МИНИСТЕРСТВО
ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Государственное образовательное учреждение высшего профессионального образования

САМАРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Д.Н. Цивинский

РАЗНООБРАЗИЕ ФОРМ УРАВНЕНИЙ ПАРНОЙ РЕГРЕССИИ

Допущено учебно-методическим объединением вузов Российской Федерации по нефтегазовому образованию в качестве учебного пособия для подготовки дипломированных специалистов по направлению 650700 «Нефтегазовое дело» и бакалавров и магистров направления 553600 «Нефтегазовое дело»

Самара 2002

УДК 519.22:681.3(076.5)

Разнообразие форм уравнений парной регрессии: Учебное пособие / Д.Н. Цивинский; Самар. гос. техн. ун-т. Самара, 2002, 80 с.

Достаточно подробно описан метод НК в двумерном пространстве: теоретические основы метода, его достоинства и недостатки, парная корреляция, линейная и нелинейная двухпараметрическая регрессия. Для облегчения понимания подробно рассмотрены более 20 специальных терминов, причём определение каждого термина представлено отдельной статьёй. Описан метод получения трёхпараметрических уравнений регрессии, совмещающий метод НК и метод сканирования. Обоснована процедура преобразования системы координат и поиска оптимальной формы уравнения парной регрессии. В сложных случаях рекомендованы параболическая регрессия и сплайн-аппроксимация. Всего рассмотрено и описано более 20 форм уравнений парной зависимости, охватывающих практически все важные случаи. Регрессионный анализ уравнения включает оценку значимости коэффициентов, построение стандартных границ корреляционного поля и проверку уравнения на адекватность.

Пример расчёта иллюстрирует все этапы получения уравнения регрессии.

Учебное пособие предназначено для студентов нефтетехнологического факультета очной и заочной форм обучения, может быть полезно студентам и аспирантам любой специальности при обработке экспериментальных данных, а также может быть использовано инженерами, занимающимися экспериментальными исследованиями.

ISBN 5-7964-0242-0.

Ил. 10. Табл. 10. Библиогр. 22 назв.

Печатается по решению редакционно-издательского совета СамГТУ.

Рецензенты: В.А.Акулов,
В.К.Давыдов

ISBN 5-7964-0242-0

Условные обозначения

Условное обозначение	Физическая величина
B	Формальный параметр
b_j	Оценка j-того коэффициента уравнения регрессии (оценка генерального параметра β_j)
D	Детерминант матрицы
DX	Математическое ожидание дисперсии случайной величины X
D^2	Выборочная дисперсия (смещённая оценка генеральной дисперсии, σ^2_x)
F	Критерий Фишера
h	Формальный параметр
i	Номер наблюдения (эксперимента), номер элемента массива, номер строки матрицы наблюдений
j	Номер коэффициента уравнения регрессии, номер фактора
k	Количество факторов (размерность факторного пространства)
l	Число связей, накладываемых на выборку (количество коэффициентов в уравнении регрессии)
m_x, m_y	Оценки математических ожиданий случайных величин X и Y
n	Размерность массива данных, число наблюдений (опытов), число *точек*
P	Вероятность события
P	Доверительная вероятность
$p(x)$	Плотность распределения вероятностей случайной величины
p_i	Вероятность i-того события
r_{xy}	Выборочный коэффициент корреляции x и y
S^2	Сумма квадратов отклонений
s_x, s_y	Выборочные квадратичные отклонения переменных X и Y (стандартные отклонения)
$s^2_{ад}$	Дисперсия адекватности

Условное обозначение	Физическая величина
$s^2_{оп}$	Дисперсия воспроизводимости
s^2_x, s^2_y	Выборочные дисперсии (несмещённые оценки генеральных параметров, $б^2_x$, $б^2_y$)
t	Критерий Стьюдента
W	Вес результата измерения (вес *экспериментальной точки*)
X, Y	Случайные величины
$\bar{x}$, $\bar{y}$	Выборочные средние значения переменных (оценки математических ожиданий M_x, M_y)
x	Независимая переменная величина (фактор)
y	Функция отклика (экспериментальное значение)
$\hat{y}$	Значение функции отклика, рассчитанное по уравнению регрессии
α	Уровень значимости
β_j	Математическое ожидание j-того коэффициента уравнения регрессии (генеральный j-тый параметр)
Γ	Гамма функция
Δ	Интервал
EX	Математическое ожидание случайной величины X
ε	Малое число
θ	Статистика критерия
M_x, M_y	Математические ожидания случайных величин X, Y
μ_2	Центральный момент второго порядка
ν	Число степеней свободы
ρ_{xy}	Генеральный коэффициент корреляции
$б^2_x$, $б^2_y$	Генеральные дисперсии случайных величин X, Y
$б_x$, $б_y$	Генеральные квадратичные отклонения
υ	Характеристика процесса (неизвестная функция)
Φ	Сумма квадратов разностей экспериментальных и расчётных значений функции отклика

ВВЕДЕНИЕ

В науке и технологии достаточно часто возникает задача поиска оптимальных условий проведения процесса или получения математической модели процесса без исследования механизма. Другими словами, необходимо решить задачу пространственно-временной организации системы при неполном знании механизма явлений, происходящих в системе, например, при оптимизации режима проводки скважины найти такое сочетание нагрузки на инструмент, скорости вращения, расхода и реологических характеристик промывочной жидкости, типоразмера инструмента и/или гидромониторного эффекта и др., чтобы получить максимальную скорость проходки. Другой пример: физическая сущность взаимодействия компонентов буровых и цементных растворов, используемых при строительстве скважин, достаточно сложна. Для целей практического приготовления этих растворов в полевых условиях достаточно уравнений, связывающих важные характеристики (структурная вязкость, напряжение сдвига, водоотдача и др.) с составом.

Эти и множество других подобных задач можно решить по разному. Можно провести исследование физической сущности процессов, происходящих в системе: изучить термодинамику физических и химических процессов, скорости химических превращений, массопередачи, теплопередачи, особенности гидродинамики. Всё это требует углублённого изучения объекта моделирования. Полученные при этом математические модели имеют как самостоятельную научную ценность, так и практическую возможность использования их для проектирования технологических установок и анализа протекающих в них процессов. Это обусловлено тем, что выявленные связи и количественные зависимости соответствуют физической сущности процессов, происходящих в системе, вскрывают причинно-следственные связи и могут быть распространены как за пределы изученной системы, так и на другие классы подобных процессов и/или систем. Это путь научного познания природы и процессов, происходящих в ней, он достаточно длителен и дорог. Экспериментатор выбирает тот или иной путь исследования, основываясь на своём опыте и интуиции. Получаемые в результате математические модели принято называть структурными или детерминистическими.

С другой стороны, структурные математические модели малопригодны для целей автоматизированного управления и оптимизации действующих производств. Практически невозможно в таких моделях учесть все особенности конкретного технологического оборудования и сырья.

В большинстве случаев задачи автоматизации системы и оптимизации процесса могут быть решены экспериментальным путём при ограниченном знании механизма явлений в системе. В таких случаях необходима формальная математическая модель системы, так называемая экспериментально-статистическая модель, формально описывающая зависимость отклика системы на изменения входов и пренебрегающая механизмом процесса.

Основы математической теории построения регрессионных моделей были заложены К. Гауссом (1794-1795) и А. Лежандром (1805-1806). К. Гаусс рассматривал проблему замены неизвестного истинного значения величины M_x её приближённым значением m_x, вычисленным по результатам экспериментального определения. Основу метода составляет требование минимума среднего значения $E(m_x-M_x)^2$ (отсюда и название - *метод наименьших квадратов*). Этому требованию соответствует оптимальность оценки m_x в смысле наилучшего соответствия истинному значению M_x. Ущерб от замены истинного неизвестного значения M_x приближённым значением m_x, вычисленным по результатам эксперимента при отсутствии систематических ошибок, пропорционален квадрату ошибки $(m_x-M_x)^2$.

Поиск оценки математического ожидания выборки предполагает, что элементы выборки независимы и нормально распределены. Но так бывает далеко не всегда, поскольку каждое событие в мире является результатом одновременного воздействия некоторого множества детерминированных факторов. В таких случаях исследователей интересует функциональная зависимость какой-либо величины, например y, от некоторого множества других, например x_1, x_2, ..., x_k. Простейшее уравнение, описывающее некоторые процессы в природе, называется уравнением линейной множественной регрессии

$$y=b_0+b_1x_1+b_2x_2+b_3x_3+\ldots+b_kx_k,$$

а собственно метод наименьших квадратов (НК) в практике экспериментальных исследований получил своё название из-за его связи с минимизацией сумм квадратов вида

$$\Phi = \sum_{i=1}^{n}\left(y_i - f(x_1, x_2, \ldots, x_k)_i\right)^2,$$

где y_i - значение *функции отклика* в i-том эксперименте, n - число

опытов, k - число независимых переменных или *факторов*, влияние которых подлежит количественной оценке; $x_{i,j}$ - значение j-того фактора в i-том эксперименте, а $f(x_1, x_2, \ldots, x_k)$ - функция, вид которой в общем случае **неизвестен**. В некотором частном случае - модель линейной **множественной** регрессии.

1. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим некоторый технологический процесс, протекающий в системе S (рис. 1.1). Как правило, в каждую технологическую систему поступают материальные (энергетические) потоки, которые можно измерять и регулировать. Эти потоки характеризуются входными параметрами x_1, x_2,..., x_k. Соответственно из системы выходят те или иные материальные (энергетические) потоки, характеризующиеся выходными параметрами y_1, y_2, y_3,... Любая система находится под воздействием неконтролируемых, случайным образом изменяющихся параметров w_1, w_2, w_3,... (это так называемый "шум"). В качестве случайных обычно рассматриваются параметры, которые по каким-либо причинам невозможно (или очень трудно) измерить.

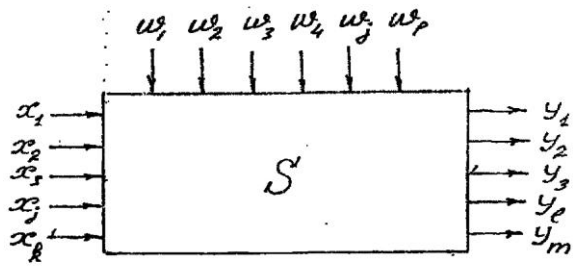

Рис. 1.1. Принципиальная схема объекта (концепция чёрного ящика).

В технологии строительства скважин, например, такими факторами будут изменения нагрузки на инструмент вследствие трения бурильной колонны о стенки скважины, флуктуации скорости вращения инструмента и расхода промывочной жидкости, различные углы встречи инструмента с прожилками твёрдых пород, вибрация бурильной колонны и др. [12]. Комплекс параметров x_1, x_2,..., x_k - основной, он определяет условия проведения процесса. Разделение входных параметров на основные и случайные достаточно условно. Случайным будет любой параметр, не вошедший в основной комплекс входных параметров.

Задача получения математической модели исследуемого процесса выглядит следующим образом: нужно получить некоторое выражение для

искомой функции:

$$y=f(x_1,\ x_2,\ldots.\ x_k), \tag{1.1}$$

где y - *функция отклика*; x_j - независимые переменные или *факторы*; k - количество факторов влияющих на процесс, причём собственно вид функции f, как правило, неизвестен. Пространство, образованное независимыми переменными, называется *факторным пространством*, а геометрический образ, соответствующий функции отклика, называется *поверхностью отклика*. Термин "функция отклика" принят для того, чтобы отличать функции, имеющие аналитическое выражение, от так называемых экспериментально-статистических зависимостей. Особенность задачи заключается в том, что исследование поверхности отклика производится при неполном знании механизма изучаемых явлений. Естественно считать, что в таком случае аналитическое выражение функции отклика неизвестно, и её представляют полиномом [1]

$$\upsilon = \beta_0 + \sum_{j=1}^{k}\beta_j x_j + \sum_{\substack{l=1\\ m=2\\ l\neq m}}^{k}\beta_{l,m}x_l x_m + \sum_{j=1}^{k}\beta_{jj}x_j^2 \ldots, \tag{1.2}$$

где υ - характеристика процесса, подлежащая оптимизации или регулированию; β_0, β_j, $\beta_{l,m}$, β_{jj},... - коэффициенты уравнения регрессии. Разложение неизвестной функции в степенной ряд эквивалентно представлению её рядом Тейлора:

$$\beta_1 = \frac{f}{x_1};\quad \beta_2 = \frac{f}{x_2};\ldots,\quad \beta_k = \frac{f}{x_k}$$

$$\beta_{1,2} = \frac{^2f}{x_1\ x_2};\quad \beta_{2,3} = \frac{^2f}{x_2\ x_3};\ldots \tag{1.3}$$

$$\beta_{1,1} = \frac{1}{2}\cdot\frac{^2f}{x^2};\quad \beta_{2,2} = \frac{1}{2}\cdot\frac{^2f}{x^2};\ldots.$$

Очевидно, что для определения коэффициентов уравнения регрессии необходимо поставить некоторое количество опытов, т.е. для нескольких комбинаций значений независимых переменных определить значения характеристики процесса υ, которые при этом обозначают латинской буквой y, и подвергнуть эти данные соответствующей матема-

тической обработке. При этом возникает следующая трудность. В ходе эксперимента неизбежны флуктуации независимых переменных x_1, x_2,..., x_k и неизбежны ошибки при измерении характеристики процесса y. Естественно, что и те и другие ошибки вносят более или менее существенное искажение в наблюдаемую картину, и в процессе обработки данных эти ошибки каким-либо образом необходимо учесть. Следующая трудность заключается в том, что экспериментатор может осуществить только некоторое конкретное, ограниченное число опытов, т.е. осуществить *выборку* из *совокупности* (под совокупностью подразумеваются все возможные опыты в исследуемом факторном пространстве). При этом экспериментатор не всегда имеет представление о форме поверхности отклика и о том, какое место занимают его опыты в факторном пространстве. Поэтому по результатам выборки можно определить только *выборочные* коэффициенты уравнения регрессии, обозначаемые латинскими буквами b_0, b_j, $b_{l,m}$, b_{jj}:

$$\hat{y} = b_0 + \sum_{j=1}^{k} b_j x_j + \sum_{\substack{l=1 \\ m=2 \\ l \neq m}}^{k} b_{l,m} x_l x_m + \sum_{j=1}^{k} b_{jj} x_j^2 \ldots \qquad (1.4)$$

где $\hat{y}$ – расчётное значение функции отклика. Выборочные коэффициенты b_0, b_j, $b_{l,m}$, b_{jj} ещё называют ***оценками*** коэффициентов β_0, β_j, $\beta_{l,m}$, β_{jj}. Для определения коэффициентов β_0, β_j, $\beta_{l,m}$, β_{jj} необходимо поставить опыты во всех точках факторного пространства, что реально невозможно. Естественно, что соответствие выборочных коэффициентов генеральным будет при удачном распределении условий опытов по объёму факторного пространства и при относительно небольших отклонениях экспериментальных значений y от неизвестной поверхности отклика v.

С познавательной точки зрения такая полиномиальная модель не представляет особого интереса. Зная оценки коэффициентов отрезка ряда Тейлора, нельзя восстановить исходную функцию, аналитическое выражение которой остаётся неизвестным исследователю, и, следовательно, невозможно получить информацию о механизме процесса. Полиномиальные модели справедливы только для объекта, на котором проводился эксперимент. Более того, в большинстве случаев эти модели справедливы только в исследованном интервале изменения факторов, и за его пределами возможны значительные расхождения с эксперимен-

тальными данными. С другой стороны, в практическом отношении полиномиальные модели очень полезны и широко используются для решения различных научных и прикладных задач, они достаточно просты в получении.

Прежде чем переходить к подробному рассмотрению метода НК, рассмотрим определения некоторых терминов. При первом чтении их можно пропустить.

Адекватность (*франц.* adequat - адекватный < *лат.* adaequatus - соответственный, тождественный, приравненный, равный; *лат.* adaequo - сравнивать, уравнивать) - соответствие, соразмерность, верность, точность, полное соответствие исследуемому предмету. В теории познания термин "адекватность" служит для обозначения верного воспроизведения объективных связей и отношений действительности в представлениях, понятиях и суждениях. В этом смысле *истина* определяется как адекватность мышления бытию.

В моделировании адекватность - количественная характеристика соответствия модели оригиналу. Критерием адекватности является однородность *дисперсий* - дисперсии *воспроизводимости* $s^2_{оп}$ и дисперсии адекватности $s^2_{ад}$:

$$s^2_{оп} = \sum_{i=1}^{n_{оп}} (y_i - \bar{y})^2 / (n_{оп} - 1); \qquad (1.5)$$

$$s^2_{ад} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n - l), \qquad (1.6)$$

где $n_{оп}$ - число опытов в *выборке* на *воспроизводимость*; n - число опытов в выборке, осуществляемой с целью получения уравнения *функции отклика*; l - *число связей*, наложенных на выборку (число параметров уравнения, определённых по выборке). Если дисперсии однородны, математическая модель адекватна, если неоднородны - неадекватна. Более или менее объективным критерием однородности дисперсий является значение опытного критерия Фишера

$$F^{оп}_{\nu_1, \nu_2} = \frac{s^2_{ад}}{s^2_{оп}}, \qquad (1.7)$$

где ν_1 - *число степеней свободы* числителя; ν_2 - число степеней свободы знаменателя. В числителе всегда должна быть большая дисперсия, в знаменателе - меньшая; дисперсия адекватности обычно больше дисперсии воспроизводимости. Это объясняется тем, что дисперсия адекватности включает в себя два вида ошибок - ошибки расчётные, обусловленные приближённым характером математической модели, и ошибки экспериментальные, а дисперсия воспроизводимости (опытная дисперсия) характеризует только ошибки эксперимента. Опытный критерий Фишера сравнивают с табличным значением критерия Фишера F^{α}, взятого с требуемым *уровнем значимости* α: - если опытный критерий Фишера меньше табличного, то дисперсии однородны и соответственно модель адекватна эксперименту, если больше - дисперсии неоднородны и модель неадекватна.

Воспроизводимость в теории и практике экспериментальных исследований - характеристика точности лабораторного или промышленного эксперимента, а также подтверждение результатов тех или иных *наблюдений* в природе и обществе другими исследователями, авторами в другое время, в тех или иных условиях. Обычно воспроизводимость характеризуется количественно (т.е. числом, в процентах или долях единицы), но может характеризовать явление или процесс и качественно. См. также *Адекватность*, *Дисперсия воспроизводимости*.

Выборка - понятие математической статистики, объединяющее результаты каких-либо однородных наблюдений [8, 16]. Выборкой в широком смысле слова называется конечная совокупность результатов наблюдений X_1, X_2,..., X_n, представляющих собой независимые одинаково распределённые случайные величины. Определённая таким образом выборка называется *случайной*, а её конкретные значения в каждом отдельном случае x_1, x_2,..., x_n - простой выборкой. В узком смысле понятие выборки связано с теорией статистического *выборочного метода* и предполагает наличие некоторой конечной *совокупности*, из которой эта выборка извлекается (рассматриваются, например повторные и бесповторные выборки). С точки зрения исследователя, осуществляющего экспериментальные исследования с целью моделирования процесса, выборкой будет называться конкретное количество анализов, опытов, измерений и т.п., а под совокупностью будет подразумеваться абстрактная бесконечность возможных анализов, опытов, измерений и т.п.

