Е. С. Уланова, В. Н. Забелин

МЕТОДЫ КОРРЕЛЯЦИОННОГО И РЕГРЕССИОННОГО АНАЛИЗА В АГРОМЕТЕОРОЛОГИИ

ЛЕНИНГРАД ГИДРОМЕТЕОИЗДАТ 1990

УДК 630 : 551 + 551.509.314

Рецензент д-р физ.-мат. наук О. Д. Сиротенко

Первая часть книги содержит основы корреляционного и регрессионного анализа. Рассмотрено применение статистических методов для нахождения линейных и нелинейных связей. Даны примеры расчета различных уравнений регрессии из агрометеорологии.

Во второй части книги главное внимание уделено более сложным регрессионным моделям, развивающим классические методы регрессионного анализа. Такие модели, как гребневой, робастной регрессии и ряда других, быстро развиваемые в последние годы, позволяют расширить применение статистических методов в агрометеорологии. Даны практические примеры использования указанных моделей и рекомендации по статистическому моделированию.

Книга может быть полезна агрометеорологам, гидрометеорологам, географам, специалистам сельского хозяйства, а также студентам гидрометеорологических, сельскохозяйственных и других вузов.

The first part of "Correlation and Regression analysis in agrometeorology" by Ulanova E. S. and Zabelin V. N. is devoted to the basic ideas of correlation and regression analysis. It is illustrated by many simple and useful examples. In the second part the main attention is paied to advanced regression methods such as ridge, robust and others, elaborated in the last years. These methods expand the region of classic regression analyse in agrometeorology.

Authors also give practical recommendations on statistical modelling of weather-crop relations.

The "Correlation and Regression analysis in agrometeorology" is meant for specialists in agrometeorology and agriculture, as well as for the students of agrometeorological and agricultural institutes.

У $\frac{3702030000\text{-}026}{069\ (02)\text{-}90}$ 40-90

ISBN 5-286-00424-5

ПРЕДИСЛОВИЕ

Агрометеорологические исследования основываются чаще всего на обширных материалах агрометеорологических наблюдений и проводятся в основном с использованием методов математической статистики.

В результате этих исследований и статистической обработки массовых материалов сопряженных наблюдений за развитием и состоянием сельскохозяйственных культур, урожайностью и агрометеорологическими условиями был найден ряд важных закономерностей, на основе которых разработаны количественные методы, составляющие научный фонд агрометеорологии.

Использование ЭВМ значительно расширило и углубило возможности статистического моделирования во всех областях науки и техники. В настоящее время практически сняты ограничения на объем и сложность проводимых вычислений. Доступность ЭВМ, оснащенных пакетами статистических программ, позволяет при минимальной математической подготовке исследователя проводить обширные расчеты с помощью достаточно сложных статистических методов. В последние годы значительное развитие получили и методы регрессионного анализа.

В то же время следует отметить явно недостаточное количество отечественной специальной литературы по математической статистике в применении к агрометеорологии. В основном это книги В. М. Обухова [22], Е. С. Улановой [29, 30] и О. Д. Сиротенко [30], изданные более 20 лет назад. Поэтому публикация настоящей работы по методам корреляционного и регрессионного анализа в агрометеорологии является крайне необходимой.

Первая часть предлагаемой книги предназначена для специалистов, начинающих изучение математической статистики. В ней изложены основы корреляционного и регрессионного анализа и рассмотрено применение статистических методов для нахождения линейных и нелинейных связей. Даны расчеты уравнений регрессии с использованием малой вычислительной техники. Приведены практические примеры из агрометеорологии.

Во второй части книги описаны последние достижения науки в развитии методов регрессионного анализа: гребневого, робастного и др. Приведены наиболее известные методы автоматического выбора предикторов и описаны критерии их отбора. Уделено внимание улучшению построенной модели с помощью анализа регрессионных остатков (ошибок) и важному, но недостаточно широко освещенному вопросу идентификации резко выделяющихся данных (выбросов).

В помощь специалистам, не владеющим матричной алгеброй, в гл. 7 приведены некоторые сведения из этой области в объеме, достаточном для понимания изложенного материала.

Первая часть книги (гл. 1—6) написана Е. С. Улановой, вторая часть (гл. 7—12) — В. Н. Забелиным.

Часть I

ОСНОВЫ КОРРЕЛЯЦИОННОГО И РЕГРЕССИОННОГО АНАЛИЗА

Глава 1. РАЗЛИЧНЫЕ ТИПЫ СВЯЗЕЙ МЕЖДУ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ

1.1. ФУНКЦИОНАЛЬНЫЕ И СТАТИСТИЧЕСКИЕ СВЯЗИ. АРГУМЕНТ И ФУНКЦИЯ. ЗАДАЧИ ТЕОРИИ КОРРЕЛЯЦИИ

Явления повседневной жизни неразрывно связаны с числами и измерениями. При наличии большого числа наблюдений и измерений возникает необходимость свести первоначальную массу данных к небольшому числу показателей, к определенной системе и форме, так как никакой человеческий ум не способен вместить в себя большое количество разрозненных числовых данных. Перед исследователями встала задача в относительно небольшом числе сводных показателей отразить важную и существенную закономерность, содержащуюся в данной массе наблюдений и измерений. Был разработан особый род научного математического метода, который называется статистическим методом или математической статистикой.

Как наука математическая статистика является одним из разделов математики и ее можно рассматривать как математику, применяемую при обработке результатов массовых наблюдений. В математической статистике, как и во всей математике, одна и та же формула может одинаково относиться к самым различным объектам. Поэтому статистические методы применяются во многих областях знаний. Однако необходимо иметь в виду, что научная статистика должна базироваться на предварительном качественном анализе и не может использоваться в отрыве от реальной природы объекта исследования.

Одним из центральных разделов математической статистики является теория корреляции, которая изучает взаимосвязь, взаимозависимость между исследуемыми величинами (латинское слово correlato означает соотношение, взаимосвязь).

Изучением зависимостей между различными явлениями занимается любая наука, так как каждое явление в природе и обществе не возникает само по себе, а находится в связи с другими явлениями.

Диалектический подход к изучению природы и общества требует рассмотрения явлений в их взаимосвязи и непрерывном изменении. Теория корреляции позволяет выразить эти взаимосвязи в количественной форме.

Наиболее простым видом связи между величинами является функциональная, когда каждому значению одной величины соот-

ветствует одно и только одно вполне определенное значение другой.

К функциональным связям относятся, например, следующие:

связь между силой тока I, напряжением E и сопротивлением R:

$$I = E/R;$$

связь между давлением p, температурой T и объёмом V газа:

$$p = RT/V$$

(R — постоянный коэффициент);

связь между радиусом R длиной C окружности:

$$C = 2\pi R.$$

Функциональные связи между переменными величинами изучаются в специальном разделе математики — математическом анализе. Они характерны для количественных соотношений в области астрономии, механики, физики. В природе же чаще всего наблюдаются нефункциональные связи, когда переменная величина y изменяется главным образом в зависимости от другой переменной x, но на изменение первой влияет также множество других дополнительных факторов, учесть которые подчас невозможно, и тогда каждому значению x соответствует несколько значений y.

Такие связи (зависимости), когда численному значению одной величины x соответствует не одно, а несколько значений другой величины y, т. е. целая статистическая совокупность значений y, группирующихся лишь около некоторой средней величины $\bar{y}_x$, называются статистическими, или корреляционными.

Считают, что y находится в корреляционной зависимости от x, если, во-первых, каждому значению аргумента x соответствует ряд распределения функции y, во-вторых, с изменением x эти ряды y закономерно изменяют свое положение.

Часто в литературе встречается и такое определение корреляции: связь между переменными величинами x и y называется статистической, или корреляционной, если различным значениям одной из них (x) соответствуют определенные групповые средние значения другой ($\bar{y}_x$) или наоборот.

В таких случаях чаще всего одна величина рассматривается как независимая переменная, называется аргументом и обозначается буквой x; другая является зависимой переменной, называется функцией и обозначается буквой y. Например, если определяют связь урожайности сельскохозяйственных культур с количеством осадков, то ясно, что в данном случае независимой переменной — аргументом (x) — будут осадки, а зависимой функцией (y) — урожайность, а не наоборот.

Однако не всегда так просто бывает выбирать зависимую и независимую переменные, поскольку часто надо найти связь ме-

жду взаимовлияющими и взаимозависимыми явлениями. В этом случае одну величину условно принимают за аргумент x, а другую — искомую — за функцию y. Например, находя связь между запасами влаги в различных слоях почвы (0—20 и 20—50 см), запасы влаги в верхнем слое (0—20 см) можно условно принять за аргумент x, а запасы влаги в слое 20—50 см — за функцию y и рассчитать уравнение относительно y (см. раздел 2.8).

Главной задачей исследования статистических связей между различными переменными величинами является выяснение на основе большого числа наблюдений поведения функции y в зависимости от изменения главного своего аргумента x при условии, что другие аргументы не изменяются. Однако в природе такого условия быть не может. Эти другие аргументы также изменяются и тем самым в различной степени затушевывают и искажают интересующие нас зависимости. Поэтому, определяя зависимость одной величины y от другой, главной (x), необходимо знать, хотя бы в общем, степень влияния на других, дополнительных, изменяющихся, но неучтенных факторов. Если эта величина мала, то, зная x, можно достаточно точно определить y. Если же влияние дополнительных факторов велико, то y и x слабо связаны между собой, и с изменением x нельзя достаточно точно определить поведение y.

Определение степени влияния главных учитываемых факторов и дополнительных неучтенных является второй задачей теории корреляции.

Первая задача теории корреляции решается путем определения формы связи и нахождения уравнения этой связи двух или нескольких переменных величин. Вторая задача решается путем расчета различных показателей тесноты связи, которые дают оценку степени рассеяния значений y для разных значений x.

1.2. ОСНОВНЫЕ ВИДЫ ЛИНЕЙНЫХ И НЕЛИНЕЙНЫХ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ И ИХ УРАВНЕНИЯ

Общий вид уравнения корреляционной связи $y=f(x)$. В наиболее точном выражении $\bar{y}_{x_i}=f(x_i)$ или $\bar{y}_i=f(x_i)$. Здесь $f(x_i)$ представляет собой определенную однозначную функцию, дающую возможность по x_i находить приближенно соответствующие средние величины $\bar{y}_i$ ($i=1, 2, 3, \ldots, n$ — знак порядкового номера наблюдений, n — общее число наблюдений). Это уравнение называется уравнением регрессии y по x, где x — аргумент, а y — функция.

Можно также уравнение находить в отношении x, когда $\bar{x}_i$ представляет собой условную среднюю значений x, вычисляемую по условному распределению x, соответствующему y_i. Тогда $\bar{x}_i=f(y_i)$ является выражением корреляционной связи $\bar{x}_i$ с y_i

и называется уравнением регрессии x по y. В данном случае $\bar{x}_i$ приобретает значения функции, а y_i — аргумента.

Таким образом, для каждой статистической зависимости можно рассматривать два различных вида связи: y по x и x по y. По этим связям находят приближенные формулы, выражающие зависимость между значениями x_i и средними значениями $\bar{y}_i$ или наоборот. Такие формулы, полученные на основе статистического анализа экспериментальных данных, называются эмпирическими.

Существуют различные методы нахождения эмпирических формул; формулы дают результаты в разной степени приближенного характера. Оценка точности статистических (корреляционных) зависимостей и полученных эмпирических формул производится на основании корреляционного анализа.

Как уже отмечалось, первой задачей теории корреляции и корреляционного анализа является выбор формы связи переменных величин, состоящий в определении вида функции $\bar{y}_i = f(x_i)$.

Из встречающихся форм корреляционных связей наиболее распространены и изучены линейные. Однако наблюдаются и нелинейные связи. Не всегда задача выбора формы связи бывает легкой. При графическом изображении статистической связи часто точки располагаются так, что можно провести несколько линий различных типов. Например, прямая линия, гипербола или парабола могут совпадать на большей части графика. Поэтому при выборе формы связи (типа линии связи) необходимо прежде всего учитывать характерные особенности линии связи, вытекающие непосредственно из самой физической сущности изучаемого явления. Таким образом, выбору должен предшествовать логический анализ, обусловленный знанием общих закономерностей исследуемых явлений.

Для выбора формы статистической связи нужно хорошо знать различные типы линий и их уравнения.

Обычно в уравнениях переменные величины, между которыми ищется связь, обозначаются последними буквами латинского алфавита: x, y, z, u, v, а постоянные коэффициенты при этих переменных (параметры уравнения) обозначаются первыми буквами алфавита: a, b, c, d и т. д.

Между различными агрометеорологическими величинами чаще всего встречаются следующие типы линий и их уравнения (рис. 1.1).

1. Прямая, проходящая через начало координат (рис. 1.1 *а*). Описывается уравнением этой прямой $y = ax$. Здесь предполагается, что $a > 0$. Имеем прямую пропорциональную зависимость y от x, в которой необходимо определить один параметр a. Линии этого типа следует выбирать в том случае, когда по смыслу $y = 0$ при $x = 0$.

2. Прямая, не проходящая через начало координат. Описывается двумя уравнениями $y = ax + b$ (рис. 1.1 *б*) и $y = -ax + b$

(рис. 1.1 *в*). Имеем соответственно линейную прямую и линейную обратную зависимость y от x, в которой необходимо определить два параметра: a и b.

3. Парабола, симметричная одной из осей координат, с вершиной в начале координат (рис. 1.1 *г*). Описывается уравнением

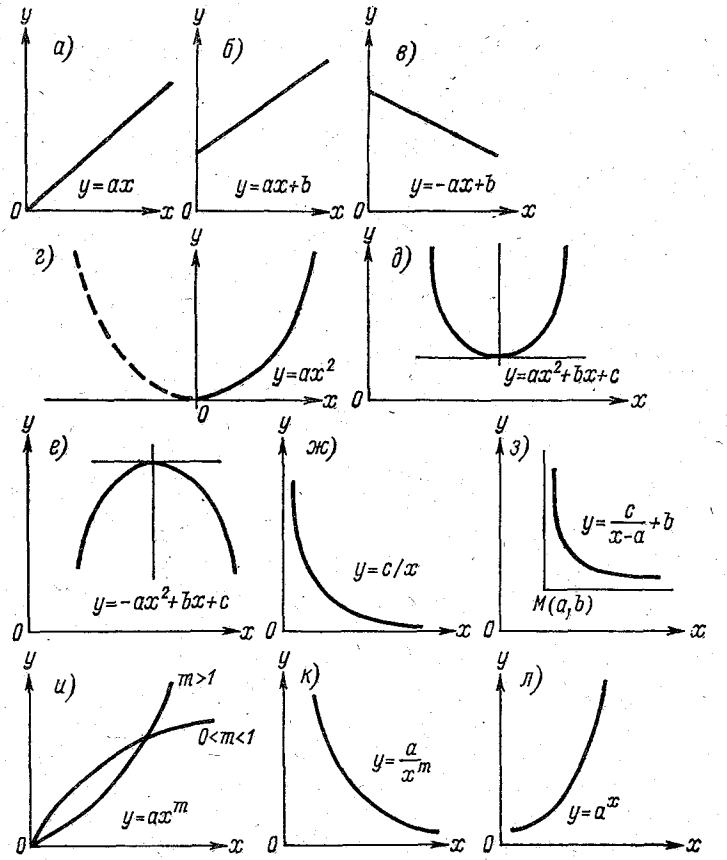


Рис. 1.1. Основные тины линий различных форм связи между переменными величинами и их уравнения.

$y=ax^2$. Такими параболами изображаются зависимости, где одна из величин, x или y, пропорциональна квадрату другой. Формула содержит один параметр a. При увеличении абсолютного значения этого параметра уменьшается «раствор» параболы.

4. Парабола, симметричная прямой, параллельной оси y. Описывается уравнением $y=ax^2+bx+c$. Функция квадратическая. Направление выпуклости параболы зависит от знака параметра a. При положительном параметре выпуклость параболы направлена вниз (рис. 1.1 *д*), при отрицательном — вверх (рис. 1.1 *е*). Линии

этого типа выбираются при наличии одного максимума или одного минимума, и кривые симметричны относительно их. В формуле необходимо определить три параметра: a, b и c.

5. Гипербола, асимптотически приближающаяся к осям координат (рис. 1.1 *ж*). Описывается уравнением $y=c/x$. Имеем обратно пропорциональную зависимость y от x, где необходимо определить один параметр c.

6. Гипербола, асимптотически приближающаяся к прямым, параллельным осям координат (рис. 1.1 *з*). Описывается уравнением $y=\frac{c}{x-a}+b$. Формула содержит три параметра. Параметры a и b являются координатами точки M. Знак параметра c зависит от расположения гиперболы в отношении асимптот.

7. Степенные кривые (рис. 1.1 *и, к*). Описываются уравнением $y=ax^m$, где m может быть целым или дробным. Частным случаем степенных функций являются параболы (рис. 1.1 *г*) и гиперболы (рис. 1.1 *ж, з*).

8. Показательная кривая, когда с возрастанием одной величины (x) значительно возрастает другая (y). Описывается уравнением $y=a^x$ (рис. 1.1 *л*).

После того как установлена форма связи, выбраны тип линии и вид общего уравнения связи, приступают к вычислению параметров уравнения данной связи и определению ее тесноты.

Глава 2. ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ ДВУХ ПЕРЕМЕННЫХ ВЕЛИЧИН

2.1. КОРРЕЛЯЦИОННОЕ ПОЛЕ. КОРРЕЛЯЦИОННАЯ ТАБЛИЦА. ЭМПИРИЧЕСКИЕ ЛИНИИ РЕГРЕССИИ

При исследовании различных явлений часто требуется установить связь между двумя переменными величинами. Как уже отмечалось, наиболее распространенной и хорошо изученной с помощью математической статистики формой связи является линейная.

Корреляция между случайными переменными величинами x и y называется линейной корреляцией, если функции регрессии $y=f_1(x)$ и $x=f_2(y)$ являются линейными. В этом случае при графическом изображении обе линии регрессии прямые. Они называются прямыми регрессии и выражаются следующими уравнениями:

линейное корреляционное уравнение регрессии y по x

$$y=ax+b; \tag{2.1}$$

линейное корреляционное уравнение регрессии x по y

$$x=ay+b. \tag{2.2}$$

Параметр a в (2.1) называется коэффициентом регрессии y по x и обозначается $\rho_{y/x}$. Аналогично ему параметр a_1 в (2.2) называется коэффициентом регрессии x по y и обозначается $\rho_{x/y}$.

Коэффициент регрессии a_1 одновременно является и угловым коэффициентом $k_{y/x}$ прямой регрессии y по x:

$$a_1 = k_{y/x} = \rho_{y/x}.$$

Коэффициент регрессии a не является угловым коэффициентом соответствующей прямой регрессии x по y. Угловым коэффициентом $k_{x/y}$ является величина, обратная коэффициенту регрессии:

$$k_{x/y} = \frac{1}{\rho_{y/x}}.$$

Расчету корреляционных уравнений, коэффициентов регрессии и показателей тесноты связи обычно предшествует первичный анализ, систематизация имеющегося материала наблюдений и его статистическая обработка.

Обширный материал данных наблюдений за двумя переменными величинами, связь между которыми надо определить, необходимо сначала проанализировать с точки зрения соответствия данных общим закономерностям изменения того или иного явления и его взаимосвязи с другими явлениями. После анализа и отбраковки ошибочных данных материал наблюдений представляется в виде таблицы-сводки, где указаны соответствующие друг другу пары значений x_i и y_i (табл. 2.1). Здесь x_i обозначает

Таблица 2.1

№ п/п	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
...	...	...
n	x_n	y_n
n	$\bar{x}$	$\bar{y}$

независимую переменную (аргумент), а y_i — зависимую переменную (функцию), i — любой порядковый номер x или y, от 1 до n, n — общее число пар наблюдений x и y.

Для определения линейности связи необходимо прежде всего построить график этой связи. Для этого данные каждой пары значений x и y в виде точки должны быть нанесены на график в прямоугольной системе координат. График кроме формы связи

позволяет увидеть и тесноту связи. Для разбора основных элементов теории корреляции в качестве примера проанализируем связь между запасами продуктивной влаги в различных слоях почвы осенью под озимой пшеницей. Эта связь была получена по данным фактических наблюдений гидрометеорологических станций за запасами влаги осенью на полях озимой пшеницы в южных районах Украины. [29].

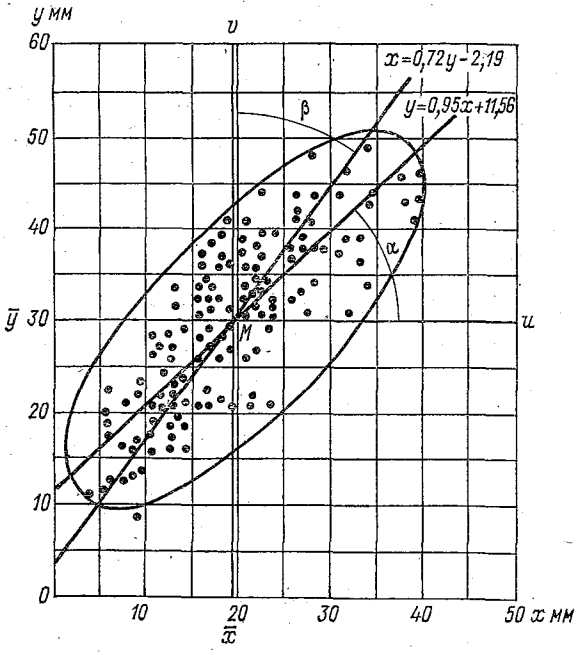


Рис. 2.1. Связь между запасами влаги в различных слоях почвы осенью под озимой пшеницей.

x — запасы влаги в слое 0—20 см, y — то же в слое 20—50 см.

Как известно, в начальный период развития и роста озимой пшеницы осенью для оценки и прогноза ее влагообеспеченности агрометеорологам важно знать запасы продуктивной влаги в верхних слоях почвы до полуметра. Так как в засушливых южных районах при продолжительной осени корневая система озимой пшеницы к моменту прекращения вегетации осенью может достигать глубины 20—40 см, то кроме запасов влаги в почве важно знать и их распределение по слоям.

В приведенном случае анализировались данные о запасах продуктивной влаги в верхнем 20-сантиметровом слое почвы и в слое от 20 до 50 см глубины (рис. 2.1).

За независимую переменную x условно были взяты запасы влаги в слое 0—20 см, а за зависимую переменную y — запасы

влаги в слое 20—50 см. Такой выбор аргумента и функции объясняется тем, что данные о запасах влаги в слое почвы 0—20 см или могут быть получены в результате наблюдений, или рассчитаны разными методами, в то время как для расчета запасов влаги в слое 20—50 см методов не было.

Следовательно, при анализе материала наблюдений в первую очередь важно определить, относительно какой величины искать уравнение. Ту величину, которая известна при расчетах и прогнозах, следует считать аргументом x, а неизвестную величину, которую надо рассчитать по искомому уравнению, — функцией y и искать уравнение в отношении y.

Получив 135 случаев пар наблюдений за запасами влаги в слоях почвы 0—20 и 20—50 см, записываем их сначала в виде простой таблицы-сводки, где под одним порядковым номером записываются пары значений x и y. Затем для выяснения линейности связи строим график, откладывая в прямоугольной системе координат по оси x значения запасов влаги в слое почвы 0—20 см, а по оси y — значения запасов влаги в слое почвы 20—50 см. Таким образом, получаем для каждого порядкового номера сводной таблицы на плоскости точки с координатами x_1y_1, x_2y_2, ..., $x_{135}y_{135}$. Это поле точек называется полем корреляции, или корреляционным полем yx (см. рис. 2.1).

Если число случаев пар наблюдений велико (больше 100), то корреляционное поле имеет вид более или менее правильного эллипса со сгущением точек в центре и сравнительно редким их расположением на периферии. Отклонение осей эллипса от координатных направлений указывает на наличие корреляции. Вытянутость же эллипса не всегда является ее объективным показателем, так как зависит от принятых масштабов по осям координат. По корреляционному полю можно судить о форме и тесноте связи.

На основе полученного графического поля корреляции при большом числе наблюдений продолжают дальнейшую систематизацию данных путем их группировки и построения корреляционной таблицы, или корреляционной решетки.

Корреляционная таблица строится по интервалам значений x и y, выбранным самим исследователем. Для этого на график, где изображено поле корреляции, наносят координатную сетку через точки, которые определяют границы выбранных интервалов значений для x и y (см. рис. 2.1). Таким образом все поле разбивается на вертикальные и горизонтальные столбцы, которые называются строями.

Вследствие пересечения строев плоскость корреляционного поля разбивается на прямоугольники или клетки. Подсчитав число точек в каждом прямоугольнике, который соответствует определенным значениям интервалов x и y, и записав эти данные в виде таблицы, получим корреляционную таблицу, или корреляционную решетку, которая удобна для проведения дальнейших операций по нахождению линий регрессий и их уравнений.

Таблица 2.2

Общий вид корреляционной таблицы, или решетки

Интервал Δy	Середина интервала Δy	Интервал Δx				m_y	$\bar{x}_{y_i}$
		Δx_1	Δx_2	$\cdots$	Δx_s		
		Середина интервала Δx					
		x_1	x_2	$\cdots$	x_s		
Δy_1	y_1	m_{11}	m_{21}	$\cdots$	m_{s1}	m_{y_1}	$\bar{x}_{y_1}$
Δy_2	y_2	m_{12}	m_{22}	$\cdots$	m_{s2}	m_{y_2}	$\bar{x}_{y_2}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Δy_k	y_k	m_{1k}	m_{2k}	$\cdots$	m_{sk}	m_{y_k}	$\bar{x}_{y_k}$
m_{x_i}		m_{x_1}	m_{x_2}	$\cdots$	m_{x_s}	$n = sk$	$\bar{x}$
$\bar{y}_{x_i}$		$\bar{y}_{x_1}$	$\bar{y}_{x_2}$	$\cdots$	$\bar{y}_{x_s}$	$\bar{y}$	

Корреляционную таблицу (табл. 2.2) можно строить непосредственно по первичной таблице-сводке, не прибегая к графическому построению корреляционного поля. В этом случае устанавливаются нужные интервалы для значений x и y и делается выборка данных по этим интервалам. В результате получают частоту (m_{xy}) сочетаний значений x и y определенных интервалов.

На пересечении каждого вертикального столбца и горизонтальной строки записывают частоту m_{xy}, показывающую, сколько раз при данном значении x встречались указанные значения y или наоборот.

В предпоследнюю строку и предпоследний столбец вписывают суммы частот по столбцам и строкам:

$$\sum m_x = \sum^{y} m_{xy}, \quad \sum m_y = \sum^{x} m_{xy}, \quad \sum m_x = \sum m_y = n,$$

(n — общее число наблюдений). Значки x и y над знаками сумм обозначают суммирование вдоль столбца или вдоль строки, т. е. $\sum^{y}$ обозначает суммирование частот y по интервалам при неизменном x, а $\sum^{x}$ — суммирование частот x по интервалам при неизменном y.

В последнюю строку и последний столбец вписывают взвешенные по частотам условные средние ($\bar{y}_x$ и $\bar{x}_y$) значений y и x по столбцам и строкам:

$$\bar{y}_x = \frac{\sum^{y} m_{xy} y}{m_x}, \quad \bar{x}_y = \frac{\sum^{x} m_{xy} x}{m_y}. \tag{2.3}$$

Суммы величин, стоящих в предпоследней строке и предпоследнем столбце, должны быть равны общему числу наблюдений n:

$$\sum^{x} m_x = n; \quad \sum^{y} m_y = n.$$

В предпоследние клетки последней строки и последнего столца вписывают подсчитанные общие средние взвешенные значения всех y и x, т. е. $\bar{y}$ и $\bar{x}$. Они могут быть вычислены как средние всех y или x, взвешенные по частотам m_{xy}, или как средние из $\bar{y}_x$ и $\bar{x}_y$, взвешенные по частотам m_x или m_y:

$$\bar{y} = \frac{\sum^{xy} m_{xy} y}{n} \text{ или } \bar{y} = \frac{\sum^{x} m_x \bar{y}_x}{n},$$

$$\bar{x} = \frac{\sum^{xy} m_{xy} x}{n} \text{ или } \bar{x} = \frac{\sum^{y} m_y \bar{x}_y}{n}. \qquad (2.4)$$

По корреляционной таблице, так же как и по полю корреляции, можно судить о форме связи и ее тесноте, так как мы по существу получаем корреляционную зависимость y от x в виде таблицы значений $\bar{y}_x$ для каждого значения x с указанием частот, а также корреляционную зависимость x от y в виде таблицы значений $\bar{x}_y$ для каждого значения y с указанием частот. Если частоты расположены вниз направо, то связь между величинами прямая, т. е. при увеличении x увеличивается y. Если же частоты расположены по диагонали вверх направо, то связь обратная, т. е. увеличением x уменьшается y.

По корреляционной таблице легко можно построить графическое поле корреляции, накладывая на координатную плоскость сетку, соответствующую интервалам таблицы, и изображая частоту каждой клетки в виде соответствующего числа точек, равномерно распределенных внутри клетки. Подобное поле корреляции, составленное на основе корреляционной таблицы по частотам интервалов, называется вторичным корреляционным полем (рис. 2.2).

Корреляционная таблица облегчает построение эмпирической и теоретической линий регрессии и расчет уравнений регрессии методом группировки данных, о чем будет сказано в разделе 2.8. Однако следует помнить, что корреляционная таблица строится обычно при большом числе пар наблюдений (больше 100).

При небольшом числе данных корреляционные таблицы по интервалам строить не рекомендуется. При обработке корреляционной таблицы считают, что число случаев в каждой клетке относится к серединам интервалов, а это при малом числе наблюдений может дать заметные ошибки.

Кроме того, интервалы не должны быть большими, иначе данные будут искажены и часть их потеряется, так как вместо фактических данных наблюдений при группировке берутся условные,

относящиеся к серединам интервалов. Доказано, что потеря информации, обусловленная группировкой, составляет менее 1 %, если интервал группировки не превосходит четвертую часть среднего квадратического отклонения σ (см. раздел 2.3). При надлежащем подборе интервала группировки ущерб в отношении точности невелик, к тому же значительно сокращаются затраты труда на обработку данных.

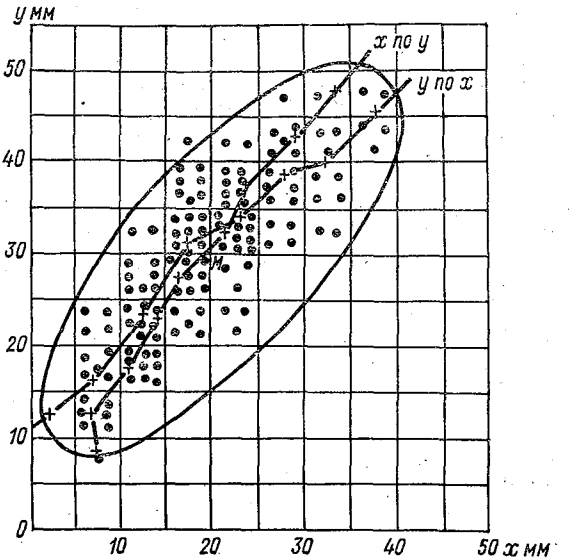


Рис. 2.2. Связь между запасами влаги в различных слоях почвы осенью под озимой пшеницей (вторичное корреляционное поле).

x — запасы влаги в слое 0—20 см, *y* — то же в слое 20—50 см.

В нашем примере, где ищется связь между запасами влаги в различных слоях почвы, число пар наблюдений x и y составляет 135 ($n = 135$), поэтому можно построить корреляционную таблицу (табл. 2.3). Это легко сделать на основе построенного поля корреляции (см. рис. 2.1).

Берем по осям x и y интервалы значений, равные 5 мм, и строим по ним сетку, разбивая все поле на клетки. Такие же интервалы берем для корреляционной таблицы, делая в ней такую же сетку, как и на графике, но с указанием значений середин интервалов по x и y. Затем подсчитываем число точек на графике в каждой клетке и вписываем это число в соответствующую клетку таблицы с теми же интервалами значений. Например, на графике в клетке при x от 10 до 15 мм и при y от 20 до 25 мм мы имеем 9 точек. В клетку таблицы для этих же интервалов x и y записываем цифру 9 и т. д. Получив частоты по столбцам и строкам, складываем их по вертикалям и горизонта-

Таблица 2.3

Пример составления таблицы корреляционной связи между запасами влаги в слоях почвы 0—20 и 20—50 см

Интервал Δy	Середина интервала Δy	Интервал Δx								m_y	$\bar{x}_y$
		0—5	5—10	10—15	15—20	20—25	25—30	30—35	35—40		
		Середина интервала Δx									
		2,5	7,5	12,5	17,5	22,5	27,5	32,5	37,5		
0—5	2,5										
5—10	7,5		1							1	7,5
10—15	12,5	1	6							7	6,8
15—20	17,5		5	9						14	10,7
20—25	22,5		4	9	5	3				21	14,1
25—30	27,5			7	9	3				19	16,5
30—35	32,5			2	9	12	4	2		29	21,6
35—40	37,5				7	8	6	4		25	23,9
40—45	42,5				1	2	5	3	3	14	29,3
45—50	47,5						1	2	2	5	33,5
m_x		1	16	27	31	28	16	11	5	135	$\bar{x}$=19
$\bar{y}_x$		12,5	16,2	22,9	30,9	33,0	38,4	39,8	44,5	$\bar{y}$=30	

лям. Например, сумма частот y при $x=7{,}5$ равна 16, а при $x=12{,}5$ сумма частот y равна 27 ($\sum^{y} m_{7,5}=16$, $\sum^{y} m_{12,5}=27$ и т. д.). Затем подсчитываем суммы частот x при различных значениях середин интервалов y. Например, $\sum^{x} m_{12,5}=7$, $\sum^{x} m_{17,5}=14$ и т. д.

Частные суммы частот x для середин различных интервалов y суммируем и получаем общую сумму частот $\sum m_y=135$. Так же суммируя все частоты y по серединам интервалов x, получаем $\sum m_x=135$. После этого приступаем к расчету средних взвешенных значений $\bar{y}_x$ и $\bar{x}_y$ по столбцам и строкам.

В табл. 2.3 имеем восемь столбцов с различной частной суммой частот y. Следовательно, необходимо найти восемь условных средних значений $\bar{y}_x$. Это достигается путем суммирования по вертикали произведений числа частот каждой клетки на соответствующее значение середины интервала y и деления этой суммы на m_x:

$$\bar{y}_{2,5}=12{,}5,$$

$$\bar{y}_{7,5}=\frac{7{,}5+6\cdot 12{,}5+5\cdot 17{,}5+4\cdot 22{,}5}{16}=16{,}2,$$

$$\bar{y}_{12,5}=\frac{9\cdot 17{,}5+9\cdot 22{,}5+7\cdot 27{,}5+2\cdot 32{,}5}{27}=22{,}9 \text{ и т. д.}$$

По горизонтали в табл. 2.3 надо рассчитать 10 условных средних значений $\bar{x}_y$, которые находим путем суммирования по горизонтали произведений числа частот каждой клетки на соответствующее значение середины интервала x и деления этой суммы на m_y. Начинаем со второй строки, так как в первой строке значений x нет:

$$\bar{x}_{7,5} = 7{,}5,$$

$$\bar{x}_{12,5} = \frac{2 \cdot 5 + 6 \cdot 7{,}5}{7} = 6{,}8,$$

$$\bar{x}_{17,5} = \frac{5 \cdot 7{,}5 + 9 \cdot 12{,}5}{14} = 10{,}7 \text{ и т. д.}$$

Получив условные средние взвешенные значения $\bar{y}_x$ и $\bar{x}_y$ по строям, находим общие средние взвешенные значения всех y и x:

$$\bar{y} = \frac{\sum^{x} m_x \bar{y}_x}{n} = \frac{12{,}5 + 16 \cdot 16{,}2 + 27 \cdot 22{,}9 + \ldots + 5 \cdot 44{,}5}{135} = 30{,}0,$$

$$\bar{x} = \frac{\sum^{y} m_y \bar{x}_y}{n} = \frac{7{,}5 + 7 \cdot 6{,}8 + 14 \cdot 10{,}7 + \ldots + 5 \cdot 33{,}5}{135} = 19{,}0$$

и записываем их в корреляционную таблицу.

Рассчитав условные средние значения $\bar{y}_x$ и $\bar{x}_y$, а также общие средние $\bar{y}$ и $\bar{x}$, можно приступить к построению эмпирических линий регрессии y по x и x по y.

По данным корреляционной таблицы построим вторичное поле корреляции (рис. 2.2), нанося в каждую клетку на графике число точек, соответственно числу частот в таблице и распределяя их равномерно по клетке, ограниченной выбранными интервалами. После этого наносим на график данные восьми условных средних значений $\bar{y}_x$ для середин интервалов x и общее среднее значение $\bar{y}$. Соединяя точки средних значений, получаем ломаную линию, которая называется эмпирической линией регрессии y по x. Наносим условные средние значения $\bar{x}_y$ для середин интервалов y, а также общее среднее значение $\bar{x}$. Соединяя точки значений всех средних, получаем эмпирическую линию регрессии x по y. Точка M на графике с координатами $\overline{xy}$ называется центром распределения.

Вторичное корреляционное поле строить не обязательно. Эмпирические линии регрессии можно построить и на первичном корреляционном поле, нанеся на него условные и общие средние значения $\bar{y}_x$, $\bar{x}_y$, $\bar{y}$, $\bar{x}$.

Если число случаев наблюдений мало, корреляционную таблицу не составляют, а эмпирические линии регрессии получают следующим образом. По таблице-сводке наблюдений строят график корреляционного поля. Оси X и Y разбивают на нужные интервалы и получают строи по x и по y. Находят условное среднее значение $\overline{y_x}$ для каждого интервала оси X и общее среднее значение всех $y/(\bar{y})$. Наносят значения $\bar{y}_x$ на график для середин интервалов x и соединяют точки ломаной линией, получая эмпирическую линию регрессии y по x. Определив по строям оси Y условные средние значения $\overline{x_y}$ и нанеся их на график для середин интервалов y, строят эмпирическую линию регрессии x по y.

Эмпирические линии регрессии получаются ломаными, поэтому проводят их сглаживание, или выравнивание, и получают плавные прямые линии регрессии, которые называются теоретическими линиями регрессии. Для наглядности они нанесены на рис. 2.1.

Теоретическую линию регрессии следует проводить после нахождения уравнения данной функции. Выравнивание по уравнению называется аналитическим выравниванием эмпирической линии регрессии.

Аппарат теории корреляции двух переменных величин для вычисления тесноты связи и уравнений связи дает в качестве суммарных характеристик следующие основные показатели:

1) средние арифметические значения $\overline{x}$ и $\overline{y}$;
2) средние квадратические отклонения σ_x и σ_y;
3) коэффициент корреляции r.

2.2. СРЕДНЯЯ АРИФМЕТИЧЕСКАЯ И ЕЕ СВОЙСТВА

Средняя арифметическая величина является простейшей и в то же время очень важной величиной, так как является первой сводной статистической характеристикой.

Средней арифметической величиной называется сумма значений признака (элемента), разделенная на число этих значений:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (2.5)$$

Такая средняя арифметическая называется простой. Если значениям x соответствуют различные частоты m, то средняя арифметическая зависит не только от значений x, но и от их частот m:

$$\bar{x} = \frac{m_1 x_1 + m_2 x_2 + \ldots + m_k x_k}{m_1 + m_2 + \ldots + m_k} = \frac{\sum_{i=1}^{k} m_i x_i}{\sum_{i=1}^{k} m_i}. \qquad (2.6)$$

Такая средняя называется средней взвешенной величиной, а частоты $m_1, m_2, \ldots, m_k$ — весами.

Значение признака (элемента), соответствующее каждой отдельной группе или интервалу, называется вариантом. Если x_1, x_2, x_3, ..., x_n — варианты, а m_1, m_2, ..., m_k — их частоты, то взвешенная средняя арифметическая равна сумме произведений вариантов на их веса (или частоты), разделенной на сумму весов.

Средние взвешенные величины мы находим при определении средних в корреляционной табл. 2.3. Если число наблюдений велико, то расчет средней арифметической получается громоздким. Его можно упростить, используя свойства средней арифметической величины.

Первое свойство. Если все значения $\bar{x}$ уменьшить на одно и то же число, то и средняя арифметическая уменьшится на это же число:

$$\overline{x-a}=\frac{\sum m_i x_i}{\sum m_i}-a=\bar{x}-a.$$

Число a может быть каким угодно, но лучше его брать из середины ряда значений x.

Второе свойство. Если все значения x разделить на одно и то же число, то средняя арифметическая тоже разделится на это же число:

$$\frac{x}{l}=\frac{1}{l}\,\frac{\sum m_i x_i}{\sum m_i}=\frac{\bar{x}}{l}.$$

Третье свойство. Средняя арифметическая сумма равна сумме средних арифметических, а средняя арифметическая разности равна разности средних арифметических:

$$\bar{x}=\frac{\sum (y+z)}{n}=\bar{y}+\bar{z};\quad \bar{x}=\overline{y-z}=\bar{y}-\bar{z}.$$

Четвертое свойство. Сумма отклонений от средней арифметической равна нулю:

$$\sum (y-\bar{x})=\overline{nx}-\overline{nx}=0.$$

Пятое свойство. Сумма квадратов отклонений от средней арифметической меньше, чем сумма квадратов отклонений от любого другого числа.

2.3. ДИСПЕРСИЯ И СРЕДНЕЕ КВАДРАТИЧЕСКОЕ ОТКЛОНЕНИЕ. ИХ СВОЙСТВА

При различном числе наблюдений важно знать не только средние величины, но и отклонения отдельных значений от средней. Отклонением называется разность между отдельным значением x_i

и средним значением $\bar{x}$. Отклонение положительно, когда x_i больше $\bar{x}$ и отрицательно, когда x_i меньше $\bar{x}$. Сумма всех положительных и отрицательных отклонений от средней арифметической, согласно четвертому свойству, будет равна нулю. Поэтому среднее отклонение нельзя использовать как характеристику рассеяния. Для этого вводят другой показатель — дисперсию (D или σ^2), которая является средней арифметической квадратов отклонений, устраняющей влияние знаков на результат.

Дисперсия характеризует рассеяние значений переменной величины около средней арифметической $\bar{x}$:

$$D = \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \text{ или } D = \sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 m_i}{\sum_{i=1}^{k} m_i}. \quad (2.7)$$

Таким образом, для вычисления дисперсии сначала все отклонения возводятся в квадрат, а потом рассчитывается средняя арифметическая квадратов отклонений.

Важной статистической характеристикой является среднее квадратическое отклонение. Средним квадратическим отклонением называется абсолютное значение корня квадратного из дисперсии $\sigma = \sqrt{D}$:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \text{ или}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 m_i}{\sum_{i=1}^{k} m_i}}. \quad (2.8)$$

В математической статистике среднее квадратическое отклонение часто называют стандартным отклонением или просто стандартом.

При нормальном распределении около $^2/_3$ всех отклонений значений величины от ее среднего арифметического не превышают по абсолютному значению среднее квадратическое отклонение. Среднее квадратическое отклонение σ_x выборочной средней x называют средней квадратической ошибкой или просто средней ошибкой (то же и для y).

Дисперсия и среднее квадратическое отклонение имеют следующие свойства.

1. Если все значения признака увеличить или уменьшить на одно и то же число a, то на то же число увеличится или уменьшится их средняя, отклонения же останутся без изменения. Сле-

довательно, останутся без изменения и среднее квадратическое отклонение и дисперсия.

2. Если все значения признака умножить на одно и то же число k, то в k раз увеличится и их средняя; следовательно, в k раз увеличатся отклонения. Квадраты же отклонений увеличатся в k^2 раз. Таким образом, среднее квадратическое отклонение возрастет в k раз, а дисперсия — в k^2 раз.

3. Если все значения признака одинаковы, то они совпадают со своей средней и отклонения равны нулю. Следовательно, и дисперсия и квадратическое отклонение равны нулю.

4. Средний квадрат отклонений значений признака от любой величины a больше дисперсии на квадрат отклонения этой величины a от среднего значения признака $\bar{x}$:

$$\frac{\sum (x+a)^2}{n} = \sigma^2 + (a - \bar{x})^2.$$

2.4. КОЭФФИЦИЕНТ ЛИНЕЙНОЙ КОРРЕЛЯЦИИ ДВУХ ПЕРЕМЕННЫХ ВЕЛИЧИН

Кроме средних величин $\bar{x}$ и $\bar{y}$ и средних квадратических отклонений σ_x и σ_y для нахождения уравнений регрессии и характеристики тесноты связи необходима еще одна величина, называемая коэффициентом линейной корреляции r.

На рис. 2.1 даны две прямые линии регрессии y по x и x по y. Направление этих прямых определяется коэффициентами регрессии: тангенсом угла между прямой регрессии y по x и осью u и тангенсом угла между прямой регрессии x по y и осью v. Обозначим эти углы соответственно α и β, тогда коэффициентами регрессии будут $\operatorname{tg}\alpha$ и $\operatorname{tg}\beta$. В уравнениях $y = ax + b$ и $x = ay + b$, $a = \operatorname{tg}\alpha$, $a = \operatorname{tg}\beta$.

Коэффициенты регрессии могут быть оба или положительными или отрицательными. В общем случае эти две прямые регрессии не совпадают. Они совпадут, если связь между y и x будет функциональной, т. е. угол φ между прямыми равен нулю.

По углу φ между прямыми регрессии можно судить о тесноте связи между y и x. Чем больше угол φ, тем слабее связь, и чем ближе угол φ к нулю, тем связь ближе к функциональной.

Для совпадающих прямых при $\varphi = 0$ имеем $\alpha = 90 - \beta$ и поэтому $\operatorname{tg}\alpha = \operatorname{tg}(90 - \beta) = \operatorname{ctg}\beta = 1/\operatorname{tg}\beta$. Отсюда $\operatorname{tg}\alpha \operatorname{tg}\beta = 1$.

Если связи между величинами нет, то y мало изменяется при изменении x, и наоборот. В этом случае α и β близки к нулю и в пределе $\operatorname{tg}\alpha \operatorname{tg}\beta = 0$.

Корень квадратный из числа $\operatorname{tg}\alpha \operatorname{tg}\beta$ принимают за критерий степени близости корреляционной связи к линейной функциональной зависимости и называют коэффициентом корреляции двух переменных величин x и y:

$$r = \sqrt{\operatorname{tg}\alpha \operatorname{tg}\beta}. \qquad (2.9)$$

В уравнениях прямых регрессии $y = ax + b$ и $x = ay + b$, где коэффициенты регрессии a и a_1 соответственно равны $\operatorname{tg}\alpha$ и $\operatorname{tg}\beta$, коэффициент корреляции равен

$$r = \sqrt{aa_1}, \tag{2.10}$$

т. е. коэффициентом линейной корреляции является среднее геометрическое значение коэффициентов регрессий.

При $r > 0$ обе прямые регрессии y по x и x по y проходят через центр распределения $M(\bar{x}, \bar{y})$ и образуют острые углы с положительными направлениями осей X и Y. В этом случае корреляция положительная, так как с увеличением одной величины возрастают соответственно условные средние другой.

При $r = 0$ в случае независимости величин x и y прямая регрессии y по x параллельна оси X, а прямая регрессии x по y параллельна оси Y. Следовательно, угол между этими прямыми равен 90°. С увеличением r наклон каждой из прямых к соответствующей оси координат возрастает, а угол между прямыми уменьшается.

При $r = 1$ обе прямые сливаются в одну, в этом случае зависимость будет функциональной.

При $r < 0$ прямые регрессии проходят через точку $M(\bar{x}, \bar{y})$ и образуют тупые углы с положительными направлениями осей координат. Угол φ между прямыми острый и всегда убывает с приближением r к единице.

При $r = -1$ обе прямые сливаются в одну, это случай обратной линейной функциональной зависимости одной величины от другой.

Таким образом, коэффициент корреляции является безразмерным и изменяется в пределах $-1 \leqslant r \leqslant 1$.

Кроме формулы (2.10), имеется еще ряд формул для расчета коэффициента корреляции r, где r выражается через средние величины $\bar{x}$, $\bar{y}$ и средние квадратические отклонения σ_x, σ_y.

Известно следующее варажение для расчета коэффициентов прямых регрессии:

$$a_{y/x} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad a_{1x/y} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{y^2} - (\bar{y})^2}. \tag{2.11}$$

Знаменатели обоих выражений обозначают соответствующие дисперсии:

$$\overline{x^2} - (\bar{x})^2 = D(x) = \sigma_x^2, \quad \overline{y^2} - (\bar{y})^2 = D(y) = \sigma_y^2, \tag{2.12}$$

откуда имеем простую формулу для r, выраженную через $\bar{x}$, $\bar{y}$, σ_x, σ_y:

$$r = \sqrt{aa_1} = \sqrt{\frac{(\overline{xy} - \bar{x}\bar{y})^2}{\sigma_x^2\sigma_y^2}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x\sigma_y}, \tag{2.13}$$

где r — линейный коэффициент корреляции между x и y; $\overline{xy} = \sum xy/n$; $\bar{x}$ и $\bar{y}$ соответственно средние арифметические признаков x и y, σ_x и σ_y — средние квадратические отклонения, найденные соответственно по признакам x и y.

Представив σ_x и σ_y в виде

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2}, \quad \sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} \tag{2.14}$$

и подставив их в (2.13), получим новую формулу для расчета r:

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{[\overline{x^2} - (\bar{x})^2]\,[\overline{y^2} - (\bar{y})^2]}}. \tag{2.15}$$

Из (2.15) видно, что между независимыми величинами корреляции не существует, так как для таких величин выполняется равенство $\overline{xy} = \bar{x}\bar{y}$.

Указанная выше формула для расчета r позволяет выразить каждый коэффициент регрессии через коэффициент корреляции.

В случае регрессии y по x

$$a = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x^2} = \frac{\sigma_y}{\sigma_x} \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} = \frac{\sigma_y}{\sigma_x} r; \tag{2.16}$$

в случае регрессии x по y

$$a_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_y^2} = \frac{\sigma_x}{\sigma_y} \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_y \sigma_x} = \frac{\sigma_x}{\sigma_y} r. \tag{2.17}$$

Таким образом, нахождение уравнений регрессии будет облегчено, если найдем значения r, σ_x и σ_y.

Для этого необходимо найти

$$\overline{xy} - \bar{x}\bar{y} = \frac{\sum xy}{n} - \bar{x}\bar{y}, \quad \sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{\frac{\sum x^2}{n} - (\bar{x})^2},$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{\frac{\sum y^2}{n} - (\bar{y})^2}.$$

Тогда

$$r = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{\sigma_x \sigma_y} \tag{2.18}$$

или

$$r = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{\sqrt{\frac{\sum x^2}{n} - (\bar{x})^2}\sqrt{\frac{\sum y^2}{n} - (\bar{y})^2}}. \tag{2.19}$$

Умножив числитель и знаменатель (2.19) на n^2, получим

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}. \qquad (2.20)$$

Несмотря на громоздкий вид, формула (2.20) удобна для расчетов, особенно при пользовании счетной машиной.

По приведенным формулам для расчета r вычисляют коэффициент корреляции для несгруппированных данных. Для нахождения коэффициента корреляции по формулам 2.19 и 2.20 дополнительно строят таблицу, аналогичную табл. 2.4, по которой удобно проводить расчеты.

При обработке статистических величин с помощью вычислительной машины табл. 2.4 перестраивают подобно табл. 2.4а.

Таблица 2.4

n_i	x	y	xy	x^2	y^2
$\sum n_i = n$	$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum y^2$

Таблица 2.4а

№ п/п	x	y
n	$\sum x$	$\sum y$
	$\sum x^2$	$\sum y^2$
	$\sum xy$	

По такому же типу при расчете на счетной машине перестраиваются и остальные таблицы, когда можно сразу получить различные необходимые степени величин и их суммы, а также произведения величин и сумму произведений.

При больших значениях x и y расчет r по приведенным выше формулам становится громоздким, так как используются произведение xy и их квадраты. Поэтому более удобно при несгруппированных данных использовать для расчета r формулу, где учитываются не сами величины, а их отклонения от средней, что позволяет проводить действия с меньшими числами, чем сами величины.

Проведем некоторые преобразования в (2.13):

$$r=\frac{\overline{xy}-\bar{x}\bar{y}}{\sigma_x\sigma_y}=\frac{\overline{xy}-\bar{x}\bar{y}}{\sqrt{\frac{\sum(x-\bar{x})^2}{n}\,\frac{\sum(y-\bar{y})^2}{n}}}=$$

$$=\frac{\overline{xy}-\bar{x}\bar{y}}{\frac{1}{n}\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}. \qquad (2.21)$$

Возьмем выражение $\frac{\sum(x-\bar{x})(y-\bar{y})}{n}$ и, раскрывая скобки под знаком суммы, преобразуем его:

$$\frac{\sum(xy-\overline{xy}-y\bar{x}+\bar{x}\bar{y})}{n}=\frac{\sum xy-\bar{y}\sum x-\bar{x}\sum y+\bar{x}\bar{y}n}{n}=$$

$$=\frac{\sum xy}{n}-\frac{\bar{y}\sum x}{n}-\frac{\bar{x}\sum y}{n}+\frac{\bar{x}\bar{y}n}{n}=\overline{xy}-\bar{y}\bar{x}-\bar{x}\bar{y}+$$

$$+\bar{x}\bar{y}=\overline{xy}-\bar{x}\bar{y}.$$

Заменяя числитель в (2.21) на выражение $\frac{\sum(x-\bar{x})(y-\bar{y})}{n}$, получим формулу для расчета r, где используются не сами величины, а их отклонения от средних:

$$r=\frac{\frac{\sum(x-\bar{x})(y-\bar{y})}{n}}{\frac{1}{n}\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}.$$

Обозначив отклонения $x-\bar{x}$ через Δx, а $y-\bar{y}$ через Δy, получим

$$r=\frac{\sum\Delta x\,\Delta y}{\sqrt{\sum\Delta x^2\sum\Delta y^2}}. \qquad (2.22)$$

Формула (2.22) для практического использования несгруппированных данных является наиболее удобной. Для расчета r по формулам (2.22) строят таблицу, аналогичную табл. 2.5.

Таблица 2.5

n	x	y	$\Delta x = x - \bar{x}$	$\Delta y = y - \bar{y}$	$\Delta x\,\Delta y$	Δx^2	Δy^2
n	$\sum x$	$\sum y$			$\sum \Delta x\,\Delta y$	$\sum \Delta x^2$	$\sum \Delta y^2$

Для расчетов r можно использовать формулу Пирсона

$$r = \frac{\sum \Delta x\,\Delta y}{n\sigma_x\sigma_y}. \qquad (2.23)$$

Выражение $\dfrac{\sum \Delta x\,\Delta y}{n} = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{n}$ называется ковариацией, или первым моментом произведений μ_{11}.

Понятие момента в математической статистике занимает важное место. Центральным моментом статистической величины называется сумма произведений тех или иных степеней отклонений x от $\bar{x}$ или y от $\bar{y}$ на соответствующую частоту (m), деленная на сумму всех частот:

$$\mu = \frac{1}{n}\sum m_i (x_i - \bar{x})^h.$$

Второй центральный момент (μ_2) является дисперсией, а первый центральный момент (μ_1) — средним отклонением от средней арифметической величины.

Формула коэффициента корреляции, выраженная через моменты, будет иметь следующий вид:

$$r = \frac{\mu_{11}}{\sigma_x\sigma_y}. \qquad (2.24)$$

Если данные сгруппированы в корреляционную таблицу по частотам m_{xy}, то r рассчитывается по следующим формулам

$$r = \frac{\dfrac{1}{n}\sum m_{xy}(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}. \qquad (2.25)$$

В числителе формулы (2.25) можно провести замену:

$$\frac{1}{n}\sum m_{xy}(x-\bar{x})(y-\bar{y})=\frac{1}{n}\sum m_{xy}(xy-\bar{x}y-\bar{y}x+\bar{x}\bar{y})=$$

$$=\frac{1}{n}\left[\sum m_{xy}xy-\bar{x}\sum m_{xy}(y-\bar{y})\sum m_{xy}x+\bar{x}\bar{y}\sum m_{xy}\right]=$$

$$=\frac{1}{n}\left[n\overline{xy}-n\bar{x}\bar{y}-n\bar{x}\bar{y}+n\bar{x}\bar{y}\right]=\overline{xy}-\bar{x}\bar{y}.$$

Заменив $\overline{xy}$ на $\frac{\sum m_x xy}{n}$, получим более удобную для расчета формулу

$$r=\frac{\frac{\sum m_x xy}{n}-\bar{x}\bar{y}}{\sqrt{\frac{\sum m_x x^2}{n}+(\bar{x})^2}\sqrt{\frac{\sum m_y y^2}{n}-(\bar{y})^2}}. \qquad (2.26)$$

Мы привели ряд встречающихся в литературе формул для вычисления коэффициента корреляции для несгруппированных и сгруппированных данных. Выбор той или иной формулы зависит от материала наблюдений и количества данных. Поэтому прежде чем приступить к расчетам, необходимо, учитывая техническую сторону расчетов, выбрать ту формулу для определения r, которая даст менее громоздкие расчеты.

2.5. СВОЙСТВА КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Первое свойство. Коэффициент корреляции не изменится, если из всех значений x и y вычесть какие-нибудь постоянные a и b и полученные результаты разделить на какие-то постоянные k и l (это свойство основано на свойствах среднего арифметического и дисперсии):

$$r=\frac{\sum m_{xy}\left[\left(\frac{x-a}{k}-\frac{\bar{x}-a}{k}\right)\left(\frac{y-b}{l}-\frac{\bar{y}-b}{l}\right)\right]}{n\frac{\sigma_x}{k}\frac{\sigma_y}{l}}=$$

$$=\frac{\sum m_{xy}\frac{x-\bar{x}}{k}\frac{y-\bar{y}}{l}}{n\frac{\sigma_x}{k}\frac{\sigma_y}{l}}=\frac{\sum m_{xy}(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}. \qquad (2.27)$$

Иначе говоря, коэффициент корреляции не изменится, если от первоначальных значений x и y перейти к новым условным значениям x' и y':

$$x'=\frac{x-a}{k},\quad y'=\frac{y-b}{l},\quad x=kx'+a,\quad y=ly'+b,$$

$$\bar{x}=k\bar{x}'+a,\quad \bar{y}=l\bar{y}'+b,\quad \sigma_x=k\sigma_{x'},\quad \sigma_y=l\sigma_{y'}.$$

Тогда

$$r_{y/x}=\frac{\sum m_{xy}\,(x-\bar{x})\,(y-\bar{y})}{n\sigma_x\sigma_y}=\frac{\sum m_{xy}k\,(x'-\overline{x'})\,l\,(y'-\overline{y'})}{nk\sigma_{x'}l\sigma_{y'}}, \quad (2.28)$$

так как $m_{xy}=m_{x'y'}$, то

$$r_{y/x}=\frac{\sum m_{x'y'}\,(x'-\overline{x'})\,(y'-\overline{y'})}{n\sigma_{x'}\sigma_{y'}}, \quad (2.29)$$

т. е.

$$r_{y/x}=r_{y'/x'}.$$

На этом свойстве основан упрощенный метод вычислений коэффициента корреляции — метод группировки, который приведен в разделе 2.8.

