О.П.Крастинь

РАЗРАБОТКА И ИНТЕРПРЕТАЦИЯ МОДЕЛЕЙ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ В ЭКОНОМИКЕ

Ольгерт Петрович Крастинь

РАЗРАБОТКА И ИНТЕРПРЕТАЦИЯ МОДЕЛЕЙ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ В ЭКОНОМИКЕ

Редактор *Н. Буртниеце*
Художник *И. Лиецкалниня*
Художественный редактор *Э. Бурова*
Технический редактор *В. Лаздиня, И. Заляйскалне*
Корректор *М. Устинова*

ИБ № 1066

Сдано в набор 25.05.82. Подписано в печать 04.02.83. ЯТ 03538. Формат 60×90/16. Бумага типогр. № 1. Гарнитура литературная. Высокая печать. 19 физ. печ. л.; 19 усл. печ. л.; 19 усл. кр.-отт.; 18,25 уч.-изд. л. Тираж 1000 экз. Заказ № 1383. Цена 1 р. 40 к. Заказное. Издательство «Зинатне», 226530 ГСП Рига, ул. Тургенева, 19. Отпечатано в производственном объединении «Полиграфистс» Государственного комитета Латвийской ССР по делам издательств, полиграфии и книжной торговли, 226050 Рига, ул. Горького, 6.

О. П. Крастинь

РАЗРАБОТКА И ИНТЕРПРЕТАЦИЯ МОДЕЛЕЙ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ В ЭКОНОМИКЕ

ЛАТВИЙСКИЙ ОРДЕНА ТРУДОВОГО КРАСНОГО ЗНАМЕНИ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМ. П. СТУЧКИ

О. П. Крастинь

РАЗРАБОТКА И ИНТЕРПРЕТАЦИЯ МОДЕЛЕЙ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ В ЭКОНОМИКЕ

РИГА «ЗИНАТНЕ» 1983

33S3
65.05
К 784

Крастинь О. П. РАЗРАБОТКА И ИНТЕРПРЕТАЦИЯ МОДЕЛЕЙ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ В ЭКОНОМИКЕ. — Рига: Зинатне, 1983. 302 с.

В монографии рассматривается комплекс проблем разработки и интерпретации моделей корреляционных связей в экономике. При постановке задачи моделирования особое внимание уделяется логическим рекомендациям, учитываемым при предварительном отборе факторов, а также оценке характера исходных данных, обоснованию формы связи, особенностям глобальных и локальных моделей и их экономическому содержанию. В группе вопросов оценки параметров статистических моделей связи наиболее подробно изучается система алгоритмов и показателей, основанных на обратных матрицах систем нормальных уравнений. Особое внимание уделяется интерпретации и применению регрессионных моделей при анализе хозяйственной деятельности и прогнозировании. Методические выводы, полученные на основе исследования связей в сельском хозяйстве, могут быть использованы при решении аналогичных задач в других отраслях народного хозяйства.

Книга рассчитана на научных работников, преподавателей, аспирантов и студентов, интересующихся вопросами моделирования связей в экономике.

Печатается по решению Редакционно-издательского совета АН Латвийской ССР от 25 июня 1981 года

К $\frac{0004020105-022}{М811(11)-83}$ 5—82

ПРЕДИСЛОВИЕ

Статистическая наука располагает многочисленными методами изучения связей и закономерностей различных явлений. Так, наряду с простыми и доступными методами (аналитическая группировка, графический анализ, различные индексы и др.) существуют сложные, тонкие методы, требующие отличных знаний математики и обязательного применения ЭВМ (факторный анализ, анализ главных компонент, некоторые виды кластерного анализа, спектральный анализ и др.).

