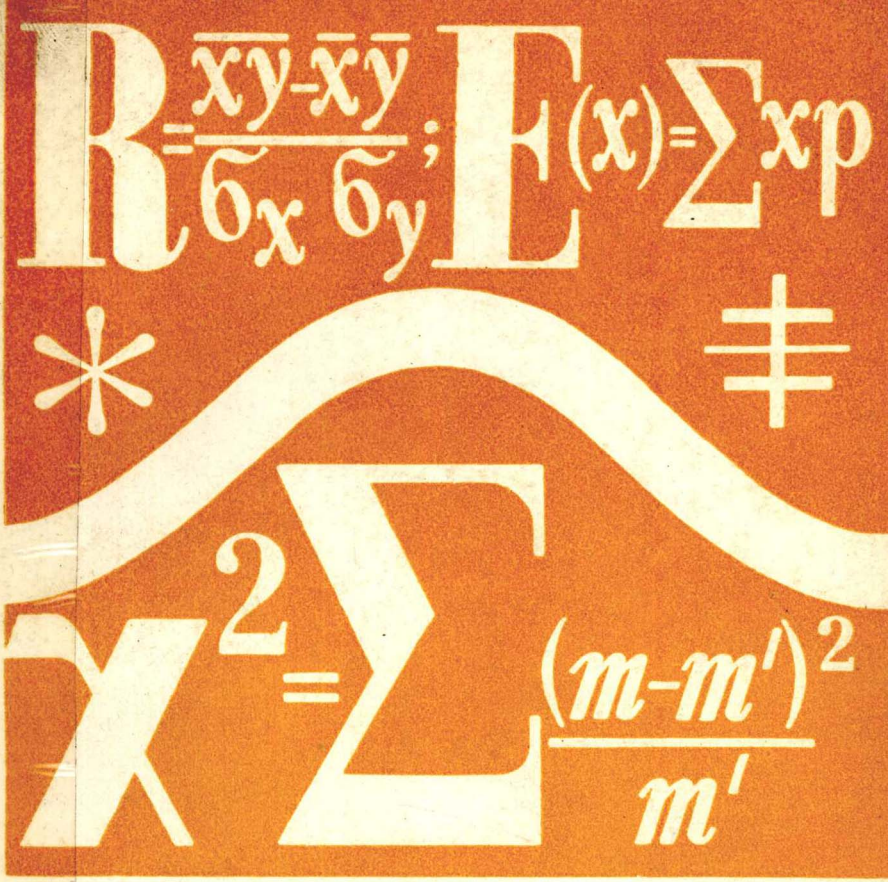
МАТЕМАТИЧЕСКАЯ СТАТИСТИКА ДЛЯ ЭКОНОМИСТОВ



О. П. Крастинь

ИЗУЧЕНИЕ СТАТИСТИЧЕСКИХ ЗАВИСИМОСТЕЙ ПО МНОГОЛЕТНИМ ДАННЫМ

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА ДЛЯ ЭКОНОМИСТОВ

О. П. Крастинь

ИЗУЧЕНИЕ СТАТИСТИЧЕСКИХ ЗАВИСИМОСТЕЙ ПО МНОГОЛЕТНИМ ДАННЫМ

МОСКВА «ФИНАНСЫ И СТАТИСТИКА» 1981

ББК 65.051
К78

Крастинь О. П.

К78 Изучение статистических зависимостей по многолетним данным. — М.: Финансы и статистика, 1981. — 136 с., ил. — (Мат. статистика для экономистов).

1 р.

Рассматриваются различные методы обработки многолетних данных «осреднение данных», «заводо-лет», ковариационный анализ и др. Даются их сравнительная оценка и методология для изучения статистических зависимостей. Показано их практическое применение.

Для статистиков, экономистов и специалистов, занимающихся применением современных математических методов в экономике.

К $\frac{10805-140}{010(01)-81}$ 23—81(С) 1702060000 ББК 65.051
31

ВВЕДЕНИЕ

Математические методы применяются для совершенствования экономического анализа, прогнозирования и планирования почти во всех отраслях народного хозяйства и на всех уровнях управления. Среди совокупности различных математических методов, которые используются в экономической работе, большое значение имеют методы математической статистики. Методы математической статистики дают возможность охарактеризовать различные свойства сложных массовых процессов и выявить их закономерности и связи. Особую популярность получили методы корреляции и регрессии. Они применяются в практической работе органов ЦСУ СССР.

