Министерство образования и науки Российской Федерации
Ивановский государственный химико-технологический университет

М.А. Лысова

МЕТОДЫ ОПТИМИЗАЦИИ И ОРГАНИЗАЦИИ ЭНЕРГО- И РЕСУРСОСБЕРЕГАЮЩИХ ХИМИКО-ТЕХНОЛОГИЧЕСКИХ СИСТЕМ. ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИЕ МОДЕЛИ

Учебное пособие

Иваново
2018

УДК 519.2

Лысова, М.А. Методы оптимизации и организации энерго- и ресурсосберегающих химико-технологических систем. Вероятностно-статистические модели: учеб. пособие / М.А. Лысова; Иван. гос. хим.-технол. ун-т. – Иваново, 2018. – 92 с.

Учебное пособие содержит теоретический материал по основным методам оптимизации и организации энерго- и ресурсосберегающих химико-технологических систем. В нём рассматривается статистическая обработка экспериментальных данных для синтеза химико-технологических систем. В каждом параграфе изложены основные теоретические сведения, приведены примеры обработки одномерных, двумерных и многомерных статистических моделей с применением различных пакетов прикладных программ. Включено большое количество заданий для выполнения лабораторных работ.

Рекомендуется студентам магистратуры направления 18.04.02 «Энерго- и ресурсосберегающие процессы в химической технологии, нефтехимии и биотехнологии».

Печатается по решению редакционно-издательского совета Ивановского государственного химико-технологического университета.

Рецензенты:

кафедра естественно-научных дисциплин Ивановской пожарно-спасательной академии ГПС МЧС России;
кандидат технических наук Е.В. Виноградова (Ивановский государственный политехнический университет).

ВВЕДЕНИЕ

Эмпирические вероятностно-статистические модели строятся на основе экспериментальных данных. Все измеряемые величины и данные лабораторных измерений являются случайными величинами. Поэтому обработку совокупности данных следует проводить вероятностно-статистическими методами.

Предметом математической статистики является изучение сложных динамических систем путем объяснения (реальных наблюдений), управления, предсказания. Математически это приводит к необходимости построения модели, позволяющей решить триединую задачу:

– упорядочить данные (полученные в результате наблюдения, сначала надо каким-либо образом обработать, представить в удобном для обозрения и анализа виде);

– оценить статистики (хотя бы приблизительно, интересующие нас характеристики наблюдаемой случайной величины);

– проверить статистические гипотезы (т. е. решение вопроса согласования результатов оценивания с опытными данными).

§ 1. Статистическая обработка экспериментальных данных для синтеза моделей химико-технологических систем

1.1. Теоретическое обоснование

Любые практические исследования связаны со сбором информации о функционировании объектов с последующим ее представлением в форме таблиц, графиков или других математических зависимостей, характеризующих свойства этого объекта. Получение информации при этом связано с организацией различных измерений на исследуемом объекте, которые искажаются случайными не учитываемыми факторами, например, погрешностью измерительных приборов, погрешностью используемой методики измерения и др. Поэтому для оценки достоверности измерения физической переменной следует использовать статистические методы.

Статистические методы оперируют со случайными величинами и позволяют оценивать погрешности измерения физических величин, а также достоверность зависимостей, построенных на их основе.

Под *случайной величиной* понимается переменная, которая в результате испытания в зависимости от случая принимает одно из возможного множества своих значений.

Случайные величины обозначаются прописными латинскими буквами: X, Y, Z и т.д., а их значения – строчными буквами. Например, случайная величина X может принимать значения $x_1,\ x_2, \ldots, x_n$.

Для *дискретной случайной величины* множество возможных значений конечно или счётно, для *непрерывной случайной величины* – несчётно.

Например, случайная величина X – суточная добыча нефти в одной скважине, т – это дискретная случайная величина, так как множество её значений счётно. Случайная величина Y – безотказное время работы прибора, ч – это непрерывная случайная величина, так как множество её значений полуинтервал $(0; +\infty)$.

Для характеристики случайной величины недостаточно иметь набор её допустимых значений. Для полной ее характеристики необходимо указать также, как часто случайная величина может принимать те или иные значения, то есть указать вероятности этих значений.

Законом распределения случайной величины называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.

Для дискретной случайной величины закон распределения может быть задан в виде таблицы или графически. Закон распределения непрерывной случайной величины нельзя задать с помощью вероятностей отдельных значений. Поэтому для непрерывных случайных величин рассматривается вероятность того, что в результате опыта случайная величина принимает значения, меньшие некоторого заданного вещественного числа x. Эта вероятность является функцией от x:

$$F(x) = P(X < x)$$

и называется *функцией распределения* случайной величины.

Для непрерывной случайной величины используется также производная функции распределения – *функция плотности* распределения случайной величины $f(x) = F'(x)$.

Для описания случайных величин часто используются их *числовые характеристики* – числа, в сжатой форме выражающие наиболее существенные черты распределения случайной величины.

Математическое ожидание $M(X)$ случайной величины X характеризует центр рассеяния случайной величины и приближенно равно среднему значению случайной величины. Математическое ожидание определяется выражениями:

$$M(X) = \sum_{i=1}^{n} x_i \cdot p_i \text{, если } X \text{ – дискретна;} \qquad (1.1)$$

$$M(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx \text{, если } X \text{ – непрерывна.} \qquad (1.2)$$

Дисперсия $\sigma^2(X)$ характеризует разброс значений случайной величины относительно её центра (математического ожидания): $\sigma^2(X) = M\left[\left(X - M(X)\right)^2\right]$ и имеет размерность квадрата случайной величины. Дисперсия определяется согласно выражениям:

$$\sigma^2(X) = \sum_{i=1}^{n} x_i^2 \cdot p_i - \left(M(X)\right)^2 \text{, если } X \text{ – дискретна;} \qquad (1.3)$$

$$\sigma^2(X) = \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx - \left(M(X)\right)^2 \text{, если } X \text{ – непрерывна.} \qquad (1.4)$$

Среднее квадратичное отклонение $\sigma(X)$ также характеризует разброс значений случайной величины, но имеет размерность самой случайной величины и определяется следующим образом:

$$\sigma(X) = \sqrt{\sigma^2(X)}\,. \qquad (1.5)$$

Некоторые законы распределения имеют важное практическое значение, в частности *нормальный закон распределения*, функция плотности которого имеет вид:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \qquad (1.6)$$

где $a = M(X)$ – математическое ожидание; $\sigma = \sigma(X)$ – среднее квадратичное отклонение.

Для нормально распределенной случайной величины числа a и σ являются параметрами функции плотности. График функции плотности нормального закона распределения называют нормальной или гауссовой кривой (рис. 1.1).

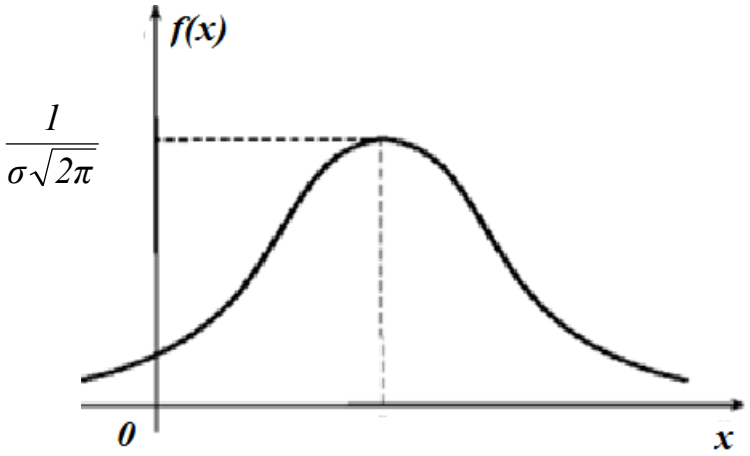


Рис. 1.1. Нормальная кривая

При изучении законов распределения случайной величины о характере закона судят по значениям, принимаемым случайной величиной. Предположим, что проводятся повторные измерения некоторой случайной величины. Назовем исходом опыта результат одного такого измерения. Тогда совокупность возможных исходов опыта представляет собой *генеральную совокупность* значений случайной величины. Числовые характеристики случайной величины, полученные по генеральной совокупности, называются *генеральными параметрами*.

На практике часто невозможно получить генеральную совокупность, а следовательно, и генеральные параметры. Поэтому проводят конечное число n опытов. Совокупность значений случайной величины при этом называется *выборкой объема* n из генеральной совокупности.

Для формирования по выборке достоверного суждения об исследуемом признаке генеральной совокупности необходимо, чтобы объём выборки был достаточно велик и предметы выборки правильно представляли генеральную совокупность, т. е. выборка должна быть *репрезентативной* (представительной). Репрезентативность выборки достигается случайным отбором результатов измерений (наблюдений), например, с применением таблицы случайных чисел.

Из случайного характера выборок следует, что суждение о генеральной совокупности также является случайным. Функция распределения, получаемая

по выборке объема n, называется *выборочной*, или *эмпирической функцией распределения*.

Пусть случайная величина X приняла m_1 раз значение x_1, m_2 раз значение x_2 и т.д., m_r раз значение x_r. Причем $m_1 + m_2 + \ldots + m_r = n$ – объем выборки. Значения случайной величины $x_1, x_2, \ldots x_r$ называются *вариантами выборки*, а соответствующие их количества в выборке $m_1, m_2, \ldots m_r$ – *частотами*.

Упорядоченная по величине последовательность выборочных значений (равные между собой члены выборки нумеруются в произвольном порядке) $x_1 \le x_2 \le \ldots \le x_r$ называется *вариационным рядом*, а расстояние $R = x_r - x_1$ между крайними членами вариационного ряда – *размахом вариационного ряда*.

Для дискретной случайной величины естественной формой эмпирического закона являются *таблицы частот* или *таблицы эмпирического* (*статистического*) *распределения выборки*, показывающие, с какой частотой наблюдалось то или иное значение величины *X*.

Относительной (или *эмпирической*) *частотой* значения x_i в выборке объёма *n* является отношение

$$p_i = \frac{m_i}{n},$$

где m_i – число повторений значения x_i (*варианты*) в выборке.

Таблица частот выглядит следующим образом:

Варианты (x_i)	x_1	x_2	. . .	x_r
Относительные частоты $\left(p_i = \frac{m_i}{n}\right)$	$\frac{m_1}{n}$	$\frac{m_2}{n}$	. . .	$\frac{m_r}{n}$

Для непрерывной случайной величины эмпирический закон распределения задают с помощью *интервальной таблицы частот*, имеющей вид:

Интервалы значений (x_i)	(c_1; c_2)	(c_2; c_3)	. . .	(c_k ; c_{k+1})
Относительные частоты ($p_i = \frac{m_i}{n}$)	$\frac{m_1}{n}$	$\frac{m_2}{n}$	. . .	$\frac{m_k}{n}$

Здесь в первой строке записаны интервалы изменения (группировки) величины X: весь диапазон изменения величины X разбит на k *частичных интервалов* (границами i-го интервала являются c_i и c_{i+1}, $h_i = c_{i+1} - c_i$ – длина i -го частичного интервала). Чаще всего используются интервалы одинаковой длины $h = h_i$, $i = \overline{1, k}$.

Для наглядного представления эмпирического распределения используются графические изображения вариационных рядов в виде: гистограммы, полигона и кумуляты. Гистограмма (рис.1.2) и полигон представляют собой графическое представление эмпирического закона распределения, а кумулята – эмпирической функции распределения случайной величины X. *Гистограммой частот* называется ступенчатый график, состоящий из прямоугольников, основаниями которых служат частичные интервалы Δx_i, а площади равны частотам m_i.

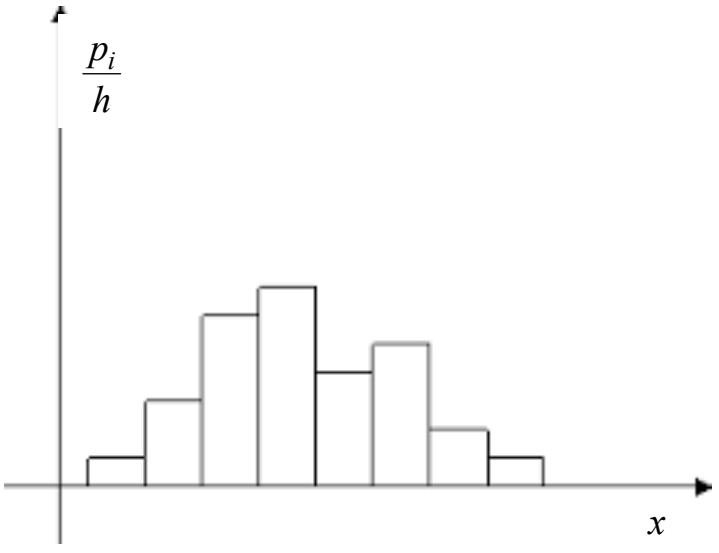


Рис. 1.2. Гистограмма

Любая оценка параметра Θ выборки (обозначим её $\widetilde{\Theta}$) представляет собой некоторую функцию $\widetilde{\Theta} = \widetilde{\Theta}(x_1, \ldots, x_n)$, зависящую от выборки и поэтому

являющуюся случайной величиной. Такие оценки $\widetilde{\Theta}$ называются *статистиками*.

Если в качестве оценки параметра $\widetilde{\Theta}$ ищется одно число, то оценка называется *точечной*.

Эмпирическим (или *выборочным*) *средним* $\bar{x}_в$ называют математическое ожидание оцениваемого признака, вычисленное для выборки по таблице эмпирического распределения:

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ , для выборки без повторений;} \qquad (1.7)$$

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot m_i \text{ , для выборки с повторениями,} \qquad (1.8)$$

где n – объём выборки; x_i – i-е значение (или среднее значение для интервала) оцениваемого признака; m_i – частота повторяющихся значений (или i-го интервала).

Эмпирической (или *выборочной*) *дисперсией* $\sigma_в^2$ называют дисперсию оцениваемого признака, вычисленную для выборки по таблице эмпирического распределения:

$$\sigma_в^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\bar{x}_в\right)^2 \text{ , для выборки без повторений;} \qquad (1.9)$$

$$\sigma_в^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \cdot m_i - \left(\bar{x}_в\right)^2 \text{ , для выборки с повторениями.} \qquad (1.10)$$

Для получения несмещённой оценки дисперсии надо ввести поправку. Получаем *исправленную выборочную дисперсию*:

$$s_в^2 = \frac{n}{n-1}\sigma_в^2. \qquad (1.11)$$

Выборочным среднеквадратическим отклонением $\sigma_в$ называют значение квадратного корня из эмпирической дисперсии оцениваемого признака:

$$\sigma_в = \sqrt{\sigma_в^2}. \qquad (1.12)$$

Аналогично несмещённой оценкой среднего квадратического отклоне-

ния является *исправленное выборочное среднее квадратическое отклонение*:

$$s_в = \sqrt{s_в^2}. \tag{1.13}$$

Однако если число наблюдений мало, то случайный характер величины $\widetilde{\Theta}$ может привести к значительному расхождению между Θ и $\widetilde{\Theta}$. Возникает задача *интервальной* оценки параметра Θ, т.е. оценки не одним числом, а интервалом $(\widetilde{\Theta}_1, \widetilde{\Theta}_2)$ так, чтобы вероятность поглощения этим интервалом параметра Θ, т.е. вероятность двойного неравенства

$$\widetilde{\Theta}_1(x_1, x_2, ..., x_n) < \Theta < \widetilde{\Theta}_2(x_1, x_2, ..., x_n),$$

была не меньше заданного числа γ $(0 \le \gamma \le 1)$.

Вероятность α называется *доверительной вероятностью* (или *уровнем доверия*), а соответствующий интервал $(\widetilde{\Theta}_1, \widetilde{\Theta}_2)$ – *доверительным интервалом* с уровнем доверия γ.

Доверительный интервал для математического ожидания a нормального распределения с уровнем доверия γ при известном среднеквадратическом отклонении σ генеральной совокупности определяется из соотношения:

$$\overline{x_в} - t \cdot \frac{\sigma}{\sqrt{n}} < a < \bar{x}_в + t \cdot \frac{\sigma}{\sqrt{n}}, \tag{1.14}$$

где t – аргумент функции Лапласа Φ(t), для которого $\Phi(t) = \frac{\gamma}{2}$ (определяется из таблицы, представленной в приложении А).

Доверительный интервал для a при неизвестном σ определяется из соотношения:

$$\bar{x} - t_\gamma \cdot \frac{s_в}{\sqrt{n}} < a < \bar{x} + t_\gamma \cdot \frac{s_в}{\sqrt{n}}, \tag{1.15}$$

где t_γ – значение распределения Стьюдента функции с $f = n - 1$ степенями свободы и уровнем доверия γ (определяется из таблицы, представленной в приложении Б); s – исправленное среднее квадратическое отклонение.

Интервальной оценкой (с надежностью γ) среднего квадратического отклонения $\sigma(X)$ нормально распределенной случайной величины X определяем по формулам:

$$\begin{gathered} s(1-q) < \sigma(X) < s(1+q) \text{ при } q < 1 \text{ или} \\ 0 < \sigma(X) < s(1+q) \text{ при } q > 1, \end{gathered} \tag{1.16}$$

где q находим по приложению В.

Под *статистической гипотезой* понимается любое предположение относительно произвольной статистики, которое можно проверить, опираясь на результаты наблюдений в случайной выборке.

В качестве статистических гипотез возможно рассмотрение предположений: о виде неизвестного распределения, параметрах генеральной совокупности известных распределений, равенстве параметров двух или нескольких распределений, независимости выборок.

Обычно исследование начинается с того, что какая-либо гипотеза, которая из неформальных соображений представляется хорошо согласующейся с ожидаемыми эмпирическими данными, объявляется *основной* (*нулевой*) и обозначается H_0. *Альтернативная* (*конкурирующая*) гипотеза, утверждающая, что гипотеза H_0 неверна, обозначается H_1. В действительности верна одна и только одна из гипотез (H_0 или H_1), но какая именно – заранее неизвестно. В этой связи строится *процедура проверки гипотезы* (*критерий согласия*), позволяющая по результатам наблюдений принимать или отвергать данную гипотезу.

Выборочное пространство X разбивается на две непересекающихся области: I_γ и $I_{1-\gamma}$. *Уровень доверия* γ (вероятность попадания в область I_γ) при-

нимается близким к единице. Величину $\alpha = 1 - \gamma$ называют *уровнем значимости* и обычно берут равной 0,05; иногда 0,01 или 0,005.

Правило проверки гипотезы формулируется следующим образом.

Если результаты наблюдений $x \in I_\gamma$, то считается, что гипотеза H_0 подтверждается эмпирическими данными, т. е. гипотеза H_0 принимается. Если же выборочное значение $x \in I_{1-\gamma}$, то предполагается, что данная гипотеза H_0 не согласуется с результатами наблюдений, т.е. H_0 отвергается.

Область $I\gamma$ называют *областью принятия гипотезы*, область $I_{1-\gamma}$ – *критической областью* (или *критерием гипотезы*).

Применение процедуры проверки гипотезы сопряжено с ошибками двух родов:

отвергнуть гипотезу, когда она верна (*ошибка первого рода*);

принять гипотезу, когда она неверна (*ошибка второго рода*).

Если выборочные данные $x \in I_{1-\gamma}$, тогда с вероятностью ошибки первого рода наблюдается случайное событие, которое противоречит гипотезе. Если вероятность такого случайного события незначительна, значит, наблюдается практически невозможное событие и данная гипотеза должна быть отвергнута.

Для проверки гипотез о виде распределения используют критерий Пирсона χ^2 (хи-квадрат). В качестве основной гипотезы (H_0) принимается утверждение о том, что исследуемая генеральная совокупность имеет заданный закон распределения. Альтернативная гипотеза H_1, как правило, не высказывается, и задача ставится так: согласуются ли данные выборки с предполагаемым законом распределения?

Для проверки гипотезы H_0 критерий Пирсона имеет вид:

$$\chi^2_{набл} = \sum_{i=1}^{k} \frac{(m_i - m_i')^2}{m_i'}, \qquad (1.17)$$

где m_i – число наблюдений, попавших в i-й интервал выборки; m_i' – ожидаемое (для предполагаемого закона распределения) число попаданий в i-й интервал; k – число интервалов.

Далее по таблице критических точек распределения χ^2 (приложение Г) для заданного уровня значимости α и числа степеней свободы $f = k - 3$ (где k – число частичных интервалов) находится критическая точка $\chi^2_{кр}$. Принимается одно из решений:

1) принять гипотезу H_0 (генеральная совокупность распределена по предполагаемому закону), если $\chi^2_{набл} < \chi^2_{кр}$;

2) отклонить гипотезу H_0, если $\chi^2_{набл} \geq \chi^2_{кр}$.

1.2. Пример выполнения лабораторной работы № 1 «Обработка результатов измерений одной случайной величины»

Дан протокол измерений случайной величины X – рабочих дебитах газовой скважины (тыс. м^3/сут.). Для этой случайной величины требуется:

1) составить интервальную таблицу частот и относительных частот;

2) построить гистограмму частот и аппроксимировать гистограмму теоретическим нормальным законом распределения;

3) методом произведений рассчитать выборочные характеристики $\bar{x}_в$, $s^2_в$, $s_в$;

4) проверить гипотезу о наличии грубых ошибок наблюдений;

5) с помощью критерия χ^2 проверить согласованность теоретического и статистического законов распределений;

6) получить точечные оценки для математического ожидания, дисперсии и среднего квадратического отклонения;

7) с надежностью $\gamma = 0{,}95$ найти доверительный интервал для математического ожидания и среднего квадратического отклонения.