Дисперсия (< *лат.* disperse - рассеянно, разбросанно, там и сям; dispersio - рассеяние, разбросанность) в математической ста-

тистике и теории вероятностей - одна из характеристик распределения вероятностей случайной величины, наиболее употребительная мера рассеяния её значений, т.е. отклонения её от среднего; дисперсия - *центральный момент* второго порядка. В теории вероятностей дисперсия DX случайной величины X определяется как математическое ожидание $E(X-M_x)^2$ квадрата отклонения X от её математического ожидания $M_x=EX$. Для случайной величины с дискретным распределением дисперсия определяется формулой

$$DX = \sum_{i=1}^{\infty} (x_i-M_x)^2 p_i, \qquad (1.8)$$

где вероятность $p_i=P(X=x_i)$ при условии, что ряд сходится; для случайной величины X с непрерывным распределением, имеющим плотность вероятности $p(x)$, - формулой

$$DX = \int_{-\infty}^{\infty} (x-M_x)^2 p(x)\,dx, \qquad (1.9)$$

если этот интеграл сходится. Дисперсия имеет важное значение в характеристике качества статистической оценки случайной величины. Наряду с дисперсией в качестве меры рассеяния (той же размерности, что и сама случайная величина) используется квадратный корень из дисперсии $\sigma=\sqrt{DX}$, называемый *квадратичным отклонением* X. Если $DX=0$, то случайная величина X принимает с вероятностью 1 единственное значение M_x. Дисперсия обладает свойством минимальности в том смысле, что

$$DX = \min_{-\infty<a<\infty} E(X-a)^2; \qquad (1.10)$$

при этом min достигается при $a=EX$.

См. также *Адекватность, Дисперсия адекватности, Дисперсия воспроизводимости.*

Дисперсия адекватности (*от лат.* disperse - рассеянно, разбросанно, там и сям; dispersio - рассеяние, разбросанность и adaequo - сравнивать, уравнивать, adaequatus - приравненный, равный) - количественная характеристика расхождения экспериментальных значений функции отклика и значений функции отклика, рассчитанных по уравнению, параметры которого определены по *выборке*. См. также *Адекватность.*

Дисперсия воспроизводимости – количественная характеристика точности эксперимента или *воспроизводимости* результатов наблюдений в природе; вычисляется по формуле:

$$s^2_{оп} = \frac{1}{n_{оп}-1} \sum_{i=1}^{n_{оп}} (y_i - \bar{y})^2, \qquad (1.11)$$

$$\bar{y} = \frac{1}{n_{оп}} \sum_{i=1}^{n_{оп}} y_i, \qquad (1.12)$$

где $n_{оп}$ – число опытов на воспроизводимость; $\nu_{оп}=n_{оп}-1$ – *число степеней свободы* дисперсии воспроизводимости. С точки зрения *метода моментов* дисперсия воспроизводимости является центральным моментом второго порядка и может быть вычислена по формулам:
для дискретной случайной величины

$$\mu_2 = \sum_{i=1}^{n} (x_i - M_x)^2 p_i = s^2_x, \qquad (1.13)$$

для непрерывной –

$$\mu_2 = \int_{-\infty}^{+\infty} (x_i - M_x)^2 p(x)dx = s^2_x. \qquad (1.14)$$

Для оценки достоверности результатов наблюдений одной дисперсии воспроизводимости недостаточно. Корень квадратный из дисперсии воспроизводимости называется *квадратичным отклонением* или *стандартным отклонением*. Начинающие исследователи обычно с трудом развивают интуитивное восприятие численного значения дисперсии или стандартного отклонения. Является ли дисперсия воспроизводимости, равная, например, 77, большой или малой? Что значит *стандартное отклонение* $0,51 \cdot 10^{-4}$? Оказывается, для интерпретации как дисперсии воспроизводимости, так и стандартного отклонения главное не получить численные значения последних, а правильно сравнить дисперсию воспроизводимости с какой-либо другой дисперсией, например *дисперсией адекватности*, или стандартное отклонение умножить на правильно выбранный *критерий Стьюдента*, чтобы получить *доверительный интервал* для неизвестного *математического ожидания*. На начальных этапах исследований стандартное отклонение $\pm s_y$ сравнивают со средним значением выборки $y_{ср}$. Если $|s_y| < y_{ср}$, то говорят о *значимом* отличии результатов наблюдений от нуля и **предварительную** *оценку* точности экспери-

мента осуществляют по отношению $s_y/y_{ср}$. Если это отношение лежит в пределах 3÷4%, то в первом приближении результаты наблюдений считают воспроизводимыми; в противном случае необходимы дополнительные изыскания.

Доверительная вероятность - вероятность достоверности принимаемой гипотезы; характеристика надёжности, полученной по *выборке оценки* того или иного параметра,

$$P=P\{|\beta-b|<\varepsilon_p\}, \qquad (1.15)$$

где β - генеральный параметр; b - его оценка; $\varepsilon_p=f(p)$ - ошибка определения генерального параметра; P - вероятность настолько большая, что событие $|\beta-b|<\varepsilon_p$ можно считать практически достоверным. Очевидно, что диапазон возможных с вероятностью P значений ошибки от замены β на b равен $\pm\varepsilon_p$. Вероятность появления ошибок больших по абсолютной величине, чем ε_p, или вероятность событий $|\beta-b|>\varepsilon_\beta$ называется *уровнем значимости:*

$$\alpha=1-P=P\{|\beta-b|>\varepsilon_p\}. \qquad (1.16)$$

Иначе выражение (1.15) может быть интерпретировано как вероятность того, что истинное значение параметра β находится в пределах:

$$b-\varepsilon_p<\beta<b+\varepsilon_p, \qquad (1.17)$$

где выборочный параметр b - по существу *случайная величина*, а ошибка его определения $(\beta-b)$ в выражениях (1.15) и (1.16) - также случайная величина. Интервал $I_p=b\pm\varepsilon_p$ называется *доверительным интервалом*. Границы интервала $b_{min}=b-\varepsilon_p$ и $b_{max}=b+\varepsilon_p$ называются *доверительными границами*. Доверительный интервал при принятой доверительной вероятности определяет точность оценки. Величина доверительного интервала зависит от принимаемой доверительной вероятности, т.е. от той вероятности, с которой гарантируется нахождение искомого параметра β внутри доверительного интервала; другими словами: чем выше гарантия надёжности оценки, тем больше величина интервала, в котором может находится генеральный параметр. См. также *Уровень значимости*.

Доверительное отклонение - *функция* от результатов наблюдений и *доверительной вероятности*, позволяющая оценить *доверительный интервал*, который с вероятностью $P=1-\alpha$ "накрывает" неизвестное значение параметра:

$$\frac{s_x}{\sqrt{n}} \cdot t_\nu^\alpha, \qquad (1.18)$$

где s_x – *квадратичное (стандартное) отклонение*; n – количество наблюдений в выборке; t – случайная величина, зависящая только от *числа степеней свободы* ν выборочной дисперсии и *уровня значимости* α, называемая *критерием Стьюдента*. Необходимо отметить, что доверительное отклонение не зависит ни от *математического ожидания* M_x, ни от *генерального параметра* $б_x$. Это *случайная величина*, зависящая только от квадратичного отклонения s_x и принимаемого исследователем *уровня значимости* α. См. также *Доверительная вероятность, Доверительные границы.*

Доверительные границы – см. *Доверительная вероятность, Доверительное отклонение, Доверительный интервал.*

Доверительный интервал – статистическая оценка параметра исследуемого вероятностного распределения, имеющая вид интервала, границами которого служат функции от результатов наблюдений и *доверительной вероятности*, который с вероятностью P "накрывает" неизвестное значение параметра. Дело в том, что значение оценки в каждом конкретном случае может отличаться от *истинного* значения параметра (*математического ожидания*), и, следовательно, в интерпретации результатов эксперимента имеется некоторая доля неопределённости. При грубых оценках величина этой неопределённости выражается с помощью *выборочной дисперсии* или *квадратичного отклонения*, т.е. вполне возможно, что неизвестный *генеральный параметр* χ находится в интервале $x_{ср} \pm s_x$; также возможно, что он находится в интервале $x_{ср} \pm 2s_x$ и т.д. Другими словами, для верной интерпретации результатов эксперимента важно установить для оценки генерального параметра χ **интервал** вместо отдельной точки $x_{ср}$, причём хотя бы одна точка этого интервала, а именно $x_{ср}$, и рассматривалась бы как "наилучшая" оценка для χ. Задача определения доверительного интервала решалась бы достаточно просто, если бы был известен закон распределения оценки $x_{ср}$ или $y_{ср}$ (1.12):

$$P\{|M_x - \bar{x}| < \varepsilon_p\} = \int_{-\varepsilon_p}^{+\varepsilon_p} p(x)\,dx = P. \qquad (1.19)$$

Также просто решалась бы эта задача при известной генеральной дисперсии $б^2_x$ (знание генеральной дисперсии $б^2_x$ позволяет оценивать доверительный интервал даже по одному наблюдению), но, к сожалению, генеральную дисперсию $б^2_x$ невозможно получить из наблюдений, её можно только оценить при помощи выборочной дисперсии s^2_x. Например, для выборки объёма n:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad s^2_x = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

где $x_{ср}$ – оценка математического ожидания M_x; s^2_x – оценка генеральной дисперсии $б^2$. Известно, что *нормальное распределение* наиболее широко распространено в природе. Известно также, что распределение случайных ошибок наблюдений подчиняется закону нормального распределения. Задача оценки доверительного интервала для математического ожидания нормально распределённой случайной величины была решена в 1908 году У. Госсетом, известным под псевдонимом Стьюдент:

$$\bar{x} - \frac{s_x}{\sqrt{n}}\cdot t < m_x < \bar{x} + \frac{s_x}{\sqrt{n}}\cdot t, \qquad (1.20)$$

где t – критерий Стьюдента, случайная величина, зависящая только от *числа степеней свободы* ν выборочной дисперсии и *уровня значимости* α. Из оценки доверительного интервала видно, что уменьшение доверительного интервала обратно пропорционально корню квадратному из числа наблюдений; другими словами, для того, чтобы повысить точность результатов наблюдений в два раза, необходимо увеличить объём выборки в четыре раза. См. также *Доверительная вероятность. Доверительное отклонение, Доверительные границы, Стьюдента критерий, Хи-квадрат критерий.*

Квадратичное отклонение, *квадратичное уклонение*, величин $x_1, x_2, \ldots, x_n$ от a – квадратный корень из выражения

$$\frac{(x_1-a)^2 + (x_2-a)^2 + \ldots + (x_n-a)^2}{n}. \qquad (1.21)$$

Наименьшее значение квадратичное отклонение имеет при $a=x_{ср}$, где $x_{ср}$ – среднее арифметическое величин $x_1, x_2, \ldots, x_n$.

$$x_{ср} = \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}. \qquad (1.22)$$

Употребляется также более общее понятие *взвешенного* квадратичного отклонения, определяемого как квадратный корень из выражения

$$\frac{W_1(x_1-a)^2 + W_2(x_2-a)^2 + \ldots + W_n(x_n-a)^2}{W_1 + W_2 + \ldots + W_n}, \qquad (1.23)$$

где числа W_1, W_2, ..., W_n называют при этом *весами*, соответствующими величинам x_1, x_2, ..., x_n. Взвешенное квадратичное отклонение достигает наименьшего значения при a, равном взвешенному среднему

$$\frac{W_1x_1 + W_2x_2 + \ldots + W_nx_n}{W_1 + W_2 + \ldots + W_n}. \qquad (1.24)$$

Такое представление о квадратичном отклонении соответствует использованию квадратичного отклонения в теории ошибок.

В теории вероятностей *квадратичное отклонение* $б_X$ случайной величины X (от её математического ожидания) определяют как квадратный корень из дисперсии $\sqrt{DX}$ и называют также *стандартным отклонением* величины X. Для любой случайной величины X с математическим ожиданием M_X и квадратичным отклонением $б_X$ вероятность отклонений X от M_X, больших по абсолютной величине $kб_X$, $k>0$, не превосходит $1/k^2$. В случае нормального распределения указанная вероятность при $k=1$ равна 0,3174, при $k=2$ - равна 0,0456 и при $k=3$ - равна 0,0027. В практических задачах, приводящих к нормальному распределению, отклонения больше, чем утроенный стандарт (*квадратичное отклонение*), практически невозможны, или, другими словами, на практике пренебрегают возможностью отклонений от среднего, больших $3б_X$ (*правило трёх сигма*).

В математической статистике квадратичное отклонение употребляют как меру качества статистических оценок и называют в этом случае *квадратичной погрешностью* (ошибкой). См. также *Стандартное отклонение*.

Корреляционный анализ - совокупность основанных на математической теории *корреляции* методов обнаружения функциональной зависимости между случайными величинами или признаками. Достаточно часто между случайными величинами существует такая связь, что с изменени-

ем одной величины меняется распределение другой. В случае парных зависимостей изменение случайной величины y, соответствующее изменению величины x, разбивается при этом на две компоненты: функциональную (связанную с зависимостью $y=f(x)$) и случайную. Если первая компонента отсутствует, то величины y и x независимы; в случае наличия зависимости $y=f(x)$ большая или меньшая точность эксперимента и измерений обусловливают больший или иеньший вклад *случайности* в общую картину результатов наблюдений, вплоть до затушёвывания следов существующей функциональной зависимости. Соотношение между функциональной и случайной компонентами определяет силу (тесноту) связи. Важнейшим показателем этой связи является *коэффициент корреляции* (1.25). Процедура корреляционного анализа включает в себя: построение *корреляционного поля* и построение корреляционной таблицы; вычисление выборочных коэффициентов корреляции и корреляционных отношений; проверку статистической гипотезы о значимости (силе) связи. При подтверждении корреляционной связи производится *регрессионный анализ* с целью установления вида функциональной зависимости $y=f(x)$. См. также *Корреляция*.

Корреляционное поле - система координат y-x с нанесёнными на координатную плоскость двумерных выборочных точек. *Корреляционное поле* - по существу график экспериментальной зависимости $y=f(x)$ (если таковая имеется). По виду расположения точек поля можно сделать предварительное заключение о наличии и форме зависимости случайных величин Y и X. Данные после визуального анализа группируют в виде корреляционной таблицы для последующей численной обработки. См. также *Корреляция, Корреляционный анализ, Регрессионный анализ*. .

Корреляция (от позднелат. correlatio - соотношение) - вероятностная или статистическая зависимость между двумя и более случайными величинами. В отличие от функциональной зависимости, корреляция возникает тогда, когда зависимость одного признака (случайной величины Y) от другого (от другой случайной величины X) осложняется случайными *факторами* или когда среди условий, от которых зависят две случайные величины, например, Y и Z, имеются общие для них обеих условия X_1, X_2,... Такого рода зависимости иногда выявляются визуально с помощью графиков и последующего *регрессионного анализа*, иногда корреляционная зависимость не очевидна, и её выявляют с помощью *корреляционного анализа*. Необходимо отметить, что статистическая зависимость, как бы ни была она сильна, никогда не может ус-

тановить *причинной связи:* предположения о причинах и *следствиях* следует искать вне статистики, например, в сфере физических сущностей явления.

В основе математической теории корреляции лежит предположение о том, что все явления в природе в большей или меньшей мере подчинены определённым вероятностным закономерностям. Зависимость между двумя случайными событиями проявляется в том, что условная вероятность наступления одного из них при наступлении другого отличается от безусловной вероятности. Или, другими словами, влияние одной из них, например X, на другую, например Y, не вполне конкретно, не жёстко функционально, хотя при возрастании случайной величины X другая, Y, имеет тенденцию возрастать или убывать. Это характеризуется тем, что при каждом фиксированном значении X в результате наблюдений получается множество значений Y, характеризующееся некоторым распределением вероятностей. В таких случаях *функция* $y=f(x)$ называется *регрессией* величины Y по X, а график такой зависимости называется *линией регрессии* Y по X. См. также *Коэффициент кореляции, Корреляционное поле.*

Коэффициент корреляции - количественная характеристика связи между случайными величинами, например, характеристикой связи случайных величин X и Y является безразмерный коэффициент кореляции

$$\rho_{xy} = \frac{E[(X-M_x)(Y-M_y)]}{б_x б_y}, \qquad (1.25)$$

где $б_x$ и $б_y$ - генеральные квадратичные отклонения величин X и Y; E - математическое ожидание. Коэффициент корреляции характеризует не всякую зависимость, а только линейную. Линейная вероятностная зависимость случайных величин заключается в том, что при возрастании одной случайной величины другая имеет тенденцию возрастать или убывать по линейному закону. Коэффициент корреляции характеризует степень тесноты линейной зависимости. Если случайные величины X и Y связаны точной линейной функциональной зависимостью $y=b_0+b_1x$, то $\rho_{xy}=\pm1$, причём знак соответствует знаку коэффицента b_1. В общем случае, когда величины X и Y связаны произвольной вероятностной зависимотью, коэффициент корреляции может иметь значение в пределах $-1<\rho_{xy}<+1$. При $\rho_{xy}>0$ существует положительная корреляционная связь между величинами X и Y, при $\rho_{xy}<0$ - отрицательная. Для независимых случайных величин $\rho_{xy}=0$.

Поскольку в результате эксперимента исследователь получает случайную выборку, то по результатам выборки определяется *выборочный* коэффициент корреляции r_{xy}. Выборочный коэффициент корреляции определяется так же, как и генеральный коэффициент ρ_{xy}, только при этом используются выборочные средние $x_{ср}$ и $y_{ср}$ и выборочные дисперсии s^2_x и s^2_y:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y} =$$

$$= \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] \cdot \left[n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \qquad (1.26)$$

Число степеней свободы для найденного таким путём коэффициента корреляции r_{xy} равно $\nu = n-2$, поскольку вычисление $x_{ср}$ и $y_{ср}$ соответствует наложению на выборку двух связей.

Коэффициент корреляции одинаково отмечает и слишком большую долю случайности (большие ошибки наблюдений), и значительную криволинейность этой связи. Между величинами X и Y может быть достаточно определённая нелинейная функциональная зависимость, а коэффициент корреляции всё же будет меньше единицы. В таких случаях можно произвести *преобразование факторного пространства* с целью нахождения системы координат, в которой линеаризуется исходная экспериментальная зависимость.

Необходимо отметить, что статистическая зависимость, как бы ни была она сильна, никогда не может установить причинной связи: наши идеи о причине должны приходить извне статистики, в конечном счёте из некоторой другой теории. Например, очевидна связь между количеством выпавших дождей и величиной урожая - дожди влияют на урожай и, совершенно определённо, урожай не воздействует на дожди. Но нет никаких ***статистических причин*** для отказа от идеи зависимости дождей от урожая: отказ сделан на основе совершенно других соображений. И даже если бы дожди и урожай были в полном функциональном соответс-

твии, то всё равно нет оснований для *обращения* этой очевидной причинной связи. См. также *Корреляция*.

Математическое ожидание, среднее значение - наиболее вероятное значение случайной величины. Математическое ожидание - одна из важнейших числовых характеристик распределения *вероятностей* случайной величины. Для случайной величины X, принимающей последовательность значений $x_1, x_2, \ldots, x_i, \ldots$ с вероятностями, равными соответственно $p_1, p_2, \ldots, p_i, \ldots$, *математическое ожидание* определяется:

для дискретной случайной величины при условии, что ряд сходится абсолютно, - формулой

$$M_x = \sum_{i=1}^{\infty} x_i p_i; \qquad (1.27)$$

для случайной величины X с непрерывным распределением, имеющим плотность вероятности $p(x)$, если интеграл сходится абсолютно, - формулой

$$M_x = \int_{-\infty}^{\infty} x p(x)\,dx. \qquad (1.28)$$

Название "математическое ожидание" происходит от понятия "ожидаемого значения выигрыша" (математического ожидания выигрыша), впервые появившегося в теории азартных игр в трудах Б. Паскаля и Х. Гюйгенса в XVII в. Но впервые в полной мере это понятие было оценено и использовано П.Л. Чебышевым (сер. XIX в.); термин "**М.О.**" ввёл П. Лаплас (1795).