Уравнения регрессии, характеризующие зависимость результатов хозяйственной работы от основных влияющих на них факторов, часто используются в качестве математико-статистических моделей при решении конкретных задач анализа и прогнозирования. В связи с этим к уравнениям регрессии предъявляются повышенные требования в отношении их экономического содержания, правильности описания реально существующих связей и особенно в отношении устойчивости во времени. Для улучшения качества определяемых моделей исследователи стараются по возможности увеличить объем обрабатываемых данных. Но возможность увеличения пространственной совокупности ограничена, поскольку образование очень больших совокупностей, как правило, нарушает требование качественной однородности. Кроме того, даже большие пространственные совокупности дают мало информации об устойчивости связей во времени. Поэтому в последние годы многие научные работники и статистики пытаются изучить раз-

личные взаимосвязи по многолетним данным, иными словами, — по данным временно-пространственной совокупности.

Из экономической и математической литературы известно несколько методов и способов получения уравнений регрессии по многолетним данным: предварительное осреднение исходных данных, способ «заводо-лет», ковариационный анализ, осреднение параметров одногодичных уравнений и др. Однако систематического анализа свойств этих методов и их сравнительной оценки нет.

Целью предлагаемой книги является последовательное и систематическое изложение методов изучения статистических зависимостей по многолетним данным. Рассматриваются различные методы и способы обработки многолетних данных с целью получения уравнений регрессии при наличии различных особенностей исходных данных. Особое значение придается сравнительной оценке этих методов.

Первая глава посвящена наиболее простому способу — применению многолетних средних данных. Во второй главе рассмотрен широкоизвестный способ «заводо-лет». В третьей главе даются введение в ковариационный анализ и сравнительная оценка разных способов при изучении парных связей.

Как показало исследование, с точки зрения математической статистики наиболее обоснованным методом изучения статистических связей по многолетним данным является ковариационный анализ. Поэтому ему посвящены главы четвертая (простой ковариационный анализ) и пятая (множественный ковариационный анализ). Но и в этих главах ковариационный анализ рассматривается в сопоставлении с другими способами, в частности со способом «заводо-лет».

Заключительная шестая глава посвящена динамическим рядам уравнений регрессии. В ней рассматривается осреднение параметров одногодичных уравнений регрессии. Этот способ не является самым корректным с точки зрения математической статистики, но зато он очень простой, удобный и всегда дает убедительные результаты, которые поддаются содержательной интерпретации. В главе изложен некоторый опыт автора в области применения регрессионного анализа.

Примеры, приведенные в книге, взяты из области сельского хозяйства. В этой отрасли в настоящее время регрессионный анализ используется наиболее широко и успешно, к этой отрасли относится большинство работ автора. Многолетние данные наиболее часто применяются при изучении сельского хозяйства, поскольку на сельское хозяйство наиболее значительно влияют неуправляемые и меняющиеся погодные условия. Однако рассмотренные в работе методы могут быть использованы и в других отраслях народного хозяйства.

Предполагается, что читатель знаком с основными методами анализа регрессии и корреляции по однородной совокупности. Мы можем рекомендовать также ряд относительно доступных для экономистов работ [4, 12 и др.].

В главах четвертой и пятой использованы некоторые рекомендации А. Е. Бригмане. Автор выражает благодарность А. Е. Бригмане за эти работы.

1. ПРИМЕНЕНИЕ МНОГОЛЕТНИХ СРЕДНИХ ДАННЫХ

1.1. ПОСТАНОВКА ВОПРОСА

В исследованиях статистических закономерностей и связей, особенно в области сельского хозяйства, часто рекомендуется использовать многолетние данные. Результаты хозяйственной деятельности в сельском хозяйстве значительно зависят от случайных благоприятных или неблагоприятных метеорологических условий [11, 15, 24]. В средних многолетних данных влияние этих случайных факторов в значительной степени взаимно погашается, поэтому многолетние средние данные значительно более устойчивы по сравнению с одногодичными [18, с. 359]. В сельскохозяйственной литературе стало почти законом, что любые выводы, сделанные на основе полевых опытов, должны быть обоснованы по крайней мере на базе пятилетних данных. Поэтому не без основания в статистической литературе часто обсуждается вопрос построения уравнений регрессии по многолетним данным.