Значения случайной величины:

484 394 375 318 471 308 252 269 390 251 352 287 464 417 304

448 291 379 315 469 381 400 342 456 383 442 365 393 385 284

274 339 322 194 377 230 332 339 414 392 244 426 311 398 404

216 373 288 297 279 457 455 320 287 484 182 322 326 322 365

359 379 378 319 251 362 295 261 339 329 251 403 417 268 375

282 403 381 306 381 377 410 182 266 308 470 363 370 327 375

524 371 295 307 408 260 395 468 252 297

Решение:

1. Записав исходные данные в порядке возрастания, получим следующий упорядоченный вариационный ряд (табл. 1.1):

Таблица 1. 1.

Упорядоченный вариационный ряд

182	251	268	287	304	318	327	352	371	378	385	400	417	457	484
182	251	269	288	306	319	329	359	373	379	390	403	417	464	524
194	252	274	291	307	320	332	362	375	379	392	403	426	468	
216	252	279	295	308	322	339	363	375	381	393	404	442	469	
230	260	282	295	308	322	339	365	375	381	394	408	448	470	
244	261	284	297	311	322	339	365	377	381	395	410	455	471	
251	266	287	297	315	326	342	370	377	383	398	414	456	484	

Найдем минимальное и максимальное значения случайной величины: $x_{\min} = 182$, $x_{\min} = 524$.

Найдем размах (разность между максимальным и минимальным значениями): $R = 524 - 182 = 342$.

Необходимо всю совокупность значений разбить на группы. Число групп можно определить по формуле Стерджесса:

$$k = 1 + 3{,}322 \cdot \lg n. \qquad (1.17)$$

Лучше брать меньшее число групп.

У нас объем выборки $n = 100$, тогда $k = 1 + 3{,}322 \cdot \lg 100 = 7{,}44$. Возьмем число групп $k = 7$.

Для нашей совокупности значений размах R=342, а число групп (интервалов) равно 7. Необходимо определить длину одного частичного интервала. Длина одного интервала должна быть конечным числом (не обязательно целым, но не иррациональным). Так как $\frac{342}{7} = 48{,}85...$ необходимо увеличить величину размаха (например, незначительно уменьшить минимальное значение случайной величины или незначительно увеличить максимальное значение случайной величины или и то и другое). В нашем примере увеличим максимальное значение случайной величины до 525. Тогда длина интервала $h = \frac{525-182}{7} = 49$.

Построим интервальную таблицу частот и относительных частот (табл. 1.2).

Таблица 1.2

Интервал $[x_i; x_{i+1})$	Среднее значение $\overline{x}_i$	Частота m_i	Относительная частота $p_i = \frac{m_i}{n}$	Плотность частоты $y_i = \frac{m_i}{h}$
182 – 231	206,5	5	0,05	0,10
231 – 280	255,5	13	0,13	0,27
280 – 329	304,5	25	0,25	0,51
329 – 378	353,5	20	0,20	0,41
378 – 427	402,5	24	0,24	0,49
427 – 476	451,5	10	0,10	0,20
476 – 525	500,5	3	0,03	0,06
Сумма	-	100	1	-

2. Построим гистограмму частот. По оси Ох откладываем значения случайной величины, по оси Оу соответствующие значения плотности частот (y_i). Ширина каждого столбика гистограммы равна длине частичного интервала, а высота – плотности частоты.

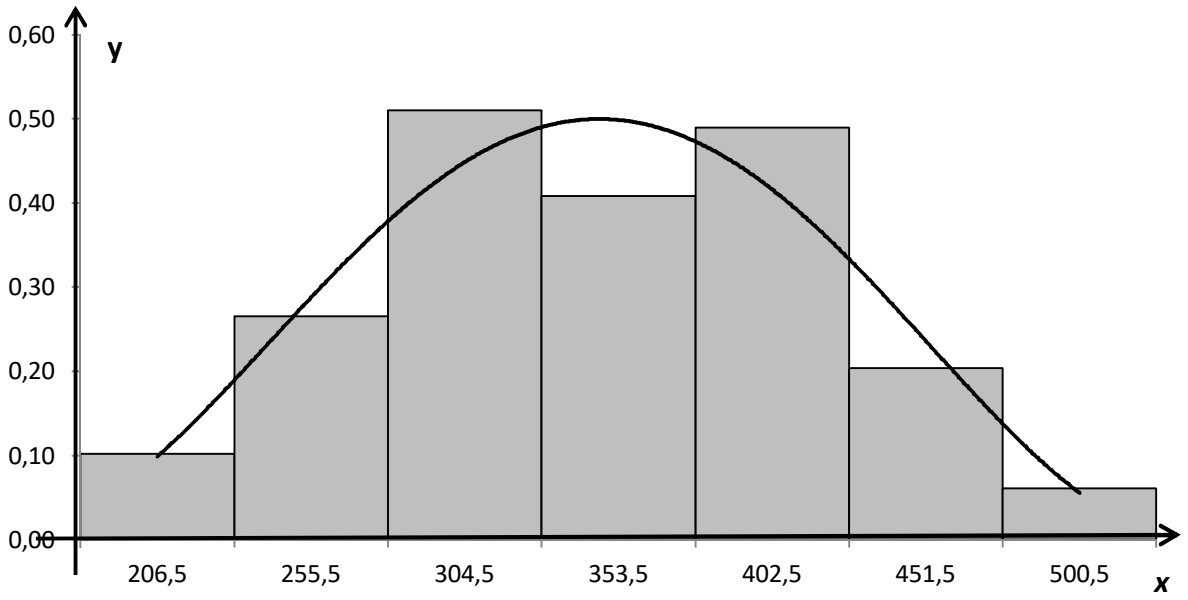


Рис.1.3. Гистограмма частот

На гистограмме изображаем кривую, проходящую через середины столбиков. По виду кривой (график кривой схож с графиком функции плотности нормального закона распределения) можно предположить, что исследуемая случайная величина подчиняется нормальному закону распределения.

3. Методом произведений вычислим числовые характеристики (табл. 1.3)

Таблица 1.3

№ п/п	**Интервал** $[x_i; x_{i+1})$	**Середина интервала** $\widetilde{x}_i$	**Частота распределения** m_i	**Условная варианта** η_i	η_i^2	$\eta_i \cdot m_i$	$\eta_i^2 \cdot m_i$
1	182 – 231	206,5	5	-2	4	-10	20
2	231 – 280	255,5	13	-1	1	-13	13
3	280 – 329	304,5	25	0	0	0	0
4	329 – 378	353,5	20	1	1	20	20
5	378 – 427	402,5	24	2	4	48	96
6	427 – 476	451,5	10	3	9	30	90
7	476 – 525	500,5	3	4	16	12	48
Σ			100			87	287

Для интервала с наибольшей частотой принимаем значение условной варианты, равной нулю. В нашем случае наибольшая частота $m_3 = 25$ соответствует третьему частичном интервалу, поэтому $\eta_3 = 0$.

a_0 =304,5 – середина частичного интервала с наибольшей частотой;

$$S_1=\sum_{i=1}^{7}\eta_i\cdot m_i=87;$$

$$S_2=\sum_{i=1}^{7}\eta_i^2\cdot m_i=287.$$

Тогда выборочное среднее: $\bar{x}_в=a_0+\frac{\Delta}{n}\cdot S_1=304{,}5+\frac{49}{100}\cdot 87=347{,}13$ тыс. м^3/сут.; выборочная дисперсия: $D_в=\frac{\Delta^2}{n}\left(S_2-\frac{S_1^2}{n}\right)=\frac{49^2}{100}\left(287-\frac{87^2}{100}\right)=$ 5073,5531; исправленная выборочная дисперсия $s_в^2=\frac{n}{n-1}\cdot\sigma_в^2=$ $=\frac{100}{99}\cdot 5073{,}5531=$ 5124,8011; выборочное среднее квадратическое отклонение: $\sigma_в=\sqrt{\sigma_в^2}=\sqrt{5073{,}5531}=71{,}23$ тыс. м^3/сут.; исправленное среднее квадратическое отклонение $s=\sqrt{S^2}=\sqrt{5124{,}8011}=71{,}59$ тыс. м^3/сут.

4. Проверим гипотезу о наличии грубых ошибок наблюдений. Выдвигаем гипотезы:

H_0 : грубой погрешности (промаха) нет;

H_1: грубая погрешность (промах) есть.

Для проверки гипотезы используем правило «трех сигм»: если $|\bar{x}_в-x_i|>3\sigma_в$, то гипотеза H_0 отвергается, принимается гипотеза H_1: грубая погрешность есть.

Проверим наименьшее $x_1=182$ и наибольшее значения $x_{100}=524$ выборки. Для $x_1=182$ имеем $|347{,}13-182|=165{,}13\le 3\cdot 71{,}23=213{,}69$, то есть для значения $x_1=182$ принимаем гипотезу H_0 об отсутствии грубых погрешностей наблюдений. Для $x_{100}=524$ имеем $|347{,}13-524|=176{,}13\le 3\cdot 71{,}23=213{,}69$, то есть также принимаем гипотезу об отсутствии грубых ошибок наблюдений.

В случае, если для какого-либо значения x_i принимается гипотеза H_1 (грубая ошибка есть), следует исключить это значение из выборки и заново построить интервальный статистический ряд и вычислить выборочные характеристики (то есть выполнить пункты 1, 2, 3 лабораторной работы).

5. При проверке гипотезы о нормальном распределении выборочной совокупности используем критерий Пирсона. Результаты вычислений представлены в табл. 1.4.

Таблица 1.4

№ п/п	x_i	x_{i+1}	z_i	z_{i+1}	$\Phi(z_i)$	$\Phi(z_{i+1})$	P_i	m_i'	m_i	$\frac{(m_i - m_i')^2}{m_i'}$
1	182	231	$-\infty$	-1,63	-0,5	-0,4484	0,0516	5,16	5	0,0050
2	231	280	-1,63	-0,94	-0,4484	-0,3264	0,1220	12,20	13	0,0525
3	280	329	-0,94	-0,25	-0,3264	-0,0987	0,2277	22,77	25	0,2184
4	329	378	-0,25	0,43	-0,0987	0,1664	0,2651	26,51	20	1,5986
5	378	427	0,43	1,12	0,1664	0,3686	0,2022	20,22	24	0,7066
6	247	476	1,12	1,81	0,3686	0,4649	0,0963	9,63	10	0,0142
7	476	525	1,81	∞	0,4649	0,5	0,0351	3,51	3	0,0741
Σ							1	100	100	2,6694

В таблице использованы следующие обозначения:

$z_i = \frac{x_i - \bar{x}_в}{\sigma_в}$; $z_{i+1} = \frac{x_{i+1} - \bar{x}_в}{\sigma_в}$ ($z_1 = -\infty$, $z_7 = +\infty$ всегда). Например, $z_2 = \frac{231 - 347{,}13}{71{,}23} = -1{,}63$.

$\Phi(z_i) = \frac{1}{\sqrt{2\pi}} \int_0^{z_i} e^{-\frac{t^2}{2}} dt$ – значение интегральной функции Лапласа (приложение А.);

$P_i = \Phi(z_{i+1}) - \Phi(z_i)$ (контроль: сумма по столбцу =1);

$m_i' = N \cdot m_i$ – теоретические частоты (контроль: сумма по столбцу =100).

Таким образом, экспериментальное значение критерия Пирсона $\chi^2_{эксп} = \sum_k \frac{(m_i - m_i')^2}{m_i'} = 2{,}6694$. По таблице критических точек распределения χ^2 (при-

ложение Г) находим критическую точку $\chi^2_{кр}(\alpha = 0{,}05;\ f = k - 3 = 7 - 3 = 4) = 9{,}5$, то есть $\chi^2_{кр} > \chi^2_{эксп}$. Значит расхождение экспериментальных и теоретических частот незначимо, а следовательно, можно сделать вывод, что данные эксперимента согласуются с гипотезой о нормальном распределении совокупности. Таким образом, за закон распределения случайной величины X – рабочих дебитах газовой скважины (тыс. м^3/сут.) можно принять нормальный закон распределения.

6. Точечной оценкой математического ожидания $M(X)$ является выборочное среднее $\bar{x}_в =$ 347,13 тыс. м^3/сут.

Несмещенной точечной оценкой дисперсии $D(X)$ является исправленная выборочная дисперсия $s^2_в$=5124,8011.

Несмещенной точечной оценкой среднего квадратического отклонения $\sigma(X)$является исправленное среднее квадратическое отклонение $s_в = 71{,}59$ тыс. м^3/сут.

7. Найдем доверительные интервалы (интервальные оценки) для математического ожидания и для среднего квадратического отклонения.

Интервальной оценкой (с надежностью γ) математического ожидания $M(X)$ нормально распределенной случайной величины X при неизвестном среднем квадратическом отклонении определяем по формуле (1.15). По условию доверительная вероятность γ = 0,95, объем выборки $n = 100$. Тогда по таблице (приложение Б) находим, $t_\gamma = 1{,}98$. Вычислим предельную ошибку $\Delta = t_\gamma \cdot \frac{s}{\sqrt{n}} =$ $1{,}98 \cdot \frac{71{,}59}{\sqrt{100}} = 14{,}20$. Таким образом, доверительный интервал для математического ожидания имеет вид $347{,}13 - 14{,}20 < M(X) < 347{,}13 + 14{,}20$, или $332{,}93 < M(X) < 361{,}33$ тыс. м^3/сут.

Интервальную оценку среднего квадратического отклонения вычисляем по формуле (1.16). По приложению В находим $q = q(n;\gamma) = q(100;0{,}95) = 0{,}143 < 1$, следовательно, используем первую формулу для нахождения доверительного интервала среднего квадратического отклонения:

$71{,}59(1-0{,}143) < \sigma(X) < 71{,}59(1+0{,}143)$, или $61{,}35 < \sigma(X) < 81{,}83$ тыс. $м^3$/сут.

1.3. Задания для выполнения лабораторной работы № 1 «Обработка результатов измерений одной случайной величины»

Дан протокол измерений случайной величины X. Требуется:

1) составить интервальную таблицу частот и относительных частот;

2) построить гистограмму частот и аппроксимировать гистограмму теоретическим нормальным законом распределения;

3) методом произведений рассчитать выборочные характеристики $\overline{x}_в$, $s_в^2$, $s_в$;

4) проверить гипотезу о наличии грубых ошибок наблюдений;

5) с помощью критерия χ^2 проверить согласованность теоретического и статистического законов распределений;

6) получить точечные оценки для математического ожидания, дисперсии и среднего квадратического отклонения;

7) с надежностью $\gamma = 0{,}95$ найти доверительный интервал для математического ожидания и среднего квадратического отклонения.

Вариант 1

Случайная величина X – продолжительность фонтанирующих скважин, мес.:

34,0	23,5	29,5	34,3	12,8	28,4	13,2	26,1	20,7
28,1	28,5	17,2	28,0	36,8	27,8	16,8	27,5	32,2
27,2	26,5	27,0	31,8	32,0	23,0	17,1	28,5	37,1
22,5	34,0	22,7	25,2	28,0	37,0	28,7	26,4	31,5
31,9	25,7	23,5	33,5	24,0	24,5	25,9	25,3	25,4
28,2	26,5	19,0	29,0	21,5	30,8	18,5	28,3	26,2
26,1	19,1	29,1	24,6	25,9	26,0	30,2	24,7	19,8
30,5	21,1	30,9	19,6	20,5	30,3	34,1	30,6	23,2
20,4	36,1	35,0	31,2	22,1	30,6	22,3	31,1	24,8
30,0	14,1	28,7	15,2	34,2	31,3	33,0	31,2	34,0
28,4	39,0	34,4	16,0	30,3	16,1	28,5	30,2	20,5
26,0								

Вариант 2

Случайная величина X – рабочие дебиты газовой скважины, тыс. $м^3$/сут:

400	379	395	382	397	388	400	391	407	292
389	397	383	403	386	393	408	394	391	403
385	398	388	395	410	395	386	398	379	400
389	397	405	397	405	397	388	400	375	397
391	392	407	392	389	402	415	392	408	394
407	392	391	403	369	392	410	392	410	395
386	398	370	395	412	395	386	397	389	400
372	395	417	395	394	400	386	402		

Вариант 3

Случайная величина X – вес метровых отрезков холста, г:

280	278	270	284	271	268	265	280	290	440
382	335	353	400	335	380	300	361	325	345
349	352	399	334	379	297	360	323	344	307
351	385	333	377	347	359	321	343	449	356
391	332	375	345	358	320	342	420	352	368
331	373	339	357	319	341	309	335	367	375
371	295	356	317	340	329	334	366	389	332
292	355	315	339	328	333	365	388	349	409
354	313	338	425	367	364	386	348	407	311
337	289	366	363	384	347	405	360	344	336
305	326	362	382	346	403	369	340	385	

Вариант 4

Случайная величина X – размеры толщины резца, мм:

11,5	10,6	9,2	10,9	11,2	9,5	10,9	10,2	11,3	9,5
9,6	11,6	10,7	9,3	10,0	12,2	9,6	10,5	10,3	11,4
11,5	10,6	9,2	10,4	11,9	9,5	10,9	10,2	11,3	9,5
10,5	9,1	10,3	12,1	9,9	10,8	10,1	11,2	9,9	10,8
9,0	10,2	8,1	9,8	10,7	10,0	11,1	9,8	10,7	10,0
10,1	8,6	9,7	10,6	10,4	11,0	9,7	10,6	10,4	11,5
12,0	9,6	10,5	10,3	11,4	9,6	10,5	10,3	8,8	9,5
9,5	10,9	10,2	11,3	9,5	10,9	10,2	8,3	9,9	10,4
10,8	10,1	11,2	9,9	10,8	10,1	12,9	9,8	10,3	10,6
10,0	11,1	9,8	10,7	10,0	8,7	9,7	10,2	10,5	11,8
11,0	9,7	10,6	10,4	12,4	9,6	10,1	10,9	11,7	9,4

Вариант 5

Случайная величина X – себестоимость единицы продукции, руб.:

86	72	67	84	75	51	77	93	50	70	71	70	58
74	55	79	82	99	69	64	66	57	73	54	88	81
49	68	63	58	56	72	53	76	85	92	67	62	57
90	71	52	87	84	48	66	61	56	89	64	51	78
83	96	70	65	60	80	63	75	76	82	47	69	64
59	79	62	74	70	81	91	69	63	58	78	61	73
68	53	46	67	62	57	77	65	72	66	52	74	66
61	56	76	46	71	68	51	73	86				

Вариант 6

Случайная величина X – количество деталей, выработанных за смену рабочими, шт.:

62	67	63	59	62	66	63	71	61	65	70	59	64
69	65	61	68	65	69	58	63	68	64	60	63	67
64	56	58	62	67	63	59	62	66	63	71	61	65
66	62	58	65	69	62	65	60	64	67	65	61	64
68	65	64	59	63	61	63	60	63	67	64	63	58
62	60	62	71	62	66	63	62	61	65	75	65	66
63	71	62	47	60	64	57	64	69	62	69	65	55
59	63	59	63	68	65	68	65	72	53	62	66	62
73	64	67	64	58	65	55	59	63	59	63	68	65
68	65	64										

Вариант 7

Случайная величина X – прочность 500-миллиметровых образцов одиночной нити основы 26 текс высшего сорта, см:

325	264	343	314	324	282	355	306	379	335
372	385	327	271	345	316	326	286	425	310
245	326	266	344	315	325	284	258	309	374
325	264	343	314	324	282	355	308	379	335
272	342	279	329	313	289	356	307	377	334
341	312	322	287	221	308	375	333	272	359
311	321	285	367	305	371	332	269	350	439
320	283	358	304	367	331	265	349	395	339
281	252	303	365	330	261	348	297	338	243
405	302	363	329	390	347	295	337	415	318
301	361	328	410	346	259	336	399	317	299
281	252	303	365	330	261	348	297	338	243

Вариант 8

Случайная величина X – суточный дебит нефти наблюдаемой скважины, т/сут:

14,8	17,4	19,2	13,1	15,1	14,5	17,1	15,0	16,1
15,6	16,7	17,2	15,0	16,1	15,6	19,1	14,7	19,1
14,7	19,1	15,1	12,5	13,2	16,0	14,6	16,6	15,7
17,3	13,8	17,0	17,4	15,5	19,2	16,1	15,1	14,6
16,9	16,0	16,6	20,5	13,4	14,7	17,5	15,7	17,6
12,7	15,5	17,3	16,2	17,7	19,1	15,2	14,4	16,5
14;8	16,8	15,7	13,5	13,9	16,2	15,9	17,4	17,2
15,4	17,8	16,1	16,5	17,1	16,0	14,2	18,0	14,9
16,3	15,8	13,6	15,9	18,5	15,2	17,8	15,4	15,0
17,7	17,0	16,0	18,1	16,4	14,1	16,3	15,0	15,8
12,7	13,7	15,9	18,0	19,6	15,3	17,9	15,3	

Вариант 9

Случайная величина X – пропускная способность нескольких участков нефтепровода, м^3/сут:

15,5	16,1	21,1	19,4	18,1	17,9	18,2	19,3	18,0
20,1	17,2	18,0	16,2	17,4	18,6	17,1	19,3	17,0
18,1	17,1	18,2	19,2	18,0	20,2	18,6	17,2	19,1

18,6	16,3	18,5	17,1	18,1	19,3	15,6	16,9	19,1
14,5	20,3	18,2	17,2	19,2	18,6	17,4	18,5	16,4
16,9	18,5	17,8	19,2	19,1	16,9	20,4	18,5	16,4
18,2	18,7	19,1	18,5	16,5	17,3	17,8	17,9	21,3
19,1	19,2	20,6	16,8	17,5	15,7	18,7	18,3	18,4
16,6	17,3	18,5	17,0	17,7	19,2	17,6	20,2	19,1
17,2	16,8	18,7	17,5	18,4	16,6	18,3	17,3	19,2
17,7	19,2	17,6	20,3	18,6	17,5	19,1	18,8	16,7
18,4	16,7	15,8	17,2	17,7	18,3			

Вариант 10

Случайная величина X – энергетические затраты на 1 м проходки при разведочном бурении нефтяных скважин в различных районах страны, тыс. руб:

21	36	28	31 .	26	27	27	24	30	36	34	25	33
30	16	36	21	19	26	41	28	24	28	36	29	25
33	34	17	20	28	36	22	44	26	24	28	37	32
25	42	26	18	19	35	26	19	43	21	22	23	38
32	24	31	32	29	19	35	24	28	42	23	27	22
39	22	24	13	26	31	22	20	31	29	41	23	31
27	40	22	27	32	26	24	22	27	28	30	32	26
29	27	31	26	11	12	27	20	14	28	19	30	20
30												

§2. Регрессионно- дисперсионный анализ двумерных моделей химико-технологических систем

2.1. Теоретическое обоснование

Задачами регрессионного анализа являются: установление формы зависимости между переменными, оценка функции регрессии, прогноз значений зависимой переменной.