Данное свойство можно также сформулировать следующим образом: линейные преобразования, сводящиеся к изменению масштаба или начала отсчета переменных величин, не изменяют коэффициента корреляции между ними.

Второе свойство. Абсолютное значение коэффициента корреляции не превосходит единицы:

$$-1\leqslant r\leqslant 1.$$

Третье свойство. При линейной функциональной связи между величинами $r=\pm 1$. В этом случае прямые регрессии y по x и x по y совпадают.

Четвертое свойство. Чем ближе $|r|$ к единице, тем теснее прямолинейная корреляция между величинами x и y.

Пятое свойство. Если регрессия y по x линейная и коэффициент корреляции равен нулю, то все групповые средние $(\bar{y}_x)$ совпадают и равны общей средней $(\bar{y})$ переменной y. В этом случае между y и x линейной корреляции нет и прямые регрессии параллельны осям координат.

Однако, когда $r=0$, между y и x возможна нелинейная корреляция, не исключаются и нелинейные функциональные связи. Таким образом, оценивая связь только по одному коэффициенту корреляции r, надо быть осторожным. Если $r=0$, это еще не означает, что связи нет, связь может быть нелинейной, которая не учитывается коэффициентом корреляции.

Считают, что величины достаточно связаны, если $r>0{,}6$. Однако можно говорить о связи и при $r<0{,}6$ если связь можно объяснить физическими иричинами.

При физически обоснованной связи приведенные свойства коэффициента корреляции позволяют считать его достаточно надежной мерой тесноты связи в условиях линейной корреляции. Правильное же истолкование его при криволинейной корреляции представляет нелегкую задачу. Для измерения тесноты криволинейной связи пользуются корреляционным отношением. В то

время как корреляционное отношение не зависит от формы кривых регрессии, коэффициент корреляции r существенно зависит от того, в какой мере линии регрессии отличаются от прямых.

2.6. УРАВНЕНИЕ ЛИНЕЙНОЙ СВЯЗИ МЕЖДУ ДВУМЯ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ

После определения коэффициента корреляции r и установления линейной корреляционной связи находят параметры уравнений этой связи. Общий вид этих уравнений:

$$y = ax + b \quad \text{и} \quad x = a_1 y + b_1.$$

Из аналитической геометрии известно, что если прямые проходят через некоторую точку с координатами $\bar{x}$ и $\bar{y}$, то уравнения этих прямых регрессии имеют вид

$$y - \bar{y} = a_{y/x}(x - \bar{x}) \quad \text{и} \quad x - \bar{x} = a_{1x/y}(y - \bar{y}). \tag{2.30}$$

Как указывалось в начале этой главы, коэффициенты a_1 и a_2 называются коэффициентами линейной регрессии, а уравнения — уравнениями регрессии.

В предыдущем разделе была рассмотрена связь между коэффициентом корреляции r и коэффициентами регрессии a_1 и a_2.

Коэффициенты регрессии в уравнениях линейной связи двух переменных равны

$$a = r\frac{\sigma_y}{\sigma_x}; \qquad a_1 = r\frac{\sigma_x}{\sigma_y}.$$

Значение параметра b для уравнения $y = ax + b$ получаем из уравнения $y - \bar{y} = a(x - \bar{x})$ путем нахождения значения выражения $\bar{y} - a\bar{x}$.

Таким образом, вычислив средние арифметические значения $\bar{x}$ и $\bar{y}$, средние квадратические отклонения σ_x и σ_y и коэффициент корреляции r, путем решения уравнений связи вида

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad \text{и} \quad x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

можем вычислить параметры a, a_1, b, b_1 уравнений

$$y = ax + b \quad \text{и} \quad x = a_1 y + b_1.$$

По этим параметрам можно будет в зависимости от изменений одной величины получать наиболее вероятное значение другой.

Обычно при определении зависимости одной величины от другой находят только одно уравнение $y = ax + b$, однако если необходимо установить зависимость x от y, то рассчитывают и второе уравнение $x = a_1 y + b_1$.

2.7. СРЕДНЯЯ И ВЕРОЯТНАЯ ОШИБКИ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ. СРЕДНЯЯ ОШИБКА УРАВНЕНИЯ РЕГРЕССИИ

При исследовании корреляционной связи переменных величин даже при большом числе наблюдений мы имеем дело с выборочной совокупностью величин из генеральной совокупности p.

Генеральной совокупностью p можно считать все те наблюдения, которые теоретически можно было бы сделать, изучая взаимосвязь двух или нескольких явлений. Практически получить генеральную совокупность при бесконечном числе членов часто бывает невозможно из-за громоздкости расчетов. Обычно имеют дело с выборочной совокупностью, т. е. когда число наблюдений n значительно меньше, чем могло быть при генеральной совокупности.

Поэтому важно знать отличие выборочных коэффициентов корреляции и выборочных коэффициентов уравнений регрессии от аналогичных величин генеральной совокупности. При очень большой выборке значения этих величин могут мало различаться.

В математической статистике существует такое положение: с вероятностью, близкой к единице, можно утверждать, что при достаточно большом объеме выборки n выборочный коэффициент линейной корреляции $r_в$ мало отличается от аналогичного генерального коэффициента корреляции $r_г$.

Средняя квадратическая ошибка коэффициента корреляции σ_r при замене генерального коэффициента корреляции $r_г$ выборочным $r_в$ равна

$$\sigma_r = \pm \frac{1 - r^2}{\sqrt{n}} \tag{2.31}$$

и

$$\sigma_r = \pm \frac{1 - r^2}{\sqrt{n - 1}} \quad \text{(для малых и средних значений } r\text{)}. \tag{2.32}$$

С вероятностью 0,954 считают, что случайная ошибка не будет превышать $2\sigma_r$, т. е.

$$r_г = r_в \pm 2\sigma_r.$$

После расчета σ_r находят отношение r/σ_r. Если $r/\sigma_r > 3$ при числе наблюдений больше 50, то можно считать, что полученный выборочный коэффициент корреляции надежен и отображает искомую связь.

Величина $r - 3\sigma_r$ является гарантийным минимумом, а величина $r + 3\sigma_r$ — гарантийным максимумом коэффициента корреляции, т. е.

$$r \pm 3\sigma_r.$$

Кроме средней квадратической ошибки коэффициента корреляции σ_r можно вычислять вероятную ошибку коэффициента корреляции E_r, которая составляет $0{,}67\sigma_r$:

$$E_r = \pm 0{,}67 \frac{1-r^2}{\sqrt{n}}. \tag{2.33}$$

Вероятное значение коэффициента корреляции заключено в пределах $r \pm E_r$, а предельное значение близко к $r \pm 4E_r$. Если $r > 4E_r$, то связь доказана.

Если выборка мала (меньше 50), то рассчитывать средние ошибки по указанным формулам не рекомендуется, в таких случаях следует оценивать корреляцию с помощью критерия Фишера [32]

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}. \tag{2.34}$$

Выражая (2.34) через десятичные логарифмы, получим

$$z = 1{,}151 \lg \frac{1+r}{1-r}. \tag{2.35}$$

Функция z подчиняется закону нормального распределения. Рассчитав z, можно найти ее ошибку по формуле

$$\sigma_z = \frac{1}{\sqrt{n-3}}, \tag{2.36}$$

где n — объём выборки.

Кроме σ_z, вычисляют еще вероятное отклонение

$$p_z = 0{,}675 \frac{1}{\sqrt{n-3}} = 0{,}675\sigma_z. \tag{2.37}$$

Допустим, что коэффициент корреляции $r = 0{,}62$. Тогда, использовав формулы (2.35)—(2.37), можно рассчитать z, σ_z и p_z: $z = 0{,}725$, $\sigma_z = 0{,}123$, $p_z = 0{,}083$.

Определяем интервал изменения z:

$$z \pm \sigma_z = 0{,}725 \pm 0{,}123 = \begin{cases} 0{,}85 — \text{верхняя граница}, \\ 0{,}60 — \text{нижняя граница}. \end{cases}$$

По верхней и нижней границам z находим границы изменения r по табл. 2.6, откуда $r = 0{,}69$ — верхняя граница и $r = 0{,}54$ — нижняя граница, т. е. $r = \begin{cases} 0{,}69, \\ 0{,}54. \end{cases}$

При корреляционной связи, где каждому значению x_i соответствует ряд значений y_j, вычисленное по уравнению $\bar{y}_i$, будет отличаться от каждого значения y_i того ряда, которое соответствовало значению x_i при наблюдениях.

Зная значение x, подставляя его в уравнение, получаем y с ошибкой или с отклонением от данных наблюдений, т. е. по-

Значения коэффициента корреляции r в

z	0,01	0,02	0,03	0,04	0,05
0,0	0,0100	0,0200	0,0300	0,0400	0,0500
0,1	0,1096	0,1194	0,1293	0,1391	0,1489
0,2	0,2070	0,2165	0,2260	0,2355	0,2449
0,3	0,3004	0,3095	0,3185	0,3275	0,3364
0,4	0,3885	0,3969	0,4053	0,4136	0,4219
0,5	0,4699	0,4777	0,4854	0,4930	0,5005
0,6	0,5441	0,5511	0,5580	0,5649	0,5717
0,7	0,6107	0,6169	0,6231	0,6291	0,6351
0,8	0,6696	0,6751	0,6805	0,6858	0,6911
0,9	0,7211	0,7259	0,7306	0,7352	0,7398
1,0	0,7658	0,7699	0,7739	0,7779	0,7818
1,1	0,8041	0,8076	0,8110	0,8144	0,8178
1,2	0,8367	0,8397	0,8426	0,8455	0,8483
1,3	0,8643	0,8668	0,8692	0,8717	0,8741
1,4	0,8875	0,8896	0,8917	0,8937	0,8957
1,5	0,9069	0,9087	0,9104	0,9121	0,9138
1,6	0,9232	0,9246	0,9261	0,9275	0,9289
1,7	0,9366	0,9379	0,9391	0,9402	0,9414
1,8	0,94783	0,94884	0,94983	0,95080	0,95175
1,9	0,95709	0,95792	0,95873	0,95953	0,96032
2,0	0,96473	0,96541	0,96609	0,96675	0,96739
2,1	0,97103	0,97159	0,97215	0,97269	0,97323
2,2	0,97622	0,97668	0,97714	0,97759	0,97803
2,3	0,98049	0,98087	0,98124	0,98161	0,98197
2,4	0,98399	0,98431	0,98462	0,98492	0,98522
2,5	0,98688	0,98714	0,98739	0,98764	0,98788
2,6	0,98924	0,98945	0,98966	0,98987	0,99007
2,7	0,99118	0,99136	0,99153	0,99170	0,99186
2,8	0,99278	0,99292	0,99306	0,99320	0,99333
2,9	0,99408	0,99420	0,99431	0,99443	0,99454

Таблица 2.6

зависимости от значений функции Фишера z

0,06	0,07	0,08	0,09	0,10
0,0599	0,0699	0,0798	0,0898	0,0997
0,1586	0,1684	0,1781	0,1877	0,1974
0,2543	0,2636	0,2729	0,2821	0,2913
0,3452	0,3540	0,3627	0,3714	0,3800
0,4301	0,4382	0,4462	0,4542	0,4621
0,5080	0,5154	0,5227	0,5299	0,5370
0,5784	0,5850	0,5915	0,5980	0,6044
0,6411	0,6469	0,6527	0,6584	0,6640
0,6963	0,7014	0,7064	0,7114	0,7163
0,7443	0,7487	0,7531	0,7574	0,7616
0,7857	0,7895	0,7932	0,7969	0,8005
0,8210	0,8243	0,8275	0,8306	0,8337
0,8511	0,8538	0,8565	0,8591	0,8617
0,8764	0,8787	0,8810	0,8832	0,8854
0,8977	0,8996	0,9015	0,9033	0,9051
0,9154	0,9170	0,9186	0,9201	0,9217
0,9302	0,9316	0,9329	0,9341	0,9354
0,9425	0,9436	0,9447	0,9458	0,94681
0,95268	0,95359	0,95449	0,95537	0,95624
0,96102	0,96185	0,96259	0,96331	0,96403
0,96803	0,96865	0,96926	0,96986	0,97045
0,97375	0,97426	0,97477	0,97526	0,97574
0,97846	0,97888	0,97929	0,97970	0,98010
0,98233	0,98267	0,98301	0,98335	0,98367
0,98551	0,98579	0,98607	0,98635	0,98661
0,98812	0,98835	0,98858	0,98881	0,98903
0,99026	0,99045	0,99064	0,99083	0,99101
0,99202	0,99218	0,99233	0,99248	0,99263
0,99346	0,99359	0,99372	0,99384	0,99396
0,99464	0,99475	0,99485	0,99495	0,99505

лучаем, как говорилось ранее, среднюю величину $\bar{y}_i$ при заданном значении x_i. Нахождение этой ошибки позволит судить о том, насколько рассеяны точки корреляционного поля относительно линии прямой регрессии.

Средняя квадратическая ошибка уравнений дегрессии y по x и x по y может быть получена по соответствующим формулам:

$$S_y = \pm\sigma_y\sqrt{1-r^2}, \tag{2.38}$$

$$S_x = \pm\sigma_x\sqrt{1-r^2}. \tag{2.39}$$

Здесь σ_y и σ_x — средние квадратические отклонения от средних арифметических величин:

$$\sigma_y = \sqrt{\frac{\sum (y-\bar{y})^2}{n}}, \quad \sigma_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n}};$$

r — коэффициент корреляции между x и y.

При $r = \pm 1$ значения S_y и S_x равны нулю. В этом случае уравнения регрессии дают точные значения y по x и x по y, т. е. между величинами имеется линейная функциональная связь.

Максимальная ошибка уравнения регрессии в три раза больше средней ошибки:

$$S_{\text{макс}} = 3S_y. \tag{2.40}$$

Средние квадратические ошибки выборочных коэффициентов регрессии a_1 и a_2 выражаются следующими формулами:

для уравнения $y = ax + b$

$$\sigma_a = \pm\frac{\sigma_y}{\sigma_x}\sqrt{\frac{1-r^2}{n}}, \tag{2.41}$$

для уравнения $y = a_1x + b_1$

$$\sigma_{a_1} = \pm\frac{\sigma_x}{\sigma_y}\sqrt{\frac{1-r^2}{n}}. \tag{2.42}$$

Таким образом, чем больше число наблюдений n и коэффициент корреляции r, тем меньше ошибки коэффициента корреляции, коэффициентов регрессии и уравнения регрессии.

2.8. ПРИМЕР РАСЧЕТА УРАВНЕНИЯ ЛИНЕЙНОЙ СВЯЗИ МЕЖДУ ДВУМЯ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ ПО СГРУППИРОВАННЫМ ДАННЫМ (СВЯЗЬ МЕЖДУ ЗАПАСАМИ ПРОДУКТИВНОЙ ВЛАГИ В РАЗЛИЧНЫХ СЛОЯХ ПОЧВЫ)

Когда число случаев пар наблюдений велико (больше 100), нахождение уравнений связи и вычисление коэффициентов корреляции для уменьшения громоздкости расчетов проводят способом группировки данных. При этом учитывают веса или частоты данных величин по корреляционной таблице.

В разделе 2.1 приведены примеры построения корреляционного поля и корреляционной таблицы для расчета уравнения связи между запасами продуктивной влаги в различных слоях почвы осенью под озимой пшеницей.

Таблица 2.3 нам необходима была для построения эмпирических линий регрессии (рис. 2.2). Для нахождения уравнения данной связи и теоретической линии регрессии $y = ax + b$ способом группировки данных по частотам используем этот же пример, но несколько видоизменим корреляционную таблицу, дополнив ее графами, необходимыми для расчета коэффициента корреляции и уравнения регрессии (табл. 2.7).

Расчеты проводят следующим образом.

В данном примере мы имеем 135 случаев парных наблюдений за запасами влаги в слоях почвы 0—20 и 20—50 см. Обозначаем через x запасы влаги в слое почвы 0—20 см, а через y — запасы влаги в слое почвы 20—50 см. Разбиваем ряд значений x и y на интервалы по 5 мм (интервалы берутся произвольно и могут быть разными для x и для y), записываем средние значения этих интервалов в графах таблицы. Затем подсчитываем число случаев парных наблюдений за запасами влаги, попадающих в соответствующие интервалы, т. е. находим частоту повторения y по данному интервалу x и наоборот. Делать это проще с помощью графика корреляционного поля, как было изложено в разделе 2.1.

Вертикальные столбцы дадут нам число частот (m_x) различных значений y по каждому данному интервалу x, а горизонтальные столбцы — число частот (m_y) различных значений x по каждому интервалу y. Более подробно о построении этой части корреляционной таблицы написано в разделе 2.1.

Согласно первому свойству, коэффициент линейной корреляции r остается неизменным, если от первоначальных значений x и y перейти к условным значениям x' и y'.

Это значительно упрощает расчеты.

Обозначим средние горизонтальные и вертикальные столбцы через 0, т. е. примем их за начальные. Следующие от них столбцы вправо и влево (для вертикальных) и вверх и вниз (для горизонтальных) по мере возрастания значений x или y будем обозначать 1, 2 и т. д, а по мере убывания —1, —2 и т. д. Таким образом, за условную единицу принимаем один интервал значений x или y, равный пяти действительным единицам.

В данном случае за начальные столбцы для x и y примем средние столбцы с интервалами 20—25 мм и обозначим их через 0. Соседние столбцы, где значения x и y возрастают, обозначим через 1, 2 и т. д., а столбцы, где значения убывают,— через —1, —2 и т. д.

Таким образом, за начало отсчета мы взяли $x_0 = 22{,}5$ мм и $y_0 = 22{,}5$ мм, обозначив их через $x' = 0$ и $y' = 0$.

Присвоив горизонтальным и вертикальным столбцам условные значения, переходим к дальнейшим расчетам корреляционной

Таблица 2.7

Корреляционная таблица для расчета уравнения связи между запасами влаги в различных слоях почвы по сгруппированным данным

Интервал Δy	Середина интервала Δy	Условные единицы y'	Интервал Δx								m_y	$m_y x'_y$	$m_y x'_y y'$	$m_y y'$	y'^2	$m_y y'^2$
			0—5	5—10	10—15	15—20	20—25	25—30	30—35	35—40						
			Середина интервала Δx													
			2,5	7,5	12,5	17,5	22,5	27,5	32,5	37,5						
			Условные единицы x'													
			—4	—3	—2	—1	0	1	2	3						
0—5	2,5	—4														
5—10	7,5	—3		1							1	—3	9	—3	9	9
10—15	12,5	—2	1	6							7	—22	44	—14	4	28
15—20	17,5	—1		5	9						14	—33	33	—14	1	14
20—25	22,5	0		4	9	5	3				21	—35	0	0	0	0
25—30	27,5	1			7	9	3				19	—23	—23	19	1	19
30—35	32,5	2			2	9	12	4	2		29	—5	—10	58	4	116
35—40	37,5	3				7	8	6	4		25	7	21	75	9	225
40—45	42,5	4				1	2	5	3	3	14	19	76	56	16	224
45—50	47,5	5						1	2	2	5	11	55	25	25	125
m_x			1	16	27	31	28	16	11	5	$n=135$	$\Sigma=-84$	$\Sigma=205$	$\Sigma=202$	$\Sigma=69$	$\Sigma=760$
$\Sigma m_x y'_x$			—2	—20	2	52	59	51	38	22	$\Sigma=202$					
$m_x y'_x x'$			8	60	—4	—52	0	51	76	66	$\Sigma=205$					
$m_x x'$			—4	—48	—54	—31	0	16	22	15	$\Sigma=-84$					
x'^2			16	9	4	1	0	1	4	9	$\Sigma=44$					
$m_x x'^2$			16	144	108	31	0	16	44	45	$\Sigma=404$					

таблицы, необходимым для нахождения r и уравнения связи $y = ax + b$.

Находим суммы частот m_y и m_x, записанных соответственно в горизонтальных и вертикальных столбцах; m_y обозначает сумму частот различных значений x для данного постоянного интервала y, а m_x — сумму частот различных значений y для данного постоянного интервала x. Сумма всех значений m_x, так же как и сумма всех значений m_y должна быть равна общему числу случаев n. В нашем примере $\sum m_x = \sum m_y = n = 135$.

Далее вычисляем $\sum m_x y'_x$ и $\sum m_y x'_y$, беря y' и x' в условных единицах. Значение $\sum m_y x'_y$ для каждого горизонтального столбца рассчитывается как алгебраическая сумма из произведений каждого числа частот m на соответствующее ему значение x' в условных единицах. Например, в первом горизонтальном столбце с частотами (m) имеется одно значение частоты, равное 1. Умножаем его на x', равное —3, и записываем полученное значение —3 в первую строку графы $\sum m_y x'_y$. Во втором горизонтальном столбце с частотами значение $m = 1$ умножаем на $x' = -4$, получаем —4. Затем второе значение частоты в этом же горизонтальном столбце $m = 6$ умножаем на $x' = -3$, получаем —18. Складываем —4 и —18. Сумму, равную —22, записываем во вторую строку графы $\sum m_y x'_y$. В третьем горизонтальном столбце частоту $m = 5$ умножаем на $x' = -3$, получаем —15; затем $m = 9$ умножаем на $x' = -2$, получаем —18. Складываем —15 и —18, получаем сумму —33, которую записываем в третью строку графы $\sum m_y x'_x$ и т. д.

Значение $\sum m_x y'_x$ для каждого вертикального столбца рассчитывается как алгебраическая сумма (с учетом знака) из произведений каждого числа частот m на соответствующее ему значение y' в условных единицах. Например, в первом вертикальном столбце частоту $m = 1$ умножаем на $y' = -2$, получаем —2 и записываем в графу $\sum m_x y'_x$ данного столбца. Во втором вертикальном столбце умножаем $1 \cdot (-3) = -3$; $6 \cdot (-2) = -12$; $5 \cdot (-1) = -5$ и $4 \cdot 0 = 0$. Сложив эти произведения, получаем сумму, равную —20, и записываем ее в графу $\sum m_x y'_x$ второго вертикального столбца и т. д. Затем находим общую алгебраическую сумму всего вертикального столбца $\sum m_y x'_y = -84$ и общую сумму всего горизонтального столбца $\sum m_x y'_x = 202$.

Далее подсчитываем вертикальные и горизонтальные столбцы значений $m_y x'^2_y$ и $m_x y'^2_x$. Они вычисляются путем умножения соответствующих значений $m_y x'_y$ (или $m_x y'_x$), полученных в предыдущем столбце, на соответствующие им условные значения y' (или x'). Например, в первой строке значение $m_y x'^2_y = -3$,

умножив его на $y' = -3$, получаем 9 и заносим эту цифру в первую строку графы $m_y x'_y y'$.

Во второй строке —22 умножаем на —2, получаем 44 и т. д. В конце всего вертикального столбца подсчитываем алгебраическую сумму: $\sum m_y x'_y y' = 205$.

Рассчитываем графу $m_x y'_x x'$, умножая значения $m_x y'_x$ на соответствующие им условные значения x'. Получаем числа $-2 \times (-4) = 8$; $-20 \cdot (-3) = 60$; $2 \cdot (-2) = -4$ и т. д. В конце этой горизонтальной графы подсчитываем алгебраическую сумму: $\sum m_x y'_x x' = 205$.

Сумма $\sum m_y x'_y y'$ должна быть равна $\sum m_x y'_x x'$. В нашем примере $\sum m_y x'_y y' = \sum m_x y'_x x' = 205$. В этом состоит проверка правильности расчетов данной корреляционной таблицы.

Коэффициент корреляции r данной связи можно рассчитать по формуле

$$r = \frac{\frac{\sum mx'y'}{n} - \overline{x'}\,\overline{y'}}{\sigma_{x'}\sigma_{y'}}.$$

Зная, что $\sum mx'y' = 205$, находим выражение

$$\frac{\sum mx'y'}{n} = \frac{205}{135} = 1{,}52.$$

Рассчитываем другие члены, входящие в формулу коэффициента корреляции: $\overline{x'}$, $\overline{y'}$, $\sigma_{x'}$ и $\sigma_{y'}$.

Средние арифметические величины $\overline{x'}$ и $\overline{y'}$ равны

$$\overline{x'} = \frac{\sum m_x x'}{n} = \frac{-84}{135} = -0{,}622, \quad \overline{y'} = \frac{\sum m_y y'}{n} = \frac{202}{135} = 1{,}496.$$

Средние квадратические отклонения $\sigma_{x'}$, $\sigma_{y'}$ находим по формулам

$$\sigma_{x'} = \sqrt{\frac{\sum m_x x'^2}{n} - (\overline{x'})^2}, \quad \sigma_{y'} = \sqrt{\frac{\sum m_y y'^2}{n} - (\overline{y'})^2}.$$

Для нахождения $\sum m_x x'^2$ берем квадраты значений x' каждого столбца и умножаем на m_x, затем суммируем эти значения. В нашем примере

$$\sum m_x x'^2 = 404, \quad \text{а} \quad \frac{\sum m_x x'^2}{n} = \frac{404}{135} = 2{,}993.$$

Согласно выполненным расчетам, $\overline{x'} = -0{,}622$. Возводим $\overline{x'}$ в квадрат: $(\overline{x'})^2 = 0{,}387$. Находим

$$\sigma_{x'} = \sqrt{\frac{\sum m_x x'^2}{n} - (\overline{x'})^2} = \sqrt{2{,}993 - 0{,}387} = \sqrt{2{,}606} = 1{,}61.$$

Подобным образом находим $\sigma_{y'}$. Берем квадраты значений y' каждого горизонтального столбца, умножаем на число равных между собой y' этого столбца, т. е. на m_y, и получаем значения $m_y y'^2$ каждого столбца. Складывая их, получаем $\sum m_y y'^2 = 760$. Отсюда находим

$$\frac{\sum m_y y'^2}{n} = \frac{760}{135} = 5{,}630.$$

По предыдущим вычислениям

$$\overline{y'} = 1{,}496, \quad (\overline{y'})^2 = 2{,}238.$$

Находим

$$\sigma_{y'} = \sqrt{\frac{\sum m_y y'^2}{n} - (\overline{y'})^2} = \sqrt{5{,}630 - 2{,}238} = \sqrt{3{,}392} = 1{,}84.$$

Таким образом, найдены все величины, необходимые для вычисления коэффициента корреляции r:

$$r_{x'/y'} = \frac{\frac{\sum m x' y'}{n} - \overline{x'}\,\overline{y'}}{\sigma_{x'}\sigma_{y'}} = \frac{1{,}52 - (-0{,}622)\,(1{,}496)}{1{,}61 \cdot 1{,}84} = 0{,}83.$$

Средняя ошибка коэффициента корреляции равна

$$\sigma_r = \pm\frac{1 - r^2}{\sqrt{n}} = \pm\frac{1 - (0{,}83)^2}{\sqrt{135}} = \pm 0{,}03;$$

тогда наиболее вероятные значения r будут находиться в интервале:

$$r \pm \sigma_r = 0{,}83 \pm 0{,}03 = \begin{cases} 0{,}86, \\ 0{,}80. \end{cases}$$

Максимальное и минимальное значения r лежат в пределах $r \pm 3\sigma_r$; $3\sigma_r = 0{,}09$:

$$r \pm 3\sigma_r = 0{,}83 \pm 0{,}09 = \begin{cases} 0{,}92, \\ 0{,}74. \end{cases}$$

Вероятная ошибка коэффициента корреляции r

$$E_r = \pm 0{,}67\sigma_r = \pm 0{,}67 \cdot 0{,}03 = \pm 0{,}02.$$

Предельные значения для r могут быть также выражены через вероятную ошибку. Они равны

$$r \pm 4E_r = 0{,}83 \pm 0{,}08 = \begin{cases} 0{,}91, \\ 0{,}75. \end{cases}$$

Обычно находится одна из ошибок или σ_r, или E_r, так как предельные значения коэффициента корреляции r, вычисленные по $\sigma_r(r \pm 3\sigma_r)$ и по $E_r(r \pm 4E_r)$, близки между собой.

Мы получили коэффициент корреляции $r = 0{,}83$ для условных единиц x' и y'.

Исходя из первого свойства коэффициента корреляции (2.5), его можно принять и для действительных значений x и y.

Рассчитаем уравнение линейной регрессии связи двух переменных величин по формуле

$$y - \bar{y} = R(x - \bar{x}),$$

где $R = r\dfrac{\sigma_y}{\sigma_x}$,

или по формуле

$$y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}).$$

Средние квадратические отклонения σ_x и σ_y у нас выражены в условных единицах. Переводим их в действительные. За условную единицу мы принимали пять действительных единиц, следовательно, для того чтобы получить σ_x и σ_y в действительных единицах их значения в условных единицах надо умножить на пять. В результате этого получаем:

$$\sigma_y = \sigma_{y'} \cdot 5 = 1{,}84 \cdot 5 = 9{,}2, \quad \sigma_x = \sigma_{x'} \cdot 5 = 1{,}61 \cdot 5 = 8{,}05.$$

Среднее арифметическое значение $\bar{x}$ в действительных единицах вычисляем по формуле

$$\bar{x} = x_0 + \overline{x'}d,$$

где x_0 — значение начального отсчета $\bar{x}$ при $x' = 0$; σ — количество действительных единиц, взятых за одну условную.

В нашем примере $x_0 = 22{,}5$, $d = 5$. Отсюда

$$\bar{x} = 22{,}5 + (-0{,}622) \cdot 5 = 19{,}39.$$

Подобным образом вычисляем $\bar{y}$ в действительных единицах

$$\bar{y} = \bar{y}_0 + \overline{y'}d.$$

В нашем примере $y_0 = 22{,}5$, $d = 5$. Отсюда

$$\bar{y} = 22{,}5 + 1{,}496 \cdot 5 = 22{,}5 + 7{,}48 = 29{,}98.$$

Находим уравнение регрессии

$$y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) = 29{,}98 + 0{,}83\,\frac{9{,}20}{8{,}05}(x - 19{,}39) =$$

$$= 29{,}98 + 0{,}95\,(x - 19{,}39).$$

Таким образом,

$$y = 0{,}95x + 11{,}56.$$

Мы получили уравнение связи между запасами влаги в различных слоях почвы [25].

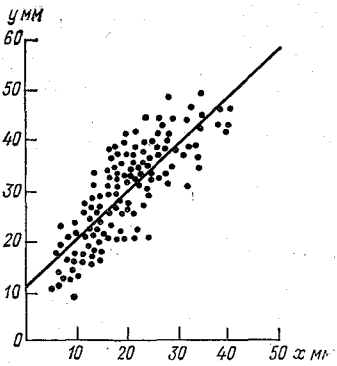


Рис. 2.3. Связь между запасами влаги в различных слоях почвы осенью под озимой пшеницей.

x — запасы влаги в слое 0—20 см, y — то же в слое 20—50 см.

Зная запасы влаги в верхнем (0—20 см) слое почвы (x), можно рассчитать запасы влаги в слое почвы 20—50 см (y) и, таким образом, иметь представление о запасах влаги в полуметровом слое почвы.

Рассчитаем среднюю квадратическую ошибку уравнения регрессии:

$$S_y = \pm\sigma_y\sqrt{1 - r^2} = \pm 9{,}2\sqrt{1 - (0{,}83)^2} = \pm 5{,}15 \text{ мм}.$$

По найденному уравнению регрессии построим теоретическую линию регрессии y по x на корреляционном поле (рис. 2.3). При $x=0$, $y=11{,}56$. Откладываем точку с такими координатами на оси Y. Зададим еще ряд значений x:

при $x=20$ $y=30{,}56$,

при $x=40$ $y=49{,}56$.

Откладываем на графике точки с данными координатами и через них проводим прямую линию. Это и есть искомая теоретическая линия регрессии, уравнение которой было найдено: $y = 0{,}95x + 11{,}56$. Можно было провести линию по двум точкам, задав значения x начала и конца расположения точек корреляционного поля.

Метод группировки данных удобен в расчетах. Он менее громоздок, однако и менее точен. Им можно пользоваться только при большом числе наблюдений (больше 100). При малом числе

наблюдений метод группировки данных применять не следует, так как это может повлечь ошибки, о которых говорилось в разделе 2.1.

2.9. ПРИМЕР РАСЧЕТА УРАВНЕНИЯ ЛИНЕЙНОЙ СВЯЗИ МЕЖДУ ДВУМЯ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ ПО НЕСГРУППИРОВАННЫМ ДАННЫМ (ЗАВИСИМОСТЬ УРОЖАЙНОСТИ ОЗИМОЙ ПШЕНИЦЫ ОТ ВЕСЕННИХ ЗАПАСОВ ВЛАГИ В ПОЧВЕ)

Уравнение регрессии и коэффициент корреляции двух переменных величин часто находят, не группируя данные и не прибегая

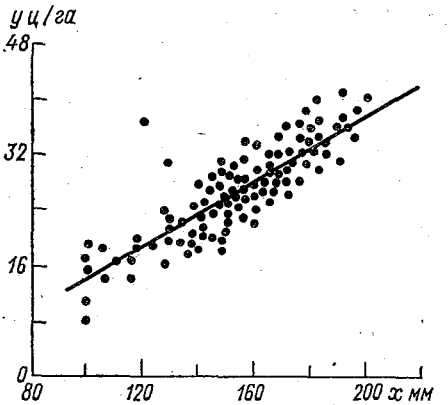


Рис. 2.4. Зависимость урожайности (y) озимой пшеницы от весенних запасов влаги (x) в метровом слое почвы при числе стеблей весной 1000—2000 на 1 м².

к условным единицам. Когда число случаев пар наблюдений не слишком велико (не больше 100), то расчеты и по несгруппированным данным могут быть не слишком громоздкими, если применить для вычисления r формулу (2.22), где величины выражены через отклонения от средних. Этот способ имеет свои преимущества в более быстром прямом пути расчета, при котором меньше возможностей допустить ошибки.

Приведем пример расчета коэффициента корреляции и уравнения регрессии по указанному способу.

Нами [31] была найдена зависимость урожайности озимой пшеницы от весенних запасов влаги в почве (рис. 2.4). Эта зависимость более четко проявляется в зоне недостаточного летнего увлажнения почвы на Украине и Северном Кавказе. В пределах запасов продуктивной влаги весной в метровом слое почвы от 100 до 200 мм, при благоприятных осенних и зимних условиях, когда весной число стеблей у озимой пшеницы на 1 м² составляет 1000—1900, при высокой агротехнике для одного и того же сорта эта зависимость урожайности озимой пшеницы от весенних запасов влаги является прямолинейной.

Проведя анализ данных об урожайности озимой пшеницы при высокой агротехнике на полях опытных станций, сельскохозяй-

ственных институтов, передовых колхозов и совхозов Украины и Северного Кавказа и наблюдений за запасами продуктивной влаги на тех же полях, мы получили большой материал сопряженных данных двух величин для каждого года — урожайности озимой пшеницы и весенних запасов продуктивной влаги в метровом слое почвы.

Урожайность — зависимая переменная величина от запасов влаги — является функцией y. Запасы влаги весной — независимая от урожайности переменная величина — является аргументом x.

Данные каждого года x и y под одним порядковым номером поместили в таблицу и получили таблицу-сводку, где было 100 сопряженных значений x и y, т. е. n — общее число случаев, равное 100.

После этого, отложив на вертикальной оси значения y — урожайности озимой пшеницы, а на горизонтальной оси значения x — запасов влаги весной, для каждой пары значений x и y получили точки на плоскости с координатами x и y. Таким образом, получили корреляционное поле из 100 случаев пар двух переменных величин x и y.

По расположению точек на графике видно, что между этими величинами существует тесная линейная корреляционная связь, для которой необходимо найти уравнение прямой линии регрессии и коэффициент корреляции, определяющий степень тесноты этой связи.

Так как запасы влаги в метровом слое почвы выражаются большими числами, то находить коэффициент корреляции и уравнение прямой регрессии удобнее всего по формуле (2.22). В этой формуле используются не сами величины, а их отклонения (Δ) от средней и тем самым исключается громоздкость расчета:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{\sum \Delta x \Delta y}{\sqrt{\sum \Delta x^2 \sum \Delta y^2}}.$$

Для облегчения расчетов по указанной формуле необходимо составить табл. 2.8 и рассчитать значения указанных в ней величин.

В графе 1 записывается порядковый номер пары наблюдений x и y. В графы 2 и 3 вносятся соответственно значения каждой пары x и y, относящиеся к одному году и к одному полю. В графах 4 и 5 рассчитываются отклонения каждого x_i и y_i от их средних арифметических величин ($\bar{x}$ и $\bar{y}$), для этого берется алгебраическая разность с учетом знака. В графах 6 и 7 вычисляются квадраты отклонений, а в графе 8 — произведение отклонений (с учетом знака). Графы 9 и 10 рассчитываются для контроля.

Таким образом, для расчета указанных граф табл. 2.8 в первую очередь следует найти средние арифметические значения $\bar{x}$ и $\bar{y}$.

Таблица 2.8

Пример расчета уравнения зависимости урожайности озимой пшеницы y от весенних запасов продуктивной влаги в почве x (корреляция двух переменных величин для несгруппированных данных)

№ п/п	x	y	$\Delta x = x - x$	$\Delta y = y - y$	Δx^2	Δy^2	$\Delta x \Delta y$	$\Delta x + \Delta y$	$(\Delta x + \Delta y)^2$
1	2	3	4	5	6	7	8	9	10
1	100	8	—53	—18,5	2809	342,25	980,5	—71,5	5112,25
2	100	10	—53	—16,5	2809	272,25	874,5	—69,5	4830,25
3	100	15	—53	—11,5	2809	132,25	609,5	—64,5	4160,25
4	100	16	—53	—10,5	2809	110,25	556,5	—63,5	4032,25
5	100	18	—53	—8,5	2809	72,25	450,5	—61,5	3782,25
6	105	14	—48	—12,5	2304	156,25	600,0	—60,5	3660,25
7	105	18	—48	—8,5	2304	72,25	408,0	—56,5	3192,25
8	110	16	—43	—10,5	1849	110,25	451,5	—53,5	2862,25
9	115	14	—38	—12,5	1444	156,25	475,0	—50,5	2550,25
10	115	16	—38	—10,5	1444	110,25	399,0	—48,5	2352,25
. . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .
95	190	36	37	9,5	1369	90,25	351,5	46,5	2162,25
96	190	36	37	9,5	1369	90,25	351,5	46,5	2162,25
97	190	36	37	9,5	1369	90,25	351,5	46,5	2162,25
98	195	34	42	7,5	1764	56,25	315,0	49,5	2450,25
99	195	38	42	11,5	1764	132,25	483,0	53,5	2862,25
100	200	40	47	13,5	2209	182,25	634,5	60,5	3660,25
Σ	15 300	2650			60 049	4777	14 705		94 236

В нашем примере

$$\bar{x}=\frac{\sum x}{n}=\frac{15\,300}{100}=153,\quad \bar{y}=\frac{\sum y}{n}=\frac{2650}{100}=26{,}5.$$

Найдя отклонения $\Delta x=x_i-\bar{x}$ и $\Delta y=y_i-\bar{y}$, их произведение $\Delta x\Delta y$, их квадраты Δx^2 и Δy^2, а также суммы этих величин, необходимо провести контроль расчетов по формуле

$$\sum (x-\bar{x})^2+\sum (y-\bar{y})^2+2\sum (x-\bar{x})(y-\bar{y})=$$
$$=\sum [(x-\bar{x})+(y-\bar{y})]^2.$$

Если расчеты в таблице верны, то значения чисел левой и правой частей формулы будут одинаковы, в противном случае необходимо провести пересчеты, пока не обнаружится ошибка и не будет соблюдено указанное равенство. Только после этого можно приступать к дальнейшим расчетам коэффициента корреляции и уравнения регрессии.

В нашем примере контроль показал правильность расчетов:

$$\sum (x-\bar{x})^2+\sum (y-\bar{y})^2+2\sum (x-\bar{x})(y-\bar{y})=$$
$$=60\,049+4777+2\cdot 14\,705=94\,236;$$
$$\sum [(x-\bar{x})+(y-\bar{y})]^2=94\,236.$$

Коэффициент корреляции r находим по следующей формуле:

$$r=\frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2\sum (y-\bar{y})^2}}=\frac{\sum \Delta x\,\Delta y}{\sqrt{\sum \Delta x^2\sum \Delta y^2}}=$$

$$=\frac{14\,705}{\sqrt{60\,049\cdot 4777}}=0{,}86.$$

Средняя ошибка коэффициента корреляции равна

$$\sigma_r=\pm\frac{1-r^2}{\sqrt{n}}=\frac{1-(0{,}86)^2}{\sqrt{100}}=\pm 0{,}026.$$

Вероятная ошибка коэффициента корреляции равна

$$E_r=\pm 0{,}67\sigma_r=\pm 0{,}67\cdot 0{,}026=\pm 0{,}017.$$

Отсюда наиболее вероятные значения коэффициента корреляции заключены в интервале

$$r\pm E_r=0{,}86\pm 0{,}017=\begin{cases}0{,}88,\\0{,}84.\end{cases}$$

Предельное значение r близко к $r\pm 4E_r$ или $r\pm 3\sigma_r$:

$$r\pm 4E_r=0{,}86\pm 0{,}07=\begin{cases}0{,}93\\0{,}79\end{cases}\text{ или }\ r\pm 3\sigma_r=0{,}86\pm 0{,}08=\begin{cases}0{,}94,\\0{,}78.\end{cases}$$

Как следует из этих расчетов, на Украине и Северном Кавказе наблюдается тесная связь урожайности озимой пшеницы с весенними запасами влаги, коэффициент корреляции которой высокий: $r=0{,}86$, а его ошибки небольшие. Даже с учетом этих ошибок предельное значение коэффициента корреляции не становится меньше 0,78, что говорит о хорошей связи между указанными величинами.

Перейдем теперь к расчету уравнения линии прямой регрессии по формуле $y-\bar{y}=R(x-\bar{x})$, где R — коэффициент уравнения регрессии

$$R_{y/x}=r\frac{\sigma_y}{\sigma_x},$$

σ_y и σ_x — средние квадратические отклонения:

$$\sigma_x=\sqrt{\frac{\sum(x-\bar{x})^2}{n}}=\sqrt{\frac{60\,049}{100}}=24{,}5,$$

$$\sigma_y=\sqrt{\frac{\sum(y-\bar{y})^2}{n}}=\sqrt{\frac{4777}{100}}=6{,}9.$$

Откуда

$$R=r\frac{\sigma_y}{\sigma_x}=0{,}86\frac{6{,}9}{24{,}5}=0{,}24.$$

Подставив значения $\bar{x}$, $\bar{y}$ и R в уравнение прямой линии, получим

$$y=0{,}24x-36{,}72+26{,}50.$$

Отсюда уравнение прямой линии, характеризующее найденную нами зависимость, окончательно имеет вид

$$y=0{,}24x-10{,}22,$$

где y — урожайность озимой пшеницы, ц/га; x — запасы продуктивной влаги в метровом слое почвы весной при переходе средней декадной температуры воздуха через 5 °С, мм.

Определим среднюю квадратическую ошибку найденного уравнения регрессии:

$$S_y=\pm\sigma_y\sqrt{1-r^2}=\pm6{,}9\sqrt{1-(0{,}86)^2}=\pm3{,}5 \text{ ц/га}.$$

Таким образом, зная весенние запасы влаги в почве, по данному уравнению можно с трехмесячной заблаговременностью независимо от будущей погоды определять виды на урожайность озимой пшеницы с ошибкой $\pm3{,}5$ ц/га.

При нахождении уравнений корреляционных связей следует указывать пределы их действия. Найденное нами уравнение, как было указано выше, действует в пределах значений весенних запасов влаги от 100 до 200 мм.

По указанному уравнению, задавая различные значения x, находим значения y и строим теоретическую линию регрессии y по x (см. рис. 2.4). Например, задаем значение $x=100$ мм, тогда $y=13{,}8$ ц/га. Отмечаем эту точку на графике. При $x=200$ мм $y=37{,}8$ ц/га получаем вторую точку на графике. Через указанные две точки проводим прямую линию. Это и будет искомая теоретическая линия регрессии, уравнение которой $y=0{,}24x-10{,}22$.

При нахождении теоретической линии регрессии достаточно задать только два значения x и, рассчитав два значения y, получить на графике две точки. Как известно, через две точки можно провести только одну прямую. Поэтому, имея две точки, мы правильно проведем искомую теоретическую линию прямой регрессии, уравнение которой нашли.

Глава 3. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ ТРЕХ ПЕРЕМЕННЫХ ВЕЛИЧИН

3.1. ПАРНЫЕ И ОБЩИЙ КОЭФФИЦИЕНТЫ МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ. УРАВНЕНИЕ СВЯЗИ МЕЖДУ ТРЕМЯ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ

При исследовании связей между различными явлениями часто можно встретиться с тем, что на одну переменную величину оказывают влияние сразу несколько переменных величин. В этой главе рассмотрим вопрос о связи трех переменных величин, т. е. когда одна зависимая переменная величина z зависит главным образом от двух других величин, независимых переменных x и y. Наиболее простой формой такой связи является линейная связь.

В случае зависимости между тремя величинами, уравнение линейной множественной корреляции будет иметь вид

$$z=ax+by+c, \tag{3.1}$$

где a, b, c — постоянные величины, параметры данного уравнения, которые необходимо определить.

Параметры уравнения можно определить двумя путями: 1) через коэффициенты корреляции данных переменных величин, 2) методом наименьших квадратов.

Рассмотрим первый способ расчета параметров уравнения через коэффициенты корреляции.

Линейное уравнение связи между тремя переменными величинами можно представить также в виде

$$z-\bar{z}=a(x-\bar{x})+b(y-\bar{y}), \tag{3.2}$$

где $\bar{z}$, $\bar{x}$ и $\bar{y}$ — средние арифметические значения величин z, x и y.

После определения вида уравнения встает задача определения тесноты связи между тремя переменными величинами, т. е. определения общего коэффициента корреляции R.

Общий коэффициент корреляции вычисляется по формуле

$$R = \sqrt{\frac{r_{zx}^2 + r_{zy}^2 - 2r_{zx}r_{zy}r_{xy}}{1 - r_{xy}^2}}, \qquad (3.3)$$

где r_{zx}, r_{zy}, r_{xy} — парные коэффициенты корреляции.

Общий коэффициент корреляции R имеет следующие свойства.

1. Значение R всегда положительно и изменяется от 0 до 1

$$0 \leqslant R \leqslant 1;$$

2. Если $R=0$, то z не может быть линейно связано с x и y; однако при этом возможна нелинейная корреляционная и даже функциональная связь z с x и y;

3. Если $R=1$, то z связано с x и y линейной функциональной связью;

4. Если R отлично от своих крайних значений (0 и 1), то при приближении R к единице теснота линейной связи z с x и y увеличивается.

Для того чтобы выделить степень влияния на полученный результат каждого фактора в отдельности и определить общий коэффициент корреляции R, надо рассчитать парные коэффициенты корреляции:

$$r_{zx} = \frac{\sum \Delta x \, \Delta z}{\sqrt{\sum \Delta x^2 \sum \Delta z^2}}, \quad r_{zy} = \frac{\sum \Delta y \, \Delta z}{\sqrt{\sum \Delta y^2 \sum \Delta z^2}},$$

$$r_{xy} = \frac{\sum \Delta x \, \Delta y}{\sqrt{\sum \Delta x^2 \sum \Delta y^2}}.$$

Здесь

$$\Delta x = x_i - \bar{x}; \quad \Delta y = y_i - \bar{y}; \quad \Delta z = z_i - \bar{z}.$$

Каждый из указанных парных коэффициентов корреляции при наличии связи между тремя переменными величинами определяет тесноту линейной связи между двумя величинами, когда третья величина остается (условно) постоянной. Парные коэффициенты корреляции r могут быть определены и по другим формулам, указанным в разделе 2.4.

Определив парные коэффициенты корреляции r_{zx}, r_{zy}, r_{xy}, а также общий коэффициент корреляции R и убедившись в достаточно надежной тесноте связи между исследуемыми величинами, переходим к определению параметров a, b и c уравнения линейной регрессии $z = ax + by + c$. Формулы для вычисления параметров a и b имеют следующий вид:

$$a = \frac{\sigma_z}{\sigma_x} \cdot \frac{r_{zx} - r_{zy}r_{xy}}{1 - r_{xy}^2}, \qquad (3.4)$$

$$b = \frac{\sigma_z}{\sigma_y} \cdot \frac{r_{zy} - r_{zx}r_{xy}}{1 - r_{xy}^2}, \qquad (3.5)$$

где σ_z, σ_x, σ_y — средние квадратические отклонения соответственно рядов z, x и y.

$$\sigma_z = \sqrt{\frac{\sum \Delta z^2}{n}}, \quad \sigma_x = \sqrt{\frac{\sum \Delta x^2}{n}}, \quad \sigma_y = \sqrt{\frac{\sum \Delta y^2}{n}}$$

(n — общее число наблюдений).

Подставив значения параметров a и b в уравнение (3.2), получим общий вид уравнения регрессии трех переменных величин:

$$z - \bar{z} = \frac{\sigma_z}{\sigma_x} \frac{r_{zx} - r_{zy}r_{xy}}{1 - r_{xy}^2}(x - \bar{x}) + \frac{\sigma_z}{\sigma_y} \frac{r_{zy} - r_{zx}r_{xy}}{1 - r_{xy}^2}(y - \bar{y}). \qquad (3.6)$$

Решив уравнение (3.6), получим окончательное уравнение множественной регрессии $z = ax + by + c$, которое будет характеризовать найденную нами связь между тремя переменными величинами.

Средняя квадратическая ошибка уравнения регрессии трех переменных величин вычисляется по формуле

$$S_z = \pm\sigma_z \sqrt{\frac{1 - r_{zx}^2 - r_{zy}^2 - r_{xy}^2 + 2r_{zx}r_{zy}r_{xy}}{1 - r_{xy}^2}}, \qquad (3.7)$$

Средняя квадратическая ошибка общего коэффициента множественной корреляции вычисляется по формуле

$$\sigma_R = \pm\frac{1 - R^2}{\sqrt{n}}. \qquad (3.8)$$

Данные, необходимые для расчета общего коэффициента множественной корреляции R и уравнения регрессии трех переменных величин вычисляются с помощью таблицы, аналогичной табл. 3.1.

Таблица 3.1

№ п/п	z	x	y	Δz	Δx	Δy	Δz^2	Δx^2	Δy^2	$\Delta z \Delta x$	$\Delta z \Delta y$	$\Delta x \Delta y$	$\Delta z + \Delta x + \Delta y$	$(\Delta z + \Delta x + \Delta y)^2$
n	Σ	Σ	Σ				Σ	Σ	Σ	Σ	Σ	Σ		Σ

3.2. ПРИМЕР РАСЧЕТА УРАВНЕНИЯ ЛИНЕЙНОЙ СВЯЗИ МЕЖДУ ТРЕМЯ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ (ЗАВИСИМОСТЬ ЗАПАСОВ ВЛАГИ В ПОЧВЕ ОТ КОЛИЧЕСТВА ОСАДКОВ В РАЗНЫЕ ПЕРИОДЫ)

Для характеристики условий роста и развития озимых культур осенью в Западной Сибири нами [29] была установлена зависимость запасов продуктивной влаги в пахотном слое почвы от количества осадков за текущий и предшествующий месяцы. Рассмотрим в качестве примера линейной связи между тремя переменными величинами зависимость запасов влаги в пахотном слое почвы в августе, когда проходит сев озимых культур в Западной Сибири, от суммы осадков в августе и июле.

Установить данную связь необходимо для того, чтобы, имея большой ряд наблюдений за осадками и небольшие ряды наблюдений за влажностью почвы, можно было по количеству осадков рассчитать запасы влаги в почве и дать агроклиматическую оценку условий увлажнения почвы в период сева озимых.

Таким образом, необходимо найти уравнение указанной связи трех переменных величин вида

$$z = ax + by + c,$$

где z — средние за август запасы продуктивной влаги в слое почвы 0—20 см, мм; x — сумма осадков в августе, мм; y — сумма осадков в июле, мм.

Для нахождения неизвестных параметров уравнения a, b, c и коэффициента множественной корреляции R, указывающего на тесноту связи, данные наблюдений необходимо расположить в табл. 3.2 и провести расчеты остальных граф.

Две последние графы в табл. 3.2 рассчитываются для контроля.

После анализа материала наблюдений и установления его пригодности для обработки в табл. 3.2, записываем данные о запасах влаги (z) и количестве осадков за август и июль (соответственно x и y) и находим их средние значения:

$$\bar{z} = \frac{\sum z}{n} = \frac{1352}{52} = 26{,}0,$$

$$\bar{x} = \frac{\sum x}{n} = \frac{2496}{52} = 48{,}0,$$

$$\bar{y} = \frac{\sum y}{n} = \frac{2964}{52} = 57{,}0.$$

После этого рассчитываем отклонения (Δ) значений каждой величины от средней:

$$\Delta z = z_i - \bar{z}, \quad \Delta x = x_i - \bar{x}, \quad \Delta y = y_i - \bar{y}.$$

Таблица 3.2

Пример расчета уравнения зависимости средних за август запасов продуктивной влаги в слое почвы 0—20 см (z) от суммы осадков за июль (y) и август (x) (множественная корреляция между тремя переменными величинами)

№ п/п	z	x	y	Δz	Δx	Δy	Δz^2	Δx^2	Δy^2	$\Delta z \Delta x$	$\Delta z \Delta y$	$\Delta x \Delta y$	$\Delta z + \Delta x + \Delta y$	$(\Delta z + \Delta x + \Delta y)^2$
1	5	5	30	—21	—43	—27	441	1849	729	903	567	1161	—91	8281
2	5	15	15	—21	—33	—42	441	1089	1764	693	882	1386	—96	9216
3	7	20	15	—19	—28	—42	361	784	1764	532	798	1176	—89	7921
4	8	20	40	—18	—28	—17	324	784	289	504	306	476	—63	3969
5	8	25	20	—18	—23	—37	324	529	1369	414	666	851	—78	6084
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
48	45	80	65	19	32	8	361	1024	64	608	152	256	59	3481
49	50	90	80	24	42	23	576	1764	529	1008	552	966	89	7921
50	50	100	90	24	52	33	576	2704	1089	1248	792	1716	109	11881
51	50	80	90	24	32	33	576	1024	1089	768	792	1056	89	7921
52	55	100	110	29	52	53	841	2704	2809	1508	1537	2756	134	17956
$\Sigma n = 52$	1352	2496	2964				8790	30298	51274	13695	15208	20023		188214

Найдя разности указанных значений, их квадраты, их произведения и суммы этих значений и записав все в таблицу, делаем контроль расчетов по формуле

$$\sum \Delta z^2 + \sum \Delta x^2 + \sum \Delta y^2 + 2\sum \Delta z\,\Delta x + 2\sum \Delta z\,\Delta y + 2\sum \Delta x\,\Delta y = \sum (\Delta x + \Delta y + \Delta z)^2;$$

$$\sum (\Delta x + \Delta y + \Delta z)^2 = 188\,214.$$

$(8790+30298+51274+2\cdot 13695+2\cdot 15208+2\cdot 20023=188214.)$

Контроль показал правильность расчетов, поэтому можно перейти к нахождению парных коэффициентов корреляции между коррелируемыми величинами r_{zx}, r_{zy}, r_{xy}:

$$r_{zx} = \frac{\sum \Delta x\,\Delta z}{\sqrt{\sum \Delta x^2 \sum \Delta z^2}} = \frac{13\,695}{\sqrt{30\,298 \cdot 8790}} = 0{,}84,$$

$$r_{zy} = \frac{\sum \Delta y\,\Delta z}{\sqrt{\sum \Delta y^2 \sum \Delta z^2}} = \frac{15\,208}{\sqrt{51\,274 \cdot 8790}} = 0{,}71,$$

$$r_{xy} = \frac{\sum \Delta x\,\Delta y}{\sqrt{\sum \Delta x^2 \sum \Delta y^2}} = \frac{20\,023}{\sqrt{30\,298 \cdot 51\,274}} = 0{,}51.$$

После этого находим общий коэффициент множественной корреляции R и его вероятную ошибку E_R:

$$R = \sqrt{\frac{r_{zx}^2 + r_{zy}^2 - 2r_{zx}r_{zy}r_{xy}}{1 - r_{xy}^2}} =$$

$$= \sqrt{\frac{(0{,}84)^2 + (0{,}71)^2 - 2\,(0{,}84 \cdot 0{,}71 \cdot 0{,}51)}{1 - (0{,}51)^2}} = 0{,}90;$$

$$E_R = \pm 0{,}67\,\frac{1 - R^2}{\sqrt{n}} = \pm 0{,}67\,\frac{1 - (0{,}90)^2}{\sqrt{52}} = \pm 0{,}02.$$

Следовательно, наиболее вероятные значения R заключены в интервале

$$R \pm E_R = 0{,}90 \pm 0{,}02 = \begin{cases} 0{,}92, \\ 0{,}88, \end{cases}$$

а предельные значения коэффициента корреляции

$$R \pm 4E_R = 0{,}90 \pm 0{,}08 = \begin{cases} 0{,}98, \\ 0{,}82. \end{cases}$$

Как видим, значения R очень высокие, т. е. связь запасов влаги в слое почвы 0—20 см с количеством осадков за август и июль очень тесная.

Найдем уравнение регрессии данной связи. Для этого рассчитаем средние квадратические отклонения:

$$\sigma_z = \sqrt{\frac{\sum \Delta z^2}{n}} = \sqrt{\frac{8790}{52}} = 13{,}0,$$

$$\sigma_x = \sqrt{\frac{\sum \Delta x^2}{n}} = \sqrt{\frac{30\,298}{52}} = 24{,}14,$$

$$\sigma_y = \sqrt{\frac{\sum \Delta y^2}{n}} = \sqrt{\frac{51\,274}{52}} = 31{,}4.$$

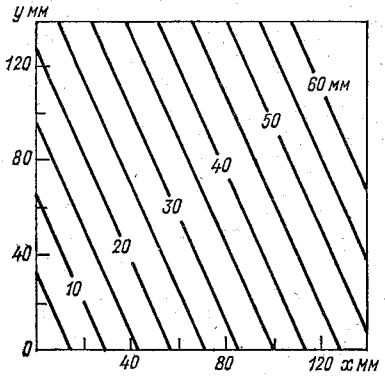


Рис. 3.1. Номограмма для определения запасов влаги в слое почвы 0—20 см в августе по количеству осадков в августе (x) и июле (y) в Западной Сибири.

Подставляя в уравнение найденные значения $\bar{z}$, $\bar{x}$, $\bar{y}$, σ_z, σ_x, σ_y, r_{zx}, r_{zy}, r_{xy}, находим искомые параметры a, b и c:

$$z - \bar{z} = \frac{\sigma_z}{\sigma_x} \frac{r_{zx} - r_{zy} r_{xy}}{1 - r_{xy}^2} (x - \bar{x}) + \frac{\sigma_z}{\sigma_y} \frac{r_{zy} - r_{zx} r_{xy}}{1 - r_{xy}^2} (y - \bar{y}),$$

$$z = 26{,}0 + 0{,}54 \cdot 0{,}65\,(x - 48) + 0{,}41 \cdot 0{,}38\,(y - 57),$$

$$z = 0{,}35x + 0{,}16y + 0{,}1. \qquad (3.9)$$

Таким образом, уравнение зависимости запасов влаги от суммы осадков за август и июль имеет вид $z = 0{,}35x + 0{,}16y + 0{,}1$, где z — средние за август запасы продуктивной влаги в слое почвы 0—20 см, мм; x — сумма осадков за август, мм; y — сумма осадков за июль, мм.