Несмещённая оценка - статистическая оценка параметра распределения вероятностей по результатам наблюдений, лишённая систематической ошибки. Например, если результаты наблюдений $x_1, x_2, \ldots, x_n$ являются взаимно независимыми случайными величинами, имеющими одинаковое нормальное распределение, заданное плотностью

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(M_x - x)^2/2\sigma^2} \qquad (1.29)$$

с неизвестными параметрами M_x и σ^2, то среднее арифметическое:

$$\bar{x} = (x_1 + x_2 + \ldots + x_n)/n; \qquad (1.30)$$

будет несмещённой оценкой для M_x. Используемая ранее для оценки δ^2 выборочная дисперсия

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (1.31)$$

не является несмещённой оценкой, так как среднее арифметическое само зависит от элементов выборки. Для устранения *смещения оценки* нужно *число степеней свободы* в выражении для s^2_x уменьшить на единицу. Несмещённой оценкой для δ^2 служит

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \qquad (1.32)$$

Опытная дисперсия – см. *Дисперсия воспроизводимости.*

Оценка – количественная характеристика параметра, получаемая по результатам *выборки*. К оценкам параметров предъявляется комплекс требований. Важнейшие среди них – *несмещённость*, *состоятельность* и *эффективность*.

Параметр (от греч. *παραμετρεω* – мерить что-либо, сопоставляя его с чем-либо, измерять что-либо по чему-либо, сравнивать что-либо по чему-либо) – величина, значения которой служат для различения элементов некоторого множества между собой. В зависимости от конкретного множества различают следующие параметры.

В математической статистике параметр – характеристика совокупности, например, математическое ожидание, дисперсия. Параметры совокупностей обычно обозначают греческими буквами, в отличие от их оценок, вычисляемых по результатам выборок и обозначаемых латинскими буквами.

В математическом моделировании параметр – величина, значения которой служат для конкретизации той или иной математической модели, например, уравнения Антуана и Андраде математически изоморфны, но первое описывает температурную зависимость давления насыщенных паров жидкости, а второе – коэффициента динамической вязкости:

$$p = \exp\left(A + \frac{B}{T+C}\right); \qquad \mu = \exp\left(A + \frac{B}{T+C}\right),$$

где в первом случае B – аналог теплоты конденсации, а во втором –

энергия активации вязкого течения; коэффициент A - предэкспоненциальный множитель; коэффициент C - формальный параметр.

Параметр в физическом моделировании и в технике - величина, являющаяся существенной характеристикой системы, технического устройства, явления или процесса. Например, в гидромеханических процессах такими величинами являются коэффициент динамической вязкости жидкой фазы, плотности жидкой и твёрдой фаз, размеры и коэффициент формы частиц твёрдой фазы и др.; для тепловых процессов такими параметрами являются удельные теплоёмкость и теплопроводность, температурный напор и т.д. Параметры могут быть постоянными и переменными (т.е. могут зависеть от времени и/или системы координат).

Регрессионный анализ (от лат. regressio - обратное движение, отход, повторение) - раздел математической статистики, объединяющий практические методы исследования регрессионной зависимости между величинами по статистическим данным (см. *Регрессия*). Проблема регрессии в математической статистике характерна тем, что о распределении изучаемых величин **нет достаточной информации.** Цель регрессионного анализа состоит в проверке *статистических гипотез* о регрессии, а содержание заключается в получении уравнения регрессии, более или менее отражающего физическую сущность изучаемого явления, и вычислении статистических оценок неизвестных параметров, входящих в уравнение регрессии. Наибольшее распространение в науке и технике имеет регрессионный анализ парной зависимости вида $y=f(x)$ вследствие простоты анализа и наглядности в графической интерпретации. При этом величина X рассматривается как *независимая переменная величина* или *фактор*, а величина Y - как зависимая переменная величина или *функция*. Нередки случаи, когда выбор переменных (X,Y) или (Y,X) достаточно произволен с точки зрения формального *описания*, но с точки зрения регрессионного анализа это имеет большое значение (см. ниже). При изучении связи между двумя величинами по результатам наблюдений $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ в соответствии с теорией регрессии предполагается, что одна из них Y со значениями y имеет некоторое *распределение вероятностей* при фиксированном значении другой переменной x с некоторыми *математическим ожиданием* и *дисперсией*. Выбор модели регрессии определяется предположениями о форме зависимости $y=f(x)$. Для установления связей между величинами y и x в эксперименте используется модель, основанная на упрощённых, но правдоподобных допущениях: величина x является **контролируемой** вели-

чиной, значения которой заранее задаются при планировании эксперимента, а наблюдаемые значения Y представимы в виде

$$y_i=\varphi(x_i,\ \beta)+\varepsilon_i,\ \ i=1,\ 2,\ \ldots,\ n, \qquad (1.33)$$

где величины ε_i характеризуют ошибки, независимые при различных измерениях и одинаково распределённые с нулевым средним и постоянной дисперсией σ^2. В действительности невозможно осуществить эксперимент, в котором отсутствовала бы ошибка измерения или поддержания на требуемом уровне независимой переменной X. Случай **неконтролируемой** переменной x (или переменная X со значениями x также имеет некоторое распределение вероятностей) отличается тем, что результаты наблюдений $(x_1,\ y_1),\ (x_2,\ y_2),\ldots,\ (x_n,\ y_n)$ представляют собой выборку из *совокупности* с некоторым двумерным распределением вероятностей. И в том, и в другом случае регрессионный анализ проводится одним и тем же способом, однако интерпретация результатов существенно различается (если обе исследуемые величины случайны, то *регрессионный анализ* в значительной степени дополняется методами *корреляционного анализа*).

Исследование регрессии, как правило, начинается с построения диаграммы рассеяния (называемого также *корреляционным полем*) точек $(x_i,\ y_i)$. Во всех случаях по форме графика $y=\varphi(x)$ можно получить предварительное представление о силе зависимости y от x или об отсутствии оной. Визуальный анализ формы графика $y=\varphi(x)$ исключительно важен, и его значение невозможно переоценить. Например, если расположение этих *точек* на графике близко к прямолинейному, то допустимо использовать в качестве приближения линейную регрессию; если вид корреляционного поля ближе к параболе, гиперболе, следует попробовать полиномиальную модель; если зависимость имеет более сложный характер (например, имеются точки максимума и (или) минимума, участки стабилизации), то чисто статистический *регрессионный анализ* недостаточно эффективен, целесообразно подобрать *детерминистическую модель*, *отражающую* физическую сущность процесса. Стандартный метод оценки линии регрессии основан на использовании полиномиальной модели ($n\geqslant 1$):

$$y=b_0+b_1x+b_2x^2+b_3x^3+\ldots+b_nx^n. \qquad (1.34)$$

Оценки $b_0,\ b_1,\ldots,\ b_n$ принято вычислять методом *наименьших квадратов*. Многочлен (1.34) характеризует так называемую *эмпирическую ли-*

нию регрессии, которая служит статистической оценкой **неизвестной истинной** линии регрессии.

Регрессионный анализ является одним из наиболее распространенных методов обработки результатов наблюдений при изучении зависимостей в биологии, химии, физике, экономике, технике, в различных технологиях и других областях. См. также *Регрессия*.

Регрессия (от лат. regressio – обратное движение, отход, повторение) в теории вероятностей и математической статистике – зависимость среднего значения какой-либо величины от некоторой другой величины или от нескольких величин. В отличие от функциональной зависимости $y=f(x)$, когда каждому значению независимой переменной x соответствует только одно значение величины y, при **регрессионной связи** одному и тому же значению x соответствует несколько значений величины y (конкретно столько, сколько раз производилось наблюдение, измерение, определение и т.п. величины y при фиксированном значении независимой величины x). Если при каждом значении $x=x_1$ наблюдается n_1 значений $y_{1,1}$, $y_{1,2}$, ..., $y_{1,n1}$ величины y, то зависимость средних арифметических

$$\bar{y}_1 = (y_{1,1}+y_{1,2}+\ldots+y_{1,n1})/n_1 \qquad (1.35)$$

от x_1 является **регрессией** в статистическом понимании этого термина. В пределе $n_1=1$. Методология регрессии основана на допущении "абсолютной" точности фиксации и "независимости" переменной X, на предположении, что случайные величины X и Y с заданным совместным распределением вероятностей связаны вероятностной зависимостью: при каждом фиксированном значении $X=x$ величина Y является случайной величиной с определённым (зависящим от значения x) условным распределением вероятностей. Уравнение $y=f(x)$, в котором x играет роль "независимой" переменной, называется **уравнением регрессии**, а соответствующий график – **линией** или **кривой регрессии** величины Y по X; переменная x называется **регрессионной переменной** или **регрессором**.

В практике экспериментальных исследований объём *данных* всегда ограничен, т.е. всегда есть дефицит информации о форме совместного распределения вероятностей, и, соответственно, более или менее реальна только задача нахождения уравнения ***приближённой*** регрессией. Основная задача исследователя при этом заключается в выборе из набора возможных функций $f(x)$, принадлежащих заданному классу, такой функции, которая минимизирует математическое ожидание $E(Y-f(x))^2$.

Такая функция называется **средней квадратической регрессией.**

Первоначально, термин "регрессия" был употреблён Ф. Гальтоном (1886) в теории наследственности в следующем специальном смысле: "возвратом к среднему состоянию" (regression to mediocrity) было названо явление, состоящее в том, что дети тех родителей, рост которых превышает среднее значение на a единиц, имеют в среднем рост, превышающий среднее значение меньше, чем на a единиц. С точки зрения терминологии термин "регрессия" следует считать неудачным.

Совокупность - понятие теории статистического *выборочного метода*. В математической статистике совокупностью называется множество каких-либо однородных элементов, из которого по определённому правилу выделяется некоторое подмножество, называемое *выборкой*. Например, при приёмочном статистическом контроле в роли совокупности выступает множество всех изделий, подлежащее общей характеризации. В простейших случаях контролируемая выборка извлекается из совокупности случайно (наугад), что с точки зрения теории вероятностей означает: если совокупность содержит N элементов и отбирается выборка из n элементов ($n<N$), то выбор должен быть осуществлён таким образом, чтобы для любой группы из n элементов вероятность быть извлечённой равнялась $n!(N-n)!/N!$.

В практике экспериментальных исследований и в математической статистике выборкой из совокупности принято также называть результаты измерений какой-либо физической величины, подверженной случайным ошибкам. В этом случае под совокупностью подразумеваются все возможные значения физической величины. Для решения практических задач бесконечное множество значений интереса не представляет; практический интерес представляют те или иные характеристики соответствующей функции распределения $F(x)$. В этом случае выборка из бесконечной совокупности представляет собой наблюдаемые значения нескольких случайных величин, по которым определяются необходимые параметры.

Состоятельность оценки - статистическая оценка параметра распределения вероятностей, обладающая тем свойством, что при увеличении числа наблюдений вероятность отклонений оценки от оцениваемого параметра на величину, превосходящее некоторое наперёд заданное число, стремится к нулю. Оценка параметра называется состоятельной, если по мере роста числа наблюдений $n\to\infty$ она стремится к математическому ожиданию оцениваемого параметра. Так, выборочное среднее и

выборочная дисперсия представляют собой состоятельные оценки соответственно математического ожидания и дисперсии нормального распределения.

Стандартное отклонение или **стандарт** - то же, что *квадратичное отклонение*. Стандартным отклонением в теории и практике экспериментальных исследований принято называть корень квадратный из выражений (1.5), (1.11), (1.32), (1.36), (1.43) и т.п. См. также *Дисперсия воспроизводимости*

Стандартные границы корреляционного поля - участок корреляционного поля $y=f(x)$, в пределах которого располагаются экспериментальные точки, ошибка определения которых не превышает соответствующего *стандартного отклонения*, вычисленного по всему массиву наблюдений $y=f(x)$. Стандартные границы вычисляются для каждой точки зависимости $y=f(x)$ с помощью стандартных отклонений параметров принятого уравнения. Естественно, что стандартные границы корреляционного поля зависят не только от точности эксперимента, но и от вида уравнения регрессии, принятого для обработки данных (см. табл.2.2). Например, для уравнения $y=b_0+b_1x$ стандартные границы для каждой i-той точки вычисляются по уравнению $y_1=(b_0\pm s_{b0})+(b_1\pm s_{b1})x_1$. Крайние значения, $y_{1,min}$ и $y_{1,max}$, и формируют стандартные границы, за пределами которых располагаются точки, ошибка определения которых превышает некоторую величину, являющуюся функцией результатов наблюдений и принятого уравнения регрессии. Для одного и того же массива данных $y=f(x)$ конфигурация и величина стандартных границ непосредственно зависят от вида уравнения регрессии, принимаемого для обработки.

Начинающие исследователи обычно с трудом принимают решение о том, какие экспериментальные точки "выпадают" из той или иной зависимости, а какие нет. Иногда задача "отсева" осложняется множественностью версий о виде зависимости $y=f(x)$. Интуитивное представление о выпадающих точках формируется в процессе анализа достаточно большого количества данных. В случаях формального выбора оптимальной формы из некоторого множества подходящих форм уравнений парной регрессии полезно одновременно с анализом степени линеаризации обращать внимание на стандартные границы корреляционного поля.

Аналогично переходу от стандартного отклонения к *доверительному отклонению* с помощью *критерия Стьюдента*, можно говорить о построении *доверительных границ* корреляционного поля:

$$y_i = (b_0 \pm s_{b_0} t^{\alpha}/\sqrt{n}) + (b_1 \pm s_{b_1} t^{\alpha}/\sqrt{n}) x_i .$$

Естественно, такая процедура усложняет расчёты, но, в отличие от стандартных границ корреляционного поля, в этом случае можно говорить с доверительной вероятности $P=1-\alpha$ достоверности отсева. Кроме этого, вычисление доверительных границ корреляционного поля исключает парадоксы, когда большая часть или даже все экспериментальные "точки" оказываются за пределами стандартных границ корреляционного поля. Такие случаи возможны, если дисперсии коэффициентов вычисляются на основе дисперсии адекватности.

Статистических гипотез проверка (*греч.* **υποθεβις** - основание, принцип, предположение, **гипотеза**; **υποτιθημι**- полагать в основание, принимать что-либо за основание, предполагать) - один из основных разделов математической статистики, объединяющий методы проверки соответствия *статистических данных* некоторой статистической *гипотезе* (гипотезе о вероятностной природе данных). Проблема проверки гипотезы статистическими методами возникает в тех случаях, когда гипотезу нельзя проверить непосредственно и приходится довольствоваться проверкой некоторых следствий, которые логически вытекают из содержания гипотезы. Если следствия, вытекающие из предполагаемой гипотезы, невозможны или противоречат физической сущности процесса, значит гипотеза неверна. С другой стороны, если те или иные события невозможны или возможны с очень малой вероятностью, но всё-таки происходят, то гипотезу также приходится отвергать. Очевидно, что, придерживаясь подобной логики рассуждений, с некоторой вероятностью можно отвергнуть гипотезу, а выявить физическую сущность изучаемого явления или показать что же происходит на самом деле невозможно.

Процедуры проверки статистических гипотез позволяют принимать или отвергать статистические гипотезы, возникающие при обработке или интерпретации результатов наблюдений (результатов измерений переменных). **Правило**, в соответствии с которым принимается или отклоняется та или иная гипотеза, называется *статистическим критерием*. Проверка статистических гипотез начинается с формулировки подходящей гипотезы об исследуемой переменной. Обычно такая гипотеза называется нулевой, обозначается H_0 и по существу является гипотезой об отсутствии различия. Например при определении оценки m_x неизвестного математического ожидания M_x подходящей нулевой гипотезой будет

предположение об отсутствии различия между нулём и арифметическим средним выборки, $x_{ср}$. Математически это записывается в виде

$$H_0: x_{ср}=0.$$

Подходящей альтернативной гипотезой в этом случае будет неравенство

$$H_1: x_{ср}\neq 0.$$

Построение критерия определяется выбором подходящей функции $\theta=f(X_1, X_2, \ldots, X_n)$ от результатов наблюдений $X_1, X_2, \ldots, X_n$, которая служит мерой расхождения между опытными и *гипотетическими* значениями. Функция θ является *случайной величиной* и называется *статистикой* критерия. Центральный момент при выборе функции θ заключается в том, что её теоретическое распределение вероятностей может быть вычислено **независимым** путём, при общем допущении, что проверяемая гипотеза верна, и что её распределение **не зависит** от характеристик гипотетического распределения. Распределение статистики θ табулируется для различных *чисел степеней свободы* ν и *уровней значимости* α. С помощью этого распределения находится *критическое значение* θ^{α} такое, что если гипотеза верна, то вероятность события $\theta>\theta^{\alpha}$ равна α (рис. 1.2).

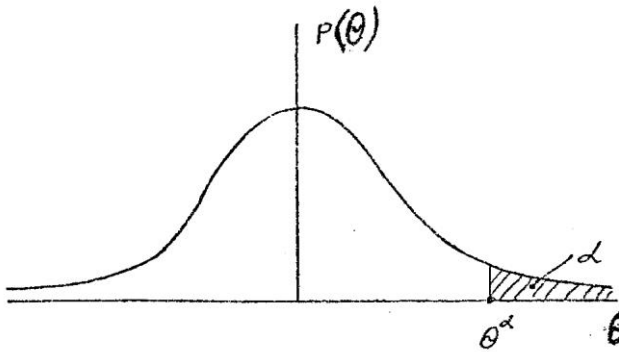

Рис. 1.2. Плотность распределения вероятностей гипотетического критерия θ. Заштрихована критическая область, составляющая α площади под кривой
(общая площадь под кривой плотности вероятностей равна 1)

Область значений $(x_1, x_2, \ldots, x_n)$, для которых $\theta(x_1, x_2, \ldots, x_n)>\theta^{\alpha}$, называется *критической областью*, это область отклонения гипотезы H_0. Если в конкретном случае обнаружится, что $\theta>\theta^{\alpha}$, то гипо-

теза отвергается; при этом считается, что *значимость* этого расхождения равна α, а *доверительная вероятность* правильности отклонения гипотезы равна $P=1-\alpha$. Если в конкретном случае $\theta<\theta^{\alpha}$, то считается, что с вероятностью P гипотеза верна. Такого рода критерии используются как для проверки параметров распределения на *значимость*, так и для проверки гипотез о самих распределениях.

В случае проверки гипотезы об отсутствии различия между нулём и оценкой математического ожидания m_x конструируется критерий в виде отношения арифметического среднего $x_{ср}$ и его стандартного отклонения s_x

$$\theta_{оп}=\frac{x_{ср}}{s_x}.$$

Очевидно, что должно быть некоторое критическое соотношение θ^{α} найденной оценки $x_{ср}$ и ошибки её определения s_x, до которого можно было бы утверждать о незначимом отличии $x_{ср}$ от нуля, а в случае превышения - о достоверности $x_{ср}$ с той или иной доверительной вероятностью $P=1-\alpha$. Очевидно также, что критерий θ^{α} должен вычисляться независимым путём.

