag{62}$$

Здесь a, b, c – неизвестные параметры, которые должны быть найдены по набору данных (56). Говорят, что временной ряд имеет линейный тренд, если $g(t) = a + b\,t$. Пример такого ряда приведен на рис. 4. Аналогично определяются квадратичный, экспоненциальный, степенной и логарифмический тренды.

Если вместо переменной $y(t)$ рассматривать новую переменную $y(t) - g(t)$, то такая операция называется устранением тренда. Цель устранения тренда состоит в приведении данных (56) к более простому виду, удобному для последующего анализа и обработки.

Случайная составляющая $\varepsilon(t)$ предполагается такой, что

$$M(\varepsilon(t)) = 0, \quad D(\varepsilon(t)) = \sigma^2 = const > 0, \tag{63}$$

а случайные величины

$$\varepsilon(t_1),\, \varepsilon(t_2), \ldots,\, \varepsilon(t_n) \tag{64}$$

взаимно независимы и имеют нормальное распределение. Условия (63) и (64) соответствуют основным предположениям МНК. Это дает возможность применить метод регрессионного анализа, описанного в первой части.

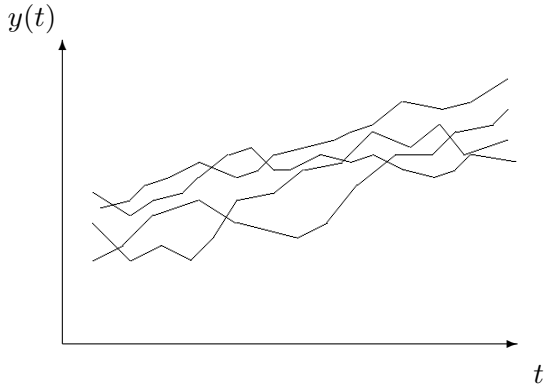

Рис. 4. Реализации случайной функции $y(t)$

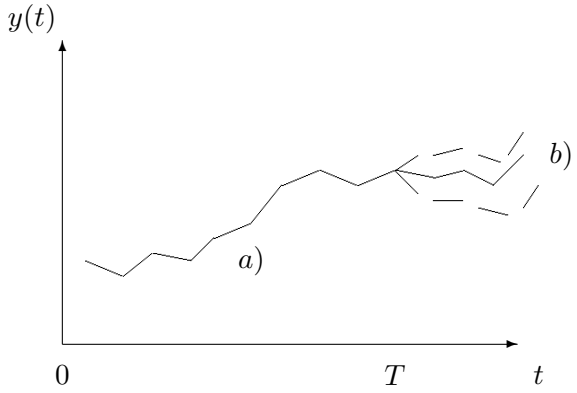

Рис. 5. Предсказание значений $y(t)$: a)–исходные данные, b)–границы доверительных интервалов

В качестве примера рассмотрим модель с квадратичным трендом

$$y(t) = a + b\,t + c\,t^2 + \varepsilon(t), \tag{65}$$

считая, что все предположения МНК выполнены. Введем новые параметры: $b_0 = a$, $b_1 = b$, $b_2 = c$. Определим новые переменные: $y = y(t)$, $x_1 = t$, $x_2 = t^2$, $\varepsilon = \varepsilon(t)$. Зависимость (65) перепишем в форме

$$y = b_0 + b_1\,x_1 + b_2\,x_2 + \varepsilon. \tag{66}$$

Пусть в наборе данных (56) представлены значения

$$y(1) = 1.5, \quad y(2) = 2.4, \quad y(3) = 0.8, \quad y(4) = 2.7, \quad y(5) = 3.0 \tag{67}$$

переменной $y(t)$, заданной в моменты времени $t = 1, 2, 3, 4, 5$. Тогда мы получим следующие выборки значений новых переменных

$$x_{11} = 1, \quad x_{12} = 2, \quad x_{13} = 3, \quad x_{14} = 4, \quad x_{15} = 5, \tag{68\,a}$$

$$x_{21} = 1, \quad x_{22} = 4, \quad x_{23} = 9, \quad x_{24} = 16, \quad x_{25} = 25, \tag{68\,b}$$

$$y_1 = 1.5, \quad y_2 = 2.4, \quad y_3 = 0.8, \quad y_4 = 2.7, \quad y_5 = 3.0. \tag{68\,c}$$

Набор данных (68) представляет собой выборки значений переменных $x_1,\ x_2,\ y$, которые используются далее в формулах раздела 1.2.3. Применение указанных формул позволяет оценить параметры $b_0,\ b_1,\ b_2$ зависимости (64), а по ним – параметры тренда модели (65).