Находим среднюю квадратическую ошибку данного уравнения регрессии

$$S_z = \pm\sigma_z \sqrt{1 - R^2} = 13{,}0 \sqrt{1 - (0{,}90)^2} = \pm 5{,}7 \text{ мм}.$$

Для удобства расчетов по найденному уравнению построим график, где по оси абсцисс (x) отложим суммы осадков за август, а по оси ординат (y) — суммы осадков за июль (рис. 3.1).

Задавая различные значения x и y, получим в поле графика значения запасов влаги (z). Эти значения будут самые различные, по ним трудно провести линии равных значений (z) через определенные интервалы, которые мы хотим получить на графике. Поэтому график лучше строить следующим образом.

Допустим, надо получить на графике значения z через интервал 5 мм. Задаем значения z через каждые 5 мм и решаем уравнение сначала относительно x при $y=0$, а затем при тех же значениях z решаем уравнение относительно y при $x=0$. Например, на графике необходимо получить линию $z=5$ мм. Решаем уравнения для данного $z=5$ мм относительно x при $y=0$, получаем $5=0{,}35x+0{,}1$, откуда $x=14{,}0$. Откладывая 14 на оси абсцисс, получаем точку с координатами $z=5$, $y=0$, $x=14$.

Теперь решаем уравнение в отношении y для этого же $z=5$ мм, но при $x=0$, получаем $5=0{,}16y+0{,}1$, откуда $y=31$. Откладывая 31 на оси y, получаем точку с координатами $z=5$, $x=0$, $y=31$. Таким образом, мы имеем две точки на осях, где $z=5$ мм. Через две точки, как известно, можно провести только одну прямую. Проводим эту прямую для $z=5$ мм.

Если надо построить график для значений z через интервал 5 мм, то дальнейшие расчеты проводят аналогичным образом для $z=10$, 15, 20 мм и т. д.

Получив точку на оси x при $z=10$, $y=0$, $x=28$ и точку на оси y при $z=10$, $x=0$, $y=62$, проводим через две эти точки линию равных значений $z=10$ мм и так же для $z=15$ мм, $z=20$ мм и т. д.

Таким образом, получаем графическое изображение зависимости, по которому легко производить расчеты запасов влаги в почве в зависимости от суммы осадков (рис. 3.1).

Уравнения связи между тремя переменными величинами можно рассчитывать также методом сгруппированных данных с введением условных единиц. В этом случае для подсчета каждого парного коэффициента корреляции r_{zx}, r_{zy}, r_{xy} составляются три отдельные корреляционные таблицы, подобно табл. 2.5. По этим таблицам кроме r_{zx}, r_{zy}, r_{xy} находят $\bar{z}$, $\bar{x}$, $\bar{y}$, σ_z, σ_x и σ_y точно таким же способом, как было подробно изложено в разделе 2.8. Затем, подставив рассчитанные величины в формулы для R и z, приведенные в настоящем разделе, получим множественный коэффициент корреляции R и уравнение регрессии трех переменных величин.

Глава 4. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ ЧЕТЫРЕХ ПЕРЕМЕННЫХ ВЕЛИЧИН

4.1. УРАВНЕНИЕ СВЯЗИ МЕЖДУ ЧЕТЫРЬМЯ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ. ПАРНЫЕ И ОБЩИЙ КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ

Часто бывает, что связь между двумя или тремя величинами недостаточно тесная и необходимо учитывать еще ряд факторов. Тогда ищут связь между четырьмя величинами или точнее ищут

зависимость одной переменной величины от трех других переменных величин. Уравнение этой зависимости будет иметь вид

$$u = ax + by + cz + d. \tag{4.1}$$

Для удобства расчетов параметров уравнения a, b, c, d и коэффициента множественной корреляции R составляется таблица, аналогичная табл. 4.1, где $\Delta u = u_i - \bar{u}$, $\Delta x = x_i - \bar{x}$, $\Delta y = y_i - \bar{y}$, $\Delta z = z_i - \bar{z}$, а $\bar{u}$, $\bar{x}$, $\bar{y}$ и $\bar{z}$ — средние арифметические величины.

Таблица 4.1

№ п/п	u	x	y	z	Δu	Δx	Δy	Δz	Δu^2	Δx^2	Δy^2	Δz^2	$\Delta u \Delta x$	$\Delta u \Delta y$	$\Delta u \Delta z$	$\Delta x \Delta y$	$\Delta x \Delta z$	$\Delta y \Delta z$	$(\Delta x + \Delta y + \Delta z + \Delta u)$	$(\Delta x + \Delta y + \Delta z + \Delta u)^2$
n	Σ	Σ	Σ	Σ					Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ		Σ

При корреляции четырех переменных величин необходимо найти шесть парных коэффициентов корреляции. Они вычисляются по формулам, аналогичным (2.22):

$$r_{ux} = \frac{\sum \Delta u \, \Delta x}{\sqrt{\sum \Delta u^2 \sum \Delta x^2}}, \quad r_{uy} = \frac{\sum \Delta u \, \Delta y}{\sqrt{\sum \Delta u^2 \sum \Delta y^2}},$$

$$r_{uz} = \frac{\sum \Delta u \, \Delta z}{\sqrt{\sum \Delta u^2 \sum \Delta z^2}}, \quad r_{xy} = \frac{\sum \Delta x \, \Delta y}{\sqrt{\sum \Delta x^2 \sum \Delta y^2}},$$

$$r_{xz} = \frac{\sum \Delta x \, \Delta z}{\sqrt{\sum \Delta x^2 \sum \Delta z^2}}, \quad r_{yz} = \frac{\sum \Delta y \, \Delta z}{\sqrt{\sum \Delta y^2 \sum \Delta z^2}}.$$

После этого находят общий коэффициент множественной корреляции четырех величин R по формуле

$$R = \sqrt{1 - \frac{D}{D_0}}, \tag{4.2}$$

где

$$D = 1 - r_{ux}^2 - r_{uy}^2 - r_{uz}^2 - r_{xy}^2 - r_{xz}^2 - r_{yz}^2 + r_{ux}^2 r_{yz}^2 + r_{xy}^2 r_{uz}^2 + r_{uy}^2 r_{xz}^2 + 2(r_{xy} r_{xz} r_{yz} + r_{yz} r_{uz} r_{uy} + r_{ux} r_{uz} r_{xz} + r_{ux} r_{uy} r_{xy}) - 2(r_{ux} r_{uz} r_{xy} r_{yz} + r_{uz} r_{uy} r_{xy} r_{xz} + r_{ux} r_{uy} r_{xz} r_{yz}); \tag{4.3}$$

$$D_0 = 1 - r_{xy}^2 - r_{yz}^2 - r_{xz}^2 + 2 r_{xy} r_{xz} r_{yz}. \tag{4.4}$$

Уравнение регрессии четырех переменных величин, выраженное через r и σ имеет общий вид

$$u-\bar{u}=\frac{\sigma_u}{\sigma_x}\left[\frac{r_{ux}\left(1-r_{yz}^2\right)-r_{uy}r_{xy}-r_{uz}r_{xz}+r_{yz}\left(r_{uy}r_{xz}+r_{uz}r_{xy}\right)}{1-r_{yz}^2-r_{xy}^2-r_{xz}^2+2r_{yz}r_{xy}r_{xz}}\right]\times$$

$$\times(x-\bar{x})+\frac{\sigma_u}{\sigma_y}\left[\frac{r_{uy}\left(1-r_{xz}^2\right)-r_{ux}r_{xy}-r_{uz}r_{yz}+r_{xz}\left(r_{ux}r_{yz}+r_{uz}r_{xy}\right)}{1-r_{yz}^2-r_{xy}^2-r_{xz}^2+2r_{yz}r_{xy}r_{xz}}\right]\times$$

$$\times(y-\bar{y})+\frac{\sigma_u}{\sigma_z}\left[\frac{r_{uz}\left(1-r_{xy}^2\right)-r_{ux}r_{xz}-r_{uy}r_{yz}+r_{xy}\left(r_{ux}r_{yz}+r_{uy}r_{xz}\right)}{1-r_{yz}^2-r_{xy}^2-r_{xz}^2+2r_{yz}r_{xy}r_{xz}}\right]\times$$

$$\times(z-\bar{z}). \qquad (4.5)$$

Определив средние квадратические отклонения

$$\sigma_u=\sqrt{\frac{\sum(u-\bar{u})^2}{n}},\quad \sigma_x=\sqrt{\frac{\sum(x-\bar{x})^2}{n}},$$

$$\sigma_y=\sqrt{\frac{\sum(y-\bar{y})^2}{n}},\quad \sigma_z=\sqrt{\frac{\sum(z-\bar{z})^2}{n}}$$

и решив общее уравнение относительно u, получим уравнение окончательного вида для связи четырех переменных $u=ax+by+cz+d$.

Средняя квадратическая ошибка уравнения регрессии рассчитывается по формуле

$$S_u=\pm\sigma_u\sqrt{1-R^2}. \qquad (4.6)$$

Как видим, расчеты уравнения связи четырех переменных не сложны, но очень громоздки. При введении еще большего числа переменных расчеты становятся еще более громоздкими. В этом случае подсчитывают парные коэффициенты корреляции между искомой величиной и каждой из величин, связь с которыми надо отыскать. Парные коэффициенты корреляции покажут, какие величины следует учесть, а какими можно пренебречь.

4.2. ПРИМЕР РАСЧЕТА УРАВНЕНИЯ ЛИНЕЙНОЙ КОРРЕЛЯЦИИ ЧЕТЫРЕХ ПЕРЕМЕННЫХ ВЕЛИЧИН (ЗАВИСИМОСТЬ ЗАПАСОВ ВЛАГИ В ПОЧВЕ ОТ КОЛИЧЕСТВА ОСАДКОВ, ТЕМПЕРАТУРЫ И ИСХОДНЫХ ЗАПАСОВ ВЛАГИ)

В агрометеорологии, как известно, прогнозирование запасов продуктивной влаги под различными культурами проводится на основе расчетов по уравнениям с четырьмя переменными. Это уравнение зависимости запасов влаги в почве к концу декады от суммы осадков и средней температуры за декаду, а также от исходных запасов влаги к началу декады.

Нами в [29] были найдены уравнения зависимости запасов влаги в почве под кукурузой к концу декады от суммы осадков и средней температуры воздуха за декаду, а также от исходных запасов влаги к началу декады. Эти зависимости были определены для метрового и пахотного слоя почвы по различным отрезкам вегетационного периода кукурузы.

Приведем пример нахождения указанного уравнения зависимости запасов влаги в метровом слое почвы к концу декады (u) от исходных запасов влаги к началу декады (x), суммы осадков (y) и средней температуры воздуха (z) за декаду в период от выбрасывания метелки до молочной спелости кукурузы. После анализа большого материала наблюдений данные заносим в табл. 4.2 и рассчитываем средние величины $\bar{u}$, $\bar{x}$, $\bar{y}$ и $\bar{z}$:

$$\bar{u} = \frac{\sum u}{n} = \frac{14\,848}{232} = 64, \quad \bar{x} = \frac{\sum x}{n} = \frac{19\,024}{232} = 82,$$

$$\bar{y} = \frac{\sum y}{n} = \frac{3944}{232} = 17, \quad \bar{z} = \frac{\sum z}{n} = \frac{5034}{232} = 21{,}7.$$

Затем находим разности значений каждой величины со средними ($\Delta u = u_i - \bar{u}$, $\Delta x = x_i - \bar{x}$ и т. д.), квадраты этих разностей, произведения разностей и суммы этих величин.

После расчета всех граф табл. 4.2 и получения сумм указанных величин по графам, приступаем к расчету шести парных коэффициентов корреляций между различными величинами:

$$r_{ux} = \frac{\sum \Delta u\, \Delta x}{\sqrt{\sum \Delta u^2 \sum \Delta x^2}} = \frac{222\,633}{\sqrt{239\,838 \cdot 308\,717}} = 0{,}82,$$

$$r_{uy} = \frac{\sum \Delta u\, \Delta y}{\sqrt{\sum \Delta u^2 \sum \Delta y^2}} = \frac{48\,989}{\sqrt{239\,838 \cdot 95\,252}} = 0{,}32,$$

$$r_{uz} = \frac{\sum \Delta u\, \Delta z}{\sqrt{\sum \Delta u^2 \sum \Delta z^2}} = \frac{-4648}{\sqrt{239\,838 \cdot 1062}} = -0{,}29,$$

$$r_{xy} = \frac{\sum \Delta x\, \Delta y}{\sqrt{\sum \Delta x^2 \sum \Delta y^2}} = \frac{-10\,144}{\sqrt{308\,717 \cdot 95\,252}} = -0{,}06,$$

$$r_{xz} = \frac{\sum \Delta x\, \Delta z}{\sqrt{\sum \Delta x^2 \sum \Delta z^2}} = \frac{-1465}{\sqrt{308\,717 \cdot 1062}} = -0{,}08,$$

$$r_{yz} = \frac{\sum \Delta y\, \Delta z}{\sqrt{\sum \Delta y^2 \sum \Delta z^2}} = \frac{-2255}{\sqrt{95\,252 \cdot 1062}} = -0{,}22.$$

Таблица 4.2

Пример расчета уравнения зависимости запасов продуктивной влаги в почве к концу декады (u) от исходных запасов влаги к началу декады (x), суммы осадков (y) и температуры воздуха (z) за декаду (множественная корреляция четырех переменных величин)

№ п/п	u	x	y	z	Δu	Δx	Δy	Δz	Δu^2	Δx^2	Δy^2	Δz^2	$\Delta n \Delta x$	$\Delta u \Delta y$	$\Delta u \Delta z$	$\Delta x \Delta y$	$\Delta x \Delta z$	$\Delta y \Delta z$
1	109	140	3	21,8	45	58	—14	0,1	2025	3364	196	0,0	2610	—630	4,5	—812	5,8	—1,4
2	131	109	62	23,2	67	27	45	1,5	4489	729	2025	2,2	1809	3015	100,5	1215	40,5	67,5
3	102	131	10	19,7	38	49	—7	—2,0	1444	2401	49	4,0	1862	—266	—76,0	—343	—98,0	14,0
4	101	154	0	20,8	37	72	—16	—0,9	1369	5184	289	0,8	2664	—629	—33,3	—1224	—64,8	15,3
5	129	101	61	18,2	65	19	44	—3,5	4225	361	1936	12,2	1235	2860	—227,5	836	—66,5	—154,0
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
228	67	93	11	19,0	3	11	—6	—2,7	9	121	36	7,3	33	—18	—8,1	—66	—29,7	16,2
229	40	62	8	21,6	—24	—20	—9	—0,1	576	400	81	0,0	480	216	2,4	180	2,0	0,9
230	45	65	7	23,0	—19	—17	—10	1,3	361	289	100	1,7	323	190	24,7	170	—22,1	—13,0
231	12	45	0	22,8	—52	—37	—17	1,1	2704	1369	289	1,2	1924	884	—57,2	629	—40,7	—18,7
232	4	12	45	24,4	—60	—70	28	2,7	3600	4900	784	7,3	4200	—1680	—162,0	—1960	—189,0	75,6
$\sum n =$ $= 232$	14 848	19 024	3944	5034					239 838	308 717	95 252	1062	222 633	48 989	—4648	—10 144	—1465	—2255

Находим общий коэффициент множественной корреляции

$$R = \sqrt{1 - \frac{D}{D_0}}.$$

Согласно (4.3) и (4.4),

$$D = 0{,}1593, \quad D_0 = 0{,}9396;$$

тогда

$$R = \sqrt{1 - \frac{0{,}1593}{0{,}9396}} = 0{,}91.$$

Вероятная ошибка коэффициента множественной корреляции

$$E_R = \pm 0{,}67 \frac{1 - R^2}{\sqrt{n}} = \pm 0{,}67 \frac{1 - 0{,}91^2}{\sqrt{232}} = \pm 0{,}007.$$

Предельные значения R равны

$$R \pm 4E_R = 0{,}91 \pm 0{,}028 = \begin{cases} 0{,}94, \\ 0{,}88. \end{cases}$$

Связь, судя по коэффициенту корреляции, очень хорошая.

Находим средние квадратические отклонения:

$$\sigma_u = \sqrt{\frac{\sum \Delta u^2}{n}} = \sqrt{\frac{239\,838}{232}} = 32{,}1,$$

$$\sigma_x = \sqrt{\frac{\sum \Delta x^2}{n}} = \sqrt{\frac{308\,717}{232}} = 36{,}5,$$

$$\sigma_y = \sqrt{\frac{\sum \Delta y^2}{n}} = \sqrt{\frac{95\,252}{232}} = 20{,}3,$$

$$\sigma_z = \sqrt{\frac{\sum \Delta z^2}{n}} = \sqrt{\frac{1062}{232}} = 2{,}1.$$

Затем приступаем к расчету уравнения регрессии по формуле (4.5) и после простых арифметических действий имеем:

$$u - 64 = 0{,}73\,(x - 82) + 0{,}54\,(y - 17) - 2{,}29\,(z - 21{,}7) \text{ или}$$

$$u - 64 = 0{,}73x - 59{,}86 + 0{,}54y - 9{,}18 - 2{,}29z + 49{,}69.$$

Окончательно уравнение искомой зависимости имеет вид

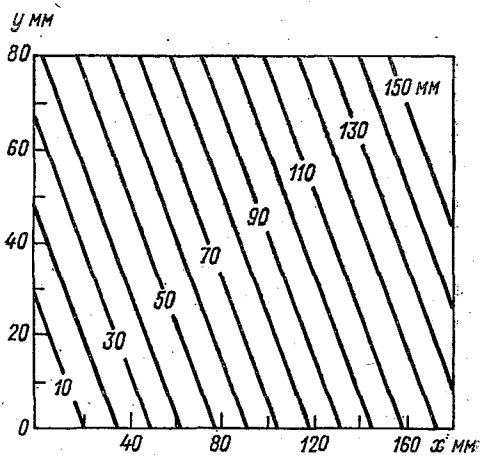
$$u = 0{,}73x + 0{,}54y - 2{,}29z + 44{,}65, \qquad (4.7)$$

где u — запасы влаги в метровом слое почвы под кукурузой к концу декады в период от выбрасывания метелки до молочной спелости; x — исходные запасы влаги в метровом слое почвы к началу декады; y — сумма осадков за декаду; z — средняя температура воздуха за декаду.

По данному уравнению (4.7) можно построить номограмму, с помощью которой легко найти u (рис. 4.1).

Однако на плоскости можно изобразить связь только трех переменных величин, поэтому графическую зависимость строят для трех величин u, x, y (при постоянном z), а для четвертой величины z рассчитывают по уравнению (4.7) поправки, которые учитывают при окончательном расчете u.

Обычно график рассчитывается для u, x и y при постоянном z, равном $\bar{z}$, т. е. при среднем арифметическом значении z. В нашем примере $\bar{z}=22{,}0$.



Средняя температура воздуха за декаду, °С	18	19	20	21	22	23	24	25	26
Поправка на запасы влаги, мм	9	7	5	2	0	-2	-5	-7	-9

Рис. 4.1. Номограмма для определения запасов влаги в метровом слое почвы под кукурузой к концу декады по исходным запасам влаги к началу декады (x), сумме осадков за декаду (y) и по средней температуре воздуха за декаду (в период от выбрасывания метелки до молочной спелости).

Таким образом, взяв $\bar{z}=22{,}0$, рассчитываем график связи трех переменных величин u, x и y, аналогично тому, как это было сделано в разделе 3.2, с той лишь разницей, что здесь берутся значения и интервалы запасов влаги для метрового слоя почвы (рис. 4.1).

Судя по данным наблюдений, использованным для расчета уравнения, запасы влаги в метровом слое почвы под кукурузой в этот период могут изменяться в пределах от 0 до 160 мм.

Возьмем интервалы для u, равные 10 мм, и найдем на графике линии, для которых u равно 10, 20, 30 мм и т. д. По оси абсцисс откладываем исходные запасы влаги к началу декады (x), по оси ординат — сумму осадков за декаду (y), в поле графика будем строить линии равных значений u с интервалом 10 мм при $\bar{z}=22{,}0$.

Задавая определенные значения u и находя одновременно при $y = 0$ и $z = 22{,}0$ значения x, а при $x = 0$ и $z = 22{,}0$ — значения y, получим на осях x и y точки, соответствующие данному значению u. Соединив эти точки, получим линию равных значений u. Таким же образом строим линии различных равных значений u с интервалом 10 мм. Получаем график связи u, x и y при $z = 22{,}0$. Следовательно, все расчеты запасов влаги по этому графику будут соответствовать температуре воздуха 22,0 °С. Для учета различных значений температуры находим поправки. Для $z = 22{,}0$ поправка на графике равна 0, так как он был построен при таких значениях температуры.

Рассчитываем уравнение (4.7) относительно u для различных значений z (через 1 °С) при одинаковых значениях x и y и таким образом получаем разные значения u только в зависимости от z.

Разности между значением u при $\bar{z} = 22$ °С и различными значениями z дадут нам поправки на температуру воздуха, которые сводятся по градациям в табличку под номограммой (см. рис. 4.1). Алгебраическая сумма значения u, снятого с графика, и поправки даст нам окончательную величину рассчитываемых запасов влаги u.

Расчеты парных коэффициентов корреляции r_{ux}, r_{uy}, r_{uz}, r_{xy}, r_{xz} и r_{yz}, а также средних квадратических отклонений σ_u, σ_x, σ_y и σ_z можно проводить также методом сгруппированных данных по частотам с введением условных единиц. Для нахождения каждого парного r рассчитываем таблицу, подобную табл. 2.7, подробное изложение расчетов которой дано в разделе 2.8. Следовательно, в общей сложности рассчитываем шесть таблиц и находим значения шести парных r и σ и средние арифметические величины $\bar{u}$, $\bar{x}$, $\bar{z}$. Затем, подставляя их в формулы для R и для u, приведенные в этом разделе, рассчитываем множественный коэффициент корреляции и уравнение регрессии четырех переменных.

4.3. ЗАВИСИМОСТЬ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ ОТ ОБЪЕМА ВЫБОРКИ И ЧИСЛА ПРЕДИКТОРОВ В УРАВНЕНИИ

Приведем данные о зависимости коэффициентов корреляции от объема выборки и числа предикторов в уравнениях, сравним их с коэффициентами корреляции в генеральной совокупности по Р. А. Фишеру [32].

Как было показано выше, коэффициент корреляции в большой степени зависит от объема выборки и при малом ее объеме может быть значительно завышенным. Принято считать на практике, что величины достаточно связаны между собой при $r \geqslant 0{,}6$. Часто это значение r принимают без учета объема выборки при малом числе случаев наблюдений.

Из рис. 4.2 *а* видно, что при простой корреляции для уравнения связи с одним предиктором коэффициент корреляции в генеральной совокупности (истинная корреляция) равен 0,6 при значении коэффициента корреляции 0,7 в выборке объемом из

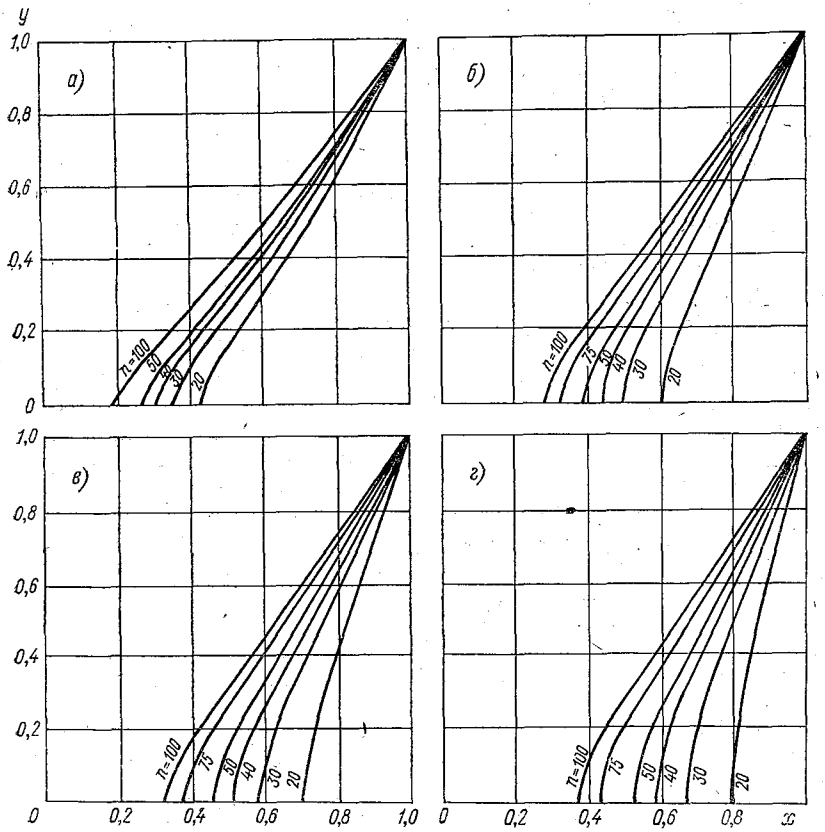


Рис. 4.2. Связь истинного коэффициента корреляции (y) с выборочным коэффициентом корреляции (x) в зависимости от длины выборки (n).

а) простая корреляция: $y=a+bx$; *б*) множественная корреляция при трех факторах: $y=a+b_1x_1+b_2x_2+b_3x_3$; *в*) множественная корреляция при пяти факторах: $y=a+b_1x_1+b_2x_2+b_3x_3+b_4x_4+b_5x_5$; *г*) множественная корреляция при семи факторах; $y=a+b_1x_1+b_2x_2+b_3x_3+b_4x_4+b_5x_5+b_6x_6+b_7x_7$

75—100 случаев наблюдений. При меньшем объеме выборок значение коэффициента корреляции должно быть больше. Так, при объеме выборки из 20 случаев наблюдений коэффициент корреляции должен быть не менее 0,8, тогда в генеральной совокупности при истинной корреляции его значение будет не менее 0,6.

При множественной корреляции для уравнения с тремя предикторами (рис. 4.2 *б*) множественный коэффициент корреляции в генеральной совокупности может быть равен 0,6 при множест-

венном коэффициенте корреляции 0,7 в выборке из 75—100 случаев и 0,82 в выборке из 20 случаев.

При множественной корреляции для уравнения с пятью или семью предикторами (рис. 4.2 *в*, *г*) множественный коэффициент корреляции в генеральной совокупности может быть равен 0,6 при множественном коэффициенте корреляции 0,7 в выборке из 75—100 случаев и 0,8 в выборке из 30—40 случаев.

Глава 5. НАХОЖДЕНИЕ ПАРАМЕТРОВ УРАВНЕНИЙ ЛИНЕЙНЫХ СВЯЗЕЙ МЕЖДУ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ

5.1. НАХОЖДЕНИЕ ПАРАМЕТРОВ УРАВНЕНИЙ ЛИНЕЙНОЙ СВЯЗИ МЕЖДУ ДВУМЯ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ

Если найдена линейная зависимость двух переменных величин, общий вид уравнения которой $y = ax + b$, то коэффициенты (параметры) уравнения, кроме определения через r, σ_x и σ_y, могут быть также найдены с помощью решения системы уравнений методом наименьших квадратов.

Допустим, что характер расположения точек на корреляционном поле показывает прямолинейную связь, теоретическая линия регрессии которой выразится уравнением $y = ax + b$. Параметры a и b, характерные для этой линии регрессии, неизвестны.

Из бесчисленного множества прямых линий, которые можно провести на плоскости по точкам корреляционного поля, следует выбрать одну, наилучшим образом соответствующую экспериментальным данным.

При корреляционной связи при одном и том же значении x будем иметь несколько значений y. Чтобы прямая регрессии ближе всего подходила к точкам, необходимы наименьшие отклонения ординат различных точек от данной прямой. Но отклонения ординат точек от прямой могут быть положительными и отрицательными в зависимости от того, где расположены точки — выше или ниже прямой. Возможен и такой случай, когда сумма отклонений $\sum (y_i - \bar{y}_x)$ окажется очень малой из-за различия знаков, а точки будут располагаться далеко от прямой. Чтобы избежать этого и исключить влияние знаков отклонений, достаточно искать наименьшее значение не суммы отклонений $\sum (y_i - \bar{y}_x)$, а суммы квадратов отклонений $\sum (y_i - \bar{y}_x)^2$.

Таким образом, для отыскания лучшей прямой регрессии данного корреляционного поля необходимо, чтобы

$$\sum (y_i - \bar{y}_x)^2 = \min, \qquad (5.1)$$

т. е. сумма квадратов отклонений фактических ординат (y_i) от ординат, вычисленных по уравнению прямой $(\bar{y}_x)$, должна быть наименьшей. В (5.1) выражено основное условие метода наименьших квадратов.

Обозначим сумму квадратов отклонений $\sum (y_i - \bar{y}_x)^2$ через f и заменяя $\bar{y}_x$ через $ax+b$, получим

$$f = \sum_{i=1}^{n} (y_i - \bar{y}_x)^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Так как величина f зависит от a и b, то ее можно рассматривать как функцию двух неизвестных параметров a и b.

Для нахождения ее минимума можно использовать известный прием дифференциального исчисления, заключающийся в отыскании двух частных производных первого порядка от функции f по a и b, в приравнивании их к нулю и определении значений a и b из полученных двух уравнений.

На основании этого будем иметь

$$\frac{\partial f}{\partial a} = 0, \quad \frac{\partial f}{\partial b} = 0 \text{ или}$$

$$\frac{\partial f}{\partial a} = 2 \sum_{i=1}^{n} (ax_i + b - y_i) x_i = 0,$$

$$\frac{\partial f}{\partial b} = 2 \sum_{i=1}^{n} (ax_i + b - y_i) = 0.$$

Раскрывая скобки и разбивая на почленные суммы, а затем вынося общие множители a и b за знаки сумм и заменяя $\sum b$ на nb, уравнения можно свести к следующему виду:

$$a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0,$$

$$a \sum_{i=1}^{n} x_i + nb - \sum_{i=1}^{n} y_i = 0. \qquad (5.2)$$

Таким образом, мы получили систему двух нормальных уравнений первой степени относительно неизвестных параметров a и b. Их можно записать также в виде

$$a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i,$$

$$a \sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i, \qquad (5.3)$$

где n — общее число случаев или число точек корреляционного поля. Значения y_i и x_i известны из наблюдений. Таким образом, параметры уравнения прямой регрессии a и b можно определить на основе данных уравнений по следующим формулам:

для коэффициента регрессии

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}; \tag{5.4}$$

для свободного члена

$$b = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}. \tag{5.5}$$

Для вычисления $\sum x_i$, $\sum y_i$, $\sum x_i^2$ и $\sum x_i y_i$ составляется таблица, аналогичная табл. 5.1.

Таблица 5.1

№ п/п	x_i	y_i	x_i^2	$x_i y_i$
n	Σx_i	Σy_i	Σx_i^2	$\Sigma x_i y_i$

Разделив числитель и знаменатель формулы (5.4) на n^2, получим

$$a = \frac{\frac{\sum x_i y_i}{n} - \frac{\sum x_i}{n}\frac{\sum y_i}{n}}{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}$$

или, что то же самое,

$$a = \frac{\overline{xy} - \bar{x}\bar{y}}{\bar{x}^2 - (\bar{x})^2}, \tag{5.6}$$

где $\overline{x^2} - (\overline{x})^2 = \sigma_x^2$, тогда

$$a = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x^2}. \tag{5.7}$$

Для взвешенных данных, если расчеты ведутся с учетом частот (m_{xy}), будем иметь следующие уравнения и формулы для определения a и b:

$$a\sum_{i=1}^{k} m_{x_i}x_i^2 + b\sum_{i=1}^{k} m_{x_i}x_i = \sum_{i=1}^{k} m_{x_i}x_iy_i,$$

$$a\sum_{i=1}^{k} m_{x_i}x_i + bn = \sum_{i=1}^{k} m_{x_i}y_i, \qquad (5.8)$$

откуда формула для коэффициента регрессии уравнения $y = ax+b$ имеет вид

$$a = \frac{n\sum_{i=1}^{k} m_{x_i}x_iy_i - \sum_{i=1}^{k} m_{x_i}x_i \sum_{i=1}^{k} m_{x_i}y_i}{n\sum_{i=1}^{k} m_{x_i}x_i^2 - \left(\sum_{i=1}^{k} m_{x_i}x_i\right)^2}; \qquad (5.9)$$

формула для свободного члена того же уравнения

$$b = \frac{\sum_{i=1}^{k} m_{x_i}y_i \sum_{i=1}^{k} m_{x_i}x_i^2 - \sum_{i=1}^{k} m_{x_i}x_i \sum_{i=1}^{k} m_{x_i}x_iy_i}{n\sum_{i=1}^{k} m_{x_i}x_i^2 - \left(\sum_{i=1}^{k} m_{x_i}x_i\right)^2}. \qquad (5.10)$$

Для расчетов параметров уравнений с учетом весов или частот составляется таблица, подобная табл. 5.2.

Таблица 5.2

№ п/п	x_i	y_i	m_{x_i}	$m_{x_i}x_i$	$m_{x_i}x_i^2$	$m_{x_i}y_i$	$m_{x_i}x_iy_i$
n				$\Sigma m_{x_i}x_i$	$\Sigma m_{x_i}x_i^2$	$\Sigma m_{x_i}y_i$	$\Sigma m_{x_i}x_iy_i$

Определив параметры a и b уравнения $y = ax+b$, строят по этому уравнению теоретическую линию регрессии, задавая различные значения x и получая рассчитанные значения y.

При расчете уравнений методом наименьших квадратов можно отсчитывать значения x и y от произвольного начала, а затем учесть это изменение отсчета в окончательной формуле. Это исключает из операций большие числа и уменьшает громоздкость расчетов.

Значения x и y можно отсчитывать от средних $\bar{x}$ и $\bar{y}$, если они не дробные числа, тогда расчеты значительно упрощаются.

В этом случае уравнение прямой регрессии будет иметь вид

$$y - \bar{y} = a'(x - \bar{x}) + b'. \qquad (5.11)$$

Нормальные уравнения для расчетов будут иметь вид

$$a' \sum (x - \bar{x}) + nb' = \sum (y - \bar{y}),$$

$$a' \sum (x - \bar{x})^2 + b' \sum (x - \bar{x}) = \sum (x - \bar{x})(y - \bar{y}); \qquad (5.12)$$

откуда

$$a' = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2},$$

$$b' = \frac{\sum (y - \bar{y})}{a' \sum (x - \bar{x})}.$$

5.2. НАХОЖДЕНИЕ ПАРАМЕТРОВ УРАВНЕНИЙ ЛИНЕЙНОЙ СВЯЗИ МЕЖДУ ТРЕМЯ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ

Линейную зависимость одной переменной величины (z) от двух других переменных величин (x и y) можно выразить уравнением

$$z = ax + by + c,$$

где a, b, c — неизвестные параметры уравнения, которые можно определить методом наименьших квадратов.

Для этого необходимо решить систему из трех нормальных уравнений

$$a \sum x + b \sum y + cn = \sum z,$$

$$a \sum xy + b \sum y^2 + c \sum y = \sum zy, \qquad (5.13)$$

$$a \sum x^2 + b \sum xy + c \sum x = \sum zx,$$

где n — общее число случаев наблюдений сочетания трех переменных величин, $\sum x$, $\sum y$ и $\sum z$ — суммы соответствующих значений каждой из переменных величин.

Для решения этих уравнений удобно пользоваться таблицей, подобной табл. 5.3. Получив из табл. 5.3 необходимые данные, составляем систему уравнений (5.13) и, решая ее, находим параметры a, b и c уравнения связи $z = ax + by + c$.

Таблица 5.3

№ п/п	z	x	y	zx	zy	xy	y^2	x^2
n	$\sum z$	$\sum x$	$\sum y$	$\sum zx$	$\sum zy$	$\sum xy$	$\sum y^2$	$\sum x^2$

5.3. НАХОЖДЕНИЕ ПАРАМЕТРОВ УРАВНЕНИЙ ЛИНЕЙНОЙ СВЯЗИ МЕЖДУ ЧЕТЫРЬМЯ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ. ПРИМЕР РАСЧЕТА УРАВНЕНИЯ ЗАВИСИМОСТИ УРОЖАЙНОСТИ ЯРОВОЙ ПШЕНИЦЫ ОТ КОЛИЧЕСТВА ОСАДКОВ В РАЗНЫЕ ПЕРИОДЫ ВЕГЕТАЦИИ И ИСПАРЕНИЯ

Линейное уравнение связи между четырьмя переменными величинами выразим формулой

$$y = a_0 + a_1 x + a_2 z + a_3 u,$$

где неизвестные параметры уравнения a_0, a_1, a_2, a_3 находятся путем решения системы четырех нормальных уравнений:

$$\begin{aligned} n a_0 + a_1 \sum x + a_2 \sum z + a_3 \sum u &= \sum y, \\ a_0 \sum x + a_1 \sum x^2 + a_2 \sum xz + a_3 \sum xu &= \sum xy, \\ a_0 \sum z + a_1 \sum xz + a_2 \sum z^2 + a_3 \sum zu &= \sum yz, \\ a_0 \sum u + a_1 \sum xu + a_2 \sum zu + a_3 \sum u^2 &= \sum yu. \end{aligned} \qquad (5.14)$$

Для вычисления параметров указанных уравнений удобно пользоваться таблицей, аналогичной табл. 5.4. Взяв необходимые данные из наблюдений и рассчитанные величины по таблице, составляем систему уравнений (5.14) и находим параметры уравнения.

При увеличении числа переменных увеличивается число членов уравнения связи, а следовательно, увеличивается и число

Таблица 5.4

№ п/п	x	z	u	y	xz	xu	zu	x^2	z^2	u^2	y^2	xy	zy	uy
n	$\sum x$	$\sum z$	$\sum u$	$\sum y$	$\sum xz$	$\sum xu$	$\sum zu$	$\sum x^2$	$\sum z^2$	$\sum u^2$	$\sum y^2$	$\sum xy$	$\sum zy$	$\sum uy$

нормальных уравнений, которые необходимо решить. Расчеты при большом числе членов (больше четырех) становятся очень громоздкими и трудоемкими, поэтому их следует проводить на электронно-вычислительных машинах.

При нахождении уравнения связи между многими переменными величинами следует брать только те величины или факторы, которые оказывают существенное влияние на искомую переменную величину, иначе учет несущественных факторов сильно загромождает расчетную работу при нахождении уравнения связи, но мало приближает к более полному изучению связи.

Приведем пример расчета параметров уравнения связи между четырьмя переменными величинами методом наименьших квадратов. Пример взят из работы В. С. Немчинова [21].

Найдем уравнение зависимости урожайности яровой пшеницы (y) от суммы осадков за период от начала сева до начала кущения пшеницы (x), от суммы осадков за период от начала кущения до начала цветения пшеницы (z) и от испарения за период от начала цветения пшеницы (u).

Для нахождения параметров уравнения указанной зависимости методом наименьших квадратов необходимо решить систему уравнений, где нужные суммы переменных величин рассчитываются по табл. 5.5.

Составив табл. 5.5 и рассчитав необходимые суммы, подставляем их значения в систему уравнений (5.14) и решаем ее:

$$25a_0 + 714a_1 + 1454a_2 + 4857a_3 = 231{,}0,$$

$$714a_0 + 32\,116a_1 + 46\,766a_2 + 121\,212a_3 = 8224{,}1,$$

$$1459a_0 + 46\,766a_1 + 106\,111a_2 + 251\,798a_3 = 15610{,}0,$$

$$4857a_0 + 121\,212a_1 + 251798a_2 + 1\,034\,455a_3 = 40229{,}7.$$

Делим все члены уравнений на коэффициенты при a_0:

$$a_0 + 28{,}56a_1 + 58{,}16a_2 + 194{,}28a_3 = 9{,}24,$$

$$a_0 + 44{,}9804a_1 + 65{,}4986a_2 + 169{,}7647a_3 = 11{,}5183,$$

$$a_0 + 32{,}0535a_1 + 72{,}7286a_2 + 172{,}5826a_3 = 10{,}6991,$$

$$a_0 + 24{,}9561a_1 + 51{,}8423a_2 + 212{,}9822a_3 = 8{,}2828.$$

Вычитаем из второго уравнения первое, из второго — третье и из третьего — четвертое. Получим три уравнения:

$$16{,}4204a_1 + 7{,}3385a_2 - 24{,}5153a_3 = 2{,}2783,$$

$$12{,}9269a_1 - 7{,}2300a_2 - 2{,}8179a_3 = 0{,}8192,$$

$$7{,}0974a_1 + 20{,}8863a_2 - 40{,}3996a_3 = 2{,}4163.$$

Делим все члены полученных уравнений на коэффициенты при a_1:

$$a_1 + 0{,}4469a_2 - 1{,}4930a_3 = 0{,}1387,$$

$$a_1 - 0{,}5637a_2 - 0{,}2180a_3 = 0{,}0634,$$

$$a_1 + 2{,}9428a_2 - 5{,}6924a_3 = 0{,}340.$$

Таблица 5.5

Таблица расчета сумм величин, необходимых для решения системы уравнений связи четырех переменных

Год	x	z	u	y	xz	xu	zu	x^2	z^2	u^2	y^2	xy	zy	uy
1905	21	76	211	10,5	1 596	4 431	16 036	441	5 776	44 521	110,25	220,5	798,0	2 215,5
1906	1	49	268	6,6	49	268	13 132	1	2 401	71 824	43,56	6,6	323,4	1 768,8
1907	44	42	163	10,5	1 848	7 172	6 846	1 936	1 764	26 569	110,25	462,0	441,0	1 711,5
1908	25	36	222	6,1	900	5 550	7 992	625	1 296	49 284	37,21	152,5	219,6	1 354,2
1909	37	58	126	11,1	2 146	4 662	7 308	1 369	3 364	15 876	123,21	410,7	643,8	1 398,6
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
1925	12	103	167	11,4	1 236	2 009	17 201	144	10 609	27 889	129,96	136,8	1 174,2	1 903,8
1926	66	74	180	13,3	4 884	11 880	13 320	4 356	5 476	32 400	176,89	877,8	984,2	2 394,0
1927	20	18	229	5,2	360	4 580	4 122	400	324	52 441	30,25	110,0	99,0	1 259,5
1928	66	71	135	14,1	4 686	8 910	9 585	4 356	5 041	18 225	198,81	930,6	1 001,1	1 908,5
1929	39	40	181	7,0	1 560	7 059	7 240	1 521	1 600	32 761	49,00	273,0	280,0	1 267,0
$\sum n = 25$	714	1454	4857	231,0	46 766	121 212	251 798	32 116	106 111	1 034 455	2512,29	8224,1	15 610,0	40 229,7

Примечание. Здесь y — урожайность яровой пшеницы, ц/га; x — сумма осадков от начала сева до начала кущения пшеницы, мм; z — сумма осадков от начала кущения до начала цветения пшеницы, мм; u — испарение от начала кущения до начала цветения пшеницы, мм.

Вычитаем из первого уравнения второе и из третьего — второе. Получим систему двух уравнений:

$$1{,}0106a_2 - 1{,}2750a_3 = 0{,}0753,$$
$$3{,}5065a_2 - 5{,}9104a_3 = 0{,}2770.$$

Делим члены уравнений на коэффициенты при a_2 и, вычитая из второго уравнения первое, находим параметр a_3:

$$\begin{array}{r} a_2 - 1{,}6855a_3 = 0{,}0790 \\ -\ a_2 - 1{,}2616a_3 = 0{,}0745 \\ \hline -0{,}4239a_3 = 0{,}0045 \end{array}$$
$$a_3 = -0{,}0106.$$

Подставляя значение параметра a_3 в одно из предыдущих уравнений, находим параметр a_2:

$$a_2 = 0{,}0745 - 1{,}2616 \cdot 0{,}0106 = 0{,}0611.$$

Подставляя значения параметров a_2 и a_3 в одно из уравнений с параметром a_1 находим параметр a_1:

$$a_1 + 0{,}0611 \cdot 0{,}4469 + 1{,}4930 \cdot 0{,}0106 = 0{,}1389,$$
$$a_1 + 0{,}0273 + 0{,}0158 = 0{,}1389,$$
$$a_1 = 0{,}0958.$$

Аналогичным образом находим параметр a_0:

$$a_0 + 28{,}56 \cdot 0{,}0958 + 58{,}16 \cdot 0{,}0611 - 194{,}28 \cdot 0{,}0106 = 9{,}24;$$
$$a_0 = 5{,}0098.$$

Подставляя полученные значения параметров в уравнение связи общего вида, получаем искомое корреляционное уравнение связи между четырьмя переменными величинами:

$$\bar{y}_{x,\,z,\,u} = 5{,}0098 + 0{,}0958x + 0{,}0611z - 0{,}0106u.$$

Указанный пример расчета параметров уравнения линейной связи между четырьмя величинами включает в себя также расчеты уравнений связи между тремя и двумя величинами (соответственно решение системы из трех и двух уравнений) методом наименьших квадратов.

Глава 6. НЕЛИНЕЙНЫЕ КОРРЕЛЯЦИОННЫЕ СВЯЗИ

6.1. НАХОЖДЕНИЕ ПАРАМЕТРОВ УРАВНЕНИЙ ПАРАБОЛИЧЕСКИХ СВЯЗЕЙ

Прямолинейная зависимость между x и y является наиболее простой формой связи. Однако часто связи между величинами являются более сложными, представляющие в графическом изо-

бражении кривую линию. Криволинейная, четко выраженная зависимость бывает видна и в корреляционной таблице по расположению частот.

Нередки случаи, когда при исследовании корреляционных связей точки $A_1(x_1y_1)$, $A_2(x_2y_2)\ldots$ $A_n(x_ny_n)$ располагаются на графике вблизи некоторой параболы. Следовательно, эмпирическую формулу необходимо искать следующего вида:

$$y = ax^2 + bx + c. \tag{6.1}$$

Это уравнение параболы второго порядка. Оно выражает параболическую зависимость, когда имеет место ускоренное возрастание или убывание одной величины (y) при равномерном возрастании другой (x). Парабола второго порядка — кривая с одним возможным максимумом или минимумом.

Метод наименьших квадратов дает возможность найти параметры указанного параболического уравнения a, b, c путем решения системы следующих нормальных уравнений:

$$\left.\begin{aligned} cn + b\sum x_i + a\sum x_i^2 &= \sum y_i, \\ c\sum x_i + b\sum x_i^2 + a\sum x_i^3 &= \sum x_iy_i, \\ c\sum x_i^2 + b\sum x_i^3 + a\sum x_i^4 &= \sum x_i^2y_i. \end{aligned}\right\} \tag{6.2}$$

Это система уравнений для несгруппированных данных простой перечневой таблицы, где для каждого значения x_i указано одно значение y_i (частота $m_{x_i} = 1$).

Определив необходимые суммы значений x и y по таблице, аналогичной табл. 6.1, и решив указанную систему уравнений, можно найти параметры уравнения искомой параболы, которая будет ближе всего располагаться около точек A_i.

Таблица 6.1

№ п/п	x_i	y_i	x_i^2	x_i^3	x_i^4	x_iy_i	$x_i^2y_i$
n	$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum x_i^3$	$\sum x_i^4$	$\sum x_iy_i$	$\sum x_i^2y_i$

Согласно основному условию наименьших квадратов, искомой параболой будет та, для которой сумма квадратов отклонений по ординате точек A_1, A_2, $A_3\ldots$ A_n от точек, лежащих на параболе, при тех же абсциссах будет наименьшей.

В случае сгруппированных данных по частотам m_{xy} система уравнений для нахождения параметров a, b и c имеет следующий вид:

$$\left.\begin{aligned} c\sum m_{x_i} + b\sum m_{x_i}x_i + a\sum m_{x_i}x_i^2 &= \sum m_{x_i}\bar{y}_{x_i},\\ c\sum m_{x_i}x_i + b\sum m_{x_i}x_i^2 + a\sum m_{x_i}x_i^3 &= \sum m_{x_i}x_i\bar{y}_{x_i},\\ c\sum m_{x_i}x_i^2 + b\sum m_{x_i}x_i^3 + a\sum m_{x_i}x_i^4 &= \sum m_{x_i}x_i^2\bar{y}_{x_i}. \end{aligned}\right\} \quad (6.3)$$

Для расчета сумм данной системы уравнений пользуются таблицей, подобной табл. 6.2. Вычисления параметров параболического уравнения громоздки, связаны с большими числами, так как переменная x берется в квадрате, в третьей и четвертой степенях. Эти вычисления значительно упрощаются, если провести преобразования и перенести начало координат в точку, наиболее близкую к вершине; тогда значения x_i и y_i значительно уменьшаются, а вычисления упрощаются:

$$x' = x - \alpha,$$

$$y' = y - \beta,$$

где α и β — координаты нового начала отсчета.

Делают и другие упрощения. Часто значения x уменьшают в несколько раз (в 10, 20 и т. д.), что значительно уменьшает громоздкость расчетов.

Кроме параболы второго порядка, при изучении связи между величинами применяются параболы более высоких порядков. Чем выше порядок параболы, тем точнее он воспроизводит опытные данные.

Уравнение параболы третьего порядка имеет вид

$$y = ax^3 + bx^2 + cx + d.$$

Для нахождения параметров уравнения a, b, c, d необходимо решить следующую систему нормальных уравнений:

$$\left.\begin{aligned} nd + c\sum x + b\sum x^2 + a\sum x^3 &= \sum y,\\ d\sum x + c\sum x^2 + b\sum x^3 + a\sum x^4 &= \sum yx,\\ d\sum x^2 + c\sum x^3 + b\sum x^4 + a\sum x^5 &= \sum yx^2,\\ d\sum x^3 + c\sum x^4 + b\sum x^5 + a\sum x^6 &= \sum yx^3. \end{aligned}\right\} \quad (6.4)$$

Расчеты необходимых сумм в уравнениях проводятся с помощью таблицы, аналогичной табл. 6.3.

При увеличении порядка параболы расчеты параметров уравнений становятся очень сложными и громоздкими.

Если мы имеем уравнение параболы некоторой степени e, то ее уравнение имеет вид $y = a_0 + a_1x + a_2x^2 + \ldots + a_ex^e$.

Таблица 6.2

№ п/п	x_i	m_{x_i}	$m_{x_i}x_i$	$m_{x_i}x_i^2$	$m_{x_i}x_i^3$	$m_{x_i}x_i^4$	$m_{x_i}\bar{y}_{x_i}$	$m_{x_i}\bar{y}_{x_i}x_i$	$m_{x_i}\bar{y}_{x_i}x_i^2$
n	$\sum x_i$	$\sum m_{x_i}$	$\sum m_{x_i}x_i$	$\sum m_{x_i}x_i^2$	$\sum m_{x_i}x_i^3$	$\sum m_{x_i}x_i^4$	$\sum m_{x_i}\bar{y}_{x_i}$	$\sum m_{x_i}\bar{y}_{x_i}x_i$	$\sum m_{x_i}\bar{y}_{x_i}x_i^2$

Таблица 6.3

n_i	x	y	x^2	x^3	x^4	x^5	x^6	yx	yx^2	yx^3
n	$\sum x$	$\sum y$	$\sum x^2$	$\sum x^3$	$\sum x^4$	$\sum x^5$	$\sum x^6$	$\sum yx$	$\sum yx^2$	$\sum yx^3$

Параметры данного уравнения a_1, $a_2 \ldots a_e$ находятся путем решения системы $e+1$ нормальных уравнений при группировке данных:

$$\left.\begin{aligned} a_0 \sum m_{x_i} + a_1 \sum m_{x_i} x_i + \ldots + a_e \sum m_{x_i} x_i^e &= \sum m_{x_i} \bar{y}_{x_i}, \\ a_0 \sum m_{x_i} x_i + a_1 \sum m_{x_i} x_i^2 + \ldots + a_e \sum m_{x_i} x_i^{e+1} &= \sum m_{x_i} x_i \bar{y}_{x_i}, \\ a_0 \sum m_{x_i} x_i^e + a_1 \sum m_{x_i} x_i^{e+1} + \ldots + a_e \sum m_{x_i} x_i^{2e} &= \sum m_{x_i} x_i^e \bar{y}_{x_i}. \end{aligned}\right\} \quad (6.5)$$

6.2. КОРРЕЛЯЦИОННОЕ ОТНОШЕНИЕ — МЕРА ТЕСНОТЫ СВЯЗИ ДЛЯ НЕЛИНЕЙНЫХ ЗАВИСИМОСТЕЙ

Вопрос тесноты связи переменных величин y и x в нелинейных зависимостях не может быть решена нахождением коэффициента корреляции r. Для этого находят корреляционное отношение (η).

Корреляционное отношение η является общим показателем тесноты связи y с x любой формы, в этом его преимущество перед коэффициентом корреляции r. При прямолинейной связи корреляционное отношение равно коэффициенту корреляции:

$$\eta = r.$$

Корреляционное отношение η называют также индексом корреляции.

Корреляционным отношением y по x называется отношение среднего квадратического отклонения условных средних $\bar{y}_x$ относительно общей средней $\bar{y}$ [межгруппового $\sigma(\bar{y}_x)$] к среднему квадратическому отклонению всех значений y относительно $\bar{y}$ (общему σ_y).

Для связи y по x

$$\eta_{y/x} = \frac{\sigma(\bar{y}_x)}{\sigma_y}, \quad (6.6)$$

$$\sigma(\bar{y}_x) = \sqrt{\frac{\sum^{x} m_x (\bar{y}_x - \bar{y})^2}{n}}, \quad \sigma_y = \sqrt{\frac{\sum^{y} m_y (y - \bar{y})^2}{n}}.$$

Для связи x по y

$$\eta_{x/y} = \frac{\sigma(\bar{x}_y)}{\sigma_x}, \quad (6.7)$$

$$\sigma(\bar{x}_y)=\sqrt{\frac{\sum^{y} m_y(\bar{x}_y-\bar{x})^2}{n}},\quad \sigma_x=\sqrt{\frac{\sum^{x} m_x(x-\bar{x})^2}{n}}.$$

Таким образом, корреляционное отношение будет разное для связи y по x и x по y:

$$\eta_{y/x}=\frac{\sigma(\bar{y}_x)}{\sigma_y}=\frac{1}{\sigma_y}\sqrt{\frac{1}{n}\sum m_x(\bar{y}_x-\bar{y})^2}=\frac{\sqrt{\sum m_x(\bar{y}_x-\bar{y})^2}}{\sqrt{\sum m_y(y-\bar{y})^2}},\tag{6.8}$$

$$\eta_{x/y}=\frac{\sigma(\bar{x}_y)}{\sigma_x}=\frac{1}{\sigma_x}\sqrt{\frac{1}{n}\sum m_y(\bar{x}_y-\bar{x})^2}=\frac{\sqrt{\sum m_y(\bar{x}_y-\bar{x})^2}}{\sqrt{\sum m_x(x-\bar{x})^2}}.\tag{6.9}$$

Следовательно, для расчета корреляционного отношения связи y по x необходимо определить квадратическое отклонение средних величин $(\bar{y})$ в строях $\sigma(\bar{y}_x)$ и общее квадратическое отклонение σ_y величин y.

Корреляционное отношение выражают также через дисперсии. В этом случае η дается следующее определение. Корреляционным отношением называется корень квадратный из отношения межгрупповой дисперсии $\sigma^2(\bar{y}_x)$ к общей дисперсии σ_y^2:

$$\eta_{y/x}=\sqrt{\frac{\sigma^2(\bar{y}_x)}{\sigma_y^2}},\tag{6.10}$$

где

$$\sigma^2(\bar{y}_x)=\frac{\sum(\bar{y}_x-\bar{y})^2 m_x}{n}$$

— межгрупповая дисперсия;

$$\sigma_y^2=\frac{\sum(y-\bar{y})^2 m_y}{n}$$

— общая дисперсия.

В двухмерной статистической совокупности могут быть три вида дисперсий.

1. Внутригрупповая дисперсия $\sigma_{x_i}(y)$ — дисперсия значений y в каждой группе по столбцам или строям корреляционной таблицы около условного среднего $\bar{y}_x$ для определенного значения x. При этом значения y меняются при неизменном x. Значит это внутригрупповое рассеяние y не зависит от x, а зависит от дру-

гих факторов. Значок x_i показывает, к какому условному распределению относится данное рассеяние y.

2. Межгрупповая дисперсия $\sigma^2(\bar{y}_x)$ — дисперсия условных средних $\bar{y}_x$ около общего среднего $\bar{y}$. В этом случае с изменением значений x условные средние $\bar{y}_x$ также меняют свои значения при переходе от одного условного распределения к другому.

3. Общая дисперсия $\sigma^2 y$ — дисперсия всех значений y (по всей таблице) около общего среднего $\bar{y}$. Таким образом, общая дисперсия $\sigma^2 y$ состоит из двух видов рассеяния y, двух дисперсий — внутригрупповой $\sigma^2_{x_i}(y)$ и межгрупповой $\sigma^2(\bar{y}_x)$: первая дисперсия не зависит от x, а вторая, наоборот, целиком зависит от x.

Корреляционное отношение η, выраженное через дисперсию, показывает, какую долю в общей мере рассеяния (дисперсии) занимает дисперсия данной величины y, возникающая вследствие влияния другой величины x.

Корреляционное отношение имеет следующие свойства.

1. Корреляционное отношение η всегда имеет значение между нулем и единицей

$$0 \leqslant \eta_{y/x} \leqslant 1.$$

2. Если корреляционное отношение равно нулю ($\eta_{y/x} = 0$ или $\eta_{x/y} = 0$), то между x и y не может быть никакой связи.

3. Если корреляционное отношение равно единице ($\eta_{y/x} = 1$ или $\eta_{x/y} = 1$), то между x и y существует точная функциональная связь.

4. С возрастанием корреляционного отношения от нуля до единицы увеличивается теснота связи x и y, переходя при $\eta = 1$ в функциональную.

5. Корреляционное отношение всегда больше или равно коэффициенту корреляции ($\eta \geqslant r$).

Средняя квадратическая ошибка корреляционного отношения равна

$$\sigma_\eta = \frac{1 - \eta^2}{\sqrt{n}}. \qquad (6.11)$$

6.3. ПРИМЕР РАСЧЕТА КОРРЕЛЯЦИОННОГО ОТНОШЕНИЯ И УРАВНЕНИЯ ПАРАБОЛИЧЕСКОЙ СВЯЗИ МЕЖДУ УРОЖАЙНОСТЬЮ ОЗИМОЙ ПШЕНИЦЫ И ВЕСЕННИМИ ЗАПАСАМИ ВЛАГИ В ПОЧВЕ

Нами был проведен анализ данных об урожайности озимой пшеницы и весенних запасах влаги в почве. Были выделены годы, когда озимая пшеница имела весной очень большое число стеблей

(2000—2600) на 1 м², и построено корреляционное поле связи урожайностей озимой пшеницы в эти годы с весенними запасами влаги (рис. 6.1).

На графике (рис. 6.1) видно, что урожайность растет при увеличении запасов влаги весной до 170—180 мм. В годы же, когда наблюдалось сочетание сильной загущенности посевов и больших запасов влаги весной, урожайность снижалась.

Эта зависимость хорошо описывается параболой.