Другой пример,- пусть имеется гипотеза о физической сущности некоторого *процесса* Y_1, Y_2, Y_3,...: для подтверждения этой гипотезы необходимо определить *параметры математической модели*, соответствующей природе изучаемого процесса, и проверить математическую модель на *адекватность*. В общем случае в результате экспериментов необходимо получить две *выборки*: выборку размерности n для определения собственно параметров математической модели y_1, y_2,..., y_n и выборку размерности $n_{оп}$ для оценки точности эксперимента $y_1, y_2, \ldots, y_{n_{оп}}$:

$$s^2_{оп}=\frac{1}{n_{оп}-1}\cdot\sum_{i=1}^{n_{оп}}(y_i-\bar{y})^2; \qquad (1.36)$$

$$\bar{y}=\frac{1}{n_{оп}}\cdot\sum_{i=1}^{n_{оп}}y_i, \qquad (1.37)$$

где $s^2_{оп}$ - несмещённая *оценка генеральной дисперсии* или выборочная дисперсия (1.5). В этом случае подходящей нулевой гипотезой H_0 будет гипотеза об отсутствии различия между дисперсией, характеризующей соответствие расчёта и эксперимента, и дисперсией, характеризующей точность собственно эксперимента, т.е. дисперсией адекватности

$s^2_{ад}$ и дисперсией воспроизводимости:

$$F^{оп}_{\nu_1,\nu_2} = \frac{s^2_{ад}}{s^2_{оп}}; \qquad (1.38)$$

$$s^2_{ад} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \qquad (1.39)$$

где F – *критерий Фишера*; $s^2_{ад}$ – *дисперсия адекватности*; ν_1 – *число степеней свободы* числителя (в данном случае *дисперсии адекватности*); ν_2 – число степеней свободы знаменателя (в данном случае *дисперсии воспроизводимости*); $y_{i,расч}$ – значения *функции отклика* рассчитанные по проверяемой модели с параметрами, определёнными по выборке $y_1, y_2, \ldots, y_n$, причём каждому $y_{i,расч}$ соответствует $y_{i,эксп}$. Опытный критерий Фишера сравнивается с табличным значением, которое получается независимым путём:

$$p(F) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\cdot\Gamma\left(\frac{\nu_2}{2}\right)} \cdot \frac{\nu_1^{\nu_2/2}\,\nu_2^{\nu_1/2}\,F^{\nu_1/2-1}}{(\nu_1 F+\nu_2)^{(\nu_1+\nu_2)/2}}. \qquad (1.40)$$

Если $F^{оп} < F^{\alpha}$, – мы попадаем в область нулевой гипотезы (рис. 1.3).

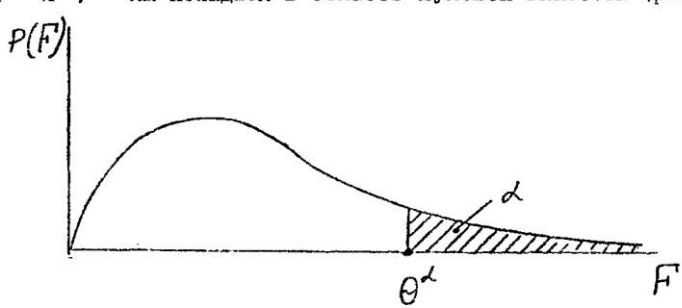

Рис. 1.3. Плотность распределения вероятностей критерия Фишера F. Заштрихована критическая область, составляющая α площади под кривой (общая площадь под кривой плотности вероятностей равна 1)

то математическая модель с уровнем значимости α адекватна, т.е. гипотеза H_0 принимается; если $F^{оп} > F^{\alpha}$, – мы попадаем в критическую область, область отклонения нулевой гипотезы, то модель неадекватна и

гипотеза H_0 отклоняется. Если $F^{оп} < F^{\alpha}$, то это ещё не означает действительного подтверждения гипотезы, так как невысокая точность экспериментальной работы приводит к завышению дисперсии воспроизводимости и соответственно к занижению опытного критерия Фишера: к такому же результату приводит уменьшение числа опытов на воспроизводимость.

При решении вопроса о принятии или отклонении какой-либо гипотезы H_0 с помощью любого критерия, основанного на результатах наблюдения, могут быть допущены ошибки двух родов. Ошибка первого рода совершается тогда, когда отвергается верная гипотеза H_0. Ошибка второго рода совершается в том случае, когда гипотеза H_0 принимается, а на самом деле верна не она, а какая-либо *альтернативная* гипотеза H_1. Вероятность допустить ошибку первого рода равна α, т.е. уровню значимости критерия; вероятность допустить ошибку второго рода равна P, т.е. доверительной вероятности. Эти ошибки не равноценны.

Уровень значимости *статистического критерия* - вероятность ошибочно отвергнуть основную проверяемую гипотезу, когда она верна. Понятие "уровень значимости" возникло в связи с задачей проверки согласованности теории с опытными данными. Если, например, в результате наблюдений регистрируются значения n случайных величин X_1, X_2, ..., X_n и требуется по этим данным проверить гипотезу H_0, согласно которой совместное распределение величин X_1, X_2, ..., X_n обладает некоторым определённым свойством, то соответствующий статистический критерий **конструируется** с помощью подходящим образом подобранной функции $\theta = f(X_1, X_2, ..., X_n)$. Эта функция обычно принимает малые значения, когда гипотеза H_0 верна, и большие значения, когда H_0 ложна; такую гипотезу ещё называют **гипотезой об отсутствии различия** или **нулевой гипотезой**. Соответствующий *критерий значимости* представляет собой правило, согласно которому значимыми считаются значения θ, превосходящие некоторое критическое значение θ^{α} (рис.1.2, рис.1.3). В свою очередь выбор величины θ^{α} определяется заданным уровнем значимости α, который в случае отклонения гипотезы H_0 совпадает с вероятностью события $\{\theta > \theta^{\alpha}\}$. Центральный момент при проверке гипотезы H_0 заключается в том, что уровнем значимости α задаются до анализа выборки на основании физической сущности задачи и последствий от ошибочного принятия решения. Диапазон значений уровней значимости, принимаемых в науке и технике, достаточно ши-

рок: 0,1; 0,05; 0,02; 0,01; 0,001; наиболее употребительно значение α=0,05. В теории статистической проверки гипотез уровень значимости наз. *вероятностью ошибки первого рода*. Вероятность такой ошибки не больше принятого уровня значимости. Например, при α=0,05 можно совершить ошибку первого рода в пяти случаях из ста. Принятие основной проверяемой гипотезы, когда она неверна, называется *ошибкой второго рода*. Фиксация уровня значимости находится целиком в компетенции исследователя, он должен решать, какой риск при отклонении истинной гипотезы является допустимым.

В геологии обычно имеют дело с обстоятельствами большой неопределённости, например, объём образцов (кернов), извлекаемых из скважин при разведочном бурении, несоизмеримо меньше объёма исследуемой залежи. Нулевой гипотезой в данном случае будет являться гипотеза об отсутствии отличия образцов исследуемой залежи от пустой породы; подтверждение гипотезы будет означать бесперспективность дальнейшего бурения, а опровержение - наличие той или иной нефтегазоносности. Если позволить допустить ошибку в одном случае из ста (α=0,01) или даже в одном случае из двадцати (α=0,05), то, имея в распоряжении керны из нескольких скважин, будет трудно отвергнуть нулевую гипотезу (т.е. доказать перспективность залежи) и возникнет необходимость во всё большем и большем объёме образцов, получить которые непросто. Принимая более скромные уровни значимости (α>0,1), можно быстрее прийти к заключению, хотя вероятность получить ошибочные выводы может оказаться очень высокой в сравнении со стандартами, принятыми в других областях.

Критерий значимости, с помощью которого гипотеза проверяется, конструируется таким образом, чтобы критическое значение критерия могло быть вычислено независимым путём в предположении, что проверяемая гипотеза верна. При этом значения собственно критерия располагаются вдоль оси абсцисс, функция представляет собой некоторую кривую, площадь под которой равна единице, а критерий значимости α точно равен площади под кривой в критической области, т.е. в области отклонения проверяемой гипотезы. Примерный характер кривых приведён на рис. 1.2 и 1.3, но в общем случае он может быть другим. Представим себе, что нефтяная компания сконструировала статистический критерий прогноза нефтегазоносности θ, состоящий из некоторых количественных переменных, позволяющих определять приоритеты при бурении. Цель применения статистического критерия - сделать более

или менее верный прогноз продуктивности скважин. Нулевая гипотеза H_0 состоит в том, что керны отобраны из совокупности бесперспективных разрезов, альтернативная H_1, - что керны отобраны из нефтяного или газового месторождения. Альтернативных гипотез может быть несколько.

Если принять уровень значимости, например α=0,05 (рис. 1.4,а).

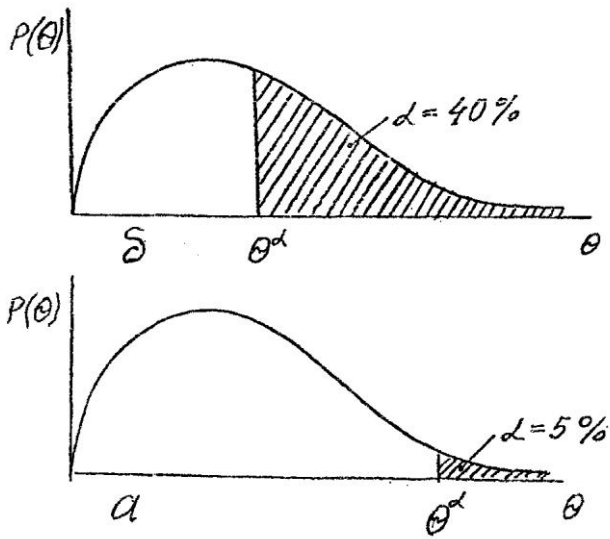

Рис. 1.4. Распределение статистики гипотетического критерия θ с критической областью α отклонения гипотезы о бесперспективности бурения (общая площадь под кривой плотности вероятностей равна 1)

то очень малая часть образцов окажется отличающейся от неперспективной породы (нулевой совокупности). Если же окажется, что образцы отличаются от неё, то это почти наверняка даст открытие месторождения при бурении. Компания получит очень высокое соотношение для числа успехов при бурении, но при этом пропустит много залежей, которые могли бы оказаться продуктивными. Другими словами, компания будет редко бурить, редко совершать ошибки первого рода, но, соответственно, оставит много месторождений неоткрытыми.

Если принять уровень значимости побольше, например α=0,40 (рис. 1.4,б), то придётся осуществлять частую сеть бурения, соответственно частота неперспективных скважин будет значительно выше, но вероятность пропуска залежи будет меньше. При такой практике

принятия решений компания будет часто бурить, часто ошибаться, но и значительно меньше месторождений нефти останется неоткрытыми.

В нефтяной промышленности случаи получения отрицательного результата при бурении перспективных площадей встречаются значительно чаще, чем последствия получения положительного результата при бурении пустых скважин. Причина этого состоит в том, что финансовый успех одного большого открытия может покрыть стоимость нескольких десятков пустых скважин. Оценка вероятности успеха одного из методов бурения в нефтяной промышленности США примерно равна 10%. Если бы эти скважины были пробурены на основе применения статистических критериев, то эта оценка соответствовала бы уровню значимости примерно α=0,9 [7].

Рассмотренный выше пример иллюстрирует выбор уровня значимости так называемого одностороннего критерия, поскольку нефть либо есть в залежи, либо нет, т.е. соответствующий критерий располагается только в положительной области. В тех случаях, когда физическая величина может принимать значения как положительные, так и отрицательные, либо критерий, соответствующий нулевой гипотезе, может располагаться в обеих областях декартовой системы координат, применяют двусторонний критерий значимости. Термин "нулевая гипотеза" возник от того, что математическое ожидание критерия, соответствующее подтверждению основной проверяемой гипотезы, равно нулю (рис. 1.5).

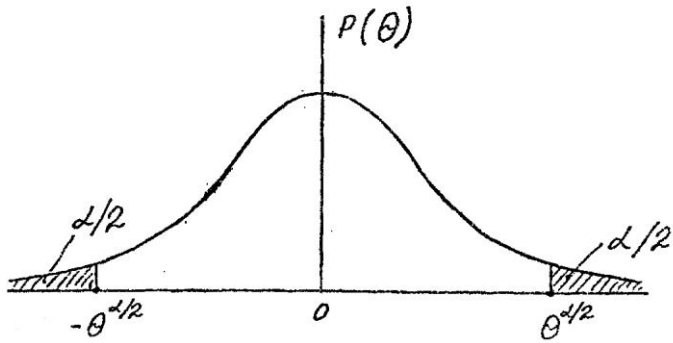

Рис. 1.5. Левая ($-\theta^{\alpha/2}$) и правая ($\theta^{\alpha/2}$) критические области гипотетического двустороннего критерия θ^{α}

Очевидно, что вероятность попадания критерия как в левую область, так и в правую одинакова, соответственно одинакова и вероятность

попадания критерия θ как в левую, так и в правую критические области. Таким образом, уровень значимости α, т.е. площадь соответствующая вероятности отклонения нулевой гипотезы, распадается на две равные области, каждая из которых равна $\alpha/2$. Критическое значение θ при этом увеличивается. Из этого следует важный вывод - приняв уровень значимости α для проверки гипотезы H_0, табличное значение критерия θ следует брать для вдвое меньшего значения, т.е. для $\alpha/2$. Соответствующие значения критериев записывют, например, так: для левой критической области $\theta^{1-\alpha/2}$, для правой - $\theta^{\alpha/2}$.

При выборе уровня значимости следует учитывать ущерб, неизбежно возникающий при использовании любого критерия значимости. Так, например, если уровень значимости чрезмерно велик, то основной ущерб будет происходить от ошибочного отклонения правильной гипотезы; если же уровень значимости мал, то ущерб будет, как правило, возникать от ошибочного принятия гипотезы, когда она ложна. Пример Дж. С. Дэвиса - крайний случай, но он показывает, что исследователь должен принимать решения на границе области риска и выбирать уровень значимости в соответствии с конкретными обстоятельствами.

Число степеней свободы характеризует информационный потенциал выборки. Это всегда целое положительное число, равное разности между числом наблюдений в выборке и числом параметров, определённых по данным выборки. Число степеней свободы обозначается греческой буквой ν, а число параметров, определённых по выборке, - латинской буквой l; таким образом $\nu=n-l$. Число параметров, определённых по выборке, ещё называется *числом связей* наложенных на выборку. Так вот, с точки зрения теории информации число степеней свободы равно числу параметров, которое ещё можно определить по выборке после той или иной обработки, а с точки зрения математической статистики ν равно числу независимых источников информации, по которым вычисляется тот или иной выборочный параметр. Дело в том, что, используя одну и ту же выборку, невозможно решить сразу две задачи: оценить параметры совокупности и применить соответствующий критерий для проверки достоверности полученных оценок без какой-либо компенсации, связанной с двукратным обращением к имеющемуся массиву наблюдений. Такой компенсацией является уменьшение знаменателя в формуле выборочной дисперсии от числа наблюдений n до числа независимых источников информации оцениваемого параметра ν. Если, например, математическое ожидание M_x оценивается по результатам пяти независимых наблюдений

$$\bar{x} = (x_1+x_2+x_3+x_4+x_5)/5, \qquad (1.41)$$

то результат имеет пять степеней свободы. Дисперсия оценивается по пяти квадратам разностей $(x_1-x_{ср})^2$, однако независимо вычисляются только четыре из этих разностей, так как определив четыре, пятую уже можно вычислить следующим образом:

$$x_5-\bar{x} = 5\bar{x}-(x_1+x_2+x_3+x_4). \qquad (1.42)$$

Поэтому имеется только четыре независимых источника информации по которым вычисляется выборочная дисперсия. Бывают случаи, когда в качестве оценки математического ожидания M_x используется величина, не зависимая от рассматриваемой выборки (например, m_x), т.е. оценка определяется независимым путём. В таких случаях для выборочной дисперсии следует пользоваться формулой (1.31):

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-m_x)^2. \qquad (1.43)$$

Эффективность оценки - свойство оценки иметь больший или меньший доверительный интервал. Оценка параметра называется эффективной, если среди нескольких оценок того же параметра она обладает наименьшей дисперсией.

Задача собственно оценки неизвестных параметров отступает на второй план перед проблемой ***априорного принятия вида функции***. Диалектика научного познания проявляется в том, что, имея массив наблюдений, исследователь на основе знаний, опыта, интуитивных соображений предполагает вид зависимости функции отклика от факторов и только после этого каким-либо методом вычисляет оценки параметров ***принятой функции***. Если в результате оказывается, что принятая функция (математическая модель) неадекватна, т.е. не соответствует результатам эксперимента при достигнутой точности опытов, то у исследователя есть три выхода: попробовать другой метод обработки данных, предположить другой вид функциональной зависимости $y=f(x_1, x_2, \ldots, x_k)$ или повторить (продолжить) эксперименты. Например, потерпев неудачу при обработке данных по уравнению $y=b_0+b_1x$, попробовать уравнение вида $y=b_0+b_1/x$, другое $y=x/(b_0+b_1x)$ и т.д. (см. табл. 2.2).

2. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ В ДВУМЕРНОМ ПРОСТРАНСТВЕ

Каждое событие в мире является результатом одновременного воздействия некоторого множества причин или детерминированных факторов, поэтому без преувеличения можно утверждать, что в большинстве случаев подходящим уравнением для статистического моделирования будет уравнение второго порядка вида (1.4) или, в особых случаях, третьего. К сожалению, получение подобного уравнения - задача не простая. Человеку, живущему в четырёхмерном пространстве-времени и привыкшему строить соответствующие мысленные модели, трудно создавать и анализировать многомерную мысленную модель, создать в своём уме образ многомерного пространства. Человеку, может быть, не проще, но эффективнее анализировать сечения многофакторной зависимости $y=f(x_1, x_2, \ldots, x_k)$, обрабатывая отдельные зависимости $y=f(x_1)$, $y=f(x_2), \ldots, y=f(x_k)$, обобщать и делать выводы впоследствии. На протяжении столетий естествоиспытатели наблюдают, что произойдёт с интересующим их явлением (функцией отклика y), если изменить независимую переменную (фактор x). Естественным результатом наблюдений является их графическое представление и попытка поиска вида функциональной связи $y=f(x)$.

2.1. Парная корреляция

Если между случайными величинами существует такая связь, что с изменением одной величины меняется распределение другой, то эту связь принято называть стохастической связью (от греч. στοχαστικός - умеющий целить, попадать; умеющий верно отгадывать, судить). Изменение случайной величины y, соответствующее изменению величины x, раскладывается при этом на две составляющих: функциональную компоненту, связанную с зависимостью $y=f(x)$, и случайную.

Например, линейная функциональная связь вида $y=bx$, представленная на рис. 2.1, может быть получена только расчётным путем. В процессе экспериментальной работы

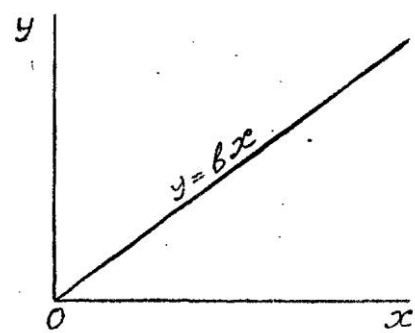

Рис.2.1. Функциональная зависимость $y=f(x)$

неизбежны ошибки определения функции отклика (рис. 2.2). На рис. 2.3 представлен случай, когда случайные величины Y и X независимы.