Виды зависимостей:

1. *Функциональная* – когда каждому значению одной переменной соответствует вполне определенное значение другой. (Применяется часто в естественных науках. Например, зависимость скорости от времени).

2. *Статистическая* – когда каждому значению одной переменной соответствует множество возможных значений (условное статистическое распределение) другой переменной.

3. *Корреляционная* – когда каждому значению одной переменной соответствует определенное условное математическое ожидание $M_x(Y)$ (математическое ожидание случайной переменной Y, вычисленное в предположении, что переменная X приняла значение $X=x$) другой переменной.

Корреляционная зависимость может быть представлена в виде:

$$M_x(Y) = \varphi(x). \qquad (2.1)$$

При этом зависимую переменную ***Y*** называют *результативным признаком* (результирующей, объясняемой, эндогенной переменной), а независимую переменную ***X*** называют *фактором* (объясняющей, входной, экзогенной переменной).

Уравнение (2.1) называется *уравнением регрессии*, функция $\varphi(x)$ – *функцией регрессии*, а её график – *линией регрессии*.

Для точного описания уравнения регрессии необходимо знать условный закон распределения зависимой переменной Y при условии, что переменная X примет значение x, т.е. $X = x$. В статистической практике такую информацию

получить не удается. В этом случае речь может идти об оценке по выборке функции регрессии. Такой оценкой является *выборочная линия регрессии*:

$$\hat{y} = \hat{\varphi}(x, b_0, b_1, \ldots, b_m), \tag{2.2}$$

где $\hat{y}$ – условное среднее значение переменной Y при фиксированном значении переменной $X = x$; $b_0, b_1, \ldots, b_m$ – параметры линии регрессии.

Уравнение (2.2) называется *выборочным уравнением регрессии*.

Истинное значение величины Y складывается из двух слагаемых:

$$y = \hat{y} + \varepsilon, \tag{2.3}$$

где y – фактическое значение результативного признака;

$\hat{y}$ – теоретическое значение результативного признака, найденное исходя из уравнения регрессии;

ε – случайная величина, характеризующая отклонения фактического значения результативного признака от теоретического, включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения (ошибка).

Различают два основных вида оценки функции регрессии $\hat{\varphi}(x)$:

1) линейная

2) нелинейная.

Уравнение парной линейной регрессии:

$$\hat{y} = a + b \cdot x. \tag{2.4}$$

Построение линейной регрессии сводится к нахождению оценок её параметров – a и b. Наиболее распространенным подходом является метод наименьших квадратов, который позволяет получить такие оценки параметров, при которых:

$$S(a,b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (a + bx_i - y_i)^2 \to \min. \tag{2.5}$$

Чтобы найти минимум функции (2.5), необходимо найти частные производные по каждому из параметров a и b и приравнять их к нулю. В результате преобразований получаем систему:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i, \end{cases} \tag{2.6}$$

решая которую находим неизвестные параметры a и b.

Параметр b называется коэффициентом регрессии. Его величина показывает, на сколько в среднем изменяется значение результативного признака Y при изменении значения фактора X на одну единицу. Формально a – значение y при $x=0$. Если признак-фактор x не может иметь нулевого значения, то вышеуказанная трактовка члена a не имеет смысла, т.е. параметр a может не иметь экономического содержания. Можно интерпретировать знак при параметре a. Если $a>0$, то относительное изменение результата происходит медленнее, чем изменение фактора, т.е. коэффициент вариации для результата у меньше коэффициента вариации по фактору x: $V_y < V_x$.

Уравнение регрессии всегда дополняется показателем тесноты связи. При использовании линейной регрессии в качестве такого показателя выступает *линейный коэффициент корреляции* r_{xy}, который определяется по формуле:

$$r_{xy} = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2}}. \tag{2.7}$$

Свойства:

1. $-1 \le r_{xy} \le 1$.

2. Чем ближе $|r_{xy}|$ к единице, тем сильнее линейная связь между признаками.

3. Если $r_{xy} > 0$, то связь между признаками прямая, а если $r_{xy} < 0$, то связь обратная.

4. При $r_{xy} = \pm 1$ имеем строгую функциональную зависимость.

5. При $r_{xy} = 0$ линейная корреляционная связь отсутствует.

Оценим статистическую значимость коэффициента линейной корреляции. Выдвигаем гипотезы:

H_0: $r_{xy} = 0$, т.е. отсутствует линейная зависимость между *x* и *y*.

H_1: $r_{xy} \neq 0$, т.е. существует линейная зависимость между *x* и *y*.

Используется t-критерий Стьюдента. Рассчитываем наблюдаемое значение критерия:

$$t_{набл} = \frac{|r_{xy}|\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}. \qquad (2.8)$$

По таблице распределения Стьюдента (приложение Б) находим критическое значение критерия: $t_{кр} = t(\alpha = 0{,}05; k = n-2)$.

Если $t_{набл} > t_{кр}$, то принимается гипотеза H_1 с вероятностью ошибки α.

Проверить значимость уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных для описания зависимой переменной.

Проверка значимости уравнения регрессии производится на основе дисперсионного анализа, который выделен в самостоятельный элемент статистического анализа.

Согласно основной идее дисперсионного анализа, общая сумма квадратов отклонений переменной y от среднего значения $\bar{y}$ раскладывается на две части – «объясненную» и «необъясненную»:

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2 ,$$

где $\sum (y - \bar{y})^2$ – общая сумма квадратов отклонений; $\sum (\hat{y} - \bar{y})^2$ – сумма квадратов отклонений, объясненная регрессией (или факторная сумма квадратов отклонений); $\sum (y - \hat{y})^2$ – остаточная сумма квадратов отклонений, характеризующая влияние неучтенных в модели факторов.

Схема дисперсионного анализа имеет вид, представленный в табл. 2.1 (n – число наблюдений, m – число независимых параметров при переменной x).

Таблица 2.1

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$\sum (y - \bar{y})^2$	*n-1*	$S^2_{общ} = \frac{1}{n-1}\sum (y - \bar{y})^2$
Факторная	$\sum (\hat{y} - \bar{y})^2$	*m*	$S^2_{факт} = \frac{1}{m}\sum (\hat{y} - \bar{y})^2$
Остаточная	$\sum (y - \hat{y})^2$	*n-m-1*	$S^2_{ост} = \frac{1}{n-m-1}\sum (y - \hat{y})^2$

Одной из наиболее эффективных оценок адекватности регрессионной модели, мерой качества уравнения регрессии, характеристикой прогностической силы анализируемой регрессионной модели является *коэффициент детерминации*, определяемый по формуле:

$$R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} . \qquad (2.9)$$

Величина R^2 показывает, какая часть вариации зависимой переменной обусловлена вариацией объясняющей переменной.

Очевидно, что $0 \le R^2 \le 1$. Чем ближе R^2 к единице, тем лучше регрессия аппроксимирует эмпирические данные, тем теснее наблюдения примыкают к линии регрессии.

Проверка значимости коэффициента детерминации и уравнения регрессии в целом. Выдвигаем гипотезы:

H_0: $R^2 = 0$, т.е. коэффициент детерминации незначим.

H_1: $R^2 \neq 0$, т.е. коэффициент детерминации значим, уравнение регрессии адекватно отражает реальные данные.

Используем F-критерий Фишера-Снедекора. Наблюдаемое значение критерия:

$$F_{набл} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}. \qquad (2.10)$$

По таблице распределения Фишера-Снедекора (приложение Д) находим критическое значение критерия: $F_{кр}(\alpha = 0{,}05; k_1 = m; k_2 = n-m-1)$. Если $F_{набл} > F_{кр}$, то признается статистическая значимость коэффициента детерминации с вероятностью $1-\alpha$.

Кроме того, чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации:

$$A = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| 100\% . \qquad (2.11)$$

Эта величина показывает, какая в среднем может быть допущена ошибка при использовании модели. Модель применима при А, меньшем 10%.

В парной линейной регрессии оценивается также *значимость отдельных параметров уравнения регрессии*, а также оценивается истинное значение параметров.

С этой целью по каждому из параметров предварительно вычисляются стандартные ошибки:

$$m_b = \sqrt{S^2 \frac{1}{\sum(x_i - \bar{x})^2}}; \; S^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}; \qquad (2.12)$$

$$m_a = \sqrt{S^2 \frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}}\,. \qquad (2.13)$$

Для оценки существенности параметра уравнения регрессии проверяется гипотеза:

$H_0 : b = 0,$

$H_1 : b \neq 0.$

Находим экспериментальное значение критерия Стьюдента $t_{эксп} = \frac{b}{m_b}$ и критическое $t_{кр} = t(\alpha = 0{,}05; k = n-2)$.

Если $t_{эксп} > t_{кр}$, гипотеза о несущественности коэффициента b отклоняется.

Аналогично проверяется значимость параметра *a*.

Доверительный интервал для коэффициента уравнения регрессии определяется так:

$$(b - t_{кр} \cdot m_b; b + t_{кр} \cdot m_b),\ (a - t_{кр} \cdot m_a; a + t_{кр} \cdot m_a). \qquad (2.14)$$

Адекватная модель может быть использована для *прогноза*.

Точечный прогноз по модели производится путем подстановки интересующего значения $x = x_p$ в уравнение регрессии:

$\hat{y}_p = a + b \cdot x_p$.

Доверительный интервал для функции регрессии, то есть для условного математического ожидания $M_x(Y)$, который с заданной надежностью (доверительной вероятностью) покрывает неизвестное значение $M_x(Y)$:

$$\hat{y}_p - t_{кр} \cdot s_{\hat{y}_p} \le M_x(Y) \le \hat{y}_p + t_{кр} \cdot s_{\hat{y}_p}, \qquad (2.15)$$

где $s_{\hat{y}_p} = \sqrt{S^2\left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$ – стандартная ошибка групповой средней.

Заметим, что прогноз значений зависимой переменной y по уравнению регрессии оправдан, если прогнозное значение $x = x_p$ объясняющей переменной не выходит за диапазон ее значений по выборке. Использование уравнения регрессии вне предела обследованного диапазона значений объясняющей переменной может привести к значительной погрешности.

Различают два класса *нелинейных регрессий*:

1. Регрессии, нелинейные относительно включенных объясняющих переменных:

– квадратичная функция: $\hat{y} = a + bx + cx^2$;

– гипербола: $\hat{y} = a + \frac{b}{x}$;

– логарифмическая функция: $\hat{y} = a + b\ln x$.

2. Регрессии, нелинейные по оцениваемым параметрам:

– степенная: $\hat{y} = ab^x$;

– показательная: $\hat{y} = ax^b$;

– экспоненциальная: $\hat{y} = e^{a+bx}$.

Нелинейные регрессии приводятся к линейным с помощью некоторых преобразований и замены переменных.

Уравнение нелинейной регрессии дополняется показателем тесноты связи – *индексом корреляции*:

$$\rho_{xy} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}}, \qquad (2.16)$$

где $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ – общая дисперсия результативного признака y;

$\sigma_{ост}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ – остаточная дисперсия.

Величина данного показателя находится в пределах $0 \le \rho_{xy} \le 1$. Чем ближе ρ_{xy} к 1, тем теснее связь рассматриваемых признаков, тем более надежно уравнение регрессии.

Квадрат индекса корреляции называется индексом детерминации и характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака:

$$\rho_{xy}^2 = 1 - \frac{\sigma_{ост}^2}{\sigma_y^2} = \frac{\sigma_{факт}^2}{\sigma_y^2}, \qquad (2.17)$$

где $\sigma_{факт}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

Линеаризация различных видов нелинейных уравнений

1. Уравнение *параболической регрессии* имеет вид:

$$\hat{y} = a + bx + cx^2.$$

Для вычисления оценок параметров параболической регрессии необходимо решить систему из трех уравнений.

$$\begin{cases} an + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i, \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i. \end{cases}$$

В качестве метода решения такой системы можно выбрать метод Крамера или метод Жордана-Гаусса (см. курс линейной алгебры). Предварительно рекомендуется составить вспомогательную таблицу вида:

N	x	y	xy	x^2	x^2y	x^3	x^4
1							
…..	…	…	…	…	…	…	…
n							
Σ							

2. Уравнение *экспоненциальной регрессии* имеет вид:

$$\hat{y} = e^{a+bx}.$$

Линеаризация происходит следующим образом. Логарифмируем: $\ln \hat{y} = \ln e^{a+bx}$, $\Rightarrow$ $\ln \hat{y} = a + bx$. Обозначим $Y = \ln y$. Тогда линейная модель следующая: $\hat{Y} = a + bx$.

Таблица для вычисления оценок параметров модели следующая:

N	x	y	Y	xY	x^2	Y^2
1						
…	…	…	…	…	…	…
Σ						

3. Уравнение *показательной регрессии* имеет вид:

$$\hat{y} = a \cdot b^x.$$

Проведем преобразования: $\ln \hat{y} = \ln a \cdot b^x$, $\ln \hat{y} = \ln a + \ln b^x$, $\ln \hat{y} = \ln a + x \cdot \ln b$. Обозначим: $Y = \ln y$, $A = \ln a$, $B = \ln b$. Получаем линейное уравнение регрессии: $\hat{Y} = A + Bx$. Вспомогательная таблица для преобразованных данных аналогична как для уравнения экспоненциальной регрессии. После вычисления оценок параметров A и B находим $a = e^A$, $b = e^B$.

4. Уравнение *логарифмической регрессии* имеет вид:

$$\hat{y} = a + b \cdot \ln x.$$

Линеаризация производится путем замены $X = \ln x$. Получаем линейную модель $\hat{y} = a + b \cdot X$. Для вычисления оценок параметров a и b составляется вспомогательная таблица вида:

N	x	y	X	Xy	X^2	y^2
1						
…	…	…	…	…	…	…
Σ						

5. Уравнение *гиперболической регрессии*:

$$\hat{y} = a + \frac{b}{x}.$$

Линеаризация производится путем замены $X = \frac{1}{x}$. Получаем линейную модель $\hat{y} = a + b \cdot X$. Для вычисления оценок параметров a и b составляется вспомогательная таблица вида:

N	x	y	X	Xy	X^2	y^2
1						
…	…	…	…	…	…	…
Σ						

2.2. Пример выполнения лабораторной работы № 2 «Регрессионно- дисперсионный анализ двумерных моделей»

Содержание работы: на основании опытных данных требуется:

1. Упорядочить данные по возрастанию. Построить поле корреляции. Сделать вывод о форме связи между признаками.

2. Вычислить линейный коэффициент парной корреляции и проверить его значимость на уровне $\alpha = 0{,}05$.

3. Найти оценки параметров линейного уравнения регрессии методом наименьших квадратов.

4. Вычислить коэффициент детерминации, проверить его значимость на уровне $\alpha = 0{,}05$.

5. Вычислить среднюю относительную ошибку аппроксимации.

6. Оценить статистическую значимость параметров уравнения регрессии на уровне $\alpha = 0{,}05$. Найти доверительные интервалы для параметров регрессии.

7. Найти прогнозное значение результативного фактора $\hat{y}_p$ при значении признака-фактора, составляющем 110% от среднего уровня $x_p = 1{,}1 \cdot \bar{x}$.

8. В предположении, что связь между признаками имеет вид $\hat{y} = a + b\sqrt{x}$, найти параметры, вычислить индекс корреляции, коэффициент детерминации, ошибку аппроксимации.

9. В предположении, что связь между признаками имеет вид $\hat{y} = ax^b$, найти параметры, вычислить индекс корреляции, коэффициент детерминации, ошибку аппроксимации.

10. На одном графике изобразить поле корреляции и три графика исследуемых зависимостей. Сделать вывод о наилучшем уравнении регрессии.

Суть лабораторной работы отражает следующая задача.

Задача. Результаты наблюдений зависимости средней заработной платы от производительности труда по цеху технологической связи. Определить форму связи между факторным и результативным признаком.

Средняя заработная плата y, тыс. дол.	0,9	1,2	1,8	2,2	2,6	2,9	3,3	3,8
Производительность труда x, тыс. дол.	1,2	3,1	5,3	7,4	9,6	11,8	14,5	18,7

Решение. 1. Изобразим полученную зависимость графически точками на координатной плоскости (рис. 2.1). Такое изображение называется *полем корреляции*.

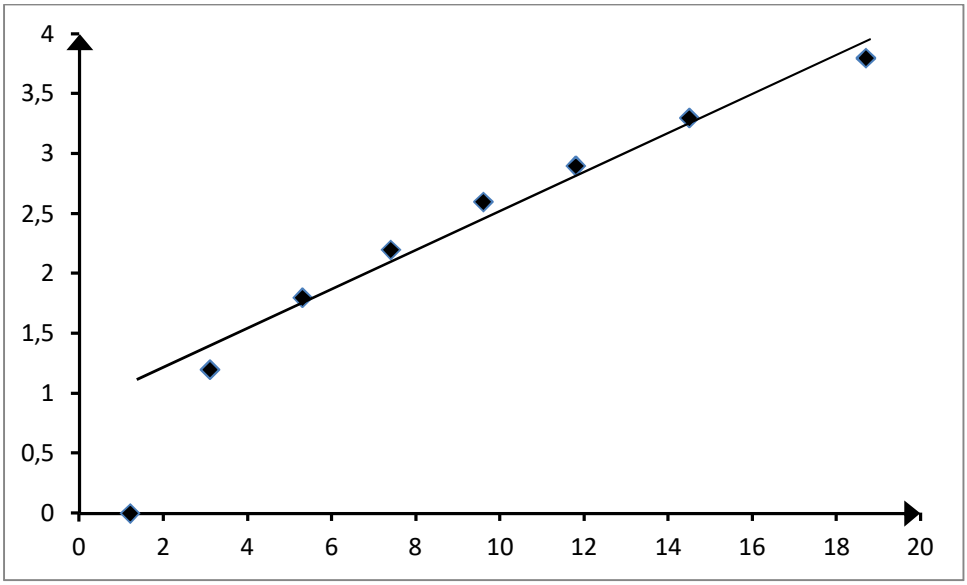


Рис. 2.1. Поле корреляции

По расположению точек можно предполагать наличие линейной корреляционной зависимости между X и Y. Поэтому уравнение регрессии будем искать в виде линейного уравнения (2.4).

2. Для вычисления оценок параметров составим таблицу 2.2.

Таблица 2.2

№ п/п	x	y	xy	x^2	y^2
1	1,2	0,9	1,08	1,44	0,81
2	3,1	1,2	3,72	9,61	1,44
3	5,3	1,8	9,54	28,09	3,24
4	7,4	2,2	16,28	54,76	4,84
5	9,6	2,6	24,96	92,16	6,76
6	11,8	2,9	34,22	139,24	8,41
7	14,5	3,3	47,85	210,25	10,89
8	18,7	3,8	71,06	349,69	14,44
Сумма	**71,6**	**18,7**	**208,71**	**885,24**	**50,83**

Вычислим коэффициент линейной корреляции по формуле (2.8), используя данные табл. 2.2:

$$r_{xy} = \frac{8 \cdot 208{,}71 - 71{,}6 \cdot 18{,}7}{\sqrt{8 \cdot 885{,}24 - 71{,}6^2} \cdot \sqrt{8 \cdot 50{,}83 - 18{,}7^2}} = 0{,}991.$$

Близость коэффициента корреляции к единице указывает на тесную линейную связь между производительностью труда и средней заработной платой.

Оценим статистическую значимость коэффициента линейной корреляции. Выдвигаем гипотезы:

H_0: $r_{xy} = 0$, т.е. отсутствует линейная зависимость между x и y.

H_1: $r_{xy} \neq 0$, т.е. существует линейная зависимость между x и y.

Используется t-критерий Стьюдента. Рассчитываем наблюдаемое значение критерия по формуле (2.8): $t_{набл} = \dfrac{0{,}991\sqrt{8-2}}{\sqrt{1-0{,}991^2}} = 18{,}13.$

По таблице распределения Стьюдента (приложение Б) находим критическое значение критерия: $t_{кр} = t(\alpha = 0{,}05; k = 6) = 2{,}45.$

Так как $t_{набл} > t_{кр}$, то принимается гипотеза H_1 с вероятностью ошибки α, то есть коэффициент корреляции статистически значим с вероятностью 0,95.

3. Для вычисления оценок параметров линейного уравнения регрессии составим систему (2.6), используя данные табл. 2.2:

$$\begin{cases} a \cdot 885{,}24 + b \cdot 71{,}6 = 208{,}71, \\ a \cdot 71{,}6 + b \cdot 8 = 18{,}7. \end{cases}$$

Решая данную систему, получаем $b = 0{,}169;\ a = 0{,}825.$

Получили уравнение $\boxed{\hat{y} = 0{,}825 + 0{,}169 \cdot x}$.

При увеличении дохода производительности труда на 1 тыс. дол., средняя заработная плата увеличивается в среднем на 0,169 тыс. дол.

Изобразим на рис. 2. 1 теоретическую линию уравнения регрессии по двум точкам: $x = 2$, $\hat{y} = 1{,}16$ и $x = 15$, $\hat{y} = 3{,}36$.