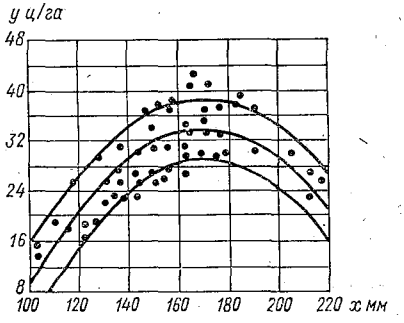


Рис. 6.1. Зависимость урожайности (y) озимой пшеницы от весенних запасов влаги (x) в метровом слое почвы в декаду перехода температуры воздуха через 5 °С весной при загущении посевов (число стеблей на 1 м² весной больше 2000).

Рассчитаем параметры уравнения для параболы второго порядка, общий вид уравнения которой

$$y = ax^2 + bx + c.$$

Будем вести расчеты методом группировки данных по частотам m_{xy}, который будет необходим и для расчета корреляционного отношения. Для определения параметров уравнения необходимо рассчитать систему уравнений вида

$$c\sum m_{x_i} + b\sum m_{x_i}x_i + a\sum m_{x_i}x_i^2 = \sum m_{x_i}\bar{y}_{x_i},$$
$$c\sum m_{x_i}x_i + b\sum m_{x_i}x_i^2 + a\sum m_{x_i}x_i^3 = \sum m_{x_i}x_i\bar{y}_{x_i},$$
$$c\sum m_{x_i}x_i^2 + b\sum m_{x_i}x_i^3 + a\sum m_{x_i}x_i^4 = \sum m_{x_i}x_i^2\bar{y}_{x_i}.$$

Следовательно, надо определить частоты m_{x_i} для y при определенных значениях x и рассчитать условные средние $\bar{y}_{x_i}$ по каждому значению середины интервала x (см. раздел 2.8). Для этого составляем корреляционную таблицу (табл. 6.4). Разбив на интервалы оси x и y (рис. 6.1), проводим вертикальные и горизонтальные линии по этим интервалам, которые дадут нам столбцы или строи табл. 6.4.

Подсчитываем на рис. 6.1 число точек в каждой клетке, соответствующей определенному интервалу x и y, и заносим это число в корреляционную таблицу — в графу с такими же интервалами. Получаем частоты m_{x_i} значений y для определенного значения середины интервала x_i.

Таблица 6.4

Корреляционная таблица связи между урожайностью озимой пшеницы (y) и весенними запасами влаги (x) при числе стеблей пшеницы 2000—2600 на 1 м² весной (связь параболическая)

Интервал y	Середина интервала y	Интервал x						Σm_{y_i}
		100—120	120—140	140—160	160—180	180—200	200—220	
		Середина интервала x						
		110	130	150	170	190	210	
12—16	14	2	0					2
16—20	18	2	2					4
20—24	22	0	3	1			2	6
24—28	26	1	3	6	1	2	2	15
28—32	30		2	3	5	1	1	12
32—36	34			1	5	0		6
36—40	38			4	2	3		9
40—44	42				3			3
Σm_{x_i}		5	10	15	16	6	5	57
$\bar{y}_{x_i}$		18,0	24,0	30,3	34,2	32,7	25,2	29

Подсчитываем сумму частот m_{x_i} или m_{xy} по вертикальным столбцам и находим для каждого столбца условное среднее $\bar{y}_{x_i}$ по формуле

$$\bar{y}_{x_i} = \frac{\sum^{y} m_{xy} y}{\sum m_{x_i}}:$$

$$\bar{y}_{x=110} = \frac{2 \cdot 14 + 2 \cdot 18 + 0 \cdot 22 + 1 \cdot 26}{5} = 18;$$

$$\bar{y}_{x=130} = \frac{0 \cdot 14 + 2 \cdot 18 + 3 \cdot 22 + 3 \cdot 26 + 2 \cdot 30}{10} = 24;$$

$$\bar{y}_{x=150} = \frac{1 \cdot 22 + 6 \cdot 26 + 3 \cdot 30 + 1 \cdot 34 + 4 \cdot 38}{15} = 30{,}3 \text{ и т. д.}$$

Затем находим общее среднее значение $\bar{y}$. По формуле $\bar{y} = \frac{\sum m_{x_i} \bar{y}_{x_i}}{n}$ находим среднее значение

$$\bar{y}_{\text{общ}} = \frac{5 \cdot 18 + 10 \cdot 24 + 15 \cdot 30{,}3 + 16 \cdot 34{,}2 + 6 \cdot 32{,}7 + 5 \cdot 25{,}2}{57} =$$

$$= \frac{1654}{57} = 29{,}02.$$

Далее по табл. 6.5 рассчитываем суммы, указанные в системе уравнений, и, подставляя их в эти уравнения, получаем:

$$57c + 9010b + 1\,466\,500a = 1654,$$

$$9010c + 1\,466\,500b + 245\,317\,000a = 266\,037,$$

$$1\,466\,500c + 245\,317\,000b + 42\,088\,570\,000a = 43\,824\,750.$$

Решаем указанную систему уравнений.

Делим каждое уравнение на коэффициенты при c:

$$c + 158{,}07b + 25728{,}07a = 29{,}02,$$

$$c + 162{,}76b + 27227{,}19a = 29{,}53,$$

$$c + 167{,}28b + 28700{,}01a = 29{,}88.$$

Вычитаем из второго и третьего уравнений первое и получаем систему из двух уравнений с двумя неизвестными:

$$4{,}69b + 1499{,}12a = 0{,}51,$$

$$9{,}21b + 2971{,}94a = 0{,}86.$$

Делим каждое уравнение на коэффициенты при b:

$$b + 319{,}642a = 0{,}109,$$

$$b + 322{,}686a = 0{,}093.$$

Вычитаем из второго уравнения первое, находим значение параметра a:

$$3{,}044a = -0{,}016,$$

$$a = -\frac{0{,}016}{3{,}044} = -0{,}0052.$$

Подставляя значение a в одно из предыдущих уравнений, находим значения параметра b:

$$b = 0{,}109 - 319{,}642\,(-0{,}0052) = 1{,}77.$$

Подставляя значения a и b в одно из уравнений с параметром c, находим значение c:

$$c = 29{,}02 - 158{,}07 \cdot (1{,}77) - 25728{,}07 - (-0{,}0052) = -116{,}978.$$

Вносим значения параметров a, b и c в уравнение параболы второй степени общего вида и получаем искомое уравнение:

$$y = -0{,}0052x^2 + 1{,}77x - 116{,}98.$$

Таблица расчета сумм величин, необходимых для опреде

x_i	m_{x_i}	$m_{x_i}x_i$	x_i^2	$m_{x_i}x_i^2$	x_i^3	$m_{x_i}x_i^3$
110	5	550	12 100	60 500	1 331 000	6 655 000
130	10	1300	16 900	169 000	2 197 000	21 970 000
150	15	2250	22 500	337 500	3 375 000	50 625 000
170	16	2720	28 900	462 400	4 913 000	78 608 000
190	6	1140	36 100	216 600	6 859 000	41 154 000
210	5	1050	44 100	220 500	9 261 000	46 305 000
Σ	57	9010		1 466 500		245 317 000

Находим теоретическую линию регрессии по данному уравнению. Задавая различные значения x, вычисляем по ним y:

при $x_1 = 110$ мм $y_1 = 14{,}8$ ц/га
„ $x_2 = 140$ „ $y_2 = 28{,}9$ „
„ $x_3 = 160$ „ $y_3 = 33{,}1$ „
„ $x_4 = 170$ „ $y_4 = 33{,}6$ „
„ $x_5 = 190$ „ $y_5 = 31{,}6$ „
„ $x_6 = 200$ „ $y_6 = 29{,}02$ „
„ $x_7 = 210$ „ $y_7 = 25{,}4$ „
„ $x_8 = 220$ „ $y_8 = 20{,}7$ „
„ $x_9 = 230$ „ $y_9 = 15{,}04$ „

Наносим полученные точки с координатами x_1, y_1, x_2y_2, ..., x_9y_9 на график и проводим по ним теоретическую кривую — параболу второго порядка (см. рис. 6.1).

Часто на графиках проводят также линии значений y, рассчитанных по уравнению с учетом ошибки уравнения ($y \pm S_y$). Между этими ограничивающими линиями бывает заключено более $^2/_3$ всех данных, вошедших в расчет уравнения связи.

Для нелинейной связи ошибка уравнения (S_y) вычисляется по формуле

$$S_y = \pm\sigma_y\sqrt{1-\eta^2},$$

где σ_y — среднее квадратическое отклонение, η — корреляционное отношение.

Найдем корреляционное отношение, показывающее тесноту найденной параболической связи:

$$\eta_{y/x} = \sqrt{\frac{\sigma^2(\bar{y}_x)}{\sigma_y^2}}.$$

Таблица 6.5

ления параметров уравнения параболы второй степени

x_i^4	$m_{x_i} x_i^4$	$\bar{y}_{x_i}$	$m_{x_i} \bar{y}_{x_i}$	$m_{x_i} x_i \bar{y}_{x_i}$	$m_{x_i} x_i^2 \bar{y}_{x_i}$
146 410 000	732 050 000	18	90	9 900	1 089 000
285 610 000	2 856 100 000	24	240	31 200	4 056 000
506 250 000	7 593 750 000	30,3	454,5	68 175	10 226 250
835 210 000	13 363 360 000	34,2	547,2	93 024	15 814 080
1 303 210 000	7 819 260 000	32,7	196,2	37 278	7 082 820
1 944 810 000	9 724 050 000	25,2	126	26 460	5 556 600
	42 088 570 000		1654	266 037	43 824 750

Рассчитываем межгрупповую дисперсию, пользуясь данными корреляционной табл. 6.4, по формуле

$$\sigma^2(\bar{y}_x) = \frac{\sum (\bar{y}_x - \bar{y})^2 m_{x_i}}{n};$$

$$\sigma^2(\bar{y}_x) = \frac{(18-29)^2 \cdot 5 + (24-29)^2 \cdot 10 + (30{,}3-29)^2 \cdot 15}{57} +$$

$$+ \frac{(34{,}2-29)^2 \cdot 16 + (32{,}7-29)^2 \cdot 6 + (25{,}2-29)^2 \cdot 5}{57} =$$

$$= \frac{1467{,}33}{57} = 25{,}74.$$

Находим общую дисперсию по формуле

$$\sigma_y^2 = \frac{\sum (y-\bar{y})^2 m_{y_i}}{n};$$

$$\sigma_y^2 = \frac{(14-29)^2 \cdot 2 + (18-29)^2 \cdot 4 + (22-29)^2 \cdot 6 + (26-29)^2 \cdot 15 + (30-29)^2 \cdot 12 + (34-29)^2 \cdot 6 + (38-29)^2 \cdot 9 + (42-29)^2 \cdot 3}{57} =$$

$$= \frac{2761}{57} = 48{,}44.$$

Рассчитываем корреляционное отношение

$$\eta = \sqrt{\frac{\sigma^2(\bar{y}_x)}{\sigma_y^2}} = \sqrt{\frac{25{,}74}{48{,}44}} = \sqrt{0{,}53} = 0{,}73.$$

Средняя ошибка корреляционного отношения равна

$$\sigma_\eta = \frac{1-\eta^2}{\sqrt{n}} = \frac{1-(0{,}73)^2}{\sqrt{57}} = \frac{0{,}47}{7{,}55} = \pm 0{,}06.$$

Таким образом, η находится в пределах $\eta \pm \sigma_\eta = \begin{cases} 0{,}79, \\ 0{,}67. \end{cases}$

Находим ошибку уравнения нелинейной регрессии по формуле

$$S_y = \pm\sigma_y\sqrt{1-\eta^2}.$$

Согласно расчетам, $\sigma_y^2 = 48{,}44$, следовательно, $\sigma_y = 6{,}96$, тогда

$$S_y = \pm 6{,}96\sqrt{1-(0{,}73)^2} = \pm 4{,}8 \text{ ц/га}.$$

Связь урожайности с весенними запасами влаги при загущении посевов (число стеблей больше 2000 на 1 м^2), как видим, несколько хуже, чем при густоте стеблей 1000—2000 на 1 м^2 весной (см. рис. 2.4).

6.4. НАХОЖДЕНИЕ ПАРАМЕТРОВ УРАВНЕНИЙ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ МЕЖДУ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ ГИПЕРБОЛИЧЕСКИХ, СТЕПЕННЫХ И ПОКАЗАТЕЛЬНЫХ КРИВЫХ

Часто связь между двумя переменными величинами выражена гиперболической кривой, уравнение которой в случае регрессии y по x имеет общий вид

$$y = \frac{a}{x} + b. \tag{6.12}$$

Как было изложено выше, метод наименьших квадратов применим для линейных и параболических связей; для других видов связей метод наименьших квадратов можно использовать для расчета параметров уравнения только после некоторых преобразований.

При нахождении параметров уравнения гиперболы обычно делают замену переменных и приводят уравнение гиперболической зависимости к линейному виду. После этого находят параметры линейного уравнения методом наименьших квадратов или через r, σ_x, σ_y, $\bar{x}$ и $\bar{y}$.

Зависимость $y = a/x + b$ можно свести к линейной, заменив $1/x$ на x'. Тогда уравнение будет иметь вид $y = ax' + b$. Это, как известно, уравнение прямой линии.

Для нахождения параметров a и b уравнения гиперболы методом наименьших квадратов необходимо значения x заменить значениями $x' = 1/x$ и составить систему необходимых уравнений.

Система нормальных уравнений для расчета параметров a и b в этом случае будет иметь вид

$$\left.\begin{aligned} nb + a\sum\frac{1}{x} &= \sum y, \\ b\sum\frac{1}{x} + a\sum\frac{1}{x^2} &= \sum\frac{y}{x} \end{aligned}\right\} \text{ или } \left.\begin{aligned} nb + a\sum x' &= \sum y, \\ b\sum x' + a\sum x'^2 &= \sum x'y. \end{aligned}\right\} \tag{6.13}$$

Расчеты, необходимые для решения данных уравнений, проводятся по таблице, аналогичной табл. 6.6. Подставляем полученные по таблице необходимые суммы в систему уравнений и решаем ее обычным образом: 1) делим каждый член обоих уравнений на коэффициенты при b и получаем два новых уравнения; 2) вычитаем из второго уравнения первое и находим параметр a; 3) подставляя в одно из уравнений полученное значение a, находим значение параметра b;

Таблица 6.6

n	x	y	$x'=\frac{1}{x}$	$(x')^2=\frac{1}{x^2}$	$x'y=\frac{y}{x}$
			$\Sigma x'$	$\Sigma x'^2$	$\Sigma x'y$

4) подставляя рассчитанные значения a и b в уравнение гиперболы, получаем окончательное искомое уравнение гиперболической связи; 5) задавая значения x в найденном уравнении гиперболы, получаем значения y и строим теоретическую кривую линию регрессии.

Параметры уравнения a и b при данной замене можно найти с помощью определителей

$$a=\frac{\begin{vmatrix}\sum y & \sum\frac{1}{x}\\ \sum\frac{y}{x} & \sum\frac{1}{x^2}\end{vmatrix}}{\begin{vmatrix}n & \sum\frac{1}{x}\\ \sum\frac{1}{x} & \sum\frac{1}{x^2}\end{vmatrix}},\qquad b=\frac{\begin{vmatrix}n & \sum y\\ \sum\frac{1}{x} & \sum\frac{y}{x}\end{vmatrix}}{\begin{vmatrix}n & \sum\frac{1}{x}\\ \sum\frac{1}{x} & \sum\frac{1}{x^2}\end{vmatrix}}.$$

Для гиперболы в случае сгруппированных данных по корреляционной таблице система уравнений имеет вид

$$\left.\begin{aligned} nb+a\sum m_{x_i}\frac{1}{x_i}&=\sum m_{x_i}\bar{y}_{x_i},\\ b\sum m_{x_i}\frac{1}{x_i}+a\sum m_{x_i}\frac{1}{x_i^2}&=\sum m_{x_i}\frac{1}{x_i}\bar{y}_{x_i}.\end{aligned}\right\}\qquad(6.14)$$

При гиперболической зависимости можно сделать другую замену. Приведя уравнение гиперболы $y=a/x+b$ к общему знаменателю, имеем $xy=a+bx$. Обозначаем xy через z и получаем урав-

нение прямой линии $z=a+bx$, параметры которого определяются системой уравнений

$$\left.\begin{aligned} na+b\sum x&=\sum z,\\ a\sum x+b\sum x^2&=\sum xz\end{aligned}\right\} \text{ или } \left.\begin{aligned} na+b\sum x&=\sum xy,\\ a\sum x+b\sum x^2&=\sum x^2y.\end{aligned}\right\} \quad (6.15)$$

Необходимые для решения уравнений суммы вычисляются по таблице, аналогичной табл. 6.7.

Таблица 6.7

№ п/п	x	y	xy	x^2	x^2y
n	Σx	Σy	Σxy	Σx^2	Σx^2y

Решая указанную систему уравнений, находим параметры a и b и, подставляя их в уравнение гиперболы, получаем искомое уравнение.

Значение параметров, или коэффициентов уравнения, a и b при данной замене можно вычислить также с помощью определителей

$$a=\frac{\begin{vmatrix}\sum xy & \sum x\\ \sum x^2y & \sum x^2\end{vmatrix}}{\begin{vmatrix}n & \sum x\\ \sum x & \sum x^2\end{vmatrix}}=\frac{\sum xy\sum x^2-\sum x\sum x^2y}{n\sum x^2-\sum x\sum x}, \quad (6.16)$$

$$b=\frac{\begin{vmatrix}n & \sum xy\\ \sum x & \sum x^2y\end{vmatrix}}{\begin{vmatrix}n & \sum x\\ \sum x & \sum x^2\end{vmatrix}}=\frac{n\sum x^2y-\sum xy\sum x}{n\sum x^2-\sum x\sum x}. \quad (6.17)$$

Корреляционная связь между двумя переменными величинами может быть выражена также степенными (6.18) и показательными (6.19) функциями

$$y=bx^a, \quad (6.18)$$

$$y=ab^x. \quad (6.19)$$

Степенные и показательные функции приводят к линейным путем логарифмирования формулы. Если степенную функцию $y=bx^a$ прологарифмировать, то получим

$$\lg y=a\lg x+\lg b.$$

Обозначив $y'=\lg y$, $x'=\lg x$, $\lg b=B$, получим уравнение прямой линии $y'=ax'+B$. Таким образом, соотношение между $\lg x$ и $\lg y$ является линейным, и можно применять уже известные методы расчета для уравнений простой прямолинейной регрессии.

При расчете параметров уравнения методом наименьших квадратов или через r и σ таблицы с данными наблюдений x и y пересчитывают на x' или y', т. е. берутся не сами величины, а их логарифмы.

При $y=b/x^a$ $\lg y=-a\lg x+\lg b$. Обозначим $y'=\lg y$, $x'=\lg x$, $B=\lg b$. Отсюда будем иметь $y'=-ax'+B$.

Если параметры a и B рассчитывают методом наименьших квадратов, то составляют следующую систему уравнений:

$$\left.\begin{aligned} n\lg b+a\sum\lg x&=\sum\lg y,\\ \lg b\sum\lg x+a\sum\lg x^2&=\sum\lg x\lg y\end{aligned}\right\} \qquad (6.20)$$

или

$$\left.\begin{aligned} nB+a\sum x'&=\sum y',\\ B\sum x'+a\sum x'^2&=\sum x'y'.\end{aligned}\right\} \qquad (6.21)$$

Необходимые суммы для решения системы уравнений (6.20) и (6.21) рассчитывают по таблице, аналогичной табл. 6.8.

Таблица 6.8

№ п/п	x	y	$x'=\lg x$	$y'=\lg y$	x'^2	$x'y'$
n			$\Sigma x'$	$\Sigma y'$	$\Sigma x'^2$	$\Sigma x'y'$

Если параметры уравнения рассчитываются через коэффициент корреляции r и средние квадратические отклонения σ_x, σ_y, то пользуются таблицей, подобной табл. 6.9.

Определив параметр B, из соотношения $\lg b=B$ находят параметр b.

Связь между двумя переменными может быть также выражена показательной (экспоненциальной) кривой, уравнение которой $y=ab^x$, где a и b — параметры уравнения, постоянные коэффициенты.

Подобными уравнениями выражаются связи между двумя величинами, когда увеличение функции y происходит значительно быстрее, чем увеличение аргумента x.

Таблица 6.9

№ п/п	x	y	$x'=\lg x$	$y'=\lg y$	$\Delta x'=x'-\bar{x}'$	$\Delta y'=y'-\bar{y}'$	$\Delta x'^2$	$\Delta y'^2$	$\Delta x'\Delta y'$
n			$\Sigma x'$	$\Sigma y'$			$\Sigma\Delta x'^2$	$\Sigma\Delta y'^2$	$\Sigma\Delta x'$ $\Sigma\Delta y'$

Путем логарифмирования уравнение кривой приводят к уравнению прямой линии:

$$\lg y = \lg a + x \lg b. \tag{6.22}$$

Обозначив $\lg y = y'$, $\lg a = A$, $\lg b = B$, получим уравнение прямой $y' = Bx + A$. Параметры данного уравнения можно определить методом наименьших квадратов. Для удобства расчетов составляется таблица, аналогичная табл. 6.10, затем рассчитываются системы уравнений:

$$\left.\begin{aligned} nA + B\sum x &= \sum y', \\ A\sum x + B\sum x^2 &= \sum xy' \end{aligned}\right\} \tag{6.23}$$

или

$$\left.\begin{aligned} n\lg a + \lg b\sum x &= \sum \lg y, \\ \lg a\sum x + \lg b\sum x^2 &= \sum x\lg y. \end{aligned}\right\} \tag{6.24}$$

Рассчитав по таблице суммы и решив систему указанных двух уравнений, находим параметры A и B. Из выражений $A = \lg a$ и $B = \lg b$ вычисляем a и b. Подставляя их в уравнение $y = ab^x$, находим искомое уравнение для данной нелинейной связи.

Таблица 6.10

№ п/п	x	y	$y'=\lg y$	x^2	$xy'=x\lg y$
n	Σx		$\Sigma y'$	Σx^2	$\Sigma xy'$

6.5. ПРИМЕР РАСЧЕТА ПАРАМЕТРОВ УРАВНЕНИЯ СТЕПЕННЫХ КРИВЫХ (ЗАВИСИМОСТЬ ПРОДОЛЖИТЕЛЬНОСТИ ПЕРИОДА ВСХОДЫ—КУЩЕНИЕ ОЗИМОЙ РЖИ ОТ ТЕМПЕРАТУРЫ ВОЗДУХА)

В агрометеорологии важным вопросом является вопрос нахождения зависимости межфазных периодов сельскохозяйственных культур от температуры воздуха. Для многих культур эта зависимость чаще всего бывает гиперболического или степенного вида.

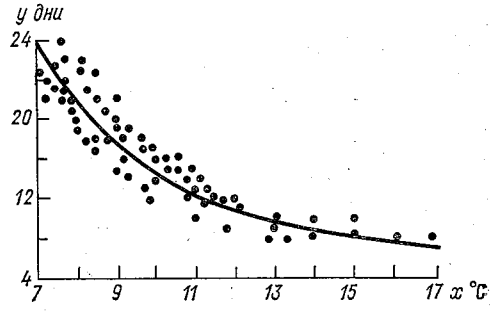


Рис. 6.2. Зависимость продолжительности периода всходы—кущение озимой ржи (y) от средней температуры воздуха за данный период (x) при достаточном увлажнении почвы (30—60 мм) в слое 0—20 см.

Рассмотрим найденную нами [31] связь продолжительности периода от всходов до начала кущения озимой ржи с температурой воздуха за этот период осенью. Прежде всего проводим анализ данных наблюдений. Так как продолжительность указанного межфазного периода зависит не только от температуры, но и от влажности почвы, то для нахождения данной зависимости исключим влияние влажности. Возьмем только те случаи, когда влажность почвы была оптимальной, т. е. когда влияние температуры было главным.

Наносим данные наблюдений на график, т. е. строим корреляционное поле (рис. 6.2).

По расположению точек на графике видно, что связь нелинейная, обратная, скорее всего, степенная (гиперболическая связь является также частным случаем степенной связи).

Уравнение обратной связи степенной кривой имеет вид

$$y = \frac{b}{x^a}. \tag{6.25}$$

Для нахождения параметров данного уравнения необходимо привести его к линейному виду. Прологарифмировав это уравнение, получим $\lg y = -a \lg x + \lg b$. Обозначаем $\lg y = y_1$, $\lg x = x_1$, $\lg b = B$ и получаем $y_1 = -ax_1 + B$ — уравнение прямой линии для обратной связи.

Если брать не сами значения x и y, а их логарифмы и наносить значения этих логарифмов в логарифмической шкале на график, то точки на графике будут располагаться в виде прямой линии, для которой надо найти уравнение связи.

Параметры искомого уравнения прямой регресии можно найти двумя путями: методом наименьших квадратов, построив таблицу, аналогичную табл. 6.8, и рассчитав систему уравнений (6.20) (см. раздел 6.4), или через коэффициент корреляции r и средние квадратические отклонения σ_x и σ_y по уравнению $y_1 - \bar{y}_1 = R(x - \bar{x}_1)$, где $R = r\sigma_{y_1}/\sigma_{x_1}$.

Проведем расчет параметров уравнения регрессии вторым путем через r и σ с помощью табл. 6.11.

В табл. 6.11 вносим под одним порядковым номером данные о средней температуре воздуха (x) за период от всходов до кущения озимой ржи и продолжительности (y) этого периода. Затем по таблицам логарифмов вычисляем логарифмы значений x и y. После этого находим суммы значений $x_1 = \lg x$ и $y_1 = \lg y$ и вычисляем средние арифметические значения $\bar{x}_1$ и $\bar{y}_1$:

$$\bar{x}_1 = \frac{65{,}5351}{66} = 0{,}9930, \quad \bar{y}_1 = \frac{78{,}6721}{66} = 1{,}1920.$$

Рассчитываем последующие графы табл. 6.11, находим отклонения каждого x_1 и y_1 от средней величины, квадраты отклонений, произведение отклонений и подсчитываем суммы этих величин (указанные действия были подробно изложены в разделе 2.9).

Получив необходимые суммы, рассчитываем коэффициент корреляции связи:

$$r = \frac{\sum \Delta x_1 \Delta y_1}{\sqrt{\sum \Delta x_1^2 \sum \Delta y_1^2}} = \frac{-0{,}9001}{\sqrt{0{,}5661 \cdot 1{,}6443}} = -0{,}93.$$

Ошибка коэффициента корреляции равна

$$\sigma_r = \pm \frac{1 - r^2}{\sqrt{n}} = \frac{1 - |(-0{,}93)^2}{\sqrt{66}} = \pm 0{,}017,$$

тогда предельные значения коэффициента корреляции

$$r \pm \sigma_r = -0{,}93 \pm 0{,}017 = \begin{cases} -0{,}95, \\ -0{,}91. \end{cases}$$

Коэффициент уравнения регресии

$$R = r \frac{\sigma_{y_1}}{\sigma_{x_1}}, \quad \sigma_{y_1} = \sqrt{\frac{\sum \Delta y_1^2}{n}},$$

$$\sigma_{x_1} = \sqrt{\frac{\sum \Delta x_1^2}{n}},$$

$$\frac{\sigma_{y_1}}{\sigma_{x_1}} = \frac{\sqrt{\sum \Delta y_1^2}}{\sqrt{\sum \Delta x_1^2}} = 1{,}70.$$

Таблица 6.11

Таблица для расчета связи между продолжительностью периода всходы—кущение озимой ржи (y) и средней температурой воздуха за данный период (x)

№ п/п	x	y	$\lg x = x_1$	$\lg y = y_1$	$\Delta x_1 = x_1 - \bar{x}_1$	$\Delta y_1 = y_1 - \bar{y}_1$	Δx_1^2	Δy_1^2	$\Delta x_1 \Delta y_1$	$\Delta x_1 + \Delta y_1$	$(\Delta x_1 + \Delta y_1)^2$
1	7,6	28	0,8808	1,4472	—0,1122	0,2552	0,0126	0,0651	—0,0286	0,1430	0,0204
2	7,7	26	0,8865	1,4150	—0,1065	0,2230	0,0113	0,0497	—0,0237	0,1165	0,0136
3	7,0	25	0,8451	1,3979	—0,1479	0,2059	0,0219	0,0424	—0,0304	0,0580	0,0034
4	7,4	25	0,8692	1,3979	—0,1238	0,2059	0,0153	0,0424	—0,0255	0,0821	0,0067
5	7,2	24	0,8573	1,3802	—0,1357	0,1882	0,0184	0,0354	—0,0255	0,0525	0,0027
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
62	13,9	7	1,1430	0,8451	0,1500	—0,3469	0,0225	0,1203	—0,0520	—0,1969	0,0388
63	15,0	10	1,1761	1,0000	0,1831	—0,1920	0,0335	0,0369	—0,0352	—0,0089	0,0000
64	15,0	8	1,1761	0,9031	0,1831	—0,2889	0,0335	0,0835	—0,0529	—0,1058	0,0112
65	16,1	8	1,2068	0,9031	0,2138	—0,2889	0,0457	0,0835	—0,0618	—0,0751	0,0056
66	17,0	8	1,2304	0,9031	0,2374	—0,2889	0,0564	0,0835	—0,0686	—0,0515	0,0026
Σ			65,5351	78,6721			0,5661	1,6443	—0,9001		0,4102

Определяем коэффициент уравнения регрессии

$$R = r\frac{\sigma_{y_1}}{\sigma_{x_1}} = -0{,}93 \cdot 1{,}70 = -1{,}58.$$

Подставляем в уравнение $y_1 - \bar{y}_1 = R(x_1 - \bar{x})$ значения R, $\bar{y}_1$, $\bar{x}_1$:

$$y_1 - 1{,}1920 = -1{,}581\,(x_1 - 0{,}9930)$$

и получаем искомое уравнение связи:

$$y_1 = 2{,}76 - 1{,}58x_1 \text{ или } \lg y = 2{,}76 - 1{,}58 \lg x.$$

По таблице антилогарифмов любого математического справочника находим b ($\lg b = 2{,}76$, $b = 575$).

Таким образом, искомое уравнение нелинейной связи можно записать $y = 575/x^{1{,}58}$. Задавая различные значения x, рассчитываем ряд значений y и строим теоретическую кривую связи. Эти расчеты ведутся по формуле $\lg y = 2{,}76 - 1{,}58 \lg x$, так как по формуле $y = 575/x^{1{,}58}$, где участвует степень 1,58, рассчитывать без логарифмирования значения y нельзя. Задаем произвольно пять значений x, получаем пять значений y:

$x_1 = 8$ $\lg y = 2{,}76 - 1{,}58 \lg 8 = 2{,}76 - 1{,}58 \cdot 0{,}90 = 1{,}34$; $y_1 = 21{,}9$;
$x_2 = 10$ $\lg y = 2{,}76 - 1{,}58 \lg 10 = 2{,}76 - 1{,}58 \cdot 1{,}00 = 1{,}18$; $y_2 = 15{,}1$;
$x_3 = 12$ $\lg y = 2{,}76 - 1{,}58 \lg 12 = 2{,}76 - 1{,}58 \cdot 1{,}08 = 1{,}05$; $y_3 = 11{,}2$;
$x_4 = 14$ $\lg y = 2{,}76 - 1{,}58 \lg 14 = 2{,}76 - 1{,}58 \cdot 1{,}15 = 0{,}94$; $y_4 = 8{,}7$;
$x_5 = 16$ $\lg y = 2{,}76 - 1{,}58 \lg 16 = 2{,}76 - 1{,}58 \cdot 1{,}20 = 0{,}86$; $y_5 = 7{,}2$.

Наносим точки со значениями x_1y_1, $x_2y_2 \ldots x_5y_5$ на корреляционное поле рис. 6.2 и проводим по ним теоретическую линию кривой регрессии, по которой впоследствии без расчета по уравнению можно снимать значения y по заданным значениям x.

Для определения логарифмов, различных степеней переменных, величин x и y и других значений следует пользоваться математическими таблицами, где указанные величины уже рассчитаны.

Кроме этого, вычисления следует проводить при помощи счетных машин, не заполняя в таблицах графы степеней и произведений по каждому порядковому номеру, а получая на машине и записывая сразу значения сумм степеней и произведений. Это значительно уменьшит объем работы по расчетам параметров уравнений корреляционных связей.

В данной части книги изложены основы методов математической статистики и статистического анализа применительно к агрометеорологии, уделено внимание вопросам, которые чаще всего бывает необходимо знать специалисту агрометеорологу. Изложенные вопросы могут быть полезны также и специалистам, работающим в области сельского хозяйства, где часто приходится иметь дело с многочисленными данными различных опытов и

наблюдений, которые необходимо подвергнуть статистическому анализу и статистической обработке.

Так как еще далеко не всегда и не везде имеется возможность вести расчеты на ЭВМ, то в первой части книги примеры статистической обработки данных наблюдений, нахождения параметров уравнений и вычисления показателей тесноты связей даны при условии применения малой вычислительной техники.

Часть II

НЕКОТОРЫЕ АСПЕКТЫ СОВРЕМЕННЫХ МЕТОДОВ РЕГРЕССИОННОГО МОДЕЛИРОВАНИЯ

Глава 7. ОПЕРАЦИИ С МАТРИЦАМИ

Как было наглядно представлено в первой части, на практике агрометеорологи часто сталкиваются с различного вида взаимоотношениями, в основе которых лежат прямые пропорциональные зависимости между двумя или несколькими величинами. Их математическое описание приводит к системам линейных алгебраических уравнений, которые удобно записывать в матричной форме. Общепринятым в статистическом анализе является язык матричной алгебры, который позволяет упрощать вычисления при наличии нескольких переменных. Существует огромное количество теоретических результатов и примеров использования на практике в различных сферах естествознания элементов матричной алгебры, однако в этой главе будут затронуты лишь самые необходимые аспекты, полезные для применения во множественном регрессионном анализе.

7.1. ОПРЕДЕЛЕНИЕ И ОСНОВНЫЕ СВОЙСТВА МАТРИЦ

Матрица — это таблица элементов, расположенных по строкам и столбцам. Элементы обычно закрывают в прямые скобки и обозначают заглавной буквой латинского алфавита, например:

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix}.$$

Каждый элемент определяется местом в матрице: номерами строки и столбца, причем первым индексом фиксируется номер строки. В матрице

$$\mathbf{B}=\begin{bmatrix} 2 & -0{,}8 \\ 4 & 0 \end{bmatrix}$$

значения элементов $b_{11}=2$, $b_{12}=-0{,}8$, $b_{21}=4$, $b_{22}=0$. Элементы матрицы **A** принято обозначать a_{ij}, т. е. указывать с помощью индекса i номер строки, а индекса j — номер столбца, в которых расположен элемент.

Матрица характеризуется своими размерами: числом строк и числом столбцов. Так, матрица **A** имеет размер (4, 3), поскольку

у нее четыре строки и три столбца, матрица **B** имеет одинаковое число столбцов и строк, ее размер — (2, 2). Матрицы с одинаковым числом столбцов и строк называются квадратными, они играют особо важную роль в регрессионном анализе. Частным случаем матриц являются векторы. Матрица, которая имеет только один столбец, называется вектором-столбцом:

$$\mathbf{C} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix}.$$

Аналогично вектору-столбцу можно определить вектор-строку как матрицу с одной строкой, например:

$$\mathbf{D} = [d_1, \quad d_2, \quad d_3].$$

Элементы вектора нумеруют одним индексом, указывающим номер места, занимаемого элементом.

Если в матрице только одна строка и один столбец, то такой элемент называют скаляром, и это — простое число.

Еще раз подчеркнем, что матрица — это наглядный способ представления элементов в виде таблицы, состоящей из строк и столбцов.

Если индексы строк и столбцов у элементов одинаковы, то для квадратной матрицы говорят о диагональных элементах, так как они лежат на ее главной диагонали.

Квадратная матрица **B**, у которой ненулевые элементы находятся только на главной диагонали и выше ее, т. е. $b_{ij}=0$ при $i > j$, называется верхней треугольной, например:

$$\mathbf{B} = \begin{bmatrix} 2 & 2 & 7 & 6 \\ \underset{(2,\,1)}{0} & 3 & 2 & 5 \\ \underset{(3,\,1)}{0} & \underset{(3,\,2)}{0} & 1 & 7 \\ \underset{(4,\,1)}{0} & \underset{(4,\,2)}{0} & \underset{(4,\,3)}{0} & 3 \end{bmatrix}.$$

В скобках у нулевых элементов указаны их индексы i и j, причем всегда $i > j$. Наоборот, при $b_{ij}=0$, когда $i < j$, матрица называется нижней треугольной.

Квадратная матрица, у которой все диагональные элементы b_{ii} не равны нулю, а все остальные элементы равны нулю, т. е. $b_{ij}=0$ при всех $i \neq j$, называется диагональной, например:

$$\mathbf{B} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 5 \end{bmatrix}.$$

Если все диагональные элементы в $b_{ii}=1$, матрица называется единичной, например:

$$\mathbf{B}=\begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix}.$$

Эта матрица выполняет при алгебраических операциях роль обычной единицы.

Нулевая матрица имеет все нулевые элементы и может быть неквадратной.

Квадратная матрица, у которой $a_{ij}=a_{ji}$, называется симметричной; например, элементы матрицы **A** симметричны относительно главной диагонали:

$$\mathbf{A}=\begin{bmatrix}5&1&21\\1&4&3\\21&3&1\end{bmatrix}$$

(так элементы $a_{21}=a_{12}=1$, $a_{31}=a_{13}=21$).

7.2. ОПЕРАЦИИ НАД МАТРИЦАМИ

Матрицы **A** и **B** считаются равными, если все соответствующие элементы одной матрицы равны элементам другой: $\mathbf{A}=\mathbf{B}$ при $a_{ij}=b_{ij}$ для всех i и j. Единичные матрицы одного размера равны между собой.

Важной операцией над матрицами является транспонирование, при котором в матрице меняются местами строки и столбцы. Транспонированная матрица **A** размера (m, n) обозначается $\mathbf{A}'$, ее размер становится (n, m), причем $a'_{ij}=a_{ji}$. Например, если

$$\mathbf{A}=\begin{bmatrix}2&-1\\1&2\\3&4\end{bmatrix},$$

то

$$\mathbf{A}'=\begin{bmatrix}2&1&3\\-1&2&4\end{bmatrix}.$$

Первый столбец матрицы **A** стал первой строкой транспонированной матрицы $\mathbf{A}'$, второй столбец — второй строкой и $a_{21}=a_{12}=-1$.

Транспонирование вектора-столбца дает в результате вектор-строку и наоборот; транспонирование верхней треугольной матрицы приводит к нижней треугольной матрице:

$$\mathbf{B}=\begin{bmatrix}1 & 2 & 1\\ 0 & 3 & 2\\ 0 & 0 & 7\end{bmatrix},\quad \mathbf{B}'=\begin{bmatrix}1 & 0 & 0\\ 2 & 3 & 0\\ 1 & 2 & 7\end{bmatrix}.$$

Естественно, если два раза «перевернуть», т. е. транспонировать матрицу, то она останется такой, какой и была до проведения операций транспонирования: $(\mathbf{A}')'=\mathbf{A}$. Например:

$$\mathbf{A}=\begin{bmatrix}2 & 3\\ 4 & 5\end{bmatrix},\quad \mathbf{A}'=\begin{bmatrix}2 & 4\\ 3 & 5\end{bmatrix},\quad (\mathbf{A}')'=\begin{bmatrix}2 & 3\\ 4 & 5\end{bmatrix}.$$

Если матрица **A** симметрична, то

$$\mathbf{A}'=\mathbf{A}.$$

Аналогично алгебраическим действиям над числами можно определить соответствующие правила действий над матрицами.

Сложение матриц можно определить как

$$\mathbf{A}+\mathbf{B}=\mathbf{C},$$

если элементы c_{ij} матрицы **C** получают как сумму соответствующих элементов a_{ij} и b_{ij}: $a_{ij}+b_{ij}=c_{ij}$.

Например, надо сложить две матрицы **A** и **B**:

$$\mathbf{A}=\begin{bmatrix}2 & 3\\ 1 & 5\\ -1 & 6\end{bmatrix},\quad \mathbf{B}=\begin{bmatrix}2 & 7\\ 0 & 3\\ 1 & -2\end{bmatrix};$$

тогда

$$\mathbf{C}=\mathbf{A}+\mathbf{B}=\begin{bmatrix}2 & 3\\ 1 & 5\\ -1 & 6\end{bmatrix}+\begin{bmatrix}2 & 7\\ 0 & 3\\ 1 & -2\end{bmatrix}=\begin{bmatrix}4 & 10\\ 1 & 8\\ 0 & 4\end{bmatrix}.$$

Очевидно, что складываемые матрицы должны быть одинакового размера.

При сложении матриц выполняются следующие свойства:

1. $\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$

$$\begin{bmatrix}1 & 2\\ 0 & -1\end{bmatrix}+\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}=\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}+\begin{bmatrix}1 & 2\\ 0 & -1\end{bmatrix}=\begin{bmatrix}1 & 3\\ 0 & -1\end{bmatrix}.$$

2. $(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$

$$\begin{bmatrix}0 & 1\\ 1 & 1\end{bmatrix}+\left(\begin{bmatrix}2 & 2\\ 0 & 0\end{bmatrix}+\begin{bmatrix}1 & 3\\ 2 & 1\end{bmatrix}\right)=\begin{bmatrix}0 & 1\\ 1 & 1\end{bmatrix}+\begin{bmatrix}3 & 5\\ 2 & 1\end{bmatrix}=\begin{bmatrix}3 & 6\\ 3 & 2\end{bmatrix};$$

$$\left(\begin{bmatrix}0 & 1\\ 1 & 1\end{bmatrix}+\begin{bmatrix}2 & 2\\ 0 & 0\end{bmatrix}\right)+\begin{bmatrix}1 & 3\\ 2 & 1\end{bmatrix}=\begin{bmatrix}2 & 3\\ 1 & 1\end{bmatrix}+\begin{bmatrix}1 & 3\\ 2 & 1\end{bmatrix}=\begin{bmatrix}3 & 6\\ 3 & 2\end{bmatrix}.$$

3. $(\mathbf{A}+\mathbf{B})'=\mathbf{A}'+\mathbf{B}'$

$$\left(\begin{bmatrix}1 & 0\\ 3 & 2\end{bmatrix}+\begin{bmatrix}-1 & 1\\ 0 & -1\end{bmatrix}\right)'=\begin{bmatrix}0 & 1\\ 3 & 1\end{bmatrix}'=\begin{bmatrix}0 & 3\\ 1 & 1\end{bmatrix};$$

$$\begin{bmatrix}1 & 0\\ 3 & 2\end{bmatrix}'+\begin{bmatrix}-1 & 1\\ 0 & -1\end{bmatrix}'=\begin{bmatrix}1 & 3\\ 0 & 2\end{bmatrix}+\begin{bmatrix}-1 & 0\\ 1 & -1\end{bmatrix}=\begin{bmatrix}0 & 3\\ 1 & 1\end{bmatrix}.$$

Операция вычитания подчиняется этим же правилам.

Если матрица $\mathbf{A}$ имеет размер (m, n), а матрица $\mathbf{B}$ — размер (n, p), то произведение матриц $\mathbf{AB}$ определяется как новая матрица размера (m, p), в которой элемент, стоящий на пересечении i-й строки и j-го столбца, равен

$$c_{ij}=\sum_{k=1}^{n} a_{ik}b_{kj}.$$

Значит, для того, чтобы получить элемент новой матрицы с индексом ij, надо сложить попарные произведения элементов i-й строки первой матрицы и соответствующих элементов j-го столбца второй матрицы. Очевидно, при этом необходимо, чтобы количество столбцов первой матрицы было равно количеству строк второй. Размер результирующей матрицы $\mathbf{C}$ будет равен числу строк матрицы $\mathbf{A}$ и числу столбцов матрицы $\mathbf{B}$. Например, пусть

$$\mathbf{A}=\begin{bmatrix}1 & 2 & 4\\ 3 & 1 & 2\end{bmatrix},\quad \mathbf{B}=\begin{bmatrix}1 & 1 & -1\\ 2 & 1 & 2\\ 1 & 3 & 4\end{bmatrix}.$$

Матрица $\mathbf{A}$ имеет три столбца и две строки, т. е. ее размер (2, 3), матрица $\mathbf{B}$ имеет размер (3, 3); число столбцов матрицы $\mathbf{A}$ равно числу строк матрицы $\mathbf{B}$ и поэтому их можно перемножать. Элементы матрицы $\mathbf{C}=\mathbf{AB}$ получают следующим образом:

$$c_{11}=a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31}=1\cdot 1+2\cdot 2+4\cdot 1=9,$$
$$c_{12}=a_{11}b_{12}+a_{12}b_{22}+a_{13}b_{32}=1\cdot 1+2\cdot 1+4\cdot 3=15,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$c_{23}=a_{21}b_{13}+a_{22}b_{23}+a_{23}b_{33}=3\cdot(-1)+1\cdot 2+2\cdot 4=7.$$

В результате полного перемножения строк и столбцов получим матрицу **C** размера (2, 3):

$$\mathbf{C}=\begin{bmatrix}9 & 15 & 19\\ 7 & 9 & 7\end{bmatrix}.$$

При умножении матриц выполняются следующие свойства.

1. Обычно $\mathbf{AB}\neq\mathbf{BA}$. Если матрицы не квадратные, то число столбцов матрицы **A** должно равняться числу строк матрицы **B**, т. е. их размер должен быть (m, k) и (k, n).

Например, пусть

$$\mathbf{A}=\begin{bmatrix}1 & 1\\ 2 & 3\end{bmatrix},\quad \mathbf{B}=\begin{bmatrix}2 & 3\\ 1 & 4\end{bmatrix};$$

тогда

$$\mathbf{AB}=\begin{bmatrix}1\cdot 2+1\cdot 1 & 1\cdot 3+1\cdot 4\\ 2\cdot 2+3\cdot 1 & 2\cdot 3+3\cdot 4\end{bmatrix}=\begin{bmatrix}3 & 7\\ 7 & 18\end{bmatrix},$$

$$\mathbf{BA}=\begin{bmatrix}2\cdot 1+3\cdot 2 & 2\cdot 1+3\cdot 3\\ 1\cdot 1+4\cdot 2 & 1\cdot 1+4\cdot 3\end{bmatrix}=\begin{bmatrix}8 & 11\\ 9 & 13\end{bmatrix},$$

т. е. $\mathbf{AB}\neq\mathbf{BA}$.

2. $(\mathbf{AB})\mathbf{C}=\mathbf{A}(\mathbf{BC})$.

3. $\mathbf{A}(\mathbf{B}+\mathbf{C})=\mathbf{AB}+\mathbf{AC}$.

Второе и третье свойства матричного умножения вполне понятные и выполняются, как для обычных чисел.

4. $(\mathbf{AB})'=\mathbf{B}'\mathbf{A}'$.

Пример выполнения четвертого свойства.

Пусть

$$(\mathbf{AB})'=\left(\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\begin{bmatrix}0 & 1 & 2\\ 1 & 2 & 1\end{bmatrix}\right)'=\begin{bmatrix}1\cdot 0+2\cdot 1 & 1\cdot 1+2\cdot 2\\ 3\cdot 0+4\cdot 4 & 3\cdot 1+4\cdot 2\\ 1\cdot 2+2\cdot 1 & 3\cdot 2+4\cdot 1\end{bmatrix}'=\left(\begin{bmatrix}2 & 5 & 4\\ 4 & 11 & 10\end{bmatrix}\right)'=\begin{bmatrix}2 & 4\\ 5 & 11\\ 4 & 10\end{bmatrix},$$

тогда

$$\mathbf{A}'\mathbf{B}'=\left(\begin{bmatrix}0 & 1 & 2\\ 1 & 2 & 1\end{bmatrix}\right)'\left(\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\right)'=\begin{bmatrix}0 & 1\\ 1 & 2\\ 2 & 1\end{bmatrix}\begin{bmatrix}1 & 3\\ 2 & 4\end{bmatrix}=$$

$$=\begin{bmatrix}0\cdot 1+1\cdot 2 & 0\cdot 3+1\cdot 4\\ 1\cdot 1+2\cdot 2 & 1\cdot 3+2\cdot 4\\ 2\cdot 1+1\cdot 2 & 2\cdot 3+1\cdot 4\end{bmatrix}=\begin{bmatrix}2 & 4\\ 5 & 11\\ 4 & 10\end{bmatrix}.$$

Существует удобный способ выяснить, можно ли перемножить матрицы и определить размер итоговой матрицы. В приведенном

ниже выражении под каждой матрицей указан ее размер (первый индекс — количество строк, второй — количество столбцов):

$$\underset{(m,\,n)}{\mathbf{A}}\ \underset{(n,\,p)}{\mathbf{B}}\ \underset{(p,\,q)}{\mathbf{C}} = \underset{(m,\,q)}{\mathbf{D}}.$$

Для того чтобы такое перемножение было осуществимо, необходимо равенство «соседних» индексов размера. Размер матрицы будет равен «внешним» индексам из цепочки $m, \ldots, q$.

Заметим, что при умножении матрицы на вектор-столбец получается вектор-столбец:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & 3 \\ 5 & 4 & -1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix};$$

$$\mathbf{Ab} = \begin{bmatrix} 1\cdot 2 + 2\cdot 3 + 1\cdot 1 \\ 0\cdot 2 + 2\cdot 3 + 3\cdot 1 \\ 5\cdot 2 + 4\cdot 3 + (-1)\cdot 1 \end{bmatrix} = \begin{bmatrix} 9 \\ 9 \\ 21 \end{bmatrix}.$$

При умножении вектора-строки на матрицу получается вектор-строка:

$$\mathbf{c} = [2 \quad 4 \quad 1], \quad \mathbf{D} = \begin{bmatrix} 2 & 1 & 1 & 2 \\ 1 & 4 & 0 & 0 \\ 3 & 5 & 1 & 0 \end{bmatrix};$$

$$\mathbf{cD} = [2\cdot 2 + 4\cdot 1 + 1\cdot 3 \quad 2\cdot 1 + 4\cdot 4 + 1\cdot 5 \quad 2\cdot 1 + 4\cdot 0 + 1\cdot 1$$
$$2\cdot 2 + 4\cdot 0 + 1\cdot 0] = [11 \quad 23 \quad 3 \quad 4].$$

При умножении вектора-строки на вектор-столбец получается скаляр:

$$\mathbf{a} = [1 \quad 0 \quad 2 \quad 4], \quad \mathbf{b} = \begin{bmatrix} 2 \\ 2 \\ 1 \\ 5 \end{bmatrix};$$

$$\mathbf{ab} = [1\cdot 2 + 0\cdot 2 + 2\cdot 1 + 4\cdot 5] = [24] = 24.$$

Операция умножения матрицы на число или скаляр определяется как умножение каждого элемента матрицы на это число. Например, если

$$\mathbf{A} = \begin{bmatrix} -2 & 0 \\ 5 & 7 \end{bmatrix}, \quad k = -2,$$

то

$$k\mathbf{A} = \begin{bmatrix} 4 & 0 \\ -10 & -14 \end{bmatrix}.$$

Результат скалярного умножения числа на матрицу можно заменить умножением на эту матрицу диагональной матрицы с элементами, равными скаляру:

$$\begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} -2 & 0 \\ 5 & 7 \end{bmatrix}=\begin{bmatrix} (-2)\cdot(-2)+0\cdot 5 & (-2)\cdot 0+0\cdot 7 \\ 0\cdot(-2)+(-2)\cdot 5 & 0\cdot 0+(-2)\cdot 7 \end{bmatrix}=$$

$$=\begin{bmatrix} 4 & 0 \\ -10 & -14 \end{bmatrix}.$$

В статистике важное значение имеет понятие обратной матрицы. Для квадратной матрицы **A** обратной называется такая матрица $\mathbf{A}^{-1}$, что их произведение дает единичную матрицу:

$$\mathbf{A}^{-1}\mathbf{A}=\mathbf{A}\mathbf{A}^{-1}=\mathbf{I}.$$

Например, для матрицы

$$\mathbf{A}=\begin{bmatrix} 2 & 3 \\ 5 & -1 \end{bmatrix}$$

обратной матрицей будет

$$\mathbf{A}^{-1}=\begin{bmatrix} 1/17 & 3/17 \\ 5/17 & -2/17 \end{bmatrix}.$$

В этом легко убедиться, умножив $\mathbf{A}^{-1}$ на **A**:

$$\mathbf{A}^{-1}\mathbf{A}=\begin{bmatrix} 1/17 & 3/17 \\ 5/17 & -2/17 \end{bmatrix}\begin{bmatrix} 2 & 3 \\ 5 & -1 \end{bmatrix}=$$

$$=\begin{bmatrix} (1/17)\cdot 2+(3/17)\cdot 5 & (1/17)\cdot 3-(3/17)\cdot 1 \\ (5/17)\cdot 2-(2/17)\cdot 5 & (5/17)\cdot 3+(2/17)\cdot 1 \end{bmatrix}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

В общем случае не у каждой квадратной матрицы есть обратная. Обратная матрица имеется у так называемых невырожденных матриц, т. е. таких, у которых определитель (детерминант) не равен нулю. Обозначим определитель матрицы **A** следующим образом:

$$\det\mathbf{A}=|\mathbf{A}|=\begin{vmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{vmatrix}.$$

Определителем матрицы **A** называется число, получаемое от суммирования произведений элементов матрицы по определенному правилу. Например, для матрицы **B** размера (2, 2)

$$\det\mathbf{B}=\begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix}=b_{11}b_{22}-b_{21}b_{12}$$

или

$$|\mathbf{B}|=\begin{vmatrix} 2 & 1 \\ 4 & 3 \end{vmatrix}=2\cdot 3-4\cdot 1=2;$$

для матрицы $\mathbf{B}$ размера (3, 3)

$$\det \mathbf{B} = \begin{vmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{vmatrix} = b_{11}b_{22}b_{33} + b_{21}b_{32}b_{13} + b_{12}b_{23}b_{31} -$$

$$- b_{31}b_{22}b_{13} - b_{21}b_{12}b_{33} - b_{32}b_{23}b_{11} \quad \text{или}$$

$$|\mathbf{B}| = \begin{vmatrix} 0 & 1 & 2 \\ 1 & 3 & 1 \\ 0 & -1 & 2 \end{vmatrix} = 0 \cdot 3 \cdot 2 - 1 \cdot 1 \cdot 2 + 1 \cdot 1 \cdot 0 -$$

$$- 0 \cdot 3 \cdot 2 + 1 \cdot 1 \cdot 0 - 1 \cdot 1 \cdot 2 = -4.$$

Не будем здесь объяснять достаточно сложную теорию определителей, лишь отметим одно их важное свойство: если какой-либо из столбцов или строк матрицы можно получить из других столбцов или строк умножением на число или сложением с другими столбцами или строками (т. е. они являются линейно зависимыми), то определитель такой матрицы равен нулю. Например, у матрицы $\mathbf{C}$ третий столбец равен первому, умноженному на 2:

$$\mathbf{C} = \begin{bmatrix} 1 & 5 & 2 \\ 2 & 1 & 4 \\ 1 & 3 & 2 \end{bmatrix}, \quad \det \mathbf{C} = 1 \cdot 1 \cdot 2 + 2 \cdot 3 \cdot 2 + 5 \cdot 4 \cdot 1 -$$

$$- 1 \cdot 1 \cdot 2 - 2 \cdot 5 \cdot 2 - 3 \cdot 4 \cdot 1 = 2 + 12 + 20 - 2 - 20 - 12 = 0;$$

у матрицы $\mathbf{B}$ первый столбец равен сумме второго и третьего столбцов

$$\mathbf{B} = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 0 & 1 \\ 5 & 4 & 1 \end{bmatrix}, \quad \det \mathbf{B} = 3 \cdot 0 \cdot 1 + 1 \cdot 4 \cdot 2 + 1 \cdot 1 \cdot 5 - 5 \cdot 0 \cdot 2 -$$

$$- 4 \cdot 1 \cdot 3 - 1 \cdot 1 \cdot 1 = 0 + 8 + 5 - 0 - 12 - 1 = 0.$$

Итак, мы ввели показатель, который позволяет выяснить, являются ли линейно зависимыми строки или столбцы у квадратной матрицы, это — детерминант. Если $\det \mathbf{A} = 0$, то матрица $\mathbf{A}$ является вырожденной и не имеет обратной матрицы, если $\det \mathbf{A} \neq 0$, матрица $\mathbf{A}$ не вырожденная и существует такая матрица $\mathbf{A}^{-1}$, что $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. Приведем пример:

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 2 \\ 2 & 5 & 3 \\ -3 & -8 & -4 \end{bmatrix}, \quad \det \mathbf{A} = 1 \cdot 5 \cdot (-4) - 2 \cdot 8 \cdot 2 -$$

$$- 3 \cdot 3 \cdot 3 + 3 \cdot 5 \cdot 2 + 8 \cdot 3 \cdot 1 + 2 \cdot 3 \cdot 4 = -20 - 32 -$$

$$- 27 + 30 + 24 + 24 = -1,$$

$$\mathbf{A}^{-1} = \begin{bmatrix} -4 & 4 & 1 \\ 1 & -2 & -1 \\ 1 & 1 & 1 \end{bmatrix}, \quad \mathbf{A}\mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Глава 8. ЛИНЕЙНЫЕ МОДЕЛИ В МАТРИЧНОЙ ФОРМЕ

8.1. ЛИНЕЙНЫЕ УРАВНЕНИЯ

В научных исследованиях, в том числе и в агрометеорологии, часто возникает необходимость обращаться к системам линейных алгебраических уравнений. Решение подобных систем уравнений является основой метода множественного регрессионного анализа. Приведенные в предыдущей главе сведения о некоторых элементах матричной алгебры помогут облегчить нахождение решений таких систем.

Пусть надо решить систему m линейных уравнений с m неизвестными:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m &= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mm}x_m &= b_m, \end{aligned}$$

где a_{ij} и b_i — известные числа, а $x_1, x_2, \ldots, x_m$ — неизвестные переменные, которые необходимо найти. Если коэффициенты при неизвестных переменных в уравнениях представить в виде матрицы $\mathbf{A}$, а неизвестные переменные и правые части уравнений соответственно в виде векторов $\mathbf{x}$ и $\mathbf{b}$, эту же систему уравнений можно записать в матричной форме

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_m \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_m \end{bmatrix} \quad \text{или} \quad \mathbf{Ax} = \mathbf{b}.$$

Решить данную систему значит для заданной матрицы коэффициентов $\mathbf{A}$ и вектора $\mathbf{b}$ найти такой вектор $\mathbf{x}$, для которого выполняется это равенство.

Рассмотрим наиболее часто встречающийся в регрессионном анализе случай, когда матрица коэффициентов является квадратной, т. е. число уравнений равно числу определяемых переменных.