Регрессионный и корреляционный анализы занимают промежуточное положение. Они достаточно просты для освоения при наличии предварительных знаний в области математики и статистики в том объеме, какой может дать экономический факультет вуза. Наиболее простые задачи могут быть решены на клавишных вычислительных машинах, хотя применение ЭВМ обеспечивает значительную экономию труда. В то же время экономико-математические модели, разработанные на основе регрессии и корреляции, достаточно адекватны для решения многих практических задач. Регрессионный и корреляционный анализы достаточно универсальны и находят применение как в общественных науках, так и в естественных и технических.

Первые три главы настоящей работы посвящены основам методов регрессии и корреляции. Новизна приводимых материалов заключается в основном в содержательной статистической интерпретации различных уже известных в математической статистике показателей, которые, однако, нередко вычисляются и применяются довольно формально. Следующие главы (основная часть книги) построены в соответствии с принципом последовательности укрупненных стадий научного исследования связей с помощью регрессии и корреляции. На первой стадии работы выполняется обоснование результативного и факторных признаков, формы связи, совокупности, по которой будут собраны исходные данные, и др. Вторая стадия связана с обработкой собранных данных. Здесь особенно актуальным является рассмотрение различных аспектов многошагового регрессионного анализа, разновидностей формул, применяемых для вычисления, и их сравнительная оценка, а также изучение применения обратной корреляционной матрицы и возможности усовершенствования исследования на ее основе.

Третья и заключительная стадии работы — оценка результатов и их применение при экономическом анализе, изучение содержательной интерпретации уравнения регрессии и его параметров, теоретического (расчетного) уровня результативного признака и других производных на их основе величин.

В последней главе выделены вопросы, связанные с применением уравнений регрессии при прогнозировании и планировании, рассмотрены вопросы устойчивости уравнений регрессии и техники прогнозирования и оценки надежности прогнозов.

Материал книги представлен таким образом, чтобы читатель, ознакомившийся с несколькими параграфами первой главы, уже имел возможность оценить основные задачи регрессии и корреляции. Такой принцип в построении работы обусловил неизбежность некоторых повторений, которые, по нашему мнению, допустимы в отношении вопросов содержательной интерпретации уравнения регрессии и его параметров, практического его применения при решении экономических задач и др.

Работа носит методический характер, поэтому большинство из приведенных примеров имеет иллюстративное значение, в таких примерах не указаны территория и время, к которым относятся статистические данные. Если же территория указывается, но отсутствует указание на время, то такой пример построен на основе реальных данных. Однако выводы, близкие к действительности, не пригодны для практических экономических выводов вследствие, например, недостаточно большой совокупности, недостаточной представительности, устаревших данных и т. д. В двух последних главах отражены результаты вполне конкретных исследований, особенно в отношении характеристики устойчивости уравнений регрессии во времени. Признаком таких материалов служит указание на территорию и время, к которым относятся эти данные.

Автор

Глава 1

ЛИНЕЙНАЯ РЕГРЕССИЯ И КОРРЕЛЯЦИЯ

1.1. ОТ ГРУППИРОВКИ К РЕГРЕССИИ

Одна из важнейших задач экономического анализа состоит в изучении связей между различными экономическими явлениями. Среди простейших способов изучения взаимосвязей, рассматриваемых общей теорией статистики, наибольшую известность имеют параллельные ряды, графический анализ и различные индексы, но самым распространенным является способ аналитической группировки.

Аналитической группировкой называется статистическая таблица, в подлежащем которой указаны интервалы какого-либо существенного признака, по которому сгруппированы единицы совокупности. В сказуемом таблицы приведены групповые средние другого признака, взаимосвязь которого с группировочным признаком предстоит изучить. С помощью такой таблицы можно оценить степень согласованности изменений группировочного признака и групповых средних другого признака. Если оба признака при переходе от одной группы к другой изменяются согласованно, то рассматриваемые признаки являются взаимосвязанными. Если согласованные изменения отсутствуют, то и взаимосвязь в данном случае не наблюдается.