4. Рассчитаем коэффициент детерминации по формуле (2.9) и проверим его значимость. Все расчеты сведем в табл. 2.3.

Таблица 2.3

№ n/n	x	y	$\hat{y}$	$\left\|\dfrac{y-\hat{y}}{y}\right\|$	$(y-\hat{y})^2$	$(y-\bar{y})^2$	$(x-\bar{x})^2$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	1,2	0,9	1,028	0,1422	0,016	2,068	60,0625

Окончание табл. 2.3

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
2	3,1	1,2	1,349	0,1242	0,022	1,295	34,2225
3	5,3	1,8	1,721	0,0439	0,006	0,289	13,3225
4	7,4	2,2	2,076	0,0564	0,015	0,019	2,4025
5	9,6	2,6	2,447	0,0588	0,023	0,069	0,4225
6	11,8	2,9	2,819	0,0279	0,007	0,316	8,1225
7	14,5	3,3	3,276	0,0073	0,001	0,925	30,8025
8	18,7	3,8	3,985	0,0487	0,034	2,137	95,0625
Σ	**71,6**	**18,7**	**18,7004**	**0,5094**	**0,125**	**7,119**	**244,42**

где $\hat{y} = 0{,}825 + 0{,}169 \cdot x$.

$$R^2 = 1 - \frac{0{,}125}{7{,}119} = 0{,}982.$$

Коэффициент детерминации показывает, что уравнением регрессии объясняется 98,2% дисперсии результативного признака, а на долю остальных факторов приходится лишь 1,8%.

Оценим качество уравнения регрессии в целом и значимость коэффициента детерминации. Наблюдаемое значение критерия вычисляем по формуле (2.10):

$$F_{набл} = \frac{0{,}982}{1 - 0{,}982} \cdot (8 - 2) = 327{,}3.$$

Табличное значение (приложение Д):

$$F_{кр}(\alpha = 0{,}05; k_1 = 1; k_2 = 8 - 2 = 6) = 5{,}99.$$

$F_{набл} > F_{кр}$, то есть принимаем гипотезу H_1, коэффициент детерминации значим, полученное уравнение регрессии адекватно отражает реальные данные.

5. Вычислим среднюю относительную ошибку аппроксимации по формуле (2.11).

Имеем: $A = \frac{100}{8} \cdot 0{,}5094 = 6{,}38\%$, что говорит о хорошем подборе модели к исходным данным.

6. Оценим статистическую значимость параметров уравнения регрессии. Найдем доверительные интервалы для параметров регрессии.

По формулам (2.12), (2.13) вычислим стандартные ошибки и экспериментальные значения критерия Стьюдента:

$$S^2 = \frac{0{,}125}{6} = 0{,}021;$$

$$m_b = \sqrt{0{,}021 \cdot \frac{1}{244{,}42}} = 0{,}0093;\ m_a = \sqrt{0{,}021 \cdot \frac{885{,}24}{8 \cdot 244{,}42}} = 0{,}0975;$$

$$t_b = \frac{0{,}169}{0{,}0093} = 18{,}17;\ \ t_a = \frac{0{,}825}{0{,}0975} = 8{,}46;$$

$$t_{кр} = t(\alpha = 0{,}05; k = n - 2 = 6) = 2{,}45.$$

Так как $t_b > t_{кр}, t_a > t_{кр}$,то признаем статистическую значимость параметров регрессии и показателя тесноты связи.

Доверительные интервалы для параметров регрессии определяем по формулам (2.14):

$$a \in [0{,}597; 1{,}075],\ b \in [0{,}145; 0{,}191].$$

7. Найдем прогнозное значение результативного фактора $\hat{y}_p$ при значении признака-фактора, составляющем 110% от среднего уровня $x_p = 1{,}1 \cdot \bar{x} = 9{,}845$, т.е. найдем среднюю заработную плату, если производительность труда составляет 9,845 тыс. долл. Прогнозное значение результативного фактора: $\hat{y}_p = 0{,}825 + 0{,}169 \cdot 9{,}845 = 2{,}488$ тыс. дол.

Прогноз для среднего значения результативного фактора по формуле (2.15).

$$s_{\hat{y}_p} = \sqrt{0{,}021 \cdot \left(\frac{1}{8} + \frac{(9{,}845 - 8{,}95)^2}{244{,}42} \right)} = 0{,}052\ \%$$

$$2{,}488 - 0{,}052 \cdot 2{,}447 \le M_x(Y) \le 2{,}488 + 0{,}052 \cdot 2{,}447;$$

$$2{,}361 \le M_x(Y) \le 2{,}615.$$

При производительности труда 9 845 тыс. дол. средняя заработная плата будет находиться в пределах от 2 361 тыс. дол. до 2 615 тыс. дол.

8. Предположим, что связь между признаками имеет вид: $\hat{y} = a + b\sqrt{x}$.

Делаем замену: $z = \sqrt{x}$ и составляем вспомогательную таблицу 2.4. После такой замены уравнение принимает вид: $\hat{y} = a + b \cdot z$, то есть является линейным уравнением регрессии. Поэтому оценку его параметров можно вычислять, составив систему (2.6).

Таблица 2.4

№ п/п	x	z	y	zy	z^2	y^2	$\hat{y}$	$\left\|\frac{y-\hat{y}}{y}\right\|$	$(y-\hat{y})^2$
1	1,2	1,095	0,9	0,986	1,2	0,81	0,7311	0,188	0,0285
2	3,1	1,761	1,2	2,113	3,1	1,44	1,3517	0,126	0,0230
3	5,3	2,302	1,8	4,144	5,3	3,24	1,8569	0,032	0,0032
4	7,4	2,720	2,2	5,985	7,4	4,84	2,2470	0,021	0,0022
5	9,6	3,098	2,6	8,056	9,6	6,76	2,5998	0,000	0,0000
6	11,8	3,435	2,9	9,962	11,8	8,41	2,9140	0,005	0,0002
7	14,5	3,808	3,3	12,566	14,5	10,89	3,2618	0,012	0,0015
8	18,7	4,324	3,8	16,433	18,7	14,44	3,7436	0,015	0,0032
Σ	**71,6**	**22,544**	**18,7**	**60,243**	**71,6**	**50,83**	**18,7059**	**0,398**	**0,0619**

Система для определения неизвестных оценок параметров a и b имеет вид аналогичный (2.6):

$$\begin{cases} a\sum_{i=1}^{n} z_i^2 + b\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} z_i y_i, \\ a\sum_{i=1}^{n} z_i + bn = \sum_{i=1}^{n} y_i, \end{cases}$$

$$\begin{cases} a \cdot 71{,}6 + b \cdot 22{,}544 = 60{,}243, \\ a \cdot 22{,}544 + b \cdot 8 = 18{,}7, \end{cases} \quad \begin{cases} a = 0{,}935, \\ b = -0{,}297. \end{cases}$$

Уравнение регрессии с квадратным корнем имеет вид: $\boxed{\hat{y} = 0{,}935 - 0{,}297\sqrt{x}}$.

Индекс корреляции вычисляем по формуле (2.16): $\rho_{xy} = \sqrt{1 - \frac{0{,}0619}{7{,}119}} = 0{,}9956$.

Индекс детерминации (формула (2.17)) $\rho^2 = 0{,}9912$ показывает, что 99,12 % вариации результативного признака объясняется вариацией признака-фактора, а 0,88% приходится на долю прочих факторов.

Средняя ошибка аппроксимации: $A = \frac{100}{8} \cdot 0{,}398 = 4{,}98\%$ показывает, что линия регрессии хорошо приближает исходные данные.

9. Предположим, что связь между признаками имеет вид: $\hat{y} = ax^b$.

Проведем преобразования: $\ln \hat{y} = \ln ax^b$, $\ln \hat{y} = \ln a + \ln x^b$, $\ln \hat{y} = \ln a + b \ln x$. Обозначим: $Y = \ln y$, $A = \ln a$, $X = \ln x$. Получаем линейное уравнение регрессии: $\hat{Y} = A + bX$.

Составляем вспомогательную таблицу 2.5 для преобразованных данных.

Таблица 2.5

N	***x***	***y***	***X***	***Y***	***XY***	***X²***	***y²***	$\hat{y}$	$\left\|\frac{y-\hat{y}}{y}\right\|$	$(y-\hat{y})^2$
1	1,2	0,9	0,182	-0,1054	-0,019	0,0332	0,0111	*0,8149*	0,0946	0,0072
2	3,1	1,2	1,131	0,1823	0,206	1,2801	0,0332	*1,3747*	0,1456	0,0305
3	5,3	1,8	1,668	0,5878	0,98	2,7812	0,3455	*1,8473*	0,0263	0,0022
4	7,4	2,2	2,001	0,7885	1,578	4,0059	0,6217	*2,2203*	0,0092	0,0004
5	9,6	2,6	2,262	0,9555	2,161	5,1156	0,9130	*2,5627*	0,0143	0,0014
6	11,8	2,9	2,468	1,0647	2,628	6,0915	1,1336	*2,8713*	0,0099	0,0008
7	14,5	3,3	2,674	1,1939	3,193	7,1511	1,4255	*3,2165*	0,0253	0,0070
8	18,7	3,8	2,929	1,335	3,91	8,5763	1,7822	*3,7004*	0,0262	0,0099
Σ	**71,6**	**18,7**	**15,315**	**6,0024**	**14,6367**	**35,0349**	**6,2658**	**18,6081**	**0,3514**	**0,0595**

Система для определения неизвестных оценок параметров A и b имеет вид аналогичный (2.6): $\begin{cases} A\sum_{i=1}^{n} X_i^2 + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_i Y_i, \\ A\sum_{i=1}^{n} X_i + bn = \sum_{i=1}^{n} Y_i, \end{cases}$

$$\begin{cases} A \cdot 35{,}0349 + b \cdot 15{,}315 = 14{,}6367, \\ A \cdot 15{,}315 + b \cdot 8 = 6{,}0024, \end{cases} \quad \begin{cases} A = -0{,}305, \\ b = 0{,}551. \end{cases}$$

Возвращаемся к исходным переменным, потенцируем: $-0{,}305 = \ln a$, $a = e^{-0.305} = 0{,}737$. Получаем показательное уравнение регрессии: $\boxed{\hat{y} = 0{,}737 \cdot x^{0{,}551}}$

Индекс корреляции: $\rho_{xy} = \sqrt{1 - \frac{0{,}0595}{7{,}119}} = 0{,}9958.$

Индекс детерминации $\rho^2 = 0{,}9916$, который показывает, что 99,16 % вариации результативного признака объясняется вариацией признака-фактора, а 0,84% приходится на долю прочих факторов.

Средняя ошибка аппроксимации $A = \frac{100}{8} \cdot 0{,}3514 = 4{,}39\%$ показывает, что линия регрессии хорошо приближает исходные данные.

10. Изобразим полученные результаты на одном графике (рис. 2.2).

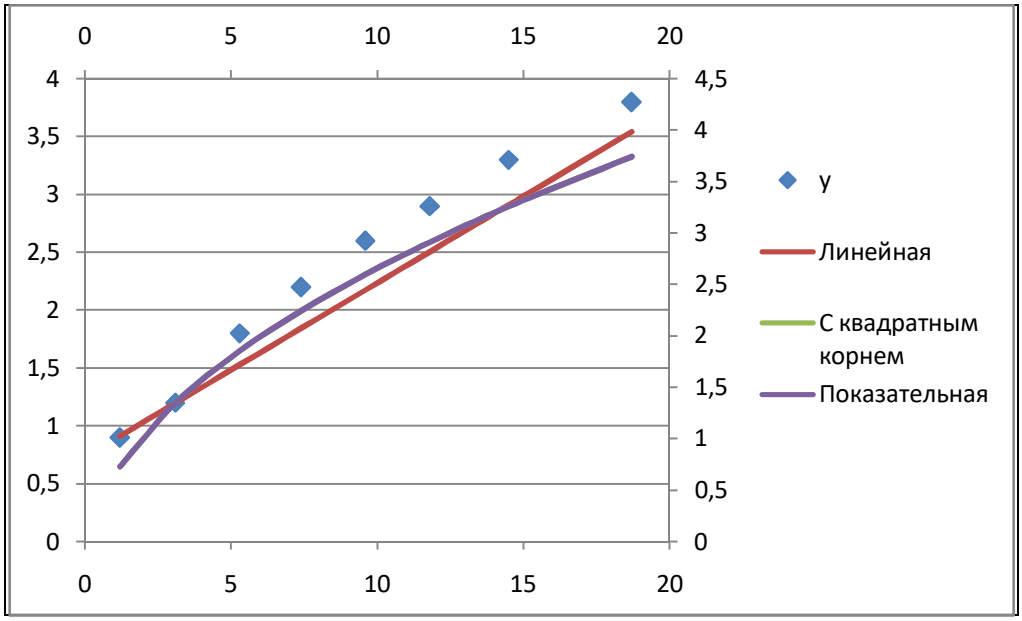


Рис. 2.2. Регрессионные модели

Сравним полученные модели по индексу детерминации и средней ошибке аппроксимации.

Таблица 2.6

Модель	Индекс детерминации	Средняя ошибка аппроксимации, %
Линейная модель $\hat{y} = 0{,}825 + 0{,}169 \cdot x$	0,982	6,38
Модель с квадратным корнем $\hat{y} = -0{,}291 + 0{,}933\sqrt{x}$	0,9912	4,98
Степенная модель $\hat{y} = 0{,}737 \cdot x^{0{,}551}$	0,9958	4,39

Наиболее хорошо исходные данные аппроксимирует степенная модель. Но так как индексы детерминации линейной и степенной модели различаются несущественно (на 0,0138), то можно использовать более простую, линейную модель.

2.3. Задания для выполнения лабораторной работы № 2 «Регрессионно- дисперсионный анализ двумерных моделей»

На основании опытных данных требуется:

1. Упорядочить данные по возрастанию. Построить поле корреляции. Сделать вывод о форме связи между признаками.

2. Вычислить линейный коэффициент парной корреляции и проверить его значимость на уровне $\alpha = 0{,}05$.

3. Найти оценки параметров линейного уравнения регрессии методом наименьших квадратов.

4. Вычислить коэффициент детерминации, проверить его значимость на уровне $\alpha = 0{,}05$.

5. Вычислить среднюю относительную ошибку аппроксимации.

6. Оценить статистическую значимость параметров уравнения регрессии на уровне $\alpha = 0{,}05$. Найти доверительные интервалы для параметров регрессии.

7. Найти прогнозное значение результативного фактора $\hat{y}_p$ при значении признака-фактора, составляющем 110% от среднего уровня $x_p = 1{,}1 \cdot \bar{x}$.

8. В предположении, что связь между признаками имеет указанный вид, найти параметры, вычислить индекс корреляции, коэффициент детерминации, ошибку аппроксимации.

9. В предположении, что связь между признаками имеет указанный вид, найти параметры, вычислить индекс корреляции, коэффициент детерминации, ошибку аппроксимации.

10. На одном графике изобразить поле корреляции, и три графика исследуемых зависимостей. Сделать вывод о наилучшем уравнении регрессии.

Вариант 1

Имеются данные между выработкой продукции x (тыс. руб.) и затратами топлива y в условных единицах:

x	7	10	60	20	45	30	15	120	40	80
y	10,0	9,0	5,4	8,0	5,8	6,3	7,5	5,0	6,0	5,2

Для пункта 8 задания считать, что связь между признаками показательная: $\hat{y} = a \cdot b^x$, для пункта 9 – логарифмическая: $\hat{y} = a + b \cdot \ln x$.

Вариант 2

Имеются данные между дебитом x (т/сут.) скважин и производительностью труда y (т/чел.):

x	14	33	147	98	76	120	55	189	200	68
y	78,6	42,4	23,8	28,6	31,6	25,8	36,4	21,2	20,0	34,2

Вариант 3

Имеются данные максимальной мощности двигателя x (л.с.) и нормы расхода моторного масла на угар и замену y (л /100 л.с.):

x	30	11	15	60	7	45	20	120	52	80
y	6,5	8,5	7,0	5,5	9,0	6,0	8,0	5,0	5,7	5,2

Для пункта 8 задания считать, что связь между признаками параболическая: $\hat{y} = a + bx + cx^2$, для пункта 9 – логарифмическая $\hat{y} = a + b \cdot \ln x$.

Вариант 4

Имеются данные распределения заводов по производственным средствам x (млн руб.) и по суточной выработке y (млн руб.):

x	1,5	3,3	20,0	7,6	15,0	12,0	10,0	19,0	5,5	17,0
y	76,0	44,0	21,0	32,0	24,0	25,8	28,6	21,2	37,0	22,2

Для пункта 8 задания считать, что связь между признаками логарифмическая $\hat{y} = a + b \cdot \ln x$, для пункта 9 – гиперболическая $\hat{y} = \dfrac{x}{ax + b}$.

Вариант 5

Имеются данные энерговооруженности x (кВт/ч) труда и выработки y (тыс. руб.):

x	1500	350	1000	750	550	150	1850	1950	1200	1700
y	36,0	21,0	31,5	26,0	23,6	20,0	42,5	60,0	33,4	38,9

Для пункта 8 задания считать, что связь между признаками гиперболическая $\hat{y} = a + \dfrac{b}{x}$, для пункта 9 – логарифмическая: $\hat{y} = a + b \cdot \ln x$.

Вариант 6

Имеются данные веса детали x (кг) и времени y (с), затрачиваемого на её закрепление:

x	1,7	19,0	5,4	15,0	10,0	7,4	3,4	20,0	12,5	17,0
y	76,0	20,2	37,0	24,0	25,5	31,4	42,5	20,1	24,8	22,2

Для пункта 8 задания считать, что связь между признаками показательная $\hat{y} = a \cdot b^x$, для пункта 9 – гиперболическая $\hat{y} = a + \dfrac{b}{x}$.

Вариант 7

Имеются данные среднегодовой стоимости основных производственных фондов x (млн руб.) и стоимости товарной продукции y (млн руб.):

x	7	10	45	20	30	15	120	80	100	60
y	100	90	58	60	63	75	49	55	50	54

Для пункта 8 задания считать, что связь между признаками экспоненциальная $\hat{y} = a \cdot e^{\frac{b}{x}}$, для пункта 9 – параболическая $\hat{y} = a + bx + cx^2$.

Вариант 8

Имеются данные затрат x (тыс. руб.) на одну тонну нефти и себестоимости y (руб.) одной тонны нефти:

x	1,4	9,8	5,5	7,6	3,3	18,9	14,7	12,0	20,0	22,0
y	1,1	2,8	2,0	2,4	1,4	3,8	3,1	3,0	3,9	4,0

Для пункта 8 задания считать, что связь между признаками логарифмическая $\hat{y} = a + b \ln x$, для пункта 9 – параболическая $\hat{y} = a + bx + cx^2$.

Вариант 9

Имеются данные механической скорости проходки x (м/сут) и количества израсходованных долот y (шт.) при бурении скважин:

x	1,60	1,20	1,32	1,68	1,35	1,41	1,69	1,80	1,90	2,07
y	34,2	19,6	27,3	32,5	31,8	30,7	42,4	42,1	45,0	55,3

Для пункта 8 задания считать, что связь между признаками показательная $\hat{y} = a \cdot b^x$, для пункта 9 – логарифмическая $\hat{y} = a + b \cdot \ln x$.

Вариант 10

Имеются данные температуры x (°С) и растворимости азотно-натриевой соли y (%):

x	22	26	45	37	28	50	56	34	60	40
y	16	17	26	24	22	31	32	18	36	20

Для пункта 8 задания считать, что связь между признаками гиперболическая $\hat{y} = a + \frac{b}{x}$, для пункта 9 – экспоненциальная $\hat{y} = a \cdot e^{bx}$.

§3. Регрессионно- дисперсионный анализ многомерных моделей химико-технологических систем

3.1. Теоретическое обоснование

Парная регрессия дает хороший результат, если влиянием других факторов можно пренебречь. В противном случае, строится уравнение множественной регрессии, когда количество независимых переменных больше одного:

$$y = f(x_1, x_2, \dots x_m) + \varepsilon, \qquad (3.1)$$

где y – зависимая переменная (результативный признак);

x_i – независимые переменные (признаки-факторы), $i = \overline{1, m}$.

Построение уравнения множественной регрессии начинается с решения вопроса о спецификации модели, который включает в себя:

1) проблему отбора факторов;

2) выбор вида уравнения регрессии.

Исследователь должен представлять природу взаимосвязи моделируемого показателя с другими экономическими явлениями.

1. Включаемые во множественную регрессию факторы должны объяснить вариацию независимой переменной. Если строится модель с набором m факторов, то для нее рассчитывается показатель детерминации R^2. При дополнительном включении в регрессию $(m+1)$-го фактора коэффициент детерминации должен возрастать, а остаточная дисперсия уменьшаться: $R^2_{m+1} \geq R^2_m, S^2_{m+1} \leq S^2_m$. Если этого не происходит, то включаемый фактор не улучшает модель и является лишним.

Коэффициенты корреляции между объясняющими переменными позволяют исключать из модели дублирующие факторы. Считается, что две переменные явно коллинеарные, т.е. находятся между собой в линейной зависимости, если $r_{x_i x_j} \geq 0{,}7$. В этом случае рекомендуется исключить один из факторов.

Причем, предпочтение отдается фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами.

Пример. При изучении зависимости $y = \hat{f}(x_1, x_2, x_3)$ матрица парных коэффициентов корреляции оказалась следующей:

Таблица 3.1

	y	x_1	x_2	x_3
y	1	-	-	-
x_1	0,8	1	-	-
x_2	0,7	0,8	1	-
x_3	0,6	0,5	0,2	1

Очевидно, что факторы x_1 и x_2 дублируют друг друга. В анализ целесообразно включить фактор x_2, а не x_1, хотя корреляция x_2 с результатом y слабее, чем корреляция фактора x_1 с y, но зато значительно слабее межфакторная корреляция $r_{x_2x_3} = 0{,}2 < r_{x_1x_3} = 0{,}5$. Поэтому в данном случае в уравнение множественной регрессии включаются факторы x_2, x_3.