Пусть дана система линейных уравнений

$$\begin{aligned} 2x_1 + 3x_2 + x_3 &= 2, \\ x_1 + 2x_2 + 2x_3 &= 1, \\ 2x + 4x - 3x &= 4. \end{aligned}$$

Ее элементы в матричной форме можно записать так:

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 2 \\ 2 & 4 & -3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Тогда данная система в матричной форме будет иметь вид

$$\mathbf{Ax}=\begin{bmatrix}2 & 3 & 1\\ 1 & 2 & 2\\ 2 & 4 & -3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=\begin{bmatrix}2x_1+3x_2+x_3\\ x_1+2x_2+2x_3\\ 2x_1+4x_2-3x_3\end{bmatrix}=\begin{bmatrix}2\\ 1\\ 4\end{bmatrix}=\mathbf{b}.$$

Подобные системы уравнений решают различными способами. Используя матричную запись, можно решить эквивалентное матричное уравнение, умножив обе части исходного уравнения на матрицу $\mathbf{A}^{-1}$:

$$\mathbf{Ax}=\mathbf{b},$$

$$\mathbf{A}^{-1}\mathbf{Ax}=\mathbf{A}^{-1}\mathbf{b},$$

$$\mathbf{Ix}=\mathbf{A}^{-1}\mathbf{b};$$

тогда

$$\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}.$$

Отсюда следует, что для численного решения данного уравнения необходимо знать матрицу $\mathbf{A}^{-1}$.

Если решение существует, то оно единственное, что следует из единственности представления обратной матрицы $\mathbf{A}^{-1}$.

Пусть дана система уравнений

$$x_1+3x_2+2x_3=2,$$

$$2x_1+5x_2+3x_3=10,$$

$$-3x_1-8x_2-4x_3=-1$$

или в матричной форме

$$\mathbf{A}=\begin{bmatrix}1 & 3 & 2\\ 2 & 5 & 3\\ -3 & -8 & -4\end{bmatrix},\quad \mathbf{b}=\begin{bmatrix}2\\ 10\\ -1\end{bmatrix},$$

$$\mathbf{x}=\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix},\quad \det\mathbf{A}\neq 0;$$

$$\mathbf{Ax}=\mathbf{b}.$$

Обратной для матрицы $\mathbf{A}$ является матрица

$$\mathbf{A}^{-1}=\begin{bmatrix}-4 & 4 & 1\\ 1 & -2 & -1\\ 1 & 1 & 1\end{bmatrix},$$

поэтому решение уравнения ищем в виде

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \begin{bmatrix} -4 & 4 & 1 \\ 1 & -2 & -1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 10 \\ -1 \end{bmatrix} =$$

$$= \begin{bmatrix} (-4)\cdot 2 + 4\cdot 10 (-1)\cdot 1 \\ 1\cdot 2 - 2\cdot 10 + 1\cdot 1 \\ 1\cdot 2 + 1\cdot 10 - 1\cdot 1 \end{bmatrix} = \begin{bmatrix} 31 \\ -17 \\ 11 \end{bmatrix};$$

получили $x_1 = 31$, $x_2 = -17$, $x_3 = 11$.

Для проверки подставим полученные неизвестные в систему уравнений

$$31 + 3\cdot(-17) + 2\cdot 11 = 2,$$

$$2\cdot 31 + 5\cdot(-17) + 3\cdot 11 = 10,$$

$$-3\cdot 31 + 8\cdot 17 - 4\cdot 11 = -1$$

и убедимся в правильности данного решения.

Проблема вырожденности матриц непосредственно связана с решением систем линейных уравнений, с числом линейно независимых уравнений в системе. Рассмотрим систему уравнений

$$\begin{aligned} x_1 + 3x_2 + x_3 &= 4, \\ 2x_1 + x_2 + 2x_3 &= 1, \\ 4x_1 + 7x_2 + 4x_3 &= 9, \end{aligned} \qquad \det \mathbf{C} = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 1 & 2 \\ 4 & 7 & 4 \end{bmatrix} = 0.$$

Анализ данной системы показывает, что третье уравнение может быть получено путем умножения первого уравнения на два и прибавления второго уравнения. Отсюда следует, что третье уравнение не несет дополнительную информацию о взаимозависимости переменных; вся информация о трех переменных содержится в первых двух уравнениях, и у системы нет единственного решения. Такая линейная зависимость уравнений ведет к вырожденности матрицы коэффициентов системы уравнений. В связи с этим вводится понятие ранга матрицы, под которым понимают число линейно независимых уравнений в системе или, что то же самое, число взаимно независимых строк в соответствующей матрице коэффициентов. Если ранг матрицы меньше, чем ее размер, матрица вырожденная. Так, у приведенной матрицы **C** размер равен (3×3), а ранг — 2.

8.2. ЛИНЕЙНЫЕ МОДЕЛИ В МАТРИЧНОЙ ФОРМЕ

В первой части книги показаны значение и роль множественных линейных регрессионных моделей. Наиболее часто они представляются в виде

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \varepsilon_i \quad (i = 1, 2, \ldots, n), \qquad (8.1)$$

где y_i — зависимая переменная, предиктант; x_i — независимая переменная, предиктор; ε_i — случайная ошибка, обычно не зависящая от других переменных, со средним нулевым значением и дисперсией σ^2, как правило, нормально распределенная; β_i — регрессионные коэффициенты, определяющие линейную связь между независимыми и зависимой переменными. Коэффициент β_0 называют свободным членом, он оценивает значение y при условии, если все x_j равны нулю.

Если имеется ряд сопряженных наблюдений за y_i и x_{ij}:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{12} & \dots & x_{1k} \\ x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots \\ x_{n2} & \dots & x_{nk} \end{bmatrix}$$

и требуется в матричной форме выразить линейную зависимость y_i от x_{ij}, то введя обозначение векторов коэффициентов и ошибок

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix},$$

можно записать

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

где $\mathbf{X}$ — матрица размером (n, k); $\mathbf{y}$ и $\boldsymbol{\varepsilon}$ — векторы размером $(n, 1)$; $\boldsymbol{\beta}$ — вектор размером $(k, 1)$. Наличие свободного члена в (8.1) требует введения в матрицу $\mathbf{X}$ столбца единиц.

Обычно количество сопряженных наблюдений намного больше, чем число независимых переменных, т. е. $n > k$, и поэтому с помощью k параметров β_i нельзя удовлетворить имеющимся n различным условиям, т. е. система не имеет своего решения. Например, система

$$2\beta_1 + \beta_2 = 1,$$

$$\beta_1 - \beta_2 = 2,$$

$$2\beta_1 - 3\beta_2 = 3$$

не имеет такого решения β_1 и β_2, чтобы одновременно при подстановке их в систему левые части уравнений были бы равны правым.

Самым распространенным приемом расчета или, как принято в статистике говорить, получения оценок $\hat{\boldsymbol{\beta}}$ неизвестных параметров линейного уравнения регрессии является метод наименьших квадратов. Суть этого метода сводится к такой «подгонке»

коэффициентов $\widehat{\beta}$, чтобы сумма квадратов ошибок $\sum \varepsilon_i^2$ системы уравнений была минимальной:

$$\sum_i \varepsilon_i^2 = \varepsilon'\varepsilon = [\varepsilon_1 \varepsilon_2 \ldots \varepsilon_n] \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix} = (y - X\widehat{\beta})(y - X\widehat{\beta}) =$$

$$= y'y - \widehat{\beta}'X'y - y'X\widehat{\beta} + \widehat{\beta}X'X\widehat{\beta} = y'y - 2\widehat{\beta}X'y + \widehat{\beta}'X'X\widehat{\beta}, \quad (8.2)$$

так как $\widehat{\beta}'X'y$ есть скаляр и не меняется при транспонировании

$$(\widehat{\beta}'X'y)' = (X'y)'\widehat{\beta} = y'X\widehat{\beta}.$$

Для того чтобы найти значение $\widehat{\beta}$, при котором сумма квадратов ошибок $\varepsilon'\varepsilon$ минимальна, продифференцируем матричное выражение (8.2) и приравняем к нулю:

$$\frac{\partial}{\partial \widehat{\beta}}(\varepsilon'\varepsilon) = -2X'y + 2X'X\widehat{\beta} = 0;$$

тогда получаем систему линейных уравнений относительно неизвестных переменных $\widehat{\beta}$

$$X'X\widehat{\beta} = X'y \quad (8.3)$$

(X и y — заданные значения) и решаем ее, умножив обе части уравнения на матрицу $(X'X)^{-1}$

$$(X'X)^{-1}(X'X)\widehat{\beta} = (X'X)^{-1}X'y,$$

$$\widehat{\beta} = (X'X)^{-1}X'y.$$

Здесь мы воспользовались свойством обратной матрицы $(X'X)^{-1} \times (X'X) = I$.

В классической постановке регрессионной задачи считается, что для каждого случая i фиксируются некоторые значения независимых переменных $x_{i1}, x_{i2}, \ldots, x_{ik}$, затем отмечается значение переменной y_i и так — n раз. Полагают, что x_{ij} известны точно, а случайной величиной является переменная y_i, при этом не существует строгой линейной зависимости между предикторами. Если это не так, и одна из переменных будет линейной комбинацией других переменных, то матрица $X'X$ окажется вырожденной, и у нее не будет обратной матрицы $(X'X)^{-1}$, которая необходима при решении уравнения (8.3).

Предполагается, что существуют истинные значения параметров β и дисперсии ошибки уравнения, которые можно определить, имея бесконечно большую выборку наблюдений.

Если считать, что ошибки одного уравнения не зависят от ошибок другого, математическое ожидание ошибок равно нулю и дисперсия ошибок не меняется, то из теории следует состоятельность, несмещенность и эффективность оценок $\hat{\beta}$. Под несмещенностью понимают отсутствие в $\hat{\beta}$ систематической погрешности, под состоятельностью — повышение точности оценивания при увеличении длины эмпирической выборки и под эффективностью — отсутствие другого приема получения коэффициентов, лучшего, чем метод наименьших квадратов.

Для расчета коэффициентов β нет необходимости делать предположение о законе распределения ошибок ε_i; это предположение играет важную роль лишь при анализе качества полученного регрессионного уравнения, его достоверности.

Глава 9. ОСНОВНЫЕ СВОЙСТВА ЛИНЕЙНЫХ МОДЕЛЕЙ

В главе 8 были рассмотрены основные положения метода наименьших квадратов и способ получения параметров линейной регрессионной модели

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m.$$

Численная оценка коэффициентов модели — только первый шаг в ее построении. Отсутствие необходимого анализа коэффициентов делает ее почти бесполезной. Обычно в статистической практике оценивают адекватность модели эмпирическим данным и реальность полученных коэффициентов регрессии. Такой этап проверки моделей основывается на статистической теории проверки гипотез и вычисления доверительных интервалов оцениваемых величин.

Линейные модели, как правило, содержат несколько параметров, поэтому можно рассматривать или сразу всю группу параметров, или каждый в отдельности.

Основой для проверки гипотез в регрессионном анализе является так называемая таблица дисперсионного анализа и использование статистики F — отношение.

В этой таблице показано, какая доля общей вариации (суммы квадратов) зависимой переменной описывается различными независимыми переменными. Такая разбивка предполагает, однако, независимость вклада каждого источника вариации, что, как правило, на практике не соблюдается. Можно также рассчитывать сумму квадратов ошибок уравнения с включением и без включения в него некоторых переменных и анализировать их разность для установления значимости включенных переменных. Однако и здесь при коррелированности предикторов возникают большие трудности.

9.1. ПРОСТАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

На простом примере с одной независимой переменной покажем основные принципы и идеи разделения вариации предиктанта и построения дисперсионной таблицы.

В табл. 9.1 приведены данные пяти пар наблюдений за урожайностью (y_i) за пять последовательных лет (x_i); графическая зависимость показана на рис. 9.1.

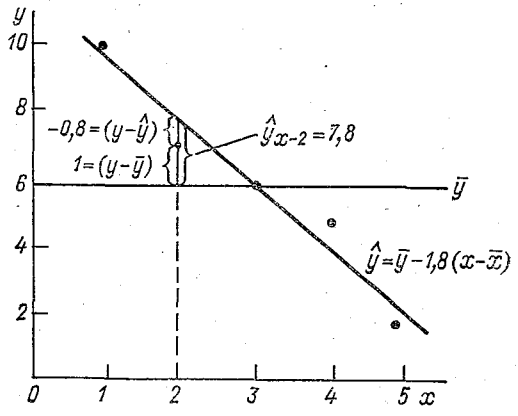


Рис. 9.1. Графическое представление разделения вариаций.

Непосредственный расчет уравнения с помощью метода, описанного в главе 8, дал такие оценки: $\hat{\beta}_0 = 11{,}4$ и $\hat{\beta}_1 = -1{,}8$, т. е. получено уравнение для расчета

$$y = 11{,}4 - 1{,}8x.$$

Отклонение расчетного значения зависимой переменной $\hat{y}$ от истинного y также представлено в табл. 9.1. На основании оче-

Таблица 9.1

Разбивка общей вариации на составляющие

x	y	$\hat{y}$	$y-\bar{y}$	$\hat{y}-\bar{y}$	$y-\hat{y}$
1	10	9,6	4	3,6	0,4
2	7	7,8	1	1,8	—0,8
3	6	6,0	0	0,0	0,0
4	5	4,2	—1	—1,8	0,8
5	2	2,4	—4	—3,6	—0,4
Сумма квадратов			34	32,40	1,60

видного геометрического смысла, следующего из рис. 9.1, можем написать тождество

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}),$$

где y_i — истинное значение переменной, $\bar{y}$ — среднее арифметическое значение y_i, $\hat{y}$ — значение, рассчитанное по модели.

Возведя обе части уравнения в квадрат и просуммировав его для всех точек графика, получим

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2. \qquad (9.1)$$

Интерпретация отдельных сумм квадратов, входящих в тождество, следующая:

$\sum (y_i - \bar{y})^2$ — сумма квадратов отклонений истинных значений зависимой переменной от расчетного значения. Здесь в качестве расчетной величины берется среднее арифметическое нредиктанта. Это выражение является как бы эталоном, от которого затем с помощью регрессионной зависимости получают какое-то улучшение предсказания при введении в модель независимых переменных. Данную величину называют обычно полной (общей) суммой квадратов (ПСК). Она имеет $n-1$ степень свободы. Одна степень свободы «теряется» из-за налагаемых на нее ограничений в методе наименьших квадратов: $\sum (y_i - \bar{y}) = 0$.

$\sum (y_i - \hat{y}_i)^2$ — сумма квадратов отклонений истинного значения зависимой переменной от предсказанного по регрессионной модели значения предиктанта. Эта величина показывает расхождение регрессионной модели с истинными данными и обычно называется суммой квадратов ошибок (СКО). Она имеет $n-2$ степени свободы из-за двух ограничений в методе наименьших квадратов: $\sum (y_i - \hat{y}_i) = 0$, $\sum x_i (y_i - \hat{y}_i) = 0$.

$\sum (\hat{y}_i - \bar{y})^2$ — сумма квадратов разности между оценками регрессионной модели и средним арифметическим значением зависимой переменной. Эту величину принято называть суммой квадратов, обусловленной регрессией, или регрессионной суммой квадратов (РСК). Она имеет одну степень свободы, так как при двух параметрах у простой регрессии есть ограничение $\sum (\hat{y}_i - \bar{y}) = 0$.

Таким образом, можно, не используя языка формул, записать ПСК = СКО + РСК.

Отметим, что подобное соотношение выполняется, если при подборе уравнения используется метод наименьших квадратов.

Из приведенного соотношения следует: СКО = ПСК — РСК.

Пусть необходимо проверить гипотезу, что $\beta_1 = 0$. Ее можно опровергнуть или принять. Обычно пишут: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$. Если наше предложение верно и нуль-гипотеза верна, то $\beta_1 \neq 0$.

Рассчитав по формуле (9.1) элементы вариации, составим таблицу дисперсионного анализа (табл. 9.2). F_0-отношение используется для проверки гипотез при предположении о нормальном распределении ошибок $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, т. е. ошибки распределены по закону $N(0, \sigma^2)$ — с нулевым средним и дисперсией σ^2.

Таблица 9.2

Таблица дисперсионного анализа для простой линейной регрессии

Источник вариации	Сумма квадратов	Степень свободы	Средний квадрат	F-отношение
Регрессия	РСК	1	РСК	$F_0 = \dfrac{\text{РСК}}{\sigma^2}$,
Отклонение от регрессии	СКО	$n-2$	$\sigma^2 = \dfrac{\text{СКО}}{n-2}$	
Полная вариация	ПСК	$n-1$		

Если верна гипотеза H_0, то F_0 будет иметь F-распределение с $(1, n-2)$ степенями свободы. Вероятность принятия гипотезы задается P — значением — площадью области под графиком плотности распределения $F(1, n-2)$ справа от числа F_0. Гипотеза H_0 отвергается, если P меньше установленного уровня значимости α. Если гипотеза H_0 принимается, то это означает, что лучше всего оценивать y_i его средним значением y для любых x_i.

Можно проверить и другие гипотезы, например: $H_0: \beta_1 = \beta_1'$, где β_1' — любое число.

Для этого используем статистику

$$t_0 = \frac{\hat{\beta}_1 - \beta_1'}{[V(\hat{\beta}_1)]^{1/2}},$$

где $V(\hat{\beta}_1) = \sigma^2 / [\sum_i (x_i - \bar{x})^2]$ — стандартная ошибка расчета коэффициента регрессии $\hat{\beta}_1$, а σ^2 — дисперсия ошибки уравнения. Если нуль-гипотеза H_0 верна, то t_0 имеет t-распределение Стьюдента с $n-2$ степенями свободы.

Стандартные регрессионные программы обычно распечатывают дисперсионную матрицу, значение статистики t_0 и стандартную ошибку коэффициентов регрессии.

Так как величина $(\hat{\beta}_1 - \beta_1')/[V(\hat{\beta}_1)]^{1/2}$ имеет t-распределение, можно указать вероятность ее попадания в интервал

$$P\{t(\alpha/2;\ n-2) \leqslant (\hat{\beta}_1 - \beta)/[V(\hat{\beta}_1)]^{1/2} \leqslant t(1-\alpha/2;\ n-2)\} = 1-\alpha. \tag{9.2}$$

Здесь $t(\alpha/2;\ n-2)$ значение t-статистики при $(\alpha/2)\cdot 100\,\%$-ном уровне значимости с $n-2$ степенями свободы. Из симметричности t-распределения следует, что

$$t(\alpha/2;\ n-2) = -t(1-\alpha/2;\ n-2).$$

Можно переписать неравенство (9.2) следующим образом:

$$P\{\hat{\beta}_1 - t(1-\alpha/2;\ n-2)[V(\hat{\beta}_1)]^{1/2} \leqslant \beta_1 \leqslant \hat{\beta}_1 + t(1-\alpha/2;\ n-2)[V(\hat{\beta}_1)]^{1/2}\} = 1-\alpha.$$

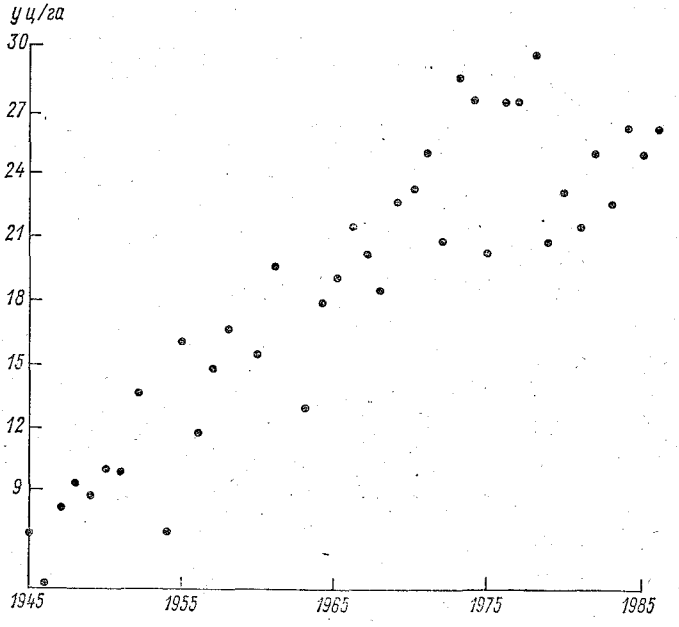


Рис. 9.2. Временной ряд урожайности (y) всех зерновых культур на Украине.

Откуда получаем, что с вероятностью $(1-\alpha)$ β_1 заключено в интервале

$$\hat{\beta}_1 - t(\alpha/2;\ n-2)[V(\hat{\beta}_1)]^{1/2} \leqslant \beta_1 \leqslant \hat{\beta}_1 + t(\alpha/2;\ n-2)[V(\hat{\beta}_1)]^{1/2}.$$

Пример. Пусть надо рассчитать значение коэффициента $\hat{\beta}_1$ в уравнении тренда урожайности всех зерновых культур по Украинской ССР с 95 %-ной вероятностью (рис. 9.2). Рассчитанное значение средней квадратической ошибки коэффициента наклона равно

$$[V(\hat{\beta}_1)]^{1/2} = 0{,}0386.$$

Для расчета 95 %-ной вероятности попадания в заданный интервал используем значение $t(\alpha;\ n-2)$. В соответствии с (9.2),

находим по таблице [13], что t (0,05; 40) = 2,02. Поэтому с 95 %-ной вероятностью истинное значение коэффициента β_1 лежит в интервале

$$0{,}508 - 2{,}02 \cdot 0{,}0386 \leqslant \beta_1 \leqslant 0{,}508 + 2{,}02 \cdot 0{,}0386;$$

$$0{,}430 \leqslant \beta_1 \leqslant 0{,}586.$$

Отсюда следует, что с вероятностью 95 % урожайность с каждым годом повышается от 0,430 до 0,586 ц/га.

Аналогичным способом можно определить достоверность попадания в интервал свободного члена уравнения.

Если распределение y близко к нормальному или не очень сильно отличается от него, то применение t-статистики для получения статистических выводов оправдано. Если это не так, необходимо различными преобразованиями изменить распределение y и привести его к нормальному закону.

9.2. ПОСЛЕДСТВИЯ НЕДОСТАТКА И ИЗБЫТКА ПЕРЕМЕННЫХ В РЕГРЕССИОННОЙ МОДЕЛИ

При построении линейных регрессионных моделей, предназначенных для прогнозирования, агрометеорологи сталкиваются с проблемой выбора необходимого подмножества переменных. Желание включить как можно больше переменных, к сожалению, часто наталкивается на трудности, связанные с ухудшением статистических свойств моделей. Использование слишком малого числа объясняющих переменных также приводит к ухудшению прогностических способностей агрометеорологических уравнений. Приведем некоторые факты, проясняющие возможные последствия некорректной спецификации регрессионной линейной модели, так как обычно обращают внимание только на влияние количества включенных в уравнение регрессоров, или предикторов, на уровень значимости множественного коэффициента корреляции, хотя главной целью является выбор подмножества переменных, которые минимизируют дисперсию и смещение отклика модели (предсказанного значения).

При статистическом моделировании, как правило, не известно точно, какие и сколько переменных должны войти в модель, поэтому ошибки в ту или иную сторону очень вероятны. Знание последствий подобных ошибок позволит избавиться от неадекватности модели.

Пусть истинная модель есть

$$\mathbf{y} = \mathbf{X}_1\beta_0 + \varepsilon$$

и выполнены основные предположения для классической модели. Предположим, что оценивается модель с расширенным набором переменных:

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon. \qquad (9.3)$$

Можно показать, что в этом случае рассчитанные коэффициенты расширенной модели будут несмещенными, т. е. в пределе стремящимися к истинным значениям. Математическое ожидание дополнительных параметров равно нулю, они также оцениваются несмещенно. В случае перебора переменных оценка β является состоятельной, $\widehat{\beta}$ в среднем квадратичном сходится к истинному значению $(\beta_0, 0)'$.

Если положить, что вектор наблюдений $(x_1, x_2)'$ состоит из некоторого набора независимых переменных, то получим два прогноза по корректной и расширенной модели:

$$y_1 = x_1'\beta_0,$$

$$y_2 = x_1'\beta_1 + x_2'\beta_2;$$

и можно показать [50], что

$$V(y_2) \geqslant V(y_1).$$

Знак равенства возможен только в случае, если

$$x_1'(X_1'X_1)^{-1} X_1'X_2 = x_2',$$

откуда следует, что при ортогональности переменных дисперсия отклика модели одинакова, если $x_2 = 0$. В случае же неортогональности предикторов равенство имеет место, когда элементы x_2 есть линейная комбинация элементов x_1.

Аналогичный результат получен и для коэффициентов модели. Если набор определяющих независимых переменных известен, то дополнительные переменные, ортогональные к основному множеству переменных, не увеличивают дисперсию коэффициентов; в случае неортогональности дисперсия коэффициентов расширенной модели возрастает, т. е. они оцениваются менее точно.

Если истинным уравнением является (9.3), а в оцениваемую модель включена лишь часть независимых переменных, составляющих матрицу X_1, тогда как переменные, составляющие X_2, отсутствуют:

$$y = X_1\gamma + \varepsilon, \quad X = \begin{pmatrix} \cdot & \cdot \\ X_1 & \cdot X_2 \\ \cdot & \cdot \end{pmatrix},$$

то полученная оценка коэффициентов

$$\hat{\gamma} = (X_1'X_1)^{-1} X_1'y$$

в общем случае является смещенной. Оценка $\hat{\gamma}$ будет несмещенной только в том случае, когда матрицы, составленные из переменных X_1 и X_2, ортогональны. Смещение зависит как от модели, которую строят, так и от истинной модели. При правильном подборе переменных смещение будет не очень значительным даже

в случае неполной модели, если $\mathbf{X}_1$ выбрано так, чтобы $\mathbf{X}_1'\mathbf{X}_2=0$, что дает отсутствие смещения в $\hat{\gamma}$.

Смещение в коэффициентах приводит к смещению прогнозов по неполной модели:

$$\hat{\mathbf{y}}=\mathbf{X}_1\hat{\gamma}=\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}=\mathbf{X}_1\beta_1+\mathbf{X}_1\mathbf{C}\beta_2+\varepsilon_1.$$

Оценки $\hat{\gamma}$ оказываются также несостоятельными.

Таким образом, наличие избыточного количества переменных в регрессионных уравнениях не может уменьшить (а фактически всегда увеличивает) дисперсию предсказанного отклика. Поэтому необходимо избегать переусложнения модели в погоне за повышением ее точности. Более простые модели имеют явные преимущества, часто они более точные. В то же время нельзя допускать невключения в модель важных, оказывающих существенное влияние на зависимую переменную предикторов, так как оценки коэффициентов уравнения окажутся несостоятельными и смещенными, что может сделать модель непригодной для прогностического использования. В любом случае надо стремиться выбирать переменные, близкие к ортогональности, т. е. слабо зависимые между собой.

На практике с фактом ухудшения предсказания по расширенной модели агрометеорологи сталкиваются достаточно часто. Наряду с возможными причинами, упомянутыми выше, на точность прогнозирования сильно влияет коррелированность входящих в регрессионную модель предикторов. При включении дополнительных, как правило, неортогональных к уже включенным в модель предикторов линейная зависимость системы предикторов возрастает, что в свою очередь приводит к ухудшению прогнозирования. Влияние коррелированности переменных на прогнозирование рассмотрено в разделе 9.3.

Проиллюстрируем эффект переусложнения на примере прогностической модели всех зерновых культур по Восточно-Казахстанской и Семипалатинской областям. В первое регрессионное уравнение для прогноза урожайности зерновых на конец июня было включено только два предиктора: показатель влагообеспеченности посевов за июнь и сумма осадков за май. Коэффициент детерминации уравнения достаточно высок: 0,653, т. е. 65,3 % дисперсии урожайности нашло объяснение через эти две переменные. Такой высокий коэффициент принято считать хорошо отражающим связь между переменными в регрессионном уравнении.

Затем было построено прогностическое уравнение с добавлением информации о влагообеспеченности следующего важного месяца вегетационного периода — июля, и сумме осадков за зимний период. Коэффициент детерминации значительно увеличился и стал равен 0,84, что свидетельствует о значимости новых факторов. Статистика t, вычисленная для обоих факторов, показала значимость коэффициентов при них на уровне 0,005. Несмотря на

это, при прогнозировании на независимой выборке средний квадрат ошибки первого уравнения в четыре с лишним раза меньше, чем второго: соответственно 0,131, и 0,560.

Неадекватность модели, возникающую из-за невключения в нее существенной переменной, покажем на примере аналогичного регрессионного уравнения по данным Павлодарской, Тургайской, Карагандинской областей. Уравнения составлялись для возможности прогноза в конце июня, в каждое из них вошло по шесть предиктантов, пять из которых были одинаковыми: это — сумма осадков за холодный период (октябрь—март), за май и за третью декаду июня, а также температура воздуха в первую и вторую декады июня. В первом уравнении шестым предиктором бралась сумма осадков за первую декаду июня, во втором — показатель влагообеспеченности июня. Многочисленными исследованиями установлено, что последний предиктор является очень важным фактором формирования урожая, его отсутствие в уравнениях приводит к их плохим прогностическим возможностям. Проверка моделей на независимых материалах дала средний квадрат ошибки первого уравнения 0,485, второго — 0,234 (в единицах дисперсии). Если исключить половину предикторов и построить регрессионное уравнение только по сумме осадков за холодный период, за май и показателю влагообеспеченности июня, то ошибка не увеличится, а станет еще меньше: 0,230.

9.3. КОРРЕЛИРОВАННОСТЬ ПРЕДИКТОРОВ И СПОСОБЫ ЕЕ ВЫЯВЛЕНИЯ

При построении регрессионных моделей обычно используют обширную информацию об агрометеорологических условиях и имеющиеся данные об элементах продуктивности сельскохозяйственных культур. Как неоднократно подчеркивалось, между этими переменными существует тесная связь. Имеют место взаимозависимости между факторами, действующими в один отрезок времени (например, декадная температура и дефицит влажности воздуха), и между одним и тем же фактором, взятым в разные моменты времени. Последняя зависимость отражает тот факт, что рассматриваемые метеорологические величины не являются независимыми переменными, а величины временных рядов, природа которых обусловлена инерционностью метеорологических процессов, их закономерной цикличностью. До недавнего времени проблеме взаимосвязи, или коррелированности, предикторов в регисионном анализе не уделялось должного внимания, хотя, как известно, сильная коррелированность может в значительной мере уменьшить эффективность использования регрессионной модели для прогноза.

Опыт применения регрессионных моделей в агрометеорологии показывает, что нередки случаи, когда из-за сильной коррелированности предикторов при множественном коэффициенте корре-

ляции 0,99 модель не пригодна для прогноза новых данных и отбраковывается [16].

Сильную коррелированность предикторов можно алгебраически представить как

$$\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \ldots + \alpha_m \mathbf{x}_m \approx \bar{0}, \qquad (9.4)$$

где $\mathbf{x}_i$ — вектор-столбец матрицы независимых переменных $\mathbf{X}$. Чем ближе левую часть с помощью подбора коэффициентов α_i можно приблизить к нулевому вектору, тем сильнее выражена коррелированность системы факторов. На практике, даже если какие-либо предикторы и связаны между собой функциональной линейной зависимостью, из-за наличия случайных ошибок в независимых переменных точное равенство в соотношении (9.4) не встречается. Можно сказать, что главный вопрос — не наличие коррелированности, а ее степень.

Линейная взаимозависимость предикторов приводит к ряду негативных последствий при построении регрессионных моделей. Прежде всего уменьшается точность оценивания, т. е. увеличиваются ошибки коэффициентов, и они сильно коррелируют друг с другом, что затрудняет анализ влияния отдельных факторов на зависимую переменную. Часто ошибки имеют знаки, противоречащие физическому смыслу. Добавление или изъятие малого количества данных наблюдений может привести к резкому изменению регрессионных коэффициентов.

Коррелированность переменных при построении регрессионных моделей — достаточно сложное и многообразное явление. Используется несколько показателей для его выявления, рассмотрим некоторые из них.

Если переменные $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ стандартизованы, то $\mathbf{X}'\mathbf{X}$ представляет корреляционную матрицу. Как уже указывалось $\det(\mathbf{X}'\mathbf{X}) = 0$, если между независимыми переменными есть линейная зависимость и $\det(\mathbf{X}'\mathbf{X}) = 1$, если они абсолютно независимы (ортогональны). Таким образом, зная возможный размах значений определителя, можно судить о степени коррелированности переменных.

Если вычислить и упорядочить по величине собственные значения матрицы $\mathbf{X}'\mathbf{X}$, т. е. $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \ldots \geqslant \lambda_m$, и составить отношение, называемое числом обусловленности:

$$p = \frac{\lambda_1}{\lambda_m},$$

можно получить полезную характеристику отклонения матрицы $\mathbf{X}'\mathbf{X}$ от идеального случая, когда все переменные ортогональны.

При построении адекватных моделей необходимо выявить наиболее коррелированные предикторы и заменить их на менее зависимые. Недостаточно использовать для этих целей корреляционную матрицу, хотя кажется очевидным, что чем больше значение парного коэффициента корреляции, тем сильнее взаимозависимость факторов. При таком подходе рассматривается не вся

совокупность предикторов. Не составляет труда построить пример, когда переменные попарно независимы, а вместе они составляют линейно зависимую систему.

Более надежным способом выявления взаимокоррелнрованности каждого предиктора со всеми остальными предикторами является анализ показателя $\mathbf{VIF}(\beta_i)$, предложенного в [44]. Он выражает степень увеличения дисперсии β_i — коэффициента в линейном регрессионном уравнении — и равен значению соответствующего диагонального элемента матрицы $(\mathbf{X'X})^{-1}$. При обычной оценке коэффициентов методом наименьших квадратов

$$\hat{\beta} = (\mathbf{X'X})^{-1}\,\mathbf{X'y}$$

дисперсия коэффициента выражается как

$$V(\beta_i) = c_{ii}\sigma^2,$$

где $\mathbf{C} = (\mathbf{X'X})^{-1}$ и c_{ii} — есть диагональный элемент матрицы $\mathbf{C}$. Можно показать, что

$$V(\beta_i) = \frac{\sigma^2}{(1 - R_i^2)}.$$

Здесь R_i — множественный коэффициент корреляции i-й независимой переменной с остальными $m - 1$ независимыми переменными; σ^2 — дисперсия ошибки уравнения. Таким образом:

$$\mathbf{VIF}(\beta_i) = \frac{1}{(1 - R_i^2)}.$$

Если x_i не коррелирована с остальными переменными и $R_i^2 = 0$, то дисперсия коэффициента равна σ^2. Если значение R_i^2 достаточно велико, то дисперсия β_i будет также большой, т. е. его расчет может быть проведен с существенной ошибкой.

Достаточно сложно дать рекомендации, когда необходимо считать систему переменных сильно коррелированной и полезно принимать меры к ее ослаблению. Одни авторы предлагают считать систему сильно коррелированной, если есть парные коэффициенты корреляции между независимыми переменными, превышающие 0,6, другие делают вывод о степени коррелированности системы на основе сопоставления парных коэффициентов корреляции между независимыми переменными с парными коэффициентами корреляции между независимыми переменными и зависимой переменной. При наличии тесной связи между предикторами систему считают достаточно серьезно подтвержденной коррелированности.

При построении агрометеорологических регрессионных моделей исследователи, как правило, сталкиваются с сильной взаимозависимостью предикторов. Числа обусловленности матриц $\mathbf{X'X}$ могут достигать больших значений, что затрудняет обращение

матриц и приводит к неточности при расчете коэффициентов моделей.

В табл. 9.3 приведены максимальные и минимальные собственные значения λ_{max} и λ_{min} и числа обусловленности для набора декадных средних областных значений сумм осадков и температуры воздуха за период с первой декады мая по вторую декаду августа.

Таблица 9.3

Максимальное и минимальное собственные значения, число обусловленности системы метеорологических факторов

Показатель	Область				
	Северо-Казахстанская	Кустанайская	Уральская	Семипалатинская	Джамбулская
λ_{max}	3,24585	3,6393	4,1570	5,47126	5,32855
λ_{min}	0,00021	0,00001	0,00132	0,00014	0,00001
p	15457,1	363900,0	3149,2	39080,9	532855,0

Во всех представленных областях значение **VIF** у некоторых коэффициентов превышало 1000.

Из таблицы видно, что система сильно коррелирована, поэтому при построении регрессионных моделей возникнут трудности с расчетом и точностью оценивания коэффициентов моделей.

Одним из самых простых способов борьбы с коррелированностью является отбрасывание избыточных и сильно зависимых факторов. Для такой процедуры очень полезным бывает показатель **VIF**, так как после применения различных методов автоматического выбора предикторов, которые в значительной степени уменьшают коррелированность независимых переменных, сильная взаимозависимость предикторов часто сохраняется.

Пример. Для объединенной совокупности данных по трем областям — Северо-Казахстанской, Кокчетавской и Кустанайской — было построено прогностическое уравнение урожайности всех зерновых культур, составляемое в конце июня. Уравнение получено после применения автоматических процедур выбора переменных по критерию C_p (см. гл. 10). В качестве предикторов брались a_1, a_2 — сумма осадков за первую и вторую декады июня, a_3 — показатель влагообеспеченности третьей декады июня, a_4 — средняя декадная сумма осадков за май, a_5 — показатель влагообеспеченности июня, a_6 — сумма осадков за холодный (октябрь—март) период. Все переменные использовались в стандартизованном виде с единичной дисперсией и нулевым средним, что приводит к увеличению степени обусловленности корреляционной матрицы.

Анализ построенной модели (табл. 9.4) показывает, что значения коэффициентов при a_1, a_3, a_6 не соответствуют реальному

Таблица 9.4

Коэффициенты прогностического регрессионного уравнения и показатели коррелированности предикторов

Показатели	Переменная					
	a_1	a_2	a_3	a_4	a_5	a_6
Коэффициенты уравнений	—0,908	0,096	—0,459	0,232	2,145	—0,592
VIF	5,88	11,63	8,20	1,28	50,0	1,22

влиянию этих предикторов на урожайность, так как они отрицательные, хотя в данном районе осадки июня, как и два других фактора, должны нести положительный вклад в регрессионное уравнение. Несмотря на это, множественный коэффициент корреляции уравнения достаточно высок (0,741) и статистически значим на 1 %-ном уровне, даже если учитывать эквивалентно независимые наблюдения для выборки урожайности, составленной по трем областям [30].

На независимых данных такая модель не сможет адекватно описывать влияние агрометеорологических условий на урожай. Обратим внимание и на большие значения показателя коррелированности предикторов. Менее точно оценены коэффициенты при a_5 и a_2.

Приведенный пример является весьма типичным, он подтверждает важность исследования степени взаимозависимости используемых потенциальных предикторов и необходимость применения методов регрессионного анализа, которые более устойчивы к коррелированности предикторов, чем классический метод наименьших квадратов (см. гл. 8).

Нет простого и однозначного совета, как избежать отрицательных последствий от коррелированности предикторов, вошедших в регрессионную модель, однако можно дать следующие рекомендации.

1. Избирать из ряда факторов наиболее важные. Если есть несколько переменных, описывающих примерно одно и то же, необходимо остановить выбор на одной переменной.

2. Воспользоваться автоматическими процедурами выбора предикторов, если априорной информации не достаточно для ограничения числа переменных (см. гл. 10). Однако такие процедуры не дают гарантии избавления от коррелированности предикторов.

3. Использовать необходимую трансформацию переменных. Например, x и x^2 могут быть сильно коррелированными, а $x - \bar{x}$ и $(x - \bar{x})^2$ — некоррелированными.

4. Применить некоторые методы обработки многомерных данных, такие, как метод главных компонент или факторный анализ.

5. Включить в модель коррелированные переменные, если это необходимо из теоретических соображений и невозможно никаким образом избавиться от коррелированности. В этом случае можно использовать метод гребневой регрессии (см. гл. 11).

9.4. ПРОВЕРКА НОРМАЛЬНОСТИ РАСПРЕДЕЛЕНИЯ ПЕРЕМЕННЫХ

Использование теории регрессионного анализа предполагает априорные знания. Так, принято считать, что зависимая переменная в агрометеорологических исследованиях — часто урожайность сельскохозяйственной культуры — распределена по нормальному закону по колоколообразной кривой (что эквивалентно нормальному распределению ошибок регрессионного уравнения). Условие нормальности распределения является очень важным для регрессионного анализа, так как на нем построена вся теория оценивания доверительных интервалов для коэффициентов модели; только при этом условии оценки, полученные методом наименьших квадратов, совпадают с оценками, полученными методом максимального правдоподобия.

Более того, вся корреляционная теория построена на основе допущения многомерного нормального распределения рассматриваемых величин. Значимость парных коэффициентов корреляции между урожайностью и метеорологическим фактором можно проверить, исходя из предположения, что они имеют двумерное нормальное распределение. То же надо предположить при рассмотрении множественного коэффициента корреляции. Эти предположения не делаются в классическом регрессионном анализе, где независимые переменные считаются неслучайными и точно известными величинами, и только зависимая переменная является недетерминированной величиной, случайной переменной.

На практике никогда не бывает распределений, в точности следующих какому-либо закону, но эмпирические данные и характер явления «подсказывают», какое распределение является более подходящим.

В [38] показано, что такой важный статистический критерий, как F, используемый для проверки значимости коэффициентов модели, малочувствителен к отклонению ошибок уравнения от нормальности только при условии «приблизительной нормальности» независимых переменных в классическом уравнении регрессии.

Исследование нормальности распределения случайных величин можно проводить с помощью графических методов — построения гистограмм, пробит-графиков и расчета статистических критериев.

Пробит-график является аналогов построений на так называемой вероятностной бумаге. При построении пробит-графика данные предварительно ранжируются. На горизонтальную ось наносятся ранжированные данные, а на вертикальной оси откладываются соответствующие им значения аргумента стандартной нор-

мальной функции распределения, т. е. $\Phi^{-1}(3j-1)/(3N+1)]$, где Φ — нормальная функция распределения. Если распределение переменной близко к нормальному, то пробит-графиком будет прямая линия.

В работе [40] отмечается, что только при объеме выборки больше 32 можно делать какие-либо выводы по пробит-графику, и предпочтительнее использовать выборки объемом не меньше 50. На примере декадных сумм осадков и температуры воздуха в Казахстане покажем, что этому вопросу надо уделять большое внимание.

Гистограммы сумм осадков показывают наличие асимметрии в распределении декадных сумм осадков. Правый «хвост» распределения обычно намного длиннее левого, что определяется естественным нулевым ограничением слева колебаний этих величин. Это хорошо видно на большинстве построенных гистограмм. Наличие нижнего, часто достигаемого предела значений сумм осадков само по себе противоречит описанию их нормальным распределением, так как в этом случае колебания переменных ничем не ограничены.

Особенно значительно отклонение распределения от гипотетического нормального вида в южных районах с недостаточным увлажнением, где среднее количество осадков мало и вероятность незначительных осадков очень велика.

Так, на рис. 9.3 показана гистограмма сумм осадков в Чимкентской, Талды-Курганской, Алма-Атинской и Джамбулской областях в третьей декаде июня. Вид гистограммы больше напоминает F-распределение при малом числе степеней свободы. Здесь нет необходимости проверять какие-либо гипотезы о соответствии эмпирического распределения нормальному, так как очевидно противное. На рис. 9.4 показан соответствующий данной гистограмме пробит-график, который далек от прямой линии.

Очень слабо напоминают колоколообразную форму нормального распределения гистограммы суммы осадков других районов Казахстана. Типично наличие асимметрии и длинного правого хвоста.

Преобразование $\sqrt{x_i}$ часто используется для приведения асимметричных распределений с одной вершиной и длинным правым хвостом к нормальному [1].

Гистограммы температуры воздуха, как правило, более симметричны, однако, анализ пробит-графиков и самих гистограмм показывает слабое их сходство с нормальной формой распределения. Для многих гистограмм характерна примерно одинаковая частота для широкого диапазона значений температуры, за пределами которого она быстро уменьшается.

Так, на рис. 9.5 и 9.6 приведены гистограмма температуры воздуха в третьей декаде июня в Северо-Казахстанской, Кокчетавской и Кустанайской областях и соответствующий ей пробит-график. В большинстве аналогичных случаев пробит-графики не со-

впадают с прямой линией, что свидетельствует об отклонении распределения температуры воздуха от нормального закона.

На рис. 9.7 показан типичный пробит-график урожайности всех зерновых культур в Восточном Казахстане. На графике

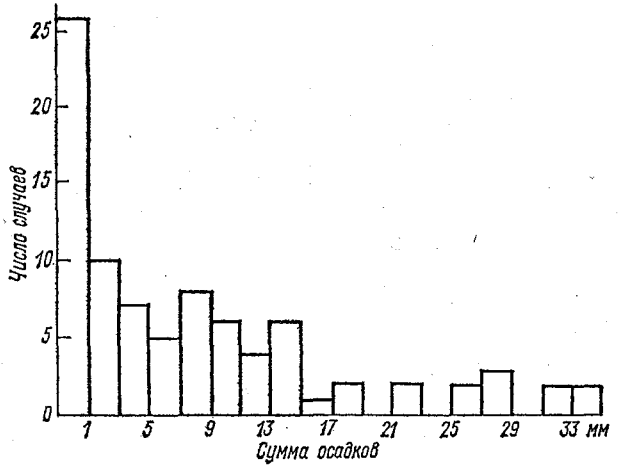


Рис. 9.3. Гистограмма сумм осадков в третьей декаде июня в Южном Казахстане.

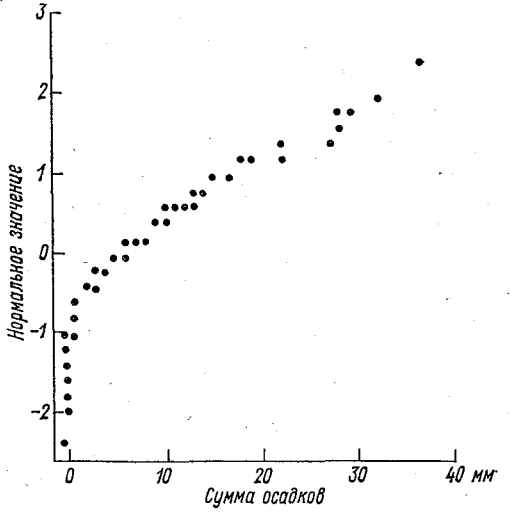


Рис. 9.4. Пробит-график сумм осадков в третьей декаде июня в Южном Казахстане.

видно отклонение точек от прямой линии, особенно заметное на концах. График построен по данным об урожайности, в которых имеет место временной тренд. После элиминирования тренда график несколько изменился (рис. 9.8). Однако и здесь хорошо заметна его нелинейность, точки значительно отклоняются на

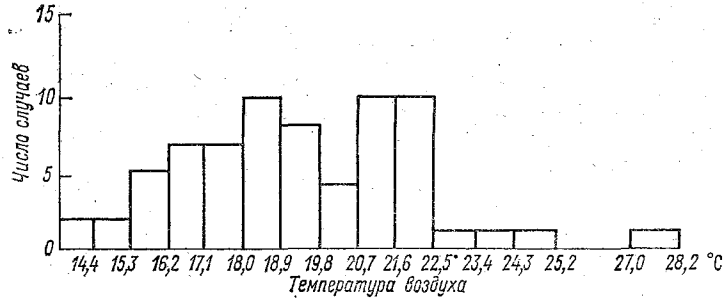


Рис. 9.5. Гистограмма сумм температуры воздуха в третьей декаде июня в Северном Казахстане.

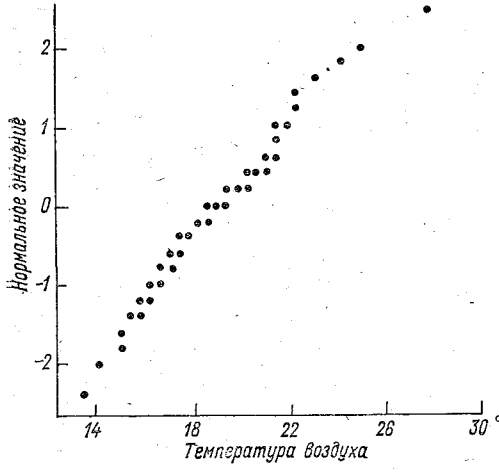


Рис. 9.6. Пробит-график температуры воздуха в третьей декаде в Северном Казахстане.

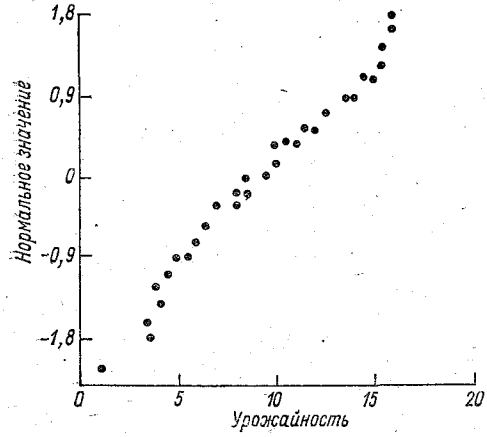


Рис. 9.7. Пробит-график урожайности зерновых культур в Восточном Казахстане.

концах. Это указывает на наличие выбросов из общей статистической совокупности.

На рис. 9.9 и 9.10 показаны гистограммы отклонений урожайности от трендов в Северном и Южном Казахстане. Из-за раз-

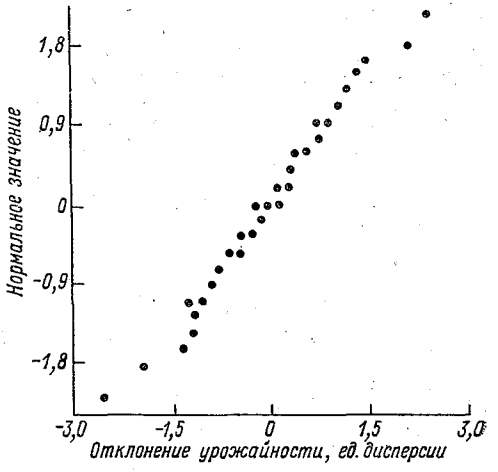


Рис. 9.8. Пробит-график отклонений урожайности зерновых культур от трендов в Восточном Казахстане.

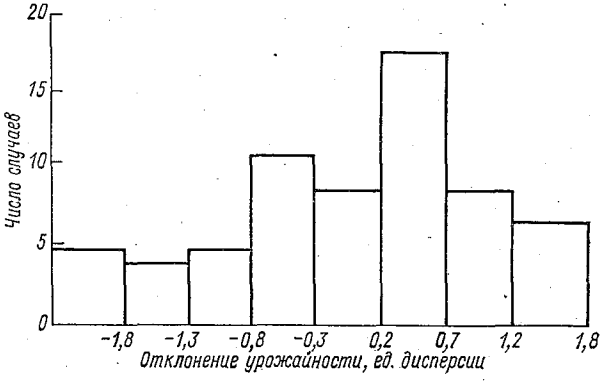


Рис. 9.9. Гистограмма отклонений урожайности зерновых культур от трендов в Северном Казахстане.

личного объема выборок длина интервалов на гистограммах неодинакова. На этих и других гистограммах отклонений урожайности от трендов можно отметить «провал» частоты около нулевого значения. Этот факт достаточно интересен и свидетельствует о едином климатическом механизме, обусловливающим колебания средней областной урожайности зерновых культур в Казахстане. Такие распределения отличаются от гауссовских также наличием асимметрии. Построение гистограмм отклонений урожайности от тренда для выборок различного объема показало достаточную устойчивость проявления такой особенности. Несо-

мненно, что в дальнейших исследованиях необходимо изучать природу возникновения таких интересных форм распределения.

Как известно, ошибки (регрессионные остатки) уравнений распределены по такому же закону (с точностью до параметров), как и сама зависимая переменная. Поэтому следует помнить, что к проверке гипотез относительно коэффициентов регрессионных моделей и степени адекватности уравнений, опирающейся на гипотезу нормального распределения, надо относится с большой

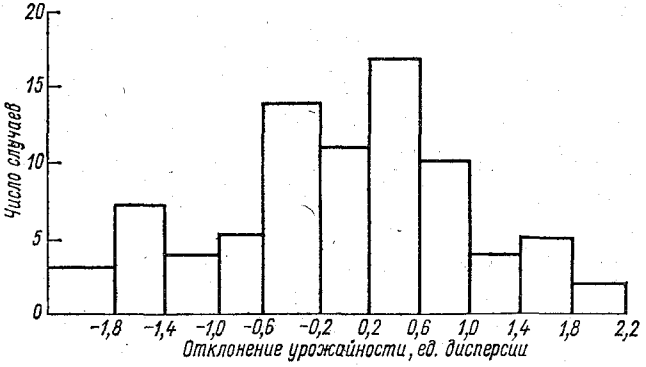


Рис. 9.10. Гистограмма отклонений урожайности зерновых культур от трендов в Южном Казахстане.

осторожностью и иметь в виду, что полученные результаты ориентировочны.

9.5. ЛИНЕЙНЫЕ ПРЕОБРАЗОВАНИЯ ПЕРЕМЕННЫХ

При расчетах на ЭВМ числа в памяти машины представляются с определенной точностью, количество значащих цифр ограничено. Это приводит при вычислениях к накоплению ошибок округления. Подчас исследователь интерпретирует не истинные соотношения между переменными, а итог накопленной суммы ошибок расчетов. В ряде программ имеется возможность увеличивать точность представления числа в ЭВМ. Приведем простой пример, иллюстрирующий возникающие проблемы. Проведем вручную следующие арифметические операции с округлением до двух значащих цифр после запятой: $500\,000\left(\frac{500}{6}-\frac{583{,}259}{7}\right)=$ $= 500\,000\,(83{,}33 - 83{,}33) = 0$. Если брать частное от деления с четырьмя значащими цифрами, то получим совсем другой результат:

$500\,000\,(83{,}3333 - 83{,}3227) = 500\,000\,(0{,}0106) = 5\,300.$

Как видим, при округлении промежуточных результатов с разной точностью, разница между двумя расчетами получилась огромная. Одна из причин ошибок округления — наличие сильно

отличающихся по значению чисел; так, например, в расчетах может быть накопленная сумма эффективных температур порядка 3000 °С и гидротермический коэффициент порядка 0,1. При наличии сильной зависимости между столбцами (переменными) в матрице **X** определитель матрицы **X′X** будет близок к нулю, что приводит также к большой потере точности расчетов коэффициентов модели, так как при вычислениях эта величина используется как делитель. В ряде программ вычисление определителя и дальнейшие расчеты прекращаются, если связь между независимыми переменными больше определенного значения.

Для того чтобы в значительной степени уменьшить ошибки округления, рекомендуется приводить переменные к одному масштабу измерения путем вычитания из переменных их средних значений и нормирования на подходящее число, например, на среднее квадратическое отклонение.

Есть еще одна немаловажная причина выполнения этих операций. Известно, что значения коэффициентов линейной модели зависят от единиц измерения независимых переменных. Можно построить уравнение зависимости урожайности от количества внесенных удобрений на один гектар посева. Если за единицу измерения количества внесенных удобрений взять кг/га, то регрессионный коэффициент будет в 1000 раз меньше, чем тот, который рассчитан для единиц измерения т/га. Из этого следует, что коэффициент определяется выбором единиц измерения. Большие трудности при интерпретации коэффициентов регрессионных уравнений возникают при одновременном использовании переменных, сильно различающихся по значению.

Для получения регрессионных коэффициентов, не зависящих от единиц измерения, используют стандартизованные переменные, т. е. такие переменные, которые приведены к одному размаху колебаний и одному среднему значению.

Стандартизованные переменные можно получить по формуле

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j},$$

из каждого наблюдения переменной вычитается соответствующее среднее $\bar{x}_j$ и результат делится на среднее квадратическое отклонение σ_j. После таких преобразований все переменные будут иметь нулевое среднее и единичную дисперсию. Если разделить z_{ij} на $\sqrt{n}$, то такая нормировка приводит матрицу **X′X** к корреляционному виду.

Уравнение множественной регрессии с переменными z_{ij} будет иметь вид

$$z_y = \beta_1' z_1 + \beta_2' z_2 + \ldots + \beta_m' z_m,$$

где z_y — стандартизованная зависимая переменная, β_i $(i = 1, 2, \ldots, m)$ — стандартизованные регрессионные коэффициенты. Обратим внимание, что в модели свободный член равен нулю.

Каждый коэффициент показывает, на сколько единиц стандартного отклонения изменится зависимая переменная z_y при изменении на одно стандартное отклонение независимой переменной при условии постоянства остальных независимых переменных.

Взаимосвязь обычных коэффициентов уравнения со стандартизованными задается соотношением

$$\beta_i' = \beta_i \frac{\sigma_{x_i}}{\sigma_y},$$

где σ_{x_i} и σ_y — средние квадратические значения соответствующего предиктора и предиктанта. Стандартизованные коэффициенты регрессионных уравнений можно сравнивать по значению и оценивать значимость независимых переменных, что сделать весьма затруднительно для обычных коэффициентов из-за различия масштабов шкал измерения. Например, в табл. 9.5 приведены обычные и стандартизованные коэффициенты уравнения для прогноза урожайности всех зерновых культур в Западном Казахстане. Относительные значения коэффициентов сильно изменились.

Таблица 9.5

Обычные и стандартизованные коэффициенты регрессионной модели для прогноза урожайности всех зерновых культур в Западном Казахстане

Фактор	Коэффициент модели	
	обычный	стандартизованный
Сумма осадков за апрель	0,065	0,267
Показатель влагообеспеченности июня	1,066	0,366
Сумма осадков за холодный период	0,014	0,474
Свободный член	—2,390	

Казавшийся малозначимым коэффициент при сумме осадков за холодный период стал самым большим по значению. Нагляднее проявилась роль включенных в уравнение (коэффициент множественной корреляции 0,84) агрометеорологических факторов.

9.6. ПОЛИНОМИАЛЬНЫЕ РЕГРЕССИОННЫЕ МОДЕЛИ

Наиболее часто встречающаяся форма нелинейных моделей — это полиномиальные модели. Линейные модели являются их частным случаем. Полиномиальная модель вида

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

называется моделью второго порядка с одной независимой переменной, так как степень независимой переменной равна двум. Графиком этой функции является парабола. Свободный член β_0 —

это значение зависимой переменной при $x=0$, коэффициенты β_1 и β_2 — соответственно линейная и нелинейная составляющие. Использовать эти уравнения для экстраполяции следует с большой осторожностью, так как вне интервала ее «подгонки» характер зависимости может резко меняться на противоположный. Можно рекомендовать полиномы второй степени для построения зависимости с одним экстремумом (рис. 9.11 *а*).

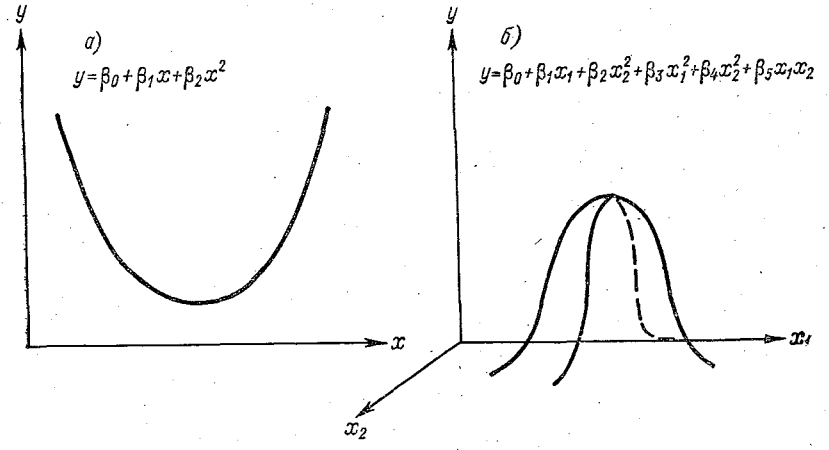


Рис. 9.11. Графическое представление полинома второй степени с одной (*а*) и двумя (*б*) переменными.