Табл. 1.1 отражает результаты группировки зависимости уро-

Таблица 1.1

АНАЛИТИЧЕСКАЯ ГРУППИРОВКА, ОТРАЖАЮЩАЯ ЗАВИСИМОСТЬ УРОЖАЙНОСТИ ЗЕРНОВЫХ ОТ НОРМ ВНЕСЕНИЯ УДОБРЕНИЙ

Номер группы i	Внесено минеральных удобрений, ц/га в пересчете на действующие вещества x_i	Число единиц совокупности (хозяйств) n_i	Урожайность зерновых, ц/га y_i
1	0,5—1,0	1	18
2	1,0—1,5	3	24
3	1,5—2,0	7	25
4	2,0—2,5	6	33
5	2,5—3,0	3	37
Всего	×	20	×
В среднем	×	×	28,8

жайности зерновых от норм внесения минеральных удобрений. Группировочным признаком здесь является факторный признак — норма внесения удобрений. Таблица свидетельствует о том, что по мере роста группировочного признака (норм внесения удобрений) увеличиваются групповые средние (урожайность). Таким образом, группировка дает возможность сделать вывод о том, что в хозяйствах, применяющих более высокие нормы минеральных удобрений, в среднем и более высокая урожайность. Рост урожайности при переходе от одной группы к другой различен.

Познавательная ценность группировки более наглядно проявляется в случае, если ее результаты отражаются графически. С этой целью строится диаграмма в прямоугольной системе координат (на оси абсцисс — значения группировочного признака, на оси ординат — групповых средних). Каждая пара сопряженных чисел $(x_i y_i)$ изображается точкой в системе координат, причем x_i является серединой интервала группировочного признака, а y_i — групповым средним другого признака в i-й группе. Соединяя стоящие рядом точки отрезками прямой, получаем ломаную линию — линию эмпирической регрессии (рис. 1.1).

Из приведенного примера видим, что аналитическая группировка является элементарным средством определения связей, существующих между экономическими явлениями. Ценность аналитических группировок заключается в первую очередь в их простоте. Таблицы и графики, отражающие результаты аналитических группировок, легко читаются и поэтому находят широкое применение. Следует, однако, указать и на некоторые недостатки, характерные для аналитических группировок. Из диаграммы 1.1 видим, что

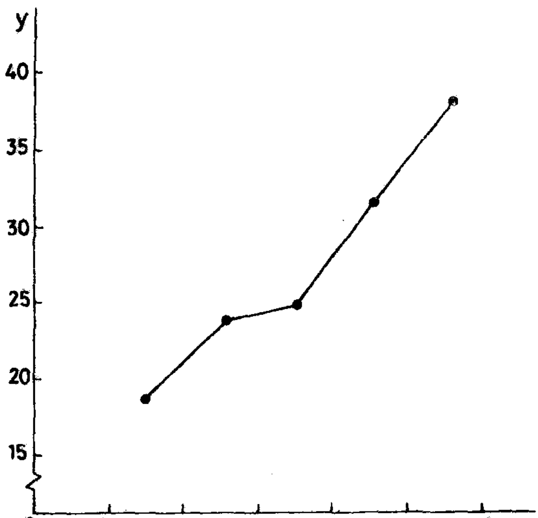

Рис. 1.1. Эмпирическая регрессия, отражающая связь норм внесения минеральных удобрений и урожайности зерновых: x — минеральные удобрения, ц/га действующих веществ; y — урожайность зерновых, ц/га.

приросту норм удобрений Δx не всегда соответствует пропорциональный прирост урожайности Δy. Например, увеличение норм удобрений с 0,75 до 1,25 ц/га связано со значительным приростом урожайности; увеличение с 1,25 до 1,75 ц/га обеспечило менее заметный прирост урожайности; при последующем увеличении — с 1,75 до 2,25 ц/га — прирост урожайности снова был значительным. Такое скачкообразное изменение урожайности по мере увеличения норм вносимых удобрений не отражает реальной зависимости урожайности от дозы внесенных удобрений. В этом легко убедиться, если в результате перегруппировки материала образовать интервалы, имеющие другие границы, или изучить те же взаимосвязи по данным других предприятий.