2. Возможны разные виды уравнений множественной регрессии: линейные и нелинейные. Линейное уравнение множественной регрессии имеет вид (3.2). Среди нелинейных моделей можно выделить следующие линеаризуемые уравнения:

- экспоненту $y = e^{a+b_1x_1+b_2x_2+\ldots+b_mx_m}$;

- гиперболу $y = \dfrac{1}{a + b_1x_1 + b_2x_2 + \ldots + b_mx_m}$.

Среди видов уравнений множественной регрессии наиболее широко используется линейная функция:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m + \varepsilon\,. \qquad (3.2)$$

Модель (3.2) называется классической нормальной моделью множественной линейной регрессии.

Включение в модель дополнительных факторов усложняет вычисления, поэтому целесообразно перейти к матричному методу с целью облегчения необходимых вычислений. Обозначим:

$B = \begin{pmatrix} b_0 \\ b_1 \\ \dots \\ b_m \end{pmatrix}$ – вектор-столбец параметров уравнения регрессии, где m – число независимых переменных;

$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$ – вектор-столбец результативного признака, где n – число наблюдений;

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$ – вектор-столбец случайных ошибок;

$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}$ – матрица значений объясняющих переменных размера $n \times (m+1)$. Заметим, что в матрицу дополнительно введен столбец, состоящий из единиц.

Тогда в матричной форме модель (3.2) примет вид:

$$\hat{Y} = B \cdot X + \varepsilon . \qquad (3.3)$$

Для оценки параметров линейного уравнения множественной регрессии применим метод наименьших квадратов.

По методу наименьших квадратов получаем систему линейных уравнений для нахождения параметров линейного уравнения множественной регрессии (3.2), которая в матричной форме выглядит следующим образом:

$$X^T \cdot X \cdot B = X^T \cdot Y \tag{3.4}$$

или

$$\begin{cases} na + b_1 \sum x_1 + b_2 \sum x_2 + \ldots + b_m \sum x_m = \sum y, \\ a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \ldots + b_m \sum x_1 x_m = \sum y x_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a \sum x_m + b_1 \sum x_1 x_m + b_2 \sum x_2 x_m + \ldots + b_m \sum x_m^2 = \sum y x_m . \end{cases} \tag{3.5}$$

В частности, для двухфакторной модели система будет иметь вид:

$$\begin{cases} na + b_1 \sum x_1 + b_2 \sum x_2 = \sum y, \\ a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 = \sum y x_1, \\ a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 = \sum y x_2 . \end{cases} \tag{3.6}$$

Решением матричного уравнения (3.4) является вектор:

$$B = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot Y \tag{3.7}$$

Запишем матрицы, входящие в уравнение (3.4). Имеем:

$$X^T \cdot X = \begin{pmatrix} n & \sum x_1 & \ldots & \sum x_m \\ \sum x_1 & \sum x_1^2 & \ldots & \sum x_1 x_m \\ \ldots & \ldots & \ldots & \ldots \\ \sum x_m & \sum x_1 x_m & \ldots & \sum x_m^2 \end{pmatrix}, \; X^T \cdot Y = \begin{pmatrix} \sum y \\ \sum y x_1 \\ \ldots \\ \sum y x_m \end{pmatrix} \tag{3.8}$$

В частности, для двухфакторной модели данные матрицы будут иметь вид:

$$X^T \cdot X = \begin{pmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 \end{pmatrix}, \; X^T \cdot Y = \begin{pmatrix} \sum y \\ \sum y x_1 \\ \sum y x_2 \end{pmatrix} \tag{3.9}$$

Средние по совокупности коэффициенты эластичности:

$$\overline{Э}_i = b_i \cdot \frac{\overline{x}_i}{\overline{y}} . \tag{3.10}$$

Они показывают, на сколько процентов в среднем изменится результат при изменении соответствующего фактора на 1%. Средние показатели эластичности можно сравнивать друг с другом и соответственно ранжировать факторы по силе их воздействия на результат.

Практическая значимость уравнения множественной регрессии оценивается с помощью показателя множественной корреляции и его квадрата – коэффициента детерминации. Показатель множественной корреляции характеризует тесноту связи рассматриваемого набора факторов с исследуемым признаком, то есть характеризует совместное влияние всех факторов на результат. *Коэффициент детерминации* вычисляется по формуле:

$$R^2_{yx_1\ldots x_m} = 1 - \frac{\sigma^2_{ост}}{\sigma^2_y} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad (3.11)$$

где $\sigma^2_{ост}$ – остаточная дисперсия;

σ^2_y – дисперсия результативного признака.

Свойства индекса множественной корреляции:

1) $0 \le R^2_{yx_1\ldots x_m} \le 1$;
2) чем ближе $R^2_{yx_1\ldots x_m}$ к единице, тем теснее связь результативного признака со всем набором исследуемых факторов;
3) величина индекса множественной корреляции должна быть больше или равна максимальному значению индекса парной корреляции: $R^2_{yx_1\ldots x_m} \ge \max_i r_{yx_i}$.

Значимость множественного уравнения регрессии оценивается с помощью F-критерия Фишера:

$$F_{набл} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}, \qquad (3.12)$$

где n – число наблюдений;

m – число параметров при переменных x_i.

Если $F_{набл} > F_{кр} = F(\alpha = 0{,}05; k_1 = m; k_2 = n - m - 1)$, то уравнение регрессии статистически значимо.

В дополнение к оценке качества множественного уравнения регрессии вычисляют среднюю относительную ошибку аппроксимации по формуле (2.12).

Для оценки значимости параметров множественной регрессии выдвигаем гипотезы:

$H_0:\ b_i = 0$, то есть коэффициент b_i незначим;

$H_1:\ b_i \neq 0$, то есть коэффициент b_i значим.

Для проверки гипотезы используется t-распределение Стьюдента. Наблюдаемое значение критерия вычисляется по формуле:

$$t_{b_i} = \frac{b_i}{s_{b_i}}, \tag{3.13}$$

где

$$s_{b_i} = s\sqrt{\left[\left(X^T X\right)^{-1}\right]_{ii}}, \tag{3.14}$$

$$s = \sqrt{\frac{\sum\left(\hat{y}_i - y_i\right)^2}{n-m-1}}. \tag{3.15}$$

По таблице распределения Стьюдента находим критическое значение критерия $t_{кр} = t(\alpha = 0{,}05; k = n-m-1)$. Если $t_{b_i} > t_{кр}$, то гипотеза H_0 отвергается и принимается гипотеза H_1 о значимости коэффициента b_i. Истинное значение параметра b_i находится в интервале

$$\left(b_i - t_{кр}\cdot s_{b_i}; b_i + t_{кр}\cdot s_{b_i}\right).$$

Наряду с интервальным оцениванием коэффициентов регрессии весьма важным для оценки точности определения зависимой переменной (прогноза) является построение доверительного интервала для функции регрессии (условного математического ожидания зависимой переменной $M_x(Y)$), найденного в предположении, что объясняющие переменные $x_1, x_2, \ldots x_m$ приняли значения,

задаваемые вектором $X_p = \begin{pmatrix} 1 \\ x_{1p} \\ x_{2p} \\ \dots \\ x_{mp} \end{pmatrix}$. Тогда доверительный интервал для $M_x(Y)$ следующий:

$$\hat{y}_p - t \cdot s_{\hat{y}_p} \le M_x(Y) \le \hat{y}_p + t \cdot s_{\hat{y}_p}, \qquad (3.16)$$

где $s_{\hat{y}_p} = s\sqrt{X_p^T \cdot \left(X^T \cdot X\right)^{-1} \cdot X_p}$ – стандартная ошибка групповой средней.

Доверительный интервал для индивидуальных значений зависимой переменной:

$$\hat{y}_p - m_{\hat{y}_p} \cdot t_{кр} \le y_p \le \hat{y}_p + m_{\hat{y}_p} \cdot t_{кр}, \qquad (3.17)$$

где $m_{\hat{y}_p} = s\sqrt{1 + X_p^T \cdot \left(X^T \cdot X\right)^{-1} \cdot X_p}$

При практическом проведении регрессионного анализа при помощи метода МНК следует обратить серьезное внимание на проблемы, связанные с выполнимостью свойств случайных отклонений моделей. Свойства оценок коэффициентов регрессии напрямую зависят от свойства случайного члена в уравнении регрессии. Для получения качественных оценок необходимо следить за выполняемостью предпосылок МНК, так как при их нарушении МНК может давать оценки с плохими статистическими свойствами. Одной из ключевых предпосылок МНК является условие постоянства дисперсий случайных отклонений. Выполнимость данной предпосылки называется *гомоскедастичностью* (постоянством дисперсии отклонений), невыполнимость данной предпосылки называется *гетероскедастичностью* (непостоянством дисперсий отклонений).

Случайные отклонения принимают произвольные значения некоторых вероятностных распределений. Но, несмотря на то, что при каждом конкретном

наблюдении случайное отклонение может быть большим либо меньшим, положительным либо отрицательным, не должно быть причины, вызывающей большие отклонения при одних наблюдениях и меньшие при других.

На рис. 3. 1 приведены два примера линейной регрессии – зависимости потребления C от дохода I: $C = \beta_0 + \beta_1 I + \varepsilon$.

В обоих случаях с ростом дохода растет среднее значение потребления. Но на рис. 3.1,а дисперсия остается одной и той же для различных уровней дохода, а на рис. 3.1,б дисперсия потребления не остается постоянной, а увеличивается с ростом дохода. Фактически это означает, что во втором случае субъекты с большим доходом в среднем потребляют больше, чем субъекты с меньшим доходом, и, кроме того, разброс в их потреблении более существенен для большего уровня дохода. Люди с большим доходом имеют больший простор для его распределения. Реалистичность данной ситуации не вызывает сомнений.

а) б)

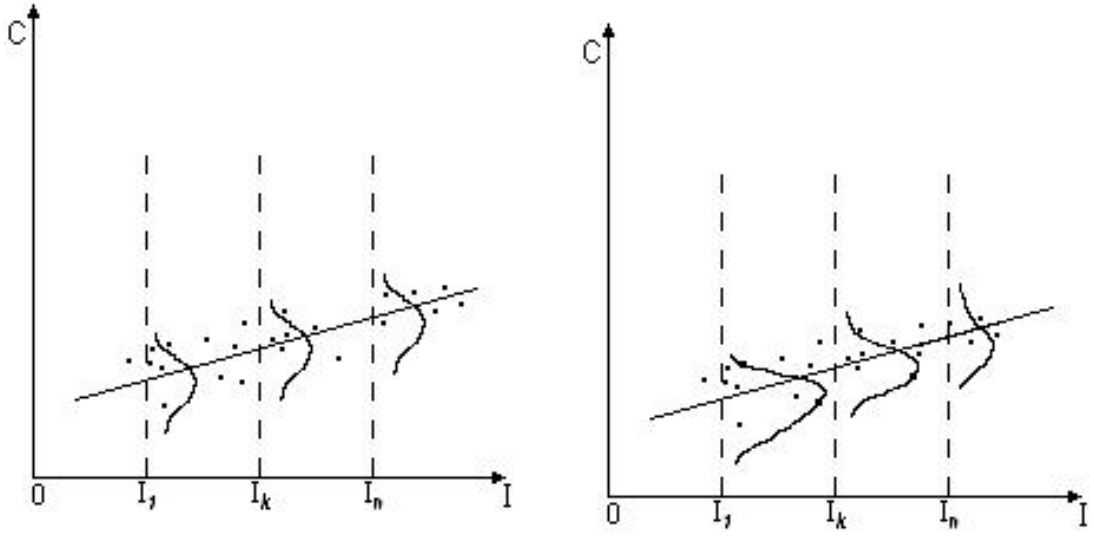


Рис. 3.1. Линейные регрессии зависимости потребления от дохода: а – дисперсия постоянна, б – дисперсия увеличивается

Гетероскедастичность приводит к тому, что выводы, полученные на основе *t*- и *F*-статистик, а также интервальные оценки будут ненадежными.

Обнаружение гетероскедастичности является довольно сложной задачей. В настоящее время существует ряд методов, позволяющих определить наличие гетероскедастичности. Наиболее часто употребляемым является тест ранговой корреляции Спирмена.

Выдвигаются гипотезы

$H_0:\ \sigma_1^2 = \ldots = \sigma_n^2$;

H_1: ошибка коррелирует с одним из факторов.

Для проверки гипотезы значения x_i и $|e_i| = |y_i - \hat{y}_i|$ ранжируются (упорядочиваются по величинам). Затем определяется коэффициент ранговой корреляции:

$$r_{x,e} = 1 - 6\frac{\sum d_i^2}{n(n^2-1)}, \tag{3.18}$$

где d_i – разность между рангами x_i и e_i, i = 1,2,…,n; n – число наблюдений.

Например, если x_{20} является 25-м по величине среди всех наблюдений, а e_{20} является 32-м, то d_{20} = 25 – 32 = -7.

Затем рассчитывается t-статистика:

$$t_{набл} = \frac{r_{x,e}\sqrt{(n-2)}}{\sqrt{1-r_{x,e}^2}}. \tag{3.19}$$

Если значение, рассчитанное по формуле (3.19), превышает критическое $t_{кр} = t(\alpha = 0{,}05; k = n - m - 1)$ (t-распределение), то необходимо отклонить гипотезу H_0 об отсутствии гетероскедастичности. В противном случае гипотеза об отсутствии гетероскедастичности принимается.

Если в модели регрессии больше, чем одна объясняющая переменная, то проверка гипотезы может осуществляться с помощью t-статистики для каждой из них отдельно.

3.2. Пример выполнения лабораторной работы
№ 3 «Регрессионно- дисперсионный анализ многомерных моделей»

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (процент от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%).

Таблица 3.2

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7,0	3,9	14,0	7	10,0	6,8	20,0
2	7,0	4,0	16,0	8	11,0	7,2	25,0
3	7,0	4,8	19,0	9	12,0	8,0	28,0
4	8,0	5,4	19,0	10	12,0	8,2	29,0
5	8,0	5,3	20,0	11	14,0	9,6	32,0
6	9,0	6,0	21,0	12	14,0	9,0	36,0

Требуется:

1. Предполагая, что между переменными y , x_1, x_2 существует линейная корреляционная зависимость, найти её аналитическое выражение, то есть найти уравнение y по x_1, x_2.

2. Вычислить коэффициенты эластичности и сделать вывод о влиянии факторных признаков на результативный признак.

3. Определить множественный коэффициент детерминации и проверить значимость полученного уравнения регрессии на уровне значимости $\alpha = 0{,}05$.

4. Оценить значимость коэффициентов регрессии и построить для них 95%-е доверительные интервалы.

5. Найти прогнозное значение результативного фактора $\hat{y}_p$ при значении признака-фактора, составляющем 110% от среднего уровня $x_p = 1{,}1 \cdot \bar{x}$.

6. Сделать вывод о возможном наличии или отсутствии гетероскедастичности в модели.

Решение

1. Уравнение регрессии ищем в виде $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$.

Обозначим $Y = \begin{pmatrix} 7{,}0 \\ 7{,}0 \\ 7{,}0 \\ \dots \\ 14{,}0 \end{pmatrix}$, $X = \begin{pmatrix} 1 & 3{,}9 & 14{,}0 \\ 1 & 4{,}0 & 16{,}0 \\ 1 & 4{,}8 & 19{,}0 \\ \dots & \dots & \dots \\ 1 & 9{,}0 & 36{,}0 \end{pmatrix}$.

Для удобства вычислений составляем таблицу:

Таблица 3.3

№ п/п	x_1	x_2	y	x_1^2	x_2^2	y^2	$x_1 \cdot x_2$	$x_1 \cdot y$	$x_2 \cdot y$
1	3,9	14,0	7	15,21	196,0	49,00	54,60	27,30	98,00
2	4	16	7	16,00	256	49	64,0	28,0	112
3	4,8	19	7	23,04	361	49	91,2	33,6	133
4	5,4	19	8	29,16	361	64	102,6	43,2	152
5	5,3	20	8	28,09	400	64	106,0	42,4	160
6	6	21	9	36,00	441	81	126,0	54,0	189
7	6,8	20	10	46,24	400	100	136,0	68,0	200
8	7,2	25	11	51,84	625	121	180,0	79,2	275
9	8	28	12	64,00	784	144	224,0	96,0	336
10	8,2	29	12	67,24	841	144	237,8	98,4	348
11	9,6	32	14	92,16	1024	196	307,2	134,4	448
12	9	36	14	81,00	1296	196	324,0	126,0	504
Сумма	**78,2**	**279**	**119**	**549,98**	**6985**	**1257**	**1953,4**	**830,5**	**2955**
Среднее	**6,517**	**23,25**	**9,917**	**45,832**	**582,083**	**104,75**	**162,783**	**69,208**	**246,25**

Используя выражения (3.9) и значения табл. 3.3, составляем матрицы:

$$X^T \cdot X = \begin{pmatrix} 12 & 78{,}2 & 279 \\ 78{,}2 & 549{,}98 & 1953{,}4 \\ 279 & 1953{,}4 & 6985 \end{pmatrix}, \quad X^T \cdot Y = \begin{pmatrix} 119 \\ 830{,}5 \\ 2955 \end{pmatrix}.$$

Вычислим матрицу $A^{-1} = \left(X^T \cdot X\right)^{-1}$ по формуле

$$A^{-1} = \frac{1}{\det A} \begin{pmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \\ A_{13} & A_{23} & A_{33} \end{pmatrix},$$

где $\det A$ – определитель матрицы A;

A_{ij} – алгебраическое дополнение элемента a_{ij} матрицы A (определитель, получаемый из определителя матрицы A вычеркиванием i-й строки и j-го столбца и умноженный на (-1) в случае, если $(i+j)$ – нечетное число).

Имеем:

$$\det A = \begin{vmatrix} 12 & 78{,}2 & 279 \\ 78{,}2 & 549{,}98 & 1953{,}4 \\ 279 & 1953{,}4 & 6985 \end{vmatrix} = 12 \cdot 549{,}98 \cdot 6985 + 78{,}2 \cdot 1953{,}4 \cdot 279 +$$

$$+ 279 \cdot 78{,}2 \cdot 1953{,}4 - 279 \cdot 549{,}98 \cdot 279 - 12 \cdot 1953{,}4 \cdot 1953{,}4 - 78{,}2 \cdot 78{,}2 \cdot 6985 =$$

21 901,34.

$$A_{11} = \begin{vmatrix} 549{,}98 & 1953{,}4 \\ 1953{,}4 & 6985 \end{vmatrix} = 25838{,}74;\ A_{12} = -\begin{vmatrix} 78{,}2 & 1953{,}4 \\ 279 & 6985 \end{vmatrix} = -1228{,}4;$$

$$A_{13} = \begin{vmatrix} 78{,}2 & 549{,}98 \\ 279 & 1953{,}4 \end{vmatrix} = -688{,}54;\ A_{21} = -\begin{vmatrix} 78{,}2 & 279 \\ 1953{,}4 & 6985 \end{vmatrix} = -1228{,}4;$$

$$A_{22} = \begin{vmatrix} 12 & 279 \\ 279 & 6985 \end{vmatrix} = 5979;\ A_{23} = -\begin{vmatrix} 12 & 78{,}2 \\ 279 & 1953{,}4 \end{vmatrix} = -1623;$$

$$A_{31} = \begin{vmatrix} 78{,}2 & 279 \\ 549{,}98 & 1953{,}4 \end{vmatrix} = -688{,}54;\ A_{32} = -\begin{vmatrix} 12 & 279 \\ 78{,}2 & 1953{,}4 \end{vmatrix} = -1623;$$

$$A_{33} = \begin{vmatrix} 12 & 78{,}2 \\ 78{,}2 & 549{,}98 \end{vmatrix} = 484{,}52.$$

Обратная матрица:

$$A^{-1} = \frac{1}{21901{,}34} \begin{pmatrix} 25838{,}74 & -1228{,}4 & -688{,}54 \\ -1228{,}4 & 5979 & -1623 \\ -688{,}54 & -1623 & 484{,}52 \end{pmatrix}.$$

По формуле (3.7) вычисляем вектор B:

$$B = \frac{1}{21901{,}34} \begin{pmatrix} 25838{,}74 & -1228{,}4 & -688{,}54 \\ -1228{,}4 & 5979 & -1623 \\ -688{,}54 & -1623 & 484{,}52 \end{pmatrix} \cdot \begin{pmatrix} 119 \\ 830{,}5 \\ 2955 \end{pmatrix} =$$

$$= \frac{1}{21901{,}34} \begin{pmatrix} 25838{,}74 \cdot 119 + (-1228{,}4) \cdot 830{,}5 + (-688{,}54) \cdot 2955 \\ (-1228{,}4) \cdot 119 + 5979 \cdot 830{,}5 + (-1623) \cdot 2955 \\ (-688{,}54) \cdot 119 + (-1623) \cdot 830{,}5 + 484{,}52 \cdot 2955 \end{pmatrix} = = \begin{pmatrix} 0{,}913 \\ 1{,}069 \\ 0{,}088 \end{pmatrix}.$$

Таким образом, $b_0 = 0{,}913$; $b_1 = 1{,}069$; $b_2 = 0{,}088$. Значит, уравнение регрессии имеет вид:

$$\boxed{\hat{y} = 0{,}913 + 1{,}069x_1 + 0{,}088x_2.}$$

2. Используя данные табл. 3.3 и найденные значения коэффициентов $b_i, i = 1, 2$, вычислим коэффициенты эластичности по формуле (3.10):

$$\overline{Э}_1 = b_1 \cdot \frac{\bar{x}_1}{\bar{y}} = 1{,}069 \cdot \frac{6{,}517}{9{,}917} = 0{,}702; \quad \overline{Э}_2 = b_2 \cdot \frac{\bar{x}_2}{\bar{y}} = 0{,}088 \cdot \frac{23{,}25}{9{,}917} = 0{,}206.$$

При увеличении количества вводимых в действие новых основных фондов на 1% от своего среднего значения выработка продукции на одного рабочего увеличивается в среднем на 0,702 %. При увеличении удельного веса рабочих высокой квалификации в общей численности рабочих на 1 % выработка продукции на одного рабочего увеличивается в среднем на 0,206 %. Значит, введение в действие новых основных фондов оказывает более существенное влияние на выработку продукции, чем увеличение удельного веса рабочих высокой квалификации в общей численности рабочих.