Пример 4. В приведенной ниже таблице представлены данные о почасовой ставке заработной платы $y(t)$ (доллары США) в легкой и текстильной промышленности за период с января 1984 г. по декабрь 1987 г. ($t = 1, 2, \ldots, 48$)

Месяц	Год			
	1984	1985	1986	1987
Январь	5.50	5.73	5.82	5.94
Февраль	5.46	5.70	5.79	5.93
Март	5.48	5.73	5.80	5.93
Апрель	5.49	5.74	5.81	5.94
Май	5.48	5.69	5.78	5.89
Июнь	5.50	5.70	5.79	5.91
Июль	5.53	5.70	5.79	5.79
Август	5.55	5.69	5.83	5.83
Сентябрь	5.63	5.75	5.91	5.91
Октябрь	5.61	5.74	5.87	5.87
Ноябрь	5.61	5.75	5.87	5.87
Декабрь	5.68	5.80	5.90	5.90

Графический анализ данных этой таблицы показывает, что переменная $y(t)$ имеет определенную тенденцию к возрастанию. Установим вид этой тенденции, опираясь на модель (61) и данные из таблицы за 1984–1986 годы (объем выборки $n = 36$). Из анализа данных всей таблицы прослеживается квадратичный тренд. Задавая $g(t) = a + bt + ct^2$ и применяя указанный выше способ, получаем, что оценки параметров равны: $\bar{a} = 5.434$, $\bar{b} = 0.019$, $\bar{c} = -0.0002$. Изучая остатки, находим, что величина d для критерия Дарбина–Уотсона равна $d = 0.922$. Для уровня значимости $\alpha = 5\%$ и двух объясняющих переменных ($k = 2$) соответствующие статистики из таблицы П3 имеют значения $d_L = 1.35$, $d_U = 1.59$. Поскольку выполнено неравенство $d < d_L$, то можно говорить о наличии положительной автокорреляции в остатках. Отсюда следует, что модель с квадратичным трендом требует доработки, в которой могут быть учтены определенные (скрытые) особенности рассматриваемых данных.

2.3. МОДЕЛЬ ВЗАИМОСВЯЗИ ДВУХ ВРЕМЕННЫХ РЯДОВ

Рассмотрим два временных ряда $y(t),\ x(t)$ и будем изучать зависимость между ними. Особенность этой задачи состоит в том, что обе переменные являются случайными функциями времени t. Как функции от t они могут изменяться в очень согласованной форме. Поэтому их совместное рассмотрение и попытка построить зависимость вида

$$y(t) = b_0 + b_1\, x(t) + \varepsilon(t) \tag{69}$$

могут привести к неверным результатам и выводам. Отсюда следует, что зависимость (69) или ее аналоги напрямую применять нельзя. Один из способов проверки наличия зависимости между $y(t)$ и $x(t)$ состоит в том, чтобы устранить влияние основной переменной t. Для этого поступают следующим образом.

Будем считать, что $y(t)$ и $x(t)$ описываются трендовыми моделями (61), а именно:

$$y(t) = g_1(t) + \varepsilon_1(t), \tag{70}$$

$$x(t) = g_2(t) + \varepsilon_2(t). \tag{71}$$

Используя результаты раздела 2.2, найдем тренды $g_1(t),\ g_2(t)$ и введем новые переменные (устранение трендов):

$$x_1 = x(t) - g_2(t), \quad y = y(t) - g_1(t). \tag{72}$$

Теперь влияние переменной t в значительной степени устранено, и можно изучать линейную зависимость

$$y = b_0 + b_1\, x_1 + \varepsilon. \tag{73}$$

Для анализа зависимости (73), ее значимости или не значимости применяются результаты разделов 1.2.2 и 1.2.4.

В приложениях используются и другие подходы, позволяющие изучать зависимость между временными рядами $y(t)$ и $x(t)$. Например, в зависимости (69) явным образом учитывают время. При таком подходе вместо формулы (69) рассматривают ее обобщение

$$y(t) = b_0 + b_1\, x(t) + b_2\, t + \varepsilon(t). \tag{74}$$

Здесь время t выступает как еще одна объясняющая переменная.