Например, С. С. Сергеев дает следующую оценку: «Наиболее значительные трудности в анализе множественной корреляции встречаются при использовании отчетных данных совхозов и колхозов, поскольку система факторов здесь очень велика и многообразна... Чтобы получить в таких условиях хотя бы немногие, но более или менее значимые результаты, стремятся ...вместо отдельных лет взять более длительные периоды, позволяющие устранить резкие колебания метеорологических условий и проявить влияние более устойчивых различий в обеспечении удобрениями, основными фондами и т. д.» [15, с. 129].

Двухгодичные данные используют в своем учебнике И. Г. Венецкий, Г. С. Кильдишев [3, с. 160—182].

Одним из методов построения уравнений регрессий по многолетним данным (их несколько) является *предварительное осреднение исходных данных* (результативного и факторных признаков) по каждой единице совокупности в отдельности. Уравнение регрессии определяется по заранее осредненным двухгодичным, трехгодичным или многолетним данным. Применяя этот способ, по каждой единице совокупности вычисляют средние величины результативного и всех факторных признаков. Эти многолетние средние величины рассматриваются как данные обычной пространственной совокупности. Уравнение регрессии и все ее оценки определяются по основным общеизвестным методам без заметных их изменений.

Мы на основе практического опыта достаточно часто обнаруживали, что такой способ часто приводит к трудно объяснимым *статистическим парадоксам.* Последние заключаются в том, что коэффициенты корреляции и регрессии, определенные по средним многолетним данным, часто выходят за пределы вариации тех же коэффициентов, но рассчитанных по одногодичным данным того же периода.

К примеру, покажем результаты вычисления трехфакторных уравнений регрессии, выражающих зависимость урожайности зерновых в центнерах с гектара от качества пашни в баллах, обеспеченности основными фондами сельскохозяйственного назначения в расчете на один гектар культивированной земли в сотнях рублей и количества внесенных удобрений всех видов в пересчете на действующие вещества на один гектар культивированной земли в центнерах в 51 колхозе Земгальской равнины Латвийской ССР [6, с. 92] (**табл.** 1.1).

По данным табл. 1.1 качество пашни, имеющее статистически значимое влияние на урожайность зерновых четыре года из пяти лет, по усредненным данным, свое влияние потеряло. Влияние основных фондов стало выше, чем в любом году из рассматриваемых лет, взятых в отдельности. Влияние удобрений по усредненным данным ближе к параметру 1967 г., когда этот показатель самый высокий. В наиболее резком виде статистический парадокс проявляется при рассмотрении свободных членов уравнений. Подобные статистические парадоксы мы встречали многократно [9].

Таблица 1.1

ПАРАМЕТРЫ МНОЖЕСТВЕННЫХ УРАВНЕНИЙ РЕГРЕССИИ, ВЫРАЖАЮЩИХ ЗАВИСИМОСТЬ УРОЖАЙНОСТИ ЗЕРНОВЫХ ОТ ТРЕХ ВЛИЯЮЩИХ НА НЕЕ ФАКТОРОВ В КОЛХОЗАХ ЗЕМГАЛЬСКОЙ РАВНИНЫ ЛАТВИЙСКОЙ ССР ЗА РЯД ЛЕТ

	1964	1965	1966	1967	1968	1964—1968
Коэффициенты регрессии при факторах:						
качество пашни	0,206	0,226	0,112*	0,181	0,151	—0,071*
основные фонды	1,19	1,33	0,11*	1,43	1,19	1,69
удобрения	3,99	4,10	4,56	5,36	3,91	5,28
Свободный член	—1,80	0,80	2,62	—2,69	4,62	12,36
Коэффициент множественной корреляции	0,729	0,727	0,603	0,694	0,741	0,704

* Коэффициент, к которому не может быть отвергнута нулевая гипотеза с вероятностью 0,95.

Аналогичное явление обнаруживали многие исследователи, например И. К. Сирожиддинов [16].

Другие авторы пришли к противоположным выводам, например М. М. Юзбашев, В. В. Короткова [23]. И. Г. Попов, Э. С. Миронова, ссылаясь на работы других авторов и на собственные исследования, считают такой способ полезным и пригодным для прогноза: «В результате проведенных нами расчетов оказалось, что производственная функция, полученная на базе усредненной информации за 5 лет для хозяйств Бурятской АССР, достаточно надежна и параметры ее не выходят за пределы области определения параметров одногодичных производственных функций, составленных для той же совокупности по тем же факторам» [13].