Эмпирическая регрессия отражает не только основную закономерность, но и различные отступления от нее. Это особенно характерно в отношении групп, которые представлены малым числом единиц совокупности. Таким образом, при рассмотрении аналитической группировки следует оценивать общий характер изменений во всей группировке. Различия групповых средних в соседних группах имеют второстепенное значение. Это обосновывает целесообразность построения не ломаной линии эмпирической регрессии, а прямой или плавной кривой, отражающей общую закономерность связей, свободную от случайностей. Такая прямая или кривая в отличие от эмпирической регрессии называется линией теоретической регрессии, а ее уравнение — уравнением регрессии. Ряд авторов вместо понятия «уравнение регрессии» используют «корреляционное уравнение», «уравнение связи» и др. Нахождение уравнения регрессии является одной из важнейших задач, решаемых методом регрессионного анализа.

1.2. ФУНКЦИОНАЛЬНАЯ И КОРРЕЛЯЦИОННАЯ СВЯЗЬ

При рассмотрении связи двух переменных величин на основе логических соображений обычно устанавливается, какая из них является причиной и какая — следствием, т. е. независимой или зависимой переменной. Например, урожайность зависит от норм внесения удобрений, а не наоборот. Поэтому количество удобрений является независимым признаком, урожайность — зависимым. В другом случае урожайность может быть независимым признаком (например, если исследуется ее связь с себестоимостью зерна, которая зависит от величины урожайности). Независимые признаки в статистике называются факторными, а зависимые — результативными. Причинную связь устанавливают на основе качественного анализа, исходя из достижений той отрасли науки, к которой относится данное исследование.

Зависимость двух переменных называется функциональной, если для каждого значения независимой переменной x_i можно найти одно определенное значение зависимой переменной y_i. В экономике, в отличие от естественных и технических наук, функциональные связи встречаются редко. В то же время экономические явления столь же редко бывают полностью независимыми друг от друга. Зависимости в экономике не наблюдаются при рассмотрении отдельных единиц совокупности, но они отчетливо проявляются при рассмотрении совокупности в целом. Для этого совокупность должна быть достаточно большой. Статистическая зависимость заключается не в изменении значения зависимой переменной от значения независимой в каждой единице совокупности, а только в совокупности в целом. Статистическая зависимость может быть различного характера. В данной работе рассматривается только корреляционная связь.

Корреляционной связью называется связь между статистическими явлениями, состоящая в изменении среднего значения одного признака в зависимости от изменения значения другого. Таким образом, при корреляционной зависимости одному значению независимого признака соответствует не одно, а несколько значений зависимого, определяющих характерное множество. Последнее может быть рассмотрено как ряд распределения. Каждому значению независимого (факторного) признака соответствует свой ряд распределения значений зависимого признака. Характер этих рядов распределения изменяется в зависимости от факторного признака. В первую очередь изменяются средние значения распределений. Таким образом, при наличии корреляционной связи по мере изменения значений независимого признака закономерно изменяются средние величины зависимого признака.

Корреляционными методами, или корреляционным и регрессионным анализом, называется совокупность математических приемов, при помощи которых исследуются и обобщаются взаимосвязи корреляционно связанных переменных.

При помощи методов регрессионного и корреляционного анализов решаются две основные задачи:

1) нахождение общей закономерности, которая характеризует зависимость двух корреляционно связанных переменных, другими словами, разработка математической модели связи;

2) определение тесноты связей.

1.3. КОРРЕЛЯЦИОННАЯ ДИАГРАММА

Определение уравнения регрессии как математической модели связи является логическим продолжением аналитической группировки. Аналитическая группировка и регрессионный анализ как два статистических метода предназначены для решения аналогичных задач изучения связей, однако это не означает, что их всегда следует использовать совместно и последовательно. Более того, использование в регрессионном анализе предварительно сгруппированных данных может быть причиной серьезных ошибок, особенно при изучении тесноты связи. Поэтому анализ регрессий и корреляций следует рассматривать как самостоятельный статистический метод.