Модели

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3$$

являются полиномами третьей степени с одной переменной. Их, как и полиномы второй степени, часто используют для аппроксимации тренда урожайности сельскохозяйственных культур.

По мере накопления статистических данных становится все более очевидным, что простое линейное повышение урожайности большинства культур в определенный период замедлилось, однако в последние годы в связи с широким внедрением интенсивных технологий опять наблюдается ее заметный рост. В такой ситуации хорошо зарекомендовавшие себя линейные и параболические зависимости стали непригодны. Здесь уместно использовать тренды третьей и более высоких степеней. Однако надо быть достаточно осторожным, так как при повышении степени (более 5) могут быстро возрастать ошибки расчетов. Полиномы высоких степеней для экстраполяции почти непригодны. При использовании полиномов выше второй степени необходимо при расчетах брать достаточно большое количество значащих цифр после запятой, иначе ошибки могут быть слишком большие.

Например, был рассчитан тренд-полином третьей степени — по данным об урожайности озимой пшеницы. Расчет уровня тренда в последний год наблюдений по уравнению с двумя зна-

чащими цифрами после запятой отличается от расчетов с пятью цифрами на большую величину — 3 ц/га.

Модель

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

является моделью второго порядка с двумя переменными. В уравнение включены не только линейные и нелинейные компоненты от каждой переменной, но и член, отражающий взаимное влияние переменных x_1 и x_2. С помощью такой модели описывают в трехмерном пространстве фигуру типа «холм» (рис. 9.11 *б*). Полиномиальные модели такого типа с двумя или большим числом переменных (при достаточности данных) — удобный инструмент для грубого описания зависимостей, встречающихся в агрометеорологии.

При описании реальных данных можно заведомо усложнить вид модели, а затем, проверив значимость коэффициентов, отбросить некоторые из них.

Отметим, что при повышении степени полинома ошибка уравнения не может увеличиваться, она, как правило, уменьшается. Так, с помощью полинома n-й степени можно добиться нулевой ошибки при его построении по $n+1$ точкам, так как полином пройдет через все точки.

Пример. По данным об урожайности всех зерновых культур в Центральночерноземном экономическом районе за 42 года был построен тренд — полином второй степени (рис. 9.12):

$$y = 1{,}656 + 0{,}971x - 0{,}0150x^2.$$

Решим вопрос, необходимо ли было повышать степень полинома до двух или можно воспользоваться линейной зависимостью. Проверяем гипотезы:

$$H_0 : \beta_2 = 0; \quad H_1 : \beta_2 \neq 0.$$

Нулевая гипотеза состоит в том, что нелинейность отсутствует. Рассчитаем t-статистику

$$t = \frac{\hat{\beta}_2}{V(\hat{\beta}_2)} = \frac{-0{,}0150}{0{,}00369} = -4{,}063.$$

Для уровня значимости 0,05 по таблице находим значение $t_0(\alpha;\ n-3)$ для 42 наблюдений: $t_0(0{,}05{,}93) = 2{,}02$.

Решающее правило таково:

если $|t| \leqslant 2{,}02$, то верна гипотеза H_0,

если $|t| > 2{,}02$, то верна гипотеза H_1.

В нашем примере $|t| = 4{,}063$, что больше 2,02; отсюда делаем вывод о необходимости использования полинома второй степени при аппроксимации тренда урожайности. Аналогично можно проверить необходимость включения в уравнение переменной любой степени.

Очень полезно бывает выяснить, используя уравнения второй степени, при каких значениях независимой переменной исследуемая функция достигает экстремума.

Если имеем квадратное уравнение

$$y = \beta_0 + \beta_1 x + \beta_2 x^2,$$

то экстремум зависимой переменной приходится на

$$x^* = -\frac{\beta_1}{2\beta_2}$$

и равен

$$y^* = \beta_0 - \frac{\beta_1^2}{4\beta_2}.$$

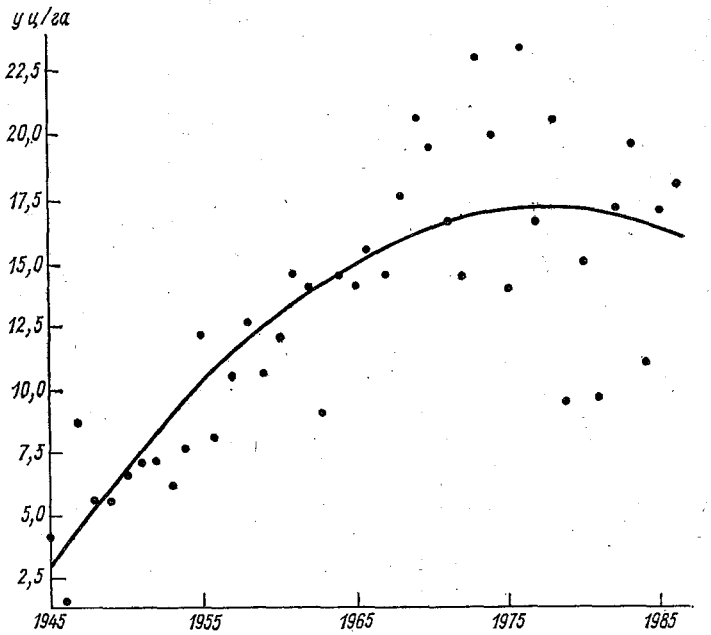


Рис. 9.12. Тренд урожайности (y) зерновых культур в Центральночерноземном районе.

Пример. На рис. 6.1 представлена параболическая зависимость урожайности озимой пшеницы от весенних запасов влаги в метровом слое почвы при сильном загущении посевов

$$y = -0{,}0052x^2 + 1{,}77x - 116{,}98.$$

Максимального значения зависимая переменная y достигает при

$$x^* = -\frac{1{,}77}{-0{,}0104} = 170{,}2 \text{ мм},$$

оно равно

$$y^* = -116{,}7 + \frac{1{,}77^2}{4 \times 0{,}0052} = 33{,}9 \text{ ц/га}.$$

Наряду с удобством и простотой расчетов полиномиальные модели имеют и недостатки. Во-первых, это наличие большего числа членов по сравнению с другими нелинейными моделями, описывающими данные с аналогичной точностью; во-вторых, это коррелированность используемых переменных (степеней исходных переменных), которая особенно сильно проявляется при небольшом размахе изменений независимой переменной и слабой нелинейности.

9.7. ЭЛИМИНИРОВАНИЕ ТРЕНДОВ УРОЖАЙНОСТИ

Формирование урожая сельскохозяйственных культур — сложный многообразный процесс, зависящий от ряда природно-климатических и экономических факторов. В настоящее время прогнозирование урожая ведется в двух взаимодополняющих направлениях, учитывающих основные группы влияющих факторов — природно-климатические и хозяйственно-экономические. При прогнозировании урожая, основанном на учете природно-климатических факторов, применяют самые различные агрометеорологические методы. При прогнозировании же, основанном на учете изменяющихся хозяйственно-экономических условий, основное внимание уделяется экстраполяции на будущее и прогнозированию именно тех условий, которые определяют общий уровень земледелия на фоне изменения природно-климатических факторов.

Анализ рядов динамики урожайности большинства сельскохозяйственных культур свидетельствует о неуклонном повышении урожайности во времени, что обусловлено объективным изменением эффективности общественного производства, применением новых методов хозяйствования. Чтобы выделить из временных рядов урожайности влияние этих существенных факторов, в агрометеорологии используют понятие тенденции, или тренда урожайности. Придавая последнему термину сходное содержание — тенденции изменения урожайности во времени,— иногда его трактуют по-другому. Одни исследователи, не учитывая изменения почвенно-климатических условий, определяют тренд при условии сохранения их среднего уровня, другие понимают под трендом функцию, описывающую общее среднее статистическое изменение уровня урожайности. В первом случае отказываются от учета короткопериодических изменений климата, во втором — достаточно сильно абстрагируются от конкретного характера как климатического, так и экономического факторов.

Использование трендов при прогнозе урожайности имеет двоякую цель: 1) элиминировать ту долю урожая, которая определяется уровнем земледелия в широком смысле слова; 2) экстраполировать динамику тренда на перспективу. Постановка этих задач

обусловлена тем, что в агрометеорологической литературе рассматривают динамический ряд временной урожайности как нестационарный процесс

$$y(t) = \varphi(t) + \varepsilon(t), \tag{9.5}$$

где t — время, $\varphi(t)$ — неслучайная функция, $\varepsilon(t)$ — случайная функция, $y(t)$ — урожайность.

Функцию $\varphi(t)$ определяют как тренд урожайности, характеризующий изменение уровня земледелия. Дискретная функция $\varepsilon(t)$ описывает случайные флуктуации урожайности под воздействием метеорологических факторов в вегетационный период конкретного года. Она используется в агрометеорологических моделях прогнозирования урожайности; так как, исходя из представления (9.5) считают, что значения $\varepsilon(t)$ обусловлены главным образом метеорологическими условиями конкретного года.

Широкий набор методов эмпирического оценивания трендов способствует их повсеместному использованию в практике оперативных агрометеорологических прогнозов и для экстраполяции тенденции урожайности на перспективу. Однако представляется, что эти две задачи требуют при решении построения трендов, обладающих специфическими свойствами.

Рассмотрим кратко один из возможных способов оценки тренда ряда урожайности при построении регрессионных прогностических моделей с использованием принципа внешнего дополнения, предложенный в [7]. Сущность метода состоит в привлечении дополнительной «внешней» информации об интересующем нас явлении при выборе полинома для интерполяции тренда урожайности. Остаточная сумма разностей является внутренним критерием, она представляет собой убывающую функцию от степени полинома. Статистические критерии значимости уменьшения остаточной суммы разностей часто не пригодны с неформальной точки зрения [10].

Сначала проверим гипотезу о том, что детерминированная функция в представлении (9.5) есть константа, т. е. подтвердим или отбросим гипотезу о стационарности динамического ряда. Для этого применим различные критерии: непараметрический критерий Вальда—Вольфовитца для общего числа серий, критерий Рамачадана—Ранганатана для сумм квадратов длин серий [34], критерий, основанный на числе инверсий во временном ряду [3]. По всем критериям и для всех областей Казахстана при уровне значимости 5 % гипотеза о наличии тренда урожайности зерновых культур отвергается. Далее попытаемся выделить, исходя из модели временного ряда (9.5), $\varepsilon(t)$, используя аналитическое выравнивание ряда урожайности $y(t)$ таким образом, чтобы полученный тренд максимально приближал случайную функцию $\varphi(t)$, которую при решении задачи о выделении метеорологически обусловленной случайной компоненты урожайности и принимаем за тренд урожайности.

В качестве меры близости тренда к $\varphi(t)$ выберем функционал $R(\varphi)$, представляющий собой коэффициент парной корреляции отклонений урожайности от трендов с агрометеорологическими условиями вегетационного периода. Функция из определенного класса, на которой $R(\varphi)$ достигает максимума, и является в указанном смысле трендом урожайности. Выбор функции осуществлялся в классе полиномов. В [7] показана эффективность применения предложенной методики при выделении метеорологически обусловленной компоненты временных рядов урожайности зерновых культур в областях Казахстана в целях создания прогностических уравнений. Для построения трендов использовались областные ряды урожайности всех зерновых культур с 1958 по 1980 г.

Расчет семейства сглаживающих полиномов осуществлялся с помощью метода наименьших квадратов. Были получены уравнения полиномов от первой до четвертой степеней. Расчеты показали, что использование полиномов более высоких степеней нецелесообразно из-за уменьшения точности вычислений и малой длины временных рядов.

Одним из приемов, используемых для обоснования правильности выбора тренда, является испытание временного ряда отклонений от тренда на случайность [33]. Были проведены также испытания для наилучших и всех других трендов. Результат оказался одинаковым — последовательности отклонений были случайными. Для проверки использовались критерии Вальда—Вольфовитца и Рамачандана—Ранганатана. Оказалось, что случайность отклонений от наилучшего тренда является необходимым, но не достаточным условием. Необходимы дополнительные критерии, одним из таких критериев может служить максимум описанного функционала $R(\varphi)$.

При выделении тренда урожайности выбор интегрального показателя, характеризующего агрометеорологические условия вегетационного периода, — важная, но не простая задача. Необходимо учитывать метеорологические условия за большой отрезок времени. В качестве такого показателя, например, можно выбрать первую главную компоненту декадных среднеобластных сумм осадков и температуры воздуха за май—август, позволяющую учесть самый важный период вегетации зерновых культур.

Расчеты показали, что для Северо-Казахстанской и Чимкентской областей первая главная компонента из-за наличия в данных аномально высоких значений осадков не является адекватной формой выражения условий вегетационного периода. Для этих областей таким показателем является хорошо зарекомендовавшая себя первая главная компонента из множества одиннадцати — соответствующих тому же периоду декадных показателей гидротермического коэффициента, полученных делением декадных сумм осадков на среднюю температуру воздуха (на главные компоненты значительное влияние оказывает преобразование масштаба данных).

Коэффициент парной корреляции используемого обобщенного показателя с отклонениями урожайности от трендов, выбранных в результате примененной процедуры, колеблется в значительной степени: от 0,31 в Целиноградской области до 0,80 в Восточно-Казахстанской. В качестве трендов, наилучшим образом выделяющих метеорологически обусловленную компоненту урожайности, для Актюбинской, Алма-Атинской, Восточно-Казахстанской, Целиноградской. Джамбулской, Карагандинской, Павлодарской, Семипалатинской, Тургайской и Уральской областей были использованы полиномы первой степени. Для Кокчетавской, Талды-Курганской областей применялись полиномы второй степени, для Чимкентской, Северо-Казахстанской, Кустанайской — полиномы третьей степени.

В табл. 9.6 приведена доля дисперсии, приходящаяся на первую главную компоненту рассматриваемых метеорологических факторов, их коэффициенты корреляции с метеорологически обусловленной долей временных рядов урожайности для областей Казахстана; кроме того, там представлены множественные коэффициенты корреляции уравнений регрессии, где в качестве независимых переменных взяты первые пять главных компонент метеорологических факторов, а в качестве зависимой переменной — урожайность всех зерновых культур. Заметим, что во всех пятнадцати областях первые пять компонент описывают примерно одинаковую долю общей дисперсии урожайности — от 60 до 69 %.

Таблица 9.6

Доля дисперсии первой главной компоненты декадных сумм осадков и температуры воздуха (σ_1^2), коэффициент корреляции первой главной компоненты (R_1) и множественный коэффициент корреляции пяти первых главных компонент (R_{1-5}) с отклонением урожайности от тренда, коэффициенты экстраполирующего тренда урожайности $y = a_0 + a_1 t$

Область	σ_1^2	R_1	R_{1-5}	a_0	a_1
Уральская	0,189	0,623	0,851	7,744	0,545
Восточно-Казахстанская	0,249	0,467	0,912	9,528	1,933
Северо-Казахстанская	0,220	0,448	0,863	8,088	3,677
Джамбулская	0,242	0,554	0,860	8,077	1,536
Карагандинская	0,185	0,480	0,872	4,592	1,387
Павлодарская	0,181	0,546	0,881	3,745	1,518
Талды-Курганская	0,283	0,551	0,945	11,119	1,833
Алма-Атинская	0,215	0,620	0,913	9,633	1,050
Целиноградская	0,237	0,308	0,856	5,858	2,370
Чимкентская	0,256	0,406	0,666	5,868	3,766
Кокчетавская	0,192	0,393	0,858	6,020	3,381
Кустанайская	0,165	0,329	0,878	6,232	2,470
Семипалатинская	0,249	0,658	0,919	7,529	—0,615
Тургайская	0,220	0,267	0,905	6,746	0,875
Актюбинская	0,173	0,426	0,834	5,579	0,458

Таким образом, в исследованиях и разработках прогностических моделей урожайности всех зерновых в областях Казахской ССР в качестве независимой переменной может использоваться освобожденная от тренда метеорологически обусловленная доля временных рядов урожайности.

Вторым моментом, связанным с понятием тренда, как уже указывалось, является построение такого тренда, который будет использован для экстраполяции тенденции урожайности на будущее. Он должен мало зависеть от случайных колебаний в последних точках временных рядов, так как небольшие изменения его формы могут при экстраполяции привести к существенным расхождениям результатов. Необходимо, чтобы экстраполирующий тренд с бо́льшим весом усваивал более «свежие» данные и полнее отражал тенденции последнего времени. При расчете экстраполирующего тренда решающая роль принадлежит исследователю. Без внешней информации о характере процессов, их тенденций трудно задать математически адекватную форму модели тренда. В нашем конкретном случае это — выбор формы тенденции роста урожайности на будущее. Несомненно, что выбранная функция должна быть растущей со временем, так как процессы, ведущие к росту урожайности сельскохозяйственных культур, не ослабевают [24].

В качестве экстраполирующего тренда урожайности на краткосрочную перспективу (3—5 лет) нами был выбран полином первой степени. Принимаем гипотезу о том, что тенденции роста урожайности на эту перспективу сохранятся на современном уровне.

Для построения линейных трендов использовался следующий подход. Каждой точке временного ряда был приписан вес, линейно возрастающий с ростом номера года. Тем самым тенденции последних лет вносили больший «вклад» в рассчитанные коэффициенты. Для устойчивости линейного тренда к резким колебаниям метеорологически обусловленной компоненты применялся подход построения робастной регрессии, описанный в разделе 11.2. Тренд в форме полинома первой степени рассчитывался итерационным методом взвешенных наименьших квадратов с весами

$$w_t(h,\ \varepsilon) = \begin{cases} 1 + \alpha t & \text{при } |\varepsilon| \leqslant h,\ t = 0{,}1;\ 0{,}2;\ \ldots;\ 2{,}3, \\ h/|\varepsilon| + \alpha t & \text{при } |\varepsilon| > h,\ t = 0{,}1;\ 0{,}2;\ \ldots;\ 2{,}3, \end{cases}$$

где α — коэффициент, выражающий повышение «значимости» новых точек временного ряда урожайности. В наших расчетах после ряда экспериментов было принято $\alpha = 0{,}5$, $h = 1$. Значение параметра h, выраженное в единицах среднего квадратического отклонения урожайности от тренда, является тем пределом отклонения метеорологически обусловленной доли урожайности от тренда, после которого влияние отклонения начинает убывать как $1/|\varepsilon|$. Для удобства расчетов порядковый номер точки временного ряда уменьшен в 10 раз.

Многие временные ряды урожайности по областям Казахстана имеют первые три—четыре точки, значительно бóльшие по значению, чем в последующее десятилетие. Это объясняется повышенным плодородием целинных земель в первые годы после их вовлечения в сельскохозяйственный оборот. Наличие такой неод-

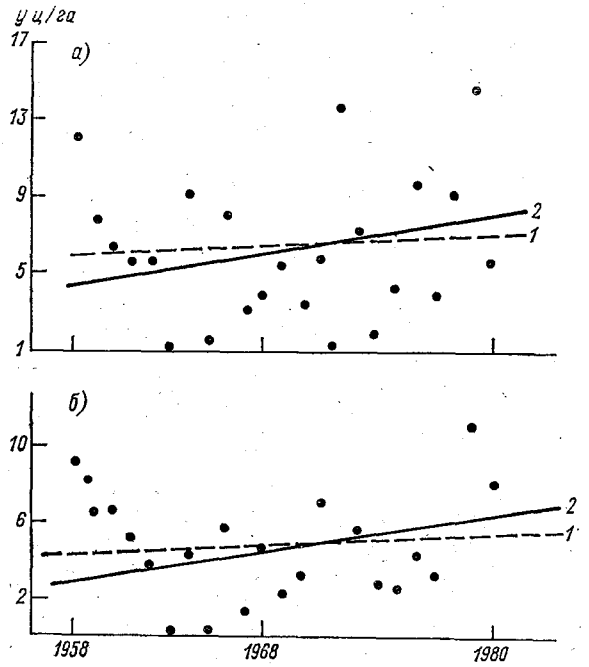


Рис. 9.13. Временные ряды урожайности (y) зерновых и их экстраполирующие тренды: обычный линейный (*1*), линейный взвешенно-робастный (*2*).

а — Карагандинская область, *б* — Павлодарская область.

нородности во временных рядах приводит к тому, что при проведении линейного экстраполяционного тренда по всему временному ряду методом наименьших квадратов общая тенденция повышения уровня земледелия в некоторых областях становится очень слабой, что противоречит реально существующей динамике за последние 10—15 лет. Подобные обычные линейные тренды не могут использоваться для экстраполяции уровня земледелия на перспективу.

При построении трендов по вышеописанной методике влияние первых лет временного ряда с аномально высокими значениями урожайности снижается. Тенденции роста урожайности ряда последних лет выражены более наглядно.

На рис. 9.13 показаны временные ряды урожайности в Павлодарской и Карагандинской областях, проведены обычный и пред-

ложенный нами линейные тренды. Коэффициент, отражающий годовой прирост тренда урожайности в Павлодарской области, у первого тренда равен 0,680, у второго — 1,518, что в 2,23 раза больше. Даже при экстраполяции уровня земледелия всего на три года вперед расхождения между ними значительны и составляют 0,84 ц/га. Аналогичная ситуация наблюдается и в Талды-Курганской области, где темпы прироста — соответственно 0,557 и 1,833, или 0,056 и 0,183 ц/га в год.

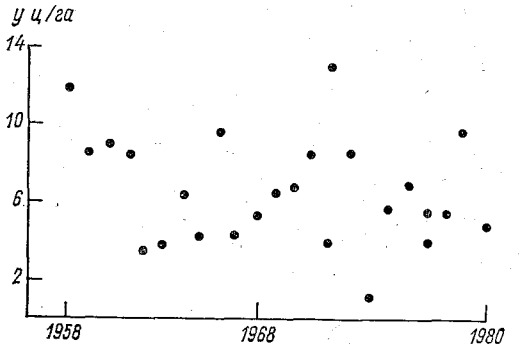


Рис. 9.14. Временной ряд урожайности (y) зерновых культур в Семипалатинской области.

В табл. 9.6 представлены коэффициенты линейных экстраполяционных трендов областных урожайностей всех зерновых. Обращает на себя внимание слабо отрицательный прирост уровня земледелия в Семипалатинской области, составляющий —0,062 ц/га в год. При обычном проведении тренда он еще больше —0,085 ц/га в год. Даже при более значительном увеличении веса «свежих» точек временно́го ряда коэффициент остается отрицательным. На рис. 9.14 показан временной ряд урожайности зерновых в Семипалатинской области. Здесь трудно выявить визуально наличие какого-либо возрастающего тренда или тенденцию к повышению уровня земледелия.

Типичным представителем второго направления прогнозирования урожайности сельскохозяйственных культур является метод, разработанный Г. В. Менжулиным. Он предлагает вычислять тренд урожайности с использованием информации о динамике основных экономических показателей сельскохозяйственного производства. Так, в [18] рассчитаны тренды урожайности пшеницы в СССР. Влияющими показателями были выбраны число зерноуборочных комбайнов и тракторов в пересчете на стандартную мощность и количество внесенных минеральных удобрений, расходы электроэнергии в сельском хозяйстве. Основным выражением, определяющим урожайность пшеницы, является

$$y_i = c + \frac{1}{a_1 + \sum_j \frac{a_j}{x_{ij}}},$$

где c, a_1, a_j — эмпирические константы, y_i, x_{ij} — урожайность и экономические показатели i-го года.

Эта форма задает зависимость урожайности, которую можно представить в виде графиков с горизонтальными асимптотами по каждому экономическому показателю, т. е. с «насыщением».

Эмпирические коэффициенты получают методом наименьших квадратов. Учитывается также наличие различных сортов возделываемой пшеницы.

Этот метод позволяет выявить особенности зависимости урожайности от технико-экономических факторов во времени. Коэффициенты c, a_1, a_j достаточно сильно зависят от выбора влияющих факторов. Подставляя в модель плановые значения экономических показателей на предстоящие годы, можно прогнозировать урожайность сельскохозяйственных культур с большой заблаговременностью.

В силу неуклонного роста уровня культуры земледелия проблема построения трендов урожайности сельскохозяйственных культур занимает значительное место в агрометеорологических исследованиях последних лет, некоторые процедуры стали традиционными. Однако при большом разнообразии конкретных условий и задач нельзя автоматически применять известные приемы во всех случаях, необходима разработка новых методов. Дополняет арсенал таких методов и позволяет точнее выделить метеорологически обусловленную компоненту рядов урожайности предлагаемый достаточно общий подход построения тренда урожайности зерновых с использованием в требуемом для решения задачи виде информации об агрометеорологических условиях сезона вегетации. Наряду с использованием метода, экстраполирующего устойчивость тренда, данный метод позволяет создавать наиболее адекватные статистические модели для прогноза урожайности сельскохозяйственных культур.

Глава 10. ПОИСК НАИЛУЧШЕГО НАБОРА ПРЕДИКТОРОВ

Исследователь, пользующийся методами регрессионного анализа, надеется, что, выбрав некоторые независимые переменные, он сможет адекватно описать изучаемый объект, например, достаточно точно предсказать будущий урожай по агрометеорологическим предикторам. Для этого необходимо правильно выбрать форму модели и отобрать наиболее информативные предикторы. Из полного множества имеющихся в распоряжении и реально влияющих на сельскохозяйственные культуры факторов среды, перечень которых может насчитывать несколько десятков переменных, необходимо выделить какое-либо подмножество. Построить модель по всем данным, как правило, невозможно из-за ее нестабильности и плохих прогнозирующих способностей.

Выбор наиболее подходящего для прогноза подмножества

переменных для регрессионной модели — достаточно сложная проблема. Во-первых, необходимо найти критерий для сравнения различных подмножеств переменных. Во-вторых, это вычислительные трудности, так как возможных вариантов сочетания предикторов может быть огромное количество. Для выбора наилучшего набора предикторов можно использовать различные критерии и численные процедуры. В идеальном виде такой набор можно представить как набор потенциальных предикторов, включающий все важнейшие переменные и полезные для аппроксимации функции плюс некоторые менее важные и совсем бесполезные для анализа независимые переменные, и считать, что независимые переменные хорошо описывают поведение зависимой переменной.

10.1. КРИТЕРИИ КАЧЕСТВА ПРИ ВЫБОРЕ МОДЕЛЕЙ

Среди используемых критериев наиболее популярным при сравнении моделей долгое время был коэффициент множественной корреляции R или коэффициент детерминации R^2. Напомним, что R^2 — квадрат коэффициента корреляции — показывает количественную связь между независимой переменной y и линейной комбинацией независимых переменных. Для уравнения с q параметрами его вычисляют по формуле

$$R_q^2 = 1 - \frac{\text{СКО}_q}{\text{ПСК}}.$$

Численно он выражает долю дисперсии независимой переменной y, объясненную с помощью регрессионного уравнения. Чем больше R_q^2, тем большую долю дисперсии могут описать переменные, включенные в модель. Однако этот критерий не пригоден для процедур отбора подмножества предикторов, так как при сравнении та модель, которая включает более широкое подмножество предикторов, всегда будет иметь большее значение R^2. При включении в регрессионное уравнение новой переменной коэффициент корреляции не может уменьшаться: он не изменяется или, как правило, увеличивается. Отсюда следует, что своего максимума коэффициент детерминации достигает при построении модели по всему множеству предикторов. Критерий R^2 можно использовать при выборе лучшего подмножества, если число предикторов фиксированно. В то же время коэффициент детерминации R^2 можно рассматривать не только как меру качества построенной модели в смысле близости точек к поверхности регрессии, но и как меру ее крутизны. Например, если зафиксировать сумму квадратов разности относительно регрессионной поверхности, то с ростом ее крутизны $\sum_i (y_i - \bar{y})^2$ будет увеличиваться, и, тем самым, будет расти R^2. Поэтому при анализе двух или более наборов данных предсказание по регрессионному уравнению с большей крутизной регрессионной поверхности может не быть более точным,

чем прогноз, полученный по уравнению с меньшим наклоном регрессионной поверхности и меньшим значением R^2. Для зависимостей, представленных на рис. 10.1, коэффициенты детерминации соответственно равны 0,81 и 0,76, хотя на первом графике разброс точек относительно линии регрессии значительно больше, чем на втором. Угол наклона прямой при $R^2 = 0{,}76$ равен 23°.

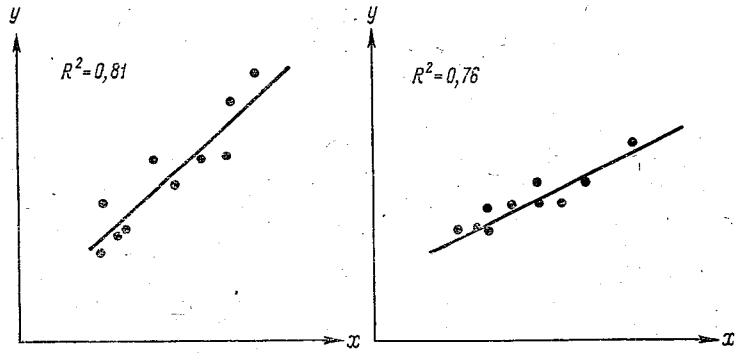


Рис. 10.1. Зависимость коэффициента детерминации (R^2) от крутизны графика прямой.

Если вращать прямую относительно неподвижной точки $(\bar{x}, \bar{y})$, через которую она проходит, то при увеличении крутизны прямой и зафиксированном значении $\sum (\hat{y}_i - y_i)^2$ величина R^2 будет принимать значения:

Угол наклона, ...° . .	0	23	30	45	60
R^2	0	0,76	0,85	0,94	0,98

Коэффициент детерминации быстро приближается к единице при стремлении угла наклона к 90°.

Математическое ожидание R^2 регрессионного уравнения в случае, когда его истинное значение равно нулю, определяется формулой

$$E(R^2) = \frac{p}{n-1},$$

где p — количество предикторов, n — длина выборки. Отсюда следует, что даже при нулевом истинном коэффициенте корреляции весьма вероятно получить большое значение R^2, если количество предикторов p сравнимо с длиной выборки n, что на практике бывает достаточно часто.

Вместо неубывающей функции от числа предикторов R^2 в качестве критерия качества моделей используют его модификацию —

скорректированный коэффициент детерминации, определяемый как

$$\bar{R}_p^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R_p^2).$$

Эта статистика, как и обычный коэффициент детерминации, является простой функцией от СКО

$$\bar{R}^2 = 1 - \frac{\text{СКО}}{\sum_i (\hat{y}_i - y_i)^2 (n-p)}.$$

Свойства этой статистики таковы, что в отличие от обычного R^2 не при всяком включении новой переменной ее значение увеличивается. Это происходит только в случае, если F-статистика при проверке гипотезы о значимости включаемой переменной больше или равна единице. В противном случае включение нового предиктора уменьшает значение $\bar{R}^2$. Наилучшим регрессионным уравнением можно считать уравнение с подмножеством переменных, обеспечивающим наибольшее значение $\bar{R}_p^2$. Обычно график достаточно гладкий, максимум выражен слабо.

В отличие от коэффициента детерминации скорректированный коэффициент при большом значении p может быть отрицательным, если n мало. Этот критерий можно с бо́льшим успехом применять для отбора наилучшего подмножества предикторов в регрессионное уравнение.

Рассмотрим еще один показатель статистической связи между двумя переменными, называемый частным коэффициентом корреляции. На практике, как правило, бывает трудно выделить «чистое» влияние каждой независимой переменной на зависимую вследствие коррелированности агрометеорологических факторов. Обычный парный коэффициент корреляции отражает влияние не только одной переменной, но и всех остальных, имеющих с ней тесную связь.

Частный коэффициент используется как мера линейной связи между зависимой переменной y и какой-либо одной из переменных $x_1, x_2, \ldots, x_p$, после удаления влияния на нее всех оставшихся переменных.

Более наглядно это можно представить, используя остатки или разности регрессионных уравнений. Пусть $\boldsymbol{\varepsilon}^y$ — вектор разности (ошибок) регрессионного уравнения

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^y + \boldsymbol{\varepsilon}^y,$$

где $\mathbf{X}$ — матрица наблюдений за переменными $x_2, x_3, x_4, \ldots, x_p$. Аналогично, пусть $\boldsymbol{\varepsilon}'$ — вектор разности уравнения

$$\mathbf{x}_1 = \mathbf{X}\boldsymbol{\beta}' + \boldsymbol{\varepsilon}',$$

где $\mathbf{X}$ — матрица наблюдений за этими же переменными x_2, x_3, x_4, ..., x_p. Тогда простой коэффициент корреляции между соответствующими элементами векторов ε^y и ε' будет частным коэффициентом корреляции между зависимой y и независимой переменными x_1, «очищенным» от влияния переменных x_2, x_3, x_4, ..., x_p. Обычно его обозначают $r_{yx_1 \cdot x_2 x_3 \ldots x_p}$. Как и парный коэффициент корреляции, частный коэффициент обладает свойством симметрии

$$r_{yx_1 \cdot x_2 x_3 \ldots x_p} = r_{x_1 y \cdot x_2 x_3 \ldots x_p}.$$

Полезно сделать следующие замечания. Если необходимо рассчитать частный коэффициент корреляции при устранении влияния только одной переменной ($r_{yx_1 \cdot x_2}$), проще всего это сделать, используя парные коэффициенты корреляции

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1 x_2}^2)}}.$$

На основе частного коэффициента корреляции, полученного при устранении влияния одной переменной, легко рассчитывается коэффициент корреляции при устранении влияния двух переменных:

$$r_{yx_1 x_2 x_3} = \frac{r_{yx_1 \cdot x_2} - r_{yx_1 \cdot x_2} r_{x_1 x_3 \cdot x_2}}{\sqrt{(1 - r_{yx_3 \cdot x_2}^2)(1 - r_{x_1 x_3 \cdot x_2}^2)}}.$$

С помощью этой рекуррентной формулы после последовательных вычислений («удаление» раз за разом очередной независимой переменной) можно получить частный коэффициент корреляции между y и x_1 при устранении влияния всех остальных переменных $r_{yx_1 \cdot x_2 x_3 \ldots x_p}$.

Такого же результата можно достичь более простым путем, используя обращение расширенной корреляционной матрицы всей системы переменных, включающей и зависимую переменную. Для этого квадрат частного коэффициента корреляции интерпретируют как долю остаточной дисперсии зависимой переменной y, «объясненную» включением дополнительной переменной в набор уже использованных в регрессионной модели переменных. Чем ближе абсолютное значение этого коэффициента к единице, тем сильнее линейная зависимость y от исследуемой переменной. Тест для проверки гипотез значимости отличия от нуля коэффициента частной корреляции между переменными x_i и x_j при устранении линейного влияния набора переменных с $H_0 : r_{ij \cdot c} = 0$ эквивалентен проверке гипотезы $H_0 : \beta_{ij} = 0$, где β_{ij} — коэффициент регрессии в зависимости x_i от x_j. Используется t-критерий:

$$t_{(n-p-1)} = \sqrt{(n - m - 1) \frac{r_{ij}^2}{1 - r_{ij}^2}},$$

где p — число переменных в наборе c. Если нулевая гипотеза истинная, то статистика подчиняется t-распределению с $(n-p-1)$ степенями свободы.

Частные коэффициенты корреляции широко применяются для ранжирования факторов по значимости их влияния на зависимую переменную при включении их в регрессионную модель шаговыми методами.

Пример. Изучалось влияние агрометеорологических факторов на урожайность зерновых в Западном Казахстане. Использовался шаговый алгоритм выбора в уравнение наиболее информативных предикторов. В качестве потенциальных предикторов были выбраны показатели агрометеорологических условий за отдельные периоды вегетации. Были рассмотрены тесно связанные между собой факторы, так как они «покрывали» пересекающиеся календарные периоды. Рассчитанные парные и частные коэффициенты корреляции при элиминировании некоторых переменных приведены в табл. 10.1.

Таблица 10.1

Парные и частные коэффициенты корреляции агрометеорологических факторов с урожайностью всех зерновых культур в Западном Казахстане при элиминировании некоторых переменных из регрессионных моделей

Фактор	Коэффициент корреляции			
	парный	частный при элиминировании переменных		
		№ 7	№ 7, 8	№ 4, 7, 8
1. Показатель влагообеспеченности за VI_3	0,49	—0,08	—0,16	—0,19
2. Показатель влагообеспеченности за VI_3—VII_1	0,37	0,12	0,36	0,21
3. Показатель влагообеспеченности за VI_2—VI_3	0,64	—0,03	—0,17	—0,05
4. Сумма осадков за апрель	0,46	0,29	0,41	
5. Сумма осадков за май	0,33	—0,34	—0,34	—0,31
6. Показатель влагообеспеченности за VI_1—VI_2	0,63	0,08	0,16	0,19
7. Показатель влагообеспеченности за июнь	0,65			
8. Сумма осадков за холодный период	0,67	0,55		

Примечание. Здесь VI_3 — третья декада июня и т. д.

Как видим, частные коэффициенты корреляции сильно зависят от того, какие переменные уже «объяснили» часть дисперсии предиктанта, т. е. от их взаимосвязанности. Естественно, что

после учета влияния показателя влагообеспеченности за июнь, роль показателя влагообеспеченности за часть месяца (№ 6) будет почти нулевой. Частный коэффициент корреляции равен 0,08, в то время как парный был достаточно высоким: 0,63.

В последнее время при выборе наилучшего подмножества предикторов часто пользуются показателем, предложенным Мэллоусом — C_p. Он основывается на средней ошибке предсказания зависимой переменной по модели, измеряет сумму квадратов смещения и дисперсию ошибки прогноза по всем n данным наблюдений и является простой функцией остаточной суммы квадратов для построенного регрессионного уравнения

$$C_p = \frac{\text{СКО}_p}{\sigma^2} - (n - 2p),$$

где σ^2 — оценка дисперсии ошибки уравнения, содержащего все переменные; n и p — соответственно длина выборки и число параметров в регрессионной модели. Известно, что модели с малым смещением имеют тенденцию к группировке около линии $C_p = p$. Уравнения со значительным смещением будут характеризоваться C_p, лежащими выше этой прямой. Графический анализ зависимости C_p от p позволяет выявлять некоторые аспекты структуры данных и трудности их интерпретации в случае, когда эти данные из-за различных причин не соответствуют предъявляемым к ним требованиям. С помощью графического представления C_p можно выбрать необходимое подмножество независимых переменных, соответствующих целям моделирования. Мэллоуз рекомендует выбирать модель с отрицательным или малым положительным значением $C_p - p$.

Оригинальный критерий для выбора наилучшего подмножества предикторов предложил Аллен [25]. Алгоритм его вычисления очень прост. На первом шаге из имеющегося набора n результатов опытов (наблюдений) удаляется первое наблюдение и регрессионная модель строится по оставшимся $n - 1$ наблюдениям. Первое наблюдение считается проверочной независимой выборкой, состоящей из одного опыта. На нем проверяется построенная модель — рассчитывается ошибка прогноза. На втором шаге первое наблюдение возвращается в «зависимую» совокупность, при этом удаляется второе по счету наблюдение. Снова строится модель по $n - 1$ наборам данных и проверяется точность прогнозирования на одном втором независимом опыте. Таким образом повторяются шаги вплоть до исключения n-го набора данных. Для каждого подмножества предикторов вычисляется средний квадрат ошибки прогнозов по всем n независимым точкам:

$$\text{PRESS} = \frac{\sum_i (y_i - \hat{y}_{-i})}{n}; \qquad (10.1)$$

предпочтение отдается модели с наименьшим значением этого показателя. При этом надо стремиться, чтобы модель не имела много параметров. Этот критерий отражает прогностические возможности уравнения на независимых данных.

В отличие от показателей C_p, $\bar{R}^2$ и R^2 эта статистика через остаточную сумму квадратов не выражается, и ее свойства недостаточно изучены. Однако эта статистика отражает новую грань модели, отличную от других критериев качества. Однако при больших выборках, как видно из (10.1), значение PRESS будет близко к остаточной сумме квадратов, так как отсутствие одного случая слабо отразится на коэффициентах уравнения.

В последние годы возможность использования статистик типа PRESS в метеорологии привлекает значительное внимание исследователей [2].

10.2. ПРОЦЕДУРЫ ОТБОРА ПРЕДИКТОРОВ

Наряду с выбором критерия качества модели второй важной проблемой является поиск лучшего уравнения или небольшой группы уравнений из огромного множества возможных. Это обусловлено экспоненциальным ростом числа подмножеств сочетаний предикторов. При небольшом количестве переменных (6 или меньше) можно без труда рассчитать критерии для всех возможных уравнений и выбрать из них лучшие. При 10 потенциальных предикторах существует 1024 возможных подмножеств переменных, а при 14 предикторах — уже 16384 (включая модель со всеми предикторами и модель, содержащую только свободный член). Даже с помощью ЭВМ построить и проверить такое большое число уравнений трудно. Один из возможных путей решения этой проблемы — использование методов поиска лучшего подмножества без проверки всех возможных уравнений регрессий.

Предложены и широко используются различные подходы, позволяющие находить наиболее информативные подмножества переменных. Их можно разделить на две группы. Более совершенным является метод псевдоперебора всех возможных сочетаний предикторов. Другим важным способом являются так называемые пошаговые процедуры, они основаны на иных принципах и более широко применяются в практике регрессионного анализа. Существует три их варианта: метод исключения, метод включения, комбинированный метод включения-исключения.

10.2.1. Метод исключения

В этом методе анализ начинается с включения в регрессионную модель всех предикторов. Затем по F или t-статистике, рассчитанной для каждого коэффициента, определяется наименее значимая переменная, которая исключается из уравнения. Снова строится модель по оставшимся переменным. Для новой модели

рассчитывается F или t-статистика и определяется наименее значимая переменная, предназначенная для удаления, и т. д. Подход поочередного исключения только одной переменной называется шаговой процедурой назад, или методом шагового исключения. Важно подчеркнуть, что удаленные переменные из дальнейшего анализа исключаются и больше не могут войти в модель. Такой подход не позволяет найти оптимальное подмножество, составленное из удаленных переменных, если такое существует, что бывает в случаях наличия между переменными сильной взаимной корреляции. Заметим, что использование F или t-статистики эквивалентно удалению переменной, минимально уменьшающей R^2 уравнения или имеющей наименьший частный коэффициент корреляции с зависимой переменной. Процесс может продолжаться до тех пор, пока позволяют заданные правила остановки:

1) при достижении определенного числа переменных в модели;

2) при значимости коэффициентов регрессии выше заданного критерия, например, F.

10.2.2. Метод включения

В отличие от метода исключения, метод включения начинается с построения уравнения, содержащего один лучший предиктор, в которое затем по одному добавляются другие переменные. Это шаговая процедура вперед. Первый лучший предиктор отбирается по максимальному (абсолютному значению) парному коэффициенту корреляции с зависимой переменной, это первый шаг. Затем добавляется переменная, удовлетворяющая одному из следующих критериев, используемых также в методе исключения: по сравнению с другими переменными она дает максимум увеличения R^2 уравнения и имеет наибольший частный коэффициент корреляции, t или F-статистику. Процесс включения продолжается до тех пор, пока включаемые переменные удовлетворят заданным критериям остановки:

1) при достижении определенного числа переменных в модели;

2) при F-статистике для всех переменных, еще не вошедших в уравнение, меньше заданного числа;

3) при сильной коррелированности систем независимых переменных, вызванной включением очередной переменной.

Обычно на практике задают какую-либо комбинацию из этих критериев.

10.2.3. Метод включения—исключения

Эта более сложная процедура является комбинацией методов включения и исключения. На каждом шаге рассматриваются следующие возможности: добавить переменную, исключить перемен-

ную, одну переменную заменить другой, завершить процесс. Принимая решение, руководствуются следующими правилами:

1) включенные в модель переменные исключаются, если они имеют F или t-статистику меньше заданного уровня;

2) переменная, включенная в модель, заменяется на другую, не включенную, если увеличивается R^2 модели;

3) переменная с максимальным значением F, включается в модель, если F превышает заданный уровень и коррелированность системы независимых переменных не становится слишком большой.

Степень допустимой коррелированности независимых переменных, вошедших в регрессионное уравнение, называется толерантностью (Т) и выражается формулой

$$T = 1 - R^2,$$

где R^2 — множественный коэффициент детерминации вводимой независимой переменной с переменными, уже входящими в модель. Если значения Т меньше заданного уровня (часто задают уровень $T = 0{,}01$), то процесс заканчивают, чтобы избежать отрицательных последствий коррелированности предиктов. Выбор Т зависит также от точности расчетов ЭВМ.

Выбор уровня F для включения ($F_{вкл}$) и исключения ($F_{искл}$) предикторов в значительной степени определяется применяемой шаговой процедурой и целями исследования. Если положить $F_{вкл}$ очень малым, скажем, 0,1, тогда почти все переменные будут включены в модель. Наиболее часто предпочитают брать $F_{вкл}$ равным 4,0, что приблизительно соответствует 5 %-ному уровню значимости F-распределения. При комбинированном методе обычно полагают $F_{вкл} = 4$, $F_{искл} = 3{,}9$.

Преимуществом шаговых методов является простота алгоритмов, большая скорость расчетов на ЭВМ (несколько секунд), возможность построения уравнения из очень большого числа (порядка сотен) потенциальных предикторов; слабостью — раздельный анализ переменных. Весьма вероятно, что переменная, которая кажется незначимой на одном шаге, становится значимой на другом или по отдельности переменные не являются значимыми, а при их совместном использовании они намного улучшат регрессионное уравнение. Полученные этими методами результаты также в значительной степени зависят от уровня заданных критериев включения, исключения и толерантности. В результате работы шаговых программ выбирается только одно уравнение, хотя иногда требуется проанализировать несколько моделей с лучшими свойствами.

10.2.4. Метод псевдоперебора

К методам псевдоперебора при выборе наилучшего подмножества предикторов относится популярный алгоритм, предложенный в [42]. Он предполагает поиск лучшего уравнения без по-

строения всех возможных регрессионных моделей и при этом учитывает структуру и взаимозависимость переменных. По существу он опирается на метод исключения Гаусса применительно к корреляционной матрице в соответствии с определенной последовательностью шагов. Математическая основа такой процедуры достаточно сложна, и поэтому здесь не будет рассматриваться.

Программы, использующие подобные алгоритмы, выдают группу лучших сочетаний предикторов для одной, двух, трех и т. д. переменных по заранее определенному критерию качества — $\overline{R}^2$ или C_p. Для нескольких сочетаний предикторов, превосходящих по используемому критерию все остальные, рассчитываются уравнения регрессии.

Можно рекомендовать совместное использование двух подходов при поиске наиболее информативных подмножеств предикторов. Вначале, при большом числе потенциальных предикторов, с помощью шаговых методов можно отобрать существенные предикторы, а затем, применяя программы псевдоперебора, выбрать оптимальные сочетания предикторов и построить регрессионные уравнения. Окончательный выбор подмножества переменных для регрессионной модели не может осуществляться только с помощью формализованных процедур. Статистические и агрометеорологические аспекты должны рассматриваться одновременно. Все известные методы поиска лучшего набора предикторов могут быть лишь вспомогательным аппаратом, облегчающим анализ фактических данных.

Формально существующий произвол в выборе лучшего подмножества предикторов должен быть, однако, ограничен ясными представлениями о роли факторов в моделируемом процессе или явлении.

Ниже приведем фрагмент распечатки программы поиска наилучшего подмножества предикторов для прогноза урожайности всех зерновых культур в Западном Казахстане с использованием критерия Мэллоуса (C_p).

Обозначение предикторов:

24 — средняя температура воздуха за третью декаду мая
26 — средняя температура воздуха за вторую декаду июня
41 — показатель влагообеспеченности третьей декады мая
44 — показатель влагообеспеченности третьей декады июня
54 — средняя декадная сумма осадков за апрель
55 — средняя декадная сумма осадков за май
70 — показатель влагообеспеченности июня
78 — сумма осадков за холодный период

Уравнения с одной переменной

R^2	$\overline{R}^2$	C_p	Переменная
0,478	0,463	23,82	70
0,457	0,442	26,11	78

0,363	0,345	36,34	26
0,244	0,223	49,30	44
0,215	0,192	52,54	54
0,105	0,080	64,44	55
0,082	0,056	66,99	41
0,075	0,049	67,70	24

Уравнения с двумя переменными

R^2	$\overline{R}^2$	C_p	Переменные	
0,637	0,616	8,50	70	78
0,608	0,585	11,66	78	54
0,596	0,572	13,01	78	41
0,540	0,513	19,08	70	55
0,523	0,495	20,96	70	54
0,498	0,469	23,63	70	26
0,492	0,462	24,28	70	41
0,488	0,458	24,78	70	24
0,481	0,451	25,45	70	44

Уравнения с тремя переменными

R^2	$\overline{R}^2$	C_p	Переменные		
0,697	0,670	3,96	70	78	54
0,694	0,667	4,26	70	78	41
0,690	0,662	4,74	78	54	26
0,679	0,650	5,96	70	78	55
0,671	0,641	6,77	78	26	41
0,671	0,641	6,78	70	78	24
0,650	0,618	9,10	78	54	24
0,649	0,617	9,22	78	54	41
0,646	0,614	9,54	70	78	44
0,640	0,607	10,20	70	78	26

Уравнения с четырьмя переменными

R^2	$\overline{R}^2$	C_p	Переменные			
0,726	0,692	2,84	70	78	54	55

Одно из пяти лучших уравнений

Переменная	Коэффициент регрессии	t-статистика
70	0,165	3,59
78	0,132	4,24
54	0,582	2,34
55	−0,534	−1,83
Свободный член	−0,172	

R^2	$\overline{R}^2$	C_p	Переменные			
0,721	0,686	3,33	70	78	54	41

Одно из пяти лучших уравнений

Переменная	Коэффициент регрессии	t-статистика
78	0,154	4,79
54	0,638	2,40
26	—0,979	—2,89
41	0,605	1,90
Свободный член	—0,525	

R^2	$\overline{R}^2$	C_p	Переменные			
0,719	0,684	3,60	70	78	54	41

Одно из пяти лучших уравнений

Переменная	Коэффициент регрессии	t-статистика
70	0,962	2,82
78	0,153	4,72
54	0,460	1,67
41	0,510	1,57
Свободный член	—0,243	

Уравнения с пятью переменными

R^2	$\overline{R}^2$	C_p	Переменная				
0,736	0,694	3,68	70	78	54	41	55

Одно из пяти лучших уравнений

Переменная	Коэффициент регрессии	t-статистика
70	0,147	3,02
78	0,143	4,39
54	0,456	1,68
41	0,372	1,12
55	—0,437	—1,44
Свободный член	—0,188	

R	$\overline{R}^2$	C_p	Переменеые				
0,734	0,692	3,90	70	78	54	24	55

Одно из пяти лучших уравнений

Переменная	Коэффициент регрессии	t-статистика
70	0,147	2,97
78	0,140	4,36
54	0,539	2,14
24	—0,411	—1,00
55	—0,447	—1,47

Свободный член —0,115

R^2	$\overline{R}^2$	C_p	Переменные				
0,734	0,691	3,95	70	78	54	26	41
0,733	0,691	3,99	70	78	54	41	44
0,728	0,684	4,59	70	78	54	44	55
0,726	0,682	4,79	70	78	54	24	26
0,726	0,682	4,80	78	54	24	26	41
0,726	0,682	4,83	70	78	54	26	55
0,719	0,674	5,57	70	78	24	26	41
0,719	0,673	5,60	70	78	54	24	41

Уравнения с шестью переменными

R^2	$\overline{R}^2$	C_p	Переменные					
0,742	0,690	5,08	70	78	54	41	44	55
0,738	0,686	5,49	70	78	54	26	41	55
0,737	0,684	5,66	70	78	54	24	26	41
0,736	0,684	5,68	70	78	54	24	41	55
0,735	0,682	5,87	70	78	54	24	26	55
0,733	0,680	6,02	70	78	54	24	26	44
0,731	0,677	6,31	78	54	24	26	41	55
0,722	0,666	7,26	70	78	24	26	41	55
0,714	0,659	8,15	78	54	24	26	44	55
0,710	0,652	8,52	78	24	26	41	44	55

Уравнения с семью переменными

R^2	$\overline{R}^2$	C_p	Переменные						
0,742	0,680	7,02	70	78	54	26	41	44	55
0,742	0,680	7,08	70	78	54	24	41	44	55
0,740	0,678	7,26	70	78	54	24	26	41	44
0,740	0,677	7,31	70	78	54	24	26	44	55
0,740	0,677	7,31	70	78	54	24	26	44	55
0,739	0,676	7,38	70	78	54	24	26	41	55
0,731	0,666	8,30	78	54	24	26	41	44	55
0,725	0,659	8,93	70	78	24	26	41	44	55
0,578	0,477	24,91	70	54	24	26	41	44	55

Уравнения с восемью переменными

R^2	$\overline{R}^2$	C_p	Переменные							
0,743	0,669	9,00	70	78	54	24	26	41	44	55

Статистики для лучшего уравнения

Статистика Мэллоуса C_p	2,84
Коэффициент детерминации R^2	0,726
Коэффициент корреляции R	0,852
СКО/$(n-p)$	0,315
Стандартная ошибка уравнения	0,566
F-статистика	21,22

Переменная	Коэффициент регрессии	Стандартная ошибка	Стандартизованный коэффициент	t-статистика	Толерантность
Свободный член	—0,172	0,060	—1,708	2,88	
70	0,165	0,460	0,567	3,59	0,342
78	0,132	0,312	0,449	4,24	0,762
54	0,582	0,248	0,239	2,34	0,823
55	—0,534	0,292	—0,248	—1,83	0,464

10.3. ИСПОЛЬЗОВАНИЕ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ В КЛАСТЕРНОМ АНАЛИЗЕ

Наряду с такими важными проблемами, как поиск количественных закономерностей влияния метеорологических условий на рост и развитие растений, корреляционный анализ можно с успехом применять для решения фундаментальных задач объективной классификации природных объектов, или кластеризации. Коэффициенты корреляции вполне пригодны для того, чтобы быть мерой расстояния между изучаемыми отношениями величин или явлений. Необходимость в простом инструменте исследований назрела уже давно.

При моделировании влияния погоды на урожайность сельскохозяйственных культур исследователь сталкивается со значительным разнообразием почвенно-климатических условий, неоднородностью структуры посевных площадей. Поэтому при наличии коротких временных рядов урожайности используют широко распространенный метод «годостанций», т. е. объединяют в одну статистическую совокупность несколько однородных по каким-либо признакам подвыборок [31]. При этом появляется возможность расширить исследуемую совокупность для устранения воздействия случайных колебаний. Объединенные данные рассматриваются как равноправные и обрабатываются как один исходный массив данных, несмотря на то, что отдельные данные наблюдений не являются статистически независимыми и при статистических оценках число степеней свободы будет меньше чем $(n-p-1)$ (n — длина объединенной выборки, p — число параметров в линейном регрессионном уравнении); но для простоты расчета часто принимают такую традиционную в агрометеорологии гипотезу. При этом полезно рассчитать количество фактически независимой информации в выборке [30].

Природно-климатические условия отдельной области характеризуются большим числом взаимозависимых факторов. Выделение ведущих из них для целей классификации по тому или иному признаку — сложная задача, особенно, если необходимо принимать во внимание и некоторые экономические показатели, как например, уровень земледелия, достигнутый в области, темпы его роста. В данном случае интересно разделение областей на группы, в каждой из которых влияние метеорологических факторов

осуществляется при сходных почвенно-климатических и экономических условиях. Географически обусловленную неоднородность формирования урожая зерновых и различие экономических условий областей можно, на наш взгляд, выявить, изучая степень пространственной связанности каких-либо временных рядов показателей условий погоды или даже урожайности. Для этого можно использовать простой метод объективной классификации, опирающийся на корреляционную матрицу. Это так называемая иерархическая процедура классификации Кинга [12]. Классификация начинается с групп, содержащих один объект, в нашем случае — области. Каждое последующее объединение уменьшает число групп на единицу. Процесс продолжается до момента, пока не образуется одна группа, объединяющая все объекты. Расстоянием между группами (мерой их связи) является средний арифметический коэффициент корреляции (Q) всех возможных пар временных рядов урожайности вошедших в группы областей

$$Q = \sum_{i=1,\ j=1}^{n_1,\ n_2} r_{ij}/(n_1 \cdot n_2),$$

где r_{ij} — коэффициент корреляции i-й и j-й областей разных групп, n_1, n_2 — количество областей в группах.

Оказалось, что принятая в качестве обобщенного показателя как почвенно-климатических, так и хозяйственно-экономических условий области урожайность всех зерновых культур позволила получить содержательное разбиение групп областей по степени их связанности. Достаточно примечателен тот факт, что не нарушен принцип географической сопряженности, т. е. выделяемые по степени близости нашего показателя области всегда имеют общую границу, что говорит о хорошем отражении климатических условий областей.

В табл. 10.2 представлены парные коэффициенты корреляции между временны́ми рядами урожайности зерновых по областям Казахстана. Особо отметим факт отрицательной корреляции урожайности зерновых между самой западной областью — Уральской — и восточными областями — Восточно-Казахстанской, Семипалатинской и Карагандинской, что ярко свидетельствует о большом различии в погодных условиях на огромной территории Казахстана и необходимости выделения однородных по природным и экономическим условиям групп областей.

На рис. 10.2 приведена дендрограмма для областей Казахской ССР, на которой показана последовательность объединения областей в однородные группы. На основе анализа географических условий и результатов, полученных с помощью объективной классификации, хорошо согласующихся с общими агрометеорологическими представлениями о возможных однородных зонах формирования урожая, были объединены в группы следующие области: в Западном Казахстане — Актюбинская и Уральская; в Восточном Казахстане — Восточно-Казахстанская и Семипалатинская;

Таблица 10.2

Корреляционная матрица рядов урожайности всех зерновых культур по областям Казахстана ($\times 10^{-2}$)

Область	Номер области														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Актюбинская	100														
2. Алма-Атинская	26	100													
3. Восточно-Казахстанская	13	50	100												
4. Карагандинская	19	54	69	100											
5. Кокчетавская	37	59	67	72	100										
6. Кустанайская	50	47	58	59	74	100									
7. Павлодарская	34	41	66	82	75	55	100								
8. Северо-Казахстанская	46	54	44	53	82	80	51	100							
9. Семипалатинская	21	44	91	79	56	53	70	34	100						
10. Уральская	76	26	—27	—26	16	9	3	22	—18	100					
11. Целиноградская	34	67	59	92	72	54	83	59	69	18	100				
12. Джамбулская	55	80	27	56	44	56	44	41	39	49	66	100			
13. Чимкентская	63	70	26	56	62	63	59	66	27	60	66	83	100		
14. Талды-Курганская	26	85	68	70	63	59	64	52	68	9	73	75	71	100	
15. Тургайская	27	48	55	77	81	64	68	63	63	2	81	47	41	55	100

в Северном Казахстане — Северо-Казахстанская, Кокчетавская и Кустанайская; в Центральном Казахстане — Целиноградская, Карагандинская, Павлодарская и Тургайская; в Южном Казах-

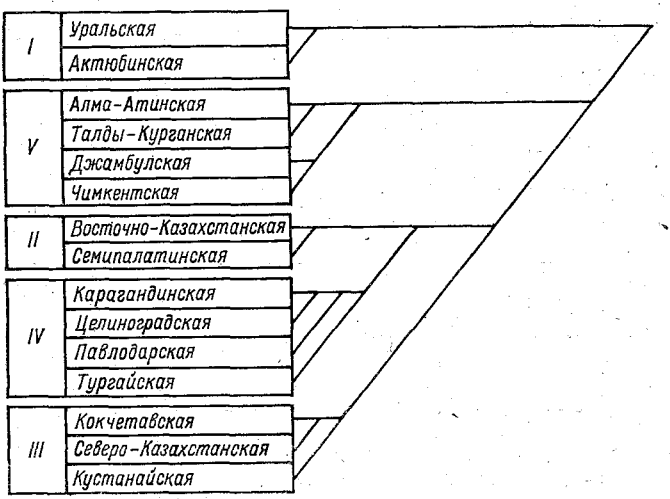


Рис. 10.2. Дендрограмма областей Казахстана по признаку сходства колебаний временных рядов урожайности.