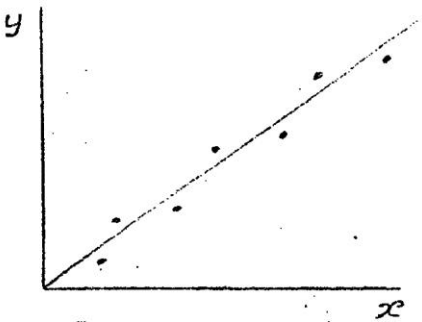

Рис.2.2. Результаты экспериментального определения неизвестной зависимости $y=f(x)$, для которой можно предполагать, что $y=bx$

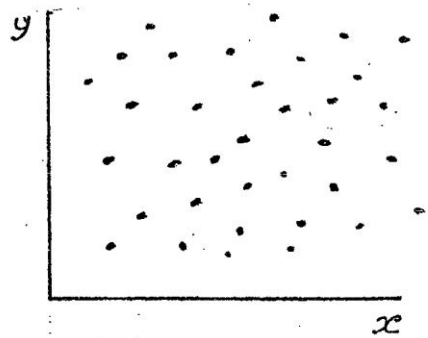

Рис.2.3. Пример отсутствия зависимости между случайными величинами y и x

В примере на рис. 2.2 линейная функциональная связь имела бы место, если бы все точки располагались на прямой регрессии. Наличие ошибок измерений приводит к тому, что связь между y и x является стохастической (вероятностной). Соотношение между функциональной и случайной компонентами определяет силу (тесноту) связи. Важнейшим показателем этой связи является коэффициент корреляции [1, 5, 17]

$$\rho_{xy} = \frac{E[(X-M_x)(Y-M_y)]}{\sigma_x \sigma_y}. \qquad (2.1)$$

Коэффициент корреляции характеризует линейную функциональную зависимость, осложнённую ошибками измерений, или отсутствие оной. Линейная стохастическая зависимость двух случайных величин заключается в том, что при возрастании одной случайной величины другая имеет тенденцию возрастать (или убывать) по линейному закону. Коэффициент корреляции характеризует степень тесноты линейной зависимости. Если случайные величины Y и X связаны линейной функциональной зависимостью $y=b_0+b_1x$ и эта связь не осложнена ошибками измерений, то $r_{yx}=\pm 1$, причём знак r_{yx} соответствует знаку коэффициента b_1. В общем случае, когда величины Y и X связаны произвольной стохастической зависимостью, коэффициент корреляции может иметь значения в пределах $-1<r_{yx}<+1$. При $r_{yx}>0$ существует положительная корреляционная связь между величинами Y и X, при $r_{yx}<0$ - отрицательная. Для независимых случайных величин $r_{yx}=0$. Практически коэффициент корреляции отмечает и больший или меньший вклад случайности (большие ошибки измерений) и большую или меньшую кривизну этой связи. Зависимость между Y и X может быть строго функци-

циональной, а коэффициент корреляции всё же будет меньше единицы.

Поскольку в результате экспериментов исследователь получает случайную выборку, то по результатам выборки определяется выборочный коэффициент корреляции r_{yx}. Выборочный коэффициент корреляции определяется так же, как и генеральный коэффициент ρ_{yx}, только при этом используются выборочные средние $x_{ср}$ и $y_{ср}$ и выборочные дисперсии s^2_y и s^2_x:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y}. \qquad (2.2)$$

После преобразований формула (2.2) приобретает вид [1]

$$r_{xy} = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] \cdot \left[n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}. \qquad (2.3)$$

Число степеней свободы для найденного таким путём коэффициента корреляции r_{xy} равно $\nu = n-2$, поскольку вычисление $x_{ср}$ и $y_{ср}$ соответствует наложению на выборку двух связей.

2.2. Линейная двухпараметрическая регрессия

2.2.1. *Обоснование метода наименьших квадратов*

Пусть имеется n пар наблюдений значений функции отклика y, полученных при фиксированных значениях независимой переменной (фактора) x. Для графического изображения этих n пар наблюдений в виде экспериментальных точек с координатами y-x на плоскости применяется система декартовых координат (рис. 2.4). Координаты точек 1-9, изображённых на рис. 2.4, приведены в табл. 2.1. Подобные результаты наблюдений могут быть получены в любой экспериментальной работе.

Задача метода наименьших квадратов (линейного регрессионного анализа) - зная положение точек 1-9 на плоскости, так провести линию регрессии, чтобы сумма квадратов отклонений $(y_i - y_{i,расч})^2$ вдоль

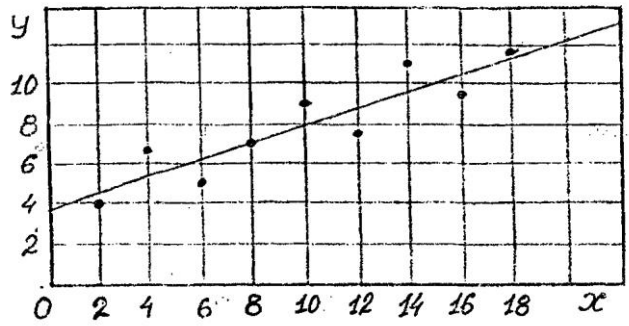

Рис. 2.4. Корреляционное поле зависимости $y=f(x)$ и результаты статистического оценивания этой корреляции уравнением регрессии $y=3,6+0,43x$

Таблица 2.1.

Результаты экспериментального определения некоторой зависимости $y=f(x)$

№	1	2	3	4	5	6	7	8	9
x	2,0	4,0	6,0	8,0	10,0	12,0	14,0	16,0	18,0
y	4,0	6,5	5,0	7,0	9,0	7,5	11,0	9,5	11,5

оси y (ординаты) этих точек от проведённой прямой ($y_{1,\text{расч}}$) была бы минимальной:

$$\Phi = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min. \qquad (2.4)$$

Для проведения вычислений по классическому методу наименьших квадратов (для проведения регрессионного анализа) к выдвигаемой гипотезе (к форме уравнения регрессии) предъявляется следующее требование: это уравнение должно быть линейным по параметрам.

Прежде чем переходить к дальнейшему необходимо обсудить смысл прилагательного "линейный" в рассматриваемой модели линейной регрессии $y=f(x)$. Эта модель должна быть *линейной относительно параметров* β_j, но не обязана быть линейной относительно $x_1, x_2, \ldots, x_k$. Дело в том, что по существу метод наименьших квадратов позволяет вычислить несмещённые и эффективные оценки параметров, являющихся **линейными функциями от наблюдений**. Другими словами, существенна линейность только относительно параметров, тогда как, возможно более "естественная" линейность функции отклика y от x с точки зрения ме-

тода НК принципиального значения не имеет. Таким образом, модель линейной регрессии включает в себя некоторое множество зависимостей y от x: $y=b_0+b_1x$, $y=b_0+b_1/x$, $y=x/(b_0+b_1x)$, $y=b_0+b_1x^2$ и т.д. Например, непосредственно полиномиальная зависимость

$$y=b_0+b_1x+b_2x^2+b_3x^3+\ldots+b_nx^n; \qquad (2.5)$$

нелинейна относительно x, но линейна относительно β_j, и с точки зрения метода НК является уравнением линейной регрессии. С другой стороны, уравнение, линейное относительно x

$$y=b_0+b_1x+b_1^2x; \qquad (2.6)$$

с точки зрения *метода* НК линейным не является, поскольку в него входят и β_1 и β^2_1. Аналогично, модель, нелинейная по своей сущности

$$y=b_0+b_1x+b_2x^2+b_3\sin x, \qquad (2.7)$$

с точки зрения *метода* НК является линейной. Противоречие "модель линейная относительно b_j, нелинейная относительно x" и наоборот - кажущееся, а необходимость соблюдения этого требования кроется в методе получения практических расчётных формул для коэффициентов b_0, b_1, b_2,... (см. ниже).

2.2.2. *Процедура метода наименьших квадратов*

Возвращаясь к процедуре проведения регрессионного анализа, предположим, что зависимость, изображённая на рис. 2.4, - линейная, т.е. при проведении парного линейного регрессионного анализа имеем дело с уравнением прямой линии. Уравнение прямой на плоскости в декартовых координатах имеет вид

$$y=b_0+b_1x, \qquad (2.8)$$

где b_0 и b_1 - коэффициенты или оценки неизвестных параметров β_0 и β_1 предполагаемой зависимости

$$v=\beta_0+\beta_1x. \qquad (2.9)$$

Сопоставляя с условием (2.4) видим, что $\hat{y}=b_0+b_1x$ и задача метода НК аналитически выражается следующим образом:

$$\Phi=\sum_{i=1}^{n}\left(y_1-(b_0+b_1x_1)\right)^2=\min. \qquad (2.10)$$

Формула (2.10) выражает сущность метода НК (принцип А. Лежандра): сумма квадратов отклонений вдоль y должна быть минимальной. Другими словами, прямая линия $y=f(x)$ будет проходить наилучшим образом через экспериментальные точки в том случае, когда сумма квадратов разностей экспериментальных значений функции отклика y и рассчитанных по искомому уравнению (2.8) будет иметь минимальное значение.

Таким образом, задача определения коэффициентов уравнения регрессии по методу НК сводится практически к определению минимума функции многих переменных. В нашем случае, если

$$\hat{y}=\varphi(x,\ b_0,\ b_1) \qquad (2.11)$$

есть функция непрерывная и дифференцируемая и требуется выбрать b_0 и b_1 так, чтобы выполнялось условие (2.10), то необходимым условием минимума $\Phi(b_0,\ b_1)$ является равенство нулю двух первых частных производных:

$$\frac{\partial \Phi}{\partial b_0}=0;\quad \frac{\partial \Phi}{\partial b_1}=0; \qquad (2.12)$$

при этом b_0 и b_1 рассматриваются как переменные величины, а x и y - как постоянные.

Подставляя в (2.12) выражение для Φ из (2.10), получим

$$\begin{cases} \dfrac{\partial \Phi}{\partial b_0}=2\sum\limits_{i=1}^{n}\{y_i-(b_0+b_1x_i)\}=0; \\ \dfrac{\partial \Phi}{\partial b_1}=2\sum\limits_{i=1}^{n}\{y_i-(b_0+b_1x_i)\}x=0. \end{cases} \qquad (2.13)$$

После простых преобразований получим

$$\begin{cases} nb_0+b_1\sum\limits_{i=1}^{n}x_i=\sum\limits_{i=1}^{n}y_i; \\ b_0\sum\limits_{i=1}^{n}x_i+b_1\sum\limits_{i=1}^{n}x_i^2=\sum\limits_{i=1}^{n}x_iy_i. \end{cases} \qquad (2.14)$$

Система (2.14) - это система нормальных уравнений. Функция $\Phi>0$ при любых b_0 и b_1, следовательно, у неё должен быть хотя бы один глобальный минимум. Поэтому, если система нормальных уравнений имеет единственное решение, то оно и является минимумом для функции Φ.

Решая систему нормальных уравнений (2.14) с помощью определителей [1], получим

$$b_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}; \qquad (2.15)$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \qquad (2.16)$$

где $n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 = D$ – главный определитель системы нормальных уравнений.

Значение коэффициента корреляции не ограничивается оценкой силы связи между величинами x и y. Из первого уравнения системы (2.14)

$$nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (2.17)$$

очевидна связь между коэффициентами b_0 и b_1

$$b_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{b_1}{n} \sum_{i=1}^{n} x_i = \bar{y} - b_1 \bar{x}. \qquad (2.18)$$

Соотношение (2.18) показывает, что между коэффициентами b_0 и b_1 существует корреляционная зависимость или, другими словами, коэффициенты b_0 и b_1 коррелированы.

2.2.3. *Уточнение метода наименьших квадратов*

Анализ расчётных формул (2.15) и (2.16) показывает, что расчёт по ним позволит восстановить зависимость, расположенную в положительной системе координат. Если исходная зависимость располагается как в положительном, так и в отрицательном секторах декартовой сис-

темы координат, формулы (2.15) и (2.16) дадут неверные результаты. В таких случаях расчёт необходимо производить по формулам:

$$A = \frac{1}{n}\cdot\sum_{i=1}^{n}|x_i|; \qquad B = \frac{1}{A^2}\cdot\sum_{i=1}^{n}x_i^2; \qquad C = \frac{1}{A}\cdot\sum_{i=1}^{n}x_i y_i;$$

$$D = \sum_{i=1}^{n} y_i; \qquad E = \frac{1}{A}\cdot\sum_{i=1}^{n}x_i; \qquad F = nB - E^2;$$

$$b_0 = \frac{BD-EC}{F}; \qquad b_1 = \frac{nC-ED}{AF}.$$

2.3. Обратная линейная регрессия

Вообще существуют ещё две модели линейной регрессии – обратная и ортогональная [9]. Общепринята зависимость (2.8), которую условно можно назвать моделью прямой регрессии. Если рассматривать зависимость $x=\varphi(y)$ и минимизировать сумму квадратов отклонений вдоль оси x, то модель

$$\hat{x} = b_0 + b_1 y \qquad (2.19)$$

будет называться обратной регрессией. Выбор уравнений (2.8) или (2.19) зависит не только от того, какая из случайных величин y или x считается фактором (независимой переменной), но и от относительной ошибки, совершаемой при измерениях y и x. По условию метода НК независимая переменная должна фиксироваться теоретически без ошибки, т.е. с большей точностью. Для выполнения этого требования в практике экспериментальной работы интервал варьирования независимой переменной принимают значительно превышающим точность измерения и регулирования. Например, при исследовании влияния температуры на скорость химической реакции, при точности регулирования температуры $\pm 0,1^\circ$C, интервал варьирования температуры в реакционной серии приблизительно 10°.

В ортогональной регрессии критерием оптимальности является минимум квадратов длин перпендикуляров, опущенных на линию регрессии. Ввиду зависимости коэффициентов получаемого уравнения от масштаба единиц измерения и непростого пересчёта ортогональная регрессия нашла достаточно ограниченное практическое применение.

2.4. Линейная однопараметрическая регрессия

Случаем одного неизвестного параметра является парная зависимость вида $y=bx$, параметр b которой по методу НК находится из следующего требования:

$$\Phi = \sum_{i=1}^{n} \left(y_i - bx_i \right)^2 = \min. \qquad (2.20)$$

Рассуждая аналогично, неизвестный параметр b уравнения $y=bx$ рассматривают как переменную величину, и условием минимума функции Φ является равенство нулю первой производной:

$$\frac{d\Phi}{db} = 0. \qquad (2.21)$$

Решением условия (2.21) является формула

$$b = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2. \qquad (2.22)$$

Расчёт по формуле (2-22) необходимо производить в тех случаях, когда по физической сущности задачи свободного члена в уравнении линейной регрессии не должно быть, т.е. при $x=0$ функция отклика y тоже равна нулю. В результате расчётов по формулам (2.15) и (2.16) то или иное значение коэффициента b_0 будет получено неизбежно. Будет ли этот коэффициент значимым - другой вопрос, но в результате корреляции коэффициентов b_0 и b_1 важный по сущности задачи коэффициент b_1 будет искажён.

2.5. Нелинейная парная регрессия

В том случае, когда при графическом изображении точек в декартовых координатах нелинейность явно просматривается на глаз или при проведении регрессионного анализа гипотеза линейности может быть отброшена, есть смысл попробовать нелинейные уравнения, например вида $y=b_0+b_1/x$, $y=1/b_0+b_1x$, $y=b_0+b_1\ln x$ и др., в зависимости от физической сущности задачи и характера кривой. Естественно, что в каждом случае нужно будет выводить расчётные формулы для коэффици-

ентов b_0 и b_1, для чего записать выражение для функции Ф, продифференцировать его дважды и решить систему нормальных уравнений. Это достаточно неудобно, кроме этого, не для каждой формы уравнения регрессии можно произвести преобразования (2.12), (2.13), и, практически, исследователи предпочитают тем или иным способом привести нелинейное уравнение к линейному виду и воспользоваться классическими формулами (2.15) и (2.16). Это удобнее, несмотря на то, что в результате линеаризующих преобразований неизбежна потеря точности.

Сущность линеаризации заключается в преобразовании декартовой системы координат с целью приведения исходной нелинейной зависимости к линейному виду. Под преобразованием системы координат подразумевается изменение масштаба по осям координат, позволяющее представить исходную кривую в виде прямой линии. Способ преобразования зависит от конкретного вида исходной зависимости или, точнее, от формы уравнения парной зависимости. Например, зная или предполагая, что исследуемая химическая реакция имеет первый порядок, данные эксперимента могут быть описаны показательной функцией $y=b_0 b_1^x$, где в качестве функции отклика y будет фигурировать концентрация c реагирующего вещества, а в качестве независимой переменной x - время τ. Для обработки данных эксперимента $c=f(\tau)$ функцию отклика y следует прологарифмировать и построить график в координатах ($\ln y$ - x) в соответствии с преобразованием

$$y=b_0 \cdot b_1^x \Rightarrow \ln y=\ln b_0+x \cdot \ln b_1. \qquad (2.23)$$

Линеаризованная форма может быть записана в виде:

$$y'=b'_0+b'_1 x' \qquad (2.24)$$

и для неё справедливы формулы (2.15) и (2.16), по которым можно вычислить коэффициенты b'_0 и b'_1. Для определения коэффициентов b_0 и b_1 следует сравнить линеаризованную форму (2.23) и уравнение (2.24): очевидно, что $\ln b_0=b'_0$, а $\ln b_1=b'_1$. Осуществив потенцирование, получим коэффициенты b_0 и b_1 искомого уравнения. Возвращаясь к задаче поиска параметров химической реакции первого порядка, получим

$$c=c_0 \cdot \exp(-k\tau),$$

где c_0 - начальная концентрация реагента, а k - константа скорости химической реакции. Очевидно, что в данном случае коэффициенты ис-

комого уравнения имеют совершенно определённый физический смысл.

В заключение следует отметить тот факт, что достаточно часто концентрации веществ численно меньше единицы, поэтому $\ln y$ в уравнении (2.23) будет величиной отрицательной и расчёт искомых параметров пратически нужно будет производить по уточнённым формулам МНК (см. раздел 2.2.3).

2.6. Выбор оптимальной формы парной регрессии

Используя метод НК, можно построить практически любые формы нелинейной парной зависимости. При этом можно исходить из физической сущности задачи, преобразовывать предполагаемое детерминистическое уравнение к виду (2.8) и вычислять коэффициенты b_0 и b_1. Далее, возвращаясь в исходные координаты, пересчитывать коэффициенты b_0 и b_1 в параметры предполагаемой модели. Практика экспериментальных исследований накопила достаточно много приёмов преобразований, среди них наиболее часто встречаются обращение, логарифмирование и возведение в степень [9]. Результаты анализа линеаризующих преобразований и возможных систем координат сведены в табл. 2.2. В ней приведены 21 форма уравнений парной регрессии и 18 наиболее распространённых систем координат для линеаризованной формы. При этом возможен формальный подход к обработке экспериментальных данных с неясной причинно-следственной зависимостью: перебор всех возможных форм уравнений парной регрессии с целью поиска той системы координат ($y'-x'$), в которой данные экспериментов будут укладываться на прямую линию.

Первая форма уравнения парной регрессии предназначена для тех случаев, когда по физической сущности задачи свободного члена в уравнении линейной регрессии не должно быть, т.е. $y=bx$.

Вторая форма - основная, она предназначена как для обработки зависимостей собственно линейных зависимостей вида $y=b_0+b_1x$, так и линеаризованных вида $y'=b'_0+b'_1x'$.