3. Определим множественный коэффициент детерминации и проверим значимость полученного уравнения регрессии y по x_1 и x_2 на уровне значимости $\alpha = 0{,}05$. Все вычисления запишем в табл. 3.4.

Таблица 3.4

№ п/п	x_1	x_2	y	$\hat{y}$	$\left\|\frac{y-\hat{y}}{y}\right\|$	$(y-\hat{y})^2$	$(y-\bar{y})^2$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	3,9	14,0	7	6,3141	0,0980	0,4705	8,5089
2	4	16	7	6,5970	0,0576	0,1624	8,5089
3	4,8	19	7	7,7162	0,1023	0,5129	8,5089
4	5,4	19	8	8,3576	0,0447	0,1279	3,6749
5	5,3	20	8	8,3387	0,0423	0,1147	3,6749
6	6	21	9	9,1750	0,0194	0,0306	0,8409
7	6,8	20	10	9,9422	0,0058	0,0033	0,0069

Окончание табл. 3.4

1	2	3	4	5	6	7	8
8	7,2	25	11	10,8098	0,0173	0,0362	1,1729
9	8	28	12	11,9290	0,0059	0,0050	4,3389
10	8,2	29	12	12,2308	0,0192	0,0533	4,3389
11	9,6	32	14	13,9914	0,0006	0,0001	16,6709
12	9	36	14	13,7020	0,0213	0,0888	16,6709
Сумма	**78,2**	**279**	**119**	**119,1038**	**0,4345**	**1,6057**	**76,9167**

где $\hat{y} = 0{,}913 + 1{,}069x_1 + 0{,}088x_2$.

Тогда $R^2 = 1 - \frac{1{,}6057}{76{,}9167} = 0{,}9791$, то есть 97б91% вариации зависимой переменной y (выработки продукции на одного работника, тыс. руб.) объясняется вариацией независимых переменных x_1 (количества введенных в действие новых основных фондов, % от стоимости на конец года) и x_2 (удельного веса рабочих высокой квалификации в общей численности рабочих, %).

Проверим значимость коэффициента детерминации $F_{набл} = \frac{0{,}9791}{1-0{,}9791} \cdot \frac{12-2-1}{2} = 210{,}81$. По таблице распределения Фишера-Снедекора (приложение Д) находим $F_{кр} = F(\alpha = 0{,}05; k_1 = 2; k_2 = 9) = 4{,}26$. Так как $F_{набл} > F_{кр}$, то принимается гипотеза о значимости коэффициента детерминации и уравнения регрессии в целом.

По формуле (2.12), используя данные табл. 3.4, найдем среднюю относительную ошибку аппроксимации:

$$A = \frac{100\%}{12} \cdot 0{,}4345 = 3{,}62\%.$$

Значение средней относительной ошибки аппроксимации менее 10%, что говорит о хорошем подборе модели к исходным данным.

Оценим значимость коэффициентов регрессии и построим для них 95%-е доверительные интервалы. По формуле (3.15), используя данные табл. 3.4, име-

ем $s=\sqrt{\frac{1{,}6057}{9}}=0{,}4295$. Матрица

$$\left(X^T\cdot X\right)^{-1}=A^{-1}=\frac{1}{21901{,}34}\begin{pmatrix}25838{,}74 & -1228{,}4 & -688{,}54\\ -1228{,}4 & 5979 & -1623\\ -688{,}54 & -1623 & 484{,}52\end{pmatrix}.$$

Тогда $s_{b_1}=0{,}4295\cdot\sqrt{\frac{5979}{21901{,}34}}=0{,}4295\cdot 0{,}522=0{,}224$. Так как $t_{b_1}=\frac{1{,}069}{0{,}224}=4{,}77>t_{кр}\left(\alpha=0{,}05;k=9\right)=2{,}26$, то коэффициент b_1 значим. Аналогично вычисляем $s_{b_2}=0{,}4295\cdot\sqrt{\frac{484{,}52}{21901{,}34}}=0{,}4295\cdot 0{,}149=0{,}064$ и $t_{b_1}=\frac{0{,}088}{0{,}064}=1{,}375<t_{кр}\left(\alpha=0{,}05;k=9\right)=2{,}26$, то есть коэффициент b_2 незначим с вероятностью 0,95.

Доверительный интервал имеет смысл построить только для коэффициента b_1: $\left(1{,}069-2{,}26\cdot 0{,}224;\ 1{,}069-2{,}26\cdot 0{,}224\right)=\left(0{,}563;1{,}575\right)$, то есть истинное значение параметра b_1 заключается в данном числовом интервале.

Оценим выработку продукции на одного работника, если введено в действие 5,5% новых основных фондов от стоимости фондов на конец года, а удельный вес рабочих высокой квалификации составил 27% в общей численности рабочих. Найти 95%-е доверительные интервалы для индивидуального и среднего значений выработки продукции на одного рабочего.

Напомним, что уравнение регрессии получено в виде $\hat{y}=0{,}913+1{,}069x_1+0{,}088x_2$.

По условию надо оценить $M_x\left(Y\right)$, где $X_p=\begin{pmatrix}1\\5{,}5\\27\end{pmatrix}$. Выборочной оценкой $M_x\left(Y\right)$ является групповая средняя, которую найдем по уравнению регрессии $\hat{y}=0{,}913+1{,}069\cdot 5{,}5+0{,}088\cdot 27=9{,}169$ тыс. руб.

Для построения доверительного интервала для $M_x(Y)$ необходимо знать дисперсию его оценки – $S^2_{\hat{y}_p}$. Выше было вычислено $s = \sqrt{\frac{1{,}6057}{9}} = 0{,}4295$. Найдем теперь

$$X_p^T \cdot \left(X^T \cdot X\right)^{-1} \cdot X_p = \begin{pmatrix} 1 & 5{,}5 & 27 \end{pmatrix} \cdot \frac{1}{21901{,}34} \begin{pmatrix} 25838{,}74 & -1228{,}4 & -688{,}54 \\ -1228{,}4 & 5979 & -1623 \\ -688{,}54 & -1623 & 484{,}52 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 5{,}5 \\ 27 \end{pmatrix} =$$

$$= \frac{1}{21901{,}34} \begin{pmatrix} 491{,}96 & -12164{,}9 & 3467 \end{pmatrix} \begin{pmatrix} 1 \\ 5{,}5 \\ 27 \end{pmatrix} = 1{,}2417 .$$

Тогда $s_{\hat{y}_p} = 0{,}4295\sqrt{1{,}2417} = 0{,}479$ тыс. руб. По таблице распределения Стьюдента (приложение Б) находим $t_{кр}(\alpha = 0{,}05; k = 9) = 2{,}26$. Доверительный интервал по формуле (3.16):

$$9{,}169 - 2{,}26 \cdot 0{,}479 \le M_x(Y) \le 9{,}169 + 2{,}26 \cdot 0{,}479 ,$$

$$8{,}086 \le M_x(Y) \le 10{,}252 .$$

Итак, с надежностью 0,95 средняя выработка продукции на одного работника при заданных значениях объясняющих переменных находится в пределах от 8,086 тыс. руб. до 10,252 тыс. руб.

Найдем доверительный интервал для индивидуального значения y_p при $X_p = \begin{pmatrix} 1 \\ 5{,}5 \\ 27 \end{pmatrix}$:

$$m_{\hat{y}_p} = 0{,}4295\sqrt{1 + 1{,}2417} = 0{,}6431;$$

и по (3.17) $9{,}169 - 0{,}6431 \cdot 2{,}26 \le y_p \le 9{,}169 + 0{,}6431 \cdot 2{,}26$,

$$7{,}716 \le y_p \le 10{,}622 \text{ тыс. руб.}$$

С надежностью 0,95 индивидуальное значение выработки продукции на одного рабочего находится в пределах от 7,716 до 10,622 тыс. руб.

Проверим наличие гетерескедастичности для переменной x_1. Необходимые вычисления заносим в табл. 3.5.

Таблица 3.5

№ п/п	x_1	Ранг x_1	y	$\hat{y}$	$\lvert e_i \rvert$	Ранг $\lvert e_i \rvert$	d_i	d_i^2
1	3,9	1	7	6,3141	0,6859	11	10	100
2	4	2	7	6,597	0,403	10	8	64
3	4,8	3	7	7,7162	0,7162	12	9	81
4	5,4	5	8	8,3576	0,3576	9	4	16
5	5,3	4	8	8,3387	0,3387	8	4	16
6	6	6	9	9,175	0,175	4	-2	4
7	6,8	7	10	9,9422	0,0578	2	-5	25
8	7,2	8	11	10,8098	0,1902	5	-3	9
9	8	9	12	11,929	0,071	3	-6	36
10	8,2	10	12	12,2308	0,2308	6	-4	16
11	9,6	12	14	13,9914	0,0086	1	-11	121
12	9	11	14	13,702	0,298	7	-4	16
Сумма								**504**

По формуле (3.18) коэффициент ранговой корреляции $r_{x_1,e} = 1 - 6 \cdot \frac{504}{12\left(12^2 - 1\right)} = -0{,}762$. По формуле (3.19) вычисляем наблюдаемое значение критерия Стьюдента $t_{набл} = \frac{-0{,}762\sqrt{12-2}}{\sqrt{1-(-0{,}762)^2}} = -3{,}72$. По таблице распределения Стьюдента (приложение Б) находим критическое значение критерия $t_{кр} = t(\alpha = 0{,}05; k = 9) = 2{,}26$. Так как $\lvert t_{набл} \rvert > t_{кр}$, то принимается гипотеза H_1 о наличии гетероскедастичности остатков переменной x_1.

Аналогичные рассуждения проводим для переменной x_2. Вычисления заносим в табл. 3.6.

Таблица 3.6

№	x_1	Ранг x_1	y	$\hat{y}$	$\lvert e_i \rvert$	Ранг $\lvert e_i \rvert$	d_i	d_i^2
1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
1	14	1	7	6,3141	0,6859	11	10	100
2	16	2	7	6,597	0,403	10	8	64
3	19	3	7	7,7162	0,7162	12	9	81
4	19	3	8	8,3576	0,3576	9	6	36
5	20	5	8	8,3387	0,3387	8	3	9
6	21	7	9	9,175	0,175	4	-3	9
7	20	5	10	9,9422	0,0578	2	-3	9

Окончание табл. 3.6

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
8	25	8	11	10,8098	0,1902	5	-3	9
9	28	9	12	11,929	0,071	3	-6	36
10	29	10	12	12,2308	0,2308	6	-4	16
11	32	11	14	13,9914	0,0086	1	-10	100
12	36	12	14	13,702	0,298	7	-5	25
Сумма								**494**

По формуле (3.18) коэффициент ранговой корреляции $r_{x_1,e} = 1 - 6 \cdot \frac{494}{12\left(12^2 - 1\right)} = -0{,}727$. По формуле (3.19) вычисляем наблюдаемое значение критерия Стьюдента $t_{набл} = \frac{-0{,}727\sqrt{12-2}}{\sqrt{1-(-0{,}727)^2}} = -3{,}35$. По таблице распределения Стьюдента (приложение Б) находим критическое значение критерия $t_{кр} = t(\alpha = 0{,}05; k = 9) = 2{,}26$. Так как $|t_{набл}| > t_{кр}$, то принимается гипотеза H_1 о наличии гетероскедастичности остатков переменной x_2.

3.3. Задания для выполнения лабораторной работы № 3 «Регрессионно- дисперсионный анализ многомерных моделей»

Требуется:

1. Предполагая, что между переменными y , x_1, x_2 существует линейная корреляционная зависимость, найти её аналитическое выражение, то есть найти уравнение y по x_1, x_2.

2. Вычислить коэффициенты эластичности и сделать вывод о влиянии факторных признаков на результативный признак.

3. Определить множественный коэффициент детерминации и проверить значимость полученного уравнения регрессии на уровне значимости $\alpha = 0{,}05$.

4. Оценить значимость коэффициентов регрессии и построить для них 95%-е доверительные интервалы.

5. Найти прогнозное значение результативного фактора $\hat{y}_p$ при значении признака-фактора, составляющем 110% от среднего уровня $x_p = 1{,}1 \cdot \bar{x}$.

6. Сделать вывод о возможном наличии или отсутствии гетероскедастичности в модели.

Вариант 1

Изучается линейная (в среднем) зависимость результативного признака y (цены товара, руб.) от двух факторных признаков (x_1 – дальности перевозки, км и x_2 – расходов на рекламу в месяц, тыс. руб.) по числовым данным, собранным на $n = 7$ однотипных фирмах.

y	48,72	53,01	51,39	73,71	67,16	69,27	42,09
x_1	12	19	17	27	21	22	10
x_2	10	14	10	11	6	7	12

Вариант 2

Изучается линейная (в среднем) зависимость результативного признака y (цены товара, руб.) от двух факторных признаков (x_1 – дальности перевозки, км и x_2 – расходов на рекламу в месяц, тыс. руб.) по числовым данным, собранным на $n = 7$ однотипных фирмах.

y	37,52	35,91	35,69	49,11	47,46	48,77	32,89
x_1	11	18	16	26	20	21	9
x_2	11	15	11	12	7	8	13

Вариант 3

Изучается линейная (в среднем) зависимость результативного признака y (цены товара, руб.) от двух факторных признаков (x_1 – дальности перевозки, км и x_2 – расходов на рекламу в месяц, тыс. руб.) по числовым данным, собранным на $n = 7$ однотипных фирмах.

y	48,72	53,01	51,39	73,71	67,16	69,27	42,09
x_1	12	19	17	27	21	22	10
x_2	10	14	10	11	6	7	12

Вариант 4

Изучается линейная (в среднем) зависимость результативного признака y (цены товара, руб.) от двух факторных признаков (x_1 – дальности перевозки, км и x_2 – расходов на рекламу в месяц, тыс. руб.) по числовым данным, собранным на $n = 7$ однотипных фирмах.

y	50,12	51,91	51,29	70,51	66,26	67,97	43,89
x_1	12	19	17	27	21	22	10
x_2	10	14	10	11	6	7	12

Вариант 5

Изучается линейная (в среднем) зависимость результативного признака y (цены товара, руб.) от двух факторных признаков (x_1 – дальности перевозки, км и x_2 – расходов на рекламу в месяц, тыс. руб.) по числовым данным, собранным на $n = 7$ однотипных фирмах.

y	36,22	38,01	37,39	56,61	52,36	54,07	29,99
x_1	11	18	16	26	20	21	9
x_2	12	16	12	13	8	9	14

Вариант 6

Изучается линейная (в среднем) зависимость результативного признака y (цены товара, руб.) от двух факторных признаков (x_1 – дальности перевозки, км и x_2 – расходов на рекламу в месяц, тыс. руб.) по числовым данным, собранным на $n = 7$ однотипных фирмах.

y	48,72	53,01	51,39	73,71	67,16	69,27	42,09
x_1	12	19	17	27	21	22	10
x_2	10	14	10	11	6	7	12

Вариант 7

Изучается линейная (в среднем) зависимость результативного признака y (цены товара, руб.) от двух факторных признаков (x_1 – дальности перевозки, км

и x_2 – расходов на рекламу в месяц, тыс. руб.) по числовым данным, собранным на $n = 7$ однотипных фирмах.

y	45,92	48,81	47,59	67,91	62,56	64,47	39,69
x_1	13	20	18	28	22	23	11
x_2	10	14	10	11	6	7	12

Вариант 8

Изучается линейная (в среднем) зависимость результативного признака y (цены товара, руб.) от двух факторных признаков (x_1 – дальности перевозки, км и x_2 – расходов на рекламу в месяц, тыс. руб.) по числовым данным, собранным на $n = 7$ однотипных фирмах.

y	47,22	50,41	49,39	70,61	65,16	67,07	40,59
x_1	13	20	18	28	22	23	11
x_2	10	14	10	11	6	7	12

Вариант 9

Изучается линейная (в среднем) зависимость результативного признака y (цены товара, руб.) от двух факторных признаков (x_1 – дальности перевозки, км и x_2 – расходов на рекламу в месяц, тыс. руб.) по числовым данным, собранным на $n = 7$ однотипных фирмах.

y	45,92	48,81	47,59	67,91	62,56	64,47	39,69
x_1	13	20	18	28	22	23	11
x_2	10	14	10	11	6	7	12

Вариант 10

Изучается линейная (в среднем) зависимость результативного признака y (цены товара, руб.) от двух факторных признаков (x_1 – дальности перевозки, км и x_2 – расходов на рекламу в месяц, тыс. руб.) по числовым данным, собранным на $n = 7$ однотипных фирмах.

y	47,62	49,41	48,79	68,01	63,76	65,47	41,39
x_1	13	20	18	28	22	23	11
x_2	10	14	10	11	6	7	12

§4. Применение MS Excel в регрессионном анализе многомерных моделей химико-технологических систем

Рассмотрим пример реализации множественного линейного регрессионного анализа в Excel.

Имеется линейная в среднем зависимость результативного признака Y – ожидаемой продолжительности жизни мужчины (в годах) от пяти факторных признаков: $x^{(1)}$ – среднего числа детей в семье, чел.; $x^{(2)}$ – ВВП на душу населения, долл.; $x^{(3)}$ – плотности населения, чел. на м^2; $x^{(4)}$ – процента грамотных, %; $x^{(5)}$ – рождаемости на 100 чел., чел. Исходные данные приведены в табл. 4.1.

Требуется:

1. Записать модель множественного линейного регрессионного анализа признака *Y*, предъявляемые к ней требования и соответствующую функцию регрессии.

2. Рассчитать с помощью программы «Корреляция» матрицу (6×6) оценок коэффициентов парной корреляции между признаками и сделать вывод о силе линейной связи результативного признака с каждым из факторов и о силе линейной связи каждой пары факторов. Найти коллинеарные регрессоры (на практике коллинеарными считаются такие регрессоры, коэффициент корреляции между которыми по модулю больше 0,7).

3. Вычислить оценки параметров модели множественной линейной регрессии с помощью программы «Регрессия» с «Выводом остатка», приняв уровень надежности, равным 95%; записать уравнение регрессии и его стандартную ошибку *(s);* используя «Остатки», вычислить среднюю относительную ошибку аппроксимации *A*.

4. Предположив выполнение условий линейного регрессионного анализа:

а) оценить статистическую значимость уравнения регрессии;

б) оценить статистическую значимость коэффициентов уравнения регрессии.

При наличии в уравнении регрессии хотя бы одного незначимого коэффициента исключить тот регрессор, при котором коэффициент незначим. Выполнить пп. 3 – 4 с оставшимися регрессорами. Процедуру пошагового исключения регрессоров продолжать до тех пор, пока не будет получено значимое уравнение регрессии со значимыми коэффициентами.

Замечание. Если после исключения регрессора уравнение становится статистически незначимым или остается значимым, но его нормированный *R*-квадрат значительно уменьшается, то этот регрессор следует «возвратить» в уравнение и исключить очередной регрессор, коэффициент при котором незначим (конечно, при наличии такого регрессора).

Систематизировать результаты пошаговой регрессии, выписав для каждого шага:

– уравнение регрессии;

– коэффициент детерминации R^2; стандартную ошибку *s*, среднюю относительную ошибку аппроксимации *A*, наблюдаемое значение *F*-статистики Фишера-Снедекора и критическое значение критерия;

– под оценками параметров – 95%-е доверительные интервалы для этих параметров;·под доверительными интервалами – числовые значения *t*-статистик.

6. Выбрать лучшее уравнение и, используя его, ответить на следующие вопросы:

а) какой процент выборочной дисперсии признака *Y* обусловлен линейным влиянием включенных в уравнение регрессоров?

б) увеличение какого регрессора на единицу его измерения (при неизменных значениях других регрессоров) ведет к наибольшему изменению среднего значения результативного признака?

в) увеличение среднего значения какого фактора на 1% (по отношению к его среднему значению) при неизменных значениях других факторов ведет к наибольшему процентному изменению среднего значения результативного признака (по отношению к его среднему значению)?