I — Западный Казахстан, II — Восточный Казахстан, III — Северный Казахстан, IV — Центральный Казахстан, V — Южный Казахстан.

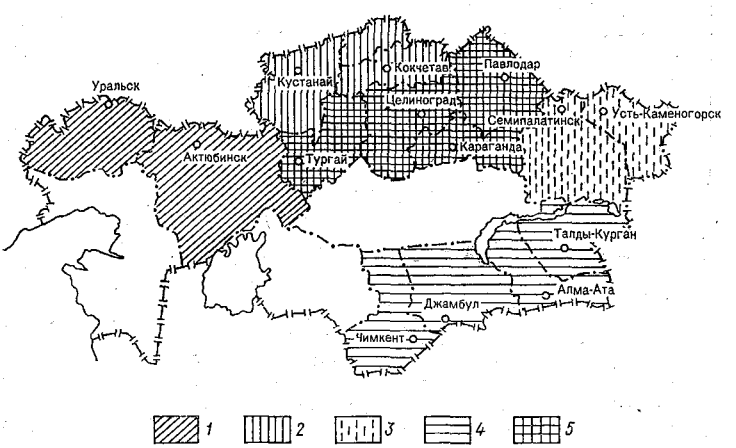


Рис. 10.3. Районирование Казахстана по степени однородности колебания временных рядов урожайности зерновых культур.

Районы: *1* — западный, *2* — северный, *3* — восточный, *4* — южный, *5* — центральный; *6* — границы областей.

стане — Алма-Атинская, Чимкентская, Джамбулская и Талды-Курганская.

На рис. 10.3 показано географическое районирование Казахстана по степени однородности колебания временных рядов уро-

жайности всех зерновых культур. Следует отметить, что полученные однородные группы областей не совпадают с группами, выделяемыми другими исследователями. Так, если в методических пособиях по составлению прогноза средней областной урожайности самых распространенных в республике зерновых культур — яровой пшеницы [11] и ячменя [6] — так же, как и у нас, одна из выделяемых групп состояла из Актюбинской и Уральской областей, то остальные из рассматриваемых там областей объединялись только в одну группу.

Представляется целесообразным при построении регрессионных прогностических уравнений использовать совокупность подвыборок с одинаковой дисперсией. Для этого отклонения урожайности от трендов по каждой области надо нормировать и привести к единичной дисперсии, тогда прогноз урожайности сельскохозяйственных культур будет выражаться в долях среднего квадратического отклонения от линии тренда. Будет учтена неоднородность условий, обусловливающая различие значения дисперсии в областях, входящих в одну группу.

Недостаточный объем имеющейся в распоряжении исследователя информации характерен для агрометеорологии, поэтому всегда остро стояла и еще долго будет актуальной проблема объединения различных статистических совокупностей в одну, более обширную, характеризующуюся какими-либо достаточно однородными данными. Пока в агрометеорологии такая проблема решается без привлечения объективных методов классификации [6]. Использованный подход иерархической классификации областей Казахстана по степени однородности колебаний обобщенного показателя агрометеорологических условий вегетационного периода — урожайности зерновых культур — привлекает тем, что позволяет ограничить степень объединения классифицируемых объектов на основе неформального анализа существа проблемы и стоящих целей. Разработанная классификация областей Казахстана для задач построения прогностических регрессионных уравнений несомненно способствует более точному оцениванию параметров моделей.

Глава 11. АЛЬТЕРНАТИВНЫЕ РЕГРЕССИОННЫЕ МОДЕЛИ

Возрастающий поток агрометеорологической информации, усложнение стоящих перед прогнозистами задач, интенсификация научных исследований требуют широкого использования вычислительной техники. Возникает необходимость в новых способах обработки статистической информации и методах построения прогностических моделей. Не всегда при анализе данных удается «прочувствовать» каждую многомерную точку пространства предикторов, разобраться в сложном комплексе влияющих факторов [17, 28]. Нередки случаи, когда при множественном коэффициенте

корреляции, равном 0,99, регрессионная модель не пригодна для прогнозирования на «свежих» данных и отбраковывается [16].

Недостаточная устойчивость метода наименьших квадратов к изменениям входной информации, повышенная чувствительность статистических процедур связаны с особенностями используемых данных. Например, сильно коррелированные переменные в линейном регрессионном анализе могут затруднить получение коэффициентов модели с заданной точностью. Еще одним источником возможных вычислительных трудностей является наличие переменных с малым коэффициентом вариации. Это относится и к так называемым «выбросам» (аномальным данным). По этим причинам необходимо изменять методы анализа, исходя из свойств реальных данных.

Невыполнение гипотез, лежащих в основе классического регрессионного анализа, на практике вызвало к жизни появление альтернативных методов регрессионного анализа. Последние двадцать лет ознаменовались значительными успехами в этом направлении. Созданы регрессионные методы, робастные (устойчивые) к отклонению ошибок регрессионного уравнения от нормального закона и к возможному присутствию аномальных данных, а также — методы гребневой регрессии для случаев сильной коррелированности предикторов и ряд других.

11.1. МОДЕЛЬ ГРЕБНЕВОЙ РЕГРЕССИИ

Метод гребневой регрессии был предложен в 1970 г., в последнее время его возможности интенсивно изучались [45, 49]. Главной целью метода гребневой регрессии (иногда его называют ридж-регрессией) является преодоление малой устойчивости оценок коэффициентов регрессионной модели, получаемых обычным методом наименьших квадратов, когда предикторы линейно взаимозависимы (коррелированы).

В отличие от метода наименьших квадратов, дающих несмещенные оценки коэффициентов уравнения, в методе гребневой регрессии оценки смещенные, но при этом они имеют меньшую дисперсию. Поэтому такие оценки могут давать более точные и приемлемые для практического использования результаты [8].

Зависимость дисперсии оценки от смещения иллюстрируется на рис. 11.1. Напомним, что коэффициенты уравнений — случайные числа и подчиняются некоторому распределению, совпадающему по форме с распределением зависимой переменной.

Проблема выбора смещенного или несмещенного оценивания непроста. В том случае, когда нежелательно получить большую ошибку коэффициентов уравнения, предпочтение обычно отдается смещенным оценкам.

Ошибка коэффициента слагается из двух составляющих: смещения коэффициента и его дисперсии. Из рис. 11.1 видно, что смещенное оценивание может быть приемлемо, если незначительным смещением оценки можно достичь большого уменьшения

дисперсии коэффициента. В этом состоит главный смысл использования метода гребневой регрессии. Метод гребневой регрессии полезен в ситуации, когда из-за сильной коррелированности предикторов значительно увеличивается дисперсия оценок регрессионных коэффициентов (см. гл. 9).

В методе гребневой регрессии «платой» за уменьшение дисперсии является смещение оценок коэффициентов. Если коррели-

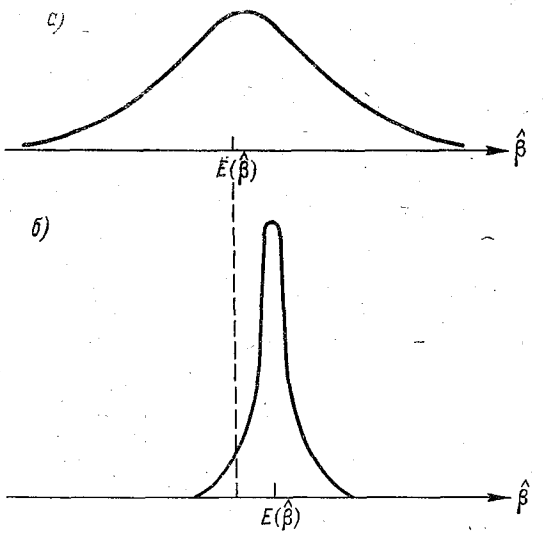


Рис. 11.1. Дисперсия и смещение оценок.

а — плотность распределения коэффициента $\hat{\beta}$ при несмещенной оценке; *б* — то же при смещенной оценке.

рованность предикторов сильная, то в большинстве случаев общая ошибка оценки коэффициентов при использовании этого метода меньше, чем при использовании традиционного метода наименьших квадратов.

В методе наименьших квадратов оценку коэффициентов можно получить по формуле

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Для простоты будем считать, что все переменные стандартизованы, т. е. $\mathbf{X}'\mathbf{X}$ — корреляционная матрица, $\hat{\beta}$ — стандартизованный коэффициент.

Оценка коэффициентов методом гребневой регрессии представляется формулой

$$\hat{\beta}^* = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y},$$

где $k\mathbf{I}$ — произведение малого положительного скаляра (числа k) и единичной матрицы. Это означает, что малая положительная константа добавлена к каждому элементу на диагонали корреляционной матрицы. Эта процедура улучшает обусловленность матрицы $\mathbf{X}'\mathbf{X}$ и делает оценки коэффициентов более устойчивыми.

При практическом использовании метода гребневой регрессии одним из основных вопросов, который необходимо решать, является выбор параметра k. Существует несколько численных методов расчета параметра, однако трудно отдать предпочтение какому-либо из них.

Наиболее часто на практике используют простой эмпирический подход, называемый методом гребневого следа. Гребневый след — это график зависимости коэффициента регрессии от параметра k. Этот метод позволяет при анализе множества предикторов выявлять те из них, которые наиболее чувствительны к изменению начальных данных, теснее связаны между собой линейной зависимостью. На графике выбирают такой параметр k, при котором коэффициенты «стабилизируются» и при дальнейшем увеличении параметра изменяются мало. Значение принятого параметра k является мерой смещения оценок от истинного значения, поэтому стараются не придавать k очень больших значений. По мере увеличения параметра k абсолютное значение коэффициентов уменьшается и стремится к нулю. Обычно k выбирают меньше 0,5. Нередко при изменении k коэффициенты уравнения меняют знаки на физически более обоснованные. Это тоже может служить ориентиром для выбора значения параметра k. Гребневой след можно использовать и для процедуры отбора предикторов в модель.

Необходимо заметить, что метод гребневой регрессии не всегда лучше метода обычной регрессии. Использование первого без понимания его возможностей и ограничений может приводить к отрицательным результатам. Очень важно иметь определенное представление о значениях коэффициентов уравнения, чувствовать их физический смысл.

На практике можно использовать и метод наименьших квадратов, если прибавлять к данным основного массива набор фиктивных значений, которые обеспечивают добавку к диагональным элементам корреляционной матрицы $\mathbf{X'X}$. Решают систему уравнений, где к матрице наблюдений за независимыми переменными добавляют квадратную матрицу $\mathbf{I}\sqrt{k}$ размера m, а к данным зависимой переменной — соответствующее количество нулей:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \\ \sqrt{k} & 0 & \dots & 0 \\ 0 & \sqrt{k} & \dots & 0 \\ 0 & 0 & \sqrt{k} & \dots \\ 0 & 0 & 0 & \dots \sqrt{k} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \\ 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}.$$

Придавая параметру k различные значения и заново решая задачу, получим зависимость регрессионных коэффициентов от параметра, т. е. гребневый след. Как правило, шаг по k выбирают небольшим, например, 0,02. Имея в распоряжении программу расчета уравнения множественной регрессии и обладая минимальными навыками программирования, можно без труда получить оценки коэффициентов гребневой регрессии. При $k=0$ имеем коэффициенты классического метода наименьших квадратов.

Пример. Строилась прогностическая модель средней областной урожайности всех зерновых культур в Северо-Казахстанской области. В качестве предикторов использовались восемь переменных: сумма осадков и температура воздуха за третью декаду мая и за три декады июня (соответственно R_1, R_2, R_3, R_4 и t_1, t_2, t_3, t_4). В соответствии с рекомендациями все переменные были стандартизованы и приведены к единичной дисперсии и нулевому среднему значению (см. гл. 9). Анализ корреляционной матрицы переменных свидетельствует о тесной линейной зависимости предикторов (табл. 11.1); коэффициент обусловленности корреляционной матрицы предикторов: $p=35{,}9$. Напомним, что при независимости переменных $p=1$.

Таблица 11.1

Корреляционная матрица системы переменных ($\times 10^{-3}$)

Переменная	R_1	R_2	R_3	R_4	t_1	t_2	t_3	t_4	y
R_1	1000								
R_2	—6	1000							
R_3	290	71	1000						
R_4	96	92	327	1000					
t_1	—19	73	—676	—589	1000				
t_2	274	—162	686	286	—351	1000			
t_3	98	—405	—617	23	431	—222	1000		
t_4	—128	—76	—417	—526	534	—435	260	1000	
y	139	217	549	586	—509	107	—313	—563	1000

На рис. 11.2 представлены гребневые следы — графики зависимости коэффициентов регрессии $\hat{\beta}_i^*$ от параметра k, построенные для значений k из интервала [0; 0,5]. Графики позволяют выявить чувствительность коэффициентов к изменению исходного набора данных, составить представление о степени обусловленности матрицы $\mathbf{X'X}$. Из общего вида графиков явствует, что коэффициенты уравнения регрессии, полученные методом наименьших квадратов, неустойчивы. Уже при небольшом добавлении к элементам диагонали числа k получились коэффициенты, сильно отличающиеся от первоначальных, при $k=0$. Значительная коррелирован-

ность предикторов привела также к несоответствию знаков коэффициентов их физическому смыслу.

Многочисленные исследования условий формирования урожая зерновых культур в Северо-Казахстанской области указывают на то, что в условиях недостаточного увлажнения сезона вегетации коэффициенты при осадках в рассматриваемый период должны

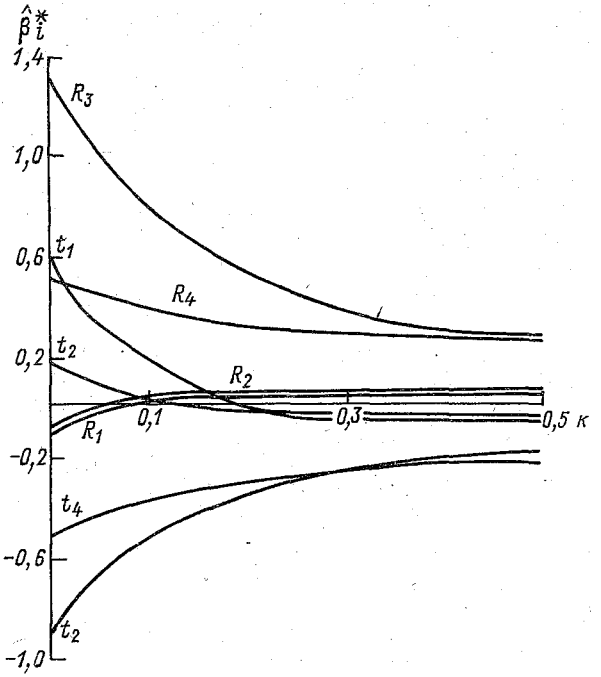


Рис. 11.2. Зависимость коэффициентов регрессионной модели от гребневого параметра k.

быть неотрицательными, а при температуре — неположительными. Однако полученная классическим методом наименьших квадратов регрессионная модель имеет вид

$$y = -0{,}105R_1 - 0{,}091R_2 + 1{,}294R_3 + 0{,}500R_4 + 0{,}559t_1 - \\ - 0{,}904t_2 + 0{,}138t_3 - 0{,}508t_4. \qquad (11.1)$$

Как видно из рис. 11,2, при значениях параметра k, больших некоторого положительного числа, коэффициенты при R_1 и R_2 меняют знак и становятся положительными.

При $k=0$ коэффициент при t_1 оказался вторым по значению положительным коэффициентом, что, как уже говорилось, не соответствует представлению о влиянии температуры воздуха на формирование урожая в этот период. С ростом параметра k коэффициент при t_1 быстро уменьшается и меняет знак на противо-

положный; изменяет знак и коэффициент при t_3. Сильнее всего проявляется влияние смещения на коэффициенте при R_3, что свидетельствует о его большой неустойчивости и тесной связи с другими переменными. Это заключение находит подтверждение в значениях коэффициентов парной корреляции, приведенных в табл. 11.1.

При $k > 0{,}5$ все коэффициенты практически постоянны. Уменьшение изменчивости коэффициентов начинает проявляться в интервале k [0,2; 0,4], поэтому стабилизирующим систему значением параметра с определенной долей субъективности выбрали $k = 0{,}3$. Этому значению k соответствует уравнение гребневой регрессии

$$y = 0{,}041R_1 + 0{,}070R_2 + 0{,}038R_3 + 0{,}288R_4 - 0{,}036t_1 - 0{,}249t_2 - \\ - 0{,}044t_3 - 0{,}249t_4 - 0{,}015. \qquad (11.2)$$

Свободный член в уравнении появляется из-за смещения среднего арифметического используемых переменных относительно нуля при применении вышеописанного метода расчета.

В отличие от уравнения (11.1), полученного классическим методом, с коэффициентами, не поддающимися непосредственной интепретации, в уравнении (11.2) коэффициенты соответствуют физическим представлениям о воздействии R_i и t_i на урожай; что также подтверждается данными табл. 11.1.

Одной из главных функций регрессионных моделей является прогнозирование на свежих данных, которые не были использованы при построении моделей. Применение метода гребневой регрессии и других альтернативных методов ставит основной целью улучшение прогностических возможностей моделей.

Для сравнения таких возможностей в уравнениях, полученных классическим методом наименьших квадратов и методом гребневой регрессиии, исходная выборка из 23 лет была разбита на два временных ряда. Первый ряд, по которому строились модели, включал 19 лет, второй — для проверки прогностических способностей моделей на независимом материале — 4 года. В качестве критерия успешности прогнозирования использовался средний квадрат ошибки прогноза. Для уравнения (11.1) он равен 2,21, для (11.2) — 0,96, что свидетельствует о преимуществах прогнозирования на независимых данных методом гребневой регрессии. Коэффициент корреляции в уравнении (11.1) достаточно высок: 0,911. Отношение среднего квадрата ошибки прогноза на независимых данных к дисперсии зависимой переменной для уравнения (11.1) равно 1,15, для гребневого уравнения (11.2) — 0,50.

Приведенный пример достаточно типичен. Использование метода гребневой регрессии при определенных условиях позволяет расширить рамки применения традиционного метода регрессионного анализа, повысить точность агрометеорологических прогнозов.

11.2. РОБАСТНАЯ РЕГРЕССИЯ

При статистической обработке данных большое значение имеют те гипотезы и допущения, которые кладутся в основу математических моделей. На практике эти гипотезы, как правило, не выполняются точно. В большинстве случаев полагают, что малое несоответствие выдвинутым предположениям приводит к незначительным погрешностям в построенной модели. Однако это не всегда соответствует действительности. В последние годы выяснилось, что многие обычно применяемые процедуры весьма чувствительны к небольшим отклонениям от положенных в их основу предположений. В связи с этим была поставлена задача создания робастных методов. Широко используемый в настоящее время термин «робастный», или, что означает, устойчивый, стал весьма распространенным в последние двадцать лет, хотя его стали применять еще в пятидесятые годы [37]. Под робастностью в широком смысле принято понимать малую чувствительность модели к небольшим отклонениям от принятых допущений. Наиболее полно рассмотрена робастность при отклонении распределения переменной от заданного закона, обычно — от нормального, или гауссовского [35, 47]. Широкое использование нормального распределения в регрессионном анализе объясняется, с одной стороны, эффективностью оценок, полученных методом наименьших квадратов, с другой — несложностью расчета. До недавнего времени не было необходимых высокопроизводительных ЭВМ, которые позволяли бы использовать более трудоемкие методы.

Приведем простой пример, показывающий неустойчивость метода наименьших квадратов к выбросам. Известно, что среднее арифметическое выборки $x_1, x_2, \ldots, x_n$ определяется как

$$\bar{x} = \frac{1}{n} \sum_{i}^{n} x_i.$$

Эта величина является эффективной оценкой, имеет наименьшую дисперсию, если величины $x_1, x_2, \ldots, x_n$ распределены по нормальному закону. Пусть необходимо вычислить среднее арифметическое выборки: 1, 0, —0,5, —1, 0,5, 12. Последнее число явно отличается от предыдущих и является выбросом. Среднее арифметическое без учета этого числа равно нулю, с его учетом — 2, что для данной выборки — явная бессмыслица. Если в качестве среднего взять медиану, которая определяется взаимным расположением упорядоченных по значению наблюдений x_i, то результат будет более реалистичным.

Чувствительность классических процедур к «длинным хвостам» распределений также очень сильно отражается на такой важной статистике, как дисперсия выборки. Выражения «устойчивость к распределению» и «устойчивость к выбросам» по смыслу значительно разнятся, хотя имеют одинаковое математическое выражение. Возможен такой подход к построению стати-

стических моделей, при котором сначала освобождают данные от выбросов, а затем применяют классические методы и процедуры к оставшимся данным. Однако на этом пути часто можно столкнуться с трудностями, так как редко удается выделить точно в многомерном регрессионном анализе имеющиеся аномальные наблюдения до тех пор, пока не использованы робастные методы оценки параметров [35]. Даже если начальный набор данных и подчинялся нормальному распределению, но был засорен выбросами, то при их удалении не исключена вероятность ошибок, когда некоторая их доля останется, а не принадлежащие к выбросам данные будут удалены. Таким образом ситуация может быть только ухудшена, и классическая теория, построенная на нормальном распределении, не сможет быть применена достаточно эффективно. Из практики известно, что процедура отбраковки части данных во многом уступает робастным процедурам, которые позволяют плавно варьировать степень влияния таких выбросов — вплоть до полного их удаления.

В связи с использованием робастных методов встают вопросы о том, как широка область применения данной статистической процедуры, устойчива ли она при отклонениях от принятых гипотез, и как ее построить, чтобы она была робастной. Важность этих методов определяется также и тем, что они дополняют классические статистические методы.

Не всегда исследователи достаточно точно знают, насколько хороши или плохи те методы, которые применяют в случае, когда данные не подчиняются строго выдвигаемым гипотезам. Рассмотрим простой и наглядный пример, поясняющий часто встречающиеся трудности в использовании классического регрессионного метода наименьших квадратов [35]. Предположим, что необходимо провести прямую линию через шесть точек; здесь возможны различные варианты (рис. 11.3). Прямая, построенная классическим методом, показана на рис. 11.3 *а*. Помещенная в табл. 11.2, соответствующая рис. 11.3 *а* ошибка этой модели $\varepsilon_i = y_i - \hat{y}_i$ показывает, что нет оснований для беспокойства по поводу адекватности модели, особенно, если сравнивать ε_i со стандартным отклонением. Аномально больших ε_i не наблюдается. Сомнение может вызвать только точка 1, где ε_i достигает максимального значения. Но вполне вероятно, что линейная модель не адекватна данным и необходимо использовать более сложную модель, например, параболу (рис. 11.3 *в*). Можно также просто отбросить точку 6, и тогда получим уравнение, представленное на рис. 11.3 *б*. Совершенно ясно, что имеющиеся данные не могут удовлетворять трем таким графикам одновременно. Из-за малой ошибки уравнения можно отдать предпочтение модели, представленной на рис. 11.3 *в*. В действительности же модельный пример был сгенерирован из пяти точек, лежащих на прямой $y = -2 - x$, к которым была добавлена случайная нормально распределенная ошибка с нулевым средним и стандартным отклонением, равным

0,6. Шестая точка, лежащая далеко от этой прямой, моделировала грубую ошибку или выброс данных. Поэтому модель на рис. 11.3 *б* адекватна и восстанавливает зависимость почти точно. В этом примере использована простая модель с двумя парамет-

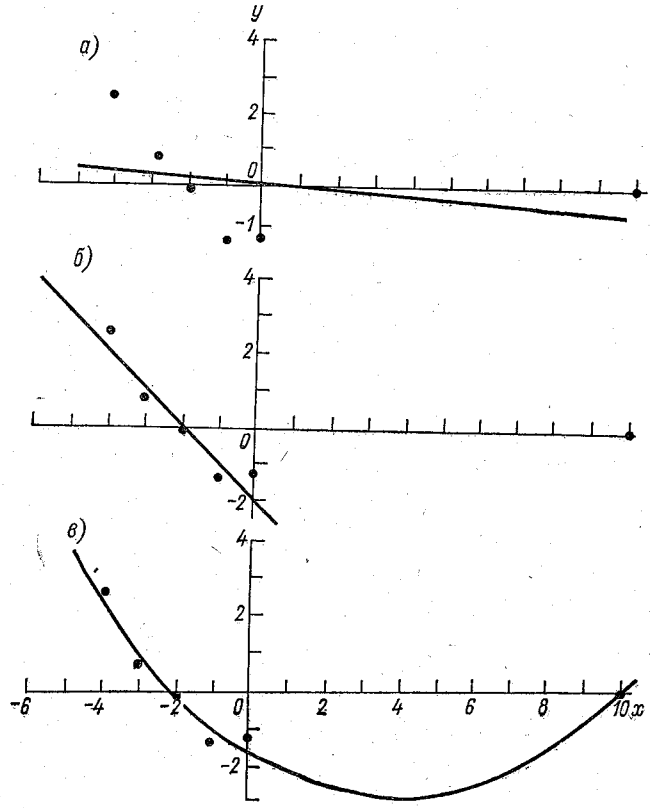


Рис. 11.3. Различные варианты построения регрессионной модели при наличии резко выделяющейся точки выброса.

а) прямая, построенная методом наименьших квадратов, *б*) то же при удалении выброса, *в*) парабола второго порядка.

рами, и рис. 11.3 позволяет без труда определить и понять источник неприятностей в точке 6, даже если соответствующие ошибки в табл. 11.2 для этого мало пригодны.

Задача значительно усложняется в случае многомерности модели. Трудность заключается в том, что выбросы невозможно выявить посредством анализа ошибок модели. Для их обнаружения необходимо найти аналитические методы определения выбросов и сильно влияющих точек, т. е. точек в многомерном пространстве переменных, наиболее существенно воздействующих на вид аппроксимирующей функции. Такие точки могут быть наиболее важными в выборке, и на них должно быть обращено особое

Таблица 11.2

Пример построения регрессионной модели

Номер точки	x	y	Рис. 11.1 *а*		Рис. 11.1 *б*		Рис. 11.1 *в*	
			$\hat{y}$	ε_i	$\hat{y}_i$	ε_i	$\hat{y}_i$	ε_i
1	—4	2,48	0,39	2,09	2,04	0,44	2,33	0,25
2	—3	0,73	0,31	0,42	1,06	—0,33	0,99	—0,26
3	—2	—0,04	0,23	—0,27	0,08	—0,12	—0,09	—0,13
4	—1	—1,44	0,15	—1,59	—0,90	—0,54	—1,00	—0,44
5	0	—1,32	0,07	—1,39	—1,87	0,55	—1,74	0,42
6	10	0	—0,75	0,75	—11,64	(11,64)	0,01	—0,01
$\hat{\sigma}$			1,55		0,55		0,41	
$\varepsilon_{max}/\hat{\sigma}$			1,35		1,00		1,08	

внимание, хотя не всегда их использование при построении регрессионной модели дает положительный результат.

К настоящему времени предложено несколько подходов к использованию идей робастности в регрессионном анализе. Рассмотрим кратко наиболее распространенный метод получения так называемых М-оценок максимального правдоподобия, предложенный Хьюбером [46]. М-оценки — это обобщение метода максимального правдоподобия, поэтому вместо того, чтобы максимизировать логарифмическую функцию правдоподобия

$$\ln L = \sum_{i=1}^{n} \ln f(x_{ij} \mid \beta),$$

минимизируют более общую функцию

$$M = \sum_{i=1}^{n} \rho(x_{ij} \mid \beta). \qquad (11.3)$$

В задачах регрессионного анализа неизвестные параметры $\vec{\beta}$ оцениваются так, чтобы ошибки уравнения ε_i были малы

$$M = \sum_{i=1}^{n} \rho(x_{ij})\,\beta \to \min. \qquad (11.4)$$

Здесь $\rho(x_{ij}|\beta) = \rho(\varepsilon_i)$ и $\varepsilon_i = y_i - \sum_{j=1}^{m} x_{ij}\beta_j$. Выбор функции ρ достаточно произволен, поэтому могут быть получены различные значения β в зависимости от вида ρ. На заданную функцию ρ налагаются определенные условия: ρ должна иметь первую и вторую производные $\Psi(\varepsilon_i)$ и $\Phi(\varepsilon_i)$, которые ограничены почти всюду.

Эти условия будут выполнены, если в качестве $\rho(\varepsilon_i)$ взять выпуклую почти везде непрерывную функцию.

Продифференцировав (11.3), получим систему линейных уравнений

$$\sum_{i=1}^{n} \Psi\left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j\right) x_{ik} = 0, \quad (k = 1, 2, \ldots, m),$$

где $\Psi = \rho'$, и решение которой эквивалентно минимизации (11.4) при выполнении условий, наложенных на ρ.

Обычно при построении робастных оценок вводят параметр масштаба s и решают систему

$$\sum_{i=1}^{n} \Psi\left(\frac{y_i - \sum_{j=1}^{m} x_{ij}\beta_j}{sk^*}\right) x_{ij} = 0 \quad (j = 1, 2, \ldots, m).$$

Как правило, параметр масштаба неизвестен, и в качестве него (оценки дисперсии ошибки уравнения) рекомендуется выбирать медиану ошибок уравнения, полученную предварительно, например, классическим методом. Константа k^* выбирается, исходя из соображений неформального характера.

Изучая «засоренные» нормально распределенные переменные, Хьюбер предложил семейство оценок, определяемых функцией ρ:

$$\rho(\varepsilon) = \begin{cases} \frac{1}{2}\varepsilon^2 & \text{при } |\varepsilon| < k^*s = h, \\ k^*s\left(|\varepsilon| - \frac{1}{2}k^*s\right) & \text{при } |\varepsilon| \geqslant k^*s = h \end{cases}$$

(h — параметр робастности), и показал некоторые оптимальные свойства оценок [48] при известном параметре s.

Выбор функций ρ, ψ и k^* — важная проблема, занимавшая многих статистиков. Можно упомянуть используемые оценки Андрюса, Рамсея, Тьюки. Рей провел практическое сравнение некоторых видов таких функций и отдал предпочтение оценкам Хьюбера [35].

Для того чтобы оценки Хьюбера были робастными, необходимо, чтобы при больших ошибках ε_i скорость роста функции $\rho(\varepsilon_i)$ была меньше скорости роста принятой в методе наименьших квадратов функции ε_i^2.

График функции Хьбера приведен на рис. 11.4. Это парабола на отрезке $[-h, h]$, продолженная далее двумя прямыми линиями. Значение h, определяющее порог, после которого происходит уменьшение скорости роста ρ, является по существу параметром, определяющим степень робастности оценок метода.

Если значение параметра h достаточно велико, то полученные регрессионные оценки совпадают с обычными оценками метода наименьших квадратов. Обычно значения h подбирают, исходя из конкретных свойств исследуемых статистических совокуп-

ностей, из представлений о «критических» отклонениях от расчетной модели.

Минимизация (11.3) с использованием хьюберовской функции ρ сводится к решению системы уравнений

$$\sum_{i=1}^{n} \Psi_h \left[\left(\frac{y_i - \sum_{j=1}^{m} x_{ij}\beta_j}{h} \right) \right] x_{ij} = 0, \quad (j = 1, 2, \ldots, m), \quad (11.5)$$

где

$$\Psi_h(\varepsilon) = \begin{cases} \varepsilon & \text{при } |\varepsilon| \leqslant h, \\ h \operatorname{sig} h(\varepsilon) & \text{при } |\varepsilon| > h. \end{cases}$$

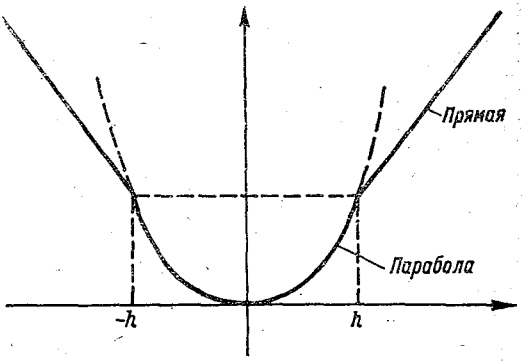


Рис. 11.4. График функции Хьюбера.

При неотрицательных значениях h уравнение (11.5) можно представить в виде

$$\sum_{i=1}^{n} x_{ij}\varepsilon_i(h)\, w(h, \varepsilon_i) = 0,$$

где

$$\varepsilon_i(h) = y_i - \sum_{j=1}^{m} x_{ij}\hat{\beta}_j(h); \quad w(h, \varepsilon_i) = \begin{cases} \Psi_h(\varepsilon_i)/\varepsilon_i & \text{при } |\varepsilon_i| \leqslant h, \\ h/|\varepsilon_i| & \text{при } |\varepsilon_i| > h, \end{cases}$$

т. е. имеем нормальную систему уравнений для взвешенного регрессионного метода наименьших квадратов с весом w, зависящим от $\widehat{\beta}_j$. Это предполагает использование итеративного взвешенного алгоритма Гаусса—Ньютона. Данный метод решения был предложен Тьюки [20].

Несмотря на то что описанный выше робастный подход к регрессии развит для симметрического распределения ошибок уравнения, Хьюбер с помощью метода Монте-Карло установил, что при практическом применении метода смещение оценок из-за несимметричности мало [48].

В агрометеорологии методы робастной регрессии не получили пока достаточного распространения, хотя выполненные с их помощью прикладные работы в других отраслях знаний дали положительные результаты и свидетельствуют о больших возможностях метода.

Приведем некоторые результаты применения метода робастной регрессии при построении прогностических регрессионных моделей урожайности всех зерновых культур в областях Казахстана. Для каждой из пяти групп областей со сходными природно-климатическими условиями строилось одно регрессионное уравнение. Предварительно с помощью автоматических процедур были отобраны наиболее информативные предикторы. Исследования закона распределения урожайности или, точнее, отклонений урожайности от временного тренда показали, что он далек от нормального. Здесь уместно напомнить, что в классической постановке регрессионной задачи ошибки уравнения распределены по такому же закону, как и зависимая переменная. Для анализа закона распределения строились гистограммы, пробит-графики. На рис. 9.9 и 9.10 показаны типичные гистограммы отклонений урожайности от соответствующих областных трендов в Северном и Южном Казахстане. Как упоминалось в разделе 9.4, анализ частот отклонений урожайности зерновых культур от трендов во всех пяти группах областей выявил достаточно устойчивую закономерность: гистограммы распределения не являются, строго говоря, унимодальными. Можно отметить наличие двух максимумов частоты, соответствующих отрицательным и положительным отклонениям урожайности от трендов. Этот факт служит иллюстрацией необходимости осторожного использования простого осреднения величин для получения «свернутой» информации о природных явлениях.

Из теории следует, что при подстановке в регрессионное уравнение средних значений предикторов получим среднее значение предиктанта. На основании свойств линейной комбинации можно сделать вывод, что если предиктант — отклонение урожайности от тренда — не подчиняется нормальному закону распределения, то и предикторы не подчиняются этому закону. Следовательно, такие широко распространенные для характеристики нормально распределенных переменных статистики, как среднее арифметическое и дисперсия, не всегда пригодны и могут ввести в заблуждение.

Вполне вероятно, что в Казахстане повторяемость так называемых средних, или климатически нормальных, агрометеорологических условий (которым должна соответствовать урожайность, близкая к уровню тренда) меньше, чем повторяемость условий, вызывающих заметное повышение или понижение средней областной урожайности всех зерновых культур. Во всех группах областей максимум частоты положительных отклонений урожайности от тренда значительно превосходит максимум частоты отрицательных отклонений. На большинстве таких гистограмм левое «крыло» длиннее правого, что свидетельствует о более сильном влиянии

неблагоприятных агрометеорологических условий на снижение урожая, чем благоприятных — на его повышение.

Устойчивость двух максимумов частоты на гистограммах говорит об их неслучайном характере и отличии распределения от гауссовского закона. Это требует осторожности в применении известных методов проверки нормальности распределения величин, в основе которых лежит нулевая гипотеза о нормальности распределения исходной выборки. Если даже и принять такую нулевую гипотезу и проверить нормальность распределения с помощью метода, изложенного в [9], с использованием отношения размаха значений выборки к среднему квадратическому отклонению, то в двух группах областей нулевая гипотеза отвергается при 10 %-ном уровне значимости.

Помимо проверки нормальности распределения используемых переменных необходимо уделять внимание выявлению данных, резко выделяющихся по какому-либо критерию. Присутствие даже малого количества таких данных значительно влияет на получаемые методом наименьших квадратов коэффициенты уравнений регрессии.

Для всех пяти групп областей строились уравнения по выбранным оптимальным предикторам. Множественные коэффициенты корреляции регрессионных уравнений, построенных обычным методом наименьших квадратов, достаточно высокие: от 0,79 до 0,85.

На основании анализа ошибок обычных регрессионных уравнений и с учетом характерных особенностей агрометеорологических данных для построения робастных моделей были выбраны два значения параметра h: 1,0 и 0,5. При $h=1{,}0$ незначительная доля данных (до 11 %) подвергалась (в малой степени) робастизирующему воздействию данной процедуры. При $h=0{,}5$ эта доля существенно возрастала (18—40 %) и влияние процедуры усиливалось. Отсюда следует, что уравнения с параметром робастности $h=1$ в меньшей степени отличаются от обычных уравнений, чем уравнения с $h=0{,}5$.

Для проверки эффективности применения робастной регрессии использовалась независимая выборка. В табл. 11.3 показано уменьшение средней квадратической ошибки прогноза при прогнозировании с помощью робастных уравнений по сравнению

Таблица 11.3

Уменьшение (%) средней квадратической ошибки прогноза на независимой выборке при использовании робастной регрессии

h	Западный Казахстан	Восточный Казахстан	Северный Казахстан	Центральный Казахстан	Южный Казахстан
1,0	7	7	6	0	1
0,5	19	13	10	1	6

с обычными классическими уравнениями при различных значениях h во всех группах областей.

По ряду наблюдений в группе областей Северного Казахстана после удаления двух лет, отнесенных к выбросам, для параметров робастности 0,8, 0,5, 0,3 были построены уравнения и на 15 % данных выборки проведена проверка на независимом материале. Оказалось, что ошибка уменьшилась по сравнению с классическим уравнением соответственно на 9, 16 и 20 %.

В большинстве случаев при наличии выбросов и отклонении распределения зависимой переменной от нормального закона использование описанного метода робастной регрессии дает положительный эффект, что особенно важно при небольшом количестве данных, когда одно-единственное «отскочившее» наблюдение может испортить регрессионную модель.

Сравнивая коэффициенты в обычном уравнении и робастном при параметре робастности $h=0{,}3$ для группы областей Южного Казахстана, можно увидеть, что они значительно отличаются друг от друга (до 24 %), а в отдельных случаях даже имеют противоположные знаки (табл. 11.4). Так, в классическое уравнение показатель влагообеспеченности июня вошел с отрицательным знаком, что не согласуется с реальным вкладом этого фактора. Изменение знака коэффициента на положительный в уравнении робастной регрессии свидетельствует о его более реалистичном и адекватном отражении действительности.

Таблица 11.4

Коэффициенты уравнений регрессии ($\times 10^{-3}$)

Метод	Показатель влагообеспеченности третьей декады июня	Сумма осадков за май	Температура воздуха за июнь	Показатель влагообеспеченности июня	Показатель влагообеспеченности июля	Сумма осадков за холодный период
Классический	308	289	—345	—22	359	135
Робастный (при $h=0{,}3$)	249	335	—289	50	445	119

Достаточно обширные исследования показали эффективность использования метода робастной регрессии в случаях отклонения распределения предиктанта от нормального закона и при наличии выбросов. Автоматическую защиту коэффициентов регресионных моделей с помощью методов робастной регрессии рекомендуется использовать при недостатке времени, отведенного для решения задачи, и при отсутствии достаточной квалификации у прогнозиста.

11.3. МОДЕЛЬ ГРЕБНЕВО-РОБАСТНОЙ РЕГРЕССИИ

Разработанные в последнее время альтернативные классическому методу наименьших квадратов методы позволяют получать хорошие результаты в реальных условиях эксперимента при нарушении той или иной основополагающей гипотезы, на которую опирается классический метод. Естественно, возникает вопрос о создании методов, которые смогли бы при нарушении нескольких гипотез одновременно (что бывает не редко) обладать устойчивыми свойствами отдельных альтернативных методов.

Полезность использования в агрометеорологических исследованиях гребневой и робастной регрессий была продемонстрирована достаточно убедительно. Публикации ВМО также ратуют за широкое применение современных статистических методов, в частности, гребневой регрессии [43]. Поэтому можно предположить, что совместное использование моделей робастной и гребневой регрессии может дать наибольший эффект. Для этой цели нами применялись подходы, описанные ранее для гребневой и робастной регрессии. При расчетах использовался взвешенный итеративный метод наименьших квадратов на основе алгоритма Гаусса—Ньютона [25]. Как уже указывалось, метод робастной регрессии с функцией потерь, предложенной Хьюбером, можно свести к хорошо известному взвешенному методу наименьших квадратов.

Для решения робастной задачи имеем систему нормальных уравнений

$$\sum_{i=1}^{n} x_{ij}\varepsilon_i(h)\, w(h,\ \varepsilon_i) = 0,$$

где

$$\varepsilon_i(h) = y_i - \sum_{j=1}^{m} x_{ij}\hat{\beta}_j(h), \quad w(h,\ \varepsilon_i) = \begin{cases} \Psi_h^{\mathrm{T}}(\varepsilon_i)/\varepsilon_i & \text{при } |\varepsilon_i| \leqslant h, \\ h/|\varepsilon_i| & \text{при } |\varepsilon_i| > h \end{cases}$$

(вес w зависит от $\hat{\beta}_j$).

На каждом шаге

$$\Delta\beta = (\mathbf{X}'w(h,\ \varepsilon_i)\,\mathbf{X})^{-1}\,\mathbf{X}'w(h,\ \varepsilon_i)\,\varepsilon_i.$$

При коррелированности предикторов, т. е. столбцов матрицы $\mathbf{X}$, предлагается добавлять стабилизирующую диагональную матрицу $\mathbf{I}\sqrt{k}$ к матрице $\mathbf{X}'w(h,\ \varepsilon_i)\,\mathbf{X}$. Это позволит сделать матрицу $\mathbf{X}'w(h,\ \varepsilon_i)\,\mathbf{X}+\mathbf{I}\sqrt{k}$ менее вырожденной и облегчит нахождение обратной матрицы.

Первоначальный выбор параметра k может осуществляться на основе анализа гребневого следа (см. раздел 11.1). Многочисленные расчеты на реальных данных показали, что гребнево-робастный метод чаще других оказывался лучшим по критерию

средней ошибки модели на независимых данных. Знаки регрессионных коэффициентов, полученные с помощью этого метода, становились физически более оправданными при стабилизирующем систему гребневом параметре k, меньшем, чем при чистом методе гребневой регрессии. Для нахождения оценок методом Гаусса—Ньютона, как правило, достаточно четырех-пяти итераций. Начальные значения коэффициентов $\hat{\beta}_j$ можно задавать равными нулю или единице.

11.4. МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Анализ главных компонент является наиболее популярным и одним из самых простых способов изучения многомерных статистических совокупностей. Он находит широкое применение в метеорологии, агрометеорологии, экологии и других естественных науках [1, 4, 26], где часто используется совместно с методами регрессионного анализа. Суть этого подхода состоит в замене сильно коррелированных переменных совокупностью других переменных, являющихся линейными комбинациями исходных и подобранных так, чтобы между ними корреляции отсутствовали. В этом случае при некоррелированности (ортогональности) переменных значительно упрощается расчет коэффициентов регрессионных моделей.

Преобразованные переменные — главные компоненты обладают еще одним важным свойством: первая компонента соответствует направлению максимально возможной вариации в пространстве переменных, вторая — направлению максимально возможной вариации в подпространстве, ортогональном первому направлению и т. д.

В основе этого метода лежит отыскание собственных чисел и соответствующих собственных векторов ковариационной или корреляционной матрицы переменных.

Если регрессионную модель представить в обычной форме

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

где $\mathbf{X}$ — матрица, составленная из данных наблюдений за считающимися случайными переменными $x_1, x_2, \ldots, x_p$, то систему уравнений для главных компонент $z_1, z_2, \ldots, z_p$, являющихся линейными комбинациями исходных переменных, можем записать в виде

$$\begin{aligned} z_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p, \\ z_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ z_p &= a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pp}x_p. \end{aligned}$$

В матричной форме главные компоненты матрицы $\mathbf{X}$ задаются так:

$$\mathbf{Z} = \mathbf{XA},$$

где матрица **Z** есть аналог матрицы **X** для новых переменных Для отыскания главных компонент надо построить такую ортогональную матрицу **A**, чтобы **Z** имела диагональную ковариационную матрицу **D**

$$\mathbf{D} = \mathbf{Z}'\mathbf{Z}.$$

Найдем матрицу **A**, удовлетворяющую условиям

$$\mathbf{A}\mathbf{A}' = \mathbf{A}'\mathbf{A} = \mathbf{I} \quad \text{и} \quad \mathbf{A}'(\mathbf{X}'\mathbf{X})\,\mathbf{A} = \mathbf{D}.$$

Пусть значения диагональных элементов матрицы **D** упорядочены, т. е. $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \ldots \geqslant \lambda_p$. Эти числа называют собственными числами матрицы **X'X**, для невырожденных матриц они больше нуля. Если есть нулевые собственные числа, то матрица вырожденная.

Столбцы в матрице **A** называют собственными векторами, а столбцы в матрице **Z** = **XA**, являющиеся новыми переменными,— главными компонентами.

Общая вариация независимых переменных определяется как

$$\sum_i x_{1i}^2 + \sum_i x_{2i}^2 + \ldots + \sum_i x_{pi}^2 = \sum_{j=1}^{p} \sum_{i=1}^{n} x_{pi}^2 = \sum_{i=1}^{p} \lambda_i$$

и отражает меру разброса точек в пространстве наблюдений. Она численно равна сумме собственных чисел матрицы **X'X**. Собственные числа отражают дисперсии главных компонент, а отношение $\lambda_1/\sum_i \lambda_i$, $\lambda_2/\sum_i \lambda_i$, ..., $\lambda_p/\sum_i \lambda_i$ представляет относительный вклад каждой компоненты в общую вариацию исходных переменных. Сумма всех вкладов равна единице.

Часто переменные $x_1, x_2, \ldots, x_p$ выражаются в различных и несопоставимых единицах измерения, поэтому трудно разделить общую вариацию на составляющие — главные компоненты. Обычно рекомендуется предварительно стандартизировать все переменные по формуле

$$\frac{x_i - \bar{x}_i}{\sqrt{n}\,\sigma}.$$

Тогда матрица **X'X** станет матрицей коэффициентов корреляции системы независимых переменных.

В этом случае для собственных значений выполняется условие

$$\lambda_1 + \lambda_2 + \ldots + \lambda_p = p.$$

Линейные преобразования исходных переменных, как, например, стандартизация, приводят к различным собственным векторам и значениям.

Наличие коррелированности у переменных означает, что можно меньшим количеством переменных описать значительную часть общей вариации переменных. Предполагают, что несколько первых главных компонент (напомним, что они ранжированы по

вкладу в общую вариацию) несут существенную информацию, а компоненты с малыми значениями λ_i содержат информационный шум и поэтому ими можно пренебречь. Вместо построения регрессионного уравнения с большим числом исходных переменных, коррелированных между собой, и оттого имеющего ряд серьезных недостатков, строят модель с небольшим количеством отобранных новых переменных — главных компонент.

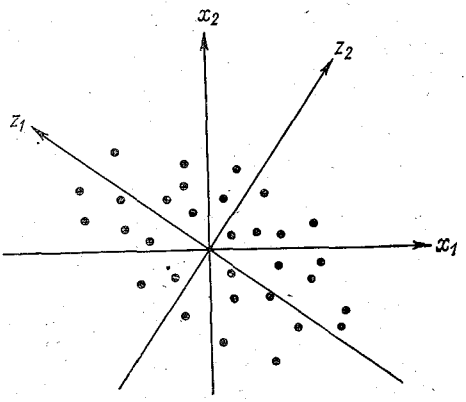


Рис. 11.5. Геометрическая интерпретация метода главных компонент в двумерном случае.

Полученным главным компонентам не всегда можно придать определенный физический смысл. Однако иногда это удается сделать. При построении модели

$$y = \alpha_1 z_1 + \alpha_2 z_2 + \ldots + \alpha_q z_q$$

ошибка коэффициента регрессии минимальна для первой главной компоненты и последовательно увеличивается для последующих компонент. Так как главные компоненты не коррелированы, то включение в уравнение новых компонент не изменяет коэффициенты у компонент, уже включенных в модель, что упрощает проведение расчетов.

При построении регрессионной модели с использованием главных компонент удобно применить процедуру пошагового отбора переменных (см. раздел 10.2). Последовательно в уравнение включают главные компоненты с наибольшими коэффициентами парной корреляции с зависимой переменной. Чаще всего такими переменными являются несколько первых главных компонент, несущих максимум информации об изменчивости исходных переменных.

Можно дать геометрическую интерпретацию анализа главных компонент. На рис. 11.5 приведено поле точек двух переменных x_1 и x_2. Надо найти такие переменные

$$z_1 = \alpha_{11} x_1 + \alpha_{12} x_2,$$
$$z_2 = \alpha_{12} x_1 + \alpha_{22} x_2$$

или, что то же самое, так повернуть ортогональные оси x_1 и x_2, чтобы их новое направление соответствовало направлениям с мак-

симальной вариацией поля признаков. Аналогичную процедуру поворота осей осуществляют в p-мерном пространстве признаков при рассмотрении p переменных. Собственные числа при этом отражают «толщину», или вариабельность, поля точек в направлении новых выбранных координатных осей.

Если при построении регрессионной модели включить в нее все главные компоненты, то получим модель, идентичную той, которая построена по всему множеству исходных переменных. Для удобства использования регрессионного уравнения по главным компонентам его преобразуют в модель по исходным переменным; например, для трех компонент при q независимых переменных:

$$y = \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 = \alpha_1 (a_{11} x_1 + \ldots + a_{q1} x_q) + \alpha_2 (a_{12} x_1 + \ldots + a_{q2} x_q) + \alpha_3 (a_{13} x_1 + \ldots + a_{q3} x_q) = (\alpha_1 a_{11} + \alpha_2 a_{12} + \alpha_3 a_{13}) x_1 + (\alpha_1 a_{21} + \alpha_2 a_{22} + \alpha_3 a_{23}) x_2 + \ldots + (\alpha_1 a_{q1} + \alpha_2 a_{q2} + \alpha_3 a_{q2}) x_q.$$

Пример. Строилась регрессионная модель зависимости урожайности зерновых культур от агрометеорологических условий в Уральском экономическом районе, где большие площади занимают озимые культуры. Предварительным анализом были отобраны данные за 37 лет о сумме осадков и температуре воздуха за ноябрь и декабрь предшествующего года (соответственно R_{XI}, R_{XII}, t_{XI}, t_{XII}), о сумме осадков за период апрель—июнь и температуре воздуха в марте и апреле (соответственно R_{IV-VI}, t_{III}, t_{IV}). Для учета постоянно растущего уровня земледелия в качестве дополнительной переменной в модель вводился фактор времени T. В табл. 11.5 приведена корреляционная матрица всех переменных.

Таблица 11.5

Корреляционная матрица системы переменных ($\times 10^{-2}$)

Переменная	y	T	R_{IV-V}	t_{III}	t_{IV}	t_{XI}	t_{XII}	R_{XI}	R_{XII}	R_{I-III}
y	100									
T	68	100								
R_{IV-V}	82	56	100							
t_{III}	21	26	12	100						
t_{IV}	8	6	—11	32	100					
t_{XI}	34	40	27	12	—1	100				
t_{XII}	—15	7	—12	9	—4	7	100			
R_{XI}	14	36	11	5	—26	25	—1	100		
R_{XII}	33	11	30	—22	—48	37	23	15	100	
R_{I-III}	35	24	9	—10	—4	16	—13	—1	3	100

Примечание. Здесь y — урожайность.

гональной с неодинаковыми элементами, поэтому прямое использование метода наименьших квадратов неправомерно. В данном случае часто прибегают к преобразованию переменных и используют обычную процедуру или взвешенный метод наименьших квадратов.

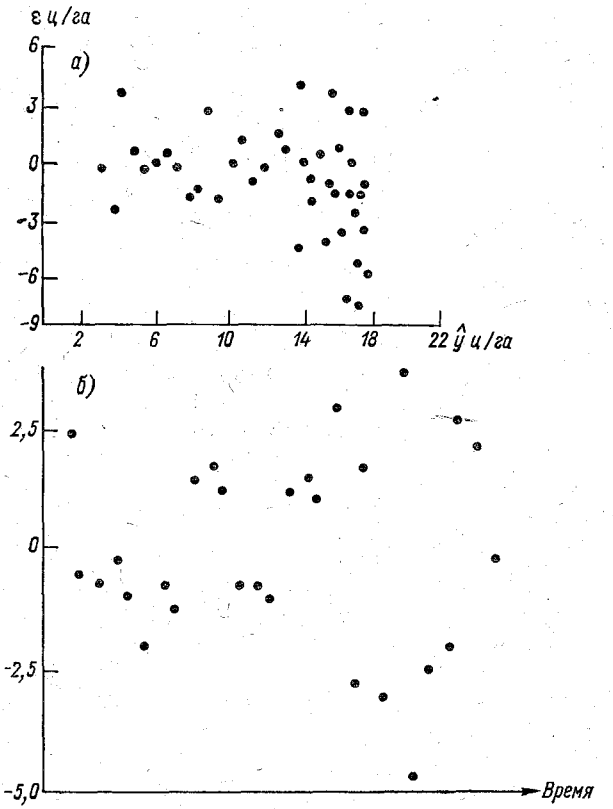


Рис. 11.6. Пример нестационарности дисперсии остатков (ε) регрессионных моделей тренда урожайности зерновых культур в Центральночерноземном районе (*а*) и тренда урожайности подсолнечника в Винницкой области (*б*).

Нормальное уравнение тогда имеет вид

$$\mathbf{X}'\mathbf{W}\mathbf{X}\beta = \mathbf{X}'\mathbf{W}\mathbf{y},$$

где $\mathbf{W}$ — диагональная матрица, у которой на главной диагонали находятся веса, обратно пропорциональные дисперсии отклонений (ошибки).

Простым примером, иллюстрирующим необходимость применения метода взвешенной регрессии, может служить построение тренда урожайности (рис. 11.6). Во многих случаях наблюдается

увеличение отклонения урожайности от уровня тренда с течением времени. На рис. 11.6*б* общий характер этих отклонений представляет собой расширяющуюся со временем полосу, что указывает на целесообразность использования взвешенного метода наименьших квадратов при построении регрессионных моделей (то же и на рис. 11.6*а*).

Для получения оценок весов можно разбить весь временной интервал на k подинтервалов и для каждого вычислить свое значение дисперсии, а затем, построив регрессионную зависимость увеличения дисперсии отклонения от времени, задать каждому году наблюдений конкретное значение w_i.

Применение взвешенного метода наименьших квадратов не ограничивается таким классическим случаем. Обычно, аппроксимируя тенденцию роста урожайности прямой линией, получают среднюю динамику прироста урожайности за все годы, хотя в ряде прикладных задач большой интерес представляет характер изменений урожайности, имеющий место в последние годы. В этом случае разумно задать годам веса таким образом, чтобы они отражали уменьшение ценности информации при ее старении.

Еще одна из важнейших областей применения взвешенного метода наименьших квадратов это — построение регрессионных моделей при наличии в данных резко отличающихся друг от друга совокупностей (групп) наблюдений. На практике не всегда возможно разделить данные на достаточно крупные массивы для раздельного построения моделей. Например, при построении прогностических моделей урожайности одни годы — с аномально плохими условиями, другие — с благоприятными. Воспользоваться только этим небольшим количеством лет для разработки регрессионных моделей для каждого типа условий обычно невозможно. Здесь уместно использовать всю имеющуюся информацию, хотя вследствие резко различающихся условий формирования урожая коэффициенты модели должны быть существенно иными. Выход из положения заключается в построении двух регрессионных моделей по всем данным: в первом случае бо́льшие веса придают аномально плохим годам выборки, во втором — благоприятным годам. При практическом использовании моделей необходимо определить, к какому типу относится год составления прогноза. Эту процедуру можно формализовать, используя методы дискриминантного анализа. Возможны различные варианты предложенного подхода. При этом ключевым моментом является выбор весов. Одним из возможных подходов в первом приближении может служить выбор весов пропорционально ошибкам невзвешенной регрессионной модели. Например, для модели, лучше отражающей засушливые условия, можно положить веса равными

$$w_i = \begin{cases} |\varepsilon_i| + 1, & \varepsilon_i \leqslant 0, \\ \dfrac{1}{\varepsilon_i + 1}, & \varepsilon_i > 0. \end{cases}$$

Использование взвешенного метода позволяет заданием веса не только уменьшать влияние какого-либо года из набора данных, но и вовсе исключить его из анализа.

11.6. МОДЕЛЬ С ОШИБКАМИ В НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ

Рассмотренные в гл. 9 оценки коэффициентов уравнения регрессии были получены в предположении, что матрица **X** не стохастична и детерминирована, т. е. погрешности в измерении ее элементов пренебрежимо малы. В реальной практике такие предположения не всегда оправданы. Более реалистичным представляется подход, в котором допускается наличие ошибок в матрице наблюдений. Наличие ошибок неизбежно, например, в случае, когда элементы матрицы **X** являются осредненными по площади значениями метеорологических величин и служат оценкой истинных средних по конечному числу метеостанций.

В классической линейной регрессии оценки, полученные методом наименьших квадратов, при принятых гипотезах были несмещенными, эффективными в классе несмещенных оценок. В случае регрессионной модели с ошибками в независимых переменных эти свойства пропадают.

В зависимости от способа получения данных выделяют три метода оценивания параметров уравнений.

1. Связь независимой переменной с зависимой переменной описывается схемой, представленной на рис. 11.7 *а*. Этой схеме соответствует модель

$$y_i = f(x_i,\ \beta_0) + \varepsilon_i$$

(индексом «0» отмечается истинное значение параметра).

2. Эксперимент описывается схемой, представленной на рис. 11.7 *б*. Известны значения x_{0i} и y_i. Этой схеме соответствует модель

$$y_i = f(x_{0i} + \nu_i,\ \beta_0) + \varepsilon_i$$

(ν_i — ошибка измерения независимых переменных).

3. Связь между независимой и зависимой переменными описывается схемой, приведенной на рис. 11.7 *в*. Известны наблюдаемые x_i, y_i. Этой схеме соответствует модель

$$y_i = f(x_{0i},\ \beta_0) + \varepsilon_i,$$

$$x_i = x_{0i} + \nu_i.$$

Модель 1 соответствует классической регрессионной схеме. Методы получения оценок модели в этом случае хорошо описаны в гл. 8.