Таким образом, есть необходимость исследования *возможностей и пределов применения средних многолетних данных в регрессионном анализе.* В результате необходимо дать ответы на следующие основные вопросы:

1) являются ли указанные статистические парадоксы результатом случайных причин или они свойственны данному способу;

2) можно ли по одногодичным коэффициентам регрессии и корреляции путем определения каких-то их средних взвешенных вычислить аналогичные характеристики, отвечающие средним многолетним данным;

3) если статистические парадоксы свойственны данному способу, в каких условиях они возникают и в каких нет.

Для решения первой и второй из поставленных задач будем использовать целенаправленное преобразование основных математических формул, добиваясь ответа в общем виде.

Для ответа на последний вопрос используем графический анализ, основанный на преобразованных корреляционных диаграммах, а также предельно упрощенных, специально подобранных примерах, позволяющих оценить и осмыслить возникновение всех промежуточных результатов.

Основные результаты этого исследования впервые были опубликованы в журнале «Вестник статистики» (1977, № 7).

1.2. КОЭФФИЦИЕНТЫ РЕГРЕССИИ И КОРРЕЛЯЦИИ

Во избежание громоздких и трудно обозримых формул в данном параграфе ограничимся *парной линейной регрессией и корреляцией.* Распространение выводов на множественную и нелинейную регрессии логически простое, но вызывает возникновение новых дополнительных проблем, обусловливающих усложнение математических выкладок.

Основными показателями, получаемыми в результате регрессионно-корреляционного анализа, являются коэффициент регрессии и коэффициент корреляции. Все остальные показатели менее важны по своему познавательному содержанию и они с привлечением других элементарных характеристик (средних арифметических, дисперсий, объема совокупности) выводятся из первых. Поэтому здесь ограничимся исследованием лишь основных показателей связей.

Коэффициент регрессии b и коэффициент корреляции r парных линейных связей вычисляются по различным формулам, дающим одинаковые результаты. В данном случае преобразуем формулы, приняв в качестве исходных данных отклонения от средних арифметических:

$$b = \frac{\Sigma (x-\overline{x})(y-\overline{y})}{\Sigma (x-\overline{x})^2} \qquad (1.1)$$

$$r=\frac{\Sigma(x-\overline{x})(y-\overline{y})}{\sqrt{\Sigma(x-\overline{x})^2\Sigma(y-\overline{y})^2}} \qquad (1.2)$$

Для дальнейших упрощений рассмотрим только два года (отчетный и базисный). Вводим для них индексы: 0; 1 и «ср» — для средних данных.

Если число единиц наблюдения в обоих периодах одно и то же, а удельные веса отдельных единиц не принимаются во внимание, как это обычно принято в регрессионно-корреляционном анализе, индексы единиц совокупности и пределы суммирования для упрощения могут быть опущены, то двухгодичные средние определяются по формулам

$$x_{ср}=\frac{x_0+x_1}{2};\quad y_{ср}=\frac{y_0+y_1}{2};$$

$$\overline{x}_{ср}=\frac{\overline{x}_0+\overline{x}_1}{2};\quad \overline{y}_{ср}=\frac{\overline{y}_0+\overline{y}_1}{2}. \qquad (1.3)$$

В формулы (1.1) и (1.2) входят следующие суммы: $\Sigma(x-\overline{x})(y-\overline{y})$; $\Sigma(x-\overline{x})^2$; $\Sigma(y-\overline{y})^2$. Напишем их для двухгодичных средних данных с использованием одногодичных на основе субституции соотношений (1.3):

$$\Sigma(x_{ср}-\overline{x}_{ср})(y_{ср}-\overline{y}_{ср})=\Sigma\left(\frac{x_0+x_1}{2}-\frac{\overline{x}_0+\overline{x}_1}{2}\right)\times$$

$$\times\left(\frac{y_0+y_1}{2}-\frac{\overline{y}_0+\overline{y}_1}{2}\right)=\frac{1}{4}\Sigma(x_0+x_1-\overline{x}_0-\overline{x}_1)\times$$

$$\times(y_0+y_1-\overline{y}_0-\overline{y}_1)=\frac{1}{4}\Sigma[(x_0-\overline{x}_0)+(x_1-\overline{x}_1)]\times$$

$$\times[(y_0-\overline{y}_0)+(y_1-\overline{y}_1)].$$

Перемножая выражения в квадратных скобках и суммируя каждый член в отдельности, получаем формулу, которая дает возможность вычислить сумму произведений смешанных отклонений средних данных, не прибегая к осреднению первичных данных:

$$\Sigma(x_{ср}-\overline{x}_{ср})(y_{ср}-\overline{y}_{ср})=\frac{1}{4}[\Sigma(x_0-\overline{x}_0)(y_0-\overline{y}_0)+$$

$$+\Sigma(x_0-\overline{x}_0)(y_1-\overline{y}_1)+\Sigma(x_1-\overline{x}_1)(y_0-\overline{y}_0)+$$

$$+\Sigma(x_1-\overline{x}_1)(y_1-\overline{y}_1)]. \qquad (1.4)$$

Аналогично преобразуем суммы $\Sigma(x-\bar{x})^2$ и $\Sigma(y-\bar{y})^2$ для двухгодичных средних, в результате чего они примут следующий вид:

$$\Sigma(x_{\text{ср}}-\bar{x}_{\text{ср}})^2=\Sigma\left(\frac{x_0+x_1}{2}-\frac{\bar{x}_0+\bar{x}_1}{2}\right)^2=\frac{1}{4}\Sigma(x_0+$$

$$+x_1-\bar{x}_0-\bar{x}_1)^2=\frac{1}{4}\Sigma[(x_0-\bar{x}_0)+(x_1-\bar{x}_1)]^2=$$

$$=\frac{1}{4}[\Sigma(x_0-\bar{x}_0)^2+2\Sigma(x_0-\bar{x}_0)(x_1-\bar{x}_1)+$$

$$+\Sigma(x_1-\bar{x}_1)^2]; \qquad (1.5)$$

$$\Sigma(y_{\text{ср}}-\bar{y}_{\text{ср}})^2=\frac{1}{4}[\Sigma(y_0-\bar{y}_0)^2+2\Sigma(y_0-\bar{y}_0)\times$$

$$\times(y_1-\bar{y}_1)+\Sigma(y_1-\bar{y}_1)^2]. \qquad (1.6)$$

Подставляя выражения (1.4), (1.5), (1.6) в формулы (1.1) и (1.2) и сокращая на $^1/_4$, получаем *окончательные формулы,* пригодные для исчисления коэффициентов регрессии и корреляции, отвечающих средним двухгодичным данным, по информации отдельных лет:

$$b_{\text{ср}}=\frac{\Sigma(x_0-\bar{x}_0)(y_0-\bar{y}_0)+\Sigma(x_0-\bar{x}_0)(y_1-\bar{y}_1)+\Sigma(x_1-\bar{x}_1)(y_0-\bar{y}_0)+}{\Sigma(x_0-\bar{x}_0)^2+2\Sigma(x_0-\bar{x}_0)(x_1-\bar{x}_1)+}\rightarrow$$

$$\rightarrow\frac{+\Sigma(x_1-\bar{x}_1)(y_1-\bar{y}_1)}{+\Sigma(x_1-\bar{x}_1)^2}; \qquad (1.7)$$

$$r_{\text{ср}}=\frac{\Sigma(x_0-\bar{x}_0)(y_0-\bar{y}_0)+\Sigma(x_0-\bar{x}_0)(y_1-\bar{y}_1)+\Sigma(x_1-\bar{x}_1)(y_0-\bar{y}_0)+}{\sqrt{[\Sigma(x_0-\bar{x}_0)^2+2\Sigma(x_0-\bar{x}_0)(x_1-\bar{x}_1)+\Sigma(x_1-\bar{x}_1)^2][\Sigma(y_0-}}\rightarrow$$

$$\rightarrow\frac{+\Sigma(x_1-\bar{x}_1)(y_1-\bar{y}_1)}{\sqrt{-\bar{y}_0)^2+2\Sigma(y_0-\bar{y}_0)(y_1-\bar{y}_1)+\Sigma(y_1-\bar{y}_1)^2]}}. \qquad (1.8)$$