Формы уравнений с 3 по 16 охватывают практически все важные случаи линеаризующих преобразований и систем координат для парной регрессии с двумя параметрами. Они предназначены для обработки строго монотонных нелинейных таблично заданных функций $y=f(x)$. При обработке нелинейных зависимостей информация вначале преобразуется в координаты, соответствующие линейному виду той или иной формы

Таблица 2.2.

Возможные формы уравнений парной регрессии и линеаризующие преобразования системы координат

	Формы уравнений парной регрессии	Линеаризующие преобразования				
		Линеаризованная форма уравнения регрессии $y'=b'_0+b'_1x'$	Декартова система координат для линейной формы		Выражения для коэффициентов b_0 и b_1	
					b'_0	b'_1
01	$y=b_1x$	$y=b_1x$	y	x	0	b_1
02	$y=b_0+b_1x$	$y=b_0+b_1x$	y	x	b_0	b_1
03	$y=b_0+b_1/x$	$y=b_0+b_1\left(\frac{1}{x}\right)$	y	$\frac{1}{x}$	b_0	b_1
04	$y=\frac{1}{b_0+b_1x}$	$\frac{1}{y}=b_0+b_1x$	$\frac{1}{y}$	x	b_0	b_1
05	$y=\frac{b_0x}{b_1+x}$	$\frac{1}{y}=\frac{1}{b_0}+\frac{b_1}{b_0}\cdot\left(\frac{1}{x}\right)$	$\frac{1}{y}$	$\frac{1}{x}$	$\frac{1}{b_0}$	$\frac{b_1}{b_0}$
06	$y=\frac{x}{b_0+b_1x}$	$\frac{x}{y}=b_0+b_1x$	$\frac{x}{y}$	x	b_0	b_1
07	$y=\frac{x}{b_0+b_1/x}$	$\frac{x}{y}=b_0+b_1\left(\frac{1}{x}\right)$	$\frac{x}{y}$	$\frac{1}{x}$	b_0	b_1
08	$y=b_0\cdot b_1^x$	$\ln y=\ln b_0+x\cdot\ln b_1$	$\ln y$	x	$\ln b_0$	$\ln b_1$
09	$y=b_0+b_1\ln x$	$y=b_0+b_1\ln x$	y	$\ln x$	b_0	b_1
10	$y=b_0\cdot x^{b_1}$	$\ln y=\ln b_0+b_1\ln x$	$\ln y$	$\ln x$	$\ln b_0$	b_1
11	$y=\frac{b_0x}{\exp(b_1x)}$	$\ln\left(\frac{x}{y}\right)=-\ln b_0+b_1x$	$\ln\frac{x}{y}$	x	$-\ln b_0$	b_1
12	$y=\frac{x}{b_0+b_1\ln x}$	$\frac{x}{y}=b_0+b_1\ln x$	$\frac{x}{y}$	$\ln x$	b_0	b_1
13	$y=\frac{1}{b_0+b_1\ln x}$	$\frac{1}{y}=b_0+b_1\ln x$	$\frac{1}{y}$	$\ln x$	b_0	b_1
14	$y=b_0\exp\left\{\frac{b_1}{x}\right\}$	$\ln y=\ln b_0+b_1\left(\frac{1}{x}\right)$	$\ln y$	$\frac{1}{x}$	$\ln b_0$	b_1
15	$y=\frac{1}{b_0+b_1\exp(-x)}$	$\frac{1}{y}=b_0+b_1\exp(-x)$	$\frac{1}{y}$	e^{-x}	b_0	b_1
16	$y=b_0+b_1\exp(-x)$	$y=b_0+b_1\exp(-x)$	y	e^{-x}	b_0	b_1
17	$y=b_0+b_1x^h$	$y=b_0+b_1x^h$	y	x^h	b_0	b_1
18	$y=\exp\left\{b_0+\left(\frac{b_1}{x+h}\right)\right\}$	$\ln y=b_0+b_1\left(\frac{1}{x+h}\right)$	$\ln y$	$\frac{1}{x+h}$	b_0	b_1
19	$y=h+b_0\cdot x^{b_1}$	$\ln(y-h)=\ln b_0+b_1\ln x$	$\ln(y-h)$	$\ln x$	$\ln b_0$	b_1
20	$y=b_0+b_1x+b_2x^2+\ldots$	–	y	x	$b_0, b_1, \ldots, b_k$	
21	Сплайн-регрессия	–	y	x	–	–

$y \to y'$, $x \to x'$ (т.е. производится линеаризация зависимости); далее линеаризованная зависимость $y'=f(x')$ обрабатывается по уточнённым формулам формы 2.

Результаты обработки линеаризованной зависимости $y'=b'_0+b'_1x'$ - дисперсия переменной $(s')^2$, коэффициент корреляции r', коэффициенты b'_0 и b'_1 и их квадратичные отклонения s'_{b0} и s'_{b1} - являются основанием для заключения, насколько близка к линейной зависимость преобразованная в координаты $(y'-x')$. Здесь следует обращать внимание на коэффициент корреляции r' и квадратичные отклонения s'_{b0} и s'_{b1}. Далее производится возврат в исходные координаты и вычисляются дисперсия переменной s^2 и коэффициенты b_0 и b_1. Критерием выбора оптимальной формы является минимум дисперсии s^2.

2.7. Нелинейная трёхпараметрическая регрессия

Особенностью форм 17, 18 и 19 является наличие трех параметров уравнений регрессии - b_0, b_1 и h. Наличие трех параметров даёт уравнению регрессии большую гибкость описания таблично заданной функции и в некоторых случаях позволяет получить уравнение регрессии с меньшей дисперсией s^2. Например, классическими аналогами формы 18 являются уравнение Антуана

$$\ln p = A + \frac{B}{T+C},$$

которое используется для описания зависимости давления насыщенных паров жидкостей p от температуры T, и уравнение Андраде

$$\ln \mu = A + \frac{B}{T+C},$$

которое хорошо описывает зависимость динамического коэффициента вязкости жидкостей и газов μ от температуры T.

Форма 17 (уравнение $y=b_0+b_1x^h$) подходит для описания параболических зависимостей в тех случаях, когда кривая $y=f(x)$ не выходит из начала координат. Для описания таких зависимостей подходит также уравнение $y=h+b_0 \cdot x^{b1}$ (форма 19), которое, по существу, подобно уравнению $y=b_0+b_1x^h$. Формы 17 и 19 различаются подбираемыми параметрами h: в форме 17 подбирается показатель степени при x, а в

форме 19 - свободный член. Формы 17 и 19 взаимно дополняют друг друга, с их помощью можно последовательно уточнять параметры уравнения.

При обработке информации по формам 17, 18 и 19 параметр h подбирается методом сканирования. Метод сканирования - простой и достаточно эффективный метод, позволяющий с требуемой степенью точности найти глобальный экстремум функции. В данном случае решается задача минимизации суммы квадратов разностей экспериментальных значений функции отклика $y_{эксп}$ и расчитанных по уравнениям 17, 18 и 19 $y_{расч}$. Поскольку функция Ф всегда положительна, она должна иметь хотя бы один глобальный минимум, при этом параметром оптимизации является h, а критериаем оптимальности - минимум дисперсии s^2. На каждом шаге оптимизации принимается значение параметра h, после чего уравнение приводится к линейному виду и по уточнённому алгоритму формы 2 вычисляются коэффициенты b'_0 и b'_1, их квадратичные отклонения s'_{b0} и s'_{b1} и дисперсия приведенной переменной $(s')^2$. Далее, как и в случае поиска оптимальной формы, производится возврат в исходные координаты и вычисляется дисперсия s^2, которая и является критерием подбора параметра h. Сканирование следует продолжать до получения минимального значения дисперсии s^2 с требуемой точностью. Комбинация параметра оптимизации h и коэффициентов b_0, b_1 и будет искомой.

Поиск оптимального значения h осуществляется следующим образом. Вначале на основании графической зависимости или пробных расчетов делается вывод об области вероятных значений h. Эта область разбивается на 10 - 20 интервалов с шагом Δh_1. В узловых точках при соотвествующих значениях h производятся линеаризующие преобразования исходной информации и обработка линеаризованной зависимости по уточнённым формулам формы 2. При текущих значениях коэффициентов b'_0, b'_1 и параметра h вычисляются значения дисперсии $(s')^2$, коэффициент корреляции r' и квадратичные отклонения s'_{b0} и s'_{b1}. Далее производится возврат в исходную систему координат $(y-x)$, и для тех же узловых значений h вычисляются значения дисперсии s^2. В общем случае при некотором значении $h=h_1$ должен существовать минимум дисперсии. Во втором цикле оптимизации область изменения h от $h_1-\Delta h_1$ до $h_1+\Delta h_1$ также разбивается на 10÷20 интервалов с шагом Δh_2. В этих

узловых точках также при соотвествующих значениях h производятся линеаризующие преобразования исходной информации, обработка линеаризованной зависимости, возврат в исходную систему координат и вычисление дисперсии s^2 с поиском s^2_{min}. Таким образом получается третья область изменения h от $h_2-\Delta h_2$ до $h_2+\Delta h_2$ в которой аналогично осуществляется третий цикл оптимизации и т.д.

Эти расчеты повторяют до получения значения h_{opt} с требуемой точностью; при найденном значении h_{opt} выполняется окончательный расчёт всех коэффициентов модели и выборочных границ.

Следует отметить, что, в общем случае, задача подбора трех коэффициентов по одной кривой не всегда может быть корректно решена вследствие наличия корреляции (взаимозависимости) между коэффициентами b_0, b_1 и параметром h уравнения регрессии. Сильная корреляция коэффициентов может привести к искажению их физического смысла и к неверным теоретическим выводам. В этом случае правильнее говорить о получении формального математического описания. Для уменьшения корреляции коэффициентов уравнения и, соответственно, исключения ошибок, следует обрабатывать одновременно несколько экспериментальных кривых, использовать различные приёмы независимого (раздельного) определения некоторых коэффициентов искомого уравнения и др.

2.8. Параболическая парная регрессия

В научных исследованиях и в технологии бывают случаи, когда не удаётся подобрать удовлетворительного математического описания таблично заданной функции каким-либо нелинейным уравнением с двумя-тремя коэффициентами (формы 3÷19, табл.2.2). В таких случаях можно попытаться получить уравнение квадратичной регрессии

$$y=b_0+b_1x+b_2x^2, \qquad (2.25)$$

кубической регрессии

$$y=b_0+b_1x+b_2x^2+b_3x^3, \qquad (2.26)$$

или в общем случае параболической регрессии (форма 20)

$$y=b_0+b_1x+b_2x^2+b_3x^3+\ldots+b_kx^k, \qquad (2.27)$$

где $k<n-2$.

Для примера рассмотрим систему нормальных уравнений для вычисления коэффициентов уравнения квадратичной регрессии. Система нормальных уравнений для поиска коэффициентов квадратичного уравнения регрессии имеет вид

$$\begin{cases} b_0 n + b_1 \sum\limits_{i=1}^{n} x_i + b_2 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} y_i; \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 + b_2 \sum\limits_{i=1}^{n} x_i^3 = \sum\limits_{i=1}^{n} x_i y_i; \\ b_0 \sum\limits_{i=1}^{n} x_i^2 + b_1 \sum\limits_{i=1}^{n} x_i^3 + b_2 \sum\limits_{i=1}^{n} x_i^4 = \sum\limits_{i=1}^{n} x_i^2 y_i; \end{cases} \qquad (2.28)$$

Решаем систему (2.28) с помощью определителей

$$b_0 = \frac{\begin{vmatrix} \sum\limits_{i=1}^{n} y_i & \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \\ \sum\limits_{i=1}^{n} x_i y_i & \sum\limits_{i=1}^{n} x_i^2 & \sum\limits_{i=1}^{n} x_i^3 \\ \sum\limits_{i=1}^{n} x_i^2 y_i & \sum\limits_{i=1}^{n} x_i^3 & \sum\limits_{i=1}^{n} x_i^4 \end{vmatrix}}{D}; \qquad (2.29)$$

$$b_1 = \frac{\begin{vmatrix} n & \sum\limits_{i=1}^{n} y_i & \sum\limits_{i=1}^{n} x_i^2 \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i y_i & \sum\limits_{i=1}^{n} x_i^3 \\ \sum\limits_{i=1}^{n} x_i^2 & \sum\limits_{i=1}^{n} x_i^2 y_i & \sum\limits_{i=1}^{n} x_i^4 \end{vmatrix}}{D}; \qquad (2.30)$$

$$b_2 = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^2 y_i \end{vmatrix}}{D}, \qquad (2.31)$$

где D – главный определитель системы нормальных уравнений,

$$D = \begin{vmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \end{vmatrix}. \qquad (2.32)$$

Очевидно, что без ЭВМ регрессионный анализ второго порядка выполнить не просто [6]. С помощью компьютера можно последовательно наращивать степень полинома, ориентируясь на дисперсию s^2. Вначале с ростом степени уравнения регрессии дисперсия уменьшается вследствие роста гибкости аппроксимирующего полинома, но начиная с некоторой степени дисперсия начинает возрастать. Дело в том, что с ростом степени полинома расчётная кривая всё ближе и ближе подходит к экспериментальным точкам (узлам таблицы), т.е. сумма квадратов разностей экспериментальных и расчётных значений функции отклика уменьшается. Но с ростом степени полинома уменьшается и число степеней свободы в знаменателе, поэтому в процессе наращивания степени полинома наблюдается минимум дисперсии, что и является условием прекращения счёта.

Форма 20 позволяет обрабатывать различные зависимости, находить производные, получать коэффициенты аппроксимирующего полинома, но следует иметь в виду, что экстраполяция найденных зависимостей за пределы исследованного интервала очень ненадёжна.

3. СПЛАЙН-АППРОКСИМАЦИЯ

Для сглаживания данных, дифференцирования экспериментальных зависимостей и интерполяции очень удобна сплайн-аппроксимация (форма 21). Совершенство метода заключается в том, что не вся кривая, а каждый отрезок кривой между экспериментальными точками аппроксимируется полиномом третьей степени. Для того, чтобы переход от точки к точке был плавным, без изломов, коэффициенты полиномов находятся из условия равенства значений собственно функций отклика, а также первой и второй производных соседних полиномов в их общей точке [16]. Естественным недостатком метода является невозможность представления сплайн-функции в виде математического выражения. Рис. 3.1 иллюстрирует возможности метода.

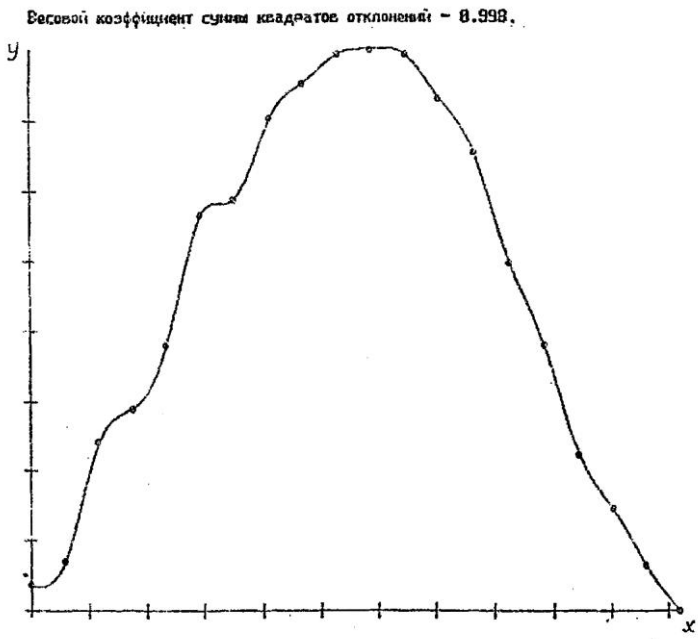

Рис. 3.1. Пример, иллюстрирующий возможности сплайн-аппросимации

4. РЕГРЕССИОННЫЙ АНАЛИЗ УРАВНЕНИЯ

Вычисление коэффициентов некоторого уравнения, произведённое по результатам экспериментальных исследований, ещё не означает, что полученное уравнение может быть использовано для решения научных и технологических задач. Практика показывает, что успешное преодоление вычислительных тонкостей не гарантирует достоверности полученного уравнения и расчёты по нему могут не соответствовать действительности. Поэтому полученное уравнение подвергают специальному исследованию, заключающемуся в проверке коэффициентов на значимость, а уравнения - на адекватность в сравнении с достигнутой точностью экспериментов. Такое исследование называется регрессионным анализом. Строго говоря, основа регрессионного анализа закладывается значительно раньше, до собственно экспериментального исследования. В процессе разработки методики эксперимента выявляются все факторы, влияющие на процесс, устраняются грубые и систематические ошибки. Отладка методики сопровождается и завершается постановкой опытов на воспроизводимость. Опытами на воспроизводимость или параллельными опытами называются опыты, поставленные в строго одинаковых условиях, обычно 5-10, иногда больше. По результатам этих опытов вычисляется дисперсия воспроизводимости:

$$s^2_{оп} = \frac{1}{n_{оп}-1} \sum_{i=1}^{n_{оп}} (y_i - \bar{y})^2, \qquad (4.1)$$

где

$$\bar{y} = \frac{1}{n_{оп}} \sum_{i=1}^{n_{оп}} y_i. \qquad (4.2)$$

В формулах (4.1) и (4.2) $n_{оп}$ - число опытов на воспроизводимость, а $n_{оп}-1=\nu_{оп}$ - число степеней свободы дисперсии воспроизводимости. Дисперсию воспроизводимости ещё называют опытной дисперсией (от Experiment - опыт, термина принятого в англо-американской литературе), но значение русского слова "воспроизводимость" больше соответствует физической сущности дисперсии (4.1).

В экспериментах прикладного, технологического направления считается нормальным, если относительная ошибка экспериментального определения параметра не превышает 2-3%. Парадокс заключается в том, что чем выше точность эксперимента, тем труднее получить уравнение,

адекватное опытным данным; другими словами, чем больше дисперсия воспроизводимости, тем большая вероятность получить уравнение, адекватное фактическому эксперименту, но в действительности не соответствующее сущности процесса или явления. Это объясняется тем, что через массив экспериментальных точек можно провести множество кривых линий, каждая из которых будет описываться соответствующим уравнением согласно тому или иному критерию оптимальности. И чем грубее эксперимент, чем больше разброс экспериментальных значений функции отклика, тем выше вероятность того, что принятая априори функция ляжет в прокрустово ложе адекватности.

4.1. Оценка значимости коэффициентов

Оценка значимости коэффициентов уравнения регрессии заключается в проверке соответствующей статистической гипотезы. Подходящей нулевой гипотезой в этом случае будет гипотеза о равенстве коэффициента уравнения регрессии нулю:

$$H_0 : b_j = 0.$$

Альтернативной гипотезой в этом случае будет неравенство

$$H_1 : b_j \neq 0.$$

Некоторая неопределённость ответа на вопрос о равенстве коэффициента b_j нулю заключается в том, что при вычислении коэффициентов уравнения регрессии b_0 и b_1 по формулам (2.14) и (2.15) ноль получить практически невозможно. Коэффициент b_j может быть очень малым числом, но не нулём, и при этом его величина будет соответствовать его сущности. И, наоборот, коэффициент b_j может быть весьма значителен по величине, но по физической сущности задачи должен быть равен нулю. Проблема может быть решена, если коэффициент b_j сравнивать с его же квадратичным отклонением s_{b_j}, которое, соответственно, может быть и очень малым числом, и очень большим, вопрос в соизмеримости b_j и s_{b_j}. Здесь возможны варианты: $|b_j| < |s_{b_j}|$ и $|b_j| > |s_{b_j}|$ (равенство $b_j = s_{b_j}$ возможно только теоретически). Если $|b_j| < |s_{b_j}|$, то это означает, что ноль попадает в доверительный интервал коэффициента b_j и нулевая гипотеза в первом приближении подтверждается.