Модель 2 описывает так называемые активные эксперименты. Эта схема хорошо соответствует реальным условиям при использовании теории планирования эксперимента.

Модель пассивных наблюдений 3 наиболее подходит для описания неконтролируемых воздействий внешней среды на биологический объект. В ней рассматриваются значения факторов x_i, сгенерированных природой, на которые влияет случайная ошибка. Исследователь наблюдает не истинное значение x_i, а ее значение с погрешностью ν_i. Как и в классическом варианте регрессии,

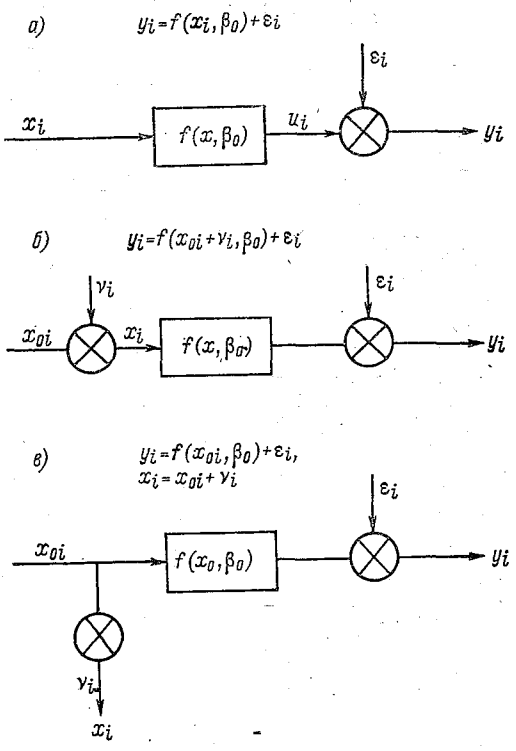


Рис. 11.7. Схемы регрессионных моделей: классической регрессии (*а*), активного эксперимента (*б*), пассивного эксперимента (*в*).

зависимая переменная наблюдается со случайной ошибкой, т. е. известны

$$y_i = v_i + \varepsilon_i,$$
$$u_i = x_{0i} + \nu_i.$$

Предполагается, что истинные значения величин v_i и x_{0i} связаны функциональным соотношением

$$v_i = f(x_{0i}, \beta_0),$$

из-за имеющихся ошибок оно переходит в так называемое структурное, приведенное выше.

Оценивая коэффициенты уравнений регрессии, необходимо принять некоторые гипотезы о свойствах ошибок:

1) все погрешности имеют нулевые средние;

2) ошибки независимых переменных не коррелированы между собой и имеют одинаковую дисперсионную матрицу;

3) ошибки зависимой и независимой переменных не коррелированы между собой;

4) ошибки зависимой переменной не коррелированы между собой.

Исходя из этих предположений, оценку $(\widehat{\beta})$ истинного значения β можно получить по формуле

$$\widehat{\beta} = \underset{\beta}{\operatorname{Argmin}}\ n^{-1} \sum_{i=1}^{n} \lambda(x_i,\ \beta)\,(y_i - \mathbf{f}(x_i,\ \beta_0)),$$

где

$$\mathbf{f}(\mathbf{x},\ \beta) = \mathbf{f}(\mathbf{x},\ \beta) - \frac{1}{2}\,\mathbf{f}''(\mathbf{x},\ \beta)\,\mathbf{C},\quad \lambda = \left(\sigma^2 + \mathbf{f}'^{\mathrm{T}}(x,\ \beta)\,\mathbf{C}\mathbf{f}'(\mathbf{x},\ \beta)\right)^{-1},$$

$$\mathbf{C} = E(\nu_i,\ \nu_{i'}').$$

Предполагается, что производные $\mathbf{f}(\mathbf{x})$ до третьего порядка равномерно ограничены по $\mathbf{x}$.

В случае линейной регрессии

$$y_i = \sum_j \beta_j x_{ij} + \varepsilon_i,\quad (i = 1,\ 2,\ \ldots,\ n),$$

$$x_i = x_{0i} + \nu_i.$$

Оценка β_j совпадает с оценками, получаемыми методом наименьших расстояний из соотношения

$$\widehat{\beta} = \underset{\beta}{\operatorname{Argmin}}\ n^{-1} \sum_{i=1}^{n} \frac{\left(y_i - \sum\limits_j \beta_j x_{ij}\right)^2}{\sigma^2 + \beta' \mathbf{C} \beta}.$$

Для простоты обычно полагают $\mathbf{C} = \gamma^2 \mathbf{I}$, что еще больше облегчает задачу расчета коэффициентов.

Имитационные исследования точности расчета коэффициентов как линейных, так и нелинейных регрессионных уравнений, построенных методами, учитывающими наличие ошибок в предикторных переменных, показали их бóльшую точность по сравнению с уравнениями, рассчитываемыми методом наименьших квадратов. Было показано [23], что смещение оценки коэффициентов регрессионного уравнения зависит как от погрешности наблюдения переменных, так и от степени их коррелированности. Небольшие ошибки в независимых переменных в случае плохой обусловленности корреляционной матрицы приводят к большому смещению оценок коэффициентов. С увеличением длины выборки влияние ошибок переменных на точность оценки параметров сказывается меньше.

Для примера приведем результаты построения регрессионного уравнения для группы, состоящей из Алма-Атинской, Чимкентской, Джамбулской и Тады-Курганской областей. В качестве независимых переменных в прогностическую модель урожайности всех зерновых культур были взяты температура воздуха за третью декаду апреля, сумма осадков и температура воздуха за вторую декаду мая, температура воздуха за третью декаду июня, показатель влагообеспеченности за период с третьей декады мая по первую декаду июня, показатель влагообеспеченности в целом за июнь, а также сумма осадков за зимний период, с октября по март, и сумма осадков за вторую и третью декады апреля.

Вышеперечисленные переменные были определены после всестороннего анализа агрометеорологических факторов, влияющих на урожай. При расчете параметров в модели с ошибками в переменных необходимо задать оценку дисперсии ошибок как независимых, так и зависимой переменной. В первом приближении дисперсию можно определить, исходя из данных о точности расчетов независимых переменных. Однако точно рассчитать дисперсию ошибок каждой переменной трудно. На основании того факта, что все переменные, выбранные для построения модели, были предварительно приведены к единичной дисперсии и нулевому среднему, для переменных была установлена одинаковая дисперсия ошибки наблюдения независимой переменной в долях дисперсии. Расчеты велись для различных уровней ошибки независимых переменных (γ^2). Для оценки дисперсии ошибки зависимой переменной использовалась дисперсия ошибки 0,368, полученная при построении обычного регрессионного уравнения методом наименьших квадратов с теми же независимыми переменными.

Рассчитанные коэффициенты регрессионных уравнений приведены в табл. 11.9. В последней строке таблицы указаны средние квадраты ошибок при проверке уравнений на независимом материале.

При $\gamma^2=0{,}005$ средний квадрат ошибки прогноза на независимых данных по сравнению с обычным регрессионным уравнением уменьшается на 9 %. Такой же практически результат и при $\gamma^2=0{,}01$. Если перевести ошибку из относительных единиц в абсолютные, то окажется, что средняя квадратическая ошибка третьей переменной по всем четырем областям равна 0,2 и 0,3 °С (соответственно при $\gamma^2=0{,}005$ и $\gamma^2=0{,}01$). Для первой переменной та же ошибка составила соответственно 0,31 и 0,44 мм. Точность оценки декадных осадков, вероятно, занижена, а точность оценки средней декадной температуры воздуха кажется вполне реальной. Отметим, что использовались модифицированные данные о сумме осадков с ограниченной дисперсией. Если для каждого типа входящих в регрессионное уравнение переменных устанавливать свой уровень точности, то уравнения станут более адекватными, их прогностические возможности, несомненно, возрастут. Задавая различные сочетания ошибок для однородных групп переменных, можно принять гипотезу, о том, что при минимуме квадрата

ошибки уравнения на «свежих» данных получим оценки ошибок, близкие к истинным.

Если сравнивать полученные разными методами коэффициенты регрессионных уравнений в табл. 11.9, можно заметить, что отрицательный коэффициент при восьмой переменной в обычном уравнении становится положительным в моделях с «ошибками». С точки зрения агрометеоролога, это вполне закономерно, так как роль запасов влаги в почве, накопленных на начало весны и зависящих главным образом от количества зимних осадков, положительна.

Таблица 11.9

Коэффициенты регрессионных уравнений при различных уровнях ошибок в переменных

№ п/п	Переменная	Коэффициент регрессии			
		при дисперсии ошибок			полученный методом наименьших квадратов
		0,005	0,01	0,05	
1	Сумма осадков за V_2	0,1673	0,1688	0,1553	0,1165
2	Температура воздуха за IV_3	0,1355	0,1390	0,1727	0,1610
3	Температура воздуха за V_2	—0,2583	—0,2581	—0,2398	—0,1430
4	Температура воздуха за VI_3	0,4779	0,4872	0,5603	0,5170
5	Показатель влагообеспеченности за V_3—VI_1	0,9876	1,0230	1,3640	1,0700
6	Показатель влагообеспеченности за VI	—0,4410	—0,4717	—0,7481	—0,4330
7	Сумма осадков за IX—III	0,0076	0,0067	0,0037	—0,0195
8	Сумма осадков за IV_2—IV_3	0,2533	0,2540	0,2514	0,2228
	$\sum_i \varepsilon_i^2/m$	0,4331	0,4351	0,5187	0,4734

Примечание. Здесь V_3 — третья декада мая и т. д.

В первый момент вызывает также сомнение наличие отрицательного коэффициента у шестой переменной. Однако если его не вводить в уравнение, то ошибка прогноза на независимой выборке становится на 8 % больше, чем при его учете. Можно попытаться найти разумное объяснение такому неожиданному явлению. Рассматриваемая группа областей включает самую южную зону Казахстана, где преобладает клин озимых культур. В середине июня там начинается массовая уборка озимых культур, и повышенная влагообеспеченность может привести к полеганию посевов; чередование дождей с жаркой погодой способствует осыпанию зерна [14]. Влагообеспеченность же первой декады июня учитывается с положительным знаком в показателе сред-

ней влагообеспеченности за третью декаду мая и первую декаду июня.

Таким образом, использование регрессионной модели, в которой учитывают ошибки в независимых переменных, повышает точность прогнозирования на независимом материале и дает возможность оценить коэффициенты уравнения «физически» более надежно.

11.7. МОДЕЛЬ С ОГРАНИЧЕНИЯМИ НА КОЭФФИЦИЕНТЫ

Развитие регрессионных методов моделирования в естественных науках в последние годы идет путем «усвоения» специфических особенностей исследуемых явлений и объектов. От сравнительно простых моделей, в которые вкладывались лишь некоторые из множества возможных факторов, стали переходить к достаточно сложным моделям, учитывающим их свойства и отношения. Такие модели иногда называют физико-статистическими, чтобы подчеркнуть использование априорной информации об интересующих нас процессах и объектах.

Одним из таких методов, позволяющим вкладывать априорную информацию в модель является регрессия с ограничениями на коэффициенты. Достаточно часто из общих соображений и на основании накопленного опыта моделирования возможно задать некоторые содержательные соотношения между коэффициентами модели или определить область их изменений. Например, пусть определенные коэффициенты уравнения будут больше заданных значений или равны между собой.

В общем случае, если имеем обычную модель

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

и требуется найти методом наименьших квадратов ее коэффициенты, на них налагают линейные ограничения типа равенств

$$\mathbf{A}\beta = \mathbf{c},$$

где $\mathbf{A}$ — заданная матрица коэффициентов в линейных соотношениях между β_j, $\mathbf{c}$ — известный вектор правых частей уравнений. При задании ограничений типа равенств для решения поставленной задачи обычно используют метод неопределенных множителей Лагранжа. В результате получают решение, удовлетворяющее обеим системам в виде

$$\hat{\beta}_0 = \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}A')^{-1}(\mathbf{c} - \mathbf{A}\hat{\beta}),$$

где $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ — обычное решение исходной регрессионной задачи без ограничений методом наименьших квадратов.

При наличии простейших ограничений на коэффициенты типа неравенств $\beta_j \geqslant k_j$, $\beta_j > k_j$ пользуются методами математического программирования, близко примыкающими к регрессионному анализу.

Впервые модель с ограничениями на параметры в агрометеорологии предложена в [27] при построении уравнения регрессии для восстановления декадных сумм осадков на отдельных станциях по данным ближайших станций на территории Литовской ССР.

Пример. Исследуемые поля сумм осадков описывались 21 метеостанцией. Считалось, что на 15 метеостанциях данных нет и их надо восстановить по данным шести метеостанций. Для проверки точности такого интерполирования использовались данные независимой выборки. Для каждой из 15 метеостанций определялись коэффициенты в линейной модели зависимости сумм осадков на метеостанции с «неизвестными» данными от данных по шести опорным станциям. На коэффициенты моделей было наложено ограничение $\beta_j \geqslant 0$.

Поставленная задача является задачей квадратического программирования с линейными ограничениями на коэффициенты типа неравенств. Она решалась методом случайного поиска минимума ошибки уравнения. Одновременно для сравнения рассчитывались регрессионные модели, в которых использовались классические подходы. Они имели несколько отрицательных коэффициентов.

Средний по 15 метеостанциям коэффициент корреляции между вычисленными и фактическими суммами осадков на независимых данных для моделей классического регрессионного анализа оказался равным 0,373, а для модели с ограничениями — 0,685. Как видно, использование даже простейших содержательных ограничений, отражающих априорные знания, резко повышает качество статистических моделей.

Глава 12. АНАЛИЗ РЕГРЕССИОННЫХ ОСТАТКОВ И ВЫБРОСОВ

Статистическое моделирование представляет собой ряд последовательных шагов, приближающих к «правильной» и полезной в использовании модели. Одним из неизбежных этапов построения моделей является ее проверка, или верификация, на пригодность для достижения поставленных целей.

Рассмотрим широко используемые на практике простые приемы, позволяющие своевременно совершенствовать модель, исправлять ее грубые изъяны. В этом значительное место отводится наглядным графическим представлениям.

При детальной оценке модели необходимо проанализировать каждый из имеющихся в наборе данных случаев (строка в матрице наблюдений **X** и соответствующее ей значение y_i), его роль и влияние на коэффициенты модели. Для этого широко используются методы анализа остатков модели. В первую очередь вы-

ясняют, достаточно ли малы ошибки и удовлетворяют ли они общим требованиям, предъявляемым к модели. Затем исследуется вопрос о необходимости различных преобразований при наличии особых случаев.

Анализ регрессионных уравнений полезен в нескольких ситуациях: 1) когда возможны грубые ошибки наблюдений, выбросы в зависимой или независимых переменных, ошибки при записи

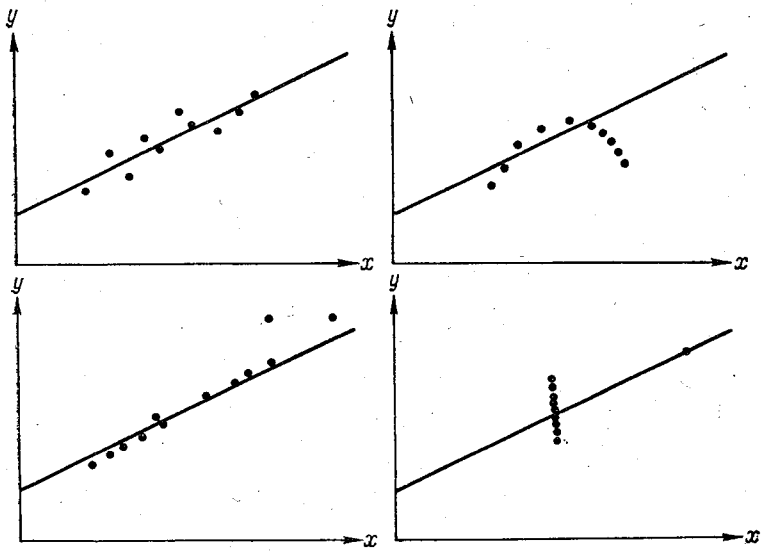


Рис. 12.1. Возможные наборы данных, описываемые одинаковыми уравнениями простой регрессии.

данных на машинные носители информации; 2) когда линейная форма модели может быть непригодна для описания данных; 3) когда необходимо изменить масштаб координатных осей или как-то преобразовать переменные; 4) когда основные гипотезы регрессионного анализа не выполняются и оценки параметров модели определены плохо.

Практика показывает, что наряду с естественными ошибками в данных наблюдений очень много ошибок бывает из-за неверного переноса данных на машинные носители. Такие ошибки могут быть замаскированы и с трудом улавливаются при простом просмотре.

Полезность исследования роли каждого случая в регрессионном анализе проиллюстрируем простыми модельными примерами. На рис. 12.1 показаны четыре простые линейные модели, имеющие одинаковые коэффициенты уравнения, дисперсию ошибок и коэффициенты корреляции. Однако видно сразу, что описываемые ими зависимости резко отличаются друг от друга. Это говорит о необходимости исследования остатков (ошибок) уравнений,

т. е. разности между наблюдаемым и предсказанным значением зависимой переменной в построенной регрессионной модели. Это позволит отыскать резкие отклонения, выбросы, и другие аномалии в данных, препятствующие получению «хороших» параметров модели.

12.1. АНАЛИЗ РЕГРЕССИОННЫХ ОСТАТКОВ

Исходя из предположения, что зависимая переменная подчиняется нормальному закону распределения, считают, что и

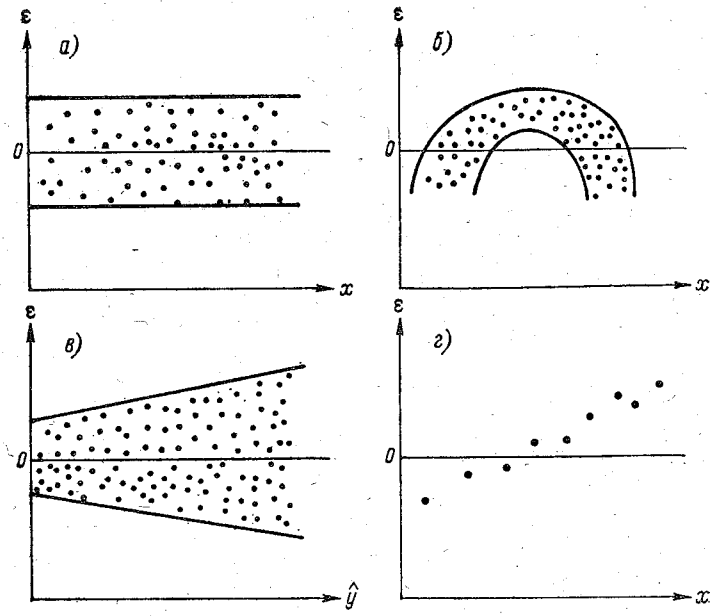


Рис. 12.2. Виды регрессионных остатков (ε).

а) при постоянной дисперсии, *б*) без учета нелинейности модели, *в*) при увеличении дисперсии, *г*) без учета линейного тренда.

ошибка линейного регрессионного уравнения также удовлетворяет этому закону. Если при этом все данные наблюдений взаимонезависимы и имеют постоянную дисперсию, то графическая связь остатков ε_i с предсказанным значением должна представлять приблизительно горизонтальную полосу одинаковой ширины на всем ее протяжении (рис. 12.2 *а*). Если график остатков не является таковым, то, вероятно, необходимо пересмотреть структуру модели, ввести новые переменные. Некоторые возможные варианты графиков представлены на рис. 12.2. При анализе зависимости строят графики зависимости остатков от предсказанного значения, от отдельных предикторов и даже от переменных, не вошедших в модель. Такие графики позволяют выявлять наличие тренда, выбросов, непостоянство дисперсии ошибок уравнения и т. д.

Согласно [25], график остатков уравнения, построенный в зависимости от переменной, имеющей параболическую форму связи с предиктантом, будет иметь общий вид, близкий к параболе.

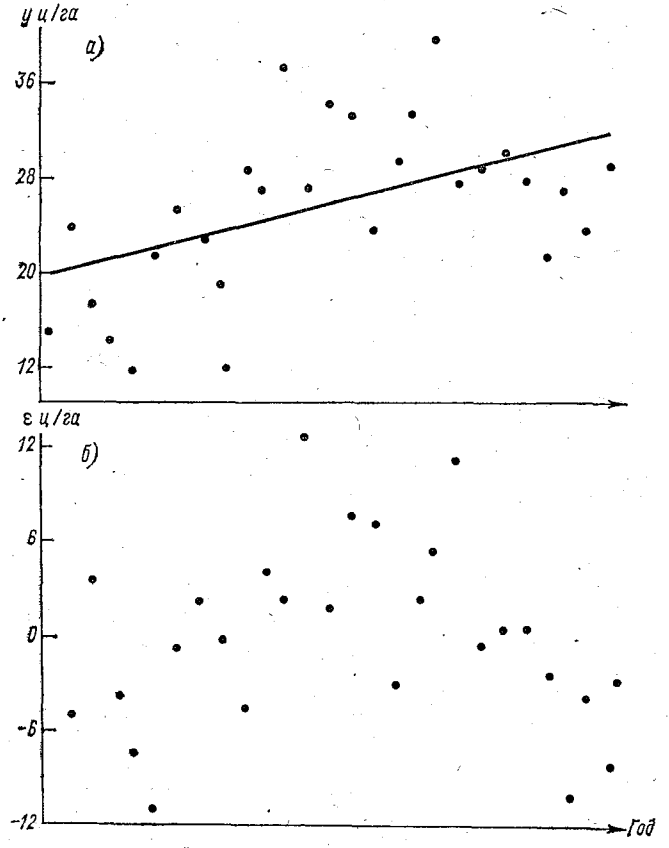


Рис. 12.3. Линейный тренд урожайности y ($а$) и регрессионные остатки ε временно́го ряда урожайности ($б$) озимой пшеницы в Кировоградской области.

На рис. 12.3 $а$ показан тренд урожайности озимой пшеницы в Кировоградской области, аппроксимированный простым линейным уравнением. Соответствующие ему ошибки приведены на рис. 12.3 $б$. Ясно видно, что график ошибок имеет нелинейную форму, как на рис. 12.2 $б$. Это свидетельствует о том, что построенная линейная модель не описывает имеющуюся нелинейность и модель надо изменить. Была построена параболическая модель, в которой ошибки ложатся приблизительно горизонтальной полосой, как на рис. 12.2 $а$, что указывает на адекватность модели на рис. 12.4 $а$.

Существует несколько статистических тестов для выявления нестабильности дисперсии ошибки уравнения. Однако наиболее просто это можно сделать с помощью графика зависимости регрессионных остатков от предсказанного значения $\hat{y}$ (см. рис. 12.2 *в*).

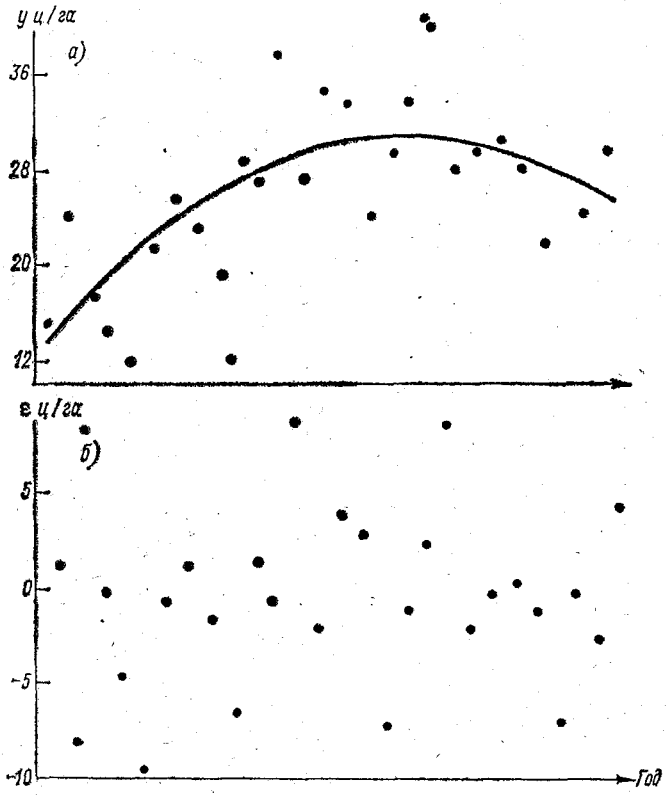


Рис. 12.4. Параболический тренд урожайности y (*а*) и регрессионные остатки ε временно́го ряда урожайности (*б*) озимой пшеницы в Кировоградской области.

При наличии нестабильности дисперсии применяют взвешенный метод наименьших квадратов или преобразуют зависимую переменную, например, $\ln y$, $y^{1/2}$ и y^{-1}. Примером нестабильности дисперсии может служить ситуация, показанная на рис. 11.6 *а*. Анализ временны́х рядов урожайности показал, что в большинстве случаев значительный рост дисперсии урожайности отмечается одновременно с повышением ее уровня.

Графики зависимости остатков от каждого предиктора, включенного в уравнение, помогают выбрать правильную функциональную форму модели и указывают на возможную необходимость включения в уравнение дополнительных членов. Из теории сле-

дует, что коэффициент корреляции между регрессионными остатками и переменными равен нулю, поэтому в такой зависимости не должно быть никакого тренда. На графике остатки должны попадать в горизонтальную полосу, если нет никаких особенностей в данных и не нарушаются основные гипотезы регрессионного анализа. Если графики будут такими, как на рис. 12.2 *б*, *в*, то это указывает на нестабильность дисперсии ошибки и требуется пре-

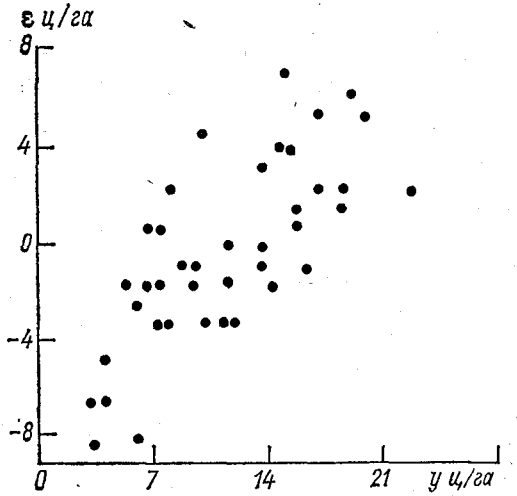


Рис. 12.5. Выделение временно́го тренда урожайности (y) всех зерновых культур в Башкирии с помощью анализа регрессионных остатков (ε).

образование типа $y/\sqrt{x_j}$ или $y\sqrt{x_j}$. Присутствие нелинейного тренда в регрессионных ошибках свидетельствует о необходимости соответствующего изменения независимости переменной или об отсутствии в модели существенного фактора.

Графики остатков позволяют выявлять временной тренд предиктанта. Если такая детерминированная составляющая имеет место, то, выстраивая регрессионные остатки в соответствии с временно́й последовательностью, ее можно обнаружить.

Было построено регрессионное уравнение для прогноза средней областной урожайности всех зерновых культур в Башкирской АССР. Пять введенных предикторов обеспечили множественный коэффициент корреляции уравнения — 0,71. Регрессионные остатки, выстроенные во временной последовательности (рис. 12.5), свидетельствуют о наличии тренда урожайности яровой пшеницы и необходимости его учета в регрессионной модели.

Еще одним полезным видом графиков для выявления формы регрессионной модели является так называемый график частич-

ных остатков. Частичные остатки можно рассчитать по выражению

$$\varepsilon^* = y - (\hat{y} - \beta_j x_j)$$

или

$$\varepsilon_i^* = y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_{j-1} x_{i,j-1} - \beta_{j+1} x_{i,j+1} - \ldots - \beta_{p-1} x_{i,p-1} \quad (i = 1, 2, \ldots, n).$$

Это ошибки уравнения, из которого исключена переменная x_j.

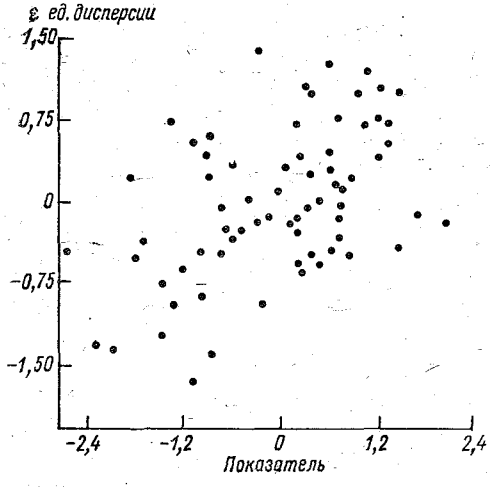


Рис. 12.6. Зависимость частичных регрессионных остатков (ε^*) от показателя влагообеспеченности июля в Северном Казахстане.

График зависимости ε^* от x_j позволяет выявить зависимость y от x_j после удаления влияния всех остальных предикторов, включенных в модель. Проще говоря, это зависимость остатков уравнения от новых переменных. Такие графики полезны как при выявлении формы зависимости, так и при определении необходимости включения дополнительных переменных в модель индикации наличия выбросов. Одним полезным свойством частичных остатков является тот факт, что угол наклона графика (при стандартизованных переменных) равен β_j — коэффициенту при x_j в модели, где эта переменная не удалена:

$$\varepsilon_i^* = \beta_j x_{ij}.$$

На рис. 12.6 показан график частичных остатков в зависимости от показателя условий вегетации в областях Северного Казахстана (Кустанайской, Кокчетавской, Северо-Казахстанской) в июле после устранения влияния уже включенных в регрессионную модель осадков за холодный период и май, а также показателей условий роста и развития посевов всех зерновых культур в июне. График показывает четко выраженную линейную связь, даже без включения последнего фактора в модель коэффициент

множественной корреляции равен 0,75. Для улучшения модели в нее необходимо включить показатель условий вегетации июля.

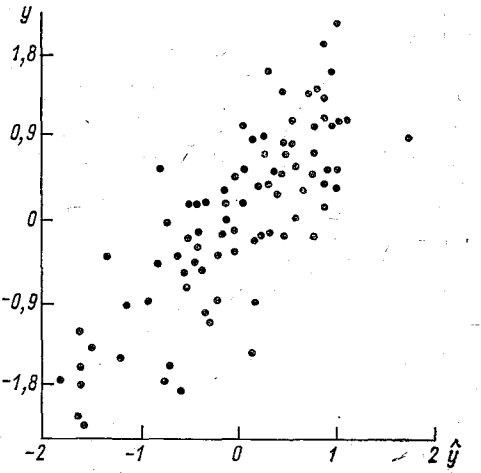


Рис. 12.7. Зависимость истинного значения предиктанта y от предсказанного $\hat{y}$ по регрессионному уравнению в Южном Казахстане.

При анализе окончательно подобранной модели необходимо убедиться, хорошо ли $\hat{y}$ приближает y и есть ли необходимость подвергать y различным трансформациям, например, Кокса-Бокса.

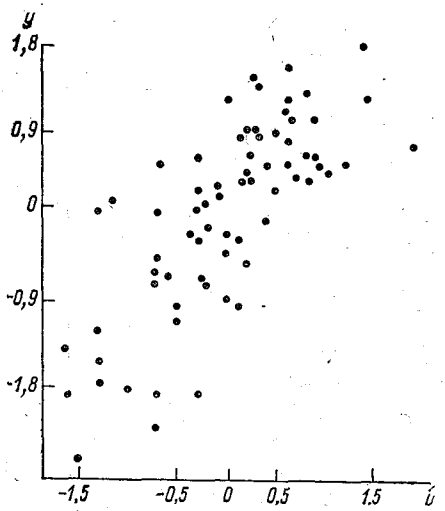


Рис. 12.8. Зависимость истинного значения предиктанта y от предсказанного $\hat{y}$ по регрессионному уравнению в Северном Казахстане.

На рис. 12.7 представлена зависимость фактической урожайности y от предсказанной по модели всех зерновых культур в Южном Казахстане. Здесь связь линейная, дополнительных преобразований y делать не надо. На рис. 12.8 показана аналогичная

связь для областей Северного Казахстана, где отмечается нелинейность и желательна трансформация предиктанта.

При построении многофакторных регрессионных уравнений, отражающих связь агрометеорологических условий и урожайности сельскохозяйственных культур, использование графического анализа остатков является наиболее простым и достаточно информативным способом проверки адекватности построенной модели. Довольно часто можно ограничиться их качественным анализом без применения статистических критериев. Такой подход в последние годы получил широкое распространение в связи с возможностью построения разнообразных графиков с помощью ЭВМ.

12.2. АНАЛИЗ ВЫБРОСОВ

В последнее время при построении статистических моделей и интерпретации данных все больше внимания уделяется «плохим», «отскакивающим», не укладывающимся в заданную структуру данным наблюдений, или, как их часто называют, выбросам. Что делать в этом случае? Одни исследователи предпочитают не «обесценивать» выборку отбрасыванием таких данных, другие, напротив, считают необходимым их удалять из исследуемой статистической совокупности при первых же подозрениях. Понятие выброса в данных исследуется в статистике очень давно. Еще Бернулли указывал на необходимость тщательного рассмотрения подобных явлений в практической деятельности [36].

Недостаточность коэффициента корреляции R для отражения свойств регрессионного уравнения, который не всегда точно представляет степень взаимозависимости факторов, не отражает имеющиеся в данных грубые ошибки и несоответствия, привела к появлению большого количества процедур, направленных на выявление подобным образом выделяющихся данных. К ним можно отнести как анализ графиков регрессионных остатков, так и анализ выбросов [5].

Помимо анализа остатков был предложен подход к выделению имеющихся в регрессионных остатках выбросов с помощью различных показателей.

Чтобы проверить влияние таких выделяющихся из всех данных точек пространств переменных на статическую модель, можно после их удаления построить модель еще раз и рассмотреть, как изменяются ее параметры и прогностические возможности. Если использовать для выявления выбросов, особенно при большом числе данных, одновременно две статистики — ε_i и $V(\varepsilon_i)$ — то задача становится достаточно непростой.

Для преодоления этой трудности в [39] предложена легко интерпретируемая мера аномальности наблюдений регрессионной модели, которая является функцией от обеих статистик: ε_i и $V(\varepsilon_i)$.

Для обычной регрессии (см. гл. 9) вектор разности записывается в виде

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix} = y - \hat{y} = y - X\hat{\beta} = (I - X)(X'X)^{-1}X'y.$$

Ковариационную матрицу вектора разности ε и отклика y можно записать так:

$$V(\varepsilon) = (I - X)(X'X)^{-1}X'\sigma^2,$$

$$V(\hat{y}) = X(X'X)^{-1}X'\sigma^2.$$

По теории, в основе которой лежит положение о нормальности распределения переменных, $(1-\alpha)\cdot 100\,\%$-ный доверительный эллипсоид для β^* (оценки вектора коэффициентов β) должен удовлетворять неравенству

$$\frac{(\beta^* - \hat{\beta})' X'X (\beta^* - \hat{\beta})}{p\sigma^2} \leqslant F(p;\ n-p;\ 1-\alpha),$$

где $\sigma^2 = \varepsilon'\varepsilon/(n-p)$, и $F(p;\ n-p;\ 1-\alpha)$ имеет центральное F-распределение со степенями свободы p и $n-p$.

Вполне разумно для определения степени влияния i-й точки данных на β вычислить оценки β после удаления этой точки. Если обозначить оценки коэффициентов наименьших квадратов $\hat{\beta}_i$ при условии, что в выборке нет i-й точки, то удобно выражать расстояние между различными оценками β в виде (12.1). Данная формула была предложена Куком в качестве меры аномальности одного случая выборки:

$$D_i = \frac{(\hat{\beta}_i - \hat{\beta})' X'X (\hat{\beta}_i - \hat{\beta})}{p\sigma^2} \quad (i = 1,\ 2,\ \ldots,\ n). \qquad (12.1)$$

Эта мера расстояния между $\hat{\beta}_i$ и $\hat{\beta}$ — расстояние Кука (D_i) — обеспечивает вероятностную интерпретацию расхождения оценок при определенном заданном уровне значимости. Например, если положить, что D_i распределено приблизительно как $F(p;\ h-p;\ 0{,}5)$, то удаление i-й точки приводит к смещению оценки, полученной методом наименьших квадратов, на границу 50 %-ной доверительной области, если за оценку β принять $\hat{\beta}$.

Если в модель входят только две переменные, можно нанести на двумерный график значения рассчитанных переменных и определить влияние каждого случая на коэффициент. Величина D_i принимает только положительные значения. Большое значение

указывает на сильно влияющий случай. В удобной для расчетов форме расстояние Кука можно записать так:

$$D_i = \frac{(\hat{y}_i - y)'(\hat{y}_i - y)}{p\sigma^2} \quad (i = 1, 2, \ldots, n).$$

В такой интерпретации D_i измеряет изменения прогностического значения $\hat{y}$ при отбрасывании одной пары наблюдений.

Можно рассчитать D_i по формуле (12.2), куда наряду с ошибкой в i-м независимом случае (ε_i^*), входит полезная статистика — расстояние Махаланобиса (M_i):

$$D_i = \frac{(\varepsilon_i^*)^2 \left[n^{-1} + M_i (n-1)\right] n}{p \cdot \text{СКО}}. \qquad (12.2)$$

Расстояние Махаланобиса показывает, насколько каждый случай или точка в p-мерном пространстве независимых переменных отклоняется от центра статистической совокупности. Оно рассчитывается по формуле

$$M_j = (\mathbf{x}_i - \bar{\mathbf{x}})' (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

и широко используется для обнаружения выбросов при анализе многомерных нормально распределенных переменных, в дискриминантном анализе. Здесь величина $(\mathbf{X}'\mathbf{X})^{-1}$ является корреляционной матрицей, $\bar{\mathbf{x}}$ — вектором средних значений.

Существуют различные модификации статистики расстояния Кука [39].

Использование статистики — расстояния Кука позволяет выявлять скрытые от исследователя и не всегда выраженные в больших разностях выбросы, которые могут быть следствием имеющихся ошибок и «засорения» как независимых, так и зависимых переменных. Например, нередко при прогнозировании урожая с определенной заблаговременностью аномальные погодные условия неучитываемого отрезка вегетационного периода могут сильно изменить состояние посевов сельскохозяйственных культур. Возникает несоответствие между урожаем и теми агрометеорологическими условиями, которые наблюдались до момента составления прогноза. Возможны и такие случаи, когда, несмотря на небольшую ошибку при прогнозе методом наименьших квадратов, метеорологические условия соответствующего года имеют очень малую вероятность появления, и их вопреки незначительности ошибки прогноза необходимо также считать непринадлежащими рассматриваемой статистической совокупности.

Появляется возможность в дополнение к робастной регрессии, ориентированной на отклонения от принятых гипотез в зависимой переменной, исключить грубые несоответствия некоторых наборов данных построенной регрессионной модели. При этом совсем необязательно, чтобы исключенный набор содержал большие ошибки и «засорение».

При обычно принятой методике построения агрометеорологических регрессионных уравнений считается, что механизм, генерирующий связь между зависимой и независимой переменными, может быть описан одним уравнением. Это, естественно, грубое упрощение, что подтверждается наличием резко выделяющихся значений статистики Кука.

После удаления данных, которые плохо укладываются в построенную модель и вследствие этого оказывают сильное влияние на коэффициенты уравнения, можно получить модель с лучшими прогностическими свойствами.

Статистика Кука в большой степени зависит от выбора подмножества предикторов, при изменении которых меняется и сама статистика. Подмножество предикторов, составленное из таких показателей, как количество декадных осадков и температура воздуха, при построении регрессионных уравнений дает бо́льшие значения статистики, чем подмножества, включающие трансформированные суммы осадков со специально уменьшенной дисперсией и ограниченной вариацией. Таким образом D_i зависит от изменения масштаба переменных.

Ниже приведены данные, показывающие уменьшение средней квадратической ошибки прогноза регрессионного уравнения, построенного после удаления выбросов по критерию Кука, в сравнении с уравнением, использующим все данные:

Район	Западный Казахстан	Восточный Казахстан	Северный Казахстан	Центральный Казахстан	Южный Казахстан
Уменьшение ошибки, % . . .	42,9	9,4	49,6	5,0	0

Длина проверочной выборки составляет от 13 до 21 случая для разных районов.

В четырех из пяти районов ошибка уменьшилась на пять и более процентов. Для двух прогностических регрессионных уравнений урожайности всех зерновых культур, построенных соответственно по данным Уральской и Актюбинской областей и Карагандинской, Павлодарской и Тургайской областей, улучшение прогностических возможностей оказалось очень большим (42,9 и 49,6 %). В первом наборе было удалено всего два случая, во втором — семь.

Из (12.2) видно, что статистика Кука зависит мультипликативно от двух факторов: 1) от ошибки прогноза по i-й точке, когда она является независимой и уравнение построено по оставшимся данным, 2) от величины отклонений условий, описываемых независимыми переменными, от средних. В большинстве случаев при использовании трансформированных переменных значения D_i, как уже упоминалось, не так велики, как при использовании обычных данных об осадках и температуре воздуха; «сильновлияющим» данным соответствуют значения D_i (при обычно принятых

уровнях значимости), не являющиеся выбросами из общей совокупности статистик D_i.

В качестве предикторов в моделях были использованы переменные, отобранные после всестороннего анализа с использованием критерия Мэллоуса (C_p).

На рис. 12.9 представлены значения статистики Кука, расстояния Махаланобиса и абсолютные ошибки уравнений для данных,

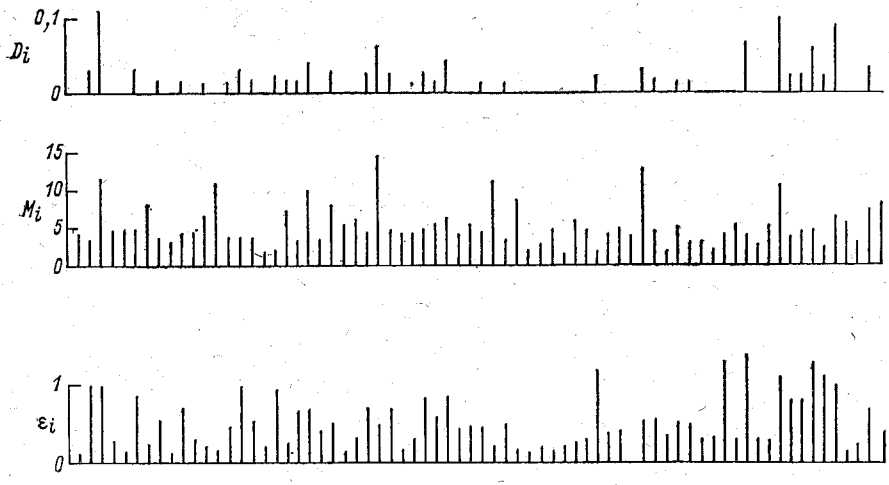


Рис. 12.9. Значения статистик: расстояние Кука (D_i), расстояние Махаланобиса (M_i), регрессионные остатки (ε_i) уравнения для прогноза урожайности зерновых культур в Центральном Казахстане (конец июля).

используемых при построении модели прогноза в конце июля в районе, включающем Целиноградскую, Кустанайскую, Павлодарскую и Тургайскую области. Два максимальных значения D_i, как видно из рисунка, не совпадают с максимальными значениями ε_i и M_i.

Из рисунков следует, что необязательно большая ошибка прогноза соответствует сильно отличающимся от среднего условиям вегетационного периода (мерой чему в многомерном случае служит расстояние Махаланобиса). Например, максимальному такому значению, содержащемуся в представленной выборке, — 14, 45 — соответствует ошибка прогноза в единицах дисперсии предиктанта 0,63. И, напротив, случаям, лежащим близко к центру факторного пространства, часто соответствует большая ошибка прогноза (например, $M=1{,}62$, $\varepsilon=-1{,}03$). Поэтому такие простые и распространенные приемы, как отбрасывание аномальных по агрометеорологическим условиям вегетационного периода лет, содержащихся в выборке, не всегда оправданы.

Первый из удаленных наборов соответствует 1962 г. в Целиноградской области. Основная причина несоответствия этих данных общему регрессионному уравнению состоит в том, что имеет место маловероятное, не нашедшее адекватного отражения в ста-

тистической регрессионной модели сочетание агрометеорологических условий разных частей вегетационного периода. Так, количество осадков за холодный период превышало в этот год норму на 13 %, что обеспечило хорошие запасы влаги в метровом слое почвы на весну. Во время массового сева — третьей декады мая — пахотный слой был отлично увлажнен обильными осадками — 32 мм, что составляет почти три нормы. В первой декаде июня осадков было немного — 8 мм (норма 12 мм) на фоне повы-

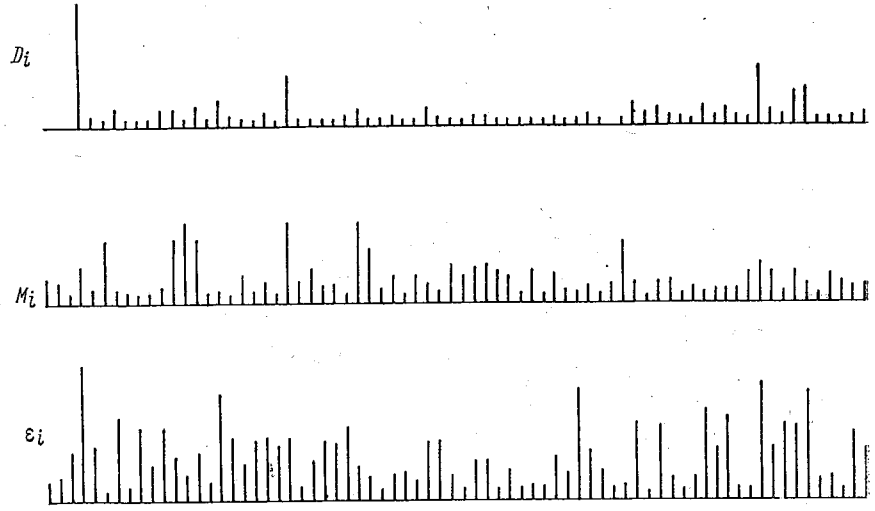


Рис. 12.10. Значение статистик: расстояние Кука (D_i), расстояние Махаланобиса (M_i), регрессионные остатки (ε_i) уравнения для прогноза урожайности зерновых культур в Центральном Казахстане (конец июня).

шенной температуры воздуха. Во второй декаде прошли обильные дожди (три нормы) при благоприятном пониженном уровне температуры (17 °С). В третьей декаде осадки не были значительны — 9 мм, но и температурный фон оставался ниже нормы — 17,5 °С, что создавало благоприятные условия для развития всех зерновых культур. Резкая смена метеорологических условий произошла в июле. Наступила поздняя летняя засуха. За весь месяц выпало всего 20 мм осадков (при норме 57 мм). Во вторую декаду июля дождей почти не было (всего 2 мм осадков). Температурный фон был высоким и в течение всех декад превышал норму на 1,8 °С и более. Колошение и цветение пшеницы и ячменя проходило в крайне неблагоприятных условиях. Большое значение имел и тот факт, что после очень хорошего периода влагообеспеченности растения попали в неблагоприятные метеорологические условия и не смогли адаптироваться к быстрой смене типа погоды. Если бы контраст погоды не был столь велик и растения развивались при средних метеорологических условиях в первую

половину вегетационного периода, то они смогли бы с бо́льшим успехом противостоять засушливым явлениям июля.

Еще большее значение имеет статистика Кука за этот год в уравнении, построенном по метеорологическим условиям веге-

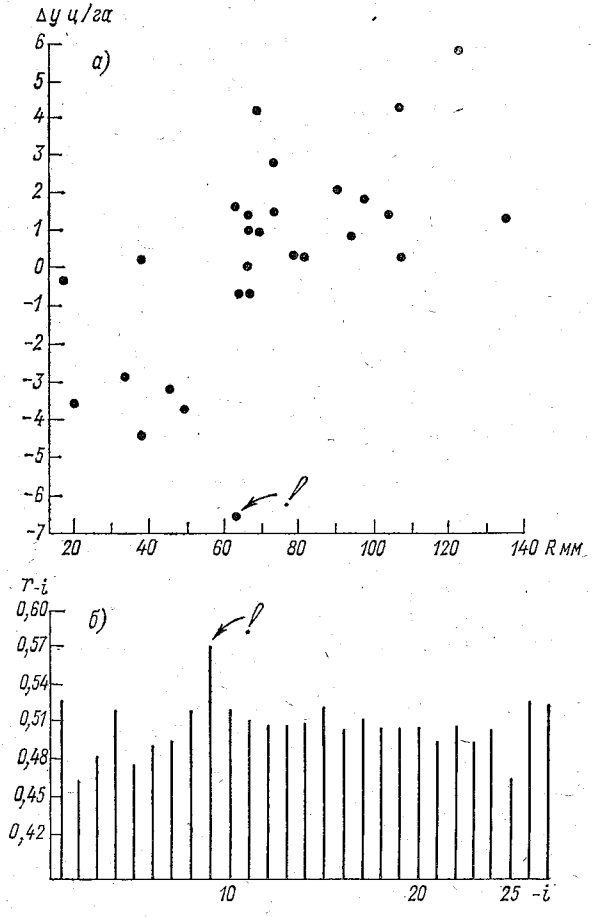


Рис. 12.11. Зависимость урожайности (y) семян подсолнечника от количества осадков (R) в июле ($а$) и коэффициент корреляции (r_i) между данными величинами при отбрасывании —i года ($б$).

тационного периода на конец июня без учета погоды июля, что видно из рис. 12.10. По прогнозу этого уравнения урожай ожидался предельно высоким. В действительности он оказался ниже нормы. В уравнении, в котором погода июля учитывается, прогноз урожайности продолжает оставаться выше нормы, хотя и не такой большой, как в первом прогнозе. Имевшая место очень

резкая смена агрометеорологических условий в середине вегетационного периода плохо укладывается в построенное регрессионное уравнение. Подобное событие нетипично для этого географического района, имеет малую вероятность осуществления, поэтому удаление этого набора данных из статистической совокупности позволяет получить коэффициенты регрессионной модели, близкие к истинным значениям.

Для анализа сильно влияющих данных можно использовать парный коэффициент корреляции, при расчете которого также поочередно отбрасывается одна точка данных. Если имеем n пар наблюдений (x_i, y_i), то рассчитываем n коэффициентов парной корреляции r_i. Теперь можно легко оценить, какая пара данных сильнее влияет на показатель корреляции.

На рис. 12.11 показана графическая связь урожайности семян подсолнечника в Кировоградской области с суммой осадков за июль. Из рисунка видно, что отмеченная «точка» лежит очень далеко от построенной прямой линии, аппроксимирующей эту зависимость. Даже не строя график, наличие такой аномальной точки можно определить по значениям r_i: 0,52, 0,46, 0,48, 0,52, 0,48, 0,49, 0,49, 0,51, 0,57, 0,51, 0,51, 0,51, 0,51, 0,51, 0,52, 0,51, 0,52, 0,51, 0,51, 0,51, 0,49, 051, 0,49, 0,51, 0,47, 0,53, 0,52.

Нетрудно рассчитать такой полезный показатель и в случае многофакторной регрессионной связи.

Совместное применение статистик, помогающих исследователю глубже понять структуру данных, наряду с новыми методами регрессионного анализа, ориентированными на учет их специфических свойств, позволит разрабатывать модели с лучшими прогностическими характеристиками.

Список литературы

1. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. 2.: Мир, 1982. 486 с.

2. Груза Г. В., Ранькова Э. Я. Вероятностные метеорологические прогнозы. Л.: Гидрометеоиздат, 1983. 271 с.

3. Денисов П. П. Методика оценки тенденций в ходе речного стока// Метеорология и гидрология. 1975. № 4. С. 101—104.

4. Джефферс Дж. Введение в системный анализ: применение в экологии. М.: Мир, 1981. 252 с.

5. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Кн. 1, 2. М.: Финансы и статистика, 1986.

6. Желтая Н. Н. Методическое пособие по составлению долгосрочных агрометеорологических прогнозов средней областной урожайности ярового ячменя в северных и западных областях Казахстана. М.: Гидрометцентр СССР, 1978. 12 с.

7. Забелин В. Н. Определение динамики урожайности зерновых культур при агрометеорологическом прогнозировании//Метеорология и гидрология. 1982. № 10. С. 103—109.

8. Забелин В. Н. Прогнозирование урожаев по взаимнокоррелированным метеорологическим факторам методом гребневой регрессии//Метеорология и гидрология. 1983. № 8. С. 104—107.

9. Закс Л. Статистическое оценивание. М.: Статистика, 1976. 776 с.

10. Ивахненко А. Г., Зайченко Ю. П., Дмитров В. Д. Принятие решений на основе самоорганизации. М.: Советское радио, 1976. 280 с.

11. Кириличева К. В. Методическое пособие по составлению агрометеорологического прогноза средней областной урожайности яровой пшеницы в основной зоне ее возделывания. М.: Гидрометцентр СССР, 1980. 25 с.

12. Классификация и кластер/Под ред. Райзин Дж. В. М.: Мир, 1980. 389 с.

13. Ликеш И., Ляга И. Основные таблицы математической статистики. М.: Финансы и статистика, 1985. 356 с.

14. Лубнин М. Г. Влияние агрометеорологических условий на работу сельскохозяйственных машин и орудий. Л.: Гидрометеоиздат, 1983. 117 с.

15. Лукомский Я. И. Теория корреляции и ее применение к анализу производства. М.: Госстатиздат, 1961. 273 с.

16. Масловская А. Д. Метод долгосрочного агрометеорологического прогноза среднереспубликанского урожая озимой пшеницы в Казахстане// Тр. КазНИГМИ. 1974. Вып. 47. С. 117—130.

17. Математические методы в агрометеорологии//Тр. ИЭМ. 1973. Вып. 3. (40). 152 с.

18. Менжулин Г. В., Николаев М. В., Савватеев С. П. Оценки экономической и погодной составляющих изменений урожайности зерновых культур//Тр. ГГИ. 1983. Вып. 280. С. 111—119.

19. Митропольский А. К. Техника статистических вычислений. М.: Физматгиз, 1961. 479 с.

20. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Вып. 2. М.: Финансы и статистика, 1982. 316 с.

21. Немчинов В. С. Сельскохозяйственная статистика с основами общей теории. М.: Сельхозгиз, 1946. 345 с.

22. Обухов В. М. Урожайность и метеорологические факторы. М.: Госпланиздат, 1949. 316 с.

23. Программное обеспечение ЭВМ. Минск: Ин-т математики АН БССР, 1983. Вып. 35. 130 с.

24. Продовольственная программа СССР на период до 1990 года и меры по ее реализации. Материалы майского пленума ЦК КПСС 1982. М.: Политиздат, 1982. 111 с.
25. Себер Дж. Линейный регрессионный анализ. М.: Мир, 1980. 456 с.
26. Сиротенко О. Д. Компонентный анализ в прогностических задачах агрометеорологии. Методическое письмо. М.: Гидрометеоиздат, 1971. 52 с.
27. Сиротенко О. Д., Удодова А. Ф., Горбачев В. А. Объективный анализ полей сумм осадков как задача квадратического программирования//Тр. ИЭМ. 1973. Вып. 3(40). С. 90—102.
28. Тьюки Дж. Анализ результатов наблюдений. М.: Мир, 1981. 693 с.
29. Уланова Е. С. Применение математической статистики в агрометеорологии для нахождения уравнений связей. М.: Гидрометеоиздат, 1964. 112 с.
30. Уланова Е. С., Сиротенко О. Д. Методы статистического анализа в агрометеорологии. Л.: Гидрометеоиздат, 1968. 198 с.
31. Уланова Е. С. Агрометеорологические условия и урожайность озимой пшеницы. Л.: Гидрометеоиздат, 1975. 302 с.
32. Фишер Р. А. Статистические методы для исследователей. М.: Госстатиздат, 1958. 248 с.
33. Френкель А. А. Математические методы анализа динамики и производительности труда. М.: Экономика, 1972. 145 с.
34. Химмельблау Д. Анализ процессов статистическими методами. М.: Мир, 1973. 973 с.
35. Хьюбер П. Робастность в статистике. М.: Мир, 1984. 303 с.
36. Bernoulli D. The most probable choice between several discrepant and the formation of the most likely induction//In C. G. Allen (1961) Biometrica. 1961. V. 48. P. 3—13.
37. Box G. E. P. Non-normality and test on variances//Biometrica. 1954. V. 10. P. 318—335.
38. Box G. E. P., Watson G. S. Robustness to non-normality of regression//Biometrica. 1962. V. 49. P. 93—106
39 Cook R. D. Detection of influental observation in linear regression// Technometrics. 1977. V. 19. P. 15—18.
40. Daniel C., Wood F. S. Fitting equations to data: computer analysis of multifactor. N. Y.: Wiley, 1980. 458 p.
41. Denby L., Mallows C. L. Two diagnostic displays for robust analysis// Technometrics. 1977. V. 19. P. 1—13.
42. Furneval G. M. Regression by leaps and bounds//Technometrics. 1974. V. 16. P. 499—511.
43. Haun J. R. Matematical models in agrometeorology. 1983. WMO. CAgM rep. 14.
44. Hawkins D. T. Identification of outliers. London, 1980. 401 p.
45. Hemmerle W. J. An explicit solution for generalized ridge regression// Technometrics. 1975. V. 17. P. 309—314.
46. Huber P. J. Robust estimation of a location parameter//Ann. Math. Statist. 1964. V. 35. P. 73—101.
47. Huber P. J. Robust-statistics: a reviev//Ann. Math. Statist. 1972. V. 43. P. 1041—1067.
48. Huber P. J. Robust regression: asymptotics conjectures and Monte-Carlo// Ann. Math. Statist. 1973. V. 1. p. 799—821.
49. Marquardt D. W. Generalized inverses, ridge regression, biased linear estimation and non-linear estimation//Technometrics V. 12. P. 591—612.
50. Walls R. E., Weeks D. L. A note on variance of a predicted response in regression//The Amer. Statist. 1969. V. 23. P. 24—26.

Предметный указатель

СОДЕРЖАНИЕ

Глава 11. Альтернативные регрессионные модели

Глава 12. Анализ регрессионных остатков и выбросов

Монография

Евгения Станиславовна Уланова, Владимир Николаевич Забелин

Методы корреляционного и регрессионного анализа в агрометеорологии

Редактор О. О. Штанникова. Художник С. В. Богородский.
Технический редактор Н. И. Перлович. Корректор Л. Б. Емельянова.

ИБ № 1797

Сдано в набор 17.10.89. Подписано в печать 22.02.90. М-19521. Формат 60×90¹/₁₆, бумага книжная. Гарнитура литературная. Печать высокая. Печ. л. 13. Кр.-отт. 13. Уч.-изд. л. 13,21. Тираж 1770 экз. Индекс ПРЛ-60. Заказ № 255. Цена 2 руб. 30 коп. Гидрометеоиздат. 199226. Ленинград, Беринга, 38

Ленинградская типография № 4 ордена Трудового Красного Знамени Ленинградского объединения «Техническая книга» им. Евгении Соколовой Государственного комитета СССР по печати. 190000, Ленинград, Прачечный переулок, 6.