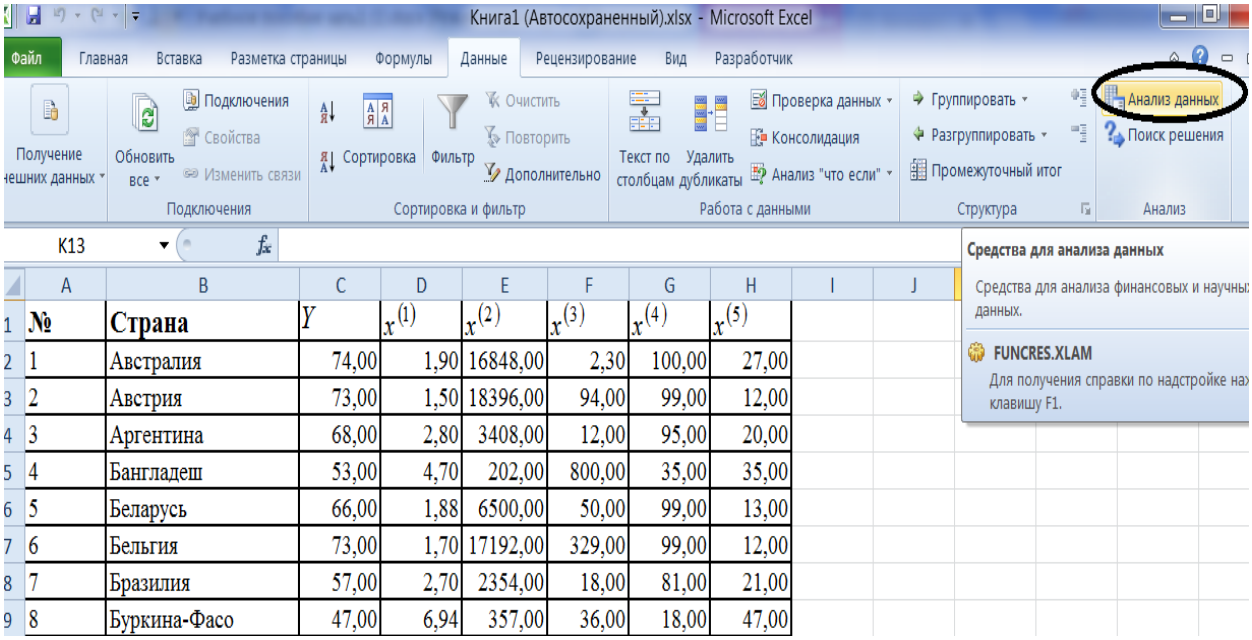
Таблица 4.1

№	Страна	Y	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
1	Австралия	74	1,9	16 848	2,3	100	27
2	Австрия	73	1,50	18 396	94,0	99	12
3	Аргентина	68	2,80	3 408	12,0	95	20
4	Бангладеш	53	4,70	202	800,0	35	35
5	Беларусь	66	1,88	6 500	50,0	99	13
6	Бельгия	73	1,70	17 192	329,0	99	12
7	Бразилия	57	2,70	2 354	18,0	81	21
8	Буркина-Фасо	47	6,94	357	36,0	18	47
9	Великобритания	74	1,83	15 974	237,0	99	13
10	Вьетнам	63	3,33	230	218,0	88	27
11	Гаити	43	5,94	383	231,0	53	40
12	Германия	73	1,47	17 539	227,0	99	11
13	Гондурас	65	4,90	1 030	46,0	73	35
14	Гонконг	75	1,40	14 641	5494,0	77	13
15	Египет	60	3,77	748	57,0	48	29
16	Замбия	44	6,68	573	11,0	73	46
17	Индия	58	4,48	275	283,0	52	29
18	Ирландия	73	1,99	12 170	51,0	98	14
19	Испания	74	1,40	13 047	77,0	95	11
20	Италия	74	1,30	17 500	188,0	97	11
21	Канада	74	1,80	19 904	2,8	97	14
22	Китай	67	1,84	377	124,0	78	21
23	Колумбия	69	2,47	1 538	31,0	87	24
24	Коста-Рика	76	3,10	2 031	64,0	93	26
25	Куба	74	1,90	1 382	99,0	94	17
26	Малайзия	66	3,51	2 995	58,0	78	29
27	Марокко	66	3,83	1 062	63,0	50	29
28	Мексика	69	3,20	3 604	46,0	87	28
29	Нидерланды	75	1,58	17 245	366,0	99	13
30	Новая Зеландия	73	2,03	14 381	13,0	99	16
31	Норвегия	74	2,00	17 755	11,0	99	13
32	ОАЭ	70	4,50	14 193	32,0	68	28
33	Польша	69	1,94	4 429	123,0	99	14
34	Португалия	71	1,50	9 000	108,0	85	12
35	Россия	64	1,83	6 680	8,8	99	13
36	Саудовская Аравия	66	6,67	6 651	7,7	62	38
37	Северная Корея	67	2,40	1 000	189,0	99	24
38	Сингапур	73	1,88	14 990	4 456,0	88	16
39	США	73	2,06	23 474	26,0	97	15
40	Таиланд	65	2,10	1 800	115,0	93	19
41	Турция	69	3,21	3 721	79,0	81	26
42	Украина	65	1,82	2 340	87,0	97	12
43	Филиппины	63	3,35	867	221,0	90	27
44	Финляндия	72	1,80	15 877	39,0	100	13
45	Франция	74	1,80	18 944	105,0	99	13
46	Чили	71	2,50	2 591	18,0	93	23
47	Швейцария	75	1,60	22 384	170,0	99	12
48	Швеция	75	2,10	16 900	19,0	99	14
49	Эфиопия	51	6,81	122	47,0	24	45
50	ЮАР	62	4,37	3 128	35,0	76	34
51	Южная Корея	68	1,65	6 627	447,0	96	16
52	Япония	76	1,55	19 860	330,0	99	11

1. Уравнение линейной множественной регрессии ищем в виде

$$\hat{Y} = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \beta_3 x^{(3)} + \beta_4 x^{(4)} + \beta_5 x^{(5)}. \qquad (4.1)$$

2. Для расчета матрицы оценок коэффициентов парной корреляции воспользуемся программой «Корреляция». Для этого в версии MS Excel 2007 и выше выберем соответствующий пункт меню «Данные» → «Анализ данных» (рис. 4.1). Если данный пункт меню отсутствует, то необходимо выбрать меню «Файл» → «Параметры» → «Надстройки» → «Перейти…». В появившемся окне установить флажок напротив «Пакет анализа».



№	Страна	Y	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
1	Австралия	74,00	1,90	16848,00	2,30	100,00	27,00
2	Австрия	73,00	1,50	18396,00	94,00	99,00	12,00
3	Аргентина	68,00	2,80	3408,00	12,00	95,00	20,00
4	Бангладеш	53,00	4,70	202,00	800,00	35,00	35,00
5	Беларусь	66,00	1,88	6500,00	50,00	99,00	13,00
6	Бельгия	73,00	1,70	17192,00	329,00	99,00	12,00
7	Бразилия	57,00	2,70	2354,00	18,00	81,00	21,00
8	Буркина-Фасо	47,00	6,94	357,00	36,00	18,00	47,00

Рис. 4.1. Выбор меню «Анализ данных»

После открытия окна «Анализ данных» выбрать пункт «Корреляция». В открывшемся окне укажем входной интервал C1:H53, в который мы ввели исходные данные (с заголовками столбцов — названиями признаков, поэтому отметим флажок «Метки в первой строке»). Укажем, что данные сгруппированы по столбцам, а результаты работы необходимо вывести на новый рабочий лист.

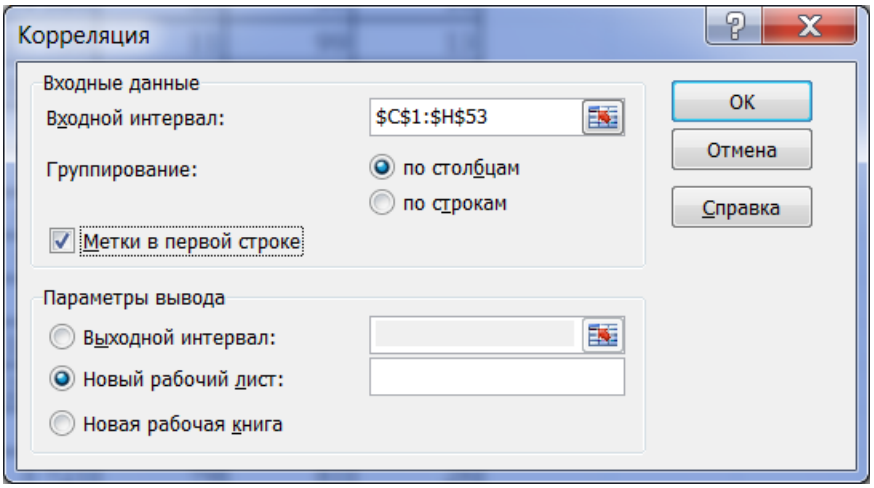
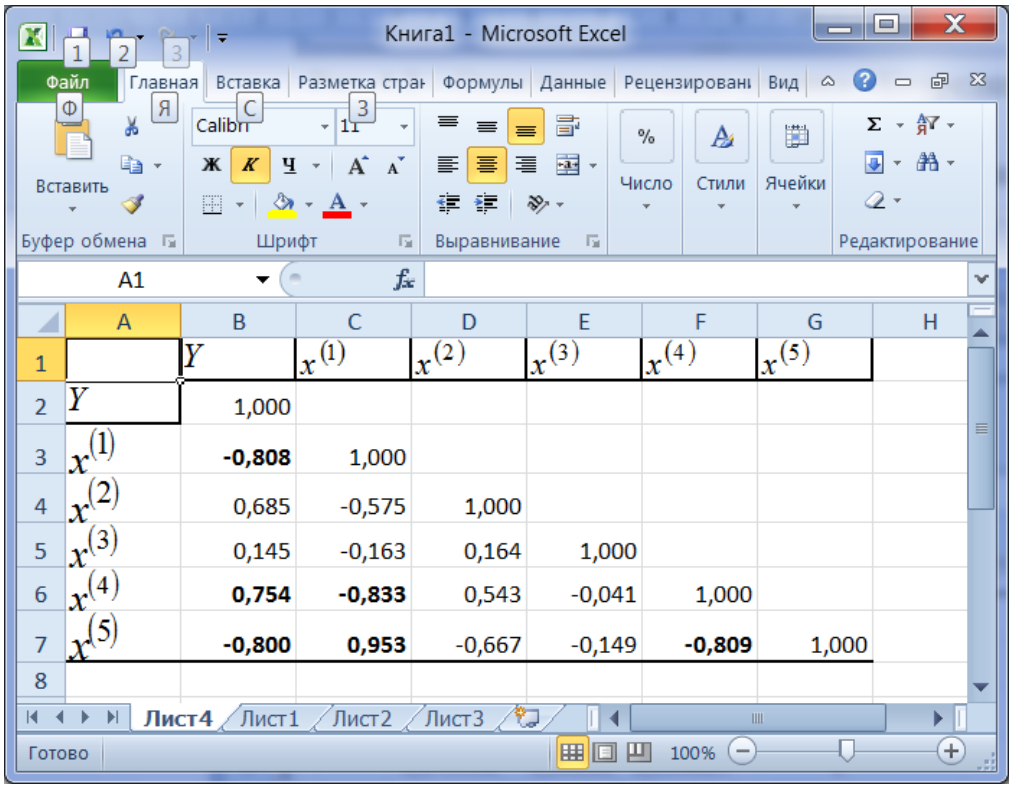


Рис. 4.2. Окно ввода данных сервиса «Корреляция»

Результаты работы сервиса «Корреляция» представлены на рис. 4.3.



	A	B	C	D	E	F	G	H
1		Y	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	
2	Y	1,000						
3	$x^{(1)}$	**-0,808**	1,000					
4	$x^{(2)}$	0,685	-0,575	1,000				
5	$x^{(3)}$	0,145	-0,163	0,164	1,000			
6	$x^{(4)}$	**0,754**	**-0,833**	0,543	-0,041	1,000		
7	$x^{(5)}$	**-0,800**	**0,953**	-0,667	-0,149	**-0,809**	1,000	
8								

Рис. 4.3. Результаты работы сервиса «Корреляция»

В результате работы программы «Корреляция» рассчитана матрица оценок коэффициентов парной корреляции (ввиду симметричности этой матрицы в результатах работы программы «Корреляция» приводится только часть матрицы — не выше главной диагонали). Жирным шрифтом выделены коэффициенты корреляции, оценки которых по модулю превосходят 0,7.

На основе анализа матрицы оценок коэффициентов парной корреляции можно сделать следующие выводы. Судя по наблюдениям, наиболее сильна линейная связь результативного признака Y (ожидаемой продолжительности жизни мужчины) с факторным признаком $x^{(1)}$ (средним числом детей в семье), с $x^{(4)}$ (процентом грамотных) и с $x^{(5)}$ (рождаемостью), поскольку модули оценок соответствующих коэффициентов парной корреляции достаточно велики: $\left|r_{Y X_1}\right|$ =0,808, $\left|r_{Y X_4}\right|$ =0,754 и $\left|r_{Y X_5}\right|$ =0,800. Линейная связь Y с $x^{(2)}$ также достаточно сильна: $\left|r_{Y X_2}\right|$= 0,6 84; связь Y с $x^{(3)}$ выражена слабее.

Достаточно сильна линейная связь между каждой парой регрессоров $x^{(1)}$, $x^{(4)}$ и $x^{(5)}$: $\left|r_{X_1 X_4}\right|$ =0,833, $\left|r_{X_1 X_5}\right|$ =0,953, $\left|r_{X_4 X_5}\right|$ =0,809 — это может свидетельствовать о коллинеарности факторов $x^{(1)}$ и $x^{(4)}$, $x^{(1)}$ и $x^{(5)}$, $x^{(4)}$ и $x^{(5)}$. Малые абсолютные значения оценок коэффициентов корреляции между остальными факторами говорят об относительно слабой линейной связи между ними.

3. Рассчитаем оценки $\beta_0, \ldots \beta_5$ параметров модели линейной регрессии и стандартную ошибку регрессии. Для этого воспользуемся программой «Регрессия», выбрав соответствующий пункт меню надстройки «Анализ данных» Microsoft Excel.

В окне ввода исходных данных программы «Регрессия» (рис. 4.4) укажем входные интервалы результативного признака Y (C1:C53) и факторных признаков $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, $x^{(4)}$, $x^{(5)}$ (D1:H53). Установим флажок «Метки» (указав, что в первой строке находятся названия переменных), очистим флажок «Константа — ноль» (чтобы в уравнении присутствовал свободный член a_0), уровень надежности $(1-\alpha)$ указывать не будем (по умолчанию он равен 95%). Укажем, что результаты работы программы необходимо вывести на новый рабочий лист. Укажем также, что необходимо вывести остатки.

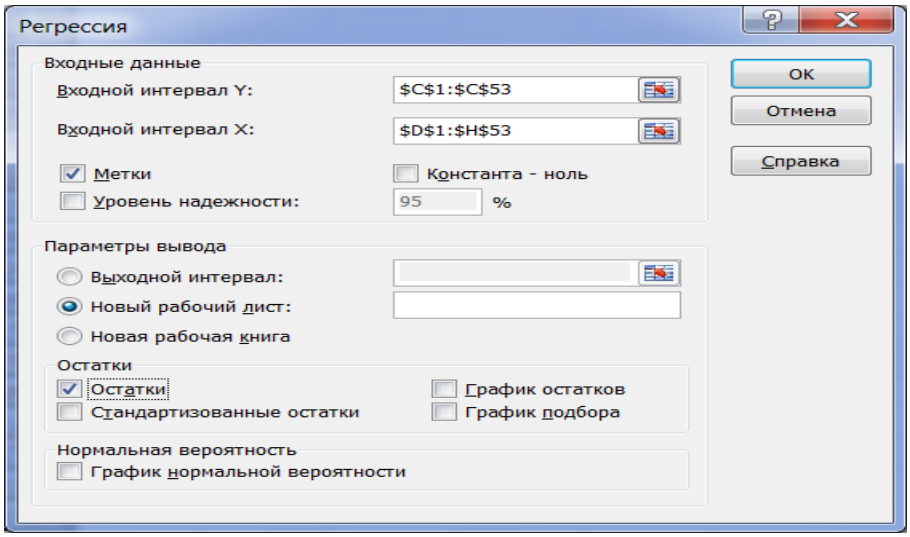


Рис. 4.4. Окно ввода данных сервиса «Регрессия»

Итог работы сервиса «Регрессия» представлен на рис. 4.5.

	A	B	C	D	E	F	G	H	I
1	**ВЫВОД ИТОГОВ**								
2									
3	*рессионная статистика*								
4	Множеств	0,86012							
5	R-квадрат	0,739807							
6	Нормиров	0,711525							
7	Стандартн	4,387314							
8	Наблюден	52							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*начимость F*			
12	Регрессия	5	2517,549	503,5097	26,15836	2,08E-12	2,417356		
13	Остаток	46	885,4321	19,24852					
14	Итого	51	3402,981						
15									
16	*Коэффициен*	*артная ош*	*татистик*	*P-Значение*	*ижние 95%*	*ерхние 95%*	*ижние 95,0*	*Верхние 95,0%*	
17	Y-пересеч	62,99303	7,080191	8,89708	1,46E-11	48,74134	77,24471	48,74134	77,24471359
18		-2,96149	1,48305	-1,99689	0,051774	-5,94671	0,023736	-5,94671	0,023736414
19		0,000351	0,000114	3,089081	0,003401	0,000122	0,00058	0,000122	0,000580254
20		0,00025	0,000695	0,359001	0,721237	-0,00115	0,001649	-0,00115	0,001648823
21		0,087079	0,058063	1,499751	0,140512	-0,02979	0,203953	-0,02979	0,203953352
22		0,116359	0,229286	0,507483	0,614241	-0,34517	0,577887	-0,34517	0,577887187
23									
24									
25									
26	ВЫВОД ОСТАТКА								
27									
28	*аблюдени*	*едсказанн*	*Остатки*						

Рис. 4.5. Результаты работы сервиса «Регрессия»

Оценки $\beta_0 = 62{,}993$, $\beta_1 = -2{,}961$, $\beta_2 = 0{,}00035$, $\beta_3 = 0{,}00025$, $\beta_4 = 0{,}087$, $\beta_5 = 0{,}116$ параметров содержатся в результатах работы программы «Регрессия» (рис. 4.5) в столбце «Коэффициенты» под заголовками «Y-пересечение» и ниже соответственно. Таким образом, оценка линейной функции регрессии такова:

$$\hat{Y} = 62{,}993 - 2{,}961x^{(1)} + 0{,}00035x^{(2)} + 0{,}00025x^{(3)} + 0{,}087x^{(4)} + 0{,}116x^{(5)}. \qquad (4.2)$$

В таблице «Вывод остатка», фрагмент которой приведен на рис. 4.5, содержится предсказанное Y — это $\hat{y}_i$, рассчитанные по построенному уравнению регрессии, и остатки — это разности $(y_i - \hat{y}_i)$. Зная эти остатки, можно, например, рассчитать среднюю относительную ошибку аппроксимации. Для этого необходимо предварительно вычислить $\left|\frac{y_i - \hat{y}_i}{y_i}\right|$, затем их сумму, а также найти среднее значение результативного признака $\bar{y}$. Для нашего примера имеем $\sum_{i=1}^{52}\left|\frac{y_i - \hat{y}_i}{y_i}\right| = 2{,}3695$. Тогда $A = \frac{100\%}{52} \cdot 2{,}3695 = 4{,}56\%$.

В таблице «Регрессионная статистика» (рис. 4.5) приведены:

– коэффициент множественной линейной детерминации $R^2 = 0{,}74$, то есть 74% вариации продолжительности жизни мужчины обусловлено линейным влиянием на нее среднего числа детей в семье, величины ВВП на душу населения, плотности населения, процента грамотных и рождаемости;

– коэффициент множественной линейной корреляции $R = 0{,}86$, то есть такова, судя по наблюдениям, степень линейной зависимости Y от $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, $x^{(4)}$, $x^{(5)}$;

– стандартная ошибка регрессии (оценка среднего квадратического отклонения) $s = 4{,}39$.

В таблице «Дисперсионный анализ» (в результатах работы программы «Регрессия» на рис. 4.5) в столбце «*df*» приводятся количества степеней свобо-

ды m = 5, n - m = 46, n - 1 = 51 соответственно случайных величин, , $\sum(\hat{y}-\bar{y})^2$, $\sum(y-\hat{y})^2$, $\sum(y-\bar{y})^2$ (см. табл. 2.3), значения которых, равные соответственно 2517,55, 885,43 и 3402,98, приводятся в столбце «*SS*»; а в столбце «*MS*» приведены значения величин $S^2_{факт}=\frac{1}{m}\sum(\hat{y}-\bar{y})^2$, $S^2_{ост}=\frac{1}{n-m-1}\sum(y-\hat{y})^2$, равные соответственно 503,51 и 19,23.

4. Проверим гипотезу о значимости коэффициента детерминации и уравнения регрессии в целом. Выдвигаем гипотезу $H_0: R^2=0$ и альтернативную ей $H_1: R^2\neq 0$. Наблюдаемое значение статистики $F_{набл}=\frac{R^2}{1-R^2}\cdot\frac{m}{n-m-1}=\frac{S^2_{факт}}{S^2_{ост}}=\frac{503{,}51}{19{,}23}=26{,}18$. Критическое значение $F_{кр}=F(\alpha=0{,}05;k_1=5;k_2=46)=2{,}4$. Так как $F_{набл}>F_{кр}$, то принимается гипотеза H_1, то есть коэффициент и уравнение в целом статистически значимы. Значение критической точки можно найти в Microsoft Excel 2010, воспользовавшись функцей =**F.ОБР.ПХ**(α, k_1, k_2) (В Microsoft Excel 2007 это функция **=FРАСПРОБР**). Проверить гипотезу H_0 можно и так: если в таблице «Дисперсионный анализ» в столбце «Значимость F» значение оказывается больше принятого уровня значимости α, то гипотезу H_0 принимают и говорят, что уравнение статистически незначимо. В нашем примере «Значимость F» равна $2{,}08\cdot10^{-12}$, что меньше $\alpha=0{,}05$, а значит уравнение значимо.

Проверим гипотезы о значимости параметров уравнения регрессии $H_0^{(j)}: \beta_j=0$ при альтернативах $H_1^{(j)}: \beta_j\neq 0$, $j=\overline{1,5}$.

В таблице (в результатах работы программы «Регрессия» на рис. 4.5) в столбце «t-статистика» приводятся значения статистики $t_{\beta_j}=\frac{\beta_j}{s_{\beta_j}}$, которая при

выполнении гипотезы имеет распределение Стьюдента с (n - m - 1) степенью свободы (см. п. 3.1).