Дисперсии коэффициентов b_0 и b_1 вычисляются по формулам

$$s^2_{b_0} = s^2_{оп}\frac{\sum_{i=1}^{n} x_i^2}{D}; \qquad (4.3)$$

$$s^2_{b_1} = s^2_{оп}\frac{n}{D}, \qquad (4.4)$$

где D - главный определитель системы нормальных уравнений. При отсутствии дисперсии воспроизводимости в формулы (4.3) и (4.4) подставляется дисперсия адекватности, но эта замена в некоторых случаях может привести к парадоксам (например, значительная часть или даже все экспериментальные точки могут оказаться за пределами *стандартных границ* корреляционного поля).

Если расчёт коэффициентов производился по уточнённым формулам МНК (см. разд. 2.2.3), то дисперсии s^2_{b0} и s^2_{b1} следует вычислять по формулам

$$s^2_{b_0} = s^2_{ад}\frac{B}{F}; \qquad s^2_{b_1} = s^2_{ад}\frac{n}{FA^2}.$$

Статистику критерия θ можно построить в виде отношения:

$$t_{оп} = \frac{|b_j|}{s_{b_j}}. \qquad (4.5)$$

Опытный t-критерий сравнивается с критерием Стьюдента, табличное значение которого получается независимым путём с помощью t-распределения, плотность вероятности которого имеет вид

$$p(t) = \frac{1}{\sqrt{\pi\nu}} \cdot \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}, \qquad (4.6)$$

где $\Gamma(\nu)$ - гамма-функция; ν - число степеней свободы. При этом следует иметь в виду, что опытный критерий может быть как положителен,

так и отрицателен (как и знак s_{b_j}). Это значит, что для проверки нулевой гипотезы нужно использовать двусторонний критерий, т.е. приняв уровень значимости α для проверки гипотезы H_0, табличное значение критерия θ следует брать для вдвое меньшего значения, т.е. для $\alpha/2$ (рис.1.5). Соответствующие значения критериев будут равны: для левой критической области $t^{1-\alpha/2}$, для правой – $t^{\alpha/2}$. Эта область значений опытного t-критерия будет соответствовать принятию нулевой гипотезы, а точнее, - неуспеху опровержения нулевой гипотезы. Области значений $t^{оп}$, выходящие за пределы $t^{1-\alpha/2} - t^{\alpha/2}$, будут соответствовать опровержению нулевой гипотезы. В данном случае это означает, что $b_j \neq 0$.

Собственно проверка на значимость заключается в сравнении опытного критерия Стьюдента с табличным:

$$t_{оп} = \frac{|b_j|}{s_{b_j}} \Leftrightarrow t^{\alpha}_{\nu_{оп}}, \qquad (4.7)$$

где табличное значение критерия Стьюдента берётся для числа степеней свободы дисперсии воспроизводимости $\nu_{оп}$ и уровня значимости $\alpha/2$. Коэффициент b_j незначим, если опытный критерий Стьюдента меньше табличного. С доверительной вероятностью $P=1-\alpha$ коэффициент b_j значимо отличается от нуля, если опытный критерий Стьюдента больше табличного. Проверку можно осуществить и с помощью *доверительного отклонения*

$$s^{\alpha}_{b_j} = s_{b_j} \cdot t^{\alpha}_{\nu_{оп}}, \qquad (4.8)$$

что более наглядно. Коэффициент b_j незначим, если

$$|b_j| < s^{\alpha}_{b_j}. \qquad (4.9)$$

4.2. Проверка уравнения регрессии на адекватность

Проверка уравнения регрессии на адекватность необходима для уверенности в том, что расчёты по полученному уравнению будут в той или иной степени соответствовать действительности. Сущность вопроса, как придать большей объективности "той или иной степени соот-

ветствия", будет рассмотрена ниже, а сейчас обсудим вопрос о проверке уравнения регрессии на адекватность. Последняя по существу сводится к проверке гипотезы об однородности дисперсии воспроизводимости $s^2_{оп}$ и дисперсии адекватности $s^2_{ад}$.

Дисперсия адекватности

$$s^2_{ад} = \frac{1}{\nu_{ад}} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \qquad (4.10)$$

где n - число опытов в выборке (т.е. число опытов, поставленных с целью вычисления коэффициентов уравнения регрессии); $\nu_{ад}$ - число степеней свободы дисперсии адекватности. В общем случае число степеней свободы $\nu = n - l$, где l - число связей, наложенных на выборку (общее число параметров определённых по выборке). В частном случае $\nu = n - k - 1$, где k - количество независимых переменных. Расчётное значение функции отклика вычисляется по уравнению (2.8) или по найденной оптимальной форме (табл. 2.2). Получение расчётных значений функции отклика заключается в подстановке в уравнение регрессии условий экспериментов (численных значений факторов) для каждого опыта, от $i=1$ до $i=n$, и выполнении соответствующих арифметических операций.

В нашем случае имеются две выборки - экспериментальная и рассчитанная:

$$y_1,\ y_2,\ \ldots,\ y_n;$$

$$\hat{y}_1,\ \hat{y}_2,\ \ldots,\ \hat{y}_n.$$

Необходимо проверить, являются ли эти две выборки частями одной и той же совокупности или это части разных совокупностей. Другими словами, если ошибка в определении оценок коэффициентов полученного уравнения настолько велика, что рассчитанные значения функции отклика формируют другую совокупность, отличную от первой, экспериментально исследуемой, то и некоторые параметры этих выборок должны отличаться. Одним из параметров, характеризующих специфические свойства выборки, является выборочная дисперсия. Дисперсии выборок из одной и той же совокупности называются однородными, из разных выборок - неоднородными. Если дисперсии $s^2_{ад}$ и $s^2_{оп}$ окажутся одно-

родными, то это будет означать, что расчёт по найденному уравнению соответствует экспериментальным данным, т.е. уравнение адекватно, если неоднородны - уравнение неадекватно и пользоваться им нельзя.

Предположим, что первая выборка взята из совокупности с дисперсией σ^2_1, а вторая - из совокупности с дисперсией σ^2_2. Проверяется нулевая гипотеза H_0 о равенстве генеральных дисперсий σ^2_1 и σ^2_2:

$$H_0: \sigma^2_1 = \sigma^2_2.$$

Для того, чтобы подтвердить эту гипотезу, необходимо доказать однородность дисперсий $s^2_{ад}$ и $s^2_{оп}$, а для того, чтобы отвергнуть, - доказать значимость различия $s^2_{ад}$ и $s^2_{оп}$. В качестве статистики θ для проверки гипотезы об однородности дисперсий $s^2_{ад}$ и $s^2_{оп}$ используется критерий Фишера, который в данном случае характеризует предельное соотношение однородных дисперсий. Распределением Фишера называется распределение случайной величины:

$$F=(s^2_1/\sigma^2_1):(s^2_2/\sigma^2_2).$$

Плотность вероятностей F-распределения описывается выражением

$$p(x) = \frac{1}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\cdot\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}\cdot\frac{x^{\nu_1/2-1}}{\left(1+\frac{\nu_1}{\nu_2}x\right)^{(\nu_1+\nu_2)/2}}, \qquad (4.11)$$

где $x>0$, $B(\nu_1,\nu_2)$ - бета-функция; ν_1 - число степеней свободы большей дисперсии; ν_2 - число степеней свободы меньшей дисперсии. То, что F-распределение зависит только от числа степеней свободы ν_1 и ν_2, достаточно очевидно, но результаты табуляции критерия Фишера $F^{\alpha}_{\nu_1,\nu_2}$ ещё зависят от принимаемого уровня значимости α (доверительной вероятности $P=1-\alpha$). В зависимости от проверяемой гипотезы уровень значимости α принимается 0,1, 0,05, 0,01 и 0,001. Доверительная вероятность P имеет непосредственное отношение к сущности вопроса *"той или иной степени соответствия" расчётов по полученному уравнению реальной действительности.* Доверительная вероятность $P=0,95$ принята для проверки большинства научных и технических гипотез. Если проверяемая гипотеза верна с уровнем значимости $\alpha=0,05$,

то это означает, что гипотеза верна с вероятностью 95%, а вероятность того, что гипотеза ошибочна, равна 5% (и не более того).

Функция распределения для $F^{\alpha}_{\nu_1,\nu_2}$ выражается через функцию распределения $B_{\nu_1,\nu_2}(x)$ бета-распределения:

$$P(F^{\alpha}_{\nu_1,\nu_2}<x) = B_{\frac{\nu_1}{2},\frac{\nu_2}{2}}\left(\frac{\frac{\nu_1}{\nu_2}x}{1+\frac{\nu_1}{\nu_2}x}\right) \qquad (4.12)$$

Это соотношение используется для вычисления значений распределения Фишера с помощью таблиц бета-распределения. Если ν_1 и ν_2 целые, то F-распределением Фишера с ν_1 и ν_2 степенями свободы называется распределение F-отношения:

$$F_{\nu_1,\nu_2} = \frac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2}, \qquad (4.13)$$

где $\chi^2_{\nu_1}$ и $\chi^2_{\nu_2}$ - независимые случайные величины, имеющие "*хи-квадрат*" распределения с ν_1 и ν_2 степенями свободы соответственно. F-распределение Фишера играет фундаментальную роль в математической статистике, и особенно как распределение отношения двух выборочных дисперсий $F=(s^2_1/\sigma^2_1):(s^2_2/\sigma^2_2)$. В условиях нулевой гипотезы $\sigma^2_1=\sigma^2_2$, $\sigma^2_1/\sigma^2_2=1$ и $F=(s^2_1/s^2_2)$, т.е. F-распределение может быть непосредственно использовано для оценки отношения выборочных дисперсий s^2_1/s^2_2:

$$F^{оп}_{\nu_{ад},\nu_{оп}} = \frac{S^2_{ад}}{S^2_{оп}}. \qquad (4.14)$$

Опытное значение критерия Фишера $F^{оп}$ сравнивается с табличным значением F^{α} для чисел степеней свободы $\nu_{ад}$ и $\nu_{оп}$ и уровня значимости α. Применительно к данному случаю табличный критерий Фишера характеризует предельное соотношение однородных дисперсий. Если опытное значение критерия Фишера меньше табличного, то дисперсии $s^2_{ад}$ и $s^2_{оп}$ являются однородными и, соответственно, гипотеза H_0 принимается. Это означает, что уравнение регрессии адекватно. Если

$F^{оп} > F^{\alpha}$, то дисперсии $s^2_{ад}$ и $s^2_{оп}$ считаются неоднородными и гипотеза H_0 отклоняется, т.е. полученное уравнение регрессии неадекватно. В этом случае возможны три выхода: попробовать другой метод обработки данных, предположить другой вид функциональной зависимости $y=f(x_1, x_2, \ldots, x_k)$, повысить точность экспериментов или продолжить эксперименты в других точках факторного пространства. Следует обратить внимание на тот факт, что уравнение регрессии вида $y=b_0+b_1x$ может оказаться адекватным и для нелинейной функции отклика в том случае, если эксперимент грубый, если велика дисперсия воспроизводимости $s^2_{оп}$.

В заключение раздела о проверке уравнения на адекватность следует обратить внимание на тот факт, что в соотношении (4.14) в числителе - дисперсия адекватности $s^2_{ад}$, а в знаменателе - дисперсия воспроизводимости $s^2_{оп}$. Было бы правильнее записать отношение большей дисперсии к меньшей, поскольку критерий Фишера в случае однородных дисперсий, в пределе, равен единице, а практически всегда больше единицы. Запись вида (4.14) общепринята потому, что практически дисперсия адекватности бывает почти всегда больше дисперсии воспроизводимости. А вот практическое превышение дисперсии адекватности над дисперсией воспроизводимости объясняется тем, что дисперсия адекватности включает в себя как ошибки эксперимента, так и ошибки, обусловленные приближённым характером уравнения регрессии. Дисперсия воспроизводимости $s^2_{оп}$ включает в себя только ошибки эксперимента.

5. ПРИМЕР РАСЧЁТА

Рассмотрим практическую методику вычисления коэффициентов уравнения регрессии по данным табл. 2.1 (см. рис. 2.4). Дисперсия воспроизводимости определена по результатам 4 параллельных опытов и равна $S^2_{оп}=3,45$.

Для вычисления коэффициентов b_0 и b_1 по формулам (2.15) и (2.16) необходимо вычислить несколько сумм. Поскольку при этом возможны ошибки, желательно их выявить в начале, для чего можно вычислить дополнительно суммы y^2_1 и $(x_1+y_1)^2$ и проверить расчёты по соотношению

$$\sum_{i=1}^{n}(x_i+y_i)^2=\sum_{i=1}^{n}x_i^2+2\cdot\sum_{i=1}^{n}x_iy_i+\sum_{i=1}^{n}y_i^2.$$

Результаты расчётов приведены в табл. 5.1. Подставляя значения сумм в формулы (2.15) и (2.16), получим

$D = 9\cdot1140 - 90^2 = 2160;$

$b_0 = (71\cdot1140 - 90\cdot813)/2160 = 3,597;$

$b_1 = (9\cdot813 - 90\cdot71)/2160 = 0,429.$

Таблица 5.1.

Методика вычисления коэффициентов регрессии

№	x	y	x^2	y^2	xy	$x+y$	$(x+y)^2$
1	2	4,0	4	16	8	6	36
2	4	6.5	16	42,25	26	10,5	110,25
3	6	5,0	36	25	30	11	121
4	8	7.0	64	49	56	15	225
5	10	9.0	100	81	90	19	361
6	12	7,5	144	56,25	90	19,5	380,25
7	14	11,0	196	121	154	25	625
8	16	9,5	256	90,25	152	25,5	650,25
9	18	11,5	324	132,25	207	29,5	870,25
Σ	90	71	1140	613	813	161	3379

Уравнение регрессии, которое с некоторой вероятностью описывает зависимость $y=f(x)$, имеет вид

$$\hat{y} = 3,597 + 0,429x.$$

5.1. *Геометрическая интерпретация коэффициентов регрессии*

Коэффициент b_0 представляет собой отрезок, отсекаемый линией регрессии на оси ординат (см. рис. 2.4) от начала координат.

Коэффициент b_1 представляет собой тангенс угла наклона линии регрессии к оси абсцисс: $tg\alpha=0,429$; $\alpha=23°23'$ (на рис. 2.4 $\alpha\neq23°23'$; дело в том, что равенство $\alpha=23°23'$ справедливо при одинаковом масштабе по осям x и y). На рис. 2.4 изображено "облако" точек с явно выраженной тенденцией: чем больше x, тем больше y (хотя и имеются весьма существенные отклонения от этого закона). Линия регрессии проведена через это "облако" точек наилучшим образом, т.е. принцип Лежандра соблюдён. Положение линии регрессии в системе координат полностью определяется коэффициентами b_0 и b_1.

5.2. *Статистическое оценивание парной корреляции*

Степень тесноты линейной связи характеризует выборочный коэффициент корреляции (1.26). Подставив значения соответствующих сумм в формулу (1.26) получим

$$r_{xy}=\frac{9\cdot813-90\cdot71}{\sqrt{(9\cdot1140-90^2)\,(9\cdot613-71^2)}}=0,9142.$$

Выборочный коэффициент корреляции $r_{xy}=0,9142$ отмечает положительную связь между величинами x и y и достаточно большую долю случайности (большие ошибки наблюдений). Необходимо проверить нулевую гипотезу $H_0:\rho_{xy}=0$. Альтернативной гипотезой будет $H_1:\rho_{xy}\neq0$, т.е. необходимо проверить, значимо ли отличается от нуля выборочный коэффициент корреляции. Нулевая гипотеза предполагает, что две переменные независимы и любое ненулевое значение r возникло из-за случайных флюктуаций. Опытный t-критерий вычисляется по формуле

$$t^{оп}=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}=\frac{0,9142\sqrt{9-2}}{\sqrt{1-0,9142^2}}=5,968>t_7^{0.05}=2,365$$

где 2,365 - двусторонний критерий Стьюдента для уровня значимости $\alpha=0,05$ и числа степеней степеней свободы $\nu=n-2=7$ (табл. П.1). Число

степеней степеней свободы коэффициента корреляции равно $\nu=n-2=7$, вследствие того, что при его вычислении на выборку накладываются две связи в виде $x_{ср}$ и $y_{ср}$. Поскольку значение опытного критерия Стьюдента попадает в критическую область (см. рис. 1.2), в область отклонения гипотезы $H_0: \rho_{xy}=0$, следует признать, что значение $r_{xy}=0,9142$ случайностью не является и между величинами x и y с доверительной вероятностью $P=0,95$ имеется положительная корреляционная связь.

5.3. *Оценка значимости коэффициентов*

Дисперсии коэффициентов b_0 и b_1 вычисляются по формулам (4.3) и (4.4). В практике экспериментальных исследований достаточно часты случаи отсутствия дисперсии воспроизводимости. Кроме этого, не всегда есть возможность ею воспользоваться для вычисления стандартных или квадратичных отклонений коэффициентов (например, при обработке данных по некоторым формам табл. 2.2), поэтому, в данном примере для вычисления квадратичных отклонений коэффициентов b_0 и b_1 воспользуемся дисперсией адекватности. Дисперсию адекватности вычислим по формуле (4.10), предварительно вычислив сумму квадратов разностей экспериментальных значений функции отклика y и рассчитанных по уравнению $y=3,597+0,429x$ (табл. 5.2):

Таблица 5.2.

Вычисление суммы квадратов разностей $(y_i-\hat{y}_i)^2$

№	x	y	$\hat{y}$	$(y_i-\hat{y}_i)$	$(y_i-\hat{y}_i)^2$
1	2	4,0	4,456	-0,455	0,2070
2	4	6,5	5,314	+1,187	1,409
3	6	5,0	6,172	-1,171	1,371
4	8	7,0	7,031	-0,029	0,000841
5	10	9,0	7,889	+1,113	1,2388
6	12	7,5	8,747	-1,245	1,55
7	14	11,0	9,606	+1,397	1,9516
8	16	9,5	10,464	-0,961	0,9235
9	18	11,5	11,322	+0,181	0,03276
Σ	90	71			8,685

$$s^2_{ад}=\frac{1}{\nu_{ад}}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\frac{8,685}{9-2}=1,2407.$$

Квадратичные (стандартные) отклонения коэффициентов b_0 и b_1

$$s^2_{b_0} = s^2_{ад} \cdot \frac{\sum_{i=1}^{n} x_1^2}{D} = 1,24 \frac{1140}{2160} = 0,6548;$$

$$s^2_{b_1} = s^2_{ад} \cdot \frac{n}{D} = 1,24 \frac{9}{2160} = 0,005169.$$

Таким образом $b_0=3,6\pm0,81$; $b_1=0,43\pm0,072$.

Опытный критерий Стьюдента по формуле (4.5):

$$t_0 = \frac{|b_0|}{s_{b_0}} = \frac{3,6}{0,8} = 4,44 > t_7^{0,05} = 2,365;$$

$$t_1 = \frac{|b_1|}{s_{b_1}} = \frac{0,43}{0,07} = 6,14 > t_7^{0,05} = 2,365,$$

где 2,365 - табличное значение двустороннего критерия Стьюдента для уровня значимости $\alpha=0,05$ и числа степеней свободы $\nu=9-2=7$ (см. табл. П.1).