Пример 5. Имеется набор данных, отражающих изменение во времени следующих переменных: $y = y(t)$ – объем реализации предприятия оптовой торговли (у.е.), $x = x(t)$ – размер торговой площади, t – время в кварталах. Требуется найти линейные тренды в динамике $x(t)$ и $y(t)$, установить или опровергнуть зависимость изменения объема реализации от размера торговой площади. Необходимые данные представлены в следующей таблице.

t	$x(t)$	$y(t)$	t	$x(t)$	$y(t)$	t	$x(t)$	$y(t)$
1	127.2	15.12	7	143.2	17.66	13	158.2	20.27
2	129.0	16.62	8	144.3	18.40	14	159.3	23.33
3	131.2	17.16	9	145.6	19.97	15	159.6	21.67
4	134.8	17.63	10	148.7	21.35	16	163.6	23.05
5	136.6	18.14	11	150.4	20.16	17	164.9	23.90
6	138.6	16.92	12	154.7	21.04	18	169.4	23.85

Будем считать, что выполнены все предположения МНК относительно случайных составляющих $\varepsilon_1(t),\ \varepsilon_2(t)$ и ε, входящих в соотношения (70), (71), (72). Уровень значимости выберем равным 5%. Для решения поставленной задачи поступим следующим образом. Предположим, что уравнения (70) и (71) содержат линейные тренды:

$$g_1(t) = a_0 + a_1\, t, \quad g_2(t) = c_0 + c_1\, t.$$

Вычисляя оценки параметров этих трендов, получаем, что $\bar{a}_0 = 15.16,\ \bar{a}_1 = 0.49,\ \bar{c}_0 = 124.4,\ \bar{c}_1 = 2.46$. Сформируем новые переменные x_1 и y по формуле (72) и построим выборки их значений. Например, значения $x_{15},\ y_5$ будут вычисляться так:

$$y_5 = y(5) - \hat{g}_1(5) = 18.14 - (15.16 + 0.49 \cdot 5) = 0.53,$$

$$x_{15} = x(5) - \hat{g}_2(5) = 136.6 - (124.4 + 2.46 \cdot 5) = -0.1.$$

В итоге будут сформированы выборки

$$x_{11},\, x_{12},\, \ldots, x_{1\,18};\quad y_1,\, y_2, \ldots, y_{18}.$$

Используя эти выборки, исследуем зависимость (73) для наших вспомогательных переменных x_1 и y. Проводя соответствующие вычисления, получаем, что оценки параметров в формуле (73) равны $\bar{b}_0 = -0.032$, $\bar{b}_1 = -0.178$. Для проверки значимости зависимости (73) по формуле (17) вычислим величину F. Она будет равна $F = 0.781$. Эту величину сравним с критическим значением F_α распределения Фишера с $f_1 = 1$ и $f_2 = 16$ степенями свободы на уровне значимости $\alpha = 0.05$ (см. таблицу П1). Критическое значение равно $F_{0.05} = 4.49$. Поскольку выполнено неравенство $F < F_{0.05}$, то зависимость (73) является не значимой. Поэтому считаем, что между переменными x_1 и y нет линейной зависимости. Отсюда делаем вывод о том, что отсутствует линейная связь между объемом реализации $y(t)$ и размером торговой площади $x(t)$. Обе переменные растут относительно согласованно, но на $y(t)$ влияют какие-то другие (не учтенные нами) факторы.

2.4. МОДЕЛЬ РАСПРЕДЕЛЕННЫХ ЛАГОВ

Модель распределенных лагов используется для построения регрессии между временными рядами $y(t)$ и $x(t)$. Особенность такой регрессии состоит в том, что зависимая переменная $y(t)$ учитывает не только текущие значения независимой переменной $x(t)$, но и ее предшествующие значения:

$$y(t) = \beta + \gamma_0 \cdot x(t) + \gamma_1 \cdot x(t-1) + \ldots + \gamma_k \cdot x(t-k) + \varepsilon(t). \tag{75}$$

Модель (75) применяется в тех случаях, когда на значения переменной $y(t)$ существенно влияют лаговые переменные $x(t-1)$, $x(t-2), \ldots, x(t-k)$. Принято, что случайная составляющая $\varepsilon(t)$, входящая в (75), удовлетворяет предположениям, указанным в разделе 2.2. Коэффициенты β, γ_0, $\gamma_1, \ldots, \gamma_k$ являются параметрами модели и должны вычисляться по набору данных (56). Модель (75) позволяет прогнозировать значения $y(t)$ на несколько шагов вперед на основе ранее полученных значений лаговых переменных. Пусть, например, по результатам обработки данных (56) установлено, что $y(t) \approx \beta + \gamma_2 \cdot x(t-2) + \gamma_3 \cdot x(t-3)$, и эта зависимость является значимой. Тогда при заданных $x(t) = x_2$, $x(t-1) = x_3$ можно найти приближенное значение $y(t+2) \approx \beta + \gamma_2 x_2 + \gamma_3 x_3$.