Формулы (1.7) и (1.8) довольно громоздкие и малопригодные для логического осмысливания и расчетов, поэтому выражения (1.4), (1.5) и (1.6) выражаем через дисперсии s^2 и ковариации cov, имея в виду общепринятые определения:

$$s_z^2=\frac{\Sigma(z-\bar{z})^2}{n};\quad \text{cov}_{zw}=\frac{\Sigma(z-\bar{z})(w-\bar{w})}{n}; \qquad (1.9)$$

$$\Sigma (x_{cp}-\overline{x}_{cp})(y_{cp}-\overline{y}_{cp})=\frac{n}{4}(\mathrm{cov}_{x_0y_0}+\mathrm{cov}_{x_0y_1}+$$
$$+\mathrm{cov}_{x_1y_0}+\mathrm{cov}_{x_1y_1}); \qquad (1.10)$$

$$\Sigma (x_{cp}-\overline{x}_{cp})^2=\frac{n}{4}(s^2_{x_0}+2\mathrm{cov}_{x_0x_1}+s^2_{x_1}); \qquad (1.11)$$

$$\Sigma (y_{cp}-\overline{y}_{cp})^2=\frac{n}{4}(s^2_{y_0}+2\mathrm{cov}_{y_0y_1}+s^2_{y_1}). \qquad (1.12)$$

Подставляя выражения (1.10), (1.11) и (1.12) в формулы (1.7) и (1.8) и сокращая на $n/4$, получаем формулы, более пригодные для решения поставленных задач:

$$b_{cp}=\frac{\mathrm{cov}_{x_0y_0}+\mathrm{cov}_{x_0y_1}+\mathrm{cov}_{x_1y_0}+\mathrm{cov}_{x_1y_1}}{s^2_{x_0}+2\mathrm{cov}_{x_0x_1}+s^2_{x_1}}; \qquad (1.13)$$

$$r_{cp}=\frac{\mathrm{cov}_{x_0y_0}+\mathrm{cov}_{x_0y_1}+\mathrm{cov}_{x_1y_0}+\mathrm{cov}_{x_1y_1}}{\sqrt{(s^2_{x_0}+2\mathrm{cov}_{x_0x_1}+s^2_{x_1})(s^2_{y_0}+2\mathrm{cov}_{y_0y_1}+s^2_{y_1})}} \qquad (1.14)$$

Учитывая, что

$$b_{zw}=\frac{\mathrm{cov}_{zw}}{s^2_w} \quad \text{и} \quad r_{zw}=\frac{\mathrm{cov}_{zw}}{s_z s_w},$$

откуда $\mathrm{cov}_{zw}=b_{zw}s^2_w$ и $\mathrm{cov}_{zw}=r_{zw}\cdot s_z\cdot s_w$, формулу (1.13) можно выразить также через коэффициенты регрессии, коэффициенты корреляции, дисперсии и стандартные отклонения одногодичных данных. Например,

$$b_{cp}=\frac{b_0s^2_{x_0}+b_1s^2_{x_1}+\mathrm{cov}_{x_0y_1}+\mathrm{cov}_{x_1y_0}}{s^2_{x_0}+2\mathrm{cov}_{x_0x_1}+s^2_{x_1}} \qquad (1.15)$$

или

$$b_{cp}=\frac{b_0s^2_{x_0}+b_1s^2_{x_1}+r_{x_0y_1}\cdot s_{x_0}s_{y_1}+r_{x_1y_0}\cdot s_{x_1}s_{y_0}}{s^2_{x_0}+2r_{x_0x_1}\cdot s_{x_0}\cdot s_{x_1}+s^2_{x_1}}. \qquad (1.16)$$

Формулу коэффициента корреляции, отражающего тесноту связей двухгодичных средних данных, см. (1.14), можно написать так:

$$r_{cp}=\frac{r_{x_0y_0}s_{x_0}s_{y_0}+r_{x_0y_1}s_{x_0}s_{y_1}+r_{x_1y_0}s_{x_1}s_{y_0}+r_{x_1y_1}s_{x_1}s_{y_1}}{\sqrt{(s^2_{x_0}+2r_{x_0x_1}s_{x_0}s_{x_1}+s^2_{x_1})(s^2_{y_0}+2r_{y_0y_1}s_{y_0}s_{y_1}+s^2_{y_1})}}. \qquad (1.17)$$

В формуле (1.15) показано, как коэффициент регрессии, соответствующий двухгодичным средним данным $b_{ср}$, определяется на основе коэффициентов регрессии, полученных по одногодичным данным b_0, b_1. Для этого необходимо дополнительно знать дисперсии факторного признака обоих периодов $s^2_{x_0}$, $s^2_{x_1}$, автоковариацию факторного признака $\mathrm{cov}_{x_0x_1}$ и два специфических показателя ковариации, названные нами *дрейфковариацией*[1] $\mathrm{cov}_{x_0y_1}$, $\mathrm{cov}_{x_1y_0}$.