Задачу регрессионного анализа можно проиллюстрировать графически, изображая корреляционную диаграмму для двух взаимосвязанных переменных в системе прямоугольных координат. На оси абсцисс наносится шкала в единицах факторного признака, на оси ординат — результативного. Отдельные единицы совокупности изображаются точками, координаты которых являются данными о двух сопряженных переменных (x_i, y_i).

На рис. 1.2 приведена корреляционная диаграмма удобрение — урожайность по данным 20 сельскохозяйственных предприятий. На рисунке видно, что отдельным значениям норм внесения удобрений не соответствует определенное значение урожайности. Однако в правой части графика точки концентрируются над осью абсцисс выше, чем в левой. Это свидетельствует о наличии корреляционной связи, заключающейся в том, что по мере повышения норм внесения удобрений урожайность в среднем повышается.

Дальнейшая задача регрессионного анализа состоит в нахождении прямой (или кривой), располагающейся по возможности ближе ко всем точкам на корреляционной диаграмме.

Для приближенного решения задачи такую линию можно провести визуально, а потом по ней найти уравнение (уравнение регрессии). Однако такой способ дает неточные результаты, причем различные у разных исследователей. Поэтому уравнение регрессии определяется как результат математических вычислений.

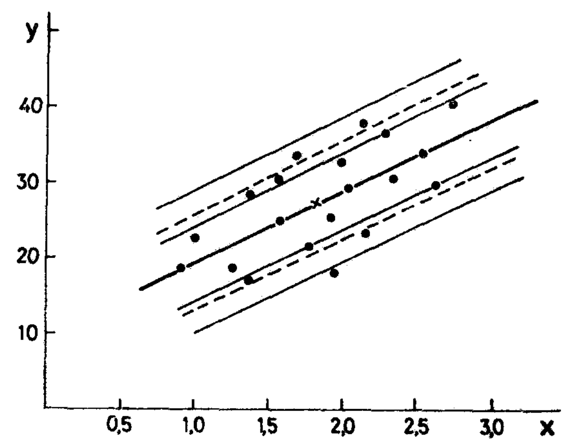

Рис. 1.2. Корреляционная диаграмма, линия регрессии и доверительные границы, отражающие связь минеральных удобрений и урожайности зерновых.

Обозначения те же, что на рис. 1.1.

1.4. ЛИНЕЙНОЕ УРАВНЕНИЕ РЕГРЕССИИ

До начала расчетов необходимо обосновать общую форму связи. Это означает выбор общего типа или класса уравнений, которые достаточно точно отражают интересующую нас связь. В данной главе принято, что связь линейна и ее с достаточной точностью отражает линейное уравнение регрессии. В экономике линейные связи и зависимости встречаются часто.

Если изучаемая связь не является линейной, то следует выбрать и обосновать другой тип уравнения. Обоснование общей формы связи является сложной задачей, которая обычно решается при совместном использовании экспертных и математических методов.

Линейное уравнение регрессии обычно записывается следующим образом:

$$\tilde{y}=a+bx, \tag{1.1}$$

где y — результативный признак; x — факторный признак; a, b — числовые константы, которые принято называть параметрами уравнения регрессии.

Уравнение, которое аналитически моделирует зависимость средней величины результативного признака от факторного признака, называется уравнением регрессии. Коэффициент b в этом уравнении называется коэффициентом регрессии, a — свободный член уравнения регрессии.

Линейное уравнение регрессии графически имеет вид прямой. Коэффициент регрессии является угловым коэффициентом, т. е. тангенсом угла наклона прямой регрессии к оси абсцисс x. При построении прямой по данному уравнению, а также при определении уравнения по заданной прямой необходимо учитывать единицы измерения взаимосвязанных признаков.

Свободный член уравнения a представляет собой длину отрезка оси ординат от начала координат до пересечения с прямой регрессии. Длина этого отрезка измеряется в единицах результативного признака.