Работа с моделью (75) включает несколько этапов. Первый из них предполагает содержательный анализ данных, в том числе и графический. На этом этапе необходимо решить вопрос о длине лага (параметр k). Второй этап связан с оценками коэффициентов выбранной модели по набору данных (56). Третий этап предполагает интерпретацию полученных оценок коэффициентов модели. Пусть в модели (75) все найденные коэффициенты γ_i имеют одинаковые знаки. Тогда модель (75) допускает четкую интерпретацию, состоящую в том, что переменная $x(t)$ оказывает влияние на переменную $y(t)$ и это влияние распространяется на определенный период времени. Здесь также используются такие понятия как средний и медианный лаг. Коэффициенты модели интерпретируются как мультипликаторы.

Если коэффициенты γ_i будут иметь различные знаки, то возможно, что модель (75) или ее порядок выбраны неудачно либо взаимосвязь между $y(t)$ и $x(t)$ имеет достаточно сложный характер, который нужно дополнительно изучать.

Пример 6. Рассмотрим набор данных, который характеризует зависимость величины $y(t)$ – объемов продаж компании за месяц от расходов на рекламу $x(t)$ (млн руб.). Эти данные представлены в следующей таблице.

t	$x(t)$	$y(t)$	t	$x(t)$	$y(t)$	t	$x(t)$	$y(t)$
1	1.52	17.955	7	1.34	15.029	13	1.24	16.872
2	1.23	16.842	8	1.38	15.626	14	1.35	16.087
3	1.78	18.448	9	1.45	16.241	15	1.68	17.496
4	1.45	18.286	10	1.14	16.100	16	1.65	18.844
5	1.48	17.388	11	1.58	17.446	17	1.56	18.904
6	1.12	16.345	12	1.45	17.351	18	1.50	19.214

Графический анализ данных из этой таблицы показывает, что динамика переменных $y(t)$ и $x(t)$ имеет согласованный характер. Для построения связи между $y(t)$ и $x(t)$ можно использовать линейную регрессионную зависимость. Зафиксируем порядок модели $k = 3$ и рассмотрим зависимость

$$y(t) = \beta + \gamma_0 \cdot x(t) + \gamma_1 \cdot x(t-1) + \gamma_2 \cdot x(t-2) + \gamma_3 \cdot x(t-3) + \varepsilon(t). \tag{76}$$

Введем новые переменные

$$y = y(t), \;\; x_1 = x(t), \;\; x_2 = x(t-1), \;\; x_3 = x(t-2), \;\; x_4 = x(t-3), \;\; \varepsilon = \varepsilon(t). \tag{77}$$

Кроме того, зададим параметры:

$$b_0 = \beta, \quad b_1 = \gamma_0, \quad b_2 = \gamma_1, \quad b_3 = \gamma_2, \quad b_4 = \gamma_3. \tag{78}$$

С учетом обозначений (77) и (78), зависимость (76) будет представлена в форме

$$y = b_0 + b_1\, x_1 + b_2\, x_2 + b_3\, x_3 + b_4\, x_4 + \varepsilon. \tag{79}$$

Оценка параметров (78), проверка всех предположений МНК и значимости зависимости (79) выполняются по формулам, которые аналогичны представленным в разделе 1.2.3. Выборки значений переменных (77) задаются из исходной таблицы с учетом сдвижки данных, начиная с $t = 4$ и далее. В частности,

$$y_1 = y(4) = 18.286, \quad x_{11} = x(4) = 1.45, \quad x_{21} = x(3) = 1.78,$$

$$x_{31} = x(2) = 1.23, \quad x_{41} = x(1) = 1.52.$$

Соответствующие оценки параметров модели (76) таковы:

$$\bar{\gamma}_0 = 4.68,\ \bar{\gamma}_1 = 3.72,\ \bar{\gamma}_2 = 2.52,\ \bar{\gamma}_3 = 1.08,\ \beta = -0.02.$$

Отсюда видно, что влияние лаговых переменных на $y(t)$ уменьшается с ростом величины лага. Иначе говоря, расходы на рекламу в предшествующие месяцы $t-1,\ t-2,\ t-3$ оказывают все меньшее воздействие на объем продаж в текущем месяце t.