В формуле (1.16) ковариации заменены на соответствующие коэффициенты автокорреляции и дрейфкорреляции. Коэффициенты дрейфкорреляции, как и коэффициенты автокорреляции, характеризуют определенные свойства дрейфов единиц совокупности во времени в корреляционной диаграмме.

Эти формулы мало пригодны для практической работы, так как вычисление автокорреляции и дрейфкорреляции переменных в условиях реальной, достаточно большой задачи требует большой вычислительной работы. Поэтому следует прийти к выводу, что не существует элементарных способов перехода от коэффициентов регрессии, вычисленных по одногодичным данным, к коэффициенту регрессии, соответствующему средним двухгодичным данным. То же относится и к коэффициентам корреляции.

Формулы (1.7), (1.8), (1.13)—(1.17) ценны тем, что дают возможность исследовать *отдельные компоненты*, определяющие числовые значения коэффициента регрессии и корреляции, вычисляемые по заранее усредненным данным. Можно убедиться, что, кроме компонент, отражающих связь в отдельные годы (b_0, b_1, r_0, r_1), в формулы входит ряд других величин, которые не учитываются при обработке одногодичных данных. При достаточно больших числовых значениях автокорреляции и дрейфкорреляции эти компоненты могут оказаться доминирующими. В таких условиях, например, коэффициент регрессии, установленный на основе средних данных, не только не будет близким к среднему из одногодичных коэффициентов регрессий, но может даже выйти за область вариации последних. Таким образом, математическая причина возникновения статистических парадоксов найдена. Но для лучшей наглядности

[1] См.: Вестник статистики, 1977, № 7, с. 42; 1977, № 9, с. 21.

ности рассмотрим то же явление на простом числовом примере, решение которого можно проследить без применения вычислительной техники.

1.3. ВОЗНИКНОВЕНИЕ СТАТИСТИЧЕСКИХ ПАРАДОКСОВ

Составление числовых примеров для иллюстрации отдельных характерных ситуаций, руководствуясь только вышерассмотренными формулами, — дело довольно сложное. Составление характерных примеров и, самое главное, выяснение ситуаций, ведущих к статистическим парадоксам, значительно облегчается, если применяются специальные корреляционные диаграммы, которые мы называем *дрейфограммами*.

На обычной корреляционной диаграмме точкой наносят сопряженные данные первой единицы совокупности за базисный период. Затем стрелкой (ее острием) показывают сопряженные данные той же единицы совокупности, но в отчетном периоде. Точку и острие стрелки соединяют прямой. Получается вектор, характеризующий перемещение (дрейф) данной единицы (например, предприятия) на корреляционной диаграмме. Аналогичным образом откладываются векторы дрейфа всех единиц рассматриваемой совокупности.

Пример 1.1. На рис. 1.1 показана дрейфограмма. Векторы подобраны таким образом, чтобы их середины лежали на биссектрисе угла первого квадранта системы прямоугольных координат, а точки и «острия стрелок» разбросаны симметрично по обе стороны этой линии.

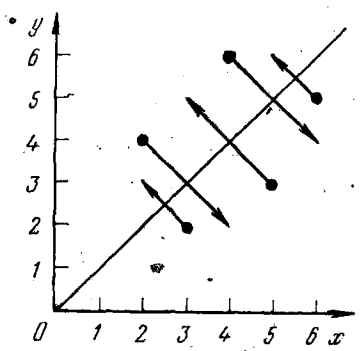


Рис. 1.1. Дрейфограмма примера 1.1

Для числовых расчетов сохраним ранее введенные символы и индексацию. Факторный признак обозначен символом x, подстрочные индексы обозначают: 0 — базисный период, 1 — отчетный период, ср — двухгодичный средний. Результативный признак представлен символом y с той же индексацией. Исходные данные при-

Таблица 1.2

ИСХОДНЫЕ ДАННЫЕ ПРИМЕРА 1.1

№ п/п	Период				Двухгодичные средние	
	базисный		отчетный			
	x_0	y_0	x_1	y_1	$x_{ср}$	$y_{ср}$
1	2	3	4	5	6	7
1	3	2	2	3	2,5	2,5
2	2	4	4	2	3	3
3	5	3	3	5	4	4
4	4	6	6	4	5	5
5	6	5	5	6	5,5	5,5
Σ	20	20	20	20	20	20
ср	4	4	4	4	4	4