Поскольку опытные значения обоих критериев Стьюдента попадают в критическую область (см. рис. 1.2), мы вынуждены отклонить нулевую гипотезу о равенстве коэффициентов b_0 и b_1 нулю и принять альтернативную, т.е. b_0 и b_1 значимо отличаются от нуля.

По аналогии с доверительным интервалом (1.20) можно записать

$$b_0 - s_{b0} t^{\alpha} < \beta_0 < b_0 + s_{b0} t^{\alpha};$$

$$b_1 - s_{b1} t^{\alpha} < \beta_1 < b_1 + s_{b1} t^{\alpha},$$

где β_0 и β_1 - генеральные параметры искомого уравнения. В нашем случае это означает, что с доверительной вероятностью 0,95 математическое ожидание коэффициента b_0 может находиться в интервале 1,708÷5,492, а математическое ожидание коэффициента b_1 - в интервале 0,2645÷0,6945.

5.4. *Построение стандартных границ корреляционного поля*

Начинающие исследователи обычно с трудом принимают решение о том, какие экспериментальные точки "выпадают" из той или иной зависимости, а какие нет. Иногда задача "отсева" осложняется множественностью версий о виде зависимости $y=f(x)$ (см. табл. 2.2). Интуитивное представление о выпадающих точках формируется в процессе анализа достаточно большого количества данных. В случаях формального выбора оптимальной формы из некоторого множества подходящих форм уравнений парной регрессии полезно одновременно с анализом степени линеаризации (см. разд. 2.6) обращать внимание на стандартные границы корреляционного поля. Стандартные границы корреляционного поля формируются в результате выбора максимальных и минимальных значений $y_{расч}$, получающихся по следующей схеме:

$$\hat{y}_1=(b_0\pm s_{b0})+(b_1\pm s_{b1})x_1.$$

Результаты расчётов и выбора минимальных и максимальных значений функции отклика представлены в табл. 5.3 и на рис. 5.1.

Таблица 5.3.

Стандартные границы корреляционного поля для выборки при обработке её по уравнению $y=(3,6\pm0,81)+(0,43\pm0,072)x$

№	x	y	$\hat{y}$	y_{min}	y_{max}
1	2	4,0	4,456	3,5	5,41
2	4	6,5	5,314	4,22	6,41
3	6	5,0	6,172	4,93	7,41
4	8	7,0	7,031	5,64	8,41
5	10	9,0	7,889	6,36	9,42
6	12	7,5	8,747	7,07	10,42
7	14	11,0	9,606	7,79	11,42
8	16	9,5	10,464	8,50	12,42
9	18	11,5	11,322	9,22	13,43

Очевидно, что стандартные границы корреляционного поля образуют своеобразный коридор, в котором находится большая часть экспериментальных точек. Точка 2 с координатами $x=4$ и $y=6,5$ находится за пределами этого коридора. Это означает, что в этом случае ошибка определения значения y превысила среднюю квадратичную ошибку эксперимента, равную $\sqrt{s^2_{ад}}=\pm1,11$. Это является формальным основанием для исключения точки 2 из таблицы данных и повторения расчёта.

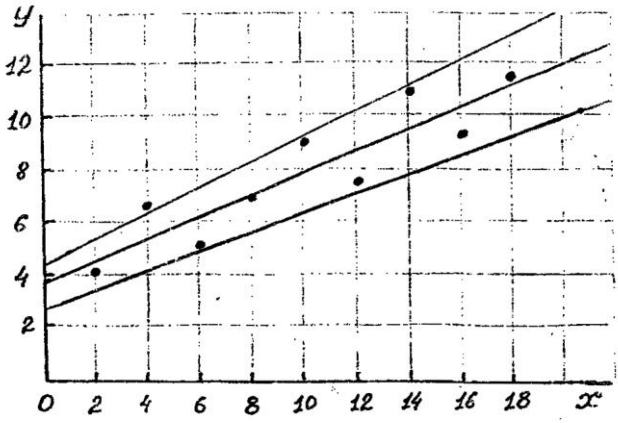

Рис. 5.1. Стандартные границы корреляционного поля зависимости $y=(3,6\pm0,81)+(0,43\pm0,072)x$

В результате расчётов параметров уравнения регрессии $y=b_0+b_1x$ без точки 2 получено уравнение

$$y=3,018+0,469x.$$

Стандартные отклонения коэффициентов $s_{b0}=\pm0,892$ и $s_{b1}=\pm0,0753$, дисперсия адекватности $s^2_{ад}=1,13$. Коэффициенты значимы, все точки находятся в пределах стандартных границ корреляционного поля (табл. 5.4).

Таблица 5.4.

Стандартные границы корреляционного поля при обработке выборки по уравнению $y=(3,018\pm0,892)+(0,469\pm0,0753)x$

№	x	y	$y_{расч}$	y_{min}	y_{max}
1	2	4,0	3,956	2,91	5,00
3	6	5,0	5,833	4,49	7,18
4	8	7,0	6,772	5,28	8,27
5	10	9,0	7,711	6,066	9,36
6	12	7,5	8,649	6,85	10,44
7	14	11,0	9,588	7,64	11,53
8	16	9,5	10,526	8,43	12,62
9	18	11,5	11,465	9,22	13,71

5.5. *Проверка уравнения регрессии на адекватность*

Проверяется нулевая гипотеза $H_0: \sigma^2_{оп} = \sigma^2_{ад}$, где $\sigma^2_{оп}$ – значение генеральной дисперсии воспроизводимости, а $\sigma^2_{ад}$ – адекватности соответственно. Альтернативной гипотезой будет их различие $H_1: \sigma^2_{оп} \neq \sigma^2_{ад}$. Равенство генеральных дисперсий $\sigma^2_{оп}$ и $\sigma^2_{ад}$ соответствует адекватности уравнения регрессии эксперименту.

Проверка уравнения регрессии на адекватность сводится к проверке однородности двух выборочных дисперсий – дисперсии адекватности и дисперсии воспроизводимости по соотношению (1.38). В нашем случае дисперсия адекватности $s^2_{ад}=1,13$ меньше дисперсии воспроизводимости $s^2_{оп}=3,45$. Поскольку таблицы критерия Фишера начинаются с единицы, в числитель следует поставить большую дисперсию

$$F^{оп}_{3,6} = \frac{s^2_{оп}}{s^2_{ад}} = \frac{3,45}{1,13} = 3,05,$$

где 3 – число степеней свободы числителя ν_1; 6 – число степеней свободы знаменателя ν_2.

Табличное значение критерия Фишера для уровня значимости $\alpha=0,05$ и соответствующих чисел степеней свободы (табл. П.2):

$$F^{0.05}_{3,6} = 5,41 > F^{оп}_{3,6} = 3,05.$$

Поскольку опытное значение критерия Фишера меньше критического, у нас нет оснований отклонить нулевую гипотезу об отсутствии различия между дисперсиями адекватности и воспроизводимости, т.е. эти дисперсии однородны. Следовательно, уравнение регрессии адекватно.

Примечание. Необходимо заметить, что процедура отсева точек, выходящих за стандартные границы корреляционного поля, не вполне корректна. Правильнее было бы строить границы корреляционного поля с той или иной доверительной вероятностью, т.е. при расчёте границ корреляционного поля использовать соответствующее значение критерия Стьюдента.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Ахназарова С.Л., Кафаров В.В.* Методы оптимизации эксперимента в химической технологии: Учеб. пособ. для хим.-технол. спец. вузов.- 2-е изд., перераб. и доп.- М.: Высш. шк., 1985.- 327 с.
2. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики.- 3-е изд., М.: 1983.
3. *Брандт З.* Статистические методы анализа наблюдений: Пер. с англ. Г.А. Погребинского / Под ред. В.Ф. Писаренко. М.: Мир, 1975. 312 с.
4. *Брандт Дж., Пирсол А.* Измерение и анализ случайных процессов. М.: Мир, 1974.
5. *Ганджумян Р.А.* Математическая статистика в разведочном бурении: Справ. пособ. - М.: Недра, 1990.- 218 с.
6. *Дьяконов В.П.* Справочник по алгоритмам и программам на языке Бейсик для персональных ЭВМ: Справочник.- М.: Наука, 1987. - 240 с.
7. *Дэвис Дж.С.* Статистический анализ данных в геологии: Пер. с англ. В 2 кн./Под ред. Д.А.Родионова.- М.: Недра, 1990.- 427 с.
8. *Зайдель А.Н.* Погрешности измерений физических величин. Л.: Наука, 1985.
9. *Каждан А.В., Гуськов О.И.* Математические методы в геологии: Учебник для вузов.- М.: Недра, 1990. - 251 с.
10. *Кендалл М.Дж., Стьюарт А.* Статистические выводы и связи: Пер. с англ.- М.: Наука. - 1973. - 900 с.
11. *Кендалл М.Дж., Стьюарт А.* Теория распределений: Пер с англ.- М.: Наука, - 1966.- 588 с.
12. *Козловский Е.А., Питерский В.М., Комаров М.А.* Кибернетика в бурении.- М.: Недра, 1982.
13. *Колемаев В.А., Калинина В.Н.* Теория вероятностей и математическая статистика. М.: ИНФРА-М, 1997.
14. Компьютерный прогноз месторождений полезных ископаемых/ В.В.Марченко, Н.В.Межеловский, Э.А.Немировский и др.- М.: Недра, 1990.- 285 с.
15. *Крамер Г.* Математические методы статистики: Пер. с англ. Изд. 2-е.- М.: Мир, 1975. - 648 с.

16. *Лаудон Т.* ЭВМ и машинные методы в геологии/Пер. с англ. Д.А.Родионова.- М.:Мир, 1981.- 318 с.
17. *Лоусон Ч., Хенсон Р.* Численное решение задач методом наименьших квадратов: Пер. с англ. М.: Наука, 1986.
18. *Закс Лотар.* Статистическое оценивание. М.: Статистика, 1976.
19. *Львовский Е.Н.* Статистические методы построения эмпирических формул: Учеб. пособ.- М.: Высш. шк., 1982.- 224 с.
20. *Тейлор Дж.* Введение в теорию ошибок. М.: Мир, 1985.
21. *Чистяков В.П.* Курс теории вероятностей. М.: Наука. 1978.
22. *Эберт К., Эдерер Х.* Компьютеры. Применение в химии: Пер. с нем. М.: Мир, 1988.

Таблица П.1.

Критические значения одностороннего критерия Стьюдента при ν степенях свободы и заданном уровне значимости α

Число степеней свободы ν	Уровень значимости α					
	0,1	0,05	0,025	0,01	0,005	0,001
1	3,078	6,314	12,706	31,821	63,657	318,310
2	1,886	2,920	4,303	6,965	9,925	22,327
3	1,638	2,353	3,182	4,541	5,841	10,215
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,893
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,785
8	1,397	1,860	2,306	2,896	3,355	4,501
9	1,383	1,833	2,262	2,821	3,250	4,297
10	1,372	1,812	2,228	2,764	3,169	4,144
11	1,363	1,796	2,201	2,718	3,105	4.025
12	1,356	1,782	2,179	2,681	3,055	3,930
13	1,350	1,771	2,160	2,650	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,377	1,746	2,120	2,583	2,921	3,686
17	1,333	1,740	2,110	2,567	2,898	3,646
18	1,330	1,734	2,101	2,552	2,878	3,610
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,080	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,500	2,807	3,485
24	1,318	1,711	2,064	2,492	2,797	3,467
25	1,316	1,708	2,060	2,485	2,787	3,450
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421
28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,310	1,697	2,042	2,457	2,750	3,385
40	1,303	1,684	2,021	2,423	2,704	3,307
60	1,296	1,671	2,000	2,390	2,660	2,232
120	1,289	1,658	1,980	2,358	2,617	3,160
∞	1,282	1,645	1,960	2,326	2,576	3,090

Таблица П.2

Критические значения распределения Фишера F^{α} для уровня значимости α=0,05

ν_2	Число степеней свободы числителя ν_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,15	6,09	6,04	6.00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4.35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3.34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,22	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,95	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Таблица П.3

Критические значения распределения Фишера F^{α} для уровня значимости α=0,025

ν_2	Число степеней свободы числителя, ν_1																	
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	499,5	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28 71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18.00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,15	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5.42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Таблица П.4

Критические значения распределения Фишера F^{α} для уровня значимости α=0,001

ν_2	Число степеней свободы числителя, ν_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4053*	5000*	5404*	5625*	5764*	5859*	5929*	5981*	6023*	6056*	6107*	6158*	6209*	6235*	6261*	6287*	6313*	6340*	6366*
2	998,5	999,0	999,2	999,2	999,3	999,3	999,4	999,4	999,4	999,4	999,4	999,4	999,4	999,5	999,5	999,5	999,5	999,5	999,5
3	167,0	148,	141,1	137,1	134,6	132,8	131,6	130,6	129,9	129,2	128,3	127,4	126,4	125,9	125,4	125,0	124,5	124,0	123,5
4	74,14	61,25	56,18	53,44	51,71	50,53	49,66	49,00	48,47	48,05	47,41	46,76	46,10	45,77	45,43	45,09	44,75	44,40	44,05
5	47,18	37,12	33,20	31,09	29,75	28,84	28,16	27,64	27,24	26,92	26,42	25,91	25,39	25,14	24,87	24,60	24,33	24,06	23,79
6	35,51	27,00	23,70	21,92	20,81	20,03	19,46	19,03	18,69	18,41	17,99	17,56	17,12	16,89	16,67	16,44	16,21	15,99	15,75
7	29,25	21,69	18,77	17,19	16,21	15,52	15,02	14,63	14,33	14,08	13,71	13,32	12,93	12,73	12,53	12,33	12,12	11,91	11,70
8	25,42	18,49	15,83	14,39	13,49	12,86	12,40	12,04	11,77	11,54	11,19	10,84	10,48	10,30	10,11	9,92	9,73	9,53	9,33
9	22,86	16,39	13,90	12,56	11,71	11,13	10,70	10,37	10,11	9,89	9,57	9,24	8,90	8,72	8,55	8,37	8,19	8,00	7,81
10	21,04	14,91	12,55	11,28	10,48	9,92	9,52	9,20	8,96	8,75	8,45	8,13	7,80	7,64	7,47	7,30	7,12	6,94	6,76
11	19,69	13,81	11,56	10,35	9,58	9,05	8,66	8,35	8,12	7,92	7,63	7,32	7,01	6,85	6,68	6,52	6,35	6,17	6,00
12	18,64	12,97	10,80	9,63	8,89	8,38	8,00	7,71	7,48	7,29	7,00	6,71	6,40	6,25	6,09	5,93	5,76	5,59	5,42
13	17,81	12,31	10,21	9,07	8,35	7,86	7,49	7,21	6,98	6,80	6,52	6,23	5,93	5,78	5,63	5,47	5,30	5,14	4,97
14	17,14	11,78	9,73	8,62	7,92	7,43	7,08	6,80	6,58	6,40	6,13	5,85	5,56	5,41	5,25	5,10	4,94	4,77	4,60
15	16,59	11,34	9,34	8,25	7,57	7,09	6,74	6,47	6,26	6,08	5,81	5,54	5,25	5,10	4,95	4,80	4,64	4,47	4,31
16	16,12	10,97	9,00	7,94	7,27	6,81	6,46	6,19	5,98	5,81	5,55	5,27	4,99	4,85	4,70	4,54	4,39	4,23	4,06
17	15,72	10,66	8,73	7,68	7,02	6,56	6,22	5,96	5,75	5,58	5,32	5,05	4,78	4,63	4,48	4,33	4,18	4,02	3,85
18	15,38	10,39	8,49	7,46	6,81	6,35	6,02	5,76	5,56	5,39	5,13	4,87	4,59	4,45	4,30	4,15	4,00	3,84	3,67
19	15,08	10,16	8,28	7,26	6,62	6,18	5,85	5,59	5,39	5,22	4,97	4,70	4,43	4,29	4,14	3,99	3,84	3,68	3,51
20	14,82	9,95	8,10	7,10	6,46	6,02	5,69	5,44	5,24	5,08	4,82	4,56	4,29	4,15	4,00	3,86	3,70	3,54	3,38
21	14,59	9,77	7,94	6,95	6,32	5,88	5,56	5,31	5,11	4,95	4,70	4,44	4,17	4,03	3,88	3,74	3,58	3,42	3,26
22	14,38	9,61	7,80	6,81	6,19	5,76	5,44	5,19	4,99	4,83	4,58	4,33	4,06	3,92	3,78	3,63	3,48	3,32	3,15
23	14,19	9,47	7,67	6,69	6,08	5,65	5,33	5,09	4,89	4,73	4,48	4,23	3,96	3,82	3,68	3,53	3,38	3,22	3,05
24	14,03	9,34	7,55	6,59	5,98	5,55	5,23	4,99	4,80	4,64	4,39	4,14	3,87	3,74	3,59	3,45	3,29	3,14	2,97
25	13,88	9,22	7,45	6,49	5,88	5,46	5,15	4,91	4,71	4,56	4,31	4,06	3,79	3,66	3,52	3,37	3,22	3,06	2,89
26	13,74	9,12	7,36	6,41	5,80	5,38	5,07	4,83	4,64	4,48	4,24	3,99	3,72	3,59	3,44	3,30	3,15	2,99	2,82
27	13,61	9,02	7,27	6,33	5,73	5,31	5,00	4,76	4,57	4,41	4,17	3,92	3,66	3,52	3,38	3,23	3,08	2,92	2,75
28	13,50	8,93	7,19	6,25	5,66	5,24	4,93	4,69	4,50	4,35	4,11	3,86	3,60	3,46	3,32	3,18	3,02	2,86	2,69
29	13,39	8,86	7,12	6,19	5,59	5,18	4,87	4,64	4,45	4,29	4,05	3,80	3,54	3,41	3,27	3,12	2,97	2,81	2,64
30	13,29	8,77	7,05	6,12	5,53	5,12	4,82	4,58	4,39	4,24	4,00	3,75	3,49	3,36	3,22	3,07	2,92	2,76	2,59
40	12,61	8,25	6,60	5,70	5,13	4,73	4,44	4,21	4,02	3,87	3,64	3,40	3,15	3,01	2,87	2,73	2,57	2,41	2,23
60	11,97	7,76	6,17	5,31	4,76	4,37	4,09	3,87	3,69	3,54	3,31	3,08	2,83	2,69	2,55	2,41	2,25	2,08	1,89
120	11,38	7,32	5,79	4,95	4,42	4,04	3,77	3,55	3,38	3,24	3,02	2,78	2,53	2,40	2,26	2,11	1,95	1,76	1,54
∞	10,83	6,91	5,42	4,62	4,10	3,74	3,47	3,27	3,10	2,96	2,74	2,51	2,27	2,13	1,99	1,84	1,66	1,45	1,00

* Эти значения надо умножить на 100

ОГЛАВЛЕНИЕ

Цивинский Дмитрий Николаевич

Разнообразие
форм уравнений парной регрессии

Редактор: С.И. Костерина
Технический редактор Г.Н. Шанькова

ЛР №620595 от 09.07.97.

Подп. в печать 30.05.01.
Формат 60*84 1/16. Бум. типогр. N2.
Печать офсетная,
Усл.кр.-отт. 4,42. Уч.-изд.л. 4,35.
Тираж 170 экз. С-351.

Самарский государственный технический
университет
443100, г. Самара, ул. Молодогвардейская,
д. 244, Главный корпус.