ПРИЛОЖЕНИЕ. СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ

Приложение содержит значения статистик для распределений Фишера, Стьюдента и Дарбина–Уотсона в наиболее распространенных случаях. Более детальные таблицы приведены в рекомендованной литературе (см. аннотацию).

Таблица П1. Критические значения F_α распределения Фишера с f_1 и f_2 степенями свободы

a) $\alpha = 0.05$

f_2	$f_1 = 1$	$f_1 = 2$	$f_1 = 3$	$f_1 = 4$	$f_1 = 5$
5	6.61	5.79	5.41	5.19	5.05
6	5.99	5.14	4.76	4.53	4.39
7	5.59	4.74	4.35	4.12	3.97
8	5.32	4.46	4.07	3.84	3.69
9	5.12	4.26	3.86	3.63	3.48
10	4.96	4.10	3.71	3.43	3.33
11	4.84	3.98	3.59	3.36	3.20
12	4.75	3.89	3.49	3.26	3.11
13	4.67	3.81	3.41	3.18	3.03
14	4.60	3.74	3.34	3.11	2.95
15	4.54	3.68	3.29	3.06	2.90
16	4.49	3.63	3.24	3.01	2.85
17	4.45	3.59	3.20	2.96	2.81
18	4.41	3.55	3.16	2.93	2.77
19	4.38	3.52	3.13	2.90	2.74
20	4.35	3.49	3.10	2.87	2.71

b) $\alpha = 0.01$

f_2	$f_1 = 1$	$f_1 = 2$	$f_1 = 3$	$f_1 = 4$	$f_1 = 5$
5	16.26	13.27	12.06	11.39	10.97
6	13.75	10.92	9.78	9.15	8.75
7	12.25	9.55	8.45	7.85	7.46
8	11.26	8.65	7.59	7.01	6.63
9	10.56	8.02	6.99	6.42	6.06
10	10.04	7.56	6.55	5.99	5.64
11	9.65	7.21	6.22	5.67	5.32
12	9.33	6.93	5.95	5.41	5.06
13	9.07	6.70	5.74	5.21	4.86
14	8.86	6.51	5.56	5.04	4.69
15	8.68	6.36	5.42	4.89	4.56
16	8.53	6.23	5.29	4.77	4.44
17	8.40	6.11	5.18	4.67	4.34
18	8.29	6.01	5.09	4.58	4.25
19	8.18	5.93	5.01	4.50	4.17
20	8.10	5.85	4.94	4.43	4.10

Таблица П2. Критические значения t_α распределения Стьюдента с f степенями свободы

f	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
14	1.761	2.145	2.977
15	1.753	2.131	2.947
16	1.746	2.120	2.921

Таблица П3. Значения статистик d_L, d_U для критерия Дарбина–Уотсона при $\alpha = 0.05$

а) $k = 1$ – одна объясняющая переменная

n	d_L	d_U	n	d_L	d_U	n	d_L	d_U
8	0.76	1.33	18	1.16	1.39	28	1.33	1.48
9	0.82	1.32	19	1.18	1.40	29	1.34	1.48
10	0.82	1.32	20	1.20	1.41	30	1.35	1.49
11	0.93	1.32	21	1.22	1.42	31	1.36	1.50
12	0.97	1.33	22	1.24	1.43	32	1.37	1.50
13	1.01	1.34	23	1.26	1.44	33	1.38	1.51
14	1.05	1.35	24	1.27	1.45	34	1.39	1.51
15	1.08	1.36	25	1.29	1.45	35	1.40	1.52
16	1.10	1.37	26	1.30	1.46	36	1.41	1.52
17	1.13	1.38	27	1.32	1.47	37	1.42	1.53

б) $k = 2$ – две объясняющие переменные

n	d_L	d_U	n	d_L	d_U	n	d_L	d_U
8	0.56	1.78	18	1.05	1.53	28	1.26	1.56
9	0.63	1.70	19	1.08	1.53	29	1.27	1.56
10	0.70	1.64	20	1.10	1.54	30	1.28	1.57
11	0.66	1.60	21	1.13	1.54	31	1.30	1.57
12	0.81	1.58	22	1.15	1.54	32	1.31	1.57
13	0.86	1.56	23	1.17	1.54	33	1.32	1.58
14	0.91	1.55	24	1.19	1.55	34	1.33	1.58
15	0.95	1.54	25	1.21	1.55	35	1.34	1.58
16	0.98	1.54	26	1.22	1.55	36	1.35	1.59
17	1.13	1.38	27	1.32	1.47	37	1.42	1.53