Таблица 1.3

ВЫЧИСЛЕНИЕ СУММ ОТКЛОНЕНИЙ И СУММ КВАДРАТОВ ОТКЛОНЕНИЙ ПРИМЕРА 1.1

№ п/п	Индексы											
	$x-\bar{x}$			$y-\bar{y}$			$(x-\bar{x})^2$			$(y-\bar{y})^2$		
	0	1	ср	0	1	ср	0	1	ср	0	1	ср
	8	9	10	11	12	13	14	15	16	17	18	19
1	—1	—2	—1,5	—2	—1	—1,5	1	4	2,25	4	1	2,25
2	—2	0	—1	0	—2	—1	4	0	1	0	4	1
3	1	—1	0	—1	1	0	1	1	0	1	1	0
4	0	2	1	2	0	1	0	4	1	4	0	1
5	2	1	1,5	1	2	1,5	4	1	2,25	1	4	2,25
Σ	0	0	0	0	0	0	10	10	6,5	10	10	6,5
s^2	x	x	x	x	x	x	2	2	1,3	2	2	1,3
s	x	x	x	x	x	x	1,414	1,414	1,140	1,414	1,414	1,140

мера 1.1, соответствующего дрейфограмме 1.1, приведены в табл. 1.2.

В табл. 1.3 и 1.4 показано вычисление сумм отклонений и сумм квадратов отклонений, а также сумм смешанных отклонений. По этим данным легко вычислить все необходимые дисперсии, стандартные отклонения и ковариации (факторные, автоковариации и дрейфковариации). Поскольку табл. 1.2—1.4 относятся к одному примеру, для удобства ссылок графы таблиц пронумерованы последовательно: с 1-й по 26-ю.

Таблица 1.4

ВЫЧИСЛЕНИЕ СУММ ПАРНЫХ ОТКЛОНЕНИЙ ПРИМЕРА 1.1

№ п. п	$(x-\bar{x})(y-\bar{y})$ Индексы 0	Индексы 1	Индексы ср	$(x_0-\bar{x_0})\times(x_1-\bar{x_1})$	$(y_0-\bar{y_0})\times(y_1-\bar{y_1})$	$(x_0-\bar{x_0})\times(y_1-\bar{y_1})$	$(x_1-\bar{x_1})\times(y_0-\bar{y_0})$
	20	21	22	23	24	25	26
1	2	2	2,25	2	2	1	4
2	0	0	1	0	0	4	0
3	—1	—1	0	—1	—1	1	1
4	0	0	1	0	0	0	4
5	2	2	2,25	2	2	4	1
Σ	3	3	6,5	3	3	10	10
cov	0,6	0,6	1,3	0,6	0,6	2	2

В графах 2 и 3 показаны исходные данные базисного периода. Используя строки Σ по графам 20, 14 и 17 и формулы (1.1) и (1.2), подсчитаем коэффициенты регрессии и корреляции для базисного года:

$$b_{y_0x_0}=\frac{3}{10}=0{,}3;\quad r_{y_0x_0}=\frac{3}{\sqrt{10\cdot 10}}=0{,}3;$$

$$a_0=\bar{y_0}-b\bar{x_0}=4-0{,}3\cdot 4=2{,}8.$$

Таким образом, данные базисного периода связаны уравнением регрессии $y_0=2{,}8+0{,}3\ x_0$. Они имеют невысокую тесноту связи ($r_0=0{,}3$).

В графах 4 и 5 показаны исходные данные отчетного периода. Используем строку Σ по графам 21, 15, 18 и те же формулы, получим:

$$b_{y_1x_1}=\frac{3}{10}=0{,}3;\quad r_{y_1x_1}=\frac{3}{\sqrt{10\cdot 10}}=0{,}3;\quad a_1=4-0,\ 3\cdot 4=2{,}8.$$

Уравнение регрессии $y_1=2{,}8+0{,}3\ x_1$ идентично уравнению базисного периода с идентичным показателем тесноты связи.

Так как все показатели регрессии и корреляции по обоим годам одинаковые, следует ожидать, что те же результаты получатся по двухгодичным средним данным. Но по данным строки Σ и графам 22, 16, 19 подсчитаем, что

$$b_{ср}=\frac{6{,}5}{6{,}5}=1;\quad r_{ср}=\frac{6{,}5}{\sqrt{6{,}5\cdot 6{,}5}}=1;\quad a=0.$$