1.5. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Чтобы определить прямую, которая наилучшим образом отражает общую закономерность связи средней величины результативного признака с факторным, необходимо обосновать определенное требование (критерий), которому должна соответствовать прямая. Поскольку прямая на корреляционной диаграмме должна находиться по возможности ближе ко всем точкам, отражающим сопряженные данные, то возникает вопрос, как измерить расстояние от этих точек до прямой. Оно может быть измерено

1) по кратчайшему пути, т. е. по перпендикуляру к линии регрессии;

2) по вертикали, т. е. по вертикальным прямым, соединяющим точки с прямой регрессии;

3) по горизонтали, т. е. по горизонтальным прямым, соединяющим точки с прямой регрессии.

Поскольку имеется три способа измерения расстояний точек до линии регрессии, получаем три критерия близости, в соответствии с которыми может быть определена прямая (графическая иллюстрация этих критериев представлена в левой части рис. 1.3). Существуют математические способы, которые дают возможность реализовать эти требования. Среди указанных в научной литературе наиболее часто обсуждается так называемый способ минимизации суммы линейных отклонений (вертикальных). Однако на практике чаще всего используется другой критерий, который реализуется методом наименьших квадратов.

Метод наименьших квадратов предусматривает требование минимизировать сумму квадратов отклонений фактически наблюденных величин результативного признака y_i от рассчитанных по уравнению регрессии величин $\tilde{y}_i$. Графически это можно представить так. Из каждой точки (x_i, y_i) строятся вертикальные отрезки

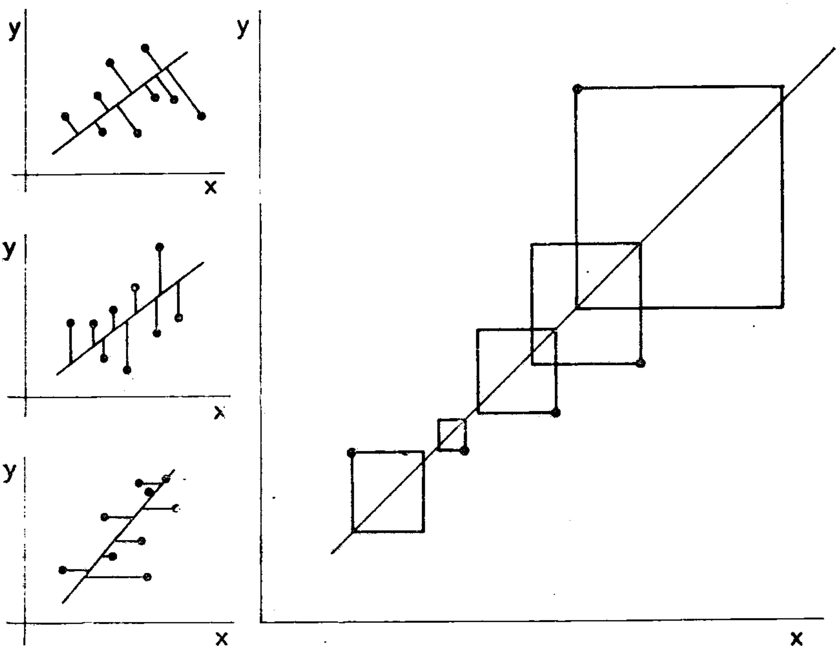

Рис. 1.3. Геометрическая интерпретация метода наименьших квадратов и других способов нахождения выравнивающей прямой.

до линии регрессии. Далее образуем квадраты, принимая эти отрезки прямых за одну из их сторон. Метод требует, чтобы сумма площадей таких квадратов была минимальной (правая часть рис. 1.3).

Метод наименьших квадратов имеет существенные преимущества перед другими методами нахождения прямой, если отклонения $y_i-\tilde{y}_i$ образуют нормальное распределение. Это одна из основных предпосылок для того, чтобы использовать закон нормального распределения для вероятностных оценок результатов моделирования. На практике эта предпосылка чаще всего выполняется приближенно. Опыт показал, что при таких условиях метод наименьших квадратов дает неплохие результаты.