В задаче значение статистики t_{β_1} = -1,99, статистики t_{β_2} = 3,09, статистики t_{β_3} = 0,36, статистики t_{β_4} = 1,50, статистики t_{β_5} = 0,51. Так как критическая точка $t_{кр}(\alpha = 0{,}05; k = 46)$, то только гипотеза $H_0^{(2)}: \beta_2 = 0$ отвергается (оценка параметра β_2 значима), а гипотезы $H_0^{(1)}: \beta_1 = 0$, $H_0^{(3)}: \beta_3 = 0$, $H_0^{(4)}: \beta_4 = 0$, $H_0^{(5)}: \beta_5 = 0$ не отвергаются (оценки параметров β_1, β_3, β_4, β_5 незначимы). Критическое значение распределения Стьюдента можно найти, использовав функцию Microsoft Excel 2010 **=СТЬЮДЕНТ.ОБР.2Х.**

В той же таблице в столбце «Р-значение» приводятся рассчитанные уровни значимости гипотез $H_0^{(j)}$ – это вероятности $p_j = 2P\{t_{\beta_j} > t_{кр}\}$. Так как p_1 = 0,051, p_2 = 0,003, p_3 = 0,721, p_4 = 0,141, p_5 = 0,614, то только гипотеза $H_0^{(2)}: \beta_2 = 0$ отвергается (гипотеза $H_0^{(j)}$ отвергается, если $p_j < \alpha$).

Эти же гипотезы можно проверить при помощи интервальных оценок параметров уравнения регрессии. Все в той же таблице в столбцах «Нижние 95%» и «Верхние 95%» приводятся нижние и верхние границы интервальных оценок параметров β_1, β_2, β_3, β_4, β_5.

Поскольку только в интервал (0,0001; 0,0006) не попадает нуль, то только гипотеза $H_0^{(2)}: \beta_2 = 0$ отвергаются, а гипотезы $H_0^{(1)}: \beta_1 = 0$, $H_0^{(3)}: \beta_3 = 0$, $H_0^{(4)}: \beta_4 = 0$, $H_0^{(5)}: \beta_5 = 0$ не отвергаются.

Таким образом, в построенном уравнении регрессии большинство коэффициентов оказались незначимы, и такое уравнение нельзя считать приемлемым.

5. Исключим из уравнения фактор $x^{(3)}$, при котором коэффициент незначим, а соответствующая этому коэффициенту абсолютная величина значения статистики t_{β_3}, равного 0,36, является наименьшей (рассчитанный уровень

значимости $p_3 = 0{,}721$ является наибольшим). С помощью сервиса «Регрессия» получаем следующий результат (рис.4.6):

	A	B	C	D	E	F	G	H	I	J
1	ВЫВОД ИТОГОВ									
2										
3	*рессионная статистика*									
4	Множеств	0,859696								
5	R-квадрат	0,739078								
6	Нормиров	0,716872								
7	Стандартн	4,346466								
8	Наблюден	52								
9										
10	Дисперсионный анализ									
11		*df*	*SS*	*MS*	*F*	*начимость F*				
12	Регрессия	4	2515,068	628,767	33,2826	3,57E-13	2,56954			
13	Остаток	47	887,9129	18,89176						
14	Итого	51	3402,981							
15										
16	*Коэффициен*	*артная ош*	*татистик*	*-Значение*	*ижние 95%*	*ерхние 95%*	*ижние 95,0*	*рхние 95,0%*		
17	Y-пересеч	63,83129	6,621849	9,639496	1,03E-12	50,50985	77,15273	50,50985	77,15273	
18		-3,07973	1,43255	-2,14982	0,036747	-5,96165	-0,19781	-5,96165	-0,19781	
19		0,000358	0,000111	3,214244	0,002366	0,000134	0,000582	0,000134	0,000582	
20		0,079976	0,054079	1,478879	0,145844	-0,02882	0,188768	-0,02882	0,188768	
21		0,121854	0,226644	0,537644	0,59336	-0,3341	0,577804	-0,3341	0,577804	
22										
23										
24										
25	ВЫВОД ОСТАТКА									

Рис. 4.6. Результат работы сервиса «Регрессия» с исключенным фактором $x^{(3)}$

Тогда оценка функции регрессии следующая:

$$\hat{Y} = 63{,}831 - 3{,}080x^{(1)} + 0{,}00036x^{(2)} + 0{,}080x^{(4)} + 0{,}122x^{(5)}. \qquad (4.3)$$

Стандартная ошибка $s = 4{,}36$, средняя относительная ошибка аппроксимации $A = 4{,}57\%$, оценка коэффициента множественной линейной корреляции равна 0,86, оценка коэффициента множественной линейной детерминации равна 0,74.

Так как $p_1 = 0{,}037$, $p_2 = 0{,}002$, $p_4 = 0{,}146$, $p_5 = 0{,}593$, то только гипотеза $H_0^{(2)} : \beta_2 = 0$ отвергается, а гипотезы $H_0^{(1)} : \beta_1 = 0$, $H_0^{(4)} : \beta_4 = 0$, $H_0^{(5)} : \beta_5 = 0$ не

отвергаются. Гипотеза $H_0 : R^2 = 0$ отвергается на 5%-м уровне значимости, так как «значимость F», равная $3{,}57 \cdot 10^{-13}$ меньше принятого уровня значимости $\alpha = 0{,}05$.

Теперь исключим из уравнения фактор $x^{(5)}$, при котором коэффициент незначим, а соответствующая этому коэффициенту абсолютная величина статистики t_{β_5}, равная 0,53, является наименьшей (рассчитанный уровень значимости $p_5 =$ 0,593 является наибольшим). Тогда оценка функции регрессии следующая:

$$\hat{Y} = 64{,}875 - 2{,}416x^{(1)} + 0{,}00033x^{(2)} + 0{,}079x^{(4)}. \qquad (4.4)$$

Стандартная ошибка s = 4,31, средняя относительная ошибка аппроксимации $A = 4{,}59\%$, оценка коэффициента множественной линейной корреляции равна 0,86, оценка коэффициента множественной линейной детерминации равна 0,74.

Так как $p_1 =$ 0,0016, $p_2 =$ 0,0014, $p_4 =$ 0,146, то только гипотеза $H_0^{(4)} : \beta_4 = 0$ отвергается, а гипотезы $H_0^{(1)} : \beta_1 = 0$, $H_0^{(2)} : \beta_2 = 0$ не отвергаются. Гипотеза $H_0 : R^2 = 0$ отвергается на 5%-м уровне значимости, так как «значимость F», равная $5{,}58 \cdot 10^{-13}$, меньше принятого уровня значимости $\alpha = 0{,}05$.

Исключим далее из уравнения фактор $x^{(4)}$, при котором коэффициент незначим, а соответствующая этому коэффициенту абсолютная величина статистики t_{β_4}, равная 1,48, является наименьшей (рассчитанный уровень значимости $p_4 =$ 0,146 является наибольшим). Тогда оценка функции регрессии следующая:

$$\hat{Y} = 73{,}671 - 3{,}223x^{(1)} + 0{,}00035x^{(2)}. \qquad (4.5)$$

Стандартная ошибка s = 4,37, средняя относительная ошибка аппроксимации $A = 4{,}59\%$, оценка коэффициента множественной линейной корреляции

равна 0,85, оценка коэффициента множественной линейной детерминации равна 0,73.

Так как $p_1 = 1{,}51 \cdot 10^{-8}$, $p_2 = 0{,}00077$, то гипотезы $H_0^{(1)} : \beta_1 = 0$, $H_0^{(2)} : \beta_2 = 0$ отвергаются. Гипотеза $H_0 : R^2 = 0$ отвергается на 5%-м уровне значимости, так как «значимость F», равная $1{,}75 \cdot 10^{-14}$, меньше принятого уровня значимости $\alpha = 0{,}05$.

Все полученные результаты сведем в табл. 4.2.

Таблица 4.2

№ п/п	Уравнение Интервальные оценки коэффициентов Наблюдаемое значение критерия Стьюдента	s	R^2	A, %	$F_{набл}$	$F_{кр}$
1	$\hat{Y} = 62{,}993 - 2{,}961x^{(1)} + 0{,}00035x^{(2)} + 0{,}00025x^{(3)} + + 0{,}087x^{(4)} + 0{,}116x^{(5)}$ (48,74; 77,24) (−5,95; 0,02) (0,0012; 0,00058) (−0,001; 0,002) (−0,029; 0,204) (−0,345; 0,578) -1,99 3,09 0,36 1,50 0,51	4,39	0,74	4,56	26,18	2,417
2	$\hat{Y} = 63{,}831 - 3{,}080x^{(1)} + 0{,}00036x^{(2)} + + 0{,}080x^{(4)} + 0{,}122x^{(5)}$ (50,51; 77,15) (−5,96; −0,19) (0,0013; 0,00058) (−0,028; 0,189) (−0,334; 0,577) -2,14 3,21 1,47 0,54	4,36	0,74	4,57	33,28	2,570
3	$\hat{Y} = 64{,}875 - 2{,}416x^{(1)} + 0{,}00033x^{(2)} + 0{,}079x^{(4)}$ (52,24; 77,51) (−3,87; −0,97) (0,0013; 0,00052) (−0,029; 0,187) -3,35 3,38 1,48	4,31	0,74	4,59	44,95	2,798
4	$\hat{Y} = 73{,}671 - 3{,}223x^{(1)} + 0{,}00035x^{(2)}$ (69,58; 77,75) (−4,42; −2,26) (0,0015; 0,00055) -6,77 3,59	4,37	0,73	4,59	64,77	3,187

6. Таким образом, наилучшим уравнением является полученное на четвертом шаге, поскольку и само уравнение, и все его коэффициенты значимы. В уравнение оказались включены факторы $x^{(1)}$ и $x^{(2)}$, линейная связь между которыми невелика $\left|r_{X_1 X_2}\right|$ = 0,575. По значению коэффициента детерминации можно сделать вывод, что 73% дисперсии продолжительности жизни мужчины связаны с линейным влиянием среднего числа детей в семье и ВВП на душу населения. Увеличение среднего числа детей на единицу (при неизменном значении $x^{(2)}$) влечет уменьшение в среднем жизни мужчины на 3,223 года. Влияние изменения ВВП на продолжительность жизни мужчины несущественно.

Кроме того, можно вычислить коэффициенты эластичности по формуле (3.10): $\overline{Э}_1 = \beta_1 \cdot \dfrac{\overline{x^{(1)}}}{\bar{y}} = -3{,}223 \cdot \dfrac{2{,}83}{67{,}48} = -0{,}135;$

$$\overline{Э}_2 = \beta_2 \cdot \frac{\overline{x^{(2)}}}{\bar{y}} = 0{,}00035 \cdot \frac{8408{,}06}{67{,}48} = 0{,}044.$$

Анализ коэффициентов эластичности показывает, что увеличение среднего числа детей в семье на 1% (при неизменном значении $x^{(2)}$) сопровождается уменьшением средней продолжительности жизни мужчины на 0,135%; увеличение среднего ВВП на душу населения на 1% влечет увеличение средней продолжительности жизни мужчины на 0,044%.

Приложение А

Таблица значений интегральной функции Лапласа $\Phi(x) = \frac{1}{2\sqrt{\pi}} \int_0^x e^{-\frac{x^2}{2}}$ **dx**

x	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*
0,00	0,0000	0,44	0,1700	0,88	0,3106	1,32	0,4066	1,76	0,4608	2,40	0,4918
0,01	0,0040	0,45	0,1736	0,89	0,3133	1,33	0,4082	1,77	0,4616	2,42	0,4922
0,02	0,0080	0,46	0,1772	0,90	0,3159	1,34	0,4099	1,78	0,4625	2,44	0,4927
0,03	0,0120	0,47	0,1808	0,91	0,3186	1,35	0,4115	1,79	0,4633	2,46	0,4931
0,04	0,0160	0,48	0,1844	0,92	0,3212	1,36	0,4131	1,80	0,4641	2,48	0,4934
0,05	0,0199	0,49	0,1879	0,93	0,3238	1,37	0,4147	1,81	0,4649	2,50	0,4938
0,06	0,0239	0,50	0,1915	0,94	0,3264	1,38	0,4162	1,82	0,4656	2,52	0,4941
0,07	0,0279	0,51	0,1950	0,95	0,3289	1,39	0,4177	1,83	0,4664	2,54	0,4945
0,08	0,0319	0,52	0,1985	0,96	0,3315	1,40	0,4192	1,84	0,4671	2,56	0,4948
0,09	0,0359	0,53	0,2019	0,97	0,3340	1,41	0,4207	1,85	0,4678	2,58	0,4951
0,10	0,0398	0,54	0,2054	0,98	0,3365	1,42	0,4222	1,86	0,4686	2,60	0,4953
0,11	0,0438	0,55	0,2088	0,99	0,3389	1,43	0,4236	1,87	0,4693	2,62	0,4956
0,12	0,0478	0,56	0,2123	1,00	0,3413	1,44	0,4251	1,88	0,4699	2,64	0,4959
0,13	0,0517	0,57	0,2157	1,01	0,3438	1,45	0,4265	1,89	0,4706	2,66	0,4961
0,14	0,0557	0,58	0,2190	1,02	0,3461	1,46	0,4279	1,90	0,4713	2,68	0,4963
0,15	0,0596	0,59	0,2224	1,03	0,3485	1,47	0,4292	1,91	0,4719	2,70	0,4965
0,16	0,0636	0,60	0,2257	1,04	0,3508	1,48	0,4306	1,92	0,4726	2,72	0,4967
0,17	0,0675	0,61	0,2291	1,05	0,3531	1,49	0,4319	1,93	0,4732	2,74	0,4969
0,18	0,0714	0,62	0,2324	1,06	0,3554	1,50	0,4332	1,94	0,4738	2,76	0,4971
0,19	0,0753	0,63	0,2357	1,07	0,3577	1,51	0,4345	1,95	0,4744	2,78	0,4973
0,20	0,0793	0,64	0,2389	1,08	0,3599	1,52	0,4357	1,96	0,4750	2,80	0,4974
0,21	0,0832	0,65	0,2422	1,09	0,3621	1,53	0,4370	1,97	0,4756	2,82	0,4976
0,22	0,00871	0,66	0,2454	1,10	0,3643	1,54	0,4382	1,98	0,4761	2,84	0,4977
0,23	0,0910	0,67	0,2486	1,11	0,3665	1,55	0,4394	1,99	0,4767	2,86	0,4979
0,24	0,0948	0,68	0,2517	1,12	0,3686	1,56	0,4406	2,00	0,4772	2,88	0,4980
0,25	0,0987	0,69	0,2549	1,13	0,3708	1,57	0,4418	2,02	0,4783	2,90	0,4981
0,26	0,1026	0,70	0,2580	1,14	0,3729	1,58	0,4429	2,04	0,4793	2,92	0,4982
0,27	0,1064	0,71	0,2611	1,15	0,3749	1,59	0,4441	2,06	0,4803	2,94	0,4984
0,28	0,1103	0,72	0,2642	1,16	0,3770	1,60	0,4452	2,08	0,4812	2,96	0,4985
0,29	0,1141	0,73	0,2673	1,17	0,3790	1,61	0,4463	2,10	0,4821	2,98	0,4986
0,30	0,1179	0,74	0,2703	1,18	0,3810	1,62	0,4474	2,12	0,4830	3,00	0,49865
0,31	0,1217	0,75	0,2734	1,19	0,3830	1,63	0,4484	2,14	0,4838	3,20	0,49931
0,32	0,1255	0,76	0,2764	1,20	0,3849	1,64	0,4495	2,16	0,4846	3,40	0,49966
0,33	0,1293	0,77	0,2794	1,21	0,3869	1,65	0,4505	2,18	0,4854	3,60	0,499 841
0,34	0,1331	0,78	0,2823	1,22	0,3883	1,66	0,4515	2,20	0,4861		
0,35	0,1368	0,79	0,2852	1,23	0,3907	1,67	0,4525	2,22	0,868	3,80	0,4999 28
0,36	0,1406	0,80	0,2881	1,24	0,3925	1,68	0,4535	2,24	0,4875		
0,37	0,1443	0,81	0,2910	1,25	0,3944	1,69	0,4545	2,26	0,4881	4,00	0,4999 68
0,38	0,1480	0,82	0,2939	1,26	0,3962	1,70	0,4554	2,28	0,4887		
0,39	0,1517	0,83	0,2967	1,27	0,3980	1,71	0,4564	2,30	0,4893	4,50	0,4999 97
0,40	0,1554	0,84	0,2995	1,28	0,3997	1,72	0,4573	2,32	0,4898		
0,41	0,1591	0,85	0,3023	1,29	0,4015	1,73	0,4582	2,34	0,4904	5,00	0,4999 97
0,42	0,1628	0,86	0,3051	1,30	0,4032	1,74	0,4591	2,36	0,4909		
0,43	0,1664	0,87	0,3078	1,31	0,4049	1,75	0,4599	2,38	0,4913	∞	0,5

Приложение Б

Критические точки распределения Стьюдента

Число степеней свободы k	**Уровень значимости α (двусторонняя критическая область)**					
	0,10	**0,05**	**0,02**	**0,01**	**0,002**	**0,001**
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	**0,025**	**0,01**	**0,005**	**0,001**	**0,0005**
	Уровень значимости α (односторонняя критическая область)					

Приложение В

Таблица значений $q = q(\gamma; n)$

n	γ			n	γ		
	0,95	0,99	0,999		0,95	0,99	0,999
5	1,37	2,67	5,64	20	0,37	0,58	0,88
6	1,09	2,01	3,88	25	0,32	0,49	0,73
7	0,92	1,62	2,98	30	0,28	0,43	0,63
8	0,80	1,38	2,42	35	0,26	0,38	0,56
9	0,71	1,20	2,06	40	0,24	0,35	0,50
10	0,65	1,08	1,80	45	0,22	0,32	0,46
11	0,59	0,98	1,60	50	0,21	0,30	0,43
12	0,55	0,90	1,45	60	1,188	0,269	0,38
13	0,52	0,83	1,33	70	0,174	0,245	0,34
14	0,48	0,78	1,23	80	0,161	0,226	0,31
15	0,46	0,73	1,15	90	0,151	0,211	0,29
16	0,44	0,70	1,07	100	0,143	0,198	0,27
17	0,42	0,66	1,01	150	0,115	0,160	0,211
18	0,40	0,63	0,96	200	0,099	0,136	0,185
19	0,39	0,60	0,92	250	0,089	0,120	0,162

Приложение Г

Критические точки распределения χ^2

Число степеней свободы *k*	**Уровень значимости α**					
	0,01	**0,025**	**0,05**	**0,95**	**0,975**	**0,99**
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Приложение Д

Критические точки распределения Фишера

(k_1 — число степеней свободы большей дисперсии, k_2 —число степеней свободы меньшей дисперсии)

Уровень значимости α =0,01

k_2	k_1											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	4052	4999	5403	5625	5764	5889	5928	5981	6022	6056	6082	6106
2	98,49	99,01	90,17	99,25	99,33	99,30	99,34	99,36	99,36	99,40	99,41	99,42
3	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,54	14,45	14,37
5	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,96	9,89
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
7	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62	6,54	6,47
8	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,74	5,67
9	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26	5,18	5,11
10	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,78	4,71
11	9,86	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,46	4,40
12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,22	4,16
13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,45

Уровень значимости α =0,05

k_2	k_1											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,5	19,00	19,16	19,25	19:30	19,33	19,36	19,37	19,38	19,39	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. Гартман, Т.Н. Основы компьютерного моделирования химико-технологических процессов: учеб. пособие / Т.Н. Гартман, Д.В. Клушин. – М.: Академкнига, 2006. – 416 с.
2. Лабутин, А.Н.Методы оптимизации химико-технологических процессов: учеб. пособие / А.Н. Лабутин, Л.С. Гордеев; Иван. хим.-технол. ин-т. – Иваново, 1983. – 78 с.
3. Математическое моделирование и оптимизация химико-технологических процессов: практическое руководство / [В.А. Холоднов и др.]. –СПб.: АНО НПО «Профессионал», 2003. – 480 с.
4. Бояринов, А.И., Кафаров, В.В. Методы оптимизации в химической технологии / А.И. Бояринов, В.В. Кафаров. – М.: Химия. 1975. – 576 с.
5. Кафаров, В.В. Методы кибернетики в химии и химической технологии / В.В. Кафаров. – М.: Химия, 1985. – 468 с.
6. Гмурман, В.Е. Руководство к решению задач по теории вероятностей и математической статистике: учеб. пособие / В.Е. Гмурман. – М.: Юрайт, 2016. – 404 с.
7. Бочкарев, В.В. Оптимизация химико-технологических процессов: учеб. пособие / В.В. Бочкарев; Томский политехнический университет. – Томск: Изд-во Томского политехнического университета, 2014. – 264 с.
8. Ефремов, Г.И. Моделирование химико-технологических процессов: учеб. пособие / Г.И. Ефремов. – М.: Инфра-М, 2016. – 256 с.
9. Сидняев, Н.И. Введение в теорию планирования эксперимента: учеб. пособие / Н.И. Сидняев, Н.Т. Вилисова. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2011. – 463 с.
10. Вершинин, В.И. Планирование и математическая обработка результатов химического эксперимента: учеб. пособие / В.И. Вершинин, Н.В. Перцев. – М.: Лань, 2017. – 236 с.

ОГЛАВЛЕНИЕ

Для заметок

Учебное издание

Лысова Марина Александровна

Методы оптимизации и организации энерго- и ресурсосберегающих химико-технологических систем. Вероятностно-статистические модели.

Учебное пособие

Редактор В.Л. Родичева

Подписано в печать 18.06.2018. Формат $60 \times 84\,{}^{1}\!/_{16}$. Бумага писчая.
Усл. печ. л. 5, 35. Уч.-изд.л. 5, 93. Тираж 50 экз. Заказ

Ивановский государственный химико-технологический университет

153000, г. Иваново, Шереметевский пр., 7