Если необходимо определить линейное уравнение $\tilde{y}_i=a+bx_i$, то условие нахождения его методом наименьших квадратов можно записать следующим образом:

$$Q=\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2\to\min, \tag{1.2}$$

где i — номер единицы наблюдения; n — число единиц в совокупности. Если суммирование проводится по всей совокупности, как в нашем случае, то для упрощения в статистике разрешается опустить номера наблюдений и пределы суммирования. После этого упрощения условие наименьших квадратов записывается так:

$$Q=\Sigma(y-\tilde{y})^2\to\min.$$

Для отыскания параметров a и b уравнения прямой $\tilde{y}=a+bx$, отвечающей требованию метода наименьших квадратов, в формуле (1.2) $\tilde{y}$ заменяют его значением $a+bx$:

$$Q=\Sigma(y-a-bx)^2. \tag{1.3}$$

Определяемые параметры a и b рассматриваем в качестве переменных величин, а сумму квадратов отклонений Q — их функции. Как известно, эта функция будет минимальной при условии, что обе частные производные равны нулю:

$$\frac{\partial Q}{\partial a}=0; \quad \frac{\partial Q}{\partial b}=0.$$

Используя приемы нахождения частных производных от функции (1.3), получаем:

$$\frac{\partial Q}{\partial a}=-2\Sigma(y-a-bx);$$

$$\frac{\partial Q}{\partial b}=-2\Sigma(y-a-bx)x.$$

Приравняв обе части к нулю и умножив на $-0{,}5$, получим:

$$\Sigma(y-a-bx)=0;$$

$$\Sigma(y-a-bx)x=0. \qquad (1.4)$$

После суммирования каждого из членов в отдельности, учитывая, что $\Sigma a=na$, переносим величины, содержащие y, в правую часть уравнений:

$$an+b\Sigma x=\Sigma y;$$

$$a\Sigma x+b\Sigma x^2=\Sigma xy. \qquad (1.5)$$

Система (1.5), в которой a и b являются неизвестными, а n, Σx, Σy, Σxy, Σx^2 известны (вычислены по данным n наблюдений x_i, y_i), называется системой нормальных уравнений, решением которой являются параметры прямой.

1.6. СОСТАВЛЕНИЕ И РЕШЕНИЕ СИСТЕМЫ НОРМАЛЬНЫХ УРАВНЕНИЙ

Для решения практических задач система нормальных уравнений используется в готовом виде (см. (1.5)).

Примем, что перед нами стоит задача изучить влияние норм внесения удобрений на урожайность зерновых. Для этой цели используем данные по 20 сельскохозяйственным предприятиям Латвийской ССР (см. табл. 1.2 и рис. 1.2). Обозначим через x_i количество внесенных удобрений (ц/га действующих веществ), а через y_i — урожайность зерновых в i-м хозяйстве.

Для определения параметров линейного уравнения регрессии согласно системе (1.5) необходимо иметь следующие величины:

$$n,\ \Sigma x_i,\ \Sigma y_i,\ \Sigma x_i^2,\ \Sigma x_i y_i.$$

Если расчет осуществляется на счетно-клавишных машинах, то рекомендуется подготовить рабочую таблицу (табл. 1.2); если же расчеты выполняются на ЭВМ, то подготавливается только исходная информация (столбцы x_i, y_i). Остальные расчеты выполняются автоматически по типовой программе. Одновременно с вычислением сумм, необходимых для составления системы нормальных уравнений, рекомендуется вычислить сумму квадратов Σy_i^2, которая понадобится для определения показателей тесноты связи.

Итоговые суммы из таблицы подставляются в систему нормальных уравнений (1.5). В результате имеем

$$20a+37{,}3b=576;\quad 37{,}3a+74{,}95b=1126{,}6.$$